

2014



Sciences.ch

Éléments de mathématiques appliquées 3^{ème} édition revue et corrigée

La science au cœur des savoirs, toutes disciplines confondues: des mathématiques à la médecine, de l'astrophysique à l'histoire des sciences...





Le pictogramme qui figure ci-contre mérite une explication. Son objet est d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit et de la connaissance le développement massif du brevetage et du droit à la propriété intellectuelle à tout va.

Toute reproduction partielle ou totale du présent support électronique est fortement recommandée (évitez le papier quand même pour des raisons écologiques évidentes...) sans autorisation de l'auteur.

Le concept de propriété intellectuelle est un crime contre l'humanité...!

Pour les personnes souhaitant citer le présent document dans leurs travaux voici si besoin un exemple inspiré de la norme ISO 690/AFNOR:

ISOZ Vincent, Éléments de mathématiques appliquées, Lausanne, Éd. Privées, 2012, 3^{ème} version, (1^{ère} version 2001, 2^{ème} version 2005), 729 p., 1'897 nb p., PDF, 21×29.7 cm (ISBN 978-2-8399-0932-7)

Nouvelle présentation
Édition Privée Sciences.ch
22 Chemin de Chandieu – 1006 Lausanne – Suisse



Avant d'imprimer, pensez à l'environnement

Le code de la propriété intellectuelle ne s'applique en aucun cas sur le contenu de ce document quels que soient les avis divergents des différents auteurs du présent PDF. Toute représentation ou reproduction du présent document par quelque procédé que ce soit constitue un pas vers l'acquisition de la connaissance motivée par les auteurs au droit de tout à chacun d'accéder à l'information.

ISBN 978-2-8399-0932-7



9 782839 909327

TABLE DES MATIÈRES

Introduction	1
<i>Méthodes</i>	<i>5</i>
Méthode de Descartes	9
<i>Vocabulaire</i>	<i>10</i>
Sur les sciences.....	10
Terminologie.....	12
Science et foi.....	14
Arithmétique.....	17
<i>Théorie de la démonstration.....</i>	<i>17</i>
La crise des fondements	18
Paradoxes	21
Raisonnement hypothético-déductif.....	22
Calcul propositionnel	23
Propositions.....	23
Connecteurs.....	25
Procédures de décision.....	31
Procédures de décisions non axiomatisées.....	32
Procédures de décisions axiomatisées.....	32
Quantificateurs	36
Calculs des prédicats	36
Grammaire	37
Langages	37
Symboles.....	38
Termes	39
Formules	41
Démonstrations.....	44
Règles de démonstration	45
<i>Nombres (scalaires)</i>	<i>59</i>
Bases numériques.....	61
Types de nombres.....	63
Nombres entiers (nombres naturels)	63
Axiomes de Peano.....	65
Nombres pairs, impairs, et parfaits	66
Nombres premiers.....	66
Nombres entiers relatifs	67
Nombres rationnels	69
Nombres irrationnels.....	71
Nombres réels	73
Nombres transfinis	75
Nombres complexes	78
Interprétation géométrique.....	84
Plan de Gauss.....	84
Formule d'Euler.....	86
Vecteur tournant.....	88
Transformations dans le plan	89

Nombres quaternions	94
Nombres algébriques.....	110
Nombres transcendants	111
Nombres abstraits.....	112
Alphabet Grec	112
Domaine de définition.....	113
Opérateurs	120
Relations binaires	120
Égalités.....	121
Comparateurs	122
Relations binaires réflexives, symétriques, antisymétriques.....	124
Classes d'équivalences	125
Lois fondamentales de l'arithmétique.....	128
Addition	129
Soustraction.....	131
Multiplication.....	133
Division.....	136
Polynômes arithmétiques	141
Valeur absolue.....	141
Règles de calcul.....	143
Théorie des nombres.....	152
Principe du bon ordre	152
Principe d'induction.....	153
Divisibilité.....	155
Division euclidienne.....	156
Plus grand commun diviseur (P.G.C.D.).....	161
Algorithme d'Euclide	162
Plus petit commun multiple (P.P.C.M.)	165
Théorème fondamental de l'arithmétique.....	167
Congruences.....	168
Classes de congruences.....	171
Fractions continues.....	174
Théorie des ensembles.....	187
Axiomes de ZF (Zermelo-Frankel)	191
Cardinaux.....	197
Produit cartésien.....	199
Bornes	200
Opérations ensemblistes.....	201
Inclusion.....	201
Intersection.....	202
Réunion/Union	203
Différence.....	205
Différence symétrique.....	205
Produit.....	206
Complémentarité.....	206
Fonctions et applications	208
Loi de composition.....	209
Loi interne/externe	209
Image et noyau	209

Fonction surjective	210
Fonction injective.....	211
Fonction bijective.....	212
Fonction composée.....	213
Théorème de Cantor-Bernstein	213
Structures.....	216
Magma	218
Monoïdes.....	218
Groupes	221
Groupe abélien	221
Groupe cyclique	221
Anneaux	226
Anneau abélien (commutatif).....	227
Anneau intègre	227
Anneau factoriel.....	227
Sous-anneau	229
Corps	229
Corps commutatif.....	229
Espaces vectoriels	231
Sous-espaces vectoriels.....	233
Algèbres	233
Homomorphismes	234
Homomorphisme de magma	234
Homomorphisme de monoïde	234
Homomorphisme d'anneau.....	234
Homomorphisme de groupe	235
Homomorphisme de corps	236
Isomorphisme	237
Endomorphisme	237
Automorphisme.....	237
Idéal.....	237
Probabilités	246
Univers des événements	246
Axiomatique de Kolmogorov	248
Événements disjoints.....	248
Événements équiprobables.....	249
Événements conjoints.....	250
Probabilités conditionnelles	254
Formule des probabilités composées	256
Formules de Bayes	257
Espérance conditionnelle	259
Réseaux bayésiens.....	262
Martingales.....	274
Analyse combinatoire	276
Arrangements simples avec répétition	276
Permutations simples sans répétition	278
Permutations simples avec répétition	279
Arrangements simples sans répétition	280
Combinaisons simples.....	281

Coefficient binomial.....	281
Formule de Pascal	282
Chaînes de Markov.....	283
Types de processus stochastiques.....	284
Matrice de transition.....	284
Statistiques	292
Échantillons.....	294
Moyennes	294
Moyenne arithmétique	295
Médiane.....	297
Moyenne quadratique.....	305
Moyenne harmonique.....	306
Moyenne géométrique.....	306
Moyenne mobile/glissante	307
Moyenne pondérée	309
Moyenne fonctionnelle.....	309
Lissage de Laplace	309
Propriété des moyennes.....	311
Types de variables.....	316
Variables discrètes	316
Fonction de répartition	317
Espérance discrète.....	318
Déviation standard discrète.....	320
Relation de Huyghens.....	323
Variable centrée réduite	325
Covariance discrète.....	326
Espérance et variance de la moyenne (erreur standard et fcp).....	331
Coefficient de corrélation.....	333
Variables continues.....	340
Densité de probabilité	337
Espérance continue	338
Variance continue	338
Postulat fondamental de la statistique	338
Indice de diversité	339
Fonctions de distribution.....	341
Fonction discrète uniforme	342
Fonction de Bernoulli.....	344
Fonction géométrique.....	345
Fonction binomiale.....	348
Fonction binomiale négative	355
Fonction hypergéométrique	359
Fonction multinomiale	365
Fonction de Poisson	372
Fonction de Gauss-Laplace	375
Somme de deux v.a. normales	382
Produit de deux v.a. normales.....	383
Loi normale centrée réduite	385
Droite de Henry.....	386
Diagramme quantile-quantile.....	390

Fonction log-normale	391
Fonction uniforme continue	395
Fonction triangulaire	398
Fonction de Pareto	401
Fonction exponentielle	407
Fonction de Cauchy.....	409
Fonction bêta.....	412
Fonction Gamma.....	415
Fonction de Khi-Deux.....	420
Fonction de Student.....	424
Fonction de Fisher-Snedecor.....	429
Fonctions de Benford	431
Estimateurs de vraisemblance	436
Estimateurs de la loi Normale	437
Estimateurs de la loi de Poisson.....	441
Estimateurs de la loi Binomiale (et géométrique)	442
Estimateurs de la loi de Weibull	443
Estimateurs de la loi Gamma	446
Facteur de correction sur population finie.....	447
Intervalle de confiance.....	450
I.C. sur la moyenne avec variance théorique connue (test Z)	451
I.C. sur la variance avec moyenne théorique connue	455
I.C. sur la variance avec moyenne empirique connue (test du Khi-2)	460
I.C. sur la moyenne avec variance empirique connue (test-T).....	461
Test binomial exact	465
I.C. pour une proportion.....	468
Test de l'égalité de deux proportions.....	471
Test des signes	473
Test de la médiane de Mood	476
Test de Poisson (à un échantillon)	479
Test de Poisson (à deux échantillons)	479
Intervalle de confiance/tolérance/prédiction	481
Loi faible des grands nombres.....	483
Inégalité de Markov	484
Inégalité de Bienaymé-Tchebychev	484
Fonction caractéristique	488
Théorème central limite	492
Tests d'hypothèses et d'adéquation	498
Orientation du test d'hypothèse	501
Puissance d'un test.....	506
Puissance du test Z à 1 échantillon	507
Puissance du test p à 1 et 2 échantillons	510
Analyse de la variance (ANOVA à 1 facteur).....	511
Test de Fisher.....	514
Test-t homoscédastique (test d'égalité de moyennes avec variances égales)	520
Test-t hétéroscédastique (test d'égalité de moyennes avec variances non égales)	521
Analyse de la variance (ANOVA à deux facteurs sans répétition)	525
Analyse de la variance (ANOVA à deux facteurs avec répétition).....	539
ANOVA multifactorielle à mesures répétées.....	545

Test C de Cochran.....	546
Test d'ajustement du Khi-2	549
Test d'ajustement de Kolmogorov-Smirnov.....	553
Test de normalité de Ryan-Joiner	559
Robustesse.....	563
Statistiques de rangs.....	564
Tests de rangs (non paramétriques).....	564
L-Statistiques	568
Test de la somme des rangs de Wilcoxon	569
Test de la somme des rangs signés de Mann-Withney.....	579
Traitement des égalités.....	585
Test de la somme des rangs signés de Wilcoxon pour 1 échantillon	587
Test de la somme des rangs signés de Wilcoxon pour 2 échantillons appariés	590
Test de Kruskal-Wallis.....	592
Test de Friedman.....	596
Statistiques des valeurs extrêmes (basées sur les rangs)	599
Test de l'étendue de Tukey	603
Coefficient de corrélation des rangs de Spearman	606
Calculs d'erreurs	610
Incertitudes relatives et absolues.....	610
Erreurs statistiques	611
Répétabilité	613
Propagation des erreurs	614
Chiffres significatifs.....	615

Algèbre621

Calcul algébrique.....622

Équations et inéquations.....	623
Équations.....	624
Inéquations.....	626
Identités remarquables	631
Triangle de Pascal	633
Binôme de Newton.....	633
Polynômes	635
Division euclidienne des polynômes.....	637
Théorème de factorisation des polynômes	638
Équations diophantiennes.....	640
Polynômes de degré 1	640
Polynômes de degré 2	641
Discriminant.....	642
Relations de Viète	643
Nombre d'or	646
Polynômes de degré 3	646
Polynômes de degré 4	650
Polynômes trigonométriques.....	652
Polynômes cyclotomiques.....	653
Polynômes de Legendre	654

Algèbre ensembliste664

Algèbre et géométrie corporelle	664
---------------------------------------	-----

Groupes cycliques	665
Groupe des racines de l'unité	666
Groupe des transformations	668
Groupe linéaire.....	668
Groupe des transformations affines	668
Groupe spécial linéaire.....	670
Groupe orthogonal	670
Groupe cercle	671
Groupe unitaire	671
Groupe spécial linéaire.....	672
Groupes de symétries	675
Groupe diédral.....	677
Orbite et stabilisateur	678
Groupe des permutations.....	679
Groupe alterné.....	685
Calcul différentiel et intégral	690
Calcul différentiel.....	690
Pente moyenne	690
Dérivée première.....	691
Fonction dérivée.....	691
Point d'inflexion	692
Théorème de Rolle	694
Théorème des accroissements finis.....	695
Règle de l'Hospital	696
Différentielles.....	697
Différentielles partielles.....	699
Différentielle totale exacte	699
Différentielle totale inexacte.....	702
Théorème de Schwarz.....	702
Dérivées usuelles.....	705
Dérivée d'une somme	709
Dérivée d'un produit.....	709
Formule de Leibniz.....	709
Dérivée d'une fonction composée	712
Dérivée d'un quotient	714
Calcul intégral	718
Intégrale définie	718
Intégrale de Riemann	719
Somme de Darboux.....	720
Intégrale indéfinie	724
Théorème fondamental du calcul intégral.....	725
Relation de Chasles.....	730
Intégrale double.....	730
Théorème de Fubini	733
Intégration par changement de variable	734
Jacobien.....	738
Matrice jacobienne.....	738
Intégration par parties	741
Primitives usuelles	742

Fonction de Dirac	764
Fonction Gamma d'Euler.....	765
Expression de la factorielle	767
Constante d'Euler-Mascheroni	768
Intégrales curvilignes	769
Intégrale curviligne d'un champ scalaire.....	769
Intégrale curviligne d'un champ vectoriel	771
Champ conservatif	771
Équations différentielles.....	774
Équations différentielles du 1 ^{er} ordre	775
Équations différentielles d'ordre 1 à variables séparées.....	775
Équations différentielles linéaires (E.D.L).....	775
Équations homogène ESSM.....	776
Méthode du polynôme caractéristique	777
Résolution de l'équation homogène de l'E.D.L à coefficients constants d'ordre 1.....	777
Résolution de l'équation homogène de l'E.D.L à coefficients non constants d'ordre 1.....	778
Résolution de l'équation homogène de l'E.D.L à coefficients non constants d'ordre 2.....	779
Méthode du facteur intégrant (d'Euler)	785
Méthode de séparation des variables.....	787
Méthode de variation de la constante	788
Systèmes d'équations différentielles	790
Méthode régulière des perturbations.....	790
Théore perturbative des équations algébriques	795
Théorie perturbative des équations différentielles	795
Suites et séries.....	806
Suites	806
Suites arithmétiques	807
Suites harmoniques	809
Suites géométriques	810
Suite de Cauchy	811
Suite de Fibonacci.....	814
Séries	815
Série numérique	815
Série de Gauss.....	816
Nombres et polynômes de Bernoulli.....	820
Séries arithmétiques	825
Séries géométriques	826
Fonction zêta et identité d'Euler.....	827
Séries de Taylor et de MacLaurin	831
Série de Maclaurin	832
Série de Taylor.....	833
Développements de Maclaurin usuels.....	836
Séries de Taylor d'une fonction à 2 variables.....	844
Forme quadratique	846
Reste de Lagrange.....	848
Formule de Taylor avec reste intégral.....	850
Séries de Fourier	851
Coefficients de Fourier.....	857
Phénomène de Gibbs.....	869

Puissance d'un signal.....	871
Théorème de Parseval	871
Transformée de Fourier.....	872
Séries de Bessel.....	880
Fonction de Bessel d'ordre 0	880
Fonction de Bessel d'ordre N	886
Équation différentielle de Bessel d'ordre N	886
Critères de convergence	887
Test de l'intégrale	887
Règle d'Alembert.....	888
Règle de Cauchy	888
Théorème de Leibniz.....	889
Convergence absolue	889
Théorème du point fixe	890
Calcul vectoriel	897
Notion de flèche	897
Ensemble de vecteurs.....	898
Pseudo vecteurs.....	899
Multiplication par un scalaire.....	900
Règle de trois	900
Espaces vectoriels	902
Combinaisons linéaires	903
Sous-espaces vectoriels.....	903
Familles génératrices.....	903
Dépendances et indépendances	904
Bases d'un espace vectoriel	905
Angles directeurs.....	906
Dimensions d'un espace vectoriel	907
Prolongements d'une famille libre.....	908
Rang d'une famille finie	909
Sommes directes	909
Espace affine	910
Espace vectoriel euclidien.....	912
Norme d'un vecteur	912
Produit scalaire vectoriel.....	913
Projection orthogonale.....	914
Inégalité de Cauchy-Schwartz	917
Inégalité triangulaire	918
Produit scalaire (général)	919
Produit vectoriel.....	920
Produit mixte.....	925
Espaces vectoriels fonctionnels.....	926
Espaces vectoriels hermitiens	927
Produit hermitien.....	929
Types d'espaces vectoriels	930
Système de coordonnées	930
Système cartésien.....	931
Système sphérique.....	932
Système cylindrique.....	936

Système polaire	938
Opérateurs différentiels	940
Gradient d'un champ scalaire	940
Gradient d'un champ de vecteurs	944
Divergences d'un champ de vecteurs	945
Rotationnels d'un champ de vecteurs	952
Laplaciens d'un champ scalaire	959
Laplacien d'un champ vectoriel	964
Identités	965
Algèbre linéaire	975
Systèmes linéaires	976
Transformations linéaires	979
Matrices	980
Matrice associée	982
Opérations sur les matrices	984
Types de matrice	986
Matrice unité	981
Matrice échelonnée	983
Matrice inversible	986
Matrice transposée	986
Matrice adjointe	987
Matrice hermitique (self-adjointe)	988
Matrice nilpotent	988
Matrice orthogonale	988
Matrice symétrique	989
Matrice anti-symétrique	989
Matrice triangulaire	989
Matrice diagonale	990
Matrice de passage	990
Déterminants	991
Propriétés des déterminants	997
Dérivée d'un déterminant	1003
Inverse d'une matrice	1005
Changements de base	1006
Matrice de passage	1006
Valeurs et vecteur propres	1008
Spectre	1008
Polynômes caractéristique	1009
Matrices de rotation	1010
Théorème spectral	1012
Calcul tensoriel	1021
Tenseur	1021
Notation indicielle	1023
Somme sur plusieurs indices	1025
Symbole de Kronecker	1025
Symbole d'antisymétrie	1026
Métrique et signature	1032
Déterminant de Gram	1035
Composantes contravariantes et covariantes	1040

Opérations dans les bases	1042
Méthode d'orthogonalisation de Schmidt	1043
Changements de bases	1043
Bases réciproques	1044
Tenseurs euclidiens	1046
Tenseurs fondamentaux	1046
Produit tensoriel de deux vecteurs	1047
Espace tensoriel	1049
Combinaisons linéaires de tenseurs	1053
Contractions des indices	1054
Tenseurs particuliers	1057
Tenseur symétrique	1057
Tenseur anti-symétrique	1059
Tenseur fondamental	1062
Coordonnées curvilignes	1063
Repère naturel en coordonnées sphériques	1066
Repère naturel en coordonnées polaires	1067
Repère naturel en coordonnées cylindriques	1068
Symboles de Christoffel	1069
Symboles de 2 ^{ème} espèce	1072
Symboles de 1 ^{ère} espèce	1072
Théorème fondamental de la géométrie riemannienne	1075
Théorème de Ricci	1078
Dérivée de covariante	1079
Identité de Ricci	1081
Tenseur de Riemann-Christoffel	1085
Identités de Bianchi	1088
Tenseur de Ricci	1089
Scalaire de Ricci	1089
Tenseur d'Einstein	1101
Identité de Bianchi contractée	1102
Identité d'Einstein	1102
Tenseur d'Einstein	1102
Calcul spinoriel	1109
Spineur unitaire	1110
Propriétés géométriques	1114
Symétries planes	1115
Rotations	1118
Matrices de Pauli	1121
Produit spinoriel	1122
Propriétés des matrices de Pauli	1123
Théorie des nœuds	1135
Représentation des tresses	1135
Groupe de tresses	1137
Représentation des nœuds	1140
Groupe de nœuds	1143
Nœud de tait	1146
Formalisation mathématique	1149
Invariant du nœud	1151

Support du noeud	1152
Noeud trivial	1153
Noeuds isotopes	1154
Noeuds équivalents	1154
Isotopie	1155
Mouvements de Reidemeister	1155
Représentation planaire	1158
Analyse fonctionnelle	1165
Représentations	1165
Représentation tabulaire	1166
Représentations graphiques	1166
Représentations planes	1167
Représentations spatiales	1169
Représentations vectorielles	1175
Propriétés des représentations graphiques	1177
Représentations analytiques	1183
Domaine naturel de définition	1184
Fonctions	1185
Dépendances	1185
Domaine d'existence	1185
Croissance et décroissance	1185
Périodicité	1185
Parité	1186
Composition	1188
Types de fonctions	1188
Fonction puissance	1188
Fonction exponentielle	1188
Fonction logarithmique	1189
Fonctions trigonométriques	1189
Fonctions polynomiales	1189
Fractions rationnelles	1189
Fonctions algébriques	1189
Fonctions en escalier	1189
Limite et continuité	1190
Asymptotes	1194
Logarithmes	1198
Base népérienne	1200
Fonction exponentielle naturelle	1203
Produit scalaire fonctionnel	1205
Analyse complexe	1213
Applications linéaires	1213
Fonction complexe	1213
Fonctions holomorphes	1221
Théorème de Cauchy-Riemann	1222
Conditions de Cauchy	1222
Orthogonalité des iso-courbes réelles et imaginaires	1227
Transformation conforme	1228
Logarithme complexe	1230
Intégration de fonctions complexes	1231

Théorème de Cauchy.....	1233
Fonctions méromorphes.....	1235
Théorème intégral de Cauchy	1235
Théorème intégral de Cauchy généralisé	1237
Série de Laurent à puissances positives	1237
Convergence d'une série.....	1238
Décomposition en chemins	1246
Chemin inverse	1248
Séries de Laurent.....	1249
Singularités.....	1257
Singularité apparente.....	1258
Singularité essentielle.....	1258
Pôle.....	1259
Théorème des résidus	1260
Pôle à l'infini	1265
Topologie	1270
Espace topologique	1270
Espace de Hausdorff	1271
Espace métrique et distance	1272
Semi-distance	1272
Distance ultramétrique	1273
Distance euclidienne	1274
Distance hölderienne.....	1274
Distance discrète	1275
Distances équivalentes	1275
Fonctions lipchitziennes.....	1276
Ensembles ouverts et fermés.....	1277
Boules.....	1278
Parties.....	1279
Boules généralisées	1281
Diamètre.....	1282
Variétés.....	1283
Variétés différentiables	1284
Théorie de la mesure	1290
Espaces mesurables	1290
Tribu.....	1290
Tribu borélienne.....	1293
Théorème de la classe monotone.....	1296
Géométrie.....	1303
Trigonométrie	1304
Radian	1304
Trigonométrie du cercle	1305
Relations remarquables	1311
Formules de Carnot.....	1313
Formules de Simpson.....	1315
Théorème du cosinus	1316
Théorème du sinus	1317
Trigonométrie hyperbolique.....	1318

Relations remarquables	1323
Trigonométrie sphérique	1324
Formule des cosinus	1326
Relations des sinus	1327
Angle solide.....	1329
Géométrie euclidienne	1337
Objets de la géométrie euclidienne	1337
Dimensions.....	1338
Constructions d'Euclide.....	1343
Droites et segments	1344
Grandeurs de même espèce	1344
Plan.....	1349
Déplacement et retournement	1349
Angles	1350
Angle saillant	1350
Angles adjacents	1350
Angles alternes/internes	1353
Mesure des angles	1353
Unités de mesure des angles	1353
Bissectrice	1354
Triangles	1358
Triangles égaux	1359
Triangles isocèles	1361
Médiatrice	1361
Lieu géométrique	1362
Triangles équilatéraux.....	1363
Triangles rectangles	1364
Triangles rectangles isocèles.....	1365
Inégalités dans les triangles.....	1365
Théorème de Pythagore.....	1367
Théorème de Thalès	1369
Parallélisme	1372
Cercles.....	1374
Axiomatique de Hilbert.....	1374
Axiomes d'associations (A).....	1377
Axiomes d'ordre (O).....	1377
Axiomes de congruence	1378
Axiomes de continuité.....	1378
Axiomes des parallèles.....	1378
Barycentre	1379
Transformations dans le plan	1383
Translation	1383
Homothétie.....	1384
Rotation.....	1387
Réflexion.....	1391
Géométries non-euclidiennes	1397
Géodésique et équation métrique	1398
Équation métrique euclidienne.....	1400
Coordonnées de Gauss	1401

Coordonnées de Riemann	1401
Espaces de Riemann.....	1401
Forme quadratique fondamentale	1403
Géométrie projective.....	1409
Perspective conique (centrale).....	1409
Images de points.....	1411
Exemple avec Adobe Flash.....	1414
Exemple avec WebGL	1425
Images de droites	1428
Perspective affines.....	1435
Perspective cavalière.....	1436
Projection orthogonale	1436
Coordonnées homogènes	1438
Géométrie analytique.....	1445
Coniques	1445
Approche algébrique	1445
Approche géométrique	1455
Foyer	1455
Paramètre de la parabole	1455
Paramètre de l'ellipse	1458
Paramétrisations	1464
Équation du plan	1464
Équation d'une droite	1466
Équation d'un cône	1471
Équation d'une sphère	1472
Équation d'un ellipsoïde	1474
Équation d'un cylindre	1476
Surface de révolution	1477
Géométrie différentielle.....	1486
Courbes paramétrées	1486
Parabole osculatrice	1491
Isoclines	1493
Trièdre de Frenet	1498
Vecteur de courbure	1500
Repère de Frenet	1501
Rayon de courbure	1501
1 ^{ère} formule de Frenet.....	1501
Cercle osculateur.....	1502
2 ^{ème} formule de Frenet.....	1504
3 ^{ème} formule de Frenet.....	1504
Trièdre de Frenet.....	1504
Nappes paramétrées	1510
Métrique d'une surface	1512
Première forme quadratique fondamentale	1512
Régularité d'une surface	1513
Formes géométriques	1522
Surfaces connues	1523
Polygones.....	1523
Polygone croisé.....	1523

Polygone concave	1524
Polygone convexe	1524
Rectangle.....	1527
Carré	1528
Triangle	1529
Triangle quelconque.....	1529
Triangle isocèle	1531
Triangle équilatéral	1532
Triangle rectangle.....	1533
Trapèze.....	1534
Parallélogramme	1534
Losange	1535
Cercle	1535
Ellipse.....	1537
Volumes connus	1539
Polyèdres.....	1540
Parallélépipède.....	1541
Pyramide	1543
Prisme droit.....	1544
Polyèdres réguliers.....	1544
Tétraèdre	1548
Hexaèdre régulier (cube).....	1550
Octaèdre	1551
Icosaèdre	1555
Dodécaèdre	1560
Corps de révolutions	1566
Cylindre.....	1567
Cône	1569
Sphère	1571
Tore	1574
Ellipsoïde	1578
Paraboloïde	1582
Tonneau à section circulaire (vertical).....	1582
<i>Théorie des graphes.....</i>	1591
Définitions.....	1591
Formule d'Euler.....	1592
Ponts de Königsberg.....	1603
Théorème d'Euler	1604
Matrice d'adjacence	1607
Catégories.....	1613
Mécanique	1621
<i>Principes</i>	1622
Systèmes d'unités	1624
MKSC	1625
Analyse dimensionnelle	1627
Notations scientifiques	1629
Temps.....	1631
Longueur	1632

Masse	1633
Énergie	1635
Charge	1636
Distributions.....	1638
Constantes	1639
Constantes universelles	1640
Constantes physiques diverses	1641
Constantes physico-chimiques diverses	1642
Constantes astrophysiques diverses	1642
Constantes de Planck.....	1643
Unités naturelles.....	1644
Cube de Okun	1645
Principes de la physique.....	1646
Principe de causalité.....	1646
Trilemne de Fries	1646
Principe de conservation de l'énergie	1646
Principe de moindre action.....	1648
Principe de Noether.....	1649
Invariance par translation dans l'espace	1651
Invariance par rotation dans l'espace	1652
Invariance par translation dans le temps	1653
Théorème de Noether.....	1654
Principe premier de Curie	1658
Espaces ponctuels.....	1659
Mécanique analytique	1670
Formalisme Lagrangien.....	1671
Coordonnées généralisées et référentiels	1672
Référentiel galiléen	1672
Repère de Copernic/Kepler	1673
Référentiel géocentrique	1673
Repère orthonormé direct.....	1673
Coordonnées généralisées	1674
Principe variationnel	1675
Intégrale d'action.....	1676
Équation d'Euler-Lagrange.....	1676
Action.....	1677
Équations d'Euler-Lagrange	1679
Identité de Beltrami.....	1681
Théorème du calcul variationnel.....	1683
Formalisme canonique	1685
Transformation de Legendre	1686
Hamiltonien.....	1686
Moments canoniques.....	1686
Fonction de Hamilton.....	1687
Équations canoniques du mouvement.....	1690
Crochets de Poisson	1692
Identité de Jacobi	1694
Transformations canoniques	1695

Mécanique classique rationnelle.....	1701
Lois de Newton	1705
Première loi (loi d'inertie)	1705
Deuxième loi (principe fondamental de la dynamique)	1707
Troisième loi (action et réaction)	1709
Conditions d'équilibre	1709
Centre de masse et masse réduite	1709
Théorème du centre de masse	1711
Théorème de Guldin.....	1713
Cinématique	1714
Position.....	1714
Vitesse	1715
Accélération	1716
Vitesse de libération.....	1718
Plan osculateur	1719
Accélération tangentielle	1720
Accélération normale	1720
Force centrifuge	1721
Principe de relativité galiléen	1721
Transformation de Galilée.....	1721
Moment cinétique.....	1724
Théorème du moment cinétique.....	1726
Moment de force	1728
Bras de levier	1728
Statique des forces.....	1731
Balistique.....	1734
Cinématique de rotation	1736
Vitesse angulaire	1736
Accélération angulaire	1738
Figures de Lissajous.....	1741
Travail et énergie.....	1742
Énergie cinétique.....	1743
Moment d'inertie	1743
Rayon de giration.....	1746
Moment d'inertie polaire	1747
Théorème d'Huygens-Steiner.....	1749
Tenseur d'inertie.....	1749
Ellipsoïde d'inertie	1755
Gyroscope	1756
Gyroscope symétrique pesant	1757
Précession	1760
Énergie potentielle gravifique	1762
Énergie potentielle d'une sphère de matière.....	1766
Conservation de l'énergie mécanique totale	1767
Conservation de la quantité de mouvement	1768
Loi de Newton généralisée.....	1769
Action lagrangienne	1771
Lagrangien mécanique	1771
Puissance	1773

Puissance d'une machine tournante	1774
Rendement	1774
Mouvements relatifs et forces d'inerties	1775
Accélération de Coriolis	1779
Principe d'Alembert	1779
Théorème de König	1782
Premier théorème de König	1782
Deuxième théorème de König	1783
Mouvements oscillants	1784
Pendule de Newton	1785
Pendule simple	1787
Pendule physique	1789
Pendule élastique	1792
Pendule conique	1795
Pendule de torsion	1797
Pendule de Foucault	1797
Pendule de Huygens	1800
Isochronisme rigoureux	1800
Courbe brachistochrone	1801
Pendule double	1808
Tribologie	1812
Frottement statique	1812
Frottement dynamique	1814
Lois de Coulomb	1815
Frottement exponentiel	1815
Frottement visqueux horizontal	1818
Frottement visqueux vertical	1820
Frottement visqueux de Stokes vertical	1821
Frottement visqueux de Stokes horizontal	1823
Mécanique ondulatoire	1832
Fonction d'onde	1832
Vitesse de phase	1832
Amplitude	1832
Équation d'onde/d'Alembert	1833
Types d'ondes	1835
Ondes périodiques	1835
Longueur d'onde	1835
Ondes harmoniques	1836
Nombre d'onde	1836
Ondes stationnaires	1837
Modes de vibrations dans un fil tendu	1839
Équation des cordes vibrantes	1840
Conditions de Dirichlet	1841
Conditions de Cauchy	1841
Conditions de Neumann	1847
Lagrangien d'une corde	1849
Densité lagrangienne	1850
Modes de vibrations dans une membrane tendue	1853
Phaseurs	1859

Vecteur d'onde.....	1860
Relation de dispersion.....	1861
Paquet d'onde	1864
Mécanique statistique.....	1869
Théorie statistique de l'information.....	1869
Quantité d'information intrinsèque.....	1870
Formule de Shannon	1872
Information mutuelle moyenne	1872
Loi de Boltzmann	1874
Distributions statistiques physiques	1878
Distribution de Maxwell	1878
Distribution de Maxwell-Boltzmann.....	1885
Fonction de Maxwell-Boltzmann.....	1890
Statistique de Maxwell-Boltzmann	1891
Fonction de partition canonique.....	1891
Distribution de Boltzmann	1892
Distribution de Fermi-Dirac	1894
Énergie et température de Fermi	1898
Distribution de Bose-Einstein	1898
Condensation de Bose-Einstein.....	1899
Loi de Fick	1903
Deuxième loi de Fick	1904
Première loi de Fick	1904
Mouvement brownien	1908
Équation de Langevin	1908
Relation de Sutherland-Einstein.....	1912
Thermodynamique.....	1918
Variables thermodynamiques	1919
Systèmes thermodynamiques	1921
Types de systèmes.....	1922
Transformation thermodynamique	1922
Types de transformations	1922
Types de cycles	1923
Variables d'état	1923
Variables extensives/intensives.....	1924
Phases.....	1925
Équations d'état	1926
Équation d'état d'un gaz parfait	1926
Équation d'état d'un liquide	1927
Équation d'état d'un solide.....	1930
Principes de la thermodynamique	1932
Principe zéro (équilibre thermique).....	1932
Principe premier (principe d'équivalence)	1932
Principe deuxième (irréversibilité).....	1933
Principe troisième (principe de Nernst)	1933
Capacités calorifiques	1934
Modèle d'Einstein (Dulong-Petit) de la capacité calorifique des solides cristallins.....	1936
Modèle de Debye de la capacité calorifique des solides cristallins	1937
Énergie interne	1942

Travail des forces mécaniques	1945
Enthalpie	1947
Loi de Laplace	1949
Relation de Mayer	1949
Coefficients thermoélastiques	1953
Chaleur	1956
Entropie	1957
Écoulement de la chaleur	1962
Cycle de Carnot	1964
Relations de Maxwell	1966
Énergie libre	1968
Enthalpie libre	1968
Équations de Gibbs	1968
Transformations de Legendre	1970
Équation de continuité	1971
Équation de la chaleur	1975
Loi de Fourier	1976
Noyau de la chaleur	1983
Rayonnement thermique	1985
Loi de Stefan-Boltzmann	1987
Loi de Planck	1993
Loi de Rayleigh-Jeans	1998
Première loi de Wien	1999
Constante de Stefan-Boltzmann	2003
Deuxième loi de Wien	2004
Constante de Wien	2004
Mécanique des milieux continus	2011
Solides	2011
Pressions	2012
Élasticité des solides	2013
Module de Young	2014
Loi de Hooke	2015
Module de cisaillement	2016
Module de glissement	2018
Coefficient de Poisson	2019
Module de compressibilité	2026
Coefficient de compressibilité	2027
Module de flexion	2027
Liquides	2033
Théorème de Pascal	2034
Viscosité	2035
Loi de Poiseuille	2037
Théorème de Bernoulli	2038
Équation de Laplace	2043
Théorème de Toricelli	2043
Effet Venturi	2044
Tube de Pitot	2045
Perte de charge (pression)	2045
Équations de Navier-Stokes	2047

Fluides incompressibles	2064
Équation d'Euler de 1ère forme	2065
Équation d'Euler de 2ème forme.....	2066
Fluide compressible (barotropes).....	2070
Fluide statique	2070
Équation de la statique des fluides (loi fondamentale de l'hydrostatique).....	2070
Nombre de Reynolds.....	2070
Approximation de Boussinesq	2073
Loi de Stokes.....	2075
Pression hydrostatique	2078
Poussée d'Archimède	2079
Vitesse du son dans un liquide	2080
Gaz	2081
Type de gaz	2081
Gaz parfait.....	2081
Loi de Boyle-Mariott	2081
Loi de Gay-Lussac	2082
Loi de Chasles.....	2082
Gaz réel	2083
Théorème du Viriel	2083
Équation de Van der Waals.....	2083
Pression cinétique	2092
Température cinétique.....	2094
Libre parcours moyen	2096
Plasmas	2098
Degré d'ionisation	2099
Fréquence plasma.....	2099
Équation hydrodynamique des électrons.....	2103
Électrodynamique	2108
Électrostatique	2109
Force électrique	2109
Loi de Coulomb	2109
Permittivité du vide.....	2110
Constante diélectrique.....	2110
Champ électrique	2110
Champ de déplacement	2111
Potentiel électrique	2111
Différence de potentiel.....	2111
Indépendance du chemin.....	2113
Équipotentiels et lignes de champ	2114
Fil rectiligne infini	2116
Dipôle électrique rigide.....	2219
Moment dipolaire.....	2121
Flux du champ électrique	2129
Flux électrique.....	2129
Loi de Gauss (du champ électrique).....	2129
Capacités	2130
Condensateur plan.....	2130

Condensateur cylindrique	2131
Condensateur sphérique	2131
Rigidité diélectrique	2133
Énergie potentielle électrostatique	2134
Magnétostatique	2140
Théorème d'Ampère	2142
Excitation magnétique	2142
Susceptibilité magnétique	2142
Perméabilité magnétique absolue	2143
Bobine solénoïdale infinie	2144
Bobine toroïdale	2145
Électro-aimant	2146
Force d'un aimant ou électro-aimant	2147
Relation de Maxwell-Ampère	2149
Loi de Biot-Savart	2150
Champ magnétique pour une boucle de courant	2152
Champ magnétique pour un fil infini	2153
Dipôle magnétique	2155
Moment magnétique dipolaire	2160
Facteur gyromagnétique	2162
Loi de Lorentz	2162
Loi de Laplace	2167
Effet Hall classique	2167
Coefficient de Hall	2168
Résistance de Hall	2169
Rayon de Larmor	2170
Pulsation gyro-synchrotron	2171
Énergie d'un dipôle magnétique	2173
Modèle de Langevin du diamagnétisme	2176
Relation de la susceptibilité diamagnétique de Langevin	2179
Modèle de Langevin du paramagnétisme	2179
Fonction de Langevin	2182
Relation de la susceptibilité paramagnétique de Langevin	2184
Électrodynamique	2189
Première équation de Maxwell	2190
Loi de Gauss (du champ électrique)	2192
Équation de Maxwell-Poisson	2193
Deuxième équation de Maxwell	2193
Loi de Gauss (du champ magnétique)	2195
Troisième équation de Maxwell	2195
Loi de Lenz-Faraday	2195
Loi de Maxwell-Faraday	2197
Bêtatron	2197
Quatrième équation de Maxwell	2200
Courant de déplacement	2202
Équation de Maxwell-Ampère	2203
Formes locales des équations de Maxwell	2203
Monopôles magnétiques	2204
Équations de Dirac-Maxwell symétrisées	2206

Équation de conservation de la charge	2207
Théorie de Jauges (électromagnétiques)	2208
Potentiel vecteur	2208
Potentiel scalaire	2209
Jauge de Lorenz	2209
Jauge de Coulomb	2210
Équation de Poisson du potentiel vecteur	2210
Jauge arbitraire	2211
Équation d'onde des potentiels électromagnétiques	2211
Quadrivecteur potentiel	2212
Quadrivecteur courant	2212
Tenseur du champ électromagnétique	2214
Lagrangien de l'interaction cours-champs	2217
Tenseur de Faraday	2220
Équations du mouvement d'une particule dans un champ électromagnétique	2220
Équations de Maxwell sous forme tensorielle	2226
Lagrangien total de l'interaction charge-champ	2231
Équation d'onde électromagnétique	2231
Équation de propagation	2232
Vitesse de propagation	2233
Équation de Helmholtz	2236
Énergie véhiculée	2237
Intensité d'une onde électromagnétique	2238
Émission	2239
Rayonnement synchrotron	2241
Potentiels de Liénard-Wiechert	2245
Électrocinétique	2258
Lois de Kirchhoff	2258
Loi des mailles	2259
Loi des nœuds	2259
Modèle de Drude	2259
Libre parcours moyen de l'électron de conduction	2260
Temps de collision moyen de l'électron de conduction	2260
Vitesse moyenne de dérive	2261
Résistivité en fonction de la température	2264
Loi d'Ohm	2265
Résistance électrique	2265
Résistances équivalentes	2267
Série	2267
Parallèle	2267
Capacités équivalentes	2268
Série	2268
Parallèle	2268
Force électromotrice	2269
Puissance électrique	2269
Champ électromoteur	2270
Force électromotrice (FEM)	2272
Loi de Faraday	2272
Loi de Lenz	2274

Inductance	2275
Coefficient d'auto-induction	2275
Effet de peau	2278
Semi-conducteurs	2281
Modèle semi-classique des bandes paraboliques	2282
Particule quasi-libre	2286
Densité de modes	2289
Sphère de Fermi	2289
Densité statistique non dégénérée des porteurs négatifs	2290
Énergie minimale de la bande de conduction	2292
Densité statistique non dégénérée des porteurs positifs	2296
Énergie maximale de la bande de valence	2297
Bandes d'énergie	2300
Bande interdite	2300
Bande de conduction	2301
Bande de valence	2301
Loi d'Ohm (des semi-conducteurs)	2303
Vitesse de Fermi	2304
Optique géométrique	2310
Sources et ombres	2310
Couleur	2313
Synthèse additive	2315
Synthèse soustractive	2316
Photométrie	2317
Flux énergétique	2318
Loi de Beer-Lambert	2318
Intensité lumineuse	2319
Émittance énergétique	2320
Luminance énergétique	2321
Loi de Lambert	2322
Loi de Kirchhoff	2323
Décomposition spectrale	2323
Loi de réfraction	2324
Principe de Fermat	2324
Loi de Snell-Descartes	2328
Loi de réfraction	2329
Effet Tcherenkov	2331
Formules de Descartes	2332
Formule de Descartes pour la réflexion sur une surface sphérique	2334
Formule de Descartes pour la réfraction au passage d'une surface sphérique	2339
Première formule de Descartes pour les lentilles minces	2342
Équation de l'opticien	2346
Deuxième formule de Descartes pour les lentilles minces (formule de conjugaison)	2346
Prisme	2349
Arc-en-ciel	2352
Angle de l'arc-en-ciel	2356
Optique ondulatoire	2362
Principe d'Huygens	2362
Loi de Malus	2363

Diffraction de Fraunhofer	2364
Cas d'une fente rectangulaire	2365
Pouvoir de résolution	2372
Cas d'un réseau de fentes rectangulaires	2373
Fentes de Young.....	2379
Polarisation de la lumière	2384
Polarisation linéaire.....	2388
Polarisation elliptique	2389
Polarisation circulaire	2391
Polarisation naturelle.....	2391
Loi de Malus de la polarisation.....	2393
Cohérence et interférence.....	2395

Atomistique2405

Physique quantique corpusculaire.....2406

Modèle de Dalton	2408
Modèle de Thomson.....	2408
Modèle de Rutherford	2409
Modèle de Bohr.....	2410
Postulats de Bohr	2411
Quantification.....	2411
Intégrale d'action	2412
Modèles des atomes hydrogénoïdes sans entraînement	2413
Rayon de Bohr	2414
Formule de Balmer.....	2416
Constante de Rydberg	2416
Modèle des atomes hydrogénoïdes avec entraînement	2418
Hypothèse du neutron	2421
Modèle de Sommerfeld et Wilson.....	2422
Nombre quantique azimutal	2424
Nombre quantique radial.....	2425
Modèle relativiste de Sommerfeld	2426
Constante de structure fine.....	2434
Moment magnétique dipolaire quantique.....	2445
Facteur gyromagnétique.....	2445
Magnéton de Bohr.....	2445
Nombre quantique magnétique	2446
Nombre quantique de moment cinétique orbital	2446
Spin	2447
Principe d'exclusion de Pauli	2449
Couches électroniques.....	2449
Nombre quantique principal.....	2450
Nombre quantique secondaire/azimutal	2450
Nombre quantique magnétique	2450
Nombre quantique de spin	2450

Physique quantique ondulatoire2459

Postulats	2460
1er postulat (état quantique).....	2461
Fonction d'onde	2461

Condition de normalisation de De Broglie.....	2461
2ème postulat (évolution temporelle d'un état quantique)	2463
Opérateur propre	2463
Fonction propre	2463
3ème postulat (observables et opérateurs)	2464
Opérateur hermitique	2464
Principe de correspondance	2465
4ème postulat (mesure d'une propriété)	2466
5ème postulat (moyenne d'une propriété)	2468
Principes d'incertitudes classiques	2469
Première relation d'incertitude classique (impulsion).....	2470
Deuxième relation d'incertitude classique (spatiale).....	2470
Troisième relation d'incertitude classique (temporelle)	2472
Algèbre quantique	2473
Opérateurs linéaires fonctionnels.....	2473
Opérateurs de quantité de mouvement.....	2476
Opérateurs adjoints et hermitiques.....	2476
Commutateurs et anti-commutateurs	2478
Relation d'anti-commutation.....	2480
Relations d'incertitudes de Heisenberg	2481
Représentatives	2482
Valeurs et fonctions propres.....	2484
Orthogonalité des fonctions propres	2485
Formalisme de Dirac	2488
Kets et Bras	2488
Comparaison formalisme	2491
Modèle de Schrödinger	2491
Onde associée de De Broglie	2491
Longueur d'onde associée	2492
Onde thermique associée de De Broglie	2493
Équation classique de Schrödinger	2493
Équation unidimensionnelle de Schrödinger	2495
Hamiltonien de Schrödinger	2496
Condition de normalisation de De Broglie.....	2499
États liés et non liés.....	2502
Équation d'évolution classique de Schrödinger.....	2503
Opérateur d'évolution.....	2503
Condition d'hermiticité.....	2507
Opérateur de Heisenberg.....	2508
Séparation des variables.....	2509
Combinaison linéaire des états.....	2511
État stationnaire	2511
Équation de continuité	2513
Implications et applications.....	2514
Particule libre	2515
Courbe de dispersion	2516
Paquet d'ondes unidimensionnelles	2516
Paquet d'ondes quasi-monochromatiques	2520
Puits de potentiel à parois rectilignes.....	2522

Énergie de Fermi	2531
Oscillateur harmonique	2532
Relation de De Broglie	2533
Opérateurs de création/destruction	2537
Effet tunnel.....	2545
Coefficient de transmission/réflexion	2549
Principe de superposition linéaire des états	2557
Théorème d'Ehrenfest	2568
Moment cinétique et spin	2572
Opérateur moment cinétique orbital.....	2573
Spin	2576
Moment cinétique total	2577
Opérateur de Casimir	2577
Couplage spin-orbite	2590
Dimensions de Planck	2594
Interprétation de Copenhague	2597
Physique quantique relativiste	2605
Équation d'évolution relativiste de Schrödinger.....	2605
Équation de Klein-Gordon libre.....	2606
Antimatière.....	2607
Équation de Klein-Gordon généralisée	2610
Équation de Dirac libre classique	2616
Équations de Weyl	2622
Bispineur	2623
Équation de Dirac.....	2624
Matrices de Dirac (matrices gamma)	2624
Solutions particulières de l'équation de Dirac	2629
Hélicité	2631
Équation de Dirac libre linéarisée	2631
Équation de Dirac généralisée	2643
Équation de Pauli	2644
Terme de Stern-Gerlach	2649
Facteur de Landé.....	2651
Moment magnétique orbital et de spin.....	2651
Énergie de Zeeman.....	2652
Fréquence de résonance	2653
Physique nucléaire	2659
L'arme nucléaire	2659
Radioactivité.....	2661
Nombre atomique.....	2661
Nombre de masse	2662
Isotones	2662
Désintégration	2663
Constante radioactive	2664
Demi-vie d'isotope	2665
Activité radioactive	2665
Équation d'activité.....	2665
Datation au carbone 14	2667
Filiation radioactive	2668

Équilibre séculaire.....	2671
Équilibre transitoire.....	2671
Non-équilibre	2672
Phénomènes radioactifs.....	2673
Énergie de liaison.....	2674
Courbe d'Aston	2674
Vallée de stabilité.....	2676
Fusion nucléaire (1).....	2677
Fission nucléaire (2).....	2677
Désintégration alpha (3).....	2679
Désintégration bêta - (4).....	2688
Désintégration bêta + (5).....	2690
Capture électronique (6).....	2691
Émission gamma (7)	2692
Conversion interne (8).....	2692
Électron de conversion.....	2692
Électron Auger	2693
Radioprotection	2695
Formule de Bethe-Bloch	2696
Effet Compton	2699
Longueur d'onde de Compton	2702
Effet photoélectrique.....	2702
Diffusion de Rutherford	2706
Paramètre d'impact.....	2707
Section efficace	2709
Section différentielle efficace de Rutherford	2711
Rayons-X et Gamma	2714
Coefficient d'atténuation massique	2715
Section efficace microscopique.....	2716
Longueur de relaxation	2717
Création paires électron-positron	2717
Modèle nucléaire "goutte liquide".....	2718
Énergie de liaison en volume	2719
Énergie de liaison superficielle	2720
Énergie de répulsion électrostatique	2721
Énergie d'asymétrie (énergie de Pauli).....	2722
Énergie de paire (énergie d'appariement).....	2727
Formule semi-empirique de von Weizsäcker.....	2727
Physique quantique des champs	2739
Photons virtuels	2741
Potentiel de Yukawa.....	2742
Champs massiques	2743
Énergie mise en jeu dans l'interaction nucléaire forte.....	2744
Énergie mise en jeu dans l'interaction nucléaire faible	2745
Champs non-massiques.....	2745
Équation d'Euler-Lagrange des champs	2746
Équations du mouvement de Heisenberg.....	2747
Principe de correspondance.....	2751
Équation de Lagrange des champs	2752

Lagrangien de Klein-Gordon	2752
Lagrangien d'interaction du champ électromagnétique avec une densité	2754
Lagrangien total du champ électromagnétique	2755
Lagrangien du champ spinoriel de Dirac libre	2756
Théories de jauge	2757
Invariance de jauge globale	2758
Invariance de jauge locale	2759
Lagrangien de l'électrodynamique des champs	2764
Physique des particules	2770
Constantes de couplage	2771
Résonance magnétique de spin	2773
Résonance magnétique électronique	2780
Résonance magnétique nucléaire	2781
Cosmologie	2785
Astronomie	2786
Équation de Drake	2786
Lois de Kepler	2788
Première loi de Kepler	2788
Deuxième loi de Kepler	2788
Troisième loi de Kepler	2789
Loi de la gravitation de Newton	2790
Potentiel gravitationnel	2793
Équation de Newton-Poisson	2794
Sphérisation des corps célestes	2795
Aplatissement des corps célestes	2796
Stabilité des atmosphères	2798
Limite de Roche	2799
Trajectoires d'orbitales képlériennes	2800
Première formule de Brinet	2801
Deuxième formule de Brinet	2806
Période orbitale képlérienne	2809
Déflexion classique de la lumière	2809
Précession du périhélie	2812
Durée de l'arc Diurne	2821
Mouvement des planètes	2827
Période synodique et sidérale	2828
Rétrogradation des planètes	2830
Points de Lagrange	2835
Positions d'équilibre du premier type	2841
Point L1 de Lagrange	2842
Point L2 de Lagrange	2847
Point L3 de Lagrange	2849
Positions d'équilibre du deuxième type	2852
Points L4, L5 de Lagrange	2852
Astrophysique	2863
Étoiles	2863
Genèse	2868
Effondrement d'un nuage interstellaire	2868

Masse de Jeans	2869
Rayon de Jeans.....	2870
Temps de chute libre	2871
Durée de vie nucléaire.....	2871
Température interne	2873
Température externe.....	2874
Luminosité	2875
Éclat	2875
Magnitude apparente.....	2876
Loi de Pogson	2877
Magnitude absolue	2877
Étoiles variables	2879
Parallaxe trigonométrique	2881
Effet Doppler-Fizeau relativiste.....	2882
Décalage spectral	2883
Vitesse apparente	2885
Limite de Chandrasekhar	2888
Limite de rupture de rotation.....	2892
Relativité restreinte.....	2898
Principes et postulats	2898
Postulat d'invariance	2899
Principe cosmologique	2899
Principe de relativité restreinte.....	2900
Transformations de Lorentz	2902
Invariants relativistes	2903
Facteur de Michelson-Morley	2906
Quadrivecteur déplacement.....	2906
Matrice de Lorentz.....	2906
Invariance de l'équation d'onde.....	2908
Interprétation hypergéométrique.....	2910
Quadrivecteur vitesse.....	2911
Quadrivecteur courant.....	2915
Quadrivecteur accélération	2916
Addition relativiste des vitesses.....	2922
Variation relativiste des longueurs.....	2923
Variation relativiste du temps	2924
Expérience de Hafele-Keating	2925
Paradoxe des jumeaux	2929
Variation relativiste de la masse	2931
Équivalence masse-énergie.....	2934
Énergie au repos	2935
Lagrangien relativiste.....	2936
Action invariante de Lorentz	2937
Quantité de mouvement relativiste.....	2938
Quadrivecteur d'énergie-impulsion.....	2941
Relation d'Einstein	2942
Force relativiste.....	2943
Quadrivecteur force	2944
Électrodynamique relativiste.....	2944

Transformation du tenseur de champ.....	2951
Espace-Temps de Minkowski	2952
Quadrivecteurs	2955
Cône d'Univers.....	2956
Relativité générale	2966
Postulats et principes	2966
Postulat d'équivalence	2966
Principe de Mach	2969
Métriques.....	2970
Équation métrique.....	2970
Tenseur métrique.....	2973
Signature	2975
Critère de Schild.....	2975
Effet Einstein	2976
Équations du mouvement.....	2978
Lagrangien géodésique	2985
Équation des géodésiques	2987
Connexion affine.....	2991
Limite newtonienne.....	2992
Tenseur d'énergie-impulsion	2994
Matrice des flux de moments	2998
Équation d'Einstein des champs	3000
Solution de Schwarzschild	3005
Coordonnées de Schwarzschild.....	3006
Métrique de Schwarzschild	3013
Vérifications expérimentales.....	3016
Précession du périhélie de Mercure	3016
Formule de Binet non relativiste	3022
Déflexion de la lumière	3029
Effet Shapiro	3034
Trous Noirs	3041
Cosmologie	3047
Modèle cosmologique newtonien.....	3047
Principe cosmologique	3047
Loi de Hubble.....	3048
Constante et temps de Hubble.....	3051
Équations de Friedmann.....	3051
Première équation de Friedman	3052
Deuxième équation de Friedman	3055
Densité critique	3056
Paramétrique de densité cosmologique	3057
Modèles cosmologique de Friedmann-Lemaître.....	3060
Espace plat ($K=0$).....	3060
Espace plat dominé par la matière	3062
Espace plat dominé par la radiation	3063
Espace sphérique ($K>0$)	3066
Espace sphérique dominé par la matière.....	3070
Espace sphérique dominé par la radiation.....	3073
Espace hyperbolique ($K<0$).....	3076

Espace hyperbolique dominé par la matière	3079
Espace hyperbolique dominé par la radiation	3082
Univers observable.....	3085
Rayonnement fossile	3097
L'Univers Trou Noir	3100
<i>Théorie des cordes</i>	3105
Équation d'onde non relativiste d'une corde transversale	3106
Équation d'onde relativiste d'une corde transversale	3108
Surface d'Univers	3108
Coordonnée de corde.....	3113
Action de Nambu-Goto	3113
Lagrangien d'une corde	3121
Chimie	3129
<i>Chimie quantique</i>	3130
Puits de potentiel rectangulaire tridimensionnel infini.....	3130
Vibrations moléculaires.....	3134
Atome hydrogénoïde	3137
Approximation de Born-Oppenheimer.....	3139
Rotateur rigide.....	3141
Polynômes associés de Legendre	3148
Harmoniques sphériques	3160
Cartes de densité et d'isodensité.....	3161
Profil de potentiel.....	3170
Potentiel électrique écranté	3170
<i>Chimie moléculaire</i>	3177
Approximation orbitale	3178
Approximation de Born-Oppenheimer.....	3179
Méthode de Slater.....	3181
Méthode C.L.O.A.....	3182
Fonction liante.....	3183
Fonction antiliante.....	3183
Liaison covalente	3184
Intégrales de résonance	3186
Intégrales coulombiennes.....	3186
Intégrales de recouvrement	3187
Équations séculaires	3187
<i>Chimie analytique</i>.....	3195
Mélanges simples	3196
Réactions	3197
Grandeurs stœchiométriques	3197
Avancement élémentaire de réaction	3198
Réactif limitant.....	3199
Taux d'avancement.....	3199
<i>Chimie thermique</i>	3206
Transformations chimiques	3206
Chaleur de réaction à volume constant	3207
Chaleur de réaction à pression constante	3207
Grandeurs molaires	3208

Masse molaire atomique	3208
Masse molaire moléculaire	3208
Titre molaire.....	3209
Titre massique associé.....	3210
Molarité.....	3210
Grandeur molaire standard.....	3210
Opérateur de Lewis	3213
Enthalpie standard de réaction	3214
Loi de Hess	3215
Relation enthalpique de Kirchhoff.....	3217

Informatique Théorique3224

Méthodes numériques 3226

Complexité	3228
NP-Complétude.....	3231
Partie Entière	3234
Algorithme d'Héron.....	3236
Algorithme d'Archimède	3237
Calcul du nombre d'Euler.....	3239
Calcul de la factorielle (formule de Stirling).....	3239
Systèmes d'équations linéaires	3240
Une équation à une inconnue	3240
Deux équations à deux inconnues	3241
Trois équations à trois inconnues.....	3242
N équations à n inconnues.....	3243
Polynômes	3243
Régressions et interpolations.....	3246
Régression linéaire à une variable explicative	3247
Droite de régression	3248
Méthode des moindres carrés.....	3249
Moindres carrés pour cas non linéaires.....	3251
Analyse de la variance de la régression bivariée	3252
Modèle linéaire gaussien.....	3256
Hypothèse de Gauss-Markov	3258
Intervalle de confiance de la pente.....	3263
Intervalle de confiance du coefficient de corrélation	3264
Test du coefficient de corrélation de Pearson	3264
Intervalle de confiance des valeurs prédictives.....	3267
Intervalle de prédiction	3271
Régression linéaire à une variable explicative forcée par l'origine	3272
Régression linéaire multiple.....	3273
Régression linéaire de polynômes simples	3278
Régression polynomiale	3279
Régression logistique (logit)	3280
Coefficient de corrélation (détermination) généralisé.....	3288
Interpolation polynomiale	3289
Courbes de Bézier (B-spline).....	3289
Méthode d'Euler	3296
Polynôme de collocation.....	3297

Méthode (d'interpolation) de Lagrange	3300
Recherche des racines	3302
Méthodes des parties proportionnelles	3303
Méthode de la bisection	3304
Méthode de la sécante (regula falsi)	3305
Méthode de Newton	3306
Dérivations numériques	3311
Intégrations numériques	3312
Méthode des rectangles	3313
Méthode des trapèzes	3314
Aires et sommes de Riemann	3314
Programmation (optimisation) linéaire	3315
Système canonique	3320
Forme standard	3320
Algorithme du simplexe	3326
Programmation (optimisation) non linéaire	3331
Méthode de Newton-Raphson	3333
Méthode de Gauss-Newton (Newton-Tangente)	3339
Méthodes de Monte-Carlo	3347
Génération de variables aléatoires	3348
Nombres pseudo-aléatoires	3348
Calcul d'une intégrale	3351
Calcul de Pi	3352
Modélisation	3353
Boostrapping	3356
Dichotomie	3361
Analyse en composantes principales (A.C.P.)	3362
Plan factoriel	3369
Analyse factorielle des correspondances (A.F.C.)	3379
Représentation en fréquences conjointes	3380
Index/Élasticité	3381
Test d'indépendance du khi-2	3384
V de Cramer	3387
phi de Cramer	3389
Test exact de Fisher	3391
Kappa d'agrément de Cohen	3396
Test de McNemar	3398
Coefficient de Yule	3400
Test de Cochran-Mantel-Haenzel	3401
Méthode des différences finies	3408
M.D.F. à une dimension spatiale	3408
M.D.F. spatio-temporelle	3408
Cellule de Yee	3413
Clustering	3414
Arbre de régression et de classification (CART)	3414
Indice d'impureté de Gini	3417
K-Means (nuées dynamiques)	3423
Dendrogramme	3431
Réseaux de neurones formels	3433

Modèle de neurone	3434
Fonction de transfert	3437
Architecture de réseau.....	3439
Algorithmes génétiques.....	3448
Codage et population initiale	3452
Opérateurs	3454
Opérateur de sélection.....	3454
Opérateur de croisement	3455
Opérateurs de mutation	3456
Fractales	3465
Fractales IFS.....	3465
Théorème de Bolzano-Weierstrass	3470
Espace métrique des fractales	3480
Fractale de Cantor	3484
Fractale du triangle de Sierpinski.....	3486
Fractale du tapis de Sierpinski	3492
Fractale spirale	3496
Fractale de Von Koch.....	3498
Fractales naturelles.....	3501
Rameau	3502
Flocon de neige	3504
Arbre	3507
Fougère	3509
Fractales à temps d'échappement	3513
Ensemble de Mandelbrot.....	3514
Ensemble de Julia.....	3518
Ensemble de Newton.....	3521
Systèmes logiques formels.....	3527
Logique stricte.....	3527
Algèbre de Boole	3528
Axiomes	3529
Théorème des constantes	3530
Théorèmes du consensus.....	3531
Théorèmes de Shannon	3532
Théorème de De Morgan	3532
Fonctions logiques	3534
Tables de Karnaugh.....	3536
Opérations arithmétiques booléennes	3539
Logique floue	3544
Ensemble flou	3547
Codes correcteurs	3554
Checksum	3556
Encodeurs	3557
Distance de Hamming.....	3557
Poids de Hamming	3559
Codes en blocs-linéaires.....	3561
Taux de codage	3562
Matrice génératrice	3563
Matrice de contrôle	3564

Codes systématiques	3567
Cryptographie	3573
Systèmes cryptographiques	3574
Principe de Kerchoffs.....	3577
Trappes	3577
Système de chiffrement à clé secrète	3578
Schéma de Feistel.....	3580
Système de chiffrement à clé publique.....	3583
Protocole de Diffie-Hellmann	3585
Système R.S.A.	3588
Théorème d'Euler	3588
Fonctions de Hachage	3600
Fonction de condensation Message Digest MD5.....	3601
Fonction de condensation Secure Hash Algorithm SHA-1	3602
Certificats d'authentification	3603
Cryptographie quantique	3606
Cryptographie alternative	3611
Automates	3616
Mise en perspective	3616
Machine de Von Neumann.....	3617
Machine de Turing	3618
Hiérarchie de Chomsky.....	3621
Langage formel	3621
Syntaxe.....	3622
Automates associés	3626
Terminologie	3626
Mots	3626
Langages	3629
Équations.....	3631
Codes.....	3632
Codes préfixes.....	3633
Algorithmes linguistiques	3634
Algorithme de Huffman	3634
Algorithme de Sardinas et Petterson	3639
Informatique quantique	3645
Polarisation du photon.....	3648
QuBit	3650
Sphère de Bloch	3656
QuBit de polarisation	3662
Base H/V	3662
Base D/A.....	3663
QuBit de spin $\frac{1}{2}$	3664
Mathématiques sociales	3674
Dynamique des populations	3675
Tables de mortalité et natalité (fonctions biométriques)	3675
Quotient de mortalité.....	3676
Ordre des vivants	3679
Renouvellement de la population.....	3683

Modèles de populations.....	3684
Modèle exponentiel.....	3684
Modèle logistique déterministe	3688
Modèle logistique chaotique	3691
Diagramme de Feigenbaum	3696
Loi de Malthus	3700
Modèle de Leslie.....	3701
Propagation des épidémies	3702
Modèle Proies-Prédateurs (de Lotka-Volterra)	3706
Modèle de capture optimale de Schaefer	3706
Modèle de Hardy-Weinberg.....	3717
Taux de croissance avec la température	3722
<i>Théorie de la décision (jeux)</i>	3728
Théorie des jeux VS Théorie de la décision.....	3728
Types de jeux	3729
Représentations de jeux.....	3730
Forme extensive d'un jeu.....	3731
Forme extensive d'une décision	3732
Forme normale d'un jeu.....	3738
Maximin	3742
Maximax	3742
Minimax	3742
Jeux répétitifs	3744
Forme ensembliste d'un jeu.....	3745
Forme graphique d'un jeu.....	3750
Jeux coopératifs et non-coopératifs.....	3753
Optimum de Pareto	3755
Équilibre de Nash.....	3755
Utilité espérée	3757
Critère de Hurwitz.....	3758
Critère de Laplace	3760
Jeux évolutionnaires.....	3761
Équilibre de Cournot.....	3767
Chaînes de Markov.....	3770
<i>Économie</i>	3779
Concepts	3779
Micro-économie	3780
Coût moyen et marginal	3785
Macro-économie	3789
Modèle de Cobb-Douglas	3789
Modèle monétaire.....	3793
Loi de Say	3795
Loi de Walras	3796
Théorie de l'offre et de la demande	3798
Théorie de la préférence.....	3798
Modèle contrarié à perte nette.....	3804
Capitalisation et actuariat	3811
Intervalles de dates.....	3813
Équivalences de taux.....	3817

Intérêt simple.....	3818
Escompte.....	3818
Intérêt composé.....	3820
Intérêt continu.....	3822
Intérêt progressif (rentes).....	3824
Rentes postnumerando.....	3825
Rentes praenumerando.....	3827
Arrondis.....	3829
Emprunts.....	3829
Emprunt à échéance fixe.....	3830
Emprunt à amortissement constant.....	3831
Emprunt à annuité constante.....	3832
Théorie moderne des portefeuilles.....	3834
Absence d'opportunité d'arbitrage.....	3836
Portefeuilles.....	3839
Actions.....	3840
Modèle d'évaluation de Durand.....	3842
Modèle d'évaluation de Gordon-Shapiro.....	3843
Obligations.....	3845
Courbe de taux.....	3847
Valeur actuelle.....	3849
Prix obligataire de non-arbitrage.....	3852
Bons de souscription.....	3854
Contrats à terme.....	3856
Pricing des contrats.....	3857
Hedging de contrats.....	3858
Options.....	3861
Types d'options simples.....	3863
Effet de levier.....	3865
Appel de marge.....	3866
Fonds de placement.....	3869
Retours et taux d'investissements.....	3869
Return On Investment (ROI).....	3870
Internal Rate of Return (IRR).....	3871
Money Weighted Rate of Return.....	3872
Time Weighted Rate of Return.....	3874
Modèle spéculatif de Bachelier.....	3876
Espérance et variance positive.....	3880
Modèle de diversification efficient de Markowitz.....	3884
Frontière efficient de Markowitz.....	3887
Portefeuille global de variance minimum.....	3894
Capital Market Line (C.M.L.).....	3894
Modèle de diversification efficient de Sharpe.....	3898
Coefficient bêta.....	3898
Coefficient alpha.....	3901
Ratio de Sharpe.....	3906
Tracking error.....	3907
Modèle d'évaluation des actifs financiers (MEDAF).....	3912
Taux de rendement certain.....	3915

Capital Market Line (C.M.L.).....	3916
Security Market Line (S.M.L.)	3919
Prime de risque	3920
Spread de crédit	3920
Modèle d'évaluation des options de Black & Scholes	3920
Équation de parité Call-Put.....	3921
Hypothèse d'efficent du marché.....	3923
Processus de Wiener	3924
Mouvement brownien standard.....	3925
Mouvement brownien généralisé	3298
Pont brownien	3932
Processus d'Itô	3935
Mouvement brownien géométrique	3935
Théorème d'Itô-Doeblin.....	3938
Modèle de Bachelier-Samuelson	3941
Équation de Black & Scholes (Merton).....	3945
Équation différentielle de Black & Scholes (à coeff. non constants)	3947
Portefeuille autofinçant sur sous-jacent risqué.....	3947
Les Grecques et autres... ..	3948
Résolution de l'E.D.P. de Black & Scholes	3951
Équation différentielle partielle à coeff. constants	3954
Forme de diffusion de l'E.D.P.....	3958
Value At Risk (V@R).....	3959
VaR relative	3960
VaR absolue.....	3963
VaR historique	3964
VaR de crédit	3965
VaR opérationnelle	3966
VaR variance-covariance.....	3968
Back-testing de la VaR (modèle binomial).....	3969
Analyse des séries temporelles (A.S.T.)	3970
Types d'erreurs	3975
Décomposition	3976
Modèles prévisionnels déterministes	3983
Moyenne mobile simple (lissage par moyenne mobile)	3984
Modèle linéaire avec coefficients saisonniers.....	3986
Lissage exponentiel simple	3989
Lissage exponentiel double à un paramètre (méthode de Brown)	3999
Lissage exponentiel double à deux paramètres de Holt (modèle additif)	4007
Lissage exponentiel triple à 3 paramètres de Holt et Winters (multiplicatif)	4012
Modèles autorégressifs.....	4025
Coefficient d'autocorrélation.....	4028
Corrélogramme	4029
Processus autorégressifs AR(p)	4033
Processus autorégressifs MA(q)	4035
Processus ARMA(p,q)	4036
Processus ARIMA(p,d,q).....	4037

Techniques de gestion	4045
Analyse du seuil de rentabilité	4047
Diagramme de Pareto	4049
Indice de Gini	4053
PERT probabiliste	4056
Loi bêta de première espèce	4057
Estimation à deux points	4061
Processus Six Sigma (lean)	4062
Rendement global combiné	4063
Modèle statistique de contrôle des salaires	4066
Gestion de stocks	4067
Stocks en avenir incertain	4069
Stock initial en gestion calendaire et à rotation nulle	4070
Modèle de Wilson	4070
Modèle de Wilson avec réapprovisionnement	4083
Relations d'optimalité	4088
Modèle de Wilson sans réapprovisionnement	4091
Relations d'optimalité	4094
Modèle de Wilson avec réapprovisionnement et rupture	4094
Relations d'optimalité	4098
Analyse de la sensibilité	4099
Biens d'équipement	4103
Amortissement linéaire	4103
Amortissement arithmétique dégressif	4104
Amortissement géométrique dégressif	4105
Choix d'investissements	4106
Valeur actuelle nette (VAN)	4106
Taux de rentabilité interne (TRI)	4109
Délai de récupération et d'amortissement	4110
Théorie des files d'attentes	4110
Modélisation des durées d'arrivées M/M/	4113
Fonction de répartition d'Erlang	4116
PASTA	4116
Modélisation des durées de service	4119
Notation de Kendall	4121
Modélisation des arrivées et départs M/M/1	4122
Relation de Little	4126
Probabilité de mise en attente M/M/K/K (formule d'Erlang-B)	4128
Probabilité de mise en attente M/M/K/inf (formule d'Erlang-C)	4132
Assurances	4136
Calcul de prime	4137
Prise en compte de l'expérience	4139
Facteur d'actualisation d'une assurance retraite	4142
Assurances de rentes	4145
Rente viagère temporaire	4146
Rente viagère différée	4147

Musique mathématique	4153
Ondes sonores longitudinales	4153
Relation de Newton-Laplace	4157
Puissance transportée par une onde sonore	4159
Mesure de l'intensité du son	4163
Ondes sphériques.....	4166
Effet Doppler.....	4166
Source fixe-Observateur en mouvement.....	4166
Source en mouvement-Observateur fixe	4168
Observateur et Source en mouvement.....	4170
Ondes de choc	4170
Gammes musicales	4171
Oscillateur harmonique	4173
Oscillateur forcé	4176
Ingénierie	4180
Génie marin & météo	4182
Horizon visuel	4182
Direction des vents	4184
Modèle atmosphérique exponentiel.....	4186
Équilibre hydrostatique	4186
Modèle atmosphérique adiabatique.....	4189
Équation hypsométrique.....	4192
Ballon-sonde.....	4193
Cyclogenèse et anticyclogenèse	4197
Cellules de Hadley	4197
Équilibre géostrophique	4198
Paramètre de Coriolis.....	4200
Nombre de Rossby	4201
Équation des vents	4202
Marées	4204
Première approche.....	4204
Force de marée statique.....	4205
Deuxième approche.....	4207
Équation de Lorenz	4212
Coefficient de Lamé.....	4216
Tenseur des taux de déformation	4219
Rouleaux de convection de Rayleigh-Bénard	4223
Nombre de Prandtl	4229
Nombre de Rayleigh	4230
Attracteurs étranges.....	4231
Vagues.....	4235
Profondeur des vagues	4246
Amplitude des vagues	4247
Génie mécanique	4253
Engrenages	4253
Rapports de transmission	4255
Association d'engrenages	4259
Résistance des matériaux (R.D.M.).....	4264

Moment quadratiques	4267
Équation de la ligne élastique	4269
Équation des poutres (Euler-Bernoulli)	4275
Énergie potentielle élastique	4281
Torsion	4282
Ressort de compression.....	4285
Flambage	4286
Formule d'Euler.....	4289
Traction	4289
Génie électrique.....	4295
Quelques symboles de base.....	4295
Courant alternatif vs Courant continu	4298
Puissance moyenne	4300
Courant RMS	4301
Puissance active	4303
Puissance réactive	4303
Puissance apparente	4303
Transformateur	4304
Circuits linéaires en régime constant.....	4307
Circuit RC série.....	4307
Constante de temps	4308
Circuit RL série	4311
Circuit RLC série	4314
Résistance critique	4315
Régime critique	4315
Régime apériodique	4317
Régime pseudo-périodique (ou des oscillations amorties).....	4319
Facteur d'amortissement	4322
Pulsation propre	4322
Circuits linéaires en régime forcé.....	4323
Filtre passe-bas passif	4326
Filtre passe-haut passif.....	4328
Intégrateur et dérivateur	4330
Génie civil	4336
Statique.....	4336
Poulies	4337
Poulie simple mobile.....	4339
Poulies composées	4340
Spirale de Cornu.....	4344
Câbles suspendus.....	4348
Câble suspendu libre (chaînette)	4348
Courbe caténaire	4351
Câble suspendu porteur (pont suspendu)	4356
Câble très tendu.....	4359
Barrages.....	4360
Génie aérospatial.....	4336
Vitesse cosmologique.....	4368
Première vitesse cosmique	4368
Deuxième vitesse cosmique	4368

Équation fondamentale de la propulsion	4371
Formule de Tsiolkovski.....	4372
Orbite géostationnaire	4379
Génie industriel	4386
Six Sigma	4387
Contrôle qualité.....	4389
Défauts/Erreurs	4390
Defects Per Unit.....	4391
Defects Per Opportunity.....	4391
Rendement Global Combiné	4393
Indices de capabilité.....	4395
Capabilité potentielle court terme	4396
Performance long terme	4398
Capabilité potentielle décentrée court terme.....	4401
Capabilité décentrée moyenne court terme	4406
Capabilité de processus de contrôle	4406
Niveaux de qualité.....	4409
Modèle de Taguchi.....	4421
Le nominal est le meilleur.....	4424
Le plus petit est le meilleur	4425
Le plus grand est le meilleur	4426
Maintenance préventive	4427
Obsolescence programmée	4427
Estimateurs empiriques.....	4428
Modèle de Weibull.....	4441
Distribution de Weibull à 2 paramètres	4442
Distribution de Weibull à 1 paramètre.....	4442
Topologie des systèmes	4447
Topologie série	4447
Topologie parallèle	4449
Topologie k/n.....	4451
Topologie série/parallèle et parallèle/série	4453
Topologie complexe	4454
Arbres de défaillances probabilistes.....	4458
Méthode du maximum de vraisemblance	4459
Modèle de survie de Kaplan-Meier.....	4461
Méthode ABC	4466
Plans d'expérience.....	4470
Plans factoriels complets.....	4479
Matrice d'expérimentation/effets	4481
Notation de Yates.....	4482
Algorithme de Yates et Hunter	4484
Matrice de Hadamard.....	4485
Plans de Plackett-Burman	4486
Plans factoriels fractionnaires	4489
Méthode de Box et Hunter.....	4492
Contrastes.....	4493
Alias.....	4493
Générateur.....	4494

Plans de résolution III	4495
Plans de résolution IV	4495
Plans de résolution V	4495
Plans et nomenclature de Taguchi.....	4498
Contrôle de réception	4514
Plan d'échantillonnage par mesure simple avec tolérance unique et σ connu.....	4517
Risque du fournisseur	4518
Risque du client	4518
Calcul des paramètres par utilisation de la norme AF-X06-023.....	4524
Niveau de qualité acceptable	4524
Plan d'échantillonnage simple par attributs.....	4526
Monographe binomial de Larson	4530
Calcul des paramètres par utilisation de la norme ISO 2859-1.....	4533
Courbe d'efficacité	4534
Cartes de contrôles (CC) de la qualité.....	4537
Règles empiriques de la WECO.....	4539
Échantillonnage.....	4541
Cartes de contrôles qualitatives (aux attributs)	4542
Carte de contrôle P.....	4542
Carte de contrôle NP	4546
Carte de contrôle C	4548
Carte de contrôle U	4551
Cartes de contrôles quantitatives (aux mesures)	4554
CC à valeurs individuelles avec limites imposées (fixes).....	4554
CC à valeurs individuelles avec moyenne et écart-type court terme	4556
CC avec moyennes basées sur l'erreur standard	4561
CC avec écarts-types (S Barre – S).....	4566
CC avec moyennes basées sur l'écart-type (X Barre – S).....	4573
CC avec étendues (R Barre – R).....	4577
CC avec moyennes basées sur l'étendue (X Barre – R).....	4580
Cartes de contrôles quantitatives autocorrélées (aux mesures).....	4584
CC à valeurs individuelles basées sur l'étendue mobile (I-EM X Barre).....	4585
CC à valeurs individuelles avec étendue mobile (I-EM Barre)	4589
CC moyennes mobiles (MA)	4593
CC Cusum avec V-Maque empirique	4596
CC EWMA (pondération exponentielle avec moyenne mobile) avec limites fixes	4603
Carte de contrôle des fréquences avec limites probabilistes	4610
Poisson arrivals see time average (PASTA)	4612
Carte de contrôle des événements rares (carte-G).....	4615
Génie logiciel	4622
Algorithme PageRank de Google.....	4622
Comptage pondéré	4624
Comptage récursif	4625
États absorbants.....	4631

Biographies	4638
Chronologie.....	4730
Humour	4763
Situations.....	4763
Mathématiques	4772
Physique	4781
Statistiques.....	4791
Économie.....	4792
Chimie	4792
Ingénierie.....	4997
Informatique	4801
Généralités.....	4808
Références	4811
Remerciements	4811
Références bibliographiques	4812
D.V.D.	4817
Liens Internet	4818
Sciences exactes	4818
Éditions-Magazines	4820
Associations.....	4823
Emplois.....	4824
TV-Radio.....	4825
Divers sciences	4826
Logiciels	4827
Citations	4830
À propos	4838

Le présent site document a été conçu de façon à ce que les connaissances requises pour le parcourir soient les plus élémentaires possibles. Il n'est donc pas nécessaire d'avoir obtenu un doctorat pour le consulter, il suffit de savoir raisonner, d'avoir l'esprit critique, d'observer et d'avoir du temps...



"La simplicité est le sceau de la vérité et celle-ci resplendit de beauté"

Omniprésente dans l'industrie (aérospatiale, imagerie, cryptographie, transports, chimie,...), ou dans les services (banques, assurances, ressources humaines, projets, logistique, architecture, télécommunications,...), la mathématique appliquée apparaît aussi dans de nombreux autres secteurs: sondages, modélisation des risques, protection des données... Elle intervient dans notre vie quotidienne (télécommunications, transports, médecine, météorologie, musique...) et contribue à la résolution de problématiques actuelles: énergie, santé, environnement, climatologie, optimisation, développement durable... Son grand succès est donc sa fantastique dispersion dans le monde réel et son intégration croissante à toutes les activités humaines. Nous allons donc vers une situation où la mathématique n'aura plus le monopole de la mathématique, mais où des économistes, gestionnaires et marchands feront tous de la mathématique.

À ce titre, ancien étudiant dans le domaine de l'ingénierie, j'ai souvent regretté l'absence d'un ouvrage unique assez complet, détaillé (sans aller dans l'extrême des puristes...) et pédagogique si possible gratuit (!) et portatif (étant personnellement un adepte des liseuses électroniques...) contenant au moins une idée non exhaustive de l'ensemble du programme de mathématique appliquée des écoles d'ingénieurs et présentant une vue d'ensemble de ce qui est utilisé dans la réalité des entreprises avec des démonstrations plus intuitives que rigoureuses mais avec suffisamment de détails afin d'éviter au lecteur des efforts inutiles. Un ouvrage aussi qui ne nécessite pas non plus de devoir s'adapter chaque fois à une nouvelle notation ou au vocabulaire spécifique à l'auteur quand il ne s'agit pas de changer carrément de langue... et où tout un chacun peut proposer des améliorations ou des compléments.

J'ai été de plus aussi frustré pendant mes études de devoir ingurgiter assez souvent des "formules" ou des "lois" soit disant (et à tort) indémontrables ou trop compliquées selon mes professeurs ou même déçu par des livres d'auteurs renommés (dont les développements sont laissés au soin du lecteur ou comme exercice...). Sur ce site Internet et le PDF associé, prédomine la volonté de ne jamais dérouter le lecteur par des formules creuses du style "il est évident que...", "on démontre facilement que...", "nous laissons le soin au lecteur de vérifier en tant qu'exercice que...", puisque tous les développements y sont présentés en détails. Mais je ne suis pas un puriste des maths! Je n'ai qu'une ambition: expliquer de la manière la plus simple possible.

Bien que je doive admettre que la démonstration de certaines relations présentées dans le cadre des cursus des écoles d'ingénieurs ne puisse se faire faute de temps dans le planning scolaire ou de place dans un livre, je ne peux accepter qu'un professeur ou un auteur dise à son étudiant (respectivement, son lecteur) que certaines lois sont indémontrables (car la plupart du temps c'est faux) ou que telle ou telle démonstration est trop compliquée sans lui donner une référence bibliographique (où l'étudiant puisse trouver l'information nécessaire à satisfaire sa curiosité) ou au moins une démonstration infiniment simplifiée et satisfaisante.

Par ailleurs, j'estime totalement archaïque le fait que certains professeurs continuent de faire prendre des notes de cours de manière massive à leurs étudiants. Il serait beaucoup plus favorable et optimal de distribuer un support de cours contenant tous les détails et ce afin de pouvoir se concentrer sur l'essentiel avec les élèves c'est-à-dire les explications orales, l'interprétation, la compréhension, le raisonnement et la mise en pratique plutôt que la copie de tableau noir à outrance... Bien évidemment donner un support de cours complet fait que certains étudiants brillent par leur absence mais... c'est tant mieux! Ainsi, ceux qui sont passionnés peuvent approfondir les sujets à la maison ou à la bibliothèque universitaire, les médiocres feront ce qu'ils ont à faire et pour le reste (élèves en difficultés mais travailleurs) ils suivront le cours donné par le professeur pour profiter de poser des questions plutôt que de recopier bêtement un tableau noir.

Pour me baser sur un modèle d'apprentissage d'un spécialiste américain, dont j'ai oublié le nom..., le présent site Internet (et son PDF associé) propose et impose les propriétés suivantes à son lecteur: découvrir, mémoriser, citer, intégrer, expliquer, reformuler, déduire, choisir, utiliser, décomposer, comparer, interpréter, juger, argumenter, modéliser, élaborer, créer, rechercher, raisonner, développer et ce dans une démarche claire, pédagogique et progressive permettant de développer l'esprit d'analyse et d'ouverture.

Alors, dans mon esprit, ce site Internet (et son PDF associé) doit pouvoir se substituer, gratuitement, à de nombreuses références et lacunes du système, permettant ainsi à tout étudiant curieux de ne pas être frustré pendant de longues années durant son cursus de formation. Sans quoi, la science de l'ingénieur pourrait alors avoir l'aspect rébarbatif d'une science figée, à l'écart de l'évolution scientifique et technique, d'une accumulation hétéroclite de connaissances et surtout de formules qui la font considérer comme un sous-produit insipide des mathématiques et qui amène dans les entreprises à de nombreux faux résultats...

Ceux qui ne voient la mathématique appliquée que comme un outil (ce qu'elle est aussi), ou comme l'ennemi des croyances religieuses, ou encore comme un domaine scolaire rébarbatif, sont légion. Il est cependant peut-être utile de rappeler que, comme le disait Galilée, "le livre de la Nature est écrit dans le langage des mathématiques" (sans vouloir faire de scientisme!). C'est dans cet esprit que ce site Internet (et son PDF associé) aborde la mathématique appliquée pour les étudiants en sciences de la Nature, de la Terre et de la Vie, ainsi que pour tous ceux qui exercent une profession liée à ces diverses matières y compris la philosophie ou pour toute personne curieuse de s'informer de l'implication des sciences dans la vie quotidienne.

Le choix de traiter l'ingénierie ici comme une branche de la mathématique appliquée provient certainement du fait que l'ensemble des domaines de la physique (anciennement dénommée "philosophie naturelle") et la mathématique sont à ce jour tellement peu discernables que la médaille de Fields (la plus haute récompense de nos jours dans le domaine de la mathématique) a été décernée en 1990 au physicien Edward Witten, qui a utilisé des idées physiques pour redémontrer un théorème mathématique. Cette tendance n'est certainement pas fortuite, car nous pouvons observer que toute science, dès qu'elle cherche à atteindre une compréhension plus détaillée du sujet qu'elle étudie, voit finalement toujours sa course aboutir dans la mathématiques pure (la voie absolue par excellence...). Ainsi, pouvons-nous présager dans un futur lointain, la convergence de toutes les sciences (pures, exactes ou sociales) vers la mathématique pour la modélisation (lire à titre d'exemple le document PDF "L'explosion des mathématiques" disponible dans la rubrique Téléchargement du site).

Il peut parfois nous paraître difficile (à cause d'une crainte aussi obscure et irrationnelle que non justifiée des sciences pures chez une importante fraction de nos contemporains) de transmettre le

sentiment de beauté mathématique de la nature, de son harmonie la plus profonde et de la mécanique parfaitement huilée de l'Univers, à ceux qui ne connaissent que les rudiments du calcul formel. Le physicien R. Feynman a parlé un jour de "deux cultures": les gens qui ont, et ceux qui n'ont pas eu une compréhension suffisante des mathématiques pour apprécier la structure scientifique de la nature. Il est bien dommage qu'il y faille cependant des mathématiques et que celles-ci aient aussi mauvaise réputation. Pour l'anecdote, on prétend qu'un roi ayant demandé à Euclide de lui enseigner la géométrie se plaignit de sa difficulté. Euclide répondit: "il n'y a pas de voie royale". Les physiciens et mathématiciens ne peuvent se convertir à un autre langage. Si vous voulez apprendre à connaître la nature, à l'apprécier à sa juste valeur, vous devez comprendre son langage. Elle ne se révèle que sous cette forme et nous ne pouvons être prétentieux au point de lui demander de changer.

Au même titre, aucune discussion intellectuelle ne vous permettra de communiquer à un sourd ce que vous ressentez vraiment en écoutant de la musique. De même, toutes les discussions du monde resteront impuissantes à transmettre une compréhension intime de la nature à ceux de "l'autre culture". Les philosophes et théologiens peuvent essayer de vous donner des idées qualitatives sur l'Univers. Le fait que la méthode scientifique (au sens plein du terme) ne puisse convaincre le monde entier de sa justesse et de sa pureté, trouve peut-être sa cause dans l'horizon limité de certaines gens qui sont amenés à s'imaginer que l'homme ou qu'un autre concept intuitif, sentimental ou arbitraire est le centre de l'Univers (principe anthropocentrique).

Certes, dans le but de partager ce savoir mathématique, il est paradoxal de vouloir augmenter, avec notre travail, la liste déjà longue des ouvrages disponibles dans les bibliothèques, dans le commerce et sur l'Internet. Néanmoins, il faut être en mesure de présenter une argumentation solide qui justifie la création d'un tel site Internet (et son PDF associé) en comparaison à des ouvrages comme ceux de Feynman, Landau ou de Bourbaki. Voici donc les quelques arguments qui paraissent cependant susceptibles d'être présentés:

1. Le grand plaisir que je prends à cette entreprise ("garder la main" et progresser).
2. La passion du partage gratuit et sans frontières de la connaissance (et en français...).
3. Le caractère évolutif et pratique d'un site Internet libre (outils de recherche efficaces).
4. Le contenu évolutif en fonction de la demande!!!
5. La présentation rigoureuse avec des démonstrations simplifiées de beaucoup de concepts.
6. La présentation du plus grand nombre d'outils mathématiques utilisés dans les entreprises.
7. La possibilité pour les étudiants et professeurs de réutiliser le contenu par copier/coller.
8. Une notation constante et fixe, dans tout l'ouvrage, pour les opérateurs mathématiques, un langage clair et rigoureux sur tous les sujets abordés (critère des 3.C.: clair, complet et concis).
9. Rassembler le maximum d'informations sur les sciences pures et exactes en un seul ouvrage électronique (portatif), homogène et rigoureux.
10. Dégager, de toutes les pseudo-vérités, les seules vérités qui se démontrent.
11. Tirer bénéfice de l'évolution des méthodes pédagogiques scolaires qui utilisent l'Internet pour chercher la solution à des problèmes de mathématiques.
12. L'amélioration spectaculaire des logiciels automatiques de traduction et de la puissance des ordinateurs qui feront de ce site Internet (et son PDF associé), je le souhaite, une référence dans les domaines des sciences dures.

Et aussi... je considère que les résultats de la recherche individuelle sont la propriété de l'humanité et qu'ils doivent être mis à la disposition de tous ceux qui explorent où que ce soit les phénomènes de la nature. De cette façon le travail de chacun profite à tous, et c'est pour toute l'humanité que s'amassent nos connaissances ce qui est dans la tendance que permet l'Internet.

Je ne cache pas, que ma contribution se limite en grande partie à ce jour à celle d'un collectionneur qui glane ses informations dans les ouvrages des maîtres ou dans les publications ou pages Internet d'anonymes et qui complète et argumente les développements en les améliorant quand ceci est encore possible. Quant à ceux qui voudraient m'accuser de plagiat, ils devraient réfléchir au fait que les théorèmes présentés dans la plupart des ouvrages payants et disponibles dans le commerce ont été découverts et rédigés par leurs illustres prédécesseurs et que leur propre apport personnel a aussi constitué, comme le mien, à mettre toutes ces informations sous une forme claire et moderne quelques centaines d'années plus tard. De plus, il peut être vu comme douteux que l'on fasse payer l'accès à une culture qui est certainement la seule véritablement valable et juste dans ce bas monde et sur lequel il n'y a ni brevet, ni droit à la propriété intellectuelle.

Ce site Internet (et son PDF associé) reflète aussi mes propres limites intellectuelles. Bien que je m'efforce d'étudier autant de domaines scientifiques et mathématiques que possible, il est impossible de tous les maîtriser. Le site Internet (et son PDF associé) indique clairement mes propres intérêts et expériences en tant que consultant, mais aussi mes points forts et mes points faibles. Je suis responsable du choix des entrées, ainsi que, bien sûr, des éventuelles erreurs et imperfections.

Après avoir tenté un ordre de présentation rigoureux (linéaire) du sujet, j'ai décidé d'arranger ce site (et son PDF associé) dans un ordre plus pédagogique (thématique) et toujours avec des exemples d'applications concrets. Il est à mon avis très difficile de parler d'un si vaste sujet dans un ordre purement mathématique en une seule vie, c'est-à-dire lorsque les notions sont introduites une à une, à partir de celles déjà connues (où chaque théorie, opérateur, outil, etc. n'apparaîtrait pas avant sa définition). Un tel plan nécessiterait de couper le site (et son PDF associé), en morceaux qui ne sont plus thématiques. J'ai donc pris la décision de présenter les choses par ordre logique et non par ordre de nécessité. Le lecteur se heurtera donc, comme le rédacteur s'y est heurté, à l'extrême complexité du sujet.

Les conséquences de ce choix sont les suivantes:

1. Il faudra parfois admettre provisoirement certaines choses, quitte à les comprendre plus tard.
2. Il sera probablement nécessaire pour le lecteur de parcourir au moins deux fois l'ensemble de l'ouvrage. Lors de la première lecture, on appréhende l'essentiel et lors de la deuxième lecture, on comprend les détails (je félicite celui qui comprendrait toutes les subtilités du premier coup).
3. Il faut accepter le fait que certains sujets se répètent et qu'il y ait de nombreuses références croisées ainsi que remarques complémentaires.

Certains savent que pour chaque théorème et modèle mathématique, il existe quasiment toujours plusieurs méthodes de démonstration. J'ai toujours tenté de choisir celle qui me semblait la plus simple (par exemple en relativité il y a la présentation algébrique et matricielle et idem en physique quantique). L'objectif étant d'arriver de toute façon au même résultat.

Ce site (et son PDF associé) étant encore en cours de finalisation, il manque forcément des vérifications de convergences, de continuité et autres... (ce qui fera grimper au plafond les mathématiciens...) ! J'ai cependant évité (ou, dans le cas contraire, je le signale) les approximations habituelles de la physique et l'utilisation de l'analyse dimensionnelle, en y ayant recours le moins possible. J'essaie également d'éviter autant que possible des sujets dont les outils mathématiques n'ont pas au préalable été présentés et démontrés avec rigueur dans le corps de l'ouvrage.

Enfin, cet exposé, perfectible, n'est pas une référence absolue et contient des erreurs. Toute remarque

est donc la bienvenue. Je m'appliquerai, dans la mesure du possible, à corriger les faiblesses signalées et à apporter les modifications nécessaires au plus vite.

En revanche, alors que la mathématique est exacte et indiscutable, la physique théorique (ses modèles), reste interprétable dans le vocabulaire commun (mais pas dans le vocabulaire mathématique) et ses conclusions toutes relatives. Je ne peux que conseiller, lorsque vous parcourrez ce site (ou son PDF associé), de lire par vous-même et de ne pas subir d'influences extérieures. Il faut avoir l'esprit très (très) critique, ne rien prendre pour acquis et tout remettre en cause sans hésitation. Par ailleurs, le mot d'ordre du bon scientifique doit être: "Doute, doute, doute..., doute encore, et vérifie toujours.". Nous tenons aussi à rappeler que "rien de ce que l'on peut voir, entendre, sentir, toucher ou goûter, n'est ce qu'il a l'air d'être", ne vous fiez dès lors pas à votre expérience quotidienne pour tirer des conclusions trop hâtives, soyez critique, cartésien, rationnel et rigoureux dans vos développements, raisonnements et conclusions!

Je tiens à préciser à ceux qui tenteraient de trouver par eux-mêmes les résultats de certains développements présents sur ce site (ou de son PDF associé), de ne pas s'inquiéter s'ils n'y arrivent pas ou s'ils doutent d'eux à cause du temps passé à la résolution d'une équation ou problème: certaines théories qui nous semblent évidentes ou simples aujourd'hui, ont mis parfois plusieurs semaines, plusieurs mois, voire plusieurs années, pour être élaborées par des mathématiciens ou physiciens de renom!

J'ai également tenté de faire en sorte que ce site (et son PDF associé) soit agréable à l'oeil et à parcourir. Les concepteurs web professionnels voudront cependant bien excuser la mauvaise qualité du code HTML / PHP (qu'ils ne verront pas en partie...) / Javascript / CSS et l'abus de l'utilisation du biseautage et estampage de Photoshop ainsi que le choix d'une interface optimisée pour une résolution de 1024 x 768 et supérieure, mais le temps me manque pour épurer le code et réaliser des finitions graphiques correctes (de plus, je privilégie plutôt la qualité du contenu que le contenant).

Enfin, j'ai choisi d'écrire cet exposé à la première personne du pluriel ("nous"). Effectivement, la mathématique-physique n'est pas une science qui s'est faite ou évoluera grâce à un travail individuel mais à l'aide d'une collaboration intensive entre personnes reliées par la même passion et le même désir du Savoir. Ainsi, en faisant usage du "nous", il est rendu hommage aux hommes de science disparus, aux contemporains et aux futurs chercheurs pour le travail qu'ils effectueront dans le but de s'approcher de la vérité et de la sagesse.

1. MÉTHODES

La science est l'ensemble des efforts systématiques (observations scrupuleuses et hypothèses vraisemblables jusqu'à preuve du contraire) pour acquérir des connaissances sur notre environnement, pour les organiser et les synthétiser en lois et théories vérifiables et ayant pour principal objectif d'expliquer le "comment" des choses (et non pas le pourquoi!) souvent par une démarche à trois étapes:

- De quoi est-ce que nous disposons?
- Par où va-t-on passer?
- Quel est notre objectif?

Les scientifiques doivent soumettre leurs idées et résultats à la vérification et la reproduction indépendante de leurs pairs. Ils doivent abandonner ou modifier leurs conclusions lorsque confrontées à

des évidences plus complètes ou différentes. La crédibilité de la science s'appuie sur ce mécanisme d'autocorrection. L'histoire de la science montre que ce système fonctionne depuis très longtemps et ce même très bien par rapport à tous les autres. Dans chaque domaine, les progrès ont été spectaculaires. Toutefois, le système a parfois des ratés qu'il faut corriger avant que les petites dérives ne s'accumulent.

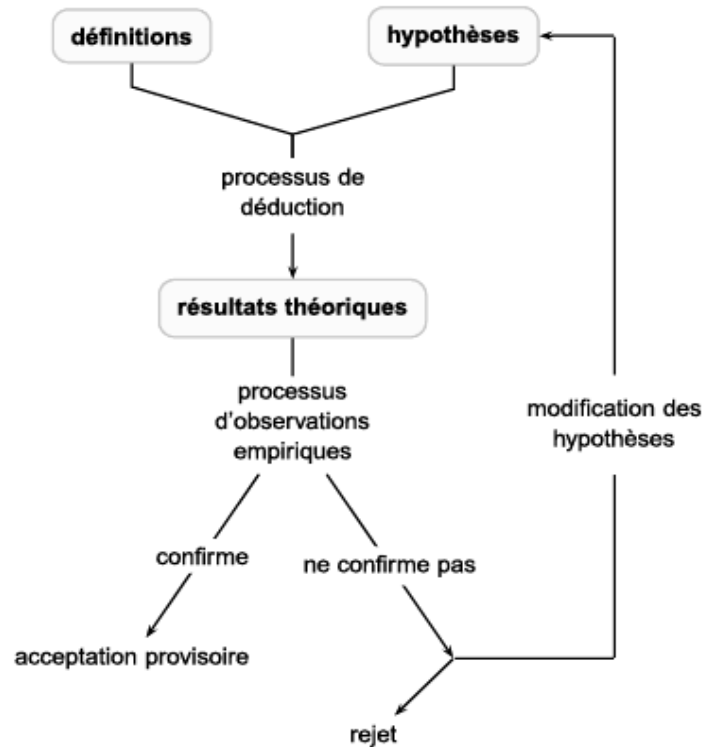
Le bémol est que les scientifiques sont des hommes. Ils ont les défauts de tous les hommes et, en particulier, la vanité, l'orgueil et la fatuité. De nos jours, il arrive que plusieurs personnes travaillant sur un même sujet depuis un certain temps développent une foi commune et croient qu'ils détiennent la vérité. Le chef de file de cette foi devient le Pape et distille des grands-messes. Le Pape qui se prend au jeu, prend sa mitre et son bâton de pèlerin pour évangéliser ses collègues hérétiques. Jusque-là, cela prête à sourire. Mais, comme dans les vraies religions, ils ont parfois la fâcheuse tendance de vouloir s'étendre au détriment de ceux qui ne croient pas. Certaines de ces "Eglises" n'hésitent pas à se comporter comme l'Inquisition. Ceux qui osent émettre une opinion différente se font incendier à chaque occasion, lors des congrès, voire sur leur lieu de travail. Certains jeunes chercheurs, en mal d'inspiration, préfèrent se convertir à cette religion dominante, pour devenir plus rapidement des dignitaires religieux à peu de frais, plutôt que des chercheurs innovants, voire iconoclastes. Le grand Pape écrit sa Bible pour diffuser sa pensée, l'impose à lire aux étudiants et aux nouveaux venus. Il formate ainsi la pensée des jeunes générations et assure son trône. C'est une attitude moyenâgeuse qui peut bloquer le progrès. Certains Papes vont jusqu'à croire que le fait d'être pris pour le pape dans un domaine leur donne automatiquement le même trône dans tous les autres domaines...

Cet avertissement, et les rappels qui vont suivre, doivent servir le scientifique à se remettre en question en faisant un bon usage de ce que nous pouvons considérer aujourd'hui comme les bonnes méthodes de travail (nous parlerons des principes de la méthode de Descartes plus loin) pour résoudre des problèmes ou développer des modèles théoriques.

Dans ce but, voici un tableau récapitulatif qui propose les différentes étapes que le scientifique devrait suivre lors de ses travaux en mathématique ou physique théorique (pour les définitions, voir juste après):

MATHÉMATIQUE	PHYSIQUE
1. Poser "l'hypothèse", la "conjecture" la "propriété" à démontrer de manière formelle ou en langage commun (les hypothèses étant notées H1., H2., ... les conjectures CJ1., CJ2.,... et les propriétés P1., P2., ...))	1. Poser correctement et de manière détaillée le ou les "problèmes" à résoudre de manière formelle ou en langage commun (les problèmes étant notés P1., P2., ...)
2. Définir les "axiomes" (sous-entendu non-démonstrables, indépendants et non-contradictoires) qui vont donner les points de départ et établir des restrictions aux développements (les axiomes étant notés A1., A2., ...).	2. Définir (ou énoncer) les "postulats" ou "principes" ou encore les "hypothèses" et "suppositions" (supposés non démontrables...) qui vont donner les points de départ et établir des restrictions aux développements (habituellement, les postulats et principes sont notés P1., P2., ... et les hypothèses H1., H2., ... en essayant d'éviter pour les postulats et principes, une confusion possible avec l'énoncé du ou des problèmes qui sont notés de la même manière).
<i>Remarque:</i> Parfois par abus, "propriétés", "conditions" et "axiomes" sont confondus alors que le concept d'axiome est beaucoup plus précis et profond. Dans la même idée, le mathématicien définit le vocabulaire spécialisé relié à des opérateurs mathématiques qui seront notés par D1., D2., D3., ...	<i>Remarque:</i> Il ne faut pas cependant oublier que la validité d'un modèle ne dépend pas du réalisme de ses hypothèses mais bien de la conformité de ses implications avec la réalité.
3. Les axiomes posés, tirer directement des "lemmes" ou des "propriétés" dont la validité en découle directement et qui préparent au développement du théorème censé valider l'hypothèse ou la conjecture de départ (les lemmes étant notés L1., L2., ... et les propriétés P1., P2.,...).	3. Une fois le "modèle théorique" développé vérifier les équations dimensionnelles pour déceler une éventuelle erreur dans les développements (ces vérifications étant notées VA1., VA2., ...).
4. Une fois le ou les "théorèmes" (notés T1., T2., ...) démontrés en tirer des "corollaires" (notés C1., C2., ...) et encore des propriétés (notées P1., P2., P3.,...).	4. Chercher les cas limites (dont les "singularités" font partie) du modèle pour en vérifier la validité intuitive (ces contrôles aux limites étant notés CL1., CL2., ...).
5. Tester la force ou l'utilité de sa ou ses conjectures ou hypothèses en démontrant la réciproque du théorème ou en la comparant avec des exemples à d'autres théories mathématiques pour voir si l'ensemble forme un tout cohérent (les exemples étant notés E1., E2., ...).	5. Tester expérimentalement le modèle théorique obtenu et soumettre le travail à comparaison avec d'autres équipes de recherche indépendantes. Le nouveau modèle doit prévoir des résultats expérimentaux observés et jamais observés (prédictions permettant de la falsifier). Si le modèle est validé alors il prend officiellement le statut de "Théorie".
6. D'éventuelles remarques peuvent être indiquées dans un ordre structuré et notées hiérarchiquement R1., R2., ...	6. D'éventuelles remarques peuvent être indiquées dans un ordre structuré et notées hiérarchiquement R1., R2., ...

Procéder comme dans le tableau ci-dessus est une base de travail possible pour travailler en mathématique et physique. Évidemment, procéder de façon propre et traditionnelle comme ci-dessus prend un petit plus de temps qu'en faisant un peu n'importe quoi, n'importe comment (c'est pour cela que la plupart des professeurs ne suivent pas ces règles, le temps leur manque cruellement pour couvrir tout le programme scolaire).



Signalons aussi une forme amusante scientifique des 8 commandements:

1. Les phénomènes tu observeras
Et jamais mesure tu ne falsifieras
(attention à l'erreur de confirmation: étudier que des phénomènes qui valident ses convictions)
2. Des hypothèses tu formuleras
Que par l'expérimentation tu testeras
3. L'expérience précisément tu décriras
Car ton collègue la reproduira
(attention au piège de la discipline narrative: coller les faits aux résultats désirés)
4. Fort de tes résultats
Une théorie tu bâtiras
5. De parcimonie tu useras
Et l'hypothèse la plus simple tu retiendras
6. Jamais vérité définitive ne sera (humilité épistémique)
Et toujours tu chercheras
7. D'une thèse non réfutable tu t'abstiendras
Car hors de la science elle restera
8. Tout échec sera pris comme une réussite
Car la science doit confirmer mais aussi infirmer

Remarques:

R1. Attention, il est très facile de faire des nouvelles théories physiques en alignant des mots. Cela s'appelle de la "philosophie" et les grecs ont pensé aux atomes en suivant cette méthode. Cela peut donc mener avec beaucoup de chance à une vraie théorie. Par contre il est bien plus difficile de faire une "**théorie prédictive**", c'est-à-dire avec des équations qui permettent de prédire le résultat d'une expérience.

R2. Toutefois ce qui sépare la mathématique de la physique est que, en mathématique, l'hypothèse est toujours vraie. Le discours mathématique n'est pas une démonstration d'une vérité extérieure à chercher, mais vise uniquement la cohérence. Ce qui doit être juste est le raisonnement.

Signalons que lorsque ces règles ne sont pas respectées, nous parlons de "**fraude scientifique**" (ce qui amène souvent à être licencié de son poste mais malheureusement on ne retire pas encore les diplômes quand cela arrive). En général, la fraude scientifique à proprement parler se présente sous trois grandes formes: le plagiat, la fabrication de données et l'altération de résultats défavorables à l'hypothèse avancée, l'omission d'exposition claire des hypothèses de travail et de données récoltées. À ces fraudes s'ajoutent des comportements qui posent problèmes concernant la qualité des travaux ou plus particulièrement l'éthique, comme ceux visant à augmenter en apparence la production (et par voie de fait la renommée du scientifique) en soumettant par exemple plusieurs fois la même publication en l'ayant un peu modifiée, les défauts de mentions de conflit d'intérêts, les expériences dangereuses, la non conservation des données primaires, etc.

1.1. MÉTHODE DE DESCARTES

Présentons maintenant les quatre principes de la méthode de Descartes qui, rappelons-le, est considéré comme le premier scientifique de l'histoire de par sa méthode d'analyse:

P1. Ne recevoir jamais aucune chose pour vraie que je ne la connusse évidemment être telle. C'est-à-dire, d'éviter soigneusement la précipitation et de ne comprendre rien de plus en mes jugements que ce qui se présenterait si clairement et si distinctement à mon esprit, que je n'eusse aucune occasion de le mettre en doute.

P2. De diviser chacune des difficultés que j'examinerais, en autant de parcelles qu'il se pourrait (observations scrupuleuses et hypothèses vraisemblables jusqu'à preuve du contraire), et qu'il serait requis pour les mieux résoudre.

P3. De conduire par ordre mes pensées, en commençant par les objets les plus simples et les plus aisés à connaître, pour monter peu à peu comme par degrés jusqu'à la connaissance des plus composés, et supposant même de l'ordre entre ceux qui ne se précèdent point naturellement les uns les autres.

P4. Faire partout des dénombrements si entiers et des revues si générales, que je fusse assuré de ne rien omettre.

2. VOCABULAIRE

La physique-mathématique, comme tout domaine de spécialisation, a son vocabulaire propre. Afin que le lecteur ne soit pas perdu dans la compréhension de certains textes qu'il pourra lire sur ce site (et son PDF associé), nous avons choisi d'exposer ici les quelques termes, abréviations et définitions fondamentaux à connaître.

Ainsi, le mathématicien aime bien terminer ses démonstrations (quand il pense qu'elles sont justes) par l'abréviation "**C.Q.F.D**" qui signifie "**Ce Qu'il Fallait Démontrer**" ou encore dans les hautes écoles par souci d'esthétisme et de traditions certains professeurs (et mêmes élèves) notent cela en latin "**Q.E.D**" qui signifie "**Quod Erat Demonstrandum**" (cela en jette...).

Et lors de définitions (elles sont nombreuses en mathématique et physique...) le scientifique fait souvent usage des terminologies suivantes:

- ... il suffit que ...
- ... si et seulement si ...
- ... nécessaire et suffisant ...
- ... signifie que ...

Les quatre ne sont pas équivalentes (identiques au sens strict). Car "il suffit que" correspond à une condition suffisante, mais pas à une condition nécessaire.

2.1. SUR LES SCIENCES

Il est important que nous définissions rigoureusement les différents types de sciences auxquelles l'être humain fait souvent référence. Effectivement, il semble qu'au 21ème siècle un abus de langage malsain s'instaure et qu'il ne devienne plus possible pour la population de distinguer la "qualité intrinsèque" d'une science d'une autre.

Remarque: Etymologiquement le mot "science" vient du latin "scientia" (connaissance) dont la racine est le verbe "scire" qui veut dire "savoir".

Cet abus de langage vient probablement du fait que les sciences pures et exactes perdent leurs illusions d'universalité et d'objectivité, dans le sens où elles s'auto-corrigent. Ceci ayant pour conséquence que certaines sciences sont reléguées au second plan et tentent d'en emprunter les méthodes, les principes et les origines pour créer une confusion quant à leurs distinctions.

En soi, la science cependant ne produit pas de vérité absolue. Par principe, une théorie scientifique est valable tant qu'elle permet de prédire des résultats mesurables et reproductibles. Mais les problèmes d'interprétation de ces résultats font partie de la philosophie naturelle.

Étant donné la diversité des phénomènes à étudier, au cours des siècles s'est constitué un nombre grandissant de disciplines comme la chimie, la biologie, la thermodynamique, etc. Toutes ces disciplines a priori hétéroclites ont pour socle commun la physique, pour langage la mathématique et comme principe élémentaire la méthode scientifique.

Ainsi, un petit rafraîchissement de mémoire semble nécessaire:

Définitions:

D1. Nous définissons par "**science pure**", tout ensemble de connaissances fondées sur un raisonnement rigoureux valable quel que soit le facteur (arbitraire) élémentaire choisi (nous disons alors "indépendant de la réalité sensible") et restreint au minimum nécessaire. Il n'y a que la mathématique (appelée souvent "reine des sciences") qui peut être classifiée dans cette catégorie.

D2. Nous définissons par "**science exacte**" ou "**science dure**", tout ensemble de connaissances fondées sur l'étude d'une observation, observation qui aura été transcrite sous forme symbolique (physique théorique). Principalement, le but des sciences exactes est non d'expliquer le "pourquoi" mais le "comment".

Remarque: Les deux définitions précédentes sont souvent incluses dans la définition de "**sciences déductives**" ou encore de "**sciences phénoménologiques**".

D3. Nous définissons par "**science de l'ingénieur**", tout ensemble de connaissances théoriques ou pratiques appliquées aux besoins de la société humaine tels que: l'électronique, la chimie, l'informatique, les télécommunications, la robotique, l'aérospatiale, biotechnologies...

D4. Nous définissons par "**science**" tout ensemble de connaissances fondées sur des études ou observations de faits dont l'interprétation n'a pas encore été retranscrite ni vérifiée avec la rigueur mathématique caractéristique des sciences qui précèdent, mais qui applique des raisonnements comparatifs statistiques. Nous incluons dans cette définition: la médecine (il faut cependant prendre garde au fait que certaines parties de la médecine étudient des phénomènes descriptifs sous forme mathématique tels que les réseaux de neurones ou autres phénomènes associés à des causes physiques connues), la sociologie, la psychologie, l'histoire, la biologie...

Selon le philosophe Karl Popper, une théorie n'est scientifiquement acceptable que si, telle qu'elle est présentée, elle peut être falsifiable, c'est à dire soumise à des tests expérimentaux. La "**connaissance scientifique**" est ainsi par définition l'ensemble des théories qui ont jusqu'alors résisté à la falsification. La science est donc par nature soumise en permanence à la remise en question.

D5. Nous définissons par "**science molle**" ou "**para-sciences**", tout ensemble de connaissances ou de pratiques qui sont actuellement fondées sur des faits invérifiables (non reproductibles scientifiquement) ni par l'expérience, ni par la mathématique. Nous incluons dans cette définition: l'astrologie, la théologie, le paranormal (qui est démolé par la science zététique), la graphologie...

D6. Nous définissons par "**sciences phénoménologiques**" ou "**sciences naturelles**", toute science qui n'est pas incluse dans les définitions précédentes (histoire, sociologie, psychologie, zoologie, biologie,...)

D7. Le "**scientisme**" est la doctrine fondamentale suivant laquelle il n'y a de vérité que dans la science.

Ce qui est intéressant dans cette doctrine, c'est que c'est certainement une des seules qui demande aux gens de devoir réfléchir par eux-mêmes et de comprendre l'environnement qui les entoure en remettant continuellement tout en question et sans ne jamais rien accepter comme acquis (...). De plus, les vraies sciences ont ceci d'extraordinaire qu'elles permettent de comprendre au-delà de ce que nous pouvons voir.

Mais enfin, la science, c'est la science, et rien de plus: une certaine mise en ordre, pas trop mal réussie, des choses qui ne conduit plus à la métaphysique comme du temps d'Aristote, mais qui n'a pas le prétention de nous livrer toute la réalité ni même le fond des choses visibles.

2.2. TERMINOLOGIE

Le tableau méthodique que nous avons présenté plus haut contient des termes qui peuvent peut-être vous sembler inconnus ou barbares. C'est la raison pour laquelle il nous semble fondamental de présenter les définitions de ces derniers, ainsi que de quelques autres tout aussi importants qui peuvent éviter des confusions malheureuses.

Définitions:

D1. Au-delà de son sens négatif, l'idée de "**problème**" renvoie à la première étape de la démarche scientifique. Formuler un problème est ainsi essentiel à sa résolution et permet de comprendre correctement ce qui fait problème et de voir ce qui doit être résolu.

Le concept de problème est intimement relié au concept "d'hypothèse" dont nous allons voir la définition ci-dessous.

D2. Une "**hypothèse**" est toujours, dans le cadre d'une théorie déjà constituée ou sous-jacente, une supposition en attente de confirmation ou d'infirmité qui tente d'expliquer un groupe de faits ou de prévoir l'apparition de faits nouveaux.

Ainsi, une hypothèse peut être à l'origine d'un problème théorique qu'il faudra formellement résoudre.

D3. Le "**postulat**" en physique correspond fréquemment à un principe (voir définition ci-dessous) dont l'admission est nécessaire pour établir une démonstration (nous sous-entendons que cela est une proposition non-démontrable).

L'équivalent mathématique (mais en plus rigoureux) du postulat est "l'axiome" dont nous verrons la définition plus loin.

D4. Un "**principe**" (parent proche du "postulat") est donc une proposition admise comme base d'un raisonnement ou une règle générale théorique qui guide la conduite des raisonnements qu'il faudra effectuer. En physique, il s'agit également d'une loi générale régissant un ensemble de phénomènes et vérifiée par l'exactitude de ses conséquences.

Remarque: le mot "principe" est utilisé avec abus dans les petites classes ou écoles d'ingénieurs par les professeurs ne sachant (ce qui est très rare), ou ne voulant (plutôt fréquent), ou ne pouvant faute de temps (quasi exclusivement), pas démontrer une relation.

L'équivalent du postulat ou du principe en mathématiques est "l'axiome" que nous définissons ainsi:

D5. Un "**axiome**" est une vérité ou proposition évidente par elle-même dont l'admission est nécessaire pour établir une démonstration.

Remarques:

R1. Nous pourrions dire que c'est quelque chose que nous posons comme une vérité pour le discours que nous nous proposons de tenir, comme une règle du jeu, et qu'elle n'a pas forcément par ailleurs une valeur de vérité universelle dans le monde sensible qui nous entoure).

R2. Les axiomes doivent toujours être indépendants entre eux (on ne doit pas pouvoir démontrer l'un à partir de l'autre), non contradictoires (nous disons également parfois qu'ils doivent être "consistants").

D6. Le "**corollaire**" est un terme malheureusement quasi inexistant en physique (à tort !) et qui est en fait une proposition résultant d'une vérité déjà démontrée. Nous pouvons également dire qu'un corollaire est une conséquence nécessaire et évidente d'un théorème (ou parfois d'un postulat en ce qui concerne la physique).

D7. Un "**lemme**" constitue une proposition déduite d'un ou de plusieurs postulats ou axiomes et dont la démonstration prépare celle d'un théorème.

Remarque: Le concept de "lemme" est lui aussi (et c'est malheureux) quasi réservé aux mathématiques.

D8. Une "**conjecture**" constitue une supposition ou opinion fondée sur la vraisemblance d'un résultat mathématique.

Remarque: Beaucoup de conjectures jouent un rôle un peu comparable à des lemmes, car elles sont des passages obligés pour obtenir d'importants résultats.

D8. Par-delà son sens faible de conjecture, une "**théorie**" ou "**théorème**" est un ensemble articulé autour d'une hypothèse et étayé par un ensemble de faits ou développements qui lui confèrent un contenu positif et rendent l'hypothèse bien fondée (ou tout au moins plausible dans le cas de la physique théorique)

D9. Une "**singularité**" est une indétermination d'un calcul qui intervient par l'apparition d'une division par le nombre zéro. Ce terme est aussi bien utilisé en mathématique qu'en physique.

D10. Une "**démonstration**" constitue un ensemble de procédures mathématiques à suivre pour démontrer le résultat déjà connu ou non d'un théorème.

D11. Si le mot "**paradoxe**" signifie étymologiquement: contraire à l'opinion commune, ce n'est cependant pas par pur goût de la provocation, mais bel et bien pour des raisons solides. Le "**sophisme**" quant à lui, est un énoncé volontairement provocateur, une proposition fausse reposant sur un raisonnement apparemment valide. Ainsi parle-t-on du fameux "paradoxe de Zénon", alors qu'il ne s'agit que d'un sophisme. Le paradoxe ne se réduit pas à de la fausseté, mais implique la coexistence de la vérité et de la fausseté, au point qu'on ne parvient plus à discriminer le vrai et le faux. Le paradoxe apparaît alors problème insoluble ou "**aporie**".

Remarque: Ajoutons que les grands paradoxes, par les interrogations qu'ils ont suscitées, ont fait progresser la science et amené des révolutions conceptuelles de grande ampleur, en mathématique comme en physique théorique (les paradoxes sur les ensembles et sur l'infini en mathématique, ceux à la base de la relativité et de la physique quantique).

2.3. SCIENCE ET FOI

Nous verrons qu'en science, une théorie est normalement incomplète, car elle ne peut décrire exhaustivement la complexité du monde réel. Il en est ainsi de toutes les théories, comme celle du Big Bang (cf. chapitre d'Astrophysique) ou de l'évolution des espèces (cf. chapitre de Dynamique Des Populations ou de Théorie Des Jeux).

Il convient de distinguer différents courants scientifiques:

- Le "**réalisme**" est une doctrine où les théories physiques ont pour objectif de décrire la réalité telle qu'elle est en soi, dans ses composantes inobservables.
- "**L'instrumentalisme**" est une doctrine où les théories sont des outils servant à prédire des observations mais qui ne décrivent pas la réalité en soi.
- Le "**fictionnalisme**" est le courant où le contenu référentiel (principes et postulats) des théories est un leurre, utile seulement pour assurer l'articulation linguistique des équations fondamentales.

Même si aujourd'hui les théories scientifiques ont le soutien de beaucoup de spécialistes, les théories alternatives ont des arguments valables et nous ne pouvons totalement les écarter. Pour autant, la création du monde en 7 jours décrite par la Bible ne peut plus être perçue comme un possible, et bien des croyants reconnaissent qu'une lecture littérale est peu compatible avec l'état actuel de nos connaissances et qu'il est plus sage de l'interpréter comme une parabole. Si la science ne fournit jamais de réponse définitive, il n'est plus possible de ne pas en tenir compte.

La foi (qu'elle soit religieuse, superstitieuse, pseudo-scientifique ou autre) a au contraire pour objectif de donner des vérités absolues d'une toute autre nature puisqu'elle relève d'une conviction personnelle invérifiable. En fait, l'une des fonctions des religions est de fournir du sens à des phénomènes qui ne sont pas explicables rationnellement. Les progrès de la connaissance entraînent donc parfois une remise en cause des dogmes religieux par la science.

A contrario, sauf à prétendre imposer sa foi (qui n'est autre qu'une conviction intimement personnelle et subjective) aux autres, il faut se défier de la tentation naturelle de qualifier de fait scientifiquement prouvé les extrapolations des modèles scientifiques au-delà de leur champ d'application.

Le mot "science" est comme nous l'avons déjà mentionné plus haut de plus en plus utilisé pour soutenir qu'il existe des preuves scientifiques là où il n'y a que croyance (certaines pages web de ce genre prolifèrent de plus en plus). Selon ses détracteurs c'est le cas du mouvement de scientologie. Selon ces derniers, nous devrions plutôt parler de "sciences occultes".

Les sciences occultes et sciences traditionnelles existent depuis l'Antiquité; elles consistent en un ensemble de connaissances et de pratiques mystérieuses ayant pour but de pénétrer et dominer les secrets de la nature. Au cours des derniers siècles, elles ont été progressivement exclues du champ de la science. Le philosophe Karl Popper s'est longuement interrogé sur la nature de la démarcation entre

science et pseudoscience. Après avoir remarqué qu'il est possible de trouver des observations pour confirmer à peu près n'importe quelle théorie, il propose une méthodologie fondée sur la réfutabilité. Une théorie doit selon lui, pour mériter le qualificatif de "scientifique", pouvoir garantir l'impossibilité de certains événements. Elle devient dès lors réfutable, donc (et alors seulement) apte à intégrer la science. Il suffirait en effet d'observer un de ces événements pour invalider la théorie, et s'orienter par conséquent sur une amélioration de celle-ci.

Enfin, citons Lavoisier: "Le physicien peut aussi, dans le silence de son laboratoire et de son cabinet, exercer des fonctions patriotiques; il peut espérer par ses travaux diminuer la masse des maux qui affligent bonheur et, n'eût-il contribué, par les routes nouvelles qu'il s'est ouvertes, qu'à prolonger de quelques années, de quelques jours, la vie moyenne des hommes, il pourrait aspirer aussi au titre glorieux de bienfaiteur de l'humanité."



Arithmétique

La mathématique est la forme ultime d'art contraint. (inconnu)

1. THÉORIE DE LA DÉMONSTRATION

Nous avons choisi de commencer l'étude de la mathématique appliquée par la théorie qui nous semble la plus fondamentale et la plus importante dans le domaine des sciences pures et exactes.

La théorie de la démonstration et du calcul propositionnel (logique) a trois objectifs dans le cadre de ce site:

1. Apprendre au lecteur comment raisonner et à démontrer et cela indépendamment de la spécialisation étudiée
2. Montrer que le processus d'une démonstration est indépendant du langage utilisé
3. Se préparer à la théorie de la logique et au théorème d'incomplétude de Gödel ainsi qu'aux automates (cf. chapitre d'Informatique Théorique).

Le théorème de Gödel est le point le plus passionnant car si nous définissons une religion comme un système de pensée qui contient des affirmations indémontrables, alors elle contient des éléments de foi, et Gödel nous enseigne que les mathématiques sont non seulement une religion, mais que c'est alors la seule religion capable de prouver qu'elle en est une!

Remarques:

R1. Il est (très) fortement conseillé de lire en parallèle à ce chapitre, ceux sur la théorie des automates et de l'algèbre de Boole disponibles dans la section d'Informatique Théorique du site.

R2. Il faut prendre cette théorie comme une curiosité sympathique mais qui n'amène fondamentalement pas grand-chose excepté des méthodes de travail/raisonnement. Par ailleurs, son objectif n'est pas de démontrer que tout est démontrable mais que toute démonstration peut se faire sur un langage commun à partir d'un certain nombre de règles.

Souvent, quand un étudiant arrive dans une classe supérieure, il a surtout appris à calculer, à utiliser des algorithmes mais relativement peu voire pas du tout à raisonner. Pour tous les raisonnements, le support visuel est un outil puissant, et les personnes qui ne voient pas qu'en traçant telle ou telle courbe ou droite la solution apparaît ou qui ne voient pas dans l'espace sont très pénalisées.

Lors des études secondaires, nous manipulons déjà des objets inconnus, mais c'est surtout pour faire des calculs, et quand nous raisonnons sur des objets représentés par des lettres, nous pouvons remplacer ceux-ci visuellement par un nombre réel, un vecteur, etc. A partir d'un certain niveau, nous demandons aux personnes de raisonner sur des structures plus abstraites, et donc de travailler sur des objets inconnus qui sont des éléments d'un ensemble lui-même inconnu, par exemple les éléments d'un groupe quelconque (cf. chapitre de Théorie Des Ensembles). Ce support visuel n'existe alors plus.

Nous demandons ainsi souvent aux étudiants de raisonner, de démontrer des propriétés, mais personne

ne leur a jamais appris à raisonner convenablement, à écrire des preuves. Si nous demandons à un étudiant de licence ce qu'est une démonstration, il a très probablement quelque difficulté à répondre. Il peut dire que c'est un texte dans lequel on trouve des mots-clés comme: "donc", "parce que", "si", "si et seulement si", "prenons un x tel que", "supposons que", "cherchons une contradiction", etc. Mais il est incapable de donner la grammaire de ces textes ni même leurs rudiments, et d'ailleurs, ses enseignants, s'ils n'ont pas suivi de cours, en seraient probablement incapables aussi.

Pour comprendre cette situation, rappelons que pour parler un enfant n'a pas besoin de connaître la grammaire. Il imite son entourage et cela marche très bien: un enfant de six ans sait utiliser des phrases déjà compliquées quant à la structure grammaticale sans avoir jamais fait de grammaire. La plupart des enseignants ne connaissent pas non plus la grammaire du raisonnement mais, chez eux, le processus d'imitation a bien marché et ils raisonnent correctement. L'expérience de la majorité des enseignants d'université montre que ce processus d'imitation marche bien chez les très bons étudiants, et alors il est suffisant, mais il marche beaucoup moins bien, voire pas du tout, chez beaucoup d'autres.

Tant que le degré de complexité est faible (notamment lors d'un raisonnement de type "équationnel"), la grammaire ne sert à rien, mais quand il augmente ou quand on ne comprend pas pourquoi quelque chose est faux, il devient nécessaire de faire un peu de grammaire pour pouvoir progresser. Les enseignants et les étudiants connaissent bien la situation suivante: dans un devoir, le correcteur a barré toute une page d'un grand trait rouge et mis "faux" dans la marge. Quand l'étudiant demande ce qui est faux, le correcteur ne peut que dire des choses du genre "ça n'a aucun rapport avec la démonstration demandée", "rien n'est juste", ..., ce qui n'aide évidemment pas l'étudiant à comprendre. Cela vient en partie, du fait que le texte rédigé par l'étudiant utilise les mots voulus mais dans un ordre plus ou moins aléatoire et qu'on ne peut donner de sens à l'assemblage de ces mots. De plus, l'enseignant n'a pas les outils nécessaires pour pouvoir expliquer ce qui ne va pas. Il faut donc les lui donner!

Ces outils existent mais sont assez récents. La théorie de la démonstration est une branche de la logique mathématique dont l'origine est la crise des fondements: il y a eu un doute sur ce que nous avons le "droit" de faire dans un raisonnement mathématique (voir la "crise des fondements" plus loin). Des paradoxes sont apparus, et il a alors été nécessaire de préciser les règles de démonstration et de vérifier que ces règles ne sont pas contradictoires. Cette théorie est apparue au début du 20ème siècle, ce qui est très peu puisque l'essentiel des mathématiques enseignées en première moitié de l'université est connu depuis le 16ème-17ème siècle.

1. LA CRISE DES FONDEMENTS

Pour les premiers Grecs, la géométrie était considérée comme la forme la plus haute du savoir, une puissante clé pour les mystères métaphysiques de l'Univers. Elle était plutôt une croyance mystique, et le lien entre le mysticisme et la religion était rendu explicite dans des cultes comme ceux des Pythagoriciens. Aucune culture n'a depuis déifié un homme pour avoir découvert un théorème géométrique! Plus tard, les mathématiques furent considérées comme le modèle d'une connaissance a priori dans la tradition aristotélicienne du rationalisme.

L'étonnement des Grecs pour les mathématiques ne nous a pas quittés, on le retrouve sous la traditionnelle métaphore des mathématiques comme "Reine des Science". Il s'est renforcé avec les succès spectaculaires des modèles mathématiques dans la science, succès que les Grecs (ignorant même la simple algèbre) n'avaient pas prévus. Depuis la découverte par Isaac Newton du calcul intégral et de la loi du carré inverse de la gravité, à la fin des années 1600, les sciences phénoménales et les plus hautes mathématiques étaient restées en étroite symbiose - au point qu'un formalisme mathématique

prédictif était devenu le signe distinctif d'une "science dure".

Après Newton, pendant les deux siècles qui suivirent, la science aspira à ce genre de rigueur et de pureté qui semblaient inhérentes aux mathématiques. La question métaphysique semblait simple: les mathématiques possédaient une connaissance a priori parfaite, et parmi les sciences, celles qui étaient capables de se mathématiser le plus parfaitement étaient les plus efficaces pour la prédiction des phénomènes. La connaissance parfaite consistait donc dans un formalisme mathématique qui, une fois atteint par la science et embrassant tous les aspects de la réalité, pouvait fonder une connaissance empirique a posteriori sur une logique rationnelle a priori. Ce fut dans cet esprit que Marie Jean-Antoine Nicolas de Caritat, marquis de Condorcet (philosophe et mathématicien français), entreprit d'imaginer la description de l'Univers entier comme un ensemble d'équations différentielles partielles se résolvant les unes après les autres.

La première faille dans cette image inspiratrice apparut dans la seconde moitié du 19^{ème} siècle, quand Riemann et Lobatchevsky prouvèrent séparément que l'axiome des parallèles d'Euclide pouvait être remplacé par d'autres qui produisaient des géométries "consistantes" (nous reviendrons sur ce terme plus loin). La géométrie de Riemann prenait modèle sur une sphère, celle de Lobatchevsky, sur la rotation d'un hyperboloïde.

L'impact de cette découverte a été obscurci plus tard par de grands chamboulements, mais sur le moment, elle fit un coup de tonnerre dans le monde intellectuel. L'existence de systèmes axiomatiques mutuellement inconsistants, et dont chacun pouvait servir de modèle à l'Univers phénoménal, remettait entièrement en question la relation entre les mathématiques et la théorie physique.

Quand on ne connaissait qu'Euclide, il n'y avait qu'une géométrie possible. On pouvait croire que les axiomes d'Euclide constituaient un genre de connaissance parfaite a priori sur la géométrie dans le monde phénoménal. Mais soudain, nous avons eu trois géométries, embarrassantes pour les subtilités métaphysiques.

Pourquoi aurions-nous à choisir entre les axiomes de la géométrie plane, sphérique et hyperbolique comme descriptions de la géométrie du réel? Parce que toutes les trois sont consistantes, nous ne pouvons en choisir aucune comme fondement a priori - le choix doit devenir empirique, basé sur leur pouvoir prédictif dans une situation donnée.

Bien sûr, les théoriciens de la physique ont longtemps été habitués à choisir des formalismes pour poser un problème scientifique. Mais il était admis largement, si ce n'est inconsciemment, que la nécessité de procéder ainsi était fonction de l'ignorance humaine, et qu'avec de la logique ou des mathématiques assez bonnes, on pouvait déduire le bon choix à partir de principes premiers, et produire des descriptions a priori de la réalité, qui devaient être confirmées après coup par une vérification empirique.

Cependant, la géométrie euclidienne, considérée pendant plusieurs centaines d'années comme le modèle de la perfection axiomatique des mathématiques, avait été détrônée. Si l'on ne pouvait connaître a priori quelque chose d'aussi fondamental que la géométrie dans l'espace, quel espoir restait-il pour une pure théorie rationnelle qui embrasserait la totalité de la nature ? Psychologiquement, Riemann et Lobatchevsky avaient frappé au cœur de l'entreprise mathématique telle qu'elle avait été conçue jusqu'alors.

De plus, Riemann et Lobatchevsky remettaient la nature de l'intuition mathématique en question. Il

avait été facile de croire implicitement que l'intuition mathématique était une forme de perception - une façon d'entrevoir le monde platonicien derrière la réalité. Mais avec deux autres géométries qui bousculaient celle d'Euclide, personne ne pouvait plus être sûr de savoir à quoi le monde ressemblait.

Les mathématiciens répondirent à ce double problème avec un excès de rigueur, en essayant d'appliquer la méthode axiomatique à toutes les mathématiques. Dans la période pré-axiomatique, les preuves reposaient souvent sur des intuitions communément admises de la "réalité" mathématique, qui ne pouvaient plus être considérées automatiquement comme valides.

La nouvelle façon de penser les mathématiques conduisait à une série de succès spectaculaires. Pourtant cela avait aussi un prix. La méthode axiomatique rendait la connexion entre les mathématiques et la réalité phénoménale toujours plus étroite. En même temps, des découvertes suggéraient que les axiomes mathématiques qui semblaient être consistants avec l'expérience phénoménale pouvaient entraîner de vertigineuses contradictions avec cette expérience.

La majorité des mathématiciens devinrent rapidement des "formalistes", soutenant que les mathématiques pures ne pouvaient qu'être considérées philosophiquement comme une sorte de jeu élaboré qui se jouait avec des signes sur le papier (c'est la théorie qui sous-tend la prophétique qualification des mathématiques de "système à contenu nul" par Robert Heinlein). La croyance "platonicienne" en la réalité des objets mathématiques, à l'ancienne manière, semblait bonne pour la poubelle, malgré le fait que les mathématiciens continuaient à se sentir comme les platoniciens durant le processus de découverte des mathématiques.

Philosophiquement, donc, la méthode axiomatique conduisait la plupart des mathématiciens à abandonner les croyances antérieures en la spécificité métaphysique des mathématiques. Elle produisit aussi la rupture contemporaine entre les mathématiques pures et appliquées. La plupart des grands mathématiciens du début de la période moderne - Newton, Leibniz, Fourier, Gauss et les autres - s'occupaient aussi de science phénoménale. La méthode axiomatique avait couvé l'idée moderne du mathématicien pur comme un super esthète, insoucieux de la physique. Ironiquement, le formalisme donnait aux purs mathématiciens un mauvais penchant à l'attitude platonicienne. Les chercheurs en mathématiques appliquées cessèrent de côtoyer les physiciens et apprirent à se mettre à leur traîne.

Ceci nous emmène au début du 20ème siècle. Pour la minorité assiégée des platoniciens, le pire était encore à venir. Cantor, Frege, Russell et Whitehead montrèrent que toutes les mathématiques pures pouvaient être construites sur le simple fondement axiomatique de la théorie des ensembles. Cela convenait parfaitement aux formalistes: les mathématiques se réunifiaient, du moins en principe, à partir d'un faisceau de petits jeux détachés d'un grand. Les platoniciens aussi étaient satisfaits, s'il en survenait une grande structure, clé de voûte consistante pour toutes les mathématiques, la spécificité métaphysique des mathématiques pouvait encore être sauvée.

D'une façon négative, pourtant, un platonicien eut le dernier mot. Kurt Gödel mit son grain de sable dans le programme formaliste d'axiomatisation quand il démontra que tout système d'axiomes assez puissant pour inclure les entiers devait être soit inconsistant (contenir des contradictions) soit incomplet (trop faible pour décider de la justesse ou de la fausseté de certaines affirmations du système). Et c'est plus ou moins où en sont les choses aujourd'hui. Les mathématiciens savent que de nombreuses tentatives pour faire avancer les mathématiques comme une connaissance a priori de l'Univers doivent se heurter à de nombreux paradoxes et à l'impossibilité de décider quel système axiomatique décrit les mathématiques réelles. Ils ont été réduits à espérer que les axiomatisations standards ne soient pas inconsistantes mais incomplètes, et à se demander anxieusement quelles contradictions ou quels

théorèmes indémontrables attendent d'être découverts ailleurs.

Cependant, sur le front de l'empirisme, les mathématiques étaient toujours un succès spectaculaire en tant qu'outil de construction théorique. Les grands succès de la physique du 20ème siècle (la relativité générale et la physique quantique) poussaient si loin hors du royaume de l'intuition physique, qu'ils ne pouvaient être compris qu'en méditant profondément sur leurs formalismes mathématiques, et en prolongeant leurs conclusions logiques, même lorsque ces conclusions semblaient sauvagement bizarres. Quelle ironie! Au moment même où la perception mathématique en venait à paraître toujours moins fiable dans les mathématiques pures, elle devenait toujours plus indispensable dans les sciences phénoménales.

À l'opposé de cet arrière-plan, l'applicabilité des mathématiques à la science phénoménale pose un problème plus épineux qu'il n'apparaît d'abord. Le rapport entre les modèles mathématiques et la prédiction des phénomènes est complexe, pas seulement dans la pratique mais dans le principe. D'autant plus complexe que, comme nous le savons maintenant, il y a des façons d'axiomatiser les mathématiques qui s'excluent!

Mais pourquoi existe-t-il seulement de bons choix de modèle mathématique ? C'est à dire, pourquoi y a-t-il un formalisme mathématique, par exemple pour la physique quantique, si productif qu'il prédit réellement la découverte de nouvelles particules observables ?

Pour répondre à cette question nous observerons qu'elle peut, aussi bien, fonctionner comme une sorte de définition. Pour beaucoup de systèmes phénoménaux, de tels formalismes prédictifs exacts n'ont pas été trouvés, et aucun ne semble plausible. Les poètes aiment marmonner sur le coeur des hommes, mais on peut trouver des exemples plus ordinaires: le climat, où le comportement d'une économie supérieure à celle d'un village, par exemple - systèmes si chaotiquement interdépendants que la prédiction exacte est effectivement impossible (pas seulement dans les faits mais en principe).

1.1. PARADOXES

Dès l'antiquité, certains logiciens avaient constaté la présence de nombreux paradoxes au sein de la rationalité. En fait, nous pouvons dire que malgré leur nombre, ces paradoxes ne sont que les illustrations d'un petit nombre de structures paradoxales. Attardons-nous à exposer à titre de culture générale les plus connus qui constituent la classe des "**propositions indécidables**".

Exemples:

E1. Le paradoxe de la classe des classes (Russell)

Il existe deux types de classes: celles qui se contiennent elles-mêmes (ou classes réflexives: la classe des ensembles non-vides, la classe des classes,...) et celles qui ne se contiennent pas elles-mêmes (ou classes irreflexives: la classe des travaux à rendre, la classe des oranges sanguines, ...). La question posée est la suivante: la classe des classes irreflexives est-elle elle-même réflexive ou irreflexive? Si elle est réflexive, elle se contient et se trouve rangée dans la classe des classes irreflexives qu'elle constitue, ce qui est contradictoire. Si elle est irreflexive, elle doit figurer dans la classe des classes irreflexives qu'elle constitue et devient ipso facto réflexive, nous sommes face à une nouvelle contradiction.

E2. Le paradoxe du bibliothécaire (Gonseth)

Dans une bibliothèque, il existe deux types de catalogues. Ceux qui se mentionnent eux-mêmes et ceux

qui ne se mentionnent pas. Un bibliothécaire doit dresser le catalogue de tous les catalogues qui ne se mentionnent pas eux-mêmes. Arrivé au terme de son travail, notre bibliothécaire se demande s'il convient ou non de mentionner le catalogue qu'il est précisément en train de rédiger. A ce moment, il est frappé de perplexité. S'il ne le mentionne pas, ce catalogue sera un catalogue qui ne se mentionne pas et qui devra dès lors figurer dans la liste des catalogues ne se mentionnant pas eux-mêmes. D'un autre côté, s'il le mentionne, ce catalogue deviendra un catalogue qui se mentionne et qui ne doit donc pas figurer dans ce catalogue, puisque celui-ci est le catalogue des catalogues qui ne se mentionnent pas.

E3. Le paradoxe du menteur (variante)

Définissons provisoirement le mensonge comme l'action de formuler une proposition fausse. Le poète crétois Epiménide affirme: "Tous les Crétois sont des menteurs", soit la proposition P . Comment décider de la valeur de vérité de P ? Si P est vraie, comme Epiménide est Crétois, P doit être fausse. Il faut donc que P soit fausse pour pouvoir être vraie, ce qui est contradictoire. P est donc fausse. Remarquons qu'on ne peut pas en déduire, comme dans le véritable paradoxe du menteur, que P doit aussi être vraie.

2. RAISONNEMENT HYPOTHETICO-DEDUCTIF

Le raisonnement hypothético-déductif est, nous le savons, la capacité qu'a l'apprenant de déduire des conclusions à partir de pures hypothèses et pas seulement d'une observation réelle. C'est un processus de réflexion qui tente de dégager une explication causale d'un phénomène quelconque (nous y reviendrons lors de nos premiers pas en physique). L'apprenant qui utilise ce type de raisonnement commence par formuler une hypothèse et essaie ensuite de confirmer ou d'infirmer son hypothèse selon le schéma synoptique ci-dessous:

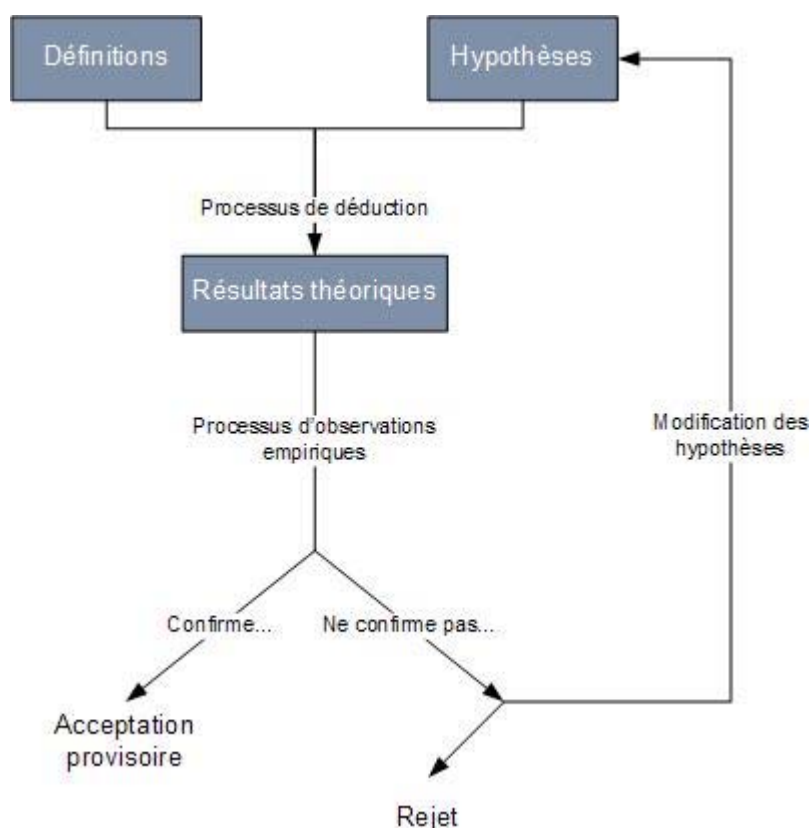


Figure: 1.1 - Diagramme synoptique du raisonnement hypothético-déductif

La procédure déductive consiste à tenir pour vrai, à titre provisoire, cette proposition première que nous appelons, en logique "le prédicat" (voir plus bas) et à en tirer toutes les conséquences logiquement nécessaires, c'est-à-dire à en rechercher les implications.

Exemple:

Soit la proposition P : "X est un homme", elle implique la proposition suivante Q : X est mortel.

L'expression $P \Rightarrow Q$ (si c'est un homme il est nécessairement mortel) est une implication prédicative (d'où le terme "prédicat"). Il n'y a pas dans cet exemple de cas où nous puissions énoncer P sans Q . Cet exemple est celui d'une implication stricte, telle que nous la trouvons dans le "syllogisme" (figure logique du raisonnement).

Remarque: Des spécialistes ont montré que le raisonnement hypothético-déductif s'élabore progressivement chez l'enfant, à partir de 6-7ans, et que ce type de raisonnement n'est utilisé systématiquement, en partant d'une fonction propositionnelle stricte qu'à partir de 11-12 ans.

3. CALCUL PROPOSITIONNEL

Le "**calcul propositionnel**" (ou "**logique propositionnelle**") est un préliminaire absolument indispensable pour aborder une formation en sciences, philosophie, droit, politique, économie, etc. Ce type de calcul autorise des procédures de décisions ou tests. Ceux-ci permettent de déterminer dans quel cas une expression (proposition) logique est vraie et en particulier si elle est toujours vraie.

Définitions:

D1. Une expression toujours vraie quel que soit le contenu linguistique des variables qui la composent est appelée une "**expression valide**", une "**tautologie**", ou encore une "**loi de la logique propositionnelle**".

D2. Une expression toujours fausse est appelée une "**contradiction**" ou "**antilogie**".

D3. Une expression qui est parfois vraie, parfois fausse est appelée une "**expression contingente**".

D4. Nous appelons "**assertion**" une expression dont nous pouvons dire sans ambiguïté si elle est vraie ou fausse.

D5. Le "**langage objet**" est le langage utilisé pour écrire les expressions logiques.

D6. Le "**métalangage**" est le langage utilisé pour parler du langage objet dans la langue courante.

Remarques:

R1. Il existe des expressions qui ne sont effectivement pas des assertions. Par exemple, l'énoncé: "cet énoncé est faux", est un paradoxe qui ne peut être ni vrai, ni faux.

R2. Soit une expression logique A . Si celle-ci est une tautologie, nous la notons fréquemment $\models A$ et si l'expression est une contradiction, nous la notons $A \models$.

3.1. PROPOSITIONS

Définition: En logique, une "**proposition**" est une affirmation qui a un sens. Cela veut dire que nous pouvons dire sans ambiguïté si cette affirmation est vraie (V) ou fausse (F). C'est ce que nous appelons le "**principe du tiers exclu**".

 Exemple:

"Je mens" n'est pas une proposition. Si nous supposons que cette affirmation est vraie, elle est une affirmation de sa propre invalidité, donc nous devrions conclure qu'elle est fausse. Mais si nous supposons qu'elle est fausse, alors l'auteur de cette affirmation ne ment pas, donc il dit la vérité, aussi la proposition serait vraie.

Définition: Une proposition en logique binaire (où les propositions sont soit vraies, soit fausses) n'est donc jamais vraie et fausse à la fois. C'est que nous appelons le "**principe de non-contradiction**".

Ainsi, une propriété sur l'ensemble E des propositions est une application P de E dans l'ensemble des "valeurs de vérité":

$$P: E \rightarrow \{V, F\} \quad (1.1)$$

Nous parlons de "**sous-ensemble associé**", lorsque la proposition engendre uniquement une partie E' de E et inversement.

 Exemple:

Dans $E = \mathbb{N}$, si $P(x)$ s'énonce " x est pair", alors $P = \{0, 2, 4, \dots, 2k, \dots\}$ ce qui est bien seulement un sous-ensemble associé de E mais de même cardinal (cf. chapitre Théorie Des Ensembles).

Définition: Soit P une propriété sur l'ensemble E . Une propriété Q sur E est une "**négation**" de P si et seulement si, pour tout $x \in E$:

- $Q(x)$ est F si $P(x)$ est V
- $Q(x)$ est V si $P(x)$ est F

Nous pouvons rassembler ces conditions dans une table dite "**table de vérité**":

P	Q
V	F
F	V

Tableau: 1.1 - Table de vérité des valeurs

Table que nous pouvons aussi trouver ou donc aussi écrire sous la forme plus explicite suivante:

P	Q
<i>Vrai</i>	<i>Faux</i>
<i>Faux</i>	<i>Vrai</i>

Tableau: 1.2 - Table de vérité des valeurs

ou encore sous forme binaire:

P	Q
1	0
0	1

Tableau: 1.3 - Table de vérité des valeurs

En d'autres termes, P et Q ont toujours des valeurs de vérité contraires. Nous noterons ce genre d'énoncé " Q est une négation de P ":

$$Q \Leftrightarrow \neg P \quad (1.2)$$

où le symbole $\neg$ est le "**connecteur de négation**".

Remarque: Les expressions doivent être des expressions bien formées (souvent abrégé "ebf"). Par définition, toute variable est une expression bien formée, alors $\neg P$ est une expression bien formée. Si P, Q sont des expressions bien formées, alors $P \Rightarrow Q$ est une expression bien formée (l'expression "je mens" n'est pas bien formée car elle se contredit elle-même).

3.2. CONNECTEURS

Il y a d'autres types de connecteurs en logique:

Soient P et Q deux propriétés définies sur le même ensemble E . $P \vee Q$ (lire " P ou Q ") est une propriété sur E définie par:

- $P \vee Q$ est vraie si au moins l'une des propriétés P, Q est vraie
- $P \vee Q$ est fausse sinon

Nous pouvons créer la table de vérité du "**connecteur OU**" ou "**connecteur de disjonction**" $\vee$:

P	Q	$P \vee Q$
V	V	V
V	F	V
F	V	V
F	F	F

Tableau: 1.4 - Table de vérité de OU

Il est facile de se convaincre que, si les parties P, Q de E sont respectivement associées aux propriétés P, Q que $P \cup Q$ (cf. chapitre Théorie Des Ensembles) est associé à $P \vee Q$.

$$\begin{aligned} P &\Leftrightarrow P \\ Q &\Leftrightarrow Q \\ P \vee Q &\Leftrightarrow P \cup Q \end{aligned} \quad (1.3)$$

Le connecteur $\vee$ est associatif. Pour s'en convaincre, il suffit de faire une table de vérité où nous

vérifions que:

$$[P \vee (Q \vee R)] \Leftrightarrow [(P \vee Q) \vee R] \quad (1.4)$$

Il existe également le "**connecteur ET**" ou "**connecteur de conjonction**" $\wedge$ pour quel que soient P, Q deux propriétés définies sur E , $P \wedge Q$ est une propriété sur E définie par:

- $P \wedge Q$ est vraie si toutes les deux propriétés P, Q sont vraies
- $P \wedge Q$ est fausse sinon

Nous pouvons créer la table de vérité du connecteur $\wedge$:

P	Q	$P \wedge Q$
V	V	V
V	F	F
F	V	F
F	F	F

Tableau: 1.5 - Table de vérité de ET

Il est également facile de se convaincre que, si les parties P, Q de E sont respectivement associées aux propriétés P, Q que $P \cap Q$ (cf. chapitre Théorie Des Ensembles) est associé à $P \wedge Q$:

$$\begin{aligned} P &\leftrightarrow P \\ Q &\leftrightarrow Q \\ P \wedge Q &\leftrightarrow P \cap Q \end{aligned} \quad (1.5)$$

Le connecteur $\wedge$ est associatif. Pour s'en convaincre, il suffit aussi de faire une table de vérité où nous vérifions que:

$$[P \wedge (Q \wedge R)] \Leftrightarrow [(P \wedge Q) \wedge R] \quad (1.6)$$

Les connecteurs $\wedge, \vee$ sont distributifs l'un sur l'autre. A l'aide d'une simple table de vérité, nous prouvons que:

$$[P \vee (Q \wedge R)] \Leftrightarrow [(P \vee Q) \wedge (P \vee R)] \quad (1.7)$$

ainsi que:

$$[P \wedge (Q \vee R)] \Leftrightarrow [(P \wedge Q) \vee (P \wedge R)] \quad (1.8)$$

Une négation de $P \vee Q$ est $[(\neg P) \wedge (\neg Q)]$ une négation de $P \wedge Q$ est $[(\neg P) \vee (\neg Q)]$ tel que pour résumer:

$$\begin{aligned} \neg(P \vee Q) &\Leftrightarrow [(\neg P) \wedge (\neg Q)] \\ \neg(P \wedge Q) &\Leftrightarrow [(\neg P) \vee (\neg Q)] \end{aligned} \quad (1.9)$$

A nouveau, ces propriétés peuvent se démontrer par une simple table de vérité.

Remarque: Pour voir les détails de tous les opérateurs logiques, le lecteur devra se rendre dans le chapitre d'Algèbre De Boole (cf. section d'Informatique Théorique) où l'identité, la double négation, l'idempotence, l'associativité, la distributivité, les relations de De Morgan sont présentées plus formellement.

Revenons maintenant sur le "**connecteur d'implication logique**" appelé aussi parfois le "**conditionnel**" noté " $\Rightarrow$ "

Remarque: Dans certains ouvrages sur le calcul propositionnel, ce connecteur est noté " $\supset$ " et dans le cadre de la théorie de la démonstration nous lui préférons souvent le symbole " $\rightarrow$ ".

Soient P, Q deux propriétés sur E . $P \Rightarrow Q$ est une propriété sur E définie par:

- $P \Rightarrow Q$ est fausse si P est vraie et Q fausse
- $P \Rightarrow Q$ est vraie sinon

En d'autres termes, P implique logiquement Q signifie que Q est vraie pour toute évaluation pour laquelle P est vraie. L'implication représente donc le "si... alors.."

Si nous écrivons la table de vérité de l'implication (attention à l'avant-dernière ligne !!!):

P	Q	$P \Rightarrow Q$
V	V	V
V	F	F
F	V	V
F	F	V

Tableau: 1.6 - Table de vérité de l'implication

Si $P \Rightarrow Q$, nous pouvons dire que pour que Q soit vraie, il suffit que P soit vraie (effectivement l'implication sera vraie si P est vraie ou fausse selon la table de vérité). Donc P est une condition suffisante de Q (mais non nécessaire!). D'un autre côté, $P \Rightarrow Q$ est équivalent à $\neg Q \Rightarrow \neg P$. Donc, si Q est fausse, il est impossible que P soit vraie (pour que l'implication reste vraie bien sûr!). Donc finalement Q est une condition nécessaire de P .

Exemples:

E1. Soit la proposition: "Si tu obtiens ton diplôme, je t'achète un ordinateur"

Parmi tous les cas, un seul correspond à une promesse non tenue: celui où l'enfant a son diplôme, et n'a toujours pas d'ordinateur (deuxième ligne dans le tableau).

Et le cas où il n'a pas le diplôme, mais reçoit quand même un ordinateur? Il est possible qu'il ait été longtemps malade et a raté un semestre, et le père a le droit d'être bon.

Que signifie cette promesse, que nous écrivons aussi: "Tu as ton diplôme $\Rightarrow$ je t'achète un ordinateur" ? Exactement ceci:

- Si tu as ton diplôme, c'est sûr, je t'achète un ordinateur (je ne peux pas ne pas l'acheter)
- Si tu n'as pas ton diplôme, je n'ai rien dit

E2. De toute proposition fausse nous pouvons déduire toute proposition (deux dernières lignes)

C'est un exemple plutôt anecdotique: dans un cours de Russell portant sur le fait que d'une proposition fausse, toute proposition peut être déduite, un étudiant lui posa la question suivante:

- "Prétendez-vous que de $2 + 2 = 5$, il s'ensuit que vous êtes le pape ? "
- "Oui", fit Russell
- "Et pourriez-vous le prouver !", demanda l'étudiant sceptique
- "Certainement", répliqua Russell, qui proposa sur le champ la démonstration suivante.

(1) Supposons que $2 + 2 = 5$

(2) Soustrayons 2 de chaque membre de l'égalité, nous obtenons $2 = 3$

(3) Par symétrie, $3 = 2$

(4) Soustrayant 1 de chaque côté, il vient $2 = 1$

Maintenant le pape et moi sommes deux. Puisque $2 = 1$, le pape et moi sommes un. Par suite, je suis le pape.

Sur ce ...

Le connecteur d'implication est essentiel en mathématiques, philosophie, etc. C'est un des fondements de toute démonstration, preuve ou déduction.

Le connecteur d'implication a comme propriétés (vérifiables à l'aide de la table de vérité ci-dessous):

$$\begin{aligned} P \Rightarrow Q &\Leftrightarrow [(\neg Q) \Rightarrow (\neg P)] \\ P \Rightarrow Q &\Leftrightarrow [(\neg P) \vee Q] \end{aligned} \quad (1.10)$$

conséquence de la dernière propriété (à nouveau vérifiable par une table de vérité):

$$\neg(P \Rightarrow Q) \Leftrightarrow [P \wedge (\neg Q)] \quad (1.11)$$

Le "**connecteur d'équivalence logique**" ou "**biconditionnel**" noté " $\Leftrightarrow$ " ou " $\Leftrightarrow$ " signifiant par définition que:

$$(P \Leftrightarrow Q) \Leftrightarrow (P \Rightarrow Q) \wedge (Q \Rightarrow P) \quad (1.12)$$

en d'autres termes, la première expression a la même valeur pour toute évaluation de la deuxième.

Ce que nous pouvons vérifier à l'aide d'une table de vérité:

P	Q	$P \Rightarrow Q$	$Q \Rightarrow P$	$P \Leftrightarrow Q$
V	V	V	V	V
V	F	F	V	F
F	V	V	F	F
F	F	V	V	V

Tableau: 1.7 - Table de vérité de l'équivalence logique

$P \Leftrightarrow Q$ signifie bien (lorsqu'il est vrai!) que " P et Q ont toujours la même valeur de vérité" ou encore " P et Q sont équivalents". C'est vrai si P et Q ont même valeur, faux dans tout cas contraire.

Bien évidemment (c'est une tautologie):

$$\neg(P \Leftrightarrow Q) \Rightarrow ((P \Rightarrow (\neg Q)) \quad (1.13)$$

La relation $P \Leftrightarrow Q$ équivaut donc à ce que P soit une condition nécessaire et suffisante de Q et à ce que Q soit une condition nécessaire et suffisante de P .

La conclusion, est que les conditions de type "nécessaire, suffisant, nécessaire et suffisant" peuvent être reformulées avec les termes "seulement si", "si", "si et seulement si".

Ainsi:

1. $P \Rightarrow Q$ traduit le fait que Q est une condition **nécessaire** pour P ou dit autrement, P est vraie **seulement si** Q est vraie (dans la table de vérité, lorsque $P \Rightarrow Q$ prend la valeur 1 on constate bien que P vaut 1 **seulement si** Q vaut 1 aussi). On dit aussi, **si** P est vraie **alors** Q est vraie.
2. $P \Leftarrow Q$ ou ce qui revient au même $Q \Rightarrow P$ traduit le fait que Q est une condition **suffisante** pour P ou dit autrement, P est vraie **si** Q est vraie (dans la table de vérité, lorsque $Q \Rightarrow P$ prend la valeur 1 on constate bien que P vaut 1 **si** Q vaut 1 aussi).
3. $P \Leftrightarrow Q$ traduit le fait que Q est une condition **nécessaire et suffisante** pour P ou dit autrement, P est vraie **si et seulement si** Q est vraie (dans la table de vérité, lorsque $P \Leftrightarrow Q$ prend la valeur 1 on constate bien que P vaut 1 **si** Q vaut 1 et **seulement si** Q vaut 1).

Remarque: L'expression "si et seulement si" correspond donc à une équivalence logique et ne peut être utilisée que pour décrire une bi-implication!!

La première étape du calcul propositionnel est donc la formalisation des énoncés du langage naturel. Pour réaliser ce travail, le calcul propositionnel fournit finalement trois types d'outils :

1. Les "variables propositionnelles" ($P, Q, R, \dots$) symbolisent des propositions simples quelconques. Si la même variable apparaît plusieurs fois, elle symbolise chaque fois la même proposition.
2. Les cinq opérateurs logiques: $\neg, \wedge, \vee, \Rightarrow, \Leftrightarrow$
3. Les signes de ponctuation se réduisent aux seules parenthèses ouvrante et fermante qui organisent la lecture de manière à éviter toute ambiguïté.

Voici un tableau récapitulatif:

Description	Symbole	Utilisation
La " négation " est un opérateur qui ne porte que sur une proposition; il est unaire ou monadique. " <i>Il ne pleut pas</i> " s'écrit $\neg P$. Cet énoncé est vrai si et seulement si P est faux (dans ce cas s'il est faux qu'il pleut). L'usage classique de la négation est caractérisé par la loi de double négation: $\neg\neg P$ est équivalent à P .	$\neg$	$\neg P$
La " conjonction " ou " produit logique " est un opérateur binaire; elle met en relation deux propositions. " <i>Tout homme est mortel ET Ma voiture perd de l'huile</i> " s'écrit $(P \wedge Q)$. Cette dernière expression est vraie si et seulement si P est vrai et Q est vrai.	$\wedge$	$(P \wedge Q)$
La " disjonction " ou " somme logique " est, elle aussi, un opérateur binaire. $(P \vee Q)$; est vraie si et seulement si P est vraie ou Q est vraie. Nous pouvons comprendre ce OU de deux façons : soit de manière inclusive, soit de manière exclusive. Dans le premier cas $(P \vee Q)$ est vrai si P est vraie, si Q est vraie ou si P et Q sont tous deux vrais. Dans le second cas, $(P \vee Q)$ est vraie si P est vraie ou si Q est vraie mais pas si les deux le sont. La disjonction du calcul propositionnel est le OU inclusif et on donne au OU exclusif le nom "d'alternative".	$\vee$	$(P \vee Q)$
" L'implication " est également un opérateur binaire. Elle correspond, en gros, au schéma linguistique " <i>Si...alors...</i> ". " <i>Si j'ai le temps, j'irai au cinéma</i> " s'écrit $P \Rightarrow Q$. $P \Rightarrow Q$ est fausse si P est vrai et Q est faux. Si le conséquent (ici Q) est vrai, l'implication $(P \Rightarrow Q)$ est vraie. Lorsque l'antécédente (ici P) est fausse, l'implication est toujours vraie. Cette dernière remarque peut être comprise si l'on se réfère à des énoncés de type : " <i>Si on pouvait mettre Paris en bouteille, on utiliserait la tour Eiffel comme bouchon.</i> " En résumé, une implication est fausse si et seulement si son antécédente est vraie et son conséquent est fausse.	$\Rightarrow$	$(P \Rightarrow Q)$
La " bi-implication " est, elle aussi, binaire : elle symbolise les expressions "... si et seulement si..." et "... est équivalent à..." L'équivalence entre deux propositions est vraie si celles-ci ont la même valeur de vérité. La bi-implication exprime donc aussi une forme	$\Leftrightarrow$	$(P \Leftrightarrow Q)$

d'identité et c'est pourquoi elle est souvent utilisée dans les définitions.		
--	--	--

Tableau: 1.8 - Récapitulatif des opérateurs

Il est possible d'établir des équivalences entre ces opérateurs. Nous avons déjà vu comment le biconditionnel pouvait se définir comme un produit de conditionnels réciproques, voyons maintenant d'autres équivalences:

$$\begin{array}{l}
 (P \Rightarrow Q) \Leftrightarrow \neg(P \wedge \neg Q) \\
 (P \Rightarrow Q) \Leftrightarrow (\neg P \vee Q) \\
 (P \vee Q) \Leftrightarrow (\neg P \Rightarrow Q) \\
 (P \wedge Q) \Leftrightarrow \neg(P \Rightarrow \neg Q)
 \end{array} \quad (1.14)$$

Remarque: Les opérateurs classiques $\wedge, \vee, \Leftrightarrow$ peuvent donc être définis à l'aide de $\neg, \Rightarrow$ grâce aux lois d'équivalence entre opérateurs.

Sont à noter également les deux relations de De Morgan (cf. chapitre d'Algèbre de Boole) :

$$\begin{array}{l}
 \neg(P \vee Q) \Leftrightarrow (\neg P \wedge \neg Q) \\
 \neg(P \wedge Q) \Leftrightarrow (\neg P \vee \neg Q)
 \end{array} \quad (1.15)$$

Elles permettent de transformer la disjonction en conjonction et vice-versa:

$$\begin{array}{l}
 (P \vee Q) \Leftrightarrow \neg(\neg P \wedge \neg Q) \\
 (P \wedge Q) \Leftrightarrow \neg(\neg P \vee \neg Q)
 \end{array} \quad (1.16)$$

3.3. PROCÉDURES DE DÉCISION

Nous avons introduit précédemment les éléments de base nous permettant d'opérer sur des expressions à partir de propriétés (variables propositionnelles) sans toutefois dire grand-chose quant à la manipulation de ces expressions. Alors, il convient maintenant de savoir qu'en calcul propositionnel il existe deux manières d'établir qu'une proposition est une loi de la logique propositionnelle. Nous pouvons soit:

1. Employer des procédures non axiomatisées
2. Recourir à des procédures axiomatiques et démonstratives

Remarque: Dans de nombreux ouvrages ces procédures sont présentées avant même la structure du langage propositionnel. Nous avons choisi de faire le contraire pensant que l'approche serait plus aisée.

3.3.1. PROCÉDURES DE DÉCISIONS NON AXIOMATISÉES

Plusieurs de ces méthodes existent mais nous nous limiterons ici à la plus simple et à la plus parlante d'entre elles, celle du calcul matriciel, souvent appelée aussi "**méthodes des tables de vérité**".

La procédure de construction est comme nous l'avons vu précédemment assez simple. Effectivement, la valeur de vérité d'une expression complexe est fonction de la valeur vérité des énoncés plus simples qui la composent, et finalement fonction de la valeur de vérité des variables propositionnelles qui la composent. En envisageant toutes les combinaisons possibles des valeurs de vérité des variables propositionnelles, nous pouvons déterminer les valeurs de vérité de l'expression complexe.

Les tables de vérité, comme nous l'avons vu, permettent donc de décider, à propos de toute proposition, si celle-ci est une tautologie (toujours vraie), une contradiction (toujours fausse) ou une expression contingente (parfois vraie, parfois fausse).

Nous pouvons ainsi distinguer quatre façons de combiner les variables propositionnelles, les parenthèses et les connecteurs:

	Nom	Description	Exemple
1	Enoncé mal formé	Non-sens. Ni vrai, ni faux	$(\forall P)Q$
2	Tautologie	Enoncé toujours vrai	$P \vee \neg P$
3	Contradiction	Enoncé toujours faux	$(P \wedge \neg P)$
4	Enoncé contingent	Enoncé parfois vrai, parfois faux	$(P \vee Q)$

Tableau: 1.9 - Combinaison de variables propositionnelles

La méthode des tables de vérité permet de déterminer le type d'expression bien formée face auquel nous nous trouvons. Elle n'exige en principe aucune invention, c'est une procédure mécanique. Les procédures axiomatisées, en revanche, ne sont pas entièrement mécaniques. Inventer une démonstration dans le cadre d'un système axiomatisé demande parfois de l'habileté, de l'habitude ou de la chance. Pour ce qui est des tables de vérité, voici la marche à suivre:

Lorsqu'on se trouve face à une expression bien formée, ou fonction de vérité, nous commençons par déterminer à combien de variables propositionnelles distinctes nous avons affaire. Ensuite, nous examinons les différents arguments qui constituent cette expression. Nous construisons alors un tableau comprenant 2^n rangées (n étant le nombre de variables) et un nombre de colonnes égal au nombre d'arguments plus des colonnes pour l'expression elle-même et ses autres composantes. Nous attribuons alors aux variables les différentes combinaisons de vérité et de fausseté qui peuvent leur être conférées (la vérité est exprimée dans la table par un 1 et la fausseté par un 0). Chacune des rangées correspond à un monde possible et la totalité des rangées constitue l'ensemble des mondes possibles. Il existe, par exemple, un monde possible dans lequel P est une proposition vraie tandis que Q est fausse.

3.3.2. PROCÉDURES DE DÉCISIONS AXIOMATISÉES

L'axiomatisation d'une théorie implique, outre la formalisation de celle-ci, que nous partions d'un nombre fini d'axiomes et que, grâce à la transformation réglée de ces derniers, nous puissions obtenir tous les théorèmes de cette théorie. Nous partons donc de quelques axiomes dont la vérité est posée (et non démontrée). Nous déterminons des règles de déduction permettant de manipuler les axiomes ou toute expression obtenue à partir de ceux-ci. L'enchaînement de ces déductions est une démonstration

qui conduit à un théorème, à une loi.

Nous allons sommairement présenter deux systèmes axiomatiques, chacun étant constitué d'axiomes utilisant deux règles dites "**règles d'inférence**" (règles intuitives) particulières:

Règle 1. Le "**modus ponens**": si nous avons prouvé A et $A \Rightarrow B$, alors nous pouvons déduire B . A est appelé la "**prémisse mineure**" et $A \Rightarrow B$ la prémisse majeure de la règle du modus ponens.

 **Exemple:** De $x > y$ et $(x > y) \Rightarrow (y < x)$ nous pouvons déduire $y < x$

Règle 2. La "**substitution**": nous pouvons dans un schéma d'axiomes remplacer une lettre par une formule quelconque, pourvu que toutes les lettres identiques soient remplacées par des formules identiques.

Donnons à titre d'exemple, deux systèmes axiomatiques: le système axiomatique de Whitehead et Russell, le système axiomatique de Lukasiewicz.

1. Le système axiomatique de Whitehead et Russell adopte comme symboles primitifs $\neg, \vee$ et définit $\Rightarrow, \wedge, \Leftrightarrow$ à partir de ces derniers de la manière suivante (relations facilement vérifiables à l'aide de tables de vérité):

$$\begin{aligned} (A \Rightarrow B) &\Leftrightarrow \neg A \vee B \\ (A \wedge B) &\Leftrightarrow \neg(\neg A \vee \neg B) \\ (A \Leftrightarrow B) &\Leftrightarrow (A \Rightarrow B) \wedge (B \Rightarrow A) \end{aligned} \quad (1.17)$$

nous avons déjà présenté plus haut quelques-uns de ces éléments.

Ce système comprend cinq axiomes, assez évidents en soi plus les deux règles d'inférence. Les axiomes sont donnés ici en utilisant des symboles non primitifs, comme le faisaient Whitehead et Russell:

$$A1. (A \vee A) \Rightarrow A$$

$$A2. B \Rightarrow (A \vee B)$$

$$A3. (A \vee B) \Rightarrow (B \vee A)$$

$$A4. (A \vee (B \vee C)) \Rightarrow (B \vee (A \vee C))$$

$$A5. (B \Rightarrow C) \Rightarrow ((A \vee B) \Rightarrow (A \vee C))$$

Remarque: Ces cinq axiomes ne sont pas indépendants les uns des autres. Le quatrième peut être obtenu à partir des quatre autres.

 **Exemple:**

Pour prouver $\neg A \vee A$, nous pouvons procéder ainsi:

(1)	$B \Rightarrow (A \vee B)$	Axiome A2
(2)	$A \Rightarrow (A \vee A)$	(1) et subst.
(3)	$(B \Rightarrow C) \Rightarrow ((A \vee B) \Rightarrow (A \vee C))$	Axiome A5 (nécessaire)
(4)	$(B \Rightarrow C) \Rightarrow ((\neg A \vee B) \Rightarrow (\neg A \vee C))$	(3) et subst.
(5)	$(B \Rightarrow C) \Rightarrow ((A \Rightarrow B) \Rightarrow (A \Rightarrow C))$	(4) et définition de $\Rightarrow$
(6)	$((A \vee A) \Rightarrow A) \Rightarrow ((A \Rightarrow (A \vee A)) \Rightarrow (A \Rightarrow A))$	(5) et subst.
(7)	$(A \Rightarrow (A \vee A)) \Rightarrow (A \Rightarrow A)$	(6) (modus ponens)
(8)	$(A \Rightarrow A) \Rightarrow (A \Rightarrow A)$	(2) et (7) et axiome A1
(9)	$A \Rightarrow A$	(8) modus ponens
(10)	$\neg A \vee A$	(9) et déf. de $\Rightarrow$

(1.18)

2. Le système axiomatique de Łukasiewicz comprend les trois axiomes suivants, plus les deux règles d'inférences (modus ponens et substitution):

$$A1. (A \Rightarrow B) \Rightarrow ((B \Rightarrow C) \Rightarrow (A \Rightarrow C))$$

$$A2. A \Rightarrow (\neg A \Rightarrow B)$$

$$A3. (\neg A \Rightarrow A) \Rightarrow A$$

Voici des preuves des deux premiers axiomes, dans le système de Whitehead et Russel. Ce sont les formules (6) et (16) de la dérivation suivante:

(1)	$(A \vee (B \vee C)) \Rightarrow (B \vee (A \vee C))$	Axiome A4
(2)	$(\neg(B \Rightarrow C) \vee (\neg(A \vee B) \vee (A \vee C)))$ $\Rightarrow (\neg(A \vee B) \vee (\neg(B \Rightarrow C) \vee (A \vee C)))$	(1) et subst.
(3)	$\neg(A \vee B) \vee (\neg(B \Rightarrow C) \vee (A \vee C))$	A4 sur (2) et modus ponens
(4)	$(A \vee B) \Rightarrow ((B \Rightarrow C) \Rightarrow (A \vee C))$	(3) par déf. de $\Rightarrow$
(5)	$(\neg A \vee B) \Rightarrow ((B \Rightarrow C) \Rightarrow (\neg A \vee C))$	(4) et subst.
(6)	$(A \Rightarrow B) \Rightarrow ((B \Rightarrow C) \Rightarrow (A \Rightarrow C))$	(5) et déf. de $\Rightarrow$
(7)	$(B \Rightarrow (A \vee B)) \Rightarrow \left(\begin{array}{l} ((A \vee B) \Rightarrow (B \vee A)) \\ \Rightarrow (B \Rightarrow (B \vee A)) \end{array} \right)$	(6) et subst.
(8)	$((A \vee B) \Rightarrow (B \vee A)) \Rightarrow (B \Rightarrow (B \vee A))$	(7) modus ponens
(9)	$B \Rightarrow (B \vee A)$	(8) modus ponens
(10)	$\neg B \Rightarrow (\neg B \vee A)$	(9) et subst.
(11)	$\neg\neg B \vee (\neg B \vee A)$	(10) et déf. de $\Rightarrow$
(12)	$(\neg\neg B \vee (\neg B \vee A)) \Rightarrow (\neg B \vee (\neg\neg B \vee A))$	Axiome A4 + subst. avec (11)
(13)	$\neg B \vee (\neg\neg B \vee A)$	(12) modus ponens
(14)	$B \Rightarrow (\neg\neg B \vee A)$	(13) et déf. de $\Rightarrow$
(15)	$B \Rightarrow (\neg B \Rightarrow A)$	(14) et déf. de $\Rightarrow$
(16)	$A \Rightarrow (\neg A \Rightarrow B)$	(15) et subst.

(1.19)

Ces axiomatisations permettent de retrouver comme théorèmes toutes les tautologies ou lois de la logique propositionnelle. De par tout ce qui a été dit jusqu'à maintenant, nous pouvons tenter de définir ce qu'est une preuve.

Définition: Une suite finie de formules $B_1, B_2, \dots, B_m$ est appelée "**preuve**" à partir des hypothèses $A_1, A_2, \dots, A_n$ si pour chaque i :

- B_i est l'une des hypothèses $A_1, A_2, \dots, A_n$
- ou B_i est une variante d'un axiome
- ou B_i est inférée (par application de la règle du modus ponens) à partir de la prémisses majeure B_j et de la prémisses mineure B_k où $j, k < i$
- ou B_i est inférée (par application de la règle de substitution) à partir d'une prémisses antérieure B_j , la variable remplacée n'apparaissant pas dans $A_1, A_2, \dots, A_n$

Une telle suite de formules, B_m étant la formule finale de la suite, est appelée plus explicitement "preuve de B_m " à partir des hypothèses $A_1, A_2, \dots, A_n$, ce que nous notons par:

$$A_1, A_2, \dots, A_n \vdash B_m \quad (1.20)$$

Remarque: Il faut noter que lorsque nous essayons de prouver un résultat à partir d'un certain nombre d'hypothèses, nous n'essayons pas de prouver les hypothèses elles-mêmes.

3.4. QUANTIFICATEURS

Nous devons compléter l'utilisation des connecteurs du calcul propositionnel par ce que nous appelons des "**quantificateurs**" si nous souhaitons pouvoir résoudre certains problèmes. Effectivement, le calcul propositionnel ne nous permet pas d'affirmer des choses générales sur les éléments d'un ensemble par exemple. Dans ce sens, la logique propositionnelle ne reflète qu'une partie du raisonnement. Le "**calcul des prédicats**" au contraire permet de manipuler formellement des affirmations telles que "il existe un x tel que [x a une voiture américaine]" ou "pour tous les x [si x est un teckel, alors x est petit]"; en somme, nous étendons les formules composées afin de pouvoir affirmer des quantifications existentielles ("il existe...") et des quantifications universelles ("pour tout...."). Les exemples que nous venons de donner font intervenir des propositions un peu particulières comme " x a une voiture américaine". Il s'agit ici de propositions comportant une variable. Ces propositions sont en fait l'application d'une fonction à x . Cette fonction, c'est celle qui associe " x a une voiture américaine" à x . Nous noterons cette fonction par "... a une voiture américaine" et nous dirons que c'est une fonction propositionnelle, car c'est une fonction dont la valeur est une proposition. Ou encore un "prédicat".

Les quantificateurs existentiels et universels vont donc de pair avec l'emploi de fonctions propositionnelles. Le calcul des prédicats est cependant limité dans les formules existentielles et universelles. Ainsi, nous nous interdisons des formules comme "il existe une affirmation de x telle que...". En fait, nous ne nous autorisons à quantifier que des "individus". C'est pour cela que la logique des prédicats est dite une "**logique du premier ordre**".

Avant de passer à l'étude du calcul des prédicats nous devons définir:

D1. Le "**quantificateur universel**": $\forall$ (pour tout)

D2. Le "**quantificateur existentiel**": $\exists$ (il existe)

Remarque: Nous utilisons parfois le symbole $\exists!$ pour dire brièvement: "il existe un et un seul":

$$\forall x \in \mathbb{R}, \exists! y \in \mathbb{R}_+, x = \ln(y) \quad (1.21)$$

Nous allons voir que la théorie de la démonstration et des ensembles est l'exacte transcription des principes et résultats de la Logique (celle avec un "L" majuscule).

4. CALCUL DES PRÉDICATS

Dans un cours de mathématiques (d'algèbre, d'analyse, de géométrie, ...), nous démontrons les propriétés de différents types d'objets (entiers, réels, matrices, suites, fonctions continues, courbes, ...). Pour pouvoir prouver ces propriétés, il faut bien sûr que les objets sur lesquels nous travaillons soient clairement définis (qu'est-ce qu'un entier, un réel, ...?).

En logique du premier ordre et, en particulier, en théorie de la démonstration, les objets que nous étudions sont les formules et leurs démonstrations. Il faut donc donner une définition précise de ce que

sont ces notions. Les termes et les formules forment la grammaire d'une langue, simplifiée à l'extrême et calculée exactement pour dire ce que nous voulons sans ambiguïté et sans détour inutile.

4.1. GRAMMAIRE

Définitions:

D1. Les "**termes**", désignent les objets dont nous voulons prouver des propriétés (nous reviendrons un peu plus loin beaucoup plus en détail sur ces derniers):

- En algèbre, les termes désignent les éléments d'un groupe (ou anneau, corps, espace vectoriel, etc.). Nous manipulons aussi des ensembles d'objets (sous-groupe, sous-espace vectoriel, etc). Les termes qui désignent ces objets, d'un autre type, seront appelés "termes du second ordre".

- En analyse, les termes désignent les réels ou (par exemple, si nous nous plaçons dans des espaces fonctionnels) des fonctions.

D2. Les "**formules**", représentent les propriétés des objets que nous étudions (nous reviendrons également beaucoup plus en détail sur ces dernières):

- En algèbre, nous pourrions écrire des formules pour exprimer que deux éléments commutent, qu'un sous-espace vectoriel est de dimension 3, etc.

- En analyse, nous écrirons des formules pour exprimer la continuité d'une fonction, la convergence d'une suite, etc.

- En théorie des ensembles, les formules pourront exprimer l'inclusion de deux ensembles, l'appartenance d'un élément à un ensemble,...

D3. Les "**démonstrations**", elles permettent d'établir qu'une formule est vraie. Le sens précis de ce mot aura lui aussi besoin d'être défini. Plus exactement, elles sont des déductions sous hypothèses, elles permettent de "mener du vrai au vrai", la question de la vérité de la conclusion étant alors renvoyée à celle des hypothèses, laquelle ne regarde pas la logique mais repose sur la connaissance que nous avons des choses dont nous parlons.

4.2. LANGAGES

En mathématique, nous utilisons, suivant le domaine, différents langages qui se distinguent par les symboles utilisés. La définition ci-dessous exprime simplement qu'il suffit de donner la liste de ces symboles pour préciser le langage.

Définition: Un "**langage**" est la donnée d'une famille (pas nécessairement finie) de symboles. Nous en distinguons de trois sortes: symboles, termes et formules.

Remarques:

R1. Nous utilisons quelques fois le mot "**vocabulaire**" ou le mot "**signature**" à la place du mot "langage".

R2. Le mot "prédicat" peut être utilisé à la place du mot "relation". Nous parlons alors de "calcul des prédicats" au lieu de "logique du premier ordre" (ce que nous avons étudié précédemment).

4.2.1. SYMBOLES

Il existe différents types de symboles que nous allons tâcher de définir:

D1. Les "**symboles de constante**" (voir remarque plus bas)

Exemple:

Le n pour l'élément neutre en théorie des ensembles (cf. chapitre de Théorie Des Ensembles)

D2. Les "**symboles de fonction**" ou "**foncteurs**". A chaque symbole de fonction est associé un entier strictement positif que nous appelons son "**arité**": c'est le nombre d'arguments de la fonction. Si l'arité est 1 (resp. 2, ..., n), nous disons que la fonction est unaire (resp. binaire, ..., n -aire)

Exemple:

Le foncteur binaire de multiplication $*$ dans les groupes (cf. chapitre de Théorie Des Ensembles).

D3. Les "**symboles de relation**". De la même manière, à chaque symbole de relation est associé un entier positif ou nul (son arité) qui correspond à son nombre d'arguments et nous parlons de relation unaire, binaire, n -aire (comme par exemple le symbole de relation "=").

D4. Les "**variables individuelles**". Dans toute la suite, nous nous donnerons un ensemble infini V de variables. Les variables seront notées comme il l'est de tradition: x, y, z (éventuellement indexées: $x_1, \dots$).

D5. A cela il faut rajouter les connecteurs et quantificateurs que nous avons longuement présentés plus haut et sur lesquels il est pour l'instant inutile de revenir.

Remarques:

R1. Un symbole de constante peut être vu comme un symbole de fonction à 0 argument (d'arité nulle).

R2. Nous considérons (sauf mention contraire) que chaque langage contient le symbole de relation binaire $=$ (lire "égal") et le symbole de relation à zéro argument dénoté $\perp$ (lire "bottom" ou "absurde") qui représente le faux. Dans la description d'un langage, nous omettrons donc souvent de les mentionner. Le symbole $\perp$ est souvent redondant. Nous pouvons en effet, sans l'utiliser, écrire une formule qui est toujours fausse. Il permet cependant de représenter le faux d'une manière canonique et donc d'écrire des règles de démonstration générales.

R3. Le rôle des fonctions et des relations est très différent. Comme nous le verrons plus loin, les symboles de fonction sont utilisés pour construire les termes (les objets du langage) et les symboles de relation pour construire les formules (les propriétés de ces objets).

4.2.2. TERMES

Les termes (nous disons aussi "termes du premier ordre") représentent les objets associés au langage.

Définitions:

Soit $\mathcal{L}$ un langage:

D1. L'ensemble $\mathcal{I}$ des termes sur $\mathcal{L}$ est le plus petit ensemble contenant les variables, les constantes et stable (on ne sort pas de l'ensemble) par l'application des symboles de fonction de $\mathcal{L}$ à des termes.

D2. Un "**terme clos**" est un terme qui ne contient pas de variables (donc par extension, seulement des constantes).

D3. Pour obtenir une définition plus formelle, nous pouvons écrire:

$$\mathcal{I}_0 = \{t\} \quad (1.22)$$

où t est une variable ou un symbole de constante et, pour tout $k \in \mathbb{N}$:

$$\mathcal{I}_{k+1} = \mathcal{I}_k \cup \{f(t_1, \dots, t_n) / t_i \in \mathcal{I}_k\} \quad (1.23)$$

où f est une fonction d'arité n (rappelons que l'arité est le nombre d'arguments de la fonction). Ainsi, pour chaque arité, il y a un degré d'ensemble de termes. Nous avons finalement:

$$\mathcal{I} = \bigcup_{k \in \mathbb{N}} \mathcal{I}_k \quad (1.24)$$

D4. Nous appellerons "**hauteur**" d'un terme t le plus petit k tel que $t \in \mathcal{I}_k$.

Remarques:

R1. La définition D4 signifie que les variables et les constantes sont des termes et que si f est un symbole de fonction n -aire et $t_1, \dots, t_n$ sont des termes alors $f(t_1, \dots, t_n)$ est un terme en soi aussi. L'ensemble $\mathcal{T}$ des termes est défini par la grammaire:

$$\boxed{\mathcal{T} = V \mid S_C \mid S_F(\mathcal{T}, \dots, \mathcal{T})} \quad (1.25)$$

Cette expression se lit de la manière suivante: un élément de l'ensemble $\mathcal{T}$ que nous sommes en train de définir est soit un élément de V (variables), soit un élément de S_C (l'ensemble des symboles de constantes), soit l'application d'un symbole de fonction $f \in S_F$ à n éléments (constantes ou variables) de $\mathcal{T}$.

Attention: le fait que f soit de la bonne arité est seulement implicite dans cette notation. De plus, l'écriture $S_F(\mathcal{T}, \dots, \mathcal{T})$ ne signifie pas que tous les arguments d'une fonction sont identiques mais simplement que ces arguments sont des éléments de $\mathcal{T}$.

R2. Il est souvent commode de voir un terme (expression) comme un arbre dont chaque noeud est étiqueté par un symbole de fonction (opérateur ou fonction) et chaque feuille par une variable ou une constante.

Dans la suite, nous allons sans cesse définir des notions (ou prouver des résultats) "par récurrence" sur la structure ou la taille d'un terme.

Définitions:

D1. Pour prouver une propriété P sur les termes, il suffit de prouver P pour les variables et les constantes et de prouver $P(f(t_1, \dots, t_n))$ à partir de $P(t_1), \dots, P(t_n)$. Nous faisons ainsi ici une "**preuve par induction sur la hauteur d'un terme**". C'est une technique que nous retrouverons dans les chapitres suivants.

D2. Pour définir une fonction Φ sur les termes, il suffit de la définir sur les variables et les constantes et de dire comment nous obtenons $\Phi(f(t_1, \dots, t_n))$ à partir de $\Phi(t_1), \dots, \Phi(t_n)$. Nous faisons ici encore une "**définition par induction sur la hauteur d'un terme**".

 Exemple:

La taille (nous disons aussi la "longueur") d'un terme t (notée $\tau(t)$) est le nombre de symboles de fonction apparaissant dans t . Formellement:

- $\tau(x) = \tau(c) = 0$ si x est une variable et c est une constante

- $\tau(f(t_1, \dots, t_n)) = 1 + \sum_{1 \leq i \leq n} \tau(t_i)$

où le 1 dans la dernière relation représente le terme f .

Remarque: La preuve par induction sur la hauteur d'un terme sera souvent insuffisante. Nous pourrions alors prouver une propriété P sur les termes en supposant la propriété vraie pour tous les termes de taille $p < n$ et en la démontrant ensuite pour les termes de taille n . Il s'agira alors d'une "preuve par récurrence sur la taille du terme" (voir de tels exemples dans le chapitre de Théorie Des Nombres).

4.2.3. FORMULES

Les formules sont construites à partir de "formules atomiques" en utilisant des connecteurs et des quantificateurs. Nous utiliserons les connecteurs et les quantificateurs suivants (qui nous sont déjà connus):

- connecteur unaire de négation: $\neg$
- connecteurs binaires de conjonction et disjonction ainsi que d'implication: $\wedge, \vee, \rightarrow$
- quantificateurs: $\exists$ qui se lit "il existe" et $\forall$ qui se lit "pour tout"

Cette notation des connecteurs est standard (elle devrait du moins). Elle est utilisée pour éviter les confusions entre les formules et le langage courant (le métalangage).

Définitions:

D1. Soit $\mathcal{L}$ un langage, les "formules atomiques" de $\mathcal{L}$ sont les formules de la forme $R(t_1, \dots, t_n)$ où R est un symbole de relation n -aire de $\mathcal{L}$ et $t_1, \dots, t_n$ sont des termes de $\mathcal{L}$. Nous notons "Atom" l'ensemble des formules atomiques. Si nous notons S_R l'ensemble des symboles de relation, nous pouvons écrire l'ensemble des termes mis en relations par l'expression:

$$Atom = S_R(\mathcal{T}, \dots, \mathcal{T}) \quad (1.26)$$

L'ensemble F des formules de la logique du premier ordre de $\mathcal{L}$ est donc défini par la grammaire (où x est une variable):

$$F = Atom \mid F \vee F \mid F \wedge F \mid F \rightarrow F \mid F \neg F \mid \exists x F \mid \forall x F \quad (1.27)$$

où il faut lire: l'ensemble des formules est le plus petit ensemble contenant les formules et tel que si F_1 et F_2 sont des formules alors $F_1 \vee F_2$, etc. sont des formules et qu'elles peuvent être en relation entre elles.

Exemple:

Les symboles de relation du langage propositionnel sont des relations d'arité 0 (même le symbole "=" est absent), les quantificateurs sont alors inutiles (puisque une formule propositionnelle ne peut pas contenir des variables). Nous obtenons alors le calcul propositionnel défini par:

$$\mathcal{C}_P = \perp \mid \neg \mathcal{C}_P \mid \mathcal{C}_P \vee \mathcal{C}_P \mid \mathcal{C}_P \wedge \mathcal{C}_P \mid \mathcal{C}_P \rightarrow \mathcal{C}_P \quad (1.28)$$

Remarquons la présence du symbole "bottom" signifiant le "faux" que nous n'avions pas mentionné lors

de notre étude de la logique propositionnelle.

Nous ferons attention à ne pas confondre termes et formules. $\sin(x)$ est un terme (fonction), $x = 3$ est une formule. Mais $\sin(x) \wedge (x = 3)$ n'est rien: nous ne pouvons, en effet, mettre un connecteur entre un terme et une formule (aucun sens).

Remarques:

R1. Pour définir une fonction Φ sur les formules, il suffit de définir Φ sur les formules atomiques.

R2. Pour prouver une propriété P sur les formules, il suffit de prouver P pour les formules atomiques.

R3. Pour prouver une propriété P sur les formules, il suffit de supposer la propriété vraie pour toutes les formules de taille $p < n$ et de la démontrer pour les formules de taille n .

D2. Une "**sous-formule**" d'une formule (ou expression) F est l'un de ses composants, in extenso une formule à partir de laquelle F est construite. Formellement, nous définissons l'ensemble $SF(F)$ des sous-formules de F par:

- Si F est atomique: $SF(F) = \{F\}$
- Si $F = F_1 \oplus F_2$ (soit une composition!) avec $\oplus \in \{\wedge, \vee, \rightarrow\}$, $SF(F) = \{F\} \cup SF(F_1) \cup SF(F_2)$
- Si $F = \neg F_1$ ou $Qx F_1$ avec $Q \in \{\forall, \exists\}$, $SF(F) = \{F\} \cup SF(F_1)$

D3. Une formule F de $\mathcal{L}$ n'utilise qu'un nombre fini de symboles de $\mathcal{L}$. Ce sous-ensemble est appelé le "**langage de la formule**" et noté $\mathcal{L}(F)$.

D4. La "**taille (ou la longueur) d'une formule**" F (notée $\tau(F)$) est le nombre de connecteurs ou de quantificateurs apparaissant dans F . Formellement:

- $\tau(F) = 0$ si F est une formule atomique
- $\tau(F_1 \oplus F_2) = 1 + \tau(F_1) + \tau(F_2)$ où $\oplus \in \{\vee, \wedge, \rightarrow\}$
- $\tau(\neg F_1) = \tau(Qx F_1) = 1 + \tau(F_1)$ avec $Q \in \{\forall, \exists\}$

D5. "**L'opérateur principal**" (nous disons aussi le "connecteur principal") d'une formule est défini par:

- Si A est atomique, alors elle n'a pas d'opérateur principal
- Si $A = \neg B$, alors $\neg$ est l'opérateur principal de A
- Si $A = B \oplus C$ où $\oplus \in \{\wedge, \vee, \rightarrow\}$, alors $\oplus$ est l'opérateur principal de A
- Si $A = QxB$ où $Q \in \{\forall, \exists\}$, alors Q est l'opérateur principal de A

D6. Soit F une formule. L'ensemble $VL(F)$ des variables libres de F et l'ensemble $VM(F)$ des variables muettes (ou liées) de F sont définis par récurrence sur $\tau(F)$.

Une occurrence d'une variable donnée est dite "**variable liée**" ou "**variable muette**" dans une formule F si dans cette même formule, un quantificateur y fait référence. Dans le cas contraire, nous disons avoir une "**variable libre**".

Remarque: Une occurrence d'une variable x dans une formule F est une position de cette variable dans la formule F . Ne pas confondre avec l'objet qu'est la variable elle-même.

Pour préciser les variables libres possibles d'une formule F , nous noterons $F[x_1, \dots, x_n]$. Cela signifie que les variables libres de F sont parmi $x_1, \dots, x_n$ in extenso si y est libre dans F , alors y est l'un des x_i mais les x_i n'apparaissent pas nécessairement dans F .

Nous pouvons définir les variables muettes ou libres de manière plus formelle:

1. Si $F = R(t_1, \dots, t_n)$ est atomique alors $VL(F)$ est l'ensemble des variables libres apparaissant dans les t_i et nous avons alors pour les variables muettes $VM(F) = \emptyset$
2. Si $F = F_1 \oplus F_2$ où $\oplus \in \{\wedge, \vee, \rightarrow\}$: $VL(F) = VL(F_1) \cup VL(F_2)$ alors $VM(F) = VM(F_1) \cup VM(F_2)$
3. si $F = \neg F_1$ alors $VL(F) = VL(F_1)$ et $VM(F) = VM(F_1)$
4. si $F = Qx F_1$ avec $Q \in \{\forall, \exists\}$: $VL(F) = VL(F_1) - \{x\}$ et $VM(F) = VM(F_1) \cup \{x\}$

Exemples:

E1. Soit $F: \forall x (x \cdot y = y \cdot x)$ alors $VL(F) = \{y\}$ et $VM(F) = \{x\}$

E2. Soit $G: \{\forall x \exists y (x \cdot z = z \cdot y)\} \wedge \{x = z \cdot z\}$ alors $VL(G) = \{x, z\}$ et $VM(G) = \{x, y\}$

D7. Nous disons que les formules F et G sont " **α -équivalentes**" si elles sont (syntaxiquement) identiques à un renommage près des occurrences liées des variables.

D8. Une "**formule close**" est une formule sans variables libres.

D9. Soit F une formule, x une variable et t un terme. $F[x := t]$ est la formule obtenue en remplaçant dans F toutes les occurrences libres de x par t , après renommage éventuel des occurrences liées de F qui apparaissent libres dans t .

Remarques:

R1. Nous noterons dans les exemples vus qu'une variable peut avoir à la fois des occurrences libres et des occurrences liées. Nous n'avons donc pas toujours $VL(F) \cap VM(F) = \emptyset$

R2. Nous ne pouvons pas renommer y en x dans la formule $\forall y(x \cdot y = y \cdot x)$ et obtenir la formule $\forall x(x \cdot x = x \cdot x)$: la variable x serait "capturée". Nous ne pouvons donc pas renommer des variables liées sans précautions: il faut éviter de capturer des occurrences libres.

5. DÉMONSTRATIONS

Les démonstrations que l'on trouve dans les ouvrages de mathématiques sont des assemblages de symboles mathématiques et de phrases contenant des mots-clés tels que: "donc", "parce que", "si", "si et seulement si", "il est nécessaire que", "il suffit de", "prenons un x tel que", "supposons que", "cherchons une contradiction", etc. Ces mots sont supposés être compris par tous de la même manière, ce qui n'est en fait, pas toujours le cas.

Dans tout ouvrage, le but d'une démonstration est de convaincre le lecteur de la vérité de l'énoncé. Suivant le niveau du lecteur, cette démonstration sera plus ou moins détaillée: quelque chose qui pourra être considéré comme évident dans un cours de licence pourrait ne pas l'être dans un cours de niveau inférieur.

Dans un devoir, le correcteur sait que le résultat demandé à l'étudiant est vrai et il en connaît la démonstration. L'étudiant doit démontrer (correctement) le résultat demandé. Le niveau de détail qu'il doit donner dépend donc de la confiance qu'aura le correcteur: dans une bonne copie, une "preuve par une récurrence évidente" passera bien, alors que dans une copie où il y eu auparavant un "évident", qui était évidemment... faux, ça ne passera pas!

Pour pouvoir gérer convenablement le niveau de détail, il faut savoir ce qu'est une démonstration complète. Ce travail de formalisation n'a été fait qu'au début de 20ème siècle!!

Plusieurs choses peuvent paraître surprenantes:

- il n'y a qu'un nombre fini de règles: deux pour chacun des connecteurs (et l'égalité) plus trois règles générales. Il n'était pas du tout évident a priori qu'un nombre fini de règles soit suffisant pour démontrer tout ce qui est vrai. Nous montrerons ce résultat (c'est essentiellement, le théorème de complétude). La preuve n'en est pas du tout triviale.

- ce sont les mêmes règles pour toutes les mathématiques et la physique: algèbre, analyse, géométrie, etc. Cela veut dire que nous avons réussi à isoler tout ce qui est général dans un raisonnement. Nous verrons plus loin qu'une démonstration est un assemblage de couples, où Γ est un ensemble de formules (les hypothèses) et A une formule (la conclusion). Quand nous faisons de l'arithmétique, de la géométrie ou de l'analyse réelle, nous utilisons, en plus des règles, des hypothèses que l'on appelle des "axiomes". Ceux-ci expriment les propriétés particulières des objets que nous manipulons (pour plus de détails sur les axiomes voir la page d'introduction du site).

Nous démontrons donc, en général, des formules en utilisant un ensemble d'hypothèses, et cet ensemble peut varier au cours de la démonstration: quand nous disons "supposons F et montrons G ", F est alors une nouvelle hypothèse que nous pourrions utiliser pour montrer G . Pour formaliser cela, nous introduisons le concept de "séquent":

Définitions:

D1. Un "séquent" est un couple (noté $\Gamma \vdash F$) où:

- Γ est un ensemble fini de formules qui représente les hypothèses que nous pouvons utiliser. Cet ensemble s'appelle aussi le "contexte du séquent".
- F est une formule. C'est la formule que nous voulons montrer. Nous dirons que cette formule est la "conclusion du séquent".

Remarques:

- R1. Si $\Gamma = \{A_1, \dots, A_n\}$ nous pourrions noter $A_1, \dots, A_n \vdash F$ au lieu de $\Gamma \vdash F$. Le signe $\vdash$ se lit "thèse" ou "démontre".
- R2. Nous noterons $\vdash F$ un séquent dont l'ensemble d'hypothèses est vide et $\Gamma_1, \dots, \Gamma_n \vdash F$ un séquent dont l'ensemble d'hypothèses est $\bigcup_i \Gamma_i$.
- R3. Nous noterons que dans le séquent $\Gamma, A \vdash B$ la formule A peut être dans Γ (elle devient alors une hypothèse).
- R4. Nous écrirons $\Gamma \not\vdash F$ pour dire que " $\Gamma \vdash F$ est non prouvable".

D2. Un séquent $\Gamma \vdash F$ est "prouvable" (ou démontrable, dérivable) s'il peut être obtenu par une application finie de règles. Une formule F est prouvable si le séquent $\vdash F$ est prouvable.

5.1. RÈGLES DE DÉMONSTRATION

Les règles de démonstration sont les briques qui permettent de construire les dérivations. Une dérivation formelle est un assemblage fini (et correct!) de règles. Cet assemblage n'est pas linéaire (ce n'est pas une suite) mais un "arbre". Nous sommes en effet souvent amenés à faire des branchements.

Nous allons présenter un choix de règles. Nous aurions pu en présenter d'autres (à la place ou en plus) qui donneraient la même notion de prouvabilité. Celles que l'on a choisies sont "naturelles" et correspondent aux raisonnements que l'on fait habituellement en mathématique. Dans la pratique courante nous utilisons, en plus des règles ci-dessous, beaucoup d'autres règles mais celles-ci peuvent se déduire des précédentes. Nous les appellerons "règles dérivées".

Il est de tradition d'écrire la racine de l'arbre (le séquent conclusion) en bas, les feuilles en haut: la nature est ainsi faite... Comme il est également de tradition d'écrire, sur une feuille de papier, de haut en bas, il ne serait pas déraisonnable d'écrire la racine en haut et les feuilles en bas. Il faut faire un choix !

Une règle se compose:

- d'un ensemble de "prémisses": chacune d'elles est un séquent. Il peut y en avoir zéro, un ou plusieurs
- du séquent conclusion de la règle
- d'une barre horizontale séparant les prémisses (en haut) de la conclusion (en bas). Sur la droite de la barre, nous indiquerons le nom de la règle.

Exemple:

$$\frac{\Gamma \vdash A \rightarrow B \quad \Gamma \vdash A}{\Gamma \vdash B} \rightarrow_e \quad (1.29)$$

Cette règle a deux prémisses ($\Gamma \vdash A \rightarrow B$ et $\Gamma \vdash A$) et une conclusion ($\Gamma \vdash B$) et se note de manière abrégée sous la forme: $\rightarrow_e$. Elle peut se lire de deux manières:

- de bas en haut: si nous voulons prouver la conclusion, il suffit par utilisation de la règle de prouver les prémisses. C'est ce qu'on fait quand nous cherchons une démonstration. Cela correspond à "l'analyse".
- de haut en bas: si nous avons prouvé les prémisses, alors nous avons aussi prouvé la conclusion. C'est ce que nous faisons quand nous rédigeons une démonstration. Cela correspond à la "synthèse".

Pour les démonstrations il existe un nombre fini de règles au nombre de 17 que nous allons définir ci-après:

1. Axiome:

$$\boxed{\frac{}{\Gamma, A \vdash A} ax} \quad (1.30)$$

De bas en haut: si la conclusion du séquent est une des hypothèses, alors le séquent est prouvable.

2. Affaiblissement:

$$\boxed{\frac{\Gamma \vdash A}{\Gamma, B \vdash A} aff} \quad (1.31)$$

Explications:

- De haut en bas: si nous démontrons A sous les hypothèses Γ , en ajoutant d'autres hypothèses on peut encore démontrer A .
- De bas en haut: il y a des hypothèses qui peuvent ne pas servir

3. Introduction de l'implication:

$$\boxed{\frac{\Gamma, A \vdash B}{\Gamma \vdash A \rightarrow B} \rightarrow_i} \quad (1.32)$$

- De bas en haut: pour montrer que $A \rightarrow B$ nous supposons A (c'est-à-dire que nous l'ajoutons aux hypothèses) et nous démontrons B .

4. Elimination de l'implication:

$$\frac{\Gamma \vdash A \rightarrow B \quad \Gamma \vdash A}{\Gamma \vdash B} \rightarrow_e \quad (1.33)$$

- De bas en haut: pour démontrer B , si nous connaissons un théorème de la forme $A \rightarrow B$ et si nous pouvons démontrer le lemme $A \rightarrow B$, il suffit de démontrer A .

5. Introduction à la conjonction:

$$\frac{\Gamma \vdash A \quad \Gamma \vdash B}{\Gamma \vdash A \wedge B} \wedge_i \quad (1.34)$$

- De bas en haut: pour montrer $A \wedge B$, il suffit de montrer A et de montrer B .

6. Elimination de la conjonction:

$$\frac{\Gamma \vdash A \wedge B}{\Gamma \vdash A} \wedge_e^g \quad (1.35) \text{ et } \frac{\Gamma \vdash A \wedge B}{\Gamma \vdash B} \wedge_e^d \quad (1.36)$$

- De haut en bas: de $A \wedge B$, nous pouvons déduire A (élimination gauche) et B (élimination droite).

7. Introduction de la disjonction:

$$\frac{\Gamma \vdash A}{\Gamma \vdash A \vee B} \vee_i^g \quad (1.37) \text{ ou } \frac{\Gamma \vdash B}{\Gamma \vdash A \vee B} \vee_i^d \quad (1.38)$$

- De bas en haut: pour démontrer $A \vee B$, il suffit de démontrer A (disjonction gauche) ou de démontrer B (disjonction droite).

8. Elimination de la disjonction:

$$\frac{\Gamma \vdash A \vee B \quad \Gamma, A \vdash C \quad \Gamma, B \vdash C}{\Gamma \vdash C} \vee_e \quad (1.39)$$

- De bas en haut: si nous voulons montrer C et que nous savons que nous avons $A \vee B$, il suffit de le montrer d'une part en supposant A , d'autre part en supposant B . C'est un raisonnement par cas.

9. Introduction de la négation:

$$\frac{\Gamma, A \vdash \perp}{\Gamma \vdash \neg A} \neg_i \quad (1.40)$$

- De bas en haut: pour montrer $\neg A$, nous supposons A et nous démontrons l'absurde ($\perp$).

10. Elimination de la négation:

$$\frac{\Gamma \vdash \neg A \quad \Gamma \vdash A}{\Gamma \vdash \perp} \neg_e \quad (1.41)$$

- De haut en bas: si nous avons montré $\neg A$ et A , alors nous avons montré l'absurde ($\perp$)

11. Absurdité classique:

$$\frac{\Gamma, \neg A \vdash \perp}{\Gamma \vdash A} \perp_e \quad (1.42)$$

- De bas en haut: pour démontrer A , il suffit de démontrer l'absurde en supposant $\neg A$.

Cette règle, est équivalente à dire: A est vraie si et seulement si il est faux que A soit fausse. Cette règle ne va pas de soi: elle est nécessaire pour prouver certains résultats (il y a des résultats que nous ne pouvons pas prouver si nous n'avons pas cette règle). Contrairement, à beaucoup d'autres, cette règle peut par ailleurs être appliquée à tout moment. Nous pouvons, en effet, toujours dire: pour prouver A je suppose $\neg A$ et je vais chercher une contradiction.

12. Introduction du quantificateur universel:

$$\frac{\Gamma \vdash A \quad x \text{ n'est pas libre dans les formules de } \Gamma}{\Gamma \vdash \forall x A} \forall_i \quad (1.43)$$

- De bas en haut: pour démontrer $\forall x A$, il suffit de montrer A en ne faisant aucune hypothèse sur x .

Remarque: pour des démonstrations cette vérification (aucune hypothèse sur x) est souvent source d'erreur.

13. Elimination du quantificateur universel:

$$\frac{\Gamma \vdash \forall x A}{\Gamma \vdash A[x := t]} \forall_e \quad (1.44)$$

- De haut en bas: de $\forall x A$, nous pouvons déduire $A[x := t]$ pour n'importe quel terme t . Ce que nous pouvons dire aussi sous la forme: si nous avons prouvé A pour tout x , alors nous pouvons utiliser A avec n'importe quel objet t (!!).

14. Introduction du quantificateur existentiel:

$$\frac{\Gamma \vdash A[x := t]}{\Gamma \vdash \exists x A} \exists_i \quad (1.45)$$

- De bas en haut: pour démontrer $\exists x A$, il suffit de trouver un objet (in extenso un terme t) pour lequel nous savons montrer $A[x := t]$.

15. Elimination du quantificateur existentiel:

$$\frac{\Gamma \vdash \exists x A \quad \Gamma, A \vdash C \quad x \text{ n'est libre ni dans les formules de } \Gamma, \text{ ni dans } C}{\Gamma \vdash C} \exists_e \quad (1.46)$$

- De bas en haut: nous démontrons qu'il existe bien un ensemble d'hypothèses tel que $\exists x A$ et partant de ce résultat comme nouvelle hypothèse, nous démontrons C . Cette formule C hérite alors de la formule $\exists x A$ et dès lors x n'est pas libre dans C car il ne l'était déjà pas dans $\exists x A$.

16. Introduction de l'égalité:

$$\frac{}{\Gamma \vdash t = t} =_i \quad (1.47)$$

De bas en haut: nous pouvons toujours montrer $t=t$. Cette règle signifie que l'égalité est réflexive (cf. chapitre Opérateurs).

17. Elimination de l'égalité:

$$\frac{\Gamma \vdash A[x := t] \quad \Gamma \vdash t = u}{\Gamma \vdash A[x := u]} =_e \quad (1.48)$$

- De haut en bas: si nous avons démontré $\Gamma \vdash A[x := t]$ et $t=u$, alors nous avons démontré $A[x := u]$. Cette règle exprime que les objets égaux ont les mêmes propriétés. Nous noterons cependant que les formules (ou relations) $t=u$ et $u=t$ ne sont pas, formellement, identiques. Il nous faudra démontrer que l'égalité est symétrique (nous en profiterons aussi pour démontrer que l'égalité est transitive).

 Exemples:

E1. Cet exemple montre que l'égalité est symétrique (un petit peu non trivial mais bon pour commencer):

$$\frac{\frac{\frac{}{x_1 = x_2 \vdash x_1 = x_1} =_i \quad \frac{}{x_1 = x_2 \vdash x_1 = x_2} ax}{x_1 = x_2 \vdash x_2 = x_1} \rightarrow_i \quad \frac{}{\vdash x_1 = x_2 \rightarrow x_2 = x_1} \forall_i \times 2}{\vdash \forall x_1, x_2 (x_1 = x_2 \rightarrow x_2 = x_1)} =_e \quad (1.49)$$

- De haut en bas: nous introduisons l'égalité $=_i$ et prouvons à partir de l'hypothèse $x_1 = x_2$ la formule $x_1 = x_1$. En même temps, nous définissons l'axiome comme quoi $x_1 = x_2$. Ensuite à partir de ces prémisses, nous éliminons l'égalité $=_e$ en substituant les termes de façon à ce qu'à partir de la supposition $x_1 = x_2$ (venant de l'axiome) nous obtenions $x_2 = x_1$. Ensuite, l'élimination de l'égalité implique automatiquement sans aucune hypothèse que $x_1 = x_2 \rightarrow x_2 = x_1$. Dès lors, il nous suffit d'introduire le quantificateur universel pour chacune des variables (donc deux fois) sans aucune hypothèse afin d'obtenir que l'égalité est symétrique.

E2. Cet exemple montre que l'égalité est transitive (c'est-à-dire si $x_1 = x_2$ et $x_2 = x_3$ alors $x_1 = x_3$). En notant F la formule $x_1 = x_2 \wedge x_2 = x_3$:

$$\frac{\frac{\frac{}{F \vdash x_1 = x_2 \wedge x_2 = x_3} ax}{F \vdash x_1 = x_2} \wedge_e \quad \frac{\frac{}{F \vdash x_1 = x_2 \wedge x_2 = x_3} ax}{F \vdash x_2 = x_3} \wedge_e}{F \vdash x_1 = x_3} =_e \quad (1.50)$$

$$\frac{}{\vdash (x_1 = x_2 \wedge x_2 = x_3) \rightarrow x_1 = x_3} \rightarrow_i \quad \frac{}{\vdash \forall x_1, x_2, x_3 ((x_1 = x_2 \wedge x_2 = x_3) \rightarrow x_1 = x_3)} \forall_i \times 3$$

Que faisons, nous ici ? Nous introduisons d'abord la formule F deux fois en tant qu'axiome afin de la décortiquer plus tard à gauche et à droite (nous n'introduisons pas l'égalité supposée déjà introduite en tant que règle). Une fois ceci fait, nous éliminons à gauche et à droite la conjonction sur la formule pour travailler sur les termes gauches et droites seuls et introduisons l'égalité sur les deux termes ce qui fait qu'à partir de la formule nous avons l'égalité transitive. Il s'ensuit que sans aucune hypothèse cela implique automatiquement que l'égalité est transitive et finalement nous disons que ceci est valable pour toute valeur des différentes variables (si la formule est vraie, alors l'égalité est transitive).

E3. L'objectif sera de démontrer que toute involution est une bijection (cf. chapitre de Théorie Des Ensembles). Soit f un symbole de fonction unaire (à une variable), nous notons (pour plus de détails voir le chapitre de Théorie Des Ensembles):

- $Inv[f]$ la formule:

$$\forall x, y \quad f(x) = f(y) \rightarrow x = y \quad (1.51)$$

qui signifie que f est injective.

- $Surj[f]$ la formule:

$$\forall y \exists x \quad f(x) = y \quad (1.52)$$

qui signifie que f est surjective

- $Bij[f]$ la formule:

$$Inv[f] \wedge Surj[f] \quad (1.53)$$

qui signifie que f est bijective.

- $Inv[f]$ la formule:

$$\forall x \quad f(f(x)) = x \quad (1.54)$$

qui signifie que f est une involution (nous notons également cela $f \circ f = Id$ c'est-à-dire que la composition de f est l'identité).

Nous aimerions savoir si:

$$Inv[f] \vdash Bij[f] \quad (1.55)$$

Nous allons présenter (en essayant que ce soit au plus clair) cette démonstration de quatre manières différentes: classique (informelle), classique (pseudo-formelle), formelle en arbre et formelle en ligne.

Méthode classique:

Nous devons montrer que si f est involutive alors elle est donc bijective. Nous avons donc deux choses à montrer (et les deux doivent être satisfaites en même temps): que la fonction est injective et surjective.

1. Montrons que l'involution est injective. Nous supposons pour cela, puisque f est involutive elle est

donc injective, tel que:

$$\forall x, y \ f(x) = f(y) \quad (1.56)$$

implique:

$$x = y \quad (1.57)$$

Or, cette supposition découle automatiquement de la définition de l'involution que:

$$\forall x \ f(f(x)) = x, \forall y \ f(f(y)) = y \quad (1.58)$$

et de l'application de f à la relation:

$$f(x) = f(y) \quad (1.59)$$

(soit trois égalités) tel que:

$$f(f(x)) = f(f(y)) \quad (1.60)$$

nous avons donc:

$$x = y \quad (1.61)$$

2. Montrons que l'involution est surjective: si elle est surjective, alors nous devons avoir:

$$\forall y \exists x \ f(x) = y \quad (1.62)$$

Or, définissons la variable x par définition de l'involution elle-même:

$$x := f(y) \quad (1.63)$$

(puisque $y = f(x) \dots$) un changement de variables après nous obtenons:

$$f(f(y)) = y \quad (1.64)$$

et donc la surjectivité est assurée.

Méthode pseudo-formelle:

Nous reprenons la même chose et nous y injectons les règles de la théorie de la démonstration:

Nous devons montrer que f involutive est donc bijective. Nous avons donc deux choses à montrer ($\wedge_i$) (et les deux doivent être satisfaites en même temps): que la fonction est injective et surjective:

$$Inj[f] \wedge Surj[f] \quad (1.65)$$

1. Montrons d'abord que l'involution est injective. Nous supposons pour cela, puisque f est involutive et donc injective, que:

$$(\forall_i) \ \forall x, y \ f(x) = f(y) \quad (1.66)$$

implique:

$$(\rightarrow_i) \quad x = y \quad (1.67)$$

Or, cette supposition découle automatiquement de la définition de l'involution que:

$$(\forall_e, ax) \quad \forall x \, f(f(x)) = x, \forall y \, f(f(y)) = y \quad (1.68)$$

et de l'application de f à la relation:

$$f(x) = f(y) \quad (1.69)$$

(soit trois égalités $=_e \times 2$) tel que:

$$f(f(x)) = f(f(y)) \quad (1.70)$$

nous avons donc:

$$x = y \quad (1.71)$$

2. Montrons que l'involution est surjective. Si elle est surjective, alors nous devons avoir:

$$\forall y \exists x \, f(x) = y \quad (\forall_i) \quad (1.72)$$

Or, définissons la variable x par définition de l'involution elle-même:

$$(\exists_i) \quad x := f(y) \quad (1.73)$$

(puisque $y = f(x) \dots$) un changement de variables après nous obtenons:

$$f(f(y)) = y \quad (1.74)$$

et donc:

$$(\forall_e, ax) \quad (1.75)$$

la surjectivité est assurée.

Méthode formelle en arbre:

Faisons cela avec la méthode graphique que nous avons déjà présentée plus haut.

1. Montrons que l'involution est injective:

Pour cela, montrons d'abord que:

$$f(x) = f(y) \vdash f(f(x)) = f(f(y)) \quad (1.76)$$

Donc:

$$\frac{\frac{\overline{Inv[f] \vdash \forall y (f(f(y)) = y)}^{ax}}{Inv[f] \vdash (f(y) = x)[y := f(x)]} \forall_e \quad \frac{\overline{f(x) = f(y) \vdash f(x) = f(y)}^{ax}}{f(x) = f(y) \vdash (f(x) = y)[x := f(y)]} \forall_e}{f(x) = f(y) \vdash (f(x) = y)[x := f(y)]} =_e \quad (1)$$

$$\frac{\frac{\overline{Inv[f] \vdash \forall x (f(f(x)) = x)}^{ax}}{Inv[f] \vdash (f(x) = y)[x := f(y)]} \forall_e \quad \frac{\overline{f(x) = f(y) \vdash f(x) = f(y)}^{ax}}{f(x) = f(y) \vdash (f(y) = x)[y := f(x)]} \forall_e}{f(x) = f(y) \vdash (f(y) = x)[y := f(x)]} =_e \quad (2) \quad (1.77)$$

$$\frac{(1) \quad (2)}{f(x) = f(y) \vdash f(f(x)) = f(f(y))} =_e \quad (3)$$

Remarque: Cette dernière relation est abrégée $=_e$ et appelée (comme d'autres existantes) "règle dérivée" car c'est un raisonnement qui est très souvent fait lors de démonstrations et un peu long à développer à chaque fois...

Dès lors:

$$\frac{\frac{\overline{Inv[f] \rightarrow Inv[f]}^{ax}}{Inv[f] \vdash f(f(x)) = x} \forall_e \quad \frac{\overline{Inv[f] \rightarrow Inv[f]}^{ax}}{Inv[f] \vdash f(f(y)) = y} \forall_e}{f(f(x)) = f(f(y)) \vdash x = y} =_e \quad (3) =_e \quad (4) \quad (1.78)$$

$$\frac{Inv[f], f(x) = f(y) \vdash x = y}{Inv[f] \vdash f(x) = f(y) \rightarrow x = y} \rightarrow_i \quad \frac{Inv[f] \vdash f(x) = f(y) \rightarrow x = y}{Inv[f] \vdash Inj[f]} \forall_i$$

2. Montrons que l'involution est surjective:

$$\frac{\frac{\overline{Inv[f] \vdash \forall x (f(f(x)) = x)}^{ax}}{Inv[f] \vdash (f(x) = y)[x := f(y)]} \forall_e}{Inv[f] \vdash \exists x (f(x) = y)} \exists_i \quad (1.79)$$

$$\frac{Inv[f] \vdash \exists x (f(x) = y)}{Inv[f] \vdash Surj[f]} \forall_i \quad (5)$$

Il s'ensuit:

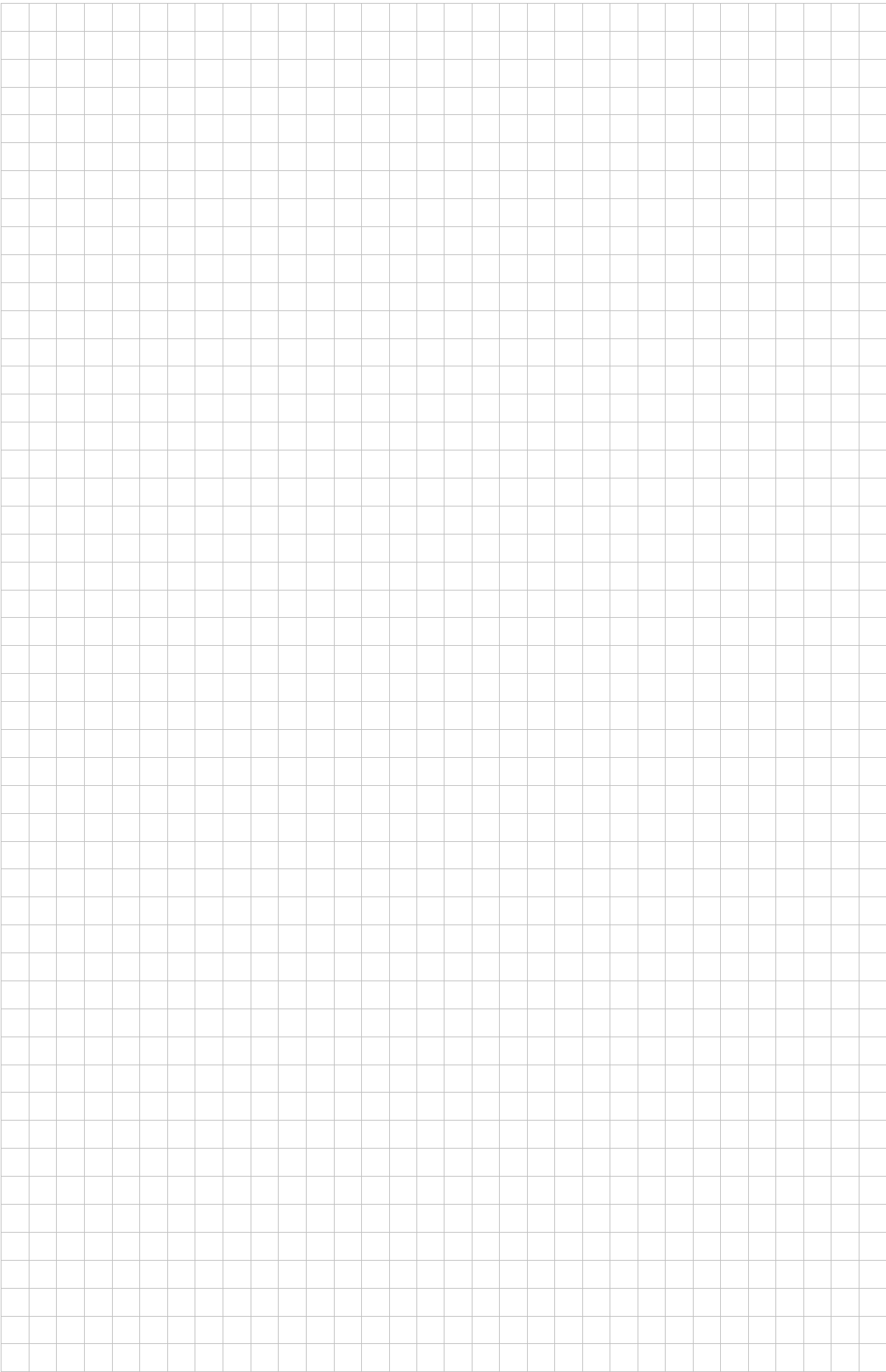
$$\frac{(5) \quad (6)}{Inv[f] \vdash Inj[f] \wedge Surj[f]} \wedge_i \quad (1.80)$$

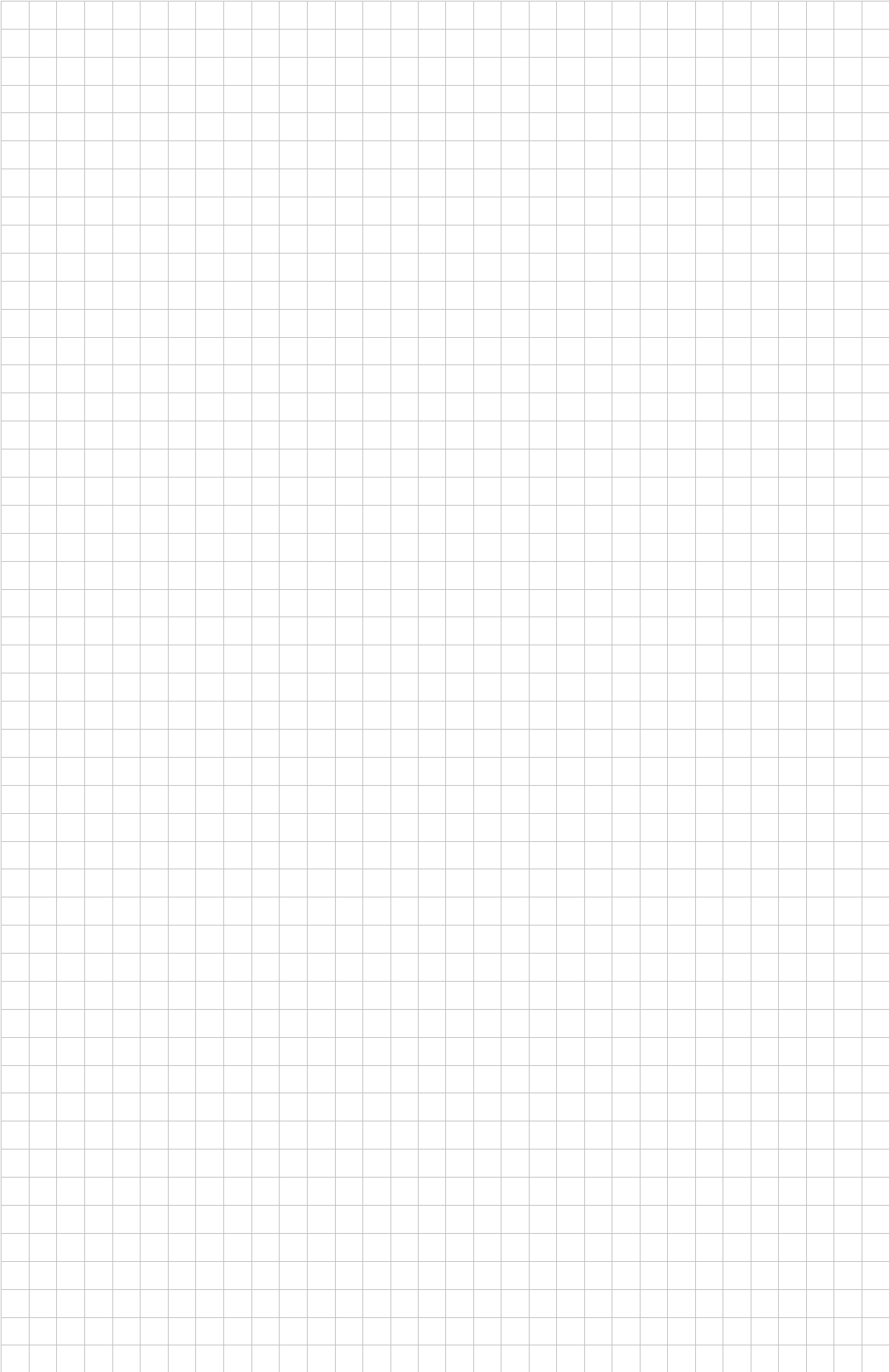
Méthode formelle en ligne:

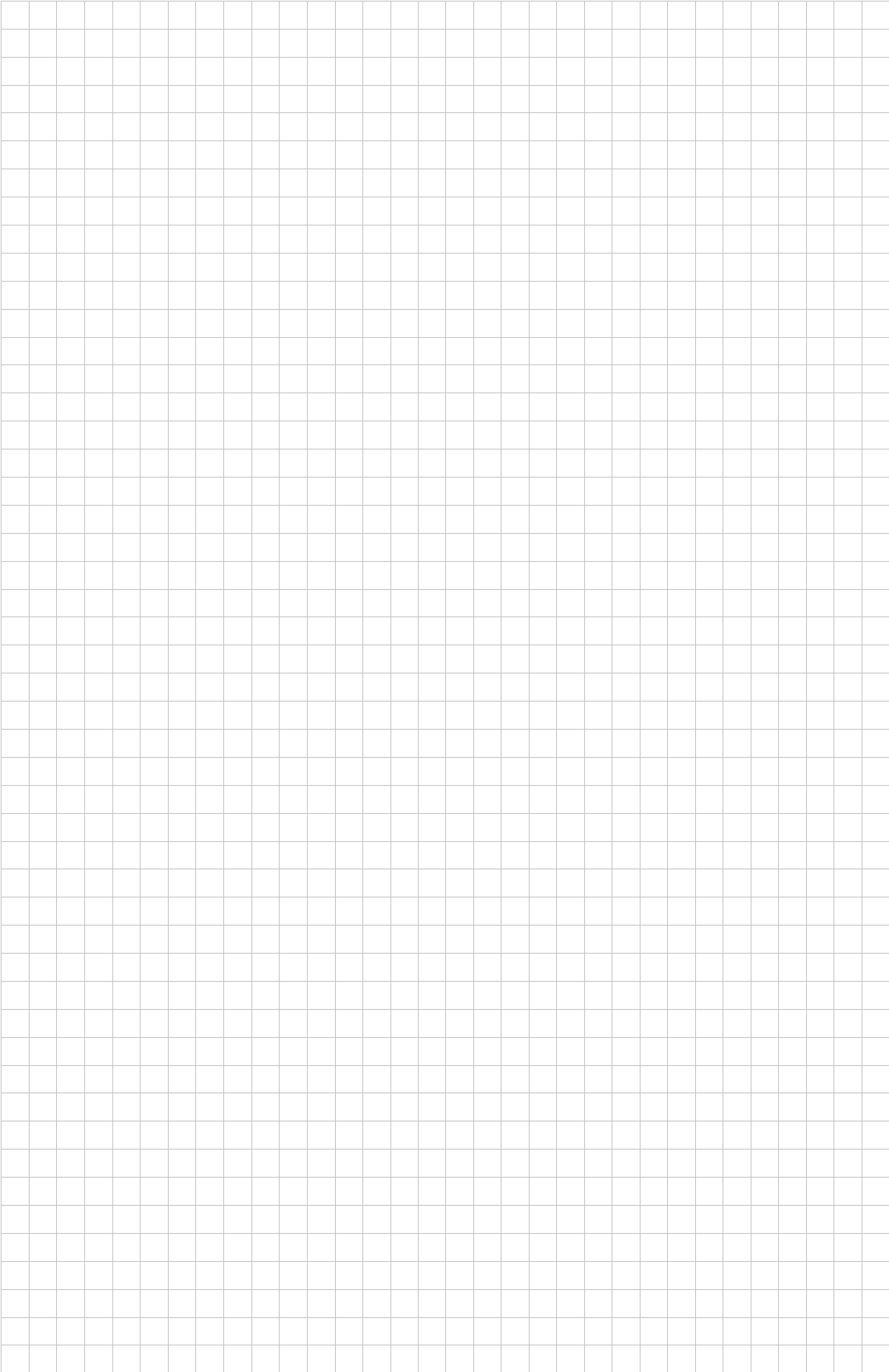
Nous pouvons faire la même chose sous une forme un peu moins... large... et plus tabulée... (cela n'en est pas moins indigeste):

$Inv[f] \vdash Bij[f]$	$\forall_i$
$1 := Inv[f] \vdash Inj[f]$	
$2 := Inv[f] \vdash Surj[f]$	
(1) $Inv[f] \vdash Inj[f]$	$\forall_i$
$Inv[f] \vdash f(x) = f(y) \rightarrow x = y$	$\rightarrow_i$
$Inv[f], f(x) = f(y) \vdash x = y$	$=_e \times 2(i)(ii)(iii)$
(i) $Inv[f] \vdash f(f(x)) = x$	$\forall_e$
$Inv[f] \vdash Inv[f]$	ax
(ii) $Inv[f] \vdash f(f(y)) = y$	$\forall_e$
$Inv[f] \vdash Inv[f]$	ax
(iii) $f(x) = f(y) \vdash f(f(x)) = f(f(y))$	$=_c$
$(1') f(x) = f(y) \vdash f(x) = f(y)$	ax
(2) $Inv[f] \vdash Surj[f]$	$\forall_i$
$Inv[f] \vdash \exists x(f(x) = y)$	$\exists_i$
$Inv[f] \vdash (f(x) = y)[x := f(y)]$	$\forall_e$
$Inv[f] \vdash \forall x(f(f(x)) = x)$	ax

(1.81)







2. NOMBRES

La base des mathématiques, mis à part le raisonnement (cf. chapitre Théorie De La Démonstration), est sans nul doute pour le commun des personnes l'arithmétique. Il est donc obligatoire que nous y fassions étape pour étudier sa provenance, quelques-unes de ses propriétés et conséquences.

Les nombres, comme les figures géométriques, constituent les bases de l'arithmétique. Ce sont aussi les bases historiques car les mathématiques ont certainement commencé par l'étude de ces objets, mais aussi les bases pédagogiques, car c'est en apprenant à compter que nous entrons dans le monde des mathématiques.

L'histoire des nombres, appelés également parfois "**scalaires**", est beaucoup trop longue pour être relatée ici, mais nous ne pouvons que vous conseiller un des meilleurs ouvrages francophones sur le sujet: Histoire Universelle des chiffres (~2'000 pages), Georges Ifrah, ISBN: 2221057791.

Cependant voici une petite bride de cette dernière qui nous semble fondamentale:

Notre système décimal actuel, de base 10, utilise les chiffres de 0 à 9, dits "**chiffres arabes**", mais au fait d'origine indienne (hindous). Effectivement, les chiffres arabes (d'origine indienne...) dans le tableau ci-dessous sont la première ligne et nous voyons qu'ils sont nettement différents des "**chiffres indiens**" de la deuxième ligne:

.	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

Tableau: 2.1 - Chiffres arabes

Il faut lire dans ce tableau: 0 "zéro", 1 "un", 2 "deux", 3 "trois", 4 "quatre", 5 "cinq", 6 "six", 7 "sept", 8 "huit", 9 "neuf". Ce système est beaucoup plus efficace que les chiffres romains (essayez de faire un calcul avec le système de notation romain vous allez voir...).

Ces chiffres ne furent introduits en Europe que vers l'an 1000. Utilisés en Inde, ils furent transmis par des Arabes au monde occidental par le pape Gerbert d'Aurillac lors de son séjour en Andalousie à la fin du 9ème siècle.

Remarque: Le mot français "**chiffre**" est une déformation du mot arabe "sifr" désignant "zéro". En italien, "zéro" se dit "zero", et serait une contraction de "zefiro", on voit là encore la racine arabe mais le zéro serait aussi d'origine indienne... Ainsi nos termes "chiffre" et "zéro" ont la même origine.

L'usage précoce d'un symbole numérique désignant "rien", au sens de "aucune quantité" ou "absence de quantité", c'est à dire notre zéro, provient du fait que les indiens utilisèrent un système dit "**système positionnel**". Dans un tel système, la position d'un chiffre dans l'écriture d'un nombre exprime la puissance de 10 et le nombre de fois qu'elle intervient... et l'absence d'une position dans ce système posait d'énormes problèmes de relecture et pouvait engendrer de grosses erreurs de calculs.

L'introduction révolutionnaire et pourtant simple du concept de rien permettait une relecture sans erreur des nombres.

L'absence d'une puissance est notée par un petit rond...: c'est le zéro. Notre système actuel est donc le "système décimal et positionnel".

Exemple:

Description du système décimal et positionnel:

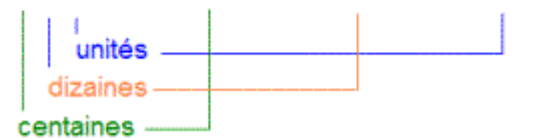
$$324 = 3 \times 10^2 + 2 \times 10^1 + 4 \times 10^0$$


Figure: 2.1 - Description système décimal et positionnel

Le nombre 324 s'écrit de gauche à droite comme étant trois centaines: 3 fois 100, deux dizaines: 2 fois 10 et quatre unités: 4 fois 1.

Remarques:

R1. Attention!! Nous différencions un chiffre d'un nombre... Le nombre est composé de chiffres et non inversement. Par ailleurs, nous différencions la partie entière de la partie décimale du nombre.

R2. Un "nombre décimal" est un nombre qui a une écriture finie en base 10.

Nous voyons parfois (et c'est conseillé) un séparateur de milliers représenté par une apostrophe ' en Suisse (posé tous les trois chiffres à partir du premier en partant de la droite pour les nombres entiers). Ainsi, nous écrirons 1'034 au lieu de 1034 ou encore 1'344'567'569 au lieu de 1344567569. Les séparateurs de milliers permettent de rapidement quantifier l'ordre de grandeur des nombres lus.

Ainsi:

- Si nous voyons uniquement une apostrophe nous saurons que le nombre est de l'ordre du millier
- Si nous voyons deux apostrophes nous saurons que le nombre est de l'ordre du million
- Si nous voyons trois apostrophes nous saurons que le nombre est de l'ordre du milliard

et ainsi de suite...

Au fait, tout nombre entier, autre que l'unité, peut être pris pour base d'un système de numérotation. Nous avons ainsi les systèmes de numérotation binaire, ternaire, quaternaire,..., décimal, duodécimal qui correspondent respectivement aux bases deux, trois, quatre,..., dix, douze.

Une généralisation de ce qui a été vu précédemment, peut s'écrire sous la forme suivante:

Tout nombre entier positif peut être représenté dans une base b sous forme de somme, où les coefficients a_i sont multipliés chacun par leur poids respectif b^i . Tel que:

$$N = a_{n-1}b^{n-1} + a_{n-2}b^{n-2} + \dots + a_1b^1 + a_0b^0 \quad (2.1)$$

Plus élégamment écrit:

$$N = \sum_{i=0}^{n-1} a_i b^i \quad (2.2)$$

avec $a_i \in [0; b - 1]$ et $N \in [0; b^n - 1]$.

Remarques:

R1. Comme très fréquemment en mathématique, nous remplacerons l'écriture des chiffres ou des nombres par des lettres latines ou grecques afin de généraliser leur représentation. Ainsi, lorsque nous parlons d'une base b la valeur b peut prendre n'importe quelle valeur entière 1, 2, 3, ...

R2. Lorsque nous prenons la valeur 2 pour b , N aura pour valeur maximale $2^n - 1$. Les nombres qui s'écrivent sous cette forme s'appellent les "**nombres de Mersenne**". Ces nombres ne peuvent être premiers (voir plus bas ce qu'est un nombre premier) que si n est premier.

Effectivement, si nous prenons (par exemples) $b = 10$ et $n = 3$ la plus grande valeur que nous pourrions avoir sera alors:

$$(b-1) \cdot 10^2 + (b-1) \cdot 10^1 + (b-1) \cdot 10^0 = 900 + 90 + 9 = 999 = 10^3 - 1 = b^3 - 1 \quad (2.3)$$

R3. Lorsque qu'un nombre est le même lu de gauche à droite ou de droite à gauche, nous parlons de "**nombre palindrome**".

1. BASES NUMÉRIQUES

Pour écrire un nombre dans un système de base b , nous devons commencer par adopter b caractères destinés à représenter les b premiers nombres $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Ces caractères sont comme nous les avons déjà définis, les "chiffres" que nous énonçons comme à l'ordinaire.

Pour la numérotation écrite, nous faisons cette convention, qu'un chiffre, placé à gauche d'un autre représente des unités de l'ordre immédiatement supérieur, ou b fois plus grandes. Pour tenir la place des unités qui peuvent manquer dans certains ordres, nous nous servons du zéro (0) et par suite, le nombre de chiffres employés peut varier.

Définition: Pour la numérotation parlée, nous convenons d'appeler "**unité simple**", "**dizaine**", "**centaine**", "**millier**", etc., les unités du premier ordre, du second, du troisième, du quatrième, etc. Ainsi les nombres 10, 11, ..., 19 se liront de même dans tous les systèmes de numérotation. Les nombres 1a, 1b, a0, b0, ... se liront dix-a, dix-bé, a-dix, bé-dix, etc. Ainsi, le nombre 5b6a71c se lira:

cinq millions bé-cent soixante-a mille sept cent dix-cé

Cet exemple est pertinent car il nous montre l'expression générale de la langue parlée que nous utilisons quotidiennement et intuitivement en base dix (faute à notre éducation).

Remarques:

R1. Les règles des opérations définies pour les nombres écrits dans le système décimal sont les mêmes pour les nombres écrits dans un système quelconque de numérotation.

R2. Pour opérer rapidement dans un système quelconque de numérotation, il est indispensable de savoir par coeur toutes les sommes et tous les produits de deux nombres d'un seul chiffre.

R3. Le choix de la base décimale semblerait dû au fait que l'humain a dix doigts.

Voyons comment nous convertissons un système de numérotation dans un ordre:

Exemple:

En base dix nous savons que 142'713 s'écrit:

$$142'713_{10} = 1 \cdot 10^5 + 4 \cdot 10^4 + 2 \cdot 10^3 + 7 \cdot 10^2 + 1 \cdot 10^1 + 3 \cdot 10^0 \quad (2.4)$$

En base deux (base binaire) le nombre 0110 s'écrit en base 10:

$$0110_2 = 0 \cdot 2^3 + 1 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0 = 6_{10} \quad (2.5)$$

et ainsi de suite...

L'inverse (pour l'exemple de la base deux) est toujours un peu plus délicat. Par exemple la conversion du nombre décimal 1'492 en base deux se fait par divisions successives par 2 des restes et donne (le principe est à peu près identique pour toutes les autres bases):

quotient	reste
1492	
746	0
373	0
186	1
93	0
46	1
23	0
11	1
5	1
2	1
1	0
0	1

(1 0 1 1 1 0 1 0 1 0 0)₂ = (1492)₁₀

Figure: 2.2 - Conversion décimal en binaire

Ainsi, pour convertir le nombre 142'713 (base décimale) en base duodécimale (base douze) nous avons (notation: q est le "quotient", et r le "reste"):

$$\frac{142'713}{12} \Rightarrow q = 11'892, r = 9 \quad (2.6)$$

$$\frac{11'892}{12} \Rightarrow q = 991, r = 0 \quad (2.7)$$

$$\frac{991}{12} \Rightarrow q = 82, r = 7 \quad (2.8)$$

$$\frac{82}{12} \Rightarrow q = 6, r = 10 \quad (2.9)$$

$$\frac{6}{12} \Rightarrow q = 0, r = 6 \quad (2.10)$$

Ainsi nous avons les restes 6, 10, 7, 0, 9 ce qui nous amène à écrire:

$$142'713_{10} = 6a709_{12} \quad (2.11)$$

Nous avons choisi pour ce cas particulier la symbolique que nous avons définie précédemment (a-dix) pour éviter toute confusion.

2. TYPES DE NOMBRES

Il existe en mathématiques une très grande variété de nombres (naturels, rationnels, réels, irrationnels, complexes, p-adiques, quaternions, transcendants, algébriques, constructibles...) puisque le mathématicien peut à loisirs en créer en ayant uniquement à poser les axiomes (règles) de manipulations de ceux-ci (cf. chapitre de Théorie Des Ensembles).

Cependant, il y en a quelques-uns que nous retrouvons plus souvent que d'autres et certains qui servent de base de construction à d'autres et qu'il conviendrait de définir suffisamment rigoureusement (sans aller dans les extrêmes) pour pouvoir savoir de quoi nous parlerons lorsque nous les utiliserons.

2.1. NOMBRES ENTIERS NATURELS

L'idée du "**nombre entier**" (nombre pour lequel il n'y a pas de chiffres après la virgule) est le concept fondamental de la mathématique et nous vient à la vue d'un groupement d'objets de même espèce (un mouton, un autre mouton, encore un autre, etc.). Lorsque la quantité d'objets d'un groupe est différente de celle d'un autre groupe nous parlons alors de groupe numériquement supérieur ou inférieur quel que soit l'espèce d'objets contenus dans ces groupes. Lorsque la quantité d'objets d'un ou de plusieurs groupes est équivalente, nous parlons alors "**d'égalité**". A chaque objet correspond le nombre "**un**" ou "**unité**" noté "1".

Pour former des groupements d'objets, nous pouvons opérer ainsi: à un objet, ajouter un autre objet, puis encore un et ainsi de suite... chacun des groupements, au point de vue de sa collectivité, est caractérisé par un nombre. Il résulte de là qu'un nombre peut être considéré comme représentant un groupement d'unités tel que chacune de ces unités corresponde à un objet de la collection.

Définition: Deux nombres sont dits "**égaux**" si à chacune des unités de l'un nous pouvons faire correspondre une unité de l'autre et inversement. Si ceci ne se vérifie pas alors nous parlons "**d'inégalité**".

Prenons un objet, puis un autre, puis au groupement formé, ajoutons encore un objet et ainsi de suite.

Les groupements ainsi constitués sont caractérisés par des nombres qui, considérés dans le même ordre que les groupements successivement obtenus, constituent la "**suite naturelle**" $\mathbb{N}$ et notée:

$$\mathbb{N} = \{0; 1; 2; 3; 4; \dots\} \quad (2.12)$$

Remarque: La présence du 0 (zéro) dans notre définition de $\mathbb{N}$ est discutable étant donné qu'il n'est ni positif ni négatif. C'est la raison pour laquelle dans certains ouvrages vous pourrez trouver une définition de $\mathbb{N}$ sans le 0.

Les constituants de cet ensemble peuvent être définis par (nous devons cette définition au mathématicien Gottlob) les propriétés (avoir lu au préalable le chapitre de Théorie Des Ensembles est recommandé...) suivantes:

P1. 0 (lire "zéro") est le nombre d'éléments (défini comme une relation d'équivalence) de tous les ensembles équivalents à (en bijection avec) l'ensemble vide.

P2. 1 (lire "un") est le nombre d'éléments de tous les ensembles équivalents à l'ensemble dont le seul élément est 1.

P3. 2 (lire "deux") est le nombre d'éléments de tous les ensembles équivalents à l'ensemble dont tous les éléments sont 0 et 1.

P4. En général, un nombre entier est le nombre d'éléments de tous les ensembles équivalents à l'ensemble des nombres entiers le précédant!

La construction de l'ensemble des entiers naturels s'est faite de la manière la plus naturelle et cohérente qui soit. Les naturels doivent leur nom à ce qu'ils avaient pour objet, aux prémices de leur existence, de dénombrer des quantités et des choses de la nature ou qui intervenaient dans la vie de l'homme.

L'originalité de l'ensemble réside dans la manière empirique dont il s'est construit car il ne résulte pas réellement d'une définition mathématique, mais davantage d'une prise de conscience par l'homme du concept de quantité dénombrable, de nombre et de lois qui traduisent des relations entre eux.

La question de l'origine de $\mathbb{N}$ est dès lors la question de l'origine des mathématiques. Et de tout temps des débats confrontant les pensées des plus grands esprits philosophiques ont tenté d'élucider ce profond mystère, à savoir si les mathématiques sont une pure création de l'esprit humain ou si au contraire l'homme n'a fait que redécouvrir une science qui existait déjà dans la nature. Outre les nombreuses questions philosophiques que cet ensemble peut susciter, il n'en est pas moins intéressant d'un point de vue exclusivement mathématique. Du fait de sa structure, il présente des propriétés remarquables qui peuvent se révéler d'une grande utilité lorsque l'on pratique certains raisonnements ou calculs.

Remarquons immédiatement que la suite naturelle des nombres entiers est illimitée (cf. chapitre de Théorie Des Nombres) mais dénombrable (nous verrons cela plus bas), car, à un groupement d'objets qui se trouve représenté par un certain nombre n , il suffira d'ajouter un objet pour obtenir un autre groupement qui sera défini par un nombre entier immédiatement supérieur $n + 1$.

Définition: Deux nombres entiers qui diffèrent d'une unité positive sont dits "**consécutifs**".

2.1.1. AXIOMES DE PEANO

Lors de la crise des fondements des mathématiques, les mathématiciens ont bien évidemment cherché à axiomatiser l'ensemble $\mathbb{N}$ et nous devons l'axiomatisation actuelle à Peano et à Dedekind.

Les axiomes de ce système comportent les symboles $<$ et $=$ pour représenter les relations "plus petit" et "égal" (cf. chapitre sur les Opérateurs). Ils comprennent d'autre part les symboles 0 pour le nombre zéro et s pour représenter le nombre "successeur". Dans ce système, 1 est noté:

$$s(0) \quad (2.13)$$

dit "successeur de zéro", 2 est noté:

$$s(s(0)) \quad (2.14)$$

Les axiomes de Peano qui construisent $\mathbb{N}$ sont les suivants (voir le chapitre de la Théorie de la Démonstration pour certains symboles):

A1. 0 est un entier naturel (permet de poser que $\mathbb{N}$ n'est pas vide).

A2. Tout entier naturel a un successeur, noté $s(n)$.

Donc s est une application injective, c'est-à-dire:

$$\forall x, y \quad s(x) = s(y) \Rightarrow x = y \quad (2.15)$$

si deux successeurs sont égaux, ils sont les successeurs d'un même nombre.

A3. $\forall x \neg (s(x) = 0)$, le successeur d'un entier naturel n'est jamais égal à zéro (ainsi $\mathbb{N}$ a un premier élément)

A4. $\varphi(0) \wedge \forall x \quad (\varphi(x) \Rightarrow \varphi(s(x))) \Rightarrow \forall x \quad \varphi(x)$, "**axiome de récurrence**" qui se doit se lire de la manière suivante: si l'on démontre qu'une propriété est vraie pour un x et son successeur, alors cette propriété est vraie pour tout x .

Donc l'ensemble de tous les nombres vérifiant les 4 axiomes est:

$$\mathbb{N} = \{0, 1, 2, 3, \dots, n, \dots\} \quad (2.16)$$

Remarque: Les axiomes de Peano permettent de construire très rigoureusement les deux opérations de base de l'arithmétique que sont l'addition et la multiplication (cf. chapitre sur les Opérateurs) et ainsi tous les autres ensembles que nous verrons par la suite.

2.2. NOMBRES PAIRS, IMPAIRS ET PARFAITS

En arithmétique, étudier la parité d'un entier, c'est déterminer si cet entier est ou non un multiple de deux. Un entier multiple de deux est un entier pair, les autres sont les entiers impairs.

Définitions:

D1. Les nombres obtenus en comptant par deux à partir de zéro, (soit 0, 2, 4, 6, 8, ...) dans cette suite naturelle sont appelés "**nombres pairs**".

Le $n^{\text{ième}}$ nombre pair est donné par la relation:

$$(n + n) = 2n \quad (2.17)$$

D2. Les nombres que nous obtenons en comptant par deux à partir de un (soit 1, 3, 5, 7, ...) dans cette suite naturelle s'appellent "**nombres impairs**".

Le $(n + 1)^{\text{ième}}$ nombre impair est donné par:

$$(2n + 1) \quad (2.18)$$

Remarque: Nous appelons "**nombres parfaits**", les nombres égaux à la somme de leurs diviseurs entiers strictement plus petits qu'eux mêmes (concept que nous verrons en détail plus tard) comme par exemple: $6=1+2+3$ et $28=1+2+4+7+14$.

2.2.1. NOMBRES PREMIERS

Définition: Un "**nombre premier**" est un entier possédant exactement 2 diviseurs (ces deux diviseurs sont donc "1" et le nombre lui-même). Dans le cas où il y a plus de 2 diviseurs on parle de "**nombre composé**".

Voici l'ensemble des nombres premiers inférieurs à 60:

$$\{2,3,5,7,11,13,17,19,23,29,31,37,41,43,47,53,59\}$$

Remarque: A noter que la définition de nombre premier exclut le chiffre "1" de l'ensemble des nombres premiers car il a un unique diviseur (lui-même) et pas deux comme le veut la définition.

Nous pouvons nous demander s'il existe une infinité de nombres premiers ? La réponse est positive et en voici une démonstration (parmi tant d'autres) par l'absurde.

Démonstration:

Supposons qu'il n'existe qu'un nombre fini de nombres premiers qui seraient:

$$p_1, p_2, \dots, p_n \quad (2.19)$$

Nous formons un nouveau nombre à partir du produit de tous les nombres premiers auquel nous ajoutons "1":

$$N = p_1 p_2 \dots p_n + 1 \quad (2.20)$$

Selon notre hypothèse initiale et le théorème fondamental de l'arithmétique (cf. chapitre de Théorie Des Nombres) ce nouveau nombre devrait être divisible par l'un des nombres premiers existants selon:

$$N = q \cdot p_i \quad (2.21)$$

Nous pouvons effectuer la division:

$$q = \frac{p_1 p_2 \dots p_n + 1}{p_i} = \frac{p_1 p_2 \dots p_n}{p_i} + \frac{1}{p_i} \quad (2.22)$$

Le premier terme se simplifie, car p_i est dans le produit. Nous notons E cet entier:

$$q = E + 1/p_i \quad (2.23)$$

Or, q et E sont deux entiers, donc $1/p_i$ doit être un entier. Mais p_i est par définition supérieur à 1. Donc $1/p_i$ n'est pas un entier.

Il y a alors contradiction et nous en concluons que les nombres premiers ne sont pas en nombre fini, mais infini.

□C.Q.F.D.

Remarques:

R1. $p_n = p_1 p_2 \dots p_n$ (le produit des n premiers nombres premiers inférieurs ou égaux à n) est appelé "**primorielle n** ".

R2. Nous renvoyons le lecteur au chapitre de Cryptographie de la section d'Informatique Théorique pour étudier quelques propriétés remarquables des nombres premiers dont la non moins fameuse fonction phi d'Euler (ou appelé aussi "fonction indicatrice").

R3. L'étude des nombres premiers est un sujet immensément vaste et certains théorèmes y relatifs sortent largement du cadre d'étude de ce site.

2.3. NOMBRES ENTIERS RELATIFS

L'ensemble $\mathbb{N}$ à quelques défauts que nous n'avons pas énoncés tout à l'heure. Par exemple, la soustraction de deux nombres dans $\mathbb{N}$ n'a pas toujours un résultat dans $\mathbb{N}$ (les nombres négatifs n'y existent pas). Autre défaut, la division de deux nombres dans $\mathbb{N}$ n'a également pas toujours un résultat dans $\mathbb{N}$ (les nombres fractionnaires n'y existent pas).

Nous pouvons dans un premier temps résoudre le problème de la soustraction en ajoutant à l'ensemble des entiers naturels, les entiers négatifs (concept révolutionnaire pour ceux qui en sont à l'origine) nous obtenons "**l'ensemble des entiers relatifs**" noté $\mathbb{Z}$ (pour "Zahl" de l'allemand):

$$\mathbb{Z} = \{ \dots; -3, -2, -1, 0, 1, 2, 3, \dots \} \quad (2.24)$$

L'ensemble des entiers naturels est donc inclus dans l'ensemble des entiers relatifs. C'est ce que nous notons sous la forme:

$$\mathbb{N} \subset \mathbb{Z} \quad (2.25)$$

et nous avons par définition (c'est une notation qu'il faut apprendre):

$$\mathbb{Z}_+ = \mathbb{N} \quad (2.26)$$

Cet ensemble a été créé à l'origine pour faire de l'ensemble des entiers naturels un objet que nous appelons un "groupe" (cf. chapitre Théorie Des Ensembles) par rapport à l'addition.

Définition: Nous disons qu'un ensemble E est un "**ensemble dénombrable**", s'il est équipotent à $\mathbb{N}$. C'est-à-dire s'il existe une bijection de (cf. chapitre Théorie Des Ensembles) $\mathbb{N}$ sur E . Ainsi, grosso modo, deux ensembles équipotents ont "autant" d'éléments au sens de leurs cardinaux (cf. chapitre de Théorie Des Ensembles), ou tout au moins la même infinité.

L'objectif de cette remarque est de faire comprendre que les ensembles $\mathbb{N}, \mathbb{Z}$ sont dénombrables.

Démonstration:

Montrons que $\mathbb{Z}$ est dénombrable en posant:

$$x_{2k} = k \text{ et } x_{2k+1} = -k - 1 \quad (2.27)$$

pour tout entier $k \geq 0$. Ceci donne l'énumération suivante:

$$0, -1, 1, -2, 2, -3, 3, \dots \quad (2.28)$$

de tous les entiers relatifs à partir des entiers naturels seuls.

□C.Q.F.D.

2.4. NOMBRES RATIONNELS

L'ensemble $\mathbb{Z}$ a aussi un défaut. Ainsi, la division de deux nombres dans $\mathbb{Z}$ n'a également pas toujours un résultat dans $\mathbb{Z}$ (les nombres fractionnaires n'y existent pas). Nous disons alors dans le langage de la théorie des ensembles que: la division n'est pas une opération interne dans $\mathbb{Z}$.

Nous pouvons ainsi définir un nouvel ensemble qui contient tous les nombres qui peuvent s'écrire sous forme de "**fraction**", c'est-à-dire du rapport d'un dividende (numérateur) et d'un diviseur (dénominateur). Quand un nombre peut se mettre sous cette forme, nous disons que c'est une "**nombre fractionnaire**":

$$\text{fraction} \left\{ \begin{array}{l} \nearrow \text{ numérateur} \\ \frac{2}{3} \leftarrow \text{ barre de fraction} \\ \searrow \text{ dénominateur} \end{array} \right.$$

Figure: 2.3 - Construction nombre fractionnaire

Une fraction peut être employée pour exprimer une partie, ou une part, de quelque chose (d'un objet, d'une distance, d'un terrain, d'une somme d'argent...).

Par définition, "**l'ensemble des nombres rationnels**" est donné par:

$$\mathbb{Q} = \left\{ \frac{p}{q} \mid (p, q) \in \mathbb{Z}, q \neq 0 \right\} \quad (2.29)$$

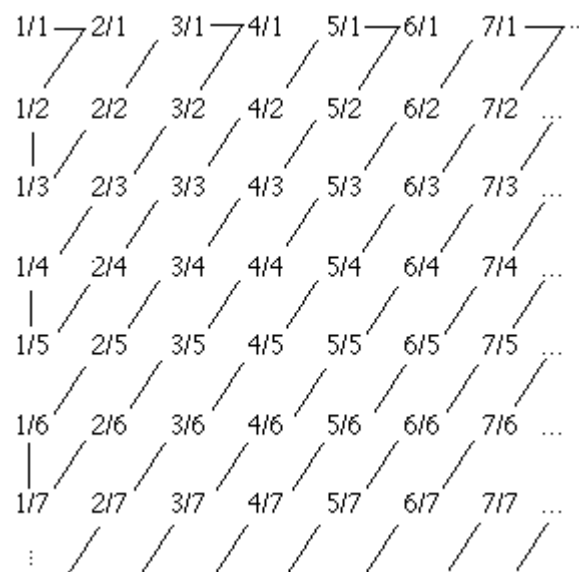
et où p et q sont des entiers sans facteurs communs (autrement dit la fraction p/q est écrite sous forme irréductible).

Nous supposons par ailleurs comme évident que:

$$\mathbb{N}, \mathbb{Z} \subset \mathbb{Q} \quad (2.30)$$

La logique de la création de l'ensemble des nombres rationnels est similaire à celle des entiers relatifs. Effectivement, les mathématiciens ont souhaité faire de l'ensemble des nombres relatifs un "groupe" par rapport à la loi de multiplication et de division (cf. chapitre de Théorie Des Ensembles).

De plus, contrairement à l'intuition, l'ensemble des nombres entiers et nombres rationnels sont équipotents. Nous pouvons nous persuader de cette équipotence en rangeant comme le fit Cantor, les rationnels dans un premier temps de la façon suivante:



L'application f est donc bijective: $\mathbb{N}$ et $\mathbb{Q}$ sont donc bien équipotents !

La définition un peu plus rigoureuse (et donc moins sympathique) de $\mathbb{Q}$ se fait à partir de $\mathbb{Z}$ en procédant comme suit (il est intéressant d'observer les notations utilisées) :

Sur l'ensemble $\mathbb{Z} \times \mathbb{Z} \setminus \{0\}$, qu'il faut lire comme étant l'ensemble construit à partir de deux éléments entiers relatifs dont on exclut le zéro pour le deuxième, on considère la relation R entre deux couples d'entiers relatifs définie par:

$$(a,b)R(a',b') \Leftrightarrow a \cdot b' = a'b \quad (2.31)$$

Nous vérifions facilement ensuite que R est une relation d'équivalence (cf. chapitre sur les Opérateurs) sur $\mathbb{Z} \times \mathbb{Z} \setminus \{0\}$.

L'ensemble des classes d'équivalences pour cette relation R noté alors $(\mathbb{Z} \times \mathbb{Z} \setminus \{0\}) / R$ est par définition $\mathbb{Q}$. C'est-à-dire que nous posons alors plus rigoureusement:

$$\mathbb{Q} = (\mathbb{Z} \times \mathbb{Z} \setminus \{0\}) / R \quad (2.32)$$

La classe d'équivalence de $(a,b) \in \mathbb{Z} \times \mathbb{Z} \setminus \{0\}$ est explicitement notée par:

$$\frac{a}{b} \quad (2.33)$$

conformément à la notation que tout le monde a l'habitude d'employer.

Nous vérifions facilement que l'addition et la multiplication qui étaient des opérations définies sur $\mathbb{Z}$ passent sans problèmes à $\mathbb{Q}$ en posant:

$$\frac{a}{b} \cdot \frac{p}{q} = \frac{a \cdot p}{b \cdot q} \quad \text{et} \quad \frac{a}{b} + \frac{p}{q} = \frac{a \cdot q + b \cdot p}{b \cdot q} \quad (2.34)$$

De plus ces opérations munissent $\mathbb{Q}$ d'une structure de corps (cf. chapitre de Théorie Des Ensembles) avec $\frac{0}{1}$ comme élément neutre additif et $\frac{1}{1}$ comme élément neutre multiplicatif. Ainsi, tout élément non nul de $\mathbb{Q}$ est inversible, en effet:

$$\frac{a}{b} \cdot \frac{b}{a} = \frac{ab}{ab} = \frac{1}{1} \quad (2.35)$$

ce qui s'écrit aussi plus techniquement:

$$(ab, ab)R(1,1) \quad (2.36)$$

Remarque: Même si nous aurions envie de définir $\mathbb{Q}$ comme étant l'ensemble $\mathbb{Z} \times \mathbb{Z} \setminus \{0\}$ où $\mathbb{Z}$ représente les numérateurs et $\mathbb{Z} \setminus \{0\}$ les dénominateurs des rationnels, ceci n'est pas possible car autrement nous aurions par exemple $(1,2) \neq (2,4)$ tandis que nous nous attendons à une égalité.

D'où le besoin d'introduire une relation d'équivalence qui nous permet d'identifier, pour revenir à l'exemple précédent, $(1,2)$ et $(2,4)$. La relation R que nous avons définie ne tombe pas du ciel, en effet le lecteur qui a manipulé les rationnels jusqu'à présent sans jamais avoir vu leur définition formelle sait que:

$$\frac{a}{b} = \frac{a'}{b'} \Leftrightarrow a \cdot b' = a' \cdot b \quad (2.37)$$

Il est donc naturel de définir la relation R comme nous l'avons fait. En particulier, en ce qui concerne l'exemple ci-dessus, $\frac{1}{2} = \frac{2}{4}$ car $(1,2)R(2,4)$ et le problème est résolu.

Outre les circonstances historiques de sa mise en place, ce nouvel ensemble se distingue des ensembles d'entiers relatifs car il induit la notion originale et paradoxale de quantité partielle. Cette notion qui a priori n'a pas de sens, trouvera sa place dans l'esprit de l'homme notamment grâce à la géométrie où l'idée de fraction de longueur, de proportion s'illustre plus intuitivement.

2.5. NOMBRES IRRATIONNELS

L'ensemble des rationnels $\mathbb{Q}$ est limité et non suffisant lui aussi. Effectivement, nous pourrions penser que tout calcul mathématique numérique avec les opérations communément connues se réduisent à cet ensemble mais ce n'est pas le cas.

Exemples:

E1. Prenons le calcul de la racine carrée de deux que nous noterons $\sqrt{2}$. Supposons que cette dernière racine soit un rationnel. Alors s'il s'agit bien d'un rationnel, nous devrions pouvoir l'exprimer comme a/b , où par de par la définition d'un rationnel a et b sont des entiers sans facteurs communs. Pour cette raison, a et b ne peuvent tous les deux être pairs. Il y a trois possibilités restantes:

1. a est impair (b est alors pair)
2. a est pair (b est alors impair)
3. a est impair (b est alors impair)

En mettant au carré, nous avons:

$$\sqrt{2} = \frac{a}{b} \quad (2.38)$$

qui peut s'écrire:

$$2b^2 = a^2 \quad (2.39)$$

Puisque le carré d'un nombre impair est impair et le carré d'un nombre pair est pair, le cas (1) est impossible, car a^2 serait impair et $2b^2$ serait pair.

Le cas (2) est aussi impossible, car alors nous pourrions écrire $a = 2c$, où c est un entier quelconque, et donc si nous le portons au carré nous avons $a^2 = 4c^2$ où nous avons un nombre pair des deux côtés de l'égalité. En remplaçant dans $2b^2 = a^2$ nous obtenons après simplification que $b^2 = 2c^2$. b^2 serait impair alors que $2c^2$ serait pair.

Le cas (3) est aussi impossible, car a^2 est donc alors impair et $2b^2$ est pair (que b soit pair ou impair!).

Il n'y a pas de solution! C'est donc que l'hypothèse de départ est fausse et qu'il n'existe pas deux entiers a et b tels que $\sqrt[3]{2} = a/b$.

E2. Démontrons, aussi par l'absurde, que le fameux nombre d'Euler e est irrationnel. Pour cela, rappelons que e (cf. chapitre d'Analyse Fonctionnelle) peut aussi être défini par la série de Taylor (cf. chapitre sur les Suites Et Séries):

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \dots + \frac{1}{n!} + \dots \quad (2.40)$$

Alors si e est rationnel, il doit pouvoir s'écrire sous la forme p/q (avec $q > 1$, car nous savons que e n'est pas entier). Multiplions les deux côtés de l'égalité par $q!$:

$$q!e = q! + \frac{q!}{1!} + \frac{q!}{2!} + \frac{q!}{3!} + \dots + \frac{q!}{q!} + \frac{q!}{(q+1)!} + \frac{q!}{(q+2)!} + \dots \quad (2.41)$$

Le premier membre $q!e$ serait alors un entier, car par définition de la factorielle:

$$q! = q(q-1)(q-2)\dots 2 \cdot 1 \quad (2.42)$$

est un entier.

Les premiers termes du second membre de la relation antécédente, jusqu'au terme $q!/q! = 1$ sont aussi des entiers car $q!/m!$ se simplifie si $q > m$. Donc par soustraction nous trouvons :

$$q!e - \left(q! + \frac{q!}{1!} + \frac{q!}{2!} + \frac{q!}{3!} + \dots + \frac{q!}{q!} \right) = \frac{q!}{(q+1)!} + \frac{q!}{(q+2)!} + \dots \quad (2.43)$$

où la série à droite devrait aussi être un entier!

Après simplification, le second membre de l'égalité devient:

$$\frac{1}{(q+1)} + \frac{1}{(q+1)(q+2)} + \dots \quad (2.44)$$

le premier terme de cette somme est strictement inférieur à $1/2$, le deuxième inférieur à $1/4$, le troisième inférieur à $1/8$, etc.

Donc, vu que chaque terme est strictement inférieur aux termes de la série harmonique suivante qui converge vers 1:

$$1/2 + 1/4 + 1/8 + \dots = 1 \quad (2.45)$$

alors par conséquent, la série n'est pas un entier puisque étant strictement inférieure à 1. Ce qui constitue une contradiction!

Ainsi, les nombres rationnels ne satisfont pas à l'expression numérique de $\sqrt[3]{2}$ comme de e (pour citer seulement ces deux exemples particuliers).

Il faut donc les compléter par l'ensemble de tous les nombres qui ne peuvent s'écrire sous forme de fraction (rapport d'un dividende et d'un diviseur entiers sans facteurs communs) et que nous appelons des "**nombres irrationnels**".

2.6. NOMBRES RÉELS

Définition: La réunion des nombres rationnels et irrationnels donne "**l'ensemble des nombres réels**".

Ce que nous notons:

$$\mathbb{Q} \subset \mathbb{R} \quad (2.46)$$

Remarque: Les mathématiciens dans leur rigueur habituelle ont différentes techniques pour définir les nombres réels. Ils utilisent pour cela des propriétés de la topologie (entre autres) et en particulier les suites de Cauchy mais c'est une autre histoire qui dépasse le cadre formel du présent chapitre.

Nous sommes évidemment amenés à nous poser la question si $\mathbb{R}$ est dénombrable ou non. La démonstration est assez simple.

Démonstration:

Par définition, nous avons vu plus haut qu'il doit y avoir une bijection entre $\mathbb{Q}$ et $\mathbb{R}$ pour dire que $\mathbb{R}$ soit dénombrable.

Pour simplifier, nous allons montrer que l'intervalle $[0,1[$ n'est alors pas dénombrable. Ceci impliquera bien sûr par extension que $\mathbb{R}$ ne l'est pas!

Les éléments de cet intervalle sont représentés par des suites infinies entre 0 et 9 (dans le système décimal):

- Certaines de ces suites sont nulles à partir d'un certain rang, d'autres non
- Nous pouvons donc identifier $[0,1[$ à l'ensemble de toutes les suites (finies ou infinies) d'entiers compris entre 0 et 9

n°1	x_{11}	x_{12}	x_{13}	x_{14}	...	x_{1p}	...
n°2	x_{21}	x_{22}	x_{23}	x_{24}	...	x_{2p}	...
n°3	x_{31}	x_{32}	x_{33}	x_{34}	...	x_{3p}	...
n°4	x_{41}	x_{42}	x_{43}	x_{44}	...	x_{4p}	...
n°5	x_{51}	x_{52}	x_{53}	x_{54}	...	x_{5p}	...
n°6	x_{61}	x_{62}	x_{63}	x_{64}	...	x_{6p}	...
	...						
			...				
n°k	x_{k1}	x_{k2}	x_{k3}	x_{k4}	...	x_{kp}	...
						...	

Tableau: 2.2 - Identification et classement de nombres réels

Si cet ensemble était dénombrable, nous pourrions classer ces suites (avec une première, une deuxième, etc.). Ainsi, la suite $x_{11}x_{12}x_{13}x_{14}...x_{1p}...$ serait classée première et ainsi de suite... comme le propose le tableau ci-dessus.

Nous pourrions alors modifier cette matrice infinie de la manière suivante: à chaque élément de la diagonale, rajouter 1, selon la règle: $0+1=1$, $1+1=2$, $8+1=9$ et $9+1=0$

n°1	x_{11} +1	x_{12}	x_{13}	x_{14}	...	x_{1p}	...
n°2	x_{21}	x_{22} +1	x_{23}	x_{24}	...	x_{2p}	...
n°3	x_{31}	x_{32}	x_{33} +1	x_{34}	...	x_{3p}	...
n°4	x_{41}	x_{42}	x_{43}	x_{44} +1	...	x_{4p}	...
n°5	x_{51}	x_{52}	x_{53}	x_{54}	...	x_{5p}	...
n°6	x_{61}	x_{62}	x_{63}	x_{64}	...	x_{6p}	...
	...						
			...				
n°k	x_{k1}	x_{k2}	x_{k3}	x_{k4}	...	x_{kp}	...
						...	

Tableau: 2.3 - Identification et classement de nombres réels

Alors considérons la suite infinie qui se trouve sur la diagonale:

- Elle ne peut être égale à la première car elle s'en distingue au moins par le premier élément
- Elle ne peut être égale à la deuxième car elle s'en distingue au moins par le deuxième élément
- Elle ne peut être égale à la troisième car elle s'en distingue au moins par le troisième élément

et ainsi de suite... Elle ne peut donc être égale à aucune des suites contenues dans ce tableau!

Donc, quel que soit le classement choisi des suites infinies de 0...9, il y en a toujours une qui échappe à ce classement! C'est donc qu'il est impossible de les numéroter... tout simplement parce qu'elles ne forment pas un ensemble dénombrable.

□C.Q.F.D.

La technique qui nous a permis d'arriver à ce résultat est connue sous le nom de "**procédé diagonal de Cantor**" (car similaire à celle utilisée pour l'équipotence entre ensemble naturel et rationnel) et l'ensemble des nombres réels est dit avoir "**la puissance du continu**" de par le fait qu'il est indénombrable.

Remarque: Nous supposons intuitif pour le lecteur que tout nombre réel peut être approché infiniment près par un nombre rationnel (pour les nombres irrationnels il suffit de s'arrêter à un nombre de décimales données et d'en trouver le rationnel correspondant). Les mathématiciens disent alors que $\mathbb{Q}$ est "**dense**" dans $\mathbb{R}$ et notent cela:

$$\bar{\mathbb{Q}} = \mathbb{R} \quad (2.47)$$

2.6.1. NOMBRES TRANSFINIS

Nous nous retrouvons donc avec un "infini" des nombres réels qui est différent de celui des nombres naturels. Cantor osa alors ce que personne n'avait osé depuis Aristote: la suite des entiers positifs est infinie, l'ensemble $\mathbb{N}$, est donc un ensemble qui a une infinité dénombrable d'éléments, alors il affirma que le cardinal (cf. chapitre de Théorie Des Ensembles) de cet ensemble était un nombre qui existait comme tel sans que l'on utilise le symbole fourre tout " ∞ ", il le nota:

$$\aleph_0 = \text{Card}(\mathbb{N}) \quad (2.48)$$

Ce symbole est la première lettre de l'alphabet hébreu, qui se prononce "aleph zéro". Cantor allait appeler ce nombre étrange, un nombre "**transfini**".

L'acte décisif est d'affirmer qu'il y a, après le fini, un transfini, c'est-à-dire une échelle illimitée de modes déterminés qui par nature sont infinis, et qui cependant peuvent être précisés, tout comme le fini, par des nombres déterminés, bien définis et distinguables les uns des autres !!

Après ce premier coup d'audace allant à l'encontre de la plupart des idées reçues depuis plus de deux mille ans, Cantor allait poursuivre sur sa lancée et établir des règles de calcul, paradoxales à première vue, sur les nombres transfinis. Ces règles se basaient, comme nous l'avons précisé tout à l'heure, sur le

fait que deux ensembles infinis sont équivalents s'il existe une bijection entre les deux ensembles.

Ainsi, nous pouvons facilement montrer que l'infini des nombres pairs est équivalent à l'infini des nombres entiers: pour cela, il suffit de montrer qu'à chaque nombre entier, nous pouvons associer un nombre pair, son double, et inversement.

Ainsi, même si les nombres pairs sont inclus dans l'ensemble des nombres entiers, il y en a une infinité $\aleph_0$ égaux, les deux ensembles sont donc équipotents. En affirmant qu'un ensemble peut être égal à une de ses parties, Cantor va à l'encontre ce qui semblait être une évidence pour Aristote et Euclide: l'ensemble de tous les ensembles est infini ! Cela va ébranler la totalité des mathématiques et va amener à l'axiomatisation de Zermelo-Fraenkel que nous verrons dans le chapitre de Théorie Des Ensembles.

A partir de ce qui précède, Cantor établit les règles de calculs suivants sur les cardinaux:

$$\aleph_0 + 1 = \aleph_0, \quad \aleph_0 + \aleph_0 = \aleph_0, \quad \aleph_0^2 = \aleph_0 \quad (2.49)$$

À première vue ces règles semblent non intuitives mais en fait elles le sont bien! En effet, Cantor définit l'addition de deux nombres transfinis comme le cardinal de l'union disjointe des ensembles correspondants.

Exemples:

E1. En notant donc $\aleph_0$ le cardinal de $\mathbb{N}$ nous avons $\aleph_0 + \aleph_0$ qui est équivalent à dire que nous sommions le cardinal de $\mathbb{N}$ union disjointe $\mathbb{N}$. Or $\mathbb{N}$ union disjointe $\mathbb{N}$ est équipotent à $\mathbb{N}$ donc $\aleph_0 + \aleph_0 = \aleph_0$ (il suffit pour s'en convaincre de prendre l'ensemble des entiers pairs et impairs tout deux dénombrables dont l'union disjointe est dénombrable).

E2. Autre exemple trivial: $\aleph_0 + 1$ correspond au cardinal de l'ensemble $\mathbb{N}$ union un point. Ce dernier ensemble est encore équipotent à $\mathbb{N}$ donc $\aleph_0 + 1 = \aleph_0$.

Nous verrons également lors de notre étude du chapitre de Théorie Des Ensembles que le concept de produit cartésien de deux ensembles dénombrables est tel que nous ayons:

$$\text{Card}(\mathbb{N} \times \mathbb{N}) = \text{Card}(\mathbb{N}^2) = [\text{Card}(\mathbb{N})]^2 \quad (2.50)$$

et donc:

$$\aleph_0^2 = \aleph_0 \quad (2.51)$$

De même (cf. chapitre de Théorie Des Ensembles), puisque $\mathbb{Z} = \mathbb{Z}^+ \cup \mathbb{Z}^-$ nous avons:

$$\aleph_0 + \aleph_0 = \aleph_0 \quad (2.52)$$

et en identifiant $\mathbb{Q}$ à $\mathbb{Z} \times \mathbb{Z}$ (rapport d'un numérateur sur un dénominateur), nous avons immédiatement:

$$\aleph_0 \times \aleph_0 = (\aleph_0)^2 = \aleph_0 \quad (2.53)$$

Nous pouvons d'ailleurs démontrer un énoncé intéressant: si nous considérons le cardinal de l'ensemble

de tous les cardinaux, il est nécessairement plus grand que tous les cardinaux, y compris lui-même (il vaut mieux avoir lu le chapitre de Théorie Des Ensembles au préalable)! En d'autres termes: le cardinal de l'ensemble de tous les ensembles de A est plus grand que le cardinal de A lui-même.

Ceci implique qu'il n'existe aucun ensemble qui contient tous les ensembles puisqu'il en existe toujours un qui est plus grand (c'est une forme équivalente du fameux ancien paradoxe de Cantor).

Dans un langage technique cela revient à considérer un ensemble non vide A et alors d'énoncer que:

$$\text{Card}(A) < \text{Card}(\mathcal{P}(A)) \quad (2.54)$$

où $\mathcal{P}(A)$ est l'ensemble des parties de A (voir le chapitre de Théorie des Ensembles pour le calcul général du cardinal de l'ensemble des parties d'un ensemble dénombrable).

C'est-à-dire par définition de la relation d'ordre $<$ (strictement inférieur), qu'il suffit de montrer qu'il n'existe pas d'application surjective $f : A \rightarrow \mathcal{P}(A)$, en d'autres termes qu'à chaque élément de l'ensemble des parties de A il ne correspond pas au moins une pré-image dans A .

Remarque: $\mathcal{P}(\mathbb{N})$ est par exemple constitué de l'ensemble des nombres impairs, pairs, premiers, et l'ensemble des naturels, ainsi que l'ensemble vide lui-même, etc. $\mathcal{P}(\mathbb{N})$ est donc l'ensemble de toutes les "patates" (pour emprunter le vocabulaire de la petite école...) possibles qui forment $\mathbb{N}$.

Démonstration (par l'absurde):

L'idée maintenant est de supposer que nous pouvons numéroter chacune des patates de $\mathcal{P}(A)$ avec au moins un élément de A (imaginez cela avec $\mathbb{N}$ ou allez voir l'exemple dans le chapitre de Théorie Des Ensembles). En d'autres termes cela revient à supposer que $f : A \rightarrow \mathcal{P}(A)$ est surjective et considérons un sous-ensemble E de A tel que:

$$E = \{x \in A \mid x \notin f(x)\} \quad (2.55)$$

c'est-à-dire l'ensemble d'éléments x de A qui n'appartiennent pas à l'ensemble numéro x (l'élément x n'appartient pas à la patate qu'il numérote... en d'autres termes).

Or, si f est surjective il doit alors exister aussi un $y \in A$ pour ce sous-ensemble E tel que:

$$f(y) = E = \{x \in A \mid x \notin f(x)\} \quad (2.56)$$

puisque E est aussi une partie de A .

Si $f(y) = E$ alors $y \notin f(y)$ mais de par la définition de E , $y \in E$ et nous avons donc une absurdité de par l'hypothèse de la surjectivité!

□C.Q.F.D.

2.7. NOMBRES COMPLEXES

Inventés au 16ème siècle entre autres par Jérôme Cardan et Raffaello Bombelli, ces nombres permettent de résoudre des problèmes n'ayant pas de solutions dans $\mathbb{R}$ ainsi que de formaliser mathématiquement certaines transformations dans le plan telles que la rotation, la similitude, la translation, etc. Pour les physiciens, les nombres complexes constituent surtout un moyen très commode de simplifier les notations. Il est ainsi très difficile d'étudier les phénomènes ondulatoires, la relativité générale ou la mécanique quantique sans recourir aux nombres et expressions complexes.

Il existe plusieurs manières de construire les nombres complexes. La première est typique de la construction telle que les mathématiciens en ont l'habitude dans le cadre de la théorie des ensembles. Ils définissent un couple de nombres réels et définissent des opérations entre ces couples pour arriver enfin à une signification du concept de nombre complexe. La deuxième est moins rigoureuse mais son approche est plus simple et consiste à définir le nombre imaginaire pur unitaire i et ensuite de construire les opérations arithmétiques à partir de sa définition. Nous allons opter pour cette deuxième méthode.

Définitions:

D1. Nous définissons le "**nombre imaginaire unitaire pur**" que nous notons i par la propriété suivante:

$$i^2 = -1 \Leftrightarrow i = \sqrt{-1} \quad (2.57)$$

D2. Un "**nombre complexe**" est un couple d'un nombre réel a et d'un nombre imaginaire ib et s'écrit généralement sous la forme suivante:

$$z = a + ib \quad (2.58)$$

a et b étant des nombres appartenant à $\mathbb{R}$.

Nous notons l'ensemble des nombres complexes $\mathbb{C}$ et avons donc par construction:

$$\mathbb{R} \subset \mathbb{C} \quad (2.59)$$

Remarque: L'ensemble $\mathbb{C}$ est identifié au plan euclidien orienté E (cf. chapitre de Calcul Vectoriel) grâce au choix d'une base orthonormée directe (nous obtenons ainsi le "plan d'Argand-Cauchy" ou plus communément "plan de Gauss" que nous verrons un peu plus loin).

L'ensemble des nombres complexes qui constitue un corps (cf. chapitre de Théorie Des Ensembles), et noté $\mathbb{C}$, est défini (de manière simple pour commencer) dans la notation de la théorie des ensembles par:

$$\mathbb{C} = \{ z = (x + i \cdot y) \mid x, y \in \mathbb{R} \} \quad (2.60)$$

En d'autres termes nous disons que le corps $\mathbb{C}$ est le corps $\mathbb{R}$ auquel nous avons "**adjoint**" le nombre imaginaire i . Ce qui se note formellement:

$$\mathbb{R}[i] \quad (2.61)$$

L'addition et la multiplication de nombres complexes sont des opérations internes à l'ensemble des

complexes (nous reviendrons beaucoup plus en détail sur certaines propriétés des nombres complexes dans le chapitre traitant de la Théorie Des Ensembles) et définies par:

$$\begin{aligned} z_1 + z_2 &= (x_1 + x_2) + i(y_1 + y_2) \\ z_1 \cdot z_2 &= (x_1 x_2 - y_1 y_2) + i(x_1 y_2 + x_2 y_1) \end{aligned} \quad (2.62)$$

La "**partie réelle**" de z est traditionnellement notée:

$$\Re(z) = x \quad (2.63)$$

La "**partie imaginaire**" de z est traditionnellement notée:

$$\Im(z) = y \quad (2.64)$$

Le "**conjugué**" ou "**conjugaison**" de z est défini par:

$$\bar{z} = x - iy \quad (2.65)$$

et est aussi parfois noté z^* (en particulier en physique quantique dans certains ouvrages!).

A partir d'un complexe et de son conjugué, il est possible de trouver ses parties réelles et imaginaires. Ce sont les relations évidentes suivantes:

$$\Re(z) = \frac{z + \bar{z}}{2} \quad \text{et} \quad \Im(z) = \frac{z - \bar{z}}{2i} \quad (2.66)$$

Le "**module**" de z (ou "norme") représente la longueur par rapport au centre du plan de Gauss (voir un peu plus bas ce qu'est le plan de Gauss) et est simplement calculé avec l'aide du théorème de Pythagore:

$$|z| = \sqrt{x^2 + y^2} = \sqrt{z \cdot \bar{z}} \quad (2.67)$$

et est donc toujours un nombre positif ou nul.

Remarque: La notation $|z|$ pour le module n'est pas innocente puisque $|z|$ coïncide avec la valeur absolue de z lorsque z est réel.

La division entre deux complexes se calcule comme (le dénominateur étant évidemment non nul):

$$\frac{z_1}{z_2} = \frac{x_1 + iy_1}{x_2 + iy_2} = \frac{x_1 + iy_1}{x_2 + iy_2} \cdot \frac{x_2 - iy_2}{x_2 - iy_2} = \frac{(x_1 x_2 + y_1 y_2) - i(x_1 y_2 - x_2 y_1)}{x_2^2 + y_2^2} \quad (2.68)$$

L'inverse d'un complexe se calculant de façon similaire:

$$\frac{1}{x + iy} = \frac{(x - iy)}{(x + iy)(x - iy)} = \frac{x - iy}{x^2 - (iy)^2} = \frac{x - iy}{x^2 + y^2} = \frac{x}{x^2 + y^2} - i \frac{y}{x^2 + y^2} \quad (2.69)$$

Nous pouvons aussi énumérer 8 importantes propriétés du module et du conjugué complexe:

P1. Nous affirmons que:

$$|z| = 0 \Leftrightarrow z = 0 \quad (2.70)$$

Démonstration:

Par définition du module $|z| = \sqrt{x^2 + y^2}$, pour que la somme $x^2 + y^2$ soit nulle, la condition nécessaire est que:

$$x = y = 0 \quad (2.71)$$

□C.Q.F.D.

P2. Nous affirmons que:

$$|z| = |-z| = |\bar{z}| \quad (2.72)$$

Démonstration:

$$|z| = \sqrt{x^2 + y^2} = \sqrt{(-x)^2 + (-y)^2} = |-z| = \sqrt{(x)^2 + (-y)^2} = |\bar{z}| \quad (2.73)$$

□C.Q.F.D.

P3. Nous affirmons que:

$$\begin{cases} |\Re(z)| \leq |z| \text{ avec égalité si et seulement si } z \text{ est réel} \\ |\Im(z)| \leq |z| \text{ avec égalité si et seulement si } z \text{ est imaginaire} \end{cases} \quad (2.74)$$

Démonstration:

Les deux inégalités ci-dessus peuvent s'écrire:

$$\begin{aligned} |a| &\leq \sqrt{a^2 + b^2} \\ |b| &\leq \sqrt{a^2 + b^2} \end{aligned} \quad (2.75)$$

donc équivalent respectivement à:

$$\begin{aligned} a^2 &\leq a^2 + b^2 \\ b^2 &\leq a^2 + b^2 \end{aligned} \quad (2.76)$$

qui sont triviales. La suite est alors triviale...

□C.Q.F.D.

P4. Nous avons:

$$\forall (z_1, z_2) \in \mathbb{C} \times \mathbb{C} \quad |z_1 z_2| = |z_1| |z_2| \quad (2.77)$$

et si:

$$z_2 \neq 0, \quad \left| \frac{z_1}{z_2} \right| = \frac{|z_1|}{|z_2|} \quad (2.78)$$

Démonstrations:

$$\begin{aligned} |z_1 z_2|^2 &= (z_1 z_2)(\overline{z_1 z_2}) = (z_1 z_2)(\overline{z_1} \overline{z_2}) = (z_1 \overline{z_1})(z_2 \overline{z_2}) = |z_1|^2 |z_2|^2 \\ \Rightarrow |z_1 z_2| &= |z_1| |z_2| \end{aligned} \quad (2.79)$$

(nous démontrerons un peu plus bas en toute généralité que $\overline{z_1 z_2} = \overline{z_1} \overline{z_2}$) et:

$$\left| \frac{z_1}{z_2} \right|^2 = \frac{z_1 \overline{z_1}}{z_2 \overline{z_2}} = \frac{|z_1|^2}{|z_2|^2} \quad (2.80)$$

□C.Q.F.D.

P5. Nous affirmons (à nouveau...) que:

$$|z|^2 = z \cdot \overline{z} \quad (2.81)$$

Démonstration:

$$|z|^2 = \sqrt{x^2 + y^2}^2 = (x + iy)(x - iy) = x^2 - ixy + iyx + y^2 = x^2 + y^2 = \overline{z} \cdot z \quad (2.82)$$

□C.Q.F.D.

P6. Nous affirmons que:

$$\forall z, z' \in \mathbb{C} \quad \overline{\overline{z}} = z, \quad \overline{z + z'} = \overline{z} + \overline{z'}, \quad \overline{z \cdot z'} = \overline{z} \cdot \overline{z'} \quad (2.83)$$

Démonstrations:

$$\overline{\overline{z}} = \overline{x + iy} = \overline{x - iy} = x + iy = z \quad (2.84)$$

et:

$$\begin{aligned} \overline{z + z'} &= \overline{(x_1 + iy_1) + (x_2 + iy_2)} = \overline{(x_1 + x_2) + i(y_1 + y_2)} = (x_1 + x_2) - i(y_1 + y_2) \\ &= (x_1 - iy_1) + (x_2 - iy_2) = \overline{z} + \overline{z'} \end{aligned} \quad (2.85)$$

et:

$$\begin{aligned} \overline{zz'} &= \overline{(x_1 + iy_1)(x_2 + iy_2)} = \overline{(x_1 x_2 - y_1 y_2) + i(x_1 y_2 + y_1 x_2)} = (x_1 x_2 - y_1 y_2) - i(x_1 y_2 + y_1 x_2) \\ &= (x_1 - iy_1)(x_2 - iy_2) = \overline{z} \overline{z'} \end{aligned} \quad (2.86)$$

Remarques:

R1. En des termes mathématiques, la première démonstration permet de montrer que la conjugaison complexe est ce que l'on appelle "involutive" (dans le sens qu'elle ne fait rien évoluer...).

R2. En des termes tout aussi mathématiques (ce n'est que du vocabulaire!), la deuxième démonstration montre que la conjugaison de la somme de deux nombres complexes est ce que nous appelons un "automorphisme du groupe" $(\mathbb{C}, +)$ (cf. chapitre de Théorie Des Ensembles).

R3. Encore une fois, pour le vocabulaire..., la troisième démonstration montre que la conjugaison du produit de deux nombres complexes est ce que nous appelons un "automorphisme du corps" $(\mathbb{C}, +, \times)$ (cf. chapitre de Théorie Des Ensembles).

P7. Nous affirmons que pour z différent de zéro:

$$\overline{\left(\frac{1}{z}\right)} = \frac{1}{\bar{z}} \quad \overline{\left(\frac{z}{z'}\right)} = \frac{\bar{z}}{\bar{z'}} \quad (2.87)$$

Nous nous restreindrons à la démonstration de la seconde relation qui est un cas général de la première (pour $z = 1$).

Démonstration:

$$\begin{aligned} \overline{\left(\frac{z}{z'}\right)} &= \frac{\overline{x_1 + iy_1}}{\overline{x_2 + iy_2}} = \frac{\overline{x_1 + iy_1} \cdot \overline{x_2 - iy_2}}{\overline{x_2 + iy_2} \cdot \overline{x_2 - iy_2}} = \frac{\overline{(x_1 x_2 + y_1 y_2) + i(y_1 x_2 - y_2 x_1)}}{x_2^2 + y_2^2} \\ &= \frac{(x_1 x_2 + y_1 y_2)}{x_2^2 + y_2^2} - i \frac{(y_1 x_2 - y_2 x_1)}{x_2^2 + y_2^2} = \frac{(x_1 x_2 + y_1 y_2) + i(y_2 x_1 - y_1 x_2)}{x_2^2 + y_2^2} \quad (2.88) \\ &= \frac{(x_1 - iy_1)(x_2 + iy_2)}{x_2^2 + y_2^2} = \frac{x_1 - iy_1}{x_2 - iy_2} \frac{x_2 + iy_2}{x_2 + iy_2} = \frac{\bar{z}}{\bar{z'}} \end{aligned}$$

□C.Q.F.D.

P8. Nous avons:

$$|z_1 + z_2|^2 = |z_1|^2 + 2\Re(z_1 \bar{z}_2) + |z_2|^2 \quad (2.89)$$

Démonstration:

$$\begin{aligned}
|z_1 + z_2|^2 &= |x_1 + iy_1 + x_2 + iy_2|^2 = |(x_1 + x_2) + i(y_1 + y_2)|^2 \\
&= \left(\sqrt{(x_1 + x_2)^2 + (y_1 + y_2)^2} \right)^2 = (x_1 + x_2)^2 + (y_1 + y_2)^2 \\
&= x_1^2 + 2x_1x_2 + x_2^2 + y_1^2 + 2y_1y_2 + y_2^2 = (x_1^2 + y_1^2) + 2(x_1x_2 + y_1y_2) + (x_2^2 + y_2^2) \\
&= |z_1|^2 + 2\Re((x_1 + iy_1)(x_2 - iy_2)) + |z_2|^2 = |z_1|^2 + 2\Re(z_1 \overline{z_2}) + |z_2|^2
\end{aligned} \tag{2.90}$$

□C.Q.F.D.

P9. Nous avons:

$$|z_1 + z_2| \leq |z_1| + |z_2| \tag{2.91}$$

pour tous complexes z_1, z_2 (rigoureusement non nuls car sinon le concept d'argument du nombre complexe que nous verrons plus loin est alors indéterminé). De plus l'égalité a lieu si et seulement si z_1 et z_2 sont colinéaires (les vecteurs sont "sur la même droite") et de même sens, autrement dit s'il existe $\lambda \in \mathbb{R}_+$ tel que $z_1 = \lambda z_2$.

Démonstration:

$$|z_1 + z_2|^2 = |z_1|^2 + 2\Re(z_1 \overline{z_2}) + |z_2|^2 \leq |z_1|^2 + 2|z_1 \overline{z_2}| + |z_2|^2 = (|z_1| + |z_2|)^2 \tag{2.92}$$

Cette inégalité peut ne pas paraître évidente à tout le monde alors développons un peu et supposons-la vraie:

$$\begin{aligned}
|z_1 + z_2|^2 &\stackrel{?}{\leq} (|z_1| + |z_2|)^2 \\
\sqrt{(x_1 + x_2)^2 + (y_1 + y_2)^2}^2 &\stackrel{?}{\leq} \left(\sqrt{x_1^2 + y_1^2} + \sqrt{x_2^2 + y_2^2} \right)^2 \\
(x_1 + x_2)^2 + (y_1 + y_2)^2 &\stackrel{?}{\leq} (x_1^2 + y_1^2) + 2\sqrt{x_1^2 + y_1^2} \sqrt{x_2^2 + y_2^2} + (x_2^2 + y_2^2) \\
x_1^2 + 2x_1x_2 + x_2^2 + y_1^2 + 2y_1y_2 + y_2^2 &\stackrel{?}{\leq} (x_1^2 + y_1^2) + 2\sqrt{x_1^2 + y_1^2} \sqrt{x_2^2 + y_2^2} + (x_2^2 + y_2^2)
\end{aligned} \tag{2.93}$$

Après simplification:

$$\begin{aligned}
x_1x_2 + y_1y_2 &\stackrel{?}{\leq} \sqrt{x_1^2 + y_1^2} \sqrt{x_2^2 + y_2^2} \\
(x_1x_2 + y_1y_2)^2 &\stackrel{?}{\leq} (x_1^2 + y_1^2)(x_2^2 + y_2^2) \\
x_1^2x_2^2 + 2x_1x_2y_1y_2 + y_1^2y_2^2 &\stackrel{?}{\leq} x_1^2x_2^2 + x_1^2y_2^2 + y_1^2x_2^2 + y_1^2y_2^2
\end{aligned} \tag{2.94}$$

et encore après simplification:

$$\begin{aligned}
 2x_1 x_2 y_1 y_2 &\stackrel{?}{\leq} x_1^2 y_2^2 + y_1^2 x_2^2 \\
 0 &\stackrel{?}{\leq} x_1^2 y_2^2 - 2x_1 x_2 y_1 y_2 + y_1^2 x_2^2 \\
 0 &\stackrel{?}{\leq} (x_1 y_2 - y_1 x_2)^2
 \end{aligned} \quad (2.95)$$

donc comme la parenthèse au carré est forcément positive ou nulle il s'ensuit:

$$0 \leq (x_1 y_2 - y_1 x_2)^2 \quad (2.96)$$

Cette dernière relation démontre donc que l'inégalité est vraie.

□C.Q.F.D.

Remarque: Il existe une forme plus générale de cette inégalité appelée "inégalité de Minkowski" présentée dans le chapitre de Calcul Vectoriel (les nombres complexes peuvent effectivement s'écrire sous la forme de vecteurs comme nous allons le voir de suite).

2.7.1. INTERPRÉTATION GÉOMÉTRIQUE

Nous pouvons aussi représenter un nombre complexe $a + ib$ ou $a - ib$ dans un plan délimité par deux axes (deux dimensions) de longueur infinie et orthogonaux entres eux. L'axe vertical représentant la partie imaginaire d'un nombre complexe et l'axe horizontal la partie réelle (voir figure ci-après).

Il y donc bijection entre l'ensemble des nombres complexes et l'ensemble des vecteurs du plan de Gauss (notion d'affixe).

Nous nommons parfois ce type de représentation "**plan de Gauss**":

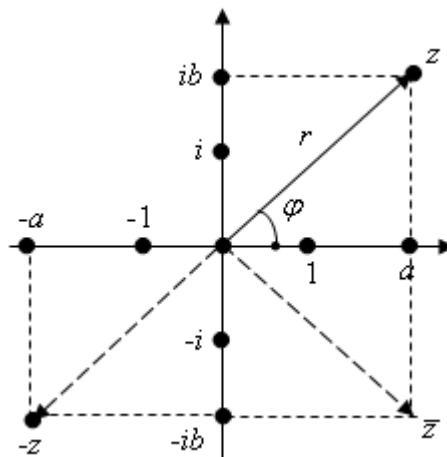


Figure: 2.5 - Plan de Gauss

et nous écrivons alors:

$$Aff(\vec{r}) = a + ib \quad (2.97)$$

Nous voyons sur ce diagramme qu'un nombre complexe a donc une interprétation vectorielle (cf.

chapitre de Calcul Vectoriel) donnée par:

$$z = a \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} b = a + ib \quad (2.98)$$

où la base canonique est définie telle que:

$$\left\{ 1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, i = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\} \quad (2.99)$$

avec:

$$r = |z| = \sqrt{a^2 + b^2} \quad (2.100)$$

Ainsi, $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ est le vecteur de la base unitaire porté par l'axe horizontal $\mathbb{R}$ et $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ est le vecteur de la base unitaire porté par l'axe imaginaire $\mathbb{R}_i$ et r est le module (la norme) positif ou nul.

Ceci est à comparer avec les vecteurs de $\mathbb{R}^2$ (cf. chapitre de Calcul Vectoriel):

$$\vec{v} = x\vec{e}_1 + y\vec{e}_2 = x \begin{pmatrix} 1 \\ 0 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} \quad (2.101)$$

avec:

$$\|\vec{v}\| = \sqrt{x^2 + y^2} \quad (2.102)$$

ce qui fait que nous pouvons identifier le plan complexe avec le plan euclidien.

Par ailleurs, les définitions du cosinus et sinus (cf. chapitre de Trigonométrie) nous donnent:

$$a = r \cos \varphi \quad b = r \sin \varphi \quad (2.103)$$

Finalement:

$$\boxed{\begin{aligned} r &= \sqrt{a^2 + b^2} \\ \varphi &= \cos^{-1} \left(\frac{a}{r} \right) = \sin^{-1} \left(\frac{b}{r} \right) = \tan^{-1} \left(\frac{b}{a} \right) \end{aligned}} \quad (2.104)$$

Ainsi:

$$z = a + ib = r \cos \varphi + i r \sin \varphi = r(\cos \varphi + i \sin \varphi) = r \cdot \text{cis } \varphi \quad (2.105)$$

complexe qui est toujours égal à lui-même modulo 2π de par les propriétés des fonctions trigonométriques:

$$z = r(\cos \varphi + i \sin \varphi) = r(\cos(\varphi + k2\pi) + i \sin(\varphi + k2\pi)) \quad (2.106)$$

avec $k \in \mathbb{N}$ et où φ est appelé "l'argument de z " et est noté traditionnellement:

$$\arg(z) \quad (2.107)$$

Les propriétés du cosinus et du sinus (cf. chapitre de Trigonométrie) nous amènent directement à écrire pour l'argument:

$$\arg(\bar{z}) = -\arg(z) \text{ et } \arg(-z) = \arg(z) + \pi \quad (2.108)$$

Nous démontrons entre autres avec les séries de Taylor (cf. chapitre des Suites Et Séries) que:

$$\cos \varphi = 1 - \frac{\varphi^2}{2!} + \frac{\varphi^4}{4!} - \dots + (-1)^k \frac{\varphi^{2k}}{(2k)!} + \dots \quad (2.109)$$

et:

$$\sin \varphi = \varphi - \frac{\varphi^3}{3!} + \frac{\varphi^5}{5!} - \dots + (-1)^k \frac{\varphi^{2k+1}}{(2k+1)!} + \dots \quad (2.110)$$

dont la somme est semblable à:

$$e^\varphi = 1 + \varphi + \frac{\varphi^2}{2!} + \frac{\varphi^3}{3!} + \dots + \frac{\varphi^k}{k!} + \dots \quad (2.111)$$

mais par contre parfaitement identique au développement de Taylor de $e^{i\varphi}$:

$$e^{i\varphi} = 1 + i\varphi - \frac{\varphi^2}{2!} - i\frac{\varphi^3}{3!} - \dots + (i)^k \frac{\varphi^k}{k!} + \dots = \cos \varphi + i \sin \varphi \quad (2.112)$$

Donc finalement, nous pouvons écrire:

$$z = r(\cos \varphi + i \sin \varphi) = r \cdot e^{i\varphi} \quad (2.113)$$

relation nommée "formule d'Euler".

Grâce à la forme exponentielle d'un nombre complexe, très fréquemment utilisée dans de nombreux domaines de la physique et de l'ingénierie, nous pouvons très facilement tirer des relations telles que (*cis* est une vieille notation qui est l'abréviation du *cos i sin* se trouvant dans la parenthèse):

$$z = r(\cos \varphi + i \sin \varphi) = r \operatorname{cis} \varphi = r e^{i\varphi}$$

$$z_1 = r_1(\cos \varphi_1 + i \sin \varphi_1) = r_1 \operatorname{cis} \varphi_1 = r_1 e^{i\varphi_1} \quad (2.114)$$

$$z_2 = r_2(\cos \varphi_2 + i \sin \varphi_2) = r_2 \operatorname{cis} \varphi_2 = r_2 e^{i\varphi_2}$$

et en supposant connues les relations trigonométriques de bases (cf. chapitre de Trigonométrie) nous avons les relations suivantes pour la multiplication de deux nombres complexes:

$$\begin{aligned} z_1 \cdot z_2 &= r_1 \cdot r_2 [\cos(\varphi_1 + \varphi_2) + i \sin(\varphi_1 + \varphi_2)] = r_1 \cdot r_2 \operatorname{cis}(\varphi_1 + \varphi_2) \\ &= r_1 e^{i\varphi_1} r_2 e^{i\varphi_2} = r_1 r_2 e^{i(\varphi_1 + \varphi_2)} \end{aligned} \quad (2.115)$$

dès lors:

$$\arg(z_1 \cdot z_2) = \arg(z_1) + \arg(z_2) \quad (2.116)$$

et donc si n est un entier positif:

$$\arg(z^n) = n \arg(z) \quad (2.117)$$

Pour le module de la multiplication (nous changeons de notation pour la lisibilité):

$$|z_1 z_2| = |r e^{i\varphi} r' e^{i\varphi'}| = |r r' e^{i(\varphi + \varphi')}| = r r' = |z_1| |z_2| \quad (2.118)$$

d'où:

$$|z^m| = |z|^m \quad (2.119)$$

Pour la division de deux nombres complexes:

$$\begin{aligned} \frac{z_1}{z_2} &= \frac{r_1}{r_2} [\cos(\varphi_1 - \varphi_2) + i \sin(\varphi_1 - \varphi_2)] = \frac{r_1}{r_2} \operatorname{cis}(\varphi_1 - \varphi_2) \\ &= \frac{r_1 e^{i\varphi_1}}{r_2 e^{i\varphi_2}} = \frac{r_1}{r_2} e^{i(\varphi_1 - \varphi_2)} \end{aligned} \quad (2.120)$$

Le module de leur division vient alors immédiatement:

$$\left| \frac{z_1}{z_2} \right| = \frac{|z_1|}{|z_2|} \quad (2.121)$$

dès lors nous avons pour l'argument:

$$\arg\left(\frac{z_1}{z_2}\right) = \arg\left(\frac{r e^{i\varphi}}{r' e^{i\varphi'}}\right) = \arg\left(\frac{r}{r'} e^{i(\varphi - \varphi')}\right) = \varphi - \varphi' = \arg(z_1) - \arg(z_2) \quad (2.122)$$

ainsi il vient immédiatement:

$$\arg(z^{-1}) = \arg\left((r e^{i\varphi})^{-1}\right) = \arg\left(\frac{1}{r} e^{-i\varphi}\right) = -\varphi = -\arg(z) \quad (2.123)$$

Pour la mise en puissance d'un nombre complexe (ou la racine):

$$z^m = r^m e^{im\varphi} = r^m [\cos(m\varphi) + i \sin(m\varphi)] = r^m \operatorname{cis}(m\varphi) \quad (2.124)$$

ce qui nous donne immédiatement un résultat déjà mentionné plus haut:

$$|z^m| = |z|^m \quad (2.125)$$

et pour l'argument:

$$\arg(z^m) = \arg\left((re^{i\varphi})^m\right) = \arg(r^m e^{im\varphi}) = m\varphi = m \arg(z) \quad (2.126)$$

Dans le cas où nous avons un module unité tel que $z = \cos(\varphi) + i \sin(\varphi)$ nous avons alors la relation:

$$(\cos \varphi + i \sin \varphi)^m = (e^{i\varphi})^m = e^{im\varphi} = \cos(m\varphi) + i \sin(m\varphi) \quad (2.127)$$

appelée "**formule de De Moivre**".

Pour le logarithme népérien d'un nombre complexe, nous avons trivialement la relation suivante sur laquelle nous reviendrons dans le chapitre d'Analyse Complexe:

$$\ln(z) = \ln(re^{i\varphi}) = \ln(r) + i\varphi \quad (2.128)$$

où $\ln(z)$ est souvent dans le cas complexe écrit $\text{Log}(z)$ avec un "L" majuscule.

Toutes les relations précédentes pourraient bien sûr être obtenues avec la forme trigonométrique des nombres complexes mais nécessiteraient alors quelques lignes supplémentaires de développements.

Remarque: Une variation sinusoïdale $f(t) = r \sin(\omega t)$ peut être représentée comme la projection (cf. chapitre de Trigonométrie) sur l'axe vertical y (axe des imaginaires de l'ensemble $\mathbb{C}$) d'un vecteur $\vec{r}$ tournant à vitesse angulaire ω autour de l'origine dans le plan xOy:

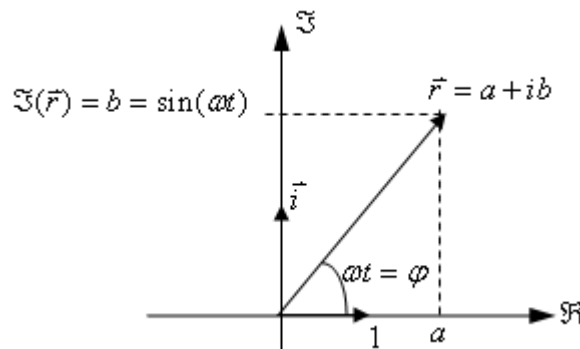


Figure: 2.6 - Représentation d'un vecteur de Fresnel

Un tel vecteur tournant s'appelle "**vecteur de Fresnel**" et peut très bien être interprété comme la partie imaginaire d'un nombre complexe donné par:

$$\vec{r} = \Re(\vec{r}) + i\Im(\vec{r}) = r(\cos(\omega t) + i \sin(\omega t)) = re^{i\omega t} = re^{i\varphi} \quad (2.129)$$

Nous retrouverons les vecteurs tournants de façon explicite lors de notre étude de la mécanique ondulatoire et optique géométrique (dans le cadre de la diffraction).

2.7.2. TRANSFORMATIONS DANS LE PLAN

Il est habituel de représenter les nombres réels comme points d'une droite graduée. Les opérations algébriques y ont leur interprétation géométrique: l'addition est une translation, la multiplication une homothétie centrée à l'origine.

En particulier nous pouvons parler de la "racine carrée d'une transformation". Une translation d'amplitude a peut être obtenue comme l'itération d'une translation d'amplitude $a/2$. De même une homothétie de rapport a peut être obtenue comme l'itérée d'une homothétie de rapport $\sqrt{a}$. En particulier une homothétie de rapport 9 est la composée de deux homothéties de rapport 3 (ou -3).

La racine carrée prend alors un sens géométrique. Mais qu'en est-il de la racine carrée de nombres négatifs? En particulier de la racine carrée de -1?

Une homothétie de rapport -1 peut être vue comme une symétrie par rapport à l'origine. Toutefois si nous voulons voir cette transformation d'une manière continue, force nous est de placer la droite dans un plan. Dès lors une homothétie de rapport -1 peut être vue comme une rotation de π radians autour de l'origine.

Du coup, le problème de la racine carrée négative se simplifie. En effet, il n'est guère difficile de décomposer une rotation de π radians en deux transformations: nous pouvons répéter soit une rotation de $\pi/2$ soit une rotation de $-\pi/2$. L'image de 1 sera la racine carrée de -1 et i est située sur une perpendiculaire à l'origine à une distance 1 soit vers le haut soit vers le bas.

Ayant réussi à positionner le nombre i il n'est plus guère difficile de disposer les autres nombres complexes dans un plan de Gauss. Nous pouvons ainsi associer à $2i$ le produit de l'homothétie (cf. chapitre de Géométrie Euclidienne) de rapport 2 par la rotation de centre O et d'angle $\pi/2$, soit une similitude centrée à l'origine. C'est ce que nous allons nous efforcer à montrer maintenant.

Soient:

$$z_1 = x_1 + iy_1 = ae^{i\alpha}, z_2 = x_2 + iy_2 = be^{i\beta} \quad (2.130)$$

et $k \in \mathbb{N}$.

Nous avons les propriétés de transformations géométriques suivantes pour les nombres complexes (voir le chapitre de Trigonométrie pour les propriétés du sinus et cosinus) que nous pouvons joyeusement combiner selon notre bon vouloir:

P1. La multiplication de z_1 par un réel λ dans le plan de Gauss correspond (trivial) à une homothétie (agrandissement) de centre O (l'intersection des axes imaginaires et réels), de rapport λ .

Démonstration:

$$\lambda z_1 = (\lambda \cdot a) e^{i\alpha} \quad (2.131)$$

□C.Q.F.D.

P2. La multiplication de z_1 par un nombre complexe de module unitaire:

$$z_0 = x_0 + iy_0 = \cos \varpi + i \sin \varpi = e^{i\varpi} \quad (2.132)$$

correspond à une rotation de centre O et d'angle du complexe z_1 .

Démonstration:

$$z_0 z_1 = e^{i\varpi} \alpha e^{i\alpha} = \alpha e^{i(\alpha+\varpi)} \quad (2.133)$$

□C.Q.F.D.

Remarque: Nous voyons alors immédiatement, par exemple, que multiplier un nombre complexe par i (c'est-à-dire $\sin(\varpi) = 1, \cos(\varpi) = 0$) correspond à une rotation de $\pi/2$.

Il est intéressant d'observer que sous forme vectorielle la rotation de centre O de z_1 par z_0 peut s'écrire à l'aide de la matrice suivante:

$$\begin{bmatrix} x_0 & -y_0 \\ y_0 & x_0 \end{bmatrix} \quad (2.134)$$

Démonstration:

Nous savons que $z_0 z_1$ est une rotation de centre O et d'angle ϖ . Il suffit de l'écrire à l'ancienne:

$$z_0 z_1 = (x_0 + iy_0)(x_1 + iy_1) = (x_0 x_1 - y_0 y_1) + i(x_0 y_1 + y_0 x_1) \quad (2.135)$$

ce qui donne sous forme vectorielle:

$$z_0 z_1 = \begin{pmatrix} x_0 x_1 - y_0 y_1 \\ 0 \end{pmatrix} + i \begin{pmatrix} 0 \\ x_0 y_1 + y_0 x_1 \end{pmatrix} \quad (2.136)$$

donc l'application linéaire est équivalente à:

$$\begin{bmatrix} x_0 & -y_0 \\ y_0 & x_0 \end{bmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 x_1 - y_0 y_1 \\ x_0 y_1 + y_0 x_1 \end{pmatrix} \quad (2.137)$$

ou encore (nous retombons sur la matrice de rotation dans le plan que nous avons dans le chapitre de Géométrie Euclidienne ce qui est un résultat remarquable!) en utilisant:

$$z = r(\cos \varphi + i \sin \varphi) = r(\cos(\varphi + k2\pi) + i \sin(\varphi + k2\pi)) \quad (2.138)$$

dans le cas particulier et arbitraire où r serait unitaire (afin d'avoir une rotation pure!):

$$z = \cos(\varphi + k2\pi) + i \sin(\varphi + k2\pi) = x_0 + iy_0 \quad (2.139)$$

nous avons immédiatement (nous avons repris les notations de l'angle tel que nous l'avons dans le chapitre de Géométrie):

$$\begin{bmatrix} x_0 & -y_0 \\ y_0 & x_0 \end{bmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{bmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 x_1 - y_0 y_1 \\ x_0 y_1 + y_0 x_1 \end{pmatrix} = \begin{pmatrix} \cos(\varphi) x_1 - \sin(\varphi) y_1 \\ \cos(\varphi) y_1 + \sin(\varphi) x_1 \end{pmatrix} \quad (2.140)$$

Remarquons que la matrice de rotation peut aussi s'écrire sous la forme :

$$\begin{bmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{bmatrix} = \cos(\varphi) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \sin(\varphi) \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = \cos(\varphi) \cdot I + \sin(\varphi) \cdot J \quad (2.141)$$

de même :

$$\begin{bmatrix} x_0 & -y_0 \\ y_0 & x_0 \end{bmatrix} = x_0 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + y_0 \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = x_0 \cdot I + y_0 \cdot J \quad (2.142)$$

□C.Q.F.D.

Ainsi nous remarquons que ces matrices de rotation ne sont pas que des applications mais sont des nombres complexes aussi (bon c'était évident dès le début mais fallait le montrer de manière esthétique et simple).

Ainsi, nous avons pour habitude de poser que :

$$1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, i = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \quad (2.143)$$

ou avec une autre notation fréquente en algèbre linéaire:

$$\mathbb{1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, i = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \quad (2.144)$$

Le corps des nombres complexes est donc isomorphe au corps des matrices réelles carrées de dimension 2 du type:

$$\begin{bmatrix} x_0 & -y_0 \\ y_0 & x_0 \end{bmatrix} \quad (2.145)$$

C'est un résultat que nous réutiliserons de nombreuses fois dans divers chapitres de ce site pour des études particulières en algèbre, géométrie et en physique quantique relativiste.

P3. La multiplication de deux complexes correspond à une homothétie ajoutée à une rotation. En d'autres termes, d'une "similitude directe".

Démonstration:

$$z_1 z_2 = a e^{i\alpha} b e^{i\beta} = (\underline{ab}) e^{i(\alpha+\beta)} \quad (2.146)$$

il s'agit donc bien d'une similitude de rapport b et d'angle β .

□C.Q.F.D.

Au contraire, l'opération suivante:

$$\overline{z_1 z_2} = a e^{i\alpha} b e^{-i\beta} = (\underline{ab}) e^{i(\alpha-\beta)} \quad (2.147)$$

sera appelée une "**similitude linéaire rétrograde**".

Par ailleurs, il en retourne trivialement la relation déjà connue suivante:

$$\arg(z_1 z_2) = \arg(z_1) + \arg(z_2) \quad (2.148)$$

Remarques:

R1. La somme de deux nombres $z_1 + z_2$ complexes ne pouvant avoir une écriture mathématique simplifiée sous quelque forme que ce soit, nous disons alors que la somme équivaut à une "**translation d'amplitude**".

R2. La combinaison d'une similitude linéaire (multiplication de deux nombres complexes) directe et d'une translation d'amplitude (addition par un troisième nombre complexe) correspond à ce que nous appelons une "**similitude linéaire directe**".

P4. Le conjugué d'un nombre complexe est géométriquement son symétrique par rapport à l'axe $\mathbb{R}$ tel que:

$$\overline{z_1} = x_1 - iy_1 = r(\cos \alpha - i \sin \alpha) = r(\cos(-\alpha) + i \sin(-\alpha)) = r e^{-i\alpha} \quad (2.149)$$

sans oublier que:

$$\cos(\varphi) = \cos(\varphi + k2\pi), \sin(\varphi) = \sin(\varphi + k2\pi) \quad (2.150)$$

Ce qui nous donne un résultat déjà connu:

$$\arg(\overline{z_1}) = -\arg(z_1) \quad (2.151)$$

D'où nous pouvons tirer la propriété suivante:

$$(\cos(\varphi + \pi) + i \sin(\varphi + \pi)) = r(-\cos(\varphi) - i \sin(\varphi)) = -r(\cos(\varphi) + i \sin(\varphi)) = -z_1 \quad (2.152)$$

d'où:

$$\arg(z_1) + \pi = \arg(-z_1) \quad (2.153)$$

P5. La négation du conjugué d'un nombre complexe est géométriquement son symétrique par rapport à l'axe des imaginaires $\mathbb{R}_i$ tel que:

$$-\bar{z}_1 = -x_1 + iy_1 = r(-\cos \alpha + i \sin \alpha) = r(\cos(\pi \pm \alpha) + i \sin(\alpha)) \quad (2.154)$$

Remarques:

R1. La combinaison de P4, P5 est appelée une "**similitude rétrograde**".

R2. L'opération géométrique qui consiste à prendre l'inverse du conjugué d'un nombre complexe (soit $\bar{z}^{-1}$) est appelée une "**inversion de pôle**".

P6. La rotation de centre c et d'angle φ est donnée par:

$$c + e^{i\varphi}(z_1 - c) \quad (2.155)$$

Explications:

Le complexe c donne un point dans le plan de Gauss qui sera le centre de rotation. La différence $z_1 - c$ donne le rayon r choisi. La multiplication par $e^{i\varphi}$ est la rotation du rayon par rapport à l'origine du plan de Gauss dans le sens inverse des aiguilles d'une montre. Finalement, l'addition par c la translation nécessaire pour ramener le rayon r tourné à l'origine du centre c . Ce qui donne schématiquement:

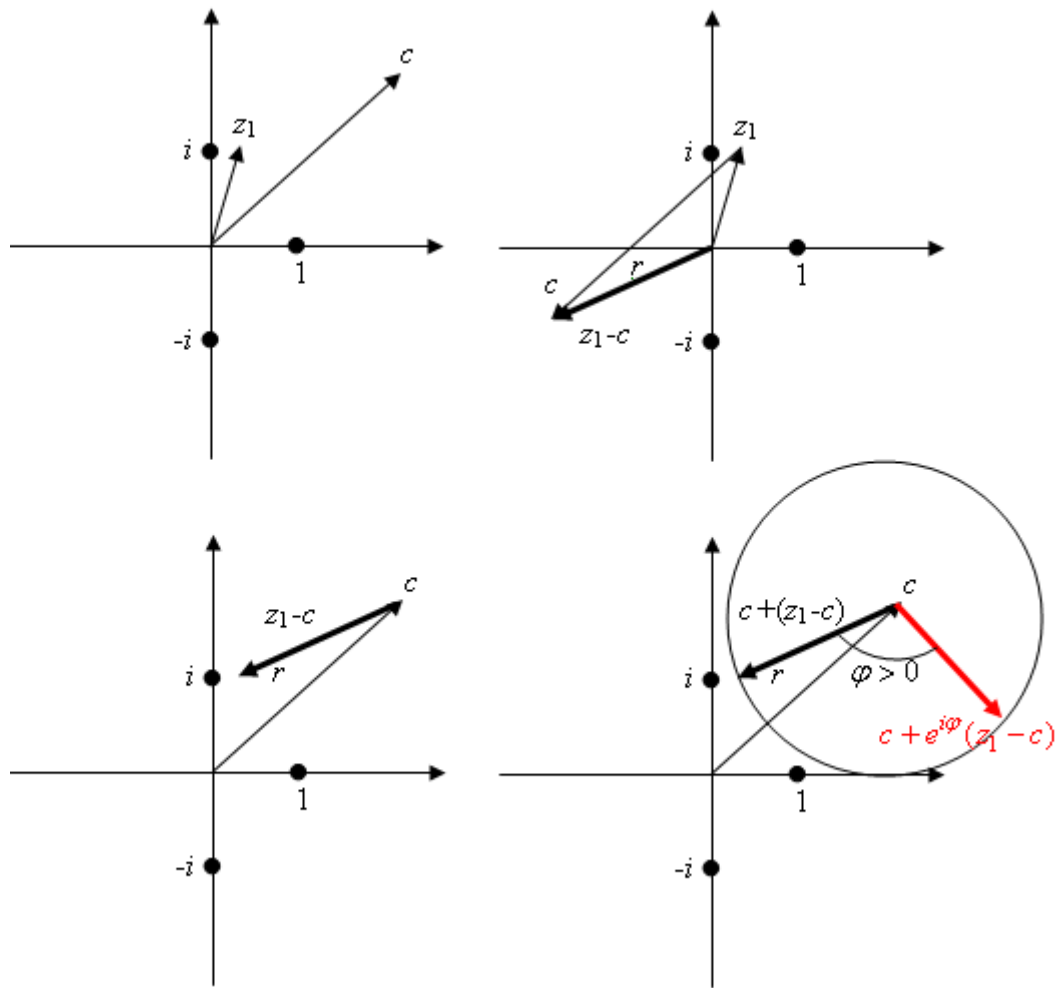


Figure: 2.7 - Représentation de la rotation complexe

P7. Sur la même idée, nous obtenons une homothétie de centre c , de rapport λ par l'opération:

$$c + \lambda(z_1 - c) \quad (2.156)$$

Explications:

La différence $z_1 - c$ donne toujours le rayon r et c un point dans le centre de Gauss. $\lambda(z_1 - c)$ donne l'homothétie du rayon par rapport à l'origine du plan de Gauss et finalement l'addition par c la translation nécessaire pour que l'homothétie soit vue comme étant faite de centre c .

2.8. NOMBRES QUATERNIONS

Appelés aussi "**hypercomplexes**", les nombres quaternions ont été inventés en 1843 par William Rowan Hamilton pour généraliser les nombres complexes.

Définition: Un quaternion est un élément $(a, b, c, d) \in \mathbb{R}^4$ et dont nous notons $\mathbb{H}$ l'ensemble qui le contient et que nous appelons "**ensemble des quaternions**".

Un "**quaternion**" peut aussi bien être représenté en ligne ou en colonne tel que:

$$(a,b,c,d) = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} \quad (2.157)$$

Nous définissons la somme de deux quaternions (a,b,c,d) et (a',b',c',d') par:

$$(a,b,c,d) + (a',b',c',d') = (a+a', b+b', c+c', d+d') \quad (2.158)$$

Il est évident (du moins nous l'espérons pour le lecteur) que $(\mathbb{H}, +)$ est un groupe commutatif (cf. chapitre de Théorie Des Ensembles), d'élément neutre $(0,0,0,0)$, l'opposé d'un élément (a,b,c,d) étant $(-a,-b,-c,-d)$

Remarque: C'est l'addition naturelle dans $\mathbb{R}^4$ vu comme $\mathbb{R}$ -espace vectoriel (cf. chapitre de Théorie Des Ensembles).

L'associativité se vérifie en appliquant les propriétés correspondantes des opérations sur $\mathbb{R}$.

Nous définissons également la multiplication:

$$(a,b,c,d) \cdot (a',b',c',d') \quad (2.159)$$

de deux quaternions (a,b,c,d) et (a',b',c',d') par l'expression:

$$\begin{pmatrix} aa' - bb' - cc' - dd' \\ (ab' + ba') + (cd' - dc') \\ (ac' + ca') - (bd' - db') \\ (da' + ad') + (bc' - cb') \end{pmatrix} \quad (2.160)$$

C'est peut-être difficile à accepter mais nous verrons un peu plus loin qu'il y a un air de famille avec les nombres complexes.

Nous pouvons remarquer que la loi de multiplication n'est pas commutative. Effectivement, en prenant la définition de la multiplication ci-dessus, nous avons:

$$\begin{aligned} (0,1,0,0) \cdot (0,0,1,0) &= (0,0,0,1) \\ (0,0,1,0) \cdot (0,1,0,0) &= (0,0,0,-1) \end{aligned} \quad (2.161)$$

Mais nous pouvons remarquer que:

$$(0,1,0,0) \cdot (0,0,1,0) = -(0,0,1,0) \cdot (0,1,0,0) \quad (2.162)$$

Remarque: La loi de multiplication est distributive avec la loi d'addition mais c'est un excellent exemple où il faut quand même prendre garde à démontrer la distributivité à gauche et à droite, puisque le produit n'est pas commutatif !

La multiplication a pour élément neutre:

$$(1,0,0,0) \quad (2.163)$$

Effectivement:

$$(1,0,0,0) \cdot (a,b,c,d) = (a,b,c,d) \cdot (1,0,0,0) = (a,b,c,d) \quad (2.164)$$

Tout élément:

$$(a,b,c,d) \in \mathbb{H}^* = \mathbb{H} - \{(0,0,0,0)\} \quad (2.165)$$

est inversible.

En effet, si (a,b,c,d) est un quaternion non nul, nous avons alors nécessairement:

$$a^2 + b^2 + c^2 + d^2 \neq 0 \quad (2.166)$$

sinon les quatre nombres a, b, c, d sont de carré nul, donc tous nuls. Soit alors le quaternion

(a_1, b_1, c_1, d_1) défini par:

$$\begin{cases} a_1 = \frac{a}{a^2 + b^2 + c^2 + d^2} \\ b_1 = \frac{-b}{a^2 + b^2 + c^2 + d^2} \\ c_1 = \frac{-c}{a^2 + b^2 + c^2 + d^2} \\ d_1 = \frac{-d}{a^2 + b^2 + c^2 + d^2} \end{cases} \quad (2.167)$$

alors en appliquant machinalement la définition de la multiplication des quaternions, nous vérifions que:

$$(a,b,c,d) \cdot (a_1,b_1,c_1,d_1) = (a_1,b_1,c_1,d_1) \cdot (a,b,c,d) = (1,0,0,0) \quad (2.168)$$

ce dernier quaternion est donc l'inverse pour la multiplication!

Montrons maintenant (pour la culture générale) que le corps des complexes $(\mathbb{C}, +, \times)$ est un sous-corps de $(\mathbb{H}, +, \times)$.

Remarque: Nous aurions pu mettre cette démonstration dans le chapitre de Théorie Des Ensembles car nous faisons usage de beaucoup de concepts qui y sont vus mais il nous a semblé un peu plus pertinent de la mettre ici.

Soit $\mathbb{H}'$ l'ensemble des quaternions de la forme $(a,b,0,0)$. Si $\mathbb{H}'$ est non vide, et si $(a,b,0,0), (a',b',0,0)$ sont des éléments de $\mathbb{H}'$ alors $(\mathbb{H}', +, \times)$ est un corps. Effectivement:

P1. Pour la soustraction (et donc l'addition):

$$(a, b, 0, 0) - (a', b', 0, 0) = (a - a', b - b', 0, 0) \in \mathbb{H}' \quad (2.169)$$

P2. La multiplication:

$$(a, b, 0, 0) \cdot (a', b', 0, 0) = (aa' - bb', ab' + ba', 0, 0) \in \mathbb{H}' \quad (2.170)$$

P3. L'élément neutre:

$$(1, 0, 0, 0) \in \mathbb{H}' \quad (2.171)$$

P4. Et finalement l'inverse:

$$(a/(a^2 + b^2), -b/(a^2 + b^2), 0, 0) \quad (2.172)$$

de $(a, b, 0, 0)$ est encore dans $\mathbb{H}'$.

Donc $(\mathbb{H}', +, \times)$ est un sous-corps de $\mathbb{H}$. Soit alors l'application:

$$f: \begin{cases} \mathbb{C} \rightarrow \mathbb{H}' \\ a + ib \mapsto (a, b, 0, 0) \end{cases} \quad (2.173)$$

f est bijective, et nous vérifions aisément que pour tous complexes z_1, z_2 , nous avons:

$$\begin{aligned} f(z_1 + z_2) &= f(z_1) + f(z_2) \\ f(z_1 z_2) &= f(z_1) f(z_2) \end{aligned} \quad (2.174)$$

Donc f est un isomorphisme de $(\mathbb{C}, +, \times)$ sur $(\mathbb{H}', +, \times)$.

Cet isomorphisme a pour intérêt (provoqué) d'identifier $\mathbb{C}$ à $\mathbb{H}'$ et d'écrire $\mathbb{C} \subset \mathbb{H}$, les lois d'addition et de soustraction sur $\mathbb{H}$ prolongeant les opérations déjà connues sur $\mathbb{C}$.

Ainsi, par convention, nous écrirons tout élément de $(a, b, 0, 0)$ de $\mathbb{H}'$ sous la forme complexe $a + ib$. En particulier 0 est l'élément $(0, 0, 0, 0)$, 1 l'élément $(1, 0, 0, 0)$ et i l'élément $(0, 1, 0, 0)$.

Nous notons par analogie et par extension j l'élément $(0, 0, 1, 0)$ et k l'élément $(0, 0, 0, 1)$. La famille $\{1, i, j, k\}$ forme une base de l'ensemble des quaternions vu comme un espace vectoriel sur $\mathbb{R}$, et nous écrirons ainsi $a + bi + cj + dk$ le quaternion (a, b, c, d) .

La notation des quaternions sous forme définie ci-avant est parfaitement adaptée à l'opération de multiplication. Pour le produit de deux quaternions nous obtenons en développant l'expression:

$$(a + bi + cj + dk) \cdot (a' + b'i + c'j + d'k) \quad (2.175)$$

16 termes que nous devons identifier à la définition d'origine de la multiplication des quaternions pour obtenir les relations suivantes:

$$\begin{aligned}
 i \cdot j &= k = -j \cdot i \\
 j \cdot k &= i = -k \cdot j \\
 k \cdot i &= j = -i \cdot k \\
 i^2 &= j^2 = k^2 = -1
 \end{aligned} \quad (2.176)$$

Ce qui peut se résumer dans un tableau:

·	1	i	j	k
1	1	i	j	k
i	i	-1	k	-j
j	j	-k	-1	i
k	k	j	-i	-1

Tableau: 2.4 - Multiplication des composantes d'un quaternion

Nous pouvons constater que l'expression de la multiplication de deux quaternions ressemble en partie beaucoup à un produit vectoriel (noté $\times$ sur ce site) et scalaire (noté $\circ$ sur ce site):

$$\begin{pmatrix}
 aa' - bb' - cc' - dd' \\
 (ab' + ba') + (cd' - dc') \\
 (ac' + ca') - (bd' - db') \\
 (da' + ad') + (bc' - cb')
 \end{pmatrix} \quad (2.177)$$

Si ce n'est pas évident (ce qui serait tout à fait compréhensible), faisons un exemple concret.

Exemple:

Soient deux quaternions sans partie réelle:

$$p = xi + yj + zk \quad q = x'i + y'j + z'k \quad (2.178)$$

et $\vec{u}, \vec{v}$ les vecteurs de $\mathbb{R}^3$ de coordonnées respectives (x, y, z) et (x', y', z') . Alors le produit:

$$p \cdot q = (0, \vec{u}) \cdot (0, \vec{v}) \quad (2.179)$$

est:

$$p \cdot q = (-xx' - yy' - zz', yz' - zy', -xz' + zx', xy' - yx') = (-\vec{u} \circ \vec{v}, \vec{u} \times \vec{v})$$

Nous pouvons aussi par curiosité nous intéresser au cas général... Soient pour cela deux quaternions:

$$p = (a, \vec{u}) \quad q = (b, \vec{v}) \quad (2.180)$$

Nous avons alors:

$$\begin{aligned}
 p \cdot q &= (a + (0, \vec{u})) \cdot (b + (0, \vec{v})) = ab + (0, a\vec{v}) + (0, b\vec{u}) + (-\vec{u} \circ \vec{v}, \vec{u} \times \vec{v}) \\
 &= (ab - \vec{u} \circ \vec{v}, a\vec{v} + b\vec{u} + \vec{u} \times \vec{v})
 \end{aligned} \quad (2.181)$$

Définition: Le centre du corps non-commutatif $(\mathbb{H}, +, \times)$ est l'ensemble des éléments de $\mathbb{H}$ commutant pour la loi de multiplication avec tous les éléments de $\mathbb{H}$.

Nous allons montrer que le centre de $(\mathbb{H}, +, \times)$ est l'ensemble des réels!

Soit $\mathbb{H}_1$ le centre de $(\mathbb{H}, +, \times)$, et (x, y, z, t) un quaternion. Nous devons avoir les conditions suivantes qui soient satisfaites:

Soit $(x, y, z, t) \in \mathbb{H}_1$ alors pour tout $(a, b, c, d) \in \mathbb{H}$ nous cherchons:

$$(x, y, z, t) \cdot (a, b, c, d) = (a, b, c, d) \cdot (x, y, z, t) \quad (2.182)$$

ce qui donne en développant:

$$\begin{aligned} xa - yb - zc - td &= ax - by - cz - dt \\ xb + ya + zd - tc &= ay + bx + ct - dz \\ xc + za - yd + tb &= az + cx - bt + dy \\ ta + xd + yc - zb &= dx + at + bz - cy \end{aligned} \quad (2.183)$$

après simplification (la première ligne du système précédent est nulle des deux côtés de l'égalité):

$$\begin{aligned} ct - dz &= 0 \\ bt - dy &= 0 \\ bz - cy &= 0 \end{aligned} \quad (2.184)$$

la résolution de ce système, nous donne:

$$y = z = t = 0 \quad (2.185)$$

Donc pour que le quaternion (x, y, z, t) soit le centre de $\mathbb{H}$ il doit être réel (sans parties imaginaires)!

Au même titre que pour les nombres complexes, nous pouvons définir un conjugué des quaternions:

Définition: Le conjugué d'un quaternion $Z = (a, b, c, d)$ est le quaternion $\bar{Z} = (a, -b, -c, -d)$

Au même titre que pour les complexes, nous remarquons que:

1. D'abord de manière évidente que si $Z = \bar{Z}$ alors cela signifie que $Z \in \mathbb{R}$.

2. Que $Z + \bar{Z} \in \mathbb{R}$

3. Qu'en développant le produit $Z \cdot \bar{Z}$ nous avons:

$$\begin{aligned} Z \cdot \bar{Z} &= (a, b, c, d) \cdot (a, -b, -c, -d) \\ &= (a^2 + b^2 + c^2 + d^2, -ab + ba - cd + dc, -ac + ca + bd - db, da - ad - bc + cb) \\ &= (a^2 + b^2 + c^2 + d^2, 0, 0, 0) = a^2 + b^2 + c^2 + d^2 \in \mathbb{R} \end{aligned} \quad (2.186)$$

que nous adopterons, par analogie avec les nombres complexes, comme une définition de la norme (ou

module) des quaternions tel que:

$$|Z| = \sqrt{Z \cdot \bar{Z}} \quad (2.187)$$

Dès lors nous avons aussi immédiatement (relation qui nous sera utile plus tard):

$$|ZZ'| = \sqrt{(ZZ') \cdot (\overline{ZZ'})} \quad (2.188)$$

Comme pour les nombres complexes (voir plus loin), il est aisé de montrer que la conjugaison est un automorphisme du groupe $(\mathbb{H}, +)$.

Effectivement, soient $Z = (a, b, c, d)$ et $Z' = (a', b', c', d')$ alors:

$$\begin{aligned} \overline{Z + Z'} &= (a + a', -b - b', -c - c', -d - d') = (a, -b, -c, -d) + (a', -b', -c', -d') \\ &= \bar{Z} + \bar{Z}' \end{aligned} \quad (2.189)$$

Il est aussi aisé de montrer qu'elle est involutive. Effectivement:

$$\overline{\bar{Z}} = (a, -(-b), -(-c), -(-d)) = (a, b, c, d) = Z \quad (2.190)$$

La conjugaison n'est par contre pas un automorphisme multiplicatif du corps $(\mathbb{H}, +, \times)$. En effet, si nous considérons la multiplication de Z, Z' et en prenons le conjugué:

$$ZZ' = \begin{pmatrix} aa' - bb' - cc' - dd' \\ (ab' + ba') + (cd' - dc') \\ (ac' + ca') - (bd' - db') \\ (da' + ad') + (bc' - cb') \end{pmatrix}, \quad \overline{ZZ'} = \begin{pmatrix} aa' - bb' - cc' - dd' \\ -(ab' + ba') - (cd' - dc') \\ -(ac' + ca') + (bd' - db') \\ -(da' + ad') - (bc' - cb') \end{pmatrix} \quad (2.191)$$

nous voyons immédiatement (ne serait-ce que pour la deuxième ligne) que nous avons:

$$\overline{\overline{ZZ'}} \neq \overline{ZZ'} \quad (2.192)$$

Revenons maintenant sur notre norme (ou module).... Pour cela, calculons le carré de la norme de $|ZZ'|$:

$$|ZZ'|^2 = (ZZ') \cdot (\overline{ZZ'}) \quad (2.193)$$

Nous savons (par définition) que:

$$ZZ' = \begin{pmatrix} aa' - bb' - cc' - dd' \\ (ab' + ba') + (cd' - dc') \\ (ac' + ca') - (bd' - db') \\ (da' + ad') + (bc' - cb') \end{pmatrix} \quad (2.194)$$

notons ce produit de manière telle que:

$$Z \cdot Z' = (\alpha, \beta, \gamma, \lambda) = Z'' \quad (2.195)$$

Nous avons alors:

$$Z'', \overline{Z''} = \alpha^2 + \beta^2 + \gamma^2 + \lambda^2 \quad (2.196)$$

en substituant il vient:

$$\begin{aligned} (ZZ') \cdot \overline{(ZZ')} &= (aa' - bb' - cc' - dd')^2 + (ab' + ba' + cd' - dc')^2 \\ &+ (ac' + ca' - bd' + db')^2 + (da' + ad' + bc' - cb')^2 \end{aligned} \quad (2.197)$$

après un développement algébrique élémentaire (honnêtement ennuyeux), nous trouvons:

$$(ZZ') \cdot \overline{(ZZ')} = (\alpha'^2 + \beta'^2 + \gamma'^2 + \lambda'^2)(\alpha^2 + \beta^2 + \gamma^2 + \lambda^2) = |Z'|^2 |Z|^2 \quad (2.198)$$

Donc:

$$(ZZ') \cdot \overline{(ZZ')} = |Z|^2 |Z'|^2 = |ZZ'|^2 \quad (2.199)$$

Remarque: La norme est donc un homomorphisme de $(\mathbb{H}, \times)$ dans $(\mathbb{R}, \times)$. Par la suite, nous noterons G l'ensemble des quaternions de norme 1.

2.8.1. INTERPRETATION MATRICIELLE

Soient q et p deux quaternions donnés, soit l'application:

$$p \rightarrow q \cdot p$$

La multiplication (à gauche) peut être faite avec une application linéaire (cf. chapitre d'Algèbre Linéaire) sur $\mathbb{H}$.

Si q s'écrit:

$$a + bi + cj + dk \quad (2.200)$$

cette application a pour matrice, dans la base $1, i, j, k$:

$$\begin{bmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{bmatrix} \quad (2.201)$$

Ce que nous vérifions bien:

$$ZZ' = \begin{bmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{bmatrix} \begin{bmatrix} a' \\ b' \\ c' \\ d' \end{bmatrix} = \begin{pmatrix} aa' - bb' - cc' - dd' \\ (ab' + ba') + (cd' - dc') \\ (ac' + ca') - (bd' - db') \\ (da' + ad') + (bc' - cb') \end{pmatrix} \quad (2.202)$$

En fait, nous pouvons alors définir les quaternions comme l'ensemble des matrices ayant la structure visible ci-dessus si nous le voulions. Cela les réduirait alors à un sous espace vectoriel de $M_4(\mathbb{R})$.

En particulier, la matrice de 1 (la partie réelle du quaternion q) n'est alors rien d'autre que la matrice identité:

$$M_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \mathbb{1} \quad (2.203)$$

de même:

$$M_i = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, M_j = \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}, M_k = \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad (2.204)$$

2.8.2. ROTATIONS

Nous allons maintenant voir que la conjugaison par un élément du groupe G des quaternions de norme unité peut s'interpréter comme une rotation pure dans l'espace!

Définition: La "**conjugaison**" par un quaternion q non nul et de norme unité est l'application S_q définie sur $\mathbb{H}$ par:

$$S_q : p \rightarrow q \cdot p \cdot q^{-1} = q \cdot p \cdot \bar{q} \quad (2.205)$$

et nous affirmons que cette application est une rotation.

Remarques:

R1. Comme q est de norme 1, nous avons bien évidemment $|q| = q \cdot \bar{q} = 1$ donc $q^{-1} = \bar{q}$. Ce quaternion peut être vu comme la valeur propre (unitaire) de l'application (matricielle) p sur le vecteur $\bar{q}$ (on se retrouve avec un concept en tout point similaire aux matrices orthogonales de rotation vues en algèbre linéaire).

R2. S_q est une application linéaire (donc si c'est bien une rotation, la rotation peut être décomposée en plusieurs rotations). Effectivement, prenons deux quaternions p_1, p_2 et λ_1, λ_2 des réels, alors nous avons:

$$S_q(\lambda_1 p_1 + \lambda_2 p_2) = q(\lambda_1 p_1 + \lambda_2 p_2)\bar{q} = \lambda_1 q p_1 \bar{q} + \lambda_2 q p_2 \bar{q} = \lambda_1 S_q(p_1) + \lambda_2 S_q(p_2) \quad (2.206)$$

Vérifions maintenant que l'application est bien une rotation pure. Comme nous l'avons vu lors de notre étude de l'algèbre linéaire et en particulier des matrices orthogonales (cf. chapitre d'Algèbre Linéaire), une première condition est que l'application conserve la norme.

Vérifions:

$$|S_q(p)| = |qp\bar{q}| = |q||p||\bar{q}| = |p| \quad (2.207)$$

Par ailleurs, nous pouvons vérifier qu'une rotation d'un quaternion purement complexe (tel qu'alors nous nous restreignons à $\mathbb{R}^3$) et la même rotation inverse sommées est nulle (le vecteur sommé à son opposé s'annulent):

$$S_q(p) + \overline{S_q(p)} = qp\bar{q} + \overline{qp\bar{q}} = qp\bar{q} + \overline{q(p\bar{q})} \quad (2.208)$$

nous vérifions trivialement que si nous avons deux quaternions q, p alors $\overline{p \cdot q} = \bar{q} \cdot \bar{p}$ dès lors:

$$\begin{aligned} S_q(p) + \overline{S_q(p)} &= qp\bar{q} + \overline{q(p\bar{q})} = qp\bar{q} + (\overline{p\bar{q}})\bar{q} = q \cdot p \cdot \bar{q} + \bar{\bar{q}} \cdot \bar{p} \cdot \bar{q} \\ &= q \cdot p \cdot \bar{q} + q \cdot \bar{p} \cdot \bar{q} = S_q(p + \bar{p}) \end{aligned} \quad (2.209)$$

pour que cette opération soit nulle, nous voyons immédiatement que nous devons restreindre p aux quaternions purement complexes. Dès lors:

$$S_q(p + \bar{p}) = S_q(0) = 0 \quad (2.210)$$

Nous en déduisons alors que p doit être purement complexe pour que l'application S_q soit une rotation et que $S_q(p)$ est un quaternion pur. En d'autres termes, cette application est stable (en d'autres termes: un quaternion pur par cette application reste un quaternion pur).

S_q restreint à l'ensemble des quaternions purement complexes est donc une isométrie vectorielle, c'est-à-dire une symétrie ou une rotation.

Nous avons vu également lors de notre étude des matrices de rotation dans le chapitre d'Algèbre

Linéaire que l'application A devait être de déterminant 1 pour que nous ayons une rotation. Voyons si c'est le cas de S_q :

Pour cela, nous calculons explicitement en fonction de:

$$q = a + bi + cj + dk \quad (2.211)$$

la matrice (dans la base canonique $\{i, j, k\}$) de S_q et nous en calculons le déterminant. Ainsi, nous obtenons les coefficients des colonnes de A en se rappelant que:

$$\begin{aligned} i \cdot j &= k = -j \cdot i \\ j \cdot k &= i = -k \cdot j \\ k \cdot i &= j = -i \cdot k \\ i^2 &= j^2 = k^2 = -1 \end{aligned} \quad (2.212)$$

et ensuite en calculant:

$$\begin{aligned}
S_q(i) &= (a + bi + cj + dk)i(a - bi - cj - dk) \\
&= \left(ai + b(i^2) + c(ji) + d(ki) \right)(a - bi - cj - dk) = (ai - b - ck + dj)(a - bi - cj - dk) \\
&= (a^2i + ab - ack + adj) - (ba - b^2i - bcj - bdk) \\
&\quad - (cak - cbj + c^2i + cd) + (daj + dbk + cd - d^2i) \\
&= \left(a^2 + b^2 - c^2 - d^2 \right)i + (\cancel{ab} - ack + adj) - (\cancel{ba} - bcj - bdk) \\
&\quad - (cak - cbj + \cancel{cd}) + (daj + dbk + \cancel{cd}) \\
&= \left(a^2 + b^2 - c^2 - d^2 \right)i + 2(ad + bc)j + 2(bd - ac)k
\end{aligned}$$

$$\begin{aligned}
S_q(j) &= (a + bi + cj + dk)j(a - bi - cj - dk) \\
&= \left(aj + b(ij) + c(j^2) + d(kj) \right)(a - bi - cj - dk) = (aj + bk - c - di)(a - bi - cj - dk) \\
&= (a^2j + abk + ac - adi) + (bak - b^2j + bci + bd) \\
&\quad - (ca - cbi - c^2j - cdk) - (dai + db - dck + d^2j) \\
&= \left(a^2 - b^2 + c^2 - d^2 \right)j + (abk + \cancel{ac} - adi) + (bak + bci + \cancel{bd}) \\
&\quad - (\cancel{ca} - cbi - cdk) - (dai + \cancel{db} - dck) \\
&= 2(bc - ad)i + \left(a^2 - b^2 + c^2 - d^2 \right)j + 2(ab + cd)k
\end{aligned} \tag{2.213}$$

$$\begin{aligned}
S_q(k) &= (a + bi + cj + dk)k(a - bi - cj - dk) \\
&= \left(ak + b(ik) + c(jk) + d(k^2) \right)(a - bi - cj - dk) = (ak - bj + ci - d)(a - bi - cj - dk) \\
&= (a^2k - abj + aci + ad) - (abj + b^2k + bc - bdi) \\
&\quad + (cai + cb - c^2k + cdj) - (da - dbi - dcj - d^2k) \\
&= \left(a^2 - b^2 - c^2 + d^2 \right)k + (-abj + aci + \cancel{ad}) - (abj + \cancel{bc} - bdi) \\
&\quad + (cai + \cancel{cb} + cdj) - (\cancel{da} - dbi - dcj) \\
&= 2(ac + bd)i + 2(cd - ab)j + \left(a^2 - b^2 - c^2 + d^2 \right)k
\end{aligned}$$

Il faut alors calculer le déterminant de la matrice (pfff...):

$$\begin{bmatrix}
a^2 + b^2 - c^2 - d^2 & 2(ad + bc) & 2(bd - ac) \\
2(bc - ad) & a^2 - b^2 + c^2 - d^2 & 2(ab + cd) \\
2(ac + bd) & 2(cd - ab) & a^2 - b^2 - c^2 + d^2
\end{bmatrix} \tag{2.214}$$

en se souvenant que (ce qui permet aussi de simplifier l'expression des termes de la diagonale comme nous pouvons le voir dans certains ouvrages):

$$a^2 + b^2 + c^2 + d^2 = 1 \tag{2.215}$$

nous trouvons que le déterminant vaut bien 1. Sinon, nous pouvons vérifier cela avec Maple:

```

>with(linalg):
> A:=linalg[matrix](3,3,[a^2+b^2-c^2-d^2,2*(a*d+b*c),2*(b*d-a*c),a^2-b^2+c^2-d^2,2*
(a*b+c*d),2*(a*c+b*d),2*(c*d-a*b),a^2-b^2-c^2+d^2]);
> factor(det(A));

```

Montrons maintenant que cette rotation est un demi-tour d'axe (l'exemple qui peut sembler particulier est général!):

D'abord, si:

$$q = xi + yj + zk \quad (2.216)$$

nous avons:

$$S_q(q) = q \cdot q \cdot \bar{q} = q \quad (2.217)$$

ce qui signifie que l'axe de rotation (x, y, z) est fixé par l'application S_q elle-même !

D'autre part, nous avons vu que si q est un quaternion purement complexe de norme 1 alors:

$$q^{-1} = \bar{q} \text{ et } \bar{\bar{q}} = -q \quad (2.218)$$

Ce qui nous donne la relation:

$$q^2 = q \cdot (-\bar{q}) = q \cdot (-(q^{-1})) = -1 \quad (2.219)$$

Ce résultat nous amène à calculer la rotation d'une rotation:

$$\begin{aligned} S_q(S_q(p)) &= q(qp\bar{q})\bar{q} = q^2 p \bar{q}^2 = S_{q^2} = (-1)p(-q)^2 \\ &= (-1)pq^2 = (-1)p(-1) = p \end{aligned} \quad (2.220)$$

Conclusion: Puisque la rotation d'une rotation est un tour complet, alors S_q est nécessairement un demi-tour :

$$S_q(p) = -p$$

par rapport (!) à l'axe (x, y, z) .

A ce stade, nous pouvons affirmer que toute rotation de l'espace peut se représenter par S_q (la conjugaison par un quaternion q de norme 1). En effet, les demi-tours engendrent le groupe des rotations, c'est-à-dire que toute rotation peut s'exprimer comme le produit d'un nombre fini de demi-tours, et donc comme la conjugaison par un produit de quaternions de norme 1 (produit qui est lui-même un quaternion de norme 1 ...).

Nous allons tout de même donner une forme explicite reliant une rotation et le quaternion qui la représente, au même titre que nous l'avons fait pour les nombres complexes.

Soit $\vec{u}(x, y, z)$ un vecteur unitaire et $\theta \in [0, 2\pi]$ un angle. Alors nous affirmons que la rotation d'axe $\vec{u}$

et d'angle θ correspond à l'application S_q , où q est le quaternion:

$$q = \cos \frac{\theta}{2} + x \sin \frac{\theta}{2} i + y \sin \frac{\theta}{2} j + z \sin \frac{\theta}{2} k = \left(\cos \frac{\theta}{2}, \sin \frac{\theta}{2} \vec{u} \right) = \begin{pmatrix} \cos \frac{\theta}{2} \\ x \sin \frac{\theta}{2} \\ y \sin \frac{\theta}{2} \\ z \sin \frac{\theta}{2} \end{pmatrix} \quad (2.221)$$

Pour que cette affirmation soit vérifiée, nous savons qu'il faut que: la norme de q soit unitaire, le déterminant de l'application S_q soit égal à l'unité, que l'application S_q conserve la norme, que l'application S_q renvoie tout vecteur colinéaire à l'axe de rotation sur l'axe de rotation.

1. La norme du quaternion proposé précédemment vaut effectivement 1:

$$\begin{aligned} |q| &= \cos^2 \frac{\theta}{2} + x^2 \sin^2 \frac{\theta}{2} + y^2 \sin^2 \frac{\theta}{2} + z^2 \sin^2 \frac{\theta}{2} = \cos^2 \frac{\theta}{2} + \sin^2 \frac{\theta}{2} (x^2 + y^2 + z^2) \\ &= \cos^2 \frac{\theta}{2} + \sin^2 \frac{\theta}{2} (x^2 + y^2 + z^2) \end{aligned} \quad (2.222)$$

et comme $\vec{u}(x,y,z)$ est unitaire alors nous avons:

$$x^2 + y^2 + z^2 = 1 \quad (2.223)$$

Donc:

$$|q| = \cos^2 \frac{\theta}{2} + \sin^2 \frac{\theta}{2} (x^2 + y^2 + z^2) = \cos^2 \frac{\theta}{2} + \sin^2 \frac{\theta}{2} = 1 \quad (2.224)$$

2. Le fait que q soit un quaternion de norme 1 amène immédiatement à ce que le déterminant de l'application S_q soit unitaire. Nous l'avons déjà montré plus haut dans le cas général de n'importe quel quaternion de norme 1 (condition nécessaire et suffisante).

3. Il en est de même pour la conservation de la norme. Nous avons déjà montré plus haut que c'était de toute façon le cas dès que le quaternion q était de norme 1 (condition nécessaire et suffisante).

4. Voyons maintenant que tout vecteur colinéaire à l'axe de rotation est projeté sur l'axe de rotation. Notons q' le quaternion purement imaginaire et unitaire $xi + yj + zk$. Nous avons alors:

$$q = \cos \frac{\theta}{2} + \sin \frac{\theta}{2} q' \quad (2.225)$$

Alors:

$$S_q(q') = qq'\bar{q} \quad (2.226)$$

mais comme q' est la restriction de q à ces éléments purs qui le constituent, cela revient à écrire:

$$S_q(q') = S_q(q) = qq\bar{q} = q \quad (2.227)$$

Montrons maintenant le choix de l'écriture $\theta/2$. Si $\vec{v} = (x_1, y_1, z_1)$ désigne un vecteur unitaire orthogonal à $\vec{u}$ (perpendiculaire à l'axe de rotation donc), et p le quaternion $x_1i + y_1j + z_1k$ alors nous avons:

$$\begin{aligned} S_q(p) &= \left(\cos \frac{\theta}{2} + \sin \frac{\theta}{2} q' \right) p \left(\cos \frac{\theta}{2} - \sin \frac{\theta}{2} q' \right) \\ &= \cos^2 \frac{\theta}{2} p + \cos \frac{\theta}{2} \sin \frac{\theta}{2} (q' p - p q') - \sin^2 \frac{\theta}{2} q' p q' \end{aligned} \quad (2.228)$$

Nous avons montré lors de la définition de la multiplication de deux quaternions que:

$$p q' = -q' p \quad (2.229)$$

nous obtenons alors:

$$\begin{aligned} S_q(p) &= \cos^2 \frac{\theta}{2} p + 2 \cos \frac{\theta}{2} \sin \frac{\theta}{2} q' p - \sin^2 \frac{\theta}{2} q' p q' \\ &= \cos^2 \frac{\theta}{2} p + 2 \cos \frac{\theta}{2} \sin \frac{\theta}{2} q' p + \sin^2 \frac{\theta}{2} q' p (-q') \\ &= \cos^2 \frac{\theta}{2} p + 2 \cos \frac{\theta}{2} \sin \frac{\theta}{2} q' p + \sin^2 \frac{\theta}{2} q' p \bar{q}' \end{aligned} \quad (2.230)$$

Nous avons également montré plus haut que:

$$S_q(p) = -p = qp\bar{q} \quad (2.231)$$

dès lors:

$$q' p \bar{q}' = S_{q'}(p) = -p \quad (2.232)$$

(le demi-tour d'axe (x, y, z)). Donc:

$$\begin{aligned} S_{q'}(p) &= \cos^2 \frac{\theta}{2} p + 2 \cos \frac{\theta}{2} \sin \frac{\theta}{2} q' p - \sin^2 \frac{\theta}{2} p \\ &= \left(\cos^2 \frac{\theta}{2} - \sin^2 \frac{\theta}{2} \right) p + 2 \cos \frac{\theta}{2} \sin \frac{\theta}{2} q' p \\ &= \cos \theta p + \sin \theta q' p \end{aligned} \quad (2.233)$$

Remarque: Nous commençons à entrevoir ici déjà l'utilité d'avoir écrit dès le début $\theta/2$ pour l'angle!

Nous savons que p est le quaternion pur assimilé à un vecteur unitaire $\vec{v}$ orthogonal à l'axe de rotation $\vec{u}$, lui-même assimilé à la partie purement imaginaire de q' . Nous remarquons alors de suite que la partie imaginaire du produit (défini!) des quaternions $q'p$ est alors égal au produit vectoriel $\vec{u} \times \vec{v} = \vec{w}$. Ce produit vectoriel engendre donc un vecteur perpendiculaire à $\vec{u}, \vec{v}$ et donc $(\vec{u}, \vec{v}, \vec{w})$.

Le couple $(\vec{v}, \vec{w})$ forme donc un plan perpendiculaire à l'axe de rotation $\vec{u}$ (c'est comme pour les nombres complexes simples dans lequel nous avons le plan de Gauss et perpendiculairement à celui-ci un axe de rotation!).

Alors finalement:

$$S_q(p) = \cos \theta p + \sin \theta q' p = \cos \theta \vec{v} + \sin \theta \vec{w} \quad (2.234)$$

Nous nous retrouvons avec une rotation basée sur un plan (mais qui a donc lieu dans l'espace!) identique à celle présentée plus haut avec les nombres complexes standards dans le plan de Gauss.

Nous savons donc maintenant comment faire n'importe quel type de rotation dans l'espace en une seule opération mathématique et ce en plus par rapport à un libre choix de l'axe !

Nous pouvons aussi maintenant mieux comprendre pourquoi l'algèbre des quaternions n'est pas commutative. Effectivement, les rotations vectorielles du plan sont commutatives mais celles de l'espace ne le sont pas comme nous le montre l'exemple ci-dessous :

Soit la configuration initiale :

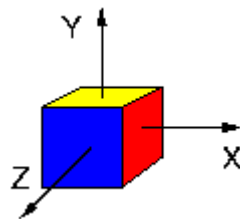


Figure: 2.8 - Situation initiale pour rotations quaternions

Alors une rotation autour de l'axe X suivie d'une rotation autour de l'axe Y :

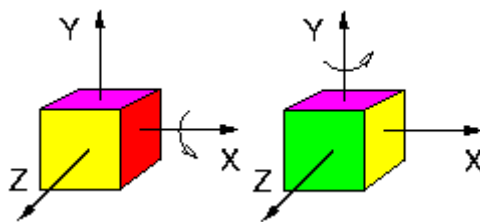


Figure: 2.9 - Exemple de rotation de quaternions

n'est pas égale à une rotation autour de l'axe Y suivie d'une rotation autour de l'axe X :

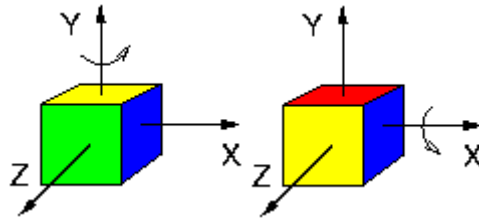


Figure: 2.10 - Exemple de non équivalence pour rotation quaternions

Les résultats obtenus seront fondamentaux pour notre compréhension des spineurs (cf. chapitre de Calcul Spinoriel)!

2.9. NOMBRES ALGÈBRIQUES ET TRANSCENDANTS

Définitions:

D1. Nous appelons "**nombre entier algébrique de degré n** ", tout nombre qui est solution d'une équation algébrique de degré n , à savoir: un polynôme de degré n (concept que nous aborderons dans la section d'Algèbre) dont les coefficients sont des entiers relatifs et dont le coefficient dominant vaut 1.

D2. Nous appelons "**nombre algébrique de degré n** ", tout nombre qui est solution d'une équation algébrique de degré n , à savoir: un polynôme de degré n dont les coefficients sont des rationnels.

Un premier résultat intéressant et particulier dans ce domaine d'étude (curiosité mathématique...) est qu'un nombre rationnel est un "nombre entier algébrique de degré n " si et seulement si c'est un entier relatif (lisez plusieurs fois au besoin...). En termes savants, nous disons alors que l'anneau $\mathbb{Z}$ est "**intégralement clos**".

Démonstration:

Nous supposons que le nombre p/q , où p et q sont deux entiers premiers entre eux (c'est-à-dire dont le rapport ne donne pas un entier ou plus rigoureusement... que le plus grand commun diviseur est 1!), est une racine du polynôme (cf. chapitre de Calcul Algébrique) suivant à coefficients entiers relatifs et dont le coefficient dominant est unitaire:

$$x^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0 \quad (2.235)$$

où l'égalité avec zéro du polynôme est implicite.

Dans ce cas:

$$p^n = -(a_{n-1}p^{n-1} + \dots + a_1pq^{n-2} + a_0q^{n-1})q \quad (2.236)$$

Puisque les coefficients sont par définition tous entiers et leurs multiples aussi dans la parenthèse, alors la parenthèse a nécessairement une valeur dans $\mathbb{Z}$.

Ainsi, q (à droite de la parenthèse) divise une puissance de p (à gauche de l'égalité), ce qui n'est possible, dans l'ensemble $\mathbb{Z}$ (car notre parenthèse a une valeur dans cet ensemble pour rappel...), que

si q vaut ± 1 (puisqu'ils étaient premiers entre eux).

Donc parmi tous les nombres rationnels, les seuls qui sont solutions d'équations polynômiales à coefficients entiers relatifs et dont le coefficient dominant est unitaire sont des entiers relatifs!

□C.Q.F.D.

Pour prendre un autre cas intéressant et particulier, il est facile de montrer qu'absolument tout nombre rationnel est un "nombre algébrique". Effectivement, si nous prenons le plus simple polynôme suivant:

$$qx - p = 0 \quad (2.237)$$

où q et p sont premiers entre eux et où q est différent de 1. Alors comme il s'agit d'une polynôme à coefficients rationnels simple, après remaniement nous avons:

$$x = \frac{p}{q} \quad (2.238)$$

Donc puisque q et p sont premiers entre eux et que q est différent de l'unité, nous avons bien que tout nombre rationnel est un "nombre algébrique de degré 1".

Ainsi, la quantité de nombres rationnels "algébriques" est plus grande que le nombre de rationnels qui sont des "entiers algébriques".

Nous avons aussi le nombre réel (et irrationnel) $\sqrt{2}$ qui est un "nombre entier algébrique de degré 2", car il est racine de:

$$x^2 - 2 = 0 \quad (2.239)$$

et le nombre complexe i qui est aussi un "nombre entier algébrique de degré 2", car il est racine de l'équation:

$$x^2 + 1 = 0 \quad (2.240)$$

etc...

Définition: Les nombres qui ne sont pas algébriques (entiers ou non!) sont transcendants.

L'ensemble de tous les nombres transcendants est non dénombrable. La preuve est simple et ne nécessite aucun développement mathématique difficile.

Effectivement, puisque les polynômes à coefficients entiers sont dénombrables, et puisque chacun de ces polynômes possède un nombre fini de zéros (voir le théorème de factorisation dans le chapitre de Calcul Algébrique), l'ensemble des nombres algébriques est dénombrable! Mais l'argument de la diagonale de Cantor (cf. chapitre de Théorie des Ensembles) établit que les nombres réels (et par conséquent les nombres complexes aussi) sont non dénombrables, donc l'ensemble de tous les nombres transcendants doit être non dénombrable.

En d'autres termes, il y a beaucoup plus de nombres transcendants que de nombres algébriques.

Les transcendants les plus connus sont π et e . Les démonstrations de leur transcendance est en cours de rédaction. Nous devrions pouvoir vous les fournir fin 2014.

2.10. NOMBRES ABSTRAITS

Le nombre peut être envisagé en faisant abstraction de la nature des objets qui constituent le groupement qu'il caractérise et ainsi qu'à la façon de codifier (chiffre arabe, romain, ou autre système universel). Nous disons alors que le nombre est "**abstrait**".

Remarque: Arbitrairement, l'être humain a adopté un système numérique majoritairement utilisé de par le monde et représenté par les symboles 1, 2, 3, 4, 5, 7, 8, 9 du système décimal et qui seront supposés connus aussi bien en écriture qu'oralement par le lecteur (apprentissage du langage).

Pour les mathématiciens, il n'est pas avantageux de travailler avec ces symboles car ils représentent uniquement des cas particuliers. Ce que cherchent les physiciens théoriciens ainsi que les mathématiciens, ce sont des "**relations littérales**" applicables dans un cas général et que les ingénieurs puissent en fonction de leurs besoins changer ces nombres abstraits par les valeurs numériques qui correspondent au problème qu'ils ont besoin de résoudre.

Ces nombres abstraits appelés aujourd'hui communément "**variables**" ou "**inconnues**", utilisées dans le cadre du "**calcul littéral**", sont très souvent représentés par:

1. L'alphabet latin:

$$a, b, c, d, e...x, y, z ; A, B, C, D, E... \quad (2.241)$$

où Les lettres minuscules du début l'alphabet latin ($a, b, c, d, e...$) sont souvent utilisées pour représenter de manière abstraite des constantes, alors que les lettres minuscules de la fin de l'alphabet latin ($...x, y, z$) sont utilisées pour représenter des entités (variables ou inconnues) dont nous recherchons la valeur.

2. L'alphabet grec:

A α	Alpha	$\Lambda \lambda$	Lambda
B β	Beta	M μ	Mu
$\Gamma \gamma$	Gamma	N ν	Nu
$\Delta \delta$	Delta	$\Xi \xi$	Xi
E ε	Epsilon	O $\omicron$	Omicron
Z ζ	Zeta	$\Pi \pi$	Pi
H η	Eta	P ρ	Rho
$\Theta \theta$	Thêta	$\Sigma \sigma$	Sigma
I ι	Iota	T τ	Tau
K κ	Kappa	Y υ	Upsilon
$\Phi \phi$	Phi	X χ	Chi
$\Psi \psi$	Psi	$\Omega \omega$	Omega

Tableau: 2.5 - Alphabet Grec

qui est particulièrement utilisé pour représenter soit des opérateurs mathématiques plus ou moins complexes (comme la somme indexée $\sum$, le variationnel δ , l'élément infinitésimal ε , le différentiel partiel ∂ , etc.) soit des variables dans le domaine de la physique (comme ω pour la pulsation, la fréquence ν , la densité ρ , etc.).

3. L'alphabet hébraïque (à moindre mesure)

Remarque: Comme nous l'avons vu, les nombres transfinis sont par exemples donnés par la lettre $\aleph_0$ "aleph".

Bien que ces symboles puissent représenter n'importe quel nombre il en existe quelques-uns qui peuvent représenter en physique des valeurs dites "constantes Universelles" comme la vitesse de la lumière c , la constante gravitationnelle G , la constante de Planck h , etc.

Nous utilisons très souvent encore d'autres symboles que nous introduirons et définirons au fur et à mesure.

Remarque: Les lettres pour représenter les nombres auraient été employées pour la première fois par Viète au 16ème siècle.

2.11. DOMAINES DE DÉFINITION

Une variable est un nombre abstrait susceptible de prendre des valeurs numériques différentes. L'ensemble de ces valeurs peut varier suivant le caractère du problème considéré.

Définitions:

D1. Nous appelons "**domaine de définition**" d'une variable, l'ensemble des valeurs numériques qu'elle est susceptible de prendre entre deux valeurs finies ou infinies appelées "**bornes**".

Soient a et b deux nombres tel que $a < b$. Alors:

D2. Nous appelons "**intervalle fermé d'extrémités a et b** ", l'ensemble de tous les nombres x compris entre ces deux valeurs incluses et nous le désignons de la façon suivante:

$$[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\} \quad (2.242)$$

D3. Nous appelons "**intervalle ouvert d'extrémités a et b** ", l'ensemble de tous les nombres x compris entre ces deux valeurs non incluses et nous le désignons de la façon suivante:

$$]a, b[= \{x \in \mathbb{R} \mid a < x < b\} \quad (2.243)$$

D4. Nous appelons "**intervalle fermé à gauche, ouvert à droite**" l'ensemble suivant:

$$[a, b[= \{x \in \mathbb{R} \mid a \leq x < b\} \quad (2.244)$$

D5. Nous appelons "**intervalle ouvert à gauche, fermé à droite**" l'ensemble suivant:

$$]a, b] = \{x \in \mathbb{R} \mid a < x \leq b\} \quad (2.245)$$

Soit sous forme résumée et imagée:

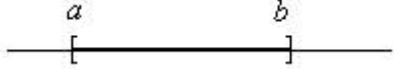
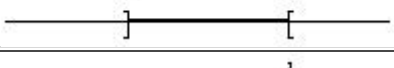
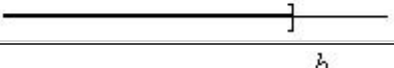
$[a, b]$		$a \leq x \leq b$	Intervalle fermé borné
$[a, b[$		$a \leq x < b$	Intervalle borné semi-fermé en a et semi-ouvert en b (ou semi-fermé à gauche et semi-ouvert à droite)
$]a, b]$		$a < x \leq b$	Intervalle borné semi-ouvert en a et semi-fermé en b (ou semi-ouvert à gauche et semi-fermé à droite)
$]a, b[$		$a < x < b$	Intervalle ouvert borné.
$] -\infty, b]$		$x \leq b$	Intervalle non borné fermé en b (ou fermé à droite)
$] -\infty, b[$		$x < b$	Intervalle non borné ouvert en b (ou ouvert à droite)
$[a, +\infty[$		$a \leq x$	Intervalle non borné fermé en a (ou fermé à gauche)
$]a, +\infty[$		$a < x$	Intervalle non borné ouvert en a (ou ouvert à gauche)

Tableau: 2.6 - Types d'intervalles et de bornes

Remarques:

R1. La notation $\{x \text{ tels que } a < x < b\}$ désigne l'ensemble des réels x strictement plus grands que a et strictement inférieurs à b .

R2. Le fait de dire qu'un intervalle est par exemple ouvert en b signifie que le réel b ne fait pas partie de celui-ci. Par contre, s'il avait été fermé alors b en aurait fait partie.

R3. Si la variable peut prendre toutes les valeurs négatives et positives possibles nous écrivons dès lors: $] -\infty, +\infty[$ où le symbole " ∞ " signifie une "infinité". Evidemment il peut y avoir des combinaisons d'intervalles ouverts et infinis à droite, fermé et limité gauche et réciproquement.

R4. Nous rappellerons ces concepts avec une autre approche lorsque nous étudierons l'algèbre (calcul littéral).

Nous disons que la variable x est "**ordonnée**" si en représentant son domaine de définition par un axe horizontal où chaque point de l'axe représente une valeur de x , alors pour chaque couple de valeurs, nous pouvons indiquer celle qui est "**antécédente**" (qui précède) et celle qui est "**conséquente**" (qui suit). Ici la notion d'antécédente ou de conséquente n'est pas liée au temps, elle exprime juste la façon d'ordonner les valeurs de la variable.

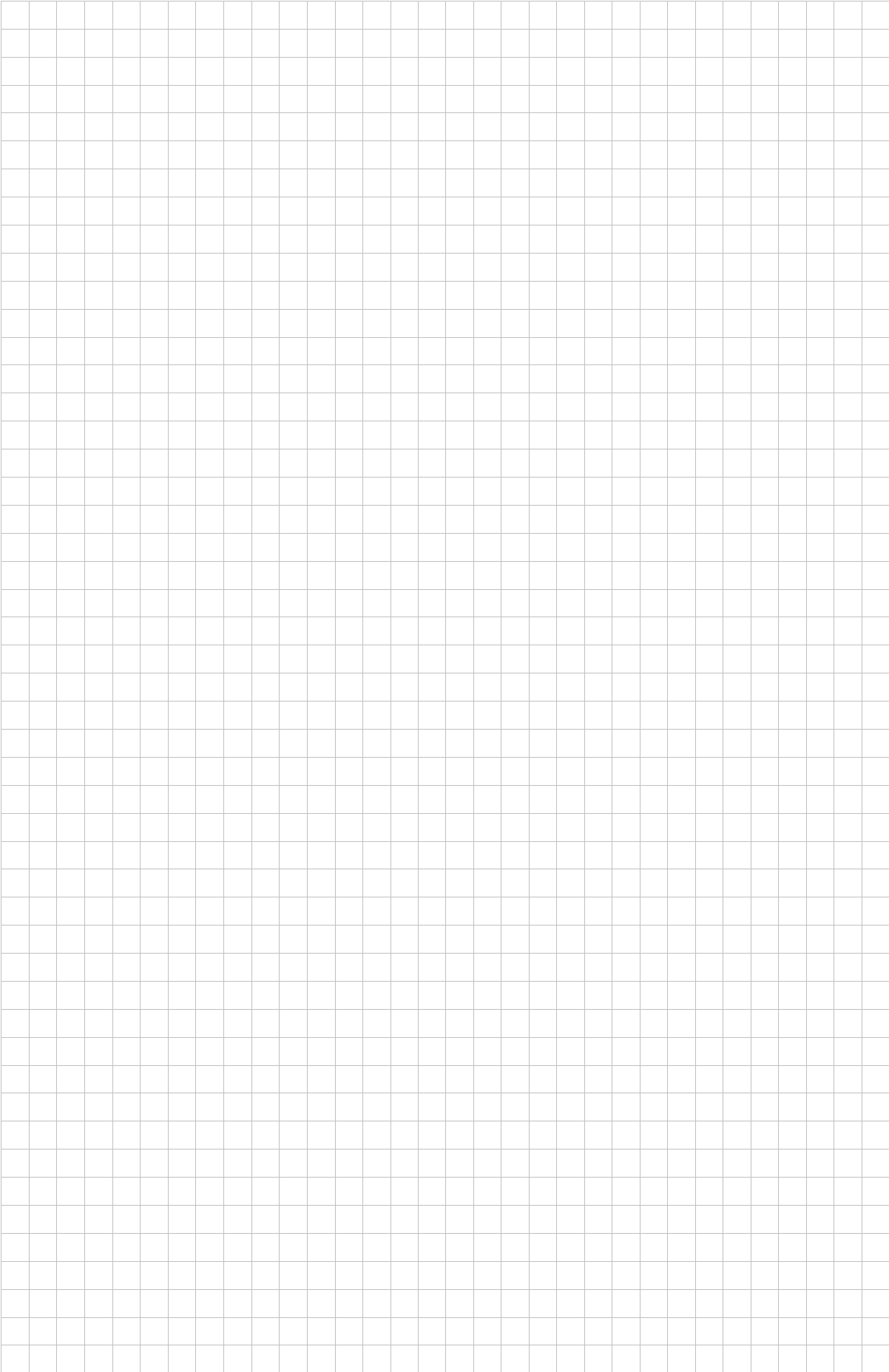
Définitions:

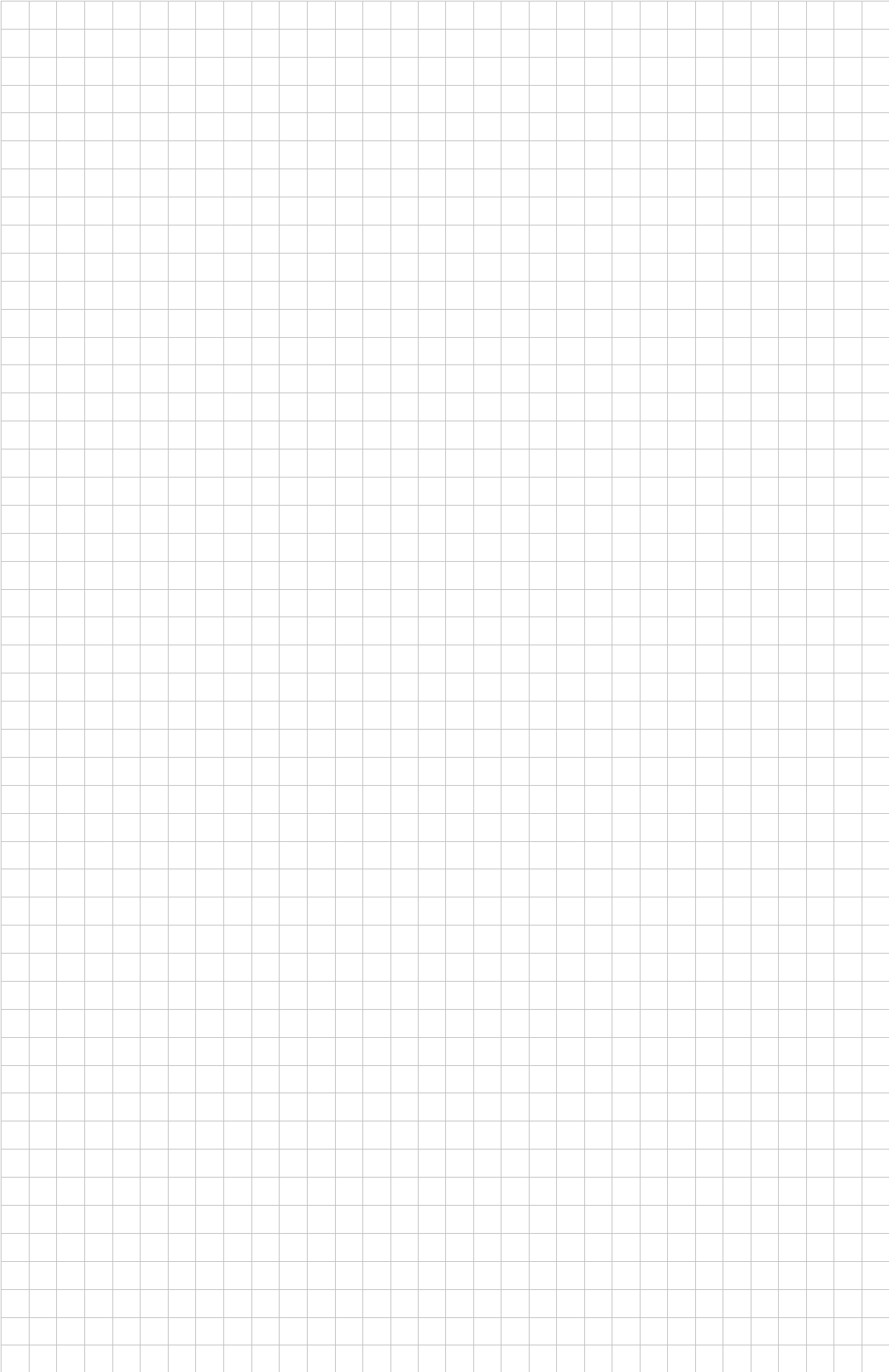
D1. Une variable est dite "**croissante**" si chaque valeur conséquente est plus grande que chaque valeur antécédente.

D2. Une variable est dite "**décroissante**" si chaque valeur conséquente est plus petite que chaque valeur antécédente.

D3. Les variables croissantes et les variables décroissantes sont appelées "**variables à variations monotones**" ou simplement "**variables monotones**".







3. OPÉRATEURS ARITHMÉTIQUES

Parler des nombres comme nous l'avons fait dans le chapitre précédent amène naturellement à considérer les opérations de calculs. Il est donc logique que nous fassions une description non exhaustive des opérations qui peuvent exister entre les nombres. Ce sera l'objectif de ce chapitre.

Nous considérerons sur ce site qu'il existe deux types d'outils fondamentaux en arithmétique (nous ne parlons pas de l'algèbre mais de l'arithmétique!):

1. Les opérateurs arithmétiques:

Il existe deux opérateurs de base (addition et soustraction) à partir desquels nous pouvons construire d'autres opérateurs: la "multiplication" et la "division".

Ces quatre opérateurs sont couramment appelés "**opérateurs rationnels**". Nous verrons ces derniers plus en détails après avoir défini les relations binaires.

Remarque: Rigoureusement l'addition suffirait si nous considérons l'ensemble commun des réels car dès lors la soustraction n'est que l'addition d'un nombre négatif.

2. Les opérateurs (relations) binaires:

Il existe 6 relations binaires fondamentales (égal, différent de, plus grand que, plus petit que, plus grand ou égal, plus petit ou égal) qui permettent de comparer des grandeurs d'éléments se trouvant à gauche et à droite (donc au nombre de deux, d'où leur nom) afin d'en tirer certaines conclusions.

Il est bien évidemment essentiel de connaître au mieux ces deux outils et leurs propriétés avant de se lancer dans des calculs plus ardu.

1. RELATIONS BINAIRES

Le concept de "**relation**" est la base de toute la mathématique dont le but est d'étudier - par observation et déduction (raisonnement), calcul et comparaison - des configurations ou relations abstraites ou concrètes de ses objets (nombres, formes, structures) en cherchant à établir les liens logiques, numériques ou conceptuels entre ces objets.

Définitions:

D1. Considérons deux ensembles non vides E et F (cf. chapitre de Théorie Des Ensembles) non nécessairement identiques. Si à certains éléments x de E nous pouvons associer par une règle mathématique précise R (non ambiguë) un élément y de F , nous définissons ainsi une "**relation fonctionnelle**" de E vers F et qui s'écrit:

$$R: E \rightarrow F \quad (3.1)$$

Ainsi, de façon plus générale, une relation fonctionnelle R peut être définie comme une règle mathématique qui associe à certains éléments x de E , certains éléments y de F .

Alors, dans ce contexte plus général, si xRy , nous disons que y est une "image" de x par R et que x est un "antécédent" ou "pré-image" de y .

L'ensemble des couples (x, y) tels que xRy soit une assertion vraie forme un "graphe" ou une "représentation" de la relation R . Nous pouvons représenter ces couples dans un repère adéquatement choisi pour faire une représentation graphique de la relation R .

Il s'agit d'un type de relations sur lequel nous reviendrons dans le chapitre d'Analyse Fonctionnelle et qui ne nous intéresse pas directement dans ce chapitre.

D2. Considérons un ensemble A non vide, si nous associons à cet ensemble (et à celui-ci uniquement!) des outils permettant de comparer les éléments le composant alors nous parlons de "relation binaire" ou "relation de comparaison" et qui s'écrit pour tout élément x et y composant A :

$$xRy \quad (3.2)$$

Ces relations peuvent aussi être représentées sous forme graphique. Dans le cas des opérateurs binaires classiques de comparaisons où A est l'ensemble des nombres naturels, relatifs, rationnels ou réels, cette forme graphique est représentée par une droite horizontale (le plus souvent...); dans le cas de la congruence (cf. chapitre de Théorie des Nombres) elle est représentée par des droites dans le plan dont les points sont donnés par la contrainte de la congruence.

Comme nous l'avons déjà dit, il existe 6 relations binaires fondamentales (égal, différent de, plus grand que, plus petit que, plus grand ou égal, plus petit ou égal). Mais nous verrons un peu plus loin que la définition rigoureuse des relations binaires permet donc de construire des outils plus abstraits (comme par exemple la congruence bien connue par les élèves de petites classes et que nous étudierons dans le chapitre de Théorie des Nombres).

1.1. ÉGALITÉS

Il est fort difficile de définir la notion "d'égalité" dans un cas général applicable à toute situation. Pour notre part, nous nous permettrons pour cette définition de nous inspirer du théorème d'extensionnalité de la théorie des ensembles (que nous verrons plus tard):

Définitions:

D1. Deux éléments sont égaux si, et seulement si, ils ont les mêmes valeurs. L'égalité est décrite par le symbole $=$ qui signifie "égal à".

Propriété (triviale): Si nous avons $a = b$, et c un nombre et $*$ une opération quelconque (telle que l'addition, la soustraction, la multiplication ou la division) alors:

$$a * c = b * c \quad (3.3)$$

Cette propriété est très utilisée pour résoudre ou simplifier des équations de type quelconque.

D2. Si deux éléments ne sont pas égaux (donc sont inégaux...), nous les relierons par le symbole $\neq$ et nous disons qu'ils sont "non égaux".

Il existe encore d'autres symboles d'égalités, qui sont une extension des deux que nous avons définis précédemment. Malheureusement, ils sont assez souvent mal utilisés (disons plutôt qu'ils sont utilisés

aux mauvais endroits) dans la plupart des ouvrages disponibles sur le marché:

$$\approx ; = ; \cong ; \equiv ; \doteq ; \propto \quad (3.4)$$

qui correspondent dans l'ordre à: presque égal (plutôt utilisé en ingénierie), asymptotiquement égal à (utilisé en analyse fonctionnelle), approximativement égal (utilisé en physique lors d'approximation de séries), identique à (utilisé aussi bien en analyse fonctionnelle qu'en physique), tend vers la limite (idem) et enfin proportionnel à (utilisé en physique ou en mathématiques financières).

1.2. COMPARETEURS

Les comparateurs sont des outils qui nous permettent de comparer et d'ordonner tout couple de nombres (et in extenso aussi des ensembles!).

La possibilité d'ordonner des nombres est presque fondamentale en mathématique. Dans le cas contraire (s'il n'était pas possible ou non imposé d'ordonner), il y aurait des tas de choses qui choqueraient nos habitudes, par exemple (certains des concepts présentés dans la phrase qui suit n'ont pas encore été vus mais nous souhaitons quand même y faire référence): plus de fonctions monotones (en particulier de suites) et lié à cela la dérivation n'indiquerait donc rien sur un "sens de variation", plus d'approche de zéros d'un polynôme par dichotomie (algorithme classique de recherche dans un ensemble ordonné partagé en deux à chaque itération), en géométrie, plus de segments ni de demi-droites, plus de demi-espace, plus de convexité, nous ne pouvons plus orienter l'espace, etc. C'est donc important de pouvoir ordonner les choses comme vous l'aurez compris.

Ainsi, pour tout $a, b, c \in \mathbb{R}$ nous écrivons lorsque a est plus grand ou égal à b :

$$a \geq b \quad (3.5)$$

et lorsque a est plus petit ou égal à b :

$$a \leq b \quad (3.6)$$

Remarque: Il est utile de rappeler que l'ensemble des réels $\mathbb{R}$ est un groupe totalement ordonné (cf. chapitre de Théorie Des Ensembles), sans quoi nous ne pourrions pas définir des relations d'ordre entre ses éléments (ce qui n'est pas le cas des nombres complexes que nous ne pouvons pas ordonner!).

Définition: Le symbole $\leq$ est une "**relation d'ordre**" (voir la définition rigoureuse plus bas!) qui signifie "**plus petit ou égal à**" et inversement le symbole $\geq$ est aussi une relation d'ordre qui signifie "**plus grand ou égal à**".

Nous avons également concernant la relation de comparaison stricte (qui n'appartient pas à la famille des relations d'ordre pour des raisons que nous précisons plus loin) les propriétés suivantes qui sont relativement intuitives:

$$a < b \text{ et } b < c \quad (3.7)$$

implique (ultérieurement noté " $\Rightarrow$ ") que:

$$a < c \quad (3.8)$$

Si:

$$a < b \text{ et } c > 0 \Rightarrow a \cdot c < b \cdot c \quad (3.9)$$

Si:

$$a < b \Rightarrow a + c < b + c \text{ et } a < b \Rightarrow a - c < b - c \quad (3.10)$$

inversement:

$$a > b \Rightarrow a + c > b + c \text{ et } a > b \Rightarrow a - c > b - c \quad (3.11)$$

Nous avons aussi:

$$0 < a < b \Rightarrow \frac{1}{a} > \frac{1}{b} \quad (3.12)$$

et inversement:

$$b < a < 0 \Rightarrow \frac{1}{a} < \frac{1}{b} \quad (3.13)$$

Nous pouvons bien évidemment multiplier, diviser, additionner ou soustraire un terme de chaque côté de la relation telle que celle-ci soit toujours vraie. Petite remarque cependant, si vous multipliez les deux membres par un nombre négatif il faudra bien évidemment changer le comparateur tel que si:

$$a < b \text{ et } c > 0 \Rightarrow a \cdot c < b \cdot c \quad (3.14)$$

et inversement:

$$a < b \text{ et } c < 0 \Rightarrow a \cdot c > b \cdot c \quad (3.15)$$

Nous avons aussi:

$$0 < a < b \text{ et } p \in \mathbb{R}^{*,+} \Rightarrow a^p < b^p \quad (3.16)$$

Soit:

$$b < a < 0 \text{ et } p \in \mathbb{R}^{*,+} \quad (3.17)$$

Si p est un nombre entier pair alors:

$$0 < a^p < b^p \quad (3.18)$$

sinon si p est impair:

$$a^p > b^p \quad (3.19)$$

Ce résultat provient simplement de la multiplication des signes puisque la puissance lorsqu'elle est non fractionnaire n'est qu'une multiplication.

Finalement:

$$0 < a < b \text{ et } n \in \mathbb{N}^* \Rightarrow \sqrt[n]{a} < \sqrt[n]{b} \quad (3.20)$$

Les relations d'ordre :

$$>, <, \leq, \geq, \gg, \ll \quad (3.21)$$

correspondent donc respectivement à: (strictement) plus grand que, (strictement) plus petit que, plus petit ou égal à, plus grand ou égal à, beaucoup plus grand que et enfin beaucoup plus petit que.

Les relations d'ordre peuvent être définies de façon un peu plus subtile et rigoureuse et abstraite et ne s'appliquent pas seulement aux comparateurs (voir par exemple la relation de congruence dans le chapitre de Théorie Des Nombres)!

Voyons cela de suite (le vocabulaire qui va suivre est aussi défini dans le chapitre de Théorie Des Ensembles):

Définition: Soit une relation binaire R d'un ensemble A vers lui-même, une relation R dans A est un sous-ensemble du produit cartésien $R \subseteq A \times A$ (c'est-à-dire que la relation binaire engendre un sous-ensemble de par les contraintes qu'elle impose aux éléments de A qui satisfont la relation) avec la propriété d'être:

P1. Une "**relation réflexive**" si $\forall x \in A$:

$$x R x \quad (3.22)$$

P2. Une "**relation symétrique**" si $\forall x, y \in A$:

$$x R y \Rightarrow y R x \quad (3.23)$$

P3. Une "**relation antisymétrique**" si $\forall x, y \in A$:

$$(x R y \text{ et } y R x) \Rightarrow x = y \quad (3.24)$$

P4. Une "**relation transitive**" si $\forall x, y, z \in A$:

$$(x R y \text{ et } y R z) \Rightarrow x R z \quad (3.25)$$

P5. Une "**relation connexe**" si $\forall x, y \in A$:

$$\forall (x, y) \in A \Rightarrow (x R y \text{ ou } y R x) \quad (3.26)$$

Les mathématiciens ont donné des noms particuliers aux familles de relations satisfaisant certaines de ces propriétés.

Définitions:

D1. Une relation est appelée "**relation d'ordre stricte**" si et seulement si elle est uniquement transitive (certains spécifient alors qu'elle est donc forcément antiréflexive mais on s'en doute...).

D2. Une relation est appelée un "**pré-ordre**" si et seulement si elle est réflexive et transitive.

D3. Une relation est appelée "**une relation d'équivalence**" si et seulement si elle est réflexive, symétrique et transitive.

D4. Une relation est appelée "**relation d'ordre**" si et seulement si elle est réflexive, transitive et antisymétrique.

D5. Une relation est appelée "**relation d'ordre total**" si et seulement si elle est réflexive, transitive, connexe et antisymétrique.

Pour les autres combinaisons il semblerait (?) qu'il n'y ait pas de désignations particulières chez les mathématiciens...

Remarque: Les relations d'ordre binaire ont toutes des propriétés similaires dans les ensembles naturels, rationnels, relatifs et réels (il n'y a pas de relation d'ordre naturelle sur l'ensemble des nombres complexes).

Si nous résumons:

Relation binaire	=	≠	>	<	≤	≥
réflexive	oui	non	non	non	oui	oui
symétrique	oui	oui	non	non	non	non
transitive	oui	non	oui	oui	oui	oui
connexe	non	non	non	non	oui	oui
antisymétrique	oui	non	non	non	oui	oui

Tableau: 3.1 - Types de relations binaires

Ainsi, nous voyons que les relations binaires $\leq, \geq$ forment avec les ensembles précités, des relations d'ordre total et qu'il est très facile de voir quelles relations binaires sont des relations d'ordre partiel, total ou d'équivalence.

Définition: Si R est une relation d'équivalence sur A . Pour $\forall x \in A$, la "**classe d'équivalence**" de x est par définition l'ensemble:

$$[x] = \{y \in A : xRy\} \quad (3.27)$$

$[x]$ est donc un sous-ensemble de A ($[x] \subseteq A$) que nous noterons aussi... par la suite R (attention donc à ne pas confondre dans ce qui suit la relation d'équivalence et le sous-ensemble...).

Nous disposons ainsi d'un nouvel ensemble qui est "**l'ensemble des classes d'équivalences**" ou "**ensemble quotient**" noté A/R . Ainsi:

$$A/R = \{[x] \mid x \in A\} \quad (3.28)$$

Il faut savoir que dans A/R nous ne regardons plus $[x]$ comme un sous-ensemble de A mais comme un élément!

Une relation d'équivalence, de manière vulgarisée sert donc à coller une seule étiquette à des éléments qui vérifient une même propriété, et à les confondre avec ladite étiquette (en sachant ce que nous faisons avec cette étiquette).

Exemple:

Dans l'ensemble des entiers relatifs $\mathbb{Z}$, si nous étudions les restes de la division par 2, nous avons que ceux-ci valent toujours soit 0 soit 1.

La classe d'équivalence de zéro est alors appelée l'ensemble des nombres entiers pairs, la classe d'équivalence de 1 est appelée l'ensemble des entiers impairs. Nous avons donc deux classes d'équivalences pour deux partitions de $\mathbb{Z}$ (gardez toujours cet exemple simple en tête pour les éléments théoriques qui suivront cela aide énormément).

Si nous nommons la première 0 et la deuxième 1, nous retrouvons les règles d'opérations entre nombres pairs et impairs:

$$0 + 0 = 0 \quad 0 + 1 = 1 \quad 1 + 1 = 0 \quad (3.29)$$

ce qui signifie respectivement que la somme de deux entiers pairs est paire, que la somme d'un pair et d'un impair est impaire et que la somme de deux impairs est paire.

Et pour la multiplication :

$$0 \cdot 0 = 0 \quad 0 \cdot 1 = 0 \quad 1 \cdot 1 = 1 \quad (3.30)$$

ce qui signifie respectivement que le produit de deux pairs est pair, le produit d'un pair et d'un impair est pair et que le produit de deux impairs est impair.

Et hop, nous avons déplacé les opérations de $\mathbb{Z}$ sur cet ensemble quotient noté $\mathbb{Z} / 2\mathbb{Z}$.

Maintenant, pour vérifier que nous avons bien affaire à une relation d'équivalence, il faudrait encore vérifier qu'elle est réflexive (xRx), symétrique (si xRy alors yRx) et transitive (si xRy et yRz alors xRz). Nous verrons comment vérifier cela quelques paragraphes plus loin car cet exemple constitue un cas très particulier de relation de congruence.

Définition: L'application $\pi : A \rightarrow A / R$ définie par $x \mapsto [x]$ est appelée "**projection canonique**". Tout élément $z \in [x]$ est alors appelé "**représentant de la classe**" $[x]$.

Considérons maintenant un ensemble E . Alors nous proposons de démontrer qu'il y a bijection entre l'ensemble des relations d'équivalence sur E et l'ensemble des partitions de E . En d'autres termes cette proposition dit qu'une relation d'équivalence sur E n'est rien d'autre qu'une partition de E .

Démonstration:

Soit R une relation d'équivalence sur E . Nous choisissons $I := E / R$ comme ensemble d'indexation des partitions et nous posons pour tout $[x] \in E / R$, $E_{[x]} := [x]$.

Il suffit de vérifier les deux propriétés suivantes de la définition des partitions pour montrer que la famille $(E_{[x]})_I$ est une partition de E :

P1. Soient $[x], [y] \in E/R$ tels que $[x] \neq [y]$ alors (trivial) $E_{[x]} \cap E_{[y]} = \emptyset$.

P2. $E = \bigcup_{[x] \in E/R} E_{[x]}$ est évident car si $x \in E$ alors $x \in [x] = E_{[x]}$.

□C.Q.F.D.

Encore une fois, il est aisé de vérifier avec l'exemple pratique de la division par 2 donné plus haut que la partition des nombres pairs et impairs satisfait ces deux propriétés.

Nous avons donc associé à la relation d'équivalence R une partition de E . Réciproquement si $(E_i)_I$ est une partition de E alors nous vérifions facilement que la relation R définie par xRy si et seulement s'il existe $i \in I$ tel que $x, y \in E_i$ est une relation d'équivalence! Les deux applications ainsi définies sont bijectives et réciproques l'une de l'autre.

Exemple:

Nous allons à présent appliquer sur un exemple un peu moins trivial que le précédent ce que nous venons de voir à la construction des anneaux $\mathbb{Z}/d\mathbb{Z}$ après quelques rappels (pour le concept d'anneau voir le chapitre de Théorie Des Ensembles).

Rappels:

R1. Soit deux nombres $n, m \in \mathbb{Z}$. Nous disons que " **n divise m** " et nous écrivons $n \mid m$ si et seulement si il existe un entier $k \in \mathbb{Z}$ tel que $m = k \cdot n$ (cf. chapitre de Théorie Des Nombres).

R2. Soit $d \geq 1$ un entier. Nous définissons la relation R par nRm si et seulement si $d \mid (n - m)$ ou dit autrement nRm si et seulement si il existe $k \in \mathbb{Z}$ tel que $n = m + k \cdot d$. Généralement nous écrivons ceci aussi $n \equiv m$ (modulo d) au lieu de nRm et nous disons que " **n est congru à m modulo d** ". Rappelons aussi que $n \equiv 0$ (modulo d) si et seulement si d divise n (cf. chapitre de Théorie Des Nombres).

Nous allons maintenant introduire une relation d'équivalence sur $\mathbb{Z}$. Démontrons que pour tout entier $d \geq 1$, la congruence modulo d est une relation d'équivalence sur $\mathbb{Z}$ (nous avons déjà démontré cela dans le chapitre de théorie des nombres lors de notre étude de la congruence mais refaisons le travail pour le plaisir).

Démonstration (contrôle des trois propriétés de l'équivalence):

P1. Réflexivité: $n \equiv n$ car $n = n + 0 \cdot d$.

P2. Symétrie: Si $n \equiv m$ alors $n = m + kd$ et donc $m = n + (-k)d$ c'est-à-dire $m \equiv n$.

P3. Transitivité: Si $n \equiv m$ et $m \equiv j$ alors $n = m + kd$ et $m = j + k'd$ donc $n = j + (k + k') \cdot d$ c'est-à-dire $n \equiv j$.

□C.Q.F.D.

Dans la situation ci-dessus, nous notons $\mathbb{Z}/d\mathbb{Z}$ l'ensemble des classes d'équivalence et noterons $[n]_d$ la classe d'équivalence de la congruence d'un entier n donnée par:

$$[n]_d = \{ \dots, n - 2d, n - d, n, n + d, n + 2d, n + 3d, \dots \} \quad (3.31)$$

(chaque différence de deux valeurs se trouvant dans les accolades est divisible par d et c'est bien ainsi une classe d'équivalence) et ainsi:

$$\mathbb{Z}/d\mathbb{Z} = \{[0]_d, [1]_d, [2]_d, \dots, [d-1]_d\} \quad (3.32)$$

En particulier (trivial car nous obtenons ainsi tout $\mathbb{Z}$):

$$\mathbb{Z}/2\mathbb{Z} = \{[0]_2, [1]_2\} \quad (3.33)$$

Ainsi, nous voyons que le premier exemple que nous avons donné avec les nombres pairs et impairs est un cas particulièrement simple des classes d'équivalence de congruence modulo 2 car elles se réduisent toutes à seulement deux classes.

Remarque: Les opérations d'addition et de multiplication définies sur $\mathbb{Z}$ définissent des opérations d'addition et de multiplication sur $\mathbb{Z}/d\mathbb{Z}$. Nous disons alors que ces opérations sont compatibles avec la relation d'équivalence et forment alors un anneau (cf. chapitre de Théorie Ensembles).

2. LOIS FONDAMENTALES DE L'ARITHMÉTIQUE

Comme nous l'avons déjà dit précédemment, il existe un opérateur de base (addition) à partir duquel il possible de définir la multiplication, la soustraction (à condition que l'ensemble de nombres soit ad hoc) et la division (même remarque que pour la soustraction) et autour desquels nous pouvons construire toute la mathématique analytique.

Bien évidemment il y a certaines subtilités à prendre en compte lorsque le niveau de rigueur augmente. Le lecteur peut alors se reporter au chapitre de Théorie Des Ensembles où ses lois fondamentales sont redéfinies avec plus de justesse.



2.1. ADDITION

Définition: L'addition de nombres entiers est une opération notée "+" qui a pour seul but de réunir en un seul nombre toutes les unités contenues dans plusieurs autres. Le résultat de l'opération se nomme "**somme**" ou "**total**". Les nombres à additionner sont appelés "**termes de l'addition**".

Remarque: Les signes d'addition "+" et de soustraction "-" sont dus à Widmann (1489).

Ainsi, $A+B+C...$ sont les termes de l'addition et le résultat est la somme des termes de l'addition.

Voici une liste de quelques propriétés intuitives que nous admettrons sans démonstrations de l'opération de l'addition:

P1. La somme de plusieurs nombres ne dépend pas de l'ordre des termes. Nous disons alors que l'addition est une "**opération commutative**". Ce qui signifie que nous avons quand A est différent de B :

$$A - B \neq B - A$$

P2. La somme de plusieurs nombres ne change pas si nous remplaçons deux ou plusieurs d'entre eux par leur résultat intermédiaire. Nous disons alors que l'addition est "**opération associative**".

P3. Le zéro est l'élément neutre de l'addition car tout nombre additionné à zéro donne ce même nombre.

P4. Suivant l'ensemble dans lequel nous travaillons, l'addition peut comporter un terme de telle façon à ce que le total soit nul. Nous disons alors qu'il existe un "**opposé**" pour l'addition.

Nous allons définir plus rigoureusement l'addition en utilisant l'axiomatique de Peano dans le cas particulier de l'ensemble des nombres entiers naturels comme nous en avons déjà fait mention dans le chapitre traitant des Nombres. Ainsi, avec ces axiomes il est possible de démontrer qu'il existe (existence) une et une seule application (unicité), notée "+", de $\mathbb{N} \times \mathbb{N}$ dans $\mathbb{N}$ vérifiant:

$$\begin{cases} \forall n \in \mathbb{N}, n + 0 = n \\ \forall n \in \mathbb{N}, s(n) = n + 1 \\ \forall p \in \mathbb{N}, \forall q \in \mathbb{N}, p + s(q) = s(p + q) \end{cases}$$

où s signifie: "successeur".

Remarque: Ce site n'ayant pas pour vocation de s'adresser à des mathématiciens, nous nous passerons de la démonstration (relativement longue) et admettrons intuitivement que l'application "+" existe et est unique... et qu'il en découle les propriétés susmentionnées.

Soient $x_1, x_2, \dots, x_n$ des nombres quelconques alors nous pouvons noter également la somme ainsi:

$$x_1 + x_2 + \dots + x_n = \sum_{i=1}^n x_i \quad (3.34)$$

en définissant des bornes supérieure et inférieure à la somme indexée (au-dessus et en-dessous de la lettre grecque majuscule "sigma").

Voici quelques rappels des propriétés relatives à cette notation condensée:

$$\sum_{i=1}^n k \cdot x_i = k \cdot \sum_{i=1}^n x_i \quad (3.35)$$

où k est une constante et:

$$\sum_{i=1}^n k = n \cdot k \quad (3.36)$$

$$\sum_{i=1}^n (x_i + y_i) = \sum_{i=1}^n x_i + \sum_{i=1}^n y_i \quad (3.37)$$

Voyons maintenant quelques cas concrets d'additions de différents nombres simples afin de mettre en pratique les bases.

Exemples:

L'addition de deux nombres relativement petits est assez facile dès que nous avons appris par coeur à compter jusqu'au nombre résultant de cette opération. Ainsi (exemples pris sur la base décimale):

$$\begin{array}{r} 5 \\ + 2 \\ \hline = 7 \end{array}, \begin{array}{r} 10 \\ + 3 \\ \hline = 13 \end{array}, \begin{array}{r} 1014 \\ + 3 \\ \hline = 1017 \end{array} \quad (3.38)$$

Pour les plus grands nombres il faut adopter une autre méthode qu'il s'agit d'apprendre par coeur. Ainsi par exemple:

$$\begin{array}{r} 9244 \\ + 3475 \\ \hline = ? \end{array} \quad (3.39)$$

Démarche: nous additionnons les colonnes (4 colonnes dans cet exemple) de droite à gauche. Pour la première colonne nous avons donc $4+5=9$ ce qui nous donne:

$$\begin{array}{r} 9244 \\ + 3475 \\ \hline = \dots\dots 9 \end{array} \quad (3.40)$$

et nous continuons ainsi pour la deuxième $4+7=11$ mais à la différence que comme nous avons un nombre supérieur à la dizaine, nous reportons le premier chiffre (de gauche) sur la colonne suivante de l'addition. Ainsi:

$$\begin{array}{r} 92^144 \\ + 3475 \\ \hline = \dots\dots 19 \end{array} \quad (3.41)$$

La troisième colonne se calcule dès lors comme $4+2+1=7$ ce qui nous donne:

$$\begin{array}{r} 92^144 \\ + 5475 \quad (3.42) \\ \hline = \dots 719 \end{array}$$

Pour la dernière colonne nous avons $9+5=14$ et à nouveau nous reportons le premier chiffre (de gauche) sur la colonne suivante de l'addition. Ainsi:

$$\begin{array}{r} ^192^144 \\ + 5475 \quad (3.43) \\ \hline = \dots 4719 \end{array}$$

et la dernière colonne donne:

$$\begin{array}{r} ^192^144 \\ + 5475 \quad (3.44) \\ \hline = 14719 \end{array}$$

Voilà comment nous procédons donc pour l'addition de nombres quelconques: nous faisons une addition par colonne de droite à gauche et si le résultat d'une addition est supérieur à la dizaine, nous reportons une unité sur la colonne suivante.

Cette méthodologie d'addition est simple à comprendre et à effectuer. Nous ne étendrons pas plus sur le sujet pour l'instant.

2.2. SOUSTRACTION

Définition: La soustraction du nombre entier A par le nombre entier B notée par le symbole "-", c'est trouver le nombre C qui, ajouté à B , redonne A .

Remarque: L'opération n'est rigoureusement parlant pas possible dans les entiers naturels $\mathbb{N}$ que si $A > B$.

Nous écrivons la soustraction sous la forme:

$$A - B = C \quad (3.45)$$

qui doit vérifier:

$$A = C + B \quad (3.46)$$

Voici quelques propriétés intuitives que nous admettrons sans démonstrations de l'opération de soustraction (bon cela découle de l'addition...):

P1. La soustraction de plusieurs nombres dépend de l'ordre des termes. Nous disons alors que la soustraction est une "**opération non-commutative**". Effectivement:

$$5 - 2 \neq 2 - 5 \quad (3.47)$$

P2. La soustraction de plusieurs nombres change si nous remplaçons deux ou plusieurs d'entre eux par leur résultat intermédiaire. Nous disons alors que la soustraction est une "**opération non-associative**". Effectivement:

$$5 - (3 - 2) \neq (5 - 3) - 2 \quad (3.48)$$

P3. Le zéro n'est pas l'élément neutre de la soustraction. Effectivement, tout nombre à qui nous soustrayons zéro donne ce même nombre, donc le zéro est neutre à droite... mais pas à gauche car tout nombre que nous soustrayons à zéro ne donne pas zéro!

P4. Suivant l'ensemble dans lequel nous travaillons, la soustraction peut comporter un terme de telle façon à ce que le total soit nul. Nous disons alors qu'il existe un "**opposé**" pour la soustraction.

Exemples:

La soustraction de deux nombres relativement petits est assez facile dès que nous avons appris par coeur à compter jusqu'à au moins le nombre résultant de cette opération. Ainsi:

$$\begin{array}{r} 5 \\ - 2 \\ \hline = 3 \end{array}, \begin{array}{r} 10 \\ - 3 \\ \hline = 7 \end{array}, \begin{array}{r} 1014 \\ - 3 \\ \hline = 1011 \end{array} \quad (3.49)$$

Pour les plus grands nombres il faut adopter une autre méthode qu'il s'agit d'apprendre par coeur (au même titre que l'addition). Ainsi par exemple:

$$\begin{array}{r} 4574 \\ - 3485 \\ \hline = ? \end{array} \quad (3.50)$$

nous soustrayons les colonnes (4 colonnes dans cet exemple) de droite à gauche. Pour la première colonne nous avons $4 - 5 = -1 < 0$ ce qui fait que nous reportons -1 sur la colonne suivante (deuxième) et écrivons $10 - 1 = 9$ en bas de la barre d'égalité de la première colonne:

$$\begin{array}{r} \\ 4574 \\ - 3485 \\ \hline = \dots\dots 9 \end{array} \quad (3.51)$$

et nous continuons ainsi pour la deuxième $7 - 8 = -1 < 0$ ce qui fait que nous reportons -1 sur la colonne suivante (troisième) et comme $-1 - 1 = -2$ nous reportons $10 - 2 = 8$ en bas de la barre d'égalité de la deuxième colonne:

$$\begin{array}{r} \\ 4574 \\ - 3485 \\ \hline = \dots\dots 89 \end{array} \quad (3.52)$$

La troisième colonne se calcule dès lors comme $5 - 7 = -2 < 0$ et nous reportons -1 sur la colonne

suivante (quatrième) et comme $-1 - 2 = -3$ nous reportons $10 - 3 = 7$ en bas de la barre d'égalité de la troisième colonne:

$$\begin{array}{r} \overset{-1}{4} \overset{-1}{5} \overset{-1}{7} 4 \\ - 3785 \\ \hline = 789 \end{array} \quad (3.53)$$

Pour la dernière colonne nous avons $4 - 3 = 1 > 0$ nous reportons donc rien sur la colonne suivante et comme $1 - 1 = 0$ nous reportons 0 en bas de la barre d'égalité de la quatrième colonne:

$$\begin{array}{r} -1-1-1 \\ 4\ 5\ 7\ 4 \\ -\quad 3785 \\ \hline \equiv 0789 \end{array} \quad (3.54)$$

Voilà comment nous procédons donc pour la soustraction de nombres quelconques. Nous faisons une soustraction par colonne de droite à gauche et si le résultat d'une soustraction est inférieur à zéro nous faisons reporter -1 sur la colonne suivante et l'addition du dernier report sur la soustraction obtenue en bas de la barre d'égalité.

La méthodologie utilisée pour la soustraction se basant sur exactement le même principe que l'addition nous ne nous étendrons pas plus sur le sujet. Cette méthode est très simple et nécessite bien sûr une certaine habitude à travailler avec les chiffres pour être totalement appréhendée.

2.3. MULTIPLICATION

Définition: La multiplication de nombres est une opération qui a pour but, étant donné deux nombres, l'un appelé "**multiplieur**" m , et l'autre "**multiplicande**" M , d'en trouver un troisième appelé "**produit**" P qui soit la somme (donc la multiplication d'écoules de la somme!) d'autant de nombres égaux au multiplicande qu'il y a d'unités au multiplieur:

$$m \times M = \sum_{i=1}^m M = P \quad (3.55)$$

Le multiplicande et le multiplicateur sont appelés les "facteurs du produit".

La multiplication s'indique à l'aide du signe "×" (anciennement) ou du point de ponctuation surélevé (notation moderne) ou sans aucun symbole tel que:

$$a \times b = a \cdot b = ab \quad (3.56)$$

Remarque: Le signe de croix "×" pour la multiplication se trouverait pour la première fois dans l'ouvrage d'Oughtred (1631) quant au point à mi-hauteur (notation moderne pour la multiplication), nous le devrions à Leibniz. Dès 1544, Stiefel, dans un de ses ouvrages n'employait aucun signe et désignait le produit de deux nombres en les plaçant l'un après l'autre.

Nous pouvons définir la multiplication en utilisant l'axiomatique de Peano dans le cas particulier des

nombres entiers naturels comme nous en avons déjà fait mention dans le chapitre traitant des Nombres. Ainsi, avec ces axiomes il est possible de démontrer qu'il existe (existence) une et une seule application (unicité), notée " $\times$ " ou plus souvent ".", de $\mathbb{N} \times \mathbb{N}$ dans $\mathbb{N}$ vérifiant:

$$\begin{cases} \forall n \in \mathbb{N}, n \cdot 0 = 0 \\ \forall p \in \mathbb{N}, \forall q \in \mathbb{N}, p \cdot (q + 1) = p \cdot q + p \end{cases} \quad (3.57)$$

Remarque: Ce site n'ayant pas pour vocation de s'adresser à des mathématiciens, nous nous passerons de la démonstration (relativement longue) et admettrons intuitivement que l'application " $\times$ " existe et est unique...

La puissance est une notation particulière d'un cas précis de multiplications. Lorsque le(s) multiplicateur(s) et multiplicande(s) sont identique(s) en valeur numérique, nous notons la multiplication (par exemple):

$$n \cdot n \cdot n \cdot n \cdot n \cdot n \cdot n \cdot n = n^8 \quad (3.58)$$

c'est ce que nous nommons la "**notation en puissance**" ou "**l'exponentiation**". Le nombre en exposant est ce que nous nommons la "**puissance**" ou "**l'exposant**" du nombre (n en l'occurrence). La notation en exposants se trouve pour la première fois dans l'ouvrage de Chuquet intitulé "Triparty en la science des nombres" (1484).

Vous pouvez vérifier par vous-même que ses propriétés sont les suivantes (par exemple):

$$n^x \cdot n^y = n^{x+y} \quad (3.59)$$

Voici quelques propriétés intuitives que nous admettrons sans démonstrations de l'opération de multiplication:

- P1. La multiplication de plusieurs nombres ne dépend pas de l'ordre des termes. Nous disons alors que la multiplication est une "**opération commutative**".
- P2. La multiplication de plusieurs nombres ne change pas si nous remplaçons deux ou plusieurs d'entre eux par leur résultat intermédiaire. Nous disons alors que la multiplication est "**opération associative**".
- P3. L'unité est l'élément neutre de la multiplication car tout multiplicande multiplié par le multiplicateur 1 est égal au multiplicande.
- P4. La multiplication peut comporter un terme de telle façon à ce que le produit soit égal à l'unité (l'élément neutre). Nous disons alors qu'il existe un "**inverse pour la multiplication**" (mais cela dépend rigoureusement dans quel ensemble de nombres nous travaillons).
- P5. La multiplication est "**distributive**", c'est-à-dire que:

$$a \cdot (b + c) = ab + ac \quad (3.60)$$

l'opération inverse s'appelant la "**factorisation**".

Introduisons encore quelques notations particulières relatives à la multiplication:

1. Soient $x_1, x_2, \dots, x_n$ des nombres quelconques (non nécessairement égaux) alors nous pouvons noter le produit ainsi:

$$x_1 \cdot x_2 \cdot \dots \cdot x_n = \prod_{i=1}^n x_i \quad (3.61)$$

en définissant des bornes supérieure et inférieure au produit indexé (au-dessus et en-dessous de la lettre grecque majuscule "Pi").

Rappel des propriétés relatives à cette notation:

$$\prod_{i=1}^n k x_i = k^n \cdot \prod_{i=1}^n x_i \quad (3.62)$$

pour tout nombre k tel que:

$$\prod_{i=1}^n k = k^n \quad (3.63)$$

Nous avons aussi par exemple:

$$\prod_{i=1}^n (x + y) = (x + y)^n \quad (3.64)$$

2. Nous définissons également la "**factorielle**" simplement (car il existe aussi une manière complexe de la définir en passant par la fonction Gamma d'Euler comme cela est fait dans le chapitre de Calcul Différentiel Et Intégral) par:

$$1 \cdot 2 \cdot 3 \cdot 4 \cdot \dots \cdot n = n! \quad (3.65)$$

Exemples:

Voyons quelques exemples simples de multiplications élémentaires. La multiplication de deux nombres relativement petits est assez facile dès que nous avons appris par coeur à compter jusqu'à au moins le nombre résultant de cette opération. Ainsi:

$$\begin{array}{r} 5 \\ \times 2 \\ \hline = 10 \end{array}, \begin{array}{r} 10 \\ \times 3 \\ \hline = 30 \end{array}, \begin{array}{r} 1014 \\ \times 3 \\ \hline = 3042 \end{array} \quad (3.66)$$

Pour les beaucoup plus grands nombres il faut adopter une autre méthode qu'il s'agit d'apprendre par coeur. Ainsi par exemple:

$$\begin{array}{r} 4574 \\ \times 8 \\ \hline = ? \end{array} \quad (3.67)$$

nous multiplions colonne par colonne et nous additionnons l'ensemble des résultats décalés d'un chiffre comme ci-dessous ($8 \times 4 = 32$, $8 \times 7 = 56$, $8 \times 5 = 40$, $8 \times 4 = 32$) ainsi nous obtenons:

$$\begin{array}{r}
 4574 \\
 \times 8 \\
 \hline
 32 \\
 56 \quad (3.68) \\
 40 \\
 32 \\
 \hline
 36592
 \end{array}$$

Cette méthodologie est très logique si vous avez bien compris comment nous construisons un chiffre en base dix. Ainsi, nous avons (nous supposons pour l'instant la distributivité comme connue):

$$\begin{aligned}
 8 \cdot 4574 &= 8 \cdot (4000 + 500 + 70 + 4) = 8 \cdot (4 \cdot 10^3 + 5 \cdot 10^2 + 7 \cdot 10^1 + 4 \cdot 10^0) \\
 &= (8 \cdot 4 \cdot 10^3) + (8 \cdot 5 \cdot 10^2) + 8 \cdot (7 \cdot 10^1) + (8 \cdot 4 \cdot 10^0) \quad (3.69) \\
 &= 32'000 + 4'000 + 560 + 32 = 36'592
 \end{aligned}$$

Pour ne pas surcharger l'écriture dans la multiplication par la méthode "verticale", nous ne représentons pas les zéros qui surchargeraient inutilement les calculs (et ce d'autant plus si le multiplicateur et/ou le multiplicande sont de très grands nombres).

2.4. DIVISION

Définition: La division de nombres entiers (pour commencer par le cas le plus simple...) est une opération, qui a pour but, étant donné deux nombres entiers, l'un appelé "**dividende**", l'autre appelé "**diviseur**", d'en trouver un troisième appelé "**quotient**" qui soit le plus grand nombre dont le produit par le diviseur puisse se retrancher (donc la division découle de la soustraction!) du dividende (la différence étant nommée le "**reste**" ou la "**congruence**").

Remarque: Dans les cas des nombre réels il n'y a jamais de reste à la fin de l'opération de division (car le quotient multiplié par le diviseur donne exactement le dividende)!

D'une façon générale dans le cadre des nombres entiers, si nous notons D le dividende, d le diviseur, Q le quotient et R le reste nous avons la relation:

$$D = Q \cdot d + R \quad (3.70)$$

en sachant que la division était initialement notée de la manière suivante:

$$D : d = \frac{D}{d} \quad (3.71)$$

Nous désignons également souvent par "**fraction**" (au lieu de "quotient"), le rapport de deux nombres ou autrement dit, la division du premier par le deuxième.

Remarque: Le signe de la division ":" est dû à Leibniz. La barre de fraction se trouve elle pour la première fois dans les ouvrages de Fibonacci (1202) et elle est probablement due aux Hindous.

Si nous divisons deux nombres entiers et que nous souhaitons un entier comme quotient et comme reste (s'il y en a un...), alors nous parlons de "**division euclidienne**".

Nous indiquons l'opération en plaçant entre les deux nombres, le dividende et le diviseur un ":" ou une barre de division "/".

Si nous avons:

$$D : d = D \cdot \frac{1}{d} = D \cdot i_D \quad (3.72)$$

nous appelons i_D l'inverse du dividende. A tout nombre est associé un inverse qui satisfait cette condition.

De cette définition il vient la notation (avec x étant un nombre quelconque différent de zéro):

$$\frac{x}{x} = x^1 \cdot x^{-1} = x^{1-1} = x^0 = 1 \quad (3.73)$$

Dans le cas de deux nombres fractionnaires, nous disons qu'ils sont "**inverses**" ou "**réciroques**", lorsque leur produit est égal à l'unité (comme la relation précédente) pour toute valeur de x , positive ou négative.

Remarques:

R1. Une division par zéro est ce que nous nommons une "**singularité**". C'est-à-dire que le résultat de la division est indéterminé.

R2. Lorsque nous multiplions le dividende et le diviseur d'une division (fraction) par un même nombre, le quotient ne change pas (il s'agit d'une fraction équivalente), mais le reste est multiplié par ce nombre.

R3. Diviser un nombre par un produit effectué de plusieurs facteurs revient à diviser ce nombre successivement par chacun des facteurs du produit et réciproquement.

Les propriétés des divisions avec les notations condensées de puissances (exponentiation) sont les suivantes (nous laisserons le soin au lecteur de le vérifier avec des valeurs numériques):

$$\frac{x \cdot x \cdot x}{y \cdot y} = x^3 \cdot y^{-2} \quad (3.74)$$

ou:

$$\frac{x \cdot x \cdot x}{x \cdot x} = \frac{x}{x} \cdot \frac{x}{x} \cdot x = 1 \cdot 1 \cdot x = x = x^3 \cdot x^{-2} = x^{3-2} = x^1 = x \quad (3.75)$$

Rappelons qu'un nombre premier (entier relatif) est un nombre qui n'a d'autres diviseurs que lui-même et l'unité (rappelons que 1 n'est pas un nombre premier). Donc tout nombre qui n'est pas premier a au moins un nombre premier comme diviseur (excepté 1 par définition!). Le plus petit des diviseurs d'un

nombre entier est donc un nombre premier (nous détaillerons les propriétés des nombres premiers relativement au sujet de la division dans le chapitre de Théorie des Nombres).

Voyons quelques propriétés de la division (certaines nous sont déjà connues car elles découlent d'un raisonnement logique des propriétés de la multiplication):

$$\begin{aligned}
 \frac{a}{b} &= \frac{c}{d} \Leftrightarrow a \cdot d = b \cdot c \\
 \frac{a \cdot d}{b \cdot d} &= \frac{a \cdot d}{b \cdot d} = \frac{a}{b} \cdot 1 = \frac{a}{b} \\
 \frac{a}{-b} &= \frac{-a}{b} = -\frac{a}{b} \\
 \frac{a}{b} + \frac{c}{b} &= \frac{a+c}{b} \quad (3.76) \\
 \frac{a}{b} + \frac{c}{d} &= \frac{a \cdot d + b \cdot c}{b \cdot d} \\
 \frac{a \cdot c}{b \cdot d} &= \frac{a \cdot c}{b \cdot d} \\
 \frac{a}{b} \div \frac{c}{d} &= \frac{a \cdot d}{b \cdot c} = \frac{a \cdot d}{b \cdot c}
 \end{aligned}$$

où la deuxième ligne est ce que nous appelons une "**amplification des termes**" et la cinquième ligne une "**mise au dénominateur commun**".

Nous avons aussi les propriétés suivantes:

P1. La division de plusieurs nombres dépend de l'ordre des termes. Nous disons alors que la division est une "**opération non-commutative**". Ce qui signifie que nous avons quand A est différent de B :

$$\frac{A}{B} \neq \frac{B}{A} \quad (3.77)$$

P2. Le résultat de la division de plusieurs nombres change si nous remplaçons deux ou plusieurs d'entre eux par leur résultat intermédiaire. Nous disons alors que la division est "**opération non-associative**":

$$\frac{\left(\frac{A}{B}\right)}{C} \neq \frac{A}{\left(\frac{B}{C}\right)} \quad (3.78)$$

P3. L'unité est l'élément neutre à droite de la division car tout dividende divisé par le diviseur 1 est égal au dividende mais l'unité n'est par contre pas neutre à gauche.

P4. La division peut comporter un terme de telle façon à ce que la division soit égale à l'unité (l'élément neutre). Nous disons alors qu'il existe un "**symétrique pour la division**".

Si a et b sont deux nombres réels positifs et non nuls nous avons:

$$a^p \cdot a^q = a^{p+q}, \frac{a^p}{a^q} = a^{p-q} \Rightarrow a^{-q} = \frac{1}{a^q} \quad (3.79)$$

$$a^p \cdot b^p = (a \cdot b)^p \Leftrightarrow \frac{a^p}{b^p} = \left(\frac{a}{b}\right)^p \quad (3.80)$$

Nous pouvons maintenant définir la racine q -ième principale d'un nombre quelconque a :

$$a^{\frac{1}{q}} = \sqrt[q]{a} \quad (3.81)$$

la dernière relation n'étant définie que pour $q \in \mathbb{N}^*$. Au niveau de la terminologie, nous avons:

$$\sqrt[q]{a} \quad (3.82)$$

qui est une racine, le nombre a est le "**radicande**" et q est l'indice de la racine. Le symbole $\sqrt{}$ est appelé le "**radical**".

De ce qui a déjà été dit pour les puissances, nous pouvons conclure aisément que:

$$\sqrt[q]{a} \cdot \sqrt[q]{b} = \sqrt[q]{a \cdot b} \Leftrightarrow \frac{\sqrt[q]{a}}{\sqrt[q]{b}} = \sqrt[q]{\frac{a}{b}} \quad (3.83)$$

et:

$$\left(a^p\right)^q = a^{p \cdot q} = \left(a^q\right)^p \quad (3.84)$$

il en ressort que:

$$a^{\frac{p}{q}} = \sqrt[q]{a^p} = \left(\sqrt[q]{a}\right)^p \text{ et } \sqrt[q]{\sqrt[q]{a}} = \sqrt[q]{a} \quad (3.85)$$

Nous avons également si $a < 0$:

$$\sqrt[q]{a^q} = a \quad (3.86)$$

si $q \in \mathbb{N}^*$ est impair et:

$$\sqrt[q]{a^q} = |a| \quad (3.87)$$

si $q \in \mathbb{N}^*$ est pair.

Si $a < 0$ et $q \in \mathbb{N}^*$ est impair, alors:

$$a^{\frac{1}{q}} = \sqrt[q]{a} \quad (3.88)$$

est le nombre réel négatif b tel que:

$$b^q = a \quad (3.89)$$

Si $q \in \mathbb{N}^*$ est pair alors bien sûr, comme nous l'avons déjà vu, la racine est complexe (cf. chapitre sur les Nombres).

Si le dénominateur d'un quotient contient un facteur de la forme $\sqrt[n]{a^k}$ avec $a \neq 0$, en multipliant le numérateur et le dénominateur par $\sqrt[n]{a^{n-k}}$, nous supprimerons la racine au dénominateur, puisque:

$$\frac{x}{\sqrt[n]{a^k}} \frac{\sqrt[n]{a^{n-k}}}{\sqrt[n]{a^{n-k}}} = \frac{x\sqrt[n]{a^{n-k}}}{\sqrt[n]{a^k} \sqrt[n]{a^{n-k}}} = \frac{x\sqrt[n]{a^{n-k}}}{\sqrt[n]{a^k a^{n-k}}} = \frac{x\sqrt[n]{a^{n-k}}}{\sqrt[n]{a^{k+n-k}}} = \frac{x\sqrt[n]{a^{n-k}}}{\sqrt[n]{a^n}} = \frac{x\sqrt[n]{a^{n-k}}}{|a|} \quad (3.90)$$

Nous appelons communément ce procédé "**rendre un dénominateur rationnel**". Nous pouvons bien sûr faire de même avec le numérateur.

Exemple:

Voyons un exemple mondialement connu de l'application de la racine qui concerne l'origine des formats papier A6, A5, A4, A3, A2, A1, A0 etc...

Ce format a au fait la propriété (c'est un objectif à l'origine) de conserver ses proportions lorsque nous plions ou coupons la feuille en deux dans sa grande dimension. Ainsi, si nous appelons L la longueur et l la largeur de la feuille, nous avons:

$$\frac{L}{l} = \frac{l}{\left(\frac{L}{2}\right)} \Rightarrow L^2 = 2l^2 \quad (3.91)$$

Il en ressort que:

$$L = \sqrt{2}l \quad (3.92)$$

Comme le format A0 à par définition une superficie de $1 \text{ [m}^2\text{]}$. Pour ce format nous avons alors:

$$L \cdot l = 1 \text{ [m}^2\text{]} \quad (3.93)$$

Nous en déduisons donc:

$$L \cdot l = \sqrt{2}l^2 = \frac{L^2}{\sqrt{2}} = 1 \text{ [m}^2\text{]} \quad (3.94)$$

et donc:

$$l = 2^{-1/4} \cdot 1 \text{ [m]} \cong 84.1 \text{ [cm]} \quad (3.95)$$

d'où nous tirons aussi:

$$L \cong 2^{1/4} \cdot 1 \text{ [m]} = 118.9 \text{ [cm]} \quad (3.96)$$

Les autres formats de déduisant donc pour rappel en divisant par deux la feuille dans sa grande dimension.

3. POLYNÔMES ARITHMÉTIQUES

Définition: Un "**polynôme arithmétique**" (à ne pas confondre avec les polynômes algébriques qui seront étudiés dans la section d'Algèbre) est un ensemble de nombres séparés les uns des autres par les opérations d'addition ou de soustraction (+ ou -).

Les composants enfermés dans le polynôme sont appelés "**termes**" du polynôme. Lorsque le polynôme contient un unique terme, nous parlons alors de "**monôme**", s'il y a deux termes nous parlons de "**binôme**", et ainsi de suite...

La valeur d'un polynôme arithmétique est égale à l'excès de la somme des termes précédés du signe + sur la somme des termes précédés du signe -.

Démonstration:

$$n_1 - n_2 + n_3 - n_4 + n_5 - n_6 + \dots - n_{i-1} + n_i = (n_1 + n_3 + n_5 + \dots + n_i) + (-1) \cdot (n_2 + n_4 + n_6 + \dots + n_{i-1})$$

(3.97)

quelles que soit les valeurs des termes.

□C.Q.F.D.

Mettre en évidence l'unité négative -1 est ce que nous appelons une "**factorisation**" ou "**mise en facteurs**". L'opération inverse, s'appelant une "**distribution**" ou "**développement**".

Le produit de plusieurs polynômes peut toujours être remplacé par un polynôme unique que nous appelons le "**produit effectué**". Nous opérons habituellement comme suit: nous multiplions successivement tous les termes du premier polynôme, en commençant par la gauche, par le premier, le second, ..., le dernier terme du second polynôme. Nous obtenons ainsi un premier produit partiel. Nous faisons, s'il y a lieu, la réduction des termes semblables. Nous multiplions ensuite chacun des termes du produit partiel successivement par le premier, le second, ..., le dernier terme du troisième polynôme en commençant par la gauche et ainsi de suite.

Le produit des polynômes $A, B, C, \dots, L, \dots$ est la somme de tous les produits de n facteurs formés avec un terme de A , un terme de B , ..., et un terme de L . S'il n'y a aucune réduction, le nombre de termes du produit est alors égal au produit des nombres de termes des facteurs.

4. VALEUR ABSOLUE

Un nombre réel est constitué de deux parties: un signe + ou - et une valeur absolue.

 **Exemples:**

E1. +7 est constitué du signe + et de la valeur absolue 7

E2. -5 est constitué du signe - et de la valeur absolue 5

La valeur absolue de +7 est donc 7, la valeur absolue de -5 est donc 5.

Définition: Pour tout nombre réel x , la "valeur absolue" de x , notée $|x|$ est donnée par:

$$\begin{cases} |x| = x & \text{si } x > 0 \\ |x| = -x & \text{si } x < 0 \\ |x| = 0 & \text{si } x = 0 \end{cases} \quad (3.98)$$

Nous remarquons que:

$$|x| = \max(-x, x) \quad (3.99)$$

Ainsi que les expressions équivalentes:

$$x \leq |x| \quad (3.100)$$

et:

$$|-x| = |x| \quad (3.101)$$

et encore:

$$\begin{aligned} |x| \leq y &\Leftrightarrow -y \leq x \leq y \\ |x| \geq y &\Leftrightarrow x \leq -y \vee x \geq y \end{aligned} \quad (3.102)$$

ces dernières étant souvent utilisées dans le cadre de la résolution des inéquations.

Indiquons qu'il est aussi utile d'interpréter l'expression $|x - y|$ comme la distance entre les deux nombres x et y sur la droite réelle. Ainsi, en munissant l'ensemble des nombres réels de la distance valeur absolue, il devient un espace métrique.

La résolution d'une inéquation telle que $|x - 3| \leq 9$ se résout alors simplement à l'aide de la notion de distance. La solution est l'ensemble des réels dont la distance au réel 3 est inférieure ou égale à 9. C'est l'intervalle de centre 3 et de rayon 9 ou autrement écrit:

$$[3 - 9, 3 + 9] = [-6, 12] \quad (3.103)$$

La valeur absolue a quelques propriétés triviales que nous énoncerons sans démonstrations:

P1. La valeur absolue de la somme algébrique de plusieurs nombres réels est inférieure ou égale à la somme des valeurs absolues des composantes de la somme:

$$|x + y| \leq |x| + |y| \quad (3.104)$$

ce que les mathématiciens appellent parfois la "**première inégalité triangulaire**".

P2. La valeur absolue de la différence est supérieure ou égale à la valeur absolue de la différence des valeurs absolues des composantes de la différence:

$$|x - y| \geq ||x| - |y|| \quad (3.105)$$

ce que les mathématiciens appellent parfois la "**deuxième inégalité triangulaire**".

P3. La valeur absolue du produit (multiplication) est égale au produit des valeurs absolues:

$$|x \cdot y| = |x| \cdot |y| \quad (3.106)$$

P4. La valeur absolue du rapport est égale au rapport des valeurs absolues:

$$\frac{|x|}{|y|} = \frac{|x|}{|y|} \quad (3.107)$$

5. RÉGLES DE CALCUL

Fréquemment en informatique (dans le développement en particulier), nous parlons de "**priorité des opérateurs**". En mathématiques nous parlons de "**priorité des ensembles d'opérations et des règles des signes**". De quoi s'agit-il exactement?

Nous avons déjà vu quelles étaient les propriétés des opérations d'addition, soustraction, multiplication, mise en puissance et division. Nous tenons donc à ce que le lecteur différencie la notion de "propriété" de celle de "priorité" (que nous allons tout de suite voir) qui sont deux notions complètement différentes!

En mathématiques, en particulier, nous définissons les priorités des symboles: $\{[()]\}$

Autrement dit:

1. Les opérations qui sont entre parenthèses () doivent être effectuées en premier dans le polynôme.
2. Les opérations qui sont entre crochets [] doivent être effectuées en second à partir des résultats obtenus des opérations qui se trouvaient entre les parenthèses ().
3. Finalement, à partir des résultats intermédiaires des opérations qui se trouvaient entre parenthèses () et crochets [], nous calculons les opérations qui se situent entre les accolades { }.

Faisons un exemple, ceci sera plus parlant.

Exemple:

Soit à calculer le polynôme:

$$\{[5 \cdot (8 + 2) + 3 \cdot [4 + (8 + 6) \cdot 2]] \cdot (1 + 9)\} \cdot 7 + 1 \quad (3.108)$$

Selon les règles que nous avons définies tout à l'heure, nous calculons d'abord tous les éléments qui sont entre parenthèses (), c'est-à-dire:

$$(8 + 2) = 10, (8 + 6) = 14, (1 + 9) = 10 \quad (3.109)$$

ce qui nous donne:

$$\{[5 \cdot 10 + 3 \cdot [4 + 14 \cdot 2]] \cdot 10\} \cdot 7 + 1 \quad (3.110)$$

Toujours selon les règles que nous avons définies tout à l'heure, nous calculons maintenant tous les éléments entre crochets en commençant toujours à calculer les termes qui sont dans les crochets [] au plus bas niveau des autres crochets []. Ainsi, nous commençons par calculer l'expression $[4 + 14 \cdot 2]$ qui se trouve dans le crochet de niveau supérieur: $[5 \cdot 10 + 3 \cdot \dots]$.

Cela nous donne $[4 + 14 \cdot 2] = 32$ et donc:

$$\{[5 \cdot 10 + 3 \cdot 32] \cdot 10\} \cdot 7 + 1 \quad (3.111)$$

Il nous reste à calculer maintenant $[5 \cdot 10 + 3 \cdot 32] = 146$ et donc:

$$\{146 \cdot 10\} \cdot 7 + 1 \quad (3.112)$$

Nous calculons maintenant l'unique terme entre accolades, ce qui nous donne:

$$\{146 \cdot 10\} = 1460 \quad (3.113)$$

Finalement il nous reste:

$$1460 \cdot 7 + 1 = 10'221 \quad (3.114)$$

Evidemment il s'agit d'un cas particulier... Mais le principe est toujours le même.

La priorité des opérateurs arithmétiques est une notion spécifique aux langages informatiques (comme nous en avons déjà fait mention) du fait qu'on ne peut dans ces derniers écrire des relations mathématiques que sur une ligne unique.

Ainsi, en informatique l'expression:

$$\frac{-a \cdot (b+c)^d}{e^f} - g \quad (3.115)$$

s'écrit (à peu de choses près):

$$-a * (b+c)^d / e^f - g \quad (3.116)$$

Un non-initié pourrait y lire:

$$\frac{-a \cdot (b+c)^d}{e^f} - g \text{ ou } \frac{[-a \cdot (b+c)]^d}{e^f} - g \text{ ou } \left\{ \frac{[-a \cdot (b+c)]^d}{e} \right\}^f - g \quad (3.117)$$

ou:

$$[-a \cdot (b+c)]^{\frac{d}{e^f}} - g \quad (3.118)$$

et encore quelques autres... ce qui vous en conviendrez, est fort dangereux car nous arriverons à des résultats différents à chaque fois (cas particuliers mis à part...) !

Ainsi, il a logiquement été défini un ordre de priorité des opérandes tel que les opérations soient effectuées dans l'ordre suivant:

1. - Négation
2. ^ Puissance
3. * / Multiplication et division
4. \ division entière (spécifique à l'informatique)
5. *Mod* Modulo (cf. chapitre de Théorie Des Nombres)
6. + - Addition et soustraction

Evidemment les règles des parenthèses (), crochets [], et accolades { } qui ont été définies en mathématiques s'appliquent à l'informatique.

Ainsi, nous obtenons dans l'ordre (nous remplaçons chaque opération effectuée par un symbole):

D'abord les termes entre parenthèses:

$$-a*(b+c)^d/e^f-g = -a*\alpha^d/e^f-g \quad (3.119)$$

Ensuite les règles de priorité des opérateurs s'appliquent dans l'ordre défini précédemment:

D'abord la négation (1):

$$-a*\alpha^d/e^f-g = \beta*\alpha^d/e^f-g \quad (3.120)$$

ensuite la puissance (2):

$$\beta*\alpha^d/e^f-g = \beta*\chi/\delta-g \quad (3.121)$$

nous appliquons la multiplication (3):

$$\beta*\chi/\delta-g = \varepsilon/\delta-g \quad (3.122)$$

et finalement la division (3):

$$\varepsilon/\delta-g = \phi-g \quad (3.123)$$

Les règles (4) et (5) ne s'appliquent pas à cet exemple particulier.

Finalement (6):

$$\phi-g = \varphi \quad (3.124)$$

Ainsi, en suivant ces règles, ni l'ordinateur, ni l'être humain ne peuvent (ne devraient) se tromper lors de l'interprétation d'une équation écrite sur une ligne unique.

En informatique, il existe cependant plusieurs opérateurs que nous ne retrouvons pas en mathématiques

et qui changent souvent de propriétés d'un langage informatique à un autre. Nous ne nous attarderons pas trop là-dessus cependant, nous avons mis ci-dessous un petit descriptif:

L'opérateur de concaténation " & " est évalué avant les opérateurs de comparaisons.

Les opérateurs de comparaison (=, <, >, ...) possèdent tous une priorité identique.

Cependant, les opérateurs les plus à gauche dans une expression, détiennent une priorité plus élevée.

Les opérateurs logiques sont évalués dans l'ordre de priorité suivant:

1. Not - 2. And - 3. Or - 4. Xor - 5. Eqv - 6. Imp

Maintenant que nous avons vu les priorités des opérateurs, quelles sont les règles des signes en vigueur en mathématiques?

D'abord, il faut savoir que ces dernières ne s'appliquent que dans le cas de la multiplication et de la division. Soient deux nombres positifs $(+x), (+y)$. Nous avons:

$$(+x) \cdot (+y) = (+x) \cdot (+y) = +(x \cdot y) \quad (3.125)$$

Autrement dit, la multiplication de deux nombres positifs est un nombre positif et ceci est généralisable à la multiplication de n nombres positifs.

Nous avons:

$$(-x) \cdot (+y) = (+x) \cdot (-y) = -(x \cdot y) \quad (3.126)$$

Autrement dit, la multiplication d'un nombre positif par un nombre négatif est négative. Ce qui est généralisable à un résultat positif de la multiplication s'il y a un nombre pair de nombres négatifs et à un résultat négatif pour un nombre impair de nombres négatifs sur la totalité n des nombres de la multiplication.

Nous avons:

$$(-x) \cdot (-y) = (+x) \cdot (+y) = +(x \cdot y) \quad (3.127)$$

Autrement dit, la multiplication de deux nombres négatifs est positive. Ce qui est généralisable à un résultat positif de la multiplication s'il y a un nombre pair de nombre négatifs et à un résultat négatif pour un nombre impair de nombres négatifs.

Pour ce qui est des divisions, le raisonnement est identique:

$$(+x)/(+y) = +(x/y) \quad \text{et} \quad (+y)/(+x) = +(y/x) \quad (3.128)$$

Autrement dit, si le numérateur et le dénominateur sont positifs, alors le résultat de la division sera positif.

Nous avons:

$$(+x)/(-y) = (-x)/(+y) = -(x/y) \quad \text{et} \quad (-y)/(+x) = (+y)/(-x) = -(y/x) \quad (3.129)$$

Autrement dit, si soit le numérateur ou le dénominateur est négatif, alors le résultat de la division sera forcément négatif.

Nous avons:

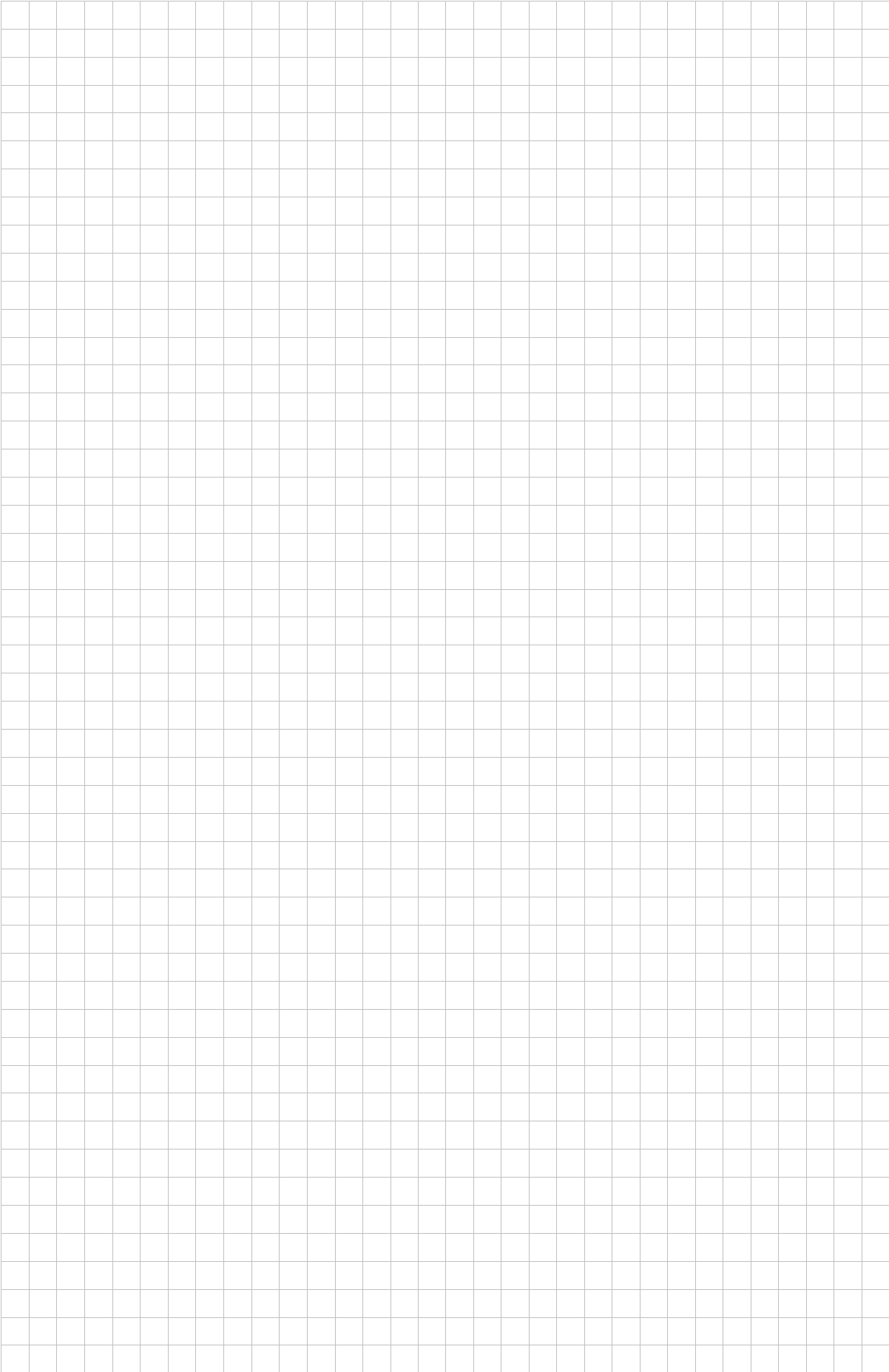
$$(-x)/(-y) = +(x/y) \text{ et } (-y)/(-x) = +(y/x) \quad (3.130)$$

Autrement dit, si le numérateur et le dénominateur sont positifs, alors le résultat de la division, sera forcément positif.

Evidemment, si nous avons une soustraction de termes, il est possible de la réécrire sous la forme:

$$x - y = x + (-1)y = -1 \cdot (-x + y) \quad (3.131)$$







4. THÉORIE DES NOMBRES

Traditionnellement, la théorie des nombres est une branche des mathématiques qui s'occupe des propriétés des nombres entiers, qu'ils soient entiers naturels ou entiers relatifs. Plus généralement, le champ d'étude de cette théorie concerne une large classe de problèmes qui proviennent naturellement de l'étude des entiers. La théorie des nombres peut être divisée en plusieurs branches d'étude (théorie algébrique des nombres, théorie calculatoire des nombres, etc.) en fonction des méthodes utilisées et des questions traitées.

Remarque: Le terme "arithmétique" était aussi utilisé pour faire référence à la théorie des nombres mais c'est un terme assez ancien, qui n'est plus aussi populaire que par le passé.

Nous avons choisi de ne présenter dans cet exposé que les sujets qui sont indispensables à l'étude de la mathématique et de la physique théorique ainsi que ceux devant faire absolument partie de la culture générale de l'ingénieur.

1. PRINCIPE DU BON ORDRE

Nous tiendrons pour acquit ce principe qui dit que tout ensemble non vide $S \subset \mathbb{N}$ contient un plus petit élément.

Nous pouvons utiliser ce théorème pour démontrer une propriété importante des nombres appelée "**propriété archimédienne**" ou "**axiome d'Archimède**" qui s'énonce ainsi:

Pour $\forall a, b \in \mathbb{N}$ où a est non nul, il existe au moins un entier positif n tel que:

$$n \cdot a \geq b \quad (4.1)$$

En d'autres termes, pour deux grandeurs inégales, il existe toujours un multiple entier de la plus petite, supérieur à la plus grande. Nous appelons "**archimédiennes**" des structures dont les éléments vérifient une propriété comparable (cf. chapitre de Théorie Des Ensembles).

Même si cela est trivial à comprendre faisons la démonstration car elle permet de voir le type de démarches utilisées par les mathématiciens quand ils doivent démontrer des éléments triviaux de ce genre...

Démonstration:

Supposons le contraire en disant que pour $\forall n \in \mathbb{N}$ nous avons:

$$n \cdot a < b \quad (4.2)$$

Si nous démontrons que cela est absurde pour tout n alors nous aurons démontré la propriété archimédienne.

Considérons alors l'ensemble:

$$S = \{b - n \cdot a \mid n \in \mathbb{N}\} \quad (4.3)$$

En utilisant le principe du bon ordre, nous en déduisons qu'il existe $s_0 \in S$ tel que $s_0 \leq s$ pour tout $s \in S$. Posons donc que ce plus petit élément est:

$$s_0 = b - n_0 a \quad (4.4)$$

et nous avons donc aussi:

$$b - (n_0 + 1)a \in S \quad (4.5)$$

Comme par hypothèse $n \cdot a < b$ nous devons alors avoir:

$$b - (n_0 + 1)a \geq b - n_0 a \quad (4.6)$$

et si nous réarrangeons et simplifions:

$$-(n_0 + 1) \geq -n_0 \quad (4.7)$$

et que nous simplifions le signe négatif nous devons donc avoir...:

$$n_0 + 1 \leq n_0 \quad (4.8)$$

d'où une contradiction évidente!

Cette contradiction amène que l'hypothèse initiale comme quoi $n \cdot a < b$ pour tout n alors est fausse et donc que la propriété archimédienne est démontrée par l'absurde.

□C.Q.F.D.

2. PRINCIPE D'INDUCTION

Soit S un ensemble de nombres naturels qui possède les deux propriétés suivantes:

P1. $1 \in S$

P2. Si $k \in S$, alors $k + 1 \in S$

Alors:

$$S = \mathbb{N} \setminus \{0\} = \mathbb{N}^* \quad (4.9)$$

Nous construisons ainsi l'ensemble des nombres naturels (se référer au chapitre de Théorie des Ensembles pour voir la construction rigoureuse de l'ensemble des nombres entiers avec les axiomes de Zermelo-Fraenkel).

Soit maintenant:

$$B = \mathbb{N}^* \setminus S \quad (4.10)$$

le symbole " $\setminus$ " signifiant "excluant". Nous voulons démontrer que:

$$B = \emptyset \quad (4.11)$$

A nouveau, même si cela est trivial à comprendre faisons la démonstration car elle permet de voir le type de démarches utilisées par les mathématiciens quand ils doivent démontrer des éléments triviaux de ce genre...

Démonstration:

Supposons le contraire, c'est-à-dire:

$$B \neq \emptyset \quad (4.12)$$

Par le principe du bon ordre, puisque $B \subset \mathbb{N}$, B doit posséder un plus petit élément que nous noterons b_0 .

Mais puisque $1 \in S$ de par (P1), nous avons que $b_0 > 1$ et bien évidemment aussi que $1 \notin B$, c'est-à-dire aussi $b_0 - 1 \in S$. En faisant appel à (P2), nous avons finalement que $b_0 \in S$, c'est-à-dire que $b_0 \notin B$, donc une contradiction.

□C.Q.F.D.

Exemple:

Nous souhaitons montrer à l'aide du principe d'induction, que la somme des n premiers carrés est égale à $n(n+1)(2n+1)/6$, c'est-à-dire que pour $n \geq 1$ nous aurions (cf. chapitre de Suites Et Séries):

$$1^2 + 2^2 + \dots + n^2 = \frac{n(n+1)(2n+1)}{6} \quad (4.13)$$

D'abord la relation ci-dessus est facilement vérifiée pour $n = 1$ nous allons montrer que $n = k + 1$ vérifie aussi cette relation. En vertu de l'hypothèse d'induction:

$$\begin{aligned} 1^2 + 2^2 + \dots + k^2 + (k+1)^2 &= \frac{k(k+1)(2k+1)}{6} + (k+1)^2 \\ &= \frac{(k+1)(k+2)(2k+3)}{6} \end{aligned} \quad (4.14)$$

nous retrouvons bien l'hypothèse de la validité de la première relation mais avec $n = k + 1$, d'où le résultat.

□C.Q.F.D.

Ce procédé de démonstration est donc d'une très grande importance dans l'étude de l'arithmétique; souvent l'observation et l'induction ont permis de soupçonner des lois qu'il eût été plus difficile de trouver par a priori. Nous nous rendons compte de l'exactitude des formules par la méthode précédente qui a donné naissance à l'algèbre moderne par les études de Fermat et de Pascal sur le triangle de Pascal (voir la section d'Algèbre)

3. DIVISIBILITÉ

Définition: Soit $A, B \in \mathbb{Z}$ avec $A \neq 0$. Nous disons que "**A divise B (sans reste)**" s'il existe un entier q (le quotient) tel que:

$$B = A \cdot q \quad (4.15)$$

auquel cas nous écrivons:

$$A|B \quad (4.16)$$

Dans le cas contraire, nous écrivons $A \nmid B$ et nous lisons "**A ne divise pas B**".

Remarques:

1. Se rappeler que le symbole $|$ est une relation alors que le symbole $/$ est une opération!
2. Il ne faut pas confondre l'expression "**A divise B**" qui signifie que le reste est obligatoirement nul et "**A est le diviseur de la division de B**" qui indique que le reste n'est pas forcément nul!

Par ailleurs, si $A|B$, nous dirons aussi que "**B est divisible par A**" ou que "**B est un multiple de A**".

Dans le cas où $A|B$ et que $1 \leq A < B$, nous dirons que A est un "**diviseur propre**" de B .

De plus, il est clair que $A|0$ quel que soit $A \in \mathbb{Z} \setminus \{0\}$ sinon quoi nous avons une singularité.

Voici maintenant quelques théorèmes élémentaires se rattachant à la divisibilité:

T1. Si $A|B$, alors $A|BC$ quel que soit $C \in \mathbb{Z}$

Démonstration:

Si $A|B$, alors il existe un entier q tel que $B = Aq$. Alors, $BC = A(qC)$ et ainsi $A|BC$.

□C.Q.F.D.

T2. Si $A|B$ et $B|C$, alors $A|C$.

Démonstration:

Si $A|B$ et $B|C$, alors il existe des entiers q et r tels que $B = Aq$ et $C = Br$. Donc, $C = A(qr)$ et ainsi $A|C$.

□C.Q.F.D.

T3. Si $A|B$ et $A|C$, alors:

$$A|(Bx + Cy), \forall x, y \in \mathbb{Z} \quad (4.17)$$

Démonstration:

Si $A|B$ et $A|C$, alors il existe des entiers q et r tels que $B = Aq$ et $C = Ar$. Il s'ensuit que:

$$Bx + Cy = (Aq)x + (Ar)y = A(qx + ry) \quad (4.18)$$

et ainsi que $A | (Bx + Cy) \forall x, y \in \mathbb{Z}$.

□C.Q.F.D.

T4. Si $A|B$ et $B|A$, alors $A = \pm B$

Démonstration:

Si $A|B$ et $B|A$, alors il existe des entiers q et r tels que $B = Aq$ et $A = Br$. Nous avons donc $B = B(qr)$ et ainsi $qr = 1$; c'est pourquoi nous pouvons avoir $q = \pm 1$ si $r = \pm 1$ et qu'ainsi $A = \pm B$

□C.Q.F.D.

T5. Si $A|B$ et $B \neq 0$ alors $|A| \leq |B|$

Démonstration:

Si $A|B$ et $B \neq 0$, alors il existe un entier $q \neq 0$ tel que $B = Aq$. Mais alors, $|B| = |A||q| \geq |A|$, puisque $|q| \geq 1$.

□C.Q.F.D.

3.1. DIVISION EUCLIDIENNE

La division euclidienne est une opération qui, à deux entiers naturels appelés dividende et diviseur, associe deux entiers appelés quotient et reste. Initialement définie aux entiers naturels non nuls, elle se généralise aux entiers relatifs et aux polynômes, par exemple.

Définition: Nous appelons "**division euclidienne**" ou "**division entière**" de deux nombres A et B l'opération consistant à diviser B par A en s'arrêtant quand le reste devient strictement inférieur à A .

Rappelons (cf. chapitre Nombres) que tout nombre qui admet exactement les deux diviseurs euclidiens (dont la division donne un reste nul) que sont 1 et lui-même est dit "**nombre premier**" (ce qui exclut le nombre 1 de la liste des nombres premiers) et que tout couple de nombres qui n'ont que 1 comme diviseur euclidien commun sont dits "**premiers entre eux**".

Soient $A, B \in \mathbb{Z}$ avec $A > 0$. Le "**théorème de la division euclidienne**" affirme qu'il existe des entiers uniques q et r tels que:

$$B = Aq + r \Leftrightarrow r = B - Aq \quad (4.19)$$

où $0 \leq r < A$. De plus, si $A \nmid B$, alors $0 < r < A$.

Démonstration:

Considérons l'ensemble:

$$S = \left\{ r = B - qA \mid q, B \in \mathbb{Z}, A \in \mathbb{Z}^*, B - qA \geq 0 \right\} \quad (4.20)$$

Il est relativement facile de voir que $S \subset \mathbb{N}^* \cup \{0\}$ et que $S \neq \emptyset$, d'où, d'après le principe du bon ordre, nous concluons que S contient un plus petit élément $r \geq 0$.

Soit q l'entier satisfaisant donc à:

$$r = B - Aq \quad (4.21)$$

Nous voulons d'abord montrer que $r < A$ en supposant le contraire (démonstration par l'absurde), c'est-à-dire que $r \geq A$. Alors, dans ce cas, nous avons:

$$B - qA = r \geq A \quad (4.22)$$

ce qui est équivalent à:

$$B - (q+1)A = r - A \geq 0 \quad (4.23)$$

mais $B - (q+1)A \in S$ et:

$$B - (q+1)A < B - qA \quad (4.24)$$

ce qui contredit le fait que:

$$r = B - qA \quad (4.25)$$

est le plus petit élément de S . Donc, $r < A$. Enfin, il est clair que si $r = 0$, nous avons $A|B$, d'où la seconde affirmation du théorème.

□C.Q.F.D.

Remarque: Dans l'énoncé de la division euclidienne, nous avons supposé que $A > 0$.
Qu'obtenons-nous lorsque $A < 0$? Dans cette situation, $-A$ est positif, et alors nous pouvons appliquer la division euclidienne à B et $-A$. Par conséquent, il existe des entiers q et r tels que:

$$B = q(-A) + r \text{ où } 0 \leq r < |A| \quad (4.26)$$

Or, cette relation peut s'écrire:

$$B = -q(A) + r \quad (4.27)$$

où bien sûr, $-q$ est un entier. La conclusion est que la division euclidienne peut s'énoncer sous la forme plus générale:

Soient $A, B \in \mathbb{Z}$, alors il existe des entiers q et r tels que:

$$B = Aq + r \quad (4.28)$$

où $0 \leq r < |A|$. De plus, si $A \nmid B$, alors $0 < r < |A|$.

Les entiers q et r sont uniques dans la division euclidienne. En effet, s'il existe deux autres entiers q_1 et r_1 tels que:

$$B = Aq_1 + r_1 \quad (4.29)$$

avec toujours $0 \leq r_1 < A$, alors:

$$A(q_1 - q) = r - r_1 \quad (4.30)$$

et ainsi $A \mid (r - r_1)$. En vertu de (T5) nous avons, si $r - r_1 \neq 0$, $|r - r_1| \geq A$.

Or, cette dernière inégalité est impossible puisque par construction $-A < r - r_1 < A$. Donc, $r = r_1$ et, puisque $A \neq 0$, alors $q_1 = q$ d'où l'unicité.

3.1.1. PLUS GRAND COMMUN DIVISEUR

Soit $a, b \in \mathbb{Z}$ tels que $ab \neq 0$. Le "**plus grand commun diviseur**" (noté "PGCD" par la suite) de a et b , noté:

$$(a, b) \quad (4.31)$$

est l'entier naturel d non nul qui satisfait aux deux propriétés suivantes:

P1. $d \mid a$ et $d \mid b$ (donc sans reste r dans la division!)

P2. si $c \mid a$ et $c \mid b$ alors $c \leq d$ et $c \mid d$ (par définition!)

Notons que 1 est toujours un diviseur commun de deux entiers arbitraires.

Exemple:

Considérons les entiers positifs 36 et 54. Un diviseur commun de 36 et 54 est un entier positif qui divise 36, et aussi 54. Par exemple, 1 et 2 sont des diviseurs communs de 36 et 54.

$$\begin{aligned} \text{Div}_{36} &= \{1, 2, 3, 4, 6, 9, 12, 18, 36\} \\ \text{Div}_{54} &= \{1, 2, 3, 6, 9, 18, 27, 54\} \end{aligned} \quad (4.32)$$

Nous avons alors l'intersection représentée par le diagramme de Venn suivant:

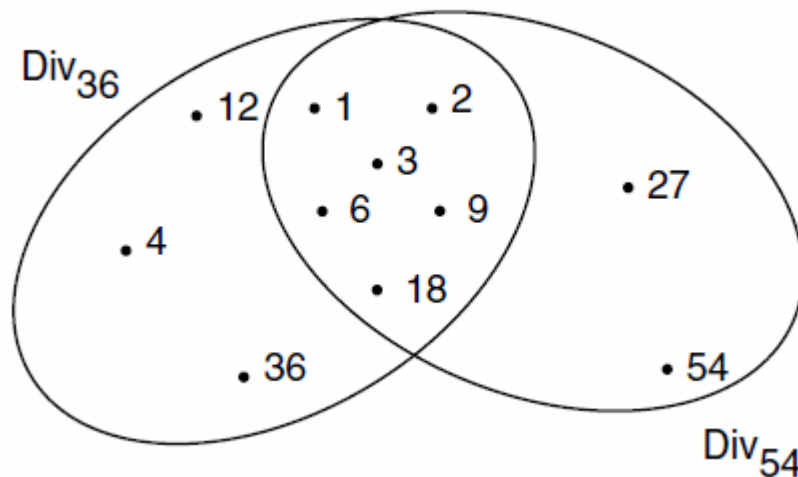


Figure: 4.1 - Diagramme de Venn des diviseurs communs

avec l'ensemble des diviseurs communs suivant:

$$\text{Div}_{36} \cap \text{Div}_{54} = \{1, 2, 3, 6, 9, 18\} \quad (4.33)$$

et donc le PGCD est:

$$\max\{1, 2, 3, 6, 9, 18\} = 18 \quad (4.34)$$

et nous constatons que l'ensemble des diviseurs communs de 36 et 54 est aussi l'ensemble des diviseurs de 18.

Cependant, il n'est pas forcément évident que le PGCD autre qu'unitaire (c'est-à-dire différent 1) de deux entiers a et b qui ne sont pas premiers entre eux existe toujours. Ce fait est démontré dans le théorème suivant (cependant, si le PGCD existe, il est de par sa définition unique!) dit "**théorème de Bézout**" qui permet aussi de démontrer d'autres propriétés intéressantes de deux nombres comme nous le verrons plus tard.

Démonstration:

Soient $a, b \in \mathbb{Z}$ tels que $ab \neq 0$. Si d divise a et d divise b (pour les deux sans reste r !) il existe alors obligatoirement des entiers relatifs x et y tels que:

$$\boxed{d = (a, b) = ax + by} \quad (4.35)$$

Cette relation est appelée "**identité de Bézout**" et il s'agit d'une équation diophantienne linéaire (cf. chapitre de Calcul Algébrique).

Evidemment, si a et b sont premiers entre eux nous savons que d vaut alors 1.

Pour démontrer l'identité de Bézout considérons d'abord l'ensemble:

$$S = \{d = ax + by \mid x, y \in \mathbb{Z}, ax + by > 0\} \quad (4.36)$$

Comme $S \subset \mathbb{N}$ et $S \neq \emptyset$, nous pouvons utiliser le principe du bon ordre et conclure que S possède un plus petit élément d . Nous pouvons alors écrire:

$$d = ax_0 + by_0 \quad (4.37)$$

pour un certain choix $x_0, y_0 \in \mathbb{Z}$. Il suffit donc de montrer que $d = (a, b)$ pour démontrer l'identité de Bézout!

Procédons via une démonstration par l'absurde en posant supposant $d \nmid a$. Alors si c'est le cas, d'après la division euclidienne, il existe $q, r \in \mathbb{Z}$ tels que $a = qd + r$, où $0 < r < d$. Mais alors:

$$r = a - qd = a - q(ax_0 + by_0) = a(1 - qx_0) + b(-qy_0) \quad (4.38)$$

Ainsi, nous avons que $r \in S$ et $r < d$, ce qui contredit le fait que d est le plus petit élément possible de S . Donc nous avons démontré ainsi non seulement que $d \mid a$ mais qu'en plus d existe toujours et, de la même façon, nous démontrons que $d \mid b$.

Comme corollaire important montrons maintenant que si $a, b \in \mathbb{Z}$ tels que $ab \neq 0$, alors:

$$S = \{ax + by \mid x, y \in \mathbb{Z}\} \quad (4.39)$$

constitue l'ensemble de tous les multiples de $d = (a, b)$:

Comme $d \mid a$ et $d \mid b$, alors nous avons forcément $d \mid ax + by$ pour tout $x, y \in \mathbb{Z}$. Soit $M = \{nd \mid n \in \mathbb{Z}\}$. Notre problème se réduit au fait à montrer que $S = M$.

Soit d'abord $s \in S$ ce qui signifie que $d \mid s$ et qui implique $s \in M$.

Soit un $m \in M$, cela voudrait donc dire que $m = nd$ pour un certain $n \in \mathbb{Z}$.

Comme $d = ax_0 + by_0$ pour un choix d'entiers quelconques $x_0, y_0 \in \mathbb{Z}$, alors:

$$m = nd = n(ax_0 + by_0) = a(nx_0) + b(ny_0) \in S \quad (4.40)$$

□C.Q.F.D.

Les hypothèses peuvent sembler compliquées mais portez plutôt votre attention un certain temps sur la dernière relation. Vous allez tout de suite comprendre!

Remarque: Si au lieu de définir le PGCD de deux entiers non nuls, nous permettons à l'un d'entre eux d'être égal à 0, disons: $a \neq b$, $b = 0$. Dans ce cas, nous avons $a|b$ et, selon notre définition du PGCD, il est clair que $(a, 0) = |a|$.

Soit $d = (a, b)$ et soit $m \in \mathbb{Z}$, alors nous avons les propriétés suivantes du PGCD:

P1. $(a, b + ma) = (a, b) = (a, -b)$

P2. $(am, bm) = |m|(a, b)$ où $m \neq 0$

P3. $\left(\frac{a}{d}, \frac{b}{d}\right) = 1$

P4. Si $g \in \mathbb{Z} \setminus \{0\}$ tel que $g|a$ et $g|b$ alors $\left(\frac{a}{g}, \frac{b}{g}\right) = \frac{1}{|g|}(a, b)$

Dans certains ouvrages, ces quatre propriétés sont démontrées en utilisant intrinsèquement la propriété elle-même. Personnellement nous nous en abstenons car faire cela est plus ridicule qu'autre chose à notre goût car la propriété est une démonstration en elle-même.

Elaborons maintenant une méthode (algorithme) qui s'avérera très importante pour calculer (déterminer) le plus grand commun diviseur de deux entiers (utile en informatique parfois).

3.1.2. ALGORITHME D'EUCLIDE

L'algorithme d'Euclide est un algorithme permettant donc de déterminer le plus grand commun diviseur de deux entiers.

Pour aborder cette méthode de manière intuitive, il faut savoir que vous devez comprendre un nombre entier comme une longueur, un couple d'entiers comme un rectangle (côtés) et leur PGCD est la taille du plus grand carré permettant de carreler (paver) ce rectangle par définition (oui si vous réfléchissez un petit moment c'est assez logique!).

L'algorithme décompose le rectangle initial en carrés, de plus en plus petits, par divisions euclidiennes successives, de la longueur par la largeur, puis de la largeur par le reste, jusqu'à un reste nul. Il faut bien comprendre cette démarche géométrique pour comprendre ensuite l'algorithme.

Exemple:

Considérons que nous cherchons le PGCD (a, b) où b vaut 21 et a vaut 15 et gardons à l'esprit que le PGCD, outre le fait qu'il divise a et b , doit laisser un reste nul! En d'autres termes il doit pouvoir diviser le reste de la division de b par a aussi!

Nous avons donc le rectangle de 21 par 15 suivant:

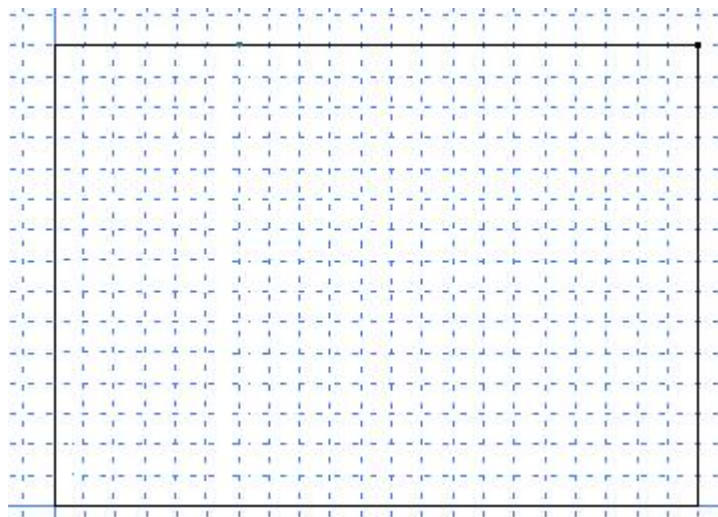


Figure: 4.2 - Première étape de l'algorithme PGCD

D'abord nous regardons si 15 est le PGCD (on commence toujours par le plus petit). Nous divisons alors 21 par 15 ce qui équivaut géométriquement à:

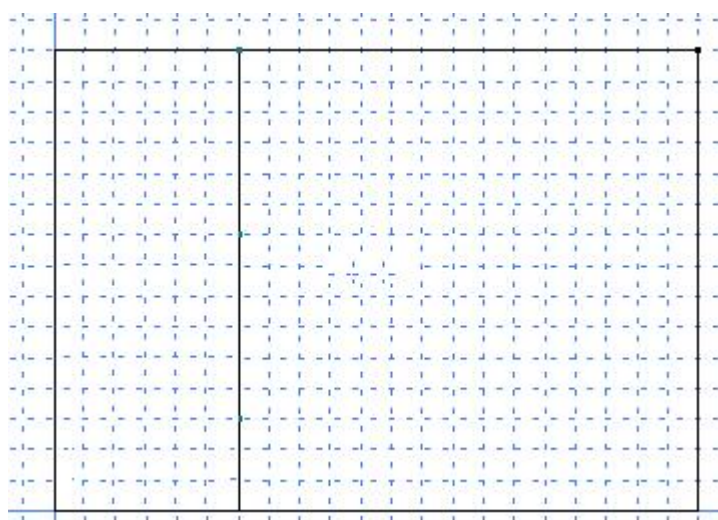


Figure: 4.3 - Deuxième étape de l'algorithme PGCD

15 n'est donc pas le PGCD (on s'en doutait...). Nous voyons immédiatement que nous n'arrivons pas à paver le rectangle avec un carré de 15 par 15.

Nous avons donc un reste de 6 (rectangle de gauche). Le PGCD comme nous le savons doit, s'il existe, par définition pouvoir diviser ce reste et laisser un reste nul.

Il nous reste donc un rectangle de 15 par 6. Nous cherchons donc maintenant à paver ce nouveau rectangle car nous savons que le PGCD est par construction inférieur ou égal à 6. Nous avons alors:



Figure: 4.4 - Troisième étape de l'algorithme PGCD

Et nous divisons donc 15 par le reste 6 (ce résultat sera inférieur à 6 et permet immédiatement de tester

si le reste sera le PGCD). Nous obtenons:



Figure: 4.5 - Quatrième étape de l'algorithme PGCD

A nouveau, nous n'arrivons pas à paver ce rectangle rien qu'avec des carrés. En d'autres termes, nous avons un reste non nul qui vaut 3. Soit un rectangle de 6 par 3. Nous cherchons donc maintenant à paver ce nouveau rectangle car nous savons que le PGCD est par construction inférieur ou égal à 3 et qu'il laissera un reste nul s'il existe. Nous avons alors géométriquement:

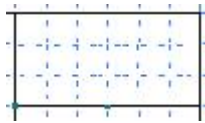


Figure: 4.6 - Cinquième étape de l'algorithme PGCD

Nous divisons 6 par 3 (ce qui sera inférieur à 3 et permet immédiatement de tester si le reste sera le PGCD):

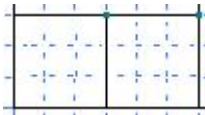


Figure: 4.7 - Sixième et dernière étape de l'algorithme PGCD

et c'est tout bon! Nous avons 3 qui laisse donc un reste nul et divise le reste 6 il s'agit donc du PGCD. Nous avons donc au final:

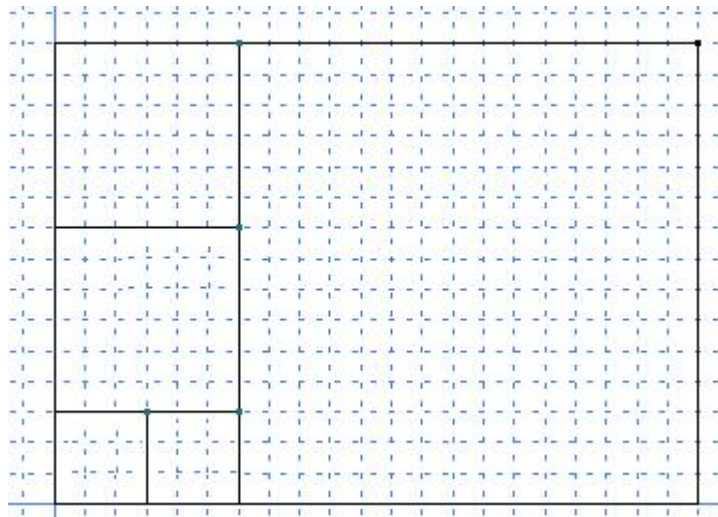


Figure: 4.8 - Résumé de l'algorithme PGCD

Maintenant, voyons l'approche formelle équivalente:

Soient $a, b \in \mathbb{Z}$, où $a > 0$. En appliquant successivement la division euclidienne (avec $b > a$), nous obtenons la suite d'équations:

$$\begin{aligned}
 b &= aq_1 + r_1 & 0 < r_1 < a \\
 a &= r_1q_2 + r_2 & 0 < r_2 < r_1 \\
 r_1 &= r_2q_3 + r_3 & 0 < r_3 < r_2 \\
 &\dots & \dots \\
 r_{j-2} &= r_{j-1}q_j + r_j & 0 < r_j < r_{j-1} \\
 r_{j-1} &= r_jq_{j+1}
 \end{aligned} \tag{4.41}$$

Si $d = (a, b)$, alors $d = r_j$.

Sinon de manière plus formelle:

Démonstration:

Nous voulons d'abord montrer que $r_j = (a, b)$. Or, d'après la propriété P1:

$$(a, b + ma) = (a, b) = (a, -b) \tag{4.42}$$

nous avons:

$$(a, b) = (a, r_1) = (r_1, r_2) = \dots = (r_{j-1}, r_j) \tag{4.43}$$

Pour démontrer la deuxième propriété de l'algorithme d'Euclide, nous écrivons l'avant-dernière équation du système sous la forme:

$$r_j = r_{j-2} - q_j r_{j-1} \tag{4.44}$$

Or, en utilisant l'équation qui précède cette avant-dernière équation du système, nous avons:

$$\begin{aligned}
 r_j &= r_{j-2} - q_j(r_{j-3} - q_{j-1}r_{j-2}) \\
 &= (1 + q_jq_{j-1})r_{j-2} + (-q_j)r_{j-3}
 \end{aligned} \tag{4.45}$$

En continuant ce processus, nous arrivons à exprimer r_j comme une combinaison linéaire de a et b .

□C.Q.F.D.

Exemple:

Calculons le plus grand commun diviseur de (429,966) et exprimons ce nombre comme une combinaison linéaire de 429 et de 966.

Nous appliquons bien évidemment l'algorithme d'Euclide:

$$\begin{aligned}
 966 &= 429 \cdot 2 + 108 \\
 429 &= 108 \cdot 3 + 105 \\
 108 &= 105 \cdot 1 + 3 \\
 105 &= 3 \cdot 35
 \end{aligned} \tag{4.46}$$

Nous en déduisons donc que:

$$r_j = d = (966, 429) = 3 \quad (4.47)$$

et, de plus, que:

$$\begin{aligned} 3 &= 108 - 105 \cdot 1 = 108 - (429 - 108 \cdot 3) = 108 \cdot 4 - 429 \\ &= (966 - 429 \cdot 2) \cdot 4 - 429 = 966 \cdot 4 - 429 \cdot 9 = 966 \cdot 4 + 429 \cdot (-9) \end{aligned} \quad (4.48)$$

Donc le PGCD est bien exprimé comme une combinaison linéaire de a et b et constitue à ce titre le PGCD.

Définition: Nous disons que les entiers $a_1, a_2, \dots, a_n$ sont "**relativement premiers**" ou "**premiers entre eux**" si:

$$(a_1, a_2, \dots, a_n) = 1 \quad (4.49)$$

3.1.3. PLUS PETIT COMMUN MULTIPLE

Définitions:

D1. Soient $a_1, a_2, \dots, a_n \in \mathbb{Z} \setminus \{0\}$, nous disons que m est un "**commun multiple**" de $a_1, a_2, \dots, a_n$ si $a_i \mid m$ pour $i = 1, 2, \dots, n$.

D2. Soient $a_1, a_2, \dots, a_n \in \mathbb{Z} \setminus \{0\}$, nous appelons "**plus petit commun multiple**" (PPCM) de $a_1, a_2, \dots, a_n$, noté:

$$[a_1, a_2, \dots, a_n] \quad (4.50)$$

le plus petit entier commun multiple positif à tous les communs multiples de $a_1, a_2, \dots, a_n$.

 **Exemple:**

Considérons les entiers positifs 3 et 5. Un multiple commun de 3 et 5 est un entier positif qui est à la fois un multiple de 3, et un multiple de 5. Autrement dit, qui est divisible par 3 et aussi par 5. Nous avons donc:

$$\begin{aligned} M_3 &= \{3, 6, 9, 12, 15, 18, 21, 24, 27, 30, \dots\} \\ M_5 &= \{5, 10, 15, 20, 25, 30, 35, 40, 45, \dots\} \end{aligned} \quad (4.51)$$

Nous avons alors l'intersection représentée par le diagramme de Venn suivant:

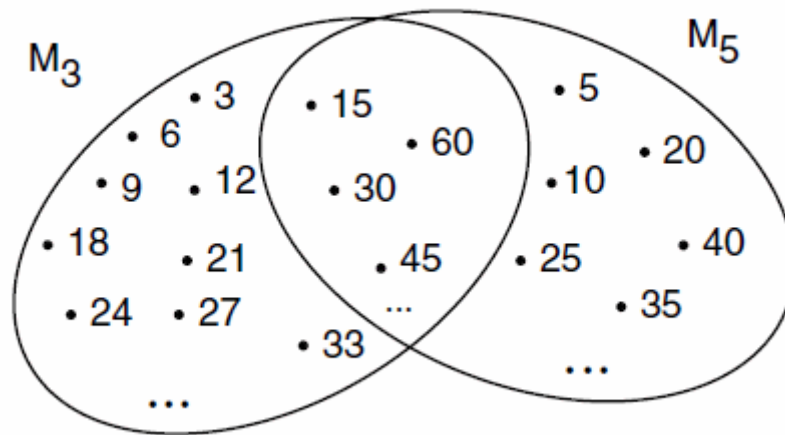


Figure: 4.9 - Diagramme de Venn des communs multiples

avec l'ensemble des communs multiples suivants:

$$M_3 \cap M_5 = \{15, 30, 45, 60, \dots\} \quad (4.52)$$

et le PPCM est alors:

$$\min \{15, 30, 45, 60, \dots\} = 15 \quad (4.53)$$

Nous constatons que l'ensemble des multiples communs de 3 et 5 est aussi l'ensemble des multiples de 15.

Remarque: Soient $a_1, a_2, \dots, a_n \in \mathbb{Z} \setminus \{0\}$. Alors, le plus petit commun multiple existe. En effet, considérons l'ensemble E des entiers naturels m qui pour tout i divisent a_i . Ce que nous noterons:

$$E = \{m \mid a_i \mid m, m \in \mathbb{N}, i = 1, 2, \dots, n\} \quad (4.54)$$

Puisque nous avons forcément $|a_1 a_2 \dots a_n| \in E$, alors l'ensemble est non vide et, d'après l'axiome du bon ordre, l'ensemble E contient un plus petit élément positif.

Voyons maintenant quelques théorèmes relatifs au PPCM:

T1. Si m est un commun multiple quelconque de $a_1, a_2, \dots, a_n$ alors $[a_1, a_2, \dots, a_n] \mid m$, c'est-à-dire que m divise chacun des a_i .

Démonstration:

Soit $M = [a_1, a_2, \dots, a_n]$. Alors, d'après la division euclidienne, il existe des entiers q et r tels que:

$$m = qM + r, 0 \leq r < M \quad (4.55)$$

Il suffit de montrer que $r = 0$. Supposons $r \neq 0$ (démonstration par l'absurde). Puisque $a_i \mid m$ et $a_i \mid M$

, alors on a $\alpha_i \mid r$ et cela pour $i = 1, 2, \dots, n$. Donc, r est un commun multiple de $\alpha_1, \alpha_2, \dots, \alpha_n$ plus petit que le PPCM. On vient d'obtenir une contradiction, ce qui prouve le théorème.

□C.Q.F.D.

T2. Si $k > 0$, alors $[ka_1, ka_2, \dots, ka_n] = k[a_1, a_2, \dots, a_n]$

La démonstration sera supposée évidente (dans le cas contraire contactez-nous et cela sera détaillé!)

T3. $[a, b] \cdot (a, b) = |ab|$

Démonstration:

Pour la démonstration, nous allons utiliser le "**lemme d'Euclide**" qui dit que si $a \mid bc$ et $(a, b) = 1$ alors $a \mid c$.

Effectivement cela se vérifie aisément car nous avons vu qu'il existe $x, y \in \mathbb{Z}$ tels que $1 = ax + by$ et alors $c = acx + bcy$. Mais $a \mid ac$ et $a \mid bc$ impliquent que $a \mid (acx + bcy)$, c'est-à-dire également que $a \mid c$.

Revenons à notre théorème:

Puisque $(a, b) = (a, -b)$ et $[a, b] = [a, -b]$, il suffit de prouver le résultat pour des entiers positifs a et b . En tout premier lieu, considérons le cas où $(a, b) = 1$. L'entier $[a, b]$ étant un multiple de a , nous pouvons écrire $[a, b] = ma$. Ainsi, nous avons $b \mid ma$ et, puisque $(a, b) = 1$, il s'ensuit, d'après le lemme d'Euclide, que $b \mid m$. Donc, $b \leq m$ et alors $ab \leq am$. Mais ab est un commun multiple de a et b qui ne peut être plus petit que le PPCM. c'est pourquoi $ab = ma = [a, b]$.

Pour le cas général, c'est-à-dire $(a, b) = d > 1$, nous avons, d'après la propriété:

$$\left(\frac{a}{d}, \frac{b}{d} \right) = 1 \quad (4.56)$$

et avec le résultat obtenu précédemment que:

$$\left[\frac{a}{d}, \frac{b}{d} \right] \cdot \left(\frac{a}{d}, \frac{b}{d} \right) = \frac{a}{d} \frac{b}{d} \quad (4.57)$$

Lorsque nous multiplions des deux côtés de l'équation par d^2 , le résultat suit et la démonstration est effectuée.

□C.Q.F.D.

3.2. THÉORÈME FONDAMENTAL DE L'ARITHMÉTIQUE

Le théorème fondamental de l'arithmétique dit que tout nombre naturel $n > 1$ peut s'écrire comme un produit de nombres premiers, et cette représentation est unique, à part l'ordre dans lequel les facteurs premiers sont disposés.

Le théorème établit l'importance des nombres premiers. Essentiellement, ils sont les briques

élémentaires de construction des entiers positifs, chaque entier positif contenant des nombres premiers d'une manière unique.

Remarque: Ce théorème est parfois appelé "**théorème de factorisation**" (un peu à tort... car d'autres théorèmes portent le même nom...).

Démonstration:

Si n est premier, et donc produit d'un unique entier premier, à savoir lui-même le résultat est vrai et la démonstration est terminée (dire qu'un nombre premier est produit de lui-même est bien évidemment un abus de langage!). Supposons que n n'est pas premier et donc strictement supérieur à 1 et considérons l'ensemble:

$$D = \{d \mid d \mid n \text{ et } 1 < d < n\} \quad (4.58)$$

Alors, $D \subset \mathbb{N}$ et, puisque n est composé, nous avons que $D \neq \emptyset$. D'après le principe du bon ordre, D possède un plus petit élément p_1 qui est premier, sans quoi le choix minimal de p_1 serait contredit. Nous pouvons donc écrire $n = p_1 n_1$. Si n_1 est premier, alors la preuve est terminée. Si n_1 est composé, alors nous répétons le même argument que précédemment et nous en déduisons l'existence d'un nombre premier p_2 et d'un entier $n_2 < n_1$ tels que $n = p_1 p_2 n_2$. En poursuivant ainsi nous arrivons forcément à la conclusion que n_k sera premier.

Donc finalement nous avons bien démontré qu'un nombre quelconque est décomposable en facteurs de nombres premiers à l'aide du principe du bon ordre.

□C.Q.F.D.

Nous ne connaissons pas à ce jour de loi simple qui permette de calculer le n -ième facteur premier (p_n). Ainsi, pour savoir si un entier m est premier, il est pratiquement plus facile à ce jour de vérifier sa présence dans une table de nombres premiers.

En fait, nous utilisons aujourd'hui la méthode suivante:

Soit un nombre m , si nous voulons déterminer s'il est premier ou non, nous calculons s'il est divisible par les nombres premiers p_n qui appartiennent à l'ensemble:

$$\{p_n \in \mathbb{N} \mid p_n < \sqrt{m}\} \quad (4.59)$$

Exemple:

L'entier 223 n'est ni divisible par 2, ni par 3, ni par 5, ni par 7, ni par 11, ni par 13. Il est inutile de continuer avec le prochain nombre premier, car $17^2 = 289 > 223$. Nous en déduisons dès lors que le nombre 223 est premier.

3.3. CONGRUENCES

Définition: Soit $m \in \mathbb{Z} \setminus \{0\}$. Si a et b ont même reste dans la division euclidienne par m nous disons

que "*a* est congru à *b* modulo *m*", et nous écrivons:

$$a \equiv b \pmod{m} \quad (4.60)$$

ou de manière équivalente il existe un nombre entier relatif *k* tel que:

$$b = a + km \quad (4.61)$$

Le lecteur pourra vérifier que cela impose que

$$m \mid (a - b) \quad (4.62)$$

soit en français.... que *m* divise la différence entre *a* et *b*. Dans le cas contraire, nous disons que "*a* est non congru à *b* modulo *m*".

Une autre manière de dire tout cela si ce n'est pas clair...:

L'étude de ces propriétés qui relient trois nombres entre eux est appelée communément "*l'arithmétique modulaire*".

Remarques:

R1. Que nous soyons bien d'accord, la congruence implique un reste nul pour la division !

R2. Nous excluons en plus de 0 aussi 1 et -1 pour les valeurs que peut prendre *m* dans la définition de la congruence dans certains ouvrages.

R3. Derrière le terme de congruence se cachent des notions semblables mais de niveaux d'abstraction différents:

- En arithmétique modulaire, nous disons donc que "deux entiers relatifs *a* et *b* sont congrus modulo *m* s'ils ont même reste dans la division euclidienne par *m*". Nous pouvons aussi dire qu'ils sont congrus modulo *m* si leur différence est un multiple de *m*.

- Dans la mesure des angles orientés, nous disons que "deux mesures sont congrues modulo $2\pi [rad]$ si et seulement si leur différence est un multiple de $2\pi [rad]$ ". Cela caractérise deux mesures d'un même angle (cf. chapitre de Trigonométrie).

- En algèbre, nous parlons de congruence modulo *I* dans un anneau commutatif (cf. chapitre de Théorie Des Ensembles) dont *I* est un idéal: "*x* est congru à *y* modulo *I* si et seulement si leur différence appartient *I*". Cette congruence est une relation d'équivalence, compatible avec les opérations d'addition et multiplication et permet de définir un anneau quotient de l'ensemble parent avec son idéal *I*.

- Nous trouvons parfois, dans l'étude de la géométrie (cf. chapitre de Géométrie Euclidienne) le terme de congru mis à la place de semblable. Il s'agit alors d'une simple relation d'équivalence sur l'ensemble des figures planes.

La relation de congruence $\equiv$ est une relation d'équivalence (cf. chapitre sur les Opérateurs), en d'autres

termes , soient $a, b, c, m \in \mathbb{Z}, m > 1$ alors la relation de congruence est:

P1. Réflexive:

$$a \equiv a \pmod{m} \quad (4.63)$$

P2. Symétrique:

$$a \equiv b \pmod{m} \Leftrightarrow b \equiv a \pmod{m} \quad (4.64)$$

P3. Transitive:

$$a \equiv b \pmod{m}, b \equiv c \pmod{m} \Rightarrow a \equiv c \pmod{m} \quad (4.65)$$

Démonstration:

Les propriétés P1 et P2 sont évidentes (si ce n'est pas le cas faites-le nous savoir nous développerons!). Nous démontrerons P3. Les hypothèses impliquent que $b = a + km, c = b + lm$. Mais alors:

$$c = b + lm = (a + km) + lm = a + (k + l)m \quad (4.66)$$

ce qui montre que a et c sont congrus modulo m .

□C.Q.F.D.

La relation de congruence $\equiv$ est compatible avec la somme et le produit (se rappeler que la puissance n'est finalement qu'une extension du produit!).

Effectivement, soient $a, b, a', b', m \in \mathbb{Z}, m > 1$ tel que $a \equiv b \pmod{m}$ et $a' \equiv b' \pmod{m}$ alors:

$$P1. a + a' \equiv b + b' \pmod{m}$$

$$P2. aa' \equiv bb' \pmod{m}$$

Démonstrations:

Nous avons:

$$a = b + km, a' = b' + lm \quad (4.67)$$

par hypothèse. Mais alors:

$$a + a' = b + b' + (l + k)m \quad (4.68)$$

ce qui démontre P1. Nous avons également:

$$aa' = bb' + blm + b'km + klm^2 = bb' + sm \quad (4.69)$$

ce qui démontre P2.

□C.Q.F.D.

Remarque: La relation de congruence se comporte sur de nombreux points comme la relation d'égalité. Néanmoins une propriété de la relation d'égalité n'est plus vraie pour celle de congruence, à savoir la simplification: si $ab \equiv ac \pmod{m}$, nous n'avons pas nécessairement $b \equiv c \pmod{m}$.

Exemple:

$$2 \cdot 1 \equiv 2 \cdot 3 \pmod{4} \text{ mais } 1 \not\equiv 3 \pmod{4}$$

Jusqu'ici, nous avons vu des propriétés des congruences faisant intervenir un seul modulus. Nous allons maintenant étudier le comportement de la relation de congruence lors d'un changement de modulus.

P1. Si $a \equiv b \pmod{m}$ et $d|m$, alors $a \equiv b \pmod{d}$

P2. Si $a \equiv b \pmod{r}$ et $a \equiv b \pmod{s}$ alors a et b sont congrus modulo $[r,s]$

Ces deux propriétés sont évidentes. Inutile d'aller dans les détails pour P1. Pour P2, puisque $b-a$ est un multiple de r et de s puisque par hypothèse:

$$\frac{b-a}{r} = k, \frac{b-a}{s} = l \Rightarrow b-a = rk = sl \quad (4.70)$$

$b-a$ est donc un multiple du PPCM de r et s , ce qui démontre P2.

De ces propriétés il vient que si nous désignons par $f(x)$ un polynôme à coefficient entiers (positifs ou négatifs):

$$f(x) = Ax^n + Bx^{n-1} + Cx^{n-2} + \dots + Kx + L \quad (4.71)$$

La congruence $a \equiv b \pmod{m}$ donnera aussi $f(a) \equiv f(b) \pmod{m}$.

Si nous remplaçons x successivement par tous les nombres entiers dans un polynôme $f(x)$ à coefficients entiers, et si nous prenons les résidus pour le module m , ces résidus se reproduisent de m en m (dans le sens où la congruence se vérifie), puisque nous avons, quel que soit l'entier m et x :

$$f(x) \equiv f(x+m) \pmod{m} \quad (4.72)$$

Nous en déduisons alors l'impossibilité de résoudre la congruence suivante:

$$f(x) \equiv r \pmod{m} \quad (4.73)$$

en nombres entiers, si r désigne l'un quelconque des non-résidus (un résidu qui ne satisfait pas la congruence).

3.3.1. CLASSES DE CONGRUENCE

Définition: Nous appelons "**classe de congruence modulo m** ", le sous-ensemble de l'ensemble $\mathbb{Z}$ défini par la propriété que deux éléments a et b de $\mathbb{Z}$ sont dans la même classe si et seulement si $a \equiv b \pmod{m}$ ou qu'un ensemble d'éléments entre eux sont congrus par ce même modulo.

Remarque: Nous avons vu dans le chapitre traitant des opérateurs qu'il s'agit en fait d'une classe d'équivalence car la congruence modulo m est, comme nous l'avons démontré plus haut, une relation d'équivalence.

Exemple:

Soit $m = 3$. Nous divisons l'ensemble des entiers en classes de congruence modulo 3. Exemple de trois ensembles dont tous les éléments sont congrus entre eux sans reste (observez bien ce que donne l'ensemble des classes!):

$$\begin{aligned} &\{\dots, -9, -6, -3, 0, 3, 6, 9, 12, \dots\} \\ &\{\dots, -8, -5, -2, 1, 4, 7, 10, 13, \dots\} \quad (4.74) \\ &\{\dots, -7, -4, -1, 2, 5, 8, 11, \dots\} \end{aligned}$$

Ainsi, nous voyons que pour chaque couple d'élément d'une classe de congruence modulo 3 existe. Cependant, nous voyons que nous ne pouvons pas prendre $-9 \equiv -8 \pmod{3}$ où -9 se trouve dans la première classe et -8 dans la seconde.

Le plus petit nombre non négatif de la première classe est 0, celui de la deuxième est 1 et celui de la dernière est 2. Ainsi, nous noterons ces trois classes respectivement $[0]_3, [1]_3, [2]_3$, le chiffre 3 en indice indiquant le modulus.

Il est intéressant de noter que si nous prenons un nombre quelconque de la première classe et un nombre quelconque de la deuxième, alors leur somme est toujours dans la deuxième classe. Ceci se généralise et permet de définir une somme sur les classes modulo 3 en posant:

$$\begin{aligned} [0]_3 + [0]_3 &= [0]_3 & [0]_3 + [1]_3 &= [1]_3 \\ [0]_3 + [2]_3 &= [2]_3 & [1]_3 + [1]_3 &= [2]_3 \quad (4.75) \\ [1]_3 + [2]_3 &= [0]_3 & [2]_3 + [2]_3 &= [1]_3 \end{aligned}$$

Ainsi que:

$$\begin{aligned} [0]_3 \cdot [0]_3 &= [0]_3 & [0]_3 \cdot [1]_3 &= [0]_3 \\ [0]_3 \cdot [2]_3 &= [0]_3 & [1]_3 \cdot [1]_3 &= [1]_3 \quad (4.76) \\ [1]_3 \cdot [2]_3 &= [2]_3 & [2]_3 \cdot [2]_3 &= [1]_3 \end{aligned}$$

Ainsi, pour tout $m > 1$, la classe de congruence de:

$$a \pmod{m} \quad (4.77)$$

est l'ensemble des entiers congrus à a modulo m (et congrus entre eux modulo m). Cette classe est notée:

$$[a]_m \quad (4.78)$$

Remarque: Le fait d'avoir mis entre parenthèses l'expression "et congrus entre eux modulo m " est dû au fait que la congruence, étant une relation d'équivalence nous avons comme nous l'avons démontré plus haut que si $b \equiv a \pmod{m}, c \equiv a \pmod{m}$, alors $b \equiv c \pmod{m}$.

Définition: L'ensemble des classes de congruence $[a]_m$ (qui forment par le fait que la congruence est une relation d'équivalence des: "classes d'équivalence"), pour un m fixe donne ce que nous appelons un "ensemble quotient" (cf. chapitre Opérateurs). Plus rigoureusement, nous parlons de "l'ensemble quotient de $\mathbb{Z}$ par la relation de congruence" dont les éléments sont les classes de congruence (ou: classes d'équivalence) et qui forment alors l'anneau $\mathbb{Z}/m\mathbb{Z}$.

Nous déduisons de la définition les deux propriétés triviales suivantes:

P1. Le nombre b est dans la classe $[a]_m$ si et seulement si $a \equiv b \pmod{m}$

P2. Les classes $[a]_m$ et $[b]_m$ sont égales si et seulement si $a \equiv b \pmod{m}$

Montrons maintenant qu'il y a exactement m différentes classes de congruence modulo m , à savoir $[0]_m, [1]_m, \dots, [m-1]_m$.

Démonstration:

Soit $m > 1$, alors tout nombre entier a est congru modulo m à un et un seul entier r de l'ensemble $\{0, 1, 2, \dots, m-1\}$ (remarquez bien, c'est important, que nous nous restreignons aux entiers positifs ou nuls sans prendre en compte les négatifs!). De plus, cet entier r est exactement le reste de la division de a par m . En d'autres termes, si $0 \leq r < m$, alors:

$$a \equiv r \pmod{m} \quad (4.79)$$

si et seulement si $a = qm + r$ où q est le quotient de a par m et r le reste. La démonstration est donc une conséquence immédiate de la définition de la congruence et de la division euclidienne.

□C.Q.F.D.

Définition: Un entier b dans une classe de congruence modulo m est appelé "représentant de cette classe" (il est clair que par la relation d'équivalence que deux représentants d'une même classe sont donc congrus entre eux modulo m).

Nous allons pouvoir maintenant définir une addition et une multiplication sur les classes de congruences. Pour définir la somme de deux classes $[a]_m, [b]_m$, il suffit de prendre un représentant de chaque classe, de faire leur somme et de prendre la classe de congruence du résultat. Ainsi (voir les exemples plus haut):

$$[a]_m + [b]_m = [a + b]_m \quad (4.80)$$

et de même pour la multiplication:

$$[a]_m \cdot [b]_m = [ab]_m \quad (4.81)$$

Par définition de la somme et du produit, nous constatons que la classe de 0 est l'élément neutre pour l'addition:

$$[a]_m + [0]_m = [a]_m \quad \forall a \in \mathbb{Z}, \forall m \in \mathbb{N} \quad (4.82)$$

et la classe de l'entier 1 est l'élément neutre pour la multiplication:

$$[a]_m \cdot [1]_m = [a]_m \quad \forall a \in \mathbb{Z}, \forall m \in \mathbb{N} \quad (4.83)$$

Définition: Un élément $[a]_m$ de $\mathbb{Z}/m\mathbb{Z}$ est "**une unité**" s'il existe un élément $[b]_m \in \mathbb{Z}/m\mathbb{Z}$ tel que $[a]_m \cdot [b]_m = 1$

Le théorème suivant permet de caractériser les classes modulo m qui sont des unités dans $\mathbb{Z}/m\mathbb{Z}$:

Théorème: Soit $[a]$ un élément de $\mathbb{Z}/m\mathbb{Z}$. Alors $[a]$ est une unité si et seulement si $(a, m) = 1$.

Démonstration:

Supposons d'abord que $(a, m) = 1$. Alors par Bézout, nous avons son identité:

$$as + mr = 1 \quad (4.84)$$

Autrement dit, as est congru à 1 modulo m . Mais ceci est équivalent à écrire par définition que $[a][s] = [1]$ ce qui montre que $[a]$ est une unité. Réciproquement, si $[a]$ est une unité, ceci implique qu'il existe une classe $[s]$ telle que $[a][s] = [1]$.

Ainsi, nous venons de démontrer que $\mathbb{Z}/m\mathbb{Z}$ constitue bien un anneau puisqu'il possède une addition, une multiplication, un élément neutre et un inverse.

□C.Q.F.D.

3.4. FRACTIONS CONTINUES

La notion de fraction continue remonte à l'époque de Fermat et atteint son apogée avec les travaux de Lagrange et Legendre vers la fin du 18ème siècle. Ces fractions sont importantes en physique car nous les retrouvons en acoustique ainsi que dans la démarche intellectuelle qui a amené Galois à créer sa théorie des groupes.

Considérons dans un premier temps le nombre rationnel a/b avec $(a, b) = 1$ avec $b > 0$ et $a > b$. Nous savons que tous les quotients q_i et les restes r_i sont dans le cadre de la division euclidienne des entiers positifs.

Rappelons l'algorithme d'Euclide vu plus haut (mais noté de manière un peu différente):

$$\begin{aligned}
 \frac{a}{b} &= q_1 + \frac{r_1}{b} \\
 \frac{b}{r_1} &= q_2 + \frac{r_2}{r_1} \\
 \frac{r_1}{r_2} &= q_3 + \frac{r_3}{r_2} \quad (4.85) \\
 &\dots \\
 \frac{r_{n-2}}{r_{n-1}} &= q_n
 \end{aligned}$$

Par substitutions successives, nous obtenons:

$$\begin{aligned}
 \frac{a}{b} &= q_1 + \frac{1}{b/r_1} = q_1 + \frac{1}{q_2 + \frac{1}{r_1/r_2}} \\
 &\dots \\
 &= q_1 + \frac{1}{q_2 + \frac{1}{q_3 + \frac{1}{\ddots \frac{1}{q_{n-1} + \frac{1}{q_n}}}}} \quad (4.86)
 \end{aligned}$$

Ce qui est aussi parfois noté:

$$\frac{a}{b} = q_1 + \frac{1}{1/\varepsilon_1} = q_1 + \frac{1}{q_2 + 1/\varepsilon_2} = \dots \quad (4.87)$$

Ainsi, tout nombre rationnel positif peut s'exprimer comme une fraction continue finie où $q_n \in \mathbb{N}$.

Exemples:

E1. Cherchons l'expression de $17/49$. Nous savons déjà que $0 < 17/49 < 1$ donc que $q_1 = 0$. Nous avons alors:

$$\begin{aligned}
 \frac{17}{49} &= 0 + \frac{1}{49/17} = 0 + \frac{1}{2 + \frac{1}{17/15}} = 0 + \frac{1}{2 + \frac{1}{1 + \frac{1}{15/2}}} = 0 + \frac{1}{2 + \frac{1}{1 + \frac{1}{7 + \frac{1}{2/1}}}} \quad (4.88)
 \end{aligned}$$

Nous voyons bien dans cet exemple que nous avons effectivement $q_n \in \mathbb{N}$. Nous pouvons également remarquer que par construction:

$$q_n = \left[\frac{1}{\varepsilon_{n-1}} \right] \quad (4.89)$$

où les crochets représentent la partie entière et nous avons aussi:

$$\frac{1}{\varepsilon_{n-1}} = q_n + \varepsilon_n \quad (4.90)$$

E2. Voyons comment extraire la racine carrée d'un nombre A par la méthode des fractions continues.

Soit a le plus grand nombre entier dont le carré a^2 est plus petit que A . On le soustrait de A . Il y a donc un reste de:

$$r = A - a^2 = (\sqrt{A} - a)(\sqrt{A} + a) \quad (4.91)$$

où nous avons utilisé une des identités remarquables vues dans le chapitre d'Algèbre. D'où en divisant les deux membres par la deuxième parenthèse, nous avons:

$$\sqrt{A} - a = \frac{r}{\sqrt{A} + a} \quad (4.92)$$

Soit:

$$\sqrt{A} = a + \frac{r}{\sqrt{A} + a} \quad (4.93)$$

Dans le dénominateur, nous remplaçons $\sqrt{A}$ par:

$$a + \frac{r}{\sqrt{A} + a} \quad (4.94)$$

Cela donne:

$$\sqrt{A} = a + \frac{r}{2a + \frac{r}{\sqrt{A} + a}} \quad (4.95)$$

etc.... on voit ainsi que le système est simple pour déterminer l'expression d'une racine en termes de fraction continue.

Le développement du nombre a/b s'appelle le "**développement du nombre a/b en fraction continue finie**" et est condensé sous la notation suivante:

$$[q_1, q_2, q_3, \dots, q_n] \quad (4.96)$$

Nous considérerons comme intuitif que tout nombre rationnel peut s'exprimer comme fraction continue finie et inversement que toute fraction continue finie représente un nombre rationnel. Par extension, un nombre irrationnel est représenté par une fraction continue infinie!

Considérons maintenant $[q_1, q_2, q_3, \dots, q_n]$ une fraction continue finie. La fraction continue:

$$C_k = [q_1, q_2, \dots, q_k] \quad (4.97)$$

où $k = 1, 2, \dots, n$ est appelée la "**k-ième réduite**" ou la "**k-ième convergente**" ou encore le "**k-ième quotient partiel**".

Avec cette notation, nous avons:

$$\begin{aligned} C_1 &= [q_1] = q_1 \\ C_2 &= [q_1, q_2] = q_1 + \frac{1}{q_2} = \frac{q_1 q_2 + 1}{q_2} \\ C_3 &= [q_1, q_2, q_3] = q_1 + \frac{1}{q_2 + \frac{1}{q_3}} = \frac{(q_1 q_2 + 1) \cdot q_3 + q_1}{q_2 q_3 + 1} \quad (4.98) \\ C_4 &= [q_1, q_2, q_3, q_4] = \frac{\{(q_1 q_2 + 1) \cdot q_3 + q_1\} \cdot q_4 + (q_1 q_2 + 1)}{(q_2 q_3 + 1) \cdot q_4 + q_2} \end{aligned}$$

Pour simplifier les expressions ci-dessus, nous introduisons les suites $\{n_i\}, \{d_i\}$ (n pour numérateur et d pour dénominateur) définies par:

$$\begin{aligned} n_0 &= 1, \quad n_1 = q_1, \quad \dots, \quad n_i = q_i n_{i-1} + n_{i-2} \\ d_0 &= 0, \quad d_1 = 1, \quad \dots, \quad d_i = q_i d_{i-1} + d_{i-2} \end{aligned} \quad (4.99)$$

à l'aide de cette construction, nous avons une petite inégalité immédiate intéressante pour un peu plus loin:

$$0 = d_0 < d_1 < d_2 < d_3 \dots \quad (4.100)$$

Avec la définition ci-dessus, nous constatons que:

$$C_1 = \frac{n_1}{d_1}, C_2 = \frac{n_2}{d_2}, C_3 = \frac{n_3}{d_3}, \dots, C_4 = \frac{n_4}{d_4} \quad (4.101)$$

Soit en généralisant:

$$C_k = [q_1, q_2, \dots, q_k] = \frac{n_k}{d_k} \quad (4.102)$$

Maintenant, montrons pour un usage ultérieur que pour $i \geq 1$, nous avons:

$$n_i d_{i-1} - d_i n_{i-1} = (-1)^i \quad (4.103)$$

Le résultat est immédiat pour $i = 1$. En supposant que le résultat est vrai pour i montrons qu'il est aussi vrai pour $i + 1$. Puisque:

$$n_{i+1}d_i - d_{i+1}n_i = (q_{i+1}n_i + n_{i-1})d_i - (q_{i+1}d_i + d_{i-1})n_i = -(d_{i-1}n_i - n_{i-1}d_i) = (-1)^{i+1} \quad (4.104)$$

alors en utilisant l'hypothèse d'induction, nous obtenons le résultat!

Nous pouvons maintenant établir une relation indispensable pour la suite. Montrons que si C_k est la k -ième réduite de la fraction continue simple finie $[q_1, q_2, \dots, q_n]$ alors:

$$C_1 < C_3 < C_5 < \dots < C_6 < C_4 < C_2 \quad (4.105)$$

Démonstration:

$$\begin{aligned} C_{k+2} - C_k &= (C_{k+2} - C_{k+1}) + (C_{k+1} - C_k) \\ &= \left(\frac{n_{k+2}}{d_{k+2}} - \frac{n_{k+1}}{d_{k+1}} \right) + \left(\frac{n_{k+1}}{d_{k+1}} - \frac{n_k}{d_k} \right) = \left(\frac{n_{k+2}d_{k+1} - n_{k+1}d_{k+2}}{d_{k+2}d_{k+1}} \right) + \left(\frac{n_{k+1}d_k - n_kd_{k+1}}{d_{k+1}d_k} \right) \\ &= \left(\frac{n_{k+2}d_{k+1} - d_{k+2}n_{k+1}}{d_{k+2}d_{k+1}} \right) + \left(\frac{n_{k+1}d_k - d_{k+1}n_k}{d_{k+1}d_k} \right) = \frac{(-1)^{k+2}}{d_{k+2}d_{k+1}} + \frac{(-1)^{k+1}}{d_{k+1}d_k} \\ &= \frac{(-1)^{k+2}d_k + (-1)^{k+1}d_{k+2}}{d_{k+2}d_{k+1}d_k} = \frac{(-1)^{k+1}(-1)d_k + (-1)^{k+1}d_{k+2}}{d_{k+2}d_{k+1}d_k} \\ &= \frac{-(-1)^{k+1}d_k + (-1)^{k+1}d_{k+2}}{d_{k+2}d_{k+1}d_k} = \frac{(-1)^{k+1}(d_{k+2} - d_k)}{d_{k+2}d_{k+1}d_k} \end{aligned} \quad (4.106)$$

puisque:

$$0 = d_0 < d_1 < d_2 < d_3 \dots \quad (4.107)$$

donc:

$$d_{k+2} - d_k > 0 \quad (4.108)$$

ce qui nous indique que le signe $C_{k+2} - C_k$ est le même que celui de $(-1)^{k+1}$.

Il en résulte que $C_{k+2} > C_k$ pour k impair, et que $C_{k+2} < C_k$ pour k pair. Il s'ensuit que:

$$C_1 < C_3 < C_5 < \dots \text{ et } C_2 > C_4 > C_6 > \dots \quad (4.109)$$

Ensuite, puisque:

$$C_k - C_{k-1} = \frac{n_k}{d_k} - \frac{n_{k-1}}{d_{k-1}} = \frac{n_kd_{k-1} - d_kn_{k-1}}{d_kd_{k-1}} = \frac{(-1)^k}{d_kd_{k-1}} \quad (4.110)$$

Donc pour k pair, nous avons $C_k > C_{k-1}$, nous en déduisons donc:

$$C_1 < C_3 < C_5 < \dots < C_6 < C_4 < C_2 \quad (4.111)$$

□C.Q.F.D.

Montrons maintenant que toute fraction continue infinie peut représenter un nombre irrationnel quelconque.

En des termes formels, si $\{q_n\}$ est une suite d'entiers tous positifs et que nous considérons $C_n = [q_1, q_2, \dots, q_n]$ alors celui-ci converge nécessairement vers un nombre réel si $n \rightarrow +\infty$.

Effectivement il n'est pas difficile d'observer (c'est assez intuitif) avec un exemple pratique que nous avons:

$$C_k - C_{k-1} \rightarrow 0 \quad (4.112)$$

lorsque $k \rightarrow +\infty$.

Maintenant, notons x un nombre réel quelconque et $q_1 = [x]$ la partie entière de ce nombre réel. Alors nous avons vu tout au début de notre étude des fractions continues que:

$$x = q_1 + \varepsilon_1 \quad (4.113)$$

Il vient donc que:

$$\varepsilon_1 = x - q_1 \quad (4.114)$$

Attardons-nous pour les nécessités du chapitre d'Acoustique sur le calcul d'une fraction continue d'un logarithme en utilisant la relation précédente!

D'abord rappelons que:

$$\frac{a}{b} = q_1 + \frac{1}{1/\varepsilon_1} = q_1 + \frac{1}{q_2 + 1/\varepsilon_2} = \dots \quad (4.115)$$

Soit (relation démontrée dans le chapitre d'Analyse Fonctionnelle):

$$x = \log_a(u) = \frac{\ln(u)}{\ln(a)} \quad (4.116)$$

avec $1 < a < u$ et $(a, u) = 1$.

Soit y_n défini par:

$$y_{-1} = u \quad y_0 = a \quad y_{n+1} = \frac{y_{n-1}}{y_n q_1} \quad (4.117)$$

Alors montrons que:

$$\varepsilon_n = \frac{\ln(y_{n-2})}{\ln(y_{n-1})} - q_n \quad (4.118)$$

En effet, pour $n = 1$ nous avons:

$$\varepsilon_1 = x - q_1 = \frac{\ln(u)}{\ln(a)} - q_1 = \frac{\ln(y_{-1})}{\ln(y_0)} - q_1 \quad (4.119)$$

pour $n = 2$ nous avons d'abord:

$$\varepsilon_1 = \frac{\ln(u)}{\ln(a)} - q_1 = \frac{\ln(u) - q_1 \ln(a)}{\ln(a)} = \frac{\ln(u) - \ln(a^{q_1})}{\ln(a)} = \frac{\ln\left(\frac{u}{a^{q_1}}\right)}{\ln(a)} = \frac{\ln\left(\frac{y_{-1}}{y_0^{q_1}}\right)}{\ln(a)} = \frac{\ln(y_1)}{\ln(a)} \quad (4.120)$$

donc:

$$\frac{1}{\varepsilon_1} = \frac{\ln(a)}{\ln(y_1)} \quad (4.121)$$

et puisque nous avons montré que:

$$\frac{1}{\varepsilon_{n-1}} = q_n + \varepsilon_n \Rightarrow \varepsilon_2 = \frac{1}{\varepsilon_1} - q_2 = \frac{\ln(a)}{\ln(y_1)} - q_2 = \frac{\ln(y_0)}{\ln(y_1)} - q_2 \quad (4.122)$$

etc... par récurrence ce qui démontre notre droit d'utiliser ce changement d'écriture.

Exemple:

Cherchons l'expression de la fraction continue de:

$$x = \log_a(u) = \log_2(3) \quad (4.123)$$

Nous savons en jouant avec la définition du logarithme que:

$$2^1 < 3 < 2^2 \quad (4.124)$$

donc:

$$\begin{aligned} \log_2(2^1) < \log_2(3) < \log_2(2^2) &\Rightarrow \log_2(2) < \log_2(3) < 2 \cdot \log_2(2) \\ \Rightarrow 1 < \frac{\log_2(3)}{\log_2(2)} < 2 &\Rightarrow 1 < \log_2(3) < 2 \end{aligned} \quad (4.125)$$

donc $q_1 = 1$. Nous avons alors:

$$\varepsilon_1 = \frac{\ln(y_{1-2})}{\ln(y_{1-1})} - q_1 = \frac{\ln(y_{-1})}{\ln(y_0)} - q_1 \quad (4.126)$$

et puisque:

$$y_{-1} = u = 3 \quad y_0 = a = 2 \quad (4.127)$$

il vient:

$$\varepsilon_1 = \frac{\ln(3)}{\ln(2)} - 1 \quad (4.128)$$

Donc nous avons le premier quotient partiel:

$$\frac{a}{b} = \log_2(3) = q_1 + \frac{1}{1/\varepsilon_1} = 1 + \frac{1}{1/\left(\frac{\ln(3)}{\ln(2)} - 1\right)} = \dots \quad (4.129)$$

Et in extenso nous avons déjà:

$$y_{0+1} = y_1 = \frac{y_{-1}}{y_0^{q_1}} = \frac{3}{2^1} = \frac{3}{2} \quad (4.130)$$

Simplifions:

$$1/\left(\frac{\ln(3)}{\ln(2)} - 1\right) = \frac{\ln(2)}{\ln\left(\frac{3}{2}\right)} = \frac{\ln(2)}{\ln(y_1)} \quad (4.131)$$

Donc le premier quotient partiel peut s'écrire:

$$\log_2(3) = 1 + \frac{1}{\frac{\ln(2)}{\ln(y_1)}} = \dots \quad (4.132)$$

et passons au deuxième quotient partiel. Nous savons déjà pour cela que:

$$1 < \frac{\ln(2)}{\ln\left(\frac{3}{2}\right)} < 2 \quad (4.133)$$

donc il est immédiat que $q_2 = 1$ et alors comme:

$$\varepsilon_n = \frac{\ln(y_{n-2})}{\ln(y_{n-1})} - q_n \quad (4.134)$$

nous avons:

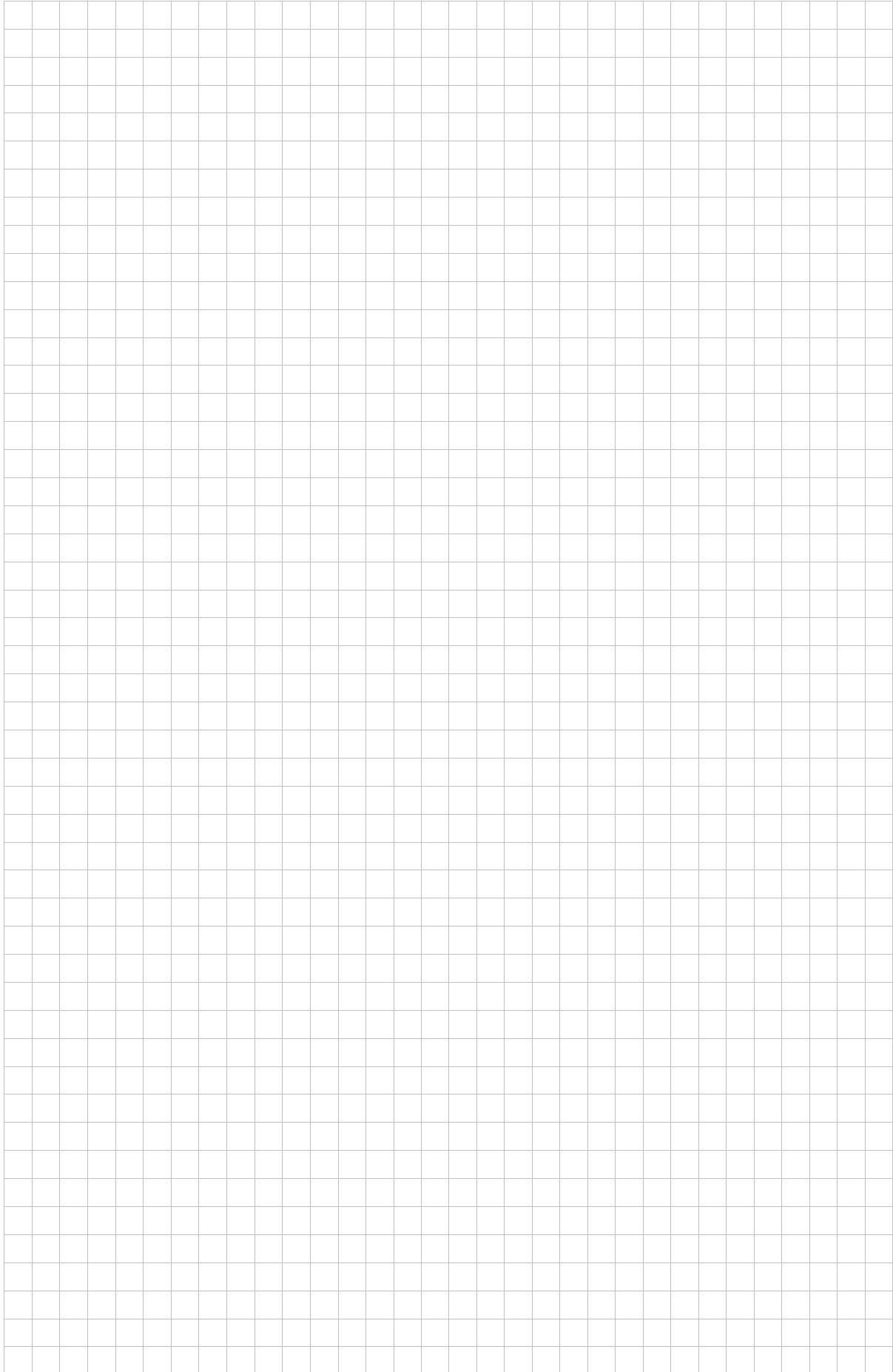
$$\varepsilon_2 = \frac{\ln(y_0)}{\ln(y_1)} - q_2 = \frac{\ln(2)}{\ln\left(\frac{3}{2}\right)} - 1 = \frac{\ln(2) - \ln\left(\frac{3}{2}\right)}{\ln\left(\frac{3}{2}\right)} = \frac{\ln\left(\frac{4}{3}\right)}{\ln\left(\frac{3}{2}\right)} \quad (4.135)$$

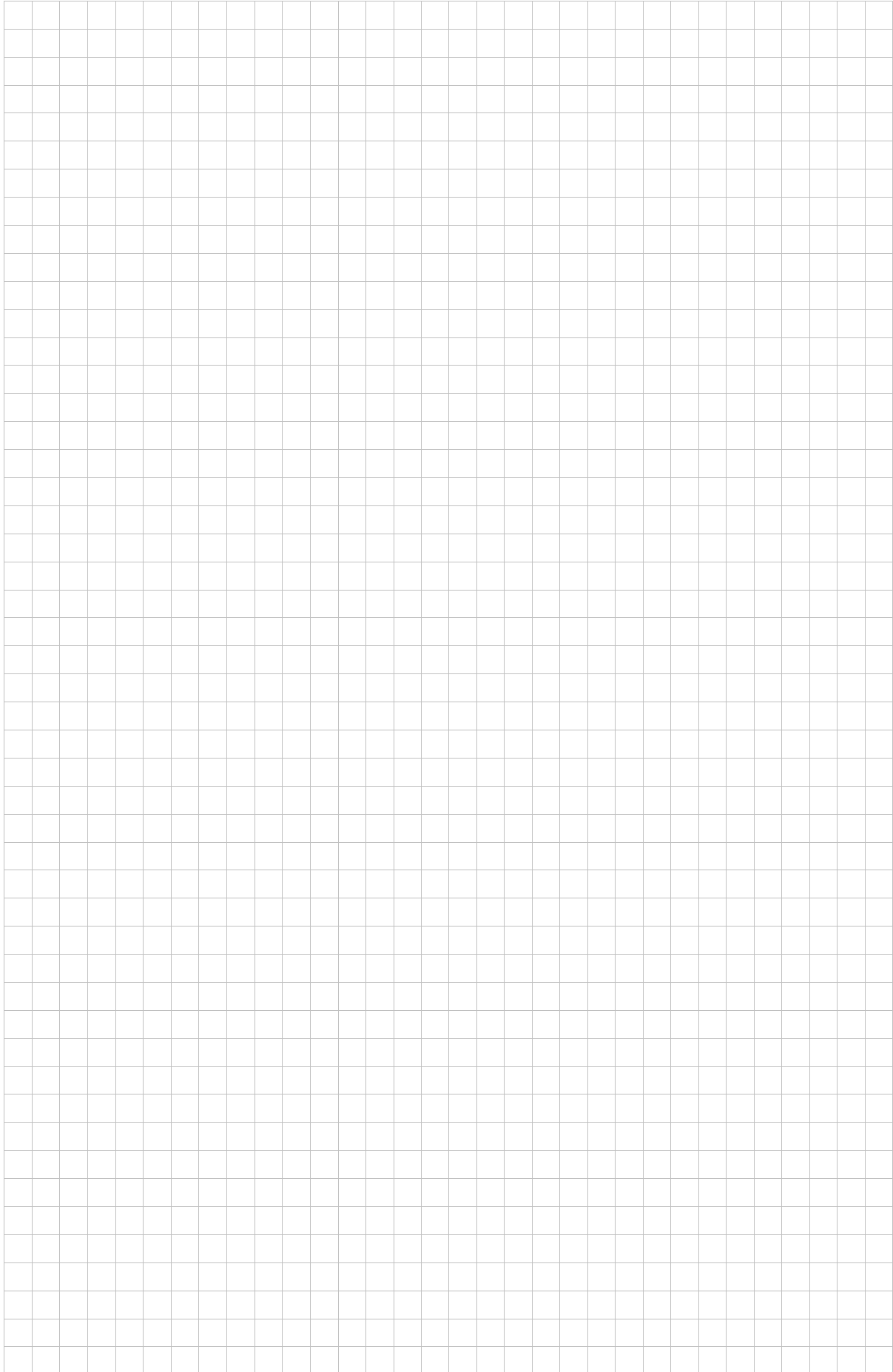
Il vient finalement:

$$\log_2(3) = q_1 + \frac{1}{q_2 + 1/\varepsilon_2} = 1 + \frac{1}{\frac{\ln\left(\frac{4}{3}\right)}{1 + 1/\frac{\ln\left(\frac{3}{2}\right)}{\ln\left(\frac{4}{3}\right)}}} = 1 + \frac{1}{\frac{\ln\left(\frac{3}{2}\right)}{1 + \frac{\ln\left(\frac{4}{3}\right)}{\ln\left(\frac{3}{2}\right)}}} \quad (4.136)$$

etc... etc.







5. THÉORIE DES ENSEMBLES

Lors de notre étude des nombres, des opérateurs, et de la théorie des nombres (dans les chapitres du même nom), nous avons assez souvent utilisé les termes "groupes", "anneaux", "corps", "homomorphisme", etc. et continuerons par la suite à le faire encore de nombreuses fois. Outre le fait que ces concepts soient d'une extrême importance, permettant de faire des démonstrations ou de construire des concepts mathématiques indispensables à l'étude de la physique théorique contemporaine (physique quantique des champs, théories des cordes, modèle standard,...), ils permettent de comprendre les composants et les propriétés de base de la mathématique et de ses opérateurs en rangeant ceux-ci par catégories distinctes. Ainsi, choisir de mettre la théorie des ensembles en tant que cinquième chapitre de ce site est un choix tout à fait discutable puisque rigoureusement c'est par là que tout commence... Cependant, nous avons besoin d'exposer quand même la théorie de la démonstration ne serait-ce que pour les notations et les méthodes dont il sera fait usage ici.

Par ailleurs, lors de l'enseignement des mathématiques modernes dans le secondaire, voire primaire (années 1970), on introduisit le langage des ensembles et l'étude préalable des relations binaires pour une approche plus rigoureuse de la notion de fonctions et d'applications (voir la définition plus loin) et de la mathématique en général.

Définition: Nous parlons de "**diagramme sagittal**" (ou de "**schéma sagittal**" du latin sagitta = flèche) pour tout schéma représentant une correspondance entre les composantes de deux ensembles reliés totalement ou partiellement par un ensemble de flèches.

Exemple:

La représentation graphique d'une fonction définie de l'ensemble $E = \{-3, -2, -1, 0, 1, 2, 3\}$ vers l'ensemble $F = \{0, 1, 2, 3, \dots, 9\}$ conduirait au diagramme sagittal ci-dessous:

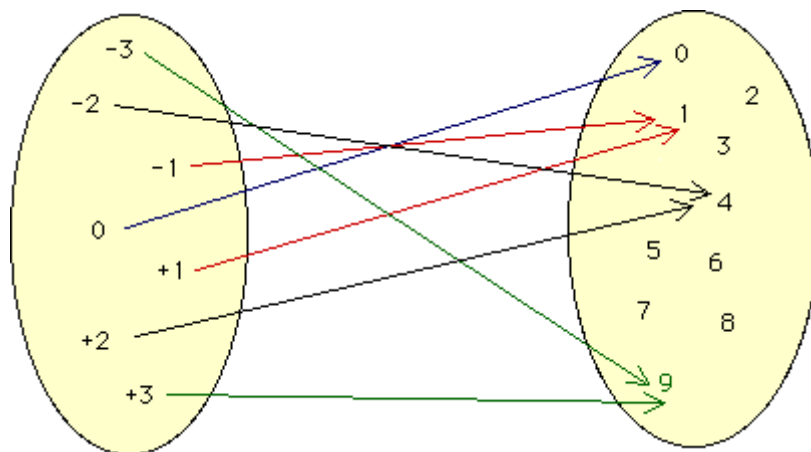


Figure: 5.1 - Fonction d'un ensemble de définition à un autre ensemble d'arrivée

Une relation de E dans E fournirait un diagramme sagittal du type:

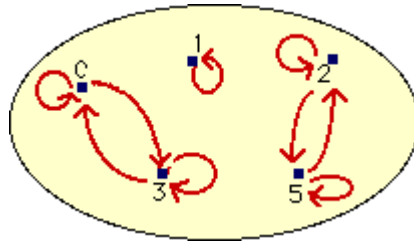


Figure: 5.2 - Fonction renvoyant dans son propre ensemble de définition

Le bouclage de chaque élément montrant une "**relation réflexive**" et la présence systématique d'une flèche retour indiquant une "**relation symétrique**".

Définition: Si l'ensemble d'arrivée est identique à l'ensemble de départ, nous disons que nous avons une "**relation binaire**".

Cependant le choix d'introduire la théorie des ensembles dans les classes d'école a une raison aussi un peu autre. Au fait, dans un souci de rigueur interne (in extenso: non liée à la réalité), une très grande partie des mathématiques a été reconstruite à l'intérieur d'un seul cadre axiomatique, dénommé donc "**théorie des ensembles**", dans le sens où chaque concept mathématique (autrefois indépendant des autres) est ramené à une définition dont tous les constituants logiques proviennent de ce même cadre: elle est considérée comme fondamentale. Ainsi, la rigueur d'un raisonnement effectué au sein de la théorie des ensembles est garantie par le fait que le cadre est "non-contradictoire" ou "consistant". Voyons les définitions qui construisent ce cadre.

Définitions:

D1. Nous appelons "**ensemble**" toute liste, collection ou rassemblement d'objets bien définis, explicitement ou implicitement.

D2. Un "**Univers**" U est un objet dont les constituants sont des ensembles.

Il faut noter que ce que les mathématiciens appellent "univers" n'est pas un ensemble! En fait il s'agit d'un modèle qui satisfait aux axiomes des ensembles.

Effectivement, nous verrons que nous ne pouvons pas parler de l'ensemble de tous les ensembles (ce n'est pas un ensemble), pour désigner l'objet qui est constitué de tous les ensembles ainsi, nous parlons d'univers.

D3. Nous appelons "**éléments**" ou "**membres de l'ensemble**" les objets appartenant à l'ensemble et nous notons:

$$p \in A \quad (5.1)$$

si p est un élément de l'ensemble A et dans le cas contraire:

$$p \notin A \quad (5.2)$$

Si B est une "**partie**" de A , ou sous-ensemble de A , nous notons cela:

$$B \subset A \text{ ou } A \supset B \quad (5.3)$$

dès lors, si pour tout:

$$\forall x, x \in B \Rightarrow x \in A \quad (5.4)$$

Nous identifions également un ensemble soit en listant ses éléments (pas toujours forcément dénombrables par ailleurs!), soit en donnant la définition de ses éléments (nombres pairs, impairs, diviseurs entiers de..., etc.).

Exemples:

E1. $A = \{1, 2, 3\}$

E2. $X = \{x \mid x \text{ est un entier positif}\}$

D3. Nous pouvons munir les ensembles d'un certain nombre de relations qui permettent de comparer leurs éléments (c'est utile parfois...) ou de comparer certaines de leurs propriétés. Ces relations sont appelées "**relations de comparaisons**" ou "**relations d'ordre**" (cf. chapitre sur les Opérateurs).

Remarques:

R1. La structure d'ensemble ordonnée a été mise en place à la base dans le cadre de la théorie des Nombres par Cantor et Dedekind.

R2. Comme nous l'avons démontré dans le chapitre sur les Opérateurs, $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}$ sont totalement ordonnés par les relations usuelles $\leq, \geq$. La relation $<$, souvent dite "**d'ordre strict**", n'est pas une relation d'ordre car non réflexive et non antisymétrique (cf. chapitre sur les Opérateurs). Par exemple, dans $\mathbb{N}$, la relation " a divise b ", souvent notée par le symbole " $\mid$ ", est un ordre partiel.

R3. Si R est un ordre sur E et F est une partie de E , la restriction à F de la relation R est un ordre sur F , dit "**ordre induit par R dans F** ".

R4. Si R est un ordre sur E , la relation R' définie par:

$$xR'y \Leftrightarrow yRx \quad (5.5)$$

est un ordre sur E , dit "**ordre réciproque**" de R . L'ordre réciproque de l'ordre usuel $\leq$ est l'ordre noté $\geq$ ainsi que l'ordre réciproque de l'ordre " a divise b " dans $\mathbb{N}$ est l'ordre " b est multiple de a ".

L'ensemble est l'être mathématique de base, dont l'existence est posée: il n'est pas défini en tant que tel, mais par ses propriétés, données par les axiomes. Il fait appel à une procédure humaine: une sorte de fonction de catégorisation, qui permet à la pensée de distinguer plusieurs éléments qualifiés d'indépendants.

Nous pouvons démontrer à partir de ces concepts, que le nombre de sous-ensembles d'un ensemble de cardinal n est 2^n .

Démonstration:

Il y a d'abord l'ensemble vide $\emptyset$, soit 0 élément choisi parmi n , in extenso C_0^n (notation du coefficient binomial non conforme à la norme ISO 31-11!) conformément à ce que nous avons vu dans le chapitre

de Probabilités:

$$C_k^n = \binom{n}{k} = \frac{A_n^k}{k!} = \frac{k!}{k!(n-k)!} \quad (5.6)$$

et ainsi de suite...

Le nombre de sous-ensembles (cardinal) de E correspond donc à la sommation de tous les coefficients binomiaux:

$$\text{card}(P(E)) = \sum_{k=0}^n C_k^n \quad (5.7)$$

Or, nous avons (cf. chapitre de Calcul Algébrique):

$$(x+y)^n = \sum_{k=0}^n C_k^n x^k y^{n-k} \quad (5.8)$$

Donc:

$$\text{card}(P(E)) = \sum_{k=0}^n C_k^n 1^k 1^{n-k} = (1+1)^n = 2^n \quad (5.9)$$

□C.Q.F.D.

Exemple:

Considérons l'ensemble $S = \{x_1, x_2, x_3\}$, nous avons l'ensemble des parties $P(S)$ constitué par:

- "L'ensemble vide": $\{\} = \emptyset$
- Les "singletons": $\{x_1\}, \{x_2\}, \{x_3\}$
- Les "duets": $\{x_1, x_2\}, \{x_1, x_3\}, \{x_2, x_3\}$
- Lui-même: $\{x_1, x_2, x_3\}$

Tel que:

$$P(S) = \{\emptyset, \{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\}, \{x_2, x_3\}, \{x_1, x_2, x_3\}\} \quad (5.10)$$

Ce qui fait bien 8 éléments!

Remarque: L'ordre dans lequel sont différenciés les éléments ne rentre pas en compte lors du comptage des parties de l'ensemble de départ.

En mathématique appliquée, nous travaillerons presque exclusivement avec des ensembles de nombres.

Nous nous restreindrons donc à l'étude des définitions et propriétés de ces derniers.

Maintenant, formalisons les concepts de base permettant de travailler avec les ensembles les plus courants que nous rencontrons dans les cursus scolaires de base.

1. AXIOMATIQUE DE ZERMELO-FRAENKEL

L'axiomatique de Zermelo-Fraenkel, abrégée "**axiomatique ZF-C**", présentée ci-dessous a été formulée par Ernst Zermelo puis précisée par Adolf Abraham Fraenkel au début du 20ème siècle et complétée par l'axiome du choix (d'où le C majuscule dans ZF-C). Elle est considérée comme la plus naturelle dans le cadre de la théorie des ensembles.

Remarque: Il existe bien d'autres axiomatiques, basées sur le concept plus général de "classe", comme celle développée par von Neumann, Bernays et Gödel (pour les notations, voir le chapitre traitant de la Théorie De La Démonstration).

Strictement et techniquement parlant, les axiomes de ZF sont des énoncés du calcul des prédicats du premier ordre (cf. chapitre de Théorie De La Démonstration) égalitaire dans un langage ayant un seul symbole primitif pour l'appartenance (relation binaire). Ce qui suit doit donc seulement être perçu comme une tentative d'exprimer en français la signification attendue de ces axiomes.

A1. Axiome d'extensionnalité:

Deux ensembles sont égaux si, et seulement si ils ont les mêmes éléments. C'est ce que nous notons:

$$A = B \Leftrightarrow (\forall x \in A, x \in B) \wedge (\forall x \in B, x \in A) \quad (5.11)$$

Donc A et B sont égaux si tout élément x de A appartient aussi à B et tout élément x de B appartient aussi à A .

A2. Axiome de l'ensemble vide:

L'ensemble vide existe, nous le notons:

$$\emptyset \quad (5.12)$$

et il n'a aucun élément, son cardinal vaut donc 0.

En réalité cet axiome peut être déduit à partir d'un autre axiome que nous verrons un peu plus loin mais il est pratique à introduire en tant que tel par commodité pédagogique dans les petites classes.

A3. Axiome de la paire:

Si A et B sont deux ensembles, alors, il existe un ensemble C contenant A et B et eux seuls comme éléments. Cet ensemble C se note alors $\{A, B\}$.

Du point de vue des ensembles considérés comme des éléments cela donne:

$$\forall A \forall B \exists C \quad (A \in C \wedge B \in C) \quad (5.13)$$

Cet axiome montre aussi l'existence du "**singleton**" (single=seul) d'un ensemble noté:

$$\{X\} \quad (5.14)$$

qui est un ensemble dont le seul élément est X (donc de cardinal unitaire). Il suffit pour cela d'appliquer l'axiome en posant l'égalité entre A et B .

A4. Axiome de la somme (dit aussi "axiome de l'union" ou encore "axiome de la réunion"):

Cet axiome permet de construire la réunion d'un ensemble; dit de façon plus commune: la réunion d'une famille quelconque d'un ensemble, est un ensemble.

Autrement dit, il existe pour tout ensemble quelconque, un ensemble qui contient exactement les éléments de tout élément de l'ensemble. La formalisation (peu intuitive) de cet axiome est la suivante:

$$\forall A \exists B \forall C \quad (C \in B \Leftrightarrow \exists D (D \in A \wedge C \in D)) \quad (5.15)$$

C'est-à-dire qu'étant donné un ensemble quelconque A , il existe un ensemble B tel que, pour tout ensemble C quelconque, C est élément de B si et seulement s'il existe un ensemble D tel que D soit un élément de A et que C soit un élément de D .

Un petit exemple particulier ne fera peut-être pas de mal...:

$$\begin{aligned} A &= \{x_1, x_2, x_3\} \\ B &= \bigcup A = \left\{ \underbrace{\overbrace{\{x_1\}}^C}_D, \underbrace{\overbrace{\{x_2\}}^C}_D, \underbrace{\overbrace{\{x_3\}}^C}_D \right\} \end{aligned} \quad (5.16)$$

Nous voyons que conformément à l'axiome, chaque D est un élément de A et que chaque C est un élément de D et ce pour chaque C appartenant à B . De même si nous prenons:

$$\begin{aligned} A &= \{x_1, x_2, x_3\} \\ B &= \bigcup A = \left\{ \underbrace{\overbrace{\{x_1, x_2\}}^C}_D, \underbrace{\overbrace{\{x_3\}}^C}_D \right\} \end{aligned} \quad (5.17)$$

L'ensemble B est donc noté:

$$\bigcup A \quad (5.18)$$

ou:

$$\bigcup_{x \in A} x \quad (5.19)$$

A5. Axiome des parties (dit aussi "axiome de l'ensemble des parties"):

Il exprime que pour tout ensemble A , l'ensemble de ses parties $P(A)$ existe.

Donc à tout ensemble A , nous pouvons associer un ensemble B qui contient exactement les parties (in extenso les sous-ensembles) C du premier:

$$\forall A \exists B \forall C \quad (C \in B \Leftrightarrow C \subset A) \quad (5.20)$$

Nous avons vu un tel exemple déjà plus haut avec $S = \{x_1, x_2, x_3\}$:

$$P(S) = \{\emptyset, \{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\}, \{x_2, x_3\}, \{x_1, x_2, x_3\}\} \quad (5.21)$$

A6. Axiome de l'infini:

Cet axiome exprime le fait qu'il existe un ensemble infini.

Pour le formaliser, nous disons qu'il existe un ensemble, dit "**ensemble autosuccesseur**" K contenant $\emptyset$ (l'ensemble vide) tel que si x appartient à K , alors $x \cup \{x\}$ appartient également à K :

$$K \text{ est autosuccesseur: } \Leftrightarrow (\emptyset \in K) \wedge (x \in K \Rightarrow x \cup \{x\} \in K) \quad (5.22)$$

Cet axiome exprime donc que l'ensemble des entiers existe. Effectivement, $\mathbb{N}$ est ainsi le plus petit ensemble autosuccesseur, au sens de l'inclusion $\mathbb{N} = \{\emptyset, \{\emptyset, \{\emptyset, \dots\}\}\}$ et par convention nous notons (où nous construisons l'ensemble des naturels):

$$\begin{aligned} 0 &= \emptyset \\ 1 &= \{\emptyset\} \\ 2 &= \{\emptyset, \{\emptyset\}\} \\ &\dots \end{aligned} \quad (5.23)$$

A7. Axiome de régularité (dit aussi "axiome de fondation"):

Le but principal de cet axiome est d'éliminer la possibilité d'avoir A comme élément de lui-même.

Ainsi, pour tout ensemble non vide A , il existe un ensemble B qui est élément de A tel qu'aucun élément de A ne soit élément de B (il faut bien différencier le niveau du langage utilisé, un ensemble et ses éléments n'ont pas le même statut) ce que nous notons:

$$A \neq \emptyset \Rightarrow \exists B \in A, A \cap B = \emptyset \quad (5.24)$$

En conséquence:

$$\forall A, \quad A \notin A \quad (5.25)$$

Démonstration:

En effet, soit A un ensemble tel que $A \in A$. Considérons le singleton $\{A\}$, ensemble dont le seul élément est A . D'après l'axiome de fondation, nous devons avoir un élément de ce singleton qui n'a aucun élément en commun avec lui. Mais le seul élément possible est A lui-même, c'est-à-dire que nous devons avoir:

$$A \cap \{A\} = \emptyset \quad (5.26)$$

Or par hypothèse, $A \in A$ et par construction $A \in \{A\}$. Donc $A \in A \cap \{A\}$, ce qui contredit l'assertion précédente. Donc:

$$A \notin A \quad (5.27)$$

□C.Q.F.D.

A8. Axiome de remplacement (dit aussi "schéma de remplacement"):

Cet axiome exprime le fait que si une formule f est une fonctionnelle alors pour tout ensemble A , il existe un ensemble B constitué exactement des images des éléments A par cette fonction.

Soient, de manière un peu plus formelle, l'ensemble A d'éléments a et la relation binaire f (qui est donc en toute généralité une fonctionnelle), il existe un ensemble B constitué des éléments b tel que $f(a,b)$ soit vraie. Si f est une fonction où b est non libre cela signifie alors que:

$$b = f(a) \text{ et } B = f(A) \quad (5.28)$$

De manière technique nous écrivons cet axiome sous la forme:

$$\forall A \forall a \in A \quad \exists! b f(a,b) \Rightarrow \exists B \forall a \in A \quad \exists b \in B f(a,b) \quad (5.29)$$

Donc pour tout ensemble A et tout élément qu'il contient, il existe un et un seul b défini par la fonctionnelle f tel qu'il existe un ensemble B où pour tout élément a appartenant à l'ensemble A il existe un b appartenant à l'ensemble B défini par la fonctionnelle f .

Voyons un exemple avec le prédicat binaire suivant qui pour la valeur de tout a de A détermine la valeur de tout b de B :

$$P(a,b) = (a = 1 \wedge b = 2) \vee (a = 3 \wedge b = 4) \quad (5.30)$$

Donc de la connaissance que a vaut 1 nous en dérivons que b vaut 2 et de manière similaire (in extenso par remplacement) si a vaut 3, nous en dérivons que b vaut 4.

Nous voyons bien au travers de ce petit exemple la relation forte qu'il y a à considérer le prédicat P comme une fonction naïve! Par ailleurs, comme il y a une infinité possible de fonctions f , le schéma de remplacement est considéré comme une infinité d'axiomes.

A9. Axiome de sélection (dit aussi "schéma de compréhension"):

Cet axiome exprime simplement que pour tout ensemble A et toute propriété P exprimable dans le langage de la théorie des ensembles, l'ensemble des éléments de A qui satisfont la propriété P existe.

Donc de manière plus formelle, à tout ensemble A et toute condition ou proposition $P(x)$, il correspond un ensemble B dont les éléments sont exactement les éléments x de A pour lesquels $P(x)$ est vraie. C'est ce que nous notons:

$$B = \{x \in A : P(x)\} \quad (5.31)$$

De manière plus complète et rigoureuse nous avons en réalité pour toute fonctionnelle f ne comportant

pas a comme variable libre:

$$\forall A \exists B \forall a \quad (a \in B \Leftrightarrow a \in A \wedge f) \quad (5.32)$$

C'est typiquement l'axiome qui nous sert à construire l'ensemble des nombres pairs:

$$\{a \in \mathbb{N} \mid \exists b \in \mathbb{N}, a = 2b\} \quad (5.33)$$

ou à démontrer l'existence de l'ensemble vide (et qui rend caduc l'axiome de l'ensemble vide) car il suffit de poser qu'il existe un ensemble satisfaisant la propriété:

$$A \neq A \quad (5.34)$$

et ce quel que soit l'ensemble A . Et seulement l'ensemble vide satisfait cette propriété de par l'axiome de sélection.

Le respect des conditions très strictes de cet axiome permet d'éliminer les paradoxes de la "**théorie naïve des ensembles**", comme le paradoxe de Russel ou le paradoxe de Cantor qui ont invalidé la théorie naïve des ensembles.

Considérons par exemple l'ensemble R de Russel de tous les ensembles qui ne s'auto-contiennent pas (notez bien que nous donnons une propriété de R sans expliciter quel est cet ensemble):

$$R = \{E : E \notin E\} \quad (5.35)$$

Le problème est de savoir si R se contient ou non. Si $R \in R$, alors, R s'auto-contient, et, par définition $R \notin R$ et inversement. Chaque possibilité est donc contradictoire.

Si maintenant nous désignons par C l'ensemble de tous les ensembles (l'Universel de Cantor), nous avons en particulier:

$$P(C) \in C \quad (5.36)$$

ce qui est impossible (i.e. par exemple avec la puissance du continu de l'ensemble de réels), d'après le théorème de Cantor.

Ces "paradoxes" (ou "antinomies syntaxiques") proviennent d'un non-respect des conditions d'application de l'axiome de sélection: pour définir E (dans l'exemple de Russel), il doit exister une proposition P qui porte sur l'ensemble R , qui doit être explicitée. La proposition définissant l'ensemble de Russel ou celui de Cantor n'indique pas quel est l'ensemble E . Elle est donc invalide!

Un exemple fort sympathique et fort connu (c'est la raison pour laquelle nous le présentons) permet de mieux comprendre (il s'agit du paradoxe de Russel dont nous avons déjà parlé plus longuement dans le chapitre de Théorie De La Démonstration):

Un jeune étudiant se rendit un jour chez son barbier. Il engagea la conversation et lui demanda s'il avait de nombreux concurrents dans sa jolie cité. De manière apparemment innocente, le barbier lui répondit: "Je n'ai aucune concurrence. En effet, de tous les hommes de la cité, je ne rase évidemment pas ceux qui se rasent eux-mêmes, mais j'ai le bonheur de raser tous ceux qui ne se rasent pas eux-mêmes."

En quoi donc, une telle affirmation si simple put-elle mettre en défaut la logique de notre jeune étudiant si malin ?

La réponse est en effet innocente, jusqu'au moment où nous décidons de l'appliquer au cas du barbier:

Se rase-t-il lui-même, oui ou non?

Supposons qu'il se rase lui-même: il entre dans la catégorie de ceux qui se rasent eux-mêmes, dont le barbier a précisé qu'il ne les rasait évidemment pas. Donc il ne rase pas lui-même.

Très bien! Supposons alors qu'il ne se rase pas lui-même: il entre alors dans la catégorie de ceux qui ne se rasent pas eux-mêmes, dont le barbier a précisé qu'il les rasait tous. Donc il se rase lui-même.

Finalement, ce malheureux barbier est dans une position étrange: s'il se rase lui-même, il ne se rase pas, et s'il ne se rase pas lui-même, il se rase. Cette logique est autodestructrice, stupidement contradictoire, rationnellement irrationnelle.

Vient alors l'axiome de sélection: Nous excluons le barbier de l'ensemble des personnes auxquelles s'applique la déclaration. Car en réalité, le problème vient du fait que le barbier est un élément de l'ensemble de tous les hommes de la cité. Ainsi, ce qui s'applique à tous les hommes ne s'applique pas au cas individuel du barbier.

A10. Axiome du choix:

Étant donné un ensemble A d'ensembles non vides mutuellement disjoints, il existe un ensemble B (l'ensemble de choix pour A) contenant exactement un élément pour chaque membre de A .

Indiquons cependant que la question de l'axiomatisation et donc des fondements se trouva quand même ébranlée de deux questions à l'époque de leur construction: quels axiomes valides doivent être choisis et dans un système d'axiomes la mathématique est-elle cohérente (ne risque-t-on pas de voir apparaître une contradiction)?

La première question fut soulevée d'abord par l'hypothèse du continu: si nous pouvons mettre deux ensembles de nombres en correspondance terme à terme, ils ont le même nombre d'éléments (cardinal). Nous pouvons mettre en correspondance les entiers avec les rationnels comme nous l'avons démontré dans le chapitre sur les Nombres, ils ont donc même cardinal, nous ne pouvons par contre mettre en correspondance les entiers avec les réels. La question est alors de savoir s'il y a un ensemble dont le nombre d'éléments serait situé entre les deux ou pas? Cette question est importante pour construire la théorie classique de l'analyse et les mathématiciens choisissent en général de dire qu'il n'y en a pas, mais nous pouvons aussi dire le contraire.

En fait l'hypothèse du continu est liée de manière plus profonde à l'axiome du choix qui peut aussi être formulé de la manière suivante: si C est une collection d'ensembles non vides alors nous pouvons choisir un élément de chaque ensemble de la collection. Si C a un nombre fini d'éléments ou un nombre dénombrable d'éléments, l'axiome semble assez évident: nous pouvons ranger les ensembles de C en les numérotant, et le choix d'un élément dans chaque ensemble est simple. Là où ça se complique c'est lorsque l'ensemble C a la puissance du continu: comment choisir des éléments s'il n'y a pas la possibilité de les numérotter?

Finalement en 1938 Kurt Gödel montre que la théorie des ensembles est cohérente sans l'axiome du

choix et sans l'hypothèse du continu aussi bien qu'avec! Et pour clore tout ça Paul Cohen montre en 1963 que l'axiome du choix et l'hypothèse du continu ne sont pas liés.

1.1. CARDINAUX

Définition: Des ensembles sont dits "**équipotents**" s'il existe une bijection (correspondance biunivoque) entre ces ensembles. Nous disons qu'ils ont alors même "**cardinal**" que la norme ISO 3111 préconise d'écrire *card* mais sur le présent site internet nous utiliserons tantôt *card* que *Card*.

Ainsi, plus rigoureusement, un cardinal (qui quantifie le nombre d'éléments contenus dans l'ensemble) est une classe d'équivalence (cf. [chapitre sur les Opérateurs](#)) pour la relation d'équipotence.

Remarque: Cantor est le principal créateur de la théorie des ensembles, sous une forme que nous qualifions aujourd'hui de "**théorie naïve des ensembles**". Mais, à côté de considérations élémentaires, sa théorie comportait des niveaux d'abstraction élevés. La vraie nouveauté de la théorie de Cantor, c'est qu'elle permet de parler de l'infini. Par exemple, une idée importante de Cantor a justement été de définir l'équipotence.

Si nous écrivons $c_1 = c_2$ en tant qu'égalité de cardinaux, nous entendons alors par là qu'il existe deux ensembles équipotents A et B tels que:

$$c_1 = \text{Card}(A) \text{ et } c_2 = \text{Card}(B) \quad (5.37)$$

Les cardinaux peuvent donc être comparés. L'ordre ainsi défini est une relation d'ordre total (cf. [chapitre sur les Opérateurs](#)) entre les cardinaux (la preuve que la relation d'ordre est totale utilise l'axiome du Choix et la preuve qu'elle soit antisymétrique est connue sous le nom de théorème de Cantor-Bernstein que nous démontrons d'ailleurs plus bas).

Dire que $c_1 < c_2$ signifie dans un vocabulaire simple que A est équipotent à une partie propre de B , mais B n'est équipotent à aucune partie propre de A . Les mathématiciens diraient que le $\text{Card}(A)$ est plus petit ou égal au $\text{Card}(B)$ s'il existe une injection de A dans B .

Nous avons vu lors de notre étude des nombres, en particulier des nombres transfinis, qu'un ensemble équipotent (ou en bijection) à $\mathbb{N}$ était dit "**ensemble dénombrable**".

Voyons cette notion un petit peu plus dans les détails:

Soit A un ensemble, s'il existe un entier n tel qu'il y ait au moins à chaque élément de A un correspondant dans l'ensemble $\{1, 2, \dots, n\}$ (au fait rigoureusement il s'agit d'une bijection... concept que nous définirons plus tard) nous disons alors que le cardinal de A , noté $\text{Card}(A)$ ou $\text{card}(A)$, est un "**cardinal fini**" et vaut n .

Dans le cas contraire, nous disons que l'ensemble A est de "**cardinal infini**" et nous posons:

$$\text{Card}(A) = +\infty \quad (5.38)$$

Un ensemble A est donc "**dénombrable**" s'il existe une bijection entre A et $\mathbb{N}$. Un ensemble de nombre A est "**au plus dénombrable**" s'il existe une bijection entre A et une partie $\mathbb{N}$. Un ensemble au plus dénombrable est donc soit de cardinal fini, soit dénombrable.

Nous vérifions dès lors les propositions suivantes:

P1. Une partie d'un ensemble dénombrable est au plus dénombrable.

P2. Un ensemble contenant un ensemble non-dénombrable n'est lui aussi pas dénombrable.

P3. Le produit de deux ensembles dénombrables est dénombrable.

Remarque: Nous pouvons restreindre un ensemble de nombres par rapport à l'élément nul et aux éléments négatifs ou positifs qu'il contient et dès lors nous notons (exemple pour l'ensemble des réels):

$$\begin{aligned}\mathbb{R}^* &= \mathbb{R} \setminus \{0\} \\ \mathbb{R}_+ &= \{x \mid x \in \mathbb{R} \text{ et } x \geq 0\} \\ \mathbb{R}_- &= \{x \mid x \in \mathbb{R} \text{ et } x \leq 0\} \quad (5.39) \\ \mathbb{R}_+^* &= \mathbb{R}_+ \setminus \{0\} \\ \mathbb{R}_-^* &= \mathbb{R}_- \setminus \{0\}\end{aligned}$$

Ces notions étant analogues pour $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$ (l'ensemble des nombres complexe n'étant pas ordonné, la deuxième et troisième ligne ne s'y applique pas).

Donc tout sous-ensemble infini de $\mathbb{N}$ est équipotent à $\mathbb{N}$ lui-même, ce qui peut sembler contre-intuitif au premier abord...!

En particulier, il y a autant d'entiers naturels pairs que d'entiers naturels quelconques (utiliser la bijection $f(n) = 2n$) de $\mathbb{N}$ vers P , où P désigne l'ensemble des entiers naturels pairs), autant d'entiers relatifs que d'entiers naturels, autant d'entiers relatifs que de nombres rationnels (voir le chapitre traitant des nombres pour les démonstrations).

Nous pouvons donc écrire:

$$\text{Card}(\mathbb{N}) = \text{Card}(\mathbb{Z}) = \text{Card}(\mathbb{Q}) = \aleph_0 \quad (5.40)$$

et plus généralement, toute partie infinie de $\mathbb{Q}$ est dénombrable.

Un résultat important: tout ensemble infini possède donc une partie infinie dénombrable.

Puisque nous avons démontré dans le chapitre traitant des nombres que l'ensemble des réels avait la "**puissance du continu**" et que l'ensemble des nombres naturels était de cardinal transfini $\aleph_0$, Cantor souleva la question s'il existait un cardinal transfini entre $\aleph_0$ et le cardinal de $\mathbb{R}$? Autrement dit, nous avons donc une quantité infinie de nombres entiers, et une quantité encore plus grande de nombres réels. Alors, existe-t-il un infini qui soit à la fois plus grand que celui des entiers et plus petit que celui des nombres réels?

Le problème se posa en notant bien évidemment $\aleph_0$ le cardinal de $\mathbb{N}$ et (nouveau) $\aleph_1$ le cardinal de $\mathbb{R}$ et en proposant de démontrer ou de contredire que:

$$\aleph_1 = 2^{\aleph_0} \quad (5.41)$$

selon la loi combinatoire qui donne le nombre d'éléments de l'ensemble que l'on peut obtenir à partir de tous les sous-ensembles d'un ensemble (tel que nous l'avons démontré précédemment).

Le reste de sa vie, Cantor essaya, en vain, de démontrer ce résultat que l'on nomma "**l'hypothèse du continu**". Il n'y réussit pas et sombra dans la folie. En 1900, au congrès international des mathématiciens, Hilbert estima qu'il s'agissait là d'un des 23 problèmes majeurs qui devraient être résolus au 20ème siècle.

Ce problème se résout d'une façon assez étonnante. D'abord, en 1938, un des plus grands logiciens du 20ème siècle, Kurt Gödel, démontra que l'hypothèse de Cantor n'était pas réfutable, c'est-à-dire qu'on ne pourrait jamais démontrer qu'elle était fausse. Puis en 1963, le mathématicien Paul Cohen boucla la boucle. Il démontra qu'on ne pourrait jamais non plus démontrer qu'elle était vraie!!! Nous pouvons conclure à juste raison que Cantor avait perdu la raison à chercher à démontrer un problème qui ne pouvait pas l'être.

1.2. PRODUIT CARTÉSIEN

Si E et F sont deux ensembles, nous appelons "**produit cartésien de E par F** " l'ensemble noté $E \times F$ (à ne pas confondre avec le produit vectoriel) formé de tous les couples possibles (e, f) où e est un élément de E et f un élément de F .

Autrement écrit:

$$E \times F = \{(e, f) \mid e \in E \wedge f \in F\} \quad (5.42)$$

Nous remarquons facilement que $E \times F$ et $F \times E$ ne sont pas les mêmes ensembles (sauf bien sûr si $E = F$).

Nous notons le produit cartésien de E par lui-même:

$$E \times E = E^2 \quad (5.43)$$

et nous disons alors E^2 est "**l'ensemble des couples d'éléments de E** ".

Nous pouvons effectuer le produit cartésien d'une suite d'ensembles $E_1 \times E_2 \times \dots \times E_n$ et ainsi obtenir l'ensemble des n -uplets $(e_1, e_2, \dots, e_n)$ où $e_1 \in E_1, e_2 \in E_2, \dots, e_n \in E_n$.

Dans le cas où tous les ensembles E_i sont identiques à E , le produit cartésien $E_1 \times E_2 \times \dots \times E_n$ se note bien évidemment E^n . Nous disons alors que E^n est "**l'ensemble des n -uplets d'éléments de E** ".

Si E et F sont finis alors le produit cartésien $E \times F$ est fini. De plus:

$$\text{Card}(E \times F) = \text{Card}(E) \cdot \text{Card}(F) \quad (5.44)$$

De là, nous voyons que si les ensembles $E_1, E_2, \dots, E_n$ sont finis alors le produit cartésien $E_1 \times E_2 \times \dots \times E_n$ est aussi fini et nous avons:

$$\text{Card}(E_1 \times E_2 \times \dots \times E_n) = \prod_{i=1}^n \text{Card}(E_i) \quad (5.45)$$

En particulier, $\text{Card}(E^n) = [\text{Card}(E)]^n$ si E est un ensemble fini.

Exemples:

E1. Si $\mathbb{R}$ est l'ensemble des nombres réels, $\mathbb{R}^2$ est alors l'ensemble des couples de réels. Dans le plan rapporté à un repère, tout point M admet des coordonnées qui sont un élément de $\mathbb{R}^2$.

E2. Lorsque nous lançons deux dés dont les faces sont numérotées de 1 à 6, chaque dé peut être symbolisé par l'ensemble $E = \{1, 2, 3, 4, 5, 6\}$. Le résultat d'un lancer est alors un élément de $E \times E$. Le cardinal de $E \times E$ est alors 36. Il y a donc 36 résultats possibles quand nous lançons 2 dés dont les faces sont numérotées de 1 à 6.

Remarque: La théorie de base des ensembles ainsi que le concept de cardinal sont à la base théorique des logiciels de bases de données relationnelles.

1.3. BORNES

Soit M un ensemble de nombres quelconques de façon à ce que $M \subset \mathbb{R}$ (exemple particulier mais fréquent) nous avons comme définitions:

D1. $x \in \mathbb{R}$ est appelé "**borne supérieure**" ou "**majorant**" de l'ensemble M , si $x \geq m$ pour $\forall m \in M$. Inversement, nous parlons de "**borne inférieure**" ou de "**minorant**" (il ne faut donc pas confondre le concept de borne avec le concept d'intervalle!).

D2. Soit $M \subset \mathbb{R}, M \neq \emptyset$. $x \in \mathbb{R}$ est appelé "**plus petite borne supérieure**" noté:

$$x = \sup M \quad (5.46)$$

de M si x est une borne supérieure de M et si pour toute borne supérieure $y \in \mathbb{R}$ nous avons $x \leq y$. Inversement, nous parlons de "**plus petite borne inférieure**" que nous notons:

$$x = \inf M \quad (5.47)$$

Les définitions sont équivalentes dans le cadre de l'analyse fonctionnelle (voir chapitre du même nom) puisque les fonctions sont définies sur des ensembles.

Effectivement, soit f une fonction dont le domaine de définition I balaie tout $\mathbb{R}$. Ce que nous notons $f: E \rightarrow \mathbb{R}$ et soit $x_0 \in \mathbb{R}$.

D1. Nous disons que f présente un "**maximum global**" en x_0 si:

$$\forall x \in E, f(x) \leq f(x_0) \quad (5.48)$$

D2. Nous disons que f présente un "**minimum global**" en x_0 si:

$$\forall x \in E, f(x) \geq f(x_0) \quad (5.49)$$

Dans chacun de ces deux cas, nous disons que f présente un "**extremum global**" en x_0 (c'est un concept que nous retrouverons souvent en mécanique analytique!).

D3. f est "**majorée**" s'il existe un réel M tel que $\forall x \in I, f(x) \leq M$. Dans ce cas, la fonction possède une borne supérieure de f sur son domaine de définition I notée traditionnellement:

$$\sup_I f \quad (5.50)$$

et elle représente donc la plus petite borne supérieure (le plus petit majorant).

D4. f est "**minorée**" s'il existe un réel M tel que $\forall x \in I, f(x) \geq M$. Dans ce cas, la fonction possède une borne inférieure de f sur son domaine de définition I notée traditionnellement:

$$\inf_I f \quad (5.51)$$

et elle représente la plus grande borne inférieure (le plus grand minorant).

D5. Nous disons que f est "**bornée**" si elle est à la fois majorée et minorée (c'est le cas des fonctions trigonométriques).

2. OPÉRATIONS ENSEMBLISTES

Nous pouvons construire à partir d'au moins trois ensembles A, B, C , l'ensemble des opérations (dont nous devons les notations à Dedekind) existant dans la théorie des ensembles (très utiles dans l'étude des probabilités et statistiques).

Remarque: Certaines des notations présentes ci-dessous se retrouveront fréquemment dans des théorèmes complexes, il est donc nécessaire de bien comprendre de quoi il en retourne.

Ainsi, nous pouvons construire les opérations ensemblistes suivantes:

2.1. INCLUSIONS

Dans le cas le plus simple, nous définissons "**l'inclusion**" par:

$$A \subset B \Leftrightarrow \forall x [x \in A \Rightarrow x \in B] \quad (5.52)$$

En langage non spécialisé voici ce qu'il faut lire: A est "inclus" (ou "fait partie", ou encore est un "sous-ensemble") dans B alors pour tout x appartenant à A chacun de ces x appartient aussi à B :

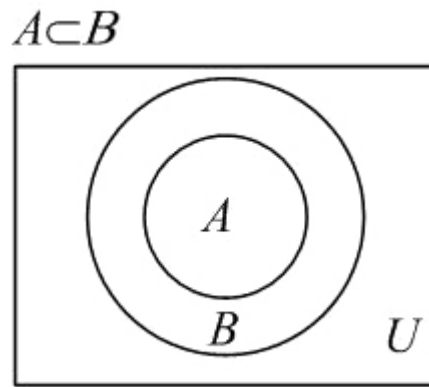


Figure: 5.3 - Exemple visuel de l'inclusion

où le U dans le coin inférieur droit de la figure représente l'univers (de Cantor).

De ceci il en découle les propriétés suivantes:

P1. Si $A \subset B$ et $B \subset A$ alors cela implique $A = B$ et réciproquement.

P2. Si $A \subset B$ et $B \subset C$ alors cela implique $A \subset C$.

2.2. INTERSECTION

Dans le cas le plus simple, nous avons:

$$A \cap B = \{x | x \in A \text{ et } x \in B\} \quad (5.53)$$

En langage non spécialisé voici ce qu'il faut lire: "**L'intersection**" des ensembles A et B consiste en l'ensemble des éléments qui se trouvent à la fois dans A et dans B :

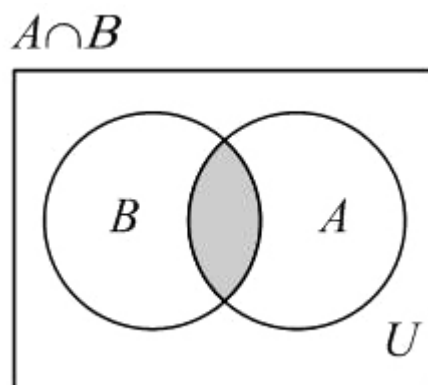


Figure: 5.4 - Exemple visuel de l'intersection

Plus généralement, si (A_i) est une famille d'ensembles indexés par $i \in I$, l'intersection des $(A_i), i \in I$ est notée:

$$\bigcap_{i \in I} A_i \quad (5.54)$$

Cette intersection est donc définie explicitement par:

$$\bigcap_{i \in I} A_i = \{x / \forall i \in I \ x \in A_i\} \quad (5.55)$$

C'est-à-dire que l'intersection de la famille d'ensembles indexés comprend tous les x qui se trouvent dans chaque ensemble de tous les ensembles de la famille.

Soient deux ensembles A et B , nous disons qu'ils sont "**disjoints**" si et seulement si:

$$A \cap B = \emptyset \quad (5.56)$$

Par ailleurs, si:

$$E \cap F = \emptyset \Rightarrow \text{Card}(E \cup F) = \text{Card}(E) + \text{Card}(F) \quad (5.57)$$

Les mathématiciens notent cela:

$$E \sqcup F \quad (5.58)$$

et l'appellent "**union disjointe**".

On plaisante parfois en disant que la connaissance se construit sur la disjonction... (ceux qui comprendront apprécieront...).

Définition: Une collection $\mathcal{S} = \{S_i\}$ d'ensembles non vides forment une "**partition**" d'un ensemble A si les propriétés suivantes sont vérifiées:

$$P1. \forall S_i, S_j \in \mathcal{S} \text{ et } i \neq j \Rightarrow S_i \cap S_j = \emptyset$$

$$P2. A = \bigcup_{S_i \in \mathcal{S}} S_i$$

Exemples:

E1. L'ensemble des nombres pairs et l'ensemble des nombres impairs forment une partition de $\mathbb{Z}$.

E2. La loi d'intersection $\cap$ est une loi commutative (voir plus loin la définition du concept de "loi") telle que:

$$A \cap B = B \cap A \quad (5.59)$$

2.3. RÉUNION/UNION

Dans le cas le plus simple, nous avons:

$$A \cup B = \{x | x \in A \text{ ou } x \in B\} \quad (5.60)$$

En langage non spécialisé voici ce qu'il faut lire: La "**réunion**" ou "**union**" des ensembles A et B consiste en l'ensemble des éléments qui se trouvent dans A et en plus dans B :

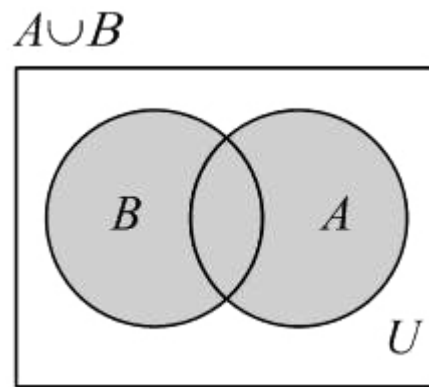


Figure: 5.5 - Exemple visuel de la réunion

Plus généralement, si (A_i) est une famille d'ensembles indexés par $i \in I$, l'union des $(A_i), i \in I$ est notée $\bigcup_{i \in I} A_i$. Cette réunion est définie par:

$$\bigcup_{i \in I} A_i = \{x / \exists i \in I \ x \in A_i\} \quad (5.61)$$

C'est-à-dire que la réunion de la famille d'ensembles indexés comprend tous les x pour lesquels il existe un ensemble indexé par i tel que x soit inclus dans cet ensemble A_i .

Nous avons les propriétés de distributivité suivantes:

$$\left(\bigcup_{i \in I} A_i \right) \cap B = \bigcup_{i \in I} (A_i \cap B) \quad (5.62)$$

$$\left(\bigcap_{i \in I} A_i \right) \cup B = \bigcap_{i \in I} (A_i \cup B) \quad (5.63)$$

La loi de réunion $\cup$ est une loi commutative (voir plus loin la définition du concept de "loi") telle que:

$$A \cup B = B \cup A \quad (5.64)$$

Nous appelons par ailleurs "**lois d'idempotences**" les relations (précisons cela pour la culture générale):

$$\begin{aligned} A \cap A &= A \\ A \cup A &= A \end{aligned} \quad (5.65)$$

et "**lois d'absorptions**" les lois:

$$\begin{aligned} A \cap (A \cup B) &= A \\ A \cup (A \cap B) &= A \end{aligned} \quad (5.66)$$

Les lois de réunion et d'intersection sont associatives telles que:

$$\begin{aligned} A \cap (B \cap C) &= (A \cap B) \cap C \\ A \cup (B \cup C) &= (A \cup B) \cup C \end{aligned} \quad (5.67)$$

et distributives telles que:

$$\begin{aligned} A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) \\ A \cup (B \cap C) &= (A \cup B) \cap (A \cup C) \end{aligned} \quad (5.68)$$

2.4. DIFFÉRENCE

Dans le cas le plus simple, nous avons:

$$A \setminus B = \{x \mid x \in A \wedge x \notin B\} \quad (5.69)$$

En langage non spécialisé voici ce qu'il faut lire: La "**différence**" des ensembles A et B consiste en l'ensemble des éléments qui se trouvent uniquement dans A (et qui excluent donc les éléments de B):

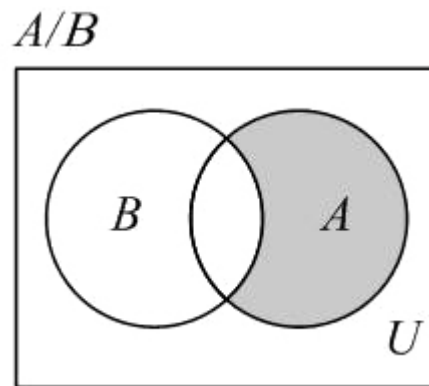


Figure: 5.6 - Exemple visuel de la différence

Si nous nous rappelons du concept de "cardinal" (voir plus haut), nous avons avec les opérations précédemment définies, la relation suivante:

$$\text{Card}(A \cup B) = \text{Card}(A) + \text{Card}(B) - \text{Card}(A \cap B) \quad (5.70)$$

d'où si $A \cap B = \emptyset$:

$$\text{Card}(A \cup B) = \text{Card}(A) + \text{Card}(B) \quad (5.71)$$

2.5. DIFFÉRENCE SYMÉTRIQUE

Soit U un ensemble. Pour tout $A, B \subseteq U$ nous définissons la différence symétrique $A \Delta B$ entre A et B par:

$$A \Delta B = (A \setminus B) \cup (B \setminus A) \quad (5.72)$$

En langage non spécialisé voici ce qu'il faut lire: La "**différence symétrique**" des ensembles A et B consiste en l'ensemble des éléments qui se trouvent uniquement dans A et de ceux se trouvant uniquement dans B (nous laissons donc de côté les éléments qui sont communs):

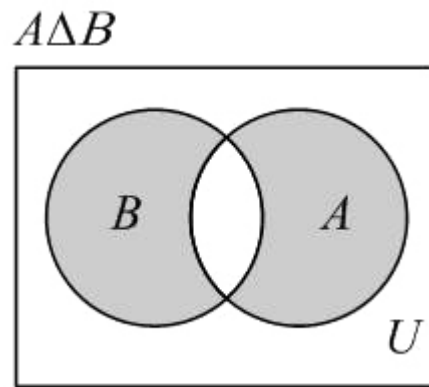


Figure: 5.7 - Exemple visuel de la différence symétrique

Les propriétés triviales sont les suivantes:

P1. $A \Delta B = B \Delta A$

P2. $A^c \Delta B^c = A \Delta B$ (pour la notion de complémentarité voir plus loin)

P3. $A \Delta B = (A \cup B) \setminus (A \cap B)$

2.6. PRODUIT

Dans le cas le plus simple, nous avons:

$$A \times B \quad (5.73)$$

En langage non spécialisé voici ce qu'il faut lire: "**l'ensemble produit**" (à ne pas confondre avec la multiplication ou le produit vectoriel) de deux ensembles A et B est l'ensemble des couples tels que:

$$(x, y), x \in A, y \in B \quad (5.74)$$

L'ensemble produit des réels $\mathbb{R} \times \mathbb{R}$ par exemple forme le plan où chaque élément est défini par une abscisse et son ordonnée. Nous retrouvons souvent les ensembles produits en mathématiques et en physique lors que nous travaillons avec des fonctions. Par exemple, une fonction de deux variables réelles qui donne un réel en sortie sera noté:

$$\begin{matrix} f(x, y) \rightarrow z \\ \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} \end{matrix} \quad (5.75)$$

mais cette notation n'est à ma connaissance pas normalisée et il en existe de nombreuses variantes.

2.7. COMPLÉMENTARITÉ

Dans le cas le plus simple, nous avons:

$$\forall A \subset B \quad \bar{A} = \{x \mid x \in B, x \notin A\} \quad (5.76)$$

En langage non spécialisé voici ce qu'il faut lire: Le "**complémentaire**" est défini comme en prenant B un ensemble et A un sous-ensemble de B alors le complémentaire de A dans B est l'ensemble des éléments qui sont dans B mais pas dans A .

Par exemple, dans la figure ci-dessous nous avons le complémentaire de A par rapport à U qui est indiqué en gris (s'il est seul il s'agit donc de l'univers seul qui l'entoure):

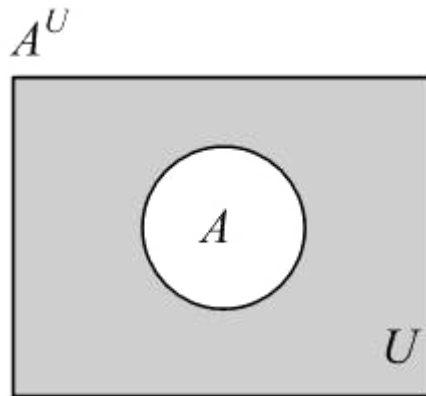


Figure: 5.8 - Exemple visuel de la complémentarité

Une autre notation très importante de la complémentarité qu'on retrouve parfois dans la littérature est la suivante:

$${}_U A^c \text{ ou } A^c \quad (5.77)$$

où dans le cas particulier à droite ci-dessus, nous pourrions aussi écrire B/A (la notation ${}_U A^c$ serait rarement utilisée car elle peut prêter à confusion dans certaines situations).

Nous avons comme propriétés pour tout A_i inclus dans B :

$$\left(\bigcup_{i \in I} A_i \right)^c = \bigcap_{i \in I} A_i^c \quad (5.78)$$

$$\left(\bigcap_{i \in I} A_i \right)^c = \bigcup_{i \in I} A_i^c \quad (5.79)$$

Voici quelques lois triviales relatives aux compléments:

$$\begin{aligned} \overline{\overline{A}} &= A \\ A \cap \overline{A} &= \emptyset \\ A \cup \overline{A} &= U \end{aligned} \quad (5.80)$$

Il existe d'autres lois très importantes en logique booléenne. Si nous considérons trois ensembles A, B, C comme représentés ci-dessous:

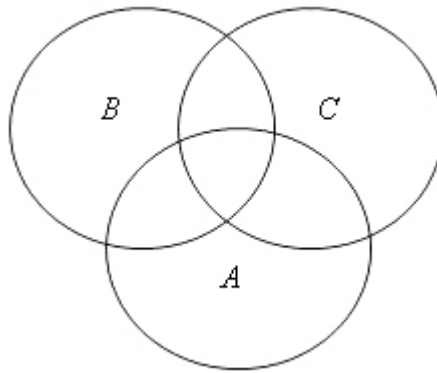


Figure: 5.9 - Exemple de trois ensembles particuliers

nous avons donc:

$$\begin{aligned} A \setminus (B \cap C) &= (A \setminus B) \cup (A \setminus C) \\ A \setminus B \setminus C &= (A \setminus B) \cap (A \setminus C) \end{aligned} \quad (5.81)$$

et les fameuses "**lois de De Morgan**" sous forme ensembliste (cf. chapitre de Systèmes Logiques Formels) et qui sont données par les relations:

$$\begin{aligned} \overline{A \cap B} &= \overline{A} \cup \overline{B} \\ \overline{A \cup B} &= \overline{A} \cap \overline{B} \end{aligned} \quad (5.82)$$

3. FONCTIONS ET APPLICATIONS

Définition: En mathématiques, une "**application**" (ou "**fonction**") notée f ou A est la donnée de deux ensembles, l'ensemble de départ E et l'ensemble d'arrivée F (ou d'**image** de E), et d'une relation associant à chaque élément x de l'ensemble de départ un et un seul élément de l'ensemble d'arrivée, que nous appelons "**image de x par f** " et que nous notons $f(x)$.

Nous appelons "images" les éléments de $f(E)$ et les éléments de E sont appelés les antécédents.

Nous disons alors que f est une application de E dans F notée:

$$f : E \rightarrow F \quad (5.83)$$

(se rappeler du premier diagramme sagittal présenté au début de ce chapitre), ou encore une application à arguments dans E et valeurs dans F .

Remarque: Le terme "fonction" est souvent utilisé pour les applications à valeurs numériques, réelles ou complexes, c'est-à-dire lorsque l'ensemble d'arrivée est $\mathbb{R}$ ou $\mathbb{C}$. Nous parlons alors de "fonction réelle", ou de "fonction complexe".

Définitions:

D1. Le "**graphe**" (ou encore "**graphique**" ou "**représentative**") d'une application $f : E \rightarrow F$ est le sous-ensemble du produit cartésien $E \times F$ constitué des couples $(x, f(x))$ pour x variant dans E . La donnée du graphe de f détermine son ensemble de départ (par projection sur la première composante

souvent notée x) et son image (par projection sur la seconde composante souvent notée y).

D2. Si le triplet $f = (E, F, \Gamma)$ est une fonction où E et F sont deux ensembles et $\Gamma \subset E \times F$ est un graphe, E et F sont respectivement la source et le but de f . Le "**domaine de définition**" ou "**ensemble de départ**" de f est:

$$D_f = \{x \in E \mid \exists y \in F, (x, y) \in \Gamma\} \quad (5.84)$$

D3. Etant donnés trois ensembles E , F et G (non vides), toute fonction de $E \times F$ vers G est appelée "**loi de composition**" de $E \times F$ à valeurs dans G .

D4. Une "**loi de composition interne**" (ou simplement "**loi interne**") dans E est une loi de composition de $E \times E$ à valeurs dans E (cas $E=F=G$).

Remarque: La soustraction dans $\mathbb{N}$ n'est pas une loi de composition interne bien qu'elle fasse partie des quatre opérations élémentaires apprises à l'école. Par contre l'addition sur $\mathbb{N}$ en est bien une.

D5. Une "**loi composition externe**" (ou simplement "**loi externe**") dans E est une loi de composition de $F \times E$ à valeurs dans E , où F est un ensemble distinct de E . En général, F est un corps, dit "**corps de scalaires**".

Exemple:

Dans le cas d'un espace vectoriel (voir définition beaucoup plus bas) la multiplication d'un vecteur (dont les composantes se basent sur un ensemble donné) par un réel est une loi de composition externe.

Remarque: Une loi de composition externe à valeurs dans E est aussi appelée "**action de F sur E** ". L'ensemble F est alors le domaine d'opérateurs. On dit aussi que F opère sur E (ayez en tête l'exemple des vecteurs précédemment cité).

D6. Nous appelons "**image de f** ", et nous notons $\text{Im}(f)$, le sous-ensemble défini par:

$$\text{Im}(f) = f(E) = \{y \in F \mid \exists x \in E, y = f(x)\} \quad (5.85)$$

Ainsi, "**L'image**" d'une application $f: E \rightarrow F$ est la collection des $f(x)$ pour x parcourant E , c'est un sous-ensemble de F .

Et nous appelons "**noyau de f** ", et nous notons $\text{Ker}(f)$, le sous-ensemble très important en mathématiques défini par:

$$\text{Ker}(f) = f^{-1}(\{0\}) = \{x \in E \mid f(x) = 0\} \quad (5.86)$$

Selon la figure (il faut bien comprendre ce concept de noyau car nous le réutiliserons de nombreuses fois pour démontrer des théorèmes ayant des applications pratiques importantes):

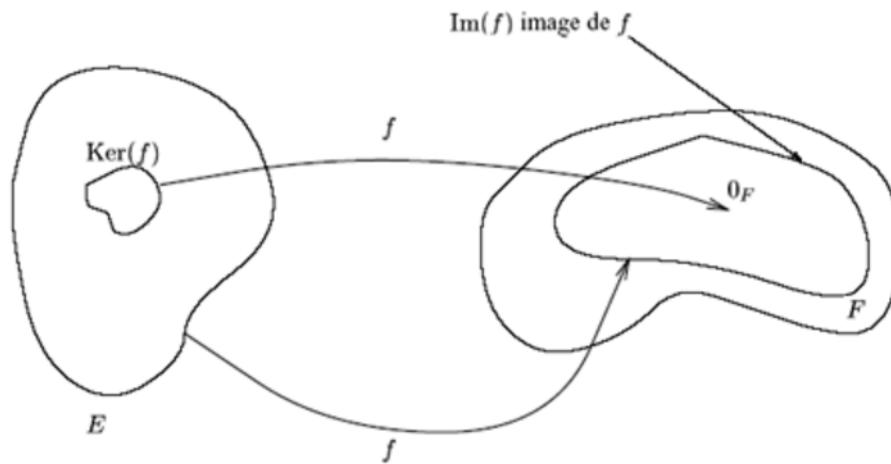


Figure: 5.10 - Représentation du concept de noyau d'une fonction

Remarques:

R1. $\text{Ker}(f)$ provient de l'allemand "Kern", signifiant tout simplement "noyau". En anglais, le noyau se dit aussi "kernel", signifiant "amande" dans le civil.

R2. Normalement les notations Im et Ker sont réservées aux homomorphismes de groupes, d'anneaux, de corps et aux applications linéaires entre espaces vectoriels ou modules etc.... (voir plus loin). Nous n'avons normalement pas l'habitude de les utiliser pour des applications quelconques entre ensembles quelconques. Mais bon...ça ne fait rien.

Les applications peuvent avoir une quantité phénoménale de propriétés dont voici celles qui font partie des connaissances générales du physicien (pour plus de renseignements sur ce qu'est une fonction, voir le chapitre traitant de l'Analyse Fonctionnelle).

Soit f une application d'un ensemble E à un ensemble F alors nous avons les propriétés suivantes:

P1. Une application est dite "**surjective**" si:

Tout élément y de F est l'image par f d'au moins (nous insistons sur le "au moins") un élément de E . Nous disons encore que c'est une "**surjection**" de E dans F . Il découle de cette définition, qu'une application $f : E \rightarrow F$ est surjective si et seulement si $F = \text{Im } f$. En d'autres termes, nous écrivons aussi cette définition ainsi:

$$\forall y \in F, \exists x \in E \quad y = f(x) \quad (5.87)$$

ce qui s'illustre par:

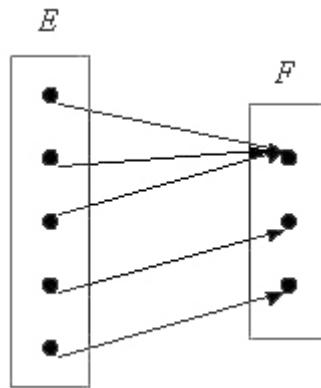


Figure: 5.11 - Représentation d'une fonction surjective

P2. Une application est dite "**injective**" si:

Tout élément y de F est l'image par f d'au plus (nous insistons sur le "au plus") un seul élément de E . Nous disons encore que f est une injection de E dans F . Il résulte de cette définition, qu'une application $f : E \rightarrow F$ est injective si et seulement si les relations $x_1, x_2 \in E$ et $f(x_1) = f(x_2)$ impliquent $x_1 = x_2$ autrement dit: une application pour laquelle deux éléments distincts ont des images distinctes est dite injective. Ou encore, une application est injective si l'une aux moins des propriétés équivalentes suivantes est vérifiée:

$$P2.1 \quad \forall (x, y) \in E^2 \quad f(x) = f(y) \Rightarrow x = y$$

$$P2.2 \quad \forall (x, y) \in E^2 \quad x \neq y \Rightarrow f(x) \neq f(y)$$

$$P2.3 \quad \forall y \in F \text{ l'équation en } x, y = f(x) \text{ a au plus une solution dans } E$$

Tout cela s'illustrant par:

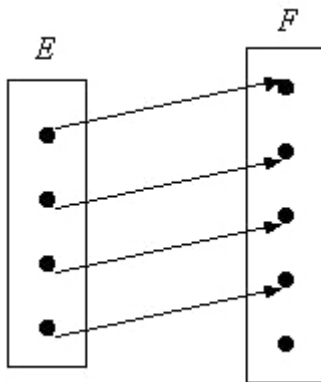


Figure: 5.12 - Représentation d'une fonction injective

P3. Une application est dite "**bijjective**" si:

Une application f de E dans F est à la fois surjective et injective. Dans ce cas, nous avons que pour tout élément y de F , l'équation $y = f(x)$ admet dans E une unique (ni "au plus", ni "au moins") pré-image x . Ce que nous écrivons aussi:

$$\forall y \in F, \exists ! x \in E \quad y = f(x) \quad (5.88)$$

ce qui s'illustre par:

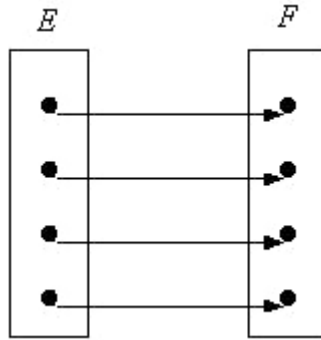


Figure: 5.13 - Représentation d'une fonction bijective

Nous sommes ainsi tout naturellement amené à définir une nouvelle application de F dans E , appelée "**fonction réciproque**" de f et notée f^{-1} , qui à tout élément y de F , fait correspondre l'élément x de E pré-image (ou solution) unique de l'équation $y = f(x)$. Autrement dit:

$$\boxed{x = f^{-1}(y)} \quad (5.89)$$

L'existence d'une application réciproque implique que le graphique d'une application bijective (dans l'ensemble des réels...) et celui de son application réciproque sont symétriques par rapport à la droite d'équation $y = x$.

Effectivement, nous remarquons que si $y = f(x)$ est équivalent à $x = f^{-1}(y)$, alors ces équations impliquent que le point (x, y) est sur le graphique de f si et seulement si le point (y, x) est sur le graphique de f^{-1} .

Exemple:

Prenons le cas d'une station de vacances où un groupe de touristes doit être logé dans un hôtel. Chaque façon de répartir ces touristes dans les chambres de l'hôtel peut être représentée par une application de l'ensemble des touristes vers l'ensemble des chambres (à chaque touriste est associée une chambre).

- Les touristes souhaitent que l'application soit injective, c'est-à-dire que chacun d'entre eux ait une chambre individuelle. Cela n'est possible que si le nombre de touristes ne dépasse pas le nombre de chambres.
- L'hôtelier souhaite que l'application soit surjective, c'est-à-dire que chaque chambre soit occupée. Cela n'est possible que s'il y a au moins autant de touristes que de chambres.
- S'il est possible de répartir les touristes de telle sorte qu'il y en ait un seul par chambre, et que toutes les chambres soient occupées: l'application sera alors à la fois injective et surjective nous dirons qu'elle est bijective.

Remarques:

R1. Il vient des définitions ci-dessus qu'une application f est bijective (ou "biunivoque") dans l'ensemble des réels si et seulement si toute droite horizontale coupe la représentation graphique de la fonction en un seul point. Nous sommes donc amenés à faire la seconde remarque suivante:

R2. Une application qui vérifie le test de la droite horizontale est continument croissante ou décroissante en tout point de son domaine de définition.

P4. Une application est dite "**fonction composée**" si:

Soit φ une application de E dans F et ψ une fonction de F dans G . L'application qui associe à chaque élément x de l'élément de E , $\psi(\varphi(x))$ de G s'appelle "**application composée**" de φ et de ψ et se note:

$$\psi \circ \varphi \quad (5.90)$$

où symbole " $\circ$ " est appelé "**rond**". Ainsi, la relation précédente s'écrit "psi rond phi" mais se lit "phi rond psi"... Ainsi:

$$(\psi \circ \varphi)(x) = \psi(\varphi(x)) \quad (5.91)$$

Soit, de plus, χ une application de G dans H . Nous vérifions aussitôt que l'opération de composition est associative:

$$\chi \circ (\psi \circ \varphi) = (\chi \circ \psi) \circ \varphi \quad (5.92)$$

Cela nous permet d'omettre les parenthèses et d'écrire plus simplement: $\chi \circ \psi \circ \varphi$

Dans le cas particulier où φ serait une application de E dans E , nous notons φ^k l'application composée $\varphi \circ \varphi \circ \dots \circ \varphi$ (k fois).

Ce qui est important dans ce que nous venons de voir dans ce chapitre, c'est que toutes les propriétés définies et énoncées ci-dessus sont applicables aux ensembles de nombres.

Voyons en un exemple très concret et très puissant:

3.1. THÉORÈME DE CANTOR-BERNSTEIN

Attention. Ce théorème, dont le résultat peut sembler évident, n'est pas forcément simple à aborder (son formalisme mathématique n'est pas très esthétique...). Nous vous conseillons de lire très lentement et de vous imaginer les diagrammes sagittaux dans la tête lors de la démonstration.

Voici l'hypothèse à démontrer: Soient X et Y deux ensembles. S'il existe une injection (voir la définition d'une fonction injective ci-dessus) de X vers Y et une autre de Y vers X , alors les deux ensembles sont en bijection (voir la définition d'une fonction bijective ci-dessus). Il s'agit donc aussi d'une relation antisymétrique.

Ce qui s'illustre par:

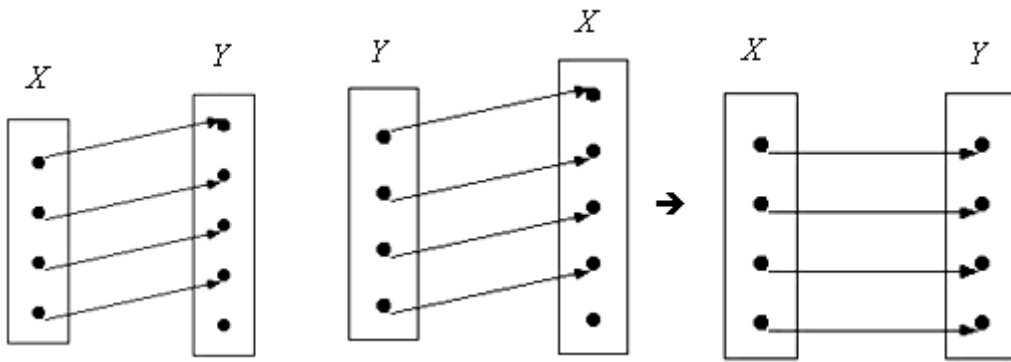


Figure: 5.14 - Représentation d'une relation antisymétrique

Pour la démonstration, nous avons besoin en toute rigueur de démontrer au préalable un lemme (évident intuitivement mais pas formellement...) dont l'énoncé est le suivant:

Soient X, Y, Z trois ensembles tels que $X \subseteq Z \subseteq Y$. Si X et Y sont en bijection, alors X et Z sont en bijection.

Un exemple d'application de ce lemme est l'ensemble des nombres naturels et des nombres rationnels qui sont en bijection. Dès lors, l'ensemble des entiers relatifs est en bijection avec l'ensemble des nombres naturels puisque:

$$\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \quad (5.93)$$

Démonstration:

D'abord, au niveau formel, créons une fonction f de Y à X telle quelle soit bijective:

$$f : Y \rightarrow X \quad (5.94)$$

Nous avons besoin pour la suite d'un ensemble A qui sera défini par l'union des images des fonctions des fonctions f (du genre $f(f(f...))$) des pré-images de l'ensemble Z dont nous excluons les éléments de X (ce que nous notons: $Z - X$). En d'autres termes (si la première forme n'est pas claire...) nous définissons l'ensemble A comme étant l'union des images de $(Z - X)$ par les applications $f \circ f \circ f \dots$. Ce que nous noterons :

$$A = \bigcup_{n=1}^{\infty} f^n(Z - X) \quad (5.95)$$

Nous avons donc par construction $A \subseteq X$. Remarquons que nous avons aussi:

$$A = \bigcup_{n=1}^{\infty} f^n(Z - X) \Rightarrow f(A) = f\left(\bigcup_{n=1}^{\infty} f^n(Z - X)\right) = \bigcup_{n=1}^{\infty} f(f^n(Z - X)) = \bigcup_{n=1}^{\infty} f^{n+1}(Z - X) \quad (5.96)$$

et en réindexant:

$$f(A) = \bigcup_{n=2}^{\infty} f^n(Z - X) \quad (5.97)$$

Nous avons alors (faire un schéma de tête des diagrammes sagittaux peut aider à ce niveau-là...):

$$f((Z - X) \cup A) = A \quad (5.98)$$

Nous pouvons démontrer élégamment cette dernière relation:

$$f((Z - X) \cup A) = f(Z - X) \cup f(A) = f(Z - X) \cup \bigcup_{n=2}^{\infty} f^n(Z - X) = \bigcup_{n=1}^{\infty} f^n(Z - X) = A \quad (5.99)$$

Comme Z peut être partitionné (rien nous en empêche!) en les deux sous-ensembles disjoints $(Z - X) \cup A$ et $(X - A)$ sans oublier que $X \subseteq Z \subseteq Y$ et $A \subseteq X$, nous posons comme une définition l'application g telle que:

$$g: Z \rightarrow X \quad (5.100)$$

tel que pour toute pré-image a nous ayons:

$$\forall a \in (Z - X) \cup A \mapsto f(a) \quad (5.101)$$

(rappelez-vous de la définition des applications notées " f ") et:

$$\forall a \in (X - A) \mapsto a \quad (5.102)$$

L'application g est alors bijective car ses restrictions à $(Z - X) \cup A$ et $(X - A)$, (qui forment une partition) sont f et l'identité qui sont par définition bijectives.

Finalement il existe bien, par construction, une bijection entre X et Z .

□C.Q.F.D.

Reprenons les hypothèses du théorème de Cantor-Bernstein:

Soit φ une injection de X vers Y et ψ une injection de Y vers X

Nous avons alors:

$$\varphi(X) \subset Y \text{ et } \psi(Y) \subset X \quad (5.103)$$

donc:

$$\psi(\varphi(X)) \subset \psi(Y) \subset X \quad (5.104)$$

Comme φ est injective, X et $\varphi(X)$ sont par définition en bijection et de même, comme ψ est injective, $\psi(\varphi(X))$ et $\varphi(X)$ sont en bijection (là il est bon de relire...).

Donc: X et $\psi(\varphi(X))$ sont eux aussi en bijection.

En utilisant le lemme sur $\psi(\varphi(X)), \psi(Y)$ et X , il vient donc que $\psi(\varphi(X))$ est en bijection $\psi(Y)$ ce qui nous donne avec ceux que nous avons vu juste précédemment, que puisque aussi $\psi(\varphi(X))$ et $\varphi(X)$ sont en bijection, alors que $\psi(Y)$ est en bijection avec $\varphi(X)$, alors X et Y sont en injection (ouf! c'est beau mais c'est aussi vicieux que simple).

□C.Q.F.D.

Ce théorème s'interprète de la manière suivante: Si nous pouvons compter une partie d'un ensemble avec la totalité des éléments d'un autre ensemble, et réciproquement, alors ils ont le même nombre d'éléments. Ce qui est évident pour des ensembles finis. Ce théorème généralise alors cette notion pour des ensembles infinis et c'est là sa force!

À partir de là, ce théorème représente l'une des briques de base pour généraliser la notion de tailles d'ensembles à des ensembles infinis.

4. STRUCTURES

L'algèbre dite "**algèbre moderne**" commence avec la théorie des structures algébriques due en partie à Carl F. Gauss et surtout à Évariste Galois. Ces structures existent en un très grand nombre mais seulement les fondamentales nous intéresseront ici. Avant de les détailler, voici un diagramme synoptique de ces principales structures et de leur hiérarchie:

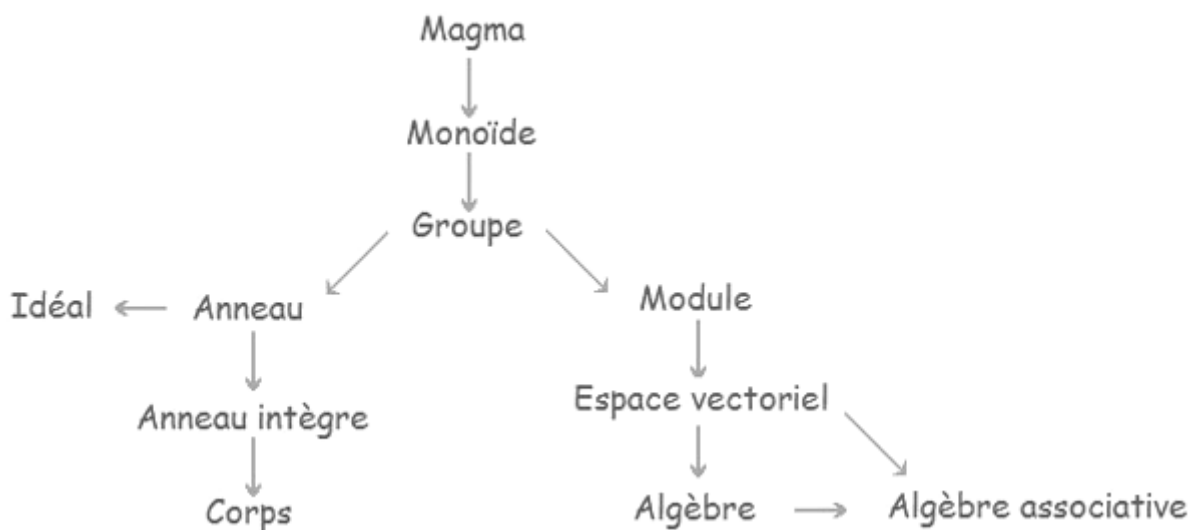


Figure: 5.15 - Diagramme synoptique des structures algébriques courantes

Remarques: Tout en haut du diagramme, la structure au nombre minimal de contraintes, en bas, un maximum. Soit, plus nous descendons, plus la structure est en quelque sorte spécialisée.

Soit pour simplifier les écritures, $*$ une loi de composition (comme l'addition, la soustraction, la multiplication ou encore la division,...)...

Remarque: Cette notation généralisée est parfois appelée "**notation stellaire**".

Définitions: Soit $*$ et $\circ$ des symboles de lois internes à un ensemble E (cela pourrait être l'addition et la multiplication pour prendre le cas le plus connu) alors:

D1. $*$ est une "loi commutative" si:

$$a * b = b * a \quad (5.105)$$

D2. $*$ est une "loi associative" si:

$$a * (b * c) = (a * b) * c \quad (5.106)$$

D3. n est "élément neutre" pour $*$ si:

$$a * n = n * a = a \quad (5.107)$$

Nous admettons par ailleurs sans démonstration (c'est intuitif) que s'il existe un élément neutre, il est unique.

D4. a' est "l'élément symétrique" (dans le sens général de l'opposé par exemple pour l'addition et l'inverse pour la multiplication) de a pour $*$ si:

$$a' * a = a * a' = n \quad (5.108)$$

Nous admettons également et sans démonstration que le symétrique de tout élément est unique.

D5. $\circ$ est une "loi distributive" par rapport à $*$ si:

$$\begin{aligned} a \circ (b * c) &= (a \circ b) * (a \circ c) \\ (a * b) \circ c &= (a \circ c) * (b \circ c) \end{aligned} \quad (5.109)$$

D6. b est "l'élément absorbant" si pour tout a et une loi $*$ nous avons:

$$a * b = 0 \quad (5.110)$$

Remarques:

R1. Si a est son propre symétrique par rapport à la loi $*$, les mathématiciens disent que a est "involutif".

R2. Si un élément b de E vérifie $a * b = b * a = b$, alors b est dit "élément absorbant" pour la loi $*$.

R3. Il faut toujours vérifier que les neutres et les symétriques le soient "à gauche" et "à droite". Ainsi, par exemple, dans $(\mathbb{Z}, -)$, l'élément 0 n'est un neutre qu'à droite car $x - 0 = x$ mais $0 - x = -x$.

4.1. MAGMA

Définition: Nous désignons un ensemble par le terme "magma" M , si les composants le constituant sont opérables par rapport à une loi interne $*$:

$$(\mathcal{M}, *) \text{ est un magma si } \Rightarrow \begin{cases} * \text{ est une opération} \\ * \text{ est une loi interne} \end{cases}$$

Remarques:

- R1. Si de plus la loi interne $*$ est commutative, nous parlons de "**magma commutatif**".
- R2. Si de plus la loi interne $*$ est associative, nous parlons de "**magma associatif**".
- R3. Si de plus la loi interne $*$ possède un élément neutre, nous parlons de "**magma unitaire**".

Il est donc important de se rappeler que si nous désignons une structure algébrique par le terme "magma" tout court, cela ne signifie en aucun cas que la loi interne est commutative, associative ou même qu'elle possède un élément neutre !

Définition: Dans un magma $(\mathcal{M}, *)$, un élément x est dit "**élément régulier**" (ou "**élément simplifiable**") à gauche si pour tout couple $(a, b) \in \mathcal{M}$ nous avons:

$$x * a = x * b \Rightarrow a = b \quad (5.111)$$

Remarque: Nous définissons de même un élément régulier à droite.

Ainsi, un élément est dit "**régulier**" s'il est régulier à droite et à gauche. Si $*$ est commutative (ce qui est le cas pour un magma commutatif), les notions d'élément régulier à gauche ou à droite coïncident.

 **Exemple:**

Dans $(\mathbb{N}, +)$ tout élément est régulier et dans $(\mathbb{N}, \times)$ tout élément non nul est régulier.

Un magma $(\mathcal{M}, *)$ est donc une structure algébrique élémentaire. Il existe des structures plus subtiles (monoïdes, groupes, anneaux, corps, espace vectoriels, etc.) dans lesquelles un ensemble est muni de plusieurs lois et de différentes propriétés. Nous allons les voir de suite et les utiliser tout au long de ce site.

4.2. MONOÏDE

Définition: Si la loi $*$ est associative et possède un élément neutre nous disons alors que le "**magma associatif unitaire**" est un "**monoïde**":

$$(\tilde{\mathcal{M}}, *) \text{ est un monoïde si } \Rightarrow \begin{cases} * \text{ est associative} \\ * \text{ Il existe dans } \tilde{\mathcal{M}} \text{ un élément neutre } \varepsilon \text{ pour } * \end{cases}$$

Remarques:

R1. Si de plus la loi interne $*$ est commutative alors nous disons alors que la structure forme un "monoïde abélien" (ou simplement "monoïde commutatif").

R2. Dans certains ouvrages nous trouvons aussi comme définition que le monoïde est un "demi-groupe" (avec une loi associative) muni d'un élément neutre.

Montrons tout de suite que l'ensemble des entiers naturels $\mathbb{N}$ est un monoïde abélien totalement ordonné (comme nous l'avons partiellement vu dans le chapitre des opérateurs) par rapport aux lois d'addition et de multiplication:

La loi d'addition $(+)$ est-elle une opération interne telle que $\forall a, b \in \mathbb{N}$ nous ayons:

$$a + b = c \quad (5.112)$$

Nous pouvons démontrer que c'est bien le cas en sachant que 1 appartient à $\mathbb{N}$ tel que:

$$\sum_{i=1}^a 1 + \sum_{i=1}^b 1 = \sum_{i=1}^{a+b} 1 \quad (5.113)$$

Donc $c \in \mathbb{N}$ et l'addition est bien une loi interne (nous disons également que l'ensemble $\mathbb{N}$ est "stable" par rapport à l'addition) et en même temps associative puisque 1 peut être additionné à lui-même par définition dans n'importe quel ordre sans que le résultat en soit altéré. Si vous vous rappelez que la multiplication est une loi qui se construit sur l'addition, alors la loi de multiplication $(\times)$ est aussi une loi interne et associative !

Nous admettons à partir d'ici qu'il est trivial que la loi d'addition est également commutative et que le zéro "0" en est l'élément neutre (n). Ainsi, la loi de multiplication est elle aussi commutative et il est trivial que "1" en est l'élément neutre (n).

Par ailleurs, pour parler déjà de quelque chose qui n'est pas directement en relation avec le monoïde... mais qui nous sera utile un peu plus loin, existe-t-il en restant dans la lignée de l'exemple précédent pour la loi d'addition $(+)$ un symétrique $\exists c$ tel que $\forall a, b \in \mathbb{N}$ nous ayons:

$$\left(\sum_{i=1}^a 1 + \sum_{i=1}^b 1 \right) + c = \left(\sum_{i=1}^{a+b} 1 \right) + c = 0 \quad (5.114)$$

avec $c \in \mathbb{N}$?

Il est assez trivial que pour que cette égalité soit satisfaite nous ayons:

$$\left(\sum_{i=1}^{a+b} 1 \right) = -c \quad (5.115)$$

soit:

$$a + b = -c \quad (5.116)$$

or les nombres négatifs n'existent pas dans $\mathbb{N}$. Ce qui nous amène aussi à la conclusion que la loi d'addition (+) n'a pas de symétrique et que la loi de soustraction (-) n'existe pas dans $\mathbb{N}$ (la soustraction étant rigoureusement l'addition d'un nombre négatif).

De même, car cela va aussi nous être utile un peu plus loin, existe-t-il pour la loi de multiplication (x) un symétrique a' tel que $\forall a \in \mathbb{N}$ nous ayons:

$$a' \cdot a = a \cdot a' = 1 \quad (5.117)$$

avec $a' \in \mathbb{N}$?

D'abord il est évident que:

$$a' = \frac{1}{a} \quad (5.118)$$

Mais excepté pour $a = 1$, le quotient $1/a$ n'existe pas dans $\mathbb{N}$. Donc nous devons conclure qu'il n'existe pas pour tout élément de $\mathbb{N}$ de symétriques pour la loi de multiplication et ainsi que la loi de division n'existe pas dans $\mathbb{N}$ et que la loi de multiplication ne forme pas un monoïde dans cet ensemble.

Synthèse:

$\mathbb{N}$ (lois)	(+)	(-)	(x)	(/)
Opération interne	oui	non	oui	non
Commutative	oui		oui	
Élément neutre	oui (zéro "0")		oui (un "1")	
Élément absorbant	non		oui (zéro "0")	
Symétrique	non		non	

Tableau: 5.1 - Lois et leurs propriétés dans l'ensemble des entiers naturels

Nous avons par exemple les propriétés suivantes relativement à l'ensemble des entiers naturels et au concept de monoïde:

P1. $(\mathbb{N}, \leq, \geq)$ est totalement ordonné (attention cette notation est un peu abusive! il suffit qu'il y ait juste une des deux relations d'ordre R pour que l'ensemble soit totalement ordonné).

P2. $(\mathbb{N}, +)$ et $(\mathbb{N}, \times)$ sont des monoïdes abéliens.

P3. L'élément zéro "0" est l'élément absorbant pour le monoïde $(\mathbb{N}, \times)$.

P4. Les lois de soustraction et division n'existent pas dans l'ensemble $\mathbb{N}$.

P5. $\mathbb{N}$ est un monoïde abélien totalement ordonné par rapport aux lois d'addition et de multiplication (attention la notation suivante est abusive car le monoïde n'est composé que d'une seule loi interne et d'une relation d'ordre R ce qui donnerait au total 4 monoïdes):

$$(\mathbb{N}, +, \times, \leq, \geq) \quad (5.119)$$

Remarques:

R1. Il est rare d'utiliser les monoïdes; car souvent, lorsque nous nous trouvons face à une structure trop pauvre pour pouvoir vraiment discuter, nous la prolongeons vers quelque chose de plus riche, comme un groupe, ou un anneau (voir plus loin) tel que l'ensemble des entiers relatifs.

R2. Dire qu'une structure algébrique est totalement ordonnée par rapport à certaines lois signifie que soit $*$ une loi, et R une relation d'ordre et a, b, c, d quatre éléments de la structure intéressée, alors si aRb et cRd implique $(a * c)R(b * d)$. Nous notons alors cette structure $(S, *, R)$ ou simplement (S, R) et en indiquant la (ou les) loi concernée.

4.3. GROUPE

Définition: Nous désignons un ensemble par le terme "**groupe**", si les composants le constituant satisfont aux trois conditions de ce que nous nommons la "**loi interne de groupe**", définie ci-dessous:

$$(G, *) \text{ est un groupe si } \Leftrightarrow \begin{cases} * \text{ est associative} \\ \text{Il existe dans } G \text{ un élément neutre pour } * \\ \text{Tout élément de } G \text{ possède un symétrique pour } * \end{cases}$$

Dans ce cas, la loi de compositions interne $*$ sera souvent (mais pas exclusivement!) notée "+" et appelée "**l'addition**", le neutre e noté "0" et le symétrique de x noté "-x".

Insistons sur le fait que la structure de groupe est probablement une des plus importantes dans la pratique de l'ingénieur et de la physique moderne en général. Raison pour laquelle il convient d'y porter une attention toute particulière (cf. chapitre d'Algèbre Ensembliste)!

Si de plus, la loi interne $*$ est également commutative, nous disons alors que le groupe est un "**groupe abélien**" ou simplement "**groupe commutatif**".

S'il existe dans G au moins un élément a tel que tout élément de G est une puissance de a ou du symétrique a' de a , nous disons que $(G, *)$ est un "**groupe cyclique de générateur a** " s'il est fini, sinon nous disons qu'il est "**monogène**" (nous reviendrons sur les groupes cycliques dans le chapitre d'Algèbre Ensembliste).

Plus généralement un groupe $(G, *)$ d'élément neutre e , non réduit uniquement à $\{e\}$ sera monogène, s'il existe un élément a de G distinct de e tel que $G = \{e, a^1, a^2, a^3, \dots, a^n, \dots\}$. Un tel groupe sera cyclique, s'il existe un entier n non nul pour lequel $a^n = e$. Le plus petit entier non nul vérifiant cette égalité est alors "**l'ordre du groupe**".

Exemple:

Montrons tout de suite que l'ensemble des entiers relatifs $\mathbb{Z}$ est un groupe abélien totalement ordonné (comme nous l'avons vu dans le chapitre des Opérateurs) par rapport aux lois d'addition et de multiplication.

D'abord pour raccourcir les développements, il est utile de rappeler que l'ensemble $\mathbb{Z}$ est un "prolongement" de $\mathbb{N}$ par le fait que nous y avons ajouté tous les nombres symétriques de signe négatif ($\mathbb{N} \subset \mathbb{Z}$).

Ainsi, en abusant toujours des notations (car normalement un groupe n'a qu'une seule loi et une seule relation d'ordre R suffit à l'ordonner):

$$(\mathbb{Z}; +, -, \leq, \geq) \quad (5.120)$$

forme un groupe abélien totalement ordonné (4 groupes au fait!) et:

$$(\mathbb{Z}; \times, \leq, \geq) \quad (5.121)$$

un monoïde abélien (deux monoïdes au fait!) totalement ordonné.

Remarquons aussi que la loi de division n'existe pas pour tout élément de l'ensemble $\mathbb{Z}$! Donc en toute généralité nous disons qu'elle n'y existe pas.

Synthèse:

$\mathbb{Z}$ (lois)	(+)	(-)	(x)	(/)
Opération interne	oui	oui	oui	non
Associative	oui	non	oui	
Commutative	oui	non	oui	
Élément neutre	oui (zéro "0")	non (0 pas neutre à gauche)	oui (un "1")	
Élément absorbant	non	non	oui (zéro "0")	
Symétrique	oui (signe opposé)	oui	non	

Tableau: 5.2 - Lois et leurs propriétés dans l'ensemble des entiers relatifs

Nous avons donc les propriétés suivantes:

P1. $(\mathbb{Z}, \leq, \geq)$ est totalement ordonné (attention à nouveau cette notation est un peu abusive! il suffit qu'il y ait juste une des deux relations d'ordre R pour que l'ensemble soit totalement ordonné).

P2. $(\mathbb{Z}, +)$ est un groupe commutatif dont zéro "0" est l'élément neutre.

P3. La loi de division n'existe pas dans l'ensemble $\mathbb{Z}$.

P4. L'ensemble $\mathbb{Z}$ est un groupe abélien totalement ordonné par rapport à la loi d'addition (attention la notation suivante est encore une fois abusive car le groupe est composé que d'une relation d'ordre R ce qui donnerait au total 2 groupes):

$$(\mathbb{Z}; +, \leq, \geq) \quad (5.122)$$

L'ensemble $\mathbb{Z}$ n'est pas un groupe commutatif totalement ordonné par rapport à la loi de multiplication:

$$(\mathbb{Z}; \times, \leq, \geq) \quad (5.123)$$

Nous voyons de suite alors que $\mathbb{Z}$ a des propriétés trop restreintes, c'est la raison pour laquelle il est intéressant de le prolonger par l'ensemble des rationnels $\mathbb{Q}$ défini de manière très simpliste... par (cf. chapitre sur les Nombres):

$$\mathbb{Q} = \left\{ \frac{p}{q} \mid (p, q) \in \mathbb{Z} / q \neq 0 \right\} \quad (5.124)$$

Ce qui signifie pour rappel que l'ensemble des rationnels est défini par l'ensemble des quotients p et q appartenant chacun à $\mathbb{Z}$ dont nous excluons à q de prendre la valeur nulle (la notation $/q$ signifiant l'exclusion).

Et nous avons évidemment:

$$\mathbb{N}, \mathbb{Z} \subset \mathbb{Q} \quad (5.125)$$

Il est dès lors évident (sans démonstration et toujours en utilisant la notation abusive déjà commentée maintes fois plus haut...) que $(\mathbb{Q}; \leq, \geq)$ est aussi totalement ordonné et aussi que $\mathbb{Q}$ est un groupe abélien totalement ordonné par rapport à la loi d'addition seulement:

$$(\mathbb{Q}; +, \geq, \leq) \quad (5.126)$$

Ce qui devient intéressant avec $\mathbb{Q}$, c'est que la loi de multiplication devient une loi interne et forme un groupe abélien commutatif dit "**groupe multiplicatif**" par rapport à $\mathbb{Q}^*$.

Démonstration:

Démontrons donc que le symétrique existe pour la loi de multiplication $(.)$ tel que:

$$a \cdot a' = n = 1 \quad (5.127)$$

Puisque dans $\mathbb{Q}^*$ tout nombre peut se mettre sous la forme:

$$a = \frac{p}{q} \quad (5.128)$$

avec $p \in \mathbb{Z}, q \in \mathbb{Z}^*$.

Alors puisque:

$$a' = \frac{1}{a} = \frac{1}{\left(\frac{p}{q} \right)} = \frac{q}{p} \quad (5.129)$$

Il existe donc un symétrique à tout rationnel dans $\mathbb{Q}^*$ pour la loi de multiplication.

□C.Q.F.D.

Par définition, ou par construction, la division existe dans $\mathbb{Q}^*$ et est une opération interne. Mais est-elle associative telle que pour $\forall p, q, r, \in \mathbb{Q}^*$ nous ayons:

$$(p/q)/r = p/(q/r) \quad (5.130)$$

Démonstration:

Au fait, la démonstration est assez triviale si nous nous rappelons que la division se définit à partir de la loi de multiplication par l'inverse et que cette dernière loi est (elle!) associative. Ainsi, il vient:

$$(p/q)/r = \left(p \cdot \frac{1}{q} \right) \cdot \frac{1}{r} = p \cdot \left(\frac{1}{q} \cdot \frac{1}{r} \right) = p \cdot \left(\frac{1}{q \cdot r} \right) \neq p/(q/r) = p / \left(q \cdot \frac{1}{r} \right) = p \cdot \frac{r}{q} \quad (5.131)$$

Donc la loi de division n'est pas associative dans $\mathbb{Q}^*$.

□C.Q.F.D.

Nous pouvons aussi nous demander si la loi de division (/) est cependant commutative tel que la relation:

$$a/b = b/a \quad (5.132)$$

pour $\forall a, b \in \mathbb{Q}^*$?

Nous voyons très bien que cela n'est pas le cas puisque nous pouvons écrire cette dernière relation sous la forme:

$$\frac{a}{b} = \frac{b}{a} = \pm 1 \Rightarrow a = \pm b \quad (5.133)$$

Synthèse:

$\mathbb{Q}$ (lois)	(+)	(-)	(x)	(/)
Opération interne	oui	oui	oui	oui
Associative	oui	non	oui	non
Commutative	oui	non	oui	non
Élément neutre	oui (zéro "0")	non (0 pas neutre à gauche)	oui (un "1")	oui ("1" neutre à droite)
Élément abs.	non	non	oui (zéro "0")	oui ("0" au numérateur)
Symétrique	oui (signe opposé)	oui (signe opposé)	non (excepté dans $\mathbb{Q}^*$)	non

Tableau: 5.3 - Lois et leurs propriétés dans l'ensemble des rationnels

Nous avons donc les propriétés suivantes:

P1. $(\mathbb{Q}, \leq, \geq)$ est totalement ordonné

P2. $(\mathbb{Q}, +), (\mathbb{Q}^*, \times)$ sont indépendamment des groupes abéliens totalement ordonnés

P3. Zéro "0" est l'élément absorbant par rapport au groupe $(\mathbb{Q}^*, \times)$

P4. L'ensemble $\mathbb{Q}$ est un groupe abélien totalement ordonné par rapport aux lois d'addition et de multiplication que nous notons:

$$(\mathbb{Q}, +, \geq, \leq) \text{ et } (\mathbb{Q}^*, \times, \geq, \leq) \quad (5.134)$$

Les mêmes propriétés sont applicables à $\mathbb{R}$ et à $\mathbb{C}$ mais à la différence que ce dernier n'est pas ordonnable.

Cependant, il peut être compréhensible que pour $\mathbb{C}$ vous soyez sceptiques. Développons donc tout cela:

Nous devons nous assurer que la somme, la différence, le produit et le quotient de deux nombres de la forme $x + iy$ donne quelque chose d'encore de cette forme.

Additionnons les nombres $a + ib$ et $c + id$ où a, b, c et d sont des réels:

$$(a + ib) + (c + id) = a + ib + c + id = a + c + ib + id = (a + c) + i(b + d) \quad (5.135)$$

Donc l'addition est bien une loi interne commutative et associative pour laquelle il existe un élément neutre et symétrique dans l'ensemble des complexes.

Soustrayons les nombres $a + ib$ et $c + id$ où a, b, c et d sont ici encore, des réels:

$$(a + ib) - (c + id) = a + ib - c - id = a - c + ib - id = (a - c) + i(b - d) \quad (5.136)$$

Donc la soustraction est une opération interne; elle n'est ni commutative, ni associative elle n'a pas d'élément neutre à gauche et pas de symétrique.

Multiplions maintenant les nombres $a + ib$ et $c + id$ où a, b, c et d là toujours, des réels. Pour parvenir à nos fins, nous emploierons la distributivité de la multiplication par rapport à l'addition.

$$\begin{aligned} (a + ib)(c + id) &= a \cdot c + a \cdot i \cdot d + i \cdot b \cdot c + i \cdot b \cdot i \cdot d \\ &= a \cdot c + i \cdot a \cdot d + i \cdot b \cdot c + i^2 \cdot b \cdot d \\ &= a \cdot c + i \cdot a \cdot d + i \cdot b \cdot c + (-1) \cdot b \cdot d = (a \cdot c - b \cdot d) + i(a \cdot d + b \cdot c) \end{aligned} \quad (5.137)$$

Donc la loi de multiplication est bien une opération interne commutative, associative et distributive (!) pour laquelle il existe un élément neutre et symétrique dans $\mathbb{C}^*$ (voir ci-après) dans l'ensemble des complexes.

Une division est avant tout une multiplication par l'inverse. Prouver qu'il existe un inverse c'est prouver qu'il existe un symétrique pour la multiplication. Inversons donc le nombre $x + iy$ où x et y sont des

réels (différents de zéro):

$$\frac{1}{x + iy} = \frac{(x - iy)}{(x + iy)(x - iy)} = \frac{x - iy}{x^2 - (iy)^2} = \frac{x - iy}{x^2 + y^2} = \frac{x}{x^2 + y^2} - i \frac{y}{x^2 + y^2} \quad (5.138)$$

Donc l'inverse d'un nombre complexe est bien une opération interne non associative et non commutative pour laquelle il existe un élément neutre, et elle est symétrique. Il en est de même pour la division, qui correspond au produit par l'inverse d'un nombre complexe.

Voyons un exemple de groupe cyclique: Dans $\mathbb{C}$, considérons $G = \{1, i, -1, -i\}$ muni de la multiplication usuelle des nombres complexes. Alors $(G, \times)$ est évidemment un groupe abélien. Un tel groupe est aussi monogène car engendré par les puissances d'un de ses éléments: i (ou bien $-i$). Ce groupe monogène étant fini, il s'agit alors d'un groupe cyclique.

4.4. ANNEAUX

L'anneau est le coeur de l'algèbre commutative qui est la structure algébrique correspondant aux concepts collégiens d'addition, de soustraction, et de multiplication.

Définition: Un groupe commutatif (ou "groupe abélien") A est un "anneau" s'il est muni d'une seconde loi de composition interne vérifiant les propriétés suivante:

$$(A, +, *) \text{ est un anneau si } \Leftrightarrow \begin{cases} (A, +) \text{ est un groupe abélien} \\ \text{La loi } * \text{ est associative} \\ \text{La loi } * \text{ est distributive par rapport à la loi } + \end{cases}$$

Comme nous le savons déjà, l'élément neutre de la première loi de composition interne $+$ est noté "0" et appelé "zéro" de l'anneau. La deuxième loi interne est souvent notée par un point à mi-hauteur et appelée la "multiplication".

Remarques:

R1. Si de plus, la deuxième loi interne de composition $*$ est également commutative, l'anneau est dit "**anneau commutatif**". Nous rencontrons aussi des anneaux non-commutatifs dans lesquels la relation de commutativité n'est pas imposée ou ne s'impose pas et alors nous devons parfois l'imposer, il faut alors renforcer la propriété de l'élément neutre de cette deuxième loi en imposant à "1" d'être un élément neutre à la fois à droite et à gauche tel que: $1a = a1 = a$ (un exemple d'anneau non-commutatif est fourni par l'ensemble des matrices $n \times n$ à coefficients dans un anneau A , par exemple $M_n(\mathbb{R})$ - voir chapitre d'Algèbre Linéaire).

R2. Si de plus, il existe dans A un élément neutre pour la deuxième loi de composition interne $*$, et que cet élément neutre est l'unité "1" nous disons alors que l'anneau est un "**anneau unitaire**" et 1 est appelé "unité" de l'anneau. Si l'anneau est commutatif et possède un élément neutre pour la deuxième loi de composition interne alors nous parlons "**d'anneau commutatif unitaire**".

R3. Si $a * b = 0 \Rightarrow (a = 0 \text{ ou } b = 0)$, quels que soient les éléments a, b de A , l'anneau est dit "**anneau intègre**" ou "**anneau sans diviseurs de zéro**" (dans le cas contraire il est bien évidemment "non intègre").

R4. Un "**anneau factoriel**" est un anneau commutatif unitaire et intègre dans lequel le théorème fondamental de l'arithmétique (cf. chapitre de Théorie des Nombres) est vérifié.

Définitions:

D1. Un élément a d'un anneau A est un "**élément unité**" s'il existe $b \in A$ tel que $ab = ba = 1$. Si un tel b existe il est unique (nous en avons vu un exemple lors de notre étude des classes de congruence en théorie des nombres).

D2. Soit A un anneau. Nous disons que A possède des diviseurs de zéro s'il existe $a, b \in A$ avec $a \neq 0, b \neq 0$ et $a \cdot b = 0$. Les éléments a et b sont appelés des "**diviseurs de zéro**".

Remarques:

R1. Il est clair qu'un anneau est intègre si et seulement si il ne possède aucun diviseur de zéro.

R2. Les notions d'unité et de diviseurs de zéro sont incompatibles mais un élément d'un anneau peut être ni l'un ni l'autre. C'est le cas, par exemple, de tous les entiers $\neq \{0, -1, 1\}$ dans $\mathbb{Z}$. Ce ne sont ni des unités, ni des diviseurs de zéro.

Nous verrons un exemple important d'anneau dans le cadre de notre étude des polynômes (cf. chapitre de Calcul Algébrique) mais nous en avons déjà vu de très importants lors de notre étude des classes de congruence dans le chapitre de théorie des nombres.

Voyons quelques exemples d'anneaux: Lors de notre étude des groupes nous avons trouvé que les structures:

$$(\mathbb{Z}, +), (\mathbb{Q}, +), (\mathbb{R}, +), (\mathbb{C}, +) \quad (5.139)$$

sont tous les quatre des groupes abéliens et les trois premiers sont en plus totalement ordonnés.

La loi de division n'étant en aucun cas associative, nous pouvons nous restreindre à étudier pour chacun des groupes précités, le couple de lois: $(+)$ et $(\times)$.

Ainsi, il vient très vite que:

$$(\mathbb{Z}, +, \times), (\mathbb{Q}, +, \times), (\mathbb{R}, +, \times), (\mathbb{C}, +, \times) \quad (5.140)$$

constituent des anneaux commutatifs unitaires et intègres.

Remarque: Nous considérerons comme évident, à ce niveau du discours, que le lecteur aura remarqué que $\mathbb{Z}$ est un "sous-anneau" de $\mathbb{Q}$ dans le sens où les opérations définies sont internes à chacun des ensembles et que les éléments neutres et identité sont identiques et qu'il existe pour chaque élément de ces ensembles un opposé qui est dans le même ensemble. Nous allons approfondir le concept de sous-anneau un peu plus loin.

Soit A un anneau. Nous avons les propriétés suivantes:

$$P1. \quad a + b = a + c \Rightarrow (b = c) \quad \forall a, b, c \in A$$

$$P2. \quad 0 \cdot a = 0 \quad \forall a \in A$$

$$P3. \quad (-1) \cdot a = -a \quad \forall a \in A$$

Démonstrations:

DM1. La propriété P1 découle de la définition D4 vu tout au début de la partie concernant les structures algébriques (tout élément possède un opposé/symétrique). En effet, nous pouvons additionner à l'égalité $a + b = a + c$ l'élément $-a$. Nous obtenons alors $(-a) + a + b = (-a) + a + c$ par l'existence de l'opposé cela donne $0 + b = 0 + c$ d'où $b = c$.

DM2. La propriété P2 découle des définitions D3 (existence de l'élément neutre), D4 (existence de l'opposé/symétrique), D5 (distributivité par rapport à l'autre loi) ainsi que de la propriété P1 ci-dessus. En effet, nous avons:

$$0 \cdot a + a = 0 \cdot a + 1 \cdot a = (0 + 1) \cdot a = 1 \cdot a = a \quad (5.141)$$

Nous avons donc $0 \cdot a + a = a$. La propriété P1 ci-dessus permet de conclure que $0 \cdot a = 0$ (nous pourrions discuter de la pertinence de ce genre de démonstration...).

DM3. La propriété P3. se montre à l'aide de P2. Nous avons:

$$0 = 0 \cdot a = (1 + (-1)) \cdot a = 1 \cdot a + (-1) \cdot a \quad (5.142)$$

en ajoutant $-a$ à cette dernière égalité, nous avons:

$$-a = (-1) \cdot a \quad (5.143)$$

□C.Q.F.D.

4.4.1. SOUS-ANNEAU

Définition: Soit A un anneau et $S \subset A$ un sous-ensemble de A . Nous disons que S est un "**sous-anneau**" de A si:

P1. $0 \in S$ (élément neutre de A est aussi celui de S)

P2. $a \in S \Rightarrow -a \in S$

P3. $a, b \in S \Rightarrow a + b \in S$

P4. $a, b \in S \Rightarrow a \cdot b \in S$

 **Exemple:**

L'anneau $\mathbb{Z}$ est un sous-anneau de $\mathbb{Q}, \mathbb{R}$

4.5. CORPS

Définition: Nous désignons un ensemble de nombres par le terme "**corps**" si:

$$(C, +, *) \text{ est un corps si } \Rightarrow \begin{cases} (C, +, *) \text{ est un anneau unitaire} \\ (C - \{0\}, *) \text{ est un groupe} \end{cases}$$

Donc un corps est un anneau non nul dans lequel tout élément non nul est inversible ou en d'autres termes: un anneau dont tous les éléments non nuls sont des unités est un corps.

Remarques:

R1. Si la loi interne $*$ est également commutative, le corps est dit "**corps commutatif**".

R2. Les quaternions (cf. chapitre sur les Nombres) forment par exemple un corps non commutatif pour l'addition et la multiplication.

Voyons des exemples de corps parmi les anneaux unitaires suivant:

$$(\mathbb{Z}, +, \times), (\mathbb{Q}, +, \times), (\mathbb{R}, +, \times), (\mathbb{C}, +, \times) \quad (5.144)$$

Il nous faut d'abord déterminer lesquels ne constituent pas des groupes par rapport à la loi interne de multiplication ($\times$).

Comme nous l'avons déjà vu dans notre étude des groupes précédemment, il est évident qu'il nous faut éliminer $(\mathbb{Z}, +, \times)$ à cause de l'existence des inverses qui n'est pas assurée dans cet ensemble.

Ainsi, les corps fondamentaux de l'arithmétique sont:

$$(\mathbb{Q}, +, \times), (\mathbb{R}, +, \times), (\mathbb{C}, +, \times) \quad (5.145)$$

et puisque la loi de multiplication ($\times$) est commutative dans ces ensembles, nous pouvons affirmer que ces corps sont également des corps commutatifs.

Nous avons souvent dans les petites classes le schéma suivant pour le corps le plus important:

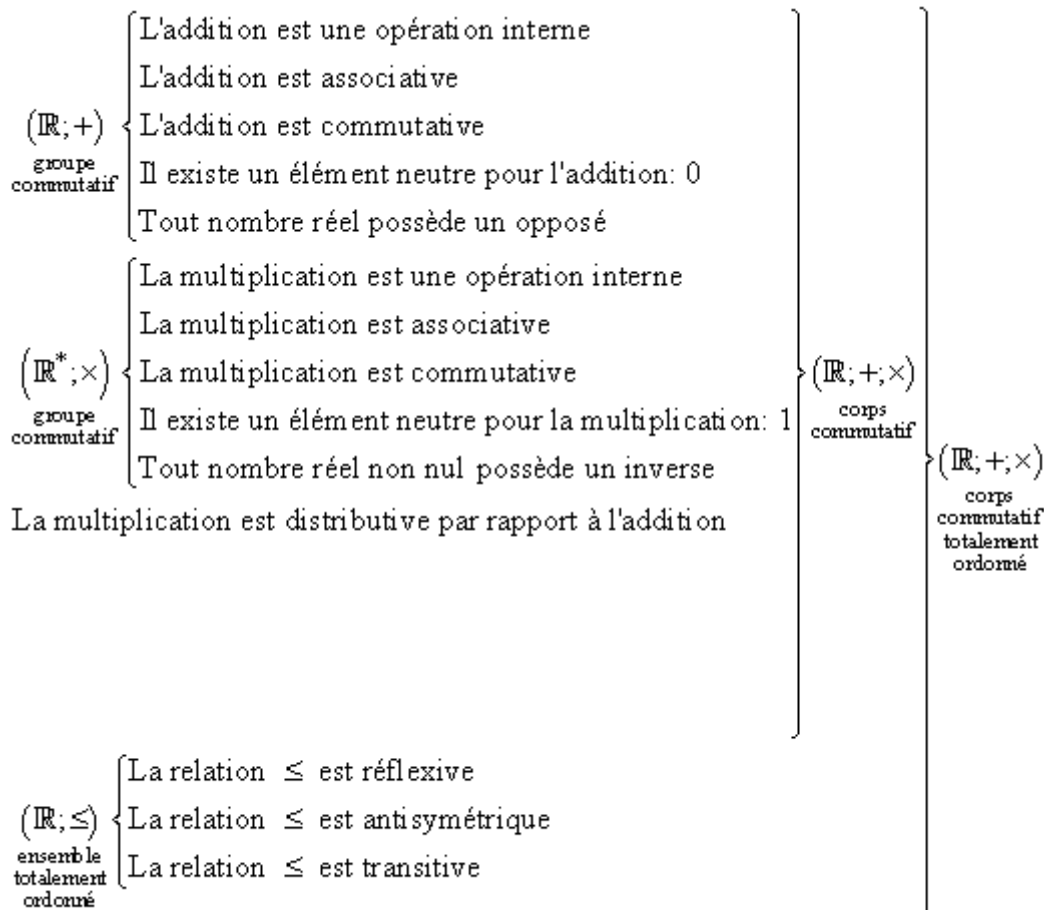


Figure: 5.16 - Propriétés classique de l'ensemble des réels

Ainsi, nous appellerons "corps" un système C de nombres réels ou complexes a tels que la somme, la différence, le produit et le quotient de deux quelconques de ces nombres a appartiennent au même système C .

Nous énonçons également cette propriété de la manière suivante: les nombres d'un corps se reproduisent par les opérations rationnelles (addition, soustraction, multiplication, division). Ainsi, il est évident que le nombre zéro ne pourra jamais former le dénominateur d'un quotient et l'ensemble des entiers ne peut former un corps car la division dans l'ensemble des nombres entiers ne donne pas nécessairement un résultat dans ce même ensemble.

4.6. ESPACES VECTORIELS

Lorsque nous définissons un "vecteur" (cf. chapitre de Calcul Vectoriel), nous faisons habituellement référence à un "espace euclidien" (cf. aussi chapitre de Calcul Vectoriel) de n dimensions de $\mathbb{R}^n$. Cependant, la notion d'espace vectoriel est beaucoup beaucoup plus vaste que ce dernier qui ne représente qu'un cas particulier.

Définition: Un "espace vectoriel (EV)" ou " K -espace vectoriel" (abrégé: K -ev) sur le corps K (nous prendrons fréquemment pour ce corps $\mathbb{R}$ ou $\mathbb{C}$) est un ensemble $(E, +, *)$ possédant les propriétés:

$$(E, +, *) \left\{ \begin{array}{l} (E, +) \text{ est un groupe abélien} \\ * \text{ est une loi externe définie par: } \begin{cases} *: E \times K \rightarrow E \\ (x, a) \rightarrow a * x \end{cases} \end{array} \right. \quad (5.146)$$

Nous avons donc deux lois de composition (en prenant les notations traditionnelles des vecteurs qui sera peut-être plus parlante et utile pour la suite...):

1. Une loi de composition interne: l'addition notée $+$ qui vérifie:

1.1. Associativité: $\forall \vec{x}, \vec{y}, \vec{z} \in E: (\vec{x} + \vec{y}) + \vec{z} = \vec{x} + (\vec{y} + \vec{z})$

1.2. Commutativité: $\forall \vec{x}, \vec{y} \in E: \vec{x} + \vec{y} = \vec{y} + \vec{x}$

1.3. Élément neutre: $\exists \vec{0} \in E: \forall \vec{x} \in E, \vec{x} + \vec{0} = \vec{x}$

1.4. Élément opposé: $\forall \vec{x} \in E, \exists \vec{x}' \in E: \vec{x} + \vec{x}' = \vec{0}$

2. Une loi de composition externe: la multiplication par un scalaire, notée $\cdot$, qui vérifie:

2.1. Associativité: $\forall \lambda, \mu \in K, \forall \vec{x} \in E: \lambda \cdot (\mu \cdot \vec{x}) = (\lambda \cdot \mu) \cdot \vec{x}$

2.2. Distributivité à droite par rapport au corps K : $\forall \lambda, \mu \in K, \forall \vec{x} \in E: (\lambda + \mu) \cdot \vec{x} = \lambda \cdot \vec{x} + \mu \cdot \vec{x}$

2.3. Distributivité à gauche par rapport à E : $\forall \lambda \in K, \forall \vec{x}, \vec{y} \in E: \lambda \cdot (\vec{x} + \vec{y}) = \lambda \cdot \vec{x} + \lambda \cdot \vec{y}$

2.4. Élément neutre (de K sur E): $\forall \vec{x} \in E: 1 \cdot \vec{x} = \vec{x}$

Remarques:

R1. Nous disons alors que l'espace vectoriel a une "**structure algébrique vectorielle**" et que ces éléments sont des "**vecteurs**", les éléments de K des "**scalaires**".

R2. Les opérations respectives s'utilisent fréquemment comme l'addition et la multiplication que nous connaissons déjà très bien sur $\mathbb{R}$, ce qui est bien commode pour nos habitudes....

R3. Dorénavant, pour distinguer les éléments du corps K et de l'ensemble E , nous noterons ceux de K par des lettres grecques et ceux de E par des lettres latines majuscules.

R4. Outre les cinq propriétés énumérées ci-dessus, il ne faut pas oublier d'ajouter les cinq autres propriétés du groupe abélien (opération interne, commutativité, associativité, élément neutre, élément inverse). Ce qui nous fait donc au total dix propriétés à respecter.

Il est inutile de démontrer que ces propriétés sont respectées pour $\mathbb{R}^n$ et, par conséquent pour $\mathbb{R}^2$. Nous pouvons cependant nous poser la question à propos de certains sous-ensembles de $\mathbb{R}^n$.

Exemples:

E1. Considérons la région rectangulaire illustrée dans la figure (a) et en perspective dans la figure (c) ci-dessous:

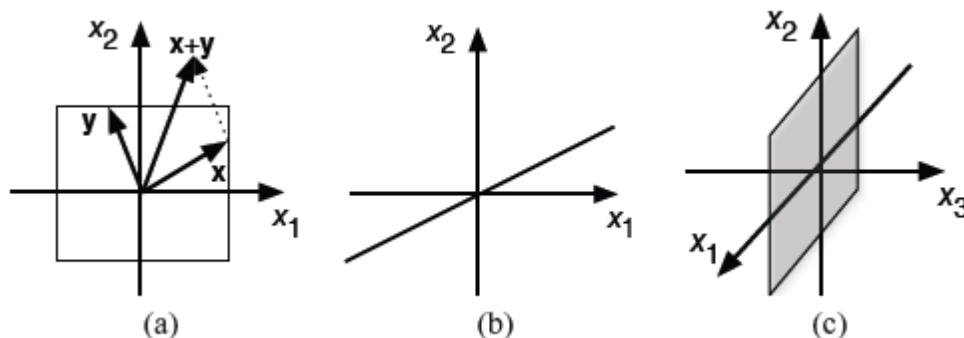


Figure: 5.17 - Exemple du concept d'espace vectoriel

Ce sous-ensemble de $\mathbb{R}^2$ n'est pas un espace vectoriel car, entre autres, la propriété d'opération interne du groupe abélien n'est pas satisfaite. En effet, si nous prenons deux vecteurs à l'intérieur du rectangle et que nous les additionnons, il se peut que le résultat sorte du rectangle. Par contre, il est facile de voir que la droite (infinie) illustrée dans la figure (b) respecte toutes les propriétés énumérées précédemment et, par conséquent, définit un espace vectoriel. Notons bien, cependant, que cette droite se doit de passer par l'origine, sinon la propriété d'élément neutre du groupe abélien ne serait pas respectée (l'élément neutre n'existant plus).

E2. Un autre exemple d'un espace vectoriel est l'ensemble $\mathcal{P}^2$ des polynômes de degré deux ou moins (cf. chapitre de Calcul Algébrique). Par exemple, deux éléments de cet espace sont:

$$\begin{aligned} x &= t^2 + 2t + 3 \\ y &= -t + 5 \end{aligned} \quad (5.147)$$

Cet ensemble respecte les 10 propriétés d'un espace vectoriel. En effet, si nous additionnons deux polynômes de degré deux ou moins, nous obtenons un autre polynôme de degré deux ou moins. Nous pouvons aussi multiplier un polynôme par un scalaire sans changer l'ordre (ou degré) de celui-ci, etc. Nous pouvons donc représenter un polynôme par des vecteurs dont les termes sont les coefficients du polynôme.

Mentionnons que nous pouvons aussi former des espaces vectoriels avec des ensembles de fonctions plus générales que des polynômes. Il importe seulement de respecter les dix propriétés fondamentales d'un espace vectoriel !

Ainsi défini, un espace vectoriel E sur K est une action de K^* sur $(E, +)$ qui est compatible avec la loi de groupe (par extension un "automorphisme" - voir la définition plus loin - sur $(E, +)$).

Définition: Soit E un espace vectoriel, nous appelons "**sous-espace vectoriel**" (SEV) F de E un sous-ensemble de E si et seulement si (notation des matheux):

$$\begin{aligned} F &\neq \emptyset \\ \forall (X, Y) \in F^2, X + Y &\in F \quad (5.148) \\ \forall \lambda \in K, \forall X \in F, \lambda X &\in F \end{aligned}$$

ou en utilisant une autre notation (celle utilisée plutôt par les physiciens):

$$\begin{aligned} F &\neq \emptyset \\ \forall (\vec{x}, \vec{y}) \in F^2, \vec{x} + \vec{y} &\in F \quad (5.149) \\ \forall \lambda \in K, \forall \vec{x} \in F, \lambda \vec{x} &\in F \end{aligned}$$

4.7. ALGÈBRES

Une "**C-algèbre** A " où C est un corps commutatif (appelée aussi souvent " K -algèbre A " pour " K örper" en allemand)), est un ensemble A muni de deux lois de composition internes $+$ (addition) et $\times$ (produit) et d'une loi externe (multiplication) à domaine d'opérateurs C (produit par un scalaire) si et seulement si:

$$(A, C, +, \cdot, \times) \text{ est une } C\text{-algèbre si} \Rightarrow \begin{cases} (A, +, \cdot) \text{ est un } C\text{-espace vectoriel} \\ (A, +, \times) \text{ est un anneau unitaire} \\ \forall \lambda \in C, \forall a, b \in A, (\lambda a) \cdot b = a \cdot (\lambda b) = \lambda(a \cdot b) \end{cases} \quad (5.150)$$

Exemples:

E1. Pour reprendre un exemple dans la lignée de celui sur les exemples vectoriels, l'espace euclidien $\mathbb{R}^3$ muni de l'addition $(+)$, de la multiplication $(\cdot)$ et du produit vectoriel $(\times)$ est une $\mathbb{R}$ -algèbre non associative et non commutative notée $(\mathbb{R}^3, \mathbb{R}, +, \cdot, \times)$.

E2. $\mathbb{C}$ est une $\mathbb{R}$ -algèbre (un nombre complexe pouvant être vu comme un vecteur à deux composantes selon ce que nous avons vu dans le chapitre des Nombres).

5. HOMOMORPHISMES

Le concept d'homomorphismes (du grec homoios = semblable et morphê = forme) a été défini par les mathématiciens car permettant de mettre en évidence des propriétés remarquables des fonctions en particulier avec leurs structures, leur noyau, et de ce que nous appelons les "idéaux" (voir plus loin). Ils nous permettront ainsi d'identifier une structure algébrique d'une autre.

Définitions:

D1. Si $(A, *)$ et $(B, \circ)$ sont deux magmas (peu importe la notation utilisée pour les lois internes), une application f de A dans B est un "**homomorphisme de magma**" ou "**morphisme de magma**" (par abus de langage nous écrivons parfois juste "homomorphisme") si:

$$f(a * b) = f(a) \circ f(b) \quad \forall a, b \in A \quad (5.151)$$

en d'autres termes, si l'image d'un composé dans A est le composé des images dans B .

D2. Si $(A, *)$ et $(B, \circ)$ sont deux monoïdes, une application f de A dans B est un "**homomorphisme de monoïde**" si:

$$\begin{aligned} f(a * b) &= f(a) \circ f(b) \quad \forall a, b \in A \\ f(1_A) &= 1_B \end{aligned} \quad (5.152)$$

où $1_A, 1_B$ sont les éléments neutres respectifs des monoïdes A, B .

D3. Si A, B sont deux anneaux, un "**homomorphisme d'anneaux**" (très important pour le chapitre de Cryptographie!) de A dans B est une application $f: A \rightarrow B$ telle que nous ayons pour tout $a, a' \in A$:

$$\begin{aligned} f(1_A) &= 1_B \\ f(a + a') &= f(a) + f(a') \\ f(a \cdot a') &= f(a)f(a') \end{aligned} \quad (5.153)$$

où $1_A, 1_B$ sont les éléments neutres des anneaux A, B par rapport à la multiplication.

Soit $f: A \rightarrow B$ un homomorphisme d'anneaux. Alors:

$$P1. f(0) = 0$$

$$P2. f(-a) = -f(a)$$

$$P3. \text{ Si } a \text{ est une unité de } A, \text{ alors } f(a) \text{ est une unité de } B \text{ et } f(a)^{-1} = f(a^{-1})$$

Démonstrations:

DM1. Par $f(a + a') = f(a) + f(a')$, nous avons $f(a + 0) = f(a) + f(0)$. En ajoutant $-f(a)$ des deux côtés de l'égalité, nous obtenons $0 = f(0)$

DM2. La propriété P2 découle aussi de $f(a + a') = f(a) + f(a')$ et de la propriété P1. En effet, nous

avons $f(0) = f(a + (-a)) = f(a) + f(-a) = 0$. En additionnant $-f(a)$ aux deux côtés de la dernière égalité, nous obtenons $f(-a) = -f(a)$.

DM3. Soient $a, b \in A$ tels que $ab = ba = 1$. Alors par $f(a \cdot b) = f(a)f(b)$ et $f(1_A) = 1_B$, nous avons $1 = f(1) = f(ab) = f(a)f(b)$ et de même $1 = f(b)f(a)$ ce qui montre que $f(b)$ est l'inverse de $f(a)$ si b est l'inverse de a .

□C.Q.F.D.

Montrons maintenant qu'un homomorphisme d'anneaux $f : A \rightarrow B$ est injectif si et seulement si l'élément 0 est la seule pré-image de 0 (et donc réciproquement), ce qui se note techniquement:

$$\text{Ker}(f) = \{0\} \quad (5.154)$$

c'est-à-dire que le noyau est trivial.

Démonstration:

La condition est clairement nécessaire. Montrons qu'elle est suffisante:

Nous supposons donc que $\text{Ker}(f) = 0$. Soit $a, a' \in A$ tel que $f(a) = f(a')$. Alors comme nous avons un homomorphisme d'anneaux nous pouvons écrire:

$$0 = f(a) - f(a') = f(a - a') \quad (5.155)$$

qui implique que $a - a' = 0$ donc que $a = a'$.

Ce qui montre que f est injectif si c'est un homomorphisme et que $\text{Ker}(f) = 0$ en est effectivement une condition suffisante.

□C.Q.F.D.

D4. Soient $(A, +)$ et $(B, *)$, deux groupes et f une application $f : A \rightarrow B$. Nous disons que f est un "**homomorphisme de groupe**" si (nous pourrions tout aussi bien mettre $*$ au lieu de $+$ dans le premier groupe et $+$ au lieu de $*$ dans le deuxième groupe, la définition resterait la même en remplaçant simplement les opérateurs respectifs!):

$$\boxed{\begin{aligned} f(x + y) &= f(x) * f(y) \quad \forall x, y \in A \\ f(1_A) &= 1_B \\ f(x^{-1}) &= (f(x))^{-1} \quad \forall x \in A \end{aligned}} \quad (5.156)$$

où $1_A, 1_B$ sont les éléments neutres respectifs des groupes A, B . Nous remarquons que la seule différence entre un homomorphisme d'anneau et un homomorphisme de groupe est que ce dernier à deux lois au lieu d'une et que nous y rajoutons le concept d'inverse.

Ceci dit, la troisième proposition ci-dessus est en fait une conséquence de la définition composée uniquement des deux premières lignes. Effectivement, considérons un homomorphisme f entre les groupes $(A, +)$ et $(B, *)$ avec 1_A et 1_B respectivement les éléments neutres de A et B .

Nous avons alors:

$$\forall x \in A: 1_B = f(1_A) = f(x + x^{-1}) = f(x) * f(x^{-1}) \quad (5.157)$$

d'où:

$$\forall x \in A: 1_B = f(x) * f(x^{-1}) \quad (5.158)$$

et donc:

$$\forall x \in A: f(x) = \frac{1_B}{f(x^{-1})} = \left(f(x^{-1})\right)^{-1} \quad (5.159)$$

D5. Soit f une application $f: A \rightarrow B$ d'un corps vers un autre. Nous disons que f est un "**homomorphisme de corps**" si f est un homomorphisme d'anneaux...

Effectivement, le fait que l'homomorphisme de corps soit le même que celui d'un anneau tient juste au fait que la différence entre les deux structures est que les éléments du corps sont tous inversibles (aucune loi ou propriété de loi ne diffère entre les deux selon leur définition).

Montrons maintenant que tout homomorphisme de corps est injectif ("**homomorphisme injectif**") en se rappelant que plus haut nous avons démontré que tout homomorphisme d'anneaux l'était!

Démonstration:

Si a est différent de 0 et $b = a^{-1}$ (nous utilisons ici la propriété que les éléments d'un corps sont inversibles!) alors:

$$1 = f(1) = f(a \cdot b) = f(a) \cdot f(b) \quad (5.160)$$

Donc lorsque a est différent de zéro $f(a)$ est différent de 0 ce qui prouve que $\text{Ker}(f) = \{0\}$ et donc que f est injective.

□C.Q.F.D.

D6. Soient A et B deux K -ev et $f: A \rightarrow B$ une application de A dans B . Nous disons que f est une "**application linéaire**" ou "**homomorphisme d'espaces vectoriels**" (il est sous-entendu que c'est relativement aux lois indiquées et pour l'application choisie) si:

$$\begin{aligned} f(x+y) &= f(x) + f(y) \quad \forall (x,y) \in A \\ f(\lambda x) &= \lambda f(x) \quad \forall x \in A, \forall \lambda \in K \end{aligned} \quad (5.161)$$

et nous notons $L(A,B)$ l'ensemble des applications linéaires.

Remarques:

R1. Nous avons déjà défini plus haut le concept d'application linéaire mais n'avons pas précisé que les deux ensembles A et B étaient des K -ev.

R2. L'application linéaire est appelée "**forme linéaire**" si et seulement si $B = K$

D7. Si l'homomorphisme est bijectif nous dirons alors que f est un "**isomorphisme**". S'il existe un isomorphisme entre A et B , nous disons que A et B sont "**isomorphes**" et nous noterons cela $A \simeq B$.

Remarque: L'isomorphisme permet au fait d'identifier deux ensembles munis d'une structure algébrique identique (que ce soit groupe, anneau, etc.) mais dont les éléments sont nommés d'une façon différente.

D8. Si l'homomorphisme f est une application uniquement interne, nous dirons alors que f est un "**endomorphisme**" (en d'autres termes, nous avons un endomorphisme si dans la définition de l'homomorphisme nous avons $A=B$).

Remarque: Si nous avons un endomorphisme f de E , f est donc restreint à $\text{Im}(f)$. Donc le terme "endomorphisme" veut juste dire que l'application f arrive dans E et pas qu'elle touche tous les éléments de E . Nous avons $f(E) \subset E$ et pas forcément $f(E) = E$ car dans ce dernier cas nous disons que f est surjective comme nous l'avons déjà vu.

D9. Si l'endomorphisme f est en plus bijectif (donc en d'autres termes si l'homomorphisme est un endomorphisme et un isomorphisme), nous dirons alors que f est un "**automorphisme**".

5.1. IDÉAL

Définition: Soit A un anneau commutatif. Un sous-ensemble $S \subset A$ est un "**idéal**" si:

P1. $\alpha + \alpha' \in S$ pour tout $\alpha, \alpha' \in S$

P2. $r \cdot \alpha \in S$ pour tout $\alpha \in S$ et tout $r \in A$

En d'autres termes, un idéal est un sous-ensemble fermé pour l'addition et stable pour la multiplication par un élément quelconque de A .

Exemple:

L'ensemble des nombres pairs est par un exemple d'idéal de l'ensemble des nombres naturels.

Remarque: Les idéaux $S = \{0\}$ et $S = A$ sont appelés les "**idéaux triviaux**".

Pour savoir si un idéal est égal à tout l'anneau, il est utile d'utiliser la propriété suivante qui spécifie que si A est un anneau et I un idéal de A , alors si $1 \in I$ nous avons $I = A$.

Démonstration:

Ceci résulte de la propriété P2 de la définition d'un idéal:

Pour tout $r \in A$, nous avons $r = r \cdot 1 \in I$ car $1 \in I$.

□C.Q.F.D.

Un premier exemple d'idéal est donné par le noyau d'un homomorphisme d'anneaux. Effectivement, démontrons que le noyau d'un homomorphisme $f: R \rightarrow S$ est un idéal de R .

Démonstration:

Soient $a, a' \in \text{Ker}(f)$. Alors:

$$f(a + a') = f(a) + f(a') = 0 + 0 = 0 \quad (5.162)$$

ce qui montre que $a + a' \in \text{Ker}(f)$. Soit $r \in R$, alors:

$$f(ra) = f(r)f(a) = f(r) \cdot 0 = 0 \quad (5.163)$$

ce qui montre que $r \cdot a \in \text{Ker}(f)$.

□C.Q.F.D.

Proposition: Soit A un anneau et soit $a \in A$. Le sous-ensemble:

$$\{ax \mid x \in A\} \quad (5.164)$$

noté (a) ou aA , est un idéal (nous allons voir un exemple concret après la prochaine définition).

Définitions:

D1. Un idéal $I \neq A$ d'un anneau A est dit "**idéal principal**" s'il existe $a \in A$ tel que $I = (a)$.

D2. Un anneau dont tous les idéaux sont principaux est dit "**anneau principal**".

Montrons maintenant que l'anneau $\mathbb{Z}$ est principal (car tous ses idéaux sont principaux).

Démonstration:

Soit I un idéal de $\mathbb{Z}$ (il est facile d'en choisir un: par exemples tous les multiples de 2 ou de 3, etc.). Soit $r \in I$ le plus petit entier positif non nul de I . Nous allons montrer que $I = r\mathbb{Z}$:

Soit a un élément quelconque de I . La division euclidienne nous permet d'écrire:

$$a = qr + r' \quad (5.165)$$

avec $0 \leq r' < r$ (nous l'avons déjà démontré).

Mais comme $r' = a - qr$ et que $a, r \in I$, par la définition d'un idéal, nous avons $r' \in I$ (la somme ou différence des éléments d'un idéal appartenant à l'idéal). Par choix de r (r' étant inférieur à r) ceci implique que $r' = 0$ et donc que $a = rq$.

Ainsi tout élément de I est un multiple de r et donc:

$$I = r\mathbb{Z} \quad (5.166)$$

□C.Q.F.D.

La démonstration ci-dessus n'utilise que la division euclidienne sur $\mathbb{Z}$. Nous pouvons alors généraliser ce résultat aux anneaux qui possèdent une division euclidienne. Ainsi, par exemple, l'anneau $k[X]$ des polynômes (cf. chapitre de Calcul Algébrique) à coefficients dans un corps k est un anneau principal car il possède une division euclidienne.

Démonstration:

Soit I un idéal de $k[X]$. Notons d le plus petit degré que puisse avoir un polynôme non nul de I . Si $d = 0$ alors $1 \in I$ et donc $I = 1 \cdot K[X] = K[X]$. Autrement, soit $a(X)$ un polynôme de degré d . Si $u(X) \in I$ alors on peut diviser $u(X)$ par $a(X)$. Il existe $q(X), r(X) \in K[X]$ tels que $\deg(r) < \deg(a) = d$ et $u(X) = q(X) \cdot a(X) + r(X)$. Donc $r(X) \in I$ ce qui entraîne $r = 0$ (autrement contradiction avec la minimalité de d). Par suite, $u(X) = q(X) \cdot a(X)$. Nous venons de montrer que $I = a(X) \cdot K[X]$

□C.Q.F.D.

Ainsi, les seuls idéaux de $\mathbb{Z}$ sont ceux de la forme $m\mathbb{Z}$. De plus si nous avons d et m qui sont des entiers > 1 . Alors $m\mathbb{Z} \subset d\mathbb{Z}$ si et seulement si $d \mid m$.

Démonstration:

Si $d \mid m$ alors il existe n avec $m = d \cdot n$. Soit $m \cdot a$ un élément de $m\mathbb{Z}$. Alors:

$$m \cdot a = d \cdot n \cdot a \in d\mathbb{Z} \quad (5.167)$$

ce qui montre que $m\mathbb{Z} \subset d\mathbb{Z}$.

Réciproquement, si $m \in d\mathbb{Z}$ ceci implique que m est de la forme $d \cdot n$ et ceci prouve que d divise m .

□C.Q.F.D.

Démontrons aussi qu'un anneau R est un corps si et seulement s'il ne possède que les idéaux triviaux $\{0\}, R$.

Démonstration:

Montrons que la condition est nécessaire: Soit I un idéal non nul de R et $r \in I$ un élément non nul. Par hypothèse (qu'il s'agit d'un corps), il est inversible, c'est-à-dire qu'il existe $t \in R$ tel que $r \cdot t = t \cdot r = 1$. Ceci implique que $1 \in I$ et donc, par un résultat obtenu plus haut $I = R$.

Réciproquement, supposons que tout idéal $I \neq R$ soit l'idéal nul. Alors si $r \in R$ est un élément non nul de R , l'idéal principal (r) doit être égal à R . Mais ceci implique que $1 \in (r)$ et donc qu'il existe $x \in R$ avec $r \cdot x = 1$ ce qui montre que r est inversible. L'anneau R est donc un corps.

□C.Q.F.D.

Cette caractérisation va nous permettre de démontrer facilement que tout homomorphisme partant d'un corps est injectif. Soit que si $f : R \rightarrow S$ est un homomorphisme où R est un corps, alors f est injectif.

Démonstration:

Nous mettons ensemble ce qui a été vu jusque-là. Nous avons démontré plus haut que le noyau $\text{Ker}(f)$ d'un homomorphisme est un idéal. Mais nous avons également démontré plus haut que nous avons soit $\text{Ker}(f) = \{0\}$ soit $\text{Ker}(f) = R$ (car l'anneau R est un corps si et seulement s'il ne possède que les idéaux triviaux).

Mais vu que $f(1) = 1 \neq 0$ (de par la définition d'un homomorphisme) il s'ensuit qu'il reste $\text{Ker}(f) = \{0\}$ (puisque nous avons démontré que si A est un anneau et I un idéal de A alors si $1 \in I$ alors $I = R$). Ceci implique par un théorème précédent (où nous avons démontré que si $\text{Ker}(f) = \{0\}$ l'homomorphisme est injectif) que... f est injective.

□C.Q.F.D.

Etudions maintenant les homomorphismes dont l'anneau de départ est $\mathbb{Z}$. Soit A un anneau et $f : \mathbb{Z} \rightarrow A$ un homomorphisme. Par définition d'un homomorphisme et par ses propriétés, il faut que $f(0) = 0$ et $f(1) = 1$. Mais il faut encore que:

$$f(n) = \underbrace{f(1 + 1 + \dots + 1)}_n = \underbrace{1_A + 1_A + \dots + 1_A}_n = n \cdot 1_A \quad (5.168)$$

pour tout $n \in \mathbb{Z}$. Ainsi f est complètement déterminé par la donnée de $f(1)$ et est donc unique. Réciproquement, nous montrons que l'application $f : \mathbb{Z} \rightarrow A$ définie par:

$$f(n) = n \cdot 1_A \quad (5.169)$$

est un homomorphisme d'anneaux. En résumé, il existe un et un seul homomorphisme de $\mathbb{Z}$ dans un anneau quelconque A .

Définition: Soient A un anneau et $f : \mathbb{Z} \rightarrow A$ l'unique homomorphisme défini précédemment. Si f est injectif, nous dirons que A est de "caractéristique nulle". Sinon, $\text{Ker}(f)$ est un idéal non trivial de $\mathbb{Z}$ et comme $\mathbb{Z}$ est dès lors principal (comme nous l'avons démontré plus haut) il est de la forme $m\mathbb{Z}$ avec $m > 0$. L'entier m est appelé la "caractéristique de A ".

Remarque: Moins formellement, la caractéristique d'un anneau est le plus petit entier positif m tel que $m \cdot 1_A = 0$. S'il n'y en a pas, alors la caractéristique est nulle.

Exemple:

L'anneau $\mathbb{Z}$ est de caractéristique nulle car l'unique homomorphisme $f: \mathbb{Z} \rightarrow \mathbb{Z}$ est l'identité. Il est donc injectif. Les injections $\mathbb{Z} \rightarrow \mathbb{Q}, \mathbb{Z} \rightarrow \mathbb{R}$ montrent que $\mathbb{Q}, \mathbb{R}$ (et $\mathbb{C}$ également) sont des corps de caractéristique nulle.

Nous nous proposons maintenant de démontrer que la caractéristique d'un anneau intègre (et en particulier d'un corps) est égale 0 ou à un premier p .

Démonstration:

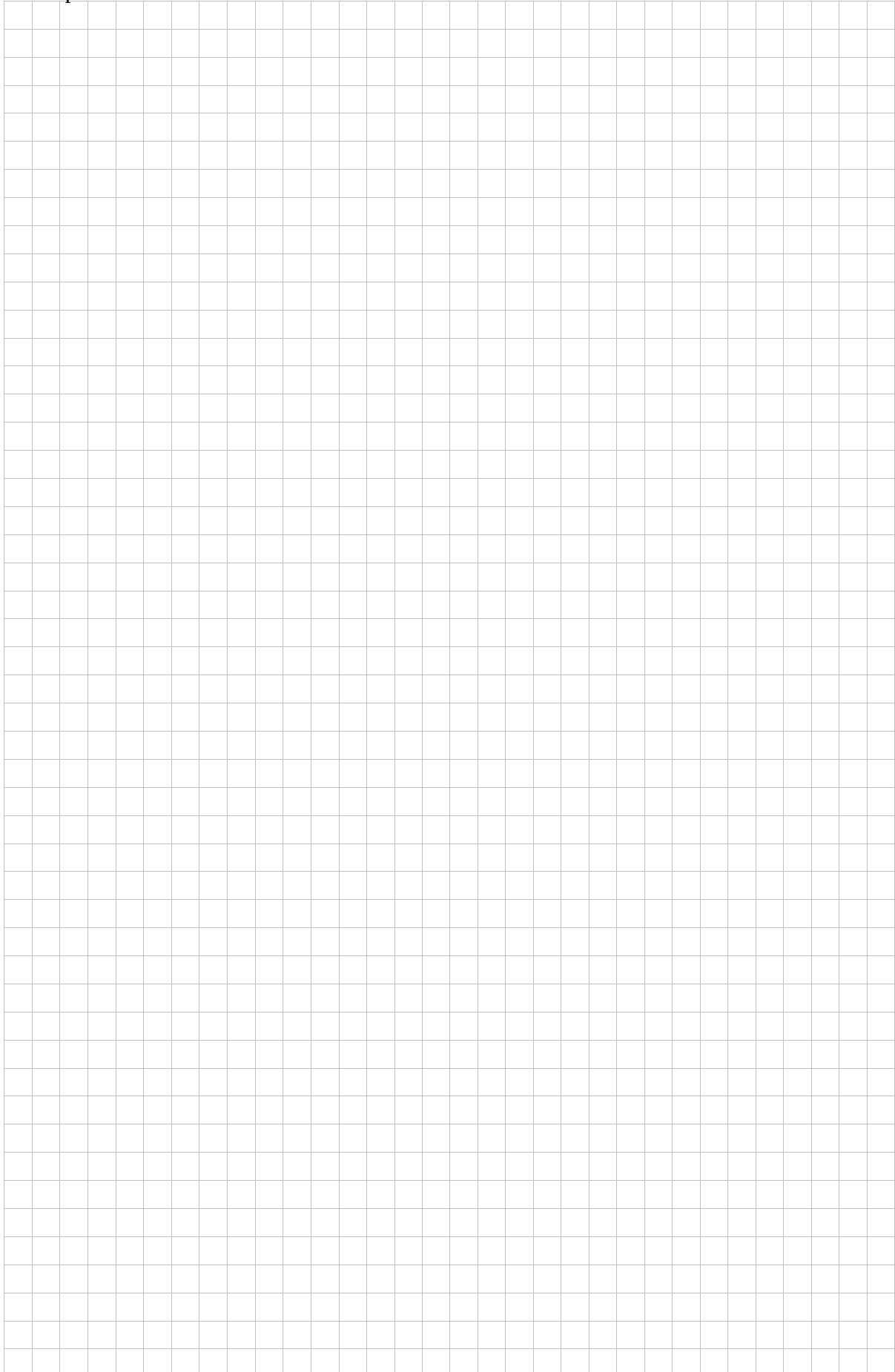
Nous montrons la contraposée. Soit A un anneau de caractéristique $m \neq 0$ avec m non premier. Il existe alors des entiers naturels $n, r < m$ tels que $m = n \cdot r$. Soit $f: \mathbb{Z} \rightarrow A$ l'unique homomorphisme (défini précédemment). Par définition (de l'idéal) de m , nous avons $f(m) = 0$ mais $f(r) \neq 0 \neq f(n)$. Mais alors $f(r)f(n) = f(r \cdot n) = f(m) = 0$ ce qui montre que A n'est pas intègre.

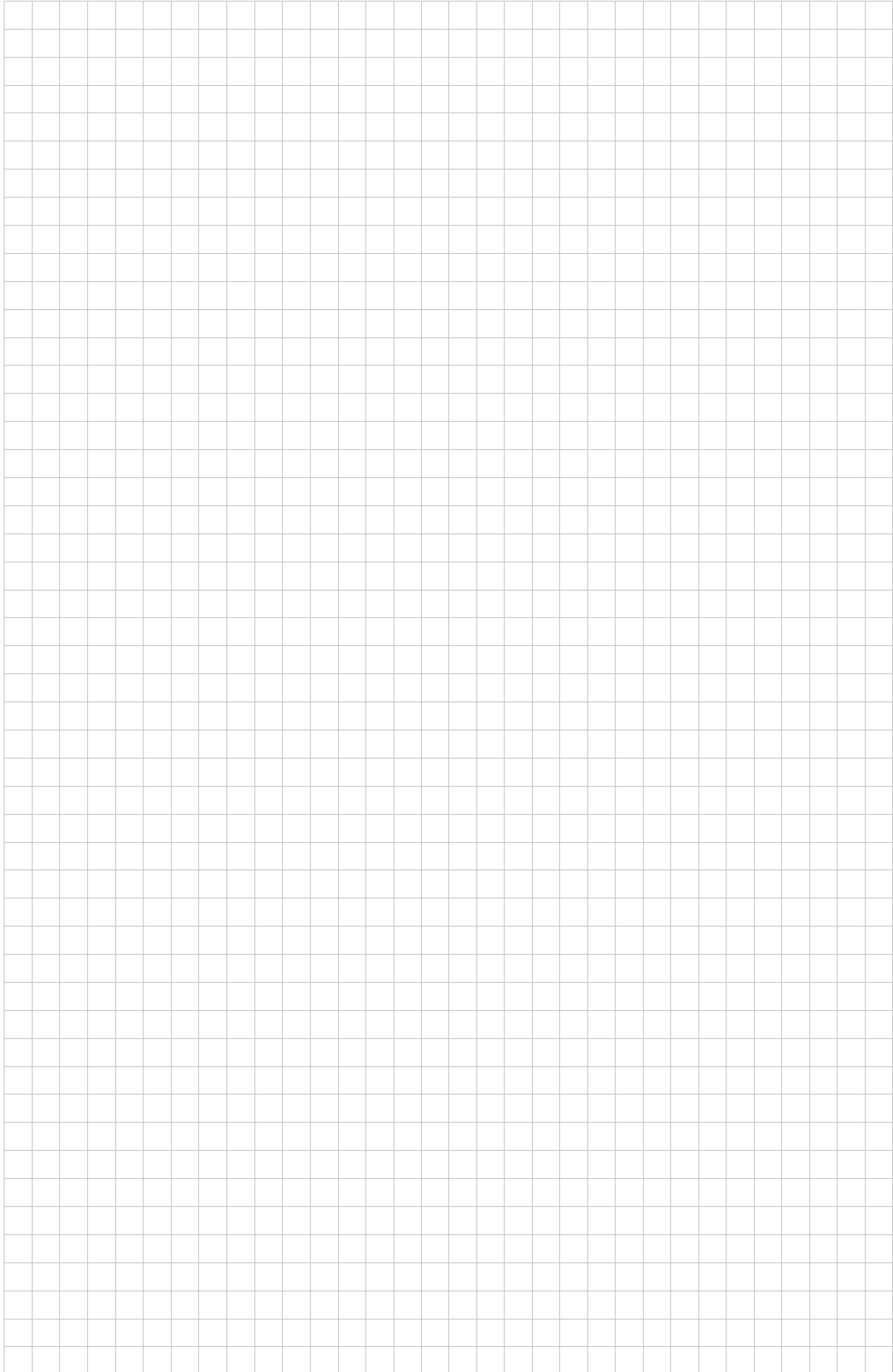
□C.Q.F.D.

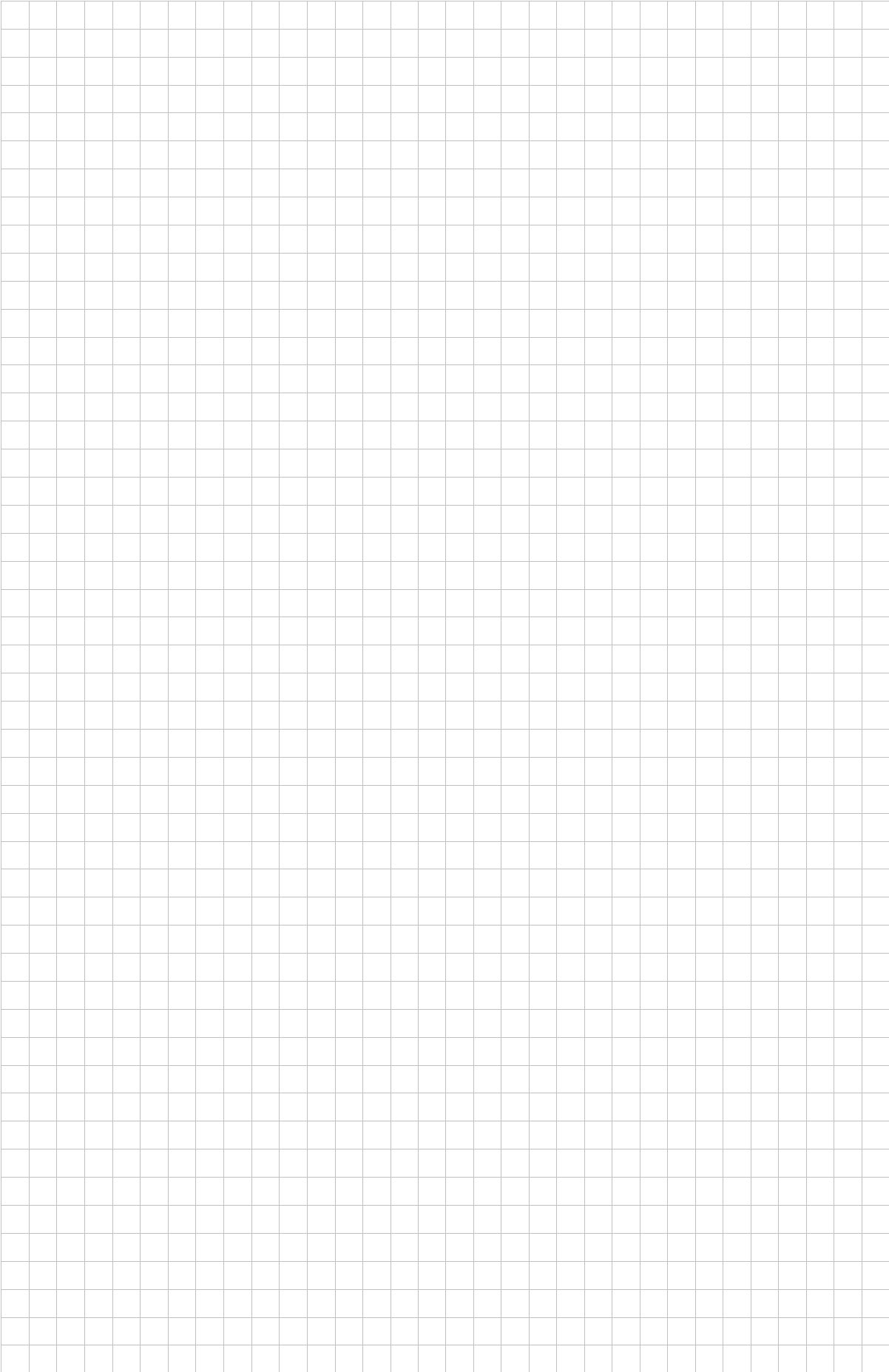
Remarque: La réciproque du théorème n'est pas vraie comme le montre l'exemple de l'anneau $\mathbb{R} \times \mathbb{R}$ où l'addition et la multiplication se font composante par composante. C'est un anneau de caractéristique nulle mais avec des diviseurs de zéro:

$$(1,0) \cdot (0,1) = 0 \quad (5.170)$$

Notes personnelles:









6. PROBABILITÉS

Le calcul des probabilités s'occupe des phénomènes aléatoires (dits plus esthétiquement: "processus stochastiques" lorsqu'ils sont dépendants du temps), c'est-à-dire de phénomènes qui ne mènent pas toujours à la même issue et qui peuvent être étudiés grâce aux nombres et à leurs conséquences et apparitions. Néanmoins, même si ces phénomènes ont des issues variées, dépendant du hasard, nous observons cependant une certaine régularité statistique.

Définitions: Il existe plusieurs manières de définir une probabilité. Principalement, nous parlons de:

D1. "**Probabilité expérimentale ou inductive**" qui est la probabilité déduite de toute la population concernée.

D2. "**Probabilité théorique ou déductive**" qui est la probabilité connue grâce à l'étude du phénomène sous-jacent sans expérimentation. Il s'agit donc d'une connaissance "a priori" par opposition à la définition précédente qui faisait plutôt référence à une notion de probabilité "à posteriori".

Comme il n'est pas toujours possible de déterminer des probabilités a priori, nous sommes souvent amenés à réaliser des expériences. Il faut donc pouvoir passer de la première à la deuxième solution. Ce passage est supposé possible en termes de limite (avec une population dont la taille tend vers la taille de la population réelle).

La modélisation formelle par le calcul des probabilités a été inventée par A.N. Kolmogorov dans un livre paru en 1933. Cette modélisation est faite à partir de l'espace de probabilités (U, A, P) que nous définirons plus loin et que nous pouvons relier à la théorie de la mesure (voir chapitre du même nom). Cependant, les probabilités ont été étudiées sur le point de vue scientifique par Fermat et Pascal au milieu du 17ème siècle.

Remarque: Si vous avez un professeur ou un formateur qui ose vous enseigner les statistiques et probabilités avec des exemples basés sur des jeux de hasard (cartes, dés, allumette, pile ou face, etc.) débarrassez-vous en ou dénoncez-le à qui de droit car cela signifierait qu'il n'a aucune expérience pratique du domaine et qu'il va vous enseigner n'importe quoi et n'importe comment (normalement les exemples devraient être basés sur l'industrie, l'économie ou la R&D, bref dans des domaines utilisés tous les jours par les entreprises mais surtout pas sur des jeux de hasard...!).

1. UNIVERS DES ÉVÉNEMENTS

Définitions:

D1. "**L'univers des événements**", ou "**univers des observables**", U est l'ensemble de toutes les issues (résultats) possibles, appelées "**événements élémentaires**", qui se présentent au cours d'une épreuve aléatoire déterminée. L'univers peut être fini (dénombrable) si les événements élémentaires sont en nombre fini ou continu (non dénombrable) s'ils sont infinis.

D2. Un "**événement**" quelconque A est un ensemble d'événements élémentaires et constitue une partie de l'univers des possibles U . Il est possible qu'un événement ne soit constitué que d'un seul événement élémentaire.

 Exemple:

Considérons l'univers de tous les groupes sanguins possible, alors l'événement A "l'individu est de rhésus positif" est représenté par:

$$A = \{A+, B+, AB+, O+\} \subset U \quad (6.1)$$

alors que l'événement B "l'individu est donneur universel" est représenté par:

$$B = \{O-\} \subset U \quad (6.2)$$

qui constitue donc un événement élémentaire.

D3. Soit U un univers et A un événement, nous disons que l'événement A "à lieu" (ou "se réalise") si lors du déroulement de l'épreuve se présente l'issue i ($i \in U$) et que ($i \in A$). Dans le cas contraire, nous disons que A "n'a pas lieu".

D4. Le sous-ensemble vide $\emptyset$ de U s'appelle "**événement impossible**". En effet, si lors de l'épreuve l'issue i se présente, nous avons toujours $i \notin \emptyset$ et l'événement $\emptyset$ n'a donc jamais lieu.

Si U est fini, ou infini dénombrable, tout sous-ensemble de U est un événement, ce n'est plus vrai si U est non dénombrable (nous verrons dans le chapitre de Statistiques pourquoi).

D5. L'ensemble U s'appelle aussi "**événement certain**". En effet, si lors de l'épreuve l'issue i se présente, nous avons toujours $i \in U$ (car U est l'univers des événements). L'événement U a donc toujours lieu.

D6. Soit A et B deux sous-ensembles de U . Nous savons que les événements $A \cup B$ et $A \cap B$ sont tous deux des sous-ensembles de U donc des événements qui sont respectivement des "**événements conjoints**" et des "**événements disjoints**".

Si deux événements A et B sont tels que:

$$A \cap B = \emptyset \quad (6.3)$$

les deux événements ne peuvent pas être réalisables pendant la même épreuve, nous disons alors qu'ils sont des "**événements incompatibles**".

Sinon, si:

$$A \cap B \neq \emptyset \quad (6.4)$$

les deux événements peuvent être réalisables dans la même épreuve (possibilité de voir un chat noir au moment où on passe sous une échelle par exemple), nous disons inversement qu'ils sont des "**événements indépendants**".

1.1. AXIOMATIQUE DE KOLMOGOROV

La probabilité d'un événement sera en quelque sorte le répondant de la notion de fréquence d'un phénomène aléatoire, en d'autres termes, à chaque événement nous allons attacher un nombre réel, appartenant à l'intervalle $[0,1]$, qui mesurera sa probabilité (chance) de réalisation. Les propriétés des fréquences que nous pouvons mettre en évidence lors d'épreuves diverses nous permettent de fixer les propriétés des probabilités.

Soit U un univers. Nous disons que nous définissons une probabilité sur les événements de U si à tout événement A de U nous associons un nombre ou une mesure $P(A)$, appelé "**probabilité a priori de l'événement A** " ou "**probabilité marginale de A** ".

A1. Pour tout événement A :

$$1 \geq P(A) \geq 0 \quad (6.5)$$

Ainsi, la probabilité de tout événement est un nombre réel compris entre 0 et 1 inclus (c'est du bon sens humain...).

A2. La probabilité de l'événement certain ou de l'ensemble (somme) des événements possibles est égale à 1:

$$P(U) = 1 \quad (6.6)$$

A3. Si $A \cap B = \emptyset$ sont deux événements incompatibles (disjoints), alors:

$$P(A \cup B) = P(A) + P(B) \quad (6.7)$$

la probabilité de la réunion ("ou") de deux événements incompatibles (ou mutuellement exclusifs) est donc égale à la somme de leurs probabilités (loi d'addition). Nous parlons alors de "**probabilité disjointe**".

Exemple:

Considérons que la probabilité dans une région donnée d'avoir sur 50 ans un tremblement de terre majeur est de 5% et que d'avoir sur la même période une inondation majeure est 10%. Nous souhaiterions savoir qu'elle est la probabilité qu'une centrale nucléaire rencontre au plus un des deux événements pendant cette même période s'ils sont bien incompatibles. Nous avons alors la probabilité qui est la somme des deux probabilités ce qui fait 15%...

Nous retrouverons un exemple de ce genre de probabilité disjointe dans le chapitre de Génie Industriel dans la méthode AMDEC (Analyse des Modes de Défaillance, de leurs Effets et de leurs Criticités) pour l'analyse de pannes des systèmes à structure complexe.

Autrement dit sous forme plus générale si $(A_i)_{i \in \mathbb{N}}$ est une suite d'événements disjoints deux à deux (A_i et A_j ne peuvent pas se produire en même temps si $i \neq j$) alors:

$$P\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} P(A_i) \quad (6.8)$$

Nous parlons alors de " **σ -additivité**" car si nous regardons de plus près les trois axiomes ci-dessus la mesure P forme une σ -algèbre (cf. chapitre de Théorie de la Mesure).

A l'opposé, si les événements ne sont pas incompatibles (ils peuvent se superposer ou autrement dit: ils ont une probabilité jointe), nous avons alors comme probabilité qu'au plus un des deux ait lieu:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (6.9)$$

Ceci signifie que la probabilité pour que l'un au plus des événements A ou B se réalise est égale à la somme des probabilités pour que se réalise A ou pour que se réalise B , moins la probabilité pour que A et B se réalisent simultanément (nous démontrerons plus loin que cela est simplement équivalent à la probabilité que les deux n'aient pas lieu en même temps!).

Exemple:

Considérons que la probabilité dans une région donnée d'avoir sur 50 ans un tremblement de terre majeur est de 5% et que d'avoir sur la même période une inondation majeure est 10% et que ces deux événements ne sont incompatibles... Nous souhaiterions savoir qu'elle est la probabilité qu'une centrale nucléaire rencontre tout au plus un des deux événements pendant cette même période. Nous avons alors la probabilité qui se calcule à partir de la relation précédente et qui donne alors 14.5%...

Et donc s'ils étaient incompatibles nous aurions $A \cap B = \emptyset$ et nous retrouverions alors bien la probabilité disjointe:

$$P(A \cup B) = P(A) + P(B) \quad (6.10)$$

Une conséquence immédiate des axiomes (A2) et (A3) est la relation entre les probabilités d'un événement A et son complémentaire, noté $\bar{A}$ (ou plus rarement conformément à la notation utilisée dans le chapitre de Théorie De La Démonstration le complémentaire peut être noté $\neg A$):

$$P(\bar{A}) = 1 - P(A) \quad (6.11)$$

Soit U un univers comportant un nombre fini n d'issues possibles:

$$U = \{i_1, i_2, \dots, i_n\} \quad (6.12)$$

où les événements:

$$I_1 = \{i_1\}, I_2 = \{i_2\}, \dots, I_n = \{i_n\} \quad (6.13)$$

sont appelés "**événements élémentaires**". Lorsque ces événements ont même probabilité, nous disons qu'ils sont "**équiprobables**". Dans ce cas, il est très facile de calculer leur probabilité. En effet, ces événements étant par définition incompatibles entre eux à ce niveau de notre discours, nous avons en vertu de l'axiome 3 des probabilités:

$$P(I_1 \cup I_2 \cup \dots \cup I_n) = P(I_1) + P(I_2) + \dots + P(I_{n-1}) + P(I_n) \quad (6.14)$$

mais puisque:

$$P(I_1 \cup I_2 \cup \dots \cup I_n) = P(U) = 1 \quad (6.15)$$

et que les probabilités du membre de droite sont par hypothèse équiprobables, nous avons:

$$P(I_1) = P(I_2) = P(I_3) = \dots = P(I_n) = \frac{1}{n} \quad (6.16)$$

Définition: Si A et B ne sont pas incompatibles mais qu'ils sont indépendants, nous savons que par leur compatibilité $A \cap B \neq \emptyset$, alors (très important en statistiques!):

$$P(A \cap B) = P(A) \cdot P(B) \quad (6.17)$$

la probabilité de l'intersection ("et") de deux événements indépendants est égale au produit de leurs probabilités (loi de multiplication). Nous parlons alors de "**probabilité conjointe**" (c'est le cas le plus fréquent) ou simplement de "**probabilité jointe**".

Exemple:

Considérons que la probabilité dans une région donnée d'avoir sur 50 ans un tremblement de terre majeur est de 5% et que d'avoir sur la même période une inondation majeure est 10%. De plus supposons que ces 2 événements ne soient pas incompatibles (en d'autres termes ils sont compatibles). Nous allons nous intéresser à leur indépendance. Ainsi, nous souhaiterions savoir qu'elle est la probabilité qu'une centrale nucléaire rencontre les deux événements en même temps, à quel que moment que ce soit, pendant cette même période. Nous avons alors la probabilité qui se calcule à partir de la relation précédente et qui donne alors 0.05%...

Autrement dit sous forme plus générale, les événements $A_1, \dots, A_n$ sont indépendants si la probabilité de l'intersection est le produit des probabilités:

$$P\left(\bigcap_{i \in \mathbb{N}} A_i\right) = \prod_{i \in \mathbb{N}} P(A_i) \quad (6.18)$$

Remarque: Attention donc à ne pas confondre "indépendants" et "incompatibles"!

Donc pour résumer jusqu'ici nous avons donc:

Type	Expression
2 événements incompatibles (disjoints)	$P(A \cup B) = P(A) + P(B)$
2 événements incompatibles (jointes)	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
2 événement non incompatibles mais indépendants	$P(A \cap B) = P(A) \cdot P(B)$

Tableau: 28.1 - Cas classiques de probabilités

Grâce à la définition précédente, nous pouvons démontrer que la probabilité pour que soit A ou soit B ait lieu (donc au moins un des deux mais pas les deux en même temps), est simplement égale à... la probabilité que les deux n'aient pas lieu en même temps:

$$\begin{aligned}
 P(A \cup B) &= P(A) + P(B) - P(A \cap B) = \\
 P(A) + P(B) - P(A)P(B) &= 1 - P(\bar{A})P(\bar{B}) = 1 - (1 - P(A))(1 - P(B))
 \end{aligned}
 \tag{6.19}$$

Nous pouvons aussi à l'aide de cette dernière définition déterminer la probabilité qu'un seul des deux événements ait lieu:

$$\begin{aligned}
 P(A \oplus B) &= P(A)P(\bar{B}) + P(B)P(\bar{A}) = P(A)(1 - P(B)) + P(B)(1 - P(A)) \\
 &= P(A) + P(B) - 2P(A)P(B) = P(A) + P(B) - 2P(A \cap B)
 \end{aligned}
 \tag{6.20}$$

Exemple:

Considérons que la probabilité dans une région donnée d'avoir sur 50 ans un tremblement de terre majeur est de 5% et que d'avoir sur la même période une inondation majeure est 10%. Nous souhaiterions savoir qu'elle est la probabilité qu'une centrale nucléaire rencontre exactement un des deux événements pendant la même période en considérant qu'ils ne peuvent avoir lieu en même temps. Nous avons alors la probabilité qui se calcule à partir de la relation précédente et qui donne alors 14%...

Il y a un domaine courant dans l'industrie dans lequel sont appliquées fréquemment les quatre relations suivantes (en anglais):

$$\begin{aligned}
 \text{AND: } &P(A) \cdot P(B) \\
 \text{OR COMPATIBLE: } &P(A) + P(B) - P(A \cap B) \\
 \text{OR INCOMPATIBLE: } &P(A) + P(B) \\
 \text{XOR : } &P(A) + P(B) - 2P(A \cap B)
 \end{aligned}
 \tag{6.21}$$

Il s'agit de "**l'analyse par arbres d'erreurs**" ou "**analyse par arbres probabilistes**" qui est utilisée pour analyser les raisons possibles de défaillance d'un système quel qu'il soit (industriel, administratif ou autre).

Pour clore cette partie du chapitre considérons la figure suivante qui montre les diagrammes de Venn (cf. chapitre de Théorie Des Ensembles) pour les 16 événements (y compris l'événement impossible) qui peuvent être décrits en termes de deux événements donnés A et B . Dans chaque cas, l'événement est représenté par la zone rouge:

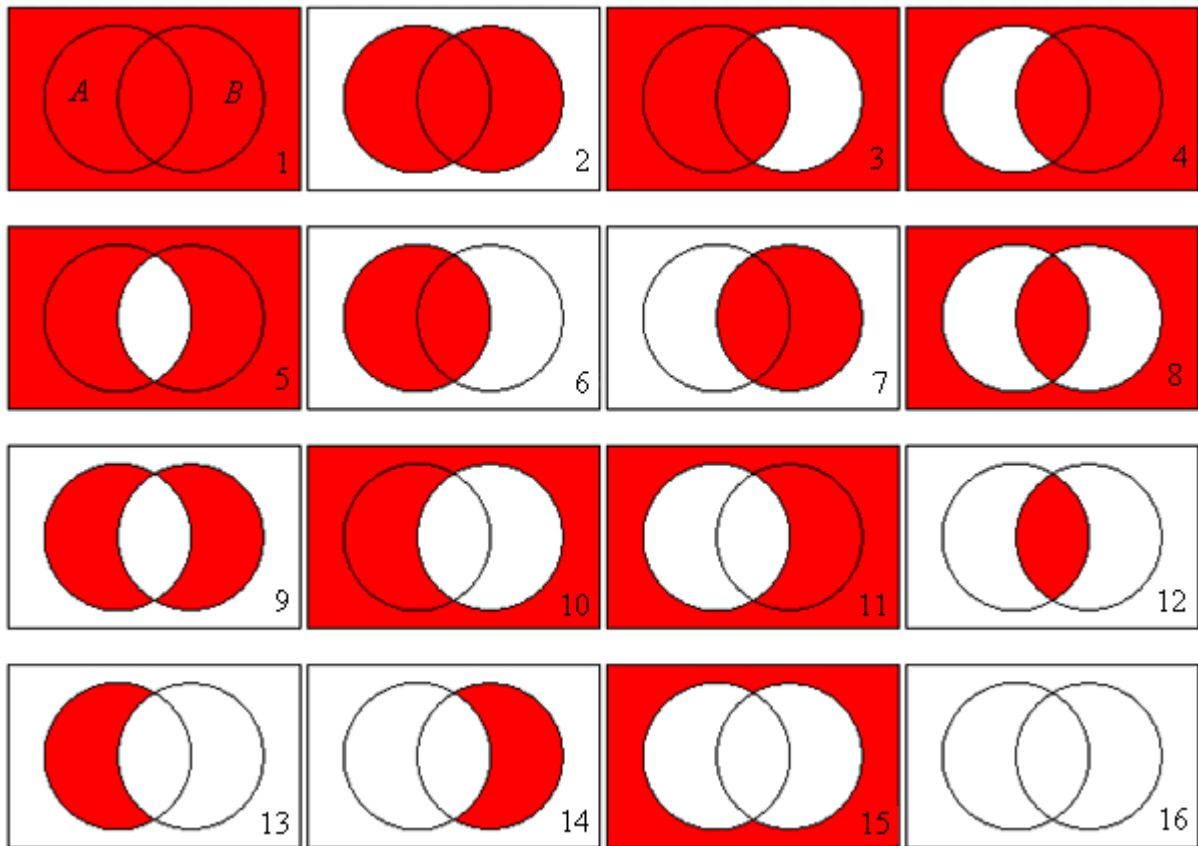


Figure: 6.1 - Diagrammes de Venn possibles pour deux événements

Considérons la situation où A représente un tremblement de terre et B représente une inondation majeure et U l'univers de tous les événements dramatiques pour une centrale nucléaire. Nous considérons que les deux événements sont indépendants. Ensuite, chacune des 16 combinaisons d'événements peuvent être décrites comme suit, soit mathématiquement ou verbalement.

1. Un tremblement de terre peut se produire ou une inondation ou rien ou l'ensemble à la fois ou tout autre événement (bref n'importe quel événement peut se produire).

$$P(U) = 1 = 100\% \quad (6.22)$$

2. $A \cup B$: Tout événement incluant un tremblement de terre, une inondation ou les deux en même temps peut se produire.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B) - P(A)P(B) \quad (6.23)$$

3. $A \cup B^c$: Tout événement incluant un tremblement de terre avec ou sans une inondation peut se produire à l'exception des événements incluant une inondation sans tremblement de terre.

$$P(A \cup B^c) = P(U) - P(B) + P(A \cap B) = 1 - P(B) + P(A)P(B) \quad (6.24)$$

4. $A^c \cup B$: Tout événement incluant une inondation avec ou sans tremblement de terre peut se produire à l'exception des événements incluant un tremblement de terre sans inondation.

$$P(A^c \cup B) = P(U) - P(A) + P(A \cap B) = 1 - P(A) + P(A)P(B) \quad (6.25)$$

5. $A^c \cup B^c$: Tout événement peut se produire sauf ceux incluant un tremblement de terre accompagné d'une inondation.

$$P(A^c \cup B^c) = P(U) - P(A \cap B) = 1 - P(A)P(B) \quad (6.26)$$

6. A : Tout événement avec un tremblement de terre peut se produire (cela inclut donc les événements associant un tremblement de terre et une inondation).

$$P(A) = P(A) \quad (6.27)$$

7. B : Tout événement avec une inondation peut se produire (cela inclut donc les événements associant une inondation et un tremblement de terre).

$$P(B) = P(B) \quad (6.28)$$

8. $(A \cap B) \cup (A^c \cap B^c)$: Tout événement peut se produire sauf ceux incluant un tremblement de terre sans inondation ou ceux incluant une inondation sans tremblement de terre.

$$\begin{aligned} P((A \cap B) \cup (A^c \cap B^c)) &= P(U) - P(A) - P(B) + 2P(A \cap B) \\ &= 1 - P(A) - P(B) + 2P(A)P(B) \end{aligned} \quad (6.29)$$

9. $(A \cap B^c) \cup (A^c \cap B)$: Tout événement incluant un tremblement de terre sans inondation ou une inondation sans tremblement de terre peut avoir lieu.

$$P((A \cap B^c) \cup (A^c \cap B)) = P(A) + P(B) - 2P(A \cap B) \quad (6.30)$$

10. B^c : Tout événement excepté ceux associés à une inondation peuvent avoir lieu.

$$P(B^c) = P(U) - P(B) = 1 - P(B) \quad (6.31)$$

11. A^c : Tout événement excepté ceux associés à un tremblement de terre peuvent avoir lieu.

$$P(A^c) = P(U) - P(A) = 1 - P(A) \quad (6.32)$$

12. $A \cap B$: Tout événement associant un tremblement de terre et une inondation peut avoir lieu.

$$P(A \cap B) = P(A)P(B) \quad (6.33)$$

13. $A \cap B^c$: Tout événement avec un tremblement de terre sans inondation peut avoir lieu.

$$P(A \cap B^c) = P(A) - P(A)P(B) \quad (6.34)$$

14. $A^c \cap B$: Tout événement avec une inondation sans tremblement de terre peut avoir lieu.

$$P(A^c \cap B) = P(B) - P(A)P(B) \quad (6.35)$$

15. $A^c \cap B^c$: Tout événement peut avoir lieu excepté ceux incluant un tremblement de terre et/ou une inondation.

$$P(A^c \cap B^c) = P(U) - P(A \cup B) = 1 - P(A) - P(B) + P(A)P(B) \quad (6.36)$$

16. $A \cap A^c$ ou $B \cap B^c$: Événement impossible.

$$P(A \cap A^c) = P(B \cap B^c) = P(\emptyset) = 0 \quad (6.37)$$

1.2. PROBABILITÉS CONDITIONNELLES

Que pouvons-nous déduire sur la probabilité d'un événement B sachant qu'un événement A est réalisé sachant qu'il existe un lien entre A et B ? En d'autres termes, s'il existe bien un lien entre A et B , la réalisation de A va modifier notre connaissance sur B et nous voulons savoir s'il est possible de définir la probabilité d'un événement conditionnellement (relativement) à un autre événement.

Ce type de probabilité est appelée "**probabilité conditionnelle**" ou "**probabilité à posteriori**" de B sachant A , et se note dans le cadre de l'étude des probabilités conditionnelles:

$$P(B/A) \quad (6.38)$$

et souvent dans la pratique pour éviter la confusion avec une possible division:

$$P(B|A) \quad (6.39)$$

et nous trouvons parfois chez les américains la notation:

$$P(B \wedge A) \quad (6.40)$$

Nous avons aussi le cas:

$$P(A|B) \quad (6.41)$$

qui est appelé "**fonction de vraisemblance de A** " ou encore "**probabilité a priori de A** " sachant B .

Historiquement, le premier mathématicien à avoir utilisé correctement la notion de probabilité conditionnelle fut Thomas Bayes (1702-1761). Aussi parlons-nous souvent de Bayes ou de bayésien dès que des probabilités conditionnelles sont en jeu: formule de Bayes, statistique bayésienne...

La notion de probabilité conditionnelle que nous allons introduire est beaucoup moins simple qu'elle ne paraît a priori et les problèmes de conditionnement sont une source inépuisable d'erreurs en tout genre (il existe de fameux paradoxes sur le sujet).

Commençons d'abord par un exemple simpliste: Supposons que nous ayons deux dés. Imaginons maintenant que nous ayons lancé seulement le premier dé. Nous voulons savoir quelle est la probabilité qu'en lançant le second dé, la somme des deux chiffres vaille une certaine valeur minimale. Ainsi, la probabilité d'obtenir cette valeur minimale fixée sachant la valeur du premier dé est totalement différente de la probabilité d'obtenir cette même valeur minimale en lançant les deux dés en même temps. Comment calculer cette nouvelle probabilité?

Formalisons la démarche:

Après le lancer du premier dé, nous avons:

$$A = \{\text{1e résultat du premier lancé est...}\} \quad (6.42)$$

Soit l'hypothèse que $B \subset A$, nous pressentons que $P(B / A)$ doit être proportionnel à $P(B)$, la constante de proportionnalité étant déterminée par la normalisation:

$$P(A / A) = 1 \quad (6.43)$$

Soit maintenant $B \subset A^c$ (B est inclus dans le complémentaire de A donc les événements sont incompatibles). Il est relativement intuitif.... que sous hypothèse précédente d'incompatibilité nous ayons la probabilité conditionnelle:

$$P(B / A) = 0 \quad (6.44)$$

Ceci nous mène aux définitions suivantes des probabilités à posteriori et respectivement à priori:

$$\boxed{P(B / A) = \frac{P(A \cap B)}{P(A)}} \quad \text{et} \quad \boxed{P(A / B) = \frac{P(A \cap B)}{P(B)}} \quad (6.45)$$

Ainsi, le fait de savoir que A est réalisé réduit l'ensemble des résultats possibles de U de B . A partir de là, seules les éventualités de $A \cap B$ ont une importance. La probabilité de A sachant B inversement (par symétrie) doit donc être proportionnelle à $P(A \cap B)$!

Le coefficient de proportionnalité qui est le dénominateur permet d'assurer l'événement certain. Effectivement, si les deux événements A et B sont indépendants (pensez à l'histoire du chat noir et de l'échelle par exemple), nous avons donc:

$$P(A \cap B) = P(A) \cdot P(B) \quad (6.46)$$

et nous voyons alors $P(B / A)$ qui vaut $P(B)$ et donc A n'apporte rien sur B et réciproquement!! Donc en d'autres termes, si A et B sont indépendants nous avons:

$$P(B/A) = P(B) \quad \text{et} \quad P(A/B) = P(A) \quad (6.47)$$

Une autre façon assez intuitive pour voir les choses est de se représenter la mesure de probabilité P comme une mesure d'aires de sous-ensembles de $\mathbb{R}^2$.

En effet, si A et B sont deux sous-ensembles de $\mathbb{R}^2$ d'aires respectives $P(A)$ et $P(B)$ alors à la question de savoir qu'elle est la probabilité qu'un point du plan appartienne à B sachant qu'il appartient à A il est assez évident de répondre que cette probabilité est donnée par:

$$\frac{\text{Surface}(A \cap B)}{\text{Surface}(A)} \quad (6.48)$$

Indiquons aussi que la définition des probabilités conditionnelles s'utilise souvent sous la forme suivante:

$$P(A \cap B) = P(A/B) \cdot P(B) = P(B/A)P(A) \quad (6.49)$$

appelée "**formule des probabilités composées**". Ainsi, la probabilité à posteriori de B sachant A peut donc aussi s'écrire sous la forme:

$$P(B/A) = \frac{P(A/B)P(B)}{P(A)} \quad (6.50)$$

Exemple:

Supposons une maladie comme la méningite. La probabilité de l'avoir sera notée $P(M) = 0.001$ (chiffre arbitraire pour l'exemple) et un signe de cette maladie comme le mal de tête sera noté $P(S) = 0.1$. Supposons connue la probabilité à posteriori d'avoir mal à la tête si nous avons une méningite:

$$P(S/M) = 0.9 \quad (6.51)$$

Le théorème de Bayes donne alors la probabilité a priori d'avoir une méningite si nous avons mal à la tête!:

$$P(M/S) = \frac{P(S/M)P(M)}{P(S)} = 0.009 \quad (6.52)$$

Pour en revenir à la théorie, notons que nous avons aussi:

$$\begin{aligned} P(A) &= P((B_1 \cup B_2 \cup \dots \cup B_n) \cap A) = P((B_1 \cap A) \cup (B_2 \cap A) \cup \dots \cup (B_n \cap A)) \\ &= P(B_1 \cap A) + P(B_2 \cap A) + \dots + P(B_n \cap A) \\ &= P(A/B_1)P(B_1) + P(A/B_2)P(B_2) + \dots + P(A/B_n)P(B_n) \\ &= \sum_{i=1}^n P(A/B_i)P(B_i) \end{aligned} \quad (6.53)$$

Nous pouvons donc connaître la probabilité de l'événement A connaissant les probabilités $P(B_i)$ élémentaires de ses causes et les probabilités conditionnelles de A pour chaque B_i :

$$P(A) = \sum_{i=1}^n P(A / B_i) P(B_i) \quad (6.54)$$

qui est appelée la "**formule des probabilités totales**" ou "**théorème des probabilités totales**". Mais aussi, pour tout j , nous avons le corollaire suivant en utilisant les résultats précédents qui nous donne suite à un événement A , la probabilité que ce soit la cause B_i qui l'ai produit:

$$P(B_j / A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(A / B_j) P(B_j)}{\sum_i P(A / B_i) P(B_i)} \quad (6.55)$$

qui est la forme générale de la "**formule de Bayes**" ou "**théorème de Bayes**" que nous utiliserons un tout petit peu en Mécanique Statistique et dans le cadre de l'étude de la théorie des files d'attente (cf. chapitre de Techniques De Gestion). Il faut savoir que les implications de ce théorème sont cependant considérables dans le quotidien, dans la médecine, dans l'industrie et dans le domaine du Data Mining informatique.

Nous retrouvons souvent dans la littérature de nombreux exemples d'applications de la relation précédente avec uniquement deux issues possibles B relativement à l'événement A . Dès lors nous avons la formule de Bayes écrite sous la forme suivante pour chacune des issues:

$$\begin{aligned} P(B_1 / A) &= \frac{P(A / B_1) P(B_1)}{P(A / B_1) P(B_1) + P(A / B_2) P(B_2)} = \frac{P(A / B_1) P(B_1)}{P(A / B_1) P(B_1) + P(A / \overline{B_1}) P(\overline{B_1})} \\ P(B_2 / A) &= \frac{P(A / B_2) P(B_2)}{P(A / B_1) P(B_1) + P(A / B_2) P(B_2)} = \frac{P(A / B_2) P(B_2)}{P(A / \overline{B_2}) P(\overline{B_2}) + P(A / B_2) P(B_2)} \end{aligned} \quad (6.56)$$

et remarquons que dans ce cas particulier (des issues binaires):

$$\begin{aligned} P(B_1 / A) + P(B_2 / A) &= P(B_1 / A) + P(\overline{B_1} / A) \\ &= \frac{P(A / B_1) P(B_1)}{P(A / B_1) P(B_1) + P(A / \overline{B_1}) P(\overline{B_1})} + \frac{P(A / \overline{B_1}) P(\overline{B_1})}{P(A / \overline{B_1}) P(\overline{B_1}) + P(A / B_1) P(B_1)} \\ &= \frac{P(A / B_1) P(B_1)}{P(A / B_1) P(B_1) + P(A / \overline{B_1}) P(\overline{B_1})} + \frac{P(A / \overline{B_1}) P(\overline{B_1})}{P(A / B_1) P(B_1) + P(A / \overline{B_1}) P(\overline{B_1})} \\ &= \frac{P(A / B_1) P(B_1) + P(A / \overline{B_1}) P(\overline{B_1})}{P(A / B_1) P(B_1) + P(A / \overline{B_1}) P(\overline{B_1})} = 1 \end{aligned} \quad (6.57)$$

ce qui est un résultat intuitif.

Pour les événements binaires, nous avons aussi (en revenant au théorème des probabilités totales vu plus haut):

$$\begin{aligned}
 P(A) &= \sum_{i=1}^2 P(A/B_i) P(B_i) = P(A/B_1) P(B_1) + P(A/B_2) P(B_2) \\
 &= P(A/B_1) P(B_1) + P(A/\overline{B_1}) P(\overline{B_1})
 \end{aligned} \tag{6.58}$$

Exemples:

E1. Une maladie affecte 10 personnes sur 10'000 (soit $0.1\% = 0.001$). Un test a été développé qui a 5% de faux positifs (personnes non atteintes pour lequel le test dit qu'ils sont atteints) mais qui détecte toujours cette maladie si une personne est atteinte. Quelle est la probabilité qu'une personne aléatoire pour laquelle le test donne un résultat positif a vraiment cette maladie?

Il y a donc sur 10'000 personnes, 500 qui seront des faux positifs et nous savons a posteriori que 10 personnes ont réellement la maladie. Alors la probabilité que quelqu'un qui a un résultat de test positif soit vraiment malade est:

$$\begin{aligned}
 P(M) &= 0.001 \\
 P(\overline{M}) &= 0.999 \\
 P(T/M) &= 1 \\
 P(T/\overline{M}) &= 0.05 \\
 P(T) &= P(T/M) P(M) + P(T/\overline{M}) P(\overline{M}) = 0.05095 \\
 P(M/T) &= \frac{P(T/M) P(M)}{P(T)} = \frac{1 \cdot 0.001}{0.05095} = 0.19627
 \end{aligned} \tag{6.59}$$

Ce résultat est souvent contre-intuitif et même scandaleux. Il met aussi en évidence pourquoi les tests de diagnostics doivent être extrêmement fiables!

E2. Deux machines M_1 et M_2 produisent respectivement 100 et 200 pièces. M_1 produit 5% de pièces défectueuses et M_2 en produit 6% (probabilités a posteriori). Quelle est la probabilité a priori pour qu'un objet défectueux ait été fabriqué par la machine M_1 ?

L'événement constaté A est donc la présence d'une pièce défectueuse et la probabilité recherchée est la probabilité a priori que celle-ci provienne de la machine M_1 .

Nous avons alors:

$$\begin{aligned}
 P(M_j/A) &= \frac{P(A/M_j) P(M_j)}{\sum_i P(A/M_i) P(M_i)} = \frac{P(A/M_1) P(M_1)}{P(A/M_1) P(M_1) + P(A/M_2) P(M_2)} \\
 &= \frac{\frac{5}{100} \cdot \frac{100}{300}}{\frac{5}{100} \cdot \frac{100}{300} + \frac{6}{100} \cdot \frac{200}{300}} = \frac{5\% \cdot \frac{1}{3}}{5\% \cdot \frac{1}{3} + 6\% \cdot \frac{2}{3}} \cong 0.29
 \end{aligned} \tag{6.60}$$

E3. D'un lot de 10 pièces dont le 30% est défectueux, nous prélevons sans remise un échantillon de taille 3. Quelle est la probabilité que la seconde pièce soit bonne (quelle que soit la première)?

Nous avons:

$$\begin{aligned}
 P(A) &= \sum_{i=1}^n P(A/B_i) P(B_i) = \sum_{i=1}^2 P(A/B_i) P(B_i) \\
 &= P(A/B_1) P(B_1) + P(A/B_2) P(B_2) = \frac{7}{9} \frac{3}{10} + \frac{6}{9} \frac{7}{10} = 70\%
 \end{aligned} \tag{6.61}$$

où $P(A/B_1)$ est la probabilité que la deuxième soit bonne sachant que la première est mauvaise et $P(A/B_2)$ est la probabilité que la deuxième soit bonne sachant que la première est bonne. $P(B_1)$ est donc la probabilité que la première soit mauvaise, $P(B_2)$ la probabilité que la première soit bonne.

E4. Terminons avec un exemple important dans les entreprises où les employés doivent plusieurs fois dans leur carrière passer des examens sous forme de questionnaire à choix multiples (Q.C.M.). Si un employé répond à une question de deux choses l'une: soit il connaît la réponse, soit il la devine. Soit p la probabilité que l'employé connaisse la réponse et donc $1-p$ celle qu'il la devine. Nous admettons que l'employé qui devine répondra correctement avec une probabilité $1/m$ où m est le nombre de réponses proposées. Quelle est alors la probabilité a priori qu'un employé connaisse (réellement) la réponse à une question à 5 choix s'il y a répondu correctement?

Soient B et A respectivement les événements "l'employé connaît la réponse" et "l'employé répond correctement à la question". Alors la probabilité à priori qu'un employé connaisse (réellement) la réponse à une question qu'il a répondu correctement est:

$$P(B/A) = \frac{P(A/B)P(B)}{P(A/B)P(B) + P(A/\bar{B})P(\bar{B})} = \frac{1 \cdot p}{1 \cdot p + \frac{1}{m}(1-p)} = \frac{5}{6} \cong 83\% \tag{6.62}$$

L'analyse bayésienne fournit donc un outil puissant de formalisation du raisonnement dans l'incertain et les exemples que nous avons montrés illustrent surtout à quel point cet outil est délicat à employer.

1.2.1. ESPÉRANCE CONDITIONNELLE

Maintenant, passons à la version continue de la probabilité conditionnelle en abordant le sujet directement avec un exemple particulier (la théorie avec le cas général étant indigeste) infiniment important dans le domaine de statistiques sociales et de la finance quantitative. Cependant, ce choix (de l'étude d'un cas particulier) implique que le lecteur ait lu au préalable le chapitre de Statistiques pour y étudier les fonctions de distributions continues et plus particulièrement celle de la loi de Pareto.

Donc voilà le scénario: Souvent, en sciences sociales ou en économie, nous trouvons dans la littérature spécialisée traitant des lois de Pareto des affirmations du type suivant (mais quasiment jamais avec une démonstration détaillée): quel que soit votre revenu, le revenu moyen de ceux qui ont un revenu supérieur au vôtre est dans un rapport constant, supérieur à 1, à votre revenu si celui-ci suit une variable aléatoire de type Pareto. Nous disons alors que la loi est isomorphe à toute partie tronquée elle-même.

Voyons de quoi il s'agit exactement:

Soit X une variable aléatoire égale au revenu et suivant une loi de Pareto de densité (cf. chapitre de Statistiques):

$$f(x) = k \frac{x_m^k}{x^{k+1}} \quad (6.63)$$

avec $k > 1, x_m > 0, x \geq x_m$ et qui a pour fonction de répartition (voir aussi le chapitre de Statistique pour la démonstration détaillée):

$$P(X \leq x) = 1 - \left(\frac{x_m}{x} \right)^k \quad (6.64)$$

La phrase commence par "quel que soit votre revenu...", choisissons donc un revenu quelconque $x_0 \geq x_m$.

À présent nous devons calculer "le revenu moyen de ceux qui ont un revenu supérieur à x_0 ". Il s'agit donc de calculer l'espérance (le revenu moyen) d'une nouvelle variable aléatoire Y qui est égale à X mais restreinte à la population des personnes ayant un revenu supérieur à x_0 :

$$Y = X |_{\{X \geq x_0\}} \quad (6.65)$$

La fonction de répartition de Y est donnée par:

$$P(Y \leq x) = P(X \leq x | X \geq x_0) \quad (6.66)$$

Cette expression est naturellement nulle si $x \leq x_0$.

Bon, jusqu'à maintenant nous n'avons fait que du vocabulaire. D'abord rappelons la relation de probabilité conditionnelle suivante vue plus haut:

$$P(B | A) = \frac{P(A \cap B)}{P(A)} \quad (6.67)$$

pour $x \geq x_0$ nous avons la loi conditionnelle à priori:

$$P(X \leq x | X \geq x_0) = \frac{P(x_0 \leq X \leq x)}{P(X \geq x_0)} \quad (6.68)$$

Avant d'aller plus loin, il faut être conscient que le numérateur et dénominateur sont indépendants mais que l'ensemble doit être toutefois considéré comme la réalisation d'une seule et unique variable aléatoire que nous noterons Y . Par ailleurs, seulement le numérateur est dépendant d'une variable. Le dénominateur peut lui être considéré comme une constante de normalisation.

Nous voyons donc que la densité de Y est donnée par la fonction:

$$f_Y(y) = \begin{cases} 0 & y < x_0 \\ \frac{f(y)}{P(X \geq x_0)} & y \geq x_0 \end{cases} \quad (6.69)$$

À présent nous pouvons calculer l'espérance de Y :

$$\begin{aligned}
 E(Y) &= \int y f_Y(y) dy = \int_{x_0}^{+\infty} y \frac{f(y)}{P(X \geq x_0)} dy = \frac{1}{P(X \geq x_0)} \int_{x_0}^{+\infty} y f(y) dy \\
 &= \frac{1}{P(X \geq x_0)} \int_{x_0}^{+\infty} y k \frac{x_m^k}{y^{k+1}} dy = \frac{1}{P(X \geq x_0)} \int_{x_0}^{+\infty} k \frac{x_m^k}{y^k} dy = \frac{k x_m^k}{P(X \geq x_0)} \int_{x_0}^{+\infty} \frac{1}{y^k} dy \quad (6.70) \\
 &= \frac{k x_m^k}{P(X \geq x_0)} \frac{1}{(k-1) y^{k-1}} \Big|_{x_0}^{+\infty} = \frac{1}{P(X \geq x_0)} \frac{k x_m^k}{(k-1) x_0^{k-1}}
 \end{aligned}$$

Sachant que:

$$P(X \geq x_0) = \int_{x_0}^{+\infty} k \frac{x_m^k}{x^{k+1}} dx = \frac{x_m^k}{x_0^k} = \left(\frac{x_m}{x_0} \right)^k \quad (6.71)$$

Nous avons au final:

$$E(Y) = \frac{k}{(k-1)} x_0 \quad (6.72)$$

$E(Y)$ représente donc le revenu moyen de ceux qui ont un revenu supérieur à x_0 et comme on peut le constater de l'égalité ci-dessus il est bien dans un rapport constant, supérieur à 1, à votre revenu x_0 .

Nous pouvons vérifier ce résultat en faisant une simulation de Monte-Carlo dans un tableur (c'est intéressant de le mentionner pour généraliser à des cas non calculable à la main). Il suffit effectivement d'y simuler l'inverse de la fonction de répartition:

$$X = \left(\frac{x_m^k}{1 - P(X \leq x)} \right)^{1/k} \quad (6.73)$$

soit dans MS Excel 11.8346 (version anglaise):

$$= (\$B\$7^{\wedge} \$B\$6 / (1 - \text{RANDBETWEEN}(1; 10000) / 10000))^{\wedge} (1 / \$B\$6)$$

et ensuite de prendre la moyenne des valeurs obtenues supérieurs ou égales à un X donné (ce qui correspondra à x_0) et vérifier que nous obtenons bien le résultat démontré précédemment!

Évidemment, nous pourrions aussi calculer la variance conditionnelle (in extenso l'écart-type conditionnel). Cela viendra peut-être un jour...

1.2.2. RÉSEAUX BAYÉSIENS

Les réseaux bayésiens sont simplement une représentation graphique d'un problème de probabilités conditionnelles qui permet de mieux visualiser l'interaction entre les différentes variables lorsque que celles-ci commencent à être en grande nombre.

C'est une technique de plus en plus utilisée dans le décisionnel assisté par logiciel (Data Mining), l'intelligence artificielle (AI) et également dans l'analyse et la gestion du risque (norme ISO 31010).

Les réseaux bayésiens sont par définition sont des graphes orientés acycliques (cf. chapitre de Théorie Des Graphes), afin qu'un événement ne puisse pas (même indirectement) influencer sa propre probabilité, avec description quantitative des dépendances entre événements.

Ces graphes servent à la fois de modèles de représentation des connaissances et de machines à calculer des probabilités conditionnelles. Ils sont surtout utilisés pour le diagnostic (médical et industriel), l'analyse de risques (diagnostics de pannes, anomalies ou accidents), la détection des spams, l'analyse de texte de voix et d'images, l'analyse d'opinions, la détection de fraudeurs ou de mauvais payeurs ainsi que dans le data mining (EGC: Extraction et Gestion de la Connaissance) en général.

Remarque: De nombreux systèmes et logiciels permettent de construire et d'analyser des réseaux bayésiens sur la base de dessins ou de d'informations existantes dans des bases de données. Solutions payantes: SQL Server, Oracle, Hugin. Solutions gratuits (à ce jour): Tanagra, Microsoft Belief Network MSBNX 1.4.2, RapidMiner. Personnellement je préfère la simplicité du petit logiciel MSBNX de Microsoft. Pour information, en 10 ans d'expérience professionnelle en tant que consultant je n'ai rencontré à ce jour qu'une seule entreprise parmi plus de 800 multinationales dans mon portefeuille qui utilisait les réseaux de bayésiens... (dans le domaine des transports).

Utiliser un réseau bayésien s'appelle faire de "**l'inférence bayésienne**". En fonction des informations observées, nous calculons la probabilité des données possibles connues mais non observées.

Pour un domaine donné (par exemple médical), nous décrivons les relations causales entre variables d'intérêt par un graphe (plus besoin de préciser qu'il est acyclique). Dans ce graphe, les relations de cause à effet entre les variables ne sont pas déterministes, mais probabilisées. Ainsi, l'observation d'une cause ou de plusieurs causes n'entraîne pas systématiquement l'effet ou les effets qui en dépendent, mais modifie seulement la probabilité de les observer.

L'intérêt particulier des réseaux bayésiens est de tenir compte simultanément de connaissances a priori d'experts (dans le graphe) et de l'expérience contenue dans les données.

Exemple de 5 variables avec relations (graphe orienté acyclique) et numérotation des états/variables (en anglais: "states"):

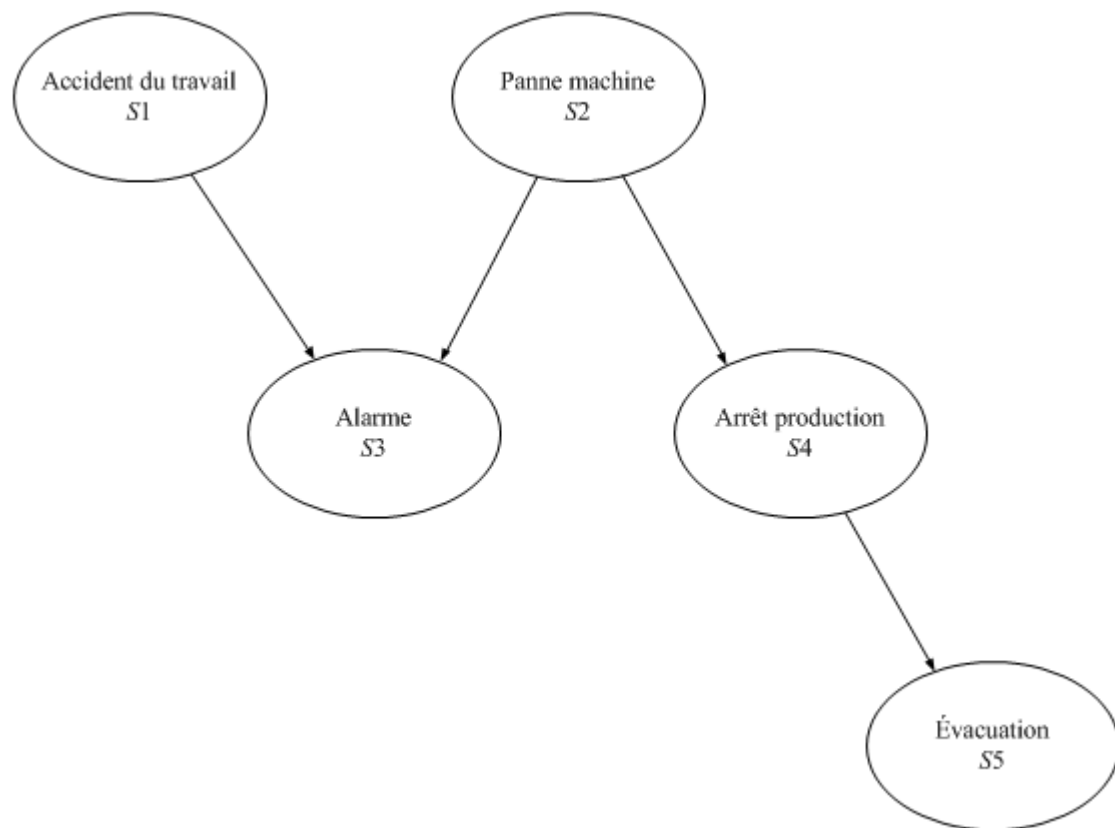


Figure: 6.2 - Exemple de réseau bayésien (acyclique orienté) à 5 états

Évidemment, la construction du graphe causal se fonde principalement sur le retour d'expériences (REX) et résulte parfois de normes ou de rapports de comités d'experts. Dans l'informatique, le graphe causal évolue automatiquement en fonction des bases de données (pensez à la librairie Amazon qui cible les publicités en fonction de vos achats passés en temps réel ou au service Genius de Apple). Cependant nous pourrions rarement penser à toutes les possibilités et il y aura aussi parfois des états cachés entre deux états qui auront été oubliés mais qui auraient permis de mieux modéliser la situation.

Imaginons dans l'exemple ci-dessus qu'à l'aide d'une base de données d'une entreprise, nous sachions que sur 100'000 jours hommes, nous avons eu dans cette entreprise 1'000 accidents du travail (soit 1% du total) et 100 pannes machines (soit 0.01% du total). Nous représentons cela alors sous la forme traditionnelle suivante:

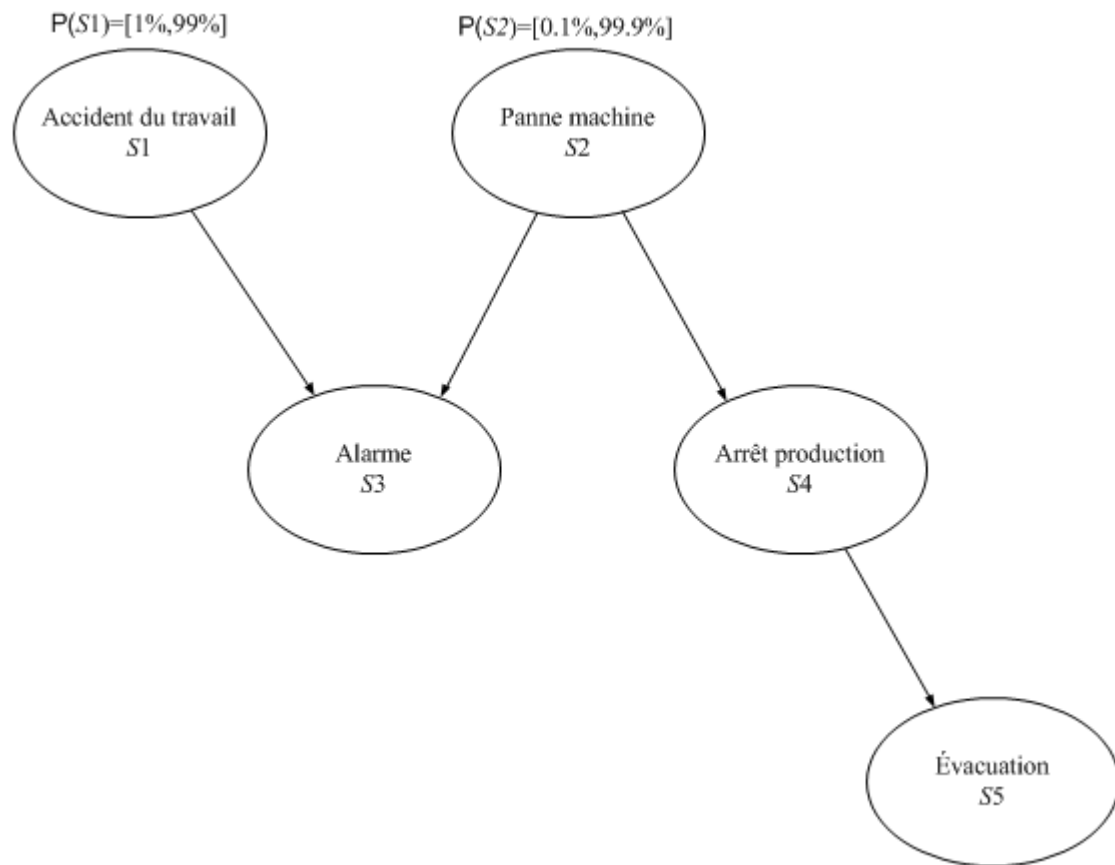


Figure: 6.3 - Réseau bayésien acyclique orienté avec probabilités de départ

où nous avons le sous-ensemble $S2, S4, S5$ qui constitue ce que les spécialistes appellent une "**connexion série ou linéaire**", le triplet $S3, S2, S4$ constitue une "**relation divergente**" (si les flèches pour ce triplet étaient inversées, nous aurions une "**relation convergente**").

Avant d'aller plus loin avec notre exemple faisons quelques constats par rapport à ces trois types de relations:

Pour toute clarté, distinguons d'abord "l'indépendance conditionnelle" de la "dépendance conditionnelle".

Nous disons que des événements A et C sont "**indépendants conditionnellement**" si étant donné un événement B l'égalité suivante est vérifiée:

$$P(A/B) = P(A/B, C) \quad (6.74)$$

Donc le qualificatif "conditionnellement" implique la présence de B et le fait que C n'influence pas la probabilité de l'événement A .

Concernant la "**dépendance conditionnelle**", nous pouvons cette fois distinguer 3 types de relations.

1. La dépendance conditionnelle du type suivant est appelée "connexion série ou linéaire" (déjà mentionnée plus haut):

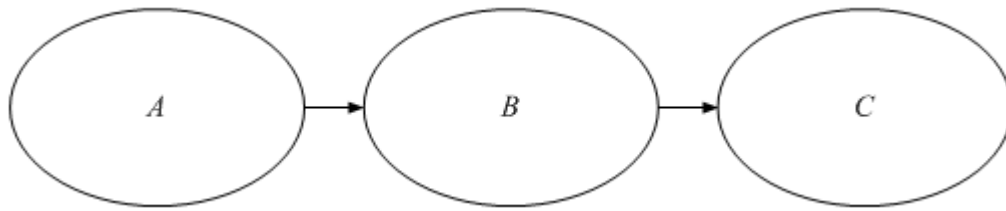


Figure: 6.4 - Réseau bayésien en connexion série/linéaire

où A , B et C sont dépendants (dans cet exemple particulier il y a 3 nœuds dépendants A , B et C mais d'une manière générale cette dépendance concernerait tous les nœuds s'il y en avait plus de 3).

En outre A et C sont dépendants mais conditionnellement à B . Mais si la variable B est connue, A n'apporte plus aucune information utile sur C (le cheminement de l'incertitude est en quelque sorte rompu) et dès lors A et C deviennent indépendants conditionnellement. Nous avons la probabilité conditionnelle qui se simplifierait donc sous la forme suivante:

$$P(C | B, A) = P(C | B) \quad (6.75)$$

2. La dépendance conditionnelle du type suivant est appelée "connexion divergente" (aussi déjà mentionnée plus haut):

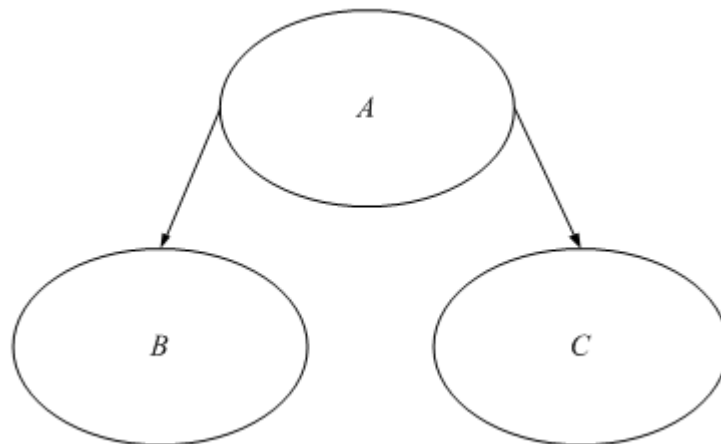


Figure: 6.5 - Réseau bayésien divergent

où l'ensemble des nœuds sont dépendants.

En outre B et C sont dépendants conditionnellement à A . Mais si A est connue, B n'apporte plus aucune information sur C (à nouveau le cheminement de l'incertitude est en quelque sorte rompu) et dès lors B et C deviennent indépendants. Nous avons donc par exemple si A est connue:

$$P(C | A, B) = P(C | A) \quad (6.76)$$

3. La dépendance conditionnelle du type suivant est appelée "connexion convergente" ou "V-Structure" (aussi déjà mentionnée plus haut):

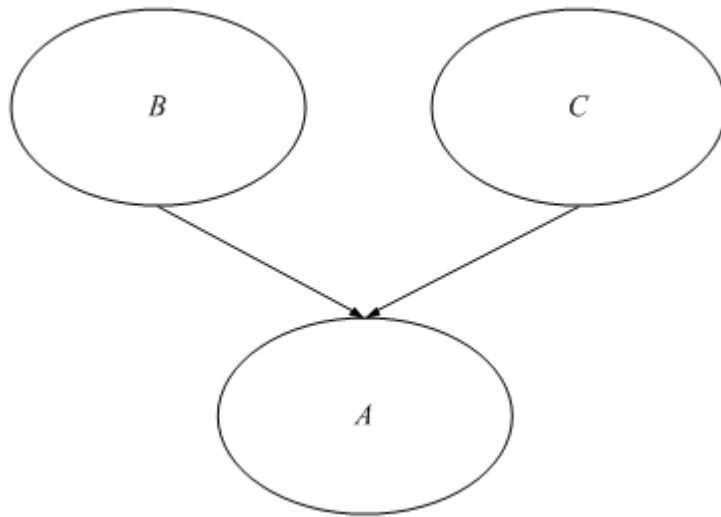


Figure: 6.6 - Réseau bayésien convergent

où cette fois les parents sont indépendants.

Donc B et C sont indépendants mais deviennent dépendants conditionnellement à A . Si A est connue, nous avons alors:

$$P(A | B, C) = P(A | B) \quad (6.77)$$

La dépendance entre les parents passe donc par l'observation de leur enfant commun.

Maintenant, pour faire un exemple concret, imaginons que notre base de données nous donne (grâce aux responsables qualité qui ont toujours su saisir les anomalies qualité) que lorsqu'une panne machine a eu lieu, 99 fois sur 100 (99%) il y a eu un arrêt total de la production (donc in extenso 1 fois sur 100: 1% il n'y pas eu d'arrêt de la production) et que sur tous les arrêts de la production 1% n'était pas dû à une panne machine. Ce que nous représentons traditionnellement sous la forme suivante:

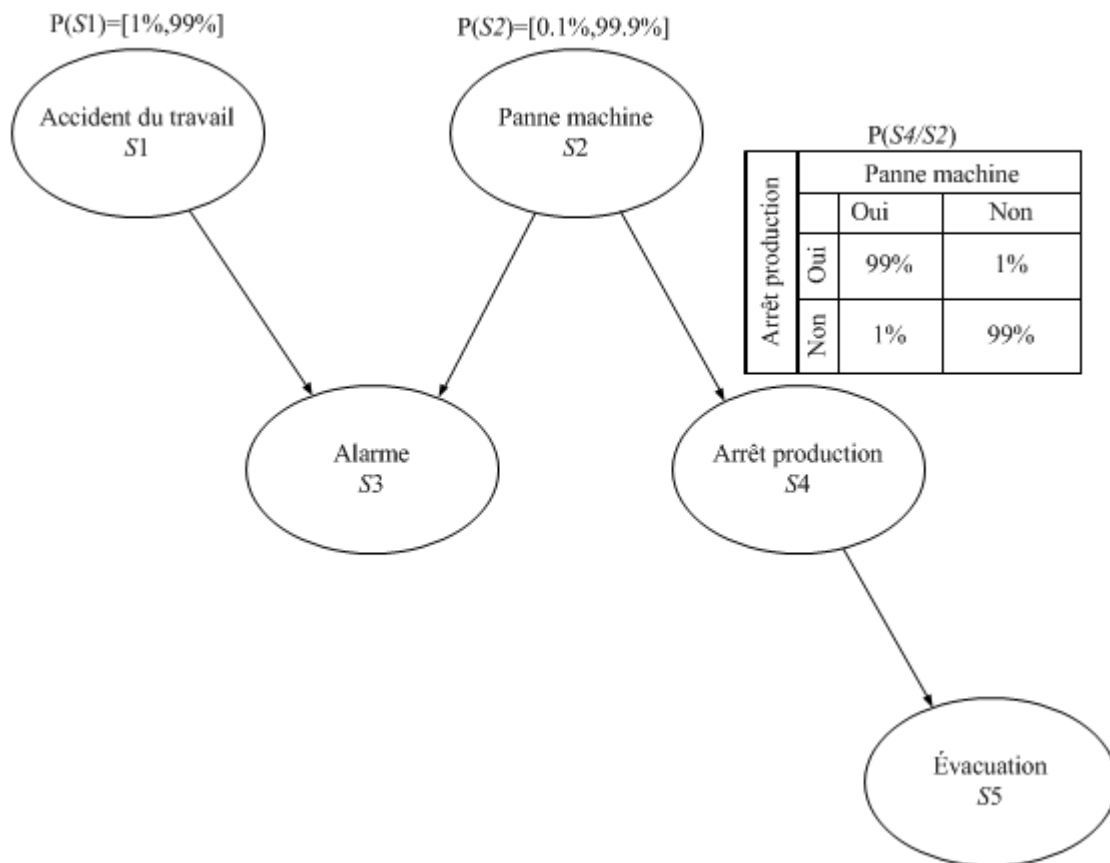


Figure: 6.7 - Réseau bayésien de 1er niveau

Donc la "**probabilité implicite**" qu'il y ait un arrêt de la production est donnée par:

$$P(S4) = P(S4 / S2) P(S2) + P(S4 / \overline{S2}) P(\overline{S2}) = 99\% \cdot 0.1\% + 1\% \cdot 99.9\% \cong 1.098\% \quad (6.78)$$

Ce chiffre représente donc la proportion implicite d'arrêts de production parmi les 100'000 jours hommes (nous pouvons donc donner une proportion de lignes de la base données représentant un arrêt production quelle que soit la cause et ce sans même avoir les détails de la base de données).

Il en découle immédiatement alors la probabilité implicite qu'il n'y ait pas d'arrêt de la production:

$$P(\overline{S4}) = 1 - P(S4) \cong 98.902\% \quad (6.79)$$

Ce qui est conforme à ce que nous donne le logiciel gratuit MSBNX 1.4.2:

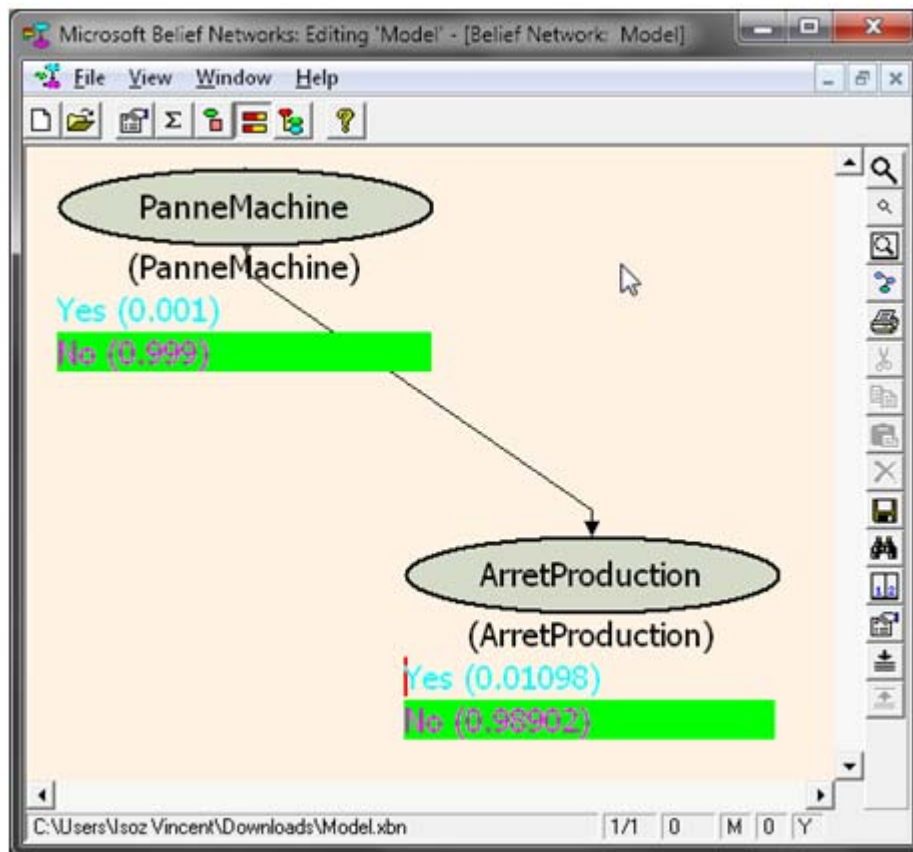


Figure: 6.8 - Début du réseau bayésien dans MSBNX 1.4.2

Maintenant supposons que nous avons observé un arrêt de la production. Quelle est la probabilité à posteriori qu'il soit dû à une panne machine? Nous avons alors:

$$P(S2 / S4) = \frac{P(S4 / S2) P(S2)}{P(S4)} = \frac{99\% \cdot 0.1\%}{1.098\%} \cong 9.02\% \quad (6.80)$$

Ce que nous pouvons aussi vérifier avec le logiciel MSBNX 1.4.2:

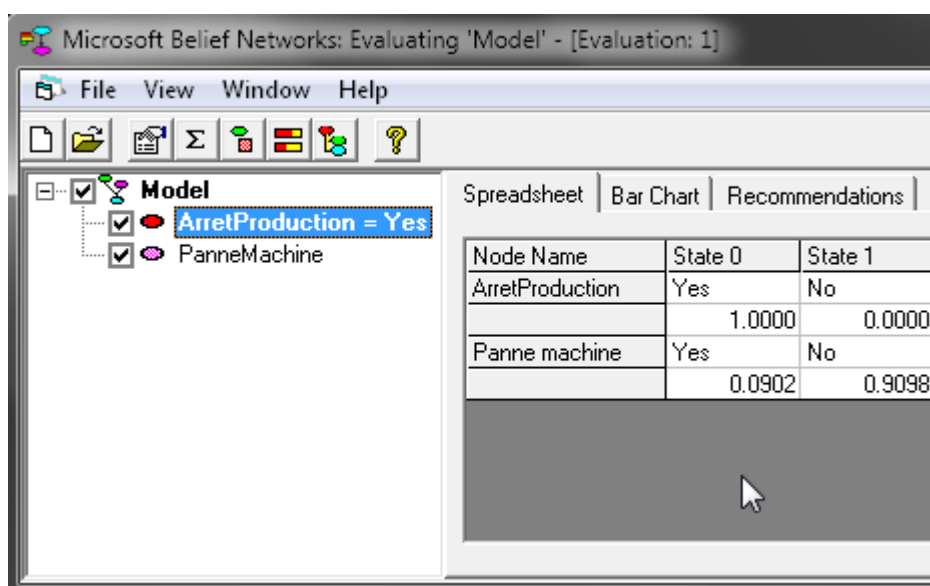


Figure: 6.9 - Probabilité a posteriori d'un arrêt dû à une panne machine dans MSBNX 1.4.2

Maintenant, imaginons que notre base de données nous donne (toujours grâce aux responsables qualité qui ont veillé à saisir les anomalies qualité) que 99 fois sur 100 (99%) lorsqu'il y a eu un arrêt de la production, il y a eu une évacuation. En revanche 5 % des évacuations ont été identifiées comme n'ayant rien à voir avec un arrêt de la production (donc 95% des évacuations sont dues à des exercices incendie OU à d'autres événements):

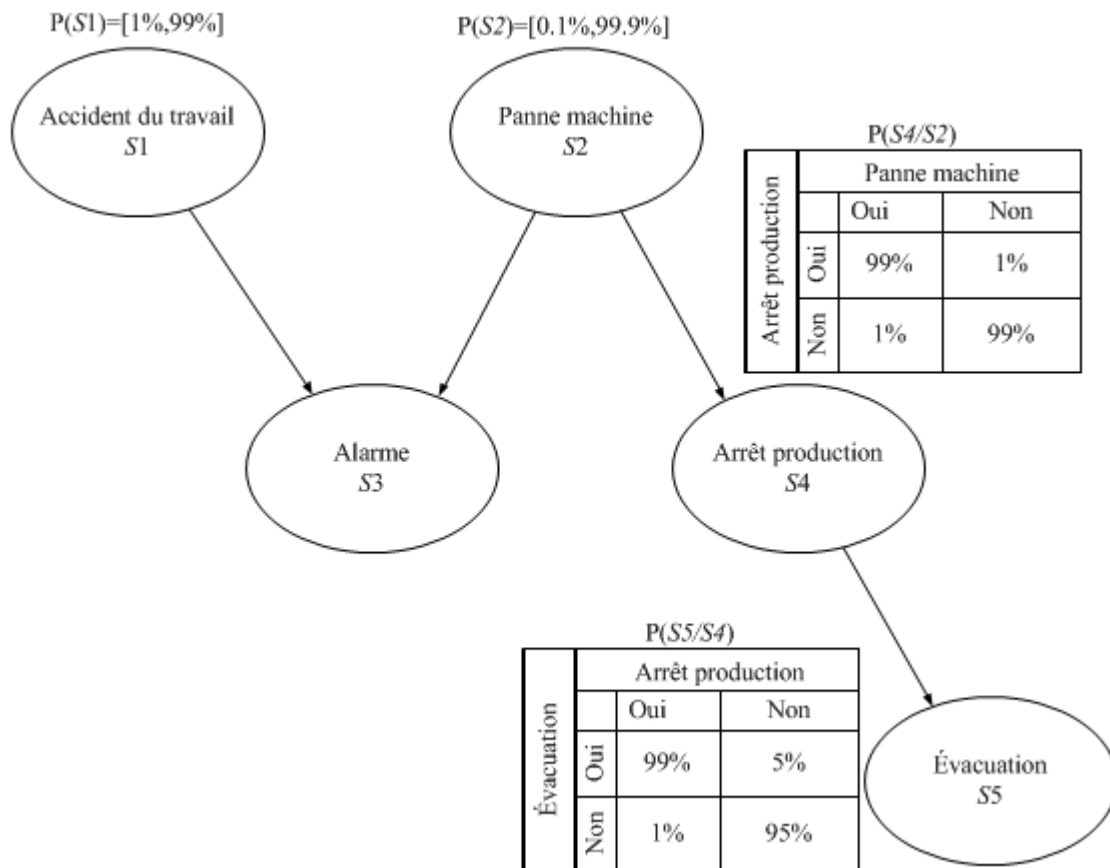


Figure: 6.10 - Réseau bayésien de deuxième niveau

Maintenant, pour calculer la probabilité implicite des évacuations a posteriori par rapport aux pannes machines, nous avons vu que lorsque nous avons une dépendance conditionnelle série, la probabilité conditionnelle ne dépendait que du parent direct. Ainsi, il vient:

$$P(S5) = P(S5 / S4) P(S4) + P(S5 / \overline{S4}) P(\overline{S4}) = 99\% \cdot 1.098\% + 5\% \cdot 98.902\% \cong 6.03\% \quad (6.81)$$

Ce qui peut se vérifier avec le logiciel MSBNX 1.4.2:

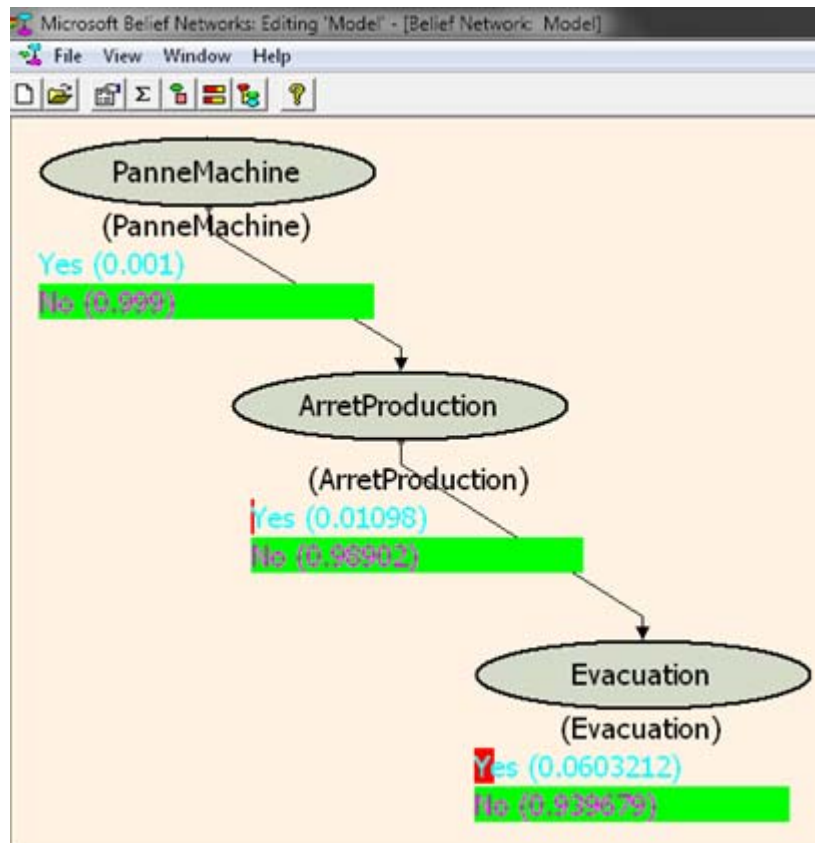


Figure: 6.11 - Probabilité implicite d'une évacuation dans MSBNX 1.4.2

Donc la probabilité implicite de l'évacuation ne dépend effectivement pas des pannes de machines.

Maintenant supposons que nous avons observé une évacuation. Nous voulons savoir quelle est la probabilité a posteriori qu'elle soit due à une panne machine! Nous avons alors:

$$P(S4 / S5) = \frac{P(S5 / S4) P(S4)}{P(S5)} = \frac{99\% \cdot 1.098\%}{6.03\%} \cong 18.02\% \quad (6.82)$$

Ce que nous pouvons aussi vérifier avec le logiciel MSBNX 1.4.2:

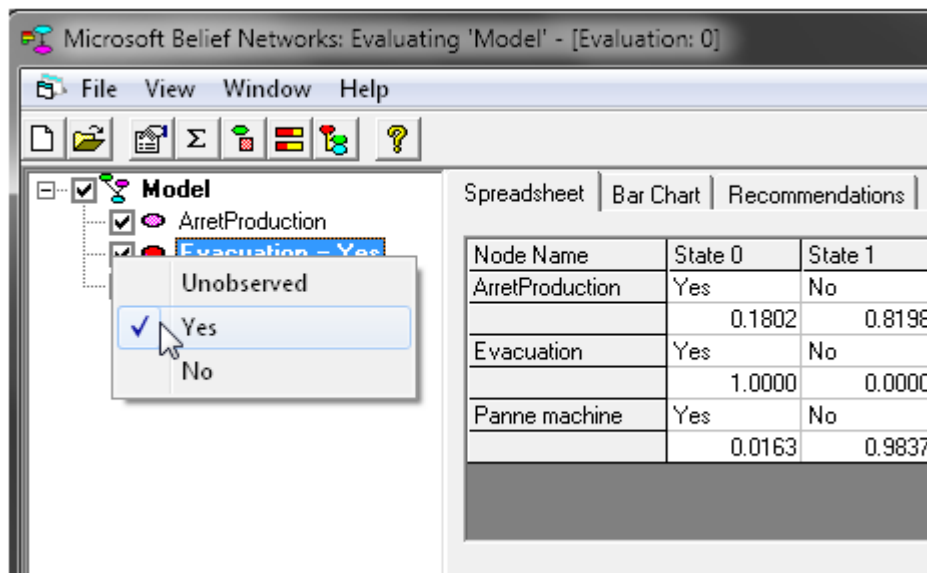


Figure: 6.12 - Probabilité a posteriori d'une évacuation due à une panne machine dans MSBNX 1.4.2

Maintenant nous étudions le cas avec l'alarme et là aussi une base de données nous permet de construire un tableau avec les différentes probabilités:

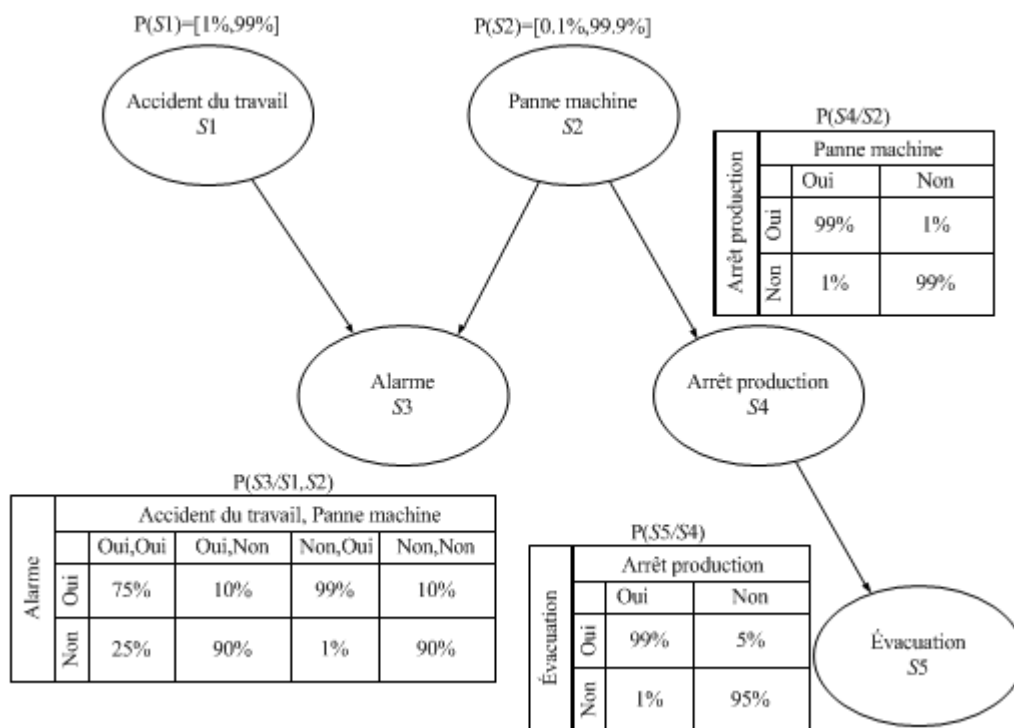


Figure: 6.13 - Réseau bayésien de deuxième niveau avec seconde branche

Maintenant, pour calculer la probabilité implicite qu'il y ait une alarme, il va falloir considérer les quatre situations possibles. Nous avons alors en utilisant le théorème des probabilités totales:

$$\begin{aligned}
 P(S3) &= P(S3/S1,S2)P(S1)P(S2) + P(S3/\bar{S1},S2)P(\bar{S1})P(S2) \\
 &+ P(S3/S1,\bar{S2})P(S1)P(\bar{S2}) + P(S3/\bar{S1},\bar{S2})P(\bar{S1})P(\bar{S2})
 \end{aligned}
 \quad (6.83)$$

Ce qui un peu plus rigoureusement devrait s'écrire:

$$\begin{aligned}
 P(S3) &= P(S3 / S1 \wedge S2) P(S1) P(S2) + P(S3 / \overline{S1} \wedge S2) P(\overline{S1}) P(S2) \\
 &+ P(S3 / S1 \wedge \overline{S2}) P(S1) P(\overline{S2}) + P(S3 / \overline{S1} \wedge \overline{S2}) P(\overline{S1}) P(\overline{S2})
 \end{aligned}
 \quad (6.84)$$

L'application numérique donne donc pour la probabilité implicite d'une alarme:

$$\begin{aligned}
 P(S3) &= 75\% \cdot 1\% \cdot 0.1\% + 99\% \cdot 99\% \cdot 0.1\% + 10\% \cdot 1\% \cdot 99.9\% + 10\% \cdot 99\% \cdot 99.9\% \\
 &\cong 10.089\%
 \end{aligned}
 \quad (6.85)$$

Ce qui se construit et se vérifie de la manière suivante avec MSBNX 1.4.2:

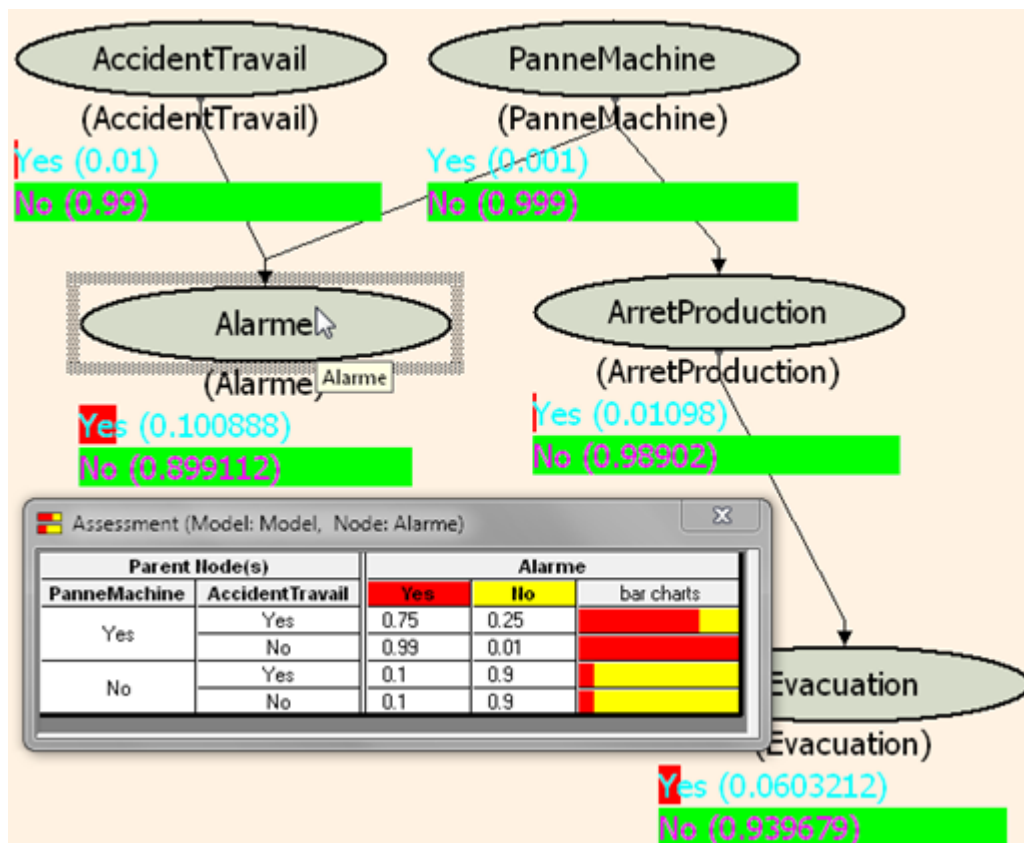


Figure: 6.14 - Probabilité implicite d'une alarme dans MSBNX 1.4.2

Concernant les notations, il peut être utile au lecteur de savoir qu'il peut parfois trouver dans la littérature:

$$P(S3 / S1 = \text{Oui}, S2 = \text{Oui}) = P(S3 / S1, S2) = P(S3 / S1 \wedge S2) \quad (6.86)$$

Remarque: Dans l'exemple particulier étudié ici les événements ont tous deux états. Mais dans la pratique cela peut aller à 3, 4 et plus. Dès lors les tableaux de croisement de probabilités deviennent vite énormes.

Comme pour les cas précédents, supposons que nous savons qu'il y a eu un accident de travail. Nous souhaitons alors calculer la probabilité a priori d'une alarme. Nous avons alors (observez que la probabilité ne dépend effectivement alors plus que de l'état S2 puisque l'état S1 est entièrement connu!):

$$\begin{aligned}
 P(S3/S1) &= P(S3/S1, S2)P(S2) + P(S3/S1, \overline{S2})P(\overline{S2}) \\
 &= 75\% \cdot 0.1\% + 10\% \cdot 99.9\% = 10.065\%
 \end{aligned}
 \tag{6.87}$$

Ce que nous pouvons aussi vérifier avec le logiciel MSBNX 1.4.2:

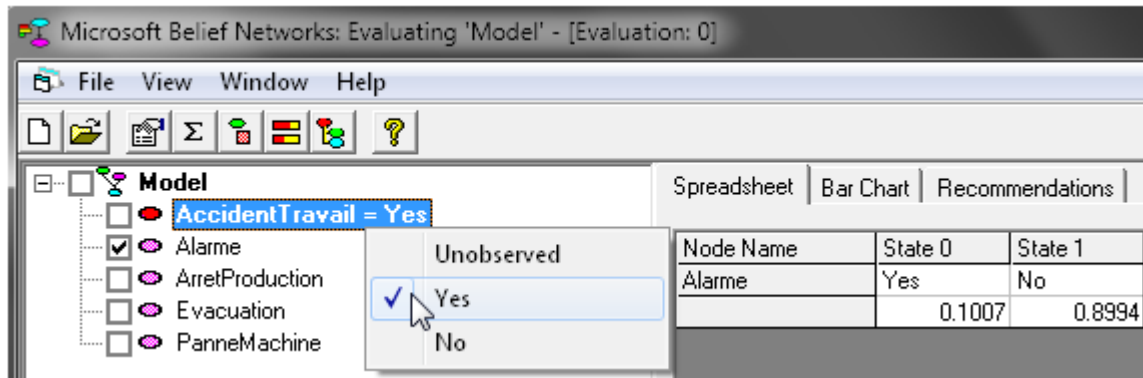


Figure: 6.15 - Probabilité implicite d'une alarme dans MSBNX 1.4.2

Ainsi, savoir qu'il y a eu un accident de travail augmente la probabilité qu'il y ait bien une alarme (nous passons d'une probabilité de 10.089% à 10.65%).

Pour terminer cet exemple, nous souhaiterions calculer les probabilités a posteriori $P(S2/S3)$ et $P(S1/S3)$. Pour cela, nous devons d'abord calculer les probabilités a priori $P(S3/S2)$ et $P(S3/S1)$ (cette dernière venant d'être calculée).

Nous avons pour la valeur manquante (ce qui se vérifie aussi facilement qu'avant avec le logiciel MSBNX 1.4.2):

$$\begin{aligned}
 P(S3/S2) &= P(S3/S2, S1)P(S1) + P(S3/S2, \overline{S1})P(\overline{S1}) \\
 &= 75\% \cdot 1\% + 99\% \cdot 99\% = 98.76\%
 \end{aligned}
 \tag{6.88}$$

Nous avons donc:

$$\begin{aligned}
 P(S3/S1) &= 10.065\% \\
 P(S3/S2) &= 98.76\%
 \end{aligned}
 \tag{6.89}$$

Nous avons maintenant tout ce qu'il faut pour calculer la probabilité a priori de $P(S3/S2)$ et $P(S3/S1)$:

$$\begin{aligned}
 P(S2/S3) &= \frac{P(S3/S2)P(S2)}{P(S3)} = \frac{98.76\% \cdot 0.1\%}{10.089\%} = 0.9789\% \\
 P(S1/S3) &= \frac{P(S3/S1)P(S1)}{P(S3)} = \frac{10.065\% \cdot 1\%}{10.089\%} = 0.9976\%
 \end{aligned}
 \tag{6.90}$$

Donc la probabilité a priori qu'il y ait une panne machine lorsque nous savons qu'il y a une alarme est de 0.979% (donc in extenso 99.021% que le déclenchement de l'alarme ne soit pas dû a priori à une panne machine). Respectivement il y a, a priori, 0.998 % de probabilité y ait un accident de travail lorsque nous savons qu'il y a une alarme (et donc 99.002 % que cela ne soit pas dû a priori à un accident de travail).

Du point de vue critique, lorsqu'il y a donc une alarme finalement nous ne pouvons pas dire grand chose. Cela est dû dans le cas présent au fait que les événements d'intérêt notable ont tous deux de faibles probabilités d'avoir lieu (accident et panne machine) et que les gens réagissent plutôt bien au niveau du déclenchement de l'alarme (sinon si les probabilités a priori étaient grandes cela signifierait que le comportement n'est pas bon puisque nous pouvons deviner (avec exaspération) à l'avance quel problème a lieu avec une certaine confiance).

Remarque: Nous n'avons pas trouvé comment vérifier ces derniers calculs avec MSNBX 1.4.2. Si quelqu'un trouve comment le faire ce serait super de nous communiquer le détail de la démarche.

Pour clore, le lecteur aura remarqué que les calculs peuvent vite devenir ennuyeux dès que le graphe devient complexe d'où l'usage de logiciels informatiques. De plus, dans le domaine bancaire qui utilise par exemple les réseaux bayésiens pour les risques de crédit, la probabilité a priori peut être plus complexe. Par exemple nous pourrions vouloir connaître la probabilité a priori qu'il y ait une panne machine sachant que nous avons une alarme et un accident de travail:

$$P(S2 / S3, S1) = P(S2 / S3 = \text{Oui}, S1 = \text{Oui}) \quad (6.91)$$

1.3. MARTINGALES

Une martingale en probabilités (il en existe une autre dans les processus stochastiques) est une technique permettant d'augmenter les chances de gain aux jeux de hasard tout en respectant les règles de jeu. Le principe dépend complètement du type de jeu qui en est la cible, mais le terme est accompagné d'une aura de mystère qui voudrait que certains joueurs connaissent des techniques secrètes mais efficaces pour tricher avec le hasard. Par exemple, de nombreux joueurs (ou candidats au jeu) cherchent LA martingale qui permettra de battre la banque dans les jeux les plus courants dans les casinos (des institutions dont la rentabilité repose presque entièrement sur la différence - même faible - qui existe entre les chances de gagner et celles de perdre).

De nombreuses martingales ne sont que le rêve de leur auteur, certaines sont en fait inapplicables, quelques-unes permettent effectivement de tricher un peu. Les jeux d'argent sont en général inéquitables: quel que soit le coup joué, la probabilité de gain du casino (ou de l'État dans le cas d'une loterie) est plus importante que celle du joueur. Dans ce type de jeu, il n'est pas possible d'inverser les chances, seulement de minimiser la probabilité de ruine du joueur.

L'exemple le plus courant est la martingale de la roulette. Elle consiste à jouer une chance simple à la roulette (noir ou rouge, paire ou impaire) de façon à gagner, par exemple, une unité dans une série de coups en doublant sa mise si l'on perd, et cela jusqu'à ce que l'on gagne. Exemple: le joueur mise 1 unité sur le rouge, si le rouge sort, il arrête de jouer et il a gagné 1 unité (2 unités de gain moins l'unité de mise), si le noir sort, il double sa mise en pariant 2 unités sur le rouge et ainsi de suite jusqu'à ce qu'il gagne.



Figure: 6.16 - Roulette de casino

Ayant une chance sur deux de gagner, il peut penser qu'il va finir par gagner ; quand il gagne, il est forcément remboursé de tout ce qu'il a joué, plus une fois sa mise de départ.

Cette martingale semble être sûre en pratique. À noter que sur le plan théorique, pour être sûr de gagner, il faudrait avoir la possibilité de jouer au cas où un nombre de fois illimité.... Ce qui présente des inconvénients majeurs:

Cette martingale est en fait limitée par les mises que le joueur peut faire car il faut doubler la mise à chaque coup tant que l'on perd: 2 fois la mise de départ, puis 4, 8, 16.... s'il perd 10 fois de suite, il doit pouvoir avancer 1024 fois sa mise initiale pour la 11e partie ! Il faut donc beaucoup d'argent pour gagner peu.

Les roulettes comportent de plus un "0" qui n'est ni rouge ni noir. Le risque de perdre lors de chaque coup est ainsi plus grand que 1/2...

De plus, pour paralyser cette stratégie, les casinos proposent des tables de jeu par tranche de mise: de 1 à 100.-, de 2 à 200.-, de 5 à 500.-, ... Impossible donc d'utiliser cette méthode sur un grand nombre de coups, ce qui augmente le risque de tout perdre.

Le black jack est un jeu qui possède des stratégies gagnantes: plusieurs techniques de jeu, qui nécessitent généralement de mémoriser les cartes, permettent de renverser les chances en faveur du joueur. Le mathématicien Edward Thorp a ainsi publié en 1962 un livre qui fut à l'époque un véritable best-seller. Mais toutes ces méthodes demandent de longues semaines d'entraînement et sont facilement décelables par le croupier (les brusques changements de montant des mises sont caractéristiques). Le casino a alors tout loisir d'écarter de son établissement les joueurs en question.

Il faut noter qu'il existe des méthodes assez évoluées. L'une d'elles repose sur les combinaisons les moins jouées. Dans les jeux où le gain dépend du nombre de joueurs gagnants (Loto...), jouer les combinaisons les moins jouées optimisera les gains. C'est ainsi que certaines personnes vendent des combinaisons qui seraient statistiquement très rarement utilisées par les autres joueurs.

Partant de ce raisonnement, on peut encore conclure qu'un joueur qui aurait réussi à déterminer ainsi les combinaisons statistiquement les moins jouées, afin d'optimiser son espérance de gain, ne sera en fait certainement pas le seul joueur à avoir obtenu par l'analyse ces fameuses combinaisons! Cela revient à dire que les numéros en théorie les moins joués sont en fait surjoués par combinaisons, le mieux serait peut-être de réaliser un savant mélange de numéros sous-joués et de numéros surjoués pour obtenir les combinaisons idéales. Une autre conclusion à tout cela est peut-être que le mieux est encore de jouer des combinaisons aléatoires qui ont finalement moins de chance d'être également choisies par les joueurs qui incorporent un facteur humain et harmonieux dans le choix de leurs nombres.

2. ANALYSE COMBINATOIRE

"L'analyse combinatoire" (techniques de dénombrement) est le domaine de la mathématique qui s'occupe de l'étude de l'ensemble des issues, événements ou faits (distinguable ou non tous distinguables) avec leurs arrangements (combinaisons) ordonnés ou non selon certaines contraintes données.

Définitions:

D1. Une suite d'objets (événements, issues, objets,...) est dite "**ordonnée**" si chaque suite composée d'un ordre particulier des objets est comptabilisée comme une configuration particulière.

D2. Une suite est donc "**non ordonnée**" si et seulement si nous intéresse la fréquence d'apparition des objets indépendamment de leur ordre.

D3. Des objets (d'une suite) sont dits "**distincts**" si leurs caractéristiques ne permettent pas de les confondre avec des autres objets.

Remarque: Nous avons choisi de mettre l'analyse combinatoire dans ce chapitre car lorsque nous calculons des probabilités, nous avons également assez souvent besoin de savoir quelle est la probabilité de tomber sur une combinaison ou un arrangement d'événements donnés sous certaines contraintes.

Souvent les étudiants ont de la peine à se rappeler de la différence entre une permutation, un arrangement et une combinaison. Voici donc un petit résumé de ce que nous allons voir:

- Permutation: On prend tous les éléments.
- Arrangement: On choisit des éléments parmi ceux de l'ensemble de départ et l'ordre intervient
- Combinaison: Idem que pour l'arrangement mais l'ordre n'intervient pas

Il existe plusieurs types d'arrangements selon les contraintes et les propriétés des éléments arrangés. Nous allons présenter et démontrer ci-dessous les 5 cas les plus répandus à partir desquels nous pouvons trouver (habituellement) tous les autres:

2.1. ARRANGEMENTS SIMPLES AVEC RÉPÉTITION

Définition: Un "**arrangement simple avec répétition**" est une suite ordonnée de longueur m de n objets distincts non nécessairement tous différents dans la suite (soit avec répétitions possibles!).

Soient A et B deux ensembles finis de cardinaux respectifs m , n tels que trivialement il y ait m façons de choisir un objet dans A (de type a) et n façons de choisir un objet dans B (de type b).

Nous avons vu dans le chapitre de Théorie Des Ensembles que si A et B sont disjoints, que:

$$\text{Card}(A \cup B) = m + n \quad (6.92)$$

Nous en déduisons donc les propriétés suivantes:

P1. Si un objet ne peut être à la fois de type a et de type b et s'il y a m façons de choisir un objet de type a et n façons de choisir un objet de type b , alors l'union des objets donne $m + n$ sélections (c'est typiquement le résultat des requêtes d'UNION en SQL, sans filtres, dans les SGBDR des entreprises).

P2. Si nous pouvons choisir un objet de type a de m façons puis un objet de type b de n façons, alors il y a selon le produit cartésien de deux ensembles (cf. chapitre de Théorie Des Ensembles):

$$\text{Card}(A \times B) = \text{Card}(A) \cdot \text{Card}(B) = m \cdot n \quad (6.93)$$

de manières choisir un seul et unique objet de type a puis un objet de type b (this is typically the result of SELECT queries in SQL, without filters, with several unrelated tables in corporate RDBMS).

Avec les mêmes notations pour m et n , nous pouvons donc choisir pour chaque élément de A , son unique image parmi les n éléments de B . Il y a donc n façons de choisir l'image du premier élément de A , puis aussi n façons de choisir l'image du deuxième élément de A , ..., puis n façons de choisir l'image du m -ème élément de A . Le nombre d'applications totales consécutives possibles de A dans B est donc égal aux m produits de n (m fois le produit cartésien du cardinal de l'ensemble B avec lui-même donc!). Ce qu'il est d'usage d'écrire (nous avons mis les différentes écritures que l'on peut trouver dans les livres scolaires):

$$\text{Card}(B^A) = \text{Card}\left(\underbrace{B \times B \times \dots \times B}_{m \text{ fois}}\right) = \text{Card}(B)^m = \overline{A}_n^m = n^m \quad (6.94)$$

où B^A est l'ensemble des applications de A dans B . La progression du nombre de possibilités est donc géométrique (et non "exponentielle" comme il est souvent dit à tort!).

Ce résultat mathématique est assimilable au résultat ordonné (un arrangement $\overline{A}_n^m$ dont l'ordre des éléments de la suite est pris en compte) de m tirages dans un sac contenant n boules différentes avec remise après chaque tirage.

Exemples:

E1. Combien de "mots" (ordonnés) de 7 lettres pouvons-nous former à partir d'un alphabet de 24 lettres distinctes (très utile pour connaître le nombre d'essais pour trouver un mot de passe par exemple)? La solution est:

$$\overline{A}_{24}^7 = 24^7 = 4'586'471'424 \quad (6.95)$$

E2. Combien de groupes d'individus aurons-nous lors d'une votation sur 5 sujets et où chacun peut être soit accepté, soit rejeté? La solution (très utilisée dans les entreprises en Suisse) est:

$$\overline{A}_2^5 = 2^5 = 32 \quad (6.96)$$

Une généralisation simple de ce dernier résultat peut consister dans l'énoncé du problème suivant:

Si nous disposons de m objets $k_1, \dots, k_m$ tels que k_i peut prendre n_i états différents alors le nombre de combinaisons possibles est:

$$A_{mn} = n_1 n_2 \dots n_i \dots n_m \quad (6.97)$$

Et si nous avons $n_1 = n_2 = \dots = n_m$ alors nous retombons sur:

$$A_{mn} = \overline{A}_n^m = n_1 n_2 \dots n_i \dots n_m = n^m \quad (6.98)$$

Exemple:

Un graphiste a créé sous le logiciel Adobe Photoshop une maquette d'une site Internet avec trois en-têtes différentes, deux variantes pour le corps, quatre variantes pour le fond, six variantes de menus et trois variantes pour le pied de pages. Le nombre total de combinaisons (compositions dans le langage du graphiste) que l'on pourra présenter au client sera de:

$$A_{mm} = 3 \cdot 2 \cdot 4 \cdot 6 \cdot 3 = 432 \quad (6.99)$$

2.2. PERMUTATIONS SIMPLES SANS RÉPÉTITION

Définition: Une "**permutation simple sans répétition**" (appelée anciennement "substitution") de n objets distincts est une suite ordonnée (différente) de ces n objets par définition tous différents dans la suite (sans répétition).

Remarque: Attention à ne pas confondre le concept de permutation (de n éléments entre eux) et d'arrangement (de n éléments parmi m)!

Le nombre de permutations de n éléments peut être calculé par récurrence: il y a n places pour un premier élément, $n-1$ pour un deuxième élément, ..., et il ne restera qu'une place pour le dernier élément restant.

Il est dès lors trivial que nous aurons un nombre de permutations donné par:

$$n \cdot (n-1) \cdot (n-2) \cdot (n-3) \dots (n-(n-1)) \quad (6.100)$$

Rappelons que le produit:

$$n \cdot (n-1) \cdot (n-2) \cdot (n-3) \dots (n-(n-1)) = \prod_{i=1}^n i \quad (6.101)$$

est appelé "**factorielle de n** " et nous la notons $n!$ pour $n \in \mathbb{N}$.

Il y a donc pour n éléments distinguables:

$$\boxed{A_n = n!} \quad (6.102)$$

permutations possibles. Ce type de calcul peut être par exemple utile en gestion de projets (calcul du nombre de manière différentes de recevoir dans une chaîne de production n pièces toutes différentes commandées chez des fournisseurs externes).

 Exemple:

Combien de "mots" (ordonnés) de 7 lettres distinctes sans répétition pouvons-nous former?

$$A_7 = 7! = 5040 \quad (6.103)$$

Ce résultat nous amène à l'assimiler au résultat ordonné (un arrangement A_n dont l'ordre des éléments de la suite est pris en compte) du tirage de toutes les boules différentes d'un sac contenant n boules distinguables sans remise.

2.3. PERMUTATIONS SIMPLES AVEC RÉPÉTITION

Définition: Lorsque nous considérons le nombre de permutations ordonnées (différentes) d'une suite de n objets distincts tous nécessairement non différents dans une quantité donnée dans la suite nous parlons de "**permutation simple avec répétition**".

Remarque: Il ne faut pas confondre cette dernière définition avec "l'arrangement avec répétition" vu plus haut!

Lorsque certains éléments ne sont pas tous distinguables dans une suite d'objets (ils sont répétitifs dans la suite), alors le nombre de permutations que nous pouvons constituer se réduit alors assez trivialement à un nombre plus petit que si tous les éléments étaient tous distinguables.

Soit n_i le nombre d'objets du type i , avec:

$$n_1 + n_2 + \dots + n_k = n \quad (6.104)$$

alors, nous notons:

$$\bar{A}_n(n_1, \dots, n_i, \dots, n_k) \quad (6.105)$$

avec $i = 1, 2, \dots, k$ le nombre de permutations possibles (pour l'instant inconnu) avec répétition (un ou plusieurs éléments répétitifs dans une suite d'éléments sont non distinguables par permutation).

Si chacune des n_i places occupées par des éléments identiques était occupée par des éléments différents, le nombre de permutations serait alors à multiplier par chacun des $n_i!$ (cas précédent).

Il vient alors que nous retombons sur la factorielle telle que:

$$\bar{A}_n(n_1, n_2, n_3, \dots, n_k) \cdot n_1! \cdot n_2! \cdot n_3! \dots \cdot n_k! = n! \quad (6.106)$$

dont nous déduisons immédiatement:

$$\boxed{\bar{A}_n(n_1, n_2, n_3, \dots, n_k) = \frac{n!}{n_1! \cdot n_2! \cdot n_3! \dots \cdot n_k!}} \quad (6.107)$$

Si les n objets sont tous différents dans la suite, nous avons alors:

$$n_1! = n_2! = \dots = n_k! = 1! = 1 \quad (6.108)$$

et nous nous retrouvons bien avec une permutation simple (sans répétition) telle que:

$$\bar{A}_n = (1, 1, \dots, 1) = \frac{n!}{1!1!\dots 1!} = n! \quad (6.109)$$

Il conviendra donc de se rappeler que les permutations avec répétition sont en plus petit nombre que celles sans répétition (évident puisque nous ne prenons pas en compte les permutations des éléments identiques entre eux!).

Exemple:

Combien de "mots" (ordonnés) pouvons-nous former avec les lettres du mot "Mississippi":

$$\bar{A}_4\left(\underset{\text{M}}{1}, \underset{\text{P}}{2}, \underset{\text{I}}{4}, \underset{\text{S}}{4}\right) = \frac{11!}{1!2!4!4!} = 34'650 \quad (6.110)$$

Ce résultat nous amène à l'assimiler au résultat ordonné (une permutation $\bar{A}_n$ dont l'ordre des éléments de la suite n'est pas pris en compte) du tirage de n boules non toutes distinguables d'un sac contenant $k \geq n$ boules avec remise limitée pour chaque boule.

2.4. ARRANGEMENTS SIMPLES SANS RÉPÉTITION

Définition: Un "**arrangement simple sans répétition**" est une suite ordonnée de p objets tous distincts pris parmi n objets distincts avec $n \geq p$.

Nous nous proposons donc maintenant de dénombrer les arrangements possibles sans répétition de p objets parmi n . Nous noterons A_n^p le nombre de ces arrangements.

Il est aisé de calculer $A_n^1 = n$ et de vérifier que $A_n^2 = n(n-1)$. Effectivement, il existe n façons de choisir le premier objet et $(n-1)$ façons de choisir le deuxième lorsque nous avons déjà le premier.

Pour déterminer A_n^p , nous raisonnons alors par récurrence. Nous supposons A_n^{p-1} connu et nous en déduisons:

$$A_n^p = A_n^{p-1} \cdot [n - (p-1)] = A_n^{p-1} \cdot (n - p + 1) \quad (6.111)$$

Dès lors:

$$A_n^p = n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdot \dots \cdot (n - (p-1)) \quad (6.112)$$

alors:

$$n! = A_n^p \cdot (n-p)! = [n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot (n-p+2) \cdot (n-p+1)] \cdot (n-p) \cdot (n-p-1) \dots \quad (6.113)$$

d'où:

$$A_n^p = \frac{n!}{(n-p)!} \quad (6.114)$$

Ce résultat nous amène à l'assimiler au résultat ordonné (un arrangement A_n^p dont l'ordre des éléments de la suite est pris en compte) du tirage de p boules distinctes d'un sac contenant n boules différentes sans remise.

Exemple:

Soit les 24 lettres de l'alphabet, combien de "mots" (ordonnés) de 7 lettres distinctes pouvons-nous former?

$$A_{24}^7 = \frac{24!}{(24-7)!} = 1'744'364'160 \quad (6.115)$$

Le lecteur aura peut-être remarqué que si nous prenons $p = n$ nous nous retrouvons avec:

$$A_n^n = \frac{n!}{(n-p)!} = \frac{n!}{0!} = n! \quad (6.116)$$

donc nous pouvons dire qu'une permutation simple de n éléments est comme un arrangement simple sans répétition avec $n = p$.

2.5. COMBINAISONS SIMPLES

Définition: Une "**combinaison simple**" ou "**choix**" est une suite non-ordonnée (dont l'ordre ne nous intéresse pas!) de p éléments tous différents (pas nécessairement dans le sens visuel du terme!) choisis parmi n objets distincts et est par définition notée sur ce site Internet C_p^n et appelée la "**binomiale**" ou "**coefficient binomial**".

Si nous permutons les éléments de chaque arrangement simple de p éléments parmi n , nous obtenons toutes les permutations simples et nous savons qu'il y en a $p!$ d'où en utilisant la convention d'écriture du présent site internet (contraire à celle préconisée par la norme ISO 31-11!):

$$C_p^n = \binom{n}{p} = \frac{A_p^n}{p!} = \frac{n!}{p! (n-p)!} \quad (6.117)$$

C'est une relation très souvent utilisée dans les jeux de hasard mais également dans l'industrie via la loi hypergéométrique (cf. chapitre de Techniques De Gestion) ainsi que dans les statistiques d'assez haut niveau comme les statistiques d'ordre (cf. chapitre de Statistiques).

Ce résultat nous amène à l'assimiler au résultat non ordonné (un arrangement C_p^n dont l'ordre des éléments de la suite n'est pas pris en compte) du tirage de p boules d'un sac contenant n boules différentes sans remise.

Remarques:

R1. Nous avons nécessairement par construction $C_p^n \leq A_p^n$.

R2. Selon les auteurs nous inversons l'indice ou le suffixe de C il faut donc être prudent!

Exemple:

Soit un alphabet de 24 lettres, combien avons-nous de choix de prendre 7 lettres parmi les 24 sans prendre en compte l'ordre dans lequel sont triées les lettres:

$$C_p^n = \frac{24!}{7!(24-7)!} = 346'104 \quad (6.118)$$

La même valeur peut être obtenue avec la fonction COMBIN() de Microsoft Excel 11.8346 (version française).

Il existe, relativement à la binomiale, une autre relation très souvent utilisée dans de nombreux cas d'études ou également de manière plus globale en physique ou analyse fonctionnelle. Il s'agit de la "**formule de Pascal**":

$$C_{p-1}^{n-1} + C_p^{n-1} = C_p^n \quad (6.119)$$

Démonstration:

$$C_{p-1}^{n-1} + C_p^{n-1} = \frac{(n-1)!}{((n-1)-(p-1))!(p-1)!} + \frac{(n-1)!}{(n-1-p)!p!} = \frac{(n-1)!}{(n-p)!(p-1)!} + \frac{(n-1)!}{(n-1-p)!p!} \quad (6.120)$$

Or $p! = p(p-1)!$ donc:

$$(p-1)! = \frac{p!}{p} \quad (6.121)$$

et de même $(n-p)(n-p-1)! = (n-p)!$:

$$(n-p-1)! = \frac{(n-p)!}{n-p} \quad (6.122)$$

Ainsi:

$$\begin{aligned} C_{p-1}^{n-1} + C_p^{n-1} &= \frac{(n-1)!}{(n-p)!(p-1)!} + \frac{(n-1)!}{(n-1-p)!p!} = \frac{(n-1)!p}{(n-p)!p!} + \frac{(n-1)!(n-p)}{(n-p)!p!} \\ &= \frac{(n-1)!}{(n-p)!p!} [p + (n-p)] = \frac{(n-1)!n}{(n-p)!p!} = \frac{n!}{(n-p)!p!} = C_p^n \end{aligned} \quad (6.123)$$

□C.Q.F.D.

Pour résumer:

Type	Expression
Arrangement simple avec répétition	$\bar{A}_n^m = n^m$
Arrangement simple sans répétition	$A_n^m = \frac{n!}{(n-m)!}$
Permutation simple sans répétition	$A_n = n!$
Permutation simple avec répétition	$\bar{A}_n(n_1, n_2, n_3, \dots, n_k) = \frac{n!}{n_1! n_2! n_3! \dots n_k!}$
Combinaison simple (cas de l'arrangement simple sans répétition où l'ordre n'est pas pris en compte)	$C_m^n = \binom{n}{m} = \frac{A_n^m}{m!} = \frac{n!}{m!(n-m)!}$

Tableau: 28.2 - Résumé des cas possibles



3. CHAÎNES DE MARKOV

Les chaînes de Markov sont des outils statistiques et probabilistes simples et puissants mais dont la forme de présentation mathématique prête parfois à l'horreur.... Nous allons tenter ici de simplifier un maximum les notations pour introduire cet outil formidable très utilisé au sein des entreprises pour gérer la logistique, les files d'attentes aux centrales d'appel ou aux caisses de magasins jusqu'à la théorie de la défaillance pour la maintenance préventive, en physique statistique ou en génie biologique (et la liste est encore longue et pour plus de détails le lecteur pourra se reporter aux chapitres concernés disponibles sur le site...).

Définitions:

D1. Nous noterons $\{X(t)\}_{t \in T}$ un processus probabiliste fonction du temps dont la valeur à chaque instant dépend de l'issue d'une expérience aléatoire. Ainsi, à chaque instant t , $X(t)$ est donc une variable aléatoire que nous désignons par "**processus stochastique**" (pour plus de détails dans le cadre de finance, voir le chapitre d'Économie).

D2. Si nous considérons un temps discret, nous notons alors $\{X_n\}_{n \in \mathbb{N}}$ un "**processus stochastique à temps discret**".

D3. Si nous supposons en outre que les variables aléatoires X_n ne peuvent prendre qu'un ensemble discret de valeurs nous parlons alors de "**processus à temps discret et à espace discret**".

Remarque: Il est tout à fait possible comme dans l'étude du télétrafic (cf. chapitre Techniques De Gestion) d'avoir un processus à temps continu et à espace d'états discrets.

Définition: $\{X_n\}_{n \in \mathbb{N}}$ est une "**chaîne de Markov**" si et seulement si:

$$P(X_n = j \mid X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \dots, X_0 = i_0) = P(X_n = j \mid X_{n-1} = i_{n-1}) \quad (6.124)$$

en d'autres termes (c'est très simple!) la probabilité pour que la chaîne soit dans un certain état à la n -ème étape du processus ne dépend que de l'état du processus à l'étape $n-1$ et pas des étapes précédentes!

Remarque: Donc en probabilités un processus stochastique vérifie la propriété markovienne ci-dessus si et seulement si la distribution conditionnelle de probabilité des états futurs, étant donné l'instant présent, ne dépend que de ce même état présent et pas des états passés. Un processus qui possède cette propriété est aussi appelé "**processus de Markov**".

Définition: Une "**chaîne de Markov homogène**" est une chaîne telle que la probabilité qu'elle a pour passer dans un certain état à la n -ième étape soit indépendante du temps. En d'autres termes, la loi de probabilité caractérisant la prochaine étape ne dépend pas du temps (de l'étape précédente), et en tout temps la loi de probabilité de la chaîne est toujours la même pour caractériser la transition à l'étape en cours.

Nous pouvons alors définir (réduire) la loi de "**probabilité de transition**" d'un état i vers un état j par:

$$p_{ij} = P(X_n = j \mid X_{n-1} = i) \quad (6.125)$$

Il est alors naturel de définir la "**matrice de transition**" ou "**matrice stochastique**":

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ \vdots & p_{32} & \ddots & \cdots \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix} \quad (6.126)$$

comme la matrice qui contient donc tous les probabilités possibles de transitions des états d'un graphe d'états orienté.

Les chaînes de Markov peuvent être représentées graphiquement sous la forme d'un graphe orienté G (cf. chapitre de Théorie Des Graphes) appelé parfois "**automate**" ayant pour sommet les points (états) i et pour arêtes les couples orientés (i, j) . Nous associons alors à chaque composante un arc orienté et sa probabilité de transition.

Exemple:

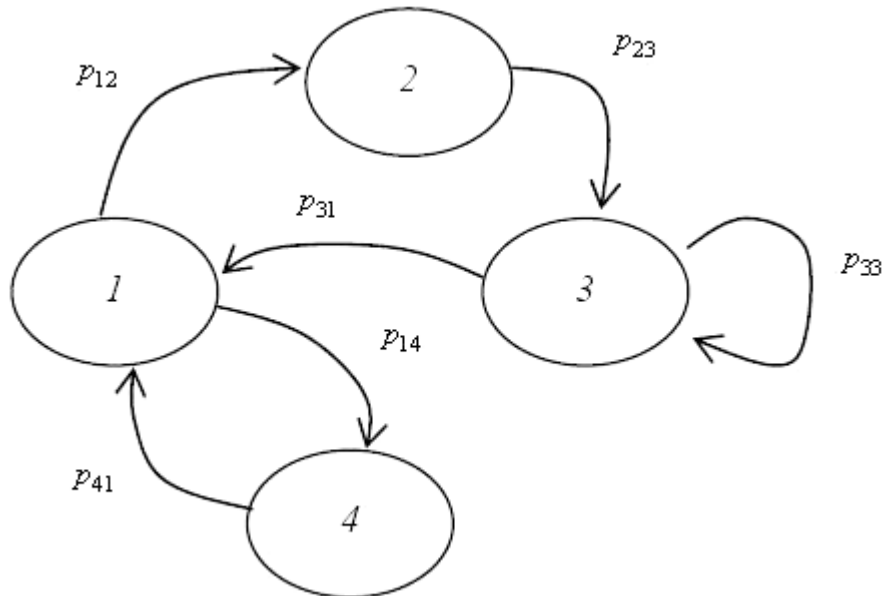


Figure: 6.17 - Exemple générique d'une chaîne de Markov

Ainsi, dans l'exemple du graphe orienté ci-dessus, les seules transitions permises par les 4 états (matrice 4×4) ci-dessus sont celles indiquées par les flèches. Ce qui fait que la matrice de transition se simplifie alors en:

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix} = \begin{bmatrix} 0 & p_{12} & 0 & p_{14} \\ 0 & 0 & p_{23} & 0 \\ p_{31} & 0 & p_{33} & 0 \\ p_{41} & 0 & 0 & 0 \end{bmatrix} \quad (6.127)$$

où le lecteur remarquera que nous avons la propriété triviale (par construction!) que la somme des termes (probabilités) d'une ligne de la matrice P est toujours unitaire (et donc que la somme des termes d'une colonne de la transposée de la matrice P est toujours unitaire aussi):

$$\sum_j p_{ij} = 1 \quad (6.128)$$

et que la matrice est positive (ce qui signifie que tous ces termes sont positifs ou nuls).

Remarque: Se rappeler que la somme des probabilités des colonnes obtenues est toujours égale à 1 pour la transposée de la matrice stochastique!!

L'analyse du régime transitoire (ou: promenade aléatoire) d'une chaîne de Markov consiste à déterminer (ou à imposer à!) la matrice-colonne (vecteur) $p(n)$ d'être dans un état donné à la n -ième étape de la promenade:

$$p(n) = \begin{pmatrix} p_1(n) \\ p_2(n) \\ \vdots \\ p_{Card(E)}(n) \end{pmatrix} \quad (6.1)$$

avec la somme des composantes qui vaut évidemment toujours 1 (car la somme des probabilités de se trouver dans un quelconque des sommets du graphe à un moment/étape donné(e) doit être égale à 100%).

Nous appelons fréquemment cette matrice-colonne "**vecteur stochastique**" ou "**mesure de probabilité sur le sommet i** ".

Démonstration:

Démontrons que la probabilité de ce vecteur stochastique est effectivement toujours unitaire.

Si $p(n)$ est un vecteur stochastique, alors son image:

$$p(n+1) = P^T \cdot p(n) \quad (6.2)$$

l'est aussi. Effectivement, $p_i(n+1) \geq 0$ car:

$$p_i(n+1) = \sum_j p_{ij} p_j(n) \quad (6.3)$$

est une somme de termes positifs ou nuls. De plus, nous trouvons:

$$\sum_i p_i(n+1) = \sum_i \sum_j p_{ij} p_j(n) = \sum_j \sum_i p_{ij} p_j(n) = \sum_j \left(\sum_i p_{ij} \right) p_j(n) = \sum_j p_j(n) = 1 \quad (6.4)$$

□C.Q.F.D.

Ce vecteur de probabilités, dont les composantes sont positives ou nulles, dépend (c'est assez intuitif) de la matrice de transition P et du vecteur de probabilités initiales $p(0)$.

Bien que cela soit démontrable (théorème de Perron-Frobenius) le lecteur pourra vérifier par un cas pratique (informatisé ou non!) que si nous choisissons un vecteur d'état $p(n)$ quelconque alors il existe pour toute matrice stochastique P un vecteur unique de probabilité noté traditionnellement π tel que:

$$P^T \cdot \pi = \pi \quad (6.5)$$

Une telle mesure de probabilité π vérifiant la relation précédente est appelée une "**mesure invariante**" ou "**mesure stationnaire**" ou encore "**mesure d'équilibre**" qui représente l'état d'équilibre du système. En termes d'algèbre linéaire (voir chapitre du même nom), pour la valeur propre 1, π est un vecteur propre de P (cf. chapitre d'Algèbre Linéaire).

Nous en verrons un exemple trivial dans le chapitre de Théorie des Graphes qui sera redéveloppé sous forme détaillée et complète ainsi que dans le chapitre de Théorie Des Jeux Et De La Décision dans le cadre de la pharmaco-économie. Mais signalons également que les chaînes de Markov sont également utilisées en météorologie par exemple ou encore dans le domaine de la casse de mots de passe informatiques:

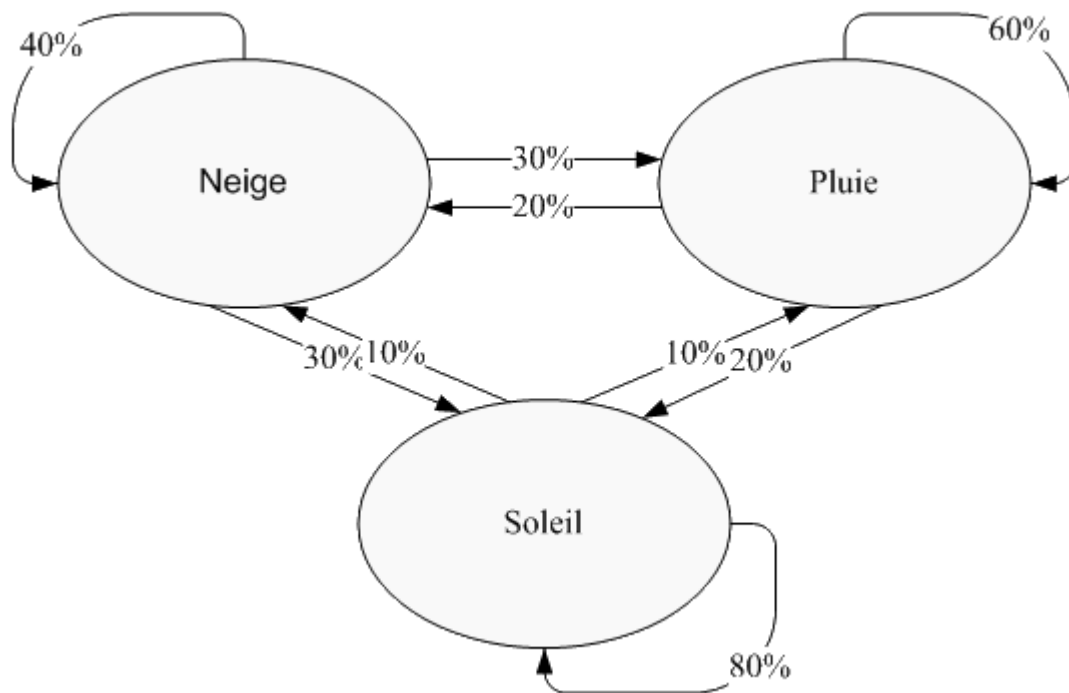


Figure: 6.1 - Exemple concret très simpliste d'une chaîne de Markov

ou dans le domaine médical, financier, des transports, du marketing, etc.

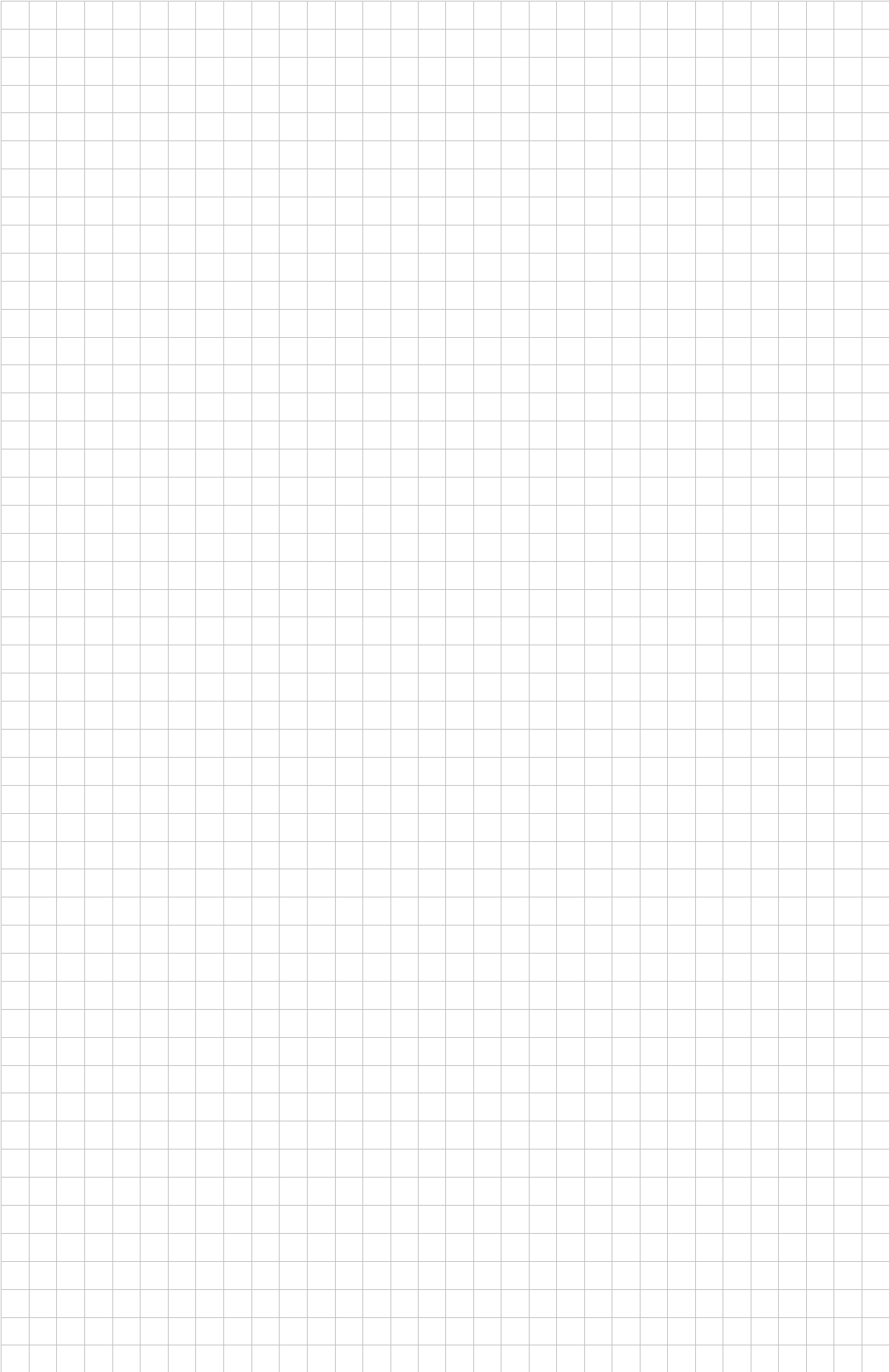
Dans le domaine du langage, à partir de l'analyse fréquentielle de séquence de mots, les ordinateurs arrivent à construire aussi des chaînes de Markov et donc à proposer une sémantique plus correcte lors de corrections grammaticales informatisées ou de transcription de écrite de présentations orales.

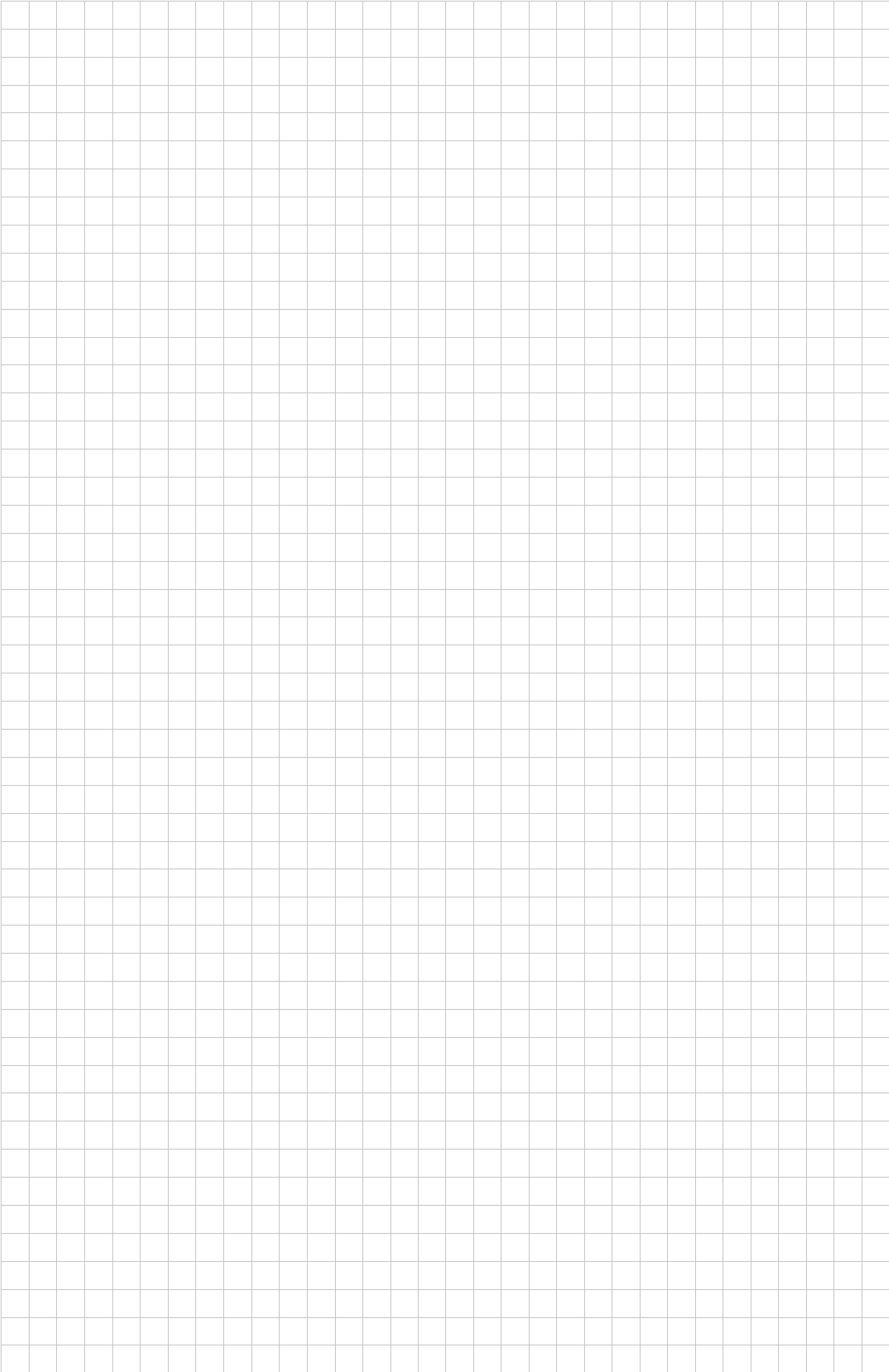
Enfin pour clore, donnons quelques définitions de vocabulaire complémentaires courantes que l'on retrouvera dans différents chapitres comme celui de Techniques de Gestion ou de Génie Industriel.

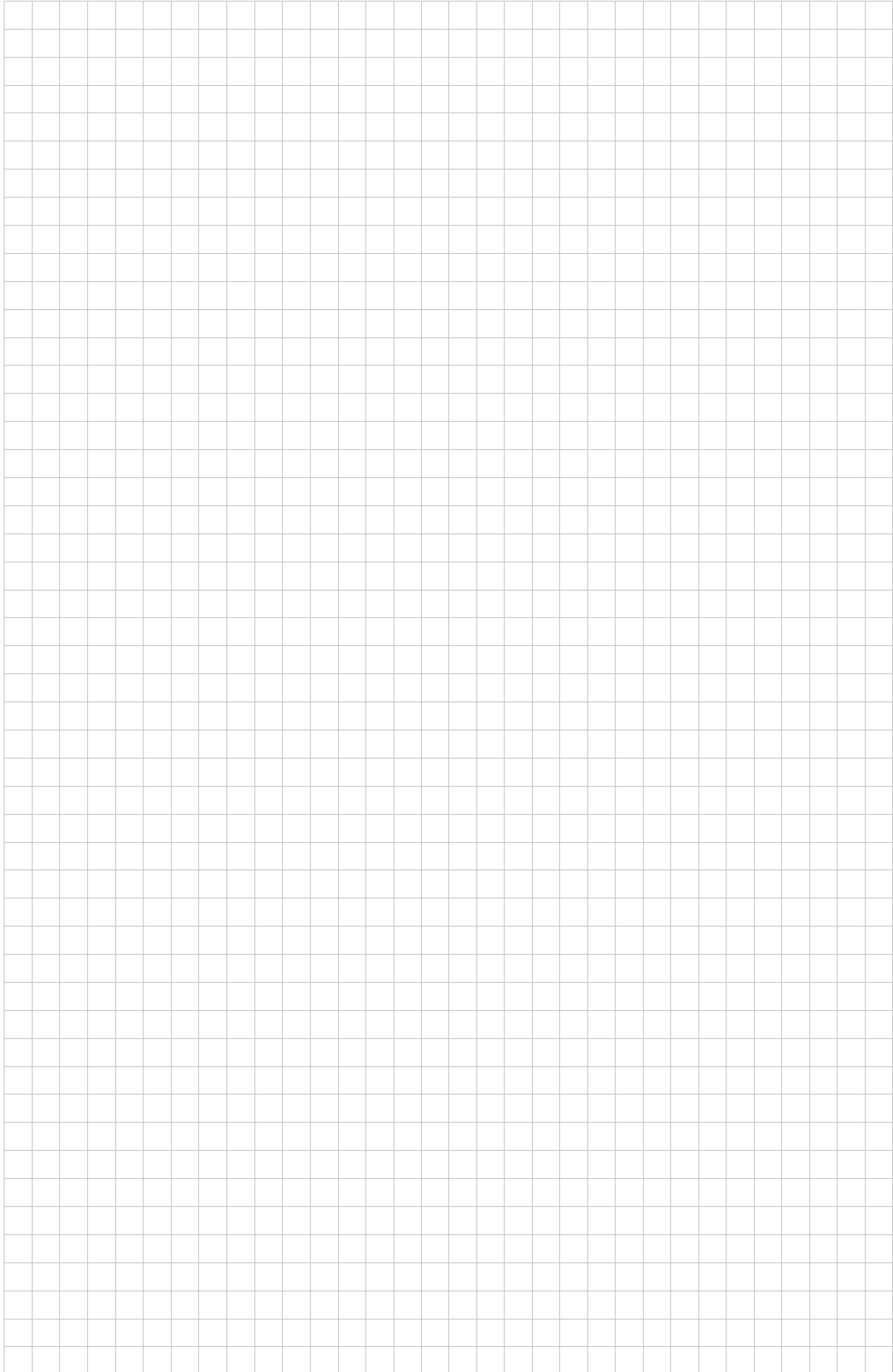
Définitions:

D1. Une chaîne de Markov est dite "**chaîne de Markov irréductible**" si tous les états sont liés aux autres (c'est le cas de la chaîne dans la figure ci-dessus).

D2. Une chaîne de Markov est dite "**chaîne de Markov absorbante**" si un quelconque des états de la chaîne absorbe les transitions (donc rien n'en sort pour dire simplement les choses!).







7. STATISTIQUES

La statistique est une science qui a pour objet le groupement méthodique de faits ou événements répétitifs qui se prêtent à une évaluation numérique ou qualitative dans le temps suivant une loi donnée. Dans l'industrie et dans l'économie en général, la statistique est une science qui permet dans un environnement incertain de faire des inférences valides.

Il faut savoir que parmi tous les domaines de la mathématique, celui qui est utilisé à la plus large échelle dans les entreprises et centres de recherches est bien la statistique et particulièrement depuis que des logiciels en facilitent grandement les calculs! Raison pour laquelle ce chapitre est un des plus gros du site internet alors que seuls les concepts élémentaires y sont présentés!

Signalons aussi que les statistiques ont très mauvaise réputation à l'université car les notations y sont souvent confuses et varient grandement d'un professeur à l'autre, d'un livre à l'autre, d'un praticien à l'autre. En toute rigueur, il faudrait se conformer au vocabulaire et notations de la norme ISO 3534-1:2006 et comme malheureusement ce chapitre a été écrit avant la publication de cette norme... un certain temps d'adaptation sera nécessaire avec qu'il y ait conformité.

Il est peut être inutile de préciser que la statistique est beaucoup utilisée en ingénierie, physique théorique, physique fondamentale, économétrie, gestion de projets ainsi que dans l'industrie des processus, dans les domaines des assurances vies et non vies, dans l'actuariat ou dans la simple analyse de banque de données (avec Microsoft Excel très souvent... malheureusement...) et la liste est encore longue. Par ailleurs, nous rencontrerons les outils présentés ici assez souvent dans les chapitres de Mécanique des Fluides, de Thermodynamique, des Techniques de Gestion, du Génie Industriel et d'Économétrie (en particulier dans ces deux dernières). Le lecteur pourra donc s'y reporter pour avoir des applications pratiques concrètes de quelques-uns des éléments théoriques les plus importants qui seront vus ici.

Signalons également que outre les quelques exemples simples donnés sur ces pages, de nombreux autres exemples applicatifs sont donnés sur le serveur d'exercices du site dans les catégories Probabilités et Statistiques, Génie Industriel, Économétrie et Techniques de Gestion.

Définition: Le but principal de la statistique est de déterminer les caractéristiques d'une population donnée à partir de l'étude d'une partie de cette population, appelée "**échantillon**" ou "**échantillon représentatif**". La détermination de ces caractéristiques doit permettre aux statistiques d'être un outil d'aide à la décision!

Remarque: Le traitement des données concerne la "**statistique descriptive**". L'interprétation des données à partir des estimateurs s'appelle "**l'inférence statistique**" (ou "**statistique inférentielle**"), et l'analyse de données en masse la "**statistique fréquentielle**" (en opposition à l'inférence bayésienne).

Lorsque nous observons un événement prenant en compte certains facteurs, il peut arriver qu'une deuxième observation ait lieu dans des conditions qui semblent identiques. En répétant ces mesures plusieurs fois sur différents objets supposés similaires, nous pouvons constater que les résultats observables sont distribués statistiquement autour d'une valeur moyenne qui est, finalement le résultat possible le plus probable. Dans la pratique, nous n'effectuons cependant parfois qu'une seule mesure et

il s'agit alors de déterminer la valeur de l'erreur que nous commettons en adoptant celle-ci comme moyenne mesurée. Cette détermination nécessite de connaître le type de distribution statistique auquel nous avons affaire et c'est ce que nous allons nous attarder (entre autres) à étudier ici (les bases du moins!). Il existe cependant plusieurs approches méthodologiques courantes (les moins courantes n'étant pas citées pour l'instant) face au hasard:

1. Une toute première consiste à ignorer purement et simplement les éléments aléatoires, pour la bonne raison que l'on ne sait pas comment les intégrer. Nous utilisons alors la "**méthode des scénarios**" appelée aussi "**simulation déterministe**". C'est typiquement un outil utilisé par les financiers ou gestionnaires non diplômés travaillant avec des outils comme Microsoft Excel (qui inclut un outil de gestion de scénarios) ou MS Project (qui inclut un outil du type scénarios déterministes du type optimiste, pessimiste, attendu).
2. Une seconde approche envisageable, quand nous ne savons pas associer des probabilités précises aux futurs événements aléatoires, est la théorie des jeux (cf. [chapitre de la Théorie Des Jeux Et De La Décision](#)) où l'on utilise des critères de sélection semi-empiriques comme le critère du maximax, du minimax, de Laplace, de Savage, etc.
3. Enfin, quand nous pouvons lier des probabilités aux événements aléatoires, soit que ces probabilités découlent de calculs ou de mesures, soit qu'elles reposent sur une expérience acquise auprès de situations antérieures de même nature que la situation actuelle, nous pouvons faire appel aux statistiques descriptives et inférentielles (contenu du présent chapitre) pour tirer des informations exploitables et pertinentes de cette masse de données acquises.
4. Une dernière approche quand nous avons connaissance de probabilités relatives aux issues intervenantes faisant suite à des choix stratégiques est l'utilisation de la théorie de la décision (cf. [chapitre de la Théorie Des Jeux Et De La Décision](#)).

Remarques:

R1. Sans la statistique mathématique, un calcul sur des données (par exemple une moyenne), n'est qu'un "**indicateur ponctuel**". C'est la statistique mathématique qui lui donne le statut d'estimateur dont on maîtrise le biais, l'incertitude et autres caractéristiques statistiques. Nous cherchons en général à ce que l'estimateur soit sans biais, convergeant et efficace (nous verrons lors de notre étude des estimateurs plus loin de quoi il s'agit exactement).

R2. Lorsque nous communiquons une statistique il devrait être obligatoire de préciser l'intervalle de confiance ainsi que la taille de l'échantillon étudié et ses caractéristiques détaillées sinon quoi elle n'a quasiment aucune valeur scientifique.

R3. Si vous avez un professeur ou un formateur qui ose vous enseigner les statistiques et probabilités avec des exemples basés sur des jeux de hasard (cartes, dés, allumette, pile ou face, etc.) débarrassez-vous en ou dénoncez-le à qui de droit car cela signifierait qu'il n'a aucune expérience pratique du domaine et qu'il va vous enseigner n'importe quoi et n'importe comment (normalement les exemples devraient être basés sur l'industrie, l'économie ou la R&D, bref dans des domaines utilisés tous les jours par les entreprises mais surtout pas sur des jeux de hasard...!).

Introduisons avant de continuer quelques définitions qui vont nous être utiles pour la suite sur le concept d'échantillons et de moyennes:

1. ÉCHANTILLONS

Lors de l'étude statistique d'ensembles d'informations, la façon de sélectionner l'échantillon est aussi importante que la manière de l'analyser. Il faut que l'échantillon soit représentatif de la population (nous ne faisons pas nécessairement référence à des populations humaines!). Pour cela, l'échantillonnage aléatoire est le meilleur moyen d'y parvenir.

Le statisticien part toujours de l'observation d'un ensemble fini d'éléments, que nous qualifions de "**population**". Les éléments observés, en nombre n , sont tous de même nature, mais cette nature peut être fort différente d'une population à l'autre.

Définitions:

D1. Nous sommes en présence d'un "**caractère quantitatif**" lorsque chaque élément observé fait explicitement l'objet d'une même mesure. A un caractère quantitatif donné, nous associons une "**variable quantitative**" continue ou discrète qui synthétise toutes les valeurs possibles que la mesure considérée est susceptible de prendre (ce type d'information étant représenté par des courbes de loi de Gauss-Laplace, de la loi bêta, de la loi de Poisson, etc.).

Remarque: Nous reviendrons sur le concept de "variable" en statistiques plus loin...

D2. Nous sommes en présence d'un "**caractère qualitatif**" lorsque chaque élément observé fait explicitement l'objet d'un rattachement unique à une "**modalité**" choisie dans un ensemble de modalités exclusives (de type: homme | femme) permettant de classer tous les éléments de l'ensemble étudié selon un certain point de vue (ce type d'information étant représenté par des diagrammes à barre, fromages, diagrammes à bulles, etc.). L'ensemble des modalités d'un caractère peut être établi a priori avant l'enquête (une liste, une nomenclature, un code) ou après enquête. Une population étudiée peut être représentée par un caractère mixte, ou ensemble de modalités tel que genre, tranche salariale, tranche d'âge, nombre d'enfants, situation matrimoniale par exemple pour un individu.

D3. Un "**échantillon aléatoire**" est un échantillon tiré au hasard dans lequel tous les individus d'une population ont la même chance, ou "**équiprobabilité**" (et nous insistons sur le fait que cette probabilité doit être égale), de se retrouver dans l'échantillon.

D4. Dans le cas contraire d'un échantillon dont les éléments n'ont pas été pris au hasard, nous disons alors que l'échantillon est "**biaisé**" (dans le cas inverse nous disons qu'il est "**non-biaisé**")

Remarque: Un petit échantillon représentatif est, de loin, préférable à un grand échantillon biaisé. Mais lorsque la taille des échantillons utilisés est petite, le hasard peut donner un résultat moins bon que celui qui est biaisé...

2. MOYENNES

La notion de "**moyenne**" ou "**tendance centrale**" (les financiers appellent cela aussi une "mesure de localisation"...) est avec la notion de "variable" à la base des statistiques.

Cette notion nous semble très familière et nous en parlons beaucoup sans nous poser trop de questions. Pourtant il existe divers qualificatifs (nous insistons sur le fait que ce ne sont que des qualificatifs!) pour distinguer la forme de la résolution d'un problème consistant à calculer la moyenne.

Il faut donc être très très prudent quant aux calculs de moyennes car il y a une fâcheuse tendance dans les entreprises à se précipiter et à utiliser systématiquement la moyenne arithmétique sans réfléchir, ce qui peut amener à de graves erreurs! Un exemple sympathique (pour faire un analogie) est qu'un nombre considérable de législations exigent seulement des seuils moyens de pollution par année alors que par exemple, fumer 1 cigarette par jour n'a pas le même impact que fumer 365 cigarette sur une année alors que les deux ont la même moyenne pris sur un an... C'est une preuve flagrante d'incompétence statistique du législateur.

Voici un petit échantillon d'erreurs courantes:

- Considérer que la moyenne arithmétique est la valeur qui coupe la population en deux parties égales.
- Considérer que la moyenne de ratios du type objectifs/réalisés est pas égale au ratio des moyennes des objectifs et des moyennées des réalisations (alors que ce n'est pas la même chose!)
- Considérer que la moyenne des salaires de différentes filiales est égale à la moyenne générale des salaires (alors que ceci n'est vrai que si et seulement si il y a le même nombre d'employés dans chaque filiale).
- Considérer que la moyenne de la moyenne des lignes d'un tableau est toujours égal à la moyenne des moyennes des colonnes (alors que ceci n'est vrai que si et seulement si le contenu des cellules est non vide).

Nous verrons ci-dessous différentes moyennes avec des exemples relatifs à l'arithmétique, au dénombrement, à la physique, à l'économétrie, à la géométrie et à la sociologie. Le lecteur trouvera d'autres exemples pratiques en parcourant l'ensemble du site.

Définitions: Soient des nombres x_i réels, nous avons alors:

D1. La "**moyenne arithmétique**" ou "**moyenne empirique**" (la plus communément connue) définie par le quotient de la somme des n valeurs observées x_i par l'effectif total n :

$$\mu_a = \frac{\sum_{i=1}^n x_i}{n} \quad (7.1)$$

et très souvent notée $\bar{x}$ ou encore $\hat{\mu}$ est pour toute loi statistique discrète ou continue un estimateur sans biais de l'espérance.

La moyenne arithmétique représente donc une mesure statistique exprimant la grandeur qu'aurait chacun des membres d'un ensemble de mesures si la somme doit être identique au produit de la moyenne par le nombre de termes.

Si plusieurs valeurs occurrent plus d'une fois dans les mesures, la moyenne arithmétique sera alors souvent notée formellement:

$$\mu_a = \frac{1}{n} \sum_{i=1}^r n_i x_i \quad (7.2)$$

et appelée "**moyenne pondérée par les effectifs**". Enfin, indiquons que dans le cadre de cette démarche, la moyenne pondérée par les effectifs prendra le nom "**d'espérance mathématique**" dans le domaine d'étude des probabilités.

Nous pouvons tout aussi bien utiliser les fréquences d'apparition des valeurs observées (dites "**fréquence des classes**"):

$$f_i = \frac{n_i}{n} \quad (7.3)$$

Nous avons alors la "**moyenne pondérée par les fréquences de classe**":

$$\mu_a = \frac{1}{n} \sum_{i=1}^r n_i x_i = \sum_{i=1}^r \frac{n_i}{n} x_i = \sum_{i=1}^r f_i x_i \quad (7.4)$$

Avant de continuer, indiquons que dans le domaine de la statistique il est souvent utile et nécessaire de regrouper les mesures/données dans des intervalles de classe de largeur donnée (voir les exemples plus loin). Il faut souvent faire plusieurs essais pour cela même s'il existe des formules semi-empiriques pour choisir le nombre de classes lorsque nous avons n valeurs à disposition. Une de ces règles semi-empiriques utilisée par de nombreux praticiens consiste à retenir le plus petit nombre entier de classes k tel que:

$$2^k \geq n \quad (7.5)$$

la largeur de l'intervalle de classe étant alors obtenue en divisant l'étendue (différence entre la valeur maximale mesurée et la minimale) par k . Par convention et en toute rigueur... (donc rarement respecté dans les notations), un intervalle de classe est fermé à gauche et ouvert à droite: [...,...[. Cette règle empirique se nomme la "**règle de Sturges**" et est basée sur le raisonnement suivant:

Nous admettons que les valeurs du coefficient binomial C_p^n donnent un histogramme idéal (nous laissons le lecteur vérifier cela simplement avec un tableau comme Microsoft Excel 11.8346 et la fonction COMBIN() qui y est disponible dans la version française). Au fur et à mesure que k devient grand l'histogramme ressemble de plus en plus à une courbe continue appelée "courbe Normale" que nous verrons plus loin.

Dès lors, en nous basant sur le théorème binomial (cf. chapitre de Calcul Algébrique), nous avons:

$$n = \sum_{p=0}^n C_p^n a^{n-p} b^p = \sum_{p=0}^n C_p^n 1^{n-p} 1^p = \sum_{p=0}^n C_p^n = (a+b)^n = (1+1)^n = 2^n \quad (7.6)$$

Ensuite, pour chaque intervalle i le praticien prendra par tradition la moyenne entre les deux bornes pour le calcul et la multipliera par la fréquence f_i de classe correspondante. Dès lors, le regroupement en fréquence de classes fait que:

1. La moyenne pondérée par les effectifs diffère de la moyenne arithmétique.

2. Vue l'approximation effectuée elle sera un moins bon indicateur que la moyenne arithmétique.

3. Elle est très sensible au choix du nombre de classes donc médiocre à ce niveau-là.

Plus loin, nous verrons deux propriétés extrêmement importantes de la moyenne arithmétique et de l'espérance mathématique qu'il vous faudra absolument comprendre (moyenne pondérée des écarts à la moyenne et la moyenne des écarts à la moyenne).

Remarque: Le "**mode**", noté *Mod* ou simplement *M*, est par définition la valeur qui apparaît le plus grand nombre de fois dans une série de valeurs. Dans Microsoft Excel 11.8346 (version française), soulignons que la fonction MODE() renvoie la première valeur dans l'ordre des valeurs ayant le plus grand nombre d'occurrences en supposant donc une distribution unimodale.

D2. La "**médiane**" ou "**moyenne milieu**", notée M_e (ou plus simplement *M*), est la valeur qui coupe une population en deux parties égales. Dans le cas d'une distribution statistique continue $f(x)$ d'une variable aléatoire X , il s'agit de la valeur qui représente 50% de probabilités cumulées d'avoir lieu tel que (nous détaillerons le concept de distribution statistique plus loin très en détails):

$$P(X \leq M_e) = P(X \geq M_e) = \int_{-\infty}^{M_e} f(x)dx = \int_{M_e}^{+\infty} f(x)dx = 0.5 \quad (7.7)$$

Dans le cas d'une série de valeurs ordonnées $x_1, x_2, \dots, x_i, \dots, x_n$, la médiane est donc de par sa définition la valeur de la variable telle que l'on ait autant d'éléments qui ont une valeur qui lui est supérieure ou égale, que d'éléments qui ont une valeur qui lui est inférieure ou égale. Elle est principalement utilisée pour les distributions asymétriques, car elle les représente mieux que la moyenne arithmétique.

Plus rigoureusement:

- Si le nombre de termes est impair, de la forme $2n+1$, la médiane de la série est le terme de rang $n+1$ (que les termes soient tous distincts ou non!).
- Si le nombre de termes est pair, de la forme $2n$, la médiane de la série est la demi-somme (moyenne arithmétique) des valeurs des termes de rang n et $n + 1$ (que les termes soient tous distincts ou non!).

Dans tous les cas, de par cette définition, il découle qu'il y a au moins 50 % des termes de la série inférieurs ou égaux à la médiane, et au moins 50% des termes de la série supérieurs ou égaux à la médiane.

Considérons par exemple la table de salaires ci-dessous:

N° Employé	Salaire	Cumul employés	%Cumul employés
1	1200	1	6%
2	1220	2	12%
3	1250	3	18%
4	1300	4	24%
5	1350	5	29%
6	1450	6	35%
7	1450	7	41%
8	1560	8	47%
9	1600	9	53%
10	1800	10	59%
11	1900	11	65%
12	2150	12	71%
13	2310	13	76%
14	2600	14	82%
15	3000	15	88%
16	3400	16	94%
17	4800	17	100%

Tableau: 7.1 - Identification de la médiane

Il y a un nombre impair $2n+1$ de valeurs. Donc la médiane de la série est le terme de rang $n+1$. Soit 1'600.- (résultat que vous donnera n'importe quel tableur informatique). La moyenne arithmétique quant à elle vaut 2'020.

En relation directe avec la médiane il est important de définir le concept suivant afin de comprendre le mécanisme sous-jacent:

Définition: Soit donnée une série statistique $x_1, x_2, \dots, x_i, \dots, x_n$, nous appelons "**dispersion des écarts absolus**" autour de x le nombre $\varepsilon'(x)$ défini par:

$$\varepsilon'(x) = \sum |x_i - x| \quad (7.8)$$

$\varepsilon'(x)$ est minimum pour une valeur de x la plus proche d'une valeur donnée x_i au sens de l'écart absolu. La médiane est la valeur qui réalise ce minimum (extrémum)! L'idée va alors consister à étudier les variations de la fonction pour trouver le rang de cet extrémum.

En effet, nous pouvons écrire:

$$\forall x \in [x_r, x_{r+1}], r \in \{1, 2, 3, \dots, n-1\} \quad \varepsilon'(x) = \sum_{i=1}^r |x_i - x| + \sum_{i=r+1}^n |x_i - x| \quad (7.9)$$

Donc par définition de la valeur x :

$$\begin{aligned}
\varepsilon'(x) &= \sum_{i=1}^r |x - x_i| + \sum_{i=r+1}^n |x_i - x| \\
&= [rx - (x_1 + x_2 + \dots + x_r)] + [(x_{r+1} + \dots + x_n) - (n-r)x] \quad (7.10) \\
&= (2r-n)x + (x_{r+1} + \dots + x_n) - (x_1 + x_2 + \dots + x_r)
\end{aligned}$$

Ce qui nous permet donc de faire sauter les valeurs absolues est simplement le choix de l'indice r qui est pris de telle manière que la série de valeurs peut en pratique toujours être coupée en deux parties: tout ce qui est inférieur à un élément de la série indexé par r et tout ce qui lui est supérieur (la médiane donc par anticipation).

$\varepsilon'(x)$ est donc une fonction affine (assimilable à l'équation d'une droite pour r et n fixés) par morceaux (discrète) où l'on peut assimiler le facteur:

$$2r-n \quad (7.11)$$

à la pente et:

$$(x_{r+1} + \dots + x_n) - (x_1 + x_2 + \dots + x_r) \quad (7.12)$$

à l'ordonnée à l'origine.

La fonction est donc décroissante (pente négative) tant que r est inférieur à $n/2$ et croissante quand r est supérieur à $n/2$. Plus précisément, nous distinguons deux cas qui nous intéressent particulièrement puisque n est un entier (elle pas donc par un extremum!):

- Si n est pair, nous pouvons poser $n = 2n'$, alors la pente peut s'écrire $2(r - n')$ et elle est nulle si $r = n'$ et dès lors puisque ce résultat n'est valable par construction que pour $\forall x \in [x_r, x_{r+1}]$ alors $\varepsilon'(x)$ est constante sur $[x_{n'}, x_{n'+1}]$ et nous avons un extrémum obligatoirement au milieu de cet intervalle (moyenne arithmétique des deux termes).

- Si n est impair, nous pouvons poser $n = 2n' + 1$ (nous coupons la série en deux parties égales), alors la pente peut s'écrire $(2r - 2n' - 1)$ et elle est donc nulle si $r = n' + 1/2$ et dès lors puisque ce résultat n'est valable que pour $\forall x \in [x_r, x_{r+1}]$ alors il est immédiat que la valeur du milieu sera la médiane $x_{n'+1}$.

Nous retrouvons donc bien la médiane dans les deux cas. Nous verrons aussi plus loin comment la médiane est définie pour une variable aléatoire continue.

Il existe un autre cas pratique où le statisticien n'a à sa disposition que des valeurs regroupées sous forme d'intervalles de classes statistiques. La procédure pour déterminer la médiane est alors différente:

Lorsque nous avons à notre disposition uniquement une variable classée, l'abscisse du point de la médiane se situe en général à l'intérieur d'une classe. Pour obtenir alors une valeur plus précise de la médiane, nous procédons à une interpolation linéaire. C'est ce que nous appelons la "**méthode d'interpolation linéaire de la médiane**".

La valeur de la médiane peut être lue sur le graphique ou calculée analytiquement. Effectivement, considérons le graphique représentant la probabilité cumulée $F(x)$ en intervalles de classe comme ci-dessous où les bornes des intervalles ont été reliées par des droites:

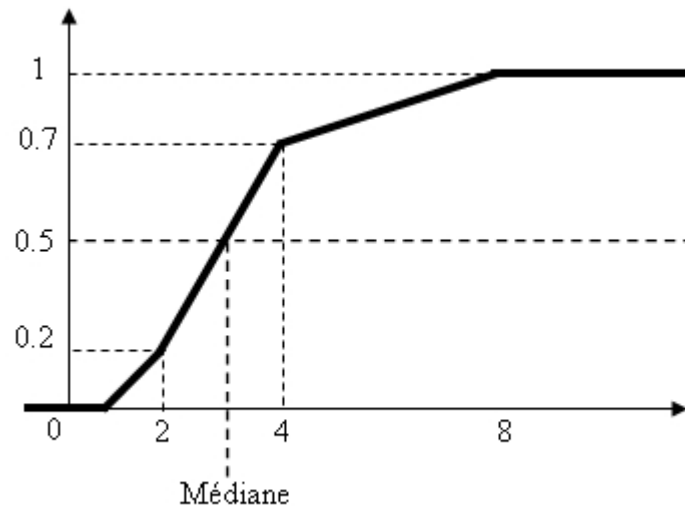


Figure: 7.1 - Représentation graphique de l'estimation par interpolation linéaire de la médiane

La valeur de la médiane M se trouve évidemment au croisement entre la probabilité de 50% (0.5) et l'abscisse. Si nous prenons dans le cadre particulier de l'exemple ci-dessus la borne supérieure de l'intervalle de classe précédant celle contenant la médiane nous avons 2 et 4 pour la borne inférieure de l'intervalle suivant. Nous avons alors en calculant la pente la relation suivante:

$$\frac{\Delta x}{\Delta y} = \frac{M - 2}{0.5 - 0.2} = \frac{4 - 2}{0.7 - 0.2} \quad (7.13)$$

Ce que nous écrivons fréquemment:

$$\frac{M_e - a}{0.5 - F(a)} = \frac{b - a}{F(b) - F(a)} \quad (7.14)$$

d'où la valeur de la médiane:

$$M_e = a + (b - a) \frac{0.5 - F(a)}{F(b) - F(a)} \quad (7.15)$$

Prenons le tableau suivant que nous retrouverons bien plus tard dans le présent chapitre:

Montant des tickets	Nombre de tickets	Nombre cumulés de tickets	Fréquences relatives cumulées
[0,50[	668	668	0.068
[50,100[	919	1'587	0.1587
[100,150[	1'498	3'085	0.3085
[150,200[	1'915	5'000	0.5000
[200,250[	1'915	6'915	0.6915
[250,300[	1'498	8'413	0.8413
[300,350[	919	9'332	0.9332
[350,400[	440	9'772	0.9772
[400 et +	228	10'000	1

Tableau: 7.2 - Identification de la classe médiane et du mode

Nous voyons que la "**classe médiane**" est dans l'intervalle [150,200] car la valeur cumulée de 0.5 s'y trouve (colonne toute à droite) mais la médiane a elle, en utilisant la relation établie précédemment, précisément une valeur de (c'est trivial dans l'exemple particulier du tableau ci-dessus mais faisons quand même le calcul...):

$$M_e = 150 + (200 - 150) \cdot \frac{0.5 - 0.3085}{0.5 - 0.3085} = 200 \quad (7.16)$$

et nous pouvons faire de même avec n'importe quel autre centile bien évidemment!

Nous pouvons également donner une définition pour déterminer la valeur modale si nous sommes seulement en possession des fréquences des classes d'intervalles. Pour cela partons du diagramme en barre des fréquences simplifié ci-dessous:

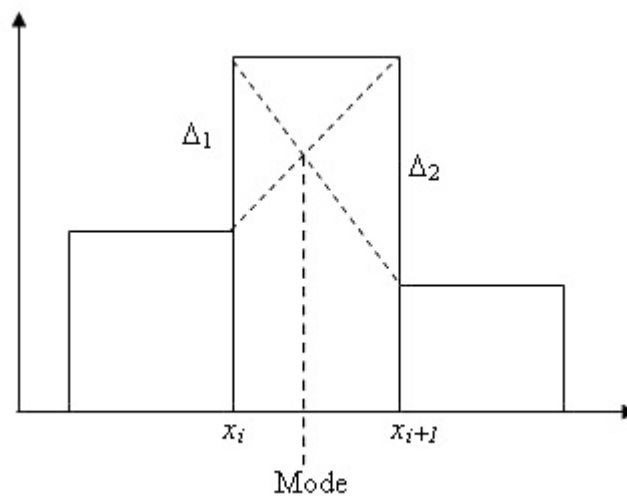


Figure: 7.2 - Représentation graphique de l'estimation par classes d'intervalles de la valeur modale

En utilisant les relations de Thalès (cf. chapitre de Géométrie Euclidienne), nous avons immédiatement, en notant M la valeur modale:

$$\frac{M - x_i}{\Delta_1} = \frac{x_{i+1} - M}{\Delta_2} \quad (7.17)$$

Comme dans une proportion, on ne change pas la valeur du rapport en additionnant les numérateurs et en additionnant les dénominateurs, il vient:

$$\frac{M - x_i}{\Delta_1} = \frac{x_{i+1} - M}{\Delta_2} = \frac{x_{i+1} - x_i}{\Delta_1 + \Delta_2} \quad (7.18)$$

Nous avons alors:

$$M = x_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} (x_{i+1} - x_i) \quad (7.19)$$

Avec l'exemple précédent cela donne alors:

$$\begin{aligned}
 M &= 150 + \frac{(1'915 - 1'498)}{(1'915 - 1'498) + (1'915 - 1'498)} (250 - 150) \\
 &= 150 + \frac{1}{2} (250 - 150) = 150 + \frac{1}{2} 100 = 200
 \end{aligned}
 \tag{7.20}$$

La question qui se pose ensuite est celle de la pertinence du choix de la moyenne, du mode ou de la médiane en termes de communication...

Un bon exemple reste celui du marché du travail où de façon générale, alors que le salaire moyen et le salaire médian sont relativement différents, les institutions de statistiques étatiques calculent la médiane que beaucoup de médias traditionnels assimilent alors explicitement au concept de "moyenne arithmétique" dans leurs communiqués.

Remarque: Pour éviter d'obtenir une moyenne arithmétique ayant peu de sens, nous calculons souvent une "**moyenne élaguée**", c'est à dire une moyenne arithmétique calculée après avoir enlevé des valeurs aberrantes à la série.

Les "**quantiles**" généralisent la notion de médiane en coupant la distribution en des ensembles donnés de parties égales (de même cardinal pourrions-nous dire...) ou autrement dit en intervalles réguliers. Nous définissons ainsi les "**quartiles**", les "**déciles**" et les "**centiles**" (ou "**percentiles**") sur la population, ordonnée dans l'ordre croissant, que nous divisons en 4, 10 ou 100 parties de même effectif.

Nous parlerons ainsi du centile 90 pour indiquer la valeur séparant les premiers 90% de la population des 10% restants.

Précisons que dans la version francophone de Microsoft Excel 11.8346 les fonctions QUARTILE(), CENTILE(), MEDIANE(), RANG.POURCENTAGE () sont disponibles et spécifions qu'il existe plusieurs variantes de calcul de ces centiles d'où une variation possible entre les résultats sur différents logiciels.

Ce concept est très important dans le cadre des intervalles de confiance que nous verrons beaucoup plus loin dans ce chapitre et très utile dans le domaine de la qualité avec l'utilisation des boîtes à moustaches (traduction de Box & Whiskers Plot ou BoxPlot) permettant de comparer ("discriminer" comme disent les spécialistes) rapidement deux populations de données ou plus et surtout d'éliminer les valeurs aberrantes (prendre comme référence la médiane sera justement plus judicieux!):

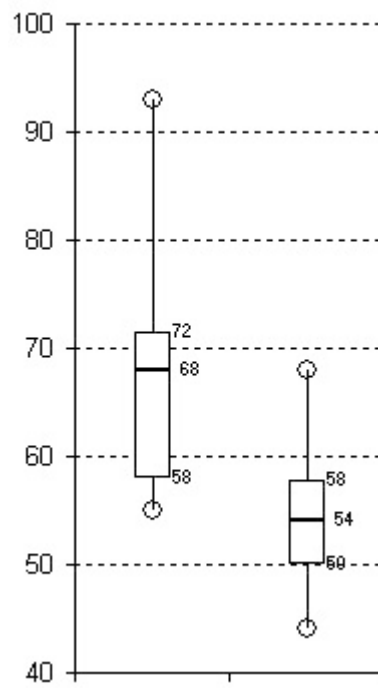
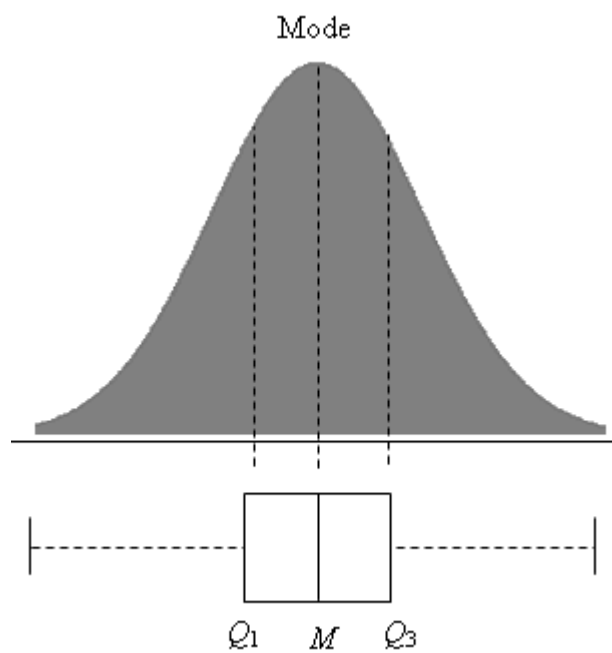


Figure: 7.3 - Box & Whiskers Plot

Une autre représentation mentale très importante des boîtes à moustache est la suivante (elle permet donc de se donner une idée de l'asymétrie de la distribution):



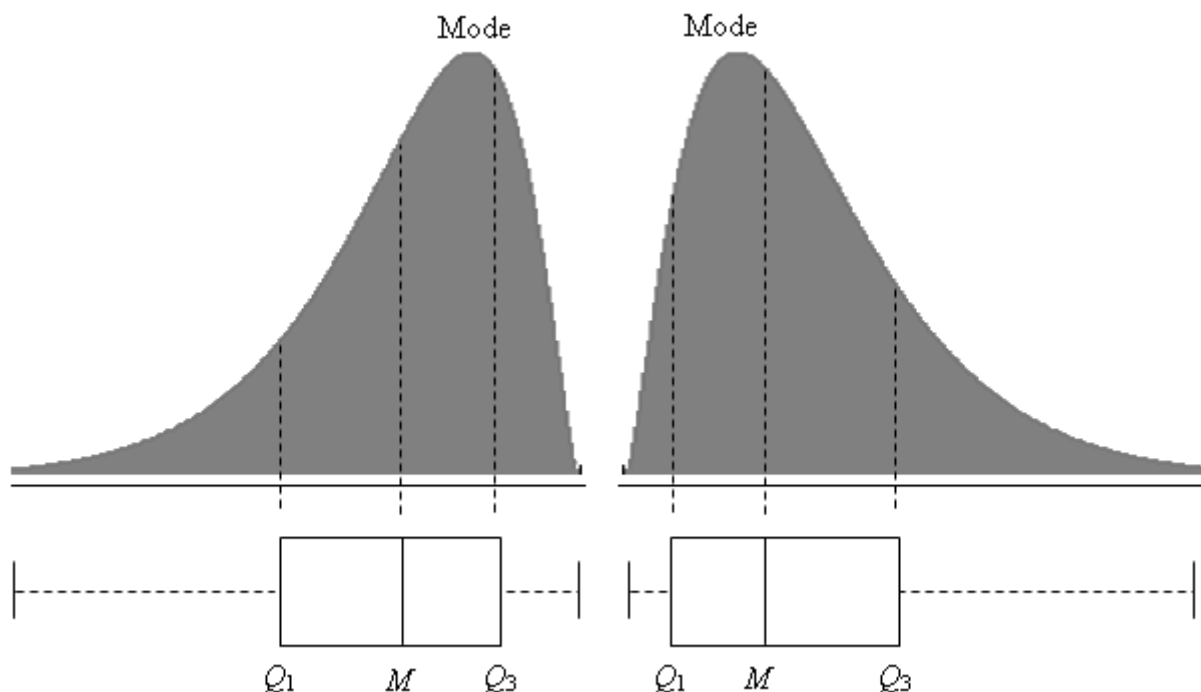


Figure: 7.4 - Représentation graphique du mode, de la médiane et des quartiles par rapport à une distribution

Les notions de médiane, valeurs aberrantes et intervalles de confiance que nous venons de démontrer et/ou de citer sont à ce point importantes qu'il existe des normes internationales pour les utiliser correctement. Citons d'abord la norme ISO 16269-7:2001 *Médiane - Estimation et intervalles de confiance* et aussi la norme ISO 16269-4:2010 *Détection et traitement des valeurs aberrantes*.

D3. Par analogie avec la médiane, nous définissons la "**médiale**" comme étant la valeur (dans l'ordre croissant des valeurs) qui partage la somme (cumuls) des valeurs en deux masses égales (donc la somme totale divisée par deux).

Dans le cas de salaires, alors que le médiane donne le 50% des salaires se trouvant en-dessous et en-dessus, la médiale donne combien de salariés se partagent (et donc le salaire partageant) la première moitié et combien de salariés se partagent la seconde moitié de l'ensemble des coûts salariaux.

Par exemple pour revenir à notre tableau sur les salaires:

N° Employé	Salaire	Cumul salaire	%Cumulé salaire
1	1200	1200	3.5%
2	1220	2420	7%
3	1250	3670	10.7%
4	1300	4970	14.5%
5	1350	6320	18.4%
6	1450	7770	22.6%
7	1450	9220	26.8%
8	1560	10780	31.4%
9	1600	12380	36.1%
10	1800	14180	41.3%
11	1900	16080	46.8%
12	2150	18230	53.1%
13	2310	20540	59.8%
14	2600	23140	67.4%
15	3000	26140	76.1%
16	3400	29540	86%
17	4800	34340	100%

Tableau: 7.3 - Identification de la médiale

La somme de tous les salaires fait donc 34'340 et la médiale est alors 17'170 (entre l'employé n°11 et 12) alors que la médiane était de 1'600. Nous voyons alors que la médiale correspond au 50% du cumul. Ce qui est un indicateur très utile dans le cadre des analyses de Pareto ou de Lorenz par exemple (cf. chapitre de Technique de Gestion).

D4. La "**moyenne quadratique**" parfois simplement notée Q qui est définie par:

$$\mu_q = \sqrt[m]{\frac{\sum_{i=1}^n x_i^m}{n}} \quad (7.21)$$

avec $m=2$.

Remarque: C'est une des moyennes les plus connues en statistiques car l'écart-type est une moyenne quadratique (voir plus loin).

Exemple:

Soit un carré de côté a , et un autre carré de côté b . La moyenne des aires des deux carrés est égale à un carré de côté:

$$\mu_q^2 = \frac{a^2 + b^2}{2} \Rightarrow \mu_q = \sqrt{\frac{a^2 + b^2}{2}} \quad (7.22)$$

D5. La "**moyenne harmonique**" parfois simplement notée H est définie par:

$$\frac{1}{\mu_h} = \frac{\sum_{i=1}^n x_i^{-1}}{n} \quad (7.23)$$

Elle est peu connue mais découle souvent de raisonnements simples et pertinents (typiquement la résistance équivalente d'un circuit électrique ayant plusieurs résistances en parallèles). Il existe une fonction MOYENNE.HARMONIQUE() dans Microsoft Excel 11.8346 (version française) pour la calculer.

Exemple:

Soit une distance d parcourue dans un sens à la vitesse v_1 et dans l'autre (ou pas) à la vitesse v_2 . La vitesse moyenne s'obtiendra en divisant la distance totale $2d$ par le temps mis à la parcourir:

$$v = \frac{2d}{t} \quad (7.24)$$

Si nous calculons le temps mis lorsqu'on parcourt d avec une vitesse v_i c'est tout simplement le quotient:

$$t_i = \frac{d}{v_i} \quad (7.25)$$

Le temps total vaut donc:

$$t = \frac{2d}{v} = t_1 + t_2 = \frac{d}{v_1} + \frac{d}{v_2} \quad (7.26)$$

La vitesse moyenne (son inverse pour être exact) sera donc bien du type harmonique:

$$\frac{1}{v} = \frac{1}{2d} \left(\frac{d}{v_1} + \frac{d}{v_2} \right) = \frac{1}{2} \left(\frac{1}{v_1} + \frac{1}{v_2} \right) \quad (7.27)$$

D6. La "**moyenne géométrique**" parfois notée simplement G est définie par:

$$\mu_g = \sqrt[n]{\prod_{i=1}^n x_i} \quad (7.28)$$

Cette moyenne est souvent oubliée mais néanmoins très connue dans le domaine de l'économétrie (surtout quand nous étudierons le rendement géométrique moyen) et de la finance d'entreprise (cf. chapitre Techniques De Gestion) raison pour laquelle il existe une fonction MOYENNE.GEOMETRIQUE() dans Microsoft Excel 11.8346 (version française) pour la calculer.

Exemple:

Supposons qu'une banque offre une possibilité de placement et prévoit pour la première année un

intérêt (c'est absurde mais c'est un exemple) avec un taux $(X - Y)\%$, mais pour la deuxième année un intérêt avec un taux $(X + Y)\%$. Au même moment une autre banque offre un intérêt à taux constant pour deux ans: $X\%$. C'est pareil, dirons-nous un peu rapidement. En fait les deux placements n'ont pas la même rentabilité.

Dans la première banque, un capital C_0 donnera au bout de la première année un intérêt:

$$(X - Y)\% \cdot C_0 \quad (7.29)$$

et la seconde année:

$$(X + Y)\% [(X - Y)\% \cdot C_0] \quad (7.30)$$

Dans l'autre banque nous aurons au bout d'un an:

$$X\% \cdot C_0 \quad (7.31)$$

et après la seconde année:

$$X\% (X\% C_0) \quad (7.32)$$

etc...

Comme vous pouvez le voir le placement ne sera pas identique si $Y \neq 0$! $X\%$ n'est donc pas la moyenne de $(X - Y)\%$ et $(X + Y)\%$.

Posons maintenant:

$$r_1 = (X + Y)\% \quad \text{et} \quad r_2 = (X - Y)\% \quad (7.33)$$

Quelle est en fait la valeur moyenne r ?

Au bout de deux ans le capital est multiplié par $r_1 \cdot r_2$. Si la moyenne vaut r il sera alors multiplié par r^2 . Nous avons donc la relation:

$$r^2 = r_1 \cdot r_2 \Leftrightarrow r = \sqrt{r_1 \cdot r_2} \quad (7.34)$$

C'est un exemple d'application où nous retrouvons donc la moyenne géométrique. L'oubli de la moyenne harmonique une erreur fréquente dans les entreprises lorsque certains employés calculent le taux moyen d'augmentation d'une valeur de référence.

D7. La "**moyenne mobile**", appelée aussi "**moyenne glissante**" est définie par:

$$MM_n = \frac{M_1 + M_2 + \dots + M_n}{n} \quad (7.35)$$

La moyenne mobile est particulièrement utilisée en économie, où elle permet de représenter une courbe de tendance d'une série de valeurs, dont le nombre de points est égal au nombre total de points de la série de valeurs moins le nombre que vous spécifiez pour la période.

Une Moyenne Mobile (*MM*) en finance est calculée à partir des moyennes des cours d'une valeur, sur une période donnée: chaque point d'une moyenne mobile sur 100 séances est la moyenne des 100 derniers cours de la valeur considérée. Cette courbe, affichée simultanément avec la courbe d'évolution des cours de la valeur, permet de lisser les variations journalières de la valeur, et de dégager des tendances.

Les moyennes mobiles peuvent être calculées sur différentes périodes, ce qui permet de dégager des tendances à court terme *MMC* (20 séances selon les habitudes de la branche), moyen terme (50-100 séances) ou long terme *MML* (plus de 100 séances).

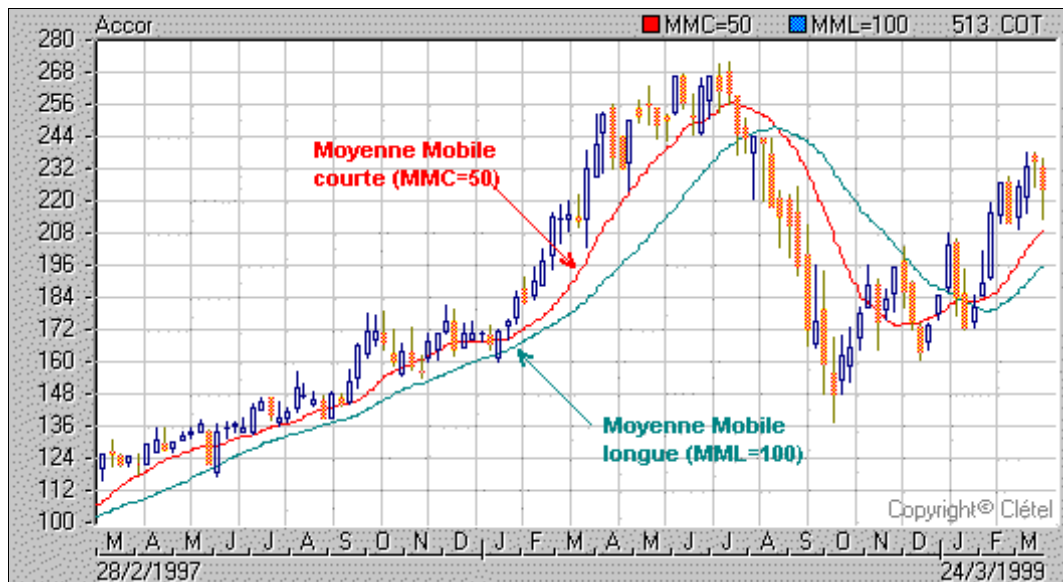


Figure: 7.5 - Représentation graphique des quelques moyennes mobiles pour 100 séances de négociation

Les croisements des moyennes mobiles par la courbe des cours (découpée avec une certaine granularité) de la valeur génèrent des signaux d'achat ou de vente (selon les professionnels) suivant le cas:

- Signal d'achat: lorsque la courbe des cours franchit la *MM*.
- Signal de vente: lorsque la courbe des cours franchit la *MM* vers le bas.

Outre la moyenne mobile, précisons qu'il existe une quantité d'autres indicateurs artificiels souvent utilisés en finance comme par exemple le "**upside/downside ratio**".

L'idée est la suivante: Si vous avez un produit financier (cf. chapitre d'Économie) actuellement de prix P_c (prix courant) pour lequel vous avez un objectif de gain haut à un prix haut correspondant que nous noterons P_h (high price) et inversement le potentiel de perte que vous estimez à un prix P_l (low price).

Alors, le rapport:

$$\frac{P_h - P_c}{P_c - P_l} = UD_R \quad (7.36)$$

donne le Upside/Downside Ratio.

Par exemple, un produit financier de 10.- avec un prix bas de 5.- et un prix haut de 15.- a donc un ratio

$UD_R = 1$ et donc un facteur spéculatif identique pour permettre le gain ou une perte de 5.-.

Un produit financier de 10.- avec un prix bas de 5.- et un prix haut de 20.- a donc un $UD_R = 2$ donc deux fois le potentiel spéculatif de gain par rapport à celui de perte.

Certaines associations boursières recommandent de refuser les UD_R inférieurs à 3. Les investisseurs ont tendance à rejeter les UD_R trop élevés pouvant être un signe de gonflage artificiel.

D8. La "**moyenne pondérée**" (dont nous avons déjà fait mention plus haut d'un cas particulier) est définie par:

$$\mu_p = \frac{\sum_{i=1}^n p_i x_i}{\sum_{i=1}^n p_i} \quad (7.37)$$

et est utilisée par exemple en géométrie pour localiser le barycentre d'un polygone, en physique pour déterminer le centre de gravité ou en statistiques pour calculer une espérance (le dénominateur étant toujours égal à l'unité en probabilités) et en gestion de projets pour estimer les durées des tâches.

Dans le cas général le poids p_i représente l'influence pondérée ou arbitraire/empirique de l'élément x_i par rapport aux autres.

D9. La "**moyenne fonctionnelle**" ou "**moyenne intégrale**" est définie par:

$$\mu_f = \frac{1}{b-a} \int_a^b f(x) dx \quad (7.38)$$

où μ_f dépend d'une fonction f d'une variable réelle intégrable (cf. chapitre de Calcul Différentiel Et Intégral) sur un intervalle $[a,b]$. Elle est très souvent utilisée en théorie du signal (électronique, électrotechnique).

2.1. LISSAGE DE LAPLACE

Pour en revenir à nos fréquences de classes vues bien plus haut et avant de continuer avec l'étude de quelques propriétés mathématiques des moyennes... il faut savoir que lorsque nous travaillons avec des lois discrètes de probabilités il arrive très (très) fréquemment que nous rencontrions un problème typique dont la source est la taille de la population. Considérons comme exemple le cas où nous avons 12 documents et que souhaiterions estimer la probabilité d'occurrence du mot "Viagra". Nous avons sur un échantillon les valeurs suivantes:

Identifiant document	Occurrences du mot
1	1
2	0
3	2
4	0
5	4
6	6
7	3
8	0
9	6
10	2
11	0
12	1

Tableau: 7.4 - Fréquences de classe du mot

Tableau que nous pouvons représenter d'une autre manière:

Occurrences du mot	Documents	Probabilité
0	4	0.33
1	2	0.17
2	2	0.17
3	1	0.083
4	1	0.083
5	0	0
6	2	0.17

Tableau: 7.5 - Fréquences de classe respective des documents

Et ici nous avons un phénomène courant. Il n'y a aucun document avec 5 occurrences du mot qui nous intéresse. L'idée (très courante dans le domaine du Data Mining) est alors d'ajouter artificiellement et empiriquement un comptage en utilisant une technique appelée "**lissage de Laplace**" qui consiste à additionner k unités à chaque occurrence et qui est courant dans le domaine du Data Mining. Dès lors le tableau devient:

Occurrences du mot	Documents	Probabilité
0	5	0.26
1	3	0.16
2	3	0.16
3	2	0.11
4	2	0.11
5	1	0.05
6	3	0.16

Tableau: 7.6 - Fréquence de classe des documents avec lissage

Évidemment ce type de technique est sujet à débat et sort du cadre scientifique... Nous avons même hésité à présenter cette technique dans le chapitre de Méthodes Numériques (avec le reste de toutes les techniques numériques empiriques)...

2.2. PROPRIÉTÉS DES MOYENNES

Voyons maintenant quelques propriétés pertinentes qui relient quelques-unes de ces moyennes ou qui sont propres à une moyenne donnée.

Les premières propriétés sont importantes donc prenez garde à bien les comprendre:

P1. Le calcul des moyennes arithmétique, quadratique et harmonique peut être généralisé à l'aide de la relation suivante:

$$\mu = \sqrt[m]{\frac{1}{n} \sum_{i=1}^n x_i^m} \quad (7.39)$$

où nous retrouvons:

1. Pour $m = 1$, la moyenne arithmétique
2. Pour $m = 2$, la moyenne quadratique
3. Pour $m = -1$, la moyenne harmonique

P2. La moyenne arithmétique a une propriété de linéarité, c'est-à-dire que (sans démonstration car simple à vérifier):

$$\overline{\lambda x + \mu} = \lambda \bar{x} + \mu \quad (7.40)$$

C'est la version statistique de la propriété de l'espérance en probabilité que nous verrons plus loin.

P3. La somme pondérée des écarts à la moyenne arithmétique est nulle.

Démonstration:

D'abord, par définition, nous savons que:

$$n = \sum_{i=1}^r n_i \quad \text{et} \quad \mu = \frac{1}{n} \sum_{i=1}^r n_i x_i \quad (7.41)$$

il s'ensuit que:

$$\sum_{i=1}^r n_i (x_i - \mu) = \sum_{i=1}^r n_i x_i - \mu \sum_{i=1}^r n_i = \sum_{i=1}^r n_i x_i - \left(\frac{1}{n} \sum_{i=1}^r n_i x_i \right) n = \sum_{i=1}^r n_i x_i - \sum_{i=1}^r n_i x_i = 0 \quad (7.42)$$

Ainsi, cet outil ne peut être utilisé comme mesure de dispersion!

Par extension la moyenne des écarts à la moyenne pondérée par les effectifs est nulle aussi:

$$\frac{\sum_{i=1}^r n_i (x_i - \mu)}{n} = 0 \quad (7.43)$$

Ce résultat est relativement important car il permettra plus loin de mieux saisir le concept d'écart-type et de variance.

P4. Soit à démontrer:

$$\mu_k < \mu_g < \mu_a < \mu_q \quad (7.44)$$

Démonstration:

Tout d'abord, nous prenons deux nombres réels non nuls x_1 et x_2 tels que $x_2 > x_1 > 0$ et nous écrivons:

1. La moyenne arithmétique:

$$\mu_a = \frac{x_1 + x_2}{2} \quad (7.45)$$

2. La moyenne géométrique:

$$\mu_g = \sqrt{x_1 x_2} \quad (7.46)$$

3. La moyenne harmonique:

$$\frac{1}{\mu_h} = \frac{\frac{1}{x_1} + \frac{1}{x_2}}{2} = \frac{\frac{x_2}{x_1 x_2} + \frac{x_1}{x_1 x_2}}{2} = \frac{x_1 + x_2}{2 x_1 x_2} \Leftrightarrow \mu_h = \frac{2 x_1 x_2}{x_1 + x_2} \quad (7.47)$$

4. La moyenne quadratique:

$$\mu_q^2 = \frac{x_1^2 + x_2^2}{2} \quad (7.48)$$

Remarque: Les comparaisons entre les moyennes précitées et la médiane ou encore les moyennes glissantes et pondérées n'ont pas de sens c'est pour cela que nous nous abstenons à les faire.

Prouvons déjà que $\mu_g > \mu_k$ par l'absurde en posant $\mu_g - \mu_k < 0$:

$$\begin{aligned}
\mu_g - \mu_h &= \sqrt{x_1 x_2} - \frac{2x_1 x_2}{x_1 + x_2} = \frac{\sqrt{x_1 x_2} x_1 + \sqrt{x_1 x_2} x_2 - 2x_1 x_2}{x_1 + x_2} < 0 \\
&\Rightarrow \sqrt{x_1 x_2} x_1 + \sqrt{x_1 x_2} x_2 - 2x_1 x_2 < 0 \\
&\Rightarrow \sqrt{x_1 x_2} x_1 + \sqrt{x_1 x_2} x_2 < 2x_1 x_2 \\
&\Rightarrow \frac{\sqrt{x_1 x_2}}{x_2} + \frac{\sqrt{x_1 x_2}}{x_1} < 2 \\
&\Rightarrow \sqrt{\frac{x_1}{x_2}} + \sqrt{\frac{x_2}{x_1}} < 2
\end{aligned} \tag{7.49}$$

Par commodité posons $y = \sqrt{\frac{x_2}{x_1}}$ nous savons que $y > 1$. Or:

$$\sqrt{\frac{x_1}{x_2}} + \sqrt{\frac{x_2}{x_1}} = \frac{1}{y} + y \tag{7.50}$$

et nous cherchons à montrer que $\sqrt{\frac{x_1}{x_2}} + \sqrt{\frac{x_2}{x_1}} < 2$ n'est pas possible. Mais ceci découle des équivalences suivantes:

$$\sqrt{\frac{x_1}{x_2}} + \sqrt{\frac{x_2}{x_1}} = \frac{1}{y} + y = \frac{y^2 + 1}{y} < 2 \Leftrightarrow y^2 + 1 - 2y < 0 \Leftrightarrow (y - 1)^2 < 0 \tag{7.51}$$

Il y a donc contradiction ce qui vérifie notre hypothèse initiale:

$$\mu_g - \mu_h > 0 \Leftrightarrow \mu_g > \mu_h \tag{7.52}$$

Regardons maintenant si $\mu_a > \mu_g$:

Sous l'hypothèse $x_2 > x_1 > 0$. Nous cherchons donc maintenant à montrer que:

$$\frac{x_1 + x_2}{2} > \sqrt{x_1 \cdot x_2} \tag{7.53}$$

Or nous avons les équivalences suivantes:

$$\frac{x_1 + x_2}{2} > \sqrt{x_1 \cdot x_2} \Leftrightarrow (x_1 + x_2)^2 > 4x_1 \cdot x_2 \Leftrightarrow x_1^2 + x_2^2 - 2x_1 \cdot x_2 > 0 \Leftrightarrow (x_1 - x_2)^2 > 0 \tag{7.54}$$

et la dernière expression est évidemment correcte.

Or le carré d'un nombre est toujours positif ce qui vérifie notre hypothèse initiale:

$$\mu_a - \mu_g > 0 \Leftrightarrow \mu_a > \mu_g \tag{7.55}$$

Nous prouvons maintenant $\mu_g > \mu_a$ et démontrons-le par l'absurde en posant $\mu_g - \mu_a < 0$:

$$\begin{aligned}
 \mu_g - \mu_a &= \sqrt{\frac{x_1^2 + x_2^2}{2}} - \frac{x_1 + x_2}{2} \\
 \Rightarrow \sqrt{\frac{x_1^2 + x_2^2}{2}} - \frac{x_1 + x_2}{2} < 0 &\Rightarrow \sqrt{\frac{x_1^2 + x_2^2}{2}} < \frac{x_1 + x_2}{2} \Rightarrow \frac{x_1^2 + x_2^2}{2} < \frac{x_1^2 + 2x_1x_2 + x_2^2}{4} \quad (7.56) \\
 \Rightarrow x_1^2 + x_2^2 < \frac{x_1^2 + 2x_1x_2 + x_2^2}{2} &\Rightarrow x_1^2 - 2x_1x_2 + x_2^2 < 0 \Rightarrow (x_1 - x_2)^2 < 0
 \end{aligned}$$

Or le carré d'un nombre est toujours positif ce qui vérifie notre hypothèse initiale:

$$\mu_g - \mu_a > 0 \Leftrightarrow \mu_g > \mu_a \quad (7.57)$$

Nous avons donc bien:

$$\mu_h < \mu_g < \mu_a < \mu_q \quad (7.58)$$

□C.Q.F.D.

Ces inégalités démontrées, nous pouvons alors passer à une figure que nous attribuons à Archimède pour placer trois de ces moyennes. L'intérêt de cet exemple est de montrer qu'il existe des relations remarquables parfois entre la statistique et la géométrie (fruit du hasard ???).

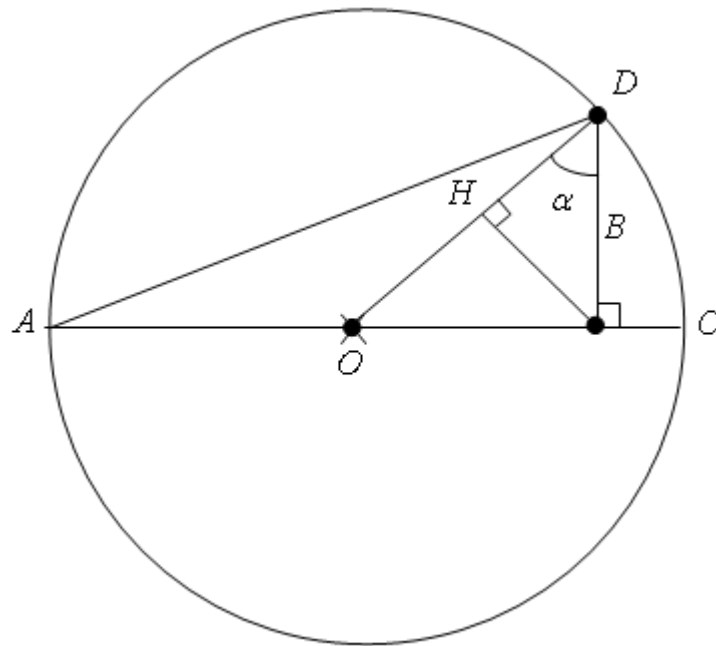


Figure: 7.6 - Point de départ pour la représentation géométrique des moyennes

Nous allons d'abord poser $a = \overline{AB}$, $b = \overline{BC}$ et O est le milieu de $\overline{AC}$. Ainsi, le cercle dessiné Ω est de centre O et de rayon $\overline{OA}$. D est l'intersection de la perpendiculaire à $\overline{AC}$ passant par B et du cercle Ω (nous choisissons l'intersection que nous voulons). H est quant à lui le projeté orthogonal de B sur $\overline{OD}$.

Archimède affirme que $\overline{OA}$ est la moyenne arithmétique de a et b et que $\overline{BD}$ est la moyenne géométrique de a et b , et $\overline{OH}$ la moyenne harmonique de a et b .

Nous démontrons donc que (trivial):

$$\overline{OA} = \frac{\overline{AC}}{2} = \frac{a+b}{2} \quad (7.59)$$

Donc $\overline{OA}$ est bien la moyenne arithmétique μ_a de a et b .

Ensuite nous avons dans le triangle rectangle ADB :

$$\overline{AD}^2 = \overline{DB}^2 + \overline{BA}^2 \quad (7.60)$$

Puis dans le triangle rectangle BDC nous avons:

$$\overline{DC}^2 = \overline{BC}^2 + \overline{DB}^2 \quad (7.61)$$

Nous additionnons alors ces deux égalités, et nous trouvons:

$$2\overline{DB}^2 + \overline{BA}^2 + \overline{BC}^2 = \overline{AD}^2 + \overline{DC}^2 \quad (7.62)$$

Nous savons que D est sur un cercle de diamètre $\overline{AC}$, donc ADC est rectangle en D , donc:

$$\overline{AD}^2 + \overline{DC}^2 = \overline{AC}^2 \quad (7.63)$$

Puis nous remplaçons $\overline{BA}$ et $\overline{BC}$ par a et b :

$$2\overline{DB}^2 + a^2 + b^2 = \overline{AC}^2 = (a+b)^2 \Leftrightarrow 2\overline{DB}^2 = 2ab \Leftrightarrow \overline{DB} = \sqrt{ab} \quad (7.64)$$

Et donc, $\overline{DB}$ est bien la moyenne géométrique μ_g de a et b .

Nous reste à prouver alors que $\overline{OH}$ est la moyenne harmonique de a et b :

Nous avons dans un premier temps (projection orthogonale):

$$\overline{DO} \circ \overline{DB} = \overline{DO} \cdot (\overline{DB} \cos \alpha) = \overline{DO} \cdot \overline{DH} = \left(\frac{a+b}{2} \right) \overline{DH} \quad (7.65)$$

Puis nous avons aussi (projection orthogonale aussi):

$$\overline{DO} \circ \overline{DB} = (\overline{DO} \cos \alpha) \cdot \overline{DB} = \overline{DB} \cdot \overline{DB} = \overline{DB}^2 \quad (7.66)$$

Nous avons donc:

$$\overline{DB}^2 = \left(\frac{a+b}{2} \right) \overline{DH} \quad (7.67)$$

et comme $\overline{DB} = \sqrt{ab}$, nous avons donc:

$$\overline{DH} = \frac{2ab}{a+b} \quad (7.68)$$

$\overline{DH}$ est donc bien la moyenne harmonique de a et b , Archimède ne s'était pas trompé.

3. TYPES DE VARIABLES

Lorsque nous avons parlé des échantillons au début de ce chapitre, nous avons fait mention de deux types d'informations: les variables quantitatives et qualitatives. Nous n'avons cependant pas précisé qu'il existait trois types de variables quantitatives très importantes qu'il convient absolument de différencier:

1. Les variables discrètes (par comptage): Sont analysées avec des lois statistiques basées sur un domaine de définition dénombrable toujours strictement positif (loi de Poisson typiquement dans l'industrie). Sont quasiment toujours représentées sous forme graphique par des histogrammes.
2. Les variables continues (par mesure): Sont analysées avec des lois statistiques basées sur un domaine de définition non dénombrable strictement positif ou pouvant prendre toute valeur positive ou négative (loi Normale typiquement dans l'industrie). Sont également quasiment toujours représentées sous forme graphique par des histogrammes avec des intervalles de classe.
3. Les variables par attribut (de classification): Il ne s'agit pas de données numériques mais de données qualitatives de type {Oui, Non}, {Réussi, Échec}, {A temps, En retard}, etc. Les données de type attribut suivent une loi Binomiale.

Comprendre les différents types de données est une discipline importante de l'ingénieur parce que cela a des conséquences importantes sur le type d'analyse, les outils et techniques qui seront employés.

Une question fréquente concernant la collecte de données est de savoir quelle est la quantité qui devrait être collectée. Au fait cela dépend du niveau de précision souhaité. Nous verrons beaucoup plus loin dans ce chapitre (avec démonstration) comment déterminer mathématiquement la quantité de données à collecter en parlant de la précision souhaitée pour un process Normal.

Voyons de près de quoi il s'agit car maintenant que le concept de moyenne nous est relativement bien connu, nous allons pouvoir aborder des calculs plus formels et qui prendront tout leur sens.

3.1. VARIABLES DISCRÈTES

Soit X une variable indépendante (un individu d'un échantillon dont la propriété est indépendante des autres individus) qui peut prendre les valeurs aléatoires discrètes $x_1, x_2, x_3, \dots$ dans $\mathbb{R}$ (réalisations du vecteur $(X_1, X_2, \dots, X_n)$) avec les probabilités respectives $p_1, p_2, p_3, \dots$ où, de par l'axiomatique des probabilités:

$$p_i \in [0,1] \quad \sum_i p_i = 1 \quad (7.69)$$

Définitions:

D1. Soit X une variable aléatoire (v.a.) numérique (quantitative). Elle est complètement décrite par la valeur de la probabilité (pour les variables discrètes) ou par la probabilité cumulée (pour les variables continues) pour qu'une réalisation de cette variable soit inférieure à x pour tout x . Cette probabilité (cumulée) est notée:

$$F(x) = P(X \leq x) \quad \forall x \in \mathbb{R} \quad (7.70)$$

avec:

$$1 \geq P(X) \geq 0 \quad (7.71)$$

où $F(x)$ s'appelle la "**fonction de répartition**" de la variable X . C'est la proportion théorique de la population considérée dont la valeur est inférieure ou égale à x . Il s'ensuit:

$$P(X > x) = 1 - F(x) \Leftrightarrow P(X \leq x) + P(X > x) = 1 \quad (7.72)$$

Plus généralement, pour toute paire de nombres a et b avec $a < b$, nous avons:

$$P(a \leq X \leq b) = F(b) - F(a) \quad (7.73)$$

D2. La "**fonction de répartition empirique**" est quant à elle définie naturellement par (nous avons indiqué les différentes notations courantes dans la littérature):

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{x_i \leq x} = \frac{1}{n} \sum_{i=1}^n \delta_{x_i \leq x} = \frac{\#(i : x_i \leq x)}{n} \in [0, 1] \quad (7.74)$$

associé à l'échantillon de variables aléatoires indépendantes et identiquement distribuées (ce que l'on nomme aussi un "**vecteur aléatoire**" noté $(X_1, X_2, \dots, X_n)$).

Il s'agit simplement du cumul normalisé à l'unité des fréquences d'apparition en-dessous d'une certaine valeur fixée (démarche que la majorité des êtres humains font naturellement en cherchant la fonction de répartition).

Donc si nous reprenons l'exemple de salaires, vu plus haut, nous avons alors par exemple pour x fixé à 1'800:

Salaires ordonnés ($x_i \leq x$)	Fréquence $1_{x_i \leq x}$
1200	1
1220	1
1250	1
1300	1
1350	1
1450	1
1450	1
1560	1
1600	1
1800	1
1900	0
2150	0
2310	0
2600	0
3000	0
3400	0
4800	0

Figure: 7.7 - Exemple de la fonction de répartition empirique

et donc:

$$\hat{F}_{17}(x \leq 1'800) = \frac{1}{17} \sum_{i=1}^{17} 1_{x_i \leq x} = \frac{1}{17} 10 \cong 59\% \quad (7.75)$$

La fonction de répartition est clairement une fonction monotone croissante (ou plus précisément "non décroissante") dont les valeurs vont de 0 à 1.

3.1.1. ESPÉRANCE ET VARIANCE DE V.A. DISCRÈTES

Définition: Nous définissons "**l'espérance mathématique**", appelée aussi "**moment d'ordre 1**", de la variable X par la relation:

$$\mu_X = E(X) = \sum_i p_i x_i \quad (7.76)$$

appelée aussi "**règle des parties**".

En d'autres termes, nous savons qu'à chaque événement de l'espace des échantillons est associé une probabilité à laquelle nous associons également une valeur (donnée par la variable aléatoire). La question étant alors de savoir quelle valeur, à long terme, nous pouvons obtenir. La valeur espérée, (l'espérance mathématique donc...) est alors la moyenne pondérée, par la probabilité, de toutes les valeurs des événements de l'espace des échantillons.

Si la probabilité est donnée par une fonction de distribution (voir les définitions des fonctions de

distribution plus bas) de la variable aléatoire, nous avons:

$$p_i = f(x_i) \Rightarrow \mu_X = E(X) = \sum_i x_i f(x_i) \quad (7.77)$$

Remarques:

R1. μ_X peut être notée μ s'il n'y pas de confusion possible.

R2. Si nous considérons chaque valeur $x_1, x_2, x_3, \dots$ comme les composantes d'un vecteur $\vec{x}$ et chaque probabilité (ou pondération) $p_1, p_2, p_3, \dots$ comme les composantes d'un vecteur $\vec{p}$ alors nous pouvons écrire l'espérance de manière technique sous la forme d'un produit scalaire souvent noté:

$$\vec{p} \circ \vec{x} = \langle \vec{p}, \vec{x} \rangle = \sum_{i=1}^n p_i x_i = E(X) \quad (7.78)$$

Voici les propriétés mathématiques les plus importantes de l'espérance pour toute variable aléatoire (quelle que soit sa loi!) ou pour toute série de variables aléatoires et que nous utiliserons souvent tout au long de ce chapitre:

P1. Multiplication par une constante:

$$E(\alpha X) = \sum_i \alpha x_i P(X = x_i) = \alpha \sum_i x_i P(X = x_i) = \alpha E(X) \quad (7.79)$$

P2. Somme de deux variables aléatoires:

$$\begin{aligned} E(X+Y) &= \sum_{i,j} [(x_i + y_j) P((X = x_i) \cap (Y = y_j))] \\ &= \sum_{i,j} [x_i P((X = x_i) \cap (Y = y_j))] + \sum_{i,j} [y_j P((X = x_i) \cap (Y = y_j))] \\ &= \sum_i x_i \left(\sum_j P((X = x_i) \cap (Y = y_j)) \right) + \sum_j y_j \left(\sum_i P((X = x_i) \cap (Y = y_j)) \right) \quad (7.80) \\ &= \sum_i x_i P\left((X = x_i) \cap \bigcup_j (Y = y_j)\right) + \sum_j y_j P\left((Y = y_j) \cap \bigcup_i (X = x_i)\right) \\ &= \sum_i x_i P(X = x_i) + \sum_j y_j P(Y = y_j) = E(X) + E(Y) \end{aligned}$$

où nous avons utilisé dans la 4ème ligne, la propriété vue dans le chapitre de Probabilités:

$$P\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} P(A_i)$$

Nous en déduisons que pour n variables aléatoires X_i , définies sur une même loi de distribution:

$$E\left(\sum_i X_i\right) = \sum_i E(X_i) \quad (7.81)$$

P3. Espérance d'une constante:

$$E(c^{te}) = \sum_i c^{te} \cdot p_i = c^{te} \sum_i p_i = c^{te} \quad (7.82)$$

Définition: Après avoir traduit la tendance par l'espérance il est intéressant de traduire la dispersion ou "**dévi**ation **standard**" autour de l'espérance par une valeur appelée "**variance de X**" ou encore "**moment centré du deuxième ordre**", notée $V(X)$ ou σ_X^2 (lire "sigma-deux") et donnée sous sa forme discrète par:

$$\sigma_X^2 = V(X) = E\left[(X - \mu_X)^2\right] = \sum_i (x_i - \mu_X)^2 p_i \quad (7.83)$$

La variance n'est cependant pas comparable directement à la moyenne, car l'unité de la variance est le carré de l'unité de la variable, ce qui découle directement de sa définition. Pour que l'indicateur de dispersion puisse être comparé aux paramètres de tendance centrale (moyenne, médiane et... mode), il suffit d'en prendre la racine carrée.

Par commodité, nous définissons ainsi "**l'écart-type**" de X , noté $\sigma(X)$, par:

$$\sigma(X) = \sigma_X = \sqrt{V(X)} \quad (7.84)$$

L'écart-type est donc la moyenne quadratique des écarts (ou "écart moyen quadratique") entre les observations et leur moyenne.

Remarques:

R1. L'écart-type σ_X de la variable aléatoire X peut être noté σ s'il n'y pas de confusion possible.

R2. L'écart-type et la variance sont, dans la littérature, souvent appelés "**paramètres de dispersion**" à l'opposé de la moyenne, mode et médiane qui sont appelés des "**paramètres de position**".

Définition: Le rapport σ / μ (exprimé en %) parfois utilisé dans les entreprises comme comparaison de la moyenne et de l'écart-type est appelée le "**coefficient de variation**" (C.V.).

Pourquoi trouvons-nous un carré (réciproquement une racine) dans cette définition de la variance? La raison intuitive est simple (la rigoureuse l'est nettement moins...). Nous avons démontré plus haut que la somme des écarts à la moyenne pondérés par les effectifs, est toujours nulle:

$$\sum_{i=1}^r n_i (x_i - \mu) = 0 \quad (7.85)$$

Or, si nous assimilons les effectifs par la probabilité en normalisant ceux-ci par rapport à n , nous

tombons sur une relation qui est la même que la variance à la différence que le terme entre parenthèse n'est pas au carré. Et nous voyons alors immédiatement le problème... la mesure de dispersion serait toujours nulle d'où la nécessité de porter cela au carré.

Nous pourrions imaginer cependant d'utiliser la valeur absolue des écarts à la moyenne, mais pour un certain nombre de raisons que nous verrons plus loin lors de notre étude des estimateurs le choix de porter au carré s'impose assez naturellement.

Signalons cependant quand même l'utilisation courante dans l'industrie deux autres indicateurs fréquents de la dispersion:

- "**L'écart absolu moyen**" (moyenne des valeurs absolues des écarts à la moyenne):

$$EM = \frac{\sum_{i=1}^N |x_i - \mu|}{N} \quad (7.86)$$

qui est un indicateur élémentaire très utilisé lorsque nous ne souhaitons pas faire de l'inférence statistique sur une série de mesures. Cet écart peut être facilement calculé dans la version française Microsoft Excel 11.8346 à l'aide de la fonction ECART.MOYEN().

- "**La déviation absolue de la médiane**" noté *MAD* (médiane des valeurs absolues des écarts à la médiane):

$$MAD = M_e(|X - M_e(X)|) \quad (7.87)$$

qui est considéré comme un indicateur plus robuste de la dispersion que ceux donnés par l'écart absolu moyen ou l'écart-type (malheureusement cet indicateur n'est pas intégré nativement dans les tableurs).

Exemple:

Considérons les mesures d'une variable aléatoire *X*:

$$(1, 1, 2, 2, 4, 6, 9) \quad (7.88)$$

dont la médiane vaut:

$$M_e(X) = M_e(1, 1, 2, 2, 4, 6, 9) = 2 \quad (7.89)$$

Les déviations absolues par rapport à la médiane sont alors:

$$|X - M_e(X)| = (1, 1, 0, 0, 2, 4, 7) \quad (7.90)$$

Mis dans l'ordre croissant, nous avons alors:

$$(0, 0, 1, 1, 2, 4, 7) \quad (7.91)$$

où nous identifions la déviation absolue de la médiane vaut:

$$MAD = M_e(0, 0, 1, 1, 2, 4, 7) = 1 \quad (7.92)$$

Dans le cas où nous avons à disposition une série de mesures, nous pouvons estimer la valeur expérimentale de la moyenne (l'espérance) et de la variance des mesures par les estimateurs suivants (il s'agit simplement au fait de l'espérance et l'écart-type d'un échantillon dont les événements sont tous équiprobables) dont la notation est particulière:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{et} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 \quad (7.93)$$

Démonstration:

$$\begin{aligned} \mu &= E(X) = \sum_{i=1}^n p_i x_i = \sum_{i=1}^n \frac{1}{n} x_i = \frac{1}{n} \sum_{i=1}^n x_i = \hat{\mu} \\ \sigma^2 &= E(X - E(X))^2 = E[(X - \mu_X)^2] = \sum_{i=1}^n (x_i - \mu_X)^2 p_i = \sum_{i=1}^n (x_i - \mu_X)^2 \frac{1}{n} \quad (7.94) \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 = \hat{\sigma}^2 \end{aligned}$$

□C.Q.F.D.

Et démontrons une petite propriété bien sympathique comme quoi la moyenne arithmétique est un optimum de la somme des carrés des écarts. Effectivement, nous avons:

$$\sum_{i=1}^n (x_i - \alpha)^2 = \sum_{i=1}^n x_i^2 - 2\alpha \sum_{i=1}^n x_i + n\alpha^2 \quad (7.95)$$

et si nous cherchons α tel que la dérivée de l'expression ci-dessus est nulle:

$$\frac{d}{d\alpha} \left(\sum_{i=1}^n x_i^2 - 2\alpha \sum_{i=1}^n x_i + n\alpha^2 \right) = 0 \quad (7.96)$$

alors α est un optimum. Nous avons alors:

$$\frac{d}{d\alpha} \left(\sum_{i=1}^n x_i^2 - 2\alpha \sum_{i=1}^n x_i + n\alpha^2 \right) = -2 \sum_{i=1}^n x_i + 2n\alpha = 0 \quad (7.97)$$

soit après réarrangement et simplification élémentaire:

$$\alpha = \frac{1}{n} \sum_{i=1}^n x_i \quad (7.98)$$

Il s'agit donc bien de la moyenne arithmétique. Maintenant pour savoir s'il s'agit d'un extrema de type

maximum ou minimum il suffit de faire la dérivée seconde et de voir que cela donne une constante positive (donc la dérivée première augmente quand α augmente). Il s'agit alors d'un bien extrema de type minimum!!!

Le terme de la somme se trouvant dans l'expression de la variance (écart-type) est appelée "**somme des carrés des écarts à la moyenne**". Nous l'appelons aussi la "**somme des carrés totale**", ou encore la "**variation totale**" dans le cadre de l'étude de l'ANOVA (voir la fin de ce chapitre).

Remarque: Il est important que le lecteur comprenne que dans ce cas l'espérance se calcule simplement en utilisant la moyenne arithmétique!

La variance peut également s'écrire sous la forme de la "**relation de Huyghens**" que nous réutiliserons plusieurs fois par la suite. Voyons de quoi il s'agit:

$$\begin{aligned} V(X) &= E[(X - \mu_X)^2] = \sum_i (x_i - \mu_X)^2 f(x_i) = \sum_i (x_i^2 - 2\mu_X x_i + \mu_X^2) f(x_i) \\ &= \sum_i x_i^2 f(x_i) - 2\mu_X \sum_i x_i f(x_i) + \mu_X^2 \sum_i f(x_i) = \sum_i x_i^2 f(x_i) - 2\mu_X^2 + \mu_X^2 \quad (7.99) \\ &= \sum_i x_i^2 f(x_i) - \mu_X^2 = E(X^2) - \mu_X^2 = E(X^2) - E(X)^2 \end{aligned}$$

Faisons maintenant un petit crochet relativement à un scénario fréquent générateur d'erreurs dans les entreprises lorsque plusieurs séries statistiques sont manipulées (cas très fréquent dans l'industrie ainsi que dans les assurances ou la finance).

Considérons deux séries statistiques portant sur le même caractère:

- $(x_1, n_1), \dots, (x_p, n_p)$, effectif total n , moyenne $\bar{x}$, écart-type σ_x
- $(y_1, m_1), \dots, (y_q, m_q)$, effectif total m , moyenne $\bar{y}$, écart-type σ_y

Nous noterons (z_k, r_k) la série statistique obtenue en regroupant les deux séries. Nous avons alors:

$$\bar{z} = \frac{\sum_{i=1}^p x_i + \sum_{i=1}^q y_i}{n+m} = \frac{n \frac{1}{n} \sum_{i=1}^p x_i + m \frac{1}{m} \sum_{i=1}^q y_i}{n+m} = \frac{n\bar{x} + m\bar{y}}{n+m} \quad (7.100)$$

Donc la moyenne des moyennes n'est pas égale à la moyenne globale (première erreur fréquente dans les entreprises) exceptée si les deux séries statistiques ont le même nombre d'effectifs!!!

Concernant l'écart-type, rappelons d'abord que nous avons:

$$\sigma_x^2 = \frac{1}{\sum_{i=1}^p n_i} \sum_{i=1}^p n_i (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^p n_i (x_i - \bar{x})^2$$

$$\sigma_y^2 = \frac{1}{\sum_{i=1}^q m_i} \sum_{i=1}^q m_i (y_i - \bar{y})^2 = \frac{1}{m} \sum_{i=1}^q m_i (y_i - \bar{y})^2 \quad (7.101)$$

Pour la suite, rappelons que nous avons démontré précédemment la relation de Huygens:

$$V(z) = \sigma_z^2 = E(z^2) - E(z)^2 = \frac{1}{k} \sum_{i=1}^k k_i z_i^2 - \bar{z}^2 \quad (7.102)$$

Il vient alors:

$$\begin{aligned} \sigma_z^2 &= \frac{1}{k} \sum_{i=1}^k k_i z_i^2 - \bar{z}^2 = \frac{1}{k} \sum_{i=1}^k k_i z_i^2 - \left(\frac{n\bar{x} + m\bar{y}}{n+m} \right)^2 = \frac{1}{n+m} \left(\sum_{i=1}^p n_i x_i^2 + \sum_{i=1}^q m_i y_i^2 \right) - \left(\frac{n\bar{x} + m\bar{y}}{n+m} \right)^2 \\ &= \frac{n \frac{1}{n} \sum_{i=1}^p n_i x_i^2 + m \frac{1}{m} \sum_{i=1}^q m_i y_i^2}{n+m} - \left(\frac{n\bar{x} + m\bar{y}}{n+m} \right)^2 \\ &= \frac{n \frac{1}{n} \sum_{i=1}^p n_i x_i^2 - n\bar{x}^2 + m \frac{1}{m} \sum_{i=1}^q m_i y_i^2 - m\bar{y}^2}{n+m} + \frac{n\bar{x}^2 + m\bar{y}^2}{n+m} - \left(\frac{n\bar{x} + m\bar{y}}{n+m} \right)^2 \\ &= \frac{n \left(\frac{1}{n} \sum_{i=1}^p n_i x_i^2 - \bar{x}^2 \right) + m \left(\frac{1}{m} \sum_{i=1}^q m_i y_i^2 - \bar{y}^2 \right)}{n+m} + \frac{n\bar{x}^2 + m\bar{y}^2}{n+m} - \left(\frac{n\bar{x} + m\bar{y}}{n+m} \right)^2 \\ &= \frac{n\sigma_x^2 + m\sigma_y^2}{n+m} + \frac{n\bar{x}^2 + m\bar{y}^2}{n+m} - \left(\frac{n\bar{x} + m\bar{y}}{n+m} \right)^2 \\ &= \frac{n\sigma_x^2 + m\sigma_y^2}{n+m} + \frac{(n\bar{x}^2 + m\bar{y}^2)(n+m) - (n\bar{x} + m\bar{y})^2}{(n+m)^2} \\ &= \frac{n\sigma_x^2 + m\sigma_y^2}{n+m} + \frac{(n^2\bar{x}^2 + mn\bar{y}^2 + mn\bar{x}^2 + m^2\bar{y}^2) - (n^2\bar{x}^2 + 2n\bar{x}m\bar{y} + m^2\bar{y}^2)}{(n+m)^2} \\ &= \frac{n\sigma_x^2 + m\sigma_y^2}{n+m} + \frac{(mn\bar{y}^2 + mn\bar{x}^2) - (2n\bar{x}m\bar{y})}{(n+m)^2} = \frac{n\sigma_x^2 + m\sigma_y^2}{n+m} + nm \frac{\bar{y}^2 + \bar{x}^2 - 2\bar{x}\bar{y}}{(n+m)^2} \\ &= \frac{n\sigma_x^2 + m\sigma_y^2}{n+m} + nm \frac{(\bar{x} - \bar{y})^2}{(n+m)^2} \end{aligned} \quad (7.103)$$

Donc nous voyons que l'écart-type global n'est pas égal à la somme des écarts-types (deuxième erreur courante dans les entreprises) excepté si les effectifs et les moyennes sont les mêmes dans les deux

séries!!!

Considérons maintenant X une variable aléatoire d'espérance μ (valeur constante et déterminée) et de variance σ^2 (valeur constante et déterminée), nous définissons la "**variable centrée réduite**" par la relation:

$$Y = \frac{X - \mu}{\sigma} \quad (7.104)$$

et l'on démontre de façon très simple en utilisant la propriété de linéarité de l'espérance et la propriété de multiplication par un scalaire de la variance (voir de suite après) que:

$$E(Y) = 0, V(Y) = 1 \quad (7.105)$$

Démonstration:

$$\begin{aligned} E(Y) &= E\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma} E(X - \mu) = \frac{1}{\sigma} (E(X) - E(\mu)) = \frac{1}{\sigma} (\mu - \mu \cdot E(1)) \\ &= \frac{1}{\sigma} (\mu - \mu \cdot E(1)) = 0 = \mu_Y \end{aligned} \quad (7.106)$$

et en utilisant la relation de Huyghens:

$$\begin{aligned} V(Y) &= E(Y^2) - \mu_Y^2 = E(Y^2) = E\left(\frac{1}{\sigma^2} (X^2 - 2X\mu + \mu^2)\right) = \frac{1}{\sigma^2} (E(X^2) - 2E(X)\mu + \mu^2) \\ &= \frac{1}{\sigma^2} (E(X^2) - 2E(X)\mu + \mu^2) = \frac{1}{\sigma^2} (E(X^2) - 2\mu^2 + \mu^2) = \frac{1}{\sigma^2} (E(X^2) - \mu^2) \\ &= \frac{1}{\sigma^2} (V(X) + \mu^2 - \mu^2) = \frac{1}{\sigma^2} V(X) = \frac{1}{\sigma^2} \sigma^2 = 1 \end{aligned} \quad (7.107)$$

□C.Q.F.D.

Ainsi, toute répartition statistique définie par une moyenne et un écart-type peut être transformée en une autre distribution statistique souvent plus simple à analyser.

Voici quelques propriétés mathématiques importantes de la variance:

P1. Multiplication par une constante:

$$V(nX) = \sum_i f(x_i) [nx_i - n\mu_X]^2 = n^2 \sum_i f(x_i) [X - \mu_X]^2 = n^2 V(X) \quad (7.108)$$

P2. Somme de deux variables aléatoires:

$$\begin{aligned}
V(X+Y) &= \sum_i \sum_j \left((x_i + y_j) - (\mu_X + \mu_Y) \right)^2 f(x_i, y_j) \\
&= \sum_i \sum_j \left((x_i - \mu_X) + (y_j - \mu_Y) \right)^2 f(x_i, y_j) \\
&= \sum_i \sum_j \left((x_i - \mu_X)^2 + 2(x_i - \mu_X)(y_j - \mu_Y) + (y_j - \mu_Y)^2 \right) f(x_i, y_j) \\
&= \sum_{i,j} (x_i - \mu_X)^2 f(x_i, y_j) + 2 \sum_{i,j} (x_i - \mu_X)(y_j - \mu_Y) f(x_i, y_j) + \sum_{i,j} (y_j - \mu_Y)^2 f(x_i, y_j) \\
&= V(X) + V(Y) + 2 \sum_{i,j} (x_i - \mu_X)(y_j - \mu_Y) f(x_i, y_j) \\
&= V(X) + V(Y) + 2E[(X - \mu_X)(Y - \mu_Y)] = V(X) + V(Y) + 2\text{cov}(X, Y)
\end{aligned} \tag{7.109}$$

3.1.2. COVARIANCE DISCRÈTE

Nous venons de voir dans la dernière relation le concept de "**covariance**" dont nous verrons une expression plus commode un peu plus bas mais donc définie par:

$$c_{X,Y} = E[(X - \mu_X)(Y - \mu_Y)] \tag{7.110}$$

Introduisons une forme plus générale et extrêmement importante de la covariance dans de nombreux domaines:

$$\begin{aligned}
V(X+Y+Z) &= \sum_i \sum_j \sum_k \left((x_i + y_j + z_k) - (\mu_X + \mu_Y + \mu_Z) \right)^2 f(x_i, y_j, z_k) \\
&= \sum_i \sum_j \sum_k \left((x_i - \mu_X) + (y_j - \mu_Y) + (z_k - \mu_Z) \right)^2 f(x_i, y_j, z_k) \\
&= \sum_i \sum_j \sum_k (A+B+C)^2 f(x_i, y_j, z_k) \\
&= \sum_i \sum_j \sum_k \left(A^2 + B^2 + C^2 + 2AB + 2BC + 2AC \right) f(x_i, y_j, z_k) \\
&= V(X) + V(Y) + V(Z) + 2 \sum_i \sum_j \sum_k (AB + BC + AC) f(x_i, y_j, z_k) \\
&= V(X) + V(Y) + V(Z) + 2\text{cov}(X, Y) + 2\text{cov}(Y, Z) + 2\text{cov}(X, Z) \\
&= V(X_1) + V(X_2) + V(X_3) + 2\text{cov}(X_1, X_2) + 2\text{cov}(X_2, X_3) + 2\text{cov}(X_1, X_3) \\
&= \sum_{i=1}^3 V(X_i) + 2 \sum_{i < j}^3 \text{cov}(X_i, X_j)
\end{aligned} \tag{7.111}$$

Donc dans le cas général:

$$V\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n V(X_i) + 2 \sum_{i < j}^n \text{cov}(X_i, X_j) \tag{7.112}$$

En utilisant la propriété de linéarité de l'espérance et le fait que:

$$\begin{aligned} E(X) &= c^{te} \\ E(c^{te}) &= c^{te} \end{aligned} \quad (7.113)$$

nous avons pour la covariance:

$$\begin{aligned} \text{cov}(X, Y) &= E[(X - E(X))(Y - E(Y))] = E[XY - E(X)Y - XE(Y) + E(X)E(Y)] \\ &= E(XY) - E(E(X)Y) - E(XE(Y)) + E(E(X)E(Y)) \\ &= E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y)E(1) \\ &= E(XY) - 2E(X)E(Y) + E(X)E(Y)E(1) = E(XY) - E(X)E(Y) \end{aligned} \quad (7.114)$$

et donc nous obtenons la relation très utilisée en statistiques et finance dans la pratique appelée **"formule de la covariance"**...:

$$c_{X,Y} = \text{cov}(X, Y) = E(XY) - E(X)E(Y) \quad (7.115)$$

qui est cependant plus connue sous la forme:

$$c_{X,Y} = \text{cov}(X, Y) = \frac{1}{n} \sum_i x_i y_i - \bar{x} \cdot \bar{y} \quad (7.116)$$

Indiquons également que si $X = Y$, ce qui équivaut donc à une covariance univariée, nous retrouvons la relation de Huyghens:

$$c_{X,X} = E(X^2) - \mu_X^2 = E[(X - \mu_X)(X - \mu_X)] = E[(X - \mu_X)^2] = V(X) = \sigma_X^2 \quad (7.117)$$

Remarque: Les statistiques peuvent être découpées selon le nombre de variables aléatoires que nous étudions. Ainsi, lorsqu'une seule variable aléatoire est étudiée, nous parlons de **"statistique univariée"**, pour deux variables aléatoires de **"statistique bivariée"** et en général, de **"statistique multivariée"**.

Si et seulement si les variables sont équiprobables, nous retrouvons la covariance dans la littérature sous la forme suivante, appelée parfois **"covariance de Pearson"**, qui découle de calculs que nous avons déjà fait antérieurement avec l'espérance:

$$c_{X,Y} = \frac{1}{n} \sum_{i=1}^n (y_i - \mu_Y)(x_i - \mu_X) \quad (7.118)$$

La covariance est un indicateur de la variation simultanée de X et Y . En effet si, en général X et Y croissent simultanément, les produits $(y_i - \mu_Y)(x_i - \mu_X)$ seront positifs (corrélés positivement), tandis que si Y décroît lorsque X croît, ces mêmes produits seront négatifs (corrélés négativement).

Signalons que si nous distribuons les termes de la dernière relation, nous avons:

$$\begin{aligned}
c_{X,Y} = \text{cov}(X,Y) &= \frac{1}{n} \sum_i (x_i - \mu_X)(y_i - \mu_Y) = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \\
&= \frac{1}{n} \sum_i x_i (y_i - \bar{y}) - \bar{x} (y_i - \bar{y}) = \frac{1}{n} \left(\sum_i x_i (y_i - \bar{y}) - \bar{x} \sum_i (y_i - \bar{y}) \right) \quad (7.119)
\end{aligned}$$

et nous avons déjà démontré que la somme des écarts à la moyenne est nulle. Dès lors nous obtenons une autre forme courante de la covariance:

$$c_{X,Y} = \text{cov}(X,Y) = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_i x_i (y_i - \bar{y}) \quad (7.120)$$

et par symétrie:

$$c_{X,Y} = \text{cov}(X,Y) = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_i y_i (x_i - \bar{x}) \quad (7.121)$$

Donc au final, dans le cas équiprobable, nous avons finalement les trois relations équivalentes importantes utilisées dans différents chapitres du présent site:

$$\begin{aligned}
c_{X,Y} = \text{cov}(X,Y) &= \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \\
c_{X,Y} = \text{cov}(X,Y) &= \frac{1}{n} \sum_i x_i y_i - \bar{x} \cdot \bar{y} \quad (7.122) \\
c_{X,Y} = \text{cov}(X,Y) &= \frac{1}{n} \sum_i x_i (y_i - \bar{y}) = \frac{1}{n} \sum_i y_i (x_i - \bar{x})
\end{aligned}$$

Dans le chapitre de Méthodes Numériques pour notre étude de la régression linéaire et de l'analyse factorielle nous aurons besoin de l'expression explicite de la propriété de bilinéarité de la variance. Pour voir en quoi cela consiste exactement, considérons trois variables aléatoires X , Y et Z et a et b deux constantes. Alors en utilisant la troisième relation donnée précédemment, nous avons:

$$\begin{aligned}
\text{cov}(Y, aX + bZ) &= \frac{1}{n} \sum_i (y_i - \bar{y})(ax_i + bz_i) \\
&= \frac{1}{n} \sum_i [(y_i - \bar{y})ax_i + (y_i - \bar{y})bz_i] = \frac{1}{n} \left[\sum_i (y_i - \bar{y})ax_i + \sum_i (y_i - \bar{y})bz_i \right] \quad (7.123) \\
&= a \frac{1}{n} \sum_i (y_i - \bar{y})x_i + b \frac{1}{n} \sum_i (y_i - \bar{y})z_i = a \text{cov}(X,Y) + b \text{cov}(Y,Z)
\end{aligned}$$

Cette dernière relation est elle aussi importante et sera utilisée dans plusieurs chapitres du site (Économie, Méthodes Numériques). Elle nous permet aussi d'obtenir directement des covariances entre des sommes de variables.

 Exemple:

Si X, Y, Z, T sont quatre variables aléatoires définies sur la même population, nous voulons calculer la covariance suivante:

$$\text{cov}(3X + 5Y, 4Z - 2T) \quad (7.124)$$

Nous allons donc développer en deux fois (raison pour laquelle nous appelons cela la "bilinearité"). D'abord par rapport au second argument (arbitrairement!):

$$\text{cov}(3X + 5Y, 4Z - 2T) = 4\text{cov}(3X + 5Y, Z) - 2\text{cov}(3X + 5Y, T) \quad (7.125)$$

et ensuite par rapport au premier:

$$\text{cov}(3X + 5Y, 4Z - 2T) = 4[3\text{cov}(X, Z) + 5\text{cov}(Y, Z)] - 2[3\text{cov}(X, T) + 5\text{cov}(Y, T)] \quad (7.126)$$

Donc au final:

$$\text{cov}(3X + 5Y, 4Z - 2T) = 12\text{cov}(X, Z) + 20\text{cov}(Y, Z) - 6\text{cov}(X, T) - 10\text{cov}(Y, T) \quad (7.127)$$

Maintenant, considérons $\vec{X}$ un vecteur de composantes $(x_1, x_2, \dots, x_n)$ et $\vec{Y}$ un autre vecteur de composantes $(y_1, y_2, \dots, y_m)$, tous deux étant des variables aléatoires, le calcul de la covariance des composantes deux à deux donne ce que l'on appelle la "**matrice des covariances**" (outil très utilisé en finance, dans la gestion en général et les méthodes numériques statistiques!).

Effectivement, si nous notons:

$$c_{X_m, X_n} = E\left[\left(x_m - \mu_{X_m}\right)\left(x_n - \mu_{X_n}\right)\right] = c_{m,n} \quad (7.128)$$

Nous pouvons dès lors écrire une matrice symétrique (le plus souvent dans la pratique elle est carrée) sous la forme:

$$c_{m,n} = \begin{bmatrix} c_{11} & \dots & \dots & c_{1n} \\ \vdots & c_{22} & & \vdots \\ \vdots & & c_{33} & \vdots \\ c_{m1} & \dots & \dots & c_{mn} \end{bmatrix} \quad (7.129)$$

Cette matrice a comme propriété remarquable que si nous prenons deux vecteurs identiques (dont les composantes sont les mêmes variables aléatoires) et que nous calculons la matrice, alors la diagonale de cette dernière donnera les variances des composantes de vecteurs (voir les exemples dans le chapitre d'économétrie)! Raison pour laquelle cette matrice est souvent appelée "**matrices des variances-covariances**" et se retrouve donc parfois notée également sous la forme suivante:

$$c_{m,n} = \begin{bmatrix} V(X_1) & \cdots & \cdots & \text{cov}(X_1 X_n) \\ \vdots & V(X_2) & & \vdots \\ \vdots & & V(X_3) & \vdots \\ \text{cov}(X_m X_1) & \cdots & \cdots & V(X_m) \end{bmatrix} = \begin{bmatrix} \sigma_{X_1}^2 & \cdots & \cdots & \sigma_{X_1 X_n} \\ \vdots & \sigma_{X_2}^2 & & \vdots \\ \vdots & & \sigma_{X_3}^2 & \vdots \\ \sigma_{X_m X_1} & \cdots & \cdots & \sigma_{X_m}^2 \end{bmatrix} \quad (7.130)$$

Cette matrice a pour avantage de montrer rapidement quelles sont les couples de variables qui ont une covariance négative et donc... dont la variance de la somme est plus petite que la somme des variances.

Remarque: Cette matrice est très importante et nous la retrouverons fréquemment dans le chapitre d'Économie lors de notre étude de la théorie du portefeuille et dans les techniques de fouille de données (data mining, clustering) dans le chapitre de Méthodes numériques (l'analyse par composantes principales).

Rappelons maintenant que nous avons un axiome en probabilités (cf. chapitre de Probabilités) qui énonçait que deux événements A, B sont indépendants si:

$$P(A \cap B) = P(A) \cdot P(B) \quad (7.131)$$

De la même façon, par extension, nous définissons l'indépendance des variables aléatoires discrètes.

Définition: Soit X, Y deux variables aléatoires discrètes. Nous disons que X, Y sont indépendantes si:

$$\forall x, y \in \mathbb{R}, P(X = x, Y = y) = P(X = x) \cdot P(Y = y) \quad (7.132)$$

Plus généralement, les variables discrètes $X_1, \dots, X_n$ sont indépendantes (en bloc) si:

$$\forall x_1, \dots, x_n \in \mathbb{R}, P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i). \quad (7.133)$$

L'indépendance de deux variables aléatoires implique que leur covariance est nulle (la réciproque est fausse!). Prouvons ceci dans le cas où les variables aléatoires ne prennent qu'un nombre fini de valeurs $\{x_i\}_I$ et $\{y_j\}_J$ respectivement, avec I, J des ensembles finis:

$$\begin{aligned} E(X \cdot Y) &= \sum_{i,j} P(X = x_i, Y = y_j) x_i y_j = \sum_{i,j} P(X = x_i) \cdot P(Y = y_j) x_i y_j \\ &= \left(\sum_i P(X = x_i) x_i \right) \cdot \left(\sum_j P(Y = y_j) y_j \right) = E(X) \cdot E(Y) \end{aligned} \quad (7.134)$$

et donc:

$$c_{X,Y} = E(X \cdot Y) - E(X)E(Y) = 0 \quad (7.135)$$

Remarque: Donc plus la covariance est faible, plus les séries sont indépendantes. A l'inverse, plus la covariance est élevée, plus les séries sont liées.

Etant donné que:

$$V(X + Y) = V(X) + V(Y) + 2c_{X,Y} \quad (7.136)$$

si X, Y sont indépendantes alors $c_{X,Y} = 0$:

$$V(X + Y) = V(X) + V(Y) \quad (7.137)$$

De manière plus générale si $X_1, \dots, X_n$ sont indépendantes (en bloc) alors pour toute loi statistique (!) nous avons:

$$V\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n V(X_i) \quad (7.138)$$

3.1.3. ESPÉRANCE ET VARIANCE DE LA MOYENNE (ERREUR STANDARD ET FCP)

Souvent en statistique, il est utile de déterminer l'écart-type de la moyenne empirique (ou en d'autres termes...: l'erreur quadratique moyenne). Voyons de quoi il s'agit:

Soit la moyenne d'une série de termes déterminés chacun par la mesure de plusieurs valeurs (il s'agit au fait de son estimateur dans un cas particulier comme nous le verrons beaucoup plus loin):

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n) \quad (7.139)$$

alors en utilisant les propriétés de l'espérance:

$$E(\bar{X}) = \frac{1}{n}(E(X_1) + E(X_2) + \dots + E(X_n)) \quad (7.140)$$

et si toutes les variables aléatoires sont identiquement distribuées et indépendantes nous avons alors:

$$E(\bar{X}) = \frac{1}{n}n\mu = \mu \quad (7.141)$$

Remarque: Nous démontrerons bien plus loin que si toutes les variables aléatoires sont identiquement distribuées et indépendantes et de variance finie, alors l'espérance suite une asymptotiquement une loi Normale.

Pour la variance, le même raisonnement s'applique:

$$V(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{1}{n^2}(V(X_1) + V(X_2) + \dots + V(X_n)) = \frac{1}{n^2}(\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2) \quad (7.142)$$

et si les variables aléatoires sont toutes identiquement distribuées et indépendantes (nous étudierons plus loin le cas très important et courant dans la pratique où cette dernière condition n'est pas satisfaite):

$$V(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \quad (7.143)$$

d'où l'écart-type de la moyenne appelé aussi "**erreur-type**", "**erreur-standard**" ou encore "**variation non systématique**":

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (7.144)$$

et il s'agit rigoureusement de l'écart-type de l'estimateur de la moyenne (c'est peut-être plus clair ainsi)!

Cette relation se trouve dans de nombreux logiciels dont les graphiques Microsoft Excel (mais il n'y a pas de fonction intégrée), écrite soit avec l'écart-type (comme ci-dessus), soit avec la notation de la variance (suffit de mettre au carré...).

Signalons que la dernière relation peut être utilisée même si la moyenne des n variables aléatoires n'est pas identique! La condition suffisante étant juste que les écarts-types soient tous égaux et c'est le cas de l'industrie (production).

Nous avons donc:

$$\begin{array}{ll} E(S_n) = n\mu & E(M_n) = \mu \\ V(S_n) = \sigma^2(S_n) = n\sigma^2 & V(M_n) = \sigma^2(M_n) = \frac{\sigma^2}{n} \end{array} \quad (7.145)$$

où S_n désigne la somme des n variables aléatoires et M_n leur moyenne.

La variable centrée réduite que nous avons introduite plus haut:

$$Y = \frac{X - \mu}{\sigma} \quad (7.146)$$

peut alors s'écrire de plusieurs manières:

$$Y = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{M_n - \mu}{\sigma/\sqrt{n}} \quad (7.147)$$

Par ailleurs, en supposant que le lecteur sache déjà ce qu'est une loi Normale $N(\mu, \sigma)$, nous démontrerons plus loin en détails car c'est extrêmement important (!) que la loi de probabilité de la variable aléatoire $\bar{X}$, moyenne de n variables aléatoires identiquement distribuées et linéairement indépendantes, est alors la loi:

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad (7.148)$$

3.1.4. COEFFICIENT DE CORRÉLATION

Maintenant, considérons X et Y deux variables aléatoires ayant pour covariance:

$$c_{X,Y} = E[(X - \mu_X)(Y - \mu_Y)] \quad (7.149)$$

Nous avons:

$$(c_{X,Y})^2 \leq V(X) \cdot V(Y) \quad (7.150)$$

nous allons démontrer cette relation immédiatement car l'utilisation de la covariance seule pour l'analyse des données n'est pas géniale car elle n'est pas à proprement parler bornée et simple d'usage (au niveau de l'interprétation). Nous allons donc construire un indicateur plus facile d'usage en entreprise.

Démonstration:

Choisissons une constante a quelconque et calculons la variance de:

$$aX + Y \quad (7.151)$$

Nous pouvons alors immédiatement écrire à l'aide des propriétés de la variance et de l'espérance:

$$V(aX + Y) = a^2 V(X) + 2ac_{X,Y} + V(Y) \quad (7.152)$$

La quantité de droite est positive et nulle en tout a par construction de la variance (de gauche). Donc le discriminant de l'expression, vue comme un trinôme en a du type:

$$\begin{aligned} P(x) &= ax^2 + bx + c = a \left[\left(x + \frac{b}{2a} \right)^2 - \frac{b^2 - 4ac}{4a^2} \right] \\ \Rightarrow \\ P(a) &= V(X)a^2 + 2c_{X,Y}a + V(Y) = V(X) \left[\left(a + \frac{2c_{X,Y}}{2V(X)} \right)^2 - \frac{(2c_{X,Y})^2 - 4V(X)V(Y)}{4V(X)^2} \right] \end{aligned} \quad (7.153)$$

Donc pour que $P(a)$ soit positif pour tout a nous avons comme seule possibilité que:

$$(2c_{X,Y})^2 - 4V(X)V(Y) \leq 0 \quad (7.154)$$

Soit après simplification:

$$(c_{X,Y})^2 \leq V(X)V(Y) \quad (7.155)$$

□C.Q.F.D.

Ce qui nous donne:

$$\frac{(c_{X,Y})^2}{V(X)V(Y)} \leq 1 \Leftrightarrow \frac{|c_{X,Y}|}{\sqrt{V(X)V(Y)}} \leq 1 \quad (7.156)$$

Finalement nous obtenons une forme de l'inégalité statistique dite "**inégalité de Cauchy-Schwarz**":

$$-1 \leq \frac{c_{X,Y}}{\sqrt{V(X)V(Y)}} \leq 1 \quad (7.157)$$

Si les variances de X et Y sont non nulles, la corrélation entre X et Y est définie par le "**coefficient de corrélation linéaire**" (il s'agit donc de la covariance standardisée afin que son amplitude ne soit pas dépendante de l'unité de mesure choisie):

$$R_{X,Y} = \frac{c_{X,Y}}{\sqrt{V(X)V(Y)}} \quad (7.158)$$

ce qui peut aussi s'écrire sous forme développée (en utilisant la relation de Huyghens):

$$R_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{V(X)V(Y)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{(E(X^2) - E(X)^2)(E(Y^2) - E(Y)^2)}} \quad (7.159)$$

ou encore plus condensée:

$$R_{X,Y} = \frac{c_{X,Y}}{\sigma(X)\sigma(Y)} \quad (7.160)$$

Signalons que normalement, la lettre R est réservée pour dire qu'il s'agit d'un estimateur du coefficient de corrélation alors que la définition ci-dessus n'est pas un estimateur et qu'en toute rigueur, nous devrions alors noter $\rho_{X,Y}$ selon les traditions d'usage.

Quels que soient l'unité et les ordres de grandeur, le coefficient de corrélation est un nombre sans unités (donc sa valeur ne dépend pas de l'unité de mesure choisie, ce qui n'est de loin pas le cas de tous les indicateurs statistiques!), compris entre -1 et 1. Il traduit la plus ou moins grande dépendance linéaire de X et Y et ou, géométriquement, le plus ou moins grand aplatissement. Nous pouvons donc dire qu'un coefficient de corrélation nul ou proche de 0 signifie qu'il n'y a pas de relation linéaire entre les caractères. Mais il n'entraîne aucune notion d'indépendance plus générale.

Quand le coefficient de corrélation est proche de 1 ou -1, les caractères sont dits fortement corrélés. Il faut prendre garde à la confusion fréquente entre corrélation et causalité. Cependant, que deux phénomènes soient corrélés n'implique en aucune façon que l'un soit cause de l'autre.

Ainsi:

- Si $R_{X,Y} = -1$ nous avons affaire à une corrélation négative dite "**corrélation négative parfaite**" (dans la cas d'une relation linéaire tous les points de mesures sont situés sur une droite de pente négative).
- Si $-1 < R_{X,Y} < 1$ nous avons affaire à une corrélation négative ou positive dite "**corrélation imparfaite**" (dans la cas d'une relation linéaire tous les points de mesures sont situés sur une droite de pente

négative ou respectivement positive).

- Si $R_{X,Y} = 0$ la corrélation est nulle... (dans la cas d'une relation linéaire tous les points de mesures sont situés sur une droite de pente nulle).
- Si $R_{X,Y} = 1$ nous avons affaire à une corrélation positive dite "**corrélation positive parfaite**" (dans la cas d'une relation linéaire tous les points de mesures sont situés sur une droite de pente positive).

L'analyse du coefficient de corrélation poursuit donc l'objectif de déterminer le degré d'association entre les différentes variables: celui-ci est souvent exprimé par le coefficient de détermination, qui est le carré du coefficient de corrélation. Le coefficient de détermination mesure donc la contribution d'une des variables à l'explication de la seconde.

En utilisant les expressions de la moyenne et de l'écart-type de variables équiprobables telles que démontrées plus haut, nous passons de:

$$R_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{(E(X^2) - E(X)^2)(E(Y^2) - E(Y)^2)}} \quad (7.161)$$

à l'estimateur du coefficient de corrélation:

$$R_{x,y} = \frac{\frac{1}{n} \sum_{k=1}^n x_k y_k - \frac{1}{n} \sum_{k=1}^n x_k \frac{1}{n} \sum_{k=1}^n y_k}{\sqrt{\left[\frac{1}{n} \sum_{k=1}^n x_k^2 - \frac{1}{n^2} \left(\sum_{k=1}^n x_k \right)^2 \right] \left[\frac{1}{n} \sum_{k=1}^n y_k^2 - \frac{1}{n^2} \left(\sum_{k=1}^n y_k \right)^2 \right]}} \quad (7.162)$$

où nous voyons que la covariance devient alors la moyenne des produits moins le produit des moyennes.

Soit après simplification:

$$R_{x,y} = \frac{\sum_{k=1}^n x_k y_k - \frac{1}{n} \sum_{k=1}^n x_k \sum_{k=1}^n y_k}{\sqrt{\left[\sum_{k=1}^n x_k^2 - \frac{1}{n} \left(\sum_{k=1}^n x_k \right)^2 \right] \left[\sum_{k=1}^n y_k^2 - \frac{1}{n} \left(\sum_{k=1}^n y_k \right)^2 \right]}} \quad (7.163)$$

Le coefficient de corrélation peut être calculé dans version française de Microsoft Excel 11.8346 avec entre autres la fonction intégrée COEFFICIENT.CORRELATION().

Remarques:

R1. Dans la littérature le coefficient de corrélation est souvent appelé "**coefficient d'échantillonnage de Pearson**" (dans le cas équiprobable) ou "**test de Bravais-Pearson**" (dans le cas non équiprobable) et lorsque nous le portons au carré, nous parlons alors de "**coefficient de détermination**".

R2. Souvent le carré de ce coefficient est un peu abusivement interprété comme le % de variation expliqué de la variable étudiée Y par la variable explicative X .

Enfin, à noter que nous avons donc la relation suivante qui est énormément utilisée dans la pratique (voir le chapitre d'Économie pour des exemples fameux!):

$$V(X+Y) = V(X) + V(Y) + 2R_{X,Y}\sqrt{V(X)V(Y)} \quad (7.164)$$

ou sa version avec l'écart-type:

$$\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2 + 2R_{X,Y}\sigma_X\sigma_Y} \quad (7.165)$$

Il s'agit d'une relation que l'on retrouve souvent en finance dans le cadre du calcul de la VaR selon la méthodologie RiskMetrics proposée par J.P. Morgan (cf. chapitre d'Économie).

 Exemple:

Une compagnie aérienne a à sa disposition 120 sièges qu'elle réserve pour des passagers en correspondance venant de deux autres vols arrivés un peu plus tôt dans la journée et en partance pour Francfort. Le premier vol arrive de Manille et le nombre de passagers à son bord suit une loi Normale de moyenne 50 et de variance 169. Le second vol arrive de Taipei et le nombre de passagers à son bord suit une loi Normale de moyenne 45 et de variance 196.

Le coefficient de corrélation linéaire entre le nombre de passagers des deux vols est mesuré comme étant:

$$R_{X,Y} = 0.5 \quad (7.166)$$

La loi que suit le nombre de passagers pour Francfort si nous supposons que la loi du couple suit elle aussi une loi Normale (selon énoncé!) est:

$$X+Y = N(\mu_X + \mu_Y, V(X+Y)) \quad (7.167)$$

avec:

$$\mu_X + \mu_Y = 50 + 45 = 95 \quad (7.168)$$

et:

$$V(X+Y) = V(X) + V(Y) + 2R_{X,Y}\sqrt{V(X)V(Y)} \quad (7.169)$$

ce qui donne:

$$V(X+Y) = 169 + 196 + 2 \cdot 0.5 \sqrt{169 \cdot 196} = 169 + 196 + 182 = 547 \quad (7.170)$$

3.2. VARIABLES CONTINUES

Définitions:

D1. Nous disons que X est une variable continue si sa "**fonction de répartition**" est continue (déjà définie plus haut). La fonction de répartition de X étant définie par:

$$F(x) = P(X \leq x) \quad \forall x \in \mathbb{R} \quad (7.171)$$

soit la probabilité cumulée que la variable aléatoire X soit plus petite ou égale à la valeur x fixée. Nous avons aussi bien évidemment:

$$1 \geq P(X) \geq 0 \quad (7.172)$$

D3. Nous appelons:

$$G(x) = 1 - F(x) = P(X > x) \quad (7.173)$$

la "**fonction de survie**" (survival function) ou "**fonction de queue**" (tail distribution function).

D2. Si de plus la fonction de répartition F de X est continûment dérivable de dérivée ρ appelée "**fonction de densité**" ou "**fonction de masse**" ou encore "**fonction de distribution**" alors nous disons que X est absolument continue et dans ce cas nous avons:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} \rho(x) dx = F(x_2) - F(x_1) \quad (7.174)$$

avec la condition de normalisation:

$$P(X < +\infty) = \int_{-\infty}^{+\infty} \rho(x) dx = 1 \quad (7.175)$$

Toute fonction de distribution de probabilité doit satisfaire l'intégrale de normalisation dans son domaine de définition!

Remarque: Il est intéressant de remarquer que la définition amène à ce que la probabilité qu'une variable aléatoire totalement continue prenne une valeur donnée est nulle! Donc ce n'est pas parce qu'un événement a une probabilité nulle qu'il ne peut arriver!!!

La moyenne ayant été définie par la somme pour une variable discrète, elle devient une intégrale pour

une variable continue:

$$E(X) = \int_{-\infty}^{+\infty} x \rho(x) dx \quad (7.176)$$

et la variance s'écrit donc:

$$V(X) = \int_{-\infty}^{+\infty} [x - E(X)]^2 \rho(x) dx \quad (7.177)$$

Nous avons alors aussi la médiane qui est logiquement redéfinie dans le cas d'une variable aléatoire continue par:

$$M_e = \int_{-\infty}^{M_e} dF(x) = \int_{-\infty}^{M_e} \rho(x) dx = \frac{1}{2} \quad (7.178)$$

et elle coïncide rarement avec la moyenne!

Souvent les statisticiens utilisent les mêmes notations pour l'espérance mathématique d'une variable continue:

$$E(X), M(X), \mu_x, \mu \quad (7.179)$$

et pour la variance:

$$V(X), \sigma_x, \sigma^2 \quad (7.180)$$

que pour une variable discrète.

Par la suite, nous calculerons ces différents termes avec développements uniquement dans les cas les plus usités.

3.3. POSTULAT FONDAMENTAL DE LA STATISTIQUE

Le but ultime de la statistique est de remonter de l'échantillon à la fonction de répartition analytique qui lui aurait donné naissance. Ce but sera présenté dans le cadre de ce site internet comme un postulat (bien que cela postulat soit très difficile à appliquer dans la pratique).

Postulat: À toute fonction de répartition empirique $\hat{F}_n(x)$ nous pouvons associer une fonction de répartition théorique $F(x)$ vers laquelle elle converge quand la taille de l'échantillon est suffisamment grande.

Si:

$$X_n = \sup \left| \hat{F}_n(x) - F(x) \right| \quad (7.181)$$

est la variable aléatoire définie comme la plus grande différence (en valeur absolue) entre $\hat{F}_n(x)$ et

$F(x)$ (observée pour toutes les valeurs de x pour un échantillon donné), alors X_n converge vers 0 presque sûrement.

Remarque: Les mathématiciens de la statistique arrivent à démontrer ce postulat sous forme d'un théorème appelé le "**théorème fondamental de la statistique**" ou "**théorème de Glivenko-Cantelli**" en ce qui concerne les fonctions continues. Personnellement, quitte à choquer les connaisseurs, je considère que cette démonstration n'en est pas une car elle est très éloignée de ce que montre l'expérience (oui c'est mon côté physicien qui ressort...) et ce résultat théorique amène un grand nombre de praticiens à faire souvent tout leur possible (exclusion de données, transformations et autres abominations) pour trouver une loi connue à laquelle ils peuvent ajuster leurs données mesurées.

4. INDICE DE DIVERSITÉ

Il arrive dans le domaine de la biologie ou de l'entreprise que l'on demande à un statisticien ou analyste de mesurer la diversité d'un certain nombre d'éléments prédéfinis. Par exemple, imaginons une multinationale ayant une gamme de produits bien définie et dont certains magasins (clients) dans le monde peuvent choisir un sous-ensemble de cette gamme pour leur commerce. La question étant alors de faire un ranking des magasins qui vendent la plus grande diversité de produits de la marque et ce en prenant en compte aussi les quantités.

Par exemple, nous avons une liste de 4 produits au total dans notre catalogue. Le hasard faisant, trois de nos clients vendent nos 4 produits mais nous souhaiterions savoir lequel en vend la plus grande diversité et ce en prenant en compte les quantités.

Nous avons les données de ventes par produit suivantes pour le client 1:

Client 1	
Produit 1	5
Produit 2	5
Produit 3	5
Produit 4	5

pour le client 2:

Client 2	
Produit 1	1
Produit 2	1
Produit 3	1
Produit 4	17

et pour le client 3:

Client 3	
Produit 1	2
Produit 2	2
Produit 3	2
Produit 4	34

Une mesure de l'information (diversité des états) qui peut être bien adaptée à cette mesure est la formule de Shannon introduite dans le chapitre de Mécanique Statistique dont l'espérance est:

$$S(X) = E\{h(x)\} = -\lambda \sum_{i=1}^n p_i \log(p_i) \quad (7.182)$$

Arbitrairement, nous prendrons $\lambda = 1$ et la base 10 pour le logarithme (ainsi, si nous avons 10 variables équiprobables, l'entropie sera unitaire par exemple...).

Dès lors il vient:

$$S = -\sum_{i=1}^n p_i \log(p_i) \quad (7.183)$$

Nous allons récrire cela de manière plus adéquate pour l'application en entreprise. Ainsi, si n est le nombre de produits et p_i est la proportion (ou "fréquence relative") de ventes du produit i parmi la totalité des ventes N nous avons alors:

$$p_i = \frac{f_i}{N} \quad (7.184)$$

Il vient alors:

$$\begin{aligned} S &= -\sum_{i=1}^n \frac{f_i}{N} \log\left(\frac{f_i}{N}\right) = -\frac{1}{N} \sum_{i=1}^n f_i [\log(f_i) - \log(N)] \\ &= \frac{\sum_{i=1}^n f_i [-\log(f_i) + \log(N)]}{N} = \frac{\log(N) \sum_{i=1}^n f_i - \sum_{i=1}^n f_i \log(f_i)}{N} \quad (7.185) \\ &= \frac{N \log(N) - \sum_{i=1}^n f_i \log(f_i)}{N} \end{aligned}$$

Nous avons alors pour le client 1:

$$\begin{aligned}
 S_1 &= \frac{N \log(N) - \sum_{i=1}^n f_i \log(f_i)}{N} \\
 &= \frac{20 \log(20) - (5 \log(5) + 5 \log(5) + 5 \log(5) + 5 \log(5))}{20} \quad (7.186) \\
 &= \frac{20 \log(20) - 20 \log(5)}{20} = \log(20) - \log(5) = \log(4) = 0.602
 \end{aligned}$$

qui est la valeur maximale possible (chaque état est équiprobable). Et pour le client 2 nous avons:

$$\begin{aligned}
 S_2 &= \frac{N \log(N) - \sum_{i=1}^n f_i \log(f_i)}{N} \quad (7.187) \\
 &= \frac{20 \log(20) - (1 \log(1) + 1 \log(1) + 1 \log(1) + 17 \log(17))}{20} = 0.255
 \end{aligned}$$

et pour le client 3:

$$\begin{aligned}
 S_2 &= \frac{N \log(N) - \sum_{i=1}^n f_i \log(f_i)}{N} \quad (7.188) \\
 &= \frac{40 \log(40) - (2 \log(2) + 2 \log(2) + 2 \log(2) + 34 \log(34))}{40} = 0.255
 \end{aligned}$$

Ainsi, le client ayant la plus grande diversité est le premier. Nous voyons aussi une propriété intéressante de la formule de Shannon à l'aide des clients 2 et 3 c'est que la quantité n'influe pas sur la diversité (puisque la seule différence entre les deux clients est la quantité qui est multipliée d'un facteur 2 et non la diversité)!

5. FONCTIONS DE DISTRIBUTIONS

Lorsque nous observons des phénomènes probabilistes, et que nous prenons note des valeurs prises par ces derniers et que nous les reportons graphiquement, nous observons toujours que les différentes mesures obtenues suivent une caractéristique courbe ou droite typique fréquemment reproductible.

Dans le domaine des probabilités et statistiques, nous appelons ces caractéristiques des "**fonctions de distribution**" car elles indiquent la fréquence avec laquelle la variable aléatoire apparaît avec certaines valeurs.

Remarque: Nous utilisons aussi simplement le terme "fonction" ou encore "loi" pour désigner ces caractéristiques.

Ces fonctions sont en pratique bornées par ce que nous appelons "**l'étendue de la distribution**", ou "**dispersion de la distribution**", qui correspond à la différence entre la donnée maximale (à droite) et la donnée minimale (à gauche) des valeurs observées:

$$E = \text{Max} - \text{Min} \quad (7.189)$$

notée souvent aussi R dans l'ingénierie de la qualité (cf. chapitre de Génie Industriel).

Si les valeurs observées se distribuent d'une certaine manière c'est qu'elles ont alors une probabilité d'avoir une certaine valeur de la fonction de distribution.

Dans la pratique industrielle (cf. chapitre de Génie Industriel), la dispersion des valeurs statistiques est importante parce qu'elle donne une indication sur la variation d'un processus (variabilité).

Définitions:

D1. La relation mathématique qui donne la probabilité qu'a une variable aléatoire d'avoir une valeur donnée de la fonction de distribution est appelée "**fonction de densité**", "**fonction de masse**" ou encore "**fonction marginale**".

D2. La relation mathématique qui donne la probabilité cumulée qu'a une variable aléatoire d'être inférieure ou égale à une certaine valeur est nommée la "**fonction de répartition**" ou "**fonction cumulée**".

D3. Des variables aléatoires sont dites "**indépendantes et identiquement distribuées**" (i.i.d.) si elles suivent toutes la même fonction de distribution et qu'elles sont indépendantes...

Remarque: Le lecteur pourra trouver la fonction de distribution de Weibull (ou "loi de Weibull") dans le chapitre traitant du Génie Industriel (section sur l'Ingénierie), et la fonction logistique dans le chapitre de Méthodes Numériques.

De telles fonctions étant très nombreuses dans la nature, nous proposons au lecteur une étude détaillée des plus connues seulement.

5.1. FONCTION DISCRÈTE UNIFORME

Si nous admettons qu'il est possible d'associer une probabilité à un événement, nous pouvons concevoir des situations où nous pouvons supposer a priori que tous les événements élémentaires sont équiprobables (c'est-à-dire qu'ils ont même probabilité). Nous utilisons alors le rapport entre le nombre de cas favorables et le nombre de cas possibles pour calculer la probabilité de tous les événements de l'Univers des événements U . Plus généralement si U est un ensemble fini d'événements équiprobables et A une partie de U nous avons sous forme ensembliste:

$$P(A) = \text{Card}(A) / \text{Card}(U) \quad (7.190)$$

Plus communément, soit e un événement pouvant avoir N issues équiprobables possibles. Alors la probabilité d'observer l'issue donnée de l'événement suit une "**fonction discrète uniforme**" (ou "**loi discrète uniforme**") donnée par la relation:

$$P_e = 1 / N \quad (7.191)$$

Ayant pour espérance (ou moyenne):

$$\mu = E(X) = \sum_i p_i x_i = P_e \sum_i x_i \quad (7.192)$$

Si nous nous mettons dans le cas particulier où $x_i = i$ avec $i = 1 \dots N$. Nous avons alors (cf. chapitre de Suites et Séries):

$$E(X) = \frac{1}{N} \sum_{i=1}^N i = \frac{1}{N} \frac{N(N+1)}{2} = \frac{N+1}{2} \quad (7.193)$$

Et pour variance:

$$\begin{aligned} V(X) &= \sum_{i=1}^N \frac{1}{N} (i - \mu)^2 = \frac{1}{N} \sum_{i=1}^N (i - \mu)^2 = \frac{1}{N} \left(\sum_{i=1}^N i^2 - 2\mu \sum_{i=1}^N i + \sum_{i=1}^N \mu^2 \right) \\ &= \frac{1}{N} \left(\sum_{i=1}^N i^2 - 2\mu \sum_{i=1}^N i + N\mu^2 \right) = \frac{1}{N} \sum_{i=1}^N i^2 - \frac{2}{N} \mu \sum_{i=1}^N i + \mu^2 \\ &= \frac{1}{N} \left(\frac{N(N+1)(2N+1)}{6} \right) - \frac{2}{N} \frac{N+1}{2} \sum_{i=1}^N i + \left(\frac{N+1}{2} \right)^2 \\ &= \frac{(N+1)(2N+1)}{6} - \frac{N+1}{N} \sum_{i=1}^N i + \frac{(N+1)^2}{4} \\ &= \frac{(N+1)(2N+1)}{6} - \frac{N+1}{N} \left(\frac{N(N+1)}{2} \right) + \frac{(N+1)^2}{4} \quad (7.194) \\ &= \frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{2} + \frac{(N+1)^2}{4} = \frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4} \\ &= \frac{2(N+1)(2N+1)}{12} - \frac{3(N+1)^2}{12} = \frac{2(N+1)(2N+1) - 3(N+1)^2}{12} \\ &= \frac{2(2N^2 + N + 2N + 1) - 3N^2 - 6N - 3}{12} = \frac{4N^2 + 6N + 2 - 3N^2 - 6N - 3}{12} \\ &= \frac{N^2 - 1}{12} \end{aligned}$$

Voici un exemple de tracé de la fonction de distribution et respectivement de répartition pour la loi discrète uniforme de paramètres $\{1,5,8,11,12\}$ (nous voyons que chaque valeur est bien équiprobable):

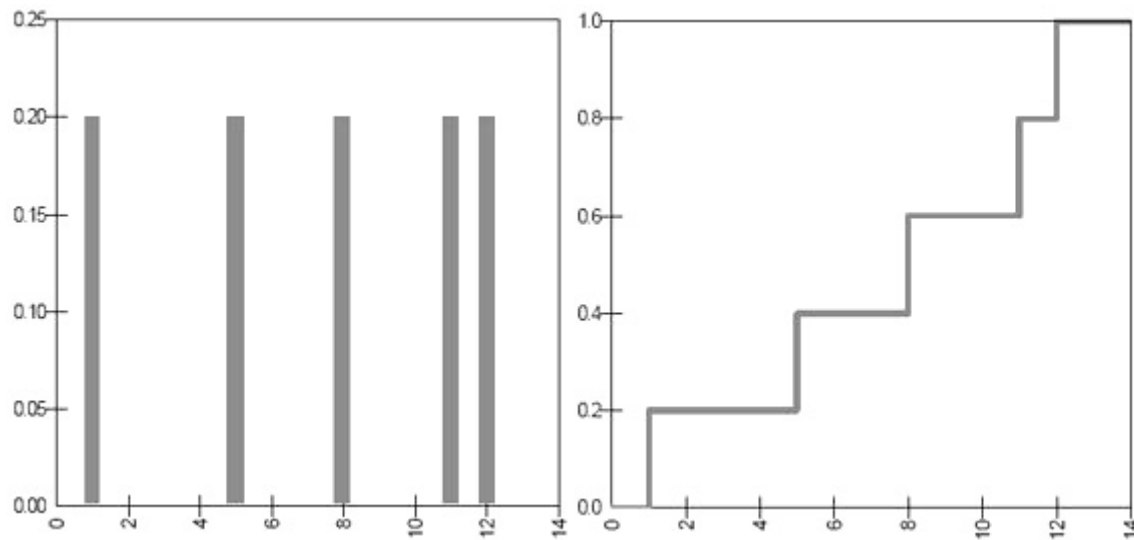


Figure: 7.8 - Loi discrète uniforme (fonction de distribution et de répartition)

5.2. FONCTION DE BERNOULLI

Si nous avons affaire à une observation binaire alors la probabilité d'un événement reste constante d'une observation à l'autre s'il n'y a pas d'effet mémoire (autrement dit: une somme de variables de Bernoulli, deux à deux indépendantes).

Nous appelons ce genre d'observations où la variable aléatoire a valeurs 0 (faux) ou 1 (vrai), avec probabilité $(1-p)$, p respectivement, des "**essais de Bernoulli**" avec "**événements contraires à probabilités contraires**".

Ainsi, une variable aléatoire X suit une "**fonction de Bernoulli**" (ou "**loi de Bernoulli**") si elle ne peut prendre que les valeurs 0 ou 1, associées aux probabilités q et p de sorte que $q + p = 1$ et:

$$P(X=0) = q \quad P(X=1) = p = (1-q) \quad (7.195)$$

L'exemple classique d'un tel processus est le jeu de pile ou face ou de tirage avec remise. Il est inutile de vérifier formellement que la probabilité cumulée est unitaire...

Remarquons que par extension, si nous considérons N événements où nous obtenons dans un ordre particulier k fois une des issues possible (réussite) et $N-k$ l'autre (échec), alors la probabilité d'obtenir une telle série (de k réussites et $N-k$ échecs ordonnés dans un ordre particulier) sera donnée par:

$$P = p^k \cdot (1-p)^{N-k} = p^k \cdot q^{N-k} \quad (7.196)$$

conformément à ce que nous avons obtenu en combinatoire dans le chapitre de Probabilités!

Voici un exemple de tracé de la fonction de répartition pour $q = 0.3$:

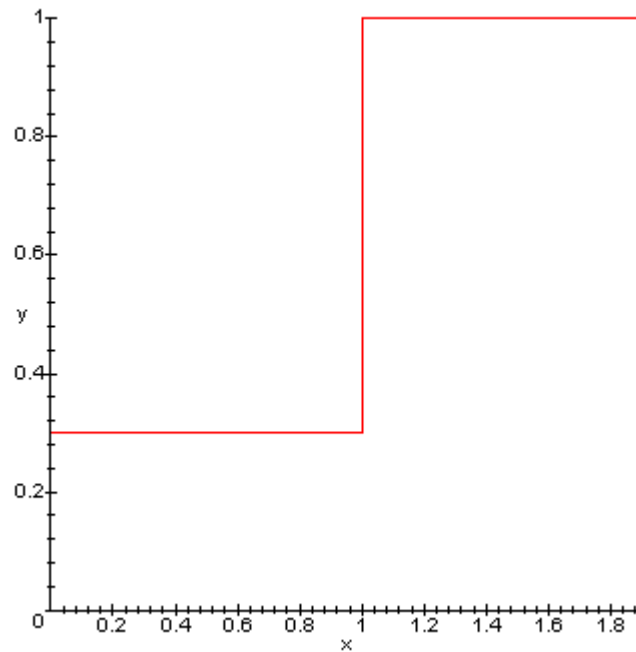


Figure: 7.9 - Loi de Bernoulli (fonction de distribution et de répartition)

La fonction de Bernoulli a donc pour espérance (moyenne):

$$\mu = E(X) = \sum p_i \cdot x_i = 1 \cdot p + 0 \cdot (1 - p) = p \quad (7.197)$$

et pour variance (nous utilisons la relation de Huyghens démontrée plus haut):

$$V(X) = \sigma^2 = E[X - E(X)]^2 = E(X^2) - E(X)^2 = E(X^2) - p^2 = p - p^2 = p(1 - p) = p \cdot q \quad (7.198)$$

Remarque: L'exemple ci-dessus n'est certes pas pertinent mais nous verrons dans le chapitre de Techniques De Gestion que la fonction de Bernoulli apparaît naturellement au début de notre étude des files d'attentes.

5.3. FONCTION GÉOMÉTRIQUE

La loi géométrique ou "**loi de Pascal**" consiste dans une épreuve de type Bernoulli, dont la probabilité de succès est p et celle d'échec $q = 1 - p$ sont constantes, que nous renouvelons de manière indépendante jusqu'au premier succès.

Si nous appelons X la variable aléatoire donnant le rang du premier succès la probabilité que $X = k$ est alors (cas particulier de la fonction de Bernoulli):

$$P(k) = q^{k-1} p \quad (7.199)$$

avec $k \in \mathbb{N}^*$.

Cette loi a pour espérance:

$$\mu = E(X) = \sum p_i x_i = \sum_k k q^{k-1} p = \sum_k k (1-p)^{k-1} p = p \sum_k k (1-p)^{k-1} \quad (7.200)$$

Or, cette dernière relation s'écrit aussi (car c'est une simple série géométrique):

$$\sum_{k=1}^{\infty} k q^{k-1} = \frac{1}{(1-q)^2} \quad (7.201)$$

Effectivement, nous avons démontré dans le chapitre sur les Suites et Séries que:

$$\sum_{k=0}^n q^k = \frac{1-q^{n+1}}{1-q} \quad (7.202)$$

En prenant la limite lorsque $n \rightarrow +\infty$ nous obtenons:

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q} \quad (7.203)$$

car $0 \leq q < 1$.

Ensuite, il suffit de dériver les deux membres de l'égalité par rapport à q et nous obtenons:

$$\sum_{k=1}^{\infty} k q^{k-1} = \frac{1}{(1-q)^2} \quad (7.204)$$

Nous avons donc le nombre moyen d'essais X qu'il faut faire pour arriver au premier succès:

$$E(X) = \sum_{k=0}^{\infty} k \cdot P(X=k) = \sum_{k=0}^{\infty} k p q^{k-1} = \frac{p}{(1-q)^2} = \frac{1}{p} \quad (7.205)$$

Calculons maintenant la variance en rappelant comme à chaque fois que (relation de Huyghens):

$$V(X) = E(X^2) - E(X)^2 \quad (7.206)$$

Commençons donc par calculer $E(X^2)$:

$$E(X^2) = \sum_{k=0}^{\infty} k^2 P(X=k) = p \sum_{k=0}^{\infty} k^2 q^{k-1} = p \sum_{k=0}^{\infty} k \cdot (k-1+1) q^{k-1} = p \sum_{k=1}^{\infty} k \cdot (k-1) q^{k-1} + p \sum_{k=0}^{\infty} k q^{k-1} \quad (7.207)$$

Le dernier terme de cette expression est l'équivalent de l'espérance calculée précédemment. Soit:

$$p \sum_{k=0}^{\infty} k q^{k-1} = \frac{1}{p} \quad (7.208)$$

Il reste à calculer:

$$p \sum_{k=1}^{\infty} k \cdot (k-1) q^{k-1} \quad (7.209)$$

Nous avons:

$$p \sum_{k=1}^{\infty} k \cdot (k-1) q^{k-1} = pq \sum_{k=2}^{\infty} k \cdot (k-1) q^{k-2} \quad (7.210)$$

Or en dérivant l'égalité:

$$\sum_{k=1}^{\infty} k q^{k-1} = \frac{1}{(1-q)^2} \quad (7.211)$$

Nous obtenons:

$$\sum_{k=1}^{\infty} k(k-1) q^{k-2} = \frac{2}{(1-q)^3} \quad (7.212)$$

Par conséquent:

$$p \sum_{k=1}^{\infty} k \cdot (k-1) q^{k-1} = pq \sum_{k=2}^{\infty} k \cdot (k-1) q^{k-2} = \frac{2pq}{(1-q)^3} = \frac{2q}{p^2} \quad (7.213)$$

Donc:

$$E(X^2) = \frac{2q}{p^2} + \frac{1}{p} \quad (7.214)$$

Pour finir:

$$V(X) = E(X^2) - E(X)^2 = \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2} \quad (7.215)$$

Exemple:

E1. Vous essayez, tard dans la nuit et dans l'obscurité, d'ouvrir une serrure au moyen d'un trousseau de 5 clés, sans porter attention, car vous êtes un peu fatigué (ou un peu éméché...) vous essayez chaque clé. Sachant qu'une seule convient, quelle est la probabilité d'utiliser la bonne clé au k -ème essai?

$$P(k) = \left(\frac{4}{5}\right)^{k-1} \left(\frac{1}{5}\right) \quad (7.216)$$

E2. Tracé de la fonction de distribution et répartition pour la fonction Géométrique de paramètre $p = 0.5$:

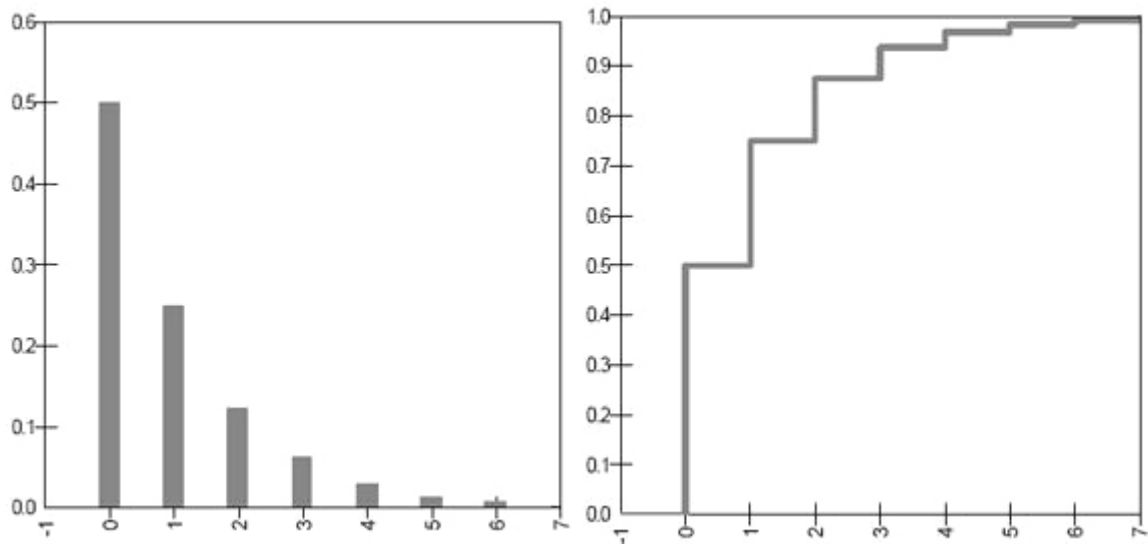


Figure: 7.10 - Loi géométrique (fonction de distribution et de répartition)

Déterminons maintenant la fonction de répartition de la loi Géométrique. Nous partons donc de:

$$P(k) = q^{k-1}p \quad (7.217)$$

nous avons alors par définition la probabilité cumulée que l'expérience réussisse dans les n premiers essais:

$$P(X \leq n) = 1 - \sum_{j=n+1}^{+\infty} pq^{j-1} = 1 - p \sum_{j=n+1}^{+\infty} q^{j-1} \quad (7.218)$$

avec n entier valant 0...1...2, etc.

Posons:

$$j-1 = n+k \Rightarrow k = n-j+1 \quad (7.219)$$

Nous avons alors:

$$\begin{aligned} P(X \leq n) &= 1 - p \sum_{k=0}^{+\infty} q^{n+k} = 1 - p \sum_{k=0}^{+\infty} q^n q^k = 1 - pq^n \sum_{k=0}^{+\infty} q^k = 1 - pq^n \frac{1}{1-q} \\ &= 1 - (1-q)q^n \frac{1}{1-q} = 1 - q^n \end{aligned} \quad (7.220)$$

5.4. FONCTION BINOMIALE

Si nous revenons maintenant à notre épreuve de Bernoulli. Plus généralement, tout N -uplet particulier formé de k succès et de $N-k$ échecs aura pour probabilité (dans le cadre d'un tirage avec remise ou sans remise si la population est grande en première approximation...):

$$P(N, k) = p^k \cdot q^{N-k} \quad (7.221)$$

d'être tiré (ou d'apparaître) quel que soit l'ordre d'apparition des échecs et réussites.

Mais, nous savons que la combinatoire permet de déterminer le nombre de N -uplets de ce type (le

nombre de manières d'ordonner les apparitions d'échecs et de réussites). Le nombre d'arrangements possibles étant, nous l'avons démontré (cf. chapitre Probabilités), donné par le coefficient binomial (notation non conforme sur ce site à la norme ISO 31-11):

$$C_k^N = \frac{N!}{k! (N-k)!} \quad (7.222)$$

Donc comme la probabilité d'obtenir une série de k succès et $N-k$ échecs particuliers est toujours identique (quel que soit l'ordre) alors il suffit de multiplier la probabilité d'une série particulière par la combinatoire (ceci étant équivalent à faire une somme):

$$P(N, k) = C_k^N \cdot P = \frac{N!}{k! (N-k)!} \cdot p^k \cdot q^{N-k} \quad (7.223)$$

pour avoir la probabilité totale d'obtenir une quelconque de ces séries possibles (puisque chacune est possible).

Remarque: Cela équivaut à l'étude d'un tirage avec remise (cf. chapitre de Probabilités) simple avec contrainte sur l'ordre ou à l'étude d'une série de succès ou d'échecs. Nous utiliserons cette relation dans le cadre de la théorie des files d'attente ou en fiabilité. Il faut noter que dans le cas de grandes populations, même si le tirage n'est pas avec remise il peut être considéré comme tel...

Écrite autrement ceci donne la "**fonction Binomiale**" (ou "**loi Binomiale**") connue aussi sous la forme de la fonction de distribution suivante:

$$P(N, k) = C_k^N p^k (1-p)^{N-k} = C_k^N p^k q^{N-k} \quad (7.224)$$

et parfois notée:

$$X \rightarrow \beta(n, p) \quad (7.225)$$

avec un petit n ou grand N (cela importe peu...) et peut être calculée dans la version française de Microsoft Excel 11.8346 à l'aide de la fonction LOI.BINOMIALE().

Nous disons parfois que la loi Binomiale est non exhaustive car la taille de la population initiale n'est pas apparente dans l'expression de la loi.

Exemple:

Nous souhaitons tester l'alternateur d'un groupe électrogène. La probabilité de défaillance à la sollicitation de ce matériel est estimée à 1 défaillance pour 1'000 démarrages.

Nous décidons d'effectuer un test de 100 démarrages. La probabilité d'observer 1 panne au cours de ce test est de:

$$\begin{aligned}
 P(N=100, k=1) &= C_k^N p^k q^{N-k} = \frac{N!}{k!(N-k)!} p^k q^{N-k} \\
 &= \frac{100!}{1!(100-1)!} \left(\frac{1}{1000}\right)^1 \left(1 - \frac{1}{1000}\right)^{99} \cong 9\%
 \end{aligned}
 \tag{7.226}$$

Nous avons bien évidemment pour la fonction de répartition (très utile dans la pratique comme le contrôle de lots de fournisseurs ou la fiabilité!):

$$\sum_{k=0}^N C_k^N \cdot p^k \cdot q^{N-k} = 1 \tag{7.227}$$

Effectivement, nous avons démontré dans le chapitre de Calcul Algébrique le "**théorème binomial**":

$$(x+y)^n = \sum_{k=0}^n C_k^n x^k y^{n-k} \tag{7.228}$$

Donc:

$$\sum_{k=0}^N C_k^N p^k (1-p)^{N-k} = (p + (1-p))^N = 1^N = 1 \tag{7.229}$$

Il vaut mieux utiliser Microsoft Excel 11.8346 (ou tout autre logiciel largement répandu) pour ne pas s'embêter à calculer ce genre de relations en utilisant la fonction CRITERE.LOI.BINOMIALE() dans la version française.

L'espérance mathématique (moyenne) de $P(N,k)$ est:

$$E(X) = \sum_{k=0}^n P(X=k)k = \sum_{k=0}^n C_k^N p^k (1-p)^{N-k} k = \sum_{k=1}^n C_k^N p^k (1-p)^{N-k} k \tag{7.230}$$

Or:

$$C_k^N = \frac{N}{k} C_{k-1}^{N-1} \Leftrightarrow k C_k^N = N C_{k-1}^{N-1} \tag{7.231}$$

d'où:

$$\begin{aligned}
 E(X) &= N \sum_{k=1}^N C_{k-1}^{N-1} p^k (1-p)^{N-k} = Np \sum_{k=1}^N C_{k-1}^{N-1} p^{k-1} (1-p)^{(N-1)-(k-1)} \\
 &= Np \sum_{k=0}^{N-1} C_k^{N-1} p^k (1-p)^{(N-1)-k} = Np (p + (1-p))^{N-1} = Np
 \end{aligned}
 \tag{7.232}$$

donne le nombre moyen de fois que l'on obtiendra l'issue souhaitée de probabilité p après N essais.

Avant de calculer la variance, introduisons la relation suivante:

$$\sum_{s=0}^{N-1} s \frac{(N-1)!}{s!(N-1-s)!} p^s q^{N-1-s} = (N-1)p \quad (7.233)$$

En effet, en utilisant les développements précédents:

$$\begin{aligned} \sum_{s=0}^{N-1} s \frac{(N-1)!}{s!(N-1-s)!} p^s q^{N-1-s} &= \sum_{s=1}^{N-1} \frac{(N-1)!}{(s-1)!(N-1-s)!} p^s q^{N-1-s} \\ &= (N-1) \sum_{s=1}^{N-1} \frac{(N-2)!}{(s-1)!(N-1-s)!} p^s q^{N-1-s} = (N-1) \sum_{j=0}^{N-2} \frac{(N-2)!}{j!(N-2-j)!} p^{j+1} q^{N-2-j} \quad (7.234) \\ &= (N-1)p \sum_{j=0}^{N-2} C_j^{N-2} p^j q^{N-2-j} = (N-1)p \left(\underbrace{p+q}_{=1} \right)^{N-2} = (N-1)p \end{aligned}$$

Commençons maintenant le (long) calcul de la variance dans lequel nous allons utiliser les résultats précédents:

$$\begin{aligned}
V(X) &= \sum_{k=0}^N \frac{N!}{k!(N-k)!} p^k q^{N-k} \cdot [k - N \cdot p]^2 = \sum_{k=0}^N \frac{N!}{k!(N-k)!} p^k q^{N-k} \cdot [k - \mu]^2 \\
&= \sum_{k=0}^N k^2 \frac{N!}{k!(N-k)!} p^k q^{N-k} + \sum_{k=0}^N \mu^2 \frac{N!}{k!(N-k)!} p^k q^{N-k} - \sum_{k=0}^N 2k\mu \frac{N!}{k!(N-k)!} p^k q^{N-k} \\
&= \sum_{k=0}^N k^2 \frac{N!}{k!(N-k)!} p^k q^{N-k} + \mu^2 \sum_{k=0}^N \frac{N!}{k!(N-k)!} p^k q^{N-k} - 2\mu \sum_{k=0}^N k \frac{N!}{k!(N-k)!} p^k q^{N-k} \\
&= \sum_{k=0}^N k^2 \frac{N!}{k!(N-k)!} p^k q^{N-k} + \mu^2 \underbrace{\sum_{k=0}^N C_k^N p^k q^{N-k}}_{=1} - 2\mu \sum_{k=0}^N k C_k^N p^k q^{N-k} \\
&= \sum_{k=0}^N k^2 \frac{N!}{k!(N-k)!} p^k q^{N-k} + \mu^2 - 2\mu \underbrace{N \sum_{k=1}^N C_{k-1}^{N-1} p^k q^{N-k}}_{=Np=\mu} \\
&= \sum_{k=0}^N k^2 \frac{N!}{k!(N-k)!} p^k q^{N-k} + \mu^2 - 2\mu^2 = \sum_{k=0}^N k^2 \frac{N!}{k!(N-k)!} p^k q^{N-k} - \mu^2 \\
&= -\mu^2 + N \sum_{k=1}^N k \frac{(N-1)!}{(k-1)!(N-k)!} p^k q^{N-k} \underset{s=k-1}{=} -\mu^2 + N \sum_{s=0}^{N-1} (s+1) \frac{(N-1)!}{s!(N-1-s)!} p^{s+1} q^{N-(s+1)} \\
&= -\mu^2 + Np \sum_{s=0}^{N-1} (s+1) \frac{(N-1)!}{s!(N-1-s)!} p^s q^{N-s-1} \\
&= -\mu^2 + Np \sum_{s=0}^{N-1} s \frac{(N-1)!}{s!(N-1-s)!} p^s q^{(N-1)-s} + Np \sum_{s=0}^{N-1} \frac{(N-1)!}{s!(N-1-s)!} p^s q^{(N-1)-s} \\
&= -\mu^2 + Np \sum_{s=0}^{N-1} \underbrace{C_s^{N-1}}_{C_s^{N-1}} p^s q^{(N-1)-s} s + Np \sum_{s=0}^{N-1} \underbrace{C_s^{N-1}}_{C_s^{N-1}} p^s q^{(N-1)-s} \\
&= -\mu^2 + Np \sum_{s=0}^{N-1} C_s^{N-1} p^s q^{(N-1)-s} s + Np \sum_{s=0}^{N-1} C_s^{N-1} p^s q^{(N-1)-s} \\
&= -\mu^2 + Np(N-1)p + Np \sum_{s=0}^{N-1} C_s^{N-1} p^s q^{N-1-s} = -\mu^2 + Np^2(N-1) + Np(p + (1-p))^{N-1} \\
&= -\mu^2 + Np^2(N-1) + Np = -\mu^2 + N^2 p^2 - Np^2 + Np = -\mu^2 + \mu^2 - Np^2 + Np \\
&= Np(1-p) = Npq
\end{aligned}$$

(7.235)

L'écart-type étant $\sigma^2 = V(k)$, nous avons:

$$\sigma = \sqrt{N \cdot p(1-p)} \quad (7.236)$$

Voici un exemple de tracé de la fonction de distribution et respectivement de répartition de la loi binomiale $\beta(10, 0.5)$:

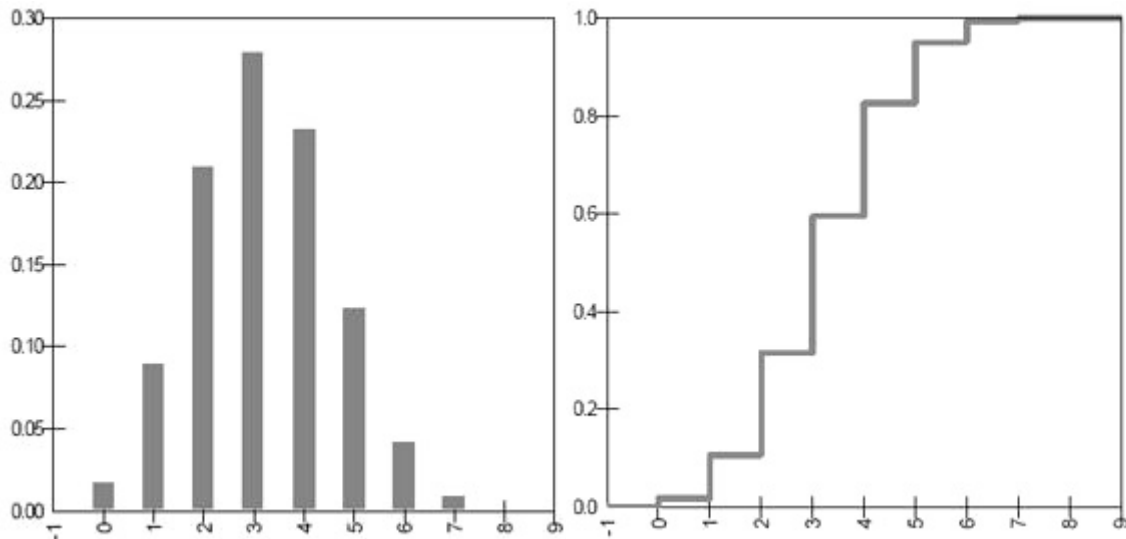


Figure: 7.11 - Loi binomiale (fonction de distribution et de répartition)

Indiquons que certaines personnes dans les entreprises ramènent le calcul de l'espérance et de l'écart-type à l'unité de N . Nous avons alors:

$$E\left(\frac{X}{N}\right) = \frac{Np}{N} = p$$

$$V\left(\frac{X}{N}\right) = \frac{1}{N^2} V(X) = \frac{1}{N^2} Npq = \frac{1}{N} p(1-p) \Leftrightarrow \sigma = \sqrt{\frac{p(1-p)}{N}} \quad (7.237)$$

Exemple:

Sur un échantillon de 100 travailleurs, 25% sont en retard au moins une fois par semaine. L'espérance du nombre de retard est alors:

$$E(X) = Np = 100 \cdot 0.25 = 25$$

$$\sigma(X) = \sqrt{Np(1-p)} = 4.33 \quad (7.238)$$

Rapporté à l'unité, cela nous donne:

$$E\left(\frac{X}{N}\right) = 0.25$$

$$\sigma\left(\frac{X}{N}\right) = \sqrt{\frac{p(1-p)}{N}} = 0.0433 \quad (7.239)$$

Pour clore concernant notre étude de loi binomiale, nous allons développer un résultat qui nous sera indispensable pour construire le test de données appariées de McNemar d'un tableau (carré) de contingence (et comme il est carré il est in extenso dichotomique) que nous étudierons dans le chapitre de Méthodes Numériques.

Nous avons besoin pour ce test de calculer la covariance de deux variables aléatoires binomiales appariées (raison pour laquelle la covariance est non nulle):

$$\text{cov}(n_i, n_j) = E(n_i n_j) - E(n_i) E(n_j) \quad (7.240)$$

Comme elles sont appariées, cela signifie que:

$$\begin{aligned} n_i + n_j &= n \\ p_i &= 1 - p_j \\ 1 - p_i &= p_j \end{aligned} \quad (7.241)$$

et donc:

$$\text{cov}(n_i, n_j) = E(n_i n_j) - np_i \cdot np_j = E(n_i n_j) - n^2 p_j (1 - p_j) \quad (7.242)$$

Maintenant, vient la difficulté qui est de calculer $E(n_i n_j)$. Pour calculer ce terme il n'existe pas à notre connaissance d'autres méthodes que de chercher la loi du couple (parfois on peut contourner cela). Dans le cas présent il s'agit d'une loi multinomiale (plus précisément: trinomiale) qu'il est d'usage d'écrire sous la forme:

$$\mathcal{M}(n, p_i, p_j, 1 - p_i - p_j) = \frac{n!}{n_i! n_j! (n - n_i - n_j)!} p_i^{n_i} p_j^{n_j} (1 - p_i - p_j)^{n - n_i - n_j} \quad (7.243)$$

Mais que nous noterons temporairement pour la suite afin de condenser l'écriture:

$$\frac{n!}{k! l! (n - k - l)!} p^k q^l r^{n - k - l} = \binom{n}{k, l} p^k q^l r^{n - k - l} \quad (7.244)$$

Nous avons donc une loi trinomiale car nous avons cherchons le nombre de fois d'avoir l'événement k , l'événement l et ni l'un ni l'autre (donc le reste du temps).

Nous avons alors:

$$\begin{aligned} E(kl) &= \sum_{k,l} kl \cdot P(kl) = \sum_{k,l} kl \cdot \frac{n!}{k! l! (n - k - l)!} p^k q^l r^{n - k - l} \\ &= \underbrace{0}_{k=0, l \neq 0} + \underbrace{0}_{l=0, k \neq 0} + \underbrace{0}_{k=0, l=0} + \sum_{k \geq 1, l \geq 1} kl \cdot \frac{n!}{k! l! (n - k - l)!} p^k q^l r^{n - k - l} \end{aligned} \quad (7.245)$$

Si $k \geq 1$ et $l \geq 1$, nous avons:

$$\begin{aligned} kl \cdot \frac{n!}{k! l! (n - k - l)!} &= kl \cdot \frac{n!}{k(k-1)! l(l-1)! (n - k - l)!} = \frac{n(n-1)(n-2)!}{(k-1)!(l-1)!(n - k - l)!} \\ &= n(n-1) \frac{(n-2)!}{(k-1)!(l-1)!(n - k - l)!} \end{aligned} \quad (7.246)$$

Maintenant utilisons cette relation dans l'espérance conjointe:

$$\begin{aligned}
E(kl) &= \sum_{k \geq 1, l \geq 1} \frac{n!}{k!l!(n-k-l)!} p^k q^l r^{n-k-l} \\
&= n(n-1) \sum_{1 \leq k \leq n, 1 \leq l \leq n} \frac{(n-2)!}{(k-1)!(l-1)!(n-k-l)!} p^k q^l r^{n-k-l} \quad (7.247) \\
&= n(n-1)pq \sum_{1 \leq k \leq n, 1 \leq l \leq n} \frac{(n-2)!}{(k-1)!(l-1)!(n-k-l)!} p^{k-1} q^{l-1} r^{n-k-l}
\end{aligned}$$

Considérons le cas où n vaut 2. Nous avons:

$$\begin{aligned}
&\sum_{1 \leq k \leq n, 1 \leq l \leq n} \frac{(n-2)!}{(k-1)!(l-1)!(n-k-l)!} p^{k-1} q^{l-1} r^{n-k-l} \\
&= \sum_{1 \leq k \leq 2, 1 \leq l \leq 2} \frac{0!}{(k-1)!(l-1)!(2-k-l)!} p^{k-1} q^{l-1} r^{2-k-l} \quad (7.248) \\
&= \frac{0!}{(1-1)!(1-1)!(2-1-1)!} p^{1-1} q^{1-1} r^{2-1-1} = 1
\end{aligned}$$

et pour n valant 3, le résultat sera aussi 1, et ainsi de suite (nous supposons afin de simplifier... que quelques exemples numériques suffiront au lecteur pour le convaincre de la généralité de cette propriété). Nous avons alors:

$$E(kl) = n(n-1)pq \underbrace{\sum_{1 \leq k \leq n, 1 \leq l \leq n} \frac{(n-2)!}{(k-1)!(l-1)!(n-k-l)!} p^{k-1} q^{l-1} r^{n-k-l}}_{=1} = n(n-1)pq \quad (7.249)$$

Donc au final:

$$\begin{aligned}
\text{cov}(n_i, n_j) &= E(n_i n_j) - n^2 p_j (1 - p_j) = n(n-1)p_i p_j - n^2 p_i p_j \quad (7.250) \\
&= -n p_i p_j
\end{aligned}$$

5.5. FONCTION BINOMIALE NÉGATIVE

La loi binomiale négative s'applique dans la même situation que la loi binomiale mais elle donne la probabilité d'avoir E échecs avant la R -ème réussite quand la probabilité de succès est p (ou inversement la probabilité d'avoir R réussites avant le E -ème échec quand la probabilité d'échec est p).

Introduisons cette fonction par l'exemple. Considérons pour cela les probabilités suivantes:

$$P(\text{succès}) = 0.2 = p \quad P(\text{échec}) = 0.8 = 1 - p = q \quad (7.251)$$

Imaginons que nous ayons fait 10 essais et que nous voulions nous arrêter à la troisième réussite et que le 10-ème essai est la troisième réussite! Nous allons noter cela:

$$[1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9] \ 10 \quad (7.252)$$

Mettons en évidence les réussites (R) et échecs (E):

$$\begin{array}{cccccccccc} [1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9] & 10 \\ [E & E & R & E & E & E & R & E & E] & R \end{array} \quad (7.253)$$

Nous avons donc 7 échecs et 3 réussites. Dans le cadre d'une expérience où les tirages sont indépendants, la probabilité que nous avons d'obtenir ce résultat particulier est alors:

$$(0.8)^7 (0.2)^3 \quad (7.254)$$

Mais l'ordre des succès et échecs dans la partie entre crochets n'a aucune importance. Donc comme nous avons 2 succès parmi 9 dans les crochets il vient que la probabilité d'obtenir le même résultat indépendamment de l'ordre est alors en utilisant la combinatoire:

$$\binom{9}{2} (0.8)^7 (0.2)^3 \cong 0.0603 \quad (7.255)$$

ce qui correspond donc à la probabilité d'avoir 7 échecs avant la 3ème réussite. Ce qui s'écrit avec Microsot Excel 14.0.6123 ou ultérieur en français (7+3=10 essais, 7 échecs dont 3 réussites):

$$=LOI.BINOMIALE.NEG.N(7;3;0.2;0)=0.0604$$

Généralisons l'écriture antérieure notant k le nombre d'échecs, N le nombre total d'essais et p la probabilité d'une réussite:

$$P(k, N, p) = \binom{N-1}{N-k-1} q^k p^{N-k} \quad (7.256)$$

Il y a plusieurs écritures possibles cependant car la relation précédente n'est pas très intuitive à mettre en pratique comme l'aura peut-être remarqué le lecteur. Ainsi, si nous notons k comme étant le nombre de succès et non le nombre d'échecs, nous avons alors (écriture la plus courante selon moi parmi d'autres équivalentes) la probabilité suivante d'avoir un certain nombre de réussites avant d'avoir un nombre k d'échecs:

$$P(k, N, p) = \binom{N-1}{k-1} q^{N-k} p^k = C_{k-1}^{N-1} p^k q^{N-k} \quad (7.257)$$

donc la comparaison avec la formulation de la loi binomiale démontrée plus haut est alors probablement évidente!

Il est cependant plus courant de noter la relation précédente en faisant disparaître N car pour l'instant l'écriture n'est toujours pas très claire. Pour cela, nous notons R le nombre de réussites, E le nombre d'échecs, p la probabilité d'une réussite et il vient alors la probabilité d'avoir R réussites après E échecs (c'est beaucoup plus clair...):

$$P(R, E, p) = \binom{E+R-1}{E-1} q^{(E+R)-E} p^E = \binom{E+R-1}{E-1} q^R p^E \quad (7.258)$$

Nous la trouvons aussi parfois sous la forme suivante en utilisant explicitement la combinatoire:

$$\begin{aligned}
 P(E, N, p) &= \binom{E+R-1}{E-1} q^R p^E = \frac{(E+R-1)!}{(E-1)!(E+R-1-(E-1))!} q^R p^E = \frac{(E+R-1)!}{(E-1)!R!} q^R p^E \\
 &= \frac{(E+R-1)!}{R!(E-1)!} q^R p^E = \binom{E+R-1}{R} q^R p^E
 \end{aligned} \quad (7.259)$$

La probabilité cumulée que nous ayons au moins R réussites avant le E -ème échec vient immédiatement:

$$\sum_{R=0}^N \binom{E+R-1}{E-1} q^R p^E \quad (7.260)$$

Remarque: Le nom de cette loi provient du fait que certains statisticiens utilisent une définition d'un coefficient combinatoire avec valeur négative pour l'expression de la fonction. Comme c'est une forme plutôt rare, nous ne souhaitons pas la démontrer. Il faut savoir aussi que cette loi est aussi connue sous le nom de "**loi de Pascal**" (au même titre que la loi géométrique...) en l'honneur de Blaise Pascal et de "**loi de Pólya**", en l'honneur de George Pólya.

Exemple:

E1. Un contrôle de qualité long terme nous a permis de calculer l'estimateur p des pièces non-conformes comme valant 2% à la sortie d'une ligne de production. Nous souhaiterions savoir la probabilité cumulée d'avoir 200 pièces bonnes avant que la 3ème pièce défectueuse apparaisse. Avec Microsoft Excel 14.0.6123 ou ultérieur en français il vient en utilisant la loi binomiale négative:

$$=LOI.BINOMIALE.NEG.N(200;3;0.02;1)=77.35\%$$

E2. Pour comparer avec la loi binomiale, demandons-nous quelle est la probabilité cumulée de tirer 198 pièces non-défectueuses parmi 201 avec Microsoft Excel 14.0.6123 ou ultérieur en français:

$$=LOI.BINOMIALE.N(198;201;0.98;1)=76.77\%$$

nous voyons donc que la différence est faible. Au fait la différence entre les deux lois est dans la pratique quasiment toujours tellement faible que nous n'utilisons alors que la loi binomiale (mais il faut quand même être prudent!).

Comme à l'habitude, déterminons maintenant la variance et l'espérance de cette loi. Commençons par l'espérance d'avoir R réussites avant le E -ème échec sachant que la probabilité d'avoir un échec est p . Pour cela nous allons utiliser une astuce très simple et géniale (tout l'art était d'y penser...). Si nous reprenons notre exemple de départ:

$$\begin{array}{cccccccccc}
 [1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9] & 10 \\
 [R & R & E & R & R & R & E & R & R] & E
 \end{array} \quad (7.261)$$

et que nous le réécrivons sous la forme suivante:

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ \underbrace{R \ R \ E}_{X_1} & \underbrace{R \ R \ R \ E}_{X_2} & \underbrace{R \ R \ E}_{X_3} \end{bmatrix} \quad (7.262)$$

Nous remarquons alors que la troisième réussite R de la première écriture peut être décomposée en la somme de trois variables aléatoires géométriques telle que:

$$R = X_1 + X_2 + \dots + X_n \quad (7.263)$$

Avec dans le cas du présent exemple particulier $n = 3$ correspondant au fait à $E = 3$. Donc en toute généralité la somme de n variables aléatoires géométriques donne toujours une loi binomiale négative si la probabilité p est égale pour chaque variable géométrique! Bref... comme nous avons démontré l'expression de l'espérance et la variance de la loi Géométrique comme étant:

$$E(X) = \frac{1}{p} \quad V(X) = \frac{1-p}{p^2} = \frac{q}{p^2} \quad (7.264)$$

Puisque les variables aléatoires sont de même paramètres et indépendantes il vient alors pour l'espérance de la loi binomiale négative:

$$\begin{aligned} E(R) &= E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n) \\ &= nE(X) = n \cdot \frac{1}{p} = E \cdot \frac{1}{p} \end{aligned} \quad (7.265)$$

Et donc pour la variance de la loi binomiale négative:

$$\begin{aligned} V(R) &= V(X_1 + X_2 + \dots + X_n) = V(X_1) + V(X_2) + \dots + V(X_n) \\ &= nV(X) = n \frac{1-p}{p^2} = E \frac{q}{p^2} \end{aligned} \quad (7.266)$$

Exemple:

Quelle est l'espérance du nombre de pièces bonnes que nous aurons avant la troisième pièce non-conforme, sachant que la probabilité d'une pièce non-conforme est de 2%?

$$E(R) = \frac{E}{p} = \frac{3}{2\%} = 150 \quad (7.267)$$

et pour la variance:

$$\sigma = \sqrt{E \frac{q}{p^2}} = \sqrt{3 \frac{(1-2\%)}{(2\%)^2}} = 85.732 \quad (7.268)$$

Voici un exemple de tracé de la fonction de distribution et répartition pour la fonction binomiale négative de paramètres $P(R, E, p) = P(3, E, 0.6)$ basé sur l'exemple du début mais avec comme seule différence d'avoir pris probabilité de réussite de 60% au lieu de 20%:

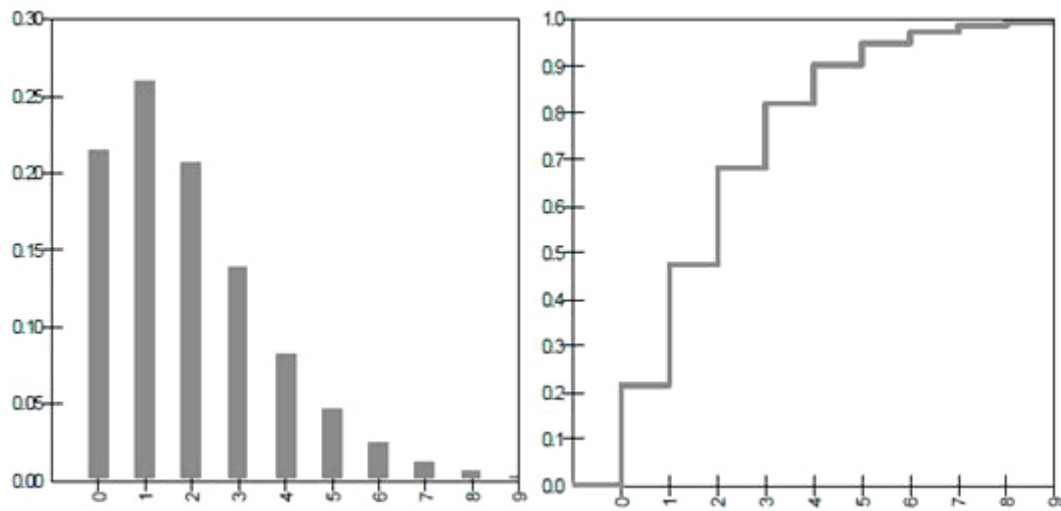


Figure: 7.12 - Loi binomiale négative (fonction de distribution et de répartition)

La distribution ci-dessus est tronquée à 9 mais continue théoriquement à l'infini. Ce qui différencie particulièrement la loi binomiale, géométrique de la loi binomiale négative sont les queues de la distribution.

5.6. FONCTION HYPERGÉOMÉTRIQUE

Nous considérons pour approcher cette fonction un exemple simple (mais guère intéressant dans la pratique) qui est celui d'une urne contenant n boules dont m sont noires et les autres m' blanches (pour plusieurs exemples concrets utilisés dans l'industrie se reporter au chapitre de Génie Industriel ou de Méthodes Numériques). Nous tirons successivement, et sans les remettre dans l'urne, p boules. Quelle est la probabilité que parmi ces p boules, il y en ait k qui soient noires (dans cet énoncé l'ordre du tirage ne nous intéresse donc pas!).

Nous parlons souvent de "tirage exhaustif" avec la loi hypergéométrique car contrairement à la loi binomiale, la taille du lot qui sert de base au tirage va apparaître dans la loi. Raison pour laquelle la loi hypergéométrique tend vers les valeurs de la loi Normale lorsque la taille du lot est petite.

Remarque: Cela équivaut à l'étude non ordonnée d'un tirage sans remise (cf. chapitre de Probabilités) avec contrainte sur les occurrences appelé parfois "tirage simultané". Nous utiliserons souvent cette fonction dans le domaine de la qualité ou de la fiabilité où les boules noires sont associées à des éléments avec défauts et les blanches à des éléments sans défauts.

Les p boules peuvent être choisies parmi les n boules de C_p^n façons (représentant donc le nombre de tirages différents possibles) avec pour rappel (cf. chapitre de Probabilités):

$$C_p^n = \binom{n}{p} = \frac{A_p^n}{p!} = \frac{n!}{p!(n-p)!} \quad (7.269)$$

Les k boules noires peuvent être choisies parmi les m noires de C_k^m façons. Les $p-k$ boules blanches peuvent être elles choisies de C_{p-k}^{n-m} façons. Il y a donc $C_k^m C_{p-k}^{n-m}$ tirages qui donnent k boules noires et $p-k$ boules blanches.

La probabilité recherchée vaut donc (nous en verrons une autre formulation possible dans le chapitre de Génie Industriel):

$$P(n, p, m, k) = H(k) = \frac{C_k^m C_{p-k}^{m-m}}{C_p^m} = \frac{\frac{m!}{k!(m-k)!} \frac{(n-m)!}{(p-k)!((n-m)-(p-k))!}}{\frac{n!}{p!(n-p)!}} \quad (7.270)$$

et est dite suivre une "**fonction Hypergéométrique**" (ou "**loi Hypergéométrique**") et peut être obtenue heureusement de manière directe dans Microsoft Excel 11.8346 avec la fonction LOI.HYPERGEOMETRIQUE().

Exemples:

E1. Nous souhaitons mettre en production un petit développement informatique de 10'000 lignes de code (n). Le retour d'expérience montre que la probabilité de défaillance est de 1 bug pour 1'000 lignes de code (soit 0.1% de 10'000 lignes) ce qui correspond à valeur de m .

Nous testons environ 50% des fonctions du logiciel au hasard avant l'envoi au client (soit l'équivalent de 5'000 lignes de code correspondant à p). La probabilité d'observer 5 bugs (k) est avec Microsoft Excel 11.8346:

$$\begin{aligned} &= \text{LOI.HYPERGEOMETRIQUE}(k;p;m;n) \\ &= \text{LOI.HYPERGEOMETRIQUE}(5;5000;0.1\%*10000;10000)=24.62\% \end{aligned}$$

E2. Dans une petite production unique d'un lot de 1'000 pièces dont nous savons que 30% en moyenne sont mauvaises à cause de la complexité des pièces par retour d'expérience d'une fabrication précédente similaire. Nous savons qu'un client va en tirer 20 au hasard pour décider d'accepter ou de rejeter le lot. Il ne rejettera pas le lot s'il trouve zéro pièce défectueuse parmi ces 20. Quelle est la probabilité d'en avoir exactement 0 de défectueuse?

$$=\text{LOI.HYPERGEOMETRIQUE}(0;20;300;1000)=0.073\%$$

et comme on exige un tirage nul, le calcul de la loi hypergéométrique se simplifie en:

$$\frac{700}{1000} \cdot \frac{699}{999} \cdot \frac{698}{998} \cdot \frac{697}{997} \cdots \frac{681}{981} = 0.073\% \quad (7.271)$$

Il n'est pas interdit de faire le calcul direct de l'espérance et de la variance la fonction hypergéométrique mais le lecteur pourra sans trop de peine imaginer que ce calcul va être... relativement indigeste. Alors nous pouvons utiliser une méthode indirecte qui de plus est intéressante.

D'abord le lecteur aura peut-être, même certainement, remarqué qu'au fait l'expérience de la loi hypergéométrique est une série d'essais de Bernoulli (sans remise bien entendu!).

Alors, nous allons tricher en utilisant dans un premier temps la propriété de linéarité de l'espérance. Définissons pour cela une nouvelle variable correspondant implicitement au fait à l'expérience de la fonction hypergéométrique (k essais de Bernoulli de suite!):

$$X = \sum_{i=1}^k X_i \quad (7.272)$$

où X_i représente la réussite d'obtenir au i -ème tirage une boule noire (soit 0 ou 1). Or, nous savons que pour tout i la variable aléatoire X_i suit une fonction de Bernoulli pour laquelle nous avons démontré lors de notre étude de la loi de Bernoulli que $E(X_i) = p$. Dès lors, de par la propriété de linéarité de l'espérance nous avons:

$$E(X) = E\left(\sum_{i=1}^k X_i\right) = k \cdot p \quad (7.273)$$

Or, dans l'essai de Bernoulli, p est la probabilité d'obtenir l'élément recherché (pour rappel...). Dans la loi hypergéométrique ce qui nous intéresse est la probabilité d'avoir une boule noire (qui sont en quantité m , avec donc m' boules blanches) par rapport à la quantité totale de boules n . Et le rapport nous donne évidemment cette probabilité. Ainsi, nous avons:

$$\mu = E(X) = k \cdot \frac{m}{m' + m} = k \cdot \frac{m}{n} \quad (7.274)$$

où k est le nombre de tirages (attention à ne pas confondre avec la notation de l'énoncé initial où il était noté par la variable p). Cette moyenne donne donc le nombre moyen de boules noires lors d'un tirage de k boules parmi n , dont m sont connues comme étant noires.

Pour déterminer la variance, nous allons utiliser la variance de la fonction de Bernoulli et la relation suivante démontrée lors de l'introduction de l'espérance et de la covariance au début de ce chapitre:

$$V(X + Y) = V(X) + V(Y) + 2\text{cov}(X, Y) = V(X) + V(Y) + 2(E(X \cdot Y) - E(X)E(Y)) \quad (7.275)$$

Dons en rappelant que nous avons $X = \sum_{i=1}^k X_i$ il vient:

$$V\left(\sum_{i=1}^k X_i\right) = \sum_{i=1}^k V(X_i) + 2 \sum_{1 \leq i < j \leq k} (E(X_i \cdot X_j) - E(X_i)E(X_j)) \quad (7.276)$$

Or, pour la loi de Bernoulli, nous avons:

$$V(X) = p \cdot q = \frac{m}{m' + m} \frac{m'}{m' + m} = \frac{m \cdot m'}{(m' + m)^2} = \frac{m \cdot m'}{n^2} \quad (7.277)$$

Alors nous avons déjà:

$$\sum_{i=1}^k V(X_i) = k \frac{m \cdot m'}{(m' + m)^2} = k \frac{m \cdot m'}{n^2} \quad (7.278)$$

Ensuite, nous avons facilement:

$$E(X_i)E(X_j) = \frac{m}{m'+m} \frac{m}{m'+m} = \left(\frac{m}{m'+m} \right)^2 \quad (7.279)$$

Le calcul de $E(X_i X_j)$ nécessite une bonne compréhension des probabilités (c'est un bon rappel!).

L'espérance $E(X_i X_j)$ est donnée (implicitement) par la somme pondérée des probabilités que deux événements aient lieu en même temps comme nous le savons. Or, nos événements sont binaires: soit c'est une boule noire (1) soit c'est une boule blanche (0). Donc tous les termes de la somme n'ayant pas deux boules noires consécutivement seront nuls!

Le problème est alors de calculer la probabilité d'avoir deux boules noires consécutives et celle-ci s'écrit donc:

$$P(X_i X_j = 1) = P((X_i = 1) \cap (X_j = 1)) = P(X_i = 1) P_{X_i=1}(X_j = 1) = \frac{m}{m'+m} \frac{m-1}{(m'+m)-1} \quad (7.280)$$

Donc nous avons finalement:

$$E(X_i X_j) = \frac{m}{m'+m} \frac{m-1}{(m'+m)-1} \quad (7.281)$$

Soit:

$$\begin{aligned} \text{cov}(X_i, X_j) &= E(X_i \cdot X_j) - E(X_i)E(X_j) = \frac{m}{m'+m} \frac{m-1}{(m'+m)-1} - \left(\frac{m}{m'+m} \right)^2 \\ &= \frac{-m \cdot m'}{(m'+m)^2 (m'+m-1)} \end{aligned} \quad (7.282)$$

Finalement:

$$\begin{aligned} V(X) &= \sum_{i=1}^k V(X_i) + 2 \sum_{1 \leq i < j \leq n} (E(X_i \cdot X_j) - E(X_i)E(X_j)) \\ &= k \frac{m \cdot m'}{(m'+m)^2} + 2 \frac{-m \cdot m'}{(m'+m)^2 (m'+m-1)} \sum_{1 \leq i < j \leq n} 1 \\ &= k \frac{m \cdot m'}{(m'+m)^2} + 2 \frac{-m \cdot m'}{(m'+m)^2 (m'+m-1)} \frac{(k-1)k}{2} \\ &= \frac{k \cdot m \cdot m' (m'+m-1) - m \cdot m' k (k-1)}{(m'+m)^2 (m'+m-1)} = \frac{k \cdot m \cdot m' (m'+m-1-k+1)}{(m'+m)^2 (m'+m-1)} \\ &= \frac{k \cdot m \cdot m' (m'+m-k)}{(m'+m)^2 (m'+m-1)} = k \frac{m}{m'+m} \frac{m'}{m'+m} \frac{m'+m-k}{m'+m-1} = k p q \frac{m'+m-k}{m'+m-1} \end{aligned} \quad (7.283)$$

où nous avons utilisé le fait que:

$$\sum_{i < j}^n \text{cov}(X_i, X_j) \quad (7.284)$$

est composé de:

$$C_2^k = \binom{k}{2} = \frac{k(k-1)}{2} \quad (7.285)$$

termes puisqu'il correspond au nombre de façons qu'il y a de choisir le couple (i, j) avec $i < j$.

Donc finalement:

$$V(X) = k p q \frac{n-k}{n-1} \quad (7.286)$$

ou:

$$\sigma = \sqrt{k p q} \sqrt{\frac{n-k}{n-1}} \quad (7.287)$$

Nous voyons qu'il s'agit du même écart-type que la loi binomiale, à la différence d'un facteur qui est noté:

$$fpc = \sqrt{\frac{n-k}{n-1}} \quad (7.288)$$

que l'on retrouve assez souvent en statistiques et qui est appelé "**facteur de correction de population**" ou en anglais "**finite population correction factor**".

Voici un exemple de tracé de la fonction de distribution et répartition pour la fonction Hypergéométrique de paramètre $(n, p, m, k) = (10, 6, 5, k)$:

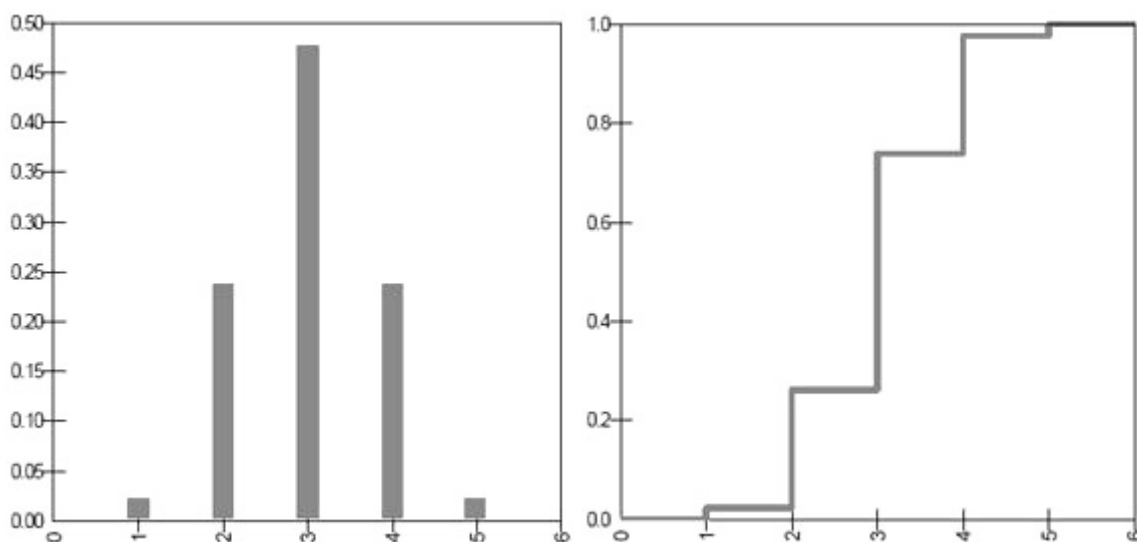


Figure: 7.13 - Loi hypergéométrique (fonction de distribution et de répartition)

Démontrons que la loi Hypergéométrique tend vers une loi binomiale puisqu'il en est fait usage de nombreuses fois dans différents chapitres du site (et particulièrement le chapitre de Génie Industriel).

Pour cela, décomposons:

$$P(n, p, m, k) = \frac{C_k^m C_{p-k}^{n-m}}{C_p^n} \quad (7.289)$$

Il vient alors:

$$\begin{aligned} \frac{C_k^m C_{p-k}^{n-m}}{C_p^n} &= \frac{\left(\frac{m!}{k!(m-k)!} \right) \left(\frac{(n-m)!}{(p-k)!((n-m)-(p-k))!} \right)}{\left(\frac{n!}{p!(n-p)!} \right)} \\ &= \frac{m!}{k!(m-k)!} \frac{(n-m)!}{(p-k)!((n-m)-(p-k))!} \frac{p!(n-p)!}{n!} \quad (7.290) \\ &= \frac{p!}{k!(p-k)!} \frac{m!}{(m-k)!} \frac{(n-m)!}{((n-m)-(p-k))!} \frac{(n-p)!}{n!} \\ &= C_k^p \frac{m!}{(m-k)!} \frac{(n-m)!}{((n-m)-(p-k))!} \frac{(n-p)!}{n!} \end{aligned}$$

Pour le deuxième terme:

$$\frac{m!}{(m-k)!} = \frac{1 \cdot 2 \cdot 3 \cdot \dots \cdot m}{1 \cdot 2 \cdot 3 \cdot \dots \cdot (m-k)} = m \cdot (m-1) \cdot \dots \cdot (m-k+1) \quad (7.291)$$

Pour $m \rightarrow +\infty$ tous les termes sont alors de l'ordre de m . Nous avons alors:

$$m \cdot (m-1) \cdot \dots \cdot (m-k+1) = m^k \quad (7.292)$$

Pour le troisième terme un développement identique en tous points au précédent permet d'obtenir:

$$\frac{(n-m)!}{((n-m)-(p-k))!} = (n-m)^{p-k} \quad (7.293)$$

Idem pour le quatrième terme:

$$\frac{(n-p)!}{n!} = n^{-p} \quad (7.294)$$

En conclusions nous avons:

$$\frac{C_k^m C_{p-k}^{n-m}}{C_p^n} \xrightarrow{m \rightarrow +\infty} C_k^p \frac{m^k (n-m)^{p-k}}{n^p} \quad (7.295)$$

Changeons d'écriture en posant p (le nombre d'individus tirés) comme étant N . Il vient alors:

$$\frac{C_k^m C_{N-k}^{n-m}}{C_N^n} \xrightarrow{m \rightarrow +\infty} C_k^N \frac{m^k (n-m)^{N-k}}{n^N} \quad (7.296)$$

Faisons un autre changement d'écriture en notant b les boules noires (black) et w les boules blanches (white). Il vient alors:

$$\frac{C_k^b C_{N-k}^{n-b}}{C_N^n} \xrightarrow{b \rightarrow +\infty} C_k^N \frac{b^k (n-b)^{N-k}}{n^N} = C_k^N \frac{b^k w^{N-k}}{n^N} \quad (7.297)$$

Enfin, notons p la proportion de boules noires et q celle de boules blanches dans le lot n . Il vient alors:

$$\begin{aligned} \frac{C_k^{np} C_{N-k}^{n-np}}{C_N^n} &\xrightarrow{np \rightarrow +\infty} C_k^N \frac{(np)^k (nq)^{N-k}}{n^N} = C_k^N (np)^k (nq)^{N-k} n^{-N} \\ &= C_k^N p^k q^{N-k} n^k n^{N-k} n^{-N} = C_k^N p^k q^{N-k} \end{aligned} \quad (7.298)$$

Nous retrouvons donc bien la loi binomiale!! En pratique, il est courant d'approximer la loi hypergéométrique de paramètres par une loi binomiale de paramètres dès que le rapport nombre d'individus tirés sur le nombre total d'individus est inférieur à 10% (c'est-à-dire lorsque l'échantillon est 10 fois plus petit que la population). Il s'ensuit que la loi hypergéométrique tend aussi (comme nous le démontrerons plus loin) vers une loi Normale lorsque la population tend vers l'infini et que l'échantillon est petit.

5.7. FONCTION MULTINOMIALE

La loi multinomiale (appelée ainsi car elle fait intervenir plusieurs fois le coefficient binomial) est une loi applicable à n événements distinguables, chacun ayant une probabilité donnée, qui surviennent une ou plusieurs fois et ce de façon non nécessairement ordonné. Il s'agit d'un cas fréquent dans les études marketing et qui nous sera utile pour construire le test statistique de McNemar beaucoup plus loin. Nous retrouvons également cette loi en finance quantitative.

Plus techniquement, considérons l'espace des événements $\Omega = \{1, 2, \dots, m\}$ muni d'une probabilité $P(\{i\}) = p_i, i = 1, 2, \dots, m$. Nous tirons n fois de suite avec remise un élément de Ω avec la probabilité $p_i, i = 1, 2, \dots, m$. Quelle est donc la probabilité d'obtenir de manière non nécessairement ordonnée l'événement 1, k_1 fois, l'événement 2, k_2 fois et ce sur une suite d'un tirage de n éléments.

Remarque: Cela équivaut à l'étude d'un tirage avec remise (cf. chapitre de Probabilités) et contraintes sur les occurrences. Donc sans contraintes nous verrons par l'exemple que nous retombons sur un tirage avec remise simple.

Nous avons vu dans le chapitre de Probabilités, que si nous prenons un ensemble d'événements ayant plusieurs issues, alors les différentes combinaisons de suites que nous pouvons obtenir en prenant p éléments choisis parmi n est:

$$C_p^n = \frac{n!}{p!(n-p)!} \quad (7.299)$$

Il y a donc:

$$C_{k_1}^n = \frac{n!}{k_1!(n-k_1)!} \quad (7.300)$$

façons différentes d'obtenir k_1 fois un certain événement. Soit une probabilité associée de:

$$P(n, k_1) = C_{k_1}^n p_1^{k_1} q_1^{n-k_1} = C_{k_1}^n p_1^{k_1} (1-p_1)^{n-k_1}$$

Maintenant, intervient la particularité de la loi multinomiale!: il n'y a pas d'échecs contrairement à la loi binomiale. Chaque "pseudo-échec" peut être considéré comme un sous tirage de k_2 parmi les $n-k_1$ éléments restants.

Ainsi le terme:

$$C_{k_1}^n p_1^{k_1} q_1^{n-k_1} \quad (7.301)$$

s'écrira sur l'ensemble de l'expérience si nous considérons un cas particulier limité à deux types d'événements:

$$C_{k_1}^n p_1^{k_1} q_1^{n-k_1} := C_{k_1}^n p_1^{k_1} \left(C_{k_2}^{n-k_1} p_2^{k_2} \right) \quad (7.302)$$

avec donc:

$$C_{k_2}^{n-k_1} = \frac{(n-k_1)!}{k_2!((n-k_1)-k_2)!} \quad (7.303)$$

qui donne le nombre de façons différentes d'obtenir k_2 fois un second événement puisque dans l'ensemble de la suite, de n éléments déjà k_1 ont été tirés ce qui fait qu'il n'en reste plus que $n-k_1$ sur lesquels nous pouvons obtenir les k_2 voulus.

Ces relations nous montrent donc qu'il s'agit d'une situation où chaque probabilité d'événement est considérée comme une sous loi binomiale (d'où son nom aussi...).

Alors nous avons dans le cas particulier de deux séries d'uplets:

$$\begin{aligned} C_{k_1}^n p_1^{k_1} C_{k_2}^{n-k_1} p_2^{k_2} &= C_{k_1}^n C_{k_2}^{n-k_1} p_1^{k_1} p_2^{k_2} \\ &= \frac{n!}{k_1!(n-k_1)!} \frac{(n-k_1)!}{k_2!((n-k_1)-k_2)!} p_1^{k_1} p_2^{k_2} \quad (7.304) \\ &= \frac{n!}{k_1!k_2!(n-k_1-k_2)!} p_1^{k_1} p_2^{k_2} \end{aligned}$$

et comme:

$$-k_1 - k_2 = -n \quad (7.305)$$

il vient:

$$P = \frac{n!}{k_1! k_2! 0!} p_1^{k_1} p_2^{k_2} = \frac{n!}{k_1! k_2!} p_1^{k_1} p_2^{k_2} \quad (7.306)$$

et nous voyons que la construction de cette loi impose donc que:

$$\sum_i p_i = 1 \quad \sum_i k_i = n \quad (7.307)$$

Ainsi, par récurrence nous avons la probabilité P recherchée appelée "**fonction Multinomiale**" (ou "**loi Multinomiale**") et donnée par:

$$P = \frac{n!}{\prod_{i=1}^m k_i!} \prod_{i=1}^m p_i^{k_i} = \frac{\left(\sum_{i=1}^m k_i\right)!}{\prod_{i=1}^m k_i!} \prod_{i=1}^m p_i^{k_i} \quad (7.308)$$

dans le logiciel Microsoft Excel 11.8346, le terme:

$$\frac{\left(\sum_{i=1}^m k_i\right)!}{\prod_{i=1}^m k_i!} \quad (7.309)$$

appelé "**coefficient multinomial**" est disponible sous le nom de la fonction MULTINOMIALE() dans la version française. Dans la littérature nous trouvons également ce terme parfois sous les formes respectives suivantes:

$$\binom{k_1 + k_2 + \dots + k_m}{k_1, k_2, \dots, k_m} \quad \binom{m}{k_1, k_2, \dots, k_m} \quad (7.310)$$

Démontrons que la loi multinomiale est bien une loi de probabilité (car nous pourrions en douter...). Si c'est bien le cas, la somme des probabilités doit être comme nous le savons, égale à l'unité.

Démonstration:

Rappelons que dans le chapitre de Calcul Algébrique nous avons démontré que (théorème binomial):

$$(x + y)^n = \sum_{k=0}^n C_k^n x^k y^{n-k} \quad (7.311)$$

Faisons maintenant un petit peu de notation:

$$(x_1 + x_2)^n = \sum_{k_1=0}^n C_k^n x_1^{k_1} x_2^{n-k_1} = \sum_{k_1=0}^n \frac{n!}{k_1!(n-k_1)!} x_1^{k_1} x_2^{n-k_1} \quad (7.312)$$

et cette fois-ci un changement de variables:

$$(x_1 + x_2)^n = \sum_{k_1=0}^n C_k^n x_1^{k_1} x_2^{n-k_1} = \sum_{k_1=0}^n \frac{n!}{k_1!k_2!} x_1^{k_1} x_2^{k_2} \quad (7.313)$$

Cette dernière relation (qui est un cas à deux termes du "théorème multinomial") va nous être utile pour démontrer que la loi multinomiale est bien une loi de probabilité. Nous prenons donc le cas particulier avec deux groupes de tirage:

$$P = \frac{n!}{\prod_{i=1}^m k_i!} p_1^{k_1} p_2^{k_2} = \frac{n!}{k_1!k_2!} p_1^{k_1} p_2^{k_2} \quad (7.314)$$

ce qui s'écrit aussi de par la construction de la loi multinomiale:

$$P = \frac{n!}{k_1!(n-k_1)!} p_1^{k_1} p_2^{(n-k_1)} \quad (7.315)$$

et donc la somme doit être égale à l'unité telle que:

$$\sum_{k_1=0}^n \frac{n!}{k_1!(n-k_1)!} p_1^{k_1} p_2^{(n-k_1)} = 1 \quad (7.316)$$

pour vérifier cela nous utilisons le théorème multinomial montré précédemment:

$$(p_1 + p_2)^n = \sum_{k_1=0}^n \frac{n!}{k_1!(n-k_1)!} p_1^{k_1} p_2^{(n-k_1)} \quad (7.317)$$

Or, comme par construction de la loi multinomiale la somme des probabilités est unitaire, nous avons bien:

$$(p_1 + p_2)^n = (1)^n = 1 = \sum_{k_1=0}^n \frac{n!}{k_1!(n-k_1)!} p_1^{k_1} p_2^{(n-k_1)} \quad (7.318)$$

□C.Q.F.D.

Exemples:

E1. Nous lançons un dé non-pipé 12 fois. Quelle est la probabilité que les six faces apparaissent le même nombre de fois (mais pas nécessairement consécutivement!) soit deux fois pour chaque:

$$P = \frac{n!}{\prod_{i=1}^m k_i!} \prod_{i=1}^m p_i^{k_i} = \frac{12!}{\prod_{i=1}^6 2!} \prod_{i=1}^6 \left(\frac{1}{6}\right)^2 = \frac{12!}{2^6 6^{12}} = 0.34\% \quad (7.319)$$

où nous voyons bien que m correspond au nombre de groupes de réussites.

E2. Nous lançons un dé non-pipé 12 fois. Quelle est la probabilité qu'une seule et unique face apparaisse 12 fois (donc que le "1" apparaisse 12 fois de suite, ou le "2", ou le "3", etc.):

$$P = \frac{n!}{\prod_{i=1}^m k_i!} \prod_{i=1}^m p_i^{k_i} = \frac{12!}{\prod_{i=1}^1 12!} \prod_{i=1}^1 \left(\frac{1}{6}\right)^{12} = \frac{1}{6^{12}} \quad (7.320)$$

Nous retrouvons donc avec ce dernier exemple un résultat connu de la binomiale.

5.8. FONCTION DE POISSON

Pour certains événements forts rares, la probabilité p est très faible et tend vers zéro. Toutefois la valeur moyenne $n \cdot p$ tend vers une valeur fixe lorsque n tend vers l'infini.

Nous partons donc d'une distribution binomiale de moyenne $\mu = n \cdot p$ que nous supposons finie lorsque n tend vers l'infini.

La probabilité de k réussites lors de n épreuves vaut (loi Binomiale):

$$P_n(k) = \binom{n}{k} p^k q^{n-k} \quad (7.321)$$

En posant $p = m/n$ (où m est temporairement la nouvelle notation pour la moyenne selon $\mu = n \cdot p$), cette expression peut s'écrire:

$$\frac{n!}{k!(n-k)!} \left(\frac{m}{n}\right)^k \left(1 - \frac{m}{n}\right)^{n-k} \quad (7.322)$$

En regroupant les termes, nous pouvons mettre la valeur de $P_n(k)$ sous la forme:

$$\frac{m^k}{k!} \left(1 - \frac{m}{n}\right)^n \frac{n!}{(n-k)! n^k \left(1 - \frac{m}{n}\right)^k} \quad (7.323)$$

Nous reconnaissons que, lorsque n tend vers l'infini, le deuxième facteur du produit a pour limite $e^{-\mu}$.

Quant au troisième facteur, puisque nous nous intéressons aux petites valeurs de k (la probabilité de réussite est très faible), sa limite pour n tendant vers l'infini vaut 1.

Cette technique de passage à la limite est parfois appelée dans ce contexte: "**théorème limite de Poisson**".

Nous obtenons ainsi la "**fonction de Poisson**" (ou "**loi de Poisson**"), appelée également parfois "**loi des événements rares**", donnée donc par:

$$P_k = \frac{\mu^k}{k!} \cdot e^{-\mu} \quad (7.324)$$

qui peut être obtenue dans Microsoft Excel 11.8346 avec la fonction LOI.POISSON() et qui dans la pratique et la littérature spécialisée est souvent notée par la lettre u .

Il s'agit bien d'une loi de probabilité puisqu'en utilisant les séries de Taylor (cf. chapitre de Suites Et Séries), nous montrons que la somme des probabilités cumulées est bien:

$$\sum_{k=0}^{+\infty} \frac{\mu^k}{k!} e^{-\mu} = e^{-\mu} \sum_{k=0}^{+\infty} \frac{\mu^k}{k!} = e^{-\mu} e^{\mu} = 1 \quad (7.325)$$

Remarque: Nous retrouverons fréquemment cette loi dans différents chapitres du site comme par exemple lors de l'étude du Génie Industriel en maintenance préventive ou encore dans le même chapitre lors de l'étude des théories des files d'attente (le lecteur peut s'y reporter pour un exemple intéressant et pragmatique) et enfin dans le domaine de l'assurance.

Voici un exemple de tracé de la fonction de distribution et répartition pour la fonction de Poisson de paramètre $\mu = 3$:

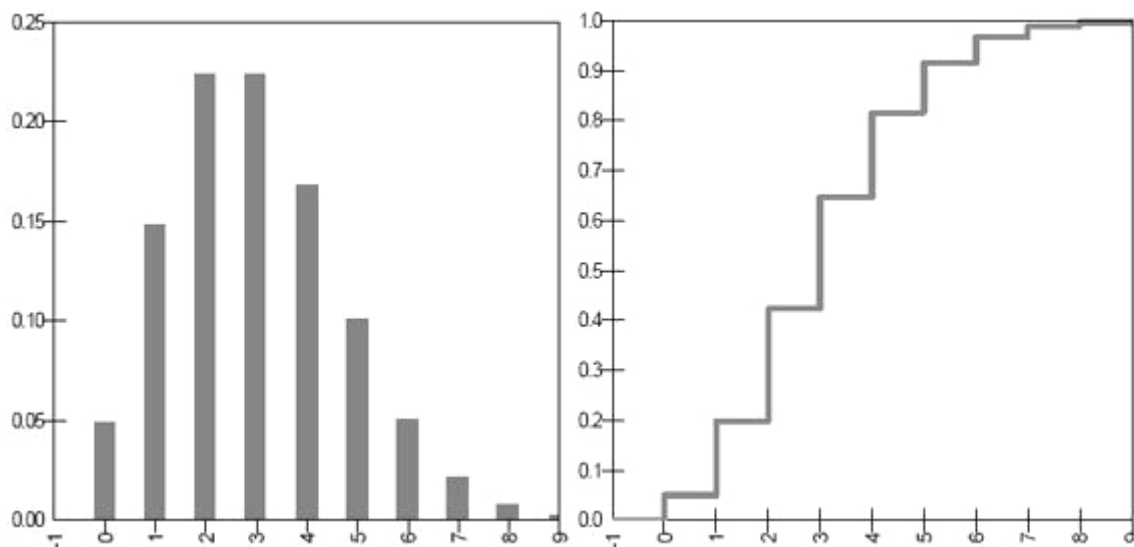


Figure: 7.14 - Loi de Poisson (fonction de distribution et de répartition)

Cette distribution est importante car elle décrit beaucoup de processus dont la probabilité est petite et constante. Elle est souvent utilisée dans la "queuing theory" (temps d'attente), test d'acceptabilité et fiabilité, et contrôles statistiques de qualité. Entre autres, elle s'applique aux processus tels que l'émission des quanta de lumière par des atomes excités, le nombre de globules rouges observés au microscope, le nombre d'appels arrivant à une centrale téléphonique. La distribution de Poisson est valable pour de nombreuses observations faites en physique nucléaire ou corpusculaire.

L'espérance (moyenne) de la fonction de Poisson est (nous utilisons la série de Taylor de l'exponentielle):

$$\mu = E(k) = \sum_{k=0}^{+\infty} k \cdot P_k = \sum_{k=0}^{+\infty} k \cdot \frac{\mu^k}{k!} \cdot e^{-\mu} = e^{-\mu} \cdot \mu \sum_{k=1}^{+\infty} \frac{\mu^{k-1}}{(k-1)!} = e^{-\mu} \cdot \mu \cdot e^{\mu} = \mu \quad (7.326)$$

et donne le nombre moyen de fois que l'on obtiendra l'issue souhaitée.

Ce résultat peut paraître déroutant.... la moyenne s'exprime par la moyenne??? Oui il ne faut simplement pas oublier que celle-ci est donnée au début par:

$$\mu = n \cdot p \quad (7.327)$$

Remarque: Pour plus de détails le lecteur peut aussi se reporter à la partie concernant les "estimateurs" dans le présent chapitre.

La variance de la fonction de distribution de Poisson est quant à elle donnée par (en utilisant à nouveau les séries de Taylor):

$$\begin{aligned} \sigma^2 = V(k) &= \sum_{k=0}^{+\infty} [k - \mu]^2 P_k = \sum_{k=0}^{+\infty} [k - \mu]^2 \frac{\mu^k}{k!} e^{-\mu} \\ &= \sum_{k=0}^{+\infty} k^2 \frac{\mu^k}{k!} e^{-\mu} - \sum_{k=0}^{+\infty} 2k\mu \frac{\mu^k}{k!} e^{-\mu} + \sum_{k=0}^{+\infty} \mu^2 \frac{\mu^k}{k!} e^{-\mu} \\ &= \sum_{k=0}^{+\infty} k^2 \frac{\mu^k}{k!} e^{-\mu} - 2\mu \sum_{k=0}^{+\infty} k \frac{\mu^k}{k!} e^{-\mu} + \mu^2 \sum_{k=0}^{+\infty} \frac{\mu^k}{k!} e^{-\mu} \\ &= \sum_{k=0}^{+\infty} k^2 \frac{\mu^k}{k!} e^{-\mu} - 2\mu \sum_{k=1}^{+\infty} \frac{k \mu^{k-1}}{(k-1)!} e^{-\mu} + \mu^2 \cdot 1 = \sum_{k=0}^{+\infty} k^2 \frac{\mu^k}{k!} e^{-\mu} - 2\mu^2 + \mu^2 \\ &= \mu e^{-\mu} \sum_{k=1}^{+\infty} k \frac{\mu^{k-1}}{(k-1)!} - \mu^2 = \mu e^{-\mu} \sum_{s=0}^{+\infty} (s+1) \frac{\mu^s}{s!} - \mu^2 \\ &= \mu e^{-\mu} \left(\sum_{s=0}^{+\infty} s \frac{\mu^s}{s!} + \sum_{s=0}^{+\infty} \frac{\mu^s}{s!} \right) - \mu^2 = \mu e^{-\mu} \left(\mu \sum_{s=1}^{+\infty} \frac{\mu^{s-1}}{(s-1)!} + e^{\mu} \right) - \mu^2 \\ &= \mu e^{-\mu} (\mu e^{\mu} + e^{\mu}) - \mu^2 = \mu e^{-\mu} \mu e^{\mu} + \mu e^{-\mu} e^{\mu} - \mu^2 = \mu^2 + \mu - \mu^2 = \mu \end{aligned} \quad (7.328)$$

toujours avec:

$$\mu = n \cdot p \quad (7.329)$$

Le fait important que pour la loi de Poisson nous ayons la variance qui soit égale à l'espérance est appelé "**propriété d'équidispersion de la de Poisson**". Il s'agit d'une propriété souvent utilisée dans la pratique comme indicateur pour identifier si des données (à support discret) sont distribuées selon une loi de Poisson.

Les lois théoriques de distributions statistiques sont établies en supposant la réalisation d'un nombre infini de mesures. Il est évident que nous ne pouvons en effectuer qu'un nombre fini N . D'où la nécessité d'établir des correspondances entre les valeurs utiles théoriques et expérimentales. Pour ces dernières nous n'obtenons évidemment qu'une approximation dont la validité est toutefois souvent

admise comme suffisante.

Maintenant démontrons une propriété importante de la loi Poisson dans le domaine de l'ingénierie que nous appelons la "stabilité par l'addition". L'idée est la suivante:

Soit deux variables aléatoires indépendantes X et Y de loi de Poisson de paramètre respectif λ et μ . Nous voulons vérifier que leur somme est aussi une loi de Poisson:

$$X + Y = P_{\lambda + \mu} \quad (7.330)$$

Voyons cela:

$$P(X + Y = k) = \sum_{i=0}^k P[(X = i) \cap (Y = k - i)] = \sum_{i=0}^k P(X = i)P(Y = k - i) \quad (7.331)$$

car les événements sont indépendants. Nous avons alors:

$$\begin{aligned} P(X + Y = k) &= \sum_{i=0}^k P(X = i)P(Y = k - i) = \sum_{i=0}^k \frac{\lambda^i e^{-\lambda}}{i!} \frac{\mu^{k-i} e^{-\mu}}{(k-i)!} \\ &= \frac{e^{-(\lambda + \mu)}}{k!} \sum_{i=0}^k \frac{k!}{i!(k-i)!} \lambda^i \mu^{k-i} \end{aligned} \quad (7.332)$$

Or, en appliquant le théorème binomial (cf. chapitre Calcul Algébrique):

$$\sum_{i=0}^k \frac{k!}{i!(k-i)!} \lambda^i \mu^{k-i} = \sum_{i=0}^k C_i^k \lambda^i \mu^{k-i} = (\lambda + \mu)^k \quad (7.333)$$

Donc au final:

$$P(X + Y = k) = (\lambda + \mu)^k \frac{e^{-(\lambda + \mu)}}{k!} \quad (7.334)$$

et donc la loi de Poisson est bien stable par l'addition.

5.9. FONCTION DE GAUSS-LAPLACE/LOI NORMALE

Cette caractéristique est la plus importante fonction de distribution en statistiques suite au résultat d'un théorème connu appelé "théorème central limite" qui comme nous le verrons, permet de démontrer (entre autres) que toute suite de variables aléatoires indépendantes de même loi ayant une espérance et un écart-type fini et non nécessairement égales converge vers une fonction de Gauss-Laplace (loi Normale).

Il est donc très important de focaliser particulièrement son attention sur les développements qui vont être faits ici!

Partons d'une fonction Binomiale et faisons tendre le nombre n d'épreuves vers l'infini. Si p est fixé au départ, la moyenne $\mu = n \cdot p$ tend également vers l'infini, de plus l'écart-type $\sigma = \sqrt{n \cdot p \cdot q}$ tend également

vers l'infini.

Remarque: Le cas où p varie et tend vers 0 tout en laissant fixe la moyenne $\mu = n \cdot p$ ayant déjà été étudié lors de la présentation de la fonction de Poisson.

Si nous voulons calculer la limite de la fonction Binomiale, il s'agira donc de faire un changement d'origine qui stabilise la moyenne, en 0 par exemple, et un changement d'unité qui stabilise l'écart-type, à 1 par exemple.

Voyons tout d'abord comment varie $P_n(k)$ en fonction de k (nombre de réussites) et calculons la différence:

$$\begin{aligned} P_n(k+1) - P_n(k) &= \binom{n}{k+1} p^{k+1} q^{n-k-1} - \binom{n}{k} p^k q^{n-k} = \binom{n}{k} p^k q^{n-k} \cdot \left[\frac{(n-k)p}{(k+1)q} - 1 \right] \\ &= P_n(k) \frac{np - k - q}{(k+1)q} \end{aligned} \quad (7.335)$$

Nous en concluons que $P_n(k)$ est une fonction croissante de k , tant que $np - k - q$ est positif (pour n , p et q fixés). Pour le voir il suffit de prendre quelques valeurs (du membre de droite de l'égalité) ou d'observer la distribution graphique de la fonction Binomiale en se souvenant bien que:

$$\mu = n \cdot p \quad (7.336)$$

Comme $q < 1$ il est par conséquent évident que la valeur de k voisine de l'espérance de la loi Binomiale $\mu = n \cdot p$ constitue le maxima de $P_n(k)$.

D'autre part la différence $P_n(k+1) - P_n(k)$ est le taux d'accroissement de la fonction $P_n(k)$. Nous pouvons alors écrire:

$$\frac{\Delta P_n(k)}{\Delta k} = \frac{P_n(k+1) - P_n(k)}{(k+1) - k} \quad (7.337)$$

comme étant la pente de la fonction.

Définissons maintenant une nouvelle variable aléatoire telle que sa moyenne soit nulle (variations négligeables) et son écart-type unitaire (une variable centrée-réduite en d'autres termes). Nous avons alors:

$$x = \frac{k - np}{\sqrt{npq}}$$

Nous avons alors aussi avec cette nouvelle variable:

$$\Delta x = \frac{(k+1) - np}{\sqrt{npq}} - \frac{k - np}{\sqrt{npq}} = \frac{[(k+1) - np] - (k - np)}{\sqrt{npq}} \quad (7.338)$$

Appelons $F(x)$ l'expression de $P_n(k)$ calculée en fonction de la nouvelle variable de moyenne nulle et

d'écart-type unitaire dont nous recherchons l'expression quand n tend vers l'infini.

Reprenons:

$$\frac{\Delta P_n(k)}{\Delta k} = \frac{np - k - q}{(k+1)q} P_n(k) = \frac{-(k - np) - q}{(k+1)q} P_n(k) \quad (7.339)$$

Afin de simplifier l'étude de cette relation quand n tend vers l'infini et k vers l'espérance $\mu = n \cdot p$, multiplions des deux côtés par $npq / \sqrt{npq}$:

$$\frac{\Delta P_n(k)}{\Delta k} \frac{npq}{\sqrt{npq}} = \frac{-(k - np) - q}{(k+1)q} \frac{npq}{\sqrt{npq}} P_n(k) \quad (7.340)$$

Réécrivons le terme de droite de l'égalité. Il vient alors:

$$\frac{-(k - np) - q}{(k+1)} \frac{np}{\sqrt{npq}} P_n(k) = \frac{[-(k - np) - q] np}{(k+1) \sqrt{npq}} P_n(k) \quad (7.341)$$

Et maintenant réécrivons le terme de gauche de la relation antérieure. Il vient:

$$\begin{aligned} \frac{\Delta P_n(k)}{\Delta k} \frac{npq}{\sqrt{npq}} &= \frac{\Delta P_n(k)}{[(k+1) - k]} \frac{npq}{\sqrt{npq}} = \frac{\Delta P_n(k)}{[(k+1) - np] - (k - np)} \frac{npq}{\sqrt{npq}} \\ &= \frac{\Delta P_n(k)}{\frac{[(k+1) - np] - (k - np)}{\sqrt{npq}}} \frac{npq}{\sqrt{npq} \sqrt{npq}} = \frac{\Delta P_n(k)}{\frac{[(k+1) - np] - (k - np)}{\sqrt{npq}}} = \frac{\Delta P_n(k)}{\Delta x} \end{aligned}$$

Après un passage à la limite pour n tendant vers l'infini nous avons dans un premier temps pour le dénominateur du deuxième terme de la relation antérieure:

$$\frac{[-(k - np) - q] np}{(k+1) \sqrt{npq}} P_n(k) \quad (7.342)$$

la simplification suivante:

$$(k+1) \sqrt{npq} \underset{n \rightarrow +\infty}{=} k \sqrt{npq} \quad (7.343)$$

Donc:

$$\frac{-[(k - np) + q] np}{k \sqrt{npq}} P_n(k) = -\frac{k - np}{k \sqrt{npq}} np P_n(k) - \frac{q np}{k \sqrt{npq}} P_n(k) \quad (7.344)$$

et dans un second temps, tenant compte du fait que les valeurs de k considérées se trouvent alors au voisinage de l'espérance np , nous obtenons:

$$\frac{k - np}{k \sqrt{npq}} np = \frac{np}{k} \frac{k - np}{\sqrt{npq}} \underset{n \rightarrow +\infty}{\cong} \frac{k - np}{\sqrt{npq}} = x \quad (7.345)$$

et:

$$\frac{qnp}{k\sqrt{npq}} \underset{n \rightarrow +\infty}{\cong} \frac{qnp}{np\sqrt{npq}} = \frac{q}{\sqrt{npq}} \underset{n \rightarrow +\infty}{=} 0 \quad (7.346)$$

Donc:

$$\frac{-[(k - np) + q]np}{k\sqrt{npq}} \underset{n \rightarrow +\infty}{=} -x \quad (7.347)$$

et comme:

$$P_n(k) \underset{n \rightarrow +\infty}{=} F(x) \quad (7.348)$$

où $F(x)$ représentera pour les quelques lignes qui vont suivre, la fonction de densité lorsque n tend vers l'infini.

Nous avons finalement:

$$\frac{dF(x)}{dx} = -xF(x) \quad (7.349)$$

Cette relation peut encore s'écrire en réarrangeant les termes:

$$\frac{1}{F(x)} dF(x) = -x dx \quad (7.350)$$

et en intégrant les deux membres de cette égalité nous obtenons (cf. chapitre de Calcul Différentiel Et Intégral):

$$\ln(F(x)) = -\frac{x^2}{2} + c^{te} \quad (7.351)$$

La fonction suivante est une des solutions de la relation précédente:

$$F(x) = \lambda \cdot e^{-\frac{x^2}{2}} \quad (7.352)$$

Effectivement:

$$\ln\left(\lambda e^{-\frac{x^2}{2}}\right) = \ln(\lambda) + \ln\left(e^{-\frac{x^2}{2}}\right) = c^{te} + \left(-\frac{x^2}{2}\right) = -\frac{x^2}{2} + c^{te} \quad (7.353)$$

La constante est déterminée par la condition que:

$$\int_{-\infty}^{+\infty} F(x) dx \quad (7.354)$$

qui représente la somme de toutes les probabilités, vaille 1. Nous pouvons montrer pour cela que:

$$\lambda = \frac{1}{\sqrt{2\pi}} \quad (7.355)$$

Démonstration:

Nous avons:

$$\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx = \int_{-\infty}^{+\infty} e^{-\left(\frac{x}{\sqrt{2}}\right)^2} dx = \sqrt{2} \int_{-\infty}^{+\infty} e^{-z^2} dz \quad (7.356)$$

Donc concentrons-nous sur le dernier terme de l'égalité. Ainsi:

$$I = \int_{-\infty}^{+\infty} e^{-x^2} dx = 2 \int_0^{+\infty} e^{-x^2} dx \quad (7.357)$$

puisque e^{-x^2} est une fonction paire (cf. chapitre d'Analyse Fonctionnelle). Écrivons maintenant le carré de l'intégrale de la manière suivante:

$$I^2 = 4 \lim_{R \rightarrow +\infty} \left[\left(\int_0^R e^{-x^2} dx \right) \left(\int_0^R e^{-y^2} dy \right) \right] = 4 \lim_{R \rightarrow +\infty} \left(\int_0^R \int_0^R e^{-(x^2+y^2)} dx dy \right) \quad (7.358)$$

et faisons un changement de variable en passant en coordonnées polaires, dès lors nous faisons aussi usage du Jacobien dans ces mêmes coordonnées (cf. chapitre de Calcul Différentiel Et Intégral):

$$\begin{aligned} I^2 &= 4 \lim_{R \rightarrow +\infty} \left(\iint e^{-r^2} r dr d\phi \right) = 4 \lim_{R \rightarrow +\infty} \left(\int_0^{\pi/2} \int_0^R e^{-r^2} r dr d\phi \right) = 4 \frac{\pi}{2} \lim_{R \rightarrow +\infty} \left(\int_0^R e^{-r^2} r dr \right) \\ &= 2\pi \left(-\frac{1}{2} e^{-r^2} \Big|_0^{+\infty} \right) = 2\pi \left(0 + \frac{1}{2} \right) = \pi \Rightarrow I = \sqrt{\pi} \end{aligned} \quad (7.359)$$

Par extension pour $e^{-\frac{x^2}{2}}$ nous avons:

$$I = \lambda^{-1} = \sqrt{2\pi} \quad (7.360)$$

□C.Q.F.D.

Nous obtenons donc la "**loi normale centrée réduite**" notée sous forme de fonction de densité de probabilité (la notation avec le F majuscule peut malheureusement porter à confusion dans le cadre du présent développement avec la fonction de répartition...):

$$F(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (7.361)$$

qui peut être calculée dans la version française Microsoft Excel 11.8346 avec la fonction

LOI.NORMALE.STANDARD() ou pour la réciproque par LOI.NORMALE.STANDARD.INVERSE().

Pour information, une variable suivant une loi Normale centrée réduite est très souvent par tradition notée Z (pour "Zentriert" en allemand).

En revenant aux variables non normées:

$$x = \frac{k - np}{\sqrt{npq}} = \frac{k - \mu}{\sigma} \quad (7.362)$$

nous obtenons donc la "**fonction Gauss-Laplace**" (ou "**loi de Gauss-Laplace**") ou également appelée "**loi Normale**" donnée sous forme de densité de probabilité par:

$$P(k, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(k-\mu)^2}{2\sigma^2}} \quad (7.363)$$

et souvent notée $N(\mu, \sigma)$. Elle peut être calculée dans la version française de Microsoft Excel 11.8346 avec la fonction LOI.NORMALE() ou pour la réciproque par LOI.NORMALE.INVERSE().

La probabilité cumulée (fonction de répartition) de valoir une certaine valeur k étant bien évidemment donnée par:

$$P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(k-\mu)^2}{2\sigma^2}} dk \quad (7.364)$$

Voici un exemple de tracé de la fonction de distribution et répartition pour la fonction Normale de paramètres $(\mu, \sigma) = (0, 1)$:

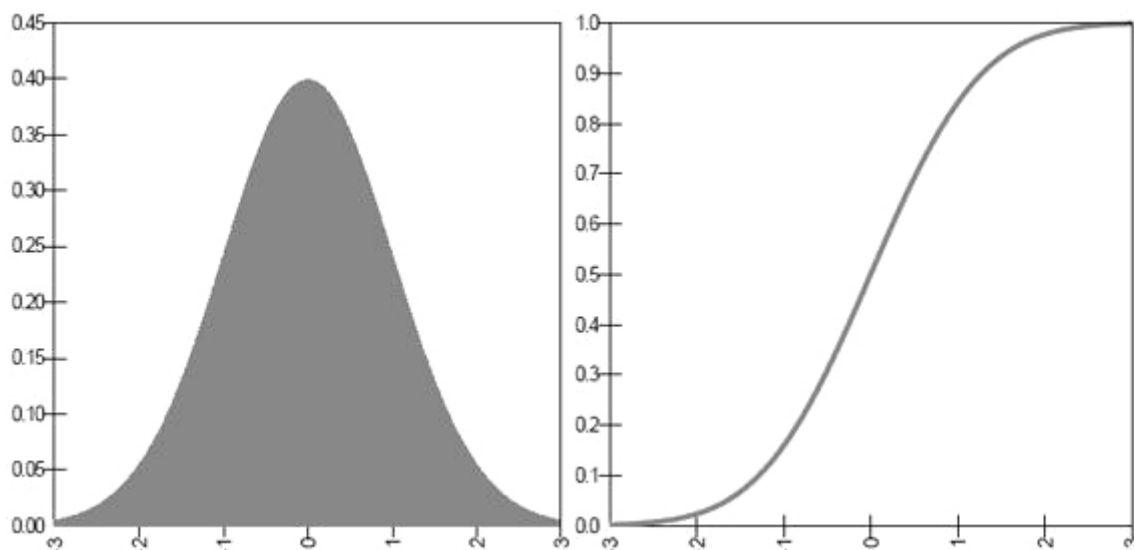


Figure: 7.15 - Loi Normale Centrée Réduite (fonction de distribution et de répartition)

Cette loi régit sous des conditions très générales, et souvent rencontrées, beaucoup de phénomènes aléatoires. Elle est par ailleurs symétrique par rapport à la moyenne μ (c'est important de s'en souvenir).

Montrons maintenant que μ représente bien l'espérance mathématique (ou la moyenne) de x (c'est un peu bête mais on peut quand même vérifier...):

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} xe^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \quad (7.365)$$

Posons:

$$u = \frac{x - \mu}{\sigma} \quad (7.366)$$

Nous avons dès lors:

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} xf(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\sigma u + \mu) e^{-\frac{1}{2}u^2} \sigma du \\ &= \sigma \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \sigma u e^{-\frac{1}{2}u^2} du + \sigma \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mu e^{-\frac{1}{2}u^2} du \quad (7.367) \\ &= \sigma \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} u e^{-\frac{1}{2}u^2} du + \frac{1}{\sqrt{2\pi}} \mu \int_{-\infty}^{+\infty} e^{-\frac{1}{2}u^2} du \end{aligned}$$

Calculons la première intégrale:

$$J = \int_{-\infty}^{+\infty} u e^{-\frac{1}{2}u^2} du = -e^{-\frac{u^2}{2}} \Big|_{-\infty}^{+\infty} = e^{-\frac{u^2}{2}} \Big|_{+\infty}^{-\infty} = 0 - 0 = 0 \quad (7.368)$$

Donc il vient au final:

$$E(X) = \frac{1}{\sqrt{2\pi}} \mu \int_{-\infty}^{+\infty} e^{-\frac{1}{2}u^2} du = \frac{1}{\sqrt{2\pi}} \mu \sqrt{2\pi} = \mu \quad (7.369)$$

Remarques:

R1. Le lecteur pourrait trouver cela déroutant dans un premier temps que le paramètre d'une fonction soit un des résultats que nous cherchons de la fonction. Ce qui dérange est la mise en pratique d'une telle chose. Au fait, tout s'éclairera lorsque nous étudierons plus loin dans ce chapitre les concepts "d'estimateurs de vraisemblance".

R2. Indiquons que dans la pratique (finance, qualité, assurance, etc.) il est fréquent de devoir calculer l'espérance uniquement pour des valeurs positives de la variable aléatoire qui est définie alors naturellement comme étant "l'espérance positive" et donnée par:

$$E^+(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{+\infty} x e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \quad (7.370)$$

Nous en verrons un exemple pratique dans le chapitre d'Économie lors de notre étude du modèle théorique de la spéculation de Louis Bachelier.

Montrons aussi (...) que σ représente bien l'écart-type de X (il convient, en d'autres termes de montrer que $V(X) = \sigma^2$) et pour cela rappelons que nous avons démontré que (relation de Huyghens):

$$V(X) = E(X^2) - E(X)^2 \quad (7.371)$$

Nous avons déjà calculé tout à l'heure $E(X) = \mu \Rightarrow E(X)^2 = \mu^2$ commençons alors par calculer $E(X^2)$:

$$E(X^2) = \int x^2 f(x) dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2 e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \quad (7.372)$$

Posons $y = (x - \mu) / \sqrt{2}\sigma$ qui conduit dès lors à:

$$\begin{aligned} E(X^2) &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} (y\sqrt{2}\sigma + \mu)^2 e^{-y^2} dy = \frac{2}{\sqrt{\pi}} \sigma^2 \int_{-\infty}^{+\infty} y^2 e^{-y^2} dy + 2\mu\sigma \sqrt{\frac{2}{\pi}} \int_{-\infty}^{+\infty} y e^{-y^2} dy \\ &\quad + \frac{\mu^2}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{-y^2} dy \end{aligned} \quad (7.373)$$

Or, nous savons:

$$\int_{-\infty}^{+\infty} y e^{-y^2} dy = 0 \quad \int_{-\infty}^{+\infty} e^{-y^2} dy = \sqrt{\pi} \quad (7.374)$$

Il reste donc à calculer la première intégrale. Pour cela, procédons par une intégration par parties (cf.

chapitre de Calcul Différentiel et Intégral):

$$\int_a^b f(t)g'(t)dt = f(t)g(t)\Big|_a^b - \int_a^b f'(t)g(t)dt$$

$$\int_{-\infty}^{+\infty} y^2 e^{-y^2} dy = \int_{-\infty}^{+\infty} y \left(y e^{-y^2} \right) dy = -y \frac{e^{-y^2}}{2} \Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} \frac{e^{-y^2}}{2} dy = \frac{\sqrt{\pi}}{2} \quad (7.375)$$

D'où:

$$E(X^2) = \frac{2}{\sqrt{\pi}} \sigma^2 \frac{\sqrt{\pi}}{2} + 2\mu\sigma \sqrt{\frac{2}{\pi}} 0 + \frac{\mu^2}{\sqrt{\pi}} \sqrt{\pi} = \mu^2 + \sigma^2 \quad (7.376)$$

Il vient finalement:

$$V(X) = E(X^2) - E(X)^2 = \sigma^2 \quad (7.377)$$

Une signification supplémentaire de l'écart-type dans la loi de Gauss-Laplace est une mesure de la largeur de la distribution telle que (cela ne peut se vérifier qu'à l'aide d'intégration à l'aide de méthodes numériques) que toute moyenne et pour tout écart-type non nul nous avons:

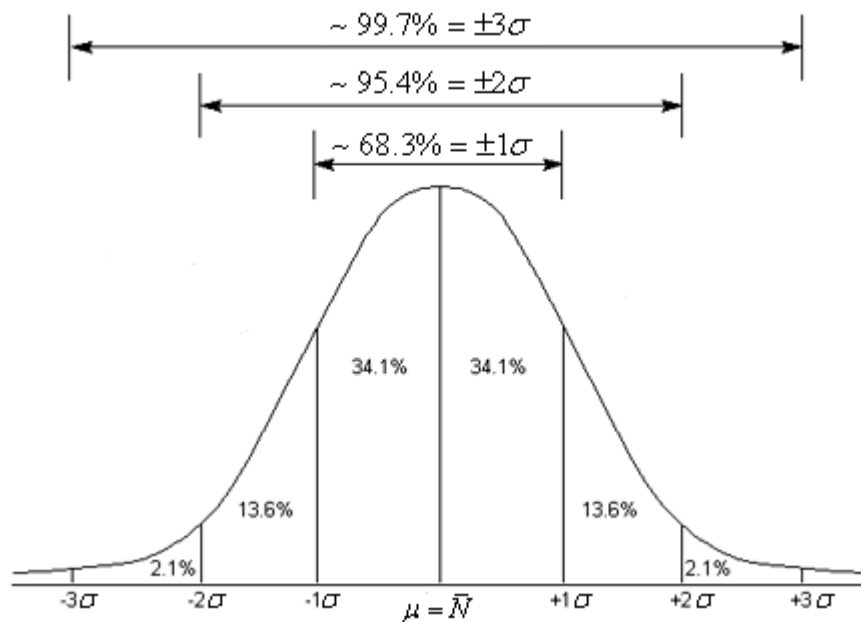


Figure: 7.16 - Intervalles sigma de la loi Normale

La largeur de l'intervalle a une très grande importance dans l'interprétation des incertitudes d'une mesure. La présentation d'un résultat comme $\bar{N} \pm \sigma$ signifie que la valeur moyenne a environ 68.3% de chance (probabilité) de se trouver entre les limites de $\bar{N} - \sigma$ et $\bar{N} + \sigma$, ou qu'elle a environ 95.4% de se trouver entre $\bar{N} - 2\sigma$ et $\bar{N} + 2\sigma$ etc.

Remarque: Ce concept est beaucoup utilisé en gestion de la qualité en entreprise particulièrement avec le concept industriel anglo-saxon Six Sigma (cf. chapitre de Génie Industriel) qui impose une maîtrise de 6σ autour de chaque côté (!) de la moyenne des côtés des pièces fabriquées (ou tout autre sujet dont on mesure la déviation).

Niveau de qualité Sigma	Taux de non-défection assuré en %	Taux de défection en parties par million
1 σ	68.26894	317'311
2 σ	95.4499	45'500
3 σ	99.73002	2'700
4 σ	99.99366	63.4
5 σ	99.999943	0.57
6 σ	99.9999998	0.002

Tableau: 7.7 - Niveau de qualité Sigma avec taux de défection/non-défection

La deuxième colonne du tableau peut facilement être obtenue avec Maple 4.00b. Par exemple pour la première ligne:

```
>S:=evalf(int(1/sqrt(2*Pi)*exp(-x^2/2),x=-1..1));
```

et la première ligne de la troisième colonne par:

```
>(1-S)*1E6;
```

Si la loi Normale était décentrée, il suffirait alors d'écrire pour la deuxième colonne:

```
>S:=evalf(int(1/sqrt(2*Pi)*exp(-(x-mu)^2/2),x=-1..1));
```

et ainsi de suite pour tout écart-type et toute moyenne on retombera sur les mêmes intervalles!!!

La loi de Gauss-Laplace n'est par ailleurs pas qu'un outil d'analyse de données mais également de génération de données. Effectivement, cette loi est une des plus importantes dans le monde des multinationales qui recourent aux outils statistiques pour la gestion du risque, la gestion de projets et la simulation lorsqu'un grand nombre de variables aléatoires sont en jeu. Le meilleur exemple d'application en étant le logiciel CrystalBall ou @Risk de Palisade (mon préféré...).

Dans ce cadre d'application (gestion de projets), il est par ailleurs très souvent fait usage de la somme (durée des tâches) ou le produit de variables aléatoires (facteur d'incertitude du client) suivant des lois de Gauss-Laplace. Voyons comment cela se calcule:

5.9.1. SOMME DE DEUX VARIABLES ALÉATOIRES NORMALES

Soient X, Y deux variables aléatoires indépendantes. Supposons que X suit la loi $N(\mu_1, \sigma_1)$ et que Y suit la loi $N(\mu_2, \sigma_2)$. Alors, la variable aléatoire $Z = X + Y$ aura une densité égale au produit de convolution de f_X et f_Y . C'est-à-dire:

$$f_Z(s) = \int_{-\infty}^{+\infty} f_X(x) f_Y(s-x) dx = \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{+\infty} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \cdot e^{-\frac{(s-x-\mu_2)^2}{2\sigma_2^2}} dx \quad (7.378)$$

ce qui équivaut à faire le produit conjoint (cf. chapitre de Probabilités) des probabilités d'apparition des deux variables continues (se rappeler le même genre de calcul sous forme discrète!)

Pour simplifier l'expression, faisons le changement de variable $t = x - \mu_1$ et posons $a = \mu_1 + \mu_2 - s$, $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$.

Comme:

$$(s-x-\mu_2)^2 = \left[-(-s+x+\mu_2)^2\right] = (-s+x+\mu_2)^2 = (t+a)^2 \quad (7.379)$$

nous obtenons:

$$\begin{aligned} f_Z(s) &= \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2\sigma_1^2}} \cdot e^{-\frac{(t+a)^2}{2\sigma_2^2}} dt = \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{+\infty} e^{-\frac{\left(\sigma t + \frac{a\sigma_1^2}{\sigma}\right)^2 + \frac{\sigma_1^2 a^2 \sigma_2^2}{\sigma^2}}{2\sigma_1^2\sigma_2^2}} dt \\ &= \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{a^2}{2\sigma^2}} \int_{-\infty}^{+\infty} e^{-\frac{\left(\sigma t + \frac{a\sigma_1^2}{\sigma}\right)^2}{2\sigma_1^2\sigma_2^2}} dt \end{aligned} \quad (7.380)$$

Nous posons:

$$u = \frac{\sigma t + \frac{a\sigma_1^2}{\sigma}}{\sqrt{2}\sigma_1\sigma_2} \Rightarrow \frac{du}{dt} = \frac{\sigma}{\sqrt{2}\sigma_1\sigma_2} \Rightarrow dt = \sqrt{2}\sigma_1\sigma_2 \frac{du}{\sigma} \quad (7.381)$$

Alors:

$$f_Z(s) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{\alpha^2}{2\sigma^2}} \int_{-\infty}^{+\infty} e^{-\frac{\left(\sigma t + \frac{\alpha\sigma_1^2}{\sigma}\right)^2}{2\sigma_1^2\sigma_2^2}} dt = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\alpha^2}{2\sigma^2}} \int_{-\infty}^{+\infty} e^{-u^2} du \quad (7.382)$$

Sachant que:

$$\int_{-\infty}^{+\infty} e^{-u^2} du = \sqrt{\pi} \quad (7.383)$$

et:

$$\alpha^2 = (-s + \mu_1 + \mu_2)^2 = [-(s - \mu_1 - \mu_2)]^2 = (s - \mu_1 - \mu_2)^2 \quad (7.384)$$

notre expression devient:

$$f_Z(s) = \frac{1}{\sigma\sqrt{2}\sqrt{\pi}} e^{-\frac{(s-\mu_1-\mu_2)^2}{2\sigma^2}} \quad (7.385)$$

Nous reconnaissons l'expression de la loi de Gauss-Laplace de moyenne $\mu_1 + \mu_2$ et d'écart type

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}.$$

Par conséquent, $X + Y$ suit la loi:

$$N\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right) \quad (7.386)$$

Le fait que la somme de deux lois Normales donne toujours une loi Normale est ce que nous nommons en statistiques la "**stabilité par la somme**" de la loi de Gauss-Laplace. Nous retrouverons ce type de propriétés pour d'autres lois que nous étudierons plus loin.

Remarque: Les familles de lois stables par addition constituent un domaine important d'étude en physique, finance et statistiques appelé "**distribution de Lévy alpha-stables**". Si le temps me le permet, je présenterai les détails de ce domaine d'étude extrêmement important dans le présent chapitre.

5.9.2. PRODUIT DE DEUX VARIABLES ALÉATOIRES NORMALES

Soient X, Y deux variables aléatoires indépendantes réelles. Nous désignerons par f_X et f_Y les densités correspondantes et nous cherchons à déterminer la densité de la variable $Z = X \cdot Y$.

Notons f la fonction de densité du couple (X, Y) . Vu que X, Y sont indépendantes (cf. chapitre de Probabilités):

$$f(x, y) = f_X(x) \cdot f_Y(y) \quad (7.387)$$

La fonction de répartition de Z est:

$$F(z) = P(Z \leq z) = P(X \cdot Y \leq z) = \iint_D f(x, y) dx dy = \iint_D f_X(x) \cdot f_Y(y) dx dy \quad (7.388)$$

où $D = \{(x, y) \mid x \cdot y \leq z\}$.

D peut se réécrire comme union disjointe (nous faisons cette opération pour anticiper lors du futur changement de variables une division par zéro):

$$D = D_1 \sqcup D_2 \sqcup D_3 \quad (7.389)$$

avec:

$$\begin{aligned} D_1 &= \{(x, y) \in \mathbb{R}^2 \mid x \cdot y \leq z \text{ et } x > 0\} \\ D_2 &= \{(x, y) \in \mathbb{R}^2 \mid x \cdot y \leq z \text{ et } x < 0\} \\ D_3 &= \{(x, y) \in \mathbb{R}^2 \mid x \cdot y \leq z \text{ et } x = 0\} \end{aligned} \quad (7.390)$$

Nous avons:

$$F(z) = \iint_{D_1} f_X(x) \cdot f_Y(y) dx dy + \iint_{D_2} f_X(x) \cdot f_Y(y) dx dy + \underbrace{\iint_{D_3} f_X(x) \cdot f_Y(y) dx dy}_{=0} \quad (7.391)$$

La dernière intégrale vaut zéro car D_3 est de mesure (épaisseur) nulle pour l'intégrale selon x .

Nous effectuons ensuite le changement de variable suivant:

$$\begin{cases} x = x \\ t = x \cdot y \end{cases} \quad (7.392)$$

Le jacobien de la transformation est:

$$\left| \begin{bmatrix} 1 & 0 \\ -t/x^2 & 1/x \end{bmatrix} \right| = \frac{1}{|x|} \quad (7.393)$$

Donc:

$$\begin{aligned} F(z) &= \int_{-\infty}^z \int_0^{+\infty} \frac{f_X(x) \cdot f_Y(t/x)}{|x|} dx dt + \int_{-\infty}^z \int_{-\infty}^0 \frac{f_X(x) \cdot f_Y(t/x)}{|x|} dx dt \\ &= \int_{-\infty}^z \int_{-\infty}^{+\infty} \frac{f_X(x) \cdot f_Y(t/x)}{|x|} dx dt \end{aligned} \quad (7.394)$$

Notons f_Z la densité de la variable Z . Par définition:

$$F(z) = \int_{-\infty}^z f_Z(t) dt \quad (7.395)$$

D'un autre côté:

$$F(z) = \int_{-\infty}^z \int_{-\infty}^{+\infty} \frac{f_X(x) \cdot f_Y(t/x)}{|x|} dx dt \quad (7.396)$$

comme nous venons de le voir. Par conséquent:

$$f_Z(t) = \int_{-\infty}^{+\infty} \frac{f_X(x) \cdot f_Y(t/x)}{|x|} dx \quad (7.397)$$

Ce qui est un peu triste c'est que dans le cas d'une loi de Gauss-Laplace (loi Normale), cette intégrale ne peut être calculée simplement que numériquement... il faut alors faire appel à des méthodes d'intégration du type Monte-Carlo (cf. chapitre de Méthodes Numériques).

D'après quelques recherches faites sur Internet cependant, mais sans certitude, cette intégrale pourrait être calculée et donnerait une nouvelle loi appelée "**loi de Bessel**".

5.9.3. LOI NORMALE CENTRÉE RÉDUITE

La fonction de Gauss-Laplace n'est pas tabulée puisqu'il faudrait autant de tables numériques que de valeurs possibles pour la moyenne μ et l'écart-type σ (qui sont donc des paramètres de la fonction comme nous l'avons vu).

C'est pourquoi, en opérant un changement de variable, la loi Normale devient la "**loi Normale centrée réduite**" où:

1. "Centrée" signifie soustraire la moyenne μ (la fonction a alors pour axe de symétrie l'axe des ordonnées).
2. "Réduite" signifie, diviser par l'écart-type σ .

Par ce changement de variable, la variable k est remplacée par la variable aléatoire centrée réduite:

$$k^* = \frac{k - \mu}{\sigma} \quad (7.398)$$

Si la variable k a pour moyenne μ et pour écart-type σ alors la variable $k^* = \frac{k - \mu}{\sigma}$ a pour moyenne 0 et pour écart-type 1.

Donc la relation:

$$P(k, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(k-\mu)^2}{2\sigma^2}} \quad (7.399)$$

s'écrit alors (trivialement) plus simplement:

$$P(k^*, 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{k^{*2}}{2}} \quad (7.400)$$

qui n'est d'autre que l'expression de la loi Normale centrée réduite souvent notée $N(0,1)$ que nous retrouverons très fréquemment dans les chapitres relatifs à la physique, la finance, la gestion et l'ingénierie!

Remarque: Calculer l'intégrale de la relation précédente entre n'importe quelles bornes n'est pas possible formellement parlant de manière exacte. Une idée possible et simple consiste alors à exprimer l'exponentielle en série de Taylor et de faire ensuite l'intégration terme par terme de la série (en s'assurant de prendre suffisamment de termes pour la convergence!).

5.9.4. DROITE DE HENRY

Souvent, dans les entreprises c'est la loi de Gauss-Laplace (Normale) qui est analysée mais des logiciels courants et facilement accessibles comme Microsoft Excel sont incapables de vérifier que les données mesurées suivent une loi Normale lorsque nous faisons de l'analyse fréquentielle (aucun outil intégré par défaut ne permet de le faire) et que nous n'avons pas les données d'origines non groupées.

L'astuce consiste alors à utiliser la variable centrée réduite qui se construit comme nous l'avons démontré plus haut avec la relation suivante:

$$k^* = \frac{k - \mu}{\sigma} \quad (7.401)$$

L'idée de la droite d'Henry est alors d'utiliser la relation linéaire entre k et k^* donnée par l'équation de la droite:

$$k^* = f(k) = \frac{k}{\sigma} - \frac{\mu}{\sigma} \quad (7.402)$$

et qui peut être tracée pour déterminer la moyenne et l'écart-type de la loi Normale.

Exemple:

Supposons que nous ayons l'analyse fréquentielle suivante de 10'000 tickets de caisse dans un supermarché:

Montant des tickets	Nombre de tickets	Nombre cumulé de tickets	Fréquences relatives cumulées
[0,50[	668	668	0.068
[50,100[	919	1'587	0.1587
[100,150[	1'498	3'085	0.3085
[150,200[	1'915	5'000	0.5000
[200,250[	1'915	6'915	0.6915
[250,300[	1'498	8'413	0.8413
[300,350[	919	9'332	0.9332
[350,400[	440	9'772	0.9772
[400 et +	228	10'000	1

Tableau: 7.8 - Intervalles de classe pour la détermination de la droite de Henry

Si nous traçons maintenant cela sous Microsoft Excel 11.8346 nous obtenons:

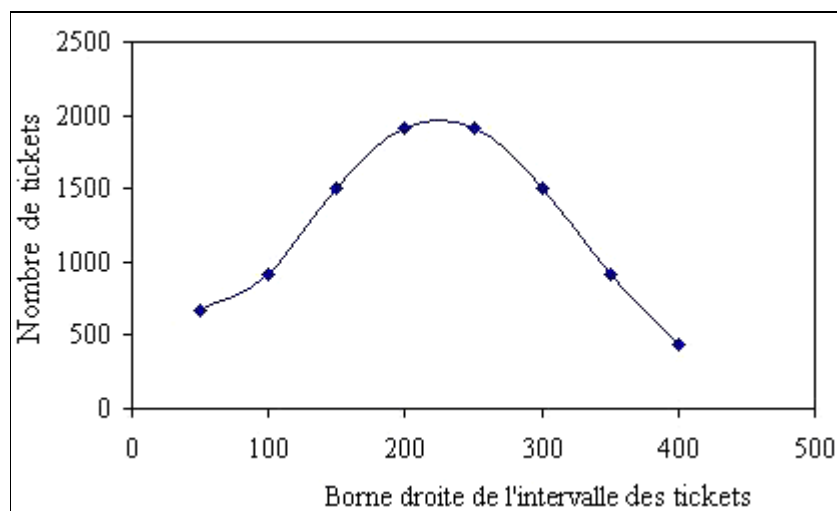


Figure: 7.17 - Distribution des ventes de tickets

Ce qui ressemble terriblement à une loi Normale d'où l'autorisation, sans trop de risques, d'utiliser dans cet exemple la technique de la droite d'Henry.

Mais que faire maintenant? Eh bien connaissant les fréquences cumulées, il ne nous reste plus qu'à calculer pour chacune d'entre elles k^* à l'aide de tables numériques ou avec la fonction `NORMSINV()` de la version anglaise de Microsoft Excel 11.8346 (car rappelons que l'intégration formelle de la fonction gaussienne n'est pas des plus faciles...).

Ceci nous donnera les valeurs de la loi Normale centrée réduite $N(0,1)$ de ces mêmes fréquences respectives cumulées (fonction de répartition). Ainsi nous obtenons (nous laissons le soin au lecteur de chercher sa table numérique ou d'ouvrir son logiciel préféré...):

Borne supérieure de l'intervalle	Fréquences relatives cumulées	Correspondance pour k^* de $N(0,1)$
50	0.068	-1.5
100	0.1587	-1
150	0.3085	-0.5
200	0.5000	0
250	0.6915	0.5
300	0.8413	1
350	0.9332	1.5
400	0.9772	2
-	1	-

Tableau: 7.9 - Fréquences relatives cumulées pour la droite de Henry

Signalons que dans le type de tableau ci-dessus, dans Microsoft Excel, les valeurs de fréquences cumulées nulles et unitaires (extrêmes) posent problèmes. Il faut alors jouer un petit peu...

Comme nous l'avons spécifié plus haut, nous avons sous forme discrète:

$$k_i^* = f(k_i) = \frac{k_i}{\sigma} - \frac{\mu}{\sigma} \quad (7.403)$$

Donc graphiquement sous Microsoft Excel 11.8346 nous obtenons grâce à notre tableau le graphique suivant (évidemment en toute rigueur on fera une régression dans les règles de l'art comme vu dans le chapitre de Méthodes Numériques avec intervalles de confiance, de prédiction et tout le toutim...):

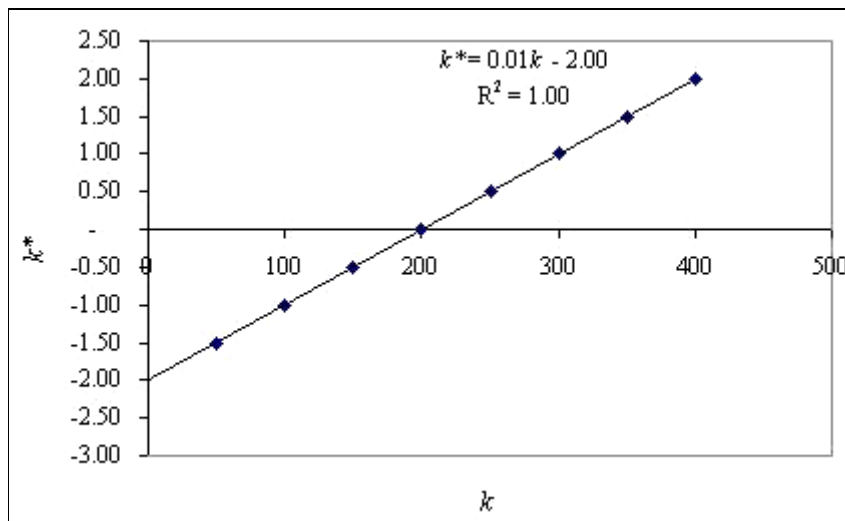


Figure: 7.18 - Forme linéarisée de la distribution

Donc à l'aide de la régression donnée par Microsoft Excel 11.8346 (ou calculée par vos soins selon les techniques de régressions linéaires vues dans le chapitre de Méthodes Numériques). Il vient:

$$k^* = \frac{k}{\sigma} - \frac{\mu}{\sigma} = 0.01k - 2 \quad (7.404)$$

dont nous déduisons immédiatement:

$$\sigma = 100 \quad \mu = 200 \quad (7.405)$$

Il s'agit donc d'une technique particulière pour une distribution particulière! Des techniques similaires plus ou moins simples (ou compliquées suivant les cas) existent pour nombre de distributions.

Voyons une autre manière approximative d'aborder le problème. Reprenons pour cet exemple notre tableau:

Tickets	Borne droite	Centre	Fréquences relatives cumulées en%
[0,50[	50	25	6.8
[50,100[	100	75	15.87
[100,150[	150	125	30.85
[150,200[	200	175	50.00
[200,250[	250	225	69.15
[250,300[	300	275	84.13
[300,350[	350	325	93.32
[350,400[	400	375	97.72
[400 et +	-	-	100

La moyenne sera maintenant calculée à l'aide de la valeur centrale des intervalles et des effectifs selon la relation vue au début de ce chapitre:

$$\hat{\mu} = \frac{\sum_{i=1}^N n_i x_i}{\sum_{i=1}^N n_i} \quad (7.406)$$

Tickets	Centre	Tickets (fréquence)	Calcul
[0,50[	25	668	16'700
[50,100[	75	919	68'925
[100,150[	125	1'498	187'250
[150,200[	175	1'915	335'125
[200,250[	225	1'915	430'875
[250,300[	275	1'498	411'950
[300,350[	325	919	298'675
[350,400[	375	440	165'000
[400 et +	-	-	-
	Somme:	9772	1'914'500
		Moyenne:	1'914'500/9'772 =195.92

La moyenne expérimentale est donc assez proche de la moyenne théorique obtenue précédemment avec la droite de Henry.

L'écart-type sera maintenant calculé à l'aide de la valeur centrale des intervalles et des effectifs selon la relation vue aussi au début de ce chapitre:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^N n_i (x_i - \bar{x})^2}{\sum_{i=1}^N n_i}} \quad (7.407)$$

Tickets	Centre	Tickets (fréquence)	Calcul
[0,50[	25	668	1997.00
[50,100[	75	919	1375.08
[100,150[	125	1'498	771.02
[150,200[	175	1'915	85.76
[200,250[	225	1'915	165.71
[250,300[	275	1'498	958.65
[300,350[	325	919	1566.93
[350,400[	375	440	1443.98
[400 et +	-	228	-
		Variance:	8364.16
		Ecart-Type:	91.45

L'écart-type expérimental est donc assez proche de l'écart-type théorique obtenu avec la méthode de la droite de Henry.

5.9.5. DIAGRAMME QUANTILE-QUANTILE

Une autre manière de juger qualitativement de l'ajustement de données expérimentales avec une loi théorique (quelle qu'elle soit!!!) est l'utilisation d'un "**diagramme quantile-quantile**".

L'idée est assez simple, il s'agit de comparer les données expérimentales, aux données théoriques supposées suivre une loi donnée. Ainsi, dans le cas de notre exemple nous avons en prenant les valeurs de la moyenne théorique (~200) et l'écart-type théorique (~100) obtenus avec la droite de Henry:

Tickets	Borne de droite expérimentale (imposée)	Fréquences relatives cumulées	Borne de droite théorique (calculée)
[0,50[	50	6.80%	50.91
[50,100[	100	15.87%	100.02
[100,150[	150	30.85%	149.99
[150,200[	200	50.00%	200
[200,250[	250	69.15%	250.01
[250,300[	300	84.13%	299.98
[300,350[	350	93.32%	350.00
[350,400[	400	97.72%	399.90
[400 et +	-	100%	-

Représenté graphiquement, cela nous donne donc le fameux diagramme quantile-quantile:

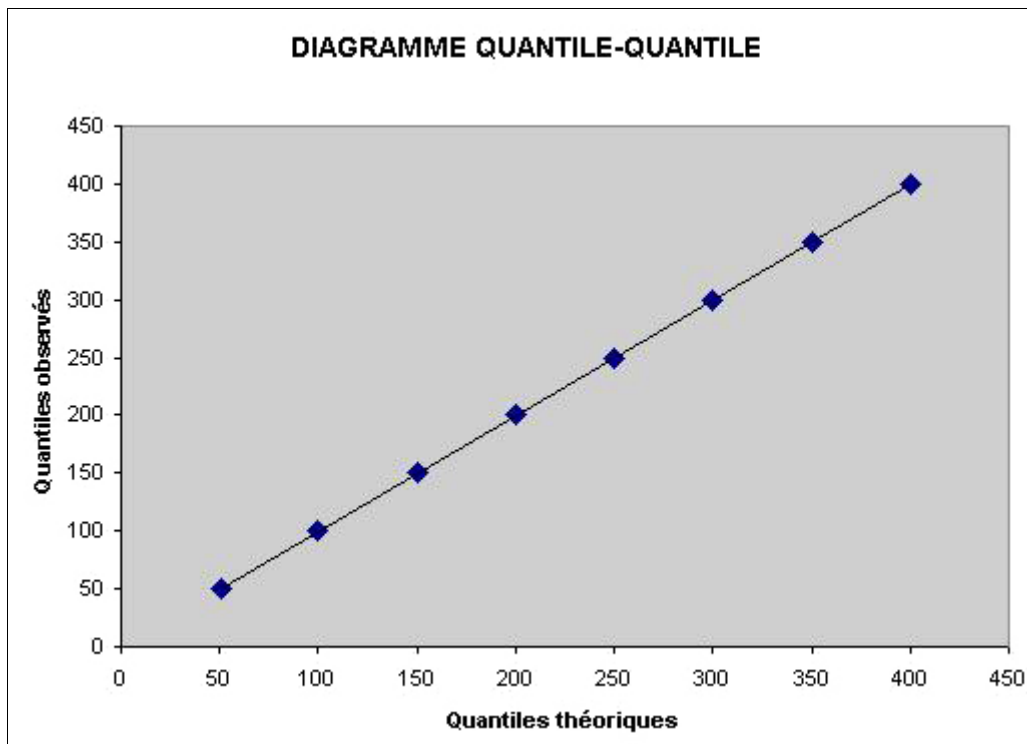


Figure: 7.19 - Diagramme quantile-quantile de la distribution

Et bien évidemment on peut comparer les quantiles observés à toute loi théorique supposée. Plus les points seront alignés sur la droite, meilleur sera l'ajustement! C'est très visuel, très simple et beaucoup utilisé par les non spécialistes en statistiques dans les entreprises.

5.10. FONCTION LOG-NORMALE

Nous disons qu'une variable aléatoire positive X suit une "**fonction log-normale**" (ou "**loi log-normale**") de paramètres $\mu, \sigma > 0$ (moments de la loi log-Normale), si et seulement si en posant:

$$y = \ln(X) \quad (7.408)$$

nous voyons que y suit une fonction de probabilité de type loi Normale de moyenne μ et de variance σ^2 (moments de la loi Normale).

In extenso, de par les propriétés des logarithmes, une variable peut être modélisée par une loi log-normale si elle est le résultat de la multiplication d'un grand nombre de petits facteurs indépendants.

La fonction de densité de X pour $x \geq 0$ est alors (cf. chapitre de Calcul Intégral):

$$f(x) = \frac{1}{\sigma x \sqrt{2\pi}} \exp\left(-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right) \quad (7.409)$$

qui peut être calculée dans la version française de Microsoft Excel 11.8346 avec la fonction LOI.LOGNORMALE() ou pour la réciproque par LOI.LOGNORMALE.INVERSE().

Ce type de scénario se retrouve fréquemment en physique, dans les techniques de maintenance ou encore en finance des marchés dans le modèle de pricing des options (voir ces chapitres respectifs du

site pour des exemples concrets). Il y a par ailleurs une remarque importante relativement à la loi log-normale dans le traitement plus loin du théorème central limite!

Montrons que la fonction de probabilité cumulée correspond bien à une loi Normale si nous faisons le changement de variable mentionné précédemment:

$$\int_0^{+\infty} f(x) dx = \int_0^{+\infty} \frac{1}{\sigma x \sqrt{2\pi}} \exp\left(-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right) dx = \frac{1}{\sigma \sqrt{2\pi}} \int_0^{+\infty} \frac{1}{x} \exp\left(-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right) dx \quad (7.410)$$

en posant:

$$y = \ln(x) \Rightarrow \frac{dy}{dx} = \frac{1}{x} \Rightarrow dx = x dy \quad (7.411)$$

et:

$$x = e^y \quad (7.412)$$

nous avons bien:

$$\begin{aligned} \int_0^{+\infty} f(x) dx &= \frac{1}{\sigma \sqrt{2\pi}} \int_0^{+\infty} \frac{1}{x} \exp\left(-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right) dx = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{1}{x} e^{-\frac{(y-\mu)^2}{2\sigma^2}} x dy \\ &= \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy \end{aligned} \quad (7.413)$$

nous tombons donc bien sur une loi Normale!

L'espérance (moyenne) de X est donnée alors par (le logarithme népérien n'étant pas défini pour $x < 0$ nous bornons l'intégrale à partir de zéro):

$$\begin{aligned} E(X) &= \int_0^{+\infty} x f(x) dx = \frac{1}{\sigma \sqrt{2\pi}} \int_0^{+\infty} \exp\left(-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right) dx \\ &= \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{(u - \mu)^2}{2\sigma^2} + u\right) du \end{aligned} \quad (7.414)$$

où nous avons effectué le changement de variable:

$$u = \ln(x) \Leftrightarrow x = e^u, x du = dx \quad (7.415)$$

L'expression:

$$-\frac{(u - \mu)^2}{2\sigma^2} + u \quad (7.416)$$

étant par ailleurs égale à:

$$-\frac{1}{2\sigma^2} \left(\left(u - (\mu + \sigma^2) \right)^2 - (\mu + \sigma^2)^2 + \mu^2 \right) \quad (7.417)$$

la dernière intégrale devient donc:

$$\begin{aligned} E(X) &= \frac{\exp\left(\frac{(\mu + \sigma^2)^2 - \mu^2}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2} \left(u - (\mu + \sigma^2) \right)^2\right) du \quad (7.418) \\ &= \exp\left(\frac{(\mu + \sigma^2)^2 - \mu^2}{2\sigma^2}\right) = \exp\left(\mu + \frac{\sigma^2}{2}\right) \end{aligned}$$

et où nous avons utilisé la propriété qui a émergée lors de notre étude de la loi Normale, c'est-à-dire que toute intégrale de la forme:

$$\int_{-\infty}^{+\infty} e^{-\frac{(x-c)^2}{2\sigma^2}} dx = \sigma\sqrt{2\pi} \quad (7.419)$$

a donc toujours la même valeur!

Pour le calcul de la variance, rappelons que pour une variable aléatoire X celle-ci est définie par:

$$V(X) = E(X^2) - E(X)^2 \quad (7.420)$$

Calculons $E(X^2)$ en procédant de manière similaire aux développements précédents:

$$\begin{aligned}
E(X^2) &= \int_0^{+\infty} x^2 f(x) dx = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{+\infty} x \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right) dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^u \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) e^u du = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2} + 2u\right) du \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2} \left(u - (\mu + 2\sigma^2)\right)^2 - (\mu + 2\sigma^2)^2 + \mu^2\right) du \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{(u - (\mu + 2\sigma^2))^2}{2\sigma^2}\right) \exp\left(\frac{(\mu + 2\sigma^2)^2 - \mu^2}{2\sigma^2}\right) du \\
&= \frac{\exp\left(\frac{(\mu + 2\sigma^2)^2 - \mu^2}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{(u - (\mu + 2\sigma^2))^2}{2\sigma^2}\right) du \\
&= \exp\left(\frac{(\mu + 2\sigma^2)^2 - \mu^2}{2\sigma^2}\right) = \exp\left(\frac{4\mu\sigma^2 + 4\sigma^4}{2\sigma^2}\right) = \exp(2\mu + 2\sigma^2) = e^{2(\mu + \sigma^2)}
\end{aligned} \tag{7.421}$$

où nous avons encore une fois le changement de variable:

$$u = \ln(x) \Leftrightarrow x = e^u, x du = dx \tag{7.422}$$

et où nous avons transformé l'expression:

$$-\frac{(u-\mu)^2}{2\sigma^2} + 2u \tag{7.423}$$

sous la forme:

$$-\frac{1}{2\sigma^2} \left((u - (\mu + 2\sigma^2))^2 - (\mu + 2\sigma^2)^2 + \mu^2 \right) \tag{7.424}$$

Donc:

$$\begin{aligned}
V(X) &= E(X^2) - E(X)^2 = \exp(2\mu + 2\sigma^2) - \left(\exp\left(\mu + \frac{\sigma^2}{2}\right) \right)^2 \\
&= \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2) = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1)
\end{aligned} \tag{7.425}$$

Voici un exemple de tracé de la fonction de distribution et répartition pour la fonction Log-Normale de paramètres $(\mu, \sigma) = (0, 1)$:

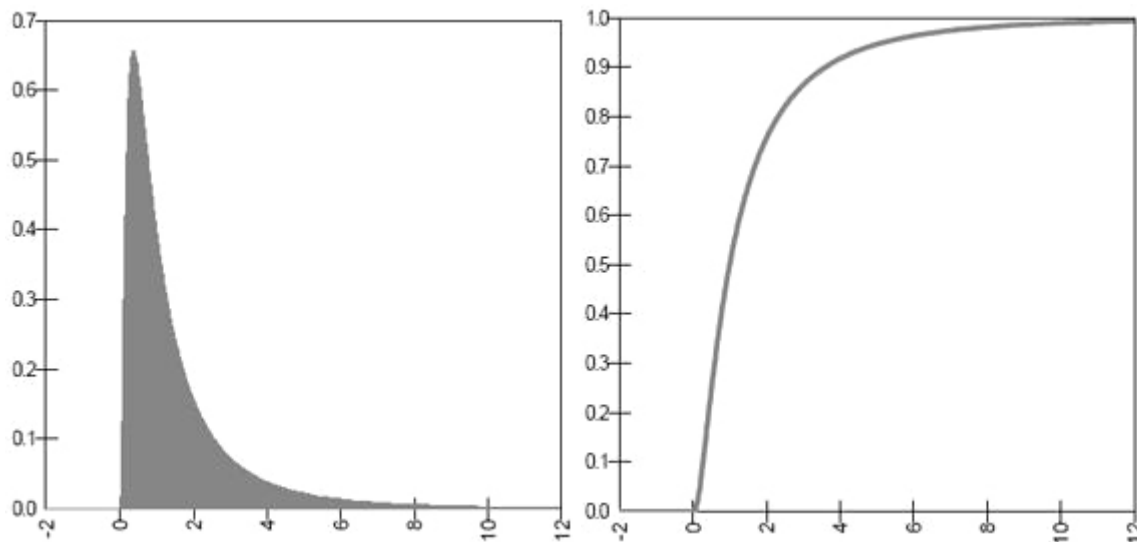


Figure: 7.20 - Loi Log-Normale (fonction de distribution et de répartition)

5.11. FONCTION UNIFORME CONTINUE

Soient $a < b$. Nous définissons la fonction de distribution de la "**fonction uniforme**" (ou "**loi uniforme**") par la relation:

$$P_{a,b}(x) = \frac{1}{b-a} \cdot 1_{[a,b]} \quad (7.426)$$

Nous avons donc pour fonction de répartition:

$$P(X \leq x) = \int \frac{1}{b-a} 1_{[a,b]} dx = \frac{1}{b-a} 1_{[a,b]} \int_a^x dx = \frac{x_{[a,b]} - a}{b-a}$$

Il s'agit bien d'une fonction de distribution car elle vérifie (intégrale simple):

$$\int_{-\infty}^{+\infty} P_{a,b}(x) dx = \int_{-\infty}^{+\infty} \frac{1_{[a,b]}}{b-a} dx = \frac{1_{[a,b]}}{b-a} \int_a^b dx = \frac{1_{[a,b]}}{b-a} x \Big|_a^b = 1_{[a,b]} \frac{b-a}{b-a} = 1 \quad (7.427)$$

La fonction uniforme a par ailleurs pour espérance (moyenne):

$$\begin{aligned} \mu = E(X) &= \int x f(x) dx = \frac{1}{b-a} \int_a^b x dx = \frac{1}{b-a} \frac{x^2}{2} \Big|_a^b = \frac{1}{b-a} \frac{b^2 - a^2}{2} \\ &= \frac{1}{b-a} \frac{(b+a)(b-a)}{2} = \frac{a+b}{2} \end{aligned} \quad (7.428)$$

et pour variance en utilisant la relation de Huyghens:

$$\begin{aligned}
 V(X) &= E(X^2) - E(X)^2 = \int_a^b x^2 f(x) dx - \left(\frac{a+b}{2} \right)^2 = \frac{1}{b-a} \int_a^b x^2 dx - \left(\frac{a+b}{2} \right)^2 \\
 &= \frac{1}{b-a} \frac{x^3}{3} \Big|_a^b - \left(\frac{a+b}{2} \right)^2 = \frac{1}{b-a} \frac{b^3 - a^3}{3} - \left(\frac{a+b}{2} \right)^2 \\
 &= \frac{1}{b-a} \frac{(b-a)(b^2 + ab + a^2)}{3} - \left(\frac{a+b}{2} \right)^2 \quad (7.429) \\
 &= \frac{1}{3} (b^2 + ab + a^2) - \frac{1}{4} (b^2 + 2ab + a^2) = \frac{4(b^2 + ab + a^2) - 3(b^2 + 2ab + a^2)}{12} \\
 &= \frac{4b^2 + 4ab + 4a^2 - 3b^2 - 6ab - 3a^2}{12} = \frac{b^2 - 2ab - a^2}{12} = \frac{(b-a)^2}{12}
 \end{aligned}$$

$1_{[a,b]}$ signifie qu'en dehors du domaine de définition $[a,b]$ la fonction de distribution est nulle. Nous retrouverons ce type de notation dans certaines autres fonctions de distribution.

Voici un exemple de tracé de la fonction de distribution et respectivement de répartition pour la loi Uniforme continue de paramètres $(\alpha, b) = (0, 1)$:

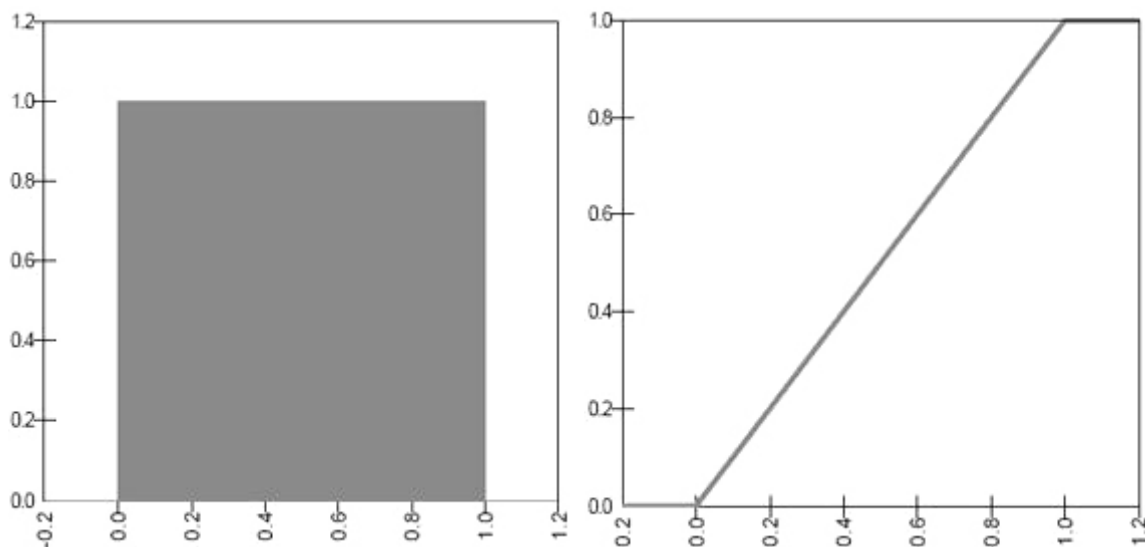


Figure: 7.21 - Loi uniforme continue (fonction de distribution et de répartition)

Remarque: Cette fonction est souvent utilisée en simulation dans les entreprises pour signaler que la variable aléatoire a des probabilités égales d'avoir une valeur comprise dans un certain intervalle (typiquement dans les rendements de portefeuilles ou encore dans l'estimation des durées des projets). Le meilleur exemple d'application étant à nouveau le logiciel CrystalBall ou @Risk qui s'intègrent dans MS Project.

Voyons un résultat intéressant de la loi Uniforme continue (et qui s'applique à la discrète aussi en fait...).

Souvent j'entends des gestionnaires (qui se jugent de haut niveau) dire que comme une mesure a une probabilité égale d'avoir lieu dans un intervalle fermé donné, alors la somme de deux variables aléatoires indépendantes du même type aussi!

Or nous allons démontrer ici que ce n'est pas le cas (si quelqu'un a une démonstration plus élégante je

suis preneur)!

Démonstration:

Considérons deux variables aléatoires indépendantes X et Y qui suivent une loi uniforme dans un intervalle fermé $[0, a]$. Nous cherchons donc la densité de leur somme qui sera notée:

$$Z = X + Y \quad (7.430)$$

Nous avons alors:

$$f_X(x) = f_Y(y) = \begin{cases} 1 & \text{si } 0 \leq x, y \leq a \\ 0 & \text{sinon} \end{cases} \quad (7.431)$$

avec la variable:

$$0 \leq z \leq 2a \quad (7.432)$$

Pour calculer la loi de la somme, rappelons que nous savons qu'en termes discrets cela équivaut à faire le produit conjoint des probabilités (cf. chapitre de Probabilités) d'apparition des deux variables continues (se rappeler le même genre de calcul sous forme discrète!)

C'est-à-dire:

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(z-y)f_Y(y)dy \quad (7.433)$$

Comme $f_Y(y) = 1$ si $0 \leq y \leq a$ et 0 sinon alors le produit de convolution précédent se réduit à:

$$f_Z(z) = \int_0^a f_X(z-y)dy \quad (7.434)$$

L'intégrant vaut par définition 0 sauf lorsque par construction $0 \leq z-y \leq a$ où il vaut alors 1.

Intéressons-nous alors aux bornes de l'intégrale dans ce dernier cas qui est bien évidemment le seul qui est intéressant....

Faisons d'abord un changement de variables en posant:

$$u = z - y \quad (7.435)$$

d'où:

$$du = -dy \quad (7.436)$$

L'intégrale s'écrit alors dans cet intervalle après ce changement de variable:

$$f_Z(z) = \int_0^a f_X(z-y)dy = - \int_z^{z-a} f_X(u)du = \int_{z-a}^z du \quad (7.437)$$

En se rappelant comme vu au début que $0 \leq z \leq 2a$, alors nous avons immédiatement si $z < 0$ et $z > 2a$ que l'intégrale est nulle.

Nous allons considérer deux cas pour cet intervalle car la convolution de ces deux fonctions rectangulaires peut se distinguer selon la situation où dans un premier temps elles se croisent (s'emboîtent), c'est-à-dire où $0 \leq z \leq a$, et ensuite s'éloignent l'une de l'autre, c'est-à-dire $a < z \leq 2a$.

- Dans le premier cas (emboîtement) où $0 \leq z \leq a$:

$$f_Z(u) = \int_{z-a}^z du = \int_0^z du = u \Big|_0^z = z \quad (7.438)$$

où nous avons changé la borne inférieure à 0 car de toute façon $f_X(u)$ est nulle pour toute valeur négative (et lorsque $0 \leq z \leq a$, $z-a$ est justement négatif ou nul!).

- Dans le deuxième cas (déboîtement) où $a < z \leq 2a$:

$$f_Z(u) = \int_{z-a}^z du = \int_{z-a}^a du = u \Big|_{z-a}^a = a - (z-a) = 2a - z \quad (7.439)$$

où nous avons changé la borne supérieure à a car de toute façon $f_X(u)$ est nulle pour toute valeur supérieure (et lorsque $a < z \leq 2a$, z est justement plus grand que a).

Donc au final, nous avons:

$$f_Z(z) = \begin{cases} z & \text{si } 0 \leq z \leq a \\ 2a - z & \text{si } a < z \leq 2a \\ 0 & \text{sinon} \end{cases} \quad (7.440)$$

□C.Q.F.D.

Il s'agit d'un cas particulier, volontairement simplifié, de la loi triangulaire que nous allons voir de suite.

Ce résultat (qui peut sembler contre intuitif) se vérifie en quelques secondes avec un tableur comme Microsoft Excel 11.8346 en utilisant la fonction ALEA.ENTRE.BORNES() et la fonction FREQUENCE() dans la version française.

5.12. FONCTION TRIANGULAIRE

Soient $a < c < b$. Nous définissons la "**fonction triangulaire**" (ou "**loi triangulaire**") par construction basée sur les deux fonctions de distribution suivantes:

$$\begin{aligned} P_{a,c}(x) &= \frac{2(x-a)}{(b-a)(c-a)} \cdot 1_{[a,c]} \\ P_{c,b}(x) &= \frac{2(b-x)}{(b-a)(b-c)} \cdot 1_{]c,b]} \end{aligned} \quad (7.441)$$

où a est souvent assimilée à la valeur optimiste, c la valeur attendue (le mode) et b la valeur pessimiste.

C'est effectivement la seule manière de l'écrire si le lecteur garde à l'esprit que le triangle de base $c-a$ doit avoir une hauteur h valant $2/(c-a)$ telle que sa surface totale soit égale à l'unité (nous allons de suite le montrer).

Voici un exemple de tracé de la fonction de distribution et répartition pour la fonction triangulaire de paramètres $(a,c,b)=(0,3,5)$:

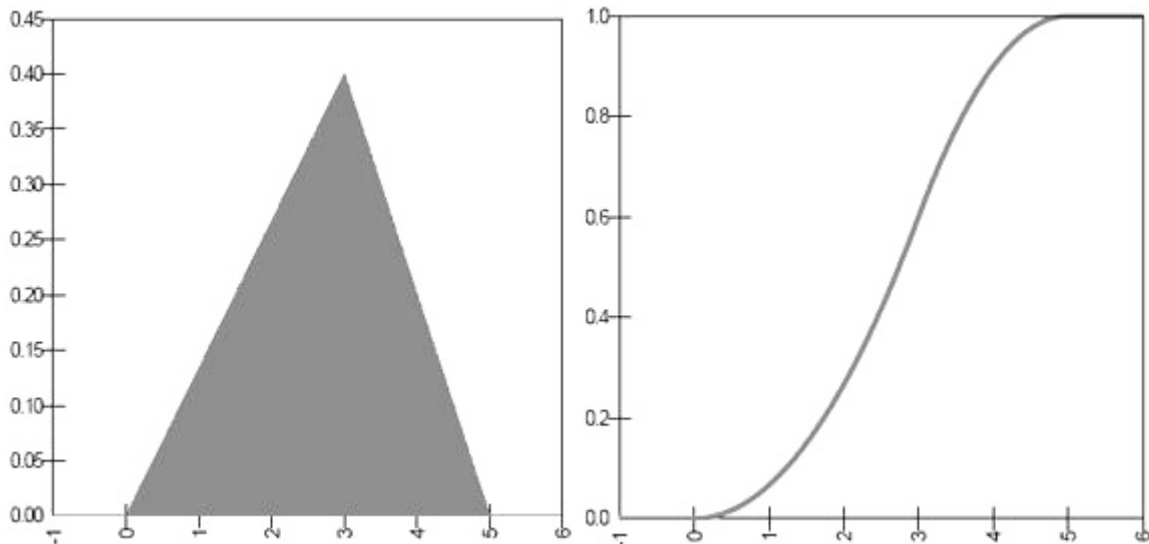


Figure: 7.22 - Loi triangulaire (fonction de distribution et de répartition)

La pente de la première droite (croissante de gauche) est donc bien évidemment:

$$\frac{2}{(b-a)(c-a)} \quad (7.442)$$

et la pente de la deuxième droite (décroissante à droite):

$$\frac{-2}{(b-a)(b-c)} \quad (7.443)$$

Cette fonction est une fonction de distribution si elle vérifie:

$$\rho = \int_{-\infty}^{+\infty} (P_{a,c}(x) + P_{c,b}(x)) dx = 1 \quad (7.444)$$

Il s'agit dans ce cas de l'aire du triangle qui rappelons-le est simplement la base multipliée par la hauteur le tout divisé par 2 (cf. chapitre sur les Formes Géométriques):

$$\rho = \frac{1}{2} \left((b-a) \cdot \frac{2}{(b-a)} \right) = 1 \quad (7.445)$$

Remarque: Cette fonction est beaucoup utilisée en gestion de projet dans le cadre de l'estimation des durées des tâches ou encore en simulations industrielles. La valeur a correspondant à la valeur optimiste, la valeur c à la valeur attendue (mode) et la valeur b à la valeur pessimiste. Le meilleur exemple d'application étant à nouveau le logiciel CrystalBall ou @Risk qui s'intègrent dans MS Project.

La fonction triangulaire a par ailleurs une espérance (moyenne):

$$\begin{aligned}
 \mu &= \int_{-\infty}^{+\infty} xf(x)dx = \int_a^c x \frac{2(x-a)}{(b-a)(c-a)} dx + \int_c^b x \frac{2(b-x)}{(b-a)(b-c)} dx \\
 &= \frac{2}{(b-a)(c-a)} \cdot \left(\frac{1}{3}x^3 - \frac{1}{2}ax^2 \right) \Big|_a^c + \frac{2}{(b-a)(b-c)} \cdot \left(\frac{1}{2}bx^2 - \frac{1}{3}x^3 \right) \Big|_c^b \\
 &= \frac{2}{(b-a)(c-a)} \cdot \left(\left(\frac{1}{3}c^3 - \frac{1}{2}ac^2 \right) - \left(\frac{1}{3}a^3 - \frac{1}{2}a^3 \right) \right) \\
 &\quad + \frac{2}{(b-a)(b-c)} \cdot \left(\left(\frac{1}{2}b^3 - \frac{1}{3}b^3 \right) - \left(\frac{1}{2}bc^2 - \frac{1}{3}c^3 \right) \right) \\
 &= 2 \frac{(b-c) \left(\left(\frac{1}{3}c^3 - \frac{1}{2}ac^2 \right) + \frac{a^3}{6} \right) + (c-a) \left(\frac{1}{6}b^3 - \left(\frac{1}{2}bc^2 - \frac{1}{3}c^3 \right) \right)}{(b-a)(c-a)(b-c)} \\
 &= 2 \frac{\frac{1}{3}bc^3 - \cancel{\frac{1}{2}abc^2} + \frac{ba^3}{6} - \cancel{\frac{1}{3}c^4} + \frac{1}{2}ac^3 - \frac{a^3c}{6}}{(b-a)(c-a)(b-c)} \\
 &\quad + 2 \frac{\frac{1}{6}cb^3 - \frac{1}{2}bc^3 + \cancel{\frac{1}{3}c^4} - \frac{ab^3}{6} + \cancel{\frac{1}{2}abc^2} - \frac{1}{3}ac^3}{(b-a)(c-a)(b-c)} \\
 &= \frac{\left(\frac{2}{3}bc^3 + ac^3 - \frac{1}{3}a^3c + \frac{1}{3}cb^3 + \frac{ba^3}{3} - \frac{ab^3}{3} - bc^3 - \frac{2}{3}ac^3 \right)}{(b-a)(c-a)(b-c)} \\
 &= \frac{\left(-\frac{1}{3}bc^3 - \frac{1}{3}a^3c + \frac{1}{3}ba^3 - \frac{1}{3}ab^3 + \frac{1}{3}cb^3 + \frac{1}{3}ac^3 \right)}{(b-a)(c-a)(b-c)} \\
 &= \frac{1}{3} \frac{(-bc^3 - a^3c + ba^3 - ab^3 + cb^3 + ac^3)}{(b-a)(c-a)(b-c)} \\
 &= \frac{1}{3} \frac{(a+b+c)(b-a)(c-a)(b-c)}{(b-a)(c-a)(b-c)} \\
 &= \frac{a+b+c}{3}
 \end{aligned}
 \tag{7.446}$$

et pour variance:

$$\begin{aligned}
\sigma^2 &= \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx = \int_a^c (x - \mu)^2 \frac{2(x-a)}{(b-a)(c-a)} dx + \int_c^b (x - \mu)^2 \frac{2(b-x)}{(b-a)(b-c)} dx \\
&= -\frac{1}{6} \frac{c(-3c^3 + 8c^2\mu + 4c^2a - 6c\mu^2 - 12c\mu a + 12\mu^2a)}{(b-a)(c-a)} + \frac{1}{6} \frac{a^2(a^2 - 4\mu a + 6\mu^2)}{(b-a)(c-a)} \\
&\quad - \frac{1}{6} \frac{(-3c^3 + 8c^2\mu + 4c^2b - 6c\mu^2 - 12c\mu b + 12\mu^2b)}{(b-a)(c-a)} + \frac{1}{6} \frac{b^2(b^2 - 4\mu b + 6\mu^2)}{(b-a)(c-a)} \\
&= \frac{1}{6}b^2 + \frac{1}{6}ba - \frac{2}{3}\mu b + \frac{1}{6}bc + \frac{1}{6}a^2 - \frac{2}{3}\mu a + \mu^2 + \frac{1}{6}ca - \frac{2}{3}c\mu + \frac{1}{6}c^2
\end{aligned} \tag{7.447}$$

on remplace μ par l'expression obtenue précédemment et on simplifie (c'est de l'algèbre élémentaire pénible...):

$$\sigma^2 = \frac{a^2 + b^2 + c^2 - ab - ac - bc}{18} \tag{7.448}$$

Nous pouvons montrer que la somme de deux variables aléatoires indépendantes chacune de loi uniforme sur $[a, b]$ (donc indépendantes et identiquement distribuées) suit une loi triangulaire sur $[2a, 2b]$ mais si elles n'ont pas les mêmes bornes, alors leur somme donne un truc qui n'a pas de nom à ma connaissance...

5.13. FONCTION DE PARETO

La "**fonction de Pareto**" (ou "**loi de Pareto**"), appelée aussi "**loi de puissance**" ou encore "**loi scalante**" est la formalisation du principe des 80-20. Cet outil d'aide à la décision détermine les facteurs (environ 20%) cruciaux qui influencent la plus grande partie (80%) de l'objectif.

Remarque: Cette loi est un outil fondamental et basique en gestion de la qualité (cf. chapitre de Génie Industriel et Techniques de Gestion). Elle est aussi utilisée en réassurance. La théorie des files d'attente s'est intéressée à cette distribution, lorsque des recherches des années 1990 ont montré que cette loi régissait aussi nombre de grandeurs observées dans le trafic Internet (et plus généralement sur tous les réseaux de données à grande vitesse).

Une variable aléatoire est dite par définition suivre une loi de Pareto si sa fonction de répartition est donnée par:

$$P(X \leq x) = 1 - \left(\frac{x_m}{x} \right)^k \tag{7.449}$$

avec x qui doit être supérieur ou égal à x_m .

La fonction de densité (fonction de distribution) de Pareto est alors:

$$\begin{aligned}
 f(x) &= \frac{d}{dx} \left(1 - \left(\frac{x_m}{x} \right)^k \right) = - \frac{d}{dx} \left(\frac{x_m}{x} \right)^k = -x_m^k \frac{d}{dx} \frac{1}{x^k} = -x_m^k \frac{d}{dx} x^{-k} = -x_m^k (-k x^{-k-1}) \\
 &= k x_m^k x^{-k-1} = k \frac{x_m^k}{x^{k+1}}
 \end{aligned} \tag{7.450}$$

avec $k \in \mathbb{R}_+$ et $x \geq x_m \geq 0$ (donc $x > 0$).

La distribution de Pareto est donc définie par deux paramètres, x_m et k (nommé "**index de Pareto**").

Cette une loi dite aussi à "**invariance d'échelle**" ou "**loi fractale**", terme définissant la propriété suivante:

$$f(c^{te} \cdot x) = k x_m^k (c^{te} \cdot x)^{-k-1} = (c^{te})^{-k-1} k x_m^k x^{-k-1} = (c^{te})^{-k-1} f(x) \propto f(x) \tag{7.451}$$

La loi de Pareto est par ailleurs bien une fonction de distribution puisque étant connue sa fonction de répartition:

$$\int_0^{+\infty} f(x) dx = \left(1 - \left(\frac{x_m}{x} \right)^k \right) \Big|_{x_m}^{+\infty} = \left(1 - \left(\frac{x_m}{+\infty} \right)^k \right) - \left(1 - \left(\frac{x_m}{x_m} \right)^k \right) = (1-0) - (1-1^k) = 1 \tag{7.452}$$

L'espérance (moyenne) est donnée par:

$$\begin{aligned}
 \mu = E(X) &= \int_{x_m}^{+\infty} x f(x) dx = \int_{x_m}^{+\infty} x k \frac{x_m^k}{x^{k+1}} dx = k x_m^k \int_{x_m}^{+\infty} \frac{1}{x^k} dx \\
 &= - \frac{k x_m^k}{k-1} \cdot \frac{1}{x^{k-1}} \Big|_{x_m}^{+\infty} = \frac{k \cdot x_m}{k-1}
 \end{aligned} \tag{7.453}$$

si $k > 1$. Si $k \leq 1$, l'espérance n'existe pas.

Pour calculer la variance, en utilisant la relation:

$$V(X) = E(X^2) - E(X)^2 \tag{7.454}$$

Nous avons:

$$E(X^2) = \int_{x_m}^{+\infty} x^2 f(x) dx = k x_m^k \int_{x_m}^{+\infty} \frac{1}{x^{k-1}} dx = - \frac{k x_m^k}{k-2} \cdot \frac{1}{x^{k-2}} \Big|_{x_m}^{+\infty} = \frac{k \cdot x_m^2}{k-2} \tag{7.455}$$

si $k > 2$. Si $k \leq 2$, $E(X^2)$ n'existe pas.

Donc si $k > 2$:

$$\sigma^2 = V(X) = \frac{k \cdot x_m^2}{k-2} - \left(\frac{k \cdot x_m}{k-1} \right)^2 = \frac{k x_m^2}{(k-1)^2 (k-2)} \quad (7.456)$$

Si $k \leq 2$, la variance n'existe pas.

Voici un exemple de tracé de la fonction de distribution et répartition pour la fonction de Pareto de paramètre $(x, x_m, k) = (x, 1, 2)$:

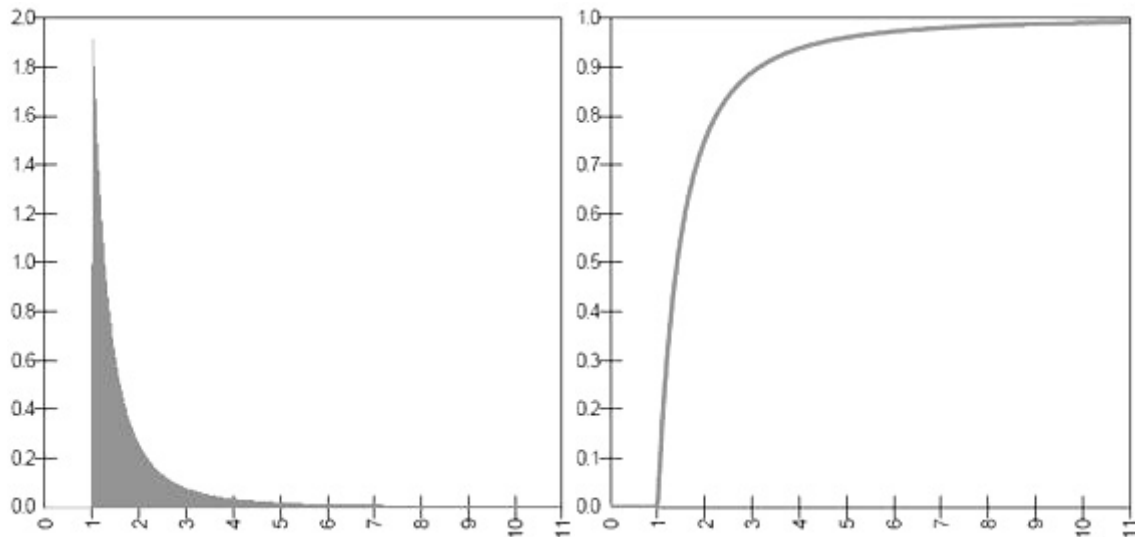


Figure: 7.23 - Loi de Pareto (fonction de distribution et de répartition)

Remarque: Il faut noter que lorsque $k \rightarrow +\infty$ la distribution s'approche de $\delta(x - x_m)$ où δ est la fonction Delta de Dirac.

Il existe une autre manière importante de déduire la famille des lois de Pareto qui permet de comprendre bien des choses concernant d'autres lois et qui est souvent présentée de la façon suivante:

Notons x_0 le seuil au-delà duquel nous calculons l'espérance de la quantité examinée, et $E(Y)$ l'espérance au-delà de ce seuil x_0 tel qu'il soit proportionnel (linéairement dépendant) au seuil choisi:

$$E(Y) = ax_0 + b \quad (7.457)$$

Cette relation fonctionnelle exprime l'idée que la moyenne conditionnelle au-delà du seuil x_0 est un multiple de ce seuil à une constante près, c'est-à-dire une fonction linéaire de ce seuil.

Ainsi, en gestion de projets par exemple, nous pourrions dire qu'une fois un certain seuil de durée dépassé, la durée espérée est un multiple de ce même seuil à une constante près.

Si une relation linéaire de ce type existe et est bien vérifiée, nous parlons alors de distribution de probabilité sous la forme d'une loi de Pareto généralisée.

Considérons l'espérance mathématique de la fonction conditionnelle bayésienne donnée par (cf. chapitre de Probabilités):

$$E(Y) = \frac{1}{P(X \geq x_0)} \int_{x_0}^{+\infty} y f(y) dy \quad (7.458)$$

Si nous notons $F(y)$, la fonction de répartition de $f(y)$, nous avons alors par définition:

$$dF(y) = f(y) dy \quad (7.459)$$

Dès lors:

$$E(Y) = \frac{1}{P(X \geq x_0)} \int_{x_0}^{+\infty} y dF(y) \quad (7.460)$$

et si nous définissons:

$$\bar{F}_X(x) = P(X \geq x_0) = 1 - F_X(x) \quad (7.461)$$

que nous pouvons assimiler à la "queue de la distribution".

Il vient:

$$E(Y) = \frac{1}{\bar{F}(x_0)} \int_{x_0}^{+\infty} y dF(y) \quad (7.462)$$

et donc nous cherchons le cas très particulier où:

$$\frac{1}{\bar{F}(x)} \int_x^{+\infty} y dF(y) = ax + b \quad (7.463)$$

c'est-à-dire:

$$\int_x^{+\infty} y dF(y) = (ax + b) \bar{F}(x) \quad (7.464)$$

En dérivant par rapport à x , nous trouvons:

$$\frac{d}{dx} \left(\int_x^{+\infty} y dF(y) \right) = \frac{d}{dx} ((ax + b) \bar{F}(x)) \quad (7.465)$$

La dérivée de l'intégrale définie ci-dessus sera la dérivée d'une constante (valorisation de l'intégrale en $+\infty$) moins la dérivée de l'intégrale de l'expression analytique en x . Nous avons donc:

$$\frac{d}{dx} \left(\int_x^{+\infty} y dF(y) \right) = \frac{d}{dx} \left(\int_x^{+\infty} y f(y) dy \right) = -x f(x) = -x \frac{dF(x)}{dx} = a \bar{F}(x) + (ax + b) \frac{d\bar{F}(x)}{dx} \quad (7.466)$$

Soit:

$$-x \frac{dF(x)}{dx} = a\bar{F}(x) + (ax+b) \frac{d\bar{F}(x)}{dx} \quad (7.467)$$

et comme:

$$dF = d(1 - \bar{F}) = -d\bar{F} \quad (7.468)$$

Il vient:

$$x \frac{d\bar{F}(x)}{dx} = a\bar{F}(x) + (ax+b) \frac{d\bar{F}(x)}{dx} \quad (7.469)$$

Après simplification et réarrangement nous obtenons:

$$a\bar{F}(x)dx = -(x(a-1)+b)d\bar{F}(x) \quad (7.470)$$

qui est donc une équation différentielle en $\bar{F}(x)$. Sa résolution fournit toutes les formes de lois de Pareto recherchées, selon les valeurs que prennent les paramètres a et b .

Pour résoudre cette équation différentielle, considérons le cas particulier où $a > 1, b = 0$. Nous avons alors:

$$a\bar{F}(x)dx = -x(a-1)d\bar{F}(x) \quad (7.471)$$

En posant:

$$k = \frac{a}{a-1} \quad (7.472)$$

Nous avons alors:

$$-\frac{1}{x}dx = \frac{1}{k} \frac{1}{\bar{F}(x)} d\bar{F}(x) \quad (7.473)$$

et donc:

$$-\ln(x) = \frac{1}{k} \ln(\bar{F}(x)) + c^{te} \quad (7.474)$$

Il vient:

$$e^{\ln\left(\frac{1}{x}\right)} = e^{\frac{1}{k} \ln(\bar{F}(x)) + c^{te}} = e^{\frac{1}{k} \ln(\bar{F}(x))} e^{c^{te}} \quad (7.475)$$

et donc:

$$\left(\frac{1}{x}\right) = \bar{F}(x)^{\frac{1}{k}} \cdot e^{c^{te}} \quad (7.476)$$

Nous avons:

$$\bar{F}(x) = \left(\frac{c^{te}}{x} \right)^k = \left(\frac{x_m}{x} \right)^k \quad (7.477)$$

Il vient alors pour la fonction de répartition:

$$F(x) = 1 - \bar{F}(x) = 1 - \left(\frac{x_m}{x} \right)^k \quad (7.478)$$

Si nous cherchons la fonction de distribution, nous dérivons par x pour obtenir:

$$f(x) = \frac{dF(x)}{dx} = k \frac{x_m^k}{x^{k+1}} \quad (7.479)$$

Il s'agit de la loi de Pareto que nous avons utilisée depuis le début et nommée "**distribution de Pareto de type I**" (nous ne montrerons pas sur ce site celles de type II).

Une chose intéressante à observer au passage est le cas de la résolution de l'équation différentielle:

$$a\bar{F}(x)dx = -(x(a-1)+b)d\bar{F}(x) \quad (7.480)$$

lorsque $a = 1, b > 0$. L'équation différentielle se réduit alors à:

$$\bar{F}(x)dx = -bd\bar{F}(x) \quad (7.481)$$

Soit:

$$-\frac{1}{b}dx = \frac{1}{\bar{F}(x)}d\bar{F}(x) \quad (7.482)$$

Après intégration:

$$-\frac{1}{b}x = \ln(\bar{F}(x)) \quad (7.483)$$

et donc:

$$\bar{F}(x) = e^{-\frac{1}{b}x} \quad (7.484)$$

Si nous faisons un petit changement de notation:

$$\bar{F}(x) = e^{-\lambda x} \quad (7.485)$$

et que nous écrivons la fonction de répartition:

$$F(x) = 1 - \bar{F}(x) = 1 - e^{-\lambda x} \quad (7.486)$$

Et en dérivant nous obtenons la fonction de distribution de la loi exponentielle:

$$F(x) = \lambda e^{-\lambda x} \quad (7.487)$$

Donc la loi exponentielle a une espérance conditionnelle seuil qui est égale à:

$$E(Y) = ax_0 + b =_{a=1} x_0 + b =_{b=\frac{1}{\lambda}} x_0 + \frac{1}{\lambda} = x_0 + \frac{1}{\lambda} = x_0 + \sigma \quad (7.488)$$

Donc l'espérance conditionnelle seuil est égale à elle-même augmenté de l'écart-type de la distribution.

5.14. FONCTION EXPONENTIELLE

Nous définissons la "**fonction exponentielle**" (ou "**loi exponentielle**") par la relation de fonction de distribution suivante:

$$P(x) = \lambda \cdot e^{-\lambda x} \cdot 1_{[0,+\infty[} \quad (7.489)$$

avec $\lambda > 0$ qui comme nous allons de suite le montrer n'est autre que l'inverse de la moyenne et où x est une variable aléatoire sans mémoire.

Au fait la loi exponentielle découle naturellement de développements très simples (voir celui dans le chapitre de Physique Nucléaire par exemple) sous des hypothèses qui imposent une constance dans le vieillissement d'un phénomène. Dans le chapitre des Techniques de Gestion, nous avons aussi démontré en détails dans la partie concernant la théorie des files d'attente, que cette loi était sans mémoire. C'est-à-dire que que la probabilité cumulée qu'un phénomène se produise entre les temps t et $t+s$ s'il ne s'est pas produit avant est la même que la probabilité qu'il se produise entre les temps 0 et s .

Remarques:

R1. Cette fonction se retrouve fréquemment en physique nucléaire (voir chapitre du même nom) ou encore en physique quantique (voir chapitre du même nom) ainsi qu'en fiabilité (cf. chapitre de Génie Industriel) ou dans la théorie des files d'attente (cf. chapitre de Techniques de Gestion).

R2. Nous pouvons obtenir cette loi dans la version française de Microsoft Excel 11.8346 avec la fonction LOI.EXPONENTIELLE().

Il s'agit par ailleurs bien d'une fonction de distribution car elle vérifie:

$$\int_{-\infty}^{+\infty} P_{\lambda}(x) dx = \int_0^{+\infty} \lambda e^{-\lambda x} dx = \lambda \int_0^{+\infty} e^{-\lambda x} dx = \lambda \left(-\frac{1}{\lambda} \right) e^{-\lambda x} \Big|_0^{+\infty} = - \left(e^{-\infty} - e^{-0} \right) = -(0 - 1) = 1 \quad (7.490)$$

La fonction exponentielle a pour espérance (moyenne) en utilisant l'intégration par parties:

$$\begin{aligned}
 \mu &= \int_{-\infty}^{+\infty} x \cdot P_{\lambda}(x) dx = \int_0^{+\infty} x \cdot \lambda \cdot e^{-\lambda x} dx = -e^{-\lambda x} \cdot x \Big|_0^{+\infty} - \int_0^{+\infty} -e^{-\lambda x} dx \\
 &= \int_0^{+\infty} e^{-\lambda x} dx = -\frac{1}{\lambda} e^{-\lambda x} \Big|_0^{+\infty} = \frac{1}{\lambda}
 \end{aligned}
 \tag{7.491}$$

et pour variance (nous utilisons à nouveau $V(X) = E(X^2) - E(X)^2$) et il ne nous reste plus qu'à calculer:

$$E(X^2) = \int_0^{+\infty} \lambda x^2 e^{-\lambda x} dx \tag{7.492}$$

Un changement de variable $y = \lambda x \Rightarrow dy = \lambda dx$ conduit à:

$$E(X^2) = \frac{1}{\lambda^2} \int_0^{+\infty} y^2 e^{-y} dy \tag{7.493}$$

Une double intégration par parties donne:

$$\begin{aligned}
 \int_a^b f(t) g'(t) dt &= f(t) g(t) \Big|_a^b - \int_a^b f'(t) g(t) dt \\
 \int_0^{+\infty} y^2 e^{-y} dy &= -y^2 e^{-y} \Big|_0^{+\infty} + 2 \int_0^{+\infty} y e^{-y} dy = 2 \left[-y e^{-y} \Big|_0^{+\infty} \right] + 2 \int_0^{+\infty} e^{-y} dy = 2
 \end{aligned}
 \tag{7.494}$$

D'où $E(X^2) = 2 / \lambda^2$ il vient dès lors:

$$V(X) = E(X^2) - E(X)^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda} \right)^2 = \frac{1}{\lambda^2} \tag{7.495}$$

Donc l'écart-type (racine carrée de la variance pour rappel) et la moyenne ont exactement la même expression!

Voici un exemple de tracé de la fonction de distribution et répartition pour la fonction exponentielle de paramètre $\lambda = 1$:

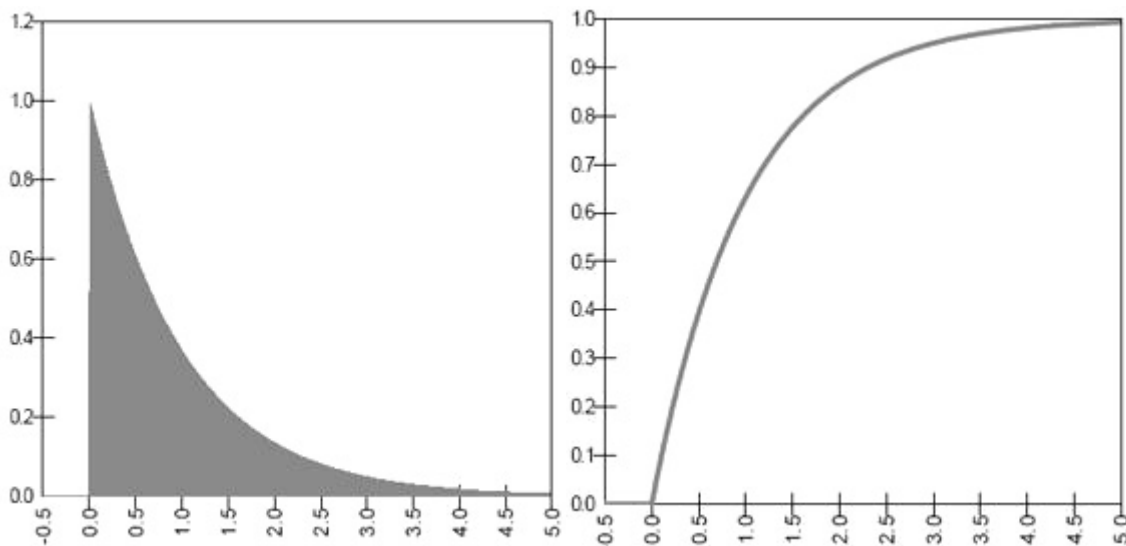


Figure: 7.24 - Loi exponentielle (fonction de distribution et de répartition)

Déterminons maintenant la fonction de répartition de la loi exponentielle:

$$\begin{aligned}
 P(x \leq X) &= \int \lambda e^{-\lambda t} dt = \lambda \int e^{-\lambda t} dt = \lambda \cdot \left(-\frac{1}{\lambda} e^{-\lambda t} \right) \Big|_0^x = -\left(e^{-\lambda x} \right) \Big|_0^x \\
 &= -\left(e^{-\lambda x} - 1 \right) = 1 - e^{-\lambda x}
 \end{aligned}
 \tag{7.496}$$

Remarque: Nous verrons plus loin que la fonction de distribution exponentielle n'est qu'un cas particulier d'une fonction plus générale qui est la fonction du Khi-deux, cette dernière aussi n'étant qu'un cas particulier d'une fonction encore plus générale qui est la fonction Gamma. Il s'agit d'une propriété très importante utilisée dans le "test de Poisson" pour les événements rares (voir plus loin aussi).

5.15. FONCTION DE CAUCHY

Soient X, Y deux variables aléatoires indépendantes suivant des lois Normales centrées réduites (variance unité et espérance nulle). La fonction de densité est donc donnée par:

$$f_X(x) = f_Y(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{7.497}$$

La variable aléatoire:

$$T = \frac{X}{|Y|} \tag{7.498}$$

(la valeur absolue intervient dans l'intégrale lors du changement variable) suit une allure caractéristique appelée "**fonction de Cauchy**" (ou "**loi de Cauchy**") ou encore "**loi de Lorentz**".

Déterminons sa fonction de densité f . Pour cela, rappelons que f est déterminée par la relation (générale):

$$\forall t \in \mathbb{R}, P(T < t) = \int_{-\infty}^t f(x) dx \quad (7.499)$$

Donc (application du calcul intégral élémentaire):

$$f(t) = \frac{d}{dt} P(T < t) \quad (7.500)$$

dans le cas où f est continue.

Etant donné que X et Y sont indépendants, la fonction de densité du vecteur aléatoire est donnée par un des axiomes des probabilités (cf. chapitre de Probabilités):

$$(x, y) \mapsto f_X(x) \cdot f_Y(y) \quad (7.501)$$

Donc:

$$P(T < t) = P\left(\frac{X}{|Y|} < t\right) = P(X < t|Y|) = \int_D f_X(x) f_Y(y) dx dy \quad (7.502)$$

où donc $D = \{(x, y) \mid x < t \cdot |y|\}$.

Cette dernière intégrale devient:

$$\int_D f_X(x) f_Y(y) dx dy = \int_{-\infty}^{+\infty} \int_{-\infty}^{t|y|} f_X(x) \cdot f_Y(y) dx dy \quad (7.503)$$

Faisons le changement de variable $x = u \cdot |y|$ dans l'intégrale intérieure. Nous obtenons:

$$P(T < t) = \int_{-\infty}^{+\infty} \int_{-\infty}^t f_X(u \cdot |y|) \cdot f_Y(y) \cdot |y| \cdot du dy = \int_{-\infty}^t \int_{-\infty}^{+\infty} f_X(u \cdot |y|) \cdot f_Y(y) \cdot |y| \cdot dy du \quad (7.504)$$

Donc:

$$f(t) = \frac{d}{dt} P(T < t) = \int_{-\infty}^{+\infty} f_X(t \cdot |y|) \cdot f_Y(y) \cdot |y| \cdot dy = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{\frac{-y^2(t^2+1)}{2}} \cdot |y| \cdot dy \quad (7.505)$$

C'est maintenant que la valeur absolue va nous être utile pour écrire:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{\frac{-y^2(t^2+1)}{2}} |y| dy = -\frac{1}{2\pi} \int_{-\infty}^0 e^{\frac{-y^2(t^2+1)}{2}} y dy + \frac{1}{2\pi} \int_0^{+\infty} e^{\frac{-y^2(t^2+1)}{2}} y dy \quad (7.506)$$

Pour la première intégrale nous avons:

$$\int_{-\infty}^0 e^{\frac{-y^2(t^2+1)}{2}} y dy = - \frac{e^{\frac{-y^2(t^2+1)}{2}}}{t^2+1} \Bigg|_{-\infty}^0 = -\frac{1}{t^2+1} + \frac{1}{t^2+1} = 0 \quad (7.507)$$

Il ne reste donc plus que la seconde intégrale et en faisant le changement de variable $v = y^2$, nous obtenons:

$$f(t) = \frac{1}{2\pi} \int_0^{+\infty} e^{\frac{-v(t^2+1)}{2}} \cdot dv = - \frac{1}{\pi(t^2+1)} e^{\frac{-v(t^2+1)}{2}} \Bigg|_0^{+\infty} = \frac{1}{\pi(t^2+1)} \quad (7.508)$$

Ce que nous noterons par la suite (afin de respecter les notations adoptées jusqu'à présent):

$$P(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2} \quad (7.509)$$

et qui n'est d'autre que la fonction de Cauchy.

Il s'agit par ailleurs bien d'une fonction de distribution car elle vérifie (cf. chapitre de Calcul Différentiel et Intégral):

$$\int_{-\infty}^{+\infty} P(x) dx = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{1}{1+x^2} dx = \frac{1}{\pi} \cdot (\arctan(+\infty) - \arctan(-\infty)) = 1. \quad (7.510)$$

Voici un exemple de tracé de la fonction de distribution de Cauchy:

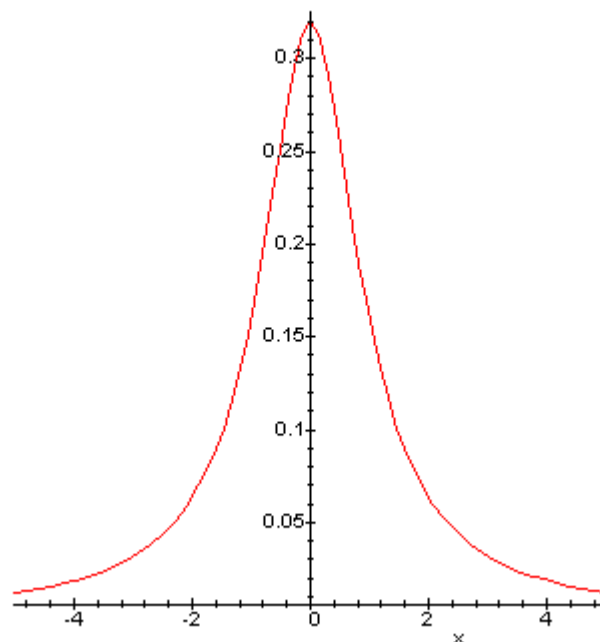


Figure: 7.25 - Loi de Cauchy (fonction de distribution)

La fonction de Cauchy a pour espérance (moyenne):

$$\begin{aligned}
\mu &= \int_{-\infty}^{+\infty} x \cdot P(x) dx = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{x}{1+x^2} dx = \frac{1}{\pi} \left(\int_{-\infty}^0 \frac{x}{1+x^2} dx + \int_0^{+\infty} \frac{x}{1+x^2} dx \right) \\
&= \frac{1}{2\pi} \left(\ln(1+x^2) \Big|_{-\infty}^0 + \ln(1+x^2) \Big|_0^{+\infty} \right) = \frac{1}{2\pi} (\ln(1) - \ln(\infty) + \ln(\infty) - \ln(1)) \quad (7.511) \\
&= \frac{1}{2\pi} (-\ln(\infty) + \ln(\infty)) = 0
\end{aligned}$$

Attention!!! Les calculs précédents ne donnent pas zéro au fait car la soustraction d'infinis est non pas nulle mais indéterminée! La loi de Cauchy n'admet donc pas d'espérance rigoureusement parlant!

Ainsi, même si nous pouvons bricoler une variance:

$$\begin{aligned}
\sigma^2 &= \int_{-\infty}^{+\infty} (x - \mu)^2 \cdot f(x) dx = \int_{-\infty}^{+\infty} x^2 \cdot P(x) dx = \frac{1}{\pi} \cdot \int_{-\infty}^{+\infty} \frac{x^2}{1+x^2} dx = \frac{1}{\pi} \cdot \int_{-\infty}^{+\infty} \left(1 - \frac{1}{1+x^2}\right) dx = \\
&\frac{2}{\pi} \cdot \lim_{t \rightarrow +\infty} \int_0^t \left(1 - \frac{1}{1+x^2}\right) dx = \frac{2}{\pi} \cdot \lim_{t \rightarrow +\infty} (t - \arctan(t)) = +\infty \quad (7.512)
\end{aligned}$$

celle-ci est absurde et n'existe rigoureusement parlant pas puisque l'espérance n'existe pas...!

5.16. LOI BÊTA

Rappelons d'abord que la fonction Gamma d'Euler est définie par la relation (cf. chapitre de Calcul Différentiel Et Intégral):

$$\Gamma(z) = \int_0^{+\infty} e^{-x} x^{z-1} dx \quad (7.513)$$

Nous avons démontré (cf. chapitre de Calcul Différentiel Et Intégral) qu'une propriété non triviale de cette fonction est que:

$$\Gamma(z+1) = z \cdot \Gamma(z) \quad (7.514)$$

Posons maintenant:

$$\Gamma(a) \cdot \Gamma(b) = \lim_{R \rightarrow +\infty} \iint_{A_R} e^{-x-y} x^{a-1} y^{b-1} dx dy \quad (7.515)$$

où:

$$A_R = \{(x, y) | x \geq 0, y \geq 0, x + y \leq R\} \quad (7.516)$$

En faisant le changement de variables:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} u \cdot v \\ v \end{pmatrix} \quad (7.517)$$

nous obtenons:

$$\Gamma(a) \cdot \Gamma(b) = \lim_{R \rightarrow +\infty} \iint_{A_R} e^{-x-y} x^{a-1} y^{b-1} dx dy = \lim_{R \rightarrow +\infty} \int_0^R e^{-u} \left(\int_0^u (u-v)^{a-1} v^{b-1} dv \right) du \quad (7.518)$$

Pour l'intégrale interne nous utilisons maintenant la substitution $v = u \cdot t$ ($0 \leq t \leq 1$) et nous trouvons alors:

$$\begin{aligned} \Gamma(a) \cdot \Gamma(b) &= \lim_{R \rightarrow +\infty} \int_0^R e^{-u} \left(\int_0^u (u-v)^{a-1} v^{b-1} dv \right) du = \lim_{R \rightarrow +\infty} \int_0^R e^{-u} u^{a+b-1} du \cdot \underbrace{\int_0^1 (1-t)^{a-1} t^{b-1} dt}_{=: B(a,b)} \\ &= B(a,b) \cdot \int_0^{+\infty} e^{-u} u^{a+b-1} du = B(a,b) \cdot \Gamma(a+b) \end{aligned} \quad (7.519)$$

La fonction B qui apparaît dans l'expression ci-dessus est appelée "**fonction bêta**" et nous avons donc:

$$B(a,b) = \frac{\Gamma(a) \cdot \Gamma(b)}{\Gamma(a+b)} \quad (7.520)$$

Maintenant que nous avons défini ce qu'était la fonction bêta, considérons deux paramètres $a > 0, b > 0$ et considérons la relation particulière ci-dessous comme étant la "**fonction de distribution bêta**" ou "**loi bêta**" (il existe plusieurs formulations de la loi bêta dont une très importante qui est étudiée en détails dans le chapitre de Techniques de Gestion):

$$P_{a,b}(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \cdot 1_{]0,1[} \quad (7.521)$$

où:

$$B(a,b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx \quad (7.522)$$

Nous vérifions d'abord que $P_{a,b}(x)$ est bien une fonction de distribution (sans trop aller dans les détails...):

$$\begin{aligned} \int_{-\infty}^{+\infty} P_{a,b}(x) dx &= \int_{-\infty}^{+\infty} \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \cdot 1_{]0,1[} dx = \int_{-\infty}^{+\infty} \frac{x^{a-1}(1-x)^{b-1}}{c^{te}} \cdot 1_{]0,1[} dx \\ &= \frac{1}{c^{te}} \int_{-\infty}^{+\infty} x^{a-1}(1-x)^{b-1} dx = \frac{1}{c^{te}} B(a,b) = \frac{1}{c^{te}} c^{te} = 1 \end{aligned} \quad (7.523)$$

Maintenant, calculons son espérance (moyenne):

$$\begin{aligned} \mu &= \int_{-\infty}^{+\infty} x \cdot P_{a,b}(x) dx = \frac{1}{B(a,b)} \cdot \int_0^1 x^a (1-x)^{b-1} dx = \frac{B(a+1,b)}{B(a,b)} \\ &= \frac{\Gamma(a+1) \cdot \Gamma(b)}{\Gamma(a+b+1)} \cdot \frac{\Gamma(a+b)}{\Gamma(a) \cdot \Gamma(b)} = \frac{a}{(a+b)} \end{aligned} \quad (7.524)$$

en utilisant la relation:

$$\Gamma(z+1) = z \cdot \Gamma(z) \quad (7.525)$$

et sa variance:

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{+\infty} (x - \mu)^2 \cdot f(x) dx = \frac{1}{B(a,b)} \cdot \int_0^1 (x - \mu)^2 \cdot x^{a-1} (1-x)^{b-1} dx \\ &= \frac{1}{B(a,b)} \cdot \left(\int_0^1 x^{a+1} (1-x)^{b-1} dx - 2\mu \int_0^1 x^a (1-x)^{b-1} dx + \mu^2 \int_0^1 x^{a-1} (1-x)^{b-1} dx \right) \\ &= \frac{1}{B(a,b)} \cdot (B(a+2, b) - 2\mu^2 B(a, b) + \mu^2 B(a, b)) = \frac{1}{B(a,b)} \cdot (B(a+2, b) - \mu^2 B(a, b)) \end{aligned} \quad (7.526)$$

En sachant que $\Gamma(z+1) = z \cdot \Gamma(z)$ et que $B(a, b) = \frac{\Gamma(a) \cdot \Gamma(b)}{\Gamma(a+b)}$ nous trouvons:

$$B(a+2, b) = \frac{\mu \cdot (a+1)}{a+b+1} \cdot B(a, b) \quad (7.527)$$

et donc:

$$\sigma^2 = \frac{\mu \cdot (a+1)}{a+b+1} - \mu^2 = \frac{ab}{(a+b)^2(a+b+1)} \quad (7.528)$$

Exemples de tracés de la fonction pour $(a, b) = (0.1, 0.5)$ en rouge, $(a, b) = (0.3, 0.5)$ en vert, $(a, b) = (0.5, 0.5)$ en noir, $(a, b) = (0.8, 0.8)$ en bleu, $(a, b) = (1, 1)$ en magenta, $(a, b) = (1.1, 1.5)$ en cyan, $(a, b) = (1, 2)$ en gris, $(a, b) = (1.5, 2)$ en turquoise, $(a, b) = (2, 2)$ en jaune, $(a, b) = (3, 3)$ en couleur or:

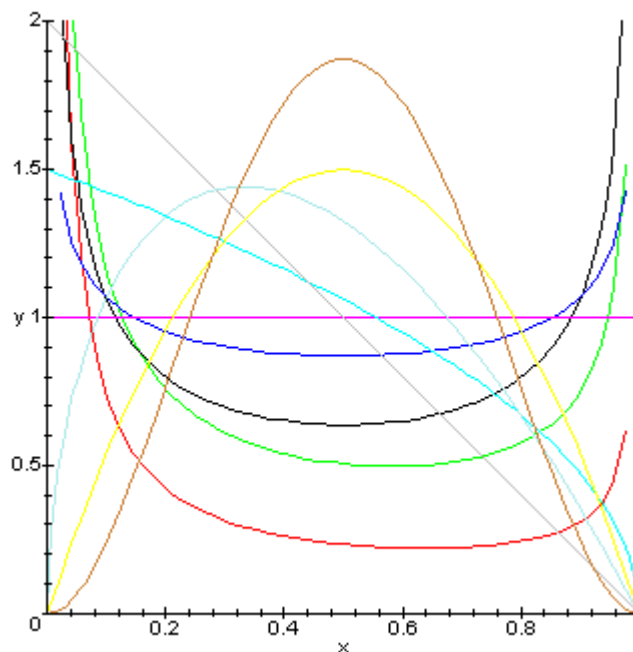


Figure: 7.26 - Loi bêta (fonctions de distribution)

et tracé de la fonction de distribution et répartition de la loi bêta de paramètres $(a, b) = (2, 3)$:

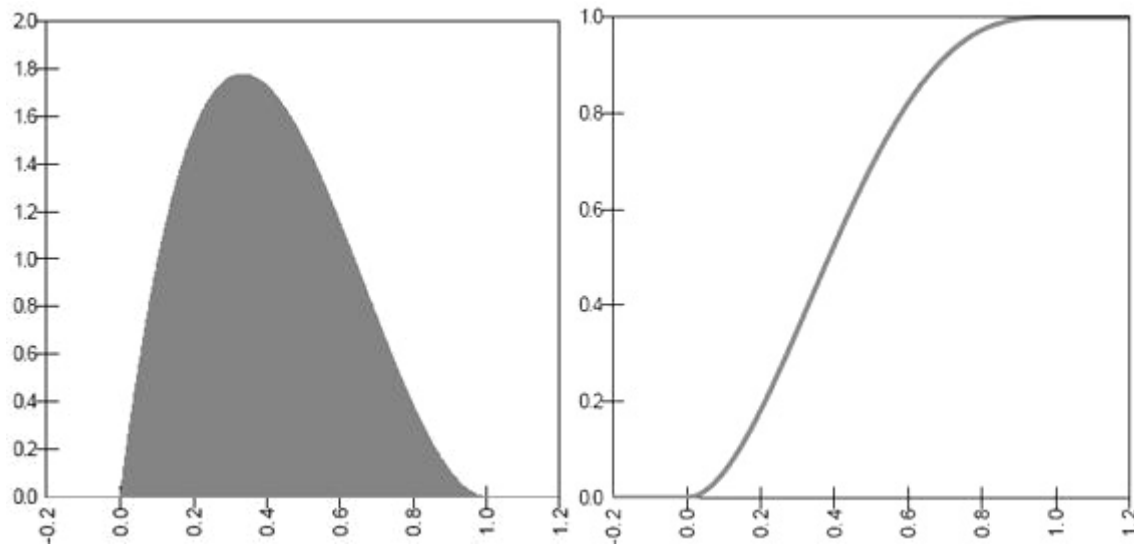


Figure: 7.27 - Loi bêta (fonction de distribution et de répartition)

5.17. FONCTION GAMMA

La fonction Gamma d'Euler étant connue, considérons deux paramètres $\alpha > 0, \lambda > 0$ et définissons la "**fonction Gamma**" (ou "**loi Gamma**") comme étant donnée par la relation:

$$P_{\alpha,\lambda}(x) = \frac{x^{\alpha-1} e^{-\lambda x}}{\int_0^{+\infty} x^{\alpha-1} e^{-\lambda x} dx} \cdot 1_{]0,+\infty[} \quad (7.529)$$

En faisant le changement de variables $t = \lambda x$ nous obtenons:

$$\int_0^{+\infty} x^{\alpha-1} e^{-\lambda x} dx = \frac{\Gamma(\alpha)}{\lambda^\alpha} \quad (7.530)$$

et pouvons alors écrire la relation sous une forme plus classique que nous trouvons fréquemment dans les ouvrages:

$$P_{\alpha,\lambda}(x) = \frac{x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} \cdot \lambda^\alpha \cdot 1_{]0,+\infty[} \quad (7.531)$$

et c'est sous cette forme que nous retrouvons cette fonction de distribution dans la version française de Microsoft Excel 11.8346 sous le nom LOI.GAMMA() et pour sa réciproque par LOI.GAMMA.INVERSE().

Si $\alpha \in \mathbb{N}$, la loi Gamma au dénominateur devient (cf. chapitre de Calcul Différentiel Et Intégral) la factorielle $(\alpha - 1)!$. La fonction Gamma peut alors s'écrire:

$$P_{\alpha,\lambda}(x) = \frac{x^{\alpha-1} \lambda^\alpha e^{-\lambda x}}{(\alpha - 1)!} = \frac{(x\lambda)^{\alpha-1}}{(\alpha - 1)!} \lambda e^{-\lambda x} \quad (7.532)$$

Cette forme particulière de la fonction de distribution de la fonction Gamma s'appelle alors la "**fonction d'Erlang**" que nous retrouvons naturellement dans la théorie des files d'attente et qui est donc très

importante dans la pratique!

Remarque: Si $\alpha = 1$ alors $\Gamma(\alpha) = 1$ et $x^{\alpha-1} = 1$ et nous retombons sur la loi exponentielle.

Ensuite, nous vérifions avec un raisonnement similaire en tout point à celui de la fonction bêta que $P_{\alpha,\lambda}(x)$ est une fonction de distribution:

$$\int_{-\infty}^{+\infty} P_{\alpha,\lambda}(x) dx = 1 \quad (7.533)$$

Exemples tracés de la fonction de distribution pour $(\alpha, \lambda) = (0.5, 1)$ en rouge, $(\alpha, \lambda) = (1, 1)$ en vert, $(\alpha, \lambda) = (2, 1)$ en noir, $(\alpha, \lambda) = (4, 2)$ en bleu, $(\alpha, \lambda) = (16, 8)$ en magenta:

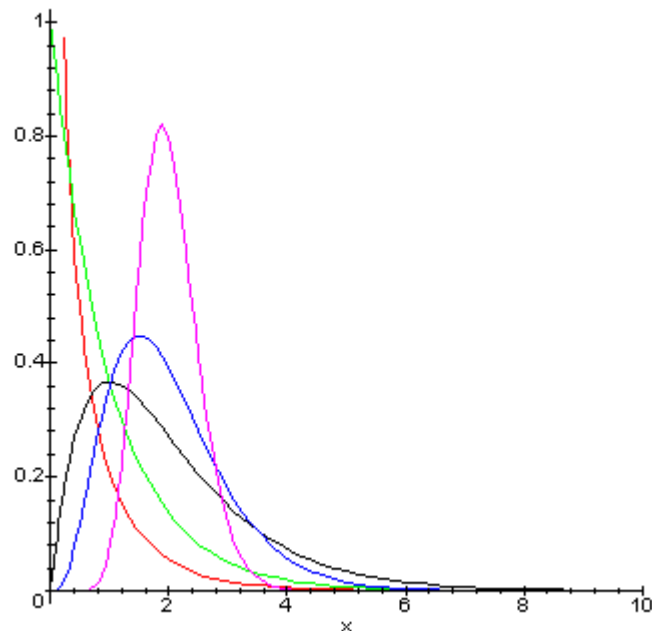


Figure: 7.28 - Loi Gamma (fonction de distribution)

et tracé de la fonction de distribution et répartition pour la fonction Gamma de paramètres $(\alpha, \lambda) = (4, 1)$:

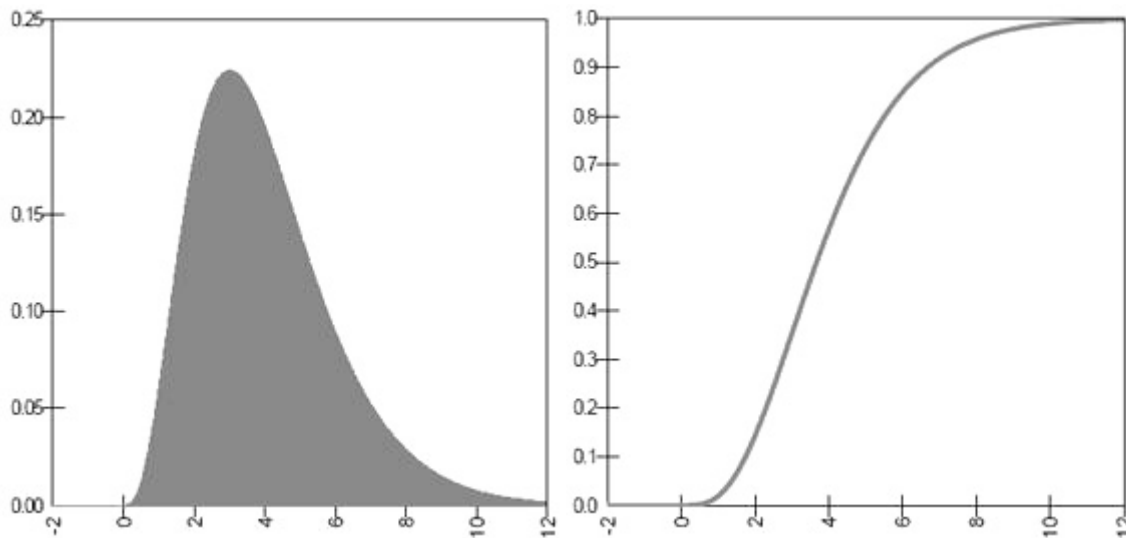


Figure: 7.29 - Loi Gamma (fonction de distribution et de répartition)

La fonction Gamma a par ailleurs pour espérance (moyenne):

$$\mu = \int_{-\infty}^{+\infty} x \cdot f(x) dx = \frac{\lambda^a}{\Gamma(a)} \int_0^{+\infty} x^a e^{-\lambda x} dx = \frac{\lambda^a}{\Gamma(a)} \cdot \frac{\Gamma(a+1)}{\lambda^{a+1}} = \frac{\lambda^a}{\Gamma(a)} \cdot \frac{a\Gamma(a)}{\lambda^{a+1}} = \frac{a}{\lambda} \quad (7.534)$$

et pour variance:

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{+\infty} (x - \mu)^2 \cdot f(x) dx = \frac{\lambda^a}{\Gamma(a)} \int_0^{+\infty} (x - \mu)^2 \cdot x^{a-1} e^{-\lambda x} dx \\ &= \frac{\lambda^a}{\Gamma(a)} \cdot \left(\int_0^{+\infty} x^{a+1} e^{-\lambda x} dx + \mu^2 \int_0^{+\infty} x^{a-1} e^{-\lambda x} dx - 2\mu \int_0^{+\infty} x^a e^{-\lambda x} dx \right) \\ &= \frac{\lambda^a}{\Gamma(a)} \cdot \left(\frac{\Gamma(a+2)}{\lambda^{a+2}} + \mu^2 \frac{\Gamma(a)}{\lambda^a} - 2\mu \frac{\Gamma(a+1)}{\lambda^{a+1}} \right) = \frac{1}{\Gamma(a) \cdot \lambda^2} \cdot (\Gamma(a+2) + a^2 \Gamma(a) - 2a\Gamma(a+1)) \\ &= \frac{1}{\Gamma(a) \cdot \lambda^2} ((a+1) \cdot a \cdot \Gamma(a) + a^2 \Gamma(a) - 2a^2 \Gamma(a)) = \frac{a}{\lambda^2} \end{aligned} \quad (7.535)$$

Démontrons une propriété de la fonction Gamma qui nous servira à établir plus tard dans ce chapitre, lors de notre étude de l'analyse de la variance et des intervalles de confiance sur des petits échantillons, une autre propriété extrêmement importante de la loi du Khi-deux.

Comme nous le savons, la fonction de densité d'une variable aléatoire suivant une fonction Gamma de paramètres $a, \lambda > 0$ est:

$$P_{a,\lambda}(x) = f(x) = \frac{e^{-\lambda x} x^{a-1}}{\Gamma(a)} \lambda^a, \cdot 1_{]0,+\infty[} \quad (7.536)$$

avec (cf. chapitre de Calcul Différentiel Et Intégral) la fonction Gamma d'Euler:

$$\Gamma(a) = \int_0^{+\infty} e^{-x} x^{a-1} dx \quad (7.537)$$

Par ailleurs, quand une variable aléatoire suit une fonction Gamma nous la notons:

$$X = \gamma(\alpha, \lambda) \quad (7.538)$$

Soient X, Y deux variables indépendantes. Montrons que si $X = \gamma(p, \lambda)$ et $Y = \gamma(q, \lambda)$ alors:

$$X + Y = \gamma(p + q, \lambda) \quad (7.539)$$

Notons f la fonction de densité du couple (X, Y) , f_X la fonction de densité de X et f_Y la fonction de densité de Y . Vu que X, Y sont indépendantes, nous avons:

$$f(x, y) = f_X(x) \cdot f_Y(y) \quad (7.540)$$

pour tout $x, y \geq 0$.

Soit $Z = X + Y$. La fonction de répartition de Z est alors:

$$F(z) = P(Z \leq z) = P(X + Y \leq z) = \iint_D f(x, y) dx dy \quad (7.541)$$

où $D = \{(x, y) \mid x + y \leq z\}$.

Remarque: Nous appelons un tel calcul une "**convolution**" et les statisticiens ont souvent à manipuler de telles entités ayant à travailler sur de nombreuses variables aléatoires qu'il faut sommer ou même multiplier.

En simplifiant:

$$F(z) = \int_{-\infty}^{+\infty} \int_{-\infty}^{z-x} f_X(x) \cdot f_Y(y) dy dx \quad (7.542)$$

Nous effectuons le changement de variable suivant:

$$\begin{cases} x = x \\ y = s - x \end{cases} \quad (7.543)$$

Le jacobien est alors (cf. chapitre de Calcul Différentiel Et Intégral):

$$J = \begin{vmatrix} \frac{\partial x}{\partial x} & \frac{\partial x}{\partial s} \\ \frac{\partial y}{\partial x} & \frac{\partial y}{\partial s} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1 \quad (7.544)$$

Donc avec la nouvelle borne d'intégration $s = x + y = x + (z - x) = z$ nous avons:

$$F(z) = \int_{-\infty}^{+\infty} \int_{-\infty}^z f_X(x) \cdot f_Y(s-x) ds dx = \int_{-\infty}^z \int_{-\infty}^{+\infty} f_X(x) \cdot f_Y(s-x) dx ds \quad (7.545)$$

Si nous notons g la fonction de densité de Z nous avons:

$$F(z) = \int_{-\infty}^z \int_{-\infty}^{+\infty} f_X(x) \cdot f_Y(s-x) dx ds = \int_{-\infty}^z g(s) ds \quad (7.546)$$

Par suite:

$$g(s) = \int_{-\infty}^{+\infty} f_X(x) \cdot f_Y(s-x) dx \quad (7.547)$$

f_X et f_Y étant nulles lorsque leur argument est négatif, nous pouvons changer les bornes d'intégration:

$$g(s) = \int_0^s f_X(x) \cdot f_Y(s-x) dx \text{ pour } s \geq 0 \quad (7.548)$$

Calculons g :

$$g(s) = \frac{\lambda^{p+q} e^{-\lambda s}}{\Gamma(p) \cdot \Gamma(q)} \int_0^s x^{p-1} (s-x)^{q-1} dx \quad (7.549)$$

Après le changement de variable $x = st$ nous obtenons:

$$g(s) = \frac{\lambda^{p+q} e^{-\lambda s}}{\Gamma(p) \cdot \Gamma(q)} s^{p+q-1} \int_0^1 t^{p-1} (1-t)^{q-1} dt = \frac{\lambda^{p+q} e^{-\lambda s}}{\Gamma(p) \cdot \Gamma(q)} s^{p+q-1} \cdot B(p, q) \quad (7.550)$$

où B est la fonction bêta que nous avons vue plus haut dans notre étude de la fonction de distribution bêta. Or nous avons aussi démontré la relation:

$$B(p, q) = \frac{\Gamma(p) \Gamma(q)}{\Gamma(p+q)} \quad (7.551)$$

Donc:

$$g(s) = \frac{\lambda^{p+q} e^{-\lambda s}}{\Gamma(p+q)} s^{p+q-1} \quad (7.552)$$

Ce qui finalement nous donne:

$$P(Z \leq z) = P(X+Y \leq z) = \int_0^z g(s) ds = \int_0^z \frac{\lambda^{p+q} e^{-\lambda s}}{\Gamma(p+q)} s^{p+q-1} ds \quad (7.553)$$

Ce qui montre que bien que si deux variables aléatoires suivent une fonction Gamma alors leur somme est aussi telle que:

$$X+Y = \gamma(p+q, \lambda) \quad (7.554)$$

Donc la fonction Gamma est stable par addition de même que le sont toutes les lois qui découlent de la loi Gamma et que nous allons aborder ci-après.

5.18. FONCTION DE KHI-DEUX (OU DE PEARSON)

La "**fonction de Khi-deux**" (appelée aussi "**loi du Khi-deux**" ou encore "**loi de Pearson**") n'est par définition qu'un cas particulier de la fonction de distribution Gamma dans le cas où $\alpha = k/2$ et $\lambda = 1/2$, avec k entier positif:

$$P_k(x) = \frac{1}{2^{k/2} \cdot \Gamma(k/2)} x^{k/2-1} e^{-x/2} \cdot 1_{[0,+\infty[} \quad (7.555)$$

Cette relation qui relie la loi du Khi-deux à la loi Gamma est importante dans la version française de Microsoft Excel 11.8346 car la fonction LOI.KHIDEUX() donne le seuil de confiance et non la loi de distribution. Il faut alors utiliser la fonction LOI.GAMMA() avec les paramètres donnés ci-dessus (à part qu'il faut prendre l'inverse de 1/2, soit 2 comme paramètre) pour avoir la fonction de distribution et de répartition.

Le lecteur qui voudra vérifier que la loi du Khi-2 est un cas particulier de la loi Gamma, pourra écrire dans la version française de Microsoft Excel 14.0.6123:

$$\begin{aligned} &= \text{LOI.KHIDEUX.N}(2*x; 2*k; \text{VRAI}) \\ &= \text{LOI.GAMMA.N}(x; k; 1; \text{VRAI}) \end{aligned}$$

Tous les calculs faits auparavant s'appliquent et nous avons alors immédiatement:

$$\mu = k, \sigma^2 = 2k \quad (7.556)$$

Exemples de tracés de la fonction de distribution pour $k = 1$ en rouge, $k = 2$ en vert, $k = 3$ en noir, $k = 4$ en bleu:

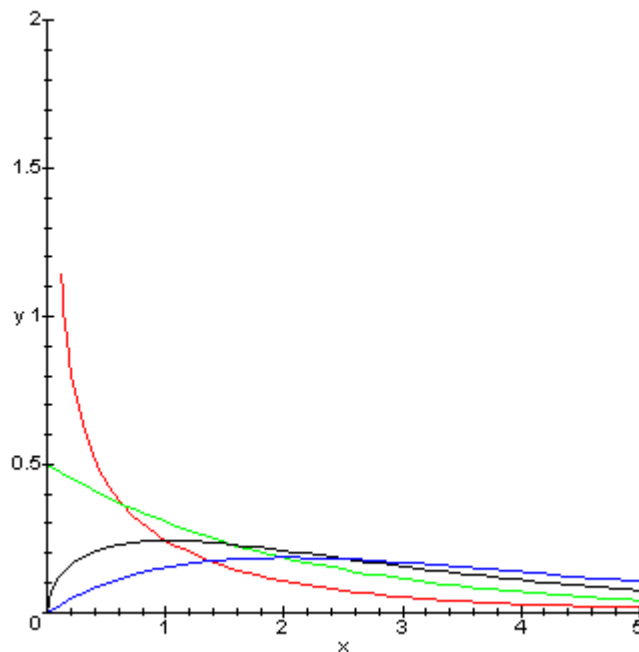


Figure: 7.30 - Loi du Khi-deux (fonctions de distribution)

et tracé de la fonction de distribution et respectivement de répartition pour la loi du Khi-deux pour $k = 2$:

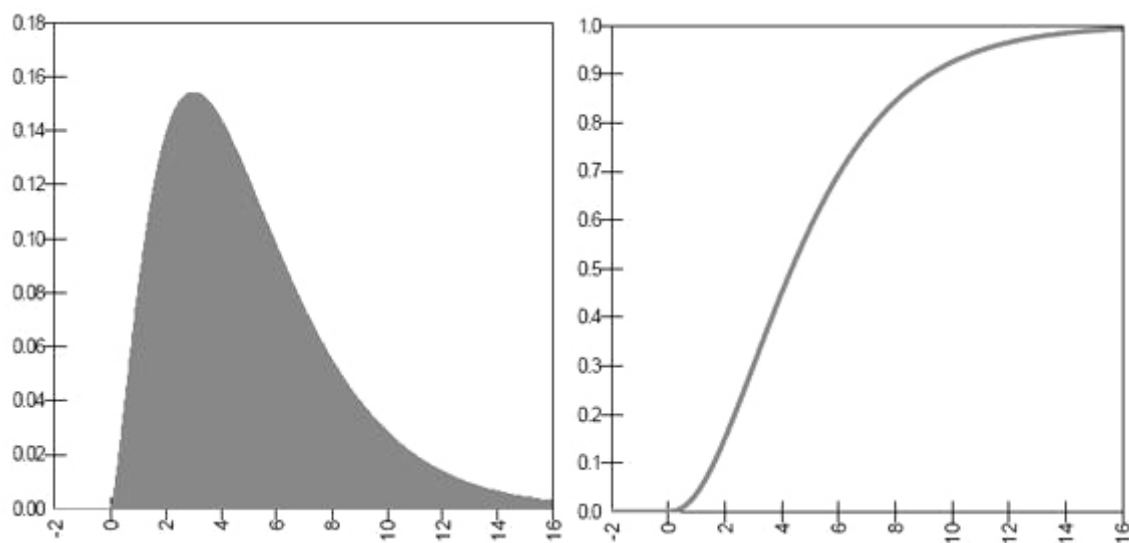


Figure: 7.31 - Loi du Khi-deux (fonction de distribution et de répartition)

Dans la littérature, il est de tradition de noter:

$$X = \chi_k^2 \text{ ou } X = \chi^2(k) \quad (7.557)$$

pour indiquer que la distribution de la variable aléatoire X est la loi du Khi-deux. Par ailleurs il est courant de nommer le paramètre k "**degré de liberté**" et de l'abréger "**ddl**".

La fonction Khi-deux découle donc de la loi Gamma et par ailleurs en prenant $k = 2$ nous retrouvons aussi la loi exponentielle (voir plus haut) pour $\lambda = 1/2$:

$$P_2(x) = \frac{1}{2^1 \Gamma(1)} x^{2/2-1} e^{-x/2} = \frac{1}{2} e^{-x/2} 1_{[0,+\infty[} \quad (7.558)$$

Par ailleurs, puisque (cf. chapitre de Calcul Différentiel Et Intégral):

$$\Gamma(1/2) = \sqrt{\pi} \quad (7.559)$$

la loi du Khi-deux avec k égal à l'unité peut s'écrire sous la forme:

$$P_2(x) = \frac{1}{2^{1/2} \Gamma(1/2)} x^{1/2-1} e^{-x/2} = \frac{1}{\sqrt{2\pi x}} e^{-x/2} 1_{[0,+\infty[} \quad (7.560)$$

Enfin, terminons avec une propriété assez importante dans les tests statistiques que nous étudierons un peu plus loin et particulièrement dans les intervalles de confiance des événements rares.

Effectivement, le lecteur pourra vérifier dans un tableur comme Microsoft Excel 14.0.6123 (version française), que nous avons:

$$\begin{aligned} &= \text{LOI.POISSON.N}(x \in \mathbb{N}; \mu; \text{VRAD}) \\ &= 1 - \text{LOI.KHIDEUX.N}(2*\mu; 2*(x+1); \text{VRAD}) \quad (7.561) \\ &= 1 - \text{LOI.GAMMA.N}(\mu; x+1; \text{VRAD}) \end{aligned}$$

Il nous faut donc démontrer cette relation entre loi du khi-2 et loi de Poisson. Voyons cela en partant de la loi Gamma:

$$P_{\alpha, \lambda}(x) = \frac{\lambda^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} \cdot \lambda^{\alpha} \cdot 1_{[0,+\infty[} \quad (7.562)$$

Si nous posons $\lambda = 1/2$ et $\alpha = k/2$ nous avons alors la loi du khi-2 à k degré de libertés:

$$P_k(x) = \frac{1}{2^{k/2} \cdot \Gamma(k/2)} x^{k/2-1} e^{-x/2} \cdot 1_{[0,+\infty[} \quad (7.563)$$

Maintenant, rappelons que nous avons vu dans le chapitre de Suites Et Séries, la série de Taylor (Maclaurin) avec reste intégral à l'ordre $n-1$ autour de 0 jusqu'à λ suivante:

$$\begin{aligned} e^{\lambda} &= \sum_{k=0}^{n-1} \frac{\lambda^k}{k!} e^x(0) + \int_0^{\lambda} e^t \frac{(\lambda-t)^{n-1}}{(n-1)!} dt = \sum_{k=0}^{n-1} \frac{\lambda^k}{k!} + \int_0^{\lambda} e^t \frac{(\lambda-t)^{n-1}}{(n-1)!} dt \\ &= \sum_{k=0}^{n-1} \frac{\lambda^k}{k!} - \int_{\lambda}^0 e^{\lambda-u} \frac{u^{n-1}}{(n-1)!} du = \sum_{k=0}^{n-1} \frac{\lambda^k}{k!} + \int_0^{\lambda} e^{\lambda-u} \frac{u^{n-1}}{(n-1)!} du \end{aligned} \quad (7.564)$$

Nous multiplions par $e^{-\lambda}$:

$$e^{-\lambda} e^{\lambda} = e^{-\lambda} \left(\sum_{k=0}^{n-1} \frac{\lambda^k}{k!} + \int_0^{\lambda} e^{\lambda-u} \frac{u^{n-1}}{(n-1)!} du \right) \Rightarrow 1 = \sum_{k=0}^{n-1} \frac{\lambda^k}{k!} e^{-\lambda} + \int_0^{\lambda} e^{-u} \frac{u^{n-1}}{(n-1)!} du \quad (7.565)$$

Et donc:

$$\sum_{k=0}^{n-1} \frac{\lambda^k}{k!} e^{-\lambda} = 1 - \int_0^{\lambda} e^{-u} \frac{u^{n-1}}{(n-1)!} du \quad (7.566)$$

Or, concentrons-nous sur le terme:

$$\int_0^{\lambda} e^{-u} \frac{u^{n-1}}{(n-1)!} du \quad (7.567)$$

et faisons un premier changement de variable:

$$\int_0^{\lambda} e^{-u} \frac{u^{n-1}}{(n-1)!} du = \int_{u=\lambda/2}^{\lambda} \frac{1}{2^{n-1} (n-1)!} x^{n-1} e^{-x/2} \frac{1}{2} dx = \int_0^{\lambda} \frac{1}{2^n (n-1)!} x^{n-1} e^{-x/2} dx \quad (7.568)$$

et un second changement de variable (attention! le k dans le changement de variable n'est pas le même que celui de la somme de la loi de Poisson...):

$$\int_0^{\lambda} \frac{1}{2^n (n-1)!} x^{n-1} e^{-x/2} dx = \int_{n=k/2}^{\lambda} \frac{1}{2^{k/2} \left(\frac{k}{2} - 1\right)!} x^{k/2-1} e^{-x/2} dx \quad (7.569)$$

Or, nous avons démontré dans le chapitre de Calcul Différentiel Et Intégral que si x est un entier strictement positif:

$$x! = \Gamma(x+1) \quad (7.570)$$

Il vient alors:

$$\left(\frac{k}{2} - 1\right)! = \Gamma\left(\left(\frac{k}{2} - 1\right) + 1\right) = \Gamma\left(\frac{k}{2}\right) \quad (7.571)$$

Nous avons finalement:

$$\int_0^{\lambda} \frac{1}{2^{k/2-1} \Gamma(k/2)} x^{k/2-1} e^{-x/2} dx \quad (7.572)$$

où nous retrouvons donc bien la fonction de distribution du khi-2 sous l'intégrale! Donc au final:

$$\sum_{k=0}^{n-1} \frac{\lambda^k}{k!} e^{-\lambda} = 1 - \int_0^{\lambda} \frac{1}{2^{k/2-1} \Gamma(k/2)} x^{k/2-1} e^{-x/2} dx \quad (7.573)$$

D'où la relation donnée plus haut pour les tableurs!

5.19. FONCTION DE STUDENT

La "**fonction de Student**" (ou "**loi de Student**") de paramètre k est définie par la relation:

$$P(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}} \quad (7.574)$$

avec k étant le degré de liberté de la loi du Khi-deux sous-jacente à la construction de la fonction de Student comme nous allons le voir.

Indiquons qu'elle peut aussi être obtenue dans la version français de Microsoft Excel 11.8346 à l'aide des fonctions LOI.STUDENT() et sa réciproque par LOI.STUDENT.INVERSE().

Il s'agit bien d'une fonction de distribution car elle vérifie également (reste à démontrer directement mais bon comme nous allons le voir elle est le produit de deux fonctions de distribution donc indirectement...):

$$\int_{-\infty}^{+\infty} P_k(x) dx = 1 \quad (7.575)$$

Voyons la démonstration la plus simple pour justifier la provenance de la loi de Student et qui nous sera en même temps très utile dans l'inférence statistique et l'analyse de la variance plus loin.

Pour cette démonstration, rappelons que:

R1. Si X, Y sont deux variables aléatoires indépendantes de densités respectives f_X, f_Y , la loi du couple (X, Y) possède une densité f vérifiant (axiome des probabilités!):

$$f(x, y) = f_X(x) \cdot f_Y(y) \quad (7.576)$$

R2. La loi $N(0, 1)$ est donnée par (voir plus haut):

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad (7.577)$$

R3. La loi χ_n^2 est donnée par (voir précédemment):

$$f(y) = \frac{1}{2^{n/2} \Gamma(n/2)} e^{-y/2} y^{n/2-1} \quad (7.578)$$

pour $y \geq 0$ et $n \geq 1$.

R4. La fonction Γ est définie pour tout $\alpha > 0$ par (cf. chapitre de Calcul Différentiel et Intégral):

$$\Gamma(\alpha) = \int_0^{+\infty} e^{-x} x^{\alpha-1} dx \quad (7.579)$$

et vérifie (cf. chapitre de Calcul Différentiel et Intégral):

$$\Gamma(\alpha-1) = \frac{\Gamma(\alpha)}{\alpha-1} \quad (7.580)$$

pour $\alpha \geq 2$.

Ces rappels étant faits, considérons maintenant X une variable aléatoire suivant la loi $N(0,1)$ et Y une variable aléatoire suivant la loi χ_n^2 .

Nous supposons X et Y indépendantes et nous considérons la variable aléatoire (c'est à l'origine l'étude historique de la loi de Student dans le cadre de l'inférence statistique qui a amené à poser cette variable dont nous justifierons l'origine plus loin):

$$T = \sqrt{n} \frac{X}{\sqrt{Y}} = \frac{X}{\sqrt{Y/n}} \quad (7.581)$$

Nous allons montrer que T suit une loi de Student de paramètre n .

Démonstration:

Notons F et f les fonctions de répartition et de densité de T et f_X, f_Y, f les fonctions de densité de X , Y et (X,Y) respectivement. Nous avons alors pour tout $t \in \mathbb{R}$:

$$F(t) = P(T \leq t) = P\left(\sqrt{n} \frac{X}{\sqrt{Y}} \leq t\right) = \iint_D f(x, y) dx dy = \iint_D f_X(x) \cdot f_Y(y) dx dy \quad (7.582)$$

où:

$$D = \left\{ (x, y) \in \mathbb{R} \times \mathbb{R}^+ \mid x \leq \frac{t\sqrt{y}}{\sqrt{n}} \right\} \quad (7.583)$$

la valeur imposée positive et non nulle de y étant due au fait qu'elle est sous une racine et en plus au dénominateur.

Ainsi:

$$F(t) = \iint_D f_X(x) \cdot f_Y(y) dx dy = \int_0^{+\infty} f_Y(y) \int_{-\infty}^{t\sqrt{y}/\sqrt{n}} f_X(x) dx dy = \int_0^{+\infty} f_Y(y) \Phi\left(t\sqrt{y}/\sqrt{n}\right) dy \quad (7.584)$$

où comme X suit une loi $N(0,1)$:

$$\Phi(u) = \int_{-\infty}^u \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \quad (7.585)$$

est la fonction de répartition de la loi Normale centrée réduite.

Nous obtenons alors la fonction de densité de T en dérivant F :

$$f(t) = F'(t) = \frac{1}{\sqrt{n}} \int_0^{+\infty} f_Y(y) f_X\left(t\sqrt{y}/\sqrt{n}\right) \sqrt{y} dy \quad (7.586)$$

car (la dérivée d'une fonction est égale à sa dérivée multipliée par sa dérivée intérieure):

$$\frac{d\Phi\left(t\sqrt{y}/\sqrt{n}\right)}{dt} = f_X\left(t\sqrt{y}/\sqrt{n}\right) \frac{\sqrt{y}}{\sqrt{n}} \quad (7.587)$$

Donc:

$$\begin{aligned} f(t) &= \frac{1}{\sqrt{n}} \int_0^{+\infty} f_Y(y) f_X\left(t\sqrt{y}/\sqrt{n}\right) \sqrt{y} dy \\ &= \frac{1}{\sqrt{n}} \int_0^{+\infty} \frac{1}{2^{n/2} \Gamma(n/2)} e^{-y/2} y^{\frac{n}{2}-1} \frac{1}{\sqrt{2\pi}} e^{-\frac{(t\sqrt{y}/\sqrt{n})^2}{2}} \sqrt{y} dy \\ &= \frac{1}{2^{n/2} \Gamma(n/2) \sqrt{2\pi} \sqrt{n}} \int_0^{+\infty} e^{-y/2} y^{\frac{n}{2}-1} e^{-\frac{(t\sqrt{y}/\sqrt{n})^2}{2}} \sqrt{y} dy \\ &= \frac{1}{2^{n/2} \Gamma(n/2) \sqrt{2\pi} \sqrt{n}} \int_0^{+\infty} e^{-\frac{y(1+t^2/n)}{2}} y^{\frac{n}{2}-1} dy \end{aligned} \quad (7.588)$$

En faisant le changement de variable:

$$u = \frac{y}{2} \left(1 + \frac{t^2}{n}\right) \Rightarrow \begin{cases} dy = \left(\frac{2}{1 + \frac{t^2}{n}} \right) du \\ y = \left(\frac{1 + \frac{t^2}{n}}{2} \right) u \end{cases} \quad (7.589)$$

nous obtenons:

$$\begin{aligned} f(t) &= \frac{1}{2^{n/2} \Gamma(n/2) \sqrt{2\pi} \sqrt{n}} \left(\frac{2}{1 + t^2/n} \right)^{\frac{n+1}{2}} \underbrace{\int_0^{+\infty} e^{-u} u^{\frac{n-1}{2}} du}_{\Gamma\left(\frac{n+1}{2}\right)} = \frac{\Gamma\left(\frac{n+1}{2}\right)}{2^{n/2} \Gamma(n/2) \sqrt{2\pi} \sqrt{n}} \left(\frac{2}{1 + t^2/n} \right)^{\frac{n+1}{2}} \\ &= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(n/2) \sqrt{n\pi}} \left(1 + \frac{t^2}{n} \right)^{-\frac{n+1}{2}} \end{aligned} \quad (7.590)$$

ce qui est bien la loi de Student de paramètre n .

Voyons maintenant quelle est l'espérance de la loi de Student:

$$T = \sqrt{n} \frac{X}{\sqrt{Y}} \quad (7.591)$$

Nous avons:

$$E(T) = E(\sqrt{n}X) \cdot E\left(\frac{1}{\sqrt{Y}}\right) \quad (7.592)$$

Mais $E\left(\frac{1}{\sqrt{Y}}\right)$ existe si et seulement si $n \geq 2$. En effet pour $n = 1$:

$$E\left(\frac{1}{\sqrt{Y}}\right) = \int_0^{+\infty} \frac{1}{\sqrt{y}} f_Y(y) dy = \frac{1}{2^{n/2} \Gamma(n/2)} \int_0^{+\infty} e^{-y/2} y^{\frac{n-3}{2}} dy \quad (7.593)$$

et:

$$\int_0^{+\infty} e^{-y/2} y^{\frac{n-3}{2}} dy = \int_0^{+\infty} \frac{e^{-y/2}}{y} dy \geq \int_0^1 \frac{e^{-y/2}}{y} dy \geq e^{-1/2} \int_0^1 \frac{1}{y} dy = +\infty \quad (7.594)$$

Tandis que pour $n \geq 2$ nous avons:

$$\int_0^{+\infty} e^{-y/2} y^{\frac{n-3}{2}} dy = 2^{\frac{n-1}{2}} \int_0^{+\infty} e^{-u} u^{\frac{n-3}{2}} du = 2^{\frac{n-1}{2}} \Gamma\left(\frac{n-1}{2}\right) < +\infty \quad (7.595)$$

Ainsi pour $n = 1$, l'espérance n'existe pas.

Donc pour $n \geq 2$:

$$E(T) = \sqrt{n} \underbrace{E(X)}_0 \cdot E\left(\frac{1}{\sqrt{Y}}\right) = 0 \quad (7.596)$$

Voyons maintenant la valeur de la variance. Nous avons donc:

$$V(T) = E(T^2) - E(T)^2 \quad (7.597)$$

Discutons de l'existence de $E(T^2)$. Nous avons trivialement:

$$E(T^2) = n E\left(\frac{X^2}{Y}\right) = n E(X^2) E\left(\frac{1}{Y}\right) \quad (7.598)$$

X suit une loi normale centrée réduite donc:

$$V(X) = 1 = E(X^2) - \underbrace{E(X)^2}_0 = E(X^2) \Rightarrow E(X^2) = 1 \quad (7.599)$$

Pour ce qui est de $E\left(\frac{1}{Y}\right)$ nous avons:

$$\begin{aligned} E\left(\frac{1}{Y}\right) &= \int_0^{+\infty} \frac{1}{y} f_Y(y) dy = \frac{1}{2^{n/2} \Gamma(n/2)} \int_0^{+\infty} e^{-y/2} y^{\frac{n}{2}-2} dy = \frac{1}{2\Gamma(n/2)} \int_0^{+\infty} e^{-u} u^{\frac{n}{2}-2} du \\ &= \frac{\Gamma\left(\frac{n}{2}-1\right)}{2\Gamma(n/2)} \end{aligned} \quad (7.600)$$

où nous avons fait le changement de variable $u = y/2$.

Mais l'intégrale définissant $\Gamma\left(\frac{n}{2}-1\right)$ converge seulement si $n \geq 3$.

Donc $E(T^2)$ existe si et seulement si $n \geq 3$ et vaut alors selon les propriétés de la loi Gamma d'Euler démontrées dans le chapitre de Calcul Différentiel et Intégral:

$$nE(T) = n \frac{\Gamma\left(\frac{n-2}{2}\right)}{2\Gamma(n/2)} = \frac{n}{n-2} \quad (7.601)$$

Ainsi pour $n \geq 3$:

$$V(T) = \frac{n}{n-2} \quad (7.602)$$

Il est par ailleurs important de remarquer que cette loi est symétrique par rapport à 0!

Exemple de tracé de la fonction de distribution et répartition pour la fonction de Student de paramètre $k = 3$:

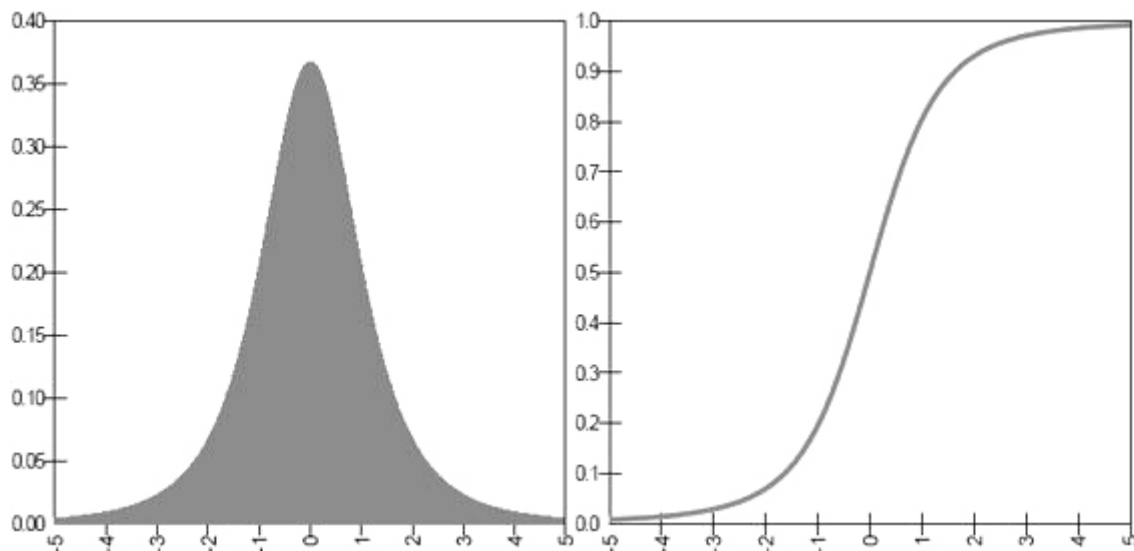


Figure: 7.32 - Loi de Student (fonction de distribution et de répartition)

5.20. FONCTION DE FISHER

La "**fonction de Fisher**" (ou "**loi de Fisher-Snedecor**") de paramètres k et l est définie par la relation:

$$P(x) = \frac{\Gamma\left(\frac{k+l}{2}\right) \left(\frac{k}{l}\right)^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right) \Gamma\left(\frac{l}{2}\right)} x^{\frac{k}{2}-1} \left(1 + \frac{k}{l}x\right)^{-\frac{(k+l)}{2}} \quad (7.603)$$

si $x \geq 0$. Les paramètres k et l sont des entiers positifs et correspondent aux degrés de liberté des deux lois du Khi-deux sous-jacentes. Cette distribution est souvent notée $F_{k,l}$ ou $F(k,l)$ et peut être obtenue dans la version française de Microsoft Excel 11.8346 par la fonction LOI.F().

Il s'agit bien d'une fonction de distribution car elle vérifie également (reste à démontrer directement mais bon comme nous allons le voir elle est le produit de deux fonctions de distribution donc indirectement...):

$$\int_0^{+\infty} F_{k,l}(x) dx = 1 \quad (7.604)$$

Voyons la démonstration la plus simple pour justifier la provenance de la loi de Fisher et qui nous sera en même temps très utile dans l'inférence statistique et l'analyse de la variance plus loin.

Pour cette démonstration, rappelons que:

R1. La loi χ_n^2 est donnée par (voir plus haut):

$$f(y) = \frac{1}{2^{n/2} \Gamma(n/2)} e^{-y/2} y^{\frac{n}{2}-1} \quad (7.605)$$

pour $y \geq 0$ et $n \geq 1$.

R2. La fonction Γ est définie pour tout $\alpha > 0$ par (cf. chapitre de Calcul Différentiel et Intégral):

$$\Gamma(\alpha) = \int_0^{+\infty} e^{-x} x^{\alpha-1} dx \quad (7.606)$$

Soient X, Y deux variables aléatoires indépendantes suivant respectivement les lois χ_n^2 et χ_m^2 .

Nous considérons la variable aléatoire:

$$F = \frac{X/n}{Y/m} \quad (7.607)$$

Nous allons donc montrer que la loi de T est la loi de Fisher-Snedecor de paramètres n, m .

Notons pour cela F et f les fonctions de répartition et de densité de T et f_X, f_Y, f les fonctions de densité de X, Y et (X, Y) respectivement. Nous avons pour tout $t \in \mathbb{R}$:

$$F(t) = P(T \leq t) = P\left(\frac{X/n}{Y/m} \leq t\right) = \iint_D f(x, y) dx dy = \iint_D f_X(x) \cdot f_Y(y) dx dy \quad (7.608)$$

où:

$$D = \left\{ (x, y) \in \mathbb{R}^+ \times \mathbb{R}^+ \mid x \leq \frac{nt}{m} y \right\} \quad (7.609)$$

où les valeurs positives imposées proviennent à l'origine d'une loi du Khi-deux pour x et y .

Ainsi:

$$F(t) = \iint_D f_X(x) \cdot f_Y(y) dx dy = \int_0^{+\infty} f_Y(y) \int_0^{\frac{nt}{m}y} f_X(x) dx dy \quad (7.610)$$

Nous obtenons la fonction de densité de T en dérivant F . D'abord la dérivée intérieure:

$$f(t) = F'(t) = \frac{n}{m} \int_0^{+\infty} f_Y(y) f_X\left(\frac{nt}{m}y\right) y dy \quad (7.611)$$

Ensuite en explicitant puisque:

$$f(y) = \frac{1}{2^{m/2} \Gamma(m/2)} e^{-y/2} y^{\frac{m}{2}-1} \text{ et } f\left(\frac{nt}{m}y\right) = \frac{1}{2^{n/2} \Gamma(n/2)} e^{-\frac{nt}{2m}y} \left(\frac{nt}{m}y\right)^{\frac{n}{2}-1} \quad (7.612)$$

nous avons alors:

$$\begin{aligned}
f(t) &= \frac{n}{m} \int_0^{+\infty} \frac{1}{2^{m/2} \Gamma(m/2)} e^{-y/2} y^{\frac{m}{2}-1} \frac{1}{2^{n/2} \Gamma(n/2)} e^{-\frac{nt}{2m}y} \left(\frac{nt}{m}y\right)^{\frac{n}{2}-1} y dy \\
&= \frac{1}{2^{n/2} \Gamma(n/2) 2^{m/2} \Gamma(m/2)} \frac{n}{m} \int_0^{+\infty} e^{-y/2} y^{\frac{m}{2}-1} e^{-\frac{nt}{2m}y} \left(\frac{nt}{m}\right)^{\frac{n}{2}-1} y^{\frac{n}{2}-1} y dy \\
&= \frac{\frac{n}{m}}{2^{n/2} \Gamma(n/2) 2^{m/2} \Gamma(m/2)} \left(\frac{nt}{m}\right)^{\frac{n}{2}-1} \int_0^{+\infty} e^{-y/2} y^{\frac{m}{2}-1} e^{-\frac{nt}{2m}y} y^{\frac{n}{2}-1} y dy \\
&= \frac{\frac{n}{m}}{2^{(n+m)/2} \Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{m}{2}\right)} \left(\frac{nt}{m}\right)^{\frac{n}{2}-1} \int_0^{+\infty} y^{\frac{n+m}{2}-1} e^{-\frac{y}{2}\left(1+\frac{nt}{m}\right)} dy
\end{aligned} \tag{7.613}$$

En faisant le changement de variable:

$$u = \frac{y}{2} \left(1 + \frac{nt}{m}\right) \Rightarrow y = u \cdot \frac{2}{\left(1 + \frac{n}{m}t\right)} \Rightarrow dy = du \frac{2}{\left(1 + \frac{n}{m}t\right)} \tag{7.614}$$

nous obtenons:

$$\begin{aligned}
f(t) &= \frac{\frac{n}{m}}{2^{(n+m)/2} \Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{m}{2}\right)} \left(\frac{nt}{m}\right)^{\frac{n}{2}-1} \frac{2^{\frac{n+m}{2}}}{\left(1 + \frac{n}{m}t\right)^{\frac{n+m}{2}}} \underbrace{\int_0^{+\infty} u^{\frac{n+m}{2}-1} e^{-u} du}_{\Gamma\left(\frac{n+m}{2}\right)} \\
&= \frac{\frac{n}{m} \Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{m}{2}\right)} \left(\frac{nt}{m}\right)^{\frac{n}{2}-1} \left(1 + \frac{n}{m}t\right)^{-\frac{n+m}{2}} = \frac{\frac{n}{m} \Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{m}{2}\right)} \left(\frac{n}{m}\right)^{\frac{n}{2}-1} t^{\frac{n}{2}-1} \left(1 + \frac{n}{m}t\right)^{-\frac{n+m}{2}} \\
&= \frac{\Gamma\left(\frac{n+m}{2}\right) \left(\frac{n}{m}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{m}{2}\right)} t^{\frac{n}{2}-1} \left(1 + \frac{n}{m}t\right)^{-\frac{n+m}{2}}
\end{aligned} \tag{7.615}$$

□C.Q.F.D.

5.21. FONCTION DE BENFORD

Cette distribution aurait été découverte une première fois en 1881 par Simon Newcomb, un astronome américain, après qu'il se fut aperçu de l'usure (et donc de l'utilisation) préférentielle des premières pages des tables de logarithmes (alors compilées dans des ouvrages). Frank Benford, aux alentours de 1938, remarqua à son tour cette usure inégale, crut être le premier à formuler cette loi qui porte indûment son nom aujourd'hui et arriva aux mêmes résultats après avoir répertorié des dizaines de milliers de données (longueurs de fleuves, cours de la bourse, etc.).

Seule explication possible: nous avons plus souvent besoin d'extraire le logarithme de chiffres commençant par 1 que de chiffres commençant par 9, ce qui implique que les premiers sont "plus nombreux" que les seconds.

Bien que cette idée lui paraisse tout à fait invraisemblable, Benford entreprend de vérifier son hypothèse. Rien de plus simple: il se procure des tables de valeurs numériques, et calcule le pourcentage d'apparition du chiffre le plus à gauche (première décimale). Les résultats qu'il obtient confirment son intuition:

Chiffre initial	Probabilité d'apparition
1	30.1 %
2	17.6 %
3	12.5 %
4	9.7 %
5	7.9 %
6	6.7 %
7	5.8 %
8	5.1 %
9	4.6 %

Tableau: 7.10 - Probabilité d'apparition d'un chiffre selon la loi de Benford

A partir de ces données, Benford trouve expérimentalement que la probabilité cumulée qu'un nombre commence par le chiffre n (excepté 0) est (nous allons le démontrer plus loin) donnée par la relation:

$$P(n) = \log_{10} \left(1 + \frac{1}{n} \right) \quad (7.616)$$

appelée "**fonction de Benford**" (ou "**loi de Benford**").

Voici un tracé de la fonction précédente:

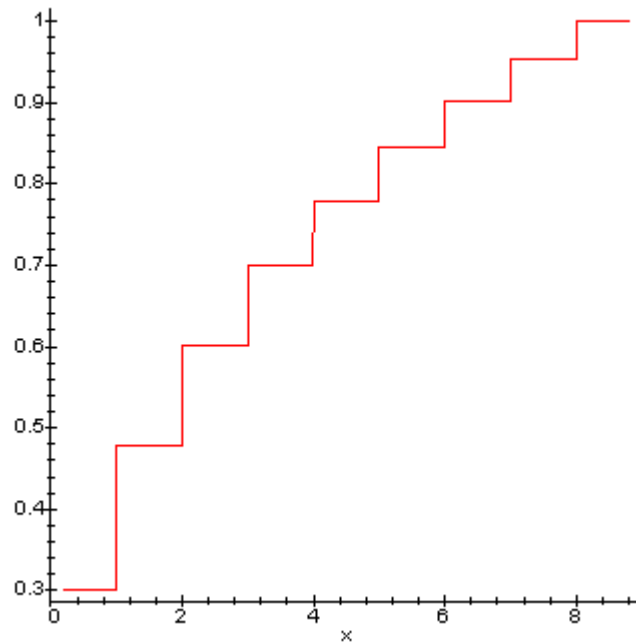


Figure: 7.33 - Tracé de la fonction de Benford (fonction de répartition)

Il convient de préciser que cette loi ne s'applique qu'à des listes de valeurs "naturelles", c'est-à-dire à des chiffres ayant une signification physique. Elle ne fonctionne évidemment pas sur une liste de chiffres tirés au hasard.

La loi de Benford a été testée sur toutes sortes de tables: longueur des fleuves du globe, superficie des pays, résultat des élections, liste des prix de l'épicerie du coin... Elle se vérifie presque à tous les coups.

Elle est évidemment indépendante de l'unité choisie. Si l'on prend par exemple la liste des prix d'un supermarché, elle fonctionne aussi bien avec les valeurs exprimées en Francs qu'avec les mêmes prix convertis en Euros.

Cet étrange phénomène est resté peu étudié et inexpliqué jusqu'à une époque assez récente. Puis une démonstration générale en a été donnée en 1996, qui fait appel au théorème de la limite centrale.

Aussi surprenant que cela puisse paraître, cette loi a trouvé une application: le fisc l'utilise aux Etats-Unis pour détecter les fausses déclarations. Le principe est basé sur la restriction vue plus haut: la loi de Benford ne s'applique que sur des valeurs ayant une signification physique.

S'il existe une distribution de probabilité universelle $P(n)$ sur de tels nombres, ils doivent être invariants sous un changement d'échelle tel que:

$$P(kn) = f(k)P(n) \quad (7.617)$$

Si:

$$\int P(n)dn = 1 \quad (7.618)$$

alors:

$$\int P(kn)dn = 1/k \quad (7.619)$$

et la normalisation de la distribution donne:

$$f(k) = 1/k \quad (7.620)$$

si nous dérivons $P(kn) = f(k)P(n)$ par rapport à k nous obtenons:

$$\begin{aligned} \frac{d}{dk} P(kn) &= \frac{d}{dk} \frac{1}{k} P(n) \\ nP'(kn) &= P(n) \frac{d}{dk} \frac{1}{k} \\ nP'(kn) &= -P(n) \frac{1}{k^2} \end{aligned} \quad (7.621)$$

en posant $k = 1$ nous avons:

$$nP'(n) = -P(n) \quad (7.622)$$

Cette équation différentielle a pour solution:

$$P(n) = 1/n \quad (7.623)$$

Cette fonction, n'est pas en premier lieu à proprement parler une fonction de distribution de probabilité (elle diverge) et deuxièmement, les lois de la physique et humaines imposent des limites.

Nous devons donc comparer cette distribution par rapport à une référence arbitraire. Ainsi, si le nombre décimal étudié contient plusieurs puissance de 10 (10 au total: 0,1,2,3,4,5,6,7,9) la probabilité que le premier chiffre non nul (décimal) soit D est donnée par la distribution logarithmique:

$$P_D = \frac{\int_D^{D+1} P(n)dn}{\int_1^{10} P(n)dn} \quad (7.624)$$

Les bornes de l'intégrale sont de 1 à 10 puisque la valeur nulle est interdite.

L'intégrale du dénominateur donne:

$$\int_1^{10} P(n)dn = \int_1^{10} \frac{1}{n} dn = \ln(10) - \ln(1) = \ln(10) \quad (7.625)$$

L'intégrale du numérateur donne:

$$\int_D^{D+1} P(n)dn = \int_D^{D+1} \frac{1}{n} dn = \ln(D+1) - \ln(D) = \ln\left(\frac{D+1}{D}\right) \quad (7.626)$$

Ce qui nous donne finalement:

$$P_D = \frac{\ln\left(\frac{D+1}{D}\right)}{\ln(10)} = \frac{\ln\left(1 + \frac{1}{D}\right)}{\ln(10)} \quad (7.627)$$

De par les propriétés des logarithmes (voir le chapitre d'Analyse fonctionnelle) nous avons:

$$P_D = \log_{10}\left(1 + \frac{1}{D}\right) \quad (7.628)$$

Cependant, la loi de Benford ne s'applique pas uniquement aux données invariantes par changement d'échelle mais également à des nombres provenant de sources quelconques. Expliquer ce cas implique une investigation plus rigoureuse en utilisant le théorème de la limite centrale. Cette démonstration a été effectuée seulement en 1996 par T. Hill par une approche utilisant la distribution des distributions.

Pour résumer une partie importante de tout ce que nous avons vu jusqu'ici, l'illustration ci-dessous est très utile car elle résume les relations 76 distributions univariées (57 continues et 19 discrètes):

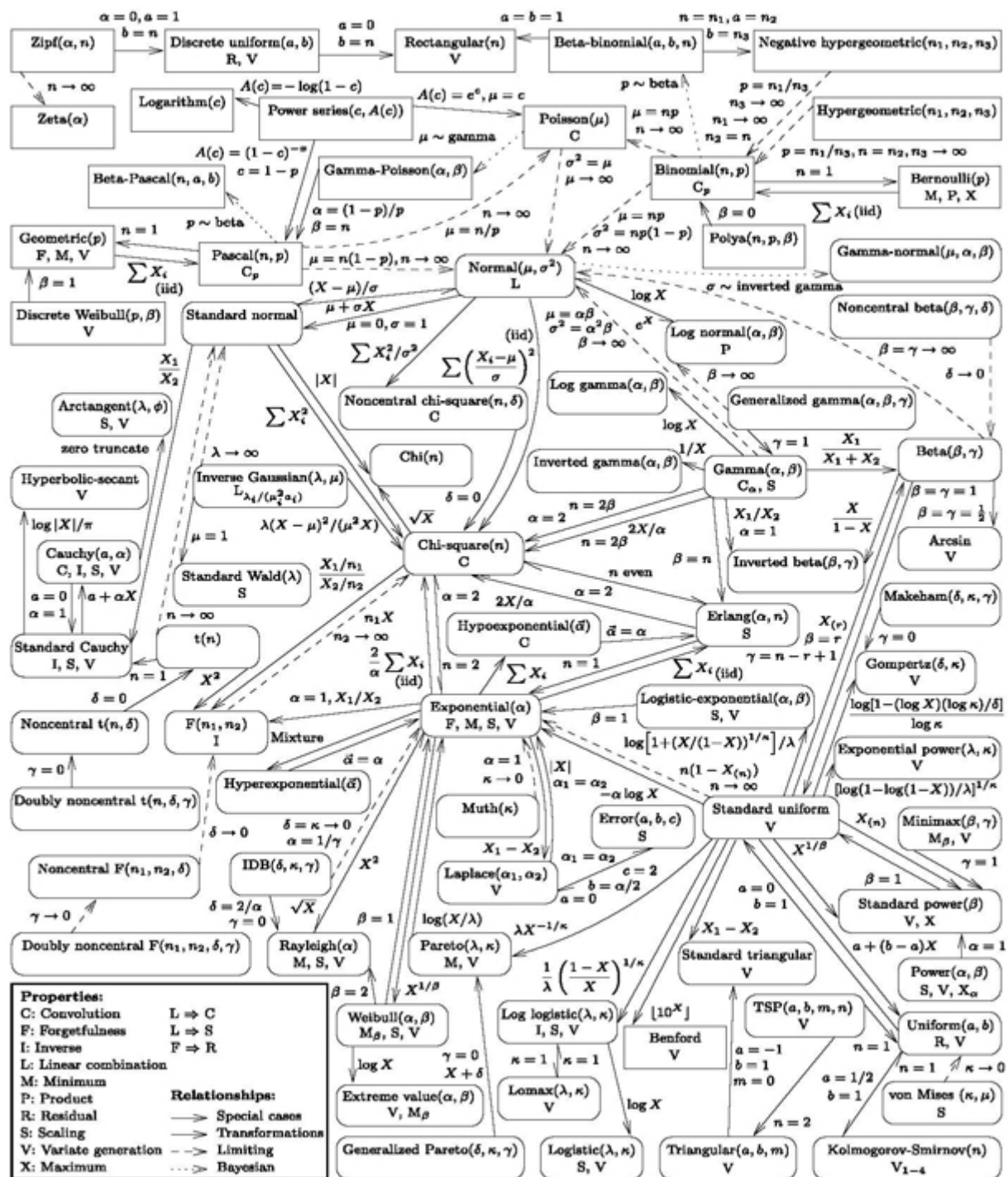


Figure: 7.34 - Relations entre distributions (Source: AMS Lawrence M. Leemis and Jacquelyn T. McQueston)

6. ESTIMATEURS DE VRAISEMBLANCE

Ce qui va suivre est d'une extrême importance en statistiques et est utilisé énormément en pratique. Il convient donc d'y accorder une attention toute particulière! Outre le fait que nous utiliserons cette technique dans la présent chapitre, nous la retrouverons dans le chapitre de Méthodes Numériques pour les techniques avancées de régressions linéaires généralisées ainsi que dans le chapitre de Génie Industriel dans le cadre de l'estimation des paramètres de fiabilité.

Nous supposons que nous disposons d'observations $x_1, x_2, x_3, \dots, x_n$ qui sont des réalisations de variables aléatoires non biaisées (dans le sens qu'elles sont choisies aléatoirement parmi un lot) indépendantes

$X_1, X_2, X_3, \dots, X_n$ de loi de probabilité inconnue mais identique.

Nous allons chercher à estimer cette loi de probabilité P inconnue à partir des observations

$x_1, x_2, x_3, \dots, x_n$.

Supposons que nous procédons par tâtonnements pour estimer la loi de probabilité P inconnue. Une manière de procéder est de se demander si les observations $x_1, x_2, x_3, \dots, x_n$ avaient une probabilité élevée ou non de sortir avec cette loi de probabilité arbitraire P .

Nous devons pour cela calculer la probabilité conjointe qu'avaient les observations $x_1, x_2, x_3, \dots, x_n$ de sortir avec $p_1, p_2, p_3, \dots, p_n$. Cette probabilité vaut (cf. chapitre de Probabilités):

$$\prod_{i=1}^n P(X_i = x_i) \quad (7.629)$$

en notant P la loi de probabilité supposée associée à $p_1, p_2, p_3, \dots, p_n$. Il faut avouer qu'il serait alors particulièrement maladroit de choisir une loi de probabilité (avec ses paramètres!) qui minimise cette quantité...

Au contraire, nous allons chercher la probabilité $p_1, p_2, p_3, \dots, p_n$ qui maximise $\prod_{i=1}^n P(X_i = x_i)$, c'est-à-dire qui rende les observations $x_1, x_2, x_3, \dots, x_n$ le plus vraisemblable possible.

Nous sommes donc amenés à chercher le (ou les) paramètre(s) θ qui maximise(nt) la quantité:

$$L_n(\theta) = \prod_{i=1}^n P_{\theta}(X_i = x_i) \quad (7.630)$$

Cette quantité L porte le nom de "**vraisemblance**". C'est une fonction du ou des paramètres θ et des observations $x_1, x_2, x_3, \dots, x_n$.

La ou les valeurs du paramètre θ qui maximisent la vraisemblance $L_n(\theta)$ sont appelées "**estimateurs du maximum de vraisemblance**" (estimateur MV).

Faisons quand même trois petits exemples (très classiques, utiles et importants dans l'industrie) avec dans l'ordre d'importance (donc pas forcément dans l'ordre de facilité...) la fonction de distribution de Gauss-Laplace (Normale), la fonction de distribution de Poisson et finalement Binomiale (et in extenso Géométrique).

Remarque: Ces trois exemples sont importants car utilisés dans les SPC (maîtrise statistiques de processus) dans différentes multinationales à travers le monde (cf. chapitre de Génie Industriel).

6.1. ESTIMATEURS DE LA LOI NORMALE

Soit $x_1, x_2, x_3, \dots, x_n$ un n -échantillon de variables aléatoires indépendamment distribuées supposées suivre une loi de Gauss-Laplace (loi Normale) de paramètres μ et σ^2 .

Nous recherchons quelles sont les valeurs des estimateurs du maximum de vraisemblance θ qui maximisent la vraisemblance $L_n(\theta)$ de la loi Normale?

Remarque: Il va de soi que les estimateurs du maximum de vraisemblance θ sont ici:

$$\theta = (\mu, \sigma^2) \quad (7.631)$$

Nous avons démontré plus haut que la densité d'une variable aléatoire gaussienne était donnée par:

$$P(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (7.632)$$

La vraisemblance est alors donnée par:

$$L(\mu, \sigma^2) = \prod_{i=1}^n P(x_i, \mu, \sigma) = \frac{1}{\sigma^n \sqrt{2\pi}^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2} \quad (7.633)$$

Maximiser une fonction ou maximiser son logarithme est équivalent donc la "**log-vraisemblance**" sera:

$$\ln [L(\mu, \sigma^2)] = -\frac{n}{2} \ln(2\pi) - n \ln(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \quad (7.634)$$

Pour déterminer les deux estimateurs de la loi Normale, fixons d'abord l'écart-type. Pour cela, dérivons $\ln [L(\mu, \sigma^2)]$ par rapport à μ et regardons pour quelle valeur de la moyenne la fonction s'annule.

Il nous reste après simplification le terme suivant qui est égal à zéro:

$$\sum_{i=1}^n (x_i - \mu) \quad (7.635)$$

Ainsi, l'estimateur du maximum de vraisemblance de la moyenne (espérance) de la loi Normale est donc après réarrangement:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad (7.636)$$

et nous voyons qu'il s'agit simplement de la moyenne arithmétique (ou appelée aussi "**moyenne empirique**").

Fixons maintenant la moyenne. L'annulation de la dérivée de $\ln [L(\mu, \sigma^2)]$ en σ^2 conduit à:

$$\frac{\partial}{\partial \sigma} \ln [L(\mu, \sigma^2)] = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2 = 0 \quad (7.637)$$

Ce qui nous permet d'écrire l'estimateur du maximum de vraisemblance pour l'écart-type (la variance lorsque la moyenne est connue selon la loi de distribution supposée elle aussi connue!):

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 \quad (7.638)$$

que certains appellent aussi "**écart-type de Pearson**"...

Cependant, nous n'avons pas encore défini ce qu'était un bon estimateur ! Ce que nous entendons par là:

- Si l'espérance d'un estimateur est égale à elle-même, nous disons que cet estimateur est "**sans biais**" et c'est bien évidemment ce que nous cherchons!
- Si l'espérance d'un estimateur n'est pas égale à elle-même, nous disons alors que cet estimateur est "**biaisé**" et c'est forcément moins bien...

Dans l'exemple précédent, la moyenne est donc non biaisée (trivial car la moyenne de la moyenne arithmétique est égale à elle-même). Mais qu'en est-il de la variance (in extenso de l'écart-type) ?

Un petit calcul simple par linéarité de l'espérance (puisque les variables aléatoires sont identiquement distribuées) va nous donner la réponse dans le cas où la moyenne théorique est approchée comme dans la pratique (industrie) par l'estimateur de la moyenne (cas le plus fréquent).

Nous avons donc le calcul de l'espérance de la "**variance empirique**":

$$\begin{aligned} E(\hat{\sigma}^2) &= E(V(X)) = E\left(\frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2\right) = E\left(\frac{1}{n} \sum_{i=1}^n (x_i^2 - 2x_i \hat{\mu} + \hat{\mu}^2)\right) \\ &= E\left(\frac{1}{n} \sum_{i=1}^n x_i^2 - 2\hat{\mu} \frac{1}{n} \sum_{i=1}^n x_i + \hat{\mu}^2\right) = E\left(\frac{1}{n} \sum_{i=1}^n x_i^2 - 2\hat{\mu}^2 + \hat{\mu}^2\right) = E\left(\frac{1}{n} \sum_{i=1}^n x_i^2 - \hat{\mu}^2\right) \quad (7.639) \\ &= \frac{1}{n} \sum_{i=1}^n E(x_i^2) - E(\hat{\mu}^2) \end{aligned}$$

Or, comme les variables sont équidistribuées:

$$E(\hat{\sigma}^2) = \frac{1}{n} \sum_{i=1}^n E(x_i^2) - E(\hat{\mu}^2) = E(x^2) - E(\hat{\mu}^2) \quad (7.640)$$

Et nous avons (relation de Huyghens):

$$V(X) = E(X^2) - E(X)^2 \quad (7.641)$$

ainsi que:

$$V(\hat{\mu}) = E(\hat{\mu}^2) - E(\hat{\mu})^2 = E(\hat{\mu}^2) - E(X)^2 \quad (7.642)$$

où la deuxième relation ne peut s'écrire que parce que nous utilisons l'estimateur du maximum de

vraisemblance de la moyenne (moyenne empirique). D'où:

$$E(\hat{\sigma}^2) = E(X^2) - E(\hat{\mu}^2) = (V(X) + E(X)^2) - (V(\hat{\mu}) + E(X)^2) = V(X) - V(\hat{\mu}) \quad (7.643)$$

et comme:

$$V(X) = \sigma^2 \text{ et } V(\hat{\mu}) = \frac{\sigma^2}{n} \quad (7.644)$$

Nous avons finalement:

$$E(\hat{\sigma}^2) = \sigma^2 - \frac{\sigma^2}{n} = \left(1 - \frac{1}{n}\right) \sigma^2 = \frac{n-1}{n} \sigma^2 \quad (7.645)$$

nous avons donc un biais de moins une fois l'erreur-standard:

$$-\frac{\sigma^2}{n} \quad (7.646)$$

nous disons alors que cet estimateur à un biais négatif (il sous-estime la vraie valeur!).

Nous noterons également que l'estimateur tend vers un estimateur sans biais (E.S.B.) lorsque le nombre d'individus tend vers l'infini $n \rightarrow \infty$. Nous disons alors que nous avons un "**estimateur asymptotiquement non biaisé**" ou "**estimateur asymptotiquement débiaisé**".

Il est important de prendre note que nous avons démontré que la variance empirique tend vers la variance théorique quand n tend vers l'infini et ce... que les données suivent une loi Normale ou non!

Remarque: Un estimateur est aussi dit "**estimateur consistant**" s'il converge en probabilité, lorsque $n \rightarrow \infty$, vers la vraie valeur du paramètre.

De par les propriétés de l'espérance, nous avons alors:

$$E(E(\hat{\sigma}^2)) = E(\hat{\sigma}^2) = E\left(\frac{n-1}{n} \sigma^2\right) \quad (7.647)$$

il vient alors:

$$\sigma = \sqrt{\frac{n}{n-1} \hat{\sigma}^2} = \sqrt{\frac{n}{n-1} \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2} \quad (7.648)$$

que certains appellent aussi "**écart-type standard**"... (à ne pas confondre avec "l'erreur-standard" que nous verrons plus loin).

Nous avons donc finalement pour résumer les deux résultats importants suivants:

1. "**L'estimateur du maximum de vraisemblance biaisé**" ou appelé également "**écart-type empirique**" ou encore "**écart-type échantillonnal**" ou encore "**écart-type de Pearson**" ... et donc donné par:

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2} \quad (7.649)$$

lorsque $n \rightarrow \infty$. Nous retrouvons cet écart-type suivant les contextes (par tradition) noté de quatre différentes façons qui sont:

$$\sigma, S, S_{n-1} \quad (7.650)$$

et même parfois (mais c'est très malheureux car cela génère alors souvent de la confusion) σ ou S .

2. "L'estimateur du maximum de vraisemblance non biaisé" ou appelé également "écart-type standard":

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2} \quad (7.651)$$

Nous retrouvons cet écart-type suivant les contextes (par tradition) noté de quatre différentes façons qui sont:

$$\sigma, S, S_{n-1} \quad (7.652)$$

Nous retrouverons ces deux dernières souvent dans les tables et dans de nombreux logiciels et que nous utiliserons plus bas dans les développements des intervalles de confiance et des tests d'hypothèses!

Par exemple, dans la version française de Microsoft Excel 11.8346 l'estimateur biaisé est donné par la fonction ECARTYPEP() et le non biaisé par ECARTTYPE().

Au total, cela nous fait donc trois estimateurs pour la même quantité!! Comme dans l'écrasante majorité des cas de l'industrie la moyenne théorique n'est pas connue, nous utilisons le plus souvent les deux dernières relations encadrées ci-dessus. Maintenant, c'est là que c'est le plus vicieux: lorsque nous calculons le biais des deux estimateurs, le premier est biaisé, le second ne l'est pas. Donc nous aurions tendance à n'utiliser que le second. Que nenni! Car nous pourrions aussi parler de la variance et de la précision d'un estimateur, qui sont aussi des critères importants pour juger de la qualité d'un estimateur par rapport à un autre. Si nous faisons le calcul de la variance des deux estimateurs, alors le premier, qui est biaisé, a une variance plus petite que le second qui est sans biais! Tout ça pour dire que le critère du biais n'est pas (et de loin) le seul à étudier pour juger de la qualité d'un estimateur.

Enfin, il est important de se rappeler que le facteur -1 du dénominateur de l'estimateur du maximum de vraisemblance non biaisé provient du fait qu'il fallait corriger l'espérance de l'estimateur biaisé à la base minoré de une fois l'erreur-standard!

6.2. ESTIMATEURS DE LA LOI DE POISSON

En utilisant la même méthode que pour la loi Normale (Gauss-Laplace), nous allons donc rechercher les estimateurs du maximum de vraisemblance de la loi de Poisson qui rappelons-le, est définie par:

$$P_k = \frac{\mu^k}{k!} e^{-\mu} \quad (7.653)$$

Dès lors, la vraisemblance est donnée par:

$$L(\mu) = \prod_{i=1}^n P(x_i, \mu) = \frac{\mu^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-\mu n} \quad (7.654)$$

Maximiser une fonction ou maximiser son logarithme est équivalent donc:

$$\ln [L(\mu)] = \ln(\mu) \sum_{i=1}^n x_i - \sum_{i=1}^n \ln(x_i!) - \mu n \quad (7.655)$$

Nous cherchons maintenant à la maximiser:

$$\frac{\partial \ln [L(\mu)]}{\partial \mu} = \frac{1}{\mu} \sum_{i=1}^n x_i - n = 0 \quad (7.656)$$

et obtenons donc son unique estimateur du maximum de vraisemblance qui sera:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad (7.657)$$

Il est tout à fait normal de retrouver dans cet exemple didactique la moyenne empirique, car c'est le meilleur estimateur possible pour le paramètre de la loi de Poisson (qui représente aussi l'espérance d'une loi de Poisson).

Sachant que l'écart-type de la distribution particulière (voir plus haut) n'est que la racine carrée de la moyenne, nous avons alors pour l'écart-type du maximum de vraisemblance biaisé:

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i} \quad (7.658)$$

Remarque: Nous montrons de la même manière des résultats identiques pour la loi exponentielle très utilisée en maintenance préventive et fiabilité!

6.3. ESTIMATEUR DE LA LOI BINOMIALE (ET GÉOMÉTRIQUE)

En utilisant la même méthode que pour la loi Normale (Gauss-Laplace) et la loi de Poisson, nous allons donc rechercher l'estimateur du maximum de vraisemblance de la loi Binomiale qui rappelons-le, est définie par:

$$P(N, k) = C_k^N \cdot p^k \cdot (1-p)^{N-k} = \frac{N!}{k!(N-k)!} \cdot p^k \cdot (1-p)^{N-k} \quad (7.659)$$

Dès lors, la vraisemblance est donnée par:

$$L(p) = \prod_{i=1}^N P(x_i, p) = C_k^N p^k (1-p)^{N-k} \quad (7.660)$$

Il convient de se rappeler que le facteur qui suit le terme combinatoire exprime déjà les variables successives selon ce que nous avons vu lors de notre étude de la fonction de distribution de Bernoulli et de la fonction binomiale. D'où la disparition du produit dans la dernière égalité précédente.

Maximiser une fonction ou maximiser son logarithme est équivalent donc:

$$\ln[L(p)] = \ln\left(C_k^N\right) + k \ln(p) + (N - k) \ln(1 - p) \quad (7.661)$$

Nous cherchons maintenant à la maximiser:

$$\frac{\partial \ln[L(p)]}{\partial p} = \frac{k}{p} - \frac{N - k}{1 - p} = 0 \quad (7.662)$$

Le lecteur au remarque que le coefficient binomial a disparu. Dès lors, nous en déduisons immédiatement que l'estimateur de la loi binomiale sera le même que celui de la loi géométrique.

Ce qui donne:

$$k(1 - p) - p(N - k) = k - kp - pN + pk = k - pN = 0 \quad (7.663)$$

d'où nous tirons l'estimateur du maximum de vraisemblance biaisé qui sera:

$$\hat{p} = \frac{k}{N} \quad (7.664)$$

Ce résultat est assez intuitif si l'on considère l'exemple classique d'une pièce de monnaie qui a une chance sur deux de tomber sur une de ces faces. La probabilité p étant le nombre de fois k où une face donnée a été observée sur le nombre d'essais total (toutes faces confondues).

Remarque: Dans la pratique, il n'est pas aussi simple d'appliquer ces estimateurs! Il faut bien réfléchir lesquels sont les plus adaptés à une expérience donnée et idéalement calculer également l'erreur quadratique moyenne (erreur-standard) de chacun des estimateurs de la moyenne (comme nous l'avons déjà fait pour la moyenne empirique plus tôt). Bref c'est un long travail de réflexion.

6.4. ESTIMATEURS DE LA LOI DE WEIBULL

Nous avons vu dans le chapitre de Génie Industriel une étude très détaillée de la loi de Weibull à trois paramètres avec son écart-type et son espérance car nous avons précisé qu'elle était assez utilisée dans le domaine de l'ingénierie de la fiabilité.

Malheureusement les trois paramètres de cette loi nous sont en pratique inconnus. A l'aide des estimateurs nous pouvons cependant déterminer l'expression de deux des trois en supposant γ comme étant nul. Cela nous donne donc la loi de Weibull dite "loi de Weibull à deux paramètres" suivante:

$$F(x, \beta, \eta) = \frac{\beta}{\eta} \left(\frac{x}{\eta}\right)^{\beta-1} \exp\left(-\left(\frac{x}{\eta}\right)^{\beta}\right) \quad (7.665)$$

avec pour rappel $\beta > 0$ et $\eta > 0$.

Dès lors la vraisemblance est donnée par:

$$\begin{aligned} L(\beta, \eta) &= \prod_{i=1}^n P(x_i, \beta, \eta) = \prod_{i=1}^n \frac{\beta}{\eta} \left(\frac{x_i}{\eta} \right)^{\beta-1} \exp \left(- \left(\frac{x_i}{\eta} \right)^{\beta} \right) \\ &= \left(\frac{\beta}{\eta} \right)^n \exp \left(- \sum_{i=1}^n \left(\frac{x_i}{\eta} \right)^{\beta} \right) \prod_{i=1}^n \left(\frac{x_i}{\eta} \right)^{\beta-1} \end{aligned} \quad (7.666)$$

Maximiser une fonction ou maximiser son logarithme est équivalent donc:

$$\begin{aligned} \ln [L(\beta, \eta)] &= \ln \left[\left(\frac{\beta}{\eta} \right)^n \exp \left(- \sum_{i=1}^n \left(\frac{x_i}{\eta} \right)^{\beta} \right) \prod_{i=1}^n \left(\frac{x_i}{\eta} \right)^{\beta-1} \right] \\ &= n \ln \left(\frac{\beta}{\eta} \right) - \frac{1}{\eta^{\beta}} \sum_{i=1}^n x_i^{\beta} + \sum_{i=1}^n \ln \left[\left(\frac{x_i}{\eta} \right)^{\beta-1} \right] = n \ln \left(\frac{\beta}{\eta} \right) - \frac{1}{\eta^{\beta}} \sum_{i=1}^n x_i^{\beta} + (\beta-1) \sum_{i=1}^n \ln \left(\frac{x_i}{\eta} \right) \\ &= n \ln \left(\frac{\beta}{\eta} \right) - \eta^{-\beta} \sum_{i=1}^n x_i^{\beta} + (\beta-1) \sum_{i=1}^n \ln \left(\frac{x_i}{\eta} \right) \end{aligned} \quad (7.667)$$

Cherchons maintenant à maximiser cela en se rappelant que (cf. chapitre de Calcul Différentiel et Intégral):

$$\frac{d}{dx} a^x = a^x \ln(a) \text{ et } \frac{d}{dx} a^{-x} = -a^x \ln(a) \quad (7.668)$$

d'où:

$$\begin{aligned} \frac{\partial \ln [L(\beta, \eta)]}{\partial \beta} &= n \frac{1}{\beta} + \frac{1}{\eta^{\beta}} \ln(\eta) \sum_{i=1}^n x_i^{\beta} - \frac{1}{\eta^{\beta}} \sum_{i=1}^n x_i^{\beta} \ln(x_i) + \sum_{i=1}^n \ln \left(\frac{x_i}{\eta} \right) \\ &= n \frac{1}{\beta} + \frac{1}{\eta^{\beta}} \left[\sum_{i=1}^n x_i^{\beta} (\ln(\eta) - \ln(x_i)) \right] + \sum_{i=1}^n \ln \left(\frac{x_i}{\eta} \right) \\ &= n \frac{1}{\beta} - \frac{1}{\eta^{\beta}} \left[\sum_{i=1}^n x_i^{\beta} (\ln(x_i) - \ln(\eta)) \right] + \sum_{i=1}^n \ln \left(\frac{x_i}{\eta} \right) \\ &= n \frac{1}{\beta} - \frac{1}{\eta^{\beta}} \sum_{i=1}^n x_i^{\beta} \ln \left(\frac{x_i}{\eta} \right) + \sum_{i=1}^n \ln \left(\frac{x_i}{\eta} \right) = 0 \end{aligned} \quad (7.669)$$

Et nous avons pour le deuxième paramètre:

$$\frac{\partial \ln [L(\beta, \eta)]}{\partial \eta} = -n \frac{1}{\eta} - \frac{\beta}{\eta^{\beta} \eta} \sum_{i=1}^n x_i^{\beta} + (1-\beta) n \frac{1}{\eta} = 0 \quad (7.670)$$

d'où:

$$- \frac{1}{\eta^{\beta}} \sum_{i=1}^n x_i^{\beta} - n = 0 \quad (7.671)$$

Finalement avec les écritures correctes (et dans l'ordre de résolution dans la pratique):

$$n \frac{1}{\hat{\beta}} - \frac{1}{\hat{\eta}^{\hat{\beta}}} \sum_{i=1}^n x_i^{\hat{\beta}} \ln \left(\frac{x_i}{\hat{\eta}} \right) + \sum_{i=1}^n \ln \left(\frac{x_i}{\hat{\eta}} \right) = 0 \text{ et } -\frac{1}{\hat{\eta}^{\hat{\beta}}} \sum_{i=1}^n x_i^{\hat{\beta}} - n = 0 \quad (7.672)$$

La résolution de ces équations implique de lourds calculs et on ne peut rien en tirer dans les tableaux classiques comme Microsoft Excel ou Calc de Open Office.

On prend alors une approche différente en écrivant notre loi de Weibull à deux paramètres ainsi:

$$P(x, \beta, \theta) = \frac{\beta}{\theta} x^{\beta-1} \exp \left(-\frac{x^{\beta}}{\theta} \right) \quad (7.673)$$

avec pour rappel $\beta > 0$ et $\theta > 0$.

Dès lors la vraisemblance est donnée par:

$$L(\beta, \eta) = \prod_{i=1}^n P(x_i, \beta, \theta) = \prod_{i=1}^n \frac{\beta}{\theta} x_i^{\beta-1} \exp \left(-\frac{x_i^{\beta}}{\theta} \right) = \left(\frac{\beta}{\theta} \right)^n \exp \left(-\sum_{i=1}^n \frac{x_i^{\beta}}{\theta} \right) \prod_{i=1}^n x_i^{\beta-1} \quad (7.674)$$

Maximiser une fonction ou maximiser son logarithme est équivalent donc:

$$\begin{aligned} \ln [L(\beta, \theta)] &= \ln \left[\left(\frac{\beta}{\theta} \right)^n \exp \left(-\sum_{i=1}^n \frac{x_i^{\beta}}{\theta} \right) \prod_{i=1}^n x_i^{\beta-1} \right] \\ &= n \ln \left(\frac{\beta}{\theta} \right) - \frac{1}{\theta} \sum_{i=1}^n x_i^{\beta} + \sum_{i=1}^n \ln (x_i^{\beta-1}) = n \ln \left(\frac{\beta}{\theta} \right) - \frac{1}{\theta} \sum_{i=1}^n x_i^{\beta} + (\beta-1) \sum_{i=1}^n \ln (x_i) \end{aligned} \quad (7.675)$$

Cherchons maintenant à maximiser cela en se rappelant que (cf. chapitre de Calcul Différentiel et Intégral):

$$\frac{d}{dx} a^x = a^x \ln(a) \text{ et } \frac{d}{dx} a^{-x} = -a^x \ln(a) \quad (7.676)$$

d'où:

$$\frac{\partial \ln [L(\beta, \theta)]}{\partial \beta} = \frac{n}{\beta} - \frac{1}{\theta} \sum_{i=1}^n x_i^{\beta} \ln (x_i) + \sum_{i=1}^n \ln (x_i) = 0 \quad (7.677)$$

Et nous avons pour le deuxième paramètre:

$$\frac{\partial \ln [L(\beta, \theta)]}{\partial \theta} = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n x_i^{\beta} = 0 \quad (7.678)$$

Il est alors immédiat que:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i^{\beta} \quad (7.679)$$

injecté dans la relation:

$$\frac{n}{\beta} - \frac{1}{\theta} \sum_{i=1}^n x_i^{\beta} \ln(x_i) + \sum_{i=1}^n \ln(x_i) = 0 \quad (7.680)$$

Il vient:

$$\frac{n}{\beta} - \frac{\sum_{i=1}^n x_i^{\beta} \ln(x_i)}{\frac{1}{n} \sum_{i=1}^n x_i^{\beta}} + \sum_{i=1}^n \ln(x_i) = 0 \quad (7.681)$$

en simplifiant:

$$\frac{\sum_{i=1}^n x_i^{\hat{\beta}} \ln(x_i)}{\sum_{i=1}^n x_i^{\hat{\beta}}} - \frac{1}{\hat{\beta}} = \frac{1}{n} \sum_{i=1}^n \ln(x_i) \quad (7.682)$$

La résolution des deux équations (dans l'ordre de haut en bas):

$$\boxed{\begin{aligned} \frac{\sum_{i=1}^n x_i^{\hat{\beta}} \ln(x_i)}{\sum_{i=1}^n x_i^{\hat{\beta}}} - \frac{1}{\hat{\beta}} - \frac{1}{n} \sum_{i=1}^n \ln(x_i) &= 0 \\ \frac{1}{n} \sum_{i=1}^n x_i^{\hat{\beta}} - \hat{\theta} &= 0 \end{aligned}} \quad (7.683)$$

peut très facilement être calculée avec l'outil Valeur Cible de Microsoft Excel ou Calc de Open Office.

6.5. ESTIMATEURS DE LA LOI GAMMA

Nous allons utiliser ici une technique appelée "**méthode des moments**" pour déterminer les estimateurs des paramètres de la loi Gamma.

Supposons que $X_1, \dots, X_n$ sont des variables aléatoires indépendantes et identiquement distribuées selon la loi Gamma avec pour densité:

$$P_{\alpha, \lambda}(x) = \frac{x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} \cdot \lambda^{\alpha} \cdot 1_{]0, +\infty[} \quad (7.684)$$

Nous cherchons à estimer α, λ . Pour cela, nous déterminons d'abord quelques moments théoriques.

Le premier moment est l'espérance qui comme nous l'avons démontré vaut:

$$E(X) = m_1 = \frac{\alpha}{\lambda} \quad (7.685)$$

et le second moment, l'espérance du carré de la variable, est comme nous l'avons démontré implicitement lors de la démonstration de la variance de la loi Gamma:

$$E(X^2) = m_2 = \frac{\alpha(\alpha+1)}{\lambda^2} \quad (7.686)$$

Nous exprimons ensuite la relation entre les paramètres et les moments théoriques:

$$\begin{aligned} m_1 &= \frac{\alpha}{\lambda} \\ m_2 &= \frac{\alpha(\alpha+1)}{\lambda^2} \end{aligned} \quad (7.687)$$

La résolution donne:

$$\begin{aligned} \alpha &= \frac{m_1^2}{m_2 - m_1^2} \\ \lambda &= \frac{m_2 - m_1^2}{m_1^2} \end{aligned} \quad (7.688)$$

Une fois ce système établie, la méthode des moments consiste à utiliser les moments empiriques, en l'occurrence pour notre exemple les deux premiers, $\hat{m}_1, \hat{m}_2$:

$$\begin{aligned} \hat{m}_1 &= \frac{X_1 + \dots + X_n}{n} \\ \hat{m}_2 &= \frac{X_1^2 + \dots + X_n^2}{n} \end{aligned} \quad (7.689)$$

que l'on pose égaux aux moments théoriques vrais... Dès lors, il vient:

$$\begin{aligned} \alpha &= \frac{\hat{m}_1^2}{\hat{m}_2 - \hat{m}_1^2} \\ \lambda &= \frac{\hat{m}_2 - \hat{m}_1^2}{\hat{m}_1^2} \end{aligned} \quad (7.690)$$

7. FACTEUR DE CORRECTION SUR POPULATION FINIE

Maintenant démontrons un autre résultat qui nous sera indispensables dans certains tests statistiques que nous verrons plus loin.

Supposons que nous avons une population de N individus que nous représentons par l'ensemble $\{1, 2, \dots, N\}$ et une variable aléatoire X qui est donc une application de $\{1, 2, \dots, N\}$ dans $\mathbb{R}$. Nous posons $x_i = X(i)$. La moyenne de X est alors donnée par:

$$\mu = \frac{1}{N} \sum x_i \quad (7.691)$$

La variance de X est par définition:

$$\sigma^2 = \frac{1}{N} \sum (x_i^2 - \mu^2) \quad (7.692)$$

Considérons à présent l'ensemble E des échantillons $(i_1, \dots, i_n)$ de taille n pris dans $\{1, 2, \dots, N\}$ avec $0 < n < N$. Chaque individu a une probabilité d'être tiré égale à:

$$\frac{1}{N} \cdot \frac{1}{N-1} \cdots \frac{1}{N-(n-1)} = \frac{(N-n)!}{N!} \quad (7.693)$$

Nous nous intéressons à la variable aléatoire $\bar{X}$ définie sur E et étant égale à la moyenne de l'échantillon. Plus précisément:

$$\bar{X}(i_1, \dots, i_n) = \frac{1}{n} \sum_{k=1}^n x_{i_k} \quad (7.694)$$

Afin de calculer la variance $V(\bar{X})$, nous allons exprimer $\bar{X}$ comme somme de variables aléatoires. En effet si nous définissons les variables X_k avec $k = 1 \dots N$ par:

$$X_k(i_1, \dots, i_n) = \begin{cases} x_k & \text{si } k \in \{i_1, \dots, i_n\} \\ 0 & \text{sinon} \end{cases} \quad (7.695)$$

Nous avons naturellement (donc de la par la définition précédente):

$$\bar{X} = \frac{1}{n} \sum_{k=1}^N X_k \quad (7.696)$$

et donc il vient:

$$V(\bar{X}) = \frac{1}{n^2} \left(\sum V(X_k) + \sum_{i \neq j} \text{cov}(X_i, X_j) \right) \quad (7.697)$$

Les variables aléatoires X_k ne sont pas indépendantes deux à deux, en effet comme nous allons le voir, leurs covariances ne sont pas nulles si N est fini. Dans le cas contraire, (covariance nulle), nous retrouvons un résultat déjà démontré plus haut:

$$V(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \quad (7.698)$$

Il nous faut donc calculer les variances $V(X_k)$ et les covariances $\text{cov}(X_i, X_j)$.

Pour ce faire nous allons utiliser la relation de Huyghens et pour cela nous allons commencer par calculer l'espérance $E(X_k)$:

$$E(X_k) = P(X_k = x_k) \cdot x_k \quad (7.699)$$

Or $P(X_k = x_k)$ est la probabilité qu'un échantillon contienne k . Cette probabilité vaut bien évidemment n/N et par suite:

$$E(X_k) = \frac{n}{N} \cdot x_k \quad (7.700)$$

De la même façon nous obtenons

$$E(X_k^2) = \frac{n}{N} \cdot x_k^2 \quad (7.701)$$

Nous pouvons donc calculer la variance $V(X_k)$:

$$V(X_k) = E(X_k^2) - E(X_k)^2 = \frac{n}{N} \cdot x_k^2 - \left(\frac{n}{N} \cdot x_k \right)^2 = \frac{n(N-n)}{N^2} \cdot x_k^2 \quad (7.702)$$

Pour calculer les covariances avons à présent besoin de calculer les espérances $E(X_i \cdot X_j)$:

$$E(X_i \cdot X_j) = P(X_i = x_i, X_j = x_j) \cdot x_i x_j \quad (7.703)$$

Or $P(X_i = x_i, X_j = x_j)$ est la probabilité qu'un échantillon contienne i et j . Cette probabilité vaut bien évidemment:

$$\frac{n}{N} \cdot \frac{n-1}{N-1} \quad (7.704)$$

et par suite:

$$E(X_i \cdot X_j) = \frac{n(n-1)}{N(N-1)} \cdot x_i x_j \quad (7.705)$$

Nous pouvons à présent calculer les covariances:

$$\begin{aligned} cov(X_i, X_j) &= E(X_i \cdot X_j) - E(X_i)E(X_j) = \frac{n(n-1)}{N(N-1)} \cdot x_i x_j - \frac{n^2}{N^2} \cdot x_i x_j \\ &= -\frac{n(N-n)}{N^2(N-1)} \cdot x_i x_j \end{aligned} \quad (7.706)$$

Nous sommes maintenant en mesure de calculer $V(\bar{X})$:

$$V(\bar{X}) = \frac{1}{n^2} \left(\sum V(X_k) + \sum_{i \neq j} Cov(X_i, X_j) \right) = \frac{1}{n^2} \left(\frac{n(N-n)}{N^2} \sum x_k^2 - \frac{n(N-n)}{N^2(N-1)} \sum_{i \neq j} x_i x_j \right) \quad (7.707)$$

Sachant que (réarrangement de la définition de la variance):

$$\sum x_i^2 = N \cdot (\sigma^2 + \mu^2) \quad (7.708)$$

Et que (réarrangement de la définition de la moyenne):

$$\sum x_i = N\mu \quad (7.709)$$

Nous avons:

$$\begin{aligned} V(\bar{X}) &= \frac{1}{n^2} \left(\frac{n(N-n)}{N} (\sigma^2 + \mu^2) - \frac{n(N-n)}{N^2(N-1)} \sum_i x_i (N\mu - x_i) \right) \\ &= \frac{1}{n^2} \left(\frac{n(N-n)}{N} (\sigma^2 + \mu^2) - \frac{n(N-n)}{N^2(N-1)} N^2 \mu^2 + \frac{n(N-n)}{N^2(N-1)} \sum_i x_i^2 \right) \quad (7.710) \\ &= \frac{1}{n^2} \left(\frac{n(N-n)}{N} (\sigma^2 + \mu^2) - \frac{n(N-n)}{(N-1)} \mu^2 + \frac{n(N-n)}{N(N-1)} (\sigma^2 + \mu^2) \right) \\ &= \dots = \sigma^2 \frac{(N-n)}{n(N-1)} \end{aligned}$$

Et donc:

$$\sqrt{V(\bar{X})} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad (7.711)$$

Le terme:

$$fcp = \sqrt{\frac{N-n}{N-1}} \leq 1 \quad (7.712)$$

que nous avons déjà rencontré lors de notre étude la loi hypergéométrique, et appelé "**facteur de correction sur population finie**" et il a pour effet de réduire l'erreur-standard d'autant plus que n est grand.

8. INTERVALLES DE CONFIANCE

Jusqu'à maintenant nous avons toujours déterminé les différents estimateurs de vraisemblance ou estimateurs simples (variance, écart-type) à partir de lois (fonctions) statistiques théoriques ou mesurées sur toute une population de données.

Définition: Un "**intervalle de confiance**" est un couple de nombres qui définit une plage de valeurs possibles avec une certaine probabilité pour un estimateur statistique donné (plage calculée à l'aide de paramètres vrais mesurés). Il s'agit du cas le plus fréquent en statistiques.

Nous allons maintenant aborder une approche un peu différente et importante dans l'industrie en se demandant maintenant quelles doivent être les tailles d'échantillons pour avoir une certaine validité (intervalle de confiance I.C.) pour les données mesurées ou encore quel écart-type ou fractile dans une

loi Normale centrée réduite (grand nombre d'individus), du Khi-deux, de Student ou de Fisher correspond à un certain intervalle de confiance (nous verrons ces deux derniers cas de faibles échantillons dans la partie traitant de l'analyse de la variance ou ANOVA) lorsque la variance ou la moyenne est connue ou respectivement inconnue sur l'ensemble ou une partie de la population donnée.

Indiquons que ces intervalles de confiance utilisent le théorème central limite démontré plus loin (afin d'éviter toute frustration) et que les développements que nous allons faire maintenant nous seront également utiles dans le domaine des Tests d'Hypothèses qui ont une place majeure en statistique!

Enfin, indiquons que de très nombreuses organisations (privées ou étatiques) font des statistiques fausses car les hypothèses et conditions d'utilisation ne sont pas rigoureusement vérifiées ou simplement omises ou pire encore, toute la base (les mesures) ne sont pas collectées dans les règles de l'art (fiabilisation de la collecte de données).

Remarque: Le praticien doit être très prudent quant à au calcul des intervalles de confiance et à l'utilisation des tests d'hypothèses dans la pratique. Raison pour laquelle, afin d'éviter toute erreur triviale d'utilisation ou d'interprétation, il est important de se référer aux normes suivantes par exemple: ISO 2602:1980 (Interprétation statistique de résultats d'essais - Estimation de la moyenne - Intervalle de confiance), ISO 2854:1976 (Interprétation statistique des données - Techniques d'estimation et tests portant sur des moyennes et des variances), ISO 3301:1975 (Interprétation statistique des données - Comparaison de deux moyennes dans le cas d'observations appariées), ISO 3494:1976 (Interprétation statistique des données -- Efficacité des tests portant sur des moyennes et des variances), ISO 5479:1997 (Interprétation statistique des données - Tests pour les écarts à la distribution normale), ISO 10725:2000 + ISO 11648-1:2003 + ISO 11648-2:2001 (Plans et procédures d'échantillonnage pour acceptation pour le contrôle de matériaux en vrac), ISO 11453:1996 (Interprétation statistique des données - Tests et intervalles de confiance portant sur les proportions), ISO 16269-4:2010 (Interprétation statistique des données Détection et traitement des valeurs aberrantes), ISO 16269-6:2005 (Interprétation statistique des données - Détermination des intervalles statistiques de tolérance), ISO 16269-8:2004 (Interprétation statistique des données - Détermination des intervalles de prédiction), ISO/TR 18532:2009 (Lignes directrices pour l'application des méthodes statistiques à la qualité et à la normalisation industrielle).

8.1. I.C. SUR LA MOYENNE AVEC VARIANCE THÉORIQUE CONNUE

Commençons par le cas le plus simple et le plus courant qui est la détermination du nombre d'individus pour avoir une certaine confiance dans la moyenne des mesures effectuées d'une variable aléatoire supposée suivre une loi Normale.

D'abord rappelons que nous avons démontré au début de ce chapitre que l'erreur-type (écart-type à la moyenne) était sous l'hypothèses de variables indépendantes et indentiquement distribuées:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (7.713)$$

Maintenant, avant d'aller plus loin, considérons X comme une variable aléatoire suivant une loi Normale de moyenne μ et d'écart-type σ . Nous souhaiterions que la variable aléatoire ait par exemple 95% de probabilité cumulée de se trouver dans un intervalle symétrique borné donné. Ce qui s'exprime donc sous la forme suivante:

$$P(\mu - \delta \leq X \leq \mu + \delta) = 0.95 \quad (7.714)$$

Remarque: Donc avec un intervalle de confiance de 95% vous aurez raison 19 fois sur 20, ou n'importe quel autre niveau de confiance ou niveau de risque α (1-niveau de confiance, soit 5%) que vous vous serez fixé à l'avance. En moyenne, vos conclusions seront donc bonnes, mais nous ne pourrons jamais savoir si une décision particulière est bonne! Si le niveau de risque est très faible mais que l'événement a quand même lieu, les spécialistes parlent alors de "**grande déviation**" ou de "**black swan**" (cygne noir). La gestion des valeurs aberrantes est traitée dans la norme ISO 16269-4:2010 *Détection et traitement des valeurs aberrantes* que tout ingénieur faisant des statistiques en entreprise se doit de respecter.

En centrant et réduisant la variable aléatoire:

$$P\left(-\frac{\delta}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{\delta}{\sigma}\right) = 0.95 \quad (7.715)$$

Notons maintenant Y la variable centrée réduite:

$$P\left(-\frac{\delta}{\sigma} \leq Y\right) + \left(1 - P\left(Y \geq \frac{\delta}{\sigma}\right)\right) = 0.95 \quad (7.716)$$

Puisque la loi Normale centrée réduite est symétrique:

$$1 - 2P\left(Y \geq \frac{\delta}{\sigma}\right) = 0.95 \quad (7.717)$$

D'où:

$$P\left(Y \geq \frac{\delta}{\sigma}\right) = 0.025 \quad (7.718)$$

A partir de là en lisant dans les tables numériques de la loi Normale centrée réduite, nous avons pour satisfaire cette égalité que:

$$\frac{\delta}{\sigma} \cong 1.96 \quad (7.719)$$

Ce qui s'obtient facilement avec la version anglaise de Microsoft Excel 11.8346 en utilisant la fonction: -NORMSINV((1-0.95)/2).

Donc:

$$\delta \cong 1.96\sigma \quad (7.720)$$

Ce qui est noté de façon traditionnelle dans le cas général autre que 95% par (Z étant la variable aléatoire correspond donc à la loi Normale centrée réduite):

$$\delta = Z\sigma \quad (7.721)$$

Or, considérons que la variable X sur laquelle nous souhaitons faire de l'inférence statistique est justement la moyenne (et nous démontrerons plus loin que celle-ci suit une loi Normale centrée réduite). Dès lors:

$$\delta = Z \sigma_{\bar{X}} = Z \frac{\sigma}{\sqrt{n}} \quad (7.722)$$

nous en tirons:

$$n = \frac{Z^2 \sigma^2}{\delta^2} \quad (7.723)$$

dont nous prenons évidemment (normalement...) la valeur entière supérieure...

Cette dernière notation est plus souvent écrite sous la forme suivante mettant mieux en évidence la largeur de l'intervalle de confiance à un niveau α sous-jacent:

$$n = \left(\frac{Z_{\alpha/2} \sigma}{\delta} \right)^2 \quad (7.724)$$

Relation appelée "**effectif de l'échantillon pour estimation par loi Normale**".

Ainsi, nous pouvons maintenant savoir le nombre d'individus à avoir pour s'assurer un intervalle de précision δ (marge d'erreur) autour de la moyenne et pour qu'un pourcentage donné des mesures se trouvent dans cet intervalle et en supposant l'écart-type théorique σ connu (ou imposé) d'avance (typiquement utilisé dans l'ingénierie de la qualité ou les instituts de sondages).

Autrement dit, nous pouvons calculer le nombre n d'individus à mesurer pour s'assurer un intervalle de confiance donné (associé à Z) de la moyenne mesurée en supposant l'écart-type théorique connu (ou imposé) et en souhaitant un précision de δ en valeur absolue sur la moyenne.

Cependant... en réalité, la variable Z provient du théorème central limite (voir plus bas) qui donne pour un échantillon de grande taille (approximativement):

$$Z = Z_n = \frac{M_n - \mu}{\sigma / \sqrt{n}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \quad (7.725)$$

En réarrangeant nous obtenons:

$$\mu = \bar{X} - Z \frac{\sigma}{\sqrt{n}} \quad (7.726)$$

et comme Z peut être négatif ou positif alors il est plus censé d'écrire cela sous la forme:

$$\mu = \bar{X} \pm Z \frac{\sigma}{\sqrt{n}} \quad (7.727)$$

Soit:

$$\bar{X} - Z \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z \frac{\sigma}{\sqrt{n}} \quad (7.728)$$

que les ingénieurs notent parfois:

$$LCL \leq \mu \leq UCL \quad (7.729)$$

avec LCL étant la lower confidence limit et UCL la upper confidence limit. C'est de la terminologie Six Sigma (cf. chapitre de Génie Industriel).

Et nous venons de voir plus avant que pour avoir un intervalle de confiance à 95% nous devrions avoir $Z=1.96$. Et puisque la loi Normale est symétrique:

$$95\% = 1 - 5\% = 1 - \alpha = 1 - 2 \cdot 2.5\% = 1 - 2 \cdot 0.025 \quad (7.730)$$

Cela se note finalement:

$$\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (7.731)$$

Comme nous l'avons déjà mentionné, et nous le démontrerons un peu plus loin, la moyenne arithmétique centrée réduite d'une série de variables aléatoires indépendantes et identiquement distribuées de variance finie suit asymptotiquement une loi Normale centrée réduite, alors l'intervalle de confiance ci-dessus a une portée très générale! Raison pour laquelle nous parlons parfois de "**d'intervalle de confiance asymptotique de la moyenne**".

Ces intervalles ont évidemment pour origine que nous travaillons très souvent en statistiques sur des échantillons et non sur toute la population disponible. L'échantillonnage choisi influe donc sur l'estimateur ponctuel. Nous parlons alors de "**fluctuation d'échantillonnage**".

Dans le cas particulier d'un I.C. (intervalle de confiance) à 95%, la dernière relation s'écrit:

$$\bar{X} - Z_{0.025} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{0.025} \frac{\sigma}{\sqrt{n}} \quad (7.732)$$

Parfois nous retrouvons l'inégalité antéprécédente sous la forme équivalente suivante:

$$\bar{X} - Z_{\alpha/2} \sigma_{\bar{X}} \leq \mu_{\bar{X}} \leq \bar{X} + Z_{\alpha/2} \sigma_{\bar{X}} \quad (7.733)$$

ou encore plus rarement sous la forme générale suivante (que l'on retrouve pour toutes les intervalles):

$$\bar{X} - ME \leq \mu \leq \bar{X} + ME \quad (7.734)$$

où ME signifie "**marge d'erreur**".

Nous sommes ainsi capables maintenant d'estimer des tailles de population nécessaires à obtenir un certain niveau de confiance α dans un résultat, soit d'estimer dans quel intervalle de confiance se trouve la moyenne théorique en connaissant la moyenne expérimentale (empirique) et l'estimateur du maximum de vraisemblance de l'écart-type. Nous pouvons bien évidemment dès lors aussi déterminer la

probabilité avec laquelle la moyenne est en dehors d'un certain intervalle... (l'un comme l'autre étant beaucoup utilisés dans l'industrie).

Enfin, signalons que du résultat précédent, nous déduisons immédiatement par la propriété de stabilité de la loi Normale (démontrée plus haut) le test suivant que nous retrouvons dans de très nombreux logiciels de statistiques:

$$Z^{calc.} = \frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (7.735)$$

appelé "test Z bilatéral sur la différence de deux moyennes" avec l'intervalle de confiance correspondant:

$$(\bar{X}_2 - \bar{X}_1) - Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_2 - \mu_1 \leq (\bar{X}_2 - \bar{X}_1) + Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (7.736)$$

Remarque: La taille de la population mère pour les relations développées plus haut n'entre pas en ligne de compte dans le calcul des intervalles de confiance ni dans celui de la taille de l'échantillon, et pour cause, elle est considérée infinie. Il faut donc faire attention à ne pas avoir parfois des tailles d'échantillons qui sont plus grandes que la population mère réelle possible...

8.2. I.C. SUR LA VARIANCE AVEC MOYENNE THÉORIQUE CONNUE

Commençons par démontrer une propriété fondamentale de la loi du Khi-deux:

Si une variable aléatoire X suit une loi Normale centrée réduite $X = N(0,1)$ alors son carré suit une loi du Khi-deux de degré de liberté 1:

$$X^2 = \chi^2(1) \quad (7.737)$$

Démonstration:

Pour démontrer cette propriété, il suffit de calculer la densité de la variable aléatoire X^2 avec $X = N(0,1)$. Or, si $X = N(0,1)$ et si nous posons $Y = X^2$, alors pour tout $y \geq 0$ nous obtenons:

$$P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq +\sqrt{y}) \quad (7.738)$$

Puisque la loi Normale centrée réduite est symétrique par rapport à 0 pour la variable aléatoire X , nous pouvons écrire:

$$P(Y \leq y) = 2 \cdot P(0 \leq X \leq \sqrt{y}) \quad (7.739)$$

En notant Φ la fonction de répartition de la loi Normale centrée réduite (sa probabilité cumulée en d'autres termes pour rappel...), nous avons:

$$P(Y \leq y) = 2 \cdot [\Phi(\sqrt{y}) - \Phi(0)] \quad (7.740)$$

et comme:

$$\Phi(0) = P(0, 0, 1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 e^{-k^2/2} dk = 0.5 \quad (7.741)$$

alors:

$$P(Y \leq y) = 2 \cdot \Phi(\sqrt{y}) - 2 \cdot 0.5 = 2 \cdot \Phi(\sqrt{y}) - 1 \quad (7.742)$$

La fonction de répartition de la variable aléatoire (probabilité cumulée) $Y = X^2$ est donc donnée par:

$$P(Y \leq y) = 2\Phi(\sqrt{y}) - 1 \quad (7.743)$$

si y est supérieur ou égal à zéro, nulle si y inférieur à zéro. Nous noterons cette répartition $f_Y(y)$ pour la suite des calculs.

Puisque la fonction de distribution est la dérivée de la fonction de répartition et que X suit une loi Normale centrée réduite alors nous avons pour la variable aléatoire X :

$$P(X = x) = \Phi'(x) = \frac{d}{dx} \Phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad (7.744)$$

et il s'ensuit pour la loi de distribution de Y (qui est donc le carré de X pour rappel!):

$$\begin{aligned} \frac{d}{dy} f_Y(y) &= \frac{d}{dy} (2\Phi(\sqrt{y}) - 1) = 2 \frac{d}{dy} \Phi(\sqrt{y}) = 2\Phi'(\sqrt{y}) \frac{d}{dy} \sqrt{y} \\ &= \frac{1}{\sqrt{y}} \Phi'(\sqrt{y}) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-y/2} = \frac{1}{\sqrt{2\pi y}} e^{-y/2} \end{aligned} \quad (7.745)$$

cette dernière expression correspond exactement à la relation que nous avons obtenue lors de notre étude de la loi du Khi-deux en imposant un degré de liberté unité.

Le théorème est donc bien démontré, à savoir que si X suit une loi Normale centrée réduite alors son carré suit une loi du Khi-deux à 1 degré de liberté tel que:

$$\boxed{(X_{N(0,1)})^2 = \chi^2(1)} \quad (7.746)$$

□C.Q.F.D.

Ce type de relation est utilisé dans les processus industriels et leur contrôle (cf. chapitre de Génie Industriel).

Nous allons maintenant utiliser un résultat démontré lors de notre étude de la loi Gamma. Nous avons effectivement vu plus haut que la somme de deux variables aléatoires suivant une loi Gamma suit aussi

une loi Gamma dont les paramètres s'additionnent:

$$X+Y = \gamma(p+q, \lambda) \quad (7.747)$$

Comme la loi du Khi-deux n'est qu'un cas particulier de la loi Gamma, le même résultat s'applique.

Pour être plus précis, cela revient à écrire:

Si $X_1, \dots, X_k$ sont des variables aléatoires indépendantes (!) et identiquement distribuées $N(0,1)$ alors par extension de la démonstration précédente où nous avons montré que:

$$\left(X_{N(0,1)}\right)^2 = \chi^2(1) \quad (7.748)$$

et de la propriété d'addition de la loi Gamma, la somme de leurs carrés suit alors une loi du Khi-deux de degré k telle que:

$$\chi^2(k) = X_1^2 + X_2^2 + \dots + X_k^2 \quad (7.749)$$

Ainsi, la loi du χ^2 à k degrés de liberté est la loi de probabilité de la somme des carrés de k variables normales centrées réduites linéairement indépendantes entre elles. Il s'agit de la propriété de linéarité de la loi du Khi-deux (implicitement de la linéarité de la loi Gamma)!

Maintenant voyons une autre propriété importante de la loi du Khi-deux: Si $X_1, \dots, X_n$ sont des variables aléatoires indépendantes et identiquement distribuées $N(\mu, \sigma^2)$ (donc de même moyenne et même écart-type et suivant une loi Normale) et si nous notons l'estimateur du maximum de vraisemblance de la variance:

$$S_*^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \quad (7.750)$$

alors, le rapport de la variable aléatoire S_*^2 sur l'écart-type supposé connu de l'ensemble de la population (dit "écart-type vrai" ou "écart-type théorique" pour bien différencier!) multiplié par le nombre d'individus n de la population suit une loi du Khi-deux de degré n telle que:

$$\chi^2(n) = \chi_n^2 = \frac{n \cdot S_*^2}{\sigma^2} \quad (7.751)$$

Ce résultat est appelé "**théorème de Cochran**" ou encore "**théorème de Fisher-Cochran**" (dans le cas particulier d'échantillons gaussiens) et nous donne donc une distribution pour les écarts-types empiriques (dont la loi parente est une loi Normale).

En utilisant la valeur de l'écart-type démontrée lors de notre étude de la loi du khi-deux nous avons donc:

$$V\left(\frac{nS_*^2}{\sigma^2}\right) = V\left(\chi^2(n)\right) = 2n \quad (7.752)$$

Mais n et σ sont imposés et sont donc considérés comme des constantes. Il vient alors:

$$\frac{n^2}{\sigma^4} V(S_*^2) = V(\chi^2(n)) = 2n \Rightarrow V(S_*^2) = \sigma_{S_*^2}^2 = \sigma^4 \frac{2}{n} \quad (7.753)$$

Et dès lors nous avons une expression de l'écart-type de l'écart-type empirique si nous connaissons l'écart-type de la population:

$$\sigma_{S_*^2} = \sigma^2 \sqrt{\frac{2}{n}} \quad (7.754)$$

Mais nous avons démontré lors de notre étude des estimateurs que:

$$\hat{\sigma} = \sqrt{\frac{n-1}{n}} \sigma \quad (7.755)$$

Dès lors il vient que:

$$\hat{\sigma}_{S_*^2} = \sqrt{\frac{n-1}{n}} \sigma_{S_*^2} = \sqrt{\frac{n-1}{n}} \sigma^2 \sqrt{\frac{2}{n}} = \sqrt{\frac{2(n-1)}{n^2}} \sigma^2 \quad (7.756)$$

Il en découle donc la relation parfois importante dans la pratique:

$$\hat{\sigma}_{S_*^2} = \sqrt{\frac{2(n-1)}{n^2}} \sigma^2 \quad (7.757)$$

Rappelons que la population parente est dite "infinie" si le tirage de l'échantillon avec remise ou encore si la taille N de la population est très supérieure à celle de n de l'échantillon

Remarques:

R1. En laboratoire, les $X_1, \dots, X_n$ peuvent être vues comme une classe d'individus d'un même produit étudié identiquement par différentes équipes de recherche avec des instruments de même précision (écart-type de mesure nul).

R2. S_*^2 est la "**variance interclasse**" également appelée "**variance expliquée**". Donc elle donne la variance d'une mesure ayant eu lieu dans les différents laboratoires.

Ce qui est intéressant c'est qu'à partir du calcul de la loi du Khi-deux en connaissant n et l'écart-type σ^2 il est possible d'estimer cette variance (écart-type) interclasse.

Pour voir que cette dernière propriété est une généralisation élémentaire de la relation:

$$\chi^2(n) = X_1^2 + X_2^2 + \dots + X_n^2 \quad (7.758)$$

il suffit de constater que la variable aléatoire $(n \cdot S_*^2) / \sigma^2$ est une somme de n carrés de $N(0,1)$ indépendants les uns des autres. Effectivement, rappelons qu'une variable aléatoire centrée réduite (voir notre étude de la loi Normale) est donnée par:

$$Y = \frac{X - \mu}{\sigma} \quad (7.759)$$

Dès lors:

$$\frac{nS_*^2}{\sigma^2} = \frac{n}{\sigma^2} \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \quad (7.760)$$

Or, puisque les variables aléatoires $X_1, \dots, X_n$ sont indépendantes et identiquement distribuées selon une loi Normale, alors les variables aléatoires:

$$\frac{X_1 - \mu}{\sigma}, \dots, \frac{X_n - \mu}{\sigma} \quad (7.761)$$

sont aussi indépendantes et identiquement distribuées mais selon une loi Normale centrée réduite.

Puisque:

$$\frac{nS_*^2}{\sigma^2} = \chi^2(n) \quad (7.762)$$

en réarrangeant nous obtenons:

$$\sigma^2 = \frac{nS_*^2}{\chi^2(n)} \quad (7.763)$$

Donc sur la population de mesures, l'écart-type vrai suit la relation donnée ci-dessus. Il est donc possible de faire de l'inférence statistique sur l'écart-type lorsque la moyenne théorique est connue (...).

Puisque la fonction du Khi-deux n'est pas symétrique, la seule possibilité pour faire l'inférence c'est de faire appel au calcul numérique et nous noterons alors l'intervalle de confiance à 95% (par exemple...) de la manière suivante:

$$\frac{nS_*^2}{\chi_{2.5\%}^2(n)} \leq \sigma^2 \leq \frac{nS_*^2}{\chi_{97.5\%}^2(n)} \quad (7.764)$$

Soit en notant $95\% = 1 - \alpha$:

$$\frac{nS_*^2}{\chi_{\alpha/2}^2(n)} \leq \sigma^2 \leq \frac{nS_*^2}{\chi_{1-\alpha/2}^2(n)} \quad (7.765)$$

le dénominateur étant alors bien évidemment la probabilité cumulée. Cette relation est rarement utilisée dans la pratique car la moyenne théorique n'est pas connue. Voyons donc le cas le plus courant:

8.3. I.C. SUR LA VARIANCE AVEC MOYENNE EMPIRIQUE

Cherchons maintenant à faire de l'inférence statistique lorsque la moyenne théorique de la population μ n'est pas connue. Pour cela, considérons maintenant la somme:

$$\chi^2(n) = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^n ((X_i - \bar{X}) + (\bar{X} - \mu))^2 \quad (7.766)$$

où pour rappel $\bar{X}$ est la moyenne empirique (arithmétique) de l'échantillon:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (7.767)$$

En continuant le développement nous avons:

$$\begin{aligned} \chi^2(n) &= \frac{1}{\sigma^2} \sum_{i=1}^n ((X_i - \bar{X}) + (\bar{X} - \mu))^2 = \frac{1}{\sigma^2} \sum_{i=1}^n ((X_i - \bar{X})^2 + 2(X_i - \bar{X})(\bar{X} - \mu) + (\bar{X} - \mu)^2) \\ &= \frac{1}{\sigma^2} \left[\sum_{i=1}^n (X_i - \bar{X})^2 + 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \bar{X}) + \sum_{i=1}^n (\bar{X} - \mu)^2 \right] \\ &= \frac{1}{\sigma^2} \left[\sum_{i=1}^n (X_i - \bar{X})^2 + 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \bar{X}) + n(\bar{X} - \mu)^2 \right] \end{aligned} \quad (7.768)$$

Or, nous avons démontré au début de ce chapitre que la somme des écarts à la moyenne était nulle. Donc:

$$\begin{aligned} \chi^2(n) &= \frac{1}{\sigma^2} \left[\sum_{i=1}^n (X_i - \bar{X})^2 + 2(\bar{X} - \mu) \cdot 0 + n(\bar{X} - \mu)^2 \right] = \frac{1}{\sigma^2} \left[\sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 \right] \\ &= \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma^2 / \sqrt{n}} \right)^2 \end{aligned} \quad (7.769)$$

et reprenons l'estimateur sans biais de la loi Normale (nous changeons de notation pour respecter les traditions et bien différencier la moyenne empirique de la moyenne théorique):

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (7.770)$$

Dès lors:

$$\chi^2(n) = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma^2 / \sqrt{n}} \right)^2 = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma^2 / \sqrt{n}} \right)^2 \quad (7.771)$$

ou autrement écrit:

$$\chi^2(n) = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{M_n - \mu}{\sigma / \sqrt{n}} \right)^2 \quad (7.772)$$

Puisque le deuxième terme (au carré) suit une loi Normale centrée réduite aussi, alors si nous le supprimons nous obtenons de par la propriété démontrée plus haut de la loi du Khi-deux:

$$\chi^2(n-1) = \frac{(n-1)S^2}{\sigma^2} \quad (7.773)$$

Ces développements nous permettent cette fois-ci de faire aussi de l'inférence sur la variance σ^2 d'une loi $N(\mu, \sigma)$ lorsque les paramètres μ et σ sont tous les deux inconnus pour l'ensemble de la population. C'est ce résultat qui nous donne, par exemple, l'intervalle de confiance:

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)} \quad (7.774)$$

lorsque la moyenne théorique μ est donc inconnue.

8.4. I.C. SUR LA MOYENNE AVEC VARIANCE EMPIRIQUE CONNUE

Nous avons démontré beaucoup plus haut que la loi de Student provenait de la relation suivante:

$$T(k) = \frac{Z}{\sqrt{U/k}} \quad (7.775)$$

si Z et U sont des variables aléatoires indépendantes et si Z suit une loi Normale centrée réduite $N(0,1)$ et U une loi du Khi-deux $\chi^2(k)$ tel que:

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(n/2)\sqrt{k\pi}} \left(1 + \frac{t^2}{n}\right)^{-\frac{k+1}{2}} \quad (7.776)$$

et rappelons que la fonction de densité (distribution) est symétrique!

Voici une application très importante du résultat ci-dessus:

Supposons que $X_1, \dots, X_n$ constituent un échantillon aléatoire de taille n issu de la loi $N(\mu, \sigma)$. Alors nous pouvons déjà écrire que selon les développements faits plus haut:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \quad (7.777)$$

Et pour U qui suit une loi $\chi^2(k)$, si nous posons $k = n - 1$ alors selon les résultats obtenus plus haut:

$$U = \frac{(n-1)S^2}{\sigma^2} = \chi^2(n-1) \quad (7.778)$$

Nous avons alors après quelques simplifications triviales:

$$\frac{Z}{\sqrt{U/k}} = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2} / (n-1)}} = \frac{\frac{\bar{X} - \mu}{1/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{n-1}}} = \frac{\frac{\bar{X} - \mu}{1/\sqrt{n}}}{\sqrt{S^2}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad (7.779)$$

Donc puisque:

$$T(k) = \frac{Z}{\sqrt{U/k}} \quad (7.780)$$

suit une loi de Student de paramètre k alors nous obtenons le "**independant one-sample t-test**" (en anglais) ou "**test-T à 1 échantillon**":

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = T(n-1) \quad (7.781)$$

quisuit aussi une loi de Student de paramètre $n-1$ et qui est très utilisé dans les laboratoires pour les tests d'étalonnages.

Ce qui nous donne aussi après réarrangement:

$$\mu = \bar{X} - \frac{S}{\sqrt{n}} T(n-1) \quad (7.782)$$

Ce qui nous permet de faire de l'inférence sur la moyenne μ d'une loi Normale d'écart-type théorique inconnu (sous-entendu qu'il n'y a pas assez de valeurs expérimentales) mais dont l'estimateur sans biais de l'écart-type est connu. C'est ce résultat qui nous donne l'intervalle de confiance:

$$\bar{X} - \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \quad (7.783)$$

où nous retrouvons les mêmes indices que pour l'inférence statistique sur la moyenne (espérance) d'une variable aléatoire d'écart-type (théorique) connu puisque la loi de Student tend asymptotiquement pour de grandes valeurs de n vers une loi Normale. Ainsi, l'intervalle précédent et l'intervalle suivant:

$$\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (7.784)$$

donneront des valeurs très proches (à la troisième décimale) pour des grandeurs de n aux alentours des 10'000 (dans la pratique on considère qu'à partir de 100 c'est identique...).

Nous déduisons immédiatement par la propriété de stabilité de la loi du Khi-deux (démontrée plus haut par le fait qu'elle découle de la loi Gamma) le test suivant que nous retrouvons dans de très nombreux logiciels de statistiques:

$$(\bar{X}_2 - \bar{X}_1) - T_{\alpha/2}(n-1) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_1}} \leq \mu_2 - \mu_1 \leq (\bar{X}_2 - \bar{X}_1) + T_{\alpha/2}(n-1) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_1}} \quad (7.785)$$

appelé "**test-T bilatéral sur la différence de deux moyennes**".

Nous pouvons bien évidemment dès lors aussi déterminer la probabilité avec laquelle la moyenne est dedans ou en dehors d'un certain intervalle... (l'une comme l'autre étant beaucoup utilisées dans l'industrie).

Le lecteur pourra s'amuser à contrôler avec la version française de Microsoft Excel 11.8346 que pour un grand nombre de mesures n , la loi de Student tend vers la loi Normale centrée réduite en comparant les valeurs des deux fonctions ci-dessous:

$$\begin{aligned} &= \text{LOI.STUDENT.INVERSE.N}(5\%/2; n-1) \\ &= \text{LOI.NORMAL.STANDARD.INVERSE.N}(5\%/2) \end{aligned}$$

Remarque: Le résultat précédent fut obtenu par William S. Gosset aux alentours de 1910. Gosset qui avait étudié la mathématique et la chimie, travaillait comme statisticien pour la brasserie Guinness en Angleterre. À l'époque, on savait que si $X_1, \dots, X_n$ sont des variables aléatoires indépendantes et identiquement distribuées alors:

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = N(0, 1) \quad (7.786)$$

Toutefois, dans les applications statistiques on s'intéressait bien évidemment plutôt à la quantité:

$$\frac{\bar{X} - \mu}{S / \sqrt{n}} \quad (7.787)$$

on se contentait alors de supposer que cette quantité suivait à peu près une loi Normale centrée réduite ce qui n'était pas une mauvaise approximation comme le montre l'image ci-dessous ($df = n - 1$):

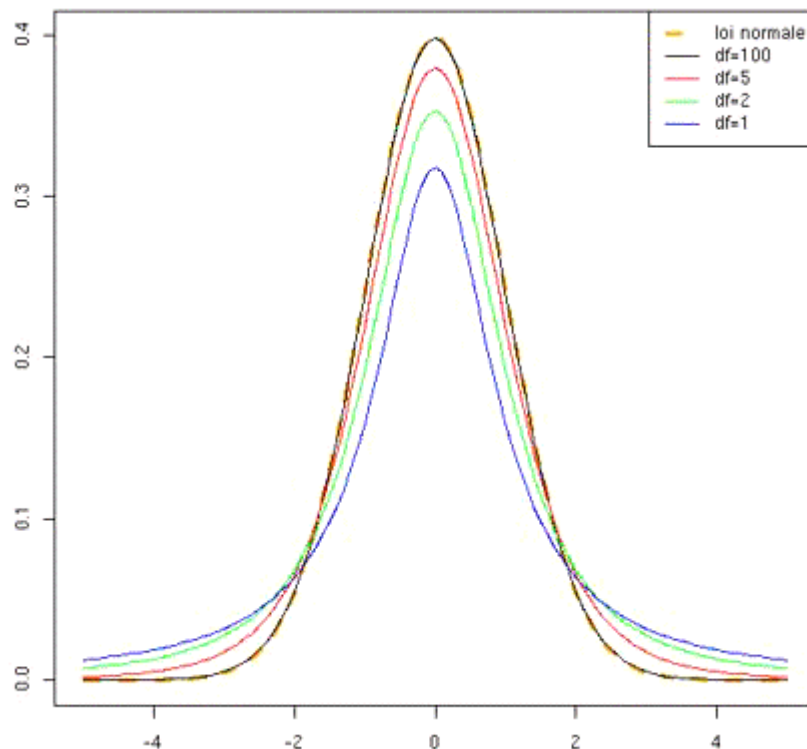


Figure: 7.35 - Comparaison entre la fonction de distribution Normale et celle de Student

Suite à de nombreuses simulations, Gosset arriva à la conclusion que cette approximation était valide seulement lorsque n est suffisamment grand (donc cela lui donnait l'indication comme quoi il devait y avoir quelque part derrière le théorème central limite). Il décida de déterminer l'origine de la distribution et après avoir suivi un cours de statistique avec Karl Pearson il obtint son fameux résultat qu'il publia sous le pseudonyme de Student. Ainsi, on appelle loi de Student la loi de probabilité qui aurait dû être appelée la loi ou fonction de Gosset.

Signalons enfin que le test de Student est très utilisé pour identifier si des variations (progressions ou l'inverse) de la moyenne des chiffres de deux populations identiques sont significatives. C'est-à-dire que

si la taille de deux échantillons dépendants est identique alors nous pouvons créer le test suivant (nous avons indiqué tous les différents types d'écritures que l'on peut retrouver dans la littérature et dans les nombreux logiciels implémentant ce test):

$$\begin{aligned}
 T(n-1) &= \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{X_{\bar{D}} - \mu_D}{S_D/\sqrt{n}} = \frac{\bar{D} - \delta_0}{S_D/\sqrt{n}} = \frac{\frac{1}{n} \sum_i d_i - \delta_0}{S_D/\sqrt{n}} = \frac{\frac{1}{n} \sum_i (X_{i2} - X_{i1}) - \delta_0}{S_D/\sqrt{n}} \\
 &= \frac{\left(\frac{1}{n} \sum_i X_{i2} - \frac{1}{n} \sum_i X_{i1} \right) - \delta_0}{S_D/\sqrt{n}} = \frac{(\bar{X}_2 - \bar{X}_1) - \delta_0}{S_D/\sqrt{n}} \quad (7.788)
 \end{aligned}$$

Avec:

$$S_D = \sqrt{\frac{1}{n-1} \sum_i (d_i - \bar{D})^2} \quad (7.789)$$

La relation antérieure est donc très utile pour comparer deux fois le même échantillon dans des situations différentes de mesure (ventes avant ou après rabais d'un article par exemple). Cette dernière relation est appelée "**test-T de deux moyennes d'échantillons appariés (ou échantillons dépendants)**".

Définition: Nous parlons "**d'échantillons appariés**" (par paires) si les échantillons de valeurs sont prises 2 fois sur les mêmes individus (donc les valeurs des paires ne sont pas indépendantes, contrairement à deux échantillons pris indépendamment).

8.5. TEST BINOMIAL EXACT

Il arrive fréquemment lors de mesures que l'on souhaite comparer si deux échantillons de petite taille pris au hasard (sans remise!) d'une population elle aussi petite... sont significativement différents ou non alors que l'on attendait une égalité parfaite (50%/50%)!

Il s'agit donc d'un test adapté aux cas suivants:

- Savoir si un échantillon d'une population préfère utiliser une technique de travail plutôt qu'une autre alors que l'on s'attend à ce que la population utilise autant l'une que l'autre
- Savoir si un échantillon d'une population a une caractéristique prédominante parmi deux possibilités alors que l'on s'attend à ce que la population soit parfaitement équilibrée.

Avant d'aller plus en détails, rappelons qu'il faut être extrêmement prudent quant à la manière d'obtenir les deux échantillons. Il faut que l'expérience soit non biaisée, cela signifie pour rappel, que le protocole de tirage ne doit en aucun cas avantager l'une au l'autre des caractéristiques de la population (si vous étudiez l'équilibre homme/femme dans une population en attirant dans le sondage des personnes grâce à un cadeau sous la forme de bijoux vous aurez alors un échantillon biaisé... car vous aurez probablement naturellement plus de femmes que d'hommes...).

Ceci étant dit, cette situation correspond donc à une loi binomiale pour laquelle nous avons démontré plus haut dans ce chapitre que la probabilité de k réussites pour une population de taille N dont la

probabilité de réussite est p (et la probabilité d'échec q donc de $1-p$) était donnée par la relation:

$$P(N, k) = C_k^N p^k q^{N-k} = C_k^N p^k (1-p)^{N-k} \quad (7.790)$$

Dans le cas qui nous intéresse, nous avons donc $p = q = 0.5$:

$$P(N, k) = C_k^N 0.5^k 0.5^{N-k} = 0.5^N C_k^N \quad (7.791)$$

tout en se rappelant que la distribution ne sera pas pour autant symétrique et ce surtout si la taille N de la population est petite.

Si nous notons maintenant x le nombre de réussites (considéré comme la taille du premier échantillon) et y le nombre d'échecs (considéré comme la taille du deuxième échantillon), nous avons alors:

$$P(N, k) = 0.5^N C_k^N = P(x+y, k) = 0.5^{x+y} C_k^{x+y} \quad (7.792)$$

Ceci étant fait, pour construire le test et de par l'asymétrie de la distribution, nous allons calculer la probabilité cumulée que k soit plus petit que le x obtenu par l'expérience et la sommer à la probabilité cumulée pour que k soit plus grand que le y obtenu par l'expérience (ce qui correspond à la probabilité cumulée des queues respectivement gauche et droite de la distribution). Cette somme correspond donc à la probabilité:

$$P = 0.5^N \sum_{k=0}^x C_k^N + 0.5^N \sum_{k=y}^N C_k^N \quad (7.793)$$

et cette dernière relation est appelée "**test binomial exact (bilatéral)**".

Si la probabilité P obtenue pour la somme est au-dessus d'une certaine probabilité cumulée fixée à l'avance, nous dirons alors que la différence avec un échantillon tiré au hasard dans une population parfaitement équilibrée n'est pas significative (en bilatéral...) et respectivement si elle est en-dessous, la différence sera donc significative et nous rejeterons l'équilibre supposé.

Ainsi, si:

$$0.5^N \sum_{k=0}^{x'} C_k^N + 0.5^N \sum_{k=y'}^N C_k^N = 0.5^N \left(\sum_{k=0}^{x'} C_k^N + \sum_{k=y'}^N C_k^N \right) \geq \alpha \quad (7.794)$$

la différence par rapport à une population équilibrée sera considérée comme non significative. Souvent on prendra au maximum α comme valant 5% (mais rarement en-dessous) ce qui correspond donc à un intervalle de confiance de 95%.

Malheureusement d'un logiciel de statistiques à l'autre les paramètres demandés ou les résultats obtenus ne seront pas nécessairement les mêmes (les tableurs n'intègrent pas de fonction spécifique pour le test binomial, il faudra souvent construire un tableau ou programmer soi-même la fonction). Par exemple, certains logiciels calculent systématiquement et imposent (ce qui est assez logique dans un sens...):

$$0.5^N \left(\sum_{k=0}^{x-1} C_k^N + \sum_{k=y+1}^N C_k^N \right) = P \quad (7.795)$$

Exemple:

D'une petite population ayant deux caractéristiques x et y particulières qui nous intéressaient et pour laquelle nous nous attendions à avoir un parfait équilibre tel que $x = y$ nous avons en réalité obtenu $x = 5$ et $y = 7$. Nous souhaiterions faire le calcul avec Microsoft Excel 11.8346 pour savoir si cette différence est significative ou non à un niveau de 5%?

Pour répondre à cette question, nous allons donc calculer la probabilité cumulée

$$0.5^N \left(\sum_{k=0}^x C_k^N + \sum_{k=y}^N C_k^N \right) = 0.5^{12} \left(\sum_{k=0}^5 C_k^{12} + \sum_{k=7}^{12} C_k^{12} \right) \quad (7.796)$$

ce qui nous donne:

k	Coeff. binomial
0	1
1	12
2	66
3	220
4	495
5	792
7	792
8	495
9	220
10	66
11	12
12	1
P	0.774

Figure: 7.36 - Valeurs du calcul des coefficients binomiaux dans Microsoft Excel 11.8346

soit explicitement:

k	Coeff. binomial
0	=COMBIN(12,E12)
1	=COMBIN(12,E13)
2	=COMBIN(12,E14)
3	=COMBIN(12,E15)
4	=COMBIN(12,E16)
5	=COMBIN(12,E17)
7	=COMBIN(12,E18)
8	=COMBIN(12,E19)
9	=COMBIN(12,E20)
10	=COMBIN(12,E21)
11	=COMBIN(12,E22)
12	=COMBIN(12,E23)
P	=0.5^12*SOMME(F12:F23)

Figure: 7.37 - Formules du calcul des coefficients binomiaux dans Microsoft Excel 11.8346

donc la probabilité cumulée étant de 0.774 (soit 77.4%) la différence par rapport à une population

équilibrée sera considérée donc comme non significative.

Remarque: Ce test est également utilisé par la majorité des logiciels de statistiques (comme Minitab) pour donner un intervalle de confiance de la conformité d'opinions par rapport à celle d'un expert. C'est ce que nous appelons une étude R&R (reproductibilité & répétabilité) par attributs (voir mon livre sur Minitab pour un exemple).

8.6. I.C. POUR UNE PROPORTION

Indiquons que certains statisticiens utilisent le fait que la loi Normale découle de la loi de Poisson qui elle-même découle de la loi Binomiale (nous l'avons démontré lorsque n tend vers l'infini et que p et q sont du même ordre) pour faire un intervalle de confiance dans le cadre de l'analyse de proportions (très utilisé dans l'analyse de la qualité dans les industries).

Pour voir cela, notons X_i la variable aléatoire définie par:

$$X_i = \begin{cases} 1 & \text{si le } i\text{-ème élément de l'échantillon possède l'attribut } A \\ 0 & \text{sinon} \end{cases} \quad (7.797)$$

où l'attribut A peut être la propriété "défectueux" ou "non défectueux" par exemple pour une analyse de pièces. Nous noterons k le nombre de réussites de l'attribut A .

La variable aléatoire $X = X_1 + X_2 + \dots + X_n$ nous l'avons démontré au début de ce chapitre, suit une loi Binomiale de paramètres n et p avec les moments:

$$\begin{aligned} \mu &= E(X) = n \cdot p \\ \sigma &= \sqrt{V(X)} = \sqrt{n \cdot p \cdot q} = \sqrt{n \cdot p \cdot (1-p)} \end{aligned} \quad (7.798)$$

Ceci étant, nous ne connaissons pas la valeur vraie de p . Nous allons donc utiliser l'estimateur de la loi Binomiale démontré plus haut:

$$\hat{p} = \frac{k}{n} = \frac{X}{n} \quad (7.799)$$

D'après les propriétés de l'espérance nous avons alors:

$$E(\hat{p}) = \mu_{\hat{p}} = E\left(\frac{X}{n}\right) = \frac{E(X)}{n} = \frac{np}{n} = p \quad (7.800)$$

Et nous avons d'après les propriétés de la variance, la relation suivante pour la variance de la moyenne empirique de la proportion:

$$V(\hat{p}) = \sigma_{\hat{p}}^2 = V\left(\frac{X}{n}\right) = \frac{V(X)}{n^2} = \frac{n \cdot p \cdot (1-p)}{n^2} = \frac{p \cdot (1-p)}{n} \quad (7.801)$$

Ce qui nous amène alors à:

$$\mu_{\hat{p}} \cong \hat{p} \text{ et } \sigma_{\hat{p}} \cong \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \quad (7.802)$$

Maintenant rappelons enfin que nous avons démontré que la loi Normale découlait de la loi Binomiale sous certaines conditions (les praticiens admettent que c'est applicable tant que $n > 50$ et $n \cdot p \geq 5$). Autrement dit, que la variable aléatoire X suivant une loi Binomiale suit une loi Normale sous certaines conditions. Évidemment, si X suit une loi Normale alors X/n aussi (et donc $\hat{p}$...). Dès lors nous pouvons centrer et réduire $\hat{p}$ afin qu'il se comporte comme la variable aléatoire centrée réduite notée Z :

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}} \quad (7.803)$$

Exemple:

Si 5% de la production annuelle d'une entreprise est défectueuse, quelle est la probabilité qu'en prenant un échantillon de 75 pièces de la ligne de production que seulement 2% ou moins soit défectueux?

Nous avons dès lors avec:

$$Z = \frac{0.02 - 0.05}{\sqrt{\frac{0.05 \cdot 0.95}{75}}} \cong -1.19 \quad (7.804)$$

La probabilité cumulée correspondante à cette valeur de la variable aléatoire est avec la version anglaise de Microsoft Excel 11.8346:

$$= \text{NORMSDIST}(-1.19) = 11.66\%$$

Nous pouvons maintenant approximer l'intervalle de confiance pour la proportion en se basant sur la loi Binomiale et son comportement asymptotiquement normal dans les conditions démontrées lors de notre introduction de la loi Normale tel que:

$$\hat{p} - Z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \leq p \leq \hat{p} + Z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \quad (7.805)$$

Avant de passer à un exemple, il est peut-être utile de préciser au lecteur que cette approximation par une loi Normale est très courante et que nous allons la rencontrer encore de nombreuses fois dans les démonstrations qui vont suivre. C'est tellement courant qu'on a même donné un nom à cette méthode...: la "**méthode de Wald**" (bon en réalité il y en a plusieurs de méthodes de Wald mais c'est la plus connue que nous utiliserons à chaque fois).

Exemple:

Prenons $\alpha = 5\%$, nous avons alors:

$$\begin{aligned} \hat{p} - Z_{0.025} \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} &\leq p \leq \hat{p} + Z_{0.025} \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \\ \hat{p} - 1.96 \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} &\leq p \leq \hat{p} + 1.96 \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \end{aligned} \quad (7.806)$$

Sur une production de 300 éléments nous en avons trouvé 8 qui étaient défectueux. Quel est donc l'intervalle de confiance?

Nous vérifions d'abord avec:

$$\hat{p} = \frac{8}{300} \cong 2.66\% \quad (7.807)$$

que:

$$n \cdot \hat{p} = 300 \cdot \frac{8}{300} = 8 \geq 5 \quad (7.808)$$

Donc il est acceptable d'utiliser l'intervalle de confiance par la loi Normale. Nous avons dès lors:

$$\begin{aligned} \frac{8}{300} - 1.96 \sqrt{\frac{\frac{8}{300} \cdot \left(1 - \frac{8}{300}\right)}{300}} &\leq p \leq \frac{8}{300} + 1.96 \sqrt{\frac{\frac{8}{300} \cdot \left(1 - \frac{8}{300}\right)}{300}} \\ 0.84\% &\leq p \leq 4.49\% \end{aligned} \quad (7.809)$$

Pour clore ce sujet, nous pouvons évidemment nous intéresser aussi au nombre d'individus (taille d'échantillon) qu'il faut avoir pour satisfaire une certaine précision d'intervalle de confiance (imposé) en ayant un écart-type imposé.

Nous avons donc selon les hypothèses susmentionnées et dans l'acceptation de l'approximation par une loi Normale que:

$$Z \cong \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{\delta}{\sqrt{\frac{p(1-p)}{n}}} \quad (7.810)$$

Et en procédant de manière identique aux développements effectués plus haut avec la loi Normale, nous obtenons:

$$n \cong \frac{Z_{\alpha/2}^2 p(1-p)}{\delta^2} = \frac{Z_{\alpha/2}^2 pq}{\delta^2} \quad (7.811)$$

dont nous prenons évidemment normalement la valeur entière supérieure dans la pratique...

Une question qui revient souvent dans la pratique concerne le fait de savoir s'il faut prendre en unilatéral ou bilatéral. Au fait cela il n'y a pas de réponse précise, tout dépend de ce que nous cherchons à mettre en évidence.

Remarque: La taille de la population mère pour les relations développées plus haut n'entre pas en ligne de compte dans le calcul des intervalles de confiance ni dans celui de la taille de l'échantillon, et pour cause, elle est considérée infinie. Il faut donc faire attention à ne pas avoir parfois des tailles d'échantillons qui sont plus grandes que la population mère réelle possible...

Exemple:

Nous souhaiterions savoir le nombre d'individus (taille d'échantillon) à prendre d'un lot de production sachant que la proportion de défectueux est imposée à 30% avec une erreur tolérée d'environ 5% entre la proportion réelle et empirique et ce afin d'obtenir un intervalle de confiance à un niveau de 95% du résultat:

$$n \cong \frac{Z_{5\%} \cdot 30\% \cdot (1 - 30\%) }{5\%^2} = \frac{1.96 \cdot 30\% \cdot 70\% }{5\%^2} = 165 \quad (7.812)$$

Remarque: La dernière relation est très très souvent utilisée en théorie des sondages (analyses pour des votations avec réponses de type: Oui/Non) où parfois la taille de l'échantillon n est imposée pour des raisons de coûts du sondage et dont nous cherchons à calculer l'incertitude δ et parfois l'inverse (l'incertitude est imposée et donc nous cherchons à connaître la taille de l'échantillon).

8.6.1. TEST DE L'ÉGALITÉ DE DEUX PROPORTIONS

Toujours dans le même contexte que l'approximation précédente de la loi Binomiale par une loi Normale, l'industrie (en particulier la biostatistique) est friande de comparer deux proportions de deux populations différentes afin de savoir si elles sont statistiquement égales ou non (autrement dit: significativement différentes ou pas).

Dès lors rappelons que nous avons démontré la stabilité de la loi Normale si deux variables aléatoires étaient indépendantes et identiquement distribuées (selon une loi Normale donc!):

$$X - Y = N\left(\mu = \mu_X - \mu_Y, \sigma = \sqrt{\sigma_X^2 + \sigma_Y^2}\right) \quad (7.813)$$

Dans le cadre des hypothèses susmentionnées il en est alors de même approximativement pour la différence de deux proportions:

$$\begin{aligned} \frac{X}{n_1} - \frac{Y}{n_2} &\cong N\left(\mu_{\bar{p}_X} - \mu_{\bar{p}_Y}, \sqrt{\sigma_{\bar{p}_X}^2 + \sigma_{\bar{p}_Y}^2}\right) \cong N\left(\hat{p}_X - \hat{p}_Y, \sqrt{\sigma_{\hat{p}_X}^2 + \sigma_{\hat{p}_Y}^2}\right) \\ &= N\left(\hat{p}_X - \hat{p}_Y, \sqrt{\frac{\hat{p}_X(1 - \hat{p}_X)}{n_1} + \frac{\hat{p}_Y(1 - \hat{p}_Y)}{n_2}}\right) \end{aligned} \quad (7.814)$$

Dès lors nous savons que cette nouvelle variable centrée réduite suit une loi Normale selon:

$$Z \cong \frac{(\hat{p}_X - \hat{p}_Y) - \mu}{\sqrt{\frac{\hat{p}_X \cdot (1 - \hat{p}_X)}{n_1} + \frac{\hat{p}_Y \cdot (1 - \hat{p}_Y)}{n_2}}} \quad (7.815)$$

et comme nous cherchons à savoir la probabilité cumulée que l'espérance théorique de la différence est nulle, cette dernière relation se réduit alors dans ce cas à:

$$Z \cong \frac{(\hat{p}_X - \hat{p}_Y)}{\sqrt{\frac{\hat{p}_X \cdot (1 - \hat{p}_X)}{n_1} + \frac{\hat{p}_Y \cdot (1 - \hat{p}_Y)}{n_2}}} \quad (7.816)$$

Évidemment nous pouvons aussi construire (comme toujours...) un intervalle de confiance à partir de cette relation.

Il semblerait cependant que cette dernière relation approximative serait d'après l'expérience plus correcte en prenant pour dénominateur:

$$Z \cong \frac{(\hat{p}_X - \hat{p}_Y)}{\sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n_1} + \frac{\hat{p} \cdot (1 - \hat{p})}{n_2}}} = \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{\hat{p} \cdot (1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (7.817)$$

où $\hat{p}$ sera pris comme le mélange de deux populations. C'est-à-dire:

$$\hat{p} = \frac{\text{somme du nombre d'événements dans chacun des deux prélèvements}}{\text{somme du nombre d'échantillons}} \quad (7.818)$$

soit (en changeant la notations des indices des proportions expérimentales):

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} \quad (7.819)$$

Ce test est aussi appelé "**test Z de l'égalité de deux proportions**"

Exemple:

Dans le cadre d'un plan d'échantillonnage (cf. chapitre de Génie Industriel) nous avons prélevé sur un premier lot de 50 individus, 48 en parfait états. Dans un second lot de 30 individus, 26 étaient en bon état.

Nous avons donc:

$$\hat{p}_X = \frac{48}{50} = 96\%, \hat{p}_Y = \frac{26}{30} = 86.66\% \quad (7.820)$$

Nous souhaiterions donc savoir si la différence est significative avec une certitude de 95% ou simplement due au hasard. Nous utilisons alors:

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{48 + 26}{50 + 30} = 92.50\% \quad (7.821)$$

et:

$$Z \cong \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{\hat{p} \cdot (1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{0.96 - 0.8666}{\sqrt{0.925 \cdot (1 - 0.925) \left(\frac{1}{50} + \frac{1}{30} \right)}} \cong 1.535 \quad (7.822)$$

Ce qui correspond à une probabilité cumulée en utilisant la version anglaise de Microsoft Excel 11.8346 de:

$$= \text{NORMSDIST}(1.535) = 93.77\%$$

Donc la différence est due au hasard (ceci dit c'est presque in extremis...). Autrement dit, elle n'est pas significative sous les contraintes énoncées.

8.6.2. TEST DES SIGNES

Nous mesurons quelque chose sur un échantillon puis, plus tard, nous mesurons la même chose sur ce même échantillon mais avec une autre méthode (donc il s'agit donc d'échantillons appariés!). Les deux classements ordonnés des mesures sont comparés et chaque observation est affectée d'un signe ("+" en cas d'élévation dans le classement, "-" en cas de descente). Celles qui restent au même niveau sont éliminées.

Selon l'hypothèse à tester, il y a autant de "+" que de "-", c'est-à-dire que la médiane de la distribution n'a pas bougé (cette affirmation peut ne pas paraître évidente à la première lecture il faut donc bien prendre du temps parfois pour réfléchir là-dessus).

L'idée étant que pour chaque couple de valeurs, il n'y a que deux signes possibles de variations, nous avons une chance sur deux (50% de probabilité) que la différence soit positive ou négative. Ce test est donc basé uniquement sur l'étude des signes des différences observées entre les paires d'individus, quelles que soient les valeurs de ces différences.

Nous pouvons souhaiter contrôler deux hypothèses:

- L'inégalité des proportions de signes doit être significative. Donc l'un des signes doit être en petit nombre par rapport à l'autre, ce qui correspond à un test unilatéral gauche (la probabilité cumulée d'avoir ce petit nombre de signes doit être inférieur à un niveau α donné).
- La proportion des deux signes doit être faiblement déséquilibrée ($P(+) = P(-) = 0.5$). Il s'agit donc dans ce cas d'un test en bilatéral (c'est le cas le plus courant) avec un certain niveau α donné.

Pour pouvoir créer un tel test, nous allons considérer l'apparition des "+" et des "-" comme un système de tirage aléatoire binaire dont l'ordre des succès n'est pas pris en compte (il s'agit donc d'une loi binomiale ou hypergéométrique) et avec remise (ce qui élimine d'emblée la loi hypergéométrique qui n'est pas symétrique et pose des problèmes d'utilisation dans la pratique...). Pour considérer un tirage aléatoire avec remise (alors qu'on ne fait pas réellement de remise), il faut que la population N soit

grande. Raison pour laquelle le test des signes considère que les valeurs appariées doivent être continues (ce qui permet in extenso d'approcher la loi hypergéométrique par la loi binomiale). Cependant certains logiciels de statistiques utilisent la loi hypergéométrique pour des soucis de précision.

Remarque: Il faut savoir que la majorité des logiciels de statistiques, font implicitement l'hypothèse lors de ce test que les données sont continues et utilisent la loi binomiale.

Exemple:

Considérons deux séries de mesures avec deux méthodes différentes. Nous souhaiterions tester l'hypothèse avec un niveau α de 5% si la différence entre les deux méthodes est significative (nous nous attendons donc à une équilibre des signes). Il s'agit donc d'un test des signes à deux échantillons (sachant qu'il est possible de faire la même chose en comparant les valeurs d'un seul et unique échantillon à sa médiane).

20.4, 25.4, 25.6, 25.6, 26.6, 28.6, 28.7, 29, 29.8, 30.5, 30.9, 31.1
20.7, 26.3, 26.8, 28.1, 26.2, 27.3, 29.5, 32, 30.9, 32.3, 32.3, 31.7

Nous avons donc les différences:

-0.3, -0.9, -1.2, -2.5, 0.4, 1.3, -0.8, -3.0, -1.1, -1.8, -1.4, -0.6

Soit:

-, -, -, -, +, +, -, -, -, -, -, -

Bon il déjà clair que le résultat va être le rejet de l'hypothèse comme quoi il n'y pas de différence. Mais faisons quand même le calcul. Comme le test est en bilatéral à un niveau de 5%, la probabilité cumulée d'avoir obtenu au moins deux signes "+" ne doit pas être inférieure à 2.5% et pas supérieure à 97.5% si l'on veut accepter (ne pas rejeter) l'hypothèse comme quoi la différence n'est pas significative.

Nous avons alors:

$$P(X \leq 2) = \sum_{k=0}^2 C_k^{12} 0.5^k (1-0.5)^{12-k} = \sum_{k=0}^2 C_k^{12} 0.5^k 0.5^{12-k} = 0.5^{12} \sum_{k=0}^2 C_k^{12} \cong 1.93\% \quad (7.823)$$

Soit avec la version française de Microsoft Excel 14.0.6123:

$$=LOI.BINOMIALE(2;12;0.5;1)=1.928\%$$

ou si nous faisons pas d'approximation en étant plus précis avec la loi hypergéométrique:

$$=LOI.HYPERGEOMETRIQUE.N(2;24/2;12;24;VRAI)=0.17\%$$

ce qui n'est guère plus brillant...

Donc la probabilité cumulée est inférieure à 2.5% et n'est de loin pas supérieure à 97.5%, nous rejetons l'hypothèse comme quoi la différence n'est pas significative.

Nous pourrions accepter l'hypothèse si nous prenions pour α la valeur:

$$p - \text{value} \cong 1.93\% \cdot 2 = 3.86\% \quad (7.824)$$

mais bon ce n'est pas le cas!

Enfin, pour terminer concernant ce test des signes (test de la médiane), indiquons que certains logiciels de statistiques proposent un intervalle de confiance de la médiane basé sur la méthode de calcul exposée précédemment (intervalle de confiance d'une loi binomiale). Cependant, nous pensons qu'il vaudrait mieux favoriser le bootstrapping comme nous l'avons vu dans le chapitre de Méthodes Numériques, nous nous abstenons donc de présenter cette technique ici. De plus il est peut utile de préciser que certains font une approximation en loi Normale (comme avec la majorité des tests mais nous nous en abstenons dans le cas présent).

8.6.3. TEST DE LA MÉDIANE DE MOOD

Nous considérons deux échantillons indépendants $(X_1, \dots, X_{n_1})$ et $(Y_1, \dots, Y_{n_2})$. Nous supposons que $(X_1, \dots, X_{n_1})$ est un échantillon indépendant et identique distribué d'une loi continue F et $(Y_1, \dots, Y_{n_2})$ est un échantillon indépendant et identiquement distribué d'une loi continue G .

Après regroupement des $n_1 + n_2$ valeurs des deux échantillons, $k = n_1 \cdot M_n$ est le nombre d'observations X_i qui sont supérieures à la médiane des $N = n_1 + n_2$ observations (la notation n'est pas géniale car elle peut faire croire à une multiplication mais bon...).

Sous l'hypothèse nulle que les variables X et Y suivent la même loi continue (c'est-à-dire $G=F$), la variable $k = n_1 \cdot M_n$ peut prendre les valeurs $0, 1, \dots, n_1$ selon la loi hypergéométrique:

$$P(N, N/2, n_1, k) = \frac{C_k^{n_1} C_{N/2-k}^{n_2}}{C_{N/2}^N} = \frac{\frac{n_1!}{k!(n_1-k)!} \frac{(N/2-k)!(n_2-(N/2-k))!}{N!}}{\frac{N!}{N/2!(N-N/2)!}} \quad (7.825)$$

Dès lors, nous pouvons calculer la probabilité cumulée en unilatéral d'avoir k . Le test de Mood est donc un test purement unilatéral.

Exemple:

Considérons les deux échantillons:

23.4, 24.4, 24.6, 24.9, 25.0, 26.2, 26.3, 26.8, 26.8, 26.9, 27.0, 27.6, 27.7

22.5, 22.9, 23.7, 24, 24.4, 24.5, 25.3, 26, 26.2, 26.4, 26.7, 26.9, 27.4

La médiane globale calculée avec Microsoft Excel 14.0.6123 est de 26.10. Nous avons au total:

$$k = n_1 \cdot M_n = 8 \quad (7.826)$$

Il vient alors avec la version française de Microsoft Excel 14.0.6123:

$$=LOI.HYPERGEOMETRIQUE.N(8;26/2;13;26;VRAI)=94.24\%$$

Donc à un seuil de 5%, nous ne rejetons pas l'hypothèse nulle (mais bon étant proche de la limite c'est un peu périlleux de conclure cela...). Si nous faisons le même calcul avec la loi Binomiale nous obtenons:

$$LOI.BINOMIALE.N(8;26/2;0.5;1)=86.65\%$$

Mais bien évidemment ici l'approximation ne s'applique pas puisque l'approximation par une loi binomiale est acceptable dans la pratique que lorsque l'échantillon est environ 10 fois plus petit que la population.

Remarque: Il existe malheureusement plusieurs versions du test de Mood. Par exemple un logiciel comme Minitab compare à l'aide d'une table de contingence... le contingent de valeurs au-dessus ou en-dessous de la médiane et fait un simple test d'indépendance du Khi-deux (test de Pearson) vu dans le chapitre de Méthodes Numériques.

8.6.4. TEST DE POISSON (1 ÉCHANTILLON)

Nous savons qu'un certain nombre d'événements rares suivent une loi de Poisson. Nous pouvons alors nous permettre comme pour toute autre loi, de calculer la probabilité cumulée dans un intervalle donné (bilatéral ou unilatéral).

Donc, si nous avons une variable aléatoire discrète suivant une loi de Poisson:

$$P_k = \frac{\mu^k}{k!} \cdot e^{-\mu} \quad (7.827)$$

Nous avons alors en unilatéral droite à un certain niveau de confiance α , la valeur de n de k la plus proche satisfaisant la condition:

$$\sum_{k=0}^n \frac{\mu^k}{k!} e^{-\mu} \rightarrow 1 - \alpha \quad (7.828)$$

Donc pour trouver la valeur de n (entier strictement positif ou nul) il faudrait inverser la somme, ce qui est peu... pratique (raison pour laquelle aucun tableur à ce jour ne propose de fonction pour la loi de Poisson inverse).

Maintenant, rappelons que nous avons vu dans le chapitre de Suites Et Séries, la série de Taylor (Maclaurin) avec reste intégral à l'ordre $n-1$ autour de 0 jusqu'à λ suivante:

$$\sum_{k=0}^{n-1} \frac{\lambda^k}{k!} e^{-\lambda} = 1 - \int_0^{\lambda} \frac{1}{2^{k/2-1} \Gamma(k/2)} x^{k/2-1} e^{-x/2} dx \quad (7.829)$$

Résultat que nous avons également donné sous la forme de fonctions pour la version française de Microsoft Excel 14.0.6123 pour que le lecteur puisse vérifier cette équivalence:

$$\begin{aligned}
 &= \text{LOI.POISSON.N}(x \in \mathbb{N}; \mu; \text{VRAI}) \\
 &= 1 - \text{LOI.KHIDEUX.N}(2 * \mu; 2 * (x + 1); \text{VRAI})
 \end{aligned}
 \quad (7.830)$$

Il vient alors que dans les tableurs, nous pouvons utiliser la loi du Khi-deux inverse pour calculer l'inverse de la loi de Poisson avec cette fois cependant une petite nuance: le résultat ne donnera pas nécessairement un nombre entier.

Si par exemple nous prenons (toujours avec la version française de Microsoft Excel 14.0.6123):

$$=1-\text{LOI.KHIDEUX.N}(2*20;2*(15+1);\text{VRAI})=0.156513135 \quad (7.831)$$

La question est alors de trouver l'écriture pour l'inverse... Celle-ci est alors donnée par:

$$=\text{KHIDEUX.INVERSE}(1-0.156513135;2*20)/2=15.53194258 \quad (7.832)$$

Finalement, l'écriture de l'inverse est assez naturelle. Ainsi, le "**test de Poisson à 1 échantillon**" à un niveau α donné en unilatéral droite peut s'écrire:

$$k \leq \text{KHIDEUX.INVERSE}(1-\alpha; 2 * (\text{nombre de mesures} + 1)) / 2 \quad (7.833)$$

Soit formellement:

$$\lambda \leq \frac{\chi^2_{1-\alpha} (2 \cdot (n_{\text{imposé}} + 1))}{2} \quad (7.834)$$

Attention cependant à une chose! Il semblerait que certains logiciels de statistiques approximent parfois un peu abusivement la loi de Poisson par une loi Normale. Dès lors, l'intervalle unilatéral se calcule à partir de:

$$\mu \leq \bar{X} + Z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \quad (7.835)$$

Mais avec la loi de Poisson, nous avons:

$$\begin{aligned}
 \mu &= \sigma^2 = \lambda \\
 \bar{X} &= \hat{\lambda}
 \end{aligned}
 \quad (7.836)$$

Il vient alors:

$$\lambda \leq \hat{\lambda} + Z_{1-\alpha} \sqrt{\frac{\lambda_{\text{imposé}}}{n}} \quad (7.837)$$

Exemple:

Une société fabrique des télévisions en quantité constante et a mesuré le nombre d'appareil défectueux produits chaque trimestre pendant les dix dernières années (donc 4 fois 10 mesures). La direction décide que le nombre maximum acceptable d'unités défectueuses est de 20 par trimestre et souhaite déterminer si l'usine satisfait à ces exigences (sous l'hypothèse que la distribution des défectueux suive une loi de Poisson) à un niveau de confiance de 5%.

Les 40 mesures nous donnent une moyenne de:

$$\bar{X} = \hat{\lambda} = 17.825 \quad (7.838)$$

Nous avons alors avec l'approximation grossière:

$$\lambda \leq 17.825 + Z_{1-5\%} \sqrt{\frac{20}{40}} \quad (7.839)$$

Soit dans un tableur comme la version française de Microsoft Excel 14.0.6123:

$$\begin{aligned} &= \text{LOI.NORMALE.STANDARD.INVERSE.N}(1-5\%)*\text{RACINE}(20/40)+17.825 \\ &\cong 18.988 \end{aligned} \quad (7.840)$$

ou:

$$\lambda \leq \frac{\chi^2_{1-5\%}(2 \cdot (20+1))}{2} \quad (7.841)$$

Soit dans un tableur comme la version française de Microsoft Excel 11.8346:

$$= \text{KHIDEUX.INVERSE}(1-5\%; 2*(20+1))/2 \cong 14.072 \quad (7.842)$$

Dans les deux cas, nous sommes en-dessous de la moyenne imposée de 20 (donc on rejette l'hypothèse nulle comme quoi le nombre de défauts est supérieur ou égal à 20). Bien évidemment, il est possible pour chacune des méthodes de déterminer quelle devrait être la probabilité cumulée (niveau de confiance) qui nous amène à la limite des 20 (donc la p -value en d'autres termes). Avec la première méthode (approximation normale), la p -value est de 0.104%.

Évidemment, dans le cas bilatéral, nous avons:

$$\frac{\chi^2_{\alpha/2}(2 \cdot (n+1))}{2} \leq \lambda \leq \frac{\chi^2_{1-\alpha/2}(2 \cdot (n+1))}{2} \quad (7.843)$$

Exemple:

Une compagnie d'aviation a eu 2 deux crashes en 1'000'000 de vols (événement très rare). Quelle est l'intervalle de confiance en bilatéral à 95% sachant qu'au niveau mondial le nombre d'accident par millions est de 0.4.

Nous avons alors:

$$\frac{\chi^2_{5\%/2}(2 \cdot (2+1))}{2} \leq \lambda \leq \frac{\chi^2_{1-5\%/2}(2 \cdot (2+1))}{2} \quad (7.844)$$

Soit pour la borne supérieure avec un tableur comme la version française de Microsoft Excel 11.8346:

$$=\text{LOI.KHIDEUX.INVERSE}(1-5\%/2; 2*(2+1))/2=7.224 \quad (7.845)$$

et pour la borne inférieure:

$$= \text{LOI.KHIDEUX.INVERSE}(1-5\%/2; 2*(2+1))/2 = 0.618 \quad (7.846)$$

Donc statistiquement, cette compagnie d'avion est moins sûre que l'ensemble des compagnies.

8.6.5. TEST DE POISSON (2 ÉCHANTILLONS)

Nous venons de voir que:

$$\frac{\chi^2_{\alpha/2} (2 \cdot (n+1))}{2} \leq \lambda \leq \frac{\chi^2_{1-\alpha/2} (2 \cdot (n+1))}{2} \quad (7.847)$$

Or, en suivant le même raisonnement que celui qui nous a amené à construire les tests de comparaison des moyennes suivant:

$$(\bar{X}_2 - \bar{X}_1) - Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_2 - \mu_1 \leq (\bar{X}_2 - \bar{X}_1) + Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (7.848)$$

ou son équivalent avec la loi de Student quand l'écart-type vrai n'est pas connu et en utilisant le fait que nous avons démontré que la loi de Poisson est stable par l'addition (et donc aussi par la soustraction), que la loi de Gamma était aussi stable par l'addition (et donc aussi par la soustraction) et la loi du Khi-deux aussi puisque ce n'est qu'un cas particulier de la loi Gamma. Nous aurions tendance à écrire un peu vite:

$$\frac{\chi^2_{\alpha/2} (2(n_2 + 1) - 2(n_1 + 1))}{2} \leq \lambda_2 - \lambda_1 \leq \frac{\chi^2_{1-\alpha/2} (2(n_2 + 1) - 2(n_1 + 1))}{2} \quad (7.849)$$

Et au fait cela constitue un piège!! Car la loi du Khi-deux a un support qui est défini comme étant strictement positif et l'intervalle de confiance peut naturellement avoir la borne de gauche qui est négative. Une solution consiste alors à utiliser le test de la différence de deux proportions que nous avons déjà étudié plus haut:

$$Z \cong \frac{(\hat{p}_X - \hat{p}_Y) - \mu}{\sqrt{\frac{\hat{p}_X \cdot (1 - \hat{p}_X)}{n_1} + \frac{\hat{p}_Y \cdot (1 - \hat{p}_Y)}{n_2}}} \quad (7.850)$$

À condition bien évidemment que les conditions permettant d'approcher le test par une loi Normale soient satisfaites (les proportions doivent être inférieures typiquement à 0.1 et les n supérieurs à 50).

Exemple:

Une compagnie d'aviation a eu 2 deux crashes en 1'000'000 de vols (événement très rare). Une autre compagnie a eu 3 crashes en 1'200'000 vols. Quel est l'intervalle de confiance en bilatéral à 95% en supposant que la différence est nulle.

Les proportions sont respectivement:

$$\hat{p}_X = \frac{2}{n_1} = \frac{2}{1'000'000} = 0.000002 \text{ et } \hat{p}_Y = \frac{3}{n_2} = \frac{3}{1'200'000} = 0.0000025 \quad (7.851)$$

Notons:

$$\begin{aligned} \Delta &= \hat{p}_X - \hat{p}_Y = -0.0000005 \\ \hat{q}_X &= 1 - \hat{p}_X \\ \hat{q}_Y &= 1 - \hat{p}_Y \end{aligned} \quad (7.852)$$

Nous avons alors:

$$\Delta - Z_{5\%/2} \sqrt{\frac{\hat{p}_X \hat{q}_X}{n_1} + \frac{\hat{p}_Y \hat{q}_Y}{n_2}} \leq \mu \leq \Delta + Z_{5\%/2} \sqrt{\frac{\hat{p}_X \hat{q}_X}{n_1} + \frac{\hat{p}_Y \hat{q}_Y}{n_2}} \quad (7.853)$$

ce qui donne un intervalle de confiance pour la différence de proportion théorique attendue:

$$-0.000004461 \leq \mu \leq 0.000003461 \quad (7.854)$$

et donc comme -0.0000005 est dans cet intervalle, nous acceptons l'hypothèse comme quoi la différence des proportions n'est pas significative au seuil de 5%.

Donc pour résumer un peu les convergences de lois dans tous ces différents tests et intervalles que nous avons vu jusqu'à maintenant, nous proposons au lecteur le schéma suivant qui clarifiera peut-être plus ou moins bien les choses:

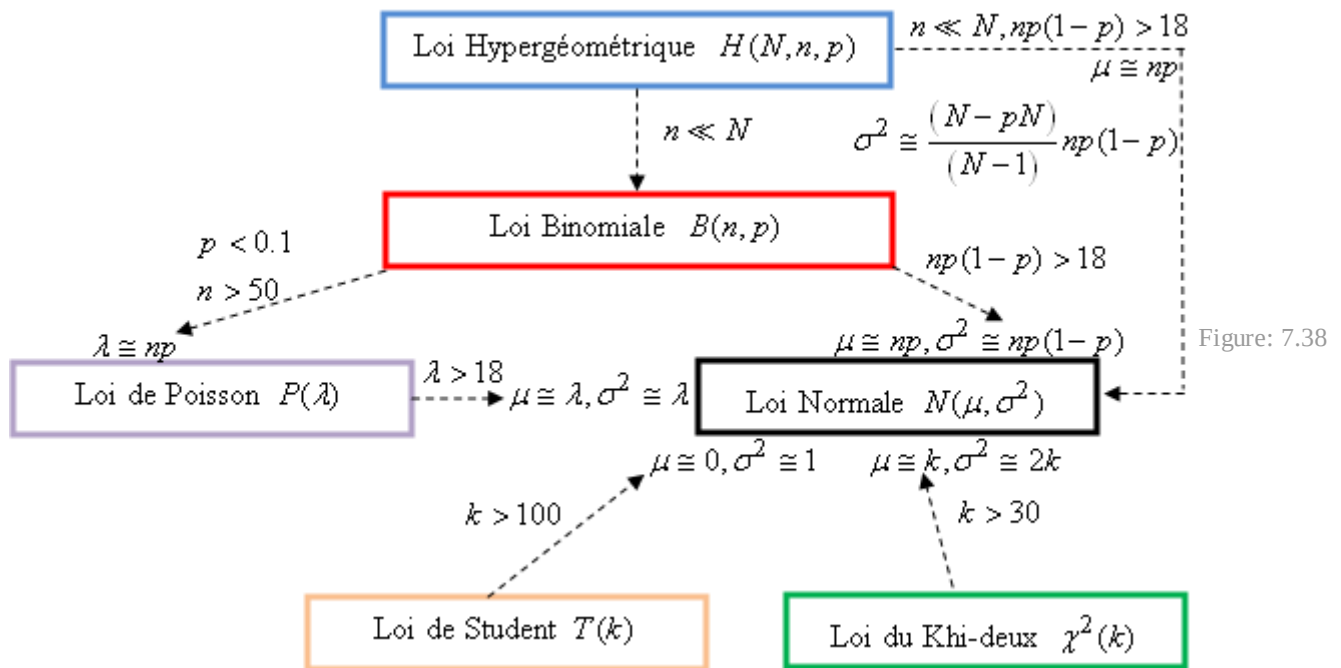


Figure: 7.38

- Convergence des différentes lois usuelles en inférence statistique élémentaire

Et aussi ce tableau où toutes les relations ont été démontrées en détail plus haut et certains déjà utilisées (d'autres le seront plus loin):

Statistique d'échantillonnage	Moyenne de la statistique	Écart-type de la statistique
Moyenne (population infinie)	μ	$\frac{\sigma}{\sqrt{n}}$
Moyenne (population finie)	μ	$\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$
Proportion (population finie)	p	$\sqrt{\frac{p(1-p)}{n}}$
Proportion (population infinie)	p	$\sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$
$\hat{\sigma}^2$ (population infinie*)	$\frac{n-1}{n} \sigma^2$	$\sqrt{\frac{2(n-1)}{n^2}} \sigma$

Tableau: 7.11 - Tableau des statistiques d'échantillonnage démontrées et utilisées en partie jusqu'à maintenant

*: Pour autant que la population parente soit distribuée normalement.

9. INTERVALLE DE CONFIANCE/TOLÉRANCE/PRÉDICTION

Nous allons ici, afin d'éviter une confusion fréquente et avant de passer à d'autres sujets plus complexes, comparer l'intervalle de confiance, l'intervalle de tolérance (souvent appelé "intervalle de fluctuation" dans certains programmes scolaires) et enfin l'intervalle de prédiction.

Définitions:

D1. "**L'intervalle de tolérance**" (ou "**intervalle de fluctuation**") est un intervalle contenant un certain pourcentage (souvent 68.26, 95.44 ou 99.73%) des individus d'une population de mesures.

D2. "**L'intervalle de confiance**" pour un échantillon de moyenne $\bar{X}$ contient l'intervalle de valeur à un niveau de confiance donné (souvent 90, 95 ou 99%) de l'espérance μ (moyenne vraie) de la population.

D3. "**L'intervalle de prédiction**" permet de déterminer un intervalle d'une valeur individuelle basée sur la connaissance de la moyenne échantillonnale et de l'écart-type de la population.

Un exemple valant mieux assez souvent mieux que mille mots, prenons le cas où la moyenne et l'écart-type de prix de 49 DVD sont:

$$\bar{X} = 31.5531 \quad S = \hat{\sigma} = 0.7992 \quad (7.855)$$

Nous avons alors:

$$\begin{aligned} [\bar{X} \pm S] &= [30.8, 32.4] \\ [\bar{X} \pm 2S] &= [30.0, 33.2] \quad (7.856) \\ [\bar{X} \pm 3S] &= [29.2, 34.2] \end{aligned}$$

correspondant respectivement à des intervalles de tolérance selon une loi Normale de 68.26, 95.44 et

99.73%.

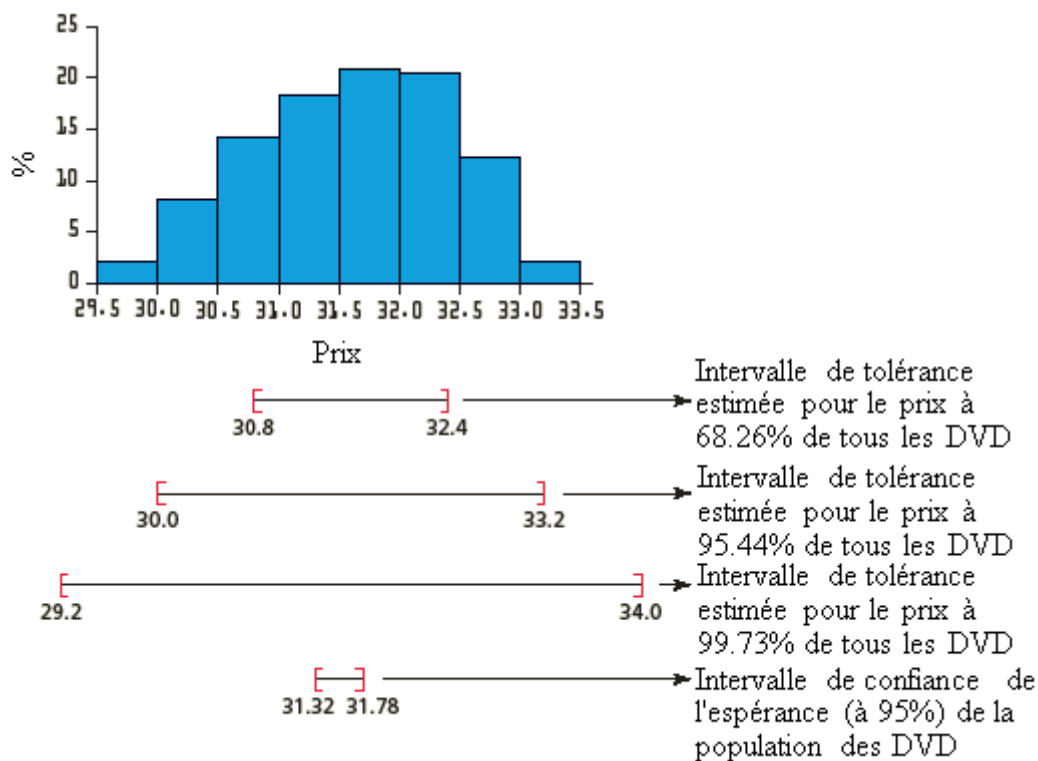
Par contre, un intervalle de confiance à 95% basé sur la relation démontrée plus haut:

$$\bar{X} - \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \quad (7.857)$$

Donne:

$$31.32 \leq \mu \leq 31.78 \quad (7.858)$$

Donc 95% de probabilité cumulée que la moyenne vraie (espérance) se trouve comprise entre 31.32 et 31.78.



Histogramme de l'échantillon des prix de 49 DVD

Maintenant passe à une notion qui curieusement est rarement traitée dans les ouvrages de statistiques. L'idée de l'intervalle de prédiction est de plutôt que de s'intéresser à l'intervalle de confiance de l'espérance basé sur une moyenne expérimentale, d'utiliser cette moyenne expérimentale (échantillonnale) comme base pour prévoir l'intervalle d'une unique valeur (et non d'une moyenne!).

Nous allons donc nous intéresser à la différence entre la moyenne et une valeur ponctuelle:

$$\bar{x} - x \quad (7.859)$$

que nous supposons proche de zéro (il vaut mieux pour avoir un produit fiable et passer les tests d'autorisation des ventes). Concernant la variance, ce qui nous intéresse ce n'est plus simplement l'écart-type de la moyenne mais l'écart-type de la différence... et comme l'échantillon est indépendant de la valeur unique nous avons:

$$\sigma_{\bar{x}-x}^2 = \sigma_{\bar{x}}^2 + \sigma_x^2 = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2 \left(1 + \frac{1}{n}\right) \quad (7.860)$$

Donc nous pouvons écrire qu'en première approximation:

$$Z \cong \frac{\bar{x} - x}{\sigma_{\bar{x}-x}} = \frac{\bar{x} - x}{\sqrt{\sigma^2 \left(1 + \frac{1}{n}\right)}} = \frac{\bar{x} - x}{\sigma \sqrt{\left(1 + \frac{1}{n}\right)}} \quad (7.861)$$

Et bien évidemment suite à ce que nous avons vu:

$$T(n-1) \cong \frac{\bar{x} - x}{S \sqrt{\left(1 + \frac{1}{n}\right)}} \quad (7.862)$$

Et donc in extenso nous pouvons construire l'intervalle de prédiction:

$$\bar{x} - T_{\alpha/2}(n-1) S \sqrt{\left(1 + \frac{1}{n}\right)} \leq x \leq \bar{x} + T_{\alpha/2}(n-1) S \sqrt{\left(1 + \frac{1}{n}\right)} \quad (7.863)$$

10. LOI FAIBLE DES GRANDS NOMBRES

Nous allons maintenant nous attarder sur une relation très intéressante en statistiques qui permet de dire pas mal de choses tout en ayant peu de données et ce quelle que soit la loi considérée (ce qui est pas mal quand même!). C'est une propriété très utilisée en simulation statistique par exemple dans le cadre de l'utilisation de Monte-Carlo.

Soit une variable aléatoire à valeurs dans $\mathbb{R}^+$. Alors nous allons démontrer la relation suivante appelée "**inégalité de Markov**":

$$\forall \lambda > 0 : P(X \geq \lambda) \leq \frac{E(X)}{\lambda} \quad (7.864)$$

avec $E(X) \leq \lambda$ dans le contexte particulier des probabilités.

En d'autres termes, nous proposons de démontrer que la probabilité qu'une variable aléatoire soit plus grande ou égale qu'une valeur λ est inférieure ou égale à son espérance divisée par la valeur considérée λ et ce quelle que soit la loi de distribution de la variable aléatoire X !

Démonstration:

Notons les valeurs de X par $\{x_1, \dots, x_n\}$, où $0 \leq x_1 < x_2 < \dots < x_n$ (c'est-à-dire triées par ordre croissant) et posons $x_0 = 0$. Nous remarquons d'abord que l'inégalité est triviale au cas où $\lambda \geq x_n \geq 0$.

Effectivement, comme X ne peut être compris qu'entre 0 et x_n par définition alors la probabilité qu'il soit supérieur à x_n est nulle. En d'autres termes:

$$P(X \geq \lambda) = 0 \quad (7.865)$$

et X étant positif, $E(X)$ l'est aussi, d'où l'inégalité pour ce cas particulier dans un premier temps.

Sinon, nous avons $0 < \lambda \leq x_n$ et il existe alors un $k \in \{1, \dots, n\}$ tel que $x_{k-1} < \lambda \leq x_k$. Donc:

$$E(X) = \sum_{i=1}^n x_i p_i \geq \sum_{i=k}^n x_i p_i \geq \lambda \sum_{i=k}^n p_i = \lambda \cdot P(X \geq \lambda) \quad (7.866)$$

□C.Q.F.D.

Exemple:

Nous supposons que le nombre de pièces sortant d'une usine donnée en l'espace d'une semaine est une variable aléatoire d'espérance 50. Si nous souhaitons estimer la probabilité cumulée que la production dépasse 75 pièces nous appliquerons simplement:

$$P(X \geq 75) \leq \frac{E(X)}{\lambda} = \frac{50}{75} = 0.6666 = 66.66\% \quad (7.867)$$

Considérons maintenant une sorte de généralisation de cette inégalité appelée "**inégalité de Bienaymé-Tchebychev**" (abrégée "**inégalité BT**") qui va nous permettre d'obtenir un résultat très très très intéressant et important un peu plus bas.

Considérons une variable aléatoire X . Alors nous allons démontrer l'inégalité de Bienaymé-Tchebychev suivante:

$$\forall \varepsilon > 0 : P(|X - E(X)| \geq \varepsilon) \leq \frac{V(X)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2} \quad (7.868)$$

qui exprime le fait que plus l'écart-type est petit, plus la probabilité que la variable aléatoire X s'éloigne de son espérance est faible.

Démonstration:

Nous obtenons cette inégalité en écrivant d'abord:

$$\forall \varepsilon > 0 : P(|X - E(X)| \geq \varepsilon) = P([X - E(X)]^2 \geq \varepsilon^2) \quad (7.869)$$

où le choix du carré va nous servir pour une simplification future.

Puis en appliquant l'inégalité de Markov (comme quoi c'est quand même utile...) à la variable aléatoire $Y = [X - E(X)]^2$ avec $\lambda = \varepsilon^2$ il vient automatiquement:

$$P(Y \geq \varepsilon^2) = P([X - E(X)]^2 \geq \varepsilon^2) \leq \frac{E(Y)}{\varepsilon^2} = \frac{E([X - E(X)]^2)}{\varepsilon^2} \quad (7.870)$$

Ensuite, en utilisant la définition de la variance:

$$E\left([X - E(X)]^2\right) = E\left([X - \mu_X]^2\right) = V(X) = \sigma^2 \quad (7.871)$$

Nous obtenons bien:

$$P(Y \geq \varepsilon^2) \leq \frac{E\left([X - E(X)]^2\right)}{\varepsilon^2} = \frac{V(X)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2} \quad (7.872)$$

□C.Q.F.D.

Si nous posons:

$$\varepsilon = t\sigma \quad (7.873)$$

l'inégalité s'écrit aussi:

$$P(|X - E(X)| \geq t\sigma) \leq \frac{1}{t^2} \quad (7.874)$$

et exprime que la probabilité qu'afin que X s'éloigne de son espérance de plus que t fois son écart-type, est inférieure à $1/t^2$. Il y a, en particulier, moins de 1 chance sur 9 pour que X s'éloigne de son espérance de plus de trois fois l'écart-type.

Exemple:

Nous reprenons l'exemple où le nombre de pièces sortant d'une usine donnée en l'espace d'une semaine est une variable aléatoire d'espérance 50. Nous supposons en plus que la variance de la production hebdomadaire est de 25. Nous cherchons à calculer la probabilité que la production de la semaine prochaine soit comprise entre 40 et 60 pièces.

Pour calculer ceci il faut d'abord se souvenir que l'inégalité de BT est basée en parties sur le terme $|X - E(X)|$ donc nous avons:

$$|40 - 50| = |60 - 50| = 10 \quad (7.875)$$

donc l'inégalité de BT nous permet bien de travailler sur des intervalles égaux en valeur absolue ce qui s'écrit aussi:

$$-10 \leq X - E(X) \leq 10 \quad (7.876)$$

Ensuite, ne reste plus qu'à appliquer simplement l'inégalité numériquement:

$$P(|X - E(X)| \geq \varepsilon) = P(|X - E(X)| \geq 10) = \frac{V(X)}{\varepsilon^2} = \frac{25}{10^2} = 0.25 = 25\% \quad (7.877)$$

Ces deux dernières inégalités vont nous permettre d'obtenir une relation très importante et puissante que nous appelons la "**loi faible des grands nombres**" (L.F.G.N.) ou encore "**théorème de Khintchine**".

Considérons une variable aléatoire X admettant une variance et $(X_n)_{n \in \mathbb{N}}$ une suite de variables

aléatoires indépendantes (donc non corrélées deux-deux) de même loi que X et ayant toutes les mêmes espérances μ et les mêmes écarts-types σ .

Ce que nous allons montrer est que si nous mesurons une même quantité aléatoire X_n de même loi au cours d'une suite d'expériences indépendantes (alors dans ce cas, nous disons techniquement que la suite $(X_n)_{n \in \mathbb{N}^*}$ de variables aléatoires est définie sur le même espace probabilisé), alors la moyenne arithmétique des valeurs observées va se stabiliser sur l'espérance de X quand le nombre de mesures est infiniment élevé.

De manière formelle ceci s'exprime sous la forme:

$$\forall \varepsilon > 0, P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2} \rightarrow 0 \quad (7.878)$$

lorsque $n \rightarrow +\infty$ c'est cela le résultat très important dont nous faisons mention plus haut! L'estimateur empirique de la moyenne tend donc pour toute loi vers l'espérance vraie si n est grand! Donc de par la même nous assurons que la moyenne empirique est un estimateur convergent de l'espérance! Ce résultat (assez intuitif) est parfois appelé "**théorème fondamental de Monte Carlo**" car il est au centre du principe des simulations du même nom (cf. chapitre de Méthodes Numériques) qui ont une importance cruciale dans l'étude des statistiques avancées.

Donc en d'autres termes la probabilité cumulée que la différence entre la moyenne arithmétique et l'espérance des variables aléatoires observées soit comprise dans un intervalle autour de la moyenne tend vers zéro quand le nombre de variables aléatoires mesurées tend vers l'infini (ce qui est finalement intuitif).

Ce résultat nous permet d'estimer l'espérance mathématique en utilisant la moyenne empirique (arithmétique) calculée sur un très grand nombre d'expériences.

Démonstration:

Nous utilisons l'inégalité de Bienaymé-Tchebychev pour la variable aléatoire (cette relation s'interprète difficilement mais permet d'avoir le résultat escompté):

$$Y = \frac{1}{n} \sum_{i=1}^n X_i \quad (7.879)$$

Et nous calculons d'abord en utilisant les propriétés mathématiques de l'espérance que nous avons démontrées plus haut:

$$E(Y) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} E\left(\sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} n\mu = \mu \quad (7.880)$$

et dans un deuxième temps en utilisant les propriétés mathématiques de la variance aussi déjà démontrées plus haut:

$$\begin{aligned}
 V(Y) &= V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^n X_i\right) \\
 &= \frac{1}{n^2} \left(\sum_{i=1}^n V(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{cov}(X_i, X_j) \right) \quad (7.881)
 \end{aligned}$$

et puisque nous avons supposé les variables non corrélées entre elles alors la covariance est nulle dès lors:

$$V(Y) = \frac{1}{n^2} \left(\sum_{i=1}^n V(X_i) \right) = \frac{1}{n^2} \left(\sum_{i=1}^n \sigma^2 \right) = \frac{1}{n^2} n \sigma^2 = \frac{1}{n} \sigma^2 \quad (7.882)$$

Donc en injectant cela dans l'inégalité BT:

$$P\left([X - E(X)]^2 \geq \varepsilon^2\right) \leq \frac{V(X)}{\varepsilon^2} \quad (7.883)$$

nous avons alors:

$$P\left([Y - E(Y)]^2 \geq \varepsilon^2\right) \leq \frac{V(Y)}{\varepsilon^2} \quad (7.884)$$

qui devient:

$$P\left(\left[\left(\frac{1}{n} \sum_{i=1}^n X_i\right) - \mu\right]^2 \geq \varepsilon^2\right) \leq \frac{\sigma^2}{n\varepsilon^2} \quad (7.885)$$

et l'inégalité tend bien vers zéro quand n au dénominateur tend vers l'infini.

□C.Q.F.D.

Signalons que cette dernière relation est souvent notée dans certains ouvrages et conformément à ce que nous avons vu au début de ce chapitre:

$$P(|M_n - m_X| > \varepsilon) \leq \frac{\sigma_X^2}{n\varepsilon^2} \quad (7.886)$$

ou encore:

$$P(|M_n - m_X| < \varepsilon) \geq 1 - \frac{\sigma_X^2}{n\varepsilon^2} \quad (7.887)$$

Donc, pour $\forall \varepsilon > 0$:

$$\lim_{n \rightarrow +\infty} P(|M_n - m_X| < \varepsilon) = 1 \quad (7.888)$$

11. FONCTION CARACTÉRISTIQUE

Avant de donner une démonstration à la manière de l'ingénieur du théorème central limite, introduisons d'abord le concept de "**fonction caractéristique**" qui tient une place centrale en statistiques.

D'abord, rappelons que la transformée de Fourier est donnée dans sa version physicienne par (cf. chapitre de Suites et Séries) la relation:

$$F(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x) e^{-ikx} dx \quad (7.889)$$

Rappelons que la transformation de Fourier est un analogue de la théorie des séries de Fourier pour les fonctions non périodiques, et permet de leur associer un spectre en fréquences. Au facteur près, il s'agit d'une "**transformée de Laplace bilatérale**" donnée par:

$$L_x[f(x)] = F(p) = \int_{-\infty}^{+\infty} f(x) e^{-px} dx \quad (7.890)$$

avec p qui est la variable complexe donnée dans le cas présent par (la partie réelle est nulle puisque la transformée de Fourier n'est que le cas particulier d'une transformée de Laplace dont la partie réelle de la variable est nulle: dont faire une transformée de Fourier c'est faire une transformée de Laplace sur l'axe des complexes uniquement):

$$p = ik \quad (7.891)$$

Nous souhaitons maintenant démontrer que si:

$$g(x) = f'(x) \text{ alors } G(k) = ikF(k) \quad (7.892)$$

En d'autres termes, nous cherchons une expression simplifiée de la transformée de Fourier de la dérivée de $f(x)$.

Démonstration:

Nous partons donc de:

$$G(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f'(x) e^{-ikx} dx \quad (7.893)$$

Une intégration par parties donne:

$$\int_{-\infty}^{+\infty} f'(x) e^{-ikx} dx = f(x) \cdot e^{-ikx} \Big|_{x=-\infty}^{x=+\infty} + ik \int_{-\infty}^{+\infty} f(x) e^{-ikx} dx \quad (7.894)$$

En imposant que, f tend vers zéro à l'infini, nous avons alors:

$$f(x) \cdot e^{-ikx} \Big|_{x=-\infty}^{x=+\infty} = 0 \quad (7.895)$$

et:

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f'(x) e^{-ikx} dx = \frac{ik}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x) e^{-ikx} dx = ik \cdot F(k) \quad (7.896)$$

C'est le premier résultat dont nous avons besoin.

□C.Q.F.D.

Maintenant, démontrons que si:

$$g(x) = xf'(x) \text{ alors } G(k) = iF'(k) \quad (7.897)$$

Démonstration:

Nous partons donc de:

$$\begin{aligned} G(k) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} xf'(x) e^{-ikx} dx = i \frac{d}{dk} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x) e^{-ikx} dx \right) = iF'(k) \\ &= i \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x) \frac{d}{dk} e^{-ikx} dx = i \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x) (-ix) e^{-ikx} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} xf'(x) e^{-ikx} dx \end{aligned} \quad (7.898)$$

C'est le deuxième résultat dont nous avons besoin.

□C.Q.F.D.

Maintenant effectuons le calcul de la transformée de Fourier de la loi Normale centrée-réduite (ce choix n'est pas innocent...):

$$y = f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (7.899)$$

Nous savons que cette dernière relation est trivialement solution de l'équation différentielle (ou bien elle vérifie):

$$y' = -xy \quad (7.900)$$

en prenant la transformée de Fourier des deux côté de l'égalité, nous avons en utilisant les deux résultats précédents:

$$\begin{aligned} g(x) &= f'(x) \text{ alors } G(k) = ikF(k) \quad (7.901) \\ g(x) &= xf'(x) \text{ alors } G(k) = iF'(k) \end{aligned}$$

Nous avons:

$$ikF(k) = -iF'(k) \quad (7.902)$$

Ou encore:

$$-kdk = \frac{1}{F(k)} dF(k) \quad (7.903)$$

Donc après intégration:

$$\begin{aligned} \int_0^k -kdk &= \int_0^k \frac{1}{F(k)} dF(k) \Rightarrow -\frac{1}{2}k^2 \Big|_0^k = \ln(F(k)) \Big|_0^k \Rightarrow -\frac{1}{2}k^2 = \ln(F(k)) - \ln(F(0)) \\ &\Rightarrow -\frac{1}{2}k^2 = \ln\left(\frac{F(k)}{F(0)}\right) \Rightarrow e^{-\frac{1}{2}k^2} = \frac{F(k)}{F(0)} \Rightarrow F(k) = F(0)e^{-\frac{1}{2}k^2} \end{aligned} \quad (7.904)$$

Puisque:

$$F(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x) e^{-ikx} dx \quad (7.905)$$

nous avons donc:

$$F(0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x) e^{-i0x} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx \quad (7.906)$$

Nous avons démontré lors de notre étude de la loi Normale que:

$$\int_{-\infty}^{+\infty} e^{-\frac{1}{2}x^2} dx = \sqrt{2\pi} \quad (7.907)$$

Donc:

$$F(0) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx = \frac{1}{2\pi} \sqrt{2\pi} = \frac{1}{\sqrt{2\pi}} \quad (7.908)$$

Nous avons alors (résultat important!):

$$F(k) = \frac{1}{\sqrt{2\pi}} e^{-\frac{k^2}{2}} \quad (7.909)$$

Introduisons maintenant la fonction caractéristique telle que définie par les statisticiens:

$$\phi_X(k) = E[e^{ikX}] = \int_{-\infty}^{+\infty} e^{ikx} f_X(x) dx \quad (7.910)$$

qui est un outil analytique important et puissant permettant d'analyser une somme de variables aléatoires indépendantes. De plus, cette fonction contient toutes les informations caractéristiques de la variable aléatoire X .

Il s'agit

Remarque: La notation n'est pas innocente puisque le $E[...]$ représente une espérance de la fonction de densité par rapport à l'exponentielle complexe.

Donc la fonction caractéristique de la variable aléatoire normale centrée réduite de distribution:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (7.911)$$

devient simple à déterminer car:

$$\phi_X(k) = \int_{-\infty}^{+\infty} e^{ikx} f_X(x) dx = \sqrt{2\pi} F(-k) \quad (7.912)$$

raison pour laquelle la fonction caractéristique de la loi Normale centrée réduite est souvent assimilée à une simple transformée de Fourier.

Et grâce au résultat précédent:

$$F(k) = \frac{1}{\sqrt{2\pi}} e^{-\frac{k^2}{2}} \Rightarrow F(-k) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(-k)^2}{2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{k^2}{2}} \quad (7.913)$$

Donc:

$$\phi_X(k) = \frac{\sqrt{2\pi}}{\sqrt{2\pi}} e^{-\frac{k^2}{2}} = e^{-\frac{k^2}{2}} \quad (7.914)$$

qui est le résultat dont nous avons besoin pour le théorème central limite.

Mais avant cela, regardons d'un peu plus près cette fonction caractéristique:

$$\phi_X(\omega) = E[e^{i\omega x}] = \int_{-\infty}^{+\infty} e^{i\omega x} f_X(x) dx \quad (7.915)$$

En développement de Maclaurin nous avons (cf. chapitre Suites et Séries) et en changeant un peu les notations:

$$\phi_X(\omega) = \int_{-\infty}^{+\infty} \left[\sum_{k=0}^{+\infty} \frac{(i\omega)^k}{k!} x^k f_X(x) \right] dx \quad (7.916)$$

et en intervertissant la somme et l'intégrale, nous avons:

$$\phi_X(\omega) = \sum_{k=0}^{+\infty} \frac{(i\omega)^k}{k!} \int_{-\infty}^{+\infty} x^k f_X(x) dx = \sum_{k=0}^{+\infty} \frac{(i\omega)^k}{k!} M_X^k \quad (7.917)$$

Cette fonction caractéristique contient donc tous les moments (terme général utilisé pour l'écart-type et l'espérance) de X .

12. THÉORÈME CENTRAL LIMITE

Le théorème central limite est un ensemble de résultats du début du 20ème siècle sur la convergence faible d'une suite de variables aléatoires en probabilité. Intuitivement, d'après ces résultats, toute somme (implicitement: la moyenne de ses variables) de variables aléatoires indépendantes et identiquement distribuées tend vers une certaine variable aléatoire. Le résultat le plus connu et le plus important est simplement appelé "théorème central limite" qui concerne une somme de variables aléatoires indépendantes avec variance existante dont le nombre tend vers l'infini et c'est celui-ci que nous allons démontrer de manière heuristique ici.

Dans le cas le plus simple, considéré ci-dessous pour la démonstration du théorème, ces variables sont continues, indépendantes et possèdent la même moyenne et la même variance. Pour tenter d'obtenir un résultat fini, il faut centrer cette somme en lui soustrayant sa moyenne et la réduire en la divisant par son écart-type. Sous des conditions assez larges, la loi de probabilité (de la moyenne) converge alors vers une loi Normale centrée réduite. L'omniprésence de la loi Normale s'explique par le fait que de nombreux phénomènes considérés comme aléatoires sont dus à la superposition de causes nombreuses.

Ce théorème de probabilités possède donc une interprétation en statistique mathématique. Cette dernière associe une loi de probabilité à une population. Chaque élément extrait de la population est donc considéré comme une variable aléatoire et, en réunissant un nombre n de ces variables supposées indépendantes, nous obtenons un échantillon. La somme de ces variables aléatoires divisée par n donne une nouvelle variable nommée la moyenne empirique. Celle-ci, une fois réduite, tend vers une variable Normale réduite lorsque n tend vers l'infini comme nous le savons.

Le théorème central limite nous dit à quoi il faut s'attendre en matière de sommes de variables aléatoires indépendantes. Mais qu'en est-il des produits? Eh bien, le logarithme d'un produit (à facteurs strictement positifs) est la somme des logarithmes des facteurs, de sorte que le logarithme d'un produit de variables aléatoires (à valeurs strictement positives) tend vers une loi Normale, ce qui entraîne une loi log-Normale pour le produit lui-même.

En elle-même, la convergence vers la loi Normale ("**normalité asymptotique**") de nombreuses sommes de variables aléatoires lorsque leur nombre tend vers l'infini n'intéresse que le mathématicien. Pour le praticien, il est intéressant de s'arrêter un peu avant la limite: la somme d'un grand nombre de ces variables est presque gaussienne, ce qui fournit une approximation souvent plus facilement utilisable que la loi exacte.

En s'éloignant encore plus de la théorie, on peut dire que bon nombre de phénomènes naturels sont dus à la superposition de causes nombreuses, plus ou moins indépendantes. Il en résulte que la loi Normale les représente de manière raisonnablement efficace.

A l'inverse, on peut dire qu'aucun phénomène concret n'est vraiment Gaussien car il ne peut dépasser certaines limites, en particulier s'il est à valeurs positives.

Démonstration:

Soit $\{X_i\}_{i=1, \dots, \infty}$ une suite (échantillon) de variables aléatoires continues (dans notre démonstration simplifiée...), indépendantes (mesures de phénomènes physiques ou mécaniques indépendants par

exemple) et identiquement distribuées, dont la moyenne μ_X et l'écart-type σ_X existent (ce qui signifie que le théorème central limite fonctionne que pour les phénomènes à variance finie!!!).

Nous avons vu au début de ce chapitre que:

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{M_n - \mu}{\sigma/\sqrt{n}} \quad (7.918)$$

sont les mêmes expressions d'une variable centrée réduite générée à l'aide d'une suite de n variables aléatoires identiquement distribuées qui par construction a donc une moyenne nulle et une variance unitaire:

$$E(Z_n) = 0 \text{ et } V(Z_n) = 1 \quad (7.919)$$

Développons la première forme de l'égalité antérieure (les 2 sont de toute façon égales!):

$$Z_n = \frac{1}{\sqrt{n}\sigma} \left[\sum_{i=1}^n (X_i - \mu_X) \right] = \frac{1}{\sqrt{n}\sigma} \left[\left(\sum_{i=1}^n X_i \right) - n\mu \right] \quad (7.920)$$

Maintenant utilisons la fonction caractéristique de la loi Normale centrée-réduite (nous allégeons par la même occasion l'écriture des estimateurs de la moyenne et de l'écart-type):

$$\phi_{Z_n}(\omega) = E[e^{i\omega Z_n}] = E \left[\exp \left(i\omega \frac{\sum_{i=1}^n (X_i - \mu)}{\sqrt{n}\sigma} \right) \right] \quad (7.921)$$

Comme les variables aléatoires X_i sont indépendantes et identiquement distribuées, il vient:

$$\phi_{Z_n}(\omega) = E \left[\exp \left(i\omega \frac{X_1 - \mu}{\sqrt{n}\sigma} \right) \right] \cdots E \left[\exp \left(i\omega \frac{X_n - \mu}{\sqrt{n}\sigma} \right) \right] = \left\{ E \left[\exp \left(i\omega \frac{X - \mu}{\sqrt{n}\sigma} \right) \right] \right\}^n \quad (7.922)$$

Un développement de Taylor (cf. chapitre de Suites Et Séries) du terme entre accolades donne au troisième ordre (développement en série de Maclaurin de l'exponentielle):

$$E \left[\exp \left(i\omega \frac{X - \mu}{\sqrt{n}\sigma} \right) \right] \cong E \left[1 + i \frac{\omega}{\sqrt{n}\sigma} (X - \mu) - \frac{\omega^2}{2!n\sigma^2} (X - \mu)^2 \right] \quad (7.923)$$

Finalement:

$$\begin{aligned}
\phi_{Z_n}(\omega) &\equiv \left\{ E \left[1 + i \frac{\omega}{\sqrt{n}\sigma} (X - \mu) - \frac{\omega^2}{2n\sigma^2} (X - \mu)^2 \right] \right\}^n \\
&= \left[\int_{-\infty}^{+\infty} \left(1 + i \frac{\omega}{\sqrt{n}\sigma} (X - \mu) - \frac{\omega^2}{2n\sigma^2} (X - \mu)^2 \right) f_X(x) dx \right]^n \\
&= \left[\int_{-\infty}^{+\infty} f_X(x) dx + \int_{-\infty}^{+\infty} iX \frac{\omega}{\sqrt{n}\sigma} f_X(x) dx - \int_{-\infty}^{+\infty} \mu i \frac{\omega}{\sqrt{n}\sigma} f_X(x) dx - \int_{-\infty}^{+\infty} \frac{\omega^2}{2n\sigma^2} (X - \mu)^2 f_X(x) dx \right]^n \\
&= \left[1 + \frac{i\omega}{\sqrt{n}\sigma} \int_{-\infty}^{+\infty} X f_X(x) dx - \frac{\mu i \omega}{\sqrt{n}\sigma} \int_{-\infty}^{+\infty} f_X(x) dx - \frac{\omega^2}{2n\sigma^2} \int_{-\infty}^{+\infty} (X - \mu)^2 f_X(x) dx \right]^n \\
&= \left[1 + \frac{i\omega}{\sqrt{n}\sigma} \mu - \frac{\mu i \omega}{\sqrt{n}\sigma} - \frac{\omega^2}{2n\sigma^2} \int_{-\infty}^{+\infty} (X - \mu)^2 f_X(x) dx \right]^n \\
&= \left[1 - \frac{\omega^2}{2n\sigma^2} \int_{-\infty}^{+\infty} (X - \mu)^2 f_X(x) dx \right]^n = \left[1 - \frac{\omega^2}{2n\sigma^2} V(X) \right]^n \\
&= \left[1 - \frac{\omega^2}{2n} \right]^n
\end{aligned}$$

(7.924)

Posons:

$$-\frac{\omega^2}{2n} = \frac{1}{x} \quad (7.925)$$

Nous avons alors:

$$\phi_{Z_n}(\omega) \equiv \left[1 + \frac{1}{x} \right]^{-\frac{\omega^2 x}{2}} = \left[\left(1 + \frac{1}{x} \right)^x \right]^{-\frac{\omega^2}{2}} \quad (7.926)$$

Et donc quand x tend vers l'infini (cf. chapitre d'Analyse fonctionnelle):

$$\phi_{Z_n}(\omega) \equiv \lim_{n \rightarrow +\infty} \left[\left(1 + \frac{1}{x} \right)^x \right]^{-\frac{\omega^2}{2}} = \left[\lim_{n \rightarrow +\infty} \left(1 + \frac{1}{x} \right)^x \right]^{-\frac{\omega^2}{2}} = e^{-\frac{\omega^2}{2}} \quad (7.927)$$

Nous retrouvons donc la fonction caractéristique de la loi Normale centrée réduite!

En deux mots, le Théorème Central Limite (TCL) dit que pour de grands échantillons, la somme centrée et réduite de n variables aléatoires indépendamment distribuées suit une loi Normale centrée et réduite. Et donc nous avons in extenso pour la moyenne empirique:

$$\frac{M_n - \mu}{\sigma / \sqrt{n}} = N(0, 1) \Leftrightarrow M_n = N(\mu, \sigma / \sqrt{n}) \quad (7.928)$$

Malgré l'immensité de son champ d'applications, le TCL n'est pas universel. Dans sa forme la plus simple, il impose en particulier à la variable considérée d'avoir des moments du premier et du deuxième ordre (moyenne et variance). Si tel n'est pas le cas, il ne s'applique plus.

L'exemple le plus simple d'échec du TLC est donné par la distribution de Cauchy, qui n'a ni moyenne, ni variance, et dont la moyenne empirique a toujours la même distribution (Cauchy) quelle que soit la taille de l'échantillon.

Maintenant, nous allons illustrer le théorème central limite dans le cas d'une suite (X_n) de variables aléatoires indépendantes discrètes suivant une loi de Bernoulli de paramètre 1/2.

Nous pouvons imaginer que X_n représente le résultat obtenu au n -ème lancé d'une pièce de monnaie (en attribuant le nombre 1 pour pile et 0 pour face). Notons:

$$\overline{X_n} = \frac{1}{n} \sum_{k=1}^n X_k \quad (7.929)$$

la moyenne. Nous avons pour tout n bien évidemment:

$$E(X_n) = 1/2 \quad V(X_n) = 1/4 \quad (7.930)$$

et donc:

$$E(\overline{X_n}) = \frac{1}{n} \sum_{k=1}^n 1/2 = 1/2 \quad V(\overline{X_n}) = \frac{1}{n^2} \sum_{k=1}^n 1/4 = \frac{1}{4n} \quad (7.931)$$

Après avoir centré et réduit $\overline{X_n}$ nous obtenons:

$$\frac{\overline{X_n} - E(\overline{X_n})}{\sqrt{V(\overline{X_n})}} = 2\sqrt{n}(\overline{X_n} - 1/2) \quad (7.932)$$

Notons Φ la fonction de répartition de la loi Normale centrée réduite.

Le théorème central limite nous dit que pour tout $t \in \mathbb{R}$:

$$\lim_{n \rightarrow \infty} P\left(2\sqrt{n}(\overline{X_n} - 1/2) \leq t\right) = \Phi(t) \quad (7.933)$$

A l'aide de Maple 4.00b nous avons tracé en bleu quelques graphiques de la fonction:

$$t \mapsto F_n(t) = P\left(2\sqrt{n}(\overline{X_n} - 1/2) \leq t\right) \quad (7.934)$$

pour différentes valeurs de n . Nous avons représenté en rouge la fonction Φ .

$n = 1$:

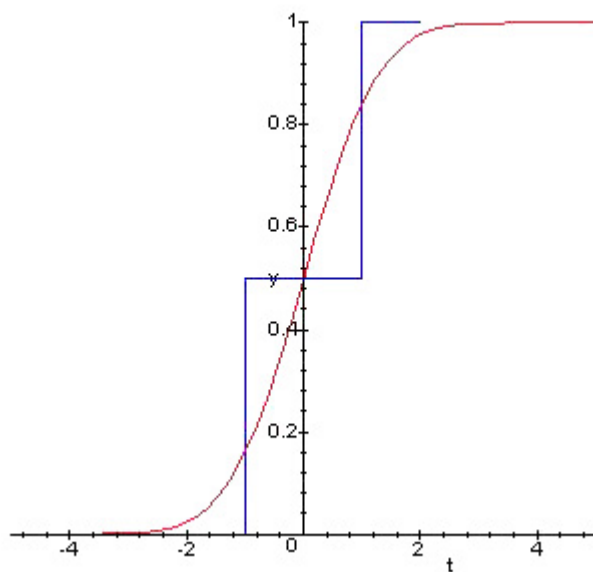


Figure: 7.39 - Première approche de la loi de Bernoulli par le loi Normale selon le TCL

$n = 2$:

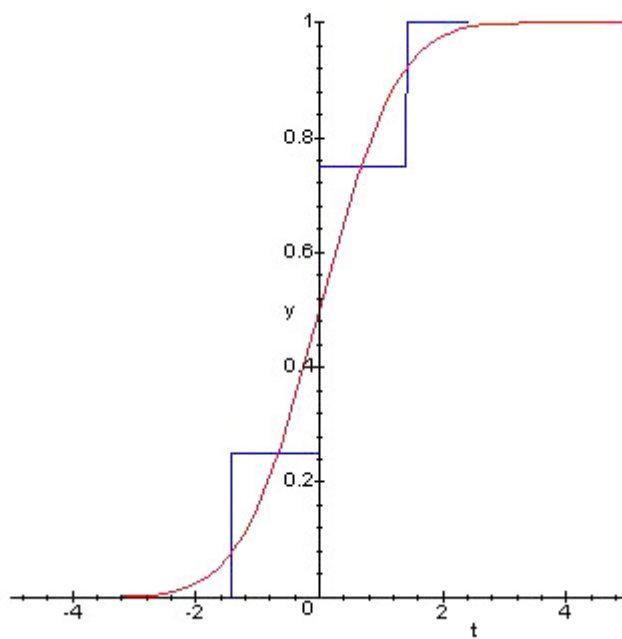


Figure: 7.40 - Deuxième approche de la loi de Bernoulli par le loi Normale selon le TCL

$n = 5$:

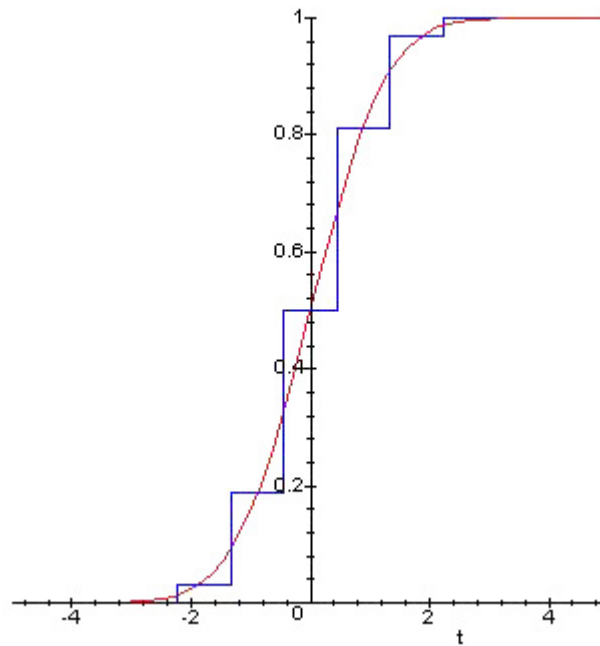


Figure: 7.41 - Troisième approche de la loi de Bernoulli par le loi Normale selon le TCL

$n = 30$:

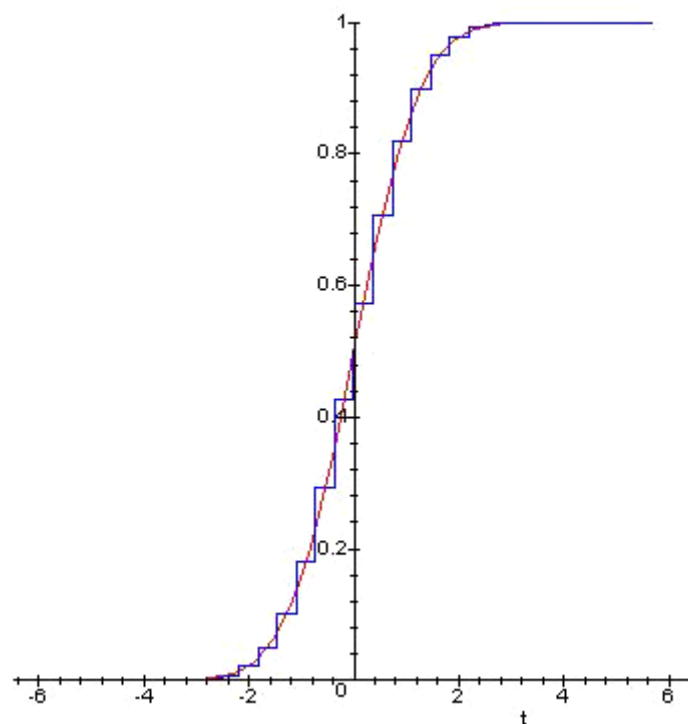


Figure: 7.42 - Quatrième approche de la loi de Bernoulli par le loi Normale selon le TCL

Ces graphiques obtenus avec Maple 4.00b à l'aide des commandes suivantes:

```
> with(stats):
> with(plots):
> e1:=plot(Heaviside(t+1)*statevalf[dcdf,binomiald[1,0.5]](trunc((t+1)/2)),t=-2..2,y=0..1,color=blue):
```

```

> e2:=plot(Heaviside(t+sqrt(2))*statevalf[dcdf,binomiald[2,0.5]]
(trunc((t*sqrt(2)+2)/2)),t=-sqrt(2)-1..sqrt(2)+1,y=0..1,color=blue):
> e3:=plot(Heaviside(t+sqrt(5))*statevalf[dcdf,binomiald[5,0.5]]
(trunc((t*sqrt(5)+5)/2)),t=-sqrt(5)-1..sqrt(5)+1,y=0..1,color=blue):
> e4:=plot(statevalf[cdf,normald](t),t=-5..5):
> e5:=plot(Heaviside(t+sqrt(30))*statevalf[dcdf,binomiald[30,0.5]]
(trunc((t*sqrt(30)+30)/2)),t=-sqrt(30)-1..sqrt(30)+1,y=0..1,color=blue):
> display({e1,e4});
> display({e2,e4});
> display({e4,e3});
> display({e5,e4});

```

montrent bien la convergence de F_n vers Φ .

En fait nous remarquons que la convergence est carrément uniforme ce qui est confirmé par le "théorème central limite de Moivre-Laplace":

Soit X_n une suite de variables aléatoires indépendantes de même loi de Bernoulli de paramètre p , $0 < p < 1$. Alors:

$$F_n(t) = P\left(2\sqrt{n}\left(\overline{X_n} - 1/2\right) \leq t\right) \quad (7.935)$$

tend uniformément vers $\Phi(t)$ sur $\mathbb{R}$ lorsque $n \rightarrow \infty$.

13. TESTS D'HYPOTHÈSE ET D'ADÉQUATION

Lors de notre étude des intervalles de confiance, rappelons que nous sommes arrivés aux quelques relations suivantes (ce n'est que l'échantillon des plus importantes démontrées plus haut!):

$$\overline{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (7.936)$$

et:

$$\frac{nS_n^2}{\chi_{\alpha/2}^2(n)} \leq \sigma^2 \leq \frac{nS_n^2}{\chi_{1-\alpha/2}^2(n)} \quad (7.937)$$

et:

$$\frac{(n-1)S^2}{\chi_{\alpha/2}^2(n-1)} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2(n-1)} \quad (7.938)$$

et enfin:

$$\overline{X} - \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \leq \mu \leq \overline{X} + \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \quad (7.939)$$

qui permettaient donc de faire de l'inférence statistique en fonction de la connaissance ou non de la

moyenne ou de la variance vraie sur la totalité ou sur un échantillon de la population. En d'autres termes de savoir dans quelles bornes se situait un moment (moyenne ou variance) en fonction d'un certain niveau de confiance α imposé. Nous avons vu que le deuxième intervalle ci-dessus ne peut être que difficilement utilisé dans la pratique (suppose la moyenne théorique connue) et nous lui préférons donc le troisième.

Nous allons également démontrer en détails plus loin les deux intervalles suivants:

$$\frac{1}{F_{\alpha/2}(n, m)} \frac{S_X^2}{S_Y^2} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq \frac{1}{F_{1-\alpha/2}(n, m)} \frac{S_X^2}{S_Y^2} \quad (7.940)$$

et:

$$\frac{1}{F_{\alpha/2}(n-1, m-1)} \frac{S_X^2}{S_Y^2} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq \frac{1}{F_{1-\alpha/2}(n-1, m-1)} \frac{S_X^2}{S_Y^2} \quad (7.941)$$

Le premier intervalle ci-dessus ne peut être lui aussi que difficilement utilisé dans la pratique (suppose la moyenne théorique connue) et nous lui préférons donc le deuxième.

Définition: Lorsque nous cherchons à savoir si nous pouvons faire confiance à la valeur d'une statistique (moyenne, médiane, variance, coefficient de corrélation, etc.) avec une certaine certitude, nous parlons de "**test d'hypothèse**" et plus particulièrement de "**test de conformité**" (nous parlons de "**test d'adéquation**" quand il s'agit de vérifier que des mesures suivent bien une loi donnée et non juste une statistique).

Les tests d'hypothèses sont destinés à vérifier si un échantillon peut être considéré comme extrait d'une population donnée ou représentatif de cette population, vis-à-vis d'un paramètre comme la moyenne, la variance ou la fréquence observée. Ceci implique que la loi théorique du paramètre soit connue au niveau de la population. Les tests d'hypothèses ne sont pas faits pour démontrer l'hypothèse nulle (exprimant généralement une égalité ou une homogénéité entre différentes populations), mais pour éventuellement la rejeter (disposons pour être exact que le rejet est plus robuste).

Par exemple, si nous souhaitons savoir avec une certaine confiance si une moyenne donnée d'un échantillon de population est réaliste par rapport à la vraie moyenne théorique inconnue, nous utiliserons le "**test-Z**" qui est simplement:

$$\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (7.942)$$

Maintenant rappelons que nous avons démontré que si nous avons deux variables aléatoires de loi:

$$N(\mu_1, \sigma_1^2) \quad N(\mu_2, \sigma_2^2) \quad (7.943)$$

alors la soustraction (différencier) des moyennes donne:

$$N(\mu_2 - \mu_1, \sqrt{\sigma_1^2 + \sigma_2^2}) \quad (7.944)$$

Donc pour la différence de deux moyennes de variables aléatoires provenant de deux échantillons de

population nous obtenons directement:

$$N\left(\bar{X}_2 - \bar{X}_1, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \quad (7.945)$$

Nous pouvons alors adapter le test-Z sous la forme:

$$Z^{calc.} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (7.946)$$

La relation qui est très utile lorsque pour deux échantillons de deux populations de données, nous voulons vérifier s'il existe une différence significative des différences des moyennes théoriques à un niveau de confiance α fixé et la probabilité associée pour avoir cette différence:

$$(\bar{X}_2 - \bar{X}_1) - Z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_2 - \mu_1 \leq (\bar{X}_2 - \bar{X}_1) + Z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (7.947)$$

Donc:

$$|Z^{calc.}| \leq |Z_{\alpha/2}| \quad (7.948)$$

Nous parlons du "**test-Z de la moyenne à deux échantillons**" et il est beaucoup utilisé dans l'industrie pour vérifier l'égalité de la moyenne de deux populations de mesures.

Et si l'écart-type théorique n'est pas connu, nous utiliserons le "**test-T**" de Student (pas mal utilisé en pharmaco-économie) démontré plus haut:

$$\bar{X} - \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \quad (7.949)$$

Dans la même idée pour l'écart-type, nous utiliserons le "**test du Khi-deux**" aussi déjà démontré plus haut:

$$\frac{(n-1)S^2}{\chi_{\alpha/2}^2(n-1)} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2(n-1)} \quad (7.950)$$

Et lorsque nous voulons tester l'égalité de la variance de deux populations nous utilisons le "**test-F**" de Fisher (démontré plus bas lors de notre étude de l'analyse de la variance):

$$\frac{1}{F_{\alpha/2}(n-1, m-1)} \frac{S_X^2}{S_Y^2} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq \frac{1}{F_{1-\alpha/2}(n-1, m-1)} \frac{S_X^2}{S_Y^2} \quad (7.951)$$

Dans la pratique il faut avoir conscience que le but d'un test est très très souvent de montrer que l'effet est significatif. Il est alors d'usage de dire que le test réussit si l'hypothèse nulle est rejetée au profit de l'hypothèse alternative. Lorsque le praticien sait que l'effet est significatif et pourtant que son test

échoue à rejeter l'hypothèse nulle on parle parfois du "**dilemme du non rejet de l'hypothèse nulle**". Comme nous le verrons un peu plus loin, l'idée est alors de calculer à posteriori la puissance du test (celle-ci étant alors appelée par certains logiciels comme SPSS: "**puissance observée**") et d'adapter la taille de l'échantillon en conséquence pour avoir une puissance acceptable selon la tradition d'usage.

13.1. ORIENTATION DU TEST D'HYPOTHÈSE

Le fait que nous obtenions l'ensemble des valeurs satisfaisant à un test borné à droite et (!) à gauche est ce que nous appelons dans le cas général un "**test bilatéral**" car il comprend le test unilatéral à gauche et unilatéral à droite. Ainsi, tous les tests susmentionnés sont dans une forme bilatérale mais nous pourrions en faire une utilisation unilatérale aussi! Nous utilisons un test unilatéral lorsque la différence attendue (ou à mettre en évidence) ne peut aller que dans un sens (typiquement dans le cas des essais cliniques ou lors d'une action corrective de contrôle qualité en industrie pour laquelle nous nous attendons à une amélioration allant dans une unique direction). Les tests unilatéraux sont parfois nommés "**test de non-infériorité**" (unilatéral gauche) ou "**test de non-supériorité**" (unilatéral droite).

Ci-dessous, nous avons représenté par exemple un test unilatéral à droite (car la région de rejet est à droite et donc la probabilité cumulée est unilatérale gauche) et un test bilatéral:

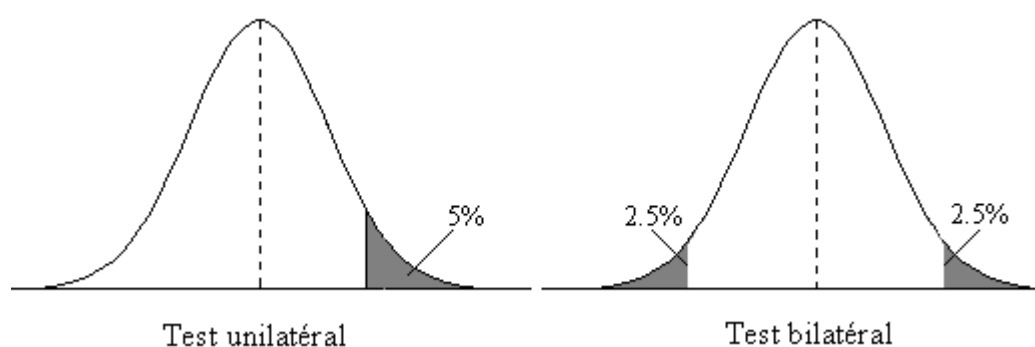


Figure: 7.43 - Illustration d'un test (ou intervalle de confiance) unilatéral à droite et bilatéral

Nous pouvons également résumer la manière de déterminer la p -value par le logigramme suivant:

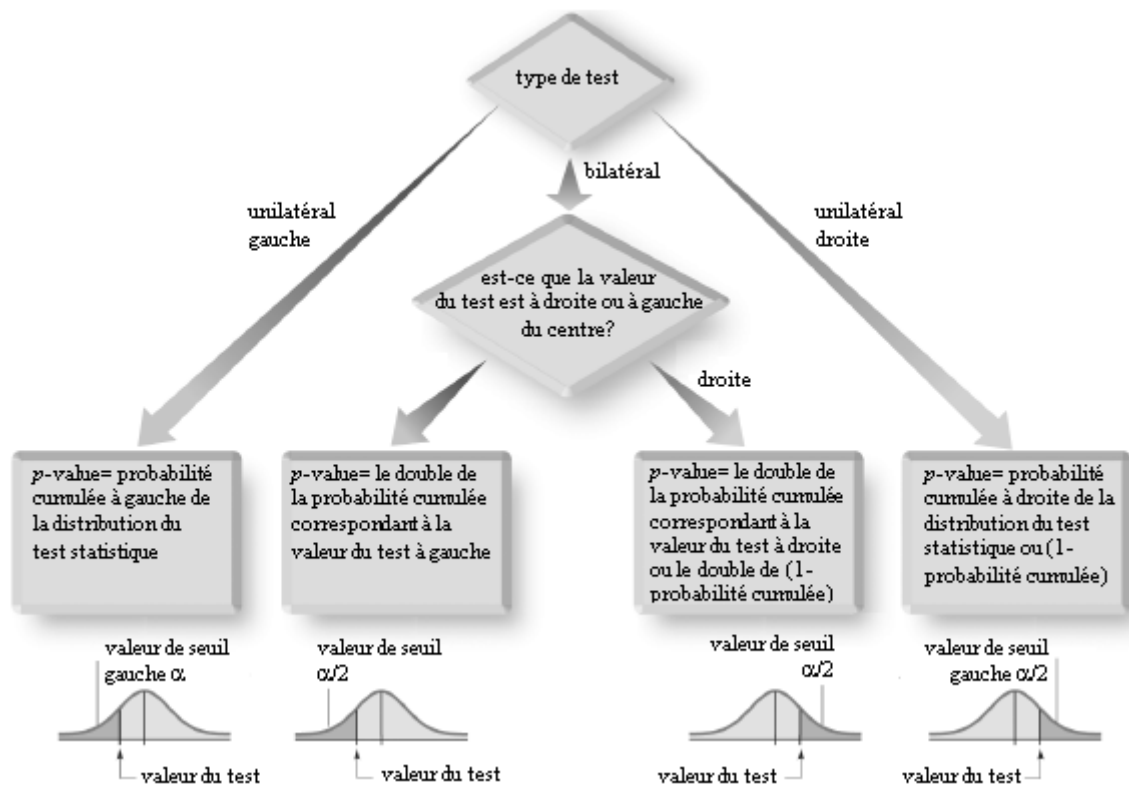


Figure: 7.44 - Figure de résumé pour déterminer la p -value lors de tests paramétriques à distribution symétrique

Signalons aussi que les tests d'hypothèses sur l'écart-type (variance), la moyenne ou la corrélation sont appelés des "**tests paramétriques**" à l'inverse des tests non paramétriques que nous verrons beaucoup plus loin.

Remarques:

R1. Il existe également une autre définition du concept de test paramétrique et non-paramétrique (un peu différente car plus précise) à voir plus loin...

R2. Attention! Certains auteurs ou professeurs parlent parfois de test "unilatéral à gauche" pour un "test unilatéral à droite"... Au fait il s'agit simplement d'un choix de vocabulaire. Si la référence pédagogique n'est pas la zone de rejet mais la zone d'acceptation, alors il est clair que les concepts de droite et gauche s'inversent...

Enfin, de nombreux logiciels calculent ce que nous appelons la " **p -value**" qui est le risque calculé (probabilité) α qu'aurait pu fixer le statisticien pour être à la limite entre l'acceptation de l'hypothèse nulle et son rejet (rappelons qu'un test qui réussit ne prouve rien). La p -value est donc une valeur fondamentale dans le domaine car elle permet de chiffrer la vraisemblance de l'hypothèse nulle H_0 (acceptation ou rejet).

Pour un test d'hypothèse, par exemple, le 5% de risque α est celui de rejeter l'hypothèse nulle H_0 alors même qu'elle est vraie. Si le risque imposé/choisi est 5% et que la p -value calculée est inférieure (dans la majorité des tests mais il faut être prudent car ce n'est pas une généralité!!!), le test échoue (rejet de l'hypothèse) en faveur d'une hypothèse alternative notée H_1 ou parfois H_A .

L'hypothèse alternative a bien évidemment elle-même son propre risque que nous notons β et sa propre p -value.

Peut-être, pour mieux comprendre, voici une illustration d'un cas particulier d'un test d'hypothèse bilatéral de la moyenne pour une variable aléatoire suivant typiquement une loi Normale (en gros c'est le même principe pour tous les tests...):

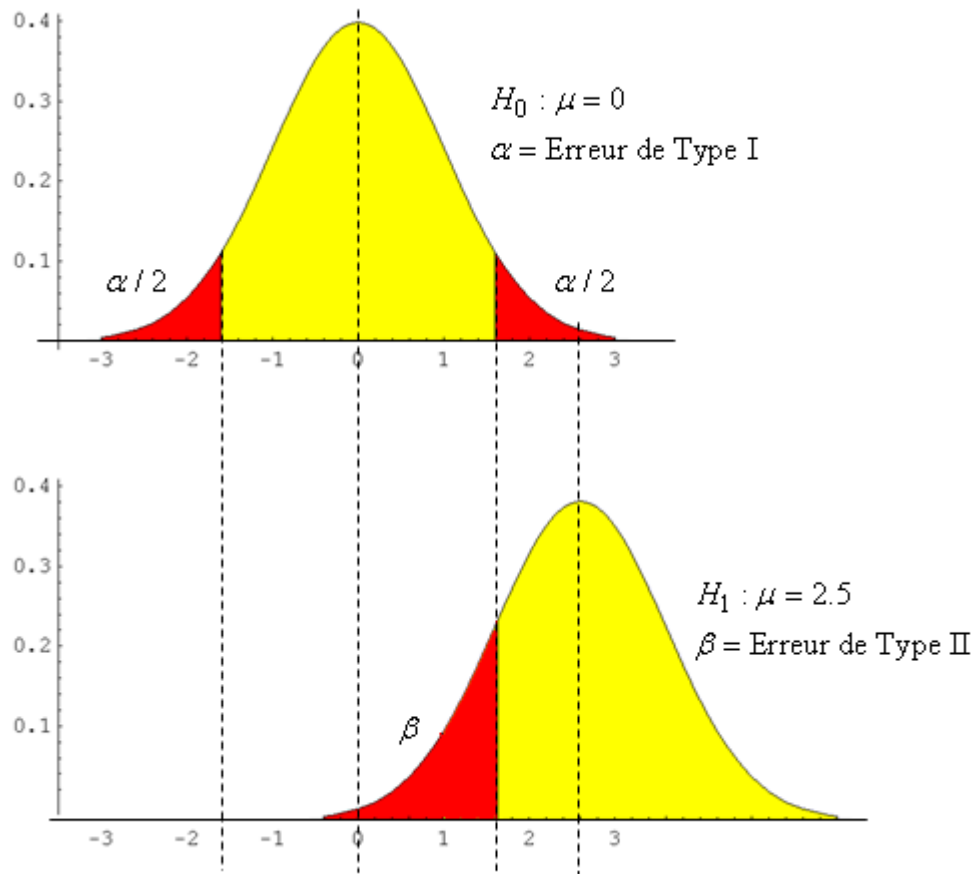


Figure: 7.45 - Hypothèse nulle et alternative d'un test bilatéral particulier

Ainsi, dans le cas présenté ci-dessus, nous voyons mieux pourquoi l'hypothèse nulle peut donc être acceptée ou rejetée en faveur de l'hypothèse alternative (qui est de même loi que l'hypothèse nulle mais juste décalée) dépendant de la valeur de référence mesurée qui sera utilisée pour le test (en l'occurrence dans le cas particulier il s'agit de la moyenne arithmétique des mesures).

Nous remarquons aussi que la zone rouge de l'hypothèse alternative, correspondant à la probabilité cumulée β , est confondue en partie avec la partie jaune de l'hypothèse nulle. Raison pour laquelle nous pouvons parfois accepter l'hypothèse nulle à tort. Nous voyons cependant que plus β serait petit, plus l'hypothèse alternative serait donc éloignée de la zone limite rouge de l'hypothèse nulle (cela correspondrait à une translation vers la droite dans le cas présent) et moins la probabilité de faire une fausse conclusion est grande. Raison pour laquelle nous parlons de "risque β " car plus celui-ci est petit, mieux c'est. In extenso, plus $1 - \beta$ est grand, moins il y a de risque de confondre l'hypothèse nulle et alternative. Raison pour laquelle $1 - \beta$ est appelé puissance **du test**" (voir plus bas la section qui est consacrée à cette notion).

Nous acceptons l'hypothèse nulle si la p -value est plus grande que 5% (0.05). Au fait, plus la p -value

est grande, mieux c'est car l'intervalle de confiance est de plus en plus petit. Si l'intervalle de confiance vient à être énorme (très proche de 100%) car la p -value est très petite alors l'analyse n'a plus vraiment de sens physiquement parlant!

Ainsi, si la p -value est faible, c'est qu'il faudrait prendre un risque faible de se tromper, donc accepter dans presque tous les cas l'hypothèse testée...

Remarque:

Nous ne devrions jamais dire que nous "acceptons" une hypothèse ou encore qu'elle est "vraie" ou "fausse" car ces termes sont trop forts et pourraient faire penser à une preuve scientifique. Nous devrions dire si nous "rejetons" ou "ne rejetons pas" l'hypothèse nulle et qu'elle est éventuellement "correcte" ou "non correcte".

Pour les test d'hypothèses bilatéraux, nous pouvons par exemple dire que nous avons (ou n'avons pas) une différence significative entre la valeur de référence mesurée et la valeur attendue. Pour les tests unilatéraux, nous pouvons dire que la valeur de référence mesurée est significativement plus grande ou plus petite que la valeur attendue.

Par ailleurs si le lecteur a bien compris la construction des tests d'hypothèses, le fait de rejeter une hypothèse à tort ("**Erreur de Type I**" ou "**Erreur de première espèce**") est donc plus robuste que de l'accepter à tort ("**Erreur de type II**" ou "**Erreur de deuxième espèce**").

Le lecteur remarquera aussi en s'aidant de la figure précédent qu'un test unilatéral a une plus forte puissance qu'un test bilatéral (à même niveau de risque bien entendu!). Ainsi, une différence non significative en test bilatéral, peut s'avérer significative en unilatéral.

Définitions:

D1. La probabilité α de l'erreur de Type I (de première espèce/faux négatif) est la probabilité de rejet de l'hypothèse nulle alors qu'elle est vraie.

D2. La probabilité β de l'erreur de Type II (de deuxième espèce/faux positif) est la probabilité de maintien de l'hypothèse nulle alors qu'elle est fausse.

Pour clore, voici les trois situations types de tests d'hypothèses sur la statistique qu'est la moyenne dans le cadre d'une distribution sous-jacente normale et dont l'espérance est dans ce cas particulier supposée nulle et de variance unitaire (car on peut très souvent ce ramener à ce cas particulier en centrant et réduisant la variable aléatoire sous-jacente):

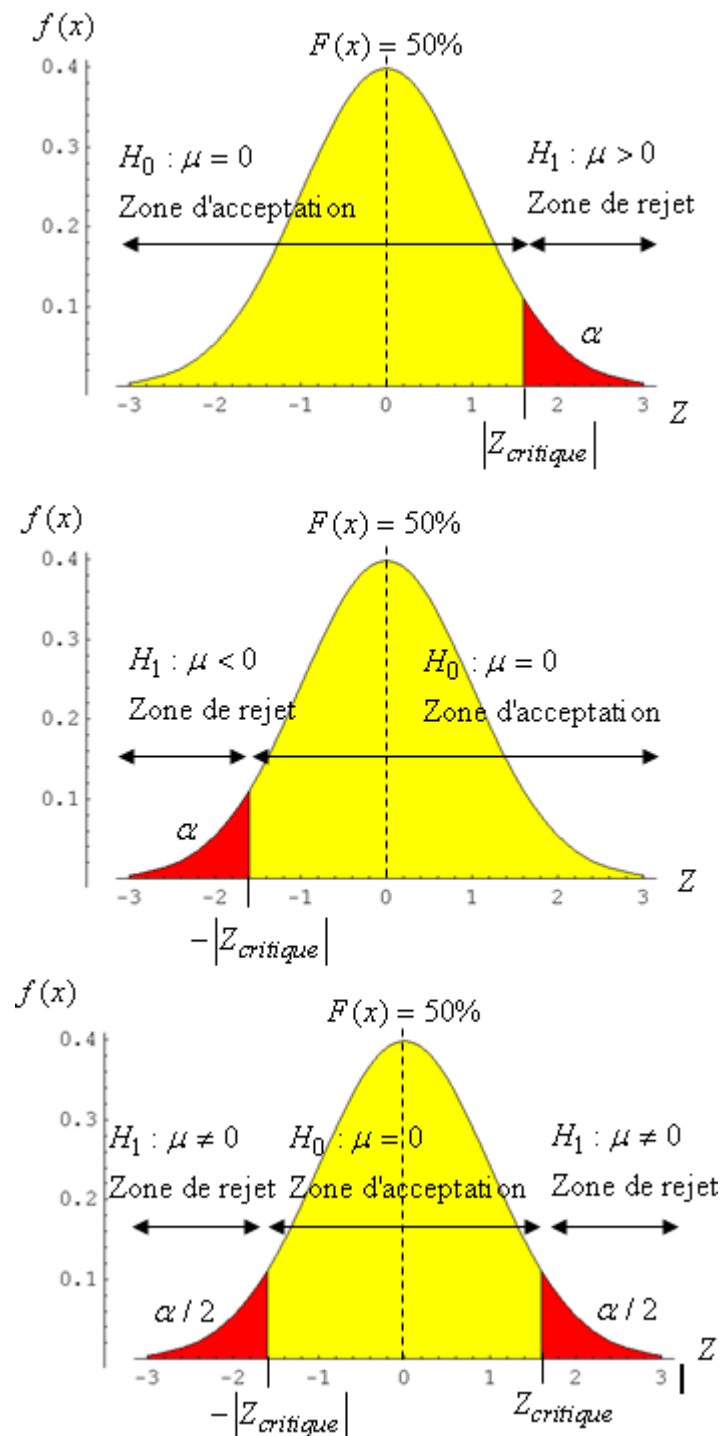


Figure: 7.46 - Les trois scénarios possibles d'un test d'hypothèse sur la moyenne

Indiquons que cela n'a aucun sens (contrairement à ce que nous pouvons parfois lire sur certains supports papier ou électronique) d'avoir les hypothèses nulles suivantes dans le cas particulier représenté ci-dessus:

$$H_0 : \mu \leq 0$$

$$H_0 : \mu \geq 0 \quad (7.952)$$

$$H_0 : \mu \neq 0$$

avec l'hypothèse alternative qui en découle automatiquement (je ne l'ai pas écrite car c'est inutile). La

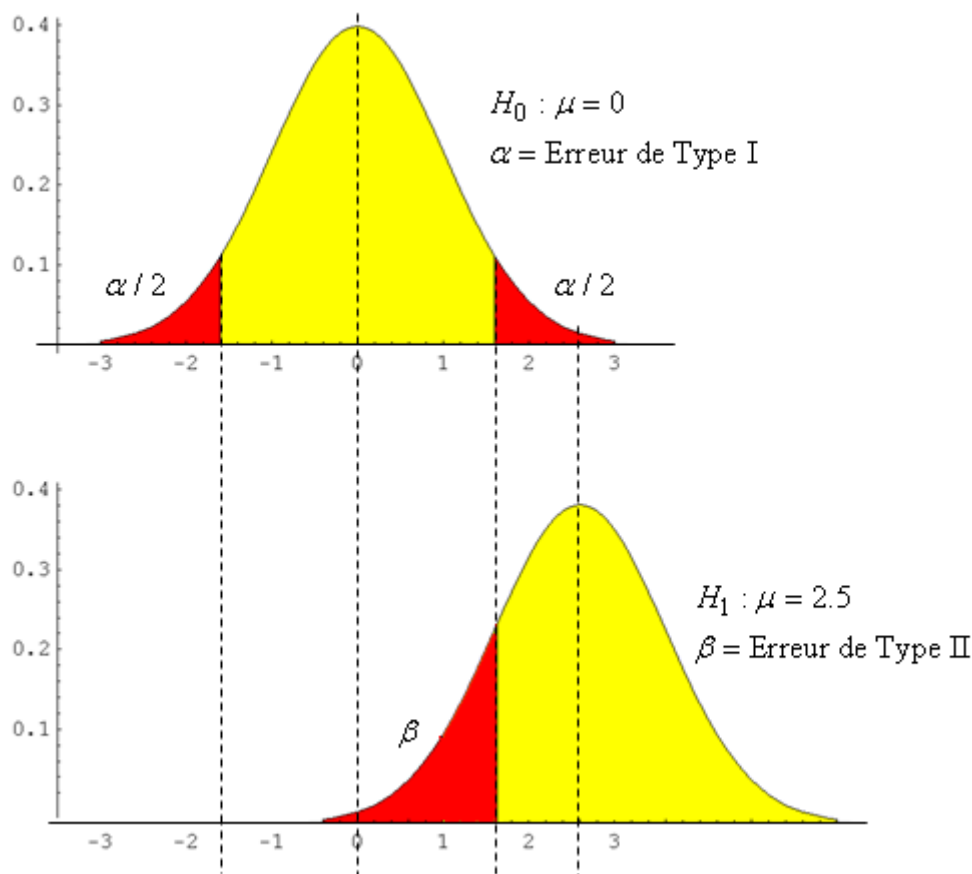
raison en est simple: comment pourriez-vous positionner votre distribution Normale centrée réduite si l'espérance n'est pas fixée...??? Raison pour laquelle l'hypothèse nulle dans le cadre des tests sur la moyenne (et d'un certain autre nombre de tests) est toujours une égalité!

Pour résumer, nous pouvons dire que si nous prenons une décision, nous pouvons nous tromper et il vaut mieux éviter de se tromper souvent. En clair, la probabilité de dire une bêtise doit être connue et de préférence petite.

13.2. PUISSANCE D'UN TEST

Lorsque l'effet est concrètement important, on imagine bien qu'il faut moins d'observations pour le démontrer que lorsqu'il est petit... mais combien au juste? A-t-on les moyens, en termes de nombre de mesures, de démontrer ce que l'on cherche? Faut-il s'y prendre autrement et changer le dispositif de son observation/expérimentation?

Pour étudier plus en détails la notion de "puissance de test" que nous avons jusque là uniquement mentionnée, rappelons la figure suivante déjà rencontrée juste un peu plus haut:



Dans l'exemple particulier ci-dessus, nous allons donc rejeter l'hypothèse nulle si $\bar{X} > 1.96$ ou si $\bar{X} < -1.96$. Imaginons que dans le cadre de l'hypothèse alternative, si nous avons mesuré 2.5, nous aurons comme puissance du test:

$$1 - \beta = 1 - P(Y_{N(2.5,1)} \leq 1.96) = 1 - P\left(Y_{N(\bar{X},1)} \leq Z_{1-\alpha/2}\right) = 1 - 0.2946 = 0.705 \quad (7.953)$$

Donc le test est relativement puissant (dans la pratique, nous considérons un test comme étant puissant si sa valeur est au-delà de 80%). Ainsi, nous remarquons que la puissance $1 - \beta$ (a posteriori!) est d'autant plus grande que la p -value sera petite (et respectivement la puissance sera à posteriori d'autant plus petite que la p -value sera grande). Donc la puissance a posteriori est en correspondance décroissante avec la p -value (dans la pratique il est cependant un peu absurde de faire ces calculs a posteriori).

13.2.1. PUISSANCE DU TEST Z À 1 ÉCHANTILLON

En toute généralité, dans le cas d'un test bilatéral, la relation précédente s'écrit donc:

$$1 - \beta = 1 - P\left(Y_{N(\bar{X}, 1)} \leq \mu + Z_{1-\alpha/2}\right) = 1 - P\left(Y_{N(0,1)} \leq \mu - \bar{X} + Z_{1-\alpha/2}\right) \quad (7.954)$$

Si l'écart-type de la moyenne n'est pas été unitaire, nous avons:

$$1 - \beta = 1 - P\left(Y_{N(0, \sigma_{\bar{X}}^2)} \leq \mu - \bar{X} + Z_{1-\alpha/2} \sigma_{\bar{X}}\right) \quad (7.955)$$

Il vient donc:

$$1 - \beta = 1 - P\left(Y_{N(0,1)} \leq \frac{\mu - \bar{X}}{\sigma_{\bar{X}}} + Z_{1-\alpha/2}\right) = 1 - P\left(Y_{N(0,1)} \leq \frac{\mu - \bar{X}}{\sigma/\sqrt{n}} + Z_{1-\alpha/2}\right) \quad (7.956)$$

autrement écrit:

$$\begin{aligned} 1 - \beta &= 1 - P\left(Y_{N(0,1)} \leq \frac{\mu - \bar{X}}{\sigma} \sqrt{n} + Z_{1-\alpha/2}\right) = 1 - P\left(Y_{N(0,1)} \leq d_{\sigma} \sqrt{n} + Z_{1-\alpha/2}\right) \\ &= P\left(Y_{N(0,1)} \leq -Z_{1-\alpha/2} - d_{\sigma} \sqrt{n}\right) = P\left(Y_{N(0,1)} \leq Z_{\alpha/2} - d_{\sigma} \sqrt{n}\right) \end{aligned} \quad (7.957)$$

C'est sous cette forme que nous retrouvons la puissance d'un test bilatéral de la moyenne (puissance Z à 1 échantillon):

$$1 - \beta = P\left(Y_{N(0,1)} \leq Z_{\alpha/2} - d_{\sigma} \sqrt{n}\right) \quad (7.958)$$

où d est parfois appelé la "**taille d'effet**" et est donc donné par:

$$d_{\sigma} = \frac{\mu - \bar{X}}{\sigma} = \frac{\delta}{\sigma} \quad (7.959)$$

et δ est nommé la "**différence**"!

Il va de soit que si la variance vraie n'est pas connue, il faut alors remplacer la loi Normale par la loi de Student tel que:

$$1 - \beta = P\left(Y_{N(0,1)} \leq T_{\alpha/2} - d_S \sqrt{n}\right) \quad (7.960)$$

avec:

$$d_S = \frac{\mu - \bar{X}}{S} = \frac{\delta}{S} \quad (7.961)$$

Remarque: Attention à un petit piège courant! Le développement ci-dessus correspond à un δ qui est donc négatif relativement à l'exemple de départ! La relation est un peu différente dans le cas où δ est positif mais cela n'a aucune importance car la puissance du test est identique valeur absolue!

Pour avoir la taille de l'échantillon c'est assez simple. Nous avons:

$$Z_{1-\beta} = Z_{\alpha/2} - d_\sigma \sqrt{n} \quad (7.962)$$

et donc en bilatéral:

$$n = \left(\frac{Z_{\alpha/2} - Z_{1-\beta}}{d_\sigma} \right)^2 \quad (7.963)$$

où nous voyons que si la puissance du test est imposée comme étant égale à 50%, ayant Z qui vaut alors 0 nous retombons (!) sur la relation de l'effectif de l'échantillon pour loi Normale démontrée bien plus haut:

$$n = \left(\frac{Z_{\alpha/2} \sigma}{\delta} \right)^2 \quad (7.964)$$

Signalons aussi que nous retrouvons parfois dans la littérature la relation antérieure sous la forme suivante:

$$n = \frac{\sigma^2 (Z_{1-\alpha/2} + Z_{1-\beta})^2}{(\mu_0 - \mu_A)^2} \quad (7.965)$$

Évidemment nous pouvons fixer d'autres paramètres pour déterminer la valeur de la variable restante. Nous pourrions par exemple chercher la valeur de la puissance du test en imposant l'écart-type, la taille de l'échantillon et le niveau de confiance, etc.

Un lecteur nous a proposé une manière très élégante de retrouver le même résultat avec beaucoup moins de développements... Effectivement, il suffit de voir sur la figure précédente que nous avons:

$$Z_\beta \frac{\sigma}{\sqrt{n}} + \bar{X} = Z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} + \mu \quad (7.966)$$

Donc nous tirons immédiatement une relation équivalent aux deux précédentes (qui donne bien évidemment le même résultat numérique):

$$n = \left(\frac{\sigma(Z_\beta - Z_{1-\alpha/2})}{\mu - \bar{X}} \right)^2 \quad (7.967)$$

Remarque: Le lecteur attentif aura peut-être remarqué que nous avons supposé dans les développements qui précèdent que l'écart-type de la moyenne vraie et alternative (estimée) est implicitement supposée être la même... Dans la pratique cela est presque tout le temps le cas, raison pour laquelle les quasi totalité des logiciels de statistiques ne demandent qu'un seul écart-type pour le calcul de la puissance du test Z à 1 échantillon. Cependant, dans certains rares logiciels universitaires, on demande l'écart-type des deux moyennes. Mais dès lors les développements ci-dessus sont différents.

Une analyse de puissance peut avoir plusieurs facettes:

1. Nous connaissons le niveau du test, la taille d'échantillon et la taille d'effet (implicitement la différence) et nous cherchons à calculer la puissance. Ceci permet de voir si notre dispositif expérimental est bien calibré.
2. Nous connaissons la puissance voulue, le niveau du test et la taille d'effet à détecter. Nous cherchons alors à calculer la taille d'échantillon nécessaire pour monter un dispositif expérimental efficace.
3. Nous connaissons la puissance voulue, le niveau du test et la taille d'échantillon et nous cherchons à vérifier qu'elle taille d'effet nous pouvons espérer mettre en évidence.

Sauf exception, nous considérerons qu'il est inutile de montrer un test si la puissance escomptée est inférieure à 80%. Cette puissance correspond à une probabilité de 80% de ne pas rejeter l'hypothèse nulle à tort, ou, ce qui revient au même de 20% d'erreur de type II.

Évidemment, il est possible de faire le même raisonnement (analytiquement quand c'est possible, sinon numériquement) avec absolument TOUS les tests d'hypothèses que nous avons vus jusqu'à maintenant. Donc au même titre qu'il y a un peu plus d'une centaine de tests d'hypothèses dans le domaine des statistiques comme nous l'avons déjà mentionné... il est évident que nous n'allons pas nous... amuser... à faire les mêmes développements pour tous ces tests mais seulement pour les grands classiques. Tant que nous avons des ordinateurs à notre disposition avec les algorithmes intégrés par des informaticiens/scientifiques, nous pouvons nous passer de refaire tous les développements qui n'apporteraient pas grand chose. Par ailleurs, la majorité des logiciels comportent des outils pour calculer la puissance de 5 à 10 tests le plus souvent.

Remarque: Nous ne traiterons pas des tests statistiques paramétriques de détection des valeurs aberrantes sur ce site comme le test Q de Dixon ou de Grubb pour la simple raison qu'ils ont une origine trop empirique et qu'ils n'ont aucun intérêt analytiquement parlant. Par contre, si des lecteurs insistent, nous pourrions mettre les détails sur ces tests avec les algorithmes détaillés de calcul des valeurs critiques en utilisant un simple tableur et la technique de Monte-Carlo pour n'importe la distribution de leur votre choix (mais pas uniquement selon la loi Normale contrairement à ce qui est écrit dans la majorité des livres).

13.3. PUISSANCE DU TEST P À 1 ET 2 ÉCHANTILLONS

De même que l'intervalle de confiance de la loi Normale avec écart-type théorique connu (c'est-à-dire sur toute la population), nous pouvons déterminer le nombre d'individus (taille d'échantillon) si nous souhaitons imposer une puissance au test de la proportion à 1 échantillon étudié plus haut. Pour cela, nous utilisons la même technique que pour la puissance du test Z. Nous écrivons alors dans un premier temps:

$$Z_{\beta} \sqrt{\frac{\hat{p}\hat{q}}{n}} + \hat{p} = Z_{1-\alpha/2} \sqrt{\frac{pq}{n}} + p \quad (7.968)$$

D'où nous déduisons:

$$n = \left(\frac{Z_{\beta} \sqrt{\hat{p}\hat{q}} - Z_{1-\alpha/2} \sqrt{pq}}{p - \hat{p}} \right)^2 \quad (7.969)$$

Donc si la puissance est de 50%, nous retrouvons bien:

$$n \cong \frac{Z_{\alpha/2}^2 p(1-p)}{\delta^2} = \frac{Z_{\alpha/2}^2 pq}{\delta^2} \quad (7.970)$$

Pour la puissance du test de la différence de deux proportions (test de la proportion à deux échantillons) dans l'objectif de déterminer la taille de l'échantillon nous sommes obligés de poser $n_1 = n_2 = n$. Dés lors, les développements obtenus lors de l'étude du test de la différence de deux proportions s'écrivent:

$$Z \cong \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{2\hat{p}(1-\hat{p})}{n}}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{2\hat{p}\hat{q}}{n}}} \quad (7.971)$$

avec:

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{n(\hat{p}_1 + \hat{p}_2)}{2n} = \frac{\hat{p}_1 + \hat{p}_2}{2} \quad (7.972)$$

De la même manière que nous l'avons fait pour le test Z et le test p à 1 échantillon, nous avons:

$$Z_{\beta} \sqrt{\frac{2\hat{p}\hat{q}}{n}} + (\hat{p}_2 - \hat{p}_1) = Z_{1-\alpha/2} \sqrt{\frac{2\hat{p}\hat{q}}{n}} + \frac{\hat{p}_1 + \hat{p}_2}{2} \quad (7.973)$$

Soit:

$$Z_{\beta} \sqrt{\frac{2\hat{p}\hat{q}}{n}} + (\hat{p}_2 - \hat{p}_1) = Z_{1-\alpha/2} \sqrt{\frac{2\hat{p}\hat{q}}{n}} + \hat{p} \quad (7.974)$$

Ce qui revient donc à supposer que la différence vraie des deux proportions est la moyenne (ce qui est

discutable...).

Mais nous avons aussi (comme les échantillons sont indépendants de par la propriété de la variance):

$$Z_{\beta} \sqrt{V(\hat{p}_1) + V(\hat{p}_2)} + (\hat{p}_2 - \hat{p}_1) = Z_{1-\alpha/2} \sqrt{\frac{2\hat{p}\hat{q}}{n}} + \hat{p} \quad (7.975)$$

Soit:

$$Z_{\beta} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n} + \frac{\hat{p}_2 \hat{q}_2}{n}} + (\hat{p}_2 - \hat{p}_1) = Z_{1-\alpha/2} \sqrt{\frac{2\hat{p}\hat{q}}{n}} + \hat{p} \quad (7.976)$$

ce qui nous donne:

$$Z_{\beta} \sqrt{\frac{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2}{n}} + (\hat{p}_2 - \hat{p}_1) = Z_{1-\alpha/2} \sqrt{\frac{2\hat{p}\hat{q}}{n}} + \hat{p} \quad (7.977)$$

Nous avons alors après réarrangement:

$$n = \left(\frac{Z_{\beta} \sqrt{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2} - Z_{1-\alpha/2} \sqrt{2\hat{p}\hat{q}}}{\hat{p} - (\hat{p}_2 - \hat{p}_1)} \right)^2 \quad (7.978)$$

13.4. ANALYSE DE LA VARIANCE (À UN FACTEUR)

L'objectif de l'analyse de la variance (contrairement à ce que son nom pourrait laisser penser) est une technique statistique permettant de comparer les moyennes de deux populations ou plus (très utilisé dans le pharma ou dans les labos de R&D ou de bancs d'essais). Cette méthode, néanmoins, doit son nom au fait qu'elle utilise des mesures de variance afin de déterminer le caractère significatif, ou non, des différences de moyennes mesurées sur les populations.

Plus précisément, la vraie signification est de savoir si le fait que des moyennes d'échantillons sont (légèrement) différentes peut être attribué au hasard de l'échantillonnage ou provient du fait qu'un facteur de variabilité engendre réellement des échantillons significativement différents (si nous avons les valeurs de toute la population, nous n'avons rien à faire!). Pour plus d'informations au niveau du vocabulaire et la mise en application, l'ingénieur et le chercheur se reporteront à la norme ISO 3534-3:1999.

Pour l'analyse de la variance appelée "**ANOVA à un facteur**" (ANalysis Of VAriance) ou "**ANAVAR à un facteur**" (ANAlyse de la VAriance), ou encore "**ANOVA à une voie**" ou plus rigoureusement "**ANOVA à un facteur fixe avec répétitions**" ou encore "**ANOVA à une variable catégorielle fixe avec répétition**", nous allons d'abord rappeler, comme nous l'avons démontré, que la loi de Fisher-Snedecor est donnée par le rapport de deux variables aléatoires indépendantes qui suivent une loi du Khi-deux et divisée par leur degré de liberté tel que:

$$\frac{U/k}{V/l} = F_{k,l}(x) = \frac{\Gamma\left(\frac{k+l}{2}\right) \left(\frac{k}{l}\right)^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right) \Gamma\left(\frac{l}{2}\right)} \frac{x^{\frac{k}{2}-1}}{\left(1 + \frac{k}{l}x\right)^{\frac{k+l}{2}}} \quad (7.979)$$

et nous allons voir maintenant son importance.

Considérons un échantillon aléatoire de taille n , disons $X_1, \dots, X_n$ issu de la loi $N(\mu_X, \sigma_X)$ et un échantillon aléatoire de taille m , disons $Y_1, \dots, Y_m$ issu de la loi $N(\mu_Y, \sigma_Y)$.

Considérons les estimateurs du maximum de vraisemblance de l'écart-type de la loi Normale traditionnellement notés dans le domaine de l'analyse de la variance par:

$$S_{*X}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu_X)^2 \quad \text{et} \quad S_{*Y}^2 = \frac{1}{m} \sum_{j=1}^m (Y_j - \mu_Y)^2 \quad (7.980)$$

Les statistiques ci-dessus sont celles que nous utiliserions pour estimer les variances si les moyennes théoriques μ_X, μ_Y étaient connues. Donc nous pouvons utiliser un résultat démontré plus haut lors de notre étude des intervalles de confiance:

$$\frac{nS_{*X}^2}{\sigma_X^2} = \chi^2(n) \quad \frac{mS_{*Y}^2}{\sigma_Y^2} = \chi^2(m) \quad (7.981)$$

Comme les Y_j sont indépendantes des X_i (hypothèse qui implique que la covariance est nulle, la réciproque n'étant pour rappel pas toujours vraie!), les variables:

$$\frac{nS_{*X}^2}{\sigma_X^2}, \frac{mS_{*Y}^2}{\sigma_Y^2} \quad (7.982)$$

sont indépendantes l'une de l'autre.

Remarque: Il existe un type d'ANOVA prévu pour le cas où les variables ne sont pas indépendantes (on parle alors de "**covariable**"). Il s'agit de l'ANCOVA qui signifie "**Analyse de la COvariance et de la VAriance**" qui utilise un mix entre la régression linéaire (cf. chapitre de Méthodes Numériques) et l'ANOVA. Le but de l'ANCOVA est de supprimer statistiquement l'effet indirect de la covariable.

Nous pouvons donc appliquer la loi de Fisher-Snedecor avec:

$$U = \frac{nS_{*X}^2}{\sigma_X^2} \quad \text{et} \quad k = n \quad (7.983)$$

ainsi que:

$$V = \frac{m S_{*Y}^2}{\sigma_Y^2} \quad \text{et} \quad l = m \quad (7.984)$$

Nous avons donc:

$$\frac{U/k}{V/l} = \frac{\frac{n S_{*X}^2}{\sigma_X^2} / n}{\frac{m S_{*Y}^2}{\sigma_Y^2} / m} = \frac{\frac{S_{*X}^2}{\sigma_X^2}}{\frac{S_{*Y}^2}{\sigma_Y^2}} = \frac{S_{*X}^2}{S_{*Y}^2} \frac{\sigma_Y^2}{\sigma_X^2} = F_{n,m}(x) \quad (7.985)$$

Soit:

$$\frac{\sigma_X^2}{\sigma_Y^2} = \frac{1}{F_{n,m}(x)} \frac{S_{*X}^2}{S_{*Y}^2} \quad (7.986)$$

Ce théorème nous permet de déduire l'intervalle de confiance du rapport de deux variances lorsque la moyenne théorique est connue. Puisque la fonction de Fisher n'est pas symétrique, la seule possibilité pour faire l'inférence c'est de faire appel au calcul numérique et nous noterons alors pour un intervalle de confiance donné le test de la manière suivante:

$$\frac{1}{F_{\alpha/2}(n,m)} \frac{S_{*X}^2}{S_{*Y}^2} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq \frac{1}{F_{1-\alpha/2}(n,m)} \frac{S_{*X}^2}{S_{*Y}^2} \quad (7.987)$$

Dans le cas où les moyennes μ_X, μ_Y sont inconnues, nous utilisons les estimateurs sans biais des variances traditionnellement notés dans le domaine de l'analyse de la variance par:

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \text{et} \quad S_Y^2 = \frac{1}{m-1} \sum_{j=1}^m (Y_j - \bar{Y})^2 \quad (7.988)$$

Pour estimer les variances théoriques, nous utilisons le résultat démontré plus haut:

$$\frac{(n-1)S_X^2}{\sigma_X^2} = \chi^2(n-1) \quad \text{et} \quad \frac{(m-1)S_Y^2}{\sigma_Y^2} = \chi^2(m-1) \quad (7.989)$$

Comme les Y_j sont indépendantes des X_i (hypothèse!), les variables:

$$\frac{(n-1)S_X^2}{\sigma_X^2}, \frac{(m-1)S_Y^2}{\sigma_Y^2} \quad (7.990)$$

sont indépendantes l'une de l'autre. Nous pouvons donc appliquer la loi de Fisher-Snedecor avec:

$$U = \frac{(n-1)S_X^2}{\sigma_X^2} \quad \text{et} \quad k = n-1 \quad (7.991)$$

ainsi que:

$$V = \frac{(m-1)S_Y^2}{\sigma_Y^2} \text{ et } l = m-1 \quad (7.992)$$

Nous avons donc:

$$\frac{U/k}{V/l} = \frac{\frac{(n-1)S_X^2}{\sigma_X^2} / (n-1)}{\frac{(m-1)S_Y^2}{\sigma_Y^2} / (m-1)} = \frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}} = \frac{S_X^2}{S_Y^2} \frac{\sigma_Y^2}{\sigma_X^2} = F_{n-1, m-1}(x) \quad (7.993)$$

Soit:

$$\frac{\sigma_X^2}{\sigma_Y^2} = \frac{1}{F_{n-1, m-1}(x)} \frac{S_X^2}{S_Y^2} \quad (7.994)$$

Ce théorème nous permet de déduire l'intervalle de confiance du rapport de deux variances lorsque la moyenne empirique est connue. Puisque la fonction de Fisher n'est pas symétrique, la seule possibilité pour faire l'inférence c'est de faire appel au calcul numérique et nous noterons alors pour un intervalle de confiance donné le "**test de Fisher**" de la manière suivante:

$$\frac{1}{F_{\alpha/2}(n-1, m-1)} \frac{S_X^2}{S_Y^2} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq \frac{1}{F_{1-\alpha/2}(n-1, m-1)} \frac{S_X^2}{S_Y^2} \quad (7.995)$$

tout en se rappelant que son utilisation nécessite implicitement des contraintes de normalité des variables étudiées.

R. A. Fisher (1890-1962) est, comme Karl Pearson, l'un des principaux fondateurs de la théorie moderne de la statistique. Fisher étudia à Cambridge où il obtint en 1912 un diplôme en astronomie. C'est en étudiant la théorie de l'erreur dans les observations astronomiques que Fisher s'intéressa à la statistique. Fisher est l'inventeur de la branche de la statistique appelée l'analyse de la variance.

Au début du 20ème siècle, R. Fischer développe donc la méthodologie des plans d'expérience (cf. chapitre de Génie Industriel). Pour valider l'utilité d'un facteur, il met au point un test permettant d'assurer que des échantillons différents sont de natures différentes. Ce test est basé sur l'analyse de la variance (des échantillons), et nommé ANOVA (analyse normalisée de la variance).

Prenons k échantillons de n valeurs aléatoires chacun. Chacune des valeurs étant considérée comme une observation ou une mesure de quelque chose ou sur la base de quelque chose (un lieu différent, ou un objet différent... bref: un seul et unique facteur de variabilité entre les échantillons!). Nous aurons donc un nombre total de N d'observations (mesures) donné par:

$$N = k \cdot n \quad (7.996)$$

si chacun des échantillons a un nombre identique de valeurs n (taille de l'échantillon) tel que $n_1 = n_2 \dots = n_k$ nous parlons alors de "**plan équilibré**" à k niveaux (ou k modalités).

Remarque: Si nous avons plusieurs facteurs de variabilité (par exemple: chaque lieu compare à lui-même plusieurs laboratoires), nous parlerons alors d'ANOVA multifactorielle. Dès lors, s'il n'y a que deux facteurs de variabilité, nous parlons d'ANOVA à deux facteurs (voir plus loin pour plus de détails).

Nous considérerons que chacun des k échantillons est issu (suit) d'une variable aléatoire suivant une loi Normale.

Facteur 1			
Échantillon 1	Échantillon 2	Échantillon...i	Échantillon k
x_{11}	x_{12}	...	x_{k1}
x_{21}	x_{22}	...	x_{2k}
...	...	x_{ij}	...
x_{n1}	x_{n2}	...	x_{kn}
Moyenne: $\bar{x}_1$	Moyenne: $\bar{x}_2$	Moyenne: $\bar{x}_i$	Moyenne: $\bar{x}_k$

Figure: 7.47 - Structure typique dite "croisée" d'une analyse de la variance à 1 facteur

En termes de test, nous voulons tester si les moyennes des k échantillons de taille n sont égales sous l'hypothèse que leurs variances sont égales. Ce que nous écrivons sous forme d'hypothèse de la manière suivante:

$$\begin{aligned} H_0 : \mu_1 = \mu_2 = \dots = \mu_k = \mu \\ H_1 : \exists i, j \mu_i \neq \mu_j \end{aligned} \quad (7.997)$$

Autrement dit: les échantillons sont représentatifs d'une même population (d'une même loi statistique). C'est-à-dire que les variations constatées entre les valeurs des différents échantillons sont dues essentiellement au hasard. Pour cela nous étudions la variabilité des résultats dans les échantillons et entre les échantillons. Il revient exactement au même de poser que (formulation qu'on retrouve dans certains articles ou ouvrages):

$$\begin{aligned} H_0 : \sigma_\mu = 0 \\ H_1 : \sigma_\mu > 0 \end{aligned} \quad (7.998)$$

Nous noterons donc pour la suite i l'indice d'échantillon (de 1 à k) et j l'indice de l'observation (de 1 à n). Donc x_{ij} sera la valeur de la j -ème observation de l'échantillon de données numéro i (nous avons choisi d'inverser la notation d'usage donc attention à ne pas vous tromper par la suite... nous sommes désolés... c'était une bêtise!).

Selon l'hypothèse susmentionnée, nous avons:

$$x_{ij} = N(\mu_i, \sigma^2) \quad (7.999)$$

Nous noterons par $\bar{x}_i$ la moyenne empirique/estimée (arithmétique) de l'échantillon i (souvent appelée "**moyenne marginale**"):

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \Rightarrow \sum_{j=1}^{n_i} x_{ij} = n_i \bar{x}_i \quad (7.1000)$$

et $\bar{\bar{x}}$ la moyenne empirique/estimée des N valeurs (soit la moyenne des $\bar{x}_i$) donnée donc par:

$$\bar{\bar{x}} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij} = \frac{1}{N} \sum_{i=1}^k (n_i \bar{x}_i) \quad (7.1001)$$

En utilisant les propriétés de l'espérance et de la variance déjà démontrées plus haut nous savons que:

$$\bar{x}_i = N(\mu_i, \sigma_i^2 / n_i) \quad \text{et} \quad \bar{\bar{x}} = N(\mu, \sigma^2 / N) \quad (7.1002)$$

avec μ qui est la moyenne des moyennes vraies μ_i :

$$\mu = \frac{1}{k} \sum_{i=1}^k \mu_i \quad (7.1003)$$

Maintenant, introduisons 3 variances:

1. La "**variance totale**" comme étant intuitivement la variance estimée sans biais en considérant l'ensemble des N observations comme un seul échantillon:

$$V_T = \frac{\sum_{ij} (x_{ij} - \bar{\bar{x}})^2}{N-1} \quad (7.1004)$$

où le terme au numérateur est appelé "**somme des carrés des écarts totaux**".

2. La "**variance entre échantillons**" (c'est-à-dire entre les moyennes des échantillons) est aussi intuitivement l'estimateur de la variance des moyennes des échantillons:

$$V_E = \frac{\sum_i (\bar{x}_i - \bar{\bar{x}})^2}{k-1} \quad (7.1005)$$

où le terme au numérateur est appelé "**somme des carrés des écarts entre échantillons**".

Comme nous avons démontré que si toutes les variables sont identiquement distribuées (même variance) et indépendantes la variance des individus vaut n fois celle de la moyenne:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \Leftrightarrow V(\bar{X})n = V(X) \quad (7.1006)$$

alors la "**variance des observations**" (variables aléatoires dans un échantillon) est donnée par:

$$V_A = n \cdot V_E = \frac{n \sum_i (\bar{x}_i - \bar{\bar{x}})^2}{k-1} \quad (7.1007)$$

Nous avons donc ci-dessus l'hypothèse de l'égalité des variances qui est exprimée sous forme mathématique pour les développements à suivre.

3. La "**variance résiduelle**" est l'effet des facteurs dits non contrôlés. C'est par définition la moyenne des variances des échantillons (en quelque sorte: l'erreur standard):

$$V_R = \frac{1}{k} \sum_i \frac{\sum_j (x_{ij} - \bar{x}_i)^2}{n-1} = \frac{\sum_{ij} (x_{ij} - \bar{x}_i)^2}{k(n-1)} \quad (7.1008)$$

où le terme au numérateur est appelé "**somme des carrés des écarts des résidus**" ou encore plus souvent "**erreur résiduelle**".

Au final, ces indicateurs sont parfois résumés sous la forme suivante:

$$\begin{aligned} Q_T &= \sum_{ij} (x_{ij} - \bar{\bar{x}})^2 & ddl_T &= N-1 = kn-1 & V_T &= \frac{Q_T}{ddl_T} \\ Q_A &= n \sum_i (\bar{x}_i - \bar{\bar{x}})^2 & ddl_A &= k-1 & V_A &= \frac{Q_A}{ddl_A} \\ Q_R &= \sum_{ij} (x_{ij} - \bar{x}_i)^2 & ddl_R &= k(n-1) & V_R &= \frac{Q_R}{ddl_R} \end{aligned} \quad (7.1009)$$

Remarquons que si les échantillons n'ont pas la même taille (ce qui est rare dans la pratique), nous avons alors:

$$\begin{aligned} Q_T &= \sum_{ij} (x_{ij} - \bar{\bar{x}})^2 & ddl_T &= N-1 & V_T &= \frac{Q_T}{ddl_T} \\ Q_A &= \sum_i n_i (\bar{x}_i - \bar{\bar{x}})^2 & ddl_A &= k-1 & V_A &= \frac{Q_A}{ddl_A} \\ Q_R &= \sum_{ij} (x_{ij} - \bar{x}_i)^2 & ddl_R &= \sum_i (n_i - 1) & V_R &= \frac{Q_R}{ddl_R} \end{aligned} \quad (7.1010)$$



Remarques:

R1. Le terme Q_T est souvent indiqué dans l'industrie par l'abréviation SST signifiant en anglais "Sum of Squares Total" ou plus rarement TSS pour "Total Sum of Squares".

R2. Le terme Q_A est souvent indiqué dans l'industrie par l'abréviation SSB signifiant en anglais "Sum of Squares Between (samples)" ou plus rarement SSk pour "Sum of Squares Between treatments".

R3. Le terme Q_R est souvent indiqué dans l'industrie par l'abréviation SSW signifiant en anglais "Sum of Squares Within (samples)" ou plus rarement SSE pour "Sum of Squares due to Errors".

Indiquons que nous voyons souvent dans la littérature (nous réutiliserons un peu plus loin cette notation):

$$\begin{aligned} Q_R &= \sum_{\tilde{y}} (x_{\tilde{y}} - \bar{x}_i)^2 = \frac{(n_i - 1)}{(n_i - 1)} \sum_{\tilde{y}} (x_{\tilde{y}} - \bar{x}_i)^2 = \frac{(n_i - 1)}{(n_i - 1)} \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{\tilde{y}} - \bar{x}_i)^2 \\ &= \sum_{i=1}^k (n_i - 1) \sum_{j=1}^{n_i} \frac{1}{(n_i - 1)} (x_{\tilde{y}} - \bar{x}_i)^2 = \sum_{i=1}^k (n_i - 1) S_i^2 \end{aligned} \quad (7.1011)$$

avec donc l'estimateur sans biais de la variance des observations:

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{\tilde{y}} - \bar{x}_i)^2 \quad (7.1012)$$

Avant d'aller plus loin, arrêtons-nous sur la variance résiduelle. Nous avons donc pour des échantillons qui ne sont pas de même taille:

$$V_R = \frac{\sum_{\tilde{y}} (x_{\tilde{y}} - \bar{x}_i)^2}{\sum_i (n_i - 1)} \quad (7.1013)$$

Ouvrons maintenant une petite parenthèse... Prenons le cas particulier de deux échantillons seulement. Nous pouvons alors écrire:

$$V_R = \frac{\sum_j (x_{1j} - \bar{x}_1)^2 + \sum_j (x_{2j} - \bar{x}_2)^2}{(n_1 - 1) + (n_2 - 1)} = \frac{\sum_j (x_{1j} - \bar{x}_1)^2 + \sum_j (x_{2j} - \bar{x}_2)^2}{n_1 + n_2 - 2} \quad (7.1014)$$

Soit en introduisant l'estimateur du maximum de vraisemblance de la variance:

$$V_R = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \quad (7.1015)$$

Nous pouvons d'ailleurs observer que dans le cas particulier où:

$$n_1 = n_2 = n \quad (7.1016)$$

alors:

$$V_R = \frac{nS_1^2 - S_1^2 + nS_2^2 - S_2^2}{n + n - 2} = \frac{n(S_1^2 + S_2^2) - (S_1^2 + S_2^2)}{2(n - 1)} = \frac{(n - 1)(S_1^2 + S_2^2)}{2(n - 1)} = \frac{S_1^2 + S_2^2}{2} \quad (7.1017)$$

Donc:

$$S_1^2 \leq V_R \leq S_2^2 \quad (7.1018)$$

Supposons maintenant que nous souhaitons comparer avec un certain intervalle de confiance la moyenne de deux populations ayant une variance différente pour savoir si elles sont de natures différentes ou non.

Nous connaissons pour le moment deux tests pour vérifier les moyennes. Le test-Z et le test-T. Comme dans l'industrie il est rare que nous ayons le temps de prendre des grands échantillons, concentrons-nous sur le deuxième que nous avons démontré plus haut:

$$\bar{X} - \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \quad (7.1019)$$

Et rappelons aussi que:

$$S_{\bar{X}} = \frac{S}{\sqrt{n}} \quad (7.1020)$$

Maintenant rappelons que nous avons démontré que si nous avons deux variables aléatoires de loi:

$$N(\mu_1, \sigma_1^2) \quad N(\mu_2, \sigma_2^2) \quad (7.1021)$$

alors la soustraction (différencier) des moyennes donne:

$$N(\mu_2 - \mu_1, \sqrt{\sigma_1^2 + \sigma_2^2}) \quad (7.1022)$$

Donc pour la différence de deux moyennes de variables aléatoires provenant de deux échantillons de population nous obtenons directement:

$$N\left(\bar{X}_2 - \bar{X}_1, \sqrt{S_{\bar{X}_1}^2 + S_{\bar{X}_2}^2}\right) = N\left(\bar{X}_2 - \bar{X}_1, \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right) \quad (7.1023)$$

Et maintenant l'idée est de prendre l'approximation (sous l'hypothèse que les variances sont égales):

$$N\left(\bar{X}_2 - \bar{X}_1, \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right) \cong N\left(\bar{X}_2 - \bar{X}_1, \sqrt{\frac{V_R}{n_1} + \frac{V_R}{n_2}}\right) = N\left(\bar{X}_2 - \bar{X}_1, \sqrt{V_R \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right) \quad (7.1024)$$

Cette approximation est appelée "**hypothèse homoscédastique**".

Nous avons alors l'intervalle de confiance:

$$\left(\bar{X}_2 - \bar{X}_1\right) - \sqrt{V_R \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} T_{\alpha/2}(n-1) \leq \mu_2 - \mu_1 \leq \left(\bar{X}_2 - \bar{X}_1\right) + \sqrt{V_R \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} T_{\alpha/2}(n-1) \quad (7.1025)$$

Comme l'idée dans la pratique est souvent de tester l'égalité des moyennes théoriques (et donc que leur différence est nulle) à partir des estimateurs connus alors:

$$\left(\bar{X}_2 - \bar{X}_1\right) - \sqrt{V_R \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} T_{\alpha/2}(n-1) \leq 0 \leq \left(\bar{X}_2 - \bar{X}_1\right) + \sqrt{V_R \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} T_{\alpha/2}(n-1) \quad (7.1026)$$

avec:

$$n = n_1 + n_2 \quad (7.1027)$$

Dans la plupart des logiciels disponibles sur le marché, le résultat est uniquement donné à partir du fait que le $T_{\alpha/2}$ que nous avons est compris dans le $T_{\alpha/2}$ correspondant à l'intervalle de confiance donné rappelons-le par:

$$T^{calc.}(n-1) = \frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)}{\sqrt{V_R \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (7.1028)$$

dans le cas de l'hypothèse homoscédastique (égalité des variances/homogénéité des variances).

Remarque: Cette dernière relation est appelée "**independent 2-sample T-test**", ou "**test-T homoscédastique**" ou encore "**test-T d'égalité des espérances de 2 observations avec variances égales**" ou encore plus simplement mais un peu abusivement "**test-T à 2 échantillons**", avec taille des échantillons différentes et variances égales. Souvent dans la littérature, les deux moyennes théoriques sont égales lors de la comparaison. Il s'en suit que nous avons alors:

$$T^{calc.}(n-1) = \frac{\bar{X}_2 - \bar{X}_1}{V_R \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (7.1029)$$

Sinon, dans le cas plus général de l'hypothèse d'hétéroscédasticité (non égalité des variances), nous écrivons explicitement:

$$T^{calc.}(n-1) = \frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (7.1030)$$

Donc:

$$|T^{calc.}(n-1)| \leq |T_{\alpha/2}(n-1)| \quad (7.1031)$$

Remarque: La relation antéprécédente est appelée "**independent two-sample T-test**", ou "**test-T hétéroscédastique**" ou encore "**test d'égalité des espérances: deux observations avec variances différentes**". Si la taille des échantillons est égale et que les variances le sont aussi et que nous supposons les deux moyennes théoriques égales lors de la comparaison, il s'ensuit que nous avons alors:

$$T^{calc.}(n-1) = \frac{\bar{X}_2 - \bar{X}_1}{S_{1,2} \sqrt{\frac{2}{n}}} \quad (7.1032)$$

Bref, fermons cette parenthèse et revenons à nos moutons... Nous en étions donc au tableau suivant:

$$\begin{aligned} Q_T &= \sum_{ij} (x_{ij} - \bar{\bar{x}})^2 & ddl_T &= N - 1 = kn - 1 & V_T &= \frac{Q_T}{ddl_T} \\ Q_A &= n \sum_i (\bar{x}_i - \bar{\bar{x}})^2 & ddl_A &= k - 1 & V_A &= \frac{Q_A}{ddl_A} \\ Q_R &= \sum_{ij} (x_{ij} - \bar{x}_i)^2 & ddl_R &= k(n - 1) & V_R &= \frac{Q_R}{ddl_R} \end{aligned} \quad (7.1033)$$

où nous avons donc dans le cas d'échantillons de même taille:

$$ddl_T = ddl_A + ddl_R = k - 1 + k(n - 1) = k - 1 + kn - k = kn - 1 \quad (7.1034)$$

Ainsi que l'erreur totale qui est la somme de l'erreur des moyennes (interclasses) et de l'erreur résiduelle (intra-classes) et ce que les échantillons soient de même taille ou non:

$$Q_T = Q_A + Q_R \quad (7.1035)$$

Effectivement:

$$\begin{aligned} Q_T &= \sum_{ij} (x_{ij} - \bar{\bar{x}})^2 = \sum_i \sum_j (x_{ij} - \bar{\bar{x}})^2 = \sum_i \sum_j ((x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{\bar{x}}))^2 \\ &= \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 + \sum_i \sum_j (\bar{x}_i - \bar{\bar{x}})^2 + 2 \sum_i \sum_j (x_{ij} - \bar{x}_i)(\bar{x}_i - \bar{\bar{x}}) \end{aligned} \quad (7.1036)$$

Or, nous avons:

$$\sum_i \sum_j (x_{ij} - \bar{x}_i) (\bar{x}_i - \bar{\bar{x}}) = \sum_i \left\{ (\bar{x}_i - \bar{\bar{x}}) \sum_j (x_{ij} - \bar{x}_i) \right\} = 0 \quad (7.1037)$$

car:

$$\sum_j (x_{ij} - \bar{x}_i) = 0, \forall i \quad (7.1038)$$

Donc:

$$Q_T = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 + \sum_i \sum_j (\bar{x}_i - \bar{\bar{x}})^2 = Q_R + Q_A \quad (7.1039)$$

Maintenant, sous l'hypothèse forte (qui va nous être indispensable un peu plus loin) que les variances vraies sont liées par la relation:

$$\sigma_R^2 = \sigma_A^2 = \sigma_T^2 = \sigma^2 \quad (7.1040)$$

et donc que leurs estimateurs respectifs sont asymptotiquement égaux... ce qui dans la pratique n'est approximativement vrai que lorsque certaines conditions sont satisfaites (raison pour laquelle il faut absolument avant de faire une ANOVA exécuter un calcul de la puissance et de l'effectif d'une ANOVA) nous avons:

$$\frac{Q_R}{\sigma^2} = \frac{\sum_j (x_{ij} - \bar{x}_i)^2}{\sigma^2} = \chi_{N-k}^2 \quad (7.1041)$$

ce qui découle immédiatement de la démonstration que nous avons faite lors de notre étude de l'inférence statistique avec la loi du Khi-deux où nous avons obtenu (pour rappel):

$$\chi_{n-1}^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{(n-1) \left(\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \right)}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \quad (7.1042)$$

Pour déterminer le nombre de degrés de liberté de la loi du Khi-deux de:

$$\frac{Q_A}{\sigma^2} = \frac{\sum_i n_i (\bar{x}_i - \bar{\bar{x}})^2}{\sigma^2} \quad (7.1043)$$

Nous allons utiliser le fait que (par le même raisonnement que pour la relation antérieure):

$$\frac{Q_T}{\sigma^2} = \frac{\sum_{ij} (x_{ij} - \bar{\bar{x}})^2}{N-1} = \chi_{N-1}^2 \quad (7.1044)$$

et que puisque $Q_T = Q_A + Q_R$, nous devons alors avoir:

$$\frac{Q_T}{\sigma^2} = \frac{Q_A + Q_R}{\sigma^2} = \frac{Q_A}{\sigma^2} + \frac{Q_R}{\sigma^2} = \chi^2_{?} + \chi^2_{N-k} = \chi^2_{N-1} \quad (7.1045)$$

Il s'ensuit de par la propriété de linéarité du Khi-deux:

$$\frac{Q_A}{\sigma^2} = \frac{\sum_i n_i (\bar{x}_i - \bar{\bar{x}})^2}{\sigma^2} = \chi^2_{k-1} \quad (7.1046)$$

Donc pour résumer nous avons:

$$\boxed{\frac{Q_A}{\sigma^2} = \chi^2_{k-1}} \quad \text{et} \quad \boxed{\frac{Q_R}{\sigma^2} = \chi^2_{N-k}} \quad (7.1047)$$

C'est maintenant qu'intervient la loi de Fisher dans l'hypothèse où les variances sont égales! Puisque:

$$\frac{(k-1) S_A^2}{\sigma^2} = \chi^2_{k-1} \quad \text{et} \quad \frac{(N-k) S_R^2}{\sigma^2} = \chi^2_{N-k} \quad (7.1048)$$

Ce que nous souhaitons faire c'est voir s'il y a une différence entre la variance des moyennes (interclasses) et la variance résiduelle (intra-classes). Pour comparer deux variances lorsque les moyennes vraies sont inconnues nous avons vu que le mieux était d'utiliser le test de Fisher. Or, nous avons démontré dans notre étude de la loi de Fisher un peu plus haut que:

$$\frac{\sigma_X^2}{\sigma_Y^2} = \frac{1}{F_{n-1, m-1}(x)} \frac{S_X^2}{S_Y^2} \quad (7.1049)$$

où dans notre cas d'étude:

$$\frac{\sigma_A^2}{\sigma_R^2} = 1 \Rightarrow F_{k-1, N-k} = \frac{S_A^2}{S_R^2} = \frac{V_A}{V_R} = \frac{\frac{\sum_i n_i (\bar{x}_i - \bar{\bar{x}})^2}{k-1}}{\frac{\sum_{ij} (x_{ij} - \bar{x}_i)^2}{N-k}} = \frac{\frac{Q_A}{k-1}}{\frac{Q_R}{N-k}} \quad (7.1050)$$

Au vu de l'hypothèse de la première égalité (qui précède l'implication, nous comprenons mieux la très grande sensibilité des résultats de l'ANOVA à la non égalité des variances.

Indiquons encore que la relation précédente:

$$\boxed{F_{k-1, N-k} = \frac{\frac{Q_A}{k-1}}{\frac{Q_R}{N-k}}} \quad (7.1051)$$

est souvent indiquée dans la littérature sous la forme suivante:

$$F_{k-1, N-k} = \frac{MSk}{MSE} \quad (7.1052)$$

où MSk est appelé "**Mean Square for treatments**" et MSE "**Mean Square for Error**". Ce rapport va donc nous donner la valeur de la variable aléatoire F (dont le support est pour rappel borné à zéro à gauche). Comme il s'agit d'une loi qui n'est pas symétrique il est peu judicieux de faire un test bilatéral. On fera ainsi plutôt un test unilatéral à droite (la région de rejet étant à droite).

L'interprétation de cette fraction est donc en gros la suivante: Il s'agit du rapport (normalisé au nombre de degrés de liberté) de la somme de l'erreur des moyennes (interclasses) et de l'erreur résiduelle (intra-classes) ou autrement dit le rapport de la variance interclasse par la variance résiduelle. Ce rapport suit donc une loi de Fisher à deux paramètres donnés par les degrés de liberté des classes respectives.

Remarque: S'il y a seulement deux populations (échantillons), il faut bien comprendre qu'à ce moment l'utilisation du test- T de Student suffit amplement et est considéré comme équivalent! Au fait, l'ANOVA est une comparaison indirecte des moyennes, Student une comparaison directe... il est donc évident de deviner lequel est le mieux dans cette situation particulière!

Tous les calculs que nous avons faits sont très souvent représentés dans les logiciels sous la forme d'une table standardisée dont voici la forme et le contenu (c'est ainsi que le présente Microsoft Excel 11.8346 ou Minitab 15.1.1 par exemple):

Source	Somme des carrés	ddl du χ^2	Moyenne des carrés	F	Valeur critique F_c
Inter-Classe	$Q_A = \sum_i n_i (\bar{x}_i - \bar{\bar{x}})^2$	$k-1$	$MSk = \frac{Q_A}{k-1}$	$\frac{MSk}{MSE}$	$P(F > F_{k-1, N-k})$
Intra-Classe	$Q_R = \sum_{\bar{y}} (x_{\bar{y}} - \bar{x}_i)^2$	$N-k$	$MSE = \frac{Q_R}{N-k}$		
Total	$Q_T = \sum_{\bar{y}} (x_{\bar{y}} - \bar{\bar{x}})^2$	$N-1$			

Tableau: 7.12 - Terminologie et paramètres traditionnels d'un Tableau ANOVA (TAV) à un facteur

ainsi, pour que l'hypothèse soit acceptée, il faut que la valeur de:

$$F_{k-1, N-k} = \frac{MSk}{MSE} \quad (7.1053)$$

soit plus petite ou égale au centile de la même loi F avec une probabilité cumulée correspondant à 1 soustrait de niveau de confiance α .

La valeur choisie du F critique est un peu malheureuse à mon avis dans les tableaux d'ANOVA (mais bon une fois que l'on sait que c'est ainsi...). Il est peut-être plus aisé de comprendre cette valeur si nous l'introduisons ainsi (le test unilatéral à droite ressort pédagogiquement mieux à mon avis):

$$P(F \leq F_{k-1, N-1}) = 1 - F_c \leq 1 - \alpha \quad (7.1054)$$

Il faut donc pour que le test ne soit pas rejeté que:

$$F_c \geq \alpha \quad (7.1055)$$

Donc la valeur critique de F correspond simplement et bêtement à la probabilité cumulée de la p -value.

Il faut cependant bien se rappeler que pour utiliser l'ANOVA, on doit donc supposer que les échantillons sont issus d'une même population (données appariées) et suivent une loi normale. Il est donc nécessaire de vérifier la normalité des distributions et l'homoscédasticité (test de Levene par exemple). Dans le cas contraire, il faut utiliser des variantes non paramétriques de l'analyse de variance (ANOVA de Kruskal-Wallis ou ANOVA de Friedman). Ces tests ne sont pas encore démontrés en détails à ce jour sur le site.

Remarques:

R1. À noter que dans la pratique, la variance inter-classe est très souvent nommée "**variance inter-laboratoires**" et la variance intra-classe est in extenso souvent nommée "**variance intra-laboratoire**".

R2. Il existe en ce début de 21ème siècle plus de 50 test ou procédures de comparaison de variances. L'opinion varie parmi les auteurs quant à leur pertinence et l'efficacité des tests d'homogénéité de variance (THV). Certains affirment que ces derniers sont indispensables à réaliser avant toute ANOVA, d'autres disent que ces tests sont de toute façon de piètre performance, l'ANOVA étant plus robuste aux écarts d'homoscédasticité que ce qui peut être détecté par les THV, particulièrement en cas de non-normalité. En fait, toutes ces questions se rapportent au problème dit de Behrens-Fisher, qui est celui de la comparaison de moyennes sans supposer l'équivariance. Cependant parmi la cinquantaine de tests existants, plusieurs études comparatives ont permis de dégager les tests suivants: Test de Bartlett, Levene et Brown-Forsythe.

13.5. ANALYSE DE LA VARIANCE (ANOVA À DEUX FACTEURS SANS RÉPÉTITION)

Nous allons voir maintenant le concept d'interaction qui est fondamental pour bien comprendre ce qu'il y a derrière l'ANOVA à deux facteurs (fixes) (ou "ANOVA à deux variables catégorielles fixes") sans et surtout avec répétition. Effectivement, ce n'est qu'avec l'ANOVA à deux facteurs avec répétition – par construction mathématique – que l'on peut statistiquement (sous certaines hypothèses) étudier objectivement si deux ou plusieurs facteurs interagissent de manière significative ensemble.

Il nous faut donc, avant de passer à la partie mathématique pure, introduire quelques notions:

Définitions:

D1. Nous disons qu'il y a "**absence d'interaction**" quand la moyenne des réponses d'un facteur en fonction de ses niveaux varie de la même amplitude et avec le même signe que la moyenne des réponses d'un autre facteur en fonction de ses niveaux. Nous disons alors que les courbes de réponses dans le diagramme des interactions sont parallèles.

Remarque: Le parallélisme des courbes de réponses est normal en situation d'absence d'interaction, car cela signifie que quel que soit le niveau de l'un ou l'autre des facteurs, la variation (si elle existe) de la réponse sera toujours de la même amplitude. Ce qui est caractéristique de l'indépendance (du moins localement).

D2. Nous disons que deux facteurs sont "**en interaction**" quand la moyenne des réponses d'un facteur en fonction de ses niveaux ne varie pas de la même amplitude ou/et pas avec le même signe que la moyenne des réponses d'un autre facteur en fonction de ses niveaux. Nous disons alors que les courbes de réponses dans le diagramme des interactions ne sont pas parallèles.

Remarque: L'absence d'interaction est une hypothèse très forte et une observation rare. Souvent, nous avons des interactions ou de fortes interactions.

Pour comprendre le concept, nous aurons recours à de petits exemples sans répétition qui permettront de se faire une idée qualitative du phénomène mais en aucun cas une approche scientifique de l'interaction.

À chaque fois nous visualiserons les situations au moyen de deux types de représentations: un graphique illustrant les effets principaux d'une part et un diagramme des interactions d'autre part.

Considérons le petit tableau suivant avec deux facteurs à deux niveaux ("**variables explicatives**") comportant donc 4 cellules ("**variables d'intérêt**"):

	Facteur 2	
Facteur 1	Niveau 1	Niveau 2
Niveau 1	3	3
Niveau 2	3	3

Tableau: 7.13 - Premier exemple d'une petite ANOVA à deux facteurs sans répétition

Nous aurons comme représentations avec un logiciel comme Minitab:

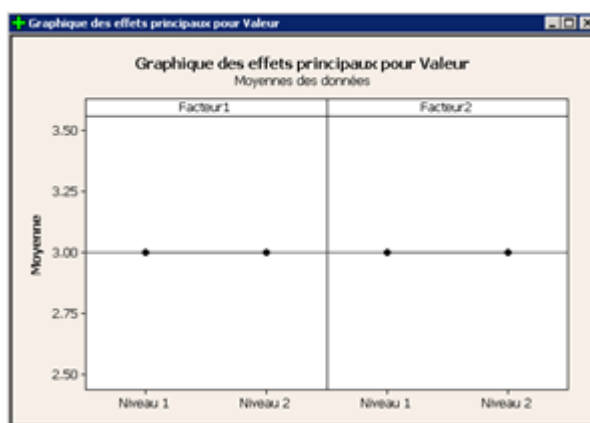


Figure: 7.48 - Graphique des effets principaux avec Minitab 15

Nous voyons bien qu'aucun facteur n'a un effet principal sur quoi que ce soit. Ce qui est relativement intuitif étant donné le contenu de tableau précédent.

Le diagramme des interactions (appelé souvent "profileur" dans l'industrie) donne lui:

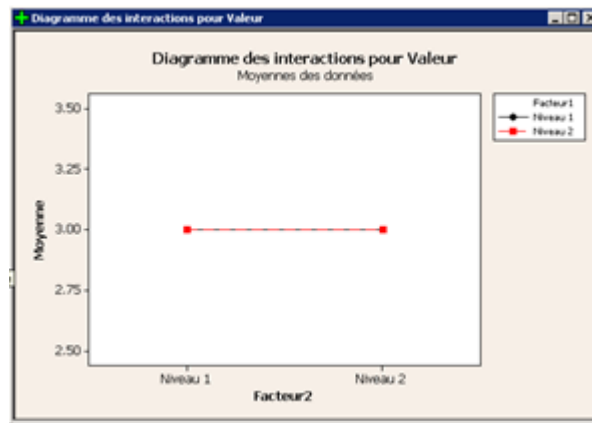


Figure: 7.49 - Diagramme des interactions avec Minitab 15

où nous pouvons constater que les facteurs n'interagissent pas entre eux (ou se neutralisent c'est selon...). Nous disons alors qu'il n'y a "(a priori) **aucun effet ni aucune interaction (localement)**". Au fait dans certaines expériences, l'absence d'interaction est une hypothèse très forte et donc souvent rare. Raison pour laquelle il faut faire attention aux mots choisis lors de l'interprétation des graphiques d'interaction (car ne pas passer par les calculs purs est délicat pour cette étape voire non scientifique!).

Maintenant considérons le tableau suivant:

	Facteur 2	
Facteur 1	Niveau 1	Niveau 2
Niveau 1	2	2
Niveau 2	4	4

Tableau: 7.14 - Deuxième exemple d'une petite ANOVA à deux facteurs sans répétition

Il nous paraît clair que le Facteur 1 à travers la prise en compte de son niveau semble avoir une influence sur la réponse. Mais voyons les différentes représentations:

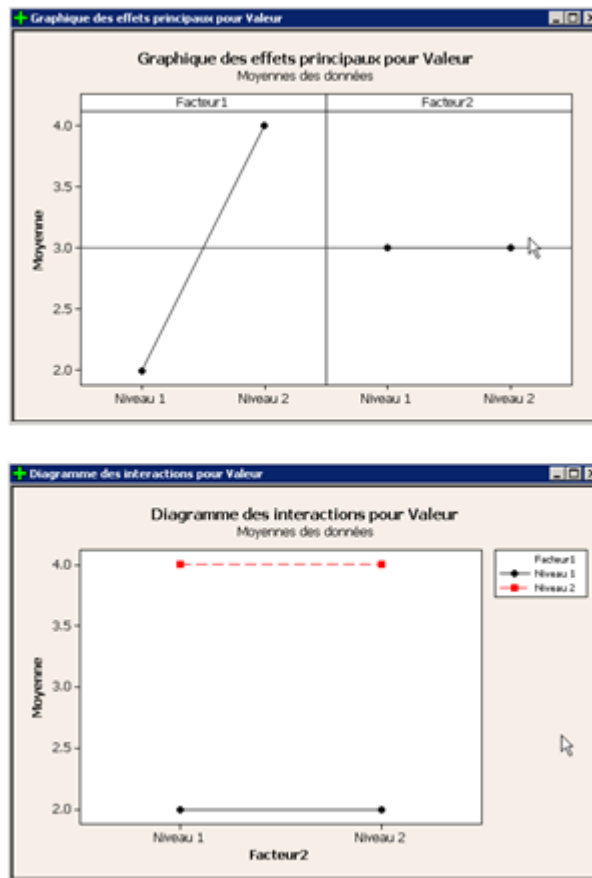


Figure: 7.50 - Graphique des effets principaux et diagramme des interactions avec Minitab 15

Détaillons plus le premier graphique comme l'a proposé un lecteur:

Ce graphique comporte 2 parties: celle de gauche analyse les effets du facteur 1 à travers ses 2 niveaux ; celle de droite en fait de même pour le facteur 2.

Examinons de plus près la partie de gauche:

Nous y voyons 2 points reliés par un segment de droite. Ici le premier point, celui pour le niveau 1, est situé à l'ordonnée 2 alors que le deuxième point, celui pour le niveau 2, est situé à l'ordonnée 4. Rappelons-nous maintenant que chaque point représente une moyenne. Ainsi l'ordonnée du premier point est bien située à la moyenne de $(2 + 2) / 2 = 2$.

Ceci étant dit et en espérant que cela a aidé à une meilleure compréhension, revenons à nos moutons...

Il apparaît assez clairement dans le graphique du dessus que seul le niveau du Facteur 1 influence la réponse, alors que le Facteur 2 n'influence en rien la réponse. Nous disons alors qu'il y a effet principal (localement) du Facteur1.

Sur le diagramme des interactions, nous avons la même information, mais sous une forme différente. Nous voyons que quel que soit le niveau du Facteur 2, les réponses sont horizontales et donc que celui-ci n'influence en rien les résultats. Nous sommes alors dans une situation où "(a priori) l'effet principal est (localement) le Facteur 1 et en absence d'interactions entre les facteurs".

Voyons maintenant le tableau suivant:

	Facteur 2	
Facteur 1	Niveau 1	Niveau 2
Niveau 1	4	2
Niveau 2	4	2

Tableau: 7.15 - Troisième exemple d'une petite ANOVA à deux facteurs sans répétition

Nous pouvons cette fois observer que le Facteur 2 a une influence mais pas le Facteur 1. Mais voyons aussi cela avec nos 2 types de représentations:

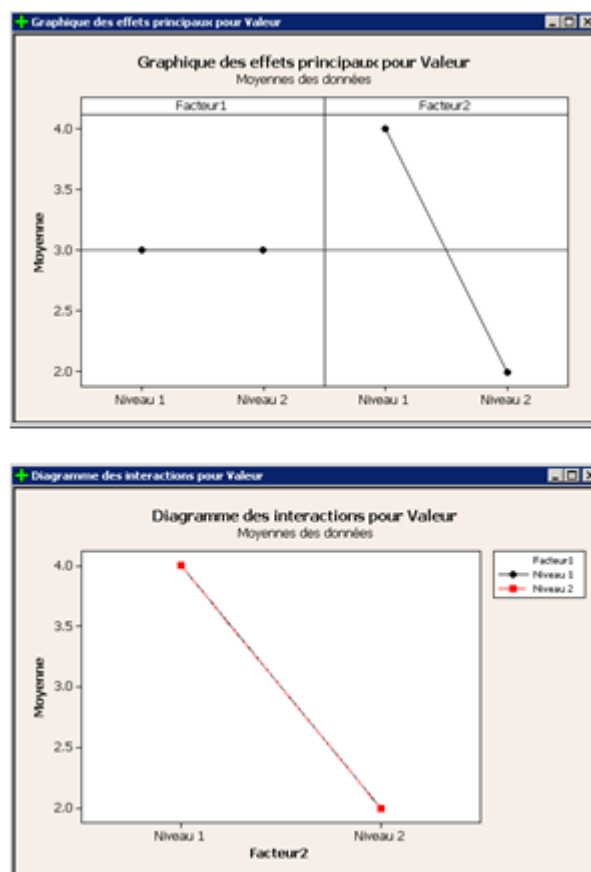


Figure: 7.51 - Graphique des effets principaux et diagramme des interactions avec Minitab 15

Nous observons bien sur le graphique que le Facteur 1 n'a aucune influence. Sur le diagramme du dessous c'est moins évident mais la superposition des deux droites montre que le Facteur 1 n'a pas d'influence. Nous disons alors qu'il y a "(a priori) effet principal (localement) du Facteur 2 et absence d'interactions entre les facteurs".

Considérons maintenant le tableau suivant:

	Facteur 2	
Facteur 1	Niveau 1	Niveau 2
Niveau 1	3	1
Niveau 2	5	3

Tableau: 7.16 - Quatrième exemple d'une petite ANOVA à deux facteurs sans répétition

Nous voyons que les deux facteurs ont une influence sur la réponse. Ce que montrent bien les deux représentations ci-dessous:

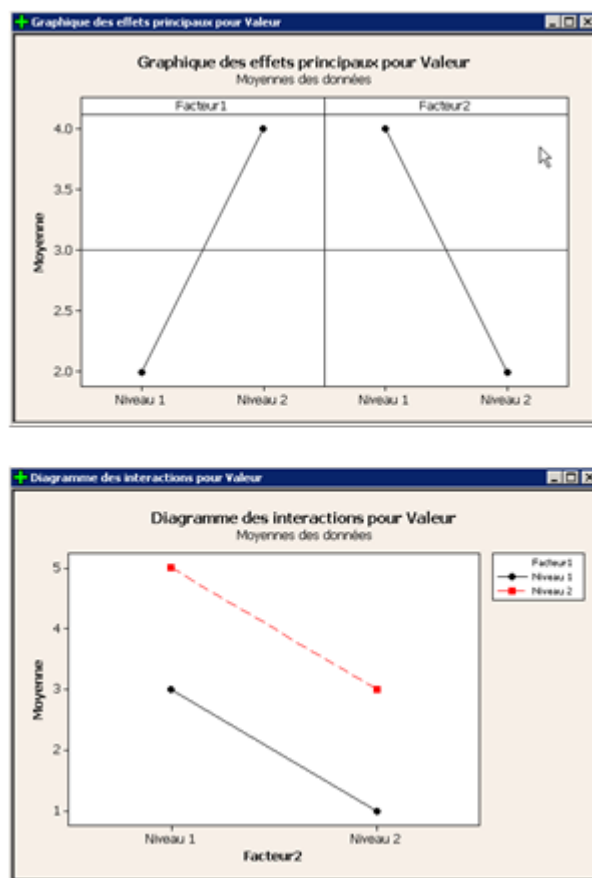


Figure: 7.52 - Graphique des effets principaux et diagramme des interactions avec Minitab 15

Nous observons bien sur le graphique du dessus que le Facteur 1 a une influence sur la réponse et qu'il en est de même du Facteur 2 (et en plus de la même amplitude quel que soit le sens!). Sur le graphique du dessous c'est moins évident mais la même conclusion est valable. Nous disons alors que "(a priori) les deux facteurs sont (localement) significatifs et sans interactions".

Passons au tableau suivant:

	Facteur 2	
Facteur 1	Niveau 1	Niveau 2
Niveau 1	2	4
Niveau 2	4	2

Tableau: 7.17 - Cinquième exemple d'une petite ANOVA à deux facteurs sans répétition

qui sous cette forme n'est pas trivial à interpréter. Mais avec les représentations nous avons tout de suite des informations plus pertinentes:

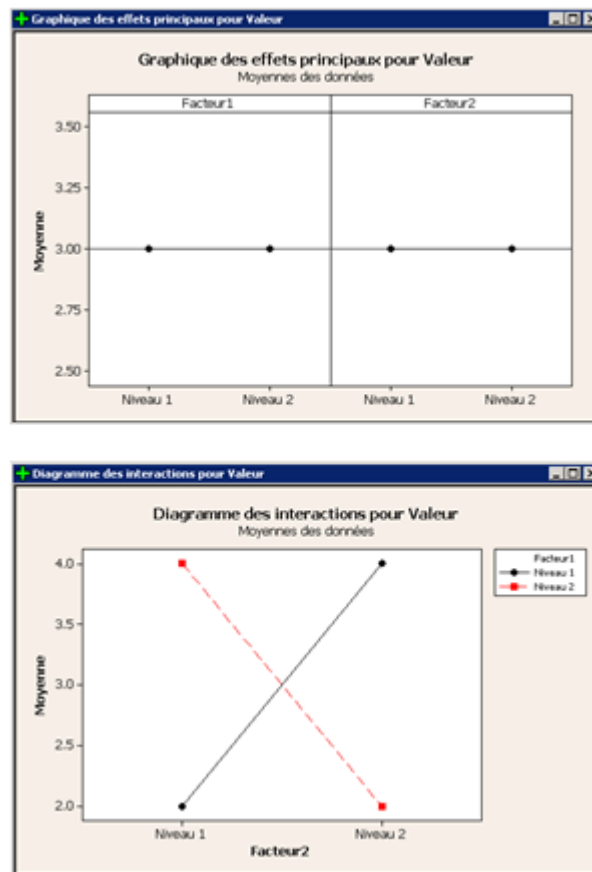


Figure: 7.53 - Graphique des effets principaux et diagramme des interactions avec Minitab 15

Nous observons bien sur le graphique ci-dessus qu'aucun des facteurs n'a d'influence sur la réponse a priori (même graphique qu'au tout début avec la même moyenne). Le diagramme du dessous nous donne une information complémentaire par contre (!!!): Les facteurs ont une influence croisée et comme cette influence croisée est de même amplitude, les effets s'annulent. Nous disons alors que les **"deux facteurs sont (localement) en interaction F1*F2"**.

Considérons maintenant le tableau suivant:

	Facteur 2	
Facteur 1	Niveau 1	Niveau 2
Niveau 1	1	3
Niveau 2	5	3

Tableau: 7.18 - Sixième exemple d'une petite ANOVA à deux facteurs sans répétition

Ce qui nous donne les deux représentations suivantes:

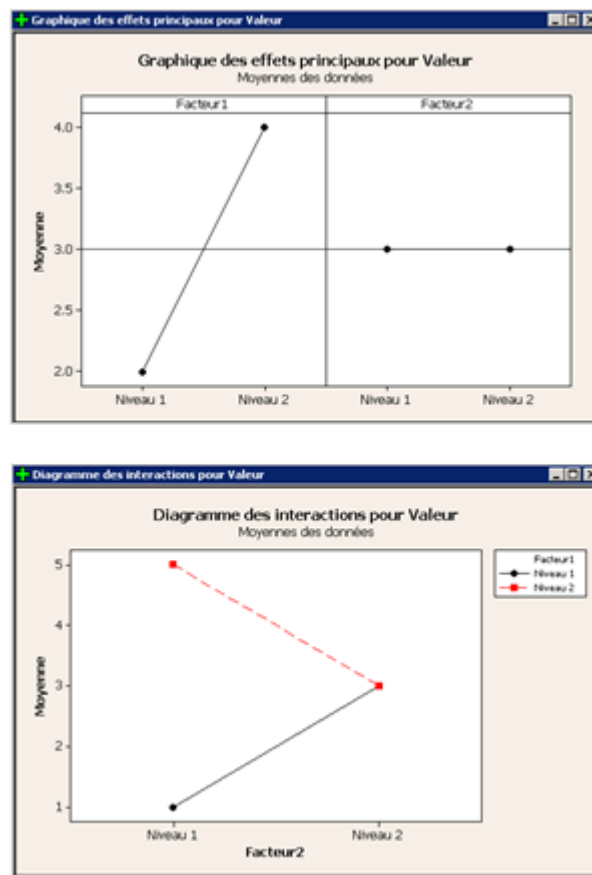


Figure: 7.54 - Graphique des effets principaux et diagramme des interactions avec Minitab 15

Nous observons bien sur le graphique du dessus que le Facteur 1 semble avoir une influence et que le Facteur 2 non (en moyenne!). Le diagramme des interactions du dessous nous donne, lui aussi, encore une fois, une information complémentaire (!!!): C'est que les facteurs sont en interaction. Nous disons alors que nous avons "(a priori) deux facteurs (localement) en interaction $F1 \cdot F2$ où l'influence du Facteur 1 est significative".

Tableau suivant:

	Facteur 2	
Facteur 1	Niveau 1	Niveau 2
Niveau 1	3	3
Niveau 2	5	1

Tableau: 7.19 - Septième exemple d'une petite ANOVA à deux facteurs sans répétition

Nous voyons que les deux facteurs ont une influence sur la réponse. Ce que montrent bien les deux représentations ci-dessous:

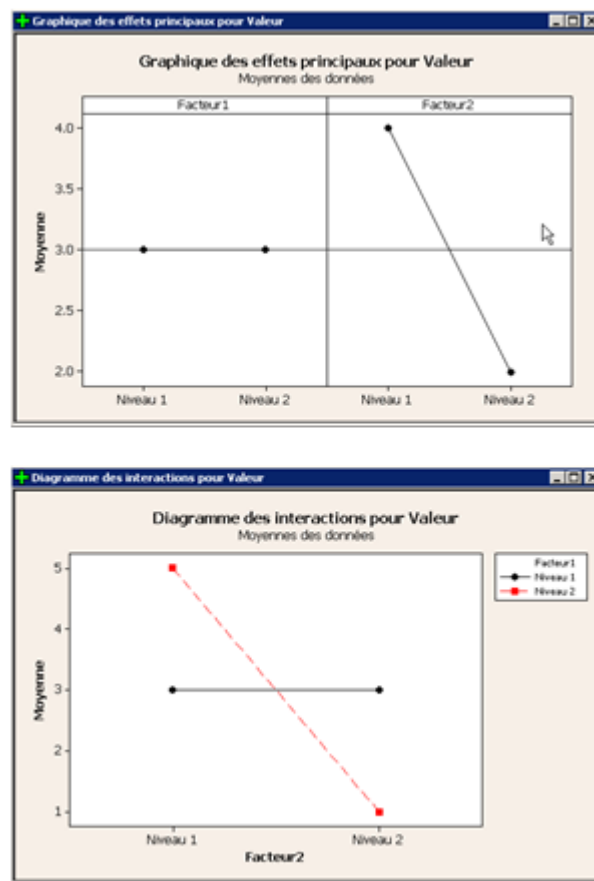


Figure: 7.55 - Graphique des effets principaux et diagramme des interactions avec Minitab 15

Nous disons ici que nous avons "(a priori) les deux facteurs (localement) en interaction $F1 \times F2$ où l'influence du Facteur 2 est significative".

Et enfin un dernier tableau

	Facteur 2	
Facteur 1	Niveau 1	Niveau 2
Niveau 1	1	1
Niveau 2	5	1

Tableau: 7.20 - Huitième exemple d'une petite ANOVA à deux facteurs sans répétition

qui nous donne les deux représentations:

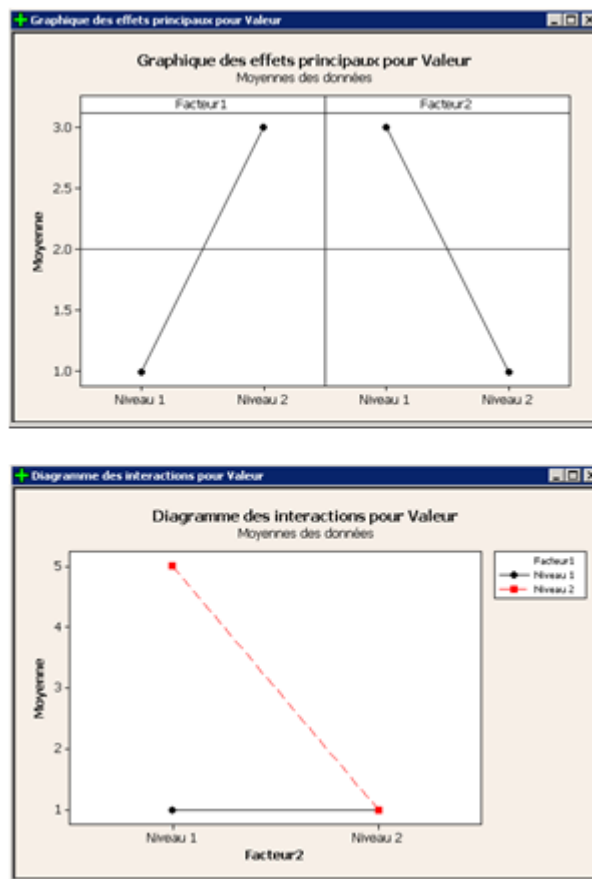


Figure: 7.56 - Graphique des effets principaux et diagramme des interactions avec Minitab 15

Nous disons ici que nous avons "(a priori) les deux facteurs (localement) en interaction $F1 \cdot F2$ où l'influence des deux facteurs est significative".

Remarque: Une croyance (communément répandue) de personnes qui manquent d'expérience dans les laboratoires consiste à penser que pour qu'une interaction soit significative il est nécessaire que les facteurs qui la composent le soient également.

Après tous ces tableaux, passons à partie mathématique:

Nous avons vu précédemment comment effectuer une analyse de la variance à un facteur. Pour rappel, cela consiste donc à faire un test d'égalité des espérances pour k échantillons indépendants de n variables aléatoires chacun (dans le cas où tous les échantillons ont donc le même nombre de mesures). Chaque échantillon étant considéré comme une expérience sur un sujet différent ou identique considéré alors comme un facteur variable indépendant!

Cependant il arrive dans la réalité que pour chaque échantillon on fasse varier un deuxième paramètre. Considéré alors comme un deuxième facteur variable. Nous parlons alors bien évidemment d'analyse de la variance à deux facteurs. De plus, nous allons considérer dans un premier temps pour simplifier les calculs que les variables aléatoires sont indépendantes! Donc un facteur n'a pas d'influence sur l'autre!!! En d'autres termes il n'y a pas d'interaction entre les facteurs. Nous parlons alors d'une "ANOVA à deux facteurs sans interactions".

Afin de déterminer la formulation du test à effectuer, rappelons que pour l'analyse de la variance à un facteur, nous avons décomposé la variance totale en la somme de la variance des moyennes

(interclasses) et de la variance résiduelle (intra-classes) telle que:

$$Q_T = Q_A + Q_R \quad (7.1056)$$

en explicitant le fait que nous comparions les échantillons $i = 1 \dots k$:

$$\begin{aligned} Q_T &= \sum_{\bar{y}} (x_{\bar{y}} - \bar{\bar{x}})^2 = \sum_i \sum_j (x_{\bar{y}} - \bar{\bar{x}})^2 = \sum_i \sum_j \left((x_{\bar{y}} - \bar{x}_i) + (\bar{x}_i - \bar{\bar{x}}) \right)^2 \\ &= \sum_i \sum_j (x_{\bar{y}} - \bar{x}_i)^2 + \sum_i \sum_j (\bar{x}_i - \bar{\bar{x}})^2 + 2 \sum_i \sum_j (x_{\bar{y}} - \bar{x}_i)(\bar{x}_i - \bar{\bar{x}}) \end{aligned} \quad (7.1057)$$

ce qui nous avait donné au final:

$$Q_T = \sum_i \sum_j (x_{\bar{y}} - \bar{x}_i)^2 + \sum_i \sum_j (\bar{x}_i - \bar{\bar{x}})^2 = Q_R + Q_A \quad (7.1058)$$

Pour l'ANOVA à deux facteurs nous partirons du tableau suivant ("Éch." est l'abréviation de "Échantillon"):

	Facteur A				
Facteur B	Éch.1	Éch.2	Éch....j	Éch.r	
Éch. 1	x_{11}	x_{12}	...	x_{1r}	Moyenne: $\bar{x}_1$
Éch.2	x_{21}	x_{22}	...	x_{2r}	Moyenne: $\bar{x}_2$
Éch. i	...	...	$x_{\bar{y}}$	...	Moyenne: $\bar{x}_i$
Éch. k	x_{k1}	x_{k2}	...	x_{kr}	Moyenne: $\bar{x}_k$
	Moyenne: $\bar{x}_1$	Moyenne: $\bar{x}_2$	Moyenne: $\bar{x}_j$	Moyenne: $\bar{x}_r$	$\bar{\bar{x}}$

Figure: 7.57 - Structure typique dite "croisée" d'une analyse de la variance à 2 facteurs sans répétition

pour lequel dans un laboratoire, le facteur maintenu fixe pendant qu'on fera varier l'autre sera appelé le "**facteur bloc**" et l'autre sera appelé le "**facteur de traitement**" et dans la pratique on fera en sorte que ce dernier ne soit pas effectué toujours dans le même ordre afin d'éliminer des éventuels effets d'inertie lors du passage d'un traitement à l'autre (les américains désignent les ANOVA à deux facteurs contrôlés sous les terme: "**randomized block design**").

Pour la suite, toute l'astuce consiste à décomposer la variance totale en comparant l'espérance des lignes (observations) indexées cette fois-ci avec $i = 1 \dots k$ et des colonnes (échantillons) indexées avec $j = 1 \dots r$ par rapport à la moyenne totale telle que:

$$\begin{aligned}
Q_T &= \sum_{ij} (x_{ij} - \bar{\bar{x}})^2 = \sum_i \sum_j (x_{ij} - \bar{\bar{x}})^2 = \sum_{ij} \left[(\bar{x}_i - \bar{\bar{x}}) + (\bar{x}_j - \bar{\bar{x}}) + (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}}) \right]^2 \\
&= \sum_{ij} \left[\left((\bar{x}_i - \bar{\bar{x}}) + (\bar{x}_j - \bar{\bar{x}}) \right)^2 + 2 \left((\bar{x}_i - \bar{\bar{x}}) + (\bar{x}_j - \bar{\bar{x}}) \right) (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}}) + (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2 \right] \quad (7.1059) \\
&= \sum_{ij} \left[(\bar{x}_i - \bar{\bar{x}})^2 + 2(\bar{x}_i - \bar{\bar{x}})(\bar{x}_j - \bar{\bar{x}}) + (\bar{x}_j - \bar{\bar{x}})^2 + 2 \left((\bar{x}_i - \bar{\bar{x}}) + (\bar{x}_j - \bar{\bar{x}}) \right) (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}}) \right. \\
&\quad \left. + (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2 \right]
\end{aligned}$$

Or, nous avons dans un premier temps:

$$\sum_{ij} 2(\bar{x}_i - \bar{\bar{x}})(\bar{x}_j - \bar{\bar{x}}) = 2 \underbrace{\sum_i (\bar{x}_i - \bar{\bar{x}})}_{=0} \underbrace{\sum_j (\bar{x}_j - \bar{\bar{x}})}_{=0} = 0 \quad (7.1060)$$

Donc il reste:

$$Q_T = \sum_{ij} \left[(\bar{x}_i - \bar{\bar{x}})^2 + (\bar{x}_j - \bar{\bar{x}})^2 + 2 \left((\bar{x}_i - \bar{\bar{x}}) + (\bar{x}_j - \bar{\bar{x}}) \right) (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}}) \right. \\
\left. + (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2 \right] \quad (7.1061)$$

Mais nous avons aussi:

$$\begin{aligned}
&\sum_{ij} 2 \left((\bar{x}_i - \bar{\bar{x}}) + (\bar{x}_j - \bar{\bar{x}}) \right) (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}}) \\
&= 2 \sum_{ij} \left((\bar{x}_i - \bar{\bar{x}}) + (\bar{x}_j - \bar{\bar{x}}) \right) (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}}) \quad (7.1062) \\
&= 2 \sum_{ij} (\bar{x}_i - \bar{\bar{x}}) (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}}) + 2 \sum_{ij} (\bar{x}_j - \bar{\bar{x}}) (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})
\end{aligned}$$

Pour la suite, indiquons d'abord que relativement à notre tableau, nous avons:

$$\begin{aligned}
\bar{\bar{x}} &= \frac{1}{r \cdot k} \sum_{ij} x_{ij} \\
\bar{x}_i &= \frac{1}{r} \sum_j x_{ij} \\
\bar{x}_j &= \frac{1}{k} \sum_i x_{ij}
\end{aligned} \quad (7.1063)$$

Il s'ensuit alors que:

$$\begin{aligned}
& \sum_{ij} (\bar{x}_i - \bar{\bar{x}}) (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}}) = \sum_i (\bar{x}_i - \bar{\bar{x}}) \left(\sum_j x_{ij} - \sum_j \bar{x}_i - \sum_j \bar{x}_j + \sum_j \bar{\bar{x}} \right) \\
& = \sum_i (\bar{x}_i - \bar{\bar{x}}) \left(r\bar{x}_i - \bar{x}_i \sum_j 1 - \sum_j \frac{1}{k} \sum_i x_{ij} + \bar{\bar{x}} \sum_j 1 \right) \\
& = \sum_i (\bar{x}_i - \bar{\bar{x}}) \left(r\bar{x}_i - \bar{x}_i r - \frac{1}{k} \sum_{ij} x_{ij} + \bar{\bar{x}} r \right) = \sum_i (\bar{x}_i - \bar{\bar{x}}) (-r\bar{\bar{x}} + \bar{\bar{x}}r) \\
& = \sum_i (\bar{x}_i - \bar{\bar{x}}) \cdot 0 = 0
\end{aligned} \tag{7.1064}$$

et il vient alors immédiatement que nous avons de même:

$$\sum_{ij} (\bar{x}_j - \bar{\bar{x}}) (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}}) = 0 \tag{7.1065}$$

Donc il reste au final:

$$\begin{aligned}
Q_T &= \sum_{ij} (\bar{x}_i - \bar{\bar{x}})^2 + \sum_{ij} (\bar{x}_j - \bar{\bar{x}})^2 + \sum_{ij} (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2 \\
&= k \sum_j (\bar{x}_j - \bar{\bar{x}})^2 + r \sum_i (\bar{x}_i - \bar{\bar{x}})^2 + \sum_{ij} (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2
\end{aligned} \tag{7.1066}$$

ce que nous noterons sur ce site de la manière condensée suivante:

$$\boxed{Q_T = Q_A + Q_B + Q_R} \tag{7.1067}$$

où Q_A, Q_B sont bien évidemment associés aux effets principaux (comparaison des moyennes marginales avec la moyenne totale).

Donc en comparaison à l'ANOVA à un facteur nous avons un terme supplémentaire pour la variance totale.

Dans l'ordre il est évident que la première somme des écarts par rapport au premier facteur colonne:

$$\sum_{ij} (\bar{x}_i - \bar{\bar{x}})^2 \tag{7.1068}$$

aura au même titre que l'ANOVA à un facteur $k - 1$ degrés de liberté. C'est-à-dire que sous les mêmes hypothèses que l'ANOVA à un facteur:

$$\frac{\sum_{ij} (\bar{x}_i - \bar{\bar{x}})^2}{\sigma^2} = \chi_{k-1}^2 \tag{7.1069}$$

La deuxième somme des écarts par rapport au deuxième facteur ligne:

$$\sum_{\tilde{y}} (\bar{x}_j - \bar{\bar{x}})^2 \quad (7.1070)$$

est nouvelle mais cependant on démontre de manière parfaitement identique au premier qu'elle aura $r - 1$ degrés de liberté. C'est-à-dire que sous les mêmes hypothèses que l'ANOVA à un facteur:

$$\frac{\sum_{\tilde{y}} (\bar{x}_j - \bar{\bar{x}})^2}{\sigma^2} = \chi_{r-1}^2 \quad (7.1071)$$

Pour la troisième somme qui suit obligatoirement aussi une loi du Khi-deux (étant donné que la variance totale suit une loi du Khi-deux et que les deux premiers termes de la somme aussi!):

$$\sum_{\tilde{y}} (x_{\tilde{y}} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2 \quad (7.1072)$$

c'est un peu plus délicat... mais il y a une astuce à la sauce physicienne...! Nous savons de par notre étude de l'ANOVA à un facteur que la somme des degrés de liberté de chaque terme doit être égale au nombre total de degrés de libertés. En d'autres termes, nous devons avoir pour l'ANOVA à deux facteurs:

$$ddl_T = N - 1 = k \cdot r - 1 = (k - 1) + (r - 1) + ??? \quad (7.1073)$$

Donc il manque bien évidemment:

$$??? = k \cdot r - 1 - [(k - 1) + (r - 1)] = k \cdot r - 1 - k + 1 - r + 1 = k \cdot r - k - r + 1 = (k - 1)(r - 1) \quad (7.1074)$$

Ainsi:

$$\frac{\sum_{\tilde{y}} (x_{\tilde{y}} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2}{\sigma^2} = \chi_{(k-1)(r-1)}^2 \quad (7.1075)$$

Donc nous avons alors le tableau suivant:

$Q_T = \sum_{\tilde{y}} (x_{\tilde{y}} - \bar{\bar{x}})^2$	$ddl_T = N - 1 = k \cdot r - 1$	$V_T = \frac{Q_T}{ddl_T}$
$Q_B = r \sum_i (\bar{x}_i - \bar{\bar{x}})^2$	$ddl_B = k - 1$	$V_B = \frac{Q_B}{ddl_B}$
$Q_A = k \sum_j (\bar{x}_j - \bar{\bar{x}})^2$	$ddl_A = r - 1$	$V_A = \frac{Q_A}{ddl_A}$
$Q_R = \sum_{\tilde{y}} (x_{\tilde{y}} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2$	$ddl_R = (k - 1)(r - 1)$	$V_R = \frac{Q_R}{ddl_R}$

(7.1076)

Enfin, le reste est exactement le même que pour l'ANOVA à un facteur simplement que nous avons deux tests à effectuer cette fois-ci qui sont:

$$F_{k-1,(k-1)(r-1)} = \frac{\frac{Q_A}{k-1}}{\frac{Q_R}{(k-1)(r-1)}} \text{ et } F_{r-1,(k-1)(r-1)} = \frac{\frac{Q_B}{r-1}}{\frac{Q_R}{(k-1)(r-1)}} \quad (7.1077)$$

Tous les calculs que nous avons faits précédemment sont très souvent représentés dans les logiciels sous la forme d'une table standardisée dont voici la forme et le contenu (c'est ainsi que le présente Microsoft Excel 11.8346 ou Minitab15.1.1 par exemple):

Somme des carrés	ddl χ^2	Moyenne des carrés	F	Valeur critique F
$Q_B = r \sum_i (\bar{x}_i - \bar{\bar{x}})^2$	r-1	$MSkB = \frac{Q_B}{r-1}$	$\frac{MSkB}{MSE}$	$P(F > F_{r-1,(k-1)(r-1)})$
$Q_A = k \sum_j (\bar{x}_j - \bar{\bar{x}})^2$	k-1	$MSkA = \frac{Q_A}{k-1}$	$\frac{MSkA}{MSE}$	$P(F > F_{k-1,(k-1)(r-1)})$
$Q_R = \sum_{ij} (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2$	(k-1)(r-1)	$MSE = \frac{Q_R}{(k-1)(r-1)}$		
$Q_T = \sum_{ij} (x_{ij} - \bar{\bar{x}})^2$	N-1			

Tableau: 7.21 - Terminologie et paramètres traditionnels d'un Tableau ANOVA (TAV) à deux facteurs sans répétition

et la condition d'acceptation de l'hypothèse d'égalité des moyennes pour chaque facteur est la même que pour l'ANOVA à un facteur (voir le serveur d'exercice pour un exemple pratique et détaillé avec Microsoft Excel 11.8346).

Évidemment, dans les développements ci-dessus, les facteurs A et B sont interchangeables dans les développements par symétrie!

13.6. ANALYSE DE LA VARIANCE À DEUX FACTEURS À MESURES RÉPÉTÉES

Jusqu'à présent nous avons examiné des ANOVA sur des expériences à un ou deux facteurs fixes (autrement dit: une ou deux variables catégorielles). Dans le cas à deux facteurs, nous avons considéré que pour chaque combinaison de facteurs nous n'avions qu'une seule mesure (cellule). Or, il peut arriver (et c'est préférable) que nous ayons plusieurs mesures pour une combinaison!

Nous qualifions ce type d'étude de "**plan expérimental à mesures répétées**" et les résultats seront traités avec une analyse de la variance à deux facteurs à mesures répétées et avec interactions! Il s'agit d'un outil extrêmement important puisqu'il permet de valider des études menées par plusieurs laboratoires (ou employés) indépendants et il est également associé à de nombreux autres outils statistiques comme celui de l'étude de la reproductibilité et de la répétabilité (Étude R&R) pour ne citer que le plus connu dans le domaine industriel.

Il faut comprendre qu'il est obligatoire dans le domaine de la statistique d'associer les interactions entre facteurs systématiquement lorsque nous avons affaire à une expérience à mesures répétées. Ceci pour la simple raison que le terme mathématique d'interaction n'apparaît que dans cette situation.

Ainsi, il peut être intuitif (avant même de le démontrer) qu'une ANOVA à deux facteurs (fixes) à mesures répétées contient une interaction double, et deux effets principaux. Une ANOVA à trois

facteurs (fixes) et à mesures répétées aura in extenso une interaction triple, trois interactions doubles et 3 effets principaux. Et ainsi de suite...

Avant de commencer, nous allons considérer le tableau de mesures suivant où l'abréviation "Éch." fait référence au mot "échantillon":

	Facteur A				
Facteur B	Éch. 1	Éch. 2	Éch ...j	Éch. r	Moyenne
Éch.1	x_{111}	x_{112}	...	x_{11r}	
Réplication 2	x_{211}	x_{212}	...	x_{21r}	
Réplication m	...	...	...	...	
Réplication n	x_{n11}	x_{n12}	...	x_{n1r}	
Moyenne Éch. 1	$\bar{x}_{11}$	$\bar{x}_{12}$	...	$\bar{x}_{13}$	$\bar{x}_{1.}$
Éch.2	x_{121}	x_{122}	...	x_{12r}	
Réplication 2	x_{221}	x_{222}	...	x_{22r}	
Réplication m	...	...	...	...	
Réplication n	x_{n21}	x_{n22}	...	x_{n2r}	
Moyenne Éch. 2	$\bar{x}_{21}$	$\bar{x}_{22}$	...	$\bar{x}_{2r}$	$\bar{x}_{2.}$
Éch.i	...	...	x_{1ij}	...	
Réplication 2	...	...	...	...	
Réplication m	...	...	...	...	
Réplication n	...	...	x_{nij}	x_{nir}	
Moyenne Éch. i	...	...	$\bar{x}_{ij}$	...	$\bar{x}_{i.}$
Éch.k	x_{1k1}	x_{1k2}	...	x_{1kr}	
Réplication 2	x_{2k1}	x_{2k2}	...	x_{2kr}	
Réplication m	...	...	...	...	
Réplication n	x_{nk1}	x_{nk2}	...	x_{nkr}	
Moyenne Éch. k	$\bar{x}_{k1}$	$\bar{x}_{k2}$	...	$\bar{x}_{kr}$	$\bar{x}_{k.}$
Moyenne	$\bar{x}_{.1}$	$\bar{x}_{.2}$	$\bar{x}_{.j}$	$\bar{x}_{.r}$	$\bar{\bar{x}}_{..}$

Figure: 7.58 - Structure typique dite "croisée" d'une analyse de la variance à 2 facteurs avec répétition

avec les propriétés habituelles des moyennes (pour rappel):

$$\begin{aligned}
 \bar{x}_{i.} &= \frac{1}{r} \sum_{j=1}^r \bar{x}_{ij} = \frac{1}{r \cdot n} \sum_{m=1}^n \sum_{j=1}^r x_{mij} \\
 \bar{x}_{.j} &= \frac{1}{k} \sum_{i=1}^k \bar{x}_{ij} = \frac{1}{k \cdot n} \sum_{m=1}^n \sum_{i=1}^k x_{mij} \\
 \bar{\bar{x}}_{..} &= \frac{1}{r} \sum_{j=1}^r \bar{x}_{.j} = \frac{1}{k} \sum_{i=1}^k \bar{x}_{i.} = \frac{1}{r \cdot n \cdot k} \sum_{m=1}^n \sum_{i=1}^k \sum_{j=1}^r x_{mij}
 \end{aligned}
 \tag{7.1078}$$

Et rappelons que pour l'ANOVA à deux facteurs sans répétitions (et donc sans interactions), toute l'astuce avait consisté à décomposer la variance totale en comparant la moyenne des lignes indexées avec $i = 1 \dots k$ et des colonnes indexées avec $j = 1 \dots r$ par rapport à la moyenne totale.

L'idée va maintenant être à peu près la même à la différence que nous allons comparer l'espérance des lignes indexées avec $i = 1 \dots k$ et des colonnes indexées avec $j = 1 \dots r$ non seulement par rapport à la moyenne totale mais aussi à celle de chaque ligne et de chaque colonne.

Pour cela nous repartons de ce que nous avons obtenu pour l'ANOVA à deux facteurs sans répétition:

$$Q_T = \sum_{ij} (x_{ij} - \bar{\bar{x}})^2 = \sum_{ij} (\bar{x}_i - \bar{\bar{x}})^2 + \sum_{ij} (\bar{x}_j - \bar{\bar{x}})^2 + \sum_{ij} (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2 \quad (7.1079)$$

mais dont la notation sera juste adaptée au contexte:

$$Q_T = \sum_{ij} (\bar{x}_{ij} - \bar{\bar{x}})^2 = r \sum_i (\bar{x}_i - \bar{\bar{x}})^2 + k \sum_j (\bar{x}_j - \bar{\bar{x}})^2 + \sum_{ij} (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2 \quad (7.1080)$$

Il est évident qu'avec cette écriture l'ANOVA à deux facteurs sans répétition deviendrait:

$$\begin{aligned} Q_T &= \sum_{ij} (x_{ij} - \bar{\bar{x}})^2 \rightarrow Q_T = \sum_{ij} (\bar{x}_{ij} - \bar{\bar{x}})^2 \\ Q_A &= r \sum_i (\bar{x}_i - \bar{\bar{x}})^2 \rightarrow Q_A = r \sum_i (\bar{x}_i - \bar{\bar{x}})^2 \\ Q_B &= k \sum_j (\bar{x}_j - \bar{\bar{x}})^2 \rightarrow Q_B = k \sum_j (\bar{x}_j - \bar{\bar{x}})^2 \end{aligned} \quad (7.1081)$$

Mais dans le cas présent, il nous faut rajouter une sommation pour les répétitions et adapter la notation pour les mesures. Donc, sans refaire tous les développements (c'est un peu culotté mais bon...), nous obtenons déjà directement:

$$Q_T = \sum_{mij} (x_{mij} - \bar{\bar{x}})^2 = nr \sum_i (\bar{x}_i - \bar{\bar{x}})^2 + nk \sum_j (\bar{x}_j - \bar{\bar{x}})^2 + \sum_{ijm} (x_{mij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2 \quad (7.1082)$$

où dans l'ordre, m est la répétition de l'échantillon i du facteur A et de l'échantillon j du facteur B .

Il vient alors bien évidemment les variances interclasses pour les facteurs A et B qui sont immédiates:

$$\begin{aligned} Q_A &= nr \sum_i (\bar{x}_i - \bar{\bar{x}})^2 \\ Q_B &= nk \sum_j (\bar{x}_j - \bar{\bar{x}})^2 \end{aligned} \quad (7.1083)$$

où Q_A, Q_B sont bien évidemment encore une fois associées aux effets principaux (comparaisons des moyennes marginales avec la moyenne totale).

Maintenant, nous allons jouer un peu en introduisant sous la somme, en plus et en moins, dans le dernier terme:

$$\sum_{ijm} \left(x_{mij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..} \right)^2 \quad (7.1084)$$

la moyenne des répétitions:

$$\bar{x}_{ij} \quad (7.1085)$$

que nous retrouverons in fine dans la somme des carrés totale:

$$\begin{aligned} \sum_{mij} \left(x_{mij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..} \right)^2 &= \sum_{mij} \left(x_{mij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..} + \bar{x}_{ij} - \bar{x}_{ij} \right)^2 \\ &= \sum_{mij} \left(\left(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..} \right) + \left(x_{mij} - \bar{x}_{ij} \right) \right)^2 = \sum_{mij} \left(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..} \right)^2 \\ &\quad + 2 \sum_{mij} \left(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..} \right) \left(x_{mij} - \bar{x}_{ij} \right) + \sum_{mij} \left(x_{mij} - \bar{x}_{ij} \right)^2 \\ &= n \sum_{ij} \left(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..} \right)^2 + 2 \sum_{mij} \left(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..} \right) \left(x_{mij} - \bar{x}_{ij} \right) + \sum_{mij} \left(x_{mij} - \bar{x}_{ij} \right)^2 \end{aligned} \quad (7.1086)$$

Bien entendu, nous reconnaissons assez vite la variance intra-classes (appelée aussi souvent "**erreur résiduelle**" ou simplement dans le cas particulier de l'ANOVA à deux facteurs avec répétition "**erreur de répétabilité**"):

$$Q_R = \sum_{mij} \left(x_{mij} - \bar{x}_{ij} \right)^2 \quad (7.1087)$$

et le terme que nous pouvons interpréter (par comparaison avec l'ANOVA à deux facteurs sans répétitions) comme étant la variance d'interaction:

$$Q_{A \times B} = n \sum_{ij} \left(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..} \right)^2 = n \sum_{ij} \left(\left(\bar{x}_{ij} - \bar{\bar{x}}_{..} \right) - \left(\bar{x}_{i.} - \bar{\bar{x}}_{..} \right) - \left(\bar{x}_{.j} - \bar{\bar{x}}_{..} \right) \right)^2 \quad (7.1088)$$

Mais si notre hypothèse est juste, le terme:

$$2 \sum_{mij} \left(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..} \right) \left(x_{mij} - \bar{x}_{ij} \right) \quad (7.1089)$$

doit s'annuler. Vérifions cela

$$\begin{aligned} \sum_{mij} \left(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..} \right) \left(x_{mij} - \bar{x}_{ij} \right) &= \sum_{mij} \left(x_{mij} - \bar{x}_{ij} \right) \left(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..} \right) \\ &= \sum_i \sum_j \sum_m \left(x_{mij} - \bar{x}_{ij} \right) \left(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..} \right) \end{aligned} \quad (7.1090)$$

et donc pour i et j fixés il vient:

$$\underbrace{(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}) \sum_m (x_{mij} - \bar{x}_{ij})}_{=0} = 0 \quad (7.1091)$$

Et donc la sommation sur tous les i et j sera aussi nulle par extension. Ceux qui ont un doute quant à l'annulation des deux termes du développement ci-dessus, pourront peut-être se rassurer en faisant une application numérique.

Donc au final:

$$\begin{aligned} Q_T &= \sum_{mij} (x_{mij} - \bar{\bar{x}})^2 \\ &= \underbrace{nk \sum_j (\bar{x}_{.j} - \bar{\bar{x}})^2}_{Q_A} + \underbrace{nr \sum_i (\bar{x}_{i.} - \bar{\bar{x}})^2}_{Q_B} + \underbrace{n \sum_{ij} (\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}})^2}_{Q_{A \times B}} + \underbrace{\sum_{mij} (x_{mij} - \bar{x}_{ij})^2}_{Q_R} \end{aligned} \quad (7.1092)$$

où pour rappel, n est donc le nombre de répétitions, r le nombre d'échantillons du facteur A et k le nombre d'échantillons du facteur B (ces deux derniers paramètres sont souvent confondus par ceux qui font les calculs à la main).

Donc en comparaison à l'ANOVA à deux facteurs sans répétition, nous avons un terme supplémentaire pour la variance totale.

Dans l'ordre il est évident que la première somme des écarts par rapport au premier facteur colonne:

$$\sum_j (\bar{x}_{.j} - \bar{\bar{x}})^2 \quad (7.1093)$$

aura au même titre que l'ANOVA à un facteur et l'ANOVA à deux facteurs sans répétition $k - 1$ degrés de liberté. C'est-à-dire que sous les mêmes hypothèses que ces deux ANOVA, nous avons:

$$\frac{\sum_j (\bar{x}_{.j} - \bar{\bar{x}})^2}{\sigma^2} = \chi_{k-1}^2 \quad (7.1094)$$

La deuxième somme des écarts par rapport au deuxième facteur ligne:

$$\sum_i (\bar{x}_{i.} - \bar{\bar{x}})^2 \quad (7.1095)$$

aura sous les mêmes hypothèses la propriété:

$$\frac{\sum_i (\bar{x}_{i.} - \bar{\bar{x}})^2}{\sigma^2} = \chi_{r-1}^2 \quad (7.1096)$$

Grâce au raisonnement effectué à l'aide de l'ANOVA à deux facteurs sans répétition, nous savons que pour le terme d'interaction:

$$\sum_{ij} (\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}})^2 \quad (7.1097)$$

nous avons:

$$\frac{\sum_{ij} (\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}})^2}{\sigma^2} = \chi_{(k-1)(r-1)}^2 \quad (7.1098)$$

Il reste à déterminer le nombre de degrés de liberté du dernier terme:

$$\sum_{mij} (x_{mij} - \bar{x}_{ij})^2 \quad (7.1099)$$

Pour ce faire, nous procédons de la même manière qu'avec l'ANOVA à deux facteurs sans répétitions. Nous savons de par notre étude de l'ANOVA à un facteur que la somme des degrés de liberté de chaque terme doit être égale au nombre total de degrés de liberté. En d'autres termes, nous devons avoir pour l'ANOVA à deux facteurs:

$$ddl_T = N - 1 = n \cdot k \cdot r - 1 = (k-1) + (r-1) + (k-1)(r-1) + ??? \quad (7.1100)$$

Donc il manque bien évidemment:

$$\begin{aligned} ??? &= nkr - 1 - [(k-1) + (r-1) + (k-1)(r-1)] = nkr - 1 - k + 1 - r + 1 - (kr - k - r + 1) \\ &= nkr - 1 - k + 1 - r + 1 - kr + k + r - 1 = nkr - kr = N - kr = kr(n-1) \end{aligned} \quad (7.1101)$$

Ainsi:

$$\frac{\sum_{mij} (x_{mij} - \bar{x}_{ij})^2}{\sigma^2} = \chi_{N-kr}^2 \quad (7.1102)$$

Donc nous avons alors le tableau suivant:

$Q_T = \sum_{mij} (x_{mij} - \bar{\bar{x}})^2$	$ddl_T = N - 1 = nkr - 1$	$V_T = \frac{Q_T}{ddl_T}$	(7.1103)
$Q_B = nr \sum_i (\bar{x}_{i.} - \bar{\bar{x}})^2$	$ddl_B = r - 1$	$V_B = \frac{Q_B}{ddl_B}$	
$Q_A = nk \sum_j (\bar{x}_{.j} - \bar{\bar{x}})^2$	$ddl_A = k - 1$	$V_A = \frac{Q_A}{ddl_A}$	
$Q_{A \times B} = n \sum_{ij} (\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}})^2$	$ddl_{A \times B} = (k-1)(r-1)$	$V_{A \times B} = \frac{Q_{A \times B}}{ddl_{A \times B}}$	
$Q_R = \sum_{mij} (x_{mij} - \bar{x}_{ij})^2$	$ddl_R = N - kr$	$V_R = \frac{Q_R}{ddl_R}$	

Enfin, le reste est exactement le même que pour l'ANOVA à deux facteurs sans réplication simplement

que nous avons trois tests à effectuer cette fois-ci qui sont:

$$F_{k-1, N-kr} = \frac{\frac{Q_A}{k-1}}{\frac{Q_R}{N-kr}}, \quad F_{r-1, N-kr} = \frac{\frac{Q_B}{r-1}}{\frac{Q_R}{N-kr}}, \quad F_{(k-1)(r-1), N-kr} = \frac{\frac{Q_{A \times B}}{(k-1)(r-1)}}{\frac{Q_R}{N-kr}} \quad (7.1104)$$

Tous les calculs que nous avons faits précédemment sont très souvent représentés dans les logiciels sous la forme d'une table standardisée donc voici la forme et le contenu (c'est ainsi que le présente Microsoft Excel 11.8346 ou Minitab 15.1.1 par exemple):

Somme des carrés	ddl χ^2	Moyenne des carrés	F	Valeur critique F
$Q_B = nr \sum_i (\bar{x}_{i.} - \bar{\bar{x}}_{..})^2$	r-1	$MS_{kB} = \frac{Q_B}{r-1}$	$\frac{MS_{kB}}{MSE}$	$P(F > F_{r-1, N-kr})$
$Q_A = nk \sum_j (\bar{x}_{.j} - \bar{\bar{x}}_{..})^2$	k-1	$MS_{kA} = \frac{Q_A}{k-1}$	$\frac{MS_{kA}}{MSE}$	$P(F > F_{k-1, N-kr})$
$Q_{A \times B}$ $= n \sum_{ij} (\bar{x}_{ij} - \bar{\bar{x}}_{i.} - \bar{\bar{x}}_{.j} + \bar{\bar{x}}_{..})^2$	(k-1)(r-1)	MS_{kAB} $= \frac{Q_{A \times B}}{(k-1)(r-1)}$	$\frac{MS_{kAB}}{MSE}$	$P(F > F_{(k-1)(r-1), N-kr})$
$Q_R = \sum_{mij} (x_{mij} - \bar{\bar{x}}_{..})^2$	N-kr	$MSE = \frac{Q_R}{N-kr}$		
$Q_T = \sum_{mij} (x_{mij} - \bar{\bar{x}}_{..})^2$	N-1			

Tableau: 7.22 - Terminologie et paramètres traditionnels d'un Tableau ANOVA (TAV) à deux facteurs avec répétition

et la condition d'acceptation de l'hypothèse d'égalité des moyennes pour chaque facteur est la même que pour l'ANOVA à un facteur (voir le serveur d'exercice pour un exemple pratique et détaillé avec Microsoft Excel 11.8346).

Évidemment, dans les développements ci-dessus, les facteurs A et B sont interchangeables dans les développements par symétrie!

13.7. ANOVA MULTIFACTORIELLE À MESURES RÉPÉTÉES

L'ANOVA multifactorielle à mesures répétées ou appelée aussi "**ANOVA multifactorielle à variables catégorielles et mesures répétées**" (et très rarement "**ANOVA équilibrée**") est simplement le nom sous lequel les spécialistes désignent les ANOVA suivantes:

- ANOVA à trois facteurs (fixes) avec ou sans répétition
- ANOVA à quatre facteurs (fixes) avec ou sans répétition
- ANOVA à cinq facteurs (fixes) avec ou sans répétition
- etc.

Évidemment, les ANOVA à un et deux facteurs (fixes) font aussi partie de la famille de l'ANOVA

multifactorielle mais elles sont rarement signalées en tant que tel dans les logiciels de statistiques et sont souvent disponibles de façon explicite dans les menus de ces mêmes logiciels (car ce sont les deux plus utilisées dans les écoles). Il faut savoir aussi que la majorité des logiciels de statistiques gèrent des ANOVA multifactorielles jusqu'à 15 facteurs fixes (variables catégorielles) à condition que le plan soit équilibré (c'est à dire que pour chaque niveau de chaque facteur, il y ait un nombre identique de mesures). Un tableur (comme Microsoft Excel) gère le plus souvent les ANOVA jusqu'à un maximum deux facteurs (fixes).

Bon maintenant le lecteur risque d'être déçu (bon je suis aussi déçu de n'avoir qu'une seule vie...) car franchement je ne souhaite pas refaire les développements vus plus haut pour les ANOVA à un facteur et deux facteurs (fixes) pour 3, 4 et ce jusqu'à 15 facteurs car cela prendrait plus de 100 pages A4 sous une forme pédagogique et claire et en plus c'est basé toujours sur la même mécanique de développement (la théorie généralisée de l'ANOVA bien qu'étant beaucoup plus courte, elle est à mon goût indigeste).

Remarque: Les ANOVA non équilibrées (non balancées) nécessitent un choix subtil de la manière de calculer les variances. En fonction de la manière de calculer, nous parlons d'ANOVA de type I, II ou III. Le choix de la méthode de calcul est même en ce début de 21ème siècle sujet à de vifs débats entre spécialistes. Raison pour laquelle nous nous abstiendrons d'étudier ce cas-là.

13.8. TEST DE C DE COCHRAN

Le test C de Cochran a pour objet la vérification de l'homogénéité des variances concernant plusieurs populations. Il s'agit d'un des tests préalables ou postérieurs (post hoc) utiles avant de faire une ANOVA balancée (équilibrée) et qui est recommandé par la norme ISO 5725 (de même que le test de Tukey que nous verrons beaucoup plus loin).

Bien que l'idée du test de Cochran soit empirique, elle est néanmoins intuitive comme le sont les définitions des tests de Grubbs et Dixon. Pourquoi alors présentons-nous sur ce site en détails le test C de Cochran alors que nous avons mentionné que nous ne le ferions pas pour le test de Grubbs et Dixon? La raison en fait est simple: le test de Grubbs et Dixon nécessite des simulations par Monte-Carlo pour déterminer les valeurs critiques de rejet ou d'acceptation de l'hypothèse nulle, alors que la valeur critique du test C de Cochran peut être obtenue relativement facilement analytiquement.

Ceci étant dit..., nous définissons le test C de Cochran par le rapport:

$$C_j = \frac{S_j^2}{\sum_{i=1}^N S_i^2} \quad (7.1105)$$

où les S_i sont les variances non biaisées des différentes sources de données au nombre de N composées chacune de n échantillons et l'hypothèse nulle est intuitivement l'égalité des variances contre l'hypothèse alternative qui est qu'une des variances est trop grande (donc mauvaise) et rejetée parce que aberrante.

La norme ISO 5725 recommande de réitérer ce test jusqu'à ce qu'il n'y ait plus aucune variance aberrante (donc trop grande ET éloignée des autres variances).

Pour déterminer la valeur critique, inversons la définition du test C de Cochran et faisons quelques manipulations algébriques élémentaires:

$$C_j^{-1} = \frac{\sum_{i=1}^N S_i^2}{S_j^2} = \frac{S_j^2 + \left(\sum_{i=1}^N S_i^2 - S_j^2 \right)}{S_j^2} = 1 + \frac{\sum_{i=1, i \neq j}^N S_i^2}{S_j^2} \quad (7.1106)$$

Nous remarquons qu'à peu de choses près, le deuxième terme de la dernière égalité ressemble presque à une loi de Fisher. Comme la loi de Fisher n'est pas stable par l'addition, il faudrait que nous trouvions une manière de transformer le terme:

$$\sum_{i=1, i \neq j}^N S_i^2 \quad (7.1107)$$

en une variance unique. L'idée est alors relativement simple mais encore fallait-il y penser... Nous savons que les S_i sont des variances non biaisés avec un facteur $1/(n-1)$. Donc si les N échantillons (niveaux) sont tous indépendants, la variance globale est alors par stabilité de la loi Normale et en reprenant les notations de l'ANOVA:

$$\begin{aligned} S_{global-j}^2 &= \frac{1}{N-1} \sum_{i=1, i \neq j}^N S_i^2 = \frac{1}{N-1} \sum_{i=1, i \neq j}^N \frac{Q_i}{n-1} = \frac{1}{(N-1)(n-1)} \sum_{i=1, i \neq j}^N Q_i \\ &= \frac{1}{(N-1)(n-1)} Q_{global-j} = \frac{Q_{global-j}}{(N-1)(n-1)} \end{aligned} \quad (7.1108)$$

Dès lors:

$$C_j^{-1} = 1 + \frac{\frac{N-1}{N-1} \sum_{i=1, i \neq j}^N S_i^2}{S_j^2} = 1 + (N-1) \frac{\frac{1}{N-1} \sum_{i=1, i \neq j}^N S_i^2}{S_j^2} = 1 + (N-1) \frac{\frac{Q_{global-j}}{(N-1)(n-1)}}{\frac{Q_j}{n-1}} \quad (7.1109)$$

Nous reconnaissons donc dans la dernière égalité le rapport de deux variances au carré. Nous avons alors identiquement à ce que nous avons démontré lors de notre étude de l'ANOVA à un facteur sans répétitions:

$$F_{n-1, (N-1)(n-1)} = \frac{S_j^2}{S_{global-j}^2} = \frac{\frac{Q_j}{n-1}}{\frac{Q_{global-j}}{(N-1)(n-1)}} \quad (7.1110)$$

et donc il vient:

$$C_j^{-1} = 1 + \frac{N-1}{F_{n-1, (N-1)(n-1)}} \quad (7.1111)$$

qui est dès lors indépendant de j et donc le test unilatéral gauche (puisque par définition le rapport du test de Cochran doit être le plus petit possible) C de Cochran aura pour valeur critique:

$$C(\alpha, N, n) = \left(1 + \frac{N-1}{F_{\alpha, n-1, (N-1)(n-1)}} \right)^{-1} \quad (7.1112)$$

Il y a cependant un petit couac avec la relation précédente. Effectivement, nous savons que nous devons itérer le test plusieurs fois. Mais bien évidemment, plus nous effectuons de test sur un échantillon de données, plus grande est la probabilité de rejeter l'hypothèse nulle à un moment ou à un autre. Ce problème est appelé "**inflation du niveau de confiance**" et dans la vie réelle nous connaissons bien son application: plus on attend, plus la probabilité qu'un événement rare ait lieu est élevée. Bien évidemment, il faut alors réduire la valeur du seuil mais dès lors cela augmente la difficulté de détecter les vrais effets. Dès lors, la démarche est la suivante.

Si nous considérons un test avec un niveau de signification $1 - \alpha$ (correspondant donc à la probabilité cumulée de ne pas faire une erreur de type I) et que nous réitérons celui-ci de manière indépendante une deuxième fois. Alors, si les tests sont indépendants, de par l'axiome des probabilités, la probabilité de ne pas faire une erreur du type I sera le produit des probabilités:

$$(1 - \alpha)(1 - \alpha) = (1 - \alpha)^2 \quad (7.1113)$$

et ainsi de suite pour n tests. Nous remarquons alors très vite que la probabilité cumulée de ne pas faire une erreur de type I décroît très vite. Par exemple, pour 10 tests réitérés indépendants avec un niveau de 5%, nous avons alors:

$$(1 - 5\%)^{10} \cong 59\% \quad (7.1114)$$

ce qui est catastrophique! Dès lors, si nous voulons un niveau de confiance sur tests réitérés d'une certaine valeur que nous noterons α_r , il paraît évident qu'il faut résoudre l'équation suivante:

$$(1 - \alpha)^n = (1 - \alpha_r) \quad (7.1115)$$

Soit (relation appelée parfois "**équation de Šidàk**"):

$$\alpha = 1 - (1 - \alpha_r)^{1/n} = 1 - \frac{1}{(1 - \alpha_r)^n} \quad (7.1116)$$

et avec un développement de Taylor au deuxième ordre il vient (cf. chapitre Suites Et Séries):

$$\alpha = 1 - (1 - \alpha_r)^{1/n} \cong 1 - (1 - n\alpha_r) = n\alpha_r \quad (7.1117)$$

que nous appelons "**approximation de Bonferroni**", parfois "**approximation de Boole**" ou encore "**approximation de Dunn**". Donc au final, nous avons:

$$C(\alpha, N, n) = \left(1 + \frac{N-1}{F_{\alpha/N, n-1, (N-1)(n-1)}} \right)^{-1} \quad (7.1118)$$

Que nous pouvons calculer avec la version anglaise de Microsoft Excel 14.0.6123 à l'aide de la formule:

$$=1/(1+(N-1)/FINV(ALPHA/N;n-1;(N-1)*(n-1))) \quad (7.1119)$$

13.9. TEST D'AJUSTEMENT DU KHI-DEUX

Nous allons étudier ici notre premier test d'ajustement non-paramétrique, un des plus connus certainement et des plus simples.

Supposons qu'une variable statistique suive une loi de probabilité P . Si nous tirons un échantillon dans la population correspondant à cette loi, la distribution observée s'écartera toujours plus ou moins de la distribution théorique, compte tenu des fluctuations d'échantillonnage.

Généralement, nous ne connaissons ni la forme de la loi P , ni la valeur de ses paramètres. C'est la nature du phénomène étudié et l'analyse de la distribution observée qui permettent de choisir une loi susceptible de convenir et d'en estimer les paramètres.

Les écarts entre la loi théorique et la distribution observée peuvent être attribués soit aux fluctuations d'échantillonnage, soit au fait que le phénomène ne suit pas, en réalité, la loi supposée.

En gros, si les écarts sont suffisamment faibles, nous admettrons qu'ils sont imputables aux fluctuations aléatoires et nous accepterons la loi retenue ; au contraire, s'ils sont trop élevés, nous en concluons qu'ils ne peuvent pas être expliqués par les seules fluctuations et que le phénomène ne suit pas la loi retenue.

Pour évaluer ces écarts et pouvoir prendre une décision, il faut:

1. Définir la mesure de la distance entre distribution empirique et distribution théorique résultant de la loi retenue.
2. Déterminer la loi de probabilité suivie par cette variable aléatoire donnant la distance.
3. Énoncer une règle de décision permettant de dire, d'après la distribution observée, si la loi retenue est acceptable ou non.

Premièrement, nous aurons pour cela besoin du théorème central limite et deuxièmement rappelons que lors de la construction de la loi Normale, nous avons montré que la variable:

$$x = \frac{k - np}{\sqrt{npq}} = \frac{k - np}{\sqrt{np(1-p)}} \quad (7.1120)$$

suivait une loi Normale centrée réduite lorsque n tendait vers l'infini (condition de Laplace) et que la probabilité p était très petite.

En pratique, l'approximation est tout à fait acceptable... dans certaines entreprises... lorsque $np > 5$ et $p \leq 0.5$ soit (c'était un des termes qui devait tendre vers zéro quand nous avons fait la démonstration):

$$\frac{q}{\sqrt{npq}} \cong 0.3 \quad (7.1121)$$

Par exemple dans les deux figures ci-dessous où nous avons représenté les lois binomiales approchées par les lois Normales associées, nous avons à gauche $n = 60, p = 1/6, np = 10$ et à droite $n = 40, p = 0.05, np = 2$:

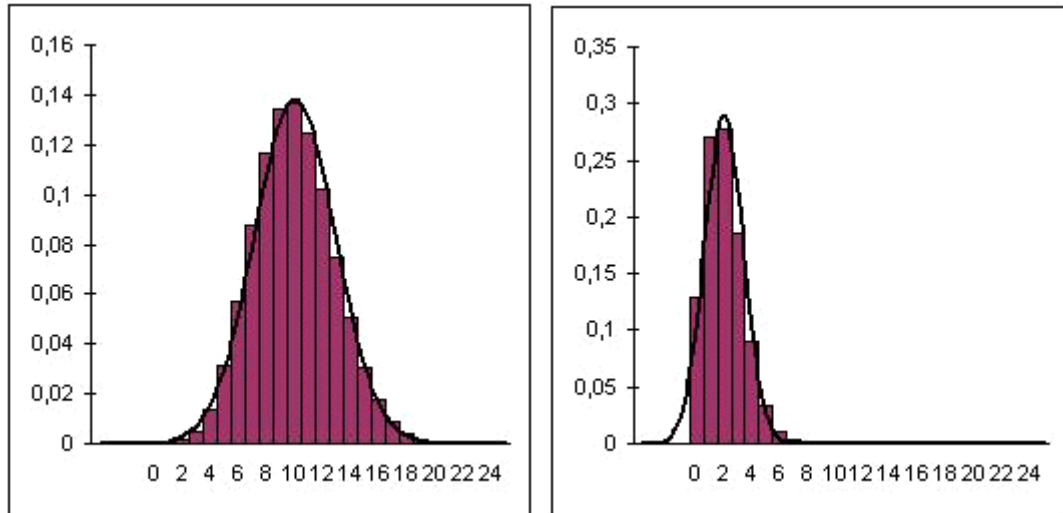


Figure: 7.59 - Approche de fonctions binomiales par fonctions Normales associées

Rappelons enfin, que nous avons démontré que la somme des carrés de n variables aléatoires normales centrées réduites linéairement indépendantes suit une loi du Khi-deux à n degrés de liberté noté $\chi^2(n)$.

Considérons maintenant une variable aléatoire X suivant une fonction de distribution théorique (continue ou discrète) P et tirons un échantillon de taille n dans la population correspondant à cette loi P .

Les n observations seront réparties suivant k modalités (classes de valeurs) $C_1, C_2, \dots, C_k$, dont les probabilités $p_1, p_2, \dots, p_k$ sont déterminées par la fonction de distribution P (se référer à l'exemple de la droite de Henry).

Pour chaque modalité C_i , l'effectif empirique est lui une variable aléatoire k_i de loi binomiale:

$$\beta(n, p_i) \quad (7.1122)$$

Cet effectif k_i correspond en effet au nombre de succès "résultat égal à la modalité C_i " de probabilité p_i , obtenus au cours des n tirages d'un lot expérimental (et non dans la population de la loi théorique comme avant).

Nous avons démontré lors de l'étude de la loi binomiale que son espérance:

$$E(k_i) = np_i \quad (7.1123)$$

représente l'effectif théorique de la modalité C_i et sa variance est:

$$V(k_i) = np_i q_i = np_i (1 - p_i) \cong np_i \quad (7.1124)$$

car p_i est relativement petite, ce qui donne $1 - p_i$ qui est assez proche de 1. Son écart-type est donc:

$$\sigma(k_i) = \sqrt{np_i} \quad (7.1125)$$

Dans ces conditions, pourvu que la modalité C_i ait un effectif théorique np_i au moins égal à 5, l'écart réduit:

$$E_i = \frac{k_i - np_i}{\sqrt{np_i}} = \frac{k_i - \mu_i}{\sigma_i} \quad (7.1126)$$

entre effectif empirique et effectif théorique peut être approximativement considéré comme une variable normale centrée réduite comme nous l'avons vu plus haut.

Nous définissons alors la variable:

$$D = \sum_{i=1}^N E_i^2 = \sum_{i=1}^N \frac{(k_i - np_i)^2}{np_i} \quad (7.1127)$$

où k_i est souvent nommée "**fréquence expérimentale**" et np_i "**fréquence théorique**".

Signalons que cette variable est aussi parfois (un peu malheureusement) notée:

$$D = \sum_{i=1}^N \frac{(n_i - n'_i)^2}{n'_i} \quad (7.1128)$$

ou le plus souvent:

$$D = \sum_{i=1}^N \frac{(O_i - E_i)^2}{E_i} \quad (7.1129)$$

Cette variable D , somme des carrés des variables E_i , nous donne une mesure de ce que nous pourrions appeler une "distance" ou "différence" ou "écart" entre distribution empirique et distribution théorique. Notons bien cependant qu'il ne s'agit pas d'une distance au sens mathématique habituel (topologique).

Rappelons que D peut donc aussi s'écrire:

$$D = \sum_{i=1}^N \frac{(k_i - np_i)^2}{np_i} = \sum_{i=1}^N \frac{(k_i - \mu_i)^2}{\sigma_i^2} \quad (7.1130)$$

D est donc la somme des carrés de N variables aléatoires normales centrées réduites liées par la seule relation linéaire:

$$k_1 + k_2 + \dots + k_N = n \quad (7.1131)$$

où n est la taille de l'échantillon. Donc D suit une loi Khi-deux mais à $N-1$ degrés de liberté, donc un

degré de moins à cause de l'unique relation linéaire qui les lie! Effectivement, rappelons que le degré de liberté indique le nombre de variables indépendantes dans la somme et non pas juste le nombre de termes sommés.

Donc:

$$\chi^2(N-1) = \sum_{i=1}^N \frac{(k_i - np_i)^2}{np_i} \quad (7.1132)$$

Nous appelons ce test un "**test non-paramétrique du Khi-deux**" ou "**test du Khi-deux de Pearson**" ou encore "**test d'ajustement du Khi-deux**" ou encore "**test de Karl Pearson**".

Ensuite, l'habitude est de déterminer la valeur de la loi du Khi-deux à $N-1$ degrés de liberté ayant 5% de probabilité d'être dépassée. Ainsi, dans l'hypothèse où le phénomène étudié suit la loi théorique P , il y a donc 95% de probabilité cumulée que la variable D prenne une valeur inférieure à celle donnée par la loi du Khi-deux.

Si la valeur de la loi du Khi-deux obtenue à partir de l'échantillon prélevé est inférieure à celle correspondant aux 95% de probabilité cumulée, nous acceptons l'hypothèse selon laquelle le phénomène suit la loi P .

Remarques:

R1. Le fait que l'hypothèse de la loi P soit acceptée ne signifie pas pour autant que cette hypothèse soit vraie, mais simplement que les informations données par l'échantillon ne permettent pas de la rejeter. De même, le fait que l'hypothèse de la loi P soit rejetée ne signifie pas nécessairement que cette hypothèse soit fausse mais que les informations données par l'échantillon conduisent plutôt à conclure à l'inadéquation d'une telle loi.

R2. Pour que la variable D suive une loi du Khi-deux, il est nécessaire que les effectifs théoriques np_i des différentes modalités C_i soient au moins égaux à 5, que l'échantillon soit tiré au hasard (pas d'autocorrélation) et qu'aucune des probabilités p_i ne soit trop proche de zéro.

Ce test d'ajustement souffre cependant d'un gros défaut: il nécessite de regrouper les mesures dans des classes C_i et dans la pratique il n'existe pas de théorème absolu (du moins à ma connaissance) pour choisir le nombre de classes (et in extenso leur largeur). C'est cette raison qui fait que le test d'ajustement (conformité) du Khi-deux est réservé pour des distributions discrètes où le problème du choix des classes en se pose pas.

Il nous faudra cependant créer des tests d'ajustement qui ne nécessitent pas l'utilisation de classes et nous allons voir de suite les outils ad hoc pour cela (test de Kolmogorov-Smirnov ou Anderson-Darling pour ne citer qu'eux).

Exemple:

Supposons que les naissances à un hôpital, pour une certaine période de temps, se répartissent comme suit:

Jour	L	M	M	J	V	S	D	Total
Observations	120	130	125	128	80	70	75	728

Tableau: 7.23 - Mesures pour l'exemple du test du Khi-deux

Nous remarquons qu'il y a eu au total 728 naissances. Nous nous posons alors la question suivant: Combien devrait il y avoir de naissances, en théorie, à chaque jour s'il n'y a pas de différence entre les jours? Ceci représente l'hypothèse nulle. En fait l'hypothèse nulle indique que les différences entre les fréquences observées et les fréquences théoriques sont relativement petites. Nous prenons donc pour acquis que si aucune différence n'existe il devrait y avoir le même nombre de naissances à chaque jour. Puisqu'il y a au total 728 naissances pour les 7 jours en théorie il devrait y avoir $728/7=104$ naissances à chaque jour. Nous avons donc maintenant le tableau suivant:

Jour	L	M	M	J	V	S	D	Total
Observations	120	130	125	128	80	70	75	728
Théorique	104	104	104	104	104	104	104	728

Tableau: 7.24 - Comparaison par rapport à l'attendu

Le total des fréquences observées est égal au total des fréquences théoriques. Il s'agit donc d'examiner la différence entre les fréquences observées et les fréquences théoriques (supposées suivre une loi uniforme) en suivant la relation du Khi-deux. En d'autres termes, nous allons faire un test d'ajustement entre une fonction de distribution empirique (observée) et la fonction de distribution uniforme. Nous avons alors:

$$\begin{aligned}
 \chi^2 &= \sum_{i=1}^N \frac{(O_i - E_i)^2}{E_i} \\
 &= \frac{(120-104)^2}{104} + \frac{(130-104)^2}{104} + \frac{(125-104)^2}{104} + \frac{(128-104)^2}{104} + \frac{(80-104)^2}{104} \\
 &\quad + \frac{(70-104)^2}{104} + \frac{(75-104)^2}{104} \\
 &= \frac{(16)^2}{104} + \frac{(26)^2}{104} + \frac{(21)^2}{104} + \frac{(24)^2}{104} + \frac{(-24)^2}{104} + \frac{(-34)^2}{104} + \frac{(-29)^2}{104} \\
 &= \frac{256}{104} + \frac{676}{104} + \frac{441}{104} + \frac{576}{104} + \frac{576}{104} + \frac{1156}{104} + \frac{841}{104} \\
 &= 2.46 + 6.5 + 4.24 + 5.54 + 5.54 + 11.12 + 8.09 = 43.49
 \end{aligned}$$

Le χ^2 est donc de 43.49. Comme tel ce chiffre signifie peu de chose. Il faut interpréter ce résultat grâce à l'aide de la table des valeurs critiques du χ^2 . On comprend qu'il est très peu probable que la fréquence observée et la fréquence théorique soit identique. Nous acceptons qu'il puisse y avoir une certaine différence (nous rejetons donc l'hypothèse).

13.10. TEST D'AJUSTEMENT DE KOLMOGOROV-SMIRNOV

En statistiques, le test de Kolmogorov-Smirnov est un test d'hypothèse basé sur une distance empirique utilisé pour déterminer si un échantillon suit bien une loi donnée connue par sa fonction de répartition continue (ou pour comparer deux échantillons et vérifier s'ils sont dépendants ou non car semblables ou dissemblables).

Pour introduire ce test, nous avons choisi l'approche de Lilliefors qui permet d'éviter des calculs complexes. Par ailleurs, les logiciels qui proposent le "**test de Lilliefors**" ne proposent pas le test Kolmogorov-Smirnov puisque ce dernier n'est correct qu'asymptotiquement (ce qui est le cas de Tangra 4.14).

Imaginez donc que nous souhaiterions construire un test non paramétrique d'ajustement qui marche aussi bien pour les lois discrètes que continues et sans souffrir du même problème que le test d'ajustement du Khi-deux (regroupement en classes).

Pour construire ce test, nous allons partir de la fonction de répartition empirique déjà définie au début de ce chapitre et donnée pour rappel par

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{x_i \leq x} = \frac{1}{n} \sum_{i=1}^n \delta_{x_i \leq x} = \frac{\#(i : x_i \leq x)}{n} \in [0,1] \quad (7.1133)$$

Notons maintenant $F(x)$, la loi vraie avec laquelle nous souhaiterions comparer $\hat{F}_n(x)$ et construisons la distance:

$$D_n(x) = \hat{F}_n(x) - F(x) \quad (7.1134)$$

Le problème avec ce choix de distance c'est... quel x faut-il alors choisir pour faire un test? Eh bien pour répondre il est simple de constater qu'il serait stupide de prendre le x pour lequel cette distance est minimale, car avoir un D_n qui peut valoir zéro n'apporte pas grand chose... Dès lors, on se reporte plutôt vers le plus grand écart en valeur absolue. Ce qui nous amène à redéfinir la distance D_n ainsi:

$$D_n(x) = \sup \left| \hat{F}_n(x) - F(x) \right| = \sup \left| \frac{1}{n} \sum_{i=1}^n 1_{x_i \leq x} - F(x) \right| \quad (7.1135)$$

où $D_n(x)$ est appelée "**distribution empirique de Kolmogorov-Smirnov**" (bon évidemment il faudrait prouver rigoureusement qu'il s'agit bien d'une distribution... mais pour l'instant c'est trop complexe au niveau du contenu du présent site, cependant cela peut se vérifier en faisant des simulations numériques). Avant d'aller plus loin relativement à la théorie, regardons un exemple pratique.

Supposons que nous ayons mesuré les cinq valeurs suivantes:

$$-1.2, 0.2, -0.6, 0.8, -1.0 \quad (7.1136)$$

soient ordonnées:

$$x_1 = -1.2, x_2 = -1.0, x_3 = -0.6, x_4 = 0.2, x_5 = 0.8 \quad (7.1137)$$

Nous voulons tester l'hypothèse nulle suivante:

$$H_0 : \hat{F}(x) = \Phi(x) \quad (7.1138)$$

où $\Phi(x)$ représente la fonction de répartition de la loi Normale centrée réduite.

La fonction de distribution empirique sera donnée par:

$$\hat{F}(x) = \begin{cases} 0 = \frac{0}{n} = \frac{0}{5} & \text{pour } x < -1.2 \\ 0.2 = \frac{1}{n} = \frac{1}{5} & \text{pour } -1.2 < x < -1 \\ 0.4 = \frac{2}{n} = \frac{2}{5} & \text{pour } -1.0 < x < -0.6 \\ 0.6 = \frac{3}{n} = \frac{3}{5} & \text{pour } -0.6 < x < +0.2 \\ 0.8 = \frac{4}{n} = \frac{4}{5} & \text{pour } +0.2 < x < +0.8 \\ 1 = \frac{5}{n} = \frac{5}{5} & \text{pour } x > 0.8 \end{cases} \quad (7.1139)$$

Ensuite, nous construisons traditionnellement le tableau suivant:

x	$\hat{F}(x)$	$\Phi(x)$	$ \hat{F}(x) - \Phi(x) $
-1.2^-	0	0.115	0.115
-1.2	0.2	0.115	0.085
-1.0^-	0.2	0.159	0.041
-1.0	0.4	0.159	0.241
-0.6^-	0.4	0.274	0.126
-0.6	0.6	0.274	0.326
$+0.2^-$	0.6	0.580	0.020
$+0.2$	0.8	0.580	0.220
$+0.8^-$	0.8	0.788	0.012
$+0.8$	1	0.788	0.212

Souvent associé au graphique comparant les fonctions de répartition empirique et théorique:

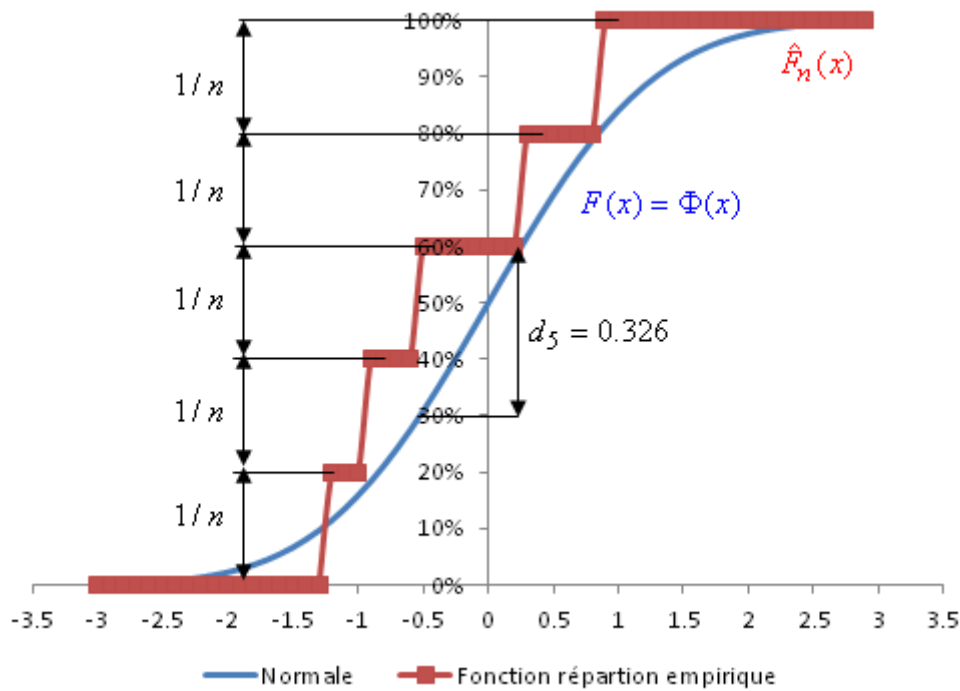


Figure: 7.60 - Représentation de l'approche du test d'ajustement de Kolmogorov-Smirnov

Nous voyons alors que l'écart maximal observé est 0.326. Nous la noterons pour la suite:

$$d_5 = 0.326 \quad (7.1140)$$

que certains logiciels comme Minitab notent par l'abréviation *KS*.

Le lecteur aura remarqué que le plus grand écart au-dessus de la courbe est mesuré par:

$$D_n^+ = \sup_{i=1}^n \left\{ \frac{i}{n} - F(x_i) \right\} = \sup_{i=1}^n \left\{ \frac{i}{n} - y_i \right\} \quad (7.1141)$$

Le plus grand écart au-dessous de la courbe est mesuré par:

$$D_n^- = \sup_{i=1}^n \left\{ F(x_i) - \frac{i-1}{n} \right\} = \sup_{i=1}^n \left\{ y_i - \frac{i-1}{n} \right\} \quad (7.1142)$$

Le plus grand écart est alors:

$$D_n = \sup_{i=1}^n \left(D_n^+, D_n^- \right) = \sup_{i=1}^n \left\{ \frac{i}{n} - y_i, y_i - \frac{i-1}{n} \right\} \quad (7.1143)$$

Mais que faire de cette valeur? À quoi la comparer? Eh bien l'idée est relativement simple et consiste à générer n valeurs (donc cinq dans le cas présent) issues de la loi de distribution $F(x)$ de l'hypothèse nulle et de les comparer à elles-mêmes. Autrement dit, il s'agit de faire une simulation de Monte-Carlo (cf. chapitre de Méthodes Numériques)

Ainsi, dans le cas présent, nous générons 5 valeurs de $N(0,1)$, ce qui nous donne par exemple la version anglaise de Microsoft Excel 11.8346 (je préfère parfois donner en anglais sinon le nom des fonction est trop long):

$$= \text{NORM.S.INV}(\text{RANDBETWEEN}(0;1000000)/1000000)$$

Nous obtenons ainsi 5 valeurs de Z (notation habituelle de la variable aléatoire d'une loi Normale centrée réduite) qui ordonnées seront par exemple:

-1.427, 0.082, 0.162, 0.294, 1.292

et nous refaisons le même tableau qu'avant:

Z	$\hat{F}(Z)$	$\Phi(Z)$	$ \hat{F}(Z) - \Phi(Z) $
-1.427 ⁻	0	0.077	0.077
-1.427	0.2	0.077	0.123
0.082 ⁻	0.2	0.533	0.333
0.082	0.4	0.533	0.133
0.162 ⁻	0.4	0.564	0.164
-0.162	0.6	0.564	0.036
+0.294 ⁻	0.6	0.616	0.016
+0.294	0.8	0.616	0.184
+1.292 ⁻	0.8	0.902	0.102
+1.292	1	0.902	0.098

Tableau: 7.25 - Tableau du test de Kolmogorov-Smirnov

Et nous avons donc l'écart maximal observé qui est de 0.333. Soit avec la version française de Microsoft Excel 14.0.6123:

	A	B	C	D	E
1	i/n	Valeurs aléatoires de Z	Valeurs de Z triées	Normale $\Phi(x)$	Delta
2	0	1.292	-1.426	0.077	0.077
3	0.2	1.292	-1.426	0.077	0.123
4	0.2	0.082	0.082	0.533	0.333
5	0.4	0.082	0.082	0.533	0.133
6	0.4	-1.427	0.162	0.564	0.164
7	0.6	-1.427	0.162	0.564	0.036
8	0.6	0.162	0.294	0.616	0.016
9	0.8	0.162	0.294	0.616	0.184
10	0.8	0.294	1.292	0.902	0.102
11	1	0.294	1.292	0.902	0.098
12				Max	0.332677

Figure: 7.61 - Calcul dans Microsoft Excel 14.0.6123

avec les formules explicites (malheureusement trop longues dans la version française du logiciel):

	A	B	C	D	E
1	//a	Valeurs aléatoires de z^2	Valeurs de z^2 triées	Normale $\Phi(z)$	Diffs
2	0	=LOI.NORMALE.STANDARD.INVERSE.N(ALEA.ENTRE.BORNES(0;1000000)/1000000)		=LOI.NORMALE.STANDARD.N(C2;1)	=ABS(D2-A2)
3	0.2	=B2		=LOI.NORMALE.STANDARD.N(C3;1)	=ABS(D3-A3)
4	0.2	=LOI.NORMALE.STANDARD.INVERSE.N(ALEA.ENTRE.BORNES(0;1000000)/1000000)		=LOI.NORMALE.STANDARD.N(C4;1)	=ABS(D4-A4)
5	0.4	=B4		=LOI.NORMALE.STANDARD.N(C5;1)	=ABS(D5-A5)
6	0.4	=LOI.NORMALE.STANDARD.INVERSE.N(ALEA.ENTRE.BORNES(0;1000000)/1000000)		=LOI.NORMALE.STANDARD.N(C6;1)	=ABS(D6-A6)
7	0.6	=B6		=LOI.NORMALE.STANDARD.N(C7;1)	=ABS(D7-A7)
8	0.6	=LOI.NORMALE.STANDARD.INVERSE.N(ALEA.ENTRE.BORNES(0;1000000)/1000000)		=LOI.NORMALE.STANDARD.N(C8;1)	=ABS(D8-A8)
9	0.8	=B8		=LOI.NORMALE.STANDARD.N(C9;1)	=ABS(D9-A9)
10	0.8	=LOI.NORMALE.STANDARD.INVERSE.N(ALEA.ENTRE.BORNES(0;1000000)/1000000)		=LOI.NORMALE.STANDARD.N(C10;1)	=ABS(D10-A10)
11	1	=B10		=LOI.NORMALE.STANDARD.N(C11;1)	=ABS(D11-A11)
12			Max		=MAX(E2:E11)

Figure: 7.62 - Fonctions explicites dans Microsoft Excel 14.0.6123

avec la petite routine VBA correspondante vite fait mal faite qui va prendre le nombre d'itérations voulues dans la cellule K1 et va mettre la distribution empirique de Kolmogorov-Smirnov dans la colonne G de la feuille active:

```
Sub simulNormal()

Application.ScreenUpdating = False
For i = 1 To Cells(1, 11)
    Calculate
    Range("B2:B11").Select
    Selection.Copy
    Range("C2").Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
    ActiveWorkbook.Worksheets("Normale").Sort.SortFields.Clear
    ActiveWorkbook.Worksheets("Normale").Sort.SortFields.Add Key:=Range("C2"), _
        SortOn:=xlSortOnValues, Order:=xlAscending, DataOption:=xlSortNormal
    With ActiveWorkbook.Worksheets("Normale").Sort
        .SetRange Range("C2:C11")
        .Header = xlGuess
        .MatchCase = False
        .Orientation = xlTopToBottom
        .SortMethod = xlPinYin
        .Apply
    End With
    Range("A1").Select
    Cells(1, 7).Value = Cells(12, 5)
Next i
Application.ScreenUpdating = True
End Sub
```

Figure: 7.63 - Code VBA Microsoft Excel 14.0.6123 pour la simulation de Monte-Carlo

Nous réitérons donc la procédure un bon millier de fois et nous obtenons la fonction de répartition (obtenue simplement en faisant un graphique de type nuage de points dans Microsoft Excel 14.0.6123 de 2'000 simulations):

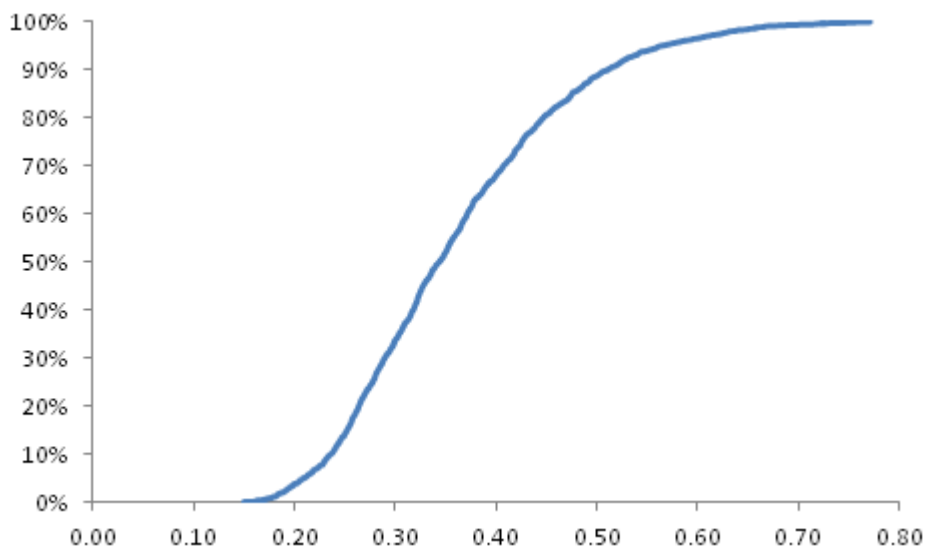


Figure: 7.64 - Fonction de répartition de Kolmogorov-Smirnov

et en appliquant un test unilatéral avec un risque $\alpha = 5\%$ nous obtenons pour le 95ème centile:

$$D_{5,\alpha} \cong 0.564 \quad (7.1144)$$

Le lecteur retrouvera la même valeur dans les tables de Kolmogorov-Smirnov disponibles dans de nombreux ouvrages. Quelques milliers de simulations suffisent donc pour retrouver les valeurs des tables!

Et maintenant, nous comparons:

$$(D_{5,\alpha} \cong 0.564) > (d_0 \cong 0.326) \quad (7.1145)$$

et donc nous ne rejetons pas l'hypothèse nulle.

Cependant, ... il faut tout de même se méfier avec seulement cinq valeurs, il est tout à fait probable que l'hypothèse nulle ne soit pas rejetée pour d'autres lois de répartition que la loi Normale.

Ainsi, comme le lecteur l'aura remarqué, pour chaque hypothèse nulle associée à une loi donnée, il faut tabuler la distribution empirique de Kolmogorov-Smirnov pour différentes valeurs de n et de α en utilisant des méthodes numériques. Dans la majorité des ouvrages on ne trouve qu'une seule table à l'aide d'un théorème puissant qui montre qu'en réalité, les valeurs critiques seront les mêmes.

Remarque: Kolmogorov et Smirnov ont démontré que lorsque n tend est très grand et que la loi de l'hypothèse nulle est continue, il n'est plus nécessaire de tabuler une table de Kolmogorov-Smirnov pour chaque loi, car nous avons alors:

$$\lim_{n \rightarrow +\infty} P(\sqrt{n}D_n \leq x) = 1 - 2 \sum_{k=1}^{+\infty} (-1)^{k-1} e^{-2k^2 x^2} \quad (7.1146)$$

donc la distribution D_n est indépendante de la loi de l'hypothèse nulle. En simulant avec la méthode de Monte-Carlo, nous observons effective une convergence lorsque n dépasse la centaine. Mais dans la pratique, la grande majorité du temps, il est impensable d'avoir un tel nombre de mesures. D'où le fait que ce résultat théorique soit peu utilisé dans la pratique et justifie l'absence de démonstration.

Pour clore, signalons au lecteur qu'il trouvera la démonstration mathématique du test d'ajustement d'Anderson-Darling dans le chapitre de Méthodes Numériques. Ce dernier test étant effectivement trop empirique pour avoir sa place ici.

13.11. TEST DE NORMALITÉ DE RYAN-JOINER

Considérons une variable aléatoire X dont nous souhaiterions vérifier la normalité ou pas. Et considérons une variable aléatoire ordonnée Y générée par une loi Normale centrée réduite. Pour comparer X et Y , nous allons centrer X et ordonner ses valeurs dans l'ordre croissant.

Pour une même taille d'échantillon, si les valeurs ordonnées de X et Y pris deux à deux suivent une même loi, la régression linéaire de l'un en fonction de l'autre doit donner un coefficient assez proche de

1. En prenant la définition du coefficient de corrélation au carré, il vient alors:

$$\rho_{X,Y}^2 = \left(\frac{\cos(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \right)^2 \quad (7.1147)$$

Y est imposé comme suivant une loi normale centrée réduite. Il vient alors:

$$\rho_{X,Y}^2 = \left(\frac{\cos(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \right)^2 = \left(\frac{E(XY) - E(X)E(Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \right)^2 = \left(\frac{E(XY)}{\sqrt{\text{Var}(X)}} \right)^2 = \frac{(E(XY))^2}{\text{Var}(X)} \quad (7.1148)$$

et si nous prenons l'estimateur du coefficient de corrélation:

$$R_{X,Y}^2 = \frac{\left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right)^2}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (7.1149)$$

Mais comme nous avons centré X, il vient:

$$R_{X,Y}^2 = \frac{\left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right)^2}{\frac{1}{n} \sum_{i=1}^n x_i^2} \quad (7.1150)$$

Soit après simplification:

$$R_{X,Y}^2 = \frac{\left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right)^2}{\frac{1}{n} \sum_{i=1}^n x_i^2} = \frac{\left(\frac{\sqrt{n}}{n} \sum_{i=1}^n x_i y_i \right)^2}{\sum_{i=1}^n x_i^2} = \frac{\left(\sum_{i=1}^n \left(\frac{y_i}{\sqrt{n}} \right) x_i \right)^2}{\sum_{i=1}^n x_i^2} = \frac{\left(\sum_{i=1}^n \alpha_i x_i \right)^2}{\sum_{i=1}^n x_i^2} \quad (7.1151)$$

Il s'agit de l'approche de Ryan-Joiner (implémentée dans Minitab) du test de Shapiro-Wilk. Les résultats des deux tests sont très similaires. Les coefficients α_i peuvent être facilement obtenus à l'aide de n'importe quel tableur à notre époque en utilisant une simulation de Monte-Carlo (cf. chapitre de Méthodes Numériques). Si un lecteur le souhaite nous détaillerons comment obtenir les α_i avec Microsoft Excel pour un n donné.

Il convient de signaler que les logiciels de statistique donne la racine carrée de la dernière égalité ci-dessous comme étant le coefficient RJ de Ryan-Joiner.

Exemple:

Considérons les 10 mesures de la colonne A déjà triées dans l'ordre croissant:

	A	B	C	D
1	Mesures	Rang (i)	NormalScore exact α_i	ZScore X_i
2	0.1685	1.00	-1.5414	-1.947246
3	0.1705	2.00	-1.0008	-0.927745
4	0.1708	3.00	-0.6536	-0.77482
5	0.1713	4.00	-0.3809	-0.519945
6	0.1728	5.00	-0.1261	0.2446801
7	0.1732	6.00	0.1205	0.4485802
8	0.1735	7.00	0.3699	0.6015052
9	0.1739	8.00	0.6501	0.8054053
10	0.1740	9.00	1.0033	0.8563803
11	0.1747	10.00	1.5340	1.2132054

Figure: 7.65 - Mesures ordonnées, rangs, coefficient de RJ et Z-score d'exemple

Les formules sont les suivantes (données en anglais car plus petites pour la capture d'écran):

	A	B	C	D
1	Mesures	Rang (i)	NormalScore exact α_i	ZScore X_i
2	0.1685	1	=Coeff_MonteCarlo!K1	=(A2-AVERAGE(\$A\$2:\$A\$11))/STDEV(\$A\$2:\$A\$11)
3	0.1705	2	=Coeff_MonteCarlo!L1	=(A3-AVERAGE(\$A\$2:\$A\$11))/STDEV(\$A\$2:\$A\$11)
4	0.1708	3	=Coeff_MonteCarlo!M1	=(A4-AVERAGE(\$A\$2:\$A\$11))/STDEV(\$A\$2:\$A\$11)
5	0.1713	4	=Coeff_MonteCarlo!N1	=(A5-AVERAGE(\$A\$2:\$A\$11))/STDEV(\$A\$2:\$A\$11)
6	0.1728	5	=Coeff_MonteCarlo!O1	=(A6-AVERAGE(\$A\$2:\$A\$11))/STDEV(\$A\$2:\$A\$11)
7	0.1732	6	=Coeff_MonteCarlo!P1	=(A7-AVERAGE(\$A\$2:\$A\$11))/STDEV(\$A\$2:\$A\$11)
8	0.1735	7	=Coeff_MonteCarlo!Q1	=(A8-AVERAGE(\$A\$2:\$A\$11))/STDEV(\$A\$2:\$A\$11)
9	0.1739	8	=Coeff_MonteCarlo!R1	=(A9-AVERAGE(\$A\$2:\$A\$11))/STDEV(\$A\$2:\$A\$11)
10	0.174	9	=Coeff_MonteCarlo!S1	=(A10-AVERAGE(\$A\$2:\$A\$11))/STDEV(\$A\$2:\$A\$11)
11	0.1747	10	=Coeff_MonteCarlo!T1	=(A11-AVERAGE(\$A\$2:\$A\$11))/STDEV(\$A\$2:\$A\$11)

Figure: 7.66 - Détails de la capture d'écran précédent avec la version anglophone de Microsoft Excel 14.0.6123

Et donc nous avons dans une feuille nommée *Coeff_MonteCarlo* des simulations de Monte-Carlo pour déterminer les 10 coefficients α_i notés traditionnellement dans le cas de 10 mesures dans les tables sous la forme suivante: $\{\alpha_{1/10}, \alpha_{2/10}, \dots, \alpha_{10/10}\}$.

D'abord il faut créer 10 colonnes avec des générations de variables aléatoires normales centrées réduites sur à peu près 10'000 lignes avec la formule suivante (donnée en anglais):

=NORM.S.INV(RANDBETWEEN(1;99999999)/100000000)

	A	B	C	D	E	F	G	H	I	J
1										
2	Val1	Val2	Val3	Val4	Val5	Val6	Val7	Val8	Val9	Val10
3	-0.99129	0.68284	0.1755	-0.68052	1.78189	-0.52315	0.76374	0.67829	-0.16533	1.61827
4	0.92486	-1.74979	0.89753	1.35076	-1.74906	1.21639	-0.27863	0.69305	-0.51966	0.79935
5	-0.26234	1.7695	0.26191	-1.1254	1.0938	0.41169	1.22103	0.88903	0.56416	-0.35126
6	1.30528	-0.49696	-1.21058	-1.06516	1.61374	1.52495	-1.17486	-1.1672	-0.02356	-0.09223
7	-1.10717	0.60102	0.30108	0.87667	-0.31205	0.96475	1.01134	0.33521	0.39329	0.13311
8	-0.88378	-1.09298	-0.12637	0.41839	-1.20593	0.2952	0.56708	-1.39583	-0.09561	0.78012
9	0.6153	-0.51951	-1.14361	1.8186	-0.4606	0.5296	-0.5367	-0.03046	1.16186	1.65168
10	0.6199	-0.08476	0.90698	-1.27954	-0.50671	-0.06454	-0.52029	0.61259	-1.56543	-0.18526
11	-0.58029	-0.70509	0.62723	-0.49299	-0.07783	0.56007	-0.69865	0.95785	-0.19304	0.49944
12	0.57066	-0.08925	-0.07208	0.95499	2.53707	1.73946	-1.5678	0.07032	0.38524	0.11992
13	-0.53362	-1.24438	0.62274	-0.53446	-1.37429	0.86534	-1.02048	0.27698	0.29341	2.26824
14	-0.33038	-0.58063	-0.74337	-0.98731	0.33377	-0.81831	0.56882	-1.38338	-0.76082	0.75053
15	0.30748	0.15149	1.27588	1.04967	0.74692	0.33037	-1.03691	-0.14088	0.85939	-0.86295

Figure: 7.67 - Génération des variables aléatoires normales centrées réduites pour les coefficient de RJ

et ensuite il faut construire les ranges de toutes ces valeurs ligne par ligne tel que:

	K	L	M	N	O	P	Q	R	S	T
1	-1.53725	-1.0033	-0.65264	-0.36952	-0.11389	0.13027	0.38655	0.66225	1.00802	1.54256
2	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
3	-0.99129	-0.68052	-0.52315	-0.16533	0.1755	0.67829	0.68284	0.76374	1.61827	1.78189
4	-1.74979	-1.74906	-0.51966	-0.27863	0.69305	0.79935	0.89753	0.92486	1.21639	1.35076
5	-1.1254	-0.35126	-0.26234	0.26191	0.41169	0.56416	0.88903	1.0938	1.22103	1.7695
6	-1.21058	-1.17486	-1.1672	-1.06516	-0.49696	-0.09223	-0.02356	1.30528	1.52495	1.61374
7	-1.10717	-0.31205	0.13311	0.30108	0.33521	0.39329	0.60102	0.87667	0.96475	1.01134
8	-1.39583	-1.20593	-1.09298	-0.88378	-0.12637	-0.09561	0.2952	0.41839	0.56708	0.78012
9	-1.14361	-0.5367	-0.51951	-0.4606	-0.03046	0.5296	0.6153	1.16186	1.65168	1.8186
10	-1.56543	-1.27954	-0.52029	-0.50671	-0.18526	-0.08476	-0.06454	0.61259	0.6199	0.90698
11	-0.70509	-0.69865	-0.58029	-0.49299	-0.19304	-0.07783	0.49944	0.56007	0.62723	0.95785
12	-1.5678	-0.08925	-0.07208	0.07032	0.11992	0.38524	0.57066	0.95499	1.73946	2.53707
13	-1.37429	-1.24438	-1.02048	-0.53446	-0.53362	0.27698	0.29341	0.62274	0.86534	2.26824
14	-1.38338	-0.98731	-0.81831	-0.76082	-0.74337	-0.58063	-0.33038	0.33377	0.56882	0.75053
15	-1.03691	-0.86295	-0.14088	0.15149	0.30748	0.33037	0.74692	0.85939	1.04967	1.27588

Figure: 7.68 - Tri des simulations pour déterminer les coefficients de RJ

avec les formules suivantes (données seulement pour les 4 premiers i faute de place dans la capture d'écran):

	K	L	M	N
1	=AVERAGE(K3:K10000)	=AVERAGE(L3:L10000)	=AVERAGE(M3:M10000)	=AVERAGE(N3:N10000)
2	O1	O2	O3	O4
3	=SMALL(\$A3:\$J3;1)	=SMALL(\$A3:\$J3;2)	=SMALL(\$A3:\$J3;3)	=SMALL(\$A3:\$J3;4)
4	=SMALL(\$A4:\$J4;1)	=SMALL(\$A4:\$J4;2)	=SMALL(\$A4:\$J4;3)	=SMALL(\$A4:\$J4;4)
5	=SMALL(\$A5:\$J5;1)	=SMALL(\$A5:\$J5;2)	=SMALL(\$A5:\$J5;3)	=SMALL(\$A5:\$J5;4)
6	=SMALL(\$A6:\$J6;1)	=SMALL(\$A6:\$J6;2)	=SMALL(\$A6:\$J6;3)	=SMALL(\$A6:\$J6;4)
7	=SMALL(\$A7:\$J7;1)	=SMALL(\$A7:\$J7;2)	=SMALL(\$A7:\$J7;3)	=SMALL(\$A7:\$J7;4)
8	=SMALL(\$A8:\$J8;1)	=SMALL(\$A8:\$J8;2)	=SMALL(\$A8:\$J8;3)	=SMALL(\$A8:\$J8;4)
9	=SMALL(\$A9:\$J9;1)	=SMALL(\$A9:\$J9;2)	=SMALL(\$A9:\$J9;3)	=SMALL(\$A9:\$J9;4)
10	=SMALL(\$A10:\$J10;1)	=SMALL(\$A10:\$J10;2)	=SMALL(\$A10:\$J10;3)	=SMALL(\$A10:\$J10;4)
11	=SMALL(\$A11:\$J11;1)	=SMALL(\$A11:\$J11;2)	=SMALL(\$A11:\$J11;3)	=SMALL(\$A11:\$J11;4)
12	=SMALL(\$A12:\$J12;1)	=SMALL(\$A12:\$J12;2)	=SMALL(\$A12:\$J12;3)	=SMALL(\$A12:\$J12;4)
13	=SMALL(\$A13:\$J13;1)	=SMALL(\$A13:\$J13;2)	=SMALL(\$A13:\$J13;3)	=SMALL(\$A13:\$J13;4)
14	=SMALL(\$A14:\$J14;1)	=SMALL(\$A14:\$J14;2)	=SMALL(\$A14:\$J14;3)	=SMALL(\$A14:\$J14;4)
15	=SMALL(\$A15:\$J15;1)	=SMALL(\$A15:\$J15;2)	=SMALL(\$A15:\$J15;3)	=SMALL(\$A15:\$J15;4)

Figure: 7.69 - Détails du tri pour la détermination des coefficients de RJ
avec la version anglophone de Microsoft Excel 14.0.6123

Pour finir, il n'y a plus qu'à calculer le coefficient de corrélation entre les colonnes C et D de la première capture d'écran:

	F
1	Coefficient de Corrélation Ryan-Joiner (RJ):
2	=CORREL(C2:C11;D2:D11)

Figure: 7.70 - Calcul final du coefficient de corrélation de RJ

Ce qui donne environ 0.963 (le carré de cette valeur étant très très proche du test de Shapiro-Wilk). Ensuite, pour savoir si on peut accepter ou rejeter l'hypothèse de normalité, il faudrait refaire la procédure avec en lieu et place mesures, des valeurs générées aussi aléatoires à partir d'une loi Normale et déterminer la valeur critique d'acceptation/rejet (normalement c'est très simple à faire mais on peut détailler sur demande).

14. ROBUSTESSE

Dans le domaine des statistiques inférentielles et tests d'hypothèses, la robustesse est un concept récurrent (les banques sont astreintes au stress testing de leurs modèles de risque). Nous en avons par ailleurs déjà fait mention plus haut...

Définitions:

D1. Un test est dit "**test robuste**" s'il reste valable alors que les hypothèses d'application ne sont pas toutes réunies. Ce peut être une taille d'échantillon un peu faible ou une loi de probabilité (loi normale pour les tests paramétriques) qui n'est pas très bien vérifiée. Par exemple, l'ANOVA est robuste par rapport à l'hypothèse de normalité mais pas par rapport à celle de l'homoscédasticité

D2. Un indicateur est dit "**indicateur robuste**" s'il est peu sensible à la présence d'outliers (le coefficient de corrélation, par exemple, n'est pas très robuste).

D3. Plus généralement, un modèle est dit "**modèle robuste**" lorsqu'il permet un prolongement des résultats (dans le temps ou pour une population). La robustesse s'applique aussi bien à une régression multiple qu'à une grille de score.

Par conséquent, à moins d'être uniquement descriptives, vos études devront respecter quelques règles pour que leurs conclusions soient généralisables.

Première condition d'une bonne robustesse: les données. Intuitivement, chacun sait qu'on ne transforme pas un cas en généralité (ce qui ne relèverait pas des statistiques mais des discussions de comptoir). Une quantité suffisante de données permet de bâtir des modèles fiables et solides. À titre d'exemple, des prévisions établies à partir d'une série chronologique montrant une saisonnalité nécessitent au moins trois ou quatre ans d'historique.

La quantité ne suffit pas, il faut la qualité. Mieux vaut s'abstenir que réaliser une étude sur des informations non fiables qui peuvent conduire à des décisions coûteuses. Par ailleurs, il convient d'éliminer ou d'imputer certaines observations (voir outliers). Si ce n'est pas possible, on se tourne vers des méthodes adaptées, par exemple celles qui utilisent la médiane plutôt que la moyenne.

15. STATISTIQUES DE RANGS

Les statistiques de rangs, appelées aussi "**statistiques d'ordre**", sont définies comme l'ensemble des techniques de calculs statistiques ou d'inférence statistiques qui ont pour objectif principal de se débarrasser de la connaissance d'une distribution paramétrée et en utilisant pour cela uniquement les rangs (ordonnés) des caractéristiques mesurées. Il s'agit d'un outil très puissant et très utilisé dans la pratique pour faire de la statistique non-paramétrée!

15.1. TESTS DE RANG (NON PARAMÉTRIQUES)

Comme nous l'avons déjà mentionné plus haut, nous parlons de tests paramétriques lorsque nous stipulons que les données sont issues d'une distribution paramétrée. Dans ce cas, les caractéristiques des données peuvent être résumées à l'aide de paramètres estimés sur l'échantillon, la procédure de test subséquente ne porte alors que sur ces paramètres.

Rappelons aux lecteurs les tests le petit nombre de tests (du moins ceux qui ont des noms particuliers) que nous avons démontrés jusqu'à maintenant (nous espérons ne pas avoir commis d'erreurs trop graves dans le listing ci-dessous):

(1) Test-T de Student

Type:

Test d'hypothèse paramétrique de type intervalle de confiance souvent utilisé en bilatéral

Concerne:

La moyenne lorsque l'écart-type théorique est inconnu

Contrainte(s):

Distribution Normale des données.

(2) Test-p de l'intervalle de confiance de proportions

Type:

Test d'hypothèse paramétrique de type intervalle de confiance souvent utilisé en bilatéral

Concerne:

La proportion de bons ou mauvais éléments dans une population

Contrainte(s):

Distribution Binomiale (et asymptotiquement) Normale des données ($n \cdot p \geq 5$).

(3) Test-p de l'égalité de deux proportions

Type:

Test d'hypothèse paramétrique de type intervalle de confiance souvent utilisé en unilatéral

Concerne:

L'égalité de deux proportions

Contrainte(s):

Distribution Binomiale et asymptotiquement Normale des données ($n \cdot p \geq 5$)

(4) Test binomial exact (égalité de deux proportions)

Type:

Test d'hypothèse paramétrique de type intervalle de confiance souvent utilisé en bilatéral

Concerne:

L'égalité de deux proportions.

Contrainte(s):

Distribution Binomiale (petit échantillon d'une grande population)

(5) Test des signes (de la médiane) de deux échantillons appariés

Type:

Test d'hypothèse non-paramétrique de type intervalle de confiance souvent utilisé en bilatéral

Concerne:

L'égalité des signes (implicitement des différences) de données appariées

Contrainte(s):

Distribution Binomiale (petit échantillon d'une grande population) mais valeurs sous-jacentes continues.

(6) Test-T de Student de deux moyennes d'échantillons appariés

Type:

Test d'hypothèse paramétrique de type intervalle de confiance souvent utilisé unilatéral.

Concerne:

La différence de deux moyennes de deux échantillons identiques

Contrainte(s):

Distribution Normale des données

(7) Test-Z

Type:

Test d'hypothèse paramétrique de type intervalle de confiance plus utilisé en bilatéral qu'en unilatéral.

Concerne:

La moyenne lorsque l'écart-type théorique est connu

Contrainte(s):

Distribution Normale des données

(8) Test-Z de la moyenne à deux échantillons

Type:

Test d'hypothèse paramétrique de type intervalle de confiance souvent utilisé en bilatéral

Concerne:

La différence de deux moyennes lorsque les écarts-types théoriques sont connus

Contrainte(s):

Distribution Normale des données

(9) Test du Khi-deux

Type:

Test d'hypothèse paramétrique de type intervalle de confiance souvent utilisé en bilatéral

Concerne:

La variance théorique

Contrainte(s):

Distribution Normale des données

(10) Test-F de Fisher

Type:

Test d'hypothèse paramétrique de type intervalle de confiance souvent utilisé en bilatéral

Concerne:

La comparaison de deux variances théoriques

Contrainte(s):

Distribution Normale des données

(11) Test-T homoscédastique

Type:

Test d'hypothèse paramétrique de type intervalle de confiance souvent utilisé en unilatéral

Concerne:

L'égalité de deux moyennes

Contrainte(s):

Distribution Normale des données et égalité des variances expérimentales

(12) Test-T hétéroscédastique*Type:*

Test d'hypothèse paramétrique de type intervalle de confiance souvent utilisé en unilatéral

Concerne:

L'égalité de deux moyennes

Contrainte(s):

Distribution Normale des données et non-égalité des variances expérimentales (cas généralisé du Test-T homoscdastique)

(13) Test de l'ANOVA à un facteur contrôlé*Type:*

Test d'hypothèse paramétrique de type intervalle de confiance utilisé uniquement en unilatéral

Concerne:

L'égalité des moyennes des échantillons (supposés implicitement appariés)

Contrainte(s):

Distribution Normale des données avec variances théoriques identiques et variances expérimentales connues et indépendance des échantillons. Les résidus doivent in extenso aussi être normalement distribués

(14) Test de l'ANOVA à deux facteurs contrôlés avec ou sans répétition*Type:*

Test d'hypothèse paramétrique de type intervalle de confiance utilisé uniquement en unilatéral

Concerne:

L'égalité des moyennes des échantillons fonction d'un paramètre variable contrôlable (ajustable).

Contrainte(s):

Distribution Normale des données avec variances théoriques identiques et variances expérimentales connues et indépendance des échantillons. Les résidus doivent in extenso aussi être normalement distribués.

(15) Test d'ajustement (dit aussi "test d'adéquation de Pearson") du Khi-deux*Type:*

Test d'ajustement paramétrique utilisé uniquement en unilatéral

Concerne:

Adéquation de valeurs expérimentales à une loi théorique

Contrainte(s):

Avoir suffisamment de classes d'intervalles et de données

Remarque: Appelé "Test de normalité" si comparé à une loi Normale.

(16) Test d'indépendance du Khi-deux*Type:*

Test d'ajustement (étudié dans le chapitre de Méthodes Numériques) paramétrique utilisé uniquement en unilatéral

Concerne:

Vérifier la dépendance ou l'indépendance (différence) de données provenant d'une table de contingence. Vérifie donc si les moyennes sont différentes ou pas entre groupes en se basant sur la contingence

Contrainte(s):

Avoir suffisamment de classes d'intervalles et de données

(17) Test de la médiane*Type:*

Test d'hypothèse non paramétrique étudié dans le chapitre de Méthodes Numériques de type intervalle de confiance toujours utilisé en bilatéral

Concerne:

La médiane

Contrainte(s):

Un nombre d'échantillons suffisant pour faire un bootstrap.

(18) Test de Mood (de la médiane)*Type:*

Test d'hypothèse non paramétrique basé sur la classification de données supérieures et inférieurs à la médiane. Les données sont aussi parfois angées dans une table de contingence après quoi on calcule un test d'indépendance du Khi-deux et le test de la médiane.

Concerne:

Vérifier la dépendance ou l'indépendance (différence) de données provenant d'une table de contingence (vérifie donc si les médianes sont différentes ou pas entre groupes en se basant sur la contingence)

Contrainte(s):

Avoir suffisamment de classes d'intervalles et de données (puisque basé sur la loi hypergéométrique et le test

d'indépendance du Khi-deux)

(19) Test de Poisson à un et deux échantillons

Type:

Tests d'hypothèses paramétrique tantôt en unilatéral ou bilatéral basé sur les événements rares (dixit la moyenne de la loi de Poisson)

Concerne:

Déterminer un intervalle de confiance pour l'occurrence d'événements rares sur une période donnée afin d'identifier une anomalie ou une différence significative par rapport à des objectifs ou des normes.

Contrainte(s):

Les événements suivent une loi de Poisson mais sont approximatés dans le cas à deux échantillons par une loi Normale...

(20) Test C de Cochran

Type:

Test d'hypothèse paramétrique utilisé en unilatéral

Concerne:

La détection d'une variance aberrante parmi un ensemble de variances avant d'effectuer une ANOVA.

Contrainte(s):

Suppose les données normalement distribuées.

(21) Test de Ryan-Joiner (version simplifiée de Shapiro-Wilk)

Type:

Test d'hypothèse non-paramétrique utilisé uniquement en unilatéral sur le coefficient de corrélation.

Concerne:

Déterminer si des mesures une fois centrées et réduites peuvent être considérées comme normalement distribuées.

Contrainte(s):

Les mesures suivent une loi Normale (in extenso centrée réduite après normalisation).

(22) Test du Kappa de Cohen

Type:

Test d'hypothèse paramétrique étudié dans le chapitre de Méthodes Numériques et utilisé uniquement en unilatéral

Concerne:

Tableau de contingence dichotomique (binaire) dont on souhaite vérifier si les agréments entre experts correspondent

Contrainte(s):

Les valeurs du tableau de contingence dichotomique suivent une loi Normale (in extenso centrée réduite après normalisation) et très très approximativement une loi du Khi-deux à un degré de liberté.

(23) Test de McNemar (données appariées)

Type:

Test d'hypothèse paramétrique étudié dans le chapitre de Méthodes Numériques utilisé uniquement en unilatéral mais avec intervalle de confiance de la différence en bilatéral

Concerne:

Tableau de contingence dichotomique (binaire) dont on souhaite vérifier si les agréments entre experts correspondent avant et après (car les données ne sont pas indépendantes)

Contrainte(s):

Les valeurs du tableau de contingence dichotomique suivent une loi Normale (in extenso centrée réduite après normalisation) et très très approximativement une loi du Khi-deux à un degré de liberté.

(24) Test exact de Fisher

Type:

Test d'ajustement (étudié dans le chapitre de Méthodes Numériques) paramétrique utilisé principalement en bilatéral

Concerne:

Vérifier si la configuration observée dans un tableau de contingence est une situation extrême par rapport aux situations possibles.

Contrainte(s):

Aucune en particulière

Tableau: 7.26 -Tests courants et démontrés jusqu'à maintenant sur le présent site Internet

Le lecteur intéressé pourra se référer à l'excellent ouvrage en langue anglaise de Gopal K. Kanji qui contient une présentation sommaire avec exemples des 100 tests statistiques paramétriques et non

paramétriques les plus utilisés dans le monde.

Les tests non paramétriques (comme les deux tests du Khi-deux déjà vus) ne font eux aucune hypothèse sur la distribution sous-jacente des données. L'étape préalable qui consistait uniquement à estimer les paramètres des distributions avant de procéder au test d'hypothèse proprement dit n'est plus nécessaire.

Lorsque les données sont quantitatives, les tests non paramétriques transforment les valeurs en rangs. L'appellation "**tests de rangs**" est alors souvent rencontrée. Lorsque les données sont qualitatives, seuls les tests non paramétriques sont utilisables.

15.1.1. L-STATISTIQUES

Avant de s'attaquer aux tests non paramétriques, donnons quelques définitions que lecteur risquerait de trouver dans la littérature hyperspécialisée et dont nous avons évité l'utilisation (du moins jusqu'à maintenant...).

La médiane, la moyenne et l'étendue suggèrent l'utilisation de combinaisons linéaires de composantes du vecteur des statistiques d'ordre.

Ainsi, notons $X_{(1)} \geq \dots \geq X_{(n)}$ une statistique d'ordre (donc les valeurs ordonnées dans l'ordre décroissant et numérotées par leur rang). Nous définissons alors la "**L-statistique**" par:

$$T_n = \sum_{i=1}^n a_{i,n} X_{(i)} \quad (7.1152)$$

et donc le premier "**L-estimateur**" le plus connu est la moyenne arithmétique pour laquelle:

$$a_{i,n} = \frac{1}{n} \quad (7.1153)$$

Le deuxième L-estimateur le plus connu est la médiane pour laquelle nous avons lorsque n est impair:

$$a_{i,n} = \begin{cases} 1 & i = \frac{n+1}{2} \\ 0 & i \neq \frac{n+1}{2} \end{cases} \quad (7.1154)$$

et lorsque n est pair:

$$a_{i,n} = \begin{cases} \frac{1}{2} & i = \frac{n}{2}, \frac{n}{2} + 1 \\ 0 & \text{sinon} \end{cases} \quad (7.1155)$$

Enfin, le troisième L-estimateur le plus connu est l'étendue pour lequel nous avons:

$$a_{i,n} = \begin{cases} +1 & i = 1 \\ -1 & i = n \\ 0 & \text{sinon} \end{cases} \quad (7.1156)$$

15.1.2. TEST DE LA SOMME DES RANGS DE WILCOXON

L'idée du "**test de la somme des rangs de Wilcoxon**" est la suivante: si nous rassemblons deux échantillons de mesures, et que nous rangeons les valeurs dans l'ordre, l'alternance des X_i (de taille n_x) et des Y_i (de taille n_y) devrait être assez régulière si les deux échantillons de loi de répartition F et respectivement G suivent la même loi de probabilité. Il s'agit donc d'un "test d'ajustement".

Il ne s'agit donc pas comme le test d'ajustement du Khi-deux de comparer des mesures à une loi théorique, mais à d'autres mesures.

Remarque: Le test de la somme des rangs de Wilcoxon est donc un test non paramétrique parce que nous n'avons pas besoin d'un quelconque indicateur de dispersion ou de position des variables aléatoires étudiées pour l'utiliser. De plus, c'est un test dit robuste dans le sens qu'il ne suppose pas la normalité des données.

Prenons un exemple avant de nous attaquer à l'aspect théorique. Voici deux échantillons de taille 10 ($n_x = n_y$) de variables quantitatives:

5.7, 3.2, 8.4, 4.1, 6.9, 5.3, 1.7, 3.2, 2.5, 7.4
8.1, 5.5, 3.4, 7.9, 4.6, 1.6, 8.5, 7.1, 8.7, 5.7 (7.1157)

Remarque: Le test de la somme des rangs de Wilcoxon peut tout à fait être utilisé pour des variables ordinales (donc catégorielles mais à condition qu'elles soient en un nombre acceptable). Typiquement, le test de la somme des rangs de Wilcoxon est aussi utilisé pour analyser la réponse à des sondages en entreprise utilisant des échelles de Likert sur 7 points.

Voici les statistiques d'ordre de l'échantillon de taille 20 ($N = n_x + n_y$) regroupé et ordonné (les 10 valeurs X_i du premier échantillon sont soulignées):

1.6, 1.7, 2.5, 3.2, 3.2, 3.4, 4.1, 4.6, 5.3, 5.5,
5.7, 5.7, 6.9, 7.1, 7.4, 7.9, 8.1, 8.4, 8.5, 8.7 (7.1158)

Les valeurs du premier échantillon X (dites "**valeurs de traitement**") ont tendance à être plus petites que celles du second Y (dites "**valeurs de contrôle**") ce que nous représentons souvent sous la forme graphique suivante (en trichant un peu avec Microsoft Excel 11.8346):

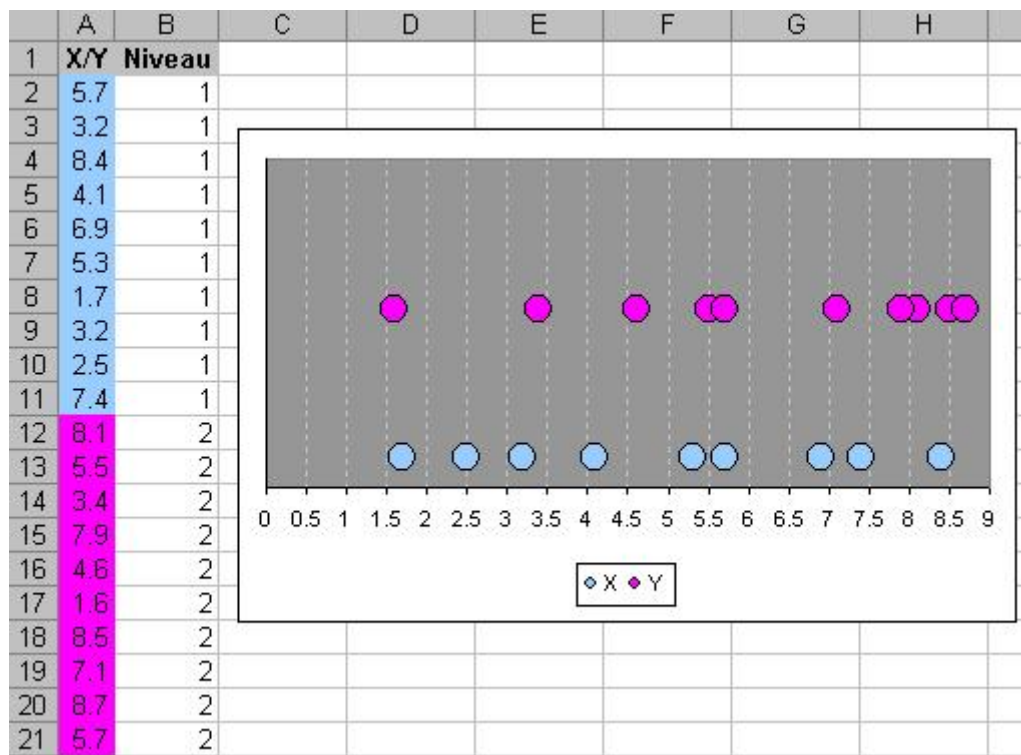


Figure: 7.71 - Comparaison des valeurs de deux échantillons dans Microsoft Excel 11.8346

L'idée est alors de chercher à savoir si cette tendance est significative. C'est-à-dire de savoir si nous avons une réelle différence du type $F < G$ entre leurs lois de répartition respectives:

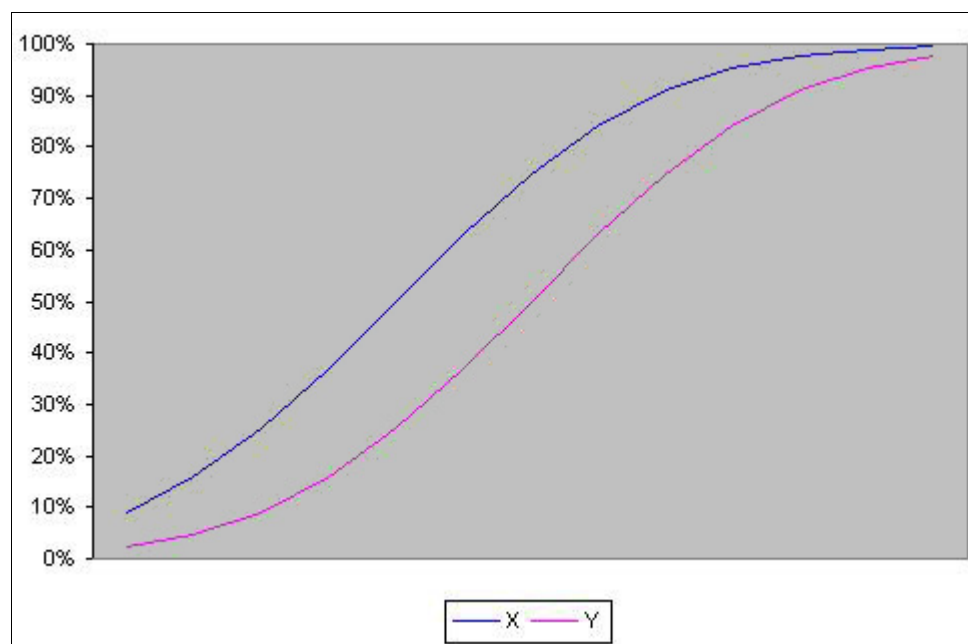


Figure: 7.72 - Exemple générique de la comparaison de deux fonctions de répartition

ou si elles peuvent être considérées comme identiques. Pour cela, il faut bien étudier la notion de "rang":

Étant donné un n -échantillon aléatoire $X_1, \dots, X_n$ d'une loi statistique continue quelconque, nous notons R_i le rang des X_i ordonnés dans une population d'échantillons. Le rang i est donc un nombre entier non nul et strictement positif compris entre 1 et N (somme des tailles de tous les échantillons).

 Exemple:

Dans:

$$1.6, \underline{1.7}, \underline{2.5}, \underline{3.2}, \underline{3.2}, 3.4, \underline{4.1}, 4.6, \underline{5.3}, 5.5, \\ 5.7, \underline{5.7}, \underline{6.9}, 7.1, \underline{7.4}, 7.9, 8.1, \underline{8.4}, 8.5, 8.7 \quad (7.1159)$$

Nous avons les "**statistiques d'ordre**" respectives:

$$X: R_2 = 1.7, R_3 = 2.5, R_4 = 3.2, R_5 = 3.2, R_7 = 4.1, \dots \\ Y: R_1 = 1.6, R_6 = 3.4, R_8 = 4.6, R_{10} = 5.5, R_{11} = 5.7, \dots \quad (7.1160)$$

Une fois le concept de "**rang**" défini, intéressons-nous à la somme dans le cadre de notre exemple avec les deux échantillons:

$$1.6, \underline{1.7}, \underline{2.5}, \underline{3.2}, \underline{3.2}, 3.4, \underline{4.1}, 4.6, \underline{5.3}, 5.5, \\ 5.7, \underline{5.7}, \underline{6.9}, 7.1, \underline{7.4}, 7.9, 8.1, \underline{8.4}, 8.5, 8.7 \quad (7.1161)$$

La somme des rangs, notée traditionnellement W_x (W pour Wilcoxon), du premier échantillon est alors:

$$W_x = 2 + 3 + 4 + 5 + 7 + 9 + 12 + 13 + 15 + 18 = 88 \quad (7.1162)$$

et pour le deuxième échantillon:

$$W_y = 1 + 6 + 8 + 10 + 11 + 14 + 16 + 17 + 19 + 20 = 122 \quad (7.1163)$$

Valeurs que nous appelons "**statistique de Wilcoxon**".

Nous pouvons donc déjà constater qu'il y a effectivement une différence qui semble a priori non négligeable en termes de rang sur les deux échantillons. Tout le problème reste maintenant à construire un outil mathématique rigoureux permettant d'en conclure un fait avec une certaine certitude.

Pour cela, introduisons d'abord la moyenne des rangs en utilisant le résultat démontré dans le chapitre de Suites et Séries en considérant un seul échantillon:

$$\bar{R} = \frac{1}{n} \sum_{i=1}^n R_i = \frac{1}{n} \sum_{i=1}^n i = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2} \quad (7.1164)$$

En calculant cela, nous remarquons assez vite qu'il s'agit de l'espérance de la loi discrète uniforme démontrée plus haut dans ce même chapitre pour une variable aléatoire discrète dont les valeurs sont comprises entre 1 et n , soit exactement la définition du rang! Ainsi, nous avons le rang qui aura pour caractéristique de moyenne et de variance pour toute la population:

$$\mu = \bar{R} = \frac{n+1}{2} \\ V(R) = \frac{n^2 - 1}{12} \quad (7.1165)$$

Pour ceux qui trouverais cette analogie douteuse voici sinon la démonstration de la variance en utilisant

la relation de Huyghens et la somme des carrées des entiers positifs démontrée dans le chapitre de Suites Et Séries:

$$\begin{aligned}
 V(R) &= E(R^2) - E(R)^2 = \sum_{i=1}^n p(R_i) R_i^2 - \left(\frac{n+1}{2}\right)^2 = \frac{1}{n} \sum_{i=1}^n R_i^2 - \left(\frac{n+1}{2}\right)^2 \\
 &= \frac{1}{n} \frac{n(2n+1)(n+1)}{6} - \frac{(n+1)^2}{4} = \frac{2(2n+1)(n+1) - 3(n^2 + 2n + 1)}{12} \quad (7.1166) \\
 &= \frac{2(2n^2 + 2n + n + 1) - 3n^2 - 6n - 3}{12} = \frac{4n^2 + 4n + 2n + 2 - 3n^2 - 6n - 3}{12} = \frac{n^2 - 1}{12}
 \end{aligned}$$

Mais évidemment pour un échantillon seul cela n'a aucun intérêt! Reprenons nos deux séries X_i, Y_i respectivement de tailles égales $n_x = 10, n_y = 10$ sans distinction:

$$\begin{aligned}
 &1.6, 1.7, 2.5, 3.2, 3.2, 3.4, 4.1, 4.6, 5.3, 5.5, \\
 &5.7, 5.7, 6.9, 7.1, 7.4, 7.9, 8.1, 8.4, 8.5, 8.7 \quad (7.1167)
 \end{aligned}$$

Nous avons alors les indicateurs statistiques de rangs sans distinction (il faut bien garder en tête que nous ne savons pas encore à ce niveau du développement si cela nous sera utile ou non):

$$\begin{aligned}
 W_{x,y} &= \sum_{i=1}^{n_x+n_y} R_i = \sum_{i=1}^{n_x+n_y} i = \frac{(n_x + n_y)(n_x + n_y + 1)}{2} = 210 \\
 \mu_{x,y} &= \bar{R}_{x,y} = \frac{(n_x + n_y) + 1}{2} = 10.5 \quad (7.1168) \\
 V(R_{x,y}) &= \frac{(n_x + n_y)^2 - 1}{12} = 33.25
 \end{aligned}$$

et les indicateurs statistiques des rangs mais cette fois-ci avec distinction:

$$\begin{aligned}
 &1.6, \underline{1.7}, \underline{2.5}, \underline{3.2}, \underline{3.2}, 3.4, \underline{4.1}, 4.6, \underline{5.3}, 5.5, \\
 &5.7, \underline{5.7}, \underline{6.9}, 7.1, \underline{7.4}, 7.9, 8.1, \underline{8.4}, 8.5, 8.7 \quad (7.1169)
 \end{aligned}$$

Nous avons alors les indicateurs statistiques locaux:

$$W_x = \sum R_{x,i} = 2+3+4+5+7+9+12+13+15+18 = 88$$

$$W_y = \sum R_{y,i} = 1+6+8+10+11+14+16+17+19+20 = 122$$

$$\bar{R}_x = \frac{1}{n} \sum R_{x,i} = \frac{2+3+4+5+7+9+12+13+15+18}{10} = 8.8 \quad (7.1170)$$

$$\bar{R}_y = \frac{1}{n} \sum R_{y,i} = \frac{1+6+8+10+11+14+16+17+19+20}{10} = 12.2$$

$$V(R_x) = \frac{1}{n_x} \sum (R_{x,i} - \bar{R}_x)^2 = 27.16$$

$$V(R_y) = \frac{1}{n_y} \sum (R_{y,i} - \bar{R}_y)^2 = 33.56$$

Ces calculs étant effectués, nous n'avons ceci dit rien de concrètement rigoureux pour l'instant en ce qui concerne le test de la somme des rangs de Wilcoxon dont l'objectif est pour rappel de vérifier si les deux échantillons suivent la même loi ou non (et donc ont in extenso les mêmes moments comme l'espérance, la variance, la médiane, etc.).

Pour avancer, considérons les n_x valeurs de l'échantillon X . Nous savons (cf. chapitre de Probabilités), qu'il y a alors:

$$C_{n_x}^{n_x+n_y} = \binom{n_x+n_y}{n_x} = \frac{(n_x+n_y)!}{n_x!((n_x+n_y)-n_x)!} = \frac{(n_x+n_y)!}{n_x!n_y!} \quad (7.1171)$$

nombre de rangements possibles des X_i dans la population des échantillons et que si le test de la somme des rangs de Wilcoxon se vérifie (c'est-à-dire: les lois de probabilités sont les mêmes pour les deux échantillons), les différents rangements sont équiprobables.

Par exemple, si nous prenons deux échantillons avec respectivement chacun deux mesures (2 variables aléatoires de traitement et 2 de contrôle), nous avons :

$$(n_x+n_y)! = 4! = 24 \quad (7.1172)$$

arrangements différents tous distincts:

$$\begin{array}{cccc} (1,2,3,4) & (2,1,3,4) & (3,1,2,4) & (4,1,2,3) \\ (1,2,4,3) & (2,1,4,3) & (3,1,4,2) & (4,1,3,2) \\ (1,3,2,4) & (2,3,1,4) & (3,2,1,4) & (4,2,1,3) \\ (1,3,4,2) & (2,3,4,1) & (3,2,4,1) & (4,2,3,1) \\ (1,4,2,3) & (2,4,1,3) & (3,4,1,2) & (4,3,1,2) \\ (1,4,3,2) & (2,4,3,1) & (3,4,2,1) & (4,3,2,1) \end{array} \quad (7.1173)$$

Mais ce n'est pas ce que nous voulons dans le cas présent car déjà nous souhaiterions pouvoir distinguer

les deux échantillons et aussi ne pas prendre en compte les arrangements qui consistent uniquement en une permutation des variables d'un même échantillon. Nous avons alors (cf. chapitre de Probabilités):

$$C_{n_x}^{n_x+n_y} = \binom{n_x+n_y}{n_x} = \frac{4!}{2!2!} = 6 \quad (7.1174)$$

combinaisons possibles! Effectivement avec deux échantillons comprenant deux variables de traitement (X) et deux variables de contrôle (Y), nous avons:

Rangs possibles Contrôles	Rangs possibles Traitements	W_s
1, 2	3, 4	7
1, 3	2, 4	6
1, 4	2, 3	5
2, 3	1, 4	5
2, 4	1, 3	4
3, 4	1, 2	3

Tableau: 7.27 - Représentation des rangs de 2 variables de traitement et de contrôle

Si l'hypothèse du test de la somme des rangs de Wilcoxon est juste, les 6 classements sont équiprobables. Nous en déduisons le tableau suivant:

Valeur de W_s	3	4	5	6	7
Probabilité	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
Cumulatif	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	$\frac{6}{6}$

Tableau: 7.28 - Probabilités associées au test de la somme des rangs Wilcoxon

Ce tableau étant construit, supposons que nous observions pour la somme des rangs des variables de traitement: $W_s = 7$. Le seuil d'un test unilatéral donnerait alors conformément au tableau ci-dessus:

$$P(W_s \geq 7) = \frac{1}{6} \cong 17\% \quad (7.1175)$$

ou si nous obtenions $W_s = 3$:

$$P(W_s \geq 3) = \frac{5}{6} \cong 83\% \quad (7.1176)$$

Nous rejeterions donc l'hypothèse d'une distribution identique entre les deux échantillons à tout seuil supérieur (ou respectivement inférieur) fixé à l'avance par la politique du laboratoire... en test unilatéral ou bilatéral (raison pour laquelle les logiciels donnent les valeurs unilatérales du test + bilatérales en même temps).

Deux choses très importantes qu'il faut remarquer pour la suite sont que:

- Premièrement dans la construction du tableau précédent (dont voici à nouveau une partie):

Valeur de W_s	3	4	5	6	7
Probabilité	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Tableau: 7.29 - Probabilités associées au test de de la somme des rangs Wilcoxon

Il y a une symétrie à la valeur 5, ce qui signifie que la loi de W_s est symétrique dans ce cas particulier. Mais si nous prenons un autre exemple avec deux échantillons comprenant respectivement deux variables de contrôle et trois de traitements (deux variables aléatoires):

Valeur de W_s	6	7	8	9	10	11	12
Probabilité	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

Tableau: 7.30 - Probabilités associées au test de de la somme des rangs Wilcoxon

le lecteur pourra vérifier que quel que soit le nombre d'échantillons et le nombre de variables de contrôle et de traitement, le tableau de probabilités ci-dessus est toujours symétrique (bon il y a une démonstration mathématique de ceci mais je la trouve peu élégante). Mais au fait c'est assez intuitif, comme les combinaisons $C_{n_x}^{n_x+n_y}$ sont indépendantes du fait que les rangs soient rangés dans l'ordre croissant ou décroissant, il est forcé qu'il y ait une symétrie.

- Deuxièmement les valeurs des variables mesurées ne rentrent pas en compte dans cette statistique paramétrique mais uniquement les valeurs tabulées des rangs avec leurs probabilités associées. Effectivement, comme vous avez pu le remarquer, nous n'avons pas eu besoin des valeurs explicites des variables aléatoires pour construire le tableau précédent.

Maintenant, sachant que la loi de W_s est symétrique et discrète nous souhaiterions calculer son espérance (nous ne nous intéresserons pas au calcul de la variance de cette loi car nous trouvons qu'elle n'apporte dans la pratique aucune information utile).

Le calcul de l'espérance est relativement facile. Effectivement, comme la loi est symétrique, l'espérance est alors la moyenne de la borne supérieure et inférieure des rangs. Voyons cela:

- La plus petite valeur possible de W_s est en supposant qu'elle est dans l'échantillon X (les algorithmes informatiques déterminent automatiquement dans quel échantillon mais de toute façon dans la pratique, les échantillons ont quasiment toujours la même taille):

$$W_{s,\min} = 1 + \dots + n_x = \frac{1}{2} n_x (n_x + 1) \quad (7.1177)$$

- La plus grande valeur possible est naturellement (se souvenir que $N = n_x + n_y$):

$$W_{s,\max} = \frac{1}{2} N(N+1) - \frac{1}{2} n_y (n_y + 1) \quad (7.1178)$$

L'espérance de la somme des rangs d'un des deux échantillons vaut alors:

$$\begin{aligned}
E(W_s) &= W_{s,\min} + \frac{W_{s,\max} - W_{s,\min}}{2} = W_{s,\max} - \frac{W_{s,\max} - W_{s,\min}}{2} \\
&= \frac{1}{2} n_x (n_x + 1) + \frac{\frac{1}{2} N(N+1) - \frac{1}{2} n_y (n_y + 1) - \frac{1}{2} n_x (n_x + 1)}{2} \\
&= \frac{2n_x (n_x + 1) + N(N+1) - n_y (n_y + 1) - n_x (n_x + 1)}{4} \\
&= \frac{n_x (n_x + 1) + N(N+1) - n_y (n_y + 1)}{4} = \frac{n_x^2 + n_x + N^2 + N - n_y^2 - n_y}{4} \quad (7.1179) \\
&= \frac{n_x^2 + n_x + N^2 + N - n_y^2 - (N - n_x)}{4} = \frac{n_x^2 + 2n_x + N^2 - (N - n_x)^2}{4} \\
&= \frac{n_x^2 + 2n_x + N^2 - (N^2 - 2Nn_x + n_x^2)}{4} = \frac{2n_x + 2Nn_x}{4} = \frac{n_x + Nn_x}{2} \\
&= \frac{1}{2} n_x (N+1)
\end{aligned}$$

Donc au final:

$$E(W_s) = \frac{1}{2} n_x (N+1) \quad (7.1180)$$

Pour le calcul de la variance, qui nous sera utile pour faire au besoin une approximation que nous verrons plus loin, apparaît (malheureusement) la covariance car la connaissance d'un des rangs donne une information partielle sur les autres. Nous avons donc:

$$V(W_s) = V\left(\sum_{i=1}^{n_x} R_i\right) = \sum_{i=1}^{n_x} V(R_i) + \sum_i \sum_{j \neq i} \text{cov}(R_i, R_j) \quad (7.1181)$$

Nous savons déjà au vu de ce que nous avons démontré juste plus haut que:

$$\sum_{i=1}^{n_x} V(R_i) = \sum_{i=1}^{n_x} \frac{N^2 - 1}{12} = \frac{n_x (N^2 - 1)}{12} = \frac{n_x ((n_x + n_y)^2 - 1)}{12} \quad (7.1182)$$

Le problème reste donc le terme avec la covariance. Pour la calculer il existe des techniques rigoureuses tenant sur plusieurs pages et une astuce qui est beaucoup plus courte. L'astuce consiste à utiliser la variable globale de rang T_i que nous noterons avec $i = 1 \dots n + m$. Comme la somme des T_i est une constante, nous avons alors:

$$\begin{aligned}
V\left(\sum_{i=1}^{n_x+n_y} T_i\right) &= V\left(\frac{N(N+1)}{2}\right) = V\left(\frac{(n_x+n_y)(n_x+n_y+1)}{2}\right) \\
&= 0 = \sum_{i=1}^{n_x+n_y} V(T_i) + \sum_{i=1}^{n_x+n_y} \sum_{\substack{j=1 \\ j \neq i}}^{n_x+n_y} \text{cov}(T_i, T_j) \\
&= (n_x+n_y) \frac{(n_x+n_y)^2 - 1}{12} + (n_x+n_y)(n_x+n_y-1) \text{cov}(T_i, T_j)
\end{aligned} \tag{7.1183}$$

Il vient alors:

$$\text{cov}(T_i, T_j) = -\frac{(n_x+n_y)^2 - 1}{12(n_x+n_y-1)} = -\frac{n_x+n_y+1}{12} \tag{7.1184}$$

Nous pouvons alors reprendre le calcul initial en remplaçant les covariances par leur expression, la dernière relation obtenue pour les covariances calculées sur les T_i s'appliquant également (ce qui n'est pas forcément intuitif... mais l'astuce fonctionne) aux R_i :

$$\begin{aligned}
V\left(\sum_{i=1}^{n_x} R_i\right) &= \frac{n_x \left((n_x+n_y)^2 - 1 \right)}{12} + n_x (n_x - 1) \left(-\frac{n_x+n_y+1}{12} \right) \\
&= \frac{1}{12} \left[n_x \left((n_x+n_y)^2 - 1 \right) - n_x (n_x - 1) (n_x+n_y+1) \right] \\
&= \frac{1}{12} \left[n_x (n_x^2 + 2n_x n_y + n_y^2 - 1) - (n_x^2 - n_x) (n_x+n_y+1) \right] \\
&= \frac{1}{12} \left[n_x^3 + 2n_x^2 n_y + n_x n_y^2 - n_x - n_x^2 (n_x+n_y+1) + n_x (n_x+n_y+1) \right] \\
&= \frac{1}{12} \left[n_x^3 + 2n_x^2 n_y + n_x n_y^2 - n_x - n_x^3 - n_x^2 n_y - n_x^2 + n_x^2 + n_x n_y + n_x \right] \\
&= \frac{1}{12} \left[n_x^2 n_y + n_x n_y^2 + n_x n_y \right] = \frac{n_x n_y (n_x+n_y+1)}{12}
\end{aligned} \tag{7.1185}$$

Soit au final:

$$V\left(\sum_{i=1}^{n_x} R_i\right) = \frac{n_x n_y (n_x+n_y+1)}{12} \tag{7.1186}$$

Ce qui est bien le même résultat que la méthode rigoureuse que l'on peut trouver dans certaines rares références.

Exemple:

E1. Passons à un cas pratique d'abord pour le cas exact. Considérons donc 2 échantillons comprenant 2

variables de traitement (X) et deux variables de contrôle (Y) (c'est un peu simpliste et absurde comme exemple mais cela facilite l'aspect pédagogique...), nous avons:

$$\begin{aligned} X &: 5.7, 3.2 \\ Y &: 8.1, 5.5 \end{aligned} \quad (7.1187)$$

Soit (la variable de traitement a donc le rangs 1 et 3 ce qui fait une somme de rang de 4):

$$\underline{3.2}, 5.5, \underline{5.7}, 8.1 \quad (7.1188)$$

Soit:

$$\begin{aligned} X &: R_1 = 3.2, R_3 = 5.7 \\ Y &: R_2 = 5.5, R_4 = 8.1 \end{aligned} \quad (7.1189)$$

Nous avons le tableau suivant comme nous l'avons montré plus haut:

Valeur de W_s	3	4	5	6	7
Probabilité	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Tableau: 7.31 - Probabilités associées au test de la somme des rangs Wilcoxon

avec dans le cas présent:

$$\begin{aligned} W_s &= 1 + 3 = 4 \\ E(W_s) &= \frac{1}{2} n_x (N+1) = \frac{1}{2} 2(4+1) = 5 \end{aligned} \quad (7.1190)$$

Si nous choisissons le seuil de confiance traditionnel à 5% en bilatéral, nous avons selon le tableau ci-dessus que:

$$P(W_s \geq 7) = \frac{1}{6} \cong 17\% \quad (7.1191)$$

$$P(W_s \leq 3) = \frac{1}{6} \cong 17\% \quad (7.1192)$$

Donc en d'autres termes nous voyons qu'il y a:

$$100\% - 2 \cdot \frac{1}{6}\% = 66.\bar{6}\% \quad (7.1193)$$

de probabilité cumulée que W_s soit compris entre 3 et 7 (la barre au-dessus du 6 signifie pour rappel que ce chiffre se répète à l'infini). Donc forcément 4 est compris dans l'intervalle bilatéral du 95%... et nous pouvons accepter l'hypothèse comme quoi les deux échantillons ne sont pas différents. La p -value correspondant en bilatéral est donc la moitié de 33.333333%.

Remarque: Au fait si nous voulions faire un exemple calculatoire manuel intéressant en jouant avec un seuil bilatéral de 5% (soit de 2.5% de chaque côté) il faudrait au moins 2 échantillons avec 4 variables aléatoires, soit 70 combinaisons de rangs possibles. En-dessous de 4 variables aléatoires par échantillons, il est évident que le test bilatéral à seuil de 95% sera tel qu'on ne rejettera jamais l'hypothèse d'égalité...

Si la taille des deux échantillons est assez grande (la majorité des praticiens considèrent que chaque échantillon doit avoir au moins 20 individus), il a été montré par simulations que nous pouvons faire l'approximation (utilisée par beaucoup de logiciels de statistiques):

$$N(0,1) = \frac{W_s - \frac{1}{2}n_x(N+1)}{\sqrt{\frac{n_x n_y (N+1)}{12}}} \quad (7.1194)$$

bien évidemment en déterminant ensuite toujours la p -value en bilatéral. Avec l'exemple précédent (n'ayant que 4 individus au total), nous avons donc:

$$N(0,1) = \frac{W_s - \frac{1}{2}n_x(N+1)}{\sqrt{\frac{n_x n_y (N+1)}{12}}} = \frac{4-5}{\sqrt{1.291}} \cong -0.7746 \quad (7.1195)$$

Ce qui correspondant à une probabilité cumulée de 21.93%. Donc la p -value correspondante en bilatéral est d'environ 44% (à comparer à la valeur d'environ 33% avec le cas exact).

15.1.3. TEST DE LA SOMME DES RANGS DE MANN-WITHNEY

Le "**test de la somme des rangs de Mann-Withney**" est au fait un test d'ajustement non-paramétrique très simple qui se déduit du test de la somme des rangs de Wilcoxon. Par ailleurs il en est inspiré à un tel point que nous l'appelons parfois dans l'industrie le "**test de Wilcoxon-Mann-Withney**" ou "**test d'ajustement de Wilcoxon-Mann-Withney**" ou encore "**test MWW**" (sans spécifier à chaque fois qu'il repose sur la somme des rangs).

Le but de ce test, identiquement au test de la somme des rangs de Wilcoxon, est de trouver un moyen de vérifier que deux échantillons indépendants non nécessairement de même taille sont issus d'une même loi ou non (in extenso sont issus d'une même population ou non) mais avec une approche différente!

Remarque: Au même titre que le test de la somme des rangs de Wilcoxon, le test de la somme des rangs de Mann-Withney peut tout à fait être utilisé pour des variables ordinales (donc catégorielles mais à condition qu'elles soient en un nombre acceptable).

Certains logiciels par ailleurs portent les choses à confusion car ils proposent le test de la somme des rangs de Wilcoxon sous le nom de test de de Mann-Withney... et inversement... et de plus n'indiquent pas ou ne proposent pas toujours le choix entre la version exacte ou approximative... Et en plus le test

de la somme des rangs de Wilcoxon et celui de la somme des rangs signés que nous verrons plus loin n'est pas différencié.... donc attention! C'est typiquement un problème dont la source est l'absence d'une norme ISO définissant la terminologie et les options qui doivent être disponibles...

Pour voir en quoi ce test consiste, construisons le tableau de rangs utilisant deux échantillons comprenant deux variables de contrôle et trois variables de traitement, nous avons alors:

Rangs possibles Contrôles	Rangs possibles Traitements	W_s
1, 2	3, 4, 5	12
1, 3	2, 4, 5	11
1, 4	2, 3, 5	10
1, 5	2, 3, 4	9
2, 3	1, 4, 5	10
2, 4	1, 3, 5	9
2, 5	1, 3, 4	8
3, 4	1, 2, 5	8
3, 5	1, 2, 4	7
4, 5	1, 2, 3	6

Tableau: 7.32 - Représentation des rangs de 3 variables de traitement et 2 de contrôle

Dont nous déduisons le tableau suivant:

Valeurs de W_s	6	7	8	9	10	11	12
Probabilité	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

Tableau: 7.33 - Probabilités associées au test de la somme des rangs de Wilcoxon

Maintenant imaginons que nous ayons une autre expérience à analyser utilisant deux échantillons comprenant trois variables de contrôle et deux variables de traitement (le symétrique du précédent donc!), nous avons alors:

Rangs possibles Contrôles	Rangs possibles Traitements	W_s
3, 4, 5	1, 2	3
2, 4, 5	1, 3	4
2, 3, 5	1, 4	5
2, 3, 4	1, 5	6
1, 4, 5	2, 3	5
1, 3, 5	2, 4	6
1, 3, 4	2, 5	7
1, 2, 5	3, 4	7
1, 2, 4	3, 5	8
1, 2, 3	4, 5	9

Tableau: 7.34 - Représentation des rangs de 2 variables de traitement et 3 de contrôle

Dont nous déduisons le tableau suivant (le lecteur remarquera que c'est exactement le même que le précédent en ce qui concerne les probabilités!!):

Valeurs de W_s	3	4	5	6	7	8	9
Probabilité	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

Tableau: 7.35 - Probabilités associées au test de de la somme des rangs Wilcoxon

Eh bien l'idée du test de Mann-Whitney est très simple:

Plutôt que de tabuler des situations symétriques, il suffit de soustraire à chaque valeur de W_s , la valeur $W_{s,\min}$ afin que chaque tableau soit identique et qu'un des deux seul soit utile. Voyons cela d'abord avec le premier tableau:

Rangs possibles Contrôles	Rangs possibles Traitements	$W_s - \frac{1}{2}n_x(n_x + 1)$
1, 2	3, 4, 5	6
1, 3	2, 4, 5	5
1, 4	2, 3, 5	4
1, 5	2, 3, 4	3
2, 3	1, 4, 5	4
2, 4	1, 3, 5	3
2, 5	1, 3, 4	2
3, 4	1, 2, 5	2
3, 5	1, 2, 4	1
4, 5	1, 2, 3	0

Tableau: 7.36 - Représentation des rangs de 3 variables de traitement et 2 de contrôle

Dont nous déduisons le tableau suivant:

Valeurs de W_{XY}	0	1	2	3	4	5	6
Probabilité	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

Tableau: 7.37 - Probabilités associées au test de la somme des rangs Wilcoxon

Maintenant imaginons que nous ayons une autre expérience à analyser utilisant deux échantillons comprenant trois variables de contrôle et deux variables de traitement, nous avons alors en utilisant la même idée:

Rangs possibles Contrôles	Rangs possibles Traitements	$W_s - \frac{1}{2}n_x(n_x + 1)$
3, 4, 5	1, 2	0
2, 4, 5	1, 3	1
2, 3, 5	1, 4	2
2, 3, 4	1, 5	3
1, 4, 5	2, 3	2
1, 3, 5	2, 4	3
1, 3, 4	2, 5	4
1, 2, 5	3, 4	4
1, 2, 4	3, 5	5
1, 2, 3	4, 5	6

Tableau: 7.38 - Représentation des rangs de 2 variables de traitement et 3 de contrôle

Dont nous déduisons cette fois-ci exactement le même tableau que précédemment:

Valeurs de W_{YX}	0	1	2	3	4	5	6
Probabilité	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

Tableau: 7.39 - Probabilités associées au test de Mann-Whitney

raison pour laquelle la littérature mentionne qu'on peut prendre celui que l'on veut!

Donc pour résumer, la variante de Mann-Whitney (dans le cas concret ici présent il s'agit de la variante dite "**variante exacte de Mann-Whitney**") consiste à tabuler pour les situations symétriques une variable notée W_{XY} définie naturellement par:

$$W_{XY} = W_s - \frac{1}{2}n_x(n_x + 1) \quad (7.1196)$$

notée aussi très souvent dans la littérature:

$$U_1 = R_1 - \frac{1}{2}n_1(n_1 + 1) \quad (7.1197)$$

car alors:

$$W_{XY} \in \{0, 1, 2, \dots\} \quad (7.1198)$$

et donc:

$$P(W_{XY} = k | n_x, n_y) = P(W_{XY} = k | n_y, n_x) \quad (7.1199)$$

Dans les tables que l'on peut trouver dans les livres, les probabilités sont données avec la valeur normalisée de U . Ainsi, si nous reprenons notre exemple précédent mais avec les notations d'usage dans la pratique (U au lieu de W_{YX}):

Valeurs de U	0	1	2	3	4	5	6
Probabilité	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

Tableau: 7.40 - Probabilités associées au test de Mann-Whitney

Nous voyons que la probabilité cumulée que $U = 2$ est de 0.4. La table précédente se trouve dans la littérature parfois sous la forme suivante:

$n_2 = 3$			
U / n_1	1	2	3
0	0.250	0.100	0.050
1	0.500	0.200	0.100
2	0.750	0.400	0.200
3	1	0.600	0.350
4		0.800	0.500
5		0.900	0.650
6		1	...

Tableau: 7.41 - Représentation classique du test de Mann-Whitney

où nous avons mis en rouge la colonne correspondant à notre exemple ($U = 2, n_1 = 2, n_2 = 3$) et en vert et gras la valeur prise comme exemple dans le paragraphe précédent. Ensuite il convient au praticien de choisir avec ces tableaux s'il souhaite faire un test bilatéral ou unilatéral.

Remarques:

R1. Il est important de se rappeler que nous avons démontré par l'exemple que nous pouvons aussi bien prendre:

$$W_{XY} = W_s - \frac{1}{2} n_x (n_x + 1) \quad (7.1200)$$

que:

$$W_{YX} = W_s - \frac{1}{2} n_y (n_y + 1) \quad (7.1201)$$

puisqu'ils génèrent les mêmes tableaux!

R2. W_{XY} est traditionnellement noté U par les praticiens comme nous l'avons vu, d'où le fait que l'on retrouve dans la littérature ce test sous le nom de "**test U de Mann-Withney**" avec les tables de probabilités associées sous le même nom. Cependant attention à ne pas confondre avec le "**test U de Wilcoxon**" appelé parfois "**test d'inversion de Wilcoxon**" qui se base sur les alternances d'apparition des valeurs des échantillons lorsque regroupés (test qui ne sera pas développé ici).

Pour voir la version approximative (asymptotique) du test U de Mann-Withney nous avons besoin de

l'espérance et de la variance. Pour cela, rappelons que nous avons donc vu que la somme des rangs normalisés était donnée par:

$$U_1 = R_1 - \frac{1}{2}n_1(n_1 + 1) \quad (7.1202)$$

mais nous pouvons aussi utiliser comme nous l'avons vu:

$$U_2 = R_2 - \frac{1}{2}n_2(n_2 + 1) \quad (7.1203)$$

et puisque:

$$R_1 + R_2 = \frac{N(N+1)}{2} \quad (7.1204)$$

avec pour rappel:

$$N = n_1 + n_2 \quad (7.1205)$$

nous avons donc:

$$\begin{aligned} U_1 + U_2 &= \left(R_1 - \frac{n_1(n_1 + 1)}{2} \right) + \left(R_2 - \frac{n_2(n_2 + 1)}{2} \right) \\ &= \frac{N(N+1)}{2} - \frac{n_1(n_1 + 1)}{2} - \frac{n_2(n_2 + 1)}{2} = \frac{(n_1 + n_2)(N+1)}{2} - \frac{n_1(n_1 + 1)}{2} - \frac{n_2(n_2 + 1)}{2} \\ &= \frac{n_1(N+1)}{2} + \frac{n_2(N+1)}{2} - \frac{n_1(n_1 + 1)}{2} - \frac{n_2(n_2 + 1)}{2} \\ &= \frac{n_1(n_1 + n_2 + 1)}{2} + \frac{n_2(n_1 + n_2 + 1)}{2} - \frac{n_1(n_1 + 1)}{2} - \frac{n_2(n_2 + 1)}{2} \\ &= \frac{n_1n_1 + n_1n_2 + n_1 + n_2n_1 + n_2n_2 + n_2 - n_1n_1 - n_1 - n_2n_2 - n_2}{2} = n_1n_2 \end{aligned} \quad (7.1206)$$

La moyenne des deux U est donc la moyenne arithmétique de la somme. Nous avons donc:

$$m_U = \frac{n_1n_2}{2} \quad (7.1207)$$

Ce qui signifie que U_1 ou U_2 doit être suffisamment différent de cette dernière moyenne pour que l'on rejette l'hypothèse nulle comme quoi les deux échantillons proviennent d'une même loi de distribution. Mais pour déterminer la p -value, nous avons qu'il nous faut aussi l'écart-type. Donc cherchons-le!

L'écart-type est le même que pour le test de la somme des rangs de Wilcoxon (puisque le deuxième terme dans l'expression des U est une constante dont la variance est nulle. Ainsi, il ne reste plus que la variance de la somme des rangs et nous avons déjà démontré plus haut qu'elle valait:

$$\sigma_U = \sqrt{\frac{n_1 n_2 (N+1)}{12}} \quad (7.1208)$$

Exemple:

Reprenons l'exemple fait avec le test de la somme des rangs de Wilcoxon mais un peu modifié (pour que l'exemple soit plus parlant) c'est-à-dire:

$$\begin{aligned} X: & 5.7, 3.2 \\ Y: & 8.1, 5.5, 1.2 \end{aligned} \quad (7.1209)$$

Soit groupé et ordonné:

$$\underline{1.2}, 3.2, \underline{5.5}, 5.7, \underline{8.1} \quad (7.1210)$$

Nous avons alors:

$$\begin{aligned} U_1 &= (2+4) - \frac{1}{2} 2(2+1) = 3 \\ U_2 &= (1+3+5) - \frac{1}{2} 3(3+1) = 3 \end{aligned} \quad (7.1211)$$

Donc nous pouvons choisir n'importe lequel pour le test vu que les deux U sont égaux. Si nous regardons le tableau créé plus haut, avec ($U=3, n_1=2, n_2=3$), nous avons donc une probabilité cumulée de 60% que U soit égal à 3. Donc nous ne rejetons pas l'hypothèse nulle (en unilatéral) comme quoi les deux échantillons proviennent de la même distribution.

L'approximation en loi Normale donne alors:

$$Z = \frac{U - m_U}{\sigma_U} = \frac{3 - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (N+1)}{12}}} = \frac{3 - \frac{3 \cdot 2}{2}}{\sqrt{\frac{3 \cdot 2 (5+1)}{12}}} = 0 \quad (7.1212)$$

Donc la probabilité cumulée est de 50% avec l'approximation Normale ce qui correspond à une p -value de 50%. Là encore nous ne rejetons pas l'hypothèse nulle.

15.1.4. TRAITEMENT DES ÉGALITÉS

Lorsque nous procédons à un test de la somme des rangs de type Wilcoxon-Mann-Whitney ou autre, des égalités de rangs peuvent se produire.

Reprenons pour l'exemple:

$$N = n_x + n_y = 2 + 3 = 5 \quad (7.1213)$$

avec les données suivantes:

Données:	17	17	17	19	21
Rang:	?	?	?	4	5

Tableau: 7.42 - Exemple de problème en cas d'égalités

Une solution conventionnelle consiste à attribuer à chaque "?" le rang moyen. Donc dans le cas présent, nous avons:

$$? = \frac{1}{3}(1+2+3) = 2 \quad (7.1214)$$

Le tableau:

Rangs possibles Contrôles	Rangs possibles Traitements	W_s
1, 2	3, 4, 5	12
1, 3	2, 4, 5	11
1, 4	2, 3, 5	10
1, 5	2, 3, 4	9
2, 3	1, 4, 5	10
2, 4	1, 3, 5	9
2, 5	1, 3, 4	8
3, 4	1, 2, 5	8
3, 5	1, 2, 4	7
4, 5	1, 2, 3	6

Tableau: 7.43 - Représentation des rangs de 3 variables de traitement et 2 de contrôle

devient alors dans ce cas particulier:

Rangs possibles Contrôles	Rangs possibles Traitements	W_s^*
2, 2	2, 4, 5	11
2, 2	2, 4, 5	11
2, 4	2, 2, 5	9
2, 5	2, 2, 4	8
2, 2	2, 4, 5	11
2, 4	2, 2, 5	9
2, 5	2, 2, 4	8
2, 4	2, 2, 5	9
2, 5	2, 2, 4	8
4, 5	2, 2, 2	6

Tableau: 7.44 - Représentation des rangs de 3 variables de traitement et 2 de contrôle

où W_s^* (remarquez la petite * en haut à droite!) représente la statistique de Wilcoxon lorsque nous sommes en présence d'égalités statistiques. La loi de W_s^* peut être plus ou moins différente de celle de W_s . Effectivement:

Statistique de Wilcoxon	6	7	8	9	10	11	12
Probabilité de W_s	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$
Probabilité de W_s^*	$\frac{1}{10}$	0	$\frac{3}{10}$	$\frac{3}{10}$	0	$\frac{3}{10}$	0

Tableau: 7.45 - Différence des deux tests statistiques en cas d'égalité ou non

15.1.5. TEST DE LA SOMME DES RANGS SIGNÉS DE WILCOXON À 1 ÉCHANTILLON

Le but du test de la "**somme des rangs signés de Wilcoxon**", appelé aussi parfois "**test de la médiane de Wilcoxon**", est d'utiliser une technique non paramétrique pour vérifier la symétrie ou non d'une distribution et donc in extenso faire une hypothèse sur la valeur de la médiane. L'idée est à la fois simple et subtile.

Le principe est que si nous comparons les différences D_i des individus d'un échantillon par rapport à la médiane, nous savons que si nous avons (par exemple) un nombre impair d'individus tous différents (non égaux), alors nous aurons 50% des données au-dessus et en-dessous de la médiane. Ensuite, pour contrôler que la distribution des valeurs des individus vérifie une certaine symétrie, l'idée (simple mais astucieuse) consiste ensuite à:

1. Calculer les différences en valeur absolue $|D_i|$ par rapport à la médiane
2. Ranger ces différences absolues par ordre croissant et leur assigner leur rang respectif
3. Calculer la somme des rangs des différences D_i qui à la base sont négatives
4. Calculer la somme des rangs des différences D_i qui à la base sont positives

et si l'échantillon a une distribution symétrique (donc la médiane est confondue alors avec la moyenne), il devrait y avoir une somme des rangs négatifs S_- qui n'est pas significativement différente de la somme des rangs positifs S_+ .

Remarque: Pour rappel, lors de notre étude des tests pour échantillons indépendants de Wilcoxon ou Mann-Whitney vus plus haut (qui n'ont pas obligatoirement la même taille), nous ordonnons ensemble les valeurs des deux échantillons et nous faisons un calcul sur les rangs de ces valeurs. Dans les tests pour échantillons appariés (donc de même taille), nous ordonnons les différences de valeurs (pas les valeurs!) et nous travaillons sur les rangs des différences.

Selon l'idée (principe) exposé plus haut, la somme des rangs qui portent le signe – vaut alors en moyenne:

$$\mu_- = \bar{R}_- = E(R_-) = E(1 \cdot X_1 + 2 \cdot X_2 + \dots + n \cdot X_n) = \sum_{i=1}^n i E(X_n) \quad (7.1215)$$

Or, nous avons déjà démontré que l'espérance de la loi binomiale est:

$$E(X_n) = np \quad (7.1216)$$

Et comme dans notre cas N vaut 1 (une seule valeur...) et p vaut $\frac{1}{2}$ (une chance sur deux d'avoir un signe négatif), il vient immédiatement en utilisant les démonstrations du chapitre de Suites Et Séries:

$$\mu_{\mathcal{L}} = \bar{R}_{\mathcal{L}} = E(R_{\mathcal{L}}) = \frac{1}{2} \sum_{i=1}^n i = \frac{1}{2} \frac{n(n+1)}{2} = \frac{n(n+1)}{4} \quad (7.1217)$$

et pour la variance en utilisant aussi les résultats du chapitre Suites Et Séries:

$$V(R_{\mathcal{L}}) = V\left(\sum_{i=1}^n iX_i\right) = \sum_{i=1}^n V(iX_i) = \sum_{i=1}^n i^2 V(X_i) \quad (7.1218)$$

et à nouveau en utilisant la variance de la loi binomiale et les résultats du chapitre Suites Et Séries:

$$V(R_{\mathcal{L}}) = \sum_{i=1}^n i^2 V(X_i) = \sum_{i=1}^n i^2 \frac{1}{4} = \frac{1}{4} \sum_{i=1}^n i^2 = \frac{1}{4} \frac{n(n+1)(2n+1)}{6} = \frac{n(n+1)(2n+1)}{24} \quad (7.1219)$$

Évidemment la somme des rangs des différences négatives (respectivement positif) sera au minimum nul et vaudra au maximum $n(n+1)/2$. Donc l'espérance dans le cas d'un test bilatéral ne doit pas être trop proche d'une de ces deux valeurs extrêmes.

Dans le cas où n est assez grand (supérieur à une trentaine), nous pouvons utiliser l'approximation de la loi Normale centrée réduite pour la variable:

$$N(0,1) = \frac{S_{\mathcal{L}} - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \quad (7.1220)$$

où $S_{\mathcal{L}}$ est donc la somme des rangs de signe négatifs.

Enfin signalons qu'empiriquement si des différences par rapport à la médiane sont nulles, elles ne seront pas prises en compte dans les rangs. Si des différences sont égales nous prendrons un rang moyen...

Exemple:

Commençons avec le cas à un échantillon comparé à sa médiane expérimentale (à l'opposé de la comparaison à une médiane hypothétisée lorsque nous considérons a priori la distribution symétrique et unimodale). Considérons que nous avons mesuré les valeurs suivantes pour le diamètre d'une pièce:

39, 20.2, 40, 32.2, 30.5, 26.5, 42.1, 45.6, 42.1, 45.6, 42.1, 29.9, 40.9

Nous souhaitons donc savoir si la médiane expérimentale calculée (valant 40 dans le cas présent) de cet échantillon peut ne pas être rejeté comme indicateur central à un niveau de confiance de 5% en bilatéral (ce qui sera le cas si le nombre de différences positifs et négatifs est assez équilibré). Nous construisons alors le tableau suivant:

Mesures	Différence	Valeur absolue	Rang	R_+	R_-
39	-1	1	2		2
20.2	-19.8	19.8	11		11
40	0	0	-		
32.2	-7.8	7.8	6		6
30.5	-9.5	9.5	8		8
26.5	-13.5	13.5	10		10
42.1	2.1	2.1	4	4	
45.6	5.6	5.6	6.5	6.5	
42.1	2.1	2.1	4	4	
45.6	5.6	5.6	6.5	6.5	
42.1	2.1	2.1	4	4	
29.9	-10.1	10.1	9		9
40.9	0.9	0.9	1	1	
		Somme:		26	46

Tableau: 7.46 - Tableau de traitement pour le test

A vue de nez l'égalité des rangs de ne s'annonce pas très bien mais allons quand même un peu plus loin...

Remarque: Suivant les ouvrages la somme des rangs ne donne pas la même valeur car il y a plusieurs techniques pour calculer des rangs de valeurs qui sont identiques... Nous avons cependant choisi celle utilisée par le logiciel Minitab qui est d'usage dans la communauté scientifique et qui correspond à celle dont nous avons déjà dictées les règles plus haut.

Si nous considérons que le nombre d'individus est suffisant..., nous utilisons l'approximation (même si dans le cas présent les conditions ne sont pas satisfaites):

$$N(0,1) = \frac{S_{-,+} - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \quad (7.1221)$$

Soit dans le cas présent:

$$N(0,1) = \frac{S_+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \cong \frac{26 - 33}{12.747} \cong -0.549 \quad (7.1222)$$

et respectivement:

$$N(0,1) = \frac{S_- - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \cong \frac{46 - 33}{12.747} \cong 1.02 \quad (7.1223)$$

Le premier cas correspond dans l'approximation à une loi Normale à une probabilité cumulée de 29.13 % obtenue avec la versions française de Microsoft Excel 14.0.6123 à l'aide de la fonction:

=LOI.NORMALE.STANDARD.N(-0.549;VRAI)

et donc à un p -value d'environ 58.26 % en bilatéral.

Le deuxième cas correspond dans l'approximation à une loi Normale à une probabilité cumulée à 84.62% obtenue avec la versions française de Microsoft Excel 14.0.6123 à l'aide de la fonction:

=LOI.NORMALE.STANDARD.N(1.02;VRAI)

ce qui correspond à une p -value d'environ 30.76 % en bilatéral (un logiciel comme Minitab donne une p -value en bilatéral de 32% puisqu'il ne fait pas l'approximation en loi Normale).

Pour le deuxième cas nous sommes à la limite mais au seuil choisi plus haut nous pouvons prudemment ne pas rejeter l'hypothèse comme quoi 40 est dans l'intervalle de confiance de la médiane (par ailleurs le test des signes amène à la même conclusion).

Remarque: Un logiciel comme Minitab bien que proposant le test de Wilcoxon à 1 échantillon de la médiane donne pour médiane la valeur de 36.5 et donne pour intervalle de confiance de la médiane les valeurs 31.1 et 42.1. Si nous appliquons la méthode de bootstrapping présentée en détails dans le chapitre de Méthodes Numériques nous obtenons comme médiane estimée 40 (pour moyenne 38.733) et comme intervalle 30.50 et 42.10... Bon dans tous les cas nous arrivons de toute façon à ne pas rejeter l'hypothèse nulle mais quand même...

15.1.6. TEST DE LA SOMME DES RANGS SIGNÉS DE WILCOXON POUR 2 ÉCHANTILLONS APPARIÉS

Le "**test de la somme des rangs signés de Wilcoxon pour 2 échantillons appariés**" est basé à 100% sur le principe du test à 1 échantillon. La seule différence est que l'hypothèse nulle ou alternative est basée sur la différence de la médiane des données prises deux à deux de chacun des échantillons. Dans la majorité des cas, l'hypothèse nulle est que la médiane des différences est nulle contre l'hypothèse alternative qu'elle est significativement différente de zéro.

Comme les développements mathématiques sont les même que pour le test à 1 échantillon, attaquons directement par un exemple.

Exemple:

Nous avons 2 logiciels (L1, L2) différents à comparer que nous voulons soumettre à 12 tâches (T1, T2, T3, ..., T12) de calculs spécifiques mais identiques pour chacun des logiciels. Nous souhaiterions savoir si les logiciels ont un temps de traitement significativement différent ou non et si oui lequel est le plus performant.

Nous avons alors le tableau suivant où le temps est en minutes et où les différences $x_i - y_i$ sont notées d_i :

Tâche	L1	L2	d_i	$ d_i $	Rang	R_+	R_-
T1	24.0	23.1	0.9	0.9	1	1	
T2	16.7	20.4	-3.7	3.7	4		4
T3	21.6	17.7	3.9	3.9	5	5	
T4	23.7	20.7	3.0	3.0	2.5	2.5	
T5	37.5	42.1	-4.6	4.6	6		6
T6	31.4	36.1	-4.7	4.7	7		7
T7	14.9	21.8	-6.9	6.9	10		10
T8	37.3	40.3	-3.0	3.0	2.5		2.5
T9	17.9	26.0	-8.1	8.1	11		11
T10	15.5	15.5	0.0	0.0	-		
T11	29.0	35.4	-6.4	6.4	9		9
T12	19.9	25.5	-5.6	5.6	8		8
					Somme:	8.5	57.5

Tableau: 7.47 - Tableau de traitement pour le test

Nous voyons déjà que le logiciel L1 est globalement plus rapide que L2 et sans utiliser les tables exactes du test des signes de Wilcoxon, nous pouvons présenter que la différence est significative.

Si nous considérons que le nombre d'individus est suffisant..., nous utilisons l'approximation (même si dans le cas présent les conditions ne sont pas satisfaites):

$$N(0,1) = \frac{S_{-,+} - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \quad (7.1224)$$

Soit dans le cas présent:

$$N(0,1) = \frac{S_+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \cong \frac{8.5 - 39}{12.747} \cong -2.392 \quad (7.1225)$$

et respectivement:

$$N(0,1) = \frac{S_- - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \cong \frac{57.5 - 39}{12.747} \cong 1.453 \quad (7.1226)$$

Le premier cas correspond dans l'approximation à une loi Normale à une probabilité cumulée de 0.84% obtenue avec avec la versions française de Microsoft Excel 14.0.6123 à l'aide de la fonction:

=LOI.NORMALE.STANDARD.N(-2.392;VRAI)

et donc à un p -value d'environ 1.68% en bilatéral.

Le deuxième cas correspond dans l'approximation à une loi Normale à une probabilité cumulée à 92.69% obtenue avec la versions française de Microsoft Excel 14.0.6123 à l'aide de la fonction:

$$=LOI.NORMALE.STANDARD.N(1.454;VRAI)$$

ce qui correspond à une p -value d'environ 14.62 % en bilatéral (un logiciel comme Minitab donne une p -value en bilatéral de 32% puisqu'il ne fait pas l'approximation en loi Normale).

Pour le premier cas, dans l'approximation par une loi Normale, nous rejetons clairement l'hypothèse nulle comme quoi la médiane des différences est nulle (donc les deux échantillons sont significativement différents). Par contre avec le deuxième cas nous avons une conclusion inverse. Évidemment, dans un telle situation il ne faudrait pas utiliser une approximation par une loi Normale d'où la situation un peu délicate...

Remarque: Avec un logiciel comme Minitab 15.1.2 qui ne propose pas le test de Wilcoxon pour échantillons appariés mais pour lequel il existe une astuce pour l'exécuter quand même, nous obtenons un p -value de 3.3%. D'autres logiciels donnent une p -value toujours inférieure à 5% (mais les valeurs diffèrent d'un logiciel à l'autre...).

15.2. TEST DE KRUSKAL-WALLIS

Le test de Kruskal-Wallis un test non paramétrique souvent assimilé (un peu rapidement...) à une ANOVA non paramétrique à une voie pour comparer si deux populations ou plus ont même médiane (hypothèse nulle) à la différence qu'il ne nécessite donc pas les hypothèses nécessaires au fonctionnement de l'ANOVA. Quand plusieurs populations comparées passent à travers ce test, ce dernier ne dit pas quelle population est significativement différente mais uniquement qu'il y en a au moins une qui l'est. En réalité, comme nous allons le démontrer, le test de Kruskal-Wallis n'est qu'une extension du test U de Mann-Whitney vu plus haut pour un nombre de populations supérieur ou égal à trois.

Pour étudier ce test, nous allons supposer que nous n'avons que deux populations et nous allons en faire une généralisation intuitive. Cette démarche est celle qu'aurait utilisée Wilcoxon avant que Kruskal et Wallis n'en fassent la démonstration générale rigoureuse.

Pour étudier ce test, rappelons d'abord que (relations dont l'origine et in extenso la démonstration ont déjà expliquées lors de notre étude du test de Mann-Withney vu plus haut) la moyenne de la somme des rangs et l'écart-type de la somme des rangs sont donnés par:

$$\begin{aligned}\mu_R = \bar{R} &= \frac{N+1}{2} \\ V(R) = \sigma_R^2 &= \frac{N^2 - 1}{12}\end{aligned}\quad (7.1227)$$

dans le cas où il n'y pas de valeurs doubles. Sous cette hypothèse, rappelons que $\bar{R}$ peut être assimilé au rang de la valeur médiane (dans le cas d'un nombre impair de mesures).

Rappelons que la moyenne des tirages de n valeurs sans remplacement parmi N sera proche d'une loi

Normale, et nous avons déjà démontré tout au début de ce chapitre que:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (7.1228)$$

et que si la population n'est pas très grande, la variance doit être corrigée par le facteur de correction sur population finie que nous avons déjà aussi démontré:

$$fcp = \sqrt{\frac{N-n}{N-1}} \quad (7.1229)$$

Dès lors, il vient:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \Rightarrow \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \frac{N-n}{N-1} \quad (7.1230)$$

Nous avons alors dans le cas qui nous concerne avec les rangs (la variance des rangs étant la variance vraie: il n'y a pas d'estimateur!):

$$\sigma_{\bar{R}}^2 = \frac{\sigma_R^2}{n} \frac{N-n}{N-1} = \frac{\left(\frac{N^2-1}{12}\right)}{n} \frac{N-n}{N-1} = \frac{(N-1)(N+1)}{12n} \frac{N-n}{N-1} = \frac{(N+1)(N-n)}{12n} \quad (7.1231)$$

Maintenant, de manière à former un variable Normale centrée réduite Z nous pouvons centrer et réduire la variable aléatoire $\hat{\bar{R}}$ obtenue par échantillonnage en écrivant:

$$Z \cong \frac{\hat{\bar{R}} - \bar{R}}{\sigma_{\bar{R}}} = \frac{\hat{\bar{R}} - \frac{N+1}{2}}{\sqrt{\frac{(N+1)(N-n)}{12n}}} \quad (7.1232)$$

où $\hat{\bar{R}}$ est donc la moyenne de la somme des rangs d'un échantillon de la population. Et au fait toute l'idée astucieuse du test de Kruskal-Wallis se trouve ici: la distribution statistique de la moyenne de la somme des rangs d'un grand nombre d'échantillons de N valeurs suit approximativement une loi Normale (revoir notre étude des limites des tirages sans remise)!

Prenons le carré :

$$Z^2 \cong \frac{12}{(N+1)(N-n)} n \left(\hat{\bar{R}} - \frac{N+1}{2} \right)^2 \cong \chi^2(1) \quad (7.1233)$$

L'approximation par la loi du Khi-deux n'étant valable que si n est assez grand comme nous en avons déjà parlé en détails lors de notre étude du test d'ajustement du Khi-deux.

Et donc la parenthèse de la première égalité est égale au carré de l'écart du rang de la médiane. Raison pour laquelle on dit souvent qu'il s'agit d'un test de la médiane (mais c'est un raccourci abusif).

Avant de continuer, insistons bien sur le fait que le scénario dans lequel nous nous trouvons est celui

d'un tirage d'un échantillon n parmi N , ce qui est équivalent à se retrouver avec deux échantillons (un de taille n et l'autre de taille $N-n$) de même loi (provenant in extenso d'une même population). Il vient alors que (relations que nous allons utiliser un peu plus loin):

$$n_i + n_j = n_1 + n_2 = n + (N - n) = N \quad (7.1234)$$

et par extension du cas à un échantillon, si nous notons R_i la somme des rangs des nombres de l'échantillon i , nous avons aussi:

$$n_i \bar{R}_i + n_j \bar{R}_j = n_1 \bar{R}_1 + n_2 \bar{R}_2 = n \bar{R}_1 + (N - n) \bar{R}_2 = N \cdot \bar{R} = N \frac{N+1}{2} = \frac{1}{2} N(N+1) \quad (7.1235)$$

Il s'ensuit que si nous notons pour la suite $\bar{R}_1 = \hat{\bar{R}} = \frac{R}{n}$ où R est donc la somme des rangs de l'échantillon $i = 1$, nous avons:

$$\bar{R}_2 = \frac{\frac{1}{2} N(N+1) - n \hat{\bar{R}}}{N - n} = \frac{\frac{1}{2} N(N+1) - R}{N - n} \quad (7.1236)$$

Si nous écrivons maintenant la relation démontrée plus haut:

$$\frac{12}{(N+1)(N-n)} n \left(\hat{\bar{R}} - \frac{N+1}{2} \right)^2 = \frac{12}{(N+1)(N-n)} n \left(\hat{\bar{R}} - \bar{R} \right)^2 \equiv \chi^2(1) \quad (7.1237)$$

sous la forme suivante (il s'agit d'un développement astucieux en marche arrière... à partir de la troisième ligne):

$$\begin{aligned}
\frac{12n}{(N+1)(N-n)} \left(\hat{\bar{R}} - \frac{N+1}{2} \right)^2 &= \frac{\hat{\bar{R}}^2 - (N+1)\hat{\bar{R}} + \frac{1}{4}(N+1)^2}{\frac{(N+1)(N-n)}{12n}} = \frac{\frac{R^2}{n^2} - (N+1)\frac{R}{n} + \frac{1}{4}(N+1)^2}{\frac{(N+1)(N-n)}{12n}} \\
&= \frac{\frac{12R^2}{n} - 12(N+1)R + 3n(N+1)^2}{(N+1)(N-n)} = \frac{12NR^2 - 12Nn(N+1)R + 3(N+1)^2 Nn^2}{N(N+1)n(N-n)} \\
&= \frac{12NR^2 - 12Nn(N+1)R + 3N^2(N+1)^2 n}{N(N+1)n(N-n)} - \frac{3(N+1)N(N+1)n(N-n)}{N(N+1)n(N-n)} \\
&= \frac{12}{N(N+1)} \left(\frac{(N-n)R^2 + \frac{1}{4}N^2(N+1)^2 n - Nn(N+1)R + nR^2}{n(N-n)} \right) - 3(N+1) \\
&= \frac{12}{N(N+1)} \left(\frac{R^2}{n} + \frac{\left(\frac{1}{2}N(N+1) - R \right)^2}{N-n} \right) - 3(N+1) = \frac{12}{N(N+1)} \left(\frac{\bar{R}_1^2}{n_1} + \frac{\bar{R}_2^2}{n_2} \right) - 3(N+1) \\
&= \frac{12}{N(N+1)} \left(\sum_{i=1}^2 \frac{\bar{R}_i^2}{n_i} \right) - 3(N+1)
\end{aligned}$$

(7.1238)

et nous retrouvons donc à la fin le fait que nous travaillions depuis le début avec deux échantillons, un de taille n et donc l'autre (in extenso par tirage) de taille $N-n$.

Le résultat précédent (qui était celui recherché depuis le début) peut être généralisé sous la forme suivante appelée "**test H de Kruskal-Wallis**" à un niveau de confiance donné en unilatéral (parfois cette relation est écrite sans les parenthèses pour la sommation ce qui peut prêter à une mauvaise lecture):

$$H = \frac{12}{N(N+1)} \left(\sum_{i=1}^c \frac{\bar{R}_i^2}{n_i} \right) - 3(N+1) \cong \chi^2(c-1) \quad (7.1239)$$

et si tous les n_i sont égaux, nous retrouvons cette relation sous la forme fréquente:

$$H = \frac{12}{nN(N+1)} \left(\sum_{i=1}^c \bar{R}_i^2 \right) - 3(N+1) \cong \chi^2(c-1) \quad (7.1240)$$

L'approximation suivant une loi du Khi-deux est cependant délicate lorsque la taille des échantillons (c) est petite (se référer à notre étude de la loi du Khi-deux).

Exemple:

Reprenons l'exemple de l'article original de Kruskal-Wallis. Nous considérons que nous avons trois machines à l'origine identiques mais dont deux ont subi quelques modifications. Nous avons mesuré la

production journalière un certain nombre de fois et avons obtenu le tableau suivant:

Standard	Rang	Modifiée 1	Rang	Modifiée 2	Rang	Somme
340	5	339	4	347	10	
345	9	333	2	343	7	
330	1	344	8	349	11	
342	6			355	12	
338	3					

n	5		3		4	12
R	24		14		40	78
R^2 / n	115.2		65.33		400	580.53

Tableau: 7.48 - Tableau d'exemple pour le test de Kruskal-Wallis

Nous avons alors bien:

$$\sum n_i = N = 12$$

$$\sum R_i = \frac{1}{2} N(N+1) = 78 \quad (7.1241)$$

et:

$$H = \frac{12}{N(N+1)} \left(\sum_{i=1}^c \frac{R_i^2}{n_i} \right) - 3(N+1) = \frac{12}{12 \cdot 13} 580.53 - 3 \cdot 13 = 5.656 \cong \chi^2(2) \quad (7.1242)$$

Or, nous avons:

$$P(\chi^2(2) \geq 5.656) = 0.059 \quad (7.1243)$$

Dans le cas présent, à un niveau de 5%, nous sommes donc à la limite avec l'approximation par une loi de Khi-deux. Comme l'ont montré Kruskal et Wallis, une simulation par Monte-Carlo donne une p -value de 0.049.

Bref, dans cette situation il conviendrait plutôt de rejeter l'hypothèse nulle comme quoi les productions sont similaires. Et donc privilégier le fait que celles-ci soient plutôt différentes. Une recommandation et de refaire le test par paire des mesures pour voir ce qui est significativement différent deux par deux.

15.3. TEST DE FRIEDMAN

Le test de Friedman, recommandé par la norme NF ISO 8587 pour l'analyse sensorielle (test de classement), considère une expérience avec deux facteurs (le premier étant considéré comme le traitement et le second comme les blocs de tests au même titre que l'ANOVA à deux facteurs contrôlés sans répétition) que l'on analyse à l'aide des rangs car les valeurs des mesures ne satisfont pas les conditions d'application d'ANOVA.

Associations, comme nous l'avons déjà fait à plusieurs reprises, la théorie à un exemple en partant du tableau suivant où huit sujets (blocs) B sous hypnose ont été soumis à quatre émotions (traitements) T .

Leur potentiel électrique épidermique a été mesuré (en millivolts) dans chaque cas (et l'ordre des traitements a été randomisé):

Émotion	1	2	3	4	5	6	7	8
Peur	23.1	57.6	10.5	23.6	11.9	54.6	21.0	20.3
Joie	22.7	53.2	9.7	19.6	13.8	47.1	13.6	23.6
Tristesse	22.5	53.7	10.8	21.1	13.7	39.2	13.7	16.3
Calme	22.6	53.1	8.3	21.6	13.3	37.0	14.8	14.8

Tableau: 7.49 - Tableau d'exemple des mesures pour le test de Friedman

L'idée centrale et subtile est de ne pas affecter un rang à l'ensemble de la population des mesures comme c'est le cas pour le test de Kruskal-Wallis (on perdrait alors le concept des blocs: in extenso du deuxième facteur) mais bien bloc par bloc tous supposés donc indépendants les uns des autres.

Remarque: Nous ne traiterons pas (au même titre que lors de notre étude du test de Kruskal-Wallis) de la situation où des mesures sont à égalité avec d'autres dans un même bloc, les démonstrations actuelles n'étant pas vraiment convaincantes.

Donc, à chaque valeur $\{x_{tb}\}_{T \times B}$ du tableau nous allons maintenant associer le rang $\{r_{tb}\}_{T \times B}$ correspondant à chaque traitement. Ce qui donnera:

Émotion	1	2	3	4	5	6	7	8
Peur	4	4	3	4	1	4	4	3
Joie	3	2	2	1	4	3	1	4
Tristesse	1	3	4	2	3	2	2	2
Calme	2	1	1	3	2	1	3	1

Tableau: 7.50 - Tableau d'exemple des rangs pour le test de Friedman

Bon maintenant que nous avons construit une sorte tableau d'ANOVA à deux facteurs contrôlés sans répétition non paramétrique que faisons-nous? Quelle est l'idée? Eh ben l'idée de base est la même que le test de Kruskal-Wallis: nous allons utiliser la propriété de la moyenne de la somme des rangs mais tout en ayant en tête que cette fois-ci la numérotation ne s'est pas faite sur l'ensemble des mesures du tableau mais bloc par bloc.

Dans le cadre de notre exemple particulier nous avons donc:

Émotion	1	2	3	4	5	6	7	8		
Peur	4	4	3	4	1	4	4	3	$R_1 = 27$	$\bar{R}_1 = 3.375$
Joie	3	2	2	1	4	3	1	4	$R_2 = 20$	$\bar{R}_2 = 2.50$
Tristesse	1	3	4	2	3	2	2	2	$R_3 = 19$	$\bar{R}_3 = 2.375$
Calme	2	1	1	3	2	1	3	1	$R_4 = 14$	$\bar{R}_4 = 1.75$

et en cas de non influence des traitements, nous nous attendons à avoir:

$$R_1 = R_2 = R_3 = R_4 = R_5 \quad (7.1244)$$

ou aussi (c'est équivalent):

$$\bar{R}_1 = \bar{R}_2 = \bar{R}_3 = \bar{R}_4 = \bar{R}_t \quad (7.1245)$$

S'il y a non influence des traitements ces quatre dernières valeurs devraient être égales et fluctuer autour de:

$$\mu_R = \bar{R} = \frac{T+1}{2} = \frac{4+1}{2} = 2.5 \quad (7.1246)$$

Nous pouvons pressentir que la fluctuation des $\bar{R}_t$ autour de μ_R doit suivre une loi Normale centrée s'il y a vraiment non influence (il existe une démonstration de cela dans l'article original de Friedman mais elle comporte des lacunes par moments et donc nous nous abstenons de la présenter). Nous pouvons également réduire la loi Normale telle que:

$$Z \equiv \frac{\bar{R}_t - \bar{R}}{\sigma_{\bar{R}}} = \frac{\bar{R}_t - \bar{R}}{\frac{\sigma_R}{\sqrt{B}}} = \frac{\bar{R}_t - \bar{R}}{\frac{\sqrt{\frac{T^2 - 1}{12}}}{\sqrt{B}}} \quad (7.1247)$$

Il n'est pas toujours intuitif que l'erreur standard soit obtenue par la division de la racine de B (du nombre de blocs) car la majorité des praticiens ont pour intuition de diviser par la racine T du nombre de traitements lorsqu'ils étudient l'aspect théorique du test de Friedman. Mais cela peut se vérifier avec une application numérique soit en se rappelant que le calcul de la variance σ_R se fait à partir des B rangs d'un traitement donné, rangs dont les valeurs (dans l'exemple ci-dessus ces valeurs sont comprises 8 fois entre 1 et 4) sont bien évidemment supposées pour un traitement donné indépendantes et identiquement distribuées.

Nous avons donc:

$$Z^2 = \frac{(\bar{R}_t - \bar{R})^2}{\sigma_{\bar{R}}^2} = \frac{(\bar{R}_t - \bar{R})^2}{\frac{\sigma_R^2}{B}} \quad (7.1248)$$

Contrairement au test de Kruskal-Wallis, nous ne faisons pas d'échantillonnage, donc nous ne devons pas corriger l'écart-type à l'aide du facteur de correction sur population finie (fcp) pour diminuer sa valeur.

L'idée de Friedmann (du moins c'est ainsi que nous allons le présenter) est de dire que l'écart-type de la somme des rangs des traitements obtenue de façon identique que lors du test de Kruskal-Wallis (dont l'origine a été détaillée lors de notre étude du test de Mann-Withney):

$$\sigma_R^2 = \frac{T^2 - 1}{12} \quad (7.1249)$$

n'est cette fois-ci qu'un estimateur de l'écart-type vrai et qu'il faut utiliser la relation entre l'estimateur non biaisé et biaisé pour corriger cette estimation (relation démontrée lors de notre étude des estimateurs):

$$\sigma_R^2 = \frac{T}{T-1} \hat{\sigma}_R^2 = \frac{T}{T-1} \frac{T^2 - 1}{12} \quad (7.1250)$$

Dès lors:

$$\sum_{t=1}^T Z_t^2 = \frac{\sum_{t=1}^T \left(\bar{R}_t - \frac{T+1}{2} \right)^2}{\frac{T}{T-1} \frac{\hat{\sigma}_R^2}{B}} = \frac{T-1}{T} \frac{\sum_{t=1}^T \left(\bar{R}_t - \frac{T+1}{2} \right)^2}{\frac{T^2 - 1}{12} \frac{1}{B}} \cong \chi^2(T-1) \quad (7.1251)$$

où nous avons retiré un degré de liberté au Khi-deux pour la raison déjà rencontrée maintes fois dans le présent chapitre.

Soit après quelques simplifications élémentaires nous obtenons le "test Q de Friedman":

$$Q = \frac{12B}{T(T+1)} \sum_{t=1}^T \left(\bar{R}_t - \frac{T+1}{2} \right)^2 \cong \chi^2(T-1) \quad (7.1252)$$

Pour en revenir à notre exemple il vient alors:

$$\begin{aligned} \frac{12B}{T(T+1)} \sum_{t=1}^T \left(\bar{R}_t - \frac{T+1}{2} \right)^2 &= \frac{12 \cdot 8}{4(4+1)} \sum_{t=1}^T \left(\bar{R}_t - \frac{4+1}{2} \right)^2 \\ &= \frac{24}{5} \left[\left(3.375 - \frac{5}{2} \right)^2 + \left(2.50 - \frac{5}{2} \right)^2 + \left(2.375 - \frac{5}{2} \right)^2 + \left(1.75 - \frac{5}{2} \right)^2 \right] = 6.45 \end{aligned} \quad (7.1253)$$

La valeur critique de $\chi^2(T-1)$ au seuil de 5% est de 7.65. Donc nous ne rejetons pas l'hypothèse comme quoi les traitements n'ont aucune influence (absence de différence entre les traitements). La probabilité cumulée correspondant à 7.65 (donc la p -value) est de 9%.

15.4. STATISTIQUES DES VALEURS EXTRÊMES

La statistique des valeurs extrêmes est un domaine très important dans la finance et l'ingénierie de la qualité (pour ne citer que les deux exemples les plus connus) qui permet d'étudier l'interpolation et la justification asymptotique des distributions. Comme le lecteur va le voir, cette statistique constitue de par sa construction un sous-domaine des statistiques d'ordre.

Remarque: Un exemple d'application pratique mondialement connu est son utilisation dans des résultats qui vont suivre dans le cadre des cartes de contrôle de la qualité (cf. chapitre de Génie Industriel).

Soient $X_1, X_2, \dots, X_n$ des variables aléatoires supposées indépendantes et identiquement distribuées de loi F et de densité f . Rappelons que nous définissons la statistique d'ordre i notée $X_{i:n}$ par:

$$X_{1,n} \leq \dots \leq X_{i,n} \leq \dots \leq X_{n,n} \quad (7.1254)$$

En posant:

$$\begin{aligned} W_n &= X_{1,n} = \min(X_1, X_2, \dots, X_n) \\ M_n &= X_{n,n} = \max(X_1, X_2, \dots, X_n) \end{aligned} \quad (7.1255)$$

Les variables W_n et M_n définissent les statistiques d'ordres extrêmes et leur écart:

$$D_n = M_n - W_n \quad (7.1256)$$

et dite "**dévi**ation **extrême**". Nous accepterons comme triviale la relation:

$$\min(X_{1,n}, X_{2,n}, \dots, X_{n,n}) = -\max(-X_{1,n}, -X_{2,n}, \dots, -X_{n,n}) \quad (7.1257)$$

Déterminons maintenant la fonction de répartition de M_n :

$$F_M(x) = P(\{M_n \leq x\}) = P\left(\bigcap_{i=1}^n (X_{i,n} \leq x)\right) \quad (7.1258)$$

car dire que $M_n \leq x$ équivaut à dire que pour chaque $X_{i,n}$ nous avons $X_{i,n} \leq x$ (pas facile à deviner qu'il faut avoir cette approche...).

Nous avons alors puisque les variables sont indépendantes (cf. chapitre de Probabilités):

$$F_M(x) = P\left(\bigcap_{i=1}^n (X_{i,n} \leq x)\right) = \prod_{i=1}^n P(X_{i,n} \leq x) = \prod_{i=1}^n F(x) = [F(x)]^n \quad (7.1259)$$

et par suite nous avons évidemment la fonction de distribution:

$$f_M(x) = n f(x) [F(x)]^{n-1} = \frac{\partial}{\partial x} F_M(x) = \frac{\partial}{\partial x} [F(x)]^n \quad (7.1260)$$

Respectivement en se basant sur la même idée:

$$\begin{aligned} F_W(x) &= P(\{W_n \leq x\}) = 1 - P(\{W_n > x\}) = 1 - P\left(\bigcap_{i=1}^n (X_{i,n} > x)\right) \\ &= 1 - \prod_{i=1}^n P(X_{i,n} > x) = 1 - \prod_{i=1}^n (1 - P(X_{i,n} \leq x)) = 1 - \prod_{i=1}^n (1 - F(x)) \quad (7.1261) \\ &= 1 - (1 - F(x))^n \end{aligned}$$

et par suite nous avons évidemment la fonction de distribution:

$$f_W(x) = n f(x) [1 - F(x)]^{n-1} = \frac{\partial}{\partial x} F_W(x) = \frac{\partial}{\partial x} (1 - (1 - F(x))^n) \quad (7.1262)$$

Il vient alors:

$$\begin{aligned}
 E(R) &= E(M_n - W_n) = E(M_n) - E(W_n) = \int_{-\infty}^{+\infty} x f_M(x) dx - \int_{-\infty}^{+\infty} x f_W(x) dx \\
 &= \int_{-\infty}^{+\infty} x (f_M(x) - f_W(x)) dx = \int_{-\infty}^{+\infty} x f_D(x) dx = \int_{-\infty}^{+\infty} x \frac{\partial}{\partial x} \left[[F(x)]^n - (1 - [1 - F(x)]^n) \right] dx
 \end{aligned} \quad (7.1263)$$

en ayant utilisé la linéarité de l'espérance et le fait que pour les deux fonctions de distribution nous travaillons sur la même variable aléatoire.

En faisant une intégration par parties (cf. chapitre de Calcul Différentiel et Intégral):

$$\begin{aligned}
 E(R) &= \bar{R} = \int_{-\infty}^{+\infty} x \frac{\partial}{\partial x} \left[[F(x)]^n - (1 - [1 - F(x)]^n) \right] dx \\
 &= x \left[[F(x)]^n - (1 - [1 - F(x)]^n) \right] \Big|_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} \left([F(x)]^n - (1 - [1 - F(x)]^n) \right) dx \quad (7.1264) \\
 &= \int_{-\infty}^{+\infty} (1 - [F(x)]^n - [1 - F(x)]^n) dx
 \end{aligned}$$

en n'oubliant pas que $F(-\infty) = 0$ et $F(+\infty) = 1$.

Maintenant considérons le cas particulier où la fonction de répartition suit une loi Normale centrée réduite:

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right) \quad (7.1265)$$

Nous avons alors:

$$\bar{R} = \int_{-\infty}^{+\infty} (1 - [F(x)]^n - [1 - F(x)]^n) dx = \int_{-\infty}^{+\infty} \left(1 - \Phi\left(\frac{x - \mu}{\sigma}\right)^n - \left[1 - \Phi\left(\frac{x - \mu}{\sigma}\right)^n \right] \right) dx \quad (7.1266)$$

Faisons un changement de variables:

$$y = \frac{x - \mu}{\sigma} \Rightarrow \frac{dy}{dx} = \frac{1}{\sigma} \Rightarrow dx = \sigma dy \quad (7.1267)$$

Nous avons alors:

$$\bar{R} = \sigma \int_{-\infty}^{+\infty} (1 - \Phi(y)^n - [1 - \Phi(y)]^n) dy \quad (7.1268)$$

et nous trouvons alors la relation donnée (99% du temps sans démonstrations) dans les livres de statistiques des procédés:

$$d_2(n) = \int_{-\infty}^{+\infty} \left(1 - \Phi(y)^n - [1 - \Phi(y)]^n \right) dy \quad (7.1269)$$

appelée "**constante de Hartley**" et donc:

$$\bar{R} = \sigma d_2(n) \quad (7.1270)$$

Cette constante est donc impossible à ce jour à calculer formellement. Soit il faut passer par des approximations en série de Taylor des termes de l'intégrale, ce qui devient un cauchemar pour n grand, soit par un calcul utilisant la méthode de Monte-Carlo (cf. chapitre de Méthodes Numériques). Comme c'est relativement long à implémenter dans un tableur, les ingénieurs qualité préfèrent utiliser des tables dans lesquelles nous trouvons par exemple:

Valeurs de n	Valeurs de $d_2(n)$ avec distribution normale
2	1.128
3	1.693
4	2.059
5	2.326
6	2.534
7	2.704
8	2.847
9	2.970
10	3.078

Tableau: 7.51 - Valeurs tabulées de la constante de Hartley

Voyons maintenant la variance de l'étendue en utilisant toujours la relation de Huyghens:

$$V(R) = E(R^2) - E(R)^2 = E(R^2) - \sigma^2 d_2^2(n) \quad (7.1271)$$

Le calcul de $E(R^2)$ est très peu digeste (du moins je n'ai rien trouvé de satisfaisant aux exigences de Sciences.ch), la plus petite démonstration complète tient sur 3 à 4 pages A4 et n'apporte formellement rien puisque nous finissons sur une intégrale non calculable à la main (par contre si quelqu'un a une démonstration simple, détaillée et élégante qu'il n'hésite pas à se manifester!). C'est pour cette raison qu'après avoir posé:

$$W = \frac{R}{\sigma} \quad (7.1272)$$

si nous écrivons comme le font de nombreux ouvrages techniques:

$$V(R) = \left[E(W^2) - d_2^2(n) \right] \sigma^2 \quad (7.1273)$$

il vient alors que:

$$\sigma_R = \sigma \sqrt{E(W^2) - d_2^2(n)} = \sigma d_3(n) \quad (7.1274)$$

Mais comme nous ne connaissons pas l'estimateur du maximum de vraisemblance non biaisé de l'écart-type σ , nous allons utiliser la relation démontrée:

$$\bar{R} = \sigma d_2(n) \quad (7.1275)$$

Pour avoir finalement un estimateur biaisé de la variance de l'étendue:

$$\hat{\sigma}_R = \frac{d_3(n)\bar{R}}{d_2(n)} = \frac{d_3\bar{R}}{d_2} \quad (7.1276)$$

Voici quelques valeurs tabulées de d_3 :

Valeurs de n	Valeurs de $d_3(n)$ avec distribution normale
2	1.128
3	1.693
4	2.059
5	2.326
6	2.534
7	2.498
8	2.459
9	2.423
10	2.931

Tableau: 7.52 - Valeurs tabulées de la constante d_3

15.4.1. TEST (DE L'ÉTENDUE) DE TUKEY

Supposons que nous avons Z_k variables aléatoires centrées réduites et indépendantes. Et notons U une variable aléatoire suivant une loi du Khi-deux à v degrés de liberté.

Définissons maintenant pour des raisons qui paraîtront évidentes un peu plus loin, "**l'étendue Studentisée**" (l'origine du nom provient de sa ressemblance avec la définition de la loi de Student) par:

$$Q_{k,v} = \frac{\max(Z_i) - \min(Z_i)}{\sqrt{U/v}} = \frac{R_Z}{\sqrt{U/v}} \quad (7.1277)$$

et tentons de déterminer si cette relation suit une loi connue et une application possible (nous retrouvons au numérateur ce que nous avons défini plus haut comme étant la "déviation extrême" mais avec une autre notation).

Pour cela, montrons que nous tombons sur la définition ci-dessus en considérant un cas un peu plus général où nous avons X_i variables aléatoires indépendantes qui suivent une loi Normale $N(\mu, \sigma^2)$ et avec l'écart type :

$$S_*^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \quad (7.1278)$$

et étudions le rapport:

$$\frac{\max(X_i) - \min(X_i)}{S_*} = \frac{R_X}{S_*} \quad (7.1279)$$

Maintenant, procédons aux transformations classiques déjà vues et démontrées et utilisées maintes fois depuis le début de ce chapitre:

$$\begin{aligned} Z_i &= \frac{X_i - \mu}{\sigma} \\ U &= \chi_v^2 = \frac{v S_*^2}{\sigma^2} \end{aligned} \quad (7.1280)$$

Et nous avons alors:

$$\begin{aligned} \frac{\max(X_i) - \min(X_i)}{S_*} &= \frac{\max\left(\frac{X_i - \mu}{\sigma}\right) - \min\left(\frac{X_i - \mu}{\sigma}\right)}{\sqrt{\frac{U}{v\sigma^2}}} \\ &= \frac{\left(\frac{1}{\sigma} \max(X_i) - \mu\right) - \left(\frac{1}{\sigma} \min(X_i) - \mu\right)}{\sqrt{\frac{U}{v\sigma^2}}} = \frac{1}{\sigma} \frac{\max(X_i) - \min(X_i)}{\frac{1}{\sigma} \sqrt{\frac{U}{v}}} \quad (7.1281) \\ &= \frac{\max(Z_i) - \min(Z_i)}{\sqrt{U/v}} = Q_{k,v} \end{aligned}$$

Donc voilà déjà pour la première étape. Pour l'instant, même si nous ne savons toujours pas si cette définition lui suit une distribution connue, nous pouvons déjà poser la définition très intéressante suivante (le terme de gauche est toujours positif):

$$\max(X_i) - \min(X_i) \leq Q_{k,v,1-\alpha} S_* \quad (7.1282)$$

ou autrement écrite:

$$R_X \leq Q_{k,v,1-\alpha} S_* \quad (7.1283)$$

et donc nous pouvons calculer quelle est la probabilité cumulée d'une étendue obtenue par mesures comparée à étendue critique $R_{X,crit}$ correspondant directement à un seuil α imposé. Ce qui nous amène à pouvoir écrire que:

$$P(R_X \leq R_{X,crit}) = 1 - \alpha \quad (7.1284)$$

Maintenant, rappelons que nous avons vu plus haut que la fonction de distribution de la déviation extrême était donnée par une relation à non connaissance non calculable analytiquement:

$$f_R(x) = \partial_x \left[[F(x)]^2 - \left(1 - [1 - F(x)]^n \right) \right] \quad (7.1285)$$

Donc la fonction de distribution $Q_{k,v,1-\alpha}$ n'est par conséquent pas assimilable à une loi connue quand $F(x)$ est une loi quelconque. Il faut donc malheureusement tabuler cette distribution par la méthode de Monte-Carlo (cf. chapitre de Méthodes Numériques) ou se référer à des tables déjà existantes.

Maintenant pour continuer, nous faisons un crochet par l'ANOVA à un facteur contrôlé que nous avons étudié. Rappelons d'abord que nous avons démontré que pour des variables aléatoires indépendantes et identiquement distribuées nous avons:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (7.1286)$$

et puisque l'ANOVA est aussi basé sur l'hypothèse que:

$$\sigma_R^2 = \sigma_A^2 = \sigma_T^2 = \sigma^2 \quad (7.1287)$$

cela implique qu'asymptotiquement les estimateurs ont la même propriété:

$$S_R^2 = S_A^2 = \sigma_T^2 = S^2 \quad (7.1288)$$

Nous savons aussi que l'écart-type de la moyenne d'un échantillon de l'ANOVA est donc donné la cadre et les hypothèses de l'ANOVA par:

$$\sigma_i = \frac{\sigma}{\sqrt{n}} \quad (7.1289)$$

Mais dans le cadre de l'ANOVA, nous avons aussi montré que sous les hypothèses imposées, nous avons:

$$S^2 = MSE = \frac{\sum_{ij} (x_{ij} - \bar{x}_i)^2}{N - k} \quad (7.1290)$$

Il vient alors que:

$$S_i = \sqrt{\frac{MSE}{n}} \quad (7.1291)$$

est un estimateur de:

$$\sigma_i = \frac{\sigma}{\sqrt{n}} \quad (7.1292)$$

Et comme nous avons démontré que:

$$\frac{(N-k)MSE}{\sigma^2} = \chi_{N-k}^2 \quad (7.1293)$$

Il vient alors que:

$$\frac{(N-k) \frac{MSE}{k}}{\frac{\sigma^2}{k}} = \chi_{N-k}^2 \quad (7.1294)$$

Dès lors, nous sommes naturellement amenés à constater que la relation que nous avons définie plus haut:

$$\max(X_i) - \min(X_i) \leq Q_{k, N-k, 1-\alpha} S^* \quad (7.1295)$$

Peut-être utilisée dans l'étude de l'ANOVA sous la forme:

$$\max(\bar{x}_i) - \min(\bar{x}_i) \leq Q_{k, N-k, 1-\alpha} \sqrt{\frac{MSE}{n}} \quad (7.1296)$$

pour faire un test préalable ou postérieur (post hoc) à une ANOVA à un facteur contrôlé pour vérifier l'hypothèse d'égalité des moyennes et identifier quelles sont les moyennes aberrantes. Donc le test de Tukey est souvent accompagné du test C de Cochran que nous avons déjà étudié plus haut lorsque nous faisons une ANOVA.

Soit, dans le cadre de l'ANOVA, nous devrions rejeter l'hypothèse d'égalité des moyennes des échantillons si:

$$\max(\bar{x}_i) - \min(\bar{x}_i) \geq Q_{k, N-k, 1-\alpha} \sqrt{\frac{MSE}{n}} \quad (7.1297)$$

ou autrement écrit:

$$\max(\bar{x}_i) - \min(\bar{x}_i) \geq R_{\mu, crit} \quad (7.1298)$$

Il est quasi immédiat alors que nous pouvons alors construire l'intervalle de confiance suivant:

$$(\bar{x}_i - \bar{x}_j) - R_{\mu, crit} \leq \mu_i - \mu_j \leq (\bar{x}_i - \bar{x}_j) + R_{\mu, crit} \quad (7.1299)$$

15.5. COEFFICIENT DE CORRÉLATION DES RANGS DE SPEARMAN

Le coefficient de corrélation des rangs de Spearman, noté R_S est le coefficient de corrélation de la suite $(R(i), S(i)), i = 1, \dots, n$, des rangs inspiré naturellement du coefficient de corrélation linéaire de Pearson vu au début de ce chapitre:

$$R_S = \frac{E[(R_i - \bar{R})(S_i - \bar{S})]}{\sqrt{V(R)V(S)}} = \frac{E(R \cdot S) - E(R)E(S)}{\sqrt{(E(R^2) - E(R)^2)(E(S^2) - E(S)^2)}} \quad (7.1300)$$

Prenons un exemple avant de nous attaquer à l'aspect théorique. Des mesures d'une population de taille 10 (nous avons repris les mêmes valeurs que celles prises pour les études des tests de rangs non paramétriques précédents):

X	Y
5.7	8.1
3.2	5.5
8.4	3.4
4.1	7.9
6.9	4.6
5.3	1.6
1.7	8.5
3.2	7.1
2.5	8.7
7.4	5.7

Tableau: 7.53 - Exemple de mesures

avec leurs rangs respectifs selon l'idée d'approche de Kendall (idée simple mais à laquelle il fallait penser!):

X	$R(i)$	Y	$S(i)$
5.7	7	8.1	8
3.2	3	5.5	4
8.4	10	3.4	2
4.1	5	7.9	7
6.9	8	4.6	3
5.3	6	1.6	1
1.7	1	8.5	9
3.2	4	7.1	6
2.5	2	8.7	10
7.4	9	5.7	5

Tableau: 7.54 - Valeurs mesurées et rangs respectifs

Maintenant démontrons que la relation donnée précédemment se simplifie drastiquement car les valeurs de R , comme celles de S , parcourent la suite des n premiers entiers. Or nous avons démontré dans le chapitre de Suites et Séries que:

$$\sum_{i=1}^n i = \frac{n(n+1)}{2} \quad (7.1301)$$

donc:

$$E(R) = E(S) = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2} \quad (7.1302)$$

Il vient alors:

$$R_S = \frac{E(R \cdot S) - \left(\frac{n+1}{2}\right)^2}{\sqrt{\left(E(R^2) - \left(\frac{n+1}{2}\right)^2\right)\left(E(S^2) - \left(\frac{n+1}{2}\right)^2\right)}} \quad (7.1303)$$

Nous avons également démontré dans le chapitre de Suites et Séries que:

$$\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6} \quad (7.1304)$$

donc:

$$E(R^2) = E(S^2) = \frac{1}{n} \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6} \quad (7.1305)$$

Il vient alors:

$$\begin{aligned} R_S &= \frac{E(R \cdot S) - \left(\frac{n+1}{2}\right)^2}{\frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2} = \frac{12E(R \cdot S) - 12\left(\frac{n+1}{2}\right)^2}{2(n+1)(2n+1) - 3(n+1)^2} \\ &= \frac{12E(R \cdot S) - 3(n+1)^2}{(2n+2)(2n+1) - 3(n^2 + 2n + 1)} = \frac{12E(R \cdot S) - 3(n^2 + 2n + 1)}{4n^2 + 2n + 4n + 2 - 3n^2 - 6n - 3} \\ &= \frac{12E(R \cdot S) - 3n^2 - 6n - 3}{n^2 - 1} = \frac{12 \frac{1}{n} \sum_{i=1}^n R_i S_i - 3n^2 - 6n - 3}{n^2 - 1} \quad (7.1306) \\ &= \frac{12 \sum_{i=1}^n R_i S_i}{n^3 - n} - \frac{3n^2 + 6n + 3}{n^2 - 1} = \frac{12 \sum_{i=1}^n R_i S_i}{n^3 - n} - \frac{3(n+1)^2}{n^2 - 1} = \frac{12 \sum_{i=1}^n R_i S_i}{n^3 - n} - \frac{3(n+1)^2}{(n+1)(n-1)} \\ &= \frac{12 \sum_{i=1}^n R_i S_i}{n(n^2 - 1)} - \frac{3(n+1)}{n-1} = \frac{6 \sum_{i=1}^n 2R_i S_i}{n(n^2 - 1)} - \frac{3(n+1)}{n-1} \end{aligned}$$

Maintenant jouons un peu pour obtenir une expression encore plus simplifiée en observant que:

$$\sum_{i=1}^n (R_i - S_i)^2 = \sum_{i=1}^n R_i^2 - 2R_i S_i + S_i^2 = \sum_{i=1}^n R_i^2 - 2 \sum_{i=1}^n R_i S_i + \sum_{i=1}^n S_i^2 \quad (7.1307)$$

il vient alors que:

$$2 \sum_{i=1}^n R_i S_i = \sum_{i=1}^n R_i^2 + \sum_{i=1}^n S_i^2 - \sum_{i=1}^n (R_i - S_i)^2 = \sum_{i=1}^n R_i^2 + \sum_{i=1}^n S_i^2 - \sum_{i=1}^n d_i^2 \quad (7.1308)$$

Nous avons alors:

$$\begin{aligned}
 R_S &= \frac{6 \sum_{i=1}^n 2R_i S_i}{n(n^2-1)} - \frac{3(n+1)}{n-1} = \frac{6 \left(\sum_{i=1}^n R_i^2 + \sum_{i=1}^n S_i^2 - \sum_{i=1}^n d_i^2 \right)}{n(n^2-1)} - \frac{3(n+1)}{n-1} \\
 &= \frac{6 \left(\sum_{i=1}^n R_i^2 + \sum_{i=1}^n S_i^2 \right)}{n(n^2-1)} - \frac{3(n+1)}{n-1} - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)}
 \end{aligned} \quad (7.1309)$$

Or, nous avons démontré que:

$$\sum_{i=1}^n i^2 = \sum_{i=1}^n R_i^2 = \sum_{i=1}^n S_i^2 = \frac{n(n+1)(2n+1)}{6} \quad (7.1310)$$

Dès lors:

$$\begin{aligned}
 R_S &= \frac{6 \left(\sum_{i=1}^n R_i^2 + \sum_{i=1}^n S_i^2 \right)}{n(n^2-1)} - \frac{3(n+1)}{n-1} - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)} \\
 &= \frac{2n(n+1)(2n+1)}{n(n^2-1)} - \frac{3(n+1)}{n-1} - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)} = \frac{2(n+1)(2n+1)}{n^2-1} - \frac{3(n+1)}{n-1} - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)} \\
 &= \frac{2(n+1)(2n+1)}{(n-1)(n+1)} - \frac{3(n+1)(n+1)}{(n-1)(n+1)} - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)} \\
 &= \frac{2(n+1)(2n+1) - 3(n+1)(n+1)}{(n-1)(n+1)} - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)} \\
 &= \frac{2(2n+1) - 3(n+1)}{(n-1)} - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)} = \frac{4n+2-3n-3}{(n-1)} - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)} \\
 &= \frac{n-1}{(n-1)} - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)} = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)}
 \end{aligned} \quad (7.1311)$$

Ainsi, nous trouvons la fameuse relation disponible dans tous les livres de Statistiques au final:

$$R_S = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)} \quad (7.1312)$$

Le coefficient de Spearman reprend les propriétés essentielles du coefficient de Pearson à savoir que:

$$-1 \leq R_g \leq +1 \quad (7.1313)$$

et prend la valeur 0 lorsque les variables sont indépendantes (en n'oubliant pas les subtilités importantes y relatives déjà mentionnées lors de notre étude du coefficient de Pearson).

Remarque: Toujours conséquence du fait qu'il soit non paramétrique, le coefficient de corrélation de Spearman peut traiter les variables intrinsèquement ordinales: un indice de satisfaction, une appréciation ou une note attribuée, etc.

16. CALCULS D'ERREURS/INCERTITUDES

Il est impossible de connaître (mesurer) la valeur exacte d'une grandeur physique expérimentalement, il est très important donc d'en déterminer l'incertitude.

Nous appelons bien évidemment "erreur", la différence entre la valeur mesurée et la valeur exacte. Cependant, comme nous ignorons la valeur exacte, nous ne pouvons pas connaître l'erreur commise quand même.... Le résultat est donc toujours incertain. C'est la raison pour laquelle nous parlons des "**incertitudes de mesure**".

Nous distinguons deux types d'incertitudes:

1. Les "erreurs systématiques": elles affectent le résultat constamment et dans le même sens (erreurs des appareils de mesures, limites de précision, etc.). Il faut alors éliminer, ou corriger le résultat, si possible !
2. Les "erreurs accidentelles" (statistiques): il faut alors répéter les mesures, calculer la moyenne et évaluer l'incertitude en utilisant les outils de statistique.

Le deuxième type d'erreurs fait un très gros usage de tous les outils statistiques que nous avons présentés jusqu'à maintenant. Nous ne reviendrons donc pas dessus et nous nous concentrerons alors uniquement sur quelques nouveaux concepts.

16.1. INCERTITUDES ABSOLUES ET RELATIVES

Si la vraie valeur d'une grandeur est x (supposée connue théoriquement) et la valeur mesurée est x_0 , alors Δx_0 est "**l'incertitude absolue**" (l'incertitude due aux appareils de mesure) ou "**erreur absolue**".

L'intervalle de confiance de la mesure est alors notée:

$$x_0 - \Delta x_0 \leq x \leq x_0 + \Delta x_0 \quad (7.1314)$$

ou:

$$x = x_0 \pm \Delta x \quad (7.1315)$$

"**L'incertitude relative**" ou "**erreur relative**" est quant à elle définie par:

$$\frac{\Delta x_0}{x_0} \quad (7.1316)$$

L'incertitude absolue permet de connaître l'approximation du dernier chiffre significatif de celle-ci. Par contre, lorsque nous désirons comparer deux mesures ayant des incertitudes absolues afin de déceler laquelle a la plus grande marge d'erreur, nous calculons l'incertitude relative de ce nombre en divisant l'incertitude absolue par le nombre, et transformons en pourcentage.

En d'autres termes, l'incertitude relative permet d'avoir une idée de la précision de la mesure en %. Si nous faisons une mesure avec une incertitude absolue de 1 [mm], nous ne saurons pas si c'est une bonne mesure ou non. Ça dépend si nous avons mesuré la taille d'une pièce de monnaie, de notre voisin, de la distance Paris-Marseille ou de la distance Terre-Lune. Bref, ça dépend de l'incertitude relative (c'est-à-dire du rapport de l'incertitude absolue sur la mesure).

16.2. ERREURS STATISTIQUES

Dans la plupart des mesures, nous pouvons estimer l'erreur due à des phénomènes aléatoires, appelée "**erreur aléatoire**", par une série de n mesures $x_1, x_2, \dots, x_i, \dots, x_n$ et ce à l'opposé de "**l'erreur systématique**" qui est la part non aléatoire de l'erreur.

L'erreur aléatoire permet d'introduire les notions de:

- Répétabilité: qui est définie comme l'étroitesse de l'accord entre les résultats de mesurages successifs d'une même grandeur, effectués avec la même méthode, par le même opérateur, avec les mêmes instruments de mesure, dans le même laboratoire, et à des intervalles de temps assez courts (voir plus un peu plus bas un traitement et une définition plus rigoureuse conforme aux normes internationales).
- Reproductibilité (parfois appelé "justesse"): qui est définie comme l'étroitesse de l'accord entre les résultats de mesurages successifs d'une même grandeur, dans le cas où les mesurages individuels sont effectués: suivant différentes méthodes, au moyen de différents instruments de mesure, par différents opérateurs dans différents laboratoires.

Ces deux notations sont toujours regroupées sous le sigle "**R&R**" ou "**Étude R&**" dans l'industrie. En général, l'accord est moins bon quand il s'agit de reproductibilité.

Remarque: Il existe des logiciels informatiques basés sur l'ANOVA à deux facteurs avec répétition comme Minitab qui génèrent des rapports très détaillés pour les analyses R&R.

Ces deux types d'erreurs peuvent être illustrés par le tir à la cible de façon plus générale:

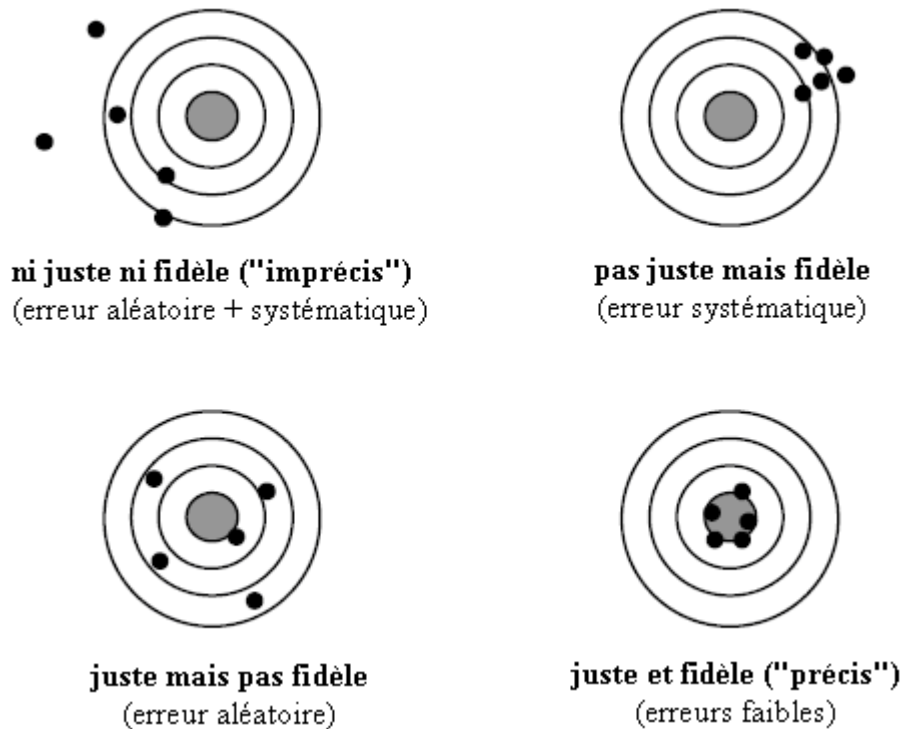


Figure: 7.73 - Types d'erreurs en ingénierie de laboratoire

Comme nous l'avons vu plus haut, la valeur moyenne arithmétique sera alors:

$$\bar{x} = \mu = \frac{1}{n} \sum_{i=1}^n x_i \quad (7.1317)$$

et l'écart moyen (estimateur biaisé démontré plus haut):

$$\hat{\sigma}^2 = \overline{\Delta x} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (7.1318)$$

et l'écart quadratique moyen ou écart-type (estimateur sans biais):

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (7.1319)$$

et nous avons démontré que l'écart-type de la moyenne était donné par:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (7.1320)$$

et comme nous l'avons vu, après un grand nombre de mesures indépendantes, la distribution des erreurs sur une mesure suit une loi Normale telle que nous puissions écrire (si nous n'avons pas assez de mesures, nous utiliserons l'I.C. basé sur la loi de Student):

$$\begin{aligned} \bar{x}_i - \sigma < x_i < \bar{x}_i + \sigma & \quad 68\% \\ \bar{x}_i - 2\sigma < x_i < \bar{x}_i + 2\sigma & \quad 95.4\% \end{aligned} \quad (7.1321)$$

bref nous pouvons réutiliser tous les outils statistiques vus jusqu'ici dans le domaine de la mesure en laboratoire ou ailleurs!

Le résultat d'une mesure doit ainsi comporter en toute rigueur 4 éléments. Par exemple:

$$\underbrace{125.3}_1 \underbrace{[Kg]}_2 \pm \underbrace{1.7}_3 \underbrace{[Kg]}_4 \underbrace{(k=2)}_4 \quad (7.1322)$$

où nous avons:

1. La valeur numérique avec un nombre correct de décimales
2. Unité de la mesure selon le standard du système international
3. Incertitude élargie de $k \cdot \sigma$ (intervalle de confiance)
4. La valeur entière du k utilisée pour l'intervalle de confiance.

16.3. RÉPÉTABILITÉ

La répétabilité r , mesure de l'écart probable entre deux mesurages sur des objets de même nature, dans un même laboratoire, sous des conditions opératoires semblables, est définie normativement (dans les normes ISO 5725:1987 et AFNOR NF X 06-041 *Fidélité des méthodes d'essai*) dans le cas monodimensionnel par:

$$P(|X_1 - X_2| < r) = p \quad (7.1323)$$

où p est une probabilité élevée, généralement égale à 95% et X_1, X_2 deux variables indépendantes et identiquement distribuées selon une loi Normale d'espérance et variance inconnues $N(\mu, \sigma^2)$. De par la stabilité de la loi Normale, il vient alors:

$$X_1 - X_2 = N(\mu, \sigma^2) - N(\mu, \sigma^2) = N(0, 2\sigma^2) \quad (7.1324)$$

Or, nous avons vu au début de ce chapitre dans le cadre de l'étude de l'intervalle de confiance de la moyenne que:

$$P\left(Y \geq \frac{\delta}{\sigma}\right) = 0.025 \quad (7.1325)$$

Donc in extenso:

$$P\left(Y \leq \frac{\delta}{\sigma}\right) = 0.975 \quad (7.1326)$$

et dès lors en utilisant les tables, nous avons:

$$\frac{\delta}{\sigma} \cong 1.96 \quad (7.1327)$$

Et donc:

$$\delta \cong 1.96\sigma \quad (7.1328)$$

Soit avec la notation des normes pour laboratoires:

$$r \cong 1.96\sigma \quad (7.1329)$$

Mais dans le cas présent, nous avons une variance doublée. Donc il vient:

$$r \cong 1.96\sqrt{2}\sigma \cong 2.77\sigma \quad (7.1330)$$

Nous retrouvons dans la relation disponible dans la norme avec le fameux coefficient de 2.77. Évidemment après il est évident que la valeur de r doit être minimisée!

16.4. PROPAGATION DES ERREURS

Soit une mesure $x \pm \Delta x$ et $y = f(x)$ une fonction de x . Quelle est l'incertitude sur y si nous connaissons uniquement l'incertitude d'un appareil de mesure mais qui ne serait pas donnée sous forme d'écart-type statistique?

Lorsque Δx est petit, $f(x)$ est remplacé au voisinage de x par sa tangente (il s'agit simplement de la dérivée bien sûr):

$$\Delta y = \left| \frac{df}{dx} \right| \Delta x \quad (7.1331)$$

mais si y dépend de plusieurs grandeurs x, z, t mesurées avec les incertitudes $\Delta x, \Delta z, \Delta t$:

$$y = f(x, z, t) \quad (7.1332)$$

l'erreur maximale possible est alors la différentielle totale exacte (cf. chapitre de Calcul Différentiel Et Intégral):

$$\Delta y = \left| \frac{\partial f}{\partial x} \right| \Delta x + \left| \frac{\partial f}{\partial z} \right| \Delta z + \left| \frac{\partial f}{\partial t} \right| \Delta t \quad (7.1333)$$

Ce que nous notons aussi souvent:

$$\Delta y = \sum_{i=1}^n \left| \frac{\partial y}{\partial x_i} \right| \Delta x_i \quad (7.1334)$$

Ce qui conduit à:

$$\left. \begin{array}{l} y = x + z \\ y = x - z \end{array} \right\} \Rightarrow \Delta y = \Delta x + \Delta z$$

(7.1335)

$$\left. \begin{array}{l} y = xz \Rightarrow \Delta y = z\Delta x + x\Delta z \\ y = \frac{x}{z} \Rightarrow \Delta y = \frac{z\Delta x + x\Delta z}{z^2} \end{array} \right\} \Rightarrow \frac{\Delta y}{y} = \frac{\Delta x}{x} + \frac{\Delta z}{z}$$

Il apparaît ainsi clairement qu'une opération mathématique ne peut améliorer l'incertitude sur les données.

Remarque: Le résultat d'une multiplication, d'une division, d'une soustraction ou d'une addition est arrondi à autant de chiffres significatifs que la donnée qui en comporte le moins.

Si l'incertitude de l'appareil de mesure est donnée sous forme statistiques (écart-type), il est évident dès lors que nous allons utiliser les propriétés de la variance déjà vues au début de ce chapitre... pour des cas simples.

16.5. CHIFFRES SIGNIFICATIFS

Dans les petites écoles (et aussi les plus grandes parfois), il est demandé de transformer une mesure exprimée en une certaine unité en une autre unité.

Par exemple, en prenant les tables, nous pouvons avoir le type de conversion suivante:

$$140 \text{ [lb]} = 140 \cdot 0.45359237 \text{ [kg/lb]} = 63.5029318 \text{ [kg]} \quad (7.1336)$$

Vient alors la question suivante (que l'élève peut avoir oublié...). Au départ d'une mesure dont la précision est de l'ordre de 1 [lb] (donc de l'ordre de 0.5 [kg]), une simple conversion d'unité pourrait-elle amener à une précision au 1/10 [mg] près ?

De cet exemple il faut donc retenir qu'une marge d'incertitude est associée à toute valeur mesurée et à toute valeur calculée à partir de valeurs mesurées.

Dans les sciences exactes, tout raisonnement, toute analyse doit prendre cette incertitude en compte.

Mais pourquoi des chiffres sont-ils significatifs et d'autres pas alors ? Parce qu'en sciences, nous ne rapportons que ce qui a objectivement été observé (principe d'objectivité). En conséquence, nous limitons l'écriture d'un nombre aux chiffres raisonnablement fiables en dépit de l'incertitude: les chiffres significatifs. La précision que des chiffres supplémentaires sembleraient apporter est alors illusoire.

Il faut alors savoir arrondir selon des règles et conventions:

- Lorsque le chiffre de rang le plus élevé qu'on laisse tomber est supérieur à 5, le chiffre précédent est augmenté de 1 (exemple: 12.66 s'arrondit à 12.7). Dans la version anglaise de Microsoft Excel 11.8346:

$$=\text{ROUND}(12.66;1)=12.7$$

- Lorsque le chiffre de rang le plus élevé qu'on laisse tomber est inférieur à 5, le chiffre précédent reste inchangé (exemple 12.64 s'arrondit à 12.6). Dans la version anglaise de Microsoft Excel 11.8346:

$$=\text{ROUND}(12.64;1)=12.6$$

- Lorsque le chiffre de rang le plus élevé qu'on laisse tomber est égal à 5, si un des chiffres qui le suivent n'est pas nul, le chiffre précédent est augmenté de 1 (exemple: 12.6502 s'arrondit à 12.7). Dans la version anglaise de Microsoft Excel 11.8346:

$$=\text{ROUND}(12.6502;1)=12.7$$

- Si le chiffre de rang le plus élevé que nous laissons tomber est un 5 terminal (qui n'est suivi d'aucun chiffre) ou qui n'est suivi que de zéros, nous augmentons de 1 le dernier chiffre du nombre arrondi s'il est impair, sinon nous le laissons inchangé (exemples: 12.75 s'arrondit à 12.8 et 12.65 à 12.6). Dans ce dernier cas, le dernier chiffre du nombre arrondi est toujours un chiffre pair. Les tableurs ne respectent pas vraiment cette dernière règle, effectivement avec la version anglaise de Microsoft Excel 11.8346 nous avons:

$$=\text{ROUND}(12.75;1)=12.8$$

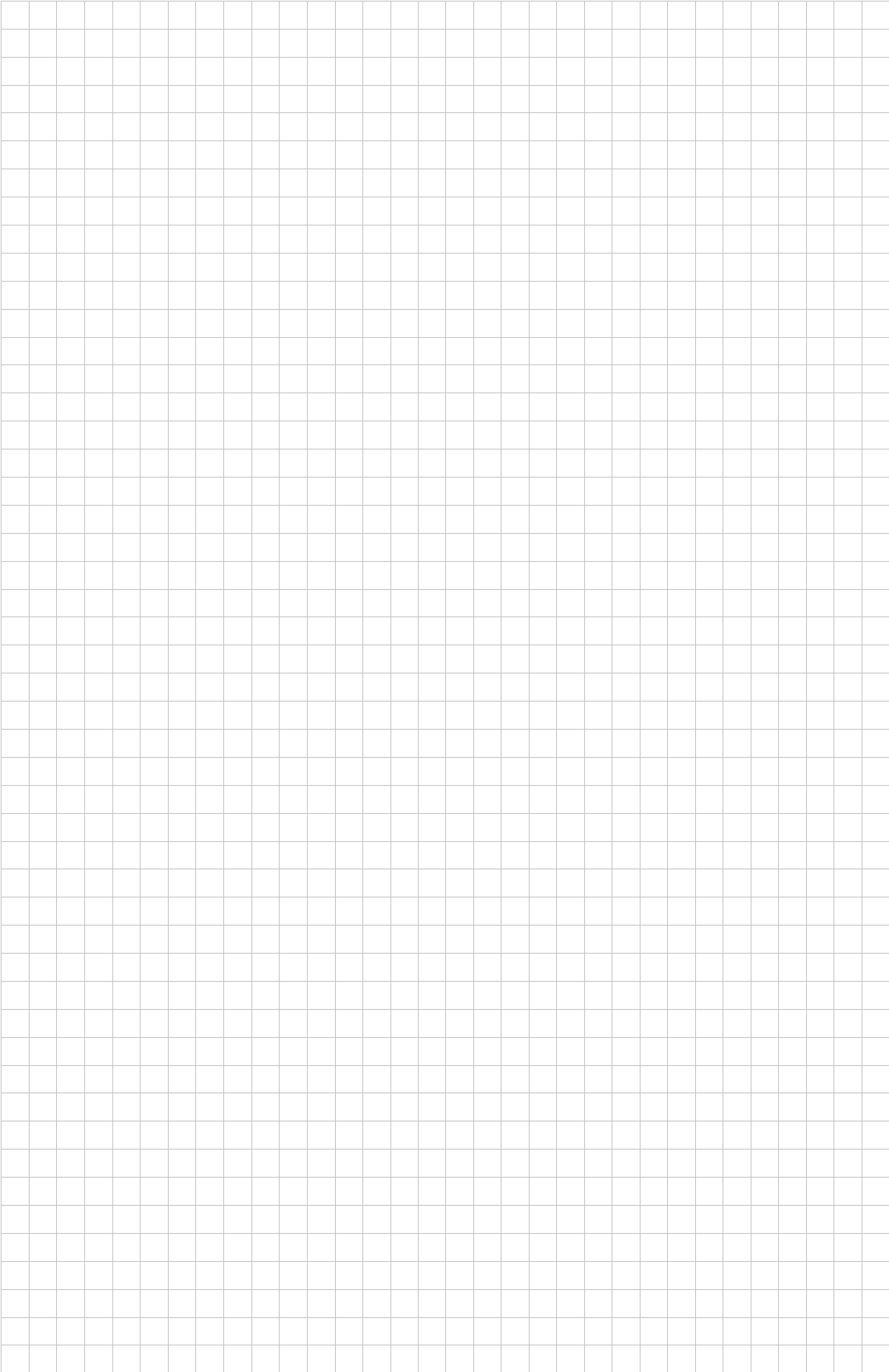
$$=\text{ROUND}(12.65;1)=12.7$$

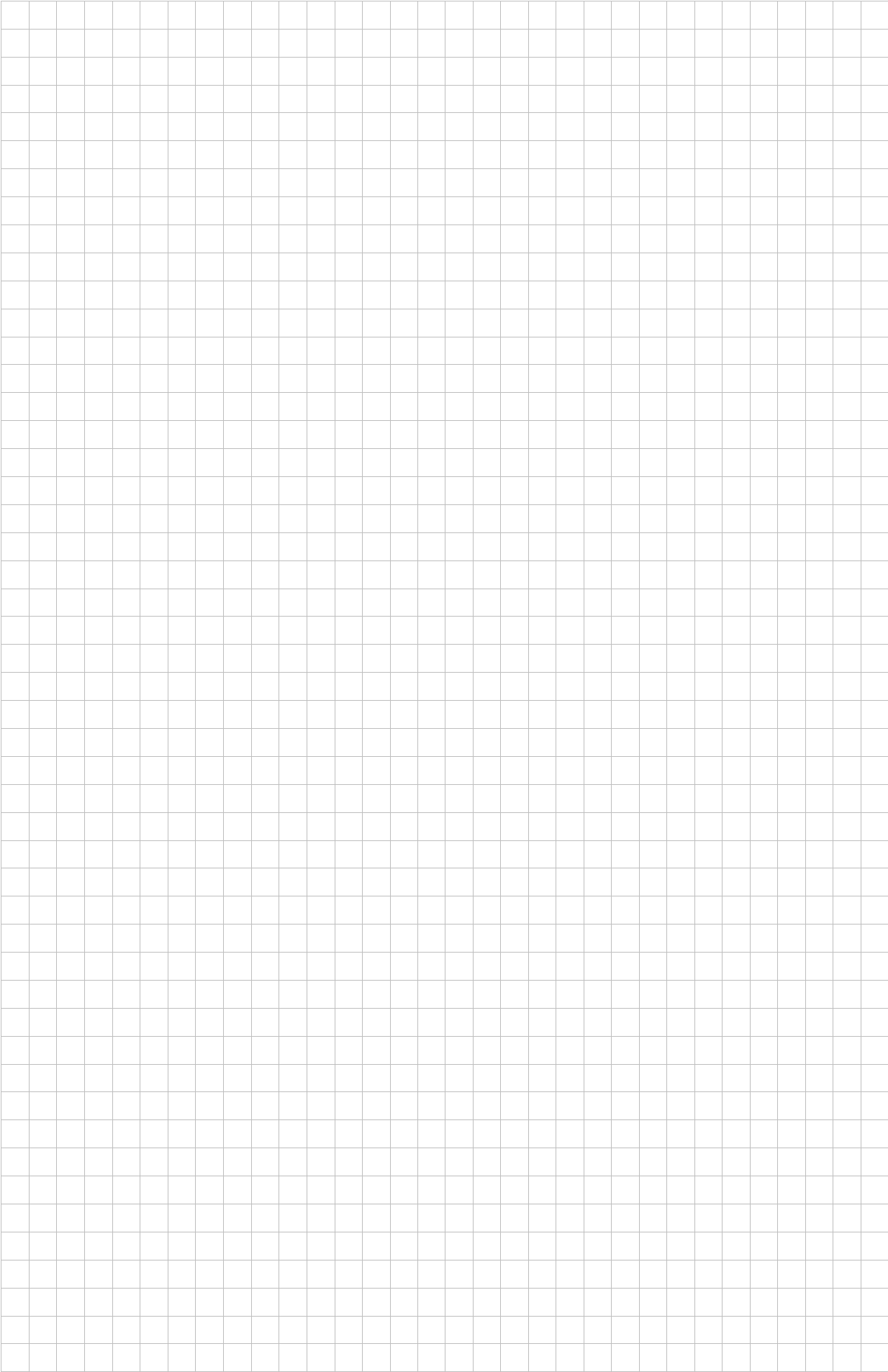
Au fait dans la pratique ces règles sont peu utilisées car les logiciels (tableurs) n'intègrent pas des fonctions adaptées. Il est alors d'usage d'arrondir simplement à la valeur de la décimale la plus proche.

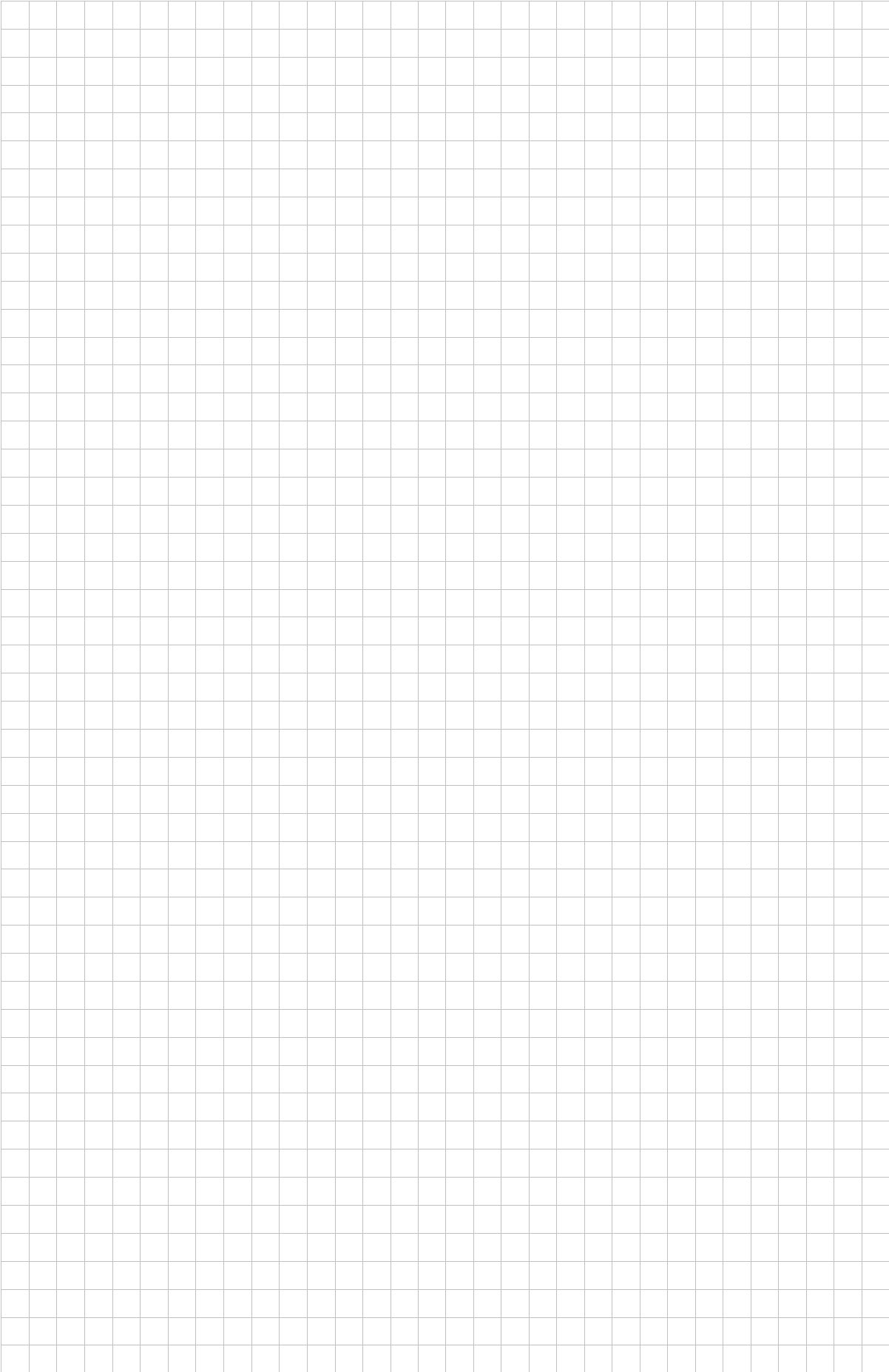
Les chiffres significatifs d'une valeur comprennent tous ses chiffres déterminés avec certitude ainsi que le premier chiffre sur lequel porte l'incertitude (ce dernier significatif occupe le même rang que l'ordre de grandeur de l'incertitude).

Souvent, les sources de données ne mentionnent pas d'intervalle de confiance (c'est-à-dire une indication +/-). Par exemple, lorsque nous écrivons $m = 25.4 \text{ [kg]}$ nous considérons conventionnellement que l'incertitude est du même ordre de grandeur que le rang du dernier chiffre significatif (soit le chiffre incertain).

En fait, seul le rang décimal de l'incertitude est implicite: sa marge réelle n'est pas précisée.









Algèbre

L'algèbre est la science du calcul des grandeurs ou structures représentées par des lettres (Larousse).

8. CALCUL ALGÈBRIQUE

Dans la section d'Arithmétique de ce site, nous avons beaucoup écrit sur différents théorèmes utilisant les nombres abstraits afin de généraliser l'étendue de la validité de ces derniers. Nous avons cependant peu abordé la façon dont nous devons manipuler ces nombres abstraits. C'est ce que nous allons voir maintenant.

Comme vous le savez peut-être déjà, le nombre peut être envisagé en faisant abstraction de la nature des objets qui constituent le groupement qu'il caractérise et ainsi qu'à la façon de codifier (chiffre arabe, romain, ou autre système...). Nous disons alors que le nombre est un "**nombre abstrait**" et lorsque nous manipulons ces types d'objets nous disons que nous faisons du "**calcul algébrique**" ou encore du "**calcul littéral**".

Définition: Le "**calcul littéral**" consiste à calculer avec des variables (c'est-à-dire avec des lettres) comme on le ferait avec des nombres.

Pour les mathématiciens il n'est souvent pas avantageux de travailler avec des valeurs numériques (1,2,3...) car ils représentent uniquement des cas particuliers. Ce que cherchent les physiciens, ingénieurs ainsi que les mathématiciens, ce sont des relations applicables universellement dans un cadre le plus général possible.

Ces nombres abstraits appelés aujourd'hui communément "**variables**" sont très souvent représentés par l'alphabet latin (pour lequel les premières lettres de l'alphabet latin a , b , c , ... désignent souvent les nombres connus, et les dernières x , y , z , ... les nombres inconnus.), l'alphabet grec (aussi beaucoup utilisé pour représenter des opérateurs mathématiques plus ou moins complexes) et l'alphabet hébraïque (dans une moindre mesure)

Bien que ces symboles puissent représenter n'importe quel nombre, il en existe cependant quelques-uns aussi bien en physique ou en mathématique qui peuvent représenter des constantes dites Universelles (vitesse de la lumière c , la constante gravitationnelle G , la valeur π , le nombre d'Euler, ...).

Remarque: Il semblerait que les lettres pour représenter les nombres ont été employées pour la première fois par Viète au milieu du 16ème siècle.

Une variable est donc susceptible de prendre des valeurs numériques différentes. L'ensemble de ces valeurs peut varier suivant le caractère du problème considéré.

Rappels (nous avons déjà défini cela dans le chapitre traitant des Nombres dans la section d'Arithmétique):

R1. Nous appelons "**domaine de définition**" d'une variable, l'ensemble des valeurs numériques qu'elle est susceptible de prendre entre deux bornes.

Soit a et b deux nombres tel que $a < b$. Alors:

R2. Nous appelons "**intervalle fermé d'extrémités a et b** ", l'ensemble de tous les nombres x compris entre ces deux valeurs et nous le désignons de la façon suivante:

$$[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\} \quad (8.1)$$

R3. Nous appelons "**intervalle ouvert d'extrémités a et b** ", l'ensemble de tous les nombres x compris entre ces deux valeurs non comprises et nous le désignons de la façon suivante:

$$]a, b[= \{x \in \mathbb{R} \mid a < x < b\} \quad (8.2)$$

R4. Nous appelons "**intervalle fermé à gauche, ouvert à droite**" la relation suivante:

$$[a, b[= \{x \in \mathbb{R} \mid a \leq x < b\} \quad (8.3)$$

R5. Nous appelons "**intervalle ouvert à gauche, fermé à droite**" la relation suivante:

$$]a, b] = \{x \in \mathbb{R} \mid a < x \leq b\} \quad (8.4)$$

Remarque: Si la variable peut prendre toutes les valeurs négatives et positives possibles nous écrivons dès lors: $] -\infty, +\infty[$ où le symbole " ∞ " signifie "**infini**". Evidemment il peut y avoir des combinaisons d'intervalles ouvert et infini à droite, fermé et limité à gauche et réciproquement.

Définition: Nous appelons "**voisinage de a** ", tout intervalle ouvert de $\mathbb{R}$ contenant a (c'est un concept simple que nous reprendrons pour définir ce qu'est une fonction continue). Ainsi:

$$\forall \varepsilon > 0,]a - \varepsilon, a + \varepsilon[\quad (8.5)$$

est un voisinage de a .

1. ÉQUATIONS ET INÉQUATIONS

L'algèbre élémentaire consiste à partir des définitions de l'addition, soustraction, multiplication, et puissance et de leurs propriétés (associativité, distributivité, commutativité, élément neutre, inverse, ...) - ce qui constitue selon l'ensemble sur lequel nous travaillons un corps ou un groupe commutatif abélien ou non (cf. chapitre Théorie des Ensembles) - à manipuler selon un but fixé des "**équations algébriques**" mettant en relation des variables et constantes.

Nous allons définir de suite après ce qu'est une équation et une inéquation mais nous souhaitons d'abord définir certaines de leurs propriétés:

Soit A et B deux polynômes (ou monômes) quelconques - voir définitions un peu plus loin - les expressions:

$$A > B, A = B, A < B, A \leq B, A \geq B \quad (8.6)$$

Vérifient les propriétés suivantes:

P1. Nous pouvons toujours ajouter ou ôter aux deux membres d'une inéquation ou équation un même polynôme en obtenant une inéquation ou équation équivalente (c'est à dire avec les mêmes solutions ou réductions). Nous disons alors que l'égalité ou l'inégalité restent "vraies" par l'opération d'addition ou de soustraction membre à membre.

P2. Si nous multiplions ou si nous divisons les deux membres d'une équation ou inéquation par un même nombre positif nous obtenons également une inéquation ou équation équivalente (nous avons déjà vu cela). Nous disons alors que l'égalité ou l'inégalité reste "vraie" par l'opération de multiplication ou division membre à membre.

P3. Si nous multiplions ou si nous divisons les deux membres d'une inéquation par un même nombre négatif et si nous inversons le sens de l'inégalité, nous obtenons alors une inéquation ou équation équivalente.

1.1. ÉQUATIONS

Définition: Une "**équation**" est une relation d'égalité entre des valeurs toutes abstraites (autrement dit: deux expressions algébriques) ou non toutes abstraites (dès lors nous parlons d'équations à une inconnue, deux inconnues, trois inconnues, ...) reliées entre elles par des opérateurs divers.

La maîtrise parfaite de l'algèbre élémentaire est fondamentale en physique-mathématique et dans l'industrie!!! Comme il existe une infinité de types d'équations, nous ne les présenterons pas ici. C'est le rôle de l'enseignant/formateur dans les classes d'entraîner le cerveau de son auditoire pendant plusieurs années (2 à 3 ans en moyenne) à résoudre énormément de configurations différentes d'équations algébriques (exposées sous forme de problèmes de tous les jours, géométriques ou purement mathématiques) et ce afin que les élèves manipulent ces dernières sans erreurs en suivant un raisonnement logique et rigoureux (ce n'est qu'en forgeant que l'on devient forgeron...!!!

En d'autres termes: Un professeur/formateur et un établissement ad hoc sont irremplaçables pour acquérir un savoir et avoir un retour d'expérience!!!

Nous avons tenté, ci-dessous, de faire une généralisation simpliste des règles de base de l'algèbre élémentaire. Cette généralisation sera d'autant plus simple à comprendre que le lecteur aura l'habitude de manipuler des quantités abstraites:

Ainsi, soit $a, b, c, d, e, \dots, x, y$ des nombres abstraits pouvant prendre n'importe quelle valeur numérique (nous restons dans le cadre des nombres classiques scolaires et industriels...).

Soit Ξ (la lettre majuscule grecque se prononçant "Xi") représentant un ou plusieurs nombres abstraits (variables) opérants entre eux d'une façon quelconque tel que nous ayons des monômes (un seul nombre abstrait) ou polynômes (poly = plusieurs) algébriques différents distinguables ou non (nous faisons donc ici une sorte d'abstrait de l'abstraction ou si vous préférez une variable de plusieurs variables).

Propriétés (il s'agit plus d'exemples au fait que de propriétés...):

P1. Nous aurons toujours $\Xi = \Xi$ si et seulement si le terme Ξ à gauche de l'égalité représente le même terme Ξ que celui qui est à droite de l'égalité. Si cette condition est satisfaite nous avons alors :

$$\Xi - \Xi = 0 \quad (8.7)$$

Sinon:

$$\Xi - \Xi = \Xi \quad (8.8)$$

où nous excluons donc les cas où tous les Ξ sont identiques entre eux (sinon nous revenons à P1).

P2. Nous avons:

$$\frac{\Xi}{\Xi} = 1 \Leftrightarrow 1 = 1 \Leftrightarrow \Xi = \Xi \quad (8.9)$$

qui vérifie la symbolique de l'équation $\Xi = \Xi$ dans le cas seulement où les éléments Ξ sont identiques entre eux (nous excluons bien évidemment le cas avec dénominateur nul).

Nous avons sinon dans le cas où tous les Ξ sont strictement différents:

$$\frac{\Xi}{\Xi} \neq \frac{\Xi}{\Xi} \quad (8.10)$$

Nous pouvons avoir:

$$\frac{\Xi}{\Xi} = \frac{\Xi}{\Xi} \quad (8.11)$$

dans le cas où une simplification (ou non) des termes contenus dans les Ξ amène à une identité de la relation binaire (non nécessairement égale à l'unité).

P3. Si tous les Ξ sont strictement identiques, alors:

$$\Xi \cdot \Xi \cdot \dots \cdot \Xi = \prod_n \Xi = \Xi^n \quad (8.12)$$

Sinon nous avons:

$$\Xi \cdot \Xi \cdot \dots \cdot \Xi \quad (8.13)$$

qui ne peut s'écrire sous forme condensée simple. Il peut aussi arriver que:

$$\Xi \cdot \Xi \cdot \dots \cdot \Xi = \Xi \quad (8.14)$$

avec le Ξ à droite de l'égalité identique à aucun, un ou encore plusieurs Ξ du membre gauche de l'égalité.

P4. Nous pouvons avoir:

$$\frac{\Xi^n}{\Xi^n} = 1 \quad (8.15)$$

sans que nécessairement les exposants du numérateur ou dénominateur soient égaux (nous excluons le dénominateur nul)

Sinon nous pouvons avoir:

$$\frac{\Xi^{\Xi}}{\Xi^{\Xi}} = \Xi \text{ ou } \frac{\Xi^{\Xi}}{\Xi^{\Xi}} = \Xi^{\Xi-\Xi} \quad (8.16)$$

mais il n'est cependant bien évidemment pas impossible d'avoir quand même $\Xi = 1$ ou $\Xi^{\Xi-\Xi} = 1$ (nous excluons le cas avec dénominateur nul)

P5. Nous avons si tous les Ξ sont strictement identiques aux dénominateurs:

$$\Xi / \Xi / \Xi / \Xi \dots \Xi = \frac{\Xi}{\Xi^n} = \Xi^{1-n} \quad (8.17)$$

Mais... il est également possible que dans l'expression précédente certains Ξ différents s'annulent cependant entre eux dès que leur division mutuelle est égale à l'unité (nous excluons le dénominateur nul).

Si tous les Ξ de la relation précédente sont identiques, la relation est égale à l'unité.

Sinon nous avons:

$$\Xi / \Xi / \Xi / \dots / \Xi = \frac{\Xi}{\Xi^n} = \Xi^{1-n} \quad (8.18)$$

mais il n'est cependant pas impossible d'avoir quand même:

$$\Xi / \Xi / \Xi / \Xi \dots \Xi = \Xi \quad (8.19)$$

avec le Ξ à droite de l'égalité identique à aucun, un ou plusieurs Ξ du membre gauche de l'égalité ou même il est tout à fait possible d'avoir:

$$\Xi / \Xi / \Xi / \Xi \dots \Xi = \Xi^{\Xi} \quad (8.20)$$

P6. Soit $\perp$ représentant indifféremment soit exclusivement l'addition ou exclusivement la soustraction nous avons (au signe près):

$$\perp \Xi \perp \Xi \perp \Xi \perp \dots \perp \Xi = n \cdot \Xi \quad (8.21)$$

si tous les Ξ sont identiques entre eux ou si la combinaison d'un nombre indéterminés de Ξ sont égaux au Ξ présent à droite de l'égalité.

Sinon quoi nous aurons:

$$\Xi \perp \Xi \perp \Xi \perp \dots \perp \Xi = \Xi \quad (8.22)$$

il peut cependant arriver que le Ξ à droite de l'égalité soit identique à aucun, un ou plusieurs Ξ du membre gauche de l'égalité.

Nous pouvons également avoir:

$$\Xi^{\Xi} \cdot \Xi^{\Xi} \cdot \dots \Xi^{\Xi} \cdot \Xi^{\Xi} = \Xi^{\sum \Xi} \quad (8.23)$$

si et seulement si les Ξ sont tous égaux (ou décomposable égaux) et les puissances Ξ non

nécessairement égales.

A partir de la connaissance des ces 7 règles/exemples de base, nous pouvons résoudre, simplifier ou montrer qu'une équation simple possède des solutions ou non par rapport à un problème ou énoncé donné.

Ainsi, soit Ψ une opérande ou une suite d'opérations quelconques sur une ou des abstractions d'abstrait Ξ et parmi tous les Ξ , une (ou plusieurs) dont la ou les valeurs numériques est ou sont inconnues (les autres étant connues). Alors, nous devons pouvoir trouver ou démontrer qu'une équation du type:

$$\Psi\Xi = \Psi\Xi \quad (8.24)$$

possède ou non des solutions.

Dans le cas d'une équation avec la valeur absolue (cf. chapitre Opérateurs Arithmétiques) du type:

$$|\Psi\Xi| = \Psi\Xi \quad (8.25)$$

avec le deuxième membre strictement positif (sinon la relation précédente serait un non sens) cela équivaut bien sûr d'après la définition de la valeur absolue à écrire:

$$\Psi\Xi = \Psi\Xi \text{ et } \Psi\Xi = -\Psi\Xi \quad (8.26)$$

Remarques:

R1. La présence de la valeur absolue dans une équation algébrique dont nous cherchons les solutions double souvent le nombre de solutions.

R2. Une équation est dite "**équation conditionnelle**", s'il y a des nombres dans l'ensemble de définition des expressions qui ne sont pas solutions (ce qui est en fait le cas le plus fréquent). Inversement, si tout nombre de l'ensemble de définition est solution de l'équation alors l'équation est dite "**équation identité**".

Nous pouvons parfois avoir à résoudre (et non à simplifier) un "**système d'équations**". Qu'est-ce que c'est ? : C'est un ensemble d'au moins 2 équations à résoudre (et non à simplifier!). La particularité du système ? : L'ensemble des solutions du système est l'intersection des solutions de toutes les équations à résoudre. Quel est son utilité ? : Elle est sans fin, ces systèmes permettent de résoudre des problèmes faisant intervenir des applications des mathématiques à d'autres domaines. A cause de la variété illimitée des applications, il est difficile d'établir des règles précises pour trouver des solutions. La marche à suivre que voici peut être utile pour autant bien sûr que le problème puisse être formulé sous forme d'équations:

1. Si le problème est posé par écrit, le lire plusieurs fois soigneusement, réfléchir aux faits donnés ainsi qu'à la quantité d'inconnues à trouver (résumer l'énoncé sur une feuille de papier est souvent plus qu'utile pour les gros problèmes!).
2. Choisir une lettre qui représente la quantité inconnue. C'est l'un des pas décisifs dans la recherche de la solution. Des phrases contenant des mots comme: trouver, quoi, combien, où, quand ; devraient vous

renseigner sur la quantité inconnue.

3. Faire éventuellement un dessin (de tête ou sur papier) avec des légendes.
4. Dresser une liste des faits connus et des relations concernant les quantités inconnues. Une relation peut être décrite par une équation dans laquelle apparaissent d'un seul ou des deux côtés du signe égal des énoncés écrits à la place des lettres ou des nombres.
5. Après avoir analysé la liste de l'étape 4, formuler une ou plusieurs équations qui décrivent précisément ce qui est énoncé avec des mots.
6. Résoudre l'équation ou le système d'équation(s) formulée(s) à l'étape 5.
7. Contrôler les solutions obtenues à l'étape 6 en se reportant à l'énoncé de départ du problème. Vérifier que la solution concorde avec les conditions de l'énoncé.

Les méthodes de résolutions des systèmes d'équations sont traitées en détails dans le chapitre de Méthodes Numériques (vous y verrez la méthode) et également dans le chapitre d'Algèbre linéaire de la présente section (vous y comprendrez pourquoi la méthode est telle quelle).

1.2. INÉQUATIONS

Précédemment nous avons vu qu'une équation était une égalité composée de différents calculs avec différents termes (dont au moins une "inconnue" ou un "chiffre abstrait"), et que "résoudre" une équation revenait à calculer la valeur de l'inconnue de l'égalité, alors que la "simplifier" revenait à minimiser mathématiquement le nombre de termes (en factorisant ou autre..) et que développer revenait à mettre à plat tous les termes.

Pourquoi avons-nous besoin de rappeler la définition d'une équation ? Tout simplement parce que pour l'inéquation, c'est le même système. La différence ? Si l'équation est une égalité, l'inéquation est une inégalité: comme l'équation, l'inéquation est composée de différents calculs avec différents termes reliés entre eux par des opérateurs quelconques, dont au moins une inconnue.

Différence entre égalité et inégalité:

- Egalité: Symbolisée par le signe =
- Inégalité: Symbolisée par les relations d'ordre d'égalités strictes et larges $<$, $\leq$, $>$, $\geq$.

Lorsque nous résolvons une inéquation, notre inconnue peut-avoir un intervalle de valeurs qui satisfont à l'inéquation. Nous disons alors que la solution de l'inéquation est un "**ensemble de valeurs**". C'est la différence fondamentale entre une égalité (plusieurs solutions) et une inégalité (intervalle de solutions) !

Rappelons les signes que nous pouvons rencontrer dans une inéquation:

$<$: Se lit "**strictement inférieur à**" ou "**strictement plus petit que**". Dans ce cas, le plus souvent, la valeur butoir numérique n'est pas comprise dans le domaine et nous pouvons représenter alors le domaine avec un crochet ouvert à gauche]... ou à droite ...[selon que la valeur butoir est positive ou négative.

$>$: Se lit "**strictement supérieur à**" ou "**strictement plus grand que**". Dans ce cas, le plus souvent, la

valeur butoir numérique n'est également pas comprise dans le domaine et nous pouvons représenter alors le domaine avec un crochet ouvert à gauche]... ou à droite ...[selon que la valeur butoir est positive ou négative.

Remarque: Attention cependant pour les deux cas précités, il existe des situations où le domaine est imposé par l'ensemble de nombres sur lequel nous travaillons (penser par exemple à une inéquation où pour certaines valeurs les solutions appartiennent à l'ensemble des complexes). Dans ce cas, les valeurs butoirs à l'ensemble de nombres sur lequel nous travaillons peuvent imposer des crochets fermés.

$\leq$: Se lit "**inférieur ou égal à**" ou "**plus petit ou égal à**". Dans ce cas, la valeur butoir numérique est comprise dans le domaine et nous pouvons représenter alors le domaine avec un crochet fermé à gauche [... ou à droite ...] (mais pas nécessairement les deux!) selon que la valeur butoir est positive ou négative.

$\geq$: Se lit "**supérieur ou égal à**" ou "**plus grand ou égal à**". Dans ce cas, la valeur butoir numérique est également comprise dans le domaine et nous pouvons représenter alors le domaine avec un crochet fermé à gauche [... ou à droite ...] (mais pas nécessairement les deux!) selon que la valeur butoir est positive ou négative.

Remarque: Nous renvoyons le lecteur au début de ce chapitre où nous avons défini la manière d'écrire des domaines de définition.

L'objectif des inéquations est la plupart du temps (excepté le côté esthétique) d'avoir au moins parmi l'ensemble des termes une valeur numérique qui permet de définir le domaine de solution (de tous les termes abstraits de l'inéquation) qui satisfait l'inéquation.

Il existe plusieurs façons de représenter les domaines de définition des variables qui satisfont à l'inéquation. Nous allons voir à travers un petit exemple quelles sont ces possibilités:

Soit une inéquation linéaire (du premier degré) en x à une seule inconnue à laquelle nous imposons une contrainte particulière arbitraire pour l'exemple (évidemment l'expression peut contenir plus de termes...):

$$x < 2 \quad (8.27)$$

nous avons dans l'inéquation ci-dessus déjà simplifié tous les termes qui étaient superflus.

Résoudre l'inégalité revient à chercher les valeurs de x inférieures à 2. Bien sûr, il n'existe pas une seule solution dans $\mathbb{R}$ mais un ensemble (intervalle) de solutions et c'est cela même le principe des inéquations!

Pour résoudre l'inéquation, nous observons d'abord le type d'inégalité imposée ("**stricte**" ou "**égal**"). Ensuite, dans les petites classes (et pas seulement parfois...) nous représentons l'ensemble $\mathbb{R}$ traditionnellement par un tableau tel que:

$-\infty$	0	$+\infty$
.....		

Tableau: 8.1 - Résolution d'inéquation

Nous savons intuitivement que la solution de notre inéquation regroupe toutes les valeurs inférieures à 2 (2 exclu des solutions) et ce jusqu'à $-\infty$. Nous écrivons alors cet intervalle ou domaine sous la forme suivante:

$$S =]-\infty, 2[\quad (8.28)$$

Ensuite, nous pouvons représenter graphiquement l'ensemble des solutions (cela aide à comprendre et prépare l'étudiant à la résolution de systèmes d'équations et d'inéquations et aux variations de fonctions). Pour cela, nous reprenons le modèle de schéma du système numérique, et y plaçons notre valeur butoir (nous n'en avons qu'une dans cet exemple mais parfois il peut y en avoir plusieurs dû au fait qu'il y a une singularité ou des racines pour certaines valeurs du domaine de définition), soit 2:

$-\infty$	0	2	$+\infty$
.....			

Tableau: 8.2 - Construction des points particuliers de l'inéquation

et enfin, nous délimitons au stylo de couleur (...) l'ensemble des solutions de $-\infty$ à 2 exclu:

$-\infty$	0	2	$+\infty$
.....		[.....	

Tableau: 8.3 - Mise en place du type de bornes de l'inéquation

A la valeur 2, nous n'oublions pas de marquer le signe[pour montrer que cette valeur est exclue des solutions. Et voilà, le tour est joué et le concept est extrapolable à des inéquations beaucoup plus complexes.

Remarques:

R1. Parfois au lieu de représenter les tableaux comme nous l'avons fait, certains professeurs (c'est un choix complètement artistique) demandent à leur élèves d'hachurer les cases du tableau et d'y dessiner de petits ronds, ou encore se servent de petites flèches, ou encore de dessiner le graphique des fonctions de l'inéquation (cette dernière méthode est certes esthétique mais prend du temps..).

R2. Dans le cadre d'inéquations de degré supérieur à 1, il faut (voir plus loin ce que cela signifie exactement) d'abord déterminer les racines de l'inéquation qui permettent de déterminer les intervalles et ensuite par essais successifs, déterminer quels intervalles sont à rejeter ou à conserver.

Nous pouvons également (au même titre que les équations) parfois avoir à résoudre un "**système d'inéquations**". Qu'est-ce que c'est ? C'est un ensemble d'au moins 2 inéquations à résoudre. La particularité du système ? : L'ensemble des solutions du système est l'intersection des solutions des toutes les inéquations à résoudre.

Autrement dit, la méthode est la même que la précédente, à la différence près que notre tableau

(représentant les domaines de solutions) comportera une ligne supplémentaire par inéquation supplémentaire dans le système plus une ligne de synthèse qui est la projection des domaines de solutions possibles du système.

Ainsi, un système à n inéquations aura un tableau récapitulatif à $n + 1$ lignes.

Mathématiquement, les domaines (car il peut y en avoir plusieurs qui sont disjoints) peuvent s'écrire comme un ensemble de domaines:

$$\mathcal{S} = \{[\dots, \dots], [\dots, \dots], \dots, [\dots, \dots]\} \quad (8.29)$$

Les systèmes d'inéquations sont très fréquents dans beaucoup de problèmes de la mathématique, physique, économétrie, etc... Il est donc important de s'entraîner à les résoudre pendant vos études avec l'aide de votre professeur.

Par exemple, voici une possible représentation du domaine de solutions d'un système d'inéquations pris du chapitre de Méthodes Numériques où nous étudions la "recherche opérationnelle".

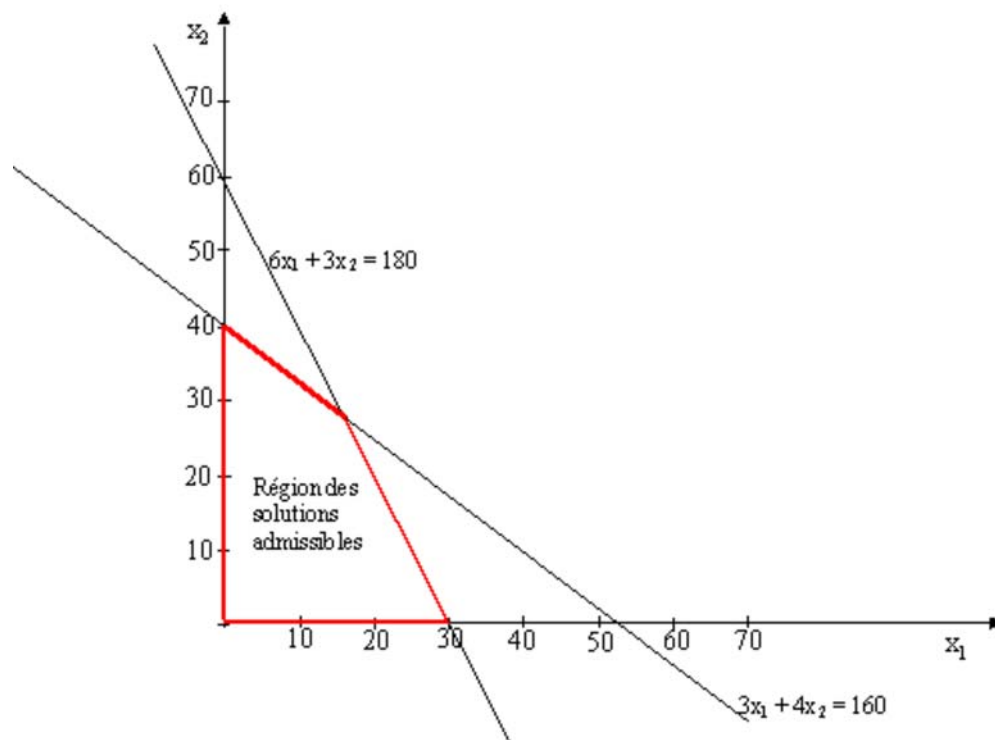


Figure: 8.1 - Représentation graphique plane d'un système d'inéquations

2. IDENTITÉS REMARQUABLES

Les identités remarquables sont des sortes de relations magiques, qui nous servent le plus souvent pour la factorisation ou la résolution d'équations algébriques.

Rappelons certaines notions qui ont déjà été vues dans le chapitre de théorie des ensembles de la section d'arithmétique (nous supposons le concept d'élément neutre connu puisque déjà défini):

Commutativité:

$$a + b = b + a \text{ et } a \cdot b = b \cdot a \quad (8.30)$$

Associativité:

$$a + (b + c) = (a + b) + c \text{ et } a \cdot (b \cdot c) = (a \cdot b) \cdot c \quad (8.31)$$

Distributivité:

$$a \cdot (b + c) = a \cdot b + a \cdot c \quad (8.32)$$

Les mêmes observations sont valables avec l'opération de soustraction bien évidemment dans les domaines de définition adéquats.

Nous pouvons vérifier avec des valeurs numériques (en remplaçant chaque nombre abstrait par un nombre choisi au hasard), ou par développement (ce serait mieux, ainsi vous êtes sûr d'avoir compris ce dont quoi nous parlions), que les identités algébriques suivantes sont vérifiées (ce sont les plus connues):

1. Identité du second degré:

$$(a + b)^2 = a^2 + 2ab + b^2 \quad (8.33)$$

2. Identité du troisième degré:

$$(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3 \quad (8.34)$$

Remarque: Nous pouvons très bien poser que $(a + (c + d))^2 = (a + b)^2$ où nous avons bien évidemment posé que $b = c + d$ (nous faisons un "abstrait d'abstraction" ou plus couramment: un "changement de variable")...:

$$(a + b + c)^2 = a^2 + 2a(b + c) + (b + c)^2 = a^2 + b^2 + c^2 + 2ab + 2bc + 2ac \quad (8.35)$$

Nous pouvons remarquer ainsi qu'en toute généralité, pour calculer le développement de $(a + b)^n$, nous utilisons le développement de $(a + b)^{n-1}$, c'est-à-dire calculé avec la valeur précédente de n .

Nous remarquons les propriétés suivantes pour a et b :

P1. Les puissances de a décroissent de n à 0 ($a^0 = 1$, donc il n'est pas noté dans le dernier terme)

P2. Les puissances de b croissent de 0 à n ($b^0 = 1$, donc il n'est pas noté dans le dernier terme)

P3. Dans chaque terme, la somme des puissances de a et b est égale à n

P4. Les coefficients multiplicateurs devant chaque terme se calculent en faisant la somme des coefficients multiplicateurs de deux termes du développement obtenu avec la valeur précédente de b (voir la figure ci-dessous).

Les coefficients binomiaux peuvent alors être obtenus par construction du "triangle de Pascal" ci-dessous:

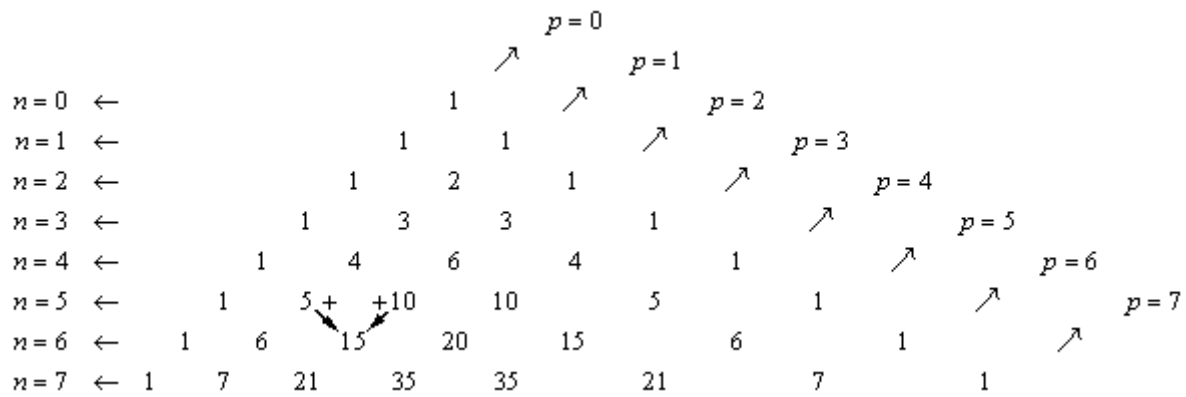


Figure: 8.2 - Construction à la main du triangle de Pascal

Dont chaque élément est donné par (cf. chapitre de Probabilités):

$$C_p^n = \frac{n!}{p!(n-p)!} \quad (8.36)$$

avec $n, p \in \mathbb{N}_+$.

Nous pouvons alors démontrer que:

$$(a+b)^n = \sum_{p=0}^n C_p^n a^{n-p} b^p \quad (8.37)$$

ce qui constitue le fameux "**binôme de Newton**" (que nous réutiliserons à de multiples endroits sur le site) ou appelé aussi "**théorème binomial**".

Démonstration:

Cette relation se démontre simplement par récurrence en supposant la relation précédente vraie et en la calculant pour le rang 1 :

$$(a+b)^1 = a+b = C_0^1 a^1 b^0 + C_1^1 a^0 b^1 = a+b \quad (8.38)$$

Montrons que si elle est vraie pour n alors elle est vraie pour $n+1$:

$$\begin{aligned}
(a+b)^{n+1} &= (a+b)(a+b)^n = (a+b) \sum_{p=0}^n C_p^n a^{n-p} b^p \\
&= \sum_{p=0}^n C_p^n a^{n+1-p} b^p + \sum_{p=0}^n C_p^n a^{n-p} b^{p+1} \\
&= C_0^n a^{n+1} + \sum_{p=1}^n C_p^n a^{n+1-p} b^p + \sum_{p=0}^{n-1} C_p^n a^{n-p} b^{p+1} + C_n^n b^{n+1} \\
&= C_0^n a^{n+1} + \sum_{p=1}^n C_p^n a^{n+1-p} b^p + \sum_{p=1}^n C_{p-1}^n a^{n+1-p} b^p + C_n^n b^{n+1} \quad (8.39) \\
&= C_0^n a^{n+1} + \sum_{p=1}^n (C_p^n + C_{p-1}^n) a^{n+1-p} b^p + C_n^n b^{n+1} \\
&= C_0^n a^{n+1} + \sum_{p=1}^n C_p^{n+1} a^{n+1-p} b^p + C_n^n b^{n+1} \\
&= \sum_{p=0}^{n+1} C_p^{n+1} a^{n+1-p} b^p
\end{aligned}$$

La relation est vraie au rang $n+1$, elle est donc vraie pour tout n .

□C.Q.F.D.

Pour ce qui est des identités remarquables avec des valeurs négatives, il est inutile d'apprendre par cœur l'emplacement du signe "-". Il suffit de faire un changement de variable et une fois le développement fait de refaire le changement de variable dans l'autre sens.

Exemple:

$$\begin{aligned}
(a-b)^2 &\underset{a \rightarrow -b}{=} (a+c)^2 = a^2 + 2ac + c^2 = a^2 + 2a(-b) + (-b)^2 \\
&= a^2 - 2ab + (-b)(-b) = a^2 - 2ab + b^2 \quad (8.40)
\end{aligned}$$

et ainsi de suite pour toute puissance n .

Nous pouvons bien sûr mélanger les genres tels que (fameux exemple particulier):

$$(a+b)(a-b) = a(a-b) + b(a-b) = a^2 - b^2 \quad (8.41)$$

et quelques relations remarquables pratiques supplémentaires qui sont souvent utilisées dans les petites classes pour les exercices:

$$\begin{aligned}
a^3 + b^3 &= (a+b)(a^2 - ab + b^2) \\
a^3 - b^3 &= (a-b)(a^2 + ab + b^2) \\
(x+a)(x+b) &= x^2 + (a+b)x + ab \quad (8.42) \\
(x+a)(x-b) &= x^2 + (a-b)x - ab \\
(x-a)(x-b) &= x^2 - (a+b)x + ab
\end{aligned}$$

et autre cas très fréquent:

$$(ax + b)(cx + d) = acx^2 + (ad + bc)x + bd \quad (8.43)$$

Remarque: Lorsqu'à partir du terme de droite (sous forme numérique simplifiée) le professeur demande à ses élèves en tant qu'exercice d'obtenir la factorisation à gauche de l'égalité, il n'existe pas d'autres moyens que de procéder par essais successifs.

Bien sûr, il y a encore un beaucoup plus grand nombre de relations utiles (dont une partie découle d'une généralisation de celles présentées ci-dessus) que le lecteur découvrira par ses propres raisonnements et en fonction de sa pratique.

Remarque: Il est bien sûr possible de multiplier des polynômes entre eux et de distribuer les termes multiplicatifs. Inversement, il est souvent demandé aux élèves des petites classes de faire la procédure inverse ("factoriser" ou "décomposer" un polynôme) afin qu'ils s'habituent à la manipulation des identités remarquables. Décomposer en un produit de facteurs est une opération importante en mathématiques, puisqu'il est ainsi possible de réduire l'étude d'expressions compliquées à l'étude de plusieurs expressions plus simples.

3. POLYNÔMES

Définition (simpliste): Nous appelons "**polynôme algébrique** $P(x)$ " une fonction de degré $n \in \mathbb{N}$ qui s'écrit:

$$P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x^1 + a_0 \quad (8.44)$$

ou de façon plus condensée par:

$$P_n(x) = \sum_{k=0}^n a_k x^k \quad (8.45)$$

Remarques:

R1. Le n en indice du $P(x)$ est parfois omis car explicitement défini dans l'énoncé.

R2. Le lecteur qui aura parcouru le chapitre de Théorie Des Ensembles, se rappellera certainement que l'ensemble des polynômes de degré n ou inférieurs forment une structure d'espace vectoriel!

Définition (ensembliste): Soit k un anneau (cf. chapitre de Théorie Des Ensembles) et $n \in \mathbb{N}^*$, "**l'anneau des polynômes**" en n indéterminées (ou variables) $k[X_1, \dots, X_n]$ est construit à partir d'un polynôme élémentaire, appelé "**monôme**" de la forme:

$$\lambda X_1^{m_1} \dots X_n^{m_n} \quad (8.46)$$

où $\lambda \in k$ est le "**coefficient du monôme**", $m_1, \dots, m_n$ sont des entiers ≥ 0 et où $X_1^{m_1} \dots X_n^{m_n}$ forme la

"**partie littérale du monôme**". Ainsi, par construction, un polynôme est une somme d'un nombre fini de monômes appelés alors "**termes du polynôme**".

Ainsi, le cas particulier commun utilisé dans les petites classes et présenté au début est $k[X]$, c'est-à-dire l'anneau des polynômes à une variable à coefficients dans k . Tout élément de $k[X]$ s'écrit donc:

$$\sum_{i=0}^n \alpha_i X^i \quad (8.47)$$

avec $\alpha_i \in k, i = 0..n$ et $n \in \mathbb{N}$.

Remarques:

R1. Notez bien que les puissances sont toujours positives (ou nulles) dans $k[X]$!!!

R2. Nous disons que deux monômes sont semblables s'ils ont la même partie littérale.

Définition: Nous nommons "**racine**" ou "**zéro de polynôme**", la ou les valeurs x telles que "**l'équation polynomiale**" $P(x) = 0$ soit satisfaite à la condition qu'au moins un des α_n avec $n > 0$ soit non nul.

Si le polynôme admet une ou plusieurs racines r_n nous pouvons alors factoriser ce dernier sous la forme (nous le démontrerons rigoureusement de manière générale plus loin):

$$P(x) = (x - r_n) \cdot (x - r_{n-1}) \cdot \dots \cdot (x - r_1) \cdot (x - r_0) \quad (8.48)$$

afin que quand x prend la valeur d'une des racines, l'expression ci-dessus soit nulle. C'est ce que nous appelons par convention "factoriser un polynôme".

Les identités algébriques sont des formes particulières de fonctions polynomiales. Considérons une constante c et une variable x et:

$$(c + x)^n = \sum_{k=0}^n C_k^n c^k \cdot x^{n-k} = C_0^n c^0 \cdot x^n + C_1^n c^1 \cdot x^{n-1} + \dots + C_{n-1}^n c^{n-1} \cdot x^1 + C_n^n c^n \quad (8.49)$$

Nous voyons que si nous posons:

$$C_0^n c^0 = a_n, C_1^n c^1 = a_{n-1}, C_{n-1}^n c^{n-1} = a_1, C_n^n c^n = a_0 \quad (8.50)$$

nous retrouvons:

$$(c + x)^n = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x^1 + a_0 \quad (8.51)$$

Définition: Le "**coefficient dominant**" d'un polynôme est le coefficient de son monôme de plus haut degré.

3.1. DIVISION EUCLIDIENNE DES POLYNÔMES

Plaçons nous à présent dans l'anneau $k[X]$. Si $P(X) \in k[X]$, nous notons $\deg(P)$ le degré du polynôme $P(X)$ à coefficients dans un anneau k (les réels ou les complexes... peu importe!)

Remarque: Par convention, $\deg(0) = -\infty$

Soit:

$$u(X) = \sum_{i=0}^k u_i X^i, \quad v(X) = \sum_{i=0}^m v_i X^i \in k[X] \quad (8.52)$$

avec $m > 0$.

Alors il existe deux polynômes uniques $q(X), r(X) \in k[X]$ tels que:

$$\deg(r) < \deg(v) = m \quad (8.53)$$

et:

$$u(X) = q(X) \cdot v(X) + r(X) \quad (8.54)$$

Démonstration:

Si $u(X) = 0$ le résultat est évident. Supposons que $u(X) \neq 0$ et montrons l'existence par récurrence sur le degré k de $u(X)$.

Si $k = 0$ alors $q(X) = 0$ (puisque $m > 0$) et donc $r(X) = u(X)$ fait l'affaire.

Supposons l'affirmation vraie pour tout $k \leq n$:

Soit $u(X)$ de degré $k = n+1$. Si $m > n+1$ alors $q(X) = 0$ et $r(X) = u(X)$ font l'affaire.

Sinon, si $m \leq n+1$ alors en écrivant:

$$u(X) - \frac{u_{n+1}}{v_m} X^{n+1-m} \cdot v(X) \quad (8.55)$$

nous réduisons $u(X)$ à un polynôme de degré $\leq n$ puisque $v(X)$ est de degré m (et qu'il existe)!

Effectivement, le terme:

$$\frac{u_{n+1}}{v_m} X^{n+1-m} \cdot v(X) \quad (8.56)$$

élimine (au moins) le terme de plus grand degré $u_{n+1} X^{n+1}$

Par hypothèse de récurrence, il existe $f(X), g(X)$ tels que:

$$u(X) - \frac{u_{n+1}}{v_m} X^{n+1-m} \cdot v(X) = f(X) \cdot v(X) + g(X) \quad (8.57)$$

avec $\deg(g) < m$. Donc:

$$u(X) = \left(\frac{u_{n+1}}{v_m} X^{n+1-m} + f(X) \right) \cdot v(X) + g(X) \quad (8.58)$$

et:

$$q(X) = \frac{u_{n+1}}{v_m} X^{n+1-m} + f(X), \quad r(X) = g(X) \quad (8.59)$$

font l'affaire.

Donc par récurrence nous observons que la division euclidienne existe dans l'anneau des polynômes $k[X]$.

□C.Q.F.D.

Remarque: Cette démonstration nous a permis dans le chapitre de théorie des ensembles de montrer que cet anneau est "principal".

3.2. THÉORÈME DE FACTORISATION DES POLYNÔMES

Nous allons maintenant démontrer une propriété importante qui est au fait à l'origine illustrée (entre autres) par les identités remarquables que nous avons vues plus haut:

Si une fonction polynôme $P(X) \in k[X]$ à coefficients dans k de degré $n \geq 1$ a une racine $x = r$ dans l'anneau k , alors nous pouvons factoriser $P(x)$ par $(x - r)$ tel que:

$$P(x) = (x - r) Q(x) \quad (8.60)$$

où Q est une fonction polynôme de degré $n-1$ (et peut donc être dans certains cas un simple monôme).

Autrement dit, "**factoriser un polynôme**", c'est l'écrire sous la forme d'un produit de polynômes. La factorisation est donc une opération qui transforme une somme en un produit.

Démonstration:

L'idée consiste à effectuer la division euclidienne de P par $(x-r)$. D'après le théorème précédent, il existerait un couple (Q, R) de polynômes tels que:

$$P(x) = (x - r) Q + R \quad (8.61)$$

et selon le résultat obtenu du théorème précédent sur la division euclidienne:

$$\deg(R) < \deg(x - r) \quad (8.62)$$

Or, $\deg(x-r) = 1$, donc $\deg(R) = 0$ (ou $-\infty$ par convention). R est donc une fonction polynôme constante. Par ailleurs, par hypothèse, r est une racine de P . Nous avons donc:

$$0 = P(r) = (r-r)Q(r) + R(r) \quad (8.63)$$

Donc $R(r) = 0$. Donc R est la fonction polynôme nulle et le théorème est pratiquement démontré. Il reste encore à prouver que $\deg(Q) = n-1$, ce qui est une conséquence immédiate de la relation:

$$\deg(P) = \deg(x-r) + \deg(Q) \quad (8.64)$$

D'où:

$$n = 1 + \deg(Q) \quad (8.65)$$

□C.Q.F.D.

De cette propriété de factoriser un polynôme vue précédemment, appelée "**théorème de factorisation**", nous pouvons donner un avant-goût d'un théorème beaucoup plus important:

Montrons que si nous avons une fonction polynôme $P(X) \in k[X]$ de degré $n \in \mathbb{N}$ à coefficients dans k , alors elle possède au plus un nombre fini n de racines (certaines étant éventuellement confondues) dans k .

Démonstration:

D'abord, puisque P a un degré, P n'est pas la fonction polynôme nulle. Ensuite, raisonnons par l'absurde:

Si la fonction P possède p racines avec $p > n$, en notant $r_1, \dots, r_p$ ces racines, nous avons, d'après le théorème de factorisation précédent (appliqué p fois):

$$P(x) = (x-r_1)(x-r_2)\dots(x-r_p)Q(x) \quad (8.66)$$

où Q est donc une fonction polynôme de degré:

$$n - p < 0 \quad (8.67)$$

Or, comme par définition un polynôme en est un si seulement son degré appartient à $\mathbb{N}$, le polynôme Q doit donc être le polynôme nul tel que:

$$Q(x) = 0 \quad (8.68)$$

Il s'ensuit que:

$$P(x) = 0 \quad (8.69)$$

ce qui contredit l'hypothèse initiale comme quoi P n'est la fonction polynôme nulle d'où:

$$p \leq n \quad (8.70)$$

3.3. ÉQUATIONS DIOPHANTIENNES

Si nous généralisons le concept de polynôme avec plusieurs variables tel que:

$$P = P(x, y, z) \quad (8.71)$$

nous appelons alors "équation diophantienne" une équation de la forme:

$$P(x, y, z) = 0 \quad (8.72)$$

où P est un polynôme à coefficients entiers (ou rationnels) dont nous cherchons les radicaux strictement dans $\mathbb{N}$ ou $\mathbb{Q}$. Des exemples classiques d'équations diophantiennes sont:

- Les triplets pythagoriciens (ou triades) tel que:

$$a^2 + b^2 = c^2 \quad (8.73)$$

- Le grand théorème de Fermat dont la conjecture dit que si n est supérieur à 2, il n'existe pas d'entiers x, y, z non nuls pour lesquels:

$$x^n + y^n = z^n \quad (8.74)$$

Pour la démonstration il faudra attendre un peu que les auteurs du site aient le temps de la comprendre également (...).

3.4. POLYNÔMES DE DEGRÉ 1

Soit:

$$P(x) = ax + b = a \left(x + \frac{b}{a} \right) \quad (8.75)$$

Si $a \neq 0$ alors le polynôme admet une unique racine simple:

$$r = -\frac{b}{a} \quad (8.76)$$

tel que $P(r) = 0$.

Remarques:

R1. Il faut toujours prendre l'habitude de vérifier l'existence de la solution dans l'équation d'origine pour s'assurer de la validation du domaine de définition de la solution. Effectivement, il existe des solutions aux développements de résolution d'une équation qui ne vérifient pas l'équation d'origine et c'est ce que nous nommons des "**solutions étrangères**" ou encore "**racines étrangères**".

R2. Si les coefficients du polynôme de degré 1 sont tous réels alors la racine est réelle.

R3. Si un des coefficients est complexe alors la racine est nécessairement complexe.

R4. Si les deux coefficients sont complexes, alors la racine est soit complexe soit réelle.

R5. Nous disons que deux équations sont équivalentes si elles admettent le même ensemble de solutions.

Voici quelques propriétés que nous considérons comme triviales et que nous admettrons donc sans démonstrations:

P1. Si nous ajoutons (ou si nous retranchons) un même nombre à chaque membre d'une équation, nous obtenons une équation qui a les mêmes solutions que l'équation dont nous sommes partis (et ce quel que soit son degré!).

P2. Si nous multiplions (ou si nous divisons) chaque membre d'une équation par un même nombre non nul, nous obtenons une équation qui a les mêmes solutions que l'équation dont nous sommes partis (et ce quel que soit son degré!).

3.5. POLYNÔMES DE DEGRÉ 2

Soit le polynôme à coefficients réels (trinôme du second degré):

$$P(x) = ax^2 + bx + c = a \left[\left(x + \frac{b}{2a} \right)^2 - \frac{b^2 - 4ac}{4a^2} \right] \quad (8.77)$$

Si nous représentons ce polynôme graphiquement sur la plan réel, cela donne:

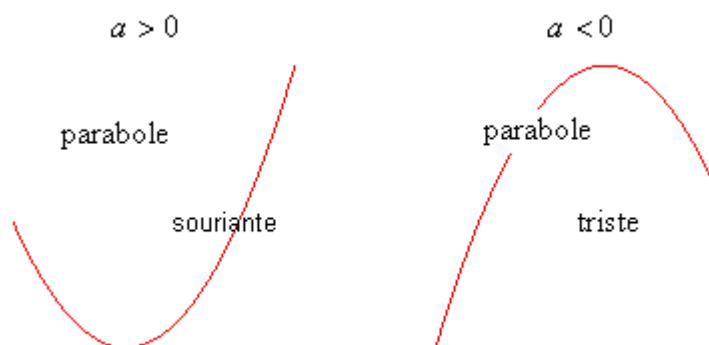


Figure: 8.3 - Représentation des polynômes en fonction du signe du terme de degré 2

Si nous dérivons cette fonction (cf. chapitre de Calcul Différentiel et Intégral) et cherchons en quel point la tangente s'annule, nous la trouvons toujours sur le point d'inflexion de la parabole (qui correspond aussi à son axe de symétrie):

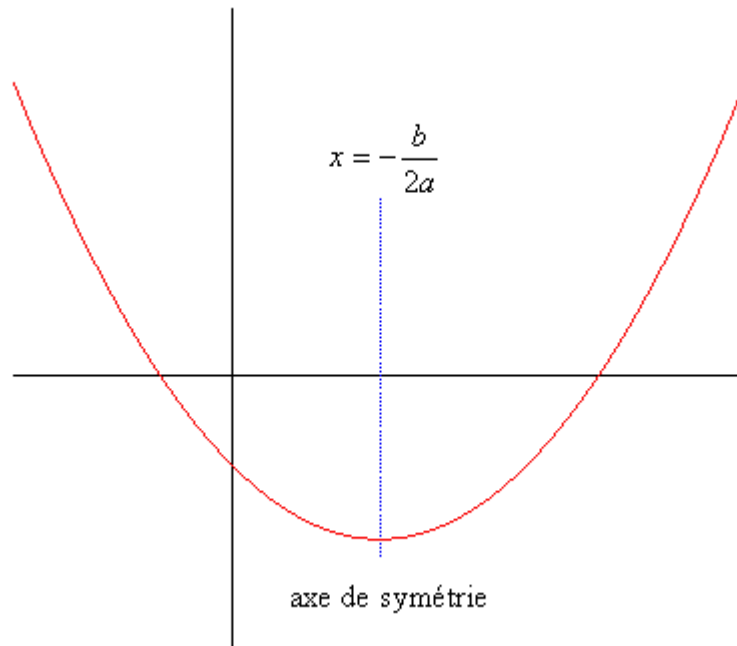


Figure: 8.4 - Point d'inflexion de la tangente

Si $a \neq 0$ alors nous avons:

$$\left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2} \Rightarrow x + \frac{b}{2a} = \pm \frac{\sqrt{b^2 - 4ac}}{2a} \Rightarrow x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad (8.78)$$

Nous avons alors une "**racine double**" (ou "**racine de multiplicité 2**") que nous notons:

$$r_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad (8.79)$$

tel que $P(r) = 0$ et où nous définissons un nouveau terme appelé "**déterminant du polynôme**" ou "**discriminant**" qui allège souvent les écritures:

$$\Delta = b^2 - 4ac \quad (8.80)$$

Remarque: Il faut aussi toujours prendre l'habitude de vérifier l'existence de la solution dans l'équation d'origine pour s'assurer de la validation du domaine de définition de la solution au cas où la solution serait "étrangère".

Si le polynôme du deuxième degré en x comporte deux racines, nous pouvons alors factoriser de manière irréductible (selon le théorème fondamental de factorisation des polynômes vu plus haut) de la manière suivante:

$$P(x) = a(x - r_1)(x - r_2) \quad (8.81)$$

Nous démontrons, à partir de l'expression des racines, sans trop de peine les relations dites "**relations de Viète**":

$$r_1 + r_2 = -\frac{b}{a} \text{ et } r_1 \cdot r_2 = \frac{c}{a} \quad (8.82)$$

Avec le signe de a et celui du discriminant Δ nous avons:

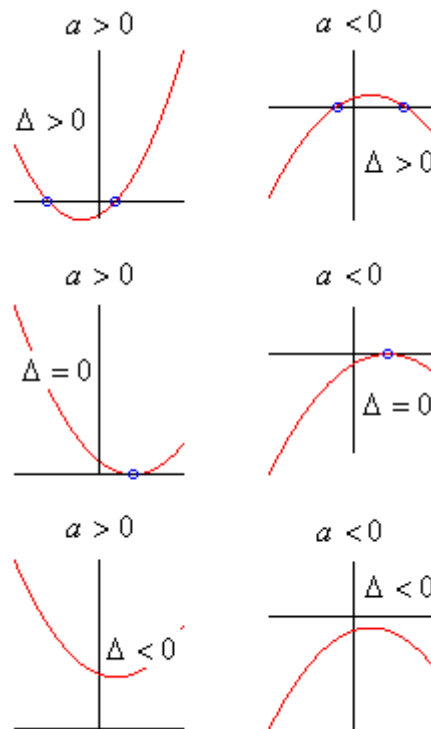


Figure: 8.5 - Caractéristique graphiques en fonction de la valeur du discriminant

Donc:

- Si $\Delta < 0$ le polynôme n'admet pas de zéros réels et ne se décompose pas en un produit de facteurs réels du premier degré mais de facteurs complexes. Ainsi (il est nécessaire d'avoir lu la partie traitant des nombres complexes dans le chapitre des Nombres de la section d'Arithmétique du site):

$$r_{1,2} = \frac{-b \pm \sqrt{(-1)(b^2 - 4ac)}}{2a} = \frac{-b \pm \sqrt{i^2(b^2 - 4ac)}}{2a} = \frac{-b \pm i\sqrt{(b^2 - 4ac)}}{2a} \quad (8.83)$$

et nous savons que nous pouvons écrire tout nombre complexe sous une forme condensée (formule d'Euler) et comme les racines complexes d'un polynôme du second degré sont conjuguées (nous connaissons ce terme) nous avons:

$$r_{1,2} = r e^{\pm i(\varphi + 2k\pi)} \quad (8.84)$$

où (rappel) r est le module des racines complexes (module égal pour les deux racines) et φ l'argument des racines complexes (égales en valeur absolue).

- Si $\Delta = 0$ alors le polynôme possède une seule solution qui est bien évidemment:

$$r_{1,2} = -\frac{b}{2a} \quad (8.85)$$

- Si $\Delta > 0$ alors le polynôme possède deux solutions définies par les relations générales que nous avons déjà données précédemment.

En ce qui concerne le cas complexe, prenons comme exemple le polynôme suivant du second degré:

$$x^2 + 1 = 0 \quad (8.86)$$

qui admet donc uniquement deux racines complexes qui sont i et $-i$. Dans le plan réel ce polynôme sera représenté avec Maple par:

```
>plot(x^2+1,x=-5..5);
```

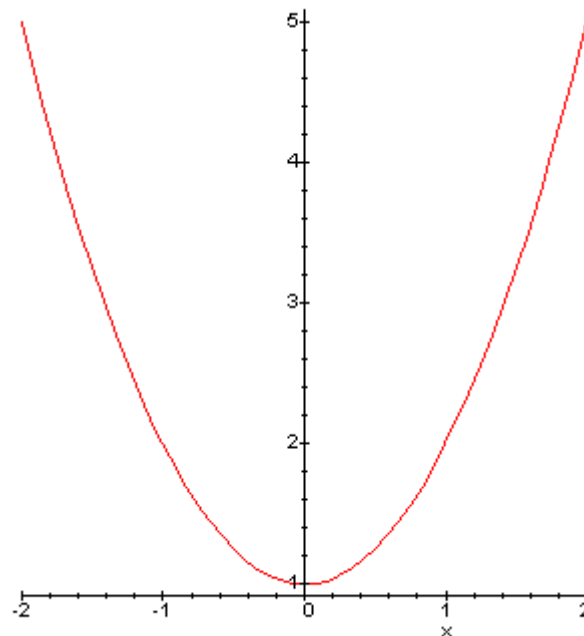


Figure: 8.6 - Exemple de tracé d'un polynôme de degré 2 qui admet que des solutions complexes

où nous voyons bien qu'il n'y a aucune solution (zéros) réelle. Alors qu'en nous plaçant dans les complexes, nous avons:

```
>plot3d(abs(-(re+I*im)^2+1),re=-2..2,im=-2..2,view=[-2..2,-2..2,-2..2],orientation=[-130,70],contours=50,style=PATCHCONTOUR,axes=frame,grid=[100,100],numpoints=10000);
```

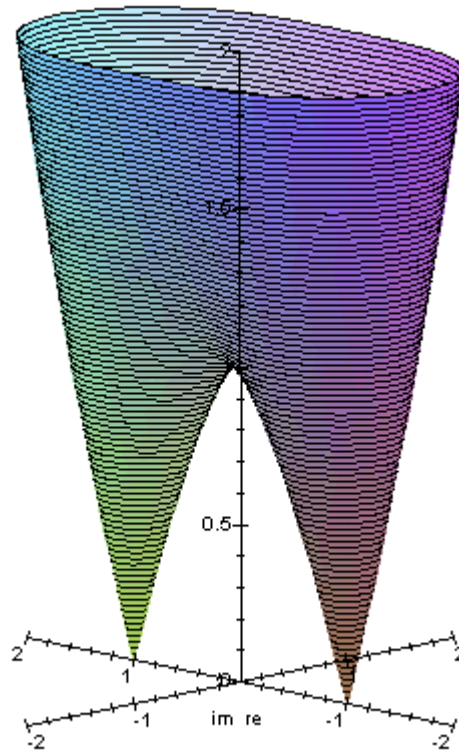


Figure: 8.7 - Le même polynôme mais en jouant avec la représentation

où les deux zéros sont bien visibles sur l'axe imaginaire en -1 et +1. Evidemment quand c'est la première fois que l'on voit une fonction représentée sur une figure en prenant en compte les valeurs complexes on essaie d'y retrouver la parabole correspondante au cas purement réel. Pour cela, il suffit de couper la surface ci-dessus en deux sur l'axe imaginaire et nous avons alors:

```
>plot3d(abs((re+I*im)^2+1),re=-2..2,im=0..2,view=[-2..2,-2..2,0..2],orientation=[-130,70],contours=50,style=PATCHCONTOUR,axes=frame,grid=[100,100],numpoints=10000);
```

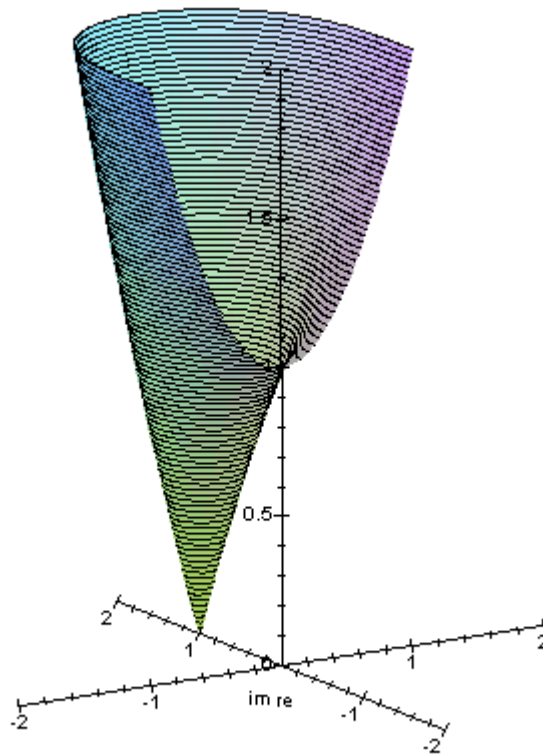


Figure: 8.8 - Un petit zoom toujours sur le même polynôme

où nous retrouvons notre parabole bien visible sur la coupe de la surface. Ainsi, nous pouvons nous demander si les valeurs complexes ne sont pas une extension naturelle de notre espace conventionnel échappant à notre sens physique commun et nos appareils de mesures.

Evidemment de ce qui a été vu jusqu'à maintenant nous en tirons que si un polynôme admet une ou plusieurs racines alors ce même polynôme est divisible par $(x - r_n)$.

3.5.1. NOMBRE D'OR

Il existe un polynôme de degré deux dont la solution est fameuse de par le monde. Ce nombre est appelé la "**divine proportion**" ou "**nombre d'or**" et se retrouve en architecture, esthétique ou encore en phyllotaxie (c'est-à-dire dans la disposition des feuilles autour de la tige des plantes).

Ce nombre vaut:

$$\frac{1+\sqrt{5}}{2} = 1.61803399 \quad (8.87)$$

et appartient à l'ensemble des nombres irrationnels car il ne peut pas s'écrire sous forme de fraction entière, mais c'est un nombre algébrique puisqu'il est la solution positive de l'équation:

$$x^2 - x - 1 = 0 \quad (8.88)$$

3.6. POLYNÔMES DE DEGRÉ 3

Bien que rare à résoudre en physique théorique ou lors de ses études, la résolution d'un polynôme du

3^{ème} degré est assez récréative et montre un bon exemple d'un raisonnement mathématique déjà mature (nous devons ces développements à Scipione del Ferro et Jérôme Cardan mathématiciens du 16^{ème} siècle...).

Soit l'équation:

$$ax^3 + bx^2 + cx + d = 0 \quad (8.89)$$

avec les coefficients tous dans $\mathbb{R}$ (pour commencer...). Dans un premier temps, le lecteur pourra voir que les raisonnements que nous avons appliqués pour les polynômes de degrés inférieurs coïncident rapidement (excepté pour des cas particuliers simplistes bien sûr...).

Nous allons contourner le problème par des changements de variables subtils mais tout à fait justifiés.

Ainsi, rien ne nous empêche de poser que:

$$\exists X \Rightarrow x = X - \frac{b}{3a} \quad (8.90)$$

et qu'en divisant le polynôme de degré 3 par a d'écrire:

$$\begin{aligned} x^3 + \frac{b}{a}x^2 + \frac{c}{a}x + \frac{d}{a} &= 0 \\ \left(X - \frac{b}{3a}\right)^3 + \frac{b}{a}\left(X - \frac{b}{3a}\right)^2 + \frac{c}{a}\left(X - \frac{b}{3a}\right) + \frac{d}{a} &= 0 \\ \left(X^3 - 3X^2\frac{b}{3a} + 3X\frac{b^2}{3^2a^2} - \frac{b^3}{27a^3}\right) + \frac{b}{a}\left(X^2 - 2X\frac{b}{3a} + \frac{b^2}{9a^2}\right) + \frac{c}{a}\left(X - \frac{b}{3a}\right) + \frac{d}{a} &= 0 \end{aligned} \quad (8.91)$$

En regroupant les termes de même ordre:

$$\begin{aligned} X^3 + X^2\left(-\frac{b}{a} + \frac{b}{a}\right) + X\left(\frac{b^2}{3a^2} - 2\frac{b^2}{3a^2} + \frac{c}{a}\right) + \left(-\frac{b^3}{27a^3} + \frac{b^3}{9a^3} - \frac{cb}{3a^2} + \frac{d}{a}\right) &= 0 \\ X^3 + X\left(\frac{c}{a} - \frac{b^2}{3a^2}\right) + \left(\frac{2b^3}{27a^3} - \frac{cb}{3a^2} + \frac{d}{a}\right) &= 0 \end{aligned} \quad (8.92)$$

et posons (rien, mais alors absolument rien ne nous l'interdit):

$$\begin{cases} (1) \ x = X - \frac{b}{3a} \\ (2) \ p = \frac{c}{a} - \frac{b^2}{3a^2} \\ (3) \ q = \frac{2b^3}{27a^3} - \frac{cb}{3a^2} + \frac{d}{a} \end{cases} \quad (8.93)$$

où (1) est connu si et seulement si X est connu et où p, q sont de toute façon connus.

Le polynôme:

$$P = X^3 + pX + q \quad (8.94)$$

étant de degré impair, il admet- comme permet de le constater tout tracé visuel d'un tel polynôme à coefficients réels - au moins une racine réelle, appelée "**racine certaine**" (vérifiez! Vous verrez bien par une représentation graphique d'un polynôme de degré impair que cela est trivial).

Maintenant, nous faisons un autre changement de variable (nous en avons tout à fait le droit) subtil:

$$X = u + v \quad (8.95)$$

en imposant la condition que u, v doivent être tels que $3uv = -p$ (rien ne nous empêche d'imposer une telle contrainte) et nous avons alors:

$$\begin{aligned} P = X^3 + pX + q &= (u + v)^3 - 3uv(u + v) + q = 0 \\ \Rightarrow u^3 + v^3 &= -q \end{aligned} \quad (8.96)$$

Dès lors nous avons:

$$\begin{cases} (1') u^3 + v^3 = -q \\ (2') u^3 v^3 = \frac{-p^3}{27} \Leftrightarrow 3uv = -p \end{cases} \quad (8.97)$$

Nous pouvons très bien faire une analogie entre les deux relations (1') et (2') et les relations de Viète que nous avons obtenues pour le polynôme de degré 2 qui rappelons-le étaient:

$$r_1 + r_2 = -\frac{b}{a} \quad \text{et} \quad r_1 \cdot r_2 = \frac{c}{a} \quad (8.98)$$

à la différence que nous avons maintenant (nous adoptons une autre notation pour ces racines intermédiaires):

$$z_1 = u^3 \quad \text{et} \quad z_2 = v^3 \quad (8.99)$$

ce qui nous donne pour le polynôme P en imposant (toujours par analogie) $a = 1$ une nouvelle équation:

$$Z^2 + qZ - \frac{p^3}{27} = 0 \quad (8.100)$$

dont z_1, z_2 sont les racines.

Cette dernière équation a pour discriminant:

$$\Delta = q^2 + 4 \frac{p^3}{27} \quad (8.101)$$

Prenons maintenant le cas par cas:

- Si $\Delta > 0$, l'équation en Z admet deux solutions z_1, z_2 dont la somme va nous donner indirectement la

valeur de X puisque par définition $X = u + v$ et $z_1 = u^3$ et $z_2 = v^3$. Nous voyons que nous avons tous les ingrédients pour trouver la première racine de l'équation initiale qui sera la racine certaine (ou "zéro certain"). Ainsi:

$$X_1 = u + v = \sqrt[3]{z_1} + \sqrt[3]{z_2} = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} \quad (8.102)$$

comme $\Delta > 0$ et que les racines supérieures sont cubiques nous avons nécessairement $X_1 \in \mathbb{R}$ si tous les coefficients de l'équation originale sont bien dans $\mathbb{R}$.

- Si $\Delta = 0$, nous le savons, l'équation en Z admet une racine double et puisque le discriminant comporte une puissance carrée de q cela signifie nécessairement que p est négatif.

Le polynôme P admet donc lui aussi une racine double et de même pour l'équation d'origine. Nous avons vu par ailleurs que pour un polynôme du second degré si le discriminant est nul les racines sont:

$$r_{1,2} = -\frac{b}{2a} \quad (8.103)$$

alors par analogie:

$$X_1 = X_2 = u + v = \sqrt[3]{z_1} + \sqrt[3]{z_2} = 2\sqrt[3]{z_{1,2}} = 2 \cdot \sqrt[3]{\frac{-q}{2}} \quad (8.104)$$

- Si $\Delta < 0$ nous devons à nouveau utiliser les nombres complexes comme nous l'avons fait lors de notre étude du polynôme de degré 2. Ainsi, nous savons que l'équation en Z admet deux solutions complexes telles que:

$$z_{1,2} = -\frac{q}{2} \pm i \sqrt{-\frac{q^2}{4} - \frac{p^3}{27}} \quad (8.105)$$

et à nouveau comme les racines sont conjuguées nous pouvons écrire sous la forme condensée:

$$z_{1,2} = r e^{\pm i(\varphi + 2k\pi)} \quad (8.106)$$

et comme:

$$X = u + v = \sqrt[3]{z_1} + \sqrt[3]{z_2} \quad (8.107)$$

nous avons donc:

$$X_k = u_k + v_k = u_k + \bar{u}_k = \sqrt[3]{r e^{i(\varphi + 2k\pi)}} + \sqrt[3]{r e^{-i(\varphi + 2k\pi)}} = \sqrt[3]{r} e^{\frac{i(\varphi + 2k\pi)}{3}} + \sqrt[3]{r} e^{-\frac{i(\varphi + 2k\pi)}{3}} \quad (8.108)$$

Comme u_k, v_k sont conjugués, nous avons nécessairement $X_k \in \mathbb{R}$.

3.7. POLYNÔMES DE DEGRÉ 4

L'équation polynomiale à résoudre ici est:

$$ax^4 + bx^3 + cx^2 + dx + e = 0 \quad (8.109)$$

avec $a \neq 0$.

Remarque: Nous devons cette méthode de résolution à l'italien Ludovico Ferrari mathématicien italien du 16ème siècle également.

Quitte à diviser par a nous avons:

$$x^4 + b'x^3 + c'x^2 + d'x + e' = 0 \quad (8.110)$$

Puis, en posant:

$$x = y - \frac{b'}{4} \quad (8.111)$$

l'équation se réduit:

$$\begin{aligned} & \left(y - \frac{b'}{4}\right)^4 + b'\left(y - \frac{b'}{4}\right)^3 + c'\left(y - \frac{b'}{4}\right)^2 + d'\left(y - \frac{b'}{4}\right) + e' \\ &= \left(y^4 - 4y^3\frac{b'}{4} + 6y^2\frac{b'^2}{4^2} - 4y\frac{b'^3}{4^3} + \frac{b'^4}{4^4}\right) + b'\left(y^3 - 3y^2\frac{b'}{4} + 3y\frac{b'^2}{4^2} - \frac{b'^3}{4^3}\right) \\ &+ c'\left(y^2 - 2y\frac{b'}{4} + \frac{b'^2}{4^2}\right) + d'\left(y - \frac{b'}{4}\right) + e' \quad (8.112) \\ &= y^4 + (b' - b')y^3 + \left(-\frac{3}{4}b'^2 + c' + \frac{3}{8}b'^2\right)y^2 + \left(\frac{3}{16}b'^3 - \frac{1}{2}cb' - \frac{1}{16}b'^3 + d'\right)y \\ &+ \left(\frac{b'^4}{256} - \frac{1}{4}d'b' - \frac{1}{64}b'^4 + \frac{1}{16}cb'^2 + e'\right) \end{aligned}$$

où nous voyons que le coefficient devant y^3 s'annule. Ainsi, tout polynôme du type:

$$P(x) = x^4 + bx^3 + cx^2 + dx + e \quad (8.113)$$

peut être écrit sous la forme suivante:

$$y^4 + c''y^2 + d''y + e'' = 0 \quad (8.114)$$

En posant:

$$\begin{aligned}
 x &= y - \frac{b'}{4} \\
 c'' &= \left(-\frac{3}{4}b'^2 + c' + \frac{3}{8}b'^2 \right) \\
 d'' &= \left(\frac{3}{16}b'^3 - \frac{1}{2}c'b' - \frac{1}{16}b'^3 + d' \right) \\
 e'' &= \left(\frac{b'^4}{256} - \frac{1}{4}d'b' - \frac{1}{64}b'^4 + \frac{1}{16}c'b'^2 + e' \right)
 \end{aligned} \quad (8.115)$$

Remarque: Si $d'' = 0$, l'équation à résoudre est en réalité une "équation bicarrée". Le changement de variable $X = y^2$ permet alors de se ramener à une équation polynomiale du deuxième degré (ce que nous savons facilement résoudre).

Nous introduisons maintenant un paramètre t (que nous choisirons judicieusement par la suite) et nous réécrivons l'équation polynomiale sous la forme suivante:

$$(x^2 + t)^2 - [(2t - c'')x^2 - d''x + (t^2 - e'')] = 0 \quad (8.116)$$

Remarque: Si le lecteur développe et distribue tous les termes de la relation précédente il retombera bien évidemment sur $x^4 + c''x^2 + d''x + e'' = 0$.

L'idée sous-jacente est d'essayer de faire en sorte que la partie entre crochets de l'expression précédente puisse s'écrire comme un carré tel que:

$$(2t - c'')x^2 - d''x + (t^2 - e'') = (x - \lambda)^2 \quad (8.117)$$

Car dans ce cas, en utilisant:

$$a^2 - b^2 = (a + b)(a - b) \quad (8.118)$$

Notre équation polynomiale peut alors s'écrire:

$$(x^2 + t)^2 - (x - \lambda)^2 = ((x^2 + t) + (x - \lambda)) \cdot ((x^2 + t) - (x - \lambda)) = 0 \quad (8.119)$$

et nous n'aurions plus qu'à résoudre deux équations polynomiales du deuxième degré (ce que nous savons déjà faire).

Or, pour que nous puissions écrire:

$$(2t - c'')x^2 - d''x + (t^2 - e'') = (x - \lambda)^2 \quad (8.120)$$

il faudrait que l'expression du deuxième degré à gauche de l'égalité n'ait qu'une seule racine. Or, nous avons vu dans notre étude des équations polynomiales du deuxième degré que cela signifiait dès lors

que le déterminant est nul:

$$\Delta = b^2 - 4ac = 0 \quad (8.121)$$

et que la racine s'exprimait par:

$$x = -\frac{b}{2a} \quad (8.122)$$

Ce qui correspond dans notre cas:

$$\Delta = d''^2 - 4(2t - c'')(t^2 - e'') = 0 \quad (8.123)$$

et donc que:

$$4(2t - c'')(t^2 - e'') = d''^2 \quad (8.124)$$

avec:

$$\lambda = -\frac{(-d'')}{2(2t - c'')} = \frac{d''}{4t - 2c''} \quad (8.125)$$

Donc finalement, si t est tel que $4(2t - c'')(t^2 - e'') = d''^2$, alors nous avons:

$$(2t - c'')x^2 - d''x + (t^2 - e'') = (2t - c'')(x - \lambda)(x - \lambda) = (2t - c'')\left(x - \frac{d''}{4t - 2c''}\right)^2 \quad (8.126)$$

puisque le théorème fondamental des polynômes nous donne pour un polynôme du deuxième degré n'ayant qu'une seule racine:

$$P(x) = ax^2 + bx + c = a(x - r_1)(x - r_2) = a(x - r_1)^2 \quad (8.127)$$

Pour conclure, il suffit de voir que trouver un nombre t vérifiant la relation:

$$4(2t - c'')(t^2 - e'') = d''^2 \quad (8.128)$$

est un problème de degré 3 que nous savons déjà résoudre par la méthode de Cardan.

De telles méthodes générales n'existent plus pour les degrés égaux ou supérieurs à 5 comme nous le verrons à l'aide de la théorie de Galois (cf. chapitre d'Algèbre Ensembliste).

3.8. POLYNÔMES TRIGONOMÉTRIQUES

Définition: Nous appelons "**polynôme trigonométrique**" de degré N toute somme finie:

$$S_N(\theta) = \sum_{n=-N}^N c_n e^{in\theta} \quad (8.129)$$

où $c_n \in \mathbb{C}$.

Un polynôme trigonométrique peut aussi être écrit en utilisant les fonctions trigonométriques usuelles grâce aux transformations suivantes:

$$S_N(\theta) = c_0 + \sum_{n=1}^N c_n e^{in\theta} + \sum_{n=1}^N c_{-n} e^{-in\theta} \quad (8.130)$$

Soit en utilisant la formule d'Euler (cf. chapitre sur les Nombres):

$$S_N(\theta) = c_0 + \sum_{n=1}^N c_n (\cos(n\theta) + i \sin(n\theta)) + \sum_{n=1}^N c_{-n} (\cos(n\theta) - i \sin(n\theta)) \quad (8.131)$$

Ce que nous pouvons réécrire aussi sous la forme:

$$S_N(\theta) = c_0 + \sum_{n=1}^N (c_n + c_{-n}) \cos(n\theta) + \sum_{n=1}^N i (c_n - c_{-n}) \sin(n\theta) \quad (8.132)$$

En posant alors:

$$a_0 = 2c_0 \quad a_n = c_n + c_{-n} \quad b_n = i(c_n - c_{-n}) \quad (8.133)$$

Il vient:

$$S_N(\theta) = \frac{a_0}{2} + \sum_{n=1}^N a_n \cos(n\theta) + \sum_{n=1}^N b_n \sin(n\theta) \quad (8.134)$$

Nous verrons longuement dans le chapitre des Suites Et Séries comment utiliser ces polynômes dans le cadre de l'étude des séries de Fourier.

3.9. POLYNÔMES CYCLOTOMIQUES

Si n est un entier naturel, nous appelons "**polynôme cyclotomique**" ce que nous notons traditionnellement Φ_n et que définissons comme étant le produit de tous les monômes $(x - \alpha)$ où α est une racine primitive n -ème de l'unité de $\mathbb{C}$.

Pour rappel une racine n -ème de l'unité (parfois appelée "**nombre de De Moivre**") est un nombre complexe dont la puissance n -ème vaut 1.

Ainsi, l'ensemble des racines n -èmes de l'unité est l'ensemble:

$$G_n = \{x \in \mathbb{C} \text{ tel que } x^n = 1\} \quad (8.135)$$

qui est un groupe cyclique (voir la Théorie Des Ensembles dans la section d'arithmétique du site et le chapitre d'Algèbre Ensembliste dans la présente section).

Nous appelons alors "**racine primitive n -ème de l'unité**" ou "**R.P.N.**" tout élément de ce groupe l'engendrant.

Les éléments de G_n sont donc du type:

$$e^{\frac{2ikx}{n}} \quad (8.136)$$

avec $k \in \mathbb{Z}$. Nous écrivons alors l'ensemble des G_n sous la forme:

$$G_n = \left\{ e^{\frac{2ikx}{n}} \text{ avec } k \in \{0..n-1\} \right\} \quad (8.137)$$

Un petit exemple de polynôme cyclotomique:

$$\Phi_4(x) = (x-i)(x+i)(x-1)(x+1) = x^4 - 1 \quad (8.138)$$

avec:

$$G_4 = \left\{ e^{\frac{2i0\pi}{4}} = 1, e^{\frac{2i1\pi}{4}} = i, e^{\frac{2i2\pi}{4}} = -1, e^{\frac{2i3\pi}{4}} = -i \right\} \quad (8.139)$$

qui sont donc les racines quatrièmes de l'unité (autrement dit chacun de ces nombres mis à la puissance 4 donne 1). Elles forment le groupe G_4 et celui-ci ne peut-être engendré que par i et $-i$ (générateur du groupe selon ce qui a été vu dans le chapitre de Théorie des Ensembles).

Donc un polynôme cyclotomique est le produit de facteurs qui s'écrit:

$$\Phi_n(x) = \prod_k \left(x - e^{\frac{2ikx}{n}} \right) \quad (8.140)$$

avec $k \in \{0..n-1\}$ et k étant premier par rapport à n .

Les polynômes ont un grand nombre de propriétés que nous n'aborderons pas ici puisque ce site ne se veut pas être un ouvrage de mathématiques supérieures.

3.10. POLYNÔMES DE LEGENDRE

Définition: les polynômes de Legendre sont définis par (lire de préférence les chapitres de calcul différentiel et intégral ainsi que d'analyse fonctionnelle avant):

$$P_n(x) = \frac{(-1)^n}{n! 2^n} \cdot \frac{d^n}{dx^n} (1-x^2)^n \quad (8.141)$$

où P_n est donc un polynôme de degré n . Nous retrouverons ces polynômes dans la résolution d'équations différentielles en physique (propagation de la chaleur, physique quantique, chimie quantique, etc.).

Démontrons que selon la définition du produit scalaire fonctionnel (cf. chapitre d'Analyse Fonctionnelle et de Calcul Vectoriel) les polynômes de Legendre sont orthogonaux.

Démonstration:

Soit P un polynôme de degré $\leq n-1$. Il suffit de montrer que $\langle P_n | P \rangle = 0$, c'est-à-dire que P_n est orthogonal à l'espace des polynômes de degré inférieur à n . Nous avons en effet:

$$\langle P_n | P \rangle = \int_{-1}^1 P_n(x) P(x) dx = \frac{(-1)^n}{n! 2^n} \cdot \int_{-1}^1 \frac{d^n}{dx^n} (1-x^2)^n P(x) dx \quad (8.142)$$

en intégrant par parties nous obtenons:

$$\begin{aligned} \langle P_n | P \rangle &= \underbrace{\frac{(-1)^n}{n! 2^n} \cdot \frac{d^{n-1}}{dx^{n-1}} (1-x^2)^n \cdot P(x)}_{=0} \Big|_{-1}^1 - \frac{(-1)^n}{n! 2^n} \cdot \int_{-1}^1 \frac{d^{n-1}}{dx^{n-1}} (1-x^2)^n \cdot \frac{d}{dx} P(x) dx \\ &= - \frac{(-1)^n}{n! 2^n} \cdot \int_{-1}^1 \frac{d^{n-1}}{dx^{n-1}} (1-x^2)^n \cdot \frac{d}{dx} P(x) dx \end{aligned} \quad (8.143)$$

Attention pour le terme nul ci-dessus, seulement le terme $(1-x^2)^n$ y est dérivé. Donc puisque x est au carré, quelque soit la dérivée la valeur sera toujours la même. Ce qui justifie que le terme soit nul.

En continuant de la sorte nous obtenons après n intégrations par parties:

$$\langle P_n | P \rangle = (-1)^n \frac{(-1)^n}{n! 2^n} \cdot \int_{-1}^1 (1-x^2)^n \cdot \underbrace{\frac{d^n}{dx^n} P(x)}_{=0} dx = 0 \quad (8.144)$$

□C.Q.F.D.

Remarque: Le terme dérivé est nul puisque le polynôme dérivé est de degré $n-1$

Voici quelques propriétés utiles en chimie quantique des polynômes de Legendre:

P1. $P_n(1) = 1$

Démonstration:

$$P_n(x) = \frac{(-1)^n}{n! 2^n} \cdot \frac{d^n}{dx^n} (1-x^2)^n = \frac{(-1)^n}{n! 2^n} \cdot \frac{d^n}{dx^n} ((1+x)^n \cdot (1-x)^n) \quad (8.145)$$

et par la formule de Leibniz (cf. chapitre de Calcul Différentiel Et Intégral) nous avons:

$$\frac{(-1)^n}{n! 2^n} \cdot \frac{d^n}{dx^n} ((1+x)^n \cdot (1-x)^n) = \frac{(-1)^n}{n! 2^n} \sum_{j=0}^n \binom{n}{j} \cdot \frac{n!}{(n-j)!} (1+x)^{n-j} \cdot \frac{(-1)^{n-j} \cdot n!}{j!} (1-x)^j \quad (8.146)$$

d'où:

$$P_n(1) = \frac{(-1)^n}{n! 2^n} 2^n \cdot (-1)^n \cdot n! = 1 \quad (8.147)$$

□C.Q.F.D.

P2. $P_n(x) = P_n(-x)$ si n est pair:

Démonstration:

Si n est pair, $\frac{d^n}{dx^n}(1-x^2)^n$ est une fonction paire et donc:

$$P_n(x) = \frac{(-1)^n}{n! 2^n} \cdot \frac{d^n}{dx^n}(1-x^2)^n \quad (8.148)$$

est paire.

□C.Q.F.D.

P3. $P_n(-x) = -P_n(x)$ si n est impair.

Démonstration:

Si n est impair, $\frac{d^n}{dx^n}(1-x^2)^n$ est impaire et donc:

$$P_n(x) = \frac{(-1)^n}{n! 2^n} \cdot \frac{d^n}{dx^n}(1-x^2)^n \quad (8.149)$$

est impaire.

□C.Q.F.D.

Nous allons à présent démontrer la validité de la relation de récurrence suivante pour les P_n (relations que nous utiliserons en physique):

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x) \quad (8.150)$$

pour $n \geq 1$.

Démonstration:

xP_n est un polynôme de degré $n+1$, il existe dès lors des $\alpha_j \in \mathbb{R}$ tel que ce polynôme peut s'exprimer comme combinaison linéaire de la famille de polynômes constituant la base orthonormale (base qui permet donc d'engendrer xP_n):

$$xP_n(x) = \sum_{j=0}^{n+1} \alpha_j P_j \quad (8.151)$$

nous pouvons dès lors écrire:

$$\langle xP_n | P_k \rangle = \left\langle \sum_{j=0}^{n+1} a_j P_j \middle| P_k \right\rangle = \sum_{j=0}^{n+1} a_j \langle P_j | P_k \rangle = a_k \|P_k\|^2 \quad (8.152)$$

mais nous choisissons $k \leq n-2$ (parce que xP_k est dès lors de degré $n-1$):

$$\langle xP_n | P_k \rangle = \int_{-1}^1 xP_n(x)P_k(x)dx = \langle P_n | xP_k \rangle = 0 \quad (8.153)$$

Donc $a_k \|P_k\|^2 = 0$ c'est-à-dire que nous devons avoir $a_k = 0$. Par suite:

$$xP_n = a_{n+1}P_{n+1} + a_nP_n + a_{n-1}P_{n-1} \quad (8.154)$$

Par les propriétés des polynômes de Legendre vues précédemment, nous pouvons écrire les égalités:

$$xP_n(x), x=1: 1 = a_{n+1} + a_n + a_{n-1} \quad (8.155)$$

et:

$$xP_n(x), x=-1: 1 = a_{n+1} - a_n + a_{n-1} \quad (8.156)$$

d'où:

$$a_n = 0 \text{ et } 1 = a_{n+1} + a_{n-1} \quad (8.157)$$

Le coefficient dominant de P_n est défini (rappelons-le) par le coefficient du monôme du plus grand degré. Ainsi, il est donné par:

$$\begin{aligned} P_n(x) &= \frac{(-1)^n}{n!2^n} \frac{d^n}{dx^n} (1-x^2)^n \Rightarrow P_n(x) = \frac{(-1)^n}{n!2^n} \frac{d^n}{dx^n} x^{2n} \\ \Rightarrow P_n(x) &= \frac{(-1)^n}{n!2^n} \frac{d^{n-1}}{dx^{n-1}} (2nx^{2n-1}) = \frac{(-1)^n}{n!2^n} \frac{d^{n-n}}{dx^{n-n}} (2n(2n-1)\dots x^{2n-n}) \\ &= \frac{(-1)^n}{n!2^n} (2n(2n-1)\dots(2n-(n-1))x^n) = \frac{(-1)^n}{n!2^n} (2n(2n-1)\dots(n+1)x^n) = \frac{(2n)!}{(n!)^2 2^n} x^n \end{aligned} \quad (8.158)$$

Donc:

$$\text{dom}(P_n(x)) = \frac{(2n)!}{(n!)^2 \cdot 2^n} \quad (8.159)$$

Remarque: Le lecteur vérifiera au besoin pour un n donné que:

$$2n(2n-1)\dots(n+1) = \frac{(2n)!}{n!} \quad (8.160)$$

La relation:

$$xP_n = a_{n+1}P_{n+1} + a_nP_n + a_{n-1}P_{n-1} \quad (8.161)$$

que nous avons obtenu ci-dessus nous impose que le coefficient dominant du polynôme de la combinaison linéaire soit égal au coefficient dominant du polynôme xP_n (nous avons éliminé le $(-1)^n$ qui se simplifie):

$$\cancel{(-1)^n} \frac{(2n)!}{(n!)^2 \cdot 2^n} = a_{n+1} \cancel{(-1)^{n+2}} \frac{(2n+2)!}{(n+1)!^2 \cdot 2^{n+1}} \quad (8.162)$$

après simplification, nous obtenons:

$$a_{n+1} = \frac{(2n)!(n+1)!^2 2^{n+1}}{(n!)^2 2^n (2n+2)!} = \frac{(n+1) \cancel{(n+1)^2}}{(2n+1) \cancel{(2n+2)}} = \frac{n+1}{2n+1} \quad (8.163)$$

et ce qui donne finalement facilement:

$$a_{n-1} = \frac{n}{2n+1} \quad (8.164)$$

La relation:

$$xP_n = a_{n+1}P_{n+1} + a_nP_n + a_{n-1}P_{n-1} \quad (8.165)$$

devient dès lors:

$$(2n+1)xP_n = (n+1)P_{n+1} + nP_{n-1} \quad (8.166)$$

□C.Q.F.D.

Voici les six premiers polynômes de Legendre:

$$P_0 = 1, P_1 = x, P_2 = 3/2x^2 - 1/2, P_3 = 5/2x^3 - 3/2x, P_4 = 35/8x^4 - 15/4x^2 + 3/8, \dots \quad (8.167)$$

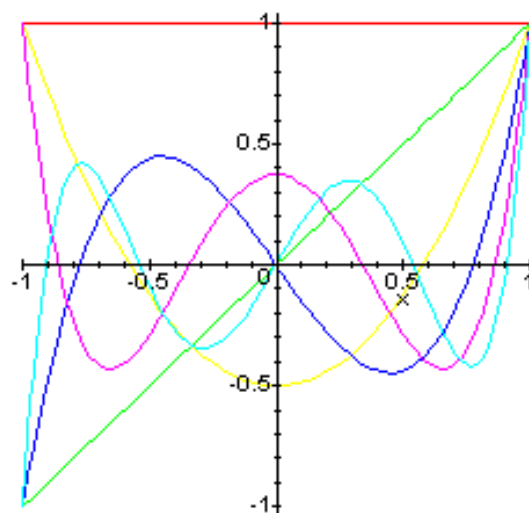


Figure: 8.9 - Quelques polynômes de Legendre

Notes personnelles:







9. ALGÈBRE (ET GÉOMÉTRIE) ENSEMBLISTE

Nous allons aborder sur ce site l'étude des structures ensemblistes de manière très pragmatique (puisque rappelons que ce site est dédié aux ingénieurs). Ainsi, il sera fait usage du minimum de formalisme et seulement les démonstrations des éléments que nous considérons comme absolument essentiels à l'ingénieur seront présentées. Par ailleurs, de nombreuses démonstrations seront faites par l'exemple et nous nous focaliserons en grande partie sur la théorie algébrique des groupes car elle a une place presque prédominante en physique plus que pour les autres structures ensemblistes.

1. ALGÈBRE ET GÉOMÉTRIE CORPORELLE

Les symétries des figures géométriques, des cristaux et de tous les autres objets de la physique macroscopique font l'objet depuis des siècles d'observations et d'études. En termes modernes, les symétries d'un objet donné forment un groupe.

Depuis le milieu du 19ème siècle, la théorie des groupes a pris une extension énorme, et ses applications à la mécanique quantique et à la théorie des particules élémentaires se sont développées tout au long du 20ème siècle.

Dans une lettre de 1877 au mathématicien Adolph Mayer, Sophus Lie écrit qu'il a créé la théorie des groupes en janvier 1873. Il s'agit bien sûr des groupes qu'il appelait "groupes continus" et qui sont appelés aujourd'hui "groupes de Lie". Lie cherchait à étendre l'usage des groupes du domaine des équations algébriques, où Galois les avait introduites, à celui des équations différentielles.

Dès 1871, la notion de générateur infinitésimal d'un groupe à un paramètre de transformations était apparue dans son oeuvre. C'est l'ensemble des générateurs infinitésimaux des sous-groupes à un paramètre d'un groupe continu qui forme ce que nous appelons aujourd'hui une algèbre de Lie.

Ce furent Wigner et Weyl qui montrèrent le rôle prééminent de la théorie des groupes, et de leurs représentations en particulier, dans la nouvelle mécanique quantique que développaient Heisenberg et Dirac. L'idée générale de la théorie des représentations est d'essayer d'étudier un groupe en le faisant agir sur un espace vectoriel de manière linéaire : nous essayons ainsi de voir le groupe comme un groupe de matrices (d'où le terme "représentation"). Nous pouvons ainsi, à partir des propriétés relativement bien connues du groupe des automorphismes de l'espace vectoriel (cf. [chapitre de Théorie des Ensembles](#)), arriver à déduire quelques propriétés du groupe qui nous intéresse.

Nous pouvons considérer la théorie des représentations de groupes comme une vaste généralisation de l'analyse de Fourier. Son développement est continu et elle a, depuis le milieu du 20ème siècle, des applications innombrables en géométrie différentielle, en théorie ergodique, en théorie des probabilités, en théorie des nombres, dans la théorie des formes automorphes, dans celle des systèmes dynamiques ainsi qu'en physique, chimie, biologie moléculaire et traitement du signal. À l'heure actuelle, des branches entières des mathématiques et de la physique en dépendent.

Avant de commencer, nous renvoyons le lecteur au chapitre traitant de la Théorie Des Ensembles pour qu'il se rappelle de la structure et des propriétés fondamentales qui constituent le groupe et également au chapitre d'Algèbre Linéaire (car nous en utiliserons quelques résultats).

1.1. GROUPES CYCLIQUES

Le groupe cyclique (dont la définition a déjà été vue dans le chapitre de Théorie des Ensembles) va nous servir de base dans le cadre de l'étude des groupes finis. Par ailleurs, plutôt que de faire des développements généralisés nous avons préféré prendre des exemples particuliers afin de présenter l'idée de groupe cyclique (approche plus adaptée à l'ingénieur).

Nous allons donc prendre l'exemple fort sympathique des heures de la montre... avec trois approches différentes qui successivement (!) permettront d'aborder un groupe cyclique simple.

- Première approche:

Imaginons donc une horloge avec une aiguille qui peut prendre 12 positions possibles (mais pas de positions intermédiaires). Nous noterons de manière spéciale les 12 positions possibles: $\bar{0}, \bar{1}, \bar{2}, \dots, \bar{11}$ (le trait au-dessus des nombres n'est pas innocent!).

Rien ne nous empêche sur l'ensemble de ces positions de définir une addition, par exemple:

$$\bar{2} + \bar{3} = \bar{5}, \bar{3} + \bar{2} = \bar{5}, \bar{5} + \bar{7} = \bar{0}, \bar{9} + \bar{5} = \bar{2}, \bar{1} - \bar{8} = \bar{5} \quad (9.1)$$

ce qui est similaire aux résultats que nous obtenons lorsque dans notre quotidien nous faisons des calculs avec notre montre.

- Deuxième approche (première extension)

Si nous observons bien notre montre, nous remarquons qu'à chaque fois que nous rajoutons 12 (ou retirons...) à une valeur des heures de notre montre alors nous tombons sur un ensemble de nombres bien déterminé qui sont aussi dans $\mathbb{Z}$. Ainsi (évidemment dans le cadre d'une montre seules les premières valeurs positives nous intéressent la plupart du temps mais ici nous faisons des maths alors nous généralisons un peu...):

$$\begin{aligned} \bar{0} &:= 0 + 12\mathbb{Z} = \{\dots, -24, -12, 0, 12, 24, \dots\} \\ \bar{1} &:= 1 + 12\mathbb{Z} = \{\dots, -11, 1, 13, 25, \dots\} \\ \bar{2} &:= 2 + 12\mathbb{Z} = \{\dots, -10, 2, 14, 26, \dots\} \\ &\dots \end{aligned} \quad (9.2)$$

Nous retrouvons ici un concept que nous avons déjà vu dans le chapitre de Théorie Des Nombres. Il s'agit de classes de congruences et l'ensemble de ces classes forme l'ensemble quotient $\mathbb{Z}/12\mathbb{Z}$. Si nous munissons cet ensemble quotient d'une loi d'addition, il est normalement facile d'observer que celle-ci est une loi interne à l'ensemble quotient, qu'elle est associative, qu'il existe un élément neutre et que chaque élément possède un symétrique (inverse).

Ainsi, cet ensemble quotient muni uniquement de la loi d'addition (sinon en ajoutant la multiplication nous pouvons former un anneau) est un groupe commutatif.

- Troisième approche (deuxième et dernière extension):

Voyons une troisième et dernière approche qui explique pourquoi le groupe quotient est cyclique.

Si nous projetons la rotation des aiguilles de notre montre (toutes les rotations dans l'algèbre

ensembliste se font traditionnellement dans le sens des aiguilles d'une montre!) dans $\mathbb{C}$ et que nous définissons:

$$x := e^{\frac{2\pi}{12}} \quad (9.3)$$

Nous avons alors $x^{12} = x^0 = 1$ et:

$$\forall a, k \in \mathbb{Z} \quad x^a = x^{a+12k} \quad (9.4)$$

ce qui explique pourquoi le groupe quotient $(\mathbb{Z}/12\mathbb{Z}, +)$ est appelé "**groupe cyclique**" (par isomorphisme de groupe selon ce qui a été vu en théorie des ensembles). Son isomorphe est noté C_{12} .

Si nous représentons dans $\mathbb{C}$ l'ensemble isomorphe C_{12} nous obtenons alors sur le cercle unité un polygone ayant n sommets comme le montre la figure ci-dessous:

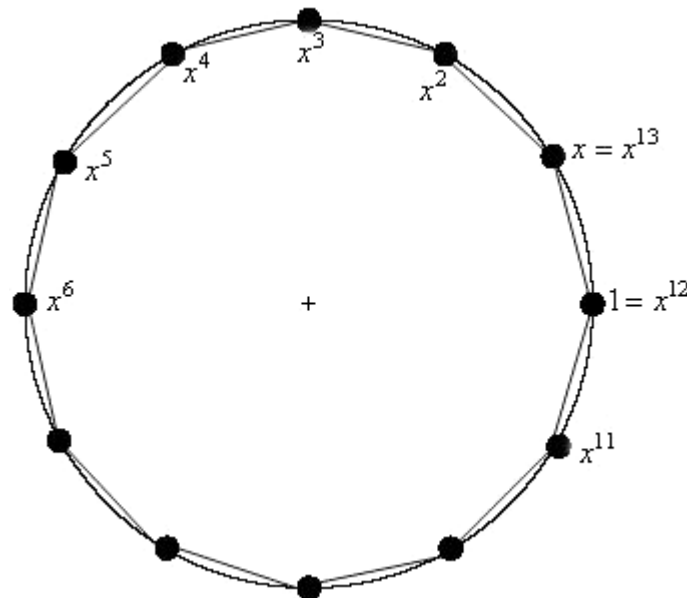


Figure: 9.1 - Groupe cyclique d'ordre 12

Par ailleurs, le nombre d'éléments composants $\mathbb{Z}/12\mathbb{Z}$ étant fini, $(\mathbb{Z}/12\mathbb{Z}, +)$ est fini. Contrairement au groupe $(\mathbb{Z}, +)$ qui est lui un groupe discret infini.

Ce concept de finitude sera peut-être plus évident avec l'exemple que nous ferons de suite après avec $\mathbb{Z}/4\mathbb{Z}$ où le lecteur observera que cet ensemble a le même nombre d'éléments que C_4 .

Remarque: Les mathématiciens appellent C_n le "**groupe des racines n -èmes de l'unité**". Une racine n -ème de l'unité (parfois appelée "**nombre de De Moivre**") est donc un nombre complexe dont la puissance n -ème vaut 1. Par ailleurs, pour un entier n donné, toutes les racines n -èmes de l'unité sont situées sur le cercle unité et sont les sommets d'un polygone régulier à n côtés ayant un sommet d'affixe 1.

Ce qui intéresse les physiciens particulièrement dans un premier temps ce sont les représentations des groupes finis (aussi les groupes continus que nous verrons plus loin). Ainsi, la représentative de $\mathbb{Z}/n\mathbb{Z}$

nous est connue puisque la rotation dans le plan complexe est donnée comme nous l'avons montrée notre étude des complexes dans le chapitre sur les Nombres:

$$\begin{bmatrix} \cos\left(\frac{2k\pi}{n}\right) & -\sin\left(\frac{2k\pi}{n}\right) \\ \sin\left(\frac{2k\pi}{n}\right) & \cos\left(\frac{2k\pi}{n}\right) \end{bmatrix} \quad (9.5)$$

avec $k \in [0, n-1]$. Cette représentative est un sous-groupe du groupe des rotations $O(2)$ sur lesquelles nous reviendrons plus loin. Le groupe des rotations du plan étant lui-même un sous-groupe du groupe linéaire $GL(2)$ (nous en donnerons une définition précise et un exemple plus loin).

Au fait, les mathématiciens sont capables de démontrer que tous les groupes quotients $\mathbb{Z}/n\mathbb{Z}$ sont cycliques à isomorphisme près avec C_n . Les mathématiciens disent aussi que $\mathbb{Z}/n\mathbb{Z}$ est un quotient fini du groupe monogène $\mathbb{Z}$.

Cette approche est par contre peut-être un peu abstraite. Alors, si le lecteur se rappelle du chapitre de Théorie Des Ensembles nous avons vu une définition bien précise de ce qu'était la cyclicité d'un groupe: Un groupe G est dit cyclique si G est engendré par la puissance d'au moins un de ses éléments $a \in G$ appelé générateur tel que:

$$G = \{a^n, n \in \mathbb{Z}\} \quad (9.6)$$

Vérifions que ce soit bien le cas pour le groupe:

$$G = \{\mathbb{Z}/4\mathbb{Z}, +\} \quad (9.7)$$

qui constitue un cas scolaire.

Nous noterons les éléments qui constituent ce groupe:

$$\{0, 1, 2, 3\} \quad (9.8)$$

Ceci étant fait, il convient de faire attention que dans la définition ensembliste du groupe cyclique nous parlons de "puissance" si la loi interne du groupe est la multiplication mais si la loi interne est l'addition, nous avons alors:

$$G = \{n \cdot a, n \in \mathbb{Z}\} \quad (9.9)$$

Le premier élément générateur du groupe:

$$G = \{\mathbb{Z}/4\mathbb{Z}, +\} \quad (9.10)$$

est l'élément 1. Effectivement:

$$\begin{aligned}
 1 &= 1 \cdot 1 = 1 \\
 1+1 &= 2 \cdot 1 = 2 \\
 1+1+1 &= 3 \cdot 1 = 3 \\
 1+1+1+1 &= 4 \cdot 1 = 0
 \end{aligned}
 \tag{9.11}$$

Le deuxième élément générateur du même groupe est 3:

$$\begin{aligned}
 3 &= 1 \cdot 3 = 3 \\
 3+3 &= 2 \cdot 3 = 2 \\
 3+3+3 &= 3 \cdot 3 = 1 \\
 3+3+3+3 &= 4 \cdot 3 = 0
 \end{aligned}
 \tag{9.12}$$

Par contre, le lecteur pourra vérifier que 2 n'est pas générateur de ce groupe!

Au fait, en ce qui concerne les groupes $G = \{\mathbb{Z} / n\mathbb{Z}, +\}$ les mathématiciens arrivent à démontrer que seuls les éléments du groupe qui sont premiers avec n sont générateurs (c'est-à-dire les éléments dont le plus grand commun diviseur est 1).

Voilà pour notre introduction aux groupes cycliques. Passons maintenant à une autre catégorie de groupes.

1.2. GROUPES DE TRANSFORMATIONS

Le groupe des rotations est celui qui intéresse le plus les physiciens surtout dans les domaines des matériaux, de la chimie, de la physique quantique et de l'art... Les mathématiciens apprécient eux l'étude des groupes de rotations dans le cadre de la géométrie bien évidemment (mais pas seulement) et les informaticiens tout autant les groupes linéaires. Nous avons d'ailleurs vu un exemple de groupe de rotations juste précédemment.

Définition: Nous appelons "**groupe linéaire d'ordre n** " et nous notons $GL(n)$ les matrices inversibles (donc le déterminant est non nul selon ce que nous avons vu dans le chapitre d'Algèbre Linéaire) dont les coefficients sont dans un corps quelconque:

$$GL(n) = \{A \in M_n \mid \det(A) \neq 0\} \tag{9.13}$$

Un exemple simple et important de groupe linéaire est celui du sous-"**groupe des transformations affines**" du plan qui est traditionnellement noté (c'est intuitif):

$$F : M(x, y) \rightarrow M'(x', y') : \begin{cases} x' = ax + by + \alpha \\ y' = cx + dy + \beta \end{cases} \tag{9.14}$$

avec $a, b, c, d, \alpha, \beta \in \mathbb{R}, ad - bc \neq 0$ (nous verrons le pourquoi du comment de l'inégalité un peu plus loin).

Prenons un exemple pratique:

$$F : M(x, y) \rightarrow M'(x', y') : \begin{cases} x' = 0.5x - 0.4y + 0.5 \\ y' = 0.8x + 0.2y - 1.5 \end{cases} \tag{9.15}$$

ce qui appliqué à un cercle donnerait:

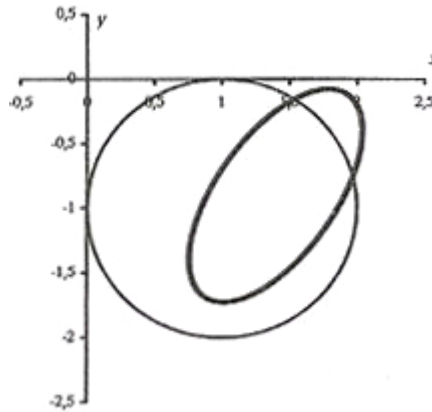


Figure: 6.2 - Transformations affines sur un cercle

Cette transformation est une manière de définir les ellipses comme images d'un cercle par une transformation affine.

Les coefficients α, β sont sans importance pour la forme de l'image. En fait, ils induisent bien évidemment des translations sur les figures. Nous pouvons donc nous en passer si nous cherchons seulement à la déformer.

Ainsi, il nous reste:

$$F: M(x, y) \rightarrow M'(x', y'): \begin{cases} x' = ax + by \\ y' = cx + dy \end{cases} \quad (9.16)$$

ce qui peut s'écrire sous forme matricielle:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (9.17)$$

La transformation se réduit donc à la matrice:

$$F = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad (9.18)$$

et comme nous l'avons vu en algèbre linéaire, la multiplication matricielle est associative mais n'est pas commutative, donc la transformation linéaire ne l'est pas non plus.

L'élément neutre est la matrice:

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (9.19)$$

et l'inverse de F est:

$$F^{-1} = \begin{bmatrix} \frac{d}{ad-bc} & \frac{-b}{ad-bc} \\ \frac{-c}{ad-bc} & \frac{a}{ad-bc} \end{bmatrix} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \quad (9.20)$$

et comme nous avons imposé $ad - bc \neq 0$ tout élément y possède donc un inverse. Ainsi, le groupe linéaire affine est non commutatif et... forme bien un groupe...

Définition: Nous appelons "**groupe spécial linéaire d'ordre n** " et nous notons $SL(n)$ les matrices inversibles dont les coefficients sont dans un corps quelconque et dont le déterminant est égal à l'unité:

$$SL(n) = \{A \in M_n \mid \det(A) = 1\} \quad (9.21)$$

Il s'agit évidemment d'un sous-groupe de $GL(n)$.

En reprenant l'exemple précédant et en se rappelant que le déterminant d'une matrice carrée bidimensionnelle est (cf. chapitre d'Algèbre Linéaire):

$$\det(F) = ad - bc = 1 = \det(F^{-1}) \quad (9.22)$$

nous remarquons bien géométriquement ce que signifie d'avoir un déterminant unitaire dans ce cas! Effectivement nous avons vu dans le chapitre d'Algèbre Linéaire lors de notre interprétation géométrique qu'avoir un déterminant équivaut à une surface. Ainsi, le fait d'avoir $ad-bc$ unitaire permet donc que quel que soit l'ordre de la transformation, nous avons l'aire qui vaut toujours 1. Ainsi, le groupe spécial linéaire conserve les surfaces.

Définition: Nous appelons "**groupe orthogonal réel d'ordre n** " et notons $O(n)$ les matrices orthogonales (cf. chapitre d'Algèbre Linéaire) $n \times n$ données par:

$$O(n) = \left\{ A \in M_n(\mathbb{R}) \mid A^T A = I_n \right\} \quad (9.23)$$

Par ailleurs, nous avons démontré dans le chapitre d'Algèbre Linéaire lors de notre étude des matrices de rotations que $A^T A = I_n$ implique $\det(A) = \pm 1$.

C'est le cas par exemple de la matrice de $O(2)$ vue précédemment (elle appartient au groupe orthogonal mais aussi au groupe des rotations que nous verrons plus loin):

$$\begin{bmatrix} \cos\left(\frac{2k\pi}{n}\right) & -\sin\left(\frac{2k\pi}{n}\right) \\ \sin\left(\frac{2k\pi}{n}\right) & \cos\left(\frac{2k\pi}{n}\right) \end{bmatrix} \quad (9.24)$$

qui est orthogonale comme il est facile de le vérifier.

Remarque: $O(1)$ est constitué aussi de l'ensemble des matrices triviales.... $[1], [-1]$

Définition: Si $A \in O(n)$ et que nous avons $\det(A) = 1$ alors nous obtenons un sous-groupe de $O(n)$

appelé "**groupe spécial orthogonal réel d'ordre n** " et noté $SO(n)$:

$$SO(n) = \left\{ A \in M_n(\mathbb{R}) \mid A^T A = I_n, \det(A) = 1 \right\} \quad (9.25)$$

La matrice de rotations donnée précédemment fait partie de ce groupe puisque son déterminant est égal à l'unité! Par ailleurs, ce groupe occupe une place très spéciale en physique et nous le retrouverons maintes fois.

Le sous-groupe $SO(2)$, appelé aussi parfois "**groupe cercle**" et noté S^1 , que nous avons aussi étudié dans le chapitre de Géométrie Euclidienne a une représentative donnée par la matrice:

$$\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \quad (9.26)$$

et occupe une place à part dans la famille des groupes $SO(n)$ avec n supérieur à l'unité. Effectivement il est le seul à être commutatif. Par ailleurs, il est isomorphe à $e^{i\theta}$ soit à $U(1)$ le groupe multiplicatif des nombres complexes de module 1. C'est aussi le groupe de symétrie propre d'un cercle et l'équivalent continu C_n .

Le sous-groupe $SO(3)$ donné par la matrice (cf. chapitre de Géométrie Euclidienne):

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta \\ 0 & \sin \theta & \cos \theta \end{bmatrix} \quad (9.27)$$

pour la rotation autour de l'axe X dans l'espace tridimensionnel n'est pas commutatif. Par ailleurs les quaternions, dont la représentative est donc $SO(3)$, forment un groupe non commutatif aussi (par rapport à la loi de multiplication) comme nous l'avons vu dans le chapitre sur les Nombres.

Par rapport à un vecteur unitaire on se rend facilement compte visuellement parlant que $SO(3)$ est un sous-groupe fermé de $GL(3)$, c'est-à-dire de l'ensemble des groupes linéaires de dimension 3.

Remarque: $SO(1)$ est constitué de la matrice $[1]$.

Définition: Nous appelons "**groupe unitaire d'ordre n** " et nous notons $U(n)$ les matrices dont les composantes sont complexes (dans le cadre de ce site le plus souvent) ou réelles et qui sont orthogonales:

$$U(n) = \left\{ A \in M_n(\mathbb{C}) \mid A^\dagger A = I_n \right\} \quad (9.28)$$

Remarquons par ailleurs que toute matrice unitaire à coefficients complexes et à une dimension... (de $U(n)$ donc...) est un nombre complexe de module unitaire, qui peut toujours s'écrire sous la forme $e^{i\mathbb{R}}$.

Nous en avons déjà vu un exemple aussi sur le site lors de notre étude des spineurs dans le chapitre de Calcul Spinoriel. Il s'agit des matrices de Pauli (utilisées dans le chapitre de Physique Quantique Relativiste) données par:

$$\sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (9.29)$$

Définition: Nous appelons "**groupe spécial unitaire d'ordre n** " et nous notons $SU(n)$ les matrices dont les coefficients sont complexes et qui sont orthogonales et dont le déterminant est unitaire:

$$SU(n) = \left\{ A \in M_n(\mathbb{C}) \mid A^\dagger A = I_n, \det(A) = 1 \right\} \quad (9.30)$$

Remarque: $U(1)$ est égal à $SU(1)$ et il s'agit donc du cercle unité complexe égal à $e^{i\mathbb{R}}$. Par ailleurs, $SO(2)$ est commutatif et isomorphe à $U(1)$ car c'est l'ensemble des rotations du plan.

Un exemple connu est toujours celui des matrices de Pauli mais simplement écrites sous la forme utilisée en Physique Quantique Relativiste (voir chapitre du même nom):

$$i\sigma_x = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}, i\sigma_y = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, i\sigma_z = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix} \quad (9.31)$$

qui font partie de $SU(2)$ et qui comme nous l'avons montré (implicitement) au début du chapitre de Calcul Spinoriel est isomorphe au groupe des quaternions $SO(3)$ de module 1 sur la sphère de dimension 3 (notée S^3). Relation que les mathématiciens appellent dans le cas présent un "**homomorphisme de revêtement**"....

Remarque: Le groupe spécial unitaire possède une importance particulière en physique des particules. Si le groupe unitaire $U(1)$ est le groupe de jauge de l'électromagnétisme (pensez au nombre complexe apparaissant dans les solutions de l'équation d'onde!), $SU(2)$ est le groupe associé à l'interaction faible, et $SU(3)$ celui de l'interaction forte. C'est par exemple grâce à la structure des représentations de $SU(3)$ que Gell-Mann a conjecturé l'existence des quarks.

Avec une approche différente de celle vue dans le chapitre de Calcul Spinoriel comment montrer que les matrices de Pauli sont les bases de $SU(2)$?

D'abord, rappelons que nous avons montré dans de Calcul Spinoriel que toute rotation dans l'espace de trois dimensions pouvait s'exprimer à l'aide de la relation:

$$R\left(\vec{L}, \frac{\theta}{2}\right) = \cos \frac{\theta}{2} \sigma_0 - i \sin \frac{\theta}{2} \vec{L} \circ \vec{\sigma} \quad (9.32)$$

Et nous avons vu dans le chapitre d'Informatique Quantique qu'une formulation explicitée et décomposée de la relation précédente était:

$$\begin{aligned}
 R_x(\theta) &= \begin{bmatrix} \cos \frac{\theta}{2} & -i \sin \frac{\theta}{2} \\ -i \sin \frac{\theta}{2} & \cos \frac{\theta}{2} \end{bmatrix} \\
 R_y(\theta) &= \begin{bmatrix} \cos \frac{\theta}{2} & -\sin \frac{\theta}{2} \\ \sin \frac{\theta}{2} & \cos \frac{\theta}{2} \end{bmatrix} \\
 R_z(\theta) &= \begin{bmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{bmatrix}
 \end{aligned} \quad (9.33)$$

et donc que tout élément de $SU(2)$ est produit de ces trois matrices qui font chacune décrire à l'extrémité d'un vecteur dans l'espace une courbe!

Maintenant, nous remarquons que ces trois matrices sont égales à:

$$R_x(\theta) = R_y(\theta) = R_z(\theta) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (9.34)$$

lorsque $\theta = 0$. Nous obtenons alors la matrice identité. Donc si nous cherchons la tangente en ce point conjoint, nous pouvons dès lors construire une base (3 vecteurs orthogonaux).

Regardons ceci:

$$\begin{aligned}
 \left. \frac{dR_x(\theta)}{d\theta} \right|_{\theta=0} &= \left. \begin{bmatrix} -\frac{1}{2} \sin \frac{\theta}{2} & -i \frac{1}{2} \cos \frac{\theta}{2} \\ -i \frac{1}{2} \cos \frac{\theta}{2} & -\frac{1}{2} \sin \frac{\theta}{2} \end{bmatrix} \right|_{\theta=0} = \begin{bmatrix} 0 & -\frac{1}{2}i \\ -\frac{1}{2}i & 0 \end{bmatrix} = -\frac{i}{2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = -\frac{i}{2} \sigma_x \\
 \left. \frac{dR_y(\theta)}{d\theta} \right|_{\theta=0} &= \left. \begin{bmatrix} -\frac{1}{2} \sin \frac{\theta}{2} & -\frac{1}{2} \cos \frac{\theta}{2} \\ \frac{1}{2} \cos \frac{\theta}{2} & -\frac{1}{2} \sin \frac{\theta}{2} \end{bmatrix} \right|_{\theta=0} = \begin{bmatrix} 0 & -\frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} = \frac{1}{i2} \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} = -\frac{i}{2} \sigma_y \\
 \left. \frac{dR_z(\theta)}{d\theta} \right|_{\theta=0} &= \left. \begin{bmatrix} -\frac{1}{2} i e^{-i\theta/2} & 0 \\ 0 & \frac{1}{2} i e^{i\theta/2} \end{bmatrix} \right|_{\theta=0} = \begin{bmatrix} -\frac{1}{2}i & 0 \\ 0 & \frac{1}{2}i \end{bmatrix} = -\frac{i}{2} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = -\frac{i}{2} \sigma_z
 \end{aligned} \quad (9.35)$$

Ainsi, $SU(2)$ admet pour base:

$$\left\{ -\frac{i}{2} \sigma_x, -\frac{i}{2} \sigma_y, -\frac{i}{2} \sigma_z \right\} \quad (9.36)$$

et ce sont en d'autres termes les générateurs infinitésimaux du groupe $SU(2)$. $SU(2)$ a donc une base qui est une Algèbre de Lie selon le vocabulaire des mathématiciens.

Ce résultat est assez remarquable... Puisque $SU(2)$ et $SO(3)$ sont isomorphes, nous pouvons alors

obtenir la base de l'Algèbre de Lie de $SO(3)$ alors avec la même méthode!!!

Voyons ceci! Nous avons vu dans le chapitre de Géométrie Euclidienne que les matrices de rotations étaient données par (nous changeons le R par un U afin de ne pas confondre avec les matrices précédentes):

$$U_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & \sin \theta \\ 0 & -\sin \theta & \cos \theta \end{bmatrix}, U_y = \begin{bmatrix} \cos \gamma & 0 & -\sin \gamma \\ 0 & 1 & 0 \\ \sin \gamma & 0 & \cos \gamma \end{bmatrix}, U_z = \begin{bmatrix} \cos \phi & \sin \phi & 0 \\ -\sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (9.37)$$

Nous remarquons à nouveau qu'en $\theta = \gamma = \phi = 0$ la courbe que fait décrire à un vecteur les trois matrices de rotations passe par:

$$U_x = U_y = U_z = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (9.38)$$

Alors de la même manière que pour $SU(2)$, nous calculons les dérivées en ces angles pour déterminer les matrices de base génératrices de $SO(3)$:

$$\begin{aligned} \left. \frac{dU_x}{d\theta} \right|_{\theta=0} &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & -\sin \theta & \cos \theta \\ 0 & -\cos \theta & -\sin \theta \end{bmatrix} \Big|_{\theta=0} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} = X_x \\ \left. \frac{dU_y}{d\gamma} \right|_{\gamma=0} &= \begin{bmatrix} -\sin \gamma & 0 & -\cos \gamma \\ 0 & 0 & 0 \\ \cos \gamma & 0 & -\sin \gamma \end{bmatrix} \Big|_{\gamma=0} = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} = X_y \\ \left. \frac{dU_z}{d\phi} \right|_{\phi=0} &= \begin{bmatrix} -\sin \phi & \cos \phi & 0 \\ -\cos \phi & -\sin \phi & 0 \\ 0 & 0 & 0 \end{bmatrix} \Big|_{\phi=0} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = X_z \end{aligned} \quad (9.39)$$

L'algèbre de Lie de $SO(3)$ admet donc pour base:

$$\{X_x, X_y, X_z\} \quad (9.40)$$

En physique, on préfère travailler avec des matrices complexes. Nous introduisons alors les matrices:

$$\{J_x = iX_x, J_y = iX_y, J_z = iX_z\} \quad (9.41)$$

Il faut alors remarquer que si nous définissons:

$$J_{\pm} = J_x \pm iJ_y = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & i \\ 0 & -i & 0 \end{bmatrix} \pm i \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ i & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & \pm 1 \\ 0 & 0 & i \\ \mp 1 & -i & 0 \end{bmatrix} \quad (9.42)$$

nous avons trivialement pour la complexe conjuguée de la matrice transposée:

$$(\mathcal{J}_{\pm})^{\dagger} = \mathcal{J}_{\mp} \quad (9.43)$$

et au fait... nous avons aussi les relations de non-commutation (ce que nous pouvons développer sur demande):

$$[\mathcal{J}_x, \mathcal{J}_y] = i\mathcal{J}_z \quad (9.44)$$

(cycl.)

et aussi la relation de commutation:

$$[\mathcal{J}_x, \|\vec{\mathcal{J}}\|^2] = 0 \quad (9.45)$$

(cycl.)

ce que satisfont aussi les matrices de Pauli et... pour rappel (ou information pour ceux qui n'ont pas encore lu le chapitre de Physique Quantique Ondulatoire) les $\mathcal{J}_i$ sont les opérateurs du moment cinétique total du système de couplage spin-orbite!!!

1.3. GROUPES DE SYMÉTRIES

Le groupe de symétries d'un objet noté X (image, signal etc. en 1D, 2D, 3D ou autre) est le groupe de toutes les isométries (une isométrie est une transformation qui conserve les longueurs) sous lesquelles il est invariant avec la composition en tant qu'opération.

Tout groupe de symétries dont les éléments ont un point fixe commun, ce qui est vrai pour tous les groupes de symétries de figures limitées, peut être représenté comme un sous-groupe du groupe orthogonal $O(n)$ en choisissant l'origine pour point fixe. Le groupe de symétries propre est alors un sous-groupe du groupe orthogonal spécial $SO(n)$, et par conséquent, il est aussi appelé le groupe de rotations de la figure.

Dans ce qui suit, nous allons interpréter la composée de deux opérations de symétries ou de rotations comme une multiplication au même titre que pour les permutations.

Définitions:

D1. Le "**groupe des symétries**", appelé aussi "**groupe des invariants**", de X est l'ensemble des symétries de X , muni de la structure de multiplication donnée par composition qui laisse X invariant.

D2. "**L'ordre**" d'un groupe est le nombre total de toutes ses symétries uniquement (y compris l'identité!).

Exemples:

E1. Le coeur:



Figure: 6.3 - Analyse du groupe de symétrie du coeur

a un groupe de symétries total à 2 éléments, à savoir l'application identité id et l'application r_v qui est la réflexion dans l'axe vertical (sous-groupe de symétries à 1 élément). Cette forme possède donc un groupe de symétries d'ordre 2. Nous observons que le symétrique est donné aussi via la relation

$$r_v \circ r_v = id$$

E2. La lettre:



Figure: 6.4 - Analyse du groupe de symétrie de phi

a un groupe de symétrie total à 4 éléments, à savoir l'application identité id , les deux réflexions r_h et r_v et la rotation par l'angle π que nous noterons t_π (sous-groupe de rotations à 1 élément). Cette forme possède donc un groupe de symétries d'ordre 3.

Dans ce groupe nous avons $r_h \circ r_v = t_\pi$ (et c'est commutatif!), $t_\pi \circ t_\pi$ est la rotation par un angle 2π , ce qui est la même application que l'application identité, donc $t_\pi \circ t_\pi = id$.

Ainsi, le groupe de symétries de cette lettre est commutatif et la loi de composition est bien interne. C'est donc bien un groupe.

E3. Le pentagone régulier:

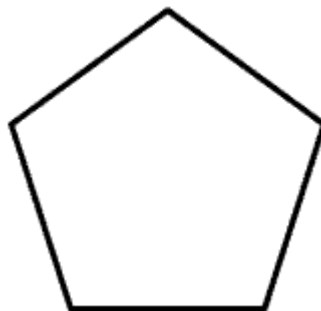


Figure: 6.5 - Analyse du groupe de symétrie du pentagone régulier

a un groupe de symétries total à 10 éléments à savoir les 5 rotations $id, t_{2\pi/5}, t_{4\pi/5}, t_{6\pi/5}, t_{8\pi/5}$ ainsi que les 5 réflexions dans les 5 axes de symétries. C'est donc un groupe de symétries d'ordre 5 correspondant au groupe cyclique $\mathbb{Z}/5\mathbb{Z}$.

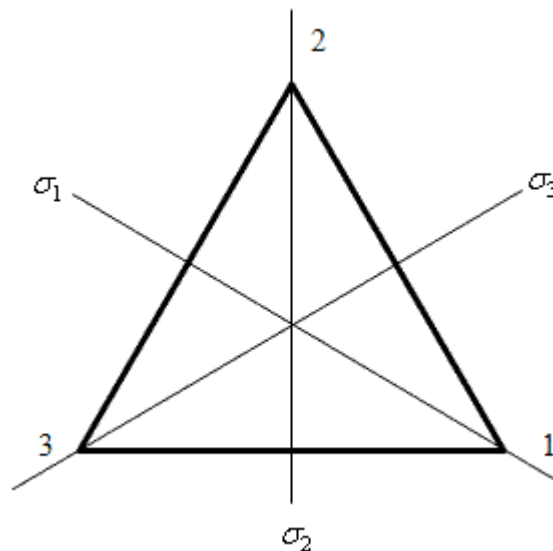
Remarque: Plus généralement, le groupe de symétries d'un n -gone régulier (si n est impair) a exactement $2n$ éléments. Ce groupe s'appelle le "**groupe diédral d'ordre n** " et est noté le plus souvent D_{2n} (il faut faire attention car certains auteurs ne multiplient pas n par le facteur 2 ce qui fait que l'indice représente alors directement l'ordre et non le nombre d'éléments).

Le pentagone a donc D_{10} pour groupe diédral et $\mathbb{Z}/5\mathbb{Z}$ en est un "**sous-groupe distingué**" (nous reviendrons plus tard sur cette notion de sous-groupe distingué).

E4. Le groupe diédral D_6 d'ordre 6 des isométries d'un triangle équilatéral (polygone régulier) a 6 éléments que nous noterons (afin que l'écriture soit moins lourde):

$$D_6 = \{id, t_{2\pi/3}, t_{4\pi/3}, \sigma_1, \sigma_2, \sigma_3\} \quad (9.46)$$

où $\sigma_1, \sigma_2, \sigma_3$ sont les symétries par rapport aux trois bissectrices (respectivement médiatrices). La table de compositions de ce groupe diédral montre aussi que ce groupe est non-commutatif:



$\nearrow D_6$	id	$t_{2\pi/3}$	$t_{4\pi/3}$	σ_1	σ_2	σ_3
id	id	$t_{2\pi/3}$	$t_{4\pi/3}$	σ_1	σ_2	σ_3
$t_{2\pi/3}$	$t_{2\pi/3}$	$t_{4\pi/3}$	id	σ_3	σ_1	σ_2
$t_{4\pi/3}$	$t_{4\pi/3}$	id	$t_{2\pi/3}$	σ_2	σ_3	σ_1
σ_1	σ_1	σ_2	σ_3	id	$t_{2\pi/3}$	$t_{4\pi/3}$
σ_2	σ_2	σ_3	σ_1	$t_{4\pi/3}$	id	$t_{2\pi/3}$
σ_3	σ_3	σ_1	σ_2	$t_{2\pi/3}$	$t_{4\pi/3}$	id

Tableau: 9.1 - Symétries du groupe diédral d'ordre 3

Nous reviendrons sur cet exemple lorsque nous introduirons un peu plus loin le concept de groupe distingué lors de notre étude des groupes de permutations et la définition des groupes distingués.

E5. Regardons un dernier exemple appliqué à la chimie en énumérant les opérations de symétries qui

laissent la molécule NH_3 (tétraèdre) invariante.

Le groupe de transformations contient 6 éléments: l'identité id , C_3 qui est la rotation de $2\pi/3$, C_3^2 la rotation de $4\pi/3$ (que nous noterons par la suite C_{-3}) toutes deux selon l'axe Z (perpendiculaire au plan XY donc...) et 3 axes $\sigma_1, \sigma_2, \sigma_3$ de symétrie/réflexion passant chacun par le milieu d'une des arêtes de base au milieu de l'arête opposée comme le montre la figure ci-dessous (pyramide vue du dessus):

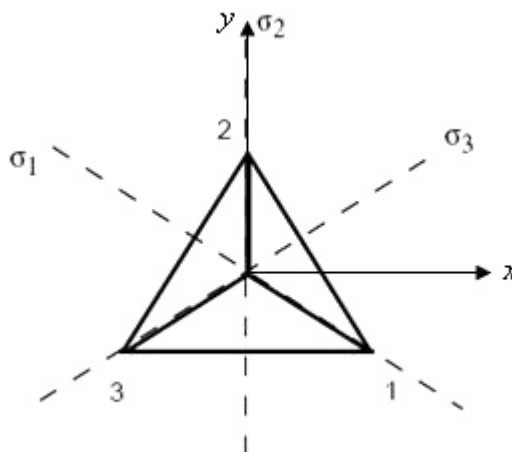


Figure: 6.6 - Opérations laissant invariant un tétraèdre

La combinaison des différents éléments de symétries montre que la table de compositions est (ce qui prouve que la loi est interne et que nous travaillons donc bien dans un groupe):

$\nearrow$	id	σ_1	σ_2	σ_3	C_3	C_{-3}
id	id	σ_1	σ_2	σ_3	C_3	C_{-3}
σ_1	σ_1	id	C_{-3}	C_3	σ_3	σ_2
σ_2	σ_2	C_3	id	C_{-3}	σ_1	σ_3
σ_3	σ_3	C_{-3}	C_3	id	σ_2	σ_1
C_3	C_3	σ_2	σ_3	σ_1	C_{-3}	id
C_{-3}	C_{-3}	σ_3	σ_1	σ_2	id	C_3

Tableau: 9.2 - Compositions de transformations du tétraèdre

Attention à l'ordre des opérations dans le tableau ci-dessus, nous appliquons d'abord l'élément de ligne puis l'élément de colonne!

Nous constatons que le groupe n'est donc pas commutatif.

1.3.1. ORBITE ET STABILISATEUR

Nous allons voir maintenant deux définitions que nous retrouverons en cristallographie (leur nom n'est pas innocent!).

Définition: L'orbite d'un élément x de E est donnée par:

$$Orb_x = \{g \cdot x, g \in G\} \quad (9.47)$$

L'orbite de x est l'ensemble des positions (dans E) susceptibles d'être occupées par l'image de x sous l'action de G . Les orbites forment évidemment une partition de E .

Exemple:

Considérons un ensemble E sur lequel agit un groupe G , par:

$$E = \{S_0, S_1, S_2, S_3, S_4, S_5\} \quad (9.48)$$

l'ensemble des 6 sommets d'un hexagone sur lequel nous faisons agir le groupe $G = \{id, t_{2\pi/3}, t_{4\pi/3}\}$.

Nous observons déjà trivialement que G est bien un groupe!

Maintenant, prenons un élément de E , par exemple S_0 .

Son orbite va donc être par définition:

$$Orb_{S_0} = \{id(S_0) = S_0, t_{2\pi/3}(S_0) = S_2, t_{4\pi/3}(S_0) = S_4\} \quad (9.49)$$

Définition: Le stabilisateur x d'un élément de E est l'ensemble:

$$Stab_x = \{g \in G \mid g \cdot x = x\} \quad (9.50)$$

des éléments qui laissent x invariant sous leur action. C'est un sous-groupe de G .

Pour reprendre notre exemple précédent. Son stabilisateur va être réduit à:

$$Stab_{S_0} = \{id\} \quad (9.51)$$

1.4. GROUPES DES PERMUTATIONS

Les groupes symétriques ont une importance non négligeable dans certains domaines de la physique quantique mais aussi en mathématiques dans le cadre de la théorie de Galois. Il convient donc d'y porter aussi une attention toute particulière.

Rappelons d'abord (cf. chapitre de Probabilités) que dans un ensemble $\{1, \dots, n\}$ il y a $n!$ permutations possibles. Les mathématiciens disent, à juste titre, qu'il y a $n!$ bijections et appellent ce nombre "**ordre du groupe de permutations**".

Prenons par exemple l'ensemble $\{1, 2, 3\}$. Cet ensemble à $3!$ permutations possibles qui sont notées dans le cadre des groupes de permutation de la manière suivante:

$$\{(1), (1\ 2), (1\ 3), (2\ 3), (1\ 2\ 3), (1\ 3\ 2)\} \quad (9.52)$$

Ce qui se lit dans l'ordre: application identité id , 1 amène sur 2 ou 2 sur un 1 (en termes de position!), 1 amène sur 3 ou 3 sur 1, 2 amène sur 3 ou 3 sur 2, 1 amène sur 2 qui amène sur 3 qui amène sur 1, 1 amène sur 3 qui amène sur 2 qui amène sur 1.

Soit de manière plus explicite:

$$\begin{aligned}
 \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} &\rightarrow (1) \\
 \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} &\rightarrow (1, 2) \\
 \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} &\rightarrow (1, 3) \\
 \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} &\rightarrow (2, 3) \\
 \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} &\rightarrow (1, 2, 3) \\
 \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} &\rightarrow (1, 3, 2)
 \end{aligned} \tag{9.53}$$

Nous pouvons observer facilement que la composition de deux permutations n'est pas commutative:

$$\begin{aligned}
 (1 \ 2) \circ (1 \ 3) &= (3, 1, 2) \\
 (1 \ 3) \circ (1 \ 2) &= (2, 3, 1)
 \end{aligned} \tag{9.54}$$

et que la composition de deux permutations est une loi interne:

$$(1 \ 2) \circ (1 \ 3 \ 2) = (1, 3, 2) = (2 \ 3) \tag{9.55}$$

avec un élément neutre qui est bien l'identité id . Nous avons donc bien un groupe non commutatif. Rappelons également au lecteur que certains éléments du groupe, s'ils sont bien choisis, peuvent former un sous-groupe. C'est l'exemple de:

$$\{(1), (1 \ 2)\} \tag{9.56}$$

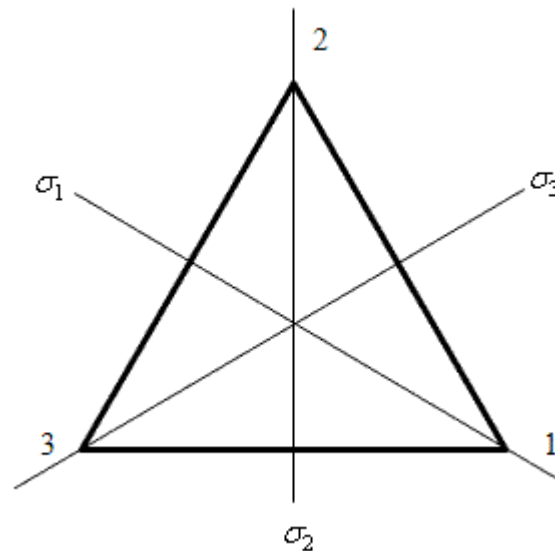
qui est un sous-groupe de $\mathcal{S}_3$ (il est facile de vérifier qu'il possède toutes les propriétés d'un groupe).

Définition: Un sous-groupe H d'un groupe G est appelé "**groupe distingué**" si, pour tout g de G et tout h de H , nous avons que ghg^{-1} est élément de H . Les mathématiciens appellent cela un "**automorphisme intérieur**"...

Voyons d'abord un exemple géométrique parlant après quoi nous reviendrons à cette définition avec $\mathcal{S}_3$.

Exemple:

Nous avons vu plus haut les éléments du groupe de symétrie diédral d'ordre 3 du triangle équilatéral. Géométriquement ils correspondent tous à des déplacements dans le plan dans lequel se trouve le triangle. Nous avons obtenu pour rappel le tableau de compositions suivant:



$\nearrow D_6$	id	$t_{2\pi/3}$	$t_{4\pi/3}$	σ_1	σ_2	σ_3
id	id	$t_{2\pi/3}$	$t_{4\pi/3}$	σ_1	σ_2	σ_3
$t_{2\pi/3}$	$t_{2\pi/3}$	$t_{4\pi/3}$	id	σ_3	σ_1	σ_2
$t_{4\pi/3}$	$t_{4\pi/3}$	id	$t_{2\pi/3}$	σ_2	σ_3	σ_1
σ_1	σ_1	σ_2	σ_3	id	$t_{2\pi/3}$	$t_{4\pi/3}$
σ_2	σ_2	σ_3	σ_1	$t_{4\pi/3}$	id	$t_{2\pi/3}$
σ_3	σ_3	σ_1	σ_2	$t_{2\pi/3}$	$t_{4\pi/3}$	id

Tableau: 9.3 - Symétries du groupe diédral d'ordre 3

D'abord, nous constatons facilement à l'aide de ce tableau que nous avons:

- Le sous-groupe formé de $\{id\}$ d'ordre 1
- Le sous-groupe formé de $\{id, t_{2\pi/3}, t_{4\pi/3}\}$ d'ordre 3
- Le sous-groupe formé de $\{id, \sigma_1\}$ d'ordre 2
- Le sous-groupe formé de $\{id, \sigma_2\}$ d'ordre 2
- Le sous-groupe formé de $\{id, \sigma_3\}$ d'ordre 2

Parmi ces 5 sous-groupes, voyons lesquels sont distingués (cela est relativement facile à visualiser à l'aide du tableau de compositions):

- Le sous-groupe formé de $\{id\}$
- Le sous-groupe formé de $\{id, t_{2\pi/3}, t_{4\pi/3}\}$

Nous allons voir maintenant une chose remarquable! En numérotant par 1, 2 et 3 les sommets du triangle équilatéral et en prenant les rotations dans le sens des aiguilles d'une montre, nous pouvons

identifier les éléments de D_6 aux éléments suivants de S_3 :

$$\begin{aligned} id &= \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} \rightarrow (1), t_{2\pi/3} = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} \rightarrow (1, 2, 3), t_{4\pi/3} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} \rightarrow (1, 3, 2) \\ \sigma_1 &= \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} \rightarrow (2, 3), \sigma_2 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} \rightarrow (1, 3), \sigma_3 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} \rightarrow (1, 2) \end{aligned} \quad (9.57)$$

et reconstruire la même table de compositions (copie de la précédente mais juste avec le changement d'écriture... hé hé!):

$\nearrow S_3$	(1)	(1 2 3)	(1 3 2)	(2 3)	(1 3)	(1 2)
(1)	(1)	(1 2 3)	(1 3 2)	(2 3)	(1 3)	(1 2)
(1 2 3)	(1 2 3)	(1 3 2)	(1)	(1 3)	(1 2)	(2 3)
(1 3 2)	(1 3 2)	(1)	(1 2 3)	(1 2)	(2 3)	(1 3)
(2 3)	(2 3)	(1 2)	(1 3)	(1)	(1 3 2)	(1 2 3)
(1 3)	(1 3)	(2 3)	(1 2)	(1 2 3)	(1)	(1 3 2)
(1 2)	(1 2)	(1 3)	(2 3)	(1 3 2)	(1 2 3)	(1)

Tableau: 9.4 - Composition du groupe distingué

Bon... ce petit interlude fermé, revenons au groupe distingué de S_3 (car il va être important pour notre introduction aux groupes de Galois) et rappelons d'abord que:

$$G = \{(1), (1\ 2), (1\ 3), (2\ 3), (1\ 2\ 3), (1\ 3\ 2)\} \quad (9.58)$$

et nous voyons que le sous-groupe distingué est formé de:

$$H = \{(1), (1\ 2\ 3), (1\ 3\ 2)\} \quad (9.59)$$

Définition: Pour tout sous-groupe H stable par les automorphismes intérieurs d'un groupe G , nous appelons "**indice de H dans G** " le quotient de l'ordre du groupe G par l'ordre du sous-groupe H et nous l'écrivons $[G/H]$.

Par exemple, l'indice du sous-groupe $\{(1), (1\ 2)\}$ dans le groupe S_3 est $6/2$ c'est-à-dire 3. Ce concept nous sera très utile lors de notre introduction aux corps de Galois plus loin.

Considérons maintenant, la permutation particulière σ pour aborder le sujet sous un angle différent mais équivalent:

$$\{1, 2, 3, 4\} \xrightarrow{\sigma} \{3, 2, 4, 1\} \quad (9.60)$$

Les mathématiciens ont pour habitude de noter cela, dans un premier temps, sous la forme:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix} \quad (9.61)$$

avec:

$$\sigma(1) = 3, \sigma(2) = 2, \sigma(3) = 4, \sigma(4) = 1 \quad (9.62)$$

Etant donné σ et τ , deux permutations, il est naturel de regarder leur composition $\tau \circ \sigma$ (rappelons que cela signifie d'abord σ , puis τ comme pour la composition de fonctions).

Ainsi, si:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix} \text{ et } \tau = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} \quad (9.63)$$

Alors:

$$\tau \circ \sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 3 & 2 \end{pmatrix} \quad (9.64)$$

et:

$$\sigma \circ \tau = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 1 & 4 \end{pmatrix} \quad (9.65)$$

Maintenant, l'idée est d'interpréter la composition comme une multiplication de permutations. Cette multiplication est alors non-commutative comme nous venons de le constater dans l'exemple précédent. Nous avons en général $\sigma \circ \tau \neq \tau \circ \sigma$.

Chaque bijection a un inverse (une fonction réciproque). Dans notre exemple il s'agit de évidemment de:

$$\sigma^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 2 & 1 & 3 \end{pmatrix} \quad (9.66)$$

Géométriquement, pour calculer l'inverse σ^{-1} d'un élément σ , il suffit de prendre la réflexion du dessin de σ dans un axe horizontal comme le montre la partie gauche de la figure ci-dessous:

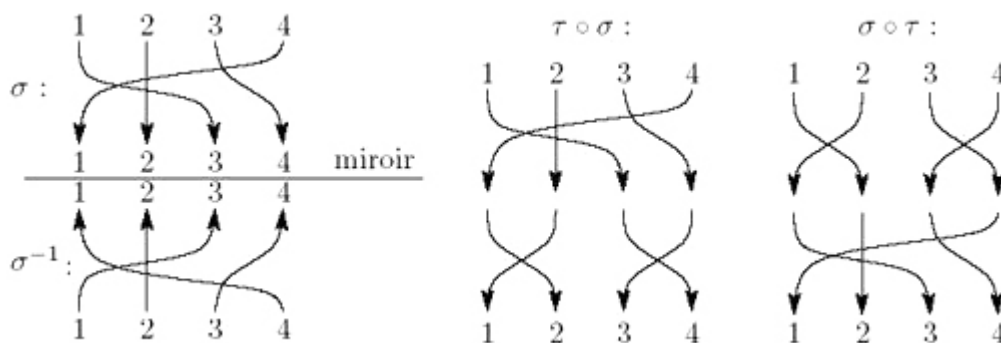


Figure: 6.7 - Exemples de composées et d'inverses de permutations

Définitions:

D1. L'ensemble des permutations d'un ensemble avec n éléments, muni de cette structure de multiplication, s'appelle le "**groupe des permutations d'ordre n** " ou "**groupe des substitutions d'ordre n** ", et se note S_n ou encore $S(n)$.

D2. Nous disons qu'un élément σ de $\mathcal{S}_n$ est un "**cycle d'ordre k** ", ou un " **k -cycle**", s'il existe $a_1, a_2, \dots, a_k \in \{1, \dots, n\}$ tel que:

- σ envoie a_1 sur a_2 , a_2 sur a_3 , ..., a_{k-1} sur a_k , et a_k sur a_1
- σ fixe tous les autres éléments de $\mathcal{S}_n$

et nous notons le cycle de la manière $\sigma = (a_1, a_2, \dots, a_k)$.

Pour mieux comprendre reprenons notre exemple de $\mathcal{S}_4$:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix} \quad (9.67)$$

Ce groupe symétrique est un 3-cycle noté $\sigma = (1 \ 3 \ 4)$ car dans l'ordre: 1 envoie sur 3, 3 envoie sur 4 et 4 envoie sur 1 (et le 2 n'étant pas mentionné il reste fixe). Nous pouvons noter cela aussi des façons suivantes équivalentes: $\sigma = (3 \ 4 \ 1)$ ou encore $\sigma = (4 \ 1 \ 3)$.

Définition: L'ordre d'un k -cycle est k (d'où le nom!).

Effectivement si nous reprenons $\sigma = (1 \ 3 \ 4)$, nous avons alors:

$$\sigma^2 = (1 \ 4 \ 3) \text{ et } \sigma^3 = id \quad (9.68)$$

Définition: Nous disons qu'une permutation σ est un "**cycle**" s'il existe $k \in \mathbb{N}$ tel que σ est un k -cycle.

Attention! Toute permutation doit s'écrire comme un produit de cycles disjoints (c'est-à-dire qu'un nombre qui apparaît dans un cycle ne doit pas apparaître dans un autre cycle). Par exemple, dans $\mathcal{S}_9$, nous avons:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 5 & 2 & 8 & 3 & 7 & 1 & 6 & 4 & 9 \end{pmatrix} = (1 \ 5 \ 7 \ 6)(3 \ 8 \ 4) \quad (9.69)$$

Donc cette permutation est un produit d'un 4-cycle et d'un 3-cycle disjoint.

Nous laisserons d'ailleurs le lecteur vérifier par lui-même que l'ordre de ce groupe $\mathcal{S}_9$ est 12...

Remarque: Les mathématiciens peuvent démontrer que si σ est un élément qui a une décomposition en c cycles disjoints de longueur $n_1, n_2, \dots, n_c$ alors l'ordre de σ est le plus petit commun multiple des ordres de tous les cycles disjoints qui le composent.

Nous supposons également intuitif que dans le vocabulaire commun, un 2-cycle dans $\mathcal{S}_n$ s'appelle aussi une "**transposition**".

Allons un petit peu plus loin. Nous nous proposons de montrer par l'exemple que l'ensemble des

transpositions engendre $\mathcal{S}_n$. Autrement, dit, toute permutation s'écrit comme un produit de transpositions.

Reprenons notre exemple (il s'agit d'une permutation paire):

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix} = (1 \ 3 \ 4) = (1 \ 3)(4 \ 1) \quad (9.70)$$

En général, un k -cycle s'écrit donc comme produit de $k-1$ transpositions.

Définition: Soit $\sigma \in \mathcal{S}_n$ une permutation. Nous disons que σ est "**permutation paire**" si, dans une écriture de σ comme produit de transpositions, il y a un nombre pair de transpositions. Nous disons que σ est "**permutation impaire**" si, dans une écriture de σ comme produit de transpositions, il y a un nombre impair de transpositions.

Finissons par un petit complément... Nous avons que $\mathcal{S}_3$ est un groupe des permutations d'ordre 3 avec donc $3!=6$ permutations possibles.

Si nous énumérons les 6 permutations nous avons vu que nous obtenons:

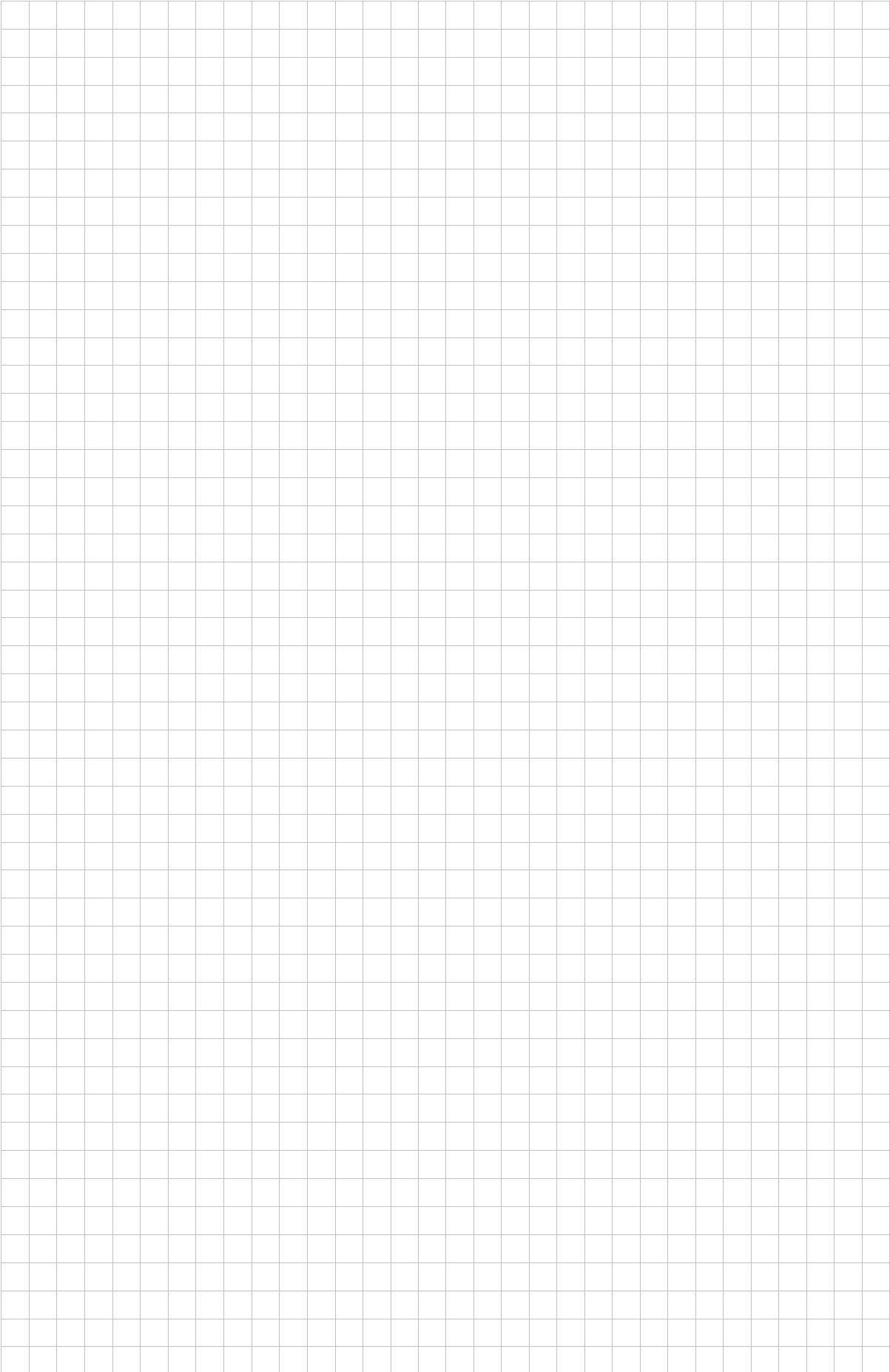
$$\{(1), (1 \ 2), (1 \ 3), (2 \ 3), (1 \ 2 \ 3), (1 \ 3 \ 2)\} \quad (9.71)$$

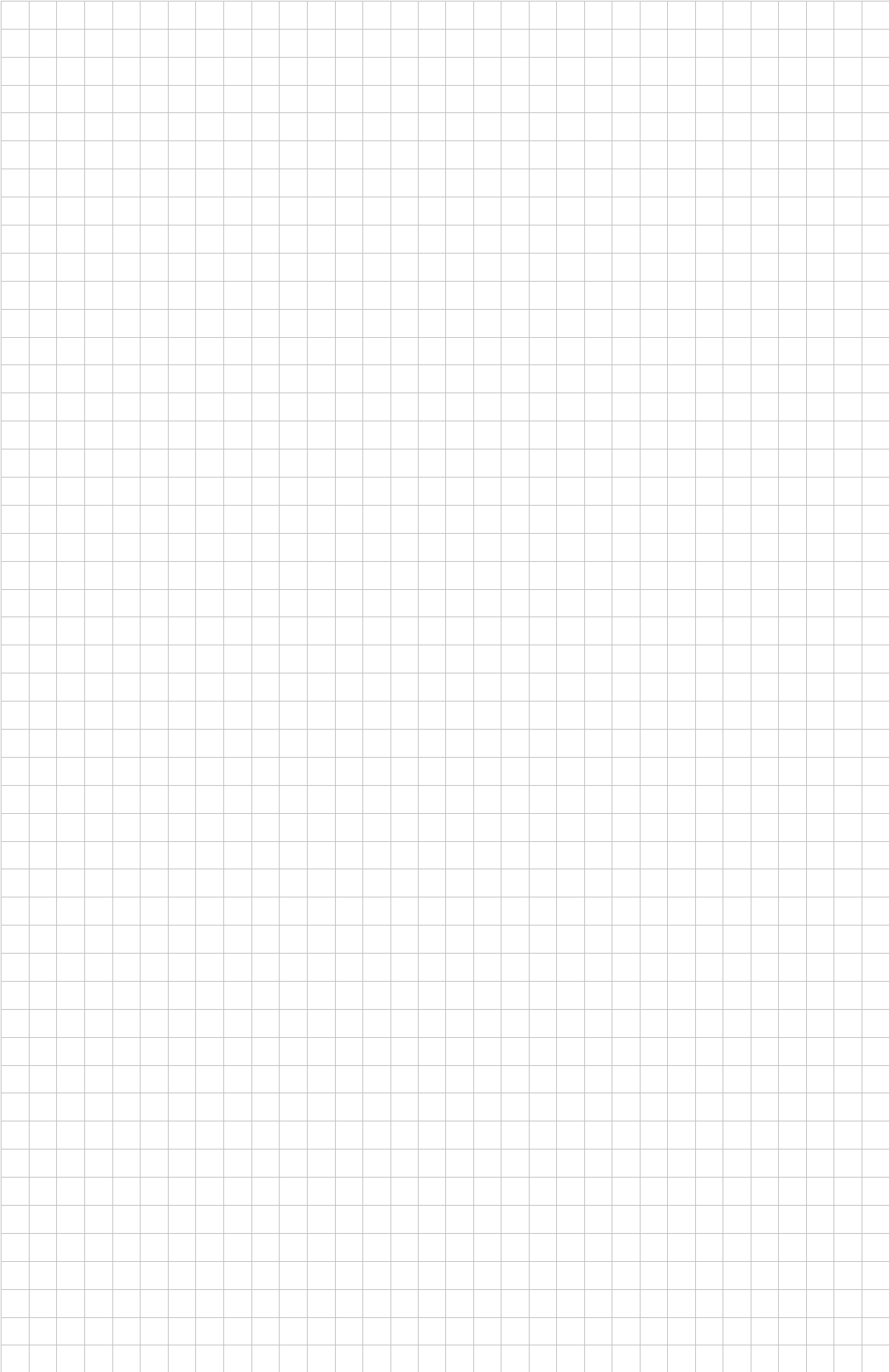
Parmi celles-ci certaines seulement peuvent être écrites comme un produit pair de transpositions:

$$(1 \ 2 \ 3) = (1 \ 2)(3 \ 1) \text{ et } (1 \ 3 \ 2) = (1 \ 3)(2 \ 1) \quad (9.72)$$

Les permutations paires forment avec la permutation identité id , un sous-groupe (non commutatif) que nous appelons le "**groupe alterné d'ordre n** " et que nous notons A_n . C'est facile de le vérifier avec l'exemple précédent.







10. CALCUL DIFFÉRENTIEL ET INTÉGRAL

Le calcul différentiel est un des domaines les plus passionnants et vastes de la mathématique, et il existe une littérature considérable (colossale) sur le sujet. Les résultats initiés par des scientifiques comme Fermat, Newton, Leibniz, Euler et compagnie depuis la fin du 17ème siècle retrouvent des implications dans absolument tous les domaines de la physique, de l'informatique, de l'électronique, de la chimie, de la finance, de la biologie et de la mathématique elle-même.

Les mathématiciens ont rédigé une telle quantité de théorèmes depuis sa naissance au milieu du 16ème siècle sur le sujet que la validation d'un échantillon de ceux-ci est parfois délicate car nécessitant à eux-seuls la vie d'un homme pour être parcourus (c'est un problème que la communauté des mathématiciens reconnaît) et vérifiés (ce qui fait que parfois personne ne les vérifie...).

Ce constat fait, nous avons choisi de ne présenter ici que les points absolument nécessaires à la compréhension des outils fondamentaux de l'ingénieur. Les puristes nous excuseront donc pour l'instant de ne pas présenter certains théorèmes qui peuvent leur sembler indispensables mais que nous rédigerons une fois le temps venu...

Nous allons principalement étudier dans ce qui va suivre ce que les mathématiciens aiment bien préciser (et ils ont raison): les cas généraux des fonctions réelles à une variable réelle. Les fonctions plus complexes (à plusieurs variables réelles ou complexes, continues ou discrètes) viendront une fois cette partie terminée.

Remarque: Nous ne nous attarderons pas à démontrer les dérivées et primitives de toutes les fonctions car comme il y a une infinité de fonctions possibles, il y a également une infinité de dérivées et de primitives. C'est le rôle des professeurs dans les instituts scolaires d'entraîner les élèves à appliquer et à comprendre le raisonnement de dérivation et d'intégration par des applications sur des fonctions connues (l'Internet ne remplacera très probablement jamais l'école à ce niveau).

1. CALCUL DIFFÉRENTIEL

Soit une fonction f réelle à une variable réelle x notée $f(x)$ (nous nous limitons à ce cas de figure pour l'instant et étudierons les dérivées partielles dans des espaces à un nombre de dimensions quelconques plus loin) continue au moins dans un intervalle où se situe l'abscisse a .

Définitions:

D1. Nous appelons "**pende moyenne**", ou encore "**coefficient directeur**" le rapport de la projection orthogonale de deux points $x_1 \neq x_2$ de la fonction f non nécessairement continue sur l'axe des abscisses et des ordonnées tel que:

$$\frac{\Delta f}{\Delta x} = \frac{f(x_2) - f(x_1)}{x_2 - x_1} = \frac{f(x_1 + h) - f(x_1)}{(x_1 + h) - (x_1)} \underset{x_1 = a}{=} \frac{f(a + h) - f(a)}{h} \quad (10.1)$$

Ce qui se représente sous forme graphique de la manière suivante avec une fonction particulière:

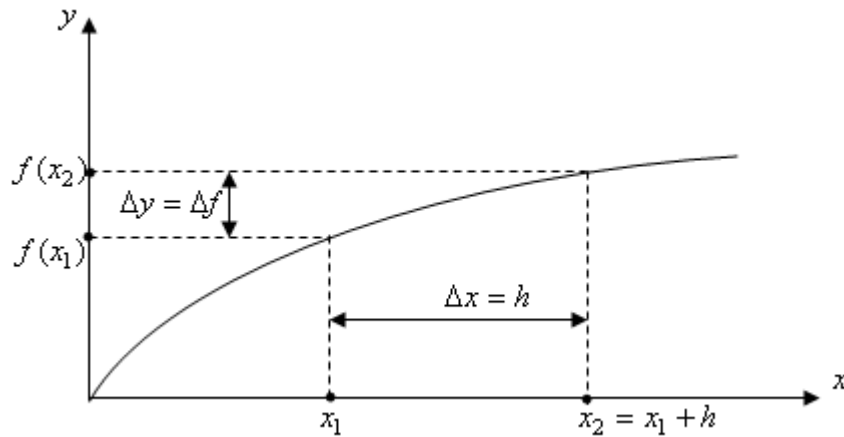


Figure: 10.1 - Exemple de calcul de la pente moyenne

Remarque: Δ signifiant "un delta" exprime le fait que nous sous-entendons une différence d'une même quantité.

Nous supposons comme évident (sans démonstration) que deux fonctions dont les pentes sont les mêmes dans un même intervalle de définition, y sont parallèles (ou confondues).

Nous démontrerons dans le chapitre de Géométrie Analytique que deux fonctions dont la multiplication des pentes vaut -1 sont perpendiculaires.

D2. Nous appelons "**nombre dérivé en a**" ou "**pente instantanée**" ou encore "**dérivée première**", la limite quand h tend vers 0 (si elle existe) du rapport de la projection orthogonale de deux points $x_1 \neq x_2$ infiniment proches de la fonction f continue (dans le sens qu'elle ne contient pas de "trous") sur l'axe des abscisses et des ordonnées tel que:

$$\frac{df_a}{dx} = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} = f'(a) \quad (10.2)$$

Une interprétation graphique donne donc bien que $f'(a)$ est le coefficient directeur (la pente de la tangente au point d'abscisse a).

Remarques:

R1. d signifiant un "**différentiel**" exprime le fait que nous sous-entendons une différence infiniment petite d'une même quantité.

R2. Nous renvoyons le lecteur au chapitre d'Analyse Fonctionnelle pour la définition de ce qu'est une fonction continue.

D3. Soit f une fonction définie sur un intervalle I et dérivable en tout point a de I , la fonction qui à tout réel a de I associe le nombre $f'(a)$ est appelée "**fonction dérivée de f sur I** " et est notée f' .

Remarque: Au niveau des notations les physiciens adoptent suivant leur humeur différentes notations possibles pour les dérivées. Ainsi, considérons la fonction réelle à une variable $f(x)$, vous trouverez dans la littérature ainsi que dans le présent site les notations suivantes pour la dérivée première:

$$\frac{d}{dx}f(x) = \frac{df(x)}{dx} = d_x f(x) = \dot{f}(x) = f'(x) \quad (10.3)$$

ou encore en considérant implicitement que f est fonction de x (ceci permet d'alléger un petit peu la tailles des développements):

$$\boxed{\frac{d}{dx}f = \frac{df}{dx} = d_x f = \dot{f} = f'} \quad (10.4)$$

Nous pouvons de la même manière définir les dérivées d'ordre 2 (dérivée d'une dérivée), les dérivées d'ordre 3 (dérivée d'une dérivée d'ordre 2) et ainsi de suite. Nous rencontrerons par ailleurs très fréquemment de telles dérivées en physique (et même en maths pour l'analyse fonctionnelle).

Précisons que les dérivées d'ordre 2 ont une interprétation très importante en physique et dans le domaines de l'optimisation (cf. chapitre de Méthodes Numériques). Effectivement, si le signe de la dérivée première est positif puis devient négatif quand x croît, alors nous devinons facilement que nous parcourons un maximum local d'une fonction (point où la dérivée est nulle) et que si le signe de la dérivée première est négatif puis devient positif quand x croît, alors nous parcourons un minimum local de la fonction (point où la dérivée est aussi nul). En d'autres termes, lorsque la pente change de signe (s'annule en changeant de signe) la fonction passe par un extremum (minimum ou maximum) et la tangente est "horizontale" en ce point: parallèle à l'axe des abscisses. Par contre, lorsque la dérivée d'ordre 2 (dérivée secondes) est nulle, c'est que la courbure de la fonction s'inverse. Nous parlons alors de "**point d'inflexion**".

Voici un exemple très ludique d'une fonction avec sa dérivées première et seconde avec Maple 4.00:

```
>plot([tanh(x),diff(tanh(x),x),diff(tanh(x),x$2)],x=-5..5,color=[red,green,blue]);
```

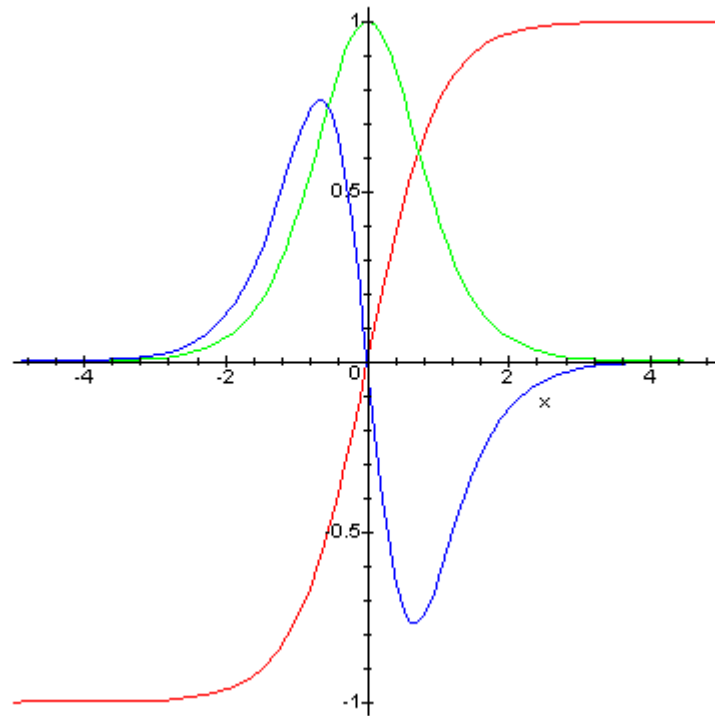


Figure: 10.2 - Plot de la fonction tangente hyperbolique, sa dérivée première et seconde

Maintenant, suite à un problème de compréhension de la part d'un lecteur dans un des chapitres du site, précisons une technique utilisée fréquemment par les physiciens. Considérons une dérivée d'ordre 2 telle que:

$$\ddot{f} = f'' = \frac{d^2 f(x)}{dx^2} = \frac{d}{dx} \left(\frac{df(x)}{dx} \right) = \frac{d}{dx} \left(\frac{d}{dx} f(x) \right) \quad (10.5)$$

Si nous regardons le d/dx comme un opérateur différentiel nous pouvons bien évidemment écrire:

$$\frac{d}{dx} \left(\frac{d}{dx} f(x) \right) = \frac{d}{dx} \frac{d}{dx} (f(x)) = \left(\frac{d}{dx} \right)^2 f(x) \quad (10.6)$$

Finalement nous avons:

$$\frac{d^2 f(x)}{dx^2} = \left(\frac{d}{dx} \right)^2 f(x) \quad (10.7)$$

et donc il vient après simplification par $f(x)$:

$$\boxed{\frac{d^2}{dx^2} = \left(\frac{d}{dx} \right)^2} \quad (10.8)$$

sinon quoi nous ne pouvons pas avoir cette égalité si l'opérateur agit explicitement sur une fonction dans une relation mathématique ou physique quelconque.

Cela peut paraître évident pour certains mais parfois moins pour d'autres... et il était visiblement utile de préciser cela car c'est souvent utilisé dans les chapitres de Relativité Restreinte, Relativité Générale, Physique Quantique Corpusculaire et Physique Quantique Ondulatoire.

Indiquons et démontrons maintenant deux propriétés intuitivement évidentes des dérivées et qui nous seront plusieurs fois indispensables pour certaines démonstrations sur ce site (comme par exemple dans le chapitre de méthodes numériques ou ici même...).

Considérons d'abord deux nombres réels $a < b$ et f une fonction à valeurs réelles continue sur $[a,b]$ et dérivable sur $]a,b[$ telle que $f(a) = f(b)$. Alors nous voulons démontrer qu'il existe bien évidemment au moins un élément c de $]a,b[$ tel que $f'(c) = 0$ (c'est typiquement le cas des fonctions polynômiales!).

Cette propriété est appelée "**théorème de Rolle**" et donc explicitement elle montre qu'il existe au moins un élément où la dérivée de f est nulle si en la parcourant nous revenons à la même valeur des images pour deux valeurs distinctes des abscisses, c'est-à-dire qu'il existe au moins un point où la tangente est horizontale.

Démonstration:

Si f est constante, c'est immédiat...

Dans le cas contraire, comme f est continue sur l'intervalle fermé borné $[a,b]$ elle admet au moins un minimum global ou maximum global compte tenu que nous nous basons sur l'hypothèse que $f(a) = f(b)$ et que f n'est pas constante. L'extrema est atteint en un point c appartenant à l'intervalle ouvert $]a, b[$ (le fait de prendre l'intervalle ouvert permet dans certains cas d'éviter d'avoir un extrema à nouveau en a ou en b).

Supposons comme premier cas que $f(c)$ est maximum global. La dérivée de la fonction f entre c et un deuxième point a alors un signe connu.

Pour h strictement positif et tel que $c+h$ appartienne à l'intervalle $[a,b]$:

$$\frac{f(c+h) - f(c)}{h} < 0 \quad (10.9)$$

En considérant la limite quand h tend vers 0, le nombre dérivé $f'(c)$ est négatif.

Pour h strictement négatif et tel que $c+h$ appartienne à l'intervalle $[a,b]$:

$$\frac{f(c+h) - f(c)}{h} > 0 \quad (10.10)$$

En considérant la limite quand h tend vers 0, le nombre dérivé $f'(c)$ est positif.

Au bout du compte, la dérivée de f est nulle au point c .

La démonstration est analogue si $f(c)$ est un minimum global, avec les signes des dérivées qui sont les opposés.

□C.Q.F.D.

Maintenant, considérons deux réels $a < b$ et $f(x)$ une fonction continue sur $[a, b]$ et dérivable sur $]a, b[$. Alors, nous nous proposons de montrer qu'il existe au moins un réel $c \in]a, b[$ tel que:

$$f(b) - f(a) = f'(c)(b - a) \quad (10.11)$$

Ce qui peut aussi s'écrire sous la forme suivante:

$$f(x_0 + k) - f(x_0) = f'(x_0 + s \cdot k)k \quad (10.12)$$

avec $s \in]0, 1[$.

Géométriquement cela signifie qu'en au moins un point c du graphe de la fonction $f(x)$, il existe une tangente de coefficient directeur:

$$\frac{f(b) - f(a)}{b - a} \quad (10.13)$$

Graphiquement cela donne:

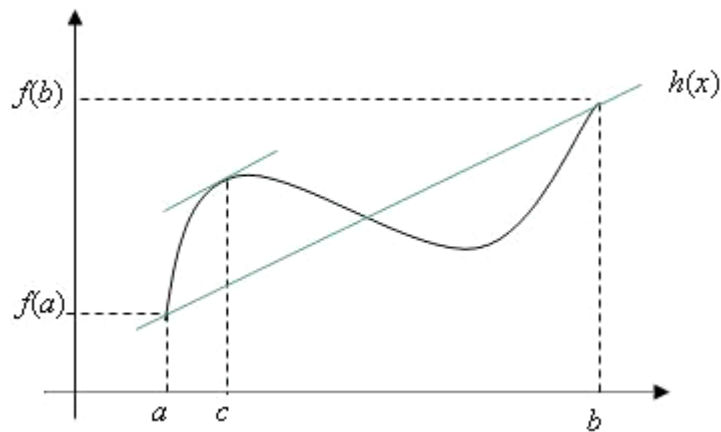


Figure: 10.3 - Représentation graphique du théorème de Rolle

Démonstration:

Nous avons d'abord:

$$h(x) = (x - a) \cdot \frac{f(b) - f(a)}{b - a} + f(a) \quad (10.14)$$

car la pente de $h(x)$ est bien évidemment $(f(b) - f(a)) / (b - a)$ et comme lorsque $x = a$ nous devons avoir $f(a)$ il s'ensuit donc la relation donnée précédemment.

Ensuite, pour démontrer qu'un tel point c existe, l'idée est de rapporter les deux points a et b à la même ordonnée ce qui en fait nous ramène au théorème de Rolle et pour cela, nous définissons une fonction g par:

$$g(x) = f(x) - h(x) = f(x) - \left[(x - a) \cdot \frac{f(b) - f(a)}{b - a} + f(a) \right] \quad (10.15)$$

qui est telle qu'effectivement $g(b) = g(a) \dots$ et en l'occurrence égal à 0 (mais cette valeur importe peu). Dès lors, le théorème de Rolle vu précédemment nous indique qu'il existe un point entre a et b où la dérivée de $g(x)$ est nulle tel que $g'(c) = 0$. Et en constatant que:

$$g'(x) = f'(x) - \frac{f(b) - f(a)}{b - a} \quad (10.16)$$

nous obtenons:

$$g'(c) = f'(c) - \frac{f(b) - f(a)}{b - a} = 0 \quad (10.17)$$

Soit après simplification:

$$f(b) - f(a) = f'(c)(b - a) \quad (10.18)$$

□C.Q.F.D.

Puisque le terme de gauche représente un accroissement fini du terme de droite, alors ce résultat est appelée "**théorème des accroissements finis**" (TAF).

A l'aide de ce petit théorème et des outils mathématiques introduits précédemment, nous pouvons construire un petit théorème fort utile et puissant en physique.

Définition: Nous appelons "**règle de L'Hôpital**" (également appelée "**règle de l'Hospital**" ou "**règle de Bernoulli**") le procédé qui utilise la dérivée dans le but de déterminer les limites difficiles à calculer de la plupart des quotients et qui apparaissent souvent en physique.

Démonstration:

Considérons deux fonctions $f(x)$ et $g(x)$ et telles que $f(a) = g(a) = 0$ alors nous pouvons écrire:

$$\frac{f(x)}{g(x)} = \frac{f(x) \frac{x-a}{x-a}}{g(x) \frac{x-a}{x-a}} = \frac{f(x) \frac{x-a}{x-a}}{\frac{x-a}{x-a} g(x)} = \frac{f(x) - f(a)}{x-a} \frac{x-a}{g(x) - g(a)} \quad (10.19)$$

Alors selon la définition de la dérivée:

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x-a} \frac{x-a}{g(x) - g(a)} = \frac{f(a+h) - f(a)}{(a+h) - a} \frac{(a+h) - a}{g(a+h) - g(a)} = \frac{f'(a)}{g'(a)} \quad (10.20)$$

□C.Q.F.D.

Nous pouvons généraliser ce résultat précédent initialement basé sur la contrainte un peu trop forte:

$$f(a) = g(a) = 0 \quad (10.21)$$

Démonstration:

Rappelons donc que selon le théorème des accroissements finis, si $f(x)$ est dérivable sur un intervalle $]a, b[$ et continue sur $[a, b]$ alors il existe un réel c dans l'intervalle $[a, b]$ tel que:

$$f'(c) = \frac{f(b) - f(a)}{b - a} \quad (10.22)$$

Si le théorème se vérifie pour deux fonctions satisfaisant aux mêmes contraintes alors nous avons deux fonctions telles que:

$$f'(c) = \frac{f(b) - f(a)}{b - a} \quad \text{et} \quad g'(c) = \frac{g(b) - g(a)}{b - a} \quad (10.23)$$

Si $g'(c)$ est non nul nous avons alors tout à fait le droit d'écrire le rapport (certains appellent cela le "théorème des accroissements fini généralisé"...):

$$\frac{f'(c)}{g'(c)} = \frac{f(b) - f(a)}{b - a} \frac{b - a}{g(b) - g(a)} = \frac{f(b) - f(a)}{g(b) - g(a)} \quad (10.24)$$

ce qui sans perdre en validité tant que c est dans l'étau $[a, x]$ peut s'écrire:

$$\frac{f'(c)}{g'(c)} = \frac{f(x) - f(a)}{g(x) - g(a)} \quad (10.25)$$

Ainsi, lorsque $x \rightarrow a$ ce qui implique que l'étau $[a, x]$ se referme et donc $c \rightarrow a$ nous avons:

$$\lim_{x \rightarrow a} \frac{f'(c)}{g'(c)} = \lim_{x \rightarrow a} \frac{f(x) - f(a)}{g(x) - g(a)} = \frac{f(a+h) - f(a)}{g(a+h) - g(a)} = \frac{f(a+h) - f(a)}{(a+h) - a} \frac{(a+h) - a}{g(a+h) - g(a)} = \frac{f'(a)}{g'(a)} \quad (10.26)$$

Ainsi, nous venons de prouver quand dans la démonstration précédente de la règle de l'Hôpital la relation:

$$\frac{f(a+h) - f(a)}{(a+h) - a} \frac{(a+h) - a}{(a+h) - g(a)} = \frac{f'(a)}{g'(a)} \quad (10.27)$$

que nous avons est vraie en toute généralité et qu'il n'est pas nécessaire que $f(a) = g(a) = 0$ soit vrai pour que le résultat soit juste!

□C.Q.F.D.

1.1. DIFFÉRENTIELLES

Nous avons indiqué précédemment ce qu'était un différentiel d . Mais il existe en fait plusieurs types de sortes de différentielles d'une fonction (remarquez que nous distinguons le genre masculin et féminin du terme):

1. Les différentiels
2. Les différentielles partielles
3. Les différentielles totales exactes
4. Les différentielles totales inexactes

Rappelons que nous appelons "différentiel df " d'une fonction f à une variable la relation donnée par (voir texte précédent):

$$\boxed{df = f' dx} \quad (10.28)$$

Cependant, pour exprimer l'effet d'un changement de toutes les variables d'une fonction f de plusieurs variables, nous devons utiliser un autre type de différentiel que nous appelons la "**différentielle totale**" (dérivée en deux sous-familles: différentielle totale exacte et différentielle totale inexacte).

Soit par exemple, une fonction $f(x, y)$ des deux variables x et y . L'accroissement df de la fonction f , pour un accroissement fini de x à $x + \Delta x$ et de y à $y + \Delta y$ est:

$$\Delta f = f(x + \Delta x, y + \Delta y) - f(x, y) \quad (10.29)$$

que nous pouvons aussi écrire:

$$\Delta f = f(x + \Delta x, y + \Delta y) - f(x, y + \Delta y) + f(x, y + \Delta y) - f(x, y) \quad (10.30)$$

ou encore:

$$\Delta f = \frac{f(x + \Delta x, y + \Delta y) - f(x, y + \Delta y)}{\Delta x} \Delta x + \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y} \Delta y \quad (10.31)$$

Pour des accroissements infiniment petits de x et y :

$$df = \lim_{\Delta x \rightarrow 0} \left(\frac{f(x + \Delta x, y + \Delta y) - f(x, y + \Delta y)}{\Delta x} \Delta x \right) + \lim_{\Delta y \rightarrow 0} \left(\frac{f(x, y + \Delta y) - f(x, y)}{\Delta y} \Delta y \right) \quad (10.32)$$

Intéressons-nous dès lors aux deux termes au passage à la limite:

$$\frac{f(x + dx, y + dy) - f(x, y + dy)}{dx} \text{ et } \frac{f(x, y + dy) - f(x, y)}{dy} \quad (10.33)$$

Le premier terme de gauche, nous le voyons, ne donne finalement que la variation en x de la fonction $f(x, y)$ en ayant y constant sur la variation. Nous notons cela dès lors (si la connaissance des variables constantes est triviale, nous ne les indiquons plus):

$$\left(\frac{\partial f}{\partial x} \right)_y = \frac{\partial f}{\partial x} = \frac{f(x + dx, y + dy) - f(x, y + dy)}{dx} \quad (10.34)$$

et de même:

$$\left(\frac{\partial f}{\partial y} \right)_x = \frac{\partial f}{\partial y} = \frac{f(x, y + dy) - f(x, y)}{dy} \quad (10.35)$$

Remarque: Quand une variable est fixée pour étudier la variation de l'autre, certains auteurs ou professeurs des anciennes générations aiment à dire: "toutes choses égales par ailleurs f varie en fonction de ... de façon". Bref, c'est un usage que l'on retrouve dans d'autres domaines (comme les régression linéaires à plusieurs variables explicatives) mais qui se perd...

Les deux expressions:

$$\left(\frac{\partial f}{\partial y}\right)_x \text{ et } \left(\frac{\partial f}{\partial x}\right)_y \quad (10.36)$$

sont ce que nous appelons des "**différentielles partielles**" ou plus simplement "**dérivée partielle**".

Il vient dès lors:

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy \quad (10.37)$$

qui est la "**différentielle de f** ". Les thermodynamiciens parlent souvent eux de la "**différentielle totale exacte de f** " ou plus simplement "**différentielle exacte de f** ".

La relation précédente est un cas particulier de ce que les mathématiciens appellent en toute généralité une "**forme différentielle**":

$$df = M(x, y)dx + N(x, y)dy \quad (10.38)$$

nous y reviendrons un peu plus loin... Il est d'usage de noter:

$$\vec{F} = \begin{pmatrix} M(x, y) \\ N(x, y) \end{pmatrix} \quad (10.39)$$

donc sous forme d'un champ vectoriel.

Il est important de se rappeler de la forme de la différentielle totale car nous la retrouverons partout dans des opérateurs particuliers en physique, dans la mécanique des fluides, dans la thermodynamique, etc.

Géométriquement, les dérivées partielles peuvent être interprétées comme suit: la fonction $f(x, y)$ définit une surface dans $\mathbb{R}^3$, dont l'intersection avec la plan $y = y_0$ est une courbe:

$$x \mapsto f(x, y_0) \quad (10.40)$$

La dérivée partielle $\partial_x f$ est alors la pente de cette courbe en tout point x . Nous avons alors naturellement la fonction suivante pour la pente au point (x_0, y_0) :

$$f(x, y_0) = f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) \quad (10.41)$$

De la même manière, la tangente à la courbe:

$$y \mapsto f(x_0, y) \quad (10.42)$$

aura pour expression:

$$f(x_0, y) = f(x_0, y_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0) \quad (10.43)$$

Le plan localement tangent au point (x_0, y_0) déterminé par ses deux tangentes est alors donné par:

$$f(x, y) = f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0) \quad (10.44)$$

où nous reconnaissons la forme de la différentielle totale exacte en réarrangeant les termes:

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy \quad (10.45)$$

Ainsi, par exemple, la surface représentée par la fonction:

$$f = e^{-x^2 - y^2} \quad (10.46)$$

est représentée ci-dessous avec les deux tangentes passant par le point:

$$(x_0, y_0) = (0.3, 0.1) \quad (10.47)$$

et dont les équations respectives sont:

$$\begin{aligned} f(x, y_0) &= f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) = e^{-x^2 - y^2} \Big|_{0.3, 0.1} - 2xe^{-x^2 - y^2} \Big|_{0.3, 0.1} (x - x_0) \\ &\cong 0.905 - 0.543(x - 0.3) = -0.543x + 1.0679 \end{aligned} \quad (10.48)$$

et:

$$\begin{aligned} f(x_0, y) &= f(x_0, y_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0) = e^{-x^2 - y^2} \Big|_{0.3, 0.1} - 2ye^{-x^2 - y^2} \Big|_{0.3, 0.1} (y - y_0) \\ &\cong 0.905 - 0.181(y - 0.1) = -0.181y + 0.9231 \end{aligned} \quad (10.49)$$

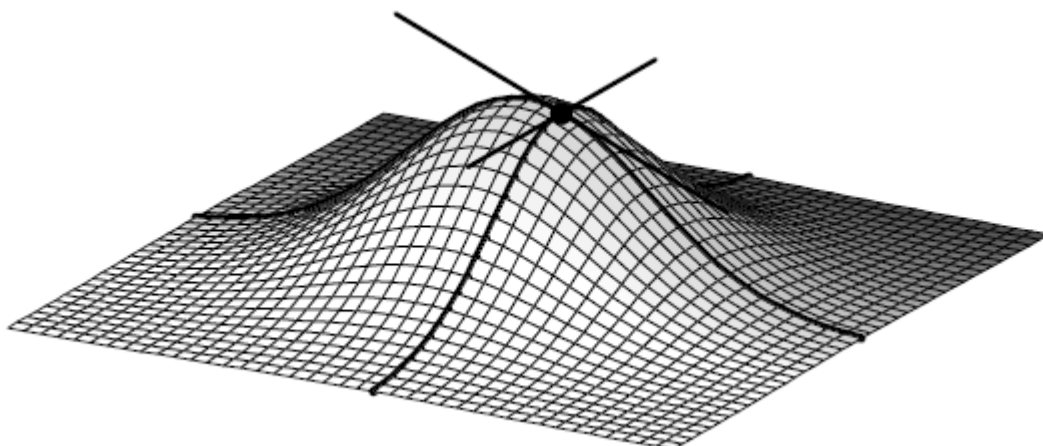


Figure: 10.4 - Les deux tangentes de la fonction au point d'intérêt

Nous avons le plan tangent en ce point qui est alors donné par:

$$\begin{aligned} f(x, y) &= f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0) \\ &\cong 0.905 - 0.543(x - 0.3) - 0.181(y - 0.1) = -0.543x - 0.181y + 1.086 \end{aligned} \quad (10.50)$$

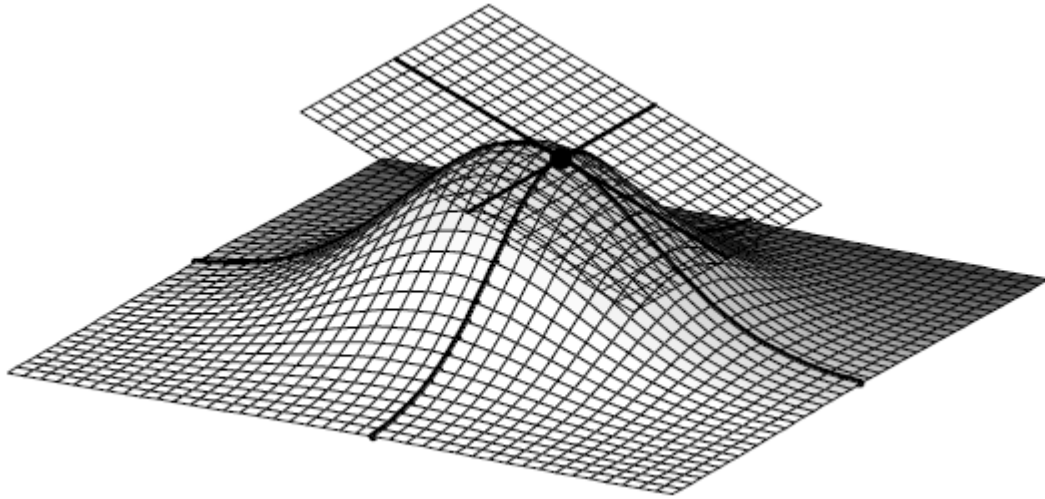


Figure: 10.5 - Les deux tangentes de la fonction au point d'intérêt avec le plan tangent

Remarque: De la même manière, pour une fonction de plus de deux variables, par exemple $f(x, y, z)$, la différentielle totale df est:

$$df = \left(\frac{\partial f}{\partial x} \right)_{y,z} dx + \left(\frac{\partial f}{\partial y} \right)_{x,z} dy + \left(\frac{\partial f}{\partial z} \right)_{x,y} dz = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \frac{\partial f}{\partial z} dz \quad (10.51)$$

Dans l'équation ci-dessus, la différentielle df a été calculée à partir de l'expression de la fonction f . Puisqu'il existe une fonction f qui vérifie l'expression de df , la différentielle df est dite alors aussi "**totale exacte**".

Profitons pour faire une indication importante sur l'utilisation des dérivées partielles par les physiciens (et donc dans les nombreux chapitres y relatifs du site). Nous avons vu plus haut que si f dépend de deux variables x, y nous avons:

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy \quad (10.52)$$

et s'il ne dépend que d'une variable nous avons alors:

$$df = \frac{\partial f}{\partial x} dx \quad (10.53)$$

et alors:

$$\frac{df}{dx} = \frac{\partial f}{\partial x} \quad (10.54)$$

raison pour laquelle les physiciens mélangent allègrement les deux notations...

Maintenant, il faut cependant savoir qu'il existe également des différentielles totales exactes qu'aucune fonction ne vérifie. Dans ce cas, nous parlons de "**différentielle totale inexacte**" et pour déterminer si une différentielle totale est exacte ou inexacte, nous utilisons les propriétés des dérivées partielles (cas très important en thermodynamique!!!).

Soit la fameuse forme différentielle générale (cela fait appel à de la géométrie différentielle):

$$dz = M(x, y)dx + N(x, y)dy \quad (10.55)$$

où $M(x, y)$ et $N(x, y)$ sont des fonctions des variables x et y . Si dz est une différentielle totale exacte, alors:

$$dz = \frac{\partial z}{\partial x} dx + \frac{\partial z}{\partial y} dy \quad (10.56)$$

Il faut donc que in extenso:

$$M(x, y) = \frac{\partial z}{\partial x} \text{ et } N(x, y) = \frac{\partial z}{\partial y} \quad (10.57)$$

ou encore, en effectuant une seconde dérivation, que:

$$\frac{\partial}{\partial y} \left(\frac{\partial z}{\partial x} \right) = \frac{\partial^2 z}{\partial y \partial x} = \frac{\partial}{\partial y} (M(x, y)) = \frac{\partial M}{\partial y} \text{ et } \frac{\partial}{\partial x} \left(\frac{\partial z}{\partial y} \right) = \frac{\partial^2 z}{\partial x \partial y} = \frac{\partial}{\partial x} (N(x, y)) = \frac{\partial N}{\partial x} \quad (10.58)$$

pour que la forme différentielle, soit une différentielle totale exacte.

Avant de continuer, nous avons besoin d'un résultat donné par le "**théorème de Schwarz**" (mais qui a été démontré à la fin du 17ème siècle par un des frères Bernoulli) qui s'énonce de la manière suivante:

Soit une fonction f , si:

$$f, \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial^2 f}{\partial x \partial y} \text{ et } \frac{\partial^2 f}{\partial y \partial x} \quad (10.59)$$

sont continues alors nous avons (il faut vraiment vérifier que ce soit le cas!) un résultat très important dans la pratique:

$$\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) = \frac{\partial^2 f}{\partial y \partial x}(x_0, y_0) \quad (10.60)$$

pour tout $(x_0, y_0) \in U$ où U est le domaine de définition où f est continue (et donc dérivable).

Démonstration:

Nous considérons l'expression:

$$H = f(x_0 + k, y_0 + h) - f(x_0 + k, y_0) - f(x_0, y_0 + h) + f(x_0, y_0) \quad (10.61)$$

Posons:

$$g(x) := f(x, y_0 + h) - f(x, y_0) \text{ et } w(y) := f(x_0 + k, y) - f(x_0, y) \quad (10.62)$$

Nous avons alors:

$$H = g(x_0 + k) - g(x_0) = w(y_0 + h) - w(y_0) \quad (10.63)$$

Par le théorème des accroissements finis:

$$\begin{aligned} H &= g(x_0 + k) - g(x_0) = k \cdot g'(x_0 + sk) \\ H &= w(y_0 + h) - w(y_0) = h \cdot w'(y_0 + th) \end{aligned} \quad (10.64)$$

avec $s, t \in]0, 1[$. En reprenant les définitions de g et w nous obtenons:

$$\begin{aligned} H &= k \cdot g'(x_0 + sk) = k \cdot \left(\frac{\partial f}{\partial x}(x_0 + sk, y_0 + h) - \frac{\partial f}{\partial x}(x_0 + sk, y_0) \right) \\ H &= h \cdot w'(y_0 + th) = h \cdot \left(\frac{\partial f}{\partial y}(x_0 + k, y_0 + th) - \frac{\partial f}{\partial y}(x_0, y_0 + th) \right) \end{aligned} \quad (10.65)$$

en appliquant à nouveau le théorème des accroissements finis aux deux membres entre parenthèses nous trouvons:

$$\begin{aligned} H &= k \cdot h \cdot \frac{\partial^2 f}{\partial y \partial x}(x_0 + sk, y_0 + \tilde{s}h) \\ H &= h \cdot k \cdot \frac{\partial^2 f}{\partial x \partial y}(x_0 + \tilde{t}k, y_0 + th) \end{aligned} \quad (10.66)$$

avec $\tilde{s}, \tilde{t} \in]0, 1[$. Pour finir:

$$\frac{\partial^2 f}{\partial y \partial x}(x_0 + sk, y_0 + \tilde{s}h) = \frac{\partial^2 f}{\partial x \partial y}(x_0 + \tilde{t}k, y_0 + th) \quad (10.67)$$

et par continuité lorsque $k, h \rightarrow 0$, nous avons:

$$\frac{\partial^2 f}{\partial y \partial x}(x_0, y_0) = \frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) \quad (10.68)$$

Plus simplement écrit:

$$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial^2 f}{\partial x \partial y} \quad (10.69)$$

Donc si f s'exprime sous forme différentielle totale exacte alors les différentielles croisées sont égales (la réciproque n'est pas forcément vraie).

□C.Q.F.D.

Par récurrence sur le nombre de variables nous pouvons démontrer le cas général (c'est long mais c'est possible, nous le ferons si besoin il y a...).

Donc finalement pour en revenir à notre problème initial, nous avons donc:

$$\frac{\partial^2 z}{\partial x \partial y} = \frac{\partial^2 z}{\partial y \partial x} \quad (10.70)$$

Ce qui nous donne finalement la "**condition de Schwarz**":

$$\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x} \quad (10.71)$$

C'est donc la condition que doit satisfaire une forme différentielle pour être une différentielle totale exacte et la condition qu'elle ne doit pas satisfaire pour être une différentielle totale inexacte!!! C'est une propriété très important pour l'étude de la Thermodynamique!

Afin de ne pas confondre les deux types de différentielles, nous utilisons le symbole δ pour représenter une différentielle totale inexacte:

$$\delta z = Mdx + Ndy \quad (10.72)$$

et d pour une différentielle totale exacte:

$$dz = Mdx + Ndy \quad (10.73)$$

La distinction est extrêmement importante car seules les différentielles totales exactes qui satisfont donc:

$$\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x} \quad (10.74)$$

ont une intégrale qui ne dépend que des bornes d'intégration (puisque toutes les variables changent en même temps). Dès lors les différentielles totales inexactes dépendent des bornes d'intégration, ce qui signifie que:

$$\int_{z_1}^{z_2} \delta z \neq z_2 - z_1 \quad (10.75)$$

et donc:

$$\oint \delta z \neq 0 \quad (10.76)$$

Alors que (voir plus loin la partie traitant des intégrales curvilignes):

$$\int_{z_1}^{z_2} dz = z_2 - z_1 = \Delta z \quad (10.77)$$

soit (voir la démonstration détaillée plus loin lorsque nous traiterons des intégrales curvilignes):

$$\oint dz = 0 \quad (10.78)$$

Autrement dit, la variation d'une fonction dont la différentielle est totale exacte, ne dépend pas du chemin suivi, mais uniquement des états initiaux et finaux car elle s'exprime comme le gradient d'une fonction (voir la démonstration par l'exemple dans le chapitre d'Électrostatique quand nous vérifierons que la différence de potentiel est indépendante du chemin). Nous appelons une telle fonction qui satisfait à une différentielle totale exacte en physique, une "**fonction d'état**" et en mathématique une "**fonction holomorphe**" (voir le chapitre d'Analyse Complexe pour plus de détails), c'est-à-dire une fonction dont la valeur ne dépend que de l'état présent et futur, et non de son histoire.

Cette distinction est très importante et particulièrement en thermodynamique où il convient de déterminer si une quantité physique est une différentielle totale exacte (une "fonction d'état" donc) ou non afin de savoir comment évoluent les systèmes.

Exemple:

Un exemple important de forme différentielle en thermodynamique, est le travail élémentaire δW d'une force exercée sur un corps en mouvement dans le plan Oxy, nous avons:

$$\delta W = \vec{F} \circ d\vec{r} = F_x dx + F_y dy \quad (10.79)$$

F_x et F_y ne dérivent pas nécessairement d'un même potentiel $U(x, y)$ tel que:

$$\frac{\partial U}{\partial x} = \frac{\partial U}{\partial y} \quad (10.80)$$

δW est donc une différentielle totale inexacte!

1.2. DÉRIVÉES USUELLES

Nous allons démontrer ici les dérivées les plus fréquentes (une petite trentaine) que nous puissions rencontrer en physique théorique et mathématique ainsi que certaines de leurs propriétés (en fait, nous allons toutes les appliquer dans les sections relatives à la Mécanique, l'Ingénierie, l'Atomistique, les Mathématiques Sociales, etc.). La liste est pour l'instant non exhaustive mais les démonstrations étant généralisées, elles peuvent s'appliquer à un grand nombre d'autres cas similaires (que nous appliquerons/rencontrerons tout au long de ce site).

1. Dérivée de x^* :

Partons d'abord d'un cas particulier, la dérivée de x^3 :

Soit donc a un réel quelconque fixé, alors:

$$\begin{aligned}\lim_{h \rightarrow 0} \frac{(a+h)^3 - a^3}{h} &= \lim_{h \rightarrow 0} \frac{a^3 + 3a^2h + 3ah^2 + h^3 - a^3}{h} = \lim_{h \rightarrow 0} \frac{3a^2h + 3ah^2 + h^3}{h} \\ &= \lim_{h \rightarrow 0} 3a^2 + 3ah + h^2 = 3a^2\end{aligned}\quad (10.81)$$

Le nombre dérivé en a de la fonction cube est donc $f'(a) = 3a^2$.

Nous pouvons généraliser ce résultat pour tout entier naturel positif ou négatif n et nous allons voir que la fonction f définie sur $\mathbb{R}$ par $f(x) = x^n$ est dérivable et que sa dérivée f' est définie par $f'(x) = nx^{n-1}$.

$$\begin{aligned}\lim_{h \rightarrow 0} \frac{(a+h)^n - a^n}{h} &= \lim_{h \rightarrow 0} \frac{\sum_{k=0}^n C_n^k a^k h^{n-k} - a^n}{h} = \lim_{h \rightarrow 0} \frac{\sum_{k=0}^{n-1} C_n^k a^k h^{n-k} + C_n^n a^n - a^n}{h} \\ &= \lim_{h \rightarrow 0} \frac{\sum_{k=0}^{n-1} C_n^k a^k h^{n-k} + a^n - a^n}{h} = \lim_{h \rightarrow 0} \frac{\sum_{k=0}^{n-1} C_n^k a^k h^{n-k}}{h} = \lim_{h \rightarrow 0} \sum_{k=0}^{n-1} C_n^k a^k h^{n-k-1} \\ &= \lim_{h \rightarrow 0} \left(\sum_{k=0}^{n-2} C_n^k a^k h^{n-k-1} + C_n^{n-1} a^{n-1} \right) = C_n^{n-1} a^{n-1} = \frac{n!}{(n-1)!} a^{n-1} = na^{n-1}\end{aligned}\quad (10.82)$$

Ainsi, nous avons (quelques exemples peuvent être utiles pour comprendre la portée de ce résultat):

$$\begin{aligned}f(x) = x^2 &\Rightarrow f'(x) = 2x^1 \\ f(x) = x^{-1} &\Rightarrow f'(x) = -x^{-2} \\ f(x) = \sqrt{x} &\Rightarrow f'(x) = \frac{1}{2} x^{-1/2}\end{aligned}\quad (10.83)$$

Nous voyons donc qu'en ayant déterminé la dérivée d'une fonction de la forme x^n , nous avons également déterminé la dérivée de toute fonction qui est mise sous cette forme tel que:

$$f(x) = \frac{1}{x^n} = x^{-n} \text{ et } f(x) = \sqrt[n]{x} = x^{1/n} \quad (10.84)$$

Cependant, les fonctions:

$$f(x) = \frac{1}{x^n}, f(x) = \sqrt[n]{x} \quad (10.85)$$

ne sont pas dérivables en $x = 0$ puisque la fonction n'y est plus définie (division par zéro). De plus, en ce qui concerne la fonction comportant la racine (puissance non entière), la dérivée n'est pas définie dans $\mathbb{R}_+^*$.

Cependant, le résultat précédent donne un résultat intéressant pour les fonctions constantes telle que:

$$f(x) = c^x = e^{x \ln c} = c^{x \ln c} \quad (10.86)$$

il n'est alors pas difficile de déterminer la dérivée qui vaut simplement:

$$f'(x) = c^x 0x^{0-1} = 0 \quad (10.87)$$

Donc la dérivée de toute fonction constante est nulle (il est important de se souvenir de ce résultat quand nous étudierons les propriétés des intégrales) !!!

2. Dérivée de la fonction $f(x)=\cos(x)$:

Soit donc a un réel quelconque fixé, alors (attention! il est utile de connaître les relations trigonométriques remarquables que nous démontrons dans le chapitre de trigonométrie dans la section de géométrie):

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{\cos(a+h) - \cos(a)}{h} &= \lim_{h \rightarrow 0} \frac{-2 \sin\left(\frac{a+h+a}{2}\right) \sin\left(\frac{a+h-a}{2}\right)}{h} \\ &= \lim_{h \rightarrow 0} \frac{-2 \sin\left(\frac{2a+h}{2}\right) \sin\left(\frac{h}{2}\right)}{h} = \lim_{h \rightarrow 0} \left(-2 \sin\left(\frac{2a+h}{2}\right) \right) \cdot \frac{1}{2} \lim_{h \rightarrow 0} \left(\frac{\sin\left(\frac{h}{2}\right)}{\frac{h}{2}} \right) = -2 \sin(a) \frac{1}{2} \quad (10.88) \\ &= -\sin(a) \end{aligned}$$

Puisque:

$$\lim_{x \rightarrow 0} \frac{\sin(x)}{x} = 1 \quad (10.89)$$

Effectivement, rappelons que la fonction $\sin(x)$ est assimilable (visuellement et mathématiquement) à une droite de fonction $f(x) = x$ au voisinage de $x \approx 0$.

Donc pour résumer:

$$f(x) = \cos(x) \Rightarrow f'(x) = -\sin(x) \quad (10.90)$$

3. Dérivée de la fonction $f(x)=\sin(x)$:

Soit donc a un réel quelconque fixé, alors (attention! il est utile de connaître les relations trigonométriques remarquables que nous démontrons dans le chapitre de Trigonométrie dans la section de géométrie):

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{\sin(a+h) - \sin(a)}{h} &= \lim_{h \rightarrow 0} \frac{2 \cos\left(\frac{a+h+a}{2}\right) \sin\left(\frac{a+h-a}{2}\right)}{h} \\ &= \lim_{h \rightarrow 0} \frac{2 \cos\left(\frac{2a+h}{2}\right) \sin\left(\frac{h}{2}\right)}{h} = \lim_{h \rightarrow 0} \left(2 \cos\left(\frac{2a+h}{2}\right) \right) \cdot \frac{1}{2} \lim_{h \rightarrow 0} \left(\frac{\sin\left(\frac{h}{2}\right)}{\frac{h}{2}} \right) = 2 \cos(a) \frac{1}{2} \quad (10.91) \\ &= \cos(a) \end{aligned}$$

Donc pour résumer:

$$f(x) = \sin(x) \Rightarrow f'(x) = \cos(x) \quad (10.92)$$

4. Dérivée de la fonction $\log_a(x)$:

La dérivée de la fonction $\log_a(x)$ est égale à $x^{-1}\log_a(e)$, c'est-à-dire si:

$$y = f(x) = \log_a(x) \quad (10.93)$$

alors:

$$y' = f'(x) = \frac{1}{x} \log_a(e) \quad (10.94)$$

Démonstration:

Si Δy est l'accroissement de la fonction $y = \log_a(x)$ pour un accroissement correspondant Δx de la variable x , alors:

$$\begin{aligned} y + \Delta y &= \log_a(x + \Delta x) \\ \Delta y &= \log_a(x + \Delta x) - \log_a(x) = \log_a \frac{x + \Delta x}{x} = \log_a \left(1 + \frac{\Delta x}{x} \right) \end{aligned} \quad (10.95)$$

et nous pouvons écrire:

$$\frac{\Delta y}{\Delta x} = \frac{1}{\Delta x} \log_a \left(1 + \frac{\Delta x}{x} \right) \quad (10.96)$$

Multiplions et divisons par x l'expression figurant dans le membre droit de la dernière égalité:

$$\frac{\Delta y}{\Delta x} = \frac{1}{x} \frac{x}{\Delta x} \log_a \left(1 + \frac{\Delta x}{x} \right) = \frac{1}{x} \log_a \left(1 + \frac{\Delta x}{x} \right)^{\frac{x}{\Delta x}} \quad (10.97)$$

Désignons la quantité $\frac{\Delta x}{x}$ par α . Il est évident que $\alpha \rightarrow 0$ quand Δx tend vers zéro pour un x donné. Par conséquent:

$$\frac{\Delta y}{\Delta x} = \frac{1}{x} \log_a (1 + \alpha)^{\frac{1}{\alpha}} \quad (10.98)$$

Or, nous retrouvons ici une autre provenance historique de la constante d'Euler (cf. chapitre d'Analyse Fonctionnelle) où:

$$\lim_{\alpha \rightarrow 0} (1 + \alpha)^{\frac{1}{\alpha}} = e \quad (10.99)$$

Ainsi:

$$y' = \lim_{\Delta x \rightarrow 0} \frac{\Delta y}{\Delta x} = \lim_{\alpha \rightarrow 0} \frac{1}{x} \log_a (1 + \alpha)^{\frac{1}{\alpha}} = \frac{1}{x} \log_a(e) \quad (10.100)$$

□C.Q.F.D.

Un cas particulier important est le cas où $a=e$. Nous avons alors:

$$y = \ln(x) \Rightarrow y' = \frac{1}{x} \quad (10.101)$$

5. Dérivée d'une somme de fonctions:

Soient u et v deux fonctions. La fonction somme $s = u + v$ est dérivable sur tout intervalle où u et v sont dérivables, sa dérivée est la fonction s' somme des fonctions dérivées u' et v' de u et v .

Ce résultat se généralise pour une somme d'un nombre quelconque fixé de fonctions.

Démonstration:

Soient a un réel fixé et u et v deux fonctions définies et dérivables en a :

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{(u+v)(a+h) - (u+v)(a)}{h} &= \lim_{h \rightarrow 0} \frac{u(a+h) + v(a+h) - u(a) - v(a)}{h} \\ &= \lim_{h \rightarrow 0} \frac{u(a+h) - u(a) + v(a+h) - v(a)}{h} = \lim_{h \rightarrow 0} \frac{u(a+h) - u(a)}{h} + \lim_{h \rightarrow 0} \frac{v(a+h) - v(a)}{h} \quad (10.102) \\ &= u'(a) + v'(a) \end{aligned}$$

Donc la dérivée d'une somme est la somme des dérivées.

□C.Q.F.D.

6. Dérivée d'un produit de fonctions:

Soient u et v deux fonctions. La fonction produit $p = u \cdot v$ est dérivable sur tout intervalle où u et v sont dérivables, sa dérivée première est la fonction p' telle que:

$$p' = u'v + uv' \quad (10.103)$$

Démonstration:

Soient a un réel fixé et u et v deux fonctions définies et dérivables en a :

$$\lim_{h \rightarrow 0} \frac{(uv)(a+h) - (uv)(a)}{h} = \lim_{h \rightarrow 0} \frac{u(a+h)v(a+h) - u(a)v(a)}{h} \quad (10.104)$$

Nous rajoutons à cette dernière relation deux termes dont la somme est nulle tels que:

$$\begin{aligned}
\lim_{h \rightarrow 0} \frac{u(a+h)v(a+h) - u(a)v(a)}{h} &= \lim_{h \rightarrow 0} \frac{u(a+h)v(a+h) - u(a)v(a+h) + u(a)v(a+h) - u(a)v(a)}{h} \\
&= \lim_{h \rightarrow 0} \frac{u(a+h)v(a+h) - u(a)v(a+h)}{h} + \lim_{h \rightarrow 0} \frac{u(a)v(a+h) - u(a)v(a)}{h} \\
&= \lim_{h \rightarrow 0} \left(\frac{u(a+h) - u(a)}{h} v(a+h) \right) + \lim_{h \rightarrow 0} \left(\frac{v(a+h) - v(a)}{h} u(a) \right) \\
&= \lim_{h \rightarrow 0} \left(\frac{u(a+h) - u(a)}{h} \right) \lim_{h \rightarrow 0} v(a+h) + \lim_{h \rightarrow 0} \left(\frac{v(a+h) - v(a)}{h} \right) \lim_{h \rightarrow 0} u(a) \\
&= u'(a)v(a) + u(a)v'(a)
\end{aligned}$$

(10.105)

□C.Q.F.D.

Mais il existe une formulation plus générale que la dérivée première d'un produit:

Considérons pour cela toujours nos deux fonctions u et v , n fois dérivables sur un intervalle I . Alors le produit uv est n fois dérivable sur I et:

$$(uv)^{(n)} = \sum_{k=0}^n C_k^n u^{(k)} v^{(n-k)} \quad (10.106)$$

et ceci constitue la "**formule de Leibniz**" que nous avons utilisée dans le chapitre de Calcul Algébrique pour l'étude des polynômes de Legendre (qui nous sont eux-mêmes indispensables pour l'étude de la chimie quantique).

La démonstration de cette expression est très proche de celle faite pour le binôme de Newton (cf. chapitre de Calcul Algébrique).

Démonstration:

Soit:

$$(uv)^{(0)} = uv \quad (10.107)$$

D'autre part:

$$(uv)^{(0)} = \sum_{k=0}^0 C_k^0 u^{(k)} v^{(0-k)} = u^{(0)} v^{(0)} = uv \quad (10.108)$$

La relation est ainsi bien initialisée.

La démonstration se fait par récurrence. Ainsi, le but est de montrer que pour $\forall n \geq 0 \in \mathbb{N}$ que si:

$$(uv)^{(n)} = \sum_{k=0}^n C_k^n u^{(k)} v^{(n-k)} \quad (10.109)$$

alors:

$$(uv)^{(n+1)} = \sum_{k=0}^{n+1} C_k^{n+1} u^{(k)} v^{(n+1-k)} \quad (10.110)$$

Nous avons donc:

$$\begin{aligned} (uv)^{(n+1)} &= \left((uv)^{(n)} \right)' = \sum_{k=0}^n C_k^n \left(u^{(k)} v^{(n-k)} \right)' = \sum_{k=0}^n C_k^n \left(\left(u^{(k)} \right)' v^{(n-k)} + u^{(k)} \left(v^{(n-k)} \right)' \right) \\ &= \sum_{k=0}^n C_k^n \left(u^{(k+1)} v^{(n-k)} + u^{(k)} v^{(n+1-k)} \right) = \sum_{k=0}^n C_k^n u^{(k+1)} v^{(n-k)} + \sum_{k=0}^n C_k^n u^{(k)} v^{(n+1-k)} \end{aligned} \quad (10.111)$$

Nous allons procéder à un changement de variable dans la première somme pour ne plus avoir le terme en $k+1$. Nous posons pour cela $j = k+1$:

$$\sum_{k=0}^n C_k^n u^{(k+1)} v^{(n-k)} = \sum_{j=1}^{n+1} C_{j-1}^n u^{(j)} v^{(n-(j-1))} \quad (10.112)$$

Si nous revenons à la lettre k , nous avons donc:

$$\sum_{k=0}^n C_k^n u^{(k+1)} v^{(n-k)} = \sum_{k=1}^{n+1} C_{k-1}^n u^{(k)} v^{(n+1-k)} \quad (10.113)$$

Nous avons donc:

$$(uv)^{(n+1)} = \sum_{k=1}^{n+1} C_{k-1}^n u^{(k)} v^{(n+1-k)} + \sum_{k=0}^n C_k^n u^{(k)} v^{(n+1-k)} \quad (10.114)$$

Nous voulons réunir les deux sommes. Pour cela, nous écartons les termes en trop dans chacune d'elles:

$$(uv)^{(n+1)} = C_{n+1-1}^n u^{(n+1)} v^{(n+1-(n+1))} + \sum_{k=1}^n C_{k-1}^n u^{(k)} v^{(n+1-k)} + \sum_{k=1}^n C_k^n u^{(k)} v^{(n+1-k)} + C_0^n u^{(0)} v^{(n+1-0)} \quad (10.115)$$

Ce qui donne donc:

$$\begin{aligned} (uv)^{(n+1)} &= u^{(n+1)} + \sum_{k=1}^n C_{k-1}^n u^{(k)} v^{(n+1-k)} + \sum_{k=1}^n C_k^n u^{(k)} v^{(n+1-k)} + v^{(n+1)} \\ &= u^{(n+1)} + \sum_{k=1}^n \left(C_{k-1}^n + C_k^n \right) u^{(k)} v^{(n+1-k)} + v^{(n+1)} \end{aligned} \quad (10.116)$$

D'après la formule de Pascal (cf. chapitre de Probabilités), nous avons:

$$C_{k-1}^n + C_k^n = C_k^{n+1} \quad (10.117)$$

Donc:

$$(uv)^{(n+1)} = u^{(n+1)} + \sum_{k=1}^n C_k^{n+1} u^{(k)} v^{(n+1-k)} + v^{(n+1)} \quad (10.118)$$

Or:

$$u^{(n+1)} = C_{n+1}^{n+1} u^{n+1} v^{(n+1-(n+1))} \quad v^{(n+1)} = C_0^{n+1} u^{(0)} v^{(n+1-0)} \quad (10.119)$$

Donc:

$$(uv)^{(n+1)} = \sum_{k=0}^{n+1} C_k^{n+1} u^{(k)} v^{(n+1-k)} \quad (10.120)$$

□C.Q.F.D.

7. Dérivée d'une fonction composée:

Soit la fonction composée $f = g \circ u$ de deux fonctions g et u dérivables, la première en $u(x)$, la seconde en x , la fonction dérivée f' est définie par $f' = (g' \circ u)u'$, c'est-à-dire:

$$f'(x) = g'(u(x)) \cdot u'(x) \quad (10.121)$$

Démonstration:

Soient a un réel fixé et u une fonction définie et dérivable en a et g une fonction définie et dérivable en $u(a)$:

$$\lim_{h \rightarrow 0} \frac{(g \circ u)(a+h) - (g \circ u)(a)}{h} = \lim_{h \rightarrow 0} \frac{g(u(a+h)) - g(u(a))}{h} \quad (10.122)$$

posons $k = u(a+h) - u(a)$, nous avons alors:

$$\lim_{h \rightarrow 0} k = 0 \quad (10.123)$$

continuons notre développement précédent:

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{g(u(a+h)) - g(u(a))}{h} &= \lim_{h \rightarrow 0} \frac{g(u(a) + k) - g(u(a))}{h} = \lim_{h \rightarrow 0} \frac{g(u(a) + k) - g(u(a))}{k} \frac{k}{h} \\ &= \lim_{h \rightarrow 0} \frac{g(u(a) + k) - g(u(a))}{k} \lim_{h \rightarrow 0} \frac{k}{h} = \lim_{h \rightarrow 0} \frac{g(u(a) + k) - g(u(a))}{k} \lim_{h \rightarrow 0} \frac{u(a+h) - u(a)}{h} \\ &= g'(u(a)) \cdot u'(a) \end{aligned} \quad (10.124)$$

□C.Q.F.D.

Donc la dérivée d'une fonction composée est donnée par la dérivée de la fonction multipliée par la "**dérivée intérieure**". Par ailleurs, ce type de dérivation est très important car souvent utilisé en physique sous la dénomination de "**dérivation en chaîne**".

Voyons de quoi il s'agit. La dernière relation obtenue peut être écrite sous une autre forme si nous posons $y = f(g(x))$ et $u = g(x)$:

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx} \quad (10.125)$$

Ce qui peut s'étendre à des cas plus compliqués par exemple si $z = f(x, y)$, $x = g(t)$, $y = h(t)$ alors:

$$\frac{\partial z}{\partial t} = \frac{\partial z}{\partial x} \frac{\partial x}{\partial t} + \frac{\partial z}{\partial y} \frac{\partial y}{\partial t} \quad (10.126)$$

8. Dérivée d'une fonction réciproque:

Si la fonction f est continue, strictement monotone sur un intervalle I , dérivable sur I , alors la fonction réciproque f^{-1} est dérivable sur l'intervalle $f(I)$ et admet pour fonction dérivée:

$$\forall y \in f(I) \quad (f^{-1})'(y) = \frac{1}{f'(f^{-1}(y))} \quad (10.127)$$

En effet, nous pouvons écrire:

$$x \xrightarrow{f} y = y = f(x) \xrightarrow{f^{-1}} x = f^{-1}(y) \quad (10.128)$$

C'est-à-dire (application identité):

$$x = f^{-1}(y) = f^{-1}(f(x)) = f^{-1} \circ f(x) \quad (10.129)$$

Par application de la dérivation des fonctions composées:

$$f^{-1}(f(x))' = f^{-1}(y)' \cdot f'(x) = (x)' = 1 \quad (10.130)$$

d'où:

$$f^{-1}(y)' = \frac{1}{f'(x)} = \frac{1}{f'(f^{-1}(y))} \quad (10.131)$$

Pour une variable x , nous poserons pour la dérivée de la fonction réciproque:

$$f^{-1}(x)' = \frac{1}{f'(y)} \quad (10.132)$$

10. Dérivée de la fonction arccos(x):

En utilisant le résultat précédent de la fonction réciproque, nous pouvons calculer la dérivée de la fonction arccos(x):

$$f^{-1}(x)' = (\arccos(x))' = \frac{1}{f'(y)} = \frac{1}{-\sin(y)} = \frac{-1}{\sin(y)} = \frac{-1}{\sqrt{1-\cos^2(y)}} = \frac{-1}{\sqrt{1-x^2}} \quad (10.133)$$

11. Dérivée de la fonction arcsin(x):

En utilisant le résultat précédent de la fonction réciproque, nous pouvons calculer la dérivée de la fonction $\arcsin(x)$:

$$f^{-1}(x)' = (\arcsin(x))' = \frac{1}{f'(y)} = \frac{1}{\cos(y)} = \frac{1}{\sqrt{1-\sin^2(y)}} = \frac{1}{\sqrt{1-x^2}} \quad (10.134)$$

12. Dérivée d'un quotient de deux fonctions:

La fonction $f = u/v$ est dérivable sur tout intervalle où les fonctions u et v sont dérivables et où la fonction v est non nulle et:

$$\left(\frac{u}{v}\right)' = \frac{u'v - uv'}{v^2} \quad (10.135)$$

Démonstration:

La fonction f peut être considérée comme le produit de deux fonctions: la fonction u et la fonction $1/v$. Une produit de deux fonctions est dérivable si chacune d'elle est dérivable, il faut donc que la fonction u soit dérivable et que la fonction $1/v$ soit également dérivable ce qui est le cas quand v est dérivable non nulle.

$$\left(\frac{u}{v}\right)' = \left(u \cdot \frac{1}{v}\right)' = u' \cdot \frac{1}{v} + u \cdot \left(\frac{1}{v}\right)' = \frac{u'}{v} + u \cdot \left(\frac{-v'}{v^2}\right) = \frac{u'v - uv'}{v^2} \quad (10.136)$$

□C.Q.F.D.

13. Dérivée de la fonction $\tan(x)$:

Par définition (cf. chapitre de Trigonométrie) nous avons:

$$f(x) = \tan(x) = \frac{\sin(x)}{\cos(x)} \quad (10.137)$$

et en appliquant donc la dérivée d'un quotient vue précédemment, nous avons:

$$f'(x) = \frac{\cos^2(x) + \sin^2(x)}{\cos^2(x)} = \frac{1}{\cos^2(x)} \quad (10.138)$$

ou encore:

$$f'(x) = \frac{\cos^2(x) + \sin^2(x)}{\cos^2(x)} = 1 + \frac{\sin^2(x)}{\cos^2(x)} = 1 + \tan^2(x) \quad (10.139)$$

14. Dérivée de la fonction $\cot(x)$:

Par définition (cf. chapitre de Trigonométrie), $\forall x \neq k \cdot \pi$ ($k \in \mathbb{Z}$):

$$f(x) = \cot(x) = \frac{1}{\tan(x)} \quad (10.140)$$

et donc (dérivée d'un quotient à nouveau):

$$f'(x) = -\frac{1 + \tan^2(x)}{\tan^2(x)} = -\frac{1}{\tan^2(x)} - 1 = -(\cot^2(x) + 1) \quad (10.141)$$

ou encore:

$$\cot'(x) = -\frac{1 + \tan^2(x)}{\tan^2(x)} = -\frac{1}{\tan^2(x)} - 1 = -\frac{\cos^2(x)}{\sin^2(x)} - 1 = -\frac{\cos^2(x) + \sin^2(x)}{\sin^2(x)} = -\frac{1}{\sin^2(x)} \quad (10.142)$$

15. Dérivée de la fonction arctan(x):

Nous utilisons les propriétés dérivées des fonctions réciproques:

$$\forall x \in \mathbb{R} \quad f'(x) = \frac{1}{1 + \tan^2(\arctan(x))} = \frac{1}{1 + x^2} \quad (10.143)$$

16. Dérivée de la fonction arccot(x):

Selon la même méthode que précédemment:

$$\forall x \in \mathbb{R} \quad f'(x) = -\frac{1}{1 + \cot^2(\arccot(x))} = -\frac{1}{1 + x^2} \quad (10.144)$$

17. Dérivée de la fonction e^x :

Nous verrons lors de notre étude des méthodes numériques (cf. chapitre de Méthodes Numériques) que le "**nombre d'Euler**" peut être calculé selon la série:

$$e^x = \sum_{k=0}^{+\infty} \frac{x^k}{k!} \quad (10.145)$$

qui converge sur $\mathbb{R}$. En dérivant terme à terme cette série qui converge, il vient:

$$\forall x \in \mathbb{R} \quad f'(x) = \sum_{k=0}^{+\infty} k \cdot \frac{x^{k-1}}{k!} = \sum_{k=1}^{+\infty} \frac{x^{k-1}}{(k-1)!} = e^x \quad (10.146)$$

Ainsi l'exponentielle est sa propre dérivée. Ainsi, nous pouvons nous permettre d'étudier les dérivées de quelques fonctions trigonométriques hyperboliques (cf. chapitre de Trigonométrie).

18. Dérivée de la fonction sinh(x):

Rappel:

$$\sinh(x) = \frac{1}{2}(e^x - e^{-x}) \quad (10.147)$$

Donc trivialement:

$$\sinh'(x) = \frac{1}{2}(e^x + e^{-x}) = \cosh(x) \quad (10.148)$$

19. Dérivée de la fonction $\cosh(x)$:

Rappel:

$$\cosh(x) = \frac{1}{2}(e^x + e^{-x}) \quad (10.149)$$

Donc trivialement:

$$\cosh'(x) = \frac{1}{2}(e^x - e^{-x}) = \sinh(x) \quad (10.150)$$

20. Dérivée de la fonction $\tanh(x)$:

Puisque par définition:

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} \quad (10.151)$$

Donc en appliquant la dérivée d'un quotient nous obtenons:

$$\tanh'(x) = \frac{\cosh^2(x) - \sinh^2(x)}{\cosh^2(x)} = \frac{1}{\cosh^2(x)} \quad (10.152)$$

Ou encore:

$$\tanh'(x) = \frac{\cosh^2(x) - \sinh^2(x)}{\cosh^2(x)} = 1 - \frac{\sinh^2(x)}{\cosh^2(x)} = 1 - \tanh^2(x) \quad (10.153)$$

21. Dérivée de la fonction $\coth(x)$:

Rappel:

$$\forall x \neq 0, \coth(x) = \frac{1}{\tanh(x)} \quad (10.154)$$

et donc:

$$\coth'(x) = -\frac{1 - \tanh^2(x)}{\tanh^2(x)} = 1 - \frac{1}{\tanh^2(x)} = 1 - \coth^2(x) \quad (10.155)$$

22. Dérivée de la fonction $\operatorname{arcsinh}(x)$:

Nous appliquons les propriétés des dérivées des fonctions réciproques:

$$\operatorname{arcsinh}'(x) = \frac{1}{\cosh(\operatorname{arcsinh}(x))} \quad (10.156)$$

Or (voir à nouveau le chapitre de Trigonométrie):

$$1 = \cosh^2(\operatorname{arcsinh}(x)) - \sinh^2(\operatorname{arcsinh}(x)) = \cosh^2(\operatorname{arcsinh}(x)) - x^2 \quad (10.157)$$

et donc:

$$1 + x^2 = \cosh^2(\operatorname{arcsinh}(x)) \quad (10.158)$$

Etant donné que $\cosh$ ne prend que des valeurs positives, nous avons:

$$\cosh(\operatorname{arcsinh}(x)) = \sqrt{1 + x^2} \quad (10.159)$$

Donc finalement:

$$\operatorname{arcsinh}'(x) = \frac{1}{\sqrt{1 + x^2}} \quad (10.160)$$

23. Dérivée de la fonction $\operatorname{arccosh}(x)$:

Nous appliquons les propriétés des dérivées des fonctions réciproques:

$$\forall x > 1 \quad \operatorname{arccosh}'(x) = \frac{1}{\sinh(\operatorname{arccosh}(x))} \quad (10.161)$$

Or selon la même méthode que précédemment:

$$1 = \cosh^2(\operatorname{arccosh}(x)) - \sinh^2(\operatorname{arccosh}(x)) = x^2 - \sinh^2(\operatorname{arccosh}(x)) \quad (10.162)$$

d'où:

$$-1 + x^2 = \sinh^2(\operatorname{arccosh}(x)) \quad (10.163)$$

Etant donné que $\operatorname{arccosh}$ ne prend que des valeurs positives nous avons alors:

$$\sinh(\operatorname{arccosh}(x)) = \sqrt{x^2 - 1} \quad (10.164)$$

Donc:

$$\forall x > 1 \quad \operatorname{arccosh}'(x) = \frac{1}{\sqrt{x^2 - 1}} \quad (10.165)$$

24. Dérivée de la fonction $\operatorname{arctanh}(x)$:

En appliquant les propriétés des dérivées des fonctions réciproques:

$$\forall x \in]-1, 1[\quad \operatorname{arctanh}'(x) = \frac{1}{1 - \tanh^2(\operatorname{arctanh}(x))} = \frac{1}{1 - x^2} \quad (10.166)$$

25. Dérivée de la fonction $\operatorname{arcoth}(x)$:

En appliquant les propriétés des dérivées des fonctions réciproques si $|x| > 1$:

$$\operatorname{arccoth}'(x) = \frac{1}{1 - \coth^2(\operatorname{arccoth}(x))} = \frac{1}{1 - x^2} \quad (10.167)$$

26. Dérivée de la fonction a^x :

Avec $a > 0$:

$$a^x = \exp(\ln(a^x)) = \exp(x \cdot \ln(a)) \quad (10.168)$$

Donc (dérivée d'une fonction composée):

$$f'(x) = \exp(x \cdot \ln(a)) \cdot \ln(a) = a^x \cdot \ln(a) \quad (10.169)$$

2. CALCUL INTÉGRAL

Nous allons aborder ici les principes élémentaires et de base du calcul intégral. La suite (avec plus de rigueur) viendra en fonction du temps qui est à la disposition des responsables du site.

2.1. INTÉGRALE DÉFINIE

L'idée première du concept d'intégral est de calculer l'aire algébrique (surface) entre une courbe et son support:

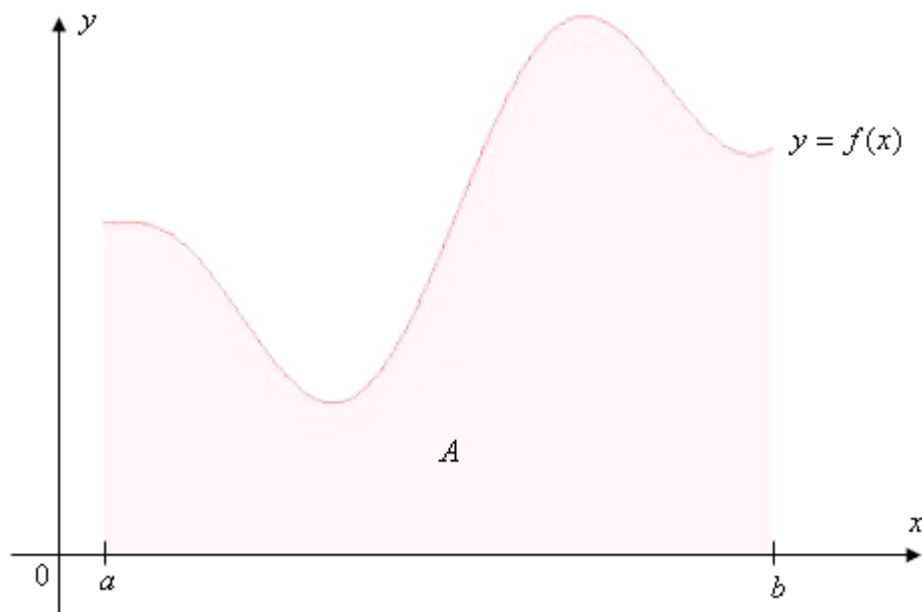


Figure: 10.6 - Aire (surface) à calculer sous une courbe continue bornée

Une valeur approchée de l'aire sous une courbe peut être obtenue par un découpage en n bandes rectangulaires verticales de même largeur. En particulier on peut réaliser un encadrement de cette aire à l'aide d'une somme majorante A_M et d'une somme minorante A_m pour un découpage donné.

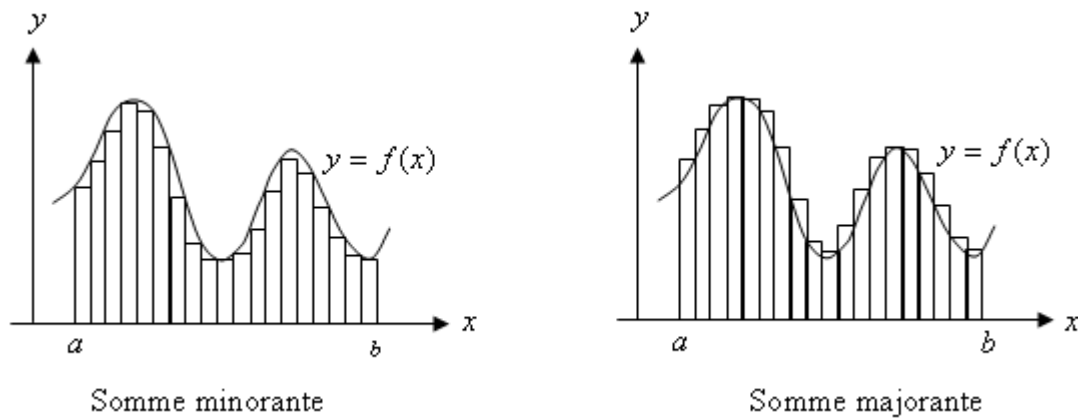


Figure: 10.7 - Représentation graphique des sommes majorante ou minorante

Supposons que le nombre n de bandes tende vers l'infini. Comme les bandes sont de même largeur, la largeur de chaque bande tend vers 0 (objectivement il n'est pas nécessaire que la largeur des sous-intervalles du découpage soit la même partout).

Si les sommes A_m et A_M ont toutes deux une limite lorsque, le nombre n de bandes, tend vers l'infini, alors l'aire A sous la courbe est comprise entre ces deux limites.

Nous avons:

$$\lim_{n \rightarrow \infty} A_m \leq A \leq \lim_{n \rightarrow \infty} A_M \quad (10.170)$$

Si ces deux limites sont égales, leur valeur est celle de l'aire sous la courbe.

D'où une première définition directe de l'intégrale définie ou dite "**intégrale de Riemann**":

Soit un intervalle $[a, b]$, divisé en n parties égales, soit f une fonction continue sur l'intervalle $[a, b]$, soit A_m , la somme algébrique minorante et soit A_M , la somme algébrique majorante. Nous appelons "**intégrale définie**" de f , depuis a jusqu'à b , notée:

$$\int_a^b f(x) dx \quad (10.171)$$

le nombre A tel que:

$$A = \lim_{n \rightarrow \infty} A_m = \lim_{n \rightarrow \infty} A_M \quad (10.172)$$

pourvu que cette limite existe. Si cette limite existe, alors nous disons que f est "**intégrable**" sur $[a, b]$ et l'intégrale définie existe. Le symbole:

$$\int_a^b \quad (10.173)$$

n'est que le symbole de la somme discrète mais appliquée aux cas d'éléments infiniment petits.

Les nombres a et b de l'intégrale sont appelés les "**bornes d'intégration**": a est la "**borne inférieure**", b est la "**borne supérieure**".

Intuitivement, il est évident que lorsque $b = a$, nous étendons la définition ainsi:

$$\int_a^a f(x) dx = 0 \quad (10.174)$$

Enfin, signalons qu'il est tout à fait possible que l'intégrale soit négative ou même complexe puisqu'il s'agit d'une surface algébrique!

Remarques:

R1. D'autres lettres que x peuvent être employées dans la notation de l'intégrale définie. Ainsi si f est intégrable sur $[a, b]$, alors $\int_a^b f(x) dx = \int_a^b f(t) dt = \int_a^b f(s) ds$ etc. C'est la raison pour laquelle la variable x de la définition est dite "**variable muette**".

R2. Comme nous le verrons plus loin, il est essentiel de ne pas confondre "intégrale définie" et "intégrale indéfinie". Ainsi, une intégrale indéfinie, notée $\int f(x) dx$ est une fonction, ou, plus précisément, une famille de fonctions appelées aussi "primitives de f " (voir plus bas) alors qu'une intégrale définie, notée $\int_a^b f(x) dx$ est une constante.

Voyons une deuxième approche de définition de l'intégrale, un peu plus rigoureuse que la précédente (suite à la demande de plusieurs lecteurs). Nous utiliserons par tradition cette fois-ci, le S de la surface au lieu du A de l'aire.

Soit f une fonction bornée sur $[a, b]$. Nous considérons une subdivision σ de son support $[a, b]$ que nous notons:

$$\sigma = \{x_0 = a < x_1 < x_2 \dots < x_{n-1} < x_n = b\} \quad (10.175)$$

où les intervalles ne sont pas obligatoirement de tailles équivalentes.

Nous pour $i = 1, 2, 3, \dots, n$:

$$m_i = \inf_{x \in [x_{i-1}, x_i]} f(x) \quad \text{et} \quad M_i = \sup_{x \in [x_{i-1}, x_i]} f(x) \quad (10.176)$$

Définitions:

D1. Nous appelons "**somme de Darboux inférieure**" associée à f et σ la surface:

$$s_{[a,b]}(f, \sigma) = \sum_{i=1}^n m_i (x_i - x_{i-1})$$

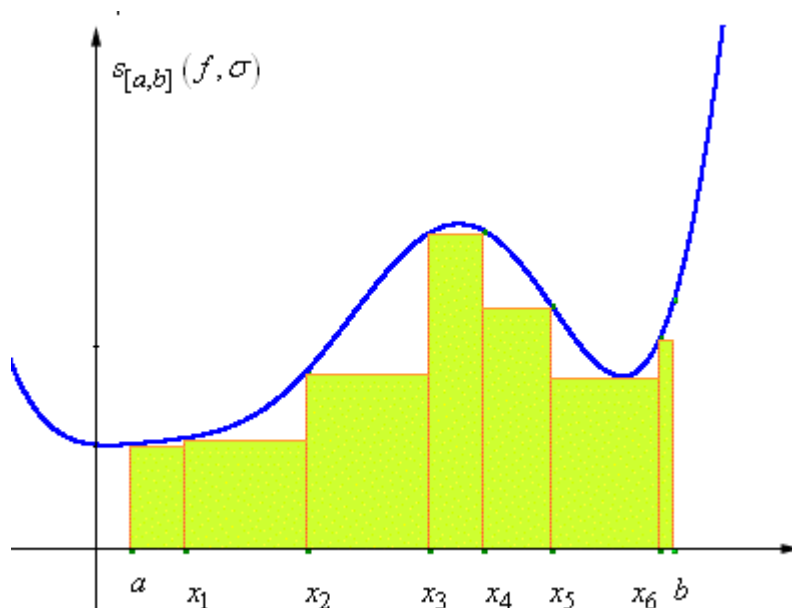


Figure: 10.8 - Principe du calcul de la somme de Darboux inférieure

D2. Nous appelons "**somme de Darboux supérieure**" associée à f et σ la surface:

$$S_{[a,b]}(f, \sigma) = \sum_{i=1}^n M_i (x_i - x_{i-1})$$

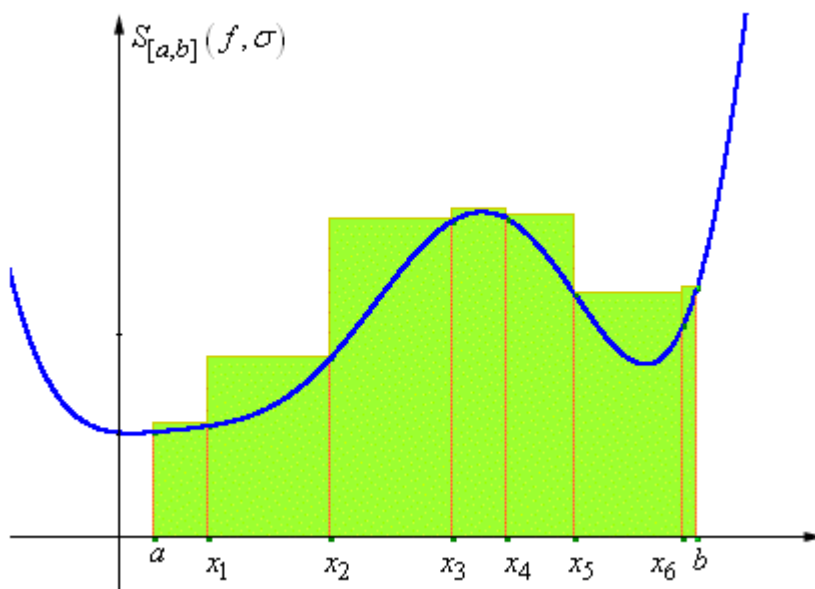


Figure: 10.9 - Principe du calcul de la somme de Darboux supérieure

Une fonction est dite "**Riemann-Intégrable sur $[a, b]$** " si et seulement si les deux surfaces susmentionnées coïncident lorsque les intervalles deviennent infiniment petit.

Le nombre correspondant à ces surfaces est alors appelé "l'intégrale de Riemann de f sur $[a, b]$ " est noté:

$$\int_a^b f(x) dx \quad (10.177)$$

L'ensemble des fonctions Riemann-intégrables sur $[a, b]$ est noté $\mathcal{R}_{a,b}$.

Les sommes de Darboux ne sont pas très utiles pour le calcul effectif d'une intégrale, par exemple à l'aide d'un ordinateur, car il est en général assez difficile de trouver les inf et sup sur les sous-intervalles. On considère plutôt:

$$\mathcal{S}_{[a,b]}(f) = \sum_{i=1}^n (x_i - x_{i-1}) f(x_{i-1}) \quad \text{ou} \quad \mathcal{S}_{[a,b]}(f) = \sum_{i=1}^n (x_i - x_{i-1}) f(x_i) \quad (10.178)$$

La somme de Riemann se définit à partir du fait que nous notons:

$$\Delta x_i = x_i - x_{i-1} \quad (10.179)$$

et que:

$$f(x_i) = f(\xi_i) \quad (10.180)$$

où:

$$\xi_i \in [x_{i-1}, x_i] \quad (10.181)$$

Dès lors:

$$S \equiv \mathcal{S}_{[a,b]}(f) = \mathcal{S}_{[a,b]}(f) = \sum_{i=1}^n f(\xi_i) \Delta x_i \quad (10.182)$$

Mais comme il faut bien choisir un ξ_i , souvent on prend soit celui à droite, soit celui à gauche, dès lors en prenant au hasard la "**méthode des rectangles à gauche**":

$$S \equiv \sum_{i=1}^n f(x_{i-1}) \Delta x_i \quad (10.183)$$

Ce qui nous donnerait pour l'exemple ci-dessous:

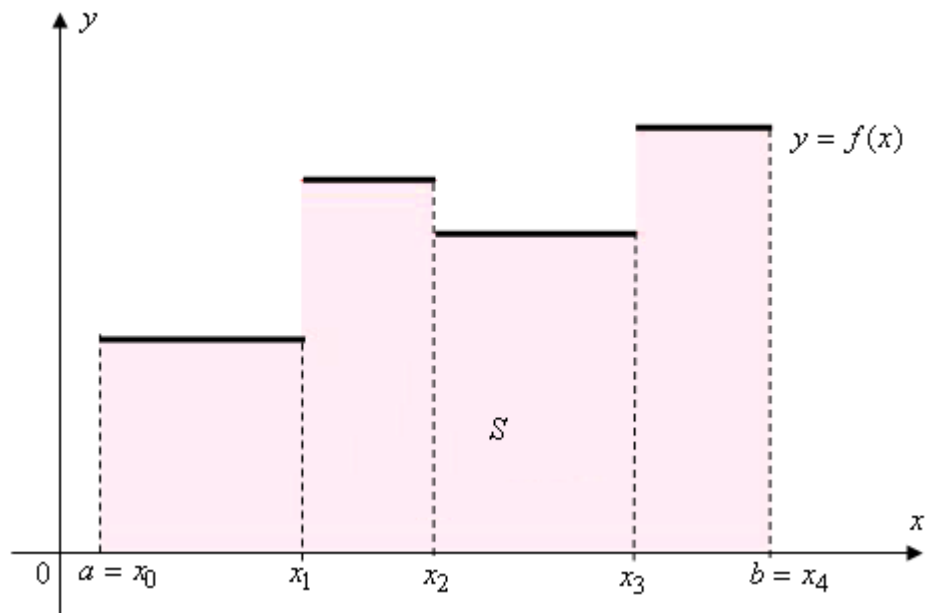
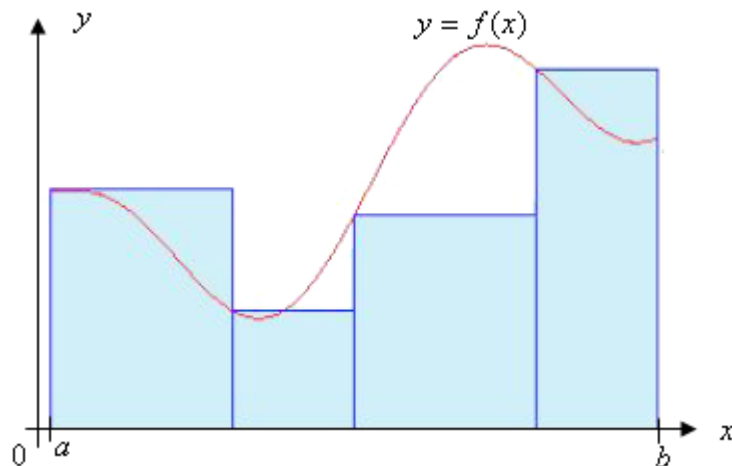


Figure: 10.10 - Principe du calcul de la méthodes des rectangles à gauche

Soit:

$$\begin{aligned}
 S &= \sum_{i=1}^4 f(x_{i-1}) \Delta x_i = f(x_0) \Delta x_1 + f(x_1) \Delta x_2 + f(x_2) \Delta x_3 + f(x_3) \Delta x_4 \\
 &= f(x_0)(x_1 - x_0) + f(x_1)(x_2 - x_1) + f(x_2)(x_3 - x_2) + f(x_3)(x_4 - x_3)
 \end{aligned}
 \tag{10.184}$$

Mais c'est facile pour une fonction en escalier... mais cela l'est moins pour une fonction continue pour laquelle nous n'obtiendrions qu'une valeur approchée de la surface réelle! L'idée est alors de prendre des intervalles de plus en plus petits:



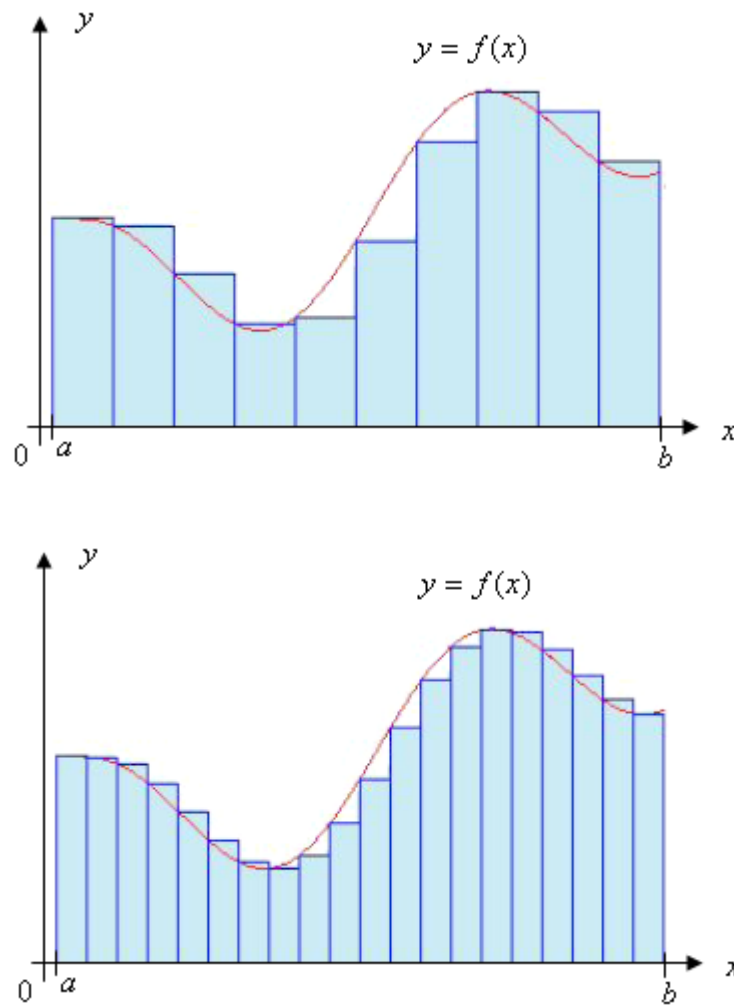


Figure: 10.11 - Principe du calcul de l'intégrale de Riemann avec la méthode des rectangles à gauche

Et dès lors, à la limite, nous obtenons la quantité voulue:

$$S = \lim_{\Delta x_i \rightarrow 0} \int_a^b f(x) dx \quad (10.185)$$

Le fait de chercher cette limite s'appelle "**calculer l'intégrale**", et plus spécifiquement de la méthode choisie: "**intégrale de Riemann**".

2.2. INTÉGRALE INDÉFINIE

Nous avons vu précédemment lors de notre étude des dérivées, le problème suivant: étant donnée une fonction $F(x)$, trouver sa dérivée, c'est-à-dire la fonction:

$$F'(x) = f(x) \quad (10.186)$$

Définition: Nous disons que la fonction $F(x)$ est une "**primitive**" ou "**intégrale indéfinie**" de la fonction $f(x)$ sur le segment $[a, b]$, si en tout point de ce segment nous avons l'égalité $F'(x) = f(x)$.

Une autre manière de voir le concept d'intégrale indéfinie est de passer par le "**théorème fondamental du calcul intégral (et différentiel)**" appelé aussi parfois "**théorème fondamental de l'analyse**" dont les 2 propriétés s'énoncent ainsi:

Soit f une fonction continue sur un intervalle fermé $[a, b]$.

P1. Si A est la fonction définie par $A(X) = \int_a^X f(t) dt$ pour tout X dans $[a, b]$, alors A est la primitive de f sur $[a, b]$ qui s'annule en a (ou en d'autres termes: $f(t)$ est la dérivée de A)

P2. Si F est une primitive de f sur $[a, b]$, alors $\int_a^b f(t) dt = F(b) - F(a)$.

Démontrons la première propriété du théorème fondamental:

Démonstration:

Soit la fonction:

$$A(X) = \int_a^X f(t) dt \quad (10.187)$$

Si f est positive et $h > 0$ (la démonstration dans le cas où $h < 0$ est proposée similaire) et comme $X > a$, nous pouvons nous représenter $A(X)$ comme l'aire sous la courbe de f depuis $t = a$ jusqu'à $t = X$.

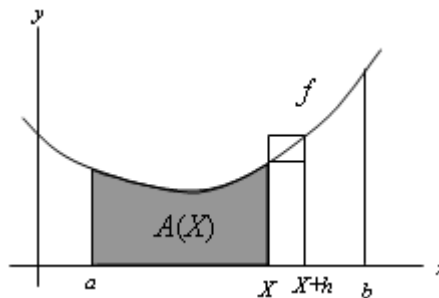


Figure: 10.12 - Représentation graphique de l'aire

Pour démontrer que A est une primitive de f , nous allons prouver que $A' = f$. Selon la définition de la dérivée:

$$A'(X) = \lim_{h \rightarrow 0} \frac{A(X+h) - A(X)}{h} \quad (10.188)$$

Etudions ce quotient: $A(X+h) - A(X)$ est représentée par l'aire de la bande de largeur h , prise en sandwich entre deux rectangles de largeur h .

Soit M le maximum de f sur l'intervalle $[X, X+h]$ et m le minimum de f sur ce même intervalle. Les aires respectives des deux rectangles sont Mh et mh .

Nous avons alors la double inégalité suivante:

$$mh \leq A(X+h) - A(X) \leq Mh \quad (10.189)$$

Comme h est positif, on peut diviser par h sans changer le sens des inégalités:

$$m \leq \frac{A(X+h) - A(X)}{h} \leq M \quad (10.190)$$

Lorsque $h \rightarrow 0^+$ et si f est une fonction continue, alors M et m ont pour limite $f(X)$, et le rapport:

$$\frac{A(X+h) - A(X)}{h} \quad (10.191)$$

qui est compris entre m et M , a bien pour limite $f(X)$.

Comme $A'(X) = f(X)$ pour tout X , ceci nous montre que la dérivée de la fonction aire est f . Ainsi A est une primitive de f . Comme $A(a) = 0$, A est bien la primitive de f qui s'annule en a .

□C.Q.F.D.

Avant de commencer la démonstration de la deuxième propriété du théorème fondamental, donnons et démontrons le théorème suivant qui va nous être indispensable: Si $F_1(x)$ et $F_2(x)$ sont deux primitives de la fonction $f(x)$ sur le segment $[a, b]$, leur différence est une constante (ce théorème est très important en physique pour ce qui est de l'étude de ce que nous appelons les "**conditions initiales**").

Démonstration:

Nous avons en vertu de la définition de la primitive:

$$\begin{aligned} F_1'(x) &= f(x) \\ F_2'(x) &= f(x) \end{aligned} \quad (10.192)$$

pour $\forall x \in [a, b]$.

Posons:

$$F_1(x) - F_2(x) = \varphi(x) \quad (10.193)$$

Nous pouvons écrire:

$$F_1'(x) - F_2'(x) = [F_1(x) - F_2(x)]' = \varphi'(x) = f(x) - f(x) = 0 \quad (10.194)$$

Il vient donc de ce que nous avons vu pendant notre étude des dérivées que $\varphi'(x) = 0$.

Nous avons alors:

$$F_1(x) - F_2(x) = c \quad (10.195)$$

□C.Q.F.D.

Il résulte de ce théorème que si nous connaissons une primitive quelconque $F(x)$ de la fonction $f(x)$, toute autre primitive de cette fonction sera de la forme:

$$F(x) + c^{\text{te}} \quad (10.196)$$

Donc finalement, nous appelons "**intégrale indéfinie**" de la fonction $f(x)$ et nous notons:

$$\int f(x) dx \quad (10.197)$$

toute expression de la forme $F(x) + c^{\text{te}}$ où $F(x)$ est une primitive de $f(x)$. Ainsi, par convention d'écriture:

$$\boxed{\int f(x) dx = F(x) + c^{\text{te}}} \quad (10.198)$$

si et seulement si $F'(x) = f(x)$.

Dans ce contexte, $f(x)$ est également appelée "**fonction à intégrer**" et $f(x)dx$, "**fonction sous le signe somme**".

Géométriquement, nous pouvons considérer l'intégrale indéfinie comme un ensemble (famille) de courbes telles que nous passons de l'une à l'autre en effectuant une translation dans le sens positif ou négatif de l'axe des ordonnées.

Revenons-en à la démonstration du point (2) du théorème fondamental de l'analyse:

Démonstration:

Soit F une primitive de f .

Puisque deux primitives diffèrent d'une constante, nous avons bien:

$$A(X) = F(X) + c^{\text{te}} \quad (10.199)$$

ce que nous pouvons écrire aussi:

$$\int_a^X f(t) dt = F(X) + c^{\text{te}} \quad (10.200)$$

pour tout X dans $[a, b]$. Le cas particulier $X = a$ donne $\int_a^a f(t) dt = 0$ et donc $F(a) + c^{\text{te}} = 0$ et $c^{\text{te}} = -F(a)$. En remplaçant, nous obtenons:

$$\int_a^X f(t) dt = F(X) - F(a) \quad (10.201)$$

Comme cette identité est valable pour tout X de l'intervalle $[a, b]$, elle est vraie en particulier pour $X = b$. D'où:

$$\boxed{\int_a^b f(t) dt = F(b) - F(a)} \quad (10.202)$$

Ce dernier résultat montre aussi quelque chose d'utile!: Il n'est pas nécessaire lorsque nous évaluons une intégrale de prendre en compte la constante de la primitive générale puisque celle-ci s'annule de par la différence des deux primitives!!

Remarques:

R1. Le théorème fondamental qui montre le lien entre primitive et intégrale a conduit à utiliser le même symbole $\int$ pour écrire une primitive (introduit par Leibniz à la fin du 17ème siècle), qui est une fonction, et une intégrale, qui elle, est un nombre.

R2. Nous avons également démontré dans le chapitre de Mécanique Analytique comment calculer à l'aide d'une intégrale la longueur d'une courbe dans le plan si la fonction $f(x)$ est explicitement connue.

Voici quelques propriétés triviales de l'intégration qu'il est bon de se rappeler car souvent utilisées ailleurs sur le site (si cela ne vous semble pas évident, contactez-nous et nous le détaillerons):

P1. La dérivée d'une intégrale indéfinie est égale à la fonction à intégrer:

$$\left(\int f(x)dx\right)' = \left(F(x) + c^{\#}\right)' = f(x) \quad (10.203)$$

P2. La différentielle d'une intégrale indéfinie est égale à l'expression sous le signe somme:

$$d\left(\int f(x)dx\right) = f(x)dx \quad (10.204)$$

P3. L'intégrale indéfinie de la différentielle d'une certaine fonction est égale à la somme de cette fonction et d'une constante arbitraire:

$$\int dF(x) = F(x) + c^{\#} \quad (10.205)$$

P4. L'intégrale indéfinie de la somme (ou soustraction) algébrique de deux ou plusieurs fonctions est égale à la somme algébrique de leurs intégrales (ne pas oublier que l'on travaille avec l'ensemble des primitives et non des primitives particulières!):

$$\int [f_1(x) + f_2(x)]dx = \int f_1(x)dx + \int f_2(x)dx \quad (10.206)$$

Démonstration:

Pour démontrer cela nous allons prouver que la dérivée du membre de gauche permet de trouver le membre de droite et inversement (réciproque) à l'aide des propriétés précédentes.

D'après P1 nous avons:

$$\left(\int [f_1(x) + f_2(x)]dx\right)' = f_1(x) + f_2(x) \quad (10.207)$$

Vérifions s'il en est de même avec le membre de droite (nous supposons connues les propriétés des dérivées que nous avons démontrées au début de ce chapitre):

$$\left(\int f_1(x) dx + \int f_2(x) dx \right)' = \left(\int f_1(x) dx \right)' + \left(\int f_2(x) dx \right)' = f_1(x) + f_2(x) \quad (10.208)$$

□C.Q.F.D.

P5. Nous pouvons sortir un facteur constant de sous le signe somme, c'est-à-dire:

$$\int c^{te} f(x) dx = c^{te} \int f(x) dx \quad (10.209)$$

Nous justifions cette égalité en dérivant les deux membres (et d'après les propriétés des dérivées):

$$\begin{aligned} \left(\int c^{te} f(x) dx \right)' &= c^{te} f(x) \\ \left(c^{te} \int f(x) dx \right)' &= c^{te} \left(\int f(x) dx \right)' = c^{te} f(x) \end{aligned} \quad (10.210)$$

P6. Nous pouvons sortir un facteur constant de l'argument de la fonction intégrée (plutôt rarement utilisée) :

$$\int f(ax) dx = \frac{1}{a} F(ax) + c^{te} \quad (10.211)$$

En effet, en dérivant les deux membres de l'égalité nous avons d'après les propriétés des dérivées:

$$\left(\int f(ax) dx \right)' = \left(\frac{1}{a} F(ax) + c^{te} \right)' = f(ax) \quad (10.212)$$

P7. L'intégration d'une fonction dont l'argument est sommé (ou soustrait) algébriquement est la primitive de l'argument sommé (respectivement soustrait):

$$\int f(x+b) dx = F(x+b) + c^{te} \quad (10.213)$$

Cette propriété se démontre également identiquement à la précédente à l'aide des propriétés des dérivées.

P8. La combinaison des propriétés P6 et P7 nous permet d'écrire:

$$\int f(ax+b) dx = \frac{1}{a} F(ax+b) + c^{te} \quad (10.214)$$

P9. Soit f une fonction continue sur $[a,b]$, nous avons pour tout c appartenant à cet intervalle:

$$\int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx \quad (10.215)$$

Ce théorème, appelé parfois "**relation de Chasles**" (de par son pendant vectoriel), découle immédiatement de la définition de l'intégrale indéfinie. F étant une primitive de f sur $[a,b]$ nous avons:

$$F(c) - F(a) + F(b) - F(c) = F(b) - F(a) \quad (10.216)$$

P10. Voilà une propriété souvent utilisée dans le chapitre de Statistiques du site (nous ne trouvons pas de moyen d'exprimer cette propriété par le langage courant donc...):

$$\frac{d}{dy} \int_{g(x)}^{h(y)} f(t) dt = \frac{d}{dy} (F(h(y)) - F(g(x))) = f(h(y)) \frac{dh(y)}{dy} - f(g(x)) \frac{dg(x)}{dy} = f(h(y)) h'(y) \quad (10.217)$$

Voyons deux propriétés qui nous seront parfois utiles pour calculer des intégrales difficiles:

P11. Si une fonction est paire (cf. chapitre d'Analyse Fonctionnelle), l'intégrale sur des bornes symétriques équivaut à:

$$\begin{aligned} \int_{-L}^{+L} f(x) dx &= \int_{-L}^0 f(x) dx + \int_0^{+L} f(x) dx = \int_{-L}^0 f(-x) dx + \int_0^{+L} f(x) dx \\ &= - \int_{u=-x}^0 f(u) du + \int_0^{+L} f(x) dx = \int_0^{+L} f(u) du + \int_0^{+L} f(x) dx = 2 \int_0^{+L} f(x) dx \end{aligned} \quad (10.218)$$

P12. Si une fonction est impaire (cf. chapitre d'Analyse Fonctionnelle), l'intégrale sur des bornes symétriques équivaut à:

$$\begin{aligned} \int_{-L}^{+L} f(x) dx &= \int_{-L}^0 f(x) dx + \int_0^{+L} f(x) dx = \int_{-L}^0 -f(-x) dx + \int_0^{+L} f(x) dx \\ &= \int_0^{+L} f(-x) dx + \int_0^{+L} f(x) dx = - \int_{u=-x}^0 f(u) du + \int_0^{+L} f(x) dx = 0 \end{aligned} \quad (10.219)$$

2.3. INTÉGRALE DOUBLE

L'idée des intégrales doubles est de mesurer le volume de la zone délimitée par le graphe d'une fonction de deux variables, au-dessus d'un domaine D du plan (ci-dessous D est rectangulaire).

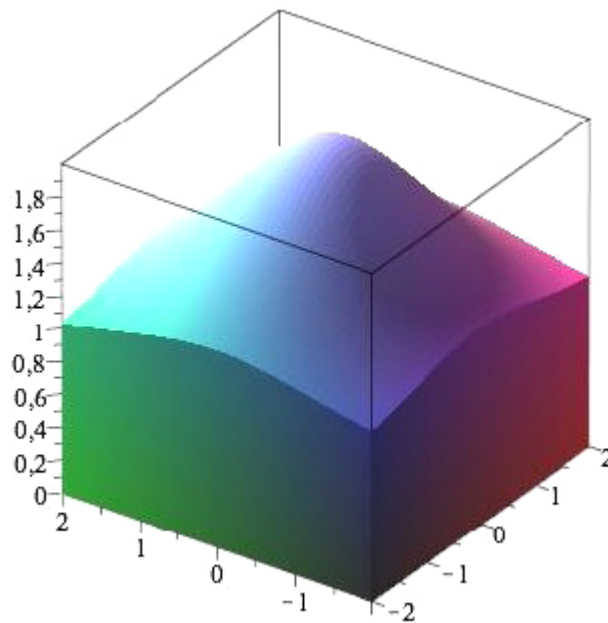


Figure: 10.13 - Exemple d'une fonction à deux variables au-dessus d'un domaine

Il va sans dire que les intégrales doubles sont extrêmement importantes aussi dans tout le domaine des mathématiques appliquées!

Là encore, l'idée est la même que l'intégrale définie. Si nous adaptons une approche simpliste, nous décomposons la fonction continue en un escalier et le volume à calculer se réduit alors à faire la somme des volumes de parallélépipèdes:

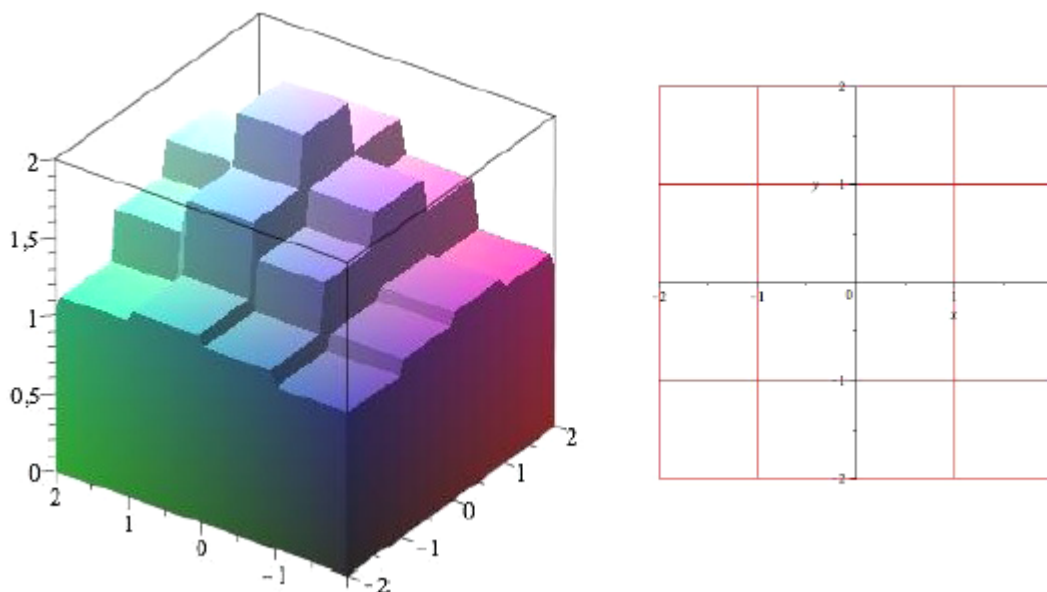


Figure: 10.14 - Décomposition du volume en parallélépipèdes grossiers

Dès lors:

$$V \cong \sum f(x_i, y_i) \Delta x_i \Delta y_i = \sum f(x_i, y_i) \Delta y_i \Delta x_i \quad (10.220)$$

Pour une fonction continue, nous procédons par approximations successives: nous calculons des sommes de Riemann pour des subdivisions de plus en plus fines du domaine D :

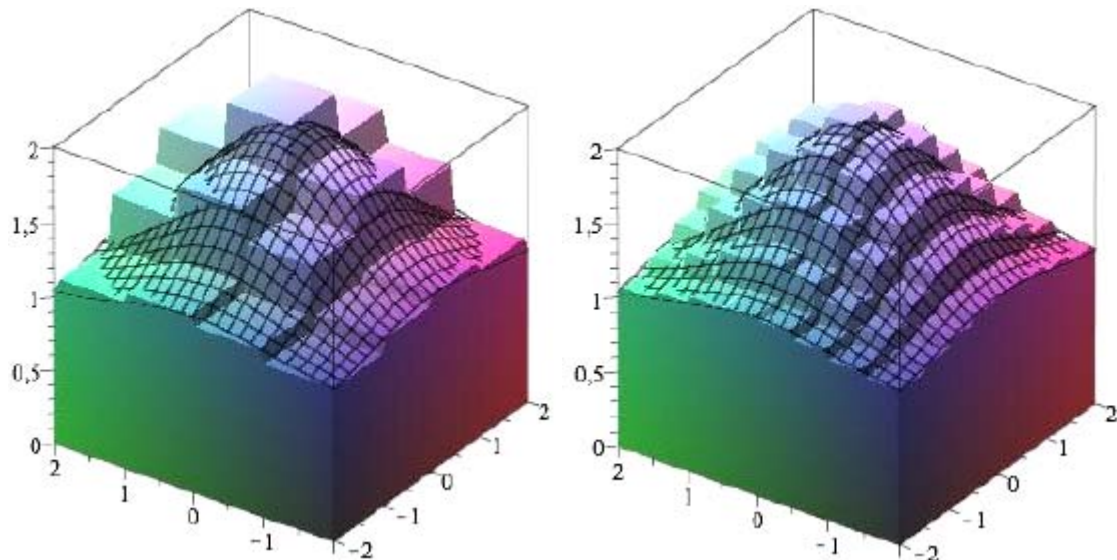


Figure: 10.15 - Décomposition du volume en parallélépipèdes de plus en plus fins

et donc à la limite:

$$V \cong \iint_D f(x, y) dx dy = \iint_D f(x, y) dy dx \quad (10.221)$$

Par contre, quand on veut intégrer sur un domaine qui n'est pas rectangulaire, les choses se compliquent à priori... Voyons comment contourner le problème.

Pour cela, nous allons construire le domaine D fermé borné de la façon suivante.

$$(x, y) \in D \Leftrightarrow \begin{cases} x \in [a, b] \\ y \in [u(x), v(x)] \end{cases} \quad (10.222)$$

où le lecteur aura remarqué que le support de y est la variable x par l'intermédiaire de deux fonctions u et v . C'est ce que nous appelons alors un "**domaine du type I**" (et donc si c'est y qui paramétrise x alors il s'agit domaine de type II).

Ce qui peut s'illustrer par la figure ci-dessous:

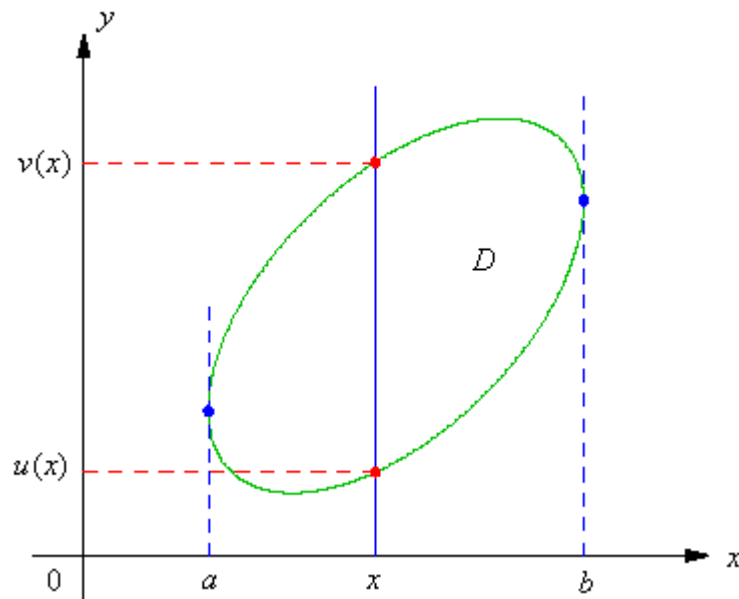


Figure: 10.16 - Exemple d'un domaine de type I

où nous remarquons que cette approche simpliste (il existe d'autres approches possibles mais qui nécessitent de faire appel à la théorie de la mesure) nécessite que le domaine soit simplement connexe (qu'il n'y ait pas de trous hors du domaine D entre $u(x)$ et $v(x)$) ou décomposé en sous-domaines disjointes simplement connexes.

Indiquons que dans la pratique (voir les différents chapitres du site et particulièrement celui sur les Formes Géométriques) les doubles intégrales se font souvent sur des volumes de révolution ce qui simplifie considérablement la paramétrisation. De plus, nous verrons plus loin qu'il est possible de changer de système de coordonnées pour simplifier encore plus les doubles intégrales, ce qui fait que la paramétrisation semble disparaître...

Bref, nous pouvons donc intégrer de la manière suivante:

$$V = \iint_D f(x,y) dx dy = \int_a^b \left(\int_{u(x)}^{v(x)} f(x,y) dy \right) dx \quad (10.223)$$

Donc nous transformons l'intégrale double en deux intégrales simples emboîtées.

2.3.1. THÉORÈME DE FUBINI

Nous allons voir un théorème important utilisé à de nombreuses reprises dans différents chapitres du site et qui permet d'inverser l'ordre d'intégration.

En se rappelant que:

$$(x,y) \in D \Leftrightarrow \begin{cases} x \in [a,b] \\ y \in [u(x),v(x)] \end{cases} \quad (10.224)$$

nous pouvons aussi utiliser:

$$(x, y) \in D \Leftrightarrow \begin{cases} y \in [c, d] \\ x \in [\alpha(y), \beta(y)] \end{cases} \quad (10.225)$$

Ainsi avec cette paramétrisation nous pouvons écrire:

$$V = \iint_D f(x, y) dx dy = \int_a^b \left(\int_{u(x)}^{v(x)} f(x, y) dy \right) dx = \int_c^d \left(\int_{\alpha(y)}^{\beta(y)} f(x, y) dx \right) dy \quad (10.226)$$

Nous pouvons ainsi changer l'ordre d'intégration, le calcul est différent, mais le résultat est le même. Mais ce n'est pas cela qui nous intéresse en réalité ici.

Considérons une fonction telle que:

$$f(x, y) = \varphi(x)\psi(y) \quad (10.227)$$

Alors:

$$V = \iint_D f(x, y) dx dy = \int_a^b \left(\int_{u(x)}^{v(x)} \varphi(x)\psi(y) dy \right) dx = \int_a^b \left(\int_{u(x)}^{v(x)} \varphi(x) \psi(y) dy \right) dx \quad (10.228)$$

Supposons que le domaine est un rectangle (nous faisons cette simplification sinon la démonstration se complique nettement). C'est-à-dire:

$$(x, y) \in D \Leftrightarrow \begin{cases} x \in [a, b] \\ y \in [c, d] \end{cases} \quad (10.229)$$

Dès lors par la propriété de linéarité des intégrales:

$$\begin{aligned} V &= \int_a^b \left(\int_{u(x)}^{v(x)} \varphi(x)\psi(y) dy \right) dx = \int_a^b \left(\int_c^d \varphi(x)\psi(y) dy \right) dx = \int_a^b \left(\varphi(x) \int_c^d \psi(y) dy \right) dx \\ &= \int_a^b \varphi(x) \left(\int_c^d \psi(y) dy \right) dx = \left(\int_c^d \psi(y) dy \right) \int_a^b \varphi(x) dx = \int_a^b \varphi(x) dx \int_c^d \psi(y) dy \end{aligned} \quad (10.230)$$

2.4. INTÉGRATION PAR CHANGEMENT DE VARIABLES

Lorsque nous ne pouvons facilement déterminer la primitive d'une fonction donnée, nous pouvons nous débrouiller par un changement de variable astucieux (parfois même très subtil) pour contourner la difficulté. Cela ne marche pas à tous les coups (car certaines fonctions ne sont pas intégrables formellement) mais il vaut la peine d'essayer avant d'avoir recours à l'ordinateur.

À nouveau, nous ne donnons que la forme générale de la méthode. C'est le rôle des professeurs dans les écoles d'entraîner les élèves à comprendre et maîtriser ce genre de techniques. De plus, les chapitres traitant des sciences exactes sur le site (physique, informatique, astrophysique, chimie, ...) regorgent d'exemples utilisant cette technique et servent ainsi implicitement d'exercices de style.

Soit à calculer l'intégrale (non bornée pour l'instant):

$$\int f(x) dx \quad (10.231)$$

bien que nous ne sachions pas calculer directement la primitive de cette fonction $f(x)$ (en tout cas nous imaginons être dans une telle situation) nous savons (d'une manière ou d'une autre) qu'elle existe (nous ne traitons pas encore des intégrales impropres à ce niveau).

La technique consiste alors dans cette intégrale à effectuer le changement de variable:

$$x = \varphi(t) \quad (10.232)$$

où $\varphi(t)$ est une fonction continue ainsi que sa dérivée, et admettant une fonction inverse. Alors $dx = \varphi'(t)dt$, démontrons que dans ce cas l'égalité:

$$\int f(x) dx = \int f(\varphi(t)) \varphi'(t) dt \quad (10.233)$$

est satisfaite.

Nous sous-entendons ici que la variable t sera remplacée après intégration du membre droit par son expression en fonction de x . Pour justifier l'égalité en ce sens, il suffit de montrer que les deux quantités considérées dont chacune n'est définie qu'à une constante arbitraire près ont la même dérivée par rapport à x . La dérivée du membre gauche est:

$$\left(\int f(x) dx \right)' = f(x) \quad (10.234)$$

Nous dérivons le membre droit par rapport à x en tenant compte que t est une fonction de x . Nous savons que:

$$\frac{dx}{dt} = \varphi'(t) \quad (10.235)$$

Nous avons par conséquent:

$$\frac{d}{dx} \left(\int f(\varphi(t)) \varphi'(t) dt \right) = \frac{d}{dt} \left(\int f(\varphi(t)) \varphi'(t) dt \right) \frac{dt}{dx} = f(\varphi(t)) \varphi'(t) \frac{1}{\varphi'(t)} = f(\varphi(t)) = f(x) \quad (10.236)$$

Les dérivées par rapport à x des deux membres de l'égalité de départ sont donc égales.

□C.Q.F.D.

Bien évidemment, la fonction $x = \varphi(t)$ doit être choisie de manière à ce que nous sachions calculer l'intégrale indéfinie figurant à droite de l'égalité.

Remarque: Il est parfois préférable de choisir le changement de variable sous la forme $t = \varphi(x)$ au lieu de $x = \varphi(t)$ car cela a une large tendance à simplifier la longueur de l'équation au lieu de l'allonger.

2.4.1. JACOBIEN

Considérons un domaine D du plan u,v limité par une courbe L . Supposons que les coordonnées x,y soient des fonctions des nouvelles variables u,v (toujours dans le cadre d'un changement de variables donc) par les relations de transformations:

$$\begin{aligned} x &= \varphi(u, v) & y &= \phi(u, v) \\ (x, y) \in D &\xrightarrow{\varphi, \phi} (\varphi, \phi) \in D' \end{aligned} \quad (10.237)$$

où les fonctions $\varphi(u, v)$ et $\phi(u, v)$ sont univoques, continues et possèdent des dérivées continues dans un certain domaine D' que nous définirons par la suite. Il correspond alors d'après les relations précédentes à tout couple de valeurs u, v un seul couple de valeur x, y et réciproquement.

Il résulte de ce qui précède qu'à tout point $P(x, y)$ du plan Oxy correspond univoquement un point $P'(u, v)$ du plan Ouv de coordonnées u, v définies par les relations précédentes. Les nombres v et u seront appelées "**coordonnées curvilignes**" de P et nous verrons des exemples concrets et schématisés de ceux-ci dans le chapitre de Calcul Vectoriel.

Si dans le plan Oxy le point P décrit la courbe fermée L délimitant le domaine D , le point correspondant décrit dans le plan Ouv un certain domaine D' . Il correspond alors à tout point de D' un point de D . Ainsi, les relations de transformations établissent une correspondance biunivoque entre les points des domaines D et D' .

Considérons maintenant dans D' une droite $u = c^{te}$. En général, les relations de transformation lui font correspondre dans le plan Oxy une ligne courbe (ou inversement). Ainsi, découpons le domaine D' par des droites $u = c^{te}$ et $v = c^{te}$ en de petits domaines rectangulaires (nous ne prendrons pas en compte dans la limite, les rectangles empiétant sur la frontière de D'). Les courbes correspondantes du domaine D découpent alors ce dernier en quadrilatère (définis par des courbes donc). Évidemment, l'inverse est applicable.

Considérons dans le plan Ouv le rectangle $\Delta s'$ limité par les droites:

$$\begin{aligned} u &= c^{te}, u + \Delta u = c^{te} \\ v &= c^{te}, v + \Delta v = c^{te} \end{aligned} \quad (10.238)$$

et le quadrilatère curviligne correspondant Δs dans le plan Oxy . Nous désignerons les aires de ces domaines partiels également par $\Delta s'$ et Δs . Nous avons évidemment:

$$\Delta s' = \Delta u \Delta v \quad (10.239)$$

Les aires $\Delta s'$ et Δs peuvent être en général différentes.

Supposons donc dans D une fonction continue $z = f(x, y)$. Il correspond à toute valeur de cette fonction du domaine D la même valeur $z = F(u, v)$ (ce qu'il faut vérifier) dans D' , où:

$$F(u, v) = f(x, y) = f(\varphi(u, v), \phi(u, v)) \quad (10.240)$$

Considérons les sommes intégrales de la fonction z dans le domaine D . Nous avons évidemment l'égalité suivante:

$$\sum f(x, y) \Delta s = \sum F(u, v) \Delta s \quad (10.241)$$

Calculons Δs , c'est-à-dire l'aire du quadrilatère curviligne P_1, P_2, P_3, P_4 dans le plan Oxy :

Déterminons les coordonnées de ses sommets:

$$\begin{array}{lll} P_1(x_1, y_1) & x_1 = \varphi(u, v) & y_1 = \phi(u, v) \\ P_2(x_2, y_2) & x_2 = \varphi(u + \Delta u, v) & y_2 = \phi(u + \Delta u, v) \\ P_3(x_3, y_3) & x_3 = \varphi(u + \Delta u, v + \Delta v) & y_3 = \phi(u + \Delta u, v + \Delta v) \\ P_4(x_4, y_4) & x_4 = \varphi(u, v + \Delta v) & y_4 = \phi(u, v + \Delta v) \end{array} \quad (10.242)$$

Nous assimilerons dans le calcul de l'aire du quadrilatère P_1, P_2, P_3, P_4 les arcs $P_1P_2, P_2P_3, P_3P_4, P_4P_1$ à des segments de droites parallèles. Nous remplacerons en outre les accroissements des fonctions par leurs différentielles. C'est dire que nous faisons abstraction des infiniment petits d'ordre plus élevé que Δu et Δv . Les relations précédentes deviennent alors:

$$\begin{array}{ll} x_1 = \varphi(u, v) & y_1 = \phi(u, v) \\ x_2 = \varphi(u, v) + \frac{\partial \varphi}{\partial u} \Delta u & y_2 = \phi(u, v) + \frac{\partial \phi}{\partial u} \Delta u \\ x_3 = \varphi(u, v) + \frac{\partial \varphi}{\partial u} \Delta u + \frac{\partial \varphi}{\partial v} \Delta v & y_3 = \phi(u, v) + \frac{\partial \phi}{\partial u} \Delta u + \frac{\partial \phi}{\partial v} \Delta v \\ x_4 = \varphi(u, v) + \frac{\partial \varphi}{\partial v} \Delta v & y_4 = \phi(u, v) + \frac{\partial \phi}{\partial v} \Delta v \end{array} \quad (10.243)$$

Sous ces hypothèses, le quadrilatère curviligne $P_1P_2P_3P_4$ peut être assimilé à un parallélogramme. Son aire Δs est approximativement égale au double de l'aire du triangle $P_1P_2P_3$, aire que nous pouvons calculer en utilisant les propriétés du déterminant (comme nous le démontrerons dans le chapitre d'Algèbre Linéaire, le déterminant dans $\mathbb{R}^2$ représente un parallélogramme alors que dans $\mathbb{R}^3$ celui-ci représente le volume d'un parallélépipède):

$$\Delta s \cong |\det(\vec{v}_1, \vec{v}_2)| = |\det(M)| = \left| \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \right| = |a_{11}a_{22} - a_{21}a_{12}| \quad (10.244)$$

Tel que (c'est là qu'il faut faire le meilleur choix pour que l'expression finale soit la plus simple et la plus esthétique, nous procédons par essais successifs et faisons enfin le choix ci-dessous):

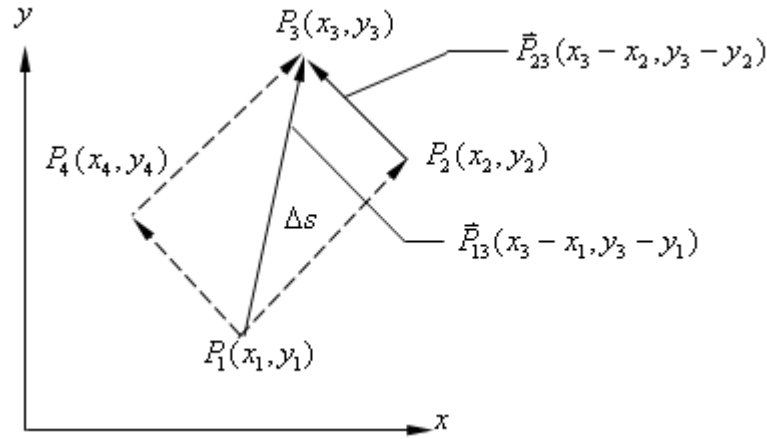


Figure: 10.17 - Représentation graphique du déterminant

Ainsi, nous avons:

$$\begin{aligned}
 \Delta s &\cong |\det(\vec{v}_1, \vec{v}_2)| = |D(\vec{P}_{13}, \vec{P}_{23})| = \left| \det \begin{bmatrix} x_3 - x_1 & x_3 - x_2 \\ y_3 - y_1 & y_3 - y_2 \end{bmatrix} \right| \\
 &= |(x_3 - x_1)(y_3 - y_2) - (x_3 - x_2)(y_3 - y_1)| \\
 &= \left| \left(\frac{\partial \varphi}{\partial u} \Delta u + \frac{\partial \varphi}{\partial v} \Delta v \right) \frac{\partial \phi}{\partial v} \Delta v - \frac{\partial \varphi}{\partial v} \Delta v \left(\frac{\partial \phi}{\partial u} \Delta u + \frac{\partial \phi}{\partial v} \Delta v \right) \right| \\
 &= \left| \frac{\partial \varphi}{\partial u} \frac{\partial \phi}{\partial v} \Delta u \Delta v - \frac{\partial \varphi}{\partial v} \frac{\partial \phi}{\partial u} \Delta u \Delta v \right| = \left| \frac{\partial \varphi}{\partial u} \frac{\partial \phi}{\partial v} - \frac{\partial \varphi}{\partial v} \frac{\partial \phi}{\partial u} \right| \Delta u \Delta v \quad (10.245) \\
 &= |\det(J)| \Delta u \Delta v = \left| \det \begin{bmatrix} \frac{\partial \varphi}{\partial u} & \frac{\partial \varphi}{\partial v} \\ \frac{\partial \phi}{\partial u} & \frac{\partial \phi}{\partial v} \end{bmatrix} \right| \Delta s'
 \end{aligned}$$

Par conséquent la relation suivante (contenant ce qu'il est d'usage d'appeler le "**déterminant fonctionnel**"):

$$\Delta s \cong \left| \det \begin{bmatrix} \frac{\partial \varphi}{\partial u} & \frac{\partial \varphi}{\partial v} \\ \frac{\partial \phi}{\partial u} & \frac{\partial \phi}{\partial v} \end{bmatrix} \right| \Delta s' \quad (10.246)$$

avec:

$$J = \begin{bmatrix} \frac{\partial \varphi}{\partial u} & \frac{\partial \varphi}{\partial v} \\ \frac{\partial \phi}{\partial u} & \frac{\partial \phi}{\partial v} \end{bmatrix} \quad (10.247)$$

qui est la "**matrice jacobienne**" (alors que son déterminant est appelé le "**jacobien**" (tout court)) de la transformation de coordonnées de $\mathbb{R}^2 \rightarrow \mathbb{R}^2$. En appliquant exactement le même raisonnement pour $\mathbb{R}^3$, la matrice jacobienne s'écrit alors (en changeant un peu les notations car sinon cela devient illisible):

$$J = \begin{bmatrix} \frac{\partial P}{\partial u} & \frac{\partial P}{\partial v} & \frac{\partial P}{\partial w} \\ \frac{\partial Q}{\partial u} & \frac{\partial Q}{\partial v} & \frac{\partial Q}{\partial w} \\ \frac{\partial R}{\partial u} & \frac{\partial R}{\partial v} & \frac{\partial R}{\partial w} \end{bmatrix} \quad (10.248)$$

Bref, à quoi cela sert-il concrètement ? Eh bien revenons à notre relation:

$$\Delta s \cong |\det(J)| \Delta s' \quad (10.249)$$

qui n'est finalement qu'approximative étant donné que dans les calculs de l'aire Δs nous avons négligé les infiniment petits d'ordre supérieur. Toutefois, plus les dimensions des domaines élémentaires Δs et $\Delta s'$ sont petites, et plus nous nous approchons de l'égalité. L'égalité ayant finalement lieu quand nous passons à la limite (finalement en maths aussi on fait des approximations... hein !), les surfaces des domaines élémentaires tendant vers zéro:

$$|\det(J)| = \lim_{\Delta s \rightarrow 0} \frac{\Delta s}{\Delta s'} \quad (10.250)$$

Appliquons maintenant l'égalité obtenue au calcul de l'intégrale double (nous pouvons faire de même avec la triple bien sûr). Nous pouvons donc finalement écrire (c'est la seule manière de poser la chose qui ait un sens):

$$\begin{aligned} \sum f(x, y) \Delta s &= \sum F(u, v) \Delta s \\ &\Downarrow \\ \sum f(x, y) \Delta s &\cong \sum F(u, v) |\det(J)| \Delta s' \end{aligned} \quad (10.251)$$

Passant à la limite, nous obtenons l'égalité stricte:

$$\boxed{\iint_D f(x, y) dx dy = \iint_{D'} F(u, v) |\det(J)| du dv} \quad (10.252)$$

Telle est la relation de transformation des coordonnées dans une intégrale double. Elle permet de ramener le calcul d'une intégrale double dans le domaine D au domaine D' , ce qui peut simplifier le problème.

De même, pour une intégrale triple, nous écrirons:

$$\iiint_D f(x, y, z) dx dy dz = \iiint_{D'} F(u, v, w) |\det(J)| du dv dw \quad (10.253)$$

Déterminons maintenant le Jacobien pour les systèmes de coordonnées les plus courants (nous renvoyons à nouveau le lecteur au chapitre de Calcul Vectoriel pour plus d'informations concernant ces systèmes):

1. Coordonnées polaires $x = r \cos \phi, y = r \sin \phi$:

$$\left| \det \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \phi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \phi} \end{bmatrix} \right| = \left| \det \begin{bmatrix} \cos \phi & -r \sin \phi \\ \sin \phi & r \cos \phi \end{bmatrix} \right| = |r| \quad (10.254)$$

Comme r est toujours positif, nous écrivons simplement:

$$|\det(J)| = r \quad (10.255)$$

2. Coordonnées cylindriques $x = r \cos \phi, y = r \sin \phi, z = z$ (cf. chapitre d'Algèbre Linéaire pour le calcul du déterminant):

$$\left| \det \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \phi} & \frac{\partial x}{\partial z} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \phi} & \frac{\partial y}{\partial z} \\ \frac{\partial z}{\partial r} & \frac{\partial z}{\partial \phi} & \frac{\partial z}{\partial z} \end{bmatrix} \right| = \left| \det \begin{bmatrix} \cos \phi & -r \sin \phi & 0 \\ \sin \phi & r \cos \phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \right| = \left| \begin{array}{l} \cos \phi r \cos \phi \cdot 1 - \cos \phi \cdot 0 \cdot 0 \\ + r \sin \phi \sin \phi \cdot 1 - r \sin \phi \cdot 0 \cdot 0 \\ + 0 \cdot \sin \phi \cdot 0 - 0 \cdot r \cos \phi \cdot 0 \end{array} \right| = |r| \quad (10.256)$$

Comme r est toujours positif, nous écrivons simplement:

$$|\det(J)| = r \quad (10.257)$$

3. En coordonnées sphériques $x = r \sin \theta \cos \phi, y = r \sin \theta \sin \phi, z = r \cos \theta$ (cf. chapitre d'Algèbre Linéaire pour le calcul du déterminant):

$$\begin{aligned} \left| \det \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \phi} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \phi} & \frac{\partial y}{\partial \theta} \\ \frac{\partial z}{\partial r} & \frac{\partial z}{\partial \phi} & \frac{\partial z}{\partial \theta} \end{bmatrix} \right| &= \left| \det \begin{bmatrix} \sin \theta \cos \phi & -r \sin \theta \sin \phi & r \cos \theta \cos \phi \\ \sin \theta \sin \phi & r \sin \theta \cos \phi & r \cos \theta \sin \phi \\ \cos \theta & 0 & -r \sin \theta \end{bmatrix} \right| \\ &= \left| \begin{array}{l} -\sin \theta \cos \phi r \sin \theta \cos \phi r \sin \theta - \sin \theta \cos \phi r \cos \theta \sin \phi 0 \\ -r \sin \theta \sin \phi \sin \theta \sin \phi r \sin \theta - r \sin \theta \sin \phi r \cos \theta \sin \phi \cos \theta \\ + r \cos \theta \cos \phi \sin \theta \sin \phi 0 - r \cos \theta \cos \phi r \sin \theta \cos \phi \cos \theta \end{array} \right| \\ &= \left| -(\sin^3 \theta \cos^2 \phi r^2 + \sin^3 \theta \sin^2 \phi r^2) - (r^2 \sin \theta \sin^2 \phi \cos^2 \theta + r^2 \cos^2 \theta \cos^2 \phi \sin \theta) \right| \\ &= \left| -\sin^3 \theta r^2 - r^2 \sin \theta \cos^2 \theta \right| = \left| -r^2 \sin \theta (\sin^2 \theta + \cos^2 \theta) \right| = \left| -r^2 \sin \theta \right| \quad (10.258) \end{aligned}$$

Comme r^2 est toujours positif, nous écrivons simplement:

$$|\det(J)| = r^2 \sin \theta \text{ avec } \sin \theta \geq 0 \quad (10.259)$$

2.5. INTÉGRATION PAR PARTIES

Lorsque nous cherchons à effectuer des intégrations, il est très fréquent que nous ayons à utiliser un outil (ou méthode de calcul) appelé "**intégration par parties**". Il existe différents degrés d'utilisation de cet outil et nous allons commencer par le plus simple et qui est le plus utilisé dans tous les chapitres traitant de physique sur le présent site.

Nous partons d'abord de la dérivée du produit de deux fonctions démontrée plus haut:

$$(f(t) \cdot g(t))' = f'(t)g(t) + f(t) \cdot g'(t) \quad (10.260)$$

nous avons donc:

$$f(t) \cdot g'(t) = (f(t) \cdot g(t))' - f'(t) \cdot g(t) \quad (10.261)$$

et il vient:

$$\int_a^b f(t) \cdot g'(t) dt = \int_a^b [(f(t) \cdot g(t))' - f'(t) \cdot g(t)] dt = \int_a^b (f(t) \cdot g(t))' dt - \int_a^b f'(t) \cdot g(t) dt \quad (10.262)$$

après une dernière simplification nous avons enfin la fameuse relation très importante:

$$\boxed{\int_a^b f(t) \cdot g'(t) dt = f(t) \cdot g(t) \Big|_a^b - \int_a^b f'(t) \cdot g(t) dt} \quad (10.263)$$

Mais nous allons parfois avoir besoin de la généralisation de cette dernière relation. Nous pouvons démontrer que si f et g sont deux applications (fonctions) de classe C^n (dérivables n fois) sur $[a,b]$ dans $\mathbb{C}$, alors :

$$\int_a^b f(t) \cdot g^{(n)}(t) dt = \left[\sum_{k=0}^{n-1} (-1)^k f^{(k)}(t) \cdot g^{(n-1-k)}(t) \right]_a^b + (-1)^n \int_a^b f^{(n)}(t) \cdot g(t) dt \quad (10.264)$$

Démonstration:

Procédons par récurrence sur n (attention ce n'est pas forcément facile à comprendre comme souvent avec les démonstrations par récurrence!).

Tout en sachant la relation vraie pour $n=1$, nous la supposons vraie pour n (comme donnée dans la relation précédente!) et nous la démontrons pour $n+1$ (donc nous devons retomber sur la relation précédente mais avec $n+1$ au lieu de n):

$$\begin{aligned}
\int_a^b f(t) \cdot g^{(n+1)}(t) dt &= f(t) \cdot g^{(n)}(t) \Big|_a^b - \int_a^b f'(t) \cdot g^{(n)}(t) dt \\
&= f(t) \cdot g^{(n)}(t) \Big|_a^b - \left[\sum_{k=0}^{n-1} (-1)^k f^{(k+1)} \cdot g^{(n-1-k)} \right]_a^b + (-1)^{n+1} \int_a^b f^{(n+1)}(t) \cdot g(t) dt \\
&= f(t) \cdot g^{(n)}(t) \Big|_a^b + \left[\sum_{k=1}^n (-1)^k f^{(k)} \cdot g^{(n-k)} \right]_a^b + (-1)^{n+1} \int_a^b f^{(n+1)}(t) \cdot g(t) dt \\
&= \left[\sum_{k=0}^n (-1)^k f^{(k)} \cdot g^{(n-k)} \right]_a^b + (-1)^{n+1} \int_a^b f^{(n+1)}(t) \cdot g(t) dt
\end{aligned} \tag{10.265}$$

Remarque (proposée par un internaute): l'astuce dans cette démonstration est de bien voir que $-(-1)^n$ donne un signe moins quand n est pair et un plus quand n est impair et de même $+(-1)^{n+1}$ donne un signe moins quand n est pair et un plus quand n est impair.

Pour $n=1$ nous retrouvons la formule bien connue et qui sera très très souvent utilisée sur tout le site:

$$\int_a^b f(t) \cdot g'(t) dt = f(t) \cdot g(t) \Big|_a^b - \int_a^b f'(t) \cdot g(t) dt \tag{10.266}$$

□C.Q.F.D.

2.6. PRIMITIVES USUELLES

Il existe en mathématique et en physique un grand nombre de primitives ou de fonctions définies sur des intégrales que nous retrouvons assez fréquemment (mais pas exclusivement). Par ailleurs, toutes les primitives démontrées ci-dessous seront utilisées dans les sections relatives à la Mécanique, l'Ingénierie, l'Atomistique, les Mathématiques Sociales, etc. Donc, comme dans n'importe quel formulaire, nous vous proposons les primitives connues mais avec les démonstrations!

Cependant, nous omettons les primitives qui découlent déjà des dérivées que nous avons démontrées plus haut. Ce qui signifie par exemple que nous supposons connues les deux primitives très importantes (certainement les plus utilisées dans l'ensemble des pages du site):

$$\begin{aligned}
\int \sin(x) dx &= -\cos(x) \\
\int \cos(x) dx &= \sin(x)
\end{aligned} \tag{10.267}$$

Sinon voici déjà une liste de quelques intégrales fréquentes (le lecteur en rencontrera de toute façon bien d'autres - développées dans les détails - lors de son parcours du site):

1. Primitive de $f(x) = \tan(x)$:

Par définition nous avons donc:

$$\int \tan(x) \, dx = \int \frac{\sin(x)}{\cos(x)} \, dx \quad (10.268)$$

Nous utilisons le changement de variable $u = \cos(x)$, $du = -\sin(x)dx$ et ainsi:

$$\int \frac{\sin(x)}{\cos(x)} \, dx = -\int \frac{1}{u} \, du = -\ln|u| = -\ln|\cos(x)| \quad (10.269)$$

Donc:

$$F(x) = -\ln|\cos(x)| \quad (10.270)$$

2. Primitive de $f(x) = \cot(x)$:

Par définition nous avons donc:

$$\int \cot(x) \, dx = \int \frac{\cos(x)}{\sin(x)} \, dx \quad (10.271)$$

Nous utilisons le changement de variable $u = \sin(x)$, $du = \cos(x)dx$ et:

$$\int \frac{\cos(x)}{\sin(x)} \, dx = \int \frac{1}{u} \, du = \ln|u| = \ln|\sin(x)| \quad (10.272)$$

Donc:

$$F(x) = \ln|\sin(x)| \quad (10.273)$$

3. Primitive de $f(x) = \arcsin(x)$:

Nous intégrons par parties:

$$\int 1 \cdot \arcsin(x) \, dx = x \cdot \arcsin(x) - \int \frac{x}{\sqrt{1-x^2}} \, dx \quad (10.274)$$

Si nous posons $u = 1 - x^2$, ce qui nous donne $du = -2x \, dx$, nous obtenons:

$$\int \frac{x}{\sqrt{1-x^2}} \, dx = -\int \frac{1}{2\sqrt{u}} \, du = -\sqrt{u} = -\sqrt{1-x^2} \quad (10.275)$$

Donc:

$$F(x) = x \cdot \arcsin(x) + \sqrt{1-x^2} \quad (10.276)$$

4. Primitive de $f(x) = \arccos(x)$:

Nous intégrons à nouveau par parties:

$$\int 1 \cdot \arccos(x) dx = x \cdot \arccos(x) + \int \frac{x}{\sqrt{1-x^2}} dx \quad (10.277)$$

Si nous posons $u = 1 - x^2$, ($du = -2x dx$), nous obtenons:

$$\int \frac{x}{\sqrt{1-x^2}} dx = - \int \frac{1}{2\sqrt{u}} du = -\sqrt{u} = -\sqrt{1-x^2} \quad (10.278)$$

Donc:

$$F(x) = x \cdot \arccos(x) - \sqrt{1-x^2} \quad (10.279)$$

5. Primitive de $f(x) = \arctan(x)$:

Nous intégrons encore une fois par parties:

$$\int 1 \cdot \arctan(x) dx = x \cdot \arctan(x) - \int \frac{x}{1+x^2} dx \quad (10.280)$$

Si nous posons $u = 1 + x^2$, ($du = 2x dx$), nous obtenons:

$$\int \frac{x}{1+x^2} dx = \int \frac{1}{2u} du = \frac{1}{2} \cdot \ln|u| = \frac{1}{2} \cdot \ln(1+x^2) \quad (10.281)$$

Donc:

$$F(x) = x \cdot \arctan(x) - \frac{1}{2} \cdot \ln(1+x^2) \quad (10.282)$$

6. Primitive de $f(x) = \operatorname{arccot}(x)$:

Encore une fois... nous intégrons par parties:

$$\int 1 \cdot \operatorname{arccot}(x) dx = x \cdot \operatorname{arccot}(x) + \int \frac{x}{1+x^2} dx \quad (10.283)$$

Si nous posons $u = 1 + x^2$, ($du = 2x dx$), nous obtenons:

$$\int \frac{x}{1+x^2} dx = \int \frac{1}{2u} du = \frac{1}{2} \cdot \ln|u| = \frac{1}{2} \cdot \ln(1+x^2) \quad (10.284)$$

Donc:

$$F(x) = x \cdot \operatorname{arccot}(x) + \frac{1}{2} \cdot \ln(1+x^2) \quad (10.285)$$

7. Primitive de $f(x) = xe^{ax}$ avec $a \in \mathbb{R} \setminus \{0\}$:

Une intégration par parties nous donne:

$$\int x e^{ax} dx = \frac{x}{a} e^{ax} - \frac{1}{a} \int e^{ax} dx = \frac{x}{a} e^{ax} - \frac{1}{a^2} e^{ax} = e^{ax} \left(\frac{1}{a} x - \frac{1}{a^2} \right) \quad (10.286)$$

Donc:

$$F(x) = e^{ax} \left(\frac{1}{a} x - \frac{1}{a^2} \right) \quad (10.287)$$

Remarque: Une autre intégrale très importante avec l'exponentielle en physique est celle que nous avons démontrée lors de notre étude de la loi de Gauss-Laplace en statistiques et probabilités (détermination de la moyenne).

8. Primitive de $f(x) = \ln(x)$:

$$\int \ln(x) dx = \int 1 \cdot \ln(x) dx \quad (10.288)$$

en intégrant par parties nous trouvons:

$$\int 1 \cdot \ln(x) dx = x \ln(x) - \int x \frac{1}{x} dx = x \ln(x) - x = x(\ln(x) - 1) \quad (10.289)$$

Donc:

$$F(x) = x \cdot (\ln(x) - 1) \quad (10.290)$$

9. Primitive de $f(x) = x \ln(ax)$ avec $a \in \mathbb{R} \setminus \{0\}$:

Une intégration par parties nous donne:

$$\int x \ln(ax) dx = \frac{1}{2} x^2 \ln(ax) - \frac{1}{2} \int x^2 \frac{1}{x} dx = \frac{1}{2} x^2 \ln(ax) - \frac{1}{4} x^2 \quad (10.291)$$

Donc:

$$F(x) = \frac{1}{2} x^2 \ln(ax) - \frac{1}{4} x^2 \quad (10.292)$$

10. Primitive de $f(x) = a^x$ pour $a > 0, a \neq 1$:

$$f(x) = a^x \Leftrightarrow f'(x) = a^x \cdot \ln(a) \quad (10.293)$$

Ainsi il vient:

$$\frac{f'(x)}{\ln(a)} = a^x \Rightarrow \int \frac{f'(x)}{\ln(a)} dx = \int a^x dx \quad (10.294)$$

Il vient:

$$\frac{1}{\ln(a)} \int f'(x) dx = \frac{f(x)}{\ln(a)} \text{ et } \int a^x dx = \int f(x) dx = F(x) \quad (10.295)$$

d'où:

$$F(x) = \frac{f(x)}{\ln(a)} = \frac{a^x}{\ln(a)} \quad (10.296)$$

11. Primitive de $f(x) = \log_a(x)$:

Pour ($a > 0, a \neq 1$) sachant que (voir les propriétés des logarithmes dans le chapitre d'analyse fonctionnelle):

$$\log_a(x) = \frac{\ln(x)}{\ln(a)} = \frac{1}{\ln(a)} \ln(x) \quad (10.297)$$

nous avons en utilisant la primitive de $\ln(x)$:

$$F(x) = \frac{x}{\ln(a)} \cdot (\ln(x) - 1) \quad (10.298)$$

12. Primitive de $f(x) = \tanh(x)$:

Nous avons:

$$\int \tanh(x) dx = \int \frac{\sinh(x)}{\cosh(x)} dx \quad (10.299)$$

Nous utilisons le changement de variable $u = \cosh(x), du = \sinh(x) dx$ et obtenons:

$$\int \frac{\sinh(x)}{\cosh(x)} dx = \int \frac{1}{u} du = \ln|u| = \ln(\cosh(x)) \quad (10.300)$$

Donc:

$$F(x) = \ln(\cosh(x)) \quad (10.301)$$

13. Primitive de $f(x) = \coth(x)$:

Nous avons donc:

$$\int \coth(x) dx = \int \frac{\cosh(x)}{\sinh(x)} dx \quad (10.302)$$

Nous utilisons le changement de variable $u = \sinh(x), du = \cosh(x) dx$ et obtenons:

$$\int \frac{\cosh(x)}{\sinh(x)} dx = \int \frac{1}{u} du = \ln|u| = \ln|\sinh(x)| \quad (10.303)$$

Donc:

$$F(x) = \ln(|\sinh(x)|) \quad (10.304)$$

14. Primitive de $f(x) = \operatorname{arcsinh}(x)$:

Nous intégrons par parties:

$$\int 1 \cdot \operatorname{arcsinh}(x) dx = x \cdot \operatorname{arcsinh}(x) - \int \frac{x}{\sqrt{1+x^2}} dx \quad (10.305)$$

Si nous posons $u = 1 + x^2$, ($du = 2x dx$) nous obtenons:

$$\int \frac{x}{\sqrt{1+x^2}} dx = \int \frac{1}{2\sqrt{u}} du = \sqrt{u} = \sqrt{1+x^2} \quad (10.306)$$

Donc:

$$F(x) = x \cdot \operatorname{arcsinh}(x) - \sqrt{1+x^2} \quad (10.307)$$

15. Primitive de $f(x) = \operatorname{arccosh}(x)$:

Nous intégrons par parties:

$$\int 1 \cdot \operatorname{arccosh}(x) dx = x \cdot \operatorname{arccosh}(x) - \int \frac{x}{\sqrt{x^2-1}} dx \quad (10.308)$$

Si nous posons $u = x^2 - 1$, ce qui nous donne $du = 2x dx$, nous obtenons:

$$\int \frac{x}{\sqrt{x^2-1}} dx = \int \frac{1}{2\sqrt{u}} du = \sqrt{u} = \sqrt{x^2-1} \quad (10.309)$$

Donc finalement:

$$F(x) = x \cdot \operatorname{arccosh}(x) - \sqrt{x^2-1} \quad (10.310)$$

16. Primitive de $f(x) = \operatorname{arctanh}(x)$:

Nous intégrons par parties:

$$\int 1 \cdot \operatorname{arctanh}(x) dx = x \cdot \operatorname{arctanh}(x) - \int \frac{x}{1-x^2} dx \quad (10.311)$$

Si nous posons $u = 1 - x^2$, ce qui nous donne $du = -2x dx$, nous obtenons:

$$\int \frac{x}{1-x^2} dx = -\int \frac{1}{2u} du = -\frac{1}{2} \ln|u| = -\frac{1}{2} \ln|1-x^2| \quad (10.312)$$

Donc finalement:

$$F(x) = x \cdot \operatorname{arctanh}(x) + \frac{1}{2} \ln(1 - x^2) \quad (10.313)$$

17. Primitive de $f(x) = \operatorname{arccoth}(x)$:

Nous intégrons par parties:

$$\int 1 \cdot \operatorname{arccoth}(x) dx = x \cdot \operatorname{arccoth}(x) - \int \frac{x}{1 - x^2} dx \quad (10.314)$$

Si nous posons $u = 1 - x^2$, ($du = -2x dx$) nous obtenons:

$$\int \frac{x}{1 - x^2} dx = - \int \frac{1}{2u} du = - \frac{1}{2} \ln|u| = - \frac{1}{2} \ln|1 - x^2| \quad (10.315)$$

Donc finalement:

$$F(x) = x \cdot \operatorname{arccoth}(x) + \frac{1}{2} \ln|1 - x^2| \quad (10.316)$$

18. Primitive de $f(x) = \sin^n(x)$ avec $n \geq 2$:

Posons $I_n = \int \sin^n x \cdot dx$. Une intégration par partie donne:

$$I_n = \int \sin x \cdot \sin^{n-1} x \cdot dx = -\cos x \cdot \sin^{n-1} x + (n-1) \int \cos^2 x \cdot \sin^{n-2} x \cdot dx \quad (10.317)$$

en remplaçant $\cos^2 x$ par $1 - \sin^2 x$ dans la dernière intégrale, nous obtenons:

$$I_n = -\cos x \cdot \sin^{n-1} x + (n-1)(I_{n-2} - I_n) \quad (10.318)$$

et donc:

$$I_n = \frac{1}{n} \left(-\cos x \cdot \sin^{n-1} x + (n-1) I_{n-2} \right) \quad (10.319)$$

19. Primitive de $f(x) = \cos^n(x)$ avec $n \geq 2$:

Dans ce cas nous avons la formule de récurrence

$$\int \cos^n x dx = \frac{1}{n} \left((n-1) \int \cos^{n-2} x dx + \sin x \cdot \cos^{n-1} x \right) \quad (10.320)$$

qui se démontre de la même façon que la relation de récurrence précédente.

20. Primitive de $f(x) = \tan^2(x)$:

Sachant que $\tan'(x) = 1 + \tan^2(x)$, nous avons:

$$\int \tan^2(x) dx = \int (\tan'(x) - 1) dx = \tan(x) - x \quad (10.321)$$

Donc:

$$F(x) = \tan(x) - x \quad (10.322)$$

21. Intégrale de $f(x) = \cot^2(x)$:

Sachant que $\cot'(x) = -1 - \cot^2(x)$, nous avons:

$$\int \cot^2(x) dx = \int (-\cot'(x) - 1) dx = -\cot(x) - x \quad (10.323)$$

Donc:

$$F(x) = -\cot(x) - x \quad (10.324)$$

22. Primitive de $f(x) = \sin^{-2}(x)$:

En utilisant les relations trigonométriques remarquables, nous avons:

$$\int \frac{1}{\sin^2(x)} dx = \int \frac{\sin^2(x) + \cos^2(x)}{\sin^2(x)} dx = \int (1 + \cot^2(x)) dx = -\cot(x) \quad (10.325)$$

selon la primitive $\cot^2(x)$. Donc:

$$F(x) = -\cot(x) \quad (10.326)$$

23. Primitive de $f(x) = \cos^{-2}(x)$:

En utilisant encore une fois les relations trigonométriques remarquables, nous avons:

$$\int \frac{1}{\cos^2(x)} dx = \int \frac{\sin^2(x) + \cos^2(x)}{\cos^2(x)} dx = \int (1 + \tan^2(x)) dx = \tan(x) \quad (10.327)$$

selon la primitive $\tan^2(x)$. Donc:

$$F(x) = \tan(x) \quad (10.328)$$

24. Primitive de $f(x) = \sin^{-1}(x)$:

Nous faisons la substitution $x = 2\arctan(t)$ ($t = \tan\left(\frac{x}{2}\right)$). Sachant que:

$$\sin(x) = \frac{2 \tan\left(\frac{x}{2}\right)}{1 + \tan^2\left(\frac{x}{2}\right)} \quad (10.329)$$

(cf. chapitre de Trigonométrie) nous obtenons alors:

$$\sin(x) = \frac{2t}{1+t^2} \text{ et } dx = \frac{2}{1+t^2} dt \quad (10.330)$$

(selon la dérivée de $\arctan(x)$). Donc:

$$\int \frac{1}{\sin x} dx = \int \frac{1}{t} dt = \ln |t| = \ln \left| \tan \frac{x}{2} \right| \quad (10.331)$$

et:

$$F(x) = \ln \left| \tan \frac{x}{2} \right| \quad (10.332)$$

25. Primitive de $f(x) = \cos^{-1}(x)$:

Sachant que $\cos x = \sin \left(x + \frac{\pi}{2} \right)$ (cf. chapitre de Trigonométrie) nous avons:

$$\int \frac{1}{\cos(x)} dx = \int \frac{1}{\sin \left(x + \frac{\pi}{2} \right)} dx \quad (10.333)$$

Nous faisons le changement de variable $x + \frac{\pi}{2} = u$ ($du = dx$):

$$\int \frac{1}{\cos(x)} dx = \int \frac{1}{\sin \left(x + \frac{\pi}{2} \right)} dx = \int \frac{1}{\sin(u)} du = \ln \left| \tan \left(\frac{u}{2} \right) \right| = \ln \left| \tan \left(\frac{x}{2} + \frac{\pi}{4} \right) \right| \quad (10.334)$$

(selon la primitive de $\sin^{-1}(x)$). Donc:

$$F(x) = \ln \left| \tan \left(\frac{x}{2} + \frac{\pi}{4} \right) \right| \quad (10.335)$$

26. Primitive de $f(x) = \frac{1}{1 + \cos(x)}$:

Nous faisons la substitution $x = 2\arctan(t)$ ($t = \tan \left(\frac{x}{2} \right)$). Sachant que (cf. chapitre de Trigonométrie):

$$\cos x = \frac{1 - \tan^2 \left(\frac{x}{2} \right)}{1 + \tan^2 \left(\frac{x}{2} \right)} \quad (10.336)$$

nous obtenons:

$$\cos x = \frac{1-t^2}{1+t^2} \text{ et } dx = \frac{2}{1+t^2} dt \quad (10.337)$$

(selon la dérivée de $\arctan(x)$). Donc:

$$\int \frac{1}{1+\cos(x)} dx = \int 1 dt = t = \tan\left(\frac{x}{2}\right) \quad (10.338)$$

et:

$$F(x) = \tan\left(\frac{x}{2}\right) \quad (10.339)$$

27. Primitive de $f(x) = \frac{1}{1-\cos(x)}$:

Nous faisons à nouveau la substitution $x = 2 \arctan(t)$ (comme précédemment). Nous trouvons alors:

$$\int \frac{1}{1-\cos x} dx = \int \frac{1}{t^2} dt = -\frac{1}{t} = -\cot\left(\frac{x}{2}\right) \quad (10.340)$$

et donc:

$$F(x) = -\cot\frac{x}{2} \quad (10.341)$$

28. Primitive de $f(x) = \frac{1}{1+\sin(x)}$:

Sachant que:

$$\sin x = \cos\left(x - \frac{\pi}{2}\right) \quad (10.342)$$

Nous avons alors:

$$\int \frac{1}{1+\sin x} dx = \int \frac{1}{1+\cos\left(x - \frac{\pi}{2}\right)} dx \quad (10.343)$$

En faisant le changement de variable:

$$u = x - \frac{\pi}{2} \text{ avec } du = dx \quad (10.344)$$

nous obtenons:

$$\int \frac{1}{1 + \sin x} dx = \int \frac{1}{1 + \cos(u)} du = \tan \frac{u}{2} = \tan \left(\frac{x}{2} - \frac{\pi}{4} \right) \quad (10.345)$$

D'où:

$$F(x) = \tan \left(\frac{x}{2} - \frac{\pi}{4} \right) \quad (10.346)$$

29. Primitive de $f(x) = \frac{1}{1 - \sin(x)}$:

Par le même raisonnement que précédemment en utilisant le cosinus nous obtenons:

$$F(x) = -\cot \left(\frac{x}{2} - \frac{\pi}{4} \right) \quad (10.347)$$

30. Primitive de $f(x) = \sinh^n(x)$ avec $n \geq 2$:

Posons:

$$I_n = \int \sinh^n(x) dx \quad (10.348)$$

Une intégration par partie donne (nous avons démontré lors des dérivées usuelles que la primitive du sinus hyperbolique était le cosinus hyperbolique):

$$I_n = \int \sinh(x) \sinh^{n-1}(x) dx = \cosh(x) \sinh^{n-1}(x) - (n-1) \int \cosh^2(x) \cdot \sinh^{n-2}(x) dx \quad (10.349)$$

en remplaçant $\cosh^2(x)$ par $1 + \sinh^2(x)$ dans la dernière intégrale, nous obtenons:

$$I_n = \cosh(x) \sinh^{n-1}(x) - (n-1)(I_{n-2} + I_n) \quad (10.350)$$

et donc:

$$I_n = \int \sinh^n(x) dx = \frac{1}{n} \left(\cosh(x) \sinh^{n-1}(x) - (n-1)I_{n-2} \right) \quad (10.351)$$

Ainsi:

$$\int \sinh^2(x) dx = \frac{1}{2} (\cosh x \cdot \sinh x - x) \quad (10.352)$$

31. Primitive de $f(x) = \cosh^n(x)$ avec $n \geq 2$:

Dans ce cas nous avons aussi la relation de récurrence:

$$I_n = \int \cosh^n(x) dx = \frac{1}{n} \left(\sinh(x) \cosh^{n-1}(x) + (n-1)I_{n-2} \right) \quad (10.353)$$

qui se démontre de la même façon que ci-dessus. Ainsi:

$$\int \cosh^2(x) dx = \frac{1}{2} (\cosh(x) \cdot \sinh(x) + x) \quad (10.354)$$

32. Primitive de $f(x) = \tanh^2(x)$:

Sachant que (démontré lors des dérivées usuelles):

$$\tanh'(x) = 1 - \tanh^2(x) \quad (10.355)$$

nous avons:

$$\int \tanh^2(x) dx = \int (1 - \tanh'(x)) dx = x - \tanh(x) \quad (10.356)$$

Donc:

$$F(x) = x - \tanh(x) \quad (10.357)$$

33. Primitive de $f(x) = \coth^2(x)$:

Sachant que (démontré lors des dérivées usuelles):

$$\coth'(x) = 1 - \coth^2(x) \quad (10.358)$$

nous avons:

$$\int \coth^2(x) dx = \int (1 - \coth'(x)) dx = x - \coth(x) \quad (10.359)$$

Donc:

$$F(x) = x - \coth x \quad (10.360)$$

34. Primitive de $f(x) = \frac{1}{\sinh^2(x)}$:

Nous avons en utilisant la primitive de $\coth^2(x)$:

$$\int \frac{1}{\sinh^2 x} dx = \int \frac{\cosh^2 x - \sinh^2 x}{\sinh^2 x} dx = \int (\coth^2 x - 1) dx = -\coth x \quad (10.361)$$

Donc:

$$F(x) = -\coth x \quad (10.362) .$$

35. Primitive de $f(x) = \frac{1}{\cosh^2(x)}$:

Nous avons en utilisant la primitive de $\tanh^2(x)$:

$$\int \frac{1}{\cosh^2(x)} dx = \int \frac{\cosh^2(x) - \sinh^2(x)}{\cosh^2(x)} dx = \int (1 - \tanh^2(x)) dx = \tanh(x) \quad (10.363)$$

Donc:

$$F(x) = \tanh(x)$$

36. Primitive de $f(x) = \frac{1}{\sinh(x)}$:

Nous faisons la substitution:

$$t = e^x \text{ avec } dx = \frac{dt}{t} \quad (10.364)$$

Nous obtenons en utilisant la dérivée $\operatorname{arctanh}(x)$:

$$\int \frac{1}{\sinh(x)} dx = \int \frac{2}{e^x - e^{-x}} dx = \int \frac{2}{t \left(t - \frac{1}{t} \right)} dt = \int \frac{2}{t^2 - 1} dt = -2 \operatorname{arctanh}(t) = -2 \operatorname{arctanh}(e^x) \quad (10.365)$$

et:

$$F(x) = -2 \operatorname{arctanh}(e^x) \quad (10.366)$$

37. Primitive de $f(x) = \frac{1}{\cosh(x)}$:

Nous faisons la substitution:

$$t = e^x \text{ avec } dx = \frac{dt}{t} \quad (10.367)$$

Nous obtenons en utilisant la dérivée $\operatorname{arctan}(x)$:

$$\int \frac{1}{\cosh x} dx = \int \frac{2}{e^x + e^{-x}} dx = \int \frac{2}{t \left(t + \frac{1}{t} \right)} dt = \int \frac{2}{t^2 + 1} dt = 2 \operatorname{arctan}(t) = 2 \operatorname{arctan}(e^x) \quad (10.368)$$

et donc:

$$F(x) = 2 \operatorname{arctan}(e^x) \quad (10.369)$$

38. Primitive de $f(x) = \frac{1}{1 + \cosh(x)}$:

Nous faisons la substitution:

$$t = e^x \text{ avec } dx = \frac{dt}{t} \quad (10.370)$$

Nous obtenons:

$$\int \frac{1}{1 + \cosh x} dx = \int \frac{2}{2 + e^x + e^{-x}} dx = \int \frac{2}{t \left(2 + t + \frac{1}{t} \right)} dt = \int \frac{2}{(1+t)^2} dt = -\frac{2}{1+t} + c^{te} = -\frac{2}{1+e^x} + c^{te} \quad (10.371)$$

Nous obtenons donc la primitive:

$$F(x) = -\frac{2}{1+e^x} + 1 = \frac{e^x - 1}{e^x + 1} = \tanh\left(\frac{x}{2}\right) \quad (10.372)$$

39. Primitive de $f(x) = \frac{1}{1 - \cosh(x)}$:

Nous faisons la substitution:

$$t = \exp(x) \text{ avec } dx = \frac{dt}{t} \quad (10.373)$$

Nous obtenons:

$$\int \frac{1}{1 - \cosh x} dx = \int \frac{2}{2 - e^x - e^{-x}} dx = \int \frac{2}{t \left(2 - t - \frac{1}{t} \right)} dt = -\int \frac{2}{(t-1)^2} dt = \frac{2}{t-1} + c^{te} = \frac{2}{e^x - 1} + c^{te} \quad (10.374)$$

Nous obtenons donc la primitive:

$$F(x) = \frac{2}{e^x - 1} + 1 = \frac{e^x + 1}{e^x - 1} = \coth\left(\frac{x}{2}\right) \quad (10.375)$$

40. Primitive de $f(x) = \frac{1}{1 + \sinh(x)}$:

Nous faisons la substitution:

$$t = e^x \text{ avec } dx = \frac{dt}{t} \quad (10.376)$$

Nous obtenons:

$$\int \frac{1}{1 + \sinh x} dx = \int \frac{2}{2 + e^x - e^{-x}} dx = \int \frac{2}{2t + t^2 - 1} dt \quad (10.377)$$

Or:

$$\frac{2}{2t+t^2-1} = \frac{2}{(t+1-\sqrt{2})(t+1+\sqrt{2})} = \frac{1}{\sqrt{2}} \left(\frac{1}{(t+1-\sqrt{2})} - \frac{1}{(t+1+\sqrt{2})} \right) \quad (10.378)$$

D'où:

$$\begin{aligned} \int \frac{2}{2t+t^2-1} dt &= \frac{1}{\sqrt{2}} \left(\int \frac{1}{(t+1-\sqrt{2})} dt - \int \frac{1}{(t+1+\sqrt{2})} dt \right) \\ &= \frac{1}{\sqrt{2}} \left(\ln |t+1-\sqrt{2}| - \ln |t+1+\sqrt{2}| \right) = \frac{1}{\sqrt{2}} \ln \left| \frac{t+1-\sqrt{2}}{t+1+\sqrt{2}} \right| = \frac{1}{\sqrt{2}} \ln \left| \frac{e^x+1-\sqrt{2}}{e^x+1+\sqrt{2}} \right| \end{aligned} \quad (10.379)$$

Donc:

$$F(x) = \frac{1}{\sqrt{2}} \ln \left| \frac{\exp(x)+1-\sqrt{2}}{\exp(x)+1+\sqrt{2}} \right| \quad (10.380)$$

41. Primitive de $f(x) = \frac{1}{1-\sinh(x)}$:

Nous faisons la substitution habituelle:

$$t = e^x \text{ avec } dx = \frac{dt}{t} \quad (10.381)$$

Nous obtenons:

$$\int \frac{1}{1-\sinh x} dx = \int \frac{2}{2-e^x+e^{-x}} dx = - \int \frac{2}{-2t+t^2-1} dt \quad (10.382)$$

Or:

$$\frac{2}{-2t+t^2-1} = \frac{2}{(t-1-\sqrt{2})(t-1+\sqrt{2})} = \frac{1}{\sqrt{2}} \left(\frac{1}{(t-1-\sqrt{2})} - \frac{1}{(t-1+\sqrt{2})} \right) \quad (10.383)$$

D'où:

$$\begin{aligned} - \int \frac{2}{-2t+t^2-1} dt &= - \frac{1}{\sqrt{2}} \left(\int \frac{1}{(t-1-\sqrt{2})} dt - \int \frac{1}{(t-1+\sqrt{2})} dt \right) \\ &= \frac{1}{\sqrt{2}} \left(\ln |t-1+\sqrt{2}| - \ln |t-1-\sqrt{2}| \right) = \frac{1}{\sqrt{2}} \ln \left| \frac{t-1+\sqrt{2}}{t-1-\sqrt{2}} \right| = \frac{1}{\sqrt{2}} \ln \left| \frac{e^x-1+\sqrt{2}}{e^x-1-\sqrt{2}} \right| \end{aligned} \quad (10.384)$$

Donc:

$$F(x) = \frac{1}{\sqrt{2}} \ln \left| \frac{e^x - 1 + \sqrt{2}}{e^x - 1 - \sqrt{2}} \right| \quad (10.385)$$

42. Primitive de $f(x) = e^{ax} \sin(bx)$ avec $a, b \in \mathbb{R}, a^2 + b^2 \neq 0$:

Une première intégration par parties donne:

$$\int e^{ax} \sin(bx) dx = \frac{1}{a} e^{ax} \sin(bx) - \frac{b}{a} \int e^{ax} \cos(bx) dx \quad (10.386)$$

Une deuxième intégration par parties donne:

$$\int e^{ax} \sin(bx) dx = \frac{1}{a} e^{ax} \sin(bx) - \frac{b}{a} \left(\frac{1}{a} e^{ax} \cos(bx) + \frac{b}{a} \int e^{ax} \sin(bx) dx \right) \quad (10.387)$$

d'où l'égalité:

$$\int e^{ax} \sin(bx) dx = \frac{1}{a} e^{ax} \sin(bx) - \frac{b}{a^2} e^{ax} \cos(bx) - \frac{b^2}{a^2} \int e^{ax} \sin(bx) dx \quad (10.388)$$

Ainsi en redistribuant la relation précédente:

$$F(x) = \int e^{ax} \sin(bx) dx = \frac{e^{ax}}{a^2 + b^2} (a \sin(bx) - b \cos(bx)) \quad (10.389)$$

43. Primitive de $f(x) = e^{ax} \cos(bx)$ avec $a, b \in \mathbb{R}, a^2 + b^2 \neq 0$:

Un raisonnement analogue à celui d'avant montre que:

$$F(x) = \frac{e^{ax}}{a^2 + b^2} (a \cos(bx) + b \sin(bx)) \quad (10.390)$$

44. Primitive de $f(x) = x \sin(ax)$ avec $a \in \mathbb{R}^*$:

Une intégration par parties nous donne:

$$F(x) = \int x \sin(ax) dx = -\frac{1}{a} x \cos(ax) + \frac{1}{a} \int \cos(ax) dx = -\frac{1}{a} x \cos(ax) + \frac{1}{a^2} \sin(ax) \quad (10.391)$$

45. Primitive de $f(x) = x \cos(ax)$ avec $a \in \mathbb{R}^*$:

Une intégration par parties nous donne:

$$F(x) = \int x \cos(ax) dx = \frac{1}{a} x \sin(ax) - \frac{1}{a} \int \sin(ax) dx = \frac{1}{a} x \sin(ax) + \frac{1}{a^2} \cos(ax) \quad (10.392)$$

46. Primitive de $f(x) = \frac{1}{(x-a)(x-b)}$ avec $a \neq b$:

Nous avons la relation suivante:

$$\frac{1}{(x-a)(x-b)} = \frac{1}{a-b} \left(\frac{1}{(x-a)} - \frac{1}{(x-b)} \right) \quad (10.393)$$

Par suite:

$$\begin{aligned} \int \frac{1}{(x-a)(x-b)} dx &= \frac{1}{a-b} \int \left(\frac{1}{(x-a)} - \frac{1}{(x-b)} \right) dx = \frac{1}{a-b} \left(\int \frac{1}{(x-a)} dx - \int \frac{1}{(x-b)} dx \right) \\ &= \frac{1}{a-b} (\ln |x-a| - \ln |x-b|) = \frac{1}{a-b} \ln \left| \frac{x-a}{x-b} \right| \end{aligned} \quad (10.394)$$

Ainsi:

$$F(x) = \frac{1}{a-b} \ln \left| \frac{x-a}{x-b} \right| \quad (10.395)$$

47. Primitive de $f(x) = \frac{1}{a^2 - x^2}$ avec $a \neq 0$:

Nous avons en utilisant le résultat précédent:

$$\int \frac{1}{a^2 - x^2} dx = \int \frac{1}{(a+x)(a-x)} dx = - \int \frac{1}{(x+a)(x-a)} dx = - \frac{1}{2a} \ln \left| \frac{x-a}{x+a} \right| = \frac{1}{2a} \ln \left| \frac{x+a}{x-a} \right| \quad (10.396)$$

Donc:

$$F(x) = \frac{1}{2a} \ln \left| \frac{x+a}{x-a} \right| \quad (10.397)$$

48. Primitive de $f(x) = \frac{1}{a^2 + x^2}$ avec $a \neq 0$:

En faisant le changement de variable:

$$x = au \text{ avec } dx = a \cdot du \quad (10.398)$$

Nous obtenons en utilisant la dérivée de $\arctan(x)$:

$$F(x) = \int \frac{1}{a^2 + x^2} dx = \frac{1}{a} \int \frac{1}{1+u^2} du = \frac{1}{a} \arctan(u) = \frac{1}{a} \arctan \left(\frac{x}{a} \right) \quad (10.399)$$

49. Soit:

$$I_{n+1} = \int \frac{1}{(1-x^2)^{n+1}} dx \quad (10.400)$$

avec $n \in \mathbb{N}$. Nous avons:

$$\begin{aligned}
 I_{n+1} &= \int \frac{1}{(1-x^2)^{n+1}} dx = \int \left(\frac{1-x^2}{(1-x^2)^{n+1}} + \frac{x^2}{(1-x^2)^{n+1}} \right) dx = \int \frac{1}{(1-x^2)^n} dx + \int \frac{x^2}{(1-x^2)^{n+1}} dx \\
 &= I_n + \int \frac{x^2}{(1-x^2)^{n+1}} dx
 \end{aligned} \tag{10.401}$$

Or cette dernière intégrale se résout par parties:

$$\begin{aligned}
 \int \frac{x^2}{(1-x^2)^{n+1}} dx &= \int \frac{x}{(1-x^2)^{n+1}} x dx = \frac{1}{2n(1-x^2)^n} x - \int \frac{1}{2n(1-x^2)^n} dx \\
 &= \frac{1}{2n(1-x^2)^n} x - \frac{1}{2n} I_n
 \end{aligned} \tag{10.402}$$

Donc:

$$I_{n+1} = I_n + \frac{1}{2n(1-x^2)^n} x - \frac{1}{2n} I_n = \frac{1}{2n(1-x^2)^n} x + \frac{(2n-1)}{2n} I_n \tag{10.403}$$

Que nous retrouvons plus fréquemment dans la littérature sous la forme:

$$2nI_{n+1} = \frac{x}{(1-x^2)^n} + (2n-1)I_n \tag{10.404}$$

Identiquement au développement suivant, nous avons pour (le signe change):

$$I_{n+1} = \int \frac{1}{(1+x^2)^{n+1}} dx \tag{10.405}$$

la relation suivante:

$$2nI_{n+1} = \frac{x}{(1+x^2)^n} + (2n-1)I_n \tag{10.406}$$

Vous pourrez trouver une application de ces deux primitives dans le modèle cosmologique newtonien de l'univers dans le chapitre d'Astrophysique ainsi que dans le chapitre de Relativité Générale dans le cadre de l'étude de l'effet Shapiro!

50. Primitive de $f(x) = \frac{1}{(1-x^2)^2}$:

Nous avons en utilisant les primitives de $\frac{1}{(1-x^2)^n}$ (vue avant) et $\frac{1}{1-x^2}$ (vue plus haut):

$$F(x) = \int \frac{1}{(1-x^2)^2} dx = \frac{1}{2} \left(\frac{x}{1-x^2} + \frac{1}{2} \ln \left| \frac{x+1}{x-1} \right| \right) \tag{10.407}$$

51. Primitive de $f(x) = \frac{1}{(1+x^2)^2}$:

Nous avons en utilisant les primitives de $\frac{1}{(1+x^2)^n}$ (vue avant) et $\frac{1}{1+x^2}$ (vue plus haut):

$$F(x) = \int \frac{1}{(1+x^2)^2} dx = \frac{1}{2} \left(\frac{x}{1+x^2} + \arctan(x) \right) \quad (10.408)$$

52. Primitive de $f(x) = \sqrt{x^2 - a^2}$ avec $a \in \mathbb{R}^*$:

Nous pouvons sans perte de généralité supposer $a > 0$. Remarquons que le domaine de définition de f est $]-\infty, -a] \cup [a, +\infty[$.

Nous allons déterminer une primitive de f uniquement sur l'intervalle $[a, +\infty[$ (car c'est celle dont nous aurons besoin dans certains chapitres).

Faisons le changement de variable:

$$x = a \cdot \cosh t \quad (10.409)$$

avec donc:

$$t = \operatorname{arccosh}\left(\frac{1}{a}x\right), dx = a \sinh(t) dt \quad (10.410)$$

où nous considérons la fonction $\cosh : \mathbb{R}_+ \rightarrow [1, +\infty[$ avec pour réciproque la fonction $\operatorname{arccosh} : [1, +\infty[\rightarrow \mathbb{R}_+$ donnée par (cf. chapitre de Trigonométrie):

$$\operatorname{arccosh}(y) = \ln\left(y + \sqrt{y^2 - 1}\right) \quad (10.411)$$

Nous obtenons alors en utilisant la primitive de $\sinh^2 x$:

$$\begin{aligned} \int \sqrt{x^2 - a^2} dx &= a^2 \int \sinh^2 t \cdot dt = \frac{a^2}{2} (\cosh t \sinh t - t) \\ &= \frac{a^2}{2} \cdot \left(\frac{1}{a} x \sinh\left(\operatorname{arccosh}\left(\frac{1}{a}x\right)\right) - \operatorname{arccosh}\left(\frac{1}{a}x\right) \right) \end{aligned} \quad (10.412)$$

or (cf. chapitre de Trigonométrie) comme:

$$\sinh^2 u = \cosh^2 u - 1 \quad (10.413)$$

Donc:

$$\sinh\left(\operatorname{arccosh}\left(\frac{1}{a}x\right)\right) = \sqrt{\cosh^2\left(\operatorname{arccosh}\left(\frac{1}{a}x\right)\right) - 1} = \sqrt{\frac{1}{a^2}x^2 - 1} \quad (10.414)$$

et en utilisant un autre résultat du chapitre de Trigonométrie:

$$\operatorname{arccosh}(x) = \ln \left(x + \sqrt{x^2 - 1} \right) \quad (10.415)$$

nous avons alors:

$$\begin{aligned} \int \sqrt{x^2 - a^2} dx &= \frac{a^2}{2} \left(\frac{x}{a} \sqrt{\frac{x^2}{a^2} - 1} - \ln \left(\frac{x}{a} + \sqrt{\frac{x^2}{a^2} - 1} \right) \right) \\ &= \frac{a^2}{2} \left(\frac{x}{a^2} \sqrt{x^2 - a^2} - \ln \left(\frac{x}{a} + \frac{1}{a} \sqrt{x^2 - a^2} \right) \right) = \frac{a^2}{2} \left(\frac{x}{a^2} \sqrt{x^2 - a^2} - \ln \left(x + \sqrt{x^2 - a^2} \right) + \ln(a) \right) \end{aligned} \quad (10.416)$$

étant donné que les primitives sont données à une constante près, nous pouvons écrire:

$$F(x) = \frac{x}{2} \sqrt{x^2 - a^2} - \frac{a^2}{2} \ln \left(x + \sqrt{x^2 - a^2} \right) \quad (10.417)$$

pour $x \geq a$. F est donc une primitive de $\sqrt{x^2 - a^2}$ sur $[a, +\infty[$.

53. Primitive de $f(x) = \sqrt{a^2 - x^2}$ avec $a \in \mathbb{R}^*$:

Nous pouvons sans perte de généralité supposer $a > 0$. Remarquons que le domaine de définition de f est $[-a, a]$.

Nous faisons la substitution:

$$x = a \sin(t) \quad (10.418)$$

avec:

$$dx = a \cos(t) dt \quad (10.419)$$

Nous obtenons:

$$\begin{aligned} \int \sqrt{a^2 - x^2} dx &= a^2 \int \sqrt{1 - \sin^2 t} \cos(t) dt = a^2 \int \cos^2(t) dt = \frac{a^2}{2} (t + \sin(t) \cos(t)) \\ &= \frac{a^2}{2} \left[\arcsin \left(\frac{x}{a} \right) + \frac{x}{a} \cos \left(\arcsin \frac{x}{a} \right) \right] \end{aligned} \quad (10.420)$$

où nous avons utilisé la primitive de $\cos^n(x)$ avec $n = 2$ démontrée plus haut. Or nous avons:

$$\cos(\arcsin y) = \sqrt{1 - \sin^2(\arcsin y)} = \sqrt{1 - y^2} \quad (10.421)$$

Donc:

$$\int \sqrt{a^2 - x^2} dx = \frac{a^2}{2} \left(\arcsin \frac{x}{a} + \frac{x}{a} \cos(\arcsin \frac{x}{a}) \right) = \frac{a^2}{2} \left(\arcsin \frac{x}{a} + \frac{x}{a} \sqrt{1 - \frac{x^2}{a^2}} \right) = \frac{a^2}{2} \arcsin \frac{x}{a} + \frac{x}{2} \sqrt{a^2 - x^2} \quad (10.422)$$

et:

$$F(x) = \frac{a^2}{2} \arcsin \frac{x}{a} + \frac{x}{2} \sqrt{a^2 - x^2} \quad (10.423)$$

54. Primitive de $f(x) = \sqrt{a^2 + x^2}$ avec $a \in \mathbb{R}^*$:

Nous pouvons sans perte de généralité supposer $a > 0$.

Faisons le changement de variable:

$$x = a \cdot \sinh t \quad (10.424)$$

avec donc:

$$t = \operatorname{arcsinh} \left(\frac{1}{a} x \right), dx = a \cosh(t) dt \quad (10.425)$$

Nous obtenons:

$$\int \sqrt{x^2 + a^2} dx = a^2 \int \cosh^2 t dt = \frac{a^2}{2} (t + \sinh t \cosh t) \quad (10.426)$$

en ayant utilisé la primitive de $\cosh^2$ démontrée plus haut.

Ainsi:

$$\int \sqrt{x^2 + a^2} dx = \frac{a^2}{2} (t + \sinh t \cosh t) = \frac{a^2}{2} \left(\operatorname{arcsinh} \left(\frac{x}{a} \right) + \frac{x}{a} \cosh(\operatorname{arcsinh} \left(\frac{x}{a} \right)) \right) \quad (10.427)$$

Mais comme nous avons vu dans le chapitre de Trigonométrie:

$$\operatorname{arcsinh} \left(\frac{x}{a} \right) = \ln \left(\frac{x}{a} + \frac{1}{a} \sqrt{x^2 + a^2} \right) = \ln \left(x + \sqrt{x^2 + a^2} \right) - \ln(a) \quad (10.428)$$

et:

$$\cosh(\operatorname{arcsinh} \left(\frac{x}{a} \right)) = \sqrt{1 + \sinh^2 \left(\operatorname{arcsinh} \left(\frac{x}{a} \right) \right)} = \sqrt{1 + \frac{x^2}{a^2}} = \frac{1}{a} \sqrt{a^2 + x^2} \quad (10.429)$$

Donc nous avons finalement:

$$F(x) = \frac{a^2}{2} \ln \left(x + \sqrt{x^2 + a^2} \right) + \frac{x}{2} \sqrt{a^2 + x^2} \quad (10.430)$$

où le $\ln(a)$ a encore une fois été omis car les primitives sont données à une constante près.

55. Primitive de $f(x) = \left(\sqrt{a^2 - x^2}\right)^{-1}$ avec $a \in \mathbb{R}^*$:

Nous pouvons sans perte de généralité supposer $a > 0$.

Nous faisons la substitution:

$$x = a \sin(t) \quad (10.431)$$

avec:

$$dx = a \cos(t) dt \quad (10.432)$$

Nous obtenons:

$$\begin{aligned} F(x) &= \int \frac{1}{\sqrt{a^2 - x^2}} dx = \int \frac{1}{\sqrt{a^2 - a^2 \sin^2 t}} a \cos t \cdot dt = \int \frac{1}{a \sqrt{1 - \sin^2 t}} a \cos t \cdot dt \\ &= \int \frac{1}{a \cos t} a \cos t \cdot dt = \int 1 \cdot dt = t = \arcsin \frac{x}{a} \end{aligned} \quad (10.433)$$

56. Primitive de $f(x) = \frac{1}{\sqrt{a^2 + x^2}}$ avec $a \in \mathbb{R}^*$:

Nous pouvons sans perte de généralité supposer $a > 0$.

Faisons le changement de variable:

$$x = a \cdot \sinh t \quad (10.434)$$

avec:

$$t = \operatorname{arcsinh}\left(\frac{1}{a} x\right), dx = a \cosh(t) dt \quad (10.435)$$

Nous obtenons de la même manière que pour les intégrales usuelles précédentes:

$$\int \frac{1}{\sqrt{x^2 + a^2}} dx = \int 1 \cdot dt = t = \operatorname{arcsinh} \frac{x}{a} \quad (10.436)$$

et sachant que (cf. chapitre de Trigonométrie):

$$\operatorname{arcsinh}(y) = \ln\left(y + \sqrt{y^2 + 1}\right) \quad (10.437)$$

Nous obtenons alors au final la primitive importante suivante:

$$F(x) = \int \frac{1}{\sqrt{x^2 + a^2}} dx = \ln \left(x + \sqrt{x^2 + a^2} \right) \quad (10.438)$$

où le $\ln(a)$ a encore une fois été omis car les primitives sont données à une constante près!

En procédant de même, mais en utilisant le cosinus hyperbolique au lieu du sinus hyperbolique, nous avons bien évidemment:

$$F(x) = \int \frac{1}{\sqrt{x^2 - a^2}} dx = \ln \left(x + \sqrt{x^2 - a^2} \right) \quad (10.439)$$

Nous réutiliserons ces deux dernières relations dans des cas pratiques importants des chapitres de Mécanique Analytique, Génie Civil (où la constante a valant 1, $\ln(a)$ est de toute façon nul!) et de Relativité Générale (où a sera non nul et donc il ne sera pas possible d'omettre la constante $\ln(a)$).

3. FONCTION DE DIRAC

La fonction de Dirac, appelée aussi "**pic de Dirac**" ou encore "**fonction delta**", joue un rôle pratique très important aussi bien en électronique et informatique qu'en physique quantique ondulatoire et physique quantique des champs (cela permet de discrétiser un continuum) ainsi que dans le domaine du génie civil (voir les chapitres du même nom pour des exemples).

Signalons avant d'aller qu'il est abusif de parler de "fonction" car une fonction est une application d'un ensemble de départ (généralement l'ensemble des réels ou complexes à une ou plusieurs dimensions) dans un ensemble d'arrivée (généralement l'ensemble des réels ou complexes à une ou plusieurs dimensions). Alors que le domaine de définition de la fonction de Dirac n'est pas un ensemble de nombre mais en toute rigueur des fonctions!

Pour la présenter simplement, considérons d'abord la fonction définie par:

$$\begin{aligned} f(x) &= \frac{1}{a} \text{ si } -\frac{a}{2} < x < \frac{a}{2} \\ f(x) &= 0 \text{ si } |x| > \frac{a}{2} \end{aligned} \quad (10.440)$$

La représentation de $y = f(x)$ est un rectangle de largeur a , de hauteur $1/a$ et de surface unité. La fonction de Dirac peut être considérée comme la limite, lorsque $a \rightarrow 0$ de la fonction $f(x)$. On a donc:

$$\begin{aligned} \delta(x) &= 0 \text{ si } x \neq 0 \\ \delta(x) &= \infty \text{ si } x = 0 \end{aligned} \quad (10.441)$$

avec:

$$\int_{-\infty}^{+\infty} \delta(x) dx = \int_{-\varepsilon}^{+\varepsilon} \delta(x) dx = 1 \quad (10.442)$$

où ε est un nombre plus grand que 0 aussi petit que nous le voulons.

Remarque : Comme le lecteur l'aura remarqué lorsque nous avons introduit la fonction $f(x)$ initiale, la fonction delta de Dirac qui en découle a donc la dimension de l'inverse d'une longueur.

Pour une fonction $g(x)$ continue en $x=0$ on a:

$$\int_{-\infty}^{+\infty} g(x) \delta(x) dx = g(0) \quad (10.443)$$

Par extension nous avons:

$$\begin{aligned} \delta(x - x_0) &= 0 \text{ si } x \neq x_0 \\ \delta(x - x_0) &= \infty \text{ si } x = x_0 \end{aligned} \quad (10.444)$$

et pour une fonction $g(x)$ continue en x_0 :

$$\int_{-\infty}^{+\infty} g(x) \delta(x - x_0) dx = g(x_0) \quad (10.445)$$

Il est alors assez aisé de définir la fonction de Dirac dans l'espace à 3 dimensions par:

$$\delta(\vec{r} - \vec{r}_0) = \delta(x - x_0) \delta(y - y_0) \delta(z - z_0) \quad (10.446)$$

4. FONCTION GAMMA D'EULER

Nous définissons la fonction Gamma d'Euler (intégrale Eulérienne de deuxième espèce) par l'intégrale suivante:

$$\Gamma_{-1}(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt \quad (10.447)$$

avec x appartenant à l'ensemble des nombres complexes dont la partie réelle est positive et non nulle (donc les réels strictement positifs sont inclus dans le domaine de définition aussi...)! Effectivement, si nous prenons des complexes avec une partie réelle nulle ou négative, l'intégrale diverge et est alors non définie!

Remarque: Nous avons déjà rencontré cette intégrale et certaines de ses propriétés (qui vont être démontrées ici) lors de notre étude des fonctions de distribution Bêta, Gamma, Khi-deux, Student et Fisher en statistiques (cf. chapitre de Statistiques). Nous utiliserons également cette intégrale en maintenance (cf. chapitre de Techniques De Gestion), en théorie des cordes (cf. chapitre de Théorie Des Cordes) et dans d'autres domaines de l'ingénierie (voir la section correspondante).

Voici un tracé graphique du module de la fonction Gamma d'Euler pour x parcourant un intervalle des nombres réels (attention dans Maple 4.00b à bien écrire GAMMA en majuscules!!!):

```
>with(plots):
> plot(GAMMA(x),x=-Pi..Pi,y=-5..5);
```

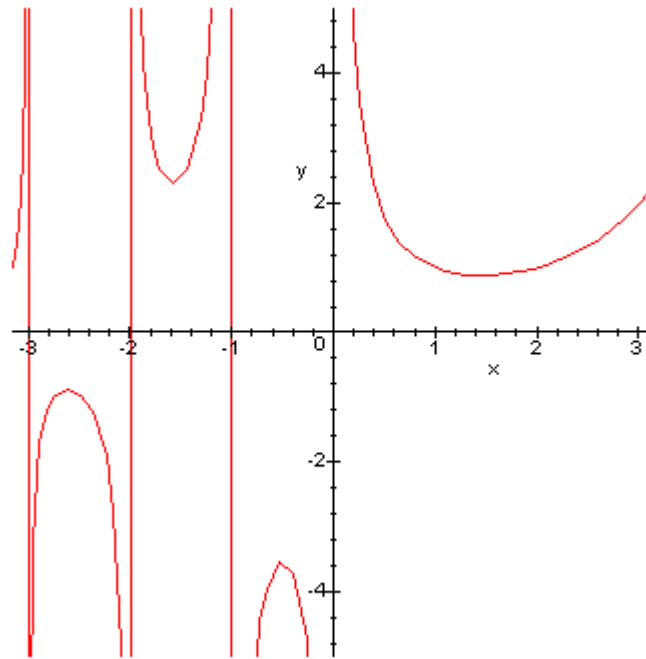


Figure: 10.18 - Plot de la fonction Gamma d'Euler dans Maple 4.00b

et la même fonction tracée avec Maple 4.00b mais dans le plan complexe cette fois-ci et toujours avec en ordonnée le module de la fonction Gamma d'Euler:

```
>with(plots):
>plot3d(abs(GAMMA(x+y*I)),x=-Pi..Pi,y=-Pi..Pi,view=0..5, grid=[30,30],orientation=[-120,45]
,axes=frame,style=patchcontour);
```

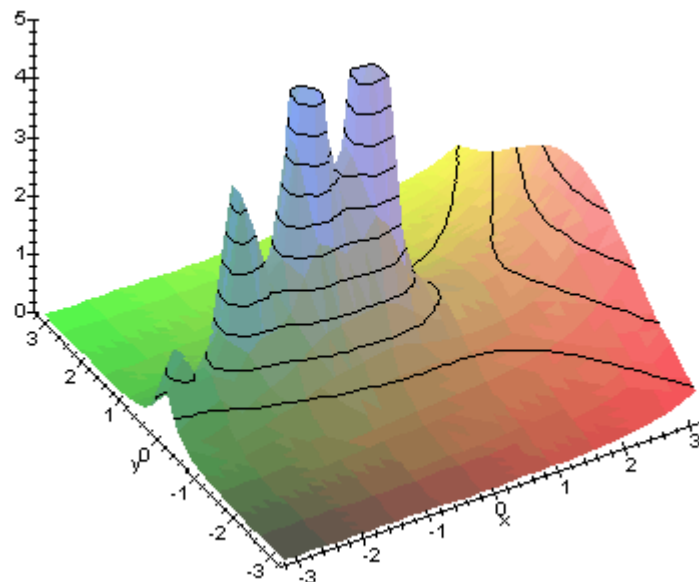


Figure: 10.19 - Plot de la fonction Gamma d'Euler dans le plan complexe avec Maple 4.00

Cette fonction est intéressante si nous imposons que la variable x appartienne aux entiers positifs et que nous l'écrivons sous la forme suivante:

$$\Gamma_0(x) = \int_0^{+\infty} t^x e^{-t} dt \quad (10.448)$$

Intégrons par partie cette dernière fonction:

$$\Gamma_0(x) = \int_0^{+\infty} t^x e^{-t} dt = -t^x e^{-t} \Big|_0^{+\infty} + x \int_0^{+\infty} t^{x-1} e^{-t} dt = -t^x e^{-t} \Big|_0^{+\infty} + x \cdot \Gamma_0(x-1) \quad (10.449)$$

Comme la fonction exponentielle décroît beaucoup plus vite que t^x nous avons alors:

$$\Gamma_0(x) = x \cdot \Gamma_0(x-1) \quad (10.450)$$

Dans la littérature, nous retrouvons fréquemment les notations suivantes (qui portent alors à confusion):

$$\begin{aligned} \Gamma_{-1}(x) &= \Gamma_0(x-1) = \Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt \\ \Gamma_0(x) &= \Gamma(x+1) = \int_0^{+\infty} t^x e^{-t} dt \end{aligned} \quad (10.451)$$

Ce qui nous amène à réécrire le résultat sous une forme plus classique:

$$\boxed{\Gamma(x+1) = x\Gamma(x)} \quad (10.452)$$

De la relation $\Gamma_0(x) = x \cdot \Gamma_0(x-1)$, il vient par récurrence:

$$\Gamma_0(x) = x(x-1)\Gamma_0(x-2) = \dots = \Gamma_0(0)x! \quad (10.453)$$

Or:

$$\Gamma_0(0) = \Gamma(1) = \int_0^{+\infty} t^0 e^{-t} dt = \int_0^{+\infty} e^{-t} dt = -e^{-t} \Big|_0^{+\infty} = -e^{-\infty} - (-e^{-0}) = 1 \quad (10.454)$$

ce qui donne:

$$\Gamma_0(x) = x \cdot (x-1) \cdot \Gamma_0(x-2) = \dots = x! \quad (10.455)$$

Donc:

$$\Gamma_0(x) = x! \quad (10.456)$$

ou autrement écrit pour $x > 0, x \in \mathbb{N}$:

$$\boxed{\Gamma(x+1) = x!} \quad (10.457)$$

Un autre résultat intéressant de la fonction gamma d'Euler est obtenu lorsque nous remplaçons t par y^2 et calculons celle-ci pour $x = 1/2$.

D'abord, nous avons:

$$\frac{dt}{dy} = \frac{dy^2}{dy} = 2y \Rightarrow dt = 2y dy \quad (10.458)$$

ensuite:

$$\Gamma_{-1}\left(\frac{1}{2}\right) = \int_0^{+\infty} t^{\frac{1}{2}-1} e^{-t} dt = \int_0^{+\infty} y^{-1} e^{-y^2} 2y dy = 2 \int_0^{+\infty} e^{-y^2} dy \quad (10.459)$$

Or, comme nous l'avons démontré dans le chapitre de Statistiques lors de notre étude de loi de Gauss-Laplace, cette dernière intégrale vaut:

$$\Gamma_{-1}\left(\frac{1}{2}\right) = 2 \int_0^{+\infty} e^{-y^2} dy = \sqrt{\pi} \quad (10.460)$$

4.1. CONSTANTE D'EULER-MASCHERONI

Ce petit texte fait juste office de curiosité relativement à la constante d'Euler e et à presque tous les outils de calcul différentiel et intégral que nous avons vus jusqu'à maintenant. C'est un très joli exemple (presque artistique) de ce que nous pouvons faire avec la mathématique dès que nous avons suffisamment d'outils à notre disposition.

De plus, cette constante est utile dans certaines équations différentielles où nous la retrouverons.

Nous avons vu dans le chapitre d'analyse fonctionnelle que la constante d'Euler e est définie par la limite:

$$e = \lim_{\alpha \rightarrow +\infty} \left(1 + \frac{1}{\alpha}\right)^{\alpha} \quad (10.461)$$

Dans un cas plus général nous pouvons très facilement démontrer de la même façon que:

$$\lim_{n \rightarrow +\infty} \left(1 - \frac{t}{n}\right)^n = e^{-t} \quad (10.462)$$

Cela suggère évidemment:

$$\Gamma_{-1}(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt = \lim_{n \rightarrow +\infty} \int_0^n t^{x-1} \left(1 - \frac{t}{n}\right)^n dt \quad (10.463)$$

par changement $t = n \cdot u$ de variable nous écrivons:

$$\begin{aligned} \Gamma_{-1}(x) &= \lim_{n \rightarrow +\infty} \int_0^1 (n \cdot u)^{x-1} (1-u)^n n \cdot du = \lim_{n \rightarrow +\infty} n^x \int_0^1 u^{x-1} (1-u)^n du \\ B(x-1, n) &= \int_0^1 u^{x-1} (1-u)^n du = \frac{u^x}{x} (1-u)^n \Big|_0^1 + \frac{1}{x} n \int_0^1 u^x (1-u)^{n-1} du \\ &= \frac{n}{x} \cdot B(x, n-1) = \frac{n(n-1)}{x(x+1)} \cdot B(x+1, n-2) = \frac{n!}{x(x+1)\dots(x+n)} \end{aligned} \quad (10.464)$$

$$\Gamma_{-1}(x) = \lim_{n \rightarrow +\infty} \frac{n! n^x}{x(x+1)\dots(x+n)}$$

Pour transformer cette expression nous pouvons écrire:

$$n^x = e^{\ln(n^x)} = e^{x \cdot \ln(n)} = e^{x(\ln(n) - 1 + 1 - 1/2 + 1/2 - \dots - 1/n + 1/n)} = e^{-x(1 + 1/2 + \dots + 1/n - \ln(n))} = e^{-x} \cdot e^{-x/2} \cdot \dots \cdot e^{-x/n} \quad (10.465)$$

Or la quantité:

$$1 + \frac{1}{2} + \dots + \frac{1}{n} - \ln(n) \quad (10.466)$$

tend vers la limite $\gamma \cong 0.5772\dots$, appelée "**constante d'Euler-Mascheroni**" ou également "**constante Gamma d'Euler**", lorsque n tend vers l'infini.

D'où:

$$\Gamma_{-1}(x) = \lim_{n \rightarrow +\infty} \frac{n! e^{-\gamma x} \cdot e^{-x} \cdot e^{-x/2} \cdot \dots \cdot e^{-x/n}}{x(x+1)\dots(x+n)} = \lim_{n \rightarrow +\infty} \frac{n! e^{-\gamma x} \cdot \prod_{k=1}^n e^{-x/k}}{x(x+1)\dots(x+n)} \quad (10.467)$$

Divisons chacun des termes du produit $(x+1)\dots(x+n)$ par l'entier correspondant pris dans $n!$, nous obtenons donc:

$$\begin{aligned} \frac{1}{\Gamma_{-1}(x)} &= x \cdot e^{\gamma x} \lim_{n \rightarrow +\infty} \frac{(x+1)\dots(x+n)}{n!} \prod_{k=1}^n e^{-x/k} = x \cdot e^{\gamma x} \lim_{n \rightarrow +\infty} \prod_{k=1}^n \left(1 + \frac{x}{k}\right) e^{-x/k} \\ &= x \cdot e^{\gamma x} \prod_{k=1}^{+\infty} \left(1 + \frac{x}{k}\right) e^{-x/k} \end{aligned} \quad (10.468)$$

5. INTÉGRALES CURVILIGNES

Les intégrales curvilignes sont aussi très importantes en physique. Le lecteur les retrouvera ainsi dans le chapitre de Mécanique Classique, Magnétostatique et Électrodynamique pour calculer le travail d'une force ou encore la "circulation d'un champ", ou encore dans le chapitre de Géométrie Euclidienne pour le calcul du centre de gravité de courbes (fonctions) pesantes, ou encore dans le chapitre Formes Géométriques pour le calculer la surface de certains corps de révolution mais aussi en Physique Quantique Corpusculaire pour la fameuse "intégrale de chemin" (qui n'est d'autre que le terme utilisé par les physiciens pour dire "intégrale curviligne") ou encore pour le calcul d'intégrales particulières utilisant le théorème des résidus démontré dans le chapitre d'Analyse Complexe ou encore pour de nombreuses transformations d'état dans le chapitre de Thermodynamique. Raison pour laquelle il n'y aura pas ici d'exemple d'application tellement ils sont nombreux dans les autres chapitres.

Avec la définition des ce intégrales, nous pourrons démontrer deux résultats très importants détaillés dans le chapitre de Calcul Vectoriel et qui sont respectivement le théorème de Green, le théorème de Stokes ou encore le théorème des résidus démontré dans le chapitre d'Analyse Complexe et déjà mentionné dans le paragraphe précédent (c'est suffisamment important pour le mentionner deux fois!).

5.1. INTÉGRALE CURVILIGNE D'UN CHAMP SCALAIRE

Considérons une courbe C paramétrée (cf. chapitre de Géométrie Différentielle) par une fonction vectorielle $\vec{r}(t)$ avec $t \in [a, b]$ de classe C^1 par morceaux (cette condition est nécessaire pour que l'on puisse intégrer sur la courbe sans problèmes).

Définitions:

D1. La courbe est dite "courbe fermée" si $\vec{r}(a) = \vec{r}(b)$

D2. La courbe est dite "courbe régulière" si $\forall t \in [a, b] \quad \vec{r}'(t) \neq 0$

Rappelons qu'une courbe paramétrée peut être écrite sous la forme suivante (toute fonction vectorielle peut être écrite sous cette forme):

$$\vec{r}(t) = x(t)\vec{i} + y(t)\vec{j} = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} \quad (10.469)$$

Considérons une fonction ou un "**champ scalaire**" $f(x, y)$ définie dans un voisinage de C . Subdivisons $[a, b]$ en n sous-intervalles Δt de même longueur tel que:

$$t_0 = a < t_1 < \dots < t_n = b \quad (10.470)$$

Nous choisissons sur chaque sous-intervalle un point $t_i^* \in [t_i, t_{i+1}]$. Soit Δs_i la longueur de l'arc de C reliant les points $(x(t_i), y(t_i))$ et $(x(t_{i+1}), y(t_{i+1}))$, l'intégrale de f de long de C est définie comme étant "**l'intégrale curviligne**" ou "**intégrale de chemin**":

$$\int_C f(x, y) ds = \lim_{n \rightarrow +\infty} \sum_{i=0}^n f(x(t_i^*), y(t_i^*)) \Delta s_i \quad (10.471)$$

Ce qui comme nous le savons, peut s'écrire (cf. chapitre de Géométrie Différentielle ou Formes Géométriques ou encore Mécanique Analytique):

$$\int_C f(x, y) ds = \int_a^b f(x(t), y(t)) \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} dt \quad (10.472)$$

et qui peut évidemment immédiatement être étendu au cas à 3 variables et plus.

Soit sous forme vectorielle:

$$\int_C f(x, y) ds = \int_a^b f(\vec{r}(t)) \left\| \frac{d\vec{r}}{dt} \right\| dt \quad (10.473)$$

L'intégrale curviligne est linéaire, c'est-à-dire que si $C = C_1 \cup C_2$ et que $C_1 \cap C_2$ est un point, alors (sans aller dans la définition rigoureuse de ce qu'est l'union de deux courbes...):

$$\int_C f(x,y)ds = \int_{C_1} f(x,y)ds + \int_{C_2} f(x,y)ds \quad (10.474)$$

5.2. INTÉGRALE CURVILIGNE D'UN CHAMP VECTORIEL

Considérons un champ de vecteur (par exemple un champ de force) tel que:

$$\vec{F} = \begin{pmatrix} F_x(x,y) \\ F_y(x,y) \end{pmatrix} = F_x(x,y)\vec{i} + F_y(x,y)\vec{j} \quad (10.475)$$

et une élément infinitésimal d'une courbe (chemin) C^1 morceaux tel:

$$d\vec{r} = \begin{pmatrix} dx \\ dy \end{pmatrix} = dx\vec{i} + dy\vec{j} \quad (10.476)$$

L'idée est alors de considérer que le produit scalaire (projection de champ vectoriel sur l'élément de chemin) représente le travail le long de l'élément différentiel:

$$\vec{F} \circ d\vec{r} \quad (10.477)$$

Par conséquent le travail sur tout le chemin sera donné par (en utilisant au passage la propriété de linéarité de l'intégrale):

$$W = \int_C \vec{F} \circ d\vec{r} = \int_C (F_x(x,y)dx + F_y(x,y)dy) = \int_C F_x(x,y)dx + \int_C F_y(x,y)dy \quad (10.478)$$

ce qui peut évidemment se généraliser à n dimensions. Indiquons que lorsque l'intégrale curviligne (de chemin) d'un champ vectoriel est étendue à une courbe fermée, nous parlons alors de "**circulation du champ vectoriel**".

Comme:

$$\frac{d\vec{r}(t)}{dt} = \vec{r}'(t) \quad (10.479)$$

Nous avons alors une écriture assez courante:

$$\int_C \vec{F} \circ d\vec{r} = \int_C \vec{F}(\vec{r}(t)) \circ \vec{r}'(t)dt = \int_C (F_x(x(t),y(t))r'_x(t) + F_y(x(t),y(t))r'_y(t))dt \quad (10.480)$$

En physique souvent les problèmes sont dans le plan et nécessitent le passage aux coordonnées polaires, ce qui en outre facilite les calculs.

Exemple:

Calculons le travail de la force de pesanteur déplaçant une masse M du point $M_1(a_1, b_1, c_1)$ au point $M_2(a_2, b_2, c_2)$ le long d'un chemin arbitraire C . Les projections de la force de pesanteur sur les axes de coordonnées sont:

$$F_x = 0, F_y = 0, F_z = -mg \quad (10.481)$$

Le travail accompli est alors:

$$W = \int_C \vec{F} \circ d\vec{r} = \int_C F_x(x, y)dx + F_y(x, y)dy + F_z(x, y)dz = - \int_C mgdz = -mg\Delta z \quad (10.482)$$

et nous retrouvons un résultat connu du chapitre de Mécanique Classique.

Une intégrale curviligne d'un champ vectoriel $\vec{F}$ le long d'une courbe C_1 est indépendante du chemin d'intégration si:

$$\int_{C_1} \vec{F} \circ d\vec{r} = \int_{C_2} \vec{F} \circ d\vec{r} \quad (10.483)$$

pour toute courbe C_2 ayant les mêmes points de départ et d'arrivée. De plus, si le champ de vecteurs satisfait (où G en physique est typiquement un potentiel):

$$\left. \begin{aligned} F_x(x, y) &= \frac{\partial G}{\partial x} \\ F_y(x, y) &= \frac{\partial G}{\partial y} \end{aligned} \right\} \Rightarrow \vec{F} = \overrightarrow{\text{grad}}(G) = \vec{\nabla}(G) \quad (10.484)$$

tel que (le lecteur y reconnaîtra une forme différentielle totale exacte):

$$\vec{F} \circ d\vec{r} = F_x(x, y)dx + F_y(x, y)dy = \frac{\partial G}{\partial x}dx + \frac{\partial G}{\partial y}dy = dG \quad (10.485)$$

Alors l'intégrale de chemin sur une courbe arbitraire dépend uniquement de la différence des valeurs de la fonction G aux deux extrémités!

Démonstration:

Si la forme différentielle du champ de vecteur satisfait bien une différentielle totale exacte, nous avons:

$$\begin{aligned}
 W_{A,B} &= \int_A^B \left(\frac{\partial G}{\partial x} dx + \frac{\partial G}{\partial y} dy \right) = \int_A^B \left(\frac{\partial G}{\partial x} \frac{dx}{dt} + \frac{\partial G}{\partial y} \frac{dy}{dt} \right) dt = \int_{t_1}^{t_2} \frac{d}{dt} G(x(t), y(t)) dt = G(t) \Big|_{t_1}^{t_2} \\
 &= G(x(t_2), y(t_2)) - G(x(t_1), y(t_1)) = G(B) - G(A)
 \end{aligned} \quad (10.486)$$

□C.Q.F.D.

Donc l'intégrale curviligne d'une différentielle totale exacte ne dépend pas du chemin d'intégration mais seulement des extrémités. Nous en déduisons également que si $\vec{F}$ dérive donc d'un potentiel scalaire et que $A = B$, l'intégrale curviligne est alors nulle.

En physique ce résultat s'interprète en disant que le travail fourni par une force $\vec{F}$ dérivant d'un potentiel scalaire s'exerçant sur une particule élémentaire lors d'un déplacement fini ne dépend pas du chemin suivi.

Définitions:

D1. Lorsque la courbe C est fermée et que l'intégrale de chemin a un résultat indépendant du sens dans lequel ce chemin est parcourue, nous utilisons la notation (la lettre sous l'intégrale pouvant évidemment varier...):

$$\oint_C \quad (10.487)$$

Si cette intégrale fermée est toujours nulle, nous disons que le champ vectoriel intégré est un "**champ conservatif**" et "**dérive d'un potentiel scalaire**" (et donc satisfait le théorème de Schwarz pour pouvoir être écrit sous forme de différentielle totale exacte) puisque ceci découle de la démonstration donnée déjà juste plus haut.

D2. Lorsque la valeur de l'intégrale de chemin fermée dépend du sens de parcours, nous utilisons la notation suivante (la lettre sous l'intégrale pouvant évidemment varier...):

$$\oint_C \quad \text{ou} \quad \oint_C \quad (10.488)$$

Ainsi, si le sens est direct (c'est-à-dire "anti-horlogique" ou encore "trigonométrique") comme la notation de gauche, son signe sera positif; si au contraire le sens est horlogique son signe sera négatif (voir la démonstration dans le chapitre d'Analyse Complexe). Nous parlons alors respectivement souvent de "sens négatif" ou "sens positif".

Ainsi, pour résumer, une intégrale curviligne (de chemin) est entièrement définie par l'expression sous le signe de l'intégrale, la forme de la courbe d'intégration et le sens d'intégration.

Remarque: Le lecteur pourra trouver des démonstrations de propriétés très importantes des intégrales curvilignes dans le chapitre de Calcul Vectoriel comme le théorème de Green-Riemann ou encore une étude d'application particulière aux fonctions holomorphes dans le chapitre d'Analyse Complexe.

6. ÉQUATIONS DIFFÉRENTIELLES

Définition: En mathématiques, une "**équation différentielle**" (E.D.) est une relation entre une ou plusieurs fonctions inconnues et leurs dérivées jusqu'à l'ordre n . "**L'ordre**" d'une équation différentielle correspond au degré maximal de différentiation auquel une des fonctions inconnues y a été soumise.

Par rapport à notre objectif d'essayer de voir comment la mathématiques décrit la réalité sensible, les équations différentielles remportent un franc succès, mais sont également la source de bien des soucis. D'abord des difficultés de modélisation (voir par exemple le système d'équation différentielles de la relativité générale...), des difficultés de résolution (il n'existe pas de méthode générale!), puis des difficultés proprement mathématiques, enfin des difficultés liées au fait que certaines équations différentielles ne sont pas stables par nature et donnent des solutions chaotiques (voir le chapitre de dynamique des populations pour des exemples simples flagrants!).

Remarque: Les équations différentielles sont utilisées pour construire des modèles mathématiques de phénomènes physiques et biologiques, par exemple pour l'étude de la radioactivité ou la mécanique céleste. Par conséquent, les équations différentielles représentent un immense champ d'étude, aussi bien en mathématiques pures qu'appliquées.

L'équation différentielle d'ordre n la plus générale peut toujours s'écrire sous la forme:

$$F(x, y, y', \dots, y^{(n)}) = 0 \quad (10.489)$$

Nous ne considérons sur ce site que le cas où x et y sont à valeur dans $\mathbb{R}$. Une solution à une telle E.D. sur l'intervalle $I \subset \mathbb{R}$ est une fonction $y \in C^n(I, \mathbb{R})$ (une fonction $y: I \rightarrow \mathbb{R}$ qui est n fois continûment dérivable) telle que pour tout $x \in I$, nous ayons:

$$F(x, y(x), y'(x), \dots, y^{(n)}(x)) = 0 \quad (10.490)$$

Remarques:

R1. Pour des raisons qui seront développées par la suite, nous disons aussi "intégrer l'E.D." au lieu de "trouver une solution à l'E.D.". La première expression se retrouve particulièrement dans la littérature anglo-saxonne.

R2. Étant donné que tout le site Internet est bourré d'exemples d'équations différentielles avec conditions initiales (on parle alors de "**problème de Cauchy**") et de méthodes de résolutions dans les chapitres sur la mécanique, la physique atomique, la cosmologie, l'économétrie, les suites et séries, etc., nous ne ferons pas d'exemples ici et ne nous intéresserons donc qu'à l'aspect théorique minimal.

6.1. ÉQUATIONS DIFFÉRENTIELLES DU 1ER ORDRE

Une équation différentielle du 1er ordre est donc une E.D. qui ne fait intervenir que la première dérivée y' .

Définition: Une équation différentielle du 1er ordre est dite "**E.D. d'ordre 1 à variables séparées**" si elle peut s'écrire sous la forme:

$$f(y)y' = g(x) \quad (10.491)$$

Une telle équation différentielle peut s'intégrer facilement. En effet, nous écrivons:

$$y' = \frac{dy}{dx} \quad (10.492)$$

Puis symboliquement:

$$f(y)dy = g(x)dx \Leftrightarrow \int f(y)dy = \int g(x)dx + c^{te} \quad (10.493)$$

Remarque: Nous écrivons ici explicitement la constante d'intégration arbitraire $c^{te} \in \mathbb{R}$ (qui est implicitement présente dans les intégrales indéfinies) pour ne pas l'oublier!

Il s'agit donc d'abord de trouver des primitives F et G de f et de g , et ensuite d'exprimer y en terme de x (et de C):

$$F(y) = G(x) + c^{te} \Leftrightarrow y = F^{-1}(G(x) + c^{te}) \quad (10.494)$$

La constante d'intégration est fixée lorsqu'on demande que pour un $x = x_0$ donné, nous ayons une valeur donnée de $y(x) = y(x_0) = y_0$. Nous parlons alors de "problème aux valeurs initiales".

6.2. ÉQUATIONS DIFFÉRENTIELLES LINÉAIRES

Définition: Une équation différentielle d'ordre n est dite "**E.D. linéaire**" (E.D.L.) si et seulement si elle est de la forme:

$$L(y) = f(x) \quad (10.495)$$

Avec:

$$L(y) = a_0(x)y + a_1(x)y' + a_2(x)y'' + \dots + a_n(x)y^{(n)} \quad (10.496)$$

Voyons maintenant une propriété qui peut sembler négligeable au premier coup d'oeil mais qui va prendre de l'importance plus loin!

Nous allons montrer que L est une application linéaire:

$$L(y+z) = \sum_{i=0}^n a_i(x)(y+z)^{(i)} = \sum_{i=0}^n a_i(x)y^{(i)} + \sum_{i=0}^n a_i(x)z^{(i)} = L(y) + L(z) \quad (10.497)$$

Et pour tout $\lambda \in \mathbb{R}$:

$$L(\lambda y) = \sum_{i=0}^n a_i(x)(\lambda y)^{(i)} = \lambda \sum_{i=0}^n a_i(x)y^{(i)} = \lambda L(y) \quad (10.498)$$

Nous disons alors que l'E.D. linéaire représente un modèle linéaire si les multiples de cette fonction (ou toute combinaison linéaire) sont aussi solution. Ainsi, en physique, pour un système linéaire, l'amplification de la cause implique une amplification de l'effet (les systèmes sont souvent linéaires dans le cas scolaires mais dans la réalité ils sont plutôt l'exception!).

Définition: L'équation différentielle (c'est la plus courante en physique):

$$L(y) = 0 \quad (10.499)$$

s'appelle "**équation homogène**" (E.H.) ou "**équation sans second membre**" (ESSM) associée à:

$$L(y) = f(x) \quad (10.500)$$

Nous allons maintenant démontrer une propriété importante des E.H.: l'ensemble $\{S_0\}$ des solutions de E.H. est le noyau de l'application linéaire L (ce qui rappelons-le signifie: $L(S_0) = 0$) et l'ensemble $\{S\}$ des solutions à $L(y) = f(x)$ est donné par:

$$S = y_p + S_0 = \{y_p + y_h \mid y_h \in \{S_0\}\} \text{ avec } L(y_p) = f(x) \quad (10.501)$$

c'est-à-dire que les solutions de la forme:

$$y = y_p + y_h \quad (10.502)$$

où y_p est une "**solution particulière**" de $L(y) = f(x)$ et y_h la "**solution homogène**", parcourent toutes les solutions de l'E.D.

Démonstration:

La première affirmation sera supposée évidente.

En ce qui concerne la 2ème partie, toute fonction de la forme $y_p + y_h$ est solution de $L(y) = f(x)$.

En effet c'est trivial et cela découle de la définition du concept de noyau (cf. chapitre de Théorie Des Ensembles):

$$L(y_p + y_h) = L(y_p) + L(y_h) = f(x) + 0 = f(x) \quad (10.503)$$

□C.Q.F.D.

Ce qu'il est important aussi de comprendre avec les E.D. linéaires avec second membre, c'est que si nous trouvons des solutions à $L(y)$ avec un second membre donné et des solutions à la même E.D. avec un autre second membre (différent!), alors la somme de toutes ces solutions, sera solution de l'E.D. avec la somme des seconds membres!!!

Il existe de nombreuses manières de résoudre les équations différentielles linéaires ou non linéaires de manière exacte ou approchée. Citons les quelques méthodes que nous analyserons plus loin par l'exemple (mais qui se trouvent déjà de très très nombreuses fois dans les chapitres de physique):

- La "**méthode du polynôme caractéristique des E.D.**" (voir plus bas) utilisée un peu dans tous les chapitre de Physique/Chimie du présent site.

- La "**méthode du facteur intégrant**" (voir plus bas) pour la culture générale mais utilisée à ce jour sur aucun cas pratique du site.
- La "**méthode de variation de la constante**" (voir plus bas) et utilisée à ce jour uniquement dans le chapitre de Génie Industriel.
- La "**méthode des perturbations des E.D.**" (voir plus bas) utile pour la physique quantique ondulatoire et le physique quantique des champs.

Signalons également d'autres méthodes très utilisées (grands classiques scolaires) mais qui sont principalement traitées au cas par cas dans les différents chapitres du site car les approches de résolution sont trop nombreuses et particulières:

- La "**méthode de séparation des variables E.D.**" (équation de la chaleur dans le chapitre de Thermodynamique, équation des vagues dans le chapitre de Génie Marin & Météo, équation d'évolution de Schrödinger dans le chapitre de Physique Quantique Ondulatoire, vibration d'un tambour dans le chapitre de Mécanique Ondulatoire) dont nous verrons un cas très particulier et simple plus bas mais pour laquelle il vaut mieux se référer aux chapitres mentionnés pour des exemples concrets.
- La "**méthode matricielle de résolution des E.D.**" et "**solution triviales des E.D.**" (modèle de Lotka-Volterra dans le chapitre de Dynamique des Populations, résonance de spin électronique ou nucléaire dans le chapitre de Physique Quantique Relativiste, modèle de Lorenz dans le chapitre de Génie Marin & Météo).
- La "**méthode de la transformée de Fourier des E.D.**" ou la "**méthode de la transformée de Laplace des E.D.**" (équation de la chaleur dans le chapitre d'Analyse Fonctionnelle, résolution de l'équation de Black & Scholes dans le chapitre d'Économie, équation de la poutre sous charge ponctuelle dans le chapitre de Génie Civil)
- Les "**méthodes numériques des E.D.**" pour résoudre les équations différentielles avec l'ordinateur quand elles n'ont pas de solutions analytiques connues (équation de la chaleur dans le chapitre de Méthodes Numériques).

6.3. MÉTHODE DU POLYNÔME CARACTÉRISTIQUE

La résolution des équations différentielles simples (à coefficients constants et sans second membre la plupart du temps...) utilise une technique faisant appel à un polynôme caractéristique de l'équation différentielle dont nous verrons les détails dans les développements à suivre sur quelques cas particuliers courants en physique.

C'est une méthode relativement simple à mettre en place lorsque nous cherchons les solutions homogènes de l'équation sans second membre (ESSM). Dans le cas contraire, celui de la présence d'un second membre, nous additionnons les solutions de l'équation homogène aux solutions particulières.

6.3.1. RÉOLUTION L'E.H. DE L'E.D.L. À COEFFICIENTS CONSTANTS D'ORDRE 1

Considérons l'E.D.L. à coefficients constants suivante:

$$\frac{dy}{dx} + ay = f(x) \quad (10.504)$$

qui est une version simplifiée de l'E.D.L à coefficients constants générale suivante:

$$a' \frac{dy}{dx} + b' y = 0 \quad (10.505)$$

où:

$$a = \frac{b'}{a'} \quad (10.506)$$

Nous écrivons son équation homogène (ESSM) associée:

$$\frac{dy}{dx} + ay = 0 \quad (10.507)$$

Ce qui peut s'écrire:

$$\frac{dy}{y} = -a dx \quad (10.508)$$

d'où:

$$\ln(y) = -ax + c^{te} \Rightarrow y_h = e^{-ax+c^{te}} = e^{c^{te}} e^{-ax} = C e^{-ax} \quad (10.509)$$

Il y a derrière cette solution homogène une infinité de solutions: à chaque valeur donnée à C correspond une solution.

Il faut encore à cette solution homogène ajouter la solution particulière y_p et nous disposons pour cela d'une collection de recettes, qui dépendent du type de la fonction $f(x)$ du second membre de l'équation. Nous les verrons au cas par cas dans les différents chapitres de Physique.

6.3.2. RÉOLUTION DE L'E.H. DE L'E.D.L. À COEFFICIENTS NON CONSTANTS DE L'E.D.L D'ORDRE 1

La solution générale des équations différentielles linéaires homogènes (ESSM) d'ordre 1 à coefficients non constants:

$$a(x)y' + b(x)y = 0 \quad (10.510)$$

peut toujours se réduire à la forme suivante:

$$y' + g(x)y = 0 \quad (10.511)$$

où:

$$g(x) = \frac{b(x)}{a(x)} \quad (10.512)$$

Bon évidemment il y a la solution $y = 0$... mais cherchons à faire mieux. Nous avons donc:

$$\frac{y'}{y} = -g(x) \quad (10.513)$$

Il vient donc:

$$\ln(y) = -G(x) + c^{te} \quad (10.514)$$

où $G(x)$ est une primitive de $g(x)$. Dès lors:

$$y = e^{c^{te}} e^{-G(x)} = A e^{-G(x)} \quad (10.515)$$

Il est aussi fréquent de retrouver ces développements sous une autre écriture un tout petit peu plus explicite qui est la suivante:

Nous repartons donc de l'équation différentielle sans seconde membre à coefficients non constante:

$$a(x) \frac{dy}{dx} + b(x)y = 0 \quad (10.516)$$

après réarrangement:

$$\frac{dy}{y} = -\frac{b(x)}{a(x)} dx \quad (10.517)$$

Et donc:

$$\ln(y_0) + c^{te} = -\int \frac{b(x)}{a(x)} dx \quad (10.518)$$

Soit:

$$y_0 = c^{te} e^{-\int \frac{b(x)}{a(x)} dx} \quad (10.519)$$

Ce résultat nous sera très utile pour calculer la transformée de Fourier d'une fonction Gaussienne (cf. chapitre Suite Et Séries), transformée de Fourier qui est indispensable pour résoudre de manière assez générale l'équation de la Chaleur (cf. chapitre de Thermodynamique), résolution qui nous permettra enfin de démontrer l'équation de Black & Scholes (cf. chapitre d'Économie).

6.3.3. RÉOLUTION L'E.H. DE L'E.D.L. À COEFFICIENTS CONSTANTS D'ORDRE 2

Considérons l'E.D.L. à coefficients constants suivante:

$$\frac{d^2 y}{dx^2} + a \frac{dy}{dx} + by = f(x) \quad (10.520)$$

qui est une version simplifiée de l'E.D.L à coefficients constants générale suivante:

$$a' \frac{d^2 y}{dx^2} + b' \frac{dy}{dx} + c' y = 0 \quad (10.521)$$

où:

$$a = \frac{b'}{a'} \quad b = \frac{c'}{a'} \quad (10.522)$$

Nous écrivons son équation homogène (ESSM) associée:

$$\frac{d^2 y}{dx^2} + a \frac{dy}{dx} + by = 0 \quad (10.523)$$

dans laquelle la fonction du second membre est nulle. Nous pouvons assez rapidement entrevoir une solution du type (en s'inspirant de la forme des solutions des E.D. du 1er ordre):

$$y = \lambda e^{K \cdot x + \tau} \quad (10.524)$$

où τ est une constante. Ce qui nous donne alors:

$$\lambda K^2 e^{K \cdot x + \tau} + a \lambda K e^{K \cdot x + \tau} + b \lambda e^{K \cdot x + \tau} = 0 \quad (10.525)$$

Ce que nous pouvons simplifier en:

$$K^2 + aK + b = 0 \quad (10.526)$$

Si notre hypothèse de départ est bonne, nous n'avons qu'à résoudre en K cette "**équation caractéristique**" (ECAR) ou "**polynôme caractéristique**" de l'équation homogène pour trouver la solution homogène:

$$K_{1,2} = \frac{-a \pm \sqrt{a^2 - 4b}}{2} \quad (10.527)$$

dont les solutions dépendent du signe du discriminant du polynôme caractéristique:

$$\Delta = a^2 - 4b \quad (10.528)$$

- Si le discriminant est strictement positif, soit $\Delta > 0$:

Alors nous savons que le polynôme caractéristique possède deux racines distinctes et nous avons alors:

$$y_h = A e^{K_1(x+\tau)} + B e^{K_2(x+\delta)} \quad (10.529)$$

où $K_1 \tau = c^{te}$ et $K_2 \delta = c^{te}$. Nous disons alors que la solution est "retardée" ou "avancée" selon les valeurs de ces constantes. Mais l'essentiel est de remarquer que si $y_h(x)$ est solution, alors $y_h(x \pm \Delta x)$ est toujours solution!

Nous parlons alors de "**solution générale de l'équation homogène**". Il y a derrière ce résultat une infinité de solutions: à chaque valeur donnée aux constantes A, B correspond une solution.

Les physiciens écrivent aussi parfois cela sous une forme particulière en posant d'abord:

$$\alpha = \frac{K_1 + K_2}{2} \quad \beta = \frac{K_1 - K_2}{2} \quad (10.530)$$

avec donc:

$$\alpha + \beta = K_1 \quad \alpha - \beta = K_2 \quad (10.531)$$

Et en utilisant les fonctions de trigonométrie hyperbolique (cf. chapitre de Trigonométrie):

$$\begin{aligned} \cosh(\beta x) &= \frac{e^{\beta x} + e^{-\beta x}}{2} \\ \sinh(\beta x) &= \frac{e^{\beta x} - e^{-\beta x}}{2} \end{aligned} \quad (10.532)$$

d'où finalement la possibilité d'écrire la solution homogène sous la forme (lorsque nous omettons l'avance ou le retard $\delta = \tau = 0$):

$$y_h = e^{\alpha x} (C_1 \cosh(\beta t) + C_2 \sinh(\beta t)) \quad (10.533)$$

Par ailleurs, montrons que les solutions de l'ESSM forment un espace vectoriel de dimension 2 (correspond donc à l'ordre de notre E.D.)!

En effet:

- La fonction zéro: $y = 0$ est solution de l'ESSM (ça c'est inutile de le démontrer... évident!).
- La somme ou soustraction des solutions reste solution (ça nous l'avons déjà démontré plus haut)
- Les éléments de la base de l'espace vectoriel (les solutions de l'ESSM) sont linéairement indépendants (ça c'est intéressant car nous en aurons besoin!).

Posons:

$$y_h = \lambda_1 f_1(x) + \lambda_2 f_2(x) \quad (10.534)$$

Alors:

$$\begin{aligned} y_h' &= \lambda_1 f_1' + \lambda_2 f_2' \\ y_h'' &= \lambda_1 f_1'' + \lambda_2 f_2'' \end{aligned} \quad (10.535)$$

Ces relations injectées dans l'ESSM sous forme généralisée:

$$a' \frac{d^2 y}{dx^2} + b' \frac{dy}{dx} + c' y = 0 \quad (10.536)$$

donne alors:

$$\begin{aligned}
 a'y_h'' + b'y_h' + c'y_h &= a'[\lambda_1 f_1'' + \lambda_2 f_2''] + b'[\lambda_1 f_1' + \lambda_2 f_2'] + c'[\lambda_1 f_1 + \lambda_2 f_2] \\
 &= \lambda_1 (a'f_1'' + b'f_1' + c'f_1) + \lambda_2 (a'f_2'' + b'f_2' + c'f_2) = \lambda_1 \cdot 0 + \lambda_2 \cdot 0 = 0
 \end{aligned}
 \tag{10.537}$$

Donc nous avons bien une structure d'espace vectoriel.

Rappelons que inversement deux fonctions sont linéairement dépendantes si:

$$\forall t \quad f_1(x) = \lambda \cdot f_2(x) \tag{10.538}$$

- Si le discriminant est nul, soit $\Delta = 0$:

L'équation caractéristique possède une racine double réelle K .

En allant un peu vite nous dirons alors:

$$y_h = Ae^{K(x+\tau)} \tag{10.539}$$

et que c'est fini... mais au fait ce serait oublier que la base vectorielle doit être formée de deux solutions indépendantes!

Donc la deuxième solution est probablement... de la forme:

$$y_{h2} = z(x)e^{K(x+\tau)} \tag{10.540}$$

Alors:

$$\begin{aligned}
 y_{h2}' &= Ke^{K(x+\tau)}z(x) + z'(x)e^{K(x+\tau)} \\
 y_{h2}'' &= K^2e^{K(x+\tau)}z(x) + Ke^{K(x+\tau)}z'(x) + Kz'(x)e^{K(x+\tau)} + z''(x)e^{K(x+\tau)} \\
 &= K^2e^{K(x+\tau)}z(x) + 2Ke^{K(x+\tau)}z'(x) + z''(x)e^{K(x+\tau)}
 \end{aligned}
 \tag{10.541}$$

Si nous l'injectons cela dans l'ESSM sous forme généralisée:

$$a'\frac{d^2y}{dx^2} + b'\frac{dy}{dx} + c'y = 0 \tag{10.542}$$

alors:

$$\begin{aligned}
 a'(K^2e^{K(x+\tau)}z + 2Ke^{K(x+\tau)}z' + z''e^{K(x+\tau)}) + b'(Ke^{K(x+\tau)}z + z'e^{K(x+\tau)}) + c'ze^{K(x+\tau)} &= 0 \\
 e^{K(x+\tau)}(a'K^2 + b'K + c')z + e^{K(x+\tau)}(2a'K + b')z' + a'e^{K(x+\tau)}z'' &= 0 \\
 (a'K^2 + b'K + c')z + (2a'K + b')z' + a'z'' &= 0
 \end{aligned}
 \tag{10.543}$$

Soit dans notre cas:

$$(K^2 + aK + b)z + (2K + a)z' + z'' = 0 \tag{10.544}$$

Or, les deux valeurs de K réelles sont justement solutions de:

$$K^2 + aK + b = 0 \quad (10.545)$$

La relation antécédente se réduit alors à:

$$(2K + a)z' + z'' = 0 \quad (10.546)$$

et comme nous sommes dans le cas d'étude où le discriminant est nul, il vient:

$$K_{1,2} = \frac{-a \pm \Delta}{2} = \frac{-a}{2} \quad (10.547)$$

Donc la relation antécédente se réduit au final à:

$$z'' = 0 \quad (10.548)$$

Nous en déduisons:

$$z'' = 0 \Rightarrow z' = c^{te} \Rightarrow z = Bx + c^{te} \quad (10.549)$$

Donc finalement:

$$y_{h2} = (Bx + c^{te})e^{K(x+\tau)} \quad (10.550)$$

Ce qui donne pour la solution générale de l'ESSM:

$$y_h = (C_1 x + C_2)e^{K(x+\tau)} \quad (10.551)$$

- Si le discriminant est négatif, soit $\Delta < 0$:

L'équation caractéristique possède deux racines complexes conjuguées (cf. chapitre d'Algèbre):

$$K_1 = \alpha + i\beta, K_2 = \alpha - i\beta \quad (10.552)$$

Dès lors:

$$\begin{aligned} y_h &= Ae^{(\alpha+i\beta)(x+\tau)} + Be^{(\alpha-i\beta)(x+\delta)} \\ &= Ae^{\alpha(x+\tau)}e^{i\beta(x+\tau)} + Be^{\alpha(x+\delta)}e^{-i\beta(x+\delta)} = e^{\alpha(x+\tau)}Ae^{i\beta(x+\tau)} + e^{\alpha(x+\delta)}Be^{-i\beta(x+\delta)} \\ &= Ae^{\alpha(x+\tau)}(\cos(\beta x + \tau\beta) + i\sin(\beta x + \tau\beta)) + Be^{\alpha(x+\delta)}(\cos(\beta x + \delta\beta) - i\sin(\beta x + \delta\beta)) \\ &= y_{h1} + y_{h2} \end{aligned} \quad (10.553)$$

Or, si nous cherchons plutôt des solutions réelles, nous pouvons toujours poser A et B égaux tels que:

$$\begin{aligned} y'_h &= \frac{1}{2}(y_{h1} + y_{h2}) + \frac{1}{2i}(y_{h1} - y_{h2}) = e^{\alpha(x+\tau)}\cos(\beta x + \tau\beta) + e^{\alpha(x+\delta)}\sin(\beta x + \delta\beta) \\ &= e^{\alpha x}e^{\alpha\tau}\cos(\beta x + \tau\beta) + e^{\alpha x}e^{\alpha\delta}\sin(\beta x + \delta\beta) = e^{\alpha x}(A'\cos(\beta x + \tau\beta) + B'\sin(\beta x + \delta\beta)) \end{aligned} \quad (10.554)$$

Et si nous posons que le retard et respectivement l'avance sont nuls ($\delta = \tau = 0$), alors nous retrouvons la relation disponible dans la plupart des livres:

$$y_h = e^{\alpha x} (A' \cos(\beta x) + B' \sin(\beta x)) \quad (10.555)$$

où A' et B' sont donc deux constantes réelles quelconques. Il existe une autre forme importante à cette dernière relation (souvent utilisée en électronique par exemple). Effectivement, Il est possible, pour tout A' et B' réels, de trouver C' et ϕ réels tels que l'égalité suivante soit vérifiée:

$$A' \cos(\beta x) + B' \sin(\beta x) = C' \cos(\beta t - \phi) \quad (10.556)$$

Nous posons:

$$C' = \sqrt{A'^2 + B'^2} \quad (10.557)$$

alors:

$$A' \cos(\beta x) + B' \sin(\beta x) = C' \left[\frac{A'}{M} \cos(\beta x) + \frac{B'}{M} \sin(\beta x) \right] \quad (10.558)$$

Il est alors possible de trouver ϕ tel que:

$$\frac{A'}{M} = \cos \phi \quad \text{et} \quad \frac{B'}{M} = \sin \phi \quad (10.559)$$

La quantité de départ s'écrit ainsi:

$$A' \cos(\beta x) + B' \sin(\beta x) = C' [\cos \phi \cos(\beta x) + \sin \phi \sin(\beta x)] = C' \cos(\beta t - \phi) \quad (10.560)$$

Finalement:

$$y_h = C' e^{\alpha x} \cos(\beta t - \phi) \quad (10.561)$$

Nous pouvons donc faire le résumé suivant:

Discriminant	Racines	Forme solution homogène
$\Delta > 0$	$K_{1,2} = \frac{-a \pm \sqrt{a^2 - 4b}}{2}$	$y_h = A e^{K_1(x+\tau)} + B e^{K_2(x+\delta)}$
$\Delta = 0$	$K_{1,2} = \frac{-a \pm \sqrt{a^2 - 4b}}{2}$	$y_h = (C_1 x + C_2) e^{K(x+\tau)}$
$\Delta < 0$	$K_{1,2} = \frac{-a \pm \sqrt{a^2 - 4b}}{2}$	$y_h = e^{\alpha x} (A' \cos(\beta x) + B' \sin(\beta x))$

Tableau: 28.1 - Solutions types de l'E.H. de l'E.D.L. à coefficients constants

6.4. MÉTHODE DU FACTEUR INTÉGRANT (D'EULER)

La technique du facteur d'intégration est utile lorsqu'il s'agit de résoudre des équations différentielles de la forme:

$$y' + f(x)y = g(x) \quad (10.562)$$

Nous n'avons pas à ce jour de cas pratique d'application de cette technique dans les autres chapitres du site. Il faut donc voir cela comme une présentation pour la culture générale.

L'idée de base étant de trouver une fonction $M(x)$, appelée "**facteur d'intégration**", par laquelle peut être multipliée notre équation différentielle pour ramener le terme de gauche de l'égalité à une simple dérivée. Par exemple, pour une équation différentielle linéaire comme celle ci-dessus, nous choisissons assez souvent le facteur d'intégration suivant (mais ce n'est de loin pas la seule possibilité et ce choix ne permet pas de tout résoudre!):

$$M(x) = e^{F(x)} \quad (10.563)$$

Nous avons alors:

$$e^{F(x)} (y' + f(x)y) = g(x)e^{F(x)} \quad (10.564)$$

ou en distribuant:

$$y'e^{F(x)} + f(x)ye^{F(x)} = g(x)e^{F(x)} \quad (10.565)$$

Ce qui peut donc être vu comme:

$$y'e^{F(x)} + F'(x)ye^{F(x)} = g(x)e^{F(x)} \quad (10.566)$$

ou encore plus fort (et c'est là que réside toute l'astuce)...:

$$(ye^{F(x)})' = g(x)e^{F(x)} \quad (10.567)$$

Nous pouvons alors prendre la primitive par rapport à x :

$$\int (ye^{F(x)})' dx = \int g(x)e^{F(x)} dx \quad (10.568)$$

et trivialement (!) nous avons la primitive de gauche qui est immédiate:

$$ye^{F(x)} = \int g(x)e^{F(x)} dx \quad (10.569)$$

Soit:

$$ye^{F(x)} = \int g(x)e^{F(x)} dx \quad (10.570)$$

Ce qui est parfois écrit:

$$y = \frac{1}{e^{\int f(x) dx}} \int g(x)e^{\int f(x) dx} dx \quad (10.571)$$

Exemple:

Considérons l'équation différentielle:

$$y' + \frac{y}{x} - 4 = 0 \quad (10.572)$$

Que nous mettrons sous la forme:

$$\frac{dy}{dx} + \frac{1}{x}y = 4 \quad (10.573)$$

Nous voyons alors que (en assumant que x soit strictement positif):

$$f(x) = \frac{1}{x} \Rightarrow F(x) = \ln(x) \quad (10.574)$$

Nous avons alors:

$$e^{\ln(x)} \left(\frac{dy}{dx} + \frac{1}{x}y \right) = e^{\ln(x)} 4 \quad (10.575)$$

Hasard faisant (l'exemple est exprès simple), nous avons cette égalité qui se simplifie puisque:

$$e^{\ln(x)} = x \quad (10.576)$$

en:

$$x \left(\frac{dy}{dx} + \frac{1}{x}y \right) = 4x \quad (10.577)$$

Ce qui peut se condenser en:

$$(xy)' = 4x \quad (10.578)$$

En intégrant:

$$\int (xy)' dx = \int 4x dx \quad (10.579)$$

Il vient alors immédiatement:

$$xy = 2x^2 + c^{te} \quad (10.580)$$

Soit:

$$y = 2x + \frac{c^{te}}{x} \quad (10.581)$$

6.5. MÉTHODE DE SÉPARATION DES VARIABLES

La méthode de séparation des variables est une technique très fréquente en physique dès que nous avons des équations différentielles du deuxième ordre. De nombreux exemples très complets et pratiques se trouvent déjà de part et d'autres dans les différents chapitres déjà précédemment mentionnés. Nous allons ici juste en présenter un cas particulier par principe juste histoire de bien faire les choses mais au minimum vital!

Considérons le cas fréquent en physique d'équation différentielle partielle du type:

$$\frac{\partial^2 U}{\partial x^2} - \frac{\partial U}{\partial y} = 0 \quad (10.582)$$

La solution de cette équation nécessite donc de trouver une fonction U qui dépende de x et de y tel que:

$$U = U(x, y) \quad (10.583)$$

En physique, l'idée consiste alors à poser que nous pouvons toujours trouver une solution dite séparable de la forme:

$$U(x, y) = X(x)Y(y) \quad (10.584)$$

Ainsi, l'équation différentielle s'écrit:

$$\frac{\partial^2 X(x)Y(y)}{\partial x^2} - \frac{\partial X(x)Y(y)}{\partial y} = 0 \quad (10.585)$$

Ce qui peut se simplifier en:

$$Y(y) \frac{\partial^2 X(x)}{\partial x^2} = X(x) \frac{\partial Y(y)}{\partial y} \quad (10.586)$$

Après réarrangement est il d'usage en physique de noter cette dernière égalité sous la forme condensée:

$$\frac{X''}{X} = \frac{Y'}{Y} \quad (10.587)$$

Cette égalité ne peut avoir lieu que si chacun des termes est une constante puisque X ne dépend que de x et Y que de y . Il vient alors:

$$\frac{X''}{X} = c^{te} \quad \frac{Y'}{Y} = c^{te} \quad (10.588)$$

Et chaque équation différentielle par alors être résolue indépendamment de l'autre et une fois les solutions trouvées on les multiplie pour donc déterminer l'expression de U .

6.6. MÉTHODE DE VARIATION DE LA CONSTANTE

L'idée de la méthode de variation de la constante est la suivante: si nous avons une solution particulière affectée de constantes, nous savons qu'en fonction des conditions initiales celles-ci sont bien déterminées. L'idée est alors de généraliser en posant que ces constantes sont des fonctions. Dans certains cas évidemment les développements mathématiques montreront que les fonctions sont obligatoirement des constantes.

L'idée sous-jacente de cette méthode, c'est de se dire que les solutions de l'équation différentielle (linéaire) avec second membre vont ressembler aux solutions de l'équation homogène. Comme le terme de droite va perturber cette solution, nous faisons varier uniquement les constantes (qui n'en seront plus), mais nous restons sur la "base" des solutions homogènes, pour chercher des solutions proches. Après, nous vérifions que ce raisonnement à la physicienne donne bien toutes les solutions de l'équation.

Voyons avant de passer au cas général un exemple simple en considérant l'équation différentielle suivante:

$$y'(t) + by(t) = 1 \quad (10.589)$$

pour laquelle la solution particulière de l'équation homogène (ESSM) est:

$$y_0(t) = Ae^{-bt} \quad (10.590)$$

La méthode de variation de la constante consiste alors à poser:

$$A = A(t) \quad (10.591)$$

et donc:

$$y_0'(t) = A'(t)e^{-bt} - bA(t)e^{-bt} \quad (10.592)$$

Or de par l'équation différentielle avec second membre, nous avons:

$$y_0'(t) + by_0(t) = 1 \Rightarrow y_0'(t) = 1 - by_0(t) = 1 - bAe^{-bt} \quad (10.593)$$

Donc en utilisant les deux dernières relations il vient:

$$\left. \begin{array}{l} y_0'(t) = A'(t)e^{-bt} - bA(t)e^{-bt} \\ y_0'(t) = 1 - bAe^{-bt} \end{array} \right\} \Rightarrow A'(t) = e^{bt} \quad (10.594)$$

et il s'ensuit que:

$$A(t) = \frac{1}{b}e^{bt} \quad (10.595)$$

où nous avons éliminé la constante d'intégration parce que nous voulons c'est une solution particulière! La solution générale particulière (pg) est alors la somme de la solution particulière homogène et celle avec la variation de la constante:

$$y_{0,pg}(t) = Ae^{-bt} + A'(t)e^{-bt} = Ae^{-bt} + \frac{1}{b}e^{bt}e^{-bt} = Ae^{-bt} + \frac{1}{b} \quad (10.596)$$

Ainsi, en généralisant l'exemple précédent, nous avons donc une équation différentielle de la forme:

$$y'(t) + by(t) = c \quad (10.597)$$

La solution particulière générale sera alors:

$$C_0 y_0(t) := C_0(t) y_0(t) \quad (10.598)$$

Nous avons alors:

$$y_0'(t) = C_0'(t) y_0(t) + C_0(t) y_0'(t) \quad (10.599)$$

d'où injecté dans l'équation différentielle d'origine:

$$\underbrace{C_0'(t) y_0(t) + C_0(t) y_0'(t)}_{y'(t)} + b(t) \underbrace{C_0(t) y_0(t)}_{y(t)} = c(t) \quad (10.600)$$

Soit après factorisation des termes semblables:

$$C_0'(t) y_0(t) + C_0(t) (y_0'(t) + b(t) C_0(t) y_0(t)) = c(t) \quad (10.601)$$

Nous avons donc la relation ci-dessus et la solution particulière à l'équation différentielle homogène (donc sans seconde membre):

$$\left. \begin{aligned} C_0'(t)y_0(t) - C_0(t)(y_0'(t) + b(t)C_0(t)y_0(t)) &= c(t) \\ y_0'(t) &= -b(t)C_0(t)y_0(t) \end{aligned} \right\} \Rightarrow C_0'(t)y_0(t) = c(t) \quad (10.602)$$

Nous trouvons donc:

$$C_0'(t) = \frac{c(t)}{y_0(t)} \quad (10.603)$$

et il suffit alors d'intégrer cette équation pour trouver $C_0(t)$. Ensuite, la solution générale particulière (pg) est alors la somme de la solution particulière homogène et de celle avec la variation de la constante.

6.7. SYSTÈMES D'ÉQUATIONS DIFFÉRENTIELLES

Voyons maintenant des développements particuliers qui vont aussi bien être utiles en physique quantique que dans la résolution de systèmes particuliers d'équations différentielles (et particulièrement une qui est connue en théorie du chaos!).

Indiquons d'abord au lecteur avant d'aller plus loin que le cas plus complexe non homogène (avec seconde membre) et avec coefficients inconnus est traité directement par l'exemple dans le chapitre de Génie Industriel lors du traitement de la fiabilité d'un système réparable sous la forme d'une chaîne de Markov avec traitement par les déterminants et valeurs/vecteurs propres.

Pour commencer cette première approche, il va nous falloir introduire le concept d'exponentiation d'une matrice:

L'ensemble des matrices $n \times n$ à coefficients dans $\mathbb{C}$ noté $M_n(\mathbb{C})$ est un espace vectoriel pour l'addition des matrices et la multiplication par un scalaire. Nous notons I la matrice identité.

Nous admettrons qu'une suite de matrices A_n converge vers une matrice A si et seulement si les suites de coefficients des matrices A_n convergent vers les coefficients correspondants de A .

Exemple:

Dans $M_2(\mathbb{C})$ la suite de matrices:

$$\begin{pmatrix} 1/n & 2 \\ 0 & \frac{n+1}{n} \end{pmatrix} \quad (10.604)$$

converge vers:

$$\begin{pmatrix} 0 & 2 \\ 0 & 1 \end{pmatrix} \quad (10.605)$$

lorsque $n \rightarrow \infty$.

Si $x \in \mathbb{C}$, nous avons vu lors de notre étude des nombres complexes (cf. chapitre sur les Nombres) que la série:

$$\sum_{k=0}^{\infty} \frac{x^k}{k!} \quad (10.606)$$

converge et sa limite est notée e^x . En fait ici il n'y a aucune difficulté à remplacer x par une matrice A puisque nous savons (nous l'avons montré lors de notre étude des nombres complexes) que tout nombre complexe peut s'écrire sous la forme suivante (le corps des nombres complexes est donc isomorphe au corps des matrices réelles carrées de dimensions 2 ayant cette forme):

$$x = a + ib = a \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} b = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \quad (10.607)$$

et qu'un nombre complexe au carré est équivalent à mettre sa forme matricielle au carré:

$$x^2 = (a + ib)(a + ib) = a^2 + aib + iba - b^2 = (a^2 - b^2) + 2abi = \begin{bmatrix} a^2 - b^2 & -2ab \\ 2ab & a^2 - b^2 \end{bmatrix} \quad (10.608)$$

Effectivement:

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix}^2 = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{bmatrix} a & -b \\ b & a \end{bmatrix} = \begin{bmatrix} a^2 - b^2 & -2ab \\ 2ab & a^2 - b^2 \end{bmatrix} \quad (10.609)$$

Nous définissons alors l'exponentielle d'une matrice $A \in M_n(\mathbb{C})$ comme la matrice limite de la suite:

$$e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!} \quad (10.610)$$

Si la matrice A est diagonale il est évident que son exponentielle est facile à calculer. En effet, si:

$$A = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} \Rightarrow A^k = \begin{bmatrix} \lambda_1^k & & 0 \\ & \ddots & \\ 0 & & \lambda_n^k \end{bmatrix} \quad (10.611)$$

Par suite:

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} \begin{bmatrix} \lambda_1^k & & 0 \\ & \ddots & \\ 0 & & \lambda_n^k \end{bmatrix} = \begin{bmatrix} \sum_{k=0}^{\infty} \frac{\lambda_1^k}{k!} & & 0 \\ & \ddots & \\ 0 & & \sum_{k=0}^{\infty} \frac{\lambda_n^k}{k!} \end{bmatrix} = \begin{bmatrix} e^{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & e^{\lambda_n} \end{bmatrix} \quad (10.612)$$

Or, il apparaît évident qu'une matrice non diagonale va être beaucoup plus compliquée à traiter! Nous allons alors utiliser la technique de diagonalisation soit une réduction des endomorphismes (cf. chapitre d'Algèbre Linéaire).

Alors, remarquons que si $S \in M_n(\mathbb{C})$ est inversible et si $A \in M_n(\mathbb{C})$ alors:

$$e^{SAS^{-1}} = Se^AS^{-1} \quad (10.613)$$

Ceci découle du fait que (penser au changement de base d'une application linéaire comme ce qui a été étudié dans le chapitre d'Algèbre Linéaire):

$$(SAS^{-1})^k = \cancel{SA} \cancel{S^{-1}S} \cancel{A} \cancel{S^{-1}S} \cancel{AS^{-1}S} \dots \cancel{S^{-1}S} \cancel{AS^{-1}} = SA^kS^{-1} \quad (10.614)$$

Donc:

$$e^{SAS^{-1}} = \sum_{k=0}^{\infty} \frac{SA^kS^{-1}}{k!} = S \left(\sum_{k=0}^{\infty} \frac{A^k}{k!} \right) S^{-1} \quad (10.615)$$

Ce développement va nous permettre de ramener le calcul de l'exponentielle d'une matrice diagonalisable à la recherche de ses valeurs propres et de ses vecteurs propres.

Exemple:

Calculons e^A où:

$$A = \begin{pmatrix} 3 & -4 \\ -6 & 1 \end{pmatrix} \quad (10.616)$$

Les valeurs propres de A sont $\lambda_1 = -3$, $\lambda_2 = 7$ et les vecteurs propres associés sont:

$$\vec{v}_1 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad (10.617)$$

Effectivement:

$$\begin{pmatrix} 3 & -4 \\ -6 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \end{pmatrix} = -3 \begin{pmatrix} 2 \\ 3 \end{pmatrix} \text{ et } \begin{pmatrix} 3 & -4 \\ -6 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 7 \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad (10.618)$$

En posant:

$$S = \begin{pmatrix} 2 & 1 \\ 3 & -1 \end{pmatrix} \quad (10.619)$$

Nous avons:

$$A = S \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} S^{-1} = S \begin{pmatrix} -3 & 0 \\ 0 & 7 \end{pmatrix} S^{-1} \quad (10.620)$$

avec:

$$S^{-1} = \begin{pmatrix} 1/5 & 1/5 \\ 3/5 & -2/5 \end{pmatrix} \quad (10.621)$$

Par conséquent:

$$e^A = e^{\begin{pmatrix} -3 & 0 \\ 0 & 7 \end{pmatrix} S^{-1}} = S e^{\begin{pmatrix} -3 & 0 \\ 0 & 7 \end{pmatrix}} S^{-1} = S \begin{pmatrix} e^{-3} & 0 \\ 0 & e^7 \end{pmatrix} S^{-1} = \frac{1}{5} \begin{pmatrix} 2e^{-3} + 3e^7 & 2e^{-3} - 2e^7 \\ 3e^{-3} - 3e^7 & 3e^{-3} + 2e^7 \end{pmatrix} \quad (10.622)$$

Maintenant, rappelons que dans le cas des nombres réels nous savons que si $x, y \in \mathbb{R}$ alors:

$$e^{x+y} = e^x e^y \quad (10.623)$$

Dans le cas des matrices nous pouvons montrer que si $A, B \in M_n(\mathbb{C})$ sont deux matrices qui commutent entre-elles c'est-à-dire telles que $AB = BA$, alors:

$$e^{A+B} = e^A e^B \quad (10.624)$$

La condition de commutativité tient au fait que l'addition dans l'exponentielle est quant à elle commutative. La démonstration est donc intuitive.

Un corollaire important de cette proposition est que pour toute matrice $A \in M_n(\mathbb{C})$, e^A est inversible. En effet les matrices A et $-A$ commutent, par conséquent:

$$e^A e^{-A} = e^{A-A} = e^0 = I \quad (10.625)$$

Nous rappelons qu'une matrice A à coefficients complexes est unitaire si:

$$AA^\dagger = I \quad (10.626)$$

La proposition suivante nous servira par la suite.

Montrons que si A est une matrice hermitienne (dite aussi "autoadjointe") (cf. chapitre d'Algèbre Linéaire) alors pour tout $t \in \mathbb{R}$, e^{itA} est unitaire.

Démonstration:

$$\left(e^{itA}\right)^\dagger = e^{\left(itA\right)^\dagger} = e^{-itA^\dagger} = e^{-itA} \quad (10.627)$$

Donc:

$$\left(e^{itA}\right)^\dagger \cdot e^{itA} = e^{-itA} e^{itA} = e^{it(A-A)} = I \quad (10.628)$$

C.Q.F.D.

Rappelons que cette condition pour une matrice autoadjointe est liée à la définition de groupe unitaire d'ordre n (cf. chapitre d'Algèbre Ensembliste).

Une des premières applications de l'exponentielle de matrices est la résolution des équations différentielles ordinaires. En effet, de l'équation différentielle linéaire ci-dessous avec comme condition initiale $y(0) = y_0$ et où A est une matrice:

$$y'(t) + Ay(t) = 0 \quad (10.629)$$

la solution est donnée (cf. chapitre de Calcul Différentiel et Intégral) par:

$$y(t) = Ce^{-At} \quad (10.630)$$

Nous retrouvons fréquemment ce genre de systèmes d'équations différentielles en biologie (dynamique des populations), en astrophysique (étude des plasmas) ou en mécanique des fluides (théorie du chaos) ainsi qu'en mécanique classique (systèmes couplés), en astronomie (orbites couplées), en électrotechnique, etc.

Exemple:

Supposons que nous ayons le système d'équations différentielles homogène (sans termes constants) suivant:

$$\begin{aligned} x' &= 3x - 4y \\ y' &= -6x + y \end{aligned} \quad (10.631)$$

La matrice associée est alors:

$$A = \begin{pmatrix} 3 & -4 \\ -6 & 1 \end{pmatrix} \quad (10.632)$$

et son exponentielle (voir les développements faits plus haut):

$$e^{-At} = \frac{1}{5} \begin{pmatrix} 2e^{-3t} + 3e^{7t} & 2e^{-3t} - 2e^{7t} \\ 3e^{-3t} - 3e^{7t} & 3e^{-3t} + 2e^{7t} \end{pmatrix} \quad (10.633)$$

La solution générale du système est donc:

$$\begin{pmatrix} x \\ y \end{pmatrix} = C_1 \begin{pmatrix} 2e^{-3t} + 3e^{7t} \\ 3e^{-3t} - 3e^{7t} \end{pmatrix} + C_2 \begin{pmatrix} 2e^{-3t} - 2e^{7t} \\ 3e^{-3t} + 2e^{7t} \end{pmatrix} \quad (10.634)$$

Nous avons donc:

$$\begin{aligned} x &= C_1 (2e^{-3t} + 3e^{7t}) + C_2 (2e^{-3t} - 2e^{7t}) = A'e^{-3t} + B'e^{7t} \\ y &= C_1 (3e^{-3t} - 3e^{7t}) + C_2 (3e^{-3t} + 2e^{7t}) = D'e^{-3t} + E'e^{7t} \end{aligned} \quad (10.635)$$

En calculant la dérivée des relations précédentes et en comparant à:

$$\begin{aligned} x' &= 3x - 4y \\ y' &= -6x + y \end{aligned} \quad (10.636)$$

nous déterminons facilement les constantes pour obtenir:

$$\begin{aligned} E' - 6B' &= 7E' \Rightarrow -6E' = 6B' \Rightarrow E' = -B' \\ -6A' + D' &= -3D' \Rightarrow 4D' = 6A' \Rightarrow D' = \frac{3}{2}A' \end{aligned} \quad (10.637)$$

ce qui nous donne finalement:

$$\begin{aligned} x &= A'e^{-3t} + B'e^{7t} \\ y &= \frac{3}{2}A'e^{-3t} - B'e^{7t} \end{aligned} \quad (10.638)$$

6.8. MÉTHODE RÉGULIÈRE DES PERTURBATIONS

Très fréquemment en physique (de pointe), un problème mathématique ne peut pas être résolu de manière exacte. Si la solution est connue il y a parfois une telle dépendance de paramètres que la solution est difficile à utiliser en tant que telle.

Il peut arriver cependant qu'un paramètre identifié de l'équation différentielle, que nous noterons par tradition avec la lettre grecque ε , soit tel que la solution soit disponible et raisonnablement simple pour $\varepsilon = 0$.

Le souci ensuite est de savoir comment la solution est altérée pour un ε non-nul mais petit quand même. Cette étude est le centre de la théorie des perturbations que nous utilisons par exemple dans le chapitre de relativité générale pour calculer la précession du périhélie de Mercure.

Comme la théorie dans le cadre général est trop complexe par rapport aux objectifs du site, nous nous proposons une approche par l'exemple d'abord avec une simple équation algébrique et ensuite avec ce qui nous intéresse: une E.D.

6.8.1. THÉORIE PERTURBATIVE DES ÉQUATIONS ALGÈBRIQUES

Considérons l'équation polynômiale suivante:

$$x^2 - 1 = \varepsilon x \quad (10.639)$$

Nous savons de par notre étude du chapitre d'analyse fonctionnelle, que cette équation polynômiale admet deux racines qui sont trivialement:

$$x_{1,2} = \frac{\varepsilon \pm \sqrt{\varepsilon^2 + 4}}{2} = \frac{\varepsilon}{2} \pm \sqrt{1 + \frac{\varepsilon^2}{4}} \quad (10.640)$$

Pour ε petit, ces racines peuvent être approximées par le premier terme en développement de série de Taylor (cf. chapitre de Suites Et Séries):

$$x_1 = 1 + \frac{\varepsilon}{2} + \frac{\varepsilon^2}{8} + O(\varepsilon^3) \quad x_2 = -1 + \frac{\varepsilon}{2} - \frac{\varepsilon^2}{8} + O(\varepsilon^3) \quad (10.641)$$

La question est de savoir si nous pouvons obtenir les deux relations précédentes sans a priori de connaissances sur la solution exacte de l'équation polynômiale initiale? La réponse est bien évidemment affirmative avec l'aide de la théorie des perturbations.

La technique se base en quatre étapes:

1. Dans la première étape, nous assumons que la solution de l'équation polynômiale est une expression du type série de Taylor en ε . Nous avons alors:

$$x = X_0 + \varepsilon X_1 + \varepsilon^2 X_2 + O(\varepsilon^3) \quad (10.642)$$

où X_0, X_1, X_2 sont bien évidemment à déterminer.

2. Dans la deuxième étape, nous injectons la solution hypothétique dans notre équation polynômiale:

$$\left(X_0 + \varepsilon X_1 + \varepsilon^2 X_2 + O(\varepsilon^3)\right)^2 - 1 = \varepsilon \left(X_0 + \varepsilon X_1 + \varepsilon^2 X_2 + O(\varepsilon^3)\right) \quad (10.643)$$

Comme:

$$\begin{aligned} x^2 &= X_0^2 + 2\left(\varepsilon X_1 + \varepsilon^2 X_2 + O(\varepsilon^3)\right)X_0 + \left(\varepsilon X_1 + \varepsilon^2 X_2 + O(\varepsilon^3)\right)^2 \\ &= X_0^2 + 2\left(\varepsilon X_1 + \varepsilon^2 X_2 + O(\varepsilon^3)\right)X_0 + \left(\left(\varepsilon X_1 + \varepsilon^2 X_2\right)^2 + 2\left(\varepsilon X_1 + \varepsilon^2 X_2\right)O(\varepsilon^3) + O^2(\varepsilon^3)\right) \\ &\cong X_0^2 + 2\left(\varepsilon X_1 + \varepsilon^2 X_2\right)X_0 + \left(\varepsilon X_1 + \varepsilon^2 X_2\right)^2 + O(\varepsilon^3) \\ &= X_0^2 + 2\left(\varepsilon X_1 + \varepsilon^2 X_2\right)X_0 + \left(\varepsilon^2 X_1^2 + 2\varepsilon^3 X_1 X_2 + \varepsilon^4 X_2^2\right) + O(\varepsilon^3) \\ &\cong X_0^2 + 2\left(\varepsilon X_1 + \varepsilon^2 X_2\right)X_0 + \left(\varepsilon^2 X_1^2\right) + O(\varepsilon^3) \\ &= X_0^2 + 2\varepsilon X_0 X_1 + \varepsilon^2 \left(X_1^2 + 2X_0 X_2\right) + O(\varepsilon^3) \end{aligned} \quad (10.644)$$

et:

$$\varepsilon x = \varepsilon X_0 + \varepsilon^2 X_1 + O(\varepsilon^3) \quad (10.645)$$

Il vient finalement que l'équation polynômiale s'écrit:

$$X_0^2 - 1 + \varepsilon(2X_0 X_1 - X_0) + \varepsilon^2(X_1^2 + 2X_0 X_2 - X_1) + O(\varepsilon^3) = 0 \quad (10.646)$$

3. Dans la troisième étape nous égalisons successivement les termes avec 0 tel que:

$$\begin{aligned}
O(\varepsilon^0): X_0^2 - 1 &= 0 \\
O(\varepsilon^1): 2X_0X_1 - X_0 &= 0 \\
O(\varepsilon^2): X_1^2 + 2X_0X_2 - X_1 &= 0 \\
O(\varepsilon^2): \dots
\end{aligned} \quad (10.647)$$

4. Quatrième et dernière étape, nous résolvons successivement les équations polynômiales ci-dessus pour obtenir:

$$X_0 = \pm 1 \Rightarrow X_1 = \frac{1}{2} \Rightarrow X_2 = \pm \frac{1}{8} \quad (10.648)$$

En injectant ces résultants dans la solution hypothétique:

$$x = X_0 + \varepsilon X_1 + \varepsilon^2 X_2 + O(\varepsilon^3) \quad (10.649)$$

il est évident d'observer que nous retombons sur la solution certaine:

$$x_1 = 1 + \frac{\varepsilon}{2} + \frac{\varepsilon^2}{8} + O(\varepsilon^3) \quad x_1 = -1 + \frac{\varepsilon}{2} - \frac{\varepsilon^2}{8} + O(\varepsilon^3) \quad (10.650)$$

6.8.2. THÉORIE PERTURBATIVE DES ÉQUATIONS DIFFÉRENTIELLES

La théorie des perturbations est aussi souvent utilisée pour résoudre un bon nombre d'équations différentielles. C'est le cas par exemple en mécanique des fluides, en relativité générale ou en physique quantique.

A nouveau, plutôt que de faire une théorie ultra abstraite et générale, voyons le concept sur un exemple tel que précédemment.

Considérons l'équation différentielle suivante:

$$y'' = -\varepsilon y' - 1 \quad (10.651)$$

ou autrement écrit:

$$y'' + \varepsilon y' = -1 \quad (10.652)$$

avec les conditions aux limites:

$$y(0) = 0, y'(0) = 1 \quad (10.653)$$

La résolution exacte est relativement facile à obtenir:

D'abord nous commençons par l'équation homogène:

$$y'' + \varepsilon y' = 0 \quad (10.654)$$

C'est donc une équation différentielle linéaire d'ordre 2 avec des coefficients constants, équation qu'il est relativement aisé de résoudre dans le cas général. Soit l'équation:

$$y'' + py' + qy = 0 \quad (10.655)$$

Supposons que la fonction y qui satisfait cette équation différentielle soit de la forme $y = e^{Kx}$ où K peut être un nombre complexe. Nous avons alors:

$$K^2 e^{Kx} + pK e^{Kx} + q e^{Kx} = 0 \text{ ou } K^2 + pK + q = 0 \quad (10.656)$$

pourvu, bien sûr, que $e^{Kx} \neq 0$. Cette dernière relation est donc l'équation quadratique auxiliaire de l'équation différentielle (polynôme caractéristique). Elle a deux solutions/racines (c'est une simple résolution d'un polynôme du deuxième degré) que nous noterons dans le cas général: K_1, K_2 . Ce qui signifie que:

$$K_1^2 e^{K_1 x} + pK_1 e^{K_1 x} + q e^{K_1 x} = 0 \text{ et } K_2^2 e^{K_2 x} + pK_2 e^{K_2 x} + q e^{K_2 x} = 0 \quad (10.657)$$

sont satisfaites pour les deux racines. Si nous faisons la somme puisque les deux sont égales à la même constante:

$$(K_1^2 e^{K_1 x} + K_2^2 e^{K_2 x}) + p(K_1 e^{K_1 x} + K_2 e^{K_2 x}) + q(e^{K_1 x} + e^{K_2 x}) = 0 \quad (10.658)$$

Ainsi, il est immédiat que la solution générale de l'équation homogène de y est du type:

$$y_h = A e^{K_1 x} + B e^{K_2 x} \quad (10.659)$$

où A, B sont bien évidemment des constantes à déterminer. Nous résolvons maintenant le polynôme caractéristique:

$$K^2 + \varepsilon K = 0 \quad (10.660)$$

Il vient immédiatement que:

$$K_{1,2} = \frac{-\varepsilon \pm \sqrt{\varepsilon^2}}{2} \Rightarrow \begin{cases} K_1 = -\varepsilon \\ K_2 = 0 \end{cases} \quad (10.661)$$

Donc:

$$y_h = A e^{-\varepsilon x} \quad (10.662)$$

Maintenant une solution particulière à:

$$y'' + \varepsilon y' = -1 \quad (10.663)$$

est relativement trivialement une solution du type:

$$y_p = Bx \quad (10.664)$$

où B est bien évidemment une constante à déterminer et qui vaut simplement une fois injectée dans l'équation différentielle:

$$y'' + \varepsilon y' = -1 \Rightarrow 0 + \varepsilon B = -1 \Rightarrow B = -\frac{1}{\varepsilon} \quad (10.665)$$

Soit:

$$y_p = -\frac{x}{\varepsilon} + c^{te} \quad (10.666)$$

D'où finalement la solution générale:

$$y = y_p + y_h = Ae^{-\varepsilon x} - \frac{x}{\varepsilon} + c^{te} \quad (10.667)$$

Ensuite, avec les conditions initiales qui sont pour rappel

$$y(0) = 0, y'(0) = 1 \quad (10.668)$$

il est très facile de trouver A :

$$y(0) = 0 \Rightarrow Ae^{-\varepsilon 0} - \frac{0}{\varepsilon} + c^{te} = A + c^{te} = 0 \Rightarrow c^{te} = -A \quad (10.669)$$

Nous avons aussi:

$$\begin{aligned} y'(0) = 1 &\Rightarrow -\varepsilon Ae^{-\varepsilon 0} - \frac{1}{\varepsilon} + c^{te} = -\varepsilon A - \frac{1}{\varepsilon} + c^{te} = 1 \\ \Rightarrow A &= -\frac{1}{\varepsilon} \left(1 - c^{te} + \frac{1}{\varepsilon} \right) = -\left(\frac{\varepsilon - \varepsilon c^{te} + 1}{\varepsilon^2} \right) \end{aligned} \quad (10.670)$$

Il est loisible de choisir que $c^{te} = 0$ ce qui nous donne:

$$A = -\left(\frac{\varepsilon + 1}{\varepsilon^2} \right) \quad (10.671)$$

Dès lors:

$$y = y_p + y_h = Ae^{-\varepsilon x} - \frac{x}{\varepsilon} + c^{te} \quad (10.672)$$

devient:

$$y = -\frac{\varepsilon + 1}{\varepsilon^2} e^{-\varepsilon x} - \frac{x}{\varepsilon} + c^{te} = -\frac{1 + \varepsilon}{\varepsilon^2} e^{-\varepsilon x} - \frac{x}{\varepsilon} + \frac{1 + \varepsilon}{\varepsilon^2} = \frac{1 + \varepsilon}{\varepsilon^2} (1 - e^{-\varepsilon x}) - \frac{x}{\varepsilon} \quad (10.673)$$

Maintenant que nous avons la solution générale, si ε est petit nous pouvons prendre le développement d'ordre 4 en série de Maclaurin de l'exponentielle (cf. chapitre de Suites Et Séries). Tel que:

$$e^{-\varepsilon x} = 1 - \varepsilon x + \frac{1}{2} \varepsilon^2 x^2 - \frac{1}{6} \varepsilon^3 x^3 + O(\varepsilon^4 x^4) \quad (10.674)$$

Injecté dans y cela donne (vous remarquerez que nous exprimons parfois explicitement par anticipation..... le terme d'ordre 5):

$$\begin{aligned} y &= \frac{1+\varepsilon}{\varepsilon^2} (1 - e^{-\varepsilon x}) - \frac{x}{\varepsilon} = \frac{1+\varepsilon}{\varepsilon^2} \left(\varepsilon x - \frac{1}{2} \varepsilon^2 x^2 + \frac{1}{6} \varepsilon^3 x^3 - O(\varepsilon^4 x^4) \right) - \frac{x}{\varepsilon} \\ &= \frac{1}{\varepsilon^2} \left(\varepsilon x - \frac{1}{2} \varepsilon^2 x^2 + \frac{1}{6} \varepsilon^3 x^3 - O(\varepsilon^4 x^4) \right) + \frac{1}{\varepsilon} \left(\varepsilon x - \frac{1}{2} \varepsilon^2 x^2 + \frac{1}{6} \varepsilon^3 x^3 - O(\varepsilon^4 x^4) \right) - \frac{x}{\varepsilon} \\ &= \frac{1}{\varepsilon} x - \frac{1}{2} x^2 + \frac{1}{6} \varepsilon x^3 - \frac{1}{24} \varepsilon^2 x^4 + x - \frac{1}{2} \varepsilon x^2 + \frac{1}{6} \varepsilon^2 x^3 - O(\varepsilon^3 x^4) - \frac{x}{\varepsilon} \\ &= x - \frac{1}{2} x^2 + \varepsilon \left(-\frac{1}{2} x^2 + \frac{1}{6} x^3 \right) + \varepsilon^2 \left(\frac{1}{6} x^3 - \frac{1}{24} x^4 \right) - O(\varepsilon^3 x^4) \end{aligned} \quad (10.675)$$

Maintenant que nous avons ce développement, ce que nous souhaitons montrer c'est qu'à partir d'un développement perturbatif nous pouvons retrouver le même résultat en série et ce sans aucune connaissance préalable sur la solution.

A nouveau, le développement pour cela se fait en 4 étapes:

1. Dans la première étape, nous assumons que la solution de l'équation différentielle est une expression du type série de Taylor en ε . Nous avons alors:

$$y = y_0 + \varepsilon y_1 + \varepsilon^2 y_2 + O(\varepsilon^3) \quad (10.676)$$

où y_0, y_1, y_2 sont bien évidemment à déterminer.

2. Dans la deuxième étape, nous injectons la solution hypothétique de notre équation différentielle dans celle-ci avec les conditions initiales et nous développons le tout.

D'abord l'équation différentielle:

$$\begin{aligned} y'' + \varepsilon y' + 1 &= (y_0'' + \varepsilon y_1'' + \varepsilon^2 y_2'' + O(\varepsilon^3)) + \varepsilon (y_0' + \varepsilon y_1' + \varepsilon^2 y_2' + O(\varepsilon^3)) + 1 \\ &= (y_0'' + 1) + \varepsilon (y_1'' + y_0') + \varepsilon^2 (y_2'' + y_1') + O(\varepsilon^3) = 0 \end{aligned} \quad (10.677)$$

ensuite les conditions initiales:

$$\begin{aligned} y(0) &= y_0(0) + \varepsilon y_1(0) + \varepsilon^2 y_2(0) + O(\varepsilon^3) = 0 \\ y'(0) &= (y_0'(0) - 1) + \varepsilon y_1'(0) + \varepsilon^2 y_2'(0) + O(\varepsilon^3) = 0 \end{aligned} \quad (10.678)$$

3. Dans la troisième étape nous égalisons successivement les termes avec 0 tel que:

$$\begin{aligned}
O(\varepsilon^0): \quad & y_0'' + 1 = 0 \quad y_0(0) = 0 \quad y_0'(0) - 1 = 0 \\
O(\varepsilon^1): \quad & y_1'' + y_0' = 0 \quad y_1(0) = 0 \quad y_1'(0) = 0 \\
O(\varepsilon^2): \quad & y_2'' + y_1' = 0 \quad y_2(0) = 0 \quad y_2'(0) = 0 \\
O(\varepsilon^3): \quad & \dots
\end{aligned} \tag{10.679}$$

4. Dans la quatrième étape nous résolvons les équations différentielles listées précédemment (si vous ne voyez pas comment nous les résolvons n'hésitez pas à nous contacter!):

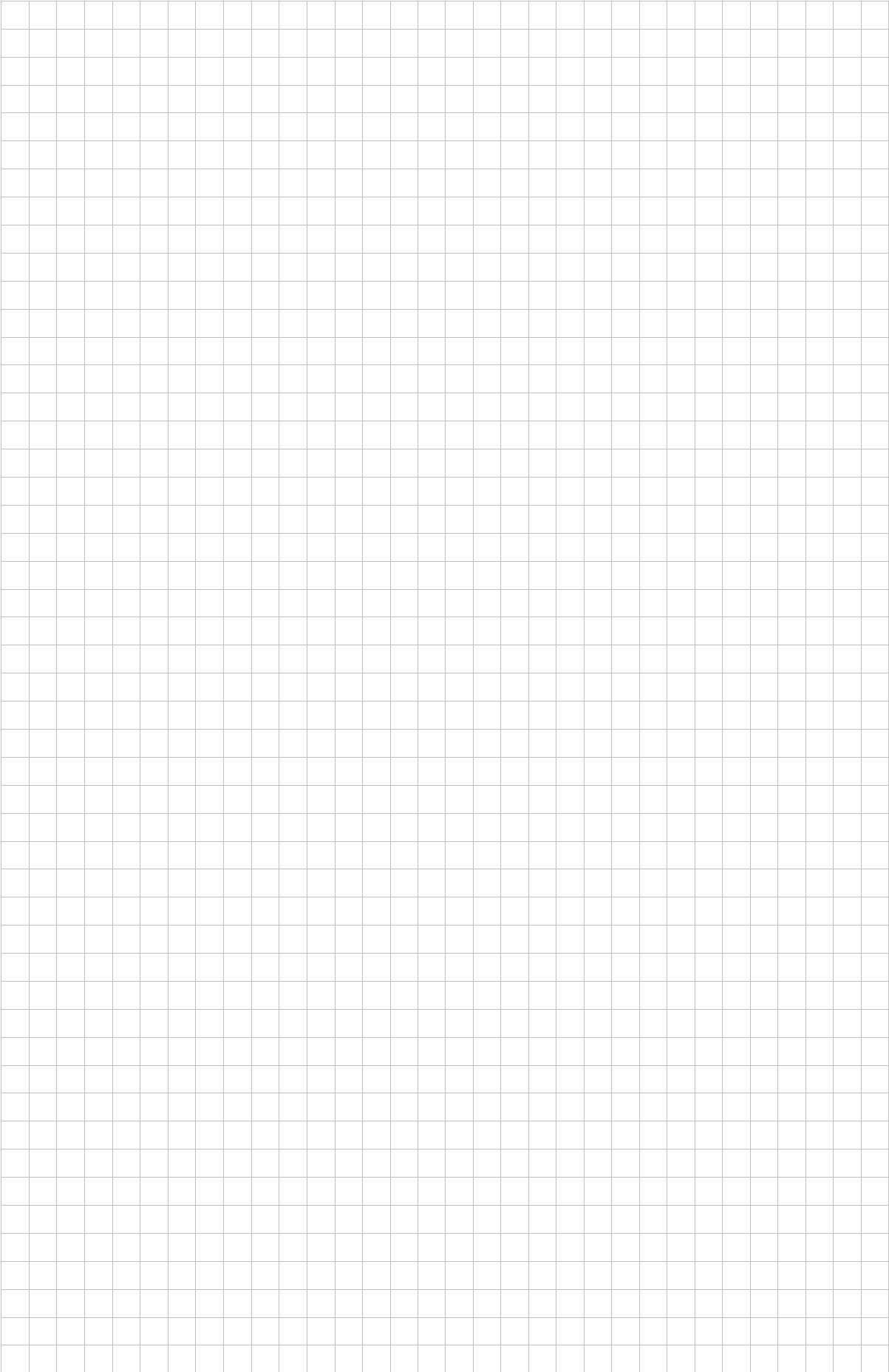
$$\begin{aligned}
y_0 &= x - \frac{1}{2}x^2 \\
y_1 &= -\frac{1}{2}x^2 + \frac{1}{6}x^3 \\
y_2 &= \frac{1}{6}x^3 - \frac{1}{24}x^4
\end{aligned} \tag{10.680}$$

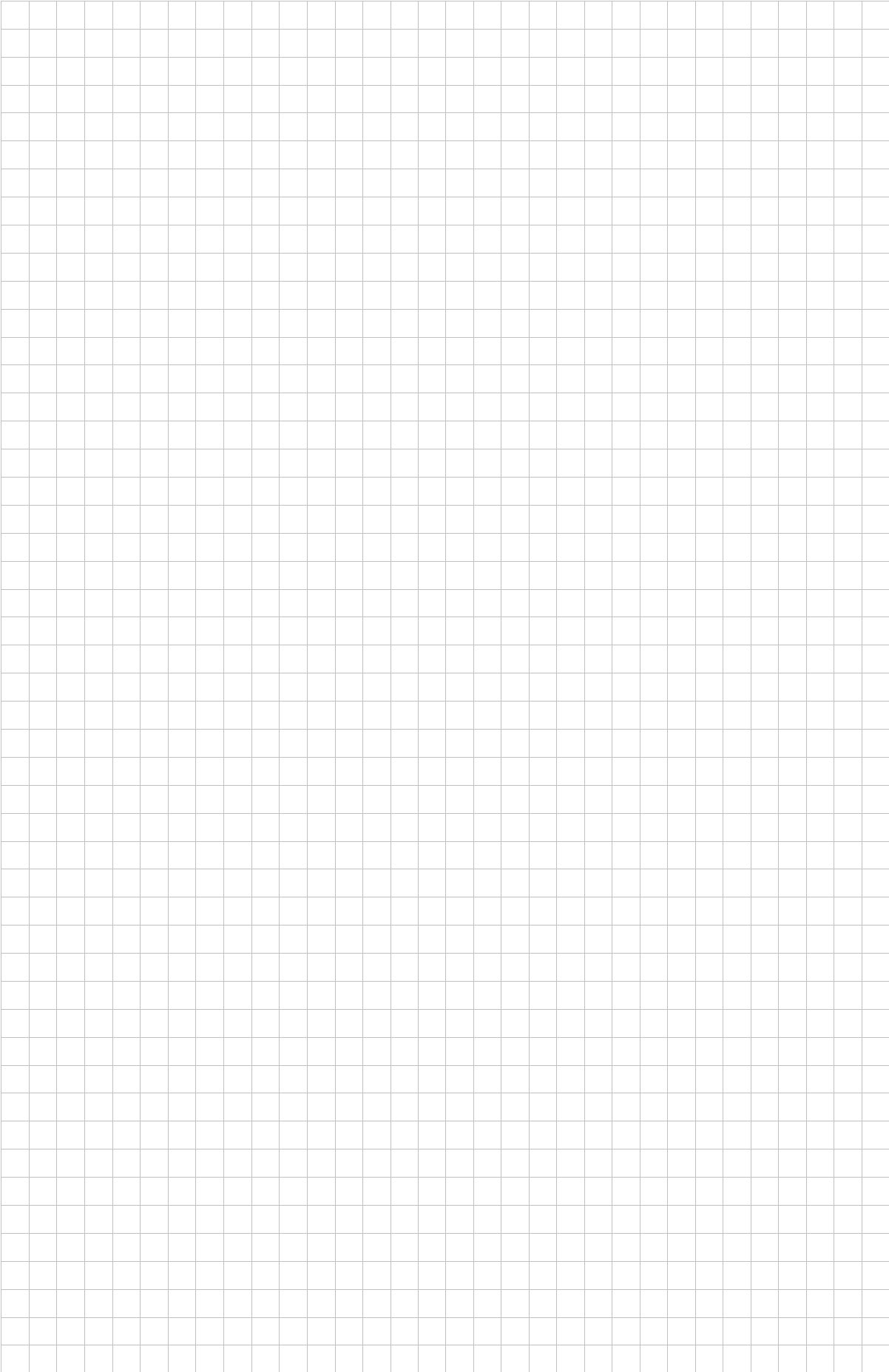
En injectant ces relations dans la solution supposée développée en série de Taylor et injectée dans l'équation différentielle:

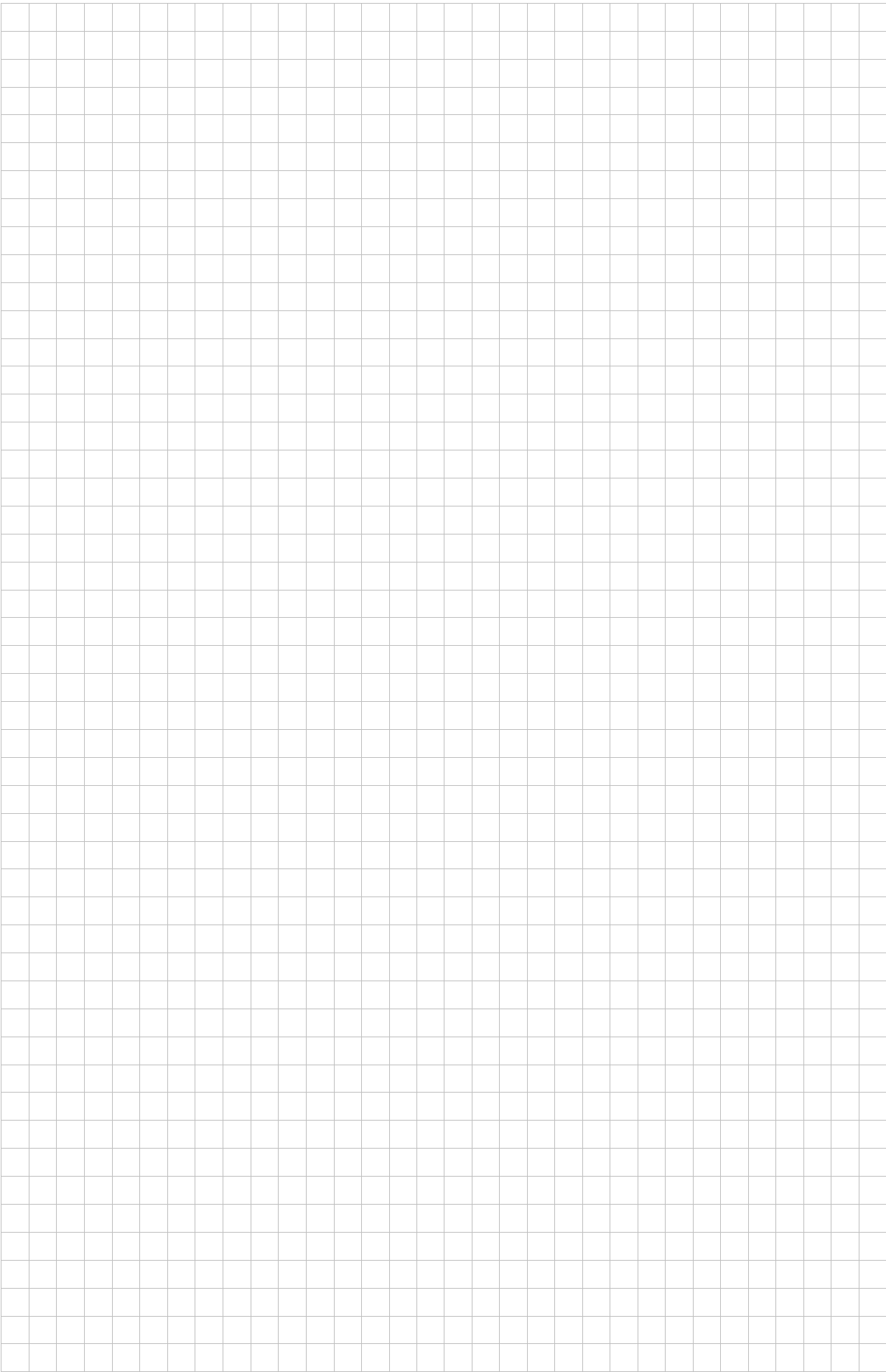
$$y = y_0 + \varepsilon y_1 + \varepsilon^2 y_2 + O(\varepsilon^3) \tag{10.681}$$

Nous retombons sur:

$$y = x - \frac{1}{2}x^2 + \varepsilon \left(-\frac{1}{2}x^2 + \frac{1}{6}x^3 \right) + \varepsilon^2 \left(\frac{1}{6}x^3 - \frac{1}{24}x^4 \right) - O(\varepsilon^3 x^4) \tag{10.682}$$







11. SUITES ET SÉRIES

Les suites et séries ont une très grande importance dans la mathématique appliquée et c'est la raison pour laquelle nous y consacrons un chapitre entier. Nous les retrouverons par ailleurs souvent dans les différents chapitres de la section de Mécanique lorsque nous aurons besoin de faire quelques approximations mineures (...) ainsi qu'en dans les chapitres d'Économie et de Techniques de Gestion. Il conviendra cependant de la part du lecteur de ne pas confondre dans ce qui va suivre le concept de "suite" de celui de "série" qui tout en étant similaires sur le fond ne s'analysent mathématiquement pas toujours de la même manière.

Nous avons souhaité dans ce chapitre rester dans des choses simples sans trop partir dans les concepts topologiques des suites et séries. Cependant, la personne intéressée par des définitions plus rigoureuses pourra se reporter dans le chapitre traitant des Fractales (section d'Informatique Théorique) et de Topologie où de nombreux concepts sur les suites sont définis (supremum, infimum, sous-suite, théorème de Bolzano-Weierstrass, etc.).

1. SUITES

Définition: Une "suite" d'un ensemble est une famille d'éléments indexée par l'ensemble des entiers naturels (cf. chapitre sur les Nombres) ou par une partie de celui-ci. De manière vulgarisée, nous disons qu'une suite est une liste d'objets mis en ordre, chacun ayant un numéro d'ordre. Nous notons classiquement une suite par:

$$(u_n) \text{ ou } (u_n)_{n \in \mathbb{N}} \quad (11.1)$$

où l'indexation se fait parfois (par tradition...) sans le 0.

Pour quelques suites, nous indiquons le premier terme u_1 (si l'indexation commence par 1 au lieu de 0), ainsi qu'une formule pour obtenir n'importe quel terme u_{n+1} à partir du terme précédent u_n quel que soit $n \geq 1$. Nous appelons une telle formulation une "définition récurrente", et la suite est dite définie "par récurrence" (et de même si elle est indexée à partir de 0 au lieu de 1).

Avant de voir quelques exemples de familles de suites qui seront utilisées dans les différents chapitres du site (Dynamiques des populations, Économie, Physique nucléaire, etc.) voyons un petit paquet de définitions comme il est de tradition en mathématique...

Définitions:

D1. Des nombres (en suite) sont en "progression arithmétique" si la différence de deux termes consécutifs est une constante r appelée la "raison".

D2. Des nombres (en suite) sont en "progression géométrique" si le rapport de deux termes consécutifs est une constante r appelée aussi la "raison".

D3. Des nombres (en suite) sont en "progression harmonique" si les inverses de deux termes consécutifs sont en progression arithmétique.

Dès lors, une "**suite**" est arithmétique, géométrique, harmonique si ses termes sont respectivement en progression arithmétique, géométrique, harmonique et b est la moyenne arithmétique, géométrique, harmonique de a et c si les nombres a, b, c sont en progression arithmétique, géométrique, harmonique.

Remarque: Pour les définitions des moyennes citées ci-dessus voir le chapitre de Statistiques

D4. Une "**suite majorée**", est une suite telle qu'il existe un réel M tel que $\forall n \in \mathbb{N}, u_n \leq M$

D5. Une "**suite minorée**", est une suite telle qu'il existe un réel m tel que $\forall n \in \mathbb{N}, u_n \geq m$

D6. Une "**suite bornée**", est une suite telle qu'elle est à la fois majorée et minorée.

D7. Une suite (u_n) est appelée "**suite croissante**" si $\forall n \in \mathbb{N}, u_{n+1} - u_n \geq 0$

D8. Une suite (u_n) est appelée "**suite décroissante**" si $\forall n \in \mathbb{N}, u_{n+1} - u_n \leq 0$

D9. Si une suite est croissante ou décroissante, nous disons qu'elle est "**monotone**".

D10. Une suite (u_n) est appelée "**suite constante**" si $\forall n \in \mathbb{N}, u_{n+1} = u_n$

1.1. SUITES ARITHMÉTIQUES

Définition: Nous disons que des nombres ou que des "**termes**" en progression forment une "**suite arithmétique**" lorsque leurs valeurs numériques diffèrent d'une valeur r appelée la "**raison**" de la suite telle que:

$$u_{k+1} = u_k + r \quad (11.2)$$

où r est donc la "**raison**" de la progression. Nous avons alors bien évidemment si l'indexation commence à partir de 0:

$$u_k = u_0 + k \cdot r \quad (11.3)$$

Ainsi, la suite:

$$n, n+1, n+2, n+3, \dots, n+i \quad (11.4)$$

où n est une constante est une suite arithmétique de raison $r = 1$.

La suite:

$$n+0 \cdot x, n+1 \cdot x, n+2 \cdot x, n+3 \cdot x, \dots, n+i \cdot x \quad (11.5)$$

est une suite arithmétique de raison $r = x$, etc.

Ainsi, si nous notons par u_n un terme quelconque de la suite (u_n) de raison r , nous avons:

$$u_0 = c^{te}, u_1 = u_0 + r, u_2 = u_1 + r = u_0 + 2r, \dots, u_n = u_0 + n \cdot r \quad (11.6)$$

Nous avons les propriétés suivantes pour un tel type de suite:

P1. Un terme dont le rang est la moyenne arithmétique des rangs de deux autres termes est la moyenne arithmétique de ces deux termes.

Démonstration:

Considérons maintenant (u_n) une suite arithmétique de raison r donnée selon le développement précédent:

$$u_n = u_0 + n \cdot r \quad (11.7)$$

et soient $a, b, k \in \mathbb{N}$ tels que $a + b = 2k$, nous avons alors:

$$u_a + u_b = u_0 + (a)r + u_0 + (b)r = 2u_0 + 2\left(\frac{a+b}{2}\right)r = 2\left(u_0 + \left(\frac{a+b}{2}\right)r\right) = 2u_{\frac{a+b}{2}} = 2u_k \quad (11.8)$$

et donc:

$$u_k = \frac{u_a + u_b}{2} \text{ avec } k = \frac{(a+b)}{2} \quad (11.9)$$

□C.Q.F.D.

P2. Pour trois termes consécutifs en progression arithmétique, le deuxième terme est la moyenne arithmétique des deux autres.

Démonstration:

$$u_k = \frac{u_a + u_b}{2} = u_{n+1} = \frac{u_n + u_{n+2}}{2} \text{ avec } k = \frac{(a+b)}{2} = \frac{(n) + (n+2)}{2} = \frac{2n+2}{2} = n+1 \quad (11.10)$$

□C.Q.F.D.

Si $u_1, u_2, u_3, u_4, \dots, u_n, \dots$ est une progression arithmétique de raison r , alors la n -ème somme partielle S'_n (c'est-à-dire, la somme des n premiers termes à la puissance 1) est donnée par:

$$S'_n = \frac{n}{2}[2u_1 + n \cdot r] \text{ ou } S'_n = \frac{n}{2}[u_1 + u_n] \quad (11.11)$$

lorsque l'indexation se fait à partir de 1.

Démonstration:

Nous pouvons écrire la série:

$$\begin{aligned} S'_n &= u_1 + u_2 + u_3 + u_4 + \dots + u_n \\ &= u_1 + (u_1 + r) + (u_1 + 2r) + (u_1 + 3r) + \dots + (u_1 + (n-1) \cdot r) \end{aligned} \quad (11.12)$$

En jouant avec la deuxième ligne, nous obtenons:

$$S_n = (u_1 + u_1 + u_1 + u_1 + \dots + u_1) + (r + 2r + 3r + \dots + (n-1) \cdot r) \quad (11.13)$$

Ce qui se simplifie encore:

$$S_n = nu_1 + r(1 + 2 + 3 + \dots + (n-1)) \quad (11.14)$$

Nous démontrerons quelques lignes plus bas que la série de Gauss simple:

$$1 + 2 + 3 + \dots + n \quad (11.15)$$

est égale à:

$$S_1 = \frac{n(n+1)}{2} \quad (11.16)$$

Nous avons alors in extenso pour:

$$1 + 2 + 3 + \dots + (n-1) \quad (11.17)$$

la relation suivante:

$$S_1 = \frac{(n-1)n}{2} \quad (11.18)$$

Il vient alors:

$$S_n = nu_1 + r \frac{(n-1)n}{2} = \frac{n}{2} [2u_1 + r(n-1)] \quad (11.19)$$

Nous voyons avec cette dernière relation que si $u_1 = r = 1$ nous retombons sur la série de Gauss simple.

Comme:

$$u_n = u_1 + r(n-1) \quad (11.20)$$

lorsque l'indexation se fait à partir de 1. Il vient alors:

$$S_n = \frac{n}{2} [u_1 + u_n] \quad (11.21)$$

□C.Q.F.D.

Nous verrons d'autres types de sommations un peu plus bas lors de notre étude des séries!

1.2. SUITES HARMONIQUES

Définition: Nous disons que des nombres $(1/a, 1/b, 1/c, \dots)$ forment une "**suite harmonique**" lorsque leurs inverses sont en progression arithmétique. Nous représentons cette progression par:

$$\frac{1}{a}, \frac{1}{b}, \frac{1}{c}, \dots, \frac{1}{h}, \frac{1}{k}, \frac{1}{l} \quad (11.22)$$

où $a, b, c, \dots, h, k, l$ désignent des termes au dénominateur en progression arithmétique de raison r . D'ailleurs, nous supposons, dans ce qui suit, qu'il n'y a aucun dénominateur nul.

En partageant cette série en groupes renfermant successivement 2^n termes, nous observons que chacun de ceux-ci est plus grand que le dernier de son groupe:

$$\frac{1}{1} > \frac{1}{2} + \frac{1}{3} > \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} \quad (11.23)$$

et que la somme des termes de chaque groupe est plus grande que $1/2$. La somme des termes de la série augmente donc indéfiniment; nous disons alors que la série est une "**série divergente**" (nous reviendrons plus en détail sur ces concepts de convergence et divergence plus bas).

1.3. SUITES GÉOMÉTRIQUES

Définition: Une "**suite géométrique**" est une suite de nombres tels que chacun d'eux est égal au précédent n multiplié par un nombre constant q que nous appelons la "**raison**" de la progression. Nous désignerons par:

$$n, q \cdot n, q^2 \cdot n, q^3 \cdot n, \dots \quad (11.24)$$

Ainsi, si nous notons par u_n un terme quelconque de la suite (u_n) , nous avons (trivial):

$$u_n = u_0 \cdot q^n \quad (11.25)$$

Voici quelques propriétés pour un tel type de suite (sans démonstration pour l'instant... sauf demande car triviales pour la plupart):

P1. (triviale) Le quotient de deux termes d'une même suite est une puissance de la raison dont l'exposant égale la différence des rangs des deux termes (simple rapport de termes de puissance).

P2. (triviale) Si nous multiplions ou divisons terme à terme deux suites géométriques, nous obtenons une troisième suite géométrique dont la raison égale le produit (respectivement le quotient) des raisons des progressions données (simple opération avec les raisons des deux séries d'origine).

P3. Dans une suite géométrique, un terme dont le rang est la moyenne arithmétique des rangs de deux autres termes est la moyenne géométrique (cf. chapitre de Statistiques) de ces deux termes (relisez plusieurs fois au besoin).

Démonstration:

Soit une suite géométrique réelle positive de raison q , nous avons:

$$u_n = u_0 q^n \quad (11.26)$$

Soient a, b deux termes de la suite géométrique, nous avons alors:

$$u_a \cdot u_b = u_0^2 \cdot q^a \cdot q^b = \left(u_0 q^{\frac{a+b}{2}} \right)^2 = u_{\frac{a+b}{2}}^2 = u_k^2 \quad (11.27)$$

et ainsi:

$$u_k = \sqrt{u_a \cdot u_b} \quad (11.28)$$

□C.Q.F.D.

Nous avons comme corolaire que pour trois termes consécutifs en progression géométrique, le deuxième terme est la moyenne géométrique des deux autres.

Démonstration:

$$u_k = \sqrt{u_a \cdot u_b} = u_{n+1} = \sqrt{u_n \cdot u_{n+2}} \quad (11.29)$$

avec:

$$k = \frac{(a+b)}{2} = \frac{(n) + (n+2)}{2} = \frac{2n+2}{2} = n+1 \quad (11.30)$$

□C.Q.F.D.

Il existe cependant quelques suites particulières qui ont des propriétés particulières que nous retrouvons très fréquemment en mathématique ou physique théorique. Sans trop entrer dans les détails, voici une petite liste (non exhaustive de ces dernières):

1.4. SUITE DE CAUCHY

Il est souvent intéressant pour le mathématicien, autant que pour le physicien, de connaître les propriétés d'une suite ayant un type de progression donnée. La propriété la plus importante étant la limite vers laquelle elle tend.

Remarque: Le lecteur qui n'est pas à l'aise avec la topologie peut sauter le texte qui va suivre en attendant... et celui qui souhaite en savoir plus sur les suites de Cauchy peut se reporter au chapitre de Topologie et particulièrement au chapitre consacré aux Fractales (section d'Informatique Théorique).

Définition: Soit (X, d) un espace métrique (cf. chapitre de Topologie), nous disons que la suite:

$$(x_n)_{n \in \mathbb{N}}, x_n \in X \quad \forall n \in \mathbb{N} \quad (11.31)$$

converge vers $x \in X$ si par définition:

$$\forall \varepsilon > 0, \exists N \in \mathbb{N} \text{ tel que } n \geq N \Rightarrow d(x_n, x) \leq \varepsilon \quad (11.32)$$

En d'autres termes plus nous avançons dans la suite, plus les points sont proches (au sens de la métrique d) les uns des autres.

Cependant la définition précédente de la convergence pose problème car la limite x doit être connue. Dans la plupart des cas intéressants, x est malheureusement inconnue. Pour sortir de cette impasse, Cauchy a l'idée de proposer la définition suivante:

Nous disons par définition que la suite $(x_n)_{n \in \mathbb{N}}$ d'éléments de X est une "**suite de Cauchy**" si:

$$\forall \varepsilon > 0, \exists N \in \mathbb{N} \text{ tel que } \forall n, m \geq N \text{ on a } d(x_n, x_m) \leq \varepsilon \quad (11.33)$$

Il est clair alors que toute suite convergente est une suite de Cauchy (bon il y a quelques subtilités auxquelles nous ne ferons pas référence pour l'instant).

Remarque: Ce critère facilite certaines démonstrations car il permet de montrer l'existence d'une limite sans faire intervenir sa valeur, en général inconnue.

Maintenant, montrons qu'une suite convergente est de Cauchy.

Démonstration:

Soit une suite u_n convergeant vers l (qui nous est inconnu donc!) et $\varepsilon > 0$ (choisi au hasard). Il existe alors selon la définition d'une suite convergente, $N \in \mathbb{N}$ tel que:

$$n \geq N \Rightarrow d(u_n, l) \leq \frac{\varepsilon}{2} \quad (11.34)$$

le choix d'écrire $\varepsilon/2$ est complètement arbitraire mais au fait nous anticipons juste le résultat de la démonstration afin que celui-ci soit plus esthétique.

Alors pour $p, q \geq N$ (au fait connaître le N en question importe peu puisque cela doit marcher pour n'importe lequel... bon n'oublions pas quand même que N dépend de $\varepsilon/2$) nous avons selon l'inégalité triangulaire (cf. chapitre de Calcul Vectoriel):

$$d(u_p, u_q) \leq d(u_p, l) + d(l, u_q) \quad (11.35)$$

et puisque $d(u_n, l) \leq \frac{\varepsilon}{2}$:

$$d(u_p, u_q) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \quad (11.36)$$

ce qui revient à écrire:

$$d(u_p, u_q) \leq \varepsilon \quad (11.37)$$

C'est peut-être un peu abstrait alors voyons un exemple avec la suite harmonique (divergente comme nous le savons déjà) $u_n = \sum_{k=1}^n \frac{1}{k}$. D'abord, rien ne nous interdit de prendre $n \geq 2$ (sinon cela va être dur de faire une différence entre deux termes...).

Dès lors nous prenons la distance euclidienne:

$$d(u_p, u_q) = d(u_{2n}, u_n) = u_{2n} - u_n = \sum_{k=1}^{2n} \frac{1}{k} - \sum_{k=1}^n \frac{1}{k} = \sum_{k=n+1}^{2n} \frac{1}{k} \quad (11.38)$$

D'abord le lecteur remarquera que dans tous les cas $k \leq 2n$ puisque compris entre $n+1$ et $2n$. Ce qui nous amène à pouvoir écrire:

$$\frac{1}{k} \geq \frac{1}{2n} \quad \forall n+1 \leq k \leq 2n \quad (11.39)$$

Donc à partir de cette égalité il vient automatiquement que chaque terme de la somme de gauche ci-dessous sera plus grand que chaque terme de la somme de droite suivant:

$$\sum_{k=n+1}^{2n} \frac{1}{k} \geq \sum_{k=n+1}^{2n} \frac{1}{2n} \quad \text{avec} \quad \sum_{k=n+1}^{2n} \frac{1}{2n} = \frac{1}{2} \quad (11.40)$$

maintenant l'idée est de voir que la somme de gauche est donc plus grande ou égale à $\varepsilon = 1/2$ et cela quel que soit n . Ainsi, l'idée c'est que nous ayons trouvé un epsilon pour lequel le critère de Cauchy est mis en défaut. Car dans le cas contraire nous aurions dû avoir:

$$\sum_{k=n+1}^{2n} \frac{1}{k} \leq \frac{1}{2} \quad (11.41)$$

donc la suite n'est pas convergente.

□C.Q.F.D.

Donc, ce n'est pas parce que des points se rapprochent les uns des autres qu'ils convergent vers un point, car ce point n'existe peut-être pas.

Exemple:

Le meilleur exemple est certainement le suivant:

Prenons $X = \mathbb{Q}$ et:

$$d(p, q) = |p - q| \quad (11.42)$$

Soit z un nombre irrationnel et $q_j \in \mathbb{Q}$, avec $|z - q_j| \leq 1/j$.

Les $q_1, q_2, \dots$ forment une suite de Cauchy. En effet:

$$|q_m - q_n| \leq |z - q_m| + |z - q_n| \leq \frac{1}{m} + \frac{1}{n} \quad (11.43)$$

et donc $|q_m - q_n| \leq \varepsilon$ si $m, n \geq 1/2\varepsilon$. Nous avons donc trouvé un N qui satisfait à notre définition d'une suite de Cauchy. Or cette suite ne converge pas dans $\mathbb{Q}$ sinon z serait rationnel.

Remarque: Les mathématiciens utilisent ce fait pour définir l'ensemble des irrationnels en utilisant quelques concepts topologique supplémentaires.

Nous venons de voir qu'une suite de Cauchy n'est pas forcément une suite convergente dans X . La réciproque toutefois est vraie: toute suite convergente est une suite de Cauchy.

1.5. SUITE DE FIBONACCI

Si nous calculons une suite de nombres commençant par 0 et 1, de telle sorte que chaque terme soit égal à la somme des deux précédents, nous pouvons former la suite:

$$0, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144... \quad (11.44)$$

par conséquent, si nous désignons les différents termes par:

$$u_0, u_1, u_2, \dots \quad (11.45)$$

nous avons la loi de formation:

$$u_{n+2} = u_{n+1} + u_n \quad (11.46)$$

La suite de Fibonacci possède des propriétés nombreuses fortes intéressantes, qui seront développées ultérieurement. Il s'agit cependant de la première "**suite récurrente**" connue (d'où le fait que nous en parlions sur ce site).

L'origine de cette suite viendrait d'un problème de lapins posé à Fibonacci en 1202. Partant d'un couple, combien de couples de lapins obtiendrons-nous après un nombre donné de mois sachant que chaque couple produit chaque mois un nouveau couple, lequel ne devient productif qu'après deux mois. Nous avons alors:

- Début: Un couple de bébés lapins qui vont grandir
 - Premier mois: Un couple de lapins adultes (qui feront des bébés le mois prochain...)
 - Deuxième mois: Un couple de lapins adultes et un couple de bébés donc 2 couples
 - Troisième mois: Deux couples de lapins adultes et un couple de bébés donc 3 couples
 - Quatrième mois: Trois couples de lapins adultes et deux couples de bébés donc 5 couples.
- etc.

Prenons un exemple réel, cette fois-ci: le coeur de certaines fleurs, les écailles d'un ananas ou d'une pomme de pin forment deux familles de spirales enroulées en sens inverse. Sur une pomme de pin, vous compterez 5 spirales dans un sens et 8 dans l'autre, sur l'ananas, 8 et 13, sur la fleur de tournesol 21 et 34. Chaque fois, nous obtenons des nombres de Fibonacci !

Une illustration de ceci consiste à faire le simple schéma suivant (appelé "**spirale de Fibonacci**") qui reproduit les nombres de Fibonacci sur un plan quadrillé:

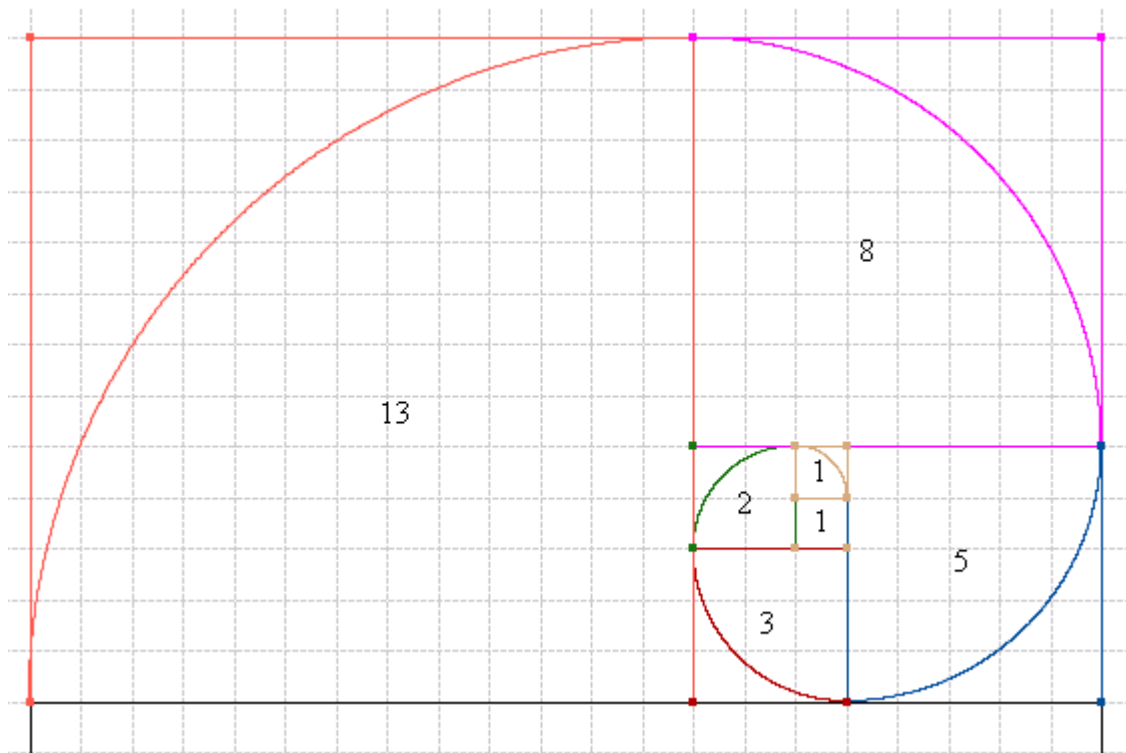


Figure: 11.1 - Spirale de Fibonacci

Nous utilisons également ce genre de suite pour montrer l'utilité du principe d'induction présenté dans le chapitre de Théorie Des Nombres se trouvant dans la section d'Arithmétique.

2. SÉRIES

Le physicien a souvent besoin pour résoudre simplement et formellement des problèmes, d'approximer certains "**termes**" (cf. chapitre de Théorie De La Démonstration) de ses équations. Pour cela, il utilisera les propriétés de certaines séries.

Il existe, une quantité phénoménale de séries et de théories gravitant autour de ces dernières, mais nous citerons en particulier les séries de Taylor (utilisées un peu partout), les séries de Fourier (théorie du signal et en mécanique ondulatoire) et les séries ou fonctions de Bessel (physique nucléaire) dont nous ferons une étude sommaire ici.

Définition: Soit donnée une suite numérique infinie:

$$a_1, a_2, a_3, \dots, a_n, \dots \quad (11.47)$$

L'expression:

$$a_1 + a_2 + a_3 + \dots + a_n + \dots \quad (11.48)$$

est appelée "**série numérique**".

Définition: La somme partielle des n premiers termes de la série est appelée "**somme partielle**" et notée s_n :

$$s_n = a_1 + a_2 + a_3 + \dots + a_n \quad (11.49)$$

Si la limite notée S suivante existe et est finie:

$$S = \lim_{n \rightarrow \infty} s_n \quad (11.50)$$

nous l'appelons la "**somme de la série**" et nous disons que la "**série converge**" (elle est donc de Cauchy). Cependant, si la limite n'existe pas, nous disons que la "**série diverge**" et n'a pas de somme (pour plus de détails voir le sous-chapitre plus loin traitant des critères de convergence).

Montrons par ailleurs que si $\sum_{n \geq 0} u_n$ est une série numérique convergente alors:

$$\lim_{n \rightarrow +\infty} u_n = 0 \quad (11.51)$$

Démonstration:

Nous supposons d'abord que $\sum_{k \geq 0} u_k$ est bien une série convergente et notons par S sa limite. Posons:

$$s_n = \sum_{k=0}^n u_k \quad s_{n-1} = \sum_{k=0}^{n-1} u_k, \quad (11.52)$$

Alors:

$$s_n - s_{n-1} = u_n \quad (11.53)$$

Or, si la série est convergente:

$$\lim_{n \rightarrow +\infty} s_n = \lim_{n \rightarrow +\infty} s_{n-1} = S \quad (11.54)$$

Donc:

$$\lim_{n \rightarrow +\infty} s_n - \lim_{n \rightarrow +\infty} s_{n-1} = \lim_{n \rightarrow +\infty} (s_n - s_{n-1}) = \lim_{n \rightarrow +\infty} u_n = S - S = 0 \quad (11.55)$$

□C.Q.F.D.

Voyons comment calculer la somme partielle des quelques séries classiques:

2.1. SÉRIES DE GAUSS

Les séries arithmétiques de Gauss sont l'expression de la somme de n premiers entiers non nuls élevés à une puissance donnée sous une forme condensée. L'application de cette forme condensée de série a une utilité pratique en physique (voir les chapitres y relatifs) lorsque l'on souhaite simplifier l'expression de certains résultats ainsi que dans les chapitre de Statistiques lors de notre étude des statistiques non paramétriques.

Gauss avait trouvé une méthode séduisante en 1786 pour déterminer cette expression lorsqu'il avait 9 ans (...):

$$\begin{aligned}
 S_1 &= 1 + 2 + 3 + \dots + (n-1) + n \\
 + \frac{S_1 &= n + (n-1) + (n-2) + \dots + 2 + 1}{2S_1 = (n+1) + (n+1) + (n+1) + \dots + (n+1) + (n+1)} \quad (11.56)
 \end{aligned}$$

En simplifiant, nous trouvons facilement:

$$\boxed{S_1 = \frac{n(n+1)}{2}} \quad (11.57)$$

pour $n \geq 0$. Indiquons que chaque somme intermédiaire de la série (1, 3, 6, 10, 15, etc.) est appelée "**nombre triangulaire**" puisqu'il est possible de le représenter sous la forme suivante:

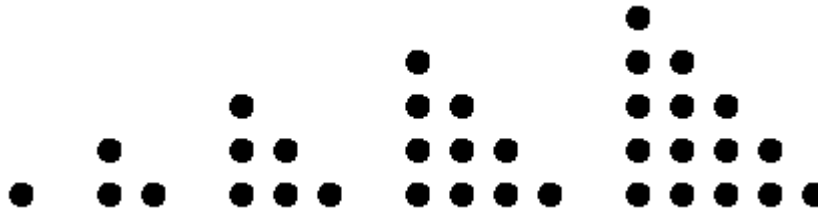


Figure: 11.2 - Nombres triangulaires 1, 3, 6, 10 et 15

Nous pouvons continuer ainsi pour des ordres supérieurs (nous les présentons non en tant qu'exercices mais parce que ces relations sont utiles!):

Calculons maintenant le cas très important que nous retrouverons dans un certain nombre d'autres chapitres (Économie, Physique Quantique Ondulatoire, etc.) et qui est la somme des n premiers carrés (toujours non nuls).

Posons pour cela:

$$\begin{aligned}
 S_1 &= 1 + 2 + 3 + \dots + n \\
 S_2 &= 1^2 + 2^2 + \dots + n^2 \quad (11.58)
 \end{aligned}$$

nous savons que (binôme de Newton):

$$(k+1)^3 - k^3 = 3k^2 + 3k + 1 \quad (11.59)$$

nous pouvons donc écrire et ajouter membre à membre les n égalités suivantes:

$$\begin{aligned}
 2^3 - 1^3 &= 3 \cdot 1^2 + 3 \cdot 1 + 1 \\
 3^3 - 2^3 &= 3 \cdot 2^2 + 3 \cdot 2 + 1 \\
 4^3 - 3^3 &= 3 \cdot 3^2 + 3 \cdot 3 + 1 \\
 &\dots \\
 (n+1)^3 - n^3 &= 3 \cdot n^2 + 3 \cdot n + 1 \\
 \hline
 (n+1)^3 - 1 &= 3 \cdot (1^2 + 2^2 + \dots + n^2) + 3 \cdot (1 + 2 + \dots + n) + n \\
 (n+1)^3 - 1 &= 3 \cdot S_2 + 3 \cdot S_1 + n \quad (11.60)
 \end{aligned}$$

Avec quelques manipulations algébriques élémentaires:

$$\begin{aligned}
 3S_2 &= (n+1)^3 - 1 - 3S_1 - n = (n+1)^3 - 1 - 3 \frac{n(n+1)}{2} - n \\
 &= (n+1)^3 - (n+1) - 3 \frac{n(n+1)}{2} = \frac{(n+1)}{2} [2(n+1)^2 - 2 - 3n] \quad (11.61)
 \end{aligned}$$

d'où:

$$S_2 = \frac{(n+1)}{6} [2n^2 + 4n + 2 - 2 - 3n] = \frac{(n+1)}{6} (2n^2 + n) \quad (11.62)$$

Finalement:

$$\boxed{S_2 = \frac{n(n+1)(2n+1)}{6}} \quad (11.63)$$

Terminons avec la somme des n premiers cubes (non nuls). Le principe étant le même que précédemment, nous posons:

$$\begin{aligned}
 S_1 &= 1 + 2 + \dots + n \\
 S_2 &= 1^2 + 2^2 + \dots + n^2 \\
 S_3 &= 1^3 + 2^3 + \dots + n^3
 \end{aligned} \quad (11.64)$$

Nous savons par ailleurs que (binôme de Newton):

$$(k+1)^4 - k^4 = 4k^3 + 6k^2 + 4k + 1 \quad (11.65)$$

Nous obtenons en faisant varier k de 1 à n , n relations que nous pouvons ajouter membre à membre:

$$\begin{aligned}
 2^4 - 1^4 &= 4 \cdot 1^3 + 6 \cdot 1^2 + 4 \cdot 1 + 1 \\
 3^4 - 2^4 &= 4 \cdot 2^3 + 6 \cdot 2^2 + 4 \cdot 2 + 1 \\
 &\dots \\
 (n+1)^4 - n^4 &= 4 \cdot n^3 + 6 \cdot n^2 + 4 \cdot n + 1 \\
 \hline
 (n+1)^4 - 1 &= 4 \cdot S_3 + 6 \cdot S_2 + 4 \cdot S_1 + n
 \end{aligned} \quad (11.66)$$

Nous avons donc:

$$(n+1)^4 - 1^4 = 4 \cdot S_3 + 6 \cdot S_2 + 4 \cdot S_1 + n \quad (11.67)$$

Ce qui donne après développement:

$$(n+1)^4 - 1^4 = 4 \cdot S_3 + 6 \cdot \frac{n(n+1)(2n+1)}{6} + 4 \cdot \frac{n(n+1)}{2} + n \quad (11.68)$$

Et après une première simplification:

$$(n+1)^4 - 1 = 4 \cdot S_3 + n(n+1)(2n+1) + 2n(n+1) + n \quad (11.69)$$

et une deuxième:

$$\begin{aligned}
4 \cdot S_3 &= -n(n+1)(2n+1) - 2n(n+1) - n + (n+1)^4 - 1 \\
&= -\left((n^2+n)(2n+1)\right) - (2n^2+2n) - n + (n^4+4n^3+6n^2+4n+1) - 1 \\
&= -(2n^3+n^2+2n^2+n) + n^4+4n^3+4n^2+n \\
&= n^4+2n^3+n^2 = n^2(n^2+2n+1) = n^2(n+1)^2
\end{aligned} \tag{11.70}$$

Le résultat final est donc:

$$S_3 = \left(\frac{n(n+1)}{2} \right)^2 \tag{11.71}$$

ou écrit autrement:

$$\boxed{S_3 = \frac{n^2(n+1)^2}{4}} \tag{11.72}$$

Évidemment, nous pouvons continuer ainsi longtemps mais à partir d'une certaine valeur de l'élévation de la puissance les choses se compliquent un petit peu (de plus, la méthode est un peu longue). Ainsi, un des membres de la famille des Bernoulli (c'était une famille de mathématiciens assez doués...) a montré une relation générale fonctionnant pour n'importe quelle puissance en définissant ce que nous appelons le "polynôme de Bernoulli".

Terminons avec un dernier cas particulier dont nous aurons besoin lors de notre étude des séries de Fourier. Nous posons:

$$S'_n = x + x^3 + x^5 + \dots + x^{2n-1} \tag{11.73}$$

Nous voulons exprimer cette expression sous forme de fraction rationnelle. Pour ce faire, nous multiplions tout par x^2 . Nous avons donc les deux expressions:

$$\begin{aligned}
S'_n &= x + x^3 + x^5 + \dots + x^{2n-1} \\
x^2 S'_n &= x^3 + x^5 + \dots + x^{2n+1}
\end{aligned} \tag{11.74}$$

Nous soustrayons la première de la deuxième:

$$S'_n(x^2 - 1) = x^{2n+1} - x \tag{11.75}$$

Finalement:

$$S'_n = \frac{x^{2n+1} - x}{x^2 - 1} \tag{11.76}$$

2.1.1. NOMBRES ET POLYNÔMES DE BERNOULLI

Comme nous venons de le voir plus haut il est possible d'exprimer la somme des n premiers entiers non nuls élevés à une puissance donnée selon (les quatre premiers ont été démontrés précédemment) les relations suivantes où nous avons posé $n = n' + 1$ avec n' le nombre de termes dont nous voulons la somme 0 non compris (d'où le signe négatif que nous n'avions pas plus haut):

$$\begin{aligned}
 S_0 &= n \\
 S_1 &= \frac{1}{2}n^2 - \frac{1}{2}n = \frac{n(n-1)}{2} \\
 S_2 &= \frac{1}{3}n^3 - \frac{1}{2}n^2 + \frac{1}{6}n = \frac{n(n-1)(2n-1)}{6} \\
 S_3 &= \frac{1}{4}n^4 - \frac{1}{2}n^3 + \frac{1}{4}n^2 = \frac{n^2(n-1)^2}{4} \\
 S_4 &= \frac{1}{5}n^5 - \frac{1}{2}n^4 + \frac{1}{3}n^3 - \frac{1}{30}n = \frac{n(n-1)(2n-1)(3n^2-3n-1)}{30} \\
 S_5 &= \frac{1}{6}n^6 - \frac{1}{2}n^5 + \frac{5}{12}n^4 - \frac{1}{12}n^2 = \frac{n^2(2n^2-2n-1)(n-1)^2}{12} \\
 &\dots
 \end{aligned} \tag{11.77}$$

Jacob Bernoulli remarqua ensuite que les polynômes S_p avaient la forme:

$$S_p = \frac{1}{p+1}n^{p+1} - \frac{1}{2}n^p + \frac{p}{12}n^{p-1} + 0 \cdot n^{p-2} + \dots \tag{11.78}$$

Dans cette expression, les nombres $(1, -1/2, 1/12, 0, \dots)$ semblent ne pas dépendre de p . Plus généralement, après tâtonnement on remarque que le polynôme peut être écrit sous la forme:

$$\begin{aligned}
 S_p &= \sum_{k=0}^p \frac{B_k}{k!} \frac{p!}{(p+1-k)!} n^{p+1-k} \\
 &= \frac{B_0}{0!} \frac{n^{p+1}}{p+1} + \frac{B_1}{1!} n^p + \frac{B_2}{2!} p n^{p-1} + \frac{B_3}{3!} p(p-1) n^{p-2} + \dots + \frac{B_p}{p!} n
 \end{aligned} \tag{11.79}$$

Ce qui donne par identification les "**nombre de Bernoulli**":

$$B_0 = 1, B_1 = -1/2, B_2 = 1/6, B_3 = 0, \dots \tag{11.80}$$

Par la suite, les mathématiciens dans leurs recherches sont tombés au hasard sur le fait que les nombres de Bernoulli pouvaient être exprimés par la série:

$$G(z) = \frac{z}{e^z - 1} = \sum_{k=0}^{\infty} B_k \frac{z^k}{k!} \text{ avec } |z| < 2\pi \tag{11.81}$$

En d'autres termes, la fonction génératrice des nombres de Bernoulli serait $G(z)$. Si nous développons les premiers termes de cette série:

$$G(z) = \left(\frac{e^z - 1}{z} \right)^{-1} = \left(1 + \frac{z}{2!} + \frac{z^2}{3!} + \frac{z^3}{4!} + \dots \right)^{-1} = 1 - \frac{z}{2} + \frac{z^2}{12} - \frac{z^4}{720} + \frac{z^6}{30240} - \frac{z^8}{1209600} + \dots \quad (11.82)$$

Démonstration:

Nous avons vu dans notre étude des nombres complexes (cf. chapitre sur les Nombres) que:

$$e^z = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \dots \quad (11.83)$$

Dès lors:

$$G(z) = \left(\frac{1 + \frac{z}{2!} + \frac{z^2}{3!} + \dots}{z} \right)^{-1} = \left(1 + \frac{z}{2!} + \frac{z^2}{3!} + \dots \right)^{-1} \quad (11.84)$$

Posons maintenant:

$$G(z) = a_0 + a_1 z + a_2 z^2 + \dots \quad (11.85)$$

Nous avons alors:

$$\left(1 + \frac{z}{2!} + \frac{z^2}{3!} + \dots \right) \cdot (a_0 + a_1 z + a_2 z^2 + \dots) = 1 \quad (11.86)$$

Nous voyons (en distribuant) que:

$$\left(1 + \frac{z}{2!} + \frac{z^2}{3!} + \dots \right) \cdot (a_0 + a_1 z + a_2 z^2 + \dots) = a_0 + \left(a_1 + \frac{a_0}{2!} \right) z + \left(a_2 + \frac{a_1}{2!} + \frac{a_0}{3!} \right) z^2 + \dots \quad (11.87)$$

par suite pour que tout cela soit égal à l'unité il faut que:

$$\begin{aligned} a_0 &= 1 \\ a_1 + \frac{a_0}{2!} &= 0 \\ a_2 + \frac{a_1}{2!} + \frac{a_0}{3!} &= 0 \\ &\dots \end{aligned} \quad (11.88)$$

De la deuxième équation nous tirons:

$$a_1 = -\frac{1}{2} \quad (11.89)$$

De la troisième équation nous tirons:

$$a_2 = -\frac{1}{6} + \frac{1}{4} = \frac{1}{12} \quad (11.90)$$

etc.

En continuant ainsi nous montrons que:

$$a_3 = 0, a_4 = -\frac{1}{720} \dots \quad (11.91)$$

Il est évident que cette méthode ne nous permet de calculer à la main que les premiers termes de cette série.

Ainsi, en se basant sur:

$$G(z) = \frac{z}{e^z - 1} = \sum_{k=0}^{\infty} B_k \frac{z^k}{k!} \quad (11.92)$$

nous trouvons que les premiers nombres de Bernoulli sont les suivants:

k	B_k
0	1
1	-1/2
2	1/6
3	0
4	-1/30
5	0
6	1/42
7	0
8	-1/30
9	0
10	5/66
11	0
12	-691/2730
13	0
14	7/6

Tableau: 11.1 - Nombres de Bernoulli

Le lecteur aura remarqué que $B_k = 0$ lorsque n est impair et différent de 1.

□C.Q.F.D.

Nous voyons bien par ailleurs, que les valeurs des nombres de Bernoulli ne peuvent pas être décrites simplement. En fait, ce sont essentiellement des valeurs de la fonction ζ de Riemann (voir plus bas) pour des valeurs entières négatives de la variable, et ces nombres sont associés à des propriétés théoriques profondes qui dépassent le cadre d'étude de ce site. Par ailleurs, les nombres de Bernoulli apparaissent également dans le développement en série de Taylor des fonctions tangentes circulaire et hyperbolique, dans la formule d'Euler-Maclaurin (voir plus bas).

Avec une petite modification, il est possible de définir les "**polynômes de Bernoulli**" $B_k(x)$ par:

$$G(z, x) = \frac{ze^{zx}}{e^z - 1} = \sum_{k=0}^{\infty} B_k(x) \frac{z^k}{k!} \quad (11.93)$$

avec donc:

$$G(z, 0) = G(z) \quad (11.94)$$

Par ailleurs, il est aisé de remarquer que:

$$\frac{\partial G(z, x)}{\partial x} = z \frac{ze^{zx}}{e^z - 1} = zG(z, x) = \sum_{k=0}^{\infty} \frac{dB_k(x)}{dx} \frac{z^k}{k!} \quad (11.95)$$

et donc il est facile d'en déduire:

$$\sum_{k=0}^{\infty} \frac{dB_k(x)}{dx} \frac{z^k}{k!} = \sum_{k=1}^{\infty} B_{k-1}(x) \frac{z^k}{(k-1)!} \quad (11.96)$$

Démonstration:

D'un côté nous avons:

$$zG(z, x) = z \sum_{k=0}^{\infty} B_k(x) \frac{z^k}{k!} = \sum_{k=0}^{\infty} B_k(x) \frac{z^{k+1}}{k!} = \sum_{k=1}^{\infty} B_{k-1}(x) \frac{z^k}{(k-1)!} \quad (11.97)$$

et d'un autre nous avons:

$$zG(z, x) = \sum_{k=0}^{\infty} \frac{dB_k(x)}{dx} \frac{z^k}{k!} \quad (11.98)$$

Donc:

$$\sum_{k=0}^{\infty} \frac{dB_k(x)}{dx} \frac{z^k}{k!} = \sum_{k=1}^{\infty} B_{k-1}(x) \frac{z^k}{(k-1)!} \quad (11.99)$$

□C.Q.F.D.

Et par identification des coefficients nous en déduisons:

$$\frac{dB_0(x)}{dx} = 0 \quad (11.100)$$

et pour $k \geq 1$:

$$\frac{dB_k(x)}{dx} = k \cdot B_{k-1}(x) \quad (11.101)$$

Il est alors aisé de déduire que les $B_k(x)$ sont des polynômes de degré k :

$$\begin{aligned} B_0(x) &= 1 \\ B_1(x) &= x - \frac{1}{2} \\ B_2(x) &= x^2 - x + \frac{1}{6} \\ B_3(x) &= x^3 - \frac{3}{2}x^2 + \frac{1}{2}x \\ B_4(x) &= x^4 - 2x^3 + x^2 - \frac{1}{30} \\ B_5(x) &= x^5 - \frac{5}{2}x^4 + \frac{5}{3}x^3 - \frac{1}{6}x \\ &\dots \end{aligned} \quad (11.102)$$

Voici un tracé de ces polynômes:

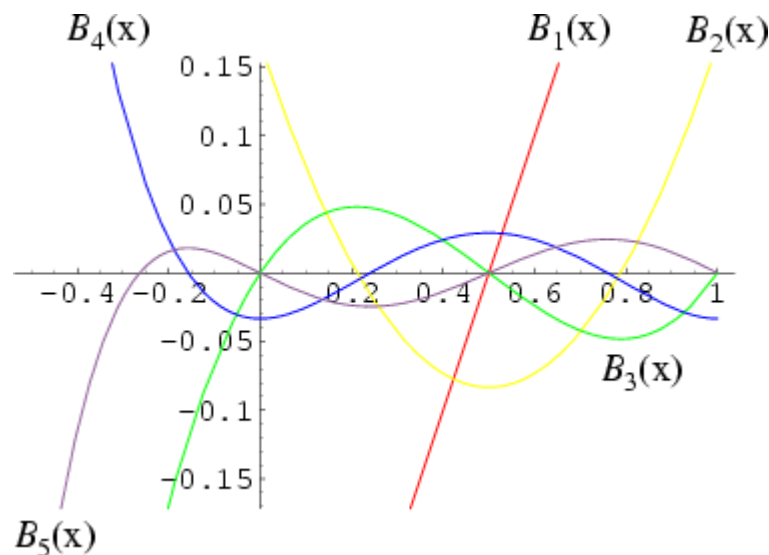


Figure: 11.3 - Quelques polynômes de Bernoulli

Ce qui est remarquable c'est qu'à l'aide des polynômes de Bernoulli, nous voyons qu'il est possible d'écrire les S_n sous la forme suivante:

$$S_p = \sum_{k=0}^{n-1} k^p = \frac{1}{p+1} (B_{p+1}(n) - B_{p+1}) \quad (11.103)$$

Certains écrivent cette relation encore autrement. Effectivement, de la relation précédente, nous pouvons écrire:

$$(p+1) \sum_{k=0}^{n-1} k^p = B_{p+1}(n) - B_{p+1}(0) \quad (11.104)$$

Et en utilisant:

$$\frac{dB_k(x)}{dx} = k \cdot B_{k-1}(x) \quad (11.105)$$

Il vient:

$$\begin{aligned} \frac{dB_{p+1}(x)}{dx} &= (p+1) \cdot B_p(x) \\ \Rightarrow B_p(x) &= \frac{1}{p+1} B'_{p+1}(x) \\ \Rightarrow \int_0^k B_p(x) dx &= \frac{1}{p+1} \int_0^k B'_{p+1}(x) dx = \frac{1}{p+1} B_{p+1}(x) \Big|_0^k = \frac{1}{p+1} [B_{p+1}(x) - B_{p+1}(0)] \end{aligned} \quad (11.106)$$

Donc nous venons de démontrer:

$$\sum_{k=0}^{n-1} k^p = \int_0^k B_p(x) dx \quad (11.107)$$

Cependant, nous pouvons maintenant nous demander ce qu'il advient de la somme partielle de suites arithmétiques et géométriques telles que présentées au début de ce chapitre.

2.2. SÉRIES ARITHMÉTIQUES

Nous avons démontré plus haut que la somme partielle de la série de Gauss (analogue à la somme des termes d'une suite arithmétique de raison $r=1$) s'écrivait donc:

$$S_1 = \frac{n(n+1)}{2} \quad (11.108)$$

si nous notons non pas n la valeur du n -ème terme mais u_n , le développement que nous avons fait pour la série de Gauss nous amène alors à:

$$S_n = \frac{n(u_n + 1)}{2} \quad (11.109)$$

et si nous notons le premier terme 1 de la série de Gauss par u_0 , nous avons alors:

$$\boxed{S_n = \frac{n(u_n + u_0)}{2}} \quad (11.110)$$

ce qui nous donne la somme partielle des n -termes d'une suite arithmétique de raison r quelconque (ou plus simplement: la somme partielle de la série arithmétique de raison r).

Remarque: Le lecteur aura observé que la raison r n'apparaît pas dans la relation. Effectivement, en reprenant (toujours) le même développement fait que pour la série de Gauss, le terme r se simplifie.

2.3. SÉRIES GÉOMÉTRIQUES

De même, avec un somme géométrique où nous avons pour rappel:

$$u_n = u_0 q^n \quad (11.111)$$

nous avons donc:

$$\begin{aligned} S_n &= u_0 + u_1 + \dots + u_{n-1} + u_n \\ qS_n &= u_1 + u_2 + \dots + u_n + u_{n+1} \end{aligned} \quad (11.112)$$

$$S_n - qS_n = u_0 + u_1 - u_1 + u_2 - u_2 + \dots + u_n - u_n - u_{n+1}$$

La dernière relation s'écrit (après simplification):

$$S_n - qS_n = (1 - q)S_n = u_0 - u_{n+1} \quad (11.113)$$

et si $q \neq 1$, nous avons:

$$S_n = \frac{u_0 - u_{n+1}}{1 - q} \quad (11.114)$$

ce qui peut s'écrire en factorisant u_0 :

$$S_n = u_0 \frac{1 - q^{n+1}}{1 - q} \quad (11.115)$$

Si q est positif et inférieur à 1, lorsque n tend vers l'infini nous avons le résultat qui sera très utilisé dans le chapitre d'Économie:

$$S_{n \rightarrow +\infty} = u_0 \frac{1}{1 - q} \quad (11.116)$$

Exemple:

Soit la suite de raison $q=2$ suivante:

$$1 \cdot 2^0 + 1 \cdot 2^1 + 1 \cdot 2^2 + 1 \cdot 2^3 + \dots + 1 \cdot 2^n = 1 + 2 + 4 + 8 + \dots + 2^n \quad (11.117)$$

pour calculer la somme des quatre premiers termes $\{1, 2, 4, 8\}$, nous prenons la puissance de 2 équivalent de $n = 3$ (le zéro n'étant pas pris en compte). Nous obtenons alors bien $S_3 = 15$.

2.3.1. FONCTION ZÊTA ET IDENTITÉ D'EULER

L'allemand Riemann a baptisé "zêta" une fonction déjà étudiée avant lui, mais qu'il examine lorsque la valeur est un nombre complexe (cf. chapitre sur les Nombres). Cette fonction se présente comme une série de puissances inverses de nombres entiers. C'est la série:

$$\zeta(s) = \frac{1}{1} + \frac{1}{2^s} + \frac{1}{3^s} + \dots \quad (11.118)$$

Remarque: Il est traditionnel de noter s la variable dont dépend cette série.

Cette série a une propriété intéressante mais si l'on reste dans le cadre des puissances entières positives et non nulles:

$$\frac{1}{x^0} + \frac{1}{x^1} + \dots + \frac{1}{x^n} + \dots = \mathcal{U}_0 \frac{(1 - q^{n+1})}{(1 - q)} = 1 \frac{1 - \left(\frac{1}{x^{n+1}}\right)}{1 - \frac{1}{x}} \quad (11.119)$$

quand $n \rightarrow +\infty$ nous avons alors:

$$\sum_{n=0}^{+\infty} \frac{1}{x^n} = \frac{1}{x^0} + \frac{1}{x^1} + \dots + \frac{1}{x^n} + \dots = \frac{1}{1 - \frac{1}{x}} \quad (11.120)$$

Si nous faisons $x = 2^s$, nous obtenons la somme des puissances inverses de 2 et de même avec $x = 3^s$ tel que:

$$\begin{aligned} 1 + \frac{1}{2^s} + \frac{1}{4^s} + \frac{1}{8^s} + \dots + \frac{1}{2^{ns}} + \dots &= \frac{1}{1 - \frac{1}{2^s}} \\ 1 + \frac{1}{3^s} + \frac{1}{9^s} + \frac{1}{27^s} + \dots + \frac{1}{3^{ns}} + \dots &= \frac{1}{1 - \frac{1}{3^s}} \end{aligned} \quad (11.121)$$

Si nous faisons le produit de ces deux expressions, nous obtenons la somme des puissances de toutes les fractions dont le dénominateur est un nombre produit de 2 et de 3:

$$\frac{1}{1 - \frac{1}{2^s}} \cdot \frac{1}{1 - \frac{1}{3^s}} = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \frac{1}{5^s} + \frac{1}{6^s} + \dots + \frac{1}{(2^a 3^b)^s} + \dots \quad (11.122)$$

Si nous prenons tous les nombres premiers à gauche, nous obtiendrons à droite tous les nombres entiers, puisque tout entier est produit de nombres premiers selon le théorème fondamental de l'arithmétique (cf. chapitre de Théorie Des Nombres), et c'est l'identité fondamentale d'Euler: ce que nous appelons maintenant la "**fonction zêta de Riemann**" est à la fois un produit fini et la somme des puissances inverses de tous les entiers:

$$\zeta(s) = \frac{1}{1 - \frac{1}{2^s}} \cdot \frac{1}{1 - \frac{1}{3^s}} \cdot \frac{1}{1 - \frac{1}{5^s}} = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \frac{1}{5^s} + \frac{1}{6^s} + \dots + \frac{1}{(2^a 3^b 5^c)^s} \quad (11.123)$$

En notation condensée, "l'identité d'Euler" est:

$$\zeta(s) = \sum_{n=1}^{+\infty} \frac{1}{n^s} = \prod_p \frac{1}{1 - \frac{1}{p^s}} = \prod_p \left(1 - p^{-s}\right)^{-1} \quad (11.124)$$

où p sont les nombres premiers.

Nous proposons maintenant au lecteur de sauter ce qui va suivre concernant la fonction zêta de Riemann et d'y revenir une fois les séries de Fourier présentées plus bas dans ce chapitre maîtrisées et comprises.

Nous supposons pour ce qui va suivre que les séries de Fourier sont maintenant connues et que l'égalité de Parseval a été étudiée (puisque'elle est aussi démontrée plus bas). Nous allons chercher à déterminer la fonction zêta de Riemann pour deux valeurs (s valant respectivement 2 et 4) qui nous seront utiles lors de la valorisation d'intégrales dans certains chapitres de la section de Mécanique.

Pour déterminer la valeur de $\zeta(2)$, nous allons exprimer la fonction:

$$f(x) = x \quad \forall x \in [-\pi, \pi] \quad (11.125)$$

sous forme de série de Fourier (ne pas oublier qu'il y a deux manières traditionnelles de définir une série de Fourier et que nous avons fait le choix des physiciens/ingénieurs!!!):

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos(nx) + b_n \sin(nx)) \quad (11.126)$$

Rappelons que comme nous l'avons démontré lors de notre étude des séries de Fourier, les coefficients de Fourier a_0, a_n, b_n s'obtiennent en résolvant:

$$a_n = \frac{2}{T} \int_{-T/2}^{T/2} f(x) \cos(nx) dx \quad b_n = \frac{2}{T} \int_{-T/2}^{T/2} f(x) \sin(nx) dx \quad (11.127)$$

et en utilisant l'intégration par parties (cf. chapitre de Calcul Différentiel Et Intégral). Nous avons alors:

$$\begin{aligned}
a_0 &= \frac{2}{T} \int_{-T/2}^{T/2} f(x) dx = \frac{2}{2\pi} \int_{-\pi}^{\pi} x dx = \frac{1}{\pi} \int_{-\pi}^{\pi} x dx = \frac{1}{2\pi} \left(x^2 \Big|_{-\pi}^{\pi} \right) = 0 \\
a_n &= \frac{2}{T} \int_{-T/2}^{T/2} f(x) \cos(nx) dx = \frac{2}{2\pi} \int_{-\pi}^{\pi} x \cos(nx) dx = \frac{1}{\pi} \left(\frac{x \sin(nx)}{n} \Big|_{-\pi}^{\pi} - \int_{-\pi}^{\pi} \frac{\sin(nx)}{n} dx \right) \\
&= \frac{1}{\pi} \left(\frac{\cos(nx)}{n^2} \Big|_{-\pi}^{\pi} \right) = 0 \\
b_n &= \frac{2}{T} \int_{-T/2}^{T/2} f(x) \sin(nx) dx = \frac{2}{2\pi} \int_{-\pi}^{\pi} x \sin(nx) dx = \frac{1}{\pi} \int_{-\pi}^{\pi} x \sin(nx) dx \\
&= \frac{1}{\pi} \left(-\frac{x \cos(nx)}{n} \Big|_{-\pi}^{\pi} - \int_{-\pi}^{\pi} \frac{\cos(nx)}{n} dx \right) = \frac{1}{\pi} \left(-\frac{x \cos(nx)}{n} \Big|_{-\pi}^{\pi} + \frac{\sin(nx)}{n} \Big|_{-\pi}^{\pi} \right) \\
&= \frac{1}{\pi} \left[-\frac{x \cos(nx)}{n} \right]_{-\pi}^{\pi} = (-1)^{n+1} \frac{2}{n}
\end{aligned} \tag{11.128}$$

Il vient alors:

$$f(x) = x = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos(nx) + b_n \sin(nx)) = \sum_{n=1}^{\infty} b_n \sin(nx) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{2}{n} \sin(nx) \tag{11.129}$$

Mais le théorème de Parseval démontré lors de notre étude des séries de Fourier nous donne aussi (se rappeler que suivant le choix de la définition de la série de Fourier et des coefficients associés, le théorème de Parseval s'exprime un peu différemment!):

$$f(x) = x = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{2}{n} \sin(nx) = \frac{a_0^2}{4} + \frac{1}{2} \sum_{n=1}^{\infty} (a_n^2 + b_n^2) \tag{11.130}$$

Il vient alors immédiatement:

$$f(x) = x = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{2}{n} \sin(nx) = \frac{1}{2} \sum_{n=1}^{\infty} b_n^2 = \frac{1}{2} \sum_{n=1}^{\infty} \left((-1)^{n+1} \frac{2}{n} \right)^2 = \frac{1}{2} \sum_{n=1}^{\infty} \frac{4}{n^2} = 2 \sum_{n=1}^{\infty} \frac{1}{n^2} \tag{11.131}$$

Mais nous avons aussi vu lors de notre démonstration du théorème de Parseval que:

$$\frac{1}{T} \int_{-T/2}^{T/2} f(t) \bar{f}(t) dt = \frac{a_0^2}{4} + \frac{1}{2} \sum_{n=1}^{\infty} (a_n^2 + b_n^2) \tag{11.132}$$

Il vient alors dans notre cas:

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} x^2 dx = 2 \sum_{n=1}^{\infty} \frac{1}{n^2} \tag{11.133}$$

Donc:

$$\frac{1}{2\pi} \frac{x^3}{3} \Big|_{-\pi}^{\pi} = 2 \sum_{n=1}^{\infty} \frac{1}{n^2} \Rightarrow \frac{1}{2\pi} \left[\frac{\pi^3}{3} - \frac{(-\pi)^3}{3} \right] = 2 \sum_{n=1}^{\infty} \frac{1}{n^2} \Rightarrow \frac{\pi^3}{3\pi} = 2 \sum_{n=1}^{\infty} \frac{1}{n^2} \quad (11.134)$$

Donc:

$$\zeta(2) = \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} \quad (11.135)$$

Pour déterminer la valeur de $\zeta(4)$, nous allons procéder de même, mais avec la fonction:

$$f(x) = x^2 \quad \forall x \in [-\pi, \pi] \quad (11.136)$$

sous forme de série de Fourier:

$$f(x) = x^2 = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos(nx) + b_n \sin(nx)) \quad (11.137)$$

Pour cela, nous allons calculer les coefficients de Fourier conformément au choix d'écriture fait lors de leur étude et en utilisant l'intégration par parties (cf. chapitre de Calcul Différentiel Et Intégral):

$$\begin{aligned} a_0 &= \frac{2}{T} \int_{-T/2}^{T/2} f(x) dx = \frac{2}{2\pi} \int_{-\pi}^{\pi} x^2 dx = \frac{1}{3\pi} x^3 \Big|_{-\pi}^{\pi} = \frac{2\pi^2}{3} \\ a_n &= \frac{2}{T} \int_{-T/2}^{T/2} f(x) \cos(nx) dx = \frac{2}{2\pi} \int_{-\pi}^{\pi} x^2 \cos(nx) dx = \frac{1}{\pi} \left(x^2 \frac{\sin(nx)}{n} \Big|_{-\pi}^{\pi} - \int_{-\pi}^{\pi} 2x \frac{\sin(nx)}{n} dx \right) \\ &= -\frac{2}{\pi n} \int_{-\pi}^{\pi} x \sin(nx) dx = -\frac{2}{\pi n} \left(-x \frac{\cos(nx)}{n} \Big|_{-\pi}^{\pi} - \int_{-\pi}^{\pi} \frac{-\cos(nx)}{n} dx \right) \\ &= -\frac{2}{\pi n} \left(-x \frac{\cos(nx)}{n} \Big|_{-\pi}^{\pi} + \frac{\sin(nx)}{n^2} \Big|_{-\pi}^{\pi} \right) = \frac{2}{\pi n} x \frac{\cos(nx)}{n} \Big|_{-\pi}^{\pi} = (-1)^{n+1} \frac{4}{n^2} \\ b_n &= \frac{2}{T} \int_{-T/2}^{T/2} f(x) \sin(nx) dx = \frac{2}{2\pi} \int_{-\pi}^{\pi} x^2 \sin(nx) dx = \frac{1}{\pi} \int_{-\pi}^{\pi} x^2 \sin(nx) dx \\ &= \frac{1}{\pi} \left(-x^2 \frac{\cos(nx)}{n} \Big|_{-\pi}^{\pi} + \int_{-\pi}^{\pi} 2x \frac{\cos(nx)}{n} dx \right) = \frac{2}{\pi n} \left(\int_{-\pi}^{\pi} x \cos(nx) dx \right) \\ &= \frac{2}{\pi n} \left(x \frac{\sin(nx)}{n} \Big|_{-\pi}^{\pi} + \int_{-\pi}^{\pi} \frac{\sin(nx)}{n} dx \right) = -\frac{2}{\pi n} \left(\int_{-\pi}^{\pi} \frac{\sin(nx)}{n} dx \right) = -\frac{2}{\pi n^2} \left(\frac{-\cos(nx)}{n} \Big|_{-\pi}^{\pi} \right) = 0 \end{aligned} \quad (11.138)$$

Il vient alors:

$$f(x) = x^2 = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(nx) = \frac{\pi^2}{3} + \sum_{n=1}^{\infty} (-1)^{n+1} \frac{4}{n^2} \cos(nx) \quad (11.139)$$

Mais le théorème de Parseval démontré lors de notre étude des séries de Fourier (voir plus bas) nous donne aussi:

$$f(x) = x^2 = \frac{\pi^2}{3} + \sum_{n=1}^{\infty} (-1)^{n+1} \frac{4}{n^2} \cos(nx) = \frac{a_0^2}{4} + \frac{1}{2} \sum_{n=1}^{\infty} (a_n^2 + b_n^2) \quad (11.140)$$

Il vient alors immédiatement:

$$f(x) = x^2 = \frac{\pi^2}{3} + \sum_{n=1}^{\infty} (-1)^{n+1} \frac{4}{n^2} \cos(nx) = \frac{\pi^4}{9} + 8 \sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{9} + 8\zeta(4) \quad (11.141)$$

Mais nous avons aussi vu lors de notre démonstration du théorème de Parseval que:

$$\frac{1}{T} \int_{-T/2}^{T/2} f(t) \bar{f}(t) dt = \frac{a_0^2}{4} + \frac{1}{2} \sum_{n=1}^{\infty} (a_n^2 + b_n^2) \quad (11.142)$$

Il vient alors dans notre cas:

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} x^4 dx = \frac{\pi^4}{9} + 8\zeta(4) \quad (11.143)$$

Donc:

$$\begin{aligned} \frac{1}{2\pi} \frac{x^5}{5} \Big|_{-\pi}^{\pi} &= \frac{\pi^4}{9} + 8\zeta(4) \Rightarrow \frac{1}{2\pi} \left[\frac{\pi^5}{5} - \frac{(-\pi)^5}{5} \right] = \frac{\pi^4}{9} + 8\zeta(4) \Rightarrow \frac{\pi^5}{5\pi} = \frac{\pi^4}{9} + 8\zeta(4) \\ \Rightarrow \frac{\pi^4}{5} - \frac{\pi^4}{9} &= 8\zeta(4) \Rightarrow \frac{9\pi^4}{45} - \frac{5\pi^4}{45} = 8\zeta(4) \Rightarrow \frac{4\pi^4}{45} = 8\zeta(4) \end{aligned} \quad (11.144)$$

Donc:

$$\zeta(4) = \sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90} \quad (11.145)$$

2.4. SÉRIES DE TAYLOR ET DE MACLAURIN

Les séries de Taylor et de Maclaurin constituent un outil pratique très puissant pour simplifier des modèles théoriques ou des calculs informatiques (modélisation de fluides ou champs dans l'espace). Elles sont utilisées énormément dans tous les domaines de la physique mais on les retrouve aussi dans l'industrie notamment en ingénierie (plans d'expérience, méthodes numériques, gestion de la qualité), statistiques (approximations d'intégrales), finance (processus stochastiques), analyse complexe... Nous conseillons donc vivement au lecteur de bien lire les développements qui vont suivre.

Soit un polynôme (à une variable):

$$P_n(x) = \sum_{k=0}^n a_k x^k \quad (11.146)$$

Nous avons trivialement pour ce dernier:

$$P(0) = a_0 \quad (11.147)$$

Soit maintenant la dérivée du polynôme $P(x)$:

$$P'(x) = \sum_{k=1}^n k a_k x^{k-1} \quad (11.148)$$

donc:

$$P'(0) = 1 \cdot a_1 \quad (11.149)$$

et ainsi de suite avec $P''(x), P'''(x), \dots$ tel que:

$$\begin{aligned} P''(x) &= \sum_{k=2}^n k(k-1) a_k x^{k-2} \Rightarrow P''(0) = 1 \cdot 2 \cdot a_2 \\ P'''(x) &= \sum_{k=3}^n k(k-1)(k-2) a_k x^{k-3} \Rightarrow P'''(0) = 1 \cdot 2 \cdot 3 \cdot a_3 \\ &\dots \\ \frac{d^n P(x)}{dx^n} &= k(k-1)(k-2)\dots(k-(n-1)) a_k x^{k-n} \Rightarrow P^n(0) = n! a_n \end{aligned} \quad (11.150)$$

Il s'ensuit que:

$$a_0 = P(0), a_1 = \frac{1}{1!} \frac{dP(0)}{dx}, a_2 = \frac{1}{2!} \frac{d^2 P(0)}{dx^2}, a_3 = \frac{1}{3!} \frac{d^3 P(0)}{dx^3}, \dots, a_n = \frac{1}{n!} \frac{d^n P(0)}{dx^n} \quad (11.151)$$

Donc finalement notre polynôme peut s'écrire:

$$P_n(x) = \sum_{k=0}^n \frac{1}{k!} \frac{d^k P(0)}{dx^k} x^k \quad (11.152)$$

relation que nous appelons "**série de Maclaurin limitée**" ou tout simplement "**série de Maclaurin**" d'ordre $k+1$.

En appliquant maintenant le même raisonnement mais en centrant le polynôme sur la valeur $x = x_0$, nous avons:

$$P_n(x) = \sum_{k=0}^n a_k (x - x_0)^k \quad (11.153)$$

et ainsi le développement précédent devient:

$$P_n(x) = \sum_{k=0}^n \frac{1}{k!} \frac{d^k P(x_0)}{dx^k} (x - x_0)^k \quad (11.154)$$

qui n'est d'autre que l'expression générale d'un polynôme exprimé sous une forme dite de "**série de Taylor limitée**" d'ordre $k+1$. Cette fonction peut être assimilée à un polynôme tant que n est fini. Mais si n est infini, comme nous le verrons plus loin, cette série converge vers la fonction dont nous cherchons la représentation sous forme de somme de termes.

Ainsi, certaines fonctions $f(x)$ pouvant être approchées par un polynôme $P(x)$ (une somme de puissances autrement dit...) centré sur la valeur x_0 peuvent être exprimées sous la forme:

$$f(x) = \sum_{k=0}^n \frac{1}{k!} \frac{d^k f(x_0)}{dx^k} (x - x_0)^k = \sum_{k=0}^n \frac{d^{(k)} f(x_0)}{k!} (x - x_0)^k \quad (11.155)$$

Relation souvent désignée sous le nom de "**théorème de Taylor**".

Par contre cette dernière relation n'est pas juste pour toutes les fonctions ne pouvant pas s'exprimer sous forme de polynômes. Dès lors nous disons que la série n'est pas convergente pour ces dernières. Nous en verrons un exemple plus bas.

La dernière relation s'écrit aussi de manière plus conventionnelle...:

$$f(x) = \sum_{k=0}^n \frac{D^k f(x_0)}{k!} (x - x_0)^k \quad (11.156)$$

Revenons brièvement à l'approximation de $f(x)$ proche et centrée en x_0 :

$$f(x) = \sum_{k=0}^n \frac{1}{k!} \frac{d^k f(x_0)}{dx^k} (x - x_0)^k \quad (11.157)$$

Certaines personnes n'aiment pas utiliser cette formulation car on risque d'oublier que l'approximation pour quelques termes n'est bonne que tant que l'on ne s'éloigne pas trop de x_0 avec x . Raison pour laquelle il arrive souvent que nous posions:

$$x = x_0 + \Delta h = x_0 + h \quad (11.158)$$

avec x_0 fixé et h variable mais petit (!) et ainsi il vient alors une forme d'écriture courante des séries de Taylor:

$$f(x_0 + h) = \sum_{k=0}^n \frac{1}{k!} \frac{d^k f(x_0)}{dx^k} (x_0 + h - x_0)^k = \sum_{k=0}^n \frac{h^k}{k!} \frac{d^k f}{dx^k}(x_0) \quad (11.159)$$

Voyons un exemple d'application avec une série de Maclaurin (avec x_0 étant nul) de la fonction $\sin(x)$ et Maple 4.00b:

```

>p[n](x) = sum((D@@i)(f)(a)/i!*(x-a)^i,i=0..n);
>p11:= taylor(sin(x),x=0,12);
>p11:= convert(p11,polynom);
>with(plots):
>tays:= plots[display](sinplot):
for i from 1 by 2 to 11 do
tpl:= convert(taylor(sin(x), x=0,i),polynom):
tays:= tays,plots[display]([sinplot,plot(tpl,x=-Pi..2*Pi,y=-2..2,
color=black,title=convert(tpl,string))]) od:
>plots[display]([tays],view=[-Pi..2*Pi,-2..2]);

```

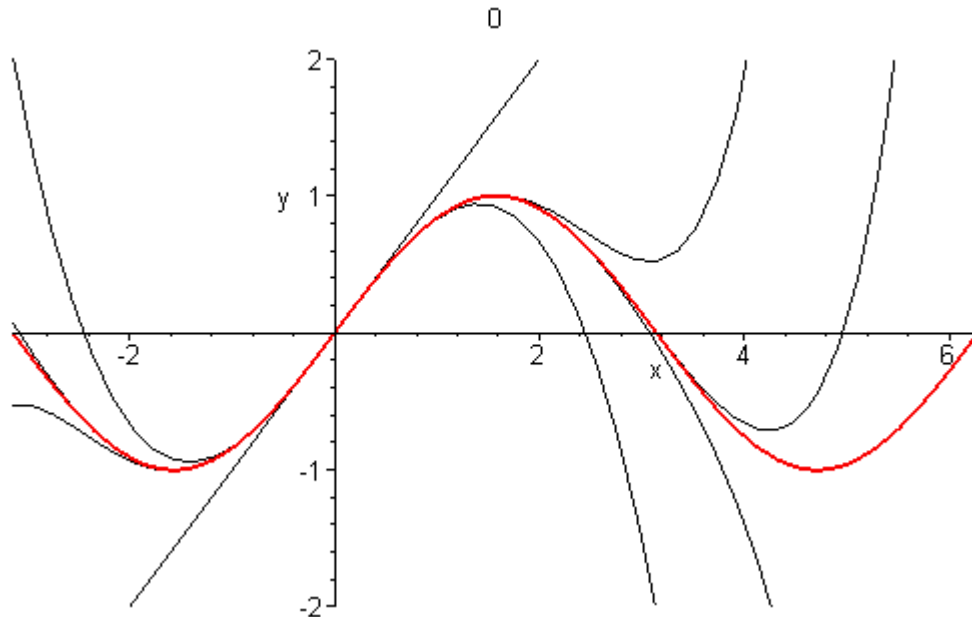


Figure: 11.4 - Approximation de la fonction sinus par un développement de Maclaurin sur Maple 4.00b

Nous voyons donc bien dans cet exemple que la série de Maclaurin ne permet que d'approcher une fonction en un point avec un nombre limités de points. Mais plus nous prenons de termes (mettre 100 termes dans l'exemple précédent) plus la validité est grande sur tout le domaine de définition de la fonction. Au fait il est possible de démontrer que la fonction $\sin(x)$ est exactement exprimable en série de Maclaurin lorsque le nombre de termes est infini. Nous disons alors que son "reste" est nul.

Par contre ceci n'est pas vrai pour toutes les fonctions. Par exemple avec la fonction:

$$f(x) = \frac{1}{1-x^2} \quad (11.160)$$

```

>p[n](x) = sum((D@@i)(f)(a)/i!*(x-a)^i,i=0..n);
>p10:= taylor(1/(1-x^2),x=0,10);
>p10:= convert(p10,polynom);
>with(plots):
>tays:= plots[display](xplot):
for i from 1 by 2 to 10 do
tpl:= convert(taylor(1/(1-x^2), x=0,i),polynom):
tays:= tays,plots[display]([xplot,plot(tpl,x=-2..2,y=-2..2,
color=black,title=convert(tpl,string))]) od:
>plots[display]([tays],view=[-2..2,-2..2]);

```

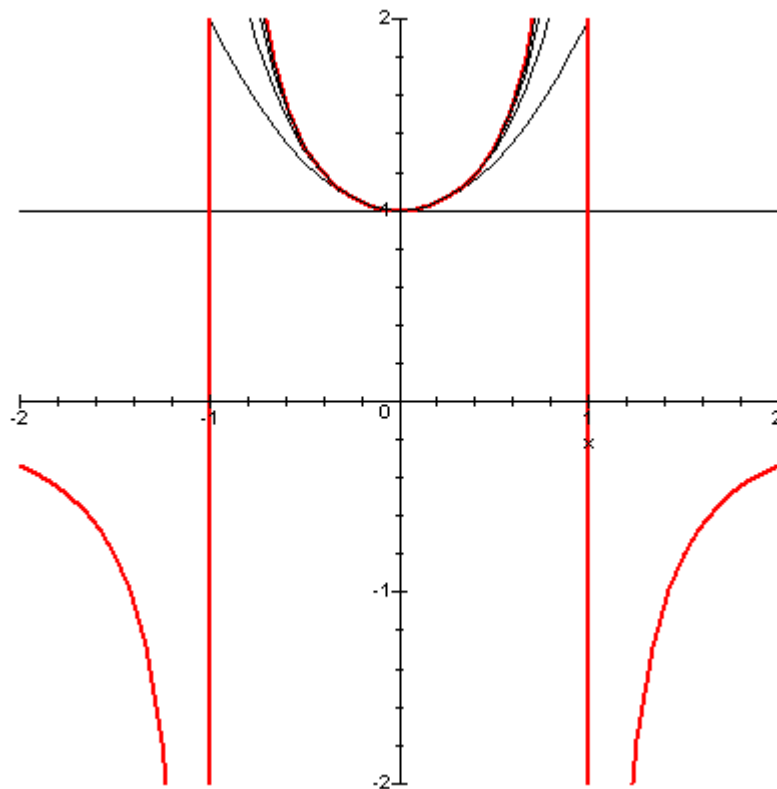



Figure: 11.5 - Contre-exemple d'approche par Maclaurin sur Maple 4.00b

Nous voyons bien ci-dessus que peu importe le nombre de termes que nous prenons, la série de Maclaurin converge seulement dans un domaine de définition compris entre $]-1,1[$. Cette intervalle est appelé le "**rayon de convergence**" et sa détermination (celle des singularités) est un point crucial dans de nombreux domaines de l'ingénierie, de la physique et de l'analyse. Nous y reviendrons plus en détails dans le chapitre d'Analyse Complexe.

Par contre nous pouvons décaler la série de Maclaurin de la fonction précédente afin d'approcher la fonction avec une série de Taylor en un autre point non singulier comme par exemple en x_0 valant 2:

```
>p[n](x) = sum((D@@i)(f)(a)/i!*(x-a)^i,i=0..n);
>p10:= taylor(1/(1-x^2),x=2,10);
>p10:= convert(p10,polynom);
>with(plots):
>tays:= plots[display](xplot);
for i from 1 by 2 to 10 do
tpl:= convert(taylor(1/(1-x^2), x=2,i),polynom):
tays:= tays,plots[display]([xplot,plot(tpl,x=0..5,y=-2..2,
color=black,title=convert(tpl,string))]) od:
>plots[display]([tays],view=[-0..5,-2..2]);
```

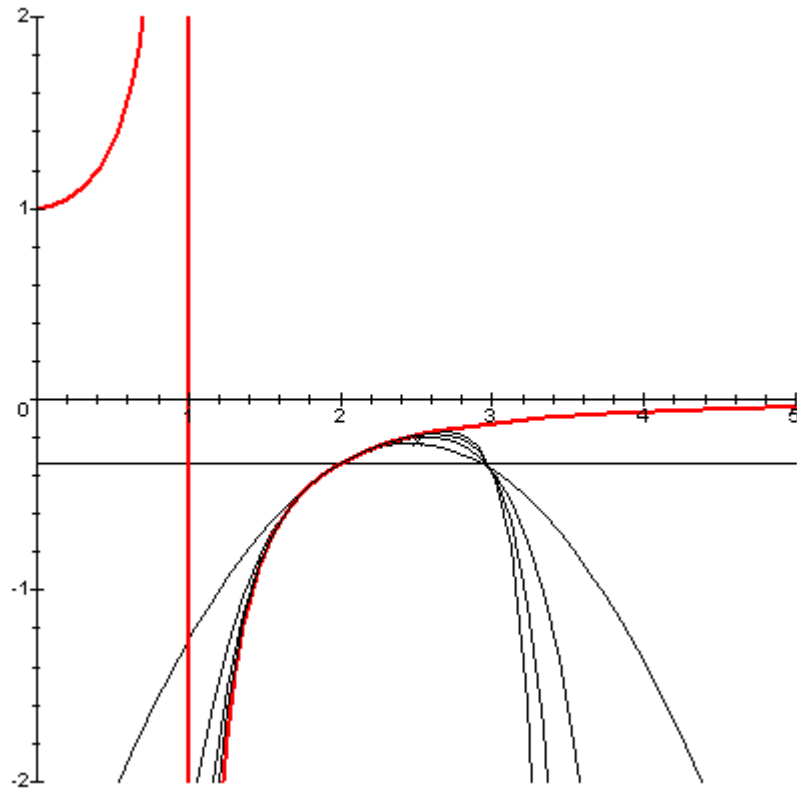


Figure: 11.6 - Décalage possible de l'approche par Maclaurin avec Maple 4.00b

Nous étudierons une généralisation au plan complexe des séries de Taylor précédentes dans le chapitre d'Analyse Complexe pour obtenir un résultat très puissant permettant aux physiciens de calculer des intégrales curvilignes compliquées.

2.4.1. DÉVELOPPEMENTS DE MACLAURIN USUELS

Nous allons démontrer ici les développements de Maclaurin les plus fréquents (une petite dizaine) jusqu'au deuxième ordre que nous puissions rencontrer en physique théorique et mathématique (en fait, nous avons développés uniquement ceux qui sont utilisés dans l'ensemble du site). La liste est pour l'instant non exhaustive mais les démonstrations étant généralisées, elles peuvent s'appliquer à un grand nombre d'autres cas (que nous appliquerons/rencontrerons tout au long de ce site).

Remarque: Les développements de Taylor (donc ailleurs qu'en zéro) étant très rare (il y en a un ou deux sur l'ensemble du site mais ils sont détaillés dans le chapitre respectif), nous les omettons.

1. Développement de Taylor-Maclaurin de $f(x) = e^x$:

D'abord rappelons que nous avons démontré dans le chapitre de Calcul Différentiel Et Intégral que:

$$\begin{aligned}
 f'(x) &= \frac{de^x}{dx} = e^x \\
 f''(x) &= \frac{d^2 e^x}{dx^2} = \frac{d}{dx} \frac{de^x}{dx} = \frac{d}{dx} e^x = e^x \quad (11.161) \\
 f'''(x) &= \frac{d^3 e^x}{dx^3} = e^x
 \end{aligned}$$

Il vient alors que:

$$\begin{aligned}
 f_3(x) \cong e^x &= \sum_{k=0}^3 \frac{1}{k!} \frac{d^k P(0)}{dx^k} x^k = \frac{1}{0!} \frac{d^0 P(0)}{dx^0} x^0 + \frac{1}{1!} \frac{d^1 P(0)}{dx^1} x^1 + \frac{1}{2!} \frac{d^2 P(0)}{dx^2} x^2 + \frac{1}{3!} \frac{d^3 P(0)}{dx^3} x^3 \\
 &\cong \frac{1}{0!} e^x \Big|_{x=0} x^0 + \frac{1}{1!} \frac{de^x}{dx} \Big|_{x=0} x^1 + \frac{1}{2!} \frac{d^2 e^x}{dx^2} \Big|_{x=0} x^2 + \frac{1}{3!} \frac{d^3 e^x}{dx^3} \Big|_{x=0} x^3 \\
 &= \frac{1}{0!} e^0 x^0 + \frac{1}{1!} e^x \Big|_{x=0} x^1 + \frac{1}{2!} e^x \Big|_{x=0} x^2 + \frac{1}{3!} e^x \Big|_{x=0} x^3 = 1 + \frac{1}{1!} e^0 x + \frac{1}{2!} e^0 x^2 + \frac{1}{3!} e^0 x^3 \\
 &= 1 + \frac{1}{1!} x + \frac{1}{2!} x^2 + \frac{1}{3!} x^3 = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!}
 \end{aligned} \quad (11.162)$$

2. Développement de Taylor-Maclaurin de $f(x) = \sin(x)$:

D'abord rappelons que nous avons démontré dans le chapitre de Calcul Différentiel Et Intégral que:

$$\begin{aligned}
 f'(x) &= \frac{d \sin(x)}{dx} = \cos(x) \\
 f''(x) &= \frac{d^2 \sin(x)}{dx^2} = \frac{d}{dx} \frac{d \sin(x)}{dx} = \frac{d}{dx} \cos(x) = -\sin(x) \quad (11.163) \\
 f'''(x) &= \frac{d^3 \sin(x)}{dx^3} = \frac{d(-\sin(x))}{dx} = -\cos(x)
 \end{aligned}$$

Il vient alors que:

$$\begin{aligned}
 f_3(x) \cong \sin(x) \\
 &\cong \sum_{k=0}^3 \frac{1}{k!} \frac{d^k P(0)}{dx^k} x^k = \frac{1}{0!} \frac{d^0 P(0)}{dx^0} x^0 + \frac{1}{1!} \frac{d^1 P(0)}{dx^1} x^1 + \frac{1}{2!} \frac{d^2 P(0)}{dx^2} x^2 + \frac{1}{3!} \frac{d^3 P(0)}{dx^3} x^3 \\
 &= \frac{1}{0!} \sin(x) \Big|_{x=0} x^0 + \frac{1}{1!} \frac{d \sin(x)}{dx} \Big|_{x=0} x^1 + \frac{1}{2!} \frac{d^2 \sin(x)}{dx^2} \Big|_{x=0} x^2 + \frac{1}{3!} \frac{d^3 \sin(x)}{dx^3} \Big|_{x=0} x^3 \\
 &= \frac{1}{0!} \sin(0) x^0 + \frac{1}{1!} \cos(x) \Big|_{x=0} x^1 - \frac{1}{2!} \sin(x) \Big|_{x=0} x^2 - \frac{1}{3!} \cos(x) \Big|_{x=0} x^3 = \frac{1}{1!} x - \frac{1}{3!} x^3 \\
 &= x - \frac{x^3}{3!}
 \end{aligned} \quad (11.164)$$

3. Développement de Taylor-Maclaurin de $f(x) = \cos(x)$:

D'abord rappelons que nous avons démontré dans le chapitre de Calcul Différentiel Et Intégral que:

$$\begin{aligned} f'(x) &= \frac{d \cos(x)}{dx} = -\sin(x) \\ f''(x) &= \frac{d^2 \cos(x)}{dx^2} = \frac{d}{dx} \frac{d \cos(x)}{dx} = -\frac{d}{dx} \sin(x) = -\cos(x) \quad (11.165) \\ f'''(x) &= \frac{d^3 \cos(x)}{dx^3} = \frac{d(-\cos(x))}{dx} = \sin(x) \end{aligned}$$

Il vient alors que:

$$\begin{aligned} f_3(x) &\cong \cos(x) \\ &\cong \sum_{k=0}^3 \frac{1}{k!} \frac{d^k P(0)}{dx^k} x^k = \frac{1}{0!} \frac{d^0 P(0)}{dx^0} x^0 + \frac{1}{1!} \frac{d^1 P(0)}{dx^1} x^1 + \frac{1}{2!} \frac{d^2 P(0)}{dx^2} x^2 + \frac{1}{3!} \frac{d^3 P(0)}{dx^3} x^3 \\ &= \frac{1}{0!} \cos(x) \Big|_{x=0} x^0 + \frac{1}{1!} \frac{d \cos(x)}{dx} \Big|_{x=0} x^1 + \frac{1}{2!} \frac{d^2 \cos(x)}{dx^2} \Big|_{x=0} x^2 + \frac{1}{3!} \frac{d^3 \cos(x)}{dx^3} \Big|_{x=0} x^3 \\ &= \frac{1}{0!} \cos(0) x^0 - \frac{1}{1!} \sin(x) \Big|_{x=0} x^1 - \frac{1}{2!} \cos(x) \Big|_{x=0} x^2 + \frac{1}{3!} \sin(x) \Big|_{x=0} x^3 = \frac{1}{0!} - \frac{1}{2!} x^2 \\ &= 1 - \frac{x^2}{2!} \end{aligned} \quad (11.166)$$

4. Développement de Taylor-Maclaurin de $f(x) = \tan(x)$:

D'abord rappelons que nous avons démontré dans le chapitre de Calcul Différentiel Et Intégral que:

$$\begin{aligned} f'(x) &= \frac{d \tan(x)}{dx} = \frac{1}{\cos^2(x)} = 1 + \tan^2(x) \\ f''(x) &= \frac{d^2 \tan(x)}{dx^2} = \frac{d}{dx} \frac{d \tan(x)}{dx} = \frac{d}{dx} \left(\frac{1}{\cos^2(x)} \right) = 2 \frac{\sin(x)}{\cos^3(x)} \\ f'''(x) &= \frac{d^3 \tan(x)}{dx^3} = 2 \frac{d \left(\frac{\sin(x)}{\cos^3(x)} \right)}{dx} = 2 \frac{d \left(\sin(x) \frac{1}{\cos^3(x)} \right)}{dx} \quad (11.167) \\ &= \frac{2}{\cos^2(x)} + 2 \sin(x) 3 \frac{\sin(x)}{\cos^4(x)} = \frac{2}{\cos^2(x)} + 6 \frac{\sin^2(x)}{\cos^4(x)} \end{aligned}$$

Il vient alors que:

$$\begin{aligned}
f_3(x) &\cong \tan(x) \\
&\cong \sum_{k=0}^3 \frac{1}{k!} \frac{d^k P(0)}{dx^k} x^k = \frac{1}{0!} \frac{d^0 P(0)}{dx^0} x^0 + \frac{1}{1!} \frac{d^1 P(0)}{dx^1} x^1 + \frac{1}{2!} \frac{d^2 P(0)}{dx^2} x^2 + \frac{1}{3!} \frac{d^3 P(0)}{dx^3} x^3 \\
&= \frac{1}{0!} \tan(x) \Big|_{x=0} x^0 + \frac{1}{1!} \frac{d \tan(x)}{dx} \Big|_{x=0} x^1 + \frac{1}{2!} \frac{d^2 \tan(x)}{dx^2} \Big|_{x=0} x^2 + \frac{1}{3!} \frac{d^3 \tan(x)}{dx^3} \Big|_{x=0} x^3 \\
&= \frac{1}{0!} \tan(0) x^0 + \frac{1}{1!} \frac{1}{\cos^2(x)} \Big|_{x=0} x^1 + \frac{1}{2!} 2 \frac{\sin(x)}{\cos^3(x)} \Big|_{x=0} x^2 + \frac{1}{3!} \left(\frac{2}{\cos^2(x)} + 6 \frac{\sin^2(x)}{\cos^4(x)} \right) \Big|_{x=0} x^3 \\
&= \frac{1}{1!} x + \frac{2}{3!} x^3 = x + \frac{x^3}{3}
\end{aligned}$$

(11.168)

5. Développement de Taylor-Maclaurin de $f(x) = \arctan(x)$:

D'abord rappelons que nous avons démontré dans le chapitre de Calcul Différentiel Et Intégral que:

$$\begin{aligned}
f'(x) &= \frac{d \arctan(x)}{dx} = \frac{1}{1+x^2} \\
f''(x) &= \frac{d^2 \arctan(x)}{dx^2} = \frac{d}{dx} \frac{d \arctan(x)}{dx} = \frac{d}{dx} \left(\frac{1}{1+x^2} \right) = -2 \frac{x}{(1+x^2)^2} \\
f'''(x) &= \frac{d^3 \arctan(x)}{dx^3} = \frac{d \left(-2 \frac{x}{(1+x^2)^2} \right)}{dx} = -2 \frac{d \left(x \frac{1}{(1+x^2)^2} \right)}{dx} \\
&= \frac{-2}{(1+x^2)^2} + (-2)x \frac{(-2) \cdot 2x}{(1+x^2)^3} = \frac{-2}{(1+x^2)^2} + \frac{8x^2}{(1+x^2)^3}
\end{aligned}$$

(11.169)

Il vient alors que:

$$\begin{aligned}
f_3(x) &\cong \arctan(x) \\
&\cong \sum_{k=0}^3 \frac{1}{k!} \frac{d^k P(0)}{dx^k} x^k = \frac{1}{0!} \frac{d^0 P(0)}{dx^0} x^0 + \frac{1}{1!} \frac{d^1 P(0)}{dx^1} x^1 + \frac{1}{2!} \frac{d^2 P(0)}{dx^2} x^2 + \frac{1}{3!} \frac{d^3 P(0)}{dx^3} x^3 \\
&= \frac{1}{0!} \arctan(x) \Big|_{x=0} x^0 + \frac{1}{1!} \frac{d \arctan(x)}{dx} \Big|_{x=0} x^1 + \frac{1}{2!} \frac{d^2 \arctan(x)}{dx^2} \Big|_{x=0} x^2 + \frac{1}{3!} \frac{d^3 \arctan(x)}{dx^3} \Big|_{x=0} x^3 \\
&= \frac{1}{0!} \arctan(0) x^0 + \frac{1}{1!} \frac{1}{1+x^2} \Big|_{x=0} x^1 + \frac{1}{2!} \frac{-2x}{(1+x^2)^2} \Big|_{x=0} x^2 + \frac{1}{3!} \left(\frac{-2}{(1+x^2)^2} + \frac{8x^2}{(1+x^2)^3} \right) \Big|_{x=0} x^3 \\
&= \frac{1}{1!} x - \frac{2}{3!} x^3 = x - \frac{x^3}{3}
\end{aligned}$$

(11.170)

6. Développement de Taylor-Maclaurin de $f(x) = \frac{1}{1+x}$:

D'abord rappelons que nous avons démontré dans le chapitre de Calcul Différentiel Et Intégral l'expression générale de la dérivée d'un quotient de deux fonctions. Il vient alors:

$$\begin{aligned}
f'(x) &= \frac{d\left(\frac{1}{1+x}\right)}{dx} = -\frac{1}{(1+x)^2} \\
f''(x) &= \frac{d^2\left(\frac{1}{1+x}\right)}{dx^2} = \frac{d}{dx} \frac{d\left(\frac{1}{1+x}\right)}{dx} = -\frac{d}{dx} \left(\frac{1}{(1+x)^2} \right) = \frac{2}{(1+x)^3} \quad (11.171) \\
f'''(x) &= \frac{d^3\left(\frac{1}{1+x}\right)}{dx^3} = \frac{d\left(\frac{2}{(1+x)^3}\right)}{dx} = -\frac{6}{(1+x)^4}
\end{aligned}$$

Il vient alors que:

$$\begin{aligned}
f_3(x) &\cong \frac{1}{1+x} \\
&\cong \sum_{k=0}^3 \frac{1}{k!} \frac{d^k P(0)}{dx^k} x^k = \frac{1}{0!} \frac{d^0 P(0)}{dx^0} x^0 + \frac{1}{1!} \frac{d^1 P(0)}{dx^1} x^1 + \frac{1}{2!} \frac{d^2 P(0)}{dx^2} x^2 + \frac{1}{3!} \frac{d^3 P(0)}{dx^3} x^3 \\
&= \frac{1}{0!} \left(\frac{1}{1+x} \right) \Big|_{x=0} x^0 + \frac{1}{1!} \frac{d \left(\frac{1}{1+x} \right)}{dx^1} \Big|_{x=0} x^1 + \frac{1}{2!} \frac{d^2 \left(\frac{1}{1+x} \right)}{dx^2} \Big|_{x=0} x^2 + \frac{1}{3!} \frac{d^3 \left(\frac{1}{1+x} \right)}{dx^3} \Big|_{x=0} x^3 \\
&= \frac{1}{0!} 1 \cdot x^0 - \frac{1}{1!} \frac{1}{(1+x)^2} \Big|_{x=0} x^1 + \frac{1}{2!} \frac{2}{(1+x)^3} \Big|_{x=0} x^2 - \frac{1}{3!} \frac{6}{(1+x)^4} \Big|_{x=0} x^3 \\
&= 1 - \frac{1}{1!} x + \frac{2}{2!} x^2 - \frac{6}{3!} x^3 = 1 - x + x^2 - x^3
\end{aligned}$$

(11.172)

Il s'ensuit immédiatement une autre série de Taylor que nous retrouverons aussi un certain nombre de fois:

$$f_2(x) \cong \frac{1}{(1+x)^2} = -\frac{d}{dx} \frac{1}{(1+x)} \cong -\frac{d}{dx} (1 - x + x^2 - x^3) = 1 - 2x + 3x^2 \quad (11.173)$$

7. Développement de Taylor-Maclaurin de $f(x) = \sqrt{1+x}$:

D'abord rappelons que nous avons démontré dans le chapitre de Calcul Différentiel Et Intégral l'expression générale de la dérivée d'un quotient de deux fonctions. Il vient alors:

$$\begin{aligned}
f'(x) &= \frac{d\sqrt{1+x}}{dx} = \frac{1}{2} \frac{1}{\sqrt{1+x}} \\
f''(x) &= \frac{d^2\sqrt{1+x}}{dx^2} = \frac{d}{dx} \frac{d\sqrt{1+x}}{dx} = \frac{d}{dx} \left(\frac{1}{2} \frac{1}{\sqrt{1+x}} \right) = -\frac{1}{4} \frac{1}{(1+x)^{3/2}} \\
f'''(x) &= \frac{d^3\sqrt{1+x}}{dx^3} = \frac{d \left(-\frac{1}{4} \frac{1}{(1+x)^{3/2}} \right)}{dx} = -\frac{1}{4} \frac{d \left(\frac{1}{(1+x)^{3/2}} \right)}{dx} = \frac{3}{8} \frac{1}{(1+x)^{5/2}}
\end{aligned}$$

(11.174)

Il vient alors que:

$$\begin{aligned}
f_3(x) &\cong \sqrt{1+x} \\
&\cong \sum_{k=0}^3 \frac{1}{k!} \frac{d^k P(0)}{dx^k} x^k = \frac{1}{0!} \frac{d^0 P(0)}{dx^0} x^0 + \frac{1}{1!} \frac{d^1 P(0)}{dx^1} x^1 + \frac{1}{2!} \frac{d^2 P(0)}{dx^2} x^2 + \frac{1}{3!} \frac{d^3 P(0)}{dx^3} x^3 \\
&= \frac{1}{0!} \sqrt{1+x} \Big|_{x=0} x^0 + \frac{1}{1!} \frac{d\sqrt{1+x}}{dx} \Big|_{x=0} x^1 + \frac{1}{2!} \frac{d^2 \sqrt{1+x}}{dx^2} \Big|_{x=0} x^2 + \frac{1}{3!} \frac{d^3 \sqrt{1+x}}{dx^3} \Big|_{x=0} x^3 \\
&= \frac{1}{0!} \sqrt{1+x} \Big|_{x=0} x^0 + \frac{1}{1!} \frac{1}{2} \frac{1}{\sqrt{1+x}} \Big|_{x=0} x^1 - \frac{1}{2!} \frac{1}{4} \frac{1}{(1+x)^{3/2}} \Big|_{x=0} x^2 + \frac{1}{3!} \frac{3}{8} \frac{1}{(1+x)^{5/2}} \Big|_{x=0} x^3 \\
&= \frac{1}{0!} + \frac{1}{1!} \frac{1}{2} x - \frac{1}{2!} \frac{1}{4} x^2 + \frac{1}{3!} \frac{3}{8} x^3 = 1 + \frac{1}{2} x - \frac{1}{2 \cdot 4} x^2 + \frac{1 \cdot 3}{2 \cdot 4 \cdot 6} x^3
\end{aligned}
\tag{11.175}$$

Il s'ensuit immédiatement une autre série de Taylor que nous retrouverons aussi un certain nombre de fois:

$$f_2(x) \cong \frac{1}{\sqrt{1+x}} = 2 \frac{d}{dx} \sqrt{1+x} \cong 2 \frac{d}{dx} \left(1 + \frac{1}{2} x - \frac{1}{2 \cdot 4} x^2 + \frac{1 \cdot 3}{2 \cdot 4 \cdot 6} x^3 \right) = 1 - \frac{1}{2} x + \frac{1 \cdot 3}{2 \cdot 4} x^2 \tag{11.176}$$

8. Développement de Taylor-Maclaurin de $f(x) = \ln(1+x)$:

D'abord rappelons que nous avons démontré dans le chapitre de Calcul Différentiel Et Intégral que la dérivée de la fonction logarithme. Il vient alors (nous retrouvons très vite les termes d'une des séries développée un peu plus haut):

$$\begin{aligned}
f'(x) &= \frac{d \ln(1+x)}{dx} = \frac{1}{1+x} \\
f''(x) &= \frac{d^2 \ln(1+x)}{dx^2} = \frac{d}{dx} \frac{d \ln(1+x)}{dx} = \frac{d}{dx} \left(\frac{1}{1+x} \right) = -\frac{1}{(1+x)^2} \\
f'''(x) &= \frac{d^3 \ln(1+x)}{dx^3} = \frac{d}{dx} \left(-\frac{1}{(1+x)^2} \right) = -\frac{1}{4} \frac{d}{dx} \left(\frac{1}{(1+x)^{3/2}} \right) = \frac{2}{(1+x)^3}
\end{aligned}
\tag{11.177}$$

Il vient alors que:

$$\begin{aligned}
f_3(x) &\cong \ln(1+x) \\
&\cong \sum_{k=0}^3 \frac{1}{k!} \frac{d^k P(0)}{dx^k} x^k = \frac{1}{0!} \frac{d^0 P(0)}{dx^0} x^0 + \frac{1}{1!} \frac{d^1 P(0)}{dx^1} x^1 + \frac{1}{2!} \frac{d^2 P(0)}{dx^2} x^2 + \frac{1}{3!} \frac{d^3 P(0)}{dx^3} x^3 \\
&= \frac{1}{0!} \ln(1+x) \Big|_{x=0} x^0 + \frac{1}{1!} \frac{d \ln(1+x)}{dx^1} \Big|_{x=0} x^1 + \frac{1}{2!} \frac{d^2 \ln(1+x)}{dx^2} \Big|_{x=0} x^2 + \frac{1}{3!} \frac{d^3 \ln(1+x)}{dx^3} \Big|_{x=0} x^3 \\
&= \frac{1}{0!} \ln(1+x) \Big|_{x=0} x^0 + \frac{1}{1!} \frac{1}{1+x} \Big|_{x=0} x^1 - \frac{1}{2!} \frac{1}{(1+x)^2} \Big|_{x=0} x^2 + \frac{1}{3!} \frac{2}{(1+x)^3} \Big|_{x=0} x^3 \\
&= \frac{1}{1!} x - \frac{1}{2!} x^2 + \frac{2}{3!} x^3 = 1+x - \frac{1}{2} x^2 + \frac{1}{3} x^3
\end{aligned}
\tag{11.178}$$

9. Considérons maintenant le cas important pour le modèle de Langevin du paramagnétisme qu'est le développement de Taylor approximé de la fonction cotangente hyperbolique (cf. chapitre de Trigonométrie), définie pour rappel par la relation:

$$\coth(x) = \frac{e^x + e^{-x}}{e^x - e^{-x}} \tag{11.179}$$

Pour cela, nous allons utiliser la notation de Landau avec des expressions du type $O(x^*)$ en se rappelant que nous venons de démontrer un peu plus haut que:

$$e^x = 1+x + \frac{x^2}{2!} + O(x^3) \tag{11.180}$$

lorsque $x \rightarrow 0$.

Pour la cotangente hyperbolique nous avons alors:

$$\begin{aligned}
\coth(x) &= \frac{e^x + e^{-x}}{e^x - e^{-x}} = \frac{1+x + \frac{x^2}{2} + O(x^3) + 1-x + \frac{x^2}{2} + O(x^3)}{1+x + \frac{x^2}{2} + \frac{x^3}{3!} + O(x^4) - 1+x - \frac{x^2}{2} + \frac{x^3}{3!} + O(x^4)} \\
&= \frac{2+x^2 + O(x^3)}{2x + \frac{x^3}{3} + O(x^4)} = \frac{2+x^2 + O(x^3)}{2x} \underbrace{\frac{1}{1+\frac{x^2}{6} + O(x^3)}}_A
\end{aligned}
\tag{11.181}$$

À présent il faut se rappeler comme nous venons de le démontrer un peu plus haut que:

$$\frac{1}{1+x} = 1-x+x^2-x^3+\dots \tag{11.182}$$

pour $|x| < 1$. Donc:

$$A = \frac{1}{1 + \left(\frac{x^2}{6} + O(x^3) \right)} = 1 - \frac{x^2}{6} + O(x^4) \quad (11.183)$$

et pour finir en remplaçant ceci dans l'expression précédente nous trouvons:

$$\begin{aligned} \coth(x) &= \frac{e^x + e^{-x}}{e^x - e^{-x}} = \frac{2 + x^2 + O(x^3)}{2x} \left(1 - \frac{x^2}{6} + O(x^4) \right) \\ &= \frac{1}{2x} \left(2 + x^2 - 2 \frac{x^2}{6} + O(x^4) \right) = \frac{1}{x} + \frac{x}{3} + O(x^3) \end{aligned} \quad (11.184)$$

2.4.2. SÉRIES DE TAYLOR D'UNE FONCTION A 2 VARIABLES

Nous allons voir ici comment approcher une fonction $f(x, y)$ de deux variables réelles par une somme de puissances (série de Taylor). Ce type d'approximation est très utilisé dans de nombreux domaines de l'ingénierie (voir chapitre de Génie Industriel et de Méthodes Numériques).

Nous cherchons donc une approximation de $f(x, y)$ au point $f(x_0 + h, y_0 + k)$. Pour cela, posons (rien ne nous interdit a priori de le faire) que:

$$x = x(t) = x_0 + ht \quad \text{et} \quad y = y(t) = y_0 + kt \quad (11.185)$$

Nous avons alors:

$$f(t) = f(x(t), y(t)) \quad (11.186)$$

La valeur de (l'astuce est là!):

$$f(1) = f(x(1), y(1)) = f(x_0 + h, y_0 + k) \quad (11.187)$$

peut être approchée en utilisant son expression en série de Taylor autour de la valeur 0 telle que:

$$\begin{aligned} f(1) &= \sum_{k=0}^n \frac{1}{k!} \frac{d^k f(t_0)}{dt^k} (t - t_0)^k = \sum_{k=0}^n \frac{1}{k!} \frac{d^k f(0)}{dt^k} (1 - 0)^k = \sum_{k=0}^n \frac{1}{k!} \frac{d^k f(0)}{dt^k} \\ &= f(0) + 1 \cdot \frac{df}{dt}(0) + \frac{1}{2!} \cdot \frac{d^2 f}{dt^2}(0) + \dots + \frac{1}{n!} \cdot \frac{d^n f}{dt^n}(0) \\ &= f(x_0, y_0) + 1 \cdot \frac{df(x(t), y(t))}{dt}(0) + \frac{1}{2!} \cdot \frac{d^2 f(x(t), y(t))}{dt^2}(0) + \dots + \frac{1}{n!} \cdot \frac{d^n f(x(t), y(t))}{dt^n}(0) \end{aligned} \quad (11.188)$$

Or, nous avons:

$$\begin{aligned}
 \frac{df(x(t), y(t))}{dt} &= \frac{\frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy}{dt} = \frac{\frac{\partial f}{\partial x} d(x_0 + ht) + \frac{\partial f}{\partial y} d(y_0 + kt)}{dt} = \frac{\frac{\partial f}{\partial x} h dt + \frac{\partial f}{\partial y} k dt}{dt} \\
 &= \frac{\partial f}{\partial x} h + \frac{\partial f}{\partial y} k = \left(h \frac{\partial}{\partial x} + k \frac{\partial}{\partial y} \right) f
 \end{aligned} \quad (11.189)$$

et:

$$\begin{aligned}
 \frac{d^2 f(x(t), y(t))}{dt^2} &= \frac{d \left(\frac{df}{dt} \right)}{dt} = \frac{\frac{\partial (df)}{\partial x} dx + \frac{\partial (df)}{\partial y} dy}{dt} \\
 &= \frac{\frac{\partial \left(\frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy \right)}{\partial x} dx + \frac{\partial \left(\frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy \right)}{\partial y} dy}{dt} \\
 &= \frac{dx \frac{\partial}{\partial x} \left(\frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy \right) + dy \frac{\partial}{\partial y} \left(\frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy \right)}{dt^2} \\
 &= \frac{\left[(dx)^2 \frac{\partial^2 f}{\partial x^2} + 2 dx dy \frac{\partial^2 f}{\partial x \partial y} \right] + \left[dx dy \frac{\partial^2 f}{\partial y \partial x} + (dy)^2 \frac{\partial^2 f}{\partial y^2} \right]}{dt^2}
 \end{aligned} \quad (11.190)$$

Selon le théorème de Schwarz (cf. chapitre de Calcul Intégral Et Différentiel):

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x} \quad (11.191)$$

Nous avons alors:

$$\begin{aligned}
 \frac{d^2 f(x(t), y(t))}{dt^2} &= \frac{(dx)^2 \frac{\partial^2 f}{\partial x^2} + 2 dx dy \frac{\partial^2 f}{\partial x \partial y} + (dy)^2 \frac{\partial^2 f}{\partial y^2}}{dt^2} = \frac{\left(\frac{\partial}{\partial x} dx + \frac{\partial}{\partial y} dy \right)^2 f}{dt^2} \\
 &= \left(h \frac{\partial}{\partial x} + k \frac{\partial}{\partial y} \right)^2 f
 \end{aligned} \quad (11.192)$$

et nous démontrons par récurrence que:

$$\frac{d^n f(x(t), y(t))}{dt^n} = \left(h \frac{\partial}{\partial x} + k \frac{\partial}{\partial y} \right)^n f \quad (11.193)$$

Nous avons alors finalement:

$$\begin{aligned}
f(x_0 + h, y_0 + k) &= f(x_0, y_0) + 1 \cdot \frac{df(x(t), y(t))}{dt}(0) + \frac{1}{2!} \frac{d^2 f(x(t), y(t))}{dt^2}(0) + \dots \\
&= f(x_0, y_0) + \frac{1}{1!} \left(h \frac{\partial f}{\partial x} + k \frac{\partial f}{\partial y} \right)(0) + \frac{1}{2!} \left(h \frac{\partial f}{\partial x} + k \frac{\partial f}{\partial y} \right)^2(0) + \dots \\
&= f(x_0, y_0) + \frac{1}{1!} \frac{\partial f}{\partial x}(x_0, y_0) h + \frac{1}{1!} \frac{\partial f}{\partial y}(x_0, y_0) k \\
&\quad + \frac{1}{2!} \left[\frac{\partial^2 f}{\partial x^2}(x_0, y_0) h^2 + 2 \frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) h k + \frac{\partial^2 f}{\partial y^2}(x_0, y_0) k^2 \right] + \dots
\end{aligned} \tag{11.194}$$

ou sous une autre forme équivalente simplifiée:

$$\begin{aligned}
f(x, y) &= f(x_0, y_0) + \frac{1}{1!} \frac{\partial f(x_0, y_0)}{\partial x} (x - x_0) + \frac{1}{1!} \frac{\partial f(x_0, y_0)}{\partial y} (y - y_0) \\
&\quad + \frac{1}{2!} \left[\frac{\partial^2 f(x_0, y_0)}{\partial x^2} (x - x_0)^2 + 2 \frac{\partial^2 f(x_0, y_0)}{\partial x \partial y} (x - x_0)(y - y_0) + \frac{\partial^2 f(x_0, y_0)}{\partial y^2} (y - y_0)^2 \right] + \dots
\end{aligned} \tag{11.195}$$

Ou encore si nous définissons une matrice H appelée "**matrice Hessienne**" donnée par:

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} \tag{11.196}$$

nous pouvons aussi écrire:

$$f(x_0 + h, y_0 + k) = f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0) h + \frac{\partial f}{\partial y}(x_0, y_0) k + \frac{1}{2} \begin{pmatrix} h & k \end{pmatrix} H \begin{pmatrix} h \\ k \end{pmatrix} \tag{11.197}$$

2.4.3. FORME QUADRATIQUE

Maintenant nous allons avoir besoin pour le chapitre de Méthodes Numériques d'énoncer une propriété importante (qui aurait tout à fait sa place uniquement dans le chapitre de Calcul Différentiel Et Intégral):

Soit f une fonction définie et dérivée sur un intervalle I et soit a un élément de I . Si f est telle que $f'(a) = 0$ alors nous disons qu'elle a un extremum local en a .

Remarque: La réciproque est fausse, la fonction x^3 en est un exemple. Sa dérivée est nulle en 0 mais il n'y a pas d'extremum local en ce point. Donc il faut être prudent!

Cependant, soit f une fonction définie et dérivée sur un intervalle I et soit a un élément de I . Si f est telle que $f'(a) = 0$ et si f' change de signe en a alors f admet un extremum local en a .

Pour maintenant revenir à notre développement de Taylor à deux variables, nous savons que si (x_0, y_0) est un extremum local de f alors dans un premier temps (cf. chapitre de Calcul Différentiel Et Intégral):

$$\frac{\partial f}{\partial x}(x_0, y_0) = \frac{\partial f}{\partial y}(x_0, y_0) = 0 \quad (11.198)$$

Cependant nous venons de voir que cette condition n'est pas suffisante à garantir que (x_0, y_0) soit un extremum local.

Reconsidérons le développement de Taylor de f ci-dessus en tenant compte de la condition précédente. Le développement se simplifie alors en:

$$f(x_0 + h, y_0 + k) = f(x_0, y_0) + \frac{1}{2} \left[\frac{\partial^2 f}{\partial x^2}(x_0, y_0) h^2 + 2 \frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) hk + \frac{\partial^2 f}{\partial y^2}(x_0, y_0) k^2 \right] \quad (11.199)$$

Nous savons alors par définition que pour que (x_0, y_0) soit un minimum local (respectivement un maximum local) il suffit que l'expression entre crochets soit positive (respectivement négative). Étant donné que les dérivées secondes de f sont continues, il suffit donc que l'expression:

$$q(h, k) := \frac{\partial^2 f}{\partial x^2}(x_0, y_0) h^2 + 2 \frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) hk + \frac{\partial^2 f}{\partial y^2}(x_0, y_0) k^2 \quad (11.200)$$

soit positive (resp. négative) quels que soient h ou k et qu'elle soit nulle que si $h = k = 0$. Nous disons alors que q est une "**forme quadratique définie positive (resp. définie négative)**".

Pour simplifier l'écriture et être conforme aux traditions posons maintenant:

$$a = \frac{\partial^2 f}{\partial x^2}(x_0, y_0), b = \frac{\partial^2 f}{\partial x \partial y}(x_0, y_0), c = \frac{\partial^2 f}{\partial y^2}(x_0, y_0), h = \Delta x, k = \Delta y \quad (11.201)$$

Nous pouvons alors réécrire q comme suit:

$$\begin{aligned} q(\Delta x, \Delta y) &= a \Delta x^2 + 2b \Delta x \Delta y + c \Delta y^2 \\ &= \frac{ac - b^2}{a} \Delta y^2 + a \left(\Delta x + \frac{b \Delta y}{a} \right)^2 = \frac{1}{a} \det(H) \Delta y^2 + a \left(\Delta x + \frac{b \Delta y}{a} \right)^2 \end{aligned} \quad (11.202)$$

Où H est toujours matrice hessienne de f évaluée en (x_0, y_0) .

Nous voyons donc que q est définie positive (minimum local) si $a > 0$ et $\det(H) > 0$, définie négative (maximum local) si $a < 0$ et $\det(H) > 0$.

En revenant aux dérivées partielles ces conditions se réécrivent comme suit:

- Définie positive (minimum local) si:

$$\frac{\partial^2 f}{\partial x^2}(x_0, y_0) > 0 \text{ et } \det(H) = \frac{\partial^2 f}{\partial x^2}(x_0, y_0) \cdot \frac{\partial^2 f}{\partial y^2}(x_0, y_0) - \left(\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) \right)^2 > 0 \quad (11.203)$$

- Définie négative (maximum local) si:

$$\frac{\partial^2 f}{\partial x^2}(x_0, y_0) < 0 \text{ et } \det(H) = \frac{\partial^2 f}{\partial x^2}(x_0, y_0) \cdot \frac{\partial^2 f}{\partial y^2}(x_0, y_0) - \left(\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) \right)^2 > 0 \quad (11.204)$$

En conclusion nous voyons que le signe du déterminant de la matrice hessienne ainsi que celui de

$\frac{\partial^2 f}{\partial x^2}$ nous permettent d'obtenir une condition suffisante pour déterminer si nous sommes en présence d'un extremum local.

2.4.4. RESTE DE LAGRANGE

Il peut y avoir un intérêt dans certaines applications numériques (cf. chapitre de Méthodes Numériques) à connaître l'erreur d'approximation du polynôme $P_n(x)$ par rapport à la fonction $f(x), \forall x$.

Définissons pour cela un "reste" $R_n(x)$, tel que:

$$R_n(x) = f(x) - P_n(x) \quad (11.205)$$

La fonction $R_n(x)$ est appelée "reste de Lagrange".

Considérons maintenant une fonction $f(x)$ qui est $(n+1)$ fois dérivable sur un intervalle qui contient x_0 . Pour une valeur x de l'intervalle, différente de x_0 , nous nous proposons de démontrer qu'il existe un nombre z situé entre x_0 et x tel que:

$$R_n(x) = \frac{f^{(n+1)}(z)}{(n+1)!} \cdot (x - x_0)^{n+1} \quad (11.206)$$

Démonstration:

Soit une fonction $g(t)$ une fonction définie par la différence d'une fonction $f(x)$ supposée connue et une approximation de Taylor de cette même fonction:

$$\begin{aligned} g(t) &= f(x) - \left(f(t) + \frac{f'(t)}{1!} (x-t) + \frac{f''(t)}{2!} (x-t)^2 + \dots + \frac{f^{(n)}(t)}{n!} (x-t)^n + \frac{f^{(n+1)}(t)}{(n+1)!} (x-t)^{n+1} \right) \\ &= f(x) - \left(f(t) + \frac{f'(t)}{1!} (x-t) + \frac{f''(t)}{2!} (x-t)^2 + \dots + \frac{f^{(n)}(t)}{n!} (x-t)^n \right) - R_n(x) \frac{(x-t)^{n+1}}{(x-x_0)^{n+1}} \end{aligned} \quad (11.207)$$

avec bien sûr:

$$R_n(x) \frac{1}{(x-x_0)^{n+1}} = \frac{f^{(n+1)}(x-t)}{(n+1)!} \quad (11.208)$$

Nous voyons que $g(t)$ s'annule bien pour la valeur $t = x$.

Dérivons maintenant $g(t)$ par rapport à t , nous trouvons:

$$\begin{aligned} g'(t) = & -f'(t) - f'(t)(-1) - f''(t)(x-t) - f''(t)(x-t)(-1) + \frac{f'''(t)}{2}(x-t)^2 \\ & + \dots - \frac{f^{(n+1)}(t)}{n!}(x-t)^n + R_n(x)(n+1) \frac{(x-t)^n}{(x-x_0)^{n+1}} \end{aligned} \quad (11.209)$$

Après simplification:

$$g'(t) = -\frac{f^{(n+1)}(t)}{n!}(x-t)^n + R_n(x)(n+1) \frac{(x-t)^n}{(x-x_0)^{n+1}} \quad (11.210)$$

Selon le théorème de Rolle (cf. chapitre de Calcul Différentiel Et Intégral), il existe une valeur $t = z$ pour laquelle la dérivée $g'(t)$ s'annule. Donc:

$$0 = -\frac{f^{(n+1)}(z)}{n!}(x-z)^n + R_n(x)(n+1) \frac{(x-z)^n}{(x-x_0)^{n+1}} \quad (11.211)$$

Nous pouvons simplifier l'équation par $(x-z)^n$:

$$R_n(x) \frac{(n+1)}{(x-x_0)^{n+1}} - \frac{f^{(n+1)}(z)}{n!} = 0 \quad (11.212)$$

ce qui s'écrit aussi:

$$R_n(x) - \frac{f^{(n+1)}(z)}{(n+1)!}(x-x_0)^{n+1} = 0 \quad (11.213)$$

et nous trouvons donc pour maximum de R_n :

$$R_n(x) = \frac{f^{(n+1)}(z)}{(n+1)!}(x-x_0)^{n+1} \quad (11.214)$$

□C.Q.F.D.

Nous voyons que plus le polynôme $P_n(x)$ est de degré élevé, plus il approxime la fonction $f(x)$ avec exactitude. Que se passe-t-il lorsque $n \rightarrow \infty$?

$$P(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x-x_0)^k \quad (11.215)$$

Supposons que $f(x)$ admette des dérivées de tout ordre (ce que nous notons C^∞) pour toutes les valeurs d'un intervalle quelconque contenant x_0 et soit R_n le reste de Lagrange de $f(x)$ en x_0 . Si, quel que soit x dans l'intervalle:

$$\lim_{n \rightarrow \infty} R_n(x) = 0 \quad (11.216)$$

alors $f(x)$ est exactement représentée par $P(x)$ sur l'intervalle.

Démonstration:

Elle découle simplement de l'expression de $P_n(x)$ lorsque $n \rightarrow \infty$.

Effectivement, si nous prenons une infinité de termes pour $P_n(x)$, la correspondance avec la fonction approchée est parfaite et donc le reste est nul.

□C.Q.F.D.

Le polynôme:

$$P(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k \quad (11.217)$$

est appelé "**polynôme de Taylor**" ou "**série de Taylor**". Si $x_0 = 0$, il est appelé "**polynôme de Maclaurin**" ou "**série de Maclaurin**".

2.4.5. FORMULE DE TAYLOR AVEC RESTE INTÉGRAL

Nous allons voir ici un théorème qui nous sera utile dans le chapitre de Statistique pour relier la loi de Poisson et la loi du Khi-2 qui est utilisée dans les logiciels statistique pour le test de Poisson des événements rares (c'est la seule application pratique utilisée dans les entreprises qui nous est connue à ce jour).

Remarque: Si quelqu'un possède une démonstration plus pédagogique dont le début fait un peu moins "formule tombée du ciel", nous sommes preneurs!

Soit f une application de $n+1$ fois dérivable dans l'intervalle $[a, b]$. Nous avons alors :

$$f(b) - f(a) = (b-a)f'(a) + \frac{(b-a)^2}{2!} f''(a) + \dots + \frac{(b-a)^n}{n!} f^{(n)}(a) + \frac{1}{n!} \int_a^b (b-t)^n f^{(n+1)}(t) dt \quad (11.218)$$

où il est important (pour la bonne compréhension de ce que nous ferons dans le chapitre de Statistiques) que le lecteur remarque dans le développement que quand la dérivée s'arrête au n -ième terme dans la série, l'intégrale (le reste) a un facteur $1/n!$, une puissance en n et une dérivée en $n+1$. Donc in extenso, comme nous allons le démontrer ci-après, si nous arrêtons le développement des termes à $n-1$, l'intégrale (le reste) aura un facteur $1/(n-1)!$, une puissance en $n-1$ et une dérivée n -ième.

Démonstration:

La démonstration se fait par récurrence. Nous considérons d'abord la formule tombée du ciel:

$$f(b) - f(a) = (b-a)f'(a) + \frac{(b-a)^2}{2!} f''(a) + \dots + \frac{(b-a)^k}{k!} f^{(k)}(a) + \frac{1}{k!} \int_a^b (b-t)^k f^{(k+1)}(t) dt \quad (11.219)$$

Nous montrons qu'elle est correcte pour $k = 0$ et ensuite nous faisons une récurrence sur k pour $k \in \{0, 1, \dots, n\}$.

Pour $k = 0$, nous avons la relation bien connue (cf. chapitre de Calcul Différentiel Et Intégral):

$$f(b) - f(a) = \int_a^b f'(t) dt \quad (11.220)$$

Supposons la propriété vraie pour $k < n$:

$$f(b) - f(a) = (b-a)f'(a) + \frac{(b-a)^2}{2!} f''(a) + \dots + \frac{(b-a)^k}{k!} f^{(k)}(a) + \frac{1}{k!} \int_a^b (b-t)^k f^{(k+1)}(t) dt \quad (11.221)$$

Nous intégrons par parties (cf. chapitre de Calcul Différentiel Et Intégral) le terme:

$$\frac{1}{k!} \int_a^b (b-t)^k f^{(k+1)}(t) dt \quad (11.222)$$

Nous avons alors:

$$\frac{1}{k!} \int_a^b (b-t)^k f^{(k+1)}(t) dt = \left[-\frac{1}{(k+1)!} (b-t)^{k+1} f^{(k+1)}(t) \right]_a^b + \frac{1}{(k+1)!} \int_a^b (b-t)^{k+1} f^{(k+2)}(t) dt \quad (11.223)$$

d'où:

$$f(b) - f(a) = (b-a)f'(a) + \frac{(b-a)^2}{2!} f''(a) + \dots + \frac{(b-a)^{k+1}}{(k+1)!} f^{(k+1)}(a) + \frac{1}{(k+1)!} \int_a^b (b-t)^{k+1} f^{(k+2)}(t) dt \quad (11.224)$$

□C.Q.F.D.

2.5. SÉRIES DE FOURIER

Nous appelons par définition "**série trigonométrique**" une série de la forme:

$$a_0 \cdot \cos(0 \cdot x) + b_0 \cdot \sin(0 \cdot x) + a_1 \cdot \cos(x) + b_1 \cdot \sin(x) + \dots + a_n \cdot \cos(nx) + b_n \cdot \sin(nx) \quad (11.225)$$

ou sous une forme plus compacte:

$$\sum_{n=0}^{\infty} (a_n \cdot \cos(nx) + b_n \cdot \sin(nx)) \quad (11.226)$$

Les constantes a_n, b_n ($n = 1, 2, \dots \in \mathbb{N}^*$) sont les coefficients de la série trigonométrique plus souvent nommés "**coefficients de Fourier**".

Remarque: Nous avons déjà fait mention de ce type de série lors de notre étude des types de polynômes existants puisque les séries de Fourier ne sont au fait que des polynômes trigonométriques. (cf. chapitre de Calcul Algébrique). Par ailleurs, nous avons vu comme exemple dans le chapitre d'Analyse Fonctionnelle lors de notre étude du produit scalaire fonctionnel que les fonctions sinus et cosinus constituaient les bases d'un espace vectoriel.

Si la série converge, sa somme est une fonction périodique $f(x)$ de période $T = 2\pi$, étant donné que $\sin(nx)$ et $\cos(nx)$ sont des fonctions périodiques de période 2π . De sorte que:

$$f(x) = f(x + 2\pi) \quad (11.227)$$

Posons maintenant le problème suivant: Nous nous donnons une fonction connue, périodique quelconque $f(x)$ continue par morceaux de période 2π . Nous nous demandons s'il existe une série trigonométrique convergeant vers $f(x)$ moyennant des conditions sur cette série.

Supposons maintenant que la fonction $f(x)$, périodique et de période 2π , puisse être effectivement représentée par une série trigonométrique convergeant vers $f(x)$ dans l'intervalle $[0, T]$, c'est-à-dire qu'elle soit la somme de cette série:

$$f(x) = \sum_{n=0}^{+\infty} (a_n \cdot \cos(nx) + b_n \cdot \sin(nx)) \quad (11.228)$$

Supposons que l'intégrale de la fonction du premier membre de cette égalité soit égale à la somme des intégrales des termes de la série ci-dessus. Ceci aura lieu, par exemple, si nous supposons que la série trigonométrique proposée converge absolument, c'est-à-dire que la série numérique suivante converge (de par la propriété bornée des fonctions trigonométriques):

$$|a_0| + |b_0| + |a_1| + |b_1| + |a_2| + |b_2| + \dots + |a_n| + |b_n| \quad (11.229)$$

La série:

$$\sum_{n=0}^{\infty} (a_n \cdot \cos(nx) + b_n \cdot \sin(nx)) \quad (11.230)$$

est alors majorable et peut être intégrée terme à terme de 0 à T (où $T = 2\pi$) ce qui nous permet de déterminer les différents coefficients de Fourier. Mais avant de commencer exposons les intégrales suivantes qui nous très seront utiles par la suite:

$$\left. \begin{array}{l} \int_0^T \cos(nx) \cdot \cos(kx) dx = 0 \\ \int_0^T \cos(nx) \cdot \sin(kx) dx = 0 \\ \int_0^T \sin(nx) \cdot \sin(kx) dx = 0 \end{array} \right\} (I) \quad \left. \begin{array}{l} \int_0^T \cos^2(kx) dx = T/2 \\ \int_0^T \cos(kx) \cdot \sin(kx) dx = 0 \\ \int_0^T \sin^2(kx) dx = T/2 \end{array} \right\} (II) \quad (11.231)$$

Avec $n, k \in \mathbb{N}$ et $n \neq k$ Avec $n, k \in \mathbb{N}$ et $n = k \neq 0$

Avant de continuer, démontrons la valeur que prennent ces six intégrales (suite à la demande des internautes). Mais d'abord, rappelons que comme $n, k \in \mathbb{N}$ alors:

$$(n+k) = w \in \mathbb{N} \text{ et } (n-k) = v \in \mathbb{Z} \quad (11.232)$$

1. Nous procédons en utilisant les relations trigonométriques remarquables (cf. chapitre de Trigonométrie) et les primitives des fonctions trigonométriques élémentaires (cf. chapitre de calcul Différentiel Et Intégral):

$$\begin{aligned} \int_0^T \cos(nx) \cos(kx) dx &= \frac{1}{2} \int_0^T \cos(nx+kx) + \cos(nx-kx) dx \\ &= \frac{1}{2} \left(\frac{\sin(nx+kx)}{n+k} - \frac{\sin(nx-kx)}{n-k} \right) \Big|_0^T = \frac{1}{2} \left(\frac{\sin((n+k)x)}{n+k} - \frac{\sin((n-k)x)}{n-k} \right) \Big|_0^T \\ &= \frac{1}{2} \left(\frac{\sin(wx)}{w} - \frac{\sin(vx)}{v} \right) \Big|_0^T = \frac{1}{2} \left(\frac{\sin(wT)}{w} - \frac{\sin(vT)}{v} \right) - \frac{1}{2} \left(\frac{\sin(w0)}{w} - \frac{\sin(v0)}{v} \right) \\ &= \frac{1}{2} \left(\frac{\sin(wT)}{w} - \frac{\sin(vT)}{v} \right) = \frac{1}{2} \left(\frac{0}{w} - \frac{0}{v} \right) = 0 \end{aligned} \quad (11.233)$$

car comme nous l'avons vu dans le chapitre de Trigonométrie $\sin(k\pi) = 0, k \in \mathbb{Z}$ et comme $T = 2\pi$, les deux différences précédentes ont tous les termes qui sont nuls tel qu'au final:

$$\int_0^T \cos(nx) \cos(kx) dx = 0 \quad (11.234)$$

2. Pour la deuxième intégrale, nous procédons selon les mêmes techniques et mêmes propriétés des fonctions trigonométriques:

$$\begin{aligned} \int_0^T \cos(nx) \sin(kx) dx &= \frac{1}{2} \int_0^T (\sin(nx+kx) + \sin(nx-kx)) dx \\ &= \frac{1}{2} \left(-\frac{\cos(nx+kx)}{n+k} - \frac{\cos(nx-kx)}{n-k} \right) \Big|_0^T = \frac{1}{2} \left(-\frac{\cos(wx)}{w} - \frac{\cos(vx)}{v} \right) \Big|_0^T \\ &= \frac{1}{2} \left(-\frac{\cos(wT)}{w} - \frac{\cos(vT)}{v} \right) - \frac{1}{2} \left(-\frac{\cos(w \cdot 0)}{w} - \frac{\cos(v \cdot 0)}{v} \right) \\ &= \frac{1}{2} \left(-\frac{1}{w} - \frac{1}{v} \right) - \frac{1}{2} \left(-\frac{1}{w} - \frac{1}{v} \right) = \frac{1}{2} \left[\left(-\frac{1}{w} - \frac{1}{v} \right) - \left(-\frac{1}{w} - \frac{1}{v} \right) \right] = 0 \end{aligned} \quad (11.235)$$

3. Et nous continuons ainsi pour la troisième, toujours selon les mêmes propriétés:

$$\begin{aligned}
\int_0^T \sin(nx) \sin(kx) dx &= \frac{1}{2} \int_0^T (\cos(nx - kx) - \cos(nx + kx)) dx \\
&= \frac{1}{2} \left(\frac{\sin(nx - kx)}{n - k} - \frac{\sin(nx + kx)}{n + k} \right) \Big|_0^T = \frac{1}{2} \left(\frac{\sin(vx)}{v} - \frac{\sin(wx)}{w} \right) \Big|_0^T \quad (11.236) \\
&= \frac{1}{2} \left(\frac{\sin(vT)}{v} - \frac{\sin(wT)}{w} \right) - \frac{1}{2} \left(\frac{\sin(v \cdot 0)}{v} - \frac{\sin(w \cdot 0)}{w} \right) = 0
\end{aligned}$$

4. Encore une fois selon les mêmes méthodes (cela devient routinier...) pour $k \neq 0$ d'abord:

$$\begin{aligned}
\int_0^T \cos^2(kx) dx &= \int_0^T \left(\frac{1 + \cos(2kx)}{2} \right) dx = \frac{1}{2} \int_0^T dx + \frac{1}{2} \int_0^T \cos(2kx) dx \\
&= \frac{1}{2} x \Big|_0^T + \frac{1}{4k} \sin(2kx) \Big|_0^T = \frac{1}{2} (T - 0) + \frac{1}{4k} (\sin(2kT) - \sin(2k \cdot 0)) \quad (11.237) \\
&= \frac{1}{2} T + \frac{1}{4k} (0 - 0) = \frac{T}{2}
\end{aligned}$$

et pour $k = 0$ il vient immédiatement:

$$\int_0^T \cos^2(0 \cdot x) dx = \int_0^T dx = T \quad (11.238)$$

5. Encore une fois... (bientôt au bout...) pour $k \neq 0$ d'abord:

$$\begin{aligned}
\int_0^T \cos(kx) \sin(kx) dx &= \frac{1}{2} \int_0^T (\sin(kx + kx) + \sin(kx - kx)) dx = \frac{1}{2} \int_0^T \sin(2kx) dx \\
&= \frac{1}{2} \left(-\frac{\cos(2kx)}{2k} \right) \Big|_0^T = -\frac{1}{4} \left(\frac{\cos(2kx)}{k} \right) \Big|_0^T = -\frac{1}{4} \left(\frac{\cos(2kT)}{k} - \frac{\cos(2k \cdot 0)}{k} \right) \quad (11.239) \\
&= -\frac{1}{4} \left(\frac{1}{k} - \frac{1}{k} \right) = 0
\end{aligned}$$

et pour $k = 0$ il vient immédiatement:

$$\int_0^T \cos(kx) \sin(kx) dx = 0 \quad (11.240)$$

6. Et enfin la dernière (...):

$$\begin{aligned}
\int_0^T \sin^2(kx) dx &= \int_0^T \left(\frac{1 - \cos(2kx)}{2} \right) dx = \frac{1}{2} \int_0^T dx - \frac{1}{2} \int_0^T \cos(2kx) dx \\
&= \frac{1}{2} x \Big|_0^T - \frac{1}{4k} \sin(2kx) \Big|_0^T = \frac{1}{2} (T - 0) - \frac{1}{4k} (\sin(2kT) - \sin(2k \cdot 0)) \quad (11.241) \\
&= \frac{1}{2} T - \frac{1}{4k} (0 - 0) = \frac{T}{2}
\end{aligned}$$

Ce petit travail fait, revenons maintenant à nos moutons... Pour déterminer les coefficients a_n multiplions les deux membres de l'égalité:

$$f(x) = \sum_{n=0}^{+\infty} (a_n \cdot \cos(nx) + b_n \cdot \sin(nx)) \quad (11.242)$$

par $\cos(kx)$:

$$f(x) \cos(kx) = \sum_{n=0}^{\infty} (a_n \cos(nx) \cos(kx) + b_n \sin(nx) \cos(kx)) \quad (11.243)$$

La série du second membre de l'égalité est majorable, étant donné que ses termes ne sont pas supérieurs en valeur absolue aux termes de la série positive convergente. Nous pouvons donc l'intégrer terme à terme sur tout segment borné de 0 à T :

$$\int_0^T f(x) \cos(kx) dx = \sum_{n=0}^{\infty} \left(a_n \int_0^T \cos(nx) \cos(kx) dx + b_n \int_0^T \sin(nx) \cos(kx) dx \right) \quad (11.244)$$

Nous avons démontré plus haut que quelles que soient les valeurs entières que prennent k ou n le deuxième terme de la parenthèse est toujours nul. Il ne reste alors plus que:

$$\int_0^T f(x) \cos(kx) dx = \sum_{n=0}^{\infty} a_n \int_0^T \cos(nx) \cos(kx) dx \quad (11.245)$$

Or, nous avons démontré plus haut que l'intégrale à droite est toujours nulle si n et k sont différents. Il ne reste alors que le cas où n et k sont égaux. C'est-à-dire:

$$\int_0^T f(x) \cos(kx) dx = a_k \int_0^T \cos(kx) \cos(kx) dx = a_k \int_0^T \cos^2(kx) dx \quad (11.246)$$

Dans cette situation, nous avons d'abord le cas particulier où k est nul. Dans ce cas:

$$\int_0^T f(x) dx = a_0 \int_0^T dx = a_0 T \quad (11.247)$$

Soit:

$$a_0 = \frac{1}{T} \int_0^T f(x) dx \quad (11.248)$$

Il est évident que le coefficient a_0 représente donc la moyenne du signal ou de sa composante continue si elle existe.

Dans le cas où k n'est pas nul, nous avons:

$$\int_0^T f(x) \cos(kx) dx = a_k \int_0^T \cos^2(kx) dx = a_k \frac{T}{2} \quad (11.249)$$

D'où nous tirons:

$$a_k = \frac{2}{T} \int_0^T f(x) \cos(kx) dx \quad (11.250)$$

Pour déterminer les coefficients b_n nous procédons de la même manière mais en multipliant cette fois-ci les deux membres de l'égalité par $\sin(kx)$:

$$f(x) \sin(kx) = \sum_{n=0}^{\infty} (a_n \cos(nx) \sin(kx) + b_n \sin(nx) \sin(kx)) \quad (11.251)$$

La série du second membre de l'égalité est majorable, étant donné que ses termes ne sont pas supérieurs en valeurs absolues aux termes de la série positive convergente. Nous pouvons donc l'intégrer terme à terme sur tout segment borné de 0 à T:

$$f(x) \sin(kx) = \sum_{n=0}^{\infty} \left(a_n \int_0^T \cos(nx) \sin(kx) dx + b_n \int_0^T \sin(nx) \sin(kx) dx \right) \quad (11.252)$$

Nous avons démontré plus haut que quelles que soient les valeurs entières que prennent k ou n le premier terme de la parenthèse est toujours nul. Il ne reste plus alors que:

$$\int_0^T f(x) \sin(kx) dx = \sum_{n=0}^{\infty} \left(b_n \int_0^T \sin(nx) \sin(kx) dx \right) \quad (11.253)$$

Or, nous avons démontré plus haut que l'intégrale à droite est toujours nulle si n et k sont différents. Il ne reste alors que le cas où n et k sont égaux. C'est-à-dire:

$$\int_0^T f(x) \sin(kx) dx = b_k \int_0^T \sin(kx) \sin(kx) dx = b_k \int_0^T \sin^2(kx) dx \quad (11.254)$$

Dans cette situation, nous avons d'abord le cas particulier où k est nul. Mais nous voyons de suite que nous avons une indétermination par zéro. Il vaut mieux alors considérer le cas général d'où nous tirons:

$$\int_0^T f(x) \sin(kx) dx = b_k \int_0^T \sin^2(kx) dx = b_k \frac{T}{2} \quad (11.255)$$

D'où nous tirons aisément que:

$$b_k = \frac{2}{T} \int_0^T f(x) \sin(kx) dx \quad (11.256)$$

Dès lors, pour la situation où k est nul le coefficient est alors nul!

Donc finalement les coefficients de Fourier sont donc déterminés par les intégrales:

$$a_0 = \frac{1}{T} \int_0^T f(x) dx \quad a_n = \frac{2}{T} \int_0^T f(x) \cos(nx) dx \quad b_n = \frac{2}{T} \int_0^T f(x) \sin(nx) dx \quad (11.257)$$

Mais comme c'est embêtant d'avoir trois résultats pour les coefficients nous allons jouer un peu avec la définition de la série de Fourier.

Effectivement en sommant de 1 à l'infini plutôt que de 0 à l'infini nous avons:

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{+\infty} (a_n \cos(nx) + b_n \sin(nx)) \quad (11.258)$$

Ce qui permet alors de n'avoir qu'à se rappeler de (a_0 inclus donc!):

$$a_n = \frac{2}{T} \int_0^T f(x) \cos(nx) dx \quad b_n = \frac{2}{T} \int_0^T f(x) \sin(nx) dx \quad (11.259)$$

Les physiciens ont quant à eux pour habitude de noter ces deux dernières relations sous la forme suivante:

$$\begin{aligned} a_k &= \frac{2}{T} \int_{-T/2}^{T/2} f(t) \cos\left(\frac{2\pi}{T} kt\right) dt = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \cos(\omega kt) dt \\ b_k &= \frac{2}{T} \int_{-T/2}^{T/2} f(t) \sin\left(\frac{2\pi}{T} kt\right) dt = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \sin(\omega kt) dt \end{aligned} \quad (11.260)$$

Cette décomposition possible de toute fonction périodique continue par morceaux approchée par une somme infinie de fonctions trigonométriques (sinus ou cosinus) consistant en une fonction fondamentale et ses harmoniques est appelée "**théorème de Fourier**" ou encore "**théorème de Fourier-Dirichlet**".

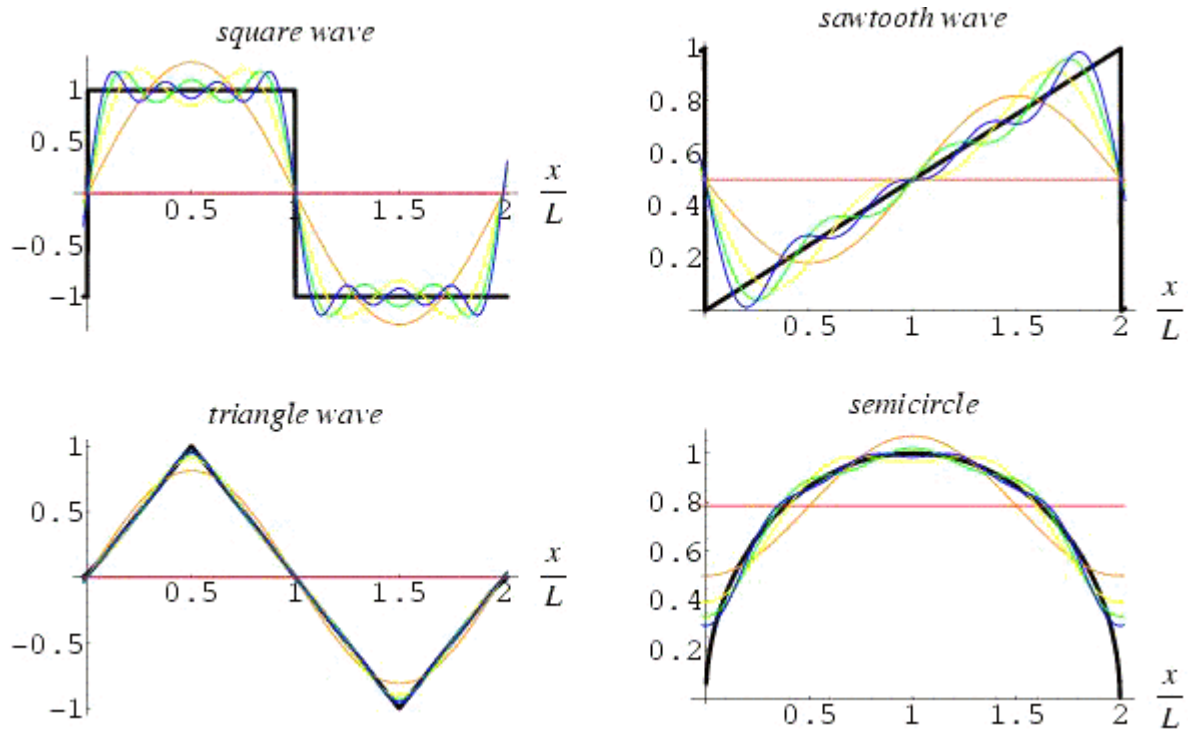


Figure: 11.7 - Exemples d'approches de fonctions par série de Fourier (source: Mathworld)

La série de Fourier permet donc implicitement de représenter toutes les fréquences contenues dans un signal périodique dont la fonction est connue mathématiquement. On se demande bien pourquoi parler des séries de Fourier quand, dans la pratique, nous ne connaissons pas vraiment la représentation mathématique de ce signal? Cela nous amènera à mieux comprendre le concept de la transformée de Fourier à temps discret, que nous verrons un peu plus loin, qui n'a nul besoin d'une représentation mathématique d'un signal continu échantillonné dans le temps.

Nous constatons par ailleurs que si $f(x)$, soit la fonction périodique dont nous cherchons l'expression en série trigonométrique de Fourier, est paire alors la série devra être paire aussi et donc ne comporter que des termes en cosinus (le cosinus étant pour rappel une fonction paire) ce qui implique que $b_n = 0$ et dans le cas contraire d'une fonction impaire $a_n = 0$ (le sinus étant pour rappel une fonction impaire)!

Il convient de noter, et c'est important pour la suite, que comme nous l'avons vu dans le chapitre de Calcul Algébrique lors de notre étude des polynômes trigonométriques, les séries de Fourier pouvaient donc s'écrire sous la forme complexe suivante (en changeant un peu les notations et en passant la somme à l'infini):

$$f(x) = \sum_{n=-\infty}^{+\infty} c_n e^{inx} \quad (11.261)$$

et nous avons vus que:

$$a_0 = 2c_0 \quad a_n = c_n + c_{-n} \quad b_n = i(c_n - c_{-n}) \quad (8.262)$$

Soit:

$$c_n = \frac{a_n - ib_n}{2} \quad (11.263)$$

Ce qui nous donne:

$$c_n = \frac{\frac{2}{T} \int_0^T f(x) \cos(nx) dx - \frac{2}{T} i \int_0^T f(x) \sin(nx) dx}{2} = \frac{1}{T} \int_0^T f(x) (\cos(nx) - i \sin(nx)) dx \quad (11.264)$$

$$= \frac{1}{T} \int_0^T f(x) e^{-inx} dx$$

Donc:

$$c_n = \frac{1}{T} \int_0^T f(x) e^{-inx} dx \quad (11.265)$$

Exemples:

E1. Lors de la décomposition d'un signal continu, nous disons abusivement que les coefficients a_n, b_n représentent chacun (implicitement) une fréquence distincte associée à une amplitude que nous visualisons sur un graphique par des lignes verticales. Ce graphique représente le spectre en fréquence du signal décomposé. Nous pouvons également adjoindre une autre représentation qui se nomme "**spectre de phase**". Ce spectre nous donne la phase du signal harmonique (en avance ou en retard de phase).

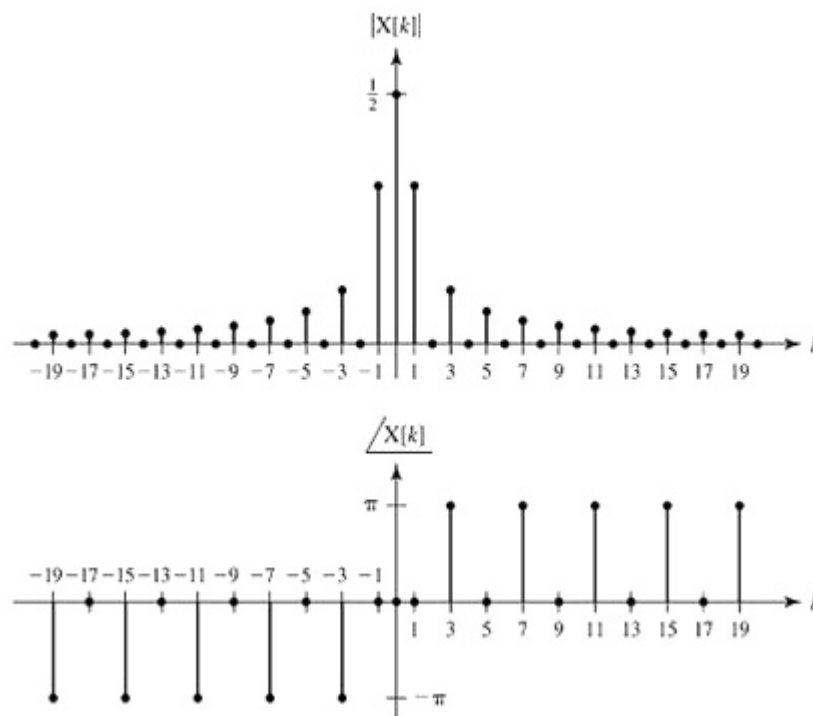


Figure: 11.8 - Exemple d'amplitudes et de fréquences associées aux différents coefficients

Voyons maintenant comment décomposer un signal périodique connu en plusieurs signaux d'amplitudes et de fréquences distinctes.

Prenons comme exemple, un signal à onde carrée périodique défini sur une période $T=2$ et d'amplitude A tel que:

$$f(t) = \begin{cases} -A & -1 < t < 0 \\ A & 0 < t < 1 \end{cases} \quad (11.266)$$

A la période $T=2$ correspond comme nous le savons une pulsation:

$$\omega = \frac{2\pi}{T} = \pi \quad (11.267)$$

Calculons en premier lieu les coefficients c_k à l'aide de l'intégrale permettant de déterminer ces coefficients (le choix des bornes de l'intégrale suppose donc que le signal est périodique par construction!) :

$$\begin{aligned} c_k &= \frac{1}{T} \int_T f(t) e^{-ik\omega t} dt = \frac{1}{2} \int_{-1}^1 f(t) e^{-ik\pi t} dt = \frac{1}{2} \left(\int_{-1}^0 -A e^{-ik\pi t} dt + \int_0^1 A e^{-ik\pi t} dt \right) \\ c_k &= \frac{-A}{2} \left(\int_{-1}^0 e^{-ik\pi t} dt - \int_0^1 e^{-ik\pi t} dt \right) = \frac{-A}{2} \left(\frac{-1}{ik\pi} \right) \left[e^{-ik\pi t} \right]_{-1}^0 + \frac{A}{2} \left(\frac{-1}{ik\pi} \right) \left[e^{-ik\pi t} \right]_0^1 \\ c_k &= \frac{A}{2ik\pi} [1 - e^{ik\pi}] - \frac{A}{2ik\pi} [e^{-ik\pi} - 1] = \frac{A}{2} \left[\frac{1 - e^{ik\pi}}{ik\pi} + \frac{e^{-ik\pi} - 1}{-ik\pi} \right] \\ c_k &= \frac{A}{ik\pi} \left[1 - \frac{1}{2} (e^{ik\pi} + e^{-ik\pi}) \right] \end{aligned} \quad (11.268)$$

En prenant $k=2$, nous avons:

$$c_2 = \frac{A}{i2\pi} \left[1 - \frac{1}{2} (e^{i2\pi} + e^{-i2\pi}) \right] = 0 \quad (11.269)$$

De même pour $k=4,6,8$ ainsi que pour tout nombre pair.

Pour ce qui est des nombres impairs, nous aurons:

$$c_1 = \frac{A}{i\pi} \left[1 - \frac{1}{2} (e^{i\pi} + e^{-i\pi}) \right] = \frac{2A}{i\pi}, \quad c_3 = \frac{A}{3i\pi} \left[1 - \frac{1}{2} (e^{i3\pi} + e^{-i3\pi}) \right] = \frac{2A}{i3\pi} \quad (11.270)$$

Les coefficients seront alors:

$$c_k = \begin{cases} 0 & k \text{ pair} \\ \frac{2A}{ik\pi} & k \text{ impair} \end{cases} \quad (11.271)$$

Il y a un seul hic dans cette relation, le coefficient c_0 ne peut être calculé selon cette relation car on peut voir que si $k=0$ dans le résultat ci-haut, nous aurons une valeur infinie et c'est du moins impossible. Le coefficient c_0 est soit nul ou non nul mais jamais infini.

Pour trouver le coefficient c_0 , nous devons calculer l'intégrale pour $k=0$. Le coefficient c_0 est alors déterminé par:

$$c_0 = \frac{1}{T} \int_T f(t) e^{-ik\omega t} dt = \frac{1}{2} \int_{-1}^1 f(t) dt = \frac{1}{2} \int_{-1}^0 f(t) dt + \frac{1}{2} \int_0^1 f(t) dt \quad (11.272)$$

$$c_0 = \frac{1}{2} \int_{-1}^0 (-A) dt + \frac{1}{2} \int_0^1 (A) dt = 0$$

Le spectre en "fréquence" (attention à l'abus de langage!) et en amplitude sera alors de la forme suivante pour $k = -5, \dots, +5$ et $A = 1$ les fréquences nulles n'étant pas représentées:

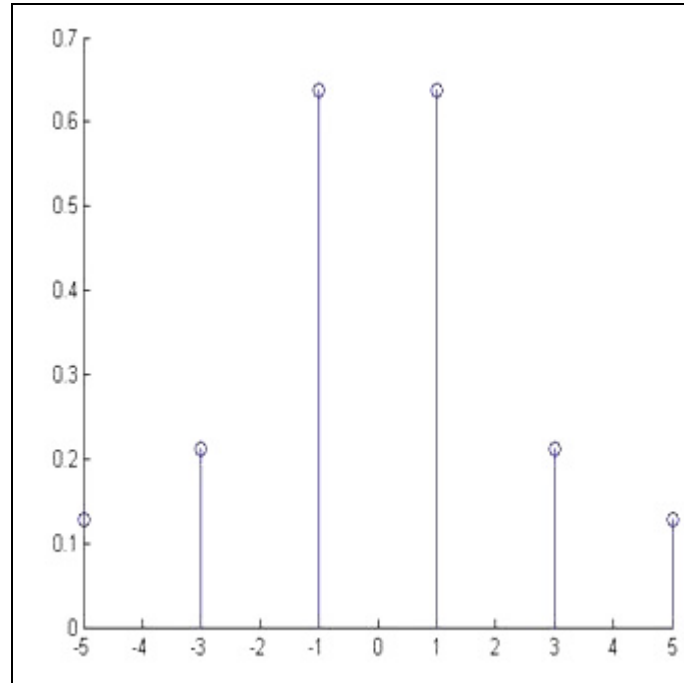


Figure: 11.9 - Spectre de fréquence des coefficients de la série de Fourier

L'abus de parler de fréquences pour les coefficients de Fourier amène donc à avoir des fréquences négatives en abscisse... mais ce n'est qu'une question de vocabulaire (il n'y a aucun rapport direct avec les vraies fréquences) auquel il faut s'habituer.

Le spectre d'amplitude et de phase se calcule selon les relations:

$$|c_k| = \sqrt{(\Re(c_k))^2 + (\Im(c_k))^2} \quad \text{et} \quad \angle \theta = \tan^{-1} \left(\frac{\Im(c_k)}{\Re(c_k)} \right) \quad (11.273)$$

Il est alors relativement aisé de remarquer que si T tend vers un nombre de plus en plus grand, les pics du spectre se rapprochent de plus en plus. Ainsi, lorsque T tend vers l'infini le spectre devient continu.

Le spectre de phase donnera ce qui suit pour les valeurs impaires:

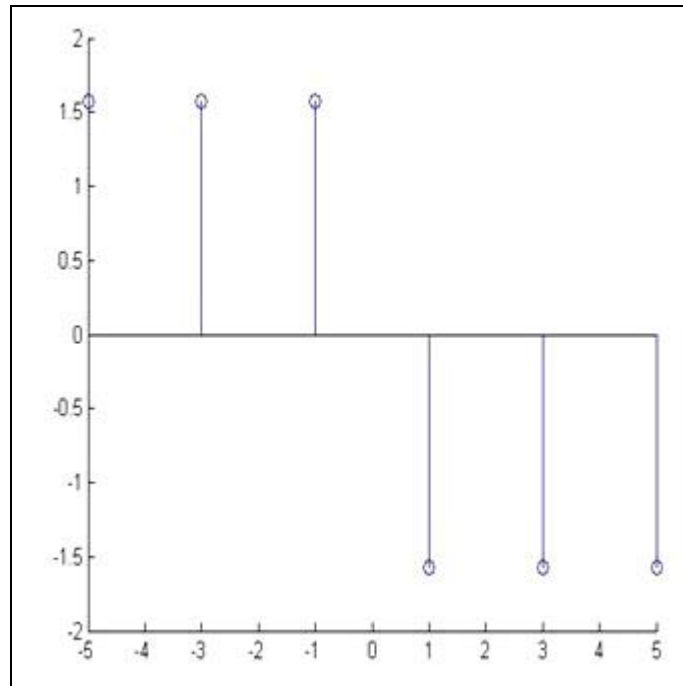


Figure: 11.10 - Spectre de phase des coefficients de la série de Fourier

Il est même possible pour l'exemple d'obtenir relativement facilement le spectre des fréquences dans Microsoft Excel 11.8346 (le lecteur pourra trouver un exemple beaucoup plus détaillé et intéressant dans le serveur d'exercices sur les Suites et Séries)!!!

Effectivement, il suffit pour cela d'échantillonner par exemple notre signal 128 fois (Microsoft Excel 11.8346 a besoin de 2^n échantillons et ne fonctionne que sous cette condition!). Nous divisons alors l'intervalle $-1 < t < 0$ en 64 échantillons et idem pour l'intervalle $0 < t < 1$:

	A	B
1	t	A
2	-1	-1
3	-0.9844	-1
4	-0.9688	-1
5	-0.9531	-1
6	-0.9375	-1
7	-0.9219	-1
8	-0.9063	-1
9	-0.8906	-1
10	-0.875	-1
11	-0.8594	-1
12	-0.8438	-1
13	-0.8281	-1
14	-0.8125	-1
15	-0.7969	-1
16	-0.7813	-1
17	-0.7656	-1
18	-0.75	-1
19	-0.7344	-1
20	-0.7188	-1
21	-0.7031	-1
22	-0.6875	-1
23	-0.6719	-1
24	-0.6563	-1
25	-0.6406	-1
26	-0.625	-1
27	-0.6094	-1
28	-0.5938	-1
29	-0.5781	-1
30	-0.5625	-1
31	-0.5469	-1
32	-0.5313	-1
33	-0.5156	-1
34	-0.5	-1
35	-0.4844	-1
36	-0.4688	-1
37	-0.4531	-1
38	-0.4375	-1
39	-0.4219	-1
40	-0.4063	-1
41	-0.3906	-1
42	-0.375	-1
43	-0.3594	-1
44	-0.3438	-1
45	-0.3281	-1
46	-0.3125	-1
47	-0.2969	-1
48	-0.2813	-1
49	-0.2656	-1
50	-0.25	-1
51	-0.2344	-1
52	-0.2188	-1
53	-0.2031	-1
54	-0.1875	-1
55	-0.1719	-1
56	-0.1563	-1
57	-0.1406	-1
58	-0.125	-1
59	-0.1094	-1
60	-0.0938	-1
61	-0.0781	-1
62	-0.0625	-1
63	-0.0469	-1
64	-0.0313	-1
65	-0.0156	-1
66	0.01563	1
67	0.03125	1
68	0.04688	1
69	0.0625	1
70	0.07813	1
71	0.09375	1
72	0.10938	1
73	0.125	1
74	0.14063	1
75	0.15625	1
76	0.17188	1
77	0.1875	1
78	0.20313	1
79	0.21875	1
80	0.23438	1
81	0.25	1
82	0.26563	1
83	0.28125	1
84	0.29688	1
85	0.3125	1
86	0.32813	1
87	0.34375	1
88	0.35938	1
89	0.375	1
90	0.39063	1
91	0.40625	1
92	0.42188	1
93	0.4375	1
94	0.45313	1
95	0.46875	1
96	0.48438	1
97	0.5	1
98	0.51563	1
99	0.53125	1
100	0.54688	1
101	0.5625	1
102	0.57813	1
103	0.59375	1
104	0.60938	1
105	0.625	1
106	0.64063	1
107	0.65625	1
108	0.67188	1
109	0.6875	1
110	0.70313	1
111	0.71875	1
112	0.73438	1
113	0.75	1
114	0.76563	1
115	0.78125	1
116	0.79688	1
117	0.8125	1
118	0.82813	1
119	0.84375	1
120	0.85938	1
121	0.875	1
122	0.89063	1
123	0.90625	1
124	0.92188	1
125	0.9375	1
126	0.95313	1
127	0.96875	1
128	0.98438	1
129	1	1

Tableau: 11.2 - Echantillonnage signal

Ce qui donne sous forme graphique (attention! pour que la transformée de Fourier discrète fonctionne bien dans Microsoft Excel 11.8346, il faut que la fréquence d'échantillonnage - correspondant au nombre de mesures dans une seconde - soit au moins 100 fois supérieure à la fréquence du signal d'origine sinon quoi le résultat peut être aberrant!):

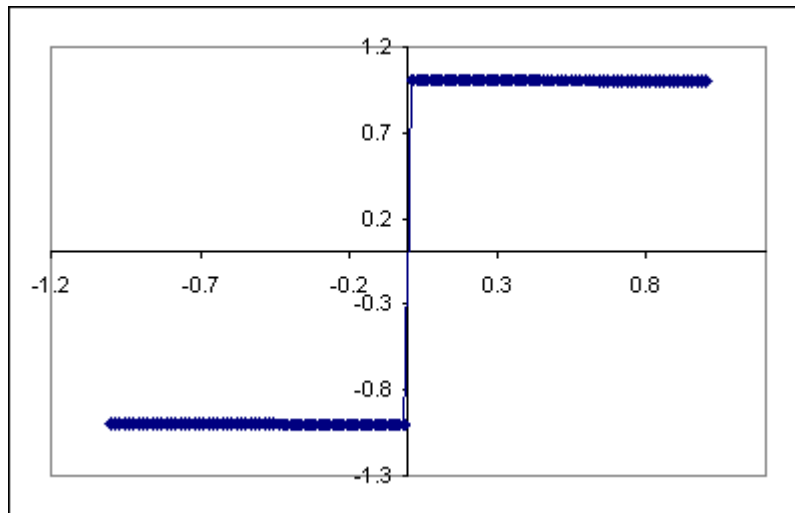


Figure: 11.11 - Représentation graphique de la série de données sous Microsoft Excel 11.8346

Ensuite, il suffit d'aller dans le menu Outils/Utilitaire d'Analyse et choisir l'option Analyse de Fourier:

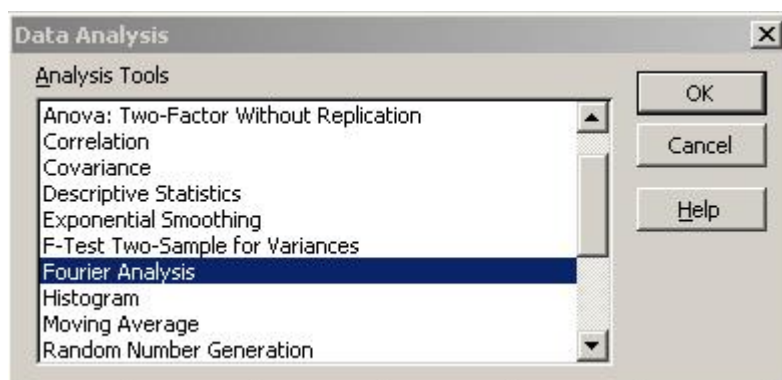


Figure: 11.12 - Caputre de la boîte de dialogue de l'utilitaire d'analyse de Microsoft Excel 11.8346

Vient ensuite la boîte de dialogue suivante qu'il faut remplir comme indiqué (on voit que l'abscisse n'a aucune importance!):

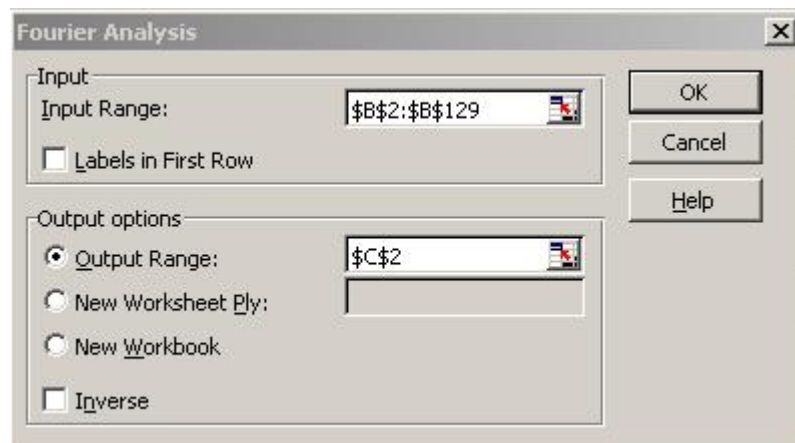


Figure: 11.13 - Paramètres de l'outil Analyse de Fourier de Microsoft Excel 11.8346

Vient alors le tableau suivant pour les coefficients:

	A	B	C
1	t	A	C_n
2	-1	-1	0
3	-0.9844	-1	-1.999999999999994+81.4709677441666i
4	-0.9688	-1	0
5	-0.9531	-1	-1.999999999999996+27.1133384847048i
6	-0.9375	-1	0
7	-0.9219	-1	-1.999999999999997+16.2155716073538i
8	-0.9063	-1	0
9	-0.8906	-1	-1.999999999999997+11.5262840102376i
10	-0.875	-1	0
11	-0.8594	-1	-1.999999999999998+8.9064044488288i
12	-0.8438	-1	0
13	-0.8282	-1	-1.999999999999998+6.2155716073538i
14	-0.8125	-1	0
15	-0.7969	-1	-1.999999999999999+3.6077381178546E-002i
16	-0.7813	-1	0
17	-0.7656	-1	-1.999999999999999+2.2707136261486i
18	-0.75	-1	0
19	-0.7344	-1	-1.999999999999999+1.4709677441667i
20	-0.7188	-1	0
21	-0.7031	-1	-1.999999999999999+0.90972442178759E-001i
22	-0.6875	-1	0
23	-0.6719	-1	-1.999999999999999+5.66401557160735E-002i
24	-0.6563	-1	0
25	-0.6406	-1	-1.999999999999999+3.43750000000000E-002i
26	-0.625	-1	0
27	-0.6094	-1	-1.999999999999999+2.12133384847049E-002i
28	-0.5938	-1	0
29	-0.5782	-1	-1.999999999999999+1.31333848470491E-002i
30	-0.5625	-1	0
31	-0.5469	-1	-1.999999999999999+8.43750000000000E-003i
32	-0.5313	-1	0
33	-0.5156	-1	-1.999999999999999+5.2628401023761E-003i
34	-0.5	-1	0
35	-0.4844	-1	-1.999999999999999+3.16666666666667E-003i
36	-0.4688	-1	0
37	-0.4531	-1	-1.999999999999999+1.97916666666667E-003i
38	-0.4375	-1	0
39	-0.4219	-1	-1.999999999999999+1.23809523809524E-003i
40	-0.4063	-1	0
41	-0.3906	-1	-1.999999999999999+7.69230769230769E-004i
42	-0.375	-1	0
43	-0.3594	-1	-1.999999999999999+4.79166666666667E-004i
44	-0.3438	-1	0
45	-0.3282	-1	-1.999999999999999+2.99074074074074E-004i
46	-0.3125	-1	0
47	-0.2969	-1	-1.999999999999999+1.86956521739130E-004i
48	-0.2813	-1	0
49	-0.2656	-1	-1.999999999999999+1.16770752941176E-004i
50	-0.25	-1	0
51	-0.2344	-1	-1.999999999999999+7.22707136261486E-005i
52	-0.2188	-1	0
53	-0.2031	-1	-1.999999999999999+4.51677075294118E-005i
54	-0.1875	-1	0
55	-0.1719	-1	-1.999999999999999+2.76148603015075E-005i
56	-0.1563	-1	0
57	-0.1406	-1	-1.999999999999999+1.72614860301508E-005i
58	-0.125	-1	0
59	-0.1094	-1	-1.999999999999999+1.07812500000000E-005i
60	-0.0938	-1	0
61	-0.0782	-1	-1.999999999999999+6.73913043478261E-006i
62	-0.0625	-1	0
63	-0.0469	-1	-1.999999999999998+0.147528863044911i
64	-0.0313	-1	0
65	-0.0156	-1	-1.999999999999993+4.90972442178759E-001i
66	0.01563	1	0
67	0.03125	1	-2-4.90972442178546E-002i
68	0.04688	1	0
69	0.0625	1	-2-0.147528863044895i
70	0.07813	1	0
71	0.09375	1	-2-0.246676472273474i
72	0.10938	1	0
73	0.125	1	-2-0.390625000000000i
74	0.14063	1	0
75	0.15625	1	-2-0.617187500000000i
76	0.17188	1	0
77	0.1875	1	-2-1.15262840102377i
78	0.20313	1	0
79	0.21875	1	-2-2.15571607353811i
80	0.23438	1	0
81	0.25	1	-2-3.90640444882881i
82	0.26563	1	0
83	0.28125	1	-2-6.21557160735381i
84	0.29688	1	0
85	0.3125	1	-2-10.9999999999999i
86	0.32813	1	0
87	0.34375	1	-2-19.9999999999999i
88	0.35938	1	0
89	0.375	1	-2-35.9999999999999i
90	0.39063	1	0
91	0.40625	1	-2-63.9999999999999i
92	0.42188	1	0
93	0.4375	1	-2-111.999999999999i
94	0.45313	1	0
95	0.46875	1	-2-199.999999999999i
96	0.48438	1	0
97	0.5	1	-2-359.999999999999i
98	0.51563	1	0
99	0.53125	1	-2-639.999999999999i
100	0.54688	1	0
101	0.5625	1	-2-1119.99999999999i
102	0.57813	1	0
103	0.59375	1	-2-1999.99999999999i
104	0.60938	1	0
105	0.625	1	-2-3599.99999999999i
106	0.64063	1	0
107	0.65625	1	-2-6399.99999999999i
108	0.67188	1	0
109	0.6875	1	-2-11199.9999999999i
110	0.70313	1	0
111	0.71875	1	-2-19999.9999999999i
112	0.73438	1	0
113	0.75	1	-2-35999.9999999999i
114	0.76563	1	0
115	0.78125	1	-2-63999.9999999999i
116	0.79688	1	0
117	0.8125	1	-2-111999.999999999i
118	0.82813	1	0
119	0.84375	1	-2-199999.999999999i
120	0.85938	1	0
121	0.875	1	-2-359999.999999999i
122	0.89063	1	0
123	0.90625	1	-2-639999.999999999i
124	0.92188	1	0
125	0.9375	1	-2-1119999.99999999i
126	0.95313	1	0
127	0.96875	1	-2-1999999.99999999i
128	0.98438	1	0
129	1	1	-2-3599999.99999999i

Tableau: 11.3 - Coefficients de Fourier

Il reste à calculer le module des nombres complexes avec la fonction `MODULE.COMPLEXE()` de Microsoft Excel 11.8346 et de diviser le résultat par 128 pour chacun des coefficients c_n mais nous voyons déjà que chaque coefficient pair est nul ce qui correspond bien au résultat théorique obtenu plus haut.

Nous avons alors en mettant l'indice n en face de chaque module:

	A	B	C	D	E
1	t	A	C_n	$ C_n $	n
2	-1	-1	0	0	0
3	-0.9844	-1	-1.99999999999994+81.4709677441666i	0.6366837	-1
4	-0.9688	-1	0	0	-2
5	-0.9531	-1	-1.99999999999996+27.1133384847048i	0.2123985	-3
6	-0.9375	-1	0	0	-4
7	-0.9219	-1	-1.99999999999997+16.2155716073538i	0.1276441	-5
8	-0.9063	-1	0	0	-6
9	-0.8906	-1	-1.99999999999997+11.5262840102376i	0.0913946	-7
10	-0.875	-1	0	0	-8
60	-0.0938	-1	0	0	-58
61	-0.0781	-1	-1.99999999999999+0.246676472273487i	0.0157434	-59
62	-0.0625	-1	0	0	-60
63	-0.0469	-1	-1.99999999999998+0.147528863044911i	0.0156675	-61
64	-0.0313	-1	0	0	-62
65	-0.0156	-1	-1.99999999999993+4.90972442178759E-	0.0156297	-63
66	0.01563	1	0	0	64
67	0.03125	1	-2-4.90972442178546E-002i	0.0156297	63
68	0.04688	1	0	0	62
69	0.0625	1	-2-0.147528863044895i	0.0156675	61
70	0.07813	1	0	0	60
71	0.09375	1	-2-0.246676472273474i	0.0157434	59
121	0.875	1	-2.000000000000003-8.9064044488288i	0.0713141	9
122	0.89063	1	0	0	8
123	0.90625	1	-2.000000000000003-11.5262840102376i	0.0913946	7
124	0.92188	1	0	0	6
125	0.9375	1	-2.000000000000005-16.2155716073538i	0.1276441	5
126	0.95313	1	0	0	4
127	0.96875	1	-2.000000000000009-27.1133384847049i	0.2123985	3
128	0.98438	1	0	0	2
129	1	1	-2.000000000000032-81.4709677441667i	0.6366837	1

Tableau: 11.4 - Module coefficients complexes

En en traçant un graphique à points (toujours dans Microsoft Excel 11.8346) un peu personnalisé des colonnes *D* et *E*, nous obtenons finalement (nous avons restreint l'axe des abscisses à [-5,+5] pour en faciliter la lecture):

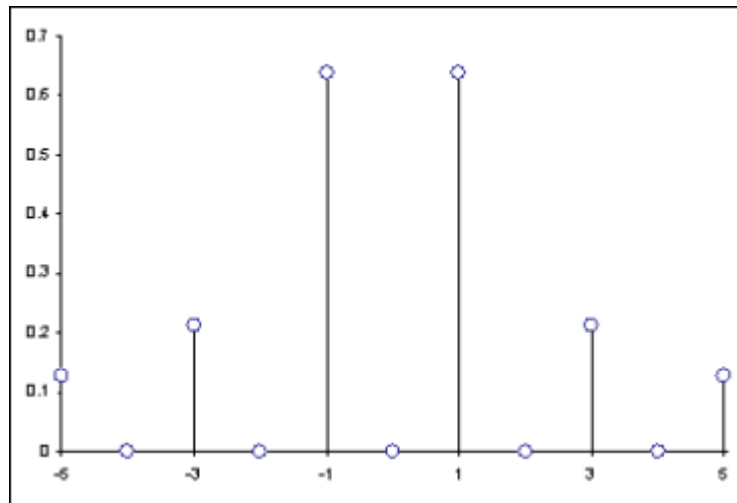


Figure: 11.14 - Spectre de fréquences de la transformée

A comparer avec les calculs théoriques (graphique déjà présenté plus haut)...

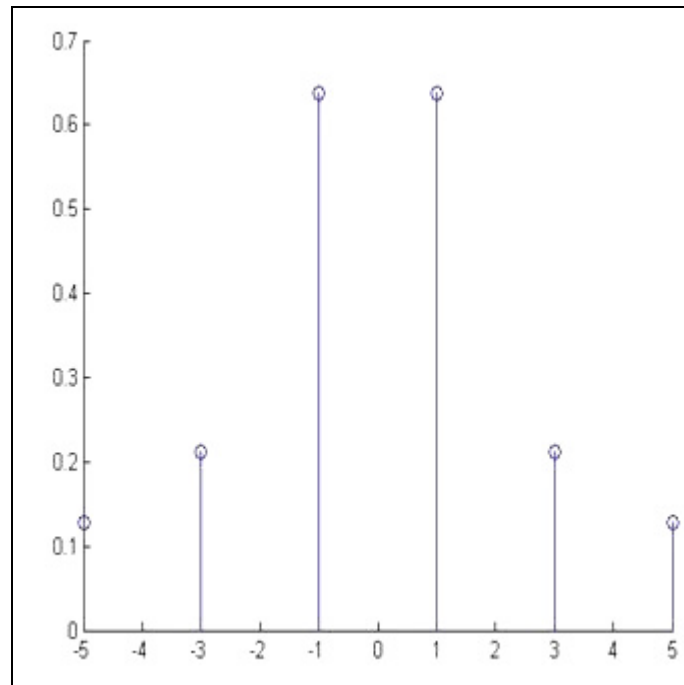


Figure: 11.15 - Spectre de fréquence calculé plus haut à la main

Nous verrons un exemple pratique et plus détaillé dans le chapitre d'Économie lors de notre étude des séries temporelles.

E2. Prenons un autre exemple identique au précédent mais sous une autre approche. Nous définissons une fonction périodique de période 2π comme suit:

$$f(x) = \begin{cases} -1 & -\pi < x < 0 \\ 1 & 0 \leq x \leq \pi \end{cases} \quad (11.274)$$

Calculons les coefficients de Fourier (nous translatons les bornes des intégrales puisque la fonction est périodique rien ne nous en empêche!):

$$\begin{aligned}
 a_n &= \frac{2}{2\pi} \int_0^{2\pi} f(x) \cos(nx) dx = \frac{1}{\pi} \left(\int_0^{\pi} 1 \cos(nx) dx + \int_{\pi}^{2\pi} -1 \cos(nx) dx \right) \\
 &= \frac{1}{\pi} \left(\int_0^{\pi} \cos(nx) dx - \int_{\pi}^{2\pi} \cos(nx) dx \right) = \frac{1}{\pi} \left(\frac{\sin(nx)}{n} \Big|_0^{\pi} - \frac{\sin(nx)}{n} \Big|_{\pi}^{2\pi} \right) = 0
 \end{aligned} \tag{11.275}$$

et:

$$\begin{aligned}
 b_n &= \frac{2}{2\pi} \int_0^{2\pi} f(x) \sin(nx) dx = \frac{1}{\pi} \left(\int_0^{\pi} 1 \sin(nx) dx + \int_{\pi}^{2\pi} -1 \sin(nx) dx \right) \\
 &= \frac{1}{\pi} \left(\int_0^{\pi} \sin(nx) dx - \int_{\pi}^{2\pi} \sin(nx) dx \right) = \frac{1}{\pi} \left(-\frac{\cos(nx)}{n} \Big|_0^{\pi} + \frac{\cos(nx)}{n} \Big|_{\pi}^{2\pi} \right) \\
 &= \frac{1}{\pi} \left(\frac{\cos(nx)}{n} \Big|_{\pi}^{2\pi} - \frac{\cos(nx)}{n} \Big|_0^{\pi} \right) = \frac{1}{\pi n} \left(\cos(nx) \Big|_{\pi}^{2\pi} - \cos(nx) \Big|_0^{\pi} \right) \\
 &= \frac{1}{\pi n} (\cos(2\pi n) - \cos(\pi n) - \cos(\pi n) + 1) \\
 &= \frac{1}{\pi n} (\cos(2\pi n) - 2\cos(\pi n) + 1)
 \end{aligned} \tag{11.276}$$

Nous remarquons que b_n vaut 0 pour n pair et vaut $4\pi/n$ pour n impair.

La série de Fourier de la fonction considérée s'écrit donc:

$$\begin{aligned}
 f(x) &= \sum_{n=0}^{+\infty} (a_n \cos(nx) + b_n \sin(nx)) = \sum_{n=0}^{+\infty} b_n \sin(nx) \\
 &= \frac{4}{\pi} \left[\frac{\sin(x)}{1} + \frac{\sin(3x)}{3} + \frac{\sin(5x)}{5} + \dots + \frac{\sin(2x+1)}{2x+1} + \dots \right]
 \end{aligned} \tag{11.277}$$

Ce qui dans Maple 4.00b s'écrit:

```
>S:=(4/Pi)*Sum(sin((2*n+1)*x)/(2*n+1),n=0..N);
```

et que nous pouvons tracer à l'aide de la fonction:

```
>plot({subs(N=4,S),subs(N=8,S),subs(N=16,S)},
x=-Pi..Pi,color=[red,green,blue],numpoints=200);
```

Ce qui donne trois traces pour 4, 8 et 16 termes de la série en rouge, vert et bleu:

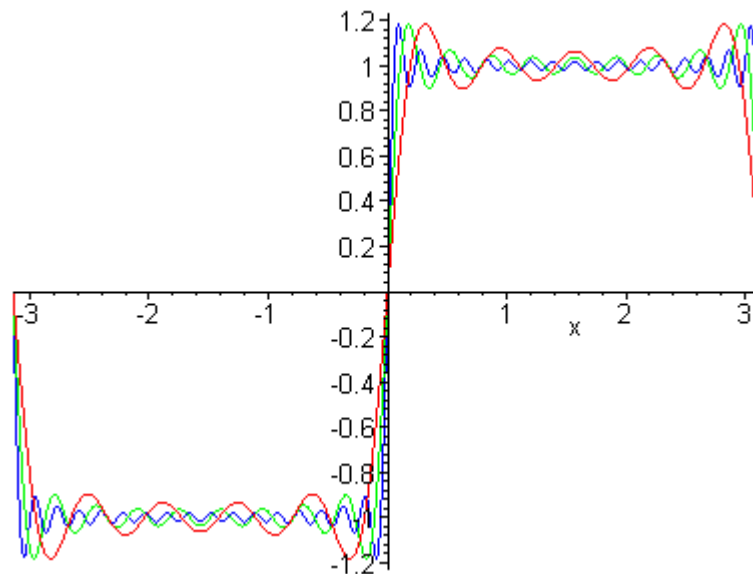


Figure: 11.16 - Exemple de série de Fourier dans Maple 4.00b avec 4, 8 et 16 termes

Pour 50 termes nous obtenons:

```
> plot(subs(N=50,S),x=-Pi..Pi,numpoints=800);
```

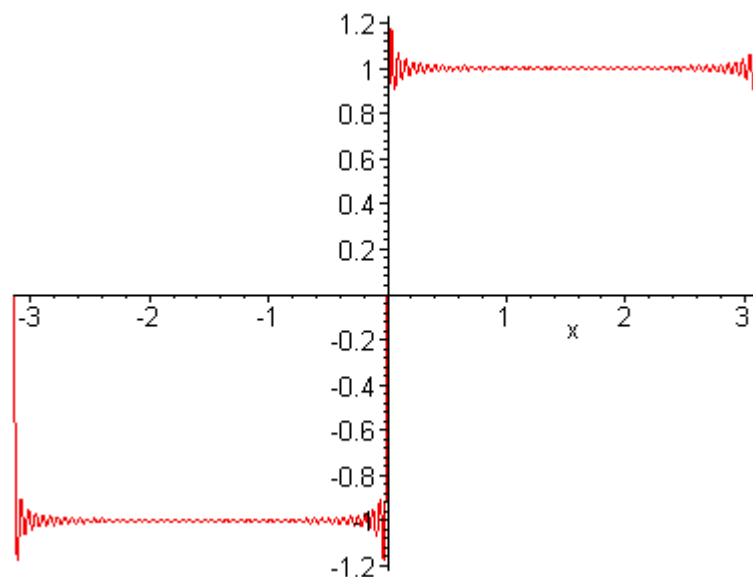


Figure: 11.17 - Exemple de série de Fourier dans Maple 4.00b avec 50 termes

Nous voyons les effets de bord appelés "**phénomène de Gibbs**". Il est possible de montrer ceux-ci se produisent à la valeur de l'abscisse correspondant à $x = \pi / 2n$ et que le pic culmine à ± 1.179 pour toute valeur de n . Voyons cela!

Nous venons donc de montrer que:

$$\begin{aligned}
 f(x) &= \sum_{n=0}^{+\infty} (a_n \cos(nx) + b_n \sin(nx)) = \sum_{n=0}^{+\infty} b_n \sin(nx) \\
 &= \frac{4}{\pi} \left[\frac{\sin(x)}{1} + \frac{\sin(3x)}{3} + \frac{\sin(5x)}{5} + \dots + \frac{\sin(2x+1)}{2x+1} + \dots \right] \quad (11.278)
 \end{aligned}$$

Ce qui peut s'écrire:

$$\begin{aligned}
 f(x) &= \frac{4}{\pi} \int_0^x [\cos(t) + \cos(3t) + \dots + \cos(2n+1)t] dt \\
 &= \frac{4}{\pi} \int_0^x \Re \left[e^{it} + e^{i3t} + \dots + e^{i(2n+1)t} \right] dt
 \end{aligned}$$

en utilisant la démonstration faite plus haut comme quoi:

$$S_n = x + x^3 + x^5 + \dots + x^{2n-1} = \frac{x^{2n+1} - x}{x^2 - 1}$$

Nous avons alors:

$$\begin{aligned}
 f(x) &= \frac{4}{\pi} \int_0^x \Re \left[e^{it} + e^{i3t} + \dots + e^{i(2n+1)t} \right] dt \\
 &= \frac{4}{\pi} \int_0^x \Re \left[\frac{e^{(2n+1)it} - e^{it}}{e^{2it} - 1} \right] dt = \frac{4}{\pi} \int_0^x \Re \left[\frac{e^{it} (e^{2nit} - 1)}{e^{it} (e^{it} - e^{-it})} \right] dt = \frac{4}{\pi} \int_0^x \Re \left[\frac{e^{2nit} - 1}{e^{it} - e^{-it}} \right] dt
 \end{aligned}$$

Rappelons que lors de notre étude des nombres complexes (cf. chapitre Nombres) nous avons démontré que:

$$\sin(x) = \frac{e^{ix} - e^{-ix}}{2i}$$

Ce qui nous amène donc à:

$$\begin{aligned}
 f(x) &= \frac{4}{\pi} \int_0^x \Re \left[\frac{e^{2nit} - 1}{e^{it} - e^{-it}} \right] dt = \frac{4}{\pi} \int_0^x \Re \left[\frac{e^{2nit} - 1}{2i \sin(t)} \right] dt \\
 &= \frac{4}{\pi} \int_0^x \Re \left[\frac{\cos(2nt) + i \sin(2nt) - 1}{2i \sin(t)} \right] dt = \frac{2}{\pi} \int_0^x \Re \left[\frac{\cos(2nt)}{i \sin(t)} + \frac{i \sin(2nt)}{i \sin(t)} - \frac{1}{i \sin(t)} \right] dt \\
 &= \frac{2}{\pi} \int_0^x \Re \left[\frac{\cos(2nt)}{i \sin(t)} + \frac{i \sin(2nt)}{i \sin(t)} - \frac{1}{i \sin(t)} \right] dt = \frac{2}{\pi} \int_0^x \Re \left[-\frac{i \cos(2nt)}{\sin(t)} + \frac{\sin(2nt)}{\sin(t)} + \frac{i}{2 \sin(t)} \right] dt \\
 &= \frac{2}{\pi} \int_0^x \Re \left[\frac{\sin(2nt)}{\sin(t)} - i \frac{\cos(2nt) + 1}{2 \sin(t)} \right] dt = \frac{2}{\pi} \int_0^x \frac{\sin(2nt)}{\sin(t)} dt
 \end{aligned}$$

Nous allons nous intéresser aux petites valeurs de x . Donc nous pouvons alors faire un développement au première ordre en Maclaurin du dénominateur (mais pas du numérateur à cause de la présence du n):

$$f(x) \cong \frac{2}{\pi} \int_0^x \frac{\sin(2nt)}{t} dt$$

Nous faisons un changement de variable:

$$f(x) \cong \frac{2}{\pi} \int_0^{2\pi x} \frac{\sin(s)}{\frac{s}{2n}} \frac{ds}{2n} = \frac{2}{\pi} \int_0^{2\pi x} \frac{\sin(s)}{s} ds = \frac{2}{\pi} \int_0^{2\pi x} \text{sinc}(x) ds$$

où nous avons utilisé la notation traditionnelle du "sinus cardinal" dans la dernière relation (car cette fraction est fréquente en physique raison pour laquelle elle a une notation spécifique).

Comme ce qui nous intéresse est de déterminer le maximum du phénomène (la perturbation) de Gibbs, nous voyons que celle-ci a lieu dans le cas particulier que nous avons présenté (voir figure ci-dessus) à chaque multiple de π et comme le dénominateur de l'expression de l'intégrale va diminuer en fonction que le multiple est plus grand, il vient que le plus grand maximum est au point où $2\pi x = \pi$ (le point 0 à l'opposé va annuler l'intégrale donc il faut l'éliminer de notre choix). Nous avons alors:

$$f_{\text{Gibbs}}(x) \cong \frac{2}{\pi} \int_0^{\pi} \frac{\sin(s)}{s} ds$$

et la valuation de cette intégrale ne peut être faire que numériquement à notre connaissance, il vient:

$$f_{\text{Gibbs}}(x) \cong 1.179$$

Soit environ 18% au-dessus de la valeur limite attendue.

2.5.1. PUISSANCE D'UN SIGNAL

Un signal périodique possède une énergie infinie et une puissance moyenne nulle (cf. chapitre d'Électrocinétique). Sa puissance moyenne sur une période est alors définie par:

$$P_f = \frac{1}{T} \int_T |f(t)|^2 dt \quad (11.279)$$

Si nous développons cette équation, nous avons:

$$\begin{aligned} P_f &= \frac{1}{T} \int_T f(t) \bar{f}(t) dt = \frac{1}{T} \int_T f(t) \left(\sum_{k=-\infty}^{\infty} \bar{c}_k e^{-ik a_b t} \right) dt \\ &= \frac{1}{T} \sum_{k=-\infty}^{\infty} \bar{c}_k \int_T f(t) e^{-ik a_b t} dt = \sum_{k=-\infty}^{\infty} \bar{c}_k c_k = \sum_{k=-\infty}^{\infty} |c_k|^2 \end{aligned} \quad (11.280)$$

Cela signifie que la puissance d'un signal à temps continu périodique est égale à la somme des coefficients de Fourier au carré. C'est ce que l'on nomme le "**théorème de Parseval**". Cela signifie que si nous avons un signal quelconque que nous pouvons décomposer en série de Fourier, nous pouvons connaître la puissance de ce signal uniquement à l'aide des coefficients spectraux.

Dans la réalité, comme nous ne pouvons déterminer mathématiquement l'expression de ce signal, nous utilisons la discrétisation ou l'échantillonnage et ensuite à l'aide d'une transformée de Fourier discrète, nous pouvons calculer la puissance de ce signal en utilisant uniquement les coefficients spectraux. Cela nous donne une caractéristique du signal.

Indiquons aussi le résultat suivant qui nous sera utile dans le chapitre de Thermodynamique pour l'étude du corps noir et qui est aussi intimement liée à des propriétés très importantes la fonction zêta de Riemann:

$$\begin{aligned}
 \sum_{k=-\infty}^{\infty} |c_k|^2 &= \sum_{k=-\infty}^{\infty} \left| \frac{a_k - ib_k}{2} \right|^2 = \frac{1}{4} \sum_{k=-\infty}^{\infty} |a_k - ib_k|^2 = \frac{1}{4} \sum_{k=-\infty}^{\infty} \left(\sqrt{a_k^2 + b_k^2} \right)^2 \\
 &= \frac{1}{4} \sum_{k=-\infty}^{\infty} (a_k^2 + b_k^2) = \frac{1}{4} \sum_{k=-\infty}^{-1} (a_k^2 + b_k^2) + \frac{1}{4} (a_0^2 + b_0^2) + \frac{1}{4} \sum_{k=1}^{\infty} (a_k^2 + b_k^2) \quad (11.281) \\
 &= \frac{1}{4} a_0^2 + 2 \cdot \frac{1}{4} \sum_{k=1}^{\infty} (a_k^2 + b_k^2) = \frac{a_0^2}{4} + \frac{1}{2} \sum_{k=1}^{\infty} (a_k^2 + b_k^2)
 \end{aligned}$$

La relation suivante:

$$\boxed{\sum_{k=-\infty}^{\infty} |c_k|^2 = \frac{a_0^2}{4} + \frac{1}{2} \sum_{k=1}^{\infty} (a_k^2 + b_k^2)} \quad (11.282)$$

est appelée "**égalité de Parseval**".

Suivant la définition de la série de Fourier et de la définition du coefficient a_0 qui en découle immédiatement, nous avons aussi fréquemment dans la littérature:

$$\sum_{k=-\infty}^{\infty} |c_k|^2 = a_0^2 + \frac{1}{2} \sum_{k=1}^{\infty} (a_k^2 + b_k^2) \quad (11.283)$$

2.5.2. TRANSFORMÉE DE FOURIER

Les séries de Fourier sont un outil très puissant pour l'analyse de signaux périodiques par exemple, mais l'ensemble des fonctions périodiques est petit comparé à l'ensemble des fonctions que nous rencontrons dans les problèmes physiques. Ainsi, allons-nous introduire un nouvel outil d'analyse extrêmement puissant qui s'étend à une classe de fonctions plus générale.

La transformée de Fourier (TF) est ainsi utilisée tant pour les signaux périodiques que pour les signaux apériodiques.

Pour cela, nous repartons de notre étude sur les séries de Fourier en notation complexe d'une fonction périodique de période T en considérant que celle-ci devient de plus en plus grande jusqu'à la faire tendre vers $T \rightarrow +\infty$. Dès lors les raies spectrales se rapprochent peu à peu pour se transformer en un spectre continu.

Ainsi, reprenons les expressions démontrées ci-avant:

$$f(x) = \sum_{n=-\infty}^{+\infty} c_n e^{inx} \quad c_n = \frac{1}{T} \int_{-T/2}^{T/2} f(x) e^{-inx} dx \quad (11.284)$$

que nous pouvons écrire de manière équivalente sous la forme traditionnelle suivante (dans laquelle il est d'usage de mettre le facteur $1/T$ plutôt dans $f(t)$):

$$f(t) = \frac{1}{T} \sum_{k=-\infty}^{+\infty} c_k e^{i \frac{2\pi k}{T} t} \quad c_k = \int_{-T/2}^{T/2} f(t) e^{-i \frac{2\pi k}{T} t} dt \quad (11.285)$$

et écrivons encore cela pour des besoins ultérieurs sous la forme suivante:

$$f(t) = \frac{1}{2\pi} \sum_{k=-\infty}^{+\infty} \frac{2\pi}{T} c_k e^{i \frac{2\pi k}{T} t} \quad c_k = \int_{-T/2}^{T/2} f(t) e^{-i \frac{2\pi k}{T} t} dt \quad (11.286)$$

et posons:

$$\omega_k = \frac{2\pi k}{T} \quad (11.287)$$

Ainsi, quand $T \rightarrow +\infty$, la pulsation tend vers zéro et nous avons $\omega_k \rightarrow \omega$ car nous passons de valeurs discrètes à valeurs continues qui parcourent l'ensemble des réels (pour tous les k). Donc de:

$$f(t) = \frac{1}{2\pi} \sum_{k=-\infty}^{+\infty} \frac{2\pi}{T} c_k e^{i \omega_k t} \quad c_k = \int_{-T/2}^{T/2} f(t) e^{-i \omega_k t} dt \quad (11.288)$$

nous passons à la limite soit:

$$T \rightarrow +\infty \quad \omega_k \rightarrow \omega \quad (11.289)$$

et cela implique aussi que:

$$\frac{2\pi}{T} = \frac{2\pi(k+1)}{T} - \frac{2\pi k}{T} = \frac{2\pi \Delta k}{T} \underset{T \rightarrow +\infty}{=} d\omega \quad (11.290)$$

Nous obtenons ainsi pour les coefficients (nous changeons de notation car l'ancienne est inadaptée):

$$F(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-i \omega t} dt \quad (11.291)$$

et pour la série infinie (dont la somme devient une intégrale):

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega) e^{i \omega t} d\omega \quad (11.292)$$

Attention!!! Pour faire la différence entre la fonction donnée et son équivalent dont nous cherchons l'expression en somme infinie, nous les noterons dorénavant différemment. Ainsi, il vient:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega) e^{i\omega t} d\omega \quad F(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt \quad (11.293)$$

Donc la série de Fourier discrète devient une fonction continue.

Définitions:

D1. Nous appelons "**transformée de Fourier (TF)**" de f la relation:

$$\mathcal{F}(f(t))(\omega) = F(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt \quad (11.294)$$

notée parfois aussi sous la forme suivante:

$$\hat{f}(\omega) = \int f(t) e^{-i\omega t} dt \quad (11.295)$$

appelée aussi parfois "**densité spectrale d'amplitude**".

D2. Nous appelons "**transformée de Fourier inverse (TFI)**" de F la relation:

$$\mathcal{F}^{-1}(F(\omega))(t) = f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega) e^{i\omega t} d\omega \quad (11.296)$$

Toute technique de transformation de ce type (car il y en a plusieurs!) s'appelle une "**transformation intégrale**".

Remarque: Il existe de nombreuses manière d'écrire la transformée de Fourier en fonction du choix de la valeur initiale de T .

Certains physiciens préfèrent symétriser ces deux expressions en mettant le même coefficient dans les deux sens, qui sera par exemple $1/\sqrt{2\pi}$. Cela donnera:

$$\mathcal{F}(f(t))(\omega) = F(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt \quad \mathcal{F}^{-1}(F(\omega))(t) = f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F(\omega) e^{i\omega t} d\omega \quad (11.297)$$

Donnons également la forme tridimensionnelle qui nous servira de nombreuses fois en mécanique ondulatoire, électrodynamique, optique ondulatoire ou encore dans les divers chapitres de physique quantique:

$$F(\vec{k}) = \frac{1}{(2\pi)^{3/2}} \iiint f(\vec{r}) e^{-i\vec{k}\cdot\vec{r}} d^3r \quad f(\vec{r}) = \frac{1}{(2\pi)^{3/2}} \iiint F(\vec{k}) e^{i\vec{k}\cdot\vec{r}} d^3k \quad (11.298)$$

Pour que les choses soient peut-être plus claires, montrons de manière générale que la transformée de Fourier $\mathcal{F}$ précédemment écrite est une isométrie (conserve la norme).

Remarquons tout d'abord que pour tout f, g nous avons le produit scalaire fonctionnel:

$$f \circ \mathcal{F}(g) = \mathcal{F}^{-1}(f) \circ g \quad (11.299)$$

Mais puisque les fonctions sont dans l'espace des complexes, comme nous l'avons vu dans le chapitre de Calcul Vectoriel, nous devons alors utiliser la notation du produit hermitien:

$$\langle f | \mathcal{F}(g) \rangle = \langle \mathcal{F}^{-1}(f) | g \rangle \quad (11.300)$$

Rappelons quand même que:

$$\langle f | g \rangle = \int_{\mathbb{R}^n} f(\vec{r}) \overline{g(\vec{r})} d^n r \quad (11.301)$$

Démonstration:

Nous voulons donc démontrer l'égalité:

$$\langle f | \mathcal{F}(g) \rangle = \langle \mathcal{F}^{-1}(f) | g \rangle \quad (11.302)$$

Soit explicitement:

$$\langle f | \mathcal{F}(g) \rangle = \int_{\mathbb{R}^n} f(\vec{r}) \overline{\mathcal{F}(g)} d^n r \quad (11.303)$$

Mais les variables à intégrer doivent être les mêmes et pour que $\mathcal{F}(g)$ soit implicitement dépendante de $\vec{r}$ il faut donc prendre la transformée de Fourier en $\vec{k}$. Tel que:

$$\mathcal{F}(g) = \frac{1}{(2\pi)^{3/2}} \iiint f(\vec{k}) e^{-i\vec{k} \cdot \vec{r}} d^3 k \quad (11.304)$$

Ainsi:

$$\begin{aligned} \langle f(\vec{r}) | \mathcal{F}(g)(\vec{r}) \rangle &= \int_{\mathbb{R}^n} f(\vec{r}) \overline{\mathcal{F}(g)} d^n r = \int_{\mathbb{R}^n} f(\vec{r}) \overline{\left(\frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} g(\vec{k}) e^{-i\vec{k} \cdot \vec{r}} d^n k \right)} d^n r \\ &= \int_{\mathbb{R}^n} f(\vec{r}) \left(\frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \overline{g(\vec{k})} e^{i\vec{k} \cdot \vec{r}} d^n k \right) d^n r \end{aligned} \quad (11.305)$$

Soit en utilisant le théorème de Fubini (cf. chapitre de Calcul Différentiel Et Intégral):

$$\langle f | \mathcal{F}(g) \rangle = \int_{\mathbb{R}^n} \overline{g(\vec{k})} \left(\frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} f(\vec{r}) e^{i\vec{k} \cdot \vec{r}} d^n r \right) d^n k = \int_{\mathbb{R}^n} \overline{g(\vec{k})} \mathcal{F}^{-1}(f)(\vec{k}) d^n k = \langle \mathcal{F}^{-1}(f) | g \rangle \quad (11.306)$$

A l'aide de ce résultat, nous avons donc aussi démontré (c'est immédiat):

$$\langle \mathcal{F}(f) | \mathcal{F}(g) \rangle = \langle \mathcal{F}^{-1}(\mathcal{F}(f)) | g \rangle = \langle f | g \rangle \quad (11.307)$$

Nous n'avons pas précisé les bornes: elles sont infinies dans chaque définition (nous intégrons sur tous les $\vec{k}$ ou $\vec{r}$ possibles).

□C.Q.F.D.

Voyons maintenant deux propriétés intéressantes de la transformée de Fourier:

P1. Si f est paire, il vient une simplification de la transformée telle que:

$$\begin{aligned}\mathcal{F}(f(x))(\omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt = \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^0 f(t) e^{-i\omega t} dt + \int_0^{+\infty} f(t) e^{-i\omega t} dt \right) \\ &= \frac{1}{\sqrt{2\pi}} \left(- \int_{+\infty}^0 f(t) e^{i\omega t} dt + \int_0^{+\infty} f(t) e^{-i\omega t} dt \right) = \frac{1}{\sqrt{2\pi}} \left(\int_0^{+\infty} f(t) e^{i\omega t} dt + \int_0^{+\infty} f(t) e^{-i\omega t} dt \right) \quad (11.308) \\ &= \frac{1}{\sqrt{2\pi}} \int_0^{+\infty} f(t) (e^{i\omega t} + e^{-i\omega t}) dt = \frac{2}{\sqrt{2\pi}} \int_0^{+\infty} f(t) \cos(\omega t) dt\end{aligned}$$

P2. Si f est impaire, nous procédons de la même manière que ci-dessus et nous obtenons:

$$\mathcal{F}(f(x))(\omega) = -\frac{2i}{\sqrt{2\pi}} \int_0^{+\infty} f(t) \sin(\omega t) dt \quad (11.309)$$

P3. Propriété très importante des transformées de Fourier qui nous sera utile en finance (cf. chapitre d'Économie) et également dans le cadre de l'étude de l'équation de la chaleur (cf. chapitre de Thermodynamique).

Rappelons d'abord que la transformée de Fourier est donnée par:

$$\mathcal{F}(f(t))(\omega) = F(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt \quad (11.310)$$

Nous souhaitons voir ce qu'il se passe si:

$$\mathcal{F}(\partial_t f(t)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \partial_x f(t) e^{-i\omega t} dt \quad (11.311)$$

En faisant une intégration par parties (cf. chapitre de Calcul Différentiel Et Intégral):

$$\int_a^b f(t) \cdot g'(t) dt = f(t) \cdot g(t) \Big|_a^b - \int_a^b f'(t) \cdot g(t) dt \quad (11.312)$$

il vient:

$$\begin{aligned}
 \mathcal{F}(\partial_t f(t)) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-i\omega t} \cdot \partial_t f(t) dt = \frac{1}{\sqrt{2\pi}} e^{-i\omega t} f(t) \Big|_{-\infty}^{+\infty} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} -i\omega e^{-i\omega t} \cdot f(t) dt \\
 &= i\omega \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-i\omega t} \cdot f(t) dt = i\omega \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt = i\omega F(\omega)
 \end{aligned}
 \tag{11.313}$$

où nous nous sommes mis dans la situation avec:

$$\lim_{t \rightarrow \pm\infty} f(t) = 0 \tag{11.314}$$

Donc:

$$\mathcal{F}(\partial_t f(t)) = i\omega F(\omega) \tag{11.315}$$

De manière générale:

$$\mathcal{F}(\partial_t^n f(t)) = (i\omega)^n F(\omega) \tag{11.316}$$

Remarque: La branche de "**l'analyse harmonique**", ou "**analyse de Fourier 2D**", est la branche des mathématiques qui étudie la représentation des fonctions ou des signaux comme superposition d'ondes de base. Elle approfondit et généralise les notions de série de Fourier et de transformée de Fourier. Les ondes de base s'appellent les harmoniques, d'où le nom de la discipline. Durant ces deux derniers siècles, elle a eu de nombreuses applications en physique et en économie sous le nom "**d'analyse spectrale**", et connaît des applications récentes notamment en traitement des signaux, mécanique quantique, neurosciences, stratigraphie, statistiques...

Exemples:

E1. Voyons donc un exemple (parmi les deux fondamentaux) d'une transformée de Fourier que nous retrouvons en physique quantique aussi bien qu'en optique ondulatoire.

Nous allons calculer la transformée de Fourier de la fonction suivante:

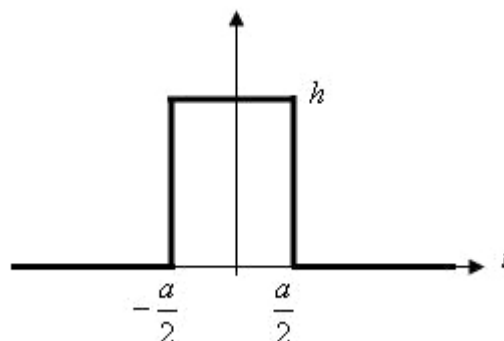


Figure: 11.18 - Fonction dont nous souhaitons calculer la TF

Nous avons donc:

$$\begin{aligned}
\mathcal{F}(f(t)) &= \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt = \int_{-\infty}^{+\infty} f(t) e^{-i2\pi f t} dt = \int_{-\infty}^{+\infty} f(t) e^{-i2\pi f t} dt \\
&= \int_{-a/2}^{+a/2} h \cdot e^{-i2\pi f t} dt = h \cdot \int_{-a/2}^{+a/2} e^{-i2\pi f t} dt = -\frac{h}{i2\pi f} \left[e^{-i2\pi f t} \right]_{-a/2}^{+a/2} \\
&= -\frac{h}{i2\pi f} \left(e^{-i\pi f a} - e^{i\pi f a} \right) = -\frac{h}{i2\pi f} \left(e^{i(-\pi f a)} - e^{i(\pi f a)} \right) = -\frac{h}{i2\pi f} (-2i \sin(\pi f a)) \\
&= h \cdot \frac{\sin(\pi f a)}{\pi f} = h \cdot a \cdot \text{sinc}(\pi f a)
\end{aligned} \tag{11.317}$$

où sinc est le sinus cardinal. Nous retombons donc sur le sinus cardinal (si nous prenons le module au carré) de la décomposition d'une onde monochromatique diffractée par une fente rectangulaire. Ainsi, il semble possible d'étudier les phénomènes de diffraction en utilisant la transformée de Fourier et ce domaine se nomme "**l'optique de Fourier**".

E2. La transformée de Fourier d'une fonction intégrable f est donnée par:

$$F(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt \tag{11.318}$$

Considérons la fonction intégrable de type Gaussienne:

$$f(t) = e^{-\alpha^2 t^2} \tag{11.319}$$

avec $\alpha > 0$ défini sur $\mathbb{R}$.

Nous voulons calculer sa transformée de Fourier car il s'agit d'un cas très important et en particulier utile pour la résolution de l'équation de la chaleur que nous traiterons dans le chapitre de Thermodynamique et pour résoudre l'équation différentielle de Black-Scholes dans le chapitre d'Économie.

L'astuce géniale, si nous voulons éviter de faire de l'analyse complexe sur 3 pages A4, consiste à remarquer que $F(\omega)$ est solution de l'équation différentielle linéaire suivante:

$$y'(\omega) + \frac{\omega}{2\alpha} y = 0 \tag{11.320}$$

où y est une fonction de ω .

En effet en dérivant $F(\omega)$ nous obtenons:

$$\begin{aligned}
F'(\omega) &= \frac{d}{d\omega} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt = \frac{d}{d\omega} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\alpha^2 t^2} \frac{d}{d\omega} e^{-i\omega t} dt \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\alpha^2 t^2} (-ite^{-i\omega t}) dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (-ite^{-\alpha^2 t^2}) e^{-i\omega t} dt
\end{aligned} \tag{11.321}$$

Une intégration par parties nous donne:

$$\begin{aligned}
 F'(\omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-i\omega t} (-ite^{-at^2}) dt = \frac{1}{\sqrt{2\pi}} \left[e^{-i\omega t} \frac{ie^{-at^2}}{2a} \right]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} (-i\omega e^{-i\omega t}) \left(\frac{ie^{-at^2}}{2a} \right) dt \\
 &= -\frac{\omega}{2a} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-i\omega t} e^{-at^2} dt = -\frac{\omega}{2a} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-at^2} e^{-i\omega t} dt
 \end{aligned} \quad (11.322)$$

On reconnaît l'expression de la transformée de Fourier de f . Par conséquent:

$$F'(\omega) = -\frac{\omega}{2a} F(\omega) \quad (11.323)$$

Ceci montre que $F(\omega)$ est bien solution de l'équation différentielle ci-dessus.

Nous avons démontré dans le chapitre de Calcul Différentiel et Intégral que la solution générale de cette équation différentielle est donnée par:

$$y = Ae^{-G(\omega)} \quad (11.324)$$

où $A \in \mathbb{R}$. Et comme dans le cas présent:

$$g(x) = \frac{\omega}{2a} \quad (11.325)$$

La primitive $G(x)$ est donc facile à calculer et il vient:

$$y = Ae^{-\frac{\omega^2}{4a}} \quad (11.326)$$

Par conséquent:

$$F(\omega) = Ae^{-\frac{\omega^2}{4a}} \quad (11.327)$$

Pour déterminer la constante A il suffit de remarquer que:

$$F(0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-at^2} e^{-i0t} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-at^2} dt = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{\pi}{a}} = \frac{1}{\sqrt{2a}} \quad (11.328)$$

et donc:

$$F(\omega) = \frac{1}{\sqrt{2a}} e^{-\frac{\omega^2}{4a}} \quad (11.329)$$

Il est alors d'usage de dire que la transformation de Fourier d'une Gaussienne est une autre Gaussienne.

2.6. SÉRIES DE BESSEL

Les fonctions de Bessel sont très utiles dans de nombreux domaines de pointe de la physique faisant intervenir des équations différentielles délicates à résoudre. Les domaines dans lesquels nous les trouvons le plus souvent sont la calorimétrie (conduction de la chaleur), la physique nucléaire (physique de réacteurs), et la mécanique des fluides.

Ces séries sont cependant très peu détaillées dans les écoles universitaires et il est souvent du rôle de l'élève de chercher les compléments d'informations dont il a besoin sur le sujet dans la bibliothèque de son école. Nous avons voulu présenter ici les développements permettant d'éviter cette démarche tout en restant chez soi devant son ordinateur (de plus les livres sur le sujet sont assez rares...).

Remarque: Nous parlons habituellement par abus de langage des "**fonctions de Bessel**" au lieu des "**séries de Bessel**".

Il existe une quantité non négligeable de fonctions de Bessel mais nous allons nous restreindre à l'étude de celles qui sont les plus utilisées en physique.

2.6.1. FONCTION DE BESSEL D'ORDRE ZÉRO

La fonction connue sous le nom de "**fonction de Bessel d'ordre zéro**", est définie par la série de puissances:

$$J_0(x) = 1 - \frac{x^2}{2^2} + \frac{x^4}{2^2 \cdot 4^2} - \frac{x^6}{2^2 \cdot 4^2 \cdot 6^2} + \dots \quad (11.330)$$

C'est lors de l'étude des propriétés de dérivation et d'intégration que Bessel a trouvé que cette série de puissance est une solution à une équation différentielle que l'on retrouve assez fréquemment en physique. C'est pourquoi elle porte son nom.

Si u_r représente le r -ème terme de la série, nous voyons aisément que:

$$\frac{u_{r+1}}{u_r} = \frac{-x^2}{(2r)^2} \quad (11.331)$$

qui tend vers zéro quand $r \rightarrow \infty$, quelle que soit la valeur de x . Cela a pour conséquence que la série converge pour toutes les valeurs de x . Comme il s'agit d'une série de puissance positive, la fonction $J_0(x)$ et toutes ses dérivées sont continues pour toutes valeurs de x , réelles ou complexes.

2.6.2. FONCTION DE BESSEL D'ORDRE N

La fonction $J_n(x)$, connue sous le nom de "**fonction de Bessel d'ordre n** ", est définie, lorsque n est un entier positif, par la série de puissance:

$$J_n(x) = \frac{x^n}{2^n n!} \left(1 - \frac{x^2}{2 \cdot (2n+2)} + \frac{x^4}{2 \cdot 4(2n+2)(2n+4)} - \dots \right) \quad (11.332)$$

qui converge pour toutes valeurs de x , réelles ou complexes.

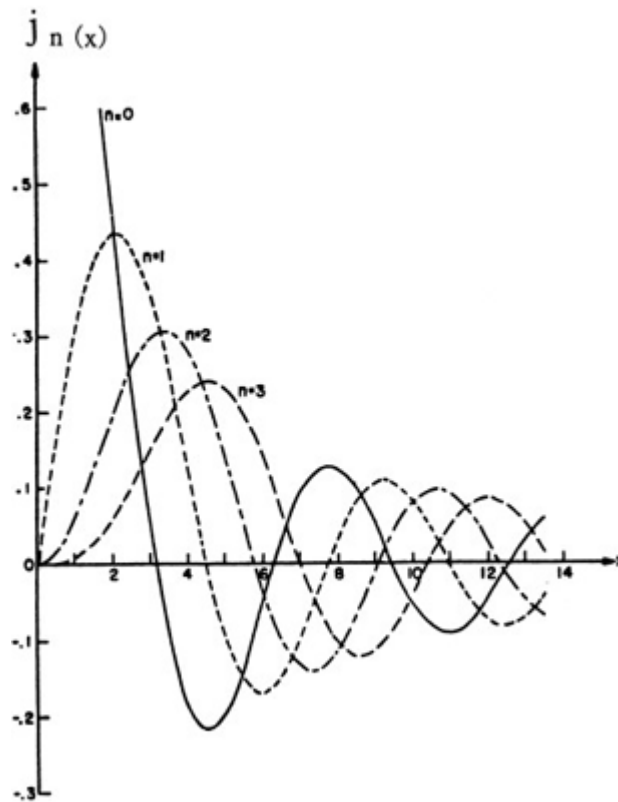


Figure: 11.19 - Tracé de quelques fonctions de Bessel

et dans Microsoft Excel 11.8346 ou Maple 4.00b la fonction précédente se trouve sous le nom BESSELJ. Par exemple pour obtenir le graphique précédent dans Maple, il suffit d'écrire:

```
> plot([BesselJ(0,x),BesselJ(1,x),BesselJ(2,x),BesselJ(3,x)],x=0..20);
```

Voyons qu'en particulier, pour $n = 1$ nous avons:

$$J_1(x) = \frac{x}{2} - \frac{x^3}{2^2 \cdot 4} + \frac{x^5}{2^2 \cdot 4^2 \cdot 6} - \frac{x^7}{2^2 \cdot 4^2 \cdot 6^2 \cdot 8} + \dots \quad (11.333)$$

et quand $n = 2$:

$$J_2(x) = \frac{x^2}{2 \cdot 4} - \frac{x^4}{2^2 \cdot 4 \cdot 6} + \frac{x^6}{2^2 \cdot 4^2 \cdot 6 \cdot 8} - \frac{x^8}{2^2 \cdot 4^2 \cdot 6^2 \cdot 8 \cdot 10} + \dots \quad (11.334)$$

Nous pouvons noter que $J_n(x)$ est une fonction paire de x quand n est pair, et impaire quand n est impair (cf. chapitre d'Analyse Fonctionnelle).

En dérivant la fonction $J_0(x)$ et en comparant le résultat avec la série $J_1(x)$, nous voyons sans trop de peine que:

$$\frac{dJ_0(x)}{dx} = -J_1(x) \quad (11.335)$$

Nous trouvons également sans trop de difficulté, la relation suivante:

$$\frac{d}{dx} xJ_1(x) = xJ_0(x) \quad (11.336)$$

En utilisant le fait que:

$$\frac{dJ_0(x)}{dx} = -J_1(x) \quad (11.337)$$

et en l'incluant dans la précédente relation, nous trouvons:

$$\frac{d}{dx} \left(x \frac{dJ_0(x)}{dx} \right) + xJ_0(x) = 0 \quad (11.338)$$

ou écrit autrement:

$$x \frac{d^2 J_0(x)}{dx^2} + \frac{dJ_0(x)}{dx} + xJ_0(x) = 0 \quad (11.339)$$

$y = J_0(x)$ est donc une solution de l'équation différentielle du second ordre:

$$\frac{d}{dx} \left(x \frac{dy}{dx} \right) + xy = 0 \quad (11.340)$$

ou écrit autrement:

$$x \frac{d^2 y}{dx^2} + \frac{dy}{dx} + xy = 0 \quad (11.341)$$

ou encore:

$$\frac{d^2 y}{dx^2} + \frac{1}{x} \frac{dy}{dx} + y = 0 \quad (11.342)$$

Une solution à une équation de Bessel de paramètre n qui n'est pas un multiple de $J_n(x)$ est appelée "**fonction de Bessel du second type**". Supposons que u soit une telle fonction et posons $v = J_0(x)$; alors d'après la relation:

$$x \frac{d^2 y}{dx^2} + \frac{dy}{dx} + xy = 0 \quad (11.343)$$

nous avons:

$$xu'' + u' + xu = 0 \text{ et } xv'' + v' + xv = 0 \quad (11.344)$$

En multipliant la première relation par v et la seconde par u et après soustraction, nous obtenons:

$$x(u''v - uv'') + u'v - uv' = 0 \quad (11.345)$$

nous avons donc également:

$$\frac{d}{dx}(u'v - uv') = u''v + u'v' - u'v' - uv'' = u''v - uv'' \quad (11.346)$$

nous pouvons donc écrire:

$$\frac{d}{dx}[x(u'v - uv')] = 0 \quad (11.347)$$

effectivement car si nous développons, nous trouvons:

$$\begin{aligned} \frac{d}{dx}[x(u'v - uv')] &= 1 \cdot u'v + xu''v + xu'v' - 1 \cdot uv' - xu'v' - xuv'' \\ &= x(u''v + u'v' - u'v' - uv'') + u'v - uv' = x(u''v - uv'') + u'v - uv' = 0 \end{aligned} \quad (11.348)$$

Pour que l'égalité:

$$\frac{d}{dx}[x(u'v - uv')] = 0 \quad (11.349)$$

soit satisfaite, nous avons:

$$x(u'v - uv') = c^{\text{te}} = B \quad (11.350)$$

En divisant par xv^2 , nous avons:

$$\frac{u'v - uv'}{v^2} = \frac{B}{xv^2} \quad (11.351)$$

ce qui est équivalent à:

$$\frac{d}{dx}\left(\frac{u}{v}\right) = \frac{B}{xv^2} \quad (11.352)$$

de suite, par intégration il vient:

$$\frac{u}{v} = A + B \int \frac{1}{xv^2} dx \quad (11.353)$$

où A est une constante. Consécutivement nous avons, puisque $v = J_0(x)$:

$$u = AJ_0(x) + BJ_0(x) \int \frac{1}{xJ_0^2(x)} dx \quad (11.354)$$

où rappelons-le, A et B sont des constantes, et $B \neq 0$ si u n'est pas un multiple de $J_0(x)$ par définition.

Si dans la dernière relation, $J_0(x)$ est remplacé par son expression en termes de série nous avons:

$$\frac{1}{xJ_0^2(x)} = \frac{1}{x} + \frac{x}{2} + \frac{5x^3}{32} + \dots \quad (11.355)$$

Pour ceux qui veulent vérifier cette dernière relation (je n'aime pas ce genre de calculs algébriques) avec Maple 4.00b il suffit d'écrire:

```
>1/x*taylor(1/(series(BesselJ(0,x),x)^2,x=0,5);
```

Dès lors:

$$\begin{aligned} J_0(x) \int \frac{1}{x J_0^2(x)} dx &= J_0(x) \left[\log(x) + \frac{x^2}{4} + \frac{5x^4}{128} + \dots \right] \\ &= J_0(x) \log(x) + \left(1 - \frac{x^2}{2^2} + \dots \right) \left(\frac{x^2}{4} + \frac{5x^4}{128} + \dots \right) \\ &= J_0(x) \log(x) + \frac{x^2}{4} - \frac{3x^4}{128} + \dots \end{aligned} \quad (11.356)$$

consécutivement si nous posons:

$$Y_0(x) = J_0(x) \log(x) + \frac{x^2}{4} - \frac{3x^4}{128} + \dots \quad (11.357)$$

où $Y_0(x)$ est une fonction de Bessel particulière du second type appelée "**fonction de Bessel-Neumann du second type d'ordre nul**".

Identiquement au fait que $J_0(x) \rightarrow 1$ quand $x \rightarrow 0$, l'expression $Y_0(x)$ à cause du terme $\log(x)$ quand x est petit tend vers $Y_0(x) \rightarrow -\infty$ quand $x \rightarrow +0$.

Finalement, il vient de ce que nous avons vu précédemment que $J_0(x)$ et $Y_0(x)$ sont des solutions indépendantes de l'équation différentielle:

$$\frac{d^2 y}{dx^2} + \frac{1}{x} \frac{dy}{dx} + y = 0 \quad (11.358)$$

La solution générale étant donc:

$$y = A J_0(x) + B Y_0(x) \quad (11.359)$$

où A, B sont des constantes arbitraires et $x > 0$ afin que $Y_0(x)$ soit réel.

Si nous remplaçons x par kx , où k est une constante, l'équation différentielle devient:

$$\frac{1}{k^2} \frac{d^2 y}{dx^2} + \frac{1}{kx} \frac{dy}{dx} + y = 0 \quad (11.360)$$

en multipliant le tout par k^2 , nous trouvons la forme générale de l'équation différentielle:

$$\frac{d^2 y}{dx^2} + \frac{1}{x} \frac{dy}{dx} + k^2 y = 0 \quad (11.361)$$

dont la solution générale est:

$$y = AJ_0(kx) + BY_0(kx) \quad (11.362)$$

où $k > 0$ afin que $Y_0(kx)$ soit réel quand $x > 0$.

Au fait, les fonctions de Bessel viennent des solutions de l'équation différentielle étudiée précédemment et solutionnées par la méthode de Frobenius. Posons:

$$Ly \equiv \left(x \frac{d^2}{dx^2} + \frac{d}{dx} + x \right) y \quad (11.363)$$

et faisons la substitution:

$$y = x^\rho + c_1 x^{\rho+1} + c_2 x^{\rho+2} + c_3 x^{\rho+3} + \dots \quad (11.364)$$

en substituant dans Ly , nous obtenons:

$$Ly = \rho^2 x^{\rho-1} + (\rho+1)^2 c_1 x^\rho + [(\rho+2)^2 c_2 + 1] x^{\rho+1} + [(\rho+3)^2 c_3 + c_1] x^{\rho+2} + \dots \quad (11.365)$$

Choisissons maintenant les c_i afin de satisfaire l'équation différentielle tels que:

$$\begin{aligned} (\rho+1)^2 c_1 &= 0 \\ (\rho+2)^2 c_2 + 1 &= 0 \\ (\rho+3)^2 c_3 + c_1 &= 0 \\ &\dots \end{aligned} \quad (11.366)$$

Dès lors, à moins que ρ soit un entier négatif, nous avons:

$$\begin{aligned} c_1 &= c_3 = c_5 = c_7 = \dots = 0 \\ c_2 &= -\frac{1}{(\rho+2)^2} \\ c_4 &= -\frac{c_2}{(\rho+4)^2} = \frac{1}{(\rho+2)^2(\rho+4)^2} \\ &\dots \end{aligned} \quad (11.367)$$

En substituant ces valeurs dans la relation:

$$y = x^\rho + c_1 x^{\rho+1} + c_2 x^{\rho+2} + c_3 x^{\rho+3} + \dots \quad (11.368)$$

nous obtenons:

$$y = x^\rho \left(1 - \frac{x^2}{(\rho+2)^2} + \frac{x^4}{(\rho+2)^2(\rho+4)^2} - \dots \right) \quad (11.369)$$

dès lors:

$$x \frac{d^2 y}{dx^2} + \frac{dy}{dx} + xy = Ly = \rho^2 x^{\rho-1} \quad (11.370)$$

si nous posons $\rho = 0$ dans l'avant-dernière relation, nous obtenons:

$$y = x^0 \left(1 - \frac{x^2}{(\rho+2)^2} + \frac{x^4}{(\rho+2)^2(\rho+4)^2} - \dots \right) = 1 - \frac{x^2}{2^2} + \frac{x^4}{2^2 4^2} - \dots = J_0(x) \quad (11.371)$$

2.6.3. ÉQUATION DIFFÉRENTIELLE DE BESSEL D'ORDRE N

Nous avons défini les séries de Bessel comme étant:

$$J_n(x) = \frac{x^n}{2^n n!} \left(1 - \frac{x^2}{2 \cdot (2n+2)} + \frac{x^4}{2 \cdot 4(2n+2)(2n+4)} - \dots \right) \quad (11.372)$$

Posons:

$$y = J_n(x) 2^n n! = x^n - \frac{x^{n+2}}{2 \cdot (2n+2)} + \frac{x^{n+4}}{2 \cdot 4(2n+2)(2n+4)} - \dots \quad (11.373)$$

et dérivons ainsi:

$$x \frac{d}{dx} \left(x \frac{dy}{dx} \right) = n^2 x^n - \frac{(n+2)^2 x^{n+2}}{2 \cdot (2n+2)} + \frac{(n+4)^2 x^{n+4}}{2 \cdot 4(2n+2)(2n+4)} - \dots \quad (11.374)$$

Mais nous avons aussi:

$$n^2 y = n^2 x^n - \frac{n^2 x^{n+2}}{2 \cdot (2n+2)} + \frac{n^2 x^{n+4}}{2 \cdot 4(2n+2)(2n+4)} - \dots \quad (11.375)$$

Par soustraction:

$$x \frac{d}{dx} \left(x \frac{dy}{dx} \right) - n^2 y = -x^{n+2} + \frac{x^{n+4}}{2 \cdot (2n+2)} - \dots = -x^2 y \quad (11.376)$$

Ce qui donne finalement:

$$x \frac{d}{dx} \left(x \frac{dy}{dx} \right) - (x^2 - n^2) y = 0 \quad (11.377)$$

Ce qui s'écrit également:

$$x^2 \frac{d^2 y}{dx^2} + x \frac{dy}{dx} + (x^2 - n^2) y = 0 \quad (11.378)$$

qui est appelée "**l'équation différentielle de Bessel d'ordre n**" ou plus simplement "**équation de Bessel**".
Au fait, la plupart des écoles ou sites Internet donnent cette équation différentielle comme une définition et pourtant il est clair qu'il y a un raisonnement rigoureux derrière cette équation.

La solution est donc du type:

$$y = 2^n n! J_n(x) \quad (11.379)$$

ce qui s'écrit encore parfois en utilisant la fonction gamma d'Euler:

$$y = 2^n \Gamma_0(n) J_n(x) = c^n J_n(x) \quad (11.380)$$

Il s'ensuit que:

$$x^2 \frac{d^2 J_n(x)}{dx^2} + x \frac{dJ_n(x)}{dx} + (x^2 - n^2) J_n(x) = 0 \quad (11.381)$$

et donc que $y = J_n(x)$ est solution de cette équation différentielle.

3. CRITÈRES DE CONVERGENCE

Lorsque nous étudions une série, l'une des questions fondamentales est celle de la convergence ou de la divergence de cette série.

Si une série converge, son terme général tend vers zéro lorsque n tend vers l'infini:

$$\lim_{n \rightarrow \infty} a_n = 0 \quad (11.382)$$

Ce critère est nécessaire mais non suffisant pour établir la convergence d'une série. Par contre, si ce critère n'est pas rempli, on est absolument sûr que la série ne converge pas (donc elle diverge!).

Trois méthodes sont proposées pour approfondir le critère de convergence:

1. Le test de l'intégrale
2. La règle d'Alembert
3. La règle de Cauchy

Dans les paragraphes suivants, nous admettrons des séries à termes positifs. Le cas de la série alternée sera vu ultérieurement.

3.1. TEST DE L'INTÉGRALE

Soit la série à termes positifs décroissants:

$$a_1 + a_2 + a_3 + \dots + a_n + \dots \quad (11.383)$$

c'est-à-dire:

$$a_1 \geq a_2 \geq a_3 \geq \dots \geq a_n \geq \dots \quad (11.384)$$

et soit une fonction continue décroissante telle que:

$$f(1) = a_1, f(2) = a_2, \dots, f(n) = a_n \quad (11.385)$$

nous pouvons alors affirmer que:

1. Si l'intégrale:

$$\int_1^{\infty} f(x) dx \quad (11.386)$$

converge, la série converge également.

2. Si l'intégrale:

$$\int_1^{\infty} f(x) dx \quad (11.387)$$

diverge, la série diverge également.

Remarque: En aucun cas l'intégrale ne donne la valeur de la somme de la série ! Le test de l'intégrale donne simplement une indication sur la convergence de la série. Avant de faire le test de l'intégrale, il est important de vérifier que les termes de la série soient strictement décroissants afin de remplir la condition $a_1 \geq a_2 \geq a_3 \geq \dots \geq a_n \geq \dots$.

3.2. RÈGLE D'ALEMBERT

Si dans une série à termes positifs:

$$a_1 + a_2 + a_3 + \dots + a_n + \dots \quad (11.388)$$

le rapport a_{n+1} / a_n (assimilable à une fonction prise en son entier) a une limite finie L lorsque $n \rightarrow \infty$:

$$L = \lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} \quad (11.389)$$

1. Si $L < 1$, la série converge

2. Si $L > 1$, la série diverge

3. Si $L = 1$ on ne peut rien dire

et nous définissons le "**rayon de convergence**" comme:

$$R = 1 / L \quad (11.390)$$

3.3. RÈGLE DE CAUCHY

Si dans une série à termes positifs:

$$a_1 + a_2 + a_3 + \dots + a_n + \dots \quad (11.391)$$

la quantité $\sqrt[n]{a_n}$ a une limite finie L lorsque $n \rightarrow \infty$ telle que:

$$L = \lim_{n \rightarrow \infty} \sqrt[n]{a_n} \quad (11.392)$$

avec à nouveau les mêmes considérations que pour la règle d'Alembert:

1. Si $L < 1$, la série converge
2. Si $L > 1$, la série diverge
3. Si $L = 1$ on ne peut rien dire

3.4. THÉOREME DE LEIBNIZ

Nous avons considéré jusqu'à présent des séries à termes positifs. Nous allons considérer dans cette partie des séries dont les termes sont alternés, c'est-à-dire des séries de la forme:

$$a_1 - a_2 + a_3 - a_4 + \dots \quad (a_n > 0) \quad (11.393)$$

Définition: Une série est dite "**série alternée**" si deux termes consécutifs de cette série sont de signe contraire.

Si dans une série alternée les termes en valeur absolue vont en décroissant:

$$a_1 > a_2 > a_3 > \dots > a_n > \dots \quad (11.394)$$

et si:

$$\lim_{n \rightarrow \infty} a_n = 0 \quad (11.395)$$

alors la série converge, sa somme est positive et n'est pas supérieure au premier terme.

Si S est la somme de la série et S_n une somme partielle, alors:

$$|S - S_n| \leq a_{n+1} \quad (11.396)$$

Remarque: Il est important de vérifier que les valeurs absolues des termes de la série soient strictement décroissantes afin de remplir la condition précédente.

3.5. CONVERGENCE ABSOLUE

Définition: Une série à termes variables est dite absolument convergente si la série formée avec la valeur absolue de ses termes converge:

$$a_1 + |-a_2| + a_3 + |-a_4| + \dots + |-a_n| + \dots \quad (a_n > 0) \quad (11.397)$$

Si une série alternée de termes est absolument convergente, la série absolue qui en découle converge aussi.

Nous pouvons généraliser la règle d'Alembert au cas des séries à termes quelconques:

$$a_1 \pm a_2 \pm a_3 \pm a_4 \pm \dots \pm a_n \pm \dots \quad (a_n > 0) \quad (11.398)$$

Ainsi, le rapport $|a_{n+1}/a_n|$ a une limite finie L lorsque pour $n \rightarrow \infty$ nous avons:

$$L = \lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| \quad (11.399)$$

toujours avec les mêmes conclusions que pour la règle d'Alembert normale.

3.6. THÉOREME DU POINT FIXE

Le théorème du point fixe n'est pas vraiment utile en physique (implicitement il est indispensable mais les physiciens utilisent souvent des outils mathématiques dont les propriétés ont déjà été validées au préalable par des mathématiciens), cependant nous le retrouvons en théorie du chaos (les vortex, tourbillons, etc...) ainsi qu'en informatique théorique (voir chapitre traitant des fractales en particulier le triangle de Sierpinski). Nous ne saurions donc que recommander au lecteur de prendre le temps de lire et de comprendre les explications et développements qui vont suivre.

Soit (X, d) , un espace métrique complet (cf. chapitre de Topologie ou des Fractales) et soit $T: X \rightarrow X$ une application strictement contractante de constante L (voir les fonctions lipschitziennes chapitre de Topologie), alors il existe un unique point $\varpi \in X$ tel que :

$$T(\varpi) = \varpi \quad (11.400)$$

ϖ est alors dit le "**point fixe**" de T (penser par exemple à $\cos(x)=x$). De plus si nous notons par:

$$T^n(x) = \underbrace{T(T(\dots T(x)\dots))}_{n\text{-fois}} \quad (11.401)$$

l'image de x par le n -ème itéré de T , nous avons alors:

$$\varpi = \lim_{n \rightarrow \infty} (T^n(x)) \quad \forall x \in X \quad (11.402)$$

et la vitesse de convergence peut d'ailleurs être estimée par:

$$d(\varpi, T^n(x)) \leq d(x, T(x)) \frac{L^n}{1-L} \quad (11.403)$$

Remarque: Vous pouvez vous amuser avec votre calculatrice de poche ou celle de votre système d'exploitation en choisissant un nombre au hasard et en en prenant le cosinus de manière itérative. Vous verrez que vous tendrez 0.74 et donc que in extenso il s'agit de la solution de $\cos(x)=x$.

Démonstration:

Soit $x \in X$. Nous considérons la suite $(T^n(x))_{n \in \mathbb{N}}$ définie comme ci-dessus. Nous allons d'abord montrer que cette suite est une suite de Cauchy (voir plus haut dans le présent chapitre ce qu'est une suite de Cauchy).

En appliquant l'inégalité triangulaire (cf. chapitre d'Analyse Vectorielle) plusieurs fois nous avons:

$$d(T^{n+k}(x), T^n(x)) \leq d(T^{n+k}(x), T^{n+k-1}(x)) + d(T^{n+k-1}(x), T^{n+k-2}(x)) + \dots + d(T^{n+1}(x), T^n(x)) \quad (11.404)$$

Or:

$$d(T^{n+1}(x), T^n(x)) \leq L \cdot d(T^n(x), T^{n-1}(x)) \leq \dots \leq L^n \cdot d(T(x), x) \quad (11.405)$$

donc:

$$d(T^{n+k}(x), T^n(x)) \leq (L^{n+k-1} + L^{n+k-2} + \dots + L^n) \cdot d(T(x), x) = L^n \cdot d(T(x), x) \cdot (1 + L + L^2 + \dots + L^{k-1})$$

et $(1 + L + L^2 + \dots + L^{k-1}) = \frac{1 - L^k}{1 - L} \leq \frac{1}{1 - L}$ ainsi $d(T^{n+k}(x), T^n(x)) \leq \frac{L^n}{1 - L} \cdot d(T(x), x) (*)$

(11.406)

pour finir:

$$0 \leq L < 1 \Rightarrow \lim_{n \rightarrow \infty} \frac{L^n}{1 - L} = 0 \quad (11.407)$$

c'est-à-dire que dans un premier temps $(T^n(x))_{n \in \mathbb{N}}$ est bien une suite de Cauchy.

(X, d) étant un espace complet nous avons que $(T^n(x))_{n \in \mathbb{N}}$ converge, et nous posons:

$$\varnothing = \lim_{n \rightarrow \infty} T^n(x) \quad (11.408)$$

A présent, nous vérifions que $\varnothing$ est bien un point fixe de T . En effet T est uniformément continue (car lipschitzienne - voir le chapitre de Topologie) donc à fortiori continue ainsi:

$$\varnothing = \lim_{n \rightarrow \infty} T^n(x) = \lim_{n \rightarrow \infty} T(T^{n-1}(x)) \underset{\text{car } T \text{ est continue}}{=} T(\lim_{n \rightarrow \infty} T^{n-1}(x)) = T(\varnothing) \quad (11.409)$$

Il reste à vérifier que $\varnothing$ est l'unique point fixe (du coup nous aurons démontré que $\varnothing$ ne dépend pas du choix de x). Supposons que nous ayons aussi $T(y) = y$ alors:

$$d(y, \varnothing) = d(T(y), T(\varnothing)) \leq L \cdot d(y, \varnothing) \Rightarrow d(y, \varnothing) = 0 \Rightarrow y = \varnothing \quad (11.410)$$

Une estimation de la vitesse de convergence est donnée par:

$$d(T^{n+k}(x), T^n(x)) \leq \frac{L^n}{1 - L} \cdot d(T(x), x) \quad (11.411)$$

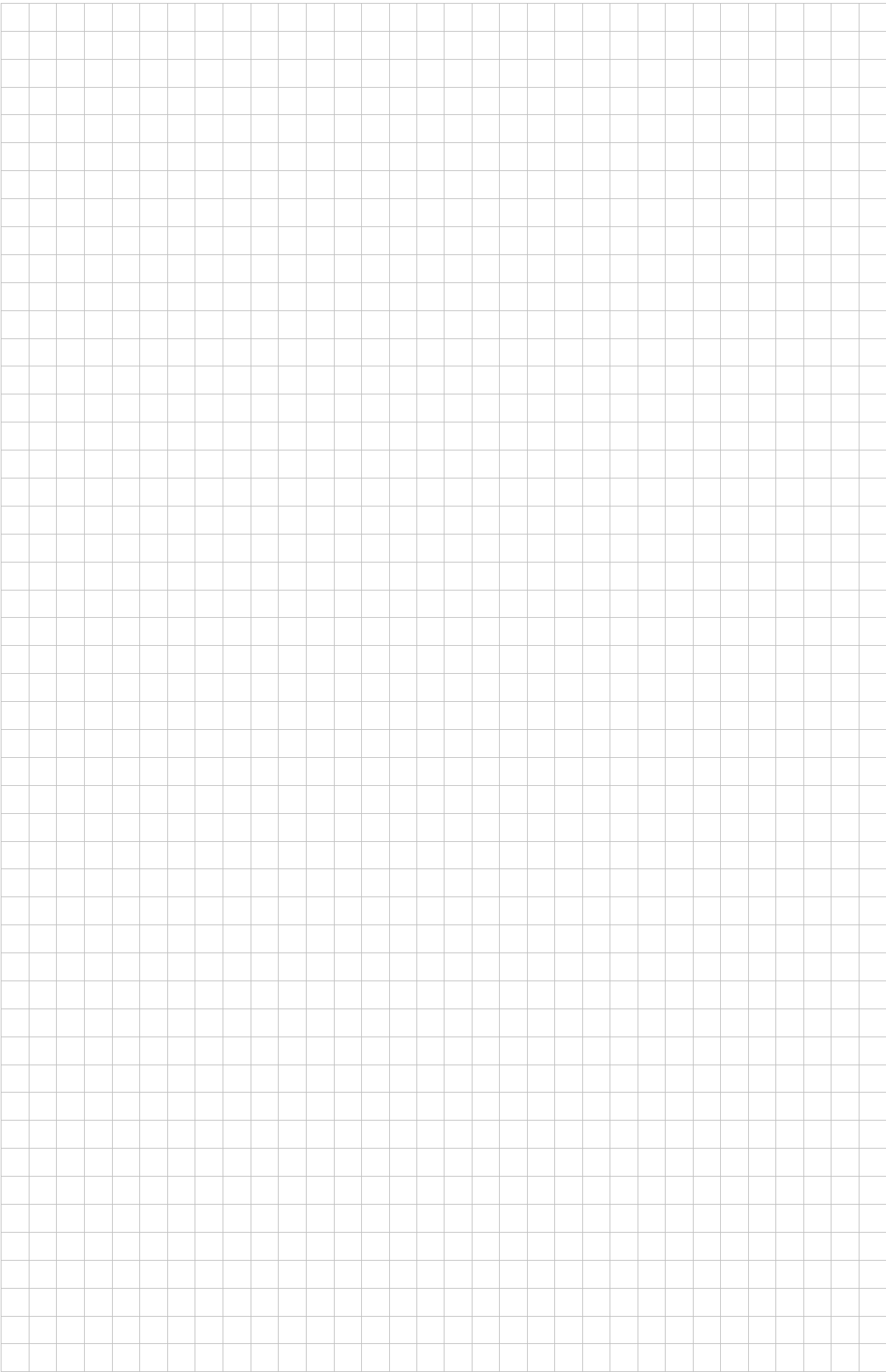
$d(,)$ est continue par rapport à chacune des variables donc:

$$\lim_{k \rightarrow \infty} d(T^{n+k}(x), T^n(x)) = d(\lim_{k \rightarrow \infty} T^{n+k}(x), T^n(x)) = d(\varnothing, T^n(x)) \quad (11.412)$$

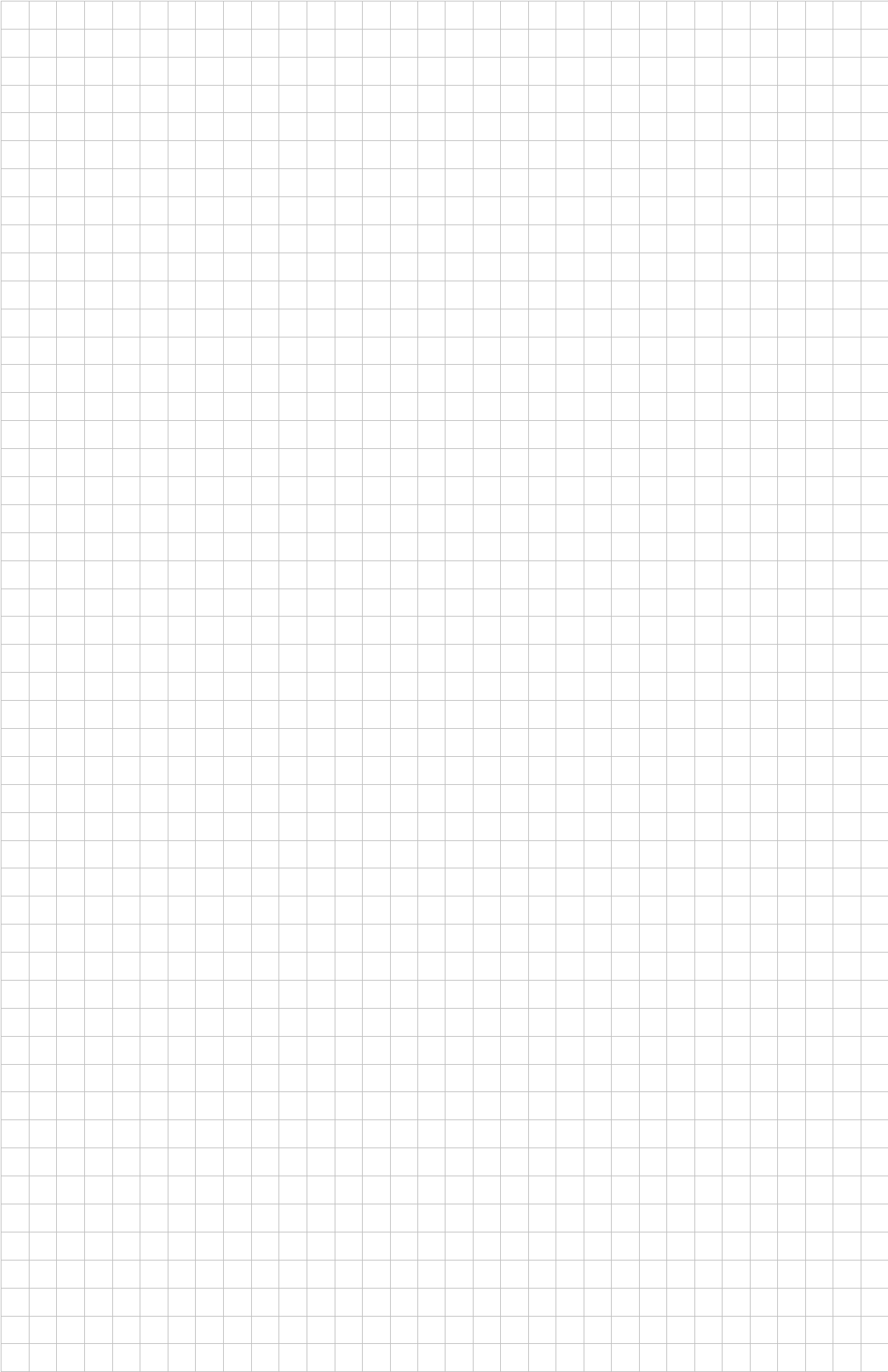
et les limites préservent les inégalités (non strictes) donc:

$$d(\varpi, T^n(x)) \leq \frac{L^n}{1-L} \cdot d(T(x), x) \quad (11.413)$$

□C.Q.F.D.







12. CALCUL VECTORIEL

Le calcul vectoriel ou "**analyse vectorielle**" est une branche des mathématiques qui étudie les champs de scalaires et de vecteurs suffisamment réguliers des espaces euclidiens (voir définition plus loin).

L'importance du calcul vectoriel provient de son utilisation intensive en physique et dans les sciences de l'ingénieur. C'est de ce point de vue que nous le présenterons, et c'est pourquoi nous nous limiterons le plus souvent au cas de l'espace usuel à trois dimensions. Dans ce cadre, un champ de vecteurs associe à chaque point de l'espace un vecteur (à trois composantes réelles), tandis qu'un champ de scalaires y associe un réel.

Remarque: Imaginons par exemple l'eau d'un lac. La donnée de sa température en chaque point forme un champ de scalaires, celle de sa vitesse en chaque point, un champ de vecteurs (voir définition plus loin).

Des notions physiques telles que la force ou la vitesse sont caractérisées par une direction, un sens et une intensité. Ce triple caractère est mis en évidence par les flèches. Celles-ci sont à l'origine de la notion de vecteur et en constituent l'exemple le plus suggestif. Bien que leur nature soit essentiellement géométrique, c'est leur aptitude à se lier les unes aux autres, donc leur comportement algébrique, qui retiendra principalement notre attention. Partagé en classes d'équivalence l'ensemble qu'elles forment représente le modèle classique d'un "espace vectoriel". Un de nos premiers objectifs est la description détaillée de ce modèle.

Remarques:

R1. Avant de lire ce qui va suivre, nous conseillons au lecteur d'avoir au moins parcouru en diagonale le chapitre traitant de la théorie des ensembles dans la section d'arithmétique. Nous y définissons ce qu'est un "espace vectoriel" en utilisant les outils de la théorie des ensembles. Ce concept bien que non absolument indispensable vaut la peine quand même de s'y attarder pour voir comment deux domaines des mathématiques s'imbriquent et aussi histoire... d'aborder les choses au moins un peu rigoureusement.

R2. L'analyse vectorielle contient beaucoup de termes et de définitions qu'il faut apprendre par coeur. Ce travail est pénible mais malheureusement nécessaire.

1. NOTION DE FLÈCHE

Nous désignerons par U l'espace ordinaire de la géométrie élémentaire et par $P, Q, \dots$ ses points. Nous appellerons "**flèche**" tout segment de droite orienté (dans l'espace).

La flèche d'origine P et d'extrémité Q sera notée $\overrightarrow{PQ}$ ou abrégée par une lettre unique (latine ou grecque) choisie arbitrairement telle que par exemple: $\vec{F}$.

Nous considérerons comme évident que toute flèche est caractérisée par sa direction, son sens (car

pour une direction donnée elle peut pointer dans deux sens), son intensité ou grandeur (longueur) et ainsi que son origine.

2. ENSEMBLE DES VECTEURS

Définitions:

D1. Nous disons que deux flèches sont "**équivalentes**" si elles ont la même direction, le même sens et la même intensité.

D2. Nous disons que deux flèches sont "**colinéaires**" si elles ont seulement la même direction.

Partageons l'ensemble des flèches en classes d'équivalence: deux flèches appartiennent à une même classe si et seulement si elles sont équivalentes.

D3. Chaque classe d'équivalence de flèches constitue un "**vecteur**" ou plus exactement un "**vecteur libre**" car son origine n'est pas prise en compte (dans le cas où son origine est bien définie, nous avons alors un "**vecteur lié**").

Rangeons, en outre, les flèches dégénérées (c'est-à-dire de la forme $\overrightarrow{PP}$) en une classe distinguée que nous appellerons "**vecteur nul**" et noterons $\vec{0}$ qui ont une direction et un sens non définis... et d'intensité nulle.

L'ensemble des vecteurs sera lui désigné par V . Il faut souligner que les éléments de V sont des classes de flèches et non pas des flèches individuelles. Il est cependant clair qu'une flèche quelconque suffit à déterminer la classe à laquelle elle appartient et il est donc naturel de l'appeler "représentant de la classe" du vecteur.

Traçons le représentant d'un vecteur $\vec{y}$ à partir de l'extrémité d'un représentant d'un vecteur $\vec{x}$. La flèche dont l'origine est celle du représentant de $\vec{x}$ et l'extrémité celle du représentant de $\vec{y}$ détermine un vecteur que nous noterons $\vec{x} + \vec{y}$. L'opération qui associe à tout couple de vecteurs leur somme s'appelle "addition vectorielle".

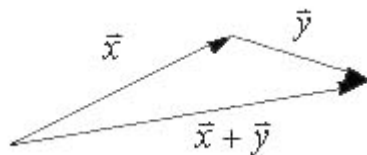


Figure: 12.1 - Exemple d'un somme de deux vecteurs

A l'aide d'une figure, il est facile de montrer que l'opération d'addition vectorielle est associative et commutative, autrement dit, que:

$$(\vec{x} + \vec{y}) + \vec{z} = \vec{x} + (\vec{y} + \vec{z}) \quad (12.1)$$

et:

$$\vec{x} + \vec{y} = \vec{y} + \vec{x} \quad (12.2)$$

Il est en outre évident que le vecteur nul $\vec{0}$ est l'élément neutre de l'addition vectorielle, autrement dit, que:

$$\vec{x} + \vec{0} = \vec{x} \text{ et } \vec{x} + (-\vec{x}) = \vec{0} \quad (12.3)$$

où $-\vec{x}$ désigne le vecteur opposé de $\vec{x}$, c'est-à-dire le vecteur dont les représentants ont la même direction et la même intensité que ceux de $\vec{x}$, mais le sens opposé. Deux vecteurs dont la somme est nulle sont alors appelés "**vecteurs opposés**" puisque la seule chose qui les différencie est leur sens...

Il s'ensuit aussi que si deux ou plusieurs vecteurs ont la même direction, la même intensité et le même sens alors ce sont des "**vecteur égaux**".

L'opération inverse de l'addition vectorielle est la soustraction vectorielle. Soustraire un vecteur revient à additionner le vecteur opposé.

Remarques:

R1. L'addition s'étend, par récurrence, au cas d'une famille finie quelconque de vecteurs. En vertu de l'associativité, ces additions successives peuvent être effectuées dans n'importe quel ordre, ce qui justifie l'écriture sans parenthèses.

R2. La multiplication entre deux vecteurs est un concept qui n'existe pas. Par contre, comme nous le verrons un peu plus loin, nous pouvons multiplier les vecteurs par certaines propriétés d'autres vecteurs que nous appelons la "norme" et encore d'autres petites choses...

2.1. PSEUDO-VECTEURS

En physique, lors de l'énoncé de ce que nous appelons le "principe de Curie", les physiciens font mention de ce qu'ils appellent des "pseudo-vecteurs". Il s'agit du vocabulaire simple pour parler de quelque chose de tout aussi trivial mais fondamentalement peu de gens en font vraiment usage. Mais il peut quand même être utile de présenter de quoi il s'agit.

Au fait, vecteurs et pseudo-vecteurs se transforment de la même manière dans une rotation ou une translation (nous verrons plus tard dans ce chapitre comment effectuer mathématiquement ces transformations). Il n'en est pas de même dans la symétrie par rapport à un plan ou à un point. Dans ces transformations nous avons par définition les propriétés suivantes:

P1. Un vecteur est transformé en son symétrique.

P2. Un pseudo-vecteur est transformé en l'opposé de son symétrique.

Voici une figure avec des exemples types (le choix des lettres représentant les vecteurs n'est pas dû au hasard; elles sont un clin d'oeil aux propriétés des champs électriques et magnétiques étudiés dans la section d'Électromagnétisme du site):

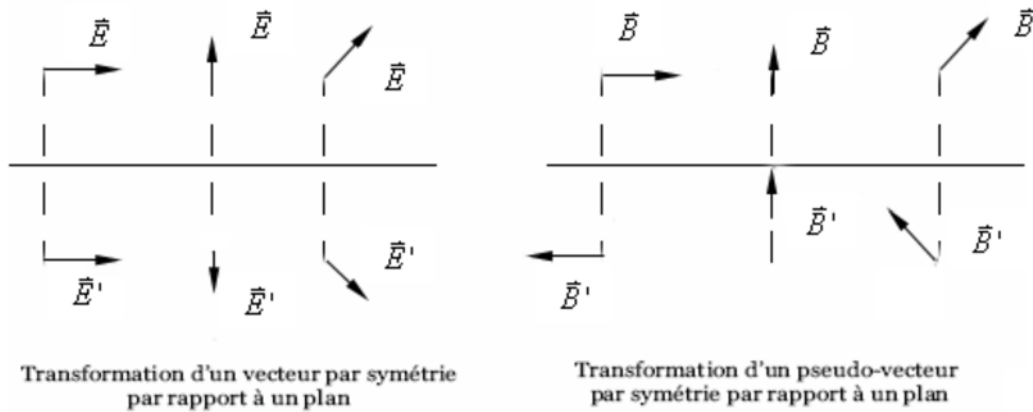


Figure: 12.2 - Différences de transformations entre un vecteur et un pseudo-vecteur

Ben voilà... c'est tout sur les pseudo-vecteurs...

2.2. MULTIPLICATION PAR UN SCALAIRE

Le vecteur $\alpha \cdot \vec{x}$ appelé "produit du nombre α par $\vec{x}$ ", est défini de la manière suivante:

Prenons une flèche représentative de $\vec{x}$ et construisons une flèche de même direction, de même sens ou de sens opposé, suivant que α est positif ou négatif, et d'intensité $|\alpha|$ fois l'intensité de la flèche initiale; la flèche ainsi obtenue est un représentant du vecteur $\alpha \cdot \vec{x}$; si $\alpha = 0$ ou $\vec{x} = \vec{0}$, nous posons $\alpha \cdot \vec{x} = \vec{0}$.

L'opération qui consiste à effectuer le produit d'un nombre par un vecteur est appelé "**multiplication par un scalaire**".

Nous vérifions aisément que la multiplication par un scalaire est associative et distributive par rapport à l'addition numérique vectorielle, autrement dit que:

$$\begin{aligned}\alpha \cdot (\beta \cdot \vec{x}) &= (\alpha \cdot \beta) \cdot \vec{x} \\ (\alpha + \beta) \cdot \vec{x} &= \alpha \cdot \vec{x} + \beta \cdot \vec{x} \quad (12.4) \\ \alpha (\vec{x} + \vec{y}) &= \alpha \vec{x} + \alpha \vec{y}\end{aligned}$$

Voyons de suite un exemple concret mondialement connu des vecteurs:

2.2.1. RÈGLE DE TROIS

Revenons un peu sur la "**règle de trois**" (appelée parfois "**règles des rapports et proportions**" ou encore "**méthode de réduction à l'unité**") souvent définie dans les petites classes de manière intuitive mais sans démonstration digne de ce nom. Cette règle est certainement l'algorithme le plus usité de par le monde qui sert à identifier un quatrième nombre quand trois sont donnés et que les quatre nombres sont linéairement dépendants.

La règle de trois est dérivée sous deux versions:

V1. Simple et directe si les grandeurs sont directement proportionnelles

V2. Simple et inverse si les grandeurs sont inversement proportionnelles

et lorsque deux variables X et Y sont proportionnelles nous le notons:

$$X \propto Y \quad (12.5)$$

Supposons maintenant que X puisse prendre les valeurs x_1, x_2 . Y prendra les valeurs linéairement dépendantes y_1, y_2 alors le rapport proportionnel suivant:

$$\boxed{\frac{x_1}{x_2} = \frac{y_1}{y_2} = c^{te}} \quad (12.6)$$

est dit "**rapport simple et direct**".

Démonstration:

Soient deux vecteurs $\vec{X} = (x_1, x_2), \vec{Y} = (y_1, y_2)$ colinéaires et donc proportionnels à un facteur près tels que:

$$\vec{X} = \lambda \vec{Y} \Leftrightarrow \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \lambda \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \Leftrightarrow \lambda = \frac{x_1}{y_1} = \frac{x_2}{y_2} \Leftrightarrow \frac{y_1}{y_2} = \frac{x_1}{x_2} \quad (12.7)$$

□C.Q.F.D.

Remarque: Si ce rapport n'est pas égal, alors il faut passer à l'utilisation d'autres outils tels que la régression et in extenso l'extrapolation.

- Le rapport proportionnel suivant:

$$\boxed{\frac{x_1}{x_2} = \frac{y_2}{y_1} = c^{te}} \quad (12.8)$$

est dit "**rapport simple et inverse**".

Démonstration:

Soient deux vecteurs $\vec{X} = (x_1, x_2), \vec{Y} = (y_1, y_2)$ colinéaires et donc proportionnels à un facteur près tel que:

$$\vec{X} = \lambda \vec{Y} \Leftrightarrow \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \lambda \begin{pmatrix} 1/y_1 \\ 1/y_2 \end{pmatrix} \Leftrightarrow \lambda = y_1 x_1 = y_2 x_2 \Leftrightarrow \frac{x_1}{x_2} = \frac{y_2}{y_1} \quad (12.9)$$

□C.Q.F.D.

Remarque: Si ce rapport n'est pas égal, alors il faut passer à la régression linéaire (cf. chapitre de Méthodes Numériques).

En gros, il suffit que nous connaissions trois variables sur les quatre pour résoudre cette simple équation

du premier degré.

Les conversions de monnaies ou d'unités de mesure se font à l'aide de la règle de trois simple directe ou indirecte. Les calculs de parités (calcul prévisionnel fait par un importateur d'un certain pays ayant ses propres unités de mesures et de monnaie, qui recherche parmi plusieurs offres étrangères (dans des systèmes d'unités de mesures et de monnaies qui diffèrent de l'importateur), laquelle est la plus avantageuse ou inversement) se font également avec la règle de trois.

Remarque: Nous appelons également "**règle conjointe simple ou inverse**", une série de règles de trois directes ou indirectes.

Dans de tels calculs, les agents du marché d'échange ont remarqué que la plupart du temps, les rapports étaient des valeurs proches de l'unité. Ils ont été ainsi naturellement amenés à définir "le pourcentage" comme étant la proportion d'une quantité ou d'une grandeur par rapport à une autre, évaluée à la centaine (en général du moins ...):

Soit un nombre $x \in \mathbb{R}$ alors sa notation en pourcentage sera:

$$x\% = x \cdot 100 \quad (12.10)$$

Soit un nombre $x \in \mathbb{R}$ alors sa notation en pour-mille sera:

$$x\text{‰} = x \cdot 1000 \quad (12.11)$$

3. ESPACES VECTORIELS

Définition: Nous appelons "**espace vectoriel**" un ensemble E d'éléments désignés par $\vec{x}, \vec{y}, \dots$ et appelés "**vecteurs**", muni d'une "**structure algébrique vectorielle**" définie par la donnée de l'addition (soustraction) vectorielle et la multiplication par un scalaire. Ces deux opérations satisfaisant les lois d'associativité, de commutativité, de distributivité, d'élément neutre et d'opposé comme nous l'avons déjà vu dans le chapitre de Théorie Des Ensembles.

Pour plus d'informations sur ce qu'est un espace vectoriel dans le sens "ensembliste" le lecteur devra se reporter au chapitre de Théorie Des Ensembles où ce concept est défini avec plus de rigueur.

Remarque: Muni de ces deux opérations, un espace vectoriel est dit "**vectorialisé**".

Pour tout entier positif n , $\mathbb{R}^n$ désignera l'ensemble des n -uplets de nombres disposés en colonne:

$$\begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_n \end{pmatrix} \quad (12.12)$$

et $\mathbb{R}^n$ est à l'évidence munie d'une structure d'espace vectoriel. Les vecteurs de cet espace seront appelés "**vecteurs-colonne**". Ils seront souvent désignés plus brièvement par:

$$(\alpha_i)_{1 \leq i \leq n} \quad (12.13)$$

ou encore plus simplement par:

$$(\alpha_i) \quad (12.14)$$

Le nombre α_i est parfois appelé "**terme**" ou "**composante d'indice i** " de (α_i) .

3.1. COMBINAISONS LINÉAIRES

Dorénavant, sauf mention explicite du contraire, les vecteurs seront les éléments d'un espace vectoriel E .

Définition: Nous appelons "**combinaison linéaire**" des vecteurs $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k$ tout vecteur de la forme:

$$\alpha_1 \vec{x}_1 + \alpha_2 \vec{x}_2 + \dots + \alpha_k \vec{x}_k \quad (12.15)$$

où $\alpha_1, \alpha_2, \dots, \alpha_k$ sont des nombres appelés "**coefficients de la combinaison linéaire**".

Le vecteur nul est combinaison linéaire de $\alpha_1 \vec{x}_1 + \alpha_2 \vec{x}_2 + \dots + \alpha_k \vec{x}_k$ avec tous les coefficients égaux à zéro. Nous parlons dès lors de "**combinaison linéaire triviale**".

Définition: Nous appelons "**combinaison convexe**", toute combinaison linéaire dont les coefficients sont non négatifs et de somme égale à 1. L'ensemble des combinaisons convexes de deux points P et Q d'un espace ponctuel P_o (ayant une origine) est le segment de droite P et Q . Pour s'en rendre compte, il suffit d'écrire:

$$\alpha P + (1 - \alpha)Q = Q + \alpha(P - Q) \quad (12.16)$$

de faire varier α de 0 à 1 et de constater que tous les points du segment sont ainsi obtenus.

Si le vecteur $\vec{x}$ est combinaison linéaire des vecteurs $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k$ et chacun de ces vecteurs est combinaison linéaire des vecteurs $\vec{y}_1, \vec{y}_2, \dots, \vec{y}_k$, alors $\vec{x}$ est combinaison linéaire de $\vec{y}_1, \vec{y}_2, \dots, \vec{y}_k$.

3.2. SOUS-ESPACES VECTORIELS

Définition: Nous appelons "**sous-espace vectoriel de E** " tout sous-ensemble de E qui est lui-même un espace vectoriel pour les opérations d'addition et de multiplication par un scalaire définies dans E .

Un sous-espace vectoriel, en tant qu'espace vectoriel, ne peut être vide puisqu'il comprend au moins un vecteur, à savoir son vecteur nul, celui-ci étant d'ailleurs forcément le vecteur nul de E . En outre, en même temps que les vecteurs $\vec{x}$ et $\vec{y}$ (s'il en contient d'autres que le vecteur nul), il comprend également toutes leurs combinaisons linéaires $\alpha \vec{x} + \beta \vec{y}$.

Inversement, nous voyons aussitôt que tout sous-ensemble jouissant de ces propriétés est un sous-espace vectoriel. Nous avons ainsi établi la proposition suivante:

Un sous ensemble S de E est un sous-espace vectoriel de E si et seulement si S est non vide et $\alpha \vec{x} + \beta \vec{y}$ appartient à S pour tout couple $(\vec{x}, \vec{y})$ de vecteurs de S et tout couple $\alpha, \beta \in \mathbb{R}$.

3.3. FAMILLES GÉNÉRATRICES

Il en découle que si nous avons une famille de vecteurs $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k)$ l'ensemble des combinaisons linéaires de $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k$ peut être un sous-espace vectoriel S de E , plus précisément le plus petit sous-espace vectoriel de E contenant $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k$.

Les vecteurs $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k$ qui satisfont à la condition ci-dessus sont appelés "**générateurs**" de S et la famille $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k)$, famille génératrice de S . Nous disons aussi que ces vecteurs ou cette famille engendrent S .

Remarque: Le sous-espace vectoriel engendré par un vecteur non nul est formé de tous les multiples de ce vecteur. Nous appelons un tel sous-espace "**droite vectorielle**". Un sous-espace vectoriel engendré par deux vecteurs non multiples l'un de l'autre est appelé "**plan vectoriel**".

3.4. DÉPENDANCES ET INDÉPENDANCES LINÉAIRES

Ce qui va suivre est très important en physique: nous conseillons donc au futur physicien de prendre vraiment le temps de bien lire les développements.

Si $\vec{e}_1, \vec{e}_2, \vec{e}_3$ sont trois vecteurs de V^3 dont les représentants ne sont pas parallèles à un même plan (par convention une flèche d'intensité nulle est parallèle à tout plan), alors tout vecteur $\vec{x}$ de V^3 peut s'écrire de manière unique sous la forme:

$$\vec{x} = \alpha_1 \vec{e}_1 + \alpha_2 \vec{e}_2 + \alpha_3 \vec{e}_3 \quad (12.17)$$

où $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$ sont des nombres.

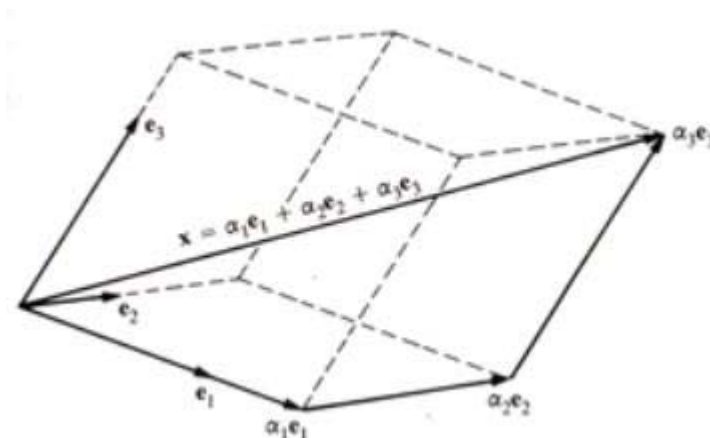


Figure: 12.3 - Exemple d'un construction d'un vecteur dans un espace à trois dimensions

En particulier, la seule possibilité d'obtenir le vecteur nul comme combinaison linéaire de $\vec{e}_1, \vec{e}_2, \vec{e}_3$ est d'attribuer la valeur triviale 0 à $\alpha_1, \alpha_2, \alpha_3$.

Réciproquement, si pour trois vecteurs $\vec{e}_1, \vec{e}_2, \vec{e}_3$ de V^3 la relation:

$$\vec{x} = \alpha_1 \vec{e}_1 + \alpha_2 \vec{e}_2 + \alpha_3 \vec{e}_3 = 0 \quad (12.18)$$

implique $\alpha_1 = \alpha_2 = \alpha_3 = 0$, aucun des vecteurs ne peut être combinaison linéaire des deux autres, autrement dit, leurs représentants ne sont pas parallèles à un même plan.

Sur la base de ces observations, nous allons étendre la notion d'absence de parallélisme à un même plan au cas d'un nombre quelconque de vecteurs d'un espace vectoriel E .

Nous disons que les vecteurs $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k$ sont "**linéairement indépendants**" si la relation:

$$\alpha_1 \vec{e}_1 + \alpha_2 \vec{e}_2 + \dots + \alpha_k \vec{e}_k = 0 \quad (12.19)$$

implique nécessairement $\alpha_1 = \alpha_2 = \dots = \alpha_k = 0$, autrement dit, si la combinaison linéaire triviale est la seule combinaison linéaire de $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k$ qui soit nulle. Dans le cas contraire, nous disons que les vecteurs $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k$ sont "**linéairement dépendants**".

Si l'attention est fixée sur la famille $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k)$ plutôt que sur les termes dont elle est constituée, nous disons que celle-ci est une "**famille libre**" ou "**famille liée**" suivant que les vecteurs $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k$ sont linéairement indépendants ou dépendants.

3.5. BASES D'UN ESPACE VECTORIEL

Définition: Nous disons qu'une famille finie de vecteurs est une base de E si et seulement si:

1. Elle est libre.
2. Elle engendre E .

D'après cette définition, toute famille libre $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k$ est une base du sous-espace vectoriel qu'elle engendre.

Exemple:

Si nous considérons $\mathbb{C}$ comme $\mathbb{R}$ -espace vectoriel (cf. chapitre de Théorie des Ensembles), alors puisque tous les éléments de $\mathbb{C}$ s'écrivent $a + ib$, les éléments qui engendrent $\mathbb{C}$ sont 1 et i (les deux sont libres).

Une base de $\mathbb{C}$ (qui est de dimension 2) comme $\mathbb{R}$ -espace vectoriel est donc la famille finie libre $\{1, i\}$.

Pour qu'une famille de vecteurs $(\vec{e}_1, \vec{e}_2, \dots, \vec{e}_n)$ soit une base de E , il faut et il suffit donc que tout vecteur $\vec{x}$ de E s'exprime de manière unique sous la forme d'une combinaison linéaire des vecteurs $\vec{e}_1, \vec{e}_2, \dots, \vec{e}_n$:

$$\vec{x} = x_1 \vec{e}_1 + x_2 \vec{e}_2 + \dots + x_n \vec{e}_n \quad (12.20)$$

La relation ci-dessus est une décomposition de $\vec{x}$ suivant la base $(\vec{e}_1, \vec{e}_2, \dots, \vec{e}_n)$ où les coefficients $x_1, x_2, \dots, x_n$ sont les composantes de $\vec{x}$ dans cette base. En présence d'une base, tout vecteur est donc entièrement déterminé par ses composantes.

Proposition:

Si $x_1, x_2, \dots, x_n$ sont les composantes de $\vec{x}$ et $y_1, y_2, \dots, y_n$ celles de $\vec{y}$, alors:

$$x_1 + y_1, x_2 + y_2, \dots, x_n + y_n \quad (12.21)$$

sont les composantes de $\vec{x} + \vec{y}$.

En d'autres termes, additionner deux vecteurs revient à additionner leurs composantes et multiplier un vecteur par un scalaire revient évidemment à multiplier ses composantes par ce même scalaire. La base est donc un outil important car elle permet d'effectuer les opérations sur les vecteurs au moyen d'opérations sur les nombres.

Exemple:

Les vecteurs-colonnes de $\mathbb{R}^n$:

$$\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \quad (12.22)$$

forment une base que nous appelons "**base canonique**" de $\mathbb{R}^n$ (nous travaillerons dans les espaces complexes dans un autre chapitre).

Remarque: Dans le cadre de l'espace à trois dimensions, les bases sont très souvent assimilées à un trièdre (effectivement si vous reliez les extrémités des trois vecteurs par des traits vous obtiendrez un trièdre imaginaire).

3.6. ANGLES DIRECTEURS

Il est clair qu'un seul angle ne peut décrire la direction d'un vecteur dans l'espace. Nous utilisons alors la notion "**d'angles directeurs**". Il s'agit de mesurer l'angle du vecteur $\vec{U}$ par rapport à chacun des axes positifs de la base:

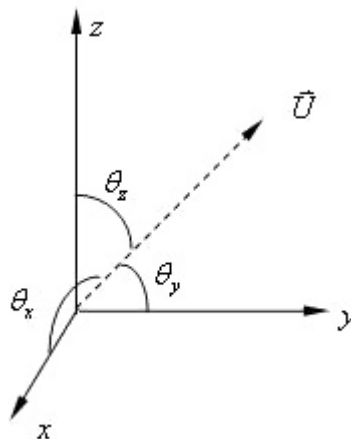


Figure: 12.4 - Représentation des angles directeurs

Si:

$$\vec{U} = (u_1, u_2, u_3) = u_1 \cdot \vec{i} + u_2 \cdot \vec{j} + u_3 \cdot \vec{k} \quad (12.23)$$

alors:

$$\begin{aligned} u_1 &= \|\vec{U}\| \cos(\theta_x) \\ u_2 &= \|\vec{U}\| \cos(\theta_y) \\ u_3 &= \|\vec{U}\| \cos(\theta_z) \end{aligned} \quad (12.24)$$

Les valeurs:

$$\cos(\theta_x), \cos(\theta_y), \cos(\theta_z) \quad (12.25)$$

sont appelées les "**cosinus directeurs**" de $\vec{U}$.

Les 3 angles mentionnés ne sont pas complètement indépendants. En effet, 2 suffisent pour déterminer complètement la direction d'un vecteur dans l'espace, le troisième pouvant se déduire de la relation suivante (obtenue à partir du calcul de la norme du vecteur, concept que nous verrons un peu plus loin):

$$\cos^2(\theta_x) + \cos^2(\theta_y) + \cos^2(\theta_z) = 1 \quad (12.26)$$

De plus, les cosinus directeurs sont les composantes scalaires d'un vecteur de norme unitaire $\vec{u}$ ayant la même direction que $\vec{U}$:

$$\frac{\vec{U}}{\|\vec{U}\|} = \vec{u} = (\cos(\theta_x), \cos(\theta_y), \cos(\theta_z)) \quad (12.27)$$

3.7. DIMENSIONS D'UN ESPACE VECTORIEL

Nous disons que E est de "**dimension finie**" s'il est engendré par une famille finie de vecteurs. Dans le cas contraire, nous disons que E est de "**dimension infinie**" (nous aborderons ce type d'espaces dans un autre chapitre). Tout espace vectoriel de dimension finie et non réduit au vecteur nul admet une base. En fait, de toute famille génératrice d'un tel espace nous pouvons extraire une base.

La dimension d'un espace vectoriel est notée:

$$\dim(E) \quad (12.28)$$

Tout espace vectoriel E de dimension finie non nulle n peut être mis en correspondance biunivoque (c'est-à-dire en bijection) avec $\mathbb{R}^n$. Il suffit de choisir une base de E et de faire correspondre à tout vecteur $\vec{x}$ de E le vecteur-colonne dont les termes sont les composantes de $\vec{x}$ dans la base choisie (c'est du blabla de mathématicien mais ce sera utile quand nous aborderons des espaces plus complexes):

$$\vec{x}(x_1, x_2, \dots, x_n) \leftrightarrow \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \quad (12.29)$$

Cette correspondance conserve les opérations d'addition et de multiplication par un scalaire que nous avons déjà vues; en d'autres termes, elle permet d'effectuer les opérations sur les vecteurs par des opérations sur les nombres.

Remarque: Nous disons alors que E et $\mathbb{R}^n$ sont "**isomorphes**" ou que la correspondance est un isomorphisme (cf. chapitre de Théorie Des Ensembles).

3.8. PROLONGEMENT D'UNE FAMILLE LIBRE

Soit $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k)$ une famille libre et $(\vec{v}_1, \vec{v}_2, \dots, \vec{v}_m)$ une famille génératrice de E . Si $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k)$ n'est pas une base de E , nous pouvons extraire une sous-famille $(\vec{v}_{i1}, \vec{v}_{i2}, \dots, \vec{v}_{il})$ de $(\vec{v}_1, \vec{v}_2, \dots, \vec{v}_m)$ de telle manière que la famille $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k, \vec{v}_{i1}, \vec{v}_{i2}, \dots, \vec{v}_{il})$ soit une base de E .

Remarque: Une telle étude a son utilité lors de passage d'espace mathématique ayant des propriétés données à un autre espace ayant des propriétés mathématiques différentes.

Démonstration:

H1. Nous supposons qu'au moins un des vecteurs $\vec{v}_i$ n'est pas combinaison linéaire des vecteurs $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k$, sinon $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k)$ engendrerait E et serait donc une base possible de E . Notons ce vecteur $\vec{v}_{i1}$. La famille $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k, \vec{v}_{i1})$ est alors une famille libre. En effet, la relation:

$$\alpha_1 \vec{x}_1 + \dots + \alpha_k \vec{x}_k + \beta_1 \vec{v}_{i1} = \vec{0} \quad (12.30)$$

implique alors tout d'abord que $\beta_1 = 0$, autrement $\vec{v}_{i1}$ serait combinaison linéaire des vecteurs $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k$, et ensuite $\alpha_1 = \dots = \alpha_k = 0$, puisque les vecteurs $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k$ sont linéairement indépendants.

Si la famille $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k, \vec{v}_{i1})$ engendre E , elle est une base possible de E et le théorème est démontré. Dans le cas contraire, le même raisonnement nous assure l'existence d'un autre vecteur $\vec{v}_{i2} \dots$. Si la nouvelle famille en découlant n'est pas une base de E , alors le procédé d'extraction de vecteurs $\vec{v}_i$ de $(\vec{v}_1, \vec{v}_2, \dots, \vec{v}_m)$ se poursuit. Lorsqu'il s'arrête, nous aurons obtenu un "prolongement" de $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k)$ en une famille libre engendrant E , c'est-à-dire une base de E .

□C.Q.F.D.

Il en retourne un corollaire: Tout espace vectoriel de dimension finie et non réduit au vecteur nul admet

une base. En fait, de toute famille génératrice d'un tel espace, nous pouvons donc extraire une base.

3.9. RANG D'UNE FAMILLE FINIE

Définition: Nous appelons "**rang d'une famille**" de vecteurs la dimension du sous-espace vectoriel de E qu'elle engendre.

Montrons que le rang d'une famille de vecteurs $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k)$ est inférieur ou égal à k et qu'il est égal à k si et seulement si cette famille est libre.

Démonstration:

Ecartons d'abord le cas trivial où le rang de la famille $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k)$ est nul. D'après le corollaire vu précédemment, nous pouvons alors extraire de cette famille une base du sous-espace vectoriel qu'elle engendre. Le rang est donc inférieur ou égal à k suivant que $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k)$ est une famille liée ou non.

□C.Q.F.D.

3.10. SOMMES DIRECTES

Définition: Nous disons que la somme $S+T$ de deux sous-espaces vectoriels S et T de E (cas particulier appliqué à un espace de dimensions 2 !) est une "**somme directe**" si:

$$S \cap T = \{\vec{0}\} \quad (12.31)$$

Dans ce cas, nous la notons:

$$S \oplus T \quad (12.32)$$

En d'autres termes, la somme de deux sous-espaces vectoriels S et T de E est directe si la décomposition de tout élément de $S+T$ en somme d'un élément de S et d'un élément de T est unique.

Ce concept de décomposition trivial va nous être très utile dans certains théorèmes dont le plus important sur ce site est certainement le théorème spectral (cf. chapitre d'Algèbre Linéaire).

De la somme directe nous pouvons introduire la notion de "**sous-espace complémentaire**" appelé encore "**sous-espace supplémentaire**" (selon les pays...):

Supposons que E soit de dimension finie. Pour tout sous-espace vectoriel S de E , il existe un sous-espace vectoriel T (non unique) de E tel que E soit somme directe de S et T . Nous disons alors que T est un "**sous-espace complémentaire**" de S dans E .

Démonstration:

Ecartons d'abord les cas triviaux où $S = \{\vec{0}\}$ et $S=E$. Le sous-espace vectoriel S admet une base $(\vec{e}_1, \vec{e}_2, \dots, \vec{e}_k)$, où k est inférieur à la dimension n de E . Par le théorème du prolongement d'une famille libre, cette base peut se prolonger en une base $(\vec{e}_1, \dots, \vec{e}_k, \vec{e}_{k+1}, \dots, \vec{e}_n)$ de E . Soit T le sous-espace vectoriel engendré par la famille $(\vec{e}_{k+1}, \dots, \vec{e}_n)$. Si $\vec{x}$ est un vecteur quelconque de E , alors $\vec{x} = \vec{s} + \vec{t}$, où $\vec{s}$ est un vecteur de S et $\vec{t}$ un vecteur de T . En outre, $S \cap T = \{\vec{0}\}$ car aucun vecteur, excepté le vecteur nul, ne

peut être combinaison linéaire des vecteurs $\vec{e}_1, \dots, \vec{e}_k$ et des vecteurs $\vec{e}_{k+1}, \dots, \vec{e}_n$. Nous en concluons donc que:

$$E = S \oplus T \quad (12.33)$$

□C.Q.F.D.

4. ESPACE AFFINE

L'espace G de la géométrie élémentaire est à la fois usuel et la source de la notion "d'espace affine" que nous allons introduire.

Cet espace G est associé à "l'espace vectoriel" géométrique V par la correspondance entre flèches et vecteurs étudiés jusqu'ici! La définition suivante ne fait que mettre en évidence les traits dominants de cette correspondance:

Définition: Soit U un ensemble non vide d'éléments que nous appellerons "points" et désignerons par les lettres $P, Q, \dots$; soit en outre E un espace vectoriel. Supposons qu'à tout couple de points (P, Q) corresponde un vecteur noté $\overrightarrow{PQ}$. Nous disons alors que U est un "espace affine" d'espace directeur (ou dit simplement abusivement de "direction") E si les conditions suivantes sont satisfaites:

C1. Pour tout point P fixé, la correspondance entre couples (P, Q) et vecteurs $\vec{x}$ est biunivoque, autrement dit, pour tout vecteur $\vec{x}$ il existe un point Q et un seul tel que $\vec{x} = \overrightarrow{PQ}$.

C2. Pour tout triplet de points (P, Q, R) :

$$\overrightarrow{PQ} + \overrightarrow{QR} = \overrightarrow{PR} \quad (12.34)$$

C'est la fameuse "relation de Chasles"-

C3. Si P est un point et $\vec{x}$ un vecteur, pour exprimer que Q est l'unique point tel que $\vec{x} = \overrightarrow{PQ}$, nous écrivons:

$$Q = P + \vec{x} \quad (12.35)$$

Bien qu'un peu abusive, cette écriture est conforme à l'usage et suggère bien le sens de l'opération qu'elle désigne.

Les propriétés suivantes découlent directement de la définition d'espace affine:

P1. $P + (\vec{x} + \vec{y}) = (P + \vec{x}) + \vec{y}$

P2. Pour tout point P , $\overrightarrow{PP} = \vec{0}$. Cela résulte de la condition $\overrightarrow{PQ} + \overrightarrow{QR} = \overrightarrow{PR}$ dans le cas où nous avons $P = Q = R$.

P3. $\overrightarrow{PQ} = -\overrightarrow{QP}$. Il suffit de poser $R=P$ dans la relation de Chasles $\overrightarrow{PQ} + \overrightarrow{QR} = \overrightarrow{PR}$.

P4. Règle du parallélogramme:

Soit le polygone de sommets (dans le sens des aiguilles d'une montre) P, P', Q', Q et arêtes $\overrightarrow{PP'}, \overrightarrow{P'Q'}, \overrightarrow{QQ'}, \overrightarrow{PQ}$:

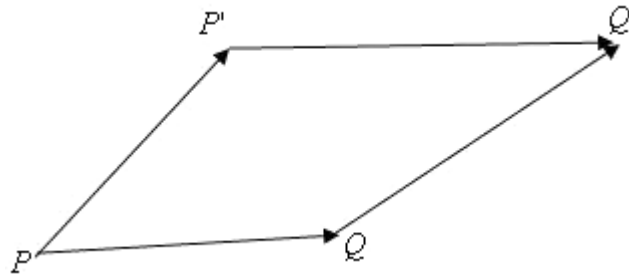


Figure: 12.5 - Polygone vectoriel

Nous avons:

$$\overrightarrow{PP'} = \overrightarrow{QQ'} \quad (12.36)$$

si et seulement si:

$$\overrightarrow{P'Q'} = \overrightarrow{PQ} \quad (12.37)$$

ce qui donnerait alors un parallélogramme!

En effet, en remplaçant R par Q' dans la relation de Chasles il vient:

$$\overrightarrow{PQ} + \overrightarrow{QQ'} = \overrightarrow{PQ'} \quad (12.38)$$

et en faisant de même mais en remplaçant R par Q' et Q par P' nous avons:

$$\overrightarrow{PP'} + \overrightarrow{P'Q'} = \overrightarrow{PQ'} \quad (12.39)$$

Nous avons alors en égalisant ces deux dernières relations:

$$\overrightarrow{PQ} + \overrightarrow{QQ'} = \overrightarrow{PP'} + \overrightarrow{P'Q'} \quad (12.40)$$

ce qui force l'égalité susmentionnée que nous voulions démontrer.

Précédemment, nous avons vu ce qui faisait qu'un espace G pouvait être muni d'une structure d'espace vectoriel (nous avons vu que nous disons que ce dernier était dès lors "vectorialisé"). Dans le cas général d'un espace affine U , le procédé est le même:

Nous choisissons un point quelconque O de U . La correspondance entre couples (O, P) et vecteurs de l'espace directeur étant alors biunivoque nous définissons l'addition de points et la multiplication d'un point par un scalaire par les opérations correspondantes sur les vecteurs de E . Muni de ces deux opérations, U devient un espace vectoriel, appelé "**vectorialisé de U relativement à O** ". Nous désignerons cet espace par $\mathcal{U}_O$ et appellerons O "**l'origine**".

Vu la manière dont les opérations ont été définies, il résulte que $\mathcal{U}_O$ est isomorphe à l'espace directeur E :

$$P = \vec{O} + \vec{x} \leftrightarrow \vec{x} \quad (12.41)$$

Toutefois, cet isomorphisme dépend du choix de l'origine O et en pratique cette origine est choisie sur base de données inhérentes aux problèmes posés. Par exemple, si une transformation affine admet un point invariant (qui ne bouge pas), il y a avantage à choisir ce point comme origine.

Remarques:

R1. Lorsque nous parlons de dimension d'un espace affine, nous parlons de la dimension de son espace directeur.

R2. L'espace G de la géométrie élémentaire est un espace affine. En effet, sa direction est l'espace géométrique V et les conditions de définition d'un espace affine sont satisfaites. Il faut bien noter qu'au couple de points (P, Q) est associé le vecteur $\overrightarrow{PQ}$ et non pas la flèche PQ . En fait, la flèche pouvant être identifiée au couple de points, nous voyons que ce que postule la définition d'un espace affine n'est rien d'autre qu'une forme abstraite de correspondance entre flèches et vecteurs.

R3. Tout espace vectoriel E peut être considéré comme un espace affine de direction E lui-même si au couple de vecteurs $(\vec{x}, \vec{y})$ est associé le vecteur $\vec{y} - \vec{x}$. En effet, les conditions de définition d'un espace affine sont dès lors satisfaites.

5. ESPACES VECTORIELS EUCLIDIENS

Avant de définir ce qu'est un espace vectoriel euclidien, nous allons au préalable définir quelques outils mathématiques et quelques concepts.

Nous pouvons, en choisissant une unité de longueur, mesurer l'intensité de chaque flèche, autrement dit, déterminer sa longueur. Nous pouvons aussi mesurer l'écart angulaire de deux flèches (ou vecteurs) quelconques d'origine commune (non nécessairement distinctes) en prenant comme unité de mesure d'angle par exemple le radian. La mesure de cet écart est alors un nombre compris entre 0 et π , appelé "angle" des deux flèches. Si les deux flèches ont même direction et même sens, leur angle est nul et si elles ont même direction et sens opposé, ce même angle est π .

Les flèches représentatives d'un même vecteur $\vec{x}$ ont toutes la même longueur. Nous désignerons cette longueur par la notation:

$$\|\vec{x}\| \quad (12.42)$$

et l'appellerons "norme" de $\vec{x}$. Il est clair que la longueur d'un vecteur est nulle si et seulement si sa norme est nulle. Nous dirons qu'un vecteur est unitaire si sa norme est 1.

Si $\vec{x}$ est un vecteur non nul:

$$\frac{1}{\|\vec{x}\|} \vec{x} \quad (12.43)$$

est un vecteur unitaire colinéaire (nécessairement...) à $\vec{x}$ dont la norme est égale à l'unité et que nous

notons $\vec{u}$.

Nous appellerons "**angle des vecteurs non nuls**" $\vec{x}$ et $\vec{y}$ l'angle de deux flèches d'origine commune représentant l'une $\vec{x}$ et l'autre $\vec{y}$.

Plus rigoureusement cependant une "**norme**" sur un espace vectoriel réel (ou complexe) E est une application $\| \cdot \| : E \rightarrow \mathbb{R}^+$ vérifiant les propriétés :

P1. Positivité:

$$\forall \vec{x} \in E \quad \|\vec{x}\| \geq 0 \quad (12.44)$$

P2. Linéarité:

$$\forall \lambda \in \mathbb{R}, \forall \vec{x} \in E \quad \|\lambda \vec{x}\| = |\lambda| \cdot \|\vec{x}\| \quad (12.45)$$

P3. Nullité:

$$\forall \vec{x} \in E \quad \|\vec{x}\| = 0 \Leftrightarrow \vec{x} = 0 \quad (12.46)$$

P4. Inégalité de Minkowski (inégalité triangulaire):

$$\forall \vec{x}, \vec{y} \in E \quad \|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\| \quad (12.47)$$

Remarques:

R1. Ces propriétés sont principalement imposées par notre approche intuitive de l'espace euclidien (espace vectoriel de dimension finie sur le corps des réels et muni d'un produit scalaire que nous verrons plus loin) et de son interprétation géométrique.

R2. Nous démontrerons un peu plus loin la propriété P4 sous la dénomination "d'inégalité triangulaire" et nous ferons une étude un peu plus générale de cette inégalité sous la dénomination "d'inégalité de Minkowski" dans le chapitre de Topologie.

5.1. PRODUIT SCALAIRE VECTORIEL

Définition: Un "**espace vectoriel euclidien**", est un espace vectoriel (réel et de dimension finie pour les puristes) possédant une opération particulière, le "**produit scalaire**" que nous noterons (notation spécifique à ce site Internet) en ce qui concerne le cas particulier des vecteurs:

$$\|\vec{x}\| \|\vec{y}\| \cos \theta = \vec{x} \circ \vec{y} \quad (12.48)$$

Remarques:

R1. Nous trouvons dans certains ouvrages (pour information) la notation $(\vec{x} | \vec{y})$ ou encore $\langle \vec{x} | \vec{y} \rangle$ lors de la généralisation de cette définition comme nous le verrons un peu plus loin.

R2. Le produit scalaire a une importance énorme dans l'ensemble du domaine des mathématiques et de la physique; ainsi nous le retrouvons dans le calcul différentiel et intégral (de par le produit scalaire fonctionnel), en topologie, en physique quantique, en analyse du signal etc. Il convient donc de bien comprendre ce qui va suivre.

R3. Le produit scalaire peut être vu comme une projection de la longueur d'un vecteur sur la longueur d'un autre.

Ce produit scalaire possède les propriétés suivantes (dont la plupart découlent de la définition même du produit scalaire) dans un espace euclidien:

P1. Commutativité: $\vec{x} \circ \vec{y} = \vec{y} \circ \vec{x}$

P2. Associativité: $\alpha(\vec{x} \circ \vec{y}) = (\alpha\vec{x}) \circ \vec{y} = \vec{x} \circ (\alpha\vec{y})$

P3. Distributivité: $\vec{x} \circ (\vec{y} + \vec{z}) = \vec{x} \circ \vec{y} + \vec{x} \circ \vec{z}$

P4. Si $\vec{x} \circ \vec{y} = 0 : \forall \vec{x} \neq \vec{0} \Rightarrow \vec{y} = \vec{0}$

P5. $\|\vec{x}\|^2 = \vec{x} \circ \vec{x}$ et $\vec{x} \circ \vec{x} > 0$ si $\vec{x} \neq \vec{0}$

P6. $(\alpha\vec{x} + \beta\vec{y}) \circ \vec{z} = \alpha(\vec{x} \circ \vec{z}) + \beta(\vec{y} \circ \vec{z})$

Seule cette dernière propriété nécessite peut-être une démonstration (de plus un des résultats de la démonstration nous servira à démontrer une autre propriété très importante du produit scalaire):

Démonstration:

Soit:

$$proj_v \vec{x} = \|\vec{x}\| \cos \theta \left(\frac{1}{\|\vec{v}\|} \vec{v} \right) \quad (12.49)$$

qui constitue la "**projection orthogonale vectorielle**" (le v en indice du *proj* signifiant "vectoriel") du vecteur $\vec{x}$ sur la normalisation à l'unité du vecteur $\vec{v}$.

A l'aide du produit scalaire, le vecteur $proj_v \vec{x}$ peut être exprimé autrement il suffit de prendre la relation que nous avons vue plus haut:

$$\|\vec{x}\| \|\vec{v}\| \cos \theta = \vec{x} \circ \vec{v} \Leftrightarrow \cos \theta = \frac{\vec{x} \circ \vec{v}}{\|\vec{x}\| \|\vec{v}\|} \quad (12.50)$$

et de l'introduire dans $proj_{\vec{v}}\vec{x}$ avec les vecteurs adéquats pour obtenir:

$$proj_{\vec{v}}\vec{x} = \|\vec{x}\| \cos \theta \left(\frac{1}{\|\vec{v}\|} \vec{v} \right) = \|\vec{x}\| \left(\frac{\vec{x} \circ \vec{v}}{\|\vec{x}\| \|\vec{v}\|} \right) \left(\frac{1}{\|\vec{v}\|} \vec{v} \right) = \frac{\vec{x} \circ \vec{v}}{\|\vec{v}\|^2} \vec{v} = \frac{\vec{x} \circ \vec{v}}{\vec{v} \circ \vec{v}} \vec{v} \quad (12.51)$$

Cette expression vaut également dans le cas où $\vec{x}$ est nul, à condition d'admettre que la projection orthogonale du vecteur nul est nulle.

La norme de $proj_{\vec{v}}\vec{x}$ s'écrit:

$$\|proj_{\vec{v}}\vec{x}\| = \frac{|\vec{x} \circ \vec{v}|}{\|\vec{v}\|^2} \|\vec{v}\| = \frac{|\vec{x} \circ \vec{v}|}{\|\vec{v}\|} \quad (12.52)$$

Si $\vec{v}$ est unitaire, les relations de projections précédentes se simplifient et deviennent:

$$proj_{\vec{v}}\vec{x} = (\vec{x} \circ \vec{v})\vec{v} \text{ et } \|proj_{\vec{v}}\vec{x}\| = |\vec{x} \circ \vec{v}| \quad (12.53)$$

Par des considérations géométriques élémentaires (distributivité du produit scalaire), il est facile de se rendre compte que:

$$proj_{\vec{v}}(\vec{x} + \vec{y}) = proj_{\vec{v}}\vec{x} + proj_{\vec{v}}\vec{y} \text{ et } proj_{\vec{v}}\alpha\vec{x} = \alpha proj_{\vec{v}}\vec{x} \quad (12.54)$$

Si nous revenons maintenant à la démonstration de:

$$(\alpha\vec{x} + \beta\vec{y}) \circ \vec{z} = \alpha(\vec{x} \circ \vec{z}) + \beta(\vec{y} \circ \vec{z}) \quad (12.55)$$

Nous avons donc dans un premier temps:

$$proj_{\vec{v}}(\alpha\vec{x} + \beta\vec{y}) = proj_{\vec{v}}\alpha\vec{x} + proj_{\vec{v}}\beta\vec{y} = \alpha proj_{\vec{v}}\vec{x} + \beta proj_{\vec{v}}\vec{y} \quad (12.56)$$

et, d'après la définition la propriété de la projection orthogonale, il vient alors immédiatement en faisant la correspondance terme à terme:

$$\begin{aligned} proj_{\vec{v}}(\alpha\vec{x} + \beta\vec{y}) &= \alpha proj_{\vec{v}}\vec{x} + \beta proj_{\vec{v}}\vec{y} \\ \frac{(\alpha\vec{x} + \beta\vec{y}) \circ \vec{z}}{\vec{z} \circ \vec{z}} \vec{z} &= \alpha \frac{\vec{x} \circ \vec{z}}{\vec{z} \circ \vec{z}} \vec{z} + \beta \frac{\vec{y} \circ \vec{z}}{\vec{z} \circ \vec{z}} \vec{z} \end{aligned} \quad (12.57)$$

d'où la propriété P6 qui s'ensuit par multiplication des deux membres de l'égalité par $\vec{z} \circ \vec{z}$ et après simplification par $\vec{z}$.

□C.Q.F.D.

Définitions:

D1. L'espace vectoriel E est dit "**espace vectoriel proprement euclidien**" si $\forall \vec{x} \in E \rightarrow \|\vec{x}\| \geq 0$.

D2. Nous disons que les vecteurs $\vec{x}$ et $\vec{y}$ sont des "**vecteurs orthogonaux**" s'ils sont non nuls et que leur produit scalaire est nul (leur angle est égal à $\pi/2$).

Une base $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ de V est dite "**base orthonormale**" si les vecteurs $\vec{e}_1, \vec{e}_2, \vec{e}_3$ sont orthogonaux deux à deux et unitaires (donc constituant une famille libre).

Remarque: Nous verrons en calcul tensoriel (nous aurions pu le faire ici aussi mais bon...) comment à partir d'un ensemble de vecteurs indépendants construire une base orthogonale. C'est ce que le lecteur pourra trouver sous le nom de "**méthode d'orthogonalisation de (Gram-)Schmidt**".

Par le raisonnement géométrique, nous voyons que tout vecteur est la somme de ses projections orthogonales sur les vecteurs d'une base orthonormale, autrement dit, si $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ est une base orthonormale:

$$\vec{x} = (\vec{x} \circ \vec{e}_1) \vec{e}_1 + (\vec{x} \circ \vec{e}_2) \vec{e}_2 + (\vec{x} \circ \vec{e}_3) \vec{e}_3 \quad (12.58)$$

Cette décomposition s'obtient également par la propriété de P6 du produit scalaire. En effet, x_1, x_2, x_3 étant les composantes de $\vec{x}$:

$$\vec{x} \circ \vec{e}_1 = (x_1 \vec{e}_1 + x_2 \vec{e}_2 + x_3 \vec{e}_3) \circ \vec{e}_1 = x_1 (\vec{e}_1 \circ \vec{e}_1) + x_2 (\vec{e}_2 \circ \vec{e}_1) + x_3 (\vec{e}_3 \circ \vec{e}_1) = x_1 \quad (12.59)$$

puisque $\vec{e}_1 \circ \vec{e}_1 = 1$ et $\vec{e}_2 \circ \vec{e}_1 = \vec{e}_3 \circ \vec{e}_1 = 0$ de même:

$$\vec{x} \circ \vec{e}_2 = x_2 \text{ et } \vec{x} \circ \vec{e}_3 = x_3 \quad (12.60)$$

d'où la décomposition.

Soit x_1, x_2, x_3 et y_1, y_2, y_3 les composantes respectives des vecteurs $\vec{x}$ et $\vec{y}$ dans une base orthonormale canonique $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ nous pouvons écrire le produit scalaire sous la forme:

$$\vec{x} \circ \vec{y} = (x_1 \vec{e}_1 + x_2 \vec{e}_2 + x_3 \vec{e}_3) \circ (y_1 \vec{e}_1 + y_2 \vec{e}_2 + y_3 \vec{e}_3) \quad (12.61)$$

de la propriété P6 du produit scalaire:

$$\vec{x} \circ \vec{y} = x_1 \vec{e}_1 \circ (y_1 \vec{e}_1 + y_2 \vec{e}_2 + y_3 \vec{e}_3) + x_2 \vec{e}_2 \circ (y_1 \vec{e}_1 + y_2 \vec{e}_2 + y_3 \vec{e}_3) + x_3 \vec{e}_3 \circ (y_1 \vec{e}_1 + y_2 \vec{e}_2 + y_3 \vec{e}_3) \quad (12.62)$$

en utilisant la propriété P1 et à nouveau P6:

$$\begin{aligned} \vec{x} \circ \vec{y} &= x_1 (y_1 \vec{e}_1 + y_2 \vec{e}_2 + y_3 \vec{e}_3) \circ \vec{e}_1 + x_2 (y_1 \vec{e}_1 + y_2 \vec{e}_2 + y_3 \vec{e}_3) \circ \vec{e}_2 + x_3 (y_1 \vec{e}_1 + y_2 \vec{e}_2 + y_3 \vec{e}_3) \circ \vec{e}_3 \\ &\Rightarrow \\ (y_1 \vec{e}_1 + y_2 \vec{e}_2 + y_3 \vec{e}_3) \circ \vec{e}_1 &= y_1 \vec{e}_1 \circ \vec{e}_1 + y_2 \vec{e}_2 \circ \vec{e}_1 + y_3 \vec{e}_3 \circ \vec{e}_1 = y_1 \|\vec{e}_1\|^2 = y_1 \\ (y_1 \vec{e}_1 + y_2 \vec{e}_2 + y_3 \vec{e}_3) \circ \vec{e}_2 &= y_1 \vec{e}_1 \circ \vec{e}_2 + y_2 \vec{e}_2 \circ \vec{e}_2 + y_3 \vec{e}_3 \circ \vec{e}_2 = y_2 \|\vec{e}_2\|^2 = y_2 \\ (y_1 \vec{e}_1 + y_2 \vec{e}_2 + y_3 \vec{e}_3) \circ \vec{e}_3 &= y_1 \vec{e}_1 \circ \vec{e}_3 + y_2 \vec{e}_2 \circ \vec{e}_3 + y_3 \vec{e}_3 \circ \vec{e}_3 = y_3 \|\vec{e}_3\|^2 = y_3 \end{aligned} \quad (12.63)$$

Ce qui nous donne finalement la décomposition:

$$\boxed{\vec{x} \circ \vec{y} = x_1 y_1 + x_2 y_2 + x_3 y_3} \quad (12.64)$$

qui constitue l'une des relations les plus importantes dans le domaine du calcul vectoriel et que nous appelons "**produit scalaire canonique**".

5.1.1. INÉGALITÉ DE CAUCHY-SCHWARZ

La relation:

$$(\vec{x} + \vec{y}) \circ (\vec{x} + \vec{y}) = \vec{x} \circ \vec{x} + \vec{x} \circ \vec{y} + \vec{y} \circ \vec{x} + \vec{y} \circ \vec{y} = \vec{x} \circ \vec{x} + \vec{y} \circ \vec{y} + 2\vec{x} \circ \vec{y} \quad (12.65)$$

s'écrit également trivialement sous la forme suivante si nous utilisons la notion de norme et la définition du produit scalaire:

$$(\vec{x} + \vec{y}) \circ (\vec{x} + \vec{y}) = \|\vec{x} + \vec{y}\|^2 = \|\vec{x}\|^2 + \|\vec{y}\|^2 + 2\vec{x} \circ \vec{y} \quad (12.66)$$

Il est intéressant de remarquer que si les deux vecteurs $\vec{x}$ et $\vec{y}$ sont orthogonaux, nous retrouvons le résultat d'un théorème fameux: le théorème de Pythagore!

Effectivement, dès lors nous avons si les deux vecteurs sont orthogonaux:

$$\vec{x} \circ \vec{y} = 0 \quad (12.67)$$

Ce qui nous donne:

$$\|\vec{x} + \vec{y}\|^2 = \|\vec{x}\|^2 + \|\vec{y}\|^2 \quad (12.68)$$

Cette relation est très importante en physique-mathématique. Il faut s'en souvenir !

Nous appelons également "**inégalité de Cauchy-Schwarz**" l'inégalité, valable pour tout choix des vecteurs $\vec{x}$ et $\vec{y}$, la relation:

$$|\vec{x} \circ \vec{y}| \leq \|\vec{x}\| \|\vec{y}\| \quad (12.69)$$

Ce qui s'écrit aussi:

$$|\vec{x} \circ \vec{y}|^2 \leq (\vec{x} \circ \vec{x}) \cdot (\vec{y} \circ \vec{y}) \quad (12.70)$$

D'abord nous considérerons comme évident que l'égalité n'a lieu qu'en cas de colinéarité des deux vecteurs.

Démonstration:

Nous nous plaçons dans le cas où $\vec{x}, \vec{y} \neq \vec{0}$. Alors, pour $\forall \lambda \in \mathbb{R}$ nous avons trivialement selon les propriétés du produit scalaire:

$$(\vec{x} + \lambda \vec{y}) \circ (\vec{x} + \lambda \vec{y}) = \vec{x} \circ \vec{x} + 2\lambda \vec{x} \circ \vec{y} + \lambda^2 \vec{y} \circ \vec{y} \geq 0 \quad (12.71)$$

Il s'agit donc d'une simple équation du deuxième degré où la variable est λ . En se rappelant de ce que nous avons vu lors de notre étude des polynômes du deuxième degré (cf. chapitre de Calcul Algébrique) la relation précédente (le fait qu'elle soit toujours supérieure ou égale à zéro) est satisfaite

que si le discriminant $b^2 - 4ac$ est négatif ou nul. En d'autres termes, si:

$$4(\vec{x} \circ \vec{y})^2 - 4(\vec{x} \circ \vec{x})(\vec{y} \circ \vec{y}) \leq 0 \quad (12.72)$$

Soit après simplification:

$$(\vec{x} \circ \vec{y})^2 - (\vec{x} \circ \vec{x})(\vec{y} \circ \vec{y}) \leq 0 \Rightarrow (\vec{x} \circ \vec{y})^2 \leq (\vec{x} \circ \vec{x})(\vec{y} \circ \vec{y}) \quad (12.73)$$

□C.Q.F.D.

Lorsque E est $\mathbb{R}^n$, l'inégalité de Cauchy-Schwarz s'écrit avec les composantes des vecteurs:

$$\sum_{i=1}^n a_i b_i \leq \sqrt{\sum_{i=1}^n a_i^2 \sum_{i=1}^n b_i^2} \quad (12.74)$$

Dans le cas particulier où $\forall i, b_i = 1$ elle devient:

$$\left(\sum_{i=1}^n a_i \right)^2 \leq n \sum_{i=1}^n a_i^2 \quad (12.75)$$

ou encore:

$$\left(\frac{1}{n} \sum_{i=1}^n a_i \right)^2 \leq \frac{1}{n} \sum_{i=1}^n a_i^2 \quad (12.76)$$

ce qui montre que le carré de la moyenne arithmétique est inférieur ou égal à la moyenne arithmétique des carrés. Ce résultat est important pour l'étude des statistiques.

Par ailleurs, en vertu de la propriété du cosinus et de l'inégalité de Cauchy-Schwarz nous pouvons écrire:

$$-1 \leq \frac{\vec{x} \circ \vec{y}}{\|\vec{x}\| \|\vec{y}\|} \leq 1 \quad (12.77)$$

relation que nous retrouvons également dans le cadre de l'étude des statistiques (cf. chapitre de Statistiques).

5.1.2. INÉGALITÉ TRIANGULAIRE

En majorant $2\vec{x} \circ \vec{y}$ par $2\|\vec{x}\|\|\vec{y}\|$ (de par l'inégalité de Cauchy-Schwarz!) et en mettant ceci dans la relation établie déjà précédemment:

$$\|\vec{x} + \vec{y}\|^2 = \|\vec{x}\|^2 + \|\vec{y}\|^2 + 2\vec{x} \circ \vec{y} \quad (12.78)$$

Nous obtenons:

$$\|\vec{x} + \vec{y}\|^2 \leq \|\vec{x}\|^2 + \|\vec{y}\|^2 + 2\|\vec{x}\|\|\vec{y}\| = (\|\vec{x}\| + \|\vec{y}\|)^2 \quad (12.79)$$

ce qui entraîne la fameuse "**inégalité triangulaire**" (très utile dans l'étude des suites et séries pour l'étude des convergences ainsi qu'en topologie):

$$\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\| \quad (12.80)$$

Remarque: La généralisation de cette inégalité relativement aux propriétés des normes telles que nous le verrons en topologie, donne ce que nous appelons "**l'inégalité de Minkowski**".

En appliquant une fois l'inégalité triangulaire aux vecteurs $\vec{x}$ et $(\vec{y} - \vec{x})$ et une autre fois aux vecteurs $(\vec{x} - \vec{y})$ et $\vec{y}$ nous obtenons la variante:

$$\left| \|\vec{x}\| - \|\vec{y}\| \right| \leq \|\vec{x} - \vec{y}\| \quad (12.81)$$

5.1.3. PRODUIT SCALAIRE (GÉNÉRAL)

Voyons maintenant une autre manière un peu plus générale (s'appliquant à des vecteurs ou fonctions), formelle et abstraite pour définir le produit scalaire tout en tentant de rester le plus simple possible (attention dans le cas général la notation du produit scalaire change!):

Définition: Soit E un espace vectoriel réel (!). Une "**forme bilinéaire symétrique positive**" sur E , est une application:

$$\phi: \begin{cases} E \times E \rightarrow \mathbb{R} \\ (\vec{x}, \vec{y}) \mapsto \langle \vec{x} | \vec{y} \rangle \end{cases} \quad (12.82)$$

qui vérifie (par définition!):

P1. La positivité:

$$\forall \vec{x} \in E \quad \phi(\vec{x}, \vec{x}) = \langle \vec{x} | \vec{x} \rangle \geq 0 \in \mathbb{R}^+ \quad (12.83)$$

P2. La nullité (définie):

$$\forall \vec{x} \in E \quad \phi(\vec{x}, \vec{x}) = \langle \vec{x} | \vec{x} \rangle = 0 \Rightarrow \vec{x} = 0 \quad (12.84)$$

P3. La symétrie :

$$\forall (\vec{x}, \vec{y}) \in E^2 \quad \phi(\vec{x}, \vec{y}) = \phi(\vec{y}, \vec{x}) \Leftrightarrow \langle \vec{x} | \vec{y} \rangle = \langle \vec{y} | \vec{x} \rangle \quad (12.85)$$

P4. La bilinéarité (forme bilinéaire) avec, dans l'ordre, la "**linéarité à gauche**" et la "**linéarité à droite**":

$$\forall (\vec{x}, \vec{y}, \vec{z}) \in E^3, \forall (\lambda, \mu) \in \mathbb{R}^2 \quad \begin{cases} \langle \lambda \vec{x} + \mu \vec{y} | \vec{z} \rangle = \lambda \langle \vec{x} | \vec{z} \rangle + \mu \langle \vec{y} | \vec{z} \rangle \\ \langle \vec{x} | \lambda \vec{y} + \mu \vec{z} \rangle = \lambda \langle \vec{x} | \vec{y} \rangle + \mu \langle \vec{x} | \vec{z} \rangle \end{cases} \quad (12.86)$$

Remarque: A nouveau, ces propriétés sont principalement imposées par notre approche intuitive de l'espace euclidien et de son interprétation géométrique.

Définition: Un espace E muni d'un produit scalaire est appelé un "**espace préhilbertien**". Si E est de dimension finie, nous parlons alors "**d'espace euclidien**".

Nous avons vu en topologie (cf. chapitre de Topologie) que les propriétés du produit scalaire sont les briques de bases pour définir une norme et donc une distance dans un espace métrique. Cette distance sera alors donnée selon ce que nous avons obtenu en topologie:

$$\forall \vec{x}, \vec{y} \in E \quad d(\vec{x}, \vec{y}) := \|\vec{x} - \vec{y}\| \quad (12.87)$$

Définition: Nous disons qu'un espace E muni d'un produit scalaire $\langle \cdot | \cdot \rangle$ est un "**espace Hilbertien**" ou "**espace de Hilbert**" si cet espace est complet pour la métrique définie ci-dessus.

En d'autres termes, avoir un espace métrique muni donc d'une distance générée par un produit scalaire est une chose. Ensuite, avoir une distance mesurable en est une autre. Un espace de Hilbert a donc des distances mesurables au sens topologique du terme car l'ensemble sur lequel on travaille est continu et n'importe quel point peut-être approché indéfiniment (imaginez avoir une règle et que vous ne puissiez pas avec cette règle approcher les points qui définissent les dimensions de votre objet... ce serait gênant...). Donc sans espace complet une grande partie des théorèmes de l'analyse fonctionnelle ne pourraient pas être utilisés dans l'étude des espaces vectoriels et cela serait très gênant en physique quantique ondulatoire par exemple...

Formellement, rappelons qu'un espace métrique est complet si toutes les suites de Cauchy (cf. chapitre des Suites Et Séries) de cet espace sont convergentes (cf. chapitre sur les Fractals) dans un espace métrique (cf. chapitre de Topologie).

5.2. PRODUIT VECTORIEL

Le produit vectoriel de deux vecteurs est une opération propre à la dimension 3. Pour l'introduire, il faut préalablement orienter l'espace destiné à le recevoir. L'orientation étant définie au moyen de la notion de "**déterminant**", nous commencerons par une brève introduction à l'étude de cette notion. Cette étude sera reprise plus tard dans le détail lors de l'analyse des systèmes linéaires dans le chapitre d'algèbre linéaire.

Définition: Nous appelons "**déterminant**" des vecteurs-colonnes de $\mathbb{R}^2$ (pour la forme générale du déterminant se reporter au chapitre d'Algèbre Linéaire):

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \quad (12.88)$$

et nous notons:

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} \quad (12.89)$$

le nombre (produit soustrait en croix):

$$a_1 b_2 - a_2 b_1 \quad (12.90)$$

Nous appelons déterminant des vecteurs-colonnes de $\mathbb{R}^3$ (cf. chapitre d'Algèbre Linéaire):

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}, \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}, \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} \quad (12.91)$$

et nous notons:

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} \quad (12.92)$$

le nombre:

$$\begin{aligned} & a_1 \begin{vmatrix} b_2 & c_2 \\ b_3 & c_3 \end{vmatrix} - a_2 \begin{vmatrix} b_1 & c_1 \\ b_3 & c_3 \end{vmatrix} + a_3 \begin{vmatrix} b_1 & c_1 \\ b_2 & c_2 \end{vmatrix} \\ &= a_1 b_2 c_3 + a_2 b_3 c_1 + a_3 b_1 c_2 - a_1 b_3 c_2 - a_2 b_1 c_3 - a_3 b_2 c_1 \end{aligned} \quad (12.93)$$

Ainsi, la fonction qui associe à tout couple de vecteurs-colonnes de $\mathbb{R}^2$ (à tout triplet de vecteurs-colonnes de $\mathbb{R}^3$) son déterminant est appelée "**déterminant d'ordre 2**" (respectivement d'ordre 3).

Le déterminant a comme propriété d'être multiplié par -1 si l'un de ses vecteurs-colonnes est remplacé par son opposé ou si deux de ses vecteurs-colonnes sont échangés (la vérification étant simple nous nous abstenons de la démonstration, sauf sur demande). En plus, le déterminant est non nul si et seulement si ses vecteurs-colonnes sont linéairement indépendants (la démonstration se trouve quelques lignes plus bas et est d'une grande importance en mathématique).

Définition: Soit x_1, x_2, x_3 et y_1, y_2, y_3 les composantes respectives des vecteurs $\vec{x}$ et $\vec{y}$ dans la base orthonormale $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$. Nous appelons "**produit vectoriel**" de $\vec{x}$ et $\vec{y}$, et nous notons indistinctement:

$$\vec{x} \times \vec{y} = \vec{x} \wedge \vec{y} \quad (12.94)$$

le vecteur:

$$(x_2 y_3 - x_3 y_2) \vec{e}_1 + (x_3 y_1 - x_1 y_3) \vec{e}_2 + (x_1 y_2 - x_2 y_1) \vec{e}_3 \quad (12.95)$$

ou sous forme de composantes:

$$\vec{x} \times \vec{y} = \begin{pmatrix} x_2 y_3 - x_3 y_2 \\ x_3 y_1 - x_1 y_3 \\ x_1 y_2 - x_2 y_1 \end{pmatrix} \quad (12.96)$$

Remarques:

R1. La première notation est la notation internationale due à Gibbs (que nous utiliserons tout au long de ce site), la deuxième est la notation française due à Burali-Forti (assez embêtante car se confond avec l'opérateur ET en logique).

R2. Il est assez embêtant de retenir par coeur les relations qui forment le produit vectoriel habituellement. Mais heureusement il existe au moins trois bons moyens mnémotechniques:

1. Le plus rapide consiste à retrouver l'une des expressions des composantes du produit vectoriel et ensuite par décrémentation des indices (en recommençant à 3 lorsque qu'on arrive à 0) de connaître toutes les autres composantes. Encore faut-il trouver un moyen simple de se souvenir d'une des composantes. Un bon moyen est la propriété mathématique suivante de deux vecteurs colinéaires permettant facilement de retrouver la troisième composante (celle selon l'axe Z):

Soit deux vecteurs colinéaires dans un même plan, alors:

$$\vec{x} = \lambda \vec{y} \Leftrightarrow \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \lambda \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \Leftrightarrow \frac{x_1}{y_1} = \frac{x_2}{y_2} \Leftrightarrow \underline{x_1 y_2 - x_2 y_1} = 0 \quad (12.97)$$

Nous retrouvons donc bien l'expression de la troisième composante du produit vectoriel de deux vecteurs (non nécessairement colinéaires... eux!).

2. La seconde mais que nous verrons lors de notre étude du calcul tensoriel consiste à utiliser le symbole d'antisymétrie (également appelé "tenseur de Levi-Civita"). Cette méthode est certainement la plus esthétique d'entre toutes mais pas nécessairement la plus rapide à développer. Nous donnons ici juste l'expression sans plus d'explications pour l'instant (elle est également utile pour l'expression du déterminant par extension):

$$(\vec{x} \times \vec{y})_i = \varepsilon_{ijk} x_j y_k \quad (12.98)$$

3. Cette dernière méthode est assez simple et triviale aussi mais elle utilise implicitement la première méthode: la i -ème composante $\vec{x} \times \vec{y}$ est le déterminant des deux colonnes privées de leur i -ème terme, le deuxième déterminant étant cependant pris avec le signe "-" tel que:

$$\vec{x} \times \vec{y} = \begin{vmatrix} x_2 & y_2 \\ x_3 & y_3 \end{vmatrix} \vec{e}_1 - \begin{vmatrix} x_1 & y_1 \\ x_3 & y_3 \end{vmatrix} \vec{e}_2 + \begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix} \vec{e}_3 \quad (12.99)$$

Il est important, même si c'est relativement simple, de se rappeler que les différents produits vectoriels pour les vecteurs d'une base orthogonale sont:

$$\begin{aligned} \vec{e}_1 \times \vec{e}_1 &= \vec{e}_2 \times \vec{e}_2 = \vec{e}_3 \times \vec{e}_3 = \vec{0} \\ \vec{e}_1 \times \vec{e}_2 &= \vec{e}_3, \vec{e}_2 \times \vec{e}_3 = \vec{e}_1, \vec{e}_3 \times \vec{e}_1 = \vec{e}_2 \end{aligned} \quad (12.100)$$

Le produit vectoriel jouit aussi propriétés suivantes que nous allons démontrer:

P1. Antisymétrie:

$$\vec{x} \times \vec{y} = -(\vec{y} \times \vec{x}) \quad (12.101)$$

P2. Linéarité:

$$(\alpha \vec{x} + \beta \vec{y}) \times \vec{z} = \alpha(\vec{x} \times \vec{z}) + \beta(\vec{y} \times \vec{z}) \quad (12.102)$$

P3. Si et seulement si $\vec{x}$ et $\vec{y}$ sont linéairement indépendants (très important !):

$$\vec{x} \times \vec{y} \neq \vec{0} \quad (12.103)$$

P4. Non associativité:

$$\vec{x} \times (\vec{y} \times \vec{z}) \neq (\vec{x} \times \vec{y}) \times \vec{z} \quad (12.104)$$

Les deux premières propriétés découlent directement de la définition et la propriété P4 se vérifie aisément en développant les composantes et en comparant les résultats obtenus.

Démontrons alors la troisième propriété qui est très importante en algèbre linéaire.

Démonstration:

Soient deux vecteurs $\vec{x}(x_1, x_2, x_3)$ et $\vec{y}(y_1, y_2, y_3)$. Si les deux vecteurs sont linéairement dépendants alors il existe α tel que nous puissions écrire:

$$\vec{y} = \alpha \cdot \vec{x} \quad (12.105)$$

Si nous développons le produit vectoriel des deux vecteurs dépendants à un facteur α près, nous obtenons:

$$\vec{x} \times \vec{y} = \vec{x} \times \alpha \vec{x} = (x_2 \alpha x_3 - x_3 \alpha x_2) \vec{e}_1 + (x_3 \alpha x_1 - x_1 \alpha x_3) \vec{e}_2 + (x_1 \alpha x_2 - x_2 \alpha x_1) \vec{e}_3 \quad (12.106)$$

Il va sans dire que le résultat ci-dessus est égal au vecteur nul $\vec{0}$ si effectivement les deux vecteurs sont linéairement dépendants.

□C.Q.F.D.

Si nous supposons maintenant que les deux vecteurs $\vec{x}$ et $\vec{y}$ sont linéairement indépendants et non nuls, nous devons démontrer que le produit vectoriel est:

P3.1. Orthogonal (perpendiculaire) à $\vec{x}$ et $\vec{y}$

P3.2. De norme $\|\vec{x}\| \|\vec{y}\| \sin \theta$, où θ est l'angle entre $\vec{x}$ et $\vec{y}$

Démonstration:

Commençons par la première propriété P3.1 (première importance en physique!):

$$\vec{x} \circ (\vec{x} \times \vec{y}) = x_1 \begin{vmatrix} x_2 & y_2 \\ x_3 & y_3 \end{vmatrix} - x_2 \begin{vmatrix} x_1 & y_1 \\ x_3 & y_3 \end{vmatrix} + x_3 \begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix} = \begin{vmatrix} x_1 & x_1 & y_1 \\ x_2 & x_2 & y_2 \\ x_3 & x_3 & y_3 \end{vmatrix} = 0 \quad (12.107)$$

ce qui montre bien que le vecteur $\vec{x}$ est perpendiculaire au vecteur résultant du produit vectoriel entre $\vec{x}$ et $\vec{y}$!

□C.Q.F.D.

Terminons avec la deuxième propriété P3.2 (aussi de première importance en physique!):

Démonstration:

Soit le carré de la norme du produit vectoriel $\|\vec{x} \times \vec{y}\|^2$. D'après la définition du produit vectoriel nous avons:

$$\begin{aligned} \|\vec{x} \times \vec{y}\|^2 &= (x_2 y_3 - x_3 y_2)^2 + (x_3 y_1 - x_1 y_3)^2 + (x_1 y_2 - x_2 y_1)^2 \\ &= (x_1^2 + x_2^2 + x_3^2)(y_1^2 + y_2^2 + y_3^2) - (x_1 y_1 + x_2 y_2 + x_3 y_3)^2 \\ &= \|\vec{x}\|^2 \|\vec{y}\|^2 - \|\vec{x}\|^2 \|\vec{y}\|^2 \cos^2 \theta = \|\vec{x}\|^2 \|\vec{y}\|^2 (1 - \cos^2 \theta) = \|\vec{x}\|^2 \|\vec{y}\|^2 \sin^2 \theta \end{aligned} \quad (12.108)$$

Donc finalement:

$$\|\vec{x} \times \vec{y}\| = \|\vec{x}\| \|\vec{y}\| \sin \theta \quad (12.109)$$

□C.Q.F.D.

Nous remarquerons que dans le cas où E est l'espace vectoriel euclidien, la norme du produit vectoriel représente l'aire (surface) du parallélogramme construit sur des représentants $\vec{x}$ et $\vec{y}$ d'origine commune.

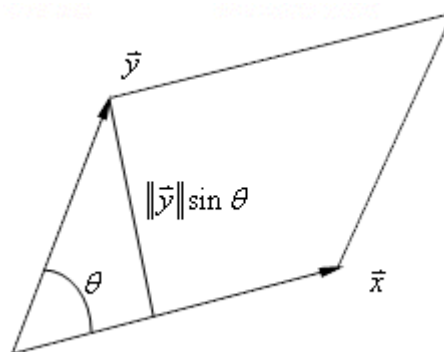


Figure: 12.6 - Représentation géométrique du produit vectoriel

Si $\vec{x}$ et $\vec{y}$ sont linéairement indépendants, le triplet $(\vec{x} \times \vec{y}, \vec{x}, \vec{y})$ et donc aussi le triplet $(\vec{x}, \vec{y}, \vec{x} \times \vec{y})$ sont directs.

En effet, (z_1, z_2, z_3) étant les composantes de $\vec{x} \times \vec{y}$ (dans la base $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$), le déterminant de passage de $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ à $(\vec{x} \times \vec{y}, \vec{x}, \vec{y})$ (par exemple) s'écrit:

$$\begin{vmatrix} z_1 & x_1 & y_1 \\ z_2 & x_2 & y_2 \\ z_3 & x_3 & y_3 \end{vmatrix} = z_1^2 + z_2^2 + z_3^2 \quad (12.110)$$

Ce déterminant est donc positif, puisqu'au moins un des z_i n'est pas nul, d'après la troisième propriété d'indépendance linéaire du produit vectoriel.

Voici encore quelques propriétés très importantes d'utilité pratique du produit vectoriel (en physique particulièrement) qui sont triviales à vérifier si les développements sont effectués (nous pouvons les faire sur demande si jamais!):

$$P1. \vec{x} \times (\vec{y} \times \vec{z}) = (\vec{x} \circ \vec{z})\vec{y} - (\vec{x} \circ \vec{y})\vec{z}$$

Remarque: Cette dernière relation est parfois appelée la "**règle de Grassmann**", ou plus couramment "**double produit vectoriel**" et il est important de noter que sans les parenthèses le résultat n'est pas unique.

$$P2. (\vec{x} \times \vec{y}) \circ (\vec{z} \times \vec{v}) = (\vec{x} \circ \vec{z})(\vec{y} \circ \vec{v}) - (\vec{x} \circ \vec{v})(\vec{y} \circ \vec{z})$$

$$P3. (\vec{x} \times \vec{y}) \circ \vec{z} = -(\vec{x} \times \vec{z}) \circ \vec{y}$$

$$P4. (\vec{x} \times \vec{y}) \circ \vec{z} = \vec{x} \circ (\vec{y} \times \vec{z})$$

$$P5. \|\vec{x} \times \vec{y}\|^2 = (\vec{x} \circ \vec{x})(\vec{y} \circ \vec{y}) - (\vec{x} \circ \vec{y})^2$$

5.3. PRODUIT MIXTE

Nous pouvons étendre la définition du produit vectoriel à un autre type d'outil mathématique que nous appelons le "**produit mixte**":

Définition: Nous appelons "**produit mixte**" des vecteurs $\vec{x}, \vec{y}, \vec{z}$ le double produit:

$$\vec{x} \circ (\vec{y} \times \vec{z}) \quad (12.111)$$

souvent condensé sous la notation suivante:

$$[\vec{x}, \vec{y}, \vec{z}] \quad (12.112)$$

D'après ce que nous avons vu lors de la définition du produit scalaire et vectoriel, le produit mixte peut également s'écrire:

$$\|\vec{x}\| \circ \|\vec{y} \times \vec{z}\| \cos \theta \quad (12.113)$$

Nous remarquerons que dans le cas où E est l'espace vectoriel euclidien $\mathbb{R}^3$, la valeur absolue du produit mixte symbolise le volume (orienté) du parallélépipède, construit sur des représentants $\vec{x}, \vec{y}, \vec{z}$ d'origine commune.

Remarque: Il est assez trivial que le produit mixte est une extension à 3 dimensions du produit vectoriel. Effectivement, dans l'expression du produit mixte, le produit vectoriel représente la surface de base du parallélépipède et le produit scalaire projette un des vecteurs sur le vecteur résultant du produit vectoriel ce qui donne la hauteur h du parallélépipède.

De par les propriétés de commutativité du produit scalaire, nous avons:

$$(\vec{x} \times \vec{y}) \circ \vec{z} = \vec{z} \circ (\vec{x} \times \vec{y}) \quad (12.114)$$

et le lecteur vérifiera sans aucune peine (nous le ferons s'il y a demande) en développant les composantes que:

$$(\vec{x} \times \vec{y}) \circ \vec{z} = \vec{x} \circ (\vec{y} \times \vec{z}) \quad (12.115)$$

Le produit mixte jouit également des propriétés que le lecteur ne devrait avoir aucun mal à vérifier en développant les composantes mis à part peut-être P3 qui découle des propriétés du produit scalaire et vectoriel (nous pouvons développer sur demande si jamais!):

$$P1. [\vec{x}, \vec{y}, \vec{z}] = [\vec{z}, \vec{x}, \vec{y}] = [\vec{y}, \vec{z}, \vec{x}] = -[\vec{y}, \vec{x}, \vec{z}] = -[\vec{z}, \vec{y}, \vec{x}] = -[\vec{x}, \vec{z}, \vec{y}]$$

$$P2. [\alpha\vec{x} + \beta\vec{y}, \vec{z}, \vec{v}] = \alpha[\vec{x}, \vec{z}, \vec{v}] + \beta[\vec{y}, \vec{z}, \vec{v}]$$

$$P3. [\vec{x}, \vec{y}, \vec{z}] \neq 0 \text{ si et seulement si } \vec{x}, \vec{y}, \vec{z} \text{ sont linéairement indépendants}$$

$$P4. (\vec{x} \times \vec{y}) \times (\vec{z} \times \vec{v}) = [\vec{x}, \vec{y}, \vec{v}] \cdot \vec{z} - [\vec{x}, \vec{y}, \vec{z}] \cdot \vec{v}$$

Remarque: Nous reviendrons sur le produit mixte lors de notre étude du calcul tensoriel car il permet d'arriver à un résultat très intéressant en particulier en ce qui concerne la relativité générale!

6. ESPACES VECTORIELS FONCTIONNELS

Soit $C_{[a,b]}^k$ l'ensemble des fonctions réelles k -fois dérivables dans l'intervalle fermé $[a, b]$. Nous désignerons les éléments de cet ensemble par les lettres $\vec{f}, \vec{g}, \dots$

La valeur de $\vec{f}$ au point t sera bien évidemment notée $\vec{f}(t)$. Dire que $\vec{f} = \vec{g}$ équivaudra donc à dire que:

$$\vec{f}(t) = \vec{g}(t), \forall t \in [a, b] \quad (12.116)$$

De manière abrégée, nous écrirons $\vec{f}(t) \equiv \vec{g}(t)$, le signe $\equiv$ indiquant ainsi que les deux membres sont égaux pour tout t de l'intervalle $[a, b]$.

Considérons les deux opérations suivantes:

$$- \vec{f}(t) + \vec{g}(t) \text{ définie par la formule } (\vec{f} + \vec{g})(t) \equiv \vec{f}(t) + \vec{g}(t)$$

$$- \alpha\vec{f} \text{ définie par la formule } (\alpha\vec{f})(t) \equiv \alpha\vec{f}(t)$$

Ces deux opérations satisfont à toutes les conditions des vecteurs d'un espace vectoriel que nous avons déjà définies au début de ce chapitre (associativité, commutativité, vecteur nul, vecteur opposé, distributivité, élément neutre) et munissent donc $C_{[a,b]}^k$ d'une structure d'espace vectoriel. Le vecteur nul de cet espace étant bien évidemment la fonction nulle et l'opposé de $\vec{f}$ étant la fonction $-\vec{f}$.

Il est intéressant de constater que $C_{[a,b]}^k$ en tant qu'espace vectoriel est une généralisation de $\mathbb{R}^n$ au cas continu. Nous pouvons en effet concevoir tout vecteur (α_i) de $\mathbb{R}^n$ sous la forme d'une fonction réelle définie dans l'ensemble $\{1, 2, \dots, n\}$: la valeur de cette fonction au point i est tout simplement α_i .

Remarque: Les polynômes de degré n et à une inconnue forment aussi un exemple d'espace vectoriel fonctionnel de dimension $n+1$ (à chaque coefficient du polynôme correspond une composante du vecteur).

Le champ d'application privilégié de la théorie abstraite du produit scalaire est constitué par les espaces vectoriels fonctionnels. Nous appelons ainsi aussi produit scalaire canonique dans $C_{[a,b]}$ l'opération définie par la relation:

$$\vec{f} \circ \vec{g} = \int_a^b \vec{f}(t) \vec{g}(t) dt \quad (12.117)$$

Cette opération définit bien un produit scalaire, les propriétés de ce dernier sont vérifiées et, en outre, l'intégrale:

$$\vec{f} \circ \vec{f} = \|\vec{f}\|^2 = \int_a^b \vec{f}^2(t) dt \quad (12.118)$$

est positive si la fonction continue $\vec{f}$ n'est pas identiquement nulle.

7. ESPACES VECTORIELS HERMITIENS

L'objectif de ce qui va suivre n'est pas de faire une étude détaillée du sujet des espaces vectoriels complexes mais juste de donner le bagage et le vocabulaire minimum nécessaire à l'étude de certaines théories physiques comme la physique quantique par exemple.

Lorsque les scalaires qui apparaissent dans la définition de la notion d'espace vectoriel sont des nombres complexes (cf. chapitre sur les Nombres), et non plus des nombres réels, nous parlons alors "d'espaces vectoriels complexes".

Remarque: Rigoureusement dans la communication courante, il devrait systématiquement être fait mention si nous parlons d'espace vectoriel réel ou d'espace vectoriel complexe...

Citons quelques exemples d'espaces vectoriels complexes :

E1. L'espace vectoriel $\mathbb{C}^n$ des vecteurs-colonnes à n termes complexes ($\mathbb{C}^1$ étant identifié à $\mathbb{C}$). Nous rencontrerons, entre autres, de tels vecteurs dans le chapitre de Physique Quantique Relativiste.

E2. L'espace vectoriel des polynômes à coefficients complexes en une indéterminée. Nous rencontrerons ce genre d'espaces dans les chapitres de Physique Quantique Ondulatoire ou encore de Chimie Quantique.

E3. L'espace vectoriel des fonctions complexes d'une variable réelle ou complexe dérivables ou non. Nous rencontrerons ce genre d'espace très fréquemment dans la section de Mécanique globalement et surtout dans les chapitres d'Électrodynamique ou encore de Mécanique Ondulatoire.

Il s'agit ici d'adapter ce que nous avons fait précédemment aux espaces vectoriels complexes. L'exemple suivant nous montre que nous ne pouvons pas transposer telles quelles les définitions précédentes. En effet, considérons l'espace vectoriel $\mathbb{C}^n$. Comme pour $\mathbb{R}^n$, nous pourrions être tenté de définir un produit scalaire sur $\mathbb{C}^n$ par:

$$\forall \vec{x}, \vec{y} \in \mathbb{C}^n, \vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \Rightarrow \langle \vec{x} | \vec{y} \rangle = x_1 y_1 + \dots + x_n y_n \quad (12.119)$$

avec $x_i, y_i \in \mathbb{C}$.

Malheureusement, nous nous apercevons que cette définition n'est pas satisfaisante car nous aurions alors:

$$\langle \vec{x} | \vec{x} \rangle = x_1^2 + \dots + x_n^2 \quad (12.120)$$

et cette quantité n'est pas en général un nombre réel dans l'espace des complexes ce qui viole la propriété de positivité du produit scalaire et donc empêche d'introduire tout concept de distance.

Nous ne pourrions donc plus définir une norme en posant $\|\vec{x}\| = \sqrt{\langle \vec{x} | \vec{x} \rangle}$. Pour que $\langle \vec{x} | \vec{x} \rangle$ soit un nombre réel positif nous voyons qu'il faudrait plutôt définir le produit scalaire comme ceci :

$$\forall \vec{x}, \vec{y} \in \mathbb{C}^n, \langle \vec{x} | \vec{y} \rangle = x_1 \overline{y_1} + \dots + x_n \overline{y_n} \quad (12.121)$$

Dans ce cas nous avons:

$$\langle \vec{x} | \vec{x} \rangle = \overline{x_1} \cdot x_1 + \dots + \overline{x_n} \cdot x_n = |x_1|^2 + \dots + |x_n|^2 \quad (12.122)$$

qui est bien un nombre réel positif. A partir de là, nous pouvons à nouveau définir une norme sur l'espace vectoriel complexe $\mathbb{C}^n$ en posant:

$$\|\vec{x}\| = \sqrt{\langle \vec{x} | \vec{x} \rangle} = \sqrt{|x_1|^2 + \dots + |x_n|^2} \quad (12.123)$$

Nous allons à présent montrer comment définir un produit scalaire sur un espace vectoriel complexe dans le cas général.

7.1. PRODUIT HERMITIEN

Définition: Soit H un espace vectoriel complexe (!). Nous appelons "produit scalaire" ou plus exactement "**produit hermitien**" sur H , une application:

$$\phi: \begin{cases} H \times H \rightarrow \mathbb{C} \\ (\vec{x}, \vec{y}) \mapsto \langle \vec{x} | \vec{y} \rangle \end{cases} \quad (12.124)$$

qui vérifie :

P1. La positivité:

$$\forall \vec{x} \in H \quad \phi(\vec{x}, \vec{x}) = \langle \vec{x} | \vec{x} \rangle \geq 0 \in \mathbb{R}^+ \quad (12.125)$$

P2. Nullité (définie):

$$\forall \vec{x} \in H \quad \phi(\vec{x}, \vec{x}) = \langle \vec{x} | \vec{x} \rangle = 0 \Leftrightarrow \vec{x} = \vec{0} \quad (12.126)$$

P3. Symétrie hermitienne :

$$\forall (\vec{x}, \vec{y}) \in H^2 \quad \phi(\vec{x}, \vec{y}) = \overline{\phi(\vec{y}, \vec{x})} \Leftrightarrow \langle \vec{x} | \vec{y} \rangle = \overline{\langle \vec{y} | \vec{x} \rangle} \quad (12.127)$$

P4. La bilinéarité (forme bilinéaire) change un peu aussi... ce qui fait que nous parlons alors de "**sesquilinearité**". Nous parlons alors, dans l'ordre, d'anti-linéarité à gauche et de linéarité à droite tel que :

$$\forall (\vec{x}, \vec{y}, \vec{z}) \in H^3, \forall (\lambda, \mu) \in \mathbb{C}^2 \quad \begin{cases} \langle \lambda \vec{x} + \mu \vec{y} | \vec{z} \rangle = \bar{\lambda} \langle \vec{x} | \vec{z} \rangle + \bar{\mu} \langle \vec{y} | \vec{z} \rangle \\ \langle \vec{x} | \lambda \vec{y} + \mu \vec{z} \rangle = \lambda \langle \vec{x} | \vec{y} \rangle + \mu \langle \vec{x} | \vec{z} \rangle \end{cases} \quad (12.128)$$

Remarques:

R1. Certains mathématiciens mettent l'anti-linéarité à droite. C'est simplement une question de convention qui n'a aucune importance.

R2. Le lecteur remarquera peut-être sans peine que si les éléments des définitions précédentes sont tous dans $\mathbb{R}$ alors la sesquilinearité se réduit à la bilinéarité et le caractère hermitien à la symétrie. Donc le produit hermitien se réduit au produit scalaire.

R3. Nous souhaitons donner pour l'instant uniquement le minimum sur le vaste sujet que sont les espaces vectoriels complexes afin que le lecteur puisse lire sans trop de peine le début du chapitre de Physique Quantique Ondulatoire.

Lorsque nous munissons un espace vectoriel complexe d'un produit scalaire alors au même titre qu'un espace vectoriel réel devient un espace vectoriel euclidien ou préhilbertien, l'espace vectoriel complexe devient donc ce que nous appelons un "**espace vectoriel hermitien**" (terme assez souvent utilisé dans le chapitre de Physique Quantique Ondulatoire).

Définition: Encore une fois, nous disons qu'un espace H muni d'un produit hermitien $\langle \cdot | \cdot \rangle$ est un "espace de Hilbert" si cet espace est complet pour la métrique définie ci-dessus.

Ainsi, les espaces de Hilbert sont une généralisation comprenant les produits scalaires et produits hermitiens des espaces euclidiens, préhilbertiens et hermitiens.

7.2. TYPES D'ESPACES VECTORIELS

Pour résumer tout cela:

- Nous appelons espace préhilbertien (réel ou complexe) tout espace vectoriel, de dimension finie ou non, muni d'un produit scalaire.
- Nous appelons espace de Hilbert (réel ou complexe) tout espace préhilbertien complet (en tant qu'espace normé).
- Nous appelons espace euclidien tout espace vectoriel réel de dimension finie muni d'un produit scalaire.
- Nous appelons espace hermitien tout espace vectoriel complexe de dimension finie muni d'un produit scalaire.

Nous savons que tout espace vectoriel (réel ou complexe) normé de dimension finie est complet. Ainsi, les espaces euclidiens et les espaces hermitiens sont des espaces de Hilbert (respectivement réels ou complexes).

8. SYSTÈMES DE COORDONNÉES

Nous allons aborder ici, l'aspect des changements de coordonnées des composantes de vecteurs non pas d'une base à une autre (pour cela il faut aller voir le chapitre d'Algèbre Linéaire) mais d'un système de coordonnées à un autre. Ce type de transformation a un fort potentiel en physique (ainsi qu'en mathématique) lorsque nous souhaitons simplifier l'étude de systèmes physiques dont les équations deviennent plus facilement manipulables dans d'autres systèmes de coordonnées.

Définition: En mathématiques, un "système de coordonnées" permet de faire correspondre à chaque point d'un espace à n dimensions, un n -uplet de scalaires.

Remarque: Dans beaucoup de cas, les scalaires considérés sont des nombres réels, mais il est possible d'utiliser des nombres complexes ou des éléments d'un quelconque corps (cf. chapitre de Théorie Des Ensembles). Plus généralement, les coordonnées peuvent provenir d'un anneau ou d'une autre structure algébrique apparentée.

Bien que nous soyons dans un chapitre et une section de mathématiques du site, nous nous permettrons dans ce qui va suivre de faire une liaison directe avec la physique pour ce qui concerne les expressions de la vitesse et de l'accélération dans différents systèmes de coordonnées (désolé pour les matheux...)

8.1. SYSTÈME DE COORDONNÉES CARTÉSIENNES

Nous ne souhaitons pas trop nous attarder sur ce système car il est bien connu de tout le monde habituellement. Rappelons cependant que la plupart du temps, en physique, les systèmes cartésiens dans lesquels nous travaillons sont $\mathbb{R}^2$ (deux dimensions spatiales), $\mathbb{R}^3$ (trois dimensions spatiales) voir $\mathbb{R}^4$ ou $\mathbb{C}^4$ (trois dimensions spatiales et une temporelle) lorsque nous travaillons en relativité.

Les systèmes cartésiens sont représentés par des vecteurs de base orthonormés (dans le sens qu'ils sont linéairement indépendants et de norme unité) qui forment une "base euclidienne" d'un espace vectoriel euclidien...

Dans $\mathbb{R}^3$ (cas le plus fréquent), il y a trois vecteurs de base notés traditionnellement:

$$\vec{e}_1, \vec{e}_2, \vec{e}_3 \quad (12.129)$$

Dans ce système, la position d'un point P (repérable par un vecteur $\vec{r}$) est définie par le triplet de nombres noté (en calcul tensoriel):

$$(x_1, x_2, x_3) \quad (12.130)$$

et en physique plus conventionnellement:

$$(x, y, z) \quad (12.131)$$

où habituellement la coordonnée (z) représente la hauteur (la verticale), la coordonnée (x) la largeur et la coordonnée (y) la longueur (évidemment ces choix sont complètement arbitraires).

Ce point P peut être repéré par un vecteur noté arbitrairement $\vec{x}$ dans la base $(\vec{e}_i)$ par la relation (utilisant la notation tensorielle):

$$\vec{x} = x^i \vec{e}_i \quad (12.132)$$

En physique, si nous travaillons avec des coordonnées, c'est toujours pour pouvoir déterminer l'emplacement d'un élément. Or, comme nous le verrons plus rigoureusement en mécanique analytique, le physicien travaille avec les notions suivantes (chaque élément dépendant souvent du temps):

- Position: $\vec{r} = (x(t), y(t), z(t), t)$

- Vitesse: $\frac{d\vec{r}}{dt} = \dot{\vec{r}} = \vec{v} = (\dot{x}(t), \dot{y}(t), \dot{z}(t), t)$

- Accélération: $\frac{d\vec{v}}{dt} = \dot{\vec{v}} = \vec{a} = (\ddot{x}(t), \ddot{y}(t), \ddot{z}(t), t)$

Maintenant voyons comment s'expriment ces différentes notions dans des systèmes tels que les coordonnées sphériques, cylindriques et polaires.

8.2. SYSTÈME DE COORDONNÉES SPHÉRIQUES

Le choix de commencer avec ce système de coordonnées n'est pas un hasard. Il a pour avantage d'être une généralisation des systèmes cylindrique et polaire dont nous retrouverons par la suite plus facilement les expressions de la position, de la vitesse et de l'accélération.

Nous représentons traditionnellement un système à coordonnées sphériques de la façon suivante:

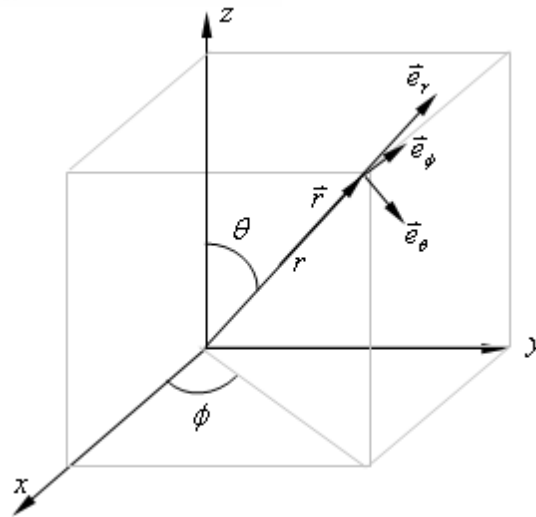


Figure: 12.7 - Représentation du système de coordonnées sphériques

Nous voyons très clairement si nous connaissons bien les relations trigonométriques élémentaires (voir le chapitre du même nom dans la section de géométrie) que nous avons les transformations:

$$\begin{aligned}
 r &= \sqrt{x^2 + y^2 + z^2}, 0 \leq r < \infty \\
 \theta &= \arccos\left(\frac{z}{r}\right) = \arccos\left(\frac{z}{\sqrt{x^2 + y^2 + z^2}}\right), 0 \leq \theta \leq \pi \\
 \phi &= \arccos\left(\frac{x}{\sqrt{x^2 + y^2}}\right) = \arccos\left(\frac{x}{r \sin \theta}\right) = \arcsin\left(\frac{y}{r \sin \theta}\right), 0 \leq \phi < 2\pi
 \end{aligned}
 \tag{12.133}$$

et inversement:

$$x = r \sin \theta \cos \phi \quad y = r \sin \theta \sin \phi \quad z = r \cos \theta \tag{12.134}$$

Maintenant il nous faut trouver les expressions qui relient les vecteurs de la base sphérique que nous choisissons de noter $\vec{e}_r, \vec{e}_\theta, \vec{e}_\phi$ avec les vecteurs de la base cartésienne $\vec{e}_x, \vec{e}_y, \vec{e}_z$.

Nous avons, comme nous pouvons le voir sur le schéma ci-dessus:

$$\begin{aligned}
\vec{e}_r &= \frac{\vec{r}}{r} = \frac{x\vec{e}_x + y\vec{e}_y + z\vec{e}_z}{r} = \frac{r \sin \theta \cos \phi \vec{e}_x + r \sin \theta \sin \phi \vec{e}_y + r \cos \theta \vec{e}_z}{r} \\
&= \sin \theta \cos \phi \vec{e}_x + \sin \theta \sin \phi \vec{e}_y + \cos \theta \vec{e}_z \\
\vec{e}_\phi &= \frac{\vec{e}_z \times \vec{e}_r}{\sin \theta} = \frac{1}{\sin \theta} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \times \begin{pmatrix} \sin \theta \cos \phi \\ \sin \theta \sin \phi \\ \cos \theta \end{pmatrix} = \frac{1}{\sin \theta} \begin{pmatrix} -\sin \theta \sin \phi \\ \sin \theta \cos \phi \\ 0 \end{pmatrix} \\
&= \frac{1}{\sin \theta} \begin{pmatrix} -\sin \theta \sin \phi \\ 0 \\ 0 \end{pmatrix} + \frac{1}{\sin \theta} \begin{pmatrix} 0 \\ \sin \theta \cos \phi \\ 0 \end{pmatrix} = -\vec{e}_x \sin \phi + \vec{e}_y \cos \phi \\
\vec{e}_\theta &= \vec{e}_\phi \times \vec{e}_r = \begin{pmatrix} -\sin \phi \\ \cos \phi \\ 0 \end{pmatrix} \times \begin{pmatrix} \sin \theta \cos \phi \\ \sin \theta \sin \phi \\ \cos \theta \end{pmatrix} = \begin{pmatrix} \cos \phi \cos \theta - 0 \cdot \sin \theta \sin \phi \\ 0 \cdot \sin \theta \cos \phi - (-\sin \phi) \cos \theta \\ -\sin \phi \cdot \sin \theta \sin \phi - \cos \phi \sin \theta \cos \phi \end{pmatrix} \\
&= \begin{pmatrix} \cos \phi \cos \theta \\ \sin \phi \cos \theta \\ -\sin \phi \cdot \sin \theta \sin \phi - \cos \phi \sin \theta \cos \phi \end{pmatrix} = \vec{e}_x \cos \theta \cos \phi + \vec{e}_y \cos \theta \sin \phi - \vec{e}_z \sin \theta \\
\end{aligned}$$

(12.135)

Indiquons qu'en divisant par $\sin \theta$ pour le deuxième vecteur de base, nous nous assurons ainsi de par les propriétés de la norme du produit vectoriel que:

$$\|\vec{e}_\phi\| = \|\vec{e}_z\| \|\vec{e}_r\| \sin \phi = 1 \quad (12.136)$$

sera bien normalisé à l'unité!

Remarque: Il est important de remarquer que le produit vectoriel de deux vecteurs de base donne toujours le troisième vecteur de base perpendiculaire (comme pour les coordonnées cartésiennes donc!).

Pour des besoins ultérieurs, déterminons les différentielles partielles de chacune de ces coordonnées:

$$\begin{aligned}
\frac{\partial \vec{e}_r}{\partial r} &= 0 \\
\frac{\partial \vec{e}_r}{\partial \theta} &= \vec{e}_x \cos \theta \cos \phi + \vec{e}_y \cos \theta \sin \phi - \vec{e}_z \sin \theta = \vec{e}_\theta \\
\frac{\partial \vec{e}_r}{\partial \phi} &= -\vec{e}_x \sin \theta \sin \phi + \vec{e}_y \sin \theta \cos \phi = (-\vec{e}_x \sin \phi + \vec{e}_y \cos \phi) \sin \theta = \vec{e}_\phi \sin \theta \\
\frac{\partial \vec{e}_\theta}{\partial r} &= 0; \frac{\partial \vec{e}_\theta}{\partial \theta} = 0; \frac{\partial \vec{e}_\theta}{\partial \phi} = -\vec{e}_x \cos \phi - \vec{e}_y \sin \phi = -(\vec{e}_r \sin \theta + \vec{e}_\phi \cos \theta) \\
\frac{\partial \vec{e}_\phi}{\partial r} &= 0 \\
\frac{\partial \vec{e}_\phi}{\partial \theta} &= -\vec{e}_x \sin \theta \cos \phi - \vec{e}_y \sin \theta \sin \phi - \vec{e}_z \cos \theta = -\vec{e}_r \\
\frac{\partial \vec{e}_\phi}{\partial \phi} &= -\vec{e}_x \cos \theta \sin \phi + \vec{e}_y \cos \theta \cos \phi = \vec{e}_\theta \cos \theta
\end{aligned} \tag{12.137}$$

Nous allons également utiliser plus tard (pour l'étude des opérateurs vectoriels) la variation $d\vec{r}$ exprimée en coordonnées sphériques:

$$\begin{aligned}
d\vec{r} &= d(r\vec{e}_r) = \vec{e}_r dr + r d\vec{e}_r = \vec{e}_r dr + r \left(\frac{\partial \vec{e}_r}{\partial r} dr + \frac{\partial \vec{e}_r}{\partial \theta} d\theta + \frac{\partial \vec{e}_r}{\partial \phi} d\phi \right) \\
&= \vec{e}_r dr + \vec{e}_\theta r d\theta + \vec{e}_\phi r \sin \theta d\phi
\end{aligned} \tag{12.138}$$

Pour exprimer la vitesse et l'accélération en coordonnées sphériques, nous aurons également besoin des dérivées par rapport au temps:

$$\begin{aligned}
\dot{\vec{e}}_r &= \frac{\partial \vec{e}_r}{\partial r} \dot{r} + \frac{\partial \vec{e}_r}{\partial \theta} \dot{\theta} + \frac{\partial \vec{e}_r}{\partial \phi} \dot{\phi} = \vec{e}_\theta \dot{\theta} + \vec{e}_\phi \dot{\phi} \sin \theta \\
\dot{\vec{e}}_\theta &= \frac{\partial \vec{e}_\theta}{\partial r} \dot{r} + \frac{\partial \vec{e}_\theta}{\partial \theta} \dot{\theta} + \frac{\partial \vec{e}_\theta}{\partial \phi} \dot{\phi} = -\vec{e}_r \dot{\theta} + \vec{e}_\phi \dot{\phi} \cos \theta \\
\dot{\vec{e}}_\phi &= \frac{\partial \vec{e}_\phi}{\partial r} \dot{r} + \frac{\partial \vec{e}_\phi}{\partial \theta} \dot{\theta} + \frac{\partial \vec{e}_\phi}{\partial \phi} \dot{\phi} = -(\vec{e}_r \sin \theta + \vec{e}_\theta \cos \theta) \dot{\phi}
\end{aligned} \tag{12.139}$$

Donc si nous faisons maintenant un peu de physique, nous avons:

$$\begin{aligned}
\vec{v} = \dot{\vec{r}} &= \begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{pmatrix} = \begin{pmatrix} d_t(r \sin \theta \cos \phi) \\ d_t(r \sin \theta \sin \phi) \\ d_t(r \cos \theta) \end{pmatrix} = \begin{pmatrix} \dot{r} \sin \theta \cos \phi + r \cos \theta \cos \phi \dot{\theta} - r \sin \theta \sin \phi \dot{\phi} \\ \dot{r} \sin \theta \sin \phi + r \cos \theta \sin \phi \dot{\theta} + r \sin \theta \cos \phi \dot{\phi} \\ \dot{r} \cos \theta - r \sin \theta \dot{\theta} \end{pmatrix} \\
&= \begin{pmatrix} \sin \theta \cos \phi \\ \sin \theta \sin \phi \\ \cos \theta \end{pmatrix} \dot{r} + \begin{pmatrix} r \cos \theta \cos \phi \\ r \cos \theta \sin \phi \\ -r \sin \theta \end{pmatrix} \dot{\theta} + \begin{pmatrix} -r \sin \theta \sin \phi \\ r \sin \theta \cos \phi \\ 0 \end{pmatrix} \dot{\phi} \\
&= \vec{e}_r \dot{r} + r \begin{pmatrix} \cos \theta \cos \phi \\ \cos \theta \sin \phi \\ -\sin \theta \end{pmatrix} \dot{\theta} + r \sin \theta \begin{pmatrix} -\sin \phi \\ \cos \phi \\ 0 \end{pmatrix} \dot{\phi} = \vec{e}_r \dot{r} + r \vec{e}_\theta \dot{\theta} + r \sin \theta \vec{e}_\phi \dot{\phi}
\end{aligned} \tag{12.140}$$

ce qui nous amène donc à (nous aurons besoin de cette relation en astrophysique):

$$\vec{v} = \dot{r}\vec{e}_r + \vec{e}_\theta r\dot{\theta} + \vec{e}_\phi r\dot{\phi}\sin\theta \quad (12.141)$$

Il est intéressant de remarquer que nous arrivons au même résultat en passant par la méthode suivante qui est peut-être moins intuitive:

$$\vec{v} = \dot{\vec{r}} = \dot{\vec{e}}_r r + \vec{e}_r \dot{r} \quad (12.142)$$

et en y substituant la dérivée obtenue plus haut:

$$\dot{\vec{e}}_r = \frac{\partial \vec{e}_r}{\partial r} \dot{r} + \frac{\partial \vec{e}_r}{\partial \theta} \dot{\theta} + \frac{\partial \vec{e}_r}{\partial \phi} \dot{\phi} = \vec{e}_\theta \dot{\theta} + \vec{e}_\phi \dot{\phi} \sin\theta \quad (12.143)$$

En ce qui concerne l'accélération nous obtenons:

$$\begin{aligned} \vec{a} = \ddot{\vec{r}} &= \begin{pmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{pmatrix} = \begin{pmatrix} d_t(\dot{r}\sin\theta\cos\phi + r\cos\theta\cos\phi\dot{\theta} - r\sin\theta\sin\phi\dot{\phi}) \\ d_t(\dot{r}\sin\theta\sin\phi + r\cos\theta\sin\phi\dot{\theta} + r\sin\theta\cos\phi\dot{\phi}) \\ d_t(\dot{r}\cos\theta - r\sin\theta\dot{\theta}) \end{pmatrix} \\ &= \begin{pmatrix} (\ddot{r}\sin\theta\cos\phi + \dot{r}\cos\theta\cos\phi\dot{\theta} - \dot{r}\sin\theta\sin\phi\dot{\phi}) + (\dot{r}\cos\theta\cos\phi\ddot{\theta} - r\sin\theta\cos\phi\ddot{\theta}\dot{\theta} \\ + r\cos\theta\cos\phi\ddot{\phi} - r\cos\theta\sin\phi\ddot{\phi}\dot{\theta} - (\dot{r}\sin\theta\sin\phi\dot{\phi} + r\cos\theta\sin\phi\dot{\phi}\dot{\theta} + r\sin\theta\cos\phi\dot{\phi}\ddot{\theta} \\ + r\sin\theta\sin\phi\ddot{\phi}) \\ (\ddot{r}\sin\theta\sin\phi + \dot{r}\cos\theta\sin\phi\dot{\theta} + \dot{r}\sin\theta\cos\phi\dot{\phi}) + (\dot{r}\cos\theta\sin\phi\ddot{\theta} - r\sin\theta\sin\phi\ddot{\theta}\dot{\theta} \\ + r\cos\theta\sin\phi\ddot{\phi} + r\cos\theta\cos\phi\ddot{\phi}\dot{\theta}) + (\dot{r}\sin\theta\cos\phi\dot{\phi} + r\cos\theta\cos\phi\dot{\phi}\dot{\theta} - r\sin\theta\sin\phi\dot{\phi}\ddot{\theta} \\ + r\sin\theta\cos\phi\ddot{\phi}) \\ (\ddot{r}\cos\theta - \dot{r}\sin\theta\dot{\theta}) - (\dot{r}\sin\theta\ddot{\theta} + r\cos\theta\ddot{\theta}\dot{\theta} + r\sin\theta\ddot{\theta}) \end{pmatrix} \\ &= (\ddot{r} - r\dot{\theta}^2) \begin{pmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{pmatrix} + (2\dot{r}\dot{\theta} + r\ddot{\theta}) \begin{pmatrix} \cos\theta\cos\phi \\ \cos\theta\sin\phi \\ -\sin\theta \end{pmatrix} \\ &\quad + (2\dot{r}\sin\theta\dot{\phi} + 2r\cos\theta\dot{\theta}\dot{\phi} + r\sin\theta\ddot{\phi}) \begin{pmatrix} -\sin\phi \\ \cos\phi \\ 0 \end{pmatrix} + r\sin\theta\dot{\phi}^2 \begin{pmatrix} -\cos\phi \\ \sin\phi \\ 0 \end{pmatrix} \\ &= (\ddot{r} - r\dot{\theta}^2)\vec{e}_r + (2\dot{r}\dot{\theta} + r\ddot{\theta})\vec{e}_\theta + (2\dot{r}\sin\theta\dot{\phi} + 2r\cos\theta\dot{\theta}\dot{\phi} + r\sin\theta\ddot{\phi})\vec{e}_\phi - r\sin\theta\dot{\phi}^2 \begin{pmatrix} \cos\phi \\ \sin\phi \\ 0 \end{pmatrix} \end{aligned} \quad (12.144)$$

Or, nous avons:

$$\begin{pmatrix} \cos \phi \\ \sin \phi \\ 0 \end{pmatrix} = \begin{pmatrix} \cos \phi \sin^2 \theta + \cos \phi \cos^2 \theta \\ \sin \phi \sin^2 \theta + \sin \phi \cos^2 \theta \\ 0 \end{pmatrix} = \sin \theta \begin{pmatrix} \sin \theta \cos \phi \\ \sin \theta \sin \phi \\ \cos \theta \end{pmatrix} + \cos \theta \begin{pmatrix} \cos \theta \cos \phi \\ \cos \theta \sin \phi \\ -\sin \theta \end{pmatrix} \quad (12.145)$$

$$= \sin \theta \vec{e}_r + \cos \theta \vec{e}_\theta$$

Donc il vient:

$$\begin{aligned} \vec{a} = & (\ddot{r} - r\dot{\theta}^2) \vec{e}_r + (2\dot{r}\dot{\theta} + r\ddot{\theta}) \vec{e}_\theta + (2\dot{r} \sin \theta \dot{\phi} + 2r \cos \theta \dot{\theta} \dot{\phi} + r \sin \theta \ddot{\phi}) \vec{e}_\phi \\ & - r \sin \theta \dot{\phi}^2 (\sin \theta \vec{e}_r + \cos \theta \vec{e}_\theta) \end{aligned} \quad (12.146)$$

Soit au final:

$$\boxed{\vec{a} = \vec{e}_r (\ddot{r} - r\dot{\theta}^2 - r\dot{\phi}^2 \sin^2 \theta) + \vec{e}_\theta (r\ddot{\theta} + 2\dot{r}\dot{\theta} - r\dot{\phi}^2 \sin \theta \cos \theta) + \vec{e}_\phi (r\ddot{\phi} \sin \theta + 2r\dot{\theta}\dot{\phi} \cos \theta + 2\dot{r}\dot{\phi} \sin \theta)} \quad (12.147)$$

8.3. SYSTÈME DE COORDONNÉES CYLINDRIQUES

Le système de coordonnées cylindriques (très utile dans l'étude des systèmes à mouvements hélicoïdaux) est assez semblable à celui des coordonnées sphériques puisqu'il peut être vu comme une tranche de la sphère. Soit le schéma:

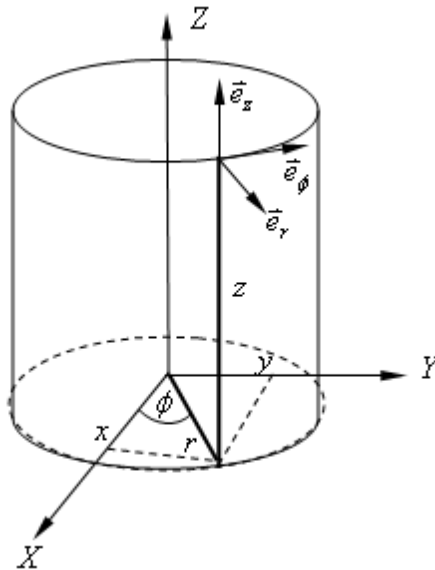


Figure: 12.8 - Représentation du système de coordonnées cylindriques

il vient sans peine qu'en coordonnées cylindriques:

$$r = \sqrt{x^2 + y^2} \quad \phi = \arccos\left(\frac{x}{r}\right) \quad \phi = \arcsin\left(\frac{y}{r}\right) \text{ et } \phi = \arctan\left(\frac{y}{x}\right) \quad (12.148)$$

et inversement:

$$x = r \cos \phi \quad y = r \sin \phi \quad z = z \quad (12.149)$$

Maintenant il nous faut trouver les expressions qui relient les vecteurs de la base cylindrique que nous choisissons de noter $\vec{e}_r, \vec{e}_\phi, \vec{e}_z$ (au lieu de $\hat{r}, \hat{\phi}, \hat{z}$ comme cela se fait traditionnellement) avec les vecteurs de la base cartésienne $\vec{e}_x, \vec{e}_y, \vec{e}_z$. Nous avons, identiquement à ce que nous avons fait pour les coordonnées sphériques:

$$\begin{aligned} \vec{e}_r &= \frac{\vec{r}}{r} = \frac{x\vec{e}_x + y\vec{e}_y}{r} = \vec{e}_x \cos \phi + \vec{e}_y \sin \phi \\ \vec{e}_\phi &= \frac{\vec{e}_z \times \vec{e}_r}{\sin \phi} = -\vec{e}_x \sin \phi + \vec{e}_y \cos \phi \\ \vec{e}_z &= \vec{e}_z \end{aligned} \quad (12.150)$$

Indiquons qu'en divisant par $\sin \phi$, nous nous assurons de par les propriétés de la norme du produit vectoriel que:

$$\|\vec{e}_\phi\| = \|\vec{e}_z\| \|\vec{e}_r\| \sin \phi = 1 \quad (12.151)$$

sera bien normalisé à l'unité! Dans le cas des coordonnées cylindriques l'angle étant de toute façon droit, nous ne serions pas obligé d'indiquer cette division, mais nous avons fait ce choix par souci d'homogénéité avec les développements précédents...

Remarque: Il est important de remarquer que le produit vectoriel de deux vecteurs de base donne toujours le troisième vecteur de base perpendiculaire (comme pour les coordonnées cartésiennes et sphériques donc!).

Pour des besoins ultérieurs, déterminons les différentielles partielles de chacune de ces coordonnées:

$$\begin{aligned} \frac{\partial \vec{e}_r}{\partial r} &= 0; \frac{\partial \vec{e}_r}{\partial \phi} = -\vec{e}_x \sin \phi + \vec{e}_y \cos \phi = \vec{e}_\phi; \frac{\partial \vec{e}_r}{\partial z} = 0 \\ \frac{\partial \vec{e}_\phi}{\partial r} &= 0; \frac{\partial \vec{e}_\phi}{\partial \phi} = -\vec{e}_x \cos \phi - \vec{e}_y \sin \phi = -\vec{e}_r; \frac{\partial \vec{e}_\phi}{\partial z} = 0 \\ \frac{\partial \vec{e}_z}{\partial r} &= 0; \frac{\partial \vec{e}_z}{\partial \phi} = 0; \frac{\partial \vec{e}_z}{\partial z} = 0 \end{aligned} \quad (12.152)$$

Nous allons également utiliser plus tard (pour l'étude des opérateurs vectoriels) la variation $d\vec{r}$ exprimée en coordonnées cylindriques:

$$d\vec{r} = d(r\vec{e}_r) = \vec{e}_r dr + r d\vec{e}_r = \vec{e}_r dr + r \left(\frac{\partial \vec{e}_r}{\partial r} dr + \frac{\partial \vec{e}_r}{\partial \phi} d\phi + \frac{\partial \vec{e}_r}{\partial z} dz \right) = \vec{e}_r dr + r \vec{e}_\phi d\phi \quad (12.153)$$

Pour exprimer la vitesse et l'accélération en coordonnées cylindriques, nous aurons également besoin des dérivées par rapport au temps:

$$\begin{aligned}
\dot{\vec{e}}_r &= \frac{\partial \vec{e}_r}{\partial r} \dot{r} + \frac{\partial \vec{e}_r}{\partial \phi} \dot{\phi} + \frac{\partial \vec{e}_r}{\partial z} \dot{z} = \vec{e}_\phi \dot{\phi} \\
\dot{\vec{e}}_\phi &= \frac{\partial \vec{e}_\phi}{\partial r} \dot{r} + \frac{\partial \vec{e}_\phi}{\partial \phi} \dot{\phi} + \frac{\partial \vec{e}_\phi}{\partial z} \dot{z} = -\vec{e}_r \dot{\phi} \quad (12.154) \\
\dot{\vec{e}}_z &= \frac{\partial \vec{e}_z}{\partial r} \dot{r} + \frac{\partial \vec{e}_z}{\partial \phi} \dot{\phi} + \frac{\partial \vec{e}_z}{\partial z} \dot{z} = \vec{0}
\end{aligned}$$

Donc si nous faisons maintenant un peu de physique, nous avons (rappelons que la composante z est indépendante des autres composantes cylindriques):

$$\vec{v} = \dot{\vec{r}} + \dot{\vec{z}} = \dot{\vec{e}}_r r + r \dot{\vec{e}}_r + \vec{e}_z \dot{z} + \vec{e}_z \dot{z} \quad (12.155)$$

ce qui nous amène à:

$$\vec{v} = \dot{r} \vec{e}_r + r \dot{\phi} \vec{e}_\phi + \dot{z} \vec{e}_z \quad (12.156)$$

Pour l'accélération nous obtenons (exactement la même démarche que pour déterminer l'expression de la vitesse):

$$\vec{a} = \ddot{\vec{r}} = \ddot{r} \vec{e}_r + \dot{r} \dot{\vec{e}}_r + \ddot{\phi} r \vec{e}_\phi + 2\dot{r} \dot{\phi} \vec{e}_\phi + \ddot{z} \vec{e}_z \quad (12.157)$$

8.4. SYSTÈME DE COORDONNÉES POLAIRES

Le système de coordonnées polaires est très semblable à celui des coordonnées cylindriques puisqu'il peut être vu comme un retranchement d'une dimension (la hauteur) du système cylindrique.

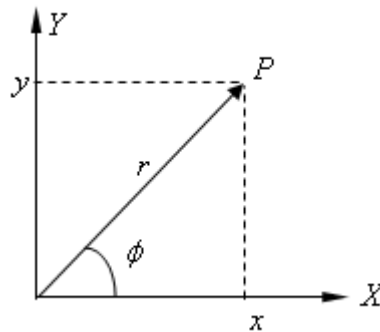


Figure: 12.9 - Représentation du système de coordonnées polaires

Ainsi, il vient sans peine qu'en coordonnées polaires:

$$r = \sqrt{x^2 + y^2} \quad \phi = \arccos\left(\frac{x}{r}\right) \quad \phi = \arcsin\left(\frac{y}{r}\right) \text{ et } \phi = \arctan\left(\frac{y}{x}\right) \quad (12.158)$$

et inversement:

$$x = r \cos \phi, y = r \sin \phi \quad (12.159)$$

Maintenant il nous faut trouver les expressions qui relient les vecteurs de la base polaire que je choisis de noter $\vec{e}_r, \vec{e}_\phi$ (au lieu de $\hat{r}, \hat{\phi}$ comme cela se fait traditionnellement) avec les vecteurs de la base

cartésienne $\vec{e}_x, \vec{e}_y, \vec{e}_z$. Nous avons identiquement à ce que nous avons fait pour les coordonnées sphériques:

$$\begin{aligned}\vec{e}_r &= \frac{\vec{r}}{r} = \frac{x\vec{e}_x + y\vec{e}_y}{r} = \vec{e}_x \cos \phi + \vec{e}_y \sin \phi \\ \vec{e}_\phi &= \frac{\vec{e}_z \times \vec{e}_r}{\sin \theta} = -\vec{e}_x \sin \phi + \vec{e}_y \cos \phi\end{aligned}\quad (12.160)$$

Explications pour la seconde ligne: en divisant par $\sin \theta$, nous nous assurons de par les propriétés de la norme du produit vectoriel que $\|\vec{e}_\phi\| = \|\vec{e}_z\| \|\vec{e}_r\| \sin \phi = 1$ sera bien normalisé à l'unité. Dans le cas des coordonnées polaires l'angle étant de toute façon droit, nous ne serions pas obligé d'indiquer cette division, mais nous avons fait ce choix par souci d'homogénéité avec les développements précédents.

Pour des besoins ultérieurs, déterminons les différentielles partielles de chacune de ces coordonnées:

$$\begin{aligned}\frac{\partial \vec{e}_r}{\partial r} &= 0; \frac{\partial \vec{e}_r}{\partial \phi} = -\vec{e}_x \sin \phi + \vec{e}_y \cos \phi = \vec{e}_\phi \\ \frac{\partial \vec{e}_\phi}{\partial r} &= 0; \frac{\partial \vec{e}_\phi}{\partial \phi} = -\vec{e}_x \cos \phi - \vec{e}_y \sin \phi = -\vec{e}_r\end{aligned}\quad (12.161)$$

Nous allons également utiliser plus tard (pour l'étude des opérateurs vectoriels) la variation $d\vec{r}$ exprimée en coordonnées polaires:

$$\begin{aligned}d\vec{r} &= d(r\vec{e}_r) = \vec{e}_r dr + r d\vec{e}_r = \vec{e}_r dr + r \left(\frac{\partial \vec{e}_r}{\partial r} dr + \frac{\partial \vec{e}_r}{\partial \phi} d\phi \right) \\ &= \vec{e}_r dr + \vec{e}_\phi r d\phi\end{aligned}\quad (12.162)$$

Pour exprimer la vitesse et l'accélération en coordonnées polaires, nous aurons également besoin des dérivées par rapport au temps:

$$\begin{aligned}\dot{\vec{e}}_r &= \frac{\partial \vec{e}_r}{\partial r} \dot{r} + \frac{\partial \vec{e}_r}{\partial \phi} \dot{\phi} = \vec{e}_\phi \dot{\phi} \\ \dot{\vec{e}}_\phi &= \frac{\partial \vec{e}_\phi}{\partial r} \dot{r} + \frac{\partial \vec{e}_\phi}{\partial \phi} \dot{\phi} = -\vec{e}_r \dot{\phi}\end{aligned}\quad (12.163)$$

Donc si nous faisons maintenant un peu de physique, nous avons:

$$\vec{v} = \dot{\vec{r}} = \dot{\vec{e}}_r r + \vec{e}_r \dot{r} \quad (12.164)$$

ce qui nous amène à:

$$\boxed{\vec{v} = \dot{r}\vec{e}_r + \vec{e}_\phi r \dot{\phi}} \quad (12.165)$$

où le premier terme est la composante radiale de la vitesse et le second la composante tangentielle de la vitesse (angulaire).

Pour l'accélération nous obtenons (exactement dans la même démarche que pour déterminer

l'expression de la vitesse):

$$\vec{a} = \vec{e}_r(\ddot{r} - \dot{\phi}^2 r) + \vec{e}_\phi(2\dot{r}\dot{\phi} + r\ddot{\phi}) \quad (12.166)$$

où le premier terme est l'accélération radiale, le second l'accélération centripète, le troisième l'accélération de Coriolis et le quatrième l'accélération tangentielle.

9. OPÉRATEURS DIFFÉRENTIELS

Définition: Définir un champ scalaire, vectoriel ou tensoriel, dans un volume V , c'est définir une application qui, à tout point $\vec{x}$ de ce volume V , associe respectivement une grandeur scalaire, vectorielle ou tensorielle.

Ainsi, l'application f qui, à tout point $\vec{x}$ de V , de coordonnées spatiales x, y, z associe la valeur scalaire $f(\vec{x}) = f(x, y, z)$ est un champ scalaire dans V .

En chaque point d'un volume traversé par un fluide en mouvement, le vecteur qui coïncide à chaque instant avec la vitesse de la particule changeante qui passe en ce point à ce même instant définit un champ vectoriel 3D, éventuellement variable dans le temps. Les champs ainsi définis constituent un outil mathématique de base dans l'ensemble de la physique.

Remarque: Lorsque nous représentons graphiquement un champ scalaire, l'ensemble des points continus de valeur égale constitue ce que l'on appelle des "isolignes" ou plus couramment "courbes de niveau".

Le gradient, la divergence et le rotationnel sont les trois principaux opérateurs différentiels linéaires du premier ordre que nous allons présenter ici. Cela signifie qu'ils ne font intervenir que des dérivées partielles (ou différentielles) premières des champs, à la différence, par exemple, du laplacien qui fait intervenir des dérivées partielles d'ordre 2.

Nous les rencontrerons en particulier dans les chapitres traitant de la mécanique des fluides, de l'électromagnétisme ainsi qu'en physique quantique ondulatoire où ils permettent d'exprimer facilement certaines propriétés.

9.1. GRADIENTS D'UN CHAMP SCALAIRE

Le gradient est un opérateur qui s'applique à un champ de scalaires et le transforme en un champ de vecteurs. Intuitivement, le gradient indique la direction de la plus grande variation du champ scalaire, et l'intensité de cette variation. Par exemple, le gradient de l'altitude est dirigé selon la ligne de plus grande pente et sa norme augmente avec la pente.

Soit un champ scalaire tridimensionnel $f(x, y, z)$, où x et y et z sont les coordonnées cartésiennes d'un point M de l'espace. Lorsque M se déplace dans l'espace selon le vecteur $d\vec{r}$ de composantes dx, dy et dz , le champ scalaire f varie de df selon la différentielle totale:

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \frac{\partial f}{\partial z} dz \quad (12.167)$$

A partir de cette relation, nous pouvons définir "l'opérateur gradient" d'un champ scalaire tel que:

$$df = \overrightarrow{\text{grad}}(f) \circ d\vec{r} = \vec{\nabla}(f) \circ d\vec{r} \quad (12.168)$$

où:

$$\overrightarrow{\text{grad}}(f) \quad (12.169)$$

est un terme vectoriel appelé le "gradient du champ scalaire f ". Pour condenser l'écriture, nous utilisons parfois le symbole $\vec{\nabla}$ nommé le "nabla du champ scalaire f ".

Le vecteur obtenu par le calcul du gradient a les quatre propriétés suivantes:

P1. Ses composantes représentent la variation (pente) de la fonction f selon les différentes directions de l'espace.

P2. Sa norme est la variation maximale de f en fonction de la distance.

P3. Sa direction est selon la variation maximale de f en fonction de la distance.

P4. Le sens indique les valeurs où f augmente.

A partir de la définition et de la différentielle totale, nous obtenons

$$\frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \frac{\partial f}{\partial z} dz = \vec{\nabla}(f) \circ d\vec{r} = \vec{\nabla}(f)_x dx + \vec{\nabla}(f)_y dy + \vec{\nabla}(f)_z dz \quad (12.170)$$

Ce qui nous amène à poser que:

$$\vec{\nabla}(f)_x = \frac{\partial f}{\partial x}, \vec{\nabla}(f)_y = \frac{\partial f}{\partial y}, \vec{\nabla}(f)_z = \frac{\partial f}{\partial z} \quad (12.171)$$

et donc que finalement l'opérateur "gradient en coordonnées cartésiennes" est donné par:

$$\vec{\nabla} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) \quad (12.172)$$

Finalement nous voyons que le gradient d'un champ scalaire $f(x, y, z)$ est le champ vectoriel dont les composantes en chaque point sont les trois dérivées du champ scalaire f par rapport aux trois coordonnées spatiales, notées ici x, y, z .

La variation de f pour un déplacement $d\vec{r}$ est donc le produit scalaire de $d\vec{r}$ par le gradient du champ f . Or, un déplacement infinitésimal $d\vec{r}$ effectué le long d'une isoligne (décrivant une isosurface), du champ scalaire tridimensionnel $f(x, y, z)$ n'engendre aucune variation df de f . Le produit scalaire évoqué est donc nul dans ce cas, ce qui implique que $d\vec{r}$ et $\vec{\nabla}(f)$ sont perpendiculaires.

En considérant cette fois un déplacement perpendiculaire aux isolignes, nous montrons facilement que le vecteur gradient de f est dirigé depuis les faibles valeurs de f vers les fortes valeurs de f . Son module étant d'autant plus grand que f varie rapidement au voisinage du point considéré.

Par sa direction, son sens et son module, le vecteur gradient d'un champ en un point comporte donc des indications sur la manière dont varie le champ autour de ce point.

Remarque: Une des conditions nécessaire et suffisante pour qu'un champ de vecteurs soit le gradient d'un champ scalaire f est que ce champ vectoriel soit irrotationnel (voir plus loin l'opérateur rotationnel d'un champ vectoriel).

Après avoir défini le gradient en coordonnées cartésiennes x, y, z nous devons nous intéresser à l'expression de cet opérateur dans d'autres systèmes de coordonnées. Il est fréquent en physique d'avoir à utiliser les coordonnées cylindriques, polaires et sphériques pour simplifier l'étude formelle de systèmes physiques. Ainsi, si nous faisons référence à notre étude des systèmes de coordonnées, nous avons (rappel) d'abord en coordonnées polaires:

$$d\vec{r} = \vec{e}_r dr + \vec{e}_\phi r d\phi \quad (12.173)$$

Or, avec la définition du gradient en coordonnées cartésiennes, nous avons en coordonnées polaires la définition suivante:

$$df = \vec{\nabla}(f) \circ d\vec{r} \quad (12.174)$$

Si nous exprimons la différentielle totale exacte (cf. chapitre de Calcul Différentiel Et Intégral) de df nous obtenons les relations suivantes:

$$df = \frac{\partial f}{\partial r} dr + \frac{\partial f}{\partial \phi} d\phi = \vec{\nabla}(f) \circ d\vec{r} = \vec{\nabla}(f) \circ (\vec{e}_r dr + \vec{e}_\phi r d\phi) \quad (12.175)$$

Ce qui nous permet d'obtenir la relation:

$$df = \frac{\partial f}{\partial r} dr + \frac{\partial f}{\partial \phi} d\phi = \vec{\nabla}(f) \circ (\vec{e}_r dr + \vec{e}_\phi r d\phi) = \vec{\nabla}(f)_r dr + \vec{\nabla}(f)_\phi r d\phi \quad (12.176)$$

donc:

$$\frac{\partial f}{\partial r} dr + \frac{\partial f}{\partial \phi} d\phi = \vec{\nabla}(f)_r dr + \vec{\nabla}(f)_\phi r d\phi \quad (12.177)$$

ce qui nous amène à:

$$\vec{\nabla}(f)_r = \frac{\partial f}{\partial r}; \vec{\nabla}(f)_\phi = \frac{1}{r} \frac{\partial f}{\partial \phi} \quad (12.178)$$

Ainsi le "**gradient en coordonnées polaires**" s'exprime comme:

$$\vec{\nabla} = \left(\frac{\partial}{\partial r}, \frac{1}{r} \frac{\partial}{\partial \phi} \right) \quad (12.179)$$

Occupons-nous maintenant de l'expression du gradient en coordonnées cylindriques. Rappelons que lors de l'étude des différents systèmes de coordonnées nous avons obtenu pour les coordonnées

cylindriques:

$$d\vec{r} = \vec{e}_r dr + \vec{e}_\phi r d\phi \quad (12.180)$$

Donc nous savons déjà que l'expression du gradient en coordonnées cylindriques sera identique à celle en coordonnées polaires à l'exception de l'ajout de la composante verticale z indépendante des autres coordonnées. Ainsi, nous obtenons l'opérateur "**gradient en coordonnées cylindriques**":

$$\vec{\nabla} = \left(\frac{\partial}{\partial r}, \frac{1}{r} \frac{\partial}{\partial \phi}, \frac{\partial}{\partial z} \right) \quad (12.181)$$

Occupons-nous maintenant de l'expression du gradient en coordonnées sphériques. Rappelons que lors de l'étude des différents systèmes de coordonnées nous avons obtenu pour les coordonnées sphériques:

$$d\vec{r} = \vec{e}_r dr + \vec{e}_\theta r d\theta + \vec{e}_\phi r \sin \theta d\phi \quad (12.182)$$

Or, avec la définition du gradient en coordonnées cartésiennes, nous avons en coordonnées sphériques la définition suivante:

$$df = \vec{\nabla}(f) \circ d\vec{r} \quad (12.183)$$

Si nous exprimons la différentielle totale de df nous obtenons les relations suivantes:

$$df = \frac{\partial f}{\partial r} dr + \frac{\partial f}{\partial \phi} d\phi + \frac{\partial f}{\partial \theta} d\theta = \vec{\nabla}(f) \circ d\vec{r} = \vec{\nabla}(f) \circ (\vec{e}_r dr + \vec{e}_\theta r d\theta + \vec{e}_\phi r \sin \theta d\phi) \quad (12.184)$$

Ce qui nous permet d'obtenir la relation (nous utilisons maintenant la notation qui use de l'opérateur "nabla"):

$$\begin{aligned} \frac{\partial f}{\partial r} dr + \frac{\partial f}{\partial \phi} d\phi + \frac{\partial f}{\partial \theta} d\theta &= \vec{\nabla}(f) \circ (\vec{e}_r dr + \vec{e}_\theta r d\theta + \vec{e}_\phi r \sin \theta d\phi) \\ &= \vec{\nabla}(f)_r dr + \vec{\nabla}(f)_\theta r d\theta + \vec{\nabla}(f)_\phi r \sin \theta d\phi \end{aligned} \quad (12.185)$$

La relation:

$$\frac{\partial f}{\partial r} dr + \frac{\partial f}{\partial \phi} d\phi + \frac{\partial f}{\partial \theta} d\theta = \vec{\nabla}(f)_r dr + \vec{\nabla}(f)_\theta r d\theta + \vec{\nabla}(f)_\phi r \sin \theta d\phi \quad (12.186)$$

Nous impose:

$$\vec{\nabla}(f)_r = \frac{\partial f}{\partial r}; \vec{\nabla}(f)_\phi = \frac{1}{r} \frac{\partial f}{\partial \phi}; \vec{\nabla}(f)_\theta = \frac{1}{r \sin \theta} \frac{\partial f}{\partial \theta} \quad (12.187)$$

Ainsi l'opérateur "**gradient en coordonnées sphériques**" s'exprime comme:

$$\vec{\nabla} = \left(\frac{\partial}{\partial r}, \frac{1}{r} \frac{\partial}{\partial \theta}, \frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} \right) \quad (12.188)$$

Nous avons donc finalement vu toutes les expressions de l'opérateur gradient dans les systèmes

cartésiens, polaires, cylindriques et sphériques.

9.2. GRADIENTS D'UN CHAMP DE VECTEURS

Le gradient d'un champ vectoriel $\vec{f}(x, y, z)$ est le champ dit "**champ tensoriel**" défini par les 9 relations suivantes en coordonnées cartésiennes:

$$\vec{\nabla}(\vec{f}) = \frac{\partial f_i}{\partial x_j} = \begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} & \frac{\partial f_1}{\partial z} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} & \frac{\partial f_2}{\partial z} \\ \frac{\partial f_3}{\partial x} & \frac{\partial f_3}{\partial y} & \frac{\partial f_3}{\partial z} \end{bmatrix} \quad (12.189)$$

Nous utiliserons un tel gradient lors de notre étude dans le chapitre de Génie Météo de l'effet Papillon dont l'origine vient de la détermination des équations de Navier-Stokes en Mécanique des Milieux Continus.

Par les 4 relations suivantes en coordonnées polaires:

$$\vec{\nabla}(\vec{f}) = \begin{bmatrix} \frac{\partial f_1}{\partial r} & \frac{1}{r} \frac{\partial f_1}{\partial \phi} \\ \frac{\partial f_2}{\partial r} & \frac{1}{r} \frac{\partial f_2}{\partial \phi} \end{bmatrix} \quad (12.190)$$

Par les 9 relations suivantes en coordonnées cylindriques:

$$\vec{\nabla}(\vec{f}) = \begin{bmatrix} \frac{\partial f_1}{\partial r} & \frac{1}{r} \frac{\partial f_1}{\partial \phi} & \frac{\partial f_1}{\partial z} \\ \frac{\partial f_2}{\partial r} & \frac{1}{r} \frac{\partial f_2}{\partial \phi} & \frac{\partial f_2}{\partial z} \\ \frac{\partial f_3}{\partial r} & \frac{1}{r} \frac{\partial f_3}{\partial \phi} & \frac{\partial f_3}{\partial z} \end{bmatrix} \quad (12.191)$$

Par les 9 relations suivantes en coordonnées sphériques:

$$\vec{\nabla}(\vec{f}) = \begin{bmatrix} \frac{\partial f_1}{\partial r} & \frac{1}{r} \frac{\partial f_1}{\partial \theta} & \frac{1}{r \sin \theta} \frac{\partial f_1}{\partial \phi} \\ \frac{\partial f_2}{\partial r} & \frac{1}{r} \frac{\partial f_2}{\partial \theta} & \frac{1}{r \sin \theta} \frac{\partial f_2}{\partial \phi} \\ \frac{\partial f_3}{\partial r} & \frac{1}{r} \frac{\partial f_3}{\partial \theta} & \frac{1}{r \sin \theta} \frac{\partial f_3}{\partial \phi} \end{bmatrix} \quad (12.192)$$

Nous avons donc finalement vu toutes les expressions du gradient d'un champ vectoriel dans les systèmes cartésiens, polaires, cylindriques et sphériques.

9.3. DIVERGENCES D'UN CHAMP DE VECTEURS

La divergence s'applique à un champ de vecteurs et le transforme en un champ de scalaires. Intuitivement, et dans les cas le plus courant, la divergence d'un champ vectoriel exprime sa tendance à provenir ou converger vers certains points.

Cependant, il faut distinguer deux contributions à la divergence que nous définirons rigoureusement un peu plus loin: l'une due aux variations de direction appelée la "**divergence directionnelle**" et l'autre due aux variations de modules (norme) appelée la "**divergence modulaire**". Ainsi, pour des champs simples, nous pouvons imaginer des cas où la divergence ne serait que modulaire et d'autres, où elle ne serait que directionnelle. Nous pourrions aussi construire un champ où les deux types de divergence coexistent, mais d'effets contraires (convergence modulaire et divergence directionnelle par exemple).

Prenons par exemple un vecteur $\vec{f}$ de l'espace et faisons lui traverser une surface S quelconque. Les physiciens assimilent alors la quantité $\vec{f}$ qui se dirige suivant la normale à la surface au flux de $\vec{f}$ à travers S .

Pour se convaincre de cette analogie nous pouvons imaginer un fluide coulant sur une surface plane, le flux à travers la surface est évidemment nul, par contre si le fluide coule verticalement à travers une surface horizontale le flux sera maximal. Il est alors immédiat de vouloir représenter le flux par le produit scalaire de $\vec{f}$ avec la normale $\vec{n}$ de la surface S .

Remarque: Il faut toujours prendre garde à la direction de $\vec{n}$ car en un point quelconque d'une surface on a en général deux normales.

Si la surface est plane la normale est la même partout mais si elle change suivant les endroits, nous nous intéresserons alors à un petit élément de surface ds .

Si un petit élément de flux est défini par:

$$d\Phi = (\vec{f} \circ \vec{n}) ds \quad (12.193)$$

alors le flux total sera donné par:

$$\Phi = \int_S (\vec{f} \circ \vec{n}) ds \quad (12.194)$$

ce qui est parfois noté (c'est un peu abusif mais pourquoi pas...):

$$\Phi = \int_S \vec{f} \circ d\vec{s} \quad (12.195)$$

Supposons maintenant que notre vecteur $\vec{f}$ déplace un point $M(x, y, z)$ de l'espace en $M'(x + dx, y + dy, z + dz)$ à travers un parallélépipède rectangle de côtés dx , dy et dz :

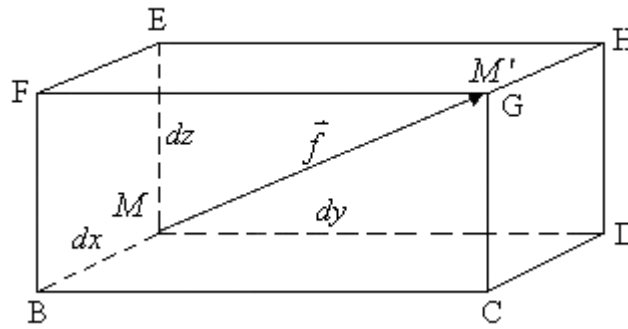


Figure: 12.10 - Déplacement du vecteur à travers un parallélépipède

Nous pouvons décomposer le mouvement (flux) à travers chaque face du parallélépipède (décompositions dans la base orthonormée). Par exemple, si nous nous intéressons à l'élément décomposé du flux à travers la face (dy, dz) décrite par les sommets BCFG nous avons bien évidemment $\vec{n} = (1, 0, 0)$.

Il nous faut encore déterminer comment représenter le flux $\vec{f}$ pour cette direction. Comme le flux est une fonction, c'est-à-dire que chacune de ces composantes peut être dépendante des trois composantes de l'espace (si nous prenons le cas particulier d'une fonction dans $\mathbb{R}^3$) nous avons:

$$\vec{f} = \begin{pmatrix} f_x(x, y, z) \\ f_y(x, y, z) \\ f_z(x, y, z) \end{pmatrix} \quad (12.196)$$

Remarque: Ceux qui ne sont pas convaincus peuvent aller lire le début du chapitre d'Électrodynamique où nous prenons le champ électrique comme (excellent) exemple.

Alors la variation du flux selon x sera donnée par:

$$\vec{f}_x = \begin{pmatrix} f_x(x+dx, y, z) \\ f_y(x+dx, y, z) \\ f_z(x+dx, y, z) \end{pmatrix} - \begin{pmatrix} f_x(x, y, z) \\ f_y(x, y, z) \\ f_z(x, y, z) \end{pmatrix} \quad (12.197)$$

ce qui nous donne:

$$d\Phi_x = (\vec{f}_x \circ \vec{n}_x) dydz = (f_x(x+dx, y, z) - f_x(x, y, z)) dydz = \frac{\partial f_x}{\partial x} dx dy dz \quad (12.198)$$

De même pour les deux autres faces:

$$\begin{aligned} d\Phi_y &= (\vec{f}_y \circ \vec{n}_y) dx dz = (f_y(x, y+dy, z) - f_y(x, y, z)) dx dz = \frac{\partial f_y}{\partial y} dy dx dz \\ d\Phi_z &= (\vec{f}_z \circ \vec{n}_z) dx dy = (f_z(x, y, z+dz) - f_z(x, y, z)) dx dy = \frac{\partial f_z}{\partial z} dz dx dy \end{aligned} \quad (12.199)$$

d'où en sommant:

$$d\Phi = \left(\frac{\partial f_x}{\partial x} + \frac{\partial f_y}{\partial y} + \frac{\partial f_z}{\partial z} \right) dx dy dz = \text{div}(\vec{f}) dV = \vec{\nabla} \circ \vec{f} dV \quad (12.200)$$

Par rapport à la première expression de Φ , le terme $dx dy dz$ est donc un élément de volume et non plus de surface. Nous avons aussi un résultat intéressant:

$$\Phi = \int_V (\vec{\nabla} \circ \vec{f}) dV = \int_S (\vec{f} \circ \vec{n}) dS \quad (12.201)$$

Remarque: Voir les exemples pratiques dans le chapitre d'Électrodynamique où par exemple pour le champ électrique la divergence est nulle pour une charge sphérique libre car les vecteurs pointent dans des directions différentes (divergence directionnelle) et les modules décroissent comme l'inverse du carré du rayon (convergence modulaire). Les deux contributions sont en oppositions et donc la divergence totale est nulle.

Le développement ci-dessus est appelé "**théorème d'Ostrogradsky**" ou "**théorème de Gauss-Ostrogradsky**" ou encore "**théorème de la divergence**" et définit en fait la divergence totale de $\vec{f}$ dans un volume comme le flux de $\vec{f}$ à travers les parois du volume (surface Gauss), ce qu'exprime bien le nom divergence.

Nous définissons l'opérateur "**divergence**" par la relation suivante (la notation tensorielle a été utilisée afin d'abrégé l'écriture) dans un espace à n dimensions:

$$\text{div}(\vec{f}) = \vec{\nabla} \circ \vec{f} = \sum_{i=1}^n \frac{\partial f_i}{\partial x_i} \quad (12.202)$$

Ainsi, nous avons pour l'opérateur "**divergence en coordonnées cartésiennes**":

$$\vec{\nabla} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) \quad (12.203)$$

Si la divergence d'un champ de vecteurs est identiquement nulle en tous les points d'un repère Eulérien, l'intégrale triple du flux de ce champ à travers un volume V sera:

$$\Phi = \iiint_V (\vec{\nabla} \circ \vec{f}) dV = 0 \quad (12.204)$$

Il en résulte que le flux de ce champ de vecteurs à travers les bords du volume est nul, c'est-à-dire que le flux entrant compense le flux sortant. Nous disons qu'un tel champ de vecteurs de divergence nulle présente un flux conservatif.

Pour déterminer l'expression de la divergence en coordonnées polaires rappelons les relations démontrées plus haut:

$$\begin{aligned}\vec{e}_r &= \frac{\vec{r}}{r} = \frac{x\vec{e}_x + y\vec{e}_y}{r} = \vec{e}_x \cos \phi + \vec{e}_y \sin \phi \\ \vec{e}_\phi &= \frac{\vec{e}_z \times \vec{e}_r}{\sin \theta} = -\vec{e}_x \sin \phi + \vec{e}_y \cos \phi\end{aligned}\quad (12.205)$$

Soit à présent $\vec{f} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ une fonction vectorielle. Nous avons :

$$\vec{f} = f_x \vec{e}_x + f_y \vec{e}_y = f_r \vec{e}_r + f_\phi \vec{e}_\phi \quad (12.206)$$

Connaissant l'expression de $\vec{e}_r, \vec{e}_\phi$ en fonction de $\vec{e}_x, \vec{e}_y$, à partir de l'expression ci-dessus nous en déduisons :

$$\begin{aligned}f_x &= \cos \phi \cdot f_r - \sin \phi \cdot f_\phi \\ f_y &= \sin \phi \cdot f_r + \cos \phi \cdot f_\phi\end{aligned}\quad (12.207)$$

La divergence de $\vec{f}$ est définie par $\vec{\nabla} \circ \vec{f} = \frac{\partial f_x}{\partial x} + \frac{\partial f_y}{\partial y}$. Nous avons :

$$\vec{\nabla} \circ \vec{f} = \vec{\nabla}(f_x) \circ \vec{e}_x + \vec{\nabla}(f_y) \circ \vec{e}_y \quad (12.208)$$

Le premier terme vaut (application du gradient en coordonnées polaires!):

$$\begin{aligned}\vec{\nabla}(f_x) \circ \vec{e}_x &= \left(\frac{\partial f_x}{\partial r} \vec{e}_r + \frac{1}{r} \frac{\partial f_x}{\partial \phi} \vec{e}_\phi \right) \circ \vec{e}_x \\ &= \frac{\partial f_x}{\partial r} \vec{e}_r \circ \vec{e}_x + \frac{1}{r} \frac{\partial f_x}{\partial \phi} \vec{e}_\phi \circ \vec{e}_x \\ &= \frac{\partial f_x}{\partial r} \cos \phi - \frac{1}{r} \frac{\partial f_x}{\partial \phi} \sin \phi\end{aligned}\quad (12.209)$$

de la même façon nous obtenons (nous pouvons détailler sur demande) :

$$\begin{aligned}\vec{\nabla}(f_y) \circ \vec{e}_y &= \frac{\partial f_y}{\partial r} \vec{e}_r \circ \vec{e}_y + \frac{1}{r} \frac{\partial f_y}{\partial \phi} \vec{e}_\phi \circ \vec{e}_y \\ &= \frac{\partial f_y}{\partial r} \sin \phi + \frac{1}{r} \frac{\partial f_y}{\partial \phi} \cos \phi\end{aligned}\quad (12.210)$$

En additionnant les deux termes et en exprimant les dérivées partielles des fonctions f_x, f_y en fonction des dérivées partielles des fonctions f_r, f_ϕ à l'aide des relations :

$$\begin{aligned}f_x &= \cos \phi \cdot f_r - \sin \phi \cdot f_\phi \\ f_y &= \sin \phi \cdot f_r + \cos \phi \cdot f_\phi\end{aligned}\quad (12.211)$$

Nous obtenons :

$$\begin{aligned}
\vec{\nabla} \circ \vec{f} &= \frac{\partial f_x}{\partial r} \cos \phi - \frac{1}{r} \frac{\partial f_x}{\partial \phi} \sin \phi + \frac{\partial f_y}{\partial r} \sin \phi + \frac{1}{r} \frac{\partial f_y}{\partial \phi} \cos \phi \\
&= \cos^2 \phi \frac{\partial f_r}{\partial r} - \cos \phi \sin \phi \frac{\partial f_\phi}{\partial r} - \frac{1}{r} \sin \phi \left(-\sin \phi f_r + \cos \phi \frac{\partial f_r}{\partial \phi} - \cos \phi f_\phi - \sin \phi \frac{\partial f_\phi}{\partial \phi} \right) \\
&\quad + \sin^2 \phi \frac{\partial f_r}{\partial r} + \sin \phi \cos \phi \frac{\partial f_\phi}{\partial r} + \frac{1}{r} \cos \phi \left(\cos \phi f_r + \sin \phi \frac{\partial f_r}{\partial \phi} - \sin \phi f_\phi + \cos \phi \frac{\partial f_\phi}{\partial \phi} \right)
\end{aligned} \quad (12.212)$$

Après simplification :

$$\vec{\nabla} \circ \vec{f} = \frac{\partial f_r}{\partial r} + \frac{1}{r} f_r + \frac{1}{r} \frac{\partial f_\phi}{\partial \phi} = \frac{1}{r} \frac{\partial(r f_r)}{\partial r} + \frac{1}{r} \frac{\partial f_\phi}{\partial \phi} \quad (12.213)$$

L'expression de l'opérateur "divergence en coordonnées polaires" est alors:

$$\vec{\nabla} = \left(\frac{1}{r} \frac{\partial(r)}{\partial r}, \frac{1}{r} \frac{\partial}{\partial \phi} \right) \quad (12.214)$$

Pour déterminer l'expression de l'opérateur divergence en coordonnées cylindriques rappelons les relations:

$$\begin{aligned}
\vec{e}_r &= \frac{\vec{r}}{r} = \frac{x\vec{e}_x + y\vec{e}_y}{r} = \vec{e}_x \cos \phi + \vec{e}_y \sin \phi \\
\vec{e}_\phi &= \frac{\vec{e}_z \times \vec{e}_r}{\sin \phi} = -\vec{e}_x \sin \phi + \vec{e}_y \cos \phi \\
\vec{e}_z &= \vec{e}_z
\end{aligned} \quad (12.215)$$

Soit à présent $\vec{f} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ une fonction vectorielle. Nous avons :

$$\vec{f} = f_x \vec{e}_x + f_y \vec{e}_y + f_z \vec{e}_z = f_r \vec{e}_r + f_\phi \vec{e}_\phi + f_z \vec{e}_z \quad (12.216)$$

Connaissant l'expression de $\vec{e}_r, \vec{e}_\phi, \vec{e}_z$ en fonction de $\vec{e}_x, \vec{e}_y, \vec{e}_z$, à partir de l'expression ci-dessus nous en déduisons :

$$\begin{aligned}
f_x &= \cos \phi \cdot f_r - \sin \phi \cdot f_\phi \\
f_y &= \sin \phi \cdot f_r + \cos \phi \cdot f_\phi \\
f_z &= f_z
\end{aligned} \quad (12.217)$$

La divergence de $\vec{f}$ est définie par $\vec{\nabla} \circ \vec{f} = \frac{\partial f_x}{\partial x} + \frac{\partial f_y}{\partial y} + \frac{\partial f_z}{\partial z}$. Nous avons :

$$\vec{\nabla} \circ \vec{f} = \vec{\nabla}(f_x) \circ \vec{e}_x + \vec{\nabla}(f_y) \circ \vec{e}_y + \vec{\nabla}(f_z) \circ \vec{e}_z \quad (12.218)$$

Le premier terme vaut (application du gradient en coordonnées cylindriques):

$$\begin{aligned}
\vec{\nabla}(f_x) \circ \vec{e}_x &= \left(\frac{\partial f_x}{\partial r} \vec{e}_r + \frac{1}{r} \frac{\partial f_x}{\partial \phi} \vec{e}_\phi + \frac{\partial f_x}{\partial z} \vec{e}_z \right) \circ \vec{e}_x \\
&= \frac{\partial f_x}{\partial r} \vec{e}_r \circ \vec{e}_x + \frac{1}{r} \frac{\partial f_x}{\partial \phi} \vec{e}_\phi \circ \vec{e}_x + \frac{\partial f_x}{\partial z} \vec{e}_z \circ \vec{e}_x \\
&= \frac{\partial f_x}{\partial r} \cos \phi - \frac{1}{r} \frac{\partial f_x}{\partial \phi} \sin \phi
\end{aligned} \quad (12.219)$$

de la même façon nous obtenons (nous pouvons détailler sur demande) :

$$\begin{aligned}
\vec{\nabla}(f_y) \circ \vec{e}_y &= \frac{\partial f_y}{\partial r} \vec{e}_r \circ \vec{e}_y + \frac{1}{r} \frac{\partial f_y}{\partial \phi} \vec{e}_\phi \circ \vec{e}_y + \frac{\partial f_y}{\partial z} \vec{e}_z \circ \vec{e}_y \\
&= \frac{\partial f_y}{\partial r} \sin \phi + \frac{1}{r} \frac{\partial f_y}{\partial \phi} \cos \phi
\end{aligned} \quad (12.220)$$

et :

$$\vec{\nabla}(f_z) \circ \vec{e}_z = \frac{\partial f_z}{\partial r} \vec{e}_r \circ \vec{e}_z + \frac{1}{r} \frac{\partial f_z}{\partial \phi} \vec{e}_\phi \circ \vec{e}_z + \frac{\partial f_z}{\partial z} \vec{e}_z \circ \vec{e}_z = \frac{\partial f_z}{\partial z} \quad (12.221)$$

En additionnant les trois termes et en exprimant les dérivées partielles des fonctions f_x, f_y, f_z en fonction des dérivées partielles des fonctions f_r, f_ϕ, f_z à l'aide des relations :

$$\begin{aligned}
f_x &= \cos \phi \cdot f_r - \sin \phi \cdot f_\phi \\
f_y &= \sin \phi \cdot f_r + \cos \phi \cdot f_\phi \\
f_z &= f_z
\end{aligned} \quad (12.222)$$

Nous obtenons :

$$\begin{aligned}
\vec{\nabla} \circ \vec{f} &= \frac{\partial f_x}{\partial r} \cos \phi - \frac{1}{r} \frac{\partial f_x}{\partial \phi} \sin \phi + \frac{\partial f_y}{\partial r} \sin \phi + \frac{1}{r} \frac{\partial f_y}{\partial \phi} \cos \phi + \frac{\partial f_z}{\partial z} \\
&= \cos^2 \phi \frac{\partial f_r}{\partial r} - \cos \phi \sin \phi \frac{\partial f_\phi}{\partial r} - \frac{1}{r} \sin \phi \left(-\sin \phi f_r + \cos \phi \frac{\partial f_r}{\partial \phi} - \cos \phi f_\phi - \sin \phi \frac{\partial f_\phi}{\partial \phi} \right) \\
&\quad + \sin^2 \phi \frac{\partial f_r}{\partial r} + \sin \phi \cos \phi \frac{\partial f_\phi}{\partial r} + \frac{1}{r} \cos \phi \left(\cos \phi f_r + \sin \phi \frac{\partial f_r}{\partial \phi} - \sin \phi f_\phi + \cos \phi \frac{\partial f_\phi}{\partial \phi} \right) + \frac{\partial f_z}{\partial z}
\end{aligned} \quad (12.223)$$

Après simplification :

$$\vec{\nabla} \circ \vec{f} = \frac{\partial f_r}{\partial r} + \frac{1}{r} f_r + \frac{1}{r} \frac{\partial f_\phi}{\partial \phi} + \frac{\partial f_z}{\partial z} = \frac{1}{r} \frac{\partial(r f_r)}{\partial r} + \frac{1}{r} \frac{\partial f_\phi}{\partial \phi} + \frac{\partial f_z}{\partial z} \quad (12.224)$$

L'expression de l'opérateur "divergence en coordonnées cylindriques" est alors:

$$\vec{\nabla} = \left(\frac{1}{r} \frac{\partial(r)}{\partial r}, \frac{1}{r} \frac{\partial}{\partial \phi}, \frac{\partial}{\partial z} \right) \quad (12.225)$$

Pour obtenir l'expression de la divergence en coordonnées sphériques, rappelons les relations:

$$\begin{aligned}\vec{e}_r &= \sin \theta \cos \phi \vec{e}_x + \sin \theta \sin \phi \vec{e}_y + \cos \theta \vec{e}_z \\ \vec{e}_\phi &= -\vec{e}_x \sin \phi + \vec{e}_y \cos \phi \\ \vec{e}_\theta &= -\vec{e}_x \cos \theta \cos \phi - \vec{e}_y \cos \theta \sin \phi - \vec{e}_z \sin \theta\end{aligned}\quad (12.226)$$

Soit à présent $\vec{f} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ une fonction vectorielle. Nous avons :

$$\vec{f} = f_x \vec{e}_x + f_y \vec{e}_y + f_z \vec{e}_z = f_r \vec{e}_r + f_\theta \vec{e}_\theta + f_\phi \vec{e}_\phi \quad (12.227)$$

Connaissant l'expression de $\vec{e}_r, \vec{e}_\theta, \vec{e}_\phi$ en fonction de $\vec{e}_x, \vec{e}_y, \vec{e}_z$, à partir de l'expression ci-dessus nous en déduisons :

$$\begin{aligned}f_x &= \sin \theta \cos \phi \cdot f_r + \cos \theta \cos \phi \cdot f_\theta - \sin \phi \cdot f_\phi \\ f_y &= \sin \theta \sin \phi \cdot f_r + \cos \theta \sin \phi \cdot f_\theta + \cos \phi \cdot f_\phi \\ f_z &= \cos \theta \cdot f_r - \sin \theta \cdot f_\theta\end{aligned}\quad (12.228)$$

La divergence de $\vec{f}$ est définie par $\vec{\nabla} \circ \vec{f} = \frac{\partial f_x}{\partial x} + \frac{\partial f_y}{\partial y} + \frac{\partial f_z}{\partial z}$. Nous avons :

$$\vec{\nabla} \circ f = \vec{\nabla}(f_x) \circ \vec{e}_x + \vec{\nabla}(f_y) \circ \vec{e}_y + \vec{\nabla}(f_z) \circ \vec{e}_z \quad (12.229)$$

Le premier terme vaut (application du gradient en coordonnées sphériques):

$$\begin{aligned}\vec{\nabla}(f_x) \circ \vec{e}_x &= \left(\frac{\partial f_x}{\partial r} \vec{e}_r + \frac{1}{r} \frac{\partial f_x}{\partial \theta} \vec{e}_\theta + \frac{1}{r \sin \theta} \frac{\partial f_x}{\partial \phi} \vec{e}_\phi \right) \circ \vec{e}_x \\ &= \frac{\partial f_x}{\partial r} \vec{e}_r \circ \vec{e}_x + \frac{1}{r} \frac{\partial f_x}{\partial \theta} \vec{e}_\theta \circ \vec{e}_x + \frac{1}{r \sin \theta} \frac{\partial f_x}{\partial \phi} \vec{e}_\phi \circ \vec{e}_x \\ &= \frac{\partial f_x}{\partial r} \sin \theta \cos \phi + \frac{1}{r} \frac{\partial f_x}{\partial \theta} \cos \theta \cos \phi - \frac{\sin \phi}{r \sin \theta} \frac{\partial f_x}{\partial \phi}\end{aligned}\quad (12.230)$$

de la même façon nous obtenons (nous pouvons détailler sur demande) :

$$\vec{\nabla}(f_y) \circ \vec{e}_y = \frac{\partial f_y}{\partial r} \sin \theta \sin \phi + \frac{1}{r} \frac{\partial f_y}{\partial \theta} \cos \theta \sin \phi + \frac{\cos \phi}{r \sin \theta} \frac{\partial f_y}{\partial \phi} \quad (12.231)$$

et :

$$\vec{\nabla}(f_z) \circ \vec{e}_z = \frac{\partial f_z}{\partial r} \cos \theta - \frac{1}{r} \frac{\partial f_z}{\partial \theta} \sin \theta \quad (12.232)$$

En additionnant les trois termes et en exprimant les dérivées partielles des fonctions f_x, f_y, f_z en fonction des dérivées partielles des fonctions f_r, f_θ, f_ϕ à l'aide des relations :

$$\begin{aligned}
 f_x &= \sin \theta \cos \phi \cdot f_r + \cos \theta \cos \phi \cdot f_\theta - \sin \phi \cdot f_\phi \\
 f_y &= \sin \theta \sin \phi \cdot f_r + \cos \theta \sin \phi \cdot f_\theta + \cos \phi \cdot f_\phi \\
 f_z &= \cos \theta \cdot f_r - \sin \theta \cdot f_\theta
 \end{aligned} \quad (12.233)$$

nous obtenons (nous pouvons développer les détails intermédiaires sur demande) :

$$\vec{\nabla} \circ \vec{f} = \frac{1}{r \sin \theta} \frac{\partial f_\phi}{\partial \phi} + \underbrace{\frac{\cos \theta}{r \sin \theta} f_\phi + \frac{1}{r} \frac{\partial f_\theta}{\partial \theta}}_{= \frac{1}{r \sin \theta} \frac{\partial (f_\theta \sin \theta)}{\partial \theta}} + \underbrace{\frac{\partial f_r}{\partial r} + \frac{2}{r} f_r}_{= \frac{1}{r^2} \frac{\partial (r^2 f_r)}{\partial r}} \quad (12.234)$$

Ainsi, l'expression de la divergence en coordonnées sphériques devient:

$$\vec{\nabla} \circ \vec{f} = \frac{1}{r \sin \theta} \frac{\partial f_\phi}{\partial \phi} + \frac{1}{r \sin \theta} \frac{\partial (f_\theta \sin \theta)}{\partial \theta} + \frac{1}{r^2} \frac{\partial (r^2 f_r)}{\partial r} \quad (12.235)$$

et donc l'opérateur de "**divergence en coordonnées sphériques**" est alors:

$$\vec{\nabla} = \left(\frac{1}{r^2} \frac{\partial (r^2)}{\partial r}, \frac{1}{r \sin \theta} \frac{\partial (\sin \theta)}{\partial \theta}, \frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} \right) \quad (12.236)$$

Nous avons donc finalement vu toutes les expressions de la divergence d'un champ vectoriel dans les systèmes cartésiens, polaires, cylindriques et sphériques.

9.4. ROTATIONNELS D'UN CHAMP DE VECTEURS

Le rotationnel d'un champ de vecteurs peut être vu (c'est une simplification!) comme le champ de vecteurs dont les lignes de champs sont perpendiculaires à celles dont nous avons calculé le rotationnel comme le montre l'exemple particulier ci-dessous:

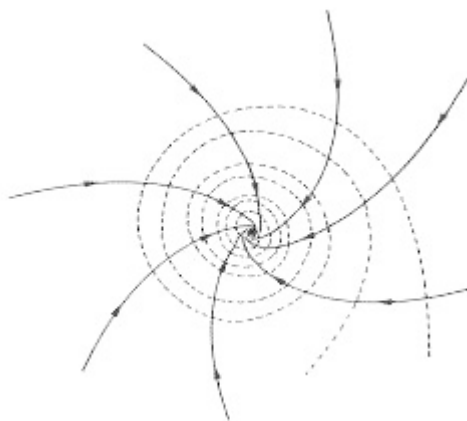


Figure: 12.11 - Exemple de rotationnel d'un champ de vecteurs

Le rotationnel transforme ainsi un champ de vecteurs en un autre champ de vecteurs. Plus difficile à se représenter précisément que le gradient et la divergence, il exprime intuitivement la tendance qu'a un champ à tourner autour d'un point (la manière dont il est tordu).

Exemples:

E1. Dans une tornade, le vent tourne autour de l'oeil du cyclone et le champ vectoriel vitesse du vent a un rotationnel non nul autour de l'oeil.

E2. Le rotationnel du champ des vitesses d'un disque qui tourne à vitesse constante est constant, dirigé selon l'axe de rotation et orienté de telle sorte que la rotation ait lieu, par rapport à lui, dans le sens direct.

Un champ de vecteurs est dit "**irrotationnel**" lorsque le rotationnel de ce champ est identiquement nul en tous les points de l'espace. Dans le cas contraire, nous disons qu'il est "**tourbillonnaire**".

Dans le cas usuel où dx représente un élément de longueur, l'unité du rotationnel est alors l'unité du champ considéré divisée par une unité de longueur. Par exemple, en mécanique des fluides: l'unité du rotationnel d'un champ de vitesse est le radian par unité de temps, comme une vitesse angulaire!

La divergence donne certaines indications sur le comportement d'un vecteur ou d'un champ de vecteurs: comment il se dirige par rapport à la normale et comment il traverse les surfaces, mais c'est insuffisant. Prenons un champ qui aurait la forme d'un cylindre et un autre champ qui aurait la forme d'une hélice de même diamètre que le cylindre. S'ils se dirigent dans la même direction leur divergence sera identique alors que les mouvements sont bien différents. Il faut donc que nous déterminions la manière dont le champ est courbé quand il traverse une surface: ceci va être déterminé par la circulation (comme le travail d'une force par exemple) du vecteur le long d'une courbe fermée, obtenue avec la somme des produits scalaires $\vec{f} \circ d\vec{r}$ sur le contour:

$$\oint \vec{f} \circ d\vec{r} \quad (12.237)$$

en fait ça revient au même de regarder comment est tordu le vecteur par rapport à la normale à la surface ce qui nous amène à définir le "**rotationnel**" ou "vecteur **tourbillon**" en écrivant:

$$\oint \vec{f} \circ d\vec{r} = \int_S \text{rot}(\vec{f}) \wedge d\vec{S} \quad (12.238)$$

qui établit donc une relation entre l'intégrale curviligne et l'intégrale de surface (on transforme donc une intégrale curviligne sur un parcours fermé en une intégrale de surface délimitée par ledit parcours).

En d'autres termes, le rotationnel se calcule en utilisant le fait que la circulation autour d'un circuit élémentaire fermé d'un champ de vecteurs est égal au flux de son rotationnel à travers la surface élémentaire immédiate engendrée par ce circuit.

Ceci est le "**théorème de Stokes**" (qui est plus rigoureusement démontrable avec un formalisme mathématique assez lourd) qui est donc en fait une définition de l'opérateur rotationnel dont nous allons chercher l'expression mathématique explicite:

Soit $\vec{f}$ un champ vectoriel défini dans un espace donné. Nous voulons donc calculer la circulation du $\vec{f}$ autour d'un contour fermé C :

$$\oint \vec{f} \circ d\vec{r} \quad (12.239)$$

Nous choisissons comme contour C le contour d'un rectangle infinitésimal de côté (dx, dy) plongé dans $\mathbb{R}^3$ et parallèle au plan xOy (remarquez que nous parcourons le contour de façon à toujours avoir la

surface à notre gauche):

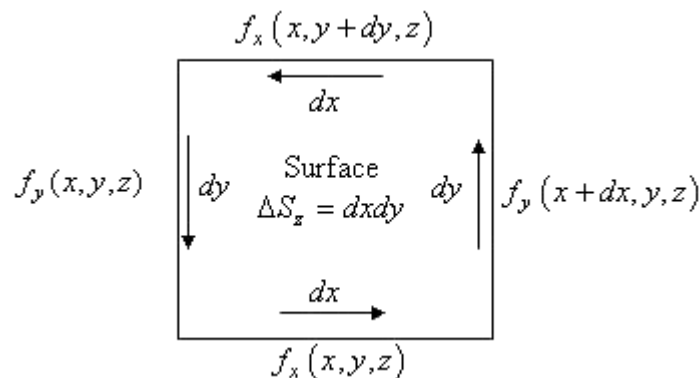


Figure: 12.12 - Contour d'intégration

Pour les deux côtés horizontaux, la contribution à la circulation est:

$$f_x(x, y, z)dx - f_x(x, y + dy, z)dx \quad (12.240)$$

Ce qui nous autorise à écrire:

$$\frac{f_x(x, y, z) - f_x(x, y + dy, z)}{dy} dx dy = -\frac{\partial f_x}{\partial y} dS_z \quad (12.241)$$

De même, pour les faces verticales:

$$\frac{f_y(x + dx, y, z) - f_y(x, y, z)}{dx} dx dy = \frac{\partial f_y}{\partial x} dS_z \quad (12.242)$$

Ainsi, nous avons la circulation selon z:

$$(\vec{f} \circ d\vec{r})_z = \left(\frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y} \right) dS_z \quad (12.243)$$

Ce qui s'écrit aussi sous la forme générale traditionnelle suivante:

$$\oint \left(f_y dx + f_x dy \right) = \int_S \left(\frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y} \right) dS_z \quad (12.244)$$

et constitue le non moins fameux "**théorème de Green**" ou appelé encore "**théorème de Green-Riemann**" que nous retrouverons dans le chapitre d'Analyse Complexe.

Et que nous écrirons dans le cas qui nous intéresse:

$$\int_S \left(\frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y} \right) dS_z = \int_S \text{rot}(\vec{f})_z dS_z \quad (12.245)$$

Par permutation circulaire nous obtenons alors:

$$\text{rot}(\vec{f}) = \left(\frac{\partial f_z}{\partial y} - \frac{\partial f_y}{\partial z} \right) \vec{e}_x + \left(\frac{\partial f_x}{\partial z} - \frac{\partial f_z}{\partial x} \right) \vec{e}_y + \left(\frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y} \right) \vec{e}_z \quad (12.246)$$

Soit sous forme vectorielle condensée:

$$\text{rot}(\vec{f}) = \begin{pmatrix} \frac{\partial f_z}{\partial y} - \frac{\partial f_y}{\partial z} \\ \frac{\partial f_x}{\partial z} - \frac{\partial f_z}{\partial x} \\ \frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y} \end{pmatrix} \quad (12.247)$$

Ce qui permet de mieux comprendre la notation, ou la définition non intuitive du rotationnel dans beaucoup d'ouvrages et qui est:

$$\text{rot}(\vec{f}) = \vec{\nabla} \times \vec{f} \quad (12.248)$$

soit le produit vectoriel de l'opérateur gradient avec le champ vectoriel.

Donc nous avons finalement démontré le théorème de Stokes qui donne bien:

$$\oint \vec{f} \circ d\vec{r} = \int_S (\vec{\nabla} \times \vec{f}) \wedge d\vec{S} \quad (12.249)$$

et en même temps le rotationnel en coordonnées cartésiennes.

Cherchons maintenant à déterminer l'expression du rotationnel en coordonnées cylindriques ((le rotationnel en coordonnées polaires n'étant pas définissable).

En réutilisant la même technique que pour le rotationnel en coordonnées cartésiennes, nous écrivons la circulation de $\vec{f}$ le long d'un contour correspondant à un petit morceau de cylindre orthogonal à (Oz) :

$P_1 P_2 P_3 P_4$.

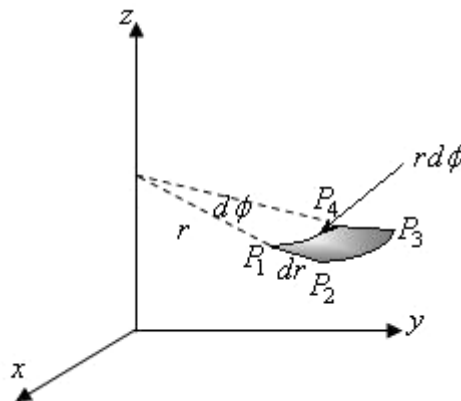


Figure: 12.13 - Représentation du morceau de cylindre

Nous avons alors en fixant z (attention le $d\vec{r}$ n'a rien à voir avec le r du rayon du cylindre... la notation peut être confuse j'en suis désolé!):

$$\begin{aligned}
(\vec{f} \circ d\vec{r})_{P_1 P_2} &= f_r(r, \phi, z) dr \\
(\vec{f} \circ d\vec{r})_{P_2 P_3} &= f_\phi(r+dr, \phi, z)(r+dr) d\phi \\
(\vec{f} \circ d\vec{r})_{P_3 P_4} &= -f_r(r, \phi+d\phi, z) dr \\
(\vec{f} \circ d\vec{r})_{P_4 P_1} &= -f_\phi(r, \phi, z) r d\phi
\end{aligned} \tag{12.250}$$

la circulation totale donne donc après regroupement des termes :

$$\begin{aligned}
\vec{f} \circ d\vec{r} &= f_r(r, \phi, z) dr + f_\phi(r+dr, \phi, z)(r+dr) d\phi - f_r(r, \phi+d\phi, z) dr - f_\phi(r, \phi, z) r d\phi \\
&= r [f_\phi(r+dr, \phi, z) - f_\phi(r, \phi, z)] d\phi + f_\phi(r+dr, \phi, z) d\phi dr - [f_r(r, \phi+d\phi, z) - f_r(r, \phi, z)] dr
\end{aligned} \tag{12.251}$$

Nous ne pouvons pas à cette étape directement comparer avec le rotationnel car il nous est difficile de faire apparaître la différentielle de la surface si nous regardons les différentielles qui apparaissent actuellement dans la circulation. Le mieux est alors de tout diviser par $rd\phi dr$:

$$\begin{aligned}
&\frac{1}{rd\phi dr} \vec{f} \circ d\vec{r} \\
&= \frac{f_\phi(r+dr, \phi, z) - f_\phi(r, \phi, z)}{dr} - \frac{1}{r} \frac{f_r(r, \phi+d\phi, z) - f_r(r, \phi, z)}{d\phi} + \frac{1}{r} f_\phi(r+dr, \phi, z) \tag{12.252} \\
&\underset{\lim_{dr \rightarrow 0}}{\lim_{d\phi \rightarrow 0}} \left(\frac{\partial f_\phi}{\partial r} + \frac{1}{r} f_\phi(r, \phi, z) - \frac{1}{r} \frac{\partial f_r}{\partial \phi} \right) = \frac{1}{r} \frac{\partial (rf_\phi)}{\partial r} - \frac{1}{r} \frac{\partial f_r}{\partial \phi}
\end{aligned}$$

Donc:

$$rot(\vec{f})_z = \frac{1}{r} \frac{\partial (rf_\phi)}{\partial r} - \frac{1}{r} \frac{\partial f_r}{\partial \phi} \tag{12.253}$$

Maintenant nous déterminons le rotationnel en fixant φ . Le problème revient à avoir donc un rectangle dans l'espace que nous parcourons pour déterminer la circulation. Or, nous savons déjà ce qu'est le résultat du rotationnel pour un rectangle en coordonnées cartésiennes:

$$rot(\vec{f})_z = \frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y} \tag{12.254}$$

à la différence que dans les coordonnées cylindriques il faut substituer z par φ , x par z , y par r , f_y par f_ϕ et finalement f_x par f_r (ce choix s'impose toujours simplement parce que la circulation se fait de telle manière que la surface soit toujours à notre gauche) . Ce qui nous donne:

$$rot(\vec{f})_\varphi = \frac{\partial f_r}{\partial z} - \frac{\partial f_z}{\partial r} \tag{12.255}$$

Il ne nous reste plus qu'à trouver la composante du rotationnel en r (soit quand r est fixé). Le calcul est alors plus délicat puisqu'il s'agit de parcourir (positivement toujours!) une surface courbée par la

variation de l'angle ϕ .

Nous avons alors en fixant r :

$$\begin{aligned}
 (\vec{f} \circ d\vec{r})_{P_1 P_2} &= f_\phi(r, \phi, z) r d\phi \\
 (\vec{f} \circ d\vec{r})_{P_2 P_3} &= f_z(r, \phi + d\phi, z) dz \\
 (\vec{f} \circ d\vec{r})_{P_3 P_4} &= -f_\phi(r, \phi, z + dz) r d\phi \\
 (\vec{f} \circ d\vec{r})_{P_4 P_1} &= -f_z(r, \phi, z) dz
 \end{aligned} \quad (12.256)$$

la circulation totale donne donc après regroupement des termes :

$$\begin{aligned}
 \vec{f} \circ d\vec{r} &= f_z(r, \phi + d\phi, z) dz - f_z(r, \phi, z) dz - r f_\phi(r, \phi, z + dz) d\phi + r f_\phi(r, \phi, z) d\phi \\
 &= [f_z(r, \phi + d\phi, z) - f_z(r, \phi, z)] dz - r [f_\phi(r, \phi, z + dz) - f_\phi(r, \phi, z)] d\phi
 \end{aligned} \quad (12.257)$$

Nous ne pouvons pas à cette étape directement comparer avec le rotationnel car il nous est difficile de faire apparaître la différentielle de la surface si nous regardons les différentielles qui apparaissent actuellement dans la circulation. Le mieux est alors tout diviser par $r d\phi dz$:

$$\frac{1}{r d\phi dz} \vec{f} \circ d\vec{r} = \frac{1}{r} \frac{f_z(r, \phi + d\phi, z) - f_z(r, \phi, z)}{d\phi} - \frac{f_\phi(r, \phi, z + dz) - f_\phi(r, \phi, z)}{dz} = \frac{1}{r} \frac{\partial f_z}{\partial \phi} - \frac{\partial f_\phi}{\partial z} \quad (12.258)$$

Donc finalement:

$$\text{rot}(\vec{f})_r = \frac{1}{r} \frac{\partial f_z}{\partial \phi} - \frac{\partial f_\phi}{\partial z} \quad (12.259)$$

Et finalement le rotationnel en coordonnées cylindriques dans sa globalité est donné par:

$$\text{rot}(\vec{f}) = \begin{pmatrix} \frac{1}{r} \frac{\partial f_z}{\partial \phi} - \frac{\partial f_\phi}{\partial z} \\ \frac{\partial f_r}{\partial z} - \frac{\partial f_z}{\partial r} \\ \frac{1}{r} \frac{\partial (r f_\phi)}{\partial r} - \frac{1}{r} \frac{\partial f_r}{\partial \phi} \end{pmatrix} \quad (12.260)$$

Le lecteur pourrait aisément vérifier que ce résultat est simplement le gradient en coordonnées cylindriques appliqué au champ vectoriel $\vec{f}$.

Pour s'en persuader, montrons maintenant directement l'expression du rotationnel en coordonnées sphériques en montrant ceci via le produit vectoriel du gradient en coordonnées sphériques avec le champ vectoriel $\vec{f}$.

D'abord rappelons que nous avons obtenu pour le gradient en coordonnées sphériques:

$$\vec{\nabla} = \left(\frac{\partial}{\partial r}, \frac{1}{r} \frac{\partial}{\partial \theta}, \frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} \right) \quad (12.261)$$

Donc il vient:

$$\vec{\nabla} \times \vec{f} = \begin{pmatrix} \frac{\partial}{\partial r} \\ \frac{1}{r} \frac{\partial}{\partial \theta} \\ \frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} \end{pmatrix} \times \begin{pmatrix} f_r \\ f_\theta \\ f_\phi \end{pmatrix} \quad (12.262)$$

ce que nous pouvons aussi écrire en décomposant à l'aide des vecteurs de base:

$$\begin{aligned} \vec{\nabla} \times \vec{f} &= \left(\vec{e}_r \frac{\partial}{\partial r} + \frac{\vec{e}_\theta}{r} \frac{\partial}{\partial \theta} + \frac{\vec{e}_\phi}{r \sin \theta} \frac{\partial}{\partial \phi} \right) \times (f_r \vec{e}_r + f_\theta \vec{e}_\theta + f_\phi \vec{e}_\phi) \\ &= \left(\vec{e}_r \frac{\partial}{\partial r} + \frac{\vec{e}_\theta}{r} \frac{\partial}{\partial \theta} + \frac{\vec{e}_\phi}{r \sin \theta} \frac{\partial}{\partial \phi} \right) \times \vec{f} = \vec{e}_r \times \frac{\partial \vec{f}}{\partial r} + \frac{\vec{e}_\theta}{r} \times \frac{\partial \vec{f}}{\partial \theta} + \frac{\vec{e}_\phi}{r \sin \theta} \times \frac{\partial \vec{f}}{\partial \phi} \\ &= \vec{e}_r \times \left(\frac{\partial f_r}{\partial r} \vec{e}_r + \frac{\partial f_\theta}{\partial r} \vec{e}_\theta + \frac{\partial f_\phi}{\partial r} \vec{e}_\phi + f_r \frac{\partial \vec{e}_r}{\partial r} + f_\theta \frac{\partial \vec{e}_\theta}{\partial r} + f_\phi \frac{\partial \vec{e}_\phi}{\partial r} \right) \\ &\quad + \frac{\vec{e}_\theta}{r} \times \left(\frac{\partial f_r}{\partial \theta} \vec{e}_r + \frac{\partial f_\theta}{\partial \theta} \vec{e}_\theta + \frac{\partial f_\phi}{\partial \theta} \vec{e}_\phi + f_r \frac{\partial \vec{e}_r}{\partial \theta} + f_\theta \frac{\partial \vec{e}_\theta}{\partial \theta} + f_\phi \frac{\partial \vec{e}_\phi}{\partial \theta} \right) \\ &\quad + \frac{\vec{e}_\phi}{r \sin \theta} \times \left(\frac{\partial f_r}{\partial \phi} \vec{e}_r + \frac{\partial f_\theta}{\partial \phi} \vec{e}_\theta + \frac{\partial f_\phi}{\partial \phi} \vec{e}_\phi + f_r \frac{\partial \vec{e}_r}{\partial \phi} + f_\theta \frac{\partial \vec{e}_\theta}{\partial \phi} + f_\phi \frac{\partial \vec{e}_\phi}{\partial \phi} \right) \end{aligned} \quad (12.263)$$

A l'aide des dérivées partielles que nous avons démontrées lors de notre introduction plus haut du système de coordonnées sphériques il vient :

$$\begin{aligned} \vec{\nabla} \times \vec{f} &= \vec{e}_r \times \left(\frac{\partial f_\theta}{\partial r} \vec{e}_r + \frac{\partial f_\phi}{\partial r} \vec{e}_\phi + 0 + 0 + 0 \right) \\ &\quad + \frac{\vec{e}_\theta}{r} \times \left(\frac{\partial f_r}{\partial \theta} \vec{e}_r + \frac{\partial f_\theta}{\partial \theta} \vec{e}_\theta + \frac{\partial f_\phi}{\partial \theta} \vec{e}_\phi + f_r \vec{e}_\theta + f_\theta (-\vec{e}_r) + 0 \right) \\ &\quad + \frac{\vec{e}_\phi}{r \sin \theta} \times \left(\frac{\partial f_r}{\partial \phi} \vec{e}_r + \frac{\partial f_\theta}{\partial \phi} \vec{e}_\theta + \frac{\partial f_\phi}{\partial \phi} \vec{e}_\phi + f_r \sin \theta \vec{e}_\phi + f_\theta \cos \theta \vec{e}_\phi + f_\phi [-(\vec{e}_r \sin \theta + \vec{e}_\theta \cos \theta)] \right) \end{aligned} \quad (12.264)$$

Les produits vectoriels avec les vecteurs colinéaires s'annulent. Il reste donc:

$$\begin{aligned} \vec{\nabla} \times \vec{f} &= \vec{e}_r \times \left(\frac{\partial f_\theta}{\partial r} \vec{e}_r + \frac{\partial f_\phi}{\partial r} \vec{e}_\phi \right) + \frac{\vec{e}_\theta}{r} \times \left(\frac{\partial f_r}{\partial \theta} \vec{e}_r + \frac{\partial f_\phi}{\partial \theta} \vec{e}_\phi + f_\theta (-\vec{e}_r) \right) \\ &\quad + \frac{\vec{e}_\phi}{r \sin \theta} \times \left(\frac{\partial f_r}{\partial \phi} \vec{e}_r + \frac{\partial f_\theta}{\partial \phi} \vec{e}_\theta + f_\phi [-(\vec{e}_r \sin \theta + \vec{e}_\theta \cos \theta)] \right) \end{aligned} \quad (12.265)$$

Comme le produit vectoriel de deux vecteurs de base donne le vecteur orthogonal correspondant

(positivement ou négativement) nous avons alors:

$$\begin{aligned}\vec{\nabla} \times \vec{f} = & \left(\frac{\partial f_\theta}{\partial r} \vec{e}_\phi - \frac{\partial f_\phi}{\partial r} \vec{e}_\theta \right) + \left(-\frac{1}{r} \frac{\partial f_r}{\partial \theta} \vec{e}_\phi + \frac{1}{r} \frac{\partial f_\phi}{\partial \theta} \vec{e}_r - \frac{f_\theta}{r} \vec{e}_\phi \right) \\ & + \left(\frac{1}{r \sin \theta} \frac{\partial f_r}{\partial \phi} \vec{e}_\theta - \frac{1}{r \sin \theta} \frac{\partial f_\theta}{\partial \phi} \vec{e}_r - \frac{f_\phi}{r} \vec{e}_\theta + \frac{f_\theta \vec{e}_r \cos \theta}{r \sin \theta} \right)\end{aligned}\quad (12.266)$$

En regroupant les termes il vient:

$$\begin{aligned}\vec{\nabla} \times \vec{f} = & \vec{e}_r \left(\frac{1}{r} \frac{\partial f_\phi}{\partial \theta} - \frac{1}{r \sin \theta} \frac{\partial f_\theta}{\partial \phi} + \frac{f_\phi \cos \theta}{r \sin \theta} \right) + \vec{e}_\theta \left(-\frac{\partial f_\phi}{\partial r} + \frac{1}{r \sin \theta} \frac{\partial f_r}{\partial \phi} - \frac{f_\phi}{r} \right) \\ & + \vec{e}_\phi \left(\frac{\partial f_\theta}{\partial r} - \frac{1}{r} \frac{\partial f_r}{\partial \theta} - \frac{f_\theta}{r} \right)\end{aligned}\quad (12.267)$$

Soit en simplifiant:

$$\vec{\nabla} \times \vec{f} = \frac{\vec{e}_r}{r \sin \theta} \left(\frac{\partial}{\partial \theta} (\sin \theta f_\phi) - \frac{\partial f_\theta}{\partial \phi} \right) + \frac{\vec{e}_\theta}{r \sin \theta} \left(\frac{\partial f_r}{\partial \phi} - \sin \theta \frac{\partial}{\partial r} (r f_\phi) \right) + \frac{\vec{e}_\phi}{r} \left(\frac{\partial}{\partial r} (r f_\theta) - \frac{\partial f_r}{\partial \theta} \right) \quad (12.268)$$

Soit finalement:

$$\vec{\nabla} \times \vec{f} = \begin{pmatrix} \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (\sin \theta f_\phi) - \frac{1}{r \sin \theta} \frac{\partial f_\theta}{\partial \phi} \\ \frac{1}{r \sin \theta} \frac{\partial f_r}{\partial \phi} - \frac{1}{r} \frac{\partial}{\partial r} (r f_\phi) \\ \frac{1}{r} \frac{\partial}{\partial r} (r f_\theta) - \frac{\vec{e}_\theta}{r} \frac{\partial f_r}{\partial \theta} \end{pmatrix} \quad (12.269)$$

9.5. LAPLACIENS D'UN CHAMP SCALAIRE

Le laplacien d'un champ scalaire $\phi(x_1, x_2, x_3)$ est le champ scalaire qui mesure la différence entre la valeur de la fonction en un point et sa moyenne autour de ce point. En d'autres termes, la dérivée partielle deuxième mesure les variations de la pente au point étudié dans un entourage immédiat et selon une dimension à la fois. Si la dérivée partielle deuxième est nulle selon x , alors la pente est constante dans un entourage immédiat et selon cette dimension, cela implique que la valeur de la fonction au point étudié est la moyenne de son entourage (selon une dimension).

Cet opérateur s'obtient à partir de la divergence du gradient et nous la notons (écriture tensorielle):

$$\Delta \cdot f = \vec{\nabla}^2 f = \text{div}(\overrightarrow{\text{grad}}(f)) = \vec{\nabla} \circ (\vec{\nabla} \cdot \phi) = \sum_{i=1}^n \frac{\partial^2 f}{\partial x_i^2} \quad (12.270)$$

Le laplacien est nul, ou assez petit, lorsque la fonction varie sans à-coup. Les fonctions vérifiant l'équation de Laplace $\Delta \cdot \phi = 0$ sont appelées "**fonctions harmoniques**".

Donc l'opérateur "**laplacien en coordonnées cartésiennes**" est:

$$\vec{\nabla}^2 = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) \quad (12.271)$$

Le laplacien d'un champ scalaire et dans d'autres systèmes de coordonnées est un peu plus long à développer. Il existe plusieurs méthodes et parmi celles existantes j'ai choisi celles dont le type de raisonnement et les outils utilisés semblaient pertinents. Il est intéressant d'aborder différentes stratégies mais bien sûr il existe des méthodes plus simples que celle présentée ci-dessous.

Soit le laplacien en coordonnées cartésiennes dans $\mathbb{R}^2$ d'un champ scalaire f :

$$\vec{\nabla}^2 f = \left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \right) \quad (12.272)$$

Pour déterminer cette expression en coordonnées polaires, nous allons utiliser la différentielle totale et la règle de chaîne en coordonnées polaires:

$$df = \frac{\partial f}{\partial r} dr + \frac{\partial f}{\partial \phi} d\phi \Rightarrow \frac{df}{dx} = \frac{\partial f}{\partial r} \frac{dr}{dx} + \frac{\partial f}{\partial \phi} \frac{d\phi}{dx} \quad (12.273)$$

donc pour une dérivée seconde:

$$\begin{aligned} \frac{d^2 f}{dx^2} &= \frac{d}{dx} \left(\frac{\partial f}{\partial r} \frac{dr}{dx} + \frac{\partial f}{\partial \phi} \frac{d\phi}{dx} \right) = \frac{\partial}{\partial x} \left(\frac{\partial f}{\partial r} \right) \frac{dr}{dx} + \frac{\partial f}{\partial r} \frac{d^2 r}{dx^2} + \frac{\partial}{\partial x} \left(\frac{\partial f}{\partial \phi} \right) \frac{d\phi}{dx} + \frac{\partial f}{\partial \phi} \frac{d^2 \phi}{dx^2} \\ &= \left(\frac{\partial^2 f}{\partial r^2} \frac{\partial r}{\partial x} + \frac{\partial^2 f}{\partial r \partial \phi} \frac{\partial \phi}{\partial x} \right) \frac{dr}{dx} + \frac{\partial f}{\partial r} \frac{d^2 r}{dx^2} + \left(\frac{\partial^2 f}{\partial \phi \partial r} \frac{\partial r}{\partial x} + \frac{\partial^2 f}{\partial \phi^2} \frac{\partial \phi}{\partial x} \right) \frac{d\phi}{dx} + \frac{\partial f}{\partial \phi} \frac{d^2 \phi}{dx^2} \end{aligned} \quad (12.274)$$

or, nous avons pour les coordonnées polaires:

$$r = \sqrt{x^2 + y^2} \text{ et } \phi = \arctan\left(\frac{y}{x}\right) \quad (12.275)$$

d'où:

$$\begin{aligned} \frac{dr}{dx} &= \frac{x}{\sqrt{x^2 + y^2}} = \frac{x}{r} \text{ et } \frac{d\phi}{dx} = \frac{-y}{x^2 + y^2} = -\frac{y}{r^2} \\ \frac{d^2 r}{dx^2} &= \frac{r - \frac{dr}{dx} x}{r^2} = \frac{r - \left(\frac{x}{r}\right)x}{r^2} = \frac{y^2}{r^3} \text{ et } \frac{d^2 \phi}{dx^2} = \frac{d}{dx} \left(-\frac{y}{r^2} \right) = -y \frac{d}{dx} \left(\frac{1}{r^2} \right) = -y \left(-\frac{2x}{r^4} \right) = \frac{2xy}{r^4} \end{aligned} \quad (12.276)$$

d'où:

$$\begin{aligned} \frac{d^2 f}{dx^2} &= \frac{\partial^2 f}{\partial r^2} \left(\frac{x}{r} \right)^2 + \frac{\partial^2 f}{\partial r \partial \phi} \left(\frac{x}{r} \right) \left(-\frac{y}{r^2} \right) + \frac{\partial f}{\partial r} \left(\frac{y^2}{r^3} \right) + \frac{\partial^2 f}{\partial \phi \partial r} \left(\frac{x}{r} \right) \left(-\frac{y}{r^2} \right) \\ &+ \frac{\partial^2 f}{\partial \phi^2} \left(-\frac{y}{r^2} \right)^2 + \frac{\partial f}{\partial \phi} \left(\frac{2xy}{r^4} \right) \end{aligned} \quad (12.277)$$

et compte tenu que les dérivées partielles secondes sont continues, alors les dérivées croisées sont égales selon le théorème de Schwarz (cf. chapitre de Calcul Différentiel Et Intégral):

$$\frac{\partial^2 f}{\partial r \partial \phi} = \frac{\partial^2 f}{\partial \phi \partial r} \quad (12.278)$$

Donc:

$$\frac{d^2 f}{dx^2} = \frac{\partial^2 f}{\partial r^2} \left(\frac{x^2}{r^2} \right) + \frac{\partial^2 f}{\partial r \partial \phi} \left(-\frac{2xy}{r^3} \right) + \frac{\partial f}{\partial r} \left(\frac{y^2}{r^3} \right) + \frac{\partial^2 f}{\partial \phi^2} \left(\frac{y^2}{r^4} \right) + \frac{\partial f}{\partial \phi} \left(\frac{2xy}{r^4} \right) \quad (12.279)$$

De façon similaire, nous aurons:

$$\frac{d^2 f}{dy^2} = \frac{\partial^2 f}{\partial r^2} \left(\frac{y^2}{r^2} \right) + \frac{\partial^2 f}{\partial r \partial \phi} \left(\frac{2xy}{r^3} \right) + \frac{\partial f}{\partial r} \left(\frac{x^2}{r^3} \right) + \frac{\partial^2 f}{\partial \phi^2} \left(\frac{x^2}{r^4} \right) + \frac{\partial f}{\partial \phi} \left(-\frac{2xy}{r^4} \right) \quad (12.280)$$

d'où l'expression du laplacien en coordonnées polaires en sommant les deux dernières expressions:

$$\vec{\nabla}^2 f = \frac{\partial^2 f}{\partial r^2} + \frac{1}{r} \frac{\partial f}{\partial r} + \frac{1}{r^2} \frac{\partial^2 f}{\partial \phi^2} \quad (12.281)$$

Donc l'opérateur "**laplacien en coordonnées polaires**" est finalement donné par:

$$\boxed{\vec{\nabla}^2 = \left(\frac{\partial^2}{\partial r^2} + \frac{1}{r} \frac{\partial}{\partial r} + \frac{1}{r^2} \frac{\partial^2}{\partial \phi^2} \right)} \quad (12.282)$$

Pour trouver l'expression du laplacien en coordonnées sphériques, nous allons utiliser l'intuition du physicien et les notions de similitude.

Nous allons tout d'abord nous aider de la figure ci-dessus pour savoir de quoi l'on parle:

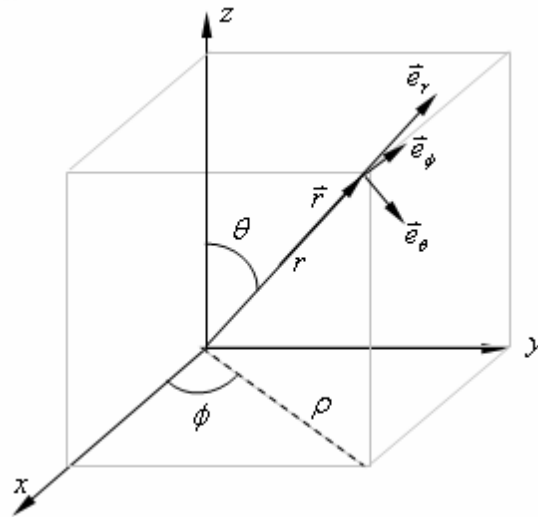


Figure: 12.14 - Représentation du système de coordonnées sphériques

Rappelons que les relations entre coordonnées cartésiennes et sphériques sont données par les relations:

$$\begin{aligned} r &= \sqrt{x^2 + y^2 + z^2} \\ \tan \phi &= y / x \\ \cos \theta &= \frac{z}{\sqrt{x^2 + y^2 + z^2}} \end{aligned} \quad (12.283)$$

Nous allons considérer maintenant les similitudes suivantes:

Coordonnées cylindriques: $x = \rho \cos \phi$ et $y = \rho \sin \phi$

Coordonnées sphériques: $z = r \cos \theta$ et $\rho = r \sin \theta$

Construisons un tableau de correspondance:

$$\begin{array}{cccc} \text{Cyl:} & x & \rho & \phi & y \\ & \updownarrow & \updownarrow & \updownarrow & \updownarrow \\ \text{Sph:} & z & r & \theta & \rho \end{array} \quad (12.284)$$

L'objectif est de jouer avec cette correspondance avec d'abord le laplacien en coordonnées cylindriques où l'on a soustrait des deux côtés de l'égalité le terme $\partial^2 f / \partial z^2$. Ainsi:

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \frac{\partial^2 f}{\partial \rho^2} + \frac{1}{\rho} \frac{\partial f}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2 f}{\partial \phi^2} \quad (12.285)$$

utilisons le tableau de correspondance et nous obtenons:

$$\frac{\partial^2 f}{\partial z^2} + \frac{\partial^2 f}{\partial \rho^2} = \frac{\partial^2 f}{\partial r^2} + \frac{1}{r} \frac{\partial f}{\partial r} + \frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2} \quad (12.286)$$

Le deuxième terme de l'égalité de cette dernière relation est l'équivalent sphérique du terme #1 du laplacien en coordonnées cylindriques:

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = \underbrace{\frac{\partial^2 f}{\partial \rho^2} + \frac{\partial^2 f}{\partial z^2}}_{\#1} + \frac{1}{\rho} \frac{\partial f}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2 f}{\partial \phi^2} \quad (12.287)$$

Maintenant examinons le terme: $\frac{1}{\rho} \frac{\partial f}{\partial \rho}$

Identiquement lorsque nous avons déterminé la relation:

$$df = \frac{\partial f}{\partial \rho} d\rho + \frac{\partial f}{\partial \phi} d\phi \Rightarrow \frac{df}{dx} = \frac{\partial f}{\partial \rho} \frac{d\rho}{dx} + \frac{\partial f}{\partial \phi} \frac{d\phi}{dx} \quad (12.288)$$

nous obtenons:

$$du = \frac{\partial u}{\partial \rho} d\rho + \frac{\partial u}{\partial \phi} d\phi \Rightarrow \frac{du}{dy} = \frac{\partial u}{\partial \rho} \frac{d\rho}{dy} + \frac{\partial u}{\partial \phi} \frac{d\phi}{dy} \quad (12.289)$$

avec:

$$\frac{d\rho}{dy} = \frac{y}{\sqrt{x^2 + y^2}} = \frac{y}{\rho} \text{ et } \frac{d\phi}{dy} = \frac{x}{x^2 + y^2} = \frac{x}{\rho^2} \quad (12.290)$$

ce qui nous permet d'écrire:

$$\frac{df}{dy} = \frac{y}{\rho} \frac{\partial f}{\partial \rho} + \frac{x}{\rho^2} \frac{\partial f}{\partial \phi} \quad (12.291)$$

si nous jouons encore avec le tableau de correspondance, nous avons:

$$\frac{df}{d\rho} = \frac{\rho}{r} \frac{\partial f}{\partial r} + \frac{z}{r^2} \frac{\partial f}{\partial \theta} \quad (12.292)$$

nous divisons cette relation des deux côtés par ρ et ainsi nous obtenons:

$$\frac{1}{\rho} \frac{df}{d\rho} = \frac{1}{r} \frac{\partial f}{\partial r} + \frac{z}{\rho r^2} \frac{\partial f}{\partial \theta} = \frac{1}{r} \frac{\partial f}{\partial r} + \frac{\cot \theta}{r^2} \frac{\partial f}{\partial \theta} \quad (12.293)$$

Nous avons donc ci-dessus l'équivalent sphérique du terme #2 du laplacien en coordonnées cylindriques:

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = \underbrace{\frac{\partial^2 f}{\partial \rho^2} + \frac{\partial^2 f}{\partial z^2}}_{\#1} + \underbrace{\frac{1}{\rho} \frac{\partial f}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2 f}{\partial \phi^2}}_{\#2} \quad (12.294)$$

Le troisième et dernier terme est très simple à déterminer. Nous remplaçons ρ par $r \sin \theta$ afin d'obtenir:

$$\frac{1}{\rho^2} \frac{\partial^2 f}{\partial \phi^2} \rightarrow \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 u}{\partial \phi^2} \quad (12.295)$$

En rassemblant tous les termes obtenus précédemment, nous obtenons enfin la forme étendue du laplacien en coordonnées sphériques si utilisé en physique:

$$\nabla^2 f = \frac{\partial^2 f}{\partial r^2} + \frac{1}{r} \frac{\partial f}{\partial r} + \frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2} + \frac{1}{r} \frac{\partial f}{\partial \theta} + \frac{\cot \theta}{r^2} \frac{\partial f}{\partial \theta} + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 f}{\partial \phi^2} \quad (12.296)$$

Nous pouvons raccourcir cette expression et factorisant les termes:

$$\tilde{\nabla}^2 f = \frac{\partial^2 f}{\partial r^2} + 2 \frac{1}{r} \frac{\partial f}{\partial r} + \frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2} + \frac{\cot \theta}{r^2} \frac{\partial f}{\partial \theta} + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 f}{\partial \phi^2} \quad (12.297)$$

Si nous condons encore un peu, nous obtenons l'expression finale de l'opérateur "**laplacien en coordonnées sphériques**" appelé aussi "**laplacien sphérique**":

$$\tilde{\nabla}^2 = \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2}{\partial \phi^2} \quad (12.298)$$

9.6. LAPLACIENS D'UN CHAMP VECTORIEL

Le laplacien d'un champ vectoriel $\vec{f}(x, y, z)$ est le champ vectoriel défini par (notation tensorielle):

$$\Delta \vec{f} = \begin{pmatrix} \sum_{i=1}^3 \frac{\partial^2 f_1}{\partial x_i^2} \\ \sum_{i=1}^3 \frac{\partial^2 f_2}{\partial x_i^2} \\ \sum_{i=1}^3 \frac{\partial^2 f_3}{\partial x_i^2} \end{pmatrix} \quad (12.299)$$

dont les composantes sont les laplaciens des composantes.

Ainsi, en coordonnées cartésiennes:

$$\Delta \vec{f} = \begin{pmatrix} \frac{\partial^2 f_1}{\partial x^2} + \frac{\partial^2 f_1}{\partial y^2} + \frac{\partial^2 f_1}{\partial z^2} \\ \frac{\partial^2 f_2}{\partial x^2} + \frac{\partial^2 f_2}{\partial y^2} + \frac{\partial^2 f_2}{\partial z^2} \\ \frac{\partial^2 f_3}{\partial x^2} + \frac{\partial^2 f_3}{\partial y^2} + \frac{\partial^2 f_3}{\partial z^2} \end{pmatrix} \quad (12.300)$$

Le laplacien d'un champ de vecteurs, appelé fréquemment "**laplacien vectoriel**", en d'autres systèmes de coordonnées est assez simple à obtenir à partir de la connaissance du laplacien d'un champ scalaire dans ces mêmes coordonnées. Ainsi, en coordonnées polaires, nous avons pour le laplacien d'un champ vectoriel la relation suivante:

$$\Delta \vec{u} = \begin{pmatrix} \frac{\partial^2 u_1}{\partial \rho^2} + \frac{1}{\rho} \frac{\partial u_1}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2 u_1}{\partial \phi^2} \\ \frac{\partial^2 u_2}{\partial \rho^2} + \frac{1}{\rho} \frac{\partial u_2}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2 u_2}{\partial \phi^2} \\ \frac{\partial^2 u_3}{\partial \rho^2} + \frac{1}{\rho} \frac{\partial u_3}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2 u_3}{\partial \phi^2} \end{pmatrix} \quad (12.301)$$

et en coordonnées cylindriques:

$$\Delta \vec{u} = \begin{pmatrix} \frac{\partial^2 u_1}{\partial \rho^2} + \frac{1}{\rho} \frac{\partial u_1}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2 u_1}{\partial \phi^2} + \frac{\partial^2 u_1}{\partial z^2} \\ \frac{\partial^2 u_2}{\partial \rho^2} + \frac{1}{\rho} \frac{\partial u_2}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2 u_2}{\partial \phi^2} + \frac{\partial^2 u_2}{\partial z^2} \\ \frac{\partial^2 u_3}{\partial \rho^2} + \frac{1}{\rho} \frac{\partial u_3}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2 u_3}{\partial \phi^2} + \frac{\partial^2 u_3}{\partial z^2} \end{pmatrix} \quad (12.302)$$

et finalement en coordonnées sphériques:

$$\Delta \vec{u} = \begin{pmatrix} \frac{\partial^2 u_1}{\partial r^2} + 2 \frac{1}{r} \frac{\partial u_1}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u_1}{\partial \theta^2} + \frac{\cot \phi}{r^2} \frac{\partial u_1}{\partial \theta} + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 u_1}{\partial \phi^2} \\ \frac{\partial^2 u_2}{\partial r^2} + 2 \frac{1}{r} \frac{\partial u_2}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u_2}{\partial \theta^2} + \frac{\cot \phi}{r^2} \frac{\partial u_2}{\partial \theta} + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 u_2}{\partial \phi^2} \\ \frac{\partial^2 u_3}{\partial r^2} + 2 \frac{1}{r} \frac{\partial u_3}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u_3}{\partial \theta^2} + \frac{\cot \phi}{r^2} \frac{\partial u_3}{\partial \theta} + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 u_3}{\partial \phi^2} \end{pmatrix} \quad (12.303)$$

9.7. IDENTITÉS

Les opérateurs différentiels scalaires et vectoriels ont des identités remarquables très simples que nous retrouverons très souvent en physique.

Voyons d'abord les relations qui n'ont aucun sens (au cas où vous tomberiez dessus sans faire exprès...):

$$\text{rot}(\text{div}(\vec{f})) \text{ ou } \vec{\nabla} \times (\vec{\nabla} \circ \vec{f}) \quad (12.304)$$

Le rotationnel d'une divergence n'existe pas puisque l'opérateur rotationnel s'applique à un champ vectoriel alors que la divergence est un scalaire.

$$\text{rot}(\Delta \cdot f) \text{ ou } \vec{\nabla} \times (\Delta \cdot f) \quad (12.305)$$

Le rotationnel d'un laplacien scalaire n'existe pas puisque l'opérateur rotationnel s'applique à un champ vectoriel alors que par construction, le laplacien est un scalaire.

Voyons maintenant quelques propriétés remarquables sans démonstrations (cependant si vous en avez besoin car vous n'y arrivez pas seul, n'hésitez pas à nous contacter, nous compléterons):

I1. Par construction le laplacien scalaire est la divergence du gradient du champ:

$$\operatorname{div}(\overrightarrow{\operatorname{grad}}(f)) = \vec{\nabla} \circ \vec{\nabla}(f) = \Delta \cdot f \quad (12.306)$$

I2. Le rotationnel du gradient est nul:

$$\operatorname{rot}[\overrightarrow{\operatorname{grad}}(f)] = \vec{\nabla} \times (\vec{\nabla} f) = \vec{0} \quad (12.307)$$

Donc si le rotationnel d'une variable vectorielle est nul, la variable peut être exprimée comme le gradient d'un potentiel scalaire! C'est une propriété très importante en électromagnétisme et en mécanique des fluides!

Démonstration:

$$\vec{\nabla} \times (\vec{\nabla} f) = \vec{\nabla} \times \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial y} \left(\frac{\partial f}{\partial z} \right) - \frac{\partial}{\partial z} \left(\frac{\partial f}{\partial y} \right) \\ \frac{\partial}{\partial z} \left(\frac{\partial f}{\partial x} \right) - \frac{\partial}{\partial x} \left(\frac{\partial f}{\partial z} \right) \\ \frac{\partial}{\partial x} \left(\frac{\partial f}{\partial y} \right) - \frac{\partial}{\partial y} \left(\frac{\partial f}{\partial x} \right) \end{pmatrix} = \vec{0} \quad (12.308)$$

□C.Q.F.D.

I3. La divergence du rotationnel d'un champ vectoriel est toujours nulle:

$$\operatorname{div}(\operatorname{rot}(\vec{f})) = 0 \text{ ou } \vec{\nabla} \circ (\vec{\nabla} \times \vec{f}) = 0 \quad (12.309)$$

Démonstration:

$$\begin{aligned} \vec{\nabla} \circ (\vec{\nabla} \times \vec{f}) &= \vec{\nabla} \circ \left(\begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \times \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} \right) = \vec{\nabla} \circ \begin{pmatrix} \partial_y f_3 - \partial_z f_2 \\ \partial_z f_1 - \partial_x f_3 \\ \partial_x f_2 - \partial_y f_1 \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \circ \begin{pmatrix} \partial_y f_3 - \partial_z f_2 \\ \partial_z f_1 - \partial_x f_3 \\ \partial_x f_2 - \partial_y f_1 \end{pmatrix} = \partial_x (\partial_y f_3 - \partial_z f_2) + \partial_y (\partial_z f_1 - \partial_x f_3) + \partial_z (\partial_x f_2 - \partial_y f_1) \\ &= \cancel{\partial_x \partial_y f_3} - \cancel{\partial_x \partial_z f_2} + \cancel{\partial_y \partial_z f_1} - \cancel{\partial_y \partial_x f_3} + \cancel{\partial_z \partial_x f_2} - \cancel{\partial_z \partial_y f_1} = 0 \end{aligned} \quad (12.310)$$

□C.Q.F.D.

I4. Le rotationnel du rotationnel d'un champ vectoriel est égal au gradient de la divergence de ce champ moins son laplacien vectoriel:

$$\operatorname{rot}(\operatorname{rot}(\vec{f})) = \overrightarrow{\operatorname{grad}}(\operatorname{div}(\vec{f})) - \Delta \vec{f} \text{ ou } \vec{\nabla} \times (\vec{\nabla} \times \vec{f}) = \vec{\nabla}(\vec{\nabla} \circ \vec{f}) - \Delta \vec{f} \quad (12.311)$$

Démonstration:

$$\begin{aligned}
\vec{\nabla} \times (\vec{\nabla} \times \vec{f}) &= \vec{\nabla} \times \left(\begin{pmatrix} \partial_x \\ \partial_y \\ \partial_z \end{pmatrix} \times \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} \right) = \vec{\nabla} \times \begin{pmatrix} \partial_y f_3 - \partial_z f_2 \\ \partial_z f_1 - \partial_x f_3 \\ \partial_x f_2 - \partial_y f_1 \end{pmatrix} = \begin{pmatrix} \partial_x \\ \partial_y \\ \partial_z \end{pmatrix} \times \begin{pmatrix} \partial_y f_3 - \partial_z f_2 \\ \partial_z f_1 - \partial_x f_3 \\ \partial_x f_2 - \partial_y f_1 \end{pmatrix} \\
&= \begin{pmatrix} \partial_y (\partial_x f_2 - \partial_y f_1) - \partial_z (\partial_z f_1 - \partial_x f_3) \\ \partial_z (\partial_y f_3 - \partial_z f_2) - \partial_x (\partial_x f_2 - \partial_y f_1) \\ \partial_x (\partial_z f_1 - \partial_x f_3) - \partial_y (\partial_y f_3 - \partial_z f_2) \end{pmatrix}
\end{aligned} \tag{12.312}$$

Il est ensuite facile de vérifier que cette dernière égalité est égale à:

$$\begin{aligned}
\vec{\nabla} (\vec{\nabla} \circ \vec{f}) - \Delta \vec{f} &= \vec{\nabla} (\partial_x f_1 + \partial_y f_2 + \partial_z f_3) - \Delta \vec{f} = \begin{pmatrix} \partial_x (\partial_x f_1 + \partial_y f_2 + \partial_z f_3) \\ \partial_y (\partial_x f_1 + \partial_y f_2 + \partial_z f_3) \\ \partial_z (\partial_x f_1 + \partial_y f_2 + \partial_z f_3) \end{pmatrix} - \Delta \vec{f} \\
&= \begin{pmatrix} \partial_x (\partial_x f_1 + \partial_y f_2 + \partial_z f_3) \\ \partial_y (\partial_x f_1 + \partial_y f_2 + \partial_z f_3) \\ \partial_z (\partial_x f_1 + \partial_y f_2 + \partial_z f_3) \end{pmatrix} - \begin{pmatrix} \frac{\partial^2 f_1}{\partial x^2} + \frac{\partial^2 f_1}{\partial y^2} + \frac{\partial^2 f_1}{\partial z^2} \\ \frac{\partial^2 f_2}{\partial x^2} + \frac{\partial^2 f_2}{\partial y^2} + \frac{\partial^2 f_2}{\partial z^2} \\ \frac{\partial^2 f_3}{\partial x^2} + \frac{\partial^2 f_3}{\partial y^2} + \frac{\partial^2 f_3}{\partial z^2} \end{pmatrix} \\
&= \begin{pmatrix} \partial_x (\partial_x f_1 + \partial_y f_2 + \partial_z f_3) - \frac{\partial^2 f_1}{\partial x^2} - \frac{\partial^2 f_1}{\partial y^2} - \frac{\partial^2 f_1}{\partial z^2} \\ \partial_y (\partial_x f_1 + \partial_y f_2 + \partial_z f_3) - \frac{\partial^2 f_2}{\partial x^2} - \frac{\partial^2 f_2}{\partial y^2} - \frac{\partial^2 f_2}{\partial z^2} \\ \partial_z (\partial_x f_1 + \partial_y f_2 + \partial_z f_3) - \frac{\partial^2 f_3}{\partial x^2} - \frac{\partial^2 f_3}{\partial y^2} - \frac{\partial^2 f_3}{\partial z^2} \end{pmatrix} \\
&= \begin{pmatrix} \partial_x (\partial_y f_2 + \partial_z f_3) - \frac{\partial^2 f_1}{\partial y^2} - \frac{\partial^2 f_1}{\partial z^2} \\ \partial_y (\partial_x f_1 + \partial_z f_3) - \frac{\partial^2 f_2}{\partial x^2} - \frac{\partial^2 f_2}{\partial z^2} \\ \partial_z (\partial_x f_1 + \partial_y f_2) - \frac{\partial^2 f_3}{\partial x^2} - \frac{\partial^2 f_3}{\partial y^2} \end{pmatrix} = \begin{pmatrix} \partial_y (\partial_x f_2 - \partial_y f_1) - \partial_z (\partial_z f_1 - \partial_x f_3) \\ \partial_z (\partial_y f_3 - \partial_z f_2) - \partial_x (\partial_x f_2 - \partial_y f_1) \\ \partial_x (\partial_z f_1 - \partial_x f_3) - \partial_y (\partial_y f_3 - \partial_z f_2) \end{pmatrix}
\end{aligned} \tag{12.313}$$

□C.Q.F.D.

I5. La multiplication de l'opérateur nabla par le produit scalaire de deux vecteurs est égale à... (voir ci-dessous), qui donne une relation très utile en mécanique des fluides:

$$\vec{\nabla} \cdot (\vec{x} \circ \vec{y}) = \vec{x} \times (\vec{\nabla} \times \vec{y}) + \vec{y} \times (\vec{\nabla} \times \vec{x}) + (\vec{y} \circ \vec{\nabla}) \cdot \vec{x} + (\vec{x} \circ \vec{\nabla}) \cdot \vec{y} \tag{12.314}$$

I6. Le produit scalaire du rotationnel d'un vecteur est la différence des opérateurs commutés tel que:

$$\vec{x} \circ (\vec{\nabla} \times \vec{y}) = \vec{y} \circ (\vec{\nabla} \times \vec{x}) - \vec{\nabla} \circ (\vec{x} \times \vec{y}) \quad (12.315)$$

Nous réutiliserons cette dernière relation lors de notre étude en électromagnétisme de la pression de radiation (entre autres).

9.8. RÉSUMÉ

Dans le cadre de ce site Internet, nous faisons usage des différentes notations présentées et résumées dans le tableau ci-dessous. Leur usage permet dans le cadre de différentes théories d'éviter des confusions avec d'autres être mathématiques. C'est embêtant certes mais il faudra faire avec.

ÊTRE MATHÉMATIQUE	NOTATIONS
Gradient d'un champ scalaire	$\vec{\nabla}(f) \quad \vec{\nabla} \cdot f \quad \vec{\nabla} f \quad \overrightarrow{\text{grad}}(f)$
Gradient d'un champ vectoriel	$\vec{\nabla}(\vec{f})$
Divergence d'un champ de vecteurs	$\vec{\nabla} \circ \vec{f}$
Rotationnel d'un champ de vecteurs	$\vec{\nabla} \times \vec{f} \quad \text{rot}(\vec{f})$
Laplacien d'un champ scalaire	$\Delta \cdot f \quad \vec{\nabla}^2 f \quad \vec{\nabla} \circ (\vec{\nabla} \cdot f)$
Laplacien d'un champ vectoriel	$\Delta \vec{f}$

Tableau: 12.1 - Résumé des opérateurs différentiels vectoriels

Et pour les pragmatiques voici un résumé des explications des opérateurs les plus importants en physique:

- Le gradient signifie "la pente" (exemple: le champ électrique est la pente du potentiel électrostatique).

Les différentes expressions de l'opérateur gradient (mises sous la forme de l'opérateur nabla) en coordonnées cartésiennes, polaires, cylindriques et sphériques sont les suivantes:

$$\vec{\nabla} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) \quad (12.316)$$

$$\vec{\nabla} = \left(\frac{\partial}{\partial r}, \frac{1}{r} \frac{\partial}{\partial \phi} \right) \quad (12.317)$$

$$\vec{\nabla} = \left(\frac{\partial}{\partial r}, \frac{1}{r} \frac{\partial}{\partial \phi}, \frac{\partial}{\partial z} \right) \quad (12.318)$$

$$\vec{\nabla} = \left(\frac{\partial}{\partial r}, \frac{1}{r} \frac{\partial}{\partial \theta}, \frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} \right) \quad (12.319)$$

- La divergence caractérise un flux de quelque chose qui vient de quelque part, d'une source, ou qui y va. Si la divergence n'est pas nulle, c'est qu'il y a concentration autour d'un point, donc la densité augmente (ou diminue, c'est selon le signe). Ça peut être la densité de charges électriques ou bien la masse volumique. D'où le fameux théorème qui dit que le flux (ce qui passe dans une surface) est égal à

l'intégrale de la divergence (ce qui reste).

Les différentes expressions de l'opérateur divergence (mises sous la forme de l'opérateur nabla) en coordonnées cartésiennes, polaires, cylindriques et sphériques sont les suivantes:

$$\vec{\nabla} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) \quad (12.320)$$

$$\vec{\nabla} = \left(\frac{1}{r} \frac{\partial}{\partial r}, \frac{1}{r} \frac{\partial}{\partial \phi} \right) \quad (12.321)$$

$$\vec{\nabla} = \left(\frac{1}{r} \frac{\partial}{\partial r}, \frac{1}{r} \frac{\partial}{\partial \phi}, \frac{\partial}{\partial z} \right) \quad (12.322)$$

$$\vec{\nabla} = \left(\frac{1}{r^2} \frac{\partial}{\partial r}, \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta}, \frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} \right) \quad (12.323)$$

- Le rotationnel caractérise l'existence d'un tourbillon (très utilisé en mécanique des fluides). S'il y a un tourbillon, on peut suivre une ligne de courant sur une courbe fermée sans qu'elle change de sens: la circulation ne sera pas nulle (elle vaut l'intégrale du rotationnel).

Les différentes expressions du rotationnel en coordonnées cartésiennes, cylindriques et sphériques sont les suivantes:

$$\text{rot}(\vec{f}) = \begin{pmatrix} \frac{\partial f_z}{\partial y} - \frac{\partial f_y}{\partial z} \\ \frac{\partial f_x}{\partial z} - \frac{\partial f_z}{\partial x} \\ \frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y} \end{pmatrix} \quad (12.324)$$

$$\text{rot}(\vec{f}) = \begin{pmatrix} \frac{1}{r} \frac{\partial f_z}{\partial \phi} - \frac{\partial f_\phi}{\partial z} \\ \frac{\partial f_r}{\partial z} - \frac{\partial f_z}{\partial r} \\ \frac{1}{r} \frac{\partial (rf_\phi)}{\partial r} - \frac{1}{r} \frac{\partial f_r}{\partial \phi} \end{pmatrix} \quad (12.325)$$

$$\vec{\nabla} \times \vec{f} = \begin{pmatrix} \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (\sin \theta f_\phi) - \frac{1}{r \sin \theta} \frac{\partial f_\theta}{\partial \phi} \\ \frac{1}{r \sin \theta} \frac{\partial f_r}{\partial \phi} - \frac{1}{r} \frac{\partial}{\partial r} (rf_\phi) \\ \frac{1}{r} \frac{\partial}{\partial r} (rf_\theta) - \frac{\partial f_r}{\partial \theta} \end{pmatrix} \quad (12.326)$$

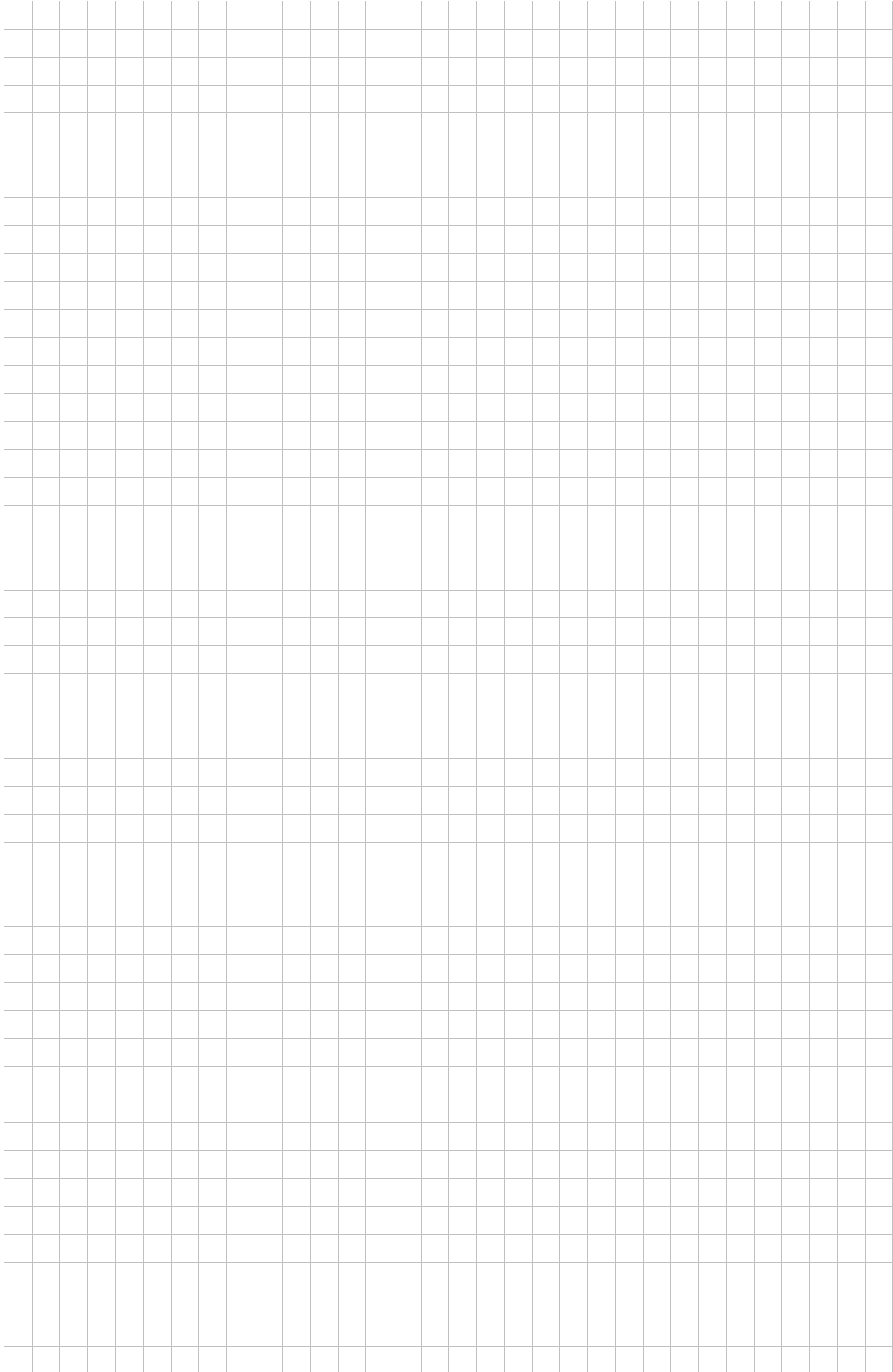
- Le laplacien d'un champ scalaire est le champ scalaire qui mesure la différence entre la valeur de la fonction en un point et sa moyenne autour de ce point. En d'autres termes, la dérivée partielle deuxième mesure les variations de la pente au point étudiée dans un entourage immédiat et selon une dimension à la fois. Si la dérivée partielle deuxième est nulle selon une direction, alors la pente est constante dans un entourage immédiat et selon cette dimension, cela implique que la valeur de la fonction au point étudié est la moyenne de son entourage (selon une dimension).

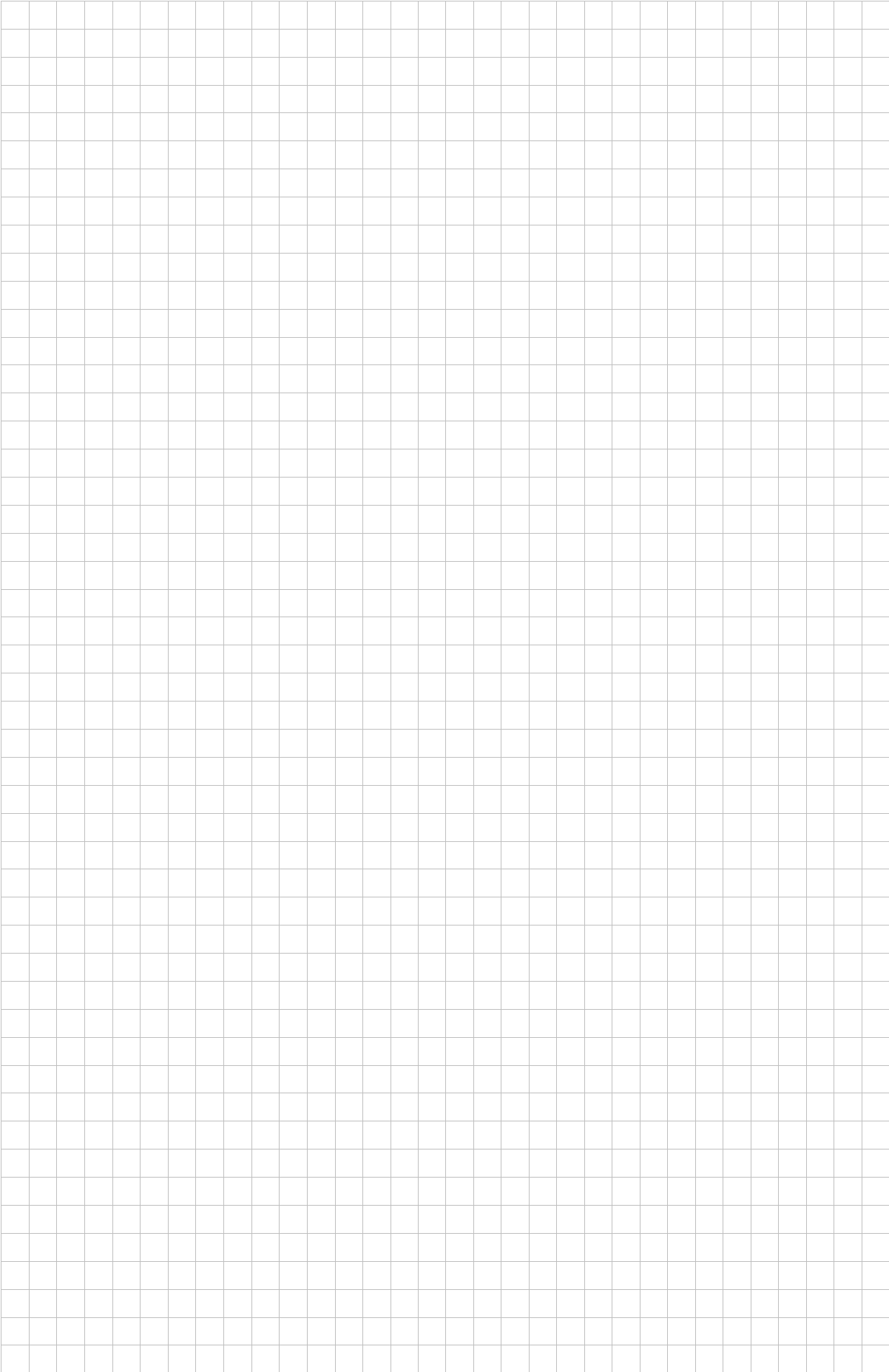
Les différentes expressions du laplacien (mises sous la forme de l'opérateur nabla) en coordonnées cartésiennes, polaires et sphériques sont les suivantes:

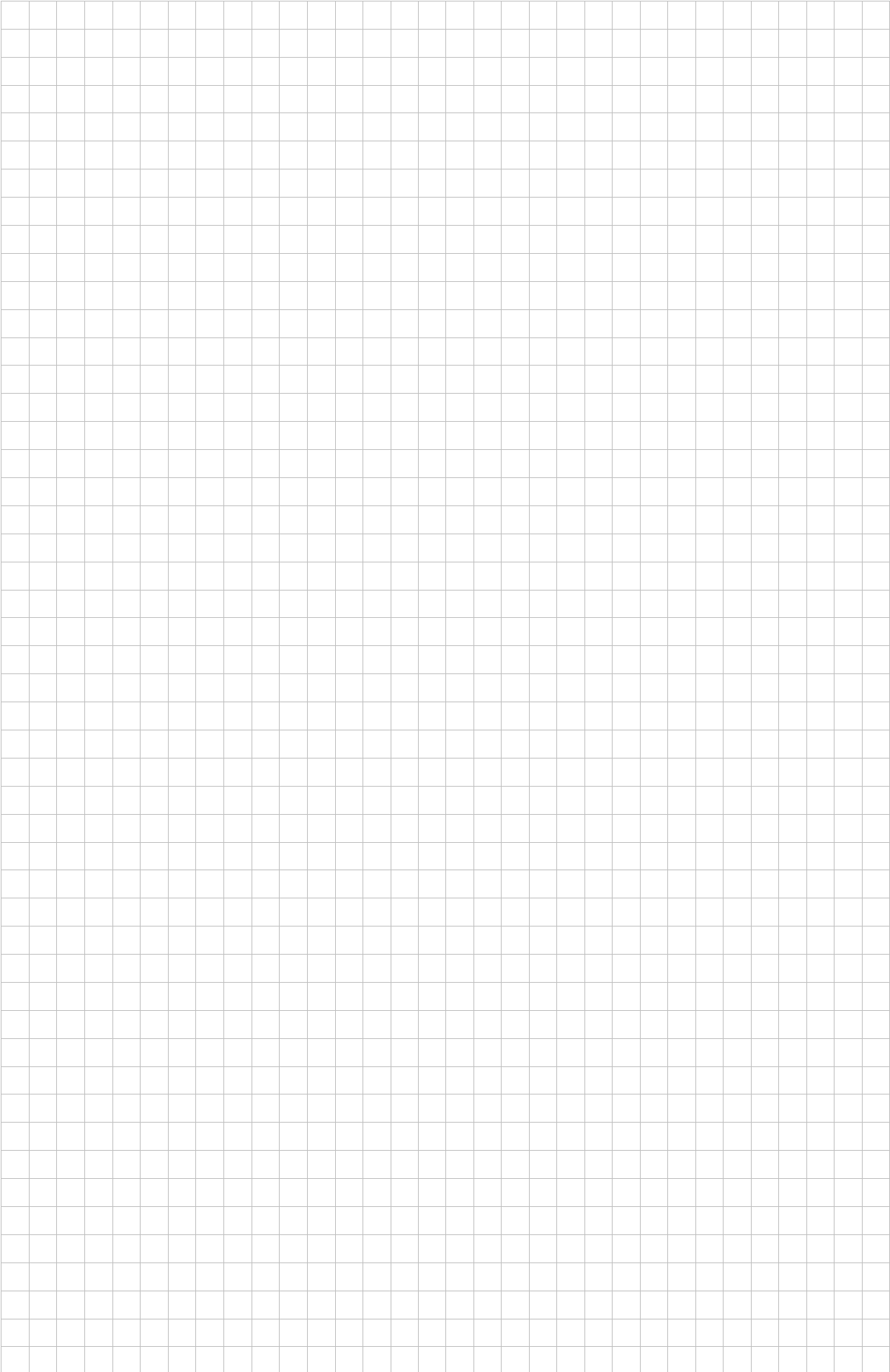
$$\bar{\nabla}^2 = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) \quad (12.327)$$

$$\bar{\nabla}^2 = \left(\frac{\partial^2}{\partial r^2} + \frac{1}{r} \frac{\partial}{\partial r} + \frac{1}{r^2} \frac{\partial^2}{\partial \phi^2} \right) \quad (12.328)$$

$$\bar{\nabla}^2 = \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2}{\partial \phi^2} \quad (12.329)$$







13. ALGÈBRE LINÉAIRE

Il y a plusieurs manières d'aborder l'algèbre linéaire. D'abord une manière pragmatique (nous commencerons par celle-ci car notre expérience nous a montré que c'est celle qui semblait le mieux marcher chez les étudiants) et une manière plus formelle que nous présenterons aussi après la première.

Ainsi, rappelons que nous avons étudié dans le chapitre de calcul algébrique comment déterminer l'intersection (si elle existe) de l'équation de deux droites (nous pouvons étendre le problème bien évidemment à plus de deux droites) dans $\mathbb{R}^2$ données par:

$$y_1 = a_1x + b_1 \text{ et } y_2 = a_2x + b_2 \quad (13.1)$$

où $a_i, b_i \in \mathbb{R}$.

En cherchant donc la valeur de x pour laquelle:

$$y_1 = y_2 \quad (13.2)$$

Ainsi nous pouvions écrire:

$$a_2x + b_2 = a_1x + b_1 \Rightarrow x = \frac{b_1 - b_2}{a_2 - a_1} \quad (13.3)$$

Cependant, il existe une autre manière de présenter le problème comme nous l'avons vu en méthodes numériques (section d'informatique théorique). Effectivement, nous pouvons écrire le problème sous la forme d'un bloc d'équations:

$$\begin{aligned} y_1 &= a_1x + b_1 \\ y_2 &= a_2x + b_2 \end{aligned} \quad (13.4)$$

et comme nous cherchons $y_1 = y_2 = y$, nous avons:

$$\begin{aligned} y - a_1x &= b_1 \\ y - a_2x &= b_2 \end{aligned} \quad (13.5)$$

Cette écriture s'appelle comme nous l'avons présenté dans le chapitre de Méthodes Numériques (section d'Informatique Théorique) un "**système linéaire**" que nous pouvons résoudre en soustrayant ou en additionnant les lignes entre elles (l'ensemble des solutions étant toujours égal), ce qui nous donne:

$$\begin{array}{rcl} y - a_1x & = & b_1 \\ + \quad y - a_2x & = & b_2 \quad | \cdot (-1) \\ \hline -a_1x + a_2x & = & b_1 - b_2 \end{array} \quad (13.6)$$

et nous voyons que nous retombons sur la solution:

$$x = \frac{b_1 - b_2}{a_2 - a_1} \quad (13.7)$$

il y donc deux manière de présenter un problème d'intersection de droites:

1. Sous forme d'équation
2. Sous forme de système

Nous allons nous intéresser dans une partie ce chapitre à la deuxième méthode qui va nous permettre à l'aide des outils vus dans le chapitre de calcul vectoriel de résoudre les intersections non plus d'une ou plusieurs droites mais d'une ou plusieurs droites et de plans, hyperplans dans respectivement $\mathbb{R}^2, \mathbb{R}^3, \dots, \mathbb{R}^n$.

Il y a cependant une condition à remplir: comme nous l'avons vu dans l'exemple précédent, nous ne pourrions pas résoudre un système d'équations à deux inconnues si nous n'avons qu'une seule équation. C'est la raison pour laquelle il faut et il suffit pour un système d'équations à n inconnues d'avoir au moins n équations. Ainsi, nous parlons de: "**systèmes de n équations à n inconnues**" et comme nous le verrons plus loin, ceci implique trivialement d'avoir une matrice carrée (le concept de "matrice" sera défini un peu plus loin). Nous démontrerons aussi que pour qu'un tel système ait des solutions non toutes nulles, il faut que nous ayons un déterminant de la matrice qui soit non nul (le concept de "déterminant" sera défini plus loin) et donc que la matrice soit inversible.

1. SYSTÈMES LINÉAIRES

Définition: Nous appelons donc "**système linéaire**", ou simplement "**système**", toute famille d'équations de la forme:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= b_2 \\ &\dots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n &= b_m \end{aligned} \quad (13.8)$$

où chaque ligne représente l'équation d'une droite, plan ou hyperplan (cf. chapitre de Géométrie Analytique) et a_{mn} les "**coefficients du système**", b_m les "**coefficients du second membre**" et les x_n les "**inconnues du système**".

Si les coefficients du second membre sont tous nuls, nous disons alors que le système est un "**système homogène**" et alors celui-ci admet au moins la solution triviale où les x_n sont tous nuls.

Nous appelons "**système homogène associé au système**", le système d'équations que nous obtenons en substituant des zéros aux coefficients du second membre.

Rappelons les éléments suivants:

- L'équation d'une droite (cf. chapitre d'Analyse Fonctionnelle) est donnée par:

$$a_1x_1 + a_2x_2 = b \quad (13.9)$$

en posant $x = x_1, y = x_2$.

- L'équation d'un plan (cf. chapitre de Géométrie Analytique) est donnée par:

$$a_1x_1 + a_2x_2 + a_3x_3 = b \quad (13.10)$$

en posant $x = x_1, y = x_2, z = x_3$.

- L'équation d'un hyperplan est très facilement (si vous ne voyez pas comment faites le nous savoir nous le préciserons) généralisable à partir de la démonstration de celle du plan et nous obtenons ainsi:

$$a_1x_1 + a_2x_2 + \dots + a_nx_n = b \quad (13.11)$$

en posant $x = x_1, y = x_2, z = x_3, \dots$

Nous écrivons souvent un système linéaire sous la forme condensée suivante:

$$\sum_{j=1}^n a_{ij}x_j = b_i, \quad i = 1, 2, \dots, m \quad (13.12)$$

Nous appelons "**solution du système**" tout n -uplet $(x_1^0, x_2^0, \dots, x_n^0)$ tel que:

$$\sum_{j=1}^n a_{ij}x_j^0 = b_i, \quad i = 1, 2, \dots, m \quad (13.13)$$

Résoudre un système signifie trouver l'ensemble des solutions de ce système. Deux systèmes à n inconnues sont dits "**systèmes équivalents**" si toute solution de l'un est solution de l'autre, autrement dit, s'ils admettent le même ensemble de solutions. Nous disons parfois que les équations d'un système sont des "**équations compatibles**" ou "**équations incompatibles**", suivant que ce système admet au moins une solution ou n'en admet aucune.

Nous pouvons également donner bien sûr une interprétation géométrique à ces systèmes. Supposons que les premiers membres des équations du système soient non nuls. Alors, nous savons que chacune de ces équations représente un hyperplan d'un espace affine (voir le chapitre de calcul vectoriel) de dimension n . Par conséquent, l'ensemble des solutions du système, regardé comme ensemble de n -uplets de coordonnées, représente une intersection finie d'hyperplans.

Exemple:

Le système d'équations suivant:

$$\begin{aligned} x_1 - 3x_3 &= -3 \\ 2x_1 - 5x_2 - x_3 &= -2 \\ x_1 + 2x_2 - 5x_3 &= 1 \end{aligned} \quad (13.14)$$

noté plus conventionnellement dans les petites classes sous la forme:

$$\begin{aligned}x - 3z &= -3 \\ 2x - 5y - z &= -2 \quad (13.15) \\ x + 2y - 5z &= 1\end{aligned}$$

Aurait comme solutions les points représentant l'intersection des trois plans définis par les trois équations. Mais comme nous pouvons le voir visuellement avec Maple à l'aide des commandes suivantes:

```
>with(plots);
>implicitplot3d({x-3*z=-3,2*x-5*y-z=-2,x+2*y-5*z=1},x=-3..3,y=-3..3,z=-3..3);
```

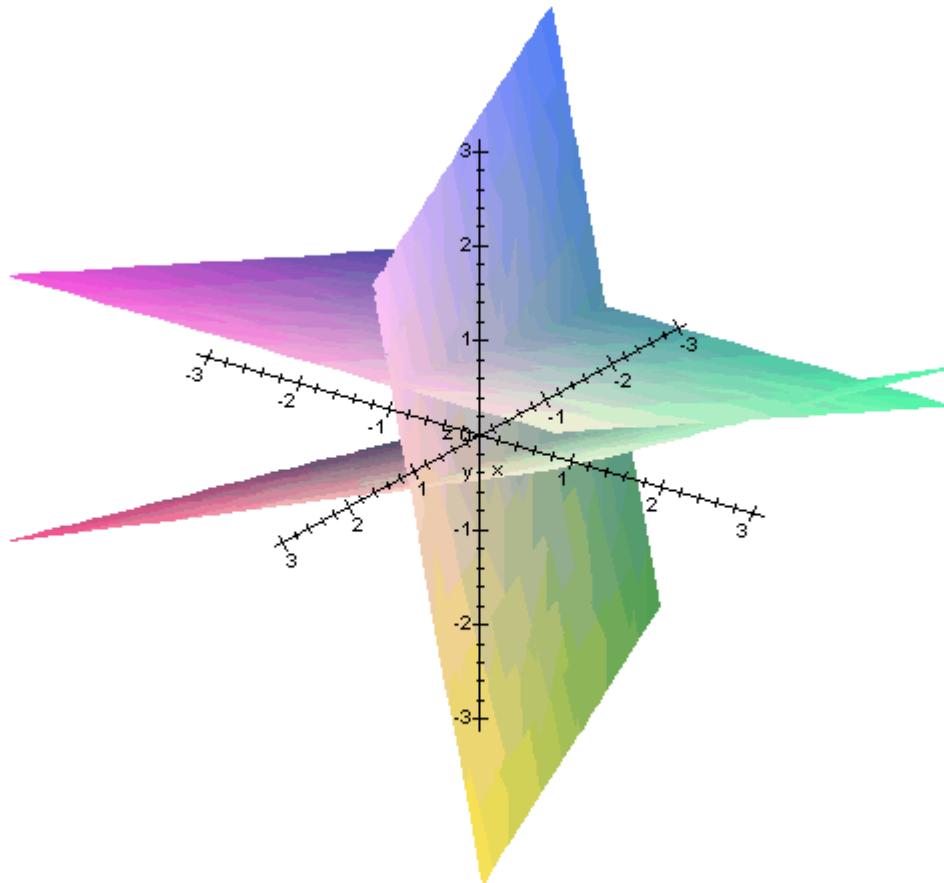


Figure: 13.1 - Représentation d'un système de 3 équations à 3 inconnues avec Maple

ce système n'a aucune solution. Ce qui peut soit se vérifier à la main, soit avec Maple en écrivant:

```
>solve({x-3*z=-3,2*x-5*y-z=-2,x+2*y-5*z=1},{x,y,z});
```

Remarque: Pour la méthode de résolution "classique" de ces systèmes, nous renvoyons le lecteur au chapitre traitant des méthodes numériques dans la section d'informatique.

C'était donc la manière pragmatique de voir les choses... passons maintenant à la seconde façon un peu plus ... mathématique (mais qui reste très simple):

2. TRANSFORMATIONS LINÉAIRES

Définition: Une "transformation linéaire" ou "application linéaire" A est une application d'un espace vectoriel E vers un espace vectoriel F telle que avec K étant $\mathbb{R}$ ou $\mathbb{C}$:

$$\begin{aligned} \forall \vec{x}_1, \vec{x}_2 \in E \quad A(\vec{x}_1 + \vec{x}_2) &= A(\vec{x}_1) + A(\vec{x}_2) \\ \forall \vec{x} \in E, \alpha \in K \quad A(\alpha \vec{x}) &= \alpha A(\vec{x}) \end{aligned} \quad (13.16)$$

plus fréquemment donné sous la forme (car l'application linéaire est souvent assimilée à une matrice):

$$\begin{aligned} \forall \vec{x}_1, \vec{x}_2 \in E \quad A(\vec{x}_1 + \vec{x}_2) &= A\vec{x}_1 + A\vec{x}_2 \\ \forall \vec{x} \in E, \alpha \in K \quad A\alpha\vec{x} &= \alpha A\vec{x} \end{aligned} \quad (13.17)$$

ceci constitue, pour rappel, un endomorphisme (cf. chapitre de Théorie Des Ensembles).

La première propriété spécifie que la transformée d'une somme de vecteurs doit être égale à la somme des transformées, pour qu'elle soit linéaire. La deuxième propriété précise que la transformée d'un vecteur auquel nous avons appliqué un facteur d'échelle (homothétie) doit aussi être égale à ce facteur appliqué sur la transformée du vecteur original. Si l'une ou l'autre de ces deux propriétés n'est pas respectée, la transformation n'est alors pas linéaire.

Nous allons maintenant montrer que toute transformation linéaire peut être représentée par une matrice:

Soient $(\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n)$ les vecteurs de base pour E et $(\vec{u}_1, \vec{u}_2, \dots, \vec{u}_m)$ ceux de F . Avec ces bases, nous pouvons représenter n'importe quels vecteurs $\vec{x} \in E, \vec{y} \in F$ avec les combinaisons linéaires suivantes (cf. chapitre de Calcul Vectoriel):

$$\vec{x} = \sum_{j=1}^n x_j \vec{v}_j \quad \vec{y} = \sum_{i=1}^m y_i \vec{u}_i \quad (13.18)$$

Soit la transformation linéaire A qui applique E sur F ($A: E \rightarrow F$). Donc $A(\vec{x}) = \vec{y}$ que nous pouvons réécrire de la façon suivante:

$$A\left(\sum_{j=1}^n x_j \vec{v}_j\right) = \sum_{i=1}^m y_i \vec{u}_i \quad (13.19)$$

mais puisque A est un opérateur linéaire par définition, nous pouvons aussi écrire:

$$\sum_{j=1}^n x_j A(\vec{v}_j) = \sum_{i=1}^m y_i \vec{u}_i \quad (13.20)$$

En considérant maintenant que les vecteurs $A(\vec{v}_j)$ sont des éléments de F , nous pouvons les réécrire en tant qu'une combinaison linéaire de ses vecteurs de base:

$$A(\vec{v}_j) = \sum_{i=1}^m a_{ij} \vec{u}_i \quad (13.21)$$

Ainsi, nous obtenons:

$$\sum_{j=1}^n x_j A(\vec{v}_j) = \sum_{j=1}^n x_j \sum_{i=1}^m a_{ij} \vec{u}_i = \sum_{i=1}^m y_i \vec{u}_i \quad (13.22)$$

En inversant l'ordre des sommations, nous pouvons écrire:

$$\sum_{i=1}^m \vec{u}_i \sum_{j=1}^n a_{ij} x_j = \sum_{i=1}^m y_i \vec{u}_i \quad (13.23)$$

et en réarrangeant cette dernière relation, nous produisons le résultat:

$$\sum_{i=1}^m \vec{u}_i \left(\sum_{j=1}^n a_{ij} x_j - y_i \right) = \vec{0} \quad (13.24)$$

Finalement, en se rappelant que les vecteurs de base $\vec{u}_i$ doivent être indépendants, nous pouvons conclure que leurs coefficients doivent forcément être nuls, donc:

$$\sum_{j=1}^n a_{ij} x_j = y_i \quad (13.25)$$

Ce qui correspond au produit de "**matrice**":

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \quad (13.26)$$

que nous pouvons noter:

$$A\vec{x} = \vec{y} \quad (13.27)$$

Autrement dit, toute transformation linéaire peut être décrite par une matrice A qu'il s'agit de multiplier avec le vecteur que nous voulons transformer, pour obtenir le vecteur résultant de la transformation.

3. MATRICES

Nous appelons donc "**matrice**" à m lignes et n colonnes, ou "**matrice de type mn** " (le premier terme correspond toujours aux lignes et le second aux colonnes, pour s'en souvenir il existe un bon moyen mnémotechnique: le président LinColn - abréviation de Ligne et Colonne...), tout tableau de nombres:

$$\begin{bmatrix} a_{11} & a_{12} \dots & a_{1j} \dots & a_{1n} \\ a_{21} & a_{22} \dots & a_{2j} \dots & a_{2n} \\ a_{i1} & a_{i2} \dots & a_{ij} \dots & a_{in} \\ a_{m1} & a_{m2} \dots & a_{mj} \dots & a_{mn} \end{bmatrix} \quad (13.28)$$

Nous désignons souvent une matrice de type $m \times n$ plus brièvement par:

$$(a_{ij})_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \quad (13.29)$$

ou simplement par (a_{ij}) .

Le nombre a_{ij} est appelé "**terme ou coefficient d'indices i, j** ". L'indice i étant appelé "**indice de ligne**" et l'indice j "**indice de colonne**".

Nous notons $M_{mn}(K)$ l'ensemble des matrices $m \times n$ dont les coefficients prennent leurs valeurs dans K (pouvant être $\mathbb{R}$ ou $\mathbb{C}$ par exemple).

Lorsque $m = n$, nous disons que (a_{ij}) est une "**matrice carrée**" d'ordre n . Dans ce cas, les termes $a_{11}, a_{22}, \dots, a_{nn}$ sont appelées "**termes diagonaux**".

Nous appelons également une matrice à une seule ligne "**matrice-ligne**" et une matrice à une seule colonne "**matrice-colonne**". Il est clair qu'une matrice colonne n'est rien d'autre qu'un "**vecteur-colonne**". Par la suite, les lignes d'une matrice seront assimilées à des matrices-lignes et les colonnes à des matrices-colonnes.

L'intérêt de la notion de matrice va apparaître tout au long des textes qui vont suivre mais la raison d'être immédiate de cette notion est simplement de permettre à certaines familles finies de nombres d'être conçues sous la forme d'un tableau rectangulaire.

Nous assignerons aux matrices des symboles propres, à savoir les lettres latines majuscules: $A, B, \dots$ et aux matrices-colonnes des symboles à savoir les lettres minuscules vectorielles $\vec{a}, \vec{b}, \dots$; nous les appellerons d'ailleurs indifféremment matrices-colonnes ou vecteurs-colonnes.

Nous appelons "**matrice nulle**", et nous la notons O , toute matrice dont chaque terme est nul. Les matrices-colonnes nulles sont également désignées par le symbole vectoriel: $\vec{0}$.

Nous appelons "**matrice unité d'ordre n** " ou "**matrice identité d'ordre n** ", et nous notons I_n , ou simplement I , la matrice carrée d'ordre n :

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & 0 \\ 0 & 0 & \dots & 1 \end{bmatrix} \quad (13.30)$$

Nous verrons plus loin que la matrice nulle joue le rôle d'élément neutre de l'addition matricielle et la matrice unité d'élément neutre de la multiplication matricielle.

Attention! Lorsque nous travaillons avec les matrices à coefficients complexes il faut toujours utiliser le terme "matrice identité" plutôt que "matrice unitaire" car dans le domaine des nombres complexes la matrice unitaire est un autre objet mathématique qu'il convient de ne pas confondre!

Nous allons maintenant revenir brièvement sur la définition de "**rang d'une famille finie**" que nous avons vue dans le chapitre de Calcul Vectoriel.

Rappel: Nous appelons "**rang**" d'une famille de vecteurs la dimension du sous-espace vectoriel de E qu'elle engendre.

Ainsi, soit $(\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n)$ les colonnes d'une matrice A , nous appelons "**rang de A** ", et nous notons $rg(A)$, le rang de la famille $(\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n)$.

Dans un langage un peu plus familier (...) le rang d'une matrice est donné par le nombre de matrices-colonnes qui ne peuvent s'exprimer par la combinaison et la multiplication par un scalaire d'autres matrices-colonnes de la même matrice.

Remarque: S'il y a des difficultés à déterminer le rang d'une matrice il existe une technique "d'échelonnage" des matrices que nous allons voir plus tard qui permet d'effectuer ce travail très rapidement.

Définition: Nous appelons "**matrice associée au système**":

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= b_2 \\ &\dots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n &= b_m \end{aligned} \quad (13.31)$$

l'objet mathématique défini par:

$$\begin{bmatrix} a_{11} & a_{12} \dots & a_{1j} \dots & a_{1n} \\ a_{21} & a_{22} \dots & a_{2j} \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{i1} & a_{i2} \dots & a_{ij} \dots & a_{in} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} \dots & a_{mj} \dots & a_{mn} \end{bmatrix} \quad (13.32)$$

c'est-à-dire la matrice A dont les termes sont les coefficients du système. Nous appelons "**matrice du second membre du système linéaire**", ou simplement "**second membre du système**", la matrice-colonne $\vec{b} = (b_i)$ dont les termes sont les coefficients du second membre de ce système. Nous appelons également "**matrice augmentée associée au système**" la matrice obtenue de A en ajoutant $\vec{b}$ comme $(n + 1)$ -ème colonne.

Si nous considérons maintenant un système de matrice associée A et de second membre $\vec{b}$. Désignons toujours par $(\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n)$ les colonnes de A . Le système s'écrit alors de manière équivalente sous la forme d'une équation vectorielle linéaire:

$$x_1\vec{a}_1 + x_2\vec{a}_2 + \dots + x_n\vec{a}_n = \vec{b} \quad (13.33)$$

Maintenant rappelons un théorème que nous avons vu en calcul vectoriel: pour que le rang d'une famille de vecteurs $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n)$ soit égal au rang de la famille augmentée $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n, \vec{y})$, il faut et il suffit que le vecteur $\vec{y}$ soit combinaison linéaire des vecteurs $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n$.

Il s'ensuit que notre système linéaire sous forme vectorielle admet au moins une solution $(x_1^0, x_2^0, \dots, x_n^0)$

si le rang de la famille $(\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n)$ est égal au rang de la famille augmentée $(\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n, \vec{b})$ et cette solution est unique si et seulement si le rang de la famille $(\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n)$ est n .

Ainsi, pour qu'un système linéaire de matrice associée A et de second membre $\vec{b}$ admette au moins une solution, il faut et il suffit que le rang de A soit égal au rang de la matrice augmentée $(A | \vec{b})$. Si cette condition est remplie, le système admet une seule solution si et seulement si le rang de A est égal au nombre d'inconnues autrement dit, les colonnes de A sont linéairement indépendantes.

Nous disons qu'une matrice est "**échelonnée**" si ses lignes satisfont aux deux conditions suivantes:

C1. Toute ligne nulle n'est suivie que de lignes nulles

C2. L'indice de colonne du premier terme non nul de toute ligne non nulle est supérieur à l'indice de colonne du premier terme non nul de la ligne qui la précède.

Une matrice échelonnée non nulle est donc de la forme:

$$\begin{bmatrix} \boxed{a_{1j_1}} & \dots & \dots & a_{1n} \\ 0 & \boxed{a_{2j_2}} & \dots & a_{2n} \\ 0 & 0 & \boxed{a_{rj_r}} & a_{rn} \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad (13.34)$$

où $j_1 < j_2 < \dots < j_r$ et $a_{1j_1}, a_{2j_2}, \dots, a_{rj_r}$ sont des termes non nuls. Bien entendu, les lignes nulles terminales peuvent manquer.

Remarque: Nous supposons relativement évident que les matrices nulles et les matrices unités sont échelonnées.

Les colonnes d'indice $j_1, j_2, \dots, j_r$ d'une matrice échelonnée sont clairement linéairement indépendantes. Envisagées comme des vecteurs-colonnes de $\mathbb{R}^r$, elles forment donc une base de cet espace vectoriel. En considérant les autres colonnes également comme des vecteurs-colonnes de $\mathbb{R}^r$, nous en déduisons qu'elles sont nécessairement combinaison linéaire de celles d'indice $j_1, j_2, \dots, j_r$ et donc que le rang de la matrice échelonnée est r .

Nous noterons que r est aussi le nombre de lignes non nulles de la matrice échelonnée et également le rang de la famille des lignes de cette matrice, puisque les lignes non nulles sont dès lors manifestement indépendantes.

Nous pouvons dès lors nous autoriser un certain nombre d'opérations élémentaires (supplémentaires) sur les lignes des matrices qui nous seront fort utiles et ce, sans changer son rang:

P1. Nous pouvons permuter les lignes.

Remarque: La matrice est juste une représentation graphique esthétique d'un système linéaire. Ainsi, permuter deux lignes ne change aucunement le système.

P2. Multiplier une ligne par un scalaire non nul

Remarque: Cela ne changeant en rien l'indépendance linéaire des vecteurs-lignes.

P3. Additionner à une ligne originelle, un multiple d'une autre

Remarque: La ligne originelle disparaîtra au profit de la nouvelle qui est indépendante de toutes les (anciennes) autres. Le système reste ainsi linéairement indépendant.

Toute matrice peut être transformée en matrice échelonnée par une suite finie d'opérations de type P1, P2, P3. C'est cette technique que nous utilisons dans le chapitre traitant des algorithmes pour résoudre les systèmes linéaires.

Il est donc évident que les opérations élémentaires sur les lignes d'une matrice ne modifient pas le rang de la famille des lignes de cette matrice. Or, nous avons observé que le rang de la famille des lignes d'une matrice échelonnée est égal au rang de la famille des colonnes, c'est-à-dire au rang de cette matrice. Nous en concluons que le rang de n'importe quelle matrice de type $m \times n$ est également le rang de la famille des lignes de cette matrice.

Comme corollaire de cette conclusion, il apparaît que:

$$rg(A) \leq \min(m, n) \quad (13.35)$$

Lors de la résolution de systèmes linéaires de m équations à n inconnues il apparaît, comme nous l'avons déjà fait remarquer tout au début de ce chapitre, qu'il doit y avoir au moins un nombre égal d'équations que d'inconnues ou plus rigoureusement: le nombre d'inconnues doit être inférieur ou égal au nombre d'équations tel que:

$$n = rg(A|b) \leq m \quad (13.36)$$

3.1. OPÉRATIONS SUR LES MATRICES

Rappelons que nous avons vu lors de notre étude du calcul vectoriel que les opérations de multiplication d'un vecteur par un scalaire, d'addition ou soustraction de vecteurs entre eux et l'opération de produit scalaire formaient dans le sens ensembliste du terme un "espace vectoriel" (voir le chapitre de théorie des ensembles) possédant ainsi aussi une "structure algébrique vectorielle". Ceci sous la condition que les vecteurs aient bien sûr les mêmes dimensions (ce constat n'étant pas valable si au lieu du produit scalaire nous prenions le produit vectoriel).

Au même titre que les vecteurs, nous pouvons multiplier une matrice par un scalaire et additionner celles-ci entre elles (tant qu'elles ont les mêmes dimensions...) mais en plus, nous pouvons aussi multiplier deux matrices entre elles sous certaines conditions que nous définirons ci-après. Cela fera également de l'ensemble des matrices dans le sens ensembliste du terme, un espace vectoriel sur le corps K et possédant ainsi aussi une "structure algébrique vectorielle".

Ainsi, un vecteur pourra aussi être vu comme une matrice particulière de dimension $m \times n = m \times 1$ et opérer dans l'espace vectoriel des matrices. En gros..., le calcul vectoriel n'est qu'un cas particulier de l'algèbre linéaire.

Définitions:

D1. Soient $A, B \in M_{mn}(\mathbb{R})$. Nous appelons "somme de A et B" la matrice $C \in M_{mn}(\mathbb{R})$ dont les coefficients sont:

$$c_{ij} = a_{ij} + b_{ij} \quad 1 \leq i \leq m, 1 \leq j \leq n \quad (13.37)$$

D2. Soient $A \in M_{mn}(\mathbb{R})$ une matrice et $\lambda \in \mathbb{R}$ un scalaire. Nous appelons "produit de A par λ " la matrice $C \in M_{mn}(\mathbb{R})$ dont les coefficients sont:

$$c_{ij} = \lambda a_{ij} \quad 1 \leq i \leq m, 1 \leq j \leq n \quad (13.38)$$

De ses deux définitions nous pouvons donc effectivement conclure que l'espace/ensemble des matrices est bien un espace vectoriel et possède ainsi une structure algébrique vectorielle.

D3. Soient E, F, G trois espaces vectoriels de bases respectives $\mathcal{E}, \mathcal{F}, \mathcal{G}$ et $f \in L(E, F), g \in L(F, G)$ deux applications linéaires (voir théorie des ensembles aussi pour un rappel).

Notons A la matrice de f relativement aux bases $\mathcal{E}, \mathcal{F}$ et B la matrice de g relativement aux bases $\mathcal{F}, \mathcal{G}$. Alors la matrice C de $g \circ f$ (voir la définition d'une fonction composée dans le chapitre d'analyse fonctionnelle) relativement aux bases $\mathcal{E}, \mathcal{G}$ est égale au produit de B par A noté BA .

$$E \xrightarrow{f} F \xrightarrow{g} G \text{ et } E \xrightarrow{g \circ f} G \quad (13.39)$$

$$E \xrightarrow{A} F \xrightarrow{B} G \text{ et } E \xrightarrow{BA} G \quad (13.40)$$

Donc soient $B \in M_{np}(\mathbb{R})$ et $A \in M_{mq}(\mathbb{R})$, nous appelons "produit matriciel" ou "multiplication matricielle" de B par A et nous notons BA , la matrice $C \in M_{mq}(\mathbb{R})$ dont les coefficients sont:

$$c_{ki} = \sum_{j=1}^n b_{kj} a_{ji} \quad 1 \leq i \leq p, 1 \leq k \leq m \quad (13.41)$$

Il est important de remarquer que contrairement à l'addition, A et B peuvent avoir des dimensions différentes. Toutefois! le nombre de lignes de A doit être égal au nombre de colonnes de B , comme l'indique l'indice n des deux matrices. Donc dans le produit BA , si B est une matrice $m \times n$, A doit être une matrice $n \times p$, quel que soit p .

En notant par des lettres latines majuscules les matrices et par les lettres grecques minuscules les scalaires, le lecteur vérifiera aisément (nous pouvons rajouter les démonstrations sur demande) les relations:

$$\begin{aligned} A(B+C) &= AB+AC \\ (A+B)C &= AC+BC \\ (\lambda A)B &= A(\lambda B) \\ (AB)C &= A(BC) \\ AB &\neq BA \end{aligned} \quad (13.42)$$

Il est surtout important de se rappeler de la dernière ligne comme quoi la multiplication matricielle n'est pas commutative.

Remarque: L'ensemble $M_{nn}(\mathbb{R})$ des matrices carrées d'ordre n à coefficients dans $\mathbb{R}$ muni de la somme et la multiplication usuelles des matrices forme un anneau (voir chapitre de théorie des ensembles). C'est vrai plus généralement si les coefficients des matrices sont pris dans un anneau quelconque: par exemple l'ensemble $M_{nn}(\mathbb{Z})$ des matrices à coefficients entiers est un anneau.

3.2. TYPE DE MATRICES

Afin de simplifier les notations et la longueur des calculs nous allons introduire ici les matrices types que le lecteur pourra rencontrer tout au long de sa lecture du site (et pas que dans la partie de mathématiques pures!).

Définitions:

D1. Soit A une matrice carrée (c'est-à-dire $A \in M_{nn}(\mathbb{R})$). La matrice A est dite "**matrice inversible**" ou "**matrice régulière**" si et seulement si A^{-1} est telle que:

$$AA^{-1} = I_n = \mathbb{1} \quad (13.43)$$

où:

$$I_n = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix} \quad (13.44)$$

si tel n'est pas le cas, nous disons que A est une "**matrice singulière**".

Cette définition est fondamentale, elle a des conséquences extrêmement importantes dans toute l'algèbre linéaire et aussi dans la physique (résolution de système linéaires, déterminant, vecteurs et valeurs propres, etc.) il convient donc de s'en souvenir.

D2. Soit:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \quad (13.45)$$

une matrice de $M_{mn}(\mathbb{R})$. Nous appelons "**matrice transposée**" de A , la matrice notée A^T (le T en exposant est selon les ouvrages en majuscule ou en minuscule), de $M_{nn}(\mathbb{R})$ définie par (nous transposons les lignes en colonnes et les colonnes en lignes):

$$A^T = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \dots & \dots & \dots & \dots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{bmatrix} \quad (13.46)$$

Voici quelques propriétés intéressantes (nous seront par ailleurs utiles plus tard lors d'un théorème fameux!) de la transposée:

$$\begin{aligned} (A^T)^T &= A \\ (\lambda A + B)^T &= \lambda A^T + B^T, \lambda \in \mathbb{R} \\ (AB)^T &= B^T A^T \\ (A^{-1})^T &= (A^T)^{-1}, \exists A^{-1} \end{aligned} \quad (13.47)$$

et aussi une propriété importante de la matrice transposée (la vérification se fait aussi par l'exemple):

$$A\vec{X} \circ \vec{Y} = \vec{X} \circ A^T \vec{Y} \quad (13.48)$$

La matrice transposée est très importante en physique et en mathématique dans le cadre de la théorie des groupes et symétries! Il convient donc aussi de se souvenir de sa définition.

D3. Soit:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \quad (13.49)$$

une matrice de $M_{mn}(\mathbb{C})$. Nous appelons "**matrice adjointe**" de A, la matrice, de $M_{mn}(\mathbb{C})$ définie par:

$$\overline{A^T} = \begin{bmatrix} \bar{a}_{11} & \bar{a}_{21} & \dots & \bar{a}_{m1} \\ \bar{a}_{12} & \bar{a}_{22} & \dots & \bar{a}_{m2} \\ \dots & \dots & \dots & \dots \\ \bar{a}_{1n} & \bar{a}_{2n} & \dots & \bar{a}_{mn} \end{bmatrix} \quad (13.50)$$

qui est donc la complexe conjuguée de la matrice transposée ou si vous préférez... la transposée de la matrice conjuguée A (dans le cas de coefficient réels... on se passera de la conjuguer!). Pour simplifier les écritures nous la notons simplement $A^\dagger$ (écriture fréquente en physique quantique et algèbre ensembliste).

Remarque: Relation triviale (qui sera souvent utilisée en physique quantique des champs):

$$(AB)^\dagger = B^\dagger A^\dagger \quad (13.51)$$

D4. Par définition, une matrice est dite "**matrice hermitique**" ou "**matrice hermitienne**" ou "**matrice self-adjointe**" ou encore "**matrice autoadjointe**"... si elle est égale à son adjointe (matrice transposée conjuguée) tel que:

$$A = A^\dagger \quad (13.52)$$

D5. Soit A une matrice carrée de $M_{nn}(\mathbb{R})$, la "**trace**" de A , notée $tr(A)$ est définie par:

$$tr(A) = \sum_{i=1}^n a_{ii} \quad (13.53)$$

Quelques relations utiles y relatives (dont nous pouvons rajouter les démonstrations détaillées sur demande):

$$\begin{aligned} tr(\lambda A + B) &= \lambda tr(A) + tr(B) \\ tr(AB) &= tr(BA), \forall A \in M_{nn}(\mathbb{R}), \forall B \in M_{nn}(\mathbb{R}) \end{aligned} \quad (13.54)$$

D6. Une matrice A est dite "**matrice nilpotente**" si en la multipliant successivement par elle-même elle peut donner zéro. En clair, s'il existe un entier k tel que:

$$A^k = 0 \quad (13.55)$$

Remarque: Pour se souvenir de ce mot, nous le décomposons en "nil" pour nulle et "potent" pour potentiel. Ainsi, quelque chose de nilpotent est donc quelque chose qui est potentiellement nul.

D7. Une matrice A est dite "**matrice orthogonale**" si ses éléments sont réels et si elle obéit à:

$$A^{-1} = A^T \quad (13.56)$$

ce qui se traduit par (où δ_{ij} est le symbole de Kronecker):

$$A^T A = \mathbb{1} = [\delta_{ij}] \quad (13.57)$$

Les vecteurs colonnes de la matrice sont donc normés à l'unité et orthogonaux entre eux (ou de même avec ses lignes!). Ainsi une matrice orthogonale représente une base orthonormée!

Remarques:

R1. C'est typiquement le cas de la matrice de la base canonique, ou de toute matrice diagonalisable.

R2. Si au lieu de prendre simplement une matrice avec des coefficients réels, nous prenons une matrice à coefficients complexes avec sa transposée complexe (matrice adjointe). Alors, nous disons que A est une "**matrice unitaire**" si elle satisfait à la relation ci-dessus!

Nous reviendrons plus tard, après avoir présenté les concepts de vecteurs et valeurs propres, sur un cas particulier et très important de matrices orthogonales (appelées "matrices de translations").

Signalons encore une autre propriété importante en géométrie, physique et statistiques des matrices orthogonales.

Soit $f(\vec{x}) = A\vec{x} + \vec{b}$, où A est une matrice orthogonale et $\vec{b} \in \mathbb{R}^n$. Alors f (respectivement A) est une isométrie. C'est-à-dire que:

$$\|f(\vec{x}) - f(\vec{y})\| = \|\vec{x} - \vec{y}\| \quad \forall \vec{x}, \vec{y} \in \mathbb{R}^n \quad (13.58)$$

Démonstration:

$$\begin{aligned} \|f(\vec{x}) - f(\vec{y})\|^2 &= (f(\vec{x}) - f(\vec{y})) \circ (f(\vec{x}) - f(\vec{y})) = (A\vec{x} - A\vec{y}) \circ (A\vec{x} - A\vec{y}) \\ &= A(\vec{x} - \vec{y}) \circ A(\vec{x} - \vec{y}) = (AA^T (\vec{x} - \vec{y})) \circ (\vec{x} - \vec{y}) = (\vec{x} - \vec{y}) \circ (\vec{x} - \vec{y}) \end{aligned} \quad (13.59)$$

et donc nous avons bien:

$$\|f(\vec{x}) - f(\vec{y})\| = \|\vec{x} - \vec{y}\| \quad (13.60)$$

Donc en d'autres termes: Les matrices orthogonales sont des applications linéaires qui conservent la norme (les distances).

□C.Q.F.D.

D8. Soit $A \in M_{\text{mm}}(\mathbb{R})$ une matrice carrée. La matrice A est dite "**matrice symétrique**" si et seulement si:

$$A^T = A \Leftrightarrow a_{ij} = a_{ji} \quad (13.61)$$

Nous retrouverons cette définition en calcul tensoriel.

D9. Soit $A \in M_{\text{mm}}(\mathbb{R})$ une matrice carrée. La matrice A est dite "**matrice anti-symétrique**" si et seulement si:

$$a_{ij} = -a_{ji} \quad (13.62)$$

ce qui impose que:

$$a_{11} = a_{22} = \dots = a_{\text{mm}} = 0 \quad (13.63)$$

Nous retrouverons cette définition dans le chapitre de Calcul Tensoriel.

D10. Soit $A \in M_{\text{mm}}(\mathbb{R})$ une matrice carrée. La matrice A est dite "**matrice triangulaire supérieure**" si et seulement si:

$$\forall i, j \in \{1, \dots, m\}, (j < i \Rightarrow a_{ij} = 0)$$

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \dots & a_{2n} \\ 0 & 0 & a_{33} & \dots \\ 0 & 0 & 0 & a_{nn} \end{bmatrix}, n = m \quad (13.64)$$

D11. Soit $A \in M_{mm}(\mathbb{R})$ une matrice carrée. La matrice A est dite "**matrice triangulaire inférieure**" si et seulement si:

$$\forall i, j \in \{1, \dots, m\}, (i < j \Rightarrow a_{ij} = 0)$$

$$A = \begin{bmatrix} a_{11} & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 \\ \dots & \dots & a_{33} & 0 \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}, n = m \quad (13.65)$$

D12. Soit $D \in M_{mm}(\mathbb{R})$, une matrice carrée. La matrice D est dite "**matrice diagonale**" si et seulement si:

$$\forall i, j \in \{1, \dots, m\}, (i \neq j \Rightarrow d_{ij} = 0) \quad (13.66)$$

La notation habituelle d'une matrice diagonale D étant:

$$D = \text{diag}(d_1, d_2, \dots, d_n) \quad d_1 = d_{11}, d_2 = d_{22}, \dots, d_n = d_{nn} \quad (13.67)$$

D13. Soient E un espace vectoriel, de dimension n et deux bases $\mathcal{B}, \mathcal{B}'$ de E :

$$\mathcal{B}(\vec{e}_1, \vec{e}_2, \dots, \vec{e}_n), \mathcal{B}'(\vec{e}'_1, \vec{e}'_2, \dots, \vec{e}'_n) \quad (13.68)$$

Nous appelons "**matrice de passage**" de la base $\mathcal{B}$ à la base $\mathcal{B}'$, et nous noterons P la matrice de $M_{nn}(K)$ dont les colonnes sont formées des composantes des vecteurs de $\mathcal{B}'$ sur la base $\mathcal{B}$ (voir plus loin traitement détaillé des changements de base pour plus d'infos).

Nous considérons le vecteur $\vec{x}(x_1, x_2, \dots, x_n)$ de E qui s'écrit dans les bases $\mathcal{B}(\vec{e}_1, \vec{e}_2, \dots, \vec{e}_n)$ et $\mathcal{B}'(\vec{e}'_1, \vec{e}'_2, \dots, \vec{e}'_n)$ suivant les relations:

$$\mathcal{B}: \vec{x} = \sum_{i=1}^n x_i \vec{e}_i, \mathcal{B}': \vec{x} = \sum_{i=1}^n x'_i \vec{e}'_i \quad (13.69)$$

Soit:

$$\vec{X} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}, \vec{X}' = \begin{pmatrix} x'_1 \\ x'_2 \\ \dots \\ x'_n \end{pmatrix} \quad (13.70)$$

le vecteur de $\mathcal{K}^*$ formé des composantes de $\vec{x}$ dans la base $\mathcal{B}$ et respectivement le vecteur formé des composantes de $\vec{x}$ dans la base $\mathcal{B}'$. Alors:

$$X' = P^{-1}X \quad (13.71)$$

relation pour laquelle la démonstration détaillée sera donnée plus loin lors de notre étude des changements de base. Nous avons également:

$$X = PX' \quad (13.72)$$

Remarques:

R1. Si lorsqu'un vecteur est donné et que sa base n'est pas spécifiée, c'est qu'il s'agit dès lors implicitement de la base canonique:

$$(\vec{e}_1, \vec{e}_2, \vec{e}_3) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (13.73)$$

qui laisse invariant la multiplication par un vecteur quelconque et lorsque la base utilisée est notée $(\vec{e}_i)$ et n'est pas spécifiée, c'est qu'il s'agit également de la base canonique.

R2. Si un vecteur est donné par rapport à la base canonique, ces composantes sont dites "**covariantes**", dans le cas contraire, où si elles sont exprimées après suite dans une autre base non canonique, alors nous disons que les composantes sont "**contravariantes**" (pour plus de précisions sur le sujet voir le chapitre de calcul tensoriel).

3.3. DÉTERMINANTS

Nous allons nous intéresser aux déterminants dans le point de vue du physicien (celui de mathématicien étant assez rébarbatif...). En physique (que ce soit en mécanique classique ou physique quantique des champs), en chimie ou en ingénierie, nous aurons fréquemment des systèmes linéaires à résoudre. Or, nous avons vu maintenant qu'un système linéaire:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= b_2 \\ &\dots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n &= b_m \end{aligned} \quad (13.74)$$

peut être écrit sous la forme:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_m \end{bmatrix} \Leftrightarrow A \cdot X = B \quad (13.75)$$

et nous savons que les seuls systèmes linéaires résolubles sont ceux qui ont autant d'équations que d'inconnues. Ainsi, la matrice A doit être une matrice carrée M_{nn} .

Si une solution existe, il existe alors une matrice-colonne (ou "vecteur") X tel que $A \cdot X = B$ ce qui implique:

$$X = A^{-1} \cdot B \quad (13.76)$$

Qu'impose cette relation ? Eh bien c'est relativement simple mais à la fois très très important: pour qu'un système linéaire ait une solution, il faut que la matrice A soit inversible ! Quel rapport avec le déterminant alors ? C'est simple: les mathématiciens ont cherché comment s'écrivaient les inverses des matrices de systèmes linéaires dont ils savaient qu'il y avait une solution et ils sont arrivés après tâtonnements successifs à déterminer une sorte de formule qui permette de vérifier si la matrice est inversible ou non. Une fois cette formule trouvée, ils ont formalisé (comme ils savent si bien le faire...), avec une très bonne rigueur, le concept entourant cette formule qu'ils ont appelé "déterminant". Ils y sont tellement bien arrivés d'ailleurs qu'on oublie parfois qu'ils ont procédé ainsi....

Remarque: Si une matrice d'un système linéaire n'est pas inversible, cela a pour conséquence qu'il existe soit aucune solution, soit une infinité de solutions (comme à l'habitude quoi...)

Nous allons ci-dessous d'abord nous intéresser à la manière de construire le déterminant en définissant un type d'application particulière. Ensuite, après avoir vu un exemple simple et interprétable du calcul d'un déterminant, nous nous attacherons à déterminer la formule de celui-ci dans le cas général. Enfin, une fois ceci fait, nous verrons quelle est la relation qui lie l'inverse d'une matrice et le déterminant.

Dans ce qui suit tous les espaces vectoriels considérés sont de dimension finie et sur le corps $\mathbb{C}$ des nombres complexes (ceux qui le préfèrent pourront prendre $\mathbb{R}$ comme corps de base, de fait nous pourrions prendre un corps quelconque).

D'abord nous allons faire un petit peu de mathématique (un peu rébarbative) avant de passer à du concret.

Soit V un espace vectoriel, nous écrirons V^n au lieu de $\underbrace{V \times \dots \times V}_{n \text{ fois}}$. $(\vec{e}_1, \dots, \vec{e}_n)$ désignera la base canonique de $\mathbb{C}^n$. $M_n(\mathbb{C})$ est l'ensemble des matrices carrées $n \times n$ à coefficients dans $\mathbb{C}$.

Définitions:

D1. Une "**application multilinéaire**" sur un espace V est par définition une application $\varphi: V^n \rightarrow \mathbb{C}$ qui est linéaire en chacune de ces composantes. C'est-à-dire:

$$\begin{aligned} & \varphi(\vec{x}_1, \dots, \vec{x}_{j-1}, \lambda \cdot \vec{u} + \mu \cdot \vec{v}, \vec{x}_{j+1}, \dots, \vec{x}_n) \\ &= \lambda \cdot \varphi(\vec{x}_1, \dots, \vec{x}_{j-1}, \vec{u}, \vec{x}_{j+1}, \dots, \vec{x}_n) + \mu \cdot \varphi(\vec{x}_1, \dots, \vec{x}_{j-1}, \vec{v}, \vec{x}_{j+1}, \dots, \vec{x}_n) \end{aligned} \quad (13.77)$$

pour tout $\lambda, \mu \in K$ et $\vec{x}_i, \vec{u}, \vec{v} \in V$ où les $\vec{x}_i$ sont des vecteurs.

Remarque: Une application multilinéaire non nulle n'est pas une application linéaire de l'espace V^n dans $\mathbb{C}$. Sauf si $n = 1$. Effectivement, cela se vérifie de par la définition de l'application linéaire versus celle de l'application multilinéaire:

$$\begin{aligned} \lambda \varphi(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n) &\neq \varphi(\lambda \vec{x}_1, \lambda \vec{x}_2, \dots, \lambda \vec{x}_n) \\ \varphi(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n) + \varphi(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n) &\neq \varphi(2\vec{x}_1, 2\vec{x}_2, \dots, 2\vec{x}_n) \end{aligned} \quad (13.78)$$

D2. Une "**application multilinéaire alternée**" sur V est par définition une application multilinéaire qui vérifie la condition suivante:

$$\varphi(\vec{x}_1, \dots, \vec{x}_j, \vec{x}_{j+1}, \vec{x}_{j+2}, \dots, \vec{x}_n) = -\varphi(\vec{x}_1, \dots, \vec{x}_{j+1}, \vec{x}_j, \vec{x}_{j+2}, \dots, \vec{x}_n) \quad (13.79)$$

pour tout $j = 1 \dots n, \vec{x}_i \in V$. Ainsi la permutation de deux vecteurs qui se suivent change le signe de φ .

Ainsi, si φ est une application multilinéaire, alors φ est alternée si et seulement si $\forall \vec{x}_i \in V, j = 1 \dots n$. nous avons:

$$\varphi(\vec{x}_1, \dots, \vec{x}_j, \vec{x}_j, \vec{x}_{j+2}, \dots, \vec{x}_n) = 0 \quad (13.80)$$

Démonstration: φ étant définie comme alternée, nous avons donc:

$$\begin{aligned} \varphi(\vec{x}_1, \dots, \vec{x}_j, \vec{x}_j, \vec{x}_{j+2}, \dots, \vec{x}_n) &= -\varphi(\vec{x}_1, \dots, \vec{x}_j, \vec{x}_j, \vec{x}_{j+2}, \dots, \vec{x}_n) \\ \Rightarrow \varphi(\vec{x}_1, \dots, \vec{x}_j, \vec{x}_j, \vec{x}_{j+2}, \dots, \vec{x}_n) + \varphi(\vec{x}_1, \dots, \vec{x}_j, \vec{x}_j, \vec{x}_{j+2}, \dots, \vec{x}_n) &= 0 \\ \Rightarrow 2\varphi(\vec{x}_1, \dots, \vec{x}_j, \vec{x}_j, \vec{x}_{j+2}, \dots, \vec{x}_n) &= 0 \\ \Rightarrow \varphi(\vec{x}_1, \dots, \vec{x}_j, \vec{x}_j, \vec{x}_{j+2}, \dots, \vec{x}_n) &= 0 \end{aligned} \quad (13.81)$$

□C.Q.F.D.

et voilà ce qui nous intéresse:

D3. Un "**déterminant**" est par définition (par imposition) une application multilinéaire alternée

$D: \underbrace{\mathbb{C}^n \times \dots \times \mathbb{C}^n}_{n \text{ fois}} \rightarrow \mathbb{C}$ vérifiant de plus:

$$D(\vec{e}_1, \dots, \vec{e}_n) = 1 \quad (13.82)$$

Remarque: Les colonnes d'une matrice carrée forment n vecteurs et nous voyons donc qu'un déterminant D sur $\mathbb{C}^n$ induit une application $\tilde{D}$ de $M_n(\mathbb{C}) \rightarrow \mathbb{C}$ (où $M_n(\mathbb{C})$ est l'espace des matrices carrées $n \times n$ à coefficients dans $\mathbb{C}$) définie par $\tilde{D}(M) = D(\vec{m}_1, \dots, \vec{m}_n)$ où $\vec{m}_i$ est la i -ème colonne de M . Par la suite, nous ferons l'abus d'écriture qui consiste à confondre D et $\tilde{D}$.

Etudions le cas $n = 2$. Si D est un déterminant, pour tout vecteur:

$$\vec{x} = \lambda \cdot \vec{e}_1 + \mu \cdot \vec{e}_2, \vec{y} = \alpha \cdot \vec{e}_1 + \beta \cdot \vec{e}_2 \quad (13.83)$$

nous avons:

$$D(\lambda \cdot \vec{e}_1 + \mu \cdot \vec{e}_2, \alpha \cdot \vec{e}_1 + \beta \cdot \vec{e}_2) \quad (13.84)$$

Comme D est multilinéaire, nous avons:

$$\begin{aligned} D(\lambda \cdot \vec{e}_1 + \mu \cdot \vec{e}_2, \alpha \cdot \vec{e}_1 + \beta \cdot \vec{e}_2) &= \lambda D(\vec{e}_1, \alpha \cdot \vec{e}_1 + \beta \cdot \vec{e}_2) + \mu D(\vec{e}_2, \alpha \cdot \vec{e}_1 + \beta \cdot \vec{e}_2) \\ &= \lambda (\alpha D(\vec{e}_1, \vec{e}_1) + \beta D(\vec{e}_1, \vec{e}_2)) + \mu (\alpha D(\vec{e}_2, \vec{e}_1) + \beta D(\vec{e}_2, \vec{e}_2)) \\ &= \lambda \cdot \alpha \cdot D(\vec{e}_1, \vec{e}_1) + \lambda \cdot \beta \cdot D(\vec{e}_1, \vec{e}_2) + \mu \cdot \alpha \cdot D(\vec{e}_2, \vec{e}_1) + \mu \cdot \beta \cdot D(\vec{e}_2, \vec{e}_2) \end{aligned} \quad (13.85)$$

et comme elle est surtout multilinéaire alternée, nous avons donc:

$$D(\lambda \cdot \vec{e}_1 + \mu \cdot \vec{e}_2, \alpha \cdot \vec{e}_1 + \beta \cdot \vec{e}_2) = \lambda \cdot \beta \cdot D(\vec{e}_1, \vec{e}_2) - \mu \cdot \alpha \cdot D(\vec{e}_1, \vec{e}_2) = \lambda \cdot \beta - \mu \cdot \alpha \quad (13.86)$$

En fait, nous venons de montrer que si un déterminant existe, il est unique et de la forme indiquée ci-dessus, il faudrait encore vérifier que l'application ainsi définie satisfait les propriétés d'un déterminant, mais ce dernier point est immédiat.

Ainsi, si $M = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ est une matrice nous avons donc:

$$D(M) = a_{11} \cdot a_{22} - a_{21} \cdot a_{12} \quad (13.87)$$

Nous retrouvons donc la forme du déterminant tel que nous en avons fait mention en calcul vectoriel.

Donnons une interprétation géométrique du déterminant. Soit $\vec{v}_1, \vec{v}_2$ deux vecteurs de $\mathbb{R}^2$.

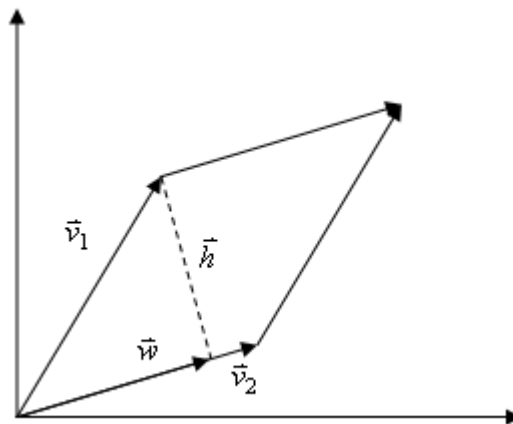


Figure: 13.2 - Interprétation géométrique du déterminant

Le vecteur $\vec{w}$ est obtenu en projetant $\vec{v}_1$ sur $\vec{v}_2$ et nous avons donc:

$$\vec{w} = \frac{\vec{v}_1 \cdot \vec{v}_2}{\|\vec{v}_2\|^2} \vec{v}_2 \text{ et } \|\vec{h}\|^2 = \|\vec{v}_1\|^2 - \|\vec{w}\|^2 = \|\vec{v}_1\|^2 - \frac{(\vec{v}_1 \cdot \vec{v}_2)^2}{\|\vec{v}_2\|^4} \|\vec{v}_2\|^2 = \|\vec{v}_1\|^2 - \frac{(\vec{v}_1 \cdot \vec{v}_2)^2}{\|\vec{v}_2\|^2} \quad (13.88)$$

L'aire du parallélogramme ci-dessus est donc:

$$S = \|\vec{v}_2\| \cdot \|\vec{h}\| = \|\vec{v}_2\| \cdot \left(\|\vec{v}_1\|^2 - \frac{(\vec{v}_1 \circ \vec{v}_2)^2}{\|\vec{v}_2\|^2} \right)^{1/2} = \left(\|\vec{v}_1\|^2 \cdot \|\vec{v}_2\|^2 - (\vec{v}_1 \circ \vec{v}_2)^2 \right)^{1/2} \quad (13.89)$$

Si $\vec{v}_1 = \lambda \vec{e}_1 + \mu \vec{e}_2$ et $\vec{v}_2 = \alpha \vec{e}_1 + \beta \vec{e}_2$ alors:

$$S^2 = (\lambda^2 + \mu^2) \cdot (\alpha^2 + \beta^2) - (\lambda\alpha + \mu\beta)^2 = \lambda^2 \cdot \beta^2 + \mu^2 \cdot \alpha^2 - 2\lambda\alpha\mu\beta = (\lambda\beta - \mu\alpha)^2 \quad (13.90)$$

et donc:

$$S = |\lambda\beta - \mu\alpha| = |D(\vec{v}_1, \vec{v}_2)| = \left| \det \begin{pmatrix} \lambda & \alpha \\ \mu & \beta \end{pmatrix} \right| \quad (13.91)$$

Ainsi le déterminant représente au signe près l'aire du parallélogramme défini par les vecteurs $\vec{v}_1, \vec{v}_2$ lorsque ceux-ci sont linéairement indépendants. Nous pouvons généraliser ce résultat à une dimension n quelconque, en particulier, pour $n = 3$, le déterminant de trois vecteurs linéairement indépendants représente le volume du parallélépipède défini par ces derniers.

Le cas plus général de l'expression du déterminant est un peu plus délicat à établir. Il faut pour cela que nous définissions une application bijective particulière mais simple que nous avons déjà rencontrée dans le chapitre Statistique.

Définition: Soit $F_n = \{1, 2, \dots, n\}, n \in \mathbb{N}^*$ nous appelons "**permutation**" de F_n toute application bijective de F_n dans F_n :

$$\sigma: \{1, 2, 3, \dots, n\} \rightarrow \{1, 2, \dots, n\} \quad (13.92)$$

Soit $\mathcal{S}_n$ l'ensemble des permutations (applications bijectives) possibles de $\{1, 2, 3, \dots, n\}$. $\mathcal{S}_n$ contient bien évidemment... (voir la combinatoire dans le chapitre de Probabilités) $n!$ éléments. La donnée d'un élément σ de $\mathcal{S}_n$ est définie par les données successives de:

$$\begin{aligned} &\sigma(1) \quad (\sigma(1) \in F_n) \\ &\sigma(2) \quad (\sigma(2) \in F_n - \{\sigma(1)\}) \\ &\dots \\ &\sigma(n) \quad (\sigma(n) \in F_n - \{\sigma(1), \sigma(2), \dots, \sigma(n-1)\}) \end{aligned} \quad (13.93)$$

Etant donnée une suite d'éléments ordonnée (croissants) d'éléments $\{1, 2, \dots, n\}, n \in \mathbb{N}^*$, nous appelons "**inversion**", toute permutation d'éléments dans la suite ordonnée (donc la suite ne sera plus ordonnée du tout...). Nous notons $I(\sigma)$ le nombre d'inversions.

Nous disons que la permutation σ est paire (impaire) si $I(\sigma)$ est pair (impair). Nous appelons "**signature**" de σ , le nombre noté $\varepsilon(\sigma)$ défini par $\varepsilon(\sigma) = (-1)^{I(\sigma)}$, c'est-à-dire:

$$\varepsilon(\sigma) = \begin{cases} +1 & \text{si la permutation } \sigma \text{ est paire} \\ -1 & \text{si la permutation } \sigma \text{ est impaire} \end{cases} \quad (13.94)$$

Nous avons maintenant les outils en place nécessaires à définir de manière générale la formule du déterminant:

Définition: Soit:

$$A = (a_{ij})_{1 \leq i, j \leq n} \in M_n(K) \quad (13.95)$$

Nous appelons "**déterminant de A**" d'une matrice carrée de dimension n , et nous notons $\det(A)$, le scalaire K défini par (nous verrons un exemple plus loin):

$$\det(A) = \sum_{\sigma \in \mathcal{S}_n} \varepsilon(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)} \quad (13.96)$$

appelé parfois "**formule de Leibniz**" ou encore "**formule de Laplace**". Cette relation a été obtenue par tâtonnements successifs et par récurrence pour de plus grandes dimensions.

 **Exemples:**

E1. Soit $A = (a_{ij})_{1 \leq i, j \leq 2} \in M_{22}(K)$, considérons les $2! = 2$ permutations des seconds indices (des entiers 1,2) pris dans leur ensemble:

$$\begin{array}{cc} 12 & 21 \end{array} \quad (13.97)$$

Nous calculons les signatures de σ . Voici le schéma de cette règle (rappel: nous disons donc... qu'il y a "inversion", si dans une permutation, un entier supérieur précède un entier inférieur):

$\sigma(1)\sigma(2)$	$\begin{array}{c} 12 \\ \sigma(1)\sigma(2) \end{array}$	$\begin{array}{c} 21 \\ \sigma(1)\sigma(2) \end{array}$
Nombre d'inversions	0	1
Permutation	Paire	Impaire
$\varepsilon(\sigma) = (-1)^{I(\sigma)}$	+1	-1

Tableau: 13.1 - Inversions et permutations d'un déterminant d'ordre 2

Donc nous avons:

$$\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \sum_{\sigma \in \mathcal{S}_2} \varepsilon(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} = \underbrace{a_{11}a_{22}}_{\varepsilon(\sigma)=+1 \begin{cases} \sigma(1)=1 \\ \sigma(2)=2 \end{cases}} - \underbrace{a_{12}a_{21}}_{\varepsilon(\sigma)=-1 \begin{cases} \sigma(1)=2 \\ \sigma(2)=1 \end{cases}} \quad (13.98)$$

Ce qui correspond bien à ce que nous avons vu initialement.

E2. Soit $A = (a_{ij})_{1 \leq i, j \leq 3} \in M_{33}(K)$, considérons les $3! = 6$ permutations des seconds indices (des entiers 1,2,3) pris dans leur ensemble:

$$123 \quad 132 \quad 213 \quad 231 \quad 312 \quad 321 \quad (13.99)$$

Nous calculons les signatures de σ . Voici le schéma de cette règle (rappel: nous disons donc... qu'il y a "inversion", si dans une permutation, un entier supérieur précède un entier inférieur):

$\sigma(1)\dots\sigma(2)\dots\sigma(3)$	123	132	213	231	312	321
Nombre d'inversions	0	1	1	2	2	3
Permutation	Paire	Impaire	Impaire	Paire	Paire	Impaire
$\varepsilon(\sigma) = (-1)^{I(\sigma)}$	+1	-1	-1	+1	+1	-1

Tableau: 13.2 - Inversions et permutations d'un déterminant d'ordre 3

Donc nous avons:

$$\det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \sum_{\sigma \in \mathcal{S}_3} \varepsilon(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} a_{3\sigma(3)} \quad (13.100)$$

$$= a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}$$

Remarque: Certaines personnes apprennent par coeur une méthode nommée "règle de Sarrus" pour calculer les déterminants d'ordre trois comme le précédent. Nous lui préférons sur ce site la formulation générale du déterminant applicable à tous les ordres.

Voyons quelques propriétés et corollaires de cette formulation du déterminant:

P1. Soit $M_n(K)$ une matrice carrée d'ordre n , nous ne changeons pas la valeur du déterminant de M_n en:

1. Effectuant une opération élémentaire sur les colonnes de M_n
2. Effectuant une opération élémentaire sur les lignes de M_n

Démonstration: Si $M_n = (a_{ij})_{i,j=1,\dots,n}$ alors M_n est composée de n vecteurs colonnes:

$$v_n = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{pmatrix} \quad (13.101)$$

Effectuer une opération élémentaire sur les colonnes de M_n revient à additionner $\lambda v_i, i \in \{1, \dots, n\}$ à une des colonnes v_n de M_n . Soit M_n' la matrice obtenue en additionnant λv à la j -ème colonne de M_n , nous avons:

$$\det(M_n') = \det(v_1, \dots, v_i, \dots, v_j + \lambda v_i, \dots, v_n) \quad (13.102)$$

Par multilinéarité (finalement la démonstration n'est vraiment pas bien dure):

$$\det(M') = \det(v_1, \dots, v_i, \dots, v_j, \dots, v_n) + \lambda \det(v_1, \dots, v_i, \dots, v_i, \dots, v_n) \quad (13.103)$$

et comme le déterminant est alterné:

$$\lambda \det(v_1, \dots, v_i, \dots, v_i, \dots, v_n) = 0 \Rightarrow \det(M_n') = \det(v_1, \dots, v_i, \dots, v_j, \dots, v_n) = \det(M_n) \quad (13.104)$$

Pour ce qui est des opérations élémentaires sur les lignes il suffit de considérer la transposée (c'est à pleurer tellement c'est simple mais il fallait y penser).

□C.Q.F.D.

P2. Soit $M_n(K)$ une matrice carrée d'ordre n et soit $\lambda \in K$:

$$\det(\lambda M_n) = \lambda^n \det(M_n) \quad (13.105)$$

Démonstration: Comme précédemment, il suffit de remarquer que si $v_1, \dots, v_n$ sont les vecteurs colonnes constituant la matrice M_n alors $\lambda v_1, \dots, \lambda v_n$ sont ceux qui constituent λM_n et:

$$\det(\lambda M_n) = \det(\lambda v_1, \dots, \lambda v_n) \quad (13.106)$$

L'application étant n -linéaire, nous aboutissons à l'égalité:

$$\det(\lambda M_n) = \lambda^n \det(v_1, \dots, v_n) \quad (13.107)$$

□C.Q.F.D.

P3. Soit $M_n(K)$ une matrice carrée d'ordre n . Nous changeons le signe du déterminant de M_n si:

1. Nous permutons deux de ses colonnes
2. Nous permutons deux de ses lignes

Démonstration: M_n est constituée des n vecteurs $v_1, \dots, v_n$. Le déterminant de M_n est égal au déterminant de ces n . Permuter deux colonnes de M_n revient à permuter les deux vecteurs correspondant. Supposons que les vecteurs permutés soit le i -ème et le j -ème, l'application déterminant étant alternée, nous avons:

$$\det(v_1, \dots, v_i, \dots, v_j, \dots, v_n) = -\det(v_1, \dots, v_j, \dots, v_i, \dots, v_n) \quad (13.108)$$

Pour ce qui est des lignes, il suffit de considérer la transposée de M_n .

□C.Q.F.D.

P4. Soit $A, B \in M_n(\mathbb{C})$ alors:

$$\det(A \cdot B) = \det(A) \cdot \det(B) \quad (13.109)$$

La démonstration peut se faire de deux manières, la première est assez indigeste et abstraite nous la laisserons aux mathématiciens (...) même si elle a l'avantage d'être générale, la seconde plus simple, consiste à vérifier cette assertion pour différentes matrices carrées.

Démonstration:

$$\begin{aligned} \det(A \cdot B) &= \det \left(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \right) = \det \left(\begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix} \right) \\ &= (a_{11}b_{11} + a_{12}b_{21})(a_{21}b_{12} + a_{22}b_{22}) - (a_{11}b_{12} + a_{12}b_{22})(a_{21}b_{11} + a_{22}b_{21}) \\ &= \cancel{a_{11}b_{11}a_{21}b_{12}} + a_{11}b_{11}a_{22}b_{22} + a_{12}b_{21}a_{21}b_{12} + \cancel{a_{12}b_{21}a_{22}b_{22}} \\ &\quad - \cancel{a_{11}b_{12}a_{21}b_{11}} - a_{11}b_{12}a_{22}b_{21} - a_{12}b_{22}a_{21}b_{11} - \cancel{a_{12}b_{22}a_{22}b_{21}} \\ &= a_{11}b_{11}a_{22}b_{22} - a_{11}b_{12}a_{22}b_{21} - a_{12}b_{22}a_{21}b_{11} + a_{12}b_{21}a_{21}b_{12} \\ &= a_{11}a_{22}b_{11}b_{22} - a_{11}a_{22}b_{12}b_{21} - a_{12}a_{21}b_{22}b_{11} + a_{12}a_{21}b_{21}b_{12} \\ \det(A) \det(B) &= (a_{11}a_{22} - a_{12}a_{21})(b_{11}b_{22} - b_{12}b_{21}) \\ &= a_{11}a_{22}b_{11}b_{22} - a_{11}a_{22}b_{12}b_{21} - a_{12}a_{21}b_{11}b_{22} + a_{12}a_{21}b_{12}b_{21} \end{aligned} \quad (13.110)$$

Les calculs donnent donc des résultats qui sont bien identiques. Nous pouvons vérifier ainsi pour des matrices carrées de dimensions supérieures.

□C.Q.F.D.

P5. Une matrice carrée $A \in M_n(\mathbb{C})$ est inversible si et seulement si $\det(A) \neq 0$.

Démonstration:

Si A est inversible, nous avons:

$$\det(A \cdot A^{-1}) = \det(\mathbb{I}) = 1 = \det(A) \cdot \det(A^{-1}) \Rightarrow \det(A) \neq 0 \quad (13.111)$$

□C.Q.F.D.

Il s'agit de la propriété la plus importante des matrices dans le cadre de la physique théorique car si A est un système linéaire, le calcul de son déterminant permet de savoir si celui-ci a des solutions uniques. Dans le cas contraire, comme nous en avons déjà fait mention, soit le système n'a aucune solution, soit une infinité !

Il faut considérer aussi un cas particulier important. Soit le système suivant:

$$A \cdot B = 0 \Rightarrow B = 0 \quad (13.112)$$

où $A \in M_n(K) \neq 0$ et $B \in M_n(K)$ à déterminer. Il est clair..., que A soit inversible ou non, la solution triviale est $B = 0$. Cependant..., imaginons un cas de physique théorique où nous avons $A \cdot B = 0$ mais

pour lequel nous savons que $A \in M_n(K) \neq 0$ et pour lequel nous imposons $B \in M_n(K) \neq 0$. Dans ce cas, il nous faut éliminer la solution triviale $B = 0$. De plus, calculer l'inverse (s'il existe) de la matrice A ne nous ramènera à rien de concret mis à part à $B = 0$ ce qui bien évidemment ne nous satisfait pas. La seule solution est alors de se débrouiller pour que les coefficients a_{ij} de la matrice A soient tels que son déterminant soit nul et donc la matrice non inversible! L'intérêt ? Eh, bien d'avoir une infinité de solutions possibles (de B donc !) qui satisfont $A \cdot B = 0$. Nous aurons besoin de cette méthodologie en mécanique quantique ondulatoire, lorsque nous déterminerons l'existence des antiparticules par l'intermédiaire de l'équation de Dirac linéarisée. Il faudra donc s'en rappeler.

P6. Deux matrices "**conjuguées**" (attention, pas dans le sens complexe du terme) ont le même déterminant.

Démonstration:

Soit $A \in M_n(\mathbb{C})$, et $P \in M_n(\mathbb{C})$ une matrice de passage d'une base à une autre (voir plus loin le traitement des changements de bases), nous avons alors:

$$\det(B) = \det(P^{-1} \cdot A \cdot P) = \underbrace{\det(P^{-1})}_{=\det(P)^{-1}} \cdot \det(A) \cdot \det(P) = \det(A) \quad (13.113)$$

□C.Q.F.D.

P7. Pour toute matrice $A \in M_n(\mathbb{C})$:

$$\det(\bar{A}) = \overline{\det(A)} \quad (13.114)$$

Démonstration:

$$\det(\bar{A}) = \sum_{\sigma \in S_n} \varepsilon(\sigma) \cdot \overline{a_{1,\sigma(1)}} \cdot \dots \cdot \overline{a_{n,\sigma(n)}} \quad (13.115)$$

Or (trivial... simple multiplication de tous les coefficients):

$$\overline{a_{1,\sigma(1)}} \cdot \dots \cdot \overline{a_{n,\sigma(n)}} = \overline{a_{\sigma^{-1}(1),1}} \cdot \dots \cdot \overline{a_{\sigma^{-1}(n),n}} \quad (13.116)$$

Puisque (trivial) $\varepsilon(\sigma^{-1}) = \varepsilon(\sigma)$ et que $x, y \in \mathbb{C}: \overline{\overline{x \cdot y}} = x \cdot y$ (cf. chapitre sur les Nombres), nous pouvons alors écrire:

$$\det(\bar{A}) = \sum_{\sigma \in S_n} \varepsilon(\sigma) \cdot \overline{a_{\sigma^{-1}(1),1}} \cdot \dots \cdot \overline{a_{\sigma^{-1}(n),n}} = \overline{\sum_{\sigma^{-1} \in S_n} \varepsilon(\sigma^{-1}) \cdot a_{\sigma^{-1}(1),1} \cdot \dots \cdot a_{\sigma^{-1}(n),n}} = \overline{\det(A)} \quad (13.117)$$

□C.Q.F.D.

P8. Pour toute matrice $A \in M_n(\mathbb{R})$:

$$\det(A^t) = \det(A) \quad (13.118)$$

Démonstration:

Ben... c'est la même chose que pour la propriété précédente mais sans les valeurs conjuguées... De fait, nous montrons de la même manière, la même propriété pour $A \in M_n(\mathbb{C})$.

□C.Q.F.D.

P9. Soit une matrice $A = (a_{ij}) \in M_n(\mathbb{C})$, nous noterons A_{ij} la matrice obtenue à partir de A en effaçant la i -ème ligne et la j -ème colonne (notation très importante à ne pas oublier pour la suite!!!). A_{ij} appartient donc à $M_{n-1}(\mathbb{C})$. Alors pour tout $i = 1..n$:

$$\det(A) = \sum_{j=1}^n (-1)^{i+j} \cdot a_{ij} \cdot \det(A_{ij}) \quad (13.119)$$

où le terme:

$$(-1)^{i+j} \det(A_{ij}) \quad (13.120)$$

est appelé le "**cofacteur**".

Démonstration:

Définissons pour cela l'application:

$$\varphi: M_n(\mathbb{C}) \rightarrow \mathbb{C}, \quad \varphi(A) := \sum_{j=1}^n (-1)^{i+j} \cdot a_{ij} \cdot \det(A_{ij}) \quad (13.121)$$

Il est facile de voir que φ est multilinéaire (il suffit de considérer $(-1)^{i+j} \cdot a_{ij}$ comme une simple constante et ensuite par extension de la définition du déterminant... trop facile...).

Montrons cependant qu'elle est alternée (dans ce cas, c'est un déterminant qui a toutes les propriétés d'un déterminant):

Soit a_k, a_{k+1} deux vecteurs colonne de A qui se suivent. Supposons que $a_k = a_{k+1}$, il faut montrer que dans ce cas $\varphi(A) = 0$ (qui découle de la définition d'une application alternée).

Nous avons premièrement (c'est obligatoire de par la définition) si nous n'effaçons aucune des colonnes j étant k ou $k+1$:

$$\det(A_{ij}) = 0 \text{ si } j \neq k, k+1 \quad (13.122)$$

et nous avons bien évidemment si nous enlevons respectivement la colonne k et la colonne $k+1$:

$$\det(A_{ik}) = \det(A_{i(k+1)}) \text{ car } A_{i,k} = A_{i,k+1} \quad (13.123)$$

Donc:

$$\varphi(A) = (-1)^{i+k} \cdot a_{ik} \cdot \det(A_{ik}) + (-1)^{i+(k+1)} \cdot a_{i(k+1)} \cdot \det(A_{i(k+1)}) = 0 \quad (13.124)$$

C'est donc OK. Elle est alternée et multilinéaire, il s'agit donc bien d'un déterminant.

Nous venons donc de montrer que φ est un déterminant et par unicité nous avons $\varphi(A) = \det(A)$ pour tout $A \in M_n(\mathbb{C})$.

□C.Q.F.D.

Voyons un exemple de cette méthode en calculant le déterminant de:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 0 \\ 1 & -1 & i \end{bmatrix} \quad (13.125)$$

Développons selon la deuxième ligne ($i = 2$). Nous obtenons:

$$\begin{aligned} \det(A) &= \begin{vmatrix} 1 & 2 & 3 \\ 2 & 1 & 0 \\ 1 & -1 & i \end{vmatrix} = \sum_{j=1}^n (-1)^{i+j} \cdot a_{ij} \cdot \det(A_{ij}) = \sum_{j=1}^3 (-1)^{2+j} \cdot a_{2j} \cdot \det(A_{2j}) \\ &= (-1)^3 \cdot 2 \cdot \det \begin{bmatrix} 2 & 3 \\ -1 & i \end{bmatrix} + (-1)^4 \cdot 1 \cdot \det \begin{bmatrix} 1 & 3 \\ 1 & i \end{bmatrix} + (-1)^5 \cdot 0 \cdot \det \begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix} \\ &= -2 \cdot (2i + 3) + i - 3 = -3i - 9. \end{aligned} \quad (13.126)$$

Développons selon la première colonne en guise de vérification (on ne sait jamais...):

$$\begin{aligned} \det(A) &= \begin{vmatrix} 1 & 2 & 3 \\ 2 & 1 & 0 \\ 1 & -1 & i \end{vmatrix} = \det \begin{bmatrix} 1 & 0 \\ -1 & i \end{bmatrix} - 2 \cdot \det \begin{bmatrix} 2 & 3 \\ -1 & i \end{bmatrix} + \det \begin{bmatrix} 2 & 3 \\ 1 & 0 \end{bmatrix} \\ &= i - 2 \cdot (2i + 3) - 3 = -3i - 9 \end{aligned} \quad (13.127)$$

Le calcul déterminé ci-dessus est donc exponentiel car si par exemple nous devons calculer le déterminant d'une matrice d'ordre 10 alors le déterminant sera développé en une somme de 10 termes, dont chacun contient le déterminant d'une matrice d'ordre 9, qui est un cofacteur de la matrice de départ. Si nous développons n'importe lequel de ces déterminants, nous obtenons une somme de 9 déterminants dont chacun contient le déterminant d'une matrice d'ordre 8. A ce stade, il y a donc 90 déterminants de matrices d'ordre 8 à calculer. Le processus pourrait se poursuivre jusqu'à ce qu'il ne reste que des déterminants d'ordre 2. Et alors là nous devinons que le nombre de matrices d'ordre 2 est très conséquent!

Définition: Soit m, n deux entiers positifs quelconques et A une matrice $m \times n$ à coefficients dans $\mathbb{C}$. Pour tout entier $k \leq \min(m, n)$ un "**mineur d'ordre k** " de A est un déterminant du type:

$$\det \begin{bmatrix} a_{i_1, j_1} & \dots & a_{i_1, j_k} \\ \vdots & \ddots & \vdots \\ a_{i_k, j_1} & \dots & a_{i_k, j_k} \end{bmatrix} \text{ avec } 1 \leq i_1 < \dots < i_k \leq m \text{ et } 1 \leq j_1 < \dots < j_k \leq n \quad (13.128)$$

Dans le cas particulier d'une matrice carrée d'ordre $n > 1$ la définition est plus simple: Le mineur M_{ij} de l'élément a_{ij} est le déterminant de la matrice d'ordre $n - 1$ obtenue en supprimant la ligne i et la colonne j . Ainsi, pour calculer le mineur d'un élément, nous supprimons la ligne et la colonne auxquelles l'élément appartient, puis nous calculons le déterminant de la matrice carrée restante.

Pour finir nous terminons en donnant une formule qui relie les coefficients de l'inverse d'une matrice $n \times n$ avec ses mineurs d'ordre $n - 1$.

3.3.1. DÉRIVÉE D'UN DÉTERMINANT

Voyons maintenant un résultat qui nous sera fort utile en relativité générale.

Soit une matrice carrée $n \times n$ avec $g_{ij} : \mathbb{R} \rightarrow \mathbb{R}$ des fonctions dérivables. Posons $g := \det(G)$ $G = (g_{ij})$. Nous voulons calculer $d_t g$. Soit g_i le i -ème vecteur colonne de la matrice G . Utilisons la formule:

$$g = \det(g_1, \dots, g_n) = \sum_{\sigma \in \mathcal{S}_n} \varepsilon(\sigma) \cdot g_{\sigma(1),1} \cdot \dots \cdot g_{\sigma(n),n} \quad (13.129)$$

Sachant que la dérivée de $g_{\sigma(1),1} \cdot \dots \cdot g_{\sigma(n),n}$ est (dérivée de n produits):

$$\begin{aligned} & \frac{dg_{\sigma(1),1}}{dt} \cdot g_{\sigma(2),2} \cdot \dots \cdot g_{\sigma(n),n} + g_{\sigma(1),1} \cdot \frac{dg_{\sigma(2),2}}{dt} \cdot g_{\sigma(3),3} \cdot \dots \cdot g_{\sigma(n),n} + \dots \\ & + g_{\sigma(1),1} \cdot \dots \cdot g_{\sigma(n-1),n-1} \cdot \frac{dg_{\sigma(n),n}}{dt} \end{aligned} \quad (13.130)$$

nous avons donc:

$$\begin{aligned} \frac{dg}{dt} &= \sum_{\sigma \in \mathcal{S}_n} \varepsilon(\sigma) \cdot \frac{dg_{\sigma(1),1}}{dt} \cdot g_{\sigma(2),2} \cdot \dots \cdot g_{\sigma(n),n} + \sum_{\sigma \in \mathcal{S}_n} \varepsilon(\sigma) \cdot g_{\sigma(1),1} \cdot \frac{dg_{\sigma(2),2}}{dt} \cdot g_{\sigma(3),3} \cdot \dots \cdot g_{\sigma(n),n} \\ &+ \dots + \sum_{\sigma \in \mathcal{S}_n} \varepsilon(\sigma) \cdot g_{\sigma(1),1} \cdot \dots \cdot g_{\sigma(n-1),n-1} \cdot \frac{dg_{\sigma(n),n}}{dt} \end{aligned} \quad (13.131)$$

Si nous regardons la première somme ci-dessus, nous remarquons que:

$$\sum_{\sigma \in \mathcal{S}_n} \varepsilon(\sigma) \cdot \frac{dg_{\sigma(1),1}}{dt} \cdot g_{\sigma(2),2} \cdot \dots \cdot g_{\sigma(n),n} = \det(g'_1, g_2, \dots, g_n) \quad (13.132)$$

où g'_1 est la dérivée du vecteur g_1 . De même pour les sommes suivantes. Ainsi,

$$\frac{dg}{dt} = \det(g'_1, g_2, \dots, g_n) + \det(g_1, g'_2, g_3, \dots, g_n) + \dots + \det(g_1, g_2, \dots, g'_n) \quad (13.133)$$

Développons encore. Considérons le terme $\det(g_1', g_2, \dots, g_n)$ ci-dessus. Si nous le développons par rapport à la première colonne, nous obtenons:

$$\det(g_1', g_2, \dots, g_n) = \sum_{i=1}^n (-1)^{1+i} \cdot \frac{dg_{i,1}}{dt} \cdot \det(G_{i,1}) \quad (13.134)$$

De même, en développant le j -ème terme de la somme ci-dessus par rapport à la j -ème colonne nous avons:

$$\det(g_1, \dots, g_j', \dots, g_n) = \sum_{i=1}^n (-1)^{j+i} \cdot \frac{dg_{i,j}}{dt} \cdot \det(G_{i,j}) \quad (13.135)$$

Si nous posons:

$$\alpha_{ij} := (-1)^{i+j} \cdot \det(G_{i,j}) \quad (13.136)$$

nous obtenons:

$$\frac{dg}{dt} = \sum_{j=1}^n \sum_{i=1}^n \alpha_{ij} \cdot \frac{dg_{i,j}}{dt} \quad (13.137)$$

ce qui en notation tensorielle (cf. chapitre de Calcul Tensoriel) s'écrit:

$$\frac{dg}{dt} = \alpha_{ij} \cdot \frac{dg_{i,j}}{dt} \quad (13.138)$$

Nous avons aussi:

$$b_{ji} = \frac{1}{\det(G)} \cdot (-1)^{i+j} \cdot \det(G_{ij}) = \frac{\alpha_{ij}}{\det(G)} = \frac{\alpha_{ij}}{g} \quad (13.139)$$

où b_{ji} est le coefficient se trouvant à la j -ème ligne, i -ème colonne de la matrice G^{-1} . Si nous notons g^{ij} le coefficient i, j de la matrice $(G^{-1})^t$ alors:

$$g^{ij} = b_{ji} \text{ et } g^{ij} = \frac{\alpha_{ij}}{g} \quad (13.140)$$

L'expression de la dérivée devient finalement:

$$\frac{dg}{dt} = g \cdot \sum_{j=1}^n \sum_{i=1}^n g^{ij} \cdot \frac{dg_{i,j}}{dt} \quad (13.141)$$

qui s'écrit en notation tensorielle:

$$\frac{dg}{dt} = g \cdot g^{ij} \cdot \frac{dg_{i,j}}{dt} \quad (13.142)$$

Ce résultat, finalement assez simple, nous sera utile dans le chapitre de Calcul Tensoriel, pour construire les outils nécessaires à l'étude de la relativité générale et à la détermination de l'équation d'Einstein des champs. Il convient donc de s'en rappeler.

3.3.2. INVERSE D'UNE MATRICE

Terminons notre étude des déterminants avec la cerise sur le gâteau en donnant une relation très importante dans de nombreux domaines de l'ingénierie, de la physique et de la mathématique qui relie les coefficients de l'inverse d'une matrice $n \times n$ avec ses mineurs d'ordre $n - 1$ (nous allons utiliser cette relation plus loin).

Soit $A \in M_n(\mathbb{C})$ une matrice inversible. Notons $A = (a_{ij})$ et $A^{-1} = (b_{ij})$. Alors:

$$b_{ij} = \frac{1}{\det(A)} \cdot (-1)^{i+j} \cdot \det(A_{ji}) \quad (13.143)$$

Démonstration:

Notons a_k le k -ème vecteur colonne de la matrice A . Sachant que $A \cdot A^{-1} = \mathbb{1}$, nous avons (trivial):

$$e_j = \sum_{k=1}^n b_{kj} a_k \quad (13.144)$$

Calculons $\det(a_1, \dots, a_{k-1}, e_j, a_{k+1}, \dots, a_n)$. D'une part en développant par rapport à la k -ème colonne nous trouvons (puisque qu'un seul des coefficients de e_j est non nul et que l'unique non nul est égal à l'unité):

$$\det(a_1, \dots, a_{k-1}, e_j, a_{k+1}, \dots, a_n) = (-1)^{j+k} \cdot \det(A_{jk}) \quad (13.145)$$

D'autre part (propriétés du déterminant):

$$\begin{aligned} \det(a_1, \dots, a_{k-1}, e_j, a_{k+1}, \dots, a_n) &= \det(a_1, \dots, a_{k-1}, \sum_{l=1}^n b_{lj} a_l, a_{k+1}, \dots, a_n) \\ &= \sum_{l=1}^n b_{lj} \cdot \underbrace{\det(a_1, \dots, a_{k-1}, a_l, a_{k+1}, \dots, a_n)}_{=0 \text{ si } l \neq k} = b_{kj} \cdot \det(A) \end{aligned} \quad (13.146)$$

Ainsi:

$$(-1)^{j+k} \cdot \det(A_{jk}) = b_{kj} \cdot \det(A) \quad (13.147)$$

c'est-à-dire:

$$b_{kj} = \frac{1}{\det(A)} \cdot (-1)^{j+k} \cdot \det(A_{jk}) \quad (13.148)$$

□C.Q.F.D.

Pour une application pratique simple, détaillée et importante dans l'industrie (car sinon dans l'ensemble du site nous inversons que très rarement de petites matrices), le lecteur pourra se reporter au chapitre de Méthodes Numériques dans la partie qui concerne la régression linéaire multiple.

4. CHANGEMENTS DE BASES

Supposons que nous passions d'une base $\mathcal{E} = (\vec{e}_1, \vec{e}_2, \dots, \vec{e}_n)$ d'un espace V^n à une autre base $\mathcal{F} = (\vec{f}_1, \vec{f}_2, \dots, \vec{f}_n)$ de ce même espace.

Décomposons les $\vec{f}_i$ dans la base $\mathcal{E}$:

$$\begin{aligned}\vec{f}_1 &= p_{11}\vec{e}_1 + p_{21}\vec{e}_2 + \dots + p_{n1}\vec{e}_n \\ \vec{f}_2 &= p_{12}\vec{e}_1 + p_{22}\vec{e}_2 + \dots + p_{n2}\vec{e}_n \\ &\dots \\ \vec{f}_n &= p_{1n}\vec{e}_1 + p_{2n}\vec{e}_2 + \dots + p_{nn}\vec{e}_n\end{aligned}\quad (13.149)$$

Définition: Nous appelons "**matrice de transition**" ou "**matrice de passage**", la matrice (l'application linéaire) qui permet de passer de $\mathcal{E} \rightarrow \mathcal{F}$ donnée par:

$$P = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1n} \\ p_{21} & p_{22} & \dots & p_{2n} \\ \vdots & & \ddots & \vdots \\ p_{n1} & p_{n2} & \dots & p_{nn} \end{bmatrix} \quad (13.150)$$

$\underbrace{\hspace{10em}}_{\begin{matrix} \vec{f}_1 & \vec{f}_2 & \dots & \vec{f}_n \end{matrix}}$

Maintenant, considérons le vecteur donné par $\vec{v} = x^i \vec{e}_i$. Alors nous nous proposons de démontrer que les composantes $y_1, y_2, \dots, y_n$ de $\vec{v}$ dans la base $\mathcal{F}$ sont données par:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = P \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \Leftrightarrow X = PY \quad (13.151)$$

ou:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = P^{-1} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \Leftrightarrow Y = P^{-1}X \quad (13.152)$$

Remarque: La matrice P est inversible, car ses colonnes sont linéairement indépendantes (ce sont les vecteurs $\vec{f}_i$ décomposés dans la base $\mathcal{E}$ et les $\vec{f}_i$ sont linéairement indépendants car ils forment une base).

Démonstration:

Prenons pour simplifier le cas $n = 2$ (la démonstration étant assez facilement généralisable...) avec $\mathcal{E} = (\vec{e}_1, \vec{e}_2)$ et $\mathcal{F} = (\vec{f}_1, \vec{f}_2)$.

Nous avons alors:

$$\begin{aligned}\vec{f}_1 &= p_{11}\vec{e}_1 + p_{21}\vec{e}_2 \\ \vec{f}_2 &= p_{12}\vec{e}_1 + p_{22}\vec{e}_2\end{aligned}\quad (13.153)$$

Nous avons donc $\vec{v} = x^i \vec{e}_i$ et nous cherchons à exprimer $\vec{v}$ dans la base $\mathcal{F}$ tel que $\vec{v} = y^i \vec{f}_i$. Nous allons chercher l'application linéaire qui relie ces deux relations telles que:

$$x^i \vec{e}_i = y^i \vec{f}_i \quad (13.154)$$

Soit écrit de manière explicite:

$$\begin{aligned}\vec{v} &= y_1 \vec{f}_1 + y_2 \vec{f}_2 = y_1 (p_{11}\vec{e}_1 + p_{21}\vec{e}_2) + y_2 (p_{12}\vec{e}_1 + p_{22}\vec{e}_2) \\ &= p_{11}y_1\vec{e}_1 + p_{21}y_1\vec{e}_2 + p_{12}y_2\vec{e}_1 + p_{22}y_2\vec{e}_2 = (p_{11}y_1 + p_{12}y_2)\vec{e}_1 + (p_{21}y_1 + p_{22}y_2)\vec{e}_2 \\ &= x_1\vec{e}_1 + x_2\vec{e}_2\end{aligned}\quad (13.155)$$

d'où:

$$\begin{cases} x_1 = p_{11}y_1 + p_{12}y_2 \\ x_2 = p_{21}y_1 + p_{22}y_2 \end{cases} \quad (13.156)$$

c'est-à-dire:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad (13.157)$$

Donc P est bien la matrice qui permet d'exprimer les composantes d'un vecteur d'une base en celles d'une autre base.

□C.Q.F.D.

Considérons maintenant une application $g: V^n \rightarrow V^n$ linéaire. Soit A sa matrice dans la base $\mathcal{E}$, et B sa matrice dans la base $\mathcal{F}$. Alors nous avons:

$$B = P^{-1}AP \quad (13.158)$$

Démonstration:

Reprenons:

$$\vec{v} = x^i \vec{e}_i = y^i \vec{f}_i \quad (13.159)$$

et posons:

$$g(\vec{v}) = x^i \vec{e}_i = y^j \vec{f}_j \quad (13.160)$$

nous avons donc une fonction qui nous amène à écrire:

$$\vec{x}' = A\vec{x} \quad \vec{y}' = B\vec{y} \quad (13.161)$$

D'autre part, nous avons (ce que nous avons démontré tout à l'heure):

$$\vec{x} = P\vec{y} \quad \vec{x}' = P\vec{y}' \quad (13.162)$$

Dès lors:

$$\vec{y}' = P^{-1}\vec{x}' = P^{-1}(A\vec{x}) = P^{-1}(A(P\vec{y})) = (P^{-1}AP)\vec{y} = B\vec{y} \quad (13.163)$$

d'où:

$$B = P^{-1}AP \quad (13.164)$$

et comme nous l'avons vu dans notre étude du déterminant, les déterminants de A , B sont égaux et donc invariants.

□C.Q.F.D.

5. VALEURS ET VECTEURS PROPRES

Définition: Une "**valeur propre**" est par définition (nous retrouverons cette définition dans l'introduction à l'algèbre quantique dans le cadre du chapitre de Physique Quantique Ondulatoire) une valeur λ appartenant à un corps K tel que soit une matrice carrée $A \in M_{\mathbb{M}}(K)$ nous avons:

$$\exists \vec{X} \in M_{\mathbb{M}1}(K) \quad \vec{X} \neq 0 \text{ et } A\vec{X} = \lambda\vec{X} \quad (13.165)$$

et réciproquement qu'un vecteur $\vec{X} \in M_{\mathbb{M}1}(K)$ est un "**vecteur propre**" si et seulement si:

$$\vec{X} \neq 0, \exists \lambda \in K, A\vec{X} = \lambda\vec{X} \quad (13.166)$$

L'avantage majeur de ces concepts sera la possibilité d'étudier une application linéaire, ou tout autre objet lié à une représentation matricielle, dans une représentation simple grâce à un changement de base sur laquelle la restriction de A est une simple homothétie.

En d'autres termes: lorsqu'une transformation (application d'une matrice) agit sur un vecteur, elle modifie la direction de ce vecteur excepté pour certaines matrices particulières qui ont des valeurs propres!

Ainsi, l'ensemble des valeurs propres d'une matrice $A \in M_{\mathbb{M}}(K)$ est appelé "**spectre de A** " et satisfait au système homogène:

$$(A - \lambda I_n)\vec{X} = \vec{0} \quad (13.167)$$

ou (peu importe cela revient au même!):

$$(\lambda I_n - A) \vec{X} = \vec{0} \quad (13.168)$$

où I_n (aussi notée $\mathbb{I}$) est une matrice diagonale unitaire (et donc aussi carrée) de dimension n . Ce système nous le savons (démontré plus haut) admet des solutions non triviales, donc $\vec{X} \neq 0$ ou $(\lambda I_n - A) \neq 0$, si et seulement si (nous verrons de nombreux exemples en physique):

$$\det(A - \lambda I_n) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{vmatrix} = 0 \quad (13.169)$$

Le déterminant $\det(A - \lambda I_n)$ est donc un polynôme en λ de degré n et peut donc avoir aux maximum n solutions/valeurs propres comme nous l'avons démontré lors de notre étude des polynômes (cf. chapitre de Calcul Algébrique) et est appelé "**polynôme caractéristique**" de A et l'équation $\det(A - \lambda I_n) = 0$ "**équation caractéristique de A** " ou "**équations aux valeurs propres**".

Pour la petite parenthèse, il est sympathique de remarquer que nous avons toujours dans le développement du $\det(A - \lambda I_n)$ la trace de la matrice $\text{tr}(A)$ et le déterminant $\det(A)$ qui apparaissent. Voyons deux exemples de cela:

$$\det(A - \lambda I_2) = \begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = \lambda^2 - \underbrace{(a_{11} + a_{22})}_{\text{tr}(A)} \lambda + \underbrace{(a_{11}a_{22} - a_{21}a_{12})}_{\det(A)} \quad (13.170)$$

et:

$$\begin{aligned} \det(A - \lambda I_3) &= \begin{vmatrix} a_{11} - \lambda & a_{12} & a_{13} \\ a_{21} & a_{22} - \lambda & a_{23} \\ a_{31} & a_{32} & a_{33} - \lambda \end{vmatrix} \\ &= -\lambda^3 + \underbrace{(a_{11} + a_{22} + a_{33})}_{\text{tr}(A)} \lambda^2 - (a_{11}a_{33} + a_{22}a_{33} + a_{11}a_{22} - a_{32}a_{23} - a_{12}a_{21} - a_{31}a_{13}) \lambda \\ &\quad + \underbrace{(a_{11}a_{22}a_{33} - a_{11}a_{32}a_{23} - a_{12}a_{21}a_{33} + a_{31}a_{23}a_{12} + a_{13}a_{21}a_{32} - a_{13}a_{31}a_{22})}_{\det(A)} \end{aligned} \quad (13.171)$$

Si nous regardons $(\lambda I_n - A)$ comme une application linéaire f , puisque ce sont les solutions non triviales qui nous intéressent, nous pouvons alors dire que les valeurs propres sont les éléments λ tels que:

$$\text{Ker}(f) = \text{Ker}(\lambda I_n - A) \neq \{\vec{0}\} \quad (13.172)$$

et que le Kernel constitue l'espace propre de A de la valeur propre λ dont les éléments non nuls sont les vecteurs propres!

En mathématiques, le concept de vecteur propre est une notion algébrique qui s'applique donc à une application linéaire d'un espace dans lui-même. Il correspond à l'étude des axes privilégiés, selon

lesquels l'application se comporte comme une dilatation, multipliant les vecteurs par une même constante. Ce rapport de dilatation/homothétie est donc la valeur propre, les vecteurs auxquels il s'applique vecteurs propres, réunis en un "espace propre".

Une autre manière de voir la chose:

- Un vecteur est dit "vecteur propre" par une application linéaire s'il est non nul et si l'application ne fait que modifier sa taille sans changer sa direction.
- Une "valeur propre" associée à un "vecteur propre" est le facteur de modification de taille, c'est à dire le nombre par lequel il faut multiplier le vecteur pour obtenir son image. Ce facteur peut être négatif (renversement du sens du vecteur) ou nul (vecteur transformé en un vecteur de longueur nulle).
- Un "espace propre" associé à une "valeur propre" est l'ensemble des vecteurs propres qui ont une même valeur propre et le vecteur nul. Ils subissent tous la multiplication par le même facteur.

Remarque: En mécanique, nous étudions les fréquences propres et les modes propres des systèmes oscillants (cf. chapitre de Mécanique Ondulatoire). En analyse fonctionnelle, une fonction propre est un vecteur propre pour un opérateur linéaire, c'est-à-dire une application linéaire agissant sur un espace de fonctions cf. chapitre d'Analyse Fonctionnelle). En géométrie ou en optique, nous parlons de directions propres pour rendre compte de la courbure des surfaces (cf. chapitre de Géométrie Non- Euclidiennes). En théorie des graphes, une valeur propre est simplement une valeur propre de la matrice d'adjacence du graphe (cf. chapitre de Théorie Des Graphes).

5.1. MATRICES DE ROTATION

Maintenant que nous avons vu ce qu'était une valeur et un vecteur propre, revenons sur un type particulier de matrices orthogonales qui nous seront particulièrement utiles dans notre étude des quaternions (cf. chapitre sur les Nombres), des groupes et symétries (cf. chapitre d'Algèbre Ensembliste) et de la physique des particules (cf. chapitre de Physique des Particules Élémentaires).

Nous notons, selon ce qui a été vu dans le chapitre d'Algèbre Ensembliste, $O(n)$ l'ensemble des matrices $n \times n$ à coefficients dans $\mathbb{R}$ orthogonales, c'est-à-dire vérifiant:

$$A^T A = \mathbb{1} \quad (13.173)$$

que nous notons aussi pour rappel:

$$A^T A = I \quad (13.174)$$

Les colonnes et les lignes d'une matrice orthogonale forment des bases orthonormées de $\mathbb{R}^n$ pour le produit scalaire habituel.

Le déterminant d'une matrice orthogonale vaut ± 1 , en effet $A^T A = I$ entraîne:

$$\det(A^T) \det(A) = \det(A^T A) = \det(I) = 1 \quad (13.175)$$

Nous notons $SO(n)$ l'ensemble des matrices orthogonales de déterminant 1. Montrons en trois points que si $A \in SO(3, \mathbb{R})$ alors A est la matrice d'une rotation par rapport à un axe passant par l'origine.

1. Toute valeur propre d'une matrice de rotation A (réelle ou complexe) est de module 1. En d'autres termes, la rotation conserve la norme:

En effet, si λ est une valeur propre de vecteur propre $\vec{x}$, nous avons:

$$|\lambda|^2 \|\vec{x}\|^2 = \|A\vec{x}\|^2 = \langle A\vec{x} | A\vec{x} \rangle = \langle \vec{x} | A^T A \vec{x} \rangle = \langle \vec{x} | \vec{x} \rangle = \|\vec{x}\|^2 \quad (13.176)$$

ou en notant le produit scalaire avec la notation habituelle du site:

$$|\lambda|^2 \|\vec{x}\|^2 = \|A\vec{x}\|^2 = A\vec{x} \circ A\vec{x} = \vec{x} \circ A^T A \vec{x} = \vec{x} \circ \vec{x} = \|\vec{x}\|^2 \quad (13.177)$$

donc $|\lambda| = 1$.

2. Il existe une droite dans l'espace qui sert d'axe de rotation et tout vecteur sur cette droite ne subit aucune rotation:

Notons $\vec{x}$ un vecteur propre normé de valeur propre 1 (c.à.d un vecteur tel que $A\vec{x} = \vec{x}$). Comme le lecteur l'aura peut-être compris (lire jusqu'au bout!), la droite engendrée par $\vec{x}$ que l'on notera $\langle \vec{x} \rangle$ constitue notre axe de rotation.

En effet, tout vecteur sur $\langle \vec{x} \rangle$ est envoyé sur lui-même par A . Dans ce cas l'espace orthogonal noté $\langle \vec{x} \rangle^\perp$ qui est de dimension deux est le plan perpendiculaire à l'axe de rotation.

3. Tout vecteur perpendiculaire à l'axe de rotation reste, après une rotation, perpendiculaire à cet axe. En d'autres termes, $\langle \vec{x} \rangle^\perp$ est invariant par A

En effet, si $\vec{w} \in \langle \vec{x} \rangle^\perp$ alors, $\vec{w} = A^T A \vec{w} = A^T \vec{w}$ et pour tout $\vec{y} \in \langle \vec{x} \rangle^\perp$:

$$A\vec{y} \circ \vec{w} = \vec{y} \circ A^T \vec{w} = \vec{y} \circ \vec{w} = 0 \quad (13.178)$$

c'est-à-dire $A\vec{y} \in \langle \vec{x} \rangle^\perp$. Donc $\langle \vec{x} \rangle^\perp$ est invariant par A .

En fin de compte, la restriction de A à l'espace $\langle \vec{x} \rangle^\perp$ est une rotation.

Exemple:

Soit $e^{i\alpha}$ (voir le chapitre sur les nombres où la rotation par les complexes est démontré) une valeur propre (dont le module est de 1 comme nous l'avons vu lors de notre étude des nombres complexes) de A restreinte à $\langle \vec{x} \rangle^\perp$.

Notons $\vec{w} = \vec{u} + i\vec{v}$ un vecteur propre avec $\vec{u}, \vec{v} \in \mathbb{R}^2$ de sorte que:

$$A\vec{w} = e^{i\alpha} \cdot \vec{w} \quad (13.179)$$

avec (comme nous l'avons déjà montré dans notre étude des nombres complexes):

$$A = \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix} \quad (13.180)$$

où nous savons de par notre étude des nombres complexes, que les vecteurs $\vec{u}, \vec{v}$ forment une base orthogonale (pas nécessairement normée!) de $\langle \vec{X} \rangle^\perp$.

Remarque: Il est par ailleurs aisé de vérifier que cette matrice est orthogonale (si ce n'est pas le cas contactez-nous et ce sera détaillé!).

5.2. THÉORÈME SPECTRAL

Voyons maintenant un théorème très important relativement aux valeurs et vecteurs propres qui se nomme le "**théorème spectral**" qui nous sera très utile à nouveau en physique et en statistiques.

Pour résumer, les mathématiciens disent dans leur langage que le théorème spectral permet d'affirmer la diagonalisabilité d'endomorphismes (de matrices) et justifie également la décomposition en valeurs propres.

Pour simplifier la démonstration, nous ne traitons ici que les matrices réelles en évitant un maximum le langage des mathématiciens.

Nous noterons dans un premier temps $M_n(\mathbb{R})$ l'ensemble des matrices $n \times n$ à coefficients réels. Nous confondons la matrice $M \in M_n(\mathbb{R})$ avec l'application linéaire induite sur l'espace vectoriel $\mathbb{R}^n$ par $\vec{v} \mapsto M\vec{v}$ ($\vec{v} \in \mathbb{R}^n$).

Rappel: Nous avons vu lors de l'étude des changements de base que si $(\vec{e}_1, \dots, \vec{e}_n)$ est une base de $\mathbb{R}^n$ et $M \in M_n(\mathbb{R})$ alors la matrice de l'application linéaire M dans la base $(\vec{e}_1, \dots, \vec{e}_n)$ est $S^{-1}MS$ où S est la matrice formée par les vecteurs colonnes $\vec{e}_1, \dots, \vec{e}_n$.

D'abord, nous vérifions simplement que si A est une matrice symétrique alors:

$$(A\vec{u}) \circ \vec{v} = \vec{u} \circ A^T \vec{v} = \vec{u} \circ A\vec{v} \quad (13.181)$$

Nous nous proposons maintenant d'étudier les propriétés suivantes d'une matrice M symétrique:

P1. Toutes les valeurs propres de M sont réelles.

Démonstration:

Soit:

$$\vec{z} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} \in \mathbb{C}^n \quad (13.182)$$

un vecteur propre a priori complexe de valeur propre $\lambda \in \mathbb{C}$. Notons:

$$\vec{z} = \begin{pmatrix} \bar{z}_1 \\ \vdots \\ \bar{z}_n \end{pmatrix} \quad (13.183)$$

le vecteur conjugué de $\vec{z}$. Nous avons alors:

$$\vec{z}^T M \vec{z} = \vec{z}^T \lambda \vec{z} = \lambda \|\vec{z}\|^2 \quad (13.184)$$

D'autre part vu que $M = \overline{M}^T$ nous avons:

$$\vec{z}^T M \vec{z} = \vec{z}^T \overline{M}^T \vec{z} = \overline{\vec{z}^T M \vec{z}} = \overline{(\vec{z}^T M \vec{z})} = \overline{\lambda \vec{z}^T \vec{z}} = \overline{\lambda} \|\vec{z}\|^2 \quad (13.185)$$

Etant donné que $\vec{z} \neq \vec{0}$ nous avons $\lambda = \overline{\lambda}$ et par suite, $\lambda \in \mathbb{R}$.

□C.Q.F.D.

P2. Deux espaces propres de M relatifs à des valeurs propres différentes sont orthogonaux (en d'autres termes, les vecteurs propres sont indépendants).

Démonstration:

Soit λ, μ deux valeurs propres distinctes de vecteurs propres correspondants $\vec{u}, \vec{v}$. Nous avons (ne pas oublier que M est symétrique!):

$$\lambda (\vec{u} \circ \vec{v}) = (\lambda \vec{u}) \circ \vec{v} = (M \vec{u}) \circ \vec{v} = \vec{u} \circ (M \vec{v}) = \vec{u} \circ (\mu \vec{v}) = \mu (\vec{u} \circ \vec{v}) \quad (13.186)$$

ainsi:

$$(\lambda - \mu) \cdot (\vec{u} \circ \vec{v}) = 0 \quad (13.187)$$

ce qui entraîne:

$$(\vec{u} \circ \vec{v}) = 0 \quad (13.188)$$

□C.Q.F.D.

Avant d'aller plus loin, il nous faut aussi démontrer que si $M \in M_n(\mathbb{R})$ est une matrice symétrique et V un sous-espace vectoriel de $\mathbb{R}^n$ invariant par M (c'est-à-dire qui vérifie pour tout $\vec{v} \in V$: $M \vec{v} \in V$) alors nous avons les propriétés suivantes:

P1. L'orthogonal de V noté $V^\perp$ (obtenu par la méthode de Gram-Schmidt vue dans le chapitre de calcul vectoriel) est aussi invariant par M .

Démonstration:

Soit $\vec{v} \in V$ et $\vec{w} \in V^\perp$ alors:

$$(M\vec{w}) \circ \vec{v} = \vec{w} \circ \left(\frac{M\vec{v}}{\varepsilon V} \right) = 0 \quad (13.189)$$

ce qui montre que $M\vec{w} \in V^\perp$.

□C.Q.F.D.

P2. Si $(\vec{w}_1, \dots, \vec{w}_k)$ est une base orthonormale de $V^\perp$ alors la matrice de la restriction de M à $V^\perp$ dans la base $(\vec{w}_1, \dots, \vec{w}_k)$ est aussi symétrique.

Démonstration:

Notons $A = (a_{ij})_{1 \leq i, j \leq k}$ la matrice de la restriction de M à $V^\perp$ dans la base $(\vec{w}_1, \dots, \vec{w}_k)$. Nous avons par définition pour tout $j = 1..k$ (puisque le vecteur résultant d'une application linéaire comme M peut s'exprimer dans sa base):

$$M\vec{w}_j = \sum_{i=1}^k a_{ij} \vec{w}_i \quad (13.190)$$

Or:

$$(M\vec{w}_j) \circ \vec{w}_m = \left(\sum_{i=1}^k a_{ij} \vec{w}_i \right) \circ \vec{w}_m = \sum_{i=1}^k a_{ij} (\vec{w}_i \circ \vec{w}_m) = a_{mj} \quad (13.191)$$

car $\vec{w}_i \circ \vec{w}_m = 0$ si $i \neq m$ dans la base orthonormale.

D'un autre coté:

$$(M\vec{w}_j) \circ \vec{w}_m = \vec{w}_j \circ (M\vec{w}_m) = \vec{w}_j \circ \left(\sum_{i=1}^k a_{im} \vec{w}_i \right) = \sum_{i=1}^k a_{im} (\vec{w}_j \circ \vec{w}_i) = a_{jm} \quad (13.192)$$

Donc $a_{m,j} = a_{j,m}$ ce qui montre que $A = A^T$.

□C.Q.F.D.

Nous allons à présent pouvoir montrer que toute matrice symétrique $M \in M_n(\mathbb{R})$ est diagonalisable. C'est-à-dire qu'il existe une matrice inversible S telle que $S^{-1}MS$ soit diagonale.

Remarque: En fait nous verrons, pour être plus précis, qu'il existe S orthogonale telle que $S^{-1}MS$ soit diagonale.

Rappel: S orthogonale signifie que $SS^T = I$ (où I est la matrice identité) ce qui équivaut à dire que les

colonnes de S forment une base orthonormale de $\mathbb{R}^n$.

Donc allons-y et pour cela considérons $M \in M_n(\mathbb{R})$ une matrice symétrique. Alors nous souhaitons démontrer qu'il existe une matrice S orthogonale telle que $S^{-1}MS$ soit diagonale (en d'autres termes, il existe une base où M est diagonalisable).

Démonstration:

Nous prouvons l'affirmation par récurrence sur n . Si $n = 1$ il n'y a rien à montrer. Supposons que l'affirmation soit vérifiée pour $k \leq n$ et prouvons là pour $k = n+1$. Soit donc $M \in M_{n+1}(\mathbb{R})$ une matrice symétrique et λ une valeur propre de M .

Nous vérifions facilement que l'espace propre $W = \text{Ker}(\lambda I_{n+1} - M)$ est invariant par M (il suffit de prendre n'importe quelle application numérique) et que par la démonstration vue plus haut que $W^\perp$ est aussi invariant par M . De plus, nous savons (cf. chapitre de calcul vectoriel), que $\mathbb{R}^{n+1}$ se décompose en somme directe $\mathbb{R}^{n+1} = W \oplus W^\perp$.

Si $\dim(W^\perp) = 0$, $\mathbb{R}^{n+1} = W$ et il suffit de prendre une base orthonormale de W pour diagonaliser M . En effet si $(\vec{w}_1, \dots, \vec{w}_{n+1})$ est une telle base, la matrice S formée par les vecteurs colonnes $\vec{w}_j$ ($j = 1..n+1$) est orthogonale et vérifie:

$$S^{-1}MS = S^{-1}\lambda S = \lambda I_{n+1} \quad (13.193)$$

$S^{-1}MS$ est bien diagonale.

Supposons donc $\dim(W^\perp) > 0$ et soit $(\vec{u}_1, \dots, \vec{u}_m)$ avec $m \leq n$ une base orthonormale de $W^\perp$. Notons A la matrice de la restriction de M à $W^\perp$ dans la base $(\vec{u}_1, \dots, \vec{u}_m)$. A est aussi symétrique (selon la démonstration d'une des propriétés précédentes).

Par hypothèse de récurrence il existe une matrice $H \in M_m(\mathbb{R})$ orthogonale telle que $H^{-1}AH$ soit diagonale.

Notons par $(\vec{w}_1, \dots, \vec{w}_{n+1-m})$ une base orthonormale de W et G la matrice formée par les vecteurs colonnes $\vec{w}_1, \dots, \vec{w}_{n+1-m}, \vec{u}_1, \dots, \vec{u}_m$. Alors, nous pouvons écrire que:

$$G^{-1}MG = \begin{bmatrix} \lambda I_{n+1-m} & 0 \\ 0 & A \end{bmatrix} \quad (13.194)$$

et G est aussi orthogonale par construction.

Considérons la matrice par blocs (matrice composée de matrices) suivante:

$$L = \begin{bmatrix} I_{n+1-m} & 0 \\ 0 & H \end{bmatrix} \in M_{n+1}(\mathbb{R}) \quad (13.195)$$

et posons $S = G \cdot L$. Il est évident que S est orthogonale car G et L le sont. Effectivement, si $G \cdot G^T = I$

et $L \cdot L^T = I$ alors (ne pas oublier que la multiplication matricielle est associative):

$$(GL)(GL)^T = (GL)(L^T G^T) = G(LL^T)G^T = GIG^T = GG^T = I \quad (13.196)$$

De plus S vérifie :

$$\begin{aligned} S^{-1}MS &= L^{-1}(G^{-1}MG)L \\ &= \begin{bmatrix} I_{n+1-m} & 0 \\ 0 & H \end{bmatrix}^{-1} \begin{bmatrix} \lambda I_{n+1-m} & 0 \\ 0 & A \end{bmatrix} \begin{bmatrix} I_{n+1-m} & 0 \\ 0 & H \end{bmatrix} = \begin{bmatrix} \lambda I_{n+1-m} & 0 \\ 0 & H^{-1}AH \end{bmatrix} \end{aligned} \quad (13.197)$$

Et alors:

$$\begin{bmatrix} \lambda I_{n+1-m} & 0 \\ 0 & H^{-1}AH \end{bmatrix} \quad (13.198)$$

est bien diagonale.

□C.Q.F.D.

Pour finir voici donc le "théorème spectral" (cas réel): Si $M \in M_n(\mathbb{R})$ une matrice symétrique alors il existe une base orthonormale formée de vecteurs propres de M .

Démonstration:

Nous avons donc vu dans les paragraphes précédents qu'il existe S orthogonale telle que $S^{-1}MS$ soit diagonale. Notons $\vec{c}_1, \dots, \vec{c}_n$ les colonnes de S . $(\vec{c}_1, \dots, \vec{c}_n)$ est une base orthonormale de $\mathbb{R}^n$ car S est orthogonale. Notant e_i le i -ème vecteur de la base canonique de $\mathbb{R}^n$ et λ_i le i -ème coefficient diagonal de $S^{-1}MS$ nous avons sans supposer directement que λ_i est une valeur propre pour l'instant:

$$\lambda_i \vec{e}_i = (S^{-1}MS) \vec{e}_i = S^{-1}M \vec{c}_i \quad (13.199)$$

en multipliant par S des deux côtés de l'égalité nous avons:

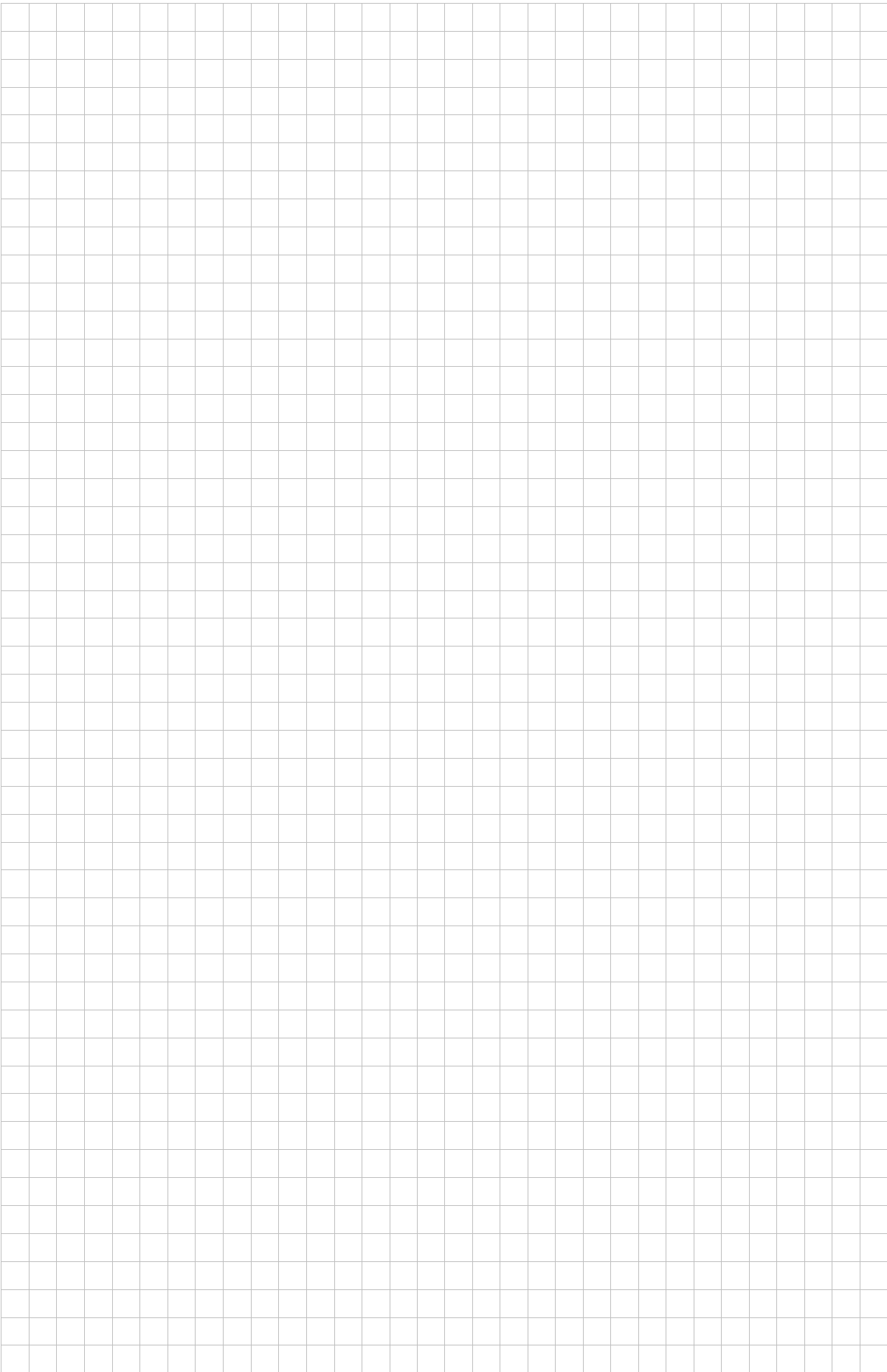
$$\lambda_i S \vec{e}_i = M \vec{c}_i \quad (13.200)$$

et donc:

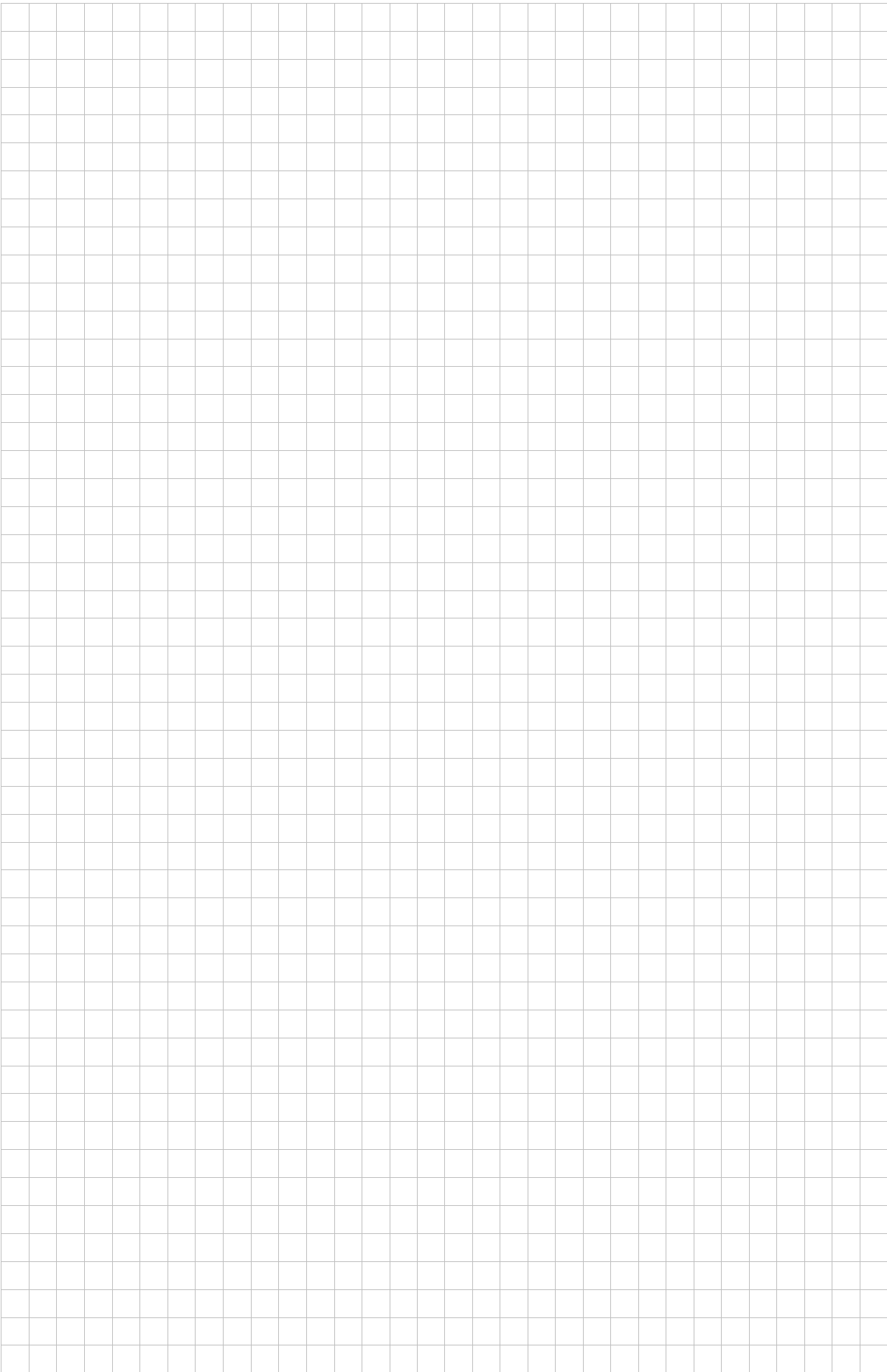
$$\lambda_i \vec{c}_i = M \vec{c}_i$$

ce qui montre que $\vec{c}_1, \dots, \vec{c}_n$ sont des vecteurs propres et λ_i les valeurs propres.

□C.Q.F.D.







14. CALCUL TENSORIEL

Le calcul vectoriel classique est une technique simple et efficace qui s'adapte parfaitement à l'étude des propriétés mécaniques et physiques de la matière dans l'espace euclidien à trois dimensions. Cependant, dans de nombreux domaines de la physique, il apparaît des grandeurs expérimentales qui ne peuvent plus être facilement représentées par de simples vecteurs-colonnes d'espaces vectoriels euclidiens. C'est le cas par exemple en mécanique des milieux continus, fluides ou solides, en électromagnétisme, relativité générale, etc.

Ainsi, dès la fin du 19^{ème} siècle, l'analyse des forces qui s'exercent à l'intérieur d'un milieu continu a conduit à mettre en évidence des grandeurs physique caractérisées par neuf nombres représentant les forces de pression ou de tension internes (cf. chapitre de Mécanique Des Milieux Continus). La représentation de ces grandeurs nécessita l'introduction d'un nouvel être mathématique qui fut appelé "tenseur", par référence à son origine physique. Par la suite, à partir de 1900, ce furent R. Ricci et T. Levi-Civita qui développèrent le calcul tensoriel puis l'étude des tenseurs permit un approfondissement de la théorie des espaces vectoriels et contribua au développement de la géométrie différentielle (voir chapitre du même nom).

Le calcul tensoriel, appelé aussi parfois "**géométrie différentielle absolue**" a également pour avantage de se libérer de tous les systèmes de coordonnées et les résultats des développements mathématiques sont ainsi invariants (énorme allégement des calculs). Il n'y a plus alors à se préoccuper dans quel référentiel il convient de travailler et cela, est très intéressant en relativité générale.

Nous conseillons par ailleurs vivement au lecteur de bien maîtriser les bases du calcul vectoriel et de l'algèbre linéaire comme elles ont été présentées auparavant. Au besoin, nous avons choisi lors de la rédaction de ce chapitre de revenir sur certains points vus dans le chapitre de Calcul Vectoriel (composantes covariantes, contravariantes,...).

Par ailleurs, si le lecteur a déjà parcouru l'étude des contraintes dans les solides (cf. chapitre de Mécanique Des Milieux Continus) ou du tenseur de Faraday (cf. chapitre d'Électrodynamique) ou du tenseur d'énergie-impulsion (cf. chapitre de Relativité Générale) ceci constituera un avantage pratique certain avant de parcourir ce qui va suivre. Par ailleurs, la rédaction des objets susmentionnés a été faite de telle manière que la notion de tenseur y soit introduite si possible (...) intuitivement.

Nous ne ferons que très peu d'exemples pratiques dans cette section. Effectivement les exemples concrets, vous l'aurez compris, viendront lorsque nous étudierons la mécanique des milieux continus, la relativité générale, la physique quantique des champs, l'électrodynamique, etc.

Un conseil peut-être: pensez matriciel, écrivez tensoriel! (vous comprendrez mieux ce petit adage une fois après avoir parcouru tout ce chapitre).

1. TENSEUR

Définition (simpliste): Les "**tenseurs**" sont des objets mathématiques généralisant les notions de vecteurs et de matrices. Ils ont été introduits, en physique, pour représenter l'état de contrainte et de déformation d'un volume soumis à des forces, d'où leur nom (tensions).

La définition rigoureuse nécessite (je pense personnellement) d'avoir d'abord lu le présent chapitre dans son intégralité. Mais sachez qu'au fait un tenseur est grosso modo comme un déterminant... (cf. chapitre d'Algèbre Linéaire). Eh oui! C'est simplement une application multilinéaire sur un espace de dimension donnée (correspondant au nombre de colonnes de la matrice/tenseurs) qui donne finalement un scalaire (d'un corps donné).

Par exemple, nous avons démontré dans le chapitre de Mécanique Des Milieux Continus que les forces normales et tangentielles dans un fluide étaient données par la relation:

$$\begin{bmatrix} N_x \\ N_y \\ N_z \end{bmatrix} = \begin{bmatrix} \sigma_{xx} & \tau_{xy} & \tau_{xz} \\ \tau_{yx} & \sigma_{yy} & \tau_{yz} \\ \tau_{zx} & \tau_{zy} & \sigma_{zz} \end{bmatrix} \begin{bmatrix} l \\ m \\ n \end{bmatrix} \quad (14.1)$$

ce qui se notait sous la forme traditionnelle condensée suivante (où nous ne distinguons plus ce qui est tangentiel de ce qui est normal donc il y a une perte de clarté):

$$N_i = \sigma_{ij} n_j \quad (14.2)$$

Nous faisons ainsi apparaître une grandeur mathématique σ_{ij} ayant 9 composantes, alors qu'un vecteur dans le même espace $\mathbb{R}^3$ en possède 3.

Cette notion est aussi beaucoup utilisée dans le chapitre de Relativité Générale où nous avons démontré que le tenseur d'énergie-impulsion dans un cas particulièrement simple est donné par:

$$T^{ij} = \begin{bmatrix} T^{00} & T^{01} & T^{02} & T^{03} \\ T^{10} & T^{11} & T^{12} & T^{13} \\ T^{20} & T^{21} & T^{22} & T^{23} \\ T^{30} & T^{31} & T^{32} & T^{33} \end{bmatrix} = \begin{bmatrix} \rho \gamma^2 c^2 & \rho \gamma^2 c v^x & \rho \gamma^2 c v^y & \rho \gamma^2 c v^z \\ \rho \gamma^2 c v^x & \rho \gamma^2 v^x v^x & \rho \gamma^2 v^x v^y & \rho \gamma^2 v^x v^z \\ \rho \gamma^2 c v^y & \rho \gamma^2 v^y v^x & \rho \gamma^2 v^y v^y & \rho \gamma^2 v^y v^z \\ \rho \gamma^2 c v^z & \rho \gamma^2 v^z v^x & \rho \gamma^2 v^z v^y & \rho \gamma^2 v^z v^z \end{bmatrix} \quad (14.3)$$

et satisfait à la relation non moins importante de conservation:

$$\nabla_\mu T^{\mu\nu} = 0 \quad (14.4)$$

Ou sinon, toujours dans le chapitre de Relativité Générale, nous avons démontré que le tenseur de la métrique de Schwarzschild est:

$$g^{\mu\nu} = \begin{bmatrix} 1/A(r) & 0 & 0 & 0 \\ 0 & -1/B(r) & 0 & 0 \\ 0 & 0 & -1/r^2 & 0 \\ 0 & 0 & 0 & -1/(r^2 \sin^2 \theta) \end{bmatrix} \quad (14.5)$$

et donne donc l'équation de la métrique (cf. chapitre de Calcul Différentiel):

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu \quad (14.6)$$

Signalons également que dans le chapitre de Relativité Restreinte nous avons démontré que le tenseur de transformation de Lorentz est donné par:

$$\begin{pmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{pmatrix} = \begin{bmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{pmatrix} \quad (14.7)$$

qui sous forme condensée donne la transformation de composantes suivantes:

$$x'^{\mu} = L^{\mu}_{\nu} x^{\nu} \quad (14.8)$$

En ce qui concerne la transformation du champ électromagnétique nous avons également démontré que le tenseur de Faraday est donné par:

$$F^{\mu\nu} = \begin{bmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & -B_z & B_y \\ E_y & B_z & 0 & -B_x \\ E_z & -B_y & B_x & 0 \end{bmatrix} \quad (14.9)$$

et permet donc de passer d'un référentiel à un autre à l'aide de la relation:

$$F'^{\mu\nu} = L^{\mu}_{\alpha} L^{\nu}_{\beta} F^{\alpha\beta} \quad (14.10)$$

Mais ce sont des tenseurs très simples qui peuvent être représentés sous formes de matrices. Il faut également savoir que ce n'est pas parce qu'une lecture d'une variable avec des indices semble indiquer que nous avons affaire à un tenseur que cela en est forcément un. Par exemple, la relation fameuse (très utilisée dans le chapitre de Relativité Générale):

$$\Gamma^i_{kj} = g^{il} \Gamma_{klj} = \frac{1}{2} g^{il} (\partial_k g_{lj} + \partial_j g_{kl} - \partial_l g_{kj}) \quad (14.11)$$

pourrait faire croire que le premier membre tout à gauche est un tenseur mais au fait il n'est rien... ce n'est qu'un symbole... d'où son nom: symbole de Christoffel (et non pas: tenseur de Christoffel).

L'intérêt des tenseurs en physique est que leurs caractéristiques sont indépendantes des coordonnées choisies. Ainsi, une relation entre tenseur dans une base sera vraie quelle que soit la base utilisée par la suite. C'est une caractéristique fondamentale pour la Relativité Générale!

2. NOTATION INDICIELLE

Nous utilisons par la suite des symboles mathématiques: coordonnées, composantes de vecteurs et tenseurs, éléments de matrice, etc., dont le nombre, dans chaque catégorie, est grand ou indéterminé. Pour distinguer les divers symboles d'une catégorie nous employons des indices. Par exemple, au lieu des variables traditionnelles x, y, z nous utiliserons éventuellement les grandeurs x_1, x_2, x_3 (comme nous l'avons déjà fait en algèbre linéaire). Cette notation devient indispensable lorsque nous avons des variables en nombre indéterminé.

Ainsi, si nous avons n variables, nous les noterons: $x_1, x_2, x_3, \dots, x_n$

Nous utilisons également des indices supérieurs, selon les besoins; par exemple, x^1, x^2, x^3 . Afin d'éviter toute confusion avec l'écriture des puissances, la quantité x^i à la puissance p sera écrite $(x^i)^p$. Lorsque le contexte écarte tout risque d'ambiguïté, l'utilisation des parenthèses n'est cependant pas fondamentalement nécessaire.

En calcul tensoriel il existe une convention de sommation qui consiste à utiliser le fait que l'indice répété, ici l'indice i , va devenir lui-même l'indication de la sommation. Nous écrivons alors, avec cette convention:

$$\sum_{i=1}^n x^i y^i = x^i y^i = x^1 y^1 + x^2 y^2 + \dots + x^n y^n \quad (14.12)$$

ce qui permet de condenser relativement bien les écritures.

Ainsi, pour représenter le système linéaire:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned} \quad (14.13)$$

nous écrirons (remarquez bien comment s'écrivent les composantes de la matrice associée!):

$$a_{ij}x_j = b_i \quad (14.14)$$

en spécifiant que c'est pour $n = 3$.

Nous voyons sur cet exemple, combien la convention de sommation permet une écriture condensée et donc puissante.

La convention de sommation s'étend à tous les symboles mathématiques comportant des indices répétés. Ainsi la décomposition d'un vecteur $\vec{x}$ sur une base $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ s'écrit pour $n = 3$ dès lors:

$$\vec{x} = x^1 \vec{e}_1 + x^2 \vec{e}_2 + x^3 \vec{e}_3 = x^i \vec{e}_i \quad (14.15)$$

En résumé, toute expression qui comporte un indice deux fois répété représente une somme sur toutes les valeurs possibles de l'indice répété.

Remarque: Nous nommons, pour des raisons évidentes que nous détaillerons plus loin, x^i la "composante contravariante" du vecteur $\vec{x}$.

2.1. SOMMATION SUR PLUSIEURS INDICES

La convention de sommation (due à Einstein) s'étend au cas où figurent, en règle générale, plusieurs indices répétés en positions supérieure et inférieure dits "**indices muets**" dans un même monôme (souvent les physiciens omettent la règle de les mettre en position opposées comme ce sera aussi le cas souvent sur ce site!). Soit par exemple, la quantité $A_j^i x^i y^j$, celle-ci représente la somme suivante pour i et j prenant les valeurs de 1 à 2:

$$A_j^i x^i y^j = A_1^1 x^1 y^1 + A_2^1 x^1 y^2 + A_1^2 x^2 y^1 + A_2^2 x^2 y^2 \quad (14.16)$$

Ainsi, nous voyons facilement qu'une expression avec deux indices de sommation qui prennent respectivement les valeurs $1, 2, \dots, n$ comportera n^2 termes; n^3 s'il y a trois indices, de sommation etc.

Il faut faire cependant attention aux substitutions avec ce genre de notation car si nous supposons que nous avons la relation:

$$A = a_{ij} x^i y^j \text{ avec } x^i = c_{ij} y^j \quad (14.17)$$

alors pour obtenir l'expression de A uniquement en fonction des variables y^j , nous ne pouvons pas écrire:

$$A = a_{ij} c_{ij} y^j y^j \quad (14.18)$$

car cela ne revient pas à la même expression après développement puisque les indices muets sont systématiquement sommés de manières identiques et rigides (nous laissons au lecteur le soin de faire ce petit exercice de style). En d'autres termes, un même indice muet ne peut pas être répété plus de 2 fois.

2.2. SYMBOLE DE KRONECKER

Ce symbole introduit par le mathématicien Kronecker, est le suivant (souvent utilisé en physique en général dans de nombreux domaines):

$$\delta_{ij} = \delta^i_j = \delta_j^i = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases} \quad (14.19)$$

Ce symbole est appelé "**symbole de Kronecker**". Il permet avantageusement d'écrire, par exemple, le produit scalaire de deux vecteurs $\vec{e}_i$ et $\vec{e}_j$, de norme unité et orthogonaux entre eux, sous la forme:

$$\vec{e}_i \circ \vec{e}_j = \delta_{ij} \quad (14.20)$$

Lors d'une sommation portant sur deux indices muets, le symbole de Kronecker annule tous les termes où les indices ont des valeurs différentes. Par exemple:

$$\delta_{ij} y_i y_j = y_i y_i \quad (14.21)$$

Nous retrouverons ce symbole dans de nombreux exemples de physique théorique (physique quantique ondulatoire, physique quantique des champs, relativité générale, mécanique des fluides, etc.).

Précisons qu'il existe une version généralisée du symbole de Kronecker:

$$\delta_{mn...p}^{ij...k} = \begin{bmatrix} \delta_m^i & \delta_n^i & \dots & \delta_p^i \\ \delta_m^j & \delta_n^j & \dots & \delta_p^j \\ \vdots & \vdots & \ddots & \vdots \\ \delta_m^k & \delta_n^k & \dots & \delta_p^k \end{bmatrix} \quad (14.22)$$

Nous avons aussi par exemple:

$$\delta_{mnp}^{ijk} = \begin{bmatrix} \delta_m^i & \delta_n^i & \delta_p^i \\ \delta_m^j & \delta_n^j & \delta_p^j \\ \delta_m^k & \delta_n^k & \delta_p^k \end{bmatrix} = \varepsilon^{ijk} \varepsilon_{mnp} \quad (14.23)$$

2.3. SYMBOLE D'ANTISYMETRIE

Un autre symbole fort utile est le "**symbole d'antisymétrie**" ou appelé aussi "**tenseur d'antisymétrie**" que nous retrouverons en Électrodynamique, en Relativité Générale et en Physique Quantique Relativiste.

Dans le cas où i, j, k prennent l'une des valeurs $\{1,2,3\}$ le symbole d'antisymétrie ε^{ijk} aura les valeurs définies suivantes:

- $\varepsilon^{ijk} = 0$, si deux quelconques des indices ou plus ont une valeur identique.
- $\varepsilon^{ijk} = 1$, si les indices sont dans l'ordre 1, 2, 3 ou proviennent d'un nombre pair de permutations des indices par rapport à l'ordre initial des indices.
- $\varepsilon^{ijk} = -1$, si les indices sont dans un ordre qui provient d'un nombre impair de permutations par rapport à l'ordre initial des indices.

Remarque: Pour se rappeler si nous avons une permutation paire (respectivement impaire) d'indices, il suffit (dans le cas particulier de 3 indices) d'observer si nous retrouvons la séquence dans la suite 123123 (respectivement 321321). Enfin, rappelons selon ce qui a été vu dans le chapitre de Probabilité, avec n indices il y a donc $n!$ permutations possibles.

En utilisant ce symbole, un déterminant d'ordre deux (cf. chapitre d'Algèbre Linéaire) s'écrit alors sous la forme avantageuse:

$$\det [a^i] = \varepsilon^{ij} a_i^1 a_j^2 \quad (14.24)$$

et le produit vectoriel (et ça c'est très pratique en relativité générale et en électrodynamique):

$$\vec{A} \times \vec{B} = \varepsilon_{ijk} A_j B_k \quad (14.25)$$

où bien sûr, j et k sont sommés et où l'indice muet i est le numéro de la ligne du vecteur résultant (en cas de demande nous ferons les développements). En particulier, le rotationnel d'un champ vectoriel est alors:

$$\vec{\nabla} \times \vec{A} = \varepsilon_{ijk} \frac{\partial A_k}{\partial x_j} = \varepsilon_{ijk} \partial_j A_k \quad (14.26)$$

Comme exemple, calculons en notation indicielle le double produit vectoriel $\vec{A} \times \vec{B} \times \vec{C}$:

$$\begin{aligned} \vec{A} \times \vec{B} \times \vec{C} &= \varepsilon_{ijk} \varepsilon_{kmn} A_j B_m C_n = (\delta_{im} \delta_{jn} - \delta_{in} \delta_{jm}) A_j B_m C_n = B_i (A_j C_j) - C_i (A_j B_j) \\ &= \vec{B} \cdot (\vec{A} \circ \vec{C}) - \vec{C} \cdot (\vec{A} \circ \vec{B}) \end{aligned} \quad (14.27)$$

où à nouveau, l'indice muet i est le numéro de la ligne du vecteur résultant. Voyons la démonstration détaillée de ces égalités (la démonstration des égalités ci-dessous n'a pas besoin de respecter l'ordre des égalités de la relation précédente).

Démonstration:

Nous avons indirectement démontré dans le chapitre de Calcul Vectoriel l'identité suivante:

$$\vec{A} \times \vec{B} \times \vec{C} = \vec{B} \cdot (\vec{A} \circ \vec{C}) - \vec{C} \cdot (\vec{A} \circ \vec{B}) \quad (14.28)$$

Remarque: Cette dernière relation est parfois appelée la "**règle de Grassmann**", ou plus couramment "**double produit vectoriel**" et il est important de noter que sans les parenthèses le résultat n'est pas unique.

Pour démontrer la relation:

$$\vec{A} \times \vec{B} \times \vec{C} = \varepsilon_{ijk} \varepsilon_{kmn} A_j B_m C_n \quad (14.29)$$

au changement d'indices près montrons d'abord que:

$$(\vec{B} \times \vec{C})_i = \varepsilon_{ijk} B_j C_k \quad (14.30)$$

ce qui nous donne:

$$\begin{pmatrix} B_2 C_3 - B_3 C_2 \\ B_3 C_1 - B_1 C_3 \\ B_1 C_2 - B_2 C_1 \end{pmatrix} = \begin{pmatrix} \varepsilon_{1jk} B_j C_k \\ \varepsilon_{2jk} B_j C_k \\ \varepsilon_{3jk} B_j C_k \end{pmatrix} \quad (14.31)$$

Ne faisons le développement que pour la première ligne (c'est déjà suffisamment long...):

$$\begin{aligned} B_2 C_3 - B_3 C_2 &= \varepsilon_{1jk} B_j C_k = \varepsilon_{11k} B_1 C_k + \varepsilon_{12k} B_2 C_k + \varepsilon_{13k} B_3 C_k \\ &= \varepsilon_{111} B_1 C_1 + \varepsilon_{112} B_1 C_2 + \varepsilon_{113} B_1 C_3 + \varepsilon_{121} B_2 C_1 + \varepsilon_{122} B_2 C_2 + \varepsilon_{123} B_2 C_3 + \varepsilon_{131} B_3 C_1 \\ &\quad + \varepsilon_{132} B_3 C_2 + \varepsilon_{133} B_3 C_3 = \varepsilon_{123} B_2 C_3 + \varepsilon_{132} B_3 C_2 = 1 B_2 C_3 - 1 B_3 C_2 = B_2 C_3 - B_3 C_2 \end{aligned} \quad (14.32)$$

C'est ce qu'il fallait montrer.

Maintenant montrons que pour la l -ième ligne nous avons bien:

$$\left(\vec{A} \times (\vec{B} \times \vec{C}) \right)_i = \varepsilon_{im} A_m \left(\varepsilon_{ijk} B_j C_k \right)_i \quad (14.33)$$

en s'aidant d'un résultat obtenu dans le chapitre de Calcul Vectoriel (produit vectoriel de trois vecteurs différents) nous avons le premier terme (la première ligne du vecteur résultant du calcul):

$$\begin{aligned} & \left(A_2 (B_1 C_2 - B_2 C_1)_3 - A_3 (B_3 C_1 - B_1 C_3)_2 \right)_1 = \varepsilon_{1mi} A_m \left(\varepsilon_{ijk} B_j C_k \right)_i \\ &= \varepsilon_{1m1} A_m (B_2 C_3 - B_3 C_2)_1 + \varepsilon_{1m2} A_m (B_3 C_1 - B_1 C_3)_2 + \varepsilon_{1m3} A_m (B_1 C_2 - B_2 C_1)_3 \\ &= \varepsilon_{111} A_1 (B_2 C_3 - B_3 C_2)_1 + \varepsilon_{121} A_2 (B_2 C_3 - B_3 C_2)_1 + \varepsilon_{131} A_3 (B_2 C_3 - B_3 C_2)_1 \\ &+ \varepsilon_{112} A_1 (B_3 C_1 - B_1 C_3)_2 + \varepsilon_{122} A_2 (B_3 C_1 - B_1 C_3)_2 + \varepsilon_{132} A_3 (B_3 C_1 - B_1 C_3)_2 \\ &+ \varepsilon_{113} A_1 (B_1 C_2 - B_2 C_1)_3 + \varepsilon_{123} A_2 (B_1 C_2 - B_2 C_1)_3 + \varepsilon_{133} A_3 (B_1 C_2 - B_2 C_1)_3 \\ &= \varepsilon_{123} A_2 (B_1 C_2 - B_2 C_1)_3 + \varepsilon_{132} A_3 (B_3 C_1 - B_1 C_3)_2 \\ &= 1 A_2 (B_1 C_2 - B_2 C_1)_3 - 1 A_3 (B_3 C_1 - B_1 C_3)_2 = A_2 (B_1 C_2 - B_2 C_1) - A_3 (B_3 C_1 - B_1 C_3) \end{aligned} \quad (14.34)$$

Il est alors immédiat que (pour i valant 1):

$$\begin{aligned} \vec{A} \times \vec{B} \times \vec{C} &= \left(\vec{A} \times \vec{B} \times \vec{C} \right)_i = B_i (A_j C_j) - C_i (A_j B_j) = \\ &= B_i (A_1 C_1 + A_2 C_2 + A_3 C_3) - C_i (A_1 B_1 + A_2 B_2 + A_3 B_3) \\ &= B_1 A_1 C_1 + B_1 A_2 C_2 + B_1 A_3 C_3 - C_1 A_1 B_1 - C_1 A_2 B_2 - C_1 A_3 B_3 \\ &= A_2 B_1 C_2 - A_2 B_2 C_1 - A_3 B_3 C_1 + A_3 B_1 C_3 = \\ &= A_2 (B_1 C_2 - B_2 C_1) - A_3 (B_3 C_1 - B_1 C_3) \end{aligned} \quad (14.35)$$

Montrons maintenant que pour i valant 1 nous avons aussi:

$$\begin{aligned} \vec{A} \times \vec{B} \times \vec{C} &= \left(\vec{A} \times \vec{B} \times \vec{C} \right)_i = \left(\delta_{im} \delta_{jn} - \delta_{in} \delta_{jm} \right) A_j B_m C_n \\ &= A_2 (B_1 C_2 - B_2 C_1) - A_3 (B_3 C_1 - B_1 C_3) \end{aligned} \quad (14.36)$$

Effectivement:

$$\begin{aligned} & \left(\vec{A} \times \vec{B} \times \vec{C} \right)_1 = \left(\delta_{1m} \delta_{jn} - \delta_{1n} \delta_{jm} \right) A_j B_m C_n \\ &= \left(\delta_{1m} \delta_{j1} - \delta_{11} \delta_{jm} \right) A_j B_m C_1 + \left(\delta_{1m} \delta_{j2} - \delta_{12} \delta_{jm} \right) A_j B_m C_2 + \left(\delta_{1m} \delta_{j3} - \delta_{13} \delta_{jm} \right) A_j B_m C_3 \\ &= \left(\delta_{1m} \delta_{11} - \delta_{11} \delta_{1m} \right) A_1 B_m C_1 + \left(\delta_{1m} \delta_{21} - \delta_{11} \delta_{2m} \right) A_2 B_m C_1 + \left(\delta_{1m} \delta_{31} - \delta_{11} \delta_{3m} \right) A_3 B_m C_1 \\ &+ \left(\delta_{1m} \delta_{12} - \delta_{12} \delta_{1m} \right) A_1 B_m C_2 + \left(\delta_{1m} \delta_{22} - \delta_{12} \delta_{2m} \right) A_2 B_m C_2 + \left(\delta_{1m} \delta_{32} - \delta_{12} \delta_{3m} \right) A_3 B_m C_2 \\ &+ \left(\delta_{1m} \delta_{13} - \delta_{13} \delta_{1m} \right) A_1 B_m C_3 + \left(\delta_{1m} \delta_{23} - \delta_{13} \delta_{2m} \right) A_2 B_m C_3 + \left(\delta_{1m} \delta_{33} - \delta_{13} \delta_{3m} \right) A_3 B_m C_3 \\ &= \left(\delta_{1m} - \delta_{1m} \right) A_1 B_m C_1 - \delta_{2m} A_2 B_m C_1 - \delta_{3m} A_3 B_m C_1 + \delta_{1m} A_2 B_m C_2 + \delta_{1m} A_3 B_m C_3 \\ &= -\delta_{2m} A_2 B_m C_1 - \delta_{3m} A_3 B_m C_1 + \delta_{1m} A_2 B_m C_2 + \delta_{1m} A_3 B_m C_3 \\ &= -\delta_{21} A_2 B_1 C_1 - \delta_{22} A_2 B_2 C_1 - \delta_{23} A_2 B_3 C_1 - \delta_{31} A_3 B_1 C_1 - \delta_{32} A_3 B_2 C_1 - \delta_{33} A_3 B_3 C_1 \\ &+ \delta_{11} A_2 B_1 C_2 + \delta_{12} A_2 B_2 C_2 + \delta_{13} A_2 B_3 C_2 + \delta_{11} A_3 B_1 C_3 + \delta_{12} A_3 B_2 C_3 + \delta_{13} A_3 B_3 C_3 \\ &= -A_2 B_2 C_1 - A_3 B_3 C_1 + A_2 B_1 C_2 + A_3 B_1 C_3 = A_2 (B_1 C_2 - B_2 C_1) - A_3 (B_3 C_1 - B_1 C_3) \end{aligned} \quad (14.37)$$

Comme deuxième exemple, montrons comment la divergence d'un rotationnel s'annule:

$$\vec{\nabla} \circ (\vec{\nabla} \times \vec{A}) = \partial_i (\varepsilon_{ijk} \partial_j A_k) = \varepsilon_{ijk} \partial_i \partial_j A_k \quad (14.38)$$

Comme de par le théorème de Schwarz $\partial_i \partial_j$ est symétrique (donc intervenir les indices n'a aucun impact) dans les indices et que ε_{ijk} est antisymétrique (par définition) dans les mêmes indices, la somme sur i et j doit nécessairement s'annuler. Par exemple, la contribution à la somme du terme $i = 1, j = 2$ est l'opposée de celle de $i = 2, j = 1$.

Remarques:

R1. Le symbole d'antisymétrie est très souvent appelé "**tenseur de Levi-Civita**" dans la littérature. Au fait, bien que ce soit bien un tenseur dans la forme de ses notations, il s'agit plus d'un outil mathématique qu'un "être" mathématique d'où la préférence de certains physiciens de le nommer "symbole" plutôt que "tenseur". Mais c'est à vous de voir...

R2. Par abus d'écriture nous n'écrivons pas le vecteur de base mais en toute rigueur, et pour éviter de l'oublier, rappelons qu'afin d'équilibrer les membres de l'égalité et dans le souci de préciser que les vecteurs sont exprimés dans la même base, nous devrions écrire:

$$\vec{A} \times \vec{B} = \varepsilon_{ijk} A_j B_k \vec{e}_i \quad \vec{\nabla} \times \vec{A} = \varepsilon_{ijk} \partial_j A_k \vec{e}_i \quad (14.39)$$

Voyons maintenant des applications concrètes de cette notation indicielle en reprenant l'exemple du changement de base que nous avons déjà vu en calcul vectoriel:

Soient deux bases $(\vec{e}_1, \vec{e}_2, \dots, \vec{e}_n)$ et $(\vec{e}'_1, \vec{e}'_2, \dots, \vec{e}'_n)$ d'un espace vectoriel euclidien E_n . Chaque vecteur d'une base peut être décomposé sur l'autre base sous la forme d'une application linéaire (matrice de changement de base - voir chapitre d'Algèbre Linéaire):

$$\vec{e}_i = A_i^k \vec{e}'_k \quad \text{et} \quad \vec{e}'_k = A_k^i \vec{e}_i \quad (14.40)$$

où nous utilisons bien évidemment la convention de sommation pour $i, k = 1, 2, \dots, n$.

Rappelons que la matrice de changement de base (ou "**matrice transformation**") doit avoir autant de colonnes que le vecteur de base a de lignes (dimensions). Petit exemple à trois dimensions:

$$\begin{pmatrix} x'_1 \\ x'_2 \\ x'_3 \end{pmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_1 a_{11} + x_2 a_{12} + x_3 a_{13} \\ x_1 a_{21} + x_2 a_{22} + x_3 a_{23} \\ x_1 a_{31} + x_2 a_{32} + x_3 a_{33} \end{pmatrix} \quad (14.41)$$

et il est évident qu'il est bien plus sympathique d'écrire cela sous la forme:

$$x'_i = A_i^k x_k \quad (14.42)$$

où donc sur A , nous avons le k qui représente la colonne de la matrice et i la ligne.

Un vecteur quelconque $\vec{x}$ de E_n peut être décomposé (nous l'avons déjà vu dans le chapitre de Calcul Vectoriel) sur chaque base de E_n sous la forme:

$$\vec{x} = x^i \vec{e}_i = x^{ik} \vec{e}'_k \quad (14.43)$$

Si nous cherchons les relations entre les composantes x^i et x^{ik} il suffit de reprendre les relations de changement de base démontrées dans le chapitre d'Algèbre Linéaire et nous avons alors:

$$\vec{x} = x^i \vec{e}_i = x^i A_i^{ik} \vec{e}'_k = x^{ik} \vec{e}'_k = x^{ik} A_k^i \vec{e}_i \quad (14.44)$$

De suite par l'unicité de la décomposition d'un vecteur sur une base, nous pouvons évaluer les coefficients des vecteurs de base et nous obtenons (il faut prendre garde à réarranger à nouveau l'ordre des termes car la multiplication matricielle n'est, en règle générale, pas commutative comme nous le savons déjà):

$$x^{ik} = A_i^{ik} x^i \text{ et } x^i = A_k^i x^{ik} \quad (14.45)$$

Il vient également la relation triviale (cf. chapitre d'Algèbre Linéaire):

$$A_k^i A_i^j = \delta_k^j \quad (14.46)$$

Effectivement faisons un exemple explicite simple avec une matrice de dimension 2:

$$\begin{bmatrix} a'_{11} & a'_{12} \\ a'_{21} & a'_{22} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a'_{11} a_{11} + a'_{12} a_{21} & a'_{11} a_{12} + a'_{12} a_{22} \\ a'_{21} a_{11} + a'_{22} a_{21} & a'_{21} a_{12} + a'_{22} a_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (14.47)$$

Une autre manière élégante de montrer en toute généralité la relation antérieure est de se rappeler du résultat démontré dans le chapitre d'Algèbre Linéaire:

$$\vec{x} = x^i \vec{e}_i = x^i A_i^{ik} \vec{e}'_k = x^{ik} \vec{e}'_k = x^{ik} A_k^i \vec{e}_i \quad (14.48)$$

et en utilisant:

$$\vec{e}'_k = A_k^i \vec{e}_i \quad (14.49)$$

Il vient alors:

$$\vec{e}_i = A_i^{ik} \vec{e}'_k = A_i^{ik} A_k^j \vec{e}_j \quad (14.50)$$

Les vecteurs de base étant linéairement indépendants, cette dernière relation implique que lorsque $i \neq j$:

$$A_i^{ik} A_k^j = 0 \quad (14.51)$$

et lorsque $i = j$:

$$A_i^{ik} A_k^j = 1 \quad (14.52)$$

Ainsi il vient:

$$A_i^k A_k^j = \delta_i^j \quad (14.53)$$

Quant au produit scalaire, les résultats obtenus avec la notation indicielle sont forts intéressants et extrêmement puissants. Nous avons déjà défini le produit scalaire dans le chapitre de Calcul Vectoriel mais voyons comment nous manipulons ce dernier avec la notation indicielle:

Considérons un espace vectoriel euclidien E_n rapporté à une base quelconque $(\vec{e}_i)$. Les vecteurs s'écrivent sur cette base (nous le savons déjà):

$$\vec{x} = x^i \vec{e}_i, \vec{y} = y^j \vec{e}_j \quad (14.54)$$

Le produit scalaire relativement à ses propriétés et à la notation indicielle s'écrit alors:

$$\vec{x} \circ \vec{y} = (x^i \vec{e}_i) \circ (y^j \vec{e}_j) = x^i y^j (\vec{e}_i \circ \vec{e}_j) = g_{ij} x^i y^j \quad (14.55)$$

Relation fondamentale pour la physique de pointe (relativité générale et théorie des cordes) qui fait apparaître le "**tenseur métrique covariant**":

$$g_{ij} = \vec{e}_i \circ \vec{e}_j \quad (14.56)$$

et pour satisfaire la propriété de commutativité du produit scalaire (cf. chapitre de Calcul Vectoriel) nous devons évidemment avoir l'égalité:

$$g_{ij} = g_{ji} \quad (14.57)$$

La relation antécédent s'écrit aussi parfois sous la forme:

$$[g] = \begin{bmatrix} g_{11} & g_{12} & \cdots \\ g_{21} & g_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} = \begin{bmatrix} \vec{e}_1 \circ \vec{e}_1 & \vec{e}_1 \circ \vec{e}_2 & \cdots \\ \vec{e}_2 \circ \vec{e}_1 & \vec{e}_2 \circ \vec{e}_2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \quad (14.58)$$

Remarque: Lorsque les vecteurs de base $(\vec{e}_i)$ forment un espace vectoriel orthogonal (pas nécessairement orthonormé) alors les quantités:

$$g_{ij} = \vec{e}_i \circ \vec{e}_j \quad (14.59)$$

sont nulles si $i \neq j$. Le produit scalaire de deux vecteurs $\vec{x}$ et $\vec{y}$ se réduit alors à:

$$\vec{x} \circ \vec{y} = g_{11}x^1y^1 + g_{22}x^2y^2 + \dots + g_{nn}x^nx^n \quad (14.60)$$

Nous avons alors dans ce cas particulier:

$$g_{ij} = \delta_{ij}(\vec{e}_i \circ \vec{e}_j) \quad (14.61)$$

et donc lorsque les vecteurs de base forment un espace vectoriel orthonormal il est clair que g_{ij} est alors égal au symbole de Kronecker seul tel que:

$$\vec{x} \circ \vec{y} = \|\vec{x} \circ \vec{y}\|^2 = \delta_{ij}x^ix^j \quad (14.62)$$

3. MÉTRIQUE ET SIGNATURE

Comme nous l'avons vu en calcul vectoriel, le produit scalaire d'un vecteur $\vec{x}$ peut permettre de définir la notion de norme d'un vecteur (et le concept de distance).

Rappelons que nous avons par définition la norme d'un vecteur qui est donnée par (cf. chapitre de Calcul Vectoriel):

$$\vec{x} \circ \vec{x} = g_{ij}x^ix^j = \|\vec{x}\|^2 \quad (14.63)$$

où les nombres g_{ij} définissent en quelque sorte une "mesure" des vecteurs; nous disons alors dans le langage du calcul tensoriel qu'ils constituent la "**métrique**" de l'espace vectoriel choisi.

Dans l'espace de la géométrie classique, la norme est un nombre qui est toujours strictement positif et qui ne devient nul que si le vecteur mesuré est égal à zéro. Par contre, l'expression précédente de la norme d'un vecteur, peut être éventuellement négative pour des nombres $g_{11}, g_{12}, \dots, g_{nn}$ quelconques (espaces complexes par exemple). Nous pouvons donc distinguer deux genres d'espaces vectoriels pré-euclidiens (espace euclidien dans lequel nous avons défini le produit scalaire) selon que la norme est positive ou non. Cependant lorsqu'en physique théorique nous souhaitons faire l'analogisme avec une structure d'espace vectoriel il faut que la condition:

$$g = \det[g_{ij}] \neq 0 \quad (14.64)$$

soit satisfaite (g_{ij} peut être écrit comme une matrice, rien ne nous l'empêche).

Explications: Nous savons que le produit scalaire doit satisfaire à la propriété de commutativité telle que:

$$\vec{x} \circ \vec{y} = g_{ij} x^i y^j = \vec{y} \circ \vec{x} = g_{ji} x^j y^i \quad (14.65)$$

D'autre part, si pour tout y^j non nul nous avons:

$$g_{ij} x^i y^j = 0 \quad (14.66)$$

cela implique $x^i = 0$ (c'est une des propriétés de la norme que nous avons vue dans le chapitre de Calcul Vectoriel). Nous pouvons alors écrire:

$$g_{ij} x^i = 0 \quad (14.67)$$

Nous nous retrouvons ici simplement avec un système de n équations à n inconnues (ne devant admettre par hypothèse que la solution $x^i = 0$), il faut et il suffit pour cela que le déterminant du système, noté g , du système soit différent de zéro (cf. chapitre d'Algèbre Linéaire). Nous devons donc avoir:

$$g = \det[g_{ij}] \neq 0 \quad (14.68)$$

C'est une des conditions pour qu'une expression assimilable à une norme sous une écriture tensorielle forme dans le cadre d'une théorie physique un espace vectoriel des états du système !!

Remarques:

- R1. Le nombre de signes + et - se trouvant dans l'expression du produit scalaire constitue une caractéristique d'un espace vectoriel donné E_n ; elle est appelée la "**signature de l'espace vectoriel**" E_n .
- R2. Une application pratique des calculs de la métrique est proposée dans le chapitre de Relativité Générale.

A partir des coefficients du tenseur métrique covariant g_{ij} définissant la métrique de l'espace E_n , nous pouvons introduire les coefficients du "**tenseur métrique contravariant**" g^{ij} définissant la métrique d'un "**espace dual**" E_n^* par la relation:

$$g_{ik} g^{kj} = \delta_i^j = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases} \quad (14.69)$$

En d'autres termes, le tenseur métrique deux fois covariant est son propre inverse par son équivalent deux fois contravariant. Nous le démontrerons explicitement plus loin en montrant lors de notre étude du déterminant de Gram que les composantes contravariantes et covariantes d'un espace euclidien sont égales et que les deux espaces ont le même nombre de dimensions.

Un cas particulier qui satisfait la relation ci-dessus est le tenseur métrique de Minkowski (cf. chapitres de Relativité Restreinte et Générale) où nous avons:

$$\eta_{\mu\nu} = \eta^{\mu\nu} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \quad (14.70)$$

Remarque: L'espace E_n est aussi appelé "espace primal".

L'espace dual est sous-tendu par n vecteurs de base $\vec{e}^i$ construite à partir des vecteurs $\vec{e}_i$ tel que:

$$\vec{e}^i = g^{ij} \vec{e}_j \quad (14.71)$$

Il est dès lors facile de voir que le produit scalaire des vecteurs $\vec{e}^i$ définit la métrique g^{ij} de l'espace dual:

$$\vec{e}^i \circ \vec{e}^j = (g^{ik} \vec{e}_k) \circ (g^{jl} \vec{e}_l) = g^{ik} g^{jl} (\vec{e}_k \circ \vec{e}_l) = g^{ik} (g^{jl} g_{kl}) = g^{ik} \delta_k^j = g^{ij} \quad (14.72)$$

tandis que les vecteurs $\vec{e}^i$ (contravariants) et $\vec{e}_j$ (covariants) sont bien orthogonaux:

$$\vec{e}^i \circ \vec{e}_j = (g^{ik} \vec{e}_k) \circ \vec{e}_j = g^{ik} (\vec{e}_k \circ \vec{e}_j) = g^{ik} g_{kj} = \delta_j^i \quad (14.73)$$

Nous pouvons exprimer aussi un vecteur dans la base duale par l'écriture suivante en remarquant bien évidemment que la position des indices muets est inversée:

$$\vec{x} = x_1 \vec{e}^1 + x_2 \vec{e}^2 + x_3 \vec{e}^3 = x_i \vec{e}^i \quad (14.74)$$

Remarque: Les composantes x_i sont nommées, pour des raisons que nous verrons plus loin, les composantes covariantes.

Ainsi nous avons finalement la possibilité de passer aussi les vecteurs d'une base à l'autre:

$$\vec{x} = x_i \vec{e}^i = x_i (g^{ik} \vec{e}_k) = (x_i g^{ik}) \vec{e}_k = x^k \vec{e}_k \quad (14.75)$$

où ce qu'il est important de retenir est que pour rendre contravariante une composante covariante, nous montons son indice:

$$x^k = g^{ik} x_i \quad (14.76)$$

et inversement, pour la rendre covariante:

$$x_i = g_{ik} x^k \quad (14.77)$$

Ainsi, toujours dans le cas de l'exemple de la métrique de Minkowski, si nous considérons le quadrivecteur contravariant:

$$x^{\mu} = \left(x^0, x^i \right) = \left(x^0, \vec{x} \right) \quad (14.78)$$

Nous avons alors:

$$x_{\mu} = \eta_{\mu\nu} x^{\nu} = \left(x_0, x_i \right) = \left(x_0, -x_i \right) = \left(x_0, -\vec{x} \right) \quad (14.79)$$

4. DÉTERMINANT DE GRAM

Voyons une autre approche pour obtenir les vecteurs de base de l'espace dual qui peut permettre par ailleurs de mieux appréhender le concept et qui nous permettra d'obtenir un résultat intéressant que nous utiliserons lors de certains calculs de la relativité générale (principalement son étude selon le formalisme lagrangien).

Nous avons donc pour $i = j = 1$:

$$\vec{e}^1 \circ \vec{e}_1 = 1 \quad (14.80)$$

Ce produit scalaire peut être vu comme une condition de normalisation pour les deux bases et les deux produits scalaires $\vec{e}^2 \circ \vec{e}_1 = 0, \vec{e}^3 \circ \vec{e}_1 = 0$ comme des conditions d'orthogonalisation. Ainsi, comme $\vec{e}_1$ est perpendiculaire à $\vec{e}^2, \vec{e}^3$ nous pouvons écrire:

$$\vec{e}_1 = c^{te} \left(\vec{e}^2 \times \vec{e}^3 \right) \quad (14.81)$$

Où c^{te} est une constante de proportionnalité. Maintenant jouons un peu avec la relation précédente:

$$\vec{e}^1 \circ \vec{e}_1 = c^{te} \vec{e}^1 \circ \left(\vec{e}^2 \times \vec{e}^3 \right) = 1 \quad (14.82)$$

Dès lors, nous obtenons:

$$c^{te} = \frac{1}{\vec{e}^1 \circ \left(\vec{e}^2 \times \vec{e}^3 \right)} = \frac{1}{\left[\vec{e}^1, \vec{e}^2, \vec{e}^3 \right]} \quad (14.83)$$

où nous voyons apparaître le produit mixte tel que nous l'avons défini dans le chapitre de Calcul Vectoriel.

Ainsi, nous obtenons très facilement:

$$\vec{e}_1 = \frac{\vec{e}^2 \times \vec{e}^3}{\left[\vec{e}^1, \vec{e}^2, \vec{e}^3 \right]} \quad (14.84)$$

ou de manière plus générale (sans démonstration car peut-être trop évident) nous avons donc pour les vecteurs covariants:

$$\vec{e}_i = \frac{\vec{e}^j \times \vec{e}^k}{\left[\vec{e}^1, \vec{e}^2, \vec{e}^3 \right]} \quad (14.85)$$

et de même pour les vecteurs contravariants (sans démonstration car peut-être trop évident?):

$$\vec{e}^i = \frac{\vec{e}_j \times \vec{e}_k}{[\vec{e}_1, \vec{e}_2, \vec{e}_3]} \quad (14.86)$$

Remarques:

R1. Le lecteur aura remarqué que les relations ci-dessus ne sont valables que pour un espace à trois dimensions.

R2. La notation des deux relations précédentes est mathématiquement un peu abusive car en réalité ce n'est pas une égalité entre deux vecteurs mais une application d'un espace vectoriel dans l'autre!

R3. Comme en physique, on considère très fréquemment des bases cartésiennes, cylindriques et sphériques orthonormées et que le dénominateur des deux relations précédentes est toujours égal à l'unité dans ces bases alors les vecteurs de bases contravariants s'identifient aux vecteurs de base covariants (et réciproquement).

Revenons maintenant sur quelque chose qui va nous sembler bien ancien... Dans le chapitre de Calcul Vectoriel, nous avons défini et étudié ce qu'étaient le produit vectoriel et le produit mixte. Nous allons voir maintenant une autre manière de représenter ceux-ci et voir que cette représentation permet d'obtenir un résultat pour le moins pertinent!

Nous avons vu dans le chapitre de Calcul Vectoriel que le produit vectoriel était donné par:

$$\vec{x} \times \vec{y} = (x_2 y_3 - x_3 y_2) \vec{e}_1 + (x_3 y_1 - x_1 y_3) \vec{e}_2 + (x_1 y_2 - x_2 y_1) \vec{e}_3 \quad (14.87)$$

Or, ce que nous n'avions pas vu et que nous pouvons constater maintenant de manière triviale c'est que cette expression n'est que le déterminant des matrices suivantes:

$$\vec{x} \times \vec{y} = \det \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix} = \det \begin{bmatrix} \vec{e}_2 \times \vec{e}_3 & \vec{e}_3 \times \vec{e}_1 & \vec{e}_1 \times \vec{e}_2 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix} \quad (14.88)$$

Mais comme nous faisons du calcul tensoriel, il nous faut maintenant proprement distinguer composantes covariantes et contravariantes. Nous allons donc réécrire cela correctement avec les composantes contravariantes:

$$\vec{x} \times \vec{y} = \det \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \end{bmatrix} \quad (14.89)$$

De même, le produit mixte peut être écrit à l'aide de cette relation et notation:

$$\begin{aligned}
 [\vec{x}, \vec{y}, \vec{z}] &= \vec{x} \circ (\vec{y} \times \vec{z}) = \vec{x} \circ \det \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ y^1 & y^2 & y^3 \\ z^1 & z^2 & z^3 \end{bmatrix} \\
 &= (x^1 \vec{e}_1 + x^2 \vec{e}_2 + x^3 \vec{e}_3) \circ \det \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ y^1 & y^2 & y^3 \\ z^1 & z^2 & z^3 \end{bmatrix}
 \end{aligned} \quad (14.90)$$

Or, en regardant l'expression du déterminant nous voyons assez facilement, sans même avoir à faire les développements que:

$$[\vec{x}, \vec{y}, \vec{z}] = \vec{e}_1 \circ (\vec{e}_2 \times \vec{e}_3) \cdot \det \begin{bmatrix} x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \\ z^1 & z^2 & z^3 \end{bmatrix} \quad (14.91)$$

Effectivement (nous calculons le déterminant en faisant usage de la démonstration du déterminant à trois composantes vue dans le chapitre d'Algèbre Linéaire):

$$\begin{aligned}
 [\vec{x}, \vec{y}, \vec{z}] &= \vec{x} \circ (\vec{y} \times \vec{z}) = \vec{e}_1 \circ (\vec{e}_2 \times \vec{e}_3) \cdot \det \begin{bmatrix} x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \\ z^1 & z^2 & z^3 \end{bmatrix} \\
 &= \vec{e}_1 \circ (\vec{e}_2 \times \vec{e}_3) \cdot (x^1 y^2 z^3 - x^1 y^3 z^2 - x^2 y^1 z^3 + x^2 y^3 z^1 + x^3 y^1 z^2 - x^3 y^2 z^1) \\
 &= \vec{e}_1 \circ \left(\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \times \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right) \cdot (x^1 y^2 z^3 - x^1 y^3 z^2 - x^2 y^1 z^3 + x^2 y^3 z^1 + x^3 y^1 z^2 - x^3 y^2 z^1) \\
 &= \vec{e}_1 \circ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \cdot (x^1 y^2 z^3 - x^1 y^3 z^2 - x^2 y^1 z^3 + x^2 y^3 z^1 + x^3 y^1 z^2 - x^3 y^2 z^1) \\
 &= \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \circ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \cdot (x^1 y^2 z^3 - x^1 y^3 z^2 - x^2 y^1 z^3 + x^2 y^3 z^1 + x^3 y^1 z^2 - x^3 y^2 z^1) \\
 &= (x^1 y^2 z^3 - x^1 y^3 z^2 - x^2 y^1 z^3 + x^2 y^3 z^1 + x^3 y^1 z^2 - x^3 y^2 z^1) \\
 &= ((x^1 y^2 z^3 - x^1 y^3 z^2) - (x^2 y^1 z^3 - x^2 y^3 z^1) + (x^3 y^1 z^2 - x^3 y^2 z^1)) \\
 &= (x^1 (y^2 z^3 - y^3 z^2) + x^2 (y^3 z^1 - y^1 z^3) + x^3 (y^1 z^2 - y^2 z^1)) \\
 &= \begin{pmatrix} x^1 \\ x^2 \\ x^3 \end{pmatrix} \circ \begin{pmatrix} y^2 z^3 - y^3 z^2 \\ y^3 z^1 - y^1 z^3 \\ y^1 z^2 - y^2 z^1 \end{pmatrix} = \begin{pmatrix} x^1 \\ x^2 \\ x^3 \end{pmatrix} \circ \left(\begin{pmatrix} y^1 \\ y^2 \\ y^3 \end{pmatrix} \times \begin{pmatrix} z^1 \\ z^2 \\ z^3 \end{pmatrix} \right) = \vec{x} \circ (\vec{y} \times \vec{z})
 \end{aligned} \quad (14.92)$$

La relation antécédente est fréquemment notée:

$$[\vec{x}, \vec{y}, \vec{z}] = \sqrt{g} \det \begin{bmatrix} x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \\ z^1 & z^2 & z^3 \end{bmatrix} \quad (14.93)$$

avec:

$$\sqrt{g} = \vec{e}_1 \circ (\vec{e}_2 \times \vec{e}_3) = [\vec{e}_1, \vec{e}_2, \vec{e}_3] \quad (14.94)$$

appelé "**volume euclidien**" (effectivement rappelons que le produit mixte est un volume!)

Remarque: Rappelons encore une fois que si les vecteurs de base sont orthonormés, qu'ils soient exprimés en coordonnées cartésiennes, cylindriques ou sphériques alors:

$$\sqrt{g} = \vec{e}_1 \circ (\vec{e}_2 \times \vec{e}_3) = 1 \quad (14.95)$$

Par ailleurs, nous avons aussi la relation non moins importante:

$$\sqrt{g} = \vec{e}_1 \circ (\vec{e}_2 \times \vec{e}_3) = \left(\frac{\vec{e}_2 \times \vec{e}_3}{[\vec{e}^1, \vec{e}^2, \vec{e}^3]} \right) \circ (\vec{e}_2 \times \vec{e}_3) = \frac{(\vec{e}^2 \times \vec{e}^3) \circ (\vec{e}_2 \times \vec{e}_3)}{[\vec{e}^1, \vec{e}^2, \vec{e}^3]} \quad (14.96)$$

En utilisant la relation vue dans le chapitre de Calcul Vectoriel:

$$(\vec{e}^2 \times \vec{e}^3) \circ (\vec{e}_2 \times \vec{e}_3) = (\vec{e}^2 \circ \vec{e}_2)(\vec{e}^3 \circ \vec{e}_3) - (\vec{e}^2 \circ \vec{e}_3)(\vec{e}^3 \circ \vec{e}_2) \quad (14.97)$$

Or, nous avons vu plus haut que $\vec{e}^i \circ \vec{e}_j = \delta_j^i$ donc:

$$(\vec{e}^2 \circ \vec{e}_2)(\vec{e}^3 \circ \vec{e}_3) - (\vec{e}^2 \circ \vec{e}_3)(\vec{e}^3 \circ \vec{e}_2) = 1 \cdot 1 - 0 \cdot 0 \quad (14.98)$$

et finalement:

$$\sqrt{g} = [\vec{e}_1, \vec{e}_2, \vec{e}_3] = \frac{1}{[\vec{e}^1, \vec{e}^2, \vec{e}^3]} \quad (14.99)$$

Ceci ayant été fait revenons à la relation du produit vectoriel:

$$\vec{x} \times \vec{y} = x^i \vec{e}_i \times y^j \vec{e}_j = \det \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \end{bmatrix} = \det \begin{bmatrix} \vec{e}_2 \times \vec{e}_3 & \vec{e}_3 \times \vec{e}_1 & \vec{e}_1 \times \vec{e}_2 \\ x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \end{bmatrix} \quad (14.100)$$

et exprimons les composantes du déterminant dans leur base duale (en coordonnées contravariantes):

$$\begin{aligned}
\vec{x} \times \vec{y} &= \det \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \end{bmatrix} = \det \begin{bmatrix} \frac{\vec{e}_2 \times \vec{e}_3}{[\vec{e}_1, \vec{e}_2, \vec{e}_3]} & \frac{\vec{e}_3 \times \vec{e}_1}{[\vec{e}_1, \vec{e}_2, \vec{e}_3]} & \frac{\vec{e}_1 \times \vec{e}_2}{[\vec{e}_1, \vec{e}_2, \vec{e}_3]} \\ x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \end{bmatrix} \\
&= \frac{1}{[\vec{e}_1, \vec{e}_2, \vec{e}_3]} \det \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \end{bmatrix} = \sqrt{g} \det \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \end{bmatrix}
\end{aligned} \tag{14.101}$$

Bien évidemment, si le produit vectoriel est exprimé en composantes covariantes alors nous avons:

$$\begin{aligned}
\vec{x} \times \vec{y} &= x_i \vec{e}^i \times y_j \vec{e}^j = \det \begin{bmatrix} \vec{e}^1 & \vec{e}^2 & \vec{e}^3 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix} = \det \begin{bmatrix} \frac{\vec{e}_2 \times \vec{e}_3}{[\vec{e}_1, \vec{e}_2, \vec{e}_3]} & \frac{\vec{e}_3 \times \vec{e}_1}{[\vec{e}_1, \vec{e}_2, \vec{e}_3]} & \frac{\vec{e}_1 \times \vec{e}_2}{[\vec{e}_1, \vec{e}_2, \vec{e}_3]} \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix} \\
&= \frac{1}{[\vec{e}_1, \vec{e}_2, \vec{e}_3]} \det \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix} = \frac{1}{\sqrt{g}} \det \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix}
\end{aligned} \tag{14.102}$$

Maintenant appliquons le produit mixte:

$$[\vec{x}, \vec{y}, \vec{z}] = x^i \vec{e}_i \circ (y^j \vec{e}_j \times z^k \vec{e}_k) = x^i \vec{e}_i \sqrt{g} \circ \det \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ y^1 & y^2 & y^3 \\ z^1 & z^2 & z^3 \end{bmatrix} \tag{14.103}$$

en connaissant l'expression du déterminant d'une matrice carrée 3×3 (cf. chapitre d'Algèbre Linéaire) il vient immédiatement:

$$[\vec{x}, \vec{y}, \vec{z}] = x^i \vec{e}_i \sqrt{g} \circ \det \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ y^1 & y^2 & y^3 \\ z^1 & z^2 & z^3 \end{bmatrix} = \sqrt{g} \det \begin{bmatrix} x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \\ z^1 & z^2 & z^3 \end{bmatrix} \tag{14.104}$$

Inversement, il vient immédiatement:

$$[\vec{x}, \vec{y}, \vec{z}] = x_i \vec{e}^i \circ (y_j \vec{e}^j \times z_k \vec{e}^k) = \frac{1}{\sqrt{g}} \det \begin{bmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{bmatrix} \tag{14.105}$$

Or, nous avons vu que dans le chapitre de Calcul Vectoriel que $x_i = \vec{x} \circ \vec{e}_i$. Il vient alors:

$$\begin{aligned}
 [\vec{x}, \vec{y}, \vec{z}] &= x_i \vec{e}^i \circ (y_j \vec{e}^j \times z_k \vec{e}^k) = \frac{1}{\sqrt{g}} \det \begin{bmatrix} \vec{x} \circ \vec{e}_1 & \vec{x} \circ \vec{e}_2 & \vec{x} \circ \vec{e}_3 \\ \vec{y} \circ \vec{e}_1 & \vec{y} \circ \vec{e}_2 & \vec{y} \circ \vec{e}_3 \\ \vec{z} \circ \vec{e}_1 & \vec{z} \circ \vec{e}_2 & \vec{z} \circ \vec{e}_3 \end{bmatrix} \\
 &= \frac{1}{[\vec{e}_1, \vec{e}_2, \vec{e}_3]} \det \begin{bmatrix} \vec{x} \circ \vec{e}_1 & \vec{x} \circ \vec{e}_2 & \vec{x} \circ \vec{e}_3 \\ \vec{y} \circ \vec{e}_1 & \vec{y} \circ \vec{e}_2 & \vec{y} \circ \vec{e}_3 \\ \vec{z} \circ \vec{e}_1 & \vec{z} \circ \vec{e}_2 & \vec{z} \circ \vec{e}_3 \end{bmatrix}
 \end{aligned} \quad (14.106)$$

et donc:

$$[\vec{x}, \vec{y}, \vec{z}][\vec{e}_1, \vec{e}_2, \vec{e}_3] = \det \begin{bmatrix} \vec{x} \circ \vec{e}_1 & \vec{x} \circ \vec{e}_2 & \vec{x} \circ \vec{e}_3 \\ \vec{y} \circ \vec{e}_1 & \vec{y} \circ \vec{e}_2 & \vec{y} \circ \vec{e}_3 \\ \vec{z} \circ \vec{e}_1 & \vec{z} \circ \vec{e}_2 & \vec{z} \circ \vec{e}_3 \end{bmatrix} \quad (14.107)$$

Cette dernière relation étant souvent appelée "**déterminant de Gram**". Un cas particulier très intéressant nous donne:

$$[\vec{e}_1, \vec{e}_2, \vec{e}_3]^2 = \det \begin{bmatrix} \vec{e}_1 \circ \vec{e}_1 & \vec{e}_1 \circ \vec{e}_2 & \vec{e}_1 \circ \vec{e}_3 \\ \vec{e}_2 \circ \vec{e}_1 & \vec{e}_2 \circ \vec{e}_2 & \vec{e}_2 \circ \vec{e}_3 \\ \vec{e}_3 \circ \vec{e}_1 & \vec{e}_3 \circ \vec{e}_2 & \vec{e}_3 \circ \vec{e}_3 \end{bmatrix} = \det \begin{bmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ g_{31} & g_{32} & g_{33} \end{bmatrix} \quad (14.108)$$

écrit autrement:

$$(\sqrt{g})^2 = g = \det(g_{ij}) \quad (14.109)$$

Ainsi, le volume euclidien est donné par ce que nous appellerons le "**déterminant fonctionnel**" du système (expression que nous retrouverons en relativité générale et théorie des cordes):

$$\sqrt{g} = \sqrt{\det(g_{ij})} \quad (14.110)$$

Si nous notons autrement le déterminant:

$$\sqrt{g} = \sqrt{|g|} \quad (14.111)$$

5. COMPOSANTES CONTRAVARIANTES ET COVARIANTES

Jusqu'à maintenant nous avons écrit les indices muets arbitrairement en exposant ou en indice selon notre bon vouloir. Cependant, cela n'est pas toujours autorisé et parfois le fait qu'un indice muet soit en exposant ou en indice a une signification bien particulière. Ceci constitue souvent la difficulté lors de l'étude de certains théorèmes, car si nous n'étudions pas ceux-là depuis le début, nous ne savons pas vraiment comment interpréter la position des indices muets. Il faut donc être extrêmement prudent à ce niveau.

Pour un espace vectoriel euclidien E_n rapporté à une base quelconque $(\vec{e}_i)$, le produit scalaire d'un vecteur $\vec{x} = x^i \vec{e}_i$ par un vecteur $\vec{e}_j$ de sa base s'écrit:

$$\vec{x} \circ \vec{e}_j = (x^i \vec{e}_i) \circ \vec{e}_j = x^i (\vec{e}_i \circ \vec{e}_j) = x^i g_{ij} = x_j \quad (14.112)$$

Donc:

$$\boxed{x_j = x^i g_{ij}} \quad (14.113)$$

Cette relation est de première importance en physique théorique et en calcul tensoriel. Il est important de s'en souvenir lorsque nous étudierons la contraction des indices plus tard (vous pouvez observer dans la relation précédente que nous avons "abaissé" dans le membre de gauche l'indice des composantes du membre droit de l'égalité).

Ces produits scalaires notés x_j , s'appellent les "**composantes covariantes**", dans la base $(\vec{e}_i)$, du vecteur $\vec{x}$. Ces composantes sont donc définies par:

$$x_j = \vec{x} \circ \vec{e}_j \quad (14.114)$$

Remarque: Cela constitue donc une projection d'un vecteur sur un des vecteurs de sa propre base.

Elles seront notées au moyen d'indices inférieurs !!! Nous verrons par la suite que ces composantes s'introduisent naturellement pour certains vecteurs de la physique, par exemple le vecteur gradient. D'autre part, la notion de composante covariante est essentielle pour les tenseurs.

Remarque: Les vecteurs de base ont toujours les indices notés en bas car ils sont leurs propres composantes covariantes (ils se projettent sur eux-mêmes par produit scalaire).

Inversement, les "**composantes contravariantes**" (autrement dit les composantes non projetées) peuvent être calculées en résolvant, par rapport aux n inconnues x^i , le système de n équations de:

$$\boxed{x_j = x^i g_{ij}} \quad (14.115)$$

Les relations précédentes montrent que les composantes covariantes x_j sont liées aux composantes x^i classiques et que les composantes contravariantes sont donc des nombres x^i tels que:

$$\vec{x} = x^i \vec{e}_i \quad (14.116)$$

Elles seront indiquées au moyen d'indices supérieurs !! L'étude des changements de base permettra de justifier encore plus l'appellation des différentes composantes.

Dans une base orthonormée canonique (cas très particulier), les composantes covariantes et contravariantes sont identiques comme nous le savons déjà suite à l'étude du déterminant de Gram. Effectivement:

$$x_j = \vec{x} \circ \vec{e}_j = (x^i \vec{e}_i) \circ \vec{e}_j = x^i (\vec{e}_i \circ \vec{e}_j) = x^i \delta_{ij} = x^j \quad (14.117)$$

Remarque: Nous voyons ci-dessus, que l'écriture incessante d'indices muets en exposants ou en indice peut parfois amener à certaines confusions et à des maux de tête sérieux...

6. OPÉRATIONS DANS LES BASES

L'intérêt du physicien pour le calcul tensoriel, est le passage de paramètres d'une base à une autre pour des raisons données (souvent dans le but soit de simplifier l'étude de problèmes ou simplement parce que les états étudiés dépendent - ou peuvent dépendre - de la géométrie de l'espace dont il est question). Il convient donc d'introduire les principaux outils qui y sont relatifs. Nous en profiterons aussi pour présenter des développements que nous aurions pu déjà aborder dans le chapitre de Calcul Vectoriel.

6.1. MÉTHODE D'ORTHOGONALISATION DE SCHMIDT

La "**méthode d'orthogonalisation de Schmidt**" (dite également de "Gram-Schmidt") permet le calcul effectif d'une base orthogonale pour tout espace vectoriel pré-euclidien E_n (nous aurions pu présenter cette méthode dans le chapitre de Calcul Vectoriel mais il nous semblait plus intéressant de la présenter dans le cadre général et esthétique du calcul tensoriel).

Pour cela, considérons un ensemble de n vecteurs linéairement indépendants $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n)$ de E_n et supposons que nous ayons pour chaque vecteur le produit scalaire (la norme):

$$\vec{x}_i \circ \vec{x}_i \neq 0 \quad (14.118)$$

Cherchons n vecteurs $\vec{e}_i$ orthogonaux entre eux. Partons pour cela de $\vec{e}_1 = \vec{x}_1$ et cherchons $\vec{e}_2$ orthogonal à $\vec{e}_1$ sous la forme:

$$\vec{e}_2 = \lambda_1 \vec{e}_1 + \vec{x}_2 \quad (14.119)$$

Le coefficient λ_1 se calcule en écrivant la relation d'orthogonalité:

$$\vec{e}_1 \circ \vec{e}_2 = \vec{x}_1 \circ (\lambda_1 \vec{x}_1 + \vec{x}_2) = 0 \quad (14.120)$$

Nous en déduisons sans trop de peine:

$$\lambda_1 = -\frac{\vec{x}_1 \circ \vec{x}_2}{\vec{x}_1 \circ \vec{x}_1} = -\frac{\vec{x}_1 \circ \vec{x}_2}{\|\vec{x}_1\|^2} \quad (14.121)$$

Le paramètre λ_1 étant déterminé, nous obtenons le vecteur $\vec{e}_2$ qui est orthogonal à $\vec{e}_1$ et non nul puisque le système $(\vec{e}_1, \vec{x}_2, \dots, \vec{x}_n)$ est linéairement indépendant.

Le vecteur $\vec{e}_3$ est cherché sous la forme:

$$\vec{e}_3 = \mu_1 \vec{e}_1 + \mu_2 \vec{e}_2 + \vec{x}_3 \quad (14.122)$$

Les deux relations d'orthogonalité: $\vec{e}_1 \circ \vec{e}_3 = 0$ et $\vec{e}_2 \circ \vec{e}_3 = 0$, permettent le calcul des coefficients μ_1 et μ_2 . Nous obtenons:

$$\mu_1 = -\frac{\vec{e}_1 \circ \vec{x}_3}{\|\vec{e}_1\|^2}; \mu_2 = -\frac{\vec{e}_2 \circ \vec{x}_3}{\|\vec{e}_2\|^2} \quad (14.123)$$

ce qui détermine le vecteur $\vec{e}_3$, orthogonal à $\vec{e}_1$ et $\vec{e}_2$, et non nul puisque le système $(\vec{e}_1, \vec{e}_2, \vec{x}_3, \dots, \vec{x}_n)$ est indépendant. En continuant le même type de calcul, nous obtenons de proche en proche un système de vecteurs $(\vec{e}_1, \vec{e}_2, \dots, \vec{e}_n)$ orthogonaux entre eux et dont aucun n'est nul.

Dans le cas où certains vecteurs seraient tels que $\vec{x}_i \circ \vec{x}_i = 0$ (leur norme est nulle), nous remplaçons $\vec{x}_i$ par $\vec{x}_i' = \vec{x}_i + \lambda \cdot \vec{x}_j$, en choisissant un vecteur $\vec{x}_j$ de telle sorte que nous obtenions $\vec{x}_i' \circ \vec{x}_i' \neq 0$.

Nous en déduisons donc que tout espace vectoriel pré-euclidien admet des bases orthogonales!

Ce système de calcul des bases est de première importance. Il permet par exemple d'étudier des systèmes physiques à partir d'un référentiel pré-euclidien dont les propriétés changent dans le temps. Ce qui est par exemple typique de la relativité générale.

6.2. CHANGEMENTS DE BASES

Soient deux bases $(\vec{e}_1, \vec{e}_2, \dots, \vec{e}_n)$ et $(\vec{e}'_1, \vec{e}'_2, \dots, \vec{e}'_n)$ d'un espace vectoriel E_n . Chaque vecteur d'une base peut être décomposé sur l'autre base sous la forme suivante (nous l'avons déjà démontré):

$$\vec{e}_i = A_i^k \vec{e}'_k \quad \text{et} \quad \vec{e}'_k = A_k^i \vec{e}_i \quad (14.124)$$

Un vecteur $\vec{x}$ de E_n peut être décomposé sur chaque base sous la forme:

$$\vec{x} = x^i \vec{e}_i = x'^k \vec{e}'_k \quad (14.125)$$

et nous avons aussi déjà démontré que:

$$x^i = A_k^i x'^k \quad \text{et} \quad x'^k = A_i^k x^i \quad (14.126)$$

Nous remarquons que les relations de transformation des composantes contravariantes d'un vecteur sont le contraire de celles des vecteurs de base, les grandeurs A et A' s'échangeant, d'où l'origine de l'appellation "**contra**"-"**variantes**" de ces composantes!

Soient x_i et x'_k les composantes covariantes du vecteur $\vec{x}$ respectivement sur les bases $(\vec{e}_i)$ et $(\vec{e}'_k)$. Remplaçons les vecteurs de base, exprimés par les relations:

$$\vec{e}_i = A_i^k \vec{e}'_k \quad \text{et} \quad \vec{e}'_k = A_k^i \vec{e}_i \quad (14.127)$$

dans l'expression de définition des composantes covariantes, il vient:

$$x_i = \vec{x} \circ \vec{e}_i = \vec{x} \circ (A_i^k \vec{e}'_k) = A_i^k (\vec{x} \circ \vec{e}'_k) = A_i^k x'_k \quad (14.128)$$

d'où la relation entre les composantes covariantes dans chaque base:

$$x_i = A_i^k x'_k \quad (14.129)$$

Nous obtenons de même:

$$x'_k = A^i_k x_i \quad (14.130)$$

Nous remarquons que les composantes covariantes se transforment comme les vecteurs de bases, d'où l'appellation de ces composantes.

6.3. BASES RÉCIPROQUES

Revenons maintenant sur le concept d'espace dual mais tel qu'il est vu dans le cadre du calcul vectoriel. Cette deuxième approche peut peut-être aider certains à mieux comprendre le concept vu précédemment mais par contre masque la raison profonde de l'origine dénominations de "covariant" et "contravariant". Pourtant c'est la présentation la plus courante dans la littérature...

Soit une base quelconque $(\vec{e}_i)$ d'un espace vectoriel euclidien E_n . Par définition, n vecteurs $\vec{e}^k$ qui vérifient les relations suivantes:

$$\vec{e}_i \circ \vec{e}^k = \delta_{ik} \quad (14.131)$$

sont appelés les "**vecteurs réciproques**" des vecteurs $\vec{e}_i$. Ils seront notés avec des indices supérieurs. Par définition, chaque vecteur réciproque $\vec{e}^k$ se doit donc d'être orthogonal à tous les vecteurs $\vec{e}_i$, sauf pour $k = i$.

Montrons d'abord que les vecteurs réciproques $\vec{e}^k$ d'une base donnée $(\vec{e}_i)$ sont linéairement indépendants. Pour cela, il faut montrer qu'une combinaison linéaire $\lambda_k \vec{e}^k$ donne un vecteur nul, si et seulement si chaque coefficient λ_k est nul.

Soit $\vec{x} = x^i \vec{e}_i$ un vecteur quelconque de E_n . Multiplions scalairement par $\vec{x}$ la combinaison linéaire précédente $\lambda_k \vec{e}^k$, on obtient:

$$(\lambda_k \vec{e}^k) \circ \vec{x} = (\lambda_k \vec{e}^k) \circ (x^i \vec{e}_i) = \lambda_k x^i (\vec{e}^k \circ \vec{e}_i) = \lambda_k x^i \delta_{ik} = \lambda_k x^k = 0 \quad (14.132)$$

Cette dernière égalité devant être vérifiée quels que soient les x^i , il est nécessaire que chaque λ_k soit nul et ainsi les vecteurs $\vec{e}^k$ sont donc linéairement indépendants (fallait déjà avoir l'idée de procéder ainsi n'est-ce pas?).

Le système de n vecteurs réciproques forme donc une base appelée la "**base réciproque**" (qui n'est autre que la base duale) de l'espace vectoriel E_n .

Exemple:

Soient trois vecteurs $\vec{e}_1, \vec{e}_2, \vec{e}_3$ formant une base (non nécessairement orthonormée!) d'un espace vectoriel euclidien. Nous décidons de noter:

$$\vec{v} = \vec{e}_1 \circ (\vec{e}_2 \times \vec{e}_3) \quad (14.133)$$

où, rappelons-le, le symbole $\times$ représente le produit vectoriel (au cas où il y aurait un petit oubli...) et l'ensemble est le produit mixte vu dans le chapitre de Calcul Vectoriel et qui représente donc un volume orienté.

Les vecteurs suivants:

$$\vec{e}^1 = \frac{\vec{e}_2 \times \vec{e}_3}{v}; \vec{e}^2 = \frac{\vec{e}_3 \times \vec{e}_1}{v}; \vec{e}^3 = \frac{\vec{e}_1 \times \vec{e}_2}{v} \quad (14.134)$$

vérifient la relation $\vec{e}_i \circ \vec{e}^k = \delta_{ik}$ et constituent le système réciproque des vecteurs $\vec{e}_1, \vec{e}_2, \vec{e}_3$. En cristallographie, ces vecteurs constituent ce que nous appelons "l'espace de Fourier associé".

Remarque: Nous reconnaissons ici les relations que nous avons déjà obtenues lors de notre étude du déterminant de Gram.

Maintenant, considérons donc un vecteur sur la base d'origine $\vec{e}_1, \vec{e}_2, \vec{e}_3$ que nous noterons donc (comme déjà vu plus haut):

$$\vec{x} = x^i \vec{e}_i \quad (14.135)$$

avec donc par définition les composantes contravariantes du vecteur qui apparaissent comme nous l'avons défini plus haut (et donc nous avons en même temps expliqué l'origine du nom). Nous avons vu aussi plus haut que chaque composante contravariant sera aussi (naturellement et par extension) donnée par:

$$x^i = \vec{x} \circ \vec{e}^i \quad (14.136)$$

De façon similaire, nous avons donc les composantes covariantes qui apparaissent:

$$\begin{aligned} \vec{x} &= x_i \vec{e}^i \\ x_i &= \vec{x} \circ \vec{e}_i \end{aligned} \quad (14.137)$$

Dans cette approche, nous définissons alors le tenseur métrique contravariant et respectivement covariant par:

$$\begin{aligned} \vec{e}_i \circ \vec{e}_j &= g_{ij} = g_{ji} \\ \vec{e}^i \circ \vec{e}^j &= g^{ij} = g^{ji} \end{aligned} \quad (14.138)$$

Il vient alors par exemple pour les composantes contravariantes (dans le cas particulier de l'espace à trois dimensions), sachant que la démarche est la même pour les composantes covariantes:

$$\begin{aligned} x^j &= \vec{x} \circ \vec{e}^j = (x_1 \vec{e}^1 + x_2 \vec{e}^2 + x_3 \vec{e}^3) \circ \vec{e}^j = x_1 (\vec{e}^1 \circ \vec{e}^j) + x_2 (\vec{e}^2 \circ \vec{e}^j) + x_3 (\vec{e}^3 \circ \vec{e}^j) \\ &= x_1 g^{1j} + x_2 g^{2j} + x_3 g^{3j} = x_j g^{ji} \end{aligned} \quad (14.139)$$

Et donc nous retrouvons la relation de transformation entre composantes covariantes et contravariantes déjà vue plus haut à la différence que cela semble plus sortir du chapeau pas définition successives et que cela cache donc l'origine de la dénomination de ces mêmes composantes. Mais peut-être que certains lecteurs préfèrent cette approche.

7. TENSEURS EUCLIDIENS

La généralisation de la notion de vecteur nous a conduits à l'étude des espaces vectoriels à n dimensions. Les tenseurs sont également des vecteurs de dimension quelconque mais qui possèdent des propriétés supplémentaires par rapport aux vecteurs.

Pour le physicien théoricien, le calcul tensoriel s'intéresse en premier lieu à la manière dont les composantes des tenseurs se transforment lors d'un changement de base des espaces vectoriels dont ils sont issus. Nous commencerons donc à étudier ces propriétés vis-à-vis des changements de base (car c'est le cas le plus intéressant).

Un tenseur est, en pratique, souvent uniquement défini et utilisé sous forme de ses composantes. Ces dernières peuvent être exprimées sous forme covariante ou contravariante comme pour tout vecteur. Mais un nouveau type de composantes va apparaître pour les tenseurs, ce sont les "composantes mixtes". Ces trois types de composantes constituent des décompositions des tenseurs euclidiens sur des bases différentes.

7.1. TENSEUR FONDAMENTAL

Au cours de la théorie vue précédemment, nous avons utilisé les quantités g_{ij} , définies à partir du produit scalaire des vecteurs de base $(\vec{e}_i)$ d'un espace vectoriel pré-euclidien E_n à n dimensions, par:

$$g_{ij} = \vec{e}_i \circ \vec{e}_j \quad (14.140)$$

Ces n^2 quantités constituent les composantes covariantes d'un tenseur appelé le "**tenseur fondamental**" ou "**tenseur métrique**".

Etudions comment varient les quantités g_{ij} lorsque nous effectuons un changement de base:

Soit $(\vec{e}'_k)$ une autre base liée à la précédente par les relations connues:

$$\vec{e}_i = A^k{}_i \vec{e}'_k \quad \text{et} \quad \vec{e}'_k = A^i{}_k \vec{e}_i \quad (14.141)$$

Substituant la relation $\vec{e}_i = A^k{}_i \vec{e}'_k$ dans l'expression de g_{ij} , il vient (nous changeons les indices comme il se doit lors d'une substitution):

$$g_{ij} = (A^k{}_i \vec{e}'_k) \circ (A^l{}_j \vec{e}'_l) = (A^k{}_i A^l{}_j) \cdot (\vec{e}'_k \circ \vec{e}'_l) \quad (14.142)$$

Dans la nouvelle base $(\vec{e}'_k)$, les produits scalaires des vecteurs de base sont donc des quantités telles que:

$$g'_{kl} = \vec{e}'_k \circ \vec{e}'_l \quad (14.143)$$

Nous avons donc finalement pour l'expression des composantes covariantes g_{ij} lors d'un changement de base:

$$g_{ij} = (A^k{}_i \ A^l{}_j) g'_{kl} \quad (14.144)$$

Identiquement nous avons:

$$g'^i{}_j = (A^i{}_k \ A^j{}_l) g_{kl} \quad (14.145)$$

De manière générale, une suite de n^2 quantités t_{ij} qui se transforment, lors d'un changement de base de E_n , selon les deux relations précédentes, à savoir:

$$t'_{ij} = (A^k{}_i \ A^l{}_j) t'_{kl} \quad \text{et} \quad t_{kl} = (A^i{}_k \ A^j{}_l) t_{ij} \quad (14.146)$$

constituent, par définition, les "**composantes covariantes d'un tenseur d'ordre deux**" (à deux indices) sur E_n .

Nous pouvons ainsi manipuler des quantités exprimant les propriétés intrinsèques des bases comme des tenseurs normaux !

7.2. PRODUIT TENSORIEL DE DEUX VECTEURS

Considérons un espace vectoriel euclidien E_n de base $(\vec{e}_i)$ et soient deux vecteurs de E_n :

$$\vec{x} = x^i \vec{e}_i \quad \text{et} \quad \vec{y} = y^j \vec{e}_j \quad (14.147)$$

Formons les produits deux à deux des composantes contravariantes x^i et y^j , soit:

$$u^{ij} = x^i y^j \quad (14.148)$$

Nous obtenons ainsi n^2 quantités, si les deux vecteurs ont le même nombre de composantes, qui constituent également les composantes contravariantes d'un tenseur d'ordre deux appelé le "**produit tensoriel**" du vecteur $\vec{x}$ par le vecteur $\vec{y}$.

Par exemple pour $\vec{x}$ de dimension 2 et $\vec{y}$ de dimension 3 nous avons:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \otimes \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} x_1 y_1 \\ x_1 y_2 \\ x_1 y_3 \\ x_2 y_1 \\ x_2 y_2 \\ x_2 y_3 \end{pmatrix} \quad (14.149)$$

Nous pouvons bien évidemment construire des produits tensoriels d'ordre trois (donc avec n^3 termes) tels qu'avec le tenseur trois fois contravariant suivant:

$$u^{ijk} = x^i y^j z^k \quad (14.150)$$

etc...

Etudions les propriétés de changement de base de ces composantes. Utilisons pour cela les relations de changement de base des composantes contravariantes d'un vecteur, à savoir:

$$x^i = A_k^i x'^k \text{ et } x'^k = A_i^k x^i \quad (14.151)$$

Remplaçons dans la relation $u^{ij} = x^i y^j$ les composantes x^i et y^j par leur expression de changement de base, il vient:

$$u^{ij} = x^i y^j = (A_k^i x'^k)(A_l^j y'^l) = A_k^i A_l^j (x'^k y'^l) = (A_k^i A_l^j) u'^{kl} \quad (14.152)$$

Les quantités u'^{kl} sont les nouvelles composantes:

$$u'^{kl} = x'^k y'^l \quad (14.153)$$

La formule de transformation des n^2 quantités u'^{ij} lors d'un changement de base de E_n est donc finalement (très similaire au tenseur métrique):

$$u'^{ij} = A_k^i A_l^j u'^{kl} \quad (14.154)$$

Une telle relation de changement de base caractérise les composantes contravariantes d'un tenseur d'ordre deux. Inversement, nous obtenons:

$$u'^{kl} = A_i^k A_j^l u'^{ij} \quad (14.155)$$

Les n^2 quantités u'^{ij} constituent donc les "**composantes contravariantes d'un tenseur d'ordre deux**".

Nous pouvons former de même les produits deux à deux des composantes covariantes x_i et y_i des vecteurs $\vec{x}$ et $\vec{y}$ soit:

$$u_{ij} = x_i y_j \quad (14.156)$$

Les formules de changement de base des composantes covariantes des vecteurs sont données par les relations suivantes que nous avons déjà démontrées précédemment:

$$x_i = A_i^k x'_k \text{ et } x'_k = A_k^i x_i \quad (14.157)$$

Substituant la première relation dans le produit $u_{ij} = x_i y_j$, il vient:

$$u_{ij} = (A_i^k x'_k)(A_j^l y'_l) = A_i^k A_j^l (x'_k y'_l) = (A_i^k A_j^l) u'_{kl} \quad (14.158)$$

C'est la relation de changement de base des composantes covariantes d'un tenseur d'ordre deux. On vérifie que l'on a:

$$u'_{kl} = A_k^i A_l^j u_{ij} \quad (14.159)$$

Identiquement nous avons bien évidemment: $u^i_k = x^i_k y^i_l$ puisque $u^i_j = x_i y_i$.

Les n^2 quantités u^i_j constituent donc les "**composantes covariantes d'un tenseur d'ordre deux**".

Formons à présent n^2 quantités en multipliant deux à deux les composantes covariantes du vecteur $\vec{x}$ par les composantes contravariantes de $\vec{y}$, nous obtenons:

$$u^j_i = x_i y^j \quad (14.160)$$

Effectuons un changement de base dans cette dernière relation en tenant compte des expressions $x_i = A^k_i x^k$ et $x^i = A^i_k x^k$, on obtient:

$$u^j_i = x_i y^j = (A^k_i x^k)(A^j_l y^l) = A^k_i A^j_l (x^k y^l) = (A^k_i A^j_l) u^l_k \quad (14.161)$$

Cette relation de changement de base caractérise les "**composantes mixtes**" d'un tenseur d'ordre deux. Inversement, on peut vérifier que l'on a:

$$u^i_k = A^i_j A^j_l u^l_k \quad (14.162)$$

Ces composantes mixtes constituent également des composantes du produit tensoriel de $\vec{x}$ par $\vec{y}$, selon une certaine base.

De manière générale, une suite de n^2 quantités u^i_j qui se transforment, lors d'un changement de base de E_n , selon les relations établies juste précédemment constituent donc, par définition, les "**composantes mixtes d'un tenseur d'ordre deux**".

7.3. ESPACES TENSORIELS

Au cours de l'étude précédente, nous avons utilisé des systèmes de n^2 nombres, créés à partir d'un espace vectoriel E_n . Lorsque ces nombres vérifient certaines relations de changement de base, nous avons appelé ces grandeurs, par définition, les "**composantes d'un tenseur**".

Nous avons vu que toute combinaison linéaire de ces composantes constitue les composantes d'autres tenseurs. Nous pouvons donc additionner entre elles les composantes des tenseurs ainsi que les multiplier par des scalaires, pour obtenir d'autres composantes de tenseurs. Ces propriétés d'addition et de multiplication font que nous allons pouvoir utiliser ces grandeurs tensorielles comme composantes de vecteurs.

D'un point de vue pratique, nous pourrions nous contenter de définir les tenseurs à partir des relations de transformation de leurs composantes lors d'un changement de base. C'est ce qui est souvent fait en physique. Cependant, la définition des tenseurs sous forme de vecteurs conduit à une meilleure compréhension de leurs propriétés et les rattache à la théorie générale des vecteurs.

Pour préciser comment nous définissons un tenseur sur une base, étudions le cas particulier d'un produit tensoriel de deux vecteurs constitués par des triplets de nombres. Considérons l'espace vectoriel euclidien E_3 dont les vecteurs sont des triplets de nombre de la forme: $\vec{x} = (x_1, x_2, x_3)$. La base orthonormée canonique de E_3 est formée de trois vecteurs:

$$\vec{e}_i = (\delta_{1i}, \delta_{2i}, \delta_{3i}) \quad (14.163)$$

avec $i = 1, 2, 3$ (jolie façon d'écrire la chose n'est-il pas...).

Des vecteurs de E_3 permettent de former les neuf quantités $x^i y^j$ que nous avons appelées les "**composantes du produit tensoriel**" des vecteurs $\vec{x}$ et $\vec{y}$.

Si nous effectuons tous les produits tensoriels possibles entre vecteurs de E_3 , nous obtenons des suites de neuf nombres qui peuvent servir à définir le vecteur suivant:

$$\vec{U} = (x^1 y^1, x^1 y^2, x^1 y^3, \dots, x^3 y^3) \quad (14.164)$$

Remarque: Nous voyons de suite avec la relation précédente que le produit tensoriel n'est dès lors pas commutatif.

Nous nous retrouvons alors avec des éléments d'un espace vectoriel E_9 à neuf dimensions, ayant pour éléments tous les multiplets formés de neuf nombres.

Ces vecteurs peuvent être décomposés, par exemple, sur une base canonique orthonormée:

$$\vec{e}_k = (\delta_{1k}, \delta_{2k}, \dots, \delta_{9k}) \quad (14.165)$$

avec $k = 1, 2, \dots, 9$.

Si nous renumérotions les quantités $x^i y^j$ selon la place qu'elles occupent dans l'expression de $\vec{U}$, soit:

$$x^i y^j = u^{\#} = u^k \quad (14.166)$$

avec $k = 1, 2, \dots, 9$ et $i, j = 1, 2, 3$, les vecteurs $\vec{U}$ s'écrivent alors:

$$\vec{U} = u^k \vec{e}_k \quad (14.167)$$

et constituent un exemple de tenseur d'ordre deux (évidemment on peut généraliser la démarche).

En quoi ces tenseurs $\vec{U}$ diffèrent-ils des vecteurs ordinaires ? Ils sont certes identiques à certains vecteurs de E_9 mais ils ont été formés à partir des vecteurs de $\vec{x}$ et $\vec{y}$ de E_3 . Pour rappeler ce fait, nous les notons:

$$\boxed{\vec{U} = \vec{x} \otimes \vec{y}} \quad (14.168)$$

et ils sont appelés "**produits tensoriels d'ordre deux**" des vecteurs $\vec{x}$ et $\vec{y}$. Le symbole $\otimes$ est donc défini de la manière dont nous avons formé les quantités $x^i y^j = u^{\#}$ et l'ordre dans lequel nous les avons classées pour former le vecteur $\vec{U}$.

Pour rappeler la dépendance entre une quantité $x^i y^j = u^{\#}$ et le vecteur de base $\vec{e}_k$ auquel il est affecté, renumérotions ces vecteurs en mettant à la place de l'indice k les deux indices i et j , relatifs aux composantes, soit:

$$\vec{e}_k = \vec{e}_{ij} \quad (14.169)$$

Ce dernier peut très bien être noté sous la forme:

$$\vec{e}_{ij} = \vec{e}_i \otimes \vec{e}_j \quad (14.170)$$

Les vecteurs $\vec{e}_i \otimes \vec{e}_j$ constituent donc une base de E_9 qui est appelée la "**base associée**".

Nous rappelons également que le produit tensoriel est non-commutatif (il est vraiment important de s'en rappeler)! Autrement dit:

$$\vec{x} \otimes \vec{y} \neq \vec{y} \otimes \vec{x} \quad (14.171)$$

Les relations précédentes nous permettent finalement d'écrire le produit tensoriel des vecteurs $\vec{x}$ et $\vec{y}$ sous la forme:

$$\vec{x} \otimes \vec{y} = x^i y^j \vec{e}_{ij} = (x^i \vec{e}_i) \otimes (y^j \vec{e}_j) = x^i y^j (\vec{e}_i \otimes \vec{e}_j) \quad (14.172)$$

L'espace vectoriel E_9 est doté d'une structure plus précise que celle de simple espace vectoriel de dimension neuf lorsque nous définissons les produits tensoriels $\vec{e}_i \otimes \vec{e}_j$ comme constituant la base de E_9 . Nous disons que E_9 est doté d'une "**structure de produit tensoriel**" ce qui nous amène à noter cet espace $E_3 \otimes E_3$ ou encore $E_3^{(2)}$.

En tant qu'élément d'un espace $E_x \otimes E_x$, un tenseur $\vec{U}$ est un vecteur de la forme générale:

$$\vec{U} = u^{ij} (\vec{e}_i \otimes \vec{e}_j) \quad (14.173)$$

Etudions ses propriétés vis-à-vis d'un changement de base de E_x tel que:

$$\vec{e}_i = A_i^k \vec{e}'_k \text{ et } \vec{e}'_k = A_k^i \vec{e}_i \quad (14.174)$$

Lors d'un tel changement, la base $(\vec{e}_i \otimes \vec{e}_j)$ associée à $\vec{e}_i$ devient une autre base $(\vec{e}'_k \otimes \vec{e}'_l)$ associée à $\vec{e}'_k$, à savoir:

$$\vec{e}'_k \otimes \vec{e}'_l = (A_k^i \vec{e}_i) \otimes (A_l^j \vec{e}_j) = (A_k^i A_l^j) \vec{e}_i \otimes \vec{e}_j \quad (14.175)$$

Par suite, le produit tensoriel $\vec{U}$ a pour composantes dans la nouvelle base:

$$\vec{U} = u^{kl} (\vec{e}'_k \otimes \vec{e}'_l) = u^{kl} (A_k^i A_l^j) \vec{e}_i \otimes \vec{e}_j = (A_i^k A_j^l) u^{kl} (A_k^i A_l^j) \vec{e}_i \otimes \vec{e}_j \quad (14.176)$$

Soit donc:

$$(x^i \vec{e}_i) \otimes (y^j \vec{e}_j) = x^i y^j (\vec{e}_i \otimes \vec{e}_j) \quad (14.177)$$

Nous avons les propriétés suivantes pour le produit tensoriel:

P1. Distributivité, à gauche et à droite, par rapport à l'addition des vecteurs:

$$\begin{aligned}
 (\vec{x} + \vec{y}) \otimes \vec{z} &= \vec{x} \otimes \vec{z} + \vec{y} \otimes \vec{z} \\
 \vec{x} \otimes (\vec{y} + \vec{z}) &= \vec{x} \otimes \vec{y} + \vec{x} \otimes \vec{z}
 \end{aligned} \quad (14.178)$$

La démonstration de ces propriétés découle simplement de la définition du produit tensoriel. Nous avons par exemple:

$$\begin{aligned}
 (\vec{x} + \vec{y}) \otimes \vec{z} &= (x^i \vec{e}_i + y^j \vec{e}_j) \otimes z^k \vec{e}_k = (x^i + y^j) \vec{e}_i \otimes z^k \vec{e}_k = (x^i + y^j) z^k \vec{e}_i \otimes \vec{e}_k \\
 &= x^i z^k \vec{e}_i \otimes \vec{e}_k + y^j z^k \vec{e}_j \otimes \vec{e}_k = \vec{x} \otimes \vec{z} + \vec{y} \otimes \vec{z}
 \end{aligned} \quad (14.179)$$

P2. Associativité avec la multiplication par une grandeur scalaire:

$$(\lambda \vec{x}) \otimes \vec{y} = \vec{x} \otimes (\lambda \vec{y}) = \lambda (\vec{x} \otimes \vec{y}) \quad (14.180)$$

Nous avons en effet:

$$(\lambda \vec{x}) \otimes \vec{y} = (\lambda x^i \vec{e}_i) \otimes y^j \vec{e}_j = \lambda x^i y^j (\vec{e}_i \otimes \vec{e}_j) = \lambda (\vec{x} \otimes \vec{y}) \quad (14.181)$$

P3. Lorsque nous choisissons une base dans chacun des espaces vectoriels $(\vec{e}_i)$ pour E_n , $(\vec{f}_j)$ pour F_m , les $n \cdot m$ éléments de G_{nm} que nous notons $\vec{e}_i \otimes \vec{f}_j$ forment également une base de G_{nm} .

Démonstration:

Déjà faite dans l'exemple particulier que nous avons utilisé au début.

□C.Q.F.D.

Remarque: En pratique, nous avons souvent à utiliser des tenseurs formés à partir de vecteurs appartenant à des espaces vectoriels identiques E_n .

Nous pouvons bien évidemment généraliser le produit tensoriel à un nombre quelconque de vecteurs. De proche en proche, compte tenu de la propriété P1, nous pouvons considérer p vecteurs

$\vec{x}_1, \vec{x}_2, \dots, \vec{x}_p$ appartenant chacun à des espaces vectoriels différents $E_{n_1}, E_{n_2}, \dots, E_{n_p}$. Si nous avons:

$$\vec{x}_1 = x^{i_1} \vec{e}_{i_1}, \vec{x}_2 = x^{i_2} \vec{e}_{i_2}, \dots \quad (14.182)$$

nous pouvons former le produit tensoriel:

$$\vec{x}_1 \otimes \vec{x}_2 \otimes \dots \otimes \vec{x}_p = x^{i_1} x^{i_2} \dots x^{i_p} (\vec{e}_{i_1} \otimes \vec{e}_{i_2} \otimes \dots \otimes \vec{e}_{i_p}) \quad (14.183)$$

avec $i_1 = \{1, \dots, n_1\}, i_2 = \{1, \dots, n_2\}, \dots, i_p = \{1, \dots, n_p\}$.

Nous construisons ainsi des produits tensoriels d'ordre p appartenant à l'espace vectoriel

$E_{n_1} \otimes E_{n_2} \otimes \dots \otimes E_{n_p}$, espace qui est muni d'une structure de produit tensoriel. Les éléments de cet espace constituent par définition des tenseurs d'ordre p .

Afin d'unifier la classification, les espaces vectoriels élémentaires, qui ne peuvent être munis d'une structure de produit tensoriel, peuvent être considérés comme ayant pour éléments des tenseurs d'ordre un. En général, nous appelons ces éléments des "**vecteurs**", réservant le nom de "**tenseurs**" à des éléments d'espaces tensoriels d'ordre égal ou supérieur à deux!

Remarque: Il est commode d'appeler "**tenseurs d'ordre zéro**" les grandeurs scalaires. Il est également rare de rencontrer des tenseurs d'ordre supérieur à 2.

Il est assez évident et nous n'en ferons pas la démonstration (excepté s'il y a une demande) que nous pouvons redéfinir absolument tous les concepts (base, décomposition sur une base, base réciproque, produit scalaire, produit tensoriel) que nous avons vus jusqu'à maintenant en considérant les tenseurs d'ordre un comme des vecteurs (il faudrait donc que nous réécrivions tout ce qui est déjà écrit ci-dessus... ce qui est inutile).

Il est aussi tout à fait possible de réitérer toutes ces définitions pour des tenseurs d'ordre supérieurs et ainsi généraliser le concept d'espace tensoriel pour toutes les dimensions.

De ces considérations, nous pouvons énoncer le "**critère de tensorialité**":

Pour que les éléments d'une suite de n^2 quantités, rapportées à une base d'un espace vectoriel $E_n^{(p)}$, puissent être considérés comme les composantes d'un tenseur, il faut et il suffit que ces quantités soit liées entre elles, dans deux bases différentes de $E_n^{(p)}$, par les relations de transformation des composantes.

Exemple:

Un vecteur peut se représenter dans une base quelconque par une suite de n composantes. Cependant, nous ne pouvons pas conclure que n'importe quelle suite de n chiffres constitue un vecteur. En effet, lorsque nous nous plaçons dans une autre base de l'espace, les composantes doivent changer également, pour représenter le même objet: nous disons alors que le vecteur est un objet intrinsèque (dont l'existence ne dépend pas du choix du repère). Il reste alors à savoir qu'un vecteur est un tenseur d'ordre 1.

7.4. COMBINAISONS LINÉAIRES DE TENSEURS

Nous pouvons former d'autres tenseurs en combinant entre elles les composantes de différents produits tensoriels définis à l'aide des vecteurs d'un même espace vectoriel. Considérons par exemple les composantes contravariantes des produits tensoriels des vecteurs $\vec{x}, \vec{y}$ et $\vec{w}, \vec{z}$:

$$u^{\vec{y}} = x^i y^j \quad v^{\vec{z}} = w^i z^j \quad (14.184)$$

Formons les quantités suivantes:

$$t^{\vec{y}} = u^{\vec{y}} + \lambda v^{\vec{z}} \quad (14.185)$$

Les n^2 quantités $t^{\vec{y}}$ vérifient également les formules générales de changement de base. Nous avons en effet, en substituant les relations de transformation des composantes contravariantes d'un produit tensoriel dans l'expression précédente:

$$t^{\vec{y}} = A_k^i A_l^j u^{kl} + \lambda A_k^i A_l^j v^{kl} = A_k^i A_l^j t^{kl} \quad (14.186)$$

Les n^2 quantités t^{ij} , vérifiant la relation de changement de base, constituent donc également des composantes contravariantes d'un tenseur d'ordre deux.

7.5. CONTRACTION DES INDICES

Considérons le produit tensoriel mixte de deux vecteurs $\vec{x}$ et $\vec{y}$ de composantes respectives contravariantes x^i et covariantes y_j . Les composantes mixtes du produit tensoriel $\vec{v}$ de ces deux vecteurs, sont:

$$v_j^i = x^i y_j \quad (14.187)$$

Effectuons l'addition des différentes composantes du tenseur $\vec{v}$ telles que $i = j$, soit:

$$v = x^i y_i \quad (14.188)$$

Nous obtenons ainsi l'expression du produit scalaire des vecteurs $\vec{x}$ et $\vec{y}$; la quantité v est un scalaire ou tenseur d'ordre zéro. Une telle addition sur des indices de variance différente constitue, par définition, l'opération de "**contraction des indices**" du tenseur $\vec{v}$. Cette opération a permis de passer d'un tenseur d'ordre deux à un tenseur d'ordre zéro; le tenseur $\vec{v}$ a été amputé d'une covariance et d'une contravariance.

Prenons également l'exemple d'un tenseur $\vec{u}$ dont les composantes mixtes sont une fois covariante et deux fois contravariantes u_k^{ij} (attention... il ne s'agit pas d'une matrice tridimensionnelle mais simplement de l'indication que les composantes de ce tenseur s'expriment à partir de trois autres variables). Considérons certaines de ses composantes telles que $j = k$, à savoir les quantités u_j^{ij} et effectuons l'addition de ces dernières. Nous obtenons alors:

$$v^i = u_1^{i1} + u_2^{i2} + \dots + u_n^{in} = \delta_j^k u_k^{ij} \quad (14.189)$$

Ces nouvelles quantités v^i forment les composantes d'un tenseur $\vec{v}$ d'ordre un (donc un vecteur!) et constituent ce que nous appelons alors les "**composantes contractées**" du tenseur $\vec{u}$ et satisfont bien évidemment aux relations de changement de base (sur demande nous pouvons faire la démonstration mais sachez qu'elle est similaire à celle que nous avons faite pour les vecteurs). Nous sommes ainsi passé d'un tenseur d'ordre trois à un tenseur d'ordre un.

Si nous partons d'un tenseur de composantes contravariantes ou covariantes, nous pouvons abaisser/élever un ou plusieurs des indices par multiplication (le cas échéant répétée) par g_{ij} ou g^{ij} (métrique diagonale unitaire et à signature positive: de type canonique) afin d'obtenir des composantes mixtes sur lesquelles nous pourrions ensuite effectuer des opérations de contraction.

Considérons un tenseur euclidien $\vec{u}$ de composantes contravariantes $u^{i_1 i_2 i_3 \dots i_p}$.

Si nous voulons effectuer une contraction sur ce tenseur, il nous faudra d'abord le transformer en un tenseur mixte. Cette transformation se fera à l'aide d'un tenseur fondamental.

Écrivons $\vec{u}$ en composantes mixtes en abaissant à la position covariante l'indice i_1 par exemple (cela revient donc à exprimer cette composante contravariante en composante covariante). Alors:

$$u_{j_1}^{i_2 i_3 \dots i_p} = g_{i_1 j_1} u_{i_1}^{i_2 i_3 \dots i_p} \quad (14.190)$$

Nous voyons bien que dans le cas présent pour descendre un indice contravariant dans un tenseur au moyen d'un tenseur fondamental, il faut d'abord aller rechercher dans les indices covariants du tenseur fondamental celui qui se retrouve en contravariant dans le tenseur d'origine et le remplacer à sa position (mais cette fois en covariance) par l'autre indice du tenseur fondamental (il en est de même lorsque l'on souhaite monter un indice dans le cas où l'on souhaiterait opérer une contraction sur un tenseur covariant).

Effectivement, rappelons que nous avons démontré que:

$$x_j = g_{ij} x^i \quad (14.191)$$

Maintenant que nous avons obtenu un tenseur à composantes mixtes, nous pouvons très bien en plus contracter les indices. Choisissons par exemple l'indice i_2 et effectuons la contraction avec l'indice j_1 , posons $i_2 = j_1 = k$ (nous ne nous intéressons plus alors qu'à certains termes particuliers), il vient alors en écrivant toute la démarche depuis le début:

$$g_{i_1 j_1} u_{i_1}^{i_2 i_3 \dots i_p} = u_{j_1}^{i_2 i_3 \dots i_p} = u_k^{i_2 i_3 \dots i_p} = v^{i_2 \dots i_p} \quad (14.192)$$

Nous obtenons donc après abaissement de l'indice et contraction, un tenseur d'ordre $p - 2$.

Remarques:

R1. La deuxième égalité de l'expression précédente est une notation abusive que l'on retrouve dans certains ouvrages (car rigoureusement il faudrait faire le calcul en deux étapes).

R2. Par suite de la symétrie des quantités g_{ij} (produit scalaire est commutatif) ce dernier tenseur est identique à celui que nous obtiendrions en abaissant à la position covariante l'indice i_2 puis en effectuant la contraction avec l'indice i_1 .

Voyons cela :

La symétrie $g_{ij} = g_{ji}$ prend ici la forme :

$$g_{i_1 j_1} = g_{i_2 j_2} \quad (14.193)$$

(cela peut paraître déroutant mais rappelons-nous que le chiffre d'une composante i indique la place de cette composante)

Donc il vient :

$$g_{j_2 i_2} u^{i_1 i_2 i_3 \dots i_p} = u^{i_1 i_2 i_3 \dots i_p} \quad (14.194)$$

et en posant $i_1 = j_2 = k$:

$$v^{i_2 i_3 \dots i_p} = u^k_{i_2} u^{k i_2 i_3 \dots i_p} \quad (14.195)$$

De manière générale, la contraction d'un tenseur permet donc de former un tenseur d'ordre $p - 2$ à partir d'un tenseur d'ordre p . Nous pouvons naturellement répéter l'opération de contraction. Ainsi, un tenseur pair, $2p$, deviendra un scalaire après p contractions et un tenseur d'ordre impair, $2p + 1$, deviendra un vecteur.

Nous pouvons étendre après cette définition de la contraction des indices, le critère de tensorialité. Nous avons vu jusqu'à maintenant, deux manières de reconnaître le caractère tensoriel d'une suite de quantités:

- la première consiste à démontrer que ces quantités sont formées par le produit tensoriel de composantes de vecteurs ou par une somme de produits tensoriels;
- la deuxième consiste à étudier la manière dont ces quantités se transforment lors d'un changement de base et à vérifier la conformité des relations de transformation;

- la troisième et nouvelle amène à poser que pour qu'un ensemble de n^{p+q} quantités, comportant p indices supérieurs et q indices inférieurs soit tensoriel, il faut et il suffit que leur produit complètement contracté par les composantes contravariantes de p vecteurs quelconques et les composantes covariantes de q vecteurs quelconques, soit une quantité (la norme au fait...) qui demeure invariante par changement de base.

8. TENSEURS PARTICULIERS

Nous pouvons être confrontés en physique théorique à des tenseurs qui ont des propriétés intéressantes. Afin d'éviter de faire un travail redondant au cas par cas, nous allons énumérer et démontrer les différentes propriétés existantes et parler de leurs possibles implications.

8.1. TENSEUR SYMÉTRIQUE

Considérons un tenseur $\vec{U}$ d'ordre deux de composantes contravariantes $u^{\vec{ij}}$. Supposons que, suivant une base $(\vec{e}_i)$, toutes ces composantes satisfassent aux relations:

$$u^{\vec{ij}} = u^{\vec{ji}} \quad (14.196)$$

Sur une autre base $(\vec{e}'_j)$, liée à la précédente par les relations de transformation connues, les nouvelles composantes de $u^{\vec{lm}}$ vérifient la relation:

$$u^{\vec{lm}} = A^{\vec{l}}_i A^{\vec{m}}_j u^{\vec{ij}} = A^{\vec{m}}_j A^{\vec{l}}_i u^{\vec{ji}} \quad (14.197)$$

Nous voyons que la propriété $u^{\vec{ij}} = u^{\vec{ji}}$ est donc une caractéristique intrinsèque du tenseur $\vec{U}$, indépendante de la base ! Nous disons alors que le tenseur est un "**tenseur symétrique**" (nous reviendrons sur cette notion un peu plus loin).

La propriété de symétrie se vérifie également pour les composantes covariantes d'un tenseur symétrique puisque nous avons:

$$u_{\vec{kl}} = g_{\vec{li}} g_{\vec{kj}} u^{\vec{ij}} = g_{\vec{kj}} g_{\vec{li}} u^{\vec{ji}} = u_{\vec{lk}} \quad (14.198)$$

Réciproquement, la symétrie des composantes covariantes entraîne celle des composantes contravariantes.

Pour des tenseurs d'ordre plus élevé, la symétrie peut être partielle, portant sur deux indices covariants ou deux indices contravariants. Ainsi, un tenseur d'ordre quatre, de composantes mixtes $u^{\vec{ijk}}_l$ peut être également symétrique en i et j , par exemple, soit:

$$u^{\vec{ijk}}_l = u^{\vec{jik}}_l \quad (14.199)$$

Nous vérifions, de même que ci-dessus, qu'une telle propriété est intrinsèque.

Un tenseur est dit "**tenseur complètement symétrique**" si toute transposition de deux indices de même variance, change la composante correspondante en elle-même. Par exemple, pour un tenseur d'ordre trois $u^{\vec{ijk}}$, complètement symétrique, nous avons les composantes suivantes qui sont égales entre elles:

$$u^{\vec{ijk}} = u^{\vec{jik}} = u^{\vec{kji}} = u^{\vec{ikj}} \quad (14.200)$$

Des exemples de tenseurs complémentaires symétriques sont le tenseur des contraintes σ_{ij} que nous verrons lors de notre étude des équations de Navier-Stokes en mécanique des fluides et les tenseurs des transformations relativistes de Lorentz que nous verrons en mécanique relativiste. Ces tenseurs sont alors dits aussi "**tenseurs totalement invariants**" (sous-entendu par changement de base).

Nous pouvons également (curiosité intéressante) obtenir une représentation géométrique des valeurs des composantes d'un tenseur symétrique d'ordre deux. Pour cela, considérons dans l'espace géométrique ordinaire des coordonnées x^i , l'équation suivante:

$$a_{ij} x^i x^j = 1 \quad (14.201)$$

où, rappelons-le, $x^i x^j$ peut être vu comme un produit tensoriel avec $i, j = 1, 2, 3$ et où les a_{ij} sont des coefficients réels donnés. Supposons que ces coefficients soient tels que:

$$a_{ij} = a_{ji} \quad (14.202)$$

L'équation précédente s'écrit alors:

$$a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2 + 2a_{12}x_1x_2 + 2a_{23}x_2x_3 + 2a_{31}x_3x_1 = 1 \quad (14.203)$$

Nous retrouvons ici l'équation d'une surface de second degré ou quadrique similaire à celle du plan que nous avons vue en géométrie plane. Nous savons par extension à la troisième dimension que ces surfaces sont des ellipsoïdes ou hyperboloïdes, selon les valeurs des quantités a_{ij} .

Étudions comment se transforment les quantités a_{ij} lorsque nous effectuons un changement de coordonnées tel que:

$$x'^k = A^k_i x^i \text{ et } x^i = A^i_k x'^k \quad (14.204)$$

L'équation de la quadrique s'écrit dans ce nouveau système de coordonnées:

$$a_{ij} A^i_k A^j_m x'^k x'^m = a'_{km} x'^k x'^m = 1 \quad (14.205)$$

d'où l'expression des coefficients dans le nouveau système d'axes:

$$a'_{km} = A^i_k A^j_m a_{ij} \quad (14.206)$$

Les coefficients a_{ij} se transforment donc comme les composantes covariantes d'un tenseur d'ordre deux. Réciproquement, si les quantités a_{ij} sont les composantes d'un tenseur symétrique, ces composantes définissent les coefficients d'une quadrique. Il existe donc une certaine équivalence entre un tenseur symétrique et les coefficients d'une quadrique. Nous dirons que l'équation de la quadrique est la "**quadrique représentative**" du tenseur symétrique.

Nous savons de par notre étude des quadriques en géométrie plane (en étendant cela au cas tridimensionnel) que nous pouvons toujours trouver un système de coordonnées par rapport auquel l'équation d'une quadrique prend une forme plus simple:

$$b_1 x_1^2 + b_2 x_2^2 + b_3 x_3^2 = 1 \quad (14.207)$$

Dans ce cas, les vecteurs de base sont portés par les axes principaux de la quadrique. Dans ce système de coordonnées, les composantes du tenseur a_{ij} se réduisent à:

$$a_{11} = b_1, a_{22} = b_2, a_{33} = b_3 \quad (14.208)$$

et $a_{ij} = 0$ pour les autres composantes. Les quantités b_i sont appelées les "**composantes principales**" du tenseur a_{ij} .

Si les quantités b_1, b_2, b_3 sont positives, la surface est un ellipsoïde, si deux quantités sont strictement positives et la troisième strictement négative, nous avons un hyperboloïde à une nappe, si deux quantités sont strictement négatives et la troisième positive, nous avons un hyperboloïde à deux nappes (pour plus d'information voir le chapitre de Géométrie Analytique).

La comparaison de l'expression de la quadrique obtenue précédemment avec l'équation classique:

$$\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} + \frac{x_3^2}{c^2} = 1 \quad (14.209)$$

où a, b, c sont les demi-axes d'un ellipsoïde montre que nous avons:

$$a = \frac{1}{\sqrt{b_1}}, b = \frac{1}{\sqrt{b_2}}, c = \frac{1}{\sqrt{b_3}} \quad (14.210)$$

8.2. TENSEUR ANTISYMETRIQUE

Lorsque les composantes contravariantes u^{ij} d'un tenseur d'ordre deux, vérifient les relations:

$$u^{ij} = -u^{ji} \quad (14.211)$$

nous disons que le tenseur est un "**tenseur antisymétrique**" (il en va de même si les composantes sont covariantes). C'est une propriété intrinsèque du tenseur qui se démontre comme pour les tenseurs symétriques, au signe "-" près. Un tenseur contravariant d'ordre deux, vérifiant donc la relation suivante sera dit "**tenseur symétrique**" (nous l'avons déjà mentionné un paquet de fois dans les paragraphes précédents):

$$u^{ij} = u^{ji} \quad (14.212)$$

Et il en va de même si les composantes sont covariantes.

Un tenseur antisymétrique doit bien évidemment satisfaire au fait que ces composantes diagonales soient nulles telles que:

$$u^{11} = u^{22} = \dots = u^{nn} = 0 \quad (14.213)$$

Si les composantes contravariantes d'un tenseur sont antisymétriques, ses composantes covariantes le sont également.

Un tenseur par exemple covariant d'ordre trois T_{ijk} sera dit symétrique en i et k si pour toutes les valeurs que peuvent prendre les indices, nous avons:

$$T_{ijk} = T_{kji} \quad (14.214)$$

Ou encore le tenseur covariant T_{ijkl} sera dit antisymétrique en i et l si pour toutes les valeurs que peuvent prendre les indices, nous avons:

$$T_{ijkl} = -T_{iljk} \quad (14.215)$$

Un tenseur $u_i^{\tilde{j}k}$ sera partiellement antisymétrique si nous avons par exemple:

$$u_i^{\tilde{j}k} = -u_i^{\tilde{k}j} \quad (14.216)$$

Il sera complètement antisymétrique si toute transposition d'indice de même variance change la composante correspondante en son opposée.

Tout tenseur $u^{\tilde{j}}$ peut être mis sous la forme d'une somme d'un tenseur symétrique et d'un tenseur antisymétrique. Nous avons en effet:

$$u^{\tilde{j}} = \frac{1}{2}(u^{\tilde{j}} + u^{\tilde{j}}) + \frac{1}{2}(u^{\tilde{j}} - u^{\tilde{j}}) \quad (14.217)$$

Le premier terme de la somme ci-dessus est un tenseur symétrique et le second, un tenseur antisymétrique.

Considérons maintenant deux vecteurs $\vec{x} = x^i \vec{e}_i$ et $\vec{y} = y^j \vec{e}_j$ d'un espace vectoriel E_n . Formons les quantités antisymétriques suivantes (nous y trouvons deux produits tensoriels):

$$u^{\tilde{j}} = x^i y^j - x^j y^i \quad (14.218)$$

où nous voyons immédiatement que les composantes $u^{\tilde{j}}$ sont celles d'un tenseur antisymétrique $\vec{U}$.

La décomposition du vecteur $\vec{U}$ dans la base $\vec{e}_i \otimes \vec{e}_j$ s'écrit:

$$\vec{U} = (\vec{x} \times \vec{y}) = (x^i y^j - x^j y^i) \vec{e}_i \otimes \vec{e}_j = \vec{x} \otimes \vec{y} - \vec{y} \otimes \vec{x} \quad (14.219)$$

Le tenseur $\vec{x} \times \vec{y}$ (noté ainsi en analogie avec le produit vectoriel pour $n = 3$) est appelé le "**produit extérieur**" des vecteurs $\vec{x}$ et $\vec{y}$. Nous disons encore que ce tenseur est un "**bivecteur**".

Le produit extérieur est donc un tenseur antisymétrique qui vérifie les propriétés suivantes:

P1. Anticommutativité: $\vec{x} \times \vec{y} = -\vec{y} \times \vec{x}$, il en résulte:

$$\vec{x} \times \vec{x} = \vec{0} \quad (14.220)$$

P2. Distributivité à gauche et à droite pour l'addition vectorielle:

$$\begin{aligned} \vec{x} \times (\vec{y} + \vec{z}) &= \vec{x} \times \vec{y} + \vec{x} \times \vec{z} \\ (\vec{x} + \vec{y}) \times \vec{z} &= \vec{x} \times \vec{z} + \vec{y} \times \vec{z} \end{aligned} \quad (14.221)$$

P3. Associativité pour la multiplication par un scalaire:

$$\alpha \vec{x} \times \vec{y} = \vec{x} \times \alpha \vec{y} = \alpha (\vec{x} \times \vec{y}) \quad (14.222)$$

P4. Les produits extérieurs:

$$\vec{e}_i \times \vec{e}_j = \vec{e}_i \otimes \vec{e}_j - \vec{e}_j \otimes \vec{e}_i \quad (14.223)$$

constituent une base de l'ensemble des bivecteurs.

Démonstration:

Un tenseur antisymétrique $\vec{U}$ d'ordre deux, élément de $E_n^{(2)}$, peut s'écrire sous la forme:

$$\vec{U} = u^{ij} \vec{e}_i \otimes \vec{e}_j = \sum_{i,j} u^{ij} \vec{e}_i \otimes \vec{e}_j + \sum_{i,j} u^{ji} \vec{e}_j \otimes \vec{e}_i \quad (14.224)$$

Échangeant, dans la dernière somme de la relation ci-dessus, le nom des indices et en tenant compte que $u^{ji} = -u^{ij}$, nous obtenons:

$$\vec{U} = \sum_{i,j} u^{ij} (\vec{e}_i \otimes \vec{e}_j - \vec{e}_j \otimes \vec{e}_i) \quad (14.225)$$

Les éléments:

$$\vec{e}_i \otimes \vec{e}_j - \vec{e}_j \otimes \vec{e}_i \quad (14.226)$$

sont linéairement indépendants puisque les vecteurs $\vec{e}_i \otimes \vec{e}_j$ le sont également. Ces éléments constituent donc une base sur laquelle les tenseurs antisymétriques peuvent être décomposés.

□C.Q.F.D.

Le nombre de vecteurs $\vec{e}_i \otimes \vec{e}_j - \vec{e}_j \otimes \vec{e}_i$ distinguables est égal au nombre de combinaisons de vecteurs pris deux à deux et distinguables parmi n tel que:

$$C_n^2 = \frac{n(n-1)}{2} < n^2 \quad (14.227)$$

Effectivement parmi les n^2 composantes, n composantes sont nulles et les $n(n-1)$ autres composantes ont des valeurs opposées deux à deux. Nous pouvons donc considérer que la moitié de ces dernières suffit à caractériser le tenseur.

Dans le cadre du produit tensoriel extérieur où nous avons:

$$u^{ij} = x^i y^j - x^j y^i \quad (14.228)$$

le nombre de composantes distinguables est également de $n(n-1)/2$ et elles sont appelées "**composantes strictes**".

Nous remarquons que pour $n = 3$, le nombre de composantes strictes du produit extérieur de deux vecteurs est aussi égal à trois. Ceci permet de former avec les composantes du bivecteur, les composantes d'un produit vectoriel $\vec{z}$.

Ainsi, un produit vectoriel n'existe donc que pour un sous-espace de bivecteurs dont le nombre de dimensions est égal à 3 et dont les pré-images sont des tenseurs antisymétriques.

Si toutes ces conditions sont satisfaites, nous disons que le vecteur $\vec{z}$ constitue le "**tenseur adjoint**" du tenseur $\vec{U}$.

8.3. TENSEUR FONDAMENTAL

Nous avons vu au début de notre étude du calcul tensoriel la définition des composantes covariantes g_{ij} du tenseur fondamental, à savoir:

$$g_{ij} = \vec{e}_i \circ \vec{e}_j \quad (14.229)$$

Ces quantités interviennent, nous le savons, dans l'expression du produit scalaire de deux vecteurs $\vec{x}$ et $\vec{y}$, de composantes contravariantes x^i et y^i , donné par la relation:

$$\vec{x} \circ \vec{y} = g_{ij} x^i y^j \quad (14.230)$$

Utilisons le critère général de tensorialité pour mettre en évidence le caractère tensoriel des g_{ij} .

L'expression précédente est un produit complètement contracté des quantités g_{ij} avec les composantes contravariantes $x^i y^j$ d'un tenseur arbitraire. Comme le produit scalaire est une quantité invariante (en l'occurrence un scalaire) par rapport aux changements de base, il en résulte que les n^2 quantités g_{ij} sont les composantes covariantes d'un tenseur.

Ce tenseur est de plus symétrique par suite de la propriété de symétrie du produit scalaire des vecteurs de base telle que:

$$g_{ij} = g_{ji} \quad (14.231)$$

Nous avons de même pour les composantes contravariantes du tenseur fondamental:

$$g^{ij} = g^{ji} \quad (14.232)$$

Si nous notons g^i_j les composantes mixtes du tenseur fondamental à lui-même:

$$g^i_j = g^{ik} g_{jk} \quad (14.233)$$

avec évidemment dans la base canonique:

$$g^i_j = \delta^i_j \quad (14.234)$$

9. COORDONNÉES CURVILIGNES

Les notions classiques de système de coordonnées peuvent être généralisées à des espaces ponctuels (voir le chapitre traitant des Principes) à n dimensions. Nous appelons "**système de coordonnées**" dans $\mathcal{E}_n$ (espace ponctuel à n dimensions donc), tout mode de définition d'un point M dans le système considéré.

Pour un système donné de coordonnées (cartésiennes, sphériques, cylindriques, polaires...), nous appelons "**ligne de coordonnées**" le "lieu" des points M lorsqu'une seule coordonnée varie, les autres étant égales à des constantes.

Étudions tout d'abord la généralisation d'un système de coordonnées relatives à un repère fixe (nous conseillons vivement au lecteur d'avoir lu au préalable la partie traitant des systèmes de coordonnées dans le chapitre de Calcul Vectoriel et la partie traitant du formalisme lagrangien dans le chapitre Principes).

Considérons un espace ponctuel $\mathcal{E}_n$ et un repère $(\vec{e}_i)$ de cet espace. Soit x^i les coordonnées rectilignes d'un point M de $\mathcal{E}_n$ par rapport à ce repère. Un système de coordonnées quelconque u^k , $k = 1, 2, \dots, n$, est obtenu en se donnant n fonctions arbitraires f^i des paramètres u^k , telles que:

$$x^i = f^i(u^1, u^2, \dots, u^n) \quad (14.235)$$

Nous supposons par la suite que ces n fonctions satisfont aux trois propriétés suivantes:

P1. Elles sont de classe supérieure ou égale à C^2 (dérivables au moins deux fois pour les besoins de la physique). Cette hypothèse implique, en tout point où elle est satisfaite, que nous avons la permutabilité des dérivations (par rapport aux deux dérivations):

$$\partial_{u^k} f^i = \partial_{x^k} f^i \quad (14.236)$$

P2. Ces fonctions sont telles que nous pouvons résoudre le système des n équations de changement de système de coordonnées par rapport aux variables u^k et les exprimer en fonction des x^i , soit:

$$u^k = g^k(x^1, x^2, \dots, x^n) \quad (14.237)$$

toujours avec $k = 1, 2, \dots, n$.

P3. Lorsque les variables u^k varient dans un domaine Δ , les variables x^i varient dans un domaine Δ' . Le jacobien des fonctions $x^i = f^i(u^1, u^2, \dots, u^n)$, défini par:

$$D(\partial_k x^i) = \begin{bmatrix} \partial_1 x^1 & \partial_1 x^2 & \dots & \partial_1 x^n \\ \partial_2 x^1 & \partial_2 x^2 & \dots & \partial_2 x^n \\ \dots & \dots & \dots & \dots \\ \partial_n x^1 & \partial_n x^2 & \dots & \partial_n x^n \end{bmatrix} \quad (14.238)$$

sera supposé différent de zéro dans le domaine Δ (ainsi que le jacobien $D(\partial_k u^i)$ des fonctions $u^k = g^k(x^1, x^2, \dots, x^n)$) et est l'inverse du jacobien de $u^k = g^k(x^1, x^2, \dots, x^n)$. Si les jacobiens existent, ils sont non nuls comme conséquence en premier lieu de la deuxième propriété ci-dessus et implicitement de la première.

Si nous fixons $(n-1)$ paramètres u^k en faisant varier un seul paramètre, u^i par exemple, nous obtenons les coordonnées $x_{(1)}^i$ d'un ensemble de points M de $\mathcal{E}_n$ qui constituent une "ligne de coordonnées".

En général, les lignes de coordonnées ne sont pas des droites mais des courbes; ces coordonnées u^k sont appelées pour cette raison des "coordonnées curvilignes". En un point M de $\mathcal{E}_n$ se croisent d'ailleurs n lignes de coordonnées.

Nous démontrons en mécanique analytique, lors de l'étude des espaces ponctuels, que les dérivées et les différentielles d'un vecteur $\overrightarrow{OM}$ de $\mathcal{E}_n$ sont indépendantes du point O d'un repère donné. Si $\mathcal{E}_n$ est rapporté à un système de coordonnées curvilignes u^k , nous écrivons:

$$\boxed{\vec{e}_k = \frac{\partial \overrightarrow{OM}}{\partial u^k} = \partial_k \overrightarrow{M}} \quad (14.239)$$

Exemple:

Un exemple de coordonnées curvilignes u^k , où chaque u^k est une fonction uniforme des coordonnées rectilignes x^k , les u^k étant de plus des fonctions continues au point courant M , est celui des coordonnées sphériques où nous avons (cf. chapitre de Calcul Vectoriel):

$$u^1 = r, u^2 = \theta, u^3 = \phi \quad (14.240)$$

Rappelons aussi que lors de notre étude du système de coordonnées sphériques en calcul vectoriel nous avons obtenu:

$$d\vec{M} = \vec{e}_r dr + \vec{e}_\theta r d\theta + \vec{e}_\phi r \sin \theta d\phi \quad (14.241)$$

Ainsi, nous voyons bien cette dépendance sous l'expression des relations suivantes:

$$\begin{aligned} \frac{\partial \vec{M}}{\partial r} &= \vec{e}_r & \frac{\partial \vec{M}}{\partial \theta} &= \vec{e}_\theta r & \frac{\partial \vec{M}}{\partial \phi} &= \vec{e}_\phi r \sin \theta \\ \left\| \frac{\partial \vec{M}}{\partial r} \right\| &= 1 & \left\| \frac{\partial \vec{M}}{\partial \theta} \right\| &= r & \left\| \frac{\partial \vec{M}}{\partial \phi} \right\| &= r \sin \theta \end{aligned} \quad (14.242)$$

Dans un espace non-euclidien, nous ne pouvons définir une base valable sur tout l'espace. Ainsi, nous construisons une base en chaque point séparément et pour cela, nous utilisons bien les coordonnées curvilignes telles qu'en chaque point M , les vecteurs de base $\vec{e}_k$ sont tangents à la ligne de coordonnées correspondante u^k via la relation donnée plus haut:

$$\vec{e}_k = \frac{\partial \overrightarrow{OM}}{\partial u^k} = \partial_k \overrightarrow{M} \quad (14.243)$$

Soient maintenant $u^1, u^2, \dots, u^n$ les coordonnées curvilignes du point M par rapport à un repère cartésien $\vec{e}_i^0$. Dans ce repère, nous avons bien évidemment:

$$\overrightarrow{OM} = x^i \vec{e}_i^0 \quad (14.244)$$

où les coordonnées cartésiennes sont des fonctions $x^i = x^i(u^1, u^2, \dots, u^n)$.

Le vecteur $\vec{e}_k$ a donc pour expression:

$$\vec{e}_k = \partial_k \left(x^i \vec{e}_i^0 \right) = \left(\partial_k x^i \right) \vec{e}_i^0 \quad (14.245)$$

A partir des composantes $\partial_k x^i$ du vecteur $\vec{e}_k$, nous pouvons former un déterminant $D(\partial_k x^i)$ qui est précisément le jacobien des fonctions x^i que nous avons défini précédemment. Puisque ce déterminant est différent de zéro (du moins imposé tel quel), il en résulte que les n vecteurs $\vec{e}_k$ sont linéairement indépendants.

Ces n vecteurs, définis par la relation:

$$\vec{e}_k = \partial_k \vec{M} \quad (14.246)$$

sont appelés la "**base naturelle**" au point M de l'espace vectoriel E_M . Ils sont colinéaires aux tangentes des n lignes coordonnées qui se coupent au point M où ils sont définis.

Nous n'insisterons pas sur le fait évident qu'à tout système de coordonnées curvilignes est associé un repère naturel dont la base est exprimée par ses mêmes coordonnées (cf. chapitre de Calcul Vectoriel).

Exemple:

En coordonnées sphériques, les vecteurs de la base naturelle sont ceux que nous avons obtenus lors de notre étude du système de coordonnées sphériques dans le chapitre de Calcul Vectoriel et qui sont orthogonaux mais non orthonormés.

Associations au point M de E_M un repère formé par le point M et par les vecteurs de la base naturelle. Ce repère est appelé le "**repère naturel**" en M du système de coordonnées u^k . Il sera noté:

$$(M, \vec{e}_k) \text{ ou } (M, \partial_k \vec{M}) \quad (14.247)$$

La différentielle du vecteur $\vec{OM}$ s'exprime alors sous la forme:

$$d\vec{M} = \partial_k \vec{M} du^k = \vec{e}_k du^k \quad (14.248)$$

Les quantités du^k constituent les composantes contravariantes du vecteur $d\vec{M}$ dans le repère naturel $(M, \vec{e}_k)$ du système de coordonnées u^k .

Considérons maintenant deux systèmes quelconques de coordonnées curvilignes u^i et $u^{i'}$, liées entre elles par les relations:

$$u^i = u^i(u^{i'}, u^{i''}, \dots, u^{i'''}) \quad u^{i'} = u^{i'}(u^1, u^2, \dots, u^n) \quad (14.249)$$

où les fonctions $u^i = u^i(u^1, u^2, \dots, u^n)$ sont supposées plusieurs fois continument dérivables par rapport aux u^{ik} et de même pour les fonctions $u^{ik} = u^{ik}(u^1, u^2, \dots, u^n)$ par rapport aux coordonnées u^i . Lorsque nous passons d'un système de coordonnées à un autre, nous disons que nous effectuons un "**changement de coordonnées curvilignes**".

Nous avons vu en relativité générale que le carré de la distance ds^2 entre deux points M et M' infiniment proches est donné par la relation:

$$ds^2 = g_{ij} dx^i dx^j \quad (14.250)$$

où les dx^i sont les composantes du vecteur $d\vec{M} = \overrightarrow{MM'}$, rapportées à un repère fixe d'un espace ponctuel $\mathcal{E}_x$. Lorsque cet espace est rapporté à un système de coordonnées curvilignes u^i , nous avons vu que la relation:

$$d\vec{M} = \partial_k \vec{M} du^k = \vec{e}_k du^k \quad (14.251)$$

montre que le vecteur $d\vec{M}$ a pour composantes contravariantes les quantités du^i par rapport au repère naturel $(M, \vec{e}_i)$. Le carré de la distance ds^2 s'écrit alors dans le repère naturel:

$$ds^2 = g_{ij} du^i du^j \quad (14.252)$$

où les quantités $g_{ij} = \vec{e}_i \circ \vec{e}_j$ sont les composantes du tenseur fondamental ou du tenseur métrique définies à l'aide d'une base naturelle. L'expression précédente s'appelle "**l'élément linéaire de l'espace ponctuel**" $\mathcal{E}_x$ ou encore la "**métrique**" de cet espace.

Les vecteurs $\vec{e}_i$ du repère naturel varient en général d'un point à un autre. C'est le cas, par exemple, des coordonnées sphériques dont les quantités g_{ij} (nous le démontrerons de suite après) sont variables !!

Une courbe Γ de $\mathcal{E}_x$ peut être définie par la donnée des coordonnées curvilignes $u^i(\alpha)$ du lieu des points $M(\alpha)$ en fonction d'un paramètre α . La distance élémentaire ds sur cette courbe Γ s'écrit alors:

$$ds = \left[g_{ij} \frac{\partial u^i}{\partial \alpha} \frac{\partial u^j}{\partial \alpha} \right]^{1/2} d\alpha \quad (14.253)$$

9.1. REPÈRE NATUREL EN COORDONNÉES SPHÉRIQUES

Déterminons la base naturelle de l'espace vectoriel E_3 associé à l'espace ponctuel $\mathcal{E}_3$ de la géométrie ordinaire, en coordonnées sphériques. Ecrivons l'expression des vecteurs $\overrightarrow{OM}$ dans un repère cartésien fixe $\vec{e}_i^0$ qui est par définition (voir le chapitre de Calcul Vectoriel pour plus de détails):

$$\overrightarrow{OM} = x^i \vec{e}_i^0 = r \sin \theta \cos \phi \vec{e}_1^0 + r \sin \theta \sin \phi \vec{e}_2^0 + r \cos \theta \vec{e}_3^0 \quad (14.254)$$

Les vecteurs de la base naturelle étant donnés par:

$$\vec{e}_k = \partial_k \vec{M} \Rightarrow \begin{cases} \vec{e}_1 = \partial_1 \vec{M} = \partial_r \vec{M} \\ \vec{e}_2 = \partial_2 \vec{M} = \partial_\theta \vec{M} \\ \vec{e}_3 = \partial_3 \vec{M} = \partial_\phi \vec{M} \end{cases} \quad (14.255)$$

Nous avons ainsi:

$$\vec{e}_1 = \partial_1 \vec{M} = \sin \theta \cos \phi \vec{e}_1^0 + \sin \theta \sin \phi \vec{e}_2^0 + \cos \theta \vec{e}_3^0 \quad (14.256)$$

La dérivée de $\vec{OM}$ par rapport à θ donne le vecteur $\vec{e}_2$:

$$\vec{e}_2 = \partial_2 \vec{M} = r \cos \theta \cos \phi \vec{e}_1^0 + r \cos \theta \sin \phi \vec{e}_2^0 - r \sin \theta \vec{e}_3^0 \quad (14.257)$$

La dérivée par rapport à ϕ donne le vecteur $\vec{e}_3$:

$$\vec{e}_3 = \partial_3 \vec{M} = -r \sin \theta \sin \phi \vec{e}_1^0 + r \sin \theta \cos \phi \vec{e}_2^0 \quad (14.258)$$

Ces trois vecteurs sont orthogonaux entre eux ainsi que nous le vérifions aisément en effectuant les produits scalaires $\vec{e}_i \circ \vec{e}_j$. Lorsqu'il en est ainsi, nous disons que les coordonnées sont des "**coordonnées curvilignes orthogonales**" (cf. chapitre de Géométrie Différentielle).

Nous retrouvons donc bien le même résultat que dans le chapitre de Calcul Vectoriel.

Ces vecteurs ne sont cependant pas tous normés, puisque nous avons:

$$g_{11} = \vec{e}_1 \circ \vec{e}_1 = 1 \quad g_{22} = \vec{e}_2 \circ \vec{e}_2 = r^2 \quad g_{33} = \vec{e}_3 \circ \vec{e}_3 = r^2 \sin^2 \theta \quad (14.259)$$

Le repère naturel, en coordonnées sphériques, est donc formé par des vecteurs variables en direction et en module en chaque point de M . Les quantités g_{ij} constituent un exemple de tenseur métrique attaché à chacun des points M de l'espace $\mathcal{E}_3$.

L'élément linéaire du plan est donné par (les détails des calculs peuvent être trouvés dans le chapitre de Relativité Générale):

$$ds^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (14.260)$$

9.2. REPÈRE NATUREL EN COORDONNÉES POLAIRES

Déterminons la base naturelle de l'espace vectoriel E_2 associé à l'espace ponctuel $\mathcal{E}_2$ de la géométrie ordinaire, en coordonnées polaires. Ecrivons l'expression des vecteurs $\vec{OM}$ dans un repère fixe cartésien $\vec{e}_i^0$ qui est par définition (voir le chapitre de Calcul Vectoriel pour plus de détails):

$$\vec{OM} = x^i \vec{e}_i^0 = r \cos \phi \vec{e}_1^0 + r \sin \phi \vec{e}_2^0 \quad (14.261)$$

Les vecteurs de la base naturelle étant donnés par:

$$\vec{e}_k = \partial_k \vec{M} \Rightarrow \begin{cases} \vec{e}_1 = \partial_1 \vec{M} = \partial_r \vec{M} \\ \vec{e}_2 = \partial_2 \vec{M} = \partial_\phi \vec{M} \end{cases} \quad (14.262)$$

Nous avons:

$$\vec{e}_1 = \partial_1 \vec{M} = \cos \phi \vec{e}_1^0 + \sin \phi \vec{e}_2^0 \quad (14.263)$$

La dérivée de $\vec{OM}$ par rapport à ϕ donne le vecteur $\vec{e}_2$:

$$\vec{e}_2 = \partial_2 \vec{M} = -r \sin \phi \vec{e}_1^0 + r \cos \phi \vec{e}_2^0 \quad (14.264)$$

Ces deux vecteurs sont orthogonaux entre eux ainsi que nous le vérifions aisément en effectuant les produits scalaires $\vec{e}_i \circ \vec{e}_j$. Nous retrouvons donc bien le même résultat que dans le chapitre de Calcul Vectoriel.

Nous avons:

$$g_{11} = \cos^2 \phi + \sin^2 \phi \quad g_{21} = g_{12} = 0 \quad g_{22} = r^2 \quad (14.265)$$

L'élément linéaire du plan est alors donné par (cf. chapitre de Relativité Générale):

$$\boxed{ds^2 = dr^2 + r^2 d\phi^2} \quad (14.266)$$

9.3. REPÈRE NATUREL EN COORDONNÉES CYLINDRIQUES

Déterminons la base naturelle de l'espace vectoriel E_3 associé à l'espace ponctuel E_3 de la géométrie ordinaire, en coordonnées cylindriques. Écrivons l'expression des vecteurs $\vec{OM}$ dans un repère fixe cartésien $\vec{e}_i^0$ qui est par définition (voir le chapitre de Calcul Vectoriel pour plus de détails):

$$\vec{OM} = x^i \vec{e}_i^0 = r \cos \phi \vec{e}_1^0 + r \sin \phi \vec{e}_2^0 + z \vec{e}_3^0 \quad (14.267)$$

Les vecteurs de la base naturelle étant donnés par:

$$\vec{e}_k = \partial_k \vec{M} \Rightarrow \begin{cases} \vec{e}_1 = \partial_1 \vec{M} = \partial_r \vec{M} \\ \vec{e}_2 = \partial_2 \vec{M} = \partial_\phi \vec{M} \\ \vec{e}_3 = \partial_3 \vec{M} = \partial_z \vec{M} \end{cases} \quad (14.268)$$

Nous avons:

$$\vec{e}_1 = \partial_1 \vec{M} = \cos \phi \vec{e}_1^0 + \sin \phi \vec{e}_2^0 \quad (14.269)$$

La dérivée de $\vec{OM}$ par rapport à ϕ donne le vecteur $\vec{e}_2$:

$$\vec{e}_2 = \partial_2 \vec{M} = -r \sin \phi \vec{e}_1^0 + r \cos \phi \vec{e}_2^0 \quad (14.270)$$

et enfin:

$$\vec{e}_3 = \partial_3 \vec{M} = \vec{e}_3^0 \quad (14.271)$$

Ces trois vecteurs sont orthogonaux entre eux ainsi qu'on le vérifie aisément en effectuant les produits scalaires $\vec{e}_i \circ \vec{e}_j$. Nous retrouvons donc encore une fois le même résultat que dans le chapitre de Calcul Vectoriel.

Nous avons:

$$g_{11} = 1 \quad g_{22} = r^2 \quad g_{33} = 1 \quad (14.272)$$

L'élément linéaire du plan est alors donné par (cf. chapitre de Relativité Restreinte):

$$ds^2 = dr^2 + r^2 d\varphi^2 + dz^2 \quad (14.273)$$

10. SYMBOLES DE CHRISTOFFEL

L'étude des champs de tenseurs constitue, pour le physicien, l'essentiel de l'analyse tensorielle. Le tenseur générique $\vec{U}$ de ce champ est une fonction du point M et nous le notons:

$$\vec{U}(M) \quad (14.274)$$

Si le tenseur $\vec{U}$ est une fonction seulement de M , le champ considéré est appelé un "**champ fixe**". Si $\vec{U}$ est, en outre, une fonction d'un ou plusieurs paramètres α autres que les coordonnées de M , nous disons alors que ce champ est variable et nous le notons:

$$\vec{U}(M, \alpha) \quad (14.275)$$

Les différentes opérations algébriques sur les tenseurs $\vec{U}(M)$ associés à un même point M ne soulèvent pas de difficulté particulière. La dérivée de $\vec{U}(M)$ par rapport à un paramètre α conduit à utiliser les résultats classiques relatifs à la dérivation des vecteurs.

Cependant, une difficulté apparaît lorsque nous cherchons à calculer la dérivée d'un tenseur $\vec{U}(M)$ par rapport aux coordonnées curvilignes. En effet, les composantes du tenseur sont définies en chaque point M par rapport à un repère naturel qui varie d'un point à un autre.

Par suite, le calcul de la variation élémentaire, appelée "**transport élémentaire**":

$$\vec{U}(M') - \vec{U}(M) \quad (14.276)$$

lorsque nous passons d'un point M à un point infiniment voisin M' ne peut se faire que si nous avons recours à une même base. Pour pouvoir comparer l'un à l'autre les tenseurs $\vec{U}(M')$ et $\vec{U}(M)$, nous sommes amenés à étudier comment varie un repère naturel, pour un système de coordonnées donné, lorsque nous passons d'un point M au point infiniment voisin M' .

Pour un système de coordonnées curvilignes x^i donné d'un espace ponctuel $\mathcal{E}_x$ un problème fondamental de l'analyse tensorielle consiste donc à déterminer, par rapport au repère naturel $(M, \vec{e}_k)$ au point M , le repère naturel $(M', \vec{e}'_k)$ au point infiniment voisin M' . Nous disons alors que nous recherchons une "**connexion affine**".

D'une part, le point M' sera parfaitement défini par rapport à M si nous déterminons le vecteur $d\vec{M}$ tel que $\overrightarrow{MM'} = d\vec{M}$. Pour des coordonnées curvilignes u^k , la décomposition d'un vecteur élémentaire $d\vec{M}$ est donnée par la relation que nous avons démontrée précédemment:

$$d\vec{M} = \partial_k \vec{M} du^k = \vec{e}_k du^k \quad (14.277)$$

les quantités du^k étant les composantes contravariantes du vecteur $d\vec{M}$ sur la base naturelle $(\vec{e}_k)$.

D'autre part, les vecteurs $\vec{e}'_k$ vont pouvoir être déterminés en calculant les variations élémentaires $d\vec{e}_k$ des vecteurs $\vec{e}_k$, par rapport au repère naturel $(M, \vec{e}_k)$, lorsque nous passons de M en M' ; nous avons alors:

$$d\vec{e}'_k = \vec{e}_k + d\vec{e}_k \quad (14.278)$$

Le calcul des vecteurs $d\vec{e}_k$ reste alors le problème essentiel à résoudre. Nous allons tout d'abord étudier un exemple de ce type de calcul en coordonnées sphériques.

Pour cela, reprenons l'expression des vecteurs $\vec{e}_k$ de la base naturelle en coordonnées sphériques, soit:

$$\begin{aligned} \vec{e}_1 &= \partial_1 \vec{M} = \sin \theta \cos \phi \vec{i} + \sin \theta \sin \phi \vec{j} + \cos \theta \vec{k} \\ \vec{e}_2 &= \partial_2 \vec{M} = r \cos \theta \cos \phi \vec{i} + r \cos \theta \sin \phi \vec{j} - r \sin \theta \vec{k} \\ \vec{e}_3 &= \partial_3 \vec{M} = -r \sin \theta \sin \phi \vec{i} + r \sin \theta \cos \phi \vec{j} \end{aligned} \quad (14.279)$$

Les vecteurs de base $\vec{i}, \vec{j}, \vec{k}$ du repère fixe cartésien étant constants en module et en direction, la différentielle du vecteur $\vec{e}_1$ s'écrit:

$$d\vec{e}_1 = (\cos \theta \cos \phi \vec{i} + \cos \theta \sin \phi \vec{j} - \sin \theta \vec{k}) d\theta + (-\sin \theta \sin \phi + \sin \theta \cos \phi \vec{j}) d\phi \quad (14.280)$$

Nous remarquons que les termes entre parenthèses représentent respectivement les vecteurs $\vec{e}_2/r$ et $\vec{e}_3/r$, d'où:

$$d\vec{e}_1 = \left(\frac{d\theta}{r} \right) \vec{e}_2 + \left(\frac{d\phi}{r} \right) \vec{e}_3 \quad (14.281)$$

Nous calculons de même, en différentiant les vecteurs $\vec{e}_2$:

$$\begin{aligned} d\vec{e}_2 &= \\ &dr \cos \theta \cos \phi \vec{i} + dr \cos \theta \sin \phi \vec{j} - dr \sin \theta \vec{k} \\ &- r \sin \theta d\theta \cos \phi \vec{i} - r \sin \theta d\theta \sin \phi \vec{j} - r \cos \theta d\theta \vec{k} \\ &- r \cos \theta \sin \phi d\phi \vec{i} + r \cos \theta \cos \phi d\phi \vec{j} \\ &= (\cos \theta \cos \phi \vec{i} + \cos \theta \sin \phi \vec{j} - \sin \theta \vec{k}) dr + (-r \sin \theta \cos \phi \vec{i} - r \sin \theta \sin \phi \vec{j} - r \cos \theta \vec{k}) d\theta \\ &+ (-r \cos \theta \sin \phi \vec{i} + r \cos \theta \cos \phi \vec{j}) d\phi \\ &= \frac{\vec{e}_2}{r} dr - r \vec{e}_1 d\theta + (-r \cos \theta \sin \phi \vec{i} + r \cos \theta \cos \phi \vec{j}) d\phi \end{aligned} \quad (14.282)$$

Avec:

$$\cot \theta = \frac{\cos \theta}{\sin \theta} \quad (14.283)$$

nous avons:

$$\vec{e}_3 \cot \theta = \vec{e}_3 \frac{\cos \theta}{\sin \theta} = -r \cos \theta \sin \phi \vec{e}_1 + r \cos \theta \cos \phi \vec{e}_2 \quad (14.284)$$

Donc finalement:

$$d\vec{e}_2 = \frac{\vec{e}_2}{r} dr - r\vec{e}_1 d\theta + \vec{e}_3 \cot \theta d\phi \quad (14.285)$$

Et:

$$\begin{aligned} d\vec{e}_3 &= -dr \sin \theta \sin \phi \vec{e}_1 + dr \sin \theta \cos \phi \vec{e}_2 \\ &\quad -r \cos \theta d\theta \sin \phi \vec{e}_1 + r \cos \theta d\theta \cos \phi \vec{e}_2 \\ &\quad -r \sin \theta \cos \phi d\phi \vec{e}_1 - r \sin \theta \sin \phi d\phi \vec{e}_2 \end{aligned} \quad (14.286)$$

Après quelques opérations algébriques élémentaires et très pertinentes (...), nous arrivons à:

$$d\vec{e}_3 = (-r \sin^2 \theta d\phi) \vec{e}_1 + (-\sin \theta \cos \theta d\phi) \vec{e}_2 + \left(\frac{dr}{r} + \cot \theta d\theta \right) \vec{e}_3 \quad (14.287)$$

Les différentielles $d\vec{e}_k$ sont ainsi décomposées sur la base naturelle $(\vec{e}_k)$. Si nous notons ω_i^k , les composantes contravariantes du vecteur $d\vec{e}_i$, celui-ci s'écrit:

$$\boxed{d\vec{e}_i = \omega_i^k \vec{e}_k} \quad (14.288)$$

Les composantes ω_i^k du vecteur $d\vec{e}_i$ sont des formes différentielles (combinaisons linéaires de différentielles). Nous avons, par exemple:

$$\omega_1^2 = \frac{d\theta}{r} \quad \omega_3^3 = \frac{dr}{r} + \cot \theta d\theta \quad (14.289)$$

Si nous notons de manière générale u^i les coordonnées sphériques, nous avons:

$$u^1 = r \quad u^2 = \theta \quad u^3 = \phi \quad (14.290)$$

Les différentielles des coordonnées sont alors notées:

$$du^1 = dr \quad du^2 = d\theta \quad du^3 = d\phi \quad (14.291)$$

et les composantes ω_i^j s'écrivent alors de manière générale:

$$\boxed{\omega_i^j = \Gamma_{ik}^j du^k} \quad (14.292)$$

où les quantités Γ_{ki}^j sont des fonctions de r, θ, ϕ qui vont être explicitement obtenues en identifiant chaque composante ω_i^j . Par exemple, la composante ω_3^3 s'écrit avec la notation de la relation précédente:

$$\omega_3^3 = \frac{dr}{r} + \cot \theta d\theta = \Gamma_{13}^3 du^1 + \Gamma_{23}^3 du^2 + \Gamma_{33}^3 du^3 \quad (14.293)$$

Identifiant les coefficients des différentielles, il vient:

$$\Gamma_{13}^3 = \frac{1}{r} \quad \Gamma_{23}^3 = \cot \theta \quad \Gamma_{33}^3 = 0 \quad (14.294)$$

En procédant de même avec les neuf composantes ω_i^j , nous obtenons les vingt-sept (...) termes Γ_{ki}^j dont les calculs détaillés pour les 27 sont donnés beaucoup plus bas dans le texte. Pour un système de coordonnées curvilignes quelconques, ces quantités Γ_{ki}^j sont appelées les "**symboles de Christoffel de deuxième espèce**" ou encore "**fonctions euclidiennes de connexion affine**".

Ainsi, pour un espace ponctuel $\mathcal{E}_3$ et un système de coordonnées curvilignes u^j quelconque, la différentielle $d\vec{e}_i = \omega_i^k \vec{e}_k$ des vecteurs $\vec{e}_i$ de la base naturelle s'écrit sur cette base:

$$d\vec{e}_i = \omega_i^k \vec{e}_k = \Gamma_{ji}^k du^j \vec{e}_k \quad (14.295)$$

Nous venons de voir, sur l'exemple des coordonnées sphériques, qu'un calcul direct permet, par identification, d'obtenir explicitement les quantités Γ_{ki}^j . Nous allons voir que nous pouvons également obtenir l'expression de ces quantités en fonction des composantes g_{ij} .

Le calcul des quantités Γ_{ki}^j en fonction des g_{ij} va nous amener à introduire d'autres symboles de Christoffel. Pour cela, écrivons les composantes covariantes, notées ω_{ji} , des différentielles $d\vec{e}_i$, soit:

$$\omega_{ji} = \vec{e}_j \circ d\vec{e}_i \quad (14.296)$$

Les composantes covariantes sont également des combinaisons linéaires des différentielles du^k que nous pouvons écrire sous la forme:

$$\omega_{ji} = \Gamma_{ji}^k du^k \quad (14.297)$$

Les quantités Γ_{ji}^k sont appelées les "**symboles de Christoffel de première espèce**".

Nous voyons très bien en parcourant à nouveau les définitions des symboles de Christoffel que:

1. Pour ce qui est des symboles de 2ème espèce, ils sont symétriques par rapport à leurs indices inférieurs :

$$\Gamma_{ji}^k = \Gamma_{ij}^k \quad (14.298)$$

2. Pour ce qui est des symboles de 1ère espèce : symétriques par rapport à leurs indices extrêmes

$$\Gamma_{kij} = \Gamma_{jik} \quad (14.299)$$

Effectivement (suite à la demande d'un lecteur), puisque nous avons:

$$d\vec{e}_i = \omega_i^k \vec{e}_k = \Gamma_{ji}^k du^j \vec{e}_k \quad (14.300)$$

Il vient alors:

$$\frac{\partial \vec{e}_i}{\partial u^j} = \vec{e}_k \Gamma_{ji}^k \quad (14.301)$$

et en permutant les indices:

$$\frac{\partial \vec{e}_j}{\partial u^i} = \vec{e}_k \Gamma_{ji}^k \quad (14.302)$$

L'identification terme à terme du développement sur un cas concret des deux dernières relations donnera (forcément) l'égalité:

$$\Gamma_{ij}^k = \Gamma_{ji}^k \quad (14.303)$$

que nous voulions prouver.

Puisque les composantes covariantes sont liées aux composantes contravariantes par les relations (contraction des indices):

$$\omega_{ji} = g_{jl} \omega_i^l = g_{jl} \Gamma_{ki}^l du^k \quad (14.304)$$

nous obtenons l'expression liant les symboles de Christoffel de chaque espèce:

$$\Gamma_{ji}^k = g_{jl} \Gamma_{ki}^l \quad (14.305)$$

Inversement:

$$\Gamma_{ki}^j = g^{jl} \Gamma_{ki}^l \quad (14.306)$$

Remarque: Diverses notations sont utilisées pour représenter les symboles de Christoffel. Les plus usuelles sont les suivantes:

- Symboles de première espèce:

$$\Gamma_{kij} = [ki, j] \quad (14.307)$$

- Symboles de deuxième espèce:

$$\Gamma_{ki}^j = [k^j j] \quad (14.308)$$

Considérons maintenant un espace ponctuel $\mathcal{E}_x$ et soit un élément linéaire ds^2 donné de cet espace:

$$ds^2 = g_{ij} du^i du^j \quad (14.309)$$

Partant de:

$$g_{ij} = \vec{e}_i \circ \vec{e}_j \quad (14.310)$$

nous obtenons par différenciation:

$$dg_{ij} = \vec{e}_i \circ d\vec{e}_j + \vec{e}_j \circ d\vec{e}_i \quad (14.311)$$

En y injectant l'expression des différentielles $d\vec{e}_j = \omega_j^i \vec{e}_i$ cela nous donne:

$$dg_{ij} = (\vec{e}_i \circ \vec{e}_l) \omega_j^l + (\vec{e}_j \circ \vec{e}_l) \omega_i^l = g_{il} \omega_j^l + g_{jl} \omega_i^l \quad (14.312)$$

où le terme ω_j^l représente la composante contravariante du vecteur $d\vec{e}_j$. On peut rendre cette composante covariante en la prémultipliant par le tenseur métrique g_{il} de manière à former ω_{ij} quantité que l'on pourra à son tour exprimer au moyen des symboles de Christoffel comme suit:

$$\omega_{ij} = g_{il} \omega_j^l = g_{il} \Gamma_{kj}^l du^k \quad (14.313)$$

substituant la relation $\Gamma_{kj}^l = g_{il} \Gamma_{kj}^i$ dans l'expression précédente (les indices utilisés dans cette relation ne sont pas ceux de l'expression en cause, mais à proprement parler cela revient au même), nous obtenons alors:

$$\omega_{ij} = \Gamma_{kij} du^k \quad (14.314)$$

La différentielle dg_{ij} s'écrit alors:

$$dg_{ij} = \omega_{ij} + \omega_{ji} = (\Gamma_{kij} + \Gamma_{kji}) du^k \quad (14.315)$$

D'autre part, la différentielle de la fonction g_{ij} s'écrit également:

$$dg_{ij} = \frac{\partial g_{ij}}{\partial u^k} du^k = \partial_k g_{ij} du^k \quad (14.316)$$

d'où en identifiant les coefficients des différentielles du^k dans ces deux dernières expressions (beaucoup plus bas dans le présent chapitre, il y a un exemple détaillé de toutes les relations qui vont suivre avec plusieurs systèmes de coordonnées):

$$\Gamma_{kij} + \Gamma_{kji} = \partial_k g_{ij} \quad (14.317)$$

Relation que le lecteur pourra (s'il doute) vérifier avec les exemples pratiques détaillés qui se trouvent bien plus bas.

Comme nous avons:

$$\Gamma_{kij} = \Gamma_{jik} \quad (14.318)$$

où il est fortement recommandé au lecteur de se rappeler pour la suite que la permutation des indices respectant cette dernière relation ne fonctionne, en général, que sur les indices extrêmes.

Nous pouvons donc écrire l'avant-dernière relation:

$$\Gamma_{jik} + \Gamma_{kji} = \partial_k g_{ij} \quad (14.319)$$

puis en effectuant une permutation circulaire sur les indices (donc il ne s'agit pas d'une permutation des indices extrêmes!), nous obtenons:

$$\begin{aligned} \Gamma_{kji} + \Gamma_{ijl} &= \partial_i g_{jk} \\ \Gamma_{ikj} + \Gamma_{jil} &= \partial_j g_{ki} \end{aligned} \quad (14.320)$$

En effectuant la somme:

$$\begin{aligned} \Gamma_{jik} + \Gamma_{kji} &= \partial_k g_{ij} \\ \Gamma_{kji} + \Gamma_{ijl} &= \partial_i g_{jk} \\ + \frac{\Gamma_{kji} + \Gamma_{ijl} + \Gamma_{jil} + \Gamma_{ikj}}{\Gamma_{jik} + \Gamma_{kji} + \Gamma_{kji} + \Gamma_{ijl}} &= \partial_k g_{ij} + \partial_i g_{jk} \end{aligned} \quad (14.321)$$

et en retranchant:

$$\begin{aligned} \Gamma_{jik} + \Gamma_{kji} + \Gamma_{kji} + \Gamma_{ijl} &= \partial_k g_{ij} + \partial_i g_{jk} \\ \Gamma_{ikj} + \Gamma_{jil} &= \partial_j g_{ki} \\ - \frac{\Gamma_{ikj} + \Gamma_{jil}}{\Gamma_{jik} + \Gamma_{kji} + \Gamma_{kji} + \Gamma_{ijl} - \Gamma_{ikj} - \Gamma_{jil}} &= \partial_k g_{ij} + \partial_i g_{jk} - \partial_j g_{ki} \end{aligned} \quad (14.322)$$

En simplifiant il vient:

$$\begin{aligned} \Gamma_{kji} + \Gamma_{kji} &= \partial_k g_{ij} + \partial_i g_{jk} - \partial_j g_{ki} \\ 2\Gamma_{kji} &= \partial_k g_{ij} + \partial_i g_{jk} - \partial_j g_{ki} \end{aligned} \quad (14.323)$$

d'où:

$$\Gamma_{kji} = \frac{1}{2} (\partial_k g_{ij} + \partial_i g_{jk} - \partial_j g_{ki}) \quad (14.324)$$

C'est l'expression des symboles de Christoffel de première espèce en fonction des dérivées partielles des composantes g_{ij} du tenseur fondamental.

Nous obtenons ceux de deuxième espèce à partir de la relation (par définition) suivante souvent appelée "**théorème fondamental de la géométrie riemannienne**":

$$\Gamma_{ij}^i = g^{ij} \Gamma_{kij} = \frac{1}{2} g^{ij} (\partial_k g_{ij} + \partial_j g_{ki} - \partial_i g_{kj}) \quad (14.325)$$

Les deux dernières expressions encadrées ci-dessus permettent le calcul effectif des symboles de Christoffel pour une métrique donnée (d'où un énorme gain en calculs). Lorsque les quantités g_{ij} sont données a priori, nous pouvons ainsi étudier les propriétés de l'espace ponctuel défini par la donnée de cette métrique, ce qui est le cas des espaces de Riemann que nous verrons plus loin.

Exemple:

Proposons-nous de calculer les Γ_{ij}^k correspondant au système de coordonnées polaires (ce sera déjà suffisamment long...) dans le plan que nous noterons cette fois-ci (contrairement au chapitre de Calcul Vectoriel) en notation indicielle:

$$\begin{aligned} x^1 &= y^1 \cos(y^2) & y^1 &= \sqrt{(x^1)^2 + (x^2)^2} \\ x^2 &= y^1 \sin(y^2) & y^2 &= \arctan\left(\frac{x^2}{x^1}\right) \end{aligned} \quad \text{avec} \quad (14.326)$$

Nous allons calculer les symboles de Christoffel à partir de notre dernière relation:

$$\Gamma_{ij}^k = g^{kl} \Gamma_{lij} = \frac{1}{2} g^{kl} (\partial_j g_{li} + \partial_i g_{lj} - \partial_l g_{ij}) \quad (14.327)$$

Occupons-nous de déterminer les composantes de la métrique. Au fait, elles sont les mêmes que celles que nous avons calculées pour les coordonnées cylindriques plus haut à la différence normalement évidente que g_{33} n'existe pas. Dès lors, nous avons:

$$g_{11} = 1 \quad g_{12} = g_{21} = 0 \quad g_{22} = (y^1)^2 \quad (14.328)$$

Calculons alors les g^{ji} . Dans cet exemple c'est assez trivial, il suffit d'appliquer la relation démontrée au début de ce chapitre:

$$g_{ik} g^{kj} = \delta_i^j \quad (14.329)$$

Nous avons alors immédiatement:

$$g^{11} = 1 \quad g^{12} = g^{21} = 0 \quad g^{22} = \frac{1}{(y^1)^2} \quad (14.330)$$

Maintenant développons l'écriture de symboles de Christoffel pour ces coordonnées:

$$\begin{aligned}
\Gamma_{11}^1 &= \frac{1}{2} g^{1l} \left(\frac{\partial g_{1l}}{\partial y^1} + \frac{\partial g_{1l}}{\partial y^1} - \frac{\partial g_{11}}{\partial y^l} \right) \quad l = 1, 2 \\
&= \frac{1}{2} g^{11} \left(\frac{\partial g_{11}}{\partial y^1} + \frac{\partial g_{11}}{\partial y^1} - \frac{\partial g_{11}}{\partial y^1} \right) + \frac{1}{2} g^{12} \left(\frac{\partial g_{12}}{\partial y^1} + \frac{\partial g_{21}}{\partial y^1} - \frac{\partial g_{11}}{\partial y^2} \right) = 0 \\
\Gamma_{11}^2 &= \frac{1}{2} g^{2l} \left(\frac{\partial g_{1l}}{\partial y^1} + \frac{\partial g_{1l}}{\partial y^1} - \frac{\partial g_{11}}{\partial y^l} \right) \quad l = 1, 2 \\
&= \frac{1}{2} g^{21} \left(\frac{\partial g_{11}}{\partial y^1} + \frac{\partial g_{11}}{\partial y^1} - \frac{\partial g_{11}}{\partial y^1} \right) + \frac{1}{2} g^{22} \left(\frac{\partial g_{12}}{\partial y^1} + \frac{\partial g_{21}}{\partial y^1} - \frac{\partial g_{11}}{\partial y^2} \right) = 0 \\
\Gamma_{12}^1 &= \frac{1}{2} g^{1l} \left(\frac{\partial g_{1l}}{\partial y^2} + \frac{\partial g_{l2}}{\partial y^1} - \frac{\partial g_{21}}{\partial y^l} \right) \quad l = 1, 2 \tag{14.331} \\
&= \frac{1}{2} g^{11} \left(\frac{\partial g_{11}}{\partial y^2} + \frac{\partial g_{12}}{\partial y^1} - \frac{\partial g_{21}}{\partial y^1} \right) + \frac{1}{2} g^{12} \left(\frac{\partial g_{12}}{\partial y^2} + \frac{\partial g_{22}}{\partial y^1} - \frac{\partial g_{21}}{\partial y^2} \right) = 0 \\
\Gamma_{12}^2 &= \frac{1}{2} g^{2l} \left(\frac{\partial g_{1l}}{\partial y^2} + \frac{\partial g_{l2}}{\partial y^1} - \frac{\partial g_{21}}{\partial y^l} \right) \quad l = 1, 2 \\
&= \frac{1}{2} g^{21} \left(\frac{\partial g_{11}}{\partial y^2} + \frac{\partial g_{12}}{\partial y^1} - \frac{\partial g_{21}}{\partial y^1} \right) + \frac{1}{2} g^{22} \left(\frac{\partial g_{12}}{\partial y^2} + \frac{\partial g_{22}}{\partial y^1} - \frac{\partial g_{21}}{\partial y^2} \right) \\
&= \frac{1}{2} g^{22} \frac{\partial g_{22}}{\partial y^1} = \frac{1}{2} \frac{1}{(y^1)^2} 2y^1 = \frac{1}{y^1}
\end{aligned}$$

d'où en raison des propriétés de symétrie:

$$\begin{aligned}
\Gamma_{21}^1 &= \Gamma_{12}^1 = 0 \\
\Gamma_{21}^2 &= \Gamma_{12}^2 = \frac{1}{y^1} \tag{14.332}
\end{aligned}$$

De même:

$$\begin{aligned}
\Gamma_{22}^1 &= \frac{1}{2} g^{1l} \left(\frac{\partial g_{2l}}{\partial y^2} + \frac{\partial g_{l2}}{\partial y^2} - \frac{\partial g_{22}}{\partial y^l} \right) \quad l = 1, 2 \\
&= \frac{1}{2} g^{11} \left(\frac{\partial g_{21}}{\partial y^2} + \frac{\partial g_{12}}{\partial y^2} - \frac{\partial g_{22}}{\partial y^1} \right) + \frac{1}{2} g^{12} \left(\frac{\partial g_{22}}{\partial y^2} + \frac{\partial g_{22}}{\partial y^2} - \frac{\partial g_{22}}{\partial y^2} \right) \\
&= -\frac{1}{2} g^{11} \frac{\partial g_{22}}{\partial y^1} = -\frac{1}{2} \cdot 1 \cdot 2y^1 = -y^1 \\
\Gamma_{22}^2 &= \frac{1}{2} g^{2l} \left(\frac{\partial g_{2l}}{\partial y^2} + \frac{\partial g_{l2}}{\partial y^2} - \frac{\partial g_{22}}{\partial y^l} \right) \quad l = 1, 2 \\
&= \frac{1}{2} g^{21} \left(\frac{\partial g_{21}}{\partial y^2} + \frac{\partial g_{12}}{\partial y^2} - \frac{\partial g_{22}}{\partial y^1} \right) + \frac{1}{2} g^{22} \left(\frac{\partial g_{22}}{\partial y^2} + \frac{\partial g_{22}}{\partial y^2} - \frac{\partial g_{22}}{\partial y^2} \right) \\
&= \frac{1}{2} g^{22} \frac{\partial g_{22}}{\partial y^2} = 0
\end{aligned} \tag{14.333}$$

En résumé:

$$\begin{aligned}
\Gamma_{11}^1 &= 0 \quad \Gamma_{12}^1 = 0 \quad \Gamma_{21}^1 = 0 \quad \Gamma_{22}^1 = -y^1 \\
\Gamma_{11}^2 &= 0 \quad \Gamma_{12}^2 = \frac{1}{y^1} \quad \Gamma_{21}^2 = \frac{1}{y^1} \quad \Gamma_{22}^2 = 0
\end{aligned} \tag{14.334}$$

11. THÉORÈME DE RICCI

Avant de lire ce qui va suivre... je tiens à rappeler au lecteur que la rédaction de ce chapitre n'est pas terminée! Ainsi, il me faut encore illustrer les notions abstraites qui vont suivre par des exemples pratiques concrets!

Ceci étant dit, nous avons donc vu dans le chapitre de Relativité Générale que les géodésiques sont les distances les plus courtes entre deux points dans n'importe quel type d'espace. Ce qui va nous intéresser maintenant, c'est d'étudier les variations d'un vecteur au cours d'un tel déplacement. Rappelons d'abord que l'équation des géodésiques pour un système de coordonnées curvilignes quelconque y^i de l'espace ponctuel $\mathcal{E}_*$ (cf. chapitre des Principes) est donnée par (cf. chapitre de Relativité Générale):

$$\frac{d^2 y^i}{ds^2} + \Gamma_{jk}^i \frac{dy^j}{ds} \frac{dy^k}{ds} = 0 \tag{14.335}$$

Considérons maintenant un vecteur $\vec{v}$ de $\mathcal{E}_*$ de composantes covariantes v_i et formons le produit scalaire des vecteurs $\vec{v}$ et $\vec{n} = dy^i / ds$ (ce dernier vecteur, noté directement ici de manière abusive avec les indices, donne les composantes tangentes à la géodésique sur laquelle circule le premier vecteur), nous avons alors la quantité suivante:

$$\vec{v} \circ \vec{n} = v_i \frac{dy^i}{ds} \tag{14.336}$$

Lors d'un déplacement le long de la géodésique, d'un point M à un point infiniment voisin M' , le scalaire subit la variation:

$$d \left(v_i \frac{dy^i}{ds} \right) = dv_k \frac{dy^k}{ds} + v_i d \left(\frac{dy^i}{ds} \right) \quad (14.337)$$

et comme:

$$\frac{d \left(\frac{dy^i}{ds} \right)}{ds} = \frac{d^2 y^i}{ds^2} \Rightarrow d \left(\frac{dy^i}{ds} \right) = \frac{d^2 y^i}{ds^2} ds \quad (14.338)$$

d'où:

$$d \left(v_i \frac{dy^i}{ds} \right) = dv_k \frac{dy^k}{ds} + v_i \frac{d^2 y^i}{ds^2} ds \quad (14.339)$$

Remplaçons dans cette dernière expression, d'une part la différentielle de dv_k par sa différentielle totale exacte:

$$dv_k = \frac{\partial v_k}{\partial y^j} dy^j = \partial_j v_k \frac{ds}{ds} dy^j = \partial_j v_k \frac{dy^j}{ds} ds \quad (14.340)$$

et d'autre part, la dérivée seconde $d^2 y^i / ds^2$ par son expression tirée de l'équation des géodésiques. Nous obtenons:

$$d \left(v_i \frac{dy^i}{ds} \right) = \partial_j v_k \frac{dy^j}{ds} \frac{dy^k}{ds} ds - v_i \Gamma_{kj}^i \frac{dy^j}{ds} \frac{dy^k}{ds} ds = \left(\partial_j v_k - v_i \Gamma_{kj}^i \right) \frac{dy^j}{ds} \frac{dy^k}{ds} ds \quad (14.341)$$

qui peut encore s'écrire:

$$d (\vec{v} \circ \vec{n}) = \left(\partial_j v_k - v_i \Gamma_{kj}^i \right) \frac{dy^j}{ds} \frac{dy^k}{ds} = Dv_k \frac{dy^k}{ds} \quad (14.342)$$

où nous avons posé:

$$Dv_k = \left(\partial_j v_k - v_i \Gamma_{kj}^i \right) \frac{dy^j}{ds} \quad (14.343)$$

qui sont par définition les différentielles absolues des composantes covariantes du vecteur $\vec{v}$. Nous définissons également la "**dérivée covariante**" (appelée également "**connexion**") par la relation:

$$\boxed{\nabla_j v_k = \partial_j v_k - v_i \Gamma_{kj}^i} \quad (14.344)$$

Remarque: Dans les ouvrages anciens ou américains ceci est souvent noté sous la forme (que nous n'utiliserons aucunement sur ce site):

$$v_{k;j} = v_{k,j} - v_i \Gamma_{kj}^i \quad (14.345)$$

faisant donc usage du ";" pour noter la dérivée covariante et de la "," pour différentielle partielle.

Puisque la dérivée du produit de deux fonctions est la somme des dérivées partielles, nous avons alors aussi:

$$\begin{aligned} \nabla_a (t_b r_c) &= r_c \nabla_a t_b + t_b \nabla_a r_c = r_c \left(\partial_a t_b - t_e \Gamma_{ab}^e \right) + t_b \left(\partial_a r_c - r_e \Gamma_{ac}^e \right) \\ &= r_c \partial_a t_b - r_c t_e \Gamma_{ab}^e + t_b \partial_a r_c - t_b r_e \Gamma_{ac}^e = r_c \partial_a t_b + t_b \partial_a r_c - r_c t_e \Gamma_{ab}^e - t_b r_e \Gamma_{ac}^e \\ &= \partial_a (t_b r_c) - r_c t_e \Gamma_{ab}^e - t_b r_e \Gamma_{ac}^e \end{aligned} \quad (14.346)$$

Si nous posons $t_b r_c \equiv \nabla_j v_i$ alors nous avons (résultat que nous utiliserons après avoir démontré le théorème de Ricci pour déterminer plus loin le tenseur d'Einstein nécessaire dans le chapitre de Relativité Générale):

$$\nabla_k (\nabla_j v_i) = \partial_k (\nabla_j v_i) - (\nabla_j v_r) \Gamma_{ik}^r - (\nabla_r v_i) \Gamma_{jk}^r \quad (14.347)$$

En coordonnées curvilignes, pour que la différentielle d'un vecteur soit un vecteur, il faut que les deux vecteurs dont nous prenons la différence se trouvent en un même point de l'espace. En d'autres termes, il faut transporter, d'une manière ou d'une autre, l'un des deux vecteurs infiniment voisins au point où se trouve le second et, seulement après faire la différence des deux vecteurs qui se trouvent maintenant en un seul et même point de l'espace. L'opération de transport parallèle doit être définie de telle sorte qu'en coordonnées cartésiennes (pour le petit exemple), la différence des composantes coïncide avec la différence ordinaire dv_k .

Ainsi, nous avons bien en coordonnées cartésiennes:

$$Dv_k = \left(\partial_j v_k - v_i \Gamma_{kj}^i \right) dy^j = dv_k \quad (14.348)$$

puisque dans ce système: $\Gamma_{kj}^i = 0$.

Ainsi, en coordonnées curvilignes la différence des composantes des deux vecteurs après le transport de l'un d'entre eux au point où se trouve l'autre est notée δv_k telle que nous ayons:

$$Dv_k = \left(\partial_j v_k - v_i \Gamma_{kj}^i \right) dy^j = dv_k - \delta v_k \quad (14.349)$$

Ceci nous amène à:

$$\delta v_k = v_i \Gamma_{kj}^i dy^j \quad (14.350)$$

Mais aussi à écrire le principe de moindre action (principe variationnel) sous la forme tensorielle:

$$Dv_k = dv_k - \delta v_k = 0 \quad (14.351)$$

Considérons maintenant un tenseur d'ordre deux, produit de deux tenseurs d'ordre un tel que (nous l'avons vu lors de notre étude des compositions de tenseurs):

$$\begin{aligned} \delta(A_i B_j) &= \delta A_{ij} = A_i \delta B_j + B_j \delta A_i = A_i B_k \Gamma_{jk}^k dy^k + B_j A_k \Gamma_{ik}^k dy^k = \\ &= (A_i B_k \Gamma_{jk}^k + B_j A_k \Gamma_{ik}^k) dy^k = \delta A_{ij} = (A_{ik} \Gamma_{jk}^k + B_{jk} \Gamma_{ik}^k) dy^k \end{aligned} \quad (14.352)$$

Donc:

$$\frac{\delta v_{ij}}{dy^k} = \partial_k v_{ij} = v_{ik} \Gamma_{jk}^k + v_{jk} \Gamma_{ik}^k \quad (14.353)$$

d'où (nous sortons les deux dernières égalités juste pour l'esthétique!):

$$\partial_k v_{ij} = v_{ik} \Gamma_{jk}^k + v_{jk} \Gamma_{ik}^k \quad (14.354)$$

Ce qui nous amène à pouvoir écrire la métrique sous sa forme variationnelle appelée "**identité de Ricci**":

$$\boxed{\partial_k g_{ij} = g_{ik} \Gamma_{jk}^k + g_{jk} \Gamma_{ik}^k} \quad (14.355)$$

Mais nous avons aussi puisque $g_{ij} = \vec{e}_i \circ \vec{e}_j$:

$$dg_{ij} = \vec{e}_i \circ d\vec{e}_j + d\vec{e}_i \circ \vec{e}_j = \vec{e}_i \circ \omega_j^k \vec{e}_k + \omega_i^k \vec{e}_k \circ \vec{e}_j = \omega_j^k (\vec{e}_i \circ \vec{e}_k) + \omega_i^k (\vec{e}_k \circ \vec{e}_j) \quad (14.356)$$

d'où l'identité:

$$dg_{ij} = \omega_j^k g_{ik} + \omega_i^k g_{jk} \quad (14.357)$$

Avec les deux relations:

$$dg_{ij} = \omega_j^k g_{ik} + \omega_i^k g_{jk} \text{ et } \delta g_{ij} = (g_{ik} \Gamma_{jk}^k + g_{jk} \Gamma_{ik}^k) dy^k \quad (14.358)$$

et la différentielle absolue (qui se généralise simplement pour un tenseur d'ordre deux):

$$Dg_{ij} = dg_{ij} - \delta g_{ij} \quad (14.359)$$

Nous avons:

$$Dg_{ij} = \omega_j^k g_{ik} + \omega_i^k g_{jk} - g_{ik} \Gamma_{jk}^k dy^k - g_{jk} \Gamma_{ik}^k dy^k \quad (14.360)$$

Or, rappelons que nous avons par définition:

$$\omega_j^k = \Gamma_{jk}^k dy^k \text{ et } \omega_i^k = \Gamma_{ik}^k dy^k \quad (14.361)$$

Donc finalement:

$$Dg_{ij} = 0 \quad (14.362)$$

La différentielle absolue sur une géodésique dans l'approximation d'un transport infinitésimal du tenseur fondamental est donc (comme nous pouvions nous y attendre) nulle. C'est le "**théorème de Ricci**". Certains physiciens théoriciens disent dès lors que "**la dérivée covariante tue la métrique**" dans le sens où la métrique ne change pas sur un différentiel d'espace.

Finalement, nous voyons aussi que pour un tenseur d'ordre deux (la métrique en particulier) nous avons:

$$Dg_{ij} = dg_{ij} - \delta g_{ij} = dg_{ij} - g_{ik}\Gamma_{jk}^k dy^k - g_{jk}\Gamma_{ik}^k dy^k \quad (14.363)$$

Nous pouvons donc écrire la différentielle absolue qui dans ce cas particulier est nulle:

$$Dg_{ij} = (\partial_k g_{ij} - g_{ik}\Gamma_{jk}^k - g_{jk}\Gamma_{ik}^k) dy^k = 0 \quad (14.364)$$

et donc:

$$\boxed{\nabla_k g_{ij} = \partial_k g_{ij} - g_{ik}\Gamma_{jk}^k - g_{jk}\Gamma_{ik}^k = 0} \quad (14.365)$$

Remarque: Il faudra se rappeler lors de la définition du tenseur d'Einstein que:

$$\nabla_k g_{ij} = 0 \text{ et } \nabla_k g_{ij} = \partial_k g_{ij} - g_{ik}\Gamma_{jk}^k - g_{jk}\Gamma_{ik}^k \quad (14.366)$$

et qu'il s'agit d'une autre manière d'exprimer qu'une variation infinitésimale sur une géodésique selon le principe de moindre action tue la métrique. Nous allons donc travailler à partir de maintenant (comme avant déjà) avec des équations différentielles non nécessairement linéaires qu'il faudra intégrer pour trouver le comportement de la matière dans un espace donné.

Déterminons maintenant une expression qui nous sera très utile en relativité générale lorsque nous déterminerons l'équation d'Einstein des champs (une autre manière d'exprimer que la dérivée covariante de la métrique est nulle):

Effectuons la multiplication contractée de:

$$\nabla_k g_{ij} = \partial_k g_{ij} - g_{ik}\Gamma_{jk}^k - g_{jk}\Gamma_{ik}^k = 0 \quad (14.367)$$

par g^{ij} , il vient alors en utilisant la relation $g^{ij}g_{jl} = \delta_l^i$ (que nous avons démontrée beaucoup plus haut) que:

$$g^{ij}\partial_k g_{ij} - g^{ij}g_{ik}\Gamma_{jk}^k - g^{ij}g_{jk}\Gamma_{ik}^k = g^{ij}\partial_k g_{ij} - \delta_j^i\Gamma_{jk}^k - \delta_i^k\Gamma_{ik}^k = 0 \quad (14.368)$$

d'où la relation:

$$g^{ij}\partial_k g_{ij} - \Gamma_{jk}^j - \Gamma_{ik}^i = 0 \quad (14.369)$$

Les quantités Γ_{jk}^j et Γ_{ik}^i représentant les mêmes sommes, nous avons alors:

$$g^{ij}\partial_k g_{ij} = 2\Gamma_{ik}^i \quad (14.370)$$

Considérons maintenant g le déterminant des quantités g_{ij} . La dérivation du déterminant nous donne:

$$\partial_k g = g g^{ij} \partial_k g_{ij} \Rightarrow \partial_k g_{ij} = \frac{\partial_k g}{g g^{ij}} \quad (14.371)$$

Démonstration:

Soit une variable quelconque que nous choisissons ici être le temps t uniquement pour simplifier les notations des calculs qui vont suivre. Lorsque la partie principale du développement sera achevée, le résultat peut être adapté à toute autre variable. Nous noterons pour la suite g_i les éléments de la j -ème colonne de g_{ij} .

Pour les développements qui vont suivre, nous définissons les notations:

$$g = \det(g_1, g_2, \dots, g_n) \quad \dot{g} = d_t g \quad \dot{g}_i = d_t g_i \quad \dot{g}_{ij} = d_t g_{ij} \quad (14.372)$$

La règle de dérivation d'un déterminant fonctionnel est (cf. chapitre d'Algèbre Linéaire):

$$\dot{g} = \det(\dot{g}_1, g_2, \dots, g_n) + \det(g_1, \dot{g}_2, \dots, g_n) + \dots + \det(g_1, g_2, \dots, \dot{g}_n) \quad (14.373)$$

En considérant le premier déterminant, en faisant appel aux mineurs pour le développement de sa première colonne:

$$\alpha_{i1} \dot{g}_{i1} \quad (14.374)$$

Pour le terme j , il vient:

$$\alpha_{ij} \dot{g}_{ij} \quad (14.375)$$

Soit:

$$\dot{g} = \alpha_{ij} \dot{g}_{ij} \text{ ou } g = \alpha_{ij} g_{ij} \quad (14.376)$$

Or, nous avons démontré bien plus haut que le tenseur métrique est son propre inverse. Donc

$$g^{ij} = g_{ij}^{-1} \quad (14.377)$$

Ce qui nous permet d'écrire:

$$g g^{ij} = \alpha_{ij} \quad (14.378)$$

et donc:

$$\dot{g} = g \dot{g}_{ij} g^{ij} \quad (14.379)$$

Ce qui s'écrit également (suite aux conventions définies au début de la démonstration):

$$d_t g = g d_t g_{ij} g^{ij} \quad (14.380)$$

où le lecteur doit donc prendre garde à ne pas mal lire en pensant typiquement que la dérivée dans le terme de droite dérive tout... alors qu'il ne dérive que g_{ij} .

Nous pouvons adopter cependant une autre variable. Soit h cette autre variable:

$$\partial_k g = g g^{ij} \partial_k g_{ij} \quad (14.381)$$

Soit en réarrangeant:

$$\partial_k g_{ij} = \left(g g^{ij} \right)^{-1} \partial_k g = \frac{1}{g} \left(g^{ij} \right)^{-1} \partial_k g \quad (14.382)$$

C'est ce que nous voulions (devions) démontrer.

□C.Q.F.D.

Maintenant en combinant:

$$\Gamma_{ih}^i = \frac{1}{2} g^{ij} \partial_h g_{ij} \quad (14.383)$$

démontré plus haut et le résultat que nous venons de démontrer:

$$\partial_h g_{ij} = \frac{1}{g} \left(g^{ij} \right)^{-1} \partial_h g \quad (14.384)$$

il vient:

$$\Gamma_{ih}^i = \frac{1}{2} g^{ij} \partial_h g_{ij} = \frac{1}{2} g^{ij} \left(\frac{1}{g} \left(g^{ij} \right)^{-1} \partial_h g \right) = \frac{1}{2g} \partial_h g \quad (14.385)$$

Nous avons donc:

$$g^{ij} \partial_k g_{ij} = \frac{\partial_k g}{g} = 2\Gamma_{ih}^i \quad (14.386)$$

Montrons qu'il est possible de dériver cette dernière relation de l'égalité importante suivante:

$$\Gamma_{ih}^i = \frac{1}{\sqrt{g}} \partial_h \sqrt{g} \quad (14.387)$$

Effectivement:

$$\Gamma_{ih}^i = \frac{1}{\sqrt{g}} \partial_h \sqrt{g} = \frac{1}{\sqrt{g}} \left(\frac{1}{2\sqrt{g}} \partial_h g \right) = \frac{1}{2g} \partial_h g \quad (14.388)$$

Cette relation ne veut pas dire grand-chose tant que nous n'en ferons pas un usage plus explicite lors de notre étude de la relativité générale (cf. chapitre de Relativité Générale).

Soit maintenant à déterminer la dérivée covariante seconde du tenseur métrique. Rappelons-nous avant d'aller plus loin (car c'est important) que nous avons obtenu:

$$\nabla_k g_{ij} = \partial_k g_{ij} - g_{ik} \Gamma_{jk}^k - g_{jk} \Gamma_{ik}^k \quad (14.389)$$

12. TENSEUR DE RIEMANN-CHRISTOFFEL

Rappelons que nous avons démontré plus haut que:

$$\nabla_k (\nabla_j v_i) = \partial_k (\nabla_j v_i) - (\nabla_j v_r) \Gamma_{ik}^r - (\nabla_r v_i) \Gamma_{jk}^r \quad (14.390)$$

Cette relation exprime sauf erreur de la part du rédacteur de ces lignes.... la dérivée covariante d'un tenseur d'ordre deux - tel que la métrique - sur un chemin géodésique dans deux directions parallèles (la dérivée covariante seconde permettant de créer la "perpendiculaire géodésique" entre les deux géodésiques infiniment proches de la dérivée covariante première). Nous appelons un tel déplacement: un "transport parallèle".

En y substituant:

$$\nabla_j v_i = \partial_j v_i - \Gamma_{ji}^l v_l \quad (14.391)$$

Nous avons alors:

$$\begin{aligned} \nabla_k (\nabla_j v_i) &= \partial_k (\partial_j v_i - \Gamma_{ji}^l v_l) - (\partial_j v_r - \Gamma_{jr}^l v_l) \Gamma_{ik}^r - (\partial_r v_i - \Gamma_{ri}^l v_l) \Gamma_{jk}^r \\ &= \partial_{kj} v_i - (\partial_k \Gamma_{ji}^l) v_l - \Gamma_{ji}^l (\partial_k v_l) - \Gamma_{ik}^r \partial_j v_r + \Gamma_{jr}^l \Gamma_{ik}^r v_l - \Gamma_{jk}^r \partial_r v_i + \Gamma_{jk}^r \Gamma_{ri}^l v_l \end{aligned} \quad (14.392)$$

Permutons maintenant les indices j et k dans l'expression précédente pour avoir une différentielle par rapport à un autre chemin:

$$\nabla_j (\nabla_k v_i) = \partial_{jk} v_i - (\partial_j \Gamma_{ki}^l) v_l - \Gamma_{ki}^l (\partial_j v_l) - \Gamma_{ij}^r \partial_k v_r + \Gamma_{kr}^l \Gamma_{ij}^r v_l - \Gamma_{kj}^r \partial_r v_i + \Gamma_{kj}^r \Gamma_{ri}^l v_l \quad (14.393)$$

En admettant que les composantes vérifient les propriétés classiques $\partial_{kj} v_i = \partial_{jk} v_i$, nous obtenons par soustraction des deux expressions précédentes:

$$\begin{aligned} \nabla_k (\nabla_j v_i) &= \partial_{kj} v_i - (\partial_k \Gamma_{ji}^l) v_l - \Gamma_{ji}^l \partial_k v_l - \Gamma_{ik}^r \partial_j v_r + \Gamma_{jr}^l \Gamma_{ik}^r v_l - \Gamma_{jk}^r \partial_r v_i + \Gamma_{jk}^r \Gamma_{ri}^l v_l \\ \nabla_j (\nabla_k v_i) &= \partial_{jk} v_i - (\partial_j \Gamma_{ki}^l) v_l - \Gamma_{ki}^l \partial_j v_l - \Gamma_{ij}^r \partial_k v_r + \Gamma_{kr}^l \Gamma_{ij}^r v_l - \Gamma_{kj}^r \partial_r v_i + \Gamma_{kj}^r \Gamma_{ri}^l v_l \\ - \frac{\nabla_k (\nabla_j v_i) - \nabla_j (\nabla_k v_i)}{\nabla_k (\nabla_j v_i) - \nabla_j (\nabla_k v_i)} &= \underbrace{(\partial_{kj} v_i - \partial_{jk} v_i)}_{=0} + (\partial_j \Gamma_{ki}^l - \partial_k \Gamma_{ji}^l) v_l + (\Gamma_{ki}^l \partial_j - \Gamma_{ji}^l \partial_k) v_l \\ &+ (\Gamma_{ij}^r \partial_k - \Gamma_{ik}^r \partial_j) v_r + (\Gamma_{jr}^l \Gamma_{ik}^r - \Gamma_{kr}^l \Gamma_{ij}^r) v_l + (\Gamma_{kj}^r - \Gamma_{jk}^r) \partial_r v_i + (\Gamma_{jk}^r \Gamma_{ri}^l - \Gamma_{kj}^r \Gamma_{ri}^l) v_l \end{aligned} \quad (14.394)$$

et puisque nous avons démontré que:

$$\Gamma_{jk}^r = \Gamma_{kj}^r \quad (14.395)$$

Nous avons donc:

$$\begin{aligned} \nabla_k (\nabla_j v_i) - \nabla_j (\nabla_k v_i) &= (\partial_j \Gamma_{ki}^l - \partial_k \Gamma_{ji}^l) v_l + (\Gamma_{ki}^l \partial_j - \Gamma_{ji}^l \partial_k) v_l + (\Gamma_{ij}^r \partial_k - \Gamma_{ik}^r \partial_j) v_r \\ &+ (\Gamma_{jr}^l \Gamma_{ik}^r - \Gamma_{kr}^l \Gamma_{ij}^r) v_l + \underbrace{(\Gamma_{kj}^r - \Gamma_{jk}^r) \partial_r v_i}_{=0} + \underbrace{(\Gamma_{jk}^r \Gamma_{ri}^l - \Gamma_{kj}^r \Gamma_{ri}^l) v_l}_{=0} \end{aligned} \quad (14.396)$$

Il reste alors:

$$\begin{aligned} \nabla_k (\nabla_j v_i) - \nabla_j (\nabla_k v_i) &= (\partial_j \Gamma_{ki}^l - \partial_k \Gamma_{ji}^l) v_l + (\Gamma_{ki}^l \partial_j - \Gamma_{ji}^l \partial_k) v_l + (\Gamma_{ij}^r \partial_k - \Gamma_{ik}^r \partial_j) v_r \\ &+ (\Gamma_{jr}^l \Gamma_{ik}^r - \Gamma_{kr}^l \Gamma_{ij}^r) v_l \end{aligned} \quad (14.397)$$

Comme le transport parallèle se fait sur des chemins de géodésiques infiniment proches, nous prenons la limite:

$$\partial_j v_l, \partial_k v_l, \partial_k v_r, \partial_j v_r \rightarrow 0 \quad (14.398)$$

ce qui sous-tend surtout le fait que le champ de vitesse est quasi égal en deux points parallèles infiniment proches.

Il reste alors:

$$\boxed{\nabla_k (\nabla_j v_i) - \nabla_j (\nabla_k v_i) \cong (\partial_j \Gamma_{ki}^l - \partial_k \Gamma_{ji}^l + \Gamma_{jr}^l \Gamma_{ik}^r - \Gamma_{kr}^l \Gamma_{ij}^r) v_l} \quad (14.399)$$

Cette relation exprime le fait que, comme la gravité, la courbure de l'espace-temps cause une accélération mutuelle entre les géodésiques. De plus, il est facile de constater, que l'accélération mutuelle entre les géodésiques est nulle si les tenseurs de Riemann-Christoffel sont nuls (typiquement en coordonnées cartésiennes, in extenso cela signifie pour un espace-temps plat). C'est exactement ce que nous attendons de la gravité: si nous n'observons aucune accélération, la courbure (nous allons de suite définir ce que c'est) est nulle et si la courbure est nulle, nous n'observons aucune accélération. Morale de l'histoire: la gravité est courbure et la courbure est gravité!!

Nous voyons que la quantité entre parenthèses est un tenseur d'ordre quatre que nous noterons sur ce site (car il y a plusieurs traditions dans la manière de le noter...):

$$\boxed{R_{i,jk}^l = \partial_j \Gamma_{ik}^l - \partial_k \Gamma_{ij}^l + \Gamma_{ik}^r \Gamma_{jr}^l - \Gamma_{ij}^r \Gamma_{kr}^l} \quad (14.400)$$

et qui résume à lui seul le transport parallèle et le fait que gravité et géométrie de l'espace sont liées ensemble.

Le tenseur $R_{i,jk}^l$ est appelé "**tenseur de Riemann-Christoffel**" ou "**tenseur de l'espace Riemannien**". La courbure d'un espace Riemannien peut aussi être caractérisée à l'aide de ce tenseur.

Si nous multiplions le tenseur $R_{i,jk}^l$ par g_{jk} , nous avons alors les données covariantes de ce tenseur telles que:

$$\boxed{R_{ij,rs} = g_{jk} R_{i,jk}^k} \quad (14.401)$$

et soient les relations suivantes que nous avons démontrées:

$$g_{jk} \Gamma_{rl}^k = \Gamma_{rjl} \quad g_{jk} \Gamma_{sl}^k = \Gamma_{sjl} \quad (14.402)$$

Dès lors, il vient:

$$\begin{aligned} g_{jk} R_{i,rs}^k &= g_{jk} \partial_r \Gamma_{is}^k - g_{jk} \partial_s \Gamma_{ir}^k + g_{jk} \Gamma_{is}^l \Gamma_{rl}^k - g_{jk} \Gamma_{ir}^l \Gamma_{sl}^k \\ &= g_{jk} \partial_r \Gamma_{is}^k - g_{jk} \partial_s \Gamma_{ir}^k + \Gamma_{is}^l \Gamma_{rjl} - \Gamma_{ir}^l \Gamma_{sjl} \end{aligned} \quad (14.403)$$

et remplaçons les quantités $g_{jk} \partial_r \Gamma_{is}^k$ par $\partial_r (g_{jk} \Gamma_{is}^k) - \Gamma_{is}^k \partial_r g_{jk}$. Nous obtenons alors:

$$\begin{aligned} g_{jk} R_{i,rs}^k &= \left(\partial_r (g_{jk} \Gamma_{is}^k) - \Gamma_{is}^k \partial_r g_{jk} \right) - \left(\partial_s (g_{jk} \Gamma_{ir}^k) - \Gamma_{ir}^k \partial_s g_{jk} \right) + \Gamma_{is}^l \Gamma_{rjl} - \Gamma_{ir}^l \Gamma_{sjl} \\ &= \partial_r (g_{jk} \Gamma_{is}^k) - \partial_s (g_{jk} \Gamma_{ir}^k) + (\Gamma_{is}^l \Gamma_{rjl} - \Gamma_{ir}^k \partial_r g_{jk}) - (\Gamma_{ir}^l \Gamma_{sjl} - \Gamma_{ir}^k \partial_s g_{jk}) \\ &= \partial_r (g_{jk} \Gamma_{is}^k) - \partial_s (g_{jk} \Gamma_{ir}^k) + (\Gamma_{is}^l \Gamma_{rjl} - \Gamma_{ir}^l \partial_r g_{jk}) - (\Gamma_{ir}^l \Gamma_{sjl} - \Gamma_{ir}^l \partial_s g_{jk}) \\ &= \partial_r (g_{jk} \Gamma_{is}^k) - \partial_s (g_{jk} \Gamma_{ir}^k) + \Gamma_{is}^l (\Gamma_{rjl} - \partial_r g_{jk}) - \Gamma_{ir}^l (\Gamma_{sjl} - \partial_s g_{jk}) \end{aligned} \quad (14.404)$$

Nous avons aussi démontré que:

$$\partial_k g_{ij} = \Gamma_{kji} + \Gamma_{kij} \Rightarrow \Gamma_{kij} - \partial_k g_{ij} = -\Gamma_{kji} \quad (14.405)$$

D'où:

$$g_{jk} R_{i,rs}^k = \partial_r (g_{jk} \Gamma_{is}^k) - \partial_s (g_{jk} \Gamma_{ir}^k) - \Gamma_{is}^l \Gamma_{rjl} + \Gamma_{ir}^l \Gamma_{sjl} \quad (14.406)$$

et comme:

$$g_{jk} \Gamma_{is}^k = \Gamma_{ijs} \quad g_{jk} \Gamma_{ir}^k = \Gamma_{ijr} \quad (14.407)$$

Nous avons:

$$g_{jk} R_{i,rs}^k = \partial_r \Gamma_{ijs} - \partial_s \Gamma_{ijr} - \Gamma_{is}^l \Gamma_{rjl} + \Gamma_{ir}^l \Gamma_{sjl} \quad (14.408)$$

et nous avons aussi démontré que:

$$\Gamma_{kji} = \frac{1}{2} (\partial_k g_{ij} + \partial_i g_{jk} - \partial_j g_{ki})$$

d'où:

$$\partial_r \Gamma_{ijs} = \frac{1}{2} (\partial_i g_{sj} + \partial_s g_{ji} - \partial_j g_{is}) \quad (14.409)$$

et en les reportant dans l'avant-dernière relation, nous obtenons:

$$\begin{aligned}
R_{ij,rs} &= g_{jk} R_{i,rs}^k = \frac{1}{2} \left(\partial_{ri} g_{sj} + \cancel{\partial_{rs} g_{ji}} - \partial_{rj} g_{is} \right) - \frac{1}{2} \left(\partial_{si} g_{rj} + \cancel{\partial_{sr} g_{ji}} - \partial_{sj} g_{ir} \right) \\
&\quad - \Gamma_{is}^l \Gamma_{rj}^l + \Gamma_{ir}^l \Gamma_{sj}^l = \frac{1}{2} \left(\partial_{ri} g_{sj} + \partial_{sj} g_{ir} - \partial_{rj} g_{is} - \partial_{si} g_{rj} \right) - \Gamma_{is}^l \Gamma_{rj}^l + \Gamma_{ir}^l \Gamma_{sj}^l
\end{aligned} \quad (14.410)$$

Nous avons donc finalement pour l'expression covariante du tenseur de Riemann-Christoffel:

$$R_{ij,rs} = \frac{1}{2} \left(\partial_{ri} g_{sj} + \partial_{sj} g_{ir} - \partial_{rj} g_{is} - \partial_{si} g_{rj} \right) - \Gamma_{is}^l \Gamma_{rj}^l + \Gamma_{ir}^l \Gamma_{sj}^l \quad (14.411)$$

Il convient de remarquer que le tenseur de Riemann-Christoffel est donc antisymétrique:

$$R_{ij,rs} = -R_{ji,rs} \quad (14.412)$$

Enfin, la permutation en bloc des indices ij et rs nous donne, par suite de la symétrie des g_{ij} et en invertissant leur ordre de dérivation:

$$R_{ij,rs} = R_{rs,ij} \quad (14.413)$$

Effectuons maintenant une permutation circulaire sur les indices j, r, s dans l'expression (obtenue donc juste un peu plus haut):

$$R_{ij,rs} = \frac{1}{2} \left(\partial_{ri} g_{sj} + \partial_{sj} g_{ir} - \partial_{rj} g_{is} - \partial_{si} g_{rj} \right) - \Gamma_{is}^l \Gamma_{rj}^l + \Gamma_{ir}^l \Gamma_{sj}^l \quad (14.414)$$

il vient:

$$\begin{aligned}
R_{ir,sj} &= \frac{1}{2} \left(\partial_{jr} g_{is} + \partial_{is} g_{jr} - \partial_{rs} g_{ij} - \partial_{ij} g_{rs} \right) - \Gamma_{ij}^l \Gamma_{sr}^l + \Gamma_{ir}^l \Gamma_{js}^l \\
R_{is,rj} &= \frac{1}{2} \left(\partial_{rs} g_{ij} + \partial_{ij} g_{rs} - \partial_{js} g_{ir} - \partial_{ir} g_{js} \right) - \Gamma_{ir}^l \Gamma_{js}^l + \Gamma_{is}^l \Gamma_{jr}^l
\end{aligned} \quad (14.415)$$

et nous avons alors (c'est très simple à contrôler avec en sommant les trois lignes ci-dessus):

$$R_{ij,rs} + R_{ir,sj} + R_{is,rj} = 0 \quad (14.416)$$

L'identité précédente est appelée "**première identité de Bianchi**". Nous avons donc aussi (nous changeons les notations des indices afin d'être plus conforme aux écritures habituelles en relativité générale):

$$\partial_\lambda R_{\alpha\beta,\mu\nu} + \partial_\nu R_{\alpha\beta,\lambda\mu} + \partial_\mu R_{\alpha\beta,\nu\lambda} = 0 \quad (14.417)$$

et par extension:

$$\nabla_\lambda R_{\alpha\beta,\mu\nu} + \nabla_\nu R_{\alpha\beta,\lambda\mu} + \nabla_\mu R_{\alpha\beta,\nu\lambda} = 0 \quad (14.418)$$

Rappelons qu'implicitement, cette relation (appelée "**deuxième identité de Bianchi**") exprime toujours simplement (si l'on peut dire...) le fait que gravité et géométrie de l'espace sont liées ensemble.

13. TENSEUR DE RICCI

Avant de voir les conséquences de l'identité de Bianchi, nous avons besoin de définir le "**tenseur de Ricci**":

$$R_{\alpha\beta} = R^{\mu}_{\alpha\mu\beta} \quad (14.419)$$

qui est donc simplement la contraction des premier et troisième indices du tenseur de Riemann-Christoffel que nous avons donné plus haut:

$$R^i_{\alpha,jk} = \partial_j \Gamma^i_{ik} - \partial_k \Gamma^i_{ij} + \Gamma^r_{ik} \Gamma^i_{jr} - \Gamma^r_{ij} \Gamma^i_{kr} \quad (14.420)$$

en d'autres termes c'est juste une notation plus condensée... et puis les lettres pour les indices supérieurs ou inférieurs ainsi que la présence de la virgule sont au libre choix de celui qui écrit (en fonction de l'humeur et surtout si le contexte permet d'éviter toute confusion).

Par exemple avec le tenseur de Riemann-Christoffel que nous venons de donner le tenseur de Ricci pourrait en fonction de l'humeur s'écrire des deux manières suivantes (nous gardons les indices avec les lettres latines):

$$\begin{aligned} R_{ik} &= R^l_{i,jk} \\ R_{ik} &= R^l_{jkl} \end{aligned} \quad (14.421)$$

D'autres contractions d'autres indices pourraient aussi être possibles mais parce que $R_{\alpha\beta,\mu\nu}$ est antisymétrique sur α, β et μ, ν alors la contraction sur ces indices reviennent à avoir $\pm R_{\alpha\beta}$.

De manière similaire, nous définissons le "**scalaire de Ricci**" (aussi parfois appelé "**scalaire de Riemann**") par la relation:

$$R = g^{\alpha\beta} R_{\alpha\beta} = g^{\alpha\beta} g^{\mu\nu} R_{\mu\alpha\nu\beta} \quad (14.422)$$

qui possède les propriétés suivantes:

- Si l'espace est plat, le scalaire de Ricci est nul
- Si l'espace est courbé comme une sphère, le scalaire de Ricci est positif
- Si l'espace est courbé comme une selle de cheval, le scalaire de Ricci est négatif

Soit explicitement (en changeant de notation pour les indices afin de bien insister sur le fait que cela n'a aucun impact!):

$$\begin{aligned}
R &= g^{ij} R_{ij} = g^{ij} R_{i\mu j}^{\mu} = g^{ij} \left(\partial_{\mu} \Gamma_{ij}^{\mu} - \partial_j \Gamma_{i\mu}^{\mu} + \Gamma_{ij}^r \Gamma_{\mu r}^{\mu} - \Gamma_{i\mu}^r \Gamma_{jr}^{\mu} \right) \\
&= g^{ij} \left(\partial_{\mu} \left(\frac{1}{2} g^{\mu l} (\partial_i g_{lj} + \partial_j g_{il} - \partial_l g_{ij}) \right) - \partial_j \left(\frac{1}{2} g^{\mu l} (\partial_i g_{l\mu} + \partial_{\mu} g_{il} - \partial_l g_{i\mu}) \right) \right) \\
&\quad + \left(\frac{1}{2} g^{rl} (\partial_i g_{lj} + \partial_j g_{il} - \partial_l g_{ij}) \right) \left(\frac{1}{2} g^{\mu l} (\partial_{\mu} g_{lr} + \partial_r g_{\mu l} - \partial_l g_{\mu r}) \right) \\
&\quad - \left(\frac{1}{2} g^{rl} (\partial_i g_{l\mu} + \partial_{\mu} g_{il} - \partial_l g_{i\mu}) \right) \left(\frac{1}{2} g^{\mu l} (\partial_j g_{lr} + \partial_r g_{jl} - \partial_l g_{jr}) \right) \quad (14.423)
\end{aligned}$$

Nous aurons des exemples pratiques concrets dans le chapitre de Relativité Générale pour les deux premiers cas mais regardons des exemples simplifiées pour les deux premiers (nous ne démontrerons par contre pas la réciproque).

Exemples:

E1. Commençons par la métrique de l'espace plat pour commencer (sans la composante temporelle). Nous avons (cf. chapitre de Relativité Générale):

$$g_{ij} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (14.424)$$

En reprenant la définition du scalaire de Ricci sous forme explicite:

$$\begin{aligned}
R &= g^{ij} \left(\partial_{\mu} \left(\frac{1}{2} g^{\mu l} (\partial_i g_{lj} + \partial_j g_{il} - \partial_l g_{ij}) \right) - \partial_j \left(\frac{1}{2} g^{\mu l} (\partial_i g_{l\mu} + \partial_{\mu} g_{il} - \partial_l g_{i\mu}) \right) \right) \\
&\quad + \left(\frac{1}{2} g^{rl} (\partial_i g_{lj} + \partial_j g_{il} - \partial_l g_{ij}) \right) \left(\frac{1}{2} g^{\mu l} (\partial_{\mu} g_{lr} + \partial_r g_{\mu l} - \partial_l g_{\mu r}) \right) \\
&\quad - \left(\frac{1}{2} g^{rl} (\partial_i g_{l\mu} + \partial_{\mu} g_{il} - \partial_l g_{i\mu}) \right) \left(\frac{1}{2} g^{\mu l} (\partial_j g_{lr} + \partial_r g_{jl} - \partial_l g_{jr}) \right) \quad (14.425)
\end{aligned}$$

Il est immédiat que R est nul puisque les dérivées partielles seront toutes nulles. Donc un espace plat a un scalaire de Ricci nul.

E2. Regardons maintenant avec la métrique du plan exprimée en coordonnées sphériques (sans la composante temporelle). Nous avons (cf. chapitre de Relativité Générale):

$$g_{ij} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2 \theta \end{bmatrix} \quad (14.426)$$

et:

$$g^{ij} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/r^2 & 0 \\ 0 & 0 & 1/r^2 \sin^2 \theta \end{bmatrix} \quad (14.427)$$

avec:

$$ds^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (14.428)$$

Nous savons que pour calculer le scalaire de Ricci (ou: courbure de Ricci), il nous faut donc calculer la contraction du tenseur de Riemann-Christoffel (soit: le tenseur de Ricci) qui lui-même dépend des symboles de Christoffel de deuxième espèce qui eux-mêmes dépendent des symboles de Christoffel de première espèce (argh!).

Nous allons donc commencer par le plus bas niveau, c'est-à-dire par déterminer tous les symboles de Christoffel de première espèce donnés pour rappel par:

$$\Gamma_{\mathfrak{ji}}^{\mathfrak{k}} = \frac{1}{2} (\partial_{\mathfrak{k}} g_{\mathfrak{ji}} + \partial_{\mathfrak{i}} g_{\mathfrak{jk}} - \partial_{\mathfrak{j}} g_{\mathfrak{ki}}) \quad (14.429)$$

Nous avons donc 3^3 , soit 27 symboles de Christoffel de première espèce possibles! Même si certains symboles sont égaux (nous l'avons démontré!), nous allons quand même tout calculer.

Commençons dans la joie et la bonne humeur...:

$$\Gamma_{111} = \frac{1}{2} (\partial_1 g_{11} + \partial_1 g_{11} - \partial_1 g_{11}) = \frac{1}{2} \partial_1 g_{11} = 0$$

$$\Gamma_{112} = \frac{1}{2} (\partial_1 g_{21} + \partial_2 g_{11} - \partial_1 g_{12}) = \frac{1}{2} \partial_2 g_{11} = 0$$

$$\Gamma_{113} = \frac{1}{2} (\partial_1 g_{31} + \partial_3 g_{11} - \partial_1 g_{13}) = \frac{1}{2} \partial_3 g_{11} = 0$$

$$\Gamma_{121} = \frac{1}{2} (\partial_1 g_{12} + \partial_1 g_{21} - \partial_2 g_{11}) = \frac{1}{2} \partial_2 g_{11} = 0$$

$$\Gamma_{122} = \frac{1}{2} (\partial_1 g_{22} + \partial_2 g_{21} - \partial_2 g_{12}) = \frac{1}{2} \partial_1 g_{22} = r$$

$$\Gamma_{123} = \frac{1}{2} (\partial_1 g_{32} + \partial_3 g_{21} - \partial_2 g_{13}) = 0$$

$$\Gamma_{131} = \frac{1}{2} (\partial_1 g_{13} + \partial_1 g_{31} - \partial_3 g_{11}) = -\frac{1}{2} \partial_3 g_{11} = 0$$

$$\Gamma_{132} = \frac{1}{2} (\partial_1 g_{23} + \partial_2 g_{31} - \partial_3 g_{12}) = 0$$

$$\Gamma_{133} = \frac{1}{2} (\partial_1 g_{33} + \partial_3 g_{31} - \partial_3 g_{13}) = \frac{1}{2} \partial_1 g_{33} = r \sin^2 \theta$$

$$\Gamma_{211} = \frac{1}{2}(\partial_2 g_{11} + \partial_1 g_{12} - \partial_1 g_{21}) = \frac{1}{2} \partial_2 g_{11} = 0$$

$$\Gamma_{212} = \frac{1}{2}(\partial_2 g_{21} + \partial_2 g_{12} - \partial_1 g_{22}) = -\frac{1}{2} \partial_1 g_{22} = -r$$

$$\Gamma_{213} = \frac{1}{2}(\partial_2 g_{31} + \partial_3 g_{12} - \partial_1 g_{23}) = 0$$

$$\Gamma_{221} = \frac{1}{2}(\partial_2 g_{12} + \partial_1 g_{22} - \partial_2 g_{21}) = \frac{1}{2} \partial_1 g_{22} = r$$

$$\Gamma_{222} = \frac{1}{2}(\partial_2 g_{22} + \partial_2 g_{22} - \partial_2 g_{22}) = \frac{1}{2} \partial_2 g_{22} = 0 \quad (14.430)$$

$$\Gamma_{223} = \frac{1}{2}(\partial_2 g_{32} + \partial_3 g_{22} - \partial_2 g_{23}) = \frac{1}{2} \partial_3 g_{22} = 0$$

$$\Gamma_{231} = \frac{1}{2}(\partial_2 g_{13} + \partial_1 g_{32} - \partial_3 g_{21}) = 0$$

$$\Gamma_{232} = \frac{1}{2}(\partial_2 g_{23} + \partial_2 g_{32} - \partial_3 g_{22}) = -\frac{1}{2} \partial_3 g_{22} = 0$$

$$\Gamma_{233} = \frac{1}{2}(\partial_2 g_{33} + \partial_3 g_{32} - \partial_3 g_{23}) = \frac{1}{2} \partial_2 g_{33} = r^2 \sin \theta \cos \theta$$

$$\Gamma_{311} = \frac{1}{2}(\partial_3 g_{11} + \partial_1 g_{13} - \partial_1 g_{31}) = \frac{1}{2} \partial_3 g_{11} = 0$$

$$\Gamma_{312} = \frac{1}{2}(\partial_3 g_{21} + \partial_2 g_{13} - \partial_1 g_{32}) = 0$$

$$\Gamma_{313} = \frac{1}{2}(\partial_3 g_{31} + \partial_3 g_{13} - \partial_1 g_{33}) = -\frac{1}{2} \partial_1 g_{33} = -r \sin^2 \theta$$

$$\Gamma_{321} = \frac{1}{2}(\partial_3 g_{12} + \partial_1 g_{23} - \partial_2 g_{31}) = 0$$

$$\Gamma_{322} = \frac{1}{2}(\partial_3 g_{22} + \partial_2 g_{23} - \partial_2 g_{32}) = \frac{1}{2} \partial_3 g_{22} = 0$$

$$\Gamma_{323} = \frac{1}{2}(\partial_3 g_{32} + \partial_3 g_{23} - \partial_2 g_{33}) = -\frac{1}{2} \partial_2 g_{33} = -r^2 \sin \theta \cos \theta$$

$$\Gamma_{331} = \frac{1}{2}(\partial_3 g_{13} + \partial_1 g_{33} - \partial_3 g_{31}) = \frac{1}{2} \partial_1 g_{33} = r \sin^2 \theta$$

$$\Gamma_{332} = \frac{1}{2}(\partial_3 g_{23} + \partial_2 g_{33} - \partial_3 g_{32}) = \frac{1}{2} \partial_2 g_{33} = r^2 \sin \theta \cos \theta$$

$$\Gamma_{333} = \frac{1}{2}(\partial_3 g_{33} + \partial_3 g_{33} - \partial_3 g_{33}) = \frac{1}{2} \partial_3 g_{33} = 0$$

Calculons maintenant tous les symboles de Christoffel de deuxième espèce dans les détails:

$$\Gamma_{ij}^i = g^{ij} \Gamma_{kij} = \frac{1}{2} g^{ij} (\partial_k g_{ij} + \partial_j g_{ki} - \partial_i g_{kj}) \quad (14.431)$$

Encore une fois, comme le tenseur métrique est diagonal, cela va nous simplifier les calculs!

Nous avons alors:

$$\begin{aligned}
 \Gamma_{12}^2 &= g^{22}\Gamma_{122} = \frac{1}{r^2} \cdot r = \frac{1}{r} \\
 \Gamma_{13}^3 &= g^{33}\Gamma_{133} = \frac{1}{r^2 \sin^2 \theta} \cdot r \sin^2 \theta = \frac{1}{r} \\
 \Gamma_{22}^1 &= g^{11}\Gamma_{212} = 1 \cdot (-r) = -r \\
 \Gamma_{21}^2 &= g^{22}\Gamma_{221} = \frac{1}{r^2} \cdot r = \frac{1}{r} \\
 \Gamma_{23}^3 &= g^{33}\Gamma_{233} = \frac{1}{r^2 \sin^2 \theta} \cdot r^2 \sin \theta \cos \theta = \frac{\cos \theta}{\sin \theta} = \frac{1}{\tan \theta} = \cot \theta \quad (14.432) \\
 \Gamma_{33}^1 &= g^{11}\Gamma_{313} = 1 \cdot (-r \sin^2 \theta) = -r \sin^2 \theta \\
 \Gamma_{33}^2 &= g^{22}\Gamma_{323} = \frac{1}{r^2} \cdot (-r^2 \sin \theta \cos \theta) = -\sin \theta \cos \theta \\
 \Gamma_{31}^3 &= g^{33}\Gamma_{331} = \frac{1}{r^2 \sin^2 \theta} \cdot r \sin^2 \theta = \frac{1}{r} \\
 \Gamma_{32}^3 &= g^{33}\Gamma_{332} = \frac{1}{r^2 \sin^2 \theta} \cdot r^2 \sin \theta \cos \theta = \frac{\cos \theta}{\sin \theta} = \frac{1}{\tan \theta} = \cot \theta
 \end{aligned}$$

Calculons maintenant les 9 composantes du tenseur de Ricci dans les détails selon:

$$R_{\alpha\beta} = R_{\alpha\mu\beta}^{\mu} = \partial_{\mu}\Gamma_{\alpha\beta}^{\mu} - \partial_{\beta}\Gamma_{\alpha\mu}^{\mu} + \Gamma_{\alpha\beta}^{\nu}\Gamma_{\mu\nu}^{\mu} - \Gamma_{\alpha\mu}^{\nu}\Gamma_{\beta\nu}^{\mu} \quad (14.433)$$

Nous avons alors (nous les calculons tous, même si nous savons que par la suite ceux qui n'ont pas $\alpha = \beta$ seront inutiles de par la métrique diagonale):

$$\begin{aligned}
 R_{11} &= R_{1\mu 1}^{\mu} = \partial_{\mu}\Gamma_{11}^{\mu} - \partial_1\Gamma_{1\mu}^{\mu} + \Gamma_{11}^{\nu}\Gamma_{\mu\nu}^{\mu} - \Gamma_{1\mu}^{\nu}\Gamma_{1\nu}^{\mu} = -\partial_1\Gamma_{1\mu}^{\mu} - \Gamma_{1\mu}^{\nu}\Gamma_{1\nu}^{\mu} \\
 &= -(\partial_1\Gamma_{11}^1 + \partial_1\Gamma_{12}^2 + \partial_1\Gamma_{13}^3) - (\Gamma_{11}^r\Gamma_{1r}^1 + \Gamma_{12}^r\Gamma_{1r}^2 + \Gamma_{13}^r\Gamma_{1r}^3) \\
 &= -(\partial_1\Gamma_{12}^2 + \partial_1\Gamma_{13}^3) - (\Gamma_{12}^r\Gamma_{1r}^2 + \Gamma_{13}^r\Gamma_{1r}^3) \\
 &= -(\partial_1\Gamma_{12}^2 + \partial_1\Gamma_{13}^3) - ((\Gamma_{12}^1\Gamma_{11}^2 + \Gamma_{12}^2\Gamma_{12}^2 + \Gamma_{12}^3\Gamma_{13}^2) + \Gamma_{13}^r\Gamma_{1r}^3) \\
 &= -(\partial_1\Gamma_{12}^2 + \partial_1\Gamma_{13}^3) - ((\Gamma_{12}^2\Gamma_{12}^2) + \Gamma_{13}^r\Gamma_{1r}^3) \\
 &= -(\partial_1\Gamma_{12}^2 + \partial_1\Gamma_{13}^3) - (\Gamma_{12}^2\Gamma_{12}^2 + (\Gamma_{13}^1\Gamma_{11}^3 + \Gamma_{13}^2\Gamma_{12}^3 + \Gamma_{13}^3\Gamma_{13}^3)) \\
 &= -(\partial_1\Gamma_{12}^2 + \partial_1\Gamma_{13}^3) - (\Gamma_{12}^2\Gamma_{12}^2 + \Gamma_{13}^3\Gamma_{13}^3) = -\left(\partial_r \frac{1}{r} + \partial_r \frac{1}{r}\right) - \left(\frac{1}{r} \frac{1}{r} + \frac{1}{r} \frac{1}{r}\right) \\
 &= -\left(-\frac{1}{r^2} - \frac{1}{r^2}\right) - \frac{2}{r^2} = \frac{2}{r^2} - \frac{2}{r^2} = 0
 \end{aligned}$$

$$\begin{aligned}
R_{12} &= R_{1\mu 2}^{\mu} = \partial_{\mu} \Gamma_{12}^{\mu} - \partial_2 \Gamma_{1\mu}^{\mu} + \Gamma_{12}^{\nu} \Gamma_{\mu\nu}^{\mu} - \Gamma_{1\mu}^{\nu} \Gamma_{2\nu}^{\mu} \\
&= \left(\partial_1 \Gamma_{12}^1 - \partial_2 \Gamma_{11}^1 + \Gamma_{12}^{\nu} \Gamma_{1\nu}^1 - \Gamma_{11}^{\nu} \Gamma_{2\nu}^1 \right) + \left(\partial_2 \Gamma_{12}^2 - \partial_2 \Gamma_{12}^2 + \Gamma_{12}^{\nu} \Gamma_{2\nu}^2 - \Gamma_{12}^{\nu} \Gamma_{2\nu}^2 \right) \\
&+ \left(\partial_3 \Gamma_{12}^3 - \partial_2 \Gamma_{13}^3 + \Gamma_{12}^{\nu} \Gamma_{3\nu}^3 - \Gamma_{13}^{\nu} \Gamma_{2\nu}^3 \right) = \left(-\partial_2 \Gamma_{13}^3 + \Gamma_{12}^{\nu} \Gamma_{3\nu}^3 - \Gamma_{13}^{\nu} \Gamma_{2\nu}^3 \right) \\
&= \left(-\partial_2 \Gamma_{13}^3 + \left(\Gamma_{12}^1 \Gamma_{31}^3 + \Gamma_{12}^2 \Gamma_{32}^3 + \Gamma_{12}^3 \Gamma_{33}^3 \right) - \Gamma_{13}^{\nu} \Gamma_{2\nu}^3 \right) \\
&= \left(-\partial_2 \Gamma_{13}^3 + \left(\Gamma_{12}^2 \Gamma_{32}^3 \right) - \Gamma_{13}^{\nu} \Gamma_{2\nu}^3 \right) = \left(-\partial_2 \Gamma_{13}^3 + \left(\Gamma_{12}^2 \Gamma_{32}^3 \right) - \left(\Gamma_{13}^1 \Gamma_{21}^3 + \Gamma_{13}^2 \Gamma_{22}^3 + \Gamma_{13}^3 \Gamma_{23}^3 \right) \right) \\
&= \left(-\partial_2 \Gamma_{13}^3 + \left(\Gamma_{12}^2 \Gamma_{32}^3 \right) - \left(\Gamma_{13}^3 \Gamma_{23}^3 \right) \right) = -\partial_2 \Gamma_{13}^3 + \Gamma_{12}^2 \Gamma_{32}^3 - \Gamma_{13}^3 \Gamma_{23}^3 \\
&= -\partial_{\theta} \frac{1}{r} + \frac{1}{r} \cot \theta - \frac{1}{r} \cot \theta = 0
\end{aligned}$$

$$\begin{aligned}
R_{13} &= R_{1\mu 3}^{\mu} = \partial_{\mu} \Gamma_{13}^{\mu} - \partial_3 \Gamma_{1\mu}^{\mu} + \Gamma_{13}^{\nu} \Gamma_{\mu\nu}^{\mu} - \Gamma_{1\mu}^{\nu} \Gamma_{3\nu}^{\mu} \\
&= \left(\partial_1 \Gamma_{13}^1 - \partial_3 \Gamma_{11}^1 + \Gamma_{13}^{\nu} \Gamma_{1\nu}^1 - \Gamma_{11}^{\nu} \Gamma_{3\nu}^1 \right) + \left(\partial_2 \Gamma_{13}^2 - \partial_3 \Gamma_{12}^2 + \Gamma_{13}^{\nu} \Gamma_{2\nu}^2 - \Gamma_{12}^{\nu} \Gamma_{3\nu}^2 \right) \\
&+ \left(\partial_3 \Gamma_{13}^3 - \partial_3 \Gamma_{13}^3 + \Gamma_{13}^{\nu} \Gamma_{3\nu}^3 - \Gamma_{13}^{\nu} \Gamma_{3\nu}^3 \right) = \left(-\partial_3 \Gamma_{12}^2 + \Gamma_{13}^{\nu} \Gamma_{2\nu}^2 - \Gamma_{12}^{\nu} \Gamma_{3\nu}^2 \right) \\
&= \left(-\partial_3 \Gamma_{12}^2 + \left(\Gamma_{13}^1 \Gamma_{21}^2 + \Gamma_{13}^2 \Gamma_{22}^2 + \Gamma_{13}^3 \Gamma_{23}^2 \right) - \Gamma_{12}^{\nu} \Gamma_{3\nu}^2 \right) \\
&= \left(-\partial_3 \Gamma_{12}^2 + \left(\Gamma_{13}^3 \Gamma_{23}^2 \right) - \Gamma_{12}^{\nu} \Gamma_{3\nu}^2 \right) = \left(-\partial_3 \Gamma_{12}^2 + \left(\Gamma_{13}^3 \Gamma_{23}^2 \right) - \left(\Gamma_{12}^1 \Gamma_{31}^2 + \Gamma_{12}^2 \Gamma_{32}^2 + \Gamma_{12}^3 \Gamma_{33}^2 \right) \right) \\
&= \left(-\partial_3 \Gamma_{12}^2 \right) = -\partial_{\phi} \frac{1}{r} = 0
\end{aligned}$$

$$\begin{aligned}
R_{21} &= R_{2\mu d}^{\mu} = \partial_{\mu} \Gamma_{21}^{\mu} - \partial_1 \Gamma_{2\mu}^{\mu} + \Gamma_{21}^r \Gamma_{\mu r}^{\mu} - \Gamma_{2\mu}^r \Gamma_{1r}^{\mu} \\
&= \left(\partial_1 \Gamma_{21}^1 - \partial_1 \Gamma_{21}^1 + \Gamma_{21}^r \Gamma_{1r}^1 - \Gamma_{21}^r \Gamma_{1r}^1 \right) + \left(\partial_2 \Gamma_{21}^2 - \partial_1 \Gamma_{22}^2 + \Gamma_{21}^r \Gamma_{2r}^2 - \Gamma_{22}^r \Gamma_{1r}^2 \right) \\
&\quad + \left(\partial_3 \Gamma_{21}^3 - \partial_1 \Gamma_{23}^3 + \Gamma_{21}^r \Gamma_{3r}^3 - \Gamma_{23}^r \Gamma_{1r}^3 \right) \\
&= \left(\partial_2 \Gamma_{21}^2 + \Gamma_{21}^r \Gamma_{2r}^2 - \Gamma_{22}^r \Gamma_{1r}^2 \right) + \left(-\partial_1 \Gamma_{23}^3 + \Gamma_{21}^r \Gamma_{3r}^3 - \Gamma_{23}^r \Gamma_{1r}^3 \right) \\
&= \left(\partial_2 \Gamma_{21}^2 + \left(\Gamma_{21}^1 \Gamma_{21}^2 + \Gamma_{21}^2 \Gamma_{22}^2 + \Gamma_{21}^3 \Gamma_{23}^2 \right) - \Gamma_{22}^r \Gamma_{1r}^2 \right) + \left(-\partial_1 \Gamma_{23}^3 + \Gamma_{21}^r \Gamma_{3r}^3 - \Gamma_{23}^r \Gamma_{1r}^3 \right) \\
&= \left(\partial_2 \Gamma_{21}^2 - \Gamma_{22}^r \Gamma_{1r}^2 \right) + \left(-\partial_1 \Gamma_{23}^3 + \Gamma_{21}^r \Gamma_{3r}^3 - \Gamma_{23}^r \Gamma_{1r}^3 \right) \\
&= \left(\partial_2 \Gamma_{21}^2 - \left(\Gamma_{22}^1 \Gamma_{11}^2 + \Gamma_{22}^2 \Gamma_{12}^2 + \Gamma_{22}^3 \Gamma_{13}^2 \right) \right) + \left(-\partial_1 \Gamma_{23}^3 + \Gamma_{21}^r \Gamma_{3r}^3 - \Gamma_{23}^r \Gamma_{1r}^3 \right) \tag{14.434} \\
&= \left(\partial_2 \Gamma_{21}^2 \right) + \left(-\partial_1 \Gamma_{23}^3 + \Gamma_{21}^r \Gamma_{3r}^3 - \Gamma_{23}^r \Gamma_{1r}^3 \right) \\
&= \partial_2 \Gamma_{21}^2 + \left(-\partial_1 \Gamma_{23}^3 + \left(\Gamma_{21}^1 \Gamma_{31}^3 + \Gamma_{21}^2 \Gamma_{32}^3 + \Gamma_{21}^3 \Gamma_{33}^3 \right) - \Gamma_{23}^r \Gamma_{1r}^3 \right) \\
&= \partial_2 \Gamma_{21}^2 + \left(-\partial_1 \Gamma_{23}^3 + \left(\Gamma_{21}^2 \Gamma_{32}^3 \right) - \Gamma_{23}^r \Gamma_{1r}^3 \right) \\
&= \partial_2 \Gamma_{21}^2 + \left(-\partial_1 \Gamma_{23}^3 + \Gamma_{21}^2 \Gamma_{32}^3 - \left(\Gamma_{23}^1 \Gamma_{11}^3 + \Gamma_{23}^2 \Gamma_{12}^3 + \Gamma_{23}^3 \Gamma_{13}^3 \right) \right) \\
&= \partial_2 \Gamma_{21}^2 + \left(-\partial_1 \Gamma_{23}^3 + \Gamma_{21}^2 \Gamma_{32}^3 - \left(\Gamma_{23}^3 \Gamma_{13}^3 \right) \right) = \partial_2 \Gamma_{21}^2 - \partial_1 \Gamma_{23}^3 + \Gamma_{21}^2 \Gamma_{32}^3 - \Gamma_{23}^3 \Gamma_{13}^3 \\
&= \partial_{\theta} \frac{1}{r} - \partial_r \cot \theta + \frac{1}{r} \cot \theta - \frac{1}{r} \cot \theta = 0
\end{aligned}$$

$$\begin{aligned}
R_{22} &= R_{2\mu 2}^{\mu} = \partial_{\mu} \Gamma_{22}^{\mu} - \partial_2 \Gamma_{2\mu}^{\mu} + \Gamma_{22}^r \Gamma_{\mu r}^{\mu} - \Gamma_{2\mu}^r \Gamma_{2r}^{\mu} \\
&= \left(\partial_1 \Gamma_{22}^1 - \partial_2 \Gamma_{21}^1 + \Gamma_{22}^r \Gamma_{1r}^1 - \Gamma_{21}^r \Gamma_{2r}^1 \right) + \left(\partial_2 \Gamma_{22}^2 - \partial_2 \Gamma_{22}^2 + \Gamma_{22}^r \Gamma_{2r}^2 - \Gamma_{22}^r \Gamma_{2r}^2 \right) \\
&+ \left(\partial_3 \Gamma_{22}^3 - \partial_2 \Gamma_{23}^3 + \Gamma_{22}^r \Gamma_{3r}^3 - \Gamma_{23}^r \Gamma_{2r}^3 \right) \\
&= \left(\partial_1 \Gamma_{22}^1 - \Gamma_{21}^r \Gamma_{2r}^1 \right) + \left(-\partial_2 \Gamma_{23}^3 + \Gamma_{22}^r \Gamma_{3r}^3 - \Gamma_{23}^r \Gamma_{2r}^3 \right) \\
&= \left(\partial_1 \Gamma_{22}^1 - \left(\Gamma_{21}^1 \Gamma_{21}^1 + \Gamma_{21}^2 \Gamma_{22}^1 + \Gamma_{21}^3 \Gamma_{23}^1 \right) \right) + \left(-\partial_2 \Gamma_{23}^3 + \Gamma_{22}^r \Gamma_{3r}^3 - \Gamma_{23}^r \Gamma_{2r}^3 \right) \\
&= \left(\partial_1 \Gamma_{22}^1 - \Gamma_{21}^2 \Gamma_{22}^1 \right) + \left(-\partial_2 \Gamma_{23}^3 + \Gamma_{22}^r \Gamma_{3r}^3 - \Gamma_{23}^r \Gamma_{2r}^3 \right) \\
&= \left(\partial_1 \Gamma_{22}^1 - \Gamma_{21}^2 \Gamma_{22}^1 \right) + \left(-\partial_2 \Gamma_{23}^3 + \left(\Gamma_{22}^1 \Gamma_{31}^3 + \Gamma_{22}^2 \Gamma_{32}^3 + \Gamma_{22}^3 \Gamma_{33}^3 \right) - \Gamma_{23}^r \Gamma_{2r}^3 \right) \\
&= \left(\partial_1 \Gamma_{22}^1 - \Gamma_{21}^2 \Gamma_{22}^1 \right) + \left(-\partial_2 \Gamma_{23}^3 + \Gamma_{22}^1 \Gamma_{31}^3 - \Gamma_{23}^r \Gamma_{2r}^3 \right) \\
&= \left(\partial_1 \Gamma_{22}^1 - \Gamma_{21}^2 \Gamma_{22}^1 \right) + \left(-\partial_2 \Gamma_{23}^3 + \Gamma_{22}^1 \Gamma_{31}^3 - \left(\Gamma_{23}^1 \Gamma_{21}^3 + \Gamma_{23}^2 \Gamma_{22}^3 + \Gamma_{23}^3 \Gamma_{23}^3 \right) \right) \\
&= \left(\partial_1 \Gamma_{22}^1 - \Gamma_{21}^2 \Gamma_{22}^1 \right) + \left(-\partial_2 \Gamma_{23}^3 + \Gamma_{22}^1 \Gamma_{31}^3 - \Gamma_{23}^3 \Gamma_{23}^3 \right) \\
&= \partial_1 \Gamma_{22}^1 - \Gamma_{21}^2 \Gamma_{22}^1 - \partial_2 \Gamma_{23}^3 + \Gamma_{22}^1 \Gamma_{31}^3 - \Gamma_{23}^3 \Gamma_{23}^3 \\
&= \partial_r (-r) - \frac{1}{r} \cdot (-r) - \partial_{\theta} \cot \theta + (-r) \frac{1}{r} - \cot^2 \theta \\
&= -1 + 1 - (-1 - \cot^2 \theta) - 1 - \cot^2 \theta = 0
\end{aligned}$$

$$\begin{aligned}
R_{23} &= R_{2\mu 3}^{\mu} = \partial_{\mu} \Gamma_{23}^{\mu} - \partial_3 \Gamma_{2\mu}^{\mu} + \Gamma_{23}^r \Gamma_{\mu r}^{\mu} - \Gamma_{2\mu}^r \Gamma_{3r}^{\mu} \\
&= \left(\partial_1 \Gamma_{23}^1 - \partial_3 \Gamma_{21}^1 + \Gamma_{23}^r \Gamma_{1r}^1 - \Gamma_{21}^r \Gamma_{3r}^1 \right) + \left(\partial_2 \Gamma_{23}^2 - \partial_3 \Gamma_{22}^2 + \Gamma_{23}^r \Gamma_{2r}^2 - \Gamma_{22}^r \Gamma_{3r}^2 \right) \\
&+ \left(\partial_3 \Gamma_{23}^3 - \partial_3 \Gamma_{23}^3 + \Gamma_{23}^r \Gamma_{3r}^3 - \Gamma_{23}^r \Gamma_{3r}^3 \right) = \left(-\Gamma_{21}^r \Gamma_{3r}^1 \right) + \left(\Gamma_{23}^r \Gamma_{2r}^2 - \Gamma_{22}^r \Gamma_{3r}^2 \right) \\
&= -\left(\Gamma_{21}^1 \Gamma_{31}^1 + \Gamma_{21}^2 \Gamma_{32}^1 + \Gamma_{21}^3 \Gamma_{33}^1 \right) + \left(\Gamma_{23}^r \Gamma_{2r}^2 - \Gamma_{22}^r \Gamma_{3r}^2 \right) \\
&= \left(\Gamma_{23}^r \Gamma_{2r}^2 - \Gamma_{22}^r \Gamma_{3r}^2 \right) = \left(\Gamma_{23}^1 \Gamma_{21}^2 + \Gamma_{23}^2 \Gamma_{22}^2 + \Gamma_{23}^3 \Gamma_{23}^2 \right) - \Gamma_{22}^r \Gamma_{3r}^2 \\
&= -\Gamma_{22}^r \Gamma_{3r}^2 = -\left(\Gamma_{22}^1 \Gamma_{31}^2 + \Gamma_{22}^2 \Gamma_{32}^2 + \Gamma_{22}^3 \Gamma_{33}^2 \right) = 0
\end{aligned}$$

$$\begin{aligned}
R_{31} &= R_{3\mu}^{\mu} = \partial_{\mu} \Gamma_{31}^{\mu} - \partial_1 \Gamma_{3\mu}^{\mu} + \Gamma_{31}^r \Gamma_{\mu r}^{\mu} - \Gamma_{3\mu}^r \Gamma_{1r}^{\mu} \\
&= \left(\partial_1 \Gamma_{31}^1 - \partial_1 \Gamma_{31}^1 + \Gamma_{31}^r \Gamma_{1r}^1 - \Gamma_{31}^r \Gamma_{1r}^1 \right) + \left(\partial_2 \Gamma_{31}^2 - \partial_1 \Gamma_{32}^2 + \Gamma_{31}^r \Gamma_{2r}^2 - \Gamma_{32}^r \Gamma_{1r}^2 \right) \\
&\quad + \left(\partial_3 \Gamma_{31}^3 - \partial_1 \Gamma_{33}^3 + \Gamma_{31}^r \Gamma_{3r}^3 - \Gamma_{33}^r \Gamma_{1r}^3 \right) \\
&= \left(\Gamma_{31}^r \Gamma_{2r}^2 - \Gamma_{32}^r \Gamma_{1r}^2 \right) + \left(\partial_3 \Gamma_{31}^3 + \Gamma_{31}^r \Gamma_{3r}^3 - \Gamma_{33}^r \Gamma_{1r}^3 \right) \\
&= \left(\left(\Gamma_{31}^1 \Gamma_{21}^2 + \Gamma_{31}^2 \Gamma_{22}^2 + \Gamma_{31}^3 \Gamma_{23}^2 \right) - \Gamma_{32}^r \Gamma_{1r}^2 \right) + \left(\partial_3 \Gamma_{31}^3 + \Gamma_{31}^r \Gamma_{3r}^3 - \Gamma_{33}^r \Gamma_{1r}^3 \right) \\
&= \left(-\Gamma_{32}^r \Gamma_{1r}^2 \right) + \left(\partial_3 \Gamma_{31}^3 + \Gamma_{31}^r \Gamma_{3r}^3 - \Gamma_{33}^r \Gamma_{1r}^3 \right) \\
&= -\left(\Gamma_{32}^1 \Gamma_{11}^2 + \Gamma_{32}^2 \Gamma_{12}^2 + \Gamma_{32}^3 \Gamma_{13}^2 \right) + \left(\partial_3 \Gamma_{31}^3 + \Gamma_{31}^r \Gamma_{3r}^3 - \Gamma_{33}^r \Gamma_{1r}^3 \right) \\
&= \left(\partial_3 \Gamma_{31}^3 + \Gamma_{31}^r \Gamma_{3r}^3 - \Gamma_{33}^r \Gamma_{1r}^3 \right) = \left(\partial_3 \Gamma_{31}^3 + \left(\Gamma_{31}^1 \Gamma_{31}^3 + \Gamma_{31}^2 \Gamma_{32}^3 + \Gamma_{31}^3 \Gamma_{33}^3 \right) - \Gamma_{33}^r \Gamma_{1r}^3 \right) \\
&= \partial_3 \Gamma_{31}^3 - \Gamma_{33}^r \Gamma_{1r}^3 = \partial_3 \Gamma_{31}^3 - \left(\Gamma_{33}^1 \Gamma_{11}^3 + \Gamma_{33}^2 \Gamma_{12}^3 + \Gamma_{33}^3 \Gamma_{13}^3 \right) \\
&= \partial_3 \Gamma_{31}^3 = \partial_{\phi} \frac{1}{r} = 0
\end{aligned}$$

$$\begin{aligned}
R_{32} &= R_{3\mu}^{\mu} = \partial_{\mu} \Gamma_{32}^{\mu} - \partial_2 \Gamma_{3\mu}^{\mu} + \Gamma_{32}^r \Gamma_{\mu r}^{\mu} - \Gamma_{3\mu}^r \Gamma_{2r}^{\mu} \\
&= \left(\partial_1 \Gamma_{32}^1 - \partial_2 \Gamma_{31}^1 + \Gamma_{32}^r \Gamma_{1r}^1 - \Gamma_{31}^r \Gamma_{2r}^1 \right) + \left(\partial_2 \Gamma_{32}^2 - \partial_2 \Gamma_{32}^2 + \Gamma_{32}^r \Gamma_{2r}^2 - \Gamma_{32}^r \Gamma_{2r}^2 \right) \\
&\quad + \left(\partial_3 \Gamma_{32}^3 - \partial_2 \Gamma_{33}^3 + \Gamma_{32}^r \Gamma_{3r}^3 - \Gamma_{33}^r \Gamma_{2r}^3 \right) \\
&= \left(-\Gamma_{31}^r \Gamma_{2r}^1 \right) + \left(\partial_3 \Gamma_{32}^3 + \Gamma_{32}^r \Gamma_{3r}^3 - \Gamma_{33}^r \Gamma_{2r}^3 \right) \\
&= -\left(\Gamma_{31}^1 \Gamma_{21}^1 + \Gamma_{31}^2 \Gamma_{22}^1 + \Gamma_{31}^3 \Gamma_{23}^1 \right) + \left(\partial_3 \Gamma_{32}^3 + \Gamma_{32}^r \Gamma_{3r}^3 - \Gamma_{33}^r \Gamma_{2r}^3 \right) \\
&= \left(\partial_3 \Gamma_{32}^3 + \Gamma_{32}^r \Gamma_{3r}^3 - \Gamma_{33}^r \Gamma_{2r}^3 \right) = \partial_3 \Gamma_{32}^3 + \left(\Gamma_{32}^1 \Gamma_{31}^3 + \Gamma_{32}^2 \Gamma_{32}^3 + \Gamma_{32}^3 \Gamma_{33}^3 \right) - \Gamma_{33}^r \Gamma_{2r}^3 \\
&= \partial_3 \Gamma_{32}^3 - \Gamma_{33}^r \Gamma_{2r}^3 = \partial_3 \Gamma_{32}^3 - \left(\Gamma_{33}^1 \Gamma_{21}^3 + \Gamma_{33}^2 \Gamma_{22}^3 + \Gamma_{33}^3 \Gamma_{23}^3 \right) \\
&= \partial_3 \Gamma_{32}^3 = \partial_{\phi} \cot \theta = 0
\end{aligned}$$

$$\begin{aligned}
R_{33} &= R_{3\mu\beta}^{\mu} = \partial_{\mu} \Gamma_{33}^{\mu} - \partial_3 \Gamma_{3\mu}^{\mu} + \Gamma_{33}^r \Gamma_{\mu r}^{\mu} - \Gamma_{3\mu}^r \Gamma_{3r}^{\mu} \\
&= (\partial_1 \Gamma_{33}^1 - \partial_3 \Gamma_{31}^1 + \Gamma_{33}^r \Gamma_{1r}^1 - \Gamma_{31}^r \Gamma_{3r}^1) + (\partial_2 \Gamma_{33}^2 - \partial_3 \Gamma_{32}^2 + \Gamma_{33}^r \Gamma_{2r}^2 - \Gamma_{32}^r \Gamma_{3r}^2) \\
&\quad + (\partial_3 \Gamma_{33}^3 - \partial_3 \Gamma_{33}^3 + \Gamma_{33}^r \Gamma_{3r}^3 - \Gamma_{33}^r \Gamma_{3r}^3) \\
&= (\partial_1 \Gamma_{33}^1 - \Gamma_{31}^r \Gamma_{3r}^1) + (\partial_2 \Gamma_{33}^2 + \Gamma_{33}^r \Gamma_{2r}^2 - \Gamma_{32}^r \Gamma_{3r}^2) \\
&= (\partial_1 \Gamma_{33}^1 - (\Gamma_{31}^1 \Gamma_{31}^1 + \Gamma_{31}^2 \Gamma_{32}^1 + \Gamma_{31}^3 \Gamma_{33}^1)) + (\partial_2 \Gamma_{33}^2 + \Gamma_{33}^r \Gamma_{2r}^2 - \Gamma_{32}^r \Gamma_{3r}^2) \\
&= (\partial_1 \Gamma_{33}^1 - \Gamma_{31}^3 \Gamma_{33}^1) + (\partial_2 \Gamma_{33}^2 + \Gamma_{33}^r \Gamma_{2r}^2 - \Gamma_{32}^r \Gamma_{3r}^2) \\
&= (\partial_1 \Gamma_{33}^1 - \Gamma_{31}^3 \Gamma_{33}^1) + (\partial_2 \Gamma_{33}^2 + (\Gamma_{33}^1 \Gamma_{21}^2 + \Gamma_{33}^2 \Gamma_{22}^2 + \Gamma_{33}^3 \Gamma_{23}^2) - \Gamma_{32}^r \Gamma_{3r}^2) \\
&= (\partial_1 \Gamma_{33}^1 - \Gamma_{31}^3 \Gamma_{33}^1) + (\partial_2 \Gamma_{33}^2 + \Gamma_{33}^1 \Gamma_{21}^2 - \Gamma_{32}^r \Gamma_{3r}^2) \\
&= (\partial_1 \Gamma_{33}^1 - \Gamma_{31}^3 \Gamma_{33}^1) + (\partial_2 \Gamma_{33}^2 + \Gamma_{33}^1 \Gamma_{21}^2 - (\Gamma_{32}^1 \Gamma_{31}^2 + \Gamma_{32}^2 \Gamma_{32}^2 + \Gamma_{32}^3 \Gamma_{33}^2)) \\
&= (\partial_1 \Gamma_{33}^1 - \Gamma_{31}^3 \Gamma_{33}^1) + (\partial_2 \Gamma_{33}^2 + \Gamma_{33}^1 \Gamma_{21}^2 - \Gamma_{32}^3 \Gamma_{33}^2) \\
&= \partial_r (-r \sin^2 \theta) - \frac{1}{r} (-r \sin^2 \theta) + \partial_{\theta} (-\sin \theta \cos \theta) + (-r \sin^2 \theta) \frac{1}{r} - \cot \theta (-\sin \theta \cos \theta) \\
&= -\sin^2 \theta + \sin^2 \theta - \partial_{\theta} (\sin \theta \cos \theta) - \sin^2 \theta + \cot \theta \sin \theta \cos \theta \\
&= -(-\sin^2 \theta + \cos^2 \theta) - \sin^2 \theta + \frac{\cos \theta}{\sin \theta} \sin \theta \cos \theta \\
&= (\sin^2 \theta - \cos^2 \theta) - \sin^2 \theta + \cos^2 \theta = 0
\end{aligned}$$

Calculons maintenant le scalaire de Ricci:

$$R = g^{\alpha\beta} R_{\alpha\beta} \quad (14.435)$$

Nous avons alors:

$$\begin{aligned}
R &= g^{\alpha\beta} R_{\alpha\beta} = g^{11} R_{11} + g^{22} R_{22} + g^{33} R_{33} \\
&= 1 \cdot R_{11} + \frac{1}{r^2} R_{22} + \frac{1}{r^2 \sin^2 \theta} R_{33} \\
&= 1 \cdot 0 + \frac{1}{r^2} \cdot 0 + \frac{1}{r^2 \sin^2 \theta} \cdot 0 = 0
\end{aligned} \quad (14.436)$$

Le scalaire de Ricci est donc nul aussi. Ce résultat peut surprendre, mais en réalité il est logique puisque nous n'avons fait que de calculer la courbure scalaire d'un espace plat exprimé en coordonnées sphériques.

E3. Imposons-nous maintenant la métrique diagonale de surface de la 2-sphère $\mathcal{S}^2$ (sans la composante temporelle). Nous avons alors conformément à ce que nous avons vu dans le chapitre de Géométrie Différentielle:

$$g_{ij} = \begin{bmatrix} r^2 & 0 \\ 0 & r^2 \sin^2 \theta \end{bmatrix} \quad (14.437)$$

et:

$$g^{ij} = \begin{bmatrix} 1/r^2 & 0 \\ 0 & 1/r^2 \sin^2 \theta \end{bmatrix} \quad (14.438)$$

Soit (cf. chapitre de Géométrie Différentielle):

$$ds^2 = r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (14.439)$$

où r est une constante!

Nous allons donc commencer par le plus bas niveau, c'est-à-dire par déterminer tous les symboles de Christoffel de première espèce donnés pour rappel par:

$$\Gamma_{\bar{k}\bar{i}} = \frac{1}{2} (\partial_{\bar{k}} g_{\bar{i}\bar{i}} + \partial_{\bar{i}} g_{\bar{k}\bar{k}} - \partial_{\bar{j}} g_{\bar{k}\bar{i}}) \quad (14.440)$$

Nous avons donc 2^3 , soit 8 symboles de Christoffel de première espèce possibles! Même si certains symboles sont égaux (nous l'avons démontré!), nous allons quand même tout calculer.

$$\begin{aligned} \Gamma_{111} &= \frac{1}{2} (\partial_1 g_{11} + \partial_1 g_{11} - \partial_1 g_{11}) = \frac{1}{2} \partial_1 g_{11} = 0 \\ \Gamma_{112} &= \frac{1}{2} (\partial_1 g_{21} + \partial_2 g_{11} - \partial_1 g_{12}) = \frac{1}{2} \partial_2 g_{11} = 0 \\ \Gamma_{121} &= \frac{1}{2} (\partial_1 g_{12} + \partial_1 g_{21} - \partial_2 g_{11}) = -\frac{1}{2} \partial_2 g_{11} = 0 \\ \Gamma_{122} &= \frac{1}{2} (\partial_1 g_{22} + \partial_2 g_{21} - \partial_2 g_{12}) = \frac{1}{2} \partial_1 g_{22} = r^2 \sin \theta \cos \theta \\ \Gamma_{211} &= \frac{1}{2} (\partial_2 g_{11} + \partial_1 g_{12} - \partial_1 g_{21}) = \frac{1}{2} \partial_2 g_{11} = 0 \\ \Gamma_{212} &= \frac{1}{2} (\partial_2 g_{21} + \partial_2 g_{12} - \partial_1 g_{22}) = -\frac{1}{2} \partial_1 g_{22} = -r^2 \sin \theta \cos \theta \\ \Gamma_{221} &= \frac{1}{2} (\partial_2 g_{12} + \partial_1 g_{22} - \partial_2 g_{21}) = \frac{1}{2} \partial_1 g_{22} = r^2 \sin \theta \cos \theta \\ \Gamma_{222} &= \frac{1}{2} (\partial_2 g_{22} + \partial_2 g_{22} - \partial_2 g_{22}) = \frac{1}{2} \partial_2 g_{22} = 0 \end{aligned} \quad (14.441)$$

Calculons maintenant tous les symboles de Christoffel de deuxième espèce dans les détails:

$$\Gamma_{\bar{i}}^{\bar{j}} = g^{\bar{i}\bar{j}} \Gamma_{\bar{k}\bar{l}} = \frac{1}{2} g^{\bar{i}\bar{j}} (\partial_{\bar{k}} g_{\bar{i}\bar{l}} + \partial_{\bar{l}} g_{\bar{k}\bar{i}} - \partial_{\bar{i}} g_{\bar{k}\bar{l}}) \quad (14.442)$$

Encore une fois, comme le tenseur métrique est diagonal, cela va nous simplifier les calculs!

Nous avons alors:

$$\begin{aligned}\Gamma_{12}^2 &= g^{22}\Gamma_{122} = \frac{1}{r^2 \sin^2 \theta} \cdot r^2 \sin \theta \cos \theta = \cot \theta \\ \Gamma_{22}^1 &= g^{11}\Gamma_{212} = \frac{1}{r^2} \cdot (-r^2 \sin \theta \cos \theta) = -\sin \theta \cos \theta \quad (14.443) \\ \Gamma_{21}^2 &= g^{22}\Gamma_{221} = \frac{1}{r^2 \sin^2 \theta} \cdot r^2 \sin \theta \cos \theta = \cot \theta\end{aligned}$$

Calculons maintenant les 4 composantes du tenseur de Ricci dans les détails selon:

$$R_{\alpha\beta} = R_{\alpha\mu\beta}^{\mu} = \partial_{\mu}\Gamma_{\alpha\beta}^{\mu} - \partial_{\beta}\Gamma_{\alpha\mu}^{\mu} + \Gamma_{\alpha\beta}^{\nu}\Gamma_{\mu\nu}^{\mu} - \Gamma_{\alpha\mu}^{\nu}\Gamma_{\beta\nu}^{\mu} \quad (14.444)$$

Nous avons alors (nous les calculons tous, même si nous savons que par la suite ceux qui n'ont pas $\alpha = \beta$ seront inutiles de par la métrique diagonale):

$$\begin{aligned}R_{11} &= R_{1\mu 1}^{\mu} = \partial_{\mu}\Gamma_{11}^{\mu} - \partial_1\Gamma_{1\mu}^{\mu} + \Gamma_{11}^{\nu}\Gamma_{\mu\nu}^{\mu} - \Gamma_{1\mu}^{\nu}\Gamma_{1\nu}^{\mu} \\ &= (\partial_1\Gamma_{11}^1 - \partial_1\Gamma_{11}^1 + \Gamma_{11}^r\Gamma_{1r}^1 - \Gamma_{11}^r\Gamma_{1r}^1) + (\partial_2\Gamma_{11}^2 - \partial_1\Gamma_{12}^2 + \Gamma_{11}^r\Gamma_{2r}^2 - \Gamma_{12}^r\Gamma_{1r}^2) \\ &= (\partial_2\Gamma_{11}^2 - \partial_1\Gamma_{12}^2 + \Gamma_{11}^r\Gamma_{2r}^2 - \Gamma_{12}^r\Gamma_{1r}^2) = (-\partial_1\Gamma_{12}^2 - \Gamma_{12}^r\Gamma_{1r}^2) \\ &= (-\partial_1\Gamma_{12}^2 - \Gamma_{12}^1\Gamma_{11}^2 - \Gamma_{12}^2\Gamma_{12}^2) = (-\partial_1\Gamma_{12}^2 - \Gamma_{12}^2\Gamma_{12}^2) \\ &= (-(-1 - \cot^2 \theta)) - \cot^2 \theta = 1\end{aligned}$$

$$\begin{aligned}R_{12} &= R_{1\mu 2}^{\mu} = \partial_{\mu}\Gamma_{12}^{\mu} - \partial_2\Gamma_{1\mu}^{\mu} + \Gamma_{12}^{\nu}\Gamma_{\mu\nu}^{\mu} - \Gamma_{1\mu}^{\nu}\Gamma_{2\nu}^{\mu} \\ &= (\partial_1\Gamma_{12}^1 - \partial_2\Gamma_{11}^1 + \Gamma_{12}^r\Gamma_{1r}^1 - \Gamma_{11}^r\Gamma_{2r}^1) + (\partial_2\Gamma_{12}^2 - \partial_2\Gamma_{12}^2 + \Gamma_{12}^r\Gamma_{2r}^2 - \Gamma_{12}^r\Gamma_{2r}^2) \\ &= (\partial_1\Gamma_{12}^1 - \partial_2\Gamma_{11}^1 + \Gamma_{12}^r\Gamma_{1r}^1 - \Gamma_{11}^r\Gamma_{2r}^1) = \Gamma_{12}^r\Gamma_{1r}^1 = \Gamma_{12}^1\Gamma_{11}^1 + \Gamma_{12}^2\Gamma_{12}^1 = 0\end{aligned}$$

$$\begin{aligned}R_{21} &= R_{2\mu 1}^{\mu} = \partial_{\mu}\Gamma_{21}^{\mu} - \partial_1\Gamma_{2\mu}^{\mu} + \Gamma_{21}^{\nu}\Gamma_{\mu\nu}^{\mu} - \Gamma_{2\mu}^{\nu}\Gamma_{1\nu}^{\mu} \\ &= (\partial_1\Gamma_{21}^1 - \partial_1\Gamma_{21}^1 + \Gamma_{21}^r\Gamma_{1r}^1 - \Gamma_{21}^r\Gamma_{1r}^1) + (\partial_2\Gamma_{21}^2 - \partial_1\Gamma_{22}^2 + \Gamma_{21}^r\Gamma_{2r}^2 - \Gamma_{22}^r\Gamma_{1r}^2) \\ &= (\partial_1\Gamma_{21}^1 - \partial_1\Gamma_{21}^1 + \Gamma_{21}^r\Gamma_{1r}^1 - \Gamma_{21}^r\Gamma_{1r}^1) \quad (14.445) \\ &= (\Gamma_{21}^r\Gamma_{1r}^1 - \Gamma_{21}^r\Gamma_{1r}^1) = (\Gamma_{21}^1\Gamma_{11}^1 + \Gamma_{21}^2\Gamma_{12}^1 - \Gamma_{21}^r\Gamma_{1r}^1) \\ &= (-\Gamma_{21}^r\Gamma_{1r}^1) = -(\Gamma_{21}^1\Gamma_{11}^1 + \Gamma_{21}^2\Gamma_{12}^1) = 0\end{aligned}$$

$$\begin{aligned}
R_{22} &= R_{2\mu 2}^{\mu} = \partial_{\mu} \Gamma_{22}^{\mu} - \partial_2 \Gamma_{2\mu}^{\mu} + \Gamma_{22}^{\nu} \Gamma_{\mu\nu}^{\mu} - \Gamma_{2\mu}^{\nu} \Gamma_{2\nu}^{\mu} \\
&= \left(\partial_1 \Gamma_{22}^1 - \partial_2 \Gamma_{21}^1 + \Gamma_{22}^r \Gamma_{1r}^1 - \Gamma_{21}^r \Gamma_{2r}^1 \right) + \left(\partial_2 \Gamma_{22}^2 - \partial_2 \Gamma_{22}^2 + \Gamma_{22}^r \Gamma_{2r}^2 - \Gamma_{22}^r \Gamma_{2r}^2 \right) \\
&= \left(\partial_1 \Gamma_{22}^1 - \partial_2 \Gamma_{21}^1 + \Gamma_{22}^r \Gamma_{1r}^1 - \Gamma_{21}^r \Gamma_{2r}^1 \right) \\
&= \left(\partial_1 \Gamma_{22}^1 + \Gamma_{22}^r \Gamma_{1r}^1 - \Gamma_{21}^r \Gamma_{2r}^1 \right) = \left(\partial_1 \Gamma_{22}^1 + \Gamma_{22}^1 \Gamma_{11}^1 + \Gamma_{22}^2 \Gamma_{12}^1 - \Gamma_{21}^1 \Gamma_{21}^1 - \Gamma_{21}^2 \Gamma_{22}^1 \right) \\
&= \left(\partial_1 \Gamma_{22}^1 - \Gamma_{21}^2 \Gamma_{22}^1 \right) = \left(\sin^2 \theta - \cos^2 \theta \right) - \cot \theta \cdot (-\sin \theta \cos \theta) \\
&= \sin^2 \theta - \cos^2 \theta + \cot \theta \sin \theta \cos \theta = \sin^2 \theta - \cos^2 \theta + \cos \theta \cos \theta = \sin^2 \theta
\end{aligned}$$

Calculons maintenant le scalaire de Ricci:

$$R = g^{\alpha\beta} R_{\alpha\beta} \quad (14.446)$$

Nous avons alors:

$$\begin{aligned}
R &= g^{\alpha\beta} R_{\alpha\beta} = g^{11} R_{11} + g^{22} R_{22} = \frac{1}{r^2} \cdot R_{11} + \frac{1}{r^2 \sin^2 \theta} R_{22} \\
&= \frac{1}{r^2} \cdot 1 + \frac{1}{r^2 \sin^2 \theta} \cdot \sin^2 \theta = \frac{1}{r^2} + \frac{1}{r^2} = \frac{2}{r^2}
\end{aligned} \quad (14.447)$$

Nous constatons que:

1. Le scalaire de Ricci est une constante. Cela signifie que l'hypersurface possède une courbure constante en tous points de la surface (nous savons que la sphère par symétrie possède une courbure constante en tous points). Elle possède donc une forme de symétrie, vis-à-vis de sa courbure. Nous avons alors affaire à une **"variété maximalement symétrique"**.

2. Ce scalaire est positif ce qui décrit un espace bombé (boule, sphère)

Remarque: Attention à ne pas confondre la valeur de la courbure de Ricci, et celle de la courbure de Gauss.

14. TENSEUR D'EINSTEIN

Appliquons une contraction à l'identité de Bianchi:

$$g^{\alpha\mu} \left(\nabla_{\lambda} R_{\alpha\beta, \mu\nu} + \nabla_{\nu} R_{\alpha\beta, \lambda\mu} + \nabla_{\mu} R_{\alpha\beta, \nu\lambda} \right) = 0 \quad (14.448)$$

Rappelons que $\nabla_{\lambda} g_{\alpha\mu} = 0$ et de même par extension que $\nabla_{\lambda} g^{\alpha\mu} = 0$. Donc finalement ceci nous amène à écrire de par la propriété des dérivées (produit en somme):

$$\nabla_{\lambda} g^{\alpha\mu} R_{\alpha\beta, \mu\nu} + \nabla_{\nu} g^{\alpha\mu} R_{\alpha\beta, \lambda\mu} + \nabla_{\mu} g^{\alpha\mu} R_{\alpha\beta, \nu\lambda} = 0 \quad (14.449)$$

et donc à obtenir:

$$\nabla_{\lambda} R_{\beta, \mu\nu}^{\mu} + \nabla_{\nu} R_{\beta, \lambda\mu}^{\mu} + \nabla_{\mu} R_{\beta, \nu\lambda}^{\mu} = 0 \quad (14.450)$$

En utilisant la propriété d'antisymétrie du tenseur de Riemann-Christoffel, nous écrivons:

$$\nabla_{\lambda} R^{\mu}_{\rho, \mu\nu} - \nabla_{\nu} R^{\mu}_{\rho, \mu\lambda} + \nabla_{\mu} R^{\mu}_{\rho, \nu\lambda} = \nabla_{\lambda} R_{\rho\nu} - \nabla_{\nu} R^{\mu}_{\rho, \mu\lambda} + \nabla_{\mu} R^{\mu}_{\rho, \nu\lambda} = 0 \quad (14.451)$$

Ce qui revient finalement à écrire de par la définition du tenseur de Ricci:

$$\nabla_{\lambda} R_{\rho\nu} - \nabla_{\nu} R_{\rho\lambda} + \nabla_{\mu} R^{\mu}_{\rho, \nu\lambda} = 0 \quad (14.452)$$

Cette dernière relation étant appelée "**identité de Bianchi contractée**".

Contractons cette relation encore une fois:

$$\begin{aligned} g^{\rho\nu} (\nabla_{\lambda} R_{\rho\nu} - \nabla_{\nu} R_{\rho\lambda} + \nabla_{\mu} R^{\mu}_{\rho, \nu\lambda}) &= \nabla_{\lambda} R^{\nu}_{\nu} - \nabla_{\nu} R^{\nu}_{\lambda} + \nabla_{\mu} R^{\mu\nu}_{\nu\lambda} = \nabla_{\lambda} R^{\nu}_{\nu} - \nabla_{\nu} R^{\nu}_{\lambda} - \nabla_{\mu} R^{\mu\nu}_{\lambda\nu} \\ &= \nabla_{\lambda} R - \nabla_{\nu} R^{\nu}_{\lambda} - \nabla_{\mu} R^{\mu}_{\lambda} = 0 \end{aligned} \quad (14.453)$$

Ce qui revient identiquement à écrire en utilisant les propriétés de la sommation d'Einstein (qui permet librement de changer les indices):

$$\nabla_{\lambda} R - \nabla_{\mu} R^{\mu}_{\lambda} - \nabla_{\mu} R^{\mu}_{\lambda} = \nabla_{\lambda} R - 2\nabla_{\mu} R^{\mu}_{\lambda} = 0 \quad (14.454)$$

Ce qui équivaut à:

$$\nabla_{\mu} R^{\mu}_{\lambda} - \frac{1}{2} \nabla_{\lambda} R = 0 \quad (14.455)$$

Comme $\nabla_{\lambda} R = g^{\mu\nu} \nabla_{\mu} R_{\nu\lambda}$, nous avons:

$$\nabla_{\mu} \left(R^{\mu}_{\lambda} - \frac{1}{2} g^{\mu}_{\lambda} R \right) = 0 \quad (14.456)$$

En montant l'indice λ par multiplication avec $g^{\nu\lambda}$, nous obtenons "**l'identité d'Einstein**":

$$\nabla_{\mu} \left(R^{\mu\nu} - \frac{1}{2} g^{\mu\nu} R \right) = 0 \quad (14.457)$$

Le "**tenseur d'Einstein**" (tenseur d'ordre deux et contravariant dans le cas présent) qui est donc une constante dans un espace Riemannien donné est dès lors défini par:

$$\boxed{G^{\mu\nu} = R^{\mu\nu} - \frac{1}{2} g^{\mu\nu} R} \quad (14.458)$$

et exprime de la façon la plus courte qui soit, le transport parallèle.

Identiquement, nous pouvons obtenir la forme covariante:

$$\boxed{G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R} \quad (14.459)$$

Le tenseur est donc construit pour une métrique uniquement Riemannienne (ce qui fait cependant quand même pas mal d'espaces possibles...), et est automatiquement non divergent:

$$\nabla_{\mu} G_{\mu\nu} = 0 \quad (14.460)$$

Nous retrouverons ce tenseur naturellement dans le chapitre de Relativité Générale lorsqu'en faisant usage du principe variationnel nous décomposerons l'action en deux termes:

1. l'action de la masse dans le champ gravitationnel
2. l'action du champ gravitationnel en l'absence de masse

En exprimant le tout dans un espace Riemannien nous obtiendrons alors la non moins fameuse équation d'Einstein des champs (sans plus d'explications dans ce chapitre):

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu} \quad (14.461)$$

les détails étant donnés dans le chapitre de Relativité Générale.

Remarque: Comme nous le voyons, nous pouvons très bien rajouter un terme constant à l'expression du tenseur d'Einstein, sans que cela ne change la nullité de sa divergence. Ce fait, utilisé en astrophysique, permet de construire des modèles d'Univers particuliers que nous traitons dans le chapitre d'Astrophysique.

Exemple:

Calculons le tenseur d'ordre 2 covariant d'Einstein:

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R \quad (14.462)$$

basé sur la métrique diagonale de surface de la 2-sphère S^2 (sans la composante temporelle):

$$g_{\mu\nu} = \begin{bmatrix} r^2 & 0 \\ 0 & r^2 \sin^2 \theta \end{bmatrix} \quad (14.463)$$

Comme la métrique est diagonale, nous avons bien démontré plus haut par l'exemple que:

$$R_{12} = R_{21} = 0 \quad (14.464)$$

et comme dans le cas présent, nous avons aussi:

$$g_{12} = g_{21} = 0 \quad (14.465)$$

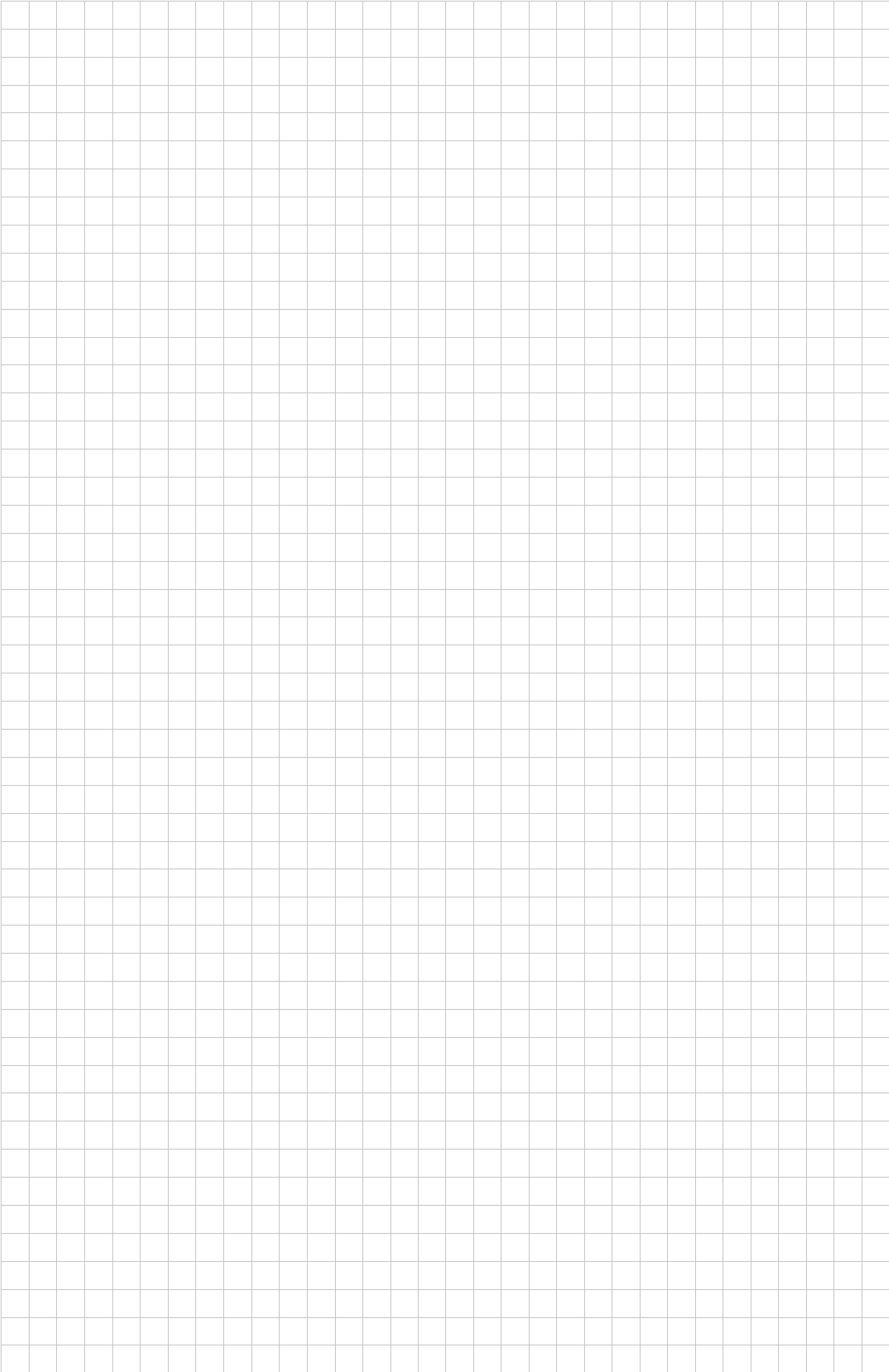
Il vient que:

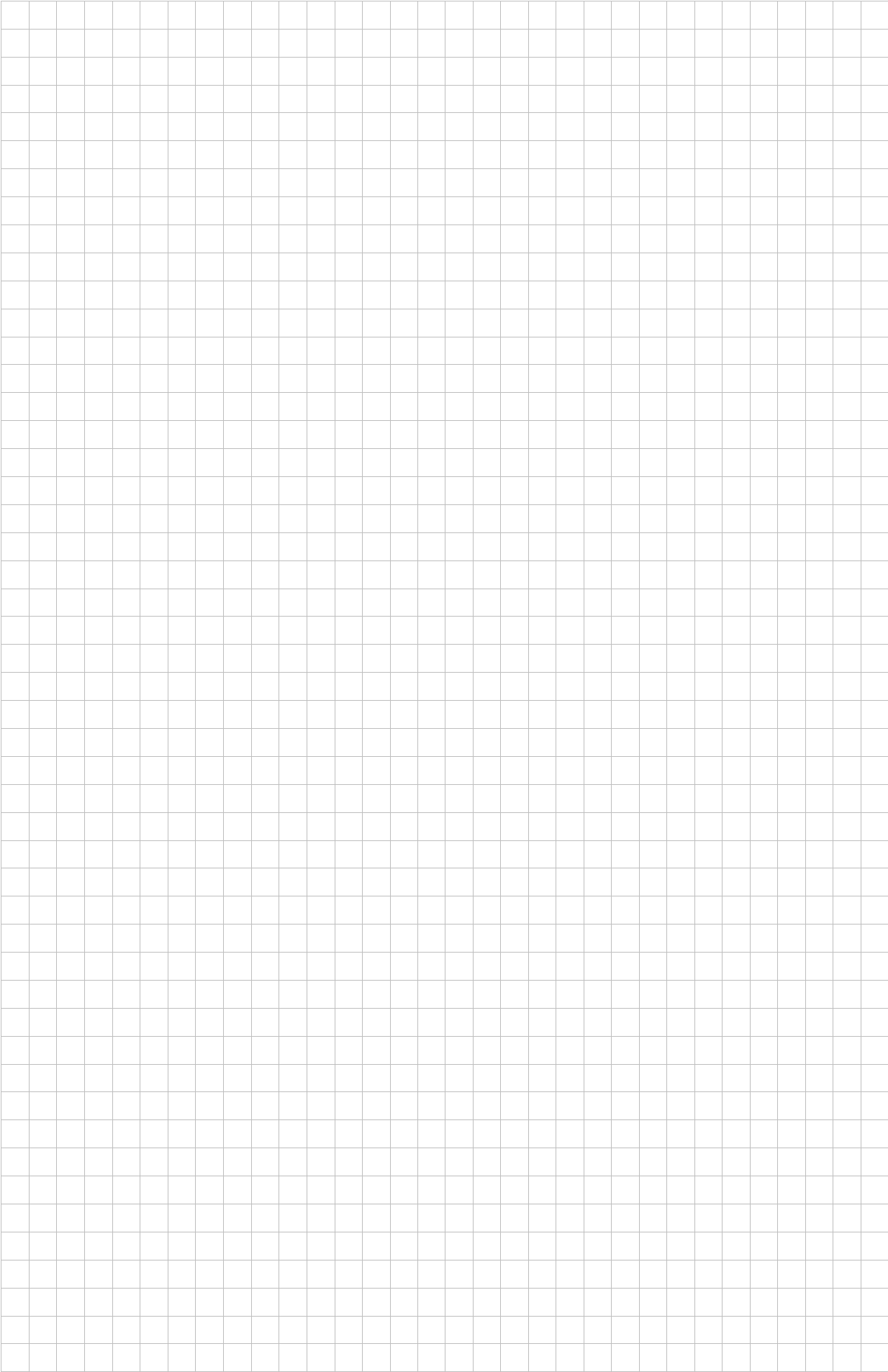
$$G_{12} = G_{21} = 0 \quad (14.466)$$

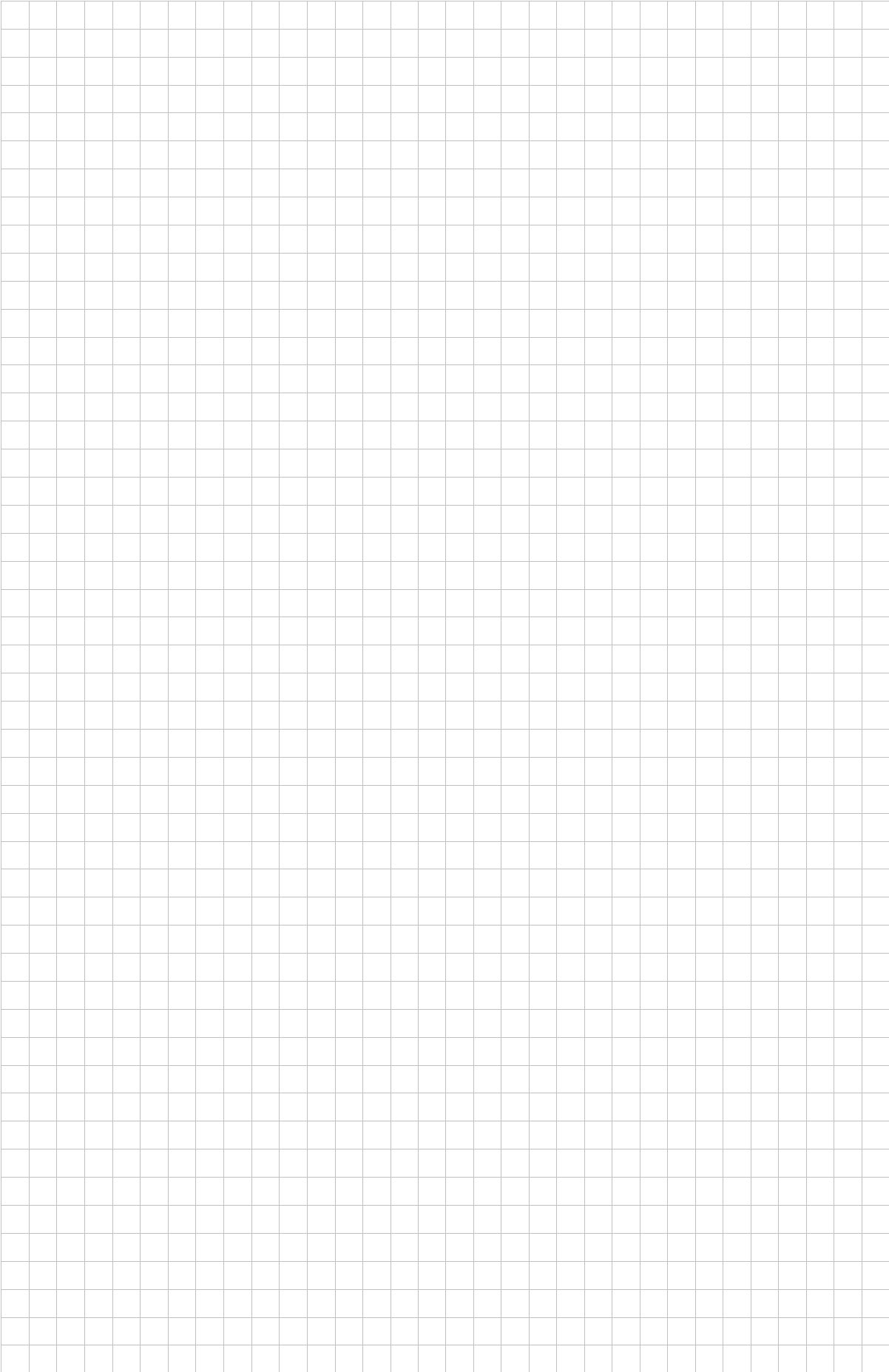
Donc nous n'avons qu'à nous concentrer sur deux composantes:

$$\begin{aligned} G_{11} &= R_{11} - \frac{1}{2} g_{11} R = 1 - \frac{1}{2} r^2 \frac{2}{r^2} = 1 - \frac{1}{2} r^2 \frac{2}{r^2} = 0 \\ G_{22} &= R_{22} - \frac{1}{2} g_{22} \frac{2}{r^2} = \sin^2 \theta - \frac{1}{2} r^2 \sin^2 \theta \frac{2}{r^2} = 0 \end{aligned} \quad (14.467)$$

ce qui vérifie bien ce que nous avons dit plus haut.







15. CALCUL SPINORIEL

Comme nous le verrons en premier en physique quantique relativiste, les spineurs jouent un rôle majeur dans la théorie quantique et en conséquence dans toute la physique contemporaine (théorie quantique des champs, modèle standard, théorie des cordes,...).

Ce fut à partir de 1927 que les physiciens Pauli, puis Dirac introduisirent les spineurs pour la représentation des fonctions d'onde (cf. [chapitre de Physique Quantique Relativiste](#)). Cependant, sous leur forme mathématique, les spineurs avaient été découverts par Élie Cartan dès 1913 lors de ses recherches sur les représentations des groupes en faisant suite à la théorie générale des espaces de Clifford (introduits par le mathématicien W.K. Clifford en 1876). Il montra, comme nous le verrons, que les spineurs fournissent au fait une représentation linéaire du groupe des rotations d'un espace à un nombre quelconque de dimensions. Ainsi, les spineurs sont donc étroitement liés à la géométrie mais leur présentation est souvent faite de manière abstraite sans signification géométrique intuitive. Ainsi, nous allons nous efforcer (comme toujours sur ce site) dans ce chapitre d'introduire de la manière la plus simple et intuitive possible les théories des spineurs.

Le formalisme spinoriel n'intéresse pas seulement la physique quantique et ses travaux, entre autres, de Roger Penrose ont montré que la théorie spinorielle était une approche extrêmement féconde de la théorie de la relativité générale. Bien que le plus couramment utilisé pour le traitement de la relativité générale soit le calcul tensoriel, Penrose a montré que dans le cas spécifique de l'espace à quatre dimensions et la métrique de Lorentz, le formalisme des spineurs à deux composantes est plus approprié.

La théorie des spineurs ou "[géométrie spinorielle](#)" est extrêmement vaste mais ce site ayant plus pour objectif de s'adresser aux physiciens, nous nous limiterons aux spineurs utiles en physique quantique ainsi que leurs propriétés y relatives.

Remarque: Nous conseillons vivement au lecteur d'avoir lu au préalable le sous-chapitre sur les quaternions (cf. [chapitre Nombres](#)), le sous-chapitre sur les rotations dans l'espace (cf. [chapitre Géométrie Euclidienne](#)) et enfin, si possible pour avoir un exemple pratique physique, le chapitre de physique quantique relativiste.

1. SPINEUR UNITAIRE

Nous allons donner ici une première définition (ou exemple) particulière simplifiée des spineurs. Ainsi, nous allons montrer qu'il est possible à partir d'un tel outil de représenter un vecteur d'un espace E_3 à trois composantes à l'aide d'un spineur à deux composantes. La méthode est extrêmement simple et celui qui a déjà lu la partie du chapitre de Physique Quantique Ondulatoire traitant de l'équation de Dirac ainsi que le chapitre d'Informatique Quantique y verra une analogie assez grandiose.

Considérons pour commencer la sphère suivante d'équation (cf. [chapitre de Géométrie Analytique](#))

$$x^2 + y^2 + z^2 = 1 \quad (15.1)$$

Et considérons le schéma suivant:

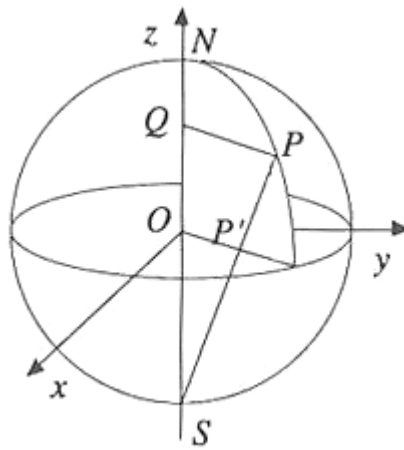


Figure: 15.1 - Sphère unitaire

Considérons-y les coordonnées (x, y, z) d'un point P de la sphère centrée en O et de rayon unité et notons N et S les points d'intersection de l'axe Oz avec la sphère.

Le point S aura par convention pour coordonnées:

$$(x = 0, y = 0, z = -1) \quad (15.2)$$

Nous obtenons une projection dite "**projection stéréographique**" P' du point P en traçant la droite SP qui traverse un plan équatorial xOy complexe $\mathbb{C}$ au point P' de coordonnées (x', y', z') .

Les triangles semblables $SP'O$ et SPQ (avec Q étant la projection orthogonale sur l'axe Oz du point P) nous donnent les relations suivantes en appliquant simplement le théorème de Thalès:

$$\frac{SP'}{SP} = \frac{x'}{x} = \frac{y'}{y} = \frac{1}{1+z} = \frac{x'+iy'}{x+iy} = \frac{x'-iy'}{x-iy} \quad (15.3)$$

Remarque: Les deux dernières relations s'obtiennent par application du théorème de Thalès (cf. chapitre de Géométrie) dans le plan équatorial complexe.

Posons maintenant:

$$\xi = x' + iy', \bar{\xi} = x' - iy' \quad (15.4)$$

Il vient, compte tenu de la relation précédente que:

$$\frac{1}{1+z} = \frac{x'+iy'}{x+iy} \quad (15.5)$$

en prenant le module au carré (voir l'étude des nombres complexes dans le chapitres des Nombres):

$$\frac{1}{|1+z|^2} = \frac{1}{(1+z)^2} = \frac{|x'+iy'|^2}{|x+iy|^2} = \frac{\xi \bar{\xi}}{x^2 + y^2} \quad (15.6)$$

et comme de l'équation de la sphère il découle:

$$x^2 + y^2 = 1 - z^2 = (1+z)(1-z) \quad (15.7)$$

nous avons finalement:

$$\frac{1}{(1+z)^2} = \frac{\xi \bar{\xi}}{x^2 + y^2} = \frac{\xi \bar{\xi}}{(1+z)(1-z)} \quad (15.8)$$

Mettons maintenant le nombre complexe ξ sous la forme $\xi = \phi / \psi$ où ϕ, ψ sont deux nombres complexes auxquels nous pouvons toujours imposer de vérifier la condition d'unitarité (rien ne nous l'interdit mais en physique cela nous arrange bien):

$$\boxed{\psi \bar{\psi} + \phi \bar{\phi} = 1} \quad (15.9)$$

Remarque: Les nombres complexes suivants satisfont donc la relation précédente:

$$\psi = \cos \alpha e^{i\phi}, \phi = \sin \alpha e^{i\gamma} \quad (15.10)$$

Rappelons avant de continuer que nous avons démontré lors de notre étude des nombres complexes que:

$$\xi = \frac{\phi}{\psi}, \bar{\xi} = \overline{\left(\frac{\phi}{\psi}\right)} = \frac{\bar{\phi}}{\bar{\psi}} \quad (15.11)$$

Dès lors il vient en injectant ces deux dernières relations dans l'équation déterminée plus haut:

$$\begin{aligned} \frac{1}{(1+z)^2} &= \frac{\frac{\phi}{\psi} \frac{\bar{\phi}}{\bar{\psi}}}{(1+z)(1-z)} \Rightarrow \bar{\psi} \psi = \bar{\phi} \phi \frac{(1+z)}{(1-z)} \Rightarrow \bar{\psi} \psi (1-z) = \bar{\phi} \phi (1+z) \\ \Rightarrow \bar{\psi} \psi - z \bar{\psi} \psi &= \bar{\phi} \phi + z \bar{\phi} \phi \Rightarrow \bar{\psi} \psi - \bar{\phi} \phi = z (\bar{\phi} \phi + \bar{\psi} \psi) = z \end{aligned} \quad (15.12)$$

d'où finalement la coordonnée verticale du point P:

$$z = \bar{\psi} \psi - \bar{\phi} \phi \quad (15.13)$$

Comme nous avons:

$$x + iy = \xi(1+z), x - iy = \bar{\xi}(1+z) \quad (15.14)$$

alors:

$$x = \frac{1}{2}(\xi + \bar{\xi})(1+z), y = -\frac{i}{2}(\xi - \bar{\xi})(1+z) \quad (15.15)$$

tenant compte des derniers développements nous avons finalement:

$$x = \psi \bar{\phi} + \bar{\psi} \phi, y = i(\psi \bar{\phi} - \bar{\psi} \phi) \quad (15.16)$$

Ainsi, à tout point P situé sur la sphère de rayon unité, nous pouvons faire correspondre un couple de nombre complexes vérifiant la relation d'unitarité imposée.

Soit sous forme complète et explicite nous avons finalement:

$$\begin{aligned}
 \overrightarrow{OP} &= \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \psi\bar{\phi} + \bar{\psi}\phi \\ i(\psi\bar{\phi} - \bar{\psi}\phi) \\ \bar{\psi}\psi - \bar{\phi}\phi \end{pmatrix} = \begin{pmatrix} \cos \alpha \sin \alpha e^{i(\beta-\gamma)} + \cos \alpha \sin \alpha e^{i(\gamma-\beta)} \\ i(\cos \alpha \sin \alpha e^{i(\beta-\gamma)} - \cos \alpha \sin \alpha e^{i(\gamma-\beta)}) \\ \cos^2 \alpha - \sin^2 \beta \end{pmatrix} \\
 &= \begin{pmatrix} \cos \alpha \sin \alpha (e^{i(\beta-\gamma)} + e^{-i(\beta-\gamma)}) \\ i \cos \alpha \sin \alpha (e^{i(\beta-\gamma)} - e^{-i(\beta-\gamma)}) \\ \cos^2 \alpha - \sin^2 \beta \end{pmatrix} \\
 &= \begin{pmatrix} \cos \alpha \sin \alpha (\cos(\beta-\gamma) + i \sin(\beta-\gamma) + \cos(-(\beta-\gamma)) + i \sin(-(\beta-\gamma))) \\ i(\cos \alpha \sin \alpha (\cos(\beta-\gamma) + i \sin(\beta-\gamma) - \cos(-(\beta-\gamma)) - i \sin(-(\beta-\gamma)))) \\ \cos^2 \alpha - \sin^2 \beta \end{pmatrix} \quad (15.17) \\
 &= \begin{pmatrix} \cos \alpha \sin \alpha (2 \cos(\beta-\gamma) + i \sin(\beta-\gamma) - i \sin(\beta-\gamma)) \\ i(\cos \alpha \sin \alpha (i \sin(\beta-\gamma) + i \sin(\beta-\gamma))) \\ \cos^2 \alpha - \sin^2 \beta \end{pmatrix} \\
 &= \begin{pmatrix} 2 \cos \alpha \sin \alpha \cos(\beta-\gamma) \\ -(2 \cos \alpha \sin \alpha \sin(\beta-\gamma)) \\ \cos^2 \alpha - \sin^2 \beta \end{pmatrix} = \begin{pmatrix} \sin(2\alpha) \cos(\beta-\gamma) \\ -\sin(2\alpha) \sin(\beta-\gamma) \\ \cos^2 \alpha - \sin^2 \beta \end{pmatrix} = \begin{pmatrix} \sin(2\alpha) \cos(\beta-\gamma) \\ -\sin(2\alpha) \sin(\beta-\gamma) \\ \cos(2\alpha) \end{pmatrix} \\
 &= \begin{pmatrix} \sin(2\alpha) \cos(\gamma-\beta) \\ \sin(2\alpha) \sin(\gamma-\beta) \\ \cos(2\alpha) \end{pmatrix}
 \end{aligned}$$

Cette dernière relation nous indique donc que 2α est l'angle entre Oz et $\overrightarrow{OP}$ (puisque l'hypoténuse de l'angle du vecteur à une norme unitaire) et donc par déduction $\gamma - \beta$ représente l'angle entre Ox et le plan (Oz, OP) :

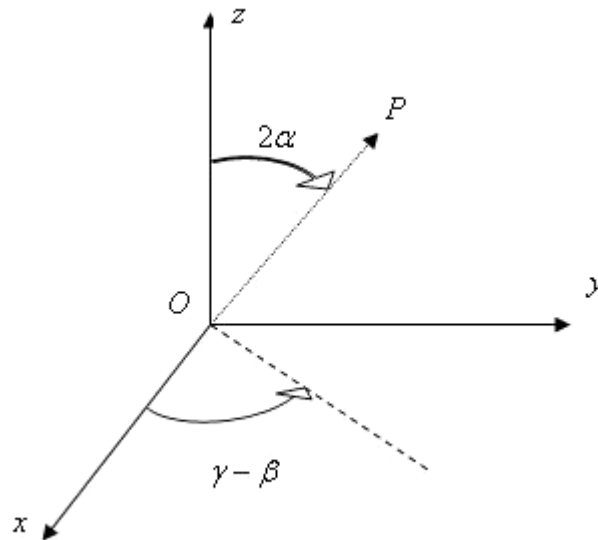


Figure: 15.2 - Représentation de la rotation

Le couple de nombres complexes de la relation antécédente constitue par définition un "spineur unitaire". Ainsi, comme nous l'avons vu, un spineur unitaire peut se mettre sous la forme:

$$\psi = \cos \alpha e^{i\phi}, \phi = \sin \alpha e^{i\gamma} \quad (15.18)$$

de même un spineur quelconque peut se mettre sous la forme un peu plus générale:

$$\psi = \rho \cos \alpha e^{i\phi}, \phi = \rho \sin \alpha e^{i\gamma} \quad (15.19)$$

Le spin ainsi mesuré l'est essentiellement à partir d'un axe orienté OZ comme nous venons de le voir avec la figure précédente.

La projection stéréographique conduit donc à représenter certains vecteurs de l'espace euclidien E_3 avec des éléments d'un espace vectoriel complexe de dimension deux qui est l'espace des spineurs.

Remarque: Cette représentation n'est pas unique car les arguments de nombres complexes ne sont (sous forme trigonométrique) déterminés qu'à une constante près.

Le lecteur qui aura déjà étudié un peu la physique quantique ondulatoire (voir chapitre du même nom) aura certainement remarqué l'étrange similarité non innocente de la condition et des relations:

$$\begin{aligned} \psi\bar{\psi} + \phi\bar{\phi} &= 1 \\ x = \psi\bar{\phi} + \bar{\psi}\phi, y &= i(\psi\bar{\phi} - \bar{\psi}\phi) \end{aligned} \quad (15.20)$$

par rapport à la condition de normalisation de de Broglie (l'intégrale sur tout l'espace de la somme des produits des fonctions d'ondes complexes conjuguées est égale à l'unité) et des développements déterminant l'équation de continuité en physique quantique ondulatoire.

Voyons maintenant pour les besoins ultérieurs, que nous pouvons trouver deux nouveaux vecteurs $\vec{v}_1, \vec{v}_2$ de l'espace euclidien E_3 , associés à un spineur unitaire (ψ, ϕ) déterminé sur la sphère unité. Ces vecteurs seront cherchés orthogonaux entre eux et de norme unité, chacun étant orthogonal au vecteur $\overrightarrow{OP}$.

Notons $\vec{v}_3 = \overrightarrow{OP}$ et $(x_i, y_i, z_i), i \in \{1, 2, 3\}$, les composantes respectives des vecteurs $\vec{v}_1, \vec{v}_2, \vec{v}_3$ sont bien sûr liés par le produit vectoriel:

$$\vec{v}_3 = \vec{v}_1 \times \vec{v}_2 \quad (15.21)$$

d'où tenant compte de l'expression des composantes $\overrightarrow{OP}$ en fonction de celles du spineur associé, ainsi que du fait $\psi\bar{\psi} + \phi\bar{\phi} = 1$, nous obtenons:

$$\begin{aligned} x_3 &= y_1 z_2 - y_2 z_1 = 1 \cdot (\psi\bar{\phi} + \bar{\psi}\phi) = (\psi\bar{\psi} + \phi\bar{\phi}) \cdot (\psi\bar{\phi} + \bar{\psi}\phi) \\ y_3 &= z_1 x_2 - z_2 x_1 = 1 \cdot i(\psi\bar{\phi} - \bar{\psi}\phi) = i(\psi\bar{\psi} + \phi\bar{\phi}) \cdot (\psi\bar{\phi} - \bar{\psi}\phi) \\ z_3 &= x_1 y_2 - x_2 y_1 = 1 \cdot (\psi\bar{\psi} - \phi\bar{\phi}) = (\psi\bar{\psi} + \phi\bar{\phi}) \cdot (\psi\bar{\psi} - \phi\bar{\phi}) \end{aligned} \quad (15.22)$$

Ecrivant l'orthogonalité des vecteurs entre eux nous obtenons bien évidemment six équations supplémentaires. Cependant l'orientation des vecteurs $\vec{v}_1, \vec{v}_2$ n'étant pas fixée, il existe une certaine indétermination sur les valeurs de leurs composantes. Choisissons des valeurs telles que:

$$x_1 + ix_2 = \psi^2 - \phi^2, y_1 + iy_2 = i(\psi^2 + \phi^2), z_1 + iz_2 = -2\psi\phi \quad (15.23)$$

Prenant les quantités complexes conjuguées des relations précédentes, nous obtenons par addition les composantes de $\vec{v}_1$:

$$x_1 = \frac{1}{2}(\psi^2 + \bar{\psi}^2 - \phi^2 - \bar{\phi}^2), y_1 = \frac{i}{2}(\psi^2 - \bar{\psi}^2 + \phi^2 - \bar{\phi}^2), z_1 = -(\psi\phi + \bar{\psi}\bar{\phi}) \quad (15.24)$$

Par soustraction, nous obtenons de même les composantes du vecteur $\vec{v}_2$:

$$x_2 = \frac{i}{2}(-\psi^2 + \bar{\psi}^2 + \phi^2 - \bar{\phi}^2), y_2 = \frac{1}{2}(\psi^2 + \bar{\psi}^2 + \phi^2 + \bar{\phi}^2), z_2 = i(\psi\phi - \bar{\psi}\bar{\phi}) \quad (15.25)$$

Nous vérifions aisément que ces valeurs redonnent bien les relations du produit vectoriel. A tout spineur unitaire (ψ, ϕ) nous pouvons donc associer trois vecteurs $\vec{v}_1, \vec{v}_2, \vec{v}_3$. Nous pouvons vérifier directement que les vecteurs ainsi calculés sont bien orthogonaux entre eux et de norme unité.

1.1. PROPRIÉTÉS GÉOMÉTRIQUES

Nous allons étudier les transformations des vecteurs de E_3 associés à un spineur afin d'en déduire les propriétés correspondantes de transformation du spineur. Les rotations dans l'espace pouvant toujours s'exprimer sous forme du produit de deux symétries planes (faire dans la tête l'expérience imaginaire), nous commençons par l'étude de ces dernières.

1.1.1. SYMÉTRIES PLANES

Considérons dans un premier temps la symétrie plane d'un vecteur:

Lors d'une symétrie par rapport à un plan P , un vecteur quelconque $\overrightarrow{OM}$ se transforme en un vecteur $\overrightarrow{OM'}$. Déterminons une matrice S qui représente cette symétrie par rapport à ce plan. Soit $\overrightarrow{OA}$ un vecteur unitaire normal au plan P et soit H le pied de la perpendiculaire abaissée d'un point M de l'espace sur le plan P .

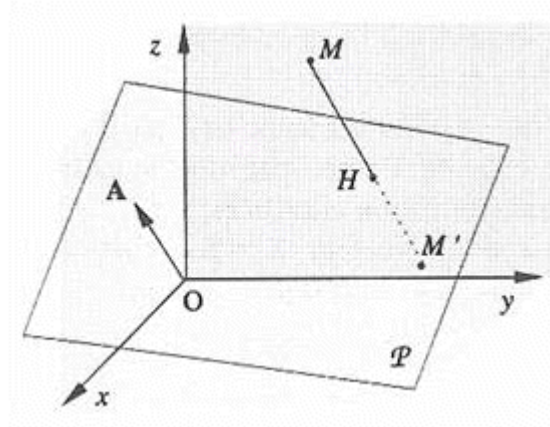


Figure: 15.3 - Symétrie par rapport à un plan

Soit M' le point symétrique de M par rapport à P , nous avons:

$$\overrightarrow{OM'} = \overrightarrow{OM} + \overrightarrow{MM'} = \overrightarrow{OM} + 2\overrightarrow{MH} = \overrightarrow{OM} - 2\overrightarrow{OA}(\overrightarrow{OM} \circ \overrightarrow{OA}) \quad (15.26)$$

Soient a_1, a_2, a_3 les composantes cartésiennes de $\overrightarrow{OA}$ et $(x, y, z), (x', y', z')$ les composantes respectives des vecteurs $\overrightarrow{OM}, \overrightarrow{OM'}$, la relation précédente nous donne les relations linéaires:

$$\begin{aligned} x' &= x - 2a_1(a_1x + a_2y + a_3z) \\ y' &= y - 2a_2(a_1x + a_2y + a_3z) \\ z' &= z - 2a_3(a_1x + a_2y + a_3z) \end{aligned} \quad (15.27)$$

La matrice S qui fait passer du vecteur $\vec{X}(x, y, z)$ au vecteur $\vec{X'}(x', y', z')$ a donc pour expression:

$$S = \begin{bmatrix} 1 - 2a_1^2 & -2a_1a_2 & -2a_1a_3 \\ -2a_2a_1 & 1 - 2a_2^2 & -2a_2a_3 \\ -2a_3a_1 & -2a_3a_2 & 1 - 2a_3^2 \end{bmatrix} \quad (15.28)$$

Gardons en mémoire ce résultat et considérons à présent deux vecteurs $(\vec{X}_1, \vec{X}_2)$, orthogonaux entre eux et unitaires, définissant comme nous l'avons vu un spineur unitaire (ψ, ϕ) . Une symétrie par rapport à un plan P transforme les vecteurs $\vec{X}_1, \vec{X}_2$ en vecteurs $\vec{X}'_1, \vec{X}'_2$ auxquels sont associés le spineur (ψ', ϕ') . Nous allons maintenant montrer que la transformation suivante du spineur (ψ, ϕ) en spineur $(\psi', \phi') = (x'_3, y'_3, z'_3)$ est:

$$\begin{bmatrix} \psi' \\ \phi' \end{bmatrix} = \begin{bmatrix} a_3 & a_1 - ia_2 \\ a_1 + ia_2 & -a_3 \end{bmatrix} \begin{bmatrix} \psi \\ \phi \end{bmatrix} = \mathcal{A} \begin{bmatrix} \psi \\ \phi \end{bmatrix} \quad (15.29)$$

et transforme précisément les vecteurs $\vec{X}_1, \vec{X}_2$ en vecteurs $\vec{X}'_1, \vec{X}'_2$, ces vecteurs se déduisant respectivement - comme nous allons le montrer - les uns des autres par une simple symétrie plane et que la matrice $\mathcal{A}$ représente bien la transformation cherchée.

La relation précédente nous donne donc:

$$\begin{aligned} \psi' &= a_3\psi + (a_1 - ia_2)\phi \\ \phi' &= (a_1 + ia_2)\psi - a_3\phi \end{aligned} \quad (15.30)$$

En tout nous avons:

$$\vec{X}'_3 = (\psi', \phi') \begin{cases} x'_1 + ix'_2 = \psi'^2 - \phi'^2 \\ y'_1 + iy'_2 = i(\psi'^2 + \phi'^2) \\ z'_1 + iz'_2 = -2\psi'\phi' \end{cases} \quad \begin{aligned} \psi' &= a_3\psi + (a_1 - ia_2)\phi \\ \phi' &= (a_1 + ia_2)\psi - a_3\phi \end{aligned} \quad (15.31)$$

Nous en déduisons:

$$\begin{aligned}
x'_3 &= \psi'^2 - \phi'^2 = (a_3\psi + (a_1 - ia_2)\phi)^2 - ((a_1 + ia_2)\psi - a_3\phi)^2 \\
&= (a_3^2\psi^2 + 2a_3\psi(a_1 - ia_2)\phi + (a_1 - ia_2)^2\phi^2) - ((a_1 + ia_2)^2\psi^2 - 2(a_1 + ia_2)\psi a_3\phi + a_3^2\phi^2) \\
&= (a_3^2\psi^2 + 2a_3\psi(a_1 - ia_2)\phi + (a_1^2 - 2a_1ia_2 - a_2^2)\phi^2) \\
&\quad - ((a_1^2 + 2a_1ia_2 - a_2^2)\psi^2 - 2(a_1 + ia_2)\psi a_3\phi + a_3^2\phi^2) \\
&= a_3^2\psi^2 + 2a_1a_3\psi\phi - 2ia_2a_3\psi\phi + a_1^2\phi^2 - 2a_1ia_2\phi^2 - a_2^2\phi^2 \\
&\quad - a_1^2\psi^2 - 2a_1ia_2\psi^2 + a_2^2\psi^2 + 2a_1a_3\psi\phi + 2ia_2a_3\psi\phi - a_3^2\phi^2 \\
&= [-a_1^2 + a_2^2 + a_3^2](\psi^2 - \phi^2) - 2a_1a_2i(\psi^2 + \phi^2) - 2a_1a_3(-2\psi\phi) \\
y'_3 &= i(\psi'^2 + \phi'^2) = i[(a_3\psi + (a_1 - ia_2)\phi)^2 + ((a_1 + ia_2)\psi - a_3\phi)^2] \\
&= i[(a_3^2\psi^2 + 2a_3\psi(a_1 - ia_2)\phi + (a_1 - ia_2)^2\phi^2) + ((a_1 + ia_2)^2\psi^2 - 2(a_1 + ia_2)\psi a_3\phi + a_3^2\phi^2)] \quad (15.32) \\
&= i(a_3^2\psi^2 + 2a_3\psi(a_1 - ia_2)\phi + (a_1^2 - 2a_1ia_2 - a_2^2)\phi^2) \\
&\quad + i((a_1^2 + 2a_1ia_2 - a_2^2)\psi^2 - 2(a_1 + ia_2)\psi a_3\phi + a_3^2\phi^2) \\
&= ia_3^2\psi^2 + 2ia_1a_3\psi\phi + 2a_2a_3\psi\phi + ia_1^2\phi^2 + 2a_1a_2\phi^2 - ia_2^2\phi^2 \\
&\quad + ia_1^2\psi^2 - 2a_1a_2\psi^2 - ia_2^2\psi^2 - 2a_1a_3\psi\phi + 2a_2a_3\psi\phi + ia_3^2\phi^2 \\
&= -2a_2a_1(\psi^2 - \phi^2) - 2a_2a_3(-2\psi\phi) + [a_1^2 - a_2^2 + a_3^2]i(\psi^2 + \phi^2) \\
z'_3 &= -2\psi'\phi' = -2(a_3\psi + (a_1 - ia_2)\phi)((a_1 + ia_2)\psi - a_3\phi) \\
&= -2(a_3\psi + a_1\phi - ia_2\phi)(a_1\psi + ia_2\psi - a_3\phi) \\
&= -2a_3a_1(\psi^2 - \phi^2) - 2a_2a_3i(\psi^2 + \phi^2) + [a_1^2 + a_2^2 - a_3^2](-2\psi\phi)
\end{aligned}$$

Par suite du fait que $\|\vec{A}\|^2 = a_1^2 + a_2^2 + a_3^2 = 1$, nous obtenons:

$$\begin{aligned}
x'_3 &= [1 - 2a_1^2]x_3 - 2a_1a_2y_3 - 2a_1a_3z_3 \\
y'_3 &= -2a_1a_2x_3 + [1 - 2a_2^2]y_3 - 2a_2a_3z_3 \quad (15.33) \\
z'_3 &= -2a_3a_1x_3 - 2a_3a_2y_3 + [1 - 2a_3^2]z_3
\end{aligned}$$

Nous retombons donc bien sur la matrice de symétrie:

$$S = \begin{bmatrix} 1 - 2a_1^2 & -2a_1a_2 & -2a_1a_3 \\ -2a_2a_1 & 1 - 2a_2^2 & -2a_2a_3 \\ -2a_3a_1 & -2a_3a_2 & 1 - 2a_3^2 \end{bmatrix} \quad (15.34)$$

Ainsi, la matrice que nous retrouverons dans le chapitre de Physique Quantique Relativiste:

$$\mathcal{A} = \begin{bmatrix} a_3 & a_1 - ia_2 \\ a_1 + ia_2 & -a_3 \end{bmatrix} \quad (15.35)$$

$\mathcal{A}$ engendre donc la transformation d'un spineur (ψ, ϕ) en un spineur (ψ', ϕ') telle que les vecteurs associées $(\vec{X}_1, \vec{X}_2)$ se déduisent respectivement de $(\vec{X}'_1, \vec{X}'_2)$ par une symétrie plane.

1.1.2. ROTATIONS

Comme nous l'avons vu dans le chapitre de géométrie euclidienne, il est possible faire une rotation d'un vecteur dans le plan ou dans l'espace à l'aide de matrices. De même, par extension, il est évident que la multiplication de deux rotations est une rotation (c'est de l'algèbre linéaire élémentaire - du moins nous le considérons tel quel).

Considérons dès lors, deux plans P, Q dont l'intersection engendre une droite (ligne) L et notons $\vec{A}(a_1, a_2, a_3)$ et $\vec{B}(b_1, b_2, b_3)$ des vecteurs unitaires portés par les normales respectives à ces deux plans sécantes en L :

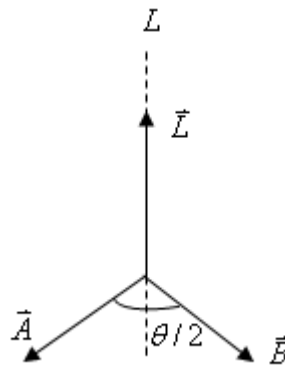


Figure: 15.4 - Intersection imagée de deux plans

Notons $\theta/2$ l'angle des vecteurs $\vec{A}, \vec{B}$ entre eux (la raison de cette notation provient de notre étude des quaternions (cf. chapitre Nombres)). Soit $\vec{L}$ le vecteur unitaire porté par la droite L résultant de l'intersection des plans P, Q et tel que:

$$\vec{A} \times \vec{B} = \vec{L} \sin\left(\frac{\theta}{2}\right) \quad (15.36)$$

Explications: $\vec{A}, \vec{B}$ sont unitaires mais pas nécessairement perpendiculaires et nous devons quand même nous assurer que $\vec{L}$ soit un vecteur unitaire (sa norme soit égale à l'unité donc). Dès lors, la relation ci-dessus nous assure que:

$$\|\vec{A}\| \|\vec{B}\| \sin\left(\frac{\theta}{2}\right) = \|\vec{L}\| \sin\left(\frac{\theta}{2}\right) \Rightarrow \|\vec{A}\| \|\vec{B}\| = \|\vec{L}\| = L_1^2 + L_2^2 + L_3^2 = 1 \quad (15.37)$$

Le produit vectoriel précédent nous donne pour les composantes de $\vec{L}$:

$$\begin{aligned}
L_1 \sin\left(\frac{\theta}{2}\right) &= a_2 b_3 - a_3 b_2 \\
L_2 \sin\left(\frac{\theta}{2}\right) &= a_3 b_1 - a_1 b_3 \\
L_3 \sin\left(\frac{\theta}{2}\right) &= a_1 b_2 - a_2 b_1
\end{aligned} \quad (15.38)$$

D'autre part, le produit scalaire s'écrit:

$$\vec{A} \circ \vec{B} = a_1 b_1 + a_2 b_2 + a_3 b_3 = \|\vec{A}\| \|\vec{B}\| \cos\left(\frac{\theta}{2}\right) = \cos\left(\frac{\theta}{2}\right) \quad (15.39)$$

Remarque: Nous allons nous servir des ces deux plans comme plans de symétrie pour nos rotations

Comme nous l'avons fait remarquer précédemment, une rotation dans E_3 peut toujours se faire avec au plus deux symétries planes. Ainsi, une rotation peut se noter par l'application (multiplication) de deux matrices de symétrie selon les résultats obtenus plus haut:

$$R\left(\vec{L}, \frac{\theta}{2}\right) = \begin{bmatrix} b_3 & b_1 - ib_2 \\ b_1 + ib_2 & -b_3 \end{bmatrix} \begin{bmatrix} a_3 & a_1 - ia_2 \\ a_1 + ia_2 & -a_3 \end{bmatrix} \quad (15.40)$$

Développant le produit de ces deux matrices et tenant compte de relations découlant du produit vectoriel et scalaire nous obtenons:

$$R\left(\vec{L}, \frac{\theta}{2}\right) = \begin{bmatrix} \cos\left(\frac{\theta}{2}\right) - iL_3 \sin\left(\frac{\theta}{2}\right) & -(iL_1 + L_2) \sin\left(\frac{\theta}{2}\right) \\ (-iL_1 + L_2) \sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) + iL_3 \sin\left(\frac{\theta}{2}\right) \end{bmatrix} \quad (15.41)$$

Ainsi, nous pouvons écrire la transformation d'un spineur (ψ, ϕ) et un spineur (ψ', ϕ') à l'aide d'une matrice de la forme:

$$R = \begin{bmatrix} a & b \\ -\bar{b} & \bar{a} \end{bmatrix} \quad (15.42)$$

dont les paramètres sont appelés "**paramètres de Cayley-Klein**".

La matrice $R\left(\vec{L}, \frac{\theta}{2}\right)$ peut être écrite sous une autre forme si nous faisons un développement limité pour des rotations infiniment petites $\theta/2 = \varepsilon/2$ (eh voilà la physique qui revient....). Ainsi, les développements de Maclaurin (cf. chapitre Suites Et Séries) nous donnent:

$$\cos\left(\frac{\varepsilon}{2}\right) = 1 - \frac{1}{2}\left(\frac{\varepsilon}{2}\right)^2 + \dots \quad \sin\left(\frac{\varepsilon}{2}\right) = \frac{\varepsilon}{2} - \frac{1}{6}\left(\frac{\varepsilon}{2}\right)^3 + \dots \quad (15.43)$$

En utilisant seuls les termes du premier ordre, la matrice de rotations s'écrit finalement:

$$R\left(\vec{L}, \frac{\varepsilon}{2}\right) = \begin{bmatrix} 1 - iL_3 \frac{\varepsilon}{2} & -(iL_1 + L_2) \frac{\varepsilon}{2} \\ (-iL_1 + L_2) \frac{\varepsilon}{2} & 1 + iL_3 \frac{\varepsilon}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \frac{\varepsilon}{2} \begin{bmatrix} -iL_3 & -(iL_1 + L_2) \\ (-iL_1 + L_2) & iL_3 \end{bmatrix} \quad (15.44)$$

Cette matrice constitue le développement limité de la matrice de rotations au voisinage de la matrice identité, cette dernière correspondant évidemment à la rotation nulle. Nous notons cette dernière également sous la forme:

$$R\left(\vec{L}, \frac{\varepsilon}{2}\right) = \sigma_0 + \varepsilon \chi(\vec{L}) \quad (15.45)$$

où la matrice σ_0 est la matrice unité d'ordre deux et $\chi(\vec{L})$ s'appelle la "**matrice infinitésimale de rotation**". Maintenant, si nous posons $L_1 = 1, L_2 = L_3 = 0$ dans $\chi(\vec{L})$ nous obtenons:

$$\chi(\vec{L}) = \chi(Ox) = \frac{1}{2} \begin{bmatrix} 0 & -i \\ -i & 0 \end{bmatrix} = -\frac{i}{2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad (15.46)$$

Comment interpréter ce résultat ? Eh bien c'est assez simple, choisir $L_1 = 1, L_2 = L_3 = 0$, nous donne un vecteur $\vec{L}$ colinéaire à l'axe Ox de $\vec{E}_3$. Dès lors, nous pouvons très bien nous imaginer les plans générant l'axe Ox qui porte $\vec{L}$. Comme $\varepsilon/2$ (in extenso $\theta/2$) est généré par les vecteurs $\vec{A}, \vec{B}$ perpendiculaires à $\vec{L}$ et donc à Ox , alors l'angle $\varepsilon/2$ (ou sa variation) représente une variation de la direction des plans normaux à $\vec{A}, \vec{B}$ qui par symétrie servent à construire la rotation (rappelons que $\vec{A}, \vec{B}$ ne sont pas nécessairement orthogonaux entre eux). Donc par extension, avoir $L_1 = 1, L_2 = L_3 = 0$ ne permet plus que de faire des rotations (symétries) autour de Ox .

De même, une rotation autour de l'axe Oy correspond à $L_2 = 1, L_1 = L_3 = 0$, ce qui donne:

$$\chi(\vec{L}) = \chi(Oy) = \frac{1}{2} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = -\frac{i}{2} \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \quad (15.47)$$

et de même avec $L_3 = 1, L_1 = L_2 = 0$ nous avons enfin:

$$\chi(\vec{L}) = \chi(Oz) = \frac{1}{2} \begin{bmatrix} -i & 0 \\ 0 & i \end{bmatrix} = -\frac{i}{2} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (15.48)$$

Les trois matrices:

$$\chi(\vec{L}) = \chi(Ox) = \frac{1}{2} \begin{bmatrix} 0 & -i \\ -i & 0 \end{bmatrix} = -\frac{i}{2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\chi(\vec{L}) = \chi(Oy) = \frac{1}{2} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = -\frac{i}{2} \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}$$

$$\chi(\tilde{L}) = \chi(Oz) = \frac{1}{2} \begin{bmatrix} -i & 0 \\ 0 & i \end{bmatrix} = -\frac{i}{2} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (15.49)$$

sont donc les matrices de rotation dans l'espace des "spineurs à deux composantes". Les physiciens et mathématiciens disent que ces matrices constituent une représentation irréductible de dimension deux du groupe "SU(2)" ou encore appelé "groupe spécial des rotations spatiales SU(2)" (cf. chapitre d'Algèbre Ensembliste).

Les matrices infinitésimales précédentes font donc apparaître de manière habile les matrices suivantes:

$$\sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \sigma_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (15.50)$$

Ces matrices sont appelées "matrices de Pauli" et nous les retrouverons dans le chapitre de Physique Quantique Ondulatoire et dans le cadre de l'étude de l'équation de Dirac et de la détermination de ses solutions explicites (utilisant les spineurs) dans le chapitre de Physique Quantique Relativiste.

En utilisant ces matrices de Pauli, la matrice de rotations infinitésimales peut finalement s'écrire:

$$R\left(\tilde{L}, \frac{\varepsilon}{2}\right) = \sigma_0 - \left(i \frac{\varepsilon}{2}\right) (L_1 \sigma_1 + L_2 \sigma_2 + L_3 \sigma_3) \quad (15.51)$$

Définissons un vecteur $\vec{\sigma}$, appelé "vecteur de Pauli", ayant pour composantes les matrices de Pauli:

$$\vec{\sigma}(\sigma_1, \sigma_2, \sigma_3) \quad (15.52)$$

L'expression $L_1 \sigma_1 + L_2 \sigma_2 + L_3 \sigma_3$ peut alors s'écrire sous forme d'une sorte de produit scalaire qui représente une somme de matrices (la flèche au-dessus du sigma est parfois omise si aucune confusion n'est possible):

$$\tilde{L} \circ \vec{\sigma} = L_1 \sigma_1 + L_2 \sigma_2 + L_3 \sigma_3 \quad (15.53)$$

Le développement limité s'écrit alors:

$$R\left(\tilde{L}, \frac{\varepsilon}{2}\right) = \sigma_0 - \left(i \frac{\varepsilon}{2}\right) \tilde{L} \circ \vec{\sigma} \quad (15.54)$$

La matrice de rotations:

$$R\left(\tilde{L}, \frac{\theta}{2}\right) = \begin{bmatrix} \cos\left(\frac{\theta}{2}\right) - iL_3 \sin\left(\frac{\theta}{2}\right) & -(iL_1 + L_2) \sin\left(\frac{\theta}{2}\right) \\ (-iL_1 + L_2) \sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) + iL_3 \sin\left(\frac{\theta}{2}\right) \end{bmatrix} \quad (15.55)$$

peut à l'aide des matrices de Pauli s'écrire sous la forme remarquable:

$$R\left(\vec{L}, \frac{\theta}{2}\right) = \cos\left(\frac{\theta}{2}\right) \sigma_0 - i \sin\left(\frac{\theta}{2}\right) \vec{L} \circ \vec{\sigma} \quad (15.56)$$

forme que nous utiliserons dans le chapitre d'Informatique Quantique pour exprimer les matrices R de manière explicite ainsi que dans le chapitre d'Algèbre Ensembliste.

Ce qui s'écrit parfois:

$$R(\vec{L}, \theta) = \cos \frac{\theta}{2} I - i \sin \frac{\theta}{2} \vec{L} \circ \vec{\sigma} \quad (15.57)$$

Ce qui peut s'écrire aussi:

$$R\left(\vec{L}, \frac{\theta}{2}\right) = \cos\left(\frac{\theta}{2}\right) \sigma_0 - L_1 i \sin\left(\frac{\theta}{2}\right) \sigma_1 - L_2 i \sin\left(\frac{\theta}{2}\right) \sigma_2 - L_3 i \sin\left(\frac{\theta}{2}\right) \sigma_3 \quad (15.58)$$

qui a donc la forme d'un quaternion de rotation d'angle θ et d'axe $\vec{L}$. D'où la raison d'avoir depuis le début choisi la notation de $\varepsilon/2$.

Il est clair, pour que l'analogie avec les quaternions soit plus forte, que les matrices 2×2 de Pauli forment un ensemble de quatre matrices linéairement indépendantes ! Tel que la base canonique pour les quaternions !

Si nous notons $\vec{L}(L_1, L_2, L_3) = \vec{X}(x, y, z)$ alors le "**produit spinoriel**" est défini finalement par:

$$L_1 \sigma_1 + L_2 \sigma_2 + L_3 \sigma_3 = x \sigma_1 + y \sigma_2 + z \sigma_3 = \vec{X} \circ \vec{\sigma} = \begin{bmatrix} z & x - iy \\ x + iy & -z \end{bmatrix} \quad (15.59)$$

Cette matrice constitue comme nous en avons déjà fait mention, au développement limité de la matrice de rotation au voisinage de la matrice identité, les composantes de $\vec{X}$ étant associées à un spineur dont la rotation se fait par la double symétrie définie par deux plans dont l'intersection est définie par le vecteur $\vec{X}$.

Nous pouvons par ailleurs remarquer la conséquence intéressante qu'une rotation de 360° ne restore pas l'objet dans sa position initiale.

Effectivement:

$$\begin{aligned} R(\vec{L}, 0) &= \cos(0) I - i \sin(0) \vec{L} \circ \vec{\sigma} = I \\ R(\vec{L}, 2\pi) &= \cos(\pi) I - i \sin(\pi) \vec{L} \circ \vec{\sigma} = -I \\ R(\vec{L}, 4\pi) &= \cos(2\pi) I - i \sin(2\pi) \vec{L} \circ \vec{\sigma} = I \end{aligned} \quad (15.60)$$

Il faut donc une rotation de 720° pour faire un tour complet! Cela correspond au spin de $\frac{1}{2}$. Il faut faire deux tours pour retrouver que l'objet réapparaisse de manière équivalente. Nous disons alors que la représentation des rotations est "**bivaluée**".

1.1.3. PROPRIÉTÉS DES MATRICES DE PAULI

Le lecteur vérifiera aisément (si ce n'est pas le cas il pourra toujours nous contacter pour que nous en rédigeons les détails) les propriétés suivantes des matrices de Pauli dont certaines seront utilisées dans le chapitre de physique quantique relativiste:

P1. Unitarité:

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_0 \quad (15.61)$$

P2. Anticommutativité:

$$\sigma_i \sigma_j = -\sigma_j \sigma_i \quad (15.62)$$

pour $i \neq j$ et $i, j = 1, 2, 3$

Les deux dernières propriétés nous donnent:

$$\sigma_i \sigma_j + \sigma_j \sigma_i = 2\delta_{ij} \sigma_0 \quad (15.63)$$

avec $i, j = 1, 2, 3$

P3. Cyclicité:

$$\sigma_1 \sigma_2 = i\sigma_3 \quad \sigma_2 \sigma_3 = i\sigma_1 \quad \sigma_3 \sigma_1 = i\sigma_2 \quad (15.64)$$

P4. Commutation:

$$\begin{aligned} \sigma_1 \sigma_2 - \sigma_2 \sigma_1 &= 2i\sigma_3 \\ \sigma_3 \sigma_1 - \sigma_1 \sigma_3 &= 2i\sigma_2 \\ \sigma_2 \sigma_3 - \sigma_3 \sigma_2 &= 2i\sigma_1 \end{aligned} \quad (15.65)$$

P5. Produit vectoriel:

Soit le carré des composantes de σ en notant abusivement par "1" la matrice unitaire (nous changeons les indices afin de vous habituer aux autres notations courantes):

$$\begin{aligned} \sigma_x \sigma_x &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 1 \\ \sigma_y \sigma_y &= \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 1 \\ \sigma_z \sigma_z &= \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 1 \end{aligned} \quad (15.66)$$

Ce qui conduit à écrire que (norme du vecteur de Pauli au carré):

$$\sigma^2 = \sigma_x^2 + \sigma_y^2 + \sigma_z^2 = 3 \cdot \mathbb{1} = 3 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (15.67)$$

Considérons maintenant les produits suivants:

$$\begin{aligned} \sigma_x \sigma_y &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix} = i \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = i \sigma_z \\ \sigma_x \sigma_y &= i \sigma_z \\ \sigma_x \sigma_z &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & i \cdot i \\ -i \cdot i & 0 \end{bmatrix} = i \begin{bmatrix} 0 & i \\ -i & 0 \end{bmatrix} = -i \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} = -i \sigma_y \\ \sigma_x \sigma_z &= -i \sigma_y \\ \sigma_z \sigma_x &= \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -i \cdot i \\ i \cdot i & 0 \end{bmatrix} = i \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} = i \sigma_y \\ \sigma_z \sigma_x &= i \sigma_y \\ \sigma_y \sigma_z &= \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix} = i \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = i \sigma_x \\ \sigma_y \sigma_z &= i \sigma_x \\ \sigma_z \sigma_y &= \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} = \begin{bmatrix} 0 & -i \\ -i & 0 \end{bmatrix} = -i \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = -i \sigma_x \\ \sigma_z \sigma_y &= -i \sigma_x \end{aligned} \quad (15.68)$$

Toutes ces relations peuvent se résumer sous la forme:

$$\boxed{\sigma_i \sigma_j = \delta_{ij} \mathbb{1} + i \varepsilon_{ijk} \sigma_k} \quad (15.69)$$

où pour rappel (cf. chapitre de Calcul Tensoriel) le symbole de Kronecker est défini par:

$$\delta_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases} \quad (15.70)$$

et le tenseur d'antisymétrie par:

$$\varepsilon_{ijk} = \begin{cases} +1 & \text{pour une permutation paire de } 1,2,3 \text{ ou aucune permutation} \\ -1 & \text{pour une permutation impaire de } 1,2,3 \\ 0 & \text{si deux indices ou plus sont les mêmes} \end{cases}$$

Nous avons aussi:

$$\begin{aligned} \sigma_x \sigma_y - \sigma_y \sigma_x &= 2i \sigma_z \\ \sigma_y \sigma_z - \sigma_z \sigma_y &= 2i \sigma_x \\ \sigma_z \sigma_x - \sigma_x \sigma_z &= 2i \sigma_y \end{aligned} \quad (15.71)$$

Nous retrouvons donc ici les composantes du produit vectoriel:

$$\vec{\sigma} \times \vec{\sigma} = 2i\vec{\sigma} \quad (15.72)$$

Maintenant voyons une identité spinorielle qui nous sera utile dans le chapitre de Physique Quantique Relativiste:

$$\begin{aligned}
 & (\vec{\sigma} \circ \vec{a}) (\vec{\sigma} \circ \vec{b}) \\
 &= (\sigma_1 a_1 + \sigma_2 a_2 + \sigma_3 a_3) (\sigma_1 b_1 + \sigma_2 b_2 + \sigma_3 b_3) \\
 &= \sum_{i,j} a_i \sigma_i b_j \sigma_j \\
 &= \begin{bmatrix} a_x & a_x - ia_y \\ a_x + ia_y & -a_x \end{bmatrix} \begin{bmatrix} b_x & b_x - ib_y \\ b_x + ib_y & -b_x \end{bmatrix} \\
 &= \begin{bmatrix} a_x b_x + (b_x + ib_y)(a_x - ia_y) & a_x (b_x - ib_y) - b_x (a_x - ia_y) \\ b_x (a_x + ia_y) - a_x (b_x + ib_y) & (b_x - ib_y)(a_x + ia_y) + a_x b_x \end{bmatrix} \\
 &= \begin{bmatrix} a_x b_x + b_x a_x + b_y a_y + i(b_y a_x - b_x a_y) & a_x b_x - ia_x b_y - b_x a_x + b_x ia_y \\ b_x a_x + b_x ia_y - a_x b_x - ia_x b_y & a_x b_x + b_x a_x + b_y a_y + i(b_x a_y - b_y a_x) \end{bmatrix} \quad (15.73) \\
 &= \begin{bmatrix} a_x b_x + a_y b_y + a_z b_z & 0 \\ 0 & a_x b_x + a_y b_y + a_z b_z \end{bmatrix} \\
 &+ \begin{bmatrix} i(b_y a_x - b_x a_y) & a_x b_x - b_x a_x + i(b_x a_y - a_x b_y) \\ b_x a_x - a_x b_x + i(b_x a_y - a_x b_y) & i(b_x a_y - b_y a_x) \end{bmatrix} \\
 &= (\vec{a} \circ \vec{b}) \mathbb{1} + \begin{bmatrix} i(b_y a_x - b_x a_y) & a_x b_x - b_x a_x + i(b_x a_y - a_x b_y) \\ b_x a_x - a_x b_x + i(b_x a_y - a_x b_y) & i(b_x a_y - b_y a_x) \end{bmatrix}
 \end{aligned}$$

Or nous avons aussi:

$$\begin{aligned}
 & i\vec{\sigma} \circ (\vec{a} \times \vec{b}) = i\vec{\sigma} \circ \begin{pmatrix} a_y b_z - a_z b_y \\ a_z b_x - a_x b_z \\ a_x b_y - a_y b_x \end{pmatrix} \\
 &= i \begin{bmatrix} a_x b_y - a_y b_x & (a_y b_z - a_z b_y) - i(a_z b_x - a_x b_z) \\ (a_y b_z - a_z b_y) + i(a_z b_x - a_x b_z) & -(a_x b_y - a_y b_x) \end{bmatrix} \quad (15.74) \\
 &= \begin{bmatrix} i(b_y a_x - b_x a_y) & a_x b_x - b_x a_x + i(b_x a_y - a_x b_y) \\ b_x a_x - a_x b_x + i(b_x a_y - b_y a_x) & i(a_y b_x - a_x b_y) \end{bmatrix}
 \end{aligned}$$

Donc finalement:

$$\boxed{(\vec{\sigma} \circ \vec{a}) (\vec{\sigma} \circ \vec{b}) = (\vec{a} \circ \vec{b}) \mathbb{1} + i\vec{\sigma} \circ (\vec{a} \times \vec{b})} \quad (15.75)$$

P6. Nous noterons que ces matrices sont aussi hermitiennes (rappelons qu'une matrice hermitienne est une matrice transposée suivie de sa conjuguée complexe selon ce que nous avons vu dans le chapitre d'Algèbre Linéaire) tel que:

$$\sigma_x = \sigma_x^\dagger \quad \sigma_y = \sigma_y^\dagger \quad \sigma_z = \sigma_z^\dagger \quad (15.76)$$

Il s'agit donc dans le langage de la physique quantique, d'opérateurs hermitiques!

Voyons maintenant quels sont les vecteurs et valeurs propres des matrices de Pauli car ce résultat est très utile en physique quantique ainsi qu'en informatique quantique!

Rappelons que lorsqu'une transformation (application d'une matrice) agit sur un vecteur, elle modifie la direction de ce vecteur excepté pour certaines matrices particulières qui ont des valeurs propres. Dans ce cas, la direction est conservée mais pas leur longueur. Cette propriété est exploitée en mécanique quantique.

Déterminons dans un premier temps, les vecteurs et valeurs propres (cf. chapitre d'Algèbre Linéaire) associées à σ_x en utilisant la méthode la plus courante:

L'équation aux valeurs propres (cf. chapitre d'Algèbre Linéaire) s'écrit donc:

$$\det(\sigma_x - \lambda I_2) = \begin{vmatrix} 0 - \lambda & 1 \\ 1 & 0 - \lambda \end{vmatrix} = 0 \quad (15.77)$$

Ce qui nous donne comme équation caractéristique:

$$\det(\sigma_x - \lambda I_2) = \lambda^2 - 1 = 0 \quad (15.78)$$

d'où les valeurs propres $\lambda = \pm 1$. Ce qui nous permet de déterminer les vecteurs propres comme suit:

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix} \Rightarrow \begin{pmatrix} y \\ x \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix} \quad (15.79)$$

Donc pour $\lambda = 1$:

$$\begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} \quad (15.80)$$

Ce qui impose que $y = x$. Le vecteur propre est donc:

$$\begin{pmatrix} x \\ y = x \end{pmatrix} = x \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \forall x \quad (15.81)$$

quelle que soit la valeur de x .

Conclusion: La direction propre du vecteur est conservée mais pas sa longueur car elle dépend de la valeur de x .

Pour $\lambda = -1$:

$$\begin{pmatrix} y \\ x \end{pmatrix} = -1 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -x \\ -y \end{pmatrix} \quad (15.82)$$

Ce qui impose que $y = -x$ et donc que le vecteur propre est:

$$\begin{pmatrix} x \\ y = -x \end{pmatrix} = x \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \forall x \quad (15.83)$$

Les vecteurs propres précédents écrits avec le formalisme de Dirac (cf. chapitre de Physique Quantique Ondulatoire) donnent pour $\lambda = 1$:

$$|v\rangle = x \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} x \\ x \end{pmatrix} \quad (15.84)$$

avec une norme de 1 puisque nous normalisons à l'unité):

$$\langle v|v\rangle = x^2 + x^2 = 2x^2 = 1 \Rightarrow x = \pm \frac{1}{\sqrt{2}} \quad (15.85)$$

Remarque: Dans le formalisme de Dirac, $\langle v|$ est le Bra et $|v\rangle$ est le Ket.

Ceci n'étant valable que pour des composantes qui sont des nombres réels. Le vecteur propre normé a donc pour expression:

$$|v\rangle = x = \pm \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad (15.86)$$

et pour $\lambda = -1$:

$$|v\rangle = x \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} x \\ -x \end{pmatrix} \quad (15.87)$$

et:

$$\langle v|v\rangle = x^2 + (-x)^2 = 2x^2 = 1 \Rightarrow x = \pm \frac{1}{\sqrt{2}} \quad (15.88)$$

et le vecteur propre normé a donc pour expression:

$$|v\rangle = \frac{\pm 1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad (15.89)$$

Déterminons maintenant, les vecteurs et valeurs propres associées à σ_y en procédant de même:

Nous avons donc pour les valeurs propres:

$$\det(\sigma_y - \lambda I_x) = \begin{vmatrix} 0 - \lambda & -i \\ i & 0 - \lambda \end{vmatrix} = 0 \Rightarrow \lambda^2 - (-i \cdot i) = 0 \Rightarrow \lambda = \pm 1 \quad (15.90)$$

Les vecteurs propres se déterminent comme suit:

$$\begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix} \Rightarrow \begin{pmatrix} -iy \\ ix \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix} \quad (15.91)$$

et donc pour $\lambda = 1$:

$$\begin{pmatrix} -iy \\ ix \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} \Rightarrow x = \frac{y}{i} = -iy \quad (15.92)$$

Le vecteur propre est dès lors:

$$\begin{pmatrix} x \\ y = ix \end{pmatrix} = \begin{pmatrix} x \\ ix \end{pmatrix} = x \begin{pmatrix} 1 \\ i \end{pmatrix}, \forall x \quad (15.93)$$

La norme associée:

$$\langle v | v \rangle = x^2 [1, -i] \begin{pmatrix} 1 \\ i \end{pmatrix} = x^2 [1 + (-i \cdot i)] = 2x^2 = 1 \Rightarrow x = \pm \frac{1}{\sqrt{2}} \quad (15.94)$$

Le vecteur propre normé a donc pour expression:

$$|v\rangle = x = \frac{\pm 1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix} \quad (15.95)$$

Pour $\lambda = -1$:

$$\begin{pmatrix} -iy \\ ix \end{pmatrix} = - \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -x \\ -y \end{pmatrix} \Rightarrow y = \frac{1}{i} x = -ix \quad (15.96)$$

Le vecteur propre est dès lors:

$$\begin{pmatrix} x \\ y = -ix \end{pmatrix} = \begin{pmatrix} x \\ -ix \end{pmatrix} = x \begin{pmatrix} 1 \\ -i \end{pmatrix}, \forall x \quad (15.97)$$

la norme associée:

$$\langle v | v \rangle = x^2 [1, i] \begin{pmatrix} 1 \\ -i \end{pmatrix} = x^2 [1^2 + (i \cdot -i)] = 2x^2 = 1 \Rightarrow x = \pm \frac{1}{\sqrt{2}} \quad (15.98)$$

Le vecteur normé a donc pour expression:

$$|v\rangle = x = \frac{\pm 1}{\sqrt{2}} \begin{pmatrix} 1 \\ -i \end{pmatrix} \quad (15.99)$$

Déterminons maintenant, les vecteurs et valeurs propres associées à σ_x en procédant de même.

Nous avons alors:

$$\det(\sigma_x - \lambda I_x) = \begin{vmatrix} 1-\lambda & 0 \\ 0 & -1-\lambda \end{vmatrix} = (1-\lambda)(-1-\lambda) = -(\lambda^2 - 1) = 0 \quad (15.100)$$

$$\lambda^2 - 1 = 0 \Rightarrow \lambda = \pm 1$$

Les vecteurs propres sont alors pour $\lambda = 1$:

$$\begin{pmatrix} x \\ -y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} \quad (15.101)$$

ce qui nous pose légèrement problème pour dire quoi que ce soit... la seule possibilité est de choisir $y = 0$ et ainsi:

$$\begin{pmatrix} x \\ 0 \end{pmatrix} = x \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (15.102)$$

et la norme associée:

$$\langle v | v \rangle = x^2 + 0^2 = x^2 = 1 \Rightarrow x = \pm 1 \quad (15.103)$$

Le vecteur propre normé a alors pour expression:

$$|v\rangle = \pm \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (15.104)$$

et pour $\lambda = -1$ nous aurons le même choix à faire en posant cette fois-ci $x = 0$ donc:

$$|v\rangle = \begin{pmatrix} 0 \\ -y \end{pmatrix} = -y \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad (15.105)$$

d'où la norme associée:

$$|v\rangle = 0^2 + (-y)^2 = y^2 = 1 \Rightarrow y = \pm 1 \quad (15.106)$$

Le vecteur propre normé a donc finalement pour expression:

$$|v\rangle = \pm \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad (15.107)$$

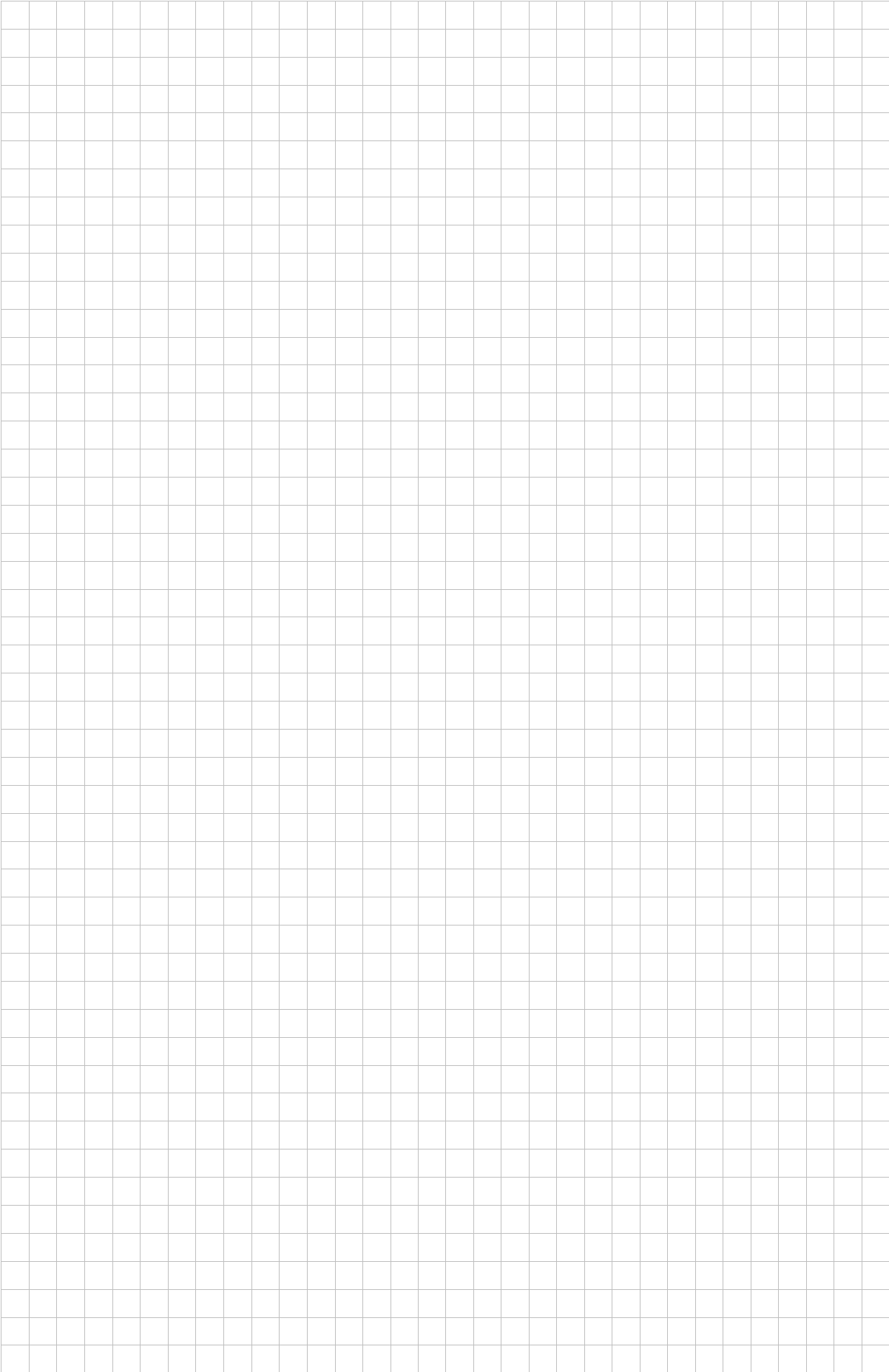
Donc les vecteurs propres normés de σ_x se trouvent sur les directions des axes de coordonnées cartésiennes. C'est pour cette raison particulière que les vecteurs propres de σ_x sont notés en informatique quantique:

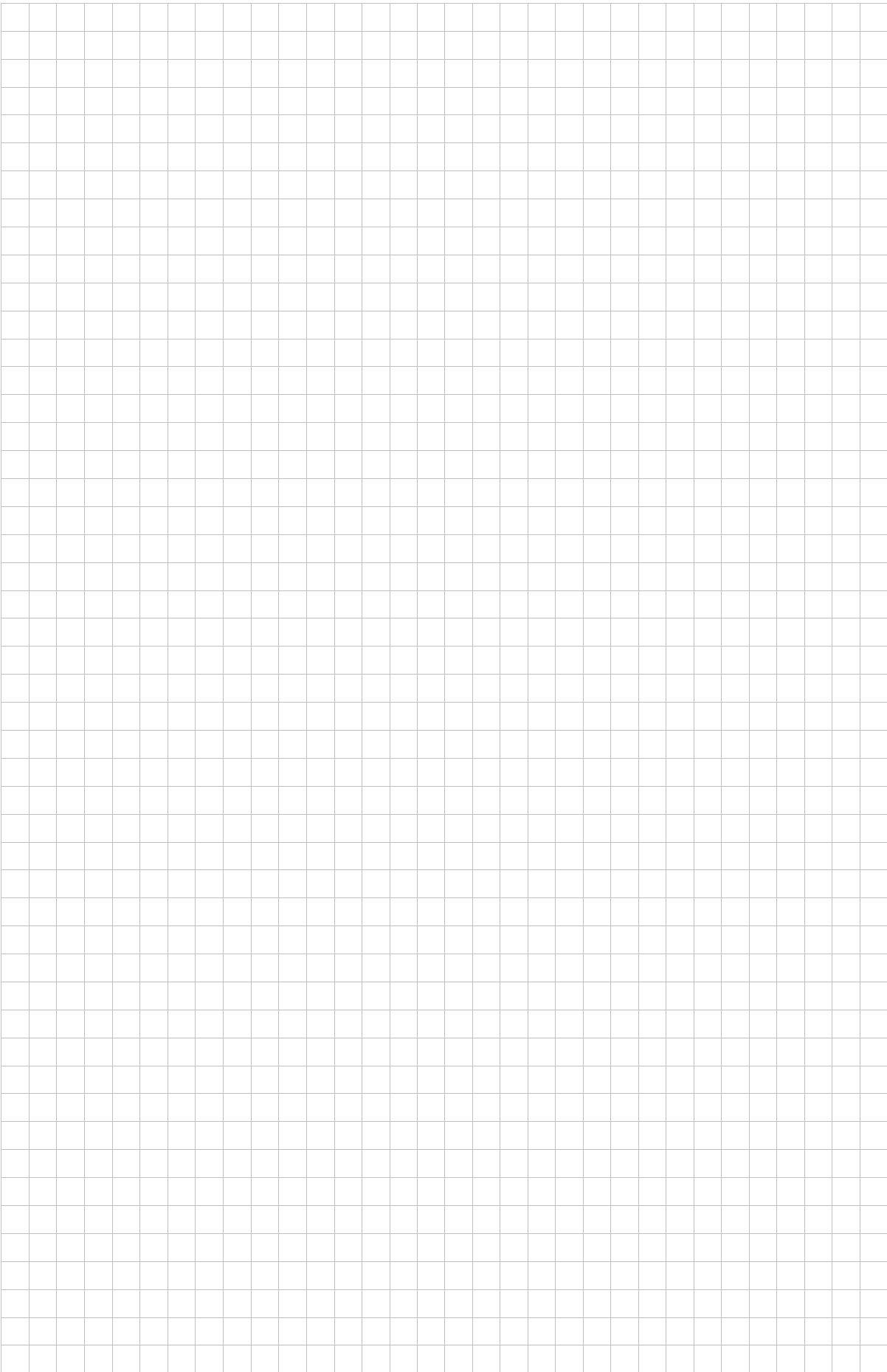
$$|+\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, |-\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad (15.108)$$

et il faut savoir que l'on note alors aussi:

$$\sigma_x |\pm\rangle = \sigma_z |\pm\rangle = \pm |\pm\rangle \quad (15.109)$$







René Descartes avait imaginé un système du monde, où l'Univers était animé par des tourbillons. À la fin du siècle dernier, Tait et Kelvin ont ressuscité cette théorie, et interprété les liaisons chimiques en imaginant des molécules nouées qui s'enlacceraient.

Les développements ultérieurs de la physique et de la chimie ont amené une autre théorie de la liaison chimique, fondée sur le partage d'électrons entre les atomes (cf. chapitre de Chimie Quantique). Mais le sujet des noeuds était lancé! Dans certains domaines, les noeuds seront... le noeud du sujet.

Remarques:

R1. La dénomination "**théorie des noeuds**" qui c'est imposée dans la communauté scientifique est assez malheureuse. Certains mathématiciens (francophones) ont choisi à juste raison d'utiliser plutôt la dénomination "**théorie des tresses et entrelacs**" qui est plus correcte et générale (puisque'un entrelac est un noeud à plusieurs composantes).

R2. Une grande partie des textes et figures ci-dessous est une reproduction, avec accord, des supports du Professeur Michael Eisermann de l'Université de Stuttgart (<http://www.igt.uni-stuttgart.de/eiserm>)

La théorie mathématique des noeuds a été lancée par les travaux de Little et Kirkman, qui cherchaient donc à donner un fondement aux idées physiques de Tait et Kelvin. Les premières classifications des noeuds utilisaient une image filiforme des noeuds, sans épaisseur, et une projection plane, comme l'ombre portée sur un écran. Un des caractères immédiatement perceptibles concerne les croisements où nous distinguons le brin au-dessus et le brin au-dessous. Les cannages, ou les tressages de bandes, reposent ainsi sur la considération des croisements, et le nombre de tels croisements sera le premier élément de classification.

1. REPRÉSENTATION DES TRESSSES

Voici un exemple de tresse (ensemble de brins) et la chaîne correspondante (une chaîne est une tresse que l'on a refermée):

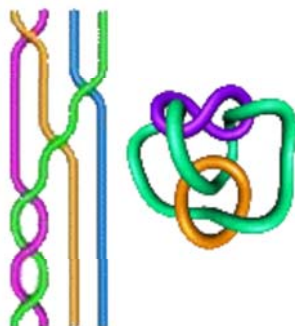


Figure: ???1 - Tresse et chaîne (diagramme de noeud) correspondante

Concentrons-nous sur la tresse à gauche de la figure ci-dessus pour commencer et demandons-nous quelle serait la meilleure manière de présenter les choses afin de pouvoir comparer des tresses? Faisons une première tentative avec la représentation suivante d'une tresse particulière:

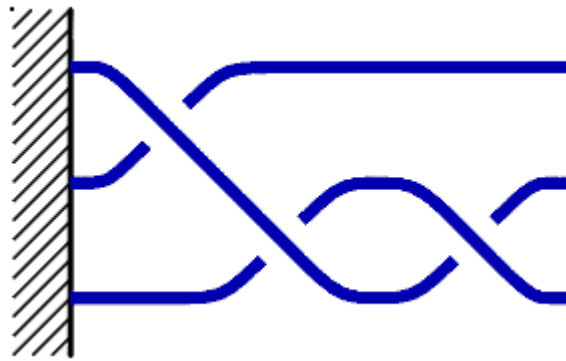


Figure: ??.2 - Présentation particulière d'une tresse

Comme les brins d'une tresse sont flexibles et peuvent bouger, nous remarquons de suite que cette représentation a un problème: toutes les tresses sont égales. Effectivement:

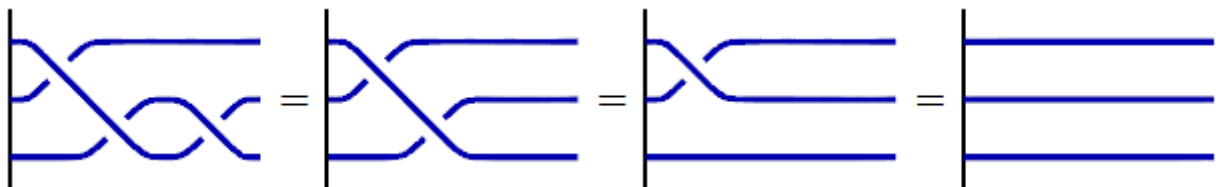


Figure: ??.3 - Exemple de problème de la représentation choisie

Le bon modèle, consiste alors peut-être à fixer les extrémités des deux côtés:



Figure: ??.4 - Autre choix de représentation

Bien évidemment, avec ce modèle les brins du milieu peuvent encore bouger comme ci-dessous et la tresse reste invariante:

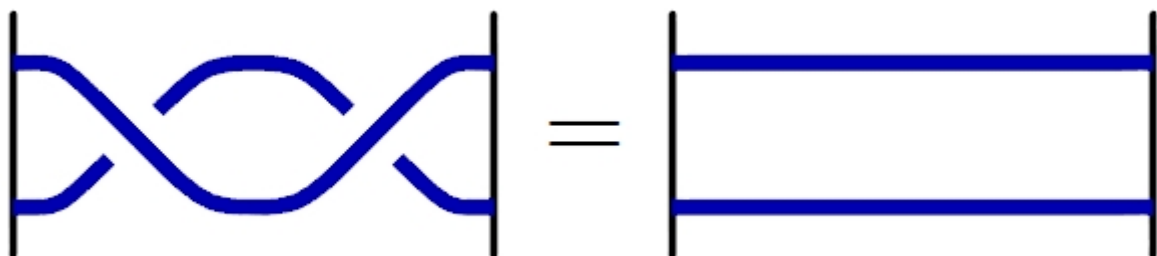


Figure: ??.5 - Premier exemple de tresse invariante

ou encore comme ici:

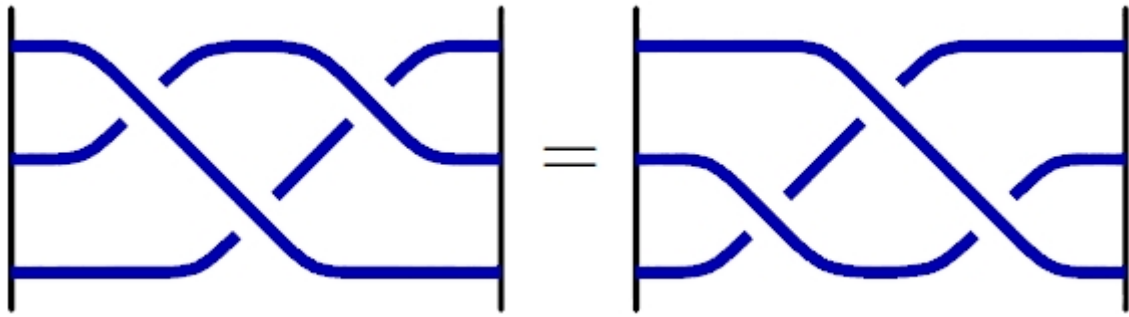


Figure: 7.6 - Deuxième exemple de tresse invariante

Nous pouvons également traduire les croisements et la tresse reste toujours invariante:

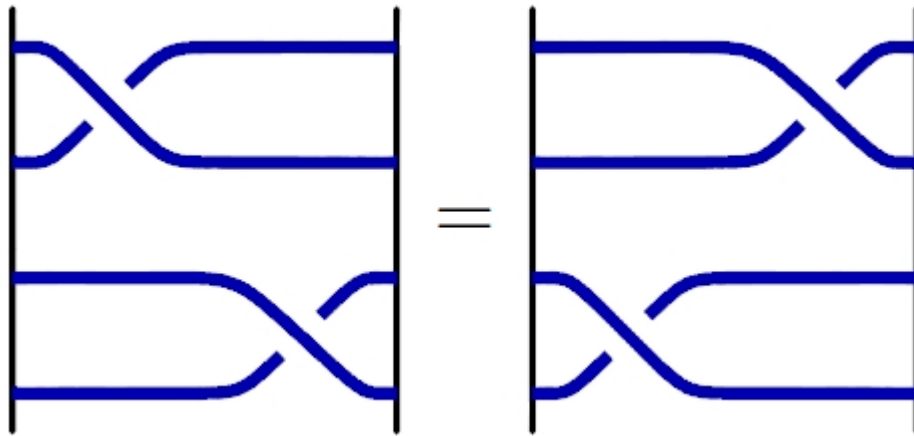


Figure: 7.7 - Translation de croisement et préservation triviale de l'invariance

La longueur n'est par ailleurs pas une variables dans ce modèle:



Figure:

7.8 - Invariance par la longueur

1.1. GROUPE DE TRESSES

Si nous définissons la multiplication de deux tresses comme étant l'opération qui consiste schématiquement à faire:

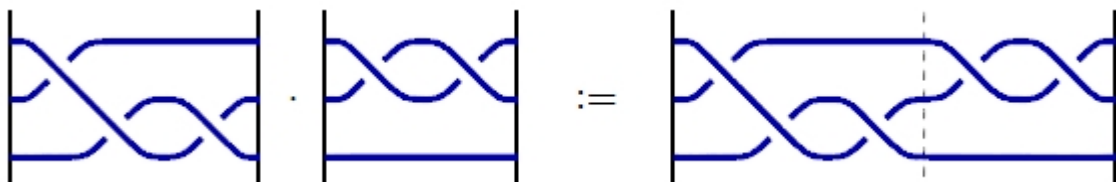


Figure: 7.9 - Exemple de multiplication de deux tresses

Nous pouvons nous poser la question si cette opération constitue un groupe du type commutatif?

Rappelons avant cela que nous avons défini la structure de groupe dans le chapitre de Théorie des ensembles de la manière suivante: Nous désignons un ensemble par le terme "**groupe**", si les composants le constituant satisfont aux trois conditions de ce que nous nommons la "**loi interne de groupe**", définie ci-dessous:

$$(G, *) \text{ est un groupe si } \Leftrightarrow \begin{cases} * \text{ est associative} \\ \text{Il existe dans } G \text{ un élément neutre pour } * \\ \text{Tout élément de } G \text{ possède un symétrique pour } * \end{cases}$$

Si de plus, la loi interne $*$ est également commutative, nous disons alors que le groupe est un "**groupe abélien**" ou simplement "**groupe commutatif**".

Commençons par le premier contrôle. Est-ce que cette représentation est associative?:

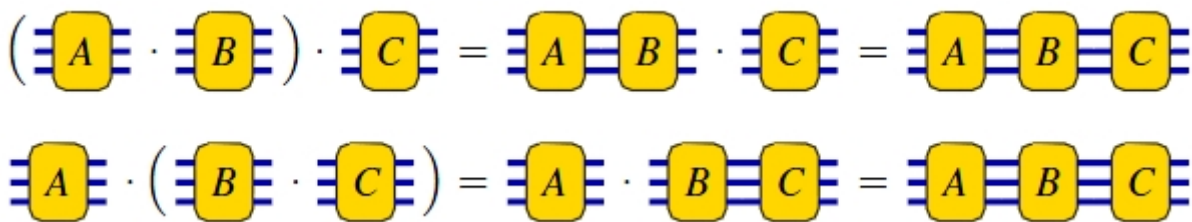


Figure: ??10 - Contrôle de l'associativité des tresses (par la multiplication)

La réponse est donc OUI quel que soit le nombre de brins!

Est-ce que cette représentation est commutative?:

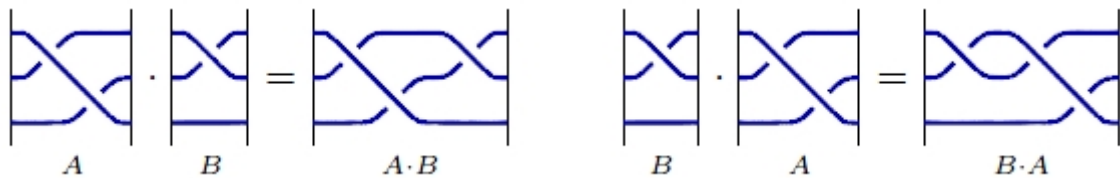


Figure: ??11 - Contrôle de la commutativité des tresses (par la multiplication)

La réponse est donc NON au-delà de 2 brins (car avec 2 brins elle est commutative)!

Admet-elle un élément neutre?:

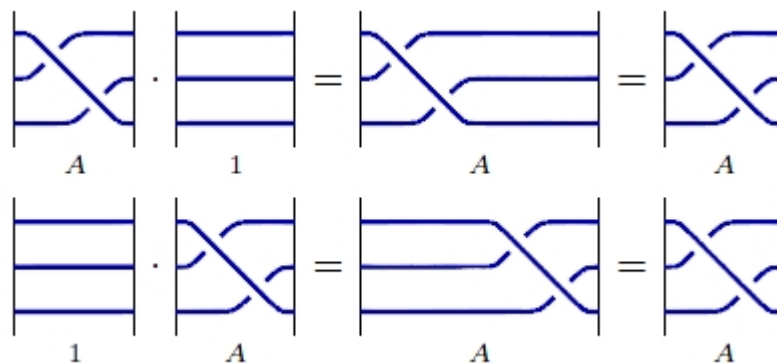


Figure: ??12 - Contrôle de l'existence d'un élément neutre des tresses (par la multiplication)

La réponse est donc OUI quel que soit le nombre de brins!

Existe-il des éléments inverses (symétriques)?:

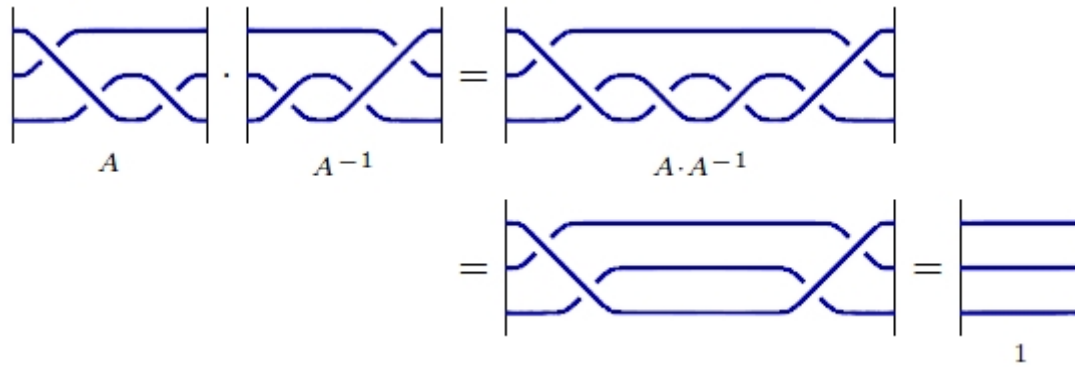


Figure: ??13 - Contrôle de l'existence d'éléments symétriques pour les tresses (par la multiplication)

La réponse est donc OUI quel que soit le nombre de brins!

Donc les tresses à n brins forme un groupe non commutatif noté $(\mathcal{B}_n, \cdot)$ où le B stylisé représente le mot "braids" (tresses: en anglais) !

Observons quelque chose d'intéressant en ce qui concerne les tresses à deux brins. Si nous indexons leurs vrilles par des nombres entiers relatifs comme ci-dessous (imaginez la tresse du milieu dont vous faites tourner les extrémités, cela donne toutes les autres vrilles de gauche et de droite):

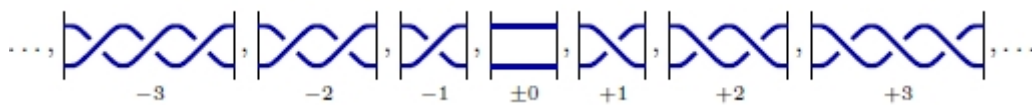


Figure: ??14 - Indexation des vrilles de tresses

avec la convention de vrille v positive et vrille v négative:

$$v\left(\begin{array}{c} \nearrow \\ \searrow \end{array}\right) = +1 \quad v\left(\begin{array}{c} \nwarrow \\ \searrow \end{array}\right) = -1$$

Figure: ??15 - Convention d'indexation

Attention au piège! (essayez de le trouver... ce n'est pas toujours facile au premier coup d'oeil):

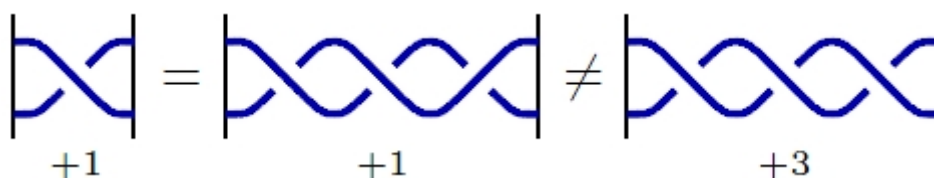


Figure: ??16 - Exemple de piège d'indexation

Effectivement, le croisement de droite la deuxième tresse ci-dessus n'est justement pas.... un croisement! Nous disons alors que le nombre de croisement n'est pas un invariant. C'est pour cette raison que ce qui compte est la "vrille" car elle ne change pas sous les mouvements de la tresse et non pas simplement un croisement!!!

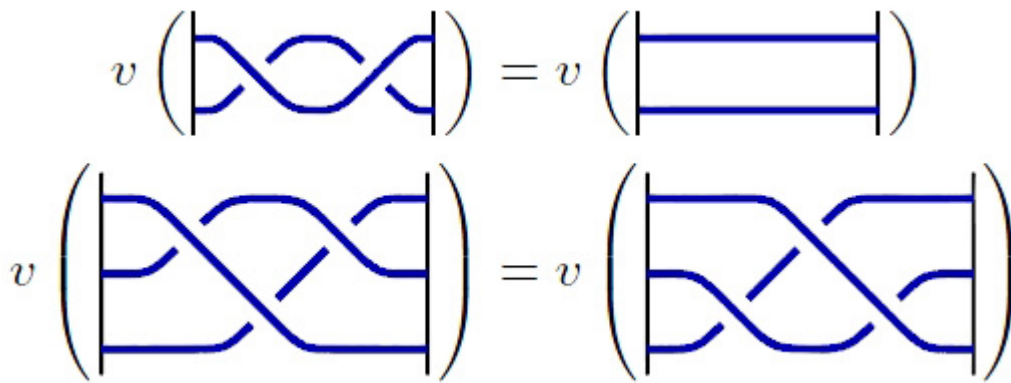


Figure: ??17 - Différence entre vrille et croisement

Nous remarquons que nous avons alors en ce qui concerne la multiplication des tresses (cf. chapitre de Théorie des Ensembles) et leur indexation sous forme du nombre de vrilles les propriétés suivantes:

$$\begin{aligned} f(A \cdot B) &= f(A) + f(B) \quad \forall A, B \in \mathcal{B}_2 \\ f(1_A) &= 1_{\mathbb{Z}} \\ f(A^{-1}) &= (f(A))^{-1} \quad \forall A \in \mathcal{B}_2 \end{aligned} \quad (??1)$$

Donc le groupe multiplicatifs des tresses à 2 brins est un homomorphisme de groupe avec la loi d'addition du groupe des entiers relatifs. De plus, nous voyons trivialement que l'application f est bijective. Nous avons dès lors un isomorphisme de groupes! Au delà de 3 brins, nous avons vu que le groupe n'était plus commutatif!

2. REPRÉSENTATION DES NOEUDS

Revenons à la première figure présentée plus haut:

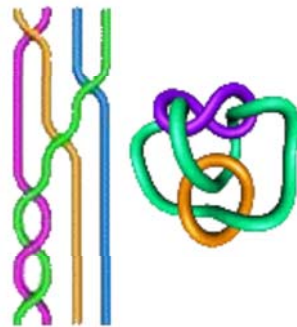


Figure: ??18 - Tresse et chaîne (diagramme de noeud) correspondante

Une première difficulté apparaît alors pour l'image de droite (qui est un "**diagramme de noeud**"): un noeud (chaîne avec un seul brin) est un objet géométrique à trois dimensions, et nous utilisons une représentation plane obtenue en projetant le noeud sur un plan. Ceci pourrait donc compliquer les choses... nous verrons!

Remarque: Tout noeud ou toute chaîne peut donc être obtenue à partir d'une tresse.

Il existe alors de nombreux points de vue possibles, et diverses apparences d'un même noeud. Deux noeuds de projections différentes sont-ils distincts (non isotopes) ? Dans la première table publiée (celle de Tait et Little) le nombre minimal de croisements est utilisé comme principe de classification (jusqu'à dix croisements) et il fallut presque un siècle pour détecter une duplication: deux noeuds identiques avaient été pris pour différents.

En pratique un noeud ressemble plutôt à ceci:



Figure: ??19 - Exemple d'un noeud dans la vie réelle

Mais les mathématiciens ont pris l'habitude de connecter les deux extrémités de la corde pour obtenir ceci:



Figure: ??20 - Représentation par un matheux

donc un noeud est toujours une boucle qui se referme sur elle-même. En d'autres termes, un noeud ouvert est toujours équivalent à un noeud fermé:

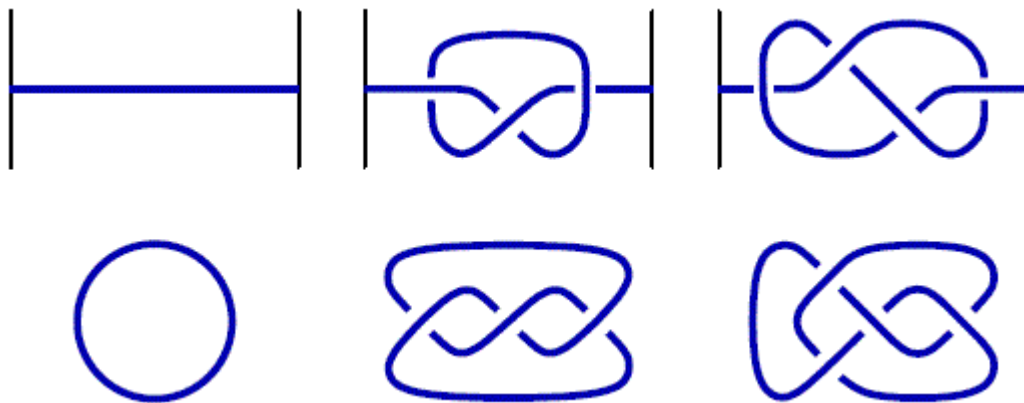


Figure: ??21 - Équivalence entre un noeud ouvert et un noeud fermé

Un noeud fermé peut paraître différent suivant l'angle sous lequel on le regarde. Ainsi, les deux noeuds ci-dessous sont deux représentations du même type de noeud appelé le "**noeud de trèfle**".



Figure: ??22 - 2 perspectives différentes du noeud de trèfle

Nous pouvons également nous poser la question si le noeud de trèfle à gauche et son miroir à droite sont égaux?:

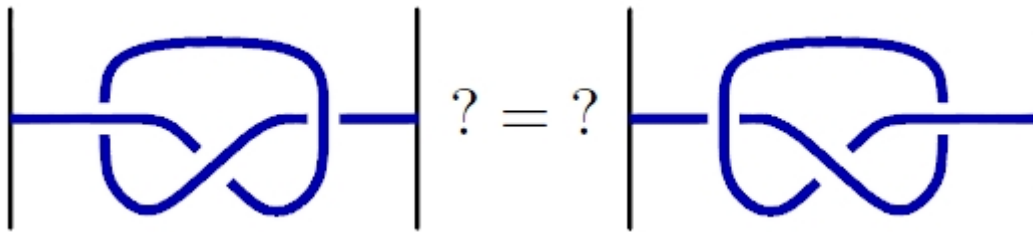


Figure: ??23 - Recherche d'égalité entre symétrie miroir du noeud de trèfle

ou la même question mais lorsque les deux noeuds sont fermés:

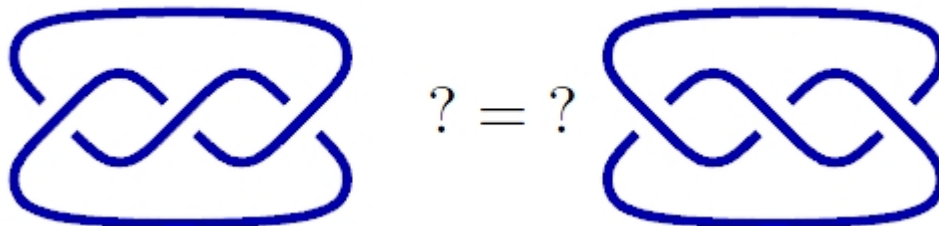


Figure: ??24 - Même situation mais en fermant le noeud

Plus difficile, Perko a montré que ces deux noeuds de la paire dite "**paire de Perko**", sont en fait le même noeud:

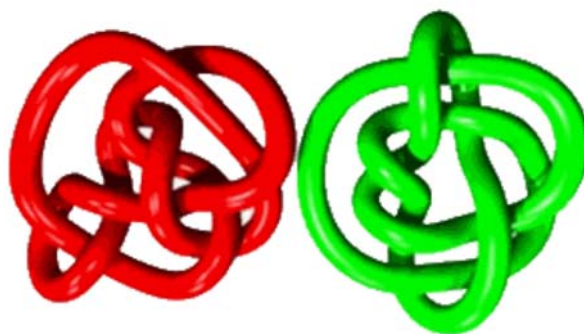


Figure: ??25 - Exemple d'une paire de Perko

Il n'est donc pas évident de savoir lorsque deux objets physiques représentent fondamentalement le même noeud. Mais, c'est cela qui intéresse les mathématiciens. Plus précisément, ils aimeraient classifier les noeuds, c'est-à-dire déterminer tous les types de noeuds qui sont fondamentalement différents et pas simplement en apparence (c'est la même idée qu'en topologie ou les mathématiciens ont réussie à démontrer que tout volume se réduisait à trois volumes primaires).

2.1. GROUPE DE NOEUDS

Au même titre que nous l'avons fait pour les tresses, regardons si les noeuds forment un groupe?

Donc nous définissons la multiplication de deux noeuds comme étant l'opération qui consiste schématiquement à faire:

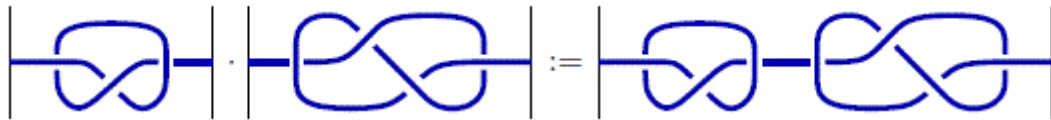


Figure: ??26 - Rappel de la définition de la multiplication de deux noeuds

Nous pouvons nous poser la question si cette opération constitue un groupe du type commutatif?

Commençons par le premier contrôle. Est-ce que cette représentation est associative?:

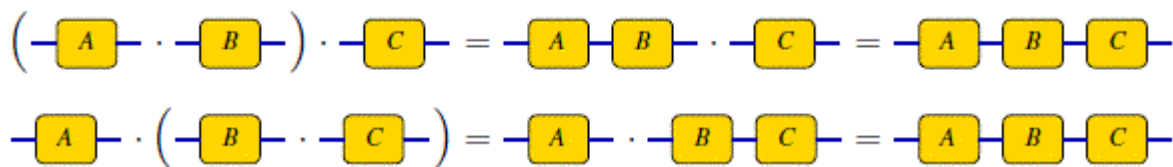


Figure: ??27 - Vérification de la commutativité de noeuds (par la multiplication)

La réponse est donc OUI quel que soit le nombre de noeuds!

Admet-elle un élément neutre?:

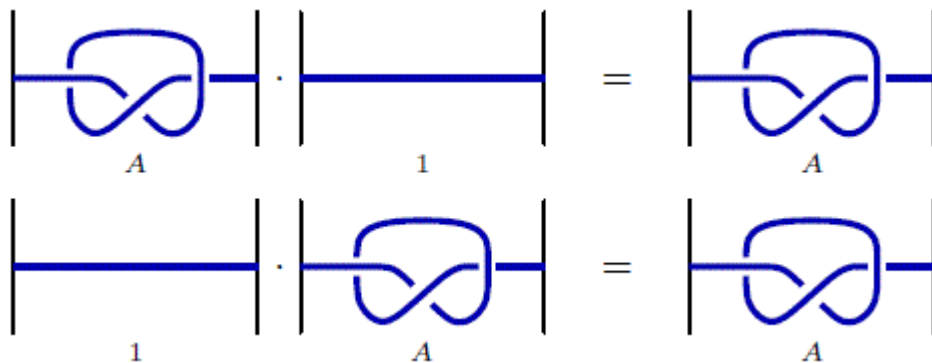


Figure: ??28 - Vérification de l'existence d'un élément neutre pour les noeuds (par la multiplication)

La réponse est donc OUI quel que soit le nombre de brins!

Est-elle commutative?:

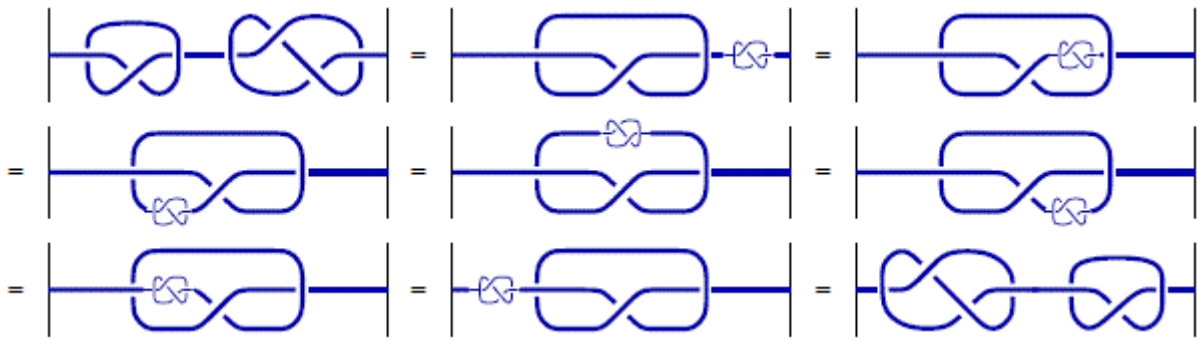


Figure: ??29 - Vérification de la commutativité de noeuds (par la multiplication)

Donc contrairement aux tresses OUI!

Existe-il des éléments inverses (symétriques)?:



Figure: ??30 - Vérification de l'existence d'un élément neutre pour les noeuds (par la multiplication)

et ça c'est une des grandes questions!

Avec le temps et les efforts des mathématiciens, trois points de vue mathématiques essentiels pour l'étude des noeuds se sont dégagés afin de résoudre différents problèmes:

1. Topologique, où le noeud est conçu comme la réunion d'un nombre fini de courbes fermées, à déformations près.
2. Algébrique, où l'on dénombre des croisements par exemple, et où l'on associe les groupes aux noeuds.
3. Géométrique, où l'on tient compte de la forme du noeud, en mesurant longueurs ou angles. En particulier, l'idée du nombre de rotations apparaît sous la forme de vrillage et de nombre d'entrelacements, et ces notions sont fondamentales pour l'étude des ADN en biologie moléculaire.

L'articulation entre ces points de vue est délicate. Il y a des évidences difficiles à prouver rigoureusement – qu'on songe au théorème de Jordan qui affirme qu'une courbe fermée plane, sans croisement, délimite un intérieur et un extérieur ! Il a fallu presque deux siècles pour définir correctement la notion de courbe, et il y a des noeuds sauvages à écarter avant toute classification. Il faut du soin pour définir correctement les déformations des noeuds, que le mathématicien appelle "**isotopies**".

L'idée mathématique la plus efficace pour l'étude des noeuds est celle de "**noeud invariant**". Un invariant d'un noeud est une caractéristique (nombre entier, nombre réel, polynôme, groupe...) qui reste inchangée lors d'une déformation. Si nous disposons d'un invariant, nous pouvons affirmer que deux noeuds sont vraiment différents quand l'invariant ne prend pas la même valeur pour les deux noeuds. Mais si deux noeuds ont le même invariant, nous ne pouvons affirmer qu'ils sont du même type (déformable l'un en l'autre). Un exemple typique de deux noeuds ayant le même invariant et pour lesquels il n'est pas trivial de dire s'ils sont isotopes est le trèfle de droite et celui de gauche lorsque l'invariant choisi est le nombre de croisements $c(D)$. Il faudrait pour cela disposer d'un système complet d'invariants.

Le mathématicien russe V. Vassiliev a introduit en 1990 une classe nouvelle d'invariants. Il reste à les rendre explicitement calculables et à prouver qu'ils forment un système complet, comme on le pense généralement.

Henri Poincaré a introduit vers 1900 la notion de groupe fondamental d'un espace, qui décrit les possibilités de parcours avec retour au point de départ. Appliqué à l'espace extérieur à un noeud, cela fournit le "**groupe du noeud**" et le "**polynôme d'Alexander**" qui lui est lié (voir plus loin dans la partie formalisation mathématique).

La même construction appliquée à ce qu'on appelle l'espace de configuration donne une définition efficace des groupes de tresses. Ces groupes ont été introduits, sous une forme intuitive, vers 1920, par le mathématicien viennois Emil Artin, un des pères de l'algèbre moderne. Dès 1937, le mathématicien russe Markov relie les noeuds et les tresses, et donne une méthode théorique pour définir, au moyen des tresses, des invariants des noeuds.

Il y a 15 ans, les groupes de tresses étaient une curiosité et leur complexité était rebutante. Puis, brusquement, ils sont devenus un thème central de la recherche scientifique. Donnons une idée de la diversité des points de vue qui mènent aux groupes de tresses:

1. En géométrie, le mathématicien russe Vladimir Arnold a classifié sous le nom de "catastrophes" des singularités de configurations géométriques, qui amènent aux espaces de configuration.
2. En algèbre, plus précisément dans la théorie des groupes, deux avatars des groupes de tresses (les groupes de Coxeter et les algèbres de Hecke) jouent un rôle central.
3. En mécanique statistique, l'étude des modèles exactement résolubles se fait grâce à l'emploi des relations de Yang-Baxter et des groupes quantiques ; le lien avec les groupes de tresses est profond.
4. La physique à deux dimensions a pris récemment beaucoup d'importance, et on en attend un modèle de la supraconductivité à haute température. La classification usuelle des particules en fermions et bosons se complique à deux dimensions. La notion nouvelle est celle d'anyon, dont le modèle mathématique est lié aux groupes de tresses.
5. La théorie quantique des champs est le modèle mathématique des particules élémentaires. Edward Witten a fait le pont entre cette théorie et les groupes de tresses, via le polynôme de Jones.

La théorie des noeuds constitue ainsi une interface très active entre physique et mathématiques. Les noeuds et les tresses fournissent aujourd'hui un outil de modélisation efficace de la physique des polymères aux cristaux liquides, en passant par la biologie moléculaire. Dans la direction opposée, les idées importées de la physique ont déclenché une révolution en mathématiques: d'un sujet un peu marginal il y a encore 15 ans, la théorie des noeuds est devenue aujourd'hui l'un des grands chantiers mathématiques.

3. NOEUDS DE TAIT

L'objectif principal de la théorie des noeuds est de classer tous les noeuds et trouve son origine dans la physique au début du 19^{ème} siècle comme nous en avons déjà fait mention. Pourquoi ?

Rappelons qu'à cette époque, les atomes étaient un mystère: pourquoi ceux-ci semblaient-ils indestructibles et pourtant existant en tellement de variétés et capables de se combiner pour donner d'innombrables autres composants ?

A cette époque, les plus belles équations de la physique (qui sont souvent de bons candidats pour expliquer des choses que nous ne comprenons pas...) étaient les équations de Maxwell, il était alors naturel (ou tentant) pour les physiciens d'essayer d'expliquer la mécanique atomique en terme d'électromagnétisme, même si nous savons aujourd'hui que cette voie était destinée à l'échec. Plus tard au milieu du 19^{ème}, les ondes électromagnétiques étaient largement conceptualisées comme des vibrations d'un milieu appelé à l'époque "éther luminophore". Le référentiel de cet éther était alors défini (ou du moins supposé) comme étant un référentiel absolu. Mais seulement... plus tard, les expérimentateurs Michelson et Morley, montrèrent que le mouvement relatif de notre planète dans cet éther était indétectable (et même pire... ils mesurèrent la constance absolue de la vitesse de lumière!). Leurs expériences amenèrent par ailleurs Einstein et Poincaré à développer leur fameuse théorie de la relativité restreinte (cf. chapitre de Relativité Restreinte).

Il était alors à l'époque très gênant de travailler avec le concept suivant: il y avait des ondes sans rien qui ondulaient ! Ceci du fait, qu'à l'époque, les physiciens avaient fortement tendance à comprendre l'électrodynamique avec des analogies mécaniques. Maxwell, par exemple, a passé pas mal de temps à conceptualiser les champs électriques et magnétiques en terme de "tubes fins, de section variable, transportant un fluide incompressible". Une raison pour cela était la forme des équations de Maxwell dans le vide (voir le chapitre d'électrodynamique):

$$\vec{\nabla} \cdot \vec{E} = 0 \quad \vec{\nabla} \times \vec{E} = \vec{0} \quad (??2)$$

ce qui est, il est vrai, similaire à la mécanique des fluides (voir chapitre du même nom) où nous avons pour un fluide incompressible sans viscosité et sans tourbillons:

$$\vec{\nabla} \cdot \vec{v} = 0 \quad \vec{\nabla} \times \vec{v} = \vec{0} \quad (??3)$$

Plus généralement, dans le cas où la rotationnel $\vec{\nabla} \times \vec{v}$ est non nul, Helmholtz montra en 1858 que les lignes de champ du vortex – définies par les lignes de $\vec{\nabla} \times \vec{v}$ - se déplacent dans la direction de $\vec{v}$ comme si elles avaient une existence propre (alors là c'est la brèche ouverte bien évidemment...!). Ces lignes ne pouvaient avoir de fin mais pouvaient former des boucles.

En 1867 le mathématicien P. G. Tait (assistant de Hamilton et un champion des quaternions) trouva une méthode ingénieuse pour démontrer cet effet en coupant un trou circulaire dans une boîte, en remplissant celle-ci avec de la fumée, et en expulsant ensuite la fumée par compression de l'air dans la boîte formant ainsi des cercles de fumée. Il montra par ailleurs ceci à son ami Kelvin qui nota l'analogie avec l'électromagnétisme et proposa une théorie dans laquelle les atomes étaient des vortex (noeuds) dans l'éther ! Il émit l'hypothèse que les différents types d'atomes correspondaient à différents types de vortex noués (oui c'est un peu tiré par les cheveux mais bon...) !

Tait essaya par la suite de classer (voir tableau plus bas) les lignes nouées en accord avec:

1. Le nombre d'entrelacements quand celles-ci sont projetées sur un plan
2. En ne représentant que les "noeuds premiers"

Définition: Un noeud peut être compliqué parce qu'il est la succession de noeuds simples:



Figure: ??31 - Exemple d'une succession de noeuds sur un noeud ouvert

Ces noeuds simples (à un brin) peuvent être séparés en coupant la corde:

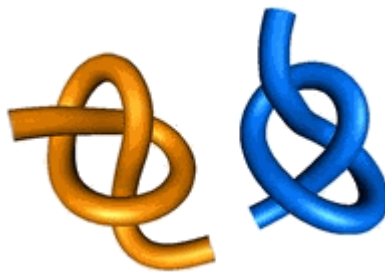


Figure: ??32 - Séparation du noeud

Un noeud premier est donc un noeud qui ne peut être séparé en noeuds plus simples: couper la corde dénoue le noeud.

Classifier les noeuds c'est donc chercher à déterminer les briques élémentaires: tous les noeuds premiers. La liste que Tait obtint dans un premier temps fut les noeuds suivants (dans l'espoir d'obtenir le tableau périodique des éléments version "éther"...) et si nous vous conseillons au besoin de vous munir d'une ficelle pour vous assurer qu'ils sont bien premiers (des fois cela est difficile mentalement):

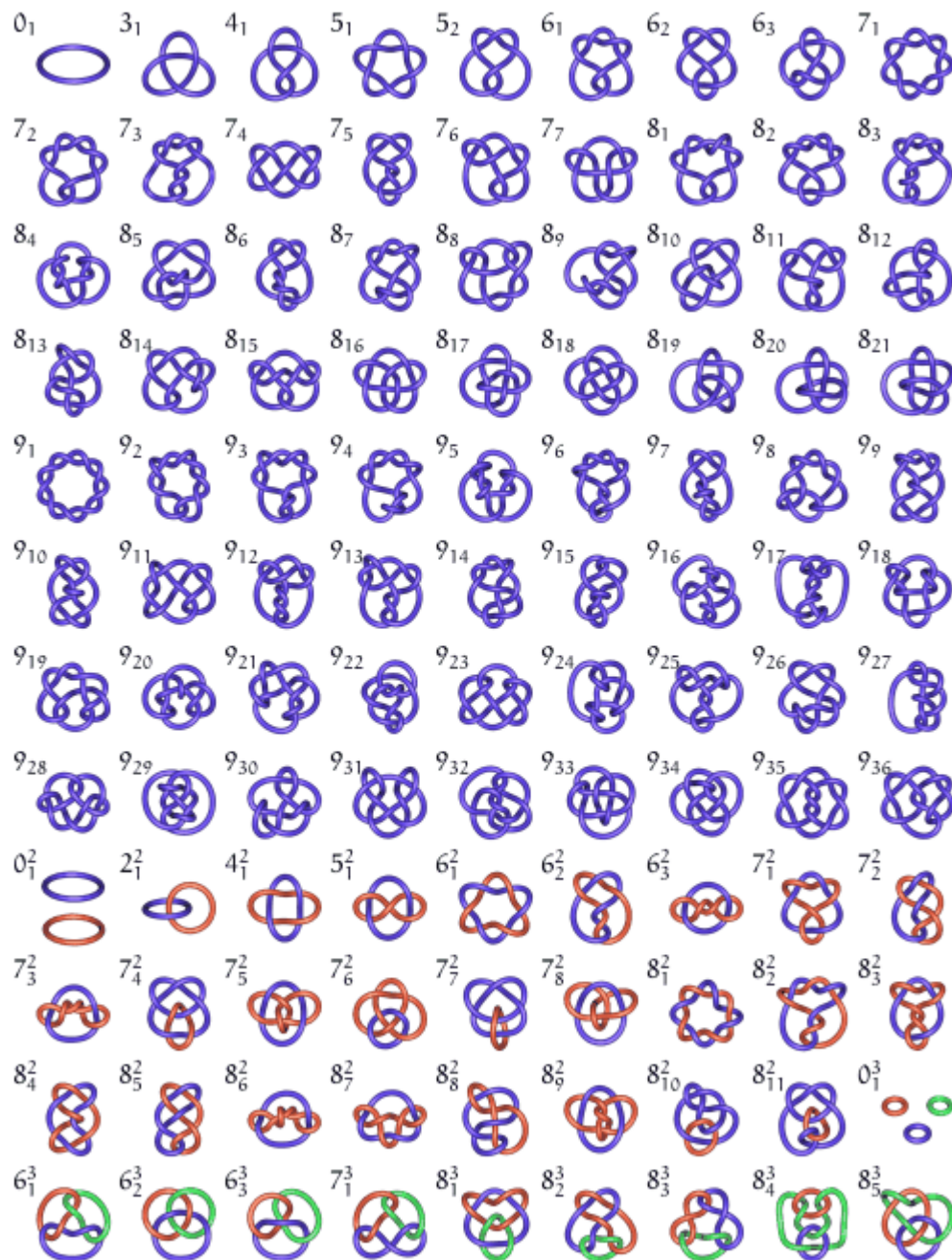


Figure: ??33 - Classification des noeuds de Tait

Remarque:

R1. Ce tableau est très important nous y ferons très souvent référence lors d'exemples applicatifs en utilisant la nomenclature qui y est proposée.

R2. Les valeurs correspondantes dans le tableau ci-dessus, appelées "**mesures de complexité du noeud**", sont notées $c(D)_{Id}^{c(B)}$ en toute généralité où $c(D)$ est le nombre de croisements, $c(B)$ le nombre de brins, et Id le numéro d'identifiant du noeud dans la classe $c(D)$.

La beauté de cette théorie des "atomes vortex" résidait dans le fait qu'il était en accord avec l'aspect continu du monde merveilleux des fluides, ou aux équations de Maxwell par extension, pour discrétiser les différents types d'atomes. Une difficulté cependant avec cette théorie était la remarquable stabilité des atomes. En 1905 Kelvin admit qu'après de nombreuses années d'échecs à tenter de prouver que le mouvement des cercles de Helmholtz était stable, que la conclusion était que ces cercles étaient essentiellement instables et devaient donc se dissiper. Curieusement c'est cette stabilité extrême des atomes qui fut une des pièces du puzzle pour construire la physique quantique corpusculaire.

De plus, avec la théorie de la relativité, le concept d'éther cher à Maxwell et ses contemporains, en particulier en ce qui concerne la théorie des noeuds, devint synonyme d'un concept inutile et avec en plus la théorie quantique la théorie des atomes vortex fut complètement oubliée. La théorie des noeuds ressurgit à cause de quelques conjectures que Tait ne fut pas capable de démontrer (nous y reviendrons) qui le furent seulement dans les années 1980 dans un tournant hasardeux de la physique théorique.

Ce que les physiciens abandonnèrent intrigua et continue d'intriguer les mathématiciens. La question de base restant la même: comment pouvons nous dire que deux noeuds sont "isotopes" (nous définirons plus loin de quoi il s'agit). Cette question est intimement reliée aux fameuses conjectures de Tait. Pour attaquer ces conjectures et les questions basiques de ressemblance des noeuds, les topologistes ont développé les noeuds invariants. Un exemple d'un noeud invariant connu et ayant eu beaucoup de succès sont les polynômes d'Alexander découverts par J. W. Alexander en 1927 (voir plus loin). Ainsi, si les polynômes de deux noeuds sont différents, ceux-ci ne sont alors pas isotopes. Malheureusement, il existe quelques noeuds ayant des polynômes d'Alexander équivalents et qui sont pourtant non isotopes...

La théorie mathématique des noeuds se développe alors pendant une cinquantaine d'années et était un peu tombée en désuétude lors du coup de tonnerre de la découverte par Jones en 1984 d'un nouvel invariant des noeuds (le polynôme de Jones). La découverte de Jones est assez exemplaire du point de vue scientifique et peut donner lieu à méditation sur l'organisation actuelle de la science. Jones n'était pas du tout un spécialiste des noeuds. Il s'intéressait à la classification des facteurs dans les algèbres de Von Neumann (analyse fonctionnelle). Il a obtenu des algèbres de matrices dont les relations de commutation (équations de Yang-Baxter) étaient proches des relations du groupe de tresses. Des tresses aux noeuds, il n'y a qu'un pas qu'il a franchi avec l'aide de Joan Birman qui est une spécialiste des noeuds.

Nous assistons ensuite à une explosion de découvertes: version purement combinatoire, nouveaux polynôme, etc. En 1989, Witten montre que le polynôme de Jones peut être obtenu à partir de la théorie quantique des champs au moyen d'une intégrale de Feynman, donnant ainsi la première définition n'utilisant pas les projections planes du noeud. D'une certaine façon la théorie de Jones-Witten est une extension non commutative du travail de Gauss. Le groupe de Lie (cf. chapitre d'Algèbre Ensembliste) qui intervient en magnétisme est $U(1)$, alors que l'invariant de Witten est une intégrale de Feynman sur un espace de $SU(2)$ -connections.

4. FORMALISATION MATHÉMATIQUE

Un noeud est modélisé mathématiquement par une application injective, différentiable et dont la dérivée ne s'annule pas, du cercle dans l'espace orienté de dimension 3 (noeud trivial).

Les deux problèmes centraux de la théorie des noeuds est de pouvoir décider de façon calculable si le noeud est trivial (peut se défaire sans couper la ficelle...) ou non (ce problème n'est pas résolu) et si deux noeuds sont vraiment équivalents.

Le premier type de problème peut être bien représenté quant au fait de déterminer si le noeud suivant est noué ou non...?:

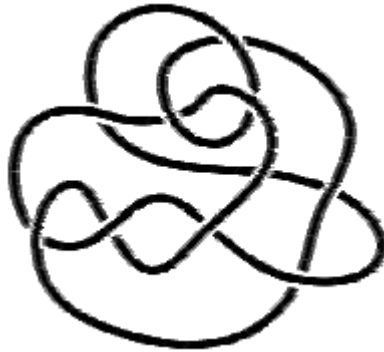


Figure: ??34 - Noeud noué?

La réponse est non comme le montre la figure ci-dessous (lire de gauche à droite et de haut en bas):

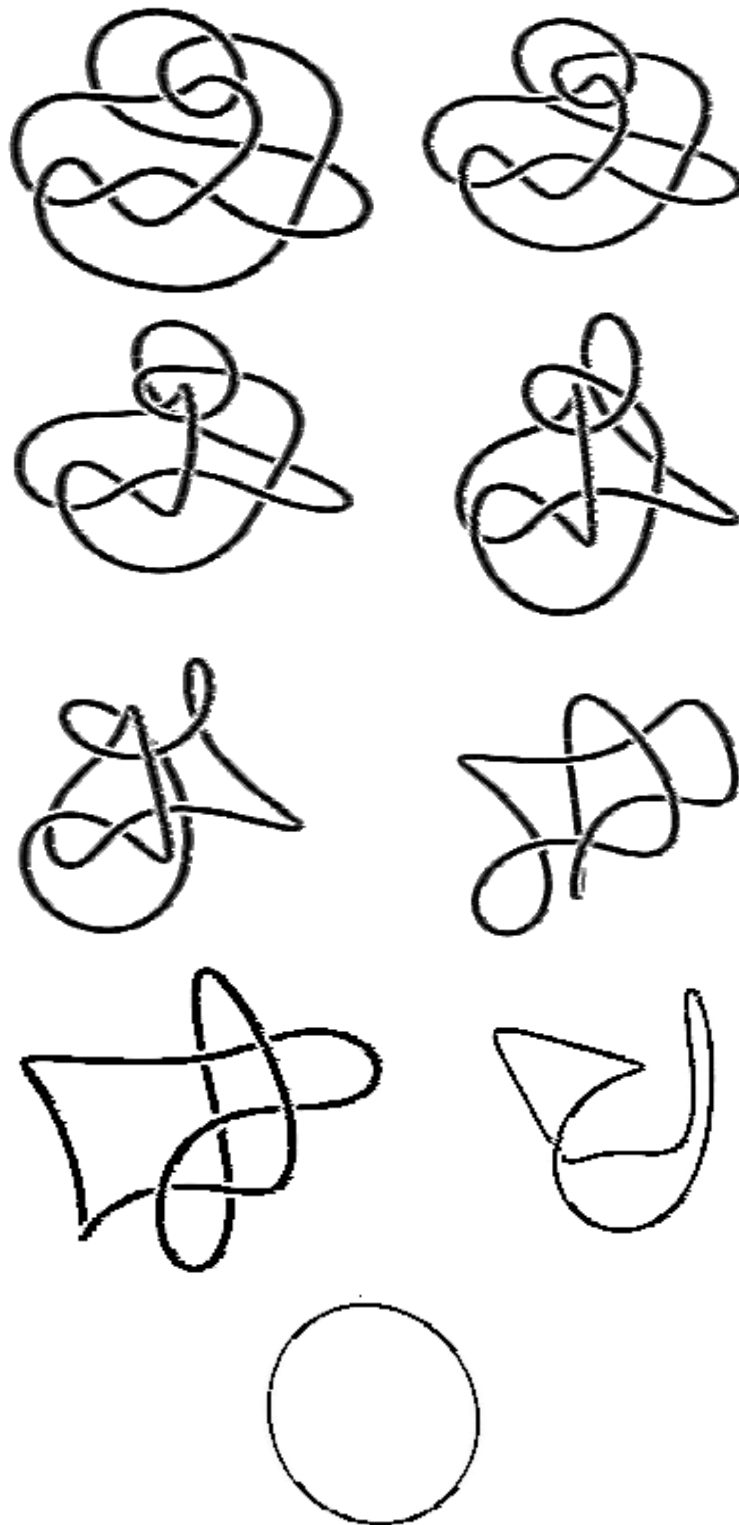


Figure: ??.35 - Eh ben non....

Le but pour le deuxième problème est d'associer aux noeuds des objets mathématiques calculables (polynômes, nombres) appelés "**invariants du noeud**" et qui sont insensibles à une déformation du noeud. Si l'invariant n'est pas égal à celui du noeud trivial qui est:

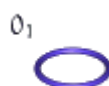


Figure: ??.36 - Noeud trivial

nous sommes sûr que le noeud n'est pas trivial. Et par exemple, le noeud non trivial le plus simple est le noeud de trèfle:

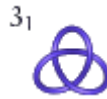


Figure: ??37 - Noeud de trèfle

Le problème est donc de trouver des invariants assez fins. Nous en décrivons deux:

1. Le nombre d'entrelacement (de 2 noeuds) du à Gauss et qui intervient en électromagnétisme
2. Le polynôme de Jones (introduit dans les années 1985 par Vaughan Jones médaille Fields 1990), qui est assez subtil pour distinguer par exemple le noeud de trèfle droite du gauche.

Nous décrivons aussi une classe générale d'invariants: les invariants de type fini ou de Vassiliev. Ces invariants définis de façon assez peu constructive sont peut-être des invariants complets, mais nous ne le savons pas à ce jour. Nous décrivons le nombre d'entrelacement de 2 noeuds comme un invariant combinatoire calculable à partir d'un diagramme des noeuds. Nous écrivons ensuite la formule intégrale classique liée au magnétisme (Gauss) pour le calculer. Nous ferons alors un petit détour par la géométrie différentielle globale des courbes de l'espace tridimensionnel pour montrer la formule de White qui relie 3 invariants géométriques associés à un ruban. Nous décrivons ensuite les nouveaux invariants polynomiaux d'un point de vue combinatoire. Le point de vue intégrales de Feynman sera évoqué.

Vassiliev a introduit une famille générale d'invariants qui contient la plupart des invariants connus et dont nous pouvons dire qu'ils sont de type fini. Nous en décrivons le principe.

Remarque: Nous avons délibérément choisi de ne pas énumérer un certain nombre de définitions (aussi nombreuses qu'en théorie des graphes) que le lecteur pourra trouver facilement dans littérature ou sur Internet. Nous nous permettons de le omettre dans le sens qu'elles ne nous seront pas utiles dans l'application de la théorie des noeuds en physique quantique des champs (et que mis à part pour ceux qui aiment faire de petits dessins elles sont inutiles).

Définitions:

D1. Un noeud peut-être défini (car il existe plusieurs définitions possibles et certaines ont de petits aléas assez embêtants...) par l'image du cercle noté S^1 par une application continue (la ficelle étant supposée telle quelle), injective (ceci évitant que la ficelle rentre dans elle-même):

$$f : S^1 \rightarrow \mathbb{R}^3 \quad (??4)$$

autrement dit, c'est une courbe sans point double, tracée dans l'espace euclidien de dimension trois.

Un noeud est donc représenté par une application injective $f \in C^1([0,1], \mathbb{R}^3)$ (imposer des noeuds de classe C^1 permet d'éviter d'avoir des courbes trop... sauvages) vérifiant que $f(0) = f(1)$ (noeud fermé). L'image de f est parfois appelée "**support du noeud f** ": c'est la réalisation "physique" du noeud dans l'espace.

Remarque: L'ensemble des noeuds sera par la suite noté par la lettre N

D2. Nous disons qu'un noeud est un "**noeud trivial**" si l'application f qui le définit se prolonge en une application du disque $D^2 \rightarrow \mathbb{R}^3$ continue et toujours injective: un noeud trivial est donc un noeud qui bord un disque plongé de $\mathbb{R}^3$.

D3. Deux noeuds γ_1, γ_2 sont équivalents s'il existe une application continue:

$$F : [1, 2] \rightarrow N \quad (??5)$$

telle que $F(1) = \gamma_1$ et $F(2) = \gamma_2$. Reste donc à trouver F qui est une courbe dans l'espace des courbes.

Pour résumer grossièrement..., un noeud est une ficelle dont nous avons soudé les deux extrémités.

D3. Le $c(K)$ d'un noeud K est le nombre entier naturel représentant le nombre minimum de croisements pour tout diagramme d'un type de noeud (c'est une mesure naturelle de complexité).

Exemple:

Le noeud 0_1 à un $c(K) = c(0_1) = 0$ nul. Il n'existerait pas de noeuds avec un $c(K)$ est à un ou à l'unité.

Une preuve consiste à énumérer tous les diagrammes possibles avec un ou deux croisement et de voir que ceux qui sont au fait sont des noeuds équivalents de type 0_1 ou des entrelacs. Le trèfle (noeud 3_1) a un $c(K) = c(3_1) = 3$.

D4. L'image miroir $\bar{K}$ d'un noeud K est obtenue par réflexion sur un plan dans $\mathbb{R}^3$. Le noeud K peut-être construit par inversion des croisements du diagramme du noeud:



Figure: ??38 - Image miroir par inversion de croisements

Nous pouvons facilement nous en convaincre en prenant une réflexion en pliant la page de lecture ou encore mieux... en s'équipant du matériel adéquat et d'un miroir (...).

Exemple:

Avec le noeud 3_1 ("**trèfle gauche**" et "**trèfle droite**"):



Figure: 39 - Miroir du noeud de trèfle

Nous remarquons par ailleurs que dans la table de Tait, les noeuds miroirs distincts ne sont pas représentés!

D5. Deux noeuds sont dits "**noeuds isotopes**" si nous pouvons passer de l'un à l'autre par des manipulations continues (le trèfle gauche et le trèfle droit ne sont par exemple pas isotopes!) et il s'agit du problème numéro 1 de la théorie des noeuds: détecter les noeuds isotopes.

D6 Deux noeuds sont dits "**noeuds équivalents**" s'ils sont isotopes ou si l'un est isotope à l'image de l'autre dans un miroir. D'après ce qui précède, chaque noeud est donc forcément équivalent à sa propre image-miroir mais seuls les noeuds réflexifs sont isotopes à leur image dans un miroir. Le noeud en huit 4_1 est un bon exemple de ce genre de noeuds, qui sont par ailleurs assez rares:



Figure: 40 - Isotopie de noeuds

D7. Un "**entrelac**" est une sous-variété (cf. chapitre de Géométrie Différentielle) compacte (cf. chapitre de Topologie), de classe C^∞ et de dimension 1.

D8. Le "**nombre de composantes connexes**" est noté $c(E)$. Si $c(E) = 1$ nous disons que E est un noeud.

La plupart du temps les entrelacs seront orientés (cf. chapitre de Théorie Des Graphes), et nous identifierons les entrelacs isotopes. Nous représentons donc les entrelacs dans le plan en les projetant et en spécifiant le type de points de croisement.

Exemple:

Entrelacs à trois composantes connexes (à gauche) et noeud de trèfle (à droite).

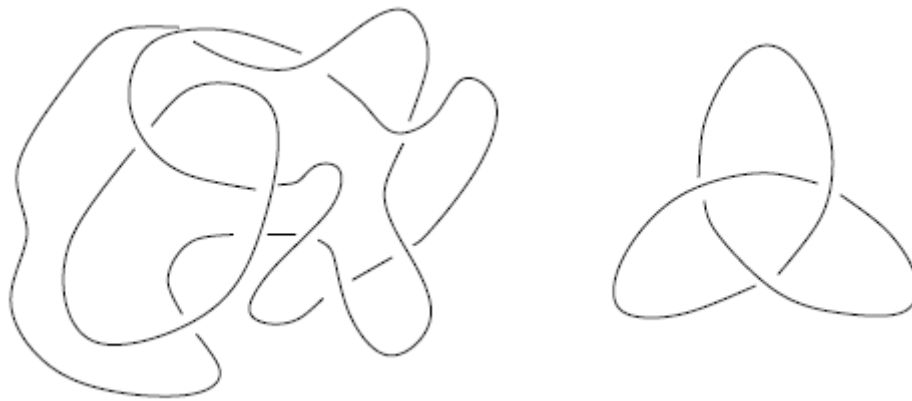


Figure: 41 - Entrelacs à trois composantes connexes

Les isotopies permettent de démêler les entrelacs.

Définition: Une "**isotopie**" est une opération qui laisse un noeud invariant. En d'autres termes, c'est un mouvement du noeud qui ne le déforme/change pas.

Elles donnent dans le plan trois types de mouvements particuliers appelés "**mouvements de Reidemeister**":

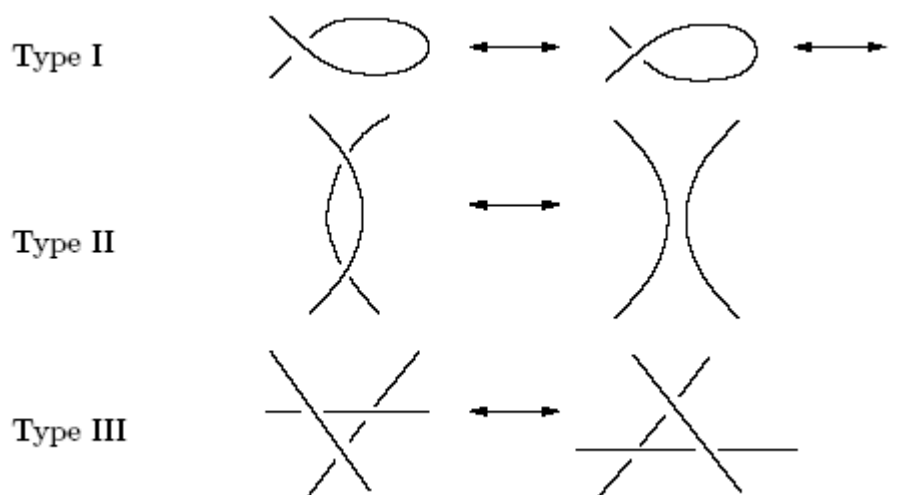


Figure: 42 - Représentation des mouvements de Reidemeister

Il s'agit donc bien de trois opérations simples permettant de changer une partie d'un noeud sans changer la nature du noeud lui-même.

Donc deux diagrammes représentent les mêmes noeuds, si nous pouvons passer de l'un à l'autre par une suite finie de mouvements de Reidemeister. Ainsi, avec l'exemple ci-dessous, nous montrons que le noeud initial, est équivalent au noeud trivial:

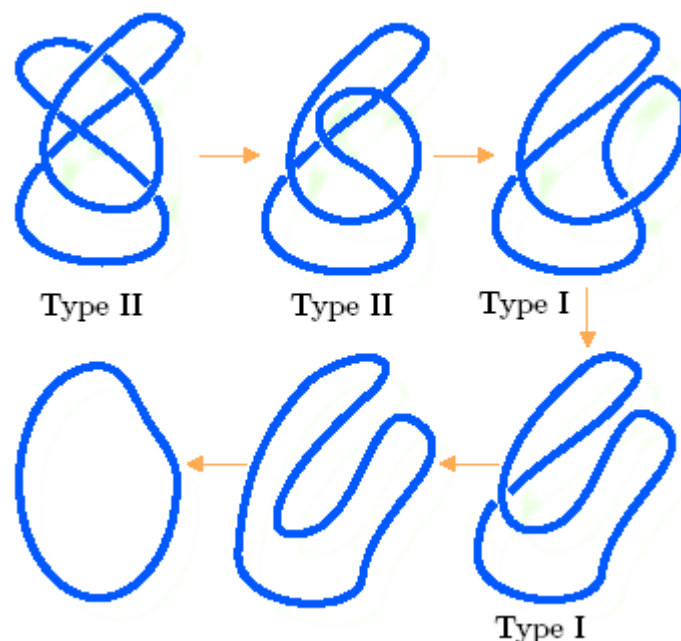


Figure: ??43 - Exemple simpliste d'application des mouvements de Reidemeister

En biologie les brins d'ARN et ADN ainsi que les filaments d'acides aminés s'enroulent selon des formes tridimensionnelles complexes (ce sont des tresses fermées: un cas plus général des noeuds). Or, souvent, au travers des microscopes, nous ne voyons qu'une projection bidimensionnelle. Les invariants permettent de remonter à des informations tridimensionnelles à partir des vues 2D que nous avons:

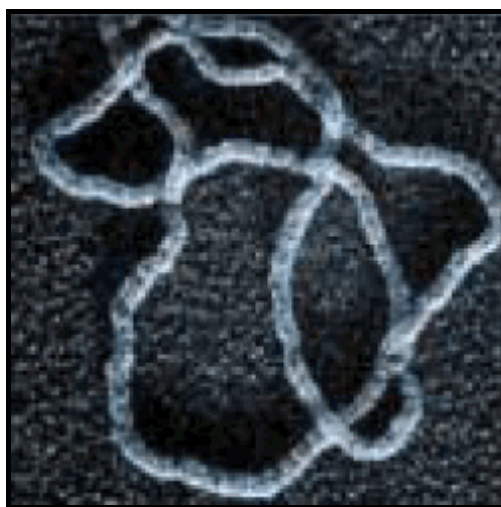


Figure: ??44 - Exemple de brin d'ARN ou d'ADN

D'autre part, les biologistes ont observé des molécules d'ADN nouées et ont constaté que la nature topologique de la molécule d'ADN, c'est-à-dire le type de noeud formé par la molécule, influe sur son fonctionnement dans les cellules en conditionnant certaines de ses propriétés chimiques. Les virus attaquent les cellules pour en changer les longues molécules d'ADN en les nouant de différentes façons. En effet, par le biais d'enzymes appelées les topoisomérases, les virus coupent et recollent différemment les brins de la molécule d'ADN de telle sorte qu'elles prennent la forme d'un noeud qui peut être très complexe. Il s'avère que le type de noeud obtenu est en quelque sorte la carte de visite du virus. Pour lutter efficacement contre les virus, il est impératif de reconnaître leur signature par leur action sur l'ADN. Par conséquent pour identifier les différents virus il faut pouvoir reconnaître les différents types de noeuds et c'est en cela que la théorie des noeuds peut aider le biologiste.

Remarque: L'identification du type de noeud en biologie moléculaire a été transformée par des centres de recherche en des jeux en ligne dont le but est de mettre la population à contribution de manière ludique pour trouver une solution au problème. Cela semble bien marcher puisqu'en fin 2011, des joueurs auraient fait une découverte pertinente concernant l'analyse de protéines.

Il existe également un théorème comme quoi deux entrelacs sont isotopes si et seulement si nous pouvons passer de l'un à l'autre par une suite finie de mouvements de Reidemeister.

Démonstration: À faire avec un lacet....

□C.Q.F.D.

D8. Trois entrelacs E_+, E_-, E_0 sont dits "**entrelacs associés**" (ou "**entrelacs en association**") s'ils ne diffèrent qu'en un point de croisement et qu'en ce point ils sont dans une des configurations suivantes:

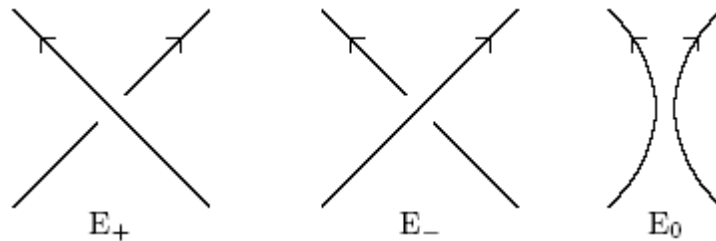


Figure: ??45 - Exemples d'entrelacs associés

Nous notons $\mathcal{E}$ l'ensemble des classes d'entrelacs et nous nous intéressons désormais à des fonctions $P: \mathcal{E} \rightarrow A$ où A sera un anneau commutatif (le lecteur n'oubliera pas que nous avons vu que les coefficients de polynômes sont des éléments d'un anneau).

Remarque: Il faut aussi se rappeler qu'un noeud est une courbe et que toute courbe peut être représentée par un polynôme d'où l'idée!

D9. P est dit "**invariant par association**" (par association des entrelacs...) si:

$$P(\bigcirc) = 1 \quad (?.?.6)$$

et s'il existe $a_+, a_-, a_0 \in A$ inversibles (!) tel que pour tout triplet d'entrelacs associées (E_+, E_-, E_0) nous ayons:

$$a_+ P(E_+) + a_- P(E_-) + a_0 P(E_0) = 0 \quad (?.?.7)$$

Nous pouvons déjà démontrer de façon assez élémentaire que si un tel polynôme existe, alors il est uniquement déterminé par les coefficients de la relation précédente. Nous résumons cela dans le théorème suivant: Si P est invariant par association, alors il est uniquement déterminé par les coefficients a_+, a_-, a_0 .

Démonstration:

Remarquons d'abord que:

$$\forall r \in \mathbb{N} \quad P(\bigcirc^r) = \left(-\frac{a_+ + a_-}{a_0} \right)^{r-1} \quad (??.8)$$

où $\bigcirc^r$ désigne le noeud composé de r cercle démêlés. En effet, la relation $P(\bigcirc) = 1$ et la propriété d'invariance de P appliquée aux entrelacs suivants:

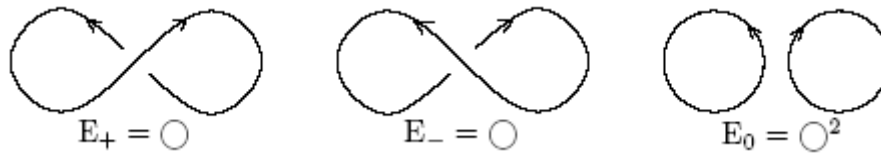


Figure: ??46 - Exemple de notation des démêlés

donne:

$$a_+ P(\bigcirc) + a_- P(\bigcirc) + a_0 P(\bigcirc^2) = 0 \quad (??.9)$$

Nous obtenons alors:

$$P(\bigcirc^2) = -\frac{a_+ P(\bigcirc) + a_- P(\bigcirc)}{a_0} = -\frac{a_+ P(\bigcirc) \cdot 1 + a_- \cdot 1}{a_0} = \left(-\frac{a_+ + a_-}{a_0} \right)^{2-1} \quad (??.10)$$

Ainsi, par récurrence sur r nous obtenons:

$$P(\bigcirc^r) = \left(-\frac{a_+ + a_-}{a_0} \right)^{r-1} \quad (??.11)$$

□C.Q.F.D.

En d'autres termes, la fonction P peut s'exprimer uniquement par ses coefficients !!! Donc nous pourrions maintenant essayer de voir si un noeud plein d'entrelacs peut être toujours se ramener à des $P(\bigcirc)$ de manière récursive.

Remarque: Il fallait y penser cependant....

4.1. REPRÉSENTATION PLANAIRE

Nous supposons que le noeud est dans l'espace euclidien de dimension trois orienté. Nous nous intéressons aux projections de ce noeud sur le plan $z = 0$. Il est claire intuitivement que nous pouvons supposer que noeud n'a aucune tangente verticale et que les points de croisement sont seulement doubles et transversaux. Une telle projection sera appelée "**bonne projection**".

Nous indiquons ensuite à chaque point de croisement quel est le brin qui passe au-dessus de l'autre. Un tel dessin représente un noeud de façon non ambiguë: nous disons que nous avons un "**diagramme de noeud**". Bien sûr, deux noeuds équivalents ont des bonnes projections qui donnent des diagrammes de noeuds différents en général (c'est là que réside une des difficultés aussi) Il est donc important de pouvoir lire l'équivalence de deux noeuds directement sur leurs diagrammes. Un diagramme du noeud du trèfle droit 3_1 :

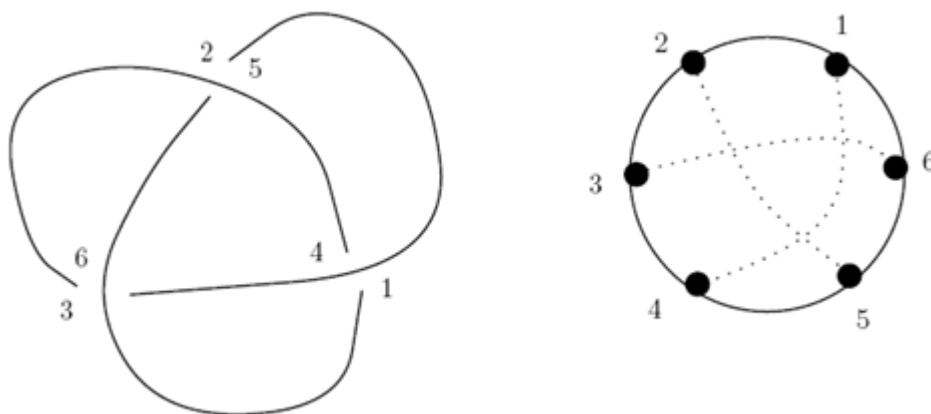


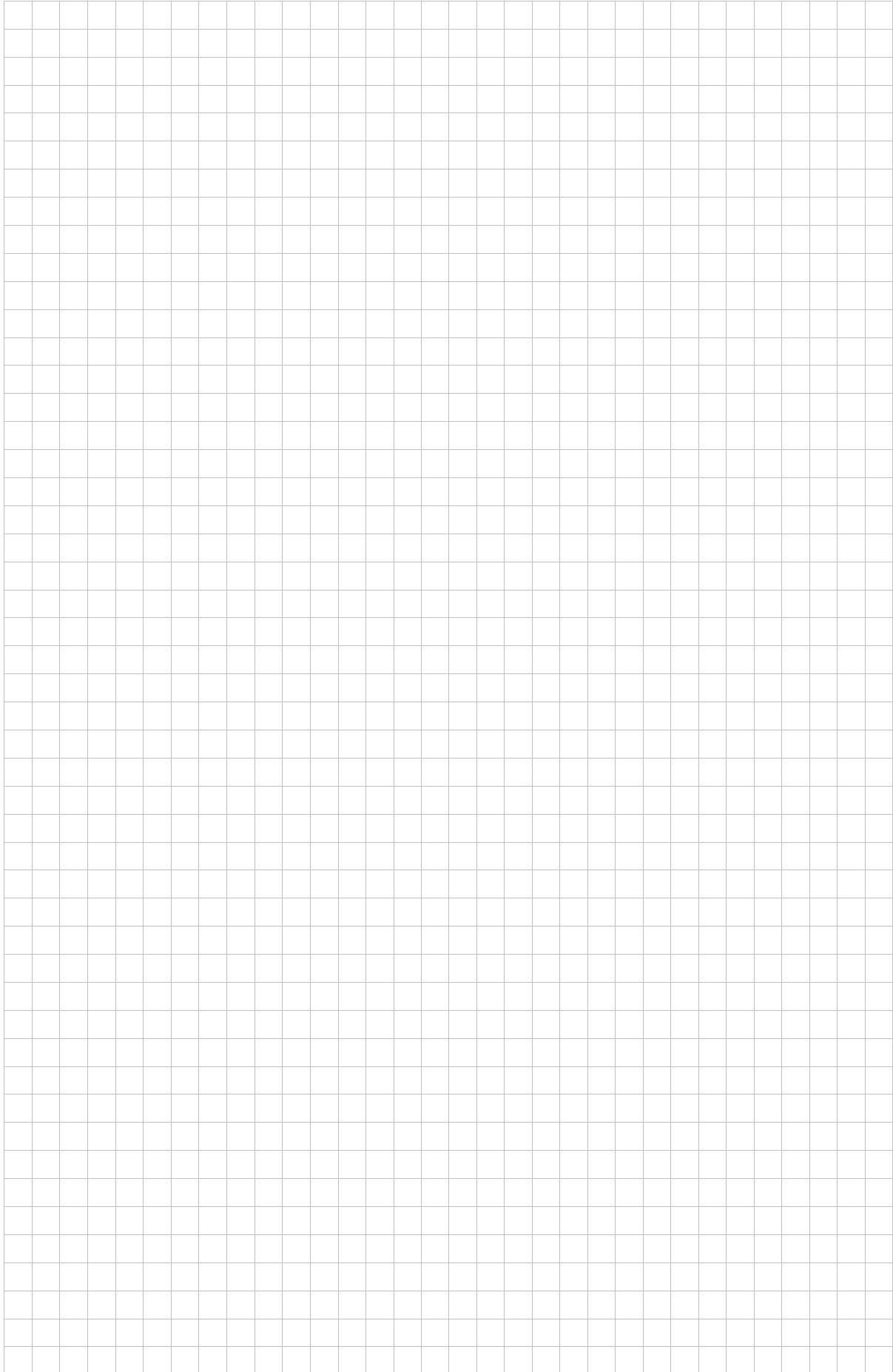
Figure: ??47 - Représentation d'un diagramme de noeud

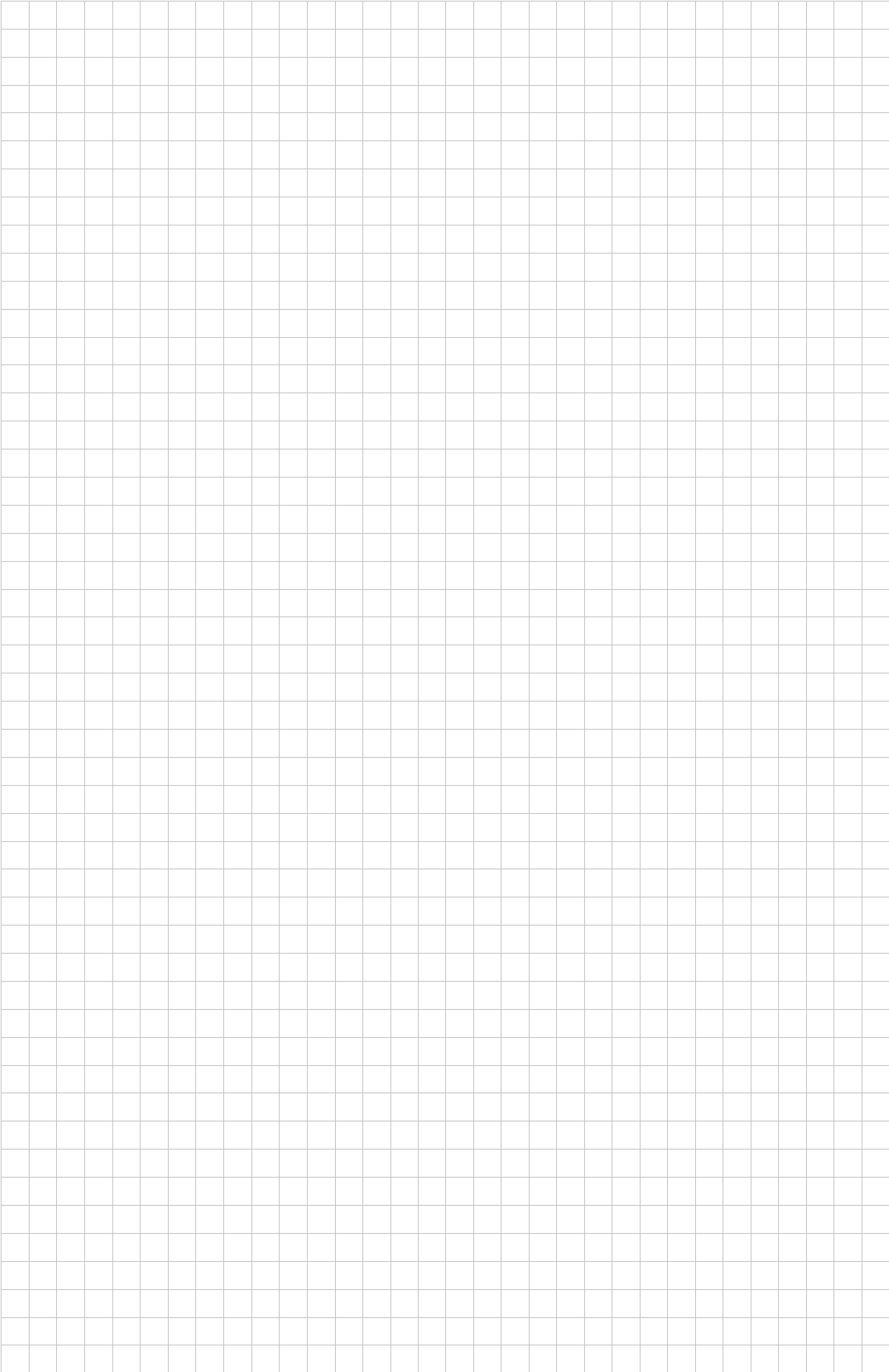
Remarque: Le noeud de trèfle est le noeud ayant le nombre minimal de croisements, à savoir 3 ; il en existe en fait deux, énantiomorphes (images l'un de l'autre par réflexion). Il s'agit au fait d'un simple noeud dont on a soudé les extrémités. Le noeud de trèfle est enfin le bord d'un ruban de Möbius à 3 demi-torsions ainsi que le noeud torique d'ordre (3,2) (3 enroulements autour du tore, sur deux tours), ainsi que celui d'ordre (2,3).

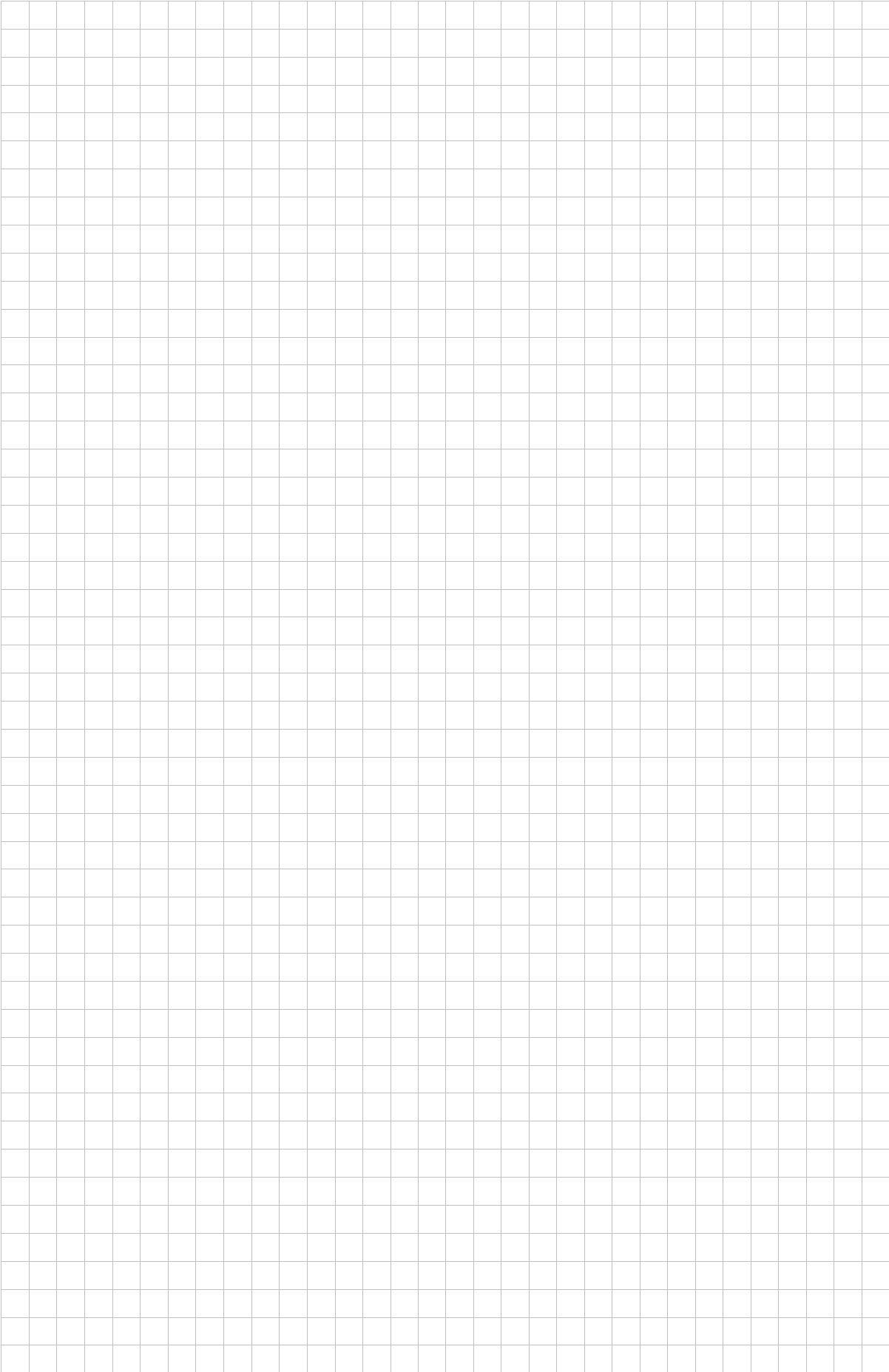
Comme un dessin n'est pas forcément facile à transmettre, ni à mettre dans un ordinateur, nous pouvons aussi donner un codage du diagramme du noeud par une matrice à coefficients entiers à trois lignes et dont le nombre de colonnes est égal au nombre de points doubles.

Nous numérotions les $2n$ points de croisement sur la courbe fermée (cercle) dans l'ordre où ils arrivent. Nous remarquons que les paires sont toutes formées d'un nombre pair et d'un nombre impair. Nous fabriquons alors les deux premières lignes de la matrice en mettant dans chaque colonne les deux numéros donnant la même projection: les impaires sur la première ligne, les pairs sur la seconde. Nous ajoutons alors à chaque colonne, un ± 1 qui indique l'orientation des 2 brins (orienté) (+1 si celui de dessous traverse de droite à gauche quand nous parcourons celui de dessus, -1 sinon). Par exemple, la matrice associée au noeud de trèfle de la figure précédente est:

$$3_1 = \begin{bmatrix} 1 & 3 & 5 \\ 4 & 6 & 2 \\ + & + & + \end{bmatrix} \quad (??12)$$









Analyse

L'analyse est formulation rigoureuse du calcul différentiel et intégral. (Wikipedia)

16. ANALYSE FONCTIONNELLE

L'analyse fonctionnelle est la branche des mathématiques et plus particulièrement de l'analyse qui est en rapport avec l'étude des espaces de fonctions. Elle prend ses racines historiques dans l'étude des transformations telles que la transformation de Fourier et dans l'étude des équations différentielles et des intégrales. À ce titre elle englobe tellement de domaines qu'il est difficile de justifier qu'elle fasse l'objet d'un chapitre car il s'agit plutôt d'un domaine d'études. Par ailleurs, c'est à cause de cette difficulté de cerner exactement le domaine qu'elle touche que le lecteur trouvera le théorème fondamental de l'analyse dans le chapitre de Calcul Différentiel Et Intégral plutôt qu'ici.

Pourquoi ceci dit utilisons-nous le terme "**analyse**" dans le cas particulier des fonctions. La raison tient historiquement à l'étude des divers phénomènes de la nature et la résolution de divers problèmes techniques et par conséquent de mathématiques, qui nous amènent souvent à considérer la variation d'une grandeur en corrélation avec la variation d'une autre ou de plusieurs autres grandeurs. Pour étudier ces variations, de nombreux outils sont à la disposition de tout à chacun:

- L'ingénieur a par exemple fréquemment recours à la représentation graphique (système d'axes cartésien, polaire, logarithmique... concepts sur lesquels nous reviendrons plus en détail) pour déterminer la relation (ou "loi") mathématique qui lie les différentes grandeurs entre elles. Certes, ce genre de méthode est (parfois...) esthétique mais les étudiants savent bien combien il est parfois pénible en laboratoire de devoir porter des points sur une feuille de papier ou à l'ordinateur. C'est malheureusement une étape nécessaire (mais dont il faudrait éviter de faire une utilisation abusive) pour comprendre comment nos prédécesseurs travaillaient et ont obtenu les résultats qui nous aident aujourd'hui dans nos avancées en physique théorique.
- Le mathématicien et le physicien théoricien ont habituellement horreur d'avoir recours aux méthodes papier-crayon-gribouillage. Quoi qu'il en soit, le rôle du mathématicien ou du physicien est de développer de nouvelles théories à l'aide d'axiomes ou de principes mathématiques ce qui ne devrait nécessiter aucunement le recours à la représentation graphique et à l'accès aux mesures expérimentales qui y sont souvent rattachées.

Remarque: Avant de commencer la lecture de ce qui va suivre, il peut être utile de rappeler au lecteur que la définition du concept de "fonction" (et les propriétés élémentaires y relatives) est donné dans le chapitre de Théorie Des Ensembles.

1. REPRÉSENTATIONS

Nous allons voir dans ce qui va suivre, dans un premier temps, comment représenter différentes grandeurs liées de façon tabulaire et graphique (eh oui! il faut bien car cela aide à comprendre) et ensuite comment analyser mathématiquement les propriétés de ces représentations uniquement à l'aide d'outils mathématiques abstraits.

Définition: Une fonction est dite "**fonction univalente**", si le nombre de ses arguments (paramètres ou variables) est égal à un. Dans le cas d'une fonction à deux arguments, nous parlons de "**fonction bivalente**", etc.

1.1. REPRÉSENTATION TABULAIRE

Parmi les modes de représentation visuelle des fonctions, la plus intuitive et la plus ancienne est celle où nous disposons dans la colonne ou la ligne d'un tableau de façon ordonnée les valeurs de la variable indépendante $x_1, x_2, \dots, x_n$ et les valeurs correspondantes, dites "**variables transformées**" de la fonction $y_1, y_2, \dots, y_n$ dans une autre colonne ou ligne alignée.

Telles sont par exemple, les tables des fonctions trigonométriques, les tables logarithmiques, etc. et au cours de l'étude expérimentale de certains phénomènes des tables qui expriment la dépendance fonctionnelle existant entre des grandeurs physiques mesurées telles que les relevés de la température de l'air enregistrés dans une station météorologique durant une journée.

Bien évidemment, ce concept est généralisable à toute fonction multivalente quel que soit son ensemble de définition.

Cependant, cette méthode est laborieuse et ne permet pas de voir directement le comportement de la fonction et donc une analyse visuelle simple et intéressante de ses propriétés. Elle a pour avantage quand même de ne pas nécessiter d'outils spéciaux ou de connaissances mathématiques poussées.

1.2. REPRÉSENTATIONS GRAPHIQUES

Les nombres naturels, relatifs, réels ou purement complexes (cf. chapitre sur les Nombres) peuvent tous être représentés le plus simplement du monde par des points sur un axe numérique (ligne droite) infini.

Pour ce faire, nous choisissons sur cet axe:

1. Un point O appelé "**origine**"
2. Un sens positif, que nous indiquons par une flèche horizontale
3. Une unité de mesure (représentée habituellement par un petit trait vertical: la "**graduation**")

Tel que:

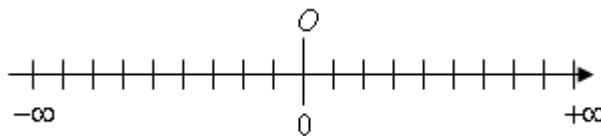


Figure: 16.1 - Exemple type de représentation d'un axe infini orienté avec origine

Le plus souvent nous disposons (par tradition) l'axe horizontalement et choisissons la direction de gauche à droite.

Remarque: Le point (lettre) O , représente très fréquemment le nombre zéro en mathématiques mais nous pourrions très bien choisir de mettre l'origine ailleurs. Par exemple, en physique le point O est souvent positionné à l'emplacement du barycentre d'un système.

Il est évident que le fait que les ensembles de nombres dont nous avons parlé sont ordonnés implique que tout nombre est représenté par un seul point de l'axe numérique. Ainsi, à deux nombres réels

distincts correspondent deux points différents de l'axe numérique.

Ainsi, il existe une correspondance biunivoque entre tous les nombres et tous les points de l'axe numérique (dans le cas des nombres réels ou complexes, il correspond non pas un nombre à chaque graduation mais un nombre à chaque point de l'axe). Ainsi, à chaque nombre correspond un point ou une graduation unique et inversement à chaque point ou graduation correspond un seul nombre dont il est l'image.

1.2.1. REPRÉSENTATIONS PLANES

Il existe outre les représentations unidimensionnelles d'autres de dimensions supérieures (ouf!) qui nous permettent de tracer non plus que des simples points sur une droite unidimensionnelle mais des fonctions d'une variable. Voyons de quoi il s'agit:

Nous pouvons à chaque valeur d'une variable x reportée sur un axe horizontal, appelé "**axe des abscisses**" ou "**axe des x** ", faire correspondre une valeur y au travers d'une fonction f :

$$y = f(x) \quad (16.1)$$

reportée sur un axe vertical, appelé "**axe des ordonnées**" ou "**axe des y** " qui passe par le croisement défini par l'origine O tel que (exemple arbitraire):

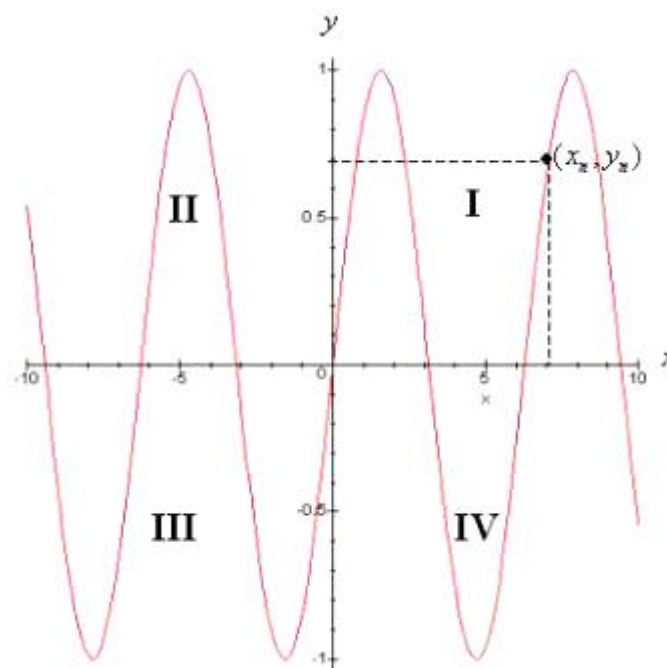


Figure: 16.2 - Exemple type d'une représentation plane avec axes orthogonaux, origine et quadrants

L'ensemble des points du plan noté sous les variantes XOY , XY ou encore xOy , Oxy , xy , dont les abscisses représentent par tradition les valeurs de la variable indépendante et les ordonnées les valeurs correspondantes de la fonction, est appelé "**graphique plan**" de cette fonction. S'il n'y a pas de confusion possible, nous dirons simplement "**graphique**".

Dans le cas d'une représentation par un système de coordonnées rectangulaires (cartésien, polaire ou logarithmique) comme la figure ci-dessus, nous pouvons observer que l'ensemble du plan des coordonnées est séparé en quatre surfaces que nous avons pour habitude d'appeler "**quadrants**".

Remarque: Lorsque nous souhaitons mettre en évidence un point particulier de la fonction représentée, nous y dessinons un petit rond tel que présenté ci-dessus.

Un autre cas classique de représentation graphique plane connu par un grand nombre d'étudiants est le tracé des polynômes (cf. chapitre de Calcul Algébrique) à coefficients réels.

Effectivement, pour résoudre les équations polynômiales du second degré (cf. chapitre de Calcul Algébrique), il est fréquent dans les petites classes que le professeur demande en plus à ses élèves de donner une expression algébrique des racines de:

$$y = f(x) = ax^2 + bx + c = 0 \quad (16.2)$$

données par, rappelons-le:

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad (16.3)$$

une résolution graphique où les deux racines (dans le cas où il y en a deux distinctes réelles) sont données par l'intersection de la parabole avec l'axe des abscisses (bien évidemment, si l'équation n'a pas de solutions, il n'y a pas d'intersections...):

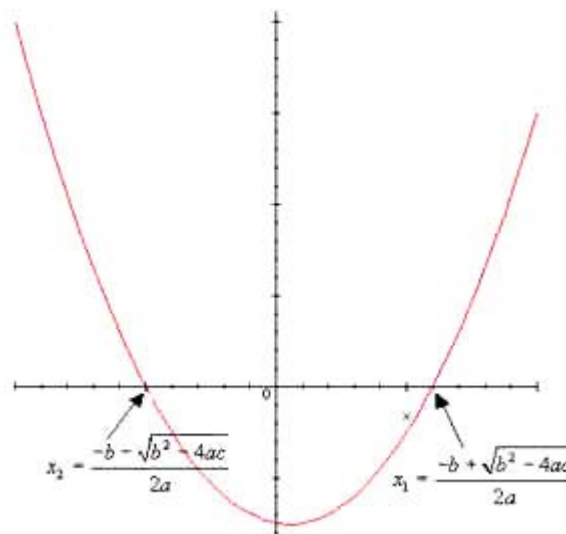


Figure: 16.3 - Représentation des racines sur un graphe planaire

La représentation graphique étant généralisable aux équations polynomiales du 3ème, 4ème et 5ème degré (nous démontrerons bien plus loin, à l'aide de la théorie de Galois qu'il n'est pas possible d'obtenir une expression algébrique générale des racines d'une équation polynomiale du 5ème degré et supérieur).

De même, les graphiques sont un outil qualitatif puissant dans le domaine des statistiques (cf. chapitre de Statistiques) comme point de départ de l'analyse de données (histogrammes, fromages, boîtes à moustaches, radars, nuages de points,...). Les hypothèses et idées qui sont générées par l'analyse graphique peuvent être investiguées avec des outils statistiques avancés.

Voici par exemple un graphique (histogramme) pris du chapitre de Génie Industriel très courant dans le domaine des statistiques et de la gestion de projets dans l'industrie mondiale:

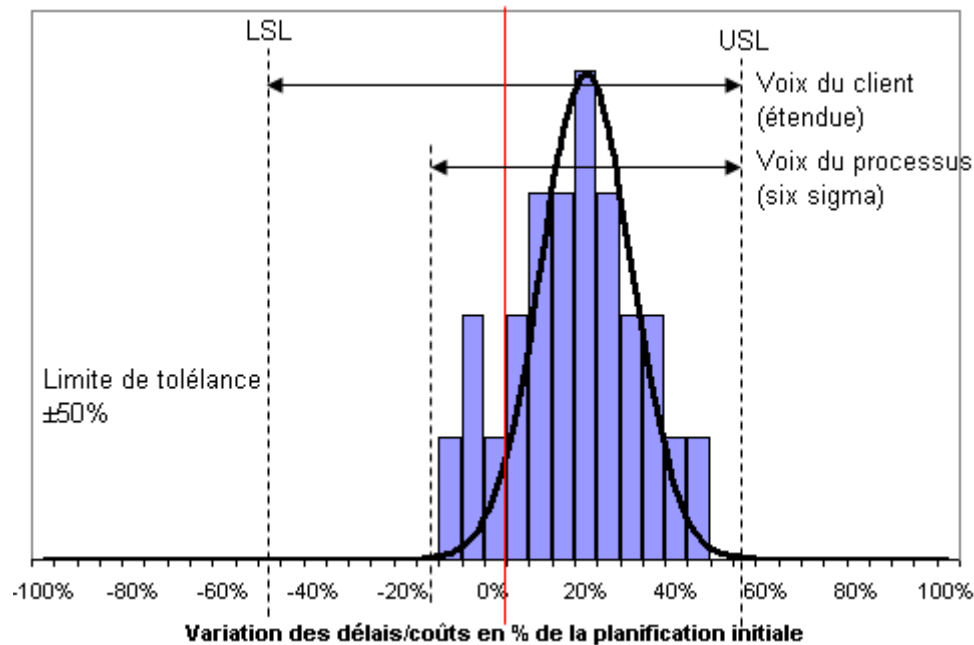


Figure: 16.4 - Exemple d'histogramme typique dans les entreprises d'ingénierie (Six Sigma)

Les histogrammes permettent d'observer les distributions et de décider de manière qualitative si elle s'ajuste à un modèle théorique particulier.

Les graphiques peuvent permettre également d'observer les changements au cours du temps de (séries temporelles, cartes de contrôle):

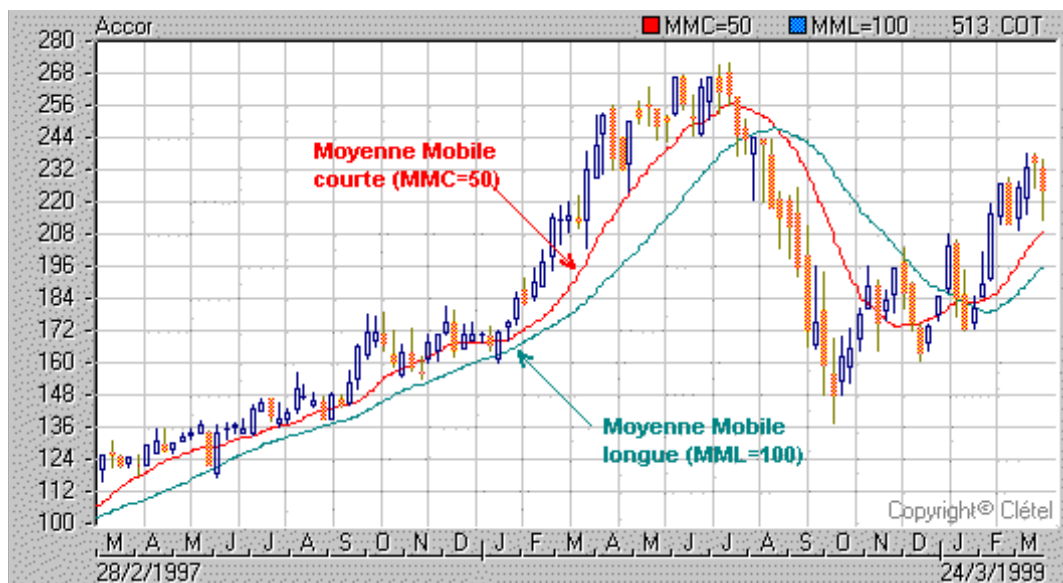


Figure: 16.5 - Exemple de série temporelle avec des moyennes mobiles dans la négociation financière

et encore à bien d'autres choses... que nous verrons tout au long des pages de ce site Internet.

1.2.2. REPRÉSENTATIONS 3D

Bien évidemment, dans le cas d'une fonction trivalente (tridimensionnelle), c'est-à-dire dont un paramètre dépend de deux autres, le principe reste le même à la différence que le nombre de quadrants double.

Cette méthode de représentation et d'analyse d'une fonction trivalente était longue à mettre en place il y a une dizaine d'années mais avec l'aide des ordinateurs en ce 21ème siècle ce problème (de temps) est assez bien résolu...

Ce type de représentation est suffisamment important en physique appliquée pour que nous nous y arrêtions un instant en faisant des exemples typiques sur plusieurs pages, des commandes les plus importantes avec Maple (même s'il existe de nombreux ouvrages sur le sujet c'est trop important pour que nous omettions ces exemples).

Ce que nous allons représenter, les mathématiciens puristes le noteraient formellement d'abord de la manière suivante (c'est bien d'avoir vu au moins une fois cette notation car vous pourriez la rencontrer dans d'autres ouvrages):

$$f : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$$

$$z = f(x, y) = \frac{12x}{(1 + x^2 + y^2)} \quad (16.4)$$

$$(x, y) \in [-10, +10] \times [-5, +5]$$

et voyons ce que cela donne donc avec Maple:

```
> restart;
> with(plots):
```

Nous prenons une fonction 3D quelconque:

```
> f:=(x,y)->12*x/(1+x^2+y^2);
```

Nous définissons le domaine d'analyse:

```
> xrange:=-10..10; yrange:=-5..5;
```

et nous faisons un plot simple:

```
> plot3d(f,xrange,yrange);
```

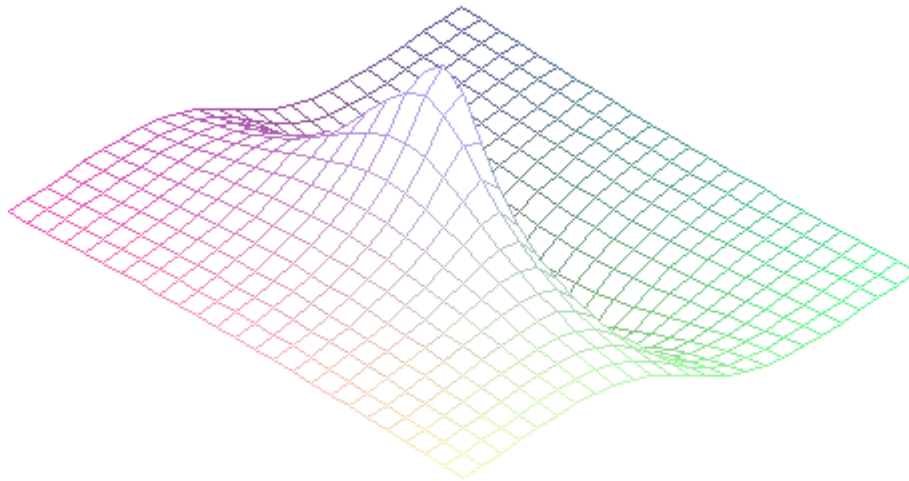



Figure: 16.6 - Exemple de représentation filiaire d'une fonction

Améliorons un peu l'aspect:

```
> plot3d(f,xrange,yrange, style=patchnogrid, grid=[80,50], shading=ZHUE, axes=FRAME, tickmarks=[3,3,3], labels=['x','y','f(x,y)'], labelfont=[TIMES,BOLD,12], title='Graphique rempli', titlefont=[TIMES,BOLD,12], scaling=unconstrained, orientation=[-107,68]);
```

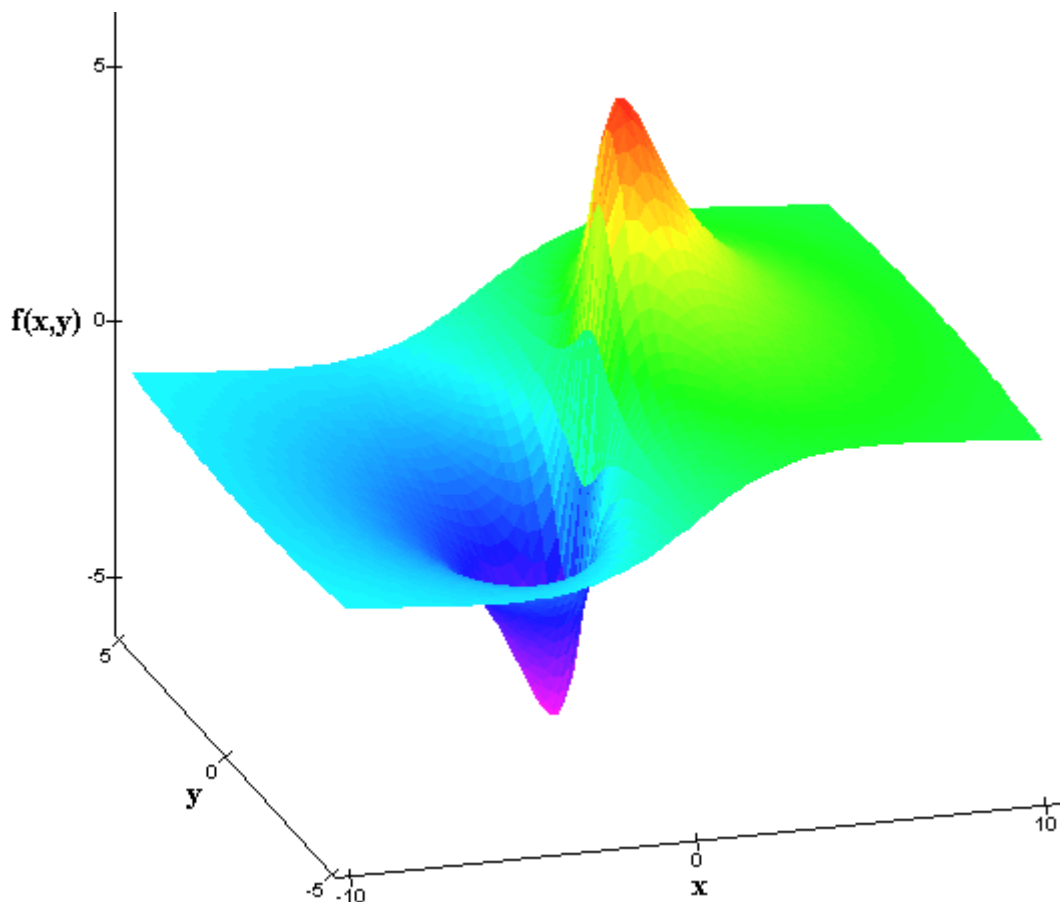


Figure: 16.7 - Représentation colorisée (gradients) d'une fonction

Traçons les courbes de niveau (cf. chapitre de Géométrie différentielle):

```
> plot3d(f,xrange,yrange,style=patchcontour);
```

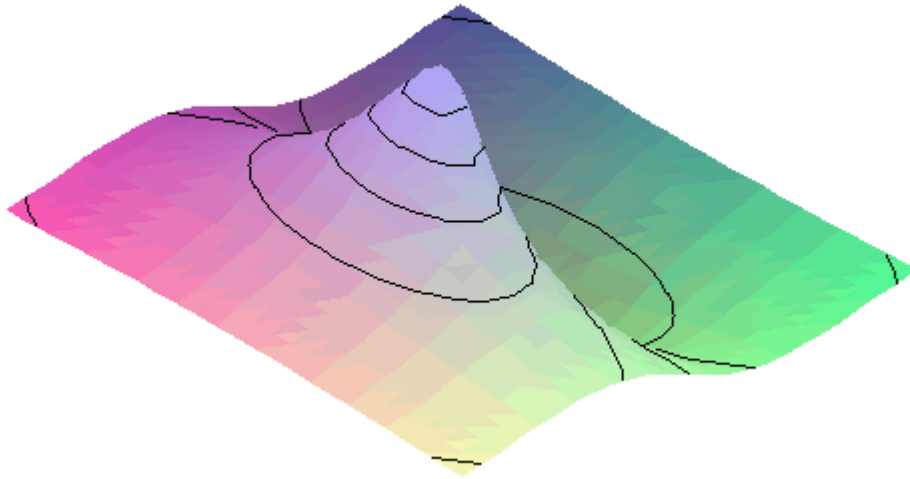


Figure: 16.8 - Représentation des courbes de niveau (isoclines) d'une fonction

Ce n'est pas très beau donc améliorons cela:

```
> plot3d(f,xrange,yrange,style=patchcontour,contours=[seq(-7+k/4,k=0..60)],grid=[80,50],shading=ZHUE,axes=FRAME, tickmarks=[3,3,3], scaling=unconstrained,orientation=[-107,68]);
```

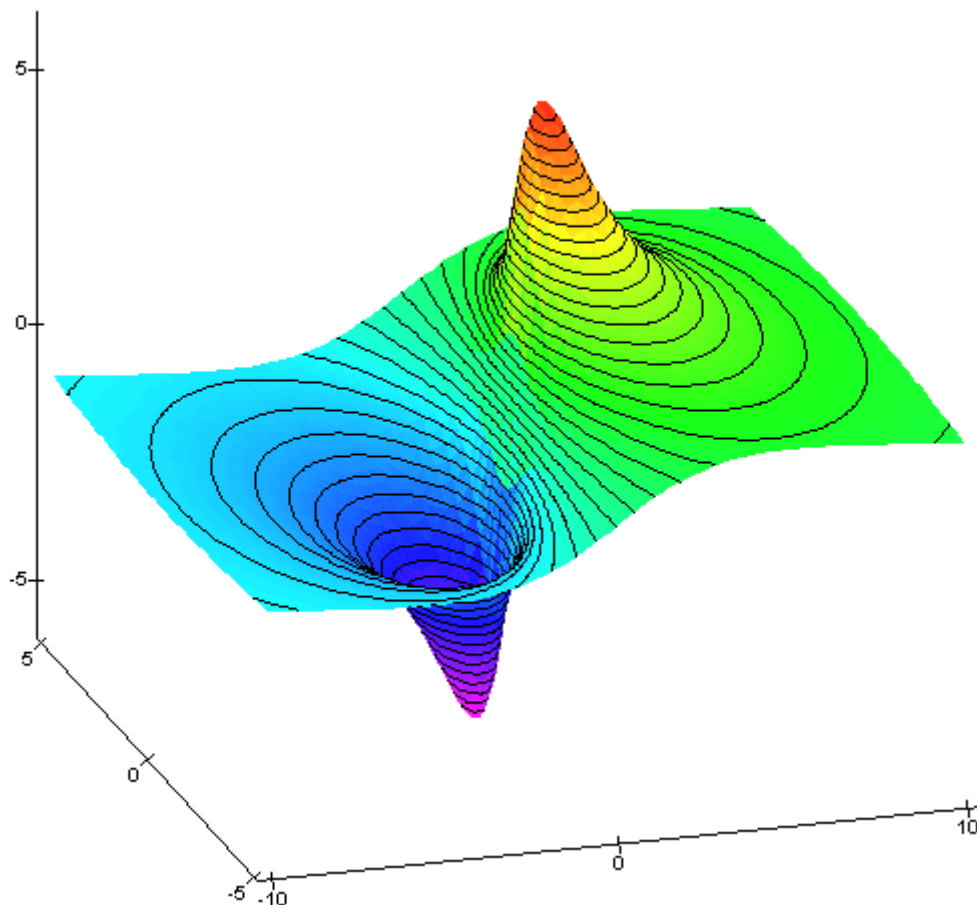


Figure: 16.9 - Représentation des gradients et isoclines d'une fonction

Avec une petite rotation pour voir du dessus:

```
> plot3d(f,xrange,yrange, style=patchcontour, contours=[seq(-7+k/4,k=0..60)], grid=[80,50],  
shading=ZHUE, axes=FRAME, tickmarks=[3,3,3], scaling=unconstrained, orientation=[-90,0]);
```

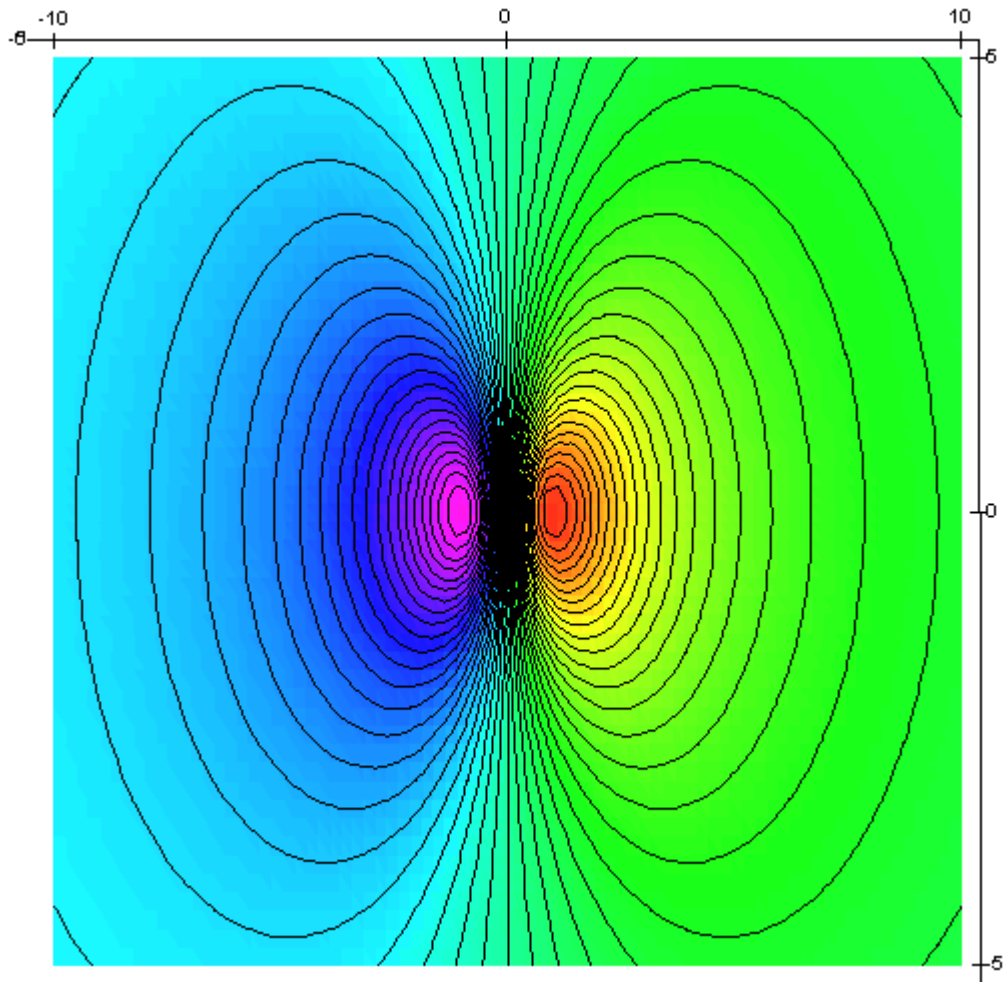


Figure: 16.10 - Représentation plane des isoclines et des gradients

Et en coupe:

```
> plot(f(x,2),x=xrange);
```

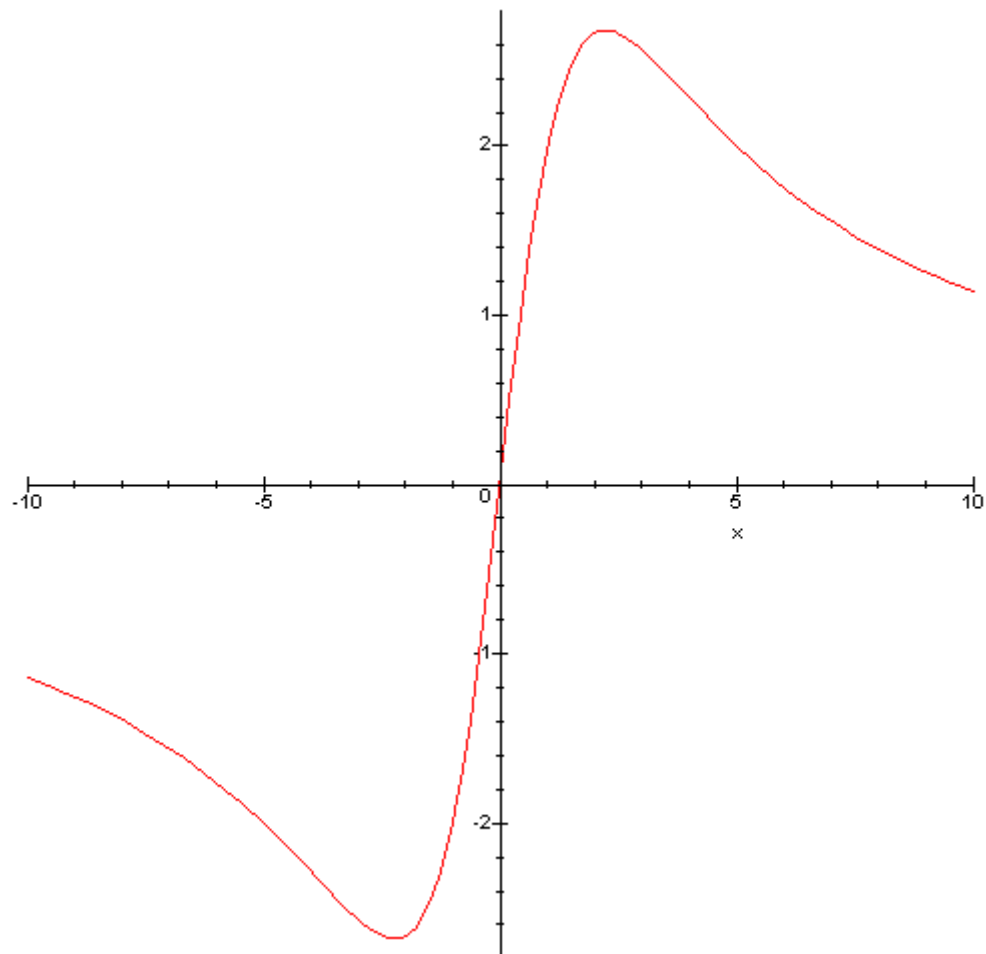


Figure: 16.11 - Représentation en coupe d'une fonction

Ou avec des coupes multiples:

```
>display([seq(plot(f(x,y),x=xrange),y=yrange) ]);
```

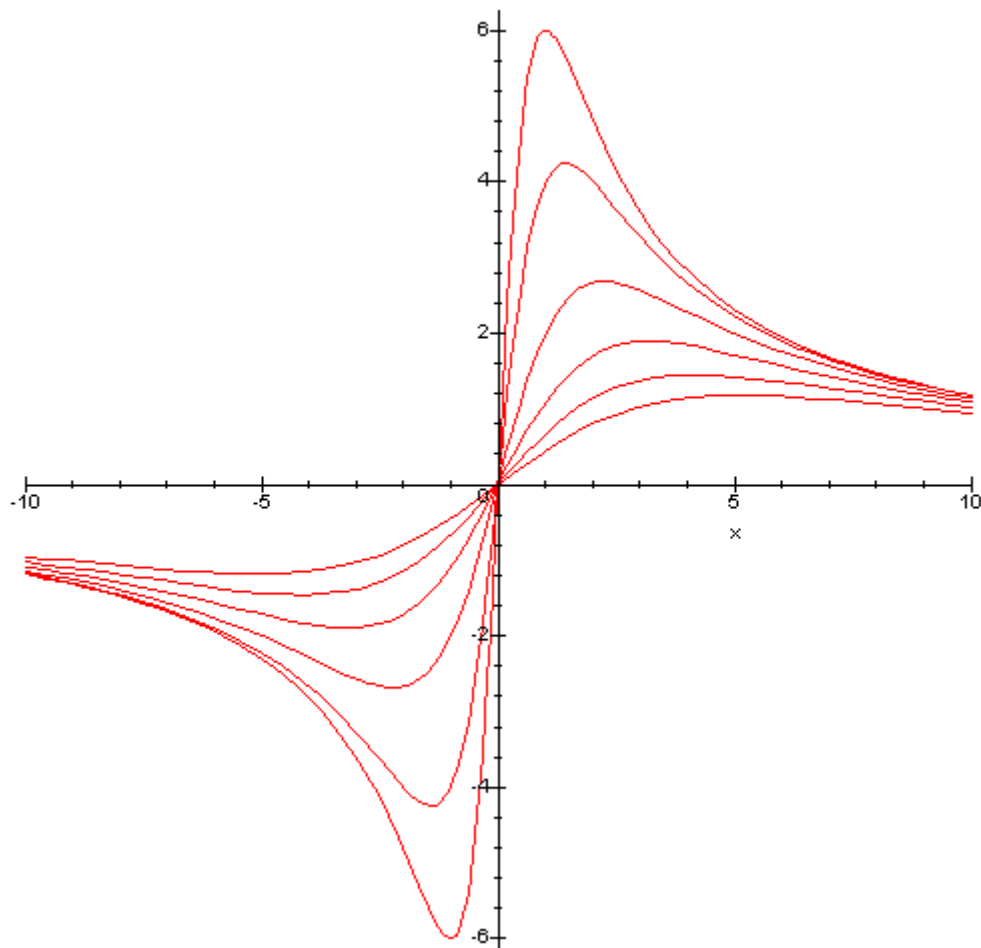


Figure: 16.12 - Représentation de coupes multiples d'une fonction

Le lecteur pourra aussi animer le précédent graphique avec la commande suivante:

```
> display([seq(display([plot(f(x,k/5),x=xrange),
textplot([6,5,cat(`y=`,convert(evalf(k/5,2),string))],font=
[TIMES,BOLD,16]))],k=-25..25)],insequence=true, title=`Animation`,titlefont=[TIMES,BOLD,18]);
```

voilà pour un exemple typiquement simple des manipulations standard d'un ingénieur dans l'entreprise utilisant des graphiques.

1.2.3. REPRÉSENTATIONS VECTORIELLES

Il est aussi fréquemment fait usage des représentations graphiques dans le cadre de la géométrie analytique pour simplifier les analyses ou faire des démonstrations de théorèmes connus sous forme visuelle (il faut cependant ne pas en abuser!).

Ainsi, nous pouvons introduire par exemple le concept de norme (cf. chapitre de Calcul Vectoriel) de manière simpliste en représentant graphiquement la distance entre deux points et en appliquant le théorème de Pythagore (cf. chapitre de Géométrie Euclidienne) qui sera supposé connu.

Ainsi, représentons trois points P_1, P_2, P_3 sur un graphique plan dans lequel a été défini un repère tel que présenté dans le graphique ci-dessous:

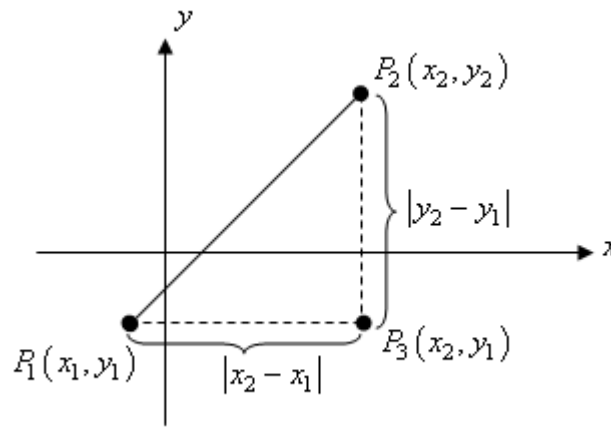


Figure: 16.13 - Mise en situation de 3 points dans un plan

Si $x_1 \neq x_2$ et $y_1 \neq y_2$ (comme sur la figure ci-dessus), les points P_1, P_2, P_3 sont les sommets d'un triangle rectangle. Par application du théorème de Pythagore (cf. chapitre de Géométrie Euclidienne):

$$[d(P_1, P_2)]^2 = [d(P_1, P_3)]^2 + [d(P_3, P_2)]^2 \quad (16.5)$$

Sur la figure, nous voyons que:

$$d(P_1, P_3) = |x_2 - x_1| \text{ et } d(P_3, P_2) = |y_2 - y_1| \quad (16.6)$$

Puisque $|x|^2 = x^2, \forall x \in \mathbb{R}$, nous pouvons écrire:

$$d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (16.7)$$

Si $x_1 = y_1 = 0$, nous nous retrouvons avec une relation appelée "**norme**", "**module**" ou encore "**distance**" que nous avons déjà définie dans le cadre de notre étude de l'analyse vectorielle (cf. chapitre de Calcul Vectoriel).

Bien évidemment, si nous considérons deux points $P_1(x_1, y_1), P_2(x_2, y_2)$, nous pouvons déterminer si un troisième point $P_3(x_3, y_3)$ est sur la médiatrice (cf. chapitre de Géométrie Euclidienne) des deux premiers et qu'il suffit pour cela que bien évidemment (par définition même de la médiatrice):

$$d(P_1, P_3) = d(P_2, P_3) \quad (16.8)$$

Comme $(x_1, y_1), (x_2, y_2)$ sont connus, nous pouvons facilement exprimer une "**expression analytique**" de la médiatrice du type:

$$\begin{aligned}
\sqrt{(x_1 - x_3)^2 + (y_1 - y_3)^2} &= \sqrt{(x_2 - x_3)^2 + (y_2 - y_3)^2} \\
(x_1 - x_3)^2 + (y_1 - y_3)^2 &= (x_2 - x_3)^2 + (y_2 - y_3)^2 \\
x_1^2 - 2x_1x_3 + \cancel{x_3^2} + y_1^2 - 2y_1y_3 + \cancel{y_3^2} &= x_2^2 - 2x_2x_3 + \cancel{x_3^2} + y_2^2 - 2y_2y_3 + \cancel{y_3^2} \\
2y_2y_3 - 2y_1y_3 &= x_2^2 - 2x_2x_3 + y_2^2 - y_1^2 + 2x_1x_3 - x_1^2 \\
y_3 2(y_2 - y_1) &= 2(x_1 - x_2)x_3 + (x_2^2 - x_1^2 + y_2^2 - y_1^2) \\
y_3 &= \frac{(x_1 - x_2)}{(y_2 - y_1)} x_3 + \frac{(x_2^2 - x_1^2 + y_2^2 - y_1^2)}{2(y_2 - y_1)} \\
y_3 &= ax_3 + b = f(x_3)
\end{aligned} \tag{16.9}$$

où a, b sont des constantes et où tout point qui satisfait cette relation, qui est en l'occurrence l'équation d'une droite, se trouve sur la médiatrice.

Par ailleurs, il est aisé de visualiser que le point milieu du segment de droite est donné par:

$$M = \left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2} \right) \tag{16.10}$$

Donc nous voyons qu'avec une simple représentation graphique, nous pouvons obtenir des résultats qui sont parfois (...) plus évidents pour les étudiants.

Profitons de cet exemple pour définir quelques concepts sur lesquels nous reviendrons et faire quelques rappels.

Définition: Toute fonction de la forme d'un polynôme (cf. chapitre de Calcul Algébrique) de degré 1 à coefficients réels constants:

$$f(x) = y = ax + b \tag{16.11}$$

est l'expression analytique de ce que nous appelons une "**droite**" de " **pente**" a et "**d'ordonnée à l'origine**" b (quand $x = 0$).

Bien évidemment, si:

$$f(x) = y = b \tag{16.12}$$

la droite est horizontale si nous la représentons graphiquement puisque y est constant pour tout x et vaut alors b . Inversement, si:

$$f(y) = x = a \tag{16.13}$$

la droite est une verticale.

1.2.4. PROPRIÉTÉS DES REPRÉSENTATIONS GRAPHIQUES

Selon le type de graphique que nous visualisons (en particulier les graphiques plans) il est possible d'extraire certaines propriétés de base. Voyons les plus importantes à connaître pour les graphiques plans d'une fonction à une variable:

$$y = f(x) \quad (16.14)$$

P1. Le graphique d'une fonction est symétrique par rapport à l'axe des ordonnées si le changement de x en $-x$ ne modifie pas la valeur de l'équation tel que:

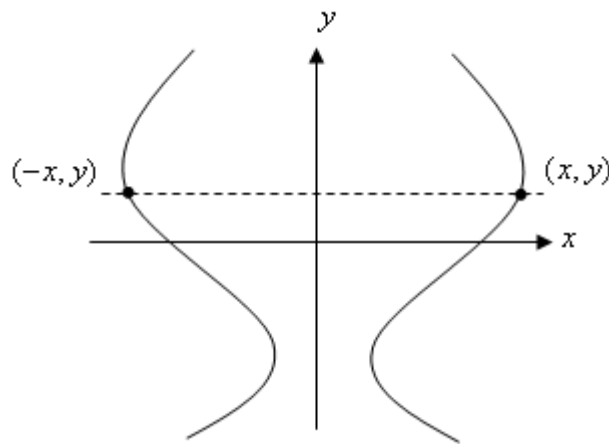


Figure: 16.14 - Exemple de symétrie par l'axe des ordonnées d'une fonction

P2. Le graphique d'une fonction est symétrique par rapport à l'axe des abscisses si le changement de y en $-y$ ne modifie pas la valeur de l'équation tel que:

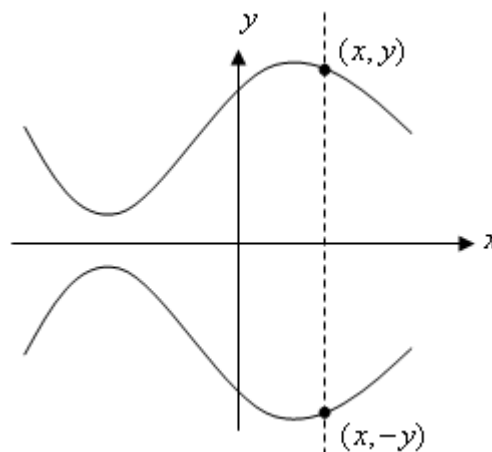


Figure: 16.15 - Exemple de symétrie par l'axe des abscisses d'une fonction

P3. Le graphique d'une fonction est symétrique par rapport à l'origine si le changement simultané de y en $-y$ et de x en $-x$ ne modifie pas la valeur de l'équation tel que:

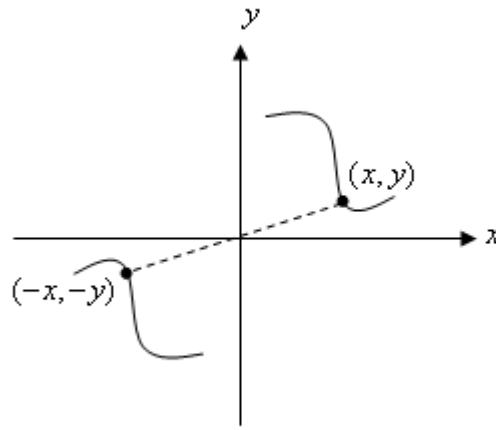


Figure: 16.16 - Exemple de symétrie par l'origine d'une fonction

P4. Soit une fonction $y = f(x)$, si nous ajoutons une constante $c^{\text{te}} > 0$ à cette fonction tel que nous écrivions:

$$y = f(x) + c^{\text{te}} \quad (16.15)$$

alors le graphique de f est déplacé (ou "translaté") verticalement vers le haut d'une distance c^{te} tel que présenté sur la figure suivante:

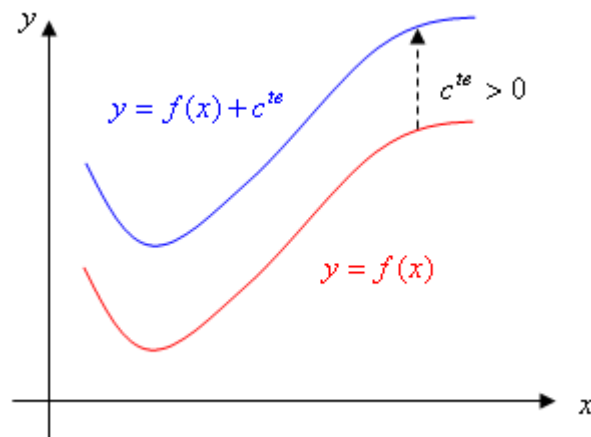


Figure: 16.17 - Exemple d'une translation verticale positive d'une fonction

Et inversement si $c^{\text{te}} > 0$ mais que:

$$y = f(x) - c^{\text{te}} \quad (16.16)$$

alors le graphique est bien évidemment translaté verticalement vers le bas:

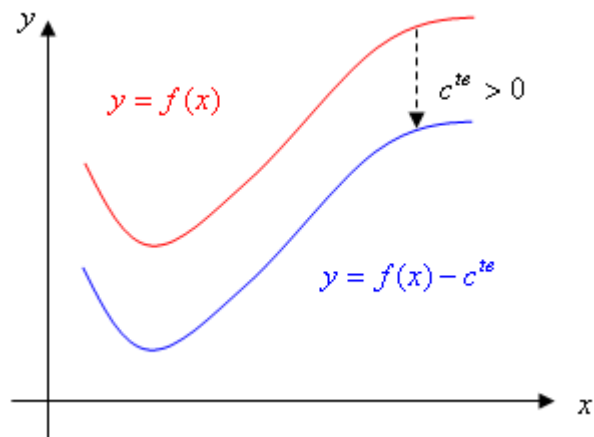


Figure: 16.18 - Exemple d'une translation verticale négative d'une fonction

Nous pouvons aussi envisager des translations horizontales de graphiques. Précisément, si $c^{te} > 0$, alors $y = f(x)$ est translaté horizontalement vers la droite si nous écrivons:

$$y = f(x - c^{te}) \quad (16.17)$$

ce qui graphiquement est représenté par:

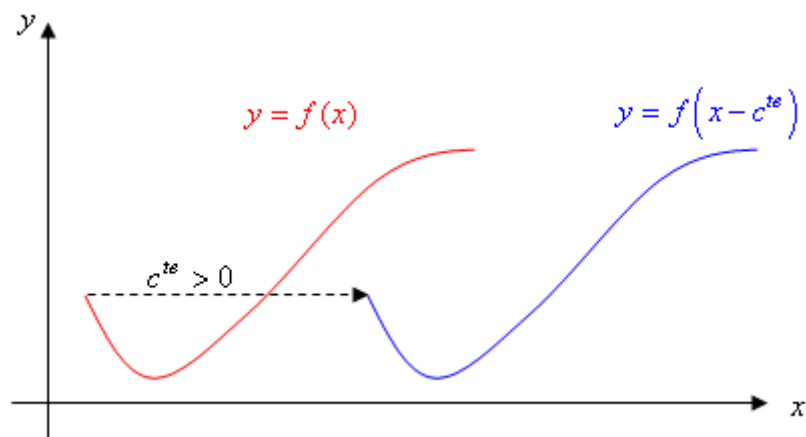


Figure: 16.19 - Exemple de translation horizontale négative d'une fonction

et inversement, translaté horizontalement vers la gauche, si nous écrivons:

$$y = f(x + c^{te}) \quad (16.18)$$

comme le montre le graphique ci-dessous:

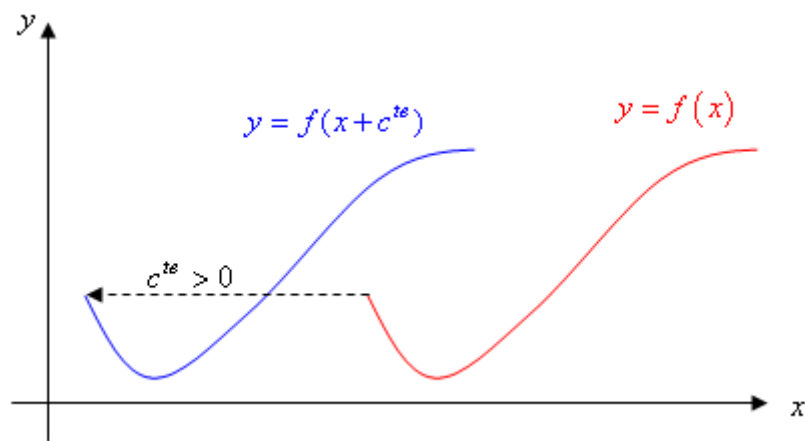


Figure: 16.20 - Exemple de translation horizontale positive d'une fonction

Pour étirer ou comprimer verticalement un graphique, il suffit de multiplier la fonction $y = f(x)$ par une constante $c^{te} > 1$ et respectivement $0 < c^{te} < 1$ tel que:

$$y = c^{te} f(x) \quad (16.19)$$

ce que nous pouvons représenter graphique par:

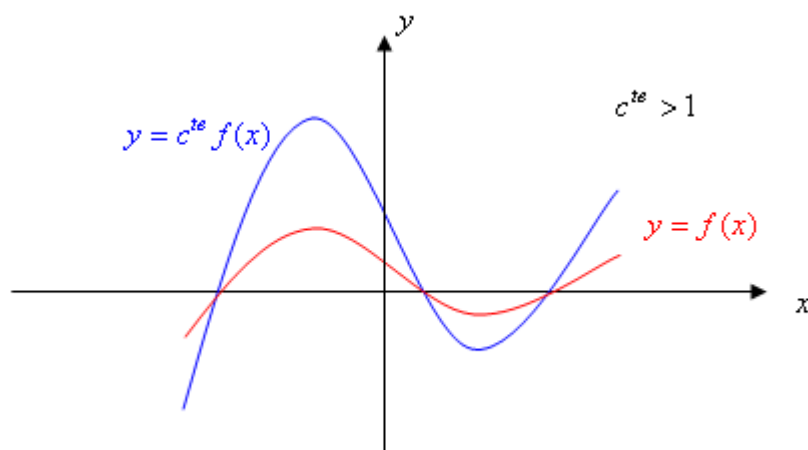


Figure: 16.21 - Exemple d'aplatissement vertical d'une fonction

et:

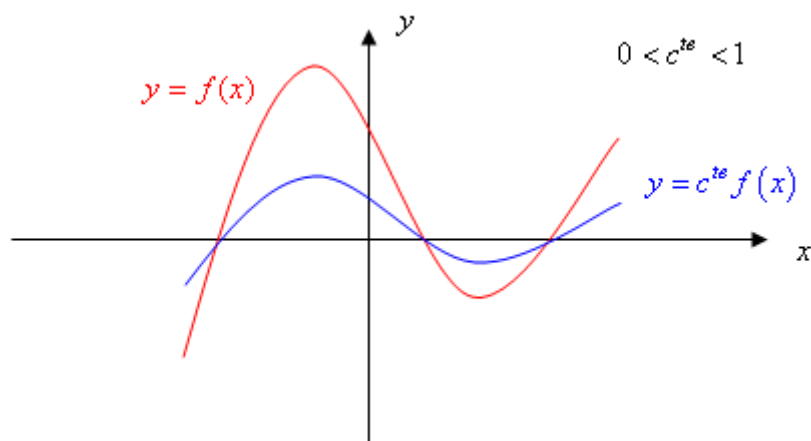


Figure: 16.22 - Exemple d'étirement vertical d'une fonction

Pour étirer ou comprimer horizontalement un graphique, il suffit de même, de multiplier la fonction $y = f(x)$ par une constante $c^{te} > 1$ et respectivement $0 < c^{te} < 1$ ou tel que:

$$y = f(c^{te} x) \quad (16.20)$$

ce que nous pouvons représenter sous forme graphique:

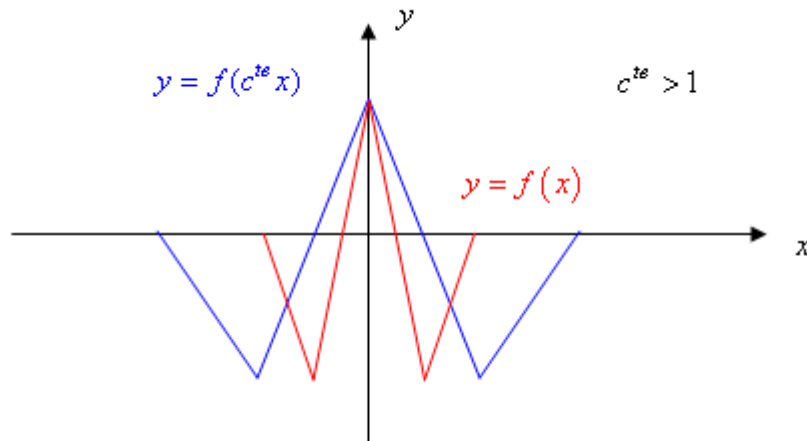


Figure: 16.23 - Exemple d'étirement horizontal d'une fonction

et:

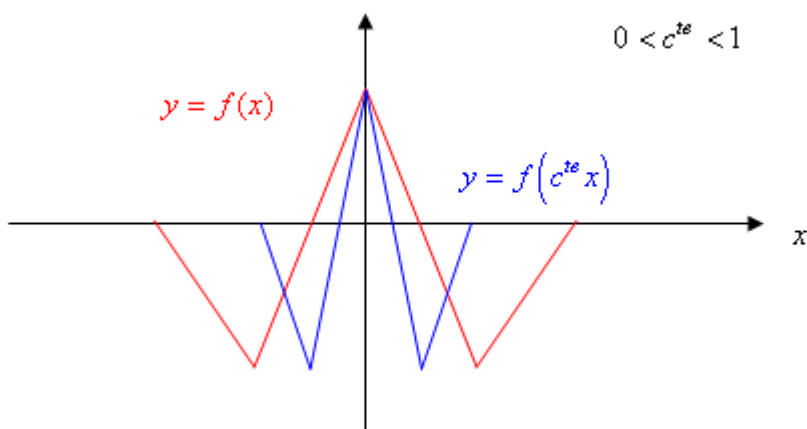


Figure: 16.24 - Exemple d'aplatissement horizontal d'une fonction

Remarque: Translater, étirer, comprimer un graphique ou lui faire subir une symétrie, c'est le transformer. Le graphique résultant de ces transformations est appelé le "**transformé**" du graphique de départ.

Définitions: Nous disons qu'une fonction f est:

- Une "**fonction croissante**" ou "**fonction croissante au sens large**" sur I si pour tout couple (x_1, x_2) , d'éléments de I tels que $x_1 \leq x_2$, nous avons $f(x_1) \leq f(x_2)$. Ce que nous notons de manière condensée:

$$\left\{ \forall (x, y) \in I^2 \wedge x \leq y \right\} \Rightarrow f(x) \leq f(y) \quad (16.21)$$

- Une "**fonction décroissante**" ou "**fonction décroissante au sens large**" sur I si pour tout couple (x_1, x_2) , d'éléments de I tels que $x_1 \leq x_2$, nous avons $f(x_1) \geq f(x_2)$. Ce que nous notons de manière condensée:

$$\left\{ \forall (x, y) \in I^2 \wedge x \leq y \right\} \Rightarrow f(x) \geq f(y) \quad (16.22)$$

Remarque: Une "**fonction est monotone**" ou "**fonction monotone au sens large**" sur I si elle est croissante ou décroissante.

-Une "**fonction strictement croissante**" sur I si pour tout couple (x_1, x_2) , d'éléments de I tels que $x_1 < x_2$, nous avons $f(x_1) < f(x_2)$. Ce que nous notons de manière condensée:

$$\left\{ \forall (x, y) \in I^2 \wedge x < y \right\} \Rightarrow f(x) < f(y) \quad (16.23)$$

- Une "**fonction strictement décroissante**" sur I si pour tout couple (x_1, x_2) , d'éléments de I tels que $x_1 < x_2$, nous avons $f(x_1) > f(x_2)$. Ce que nous notons de manière condensée:

$$\left\{ \forall (x, y) \in I^2 \wedge x < y \right\} \Rightarrow f(x) > f(y) \quad (16.24)$$

Remarque: Nous disons qu'une "**fonction est strictement monotone**" sur I si elle est strictement croissante sur I ou strictement décroissante sur I .

1.3. REPRÉSENTATIONS ANALYTIQUES

Le mode de représentation analytique est de loin le plus utilisé et consiste à représenter toute fonction en une "**expression analytique**" qui est la notation mathématique symbolique et abstraite de l'ensemble des opérations mathématiques connues que l'on doit appliquer dans un certain ordre à des nombres et des lettres exprimant des grandeurs constantes ou variables que nous cherchons à analyser.

Remarquons que par ensemble des opérations mathématiques connues, nous envisageons non seulement les opérations mathématiques vues dans la section arithmétique (addition, soustraction, extraction de la racine, etc.) mais également toutes les opérations qui seront définies au fur et à mesure dans le présent site internet.

Si la dépendance fonctionnelle $y = f(x)$ est telle que f est une expression analytique, nous disons alors que la "**fonction y de x** " est "**donnée analytiquement**". Voici quelques exemples d'expressions analytiques simples:

$$y = x^4 - 2, \quad y = \frac{x+1}{x-1}, \quad y = \sqrt{1-x^2} \quad (16.25)$$

Lorsque nous avons déterminé l'équation de la médiatrice, nous avons obtenu une expression

analytique de la droite visuelle qui la caractérise sous la forme d'une fonction du type:

$$f(x) = y = ax + b \quad (16.26)$$

qui rappelons-le, est donc l'expression analytique l'équation d'une droite, appelée également "**équation linéaire**" ou "**fonction affine**", sur un plan dont deux points $P_1(x_1, y_1), P_2(x_2, y_2)$ sont connus, la pente est donnée par le rapport de l'accroissement vertical sur l'accroissement horizontal tel que:

$$a = \frac{y_2 - y_1}{x_2 - x_1} \equiv \frac{\Delta y}{\Delta x} \quad (16.27)$$

Une application sympathique et triviale consiste à démontrer analytiquement que deux droites non verticales sont parallèles si et seulement si elles ont la même pente. Ainsi, soit deux droites données par les équations:

$$y_1 = a_1x + b_1 \quad y_2 = a_2x + b_2 \quad (16.28)$$

Les droites se coupent en un point (x, y) si et seulement si les valeurs de y sont égales pour un certain x , c'est-à-dire:

$$\begin{aligned} a_1x + b_1 &= a_2x + b_2 \\ \Rightarrow (a_1 - a_2)x &= b_2 - b_1 \end{aligned} \quad (16.29)$$

La dernière équation peut être résolue par rapport à x si et seulement si $a_1 - a_2 \neq 0$. Nous avons donc montré que les droites y_1, y_2 se coupent si et seulement si $a_1 \neq a_2$. Donc, elles ne se coupent pas (elles sont parallèles) si et seulement si $a_1 = a_2$.

De façon assez simple en appliquant le théorème de Pythagore, il n'est pas compliqué de déterminer que l'équation d'un cercle de centre $C(h, k)$ a pour équation (nous avons pour habitude en mathématiques de ne pas expliciter y pour l'équation du cercle ainsi, l'équation de ce dernier est visuellement beaucoup plus esthétique et parlante)

$$r = \sqrt{(x - h)^2 + (y - k)^2} \quad (16.30)$$

Dans ces exemples les fonctions sont exprimées analytiquement par une seule formule (égalité entre deux expressions analytiques) qui définit dans un même temps le "domaine naturel de définition" des fonctions.

Définition: Le "**domaine naturel de définition**" d'une fonction donnée par une expression analytique est l'ensemble des valeurs x pour lesquelles l'expression du membre de droite a une valeur bien déterminée.

Par exemple, la fonction:

$$y = \frac{x + 1}{x - 1} \quad (16.31)$$

est définie pour toutes les valeurs de x , excepté la valeur $x = 1$ où nous avons une singularité (division par zéro).

Remarque: Il existe une infinité de fonctions et nous ne pouvons toutes les exposer ici, cependant nous en rencontrerons plus d'un millier sur l'ensemble du site et cela devrait amplement suffire à se faire une idée de leur étude.

2. FONCTIONS

Définitions:

D1. Nous disons que y est une fonction de x et nous écrivons $y = f(x)$, $y = \varphi(x)$, etc., si à chaque valeur de la variable x appartenant à un certain domaine de définition (ensemble) D , correspond une valeur de la variable y dans un autre domaine de définition (ensemble) E . Ce que nous notons:

$$D \xrightarrow{f} E \quad f: D \rightarrow E \quad (16.32)$$

La variable x est appelée "**variable indépendante**" ou "**variable d'entrée**" et y "**variable dépendante**".

La dépendance entre les variables x et y s'appelle une "**dépendance fonctionnelle**". La lettre f , qui entre dans la notation symbolique de la dépendance fonctionnelle, indique qu'il faut appliquer certaines opérations à x pour obtenir la valeur correspondante y .

Nous écrivons parfois:

$$y = y(x) \quad (16.33)$$

au lieu de:

$$y = f(x) \quad (16.34)$$

Dans ce dernier cas la lettre y exprime en même temps la valeur de la fonction et le symbole des opérations appliquées à x .

Remarque: Comme nous l'avons vu lors de notre étude du chapitre de Théorie Des Ensembles, une application (ou fonction) peut-être injective, bijective ou surjective. Il convient donc que le lecteur pour qui ces notions ne sont pas connues aille en priorité lire ces définitions.

D2. L'ensemble des valeurs x pour lesquelles la valeur de la fonction y est donnée par la fonction $f(x)$ est appelé "**domaine d'existence**" de la fonction (ou domaine de définition de la fonction).

D3. La fonction $y = f(x)$ est dite "**fonction croissante**" si à une plus grande valeur de la variable indépendante correspond une plus grande valeur de la fonction (de l'image). Nous définissons de manière analogue mais inverse la "**fonction décroissante**".

D4. Une "**fonction constante**" est une fonction pour laquelle à toute valeur de la variable indépendante correspond toujours une même image constante.

D5. La fonction $y = f(x)$ est dite "**fonction périodique**" s'il existe un nombre constant c^* tel que la valeur de la fonction ne change pas quand nous ajoutons (ou que nous retranchons) le nombre c^* à la variable indépendante tel que:

$$y = f(x + c^{te}) = f(x) \quad (16.35)$$

Ce qui correspond à une translation selon x . La plus petite constante satisfaisant à cette condition est appelée "**période**" de la fonction. Elle est fréquemment notée T en physique.

D6. En calcul différentiel et intégral, l'expression:

$$\frac{f(x+h) - f(x)}{h} \quad (16.36)$$

avec $h \neq 0$ est d'un intérêt particulier. Nous l'appelons un "**quotient d'accroissement**" (nous reviendrons beaucoup plus en détail sur ce sujet lors de notre étude du calcul différentiel et intégral).

D7. Nous utilisons certaines propriétés des fonctions pour faciliter leur représentation graphique et leur analyse. En particulier, une fonction $f(x)$ est dite "**fonction paire**" si:

$$f(x) = f(-x) \quad (16.37)$$

pour tout x dans son domaine de définition. Une fonction est dite "**fonction impaire**" si:

$$f(-x) = -f(x) \quad (16.38)$$

pour tout x dans son domaine de définition.

Ainsi, pour résumer une fonction paire est une fonction qui ne dépend pas du signe de la variable et une fonction impaire change de signe quand nous changeons le signe de la variable (la spirale de Cornus dans le chapitre de Génie Civil est un bon exemple pratique de fonction impaire). Ce concept nous sera très utile pour simplifier certaines expressions très utiles en physique (comme les transformées de Fourier des fonctions paires ou impaires par exemple ou encore le calcul de certaines intégrales!).

Montrons maintenant que toute fonction $f(x)$ est la somme d'une fonction paire $g(x)$ et d'une fonction impaire $h(x)$.

Remarque: Ce type de théorème qui consiste à relier un concept général par un cas particulier et son opposé se retrouve souvent en mathématiques. Nous retrouverons de tels exemples en calcul tensoriel avec les tenseurs symétriques et antisymétriques (cf. chapitre de Calcul Tensoriel) ou encore en physique quantique avec les opérateurs hermitiques et antihermitiques (cf. chapitre de Physique Quantique Ondulatoire).

Démonstration:

Posons:

$$f(x) = g(x) + h(x) \quad (16.39)$$

alors:

$$f(-x) = g(x) - h(x) \quad (16.40)$$

Si nous sommes, nous avons dès lors:

$$g(x) = \frac{f(x) + f(-x)}{2} \quad (16.41)$$

et en soustrayant:

$$h(x) = \frac{f(x) - f(-x)}{2} \quad (16.42)$$

Il existe donc bien une décomposition paire et impaire de toute fonction.

□C.Q.F.D.

Enfin, il est important de remarquer que:

- Le produit de deux fonctions paires est une fonction paire
- Le produit deux fonctions impaires est une fonction paire
- Le produit d'une fonction paire et impaire est une fonction impaire

Voyons la brève démonstration pour la dernière propriété car nous en aurons besoin dans la section de Géométrie.

Soient g une fonction paire et h une fonction impaire telles que:

$$\begin{aligned} g(-x) &= g(x) \\ h(-x) &= -h(x) \end{aligned} \quad (16.43)$$

Dès lors:

$$q(-x) = g(-x)h(-x) = g(x)(-h(x)) = -g(x)h(x) = -q(x) \quad (16.44)$$

D8. De façon générale, si $f(x)$ et $g(x)$ sont des fonctions quelconques, nous utilisons la terminologie et les notations données dans le tableau suivant:

Terminologie	Valeur de la fonction
Somme $f + g$	$(f + g)(x) = f(x) + g(x)$
Différence $f - g$	$(f - g)(x) = f(x) - g(x)$
Produit $f \cdot g$	$(f \cdot g)(x) = f(x)g(x)$
Quotient $\frac{f}{g}$	$\left(\frac{f}{g}\right)(x) = \frac{f(x)}{g(x)}$

Tableau: 16.1 - Terminologie concernant les fonctions

Les domaines de définition de $f + g$, $f - g$, $f \cdot g$ sont l'intersection I des domaines de définition de $f(x)$ et de $g(x)$, c'est-à-dire les nombres qui sont communs aux deux domaines de définition. Le domaine

de définition de $f \cdot g^{-1}$ est quant à lui le sous-ensemble de I comprenant tous les x de I tels que $g(x) \neq 0$.

D9. Soit y une fonction f de u et u une fonction g de la variable x , alors y dépend de x et nous avons ce que nous appelons une "**fonction composée**" et que nous notons:

$$y = f(g(x)) \text{ ou } (f \circ g)(x) \quad (16.45)$$

Pour la dernière notation, il faut lire "**f rond g**" et ne pas confondre le "rond" avec la notation du produit scalaire que nous verrons lors de notre étude du calcul vectoriel (cf. chapitre de Calcul Vectoriel).

Le domaine de définition de la fonction composée est soit identique au domaine tout entier de définition de la fonction $u = g(x)$, soit à la partie de ce domaine dans laquelle les valeurs de u sont telles que les valeurs correspondantes $f(u)$ appartiennent au domaine de définition de cette fonction.

Le principe de fonction composée peut être appliqué non seulement une fois, mais un nombre arbitraire de fois.

Si x ne dépend pas d'une autre variable (ou qu'elle n'est pas elle-même une fonction composée), nous disons alors que $\varphi(x)$ est une "**fonction élémentaire**".

Les principales fonctions élémentaires sont des fonctions dont l'expression est l'une des suivantes:

1. La "**fonction puissance**":

$$f(x) = x^m \quad (16.46)$$

où m est un nombre positif différent de 1 (sinon il s'agit d'une fonction linéaire).

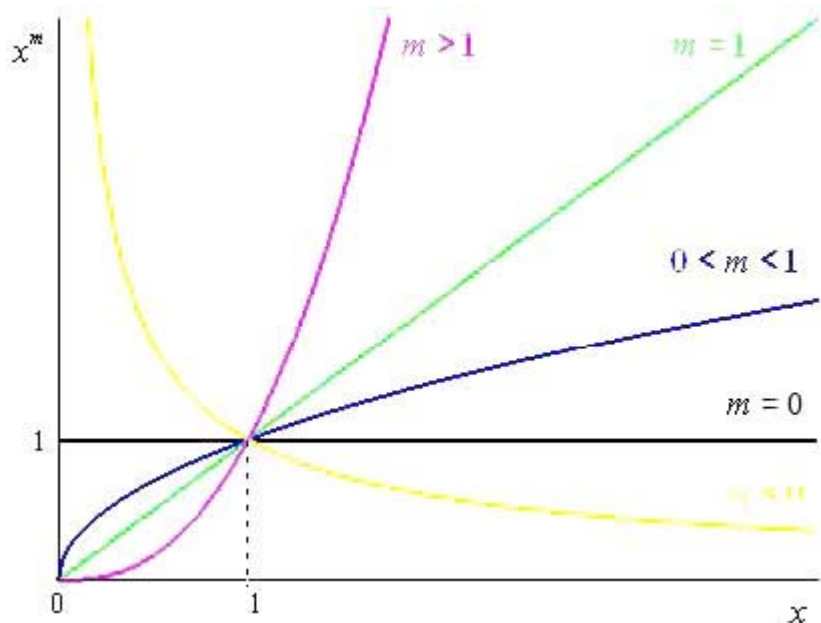


Figure: 16.25 - Différents tracés d'une fonction puissance simple

2. La "**fonction exponentielle**":

$$y = a^x \quad (16.47)$$

où a est un nombre positif différent de 1.

3. La "**fonction logarithmique**":

$$y = \log_a(x) \quad (16.48)$$

où la base du logarithme est un nombre positif a différent de l'unité (cette fonction sera définie rigoureusement un peu plus loin).

Remarque: Les fonctions exponentielles et logarithmiques sont appelées parfois des "**fonctions transcendantes**".

4. Les "**fonctions trigonométriques**" (cf. chapitre de Trigonométrie):

$$y = \sin(x), y = \cos(x), y = \tan(x), y = \cot(x), y = \sec(x), y = \operatorname{cosec}(x) \dots \quad (16.49)$$

5. Les "**fonctions polynomiales**":

$$y = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x^1 + a_0 \quad (16.50)$$

où $a_0, a_1, \dots, a_n$ sont des nombres constants appelés coefficients et n est un entier positif que nous appelons "**degré du polynôme**" (cf. chapitre de Calcul Algébrique). Il est évident que cette fonction est définie pour toutes les valeurs de x , c'est-à-dire qu'elle est définie dans un intervalle infini.

6. Les "**fractions rationnelles**" qui sont des divisions de polynômes (cf. chapitre de Calcul Algébrique):

$$y = \frac{a_0 x^n + a_1 x^{n-1} + \dots + a_n}{b_0 x^n + b_1 x^{n-1} + \dots + b_n} \quad (16.51)$$

Remarque: Deux fractions rationnelles sont égales, si l'une s'obtient de l'autre en multipliant le numérateur et le dénominateur par un même polynôme.

7. Les "**fonctions algébriques**" sont définies par le fait que la fonction $y = f(x)$ est le résultat d'opérations d'addition, de soustraction, de multiplication, de division, de variables élevées à une puissance rationnelle non entière.

Remarque: Il existe cependant un très grand nombre de fonctions que nous rencontrerons dans les différents chapitres du site. Citons par exemple les "fonctions de Bessel" (cf. chapitre des Suites Et Séries), les "fonctions lipschitziennes" (cf. chapitre de Topologie), les "fonctions de Dirac" (cf. chapitre de Calcul Différentiel Et Intégral), les "fonctions de répartition et de distribution" (cf. chapitre de Statistiques), la "fonction gamma d'Euler" (cf. chapitre de Calcul Différentiel Et Intégral), etc.

8. Une application $f: [a, b] \rightarrow \mathbb{R}$ est dite "**fonction en escalier**" si et seulement si, il existe une

subdivision $(a_i)_{0 \leq i \leq n}$ de $[a, b]$ tel que $a_0 = a$ et $a_n = b$ et $(\lambda_0, \dots, \lambda_n) \in \mathbb{R}^n$ tels que:

$$\forall i \in \{0, \dots, n-1\}, \forall x \in]a_i, a_{i+1}[\Leftrightarrow f(x) = \lambda_i$$

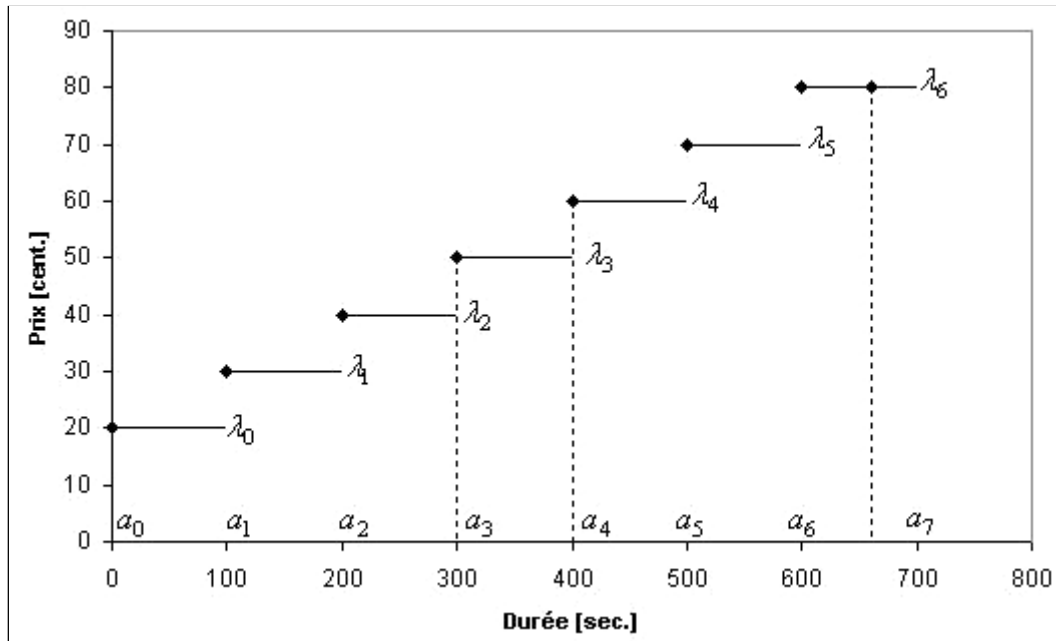


Figure: 16.26 - Exemple d'une fonction en escalier courante

2.1. LIMITE ET CONTINUITÉ DES FONCTIONS

Nous allons considérer maintenant des variables ordonnées d'un type spécial, que nous définissons par la relation "la variable tend vers une limite". Dans la suite de ce cours, la notion de limite d'une variable va jouer un rôle fondamental, étant intimement liée aux notions de base de l'analyse mathématique, la dérivée, l'intégrale, etc.

Définition: Le nombre a est appelé la "**limite**" de la grandeur variable x , si, pour tout nombre arbitrairement petit $\varepsilon > 0$ avons:

$$|x - a| < \varepsilon \quad (16.52)$$

Si le nombre a est la limite de la variable x , nous disons que "**x tend vers la limite a**".

Nous pouvons définir également la notion de limite en partant de considérations géométriques (cela peut aider à mieux comprendre... quoique pas toujours...):

Le nombre constant a est la limite de la variable x , si pour tout voisinage donné, aussi petit qu'il soit, de centre a et de rayon ε , nous pouvons trouver une valeur x telle que tous les points correspondant aux valeurs suivantes de la variable appartiennent à ce voisinage (notions que nous avons défini précédemment). Géométriquement nous représentons cela ainsi:

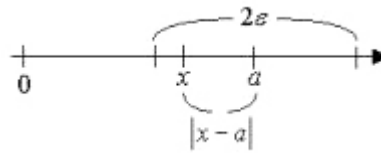


Figure: 16.27 - Notion géométrique de limite

Remarque: Il devrait être trivial que la limite d'une grandeur constante est égale à cette constante, puisque l'inégalité $|x - c| = |c - c| = 0 < \varepsilon$ est toujours satisfaite pour ε arbitraire.

Il découle également de la définition de la limite qu'une grandeur variable ne peut pas avoir n'importe comment deux limites. En effet, si:

$$\lim_{x \rightarrow a} x = a \text{ et } \lim_{x \rightarrow b} x = b \quad (16.53)$$

avec $a < b$, x doit satisfaire simultanément aux deux inégalités suivantes:

$$|x - a| < \varepsilon \text{ et } |x - b| < \varepsilon \quad (16.54)$$

pour ε arbitrairement choisi. Mais si nous faisons une représentation géométrique identique à la précédente, nous voyons assez aisément que cela est impossible si:

$$\varepsilon < \frac{b - a}{2} \quad (16.55)$$

Il ne faut également pas s'imaginer que chaque variable doit nécessairement avoir une limite.

Définition: La variable x tend vers l'infini, si pour chaque nombre positif donné M , nous indiquons une valeur de x à partir de laquelle toutes les valeurs conséquentes de la variable (valeurs de la variable appartenant dans le voisinage défini à partir de valeur indiquée précédemment) x vérifient l'inégalité $|x| > M$.

Nous pouvons vérifier ce genre de cas dans les suites arithmétiques, géométriques, ou harmoniques où chaque terme de la progression est une valeur que prend la variable x .

La variable x "**tend vers plus l'infini**", ou $x \rightarrow +\infty$ si pour $M > 0$ arbitraire, à partir d'une certaine valeur, toutes les valeurs conséquentes de la variable vérifient l'inégalité $M < x$. C'est typiquement le genre de considération que nous avons pour des progressions divergentes vers l'infini où à partir d'un certain terme de valeur égale à M tous les termes suivants sont supérieurs à M .

La variable x "**tend vers moins l'infini**" ou $x \rightarrow -\infty$ si pour $M > 0$ arbitraire, à partir d'une certaine valeur, toutes les valeurs suivantes de la variable vérifient l'inégalité $x < -M$.

Définition: Soit $y = f(x)$ une fonction définie dans un voisinage du point a ou en certains points de ce voisinage. La fonction $y = f(x)$ tend vers la limite b ($y \rightarrow b$) lorsque x tendant vers a ($x \rightarrow a$), si pour chaque nombre positif ε , aussi petit qu'il soit, nous pouvons indiquer un nombre positif δ tel que tous les x différents de a et vérifiant l'inégalité $|x - a| < \delta$ satisfont également:

$$|f(x) - b| < \varepsilon \quad (16.56)$$

L'inégalité $|x - a| < \delta$ permet d'exprimer le côté (ou le sens) depuis lequel nous venons avec notre x . Car sur le système d'axe représentant des valeurs ordonnées, nous pouvons, pour une valeur donnée, venir de sa gauche ou de sa droite pour se rapprocher d'elle (imaginez-vous au besoin, un bus qui peut venir depuis un côté ou un autre de la route tant que la distance qui le sépare à l'arrêt qui nous intéresse est inférieure à δ).

Si b est la limite de la fonction $f(x)$ quand $x \rightarrow a$, nous écrivons alors sur ce site en tous les cas:

$$\lim_{x \rightarrow a} f(x) = b \quad (16.57)$$

Pour définir le côté depuis lequel nous venons en appliquant la limite, nous utilisons une notation particulière (rappelons que cela permet de connaître de quel côté de la route vient notre bus).

Ainsi, si $f(x)$ tend vers la limite b_1 quand x tend vers un nombre a en ne prenant que des valeurs plus petites que a , nous écrirons alors:

$$\lim_{x \rightarrow a_-} f(x) = b_1 \quad (16.58)$$

(remarquez le petit $-$ en indice) et nous appellerons b_1 la "**limite à gauche**" de la fonction $f(x)$ au point a (car rappelez-vous que l'axe des ordonnées va de $-\infty$ à $+\infty$, donc les petites valeurs par rapport à une valeur donnée, se trouvent à gauche). Si x prend des valeurs plus grandes que a , nous écrirons alors:

$$\lim_{x \rightarrow a_+} f(x) = b_2 \quad (16.59)$$

(remarquez le petit $+$ en indice) et nous appellerons b_2 la "**limite à droite**" de la fonction au point a .

Définition: La fonction $f(x)$ tend vers la limite b quand $x \rightarrow \infty$ si pour chaque nombre positif ε aussi petit qu'il soit nous pouvons indiquer un nombre positif N tel que pour toutes les valeurs de x vérifiant l'inégalité $|x| > N$, l'inégalité $|f(x) - b| < \varepsilon$ est satisfaite.

Exemple:

Montrons que (nous supposons le résultat connu pour l'instant):

$$\lim_{x \rightarrow \infty} \left(\frac{x+1}{x} \right) = 1 \quad (16.60)$$

Il faut démontrer que, quel que soit ε , l'inégalité $\left| 1 + \frac{1}{x} - 1 \right| < \varepsilon$ sera satisfaite dès que $|x| > N$, où N est défini par le choix de ε . L'inégalité précédente est évidemment équivalente à $\left| \frac{1}{x} \right| < \varepsilon$, qui est satisfaite si nous avons x :

$$|x| > \frac{1}{\varepsilon} = N \quad (16.61)$$

Nous admettons que l'exemple et la méthode sont discutables, mais nous verrons plus tard les outils mathématiques adéquats pour arriver rigoureusement, sans magouilles et hypothèses de départ, au résultat obtenu précédemment.

La signification des symboles $x \rightarrow +\infty$ et $x \rightarrow -\infty$, rend évidente celle des expressions:

$$f(x) \text{ tend vers } b \text{ quand } x \rightarrow +\infty$$

et:

$$f(x) \text{ tend vers } b \text{ quand } x \rightarrow -\infty$$

que nous notons symboliquement par:

$$\lim_{x \rightarrow +\infty} f(x) = b \text{ et } \lim_{x \rightarrow -\infty} f(x) = b \quad (16.62)$$

Nous avons étudié les cas où la fonction $f(x)$ tend vers une certaine limite b quand $x \rightarrow a$ ou $x \rightarrow \infty$. Considérons maintenant le cas où la fonction $y = f(x)$ tend vers l'infini quand la variable x varie d'une certaine manière.

Définition: La fonction $f(x)$ tend vers l'infini quand $x \rightarrow a$, autrement dit $f(x)$ est infiniment grande quand $x \rightarrow a$, si pour chaque nombre positif M , aussi grand qu'il soit, nous pouvons trouver un nombre $\delta > 0$ tel que pour toutes les valeurs de x différentes de a et vérifiant la condition $|x - a| < \delta$, l'inégalité $|f(x)| > M$ est satisfaite.

Si $f(x)$ tend vers l'infini quand $x \rightarrow a$, nous écrivons:

$$\lim_{x \rightarrow a} f(x) = \infty \quad (16.63)$$

Si $f(x)$ tend vers l'infini quand $x \rightarrow a$, en ne prenant que des valeurs positives ou que des valeurs négatives, nous écrivons respectivement:

$$\lim_{x \rightarrow a} f(x) = +\infty \text{ et } \lim_{x \rightarrow a} f(x) = -\infty \quad (16.64)$$

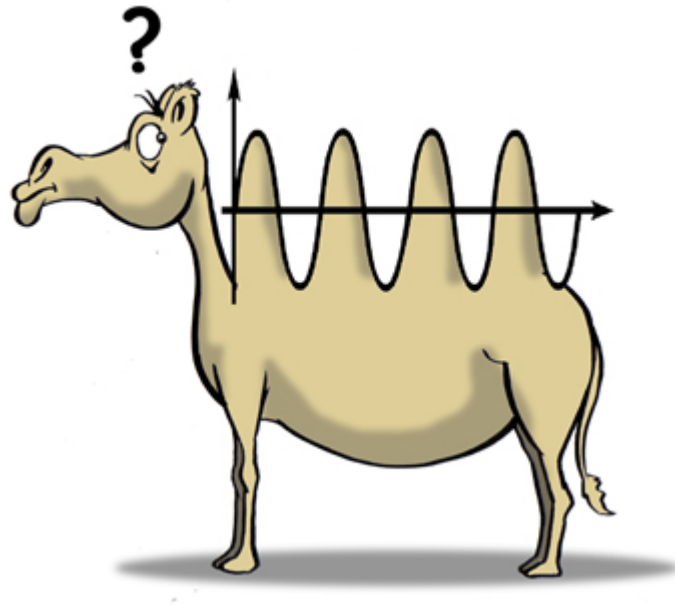
Si la fonction $f(x)$ tend vers l'infini quand $x \rightarrow \infty$ on écrit:

$$\lim_{x \rightarrow \infty} f(x) = +\infty \quad (16.65)$$

et en particulier, nous pouvons avoir:

$$\lim_{x \rightarrow +\infty} f(x) = +\infty, \lim_{x \rightarrow -\infty} f(x) = +\infty, \lim_{x \rightarrow +\infty} f(x) = -\infty, \lim_{x \rightarrow -\infty} f(x) = -\infty \quad (16.66)$$

Il peut arriver que la fonction $f(x)$ ne tende ni vers une limite finie, ni vers l'infini quand $x \rightarrow a$ (par exemple $f(x) = \sin(x)$), la fonction est alors bornée (cf. chapitre de Théorie des Ensembles).



Maintenant que nous avons grosso modo eu un aperçu du concept de limite, nous allons donner une définition extrêmement importante qui est un des piliers de beaucoup de domaines de la mathématique et de la physique.

Définition: Soit f une fonction définie sur $I \subseteq \mathbb{R}$. Soit $x_0 \in I$, nous disons que nous avons une "fonction continue" en x_0 si et seulement si:

$$\lim_{x \rightarrow x_0} f(x) = f(x_0) \quad (16.67)$$

c'est-à-dire si (il faut pouvoir arriver à y lire le fait qu'on s'approche de manière infiniment petite d'une limite ce qui permet d'assurer la continuité) que $\forall \varepsilon > 0, \exists \alpha > 0$ tel que $x \in I$ alors:

$$|x - x_0| < \alpha \Rightarrow |f(x) - f(x_0)| \leq \varepsilon \quad (16.68)$$

Remarque: f est "continue à droite" (resp. à gauche) si nous rajoutons la condition $x \geq x_0$ (resp. $x \leq x_0$).

Nous avons les corollaires triviaux suivants:

C1. f est continue en x_0 si et seulement si f est continue à droite et à gauche en x_0

C2. f est continue sur I si et seulement si f est continue en tout point de I .

2.2. ASYMPTOTES

Le terme d'asymptote est utilisé en mathématiques pour préciser des propriétés éventuelles d'une branche infinie de courbe à accroissement tendant vers l'infinésimal.

L'étude du comportement asymptotique est particulièrement développée dans les études de fonctions. Dans le domaine scientifique, il arrive fréquemment d'étudier des fonctions dépendant du temps

(évolution de populations, réaction chimique ou nucléaire, graphique de température, oscillation d'un amortisseur). Un des objectifs du chercheur est alors de connaître l'état à la fin de l'expérience, c'est-à-dire lorsqu'un grand intervalle de temps s'est écoulé. L'objectif n'est alors pas de connaître les variations intermédiaires, mais de déterminer le comportement stable, à l'infini du phénomène mesuré. Le chercheur étudie donc le comportement asymptotique de sa fonction avec les outils que les mathématiques lui offrent.

Définitions:

D1. Lorsque la limite d'une fonction $f(x)$ tend vers une constante $c^{\pm}$ quand $x \rightarrow \pm\infty$, alors la représentation graphique de cette fonction nous amène à dessiner une droite horizontale que nous appelons "**asymptote horizontale**" et dont l'équation est:

$$A(x) = c^{\pm} \quad (16.69)$$

D2. Lorsque la limite d'une fonction $f(x)$ tend vers $\pm\infty$ quand $x \rightarrow \alpha_{\pm}$, alors la représentation graphique de cette fonction nous amène à dessiner une droite verticale que nous appelons "**asymptote verticale**" et dont l'équation est:

$$A(y) = \alpha_{\pm} \quad (16.70)$$

Exemple:

La courbe représentative de la fonction $f(x)=1/(x-1)$ admet la droite d'équation $x = 1$ comme asymptote verticale et $y = 0$ comme asymptote horizontale:

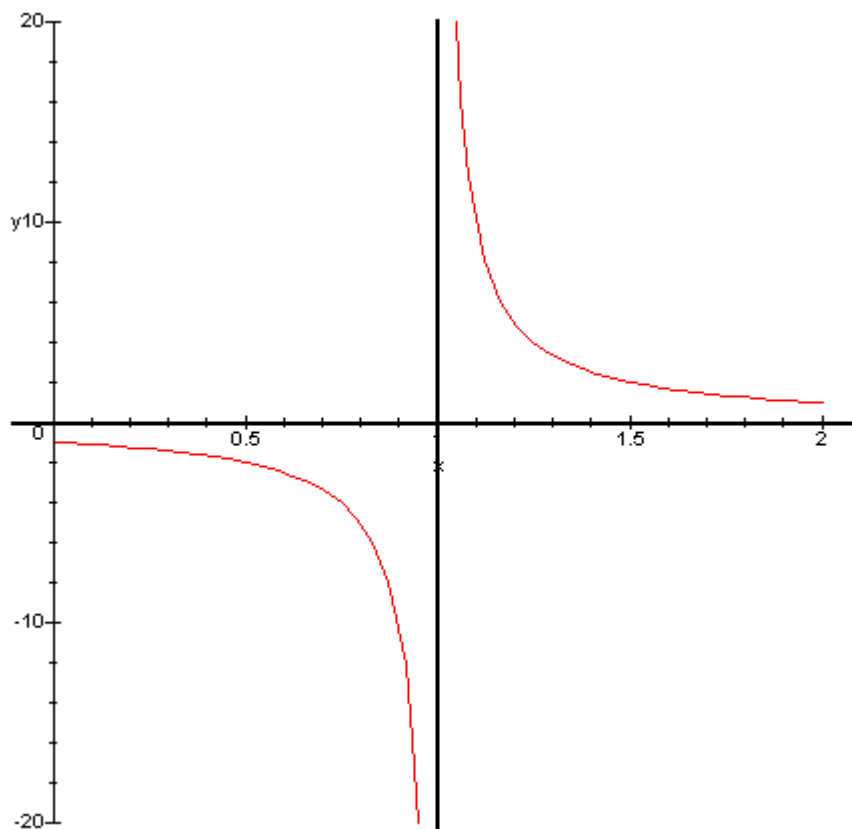


Figure: 16.28 - Représentation graphique plane d'une asymptote verticale et horizontale

D3. La droite d'équation $y = ax + b$ est une "asymptote oblique" à la courbe de la fonction $f(x)$ si:

$$\lim_{x \rightarrow \pm\infty} f(x) - (ax + b) = 0 \quad (16.71)$$

les valeurs de a et de b peuvent se retrouver facilement à l'aide des relations suivantes :

$$\begin{aligned} a &= \lim_{x \rightarrow \pm\infty} \frac{f(x)}{x} \\ b &= \lim_{x \rightarrow \pm\infty} f(x) - ax \end{aligned} \quad (16.72)$$

Remarque: Attention une courbe peut admettre deux asymptotes obliques distinctes en $+\infty$ et en $-\infty$

Pour rechercher une asymptote oblique éventuelle, il faut déjà être sûr que la fonction f admet une limite infinie en $+\infty$ ou en $-\infty$ ensuite nous cherchons la limite en $+\infty$ ou en $-\infty$ de $f(x)/x$.

Trois cas sont à considérer :

C1. La courbe représentative de f a pour direction asymptotique la droite d'équation $y = ax$:

$$\lim_{x \rightarrow \pm\infty} \frac{f(x)}{x} = a \quad (16.73)$$

 **Exemple:**

La fonction $f(x) = x + 1/x$ possède entre autres une asymptote d'équation $y = x$:

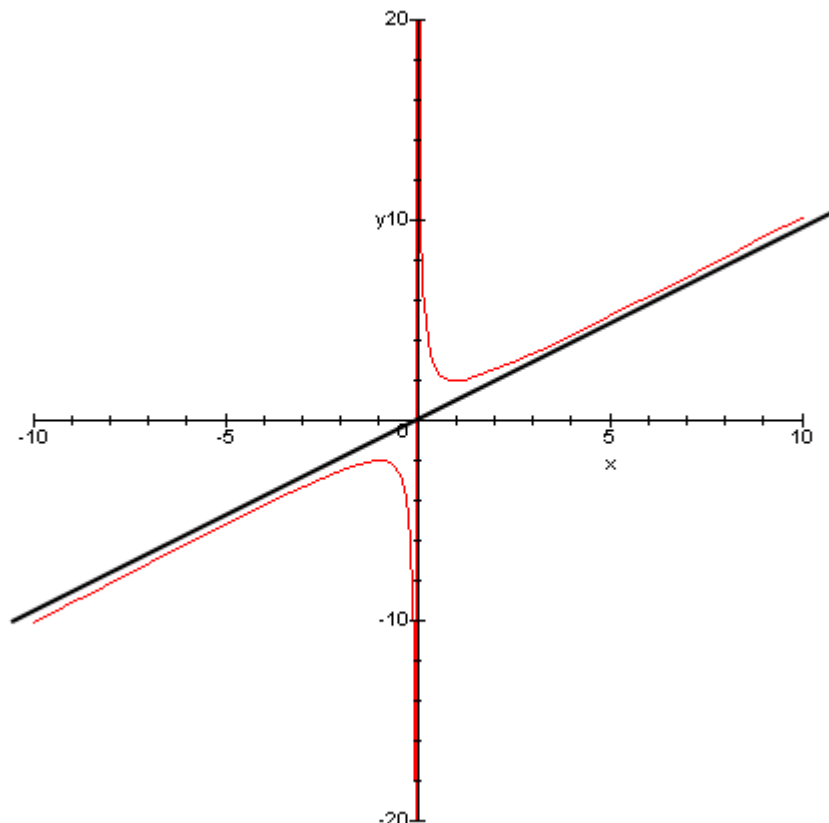


Figure: 16.29 - Représentation graphique plane d'une asymptote oblique

C2. La courbe représentative de f admet une branche infinie (cette branche infinie n'admet pas d'asymptote) et l'axe des abscisses en est la direction asymptotique (fonction racine carrée par exemple)

$$\lim_{x \rightarrow \pm\infty} \frac{f(x)}{x} = 0 \quad (16.74)$$

Exemple:

Les fonctions $f(x) = \sqrt{x}$ (en rouge) ou $\ln(x)$ (en vert) ont une limite $f(x)/x$ nulle et possèdent donc toutes deux une "branche parabolique" de direction Ox .

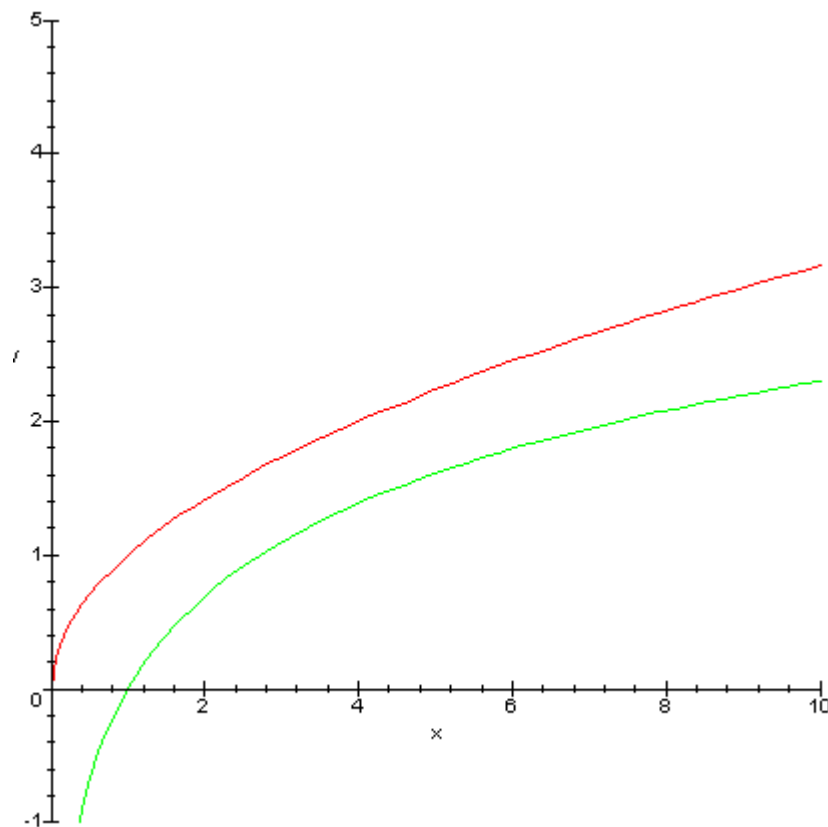


Figure: 16.30 - Exemples de branches paraboliques pour deux fonctions

C3. La courbe représentative de f admet une branche infinie (cette branche infinie n'admet pas d'asymptote) et l'axe des ordonnées en est la direction asymptotique. (nous parlons aussi de "branche parabolique" voire de "fonction carrée")

$$\lim_{x \rightarrow \pm\infty} \frac{f(x)}{x} = \pm\infty \quad (16.75)$$

Exemple:

La fonction $f(x) = x^2$ a une limite $f(x)/x$ infinie et possède donc une "branche parabolique" de direction Oy .

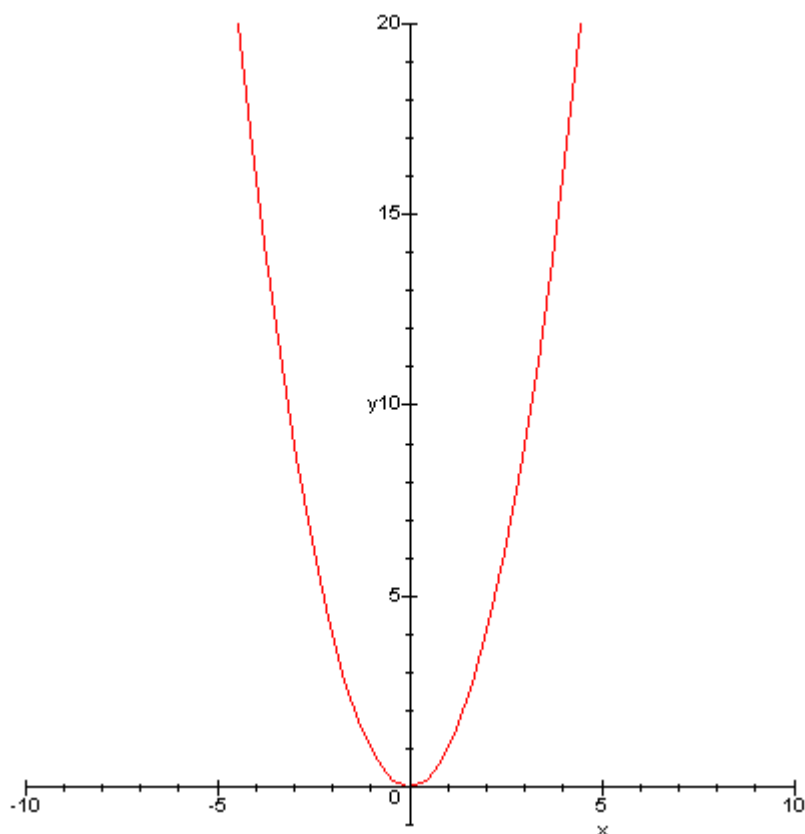


Figure: 16.31 - Exemple de parabole

3. LOGARITHMES

Nous avons longuement hésité à mettre la définition des logarithmes dans le chapitre traitant du calcul algébrique. Après un moment de réflexion, nous avons décidé qu'il valait mieux la mettre ici car pour bien la comprendre, il faut avoir connaissance des concepts de limite, domaine de définition et fonction exponentielle. Nous espérons que notre choix vous conviendra au mieux.

Soit la fonction exponentielle (bijective) de base quelconque a , où $a \in \mathbb{R}_+^* / 1$ notée:

$$f(x) = a^x \quad (16.76)$$

pour laquelle il correspond à chaque nombre réel x , exactement un nombre positif a^x (l'ensemble image de la fonction est dans $\mathbb{R}$) tel que les règles de calcul des puissances soient applicables (cf. chapitre de Calcul Algébrique).

Nous savons que pour une telle fonction, que si $a > 1$, alors $f(x)$ est croissante et positive dans $\mathbb{R}$, et si $0 < a < 1$, alors $f(x)$ est décroissante et positive dans $\mathbb{R}$.

Remarques:

R1. Si $a > 1$, lorsque x décroît vers des valeurs négatives, le graphique de $f(x)$ tend vers l'axe des x . Ainsi, l'axe des x est une asymptote horizontale. Lorsque x croît par valeurs positives, le graphique monte rapidement. Ce type de variation est caractéristique de la "**loi de croissance exponentielle**" et $f(x)$ est quelques fois appelée "**fonction de croissance**". Si $0 < a < 1$, lorsque x croît, le graphique tend asymptotiquement vers l'axe des x . Ce type de variation est connu sous le nom de "**décroissance exponentielle**".

R2. En étudiant a^x , nous excluons le cas $a \leq 0$ et $a = 1$. Notons que si $a < 0$, alors a^x n'est pas un nombre réel pour de nombreuses valeurs de x (nous rappelons que l'ensemble image est contraint à $\mathbb{R}$). Si $a = 0$, $a^0 = 0^0$ n'est pas défini. Enfin, si $a = 1$, alors $a^x = 1$ pour tout x et le graphique de $f(x)$ est une droite horizontale.

Puisque la fonction exponentielle $f(x)$ est bijective alors il existe une fonction réciproque $f^{-1}(x)$ et appelée "**fonction logarithme**" de base a notée:

$$f^{-1}(y) = \log_a(y) \quad (16.77)$$

Et donc:

$$x = \log_a(y) \quad (16.78)$$

si et seulement si $y = a^x$.

En considérant $\log_a(x)$ comme un exposant, nous avons les propriétés suivantes:

Propriétés	Justification
$\log_a 1 = 0$	$a^0 = 1$
$\log_a a = 1$	$a^1 = a$
$\log_a a^x = x$	$a^x = a^x$
$a^{\log_a(x)} = x$	$a^{\log_a(x)} = a^y = x$

Tableau: 16.2 - Propriétés du logarithme en base a

Remarques:

R1. Le mot "logarithme" signifie "nombre du logos", "logos" signifiant "raison" ou "rapport".

R2. Les fonctions logarithme et exponentielle sont définies par leur base (le nombre a). Lorsqu'on utilise une base de puissance de 10 (10, 100, 1000,...) nous parlons alors de "**système vulgaire**" car ils ont pour logarithme des nombres entiers successifs.

R3. La partie entière du logarithme s'appelle la "**caractéristique**".

Il existe deux types de logarithmes que nous retrouvons presque exclusivement en mathématiques et en physique : le logarithme en base dix et le logarithme en base e (ce dernier étant fréquemment appelé "**logarithme naturel**" ou plus exactement pour des raisons historiques justifiées "**logarithme népérien**").

D'abord celui en base 10:

$$f(x) = \log_{10}(x) \quad (16.79)$$

abusivement noté:

$$f(x) = \log(x) \quad (16.80)$$

et celui en base (eulérienne) e :

$$f(x) = \log_e(x) \quad (16.81)$$

historiquement noté:

$$f(x) = \ln(x) \quad (16.82)$$

le "n" signifiant "népérien".

Remarque: Historiquement, c'est à John Napier (1550-1617) dont le nom latinisé est "Neper" que l'on doit l'étude des logarithmes et le nom aux "logarithmes népériens".

En français pour la fonction logarithmique en base 10 il faut pour calculer:

$$f(x) = \log(x) \quad (16.83)$$

se poser la question suivante: à quelle puissance $n \in \mathbb{R}$ devons-nous élever 10 pour obtenir x ?

Formellement, cela consiste à résoudre l'équation:

$$10^n = x \quad (16.84)$$

ou autrement écrit:

$$10 = \sqrt[n]{x} \quad (16.85)$$

avec x étant connu et donc en base 10:

$$\log(x) = n \quad (16.86)$$

Pour la fonction logarithmique en base eulérienne e (ou dite "**base népérienne**") il faut pour calculer:

$$f(x) = \ln(x) \quad (16.87)$$

se poser aussi la question suivante: à quelle puissance $n \in \mathbb{R}$ devons nous élever le nombre e pour obtenir x ?

Formellement, cela consiste à résoudre l'équation:

$$e^n = x \quad (16.88)$$

ou autrement écrit:

$$e = \sqrt[n]{x} \quad (16.89)$$

avec x étant connu et donc:

$$\ln(x) = n \quad (16.90)$$

Techniquement, nous disons alors que la fonction exponentielle (voir plus bas les détails):

$$f(x) = e^x \quad (16.91)$$

est la bijection réciproque de la fonction $\ln(x)$.

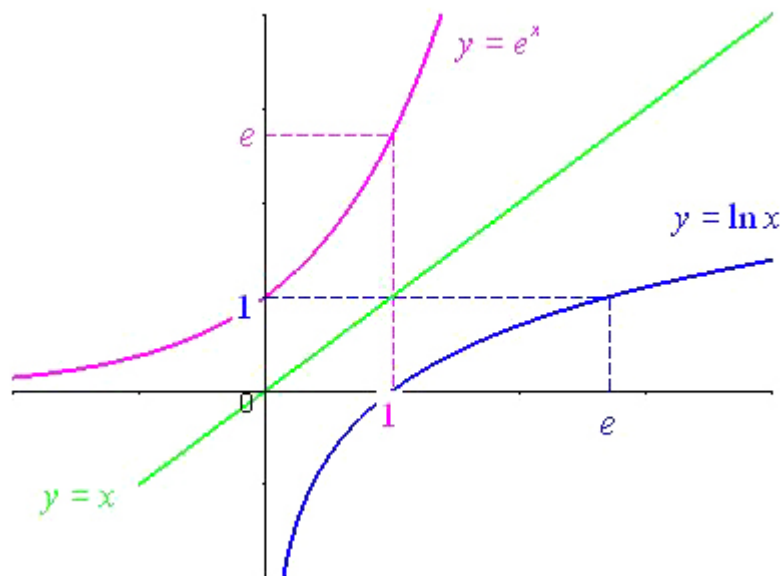


Figure: 16.32 - Représentation graphique plane de la bijection entre le logarithme népérien et l'exponentielle

Mais quel est donc ce nombre "eulérien" appelé également "**nombre d'Euler**" ? Pourquoi le retrouve-t-on si souvent en physique et en mathématiques? D'abord déterminons l'origine de sa valeur:

Pour cela, il nous faut déterminer la limite (dont l'origine historique semblerait être l'étude de problèmes financiers par Euler) de :

$$\sqrt[n]{1+n} = \left(1+n\right)^{\frac{1}{n}} = \left(1+\frac{1}{\alpha}\right)^{\alpha} \quad (16.92)$$

avec $\alpha \in \mathbb{N}$ et quand $\alpha \rightarrow \infty$.

Remarque: Le deuxième terme de l'égalité est donc typiquement le type d'expression que nous retrouvons dans les intérêts composés en finance (cf. chapitre d'Économie) ou dans tout autre type d'accroissement à facteur égal. Et ce qui nous intéresse dans le cas présent c'est quand ce type d'accroissement tend vers l'infini.

L'intérêt que nous avons à poser le problème ainsi c'est que si nous faisons tendre $\alpha \rightarrow \infty$ la fonction écrite précédemment tend vers e et cette fonction a pour propriété particulière de pouvoir se calculer plus ou moins facilement pour des raisons historiques à l'aide du binôme de Newton.

Donc d'après le développement du binôme de Newton (cf. chapitre de Calcul Algébrique) nous pouvons écrire :

$$\left(1+\frac{1}{\alpha}\right)^{\alpha} = 1 + \frac{\alpha}{1!} \frac{1}{\alpha} + \frac{\alpha(\alpha-1)}{2!} \frac{1}{\alpha^2} + \frac{\alpha(\alpha-1)(\alpha-2)}{3!} \frac{1}{\alpha^3} + \dots + \frac{\alpha(\alpha-1)\dots(\alpha-(\alpha-1))}{\alpha!} \frac{1}{\alpha^{\alpha}} \quad (16.93)$$

Ce développement, est similaire au développement de Taylor de certaines fonctions pour des cas particuliers de valeurs de développement (d'où la raison pour laquelle nous retrouvons ce nombre eulérien dans beaucoup d'endroits que nous découvrirons au fur et à mesure).

En effectuant certaines transformations algébriques évidentes, nous trouvons :

$$\left(1+\frac{1}{\alpha}\right)^{\alpha} = 1 + 1 + \frac{1}{2!} \left(1 - \frac{1}{\alpha}\right) + \frac{1}{3!} \cdot \left(1 - \frac{1}{\alpha}\right) \left(1 - \frac{2}{\alpha}\right) + \dots + \frac{1}{\alpha!} \left(1 - \frac{1}{\alpha}\right) \left(1 - \frac{2}{\alpha}\right) \dots \left(1 - \frac{\alpha-1}{\alpha}\right) \quad (16.94)$$

Nous voyons de cette dernière égalité que la fonction $\left(1+\frac{1}{\alpha}\right)^{\alpha}$ est croissante quand α croît. En effet, quand nous passons de la valeur α à la valeur $\alpha+1$ chaque terme de cette somme augmente :

$$\frac{1}{2!} \left(1 - \frac{1}{\alpha}\right) < \frac{1}{2!} \left(1 - \frac{1}{\alpha+1}\right), \text{ etc.} \quad (16.95)$$

Montrons que la grandeur variable $\left(1+\frac{1}{\alpha}\right)^{\alpha}$ est bornée. En remarquant que :

$$\left(1 - \frac{1}{\alpha}\right) < 1, \left(1 - \frac{1}{\alpha}\right) \left(1 - \frac{2}{\alpha}\right) < 1, \text{ etc.} \quad (16.96)$$

Nous obtenons donc par analogie avec l'expression étendue en binôme de Newton déterminée plus haut, la relation d'ordre suivante:

$$\left(1 + \frac{1}{\alpha}\right)^\alpha < 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \dots + \frac{1}{\alpha!} \quad (16.97)$$

D'autre part:

$$\frac{1}{3!} < \frac{1}{2^2}; \frac{1}{4!} < \frac{1}{2^3}; \frac{1}{\alpha!} < \frac{1}{2^{\alpha-1}} \quad (16.98)$$

Nous pouvons donc écrire l'inégalité:

$$\left(1 + \frac{1}{\alpha}\right)^\alpha < 1 + 1 + \underbrace{\frac{1}{2} + \frac{1}{2^2} + \dots + \frac{1}{2^{\alpha-1}}}_{< 1} \quad (16.99)$$

Les termes soulignés constituent une progression géométrique de raison $q = 1/2$ (cf. chapitre de Suites et Séries) et dont le premier terme est 1. Par suite en utilisant les résultant obtenus dans le chapitre de Suites et Séries, nous pouvons écrire:

$$\left(1 + \frac{1}{\alpha}\right)^\alpha < 1 + \left[1 + \frac{1}{2} + \frac{1}{2^2} + \dots + \frac{1}{2^{\alpha-1}}\right] = 1 + \frac{1 - (1/2)^\alpha}{1 - (1/2)} = 1 + \left[2 - \left(\frac{1}{2}\right)^{\alpha-1}\right] < 3 \quad (16.100)$$

Par conséquent, nous avons:

$$2 \leq \left(1 + \frac{1}{\alpha}\right)^\alpha < 3 \quad (16.101)$$

Nous avons donc prouvé que la fonction $\left(1 + \frac{1}{\alpha}\right)^\alpha$ est bornée.

La limite:

$$e = \lim_{\alpha \rightarrow +\infty} \left(1 + \frac{1}{\alpha}\right)^\alpha \quad (16.102)$$

tend donc vers cette valeur bornée qui est le nombre e dont la valeur est :

$$e = 2.7182818284... \quad (16.103)$$

Remarque: Comme nous l'avons démontré dans le chapitre traitant des Nombres, ce nombre est irrationnel.

Nous pouvons alors définir la "**fonction exponentielle naturelle**" (réciproque de la fonction logarithme népérien) par:

$$f(x) = e^x \quad (16.104)$$

ou également parfois notée:

$$f(x) = \exp(x) \quad (16.105)$$

Le nombre e et la fonction qui permet de le déterminer sont très utiles. Nous les retrouvons dans tous les domaines de la mathématique et de la physique et donc dans la quasi-totalité des chapitres de ce site.

Les logarithmes ont plusieurs propriétés. Les voici (nous nous référons à une base X donnée):

$$\log_X(X^m \cdot X^n) = \log_X(X^{m+n}) = \log_X(X^m) + \log_X(X^n) = m + n \quad (16.106)$$

Si nous posons $X^m = a$ et $X^n = b$ nous avons donc:

$$\log_X(a \cdot b) = \log_X(a) + \log_X(b) \quad (16.107)$$

Si nous avons le cas particulier $a = b$ alors:

$$\log_X(a \cdot a) = \log_X(a) + \log_X(a) = 2 \cdot \log_X(a) \quad (16.108)$$

Cherchons à exprimer:

$$\log_X\left(\frac{a}{b}\right) \quad (16.109)$$

sous une forme différente. Posons $t = a/b \Leftrightarrow a = t \cdot b$ ce qui nous amène au développement:

$$\begin{aligned} \log_X(a) &= \log_X(t \cdot b) = \log_X(t) + \log_X(b) \\ \Rightarrow \log_X(t) &= \log_X(a) - \log_X(b) = \log_X\left(\frac{a}{b}\right) \end{aligned} \quad (16.110)$$

Cherchons à exprimer maintenant $\log_X\left(x^{\frac{1}{n}}\right)$ avec $n \in \mathbb{N}^*$ sous une forme différente. Posons:

$$t = x^{\frac{1}{n}} \Leftrightarrow x = t^n \quad (16.111)$$

ce qui nous amène à:

$$\log_X(x) = \log_X(t^n) = n \cdot \log_X(t) \Rightarrow \log_X(t) = \log_X(x^{\frac{1}{n}}) = \frac{1}{n} \log_X(x) \quad (16.112)$$

Il y a une relation assez utilisée en physique relativement aux changements de bases logarithmiques. La première relation est triviale et découle des propriétés algébriques des logarithmes:

$$a^x = e^{x \ln(a)} = 10^{x \log_{10}(a)} = b^{x \log_b(a)} \quad (16.113)$$

La seconde relation:

$$\log_a(u) = \frac{\ln(u)}{\ln(a)} = \frac{\log_{10}(u)}{\log_{10}(a)} = \frac{\log_b(u)}{\log_b(a)} \quad (16.114)$$

est un peu moins triviale et nécessite peut-être une démonstration (nous en aurons besoin lors de notre étude des fractions continues dans le chapitre de Théorie des nombres).

Démonstration:

Nous avons d'abord les équations équivalentes (de la première relation ci-dessus):

$$w = \log_a(u) \text{ et } a^w = u \quad (16.115)$$

et nous procédons comme suit:

$$\begin{aligned} a^w &= u \\ \log_b(a^w) &= \log_b(u) \quad (16.116) \\ w \log_b(a) &= \log_b(u) \end{aligned}$$

Ce qui nous amène finalement à:

$$w = \frac{\log_b(u)}{\log_b(a)} = \log_a(u) \quad (16.117)$$

□C.Q.F.D.

4. PRODUIT SCALAIRE FONCTIONNEL

Le produit scalaire fonctionnel (analogie très forte avec le produit scalaire vectoriel vu dans le chapitre de calcul vectoriel) peut paraître inutile lorsqu'il est étudié pour la première fois hors d'un contexte appliqué mais il connaît au fait de nombreuses applications pratiques. Nous en ferons par exemple directement usage dans le chapitre de physique quantique ondulatoire et de chimie quantique ou encore dans le cadre plus important encore des polynômes trigonométriques via les séries et transformées de Fourier (cf. chapitre sur les Suites Et Séries) que nous retrouvons partout dans la physique contemporaine.

Cependant, si le lecteur n'a pas encore parcouru le chapitre de calcul vectoriel et la partie y traitant du produit scalaire vectoriel, nous ne saurions que trop recommander sa lecture sans quoi ce qui va suivre risque d'être un peu incompréhensible.

Nous nous plaçons dans l'espace $\mathcal{C}([a, b], \mathbb{R})$ des fonctions continues de l'intervalle $[a, b]$ dans $\mathbb{R}$ muni du produit scalaire défini par (nous retrouvons la notation spécifique du produit scalaire dans sa version fonctionnelle comme nous en avons fait mention lors de notre définition du produit scalaire vectoriel):

$$\langle f | g \rangle = \int_a^b f(x) \cdot g(x) dx \quad (16.118)$$

Une famille de polynômes orthogonale, comme nous pouvons en faire l'analogie avec le produit scalaire

vu dans le chapitre de Calcul Vectoriel, est donc une famille $(p_0, \dots, p_n, \dots)$ de polynômes tels que:

$$\langle p_j | p_k \rangle = 0 \text{ si } j \neq k \quad (16.119)$$

Nous rappelons qu'une famille orthogonale est libre (cf. chapitre de Calcul Vectoriel).

Le développement suivant va nous rappeler le procédé de Gram-Schmidt (cf. chapitre de Calcul Tensoriel) pour construire une famille orthogonale:

Soit $(p_0, \dots, p_n, \dots)$ une famille de polynômes linéairement indépendants définis sur $[a, b]$ et V l'espace vectoriel engendré par cette famille. La famille $(y_0, \dots, y_n, \dots)$ définie par récurrence de la manière suivante:

$$y_n = p_n - \sum_{j=0}^{n-1} \langle p_n | y_j \rangle \frac{y_j}{\|y_j\|^2} \text{ pour } n \geq 1 \quad (16.120)$$

et $y_0 = p_0$ est orthogonale et engendre V .

Démonstration:

Montrons par récurrence sur n que $(y_0, \dots, y_n, \dots)$ est une famille orthogonale qui engendre le même espace que $(p_0, \dots, p_n, \dots)$. L'assertion est vérifiée pour $n = 0$. Supposons l'assertion vérifiée pour $n \geq 0$, pour $0 \leq k \leq n$ nous avons:

$$\begin{aligned} \langle y_{n+1} | y_k \rangle &= \left\langle p_{n+1} - \sum_{j=0}^n \langle p_{n+1} | y_j \rangle \frac{y_j}{\|y_j\|^2} \middle| y_k \right\rangle = \langle p_{n+1} | y_k \rangle - \sum_{j=0}^n \langle p_{n+1} | y_j \rangle \langle y_j | y_k \rangle \cdot \frac{1}{\|y_j\|^2} \\ &= \langle p_{n+1} | y_k \rangle - \langle p_{n+1} | y_k \rangle \cdot \langle y_k | y_k \rangle \cdot \frac{1}{\|y_k\|^2} = \langle p_{n+1} | y_k \rangle - \langle p_{n+1} | y_k \rangle = 0 \end{aligned} \quad (16.121)$$

$(y_0, \dots, y_n, \dots)$ est donc orthogonale. Pour finir, l'égalité:

$$y_{n+1} = p_{n+1} - \sum_{j=0}^n \langle p_{n+1} | y_j \rangle \frac{y_j}{\|y_j\|^2} \quad (16.122)$$

montre que $(y_0, \dots, y_{n+1})$ et $(p_0, \dots, p_{n+1})$ engendrent le même espace. $(y_0, \dots, y_{n+1})$ est donc bien une famille orthogonale qui engendre V .

□C.Q.F.D.

Exemple:

Considérons l'exemple très important en physique moderne qui est l'ensemble $P_{2\pi}$ des fonctions continues 2π -périodiques qui forme un espace vectoriel (cf. chapitre de Calcul Vectoriel).

Nous définissons donc le produit scalaire de deux fonctions de cet ensemble par:

$$\forall f, g \in P_{2\pi} \quad \langle f | g \rangle = \int_{-\pi}^{\pi} f(x)g(x)dx = \int_0^{2\pi} f(x)g(x)dx \quad (16.123)$$

Le but de cette étude est de construire une base de $P_{2\pi}$ sur laquelle nous pouvons décomposer toute fonction 2π -périodique.

L'idée la plus simple est alors de se servir des fonctions trigonométriques sinus et cosinus:

$$\begin{aligned} &\cos(x), \cos(2x), \cos(3x), \dots, \cos(nx) \\ &\sin(x), \sin(2x), \sin(3x), \dots, \sin(nx) \end{aligned} \quad (16.124)$$

Les relations ci-dessous montrent que les bases choisies ci-dessus sont orthogonales et forment donc une famille libre, de plus c'est une famille génératrice de l'espace vectoriel $P_{2\pi}$ car comme nous le démontrerons lors de notre étude des séries de Fourier (cf. chapitre sur les Suites Et Séries), nous avons les valeurs suivantes:

$$\begin{aligned} &\int_0^{2\pi} \cos(kx) \cos(mx) dx = \delta_{km} \pi \\ &\int_0^{2\pi} \sin(kx) \sin(mx) dx = \delta_{km} \pi \quad (16.125) \\ &\int_0^{2\pi} \cos(kx) \sin(mx) dx = 0 \end{aligned}$$

où δ_{km} est le symbole de Kronecker (cf. chapitre de Calcul Tensoriel).

Remarque: Si le lecteur se rappelle que pour une variable aléatoire X définie sur tout $\mathbb{R}$, son espérance se calculait comme étant (cf. chapitre de Statistiques):

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx \quad (16.126)$$

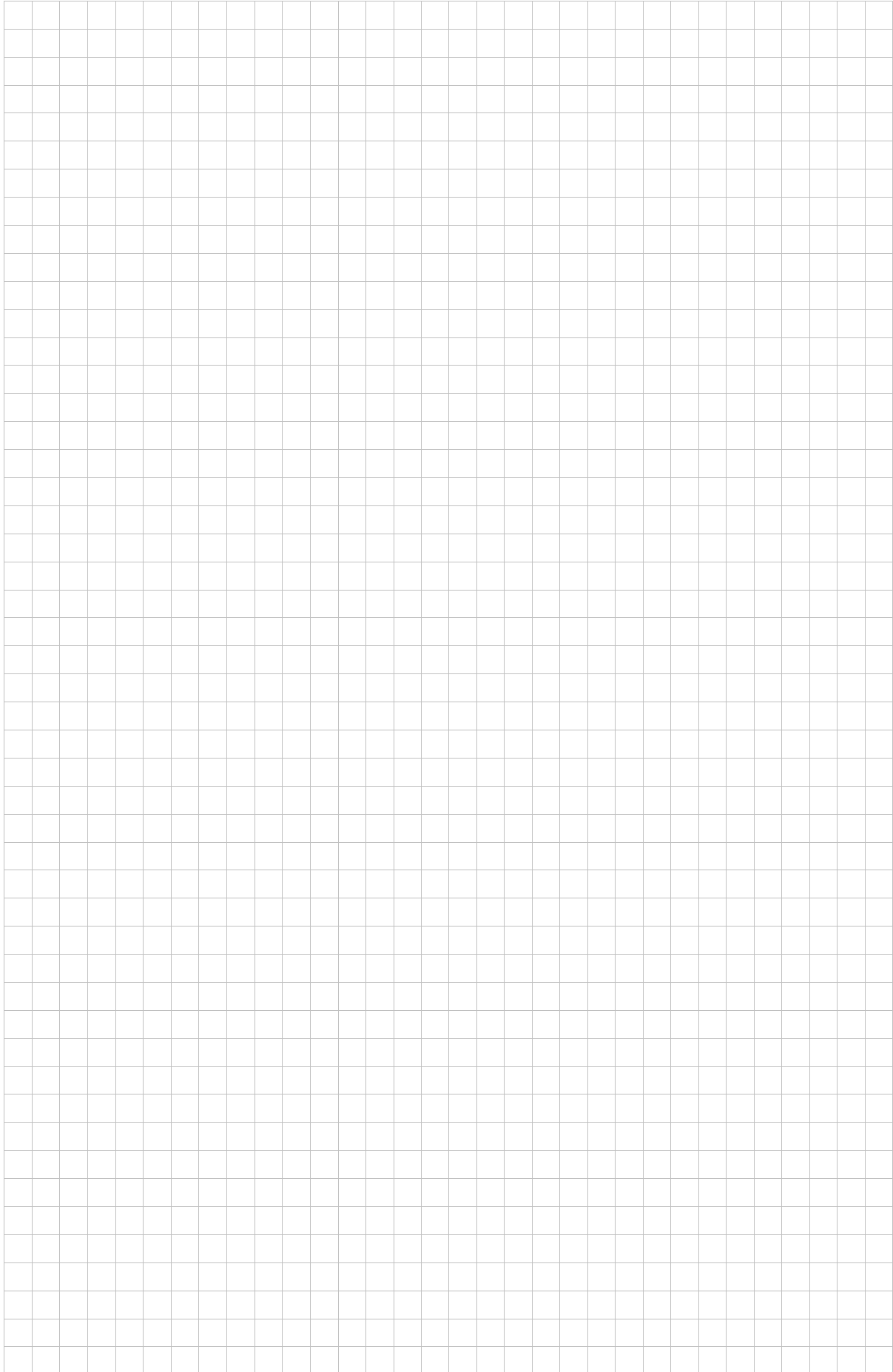
Donc nous pouvons assimiler:

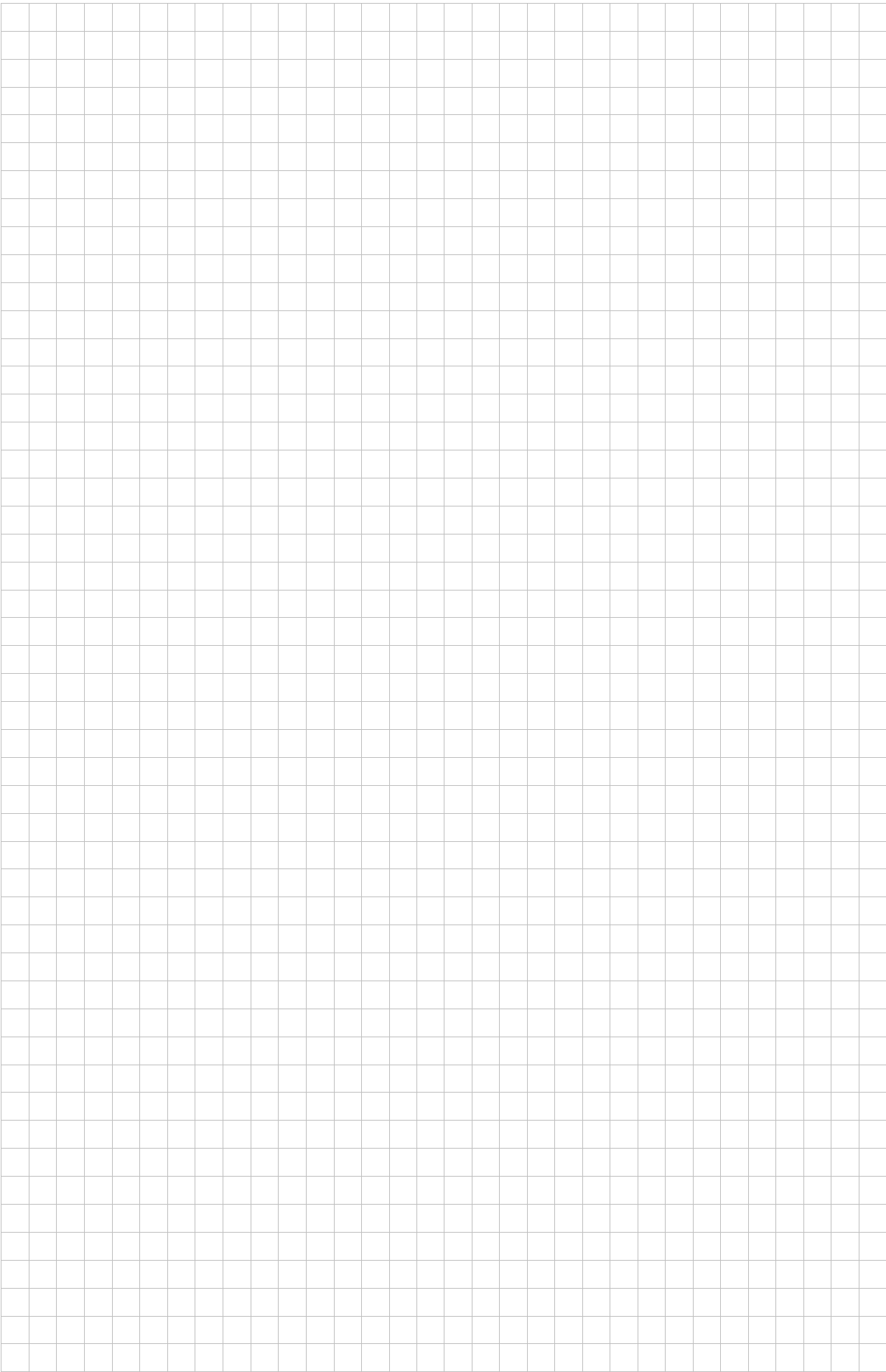
$$E(g(x)) = \int_{-\infty}^{+\infty} g(x)f(x)dx \quad (16.127)$$

où:

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \quad (16.128)$$

à l'espérance de la fonction $g(x)$! Analogie parfois forte intéressante dans la pratique!







17. ANALYSE COMPLEXE

Avant de commencer ce chapitre portant sur l'étude du calcul différentiel et intégral dans le cas généralisé de l'ensemble des nombres complexes, je tiens à signaler que je me suis beaucoup inspiré du PDF de E. Hairer (avec son autorisation) au niveau des illustrations. Le présent texte contient également de nombreuses phrases et développements repris, homogénéisés et simplifiés (au risque d'en faire grimper certains aux rideaux...) conformément aux notations et objectifs pédagogiques du reste de ce site Internet.

Le sujet de l'analyse complexe est donc l'étude des fonctions $\mathbb{C} \rightarrow \mathbb{C}$ et de leur différentiabilité (qui est différente de celle dans $\mathbb{R}^2$). Les "fonctions holomorphes" (c'est-à-dire différentiables dans un sous-ensemble de $\mathbb{C}$) possèdent nous le verrons des propriétés surprenantes, élégantes qui peuvent être réutilisées dans le cas des fonctions de $\mathbb{R}^2$ et qui ont des applications importantes en physique.

Avant de commencer expliquons l'intérêt de ce domaine de manière simplifiée!

Nous avons étudié dans la section d'Algèbre une partie du calcul différentiel et intégral avec quelques théorèmes utiles et importants pour la physique et l'ingénierie. Cependant, en restant dans $\mathbb{R}$ ou $\mathbb{R}^2$ la liste des théorèmes s'épuise en quelque sorte et on finit par ne plus trouver grand-chose de pertinent dans la pratique qui permette de simplifier le calcul d'intégrales que l'on retrouve souvent dans l'industrie. Alors, quand on sait que $\mathbb{R} \subset \mathbb{C}$ (donc l'ensemble des complexes généralise celui des réels) et que l'on peut construire aussi une correspondance $\mathbb{R}^2 \rightarrow \mathbb{C}$ comme nous allons le voir, de nouveaux théorèmes apparaissent avec des résultats très intéressants que l'on peut exploiter pour les intégrales dans $\mathbb{R}$ ou $\mathbb{R}^2$!! C'est cette raison qui fait que l'ingénieur a besoin de connaître l'analyse complexe!

Après l'étude de ce domaine particulier de la mathématique, il est fréquent de dire que le plus court chemin entre deux vérités du domaine réel passe souvent par le domaine complexe.

1. APPLICATIONS LINÉAIRES

Une bonne introduction à l'analyse complexe et à sa représentation consiste à étudier dans un premier temps (à titre pédagogique principalement) le cas particulier des applications linéaires complexes. Voyons cela:

Soient $U \subset \mathbb{C}$, un ensemble et $V \subset \mathbb{C}$ un autre ensemble. Une fonction qui associe à chaque $z \in U$ un $w \in V$:

$$w = f(z) = f(x+iy) \in V \quad (17.1)$$

est une "fonction complexe":

$$f: U \rightarrow V \quad (17.2)$$

Ce qui est important c'est de comprendre et remarquer que nous pouvons identifier:

$$z = x+iy \rightarrow (x, y) \in \mathbb{R}^2 \quad (17.3)$$

et:

$$w = f(z) = u(x, y) + iv(x, y) \rightarrow (u, v) \in \mathbb{R}^2 \quad (17.4)$$

Nous arrivons alors à deux fonctions de deux variables réelles x, y :

$$u(x, y), v(x, y) \quad (17.5)$$

qui sont les coordonnées du point w .

Définition: Une application est dite $\mathbb{C}$ -linéaire si par exemple une fonction du type:

$$w = f(z) = c \cdot z \quad (17.6)$$

où c est un nombre complexe fixé et z un nombre complexe quelconque, satisfait:

$$f(cz_1 + cz_2) = cf(z_1) + cf(z_2) \quad (17.7)$$

Nous avons vu et démontré dans le chapitre Nombres lors de notre étude des nombres complexes, que la multiplication de deux nombres complexes pouvait être équivalente à une rotation orthogonale suivie d'une homothétie et que cette même multiplication pouvait être représentée sous forme matricielle! Or la transcription sous forme matricielle implique comme nous l'avons vu dans le chapitre d'Algèbre Linéaire automatiquement la linéarité!

Donc lecteur pourra facilement vérifier qu'une matrice de rotation/homothétie est un exemple d'une application $\mathbb{C}$ -linéaire (sur demande nous pouvons détailler) que nous écrirons dorénavant:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = s \cdot \begin{bmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (17.8)$$

Ce qui se représente typiquement de la manière suivante (on y observe bien une rotation et une homothétie qui conservent les angles et les proportions):

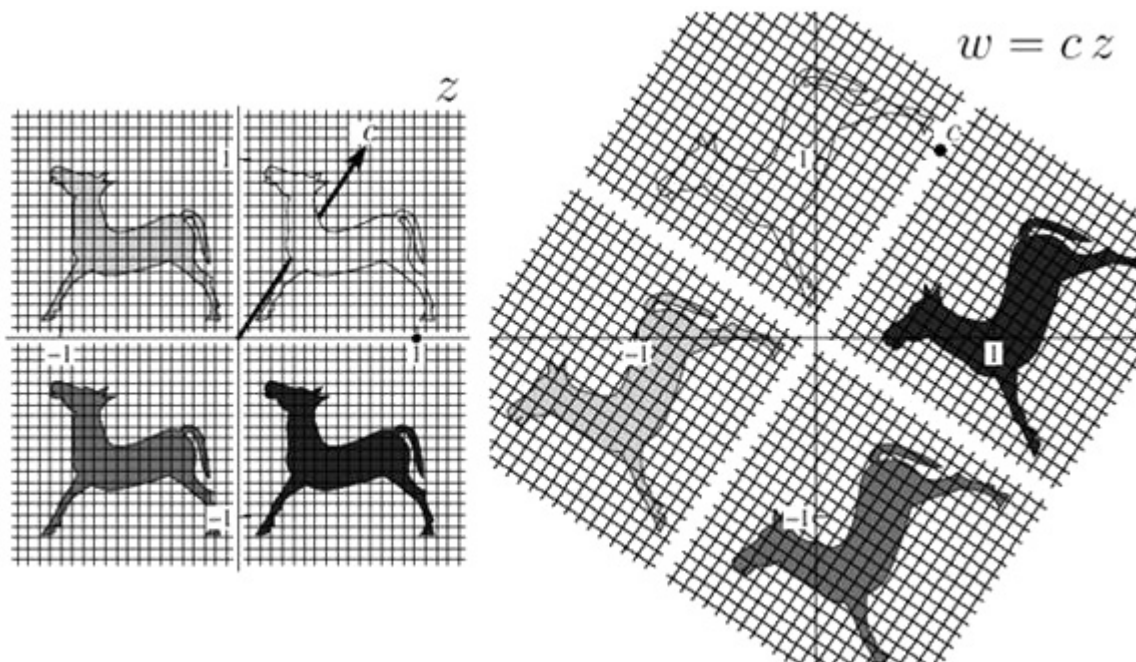


Figure: 17.1 - Exemple d'application d'une fonction complexe plane

C'est le fait que les proportions et que les angles soient conservés qui fait d'une fonction complexe qu'elle est $\mathbb{C}$ -linéaire. Dans le cas contraire, nous dirions que la fonction est $\mathbb{R}$ -linéaire.

Donc une matrice 2×2 représente une application $\mathbb{C}$ -linéaire seulement si elle est de la forme:

$$A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \quad (17.9)$$

Voyons des exemples de fonctions $\mathbb{C}$ -linéaires assez remarquables.

Exemples:

E1.

$$f(z) = z^2 \quad (17.10)$$

En coordonnées réelles cela donne:

$$u + iv = (x + iy)^2 = x^2 - y^2 + 2ixy \Rightarrow \begin{cases} u = x^2 - y^2 \\ v = 2xy \end{cases} \quad (17.11)$$

Ainsi, regardons ce que fait cette fonction avec les points du plan complexe qui sont confondus avec les lignes verticales de ce même plan (ce qui vient alors à poser $x = a$). Nous avons alors:

$$u = a^2 - y^2 \quad v = 2ay \quad (17.12)$$

et en éliminant y , nous trouvons l'équation d'une parabole ou plutôt d'une famille de paraboles (pour plusieurs valeurs de a) qui sont ouvertes à gauche du plan complexe image:

$$u = a^2 - \frac{v^2}{4a^2} \quad (17.13)$$

Si nous faisons la même analyse pour les points du plan complexe qui sont confondus avec lignes horizontales de ce même plan, nous trouvons également, en posant $y = b$, une famille de paraboles (pour plusieurs valeurs de b) qui sont ouvertes à droite du plan complexe image.

Voici une représentation du plan complexe image sur lequel nous avons dessiné une tête de chat:

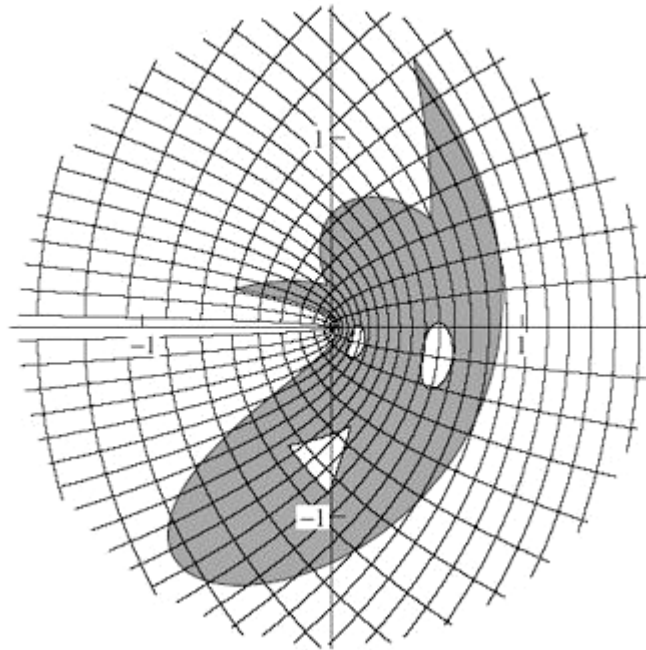


Figure: 17.2 - Représentation complexe de l'image de la fonction exemple

et si nous regardons le plan complexe pré-image correspondant, nous avons alors deux têtes de chats qui apparaissent:

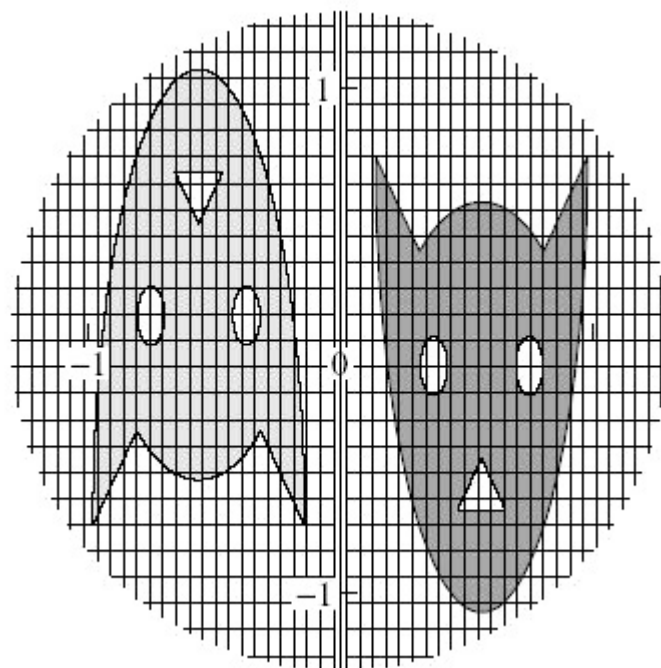


Figure: 17.3 - Représentation des pré-images de la fonction exemple

L'apparition de ces deux têtes de chats vient du fait que cette fonction possède 2 pré-images possibles pour chaque point image (c'est donc une fonction surjective).

E2. Une autre fonction intéressante est la "**transformation de Cayley**" utilisée dans certains domaines de la physique et définie par:

$$f(z) = \frac{z+1}{z-1} \quad (17.14)$$

ayant comme domaine de définition: $\mathbb{C}/\{1\}$.

On remarquera qu'il s'agit d'une fonction involutive puisque:

$$f(f(z)) = f\left(\frac{z+1}{z-1}\right) = \frac{\frac{z+1}{z-1} + 1}{\frac{z+1}{z-1} - 1} = \frac{\frac{z+1+(z-1)}{z-1}}{\frac{z+1-(z-1)}{z-1}} = \frac{z+1+(z-1)}{z+1-(z-1)} = \frac{2z}{2} = z \quad (17.15)$$

et comme nous avons démontré dans le chapitre de Théorie De La Démonstration que toute fonction involutive est à la fois injective et surjective, alors la transformation de Cayley est une fonction bijective.

Cette fonction transforme l'axe des imaginaires iy en cercle unité (et inversement puisqu'elle est involutive). Voyons cela:

$$\begin{aligned} f(z) = f(iy) &= \frac{iy+1}{iy-1} = \frac{iy+1}{iy-1} \cdot \frac{-iy-1}{-iy-1} = \frac{y^2 - iy - iy - 1}{y^2 - iy + iy + 1} = \frac{y^2 - 2iy - 1}{1 + y^2} \\ &= \frac{y^2 - 1}{1 + y^2} - i \frac{2y}{1 + y^2} = u + iv \end{aligned} \quad (17.16)$$

où:

$$u = \frac{y^2 - 1}{1 + y^2} \quad v = -\frac{2y}{1 + y^2} \quad (17.17)$$

satisfont:

$$u^2 + v^2 = \frac{(y^2 - 1)^2 + (-2y)^2}{(1 + y^2)^2} = \frac{(y^4 - 2y^2 + 1) + (4y^2)}{1 + 2y^2 + y^4} = 1 \quad (17.18)$$

Soit:

$$u^2 + v^2 = 1 \quad (17.19)$$

Il s'agit donc bien de l'équation d'un cercle.

E3. Comme autre exemple de fonction, prenons la "**transformation de Joukowski**" définie par:

$$f(z) = \frac{1}{2} \left(z + \frac{1}{z} \right) \quad (17.20)$$

Si le domaine de définition donné est construit en coordonnées polaires regardons comment un cercle ou une ellipse se transforme:

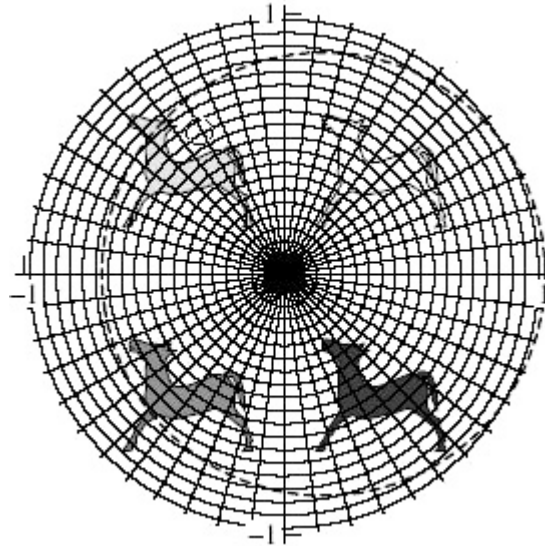


Figure: 17.4 - Transformation en coordonnées polaires d'une ellipse avec la fonction exemple

Alors le plan image sera:

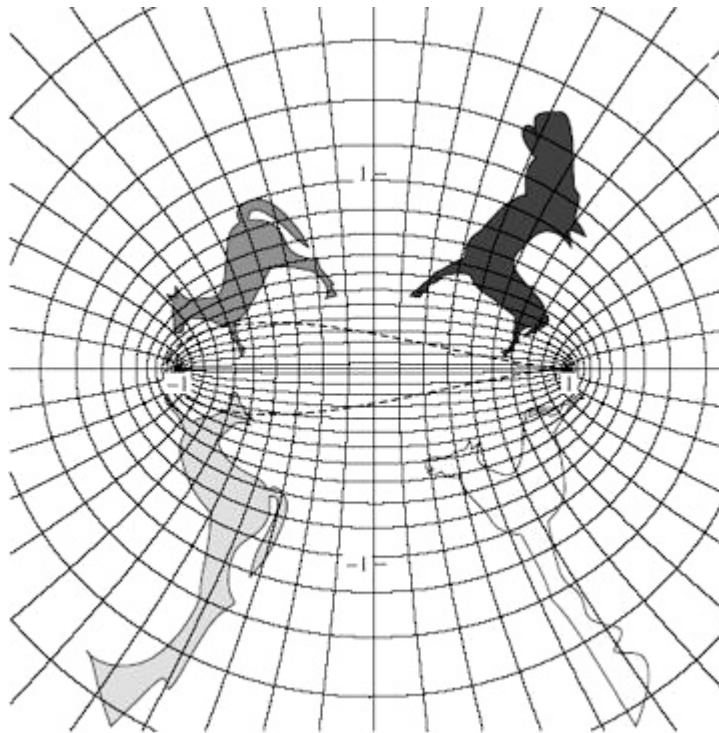


Figure: 17.5 - Résultat de la transformation en coordonnées polaires

Elle transforme donc respectivement les cercles centrés en 0 et les rayons passant par 0 en une famille d'ellipses et d'hyperboles cofocales. Pour démontrer ce fait, nous utilisons donc les coordonnées polaires complexes (formule d'Euler) vues dans le chapitre sur les Nombres:

$$z = r(\cos \varphi + i \sin \varphi) \quad (17.21)$$

et:

$$\begin{aligned}
 z^{-1} &= \frac{1}{r(\cos \varphi + i \sin \phi)} = r^{-1} \frac{1}{(\cos \varphi + i \sin \phi)} \frac{(\cos \varphi - i \sin \phi)}{(\cos \varphi - i \sin \phi)} = r^{-1} \frac{\cos \varphi - i \sin \phi}{\cos^2 \varphi + \sin^2 \phi} \quad (17.22) \\
 &= r^{-1} (\cos \varphi - i \sin \phi)
 \end{aligned}$$

Nous avons alors:

$$f(z) = \frac{1}{2} \left(r(\cos \varphi + i \sin \phi) + \frac{1}{r}(\cos \varphi - i \sin \phi) \right) = \left(\frac{r}{2} + \frac{1}{2r} \right) \cos \varphi + i \left(\frac{r}{2} - \frac{1}{2r} \right) \sin \phi \quad (17.23)$$

d'où:

$$u = \left(\frac{r}{2} + \frac{1}{2r} \right) \cos \varphi \quad v = \left(\frac{r}{2} - \frac{1}{2r} \right) \sin \phi \quad (17.24)$$

et nous voyons immédiatement que:

$$\frac{u^2}{\left(\frac{r}{2} + \frac{1}{2r} \right)^2} + \frac{v^2}{\left(\frac{r}{2} - \frac{1}{2r} \right)^2} = 1 \quad (17.25)$$

a la forme de l'équation d'une ellipse (cf. chapitre de Géométrie Analytique) et nous avons de même:

$$\begin{aligned}
 \frac{u^2}{\cos^2 \varphi} - \frac{v^2}{\sin^2 \phi} &= \left(\frac{r}{2} + \frac{1}{2r} \right)^2 - \left(\frac{r}{2} - \frac{1}{2r} \right)^2 = \left(\frac{r^2}{4} + 2 \frac{r}{2} \frac{1}{2r} + \frac{1}{4r^2} \right) - \left(\frac{r^2}{4} - 2 \frac{r}{2} \frac{1}{2r} + \frac{1}{4r^2} \right) \quad (17.26) \\
 &= 2 \frac{r}{2} \frac{1}{2r} + 2 \frac{r}{2} \frac{1}{2r} = \frac{1}{2} + \frac{1}{2} = 1
 \end{aligned}$$

qui est l'équation d'une hyperbole (cf. chapitre de Géométrie Analytique).

Cette fonction trouve son utilité dans le cas où si nous plaçons astucieusement un cercle (passant par le point $z = 1$ comme dans le cas de la première figure) le plan représenté en coordonnées polaires avec un trait discontinu pourrait ressembler à une aile d'avion. Ce qui permettait à une époque (mais la technique est obsolète aujourd'hui) en aérodynamique de transposer l'étude d'un champ de vecteurs du profil d'une aile d'avion à l'étude du profil d'un cercle et de faire par la suite la transformation de Joukowski.

Effectivement, voyons une partie de cela avec Maple 4.00b:

```

> assume(x,real,y,real);
> z:=x+I*y;
> F:=1/2*(z+1/z);

> u:=Re(F);
> u:=evalc(u);

> v:=Im(F);
> v:=evalc(v);

```

```
> with(plots):with(plottools):
> p1:=disk([0,0],1,color=black):
> p2:=implicitplot({seq(v=b/8,b=-10..10)},x=-4..4,y=-2..2,color=black):
> display([p2,p1],scaling=constrained);
```

Nous obtenons alors...:

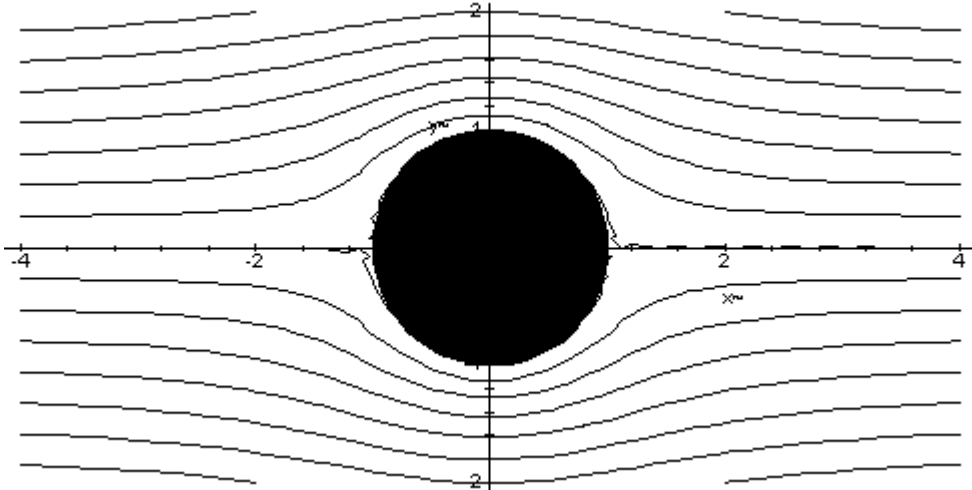


Figure: 17.6 - Exemple d'application importante de la fonction exemple

E4. Faisons un dernier exemple avec la fonction $1/z$ encore avec Maple 4.00b. Si vous y saisissez les commandes suivantes (suffisamment explicites de par leur nom pour comprendre ce qu'elles font):

```
> assume(x,real,y,real);
> z:=x+I*y;
> F:=1/z;

> u:=Re(F);
> u:=evalc(u);

> v:=Im(F);
> v:=evalc(v);

> with(plots):
> p1:=implicitplot({seq(u=a,a=-5..5)},x=-1..1,y=-1..1,numpoints=1000):
> p2:=implicitplot({seq(v=b,b=-5..5)},x=-1..1,y=-1..1,numpoints=1000,color=green):
> display([p1,p2],scaling=constrained);
```

Nous obtenons!:

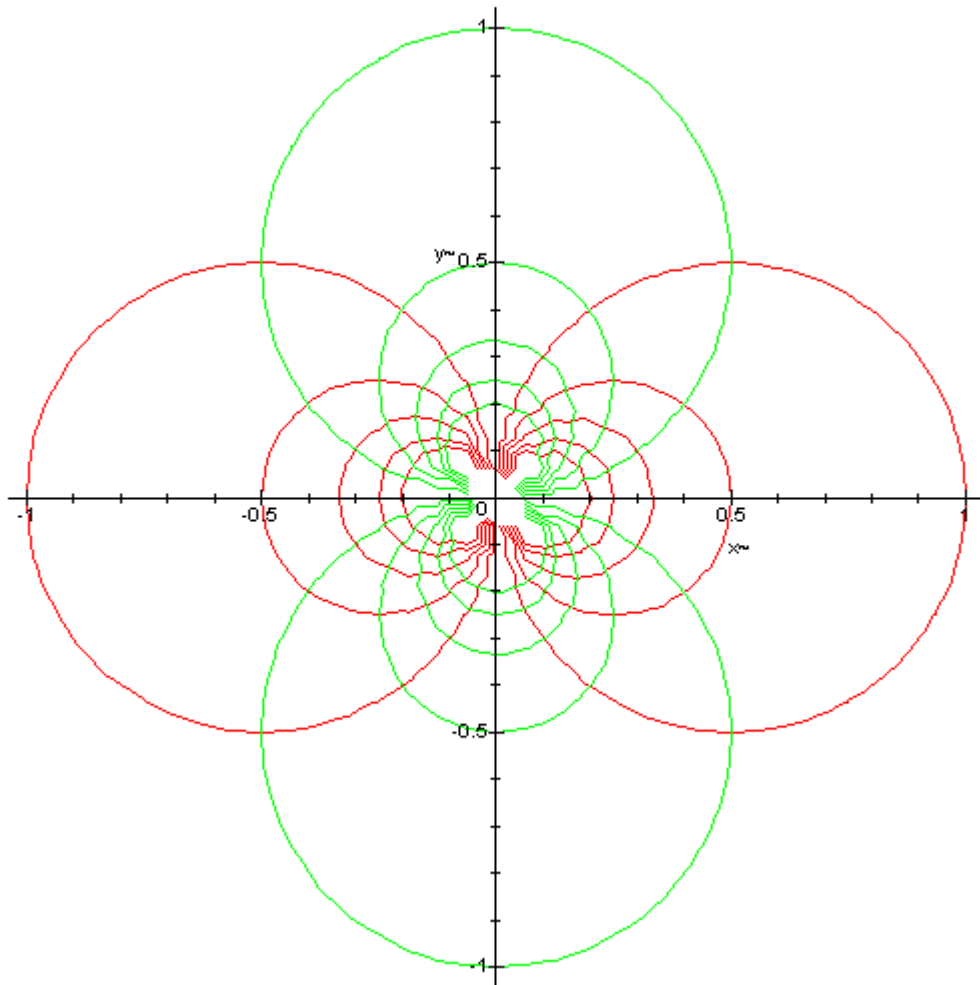


Figure: 17.7 - Autre exemple d'application importante avec une fonction complexe

voilà pour ceux qui souhaiteraient faire eux-mêmes des figures de fonctions complexes!

2. FONCTIONS HOLOMORPHES

La définition de la dérivation par rapport à une variable complexe est naturellement formellement identique à la dérivation par rapport à une variable réelle.

Nous avons alors, si la fonction est dérivable en z_0 :

$$\frac{df_h}{dz} = \lim_{z \rightarrow z_0} \frac{f(z) - f(z_0)}{z - z_0} = \lim_{h \rightarrow 0} \frac{f(z_0 + h) - f(z_0)}{(z_0 + h) - z_0} = \lim_{h \rightarrow 0} \frac{f(z_0 + h) - f(z_0)}{h} = f'(z) \quad (17.27)$$

et nous disons (abusivement dans le cadre de ce site) que la fonction est "**holomorphe**" (alors que dans $\mathbb{R}$ on dit "dérivable") ou "**analytique**" dans son domaine de définition ou dans un sous-ensemble de celui-ci si elle y est dérivable en chaque point.

Remarque:

R1. Une fonction complexe se dérive comme une fonction réelle, il suffit de poser z comme étant $x...$ à condition que ce que nous allons voir dans ce qui va suivre soit respecté!

R2. Au fait si la fonction est holomorphe dans un sous-ensemble du plan complexe, nous verrons un peu plus loin lors de notre étude de la convergence des séries de puissances qu'il s'agit toujours d'un sous-ensemble ouvert.

D'une manière équivalente, nous disons que la fonction f est $\mathbb{C}$ -différentiable en z_0 si la limite suivante existe dans $\mathbb{C}$:

$$\lim_{h \rightarrow 0} \frac{f(z_0 + h) - f(z_0)}{h} = f'(z_0) \quad (17.28)$$

Présentons maintenant un théorème central pour l'analyse complexe appelé "**théorème de Cauchy-Riemann**".

Si la fonction:

$$f(z) = f(x, y) = f(x + iy) = u(x, y) + iv(x, y) \quad (17.29)$$

est $\mathbb{C}$ -différentiable, en $z_0 = x_0 + iy_0$, alors nous avons:

$$\boxed{\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}} \quad (17.30)$$

qui est un peu l'équivalent du théorème de Schwarz dans $\mathbb{R}$ vu dans le chapitre de Calcul Différentiel Et Intégral. Ces deux dernières relations sont appelées "**conditions de Cauchy**". Donc ce sont les deux conditions que doivent vérifier une fonction complexe pour être dérivable en z_0 . Ainsi, il est possible d'utiliser ces deux relations pour étudier les points où la fonction n'est pas analytique.

Si ces conditions sont justes (ce que nous allons de suite démontrer), alors nous en déduisons que u et v doivent être toutes deux des fonctions harmoniques de x et y .

Démonstration:

Puisque:

$$\lim_{z \rightarrow z_0} \frac{f(z) - f(z_0)}{z - z_0} = f'(z_0) \quad (17.31)$$

En choisissant:

$$z = z_0 + x \quad (17.32)$$

avec $x \in \mathbb{R}$, nous obtenons:

$$f'(z_0) = \lim_{x \rightarrow x_0} \frac{f(x_0 + x, y_0) - f(x_0, y_0)}{x} = \lim_{x \rightarrow x_0} \frac{f(x_0 + x, y_0) - f(x_0, y_0)}{x} \quad (17.33)$$

et quand x tend vers une petite valeur dx , nous avons (cf. chapitre de Calcul Différentiel Et Intégral):

$$\begin{aligned} f'(z_0) &= \lim_{dx \rightarrow 0} \frac{f(x_0 + dx, y_0) - f(x_0, y_0)}{dx} = \frac{\partial f(z_0)}{\partial x} = \frac{\partial(u(x_0, y_0) + iv(x_0, y_0))}{\partial x} \\ &= \frac{\partial}{\partial x} u(x_0, y_0) + i \frac{\partial}{\partial x} v(x_0, y_0) \end{aligned} \quad (17.34)$$

et en choisissant:

$$z = z_0 + iy \quad (17.35)$$

avec $y \in \mathbb{R}$, nous obtenons:

$$f'(z_0) = \lim_{y \rightarrow y_0} \frac{f(x_0, y_0 + iy) - f(x_0, y_0)}{iy} = \lim_{y \rightarrow y_0} \frac{f(x_0, y_0 + iy) - f(x_0, y_0)}{iy} \quad (17.36)$$

et quand y tend vers une petite valeur dy nous avons (cf. chapitre de Calcul Différentiel Et Intégral):

$$\begin{aligned} f'(z_0) &= \lim_{dy \rightarrow 0} \frac{f(x_0, y_0 + idy) - f(x_0, y_0)}{idy} = \frac{1}{i} \frac{\partial f(z_0)}{\partial y} = -i \frac{\partial f(z_0)}{\partial y} \\ &= -i \frac{\partial(u(x_0, y_0) + iv(x_0, y_0))}{\partial y} = -i \left(\frac{\partial}{\partial y} u(x_0, y_0) + i \frac{\partial}{\partial y} v(x_0, y_0) \right) \\ &= \frac{\partial}{\partial y} v(x_0, y_0) - i \frac{\partial}{\partial y} u(x_0, y_0) \end{aligned} \quad (17.37)$$

Nous avons donc maintenant:

$$\begin{aligned} \left(\frac{\partial f}{\partial x} \right)_y &= \frac{\partial f}{\partial x} = \frac{\partial}{\partial x} u(x_0, y_0) + i \frac{\partial}{\partial x} v(x_0, y_0) \\ \left(\frac{\partial f}{\partial y} \right)_x &= \frac{\partial f}{\partial y} = \frac{\partial}{\partial y} v(x_0, y_0) - i \frac{\partial}{\partial y} u(x_0, y_0) \end{aligned} \quad (17.38)$$

Or nous avons démontré dans le chapitre de Calcul Différentiel Et Intégral le théorème de Schwarz suivant:

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x} \quad (17.39)$$

Dès lors:

$$\begin{aligned}\frac{\partial^2 f}{\partial y \partial x} &= \frac{\partial}{\partial y} \left(\frac{\partial}{\partial x} u(x_0, y_0) + i \frac{\partial}{\partial x} v(x_0, y_0) \right) \\ \frac{\partial^2 f}{\partial x \partial y} &= \frac{\partial}{\partial x} \left(\frac{\partial}{\partial y} v(x_0, y_0) - i \frac{\partial}{\partial y} u(x_0, y_0) \right)\end{aligned}\quad (17.40)$$

Soit:

$$\frac{\partial}{\partial y} \left(\frac{\partial}{\partial x} u(x_0, y_0) + i \frac{\partial}{\partial x} v(x_0, y_0) \right) = \frac{\partial}{\partial x} \left(\frac{\partial}{\partial y} v(x_0, y_0) - i \frac{\partial}{\partial y} u(x_0, y_0) \right) \quad (17.41)$$

Ce qui peut s'écrire:

$$\frac{\partial}{\partial y} \left(\frac{\partial}{\partial x} u(x_0, y_0) + i \frac{\partial}{\partial x} v(x_0, y_0) \right) - \frac{\partial}{\partial x} \left(\frac{\partial}{\partial y} v(x_0, y_0) - i \frac{\partial}{\partial y} u(x_0, y_0) \right) = 0 \quad (17.42)$$

Une solution triviale est d'avoir:

$$\begin{aligned}\frac{\partial}{\partial x} u(x_0, y_0) + i \frac{\partial}{\partial x} v(x_0, y_0) &= 0 \\ \frac{\partial}{\partial y} v(x_0, y_0) - i \frac{\partial}{\partial y} u(x_0, y_0) &= 0\end{aligned}\quad (17.43)$$

Soit la possibilité d'écrire:

$$\frac{\partial}{\partial x} u(x_0, y_0) + i \frac{\partial}{\partial x} v(x_0, y_0) = \frac{\partial}{\partial y} v(x_0, y_0) - i \frac{\partial}{\partial y} u(x_0, y_0) \quad (17.44)$$

En identifiant parties réelles et imaginaires, nous terminons la démonstration!

□C.Q.F.D.

Donc pour que f soit dérivable au sens complexe (holomorphe) en un point, il suffit qu'elle y soit différentiable comme fonction de deux variables réelles ($\mathbb{R}$ -différentiable en (x_0, y_0)) et que ses dérivées premières partielles en ce point vérifient les équations de Cauchy-Riemann.

Par contre, pour qu'elle soit $\mathbb{C}$ -différentiable, il faut que les équations de Cauchy-Riemann soient valables en tous les points du plan complexe (on parle alors parfois de "**fonctions entières**") et non pas seulement dans un sous-domaine de celui-ci! Dans le cas contraire, elle contient donc des "**singularités**" et nous parlons alors de "**fonction méromorphe**" (qui est donc une fonction holomorphe sauf sur les points de singularités).

Remarque: Géométriquement, nous montrerons plus tard qu'une fonction holomorphe a une interprétation possible dans le sens qu'elle est une transformation conforme (conserve les angles).

Signalons donc que si $f(z)$ est $\mathbb{C}$ -différentiable alors elle peut être développée en série de Taylor aussi (cf. chapitre de Suites et Séries):

$$f(z) = \sum_{n=0}^{+\infty} \frac{1}{n!} \frac{d^n f(z_0)}{dz^n} (z - z_0)^n \quad (17.45)$$

Remarquons une chose importante aussi. Si nous réécrivons:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y} \quad (17.46)$$

sous la forme suivante:

$$\begin{aligned} \frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} &= 0 \\ \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} &= 0 \end{aligned} \quad (17.47)$$

Nous disons alors que la fonction f est irrotationnelle (cf. chapitre de Calcul Vectoriel) puisque la première relation peut être vue comme:

$$\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} = \text{rot}(\vec{f})_z = 0 \quad (17.48)$$

ce qui est une analogie non anodine! Enfin, la deuxième relation:

$$\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} = 0 \quad (17.49)$$

permet également de dire par analogie (mais cela s'arrête à une simple analogie!) que la fonction f est non divergente (cf. chapitre de Calcul Vectoriel) ce qui est bon moyen mnémotechnique de s'en souvenir.

Mettons également autre chose en évidence. Si nous reprenons les deux équations de Cauchy-Riemann:

$$\begin{aligned} \frac{\partial u}{\partial x} &= \frac{\partial v}{\partial y} \\ \frac{\partial u}{\partial y} &= -\frac{\partial v}{\partial x} \end{aligned} \quad (17.50)$$

et que nous les dérivons encore une fois ainsi:

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} &= \frac{\partial^2 v}{\partial x \partial y} \\ \frac{\partial^2 u}{\partial y^2} &= -\frac{\partial^2 v}{\partial y \partial x} \end{aligned} \quad (17.51)$$

et que nous sommons ces deux relations, nous avons alors:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad (17.52)$$

Il en est de même avec v . Nous avons alors:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0 \quad (17.53)$$

Et nous connaissons très bien cette forme d'équations (équation de Maxwell-Poisson dans le chapitre d'Électrodynamique et de Newton-Poisson dans celui d'Astronomie...). Il s'agit d'une équation d'onde appelée aussi "**équation de Laplace**" (rien à voir avec celle vue lors de notre étude de l'hydrostatique!) et donnée par le laplacien scalaire (cf. chapitre de Calcul Vectoriel):

$$\Delta \cdot u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad \Delta \cdot v = \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0 \quad (17.54)$$

Il est alors de tradition de dire que u est harmonique et nous pouvons arriver bien évidemment au même résultat avec v ! Bon évidemment... nous le savions, puisque nous avons déjà étudié dans le chapitre sur les Nombres que les parties réelles et imaginaires d'un nombre complexe pouvaient être mises sous forme trigonométrique.

Grâce à cette découverte, Riemann a ouvert l'application des fonctions holomorphes à de nombreux problèmes de la physique, puisque ces dernières équations sont satisfaites par le potentiel gravitationnel (équation de Newton-Poisson dans le chapitre d'Astronomie), par les champs électriques et magnétiques (équation de Maxwell-Poisson dans le chapitre d'Électrodynamique), par la chaleur en équilibre (par encore d'exemples sur le site) et par les mouvements sans rotationnel de certains fluides (pas encore d'exemples non plus sur le site).

Exemple:

Le potentiel d'un dipôle peut être décrit par la fonction holomorphe:

$$f(z) = \frac{1}{z+1} + \frac{1}{z-1} \quad (17.55)$$

La figure ci-dessous:

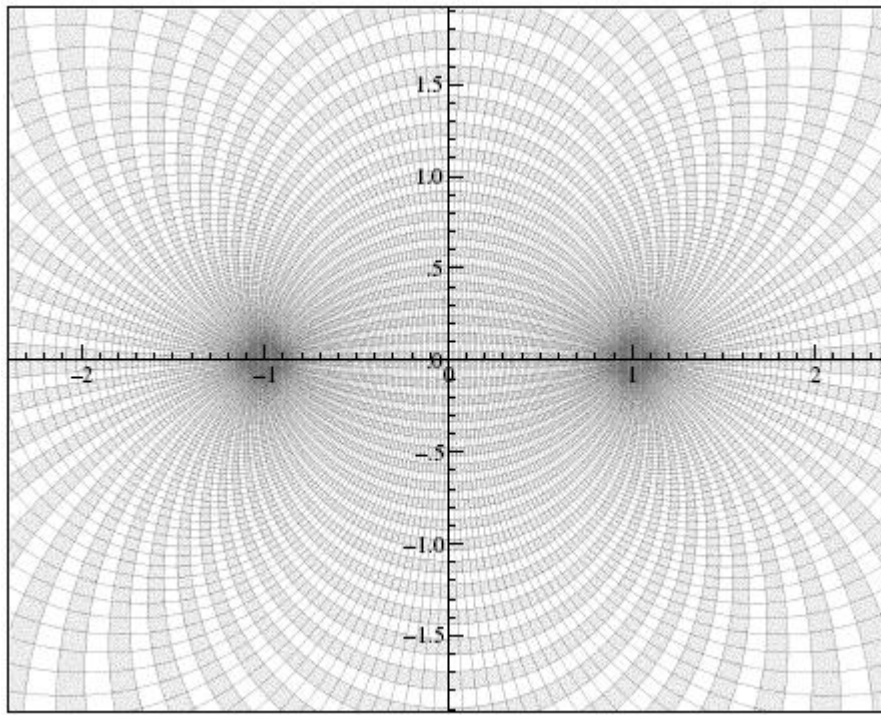


Figure: 17.8 - Représentation plane d'une fonction holomorphe bien connue...

montre les courbes de niveau des fonctions harmoniques $u(x, y)$ et $v(x, y)$ données comme parties réelle et complexe de la fonction $f(z)$ de cet exemple.

2.1. ORTHOGONALITÉ DES ISO-COURBES RÉELLES ET IMAGINAIRES

Nous allons maintenant démontrer une propriété sympathique que les fonctions qui satisfont les conditions de Cauchy (donc les fonctions analytiques!) ont. Effectivement, rappelez-vous que nous avons vu plus haut la fonction:

$$f(z) = z^2 \quad (17.56)$$

qui donnait donc le diagramme suivant:

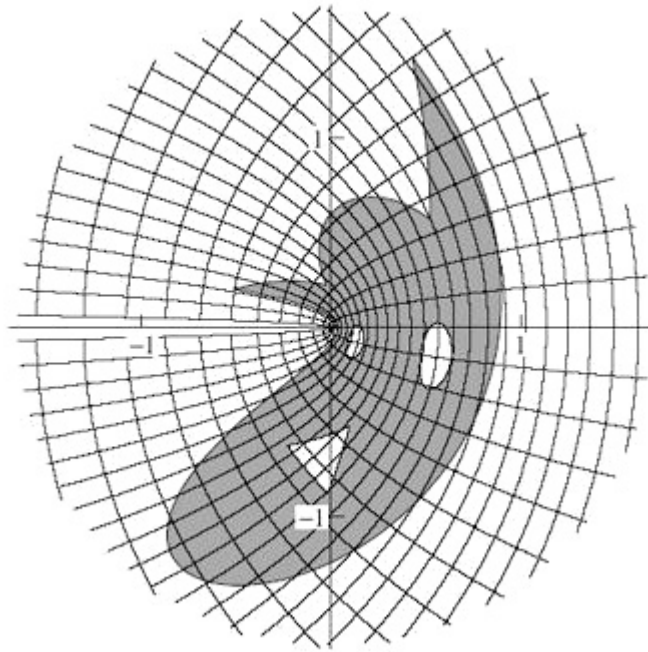


Figure: 17.9 - Rappel de la représentation plane des images d'une fonction complexe vue plus haut

Eh bien les fonctions satisfaisant les conditions de Cauchy ont la propriété géométrique simple suivante: les lignes dont la partie réelle de la fonction est constante $\Re(f(z)) = c^{te}$ et les lignes dont la partie imaginaire est constante $\Im(f(z)) = c^{te}$ sont orthogonales les unes aux autres (pensez à la forme trigonométrique des nombres complexes cela aide à mieux visualiser!).

En d'autres termes, les fonctions complexes analytiques sont des fonctions de transformation d'un domaine du plan dans un plan où les angles sont conservés. Nous disons alors que la fonction est une **"transformation conforme"**.

Pour la démonstration rappelons que nous avons démontré dans le chapitre de Calcul Vectoriel que le gradient d'une fonction f de $\mathbb{R}^2$ est donné par:

$$\vec{\nabla}(f) = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right) \quad (17.57)$$

et dans le cadre de notre étude des isoclines dans le chapitre de Géométrie Différentielle que le vecteur tangent aux isoclines de la fonction f sera toujours parallèle au vecteur du plan:

$$\vec{t} = \left(\frac{\partial f}{\partial y}, -\frac{\partial f}{\partial x} \right) \quad (17.58)$$

et que ces deux derniers vecteurs sont perpendiculaires tels que:

$$\vec{\nabla}(f) \circ \vec{t} = 0 \quad (17.59)$$

Assimilons maintenant le vecteur (parallèle) tangent $\vec{t}_u$ aux isoclines réelles:

$$\Re(f(z)) = u(x, y) = c^{te} \quad (17.60)$$

avec:

$$\vec{\xi}_u = \left(\frac{\partial u}{\partial y}, -\frac{\partial u}{\partial x} \right) \quad (17.61)$$

et le vecteur normal aux isoclines imaginaires:

$$\Im(f(z)) = v(x, y) = c^{te} \quad (17.62)$$

avec le gradient de v de composantes:

$$\vec{\nabla}(v) = \left(\frac{\partial v}{\partial x}, \frac{\partial v}{\partial y} \right) \quad (17.63)$$

En utilisant les conditions de Cauchy démontrées plus haut, nous avons pour cette dernière relation:

$$\vec{\nabla}(v) = \left(\frac{\partial v}{\partial x}, \frac{\partial v}{\partial y} \right) = \left(-\frac{\partial u}{\partial y}, \frac{\partial u}{\partial x} \right) \quad (17.64)$$

En comparant:

$$\vec{\xi}_u = \left(\frac{\partial u}{\partial y}, -\frac{\partial u}{\partial x} \right) \text{ et } \vec{\nabla}(v) = \left(-\frac{\partial u}{\partial y}, \frac{\partial u}{\partial x} \right) \quad (17.65)$$

nous voyons donc que $\vec{\xi}_u$ et $\vec{\nabla}(v)$ sont parallèles (colinéaires). Et puisque $\vec{\xi}_u$ est colinéaire aux isoclines réelles et que $\vec{\nabla}(v)$ est perpendiculaire aux isoclines imaginaires, nous avons terminé notre démonstration.

Le lecteur pourra prendre comme exemple la fonction:

$$f(z) = z^2 \quad (17.66)$$

détaillée mathématiquement et schématiquement plus haut! Mais pour changer un peu, prenons un exemple qui nous accompagnera tout au long du reste de ce chapitre et qui est la fonction holomorphe suivante:

$$f(z) = \frac{1}{1+z^2} \quad (17.67)$$

```

> assume(x, real, y, real);
> z := 1 / (1 + (x + I*y)^2);
> F := 1/z;
> u := Re(F);
> u := evalc(u);
> v := Im(F);
> v := evalc(v);
> with(plots):
> p1 := implicitplot({seq(u=a, a=-5..5)}, x=-5..5, y=-5..5, numpoints=1000);
> p2 := implicitplot({seq(v=b, b=-5..5)}, x=-5..5, y=-5..5, numpoints=1000, color=green);
> display([p1, p2]);

```

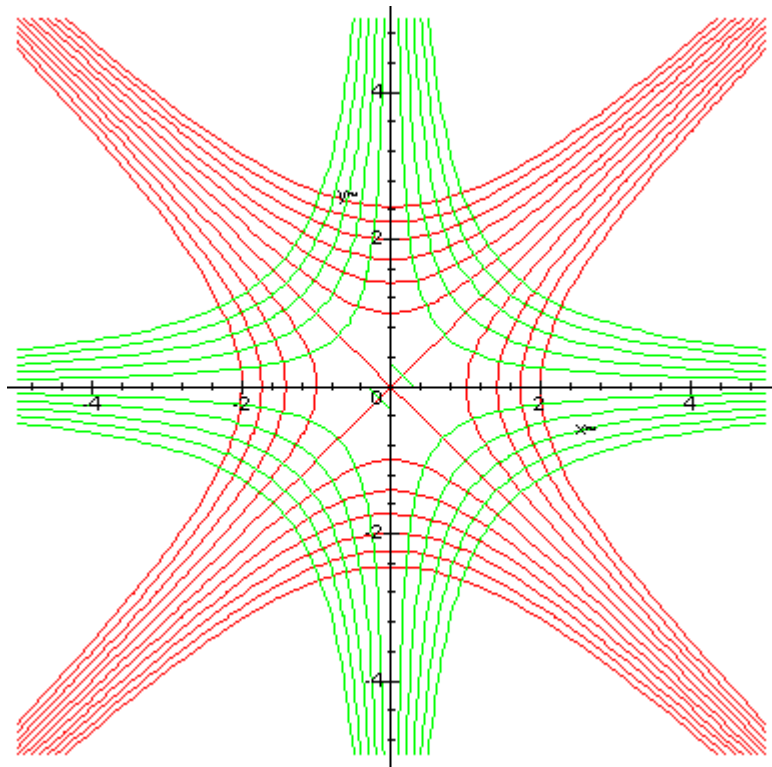


Figure: 17.10 - Représentation plane d'une fonction holomorphe importante

3. LOGARITHME COMPLEXE

Nous devons trouver pour toutes les fonctions construites dans $\mathbb{R}$ leur équivalent dans $\mathbb{C}$ tout en sachant que si nous réduisons le cas de $\mathbb{C}$ à $\mathbb{R}$ nous devons retomber sur nos pattes!

Pour cela, commençons par la fonction la plus classique et scolaire qui est donc le logarithme.

De la même manière que nous avons construit le logarithme comme étant par définition (!) la fonction réciproque de l'exponentielle naturelle e^x dans le chapitre d'Analyse Fonctionnelle, nous partons d'abord de:

$$f(z) = e^z \quad (17.68)$$

où z est donc un nombre complexe et nous allons définir le logarithme complexe qui doit se réduire au logarithme naturel si z n'a pas de partie imaginaire!

Donc par définition le logarithme complexe sera:

$$w = e^z \Leftrightarrow z = \text{Log}(w) \quad (17.69)$$

et sur l'ensemble de ce site, le logarithme complexe sera différencié du logarithme réel par un *L* majuscule!

Écrivons z et w sous la forme d'Euler vue dans le chapitre sur les Nombres:

$$\begin{aligned} w = x + iy = r(\cos \theta + i \sin \theta) &= |w| \cdot e^{i \arg(w)} = |w| (\cos(\arg(w)) + i \sin(\arg(w))) \\ z = u + iv \end{aligned} \quad (17.70)$$

Nous avons alors:

$$w = e^z \Rightarrow r(\cos \theta + i \sin \theta) = e^{u+iv} = e^u e^{iv} = e^u (\cos v + i \sin v) \quad (17.71)$$

Par correspondance, nous trouvons immédiatement:

$$r = |w| = e^u \Rightarrow u = \ln(r) \quad \text{et} \quad v = \arg(w) = \theta + 2\pi k \quad (17.72)$$

avec $k \in \mathbb{Z}$. Il vient alors:

$$z = u + iv = \ln(r) + i(\theta + 2\pi k) = \text{Log}(w) \quad (17.73)$$

Donc:

$$z = \text{Log}(w) = \ln(r) + i(\theta + 2\pi k) \quad (17.74)$$

ou autrement écrit:

$$\text{Log}(w) = \ln |w| + i \arg(w) \quad (17.75)$$

Donc si w n'a pas de partie imaginaire, nous retombons bien sur nos pattes puisque $\arg(w)$ devient nul.

Une grosse différence est mise en avant donc entre le logarithme des nombres complexes et réels: ces premiers peuvent prendre plusieurs valeurs à cause de l'argument.

Nous vérifions bien par ailleurs maintenant que:

$$e^{\text{Log}(w)} = e^{\ln |w| + i \arg(w)} = e^{\ln |w|} e^{i \arg(w)} = |w| e^{i \arg(w)} = w \quad (17.76)$$

4. INTÉGRATION DE FONCTIONS COMPLEXES

Nous venons de voir précédemment comment vérifier si une fonction complexe $f(z)$ était dérivable (elle doit au moins respecter les équations de Cauchy-Riemann) en tout point.

Maintenant voyons le cas contraire qui est... l'intégration!

Nous avons bien évidemment en reprenant les notations vues dans le chapitre de Calcul Différentiel Et Intégral:

$$F(z) = \int f(z) dz \quad (17.77)$$

soit sous forme explicite:

$$\begin{aligned} F(z) &= \int f(x+iy) d(x+iy) = \int (u+iv)(dx+idy) \\ &= \left[\int (u dx - v dy) \right] + i \left[\int (v dx + u dy) \right] \\ &= \int u dx - \int v dy + i \int v dx + i \int u dy \end{aligned} \quad (17.78)$$

Bon cette expression établie, donnons une petite explication quant à sa lecture:

1. Nous savons que u et v dépendent tous les deux dans le cas général de x et y .
2. Nous savons que u et v représentent (voir exemples au début du chapitre) des courbes fermées ou ouvertes ainsi que des droites lorsque x (ou respectivement y) est fixé et que l'autre variable associée, elle varie!

Donc chacun des termes comportant une intégrale dans l'expression écrite ci-dessus est une intégrale curviligne sur une famille de courbes ouvertes ou fermées (dont un cas particulier est des droites...)!

Cette intégrale peut être évaluée en utilisant le théorème de Green dans le plan (cf. chapitre de Calcul Vectoriel) si nous considérons le cas particulier d'un chemin curviligne fermé tel que:

$$F(z) = \oint (u dx - v dy) + i \oint (v dx + u dy) \quad (17.79)$$

Abordons d'abord la partie réelle:

Nous avons effectivement démontré (il est très fortement conseillé de relire ce théorème de Green) dans le chapitre de Calcul Vectoriel que:

$$\oint (f_x dx + f_y dy) = \int_S \left(\frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y} \right) dS_z = \int_S \left(\frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y} \right) dx dy \quad (17.80)$$

Ce qui s'écrit dans notre situation:

$$\oint (u dx + (-v) dy) = \int_S \left(\frac{\partial(-v)}{\partial x} - \frac{\partial(u)}{\partial y} \right) dx dy = - \int_S \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \right) dx dy \quad (17.81)$$

Or, si la fonction est holomorphe et satisfait donc aux équations de Cauchy-Riemann nous avons immédiatement:

$$\oint (u dx + (-v) dy) = - \underbrace{\int_S \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \right) dx dy}_{=0} = 0 \quad (17.82)$$

Ainsi notre intégrale se réduit dans le cas particulier d'un chemin fermé à:

$$F(z) = i \oint (v dx + u dy) \quad (17.83)$$

Et... réutilisons le théorème de Green:

$$\oint (v dx + u dy) = \int_S \left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} \right) dx dy \quad (17.84)$$

Or, si la fonction est holomorphe (donc pour rappel dérivable en tout point du plan complexe ou d'un sous-ensemble ouvert de celui-ci) et satisfait donc aux équations de Cauchy-Riemann nous avons immédiatement:

$$\int_S \underbrace{\left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} \right)}_{=0} dx dy = 0 \quad (17.85)$$

et nous obtenons ainsi le "**théorème de Cauchy**", ou "**théorème de Cauchy-Goursat**" dans sa version généralisée aux fonctions non continues, qui dit que si une fonction est holomorphe (satisfaisant donc les équations de Cauchy-Riemann) et intégrée sur un contour fermé alors:

$$\boxed{\oint f(z) dz = \oint (u + iv) dz = 0} \quad (17.86)$$

Comme corollaire (sans démonstration), toute fonction qui satisfait à la relation précédente est holomorphe (dans tout le plan complexe ou un sous-ensemble ouvert de celui-ci).

Ce résultat permet dans certains domaines comme la physique quantique des champs (on pense au potentiel de Yukawa qui n'est pas traité pour l'instant sur ce site) de calculer des intégrales définies réelles compliquées en utilisant la propriété ci-dessus. L'idée étant lors du choix du contour fermé de l'intégrale curviligne de s'arranger à faire apparaître l'intégrale définie réelle recherchée comme étant une partie seulement du chemin (en la généralisant au cas complexe) et de par l'égalité avec zéro en déduire sa valeur grâce aux autres parties de l'intégrale du chemin choisi (parties qui sont évidemment simples à calculer).

En d'autres termes, il s'agit de la calculer par différence. Toute la difficulté résidant dans la pratique à trouver la fonction $f(z)$ et le contour fermé permettant de faire apparaître la fonction $f(x)$ de l'intégrale définie recherchée...

À l'aide de ce résultat, faisons un exemple scolaire important qui nous sera utile par la suite (mais qui n'a aucun rapport avec le cas du calcul d'une intégrale définie réelle).

Calculons:

$$\oint \frac{dz}{z} \quad (17.87)$$

Pour cela, nous allons utiliser la simplification qui consiste à se rappeler (cf. chapitre Nombres) que:

$$z = r e^{i\theta} \quad (17.88)$$

Donc:

$$dz = r i e^{i\theta} d\theta \quad (17.89)$$

Nous pouvons alors écrire l'intégrale curviligne comme:

$$\oint \frac{dz}{z} = \int \frac{r i e^{i\theta} d\theta}{r e^{i\theta}} = i \int d\theta \quad (17.90)$$

Or comme sur un chemin fermé dérivable en tout point (donc sans sommets) l'angle à parcourir pour faire un tour complet ira nécessairement de 0 à 2π . Il vient alors:

$$\oint \frac{dz}{z} = i \int_0^{2\pi} d\theta = 2\pi i \quad (17.91)$$

Avant de continuer remarquons un fait intéressant et important: Une intégrale (on ne parle pas de la primitive mais de l'intégrale!) d'une fonction du type $1/x$ dans $\mathbb{R}$ ne serait pas calculable. Or si nous généralisons le concept à $\mathbb{C}$, nous voyons que nous contourons... (le jeu de mots...) la singularité via une intégrale curviligne qui entoure la singularité. Et... et... dans notre calcul précédent z pourrait tout à fait n'avoir que la valeur réelle et pas l'imaginaire. Donc l'intégrale de $1/x$ devient alors calculable et a un résultat dans les complexes ce qui est remarquable!

Certains mathématiciens interprètent cela en figurant que $1/x$ est une projection plane d'un espace tridimensionnel dont l'axe imaginaire est perpendiculaire au plan $\mathbb{R}^2$. D'où le fait que $1/x$ soit intégrable dans $\mathbb{C}$ mais bon c'est une interprétation...

Enfin, indiquons que $1/z$ est holomorphe sur tout le plan complexe excepté en 0 (la dérivée étant la même que pour $1/x$). Elle n'est donc pas $\mathbb{C}$ -différentiable!

Ceci étant fait, faisons un cas important et similaire avec l'intégrale curviligne suivante:

$$\oint \frac{dz}{z - z_0} \quad (17.92)$$

où z_0 est un nombre complexe constant. Posons:

$$z - z_0 = \rho e^{i\varphi} \Rightarrow \frac{dz}{d\varphi} = \rho i e^{i\varphi} \Rightarrow dz = \rho i e^{i\varphi} d\varphi \quad (17.93)$$

Nous pouvons alors écrire si nous faisons qu'un seul tour dans le sens inverse des aiguilles (sens antihorlogique ou trigonométrique) d'une montre:

$$\oint \frac{dz}{z - z_0} = \int \frac{\rho i e^{i\varphi} d\varphi}{\rho e^{i\varphi}} = i \int d\varphi = 2\pi i \quad (17.94)$$

qui n'est valable à nouveau que si notre chemin d'intégration évite z_0 sinon quoi il y a une singularité. Cette dernière intégrale est donc une petite généralisation simpliste de la précédente.

Maintenant montrons le théorème important qui nous intéresse au fait depuis le début de ce chapitre en utilisant les nombreux résultats démontrés jusqu'ici!

Nous savons que si une fonction $f(z)$ satisfait aux équations de Cauchy-Riemann, alors si nous évitons soigneusement la valeur z_0 (comme dans les calculs précédents), l'expression:

$$\frac{f(z) - f(z_0)}{z - z_0} \quad (17.95)$$

est aussi dérivable en tout point excepté en z_0 (donc l'expression n'y est plus holomorphe) qui est appelé une "**singularité**".

Effectivement, prendre une fonction holomorphe $f(z)$ satisfaisant Cauchy-Riemann et en soustraire une constante ($f(z_0)$) ne change en rien le fait que l'expression (en l'occurrence le numérateur dans la relation précédente) restera holomorphe. Enfin, multiplier celle-ci par une fraction (dénominateur de la relation précédente) qui est elle aussi holomorphe donne une fonction holomorphe. Mais des singularités peuvent alors apparaître, nous parlons alors de "**fonctions méromorphes**" (il s'agit du rapport de deux fonctions holomorphes).

Remarque: Une fonction méromorphe est une fonction holomorphe dans tout le plan complexe, sauf éventuellement sur un ensemble de points isolés dont chacun est un pôle pour la fonction. (Voir plus loin pour la notion de pôle)

Dès lors, si nous en prenons l'intégrale curviligne sur un chemin fermé évitant de passer par z_0 , le théorème de Cauchy nous donne immédiatement (voir la démonstration plus haut):

$$\oint \frac{f(z) - f(z_0)}{z - z_0} dz = 0 \quad (17.96)$$

Or, ceci s'écrit aussi après réarrangement des termes:

$$\oint \frac{f(z)}{z - z_0} dz = f(z_0) \oint \frac{1}{z - z_0} dz \quad (17.97)$$

Soit:

$$f(z_0) = \frac{1}{\oint \frac{1}{z - z_0} dz} \oint \frac{f(z)}{z - z_0} dz \quad (17.98)$$

Or, nous avons démontré plus haut que:

$$\oint \frac{dz}{z - z_0} = 2\pi i \quad (17.99)$$

Il vient alors le résultat appelé "**théorème intégral de Cauchy**", ou plus rarement "**formule de Cauchy**", (dont il existe une forme généralisée que nous démontrerons plus bas):

$$\boxed{f(z_0) = \frac{1}{2\pi i} \oint \frac{f(z)}{z - z_0} dz} \quad (17.100)$$

Au fait, dans la pratique toute la subtilité est de pouvoir ramener une fonction $g(z)$ holomorphe (qui satisfait donc les équations de Cauchy-Riemann) en la manipulant à une forme du type:

$$g(z) = \frac{f(z)}{z - z_0} \quad (17.101)$$

quand c'est possible... alors le calcul de son intégrale curviligne (de chemin fermé) devient extrêmement simple puisqu'elle sera égale à:

$$2\pi i f(z_0) \quad (17.102)$$

de par le théorème intégral de Cauchy!

Remarques:

R1. Nous savons donc calculer la valeur d'une intégrale curviligne d'une expression non holomorphe mais dont le numérateur lui l'est.

R2. Attention! Le signe de la valeur d'une intégrale curviligne va dépendre du sens dans lequel on parcourt son chemin d'intégration. Si le sens est direct (c'est-à-dire "**antihorlogique**" ou encore "**trigonométrique**") son signe sera positif; si au contraire le sens est horlogique son signe sera négatif. Vous vous dites sans doute que cette précision est sans intérêt vu que cette valeur est généralement nulle. Certes, mais nous verrons plus loin l'importance de cette spécification lorsqu'il sera question du calcul des résidus.

Il y a une relation équivalente pour la dérivée $f'(z_0)$ à celle donnée par le théorème intégral de Cauchy. Voyons cela:

$$\begin{aligned} f'(z_0) &= \lim_{h \rightarrow 0} \frac{f(z_0 + h) - f(z_0)}{h} \\ &= \lim_{h \rightarrow 0} \left(\frac{f(z_0 + h)}{h} - \frac{f(z_0)}{h} \right) = \lim_{h \rightarrow 0} \frac{1}{h} (f(z_0 + h) - f(z_0)) \\ &= \lim_{h \rightarrow 0} \frac{1}{h} \left(\frac{1}{2\pi i} \oint \frac{f(z)}{z - (z_0 + h)} dz - \frac{1}{2\pi i} \oint \frac{f(z)}{z - z_0} dz \right) \\ &= \lim_{h \rightarrow 0} \frac{1}{h} \left(\frac{1}{2\pi i} \oint \frac{f(z)}{z - z_0 - h} dz - \frac{1}{2\pi i} \oint \frac{f(z)}{z - z_0} dz \right) = \lim_{h \rightarrow 0} \frac{1}{2\pi i h} \left(\oint \frac{f(z) dz}{z - z_0 - h} - \oint \frac{f(z) dz}{z - z_0} \right) \quad (17.103) \\ &= \lim_{h \rightarrow 0} \frac{1}{2\pi i h} \left(\oint \frac{f(z) dz}{z - z_0 - h} \frac{(z - z_0)}{(z - z_0)} - \oint \frac{f(z) dz}{z - z_0} \frac{(z - z_0 - h)}{(z - z_0 - h)} \right) \\ &= \lim_{h \rightarrow 0} \frac{1}{2\pi i h} \oint \frac{f(z) [(z - z_0) - (z - z_0 - h)] dz}{(z - z_0 - h)(z - z_0)} = \lim_{h \rightarrow 0} \frac{1}{2\pi i h} \oint \frac{h f(z) dz}{(z - z_0 - h)(z - z_0)} \end{aligned}$$

Donc:

$$f'(z_0) = \frac{1}{2\pi i} \oint \frac{f(z)dz}{(z-z_0)^2} \quad (17.104)$$

en continuant ainsi, nous avons:

$$\begin{aligned} f''(z_0) &= \frac{2}{2\pi i} \oint \frac{f(z)dz}{(z-z_0)^3} \\ f'''(z_0) &= \frac{6}{2\pi i} \oint \frac{f(z)dz}{(z-z_0)^4} \end{aligned} \quad (17.105)$$

Bref, nous remarquons donc que:

$$f^{(n)}(z_0) = \frac{d^n f(z_0)}{dz^n} = \frac{n!}{2\pi i} \oint \frac{f(z)}{(z-z_0)^{n+1}} dz \quad (17.106)$$

qui n'est autre que le "**théorème intégral de Cauchy généralisé**".

Ce résultat est très puissant car il montre que les fonctions holomorphes sont infiniment dérivables (à cause du dénominateur), soit analytiques, et il est beaucoup plus difficile de trouver un théorème équivalent avec des conditions aussi simples pour les fonctions réelles.

Si nous revenons maintenant à notre développement de Taylor d'une fonction complexe:

$$f(z) = \sum_{n=0}^{+\infty} \frac{1}{n!} \frac{d^n f(z_0)}{dz^n} (z-z_0)^n \quad (17.107)$$

hummm... et que voyons-nous ici? Eh bien ceci!:

$$\begin{aligned} f(z) &= \sum_{n=0}^{+\infty} \frac{1}{n!} \frac{d^n f(z_0)}{dz^n} (z-z_0)^n = \sum_{n=0}^{+\infty} \frac{1}{n!} \frac{n!}{2\pi i} \oint \frac{f(z)dz}{(z-z_0)^{n+1}} (z-z_0)^n \\ &= \sum_{n=0}^{+\infty} \frac{1}{2\pi i} \oint \frac{f(z)}{(z-z_0)^{n+1}} dz (z-z_0)^n = \sum_{n=0}^{+\infty} c_n (z-z_0)^n \end{aligned} \quad (17.108)$$

Il en découle la relation suivante appelée "**série de Laurent à puissances positives**" (il en existe une version plus généralisée que nous allons démontrer plus loin):

$$f(z) = \sum_{n=0}^{+\infty} c_n (z-z_0)^n \quad (17.109)$$

qui donne donc l'expression formelle d'une fonction complexe sous forme de série infinie de puissances entières à proximité d'un point z_0 du plan complexe avec donc:

$$c_n = \frac{1}{2\pi i} \oint \frac{f(z)}{(z-z_0)^{n+1}} dz = \frac{f^{(n)}(z_0)}{n!} \quad (17.110)$$

En se rappelant que $d^n f(z_0)/dz^n$ s'écrit de manière équivalente $f^{(n)}(z_0)$, nous constatons que l'ensemble des deux relations précédentes nous redonne le développement en série de Taylor que nous avons obtenu en analyse réelle (cf. chapitre de Suites et Séries) et qui était:

$$f(z) = \sum_{n=0}^{+\infty} \frac{1}{n!} \frac{d^n f(z_0)}{dz^n} (z - z_0)^n \quad (17.111)$$

Ainsi, les séries de Taylor ne sont qu'un cas particulier des séries de Laurent.

Ceci est assez remarquable comme résultat car cela montre aussi que nous pouvons utiliser l'intégrale curviligne sur le plan complexe pour calculer les coefficients c_n de la série de Laurent au lieu de calculer les dérivées d'ordre n de la fonction f si ces dernières s'avéraient trop compliquées à déterminer. Ou inversement... calculer une simple dérivation au lieu de calculer une intégrale curviligne casse-tête (typiquement le cas en physique) en utilisant le fait que:

$$\oint \frac{f(z)}{(z - z_0)^{n+1}} dz = 2\pi i \frac{f^{(n)}(z_0)}{n!} \quad (17.112)$$

Le seul point malheureux étant que cette dernière relation n'est calculable que si nous arrivons à mettre la fonction dans l'intégrale curviligne sous la forme:

$$\frac{f(z)}{(z - z_0)^{n+1}} \quad (17.113)$$

où n est un entier positif ou nul. Ceci est franchement loin d'être aisé dans la grande majorité des cas! L'idée serait alors de trouver un chemin général pour l'intégrale curviligne, valable pour toute fonction $f(z)$ tel que ce dénominateur (qui contient en plus une singularité en z_0) disparaisse. Ce serait l'idéal... mais il nous faut une piste... et celle-ci va venir de l'étude de la convergence des séries de puissances complexes. Voyons de quoi il s'agit avec une approche qualitative!

4.1. CONVERGENCE D'UNE SÉRIE

Nous avons vu dans le chapitre de Suites et Séries que nombre de fonctions réelles pouvaient être exprimées en série de Maclaurin (cas particulier des séries de Taylor en $x_0 = 0$) sous la forme:

$$f(x) = \sum_{k=0}^{\infty} a_k x^k \quad (17.114)$$

Nous y avons également montré, uniquement par l'exemple, que ce développement en série de puissances infinie n'était valable pour certaines fonctions réelles que dans un certain domaine de définition appelé "rayon de convergence".

Même si ce rayon de convergence peut être déterminé plus ou moins facilement au cas par cas, il y a certains exemples déroutants qui ne pouvaient pas au début du 19ème siècle être compris sans l'analyse complexe.

Voyons un exemple simple pour comprendre de quel type de problème il s'agit. Considérons pour cela les deux fonctions:

$$g(x) = \frac{1}{1-x^2} \quad \text{et} \quad h(x) = \frac{1}{1+x^2} \quad (17.115)$$

et avant de continuer notre exemple, rappelons que nous avons démontré dans le chapitre de Suites et Séries la relation:

$$S_n = u_0 + u_1 + u_2 + \dots + u_n = \frac{u_0 - u_{n+1}}{1-q} = u_0 \frac{1-q^{n+1}}{1-q} \quad (17.116)$$

relative à une série géométrique, c'est-à-dire une série dont les termes sont du type:

$$u_n = u_0 q^n \quad (17.117)$$

Il vient dès lors immédiatement si $n \rightarrow \infty$ et $q \in]-1, +1[$:

$$S_\infty = u_0 + u_1 + u_2 + \dots + u_\infty = \sum_{k=0}^{\infty} u_k = u_0 \frac{1-q^{\infty+1}}{1-q} = u_0 \frac{1-0}{1-q} = \frac{u_0}{1-q} \quad (17.118)$$

Si $u_0 = 1$, nous avons:

$$\sum_{k=0}^{\infty} u_k = \sum_{k=0}^{\infty} u_0 q^k = \sum_{k=0}^{\infty} q^k = \frac{1}{1-q} \quad (17.119)$$

Donc si nous changeons la notation, nous avons:

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n \quad (17.120)$$

Il vient alors immédiatement:

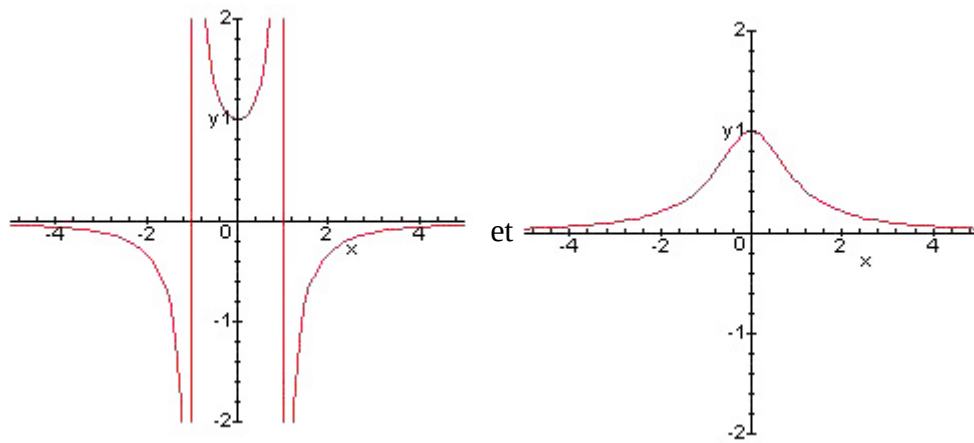
$$g(x) = \sum_{n=0}^{\infty} (x^2)^n = \sum_{n=0}^{\infty} x^{2n} \quad \text{et} \quad h(x) = \sum_{n=0}^{\infty} (-x^2)^n = \sum_{n=0}^{\infty} (-1)^n x^{2n} \quad (17.121)$$

Donc les deux fonctions $g(x)$ et $h(x)$ précédentes sont définies pour un développement en série infinie de puissances uniquement dans un rayon de convergence $x \in]-1, +1[$.

Nous obtiendrions donc le même résultat en faisant un développement en série de Maclaurin!

Nous voyons trivialement qu'il y a pour $g(x)$ deux singularités qui sont $x = \{-1, +1\}$ par contre, basiquement nous n'en voyons pas trivialement pour $h(x)$ si nous raisonnons uniquement dans $\mathbb{R}$ donc il peut être difficile pour cette dernière fonction de comprendre l'origine du rayon de convergence.

Effectivement, si nous traçons ces deux fonctions dans $\mathbb{R}$ avec Maple 4.00b nous obtenons respectivement:

Figure: 17.11 - Représentation dans Maple 4.00b des fonctions g et h

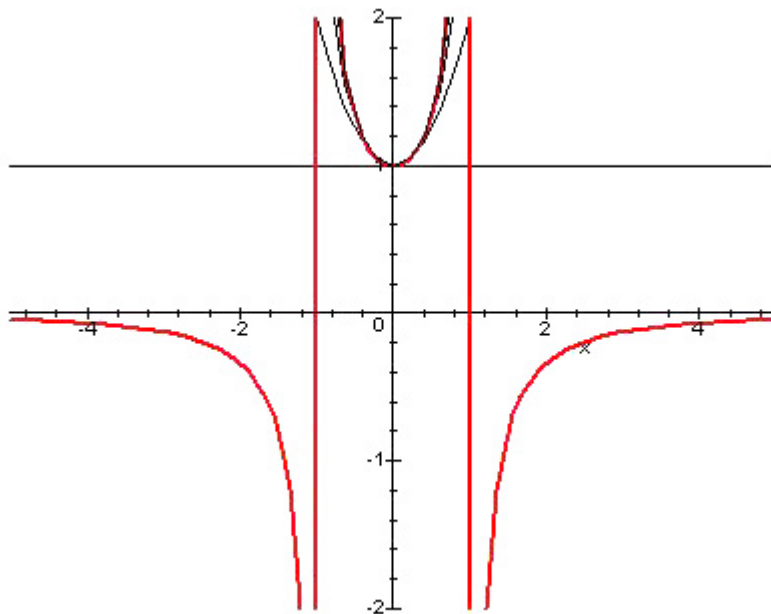
d'où le problème de savoir pourquoi il y a quand même implicitement un rayon de convergence $x \in]-1, +1[$ pour $h(x)$???

Une manière encore plus flagrante de mettre en évidence le problème, c'est de montrer l'approche de ces deux fonctions par un développement en série de Maclaurin avec dix termes:

Pour $g(x)$ nous avons:

with(plots):

```
> xplot:= plot(1/(1-x^2),x=-5..5,thickness=2,color=red);
> tays:= plots[display](xplot);
> for i from 1 by 2 to 10 do
>   tpl:= convert(taylor(1/(1-x^2), x=0,i),polynom);
>   tays:= tays,plots[display]([xplot,plot(tpl,x=-5..5,y=-2..2,
>   color=black,title=convert(tpl,string))]) od;
> plots[display]([tays],view=[-5..5,-2..2]);
```

Figure: 17.12 - Représentation plane de la fonction g pour visualiser le problème

où nous voyons bien que la série de Maclaurin (ou l'expression en série de puissances) ne converge pas en dehors de $x \in]-1, +1[$ ce qui peut être intuitif à cause des deux singularités.

Pour $h(x)$ nous avons par contre:

```
>with(plots):
> xplot:= plot(1/(1+x^2),x=-5..5,thickness=2,color=red):
> tays:= plots[display](xplot):
> for i from 1 by 2 to 10 do
>   tpl:= convert(taylor(1/(1+x^2), x=0,i),polynom):
>   tays:= tays,plots[display]([xplot,plot(tpl,x=-5..5,y=-2..2,
>   color=black,title=convert(tpl,string))]) od:
> plots[display]([tays],view=[-5..5,-2..2]);
```

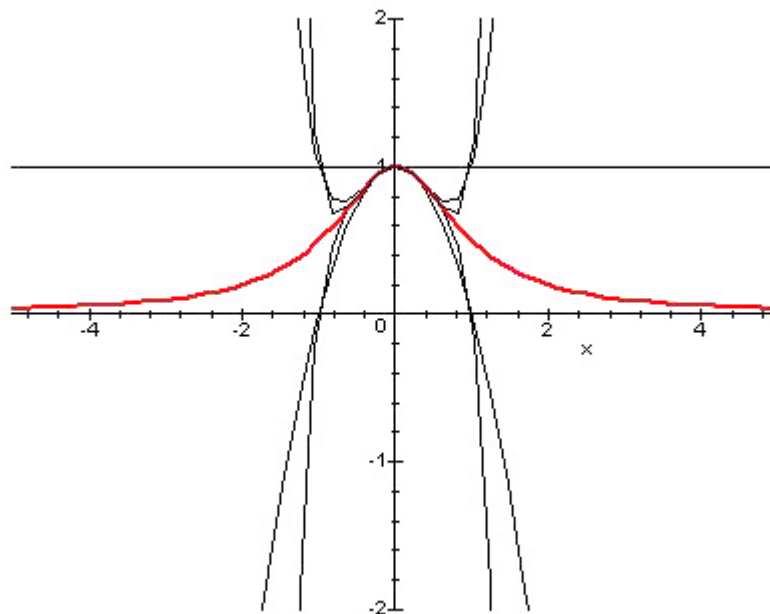


Figure: 17.13 - Étonnamment, ici la série de Maclaurin ne converge pas

où nous voyons bien que la série de Maclaurin (ou l'expression en série de puissances) ne converge pas non plus en dehors de $x \in]-1, +1[$ ce qui était déstabilisant et contre-intuitif au début de l'histoire de l'analyse réelle.

Aujourd'hui même un élève du secondaire sait qu'il est possible de raisonner aussi dans $\mathbb{C}$ et que $\mathbb{R} \subset \mathbb{C}$. Donc l'analyse réelle n'est qu'un cas particulier et restreint de l'analyse complexe.

La singularité pour $h(x)$ dans $\mathbb{C}$ vient du fait que celle-ci s'écrit :

$$h(z) = \frac{1}{1+z^2} \quad (17.122)$$

et qu'il y a donc deux singularités pour $z = \{-i, +i\}$ ce que nous voyons bien si nous représentons:

$$|h(z)| = \left| \frac{1}{1+z^2} \right| \quad (17.123)$$

avec Maple 4.00b (heureusement que nous avons maintenant l'équivalent d'un microscope dans la mathématique avec Maple...):

```
>plot3d(abs(1/(1+(re+I*im)^2)),re=-3..3,im=-3..3,view=[-2..2,-2..2,-2..2],orientation=[-130,70],contours=50,style=PATCHCONTOUR,axes=frame,grid=[100,100],numpoints=10000);
```

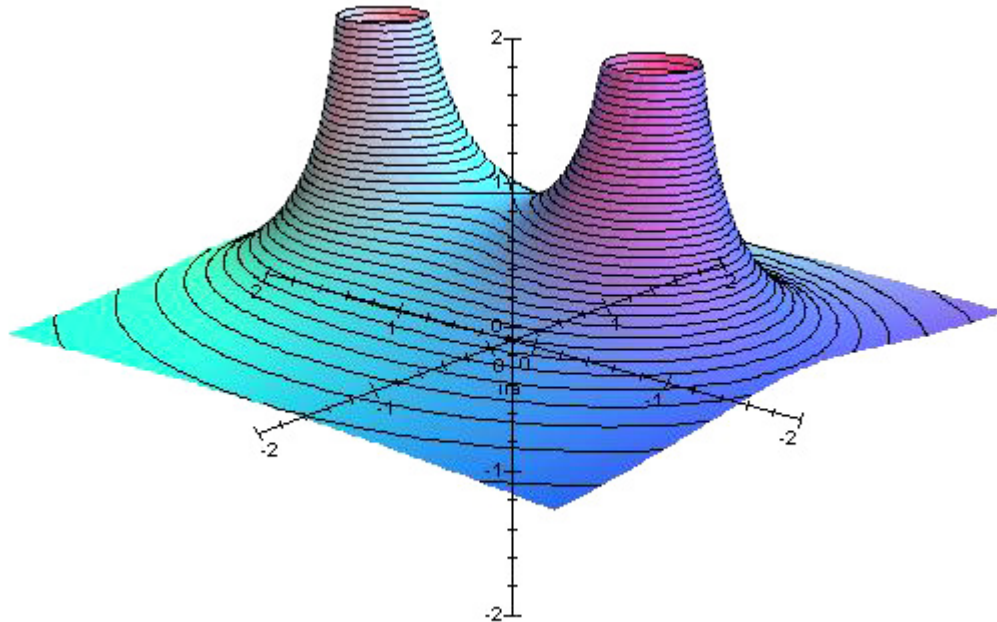


Figure: 17.14 - Représentation complexe de la fonction h pour mettre en évidence la raison de la divergence

où nous discernons les deux singularités sur l'axe imaginaire et la fonction $h(x)$ sur l'axe réel (entre les deux pics). Donc lorsque nous développons une fonction en série, nous concluons que son rayon de convergence est défini par tout le plan complexe et non par l'axe traditionnel de l'analyse réelle.

Il est ainsi plus naturel de comprendre pourquoi nous parlions dans le chapitre de Suites Et Séries de "rayon" car vu du dessus, nous avons dans le plan complexe:

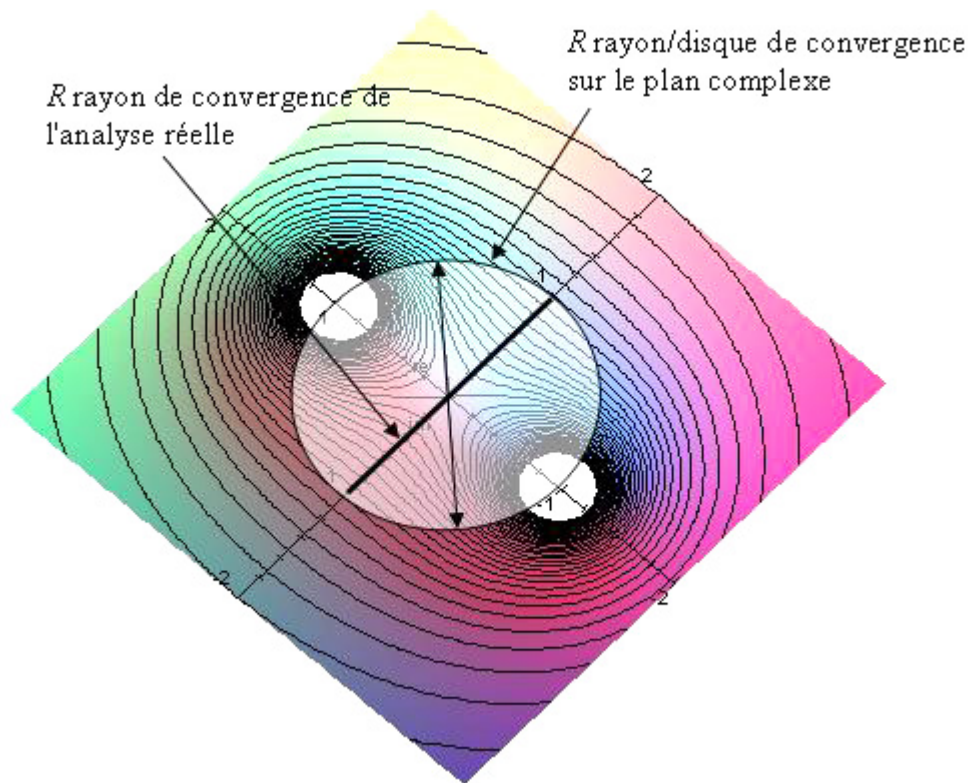


Figure: 17.15 - Représentation des différents rayons de convergence

d'où le fait que nous parlons tantôt de disque de convergence (ouvert) et tantôt de rayon de convergence (ouvert). Par ailleurs, nous remarquons sur le graphique que le domaine de convergence est connexe (tout couple de points du domaine de convergence peut être relié par une droite qui est dans le domaine de convergence).

Remarque: Rappelons qu'un sous-ensemble, intervalle ou disque "ouvert" signifie que nous n'en prenons pas les bords.

Nous comprenons alors mieux pourquoi la série de Taylor ne convergeait pas trivialement pour $h(x)$: elle doit converger sur tout le disque de convergence du plan complexe et pas seulement converger sur l'axe réel!

De tout ceci, nous déduisons que notre série de Laurent à puissances positives démontrée plus haut:

$$f(z) = \sum_{n=0}^{+\infty} c_n (z - z_0)^n \quad (17.124)$$

ne converge pas forcément, sans surprise..., sur tout le plan complexe (au même titre que les séries de Taylor sur la droite réelle puisqu'il s'agit de l'équivalent!) mais parfois uniquement dans un sous-domaine (connexe?) ouvert de ce plan autour de z_0 (qui dans l'exemple particulier pris ici valait donc: 0).

Avec notre fonction $h(x)$ exprimée en utilisant un développement de Maclaurin sur 5 termes, nous voyons immédiatement avec Maple 4.00b que sur les bords du carré inscrit au disque de convergence, la série ne converge plus et nous y devinons le début des deux singularités:

```
>plot3d(abs(1-(re+I*im)^2+(re+I*im)^4-(re+I*im)^6+(re+I*im)^8),re=-0.7..0.7,im=-0.7..0.7,view=[-1.5..1.5,-1.5..1.5,0..1.5],orientation=[-130,70],contours=50,style=PATCHCONTOUR,axes=frame,grid=[100,100],numpoints=10000);
```

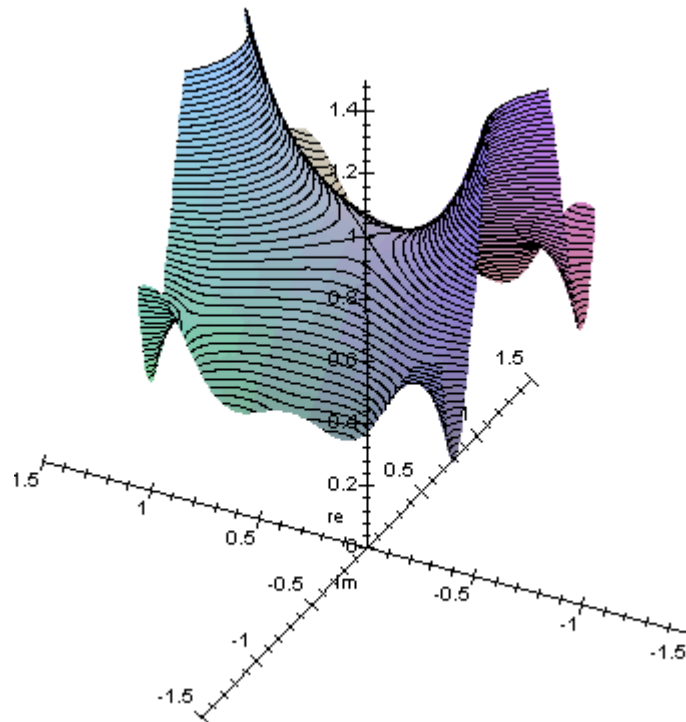


Figure: 17.16 - Zoom sur la représentation complexe afin de comprendre la raison de la divergence

un peu en dehors du disque de convergence, nous avons évidemment un peu n'importe quoi:

```
>plot3d(abs(1-(re+I*im)^2+(re+I*im)^4-(re+I*im)^6+(re+I*im)^8),re=-3..3,im=-3..3,view=[-1.5..1.5,-1.5..1.5,0..1.5],orientation=[-130,70],contours=50,style=PATCHCONTOUR,axes=frame,grid=[100,100],numpoints=10000);
```

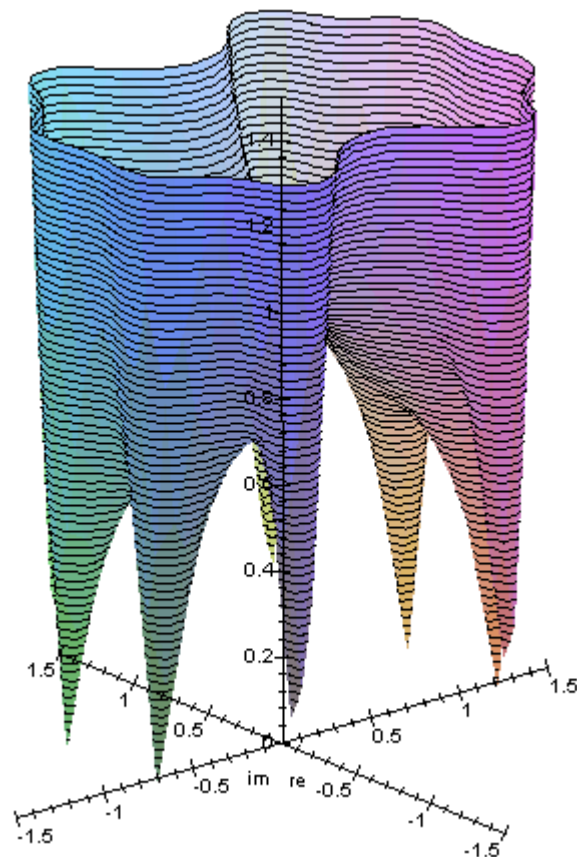


Figure: 17.17 - Cela diverge... (stalactites???)

Il y a quand même quelque chose d'intéressant à essayer... puisque nous sommes maintenant sur un plan, et non plus sur une droite, il nous est possible de faire le développement de Taylor autour d'une singularité z_0 en déformant le disque convexe en une couronne simplement connexe telle que présentée ci-dessous (la couronne étant aussi la géométrie simplement connexe la plus simple découlant de la déformation d'un disque):

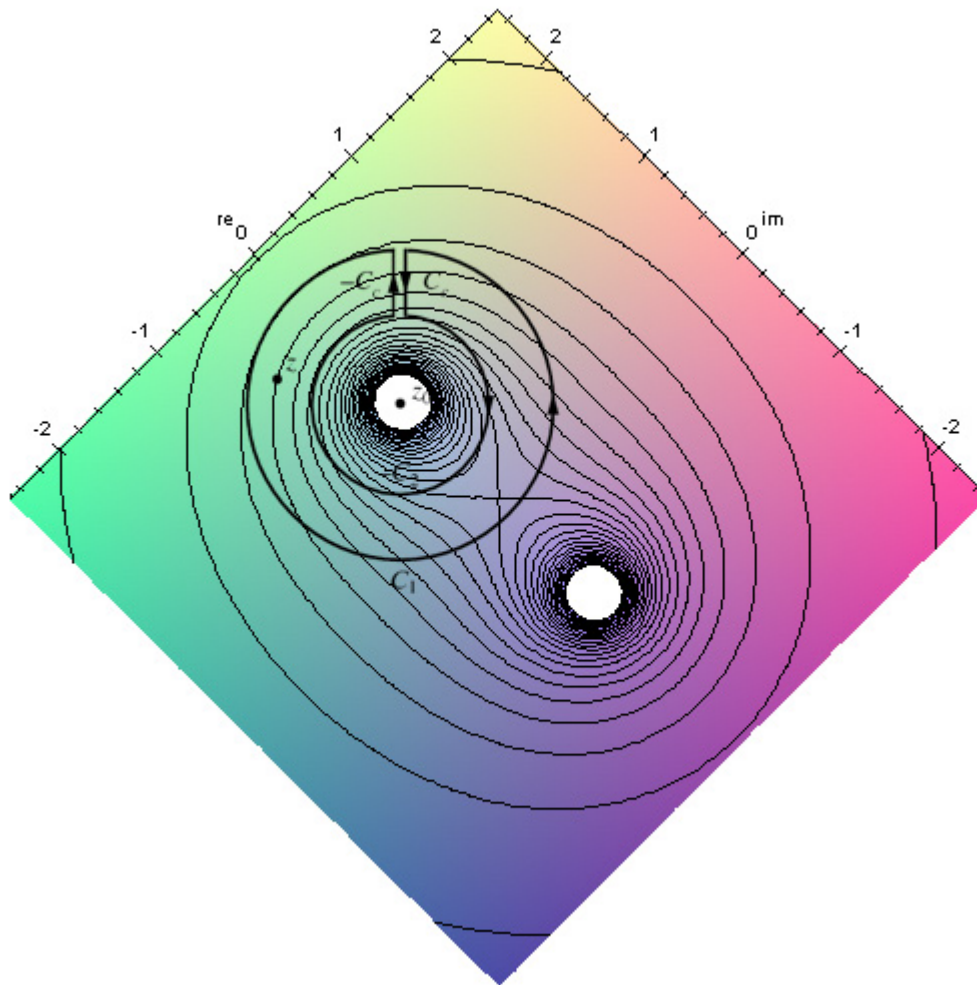


Figure: 17.18 - Représentation de la déformation d'un disque en une couronne

L'intérêt de ceci est de pouvoir déformer le domaine de convergence sur tout le plan complexe en évitant (contournant) toutes les singularités. Ainsi, contrairement aux séries de Taylor qui ne sont valables que sur un intervalle de l'axe des abscisses, nous aurions un nouveau type de série décrivant une fonction absolument partout, c'est-à-dire avant ET après (donc autour...) les singularités!

Donc évidemment nous allons imposer que dans la couronne déformée ci-dessus la fonction soit toujours holomorphe et analytique (comme dans le disque convexe initial). Avant de déterminer ce sur quoi nous allons tomber (série de Laurent généralisée), il nous faut d'abord faire une étude de la décomposition d'une intégrale en chemins:

5. DÉCOMPOSITIONS EN CHEMINS

Les intégrales curvilignes comme celles données précédemment peuvent aussi être écrites sous une autre forme assez classique et souvent utilisée dans la pratique.

Voyons cela. D'abord, nous venons de démontrer dans le cas particulier d'une fonction holomorphe que:

$$\oint f(z) dz = 0 \quad (17.125)$$

mais un chemin fermé peut être vu comme un chemin ayant un aller et un retour:

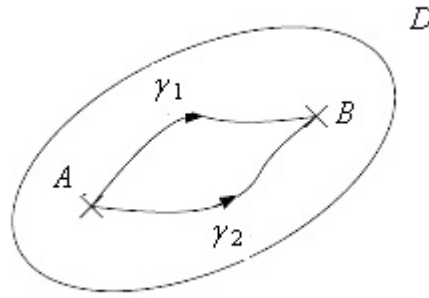


Figure: 17.19 - Représentation d'un chemin fermé avec son aller-retour

Nous avons alors:

$$\int_{\gamma_1} f(z) dz - \int_{\gamma_2} f(z) dz = 0 \Leftrightarrow \int_{\gamma_1} f(z) dz = \int_{\gamma_2} f(z) dz \quad (17.126)$$

Et maintenant vient ce qui nous intéresse... pour cela concentrons-nous sur une des intégrales curvilignes du type:

$$\int_{\gamma} f(z) dz \quad (17.127)$$

Nous savons (1ère forme d'expression) que tout nombre complexe z du type:

$$z = x + iy \quad (17.128)$$

peut être (2ème type d'expression) écrit sous la forme:

$$z(x, y) = x + iy = r(\cos \theta + i \sin \theta) = re^{i\theta} = z(r, \theta) \quad (17.129)$$

et pour intégrer sur un chemin, rien ne nous empêche d'en choisir un où r serait fixe (le module) et θ variable (nous n'aurions pas pu faire cela avec la première forme d'expression car en ne faisant varier que la partie imaginaire ou réelle, nous ne pouvons pas obtenir de courbe alors que cela est possible avec la forme d'Euler d'un nombre complexe)!

Nous avons alors:

$$z(\theta) = re^{i\theta} \quad (17.130)$$

Nous pouvons dès lors écrire:

$$\int_{\gamma} f(z) dz = \int_{\gamma} f(z(\theta)) dz(\theta) \quad (17.131)$$

et comme:

$$z'(\theta) = \frac{dz}{d\theta} \Rightarrow dz = z'(\theta) d\theta \quad (17.132)$$

d'où:

$$\int_{\gamma} f(z) dz = \int_{\gamma} f(z(\theta)) z'(\theta) d\theta \quad (17.133)$$

ce que nous retrouvons souvent sous la forme suivante dans la littérature:

$$\boxed{\int_{\gamma} f(z) dz = \int_{\gamma} f(z(t)) \dot{z}(t) dt} \quad (17.134)$$

Cette relation va nous être maintenant utile à démontrer un résultat nécessaire pour notre étude de la couronne.

5.1. CHEMIN INVERSE

Si C est une courbe allant d'un point P à un point Q , nous notons alors C^{-} la même courbe mais parcourue de Q à P .

Paramétrisons C^{-} :

Si $C(t)$ est la courbe définie sur $[a, b]$ nous définissons $C^{-}(t)$ la courbe définie sur $[a, b]$ par:

$$C^{-}(t) = C(a+b-t) \quad (17.135)$$

En effet, nous avons alors bien avec cette paramétrisation:

$$C^{-}(a) = C(a+b-a) = C(b) \text{ et } C^{-}(b) = C(a+b-b) = C(a) \quad (17.136)$$

et lorsque t croît de a à b , $a+b-t$ décroît de b à a . C^{-} n'est donc que C mais parcourue dans le sens inverse.

Nous avons alors en utilisant la dernière démonstration:

$$\int_{C^{-}} f(C^{-}) dC^{-} = \int_b^a f(C^{-}(t)) \cdot C^{-'}(t) dt = \int_b^a f(C(a+b-t)) \cdot C'(a+b-t) dt \quad (17.137)$$

Posons:

$$u = a+b-t \quad (17.138)$$

d'où:

$$du = -dt \quad (17.139)$$

Nous avons alors:

$$\begin{aligned}
 \int_{C^-} f(C^-) dC^- &= \int_b^a f(C(a+b-t)) \cdot C'(a+b-t) dt = \int_a^b f(C(u)) \cdot C'(u) (-du) \\
 &= - \int_C f(C(u)) \cdot C'(u) du = - \int_C f(C) dC
 \end{aligned}
 \tag{17.140}$$

Donc si C^- et C sont les chemins d'une même fonction mais parcourus dans le sens inverse, nous avons en reprenant notre notation conventionnelle (attention dans le deuxième terme il est implicite que la paramétrisation est différente du premier!):

$$\int_{\gamma} f(z) dz = - \int_{\gamma^-} f(z) dz \tag{17.141}$$

Soit:

$$\boxed{\int_{\gamma} f(z) dz + \int_{\gamma^-} f(z) dz = 0} \tag{17.142}$$

Raison pour laquelle nous disons souvent que le signe de la valeur d'une intégrale curviligne va dépendre du sens dans lequel on parcourt son chemin d'intégration. Si le sens est direct (c'est-à-dire "**antihorlogique**" ou encore "**trigonométrique**") son signe sera positif; si au contraire le sens est horlogique son signe sera négatif (cf. chapitre de Calcul Différentiel Et Intégral).

6. SÉRIES DE LAURENT

Cette dernière relation obtenue, nous pouvons revenir à notre déformation du disque de convergence en une couronne. Nous rappelons que l'idée étant initialement d'avoir l'expression analytique d'une fonction sous forme d'une série de puissances infinie dans un domaine restreint autour d'un point singulier et tout ceci... afin de pouvoir calculer pour les physiciens des intégrales curvilignes complexes en passant par une méthode utilisant les propriétés des séries complexes!

Commençons donc par le point (2), qui nous mènera plus facilement au point (1), en faisant un zoom sur notre couronne:

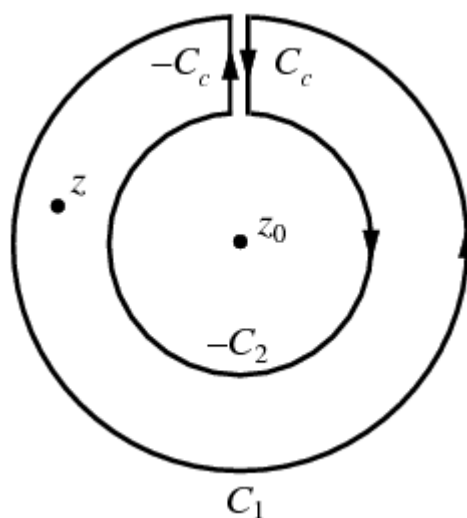


Figure: 17.20 - Zoom sur la couronne du début

Nous avons donc si la fonction f est analytique et holomorphe dans la couronne de rayon extérieur R et rayon intérieur r , l'intégrale curviligne suivante dans toute la couronne:

$$f(z) = \frac{1}{2\pi i} \oint \frac{f(z')}{z' - z} dz' \quad (17.143)$$

où nous notons z le point où nous souhaitons connaître la fonction et z' la variable dont dépend f . Ce changement de notation se justifiera par la suite pour une raison purement pratique.

La couronne peut donc être décomposée en 4 chemins:

$$f(z) = \frac{1}{2\pi i} \int_{C_1} \frac{f(z')}{z' - z} dz' + \frac{1}{2\pi i} \int_{C_c} \frac{f(z')}{z' - z} dz' + \frac{1}{2\pi i} \int_{-C_c} \frac{f(z')}{z' - z} dz' + \frac{1}{2\pi i} \int_{-C_2} \frac{f(z')}{z' - z} dz' \quad (17.144)$$

Si les deux segments C_c et $-C_c$ sont infiniment proches, ils correspondent alors à un même chemin parcouru une fois dans un sens positif et une fois dans le sens négatif. Or, nous avons démontré plus haut que:

$$\int_{\gamma} f(z) dz = - \int_{\gamma^-} f(z) dz \quad (17.145)$$

Il en découle donc que:

$$\int_{C_c} \frac{f(z')}{z' - z} dz' = - \int_{-C_c} \frac{f(z')}{z' - z} dz' \quad (17.146)$$

Ce qui nous amène à écrire:

$$\begin{aligned} f(z) &= \frac{1}{2\pi i} \int_{C_1} \frac{f(z')}{z' - z} dz' + \frac{1}{2\pi i} \int_{C_c} \frac{f(z')}{z' - z} dz' - \frac{1}{2\pi i} \int_{C_c} \frac{f(z')}{z' - z} dz' + \frac{1}{2\pi i} \int_{-C_2} \frac{f(z')}{z' - z} dz' \\ &= \frac{1}{2\pi i} \int_{C_1} \frac{f(z')}{z' - z} dz' - \frac{1}{2\pi i} \int_{C_2} \frac{f(z')}{z' - z} dz' = \frac{1}{2\pi i} \int_{\gamma_R} \frac{f(z')}{z' - z} dz' - \frac{1}{2\pi i} \int_{\gamma_r} \frac{f(z')}{z' - z} dz' = f_1(z) + f_2(z) \end{aligned} \quad (17.147)$$

où nous avons mis un "+" entre les deux derniers termes, car comme nous allons le voir de suite, le critère de convergence associé à la notation traditionnelle condensée de ce domaine d'étude, fait émerger automatiquement le signe "-".

Pour les deux intégrales f_1, f_2 , nous savons que la fraction peut s'écrire sous la forme d'une série géométrique déjà vue plus haut. Effectivement:

$$\frac{1}{z' - z} = \frac{1}{(z' - z_0) + (z_0 - z)} = \frac{1}{z' - z_0} \frac{1}{1 - \frac{z - z_0}{z' - z_0}} \quad (17.148)$$

en assimilant:

$$x = \frac{z - z_0}{z' - z_0} \quad (17.149)$$

où comme nous l'avons vu, la convergence impose que:

$$|z - z_0| < |z' - z_0| \quad (17.150)$$

afin que x soit en valeur absolue inférieure à 1.

Nous voyons alors apparaître la série géométrique infinie:

$$\frac{1}{z' - z} = \frac{1}{z' - z_0} \left[1 + \left(\frac{z - z_0}{z' - z_0} \right) + \left(\frac{z - z_0}{z' - z_0} \right)^2 + \dots + \left(\frac{z - z_0}{z' - z_0} \right)^n + \dots \right] \quad (17.151)$$

Soit:

$$\frac{1}{z' - z} = \frac{1}{z' - z_0} + \frac{z - z_0}{(z' - z_0)^2} + \frac{(z - z_0)^2}{(z' - z_0)^3} + \dots + \frac{(z - z_0)^n}{(z' - z_0)^{n+1}} + \dots \quad (17.152)$$

Pour revenir à:

$$f_1(z) = \frac{1}{2\pi i} \int_{\gamma_R} \frac{f(z')}{z' - z} dz' \quad (17.153)$$

nous avons en tout point z à l'intérieur du cercle de rayon R dont le bord est décrit par la variable z' et de centre z_0 la convergence qui est assurée car:

$$|z - z_0| < |z' - z_0| \quad (17.154)$$

Nous pouvons alors écrire:

$$f_1(z) = \frac{1}{2\pi i} \int_{\gamma_R} f(z') \left[\frac{1}{z' - z_0} + \frac{z - z_0}{(z' - z_0)^2} + \frac{(z - z_0)^2}{(z' - z_0)^3} + \dots + \frac{(z - z_0)^n}{(z' - z_0)^{n+1}} + \dots \right] dz' \quad (17.155)$$

Intégrant terme à terme, nous mettons en évidence le développement (déjà connu):

$$f_1(z) = \sum_{n=0}^{\infty} c_n (z - z_0)^n \quad (17.156)$$

avec la définition des coefficients c_n , où n est un entier positif ou nul:

$$c_n = \frac{1}{2\pi i} \int_{\gamma_R} \frac{f(z')}{(z' - z_0)^{n+1}} dz' \quad (17.157)$$

Ce développement peut faire penser au développement de Taylor au sens où seules des puissances positives (ou nulles) de $(z' - z_0)$ apparaissent, mais il n'en est pas un dans le cas de la couronne! En effet, c_n ne peut pas être écrit cette fois-ci comme:

$$c_n = \frac{f^{(n)}(z_0)}{n!} \quad (17.158)$$

puisque, par hypothèse, $f(z)$ est supposée analytique dans la couronne seulement, et peut donc fort bien ne pas l'être à l'intérieur du petit cercle de rayon r , en particulier en z_0 , auquel cas $f^{(n)}(z_0)$ peut tout simplement ne pas exister (répétons que z est strictement contraint à se trouver dans la couronne, soit $r < |z| < R$). Nous verrons plus loin ce qui se passe quand $f(z)$ est holomorphe dans ce disque et que, notamment, z_0 n'est pas un point singulier.

Il nous faut encore traiter f_2 . Nous faisons alors le même type de développement que pour f_1 , avec la différence que maintenant:

$$|z - z_0| > |z' - z_0| \quad (17.159)$$

lorsque z' parcourt le petit cercle de rayon r . Pour faire apparaître une série géométrique, il faut écrire cette fois-ci:

$$\frac{1}{z' - z} = -\frac{1}{z - z'} = -\frac{1}{(z - z_0) - (z' - z_0)} = -\frac{1}{z - z_0} \frac{1}{1 - \frac{z' - z_0}{z - z_0}} \quad (17.160)$$

d'où:

$$\frac{1}{z' - z} = -\frac{1}{z - z_0} \frac{1}{1 - \frac{z' - z_0}{z - z_0}} = -\frac{1}{z - z_0} \left[1 + \left(\frac{z' - z_0}{z - z_0} \right) + \left(\frac{z' - z_0}{z - z_0} \right)^2 + \dots + \left(\frac{z' - z_0}{z - z_0} \right)^n + \dots \right] \quad (17.161)$$

Soit:

$$\frac{1}{z' - z} = -\frac{1}{z - z_0} - \frac{z' - z_0}{(z - z_0)^2} - \frac{(z' - z_0)^2}{(z - z_0)^3} - \dots - \frac{(z' - z_0)^n}{(z - z_0)^{n+1}} - \dots \quad (17.162)$$

Intégrant terme à terme, nous mettons en évidence le développement (nouveau):

$$f_2(z) = \sum_{n=1}^{\infty} c_{-n} (z - z_0)^{-n} \quad (17.163)$$

avec:

$$c_{-n} = \frac{1}{2\pi i} \int_{\gamma_r} f(z') (z' - z_0)^{-n-1} dz' \quad (17.164)$$

En changeant n en $-n$ dans la sommation pour f_2 , nous avons pour la somme $f_1(z) + f_2(z)$:

$$f(z) = \sum_{n=-\infty}^{+\infty} c_n (z - z_0)^n \quad (17.165)$$

avec pour l'instant deux c_n distincts:

$$c_{n \geq 0} = \frac{1}{2\pi i} \int_{\gamma_R} \frac{f(z')}{(z' - z_0)^{n+1}} dz' \quad \text{et} \quad c_{n < 0} = \frac{1}{2\pi i} \int_{\gamma_r} f(z') (z' - z_0)^{-n-1} dz' \quad (17.166)$$

Nous allons voir maintenant que des deux relations peuvent être réunies en une seule!

Si nous observons bien ces deux dernières relations, nous constatons qu'elles ne dépendent nullement de z (!) et c'est bien normal puisque les c_n sont les coefficients du développement en série de $f(z)$ et ceux-ci sont les mêmes en n'importe quel point du domaine de définition de la fonction où celle-ci est analytique!

Donc les deux contours (cercles) peuvent être fusionnés en un seul cercle tant que celui-ci est situé dans la couronne et a pour centre z_0 :

$$c_n = \frac{1}{2\pi i} \int_{\gamma_R} \frac{f(z')}{(z' - z_0)^{n+1}} dz' \quad (n = 0, \pm 1, \pm 2, \dots) \quad (17.167)$$

Par ailleurs, le lecteur attentif aura remarqué que ce contour n'a même pas besoin d'être un cercle finalement. Il peut être quelconque tant qu'il est fermé et qu'il se trouve dans un domaine analytique!

Ainsi, on obtient les deux relations:

$$\boxed{\begin{aligned} f(z) &= \sum_{n=-\infty}^{+\infty} c_n (z - z_0)^n \\ c_n &= \frac{1}{2\pi i} \int_{\gamma_R} \frac{f(z')}{(z' - z_0)^{n+1}} dz' \quad (n = 0, \pm 1, \pm 2, \dots) \end{aligned}} \quad (17.168)$$

Les deux relations précédentes définissent la "**série de Laurent**" généralisée. Elle est remarquable et se distingue d'une série de Taylor au sens où elle contient toutes les puissances entières positives et négatives et les coefficients c_n ne sont pas a priori exprimables avec les dérivées de f .

La série de puissances $n \geq 0$ est appelée "**partie régulière**", celle des puissances négatives porte le nom de "**partie principale**".

La série des puissances négatives converge uniformément partout à l'extérieur de γ_r , celle des puissances positives à l'intérieur de γ_R . Au total le développement de Laurent converge uniformément dans le domaine commun, qui est la couronne et donc aussi sur le chemin unique γ .

Montrons maintenant un point que nous avons mentionné plus haut. Si le cercle ne contient pas de singularité, alors tous les coefficients:

$$c_{n<0} = \frac{1}{2\pi i} \int_{\gamma} f(z') (z' - z_0)^{-n-1} dz' \quad (17.169)$$

sont nuls. Notons d'abord que $-n-1$ est un nombre entier positif ou nul que nous noterons p tel que:

$$c_{n<0} = \frac{1}{2\pi i} \int_{\gamma} f(z') (z' - z_0)^p dz' \quad p \in \mathbb{N} \quad (17.170)$$

Nous avons alors l'intégrant suivant dans un chemin fermé:

$$f(z') (z' - z_0)^p \quad (17.171)$$

Or, si nous enlevons la singularité cela impose que $f(z')$ est holomorphe (et de toute façon c'est imposé par tous les développements initiaux sur les séries de Laurent).

$(z' - z_0)^p$ est un polynôme à puissance positive et non nulle et comme nous le savons, tout polynôme satisfaisant ces conditions est dérivable au moins une fois sans faire apparaître de singularité. Ainsi ce terme est aussi holomorphe.

En admettant que le produit de deux fonctions holomorphes est une fonction holomorphe et que le contour γ est fermé, nous avons alors en utilisant le résultat suivant démontré plus haut (pour une fonction holomorphe):

$$\oint_{\gamma} f(z) dz = 0 \quad (17.172)$$

la conséquence immédiate suivante:

$$\int_{\gamma} f(z') (z' - z_0)^p dz' = 0 \Rightarrow c_{n<0} = \frac{1}{2\pi i} \int_{\gamma} f(z') (z' - z_0)^p dz' = 0 \quad (17.173)$$

s'il n'y a pas de singularité dans le petit cercle de la couronne. Nous retrouvons alors dans ce cas un développement avec les seules puissances positives, les $c_{n \geq 0}$ étant cette fois équivalents à:

$$c_n = \frac{f^{(n)}(z_0)}{n!} \quad (17.174)$$

conformément au théorème intégral de Cauchy généralisé démontré plus haut. A contrario, nous voyons bien que c'est la partie principale (quand elle existe) qui contient l'information sur le fait que f n'est pas a priori holomorphe dans le petit disque. L'existence de puissances négatives montre que f n'est visiblement pas bornée en z_0 .

La classification des singularités d'une fonction se fondera précisément sur la considération des caractéristiques de la partie principale du développement de Laurent centré sur un point singulier de cette fonction.

 **Exemple:**

Voyons donc à quoi ressemble la série de Laurent de notre fonction:

$$f(z) = \frac{1}{1+z^2} \quad (17.175)$$

sur un domaine simplement connexe qui serait la couronne entourant la singularité i par exemple (nous aurions pu choisir la deuxième singularité $-i$ mais il fallait bien en prendre une...). Ce qui équivaut donc à chercher le développement en série de puissances de $z - i$.

Nous allons procéder de la manière suivante:

$$f(z) = \frac{1}{1+z^2} = \frac{1}{(z-i)(z+i)} \quad (17.176)$$

Nous allons utiliser pour la suite:

$$\frac{1}{z+i} = \frac{1}{2i + (z-i)} = -\frac{i}{2} \frac{1}{1 - \frac{i}{2}(z-i)} \quad (17.177)$$

La deuxième fraction peut être exprimée en série géométrique si comme nous l'avons déjà vu:

$$\left| \frac{i}{2}(z-i) \right| < 1 \quad (17.178)$$

Il vient alors:

$$\frac{1}{1 - \frac{i}{2}(z-i)} = 1 + \frac{i}{2}(z-i) + \left(\frac{i}{2}(z-i)\right)^2 + \left(\frac{i}{2}(z-i)\right)^3 + \dots \quad (17.179)$$

Multiplions les deux membres de cette égalité par $-i/2$ et divisons les ensuite par $z - i$ (le deuxième terme du dénominateur de la fraction initiale) pour obtenir pour le terme de gauche:

$$\begin{aligned} \frac{1}{1 - \frac{i}{2}(z-i)} \frac{-i}{2} \frac{1}{z-i} &= \frac{-i}{2-i(z-i)} \frac{i}{2} \frac{1}{z-i} = \frac{1}{2i + (z-i)} \frac{1}{z-i} = \frac{1}{2iz + 2 + z^2 - 2zi - 1} \\ &= \frac{1}{z^2 + 1} \end{aligned} \quad (17.180)$$

et pour le terme de droite, nous avons:

$$-\frac{i}{2} \frac{1}{z-i} - \left(\frac{i}{2}\right)^2 - \left(\frac{i}{2}\right)^3 (z-i) - \left(\frac{i}{2}\right)^4 (z-i)^2 - \dots \quad (17.181)$$

Nous avons alors au final pour notre série géométrique:

$$\frac{1}{z^2 + 1} = -\frac{i}{2} \frac{1}{z-i} - \left(\frac{i}{2}\right)^2 - \left(\frac{i}{2}\right)^3 (z-i) - \left(\frac{i}{2}\right)^4 (z-i)^2 - \dots \quad (17.182)$$

Nous voyons donc sur cette série de Laurent autour de i de la fonction holomorphe $f(z)$ dans la couronne, apparaître les coefficients:

$$c_{-1} = -\frac{i}{2}, c_0 = -\left(\frac{i}{2}\right)^2, c_1 = -\left(\frac{i}{2}\right)^3, \dots \quad (17.183)$$

et nous avons avec Maple 4.00b:

```
>plot3d(abs(-1/2*1/((re+I*im)-I)-(I/2)^2-(I/2)^3*(re-I*im)-(I/2)^4*(re-I*im)^2-(I/2)^5*(re-I*im)^3)
, re=-1.5..1.5, im=-1.5..1.5, view=[-2..2,-2..2,-1..2], orientation=[-130,70]
, contours=50, style=PATCHCONTOUR, axes=frame, grid=[100,100], numpoints=10000);
```

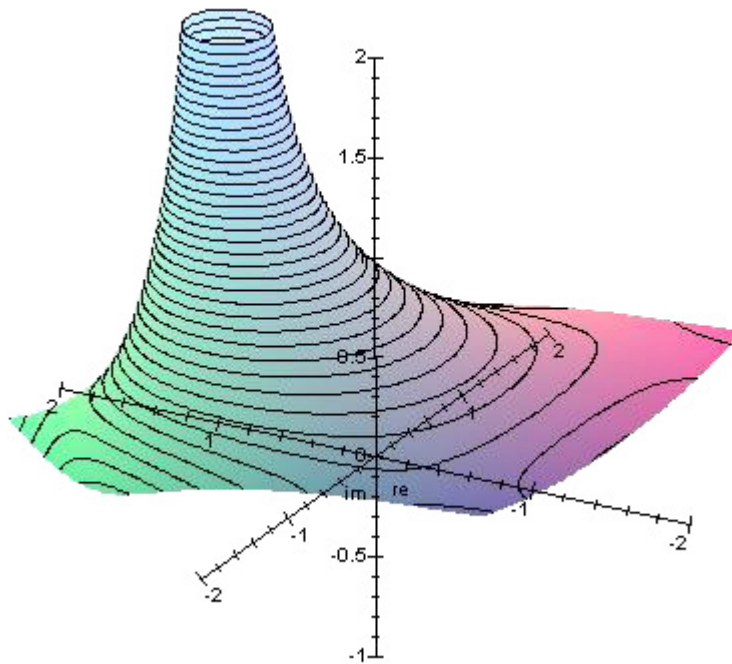


Figure: 17.21 - Représentation de la série de Laurent

où nous voyons que la série de Laurent nous permet d'exprimer $f(z)$ dans un voisinage proche de la singularité i en prenant 5 termes.

Idem si nous faisons la somme des deux séries de Laurent pour les deux singularités avec 7 termes:

```
>plot3d(abs(-1/2*1/((re+I*im)-I)-(I/2)^2-(I/2)^3*(re-I*im)-(I/2)^4*(re-I*im)^2-(I/2)^5*(re-I*im)^3 -(I/2)^6*
(re-I*im)^4-(I/2)^7*(re-I*im)^5+1/2*1/((re+I*im)+I)+(I/2)^2+
(I/2)^3*(re+I*im)+(I/2)^4*(re+I*im)^2+(I/2)^5*(re+I*im)^3+(I/2)^6*(re+I*im)^4
+(I/2)^7*(re+I*im)^5), re=-1.5..1.5, im=-1.5..1.5, view=[-2..2,-2..2,-1..2], orientation=[130,70],
contours=50, style=PATCHCONTOUR, axes=frame, grid=[100,100], numpoints=10000);
```

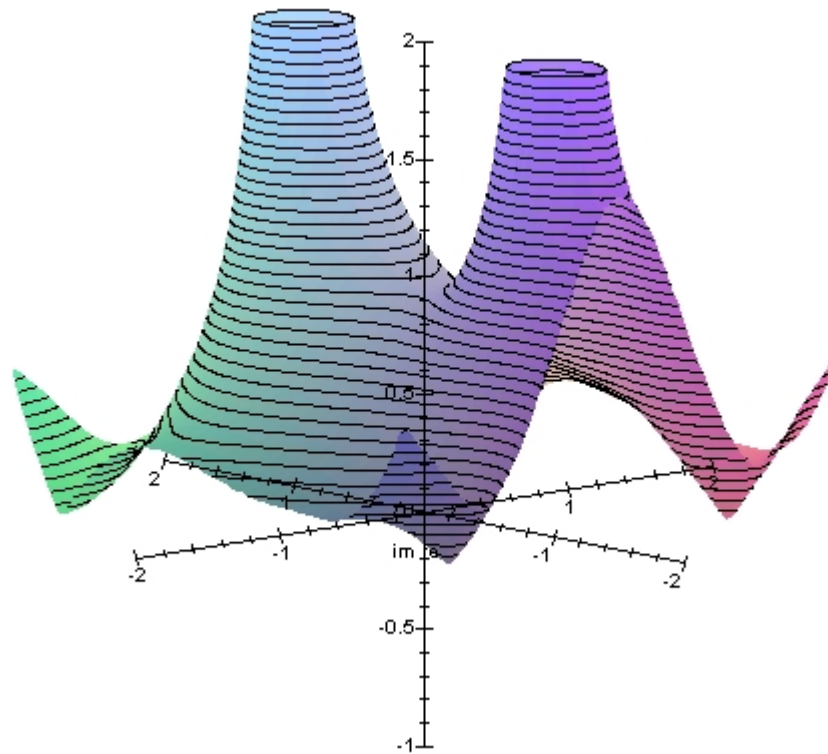


Figure: 17.22 - Somme des deux séries de Laurent pour les deux singularités

et nous voyons que très vite en dehors des deux singularités tout diverge puisque les séries ne convergent que dans une couronne où la fonction y est holomorphe. Mais cela donne déjà une bonne idée visuelle des choses.

7. SINGULARITÉS

Nous avons donc vu juste précédemment qu'il était possible de calculer l'intégrale curviligne d'une fonction, sous condition d'analyticit , sur le contour d'une singularit . Notre objectif va maintenant  tre d'am liorer cette approche.

Nous avons d j  mentionn  et mis en  vidence dans nos d monstrations que l'int grand dans le "th or me int gral de Cauchy"  tait de la forme:

$$\frac{f(z)}{z - z_0} \quad (17.184)$$

o  $f(z)$ est bien d finie en z_0 .

Le point $z = z_0$ est bien  videmment une singularit  de $g(z)$ et celle-ci n'y est donc pas d finie.

Comme nous l'avons vu lors de notre d monstration des s ries de Laurent, $g(z)$ peut  tre exprim e sous forme d'une s rie de Laurent positive dans un disque de convergence (ou ce qui revient au m me: en s rie de Laurent dans une couronne non centr e sur une singularit ...) sous la forme:

$$\begin{aligned}
 g(z) &= \frac{1}{z-z_0} \left(f(z_0) + \frac{1}{1!} f'(z_0)(z-z_0)^1 + \frac{1}{2!} f''(z_0)(z-z_0)^2 + \dots \right) \\
 &= \frac{f(z_0)}{z-z_0} + f'(z_0) + \frac{1}{2!} f''(z_0)(z-z_0) + \dots
 \end{aligned}
 \tag{17.185}$$

Avant de continuer, il est d'usage en mathématiques de définir un petit vocabulaire conventionnel en ce qui concerne cette fois-ci les éventuelles singularités de $f(z)$!

Rappelons au préalable que nous savons, et nous avons démontré, que toutes les informations sur les singularités de $f(z)$ sont contenues dans la partie principale de la série de Laurent (les puissances négatives) définie sur la couronne entourant z_0 :

$$\sum_{n=1}^{\infty} c_{-n} (z-z_0)^{-n} \tag{17.186}$$

La classification ci-après porte sur les "**singularités isolées**", c'est-à-dire un point singulier où $f(z)$ est analytique partout dans le voisinage de z_0 excepté en z_0 . Cette classification, qui nous le verrons permettra de distinguer 3 types de points singuliers, nous sera utile lors du développement de la théorie des résidus plus loin.

Définitions:

D1. Lorsque la limite de la fonction $|f(z)|$ existe en z_0 , nous disons que la singularité est un "**point singulier éliminable**" ou une "**singularité apparente**".

Par exemple:

$$f(z) = \frac{\sin(z)}{z} \tag{17.187}$$

ne semble pas être définie en $z = z_0 = 0$ mais nous avons un numérateur ayant une série de Laurent sans puissances négatives (donc une simple série de Taylor). Il vient alors en faisant la série de Maclaurin (donc la série de Taylor en $z = z_0 = 0$ en d'autres termes...):

$$f(z) = \frac{\sin(z)}{z} = \frac{1}{z} \left(z - \frac{z^3}{3!} + \dots \right) = 1 - \frac{z^2}{3!} + \dots \tag{17.188}$$

Nous voyons que $f(z)$ n'a finalement aucun terme en puissance négative et donc que nous avons éliminé la singularité (ou qu'elle n'en contient au fait pas... ce qui est facilement vérifiable avec Maple 4.00b).

Donc une autre manière équivalente de définir une singularité éliminable, est de dire que le développement de Laurent de la fonction ne contient aucun terme en puissance négative.

D2. Lorsqu'en z_0 la limite de $|f(z)|$ n'existe pas, nous parlons de "**singularité essentielle**".

Par exemple, $z_0 = 0$ est une singularité essentielle pour la fonction:

$$f(z) = e^{\frac{1}{z}} \quad (17.189)$$

En effet, si z tend vers zéro en venant de l'axe réel positif, la fonction diverge, plus précisément, elle tend vers $+\infty$. Si z vient du côté $\mathbb{R}_-$, la fonction tend vers zéro comme le montre bien le tracé Maple 4.00b suivant:

```
>plot3d(abs(exp(1/(re+I*im))),re=-5..5,im=-5..5,view=[-3..3,-3..3,-0.5..3],orientation=[-130,70],contours=50,style=PATCHCONTOUR,axes=frame,grid=[100,100],numpoints=10000);
```

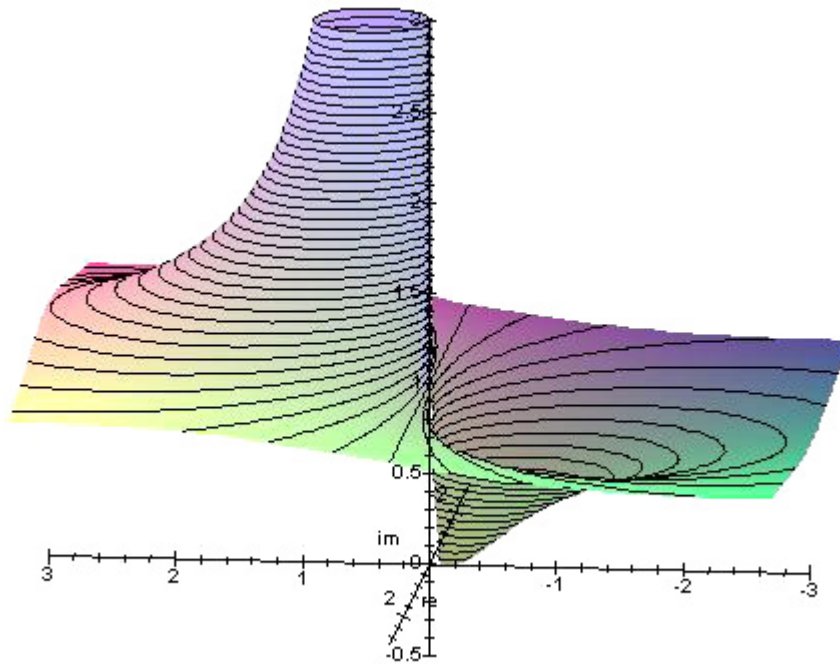


Figure: 17.23 - Représentation de la divergence de la fonction complexe choisie

Effectivement:

$$f(z) = e^{\frac{1}{z}} = \sum_{n=0}^{\infty} \frac{\left(\frac{1}{z}\right)^n}{n!} = \sum_{n=0}^{\infty} \frac{z^{-n}}{n!} \quad (17.190)$$

Donc une autre manière équivalente de définir une singularité essentielle, est de dire qu'il y a un nombre infini de termes à puissances négatives dans la partie principale de la série de Laurent.

D3. Lorsqu'en z_0 la limite de $|f(z)|$ est $+\infty$, nous parlons de "pôle".

Il s'agit de la dernière catégorie dans laquelle nous pouvons ranger une fonction qui n'est classable ni dans la première, ni dans la deuxième définition précédente.

Donc une autre manière équivalente de définir un "pôle", est de dire qu'il y a un nombre fini de termes à puissances négatives dans la partie principale de la série de Laurent. Si ce nombre de termes est k , alors nous parlons de "pôle d'ordre k ".

Remarques:

R1. Nous disons parfois qu'une "singularité essentielle" est un pôle d'ordre infini.

R2. Un pôle d'ordre 1 est appelé un "pôle simple". Un pôle d'ordre 2 est un "pôle double"...

Si nous reprenons notre exemple:

$$f(z) = \frac{1}{1+z^2} \quad (17.191)$$

Nous avons démontré plus haut que la série de Laurent de cette fonction en $z_0 = i$ était:

$$\frac{1}{z^2+1} = -\frac{i}{2} \frac{1}{z-i} - \left(\frac{i}{2}\right)^2 - \left(\frac{i}{2}\right)^3 (z-i) - \left(\frac{i}{2}\right)^4 (z-i)^2 - \dots \quad (17.192)$$

Cette a donc un pôle d'ordre 1 en $z_0 = i$ (et in extenso, nous devinons qu'elle en a un aussi en $z_0 = -i$).

8. THÉORÈME DES RÉSIDUS

Partons d'une fonction $f(z)$ dont le pôle est d'ordre inférieur ou égal à k .

Rendons-là analytique:

$$\phi(z) = (z - z_0)^k f(z) \quad (17.193)$$

(c'est-à-dire que nous avons pris une fonction $f(z)$ que nous avons rendue analytique après élimination de ses pôles supposés en un nombre fini - ordre - inférieur ou égal à k en z_0). Cette fonction $\phi(z)$ a alors un développement en série de Laurent dans un disque centré sur z_0 .

Comme nous l'avons vu plus haut, nous pouvons alors en utilisant la relation ci-dessous:

$$\phi^{(k)}(z_0) = \frac{d^k \phi(z_0)}{dz^k} = \frac{k!}{2\pi i} \oint \frac{\phi(z)}{(z - z_0)^{k+1}} dz \quad (17.194)$$

écrire:

$$\phi^{(k-1)}(z_0) = \frac{(k-1)!}{2\pi i} \oint \frac{\phi(z)}{(z - z_0)^k} dz \quad (17.195)$$

En utilisant $f(z)$ sous l'intégrale, il vient:

$$\phi^{(k-1)}(z_0) = \frac{(k-1)!}{2\pi i} \oint f(z) dz \quad (17.196)$$

Il faut bien analyser cette relation et comprendre qu'elle relie l'intégrale d'une fonction ayant des singularités avec la valeur en un point d'une fonction analytique n'ayant plus de singularités!

Cette dernière relation peut se réécrire en réarrangeant les termes:

$$\oint f(z) dz = \frac{2\pi i}{(k-1)!} \phi^{(k-1)}(z_0) \quad (17.197)$$

Et en exprimant $\phi^{(k)}(z_0)$ en utilisant (autorisé car cette dernière fonction est analytique) le fait que:

$$\phi^{(k)}(z_0) = \frac{d^k \phi(z_0)}{dz^k} \quad (17.198)$$

Nous avons:

$$\oint f(z) dz = \frac{2\pi i}{(k-1)!} \frac{d^{k-1} \phi(z_0)}{dz^{k-1}} \quad (17.199)$$

Soit en explicitant à nouveau $\phi(z)$:

$$\oint f(z) dz = \frac{2\pi i}{(k-1)!} \frac{d^{k-1}}{dz^{k-1}} \left[(z - z_0)^k f(z) \right]_{z=z_0} \quad (17.200)$$

Cette dernière relation n'est pour rappel valable que pour UNE singularité isolée z_0 (au cas où vous auriez oublié les concepts introduits lors de notre présentation des singularités) et où k vaut au minimum 1!

Les mathématiciens définissent alors:

$$\text{Res}[f(z); z_0] = \lim_{z \rightarrow z_0} \frac{1}{(k-1)!} \frac{d^{k-1}}{dz^{k-1}} \left[(z - z_0)^k f(z) \right] \quad (17.201)$$

comme étant le résidu de la fonction $f(z)$ au point $z = z_0$ étant une singularité isolée et ayant un pôle d'ordre k . Ou respectivement:

$$\text{Res}[f(z); z_0] = \frac{1}{2\pi i} \oint f(z) dz \quad (17.202)$$

où l'intégrale curviligne est donc centrée en z_0 .

Remarquons que le terme à droite de l'égalité dans la relation précédente correspond au coefficient c_{-1} de la série de Laurent. Effectivement:

$$c_n = \frac{1}{2\pi i} \int_{\gamma} \frac{f(z')}{(z' - z_0)^{n+1}} dz' \Rightarrow c_{-1} = \frac{1}{2\pi i} \int_{\gamma} \frac{f(z')}{(z' - z_0)^{-1+1}} dz' = \frac{1}{2\pi i} \int_{\gamma} f(z') dz' \quad (17.203)$$

D'où:

$$\boxed{\operatorname{Res}[f(z); z_0] = c_{-1}} \quad (17.204)$$

Remarque: Il vient donc qu'en une singularité isolée éliminable, le résidu est nul puisque comme nous l'avons vu plus haut, l'intégrale curviligne entourant un domaine sans singularité est nulle!

Bref, la relation:

$$\oint f(z) dz = 2\pi i \operatorname{Res}[f(z); z_0] = 2\pi i c_{-1} \quad (17.205)$$

est très intéressante pour le physicien... car il y a donc une manière élégante lui permettant de calculer l'intégrale curviligne d'une fonction $f(z)$ non analytique ayant une unique singularité isolée et ce juste en connaissant l'ordre de ses pôles!

Par exemple si une fonction $f(z)$ n'a qu'un pôle d'ordre 1, il vient alors:

$$\oint f(z) dz = \frac{2\pi i}{(1-1)!} \frac{d^{1-1}}{dz^{1-1}} \left[(z-z_0)^1 f(z) \right]_{z=z_0} = 2\pi i \lim_{z \rightarrow z_0} [(z-z_0) f(z)] \quad (17.206)$$

et nous remplaçons donc z_0 par la valeur voulue dans la parenthèse $(z-z_0)$ et ensuite nous calculons la limite du terme entre crochets!

Maintenant pour aller plus loin, rappelons que le contour de l'intégrale curviligne:

$$\oint f(z) dz \quad (17.207)$$

et le chemin curviligne γ de l'intégrale:

$$c_n = \frac{1}{2\pi i} \int_{\gamma} \frac{f(z')}{(z'-z_0)^{n+1}} dz' \quad (17.208)$$

sont au fait confondus (identiques) et les coefficients c_n ne dépendent pas de z ! La seule contrainte du chemin est qu'il soit fermé et dans un domaine analytique centré sur un point.

Donc si nous avons plusieurs singularités isolées, entourées par des chemins curvilignes reliés tels que présenté ci-dessous sur le plan complexe d'une fonction ayant un pôle d'ordre 3 (donc trois singularités z_0, z_1, z_2 non éliminables):

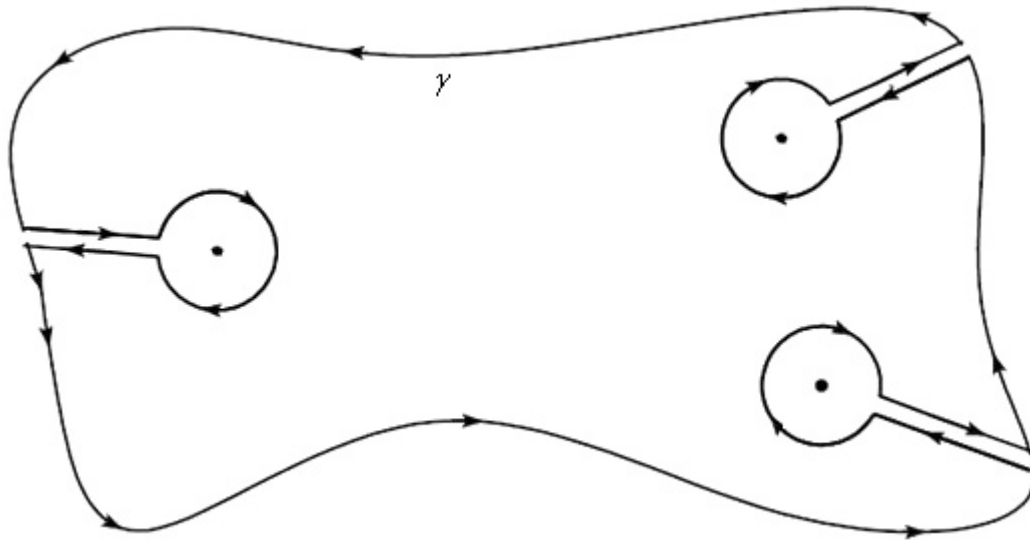


Figure: 17.24 - Singularités multiples isolées entourées par des chemins curvilignes

nous n'avons alors toujours qu'un seul chemin curviligne fermé mais dont les différentes singularités isolées sont reliées par des traverses où comme nous le savons, les chemins qui s'opposent, s'annulent! Et rappelons que les coefficients c_n sont les mêmes partout sur tout le chemin puisque celui-ci est dans un domaine analytique.

Nous avons alors la version généralisée du théorème des résidus pour une fonction f ayant n singularités isolées:

$$\oint f(z) dz = 2\pi i \sum_{i=1}^n \text{Res}[f(z), z_i] \quad (17.209)$$

$$\text{Res}[f(z), z_i] = \lim_{z \rightarrow z_i} \frac{1}{(k_i - 1)!} \frac{d^{k_i-1}}{dz^{k_i-1}} \left[(z - z_i)^{k_i} f(z) \right]$$

avec cette approche rigoureuse digne des ingénieurs... qui notent cette dernière relation parfois:

$$\oint f(z) dz = 2\pi i \sum r_i \quad (17.210)$$

où r est donc un résidu. C'est un résultat important dans le domaine de résolution d'équations différentielles associées à certaines transformées de Laplace inverses (cf. chapitre d'Analyse). Ce résultat intermédiaire nous permettra d'en obtenir un autre un peu plus loin dont l'importance est majeure pour le chapitre de Physique Quantique Corpusculaire.

Exemple:

Reprenons notre fonction:

$$f(z) = \frac{1}{1+z^2} \quad (17.211)$$

Nous savons qu'elle a un pôle d'ordre 1 en $z_0 = i$ et un pôle d'ordre 1 $z_0 = -i$. Donc si nous prenons cette fois la série de Laurent dans un chemin qui entoure les deux singularités (et non plus qu'une seule) nous avons alors une fonction avec un pôle d'ordre 2.

Il vient alors pour ce cas particulier:

$$\oint f(z) dz = 2\pi i \sum_{i=1}^n \text{Res}[f(z); z_i] \quad (17.212)$$

avec donc n valant 2.

Nous avons alors:

$$\begin{aligned} \text{Res}[f(z); z_1] &= \text{Res}[f(z); i] = \lim_{z \rightarrow i} \frac{1}{(k_1 - 1)!} \frac{d^{k_1-1}}{dz^{k_1-1}} \left[(z - z_1)^{k_1} f(z) \right] \\ &= \lim_{z \rightarrow i} \frac{1}{(1-1)!} \frac{d^{1-1}}{dz^{1-1}} \left[(z - i)^1 \frac{1}{z^2 + 1} \right] = \lim_{z \rightarrow i} \left[(z - i)^1 \frac{1}{(z + i)(z - i)} \right] \\ &= \lim_{z \rightarrow i} \left[\frac{1}{z + i} \right] = -\frac{1}{2}i \end{aligned} \quad (17.213)$$

et:

$$\begin{aligned} \text{Res}[f(z); z_2] &= \text{Res}[f(z); -i] = \lim_{z \rightarrow -i} \frac{1}{(k_2 - 1)!} \frac{d^{k_2-1}}{dz^{k_2-1}} \left[(z - z_2)^{k_2} f(z) \right] \\ &= \lim_{z \rightarrow -i} \frac{1}{(1-1)!} \frac{d^{1-1}}{dz^{1-1}} \left[(z - (-i))^1 \frac{1}{z^2 + 1} \right] = \lim_{z \rightarrow -i} \left[(z + i)^1 \frac{1}{(z + i)(z - i)} \right] \\ &= \lim_{z \rightarrow -i} \left[\frac{1}{z - i} \right] = \frac{1}{2}i \end{aligned} \quad (17.214)$$

Nous pouvons vérifier cela avec Maple 4.00b:

```
>readlib(singular);
> singular(1/(1+z^2),z);
> readlib(residue);
> residue(1/(1+z^2),z=-I);
> residue(1/(1+z^2),z=I);
```

et dès lors:

$$\oint \frac{1}{1+z^2} dz = 2\pi i \sum_{i=1}^2 \text{Res}[f(z); z_i] = 0 \quad (17.215)$$

Au fait dans le cas présent, le théorème des résidus est nul car la fonction n'a pas de pôles à l'infini ce qui se vérifie puisque dans notre exemple:

$$\lim_{z \rightarrow \pm i\infty} |f(z)| \neq \infty \quad (17.216)$$

Les physiciens quant à eux diraient que la force ne travaille pas sur le chemin...!

8.1. PÔLE À L'INFINI

Nous avons dit juste précédemment que toute fonction qui n'avait pas de pôle à l'infini avait donc la somme des résidus de tous ces pôles qui était nulle. Ce résultat est très important en physique et mérite d'être approfondi!

Il est assez facile de reconnaître le nombre de pôles... mais pour reconnaître les pôles à l'infini on risque de se faire prendre au piège.

Considérons l'expression $f(z)dz$. Si z est au voisinage de l'infini alors $1/z$ se trouve au voisinage de 0. Posons:

$$t = \frac{1}{z} \quad (17.217)$$

Nous avons alors:

$$f(z)dz = -\frac{1}{t^2} f\left(\frac{1}{t}\right) dt \quad (17.218)$$

Donc le résidu à l'infini est tel que:

$$\text{Res}[f(z); \infty] = R[g(t); 0] \quad (17.219)$$

avec:

$$g(t) = -\frac{1}{t^2} f\left(\frac{1}{t}\right) \quad (17.220)$$

Avec donc:

$$\boxed{\text{Res}[g(t); 0] = \lim_{t \rightarrow 0} \frac{1}{(k-1)!} \frac{d^{k-1}}{dt^{k-1}} \left[(t-0)^k g(t) \right] = -\lim_{t \rightarrow 0} \frac{1}{(k-1)!} \frac{d^{k-1}}{dt^{k-1}} \left[t^k \frac{1}{t^2} f\left(\frac{1}{t}\right) \right]} \quad (17.221)$$

Cette dernière relation nous sera indispensable dans le chapitre de Physique Quantique Corpusculaire pour construire le modèle relativiste de l'atome d'hydrogène de Sommerfeld car nous aurons à y calculer une intégrale curviligne ayant un pôle.

Voyons un exemple avec la fonction qui nous accompagne depuis le début de ce chapitre. C'est-à-dire:

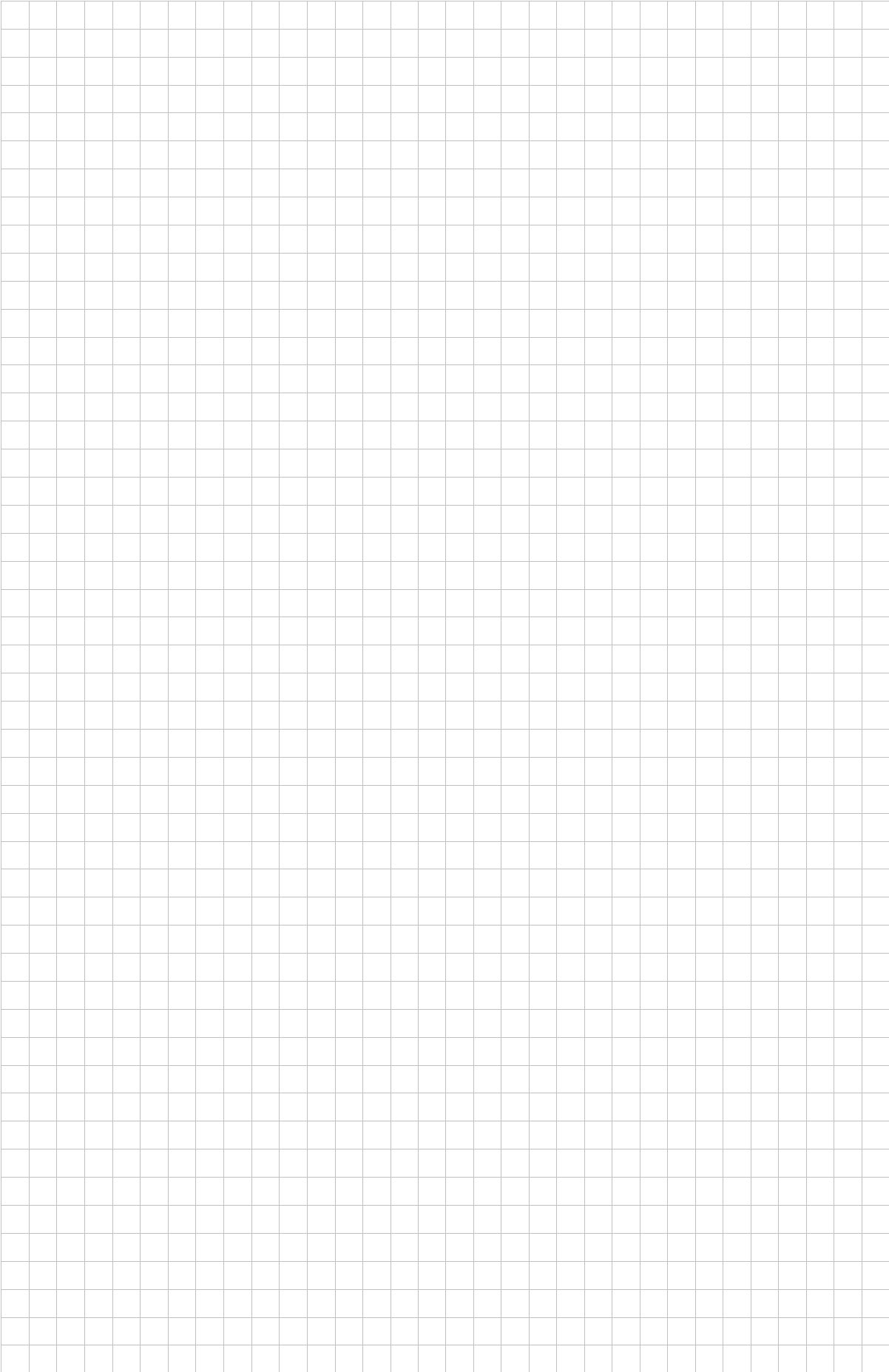
$$f(z) = \frac{1}{z^2 + 1} \quad (17.222)$$

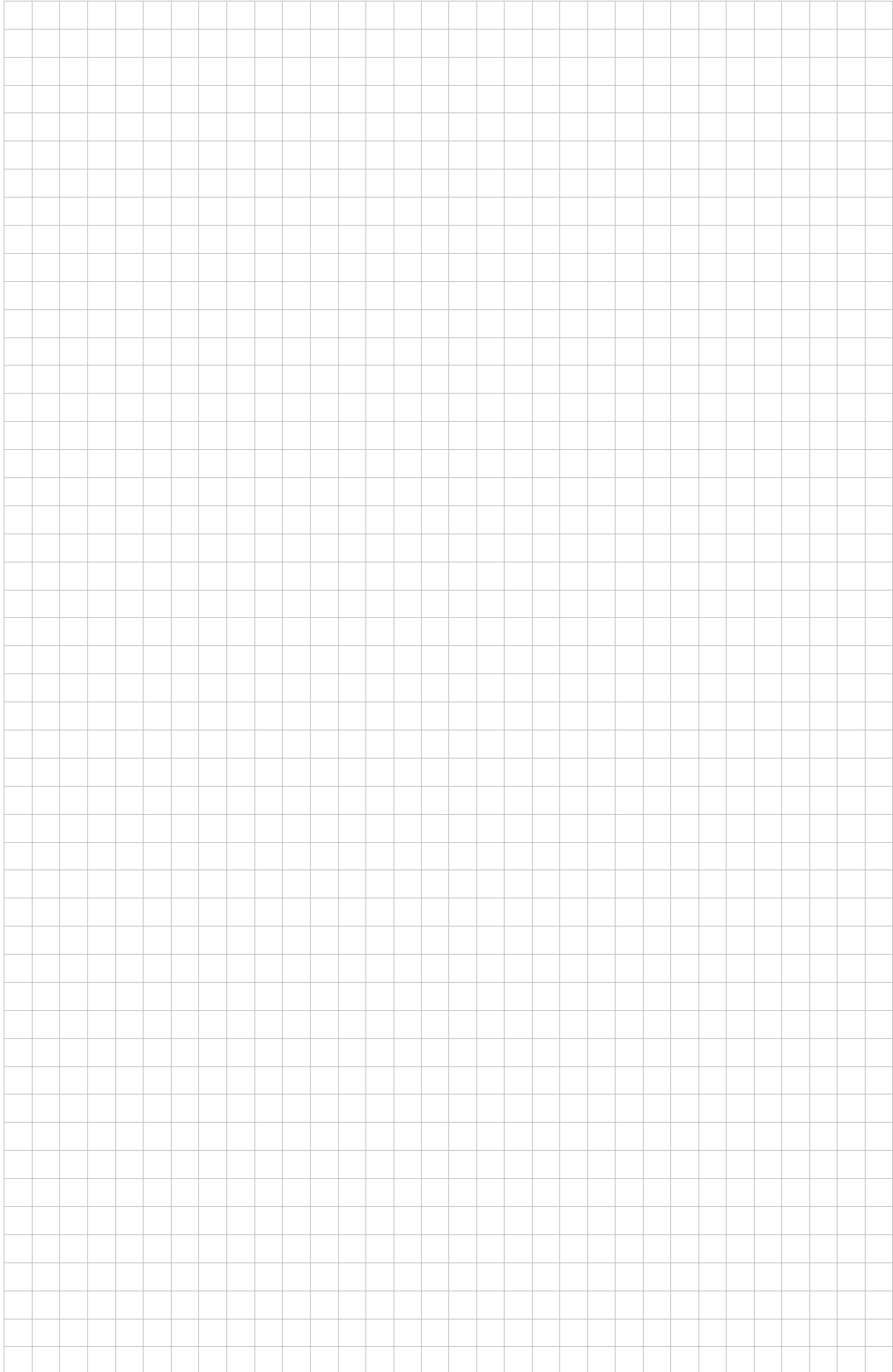
Il vient alors:

$$g(t) = -\frac{1}{t^2} f\left(\frac{1}{t}\right) = -\frac{1}{t^2} \frac{t^2}{1+t^2} = -\frac{1}{1+t^2} \quad (17.223)$$

Or nous reconnaissons immédiatement la fonction initiale au signe près et qui n'a donc pas de pôle en 0. Donc $f(z)$ n'a pas de pôle à l'infini.







18. TOPOLOGIE

La topologie (du grec: discours du lieu) est un domaine extrêmement vaste des mathématiques dont il est difficile de définir avec exactitude l'objet dont elle fait l'étude tellement les domaines où elle existe sont variés (topologie de la droite réelle, topologie des graphes, topologie différentielle, topologie complexe, topologie symplectique,...).

Ce que nous pouvons dire dans un premier temps, c'est que dans ses fondements la topologie est très intimement liée à la théorie des ensembles, à l'étude de convergence des suites et séries, à l'analyse fonctionnelle, à l'analyse complexe, au calcul intégral et différentiel, au calcul vectoriel et à la géométrie pour ne citer que les cas les plus importants se trouvant sur le présent site web.

L'origine de la topologie provient des problèmes qu'ont posé les progrès de l'analyse fonctionnelle dans l'étude rigoureuse des fonctions continues, de leur dérivabilité, de leurs limites en un point (fini ou non), de l'existence d'extremums, etc. dans des espaces de dimensions supérieures (au fait, implicitement la topologie a pour objectif de créer des outils qui permettent facilement d'étudier les propriétés des fonctions dans toutes les dimensions). Tous ces concepts, demandaient pour le mathématicien une définition rigoureuse de l'idée intuitive de proximité, tout particulièrement lors d'opérations sur ces fonctions.

Nous allons essayer de dégager les structures qui permettent de parler de limite et de continuité. L'exemple fondamental que nous prendrons est le cas de $\mathbb{R}$ (la droite de $\mathbb{R}$ pour être rigoureux...).

1. ESPACE TOPOLOGIQUE

Les espaces topologiques forment le socle conceptuel sur lequel les notions de limite, de continuité ou d'équivalence sont définies.

Le cadre est suffisamment général pour s'appliquer à un grand nombre de situations différentes: ensembles finis, discrets, espaces de la géométrie, espaces numériques à n dimensions, espaces fonctionnels les plus complexes. Ces concepts apparaissent dans presque toutes les branches des mathématiques, ils sont donc centraux dans la vision moderne des mathématiques.

Si nous pensons à la droite achevée (droite finie), afin d'étudier les concepts susmentionnés, il va falloir que nous mesurons (imaginons...) des morceaux de celle-ci à la règle. Or, les mesures prises de certains intervalles ou de l'ensemble de la droite doivent pouvoir présenter certaines propriétés minimales que nous allons énoncer tout de suite.

Définition: Soit un ensemble non vide X (la longueur d'une règle de plastique par exemple). Une "topologie $\mathcal{T}$ " ou "espace topologique $(X, \mathcal{T})$ " sur X est une famille $\mathcal{T}$ de parties de X (de longueur de notre règle...) appelées "ouverts" V (comme les intervalles ouverts vus dans le chapitre d'Analyse Fonctionnelle) telle que les axiomes suivants soient vérifiés:

A1. L'ensemble vide et X sont considérés comme des ouverts O et appartiennent obligatoirement à la famille de la topologie $\mathcal{T}$ (ces deux ouverts seuls constituent par ailleurs la "topologie grossière" la plus minimale satisfaisant à tous les axiomes):

$$\emptyset \in \mathcal{T} \text{ et } X \in \mathcal{T} \quad (18.1)$$

En d'autres termes, si nous imaginons notre règle en plastique, la mesure nulle (rigoureusement parlant: l'ensemble vide) doit appartenir à topologie définie sur la règle ainsi que la règle elle-même (vue comme sous-ensemble).

A2. Toute intersection finie d'ouverts de $\mathcal{T}$ est un ouvert de $\mathcal{T}$:

$$V_1, V_2, \dots, V_n \in \mathcal{T} \text{ implique } \bigcap_{i=1}^n V_i \in \mathcal{T} \quad (18.2)$$

A3. Toute réunion d'ouverts de $\mathcal{T}$ est un ouvert de $\mathcal{T}$:

$$V_1, V_2, \dots, V_n \in \mathcal{T} \text{ implique } \bigcup_{i=1}^n V_i \in \mathcal{T} \quad (18.3)$$

Remarques:

R1. Les mathématiciens notent fréquemment par la lettre O la famille des ouverts et F la famille des fermés. Convention que nous ne suivrons donc pas ici.

R2. Les "**fermés**" d'une topologie sont les complémentaires des ouverts. Par conséquent, la famille des fermés contient entre autres X et l'ensemble vide...

R3. Il n'y a pas de différence entre partie et sous-ensemble d'un ensemble.

Le couple $(X, \mathcal{T})$ forme un "**espace de Hausdorff**" ou "**espace séparé**" si de plus la propriété suivante dite "**axiome de Hausdorff**" est vérifiée:

A4. $\forall p, q \in X$ avec $p \neq q$, $\exists V_p, V_q \in \mathcal{T}$ tels que $p \in V_p, q \in V_q$ et $V_p \cap V_q = \emptyset$

Remarques:

R1. Un exemple bien connu d'espace topologique est $\mathbb{R}$ muni de l'ensemble F engendré par les intervalles ouverts (par la loi d'union), c'est-à-dire les intervalles $]a, b[$.

R2. Nous verrons une application très concrète des espaces de Hausdorff lors de notre étude des fractales dans la section d'Informatique Théorique.

Définition: Si nous notons (X, O) un espace topologique, O désignant les ouverts de X , une "**base**", au sens topologique, de (X, O) est une partie B de O telle que tout ouvert de O soit réunion d'ouverts de B (c'est la même idée que les espaces vectoriels au fait mais appliquée à des ensembles... rien de bien méchant).

2. ESPACE MÉTRIQUE ET DISTANCE

Définition: Un "espace métrique" noté (X, d) ou encore X_d est par définition un ensemble X muni d'une application $d : X \times X \rightarrow \mathbb{R}$, appelée "distance" ou "métrique", qui satisfait les axiomes suivants:

A1. $\forall x \neq y \quad d(x, y) > 0$ (positivité)

A2. $\forall x, y \in X \quad d(x, y) = 0 \Leftrightarrow x = y$ (axiome de séparation)

A3. $\forall x, y, z \in X \quad d(x, z) \leq d(x, y) + d(y, z)$ (inégalité triangulaire)

A4. $\forall x, y \in X \quad d(x, y) = d(y, x)$ (axiome de symétrie)

Remarques:

R1. Certains lecteurs verront tout de suite que certaines de ces propriétés ont déjà été vues dans d'autres chapitres du site lors de l'étude des distances entre points fonctionnels et lors de l'étude des normes (inégalité triangulaire démontrée dans le chapitre de Calcul Vectoriel - la symétrie, la nullité, la positivité, la séparation dans le chapitre d'Analyse Fonctionnelle).

R2. Certains auteurs omettent l'axiome A1 ce qui est rigoureusement juste car découle trivialement de A3.

R3. Un espace métrique sera en général noté (X, d) ou bien encore X_d . Nous pouvons également le noter simplement X si la distance d ne peut être confondue.

La "fonction distance" de $\forall x, y \in X$ est donc notée habituellement dans le sens le plus général qui soit en mathématiques:

$$d(x, y) \quad (18.4)$$

Si nous n'imposons pas l'axiome A2, nous disons que d est une "semi-distance" sur X et si nous autorisons une semi-distance d à prendre la valeur $+\infty$, nous préférons dire que d est un "écart".

Remarques:

R1. Si une distance d vérifie la propriété:

$$\forall x, y, z \in X \quad d(x, z) \leq \max\{d(x, y), d(y, z)\} \quad (18.5)$$

propriété plus contraignante que l'inégalité triangulaire dans certains espaces, nous disons que d est "**ultramétrique**".

Un exemple de distance ultramétrique est l'arbre généalogique:

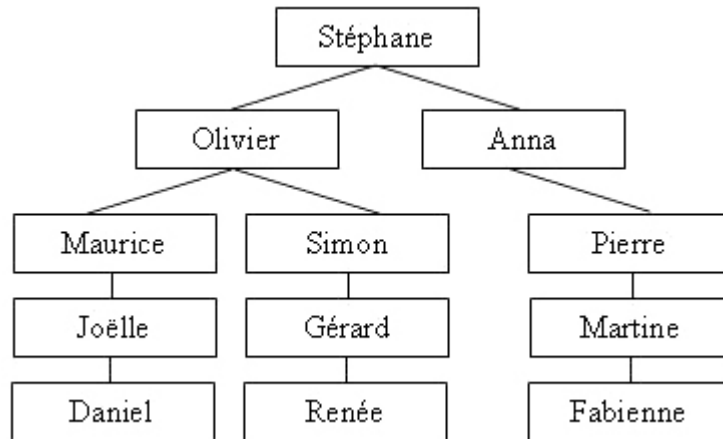


Figure: 18.1 - Exemple de distance ultramétrique d'un organigramme

Nous avons les distances suivantes:

$$\begin{aligned}
 d(\text{Daniel}, \text{Stéphane}) &= 4 \\
 d(\text{Renée}, \text{Stéphane}) &= 4 \\
 d(\text{Renée}, \text{Fabienne}) &= 8 \\
 d(\text{Daniel}, \text{Fabienne}) &= 8 \\
 d(\text{Daniel}, \text{Renée}) &= 6
 \end{aligned} \quad (18.6)$$

Nous remarquons que les distances ne s'additionnent pas, mais que nous avons par contre:

$$d(x, z) \leq \max\{d(x, y), d(y, z)\} \quad (18.7)$$

Ainsi:

$$\begin{aligned}
 d(\text{Daniel}, \text{Fabienne}) &\leq \max\{d(\text{Daniel}, \text{Renée}), d(\text{Renée}, \text{Fabienne})\} \\
 8 &\leq \max\{6, 8\}
 \end{aligned} \quad (18.8)$$

R2. Soit (E, d) un espace métrique et soit $F \neq \emptyset$ une partie de l'ensemble E . L'espace métrique (F, δ) où δ désigne la restriction $d|_{F \times F}$ de d à $F \times F \subset E \times E$ est appelé "**sous-espace métrique**" de (E, d) (il convient de vérifier que la distance d est équivalente à la distance δ). Dans ce cas, nous disons aussi que F est munie de la distance induite par celle de E . Nous notons simplement d la distance induite.

Exemples:

E1. Si nous prenons pour X le plan, ou bien l'espace à trois dimensions de la géométrie euclidienne et une unité de longueur, la "distance" au sens usuel du terme est bien une distance au sens des 5 axiomes précédemment cités. Dans ces espaces, les trois points A, B, C satisfont comme nous l'avons démontré dans le chapitre de Calcul Vectoriel:

$$d(A, B) \leq d(A, C) + d(C, B) \quad d(A, B) \geq |d(A, C) - d(C, B)| \quad (18.9)$$

avec les autres inégalités obtenues par permutation circulaire de A, B, C . Ces inégalités sont bien connues par exemple entre les longueurs des côtés d'un triangle.

E2. Si nous prenons $X = \mathbb{R}^n$, $n \in \mathbb{N} \geq 1$ et que nous dotons $\mathbb{R}^n$ d'une structure d'espace vectoriel euclidienne (et non pas non-euclidienne) et que nous prenons deux points:

$$x = (x_1, x_2, \dots, x_n) \quad y = (y_1, y_2, \dots, y_n) \quad (18.10)$$

dans $\mathbb{R}^n$, la distance est donnée alors par (nous avons déjà démontré cela dans les chapitres d'Analyse Fonctionnelle et de Calcul Vectoriel):

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (18.11)$$

Cette distance satisfait aux 5 axiomes de la distance et nous l'appelons la "**distance euclidienne**". Nous pouvons prendre (c'est une propriété intéressante pour la culture générale), que toute relation de la forme:

$$d(x, y) = \sqrt[p]{\sum_{i=1}^n (x_i - y_i)^p} \quad (18.12)$$

est aussi une distance dans $\mathbb{R}^n$ (sans démonstration). Dans le cas particulier avec $n = 1$, nous avons bien évidemment:

$$d(x, y) = |x - y| = \max\{x - y, y - x\} \quad (18.13)$$

qui est la distance usuelle sur $\mathbb{R}$.

Les mathématiciens font encore plus fort en généralisant encore plus (la démonstration a peu d'intérêt pour l'instant) la relation antérieure (en prenant en compte la définition même de la distance) sous la forme:

$$\forall p \in [1, +\infty] \subset \mathbb{R} : \delta_p(x, y) = \sqrt[p]{\sum_{i=1}^n (d(x_i, y_i))^p} \quad (18.14)$$

qui est appelée "**distance höldérienne**".

Remarque: Suite à l'intervention d'un lecteur nous précisons qu'en toute rigueur l'inclusion ci-dessus devrait être notée $[1, +\infty] \subset \mathbb{R}$ où $\mathbb{R}$ est la droite achevée (précision également valable pour l'inégalité de Minkowski ci-dessous).

Au même titre pour l'inégalité triangulaire, donnée alors par (cf. chapitre de Calcul Vectoriel):

$$\vec{x}, \vec{y} \in \mathbb{R}^3 : \|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\| \Leftrightarrow \sqrt[2]{\sum_{i=1}^3 (x_i + y_i)^2} \leq \sqrt[2]{\sum_{i=1}^3 x_i^2} + \sqrt[2]{\sum_{i=1}^3 y_i^2} \quad (18.15)$$

La généralisation, de par la vérification de l'existence de la distance höldérienne, nous donne la vraie "**inégalité de Minkowski**":

$$x, y \in \mathbb{R}, p \in [1, +\infty[\subset \mathbb{R} : \sqrt[p]{\sum_{i=1}^n (x_i + y_i)^p} \leq \sqrt[p]{\sum_{i=1}^n x_i^p} + \sqrt[p]{\sum_{i=1}^n y_i^p} \quad (18.16)$$

E3. Si nous prenons $X = \mathbb{C}$, nous considérerons la distance:

$$d(z, z') = |z - z'| \quad (18.17)$$

Ainsi, si $z = a + ib = (a, b)$ et $z' = a' + ib' = (a', b')$ nous avons le module qui de la même manière que la norme dans $\mathbb{R}^2$, forme une distance:

$$d(z, z') = \sqrt{(a - a')^2 + (b - b')^2} \quad (18.18)$$

E4. Considérons aussi $E \neq \emptyset$ un ensemble arbitraire. Posons:

$$\delta(x, y) = 1 \text{ si } x \neq y \text{ et } \delta(x, y) = 0 \text{ si } x = y \quad (18.19)$$

Il est assez facile de vérifier que cette distance vérifie les 5 axiomes et qu'elle est de plus ultramétrique. Cette distance est appelée "**distance discrète**" et le lecteur remarquera que, par analogie, nous avons opté d'exprimer cette distance par le symbole de la fonction Dirac δ (ce n'est pas innocent !!) plutôt que le traditionnel d .

2.1. DISTANCES ÉQUIVALENTES

Parfois, deux distances différentes d et δ sur un même ensemble E sont assez ressemblantes pour que les espaces métriques liés $(E, d), (E, \delta)$ possèdent les mêmes propriétés pour certains objets mathématiques définis par d d'une part, et par δ d'autre part. Il existe plusieurs notions de ressemblances dont voici une première (avant les autres qui nécessitent des outils mathématiques que nous n'avons pas encore définis):

Définition: Soient d et δ deux distances sur un même ensemble E , d et δ sont dites "**distances équivalentes**" s'il existe deux constantes réelles $c > 0, C > 0$ telles que

$$\forall x, y \in E : c \cdot d(x, y) \leq \delta(x, y) \leq C \cdot d(x, y) \quad (18.20)$$

soit:

$$c \leq \frac{\delta(x, y)}{d(x, y)} \leq C \quad (18.21)$$

avec $c \leq C$. Nous noterons par ailleurs cette équivalence $d \sim \delta$.

L'intérêt de cette définition est le suivant: si nous avons convergence pour l'une des métriques, alors nous aurons la convergence pour l'autre aussi. Plus clairement:

$$d(x_n, x) \rightarrow 0 \Leftrightarrow \delta(x_n, x) \rightarrow 0 \quad (18.22)$$

in extenso:

$$x_n \xrightarrow{d} x \Leftrightarrow x_n \xrightarrow{\delta} x \quad (18.23)$$

2.2. FONCTIONS LIPSCHITZIENNES

Relativement aux définitions précédentes, nous pouvons maintenant assigner quelques propriétés supplémentaires aux fonctions telles que nous les avons énoncées dans le chapitre de Théorie Des Ensembles:

Soient (E, d) et (F, δ) des espaces métriques, et soit $f: E \rightarrow F$ une fonction. Nous définissons les propriétés suivantes:

P1. Nous disons que f est une "**isométrie**" si (c'est plutôt intuitif...!):

$$\forall x, y \in E: \delta[f(x), f(y)] = d(x, y) \quad (18.24)$$

Si nous prenons la distance usuelle, la fonction k -lipschitzienne s'écrit alors:

$$\forall x, x' \in E, |f(x) - f(x')| \leq k|x - x'| \quad (18.25)$$

ce que nous pouvons écrire également:

$$\frac{|f(x) - f(x')|}{|x - x'|} \leq k \quad (18.26)$$

Ou ce qui revient au même: toutes les cordes tracées entre 2 points quelconques du graphe ont un coefficient directeur (dérivée) compris entre $-k$ et k .

Par exemple, la fonction $\sin(x)$ est 1-lipschitzienne (car la dérivée du cosinus est en valeur absolue comprise entre 0 et 1).

P2. Nous disons que deux espaces métriques sont "**isométriques**" s'il existe une isométrie surjective de l'un sur l'autre (ce qui est assez rassurant en géométrie...).

P3. f est dite "**L-lipschitzienne**" de constante (ou "de rapport") L s'il existe $L \geq 0 \in \mathbb{R}$ tel que:

$$\forall x, y \in E: \delta[f(x), f(y)] \leq Ld(x, y) \quad (18.27)$$

Si $L = 1$, nous disons que f est "**contractante**" (ou une "contraction"), et si $L < 1$, nous disons que f est

strictement contractante.

P4. Toute fonction f lipchitzienne est uniformément continue (voir plus loin le concept "d'uniforme continue") si elle vérifie:

$$\forall \varepsilon > 0, \exists \lambda > 0, \forall x, y \in E : d(x, y) \leq \lambda \Rightarrow \delta[f(x), f(y)] \leq \varepsilon \quad (18.28)$$

avec $\lambda = \varepsilon / L$ et $L \neq 0$ (la réciproque n'est pas vraie: toute fonction uniformément continue n'est pas nécessairement continue). En d'autres termes, si nous pouvons rapprocher deux points aussi près que nous voulons dans un espace, nous le pouvons aussi dans l'autre (ce qui assure en quelque sorte la dérivation).

Remarques:

R1. Une isométrie est toujours injective car:

$$f(x) = f(y) \Rightarrow \delta[f(x), f(y)] = 0 = d(u, v) \Rightarrow u = v \quad (18.29)$$

mais elle n'est pas en général surjective.

R2. Si (E, d) et (F, δ) sont isométriques, du point de vue de la théorie des espaces métriques ils sont indiscernables, puisque toutes leurs propriétés sont les mêmes, mais leurs éléments peuvent être de nature très différente (suites dans l'un et fonctions dans l'autre par exemple).

3. ENSEMBLES OUVERTS ET FERMÉS

Définition: Considérons un ensemble E muni d'une distance d . Un sous-ensemble U de E est dit "**sous-ensemble ouvert**" si, pour chaque élément de U , il existe une distance r non nulle pour laquelle tous les éléments de E dont la distance à cet élément est inférieure ou égale à r , appartiennent à U , ce qui se traduit en langage mathématique:

$$U \text{ ouvert de } E \Leftrightarrow \forall x \in U, \exists r \in \mathbb{R}^+ / \forall y \in E : d(x, y) \leq r \Rightarrow y \in U \quad (18.30)$$

Remarque: Le symbole $/$ signifie dans ce contexte "satisfait la propriété".

Cette définition peut sembler complexe mais en fait, sa signification concrète est plus simple qu'il n'y paraît. En fait, selon cette définition, un ensemble ouvert dans un espace topologique n'est rien d'autre qu'un ensemble de points contigus et sans bords.

L'absence de bord découle de la condition $r \neq 0$. En effet, en raisonnant par l'absurde, si un ensemble ouvert U avait un bord, alors pour chaque point situé sur celui-ci (le bord) il serait toujours possible de trouver un point n'appartenant pas à U aussi proche que l'on veut de lui. Il s'ensuit que la distance r nécessaire devient donc nulle.

Définitions:

D1. Un "**sous-ensemble fermé**" est un "ouvert avec bord".

D2. Un "**voisinage**" d'un point de E est une partie de E contenant un ouvert contenant ce point.

La définition d'un ensemble ouvert peut être simplifiée en introduisant une notion supplémentaire, celle de "**boule ouverte**":

3.1. BOULES

Soit x un élément de E :

Définition: Une "**boule ouverte de centre x et de rayon $r > 0$** " ou "**boule métrique de rayon r centrée en x** " est le sous-ensemble de tous les points de E dont la distance à x est inférieure à r , ce que nous écrivons:

$${}_o B_x^r(E) = \{y \in E \mid d(x, y) < r\} \quad (18.31)$$

Un ensemble ouvert peut également être défini comme un ensemble pour lequel il est possible de définir une boule ouverte en chaque point.

Remarques:

R1. Les ouverts ainsi définis, forment ce que nous appelons une "**topologie induite**" par la distance d ou aussi "**topologie métrique**".

R2. Nous appelons une "**couverture ouverte**" U de E , un ensemble d'ouverts de E dont la réunion est E .

Définition: Une "**boule fermée**" est similaire à une boule ouverte mais diffère dans le sens que nous y incluons les éléments situés à la distance r du centre:

$${}_f B_x^r(E) = \{y \in E \mid d(x, y) \leq r\} \quad (18.32)$$

Remarque: Pour $0 < r < r'$ les inclusions ${}_o B_x^r \subset {}_f B_x^r \subset B(x, r')$ sont des conséquences directes de la définition de boule ouverte et fermée.

Exemple:

La distance usuelle dans $\mathbb{R}$ est donnée par $d(x, y) = |x - y|$. Les boules y sont de simples intervalles.

Pour $x \in \mathbb{R}$ et $r \in \mathbb{R}_+^*$, nous avons:

$${}_o B_x^r =]x - r, x + r[\text{ et } {}_f B_x^r = [x - r, x + r] \quad (18.33)$$

Définition: Une "**sphère**" est donnée par:

$$S_x^r(E) = \{y \in E \mid d(x, y) = r\} \quad (18.34)$$

Remarque: Puisque par définition, $r > 0$, les boules ouvertes et fermées ne sont pas vides car elles contiennent au moins leur centre. Par contre, une sphère peut être vide.

Exemple:

Avec $\mathbb{R}^n, \mathbb{C}^n$ nous avons vu dans les exemples précédents que nous pouvions définir différentes distances. Pour les distinguer, nous les notons:

$$\begin{aligned} d_1(x, y) &= \sum_{i=1}^n |y_i - x_i| \\ d_2(x, y) &= \sqrt{\sum_{i=1}^n (y_i - x_i)^2} \\ d_\infty &= \max\{|y_i - x_i|, x_i - y_i\} \end{aligned} \quad (18.35)$$

Alors, dans $\mathbb{R}^2$ les boules fermées de centre O et de rayon unité équivalentes aux trois formulations précédentes, ont la forme suivante (rappel: $0 < r \leq 1$ dans cet exemple):

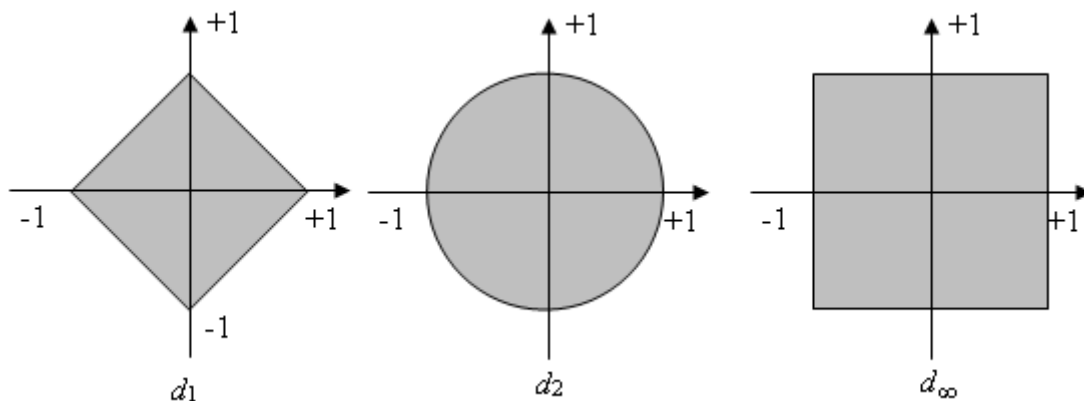


Figure: 18.2 - Exemples de boules fermées de rayon unité

3.2. PARTIES

Maintenant que nous avons défini les concepts de boules, nous pouvons enfin définir rigoureusement les concepts d'intervalles ouverts et fermés (qui dans un espace à plus d'une dimension sont nommés "parties") dont nous avons fait si souvent usage dans les chapitres d'Analyse Fonctionnelle et de Calcul Intégral Et Différentiel.

Définition: Soit (X, d) un espace métrique. Nous disons qu'une partie A de X est "**bornée**" s'il existe une boule fermée ${}_f B_{x_0}^r(X)$ telle que $A \subset {}_f B_{x_0}^r(X)$:

$$\forall x \in A, d(x_0, x) \leq r \quad (18.36)$$

Compte tenu de la remarque précédente sur les inclusions des boules, il est clair que nous pouvons remplacer l'adjectif "fermée" par "ouverte". De plus l'inégalité triangulaire entraîne que le caractère borné de A ne dépend pas du choix de x_0 (avec un x'_0 il suffit de remplacer r par $r' = r + d(x_0, x'_0)$).

Définitions:

D1. Soit X un ensemble et (Y, d) un espace métrique. Si X est un ensemble, nous disons qu'une fonction $f : X \rightarrow Y$ est "**bornée**" si son image $f(X)$ est bornée (cas de la fonction sinus ou cosinus par exemple).

D2. Soit (E, d) un espace métrique, et soit A une partie non vide de E . Pour tout $u \in E$ nous notons $d(u, A)$ et nous appelons "**distance de u à A** ", le nombre réel positif:

$$d(u, A) = \inf_{a \in A} d(u, a) = \inf \{d(u, a) \mid a \in A\} \quad (18.37)$$

Nous prolongeons la notion en posant:

$$\forall u \in E : d(u, \emptyset) = +\infty \quad (18.38)$$

Si A et B sont deux parties de E nous avons respectivement (c'est peut-être plus compréhensible ainsi...):

$$d(A, B) = \inf \{d(x, y), x \in A, y \in B\} \quad (18.39)$$

Il faut faire ici attention à bien interpréter $d(A, B)$ comme l'infimum de la distance entre les ensembles A et B , car la distance entre les parties ne définit pas toujours une distance sur la partie $\mathcal{P}(\mathbb{R})$. Effectivement, Si nous prenons $A = \{0\} \subset \mathbb{R}$ et $B = \left\{\frac{1}{n+1}, n \in \mathbb{N}\right\}$ nous avons $d(A, B) = 0$ quand $n \rightarrow +\infty$ tandis que $A \neq B$.

Remarques:

R1. Si le lecteur a bien compris la définition du concept de "parties" il remarquera qu'il n'existe pas nécessairement toujours un $a \in A$ tel que $d(u, A) = d(u, a)$. En conséquence, nous écrivons trivialement:

$$\forall \varepsilon > 0, \exists \alpha = \alpha(\varepsilon) \in A : d(u, A) \leq d(u, \alpha) < d(u, A) + \varepsilon \quad (18.40)$$

De plus, si un tel α existe, il n'est bien évidemment pas nécessairement unique.

R2. Il convient peut être de rappeler que cette distance satisfait également les 5 axiomes des distances.

D3. Soit (E, d) un espace métrique, et soit A une partie de E . Nous appelons "**adhérence**" de A et notons $\text{adh}(A)$ le sous-ensemble de E défini par:

$$\text{adh}(A) = \{u \in E \mid d(u, A) = 0\} \quad (18.41)$$

Par exemple, l'adhérence du sous-ensemble des rationnels $\mathbb{Q}$ (la partie A) de $\mathbb{R}$ (l'espace métrique E) est un sous-ensemble de $\mathbb{R}$ lui-même puisque tout nombre réel est limite de rationnels.

En particulier, puisque $\forall u \in A : d(u, \emptyset) = +\infty$, nous avons $\text{adh}(\emptyset) = \emptyset$, et puisque

$\forall u \in E : d(u, E) = 0$, nous avons $adh(E) = E$.

Remarques:

R1. Tout élément de l'ensemble $adh(A)$ est dit "**point adhérent**" à A

R2. Nous disons qu'une partie A de E est une "**partie fermée**" si elle est égale à son adhérence

R3. Nous disons qu'une partie A de E est une "**partie ouverte**" si son complémentaire par rapport à E :

$$\complement_E A = E \setminus A \quad (18.42)$$

est fermé.

Il s'ensuit que (de par les définitions):

$$u \in adh(A) \Leftrightarrow \forall r > 0 : {}_r B_u^r \cap A \neq \emptyset \quad (18.43)$$

$$A \text{ est ouverte} \Leftrightarrow \forall u \in A, \exists r > 0 : {}_r B_u^r \subset A \quad (18.44)$$

avec quelques propriétés:

P1. (triviale) Si $A \subset E$ et $B \subset E$ vérifient $A \subset B$, nous avons:

$$\forall u \in E : d(u, B) \leq d(u, A) \quad (18.45)$$

P2. (triviale) Pour tout $A \subset E$, tout $u \in E$:

$$d(u, A) = d(u, adh(A)) \quad (18.46)$$

Dernière propriété qui a pour corollaire (trivial):

Si pour tout $u \in E$ nous avons $d(u, A) = d(u, B)$, $A, B \neq \emptyset$, nous avons alors:

$$adh(A) = adh(B) \quad (18.47)$$

3.3. BOULES GÉNÉRALISÉES

La notion de distance d'un point à un ensemble permet d'étendre les notions de boule et de sphère.

Définitions:

D1. Soit $A \neq \emptyset$ et soit un $r > 0$. Nous appelons "**boule ouverte généralisée**" de centre A et de rayon r , l'ensemble suivant:

$${}_0 B_A^r = \{y \in E \mid d(y, A) < r\} \quad (18.48)$$

Respectivement "**boule fermée généralisée**":

$${}_f B_A^r = \{y \in E \mid d(y, A) \leq r\} \quad (18.49)$$

Respectivement "sphère généralisée":

$$S_A^r = \{y \in E \mid d(y, A) = r\} \quad (18.50)$$

D2. Soit (E, d) un espace métrique et soient A, B deux parties non vides de E . Nous notons $g(A, B)$ et appelons "gap" (qui signifie "écartement" ou "espacement" en français) de A à B , le nombre réel supérieur ou égal à zéro:

$$\begin{aligned} g(A, B) &= \inf_{a \in A} g(a, B) = \inf_{b \in B} g(b, A) = \inf_{a \in A} \inf_{b \in B} g(a, b) \\ &= \inf_{b \in B} \inf_{a \in A} g(a, b) = \inf \{d(a, b) \mid (a, b) \in A \times B\} \end{aligned} \quad (18.51)$$

Remarque: L'inégalité triangulaire $g(A, B) \leq g(A, C) + g(C, B)$ n'est pas valide dans le cadre des gap. Il suffit pour le démontrer, d'un seul et unique exemple qui contredirait l'inégalité.

Exemple:

Dans $\mathbb{R}$ prenons $A = \{0, 1\}, B = \{2, 3\}, C = \{1, 3\}$ nous avons alors:

$$g(A, B) = 1 > g(A, C) + g(C, B) = 0 + 0 = 0 \quad (18.52)$$

Il y a donc bien contradiction.

3.4. DIAMÈTRE

Définition: Soit (E, d) un espace métrique et soit A une partie non vide de E . Nous notons $\text{diam}(A)$ et nous appelons "diamètre" de A , le nombre réel positif non nul:

$$\text{diam}(A) = \sup \{d(a, b) \mid (a, b) \in A \times A\} \quad (18.53)$$

Toute partie non vide A d'un espace métrique vérifiant $\text{diam}(A) < +\infty$ sera aussi dite "bornée".

Remarque: Nous considérons la partie vide $\emptyset$ comme un borné de diamètre A

Si l'espace métrique (E, d) tout entier est borné, nous disons que la distance d est bornée. Par exemple, la distance discrète est bornée, la distance usuelle sur $\mathbb{R}$ ne l'est pas.

Nous avons aussi les propriétés suivantes (les deux premières sont normalement triviales, la troisième découle de la définition du diamètre):

$$P1. \text{diam}(A) = 0 \Leftrightarrow A = \{a\} \text{ ou } A = \emptyset$$

$$P2. A \subset B \Rightarrow \text{diam}(A) \leq \text{diam}(B)$$

$$P3. \text{diam}({}_f B_x^r) \leq 2r, \text{diam}({}_o B_x^r) \leq 2r, \text{diam}(S_x^r) \leq 2r$$

Attention concernant cette propriété il faut perdre l'habitude de penser avec la distance euclidienne. Le premier piège fréquent est de penser que le deuxième diamètre (celui de la boule ouverte) devrait être strictement inférieur mais ce serait oublier que le bord n'a pas d'épaisseur rigoureusement parlant!

Il y a aussi $\text{diam}(S_x^r) \leq 2r$ qui pose souvent problème. Pour s'en convaincre il suffit de prendre la distance discrète (qui vaut si deux points sont pas confondus 0, sinon 1). Ainsi, dans un espace métrique où nous prenons S_x^r avec $r = 1$, nous avons $\text{diam}(S_x^r) = 1$ (c'est un cas intéressant car complètement contre-intuitif).

$$P4. \text{diam}(A \cup B) \leq \text{diam}(A) + g(A, B) + \text{diam}(B)$$

Pour s'en convaincre, dans $\mathbb{R}$ prenons $A = B$, nous avons alors (infériorité stricte triviale):

$$(\text{diam}(A \cup B) = \text{diam}(A)) \leq \text{diam}(A) + g(A, B) + \text{diam}(B) \quad (18.54)$$

$$P5. A \text{ est borné si et seulement si } \exists r > 0, \exists x \in E : A \subset B_x^r$$

Définition: Nous appelons "**excès de Hausdorff**" ou "**distance de Hausdorff**" de A sur B :

$$e(A, B) = \sup\{d(a, B) \mid a \in A\} \quad (18.55)$$

que l'on retrouve aussi souvent noté dans la littérature avec la notation:

$$e(A, B) = d_H(A, B) \quad (18.56)$$

Exemple:

Prenons $A \subset \mathbb{R}^2$ le cercle unité centré à l'origine et pour B le carré qui lui est circonscrit. Des notions de géométrie élémentaire nous amène évidemment à constater que la distance de Hausdorff entre le cercle et le carré est donc:

$$e(A, B) = d_H(A, B) = \sqrt{2} - 1 \quad (18.57)$$

Remarque: Nous avons en général $e(B, A) \neq e(A, B)$ et ces quantités peuvent ne pas être finies.

4. VARIÉTÉS

Nous introduisons maintenant les "variétés". Ce sont des espaces topologiques qui sont "localement comme $\mathbb{R}^n$ " (notre espace par exemple...).

Définitions:

D1. Une "**variété topologique de dimension n** " est un espace de Hausdorff M tel que pour tout $p \in M$ il existe un voisinage ouvert $U \subset M$ avec $p \in U$, un voisinage ouvert $U' \subset \mathbb{R}^n$ et un homéomorphisme:

$$\varphi: U \rightarrow U' \quad (18.58)$$

D2. Un "**homéomorphisme**" entre deux espaces est une bijection continue dont l'inverse est également continue.

D3. Les couples (U, φ) sont appelés des "**cartes**", U étant le "**domaine de la carte**" et φ "**l'application de coordonnées**". Au lieu de "carte" nous disons parfois aussi "**système de coordonnées**".

Remarque: Nous noterons par $\dim M$ la dimension d'une variété topologique. Ainsi:

$$\dim M^{\mathbb{R}} = n \quad (18.59)$$

D4. Soit M une variété topologique de dimension n . Une famille A de cartes de M est appelée un "**atlas**" si pour tout $x \in M$, il existe une carte $(U, \varphi) \in A$ telle que $x \in U$.

Remarque: Notons que si $(U_1, \varphi_1), (U_2, \varphi_2)$ sont deux cartes de M telles que (ne vérifiant pas l'axiome de Hausdorff) $U_1 \cap U_2 \neq \emptyset$, alors l'application de changement de cartes:

$$\varphi_2 \circ \varphi_1^{-1} : \varphi_1(U_1 \cap U_2) \rightarrow \varphi_2(U_1 \cap U_2) \quad (18.60)$$

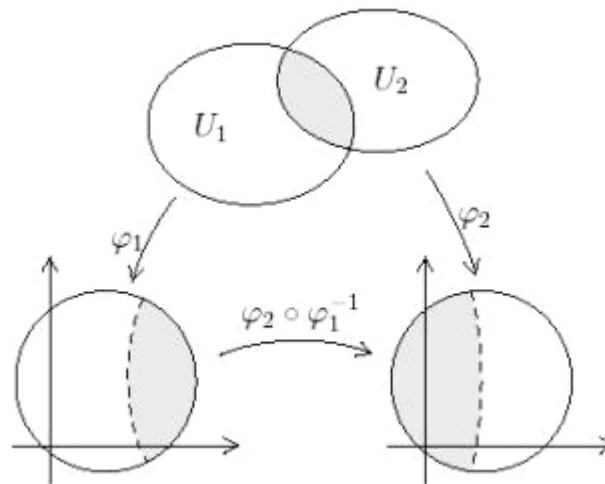


Figure: 18.3 - Exemple de changement de cartes

est un homéomorphisme.

4.1. VARIÉTÉS DIFFÉRENTIABLES

Définitions:

D1. Une "**variété différentiable**" est un espace topologique M où les applications $\varphi_2 \circ \varphi_1^{-1}$ sont des fonctions de classe C^∞ .

D2. Un "**difféomorphisme**" est une application $f : \Omega \rightarrow \Omega'$ où Ω, Ω' sont des domaines ouverts de $\mathbb{R}^n$ et si f est un homéomorphisme et en plus si Ω, Ω' sont différentiables.

Remarque: "différentiable" dans ce contexte signifiera toujours différentiable de classe C^∞

D3. Soit une variété topologique $M = M^*$ (pour simplifier l'écriture), deux cartes $(U_1, \varphi_1), (U_2, \varphi_2)$ de M sont des "**cartes compatibles**" (plus précisément, compatibles de classe C^∞), si l'une des deux propriétés suivantes est vérifiée:

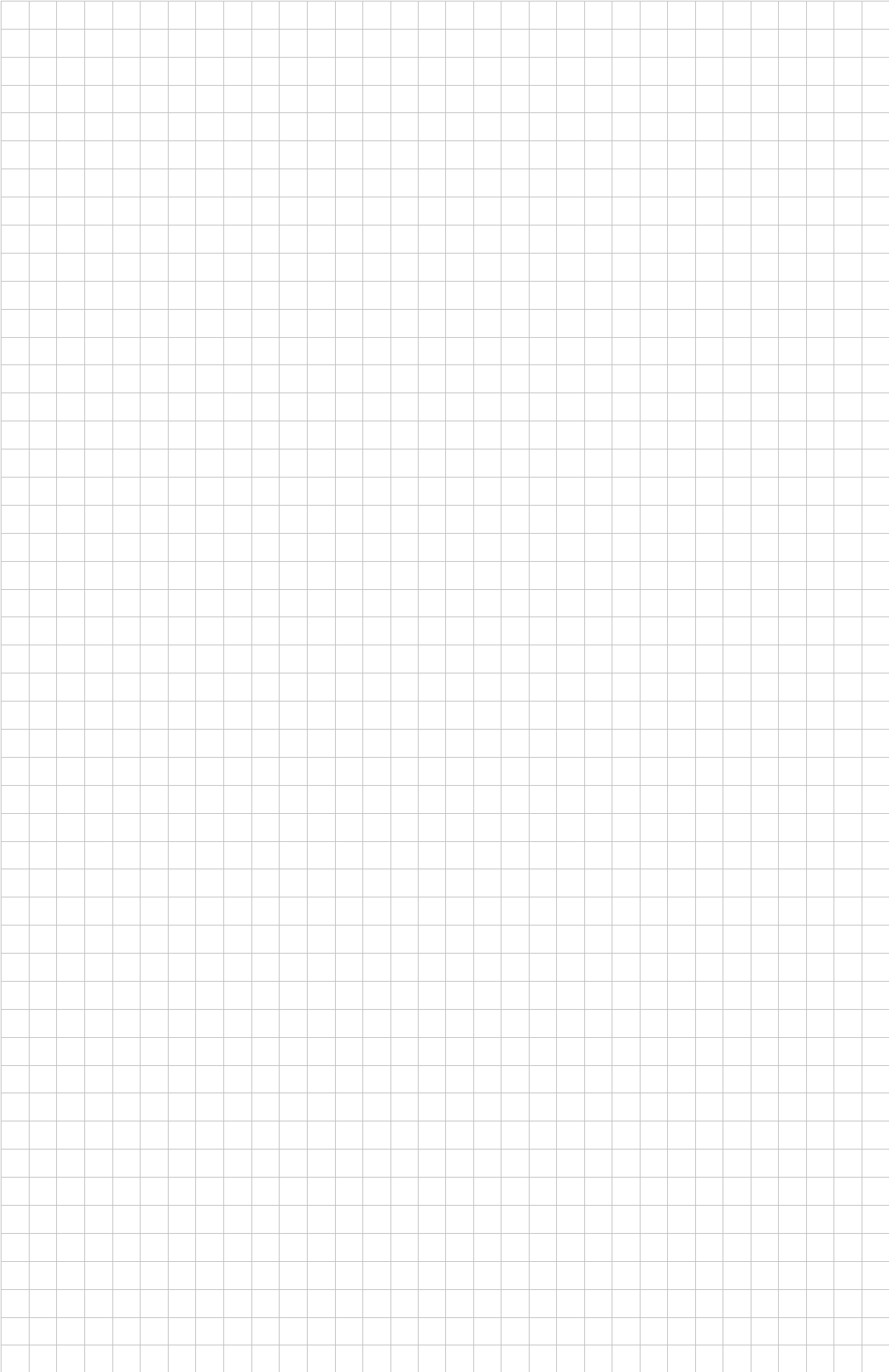
P1. $U_1 \cap U_2 \neq \emptyset$ et l'application $\varphi_2 \circ \varphi_1^{-1}$ de changement de cartes est un difféomorphisme

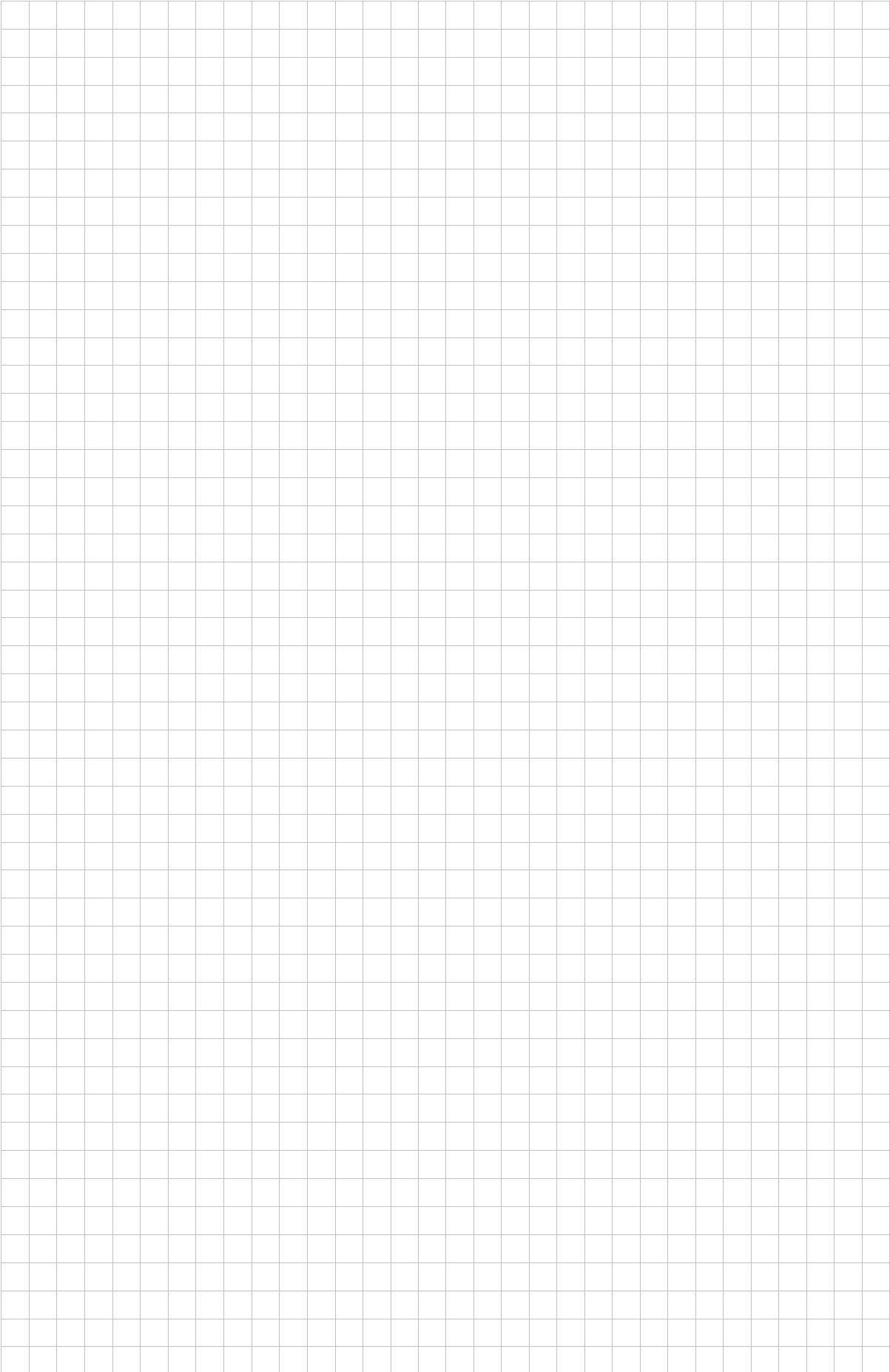
P2. $U_1 \cap U_2 = \emptyset$

Un atlas A de M est différentiable si toutes les cartes de A sont compatibles entre elles.

D4. Une "**variété différentiable**" est un couple (M, A) où M est une variété topologique et A un atlas différentiable de M .

Remarque: Étant donné un atlas différentiable, il est parfois nécessaire de le compléter: nous disons qu'une carte de M est compatible avec un atlas différentiable si elle est compatible avec chaque carte de A . Un atlas de A est un "**atlas maximal**" si toute carte compatible avec A appartient déjà à A . Un atlas maximal est appelé une "**structure différentiable**".







19. THÉORIE DE LA MESURE (ET DE L'INTÉGRATION)

La mesure, au sens topologique, va nous permettre de généraliser la notion élémentaire de mesure d'un segment, ou d'une aire (au sens de Riemann, par exemple) et est indissociable de la nouvelle théorie de l'intégration que Lebesgue mettra en place de 1901 à 1902 et que nous allons aborder ici afin de construire des outils mathématiques beaucoup plus puissant que l'intégrale simple de Riemann (cf. chapitre de Calcul Différentiel Et Intégral).

La théorie de la mesure va également nous permettre de définir avec rigueur le concept de mesure (peu importe la mesure de quoi) et ainsi de revenir sur des résultats importants de l'étude des probabilités (cf. chapitre de Probabilités). Effectivement, nous allons voir (nous définirons le vocabulaire qui suit plus loin) pourquoi $(U, \mathcal{A}, P)$ est un "espace de probabilités" où $\mathcal{A}$ est au fait une tribu sur U et P une mesure sur l'espace mesurable $(U, \mathcal{A})$.

Avertissement: Le niveau d'abstraction et de volonté requis pour la lecture et la compréhension de ce chapitre est assez élevé. Il faut être à l'aise avec les notions vues en théorie des ensembles ainsi qu'en topologie.

1. ESPACES MESURABLES

Quand en mathématiques, nous dérivons, intégrons ou comptons, nous effectuons de manière implicite une mesure d'un objet ou ensemble d'objets. Rigoureusement, les mathématiciens souhaitent définir comment peut être structuré la chose mesurée, comment faire une mesure de celle-ci et les propriétés en découlant!

Définitions:

D1. Soit E un ensemble, une "tribu" (ou " σ -algèbre") sur E est une famille $\mathcal{A}$ de parties de E vérifiant les axiomes suivants:

A1. $E \in \mathcal{A}$ (voir exemples plus bas - E étant considéré comme élément)

A2. Si A est un élément d'une tribu alors $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$. Ce qui signifie que $\mathcal{A}$ est "stable par passage au complémentaire". Cet axiome implique que l'ensemble vide $\emptyset$ est toujours un élément d'une tribu!

A3. Pour toute suite (A_n) d'éléments de $\mathcal{A}$ nous avons $\bigcup_{n=0}^{\infty} A_n \in \mathcal{A}$. Nous disons alors que $\mathcal{A}$ est "stable par union dénombrable".

Par exemple, la graduation d'une simple règle de mesure... satisfait ces trois axiomes.

Remarques:

R1. Nous écrivons $E \in \mathcal{A}$ car nous considérons avec cette notation E non plus comme un sous-ensemble de $\mathcal{A}$ mais comme un élément de $\mathcal{A}$!

R2. Les cas non dénombrables sont typiques de la topologie, de la statistique ou du calcul intégral!

D2. Le couple $(E, \mathcal{A})$ est appelé "**espace mesurable**" et nous disons que les éléments de $\mathcal{A}$ sont des "**ensembles mesurables**".

D3. Si dans le troisième axiome nous imposons que $\mathcal{A}$ soit stable par union finie (non dénombrable) nous imposons alors la notion plus générale "**d'algèbre**". Ainsi, une tribu est nécessairement contenue dans une algèbre (mais le contraire n'est pas vrai car justement l'axiome est plus fort).

Remarque: Dans le domaine des probabilités, E est assimilé à l'Univers des événements et $\mathcal{A}$ à une famille d'événements et nous parlons "**d'espace probabilisable**" ou... "**d'espace mesurable**".

Exemples:

E1. Soit $E = \{1, 2\}$ un ensemble de cardinal 2. Les deux seules tribus $\mathcal{A}$ qui satisfont les trois axiomes sont:

$$\mathcal{A}_1 = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\} \quad \mathcal{A}_2 = \{\emptyset, \{1, 2\}\} \quad (19.1)$$

Il n'y a pas d'autres tribus pour l'ensemble E donné que ces deux (la grossière, et la maximale), car il ne faut pas oublier que l'union de chacun des éléments de la tribu doit aussi être dans la tribu (axiome A3), ainsi que le complémentaire d'un élément (axiome A2).

Nous voyons par ailleurs de cet exemple que si E est un ensemble $\{\emptyset, E\}$ est bien une tribu!

E2. L'ensemble des parties de E , noté $\mathcal{P}(E)$ est aussi une tribu (dixit l'exemple 1).

Une tribu $\mathcal{A}$ est aussi "**stable par union des complémentaires dénombrables**". En effet si (A_i) est une suite d'éléments de $\mathcal{A}$ nous avons (trivial en prenant comme référence l'exemple E1):

$$\bigcup_n A_i^c \in \mathcal{A} \text{ et } \left(\bigcup_n A_i^c \right)^c \in \mathcal{A} \quad (19.2)$$

Une tribu est aussi "**stable par intersections dénombrables**" (trivial en prenant comme référence l'exemple E1):

$$\bigcap_{i=1}^{\infty} A_i \in \mathcal{A} \quad (19.3)$$

ce qui amène à ce qu'une tribu est stable par unions et intersections dénombrables. En particulier, si nous prenons deux éléments d'une tribu $A, B \in \mathcal{A}$ alors $A \setminus B \in \mathcal{A}$. Avec pour rappel (cf. chapitre de Théorie Des Ensembles):

$$\boxed{A \setminus B = A \cap B^c} \quad (19.4)$$

Remarque: Nous voyons aisément avec l'exemple E1 que si $(\mathcal{A}_i)_I$ est une famille de tribus sur E alors

$\bigcap_I \mathcal{A}_i$ est une tribu (la vérification des trois axiomes est immédiate)

Bon c'est bien joli de jouer avec des patates et sous-patates... et leurs complémentaires mais passons à la suite.

Définition: Soit E un ensemble et $\mathcal{B}$ une famille de sous-ensembles de $\mathcal{P}(E)$ tel que $\mathcal{B} \subseteq \mathcal{P}(E)$. Soit $\mathcal{S}$ la famille de toutes les tribus contenant (entre autres) $\mathcal{B}$. $\mathcal{S}$ n'est évidemment pas vide car $\mathcal{P}(E) \in \mathcal{S}$.

Nous notons par définition:

$$\sigma(\mathcal{B}) = \bigcap_{\mathcal{A} \in \mathcal{S}} \mathcal{A} \quad (19.5)$$

la "**tribu engendrée**" par $\mathcal{B}$. $\sigma(\mathcal{B})$ est donc par définition la plus petite tribu contenant $\mathcal{B}$ (et par extension la plus petite tribu de E).

Voici deux exemples qui permettent de vérifier si ce qui précède a été compris et qui permettent de mettre en évidence des résultats importants pour la suite.

Exemples:

E1. Soit E un ensemble, $A \subset E, A \neq E$ et $\mathcal{B} = \{A\}$ alors (lorsque A est vu comme un sous-ensemble de E comme le précise l'énoncé et non comme une famille de sous-ensembles!):

$$\sigma(\mathcal{B}) = \{\emptyset, E, A, A^c\} \quad (19.6)$$

E2. Si $\mathcal{A}$ est une tribu sur E alors:

$$\sigma(\mathcal{A}) = \mathcal{A} \quad (19.7)$$

E3. Soit $E = \{1, 2, 3, 4\}$ et $\mathcal{A} = \{\{1, 2\}, \{3\}\}$ nous avons dès lors (prenez bien garde car maintenant $\mathcal{A}$ est une famille de sous-ensemble et non simplement un unique sous-ensemble!) la tribu engendrée suivante:

$$\sigma(\mathcal{A}) = \{\{\emptyset\}, \{3\}, \{4\}, \{1, 2\}, \{3, 4\}, \{1, 2, 3\}, \{1, 2, 4\}, \{1, 2, 3, 4\}\} \quad (19.8)$$

Plutôt que de déterminer cette tribu en cherchant la plus petite tribu de $\mathcal{P}(E)$ contenant $\mathcal{A}$ (ce qui serait laborieux) nous jouons avec les axiomes définissant une tribu pour facilement trouver celle-ci.

Ainsi, nous trouvons donc bien dans $\sigma(\mathcal{A})$ au moins l'ensemble vide obligatoire $\{\emptyset\}$ ainsi que:

$$E = \{1, 2, 3, 4\} \quad (19.9)$$

selon l'axiome A1 et:

$$\mathcal{A} = \{\{1, 2\}, \{3\}\} \quad (19.10)$$

lui-même par définition de $\sigma(\mathcal{A})$ et les complémentaires de:

$$\{3\}^c = \{1, 2, 4\}, \{1, 2\}^c = \{3, 4\}, \{(\{1, 2\} \cup \{3\})^c = \{4\}\} \quad (19.11)$$

selon l'axiome A2 ainsi que les unions:

$$\bigcup_i \mathcal{A}_i = \{\{1, 2\} \cup \{3\} = \{1, 2, 3\}, \{1, 2\} \cup \{1, 2, 3\} = \{1, 2, 3\}, \dots\} \quad (19.12)$$

selon l'axiome A3.

Définition: Soit E un espace topologique (cf. chapitre de Topologie). Nous notons $\mathcal{B}(E)$ la tribu engendrée par les ouverts de E . $\mathcal{B}(E)$ est appelée la "**tribu borélienne**" sur E . Les éléments de $\mathcal{B}(E)$ sont appelés les "**boréliens**" de E .

Remarques:

R1. La notion de borélien est surtout intéressante car elle est nécessaire à la définition de la "tribu de Lebesgue" et par suite à "la mesure de Lebesgue" qui nous amènera à définir "l'intégrale de Lebesgue".

R2. La tribu $\mathcal{B}(E)$ étant stable par passage au complémentaire, elle contient aussi tous les fermés!

R3. Si E est un espace topologique à base dénombrable, $\mathcal{B}(E)$ est engendré par les ouverts de la base.

Exemple:

Si $\mathbb{R}$ désigne l'espace des réels muni de la topologie euclidienne (cf. chapitre de Topologie), la famille des intervalles ouverts à extrémités rationnelles est une "**base dénombrable**" (étant données les extrémités...) de $\mathbb{R}$ et donc engendre $\mathcal{B}(\mathbb{R})$. Même remarque pour $\mathbb{R}^d$, $d \in \mathbb{N}^*$, avec comme base dénombrable la famille des pavés ouverts à extrémités rationnelles.

Considérons maintenant $\mathcal{S}$ un ensemble dense (cf. chapitre de Topologie) dans $\mathbb{R}$. Les familles suivantes engendrent $\mathcal{B}(\mathbb{R})$:

$$\{]-\infty, a[, a \in \mathcal{S}\}; \{]-\infty, a], a \in \mathcal{S}\}; \{]a, +\infty[, a \in \mathcal{S}\}; \{[a, +\infty[, a \in \mathcal{S}\} \quad (19.13)$$

Démonstration:

Soit (la famille des ouverts):

$$\mathcal{F} := \{]-\infty, a[, a \in \mathcal{S}\} \quad (19.14)$$

Nous avons évidemment:

$$\sigma(\mathcal{F}) \subseteq \mathcal{B}(\mathbb{R}) \quad (19.15)$$

De plus:

$$[a, +\infty[\cap]-\infty, b[= [a, b[\text{ si } a < b \quad (19.16)$$

Donc les intervalles du type $[a, b[$ avec a et b dans $\mathcal{S}$ appartiennent aussi à $\sigma(\mathcal{F})$. Donc, si nous généralisons, avec $x < y$, il existe une suite (a_n) d'éléments de $\mathcal{S}$ décroissant vers x et une suite (b_n) d'éléments de $\mathcal{S}$ croissant vers y tel que:

$$]x, y[= \bigcup_n [a_n, b_n[\in \sigma(\mathcal{F}) \quad (19.17)$$

ce qui entraîne au même titre (que $E \in \mathcal{A}$) que $\mathcal{B}(\mathbb{R}) \subseteq \sigma(\mathcal{F})$. Les autres cas se traitent de manière analogue.

□C.Q.F.D.

Soit $(E, \mathcal{A})$ un espace mesurable et $A \subseteq E$ (et $A \in \mathcal{A}$) (où A est donc considéré comme un sous-ensemble et non comme un élément !). La famille $\{A \cap B \mid B \in \mathcal{A}\}$ est une tribu sur A appelée "**tribu trace**" de $\mathcal{A}$ sur A , nous la noterons $A \cap \mathcal{A}$. De plus, si $A \in \mathcal{A}$, la tribu trace est formée par les ensembles mesurables contenus dans A .

Démonstration: Nous allons faire une démonstration par l'exemple (...). Nous vérifions les trois points de la définition d'une tribu:

1. $A = E \cap A \Rightarrow A \in A \cap \mathcal{A}$
2. Soit $B \in \mathcal{A}$, $A \setminus (A \cap B) = A \setminus B = A \cap B^c$ et donc $A \setminus (A \cap B) \in A \cap \mathcal{A}$

Exemple:

Soit $E = \{1, 2, 3\}$ alors (une tribu parmi d'autres - ne pas oublier la stabilité par union !):

$$\mathcal{A} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\} \quad (19.18)$$

Choisissons $A = \{1, 2\}$, $B = \{2, 3\}$ (il est évident que $\{A \cap B \mid B \in \mathcal{A}\}$ est une tribu sur A).

Dès lors:

$$\begin{aligned} A \setminus (A \cap B) &= \{1, 2\} - ((1, 2) \cap \{2, 3\}) = \{1, 2\} - \{2\} = \{1\} \\ A \setminus B &= \{1, 2\} - \{2, 3\} = \{1\} \\ A \cap B^c &= \{1, 2\} \cap \{2, 3\}^c = \{1, 2\} \cap \{1\} = \{1\} \end{aligned} \quad (19.19)$$

et nous avons bien $\{1\} \in A$ ainsi que $A \in A \cap \mathcal{A}$.

3. Soit $(A \cap B_n)$ une suite d'éléments de $A \cap \mathcal{A}$ ($B_n \in \mathcal{A}$) alors:

$$\bigcup_n (A \cap B_n) = A \cap \underbrace{\left(\bigcup_n B_n \right)}_{\in \mathcal{A}} \in A \cap \mathcal{A} \quad (19.20)$$

La dernière assertion de la proposition sera supposée évidente.

□C.Q.F.D.

Soit maintenant E un ensemble, $\mathcal{C}$ une famille de parties de E et $A \subseteq E$ non vide. Nous notons $A \cap \mathcal{C}$ la trace de $\mathcal{C}$ sur A et $\sigma_A(A \cap \mathcal{C})$ la tribu engendrée par $A \cap \mathcal{C}$ sur A . Alors:

$$A \cap \sigma(\mathcal{C}) = \sigma_A(A \cap \mathcal{C}) \quad (19.21)$$

 Exemple:

Soit l'ensemble $E = \{1, 2, 3, 4\}$, $C = \{\{1, 2\}, \{3\}\}$, $A = \{3, 4\}$ alors:

$$A \cap C = \{\emptyset, \{3\}\} \quad \sigma_A(A \cap C) = \sigma_A(\{\emptyset, \{3\}\}) = \{\emptyset, \{3\}, \{4\}, \{3, 4\}\} \quad (19.22)$$

Et vérifions $A \cap \sigma(C) = \sigma_A(A \cap C)$:

$$\begin{aligned} A \cap \sigma(C) &= \{3, 4\} \cap \sigma(\{1, 2\}, \{3\}) \\ &= \{3, 4\} \cap \{\emptyset, \{3\}, \{4\}, \{1, 2\}, \{3, 4\}, \{1, 2, 3\}, \{1, 2, 4\}, \{1, 2, 3, 4\}\} \quad (19.23) \\ &= \{\emptyset, \{3\}, \{4\}, \{3, 4\}\} \end{aligned}$$

Donc l'égalité est vérifiée.

Un corollaire trivial de cette égalité est que si nous considérons un espace topologique E et $A \subseteq E$ muni de la topologie induite. Alors:

$$A \cap \mathcal{B}(E) = \mathcal{B}(A) \quad (19.24)$$

Rappelons (cf. chapitre de Théorie Des Ensembles) que si nous avons E qui est un ensemble, alors pour tout $A, B \subseteq E$ nous définissons la différence symétrique $A \Delta B$ entre A et B par:

$$A \Delta B = (A \setminus B) \cup (B \setminus A) \quad (19.25)$$

Les propriétés triviales sont les suivantes:

P1. Une algèbre est stable par différence symétrique ($A, B \in \mathcal{A}$ nous avons $A \Delta B \in \mathcal{A}$)

P2. $A \Delta B = B \Delta A$

P3. $A^c \Delta B^c = A \Delta B$

P3. $A \Delta B = (A \cup B) \setminus (A \cap B)$

Si $\mathcal{B}$ est une algèbre sur E , alors $(\mathcal{B}, \Delta, \cap)$ est un "**anneau de Boole**" (ou algèbre de Boole mais attention avec le terme "algèbre" qui peut prêter à confusion avec la théorie des ensembles) avec $\emptyset$ et E comme élément neutre "additif" (Δ) respectivement "multiplicatif" ($\cap$).

Pour des rappels sur les éléments cités dans le paragraphe précédent, le lecteur pourra se reporter au chapitre de Théorie Des Ensembles et le chapitre d'Algèbre De Boole (cf. chapitre de Systèmes Logiques Formels)

Démonstration: Δ ("addition") est associative car en développant nous obtenons (cela se vérifie en faisant un diagramme sagittal au besoin - les "patates"):

$$A \Delta (B \Delta C) = (A \cap B \cap C) \cup (A \cap B^c \cap C^c) \cup (A^c \cap B \cap C^c) \cup (A^c \cap B^c \cap C) \quad (19.26)$$

et cette dernière expression est stable par permutation (commutation) de A et C (même méthode de vérification). Donc:

$$A \Delta (B \Delta C) = C \Delta (B \Delta A) = (B \Delta A) \Delta C = (A \Delta B) \Delta C \quad (19.27)$$

Nous vérifions que $\emptyset$ est bien neutre par rapport à la différence symétrique (la démonstration que E est neutre par rapport à l'inclusion est évidente). Il est trivial que:

$$A \Delta \emptyset = \emptyset \Delta A = A \text{ et que } A \Delta A = \emptyset \quad (19.28)$$

$(\mathcal{B}, \Delta, \emptyset)$ est donc bien un groupe abélien par rapport à la loi Δ (différence symétrique).

□C.Q.F.D.

Pour finir $\cap$ est distributif par rapport à Δ . En effet:

$$\begin{aligned} A \cap (B \Delta C) &= A \cap (B \setminus C \cup C \setminus B) = (A \cap B \setminus C) \cup (A \cap C \setminus B) \\ &= (A \cap B \setminus (A \cap C)) \cup (A \cap C \setminus (A \cap B)) = (A \cap B) \Delta (A \cap C) \end{aligned} \quad (19.29)$$

Ce qui fait bien de $(\mathcal{B}, \Delta, \cap)$ un anneau (qui de plus est un anneau commutatif).

1.1. THÉORÈME DE LA CLASSE MONOTONE

Définition: Soit E un ensemble. Une "classe monotone" sur E est une famille $\mathcal{C}$ de parties de E vérifiant:

A1. $E \in \mathcal{C}$

A2. $A, B \in \mathcal{C}$ et $A \subseteq B \Rightarrow B \setminus A \in \mathcal{C}$

A3. Si (A_n) est une suite croissante (attention au terme "croissant") d'éléments de $\mathcal{C}$ alors $\bigcup_{i=1}^{\infty} A_i \in \mathcal{C}$
(stable par union dénombrable croissant)

Remarques:

R1. Une suite croissante d'ensembles c'est: $A_1 \subseteq A_2 \subseteq A_3 \dots$

R2. Les deux premiers points impliquent que $\mathcal{C}$ est stable par passage au complémentaire.

R3. Les axiomes (2) et (3) (plus le (1)) amènent que la classe monotone est stable par intersection décroissante. Une manière de vérifier c'est de prendre le complémentaire de chaque élément de la suite croissante pour tomber sur la suite décroissante et inversement.

R4. L'axiome 3 des classes monotones étant un peu plus restrictif (plus "fort") que l'axiome 3 des tribus (puisque nous y imposons une suite croissante). Cela implique que toute tribu est une classe monotone (toute union dénombrable de la tribu étant dans la tribu ce qui est une condition plus forte que la suite croissante) !

De la même manière que pour les tribus, si nous considérons une famille $(\mathcal{C}_i)_I$ de classes monotones sur E . Alors $\bigcap_I \mathcal{C}_i$ est une classe monotone (la démonstration se vérifie immédiatement par les trois axiomes précédemment cités).

Exemple:

Si E est un ensemble, $\mathcal{P}(E)$ est une classe monotone sur E . Plus généralement, une tribu est une classe monotone.

De manière équivalente aux tribus, considérons un ensemble E et $\mathcal{C} \subseteq \mathcal{P}(E)$. Soit $\mathcal{S}$ la famille de toutes les classes monotones contenant $\mathcal{C}$. $\mathcal{S}$ n'est pas vide car $\mathcal{P}(E) \in \mathcal{S}$. Nous notons:

$$\mathcal{M}(\mathcal{C}) = \bigcap_{T \in \mathcal{S}} T \quad (19.30)$$

la classe monotone engendrée par $\mathcal{C}$. $\mathcal{M}(\mathcal{C})$ est la plus petite classe monotone contenant $\mathcal{C}$ (et satisfaisant bien évidemment aux axiomes).

Remarque: Si E est un ensemble et $\mathcal{C} \subseteq \mathcal{P}(E)$ alors $\mathcal{M}(\mathcal{C}) \subseteq \sigma(\mathcal{C})$, car $\sigma(\mathcal{C})$ est une classe monotone (et aussi une tribu) contenant $\mathcal{C}$ et donc elle contient aussi $\mathcal{M}(\mathcal{C})$ (voir les exemples avec les tribus).

Le théorème (de la classe monotone) s'énonce ainsi: soit E un ensemble. Si $\mathcal{C}$ est une famille de parties de E que nous imposons stable par intersections finies alors $\mathcal{M}(\mathcal{C}) = \sigma(\mathcal{C})$ (nous devons donc prouver que la tribu minimale de $\mathcal{C}$ est égale à la classe monotone minimale de $\mathcal{C}$). Si nous n'imposons pas que $\mathcal{C}$ soit stable par intersections finies nous n'aurions pas nécessairement l'égalité.

Démonstration:

$\mathcal{M}(\mathcal{C}) \subseteq \sigma(\mathcal{C})$ comme déjà dit (c'est quasiment trivial). Nous allons montrer dans un premier temps que $\mathcal{M}(\mathcal{C})$ est donc aussi une tribu sur E . Pour ceci il suffit de montrer que $\mathcal{M}(\mathcal{C})$ est (aussi) stable par union dénombrable (et non nécessairement par une suite croissante d'éléments!).

Considérons les familles suivantes pour la démonstration:

$$\mathcal{M}_1 = \{A \in \mathcal{M}(\mathcal{C}) \mid A \cap B \in \mathcal{M}(\mathcal{C}) \text{ pour tout } B \in \mathcal{C}\} \quad (19.31)$$

$$\mathcal{M}_2 = \{A \in \mathcal{M}(\mathcal{C}) \mid A \cap B \in \mathcal{M}(\mathcal{C}) \text{ pour tout } B \in \mathcal{M}(\mathcal{C})\} \quad (19.32)$$

Par les définitions précédentes $\mathcal{M}_1 \subseteq \mathcal{M}(\mathcal{C})$ mais $\mathcal{C}$ étant (imposé) stable par intersections finies entraîne $\mathcal{C} \subseteq \mathcal{M}_1$ et donc (c'est le même raisonnement que pour les tribus):

$$\mathcal{C} \subseteq \mathcal{M}_1 \subseteq \mathcal{M}(\mathcal{C}) \quad (19.33)$$

$\mathcal{M}_1$ est une classe monotone en effet $E \in \mathcal{M}_1$, si $A_1, A_2 \in \mathcal{M}_1$ et que $A_1 \subseteq A_2$ (axiome 2) alors:

$$\forall B \in \mathcal{C}, \underbrace{A_1 \cap B}_{\in \mathcal{M}(\mathcal{C})} \subseteq \underbrace{A_2 \cap B}_{\in \mathcal{M}(\mathcal{C})} \quad (19.34)$$

et donc (ce qui appuie le fait que les autres éléments (A_n) satisfont la relation précédente):

$$\forall B \in \mathcal{C}, (A_2 \cap B) \setminus (A_1 \cap B) \in \mathcal{M}(\mathcal{C}) \Rightarrow (A_2 \setminus A_1) \cap B \in \mathcal{M}(\mathcal{C})$$

Si (A_n) est une suite croissante d'éléments de $\mathcal{M}_1$ alors:

$$\forall B \in \mathcal{C}, \left(\bigcup A_n \right) \cap B = \bigcup \underbrace{(A_n \cap B)}_{\in \mathcal{M}(\mathcal{C})} \in \mathcal{M}(\mathcal{C}) \quad (19.35)$$

car $(A_n \cap B)$ est une suite croissante.

Ainsi $\mathcal{M}_1$ est bien une classe monotone et par $\mathcal{C} \subseteq \mathcal{M}_1 \subseteq \mathcal{M}(\mathcal{C})$, nous avons donc:

$$\mathcal{M}_1 = \mathcal{M}(\mathcal{C}) \quad (19.36)$$

Cette dernière égalité implique $\mathcal{C} \subseteq \mathcal{M}_2$. Comme pour $\mathcal{M}_1$, nous montrons que $\mathcal{M}_2$ est une classe monotone et donc $\mathcal{M}(\mathcal{C}) = \mathcal{M}_2$, ce qui veut dire par extension que $\mathcal{M}(\mathcal{C})$ est donc stable par intersections finies.

$\mathcal{M}(\mathcal{C})$ étant stable par passage au complémentaire ceci entraîne que $\mathcal{M}(\mathcal{C})$ est, nous venons de le montrer, stable par unions finies (alors que nous voulons démontrer que c'est stable par union dénombrable).

Soit à présent une suite (A_n) d'éléments de $\mathcal{M}(\mathcal{C})$. Nous considérons la suite:

$$B_n = \bigcup_{i=1}^n A_i \quad (19.37)$$

(B_n) est une suite croissante d'éléments de $\mathcal{M}(\mathcal{C})$, donc:

$$\bigcup_{n=1}^{\infty} B_n \in \mathcal{M}(\mathcal{C}) \text{ mais } \bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n \quad (19.38)$$

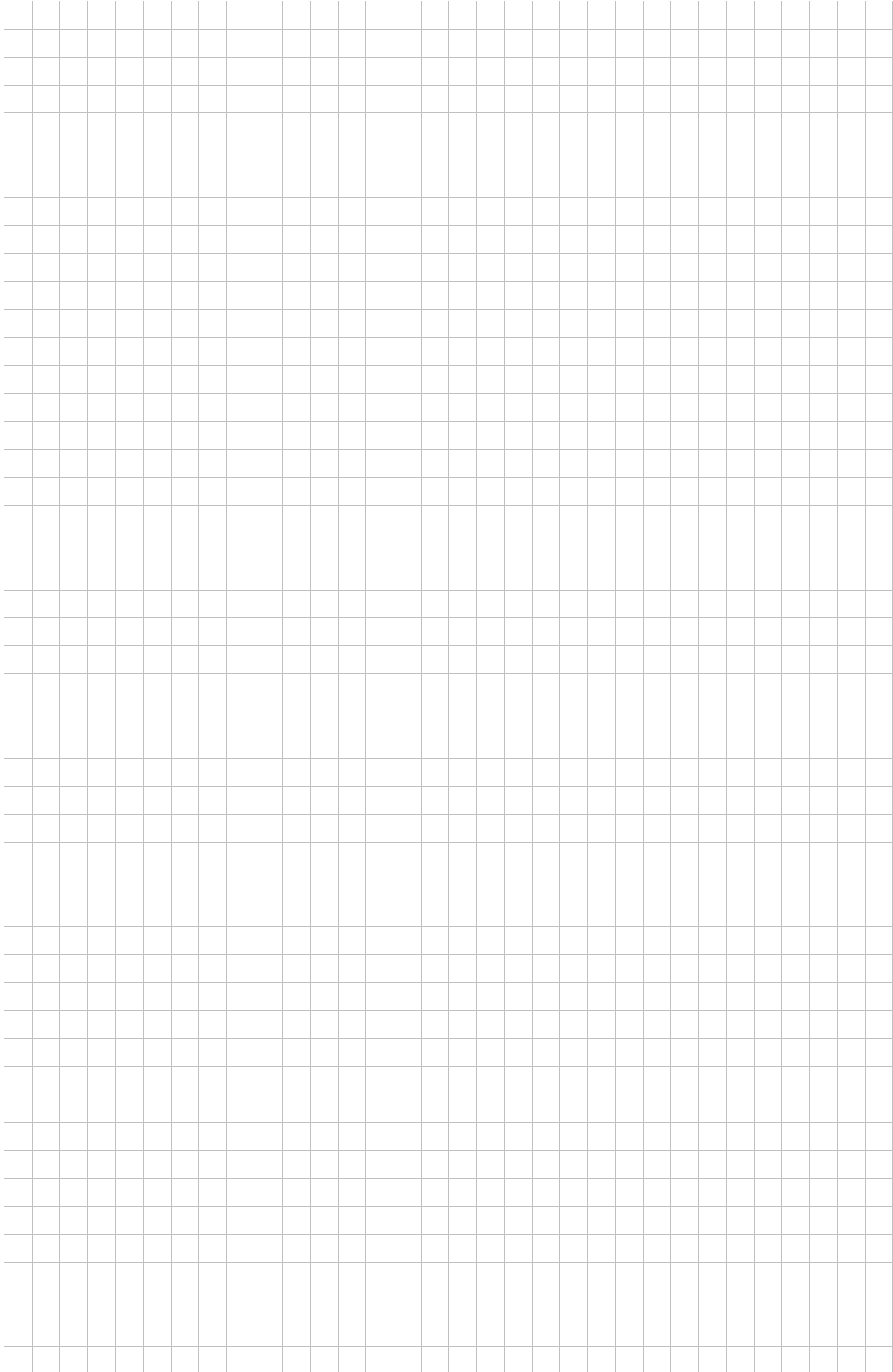
Donc:

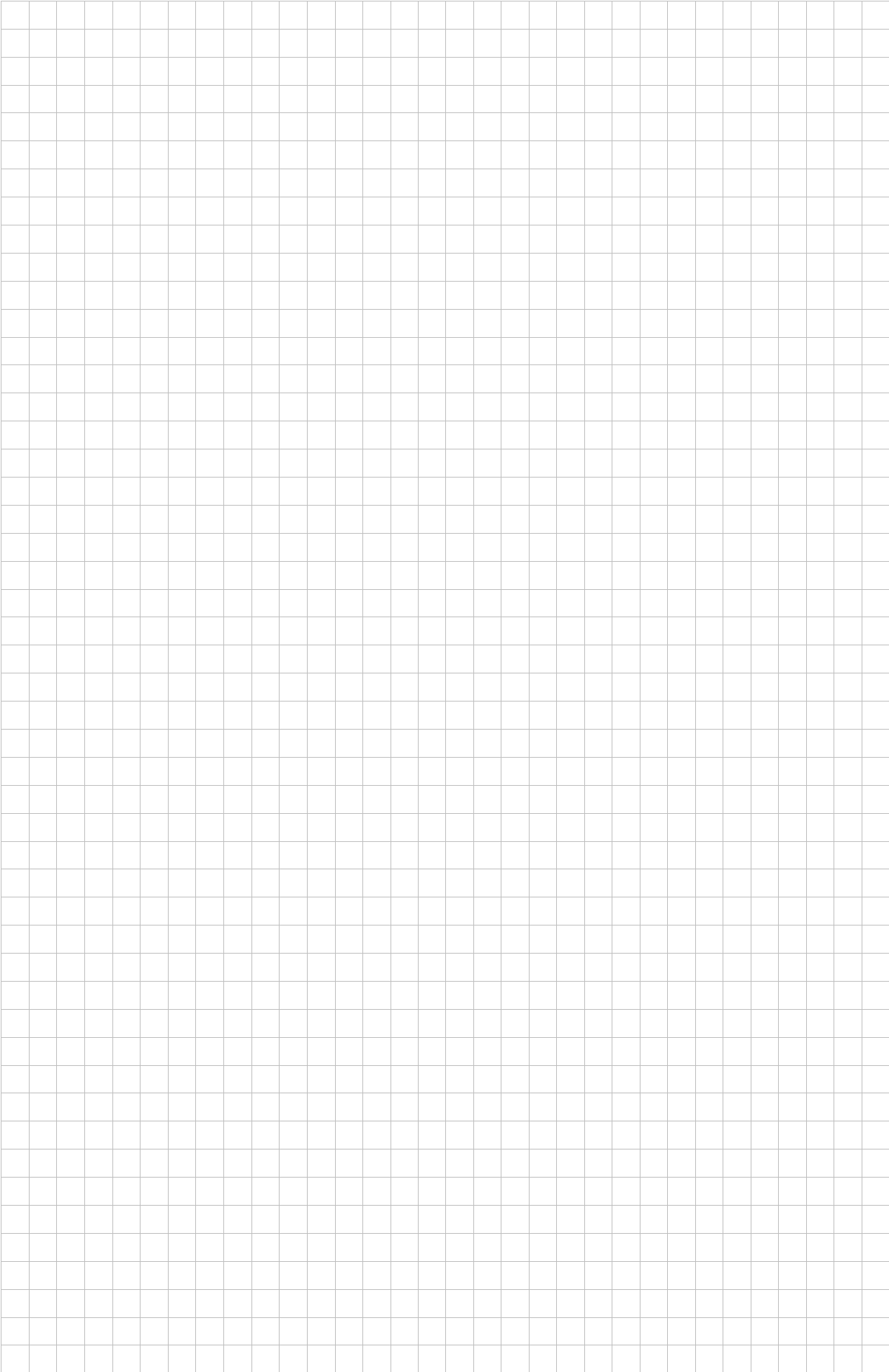
$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{M}(\mathcal{C}) \quad (19.39)$$

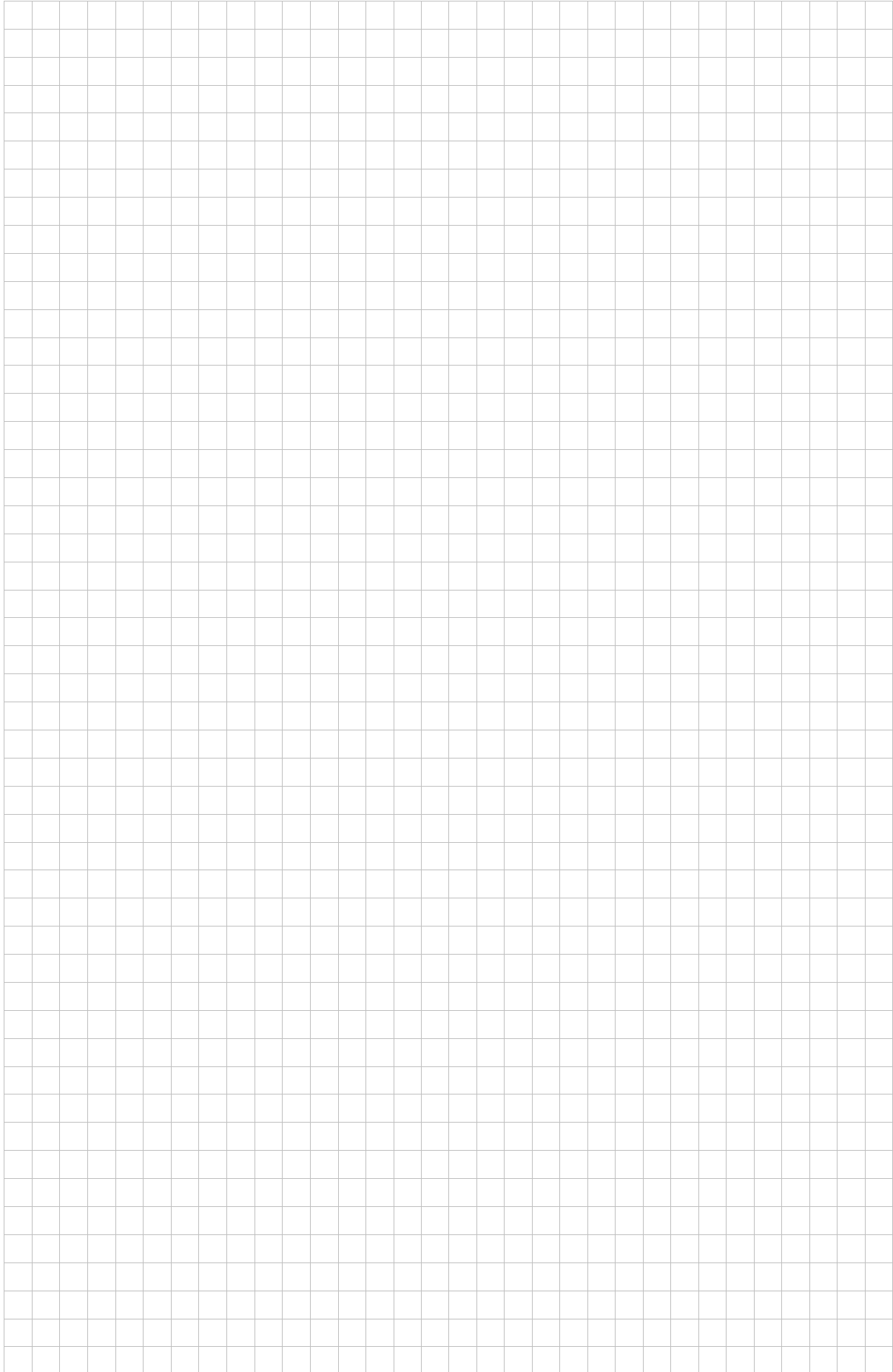
Ainsi $\mathcal{M}(\mathcal{C})$ est stable par union dénombrable et enfin $\mathcal{M}(\mathcal{C})$ est une tribu. Or comme $\mathcal{C} \subseteq \mathcal{M}(\mathcal{C}) \subseteq \sigma(\mathcal{C})$ cela nous amène donc à $\mathcal{M}(\mathcal{C}) = \sigma(\mathcal{C})$.

□C.Q.F.D.

Nous verrons plus tard quelques applications importantes de ce théorème.









Géométrie

La géométrie est la discipline mathématique ayant pour objet l'étude rigoureuse des espaces et des formes. (Larousse)

20. TRIGONOMETRIE

La trigonométrie fait partie intégrante de la science de la géométrie. Cette première ayant pour racine "mesure de la terre" la trigonométrie a pour racine "mesure des corps à trois angles (trigones)".

Remarques:

R1. Il existe actuellement trois trigonométries connues (définies) couramment utilisées en mathématique: la trigonométrie du cercle (assimilée à l'étude des "fonctions circulaires"), la trigonométrie hyperbolique et la trigonométrie sphérique. Nous proposons dans le présent texte une tentative d'approche relativement rigoureuse de toutes les relations les plus connues dans ces trois domaines.

R2. Nous ne traiterons par contre pas ici des trigonométries quadratique et rhombique qui sont utilisées par les électroniciens et qui n'ont peu voire pas d'intérêt en physique théorique. La même remarque est valable pour la trigonométrie lemniscatique qui est en relation avec les mathématiques pures et en particulier la fonction zêta de Riemann.

R3. Le lecteur qui chercherait la démonstration des dérivées et intégrales des fonctions trigonométriques définies ci-après devra se reporter au chapitre de Calcul Différentiel Et Intégral (cf. section d'Algèbre) où les dérivées et intégrales des fonctions usuelles que nous pouvons trouver dans les formulaires sont toutes démontrées.

1. RADIAN

Quand nous parlons de trigonométrie, la première chose qui devrait venir à l'esprit et s'imposer comme standard de mesures d'angles plans (voir le chapitre de géométrie plane pour la définition du concept d'angle) est la notion de "radians".

Définition: 1 "radian" (noté [rad]) est l'angle plan décrit par une sécante à un cercle, passant par son centre, tel que l'arc de cercle ainsi défini par l'axe horizontal passant par le centre du cercle et la sécante soit d'égale longueur au rayon de ce cercle.

Par exemple, pour un cercle de rayon $R = 1$ donc de circonférence (ou périmètre P) $P = 2 \cdot \pi$ la longueur de l'arc de cercle défini par une sécante ayant un angle de 1 radian par rapport à l'horizontale passant par le centre du cercle sera égale à 1.

Dès lors il vient que l'angle pour "un tour" du cercle sera de:

$$2\pi [\text{rad}] \quad (20.1)$$

L'exemple précédent se généralise à un cercle de rayon R quelconque car l'angle pour un tour complet sera toujours $2\pi [\text{rad}]$, pour un demi-tour de $\pi [\text{rad}]$ et pour un quart de $\pi/2 [\text{rad}]$...

Malheureusement dans les écoles, les professeurs du primaire apprennent encore aux enfants à mesurer les angles en degrés. Heureusement la conversion à faire n'est pas trop difficile... (c'est une simple règle de trois).

Soient r la mesure d'un angle en radians, d la mesure du même angle en degrés et g la mesure du même angle en grades (vieille unité) nous avons par définition:

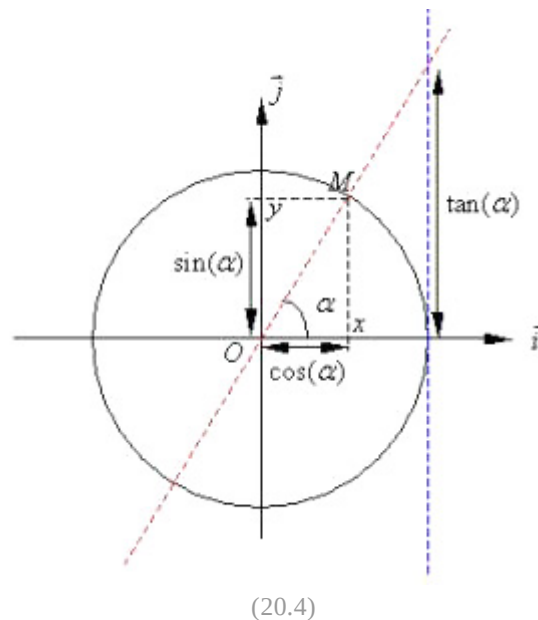
$$\frac{r}{2\pi} = \frac{d}{360} = \frac{g}{400} \quad (20.2)$$

Les astronomes et les astrophysiciens aiment bien parler en minutes ou secondes d'arc telles que:

$$1^\circ = 60' \text{ (minutes d'arc)} \quad 1' = 60'' \text{ (secondes d'arc)} \quad (20.3)$$

2. TRIGONOMÉTRIE DU CERCLE

Soit la figure ci-dessous représentant un cercle quelconque centré à l'origine dans une base directe:



De par l'application du théorème de Pythagore (cf. chapitre de Géométrie Euclidienne), nous y avons:

$$\|\vec{OM}\| = R = \sqrt{x^2 + y^2} \quad (20.5)$$

avec R étant le rayon du cercle.

A partir de cette représentation, nous pouvons définir des êtres mathématiques nommés "**fonctions trigonométriques du cercle**" appelées aussi parfois par les anciens (...) "**fonctions cyclométriques**" telles que (pour les plus importantes):

$\cos(\alpha) = x / R$	$\arccos(x / R) = \alpha$	$\sec(\alpha) = \cos^{-1}(\alpha)$	(20.6)
$\sin(\alpha) = y / R$	$\arcsin(y / R) = \alpha$	$\csc(\alpha) = \sin^{-1}(\alpha)$	
$\tan(\alpha) = y / x$	$\arctan(y / x) = \alpha$	$\cot(\alpha) = x / y$	

Il faut faire attention car suivant les auteurs arccos, arcsin et arctant seront notés respectivement $\cos^{-1}$,

$\sin^{-1}$, $\tan^{-1}$ (cette dernière écriture pouvant prêter à confusion avec l'inverse la fonction trigonométrique correspondante!).

Remarques:

R1. Lisez "**cosinus**" pour "cos", "**sinus**" pour "sin", "**tangente**" pour "tan", "**cotangente**" pour "cot", "**sécante**" pour "sec", "**cosécante**" pour "csc".

R2. Lorsque le contexte le permet et qu'il ne peut y avoir d'ambiguïté, les parenthèses après le nom de la fonction trigonométrique peuvent être omises (c'est souvent le cas en physique).

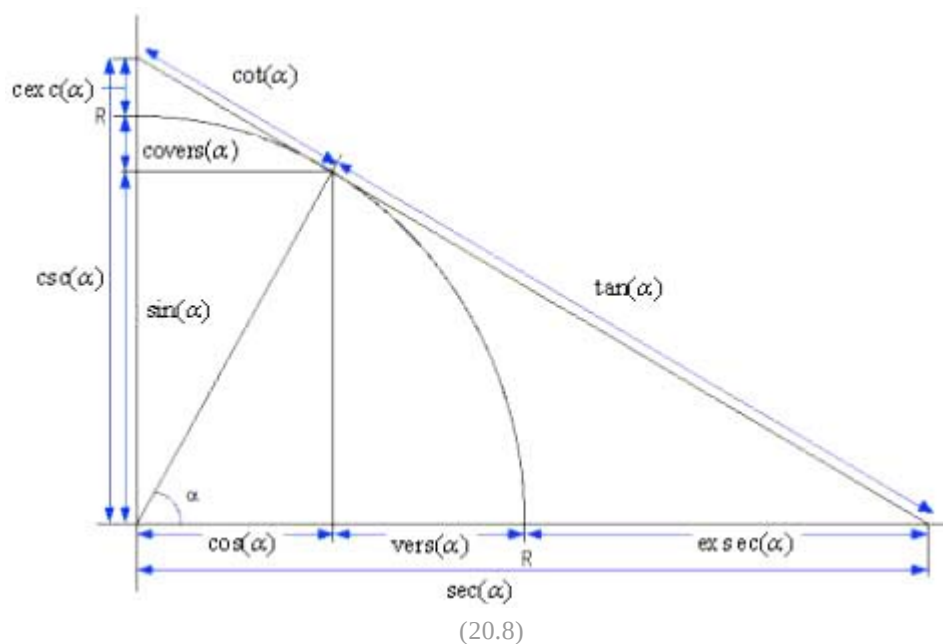
R3. Les fonctions arc... sont donc les fonctions réciproques des fonctions trigonométriques (fonctions bijectives) correspondantes!

A partir de ces fonctions, nous pouvons faire des combinaisons et tirer des relations remarquables très simples mais dont l'utilité profonde est discutable (et qui sont très très peu utilisées) telles que:

$$\begin{aligned} \text{covers}(\alpha) &= 1 - \sin(\alpha) \\ \text{vers}(\alpha) &= 1 - \cos(\alpha) \\ \text{cexc}(\alpha) &= \csc(\alpha) - 1 \\ \text{exsec}(\alpha) &= \sec(\alpha) - 1 \end{aligned} \quad (20.7)$$

mais vous ne les rencontrerez jamais sur ce site Internet car je ne fais personnellement jamais usage de cette notation (c'est plutôt d'usage dans certains ouvrages américains).

Dont voici un superbe schéma... qui résume le tout:



Propriétés:

P1. Si nous nous plaçons dans l'étude du cercle dit "**cercle trigonométrique**", il faut poser pour les

définitions ci-dessus $R = 1$. Ainsi, apparaît plus nettement le sens physique de ces définitions et il en découlera un nombre de propriétés et d'applications directement exploitables dans la physique théorique et la mathématique pure.

Effectivement, si $R = 1$ nous avons trivialement:

$$\overrightarrow{OM} = \cos \alpha \cdot \vec{i} + \sin \alpha \cdot \vec{j} \quad (20.9)$$

et en appliquant le théorème de Pythagore (cf. chapitre de Géométrie Euclidienne):

$$\|\overrightarrow{OM}\| = 1 = \sqrt{\cos^2 \alpha + \sin^2 \alpha} \quad (20.10)$$

d'où:

$$\cos^2 \alpha + \sin^2 \alpha = 1 \quad (20.11)$$

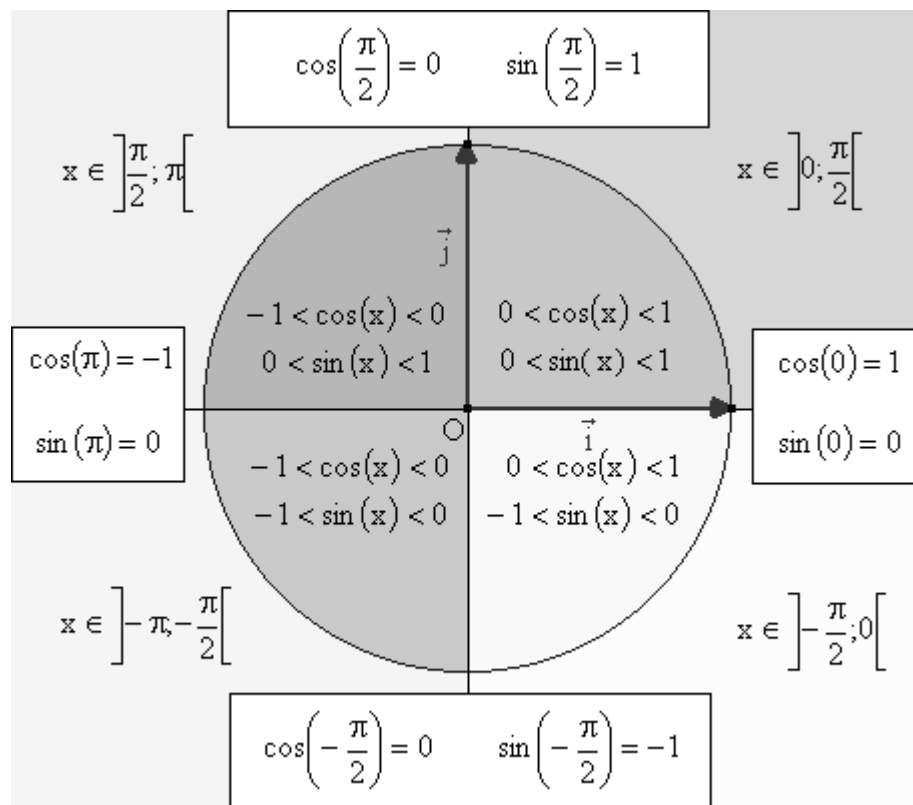
P2. Si α est un réel, et $\forall k \in \mathbb{N}$, les réels α et $\alpha + 2k\pi$ sont associés au même point M de par la périodicité du cercle trigonométrique. En effet, α et $\alpha + 2k\pi$ sont deux mesures du même angle orienté. Ainsi:

$$\forall \alpha \in \mathbb{R}, \forall k \in \mathbb{N}: \begin{aligned} \cos(\alpha \pm 2k\pi) &= \cos(\alpha) \\ \sin(\alpha \pm 2k\pi) &= \sin(\alpha) \end{aligned} \quad (20.12)$$

Idem pour toutes les fonctions trigonométriques qui découlent de la définition des fonctions sinus et cosinus.

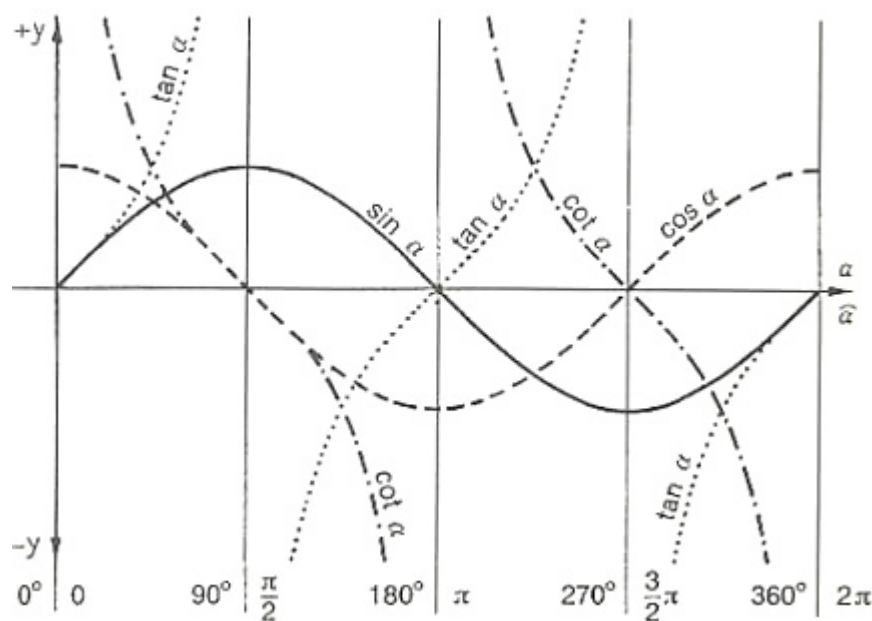
Remarque: Dans la mesure des "**angles orientés**", nous disons que deux mesures sont congrues modulo 2π [rad] si et seulement si leur différence est un multiple de 2π [rad]. Cela caractérise deux mesures d'un même angle.

Par définition, le sinus et le cosinus de tout nombre réel font partie de l'intervalle $[-1, 1]$. Plus précisément, la position de M nous permet d'en savoir plus sur le cosinus et le sinus de α . Ainsi:



(20.13)

Il existe également une autre représentation des fonctions trigonométriques du cercle, un peu plus technique au sens visuel mais assez importante pour bien comprendre, plus tard, la mécanique ondulatoire:



(20.14)

Le lecteur devrait à ce point remarquer sans trop de peine les propriétés suivantes (très souvent utilisées en physique!):

$$\begin{aligned} \sin(n \cdot \pi) &= 0 \quad \forall n \in \mathbb{Z} & \cos(n \cdot \pi) &= (-1)^n \quad \forall n \in \mathbb{Z} \\ \sin\left((2n+1) \cdot \frac{\pi}{2}\right) &= (-1)^n \quad \forall n \in \mathbb{Z} & \cos\left((2n+1) \cdot \frac{\pi}{2}\right) &= 0 \quad \forall n \in \mathbb{Z} \end{aligned} \quad (20.15)$$

et reconnaître facilement que le sinus est une fonction impaire et la fonction cosinus une fonction paire (constat qui nous sera souvent utile dans divers développements mathématiques sur les séries trigonométriques).

Nous avons vu au début de ce chapitre, que de par la définition des fonctions trigonométriques nous avons:

$$\tan \alpha = \frac{\sin \alpha}{\cos \alpha}, \cot \alpha = \frac{1}{\tan \alpha} = \frac{\cos \alpha}{\sin \alpha} \quad (20.16)$$

et également:

$$\cos^2 \alpha + \sin^2 \alpha = 1 \Rightarrow \frac{\cos^2 \alpha}{\sin^2 \alpha} + 1 = \frac{1}{\sin^2 \alpha} \Rightarrow \cot^2 \alpha + 1 = \frac{1}{\sin^2 \alpha} \quad (20.17)$$

De façon exactement identique nous démontrons que:

$$\tan^2 \alpha + 1 = \frac{1}{\cos^2 \alpha} \quad (20.18)$$

A partir de ces dernières relations nous tirons sans trop de peine que:

$$\sin \alpha = \sqrt{1 - \cos^2 \alpha} = \frac{\tan \alpha}{\sqrt{1 + \tan^2 \alpha}} = \frac{1}{\sqrt{1 + \cot^2 \alpha}} \quad (20.19)$$

identiquement nous aurons:

$$\cos \alpha = \sqrt{1 - \sin^2 \alpha} = \frac{\cot \alpha}{\sqrt{1 + \cot^2 \alpha}} = \frac{1}{\sqrt{1 + \tan^2 \alpha}} \quad (20.20)$$

par le raisonnement inverse nous tirons tout aussi facilement que:

$$\tan \alpha = \frac{\sin \alpha}{\sqrt{1 - \sin^2 \alpha}} = \frac{\sqrt{1 - \cos^2 \alpha}}{\cos \alpha} \quad \text{et} \quad \cot \alpha = \frac{\cos \alpha}{\sqrt{1 - \cos^2 \alpha}} = \frac{\sqrt{1 - \sin^2 \alpha}}{\sin \alpha} \quad (20.21)$$

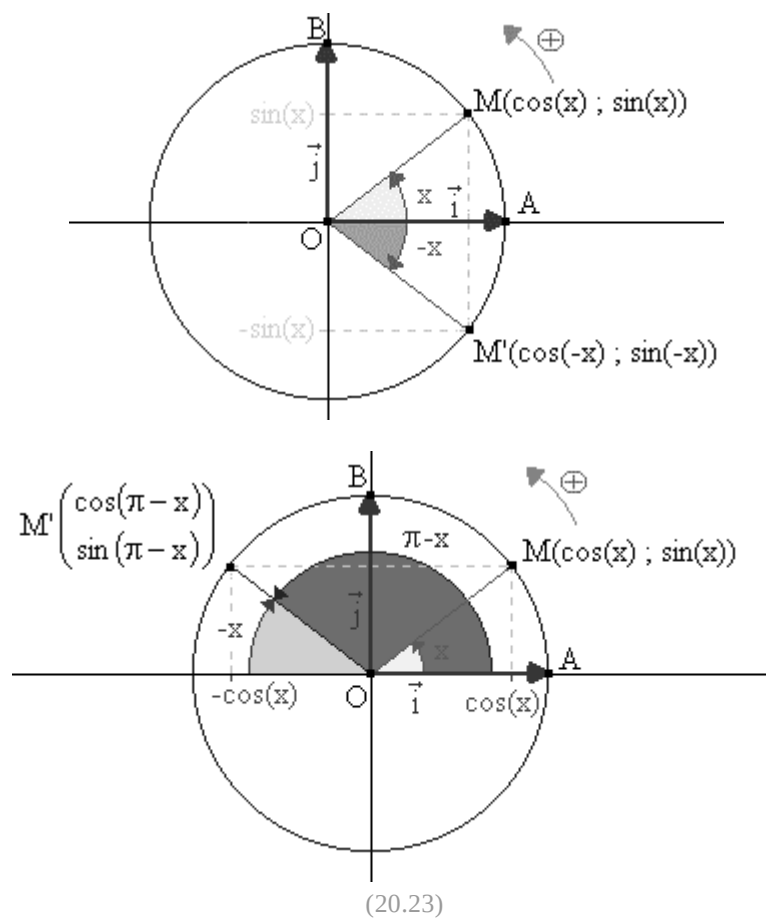
Il vient également sans difficultés en observant le cercle trigonométrique que:

$$\begin{aligned}\sin(-\alpha) &= -\sin \alpha, \quad \sin(\pi - \alpha) = \sin \alpha \\ \sin(\pi + \alpha) &= -\sin \alpha, \quad \sin(\pi/2 \pm \alpha) = \pm \cos \alpha\end{aligned}$$

$$\begin{aligned}\cos(-\alpha) &= \cos \alpha, \quad \cos(\pi \pm \alpha) = -\cos \alpha \\ \cos(\pi/2 - \alpha) &= \sin \alpha, \quad \cos(\pi/2 + \alpha) = -\sin \alpha \quad (20.22)\end{aligned}$$

$$\begin{aligned}\tan(-\alpha) &= -\tan \alpha, \quad \tan(\pi - \alpha) = -\tan \alpha \\ \tan(\pi + \alpha) &= \tan \alpha, \quad \tan(\pi/2 - \alpha) = \cot \alpha \\ \tan(\pi/2 + \alpha) &= -\cot \alpha\end{aligned}$$

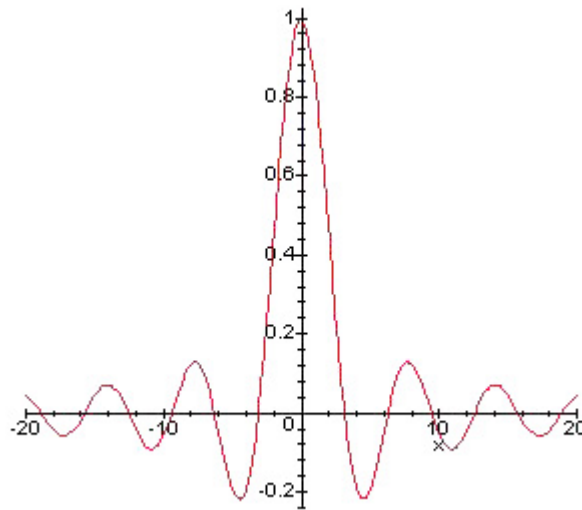
Voici les schémas qui résument la manière d'analyser quelques-unes de ces propriétés (pour les autres relations, la méthode est identique):



Introduisons maintenant une dernière relation que nous retrouvons en optique ondulatoire ou encore dans le cadre des transformées de Fourier qui est le "**sinus cardinal**":

$$\text{sinc}(x) = \frac{\sin(x)}{x} \quad (20.24)$$

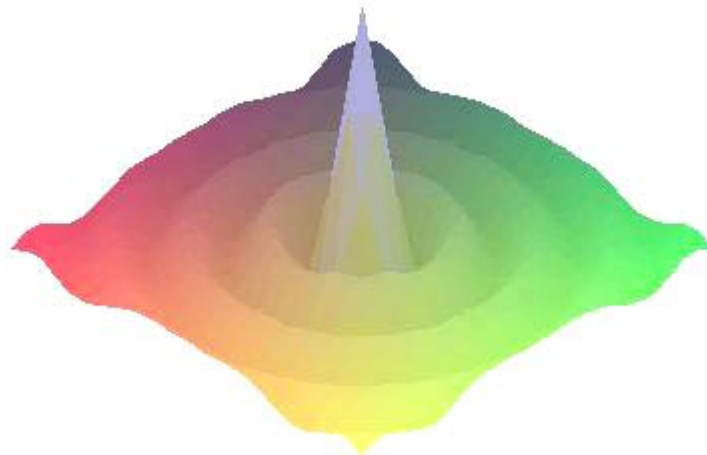
représenté par:



(20.25)

C'est surtout sa forme 3D qui est connue car souvent utilisée pour des raisons de marketing faisant penser à une goutte d'eau tombant dans un récipient d'eau (avec Maple) et c'est toujours joli à regarder...:

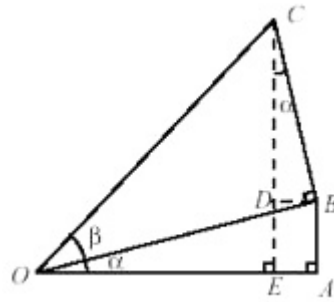
```
plot3d(sin(sqrt(x^2+y^2))/(sqrt(x^2+y^2)),x=-20..20,y=-20..20);
```



(20.26)

2.1. RELATIONS REMARQUABLES

Le dessin ci-dessous va nous permettre d'établir des relations qui permettront de résoudre des équations impliquant des fonctions trigonométriques (toutes ces relations sont de première importance en physique pour la simplification de la résolution de problèmes).



(20.27)

Nous noterons sur le schéma la relation suivante: $\overline{DE} = \overline{AB}$

Donc:

$$\begin{aligned} \sin(\alpha + \beta) &= \frac{\overline{CD} + \overline{DE}}{\overline{OC}} = \frac{\overline{CD}}{\overline{OC}} + \frac{\overline{DE}}{\overline{OC}} = \frac{\overline{CD}}{\overline{OC}} \cdot \frac{\overline{CB}}{\overline{CB}} + \frac{\overline{DE}}{\overline{OC}} \cdot \frac{\overline{OB}}{\overline{OB}} \\ &= \frac{\overline{CB}}{\overline{OC}} \cdot \frac{\overline{CD}}{\overline{CB}} + \frac{\overline{OB}}{\overline{OC}} \cdot \frac{\overline{AB}}{\overline{OB}} = \sin(\beta) \cdot \cos(\alpha) + \cos(\beta) \cdot \sin(\alpha) \end{aligned} \quad (20.28)$$

En résumé:

$$\sin(\alpha + \beta) = \sin(\alpha) \cdot \cos(\beta) + \cos(\alpha) \cdot \sin(\beta) \quad (20.29)$$

Ce qui implique trivialement si $\alpha = \beta$:

$$\sin(2\alpha) = 2\sin(\alpha) \cos(\alpha) \quad (20.30)$$

et:

$$\sin(\alpha - \beta) = \sin(\alpha + (-\beta)) = \sin(\alpha) \cdot \cos(-\beta) + \cos(\alpha) \cdot \sin(-\beta) \quad (20.31)$$

d'où:

$$\sin(\alpha - \beta) = \sin(\alpha) \cdot \cos(\beta) - \cos(\alpha) \cdot \sin(\beta) \quad (20.32)$$

Nous avons également:

$$\begin{aligned} \sin([\alpha + \beta] + \pi/2) &= \cos(\alpha + \beta) = \sin(\alpha) \cdot \cos(\beta + \pi/2) + \cos(\alpha) \cdot \sin(\beta + \pi/2) \\ &= \sin(\alpha) \cdot (-\sin \beta) + \cos(\alpha) \cdot \cos(\beta) \end{aligned} \quad (20.33)$$

d'où:

$$\cos(\alpha + \beta) = \cos(\alpha) \cdot \cos(\beta) - \sin(\alpha) \cdot \sin(\beta) \quad (20.34)$$

Ce qui implique trivialement si $\alpha = \beta$:

$$\cos(2\alpha) = \cos^2(\alpha) - \sin^2(\alpha) \quad (20.35)$$

Avec la relation déjà démontrée $\cos^2(\alpha) + \sin^2(\alpha) = 1$ nous obtenons également les relations très

importantes:

$$\cos(2\alpha) = \cos^2(\alpha) - \sin^2(\alpha) = 1 - 2\sin^2(\alpha) = 2\cos^2(\alpha) - 1 \quad (20.36)$$

Relations avec lesquelles nous obtenons très facilement les "formules de Carnot":

$$\sin^2(\alpha) = \frac{1 - \cos(2\alpha)}{2}, \cos^2(\alpha) = \frac{1 + \cos(2\alpha)}{2} \quad (20.37)$$

et:

$$\cos(\alpha - \beta) = \cos(\alpha + (-\beta)) = \cos(\alpha) \cdot \cos(-\beta) - \sin(\alpha) \cdot \sin(-\beta) \quad (20.38)$$

d'où:

$$\cos(\alpha - \beta) = \cos(\alpha) \cdot \cos(\beta) + \sin(\alpha) \cdot \sin(\beta) \quad (20.39)$$

Nous avons aussi:

$$\begin{aligned} \tan(\alpha + \beta) &= \frac{\sin(\alpha + \beta)}{\cos(\alpha + \beta)} = \frac{\sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)}{\cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)} \\ &= \frac{\frac{\sin(\alpha)\cos(\beta)}{\cos(\alpha)\cos(\beta)} + \frac{\cos(\alpha)\sin(\beta)}{\cos(\alpha)\cos(\beta)}}{\frac{\cos(\alpha)\cos(\beta)}{\cos(\alpha)\cos(\beta)} - \frac{\sin(\alpha)\sin(\beta)}{\cos(\alpha)\cos(\beta)}} = \frac{\tan(\alpha) + \tan(\beta)}{1 - \tan(\alpha)\tan(\beta)} \end{aligned} \quad (20.40)$$

Ceci, pour en arriver à la relation:

$$\tan(\alpha + \beta) = \frac{\tan(\alpha) + \tan(\beta)}{1 - \tan(\alpha) \cdot \tan(\beta)} \quad (20.41)$$

qui implique:

$$\tan(2\alpha) = \frac{2\tan(\alpha)}{1 - \tan^2(\alpha)} \quad (20.42)$$

et évidemment:

$$\tan(\alpha - \beta) = \tan(\alpha + (-\beta)) = \frac{\tan(\alpha) + \tan(-\beta)}{1 - \tan(\alpha) \cdot \tan(-\beta)} = \frac{\tan(\alpha) - \tan(\beta)}{1 + \tan(\alpha) \cdot \tan(\beta)} \quad (20.43)$$

d'où:

$$\tan(\alpha - \beta) = \frac{\tan(\alpha) - \tan(\beta)}{1 + \tan(\alpha) \cdot \tan(\beta)} \quad (20.44)$$

Nous obtenons également de manière triviale à partir des relations précédentes (nous faisons un petit mélange et nous secouons...):

$$\tan^2(\alpha) = 1 - \frac{2 \tan(\alpha)}{\tan(2\alpha)} = \frac{1 - \cos(2\alpha)}{1 + \cos(2\alpha)} \text{ et } \tan(\alpha) = \frac{1 - \cos(2\alpha)}{\sin(2\alpha)} = \frac{\sin(2\alpha)}{1 + \cos(2\alpha)} \quad (20.45)$$

Nous avons aussi:

$$\sin(2\alpha) = 2 \sin(\alpha) \cos(\alpha) = 2 \sin(\alpha) \frac{\cos(\alpha)}{\cos(\alpha)} \cos(\alpha) = 2 \tan(\alpha) \cos^2(\alpha) \quad (20.46)$$

avec:

$$\frac{1}{\cos^2(\alpha)} = \tan^2(\alpha) + 1 \quad (20.47)$$

d'où:

$$\sin(2\alpha) = \frac{2 \tan(\alpha)}{\tan^2(\alpha) + 1} \Leftrightarrow \sin(\alpha) = \frac{2 \tan\left(\frac{\alpha}{2}\right)}{\tan^2\left(\frac{\alpha}{2}\right) + 1} \quad (20.48)$$

de manière similaire nous obtenons:

$$\cos(2\alpha) = \cos^2(\alpha) - \sin^2(\alpha) = \cos^2(\alpha) (1 - \tan^2(\alpha)) \quad (20.49)$$

avec:

$$\frac{1}{\cos^2(\alpha)} = \tan^2(\alpha) + 1 \quad (20.50)$$

d'où:

$$\cos(2\alpha) = \frac{1 - \tan^2(\alpha)}{\tan^2(\alpha) + 1} \Leftrightarrow \cos(\alpha) = \frac{1 - \tan^2\left(\frac{\alpha}{2}\right)}{\tan^2\left(\frac{\alpha}{2}\right) + 1} \quad (20.51)$$

Déterminons maintenant les formules trigonométriques complémentaires appelées "**formules de Simpson**" ou "**formules d'addition**" qui permettent d'exprimer la somme de sinus et/ou de cosinus en produit de sinus et/ou cosinus.

Soient les relations déjà démontrées précédemment:

$$(1) \cos(\alpha + \beta) = \cos(\alpha) \cdot \cos(\beta) - \sin(\alpha) \cdot \sin(\beta)$$

$$(20.52)$$

$$(2) \cos(\alpha - \beta) = \cos(\alpha) \cdot \cos(\beta) + \sin(\alpha) \cdot \sin(\beta)$$

Posons $\alpha + \beta = p$ et $\alpha - \beta = q$ d'où:

$$\alpha = \frac{p+q}{2} \text{ et } \beta = \frac{p-q}{2} \quad (20.53)$$

Nous obtenons par sommation de (1) et (2):

$$\cos(p) + \cos(q) = 2 \cos\left(\frac{p+q}{2}\right) \cos\left(\frac{p-q}{2}\right) \quad (20.54)$$

et par différence:

$$\cos(p) - \cos(q) = -2 \sin\left(\frac{p+q}{2}\right) \sin\left(\frac{p-q}{2}\right) \quad (20.55)$$

De la même manière nous obtenons:

$$\sin(p) + \sin(q) = 2 \sin\left(\frac{p+q}{2}\right) \cos\left(\frac{p-q}{2}\right) \quad (20.56)$$

et par différence:

$$\sin(p) - \sin(q) = 2 \sin\left(\frac{p-q}{2}\right) \cos\left(\frac{p+q}{2}\right) \quad (20.57)$$

et inversement nous retombons très facilement sur les relations:

$$\begin{aligned} \cos(\alpha) \cos(\beta) &= \frac{1}{2} [\cos(\alpha + \beta) + \cos(\alpha - \beta)] \\ \sin(\alpha) \sin(\beta) &= \frac{1}{2} [\cos(\alpha - \beta) - \cos(\alpha + \beta)] \\ \sin(\alpha) \cos(\beta) &= \frac{1}{2} [\sin(\alpha + \beta) + \sin(\alpha - \beta)] \end{aligned} \quad (20.58)$$

Toutes ces relations nous seront utiles lors de notre étude de la physique générale et particulièrement dans le cas de calcul d'intégrales.

Remarque: Les relations suivantes démontrées précédemment:

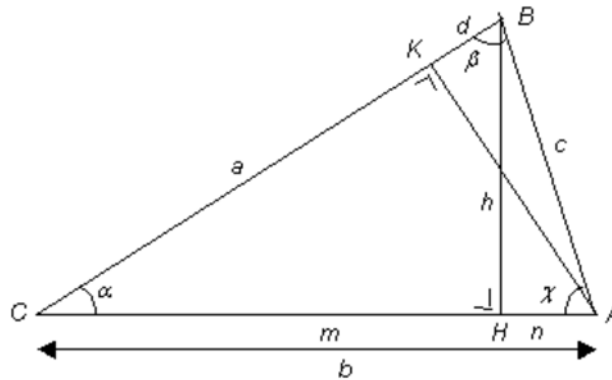
$$\begin{aligned} \sin p + \sin q &= 2 \sin \frac{p+q}{2} \cos \frac{p-q}{2} & \tan p + \tan q &= \frac{\sin(p+q)}{\cos p \cos q} \\ \sin p - \sin q &= 2 \sin \frac{p-q}{2} \cos \frac{p+q}{2} & & \\ \cos p + \cos q &= 2 \cos \frac{p+q}{2} \cos \frac{p-q}{2} & \tan p - \tan q &= \frac{\sin(p-q)}{\cos p \cos q} \\ \cos p - \cos q &= -2 \sin \frac{p+q}{2} \sin \frac{p-q}{2} & & \end{aligned} \quad (20.59)$$

sont souvent appelées "**formules de Simpson**".

2.1.1. THÉORÈME DU COSINUS

Démontrons encore le théorème du cosinus, utile en géométrie:

Dans un triangle quelconque, le carré de l'un des trois côtés est égal à la somme des carrés des deux autres diminuée du double produit de ces deux côtés par le cosinus de l'angle compris entre eux:



(20.60)

Démonstration:

$$\begin{aligned}
 a^2 &= h^2 + m^2 \\
 - \quad c^2 &= h^2 + n^2 \\
 \hline
 a^2 - c^2 &= m^2 - n^2 \\
 a^2 - c^2 &= (m+n)(m-n) \\
 a^2 - c^2 &= b(b-2n) = b^2 - 2nb
 \end{aligned}
 \tag{20.61}$$

mais dans le triangle ABH, rectangle en H, nous avons la relation $n/c = \cos(\chi)$ d'où:

$$a^2 - c^2 = b^2 - 2bc \cos(\chi) \tag{20.62}$$

Nous obtenons donc une des relations du "**théorème du cosinus**":

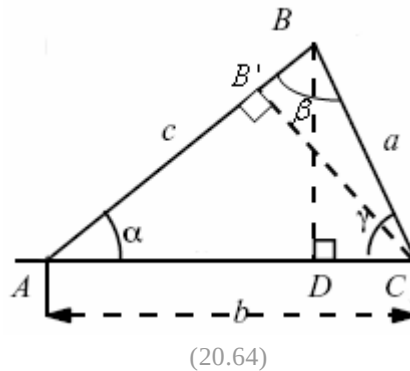
$$a^2 = b^2 + c^2 - 2bc \cos(\chi) \tag{20.63}$$

Par permutation circulaire, nous obtenons les deux autres relations connues.

Remarque: Le théorème du cosinus est parfois appelé "**formule d'Al-Kashi**"; par ailleurs si a est l'hypoténuse et son angle opposé un angle droit tel que $\cos(\chi)$ est nul, nous retrouvons donc le théorème de Pythagore. Voici pourquoi nous appelons parfois la formule d'Al-Kashi "**formule de Pythagore généralisée**".

2.1.2. THÉORÈME DU SINUS

Soit le triangle quelconque dont nous traçons deux hauteurs:



Dans le triangle ci-dessus nous avons les relations:

$$\sin(\alpha) = \frac{\overline{DB}}{c} \quad \sin(\gamma) = \frac{\overline{DB}}{a} \quad (20.65)$$

ce qui nous conduit à l'expression:

$$\overline{DB} = c \sin(\alpha) = a \sin(\gamma) \quad (20.66)$$

d'où:

$$\frac{a}{\sin(\alpha)} = \frac{c}{\sin(\gamma)} \quad (20.67)$$

Par un raisonnement similaire nous avons:

$$\sin(\alpha) = \frac{\overline{CB'}}{b} \quad \sin(\beta) = \frac{\overline{CB'}}{a} \quad (20.68)$$

Ce qui donne:

$$\frac{a}{\sin(\alpha)} = \frac{b}{\sin(\beta)} \quad (20.69)$$

Le tout combiné nous fournit le "**théorème des sinus**" dont le plus bel exemple d'application sur ce site est certainement la détermination des points de Lagrange L4 et L5 dans le chapitre d'astronomie:

$$\frac{a}{\sin(\alpha)} = \frac{b}{\sin(\beta)} = \frac{c}{\sin(\gamma)} \quad (20.70)$$

Evidemment, il n'y a pas ici toutes les relations trigonométriques (du cercle) existantes comme nous l'avons déjà dit, mais au moins les plus importantes qu'il faut savoir retrouver lors de l'étude de systèmes physiques.

3. TRIGONOMETRIE HYPERBOLIQUE

Nous avons démontré dans le chapitre d'Analyse Fonctionnelle que toute fonction $f(x)$ peut se décomposer en une fonction paire et impaire telle que:

$$f(x) = \frac{f(x) - f(-x)}{2} + \frac{f(x) + f(-x)}{2} \quad (20.71)$$

Ainsi, pour la fonction $f(x) = e^x$, nous obtenons:

$$e^x = \frac{e^x - e^{-x}}{2} + \frac{e^x + e^{-x}}{2} \quad (20.72)$$

Rappelons que lors de notre étude des nombres complexes nous avons démontré que:

$$\cos(x) = \frac{e^{ix} + e^{-ix}}{2} \quad \sin(x) = \frac{e^{ix} - e^{-ix}}{2i} \quad (20.73)$$

Nous définissons alors par analogie le sinus et le cosinus hyperbolique (nous démontrerons la provenance de ce terme plus loin) par:

$$\left| \cosh(x) = \frac{e^x + e^{-x}}{2} \quad \sinh(x) = \frac{e^x - e^{-x}}{2} \right| \quad (20.74)$$

et nous pouvons donc écrire:

$$e^x = \sinh(x) + \cosh(x) \quad (20.75)$$

Relation que nous pouvons à nouveau mettre en analogie avec:

$$e^{ix} = \cos(x) + i \sin(x) \quad (20.76)$$

Chose intéressante, nous pouvons travailler en trigonométrie avec des angles complexes. Effectivement, si nous posons $x = a + ib$, nous avons alors:

$$e^{a+ib} = \sinh(a + ib) + \cosh(a + ib) \quad (20.77)$$

Or:

$$e^{a+ib} = e^a e^{ib} = r e^{ib} = r(\cos(b) + i \sin(b)) \quad (20.78)$$

Donc:

$$e^a(\cos(b) + i \sin(b)) = \sinh(a + ib) + \cosh(a + ib) \quad (20.79)$$

Donc la fonction hyperbolique d'un angle complexe existe et l'image en est un nombre complexe aussi. Nous pouvons ainsi voir abusivement la géométrie hyperbolique comme une sorte de généralisation de la trigonométrie du cercle aux angles réels et complexes.

Par opposition à la trigonométrie du cercle, le lecteur remarquera et vérifiera facilement que nous

avons:

$$\cos^2(\alpha) + \sin^2(\alpha) = 1 \quad \cosh^2(\alpha) - \sinh^2(\alpha) = 1 \quad (20.80)$$

Démonstration:

$$\begin{aligned} \cosh^2(x) &= \left(\frac{e^x + e^{-x}}{2} \right)^2 = \frac{e^{2x} + 2e^x e^{-x} + e^{-2x}}{4} = \frac{1}{2} + \frac{e^{2x} + e^{-2x}}{4} \\ &= 1 - \frac{1}{2} + \frac{e^{2x} + e^{-2x}}{4} = 1 + \frac{e^{2x} - 2 + e^{-2x}}{4} = 1 + \frac{e^{2x} - 2e^x e^{-x} + e^{-2x}}{4} = 1 + \left(\frac{e^x - e^{-x}}{2} \right)^2 \\ &= 1 + \sinh^2(x) \end{aligned}$$

Nous avons donc:

$$\boxed{\cosh^2(x) - \sinh^2(x) = 1}$$

□C.Q.F.D.

Recherchons maintenant les fonctions réciproques des fonctions sinus et cosinus hyperboliques (que nous utiliserons parfois en physique ou en mécanique). Pour cela rappelons que:

$$y = \cosh(x) = \frac{e^x + e^{-x}}{2} \quad (20.81)$$

et que la recherche de la fonction réciproque consiste toujours à isoler x .

Donc:

$$2y - e^{-x} - e^x = 2 \cosh(x) - e^{-x} - e^x = 0 \quad (20.82)$$

c'est-à-dire:

$$(e^x)^2 - 2ye^x + 1 = 0 \quad (20.83)$$

en résolvant ce polynôme du deuxième degré en e^x puis en prenant le logarithme nous obtenons:

$$e_{1,2}^x = \frac{2y \pm \sqrt{4y^2 - 4}}{2} = y \pm \sqrt{y^2 - 1} \quad (20.84)$$

Or comme $e^x > 0$ nous devons rejeter la solution avec le signe "-". Il vient alors:

$$e^x = y + \sqrt{y^2 - 1} \quad (20.85)$$

d'où:

$$x = \ln \left(y + \sqrt{y^2 - 1} \right) \quad (20.86)$$

En procédant de même pour:

$$y = \sinh(x) = \frac{e^x - e^{-x}}{2} \quad (20.87)$$

Donc:

$$2y + e^{-x} - e^x = 2 \sinh(x) + e^{-x} - e^x = 0 \quad (20.88)$$

c'est-à-dire:

$$\left(e^x \right)^2 - 2y e^x - 1 = 0 \quad (20.89)$$

en résolvant ce polynôme du deuxième degré en e^x puis en prenant le logarithme nous obtenons:

$$e_{1,2}^x = \frac{2y \pm \sqrt{4y^2 + 4}}{2} = y \pm \sqrt{y^2 + 1} \quad (20.90)$$

Or comme $e^x > 0$ nous devons rejeter la solution avec le signe "-". Il vient alors:

$$e^x = y + \sqrt{y^2 + 1} \quad (20.91)$$

d'où:

$$x = \ln \left(y + \sqrt{y^2 + 1} \right) \quad (20.92)$$

Ainsi:

$$\boxed{\begin{aligned} \operatorname{arccosh}(x) &= \ln \left(x + \sqrt{x^2 - 1} \right) \\ \operatorname{arcsinh}(x) &= \ln \left(x + \sqrt{x^2 + 1} \right) \end{aligned}} \quad (20.93)$$

Attention car suivant les auteurs $\operatorname{arccosh}$, $\operatorname{arcsinh}$ se notent $\operatorname{argcosh}$, $\operatorname{argsinh}$ ou encore $\cosh^{-1}$, $\sinh^{-1}$. (cette dernière écriture pouvant prêter à confusion avec l'inverse la fonction hyperbolique correspondante!).

Pour étudier une représentation géométrique simple posons maintenant:

$$\cos(\alpha) = X, \sin(\alpha) = Y, \cosh(\alpha) = X', \sinh(\alpha) = Y' \quad (20.94)$$

avec une restriction à $\alpha \in \mathbb{R}$ et donc:

$$X = [-1, 1], Y = [-1, 1], X' = [1, +\infty], Y' = [-\infty, +\infty] \quad (20.95)$$

Donc nous pouvons écrire:

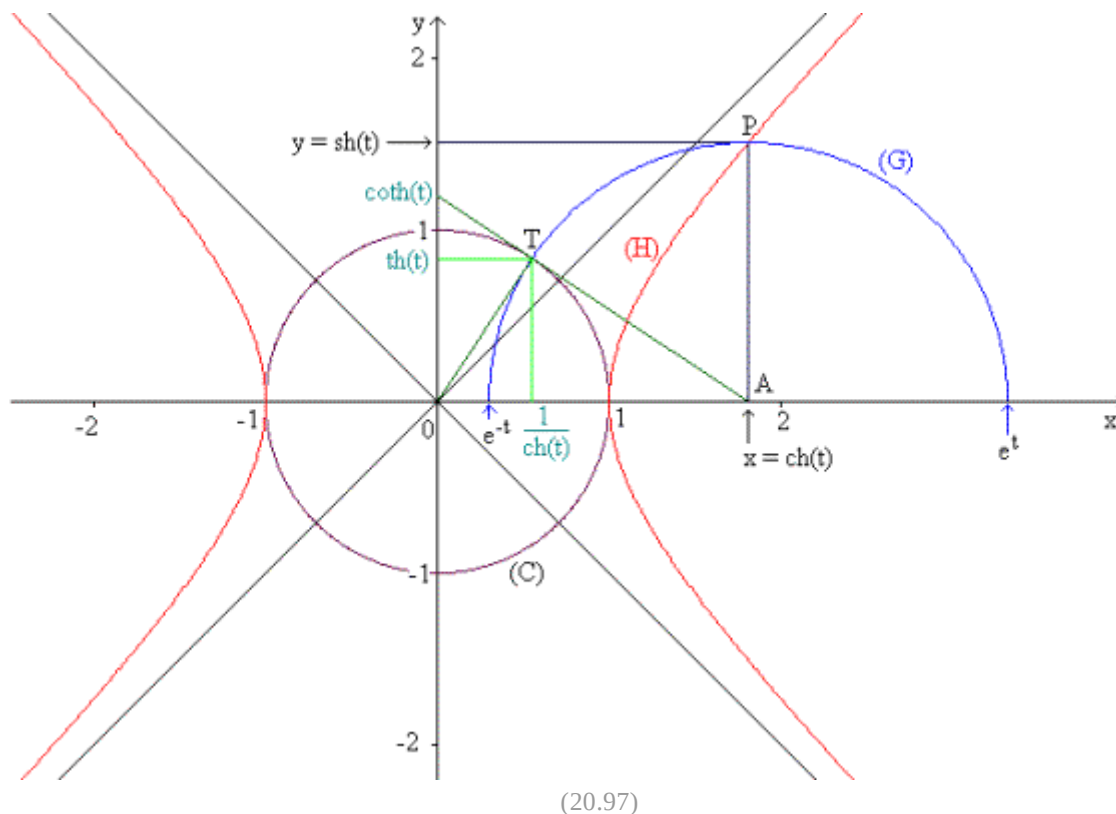
$$X^2 + Y^2 = 1 \quad X^2 - Y^2 = 1 \quad (20.96)$$

Or, comme nous le verrons lors de notre étude des coniques dans le chapitre de Géométrie Analytique:

1. La première de ces deux relations, constitue pour l'ensemble de définition donné, un cercle de rayon unité centré à l'origine. Le lecteur remarquera qu'il est assez curieux pour la trigonométrie du cercle d'obtenir un cercle...
2. La deuxième de ces deux relations, constitue pour l'ensemble de définition donné, une hyperbole équilatérale orientée selon l'axe X dont le sommet est $S(1,0)$ et de foyer $F(\sqrt{2},0)$. Le lecteur remarquera à nouveau qu'il est assez curieux pour la trigonométrie hyperbolique d'obtenir une hyperbole...

Ces deux dernières constatations devraient permettre, nous l'espérons, au lecteur de comprendre l'origine du nom de la trigonométrie hyperbolique et de constater que l'étude la trigonométrie hyperbolique sur l'hyperbole est l'analogue de l'étude de la trigonométrie du cercle sur le cercle.

Si nous représentons le cercle trigonométrique et l'hyperbole trigonométrique et rajoutons quelques informations complémentaires, voici ce que nous obtenons:



Explications:

Pour tracer à la règle et au compas le point $P(x,y)$ de l'hyperbole, nous nous donnons x , donc le point $A(x,0)$. Nous traçons la tangente au cercle (C) qui passe par $A(x,0)$ ce qui nous donne le point de tangence T . Nous traçons le cercle (G) de centre $A(x,0)$ et passant par T . Ce cercle coupe l'hyperbole au point $P(x,y)$ à la perpendiculaire en $A(x,0)$ à Ox .

Nous voyons apparaître sur la figure plusieurs valeurs des fonctions hyperboliques correspondant à $x = \cosh(t), y = \sinh(t)$ mais aussi $\tanh(t), \coth(t)$ etc. Entre autres, le cercle (G) coupe l'axe Ox en deux points dont les abscisses sont e^t et e^{-t} .

Si le lecteur veut s'en assurer au moyen de la figure, il pourra contrôler qu'en tout point de l'hyperbole, les relations (entre autres):

$$\cosh(t) = \frac{e^t + e^{-t}}{2} \quad \sinh(t) = \frac{e^t - e^{-t}}{2} \quad \cosh^2(t) - \sinh^2(t) = 1 \quad (20.98)$$

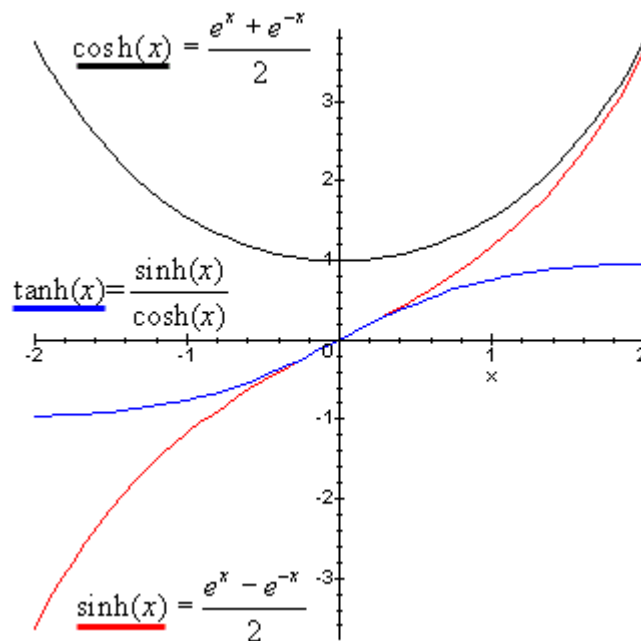
sont toujours vérifiées.

Si nous traçons maintenant sur un graphique:

$$\sinh(x) = \frac{e^x - e^{-x}}{2} \quad \cosh(x) = \frac{e^x + e^{-x}}{2} \quad \tanh(x) = \frac{\sinh(x)}{\cosh(x)} \quad (20.99)$$

Nous obtenons (ça c'est juste pour avoir vu une fois à quoi ressemblent ces fonctions) avec Maple:

```
>plot([sinh(x),cosh(x),tanh(x)],x=-2..2,color=[red,black,blue]);
```



(20.100)

Nous retrouverons la fonction $\cosh(x)$ dans le chapitre de Génie Civil par exemple dans le cadre des câbles suspendus. Nous retrouverons aussi les fonctions $\sinh(x)$ et $\tanh(x)$ dans le cadre de l'étude des vagues de gravité dans le chapitre de Génie Marin Et Météo.

3.1. RELATIONS REMARQUABLES

Soit par définition:

$$\cosh(x) = \frac{e^x + e^{-x}}{2} \quad \sinh(x) = \frac{e^x - e^{-x}}{2} \quad (20.101)$$

et:

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1} \quad \coth(x) = \frac{1}{\tanh(x)} = \frac{\sinh(x)}{\cosh(x)} = \frac{e^{2x} + 1}{e^{2x} - 1} \quad (20.102)$$

A partir de ces définitions et à l'aide des opérations élémentaires d'algèbre nous pouvons déterminer les relations remarquables suivantes (c'est beaucoup plus facile que la détermination de relations remarquables de la trigonométrie du cercle, donc sauf demande nous donnons ces relations sans démonstration):

$$\begin{aligned} \cosh(x) + \sinh(x) &= e^x & \cosh(x) - \sinh(x) &= e^{-x} \\ \cosh(-x) &= \cosh(x) & \sinh(-x) &= -\sinh(x) \\ \tanh(-x) &= -\tanh(x) & \coth(-x) &= -\coth(x) \end{aligned} \quad (20.103)$$

Egalement:

$$\cosh^2(x) = \frac{1}{1 - \tanh^2(x)} \quad \sinh^2(x) = \frac{\tanh^2(x)}{1 - \tanh^2(x)} \quad (20.104)$$

Et nous avons les relations d'addition:

$$\begin{aligned} \cosh(a+b) &= \cosh(a)\cosh(b) + \sinh(a)\sinh(b) \\ \cosh(a-b) &= \cosh(a)\cosh(b) - \sinh(a)\sinh(b) \\ \sinh(a+b) &= \sinh(a)\cosh(b) + \cosh(a)\sinh(b) \\ \sinh(a-b) &= \sinh(a)\cosh(b) - \cosh(a)\sinh(b) \end{aligned} \quad (20.105)$$

Suite à la demande d'un étudiant, démontrons les première et troisième relations ci-dessus:

Pour la première:

$$\begin{aligned} \cosh(a)\cosh(b) + \sinh(a)\sinh(b) &= \frac{e^a + e^{-a}}{2} \frac{e^b + e^{-b}}{2} + \frac{e^a - e^{-a}}{2} \frac{e^b - e^{-b}}{2} \\ &= \frac{1}{4} \left[(e^{a+b} + e^{a-b} + e^{-a+b} + e^{-a-b}) + (e^{a+b} - e^{a-b} - e^{-a+b} + e^{-a-b}) \right] \\ &= \frac{1}{4} (2e^{a+b} + 2e^{-a-b}) = \frac{1}{2} (e^{a+b} + e^{-a-b}) = \cosh(a+b) \end{aligned} \quad (20.106)$$

et la troisième:

$$\begin{aligned} \sinh(a)\cosh(b) + \cosh(a)\sinh(b) &= \frac{e^a - e^{-a}}{2} \frac{e^b + e^{-b}}{2} + \frac{e^a + e^{-a}}{2} \frac{e^b - e^{-b}}{2} \\ &= \frac{1}{4} \left[(e^{a+b} + e^{a-b} - e^{-a+b} - e^{-a-b}) + (e^{a+b} - e^{a-b} + e^{-a+b} - e^{-a-b}) \right] \\ &= \frac{1}{4} (2e^{a+b} - 2e^{-a-b}) = \frac{1}{2} (e^{a+b} - e^{-a-b}) = \sinh(a+b) \end{aligned} \quad (20.107)$$

Signalons encore d'autres relations remarquables:

$$\begin{aligned}
\cosh(p) + \cosh(q) &= 2 \cosh\left(\frac{p+q}{2}\right) \cosh\left(\frac{p-q}{2}\right) \\
\cosh(p) - \cosh(q) &= 2 \sinh\left(\frac{p+q}{2}\right) \sinh\left(\frac{p-q}{2}\right) \\
\sinh(p) + \sinh(q) &= 2 \sinh\left(\frac{p+q}{2}\right) \cosh\left(\frac{p-q}{2}\right) \\
\sinh(p) - \sinh(q) &= 2 \cosh\left(\frac{p+q}{2}\right) \sinh\left(\frac{p-q}{2}\right)
\end{aligned}
\tag{20.108}$$

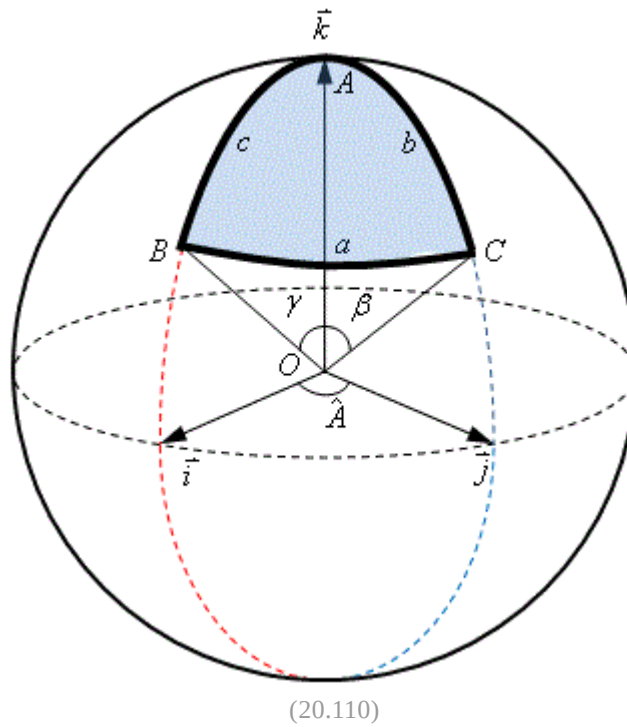
et encore:

$$\begin{aligned}
\tanh(a+b) &= \frac{\tanh(a) + \tanh(b)}{1 + \tanh(a)\tanh(b)} \\
\tanh(a-b) &= \frac{\tanh(a) - \tanh(b)}{1 - \tanh(a)\tanh(b)} \\
\tanh(p) + \tanh(q) &= \frac{\sinh(p+q)}{\cosh(p)\cosh(q)} \\
\tanh(p) - \tanh(q) &= \frac{\sinh(p-q)}{\cosh(p)\cosh(q)}
\end{aligned}
\tag{20.109}$$

4. TRIGONOMÉTRIE SPHÉRIQUE

L'objectif de la trigonométrie sphérique est de déterminer les relations remarquables existantes entre les angles et les côtés de formes projetées (dites également "**formes géodésiques**" car suivant la courbure de l'espace) sur la surface d'une sphère. Pour déterminer ces relations, nous allons nous intéresser au cas particulier d'une sphère de rayon unité et des relations entre les côtés d'un triangle (élément de surface plane élémentaire) et les différents angles existants. Nous verrons que les résultats sont au fait indépendants du rayon de la sphère et de la forme considérée initialement.

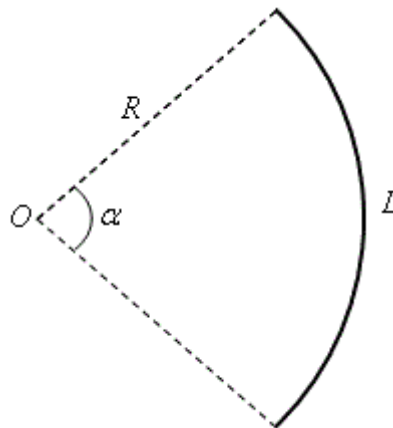
Soit la figure sur laquelle se trouve un triangle géodésique de sommets A, B, C d'angles d'ouverture respectifs $\hat{A}, \hat{B}, \hat{C}$ et de côtés opposés a, b, c et trois vecteurs $\vec{i}, \vec{j}, \vec{k}$ unitaires tels que $\vec{i} \perp \vec{k}, \vec{j} \perp \vec{k}$ et que l'extrémité de $\vec{k}$ soit confondue avec le sommet A :



L'angle entre les points B et C , noté α , n'a pas pu être représenté sur le schéma ci-dessus faute de place.

Rappelons que le périmètre d'un cercle de rayon unité sur la sphère de rayon unité vaut bien évidemment $P = 2\pi R$. Le périmètre du cercle en fonction de l'angle d'ouverture de ce dernier étant donné par (relation très très souvent utilisée en physique!!!):

$$L = 2\pi R \cdot \frac{\alpha}{2\pi} = R\alpha \quad (20.111)$$



Si le cercle est de rayon $R = 1$ (comme c'est le cas pour notre sphère), le calcul de la longueur d'arc se simplifie et devient:

$$L = \alpha \quad (20.112)$$

Nous garderons cette contrainte du rayon unité pour la suite afin de simplifier les expressions que nous allons obtenir par la suite.

Conséquence relativement aux points sur notre sphère; les côtés du triangle sont donnés par:

$$a = \alpha \quad b = \beta \quad c = \gamma \quad (20.113)$$

Considérons maintenant le produit scalaire (cf. chapitre de Calcul Vectoriel):

$$\overrightarrow{OB} \circ \overrightarrow{OC} = \|\overrightarrow{OB}\| \cdot \|\overrightarrow{OC}\| \cos(\alpha) \quad (20.114)$$

et comme $\|\overrightarrow{OB}\| = \|\overrightarrow{OC}\| = 1$ (rayon unité) nous avons:

$$\cos(\alpha) = \cos(a) = \overrightarrow{OB} \circ \overrightarrow{OC} \quad (20.115)$$

Si nous décomposons les deux vecteurs $\overrightarrow{OB}$ et $\overrightarrow{OC}$ sur les vecteurs tangents unités nous avons:

$$\begin{aligned} \overrightarrow{OB} &= \cos(\gamma) \vec{k} + \sin(\gamma) \vec{i} \\ \overrightarrow{OC} &= \cos(\beta) \vec{k} + \sin(\beta) \vec{j} \end{aligned} \quad (20.116)$$

Ce qui nous donne:

$$\cos(a) = [\cos(\gamma) \vec{k} + \sin(\gamma) \vec{i}] \circ [\cos(\beta) \vec{k} + \sin(\beta) \vec{j}] \quad (20.117)$$

ce qui donne (distributivité du produit scalaire):

$$\cos(a) = \cos(\gamma) \cos(\beta) \vec{k} \circ \vec{k} + \cos(\gamma) \sin(\beta) \vec{k} \circ \vec{j} + \sin(\gamma) \cos(\beta) \vec{i} \circ \vec{k} + \sin(\gamma) \sin(\beta) \vec{i} \circ \vec{j} \quad (20.118)$$

Comme $\vec{k} \circ \vec{k} = 1$ et $\vec{k} \circ \vec{i} = \vec{k} \circ \vec{j} = 0$, la relation précédente se réduit à:

$$\cos(a) = \cos(\gamma) \cos(\beta) + \sin(\gamma) \sin(\beta) \vec{i} \circ \vec{j} \quad (20.119)$$

et comme:

$$\vec{i} \circ \vec{j} = \cos(\hat{A}) \quad (20.120)$$

Nous avons:

$$\cos(a) = \cos(\gamma) \cos(\beta) + \sin(\gamma) \sin(\beta) \cos(\hat{A}) \quad (20.121)$$

relation dite "**relation fondamentale**" ou "**formule des cosinus**" que nous pouvons donc (de par le rayon unité) tout aussi bien écrire:

$$\boxed{\cos(a) = \cos(c) \cos(b) + \sin(c) \sin(b) \cos(\hat{A})} \quad (20.122)$$

Cette dernière relation est invariante par permutation circulaire des variables $a, c, b, \hat{A}$. Il est aussi intéressant de remarquer avant de continuer que si le triangle sphérique est à angle droit en A, la relation précédente se simplifie en:

$$\cos(a) = \cos(c) \cos(b) \quad (20.123)$$

et si le triangle est suffisamment petit par rapport au rayon et que nous faisons un développement de Taylor proche de 0 au deuxième ordre (cf. chapitre Suites et Séries) pour chacun des termes il vient:

$$1 - \frac{a^2}{2} = \left(1 - \frac{c^2}{2}\right) \left(1 - \frac{b^2}{2}\right) \quad (20.124)$$

soit:

$$1 - \frac{a^2}{2} = 1 - \frac{c^2}{2} - \frac{b^2}{2} + \underbrace{\frac{c^2 b^2}{2}}_{\rightarrow 0} \quad (20.125)$$

après simplification:

$$a^2 = b^2 + c^2 \quad (20.126)$$

nous retrouvons le théorème de Pythagore (cf. chapitre de Géométrie Euclidienne). Donc:

$$\cos(a) = \cos(c) \cos(b) \quad (20.127)$$

est le pendant en géométrie sphérique (géométrie non-euclidienne) du théorème de Pythagore de la géométrie plane (géométrie euclidienne).

Cette parenthèse fermée, revenons en à nos moutons. Les sinus de tous les angles étant positifs (puisque inférieurs à π), nous pouvons écrire:

$$\begin{aligned} \frac{\sin(\hat{A})}{\sin(a)} &= \frac{\sqrt{1 - \cos^2(\hat{A})}}{\sin(a)} = \frac{\sqrt{1 - \left(\frac{\cos(a) - \cos(b) \cos(c)}{\sin(b) \sin(c)}\right)^2}}{\sin(a)} \\ &= \frac{\sqrt{(1 - \cos^2(b))(1 - \cos^2(c)) - \cos^2(a) - \cos^2(b) \cos^2(c) + 2 \cos(a) \cos(b) \cos(c)}}{\sin(a) \sin(b) \sin(c)} \\ &= \frac{\sqrt{1 - \cos^2(a) - \cos^2(b) - \cos^2(c) + 2 \cos(a) \cos(b) \cos(c)}}{\sin(a) \sin(b) \sin(c)} \end{aligned} \quad (20.128)$$

Cette dernière relation est bien évidemment également invariante par permutation circulaire des variables $a, b, c, \hat{A}$. Donc nous obtenons une relation remarquable du triangle sphérique, appelée "relation des sinus" ou "formule des sinus":

$$\frac{\sin(\hat{A})}{\sin(a)} = \frac{\sin(\hat{B})}{\sin(b)} = \frac{\sin(\hat{C})}{\sin(c)} \quad (20.129)$$

Comme la trigonométrie sphérique est souvent utilisée pour des repérages terrestres, avec souvent 2 cercles très particuliers et orthogonaux: l'équateur terrestre et un méridien ou un parallèle quelconque, ce cas revêt un intérêt particulier. Le lecteur pourra s'exercer à retrouver les relations ci-dessous. Dans

le cas d'un triangle rectangle en A nous avons bien évidemment:

$$\frac{1}{\sin(a)} = \frac{\sin(\hat{B})}{\sin(b)} = \frac{\sin(\hat{C})}{\sin(c)} \quad (20.130)$$

Toutes les relations que nous avons déterminées jusqu'à maintenant nous permettent dans le cas où $\hat{A} = \pi/2$ et $R = 1$ de tirer des relations très intéressantes pour la géophysique:

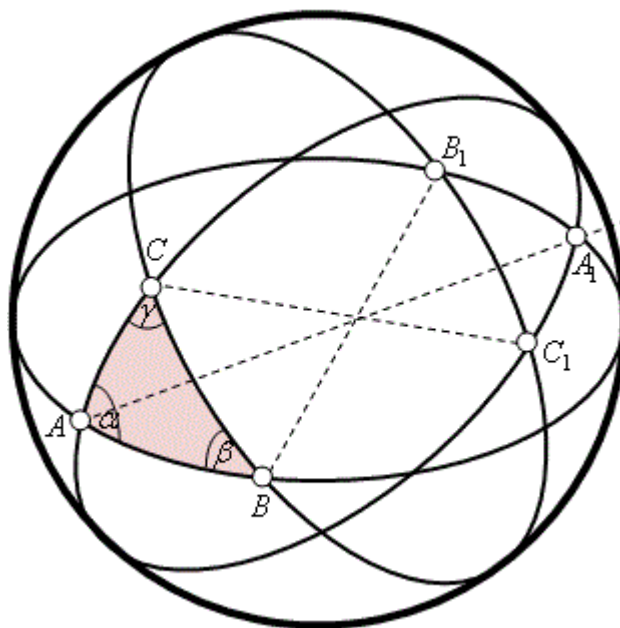
$$\begin{aligned} \sin(b) &= \sin(a) \sin(\hat{B}) \\ \sin(c) &= \sin(a) \sin(\hat{C}) \\ \cos(a) &= \cos(b) \cos(c) \quad (20.131) \\ \cos(\hat{C}) &= \sin(\hat{B}) \cos(c) \\ \cos(\hat{B}) &= \sin(\hat{C}) \cos(b) \end{aligned}$$

Evidemment, nous n'avons pas présenté ici toutes les relations de trigonométrie sphérique existantes, mais au moins les plus importantes qu'il faut savoir retrouver.

Remarque: Nous définissons "**l'excédent**" ou "**excès sphérique**" par le nombre:

$$E = \hat{A} + \hat{B} + \hat{C} - \pi \quad (20.132)$$

Pendant que nous y sommes, profitons-en pour calculer un problème classique qui est celui de la surface d'un triangle sur une sphère. Soit la figure:



(20.133)

Si nous prolongeons les arcs de géodésique AC et AB jusqu'à A_1 , nous obtenons une tranche de sphère

dont la surface S_1 est proportionnelle à l'angle α , en A. Si cet angle valait 2π , nous aurions toute la sphère et la surface vaudrait $4\pi R^2$. Comme l'angle vaut α , la proportionnalité nous dit que S_1 vaut:

$$S_1 = \frac{\alpha}{2\pi} 4\pi R^2 = \alpha 2R^2 \quad (20.134)$$

De la même manière, si nous prolongeons les arcs BC et BA jusqu'à B_1 et si nous prolongeons les arcs CA et CB jusqu'à C_1 , nous obtenons deux autres tranches dont les surfaces S_2 et S_3 valent:

$$S_2 = \beta 2R^2 \quad S_3 = \gamma 2R^2 \quad (20.135)$$

Supposons maintenant que nous additionnions ces trois surfaces:

$$S_1 + S_2 + S_3 = 2R^2 (\alpha + \beta + \gamma) \quad (20.136)$$

nous obtenons alors la moitié de la sphère $S_{\odot}$ (regarder la figure pour vous le représenter mentalement) plus 2 fois le triangle géodésique de surface S en rose sur la figure (car pris en compte 2 fois en trop).

Il faut enlever deux fois la surface de ce triangle bleu pour obtenir la surface de la demi-sphère:

$$\frac{S_{\odot}}{2} = 2R^2 (\alpha + \beta + \gamma) - 2S \quad (20.137)$$

Donc:

$$S_{\odot} = 4R^2 (\alpha + \beta + \gamma) - 4S \quad (20.138)$$

comme $S_{\odot} = 4\pi R^2$, nous avons:

$$S = \frac{4R^2 (\alpha + \beta + \gamma) - S_{\odot}}{4} = \frac{4R^2 (\alpha + \beta + \gamma) - 4\pi R^2}{4} \quad (20.139)$$

Après simplification nous en déduisons que la surface S du triangle ABC vaut::

$$S = R^2 (\alpha + \beta + \gamma - \pi) = R^2 \varepsilon \quad (20.140)$$

où ε est un angle solide.

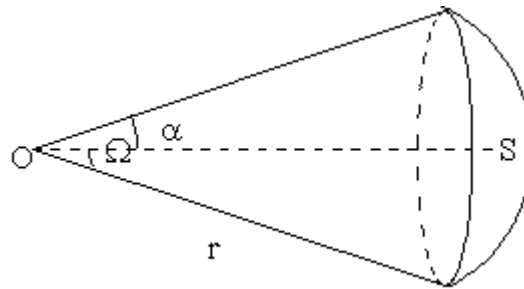
Il est assez simple de généraliser ce concept à d'autres formes du même acabit (en particulier celles composées de triangles...).

4.1. ANGLE SOLIDE

En géométrie spatiale, se pose le problème du concept d'angle d'ouverture d'une portion de l'espace (en extension à l'angle dit "angle plan"). Nous définissons alors "l'angle solide" Ω par la mesure de la portion d'espace limitée par une surface conique de sommet O et nous l'exprimons en stéradian, obtenue par le rapport:

$$\Omega = \frac{S}{r^2} \quad (20.141)$$

S étant l'aire de la calotte découpée par le cône sur une sphère de rayon r .



(20.142)

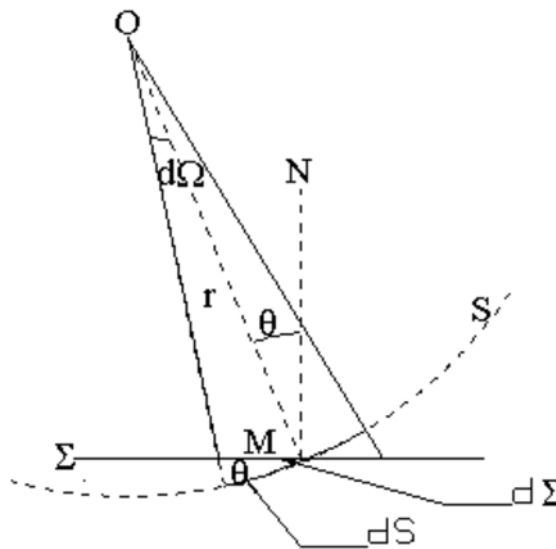
Si α est le demi-angle du cône, nous obtenons pour ce rapport (pour le calcul de la calotte d'une surface sphérique voir le chapitre traitant des formes géométriques):

$$\Omega = 2\pi(1 - \cos \alpha) \quad (20.143)$$

D'où l'on conclut que l'angle solide total vaut par définition:

$$\Omega_{\text{tot}} = 4\pi \quad (20.144)$$

Nous pouvons également calculer "l'angle solide élémentaire" tel que représenté ci-dessous:



(20.145)

Soit un angle solide élémentaire $d\Omega$ et OM l'axe du cône. Nous posons:

$$OM = r \quad (20.146)$$

Nous considérons une surface quelconque Σ passant par le point M . $d\Omega$ découpe sur cette surface une portion $d\Sigma$.

Si nous traçons la sphère S de centre O et de rayon r , cet angle solide découpe sur cette sphère une calotte d'aire dS :

$$dS = r^2 \cdot d\Omega \quad (20.147)$$

Soit MN la normale à $d\Sigma$ qui fait un angle θ avec OM . Nous avons, en assimilant dS et $d\Sigma$ à des portions de plan:

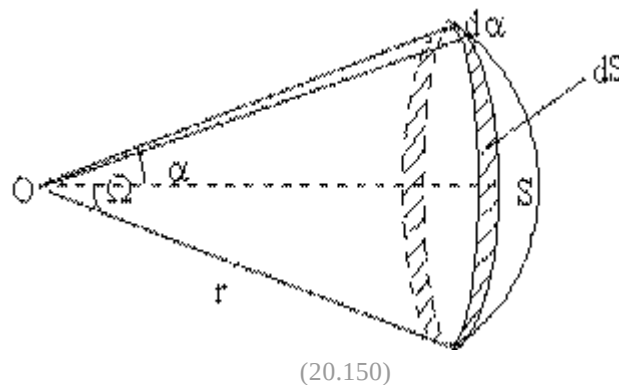
$$dS = \cos \theta d\Sigma \quad (20.148)$$

d'où:

$$d\Omega = \frac{d\Sigma \cos \theta}{r^2} \quad (20.149)$$

Ce concept d'angle solide nous sera très utile en particulier dans le domaine de la physique théorique qui traite du rayonnement thermique (cf. chapitres d'Optique et de Thermodynamique).

Nous pouvons encore calculer à partir des concepts précédents, l'angle solide élémentaire de révolution tel que présenté sur la figure ci-dessous:



Il est compris entre deux angles solides de révolution dont les demi-angles au sommet diffèrent de $d\alpha$.

$$d\Omega = d(2\pi(1 - \cos \alpha)) = 2\pi \sin \alpha d\alpha \quad (20.151)$$

où:

$$d\Omega = \frac{dS}{r^2} \quad (20.152)$$

Démonstration:

Dans le chapitre traitant des Formes Géométriques (cf. section de Géométrie) nous avons démontré les différentes manières de calculer la surface d'une sphère. De ces calculs il avait été déduit que la surface élémentaire à R constant était:

$$dS = r^2 \sin \theta d\theta d\phi \quad (20.153)$$

et puisque:

$$d\Omega = \frac{dS}{r^2} \quad (20.154)$$

l'angle solide élémentaire s'écrit alors:

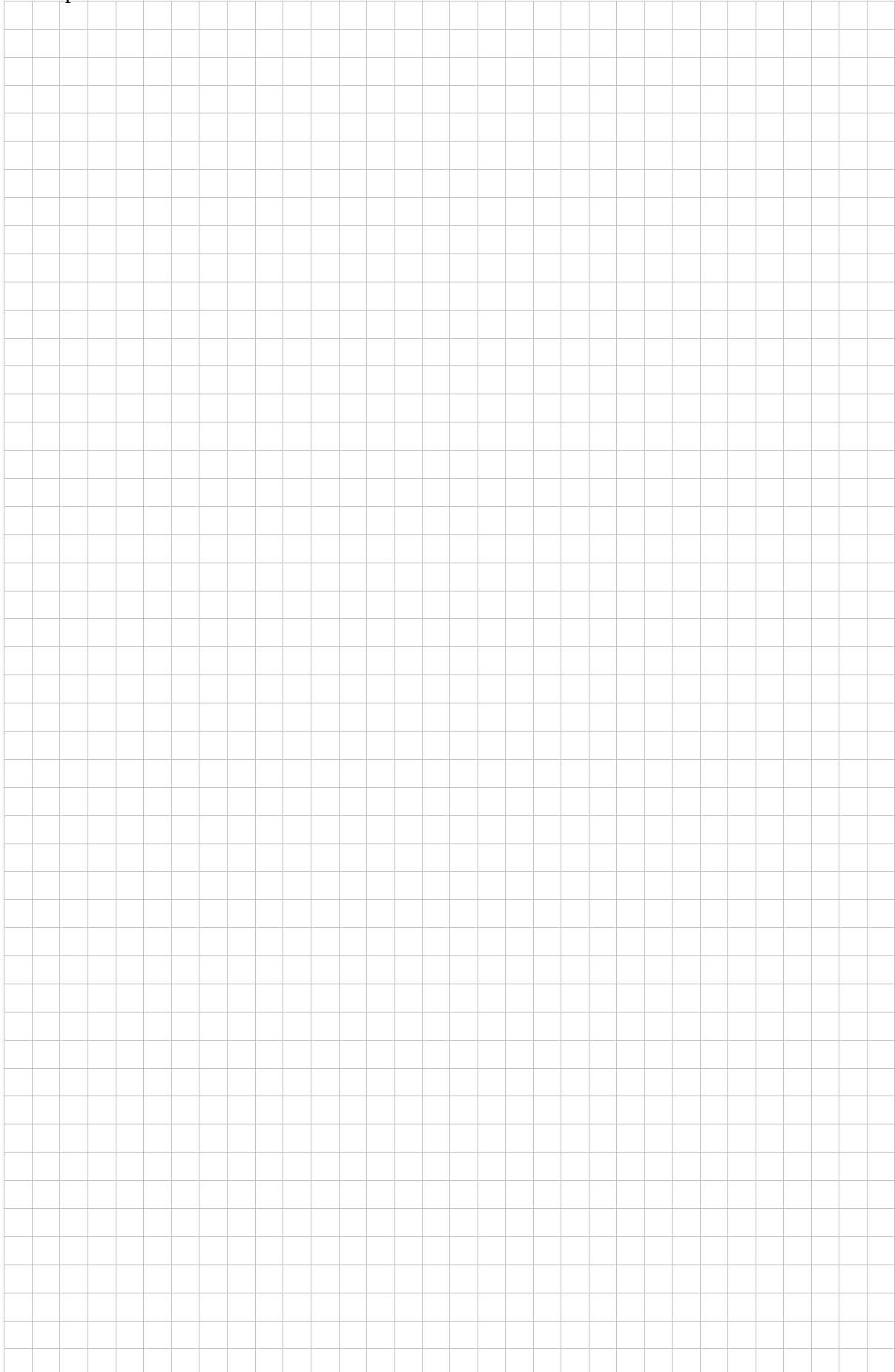
$$d\Omega = \sin \theta d\theta d\phi \quad (20.155)$$

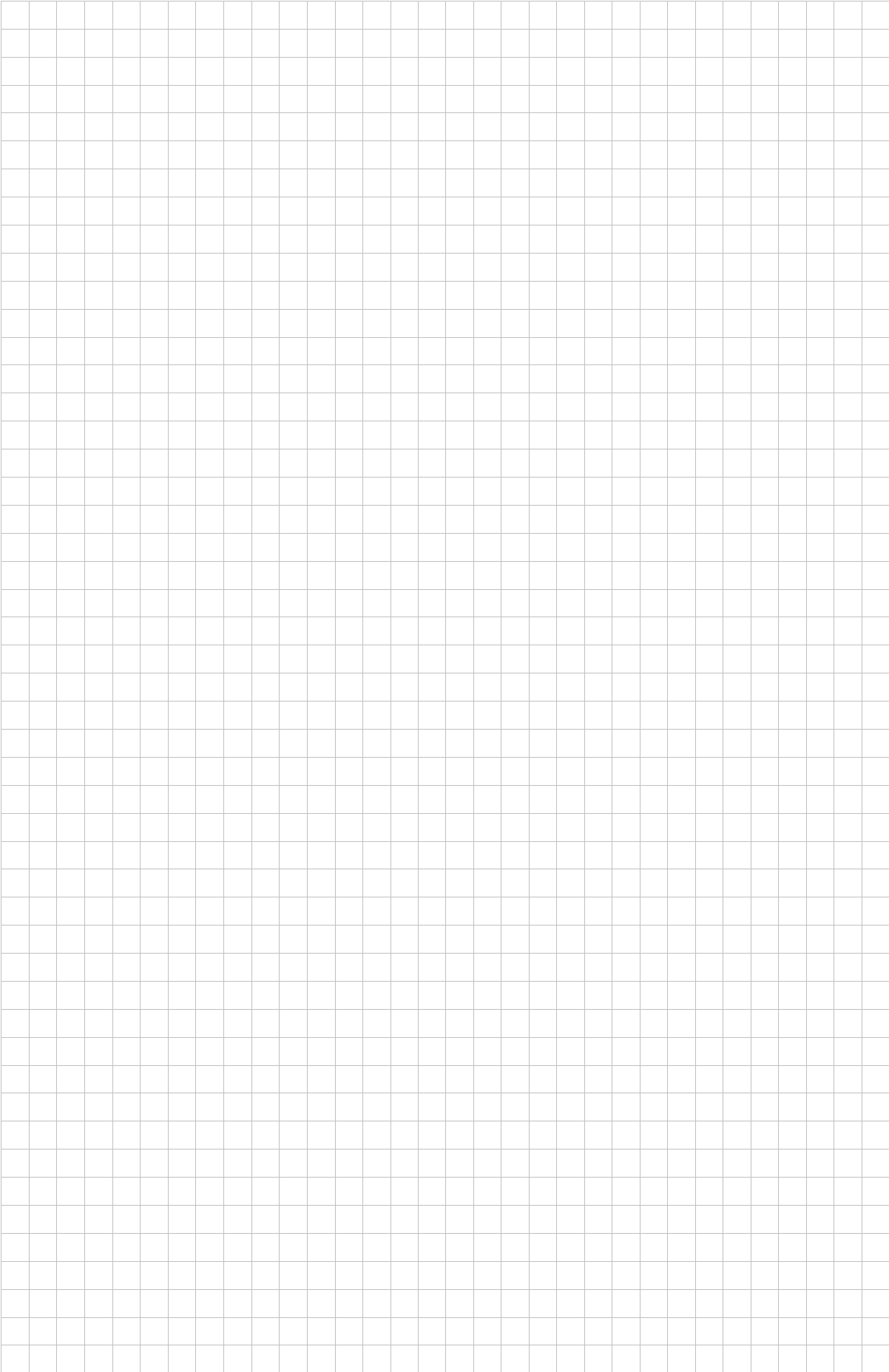
Ainsi, l'angle solide délimité par un cône de révolution, d'angle au sommet α vaut :

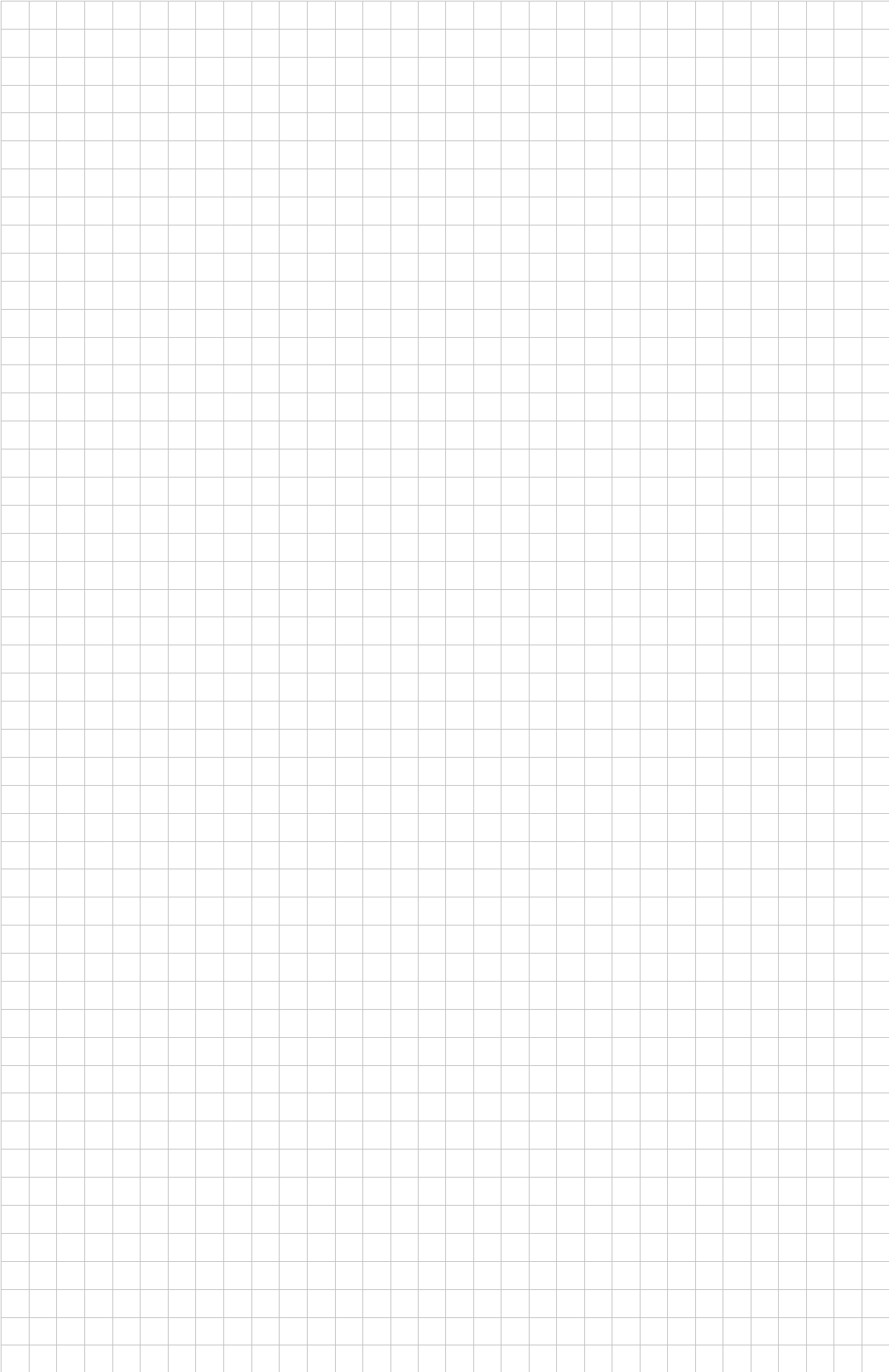
$$\begin{aligned} \Omega &= \int d\Omega = \int_0^{2\pi} d\phi \int_0^\alpha \sin \theta d\theta = \phi \Big|_0^{2\pi} \cdot (-\cos \theta) \Big|_0^\alpha = 2\pi(-\cos(\alpha) - (-\cos(0))) \\ &= 2\pi(-\cos(\alpha) + 1) = 2\pi(1 - \cos \alpha) \end{aligned} \quad (20.156)$$

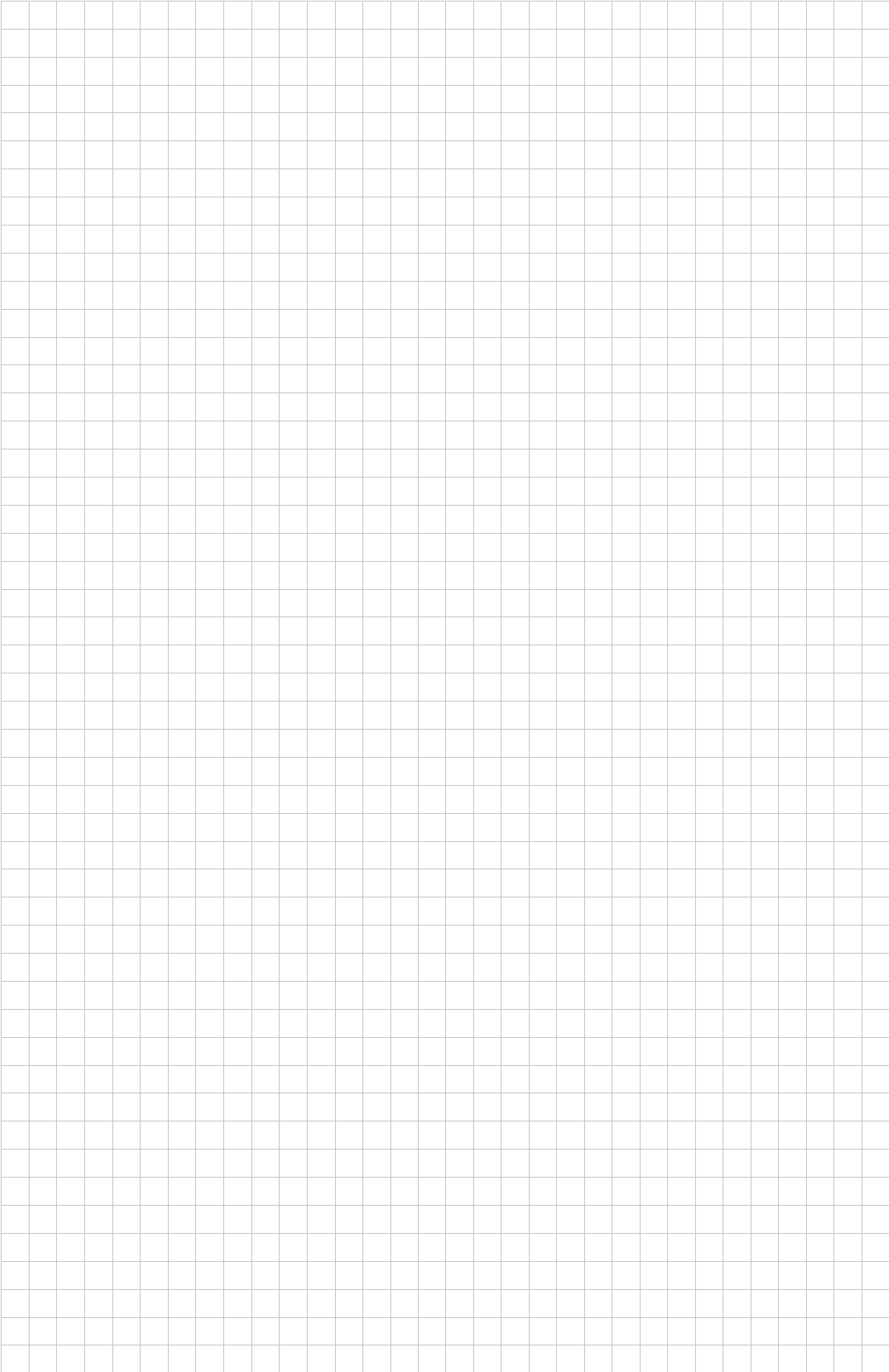
□C.Q.F.D.

Notes personnelles:









21. GÉOMÉTRIE EUCLIDIENNE

L'objet de la "géométrie euclidienne" (appelée plus communément "géométrie plane") est, en principe, l'étude des formes et des propriétés des corps naturels. La géométrie n'est cependant pas une science expérimentale, puisque son objet est, non pas d'étudier certains aspects de la nature, mais une reproduction nécessairement arbitraire de celle-ci.

Nous allons dans ce chapitre présenter implicitement, dans un premier temps, les cinq postulats de la géométrie euclidienne (dont les quatre premiers sont considérés aujourd'hui comme des axiomes) et ensuite développer autour de ceux-ci la géométrie de base que le lecteur aura besoin pour l'étude du reste du site. Une fois ceci fait, nous résumerons notre étude en présentant de manière explicite les cinq postulats d'Euclide et ensuite les axiomes de Hilbert.

Remarque: Nous avons tenté de préserver au mieux les notations propres à Euclide en ayant toutefois une approche plus moderne de certains concepts et de présenter uniquement ceux qui sont utiles à l'ingénieur sur le marché du travail.

OBJETS DE LA GÉOMÉTRIE EUCLIDIENNE

Avant d'énoncer les cinq postulats, il nous semble bon de définir quelques concepts intuitifs au préalable:

D1. La notion expérimentale la plus simple est celle de "**volume**". Nous disons qu'un corps occupe un certain volume lorsqu'il occupe dans l'espace à trois dimensions une certaine place (pour des espaces à des dimensions supérieures, nous parlons d'hyper-volumes).

D2. Nous admettons comme une chose évidente qu'un volume est limité par une "**surface**"; mais si l'existence du volume est physiquement contrôlable et mesurable, la surface est une création de l'esprit; c'est quelque chose d'analogue à une baudruche, par exemple, enveloppant un volume quelconque, mais d'analogue seulement. C'est un être géométrique à deux dimensions sans épaisseur.

D3. Lorsqu'une surface est limitée, cette limite est une "**ligne**". Ici encore, la ligne est une création de l'esprit, une ligne n'a pas d'existence expérimentale; c'est quelque chose d'analogue à la figure formée par un fil de fer. Etre géométrique encore mais d'une dimension sans hauteur ni largeur.

D4. Une "**droite**" est définie comme la ligne de plus court chemin joignant deux points sur une surface.

D5. Quand une ligne est limitée, sa limite est un "**point**": le point est quelque chose d'analogue à l'intersection de deux fils tendus. C'est encore une création de l'esprit, un être géométrique.

Remarque: Il est d'usage, en géométrie, de représenter un point par une lettre $A, B, \dots$; une ligne, ou une surface par une lettre entre parenthèses (mais cela est rarement respecté car nous supposons souvent que le lecteur sait de quoi nous parlons). Nous disons alors, par exemple: la ligne (L), la surface (S).

D6. L'expression le "**segment**" AB désigne en général une ligne limitée par les points A et B . Nous dirons qu'un point M est sur le segment AB , pour traduire le fait suivant: tout segment AB peut être séparé d'une infinité de façons en deux morceaux limités par A et M d'une part, par M et B d'autre

part - fait inspiré au géomètre par la possibilité de couper expérimentalement un bout de fil de fer en deux, et cela d'une infinité de façons (nous y reviendrons lors plus loin)

Remarque: L'expression: "la ligne (L) est tracée sur une surface (S)" signifie que la surface (S) pourrait être divisée en plusieurs morceaux, de manière que la ligne (L) soit la frontière ou une partie de frontière d'un de ces morceaux. Cette définition est inspirée du fait qu'il est possible de découper une étoffe, par exemple, en suivant avec des ciseaux un trait quelconque tracé sur cette étoffe.

Lorsqu'une ligne (L) est tracée sur une surface (S), tout point M qui est situé sur la ligne (L) est aussi, par définition, situé sur la surface (S). Nous disons alors que c'est un "**point de cette surface**".

D7. Nous appelons "**angle**" (ou "**angle plan**") ou plus rigoureusement "**angle rectiligne**" la portion de plan limitée par deux demi-droites (voir plus loin la définition d'une demi-droite)

DIMENSIONS

Nous avons parlé de volume, surface et de ligne auxquelles nous pouvons associer des dimensions. Mais qu'est-ce une dimension au fait ? Nous allons tâcher d'essayer de définir au mieux cette dernière mais d'abord, il faut savoir qu'il existe en géométrie plusieurs types de dimensions. La plus connue et commune est celle que nous appelons la "**dimension topologique**".

Par exemple, le point (abstraction mathématique et géométrique) a une dimension topologique de 0, la courbe (trait continu d'épaisseur nulle) une dimension de 1, la surface une dimension de 2, un volume une dimension de 3 et un hyper-volume une dimension 4 (pour représenter un hyper-volume, prenez un volume dessiné sur papier (...) et faites-en une translation et reliez les sommets). Ce sont toutes des valeurs entières par définition:

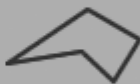
Objet	Représentation	Dimension
un point	•	0
une ligne (droite, courbe...)		1
une figure plane (quadrilatère, cercle...)		2
un solide (parallélépipède, sphère...)		3

Tableau: 21.1 - Objets, représentations et dimensions types

Pour calculer la dimension de certains objets, nous allons utiliser la méthode de la géométrie métrique plane qui consiste à prendre un étalon de cet objet, c'est-à-dire cet objet lui-même mais en plus petit, et à le reporter sur notre objet un certain nombre de fois:

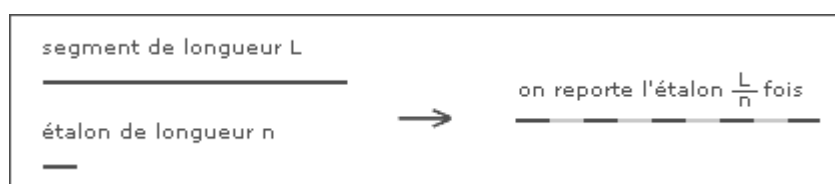


Figure: 21.1 - Concept d'étalon unidimensionnel

Soit L la longueur totale du segment. Nous allons prendre un étalon de longueur n que nous allons

reporter sur le segment. Cet étalon sera reporté L/n fois. Nous remarquons que:

$$\frac{L}{n} = \left(\frac{L}{n} \right)^1 \quad (21.1)$$

Nous pouvons appliquer le même raisonnement à une surface:

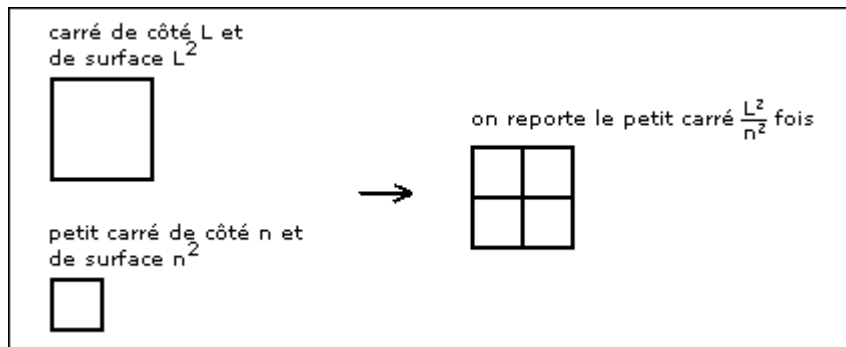


Figure: 21.2 - Concept d'étalon bidimensionnel

Soit L^2 la surface totale du carré. Cette fois, nous prenons un autre carré, plus petit, de côté n et de surface n^2 . Nous reportons le petit carré sur le grand L^2/n^2 fois pour obtenir la surface du grand carré. Nous remarquons que:

$$\frac{L^2}{n^2} = \left(\frac{L}{n} \right)^2 \quad (21.2)$$

Au travers de ces deux exemples, nous avons fait apparaître le nombre 1 pour le segment, et le nombre 2 pour le carré, l'idée est la même avec un volume mais en utilisant un cube. Ces nombres sont la "**dimension**" de l'objet.

Ainsi, pour résumer:

- Les lignes sont de dimension 1, car pour mesurer une longueur avec une division plus fine d'un facteur n d'un segment étalon, le nombre de subdivisions sera multiplié par le même facteur. Donc il y a une relation de puissance unitaire entre la subdivision et la mesure.
- Les surfaces sont de dimensions 2, car pour mesurer une surface avec une division plus fine d'un facteur n d'un carré étalon, le nombre de subdivisions sera donné par la puissance carrée de n . Donc il y a une relation de puissance deux entre la subdivision et la mesure (si on prend des carrés deux fois plus petits pour couvrir une surface, il en faudra quatre fois plus).
- Les volumes sont de dimension 3, car pour mesurer un volume avec une division plus fine d'un facteur n d'un cube étalon, le nombre de subdivisions sera donné par la puissance cubique de n . Donc il y a une relation de puissance trois entre la subdivision et la mesure (si on prend des cubes deux fois plus petits pour remplir un volume, il en faudra 8 fois plus).

Généralisons: soit N le nombre de fois que nous reportons l'étalon de longueur n sur notre objet de longueur L , et soit d la dimension de l'objet, nous avons:

$$N = \left(\frac{L}{n} \right)^d \Leftrightarrow d = \frac{\ln(N)}{\ln\left(\frac{L}{n}\right)} \quad (21.3)$$

Dans le cas des fractales (cf. chapitre Fractales) les dimensions sont variables et fractionnaires. Considérons la courbe de Von Koch (par exemple) après une itération de la suite la définissant:



Figure: 21.3 - Courbe de Von Koch

Soit L sa taille tel que $L = 1$. Pour calculer sa dimension nous prenons l'élément fondamental de la courbe (ci-dessous en rouge):



Figure: 21.4 - Choix de l'étalon pour la courbe de Von Koch

Soit n la taille de cet étalon tel que $n = 1/3$. Nous voyons très bien que nous pouvons le reporter 4 fois sur la courbe. Donc:

$$d = \frac{\ln(4)}{\ln\left(\frac{1}{1/3}\right)} = \frac{\ln(4)}{\ln(3)} \cong 1.26... \quad (21.4)$$

La dimension de la fractale de Von Koch a donc une valeur fractionnaire et est plus une courbe (proche de 1), qu'une surface (qui est de dimension 2). Les choux-fleurs sont par exemple des fractales de dimension 2.33...

Nous pouvons donc calculer la dimension de n'importe quels objets fractals à la condition de connaître leurs mesures.

Ne nous hasardons pas à aller chercher des objets complexes dans quelques galaxies alors que la fractale la plus connue se trouve dans votre assiette (la seconde étant vos poumons...). Eh oui! Le chou-fleur est bien une fractale (au même titre que vos poumons)! Vous avez sûrement déjà remarqué que quand nous découpons le chou-fleur (chose pas conseillée à tenter pour vos poumons...), nous le cassons au lieu de le couper, et ça donne plein de petits choux-fleurs, qui eux même peuvent donner d'autres plus petits choux fleurs. Cette particularité d'autosimilarité à différentes échelles fait du chou-fleur une fractale.

Calculons à présent la dimension fractale du chou-fleur. Quand nous cassons le chou-fleur, nous obtenons entre 12 et 14 branches qui ressemblent au chou-fleur entier à une dilatation près. Cette dilatation est, si nous la calculons, de facteur 3. Donc, selon la formule ci-dessus, la dimension fractale du chou-fleur est d'environ:

$$d = \frac{\ln(13)}{\ln(3)} \cong 2.33... \quad (21.5)$$

Donc le chou-fleur est plus une surface qu'un volume.

Faisons un dernier exemple avec la fractale du triangle de Sierpinski (cf. chapitre Fractales):

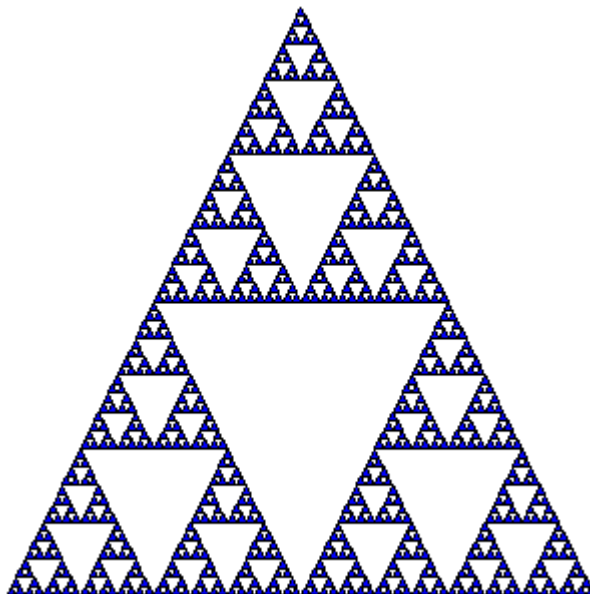


Figure: 21.5 - Triangle de Sierpinski

D'abord, il est évident que nous avons besoin d'un carré pour le recouvrir complètement. Avec des carrés deux fois plus petits, on aura besoin de trois carrés:

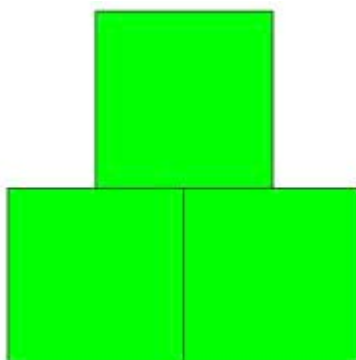


Figure: 58.6 - Trois carrés pour recouvrir la fractale de Sierpinski

Si nous divisons encore une fois la taille des carrés par deux, il nous faudra 9 carrés pour recouvrir le triangle:

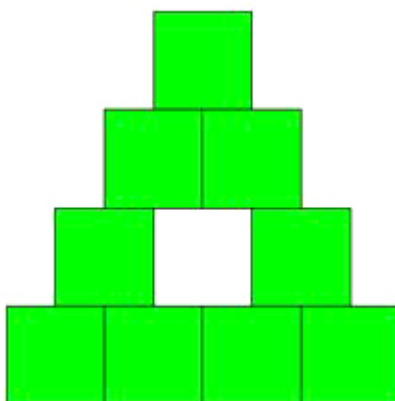


Figure: 58.7 - Neuf carrés pour recouvrir la fractale de Sierpinski

Si nous divisons encore la taille des carrés par deux il nous en faudra 27:

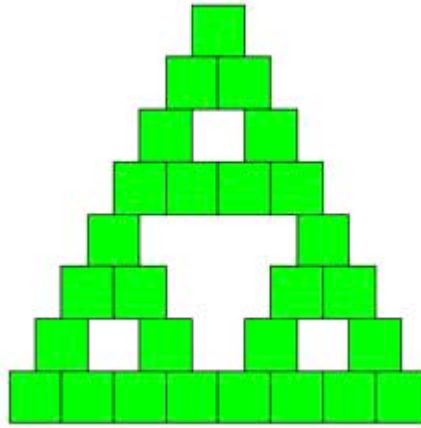


Figure: 58.8 - Plus de carrés encore...

Et donc nous remarquons que:

$$3 = 2^{\frac{\ln(3)}{\ln(2)}}, 9 = 4^{\frac{\ln(3)}{\ln(2)}}, 27 = 8^{\frac{\ln(3)}{\ln(2)}} \quad (21.6)$$

et donc:

$$d = \frac{\ln(3)}{\ln(2)} \cong 1.585... \quad (21.7)$$

Donc la fractale de Sierpinski est plus une surface qu'une courbe.

Il existe également d'autres dimensions. Prenons pour exemple, les "**dimensions d'homothétie**" dont voici quelques exemples simples (voir plus loin la définition rigoureuse de "l'homothétie"):

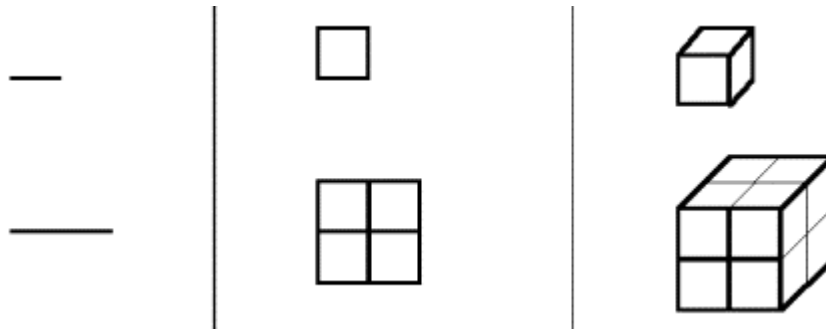


Figure: 21.9 - Représentation des dimensions d'homothétie

Le segment (tout à gauche), de dimension 1 a par homothétie, vu sa longueur, doublé et nous remarquons que:

$$d = 1, H = 2, N = 2^1 = 2 \quad (21.8)$$

Le carré (au milieu), de dimension 2 a par homothétie, vu sa surface, doublé et nous remarquons que:

$$d = 2, H = 2, N = 2^2 = 4 \quad (21.9)$$

Le cube (tout à droite), de dimension 3 a par homothétie vu son volume doublé et nous remarquons que:

$$d = 3, H = 2, N = 2^3 = 8 \quad (21.10)$$

Le facteur de duplication d'échelle (homothétie) est donc égal à:

$$2^{\text{Dimension Topologique}} \quad (21.11)$$

Comme vous pouvez le voir, il s'agit toujours d'une valeur entière mais d'un autre type de dimension.

Le concept de dimensions ayant été introduit intéressons-nous maintenant aux postulats d'Euclide qui pourront paraître vagues dans un premier temps mais qui seront détaillés au fur et à mesure de notre lecture.

CONSTRUCTIONS D'EUCLIDE

La construction de la géométrie plane d'Euclide se fonde sur cinq postulats (dont les quatre premiers sont considérés aujourd'hui comme des axiomes comme nous en avons déjà fait mention):

P1. Conduire une droite d'un point quelconque à un point quelconque.

Sous forme moderne nous dirions que par deux points distincts A et B , il passe une droite et il n'en passe qu'une seule.

Autrement dit: Deux droites (D) et (D') qui ont deux points communs sont confondues, tout point de l'une est un point de l'autre et réciproquement.

Il résulte de ce postulat que deux droites (D) et (D'), ou bien n'ont aucun point commun, ou bien ont un seul point commun qui s'appelle "**point d'intersection**" et sont alors "**sécantes**" et "**distinctes**", ou bien ont plus d'un point commun et sont alors "**confondues**".

P2. Prolonger indéfiniment, selon sa direction, une droite finie.

Sous forme moderne nous dirions que tout segment AB est prolongeable en une droite passant par A et B (compte tenu du premier axiome, elle est unique dans une géométrie euclidienne)

P3. D'un point quelconque, et avec un intervalle quelconque, décrire une circonférence de cercle.

Sous forme moderne nous dirions pour tout point A et tout point B distinct de A , nous pouvons décrire un cercle de centre A passant par B .

P4. Tous les angles droits sont égaux entre eux.

Sous forme moderne nous dirions qu'à chaque angle $\widehat{AOB}$ du plan correspond sa mesure θ , effectuée avec une unité choisie une fois pour toutes où θ est un nombre positif, inférieur à 2π . Réciproquement, soit θ un nombre positif quelconque compris entre 0 et 2π , nous admettrons qu'il existe une infinité d'angles $\widehat{AOB}$ égaux entre eux dont la mesure avec l'unité d'angle choisie soit θ .

P5. Si une droite, tombant sur deux droites, fait les angles intérieurs du même côté plus petits que deux droits, ces droites, prolongées à l'infini, se rencontreront du côté où les angles sont plus petits que deux droits.

Sous forme moderne nous dirions que: étant donnés une droite et un point, il existe une unique droite passant par ce point et ne coupant pas la droite initiale.

La construction d'Euclide permet donc le développement de la notion de mesure de longueur, d'aire, de volume, d'angle comme nous allons le voir plus loin.

Les deux théorèmes fondamentaux de la géométrie euclidienne sont le théorème de Pythagore et celui de Thalès comme nous en verrons la démonstration plus loin. Un peu d'analyse permet d'aller plus loin avec la trigonométrie que nous avons déjà développée dans le chapitre précédant.

DROITE ET SEGMENTS

Dans un premier temps, la figure géométrique la plus simple (mis à part le point...) en géométrie euclidienne est la "ligne droite" et celle-ci est directement concernée par deux premiers postulats d'Euclide.

Définitions:

D1. La "**ligne droite**" est l'image donnée par un fil tendu d'épaisseur nulle et de longueur infinie.

Remarque: Nous pouvons également définir la "**ligne droite**" comme une infinité de points mis à côté les uns des autres dans une même direction sur un plan.

D2. Nous appelons "**demi-droite**" la portion de droite limitée à un point O appelé "**origine**".

Remarque: L'expression "la demi-droite OA " désigne la demi-droite d'origine O , point nommé le "premier", qui contient le point A .

D3. Nous disons que deux demi-droites OA , OB sont des "**demi-droites opposées**" lorsqu'elles constituent la droite AB toute entière.

D4. Nous appelons "**segment**" AB une portion de droite limitée par deux points A et B . Ces points sont appelés les "**extrémités**" du segment.

GRANDEURS DE MÊME ESPÈCE

Nous disons que des figures géométriques (sous-entendu des droites) sont des "**grandeurs de même espèce**" lorsqu'il est possible de définir:

1. Dans quel cas une figure (A) sera dite égale à une figure (B) et, si elles sont inégales, laquelle est la plus petite.
2. Ce que nous devons entendre par somme d'une figure (A) et d'une figure (B).

Les définitions choisies doivent être telles que si (A) est déclaré plus petit que (B) et (B) plus petit que (C), (A) soit déclaré aussi plus petit que (C).

Il faut, en outre, que la figure appelée "somme de (A) et de (B)" soit égale à celle qui est appelée somme de (B) et de (A).

Enfin, la substitution dans une comparaison, une égalité ou une somme, d'une figure par une figure égale ne doit pas modifier le résultat des opérations.

Pour faire comprendre ce que sont des grandeurs de même espèce, prenons l'exemple des segments

de droite:

- Nous admettrons qu'il est possible de décider de l'égalité de deux segments AB et $A'B'$ lorsque nous pouvons les faire coïncider.

- Nous admettrons aussi qu'il est possible de remplacer le segment $A'B'$ par un segment égal AD et porté sur la demi-droite AB .

- Enfin, nous admettrons qu'il est possible de distinguer entre trois points A, B, C , pris au hasard sur une droite, lequel est entre les deux autres.

Nous convenons alors de dire que le segment $A'B'$ est plus petit que le segment AB , ce qui s'écrit en abrégé:

$$A'B' < AB \quad (21.12)$$

lorsque le point C , obtenu en portant sur la demi-droite AB un segment AC égal à $A'B'$, tombe entre A et B .

Si le point C était en B , les segments AB et $A'B'$ seraient égaux et nous écririons alors:

$$A'B' = AB \quad (21.13)$$

Nous convenons d'appeler la "**somme de deux segments**" $AB, A'B'$, le segment AC obtenu en portant sur la demi-droite opposée à la demi-droite BA un segment BC égal à $A'B'$. Nous traduisons cette opération en écrivant:

$$AC = AB + A'B' \quad (21.14)$$

Considérons toujours des grandeurs de même espèce... Ajouter entre elles plusieurs de ces grandeurs, c'est ajouter l'une d'elles à une autre, la somme obtenue à une troisième, etc. Par exemple, ajouter les segments AB, BC, CD , c'est ajouter AB et BC ce qui donne AC , ensuite AC et CD ce qui donne AD . Nous résumons l'opération en écrivant:

$$AD = AB + BC + CD \quad (21.15)$$

Multiplier une grandeur par un nombre entier n , c'est ajouter n grandeurs égales à celle-là. Par exemple, si nous avons $AB = BC = CD$, la relation précédente s'écrirait:

$$AD = 3 \cdot AB = 3 \cdot BC = 3 \cdot CD \quad (21.16)$$

Nous allons définir ce que nous appelons "**comparer deux grandeurs (A) et (B) de même espèce**":

Choisissons arbitrairement une grandeur (C) de même espèce que (A) et que (B) et plus petite que chacune d'elles. Formons une suite de grandeurs telles que:

$$(C_1) = (C), (C_2) = 2 \cdot (C), \dots, (C_p) = p \cdot (C) \quad (21.17)$$

Nous constatons que la grandeur (A) s'intercale entre deux grandeurs (C_p) et (C_{p+1}) et que (B) se trouve entre deux autres (C_q) et (C_{q+1}) par construction.

Nous disons alors par définition que le rapport des grandeurs (A) et (B) est un nombre $(A)/(B)$ positif

compris entre:

$$(C_p)/(C_{q+1}) \text{ et } (C_{p+1})/(C_q) \quad (21.18)$$

Prenons alors, par exemple, deux segments quelconques AB , $A'B'$ mais différents (par exemple 1.2 [cm] et 3.5 [cm]):

Pour réaliser l'opération précédente, utilisons une règle graduée dont l'unité sera la grandeur C , un segment arbitrairement choisi (par exemple de 1 [cm]).

Nous appliquons le zéro de la règle en A , et B s'intercalera par construction entre deux graduations de la règle numérotée p et $p+1$ à moins que la grandeur C soit égale à AB ... (soit avec le choix pris comme exemple, B s'intercalera entre la 1ère et la 2ème).

Pour $A'B'$, nous appliquons le zéro de la règle aussi en A' , et B' s'intercalera aussi entre deux grandeurs de la règle numérotée q et $q+1$ à moins que la grandeur C soit égale à $A'B'$... (soit avec le choix pris comme exemple, B' s'intercalera entre la 3ème et la 4ème).

Nous exprimerons le résultat de ces mesures et leur rapport en écrivant:

$$\frac{p}{q+1} < \frac{AB}{A'B'} < \frac{p+1}{q} \quad (21.19)$$

où le terme tout à gauche s'appelle une "**mesure par défaut**" et celui à l'opposé une "**mesure par excès**".

Ainsi avec les mesures prises comme exemple nous avons donc:

$$\frac{1}{3+1} < \frac{1.2}{3.5} < \frac{1+1}{3} \Rightarrow 0.25 < 0.342 < 0.\overline{6} \quad (21.20)$$

Définition: Nous appelons "**mesure d'une grandeur**" (A) le nombre positif qui mesure le rapport de cette grandeur et d'une grandeur (U) arbitrairement choisie et que nous appelons "**l'unité**", la mesure de l'unité étant "1", par définition.

Nous pouvons démontrer que si a est la mesure de (A), b celle de (B) évaluées toutes deux avec une même unité (U), le nombre (A)/(B) est égal au rapport a/b . Ce rapport étant indépendant de l'unité choisie.

Remarque: Nous disons que (B) est une "**partie aliquote**" de (A) si le rapport (A)/(B) est un nombre entier.

Nous conviendrons une fois pour toute, qu'en géométrie toutes les grandeurs de même espèce qui interviennent dans une figure donnée sont mesurées avec la même unité.

Soit (A), (B), (C), ..., (S) des grandeurs de même espèce et (A'), (B'), (C'), ..., (S'), des grandeurs de même espèce, mais qui ne sont pas nécessairement de même espèce que les précédentes. Nous disons que ces grandeurs sont "**homologues**", si nous pouvons les grouper deux à deux, (A') homologue de (A), (B') homologue de (B), ..., etc., de manière que les conditions suivantes soient réalisées:

- Si (A) est égal à (B), (A') est égal à (B');

- Si (A) est plus petit que (B) , (A') est plus petit que (B') ;
- Si (S) est la somme de (A) et de (B) , (S') est la somme de (A') et de (B') .

Pour calculer le rapport $(A)/(B)$, formons la suite $(C_1), (C_2), \dots, (C_p)$ et pour calculer le rapport $(A')/(B')$, formons la suite $(C'_1), (C'_2), \dots, (C'_p)$, obtenue comme la précédente, mais à partir de la grandeur (C') homologue de (C) .

Il est évident que si (A) s'intercale entre (C_p) et (C_{p+1}) , (A') s'intercalera entre (C'_p) et (C'_{p+1}) ; de même, si (B) s'intercale entre (C_q) et (C_{q+1}) . Les rapports $(A)/(B)$ et $(A')/(B')$ seront encadrés par les mêmes nombres:

$$\frac{p}{q+1} \quad \text{et} \quad \frac{p+1}{q} \quad (21.21)$$

Par conséquent: Le rapport de deux grandeurs (A) et (B) est égal au rapport des grandeurs homologues (A') et (B') .

Si, en particulier, les grandeurs $(A), (B), \dots$ sont mesurées avec une unité (U) , et si les grandeurs $(A'), (B'), \dots$ sont mesurées avec une unité (U') , homologue de (U) , les rapports égaux:

$$(A)/(U) \text{ et } (A')/(U') \quad (21.22)$$

ne sont autres que les mesures de (A) et de (A') . Par conséquent:

Les mesures de deux grandeurs homologues (A) et (A') sont égales, à condition que les unités choisies pour les mesurer soient des grandeurs homologues.

Considérons maintenant sur une demi-droite Ox un point M . Soit x la mesure (selon la définition précédente) de OM . A chaque point M de la demi-droite correspond un nombre positif x et un seul; nous admettrons qu'à un nombre positif x arbitrairement choisi correspond réciproquement un point M de la demi-droite et un seul.

Une conséquence de cette hypothèse est qu'il existe un point, et un seul, C qui divise le segment OM en parties égales. Ce point est le point de la demi-droite OM , tel que:

$$OC = \frac{OM}{2} \quad (21.23)$$

Nous l'appelons "milieu du segment" OM .

Théorème: Il existe un point M et un seul, situé sur le segment AB tel que la mesure du rapport MA/MB soit égal à un nombre positif donné λ .

Remarque: Si $\lambda = 1$, ce point est le milieu du segment.

Démontrons cette unicité: Soit M un point quelconque du segment AB ; soit x , la mesure de AM , et a la mesure de AB : la mesure de MB sera $a-x$ puisque M est placé entre A et B . Nous aurons alors:

$$\frac{MA}{MB} = \frac{x}{a-x} \quad (21.24)$$

Pour que ce rapport soit égal à λ , il faut, et il suffit, que x soit solution de l'équation:

$$\frac{x}{a-x} = \lambda \quad (21.25)$$

Or, cette équation admet la seule solution:

$$x = \frac{a\lambda}{1+\lambda} \quad (21.26)$$

À cette valeur positive et inférieure à a de x correspond un point M et un seul de la demi-droite AB tel que $MA=x$. Ce point M satisfait, et satisfait seul, aux conditions imposées.

Théorème: Il existe un point M et un seul de la droite AB , situé en dehors du segment AB , tel que le rapport MA/MB soit égal à un nombre donné et défini λ différent de 1.

Démontrons aussi cette unicité:

1. Supposons $\lambda > 1$. Soit M un point quelconque de la droite AB situé en dehors du segment AB : ou bien A est sur le segment MB , ou bien B est sur le segment MA . Si A est sur le segment MB , nous avons nécessairement $MB > MA$, donc:

$$\frac{MA}{MB} < 1 < \lambda \quad (21.27)$$

Le point M ne répond donc pas à la question de l'unicité.

Si B est sur le segment MA , $MA=x$, $AB=a$, $MB=x-a$. Nous aurons donc:

$$\frac{MA}{MB} = \frac{x}{x-a} \quad (21.28)$$

Pour que ce rapport soit égal à λ , il faut, et il suffit que x soit solution de l'équation:

$$\frac{x}{x-a} = \lambda \quad (21.29)$$

Cette équation admet comme seule solution:

$$x = \frac{a\lambda}{\lambda-1} \quad (21.30)$$

qui donne une valeur de x qui sera toujours positive et supérieure à λ . À cette valeur positive et supérieure à a de x correspond un point M et un seul de la demi-droite AB . Ce point M satisfait, et satisfait seul aux conditions imposées d'unicité.

2. Supposons maintenant $\lambda < 1$. Nous chercherons le point M pour lequel (nous avons simplement inversé le rapport):

$$\frac{MB}{MA} = \frac{1}{\lambda} \quad (21.31)$$

est un nombre supérieur à 1. Il y en a un et un seul d'après le point (1). Il satisfait seul aux conditions imposées.

Remarque: Il n'existe aucun point M situé en dehors du segment AB pour lequel $MA/MB=1$. En effet, si A est sur le segment MB , nous avons, quel que soit M , $MA/MB < 1$, et si B est sur le segment MA , $MA/MB > 1$.

PLAN

Passons maintenant à l'étude d'un objet géométrique de dimension supérieure à celle de la droite qu'est le plan et la surface.

Considérons une surface finie (S) et deux points A et B de cette surface. Deux cas peuvent se présenter:

1. Il y a des points de la droite AB qui ne sont pas sur la surface (S). Nous dirons dans ce cas: la droite AB coupe la surface; les points communs à la droite AB et à la surface (S) sont les points d'intersection de la surface (S) et de la droite AB . Parmi ces points communs se trouvent, en particulier les points A et B .
2. Tous les points de la droite AB sont des points de la surface (S). Nous disons alors que la droite AB est sur la surface (S).

Définition: Nous appelons "**plan**" la surface telle que toute droite AB qui joint deux points arbitrairement choisis sur la surface, soit sur la surface.

Nous admettrons qu'une pareille surface existe et que par trois points ABC , non alignés, il passe un plan et un seul. L'étude des plans sera faite ultérieurement: nous nous consacrons actuellement à l'étude des figures géométriques tracées dans un plan donné, figures dites "**figures planes**". Leur étude constitue la "**géométrie plane**".

Remarque: Dans la pratique, les figures sont tracées soit sur une feuille de papier, soit sur la surface du tableau noir.

Le plan étant défini, nous pouvons déjà nous intéresser à des opérations de base concernant les figures du plan que nous détaillerons plus tard rigoureusement.

DÉPLACEMENTS ET RETOURNEMENTS

Soit (F) un dessin effectué sur un tableau plan: effectuons sur un papier transparent, dont le recto est appliqué sur le plan du tableau, un calque (F_1) du dessin (F). Effectuons, en appliquant ce calque en un autre point du tableau, un nouveau dessin (F') identique à (F).

Deux cas sont à considérer:

1. Si le recto du papier transparent est demeuré appliqué sur le tableau, le dessin (F') se déduit de (F) par une opération appelée "**déplacement**" ou "**translation**".

2. Si, au contraire, le papier a été retourné, et si c'est le verso qui est appliqué sur le tableau, l'opération s'appelle "**retournement**" ou "**symétrie**".

A chaque point A du dessin (F), l'une quelconque de ces deux opérations fait correspondre un point A' du dessin (F'), que nous appelons "**l'homologue**" de A . Un segment AB de (F) vient en coïncidence avec un segment $A'B'$ homologue de (F'); ces deux segments sont par définition égaux et cela quels que soient A et B .

Définition: Nous disons communément que le dessin (F) est "**superposable**" au dessin (F') et que ces dessins représentent des figures égales.

ANGLES

Nous avons déjà brièvement défini le concept "d'angle" dans le chapitre précédent traitant de la trigonométrie. Nous avons maintenant en plus le quatrième postulat d'Euclide à notre disposition concernant ce concept.

Nous allons maintenant revenir plus en détails et en voir les concepts sous-jacents qui vont nous permettre d'aborder plus loin un objet particulièrement utile qu'est la bissectrice:

Définitions:

D1. Nous appelons "**angle**" (ou "**angle plan**") ou plus rigoureusement "**angle rectiligne**" la portion de plan limitée par deux demi-droites OA , OB , par exemple. Le point O s'appelle "**sommet**" de l'angle, les demi-droites OA , OB , s'appellent les "**côtés**" de l'angle.

D2. Nous appelons "angle formé par deux segments AB , AC ", l'angle de sommet A dont les côtés sont les demi-droites AB , AC .

D3. Les demi-droites OA , OB divisent le plan en deux régions: elles définissent donc deux angles:

1. L'un constitué par la région couverte de hachures (voir la figure à l'extrémité gauche ci-dessous) s'appelle "**angle saillant**".

2. L'autre, constitué par la région couverte de hachures (voir la figure au centre ci-dessous) s'appelle "**angle rentrant**".

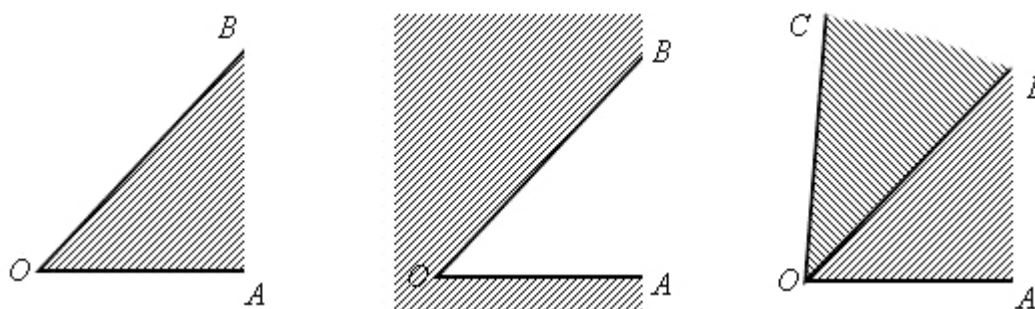


Figure: 21.10 - Angle saillant, rentrant, angles adjacents

La notation $\widehat{BOA}$ ou $\widehat{AOB}$ désigne un de ces deux angles: la lettre qui indique le sommet doit être (normalement) écrite au milieu (souvent on ne mentionne pas le sommet si le contexte est évidente). Lorsqu'aucune précision n'accompagne cette notation, elle représente par définition l'angle saillant!!

D4. Nous appelons "**angles adjacents**" deux angles qui ont le sommet et un côté communs et qui sont placés de part et d'autre de ce côté commun. Sur la figure ci-dessus à l'extrême droite, les angles

saillants $\widehat{AOB}$, $\widehat{BOC}$, sont adjacents.

Soit, $\widehat{AOB}$ et $\widehat{A'O'B'}$ deux angles d'un même plan. Nous avons admis précédemment qu'il existe un déplacement qui amène O' en O et A' en un point A de la demi-droite OA ; ce déplacement amène $O'B'$ soit en OB_1 , de manière que les deux angles $\widehat{AOB}$, $\widehat{AOB_1}$ ne soient pas adjacents, soit en OB_2 , de manière que $\widehat{AOB}$ et $\widehat{AOB_2}$ soient adjacents. Dans ce dernier cas, un demi-tour supplémentaire autour de OA amènera $\widehat{AOB_2}$ dans la position $\widehat{AOB_1}$. Ce déplacement et ce retournement, s'il y a lieu, remplace $\widehat{A'O'B'}$ par un angle $\widehat{AOB_1}$, égal par définition.

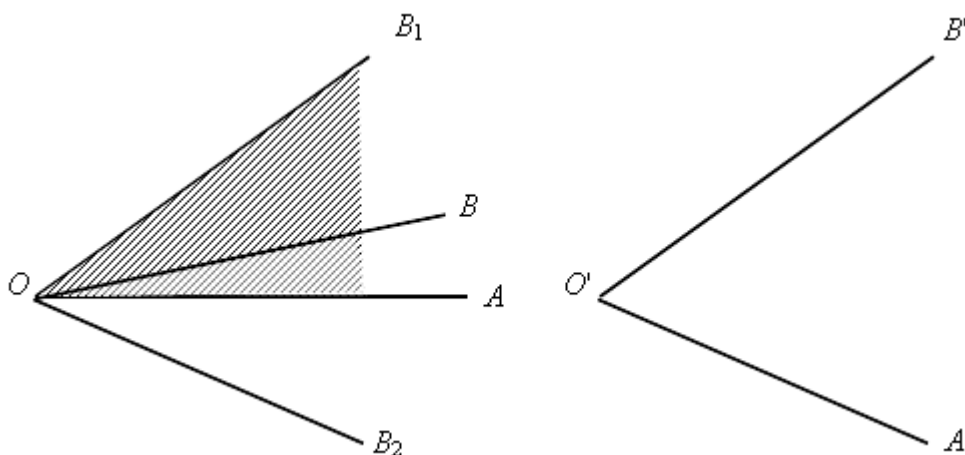


Figure: 21.11 - Représentation du déplacement de deux angles

Si OB_1 est confondu avec OB , il y a des points de l'un des deux angles $\widehat{AOB_1}$ et $\widehat{AOB}$ qui sont égaux, tout point de l'un étant un point de l'autre: nous dirons, dans ce cas, que les angles $\widehat{AOB}$, $\widehat{A'O'B'}$ sont des "angles égaux", ce qui s'exprime par l'égalité:

$$\widehat{AOB} = \widehat{A'O'B'} \quad (21.32)$$

Si OB_1 n'est pas confondu avec OB , il y a des points de l'un des deux angles $\widehat{AOB_1}$, par exemple, qui ne sont pas des points de AOB . Sur la figure ci-dessus, l'angle $\widehat{AOB}$ est couvert de hachures, l'angle $\widehat{AOB_1}$ également; les points dont nous parlons sont ceux de l'angle $\widehat{BOB_1}$ couvert une seule fois de hachures. Nous conviendrons de dire que l'angle $\widehat{AOB_1}$ et, par conséquent, l'angle égal à $\widehat{A'O'B'}$ sont, dans ce cas, plus grands que l'angle $\widehat{AOB}$, ce qui s'exprime par l'inégalité:

$$\widehat{A'O'B'} > \widehat{AOB} \quad (21.33)$$

Maintenant que nous sommes en mesure de comparer des angles, étudions comment nous pouvons sommer (et donc respectivement soustraire) ceux-ci.

Etudions d'abord le cas de la somme de deux angles $\widehat{AOB}$, $\widehat{BOC}$ adjacents. Deux cas peuvent se présenter suivant que les angles sont saillants ou rentrants:

1. Soient $\widehat{AOB}$ et $\widehat{BOC}$ les deux angles à additionner (voir figure ci-dessous à gauche). Par

définition, la somme de ces angles est l'angle $\widehat{AOC}$, ce que nous exprimons par l'égalité:

$$\widehat{AOC} = \widehat{AOB} + \widehat{BOC} \quad (21.34)$$

2. Soit $\widehat{AOB}$ un angle saillant à additionner à l'angle rentrant $\widehat{BOC}$. Si nous couvrons de hachures successivement les deux angles (voir figure ci-dessous à droite), l'angle saillant $\widehat{AOC}$ se trouve couvert deux fois:

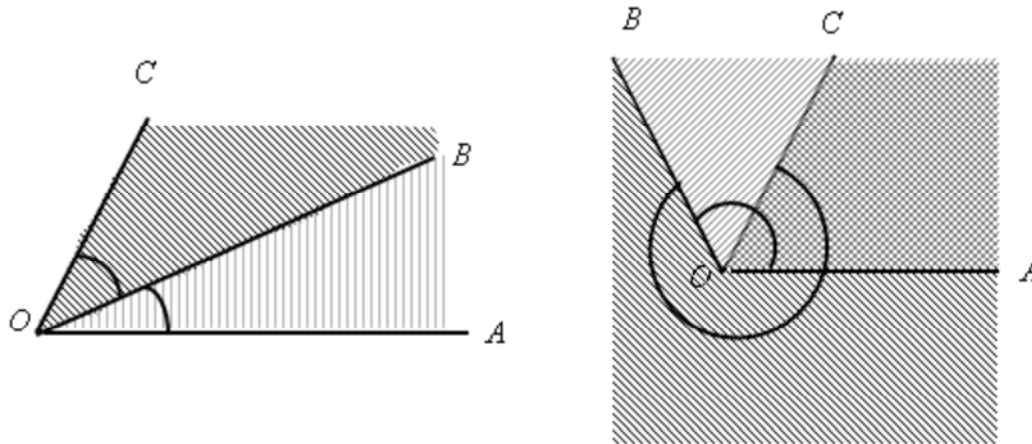


Figure: 21.12 - Somme de deux angles adjacents

Dans ce cas, la somme des angles $\widehat{AOB}$ (sailant) et $\widehat{BOC}$ (rentrant) est donc égale à $\widehat{AOC}$ augmenté de deux "angles plats" (voir plus loin la définition), ce qui s'exprime en écrivant:

$$\widehat{AOB} + \widehat{BOC} = \widehat{AOC} + 2 \text{ angles plats} \quad (21.35)$$

Remarque: Ceci peut paraître confus à certains mais ceux qui auront déjà parcouru le chapitre de Trigonométrie savent que les angles du cercle trigonométriques sont égaux à eux-mêmes modulo $2\pi [\text{rad}]$.

Étudions maintenant le cas de la somme de deux angles quelconques:

La somme de deux angles $\widehat{AOB}$, $\widehat{A'O'B'}$ est par définition, égale à la somme de l'angle $\widehat{AOB}$ et d'un angle $\widehat{BOC}$ égal à l'angle $\widehat{A'O'B'}$ et adjacent à l'angle $\widehat{AOB}$.

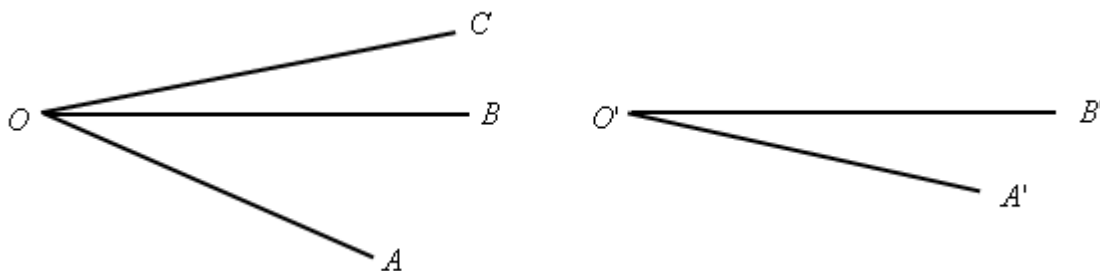


Figure: 21.13 - Somme de deux angles quelconques

Un pareil angle est obtenu par un déplacement qui amène O' en O et A' en un point OA , suivi ou non

d'un retournement autour de OA .

Étudions comme dernier cas la somme de plus de deux angles:

La somme de plusieurs angles $\widehat{AOB}$, $\widehat{A'O'B'}$, etc., est, par définition, égale à la somme obtenue en ajoutant le premier au second, cette somme au troisième, et ainsi de suite.

Soit $\widehat{AOB}$ le premier angle, $\widehat{BOC}$ le deuxième angle, $\widehat{KOL}$ un angle égal au dernier des angles à ajouter et adjacent au précédent. Le résultat de l'addition sera $\widehat{AOL}$ augmenté d'autant de fois deux angles plats que le plan a été recouvert au cours des opérations. On constate aisément que ce résultat ne dépend pas de l'ordre des angles à ajouter $\widehat{AOB}$, $\widehat{A'O'B'}$, etc.

D5. Deux angles formés par deux droites coupées par une sécante sont dits "**angles alternes-internes**" si :

1. ils sont situés de part et d'autre de la sécante;
2. ils sont situés entre les deux droites;
3. ils ne sont pas angles adjacents.

Dans l'exemple ci-dessous, les droites (X) et (Y) sont coupées respectivement en A et en B par la sécante (S).

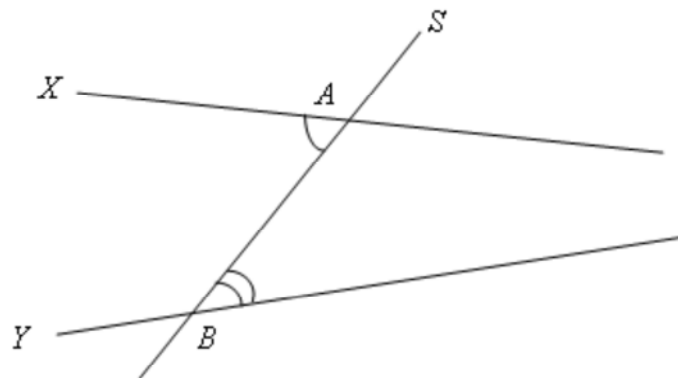


Figure: 21.14 - Exemple particulier d'angle alterne-interne

et les deux angles représentés sont les angles alternes-internes.

MESURES DES ANGLES

Nous avons défini l'égalité et la somme de deux ou plusieurs angles. Ces définitions satisfont aux conditions de grandeurs de même espèce que nous avons déjà vues précédemment.

Choisissons donc arbitrairement un angle du plan $\widehat{aob}$, qui sera l'unité d'angle pour le plan: la mesure du rapport $\widehat{AOB} / \widehat{aob}$, effectuée comme il a été expliqué précédemment lors de notre discussion sur les grandeurs de même espèce, sera un nombre positif θ , appelé par définition "mesure de l'angle $\widehat{AOB}$, avec l'unité choisie $\widehat{aob}$ ".

Nous désignons par la lettre grecque minuscule "pi" le nombre irrationnel:

$$\pi = 3.14158\dots \quad (21.36)$$

qui est par définition la mesure d'un "angle plat" (nous verrons qu'elle en est l'unité un peu plus loin).

Remarque: Comme tous les angles du plan sont plus petits que deux angles plats, le nombre θ (qui est la mesure de $\widehat{AOB}$) doit être inférieur à 2π .

Ayant défini l'angle plat, nous pouvons maintenant définir d'autres types d'angles d'usage courant:

Définitions:

D1. Nous appelons "angle droit", tout angle égal à la moitié d'un angle plat.

D2. Nous disons que deux angles sont des "angles perpendiculaires", notés $\perp$, lorsqu'ils sont tous les deux droits et adjacents.

D3. Nous appelons "angle aigu" tout angle inférieur à un angle droit et "angle obtus" tout angle supérieur à un angle droit.

D4. Nous appelons "angle orienté" ou "angle vectoriel", l'angle défini entre deux vecteurs ou droites (cf. chapitre de Calcul Vectoriel) de même origine et dont la valeur mesurée dans le sens inverse des aiguilles d'une montre sera pris comme positif et négatif s'il est pris dans le sens des aiguilles d'une montre.

L'exemple le plus connu d'angle orienté est celui du cercle trigonométrique:

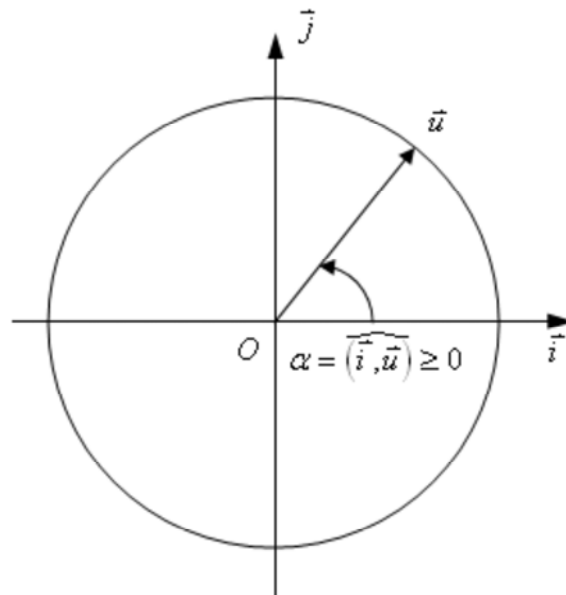


Figure: 21.15 - Angle orienté positif ou nul

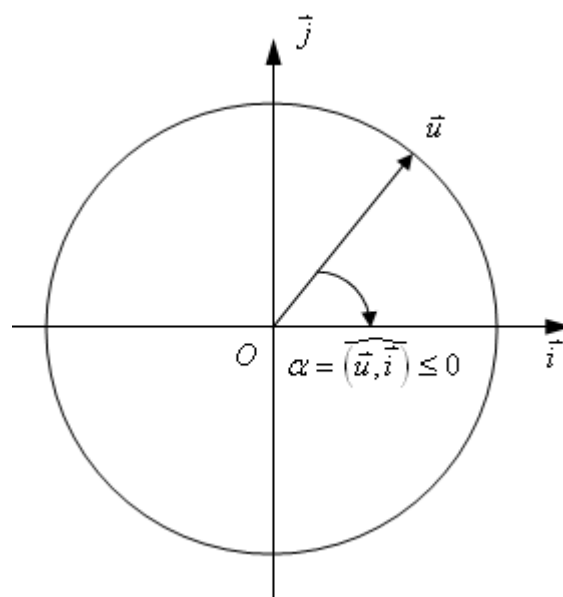


Figure: 21.16 - Angle orienté négatif ou nul

D5. Nous disons que deux angles sont des "**angles supplémentaires**" lorsque leur somme vaut deux angles droits (soit un angle plat).

D6. Nous disons que deux angles sont des "**angles complémentaires**" lorsque leur somme vaut un angle droit.

Considérons maintenant les symboles α, β, γ comme les mesures de plusieurs angles $\widehat{AOB}$, $\widehat{BOC}$, $\widehat{COD}$.

Nous n'insistons pas sur le fait évident que les égalités $\widehat{AOB} = \widehat{BOC}$ ou $\alpha = \beta$ sont équivalentes, ainsi que les inégalités $\widehat{AOB} < \widehat{BOC}$, $\alpha < \beta$. Ces remarques de bon sens s'imposeront chaque fois que des grandeurs de même espèce auront été mesurées, bien entendu avec la même unité.

En revanche, nous insisterons sur le fait que, d'après la définition même de la somme de plusieurs angles, $\alpha + \beta + \gamma$ est la mesure de l'angle $\widehat{AOD}$ augmentée d'autant de fois deux angles plats que le plan a été recouvert au cours des opérations d'addition.

$$\alpha + \beta + \gamma = \widehat{AOD} + 2n\pi \quad (21.37)$$

Le nombre entier n qui s'introduit dans ce calcul a une valeur qui pourrait être précisée, mais qui n'a pour le géomètre aucune importance, comme nous pourrions le constater ultérieurement. Nous décidons donc de ne pas écrire en géométrie le $2n\pi$ inutile (mais sous-entendu). Egalement, nous décidons par convention d'écrire:

$$\widehat{AOD} = \theta \quad (21.38)$$

Ainsi, nous avons $\alpha + \beta + \gamma = \theta$ cette convention d'écriture signifie que θ est la mesure de l'angle $\widehat{AOD}$ si nous avons $0 \leq \theta < 2\pi$.

Si θ est supérieur à 2π , l'égalité signifie que la mesure de l'angle $\widehat{AOD}$ est $\theta - 2k\pi$, k étant un nombre entier positif ou nul choisi de manière que nous ayant $0 \leq \theta - 2k\pi < 2\pi$.

Remarques:

R1. Il existe un nombre entier positif k et un seul, tel que nous ayons $0 \leq \theta - 2k\pi < 2\pi$, c'est-à-dire:

$$\frac{\theta}{2\pi} - 1 < k \leq \frac{\theta}{2\pi} \quad (21.39)$$

car les deux nombres $\frac{\theta}{2\pi} - 1$ et $\frac{\theta}{2\pi}$ sont positifs et différents de 1.

R2. Dire que α est la mesure de l'angle $\widehat{AOD}$ suppose que nous ayons $0 \leq \alpha < 2\pi$ et entraîne l'égalité $\widehat{AOD} = \alpha$. Mais écrire $\widehat{AOD} = \alpha$ n'entraîne pas nécessairement, que α soit la mesure de $\widehat{AOD}$; il faut, en outre, que l'on ait $0 \leq \alpha < 2\pi$.

UNITÉS DE MESURE DES ANGLES

Nous avons défini l'angle plat comme étant égal à π sans spécifier l'unité. C'est ce que nous allons maintenant nous appliquer à faire. Il existe (encore) plusieurs unités de mesures d'angle dont voici la liste:

Définitions:

D1. Nous appelons "**degré**" la 180^{ème} partie de l'angle plat.

Tous les calculs anciens sont effectués en degrés; les sous-multiples du degré sont: la "**minute sexagésimale**", égale au 60^{ème} du degré, et la "**seconde sexagésimale**", égale au 60^{ème} de la minute sexagésimale.

La notation $\theta = 30^\circ 18' 11''$ se lit: trente degrés, dix-huit minutes, onze secondes. Ce type de mesure est courant encore en astronomie.

Remarque: Nous utilisons encore aujourd'hui couramment le degré dans les écoles mais sans la notation usant des minutes et secondes (pas commode pour la somme des angles). Nous notons alors l'angle en degrés avec une notation décimale comme par exemple $\theta = 45.34354^\circ$.

D2. Nous appelons "**grade**" la 200^{ème} partie de l'angle plat.

Le grade est également une ancienne unité d'angle. Ses sous-multiples sont: la "**minute centésimale**", égale au 100^{ème} du grade, et la "**seconde centésimale**", égale au 100^{ème} de la minute centésimale.

La notation $\theta = 40.1824 \text{ gr}$ se lit: quarante grades, dix-huit minutes, vingt-quatre secondes.

D3. Nous appelons "**1 radian**" (noté [rad]) l'angle plan décrit par une sécante à un cercle, passant par son centre, tel que l'arc de cercle ainsi défini par l'axe horizontal passant par le centre du cercle et la sécante soit d'égale longueur au rayon de ce cercle (cf. chapitre de Trigonométrie).

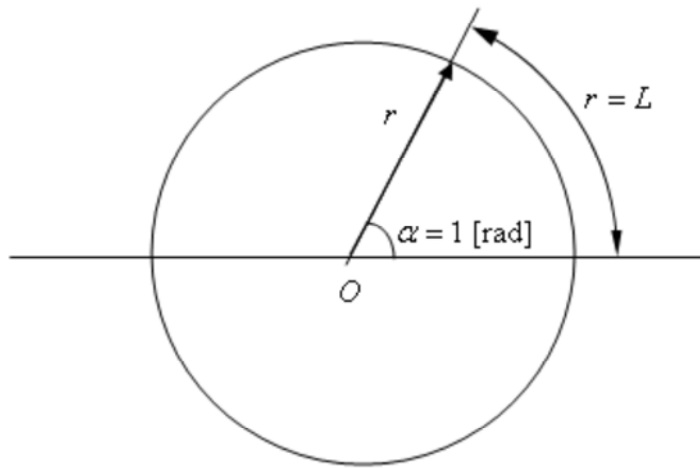


Figure: 21.17 - Définition du radian

Ainsi, en radians, un angle plat est égal à π et tous les autres angles sont des multiples réels de cette constante.

BISSECTRICE

Maintenant que nous savons comparer, additionner et mesurer des angles nous allons pouvoir nous pencher sur un concept important en géométrie qu'est celui de "bissectrice" et de quelques propriétés y relatives que nous réutiliserons plus loin pour des théorèmes importants.

Définitions:

D1. Nous appelons "**bissectrice**" la droite qui divise un angle en deux parties égales.

D2. Nous appelons "**demi-bissectrice**" la demi-droite qui divise un angle en deux parties égales.

Maintenant voyons quelques propriétés importantes de la bissectrice:

Deux droites AB et CD qui se coupent forment comme nous le savons déjà intuitivement, quatre angles:

$$\widehat{AOC}, \widehat{COB}, \widehat{BOD}, \widehat{DOA} \quad (21.40)$$

Les angles $\widehat{AOC}$, $\widehat{BOD}$ de même que les angles $\widehat{BOC}$, $\widehat{DOA}$ dont les côtés sont opposés, sont dits "**angles opposés par le sommet**".

Trivialement, si θ est la mesure de l'angle $\widehat{AOC}$, la mesure de l'angle adjacent $\widehat{COB}$ est $\pi - \theta$; celle de l'angle $\widehat{DOB}$ adjacent au précédent est $\pi - (\pi - \theta)$; celle de $\widehat{DOA}$ est $\pi - \theta$.

Soit OE la demi-bissectrice de l'angle $\widehat{AOC}$, OG celle de $\widehat{COB}$, OF celle de $\widehat{BOD}$, OH celle de $\widehat{DOA}$, nous aurons:

$$\widehat{COE} = \frac{\theta}{2}, \widehat{COG} = \frac{\pi - \theta}{2} \quad (21.41)$$

et, par conséquent:

$$\widehat{GOE} = \frac{\pi}{2} \quad (21.42)$$

Nous aurions de même:

$$\widehat{GOF} = \widehat{FOH} = \widehat{HOE} = \pi/2 \quad (21.43)$$

Nous résumons ainsi tous ces résultats et propriétés de mesure:

P1. Deux angles opposés par le sommet sont égaux.

P2. Deux angles opposés par le sommet ont même bissectrice.

P3. Les bissectrices de deux angles adjacents supplémentaires sont rectangulaires.

P4. Les bissectrices des angles formés par deux sécantes sont deux droites rectangulaires (à angle droit).

TRIANGLES

Nous avons étudié jusqu'à présent le concept de dimensions, de point, de segment, de ligne, d'angle et de plan ouvert (infini). Cependant un plan peut-être délimité par plusieurs lignes pour obtenir ainsi des formes géométriques (planes) dont les plus simples peuvent être considérées comme les triangles.

Définition: Nous appelons "**triangle**" la figure formée par trois segments AB , BC , CA , les points A, B, C n'étant pas alignés. Les segments AB , BC , CA , sont les "**côtés**" du triangle. Les points A, B, C sont les "**sommets**" du triangle. L'angle saillant $\widehat{BAC}$, qui contient tous les points du côté BC , s'appelle angle $\hat{A}$ du triangle et BC est alors son "**côté opposé**".

Remarque: Nous employons la notation $\hat{A}$ lorsqu'aucune confusion n'est possible; à défaut, nous utilisons la notation $\widehat{BAC}$ avec le même sens.

Il y a six éléments dans un triangle, à savoir: trois angles $\hat{A}, \hat{B}, \hat{C}$ et trois côtés AB, BC, CA .

Nous désignerons par $a = BC$, $b = AC$, $c = AB$ les longueurs des côtés mesurées avec la même unité; par $\hat{A}, \hat{B}, \hat{C}$, les mesures des angles.

La somme des angles d'un triangle plan est toujours égale à 180° (ou π radians). La démonstration est assez simple.

Démonstration:

Sur la figure ci-dessous, ABC est un triangle quelconque, et D la parallèle à (BC) qui passe par A . Nous observons:

1. Les angles bleus ont même mesure car ils sont alternes-internes (les droites (BC) et D étant parallèles).

Remarque: Pour la démonstration de l'égalité des angles alternes-internes voir plus loin le quatrième axiome d'Euclide.

2. De même, les angles verts ont même mesure car ils sont alternes-internes.
3. Nous remarquons que la somme des angles bleu + rouge + vert forme un angle plat en A, puisque D est une droite. Donc angle bleu + angle rouge + angle vert = 180° .
4. D'après les égalités d'angles constatées en (1) et (2), nous déduisons que:

$$\hat{A} + \hat{B} + \hat{C} = 180^\circ$$

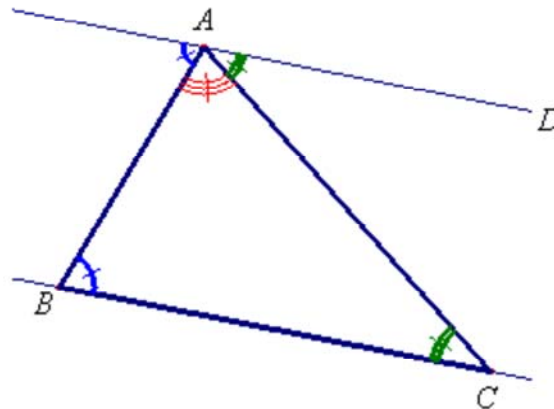


Figure: 21.18 - Somme des angles d'un triangle plan

Cette démonstration étant valable quel que soit le triangle tracé dans le plan.

□C.Q.F.D.

TRIANGLES ÉGAUX

Définitions: Nous disons que deux triangles sont des "**triangles égaux**" lorsque nous pouvons par un déplacement soit par un retournement ou les deux combinés, superposer tous les sommets du premier triangle au deuxième. Nous disons alors aussi que les triangles sont des "**triangles homologues**" ou "**triangles isométriques**".

De cette définition, il vient que deux triangles sont égaux lorsque soit:

1. Ils ont un côté égal et deux angles égaux;
2. Ils ont un angle de même mesure compris entre deux côtés de même longueur.

Démonstrations:

Premier cas d'égalité: Deux triangles qui ont un côté égal $BC=B'C'$, compris entre deux angles égaux $\hat{B} = \hat{B}'$, $\hat{C} = \hat{C}'$, sont égaux.

En d'autres termes, si 2 triangles ont 2 angles égaux 2 à 2, alors les 3èmes angles le sont aussi. Ceci étant dit nous pouvons maintenant prendre comme côté égal celui qui est situé entre 2 angles égaux

sans perte de généralité.

Puisque $BC=B'C'$ (voir figure ci-dessous), il existe un déplacement qui amène B' en B et C' en C . Ce déplacement amène A' en A_1 situé du même côté par rapport à la droite BC , ou en A_2 symétrique de A_1 par rapport à cette droite.

Les deux demi-droites BA et BA_1 font par hypothèse, avec BC le même angle $\hat{B} = \hat{B}'$. Comme elles sont par construction d'un même côté de BC , elles sont confondues. Les deux demi-droites CA et CA_1 sont confondues pour la même raison et parce que $\hat{C} = \hat{C}'$. A_1 est donc confondu avec A . Les deux triangles $ABC, A'B'C'$ sont donc égaux.

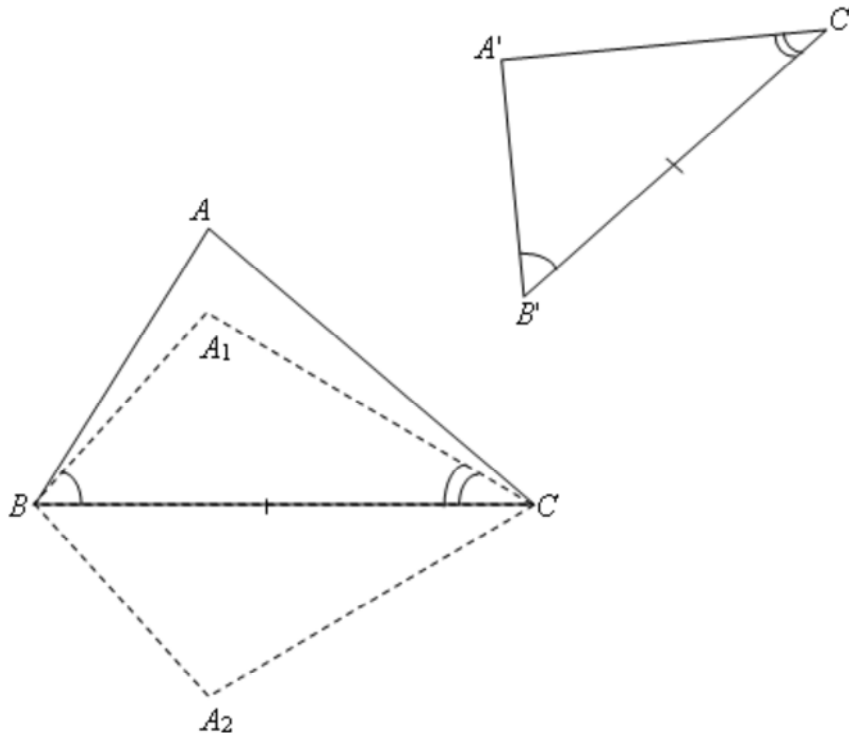


Figure: 21.19 - Deux triangles ayant côté égal et deux angles égaux

Deuxième cas d'égalité: Deux triangles qui ont un angle égal $\hat{A} = \hat{A}'$ compris entre deux côtés égaux $AB=A'B', AC=A'C'$ sont égaux.

Puisque $AB=A'B'$ (voir figure ci-dessous) il existe un déplacement situé par rapport à AB du même côté que le point. S'il l'amenait en C_2 symétrique de C_1 par rapport à AB , un demi-tour autour de AB l'amènerait en C_1 . Les demi-droites AC, AC_2 situées d'un même côté de AB font, par hypothèse le même angle avec AB , puisque $\hat{A} = \hat{A}'$. Elles sont donc confondues. L'hypothèse $AC = AC_1$ entraîne alors que C_1 et C sont confondus. Les deux triangles $ABC, A'B'C'$ sont donc égaux.

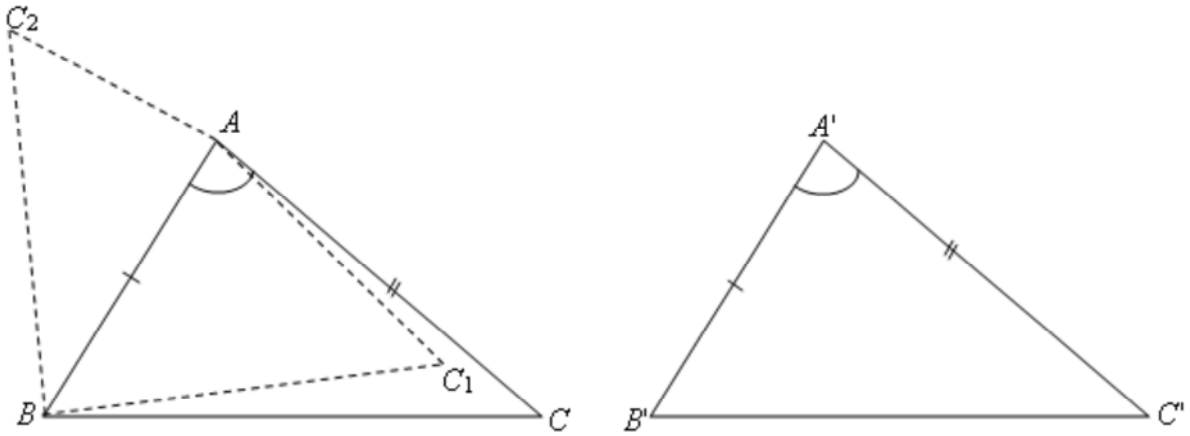


Figure: 21.20 - Déplacement montrant l'égalité des 2 triangles

□C.Q.F.D.

TRIANGLES ISOCÈLES

Définition: Nous disons qu'un triangle ABC est un "**triangle isocèle**" lorsque deux de ses côtés sont AB et AC sont égaux ("iso" signifiant "même"). Le troisième côté BC est alors appelé la "**base**" de ce triangle.

Remarque: Nous disons qu'un triangle est "**scalène**" quand il possède 3 côtés inégaux.

Définition: Nous appelons "**médiatrice**" d'un segment BC , la perpendiculaire à la droite BC au point H de cette droite, milieu de BC .

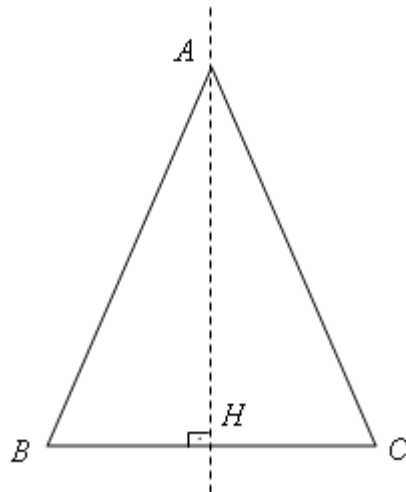


Figure: 21.21 -Représentation de la médiatrice

Théorème: Dans un triangle isocèle ABC comme représenté ci-dessus:

1. Les angles $\hat{B}$ et $\hat{C}$ opposés aux côtés égaux sont égaux.
2. La médiatrice de BC et la bissectrice de l'angle $\hat{A}$ sont confondues (figure ci-dessus).

Démonstration:

Les deux triangles BAH , CAH définis par la bissectrice de $\hat{A}$ ont un angle égal $\widehat{BAH} = \widehat{CAH}$ par construction compris entre deux côtés égaux: AH qui est commun et $AB=AC$ par hypothèse. Comme les angles $\widehat{BHA} = \widehat{CHA}$ sont droits et égaux et que la somme des angles d'un triangle est égale à un angle plat, alors les angles $\widehat{HBA}$ et $\widehat{ACH}$ sont donc égaux.

□C.Q.F.D.

Théorème: Le lieu géométrique des points M équidistants (à même distance) de deux points B et C donnés est la médiatrice (D) du segment BC .

Remarque: Nous appelons "lieu géométrique" d'un point M , assujetti à des conditions, l'ensemble des positions occupées par le point M .

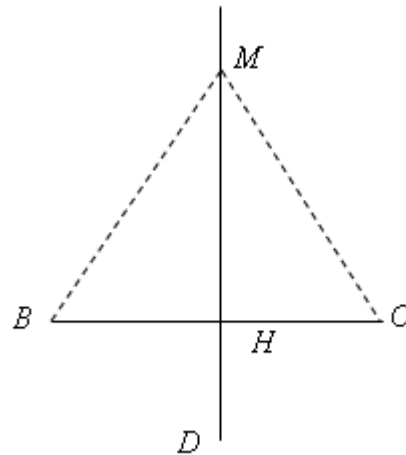
Démonstration:

Figure: 21.22 - Représentation du lieu géométrique

1. Tout point du lieu est sur la droite (D). Autrement dit, l'hypothèse $MB=MC$ entraîne que M est la médiatrice de BC . En effet, si $MB=MC$, le triangle MBC est isocèle et le sommet M est sur la médiatrice de BC .
2. Tout point de (D) est un point du lieu. Ce qui revient à dire que si M est sur la médiatrice de BC , nous avons $MB=MC$. En effet, si M est sur la droite (D) qui rencontre en H , milieu de BC , la droite BC , les triangles MHB , MHC sont égaux (deuxième cas d'égalité: $\widehat{BHM} = \widehat{CHM}$ parce que ces angles sont droits, HM commun; $HB=HC$ parce que H est le milieu de BC): les côtés MB , MC sont donc égaux et nous avons bien $MB=MC$. Le point M est un point du lieu.

□C.Q.F.D.

A l'aide de ce théorème nous pouvons en énoncer un second: Par un point A pris hors d'une droite BC , nous pouvons mener à cette droite une seule et unique perpendiculaire:

Démonstration:

1. Soit un triangle ABC , faisons subir à ce triangle un demi-tour autour de BC (symétrie horizontale): A vient en A' symétrique de A par définition, par rapport à BC . Puisque les figures ABC , $A'BC$ sont égales, $AB=A'B$ et $AC=A'C$. BC est donc la médiatrice de AA' et les droites BC et AA' sont perpendiculaires. AA' est donc bien une perpendiculaire à BC menée par A .

2. Nous ne pouvons en mener plusieurs: soit AH une perpendiculaire menée de A à BC , elle rencontre la droite BC en H qui est différent soit de B , soit de C . Supposons que H soit différente de B . $\widehat{AHB}$, $\widehat{A'HB}$ qui se déduisent l'un de l'autre par retournement, sont égaux, et, comme chacun d'eux est droit, l'angle $\widehat{AHA'}$ est plat. La droite AH est confondue avec la droite AA' et est donc bien la seule perpendiculaire à BC qui passe par A .

□C.Q.F.D.

Définitions:

D1. Nous appelons "**projection orthogonale**" d'un point A sur une droite BC le point H où la perpendiculaire menée par A à cette droite la rencontre. Le point H s'appelle aussi "**pied**" de cette perpendiculaire.

D2. Nous appelons "**distance géométrique**" du point A à la droite BC la longueur du segment AH .

Puisque BC est la médiatrice de AA' , H est le milieu de AA' ; donc: Un point A et son symétrique A' par rapport à une droite (D) sont caractérisés par les deux propriétés suivantes que nous ne démontrerons pas car suffisamment intuitives:

P1. AA' est perpendiculaire à (D) .

P2. Le milieu de AA' est sur (D) .

La droite AB , qui joint le point A à un point de la droite BC , autre que le pied H de la perpendiculaire menée de A à cette droite, s'appelle "**droite oblique**". Le point B s'appelant "**pied l'oblique**".

TRIANGLES ÉQUILATERAUX

Définition: Nous disons qu'un triangle ABC est un "**triangle équilatéral**" lorsque tous ses côtés sont de longueur égales ou que tous ses angles $\hat{A}$, $\hat{B}$, $\hat{C}$ sont égaux. Chacun de ces côtés est donc une base.

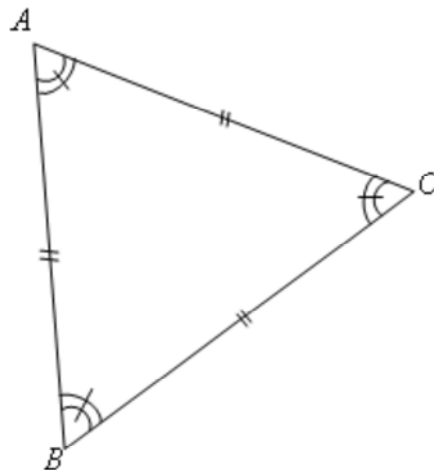


Figure: 21.23 -Exemple de triangle équilatéral

Remarque: Nous prenons pour habitude d'annoter les côtés égaux par deux traits parallèles disposés au milieu des côtés.

Comme la somme des trois angles de ce triangle doit faire 180° (en degrés) et que les trois angles ont même mesure, chacun d'eux mesure donc: $180^\circ/3$ soit 60° .

TRIANGLES RECTANGLES

Définition: Un "triangle rectangle" est un triangle ABC qui a un angle droit:

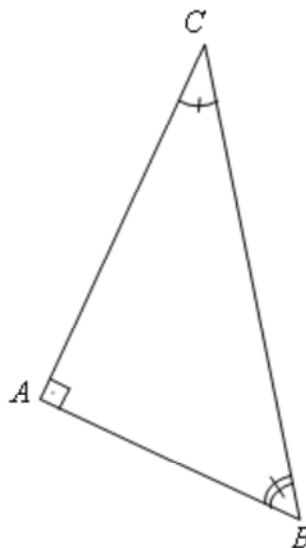


Figure: 21.24 -Exemple de triangle rectangle

Dire que le triangle est rectangle en A signifie que c'est en $\hat{A}$ que se trouve l'angle droit.

Remarque: Dans un triangle rectangle, le côté le plus grand est toujours le côté opposé à l'angle droit. Nous l'appelons "l'hypoténuse". Nous démontrons cette propriété avec le théorème de Pythagore (voir plus bas).

TRIANGLES RECTANGLES-ISOCÈLES

Définition: Un "triangle rectangle-isocèle" ABC est à la fois rectangle et isocèle, ce qui signifie qu'il a à la fois un angle droit et deux côtés de même longueur.

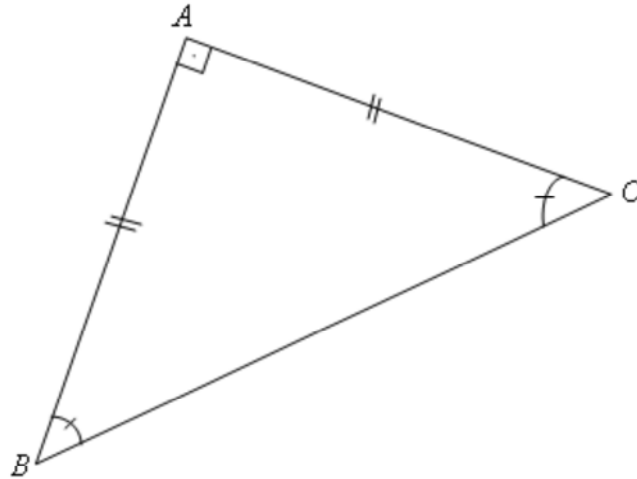


Figure: 21.25 -Exemple de triangle rectangle-isocèle

Remarque: Le sommet principal correspond à l'angle droit. En effet, comme BC , l'hypoténuse, doit être le côté le plus grand, ce sont les côtés AB et AC qui ont même longueur (plus petite).

INÉGALITÉS DANS LES TRIANGLES

Voyons maintenant quelques inégalités (propriétés) intéressantes dans le triangle.

P1. Montrons d'abord que dans tout triangle, un côté opposé à un angle droit ou obtus (supérieur à 90° donc...) est supérieur à chacun des deux autres côtés du triangle.

Démonstration:

Considérons le triangle ABC ci-dessous dans lequel $\hat{A} \geq 90^\circ$ et soit Cx le prolongement du côté BC . Portons, sur la demi-droite Bx , une longueur $BD = BA$ pour construire un triangle isocèle dans le triangle initial:

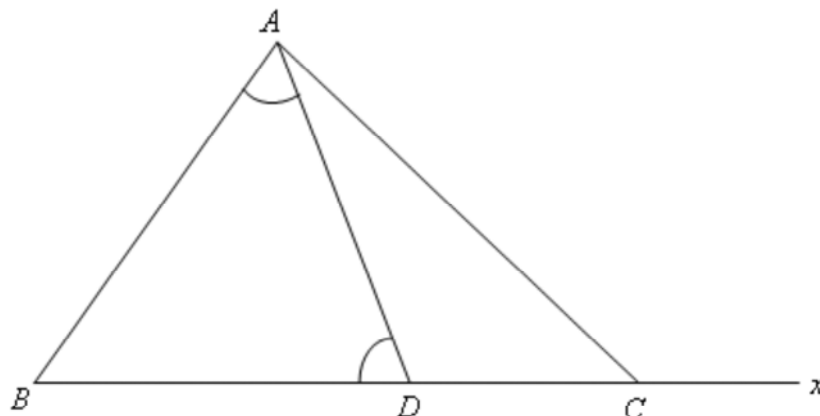


Figure: 21.26 - Construction d'un triangle isocèle

Le triangle BAD est donc un triangle isocèle de base AD dont l'angle à la base BAD est bien évidemment aigu (inférieur à 90°).

Donc par construction $\widehat{BAD} < \widehat{BAC}$. La droite AD est intérieure par construction à l'angle $\widehat{BAC}$ et par suite:

$$BC > BD \quad (21.44)$$

et comme $BD = BA$ par construction nous avons donc:

$$BC > BA \quad (21.45)$$

Ce qui termine notre démonstration. Car la démarche est la même pour montrer que $BC > CA$.

□C.Q.F.D.

P2. Dans tout triangle dont les côtés ont des longueurs strictement croissantes, un côté est toujours inférieur à la somme des deux autres.

Démonstration:

Supposons que dans le triangle ABC ci-dessous les côtés $BC = a, CA = b, AB = c$ soient tels que $a > b > c$:

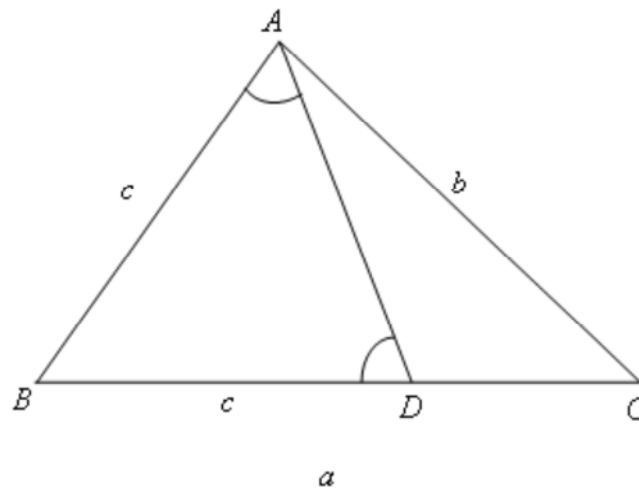


Figure: 21.27 - Exemple de triangle avec côtés strictement croissants

Soit D le point du côté BC tel que $BD = BA = c$ soient les côtés du triangle isocèle construit ABD . Nous obtenons:

$$DC = BC - BD = a - c \quad (21.46)$$

Le triangle ABD étant isocèle, l'angle à la base $\widehat{ADB}$ est aigu et son supplément $\widehat{ADC}$ est obtus. Dans le triangle ADC , nous obtenons d'après la propriété P1 précédente que $DC < CA$, c'est-à-dire:

$$a - c < b \quad (21.47)$$

ou:

$$a < b + c \quad (21.48)$$

il s'agit de la fameuse "inégalité triangulaire" sous forme géométrique. Nous la retrouverons dans de nombreux autres chapitres du site dans des espaces et des concepts mathématiques plus abstraits.

La propriété est alors immédiate pour les autres côtés b et c par permutation de la méthode:

$$\begin{aligned} a &< b + c \\ b &< a + c \quad (21.49) \\ c &< a + b \end{aligned}$$

□C.Q.F.D.

P3. Dans tout triangle un côté quelconque est supérieur à la différence des deux autres.

Démonstration:

Supposons que nous ayons $a > b > c$. En retranchant c aux deux membres de l'inégalité:

$$a + c > b \quad (21.50)$$

il vient immédiatement:

$$a > b - c \quad (21.51)$$

La propriété est alors immédiate pour les autres côtés b et c par permutation de la méthode:

$$\begin{aligned} a &> b - c \\ b &> a - c \quad (21.52) \\ c &> a - b \end{aligned}$$

En définitive puisque:

$$a > b - c \text{ et } a < b + c \quad (21.53)$$

pour tout triangle à côtés croissants nous avons:

$$\boxed{b - c < a < b + c} \quad (21.54)$$

□C.Q.F.D.

THÉORÈME DE PYTHAGORE

Maintenant que nous avons vu ce qu'était le triangle et certaines de ses propriétés ainsi que le concept d'angle, nous pouvons démontrer le fameux "théorème de Pythagore" (qui donne donc la relation que doivent satisfaire trois nombres qui représentent les côtés d'un triangle rectangle) et faire de la trigonométrie du cercle (cf. chapitre de Trigonométrie).

Il existe plusieurs démonstrations de ce théorème dont en voici une parmi tant d'autres:

Démonstration:

Soit un carré (4 angles droit) dans lequel est inscrit un autre carré, nous déterminons la surface du carré inscrit à partir des triangles rectangles résultants de l'espace vide entre les deux carrés tel que présenté sur la figure ci-dessous:

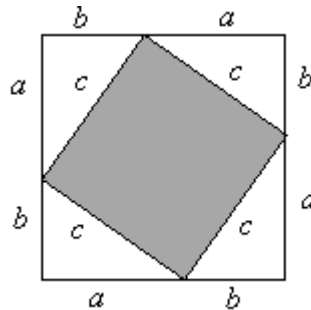


Figure: 21.28 - Construction pour la démonstration du théorème de Pythagore dans le plan

La surface du carré blanc est bien sûr:

$$(a+b)^2 = a^2 + 2ab + b^2 \quad (21.55)$$

Pour avoir la surface du carré gris on peut soustraire au carré blanc la surface des 4 triangles rectangles (d'une surface de moitié de celle d'un quadrilatère de même longueur et hauteur), chacun de surface:

$$ab/2 \quad (21.56)$$

La surface du carré gris est donc finalement:

$$a^2 + 2ab + b^2 - 4 \cdot \frac{ab}{2} = a^2 + b^2 \quad (21.57)$$

Le résultat obtenu étant équivalent au carré du côté de la surface grise, nous avons au final le résultat du fameux "**théorème de Pythagore**":

$$\boxed{a^2 + b^2 = c^2} \quad (21.58)$$

Dans un triangle rectangle, le carré de l'hypoténuse (côté opposé à l'angle droit) est égal à la somme des carrés des côtés de l'angle droit.

□C.Q.F.D.

Dans le cas particulier où nous avons trois nombres entiers a , b et c qui satisfont le théorème de Pythagore (il existe une infinité de combinaisons de nombres entiers satisfaisant le théorème de Pythagore), nous parlons alors de "**triplet pythagoricien**".

Remarque: C'est au chinois Tchao Kiung K'ing (2ème siècle) que l'on devrait cette démonstration.

Il est souvent fait mention dans les petites classe de la réciproque du théorème de Pythagore qui nous

dit: Dans un triangle, si le carré d'un côté est égal à la somme des carrés des 2 autres côtés, alors ce triangle est rectangle et l'hypoténuse sera alors le plus long côté du triangle.

THÉORÈME DE THALÈS

Ayant démontré le théorème de Pythagore et maintenant que les concepts de parallèles, segments, angles et autres nous sont connus, nous pouvons enfin démontrer le théorème de Thalès dont voici une possible démonstration qui nécessite d'abord le développement de deux lemmes:

L1. Triangles de même surface

Soit la figure:

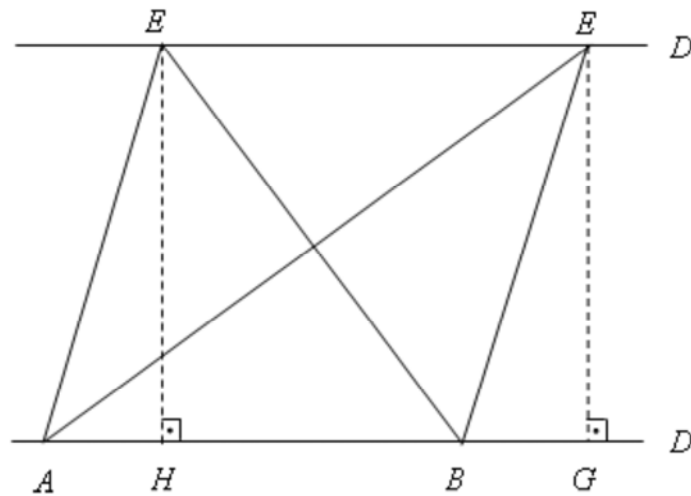


Figure: 21.29 - Construction pour la démonstration du théorème de Thalès

Nous avons:

$$\begin{aligned} D &\parallel D' \\ \overline{EH} &\perp D \quad (21.59) \\ \overline{FG} &\perp D \end{aligned}$$

$EFGH$ est un rectangle car ses côtés sont parallèles deux à deux et il a au moins deux angles droits. Donc ses côtés opposés ont même longueur: $EH=FG$.

$\overline{EH}$ est la hauteur relative à $\overline{AB}$ dans le triangle EAB et FG est la hauteur relative à $\overline{AB}$ dans le triangle FAB .

La surface du triangle ne dépend que de la longueur du côté et de la longueur de la hauteur relative à ce côté. Pour les deux triangles EAB et FAB , ces longueurs sont égales, donc ils ont la même surface.

Conclusion: Si deux triangles ont un côté commun et si les troisièmes sommets sont sur une parallèle à ce côté commun, alors ils ont la même surface.

L2. Rapports égaux

Soit le rapport de proportions ("calcul proportionnel" ou "produit en croix"):

$$\frac{a}{b} = \frac{c}{d} \quad (21.60)$$

alors:

$$ad = bc \quad (21.61)$$

Si $ad=bc$, alors $ad+cd=bc+cd$ (nous ajoutons un même nombre positif ou négatif aux deux membres). D'où après factorisation:

$$d(a+c) = c(b+d) \quad (21.62)$$

et en appliquant inversement la règle des produits en croix:

$$\frac{a+c}{b+d} = \frac{c}{d} \quad (21.63)$$

Exposons maintenant en quoi consiste le théorème de Thalès:

Soit la figure:

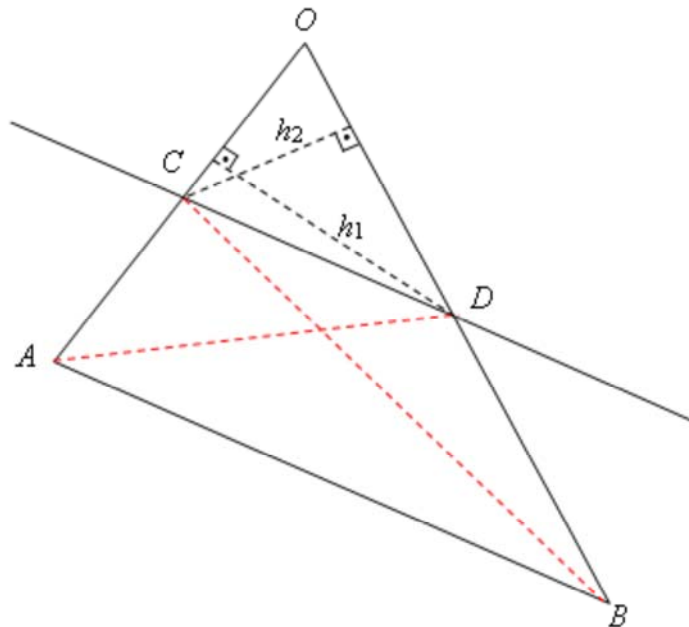


Figure: 21.30 - Première approche du théorème de Thalès

Avec:

$$\begin{aligned} AB &\parallel CD \\ h_1 &\perp OA \\ h_2 &\perp OB \end{aligned} \quad (21.64)$$

Nous avons montré précédemment que si deux triangles ont un côté commun et si les troisièmes sommets sont sur une parallèle, alors ils ont la même surface. Donc les triangles ACD et BCD ont la même surface.

En ajoutant à chacune de ces deux surfaces celle du triangle OCD , nous obtenons que les triangles ODA et OCB ont la même surface.

Nous en déduisons qu'en utilisant à nouveau le rapport en croix:

$$\frac{S_{OCD}}{S_{OAD}} = \frac{S_{OCD}}{S_{OCB}} \quad (21.65)$$

Soit h_1 la hauteur issue de D dans le triangle OCD et h_2 la hauteur issue de C dans le triangle OCD :

$$\frac{S_{OCD}}{S_{OAD}} = \frac{OC \cdot \frac{h_1}{2}}{OA \cdot \frac{h_1}{2}} = \frac{OC}{OA} \quad \text{et} \quad \frac{S_{OCD}}{S_{OCB}} = \frac{OD \cdot \frac{h_2}{2}}{OB \cdot \frac{h_2}{2}} = \frac{OD}{OB} \quad (21.66)$$

Conclusion:

$$\boxed{\frac{OC}{OA} = \frac{OD}{OB}} \quad (21.67)$$

Soit maintenant la figure:

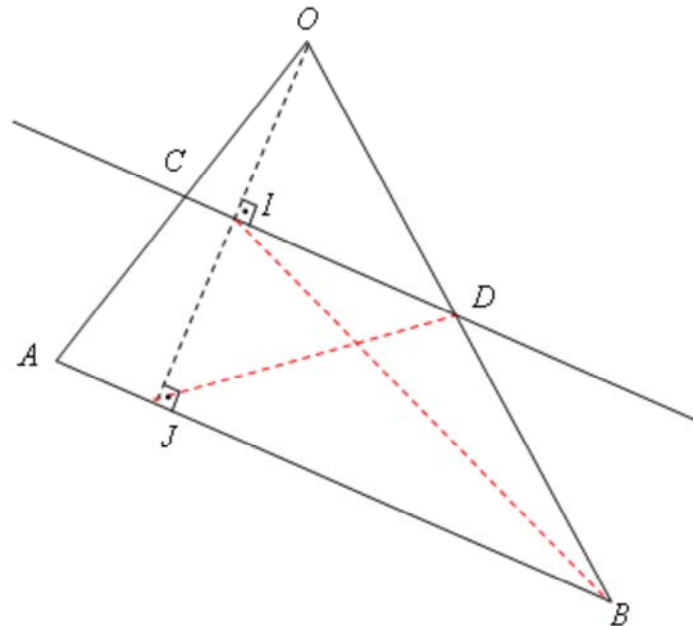


Figure: 21.31 - Deuxième approche du théorème de Thalès

Les triangles IJD et IDB ont la même surface d'après le lemme 1, ainsi que les triangles OJD et OIB donc :

$$S_{OJD} = \frac{1}{2} OJ \cdot DI \quad S_{OJB} = \frac{OJ \cdot JB}{2} - \frac{JI \cdot JB}{2} = \frac{1}{2} OI \cdot BJ \quad (21.68)$$

d'où:

$$OJ \cdot DI = OI \cdot BJ \quad (21.69)$$

et donc:

$$\boxed{\frac{OI}{OJ} = \frac{DI}{BJ}} \quad (21.70)$$

De la même manière dans les triangles OIA et OCJ , nous obtenons:

$$\frac{OI}{OJ} = \frac{CI}{AJ} \quad (21.71)$$

D'après le lemme 2, comme:

$$\boxed{\frac{OI}{OJ} = \frac{DI}{BJ} = \frac{CI}{AJ}} \quad (21.72)$$

alors:

$$\frac{OI}{OJ} = \frac{DI}{BJ} = \frac{CI}{AJ} = \frac{DI + CI}{BJ + AJ} = \frac{DC}{BA} \quad (21.73)$$

En appliquant au triangle OJB , ce que nous avons fait selon la première approche, nous avons aussi:

$$\frac{OI}{OJ} = \frac{OD}{OB} \quad (21.74)$$

Donc finalement en reprenant tous les résultats obtenus:

$$\boxed{\frac{OI}{OJ} = \frac{DC}{BA} = \frac{OD}{OB}} \quad (21.75)$$

qui constitue le "**théorème de Thalès**" des rapports.

PARALLÉLISME

Définition: Nous appelons "**parallèles**" deux droites également distantes l'une de l'autre sur toute leur longueur.

Ce concept est directement relié au cinquième postulat d'Euclide et est souvent considéré comme le plus important ayant été montré qu'il ne peut être considéré comme un axiome car n'étant pas respecté dans les géométries non-euclidiennes (cf. chapitre de Géométries Non-Euclidiennes)

Les conséquences de ce postulat sont les suivantes dans une géométrie euclidienne:

1. Si deux droites (AB) et (CD) sont parallèles, toute droite $(E'F')$ qui coupe l'une coupe l'autre.

Démonstration:

Soit F le point commun à la droite (CD) et à la droite $(E'F')$: si la droite $(E'F')$ ne coupait pas la droite (AB) , elle lui serait parallèle, et par le point F passeraient deux droites (CD) et $(E'F')$ parallèles à une

même troisième (AB), ce qui n'est pas le cas. Donc, la droite ($E'F'$), coupe la droite (AB).

□C.Q.F.D.

2. Deux droites (AB) et (CD) parallèles à une même troisième ($E'F'$) sont parallèles entre elles.

Démonstration:

Si la droite (CD) n'était pas parallèle à la droite (AB), elle la couperait: elle couperait aussi la droite ($E'F'$) parallèle à la droite (AB), elle ne serait donc pas parallèle à ($E'F'$).

□C.Q.F.D.

Théorème: Si deux droites parallèles sont coupées par une sécante:

1. Les angles alternes-internes sont égaux;
2. Les angles alternes-externes sont égaux;
3. Les angles correspondants sont égaux.

Démonstration:

Soient deux parallèles AB et CD et la sécante EF :

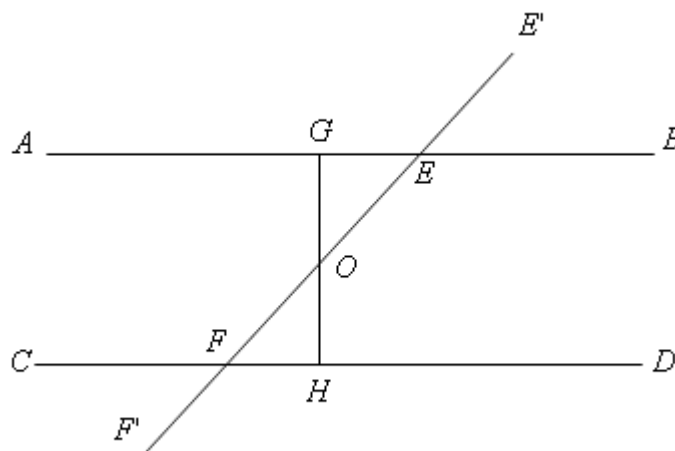


Figure: 21.32 - Deux droites parallèles coupées par une sécante

1. Par le milieu O de EF menons la perpendiculaire GH à AB , qui est aussi perpendiculaire à CD . Les triangles rectangles EOG et FOH ont un angle aigu égal à, $\widehat{GOE} = \widehat{FOH}$ et l'hypoténuse égale, $OF=OE$. Ils sont égaux, et les angles $\widehat{OFH}$ et $\widehat{GEO}$ sont égaux.

2. Les angles alternes-externes $\widehat{E'EB}$ et $\widehat{F'FC}$ sont égaux, car $\widehat{E'EB}$ est opposé par le sommet à l'angle $\widehat{GEO}$, qui est alterne-interne avec l'angle $\widehat{OFH}$.

□C.Q.F.D.

Réciproque: Si deux droites sont coupées par une sécante qui forme avec ces droites:

- Soit deux angles-alternes internes égaux;
- Soit deux angles-alternes externes égaux;
- Soit deux angles correspondants égaux, ces deux droites sont parallèles.

alors ces deux droites sont parallèles.

Remarque: Pour démontrer le parallélisme de deux droites, il faut et il suffit que les angles alternes-internes, alternes-externes ou correspondants, formés par ces deux droites avec une sécante, soient égaux.

CERCLES

Définition: Nous appelons "**cercle**" le lieu géométrique des points M du plan qui sont à une distance donnée R , appelée "**rayon**" de ce cercle, d'un point fixe O , appelé "**centre**" de ce cercle.

$$OM=R \quad (21.76)$$

Remarque: Le mot "rayon" désigne soit le segment OM , soit sa mesure R .

Les cercles sont directement concernés par le troisième postulat d'Euclide que nous avons énoncé plus haut.

Nous appelons "**diamètre**" d'un cercle toute droite qui passe par le centre O du cercle. Tout diamètre rencontre le cercle en deux points A et B , définis par $OA=OB=R$, que nous appelons "**extrémités du diamètre**". Nous réservons la notation "diamètre AB " pour le diamètre d'extrémités A et B . Nous disons que deux points d'un cercle sont "**diamétralement opposés**" quand ils sont les deux extrémités d'un même diamètre.

Un cercle divise le plan en deux régions: celle qui contient le centre, que nous appelons "**région intérieure**", et celle qui ne le contient pas, que nous appelons "**région extérieure**".

Théorème: La condition nécessaire et suffisante pour qu'un point P soit strictement intérieur à un cercle (O), de centre (O) et de rayon R , est $OP < R$.

Démonstration:

1. La condition est nécessaire: Si, par hypothèse, P est à l'intérieur du cercle (O), il est situé soit en O , soit entre les extrémités A et B du diamètre délimité par le lieu géométrique des points M . S'il est en O , la proposition est évidente, s'il n'est pas en O , il est entre O et A par exemple, et nous avons $OP < OA$, c'est-à-dire $OP < R$.

2. La condition est suffisante: Si, par hypothèse $OP < R$, P se trouve entre les extrémités A et B des lieux géométriques définis par les points M , donc à l'intérieur du cercle (O).

□C.Q.F.D.

Corollaire: la condition nécessaire et suffisante pour que P soit extérieur au cercle (O) est $OP > R$.

Nous appelons "corde" CD d'un cercle le segment dont les extrémités C et D sont sur ce cercle.

Théorème: La médiatrice d'une corde CD est un diamètre.

Démonstration:

La médiatrice (Δ) de CD (voir figure ci-dessous), corde du cercle (O) de centre O et de rayon R , passe par le point O parce que comme nous l'avons démontré lors de notre étude des triangles nous avons $OC = OD = R$.

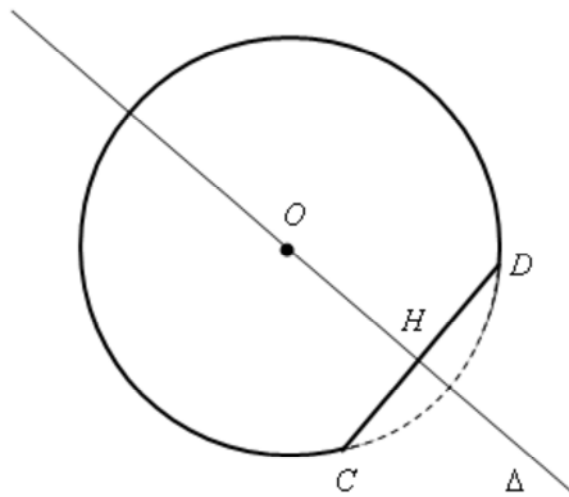


Figure: 21.33 - La médiatrice d'une corde

□C.Q.F.D.

Corollaire: La perpendiculaire menée par le centre O d'un cercle à une corde CD passe par le milieu H de cette corde.

Théorème: Par trois points A, B, C non alignés, il passe un cercle et un seul.

Démonstration (voir figure ci-dessous):

Tracer la médiatrice (D) de AB et la médiatrice (Δ) de AC . Si (D) et (Δ) étaient parallèles, la perpendiculaire AB à (D) serait perpendiculaire à (Δ), donc confondue avec AC . ABC seraient alignés. Donc (D) et (Δ) non parallèles se coupent en un point O :

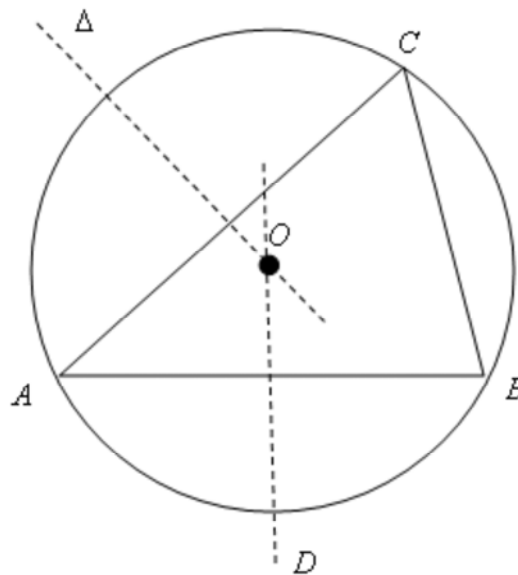


Figure: 21.34 - Triangle inscrit dans un cercle

1. Il passe un cercle par A, B, C : le point O étant sur (D) , médiatrice de AB , $OA=OB$ par définition: le point O étant sur (Δ) , médiatrice de AC , $OA=OC$. Le cercle (O) , de centre O et de rayon OA , passe par B (puisque $OA=OB$) et par C (puisque $OA=OC$). Il passe donc par A, B, C .
2. Il ne passe A, B, C qu'un seul cercle: S'il passait par A, B, C un cercle différent du cercle (O) de centre O et de rayon OA , son centre O' se trouverait sur la médiatrice de AB et de AC qui sont deux cordes de ce cercle; il serait donc confondu avec O .

□C.Q.F.D.

Remarque: La médiatrice de BC , corde du cercle (O) , passe par le point O . Nous pouvons donc dire (résultat important) que les trois médiatrices des côtés d'un triangle ABC concourent.

AXIOMES DE HILBERT

Euclide a rassemblé toutes les connaissances géométriques de son temps sous la forme de ces cinq postulats. Il a laissé son nom à la géométrie euclidienne qui utilise son cinquième postulat, à la géométrie non-euclidienne qui ne l'utilise pas, et aux espaces euclidiens.

Cette base postulée est néanmoins imparfaite, pour démontrer rigoureusement les théorèmes associés à cette géométrie, il est nécessaire d'admettre comme vrai des hypothèses supplémentaires implicites. David Hilbert construisit une axiomatique correspondante à l'idée que se faisait Euclide de la géométrie en ajoutant les hypothèses ad hoc.

Les axiomes de Hilbert sont eux regroupés en cinq catégories: l'association, l'ordre, la congruence, la continuité et les parallèles.

Trois concepts sont associés à cette axiomatique:

1. Celui de l'association définit le mot "contient", il correspond aux notions "est élément de" et "est inclus dans" de la théorie des ensembles.

2. Celui de "l'ordre" correspond à "une relation binaire" entre un couple de points et un point, il apparaît dans les expressions "entre" et permet de définir les segments.

3. La congruence, qui correspond à trois "relations d'équivalence" pour les couples de points, les triangles et les angles.

Remarque: Les points, droites et plans sont considérés comme distincts par défaut.

Voici donc les "axiomes de Hilbert":

AXIOMES D'ASSOCIATIONS (A)

A.A1. Soient deux points, il existe une droite passant par ces deux points.

A.A2. Soient deux points, il n'existe qu'une unique droite passant par ces deux points (in extenso la droite décrite en A.A1) est unique.

A.A3. Une droite contient au moins deux points, et pour une droite donnée, il existe au moins un point non contenu dans la droite.

A.A4. Soient trois points non contenus dans une droite, il existe un plan contenant ces trois points. Tout plan contient au moins un point.

A.A5. Soient trois points non contenus dans une droite, il n'existe qu'un unique plan contenant ces trois points.

A.A6. Soient deux points contenus dans une droite D et dans un plan A , alors A contient tous les points de D .

A.A7. Si deux plans A et B contiennent tous deux un point C , alors l'intersection de A et B contient au moins un autre point.

A.A8. Il existe au moins quatre points non coplanaires.

AXIOMES D'ORDRE (O)

A.O1. Si un point B est entre les points A et C , B est aussi entre les points C et A , et il existe une droite contenant les trois points A, B, C .

A.O2. Soient deux points A et C , il existe un point B élément de la droite AC tel que C se situe entre A et B .

A.O3. Soient trois points contenus dans une droite, alors un et un seul se situe entre les deux autres.

A.O4. ("Axiome de Pasch") Soient trois points A, B, C non colinéaires, et soit une droite D contenue dans le plan ABC mais ne contenant aucun des points A, B, C : Si D contient un point du segment AB , alors D contient aussi soit un point du segment AC soit un point du segment BC .

AXIOMES DE CONGRUENCE (G)

Remarque: Intuitivement "congruent" signifie en géométrie "superposable".

A.G1. Soient deux points A, B et un point A' élément d'une droite d , il existe deux et deux uniques points C et D , tel que A' se situe entre C et D , et AB est congru à $A'C$ et AB est congru à $A'D$.

A.G2. La relation de congruence est transitive, c'est-à-dire, si AB est congru à CD et si CD est congru à EF , alors AB est congru à EF .

A.G3. Soit une droite d contenant les segments adjacents $[AB]$ et $[BC]$, et soit une droite d' contenant les segments adjacents $[A'B']$ and $[B'C']$. Si $[AB]$ est congru à $[A'B']$ et $[BC]$ est congru à $[B'C']$ alors $[AC]$ est congru à $[A'C']$.

A.G4. Soit un angle ABC et une demi-droite $B'C'$, il existe deux et seulement deux demi-droites, $B'D$ et $B'E$, tel que l'angle $DB'C'$ est congru à l'angle ABC et l'angle $EB'C'$ est congru à l'angle ABC .

Corollaire: Tout angle est congru à lui-même.

A.G5. Soient deux triangles ABC et $A'B'C'$ tels que AB est congru à $A'B'$, AC est congru à $A'C'$, et l'angle BAC est congru à l'angle $B'A'C'$, alors le triangle ABC est congru au triangle $A'B'C'$.

Remarque: Ces axiomes permettent de comparer les segments, et aussi les angles de définir le milieu d'un segment, les droites orthogonales, de parler de triangles équilatéraux, isocèles, etc... Ils permettent également de définir rigoureusement les déplacements dont Euclide faisait si souvent usage sans les avoir définis.

AXIOMES DE CONTINUITÉ (C)

A.C1. ("**Axiome d'Archimède**") Soient $[AB]$ et $[CD]$ deux segments quelconques. Alors il existe toujours une suite finie de points $A_1, \dots, A_n$ appartenant à la droite contenant le segment $[AB]$ et tels que $A < A_1 < \dots < A_{n-1} < B < A_n$ qui peuvent satisfaire $AA_1 \equiv A_1A_2 \equiv \dots \equiv A_{n-1}A_n \equiv CD$.

A.C2. ("**Axiome de Cantor**") Si (A_n) et (B_n) sont deux suites infinies de points telles que $]A_{i+1}, B_{i+1}[\subset]A_i, B_i[$ et telles que $\forall]C, D[, \exists i \Rightarrow]A_i, B_i[\equiv]C, D[$, alors il existe un point X appartenant à tous les segments $]A_i, B_i[$. En d'autres termes: soit une suite de segments emboîtés dont la longueur tend vers 0 alors il y a un point commun à tous les segments.

Remarque: Ces axiomes permettent d'établir une correspondance entre les points d'une droite et l'ensemble des nombres réels.

AXIOMES DES PARALLÈLES (P)

A.P1. Soit d une droite et P un point n'appartenant pas à d . Il passe une et une seule droite d' par P qui soit parallèle à d .

Autre formulation équivalente:

A.P1. Soit une droite d , un point P non inclus dans d , alors il existe un plan contenant d et P . Ce plan contient une et une unique droite contenant P et ne contenant aucun point de d .

Nous ne pouvons pas réellement démontrer la non-contradiction logique de l'ensemble de ces axiomes. Cependant nous savons deux choses si nous faisons un parallèle avec ce que nous avons étudié dans la section d'Arithmétique et d'Algèbre du site (en particulier les chapitres sur la Théorie Des Ensembles, l'Analyse Fonctionnelle, les Suites Et Séries):

1. Si ces axiomes sont contradictoires, alors la théorie des nombres réels est contradictoire.
2. Si le système d'axiomes obtenu en supprimant l'axiome de Cantor est contradictoire, alors la théorie des nombres rationnels est contradictoire.

Ainsi, la confiance qu'on a dans la solidité de ces axiomes repose sur celle qu'on a dans la théorie des nombres réels, qui est très grande.

BARYCENTRE

Maintenant que nous avons abordé le minimum de la construction d'Euclide et d'Hilbert de la géométrie, nous pouvons passer à un niveau supérieur pour faire de l'analyse de propriétés des formes géométriques. Nous commencerons donc par étudier le concept de "barycentre", appelé également mais plus rarement "centroïde".

Remarques:

R1. La définition du barycentre nécessite certains des outils mathématiques définis dans le chapitre de Calcul Vectoriel . La lecture de ce chapitre est donc recommandée si le lecteur souhaite comprendre ce qui va suivre.

R2. Les développements qui vont suivre sont aussi bien utilisés en géométrie qu'en physique!

Définition: Nous appelons "**barycentre**" ou "**centroïde**" des points A_i du plan ou de l'espace affectés respectivement des coefficients $\alpha_1, \alpha_2, \dots, \alpha_n$ (où les α_i sont des réels tels que $\sum \alpha_i \neq 0$) l'unique point G tel que:

$$\sum_{i=1}^n \alpha_i \overrightarrow{GA_i} = \vec{0} \quad (21.77)$$

Le couple noté (A_i, α_i) est appelé "**point pondéré**" ("**point massif**" en physique quand $\alpha_i > 0$ représente une masse).

Remarques:

R1. En mécanique, le "**centre d'inertie**" d'un corps correspond au barycentre des particules qui composent le corps en question. Chaque particule étant pondérée par sa masse propre. C'est donc le point par rapport auquel la masse est uniformément répartie. Si la densité est constante, le centre d'inertie se confond avec le barycentre.

R2. Le "**centre de gravité**" d'un corps correspond au barycentre des particules qui composent le corps en question, chaque particule étant pondérée par son poids propre! Très souvent en mécanique, la dimension des corps étant faible devant la rotondité de la terre, on considère un champ de gravité uniforme. Sous cette hypothèse, le centre de gravité et le centre d'inertie sont confondus.

R3. Lorsque pour tout point massif (A_i, α_i) nous avons $\alpha_i = \alpha_j \quad \forall i, j \in \mathbb{N}^*$, nous parlons alors de "**isobarycentre**".

Pour un point O arbitraire, nous avons bien évidemment par simple addition vectorielle:

$$\sum_{i=1}^n \alpha_i \overrightarrow{GA_i} = \sum_{i=1}^n \alpha_i (\overrightarrow{OA_i} - \overrightarrow{OG}) = \sum_{i=1}^n \alpha_i \overrightarrow{OA_i} - \left(\sum_{i=1}^n \alpha_i \right) \overrightarrow{OG} = \vec{0} \quad (21.78)$$

d'où le résultat majeur:

$$\boxed{\overrightarrow{OG} = \frac{\sum_{i=1}^n \alpha_i \overrightarrow{OA_i}}{\sum_{i=1}^n \alpha_i}} \quad (21.79)$$

En passant à la limite, si le domaine est continu, nous avons:

$$\boxed{\overrightarrow{OG} = \frac{\int \overrightarrow{OM} dl}{\int dl} = \frac{\int \overrightarrow{OM} dl}{L}} \quad (21.80)$$

Nous pouvons très bien également travailler avec les éléments de surface ou de volumes (pour ne faire mention que des plus triviaux) pour déterminer le barycentre:

$$\overrightarrow{OG} = \frac{\iint \overrightarrow{OM} dS}{\int dS} = \frac{\iint \overrightarrow{OM} dS}{S} \quad \text{et} \quad \overrightarrow{OG} = \frac{\iiint \overrightarrow{OM} dV}{\int dV} = \frac{\iiint \overrightarrow{OM} dV}{V} \quad (21.81)$$

Dans l'espace muni d'un repère $(O, \vec{i}, \vec{j}, \vec{k})$ en notant (x_i, y_i, z_i) les coordonnées du point pondéré (A_i, α_i) et (x_G, y_G, z_G) celles de G , nous avons alors:

$$\boxed{x_G = \frac{\sum_{i=1}^n \alpha_i x_i}{\sum_{i=1}^n \alpha_i}, y_G = \frac{\sum_{i=1}^n \alpha_i y_i}{\sum_{i=1}^n \alpha_i}, z_G = \frac{\sum_{i=1}^n \alpha_i z_i}{\sum_{i=1}^n \alpha_i}} \quad (21.82)$$

Voyons quelques propriétés du barycentre:

P1. Soit $(A_1, \alpha_1), (A_2, \alpha_2), \dots, (A_n, \alpha_n)$, n points pondérés. Si $\sum \alpha_i \neq 0$, nous avons alors pour tout point M :

$$\sum_{i=1}^n \alpha_i \overrightarrow{MA_i} = \left(\sum_{i=1}^n \alpha_i \right) \overrightarrow{MG} \quad (21.83)$$

Démonstration:

$$\begin{aligned} \sum_{i=1}^n \alpha_i \overrightarrow{MA_i} &= \alpha_1 \overrightarrow{MA_1} + \alpha_2 \overrightarrow{MA_2} + \dots + \alpha_n \overrightarrow{MA_n} \\ &= \alpha_1 (\overrightarrow{MG} + \overrightarrow{GA_1}) + \alpha_2 (\overrightarrow{MG} + \overrightarrow{GA_2}) + \dots + \alpha_n (\overrightarrow{MG} + \overrightarrow{GA_n}) \\ &= (\alpha_1 + \alpha_2 + \dots + \alpha_n) \overrightarrow{MG} + \alpha_1 \overrightarrow{GA_1} + \alpha_2 \overrightarrow{GA_2} + \dots + \alpha_n \overrightarrow{GA_n} \\ &= \left(\sum_{i=1}^n \alpha_i \right) \overrightarrow{MG} + \sum_{i=1}^n \alpha_i \overrightarrow{GA_i} \end{aligned} \quad (21.84)$$

Puisque par définition du barycentre:

$$\sum_{i=1}^n \alpha_i \overrightarrow{GA_i} = \vec{0} \quad (21.85)$$

nous avons alors bien:

$$\sum_{i=1}^n \alpha_i \overrightarrow{MA_i} = \left(\sum_{i=1}^n \alpha_i \right) \overrightarrow{MG} + \sum_{i=1}^n \alpha_i \overrightarrow{GA_i} = \left(\sum_{i=1}^n \alpha_i \right) \overrightarrow{MG} \quad (21.86)$$

□C.Q.F.D.

P2. Pour $\forall k \in \mathbb{R}^*$, les points pondérés $(A_1, \alpha_1), (A_2, \alpha_2), \dots, (A_n, \alpha_n)$ et $(A_1, k\alpha_1), (A_2, k\alpha_2), \dots, (A_n, k\alpha_n)$ ont même barycentre car (invariance du barycentre):

$$\sum_{i=1}^n \alpha_i \overrightarrow{GA_i} = \vec{0} \Leftrightarrow \sum_{i=1}^n (k\alpha_i) \overrightarrow{GA_i} = \vec{0} \quad (21.87)$$

La démonstration est évidente (si vous ne voyez pas, contactez-nous).

P3. Le barycentre G de n points pondérés est invariant quand on remplace p d'entre eux, par leur barycentre G' , affecté de la condition $\sum_{i=1}^p \alpha_i \neq 0$ de leurs coefficients, G est alors le barycentre de:

$$\left(G'_i, \sum_{i=1}^p \alpha_i \right), (A_{p+1}, \alpha_{p+1}), \dots, (A_n, \alpha_n) \quad (21.88)$$

Démonstration:

Si G' est le barycentre des points pondérés $(A_1, \alpha_1), (A_2, \alpha_2), \dots, (A_p, \alpha_p)$ alors:

$$\sum_{i=1}^p \alpha_i \overrightarrow{MA_i} = \left(\sum_{i=1}^p \alpha_i \right) \overrightarrow{MG'} \quad (21.89)$$

Pour le cas particulier où $M = G$:

$$\sum_{i=1}^p \alpha_i \overrightarrow{GA_i} = \left(\sum_{i=1}^p \alpha_i \right) \overrightarrow{GG'} \quad (21.90)$$

Or G étant le barycentre des n points pondérés $(A_1, \alpha_1), (A_2, \alpha_2), \dots, (A_n, \alpha_n)$ donc:

$$\sum_{i=1}^n \alpha_i \overrightarrow{GA_i} = \vec{0} \Leftrightarrow \sum_{i=1}^p \alpha_i \overrightarrow{GA_i} + \sum_{i=p+1}^n \alpha_i \overrightarrow{GA_i} = \left(\sum_{i=1}^p \alpha_i \right) \overrightarrow{GG'} + \sum_{i=p+1}^n \alpha_i \overrightarrow{GA_i} = \vec{0} \quad (21.91)$$

Comme $\sum_{i=1}^p \alpha_i + \sum_{i=p+1}^n \alpha_i \neq 0$ l'égalité précédente prouve bien que G est le barycentre des points pondérés:

$$\left(G'_i, \sum_{i=1}^p \alpha_i \right), (A_{p+1}, \alpha_{p+1}), \dots, (A_n, \alpha_n) \quad (21.92)$$

□C.Q.F.D.

P4. Si $\sum \alpha_i = 0$, pour tous points M, N :

$$\sum_{i=1}^n \alpha_i \overrightarrow{MA_i} = \sum_{i=1}^n \alpha_i \overrightarrow{NA_i} = c^{te} \quad (21.93)$$

Démonstration:

Pour $M \neq N$ calculons:

$$\begin{aligned} \sum_{i=1}^n \alpha_i \overrightarrow{MA_i} - \sum_{i=1}^n \alpha_i \overrightarrow{NA_i} &= \alpha_1 \overrightarrow{MA_1} + \alpha_2 \overrightarrow{MA_2} + \dots + \alpha_n \overrightarrow{MA_n} - \alpha_1 \overrightarrow{NA_1} - \alpha_2 \overrightarrow{NA_2} - \dots - \alpha_n \overrightarrow{NA_n} \\ &= \alpha_1 (\overrightarrow{MA_1} - \overrightarrow{NA_1}) + \alpha_2 (\overrightarrow{MA_2} - \overrightarrow{NA_2}) + \dots + \alpha_n (\overrightarrow{MA_n} - \overrightarrow{NA_n}) = \alpha_1 \overrightarrow{MN} + \alpha_2 \overrightarrow{MN} + \dots + \alpha_n \overrightarrow{MN} \\ &= (\alpha_1 + \alpha_2 + \dots + \alpha_n) \overrightarrow{MN} = \left(\sum_{i=1}^n \alpha_i \right) \overrightarrow{MN} = \vec{0} \end{aligned} \quad (21.94)$$

puisque $\sum_{i=1}^n \alpha_i = 0$, nous avons alors:

$$\sum_{i=1}^n \alpha_i \overrightarrow{MA_i} = \sum_{i=1}^n \alpha_i \overrightarrow{NA_i} = \vec{0} \quad (21.95)$$

□C.Q.F.D.

Remarques: Quand un corps a une certaine symétrie, les calculs se simplifient car le barycentre doit coïncider avec l'élément de symétrie. Si un corps, comme une sphère, un parallélépipède, etc., a un centre de symétrie, le barycentre est confondu avec lui. Si le corps a seulement un axe de symétrie, le barycentre est alors sur cet axe.

TRANSFORMATIONS

Les transformations dans le plan (et plus) sont habituellement définies rigoureusement à l'aide de la théorie des groupes (cf. chapitre d'Algèbre Ensembliste). Mais dans le cadre de la géométrie euclidienne, cette approche ne nous intéressera pas. Nous ferons donc dans ce chapitre uniquement une approche très peu formelle (donc plus intuitive) des transformations élémentaires dans le plan que sont: la translation, l'homothétie et la rotation.

Remarque: Par définition, "**l'isométrie**" est une transformation qui conserve les distances et les aires. Comme nous le verrons ci-après, la translation, la rotation et la réflexion sont des isométries, l'homothétie n'étant elle pas du tout une isométrie dans le plan.

TRANSLATION

Soit une droite dans un plan P sur laquelle deux points A et B définissent un segment de la droite noté $\overline{AB}$.

Définition: Une "**translation**" T ("déplacement" dans une direction donnée comme disait Euclide) de ce segment de droite associe à chaque point A et B de nouveaux points $A'B'$ tels que $\overline{AB} = \overline{A'B'}$. Nous pouvons donc restreindre la notion de translation à un point uniquement tel que nous puissions écrire mathématiquement:

$$T: P \rightarrow P \Rightarrow T(A) = A' \quad (21.96)$$

Autrement dit, une fonction de transformation de type translation de l'ensemble du plan dans lui-même associe à chaque pré-image au plus une seule et unique image. La translation est donc une fonction bijective. Nous pouvons donc définir une application de transformation réciproque notée T^{-1} telle que (rappel de ce qui a été vu en arithmétique):

$$T^{-1}(A') = A \quad (21.97)$$

Nous disons par définition qu'un point est "**invariant par translation**" si et seulement si:

$$T(A) = A = A' \quad (21.98)$$

Dans un autre type de formalisme, le déplacement du point A au point B selon le vecteur $\overrightarrow{AB}$ est appelé "**translation du vecteur**" $\overrightarrow{AB}$ (cf. chapitre de Calcul Vectoriel). Elle se traduit mathématiquement par la somme des coordonnées du point et de la matrice des coordonnées du vecteur.

Par exemple dans un espace à trois dimensions:

$$(x, y, z) + (x', y', z') = (x + x', y + y', z + z') \quad (21.99)$$

La translation n'étant pas une transformation linéaire (cf. chapitre d'Algèbre Linéaire), nous ne pouvons nous autoriser à la représenter par la multiplication d'une matrice carrée comme nous le verrons pour les autres transformations suivantes.

Il faut pour cela passer alors par un artifice consistant à utiliser un système appelé les coordonnées homogènes (cf. chapitre de Géométrie Projective) où les points du plan sont représentés par un vecteur à trois composantes (et respectivement ceux de l'espace par un vecteur à quatre dimensions):

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \Rightarrow \begin{pmatrix} x_h \\ y_h \\ w \end{pmatrix} \text{ avec } \begin{matrix} x_h = x \cdot w \\ y_h = y \cdot w \end{matrix} \quad (21.100)$$

Dans le cadre de l'étude de la translation nous posons $w = 1$ car dans ce cas:

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{bmatrix} 1 & 0 & T_x \\ 0 & 1 & T_y \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \cdot x + 0 \cdot y + T_x \cdot 1 \\ 0 \cdot x + 1 \cdot y + T_y \cdot 1 \\ 0 + 0 + 1 \cdot 1 \end{pmatrix} = \begin{pmatrix} x + T_x \\ y + T_y \\ 1 \end{pmatrix} \quad (21.101)$$

Ce système de coordonnées homogènes est applicable à toutes les autres transformations que nous verrons par la suite en rajoutant à chaque fois une coordonnée (cf. chapitre de Géométrie Projective).

Remarque: Une translation envoie une droite sur une droite parallèle (parallèle à l'originale bien évidemment!).

HOMOTHÉTIE

Soit une forme quelconque dans le plan (point, droite, ovale, polygone,...), un nombre R , et un point C placé à un endroit prédéfini.

Définition: Une "**homothétie**" (appelée aussi "**changement d'échelle**") H de rapport R et de centre C est l'application qui à chaque point M de la forme associe au segment $\overline{CM}$ un nouveau point colinéaire à $\overline{CM}$ mais disposé à une distance supérieure ou inférieure de rapport R par rapport au centre C tel que $\overline{CM'} = R \cdot \overline{CM}$

Nous pouvons restreindre la notion d'homothétie à un segment de droite tel que nous puissions écrire mathématiquement:

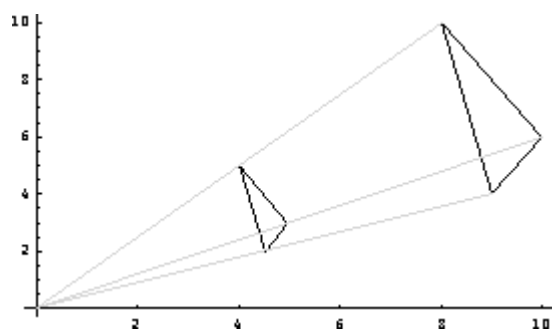
$$H : P \rightarrow P \Rightarrow H(\overline{CM}) = R \cdot \overline{CM} = \overline{CM'} \quad (21.102)$$

Autrement dit, une fonction de transformation de type homothétie de l'ensemble du plan dans lui-même associe à chaque pré-image au plus une seule et unique image. L'homothétie est donc une fonction bijective. On peut donc définir une application de transformation réciproque notée H^{-1} telle que:

$$H^{-1}(R \cdot \overline{CM}) = \overline{CM} \quad (21.103)$$

Si $R \neq 1$, alors C est le seul point invariant. Si $R = 1$, alors tous les points sont invariants et l'homothétie est dite de type "**homothétie identité**". Si $R = -1$, alors nous disons alors que nous avons une "**symétrie centrale**". La symétrie centrale est donc une rotation de 180° .

Dans un autre type de formalisme, une homothétie de centre O et de rapport k , associe au point A un point B tel que $OB = k \cdot OA$. Le point B se trouvant sur la droite OA et à une distance $OB = k \cdot OA$. Le signe de k détermine la position de B par rapport à O :



$$OB = k \cdot OA \Leftrightarrow (x', y', z') = k \cdot (x, y, z)$$

Figure: 21.35 - Exemple d'homothétie de centre O

Nous nous permettons maintenant de faire un petit passage dans la géométrie spatiale (le saut n'étant pas bien grand et nécessitant juste la connaissance du calcul vectoriel et de l'algèbre linéaire):

Nous pouvons également remplacer le scalaire k par une matrice carrée telle que:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} ax + by + cz \\ dx + ey + fz \\ gx + hy + iz \end{pmatrix} = \begin{pmatrix} kx \\ ky \\ kz \end{pmatrix} \quad (21.104)$$

Une solution triviale pour obtenir une homothétie est de poser que $b = c = d = f = g = h = 0$ d'où la forme matricielle diagonale évidente de k :

$$k = \begin{bmatrix} k & 0 & 0 \\ 0 & k & 0 \\ 0 & 0 & k \end{bmatrix} \quad (21.105)$$

Cette matrice est appelée "**matrice de transformation par homothétie de centre O (origine du repère) et de rapport k** " et donc l'homothétie étant une matrice diagonale commute avec toute application linéaire.

Dans ce cas présenté précédemment, l'homothétie conserve les formes dans tous les axes (sa géométrie est invariante par transformation) si nous utilisons en effet le même facteur k pour tous les axes. Mais nous pourrions également utiliser la matrice suivante:

$$k = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix} \quad (21.106)$$

qui elle déformerait l'objet selon un facteur différent pour chaque axe.

Ou encore faire un cisaillement dans le plan qui déforme la géométrie selon l'axe x , par exemple avec $k = 2$:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x + 2y \\ y \end{pmatrix}$$

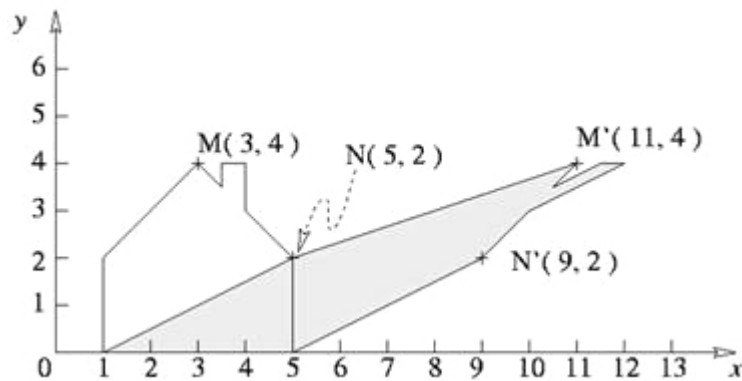


Figure: 21.36 - Homothétie avec cisaillement

La transformation inverse de l'homothétie est bien évidemment l'homothétie de centre O et de rapport k^{-1} soit sous la forme d'une matrice:

$$k^{-1} = \begin{bmatrix} 1/k & 0 & 0 \\ 0 & 1/k & 0 \\ 0 & 0 & 1/k \end{bmatrix} \quad (21.107)$$

Lorsque le centre d'homothétie ne coïncide pas avec l'origine du repère choisi (ce qui arrive quasiment tout le temps), la procédure de calcul des coordonnées du point image est très simple. Il faut:

1. Réaliser une translation pour faire correspondre le centre de l'homothétie avec l'origine du repère et appliquer cette translation à tous les points en jeu.
2. Réaliser l'homothétie proprement dite comme décrit précédemment (le centre est l'origine du repère).
3. Réaliser la translation inverse pour ramener le centre et l'image à sa place.

Pour clore, signalons que les opérations successives de translation et d'homothétie ne sont pas commutatives dans le cas général. Dans la pratique, on effectue souvent d'abord l'homothétie et ensuite la translation.

ROTATION

Soit une forme quelconque dans le plan (point, droite, ovale, polygone), un nombre α , et un point C placé à un endroit prédéfini.

Définition: Une "rotation" R d'angle α et de centre C est l'application qui à chaque point M de la forme associe au segment $\overline{CM}$ un nouveau point mais ayant subi une rotation positive ou négative d'angle α et de centre C telle que $\overline{CM}$ et $\overline{CM'}$ ont même longueur mais pas même direction.

De cette définition, il ressort que l'axe de rotation d'un objet est le lieu de points de cet objet qui restent immobiles.

Remarque: La rotation est également, de manière plus savante, une application bijective dans le plan, nous pouvons donc également définir une application de transformation réciproque notée R^{-1} .

Si $\alpha \neq 0$, alors C est le seul point invariant. Si $\alpha = 2k\pi$ (avec $k \in \mathbb{Z}$), alors tous les points sont invariants et la rotation est dite de type "rotation identité". Si nous choisissons un système d'axes perpendiculaires adéquat tel que leur intersection se confonde avec C et que $\alpha = \pm\pi$ alors R est dite une "rotation de symétrie centrale".

Dans un autre type de formalisme, la rotation s'exprime de manière beaucoup plus rigoureuse. Nous allons nous aider du dessin d'un cercle de rayon unité (donc dans le plan) pour étudier ce type de transformation. Nous allons considérer le premier cas où l'origine du repère et de la translation sont confondus:

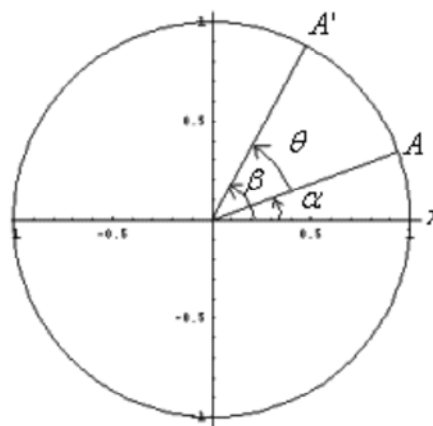


Figure: 21.37 - Exemple de rotation dans le plan

où A' est l'image A par la rotation de centre O et d'angle θ .

Nous avons dans le plan pour le point A (cf. chapitre de Trigonométrie):

$$x = r \cos \alpha, y = r \sin \alpha \quad (21.108)$$

et identiquement pour le point A' :

$$x' = r \cos \beta, y' = r \sin \beta \quad (21.109)$$

avec $\beta = \alpha + \theta$.

Ce qui nous amène à écrire:

$$\begin{aligned}
 \frac{x}{\cos \alpha} &= \frac{x'}{\cos(\alpha + \theta)} = r \Rightarrow x' = x \cdot \frac{\cos(\alpha + \theta)}{\cos \alpha} \\
 x' &= x \cdot \frac{\cos(\alpha + \theta)}{\cos \alpha} = x \frac{\cos(\alpha) \cos(\theta) - \sin(\alpha) \sin(\theta)}{\cos \alpha} \\
 x' &= x \cos(\theta) - x \frac{\sin(\alpha) \sin(\theta)}{\cos \alpha} = x \cos(\theta) - x \frac{y \sin(\theta)}{r \cos \alpha} \\
 &= x \cos(\theta) - r \cos \alpha \frac{y \sin(\theta)}{r \cos \alpha} = x \cos \theta - y \sin \theta
 \end{aligned} \tag{21.110}$$

Identiquement (en se basant sur le fait que les relations trigonométriques élémentaires présentées dans le chapitre de Trigonométrie du site sont connues), nous trouvons:

$$y' = x \sin \theta + y \cos \theta \tag{21.111}$$

ce qui nous permet d'écrire la matrice de rotation dans le plan (en s'imaginant que l'axe Z sort de la feuille):

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \cos \theta - y \sin \theta \\ x \sin \theta + y \cos \theta \end{pmatrix} \tag{21.112}$$

La transformation inverse consiste très simplement en la rotation de centre O (le même qu'auparavant) et d'angle $-\theta$ soit (nous utilisons à nouveaux les relations trigonométriques évidentes des angles opposés):

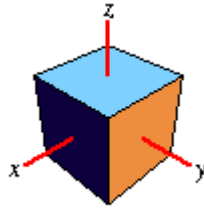
$$\begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \tag{21.113}$$

Lorsque nous souhaitons procéder à une rotation autour d'un point quelconque, tout comme pour l'homothétie, il convient de réaliser une translation de vecteur $\overrightarrow{OH}$ (H étant l'origine du repère de l'homothétie) pour faire confondre O et H , puis de réaliser la rotation simple autour de H , et enfin de ramener O (confondu alors avec H) à son point de départ.

Lors de la rotation d'un objet dans l'espace (nous profitons de la lancée... car nous en aurons besoin dans plusieurs chapitres relatifs à la physique), la transformation est assez similaire à la précédente.

Effectivement, lors d'une rotation d'angle ϕ , autour de l'axe Z la coordonnée z ne change pas. Ce qui nous amène à écrire la matrice de rotation dans l'espace tridimensionnel par rapport au plan x, y comme étant:

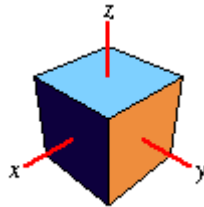
$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{bmatrix} \cos \phi & \sin \phi & 0 \\ -\sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = R_Z(\phi) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \cos \phi + y \sin \phi \\ -x \sin \phi + y \cos \phi \\ z \end{pmatrix} \tag{21.114}$$



La philosophie est ensuite toujours la même relativement aux autres axes:

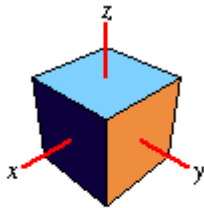
Rotation autour l'axe X d'angle θ :

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & \sin \theta \\ 0 & -\sin \theta & \cos \theta \end{bmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = R_X(\theta) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \\ y \cos \theta + z \sin \theta \\ -y \sin \theta + z \cos \theta \end{pmatrix} \quad (21.115)$$



Rotation autour de l'axe Y d'angle γ :

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{bmatrix} \cos \gamma & 0 & -\sin \gamma \\ 0 & 1 & 0 \\ \sin \gamma & 0 & \cos \gamma \end{bmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = R_Y(\gamma) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \cos \gamma - z \sin \gamma \\ y \\ x \sin \gamma + z \cos \gamma \end{pmatrix} \quad (21.116)$$



Nous avons donc finalement trois matrices de rotation correspondant chacune à un des plans (xy, yz, xz) de l'espace tridimensionnel.

Ces trois matrices font partie du groupe des matrices d'ordre trois, noté "**SO(3)**" et appelé par les physiciens et mathématiciens "**groupe de rotations spatiales SO(3)**". Une rotation quelconque peut donc être représentée par la matrice produit résultant du produit de ces trois matrices.

Toute rotation consiste ensuite en une composition de ces trois rotations mais il est important que le lecteur se souvienne du chapitre d'Algèbre Linéaire où nous avons vu que la multiplication matricielle n'est pas commutative. Ainsi, tourner autour de l'axe X de 90° et ensuite autour de l'axe Z de 90° n'est pas équivalent à faire tourner d'abord selon l'axe Z et ensuite selon l'axe X du même angle comme le montre l'image ci-dessous:

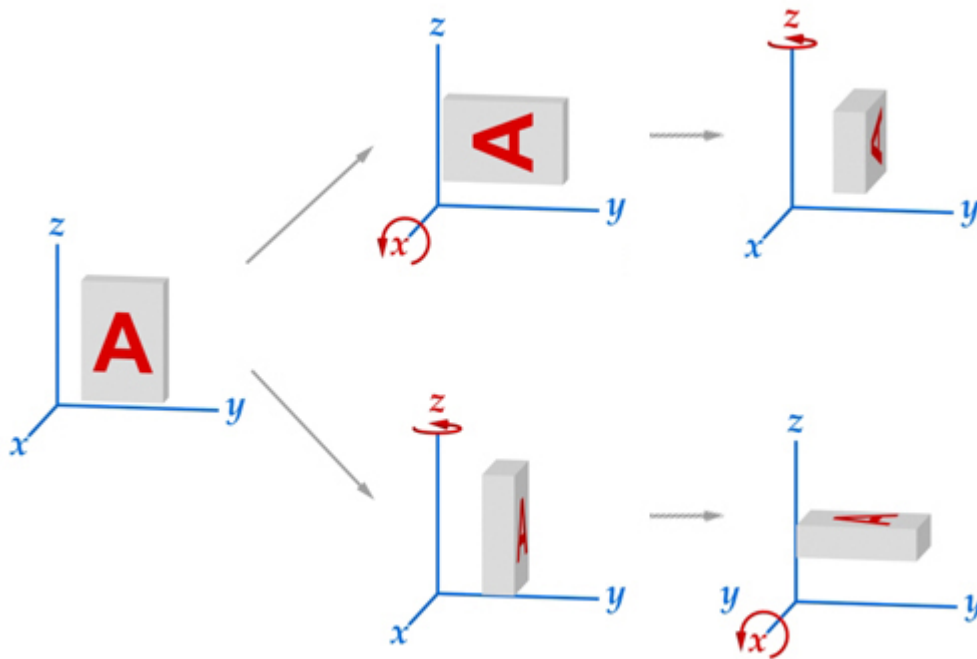


Figure: 21.38 - Exemple de non commutativité de la matrice de rotation

Enfin, pour obtenir la matrice qui compose un cas particulier des trois rotations voici par exemple les commandes à fournir dans Maple 4.00b:

```
>X:=array([[1,0,0],[0,cos(theta),sin(theta)],[0,-sin(theta),cos(theta)]]);
>Y:=array([[cos(lambda),0,-sin(lambda)],[0,1,0],[sin(lambda),0,cos(lambda)]]);
>Z:=array([[cos(phi),sin(phi),0],[-sin(phi),cos(phi),0],[0,0,1]]);
>evalm(X&*Y&*Z);
```

Si nous cherchons à réaliser la composition d'une rotation R et d'une homothétie d'échelle H (dans cet ordre) la matrice de transformation sera:

$$[H] \cdot [R] \quad (21.117)$$

Remarques:

R1. Nous rappelons (cf. chapitre d'Algèbre Linéaire) que la multiplication de 2 matrices n'est pas commutative.

R2. La similitude directe de centre C , de rapport R et d'angle α est la composée de l'homothétie de centre C et de rapport R et de la rotation de centre C et d'angle α . Nous renvoyons le lecteur au chapitre sur les Nombres pour revoir que les nombres complexes permettent formellement d'opérer avec les opérations d'addition et de multiplication à des similitudes (directes ou rétrogrades).

R3. Nous pouvons faire des rotations beaucoup plus puissantes et variables à l'aide des nombres quaternions (ou "hypercomplexes"). Pour plus d'informations le lecteur se reportera au chapitre sur les Nombres.

RÉFLEXION

Définition: La "**réflexion**", appelée également "**symétrie axiale**", notée S_{Δ} (en géométrie), par rapport à la droite Δ est l'application qui associe à chaque point M extérieur à Δ le point M' telle que Δ soit la médiatrice de MM' . Si M appartient à Δ , alors $M = M'$.

Mathématiquement cela s'écrit:

$$S_{\Delta} : P \rightarrow P \Rightarrow S_{\Delta}(M) = M' \quad (21.118)$$

Autrement dit, une fonction de transformation de type réflexion de l'ensemble du plan dans lui-même associe à chaque pré-image au plus une seule et unique image. La réflexion est donc une fonction bijective. Nous pouvons donc définir une application de transformation réciproque notée S_{Δ}^{-1} telle que:

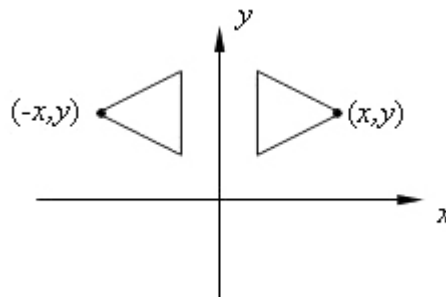
$$S_{\Delta}^{-1}(M') = M \quad (21.119)$$

Remarque: Tous les points de Δ sont trivialement invariants par la réflexion dans le plan.

Sous forme matricielle les réflexions du plan sont extrêmement simples à formaliser en utilisant l'algèbre linéaire (voir chapitre du même nom) comme le montrent les exemples ci-dessous:

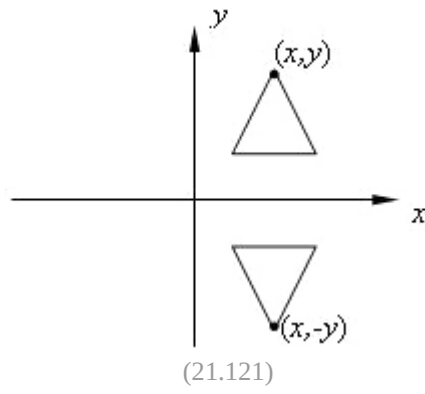
- Réflexion par rapport à l'axe des Y:

$$S_{\Delta} = \begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -x \\ y \end{pmatrix} \quad (21.120)$$



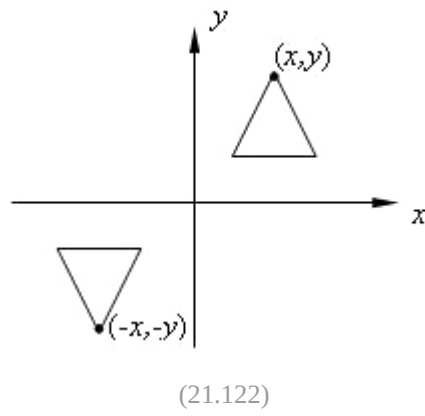
- Réflexion par rapport à l'axe des X:

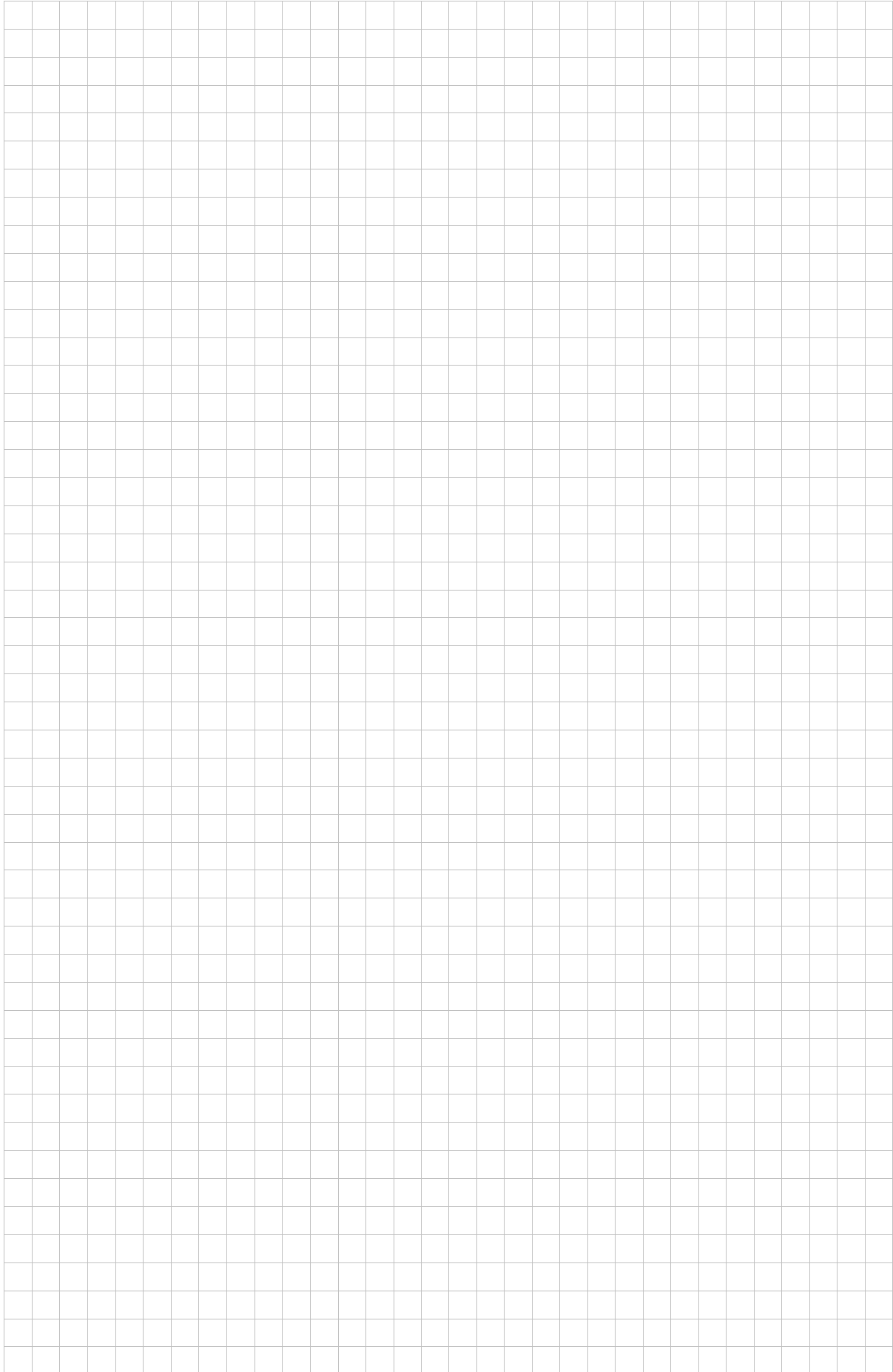
$$S_{\Delta} = \begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ -y \end{pmatrix}$$

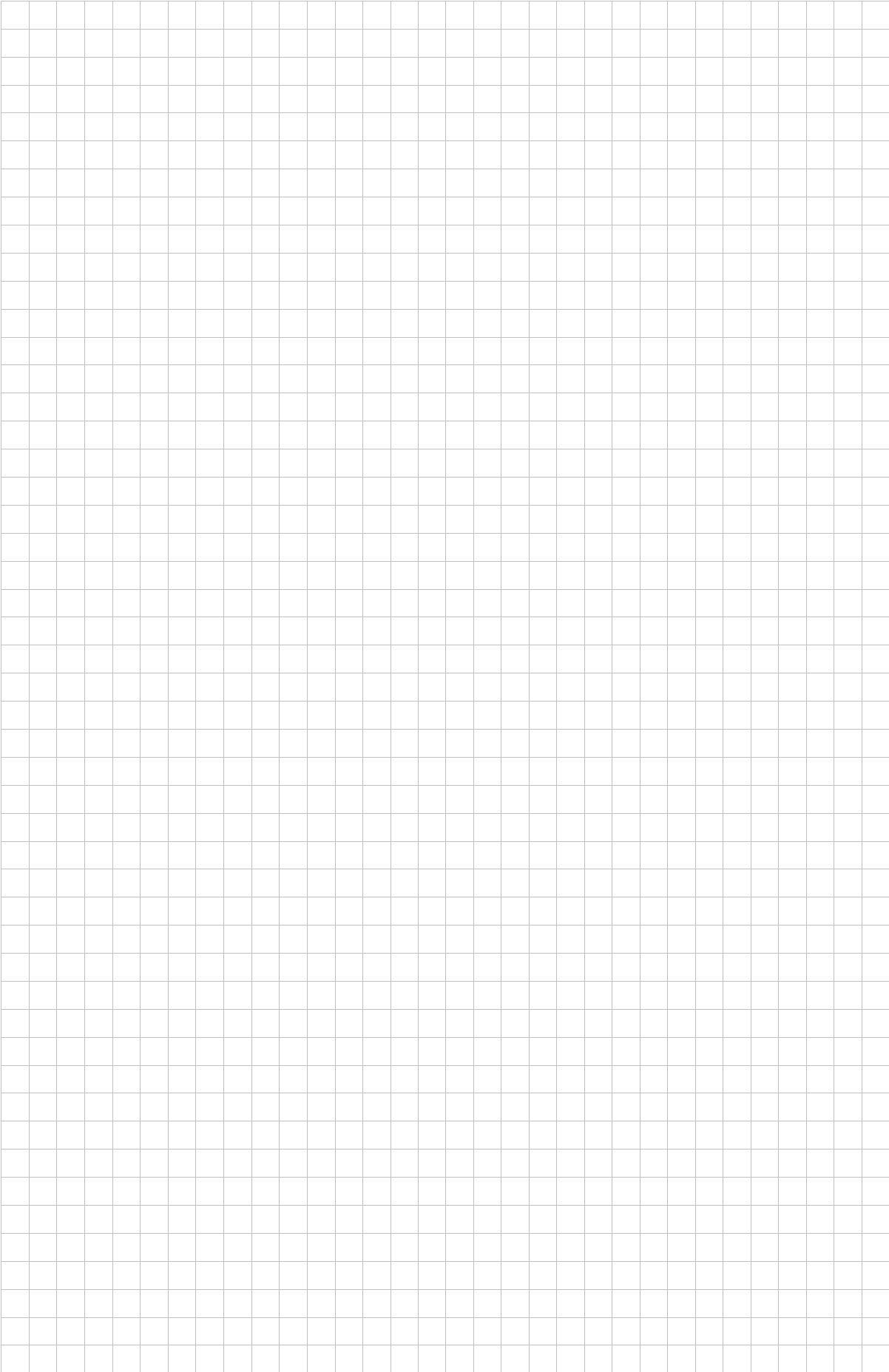


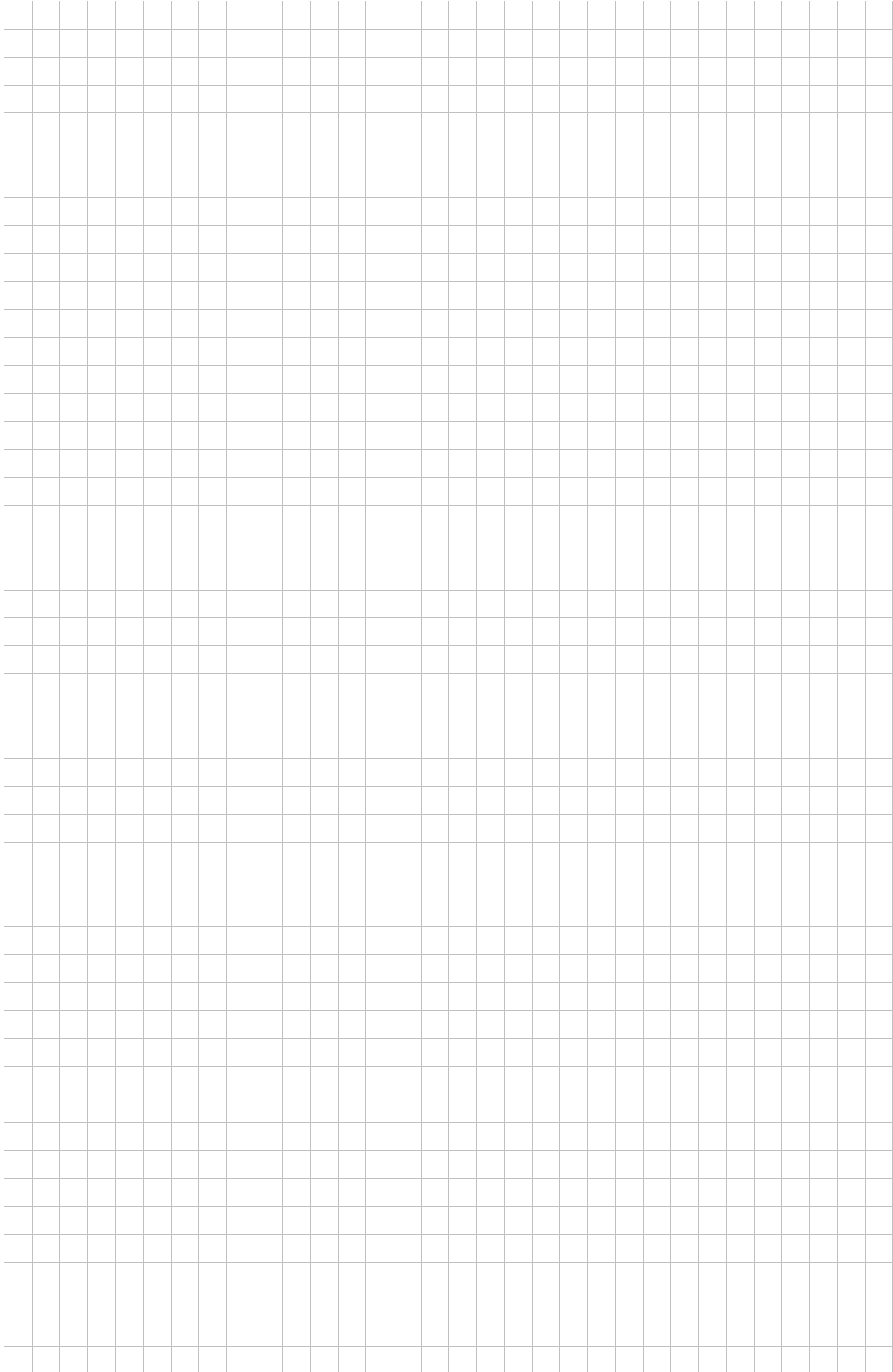
- Réflexion par rapport à l'origine:

$$\mathcal{S}_{\Delta} = \begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -x \\ -y \end{pmatrix}$$









22. GÉOMÉTRIES NON-EUCLIDIENNES

Les géométries non-euclidiennes sont toutes les géométries qui satisfont non nécessairement tous les axiomes de Hilbert (cf. chapitre de Géométrie Euclidienne) mais sans en contredire aucun (contrairement aux anciens axiomes d'Euclide et en particulier celui sur les parallèles).

Une représentation particulière de ce type de géométrie consiste à définir les points comme étant répartis sur la surface d'une sphère (ce sont les intersection des diamètres de la sphère avec la surface), et les lignes, pour généraliser le concept de droite, (nous disons maintenant "géodésique"), comme les intersections de la surface de la sphère avec les plans contenant le centre de la sphère. Deux points définissent alors de façon unique une ligne et un point est toujours donné par l'intersection de deux lignes. Cependant, dans cette géométrie, si nous nous donnons une ligne AB et un point P , il n'existe aucune ligne passant par P et ne coupant pas AB . Ainsi, le cinquième postulat d'Euclide n'est pas satisfait car en P nous ne pouvons tracer aucune parallèle à AB .

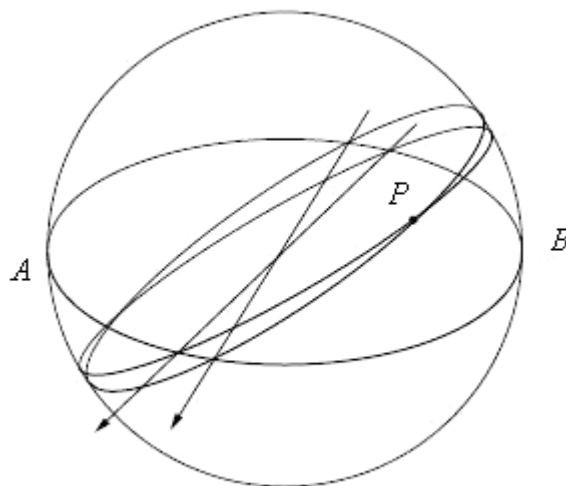


Figure: 22.1 - Exemple illustré de la violation du 5ème postulat d'Euclide

Remarque: Avant d'aborder ce chapitre, nous recommandons vivement au lecteur d'avoir lu et si possible compris les chapitres traitant du Calcul Tensoriel, de Trigonométrie et de Géométrie Euclidienne car nous allons utiliser grand nombre de résultats, non nécessairement triviaux, que nous avons pu y démontrer.

Dans le chapitre de Géométrie Euclidienne, nous avons étudié un certain nombre de théorèmes relatifs aux plans. Insistons maintenant sur le fait que le "plan" et une figure bidimensionnelle dont la courbure est nulle et plongée dans un espace à 3 dimensions (donc le plan peut dès lors s'orienter). Ceci précisé, il convient peut-être de définir plus rigoureusement ce qu'est le concept intuitif de "courbure".

Définition: Une figure est dite "**courbe**" s'il existe au moins en un point se situant sur la ou les droites, surfaces, volumes, ... la délimitant une tangente non confondue au délimiteur et donc tangente en un seul point.

C'est Gauss qui en 1824 avait formulé la possibilité qu'il existe des géométries alternatives à celles d'Euclide. Nous distinguons les géométries à "**courbure négative**", comme celle du russe Nicolaï Lobatchevski (1829) et Bolyai (1832) (somme des angles d'un triangle inférieure à 180° , nombre

infini de parallèles possibles à une droite par un point), des géométries à "courbure positive" comme celle de Riemann (1867) (somme des angles d'un triangle supérieure à 180° , parallèles se rejoignant aux pôles).

Nous allons voir dans ce chapitre différentes géométries non-euclidiennes dont les plus connues sont les "géométries Riemanniennes" (à courbure constante) et les "géométries de Lobatchevski" (de type hyperbolique donc à courbure non-constante).

Remarque: La géométrie communément appelée "géométrie de Riemann" est un espace sphérique à trois dimensions, espace fini et cependant sans bornes, à courbure régulière, alternative au postulat Euclidien des parallèles.

L'intérêt de l'étude de ces géométries est que nous ne pouvons déterminer si l'Univers dans lequel nous vivons est fait d'un type de géométrie plutôt que d'un autre car étant donné notre taille (physique), plongés que nous sommes dans quelque géométrie que ce soit à faible courbure, toute surface de l'espace nous semble localement euclidienne (deux droites parallèles ne se coupent pas). Cependant, la relativité générale, qui fait usage à outrance du calcul tensoriel (généralisation de n'importe quelle géométrie) montre qu'il existe des zones de l'espace où la géométrie est très fortement courbée et donc localement non-euclidienne et seulement l'étude de ce genre de géométries nous permet de tirer des théories expliquant des observations qui ne sont pas exploitables uniquement avec l'intuition humaine.

Avant de nous attaquer de manière formelle et abstraite à certaines géométries non-euclidiennes nous allons d'abord faire une introduction pragmatique et particulière de certains concepts qui ne nous sont pas totalement étrangers car déjà traités dans d'autres chapitres de manière théorique. Une fois cette introduction faite, qui nous sera très utile pédagogiquement parlant, nous aborderons les concepts vus plus rigoureusement.

GÉODÉSIQUE ET EQUATION MÉTRIQUE

Revenons donc sur les concepts de géodésique et courbure dont nous avons souvent fait mention dans le chapitre de Calcul Tensoriel (le fait de ne pas avoir lu ce chapitre ne pose aucun problème normalement à la compréhension de ce qui va suivre).

Considérons la surface bidimensionnelle d'une sphère de rayon R . Étant donnés deux points B et C diamétralement opposés, nous cherchons la plus courte distance s mesurée sur la sphère entre B et C . La courbe obtenue est comme nous le savons une "géodésique", notion qui généralise donc, pour une surface arbitraire, la notion de droite du plan.

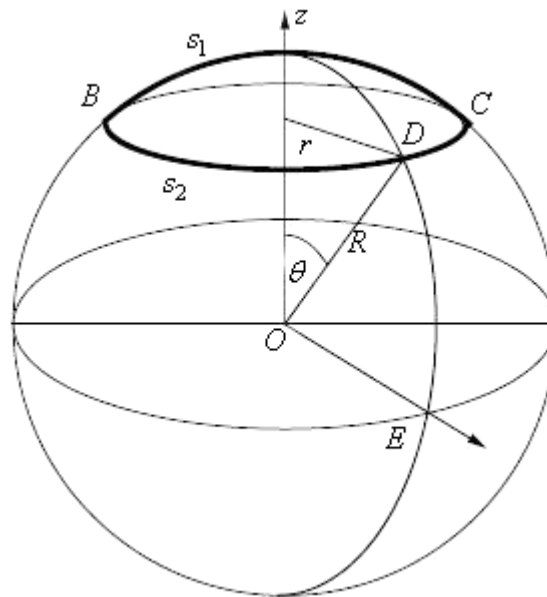


Figure: 22.2 - Illustration du problème pour la recherche de la géodésique

Remarque: Nous supposons comme intuitif que la longueur d'une courbe de l'espace tridimensionnel euclidien est toujours supérieure ou égale à la longueur de toute projection plane de cette courbe. La courbe géodésique est donc nécessairement une courbe plane.

Le rayon entre l'axe Oz et l'un des points B ou C est trivialement donné par un peu de trigonométrie élémentaire:

$$r = R \sin \theta \quad (22.1)$$

Et donc la moitié du périmètre du cercle à hauteur de B et C sera donné par:

$$s_2 = \frac{P_2}{2} = \frac{2\pi R \sin \theta}{2} = \pi R \sin \theta \quad (22.2)$$

Et nous avons démontré dans le chapitre de Trigonométrie que le périmètre d'un cercle en fonction de l'angle d'ouverture de ce dernier étant donnée par:

$$L = R \cdot \alpha \quad (22.3)$$

Il vient donc automatiquement:

$$s_1 = R \cdot (2\theta) \quad (22.4)$$

Comme $\pi \sin \theta \geq 2\theta$ sur l'intervalle $[0, \pi/2]$ alors $s_2 \geq s_1$ (il y a égalité en $\theta = 0$ et $\theta = \pi/2$). Les géodésiques de la sphère sont donc les arcs de grands cercles, trajets empruntés par les avions pour les vols intercontinentaux, et correspondent aux lignes obtenues entre la surface de la sphère et un plan passant par le centre de celle-ci.

Les propriétés géométriques des figures tracées sur la surface d'une sphère ne sont donc plus celles de la géométrie euclidienne. Ainsi, le plus court chemin d'un point B à un point C , sur la surface sphérique, est constitué par un arc de grand cercle passant par les points B et C . Les arcs de grand cercle jouent le même rôle pour la sphère que les droites dans le plan. Ce sont les "géodésiques" de

la sphère.

Considérons maintenant deux surfaces bidimensionnelles: la surface de la sphère et celle du cylindre. Etant donnés deux points B et C , nous traçons la courbe géodésique entre ces points:

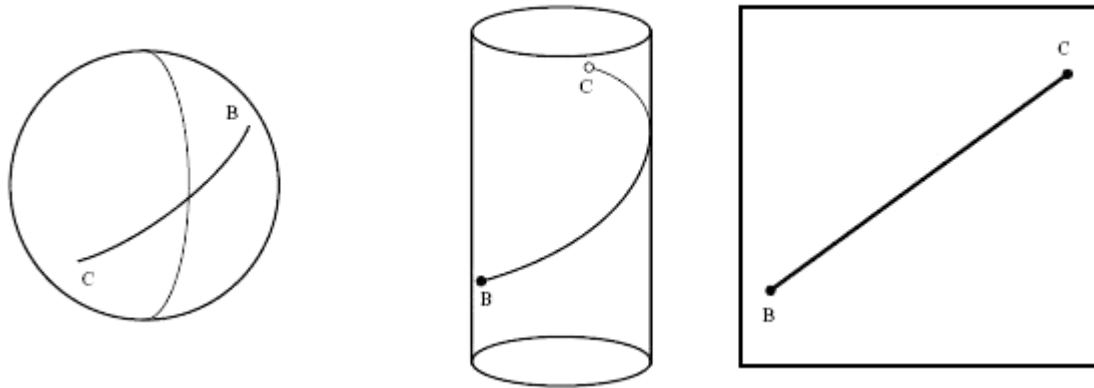


Figure: 22.3 - Représentation plan de la surface latérale d'un cylindre

Le cylindre peut être découpé parallèlement à son axe et déplié à plat. La géodésique apparaît ainsi comme une droite du plan. Nous disons alors que le cylindre est "intrinsèquement plat" (même si sa topologie diffère de celle du plan, il faut en particulier ici éviter que la coupure ne traverse la géodésique). Ce n'est évidemment intuitivement pas le cas de la surface de la sphère.

Dans le cas de la surface cylindrique, nous pouvons définir les coordonnées cartésiennes du plan $B(y_1, z_1)$ et $C(y_2, z_2)$ permettant d'écrire la longueur s de la courbe (droite) BC sous la forme du théorème de Pythagore:

$$s^2 = (y_2 - y_1)^2 + (z_2 - z_1)^2 \quad (22.5)$$

La métrique du plan est euclidienne et sous forme infinitésimale nous obtenons "l'équation métrique euclidienne":

$$ds^2 = dy^2 + dz^2 \quad (22.6)$$

Sur le cylindre, le changement de variable $y = r\theta$ donne:

$$s^2 = (r\theta_2 - r\theta_1)^2 + (z_2 - z_1)^2 = r^2(\theta_2 - \theta_1)^2 + (z_2 - z_1)^2 \quad (22.7)$$

Ou sous forme locale:

$$ds^2 = r^2 d\theta^2 + dz^2 \quad (22.8)$$

La surface du cylindre peut ainsi être représentée par des coordonnées cartésiennes analogues à celles du plan, la métrique de la surface du cylindre étant euclidienne sous forme infinitésimale et sous forme globale.

Remarque: La relation précédente correspond à ce que nous avons obtenu dans le chapitre de Calcul Tensoriel pour l'équation métrique en coordonnées polaires.

Nous pouvons nous intéresser maintenant au problème d'écrire l'analogue du théorème de Pythagore pour une surface sphérique. L'impossibilité de découper la sphère et de l'aplatir pour épouser un plan suggère des difficultés...

C'est la raison pour laquelle l'équation de la métrique ne peut s'écrire sous forme générale comme le théorème de Pythagore. Effectivement, nous avons vu dans le chapitre de Calcul Tensoriel que celle-ci était donnée par:

$$ds^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (22.9)$$

Cependant, localement (c'est-à-dire dans une région de petite dimension devant le rayon de la sphère), les propriétés de la sphère peuvent être décrites par des coordonnées cartésiennes d'un plan tangent à sa surface (c'est la propriété essentielle des espaces de Riemann!) tel que l'équation métrique soit localement euclidienne:

$$ds^2 = r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (22.10)$$

En posant:

$$d\xi = \sqrt{g_{\theta\theta}} d\theta, d\eta = \sqrt{g_{\phi\phi}} d\phi \quad (22.11)$$

il vient alors:

$$ds^2 = g_{\theta\theta} d\theta^2 + g_{\phi\phi} d\phi^2 = d\xi^2 + d\eta^2 \quad (22.12)$$

Avec:

$$g_{\theta\theta} = r^2 \quad g_{\phi\phi} = r^2 \sin^2 \theta \quad (22.13)$$

Alors que θ, ϕ sont les "**coordonnées de Gauss**", ξ, η sont les "**coordonnées de Riemann**" du plan localement tangent.

Cette petite présentation ayant été faite, passons à un cadre plus général en nous intéressant aux espaces de Riemann.

ESPACES DE RIEMANN

Pour mieux comprendre ce qu'est un espace de Riemann, nous allons de suite passer par un petit exemple d'une surface à deux dimensions (exemple très classique):

Considérons une sphère de rayon R , de surface S , située dans l'espace ordinaire à trois dimensions. Les coordonnées cartésiennes x, y, z d'un point M de la surface S peuvent s'exprimer, par exemple, en fonction des coordonnées sphériques (r, θ, ϕ) . La sphère est entièrement décrite pour un rayon donné et $0 \leq \phi < 2\pi$ et $0 \leq \theta < \pi$.

Trois tels paramètres, permettant de déterminer un point sur la surface d'une sphère, sont nous le savons (cf. chapitre de Calcul Tensoriel) des coordonnées curvilignes sur la surface ou également dites "**coordonnées de Gauss**" (Gauss étant un des premiers mathématiciens à s'intéresser à l'étude des corps plongés dans les espaces non-euclidiens). D'autres paramètres quelconques u, v, w peuvent évidemment être choisis comme coordonnées curvilignes sur la surface.

L'élément linéaire de la surface ds^2 , carré de la distance entre deux points infiniment voisins M, M' , s'écrit en fonction des coordonnées sphériques, comme nous l'avons démontré dans le chapitre de Calcul Tensoriel:

$$ds^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (22.14)$$

Nous obtenons ainsi une expression de l'élément linéaire en fonction des trois seules coordonnées de Gauss (r, θ, ϕ) . Nous pourrions bien sûr imposer une étude locale (plan tangent) afin que l'élément linéaire ne soit plus fonction que de (θ, ϕ) comme nous l'avons vu plus haut:

$$ds^2 = r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (22.15)$$

Écrire à l'aide des trois paramètres, la surface de la sphère (considérée comme un espace à deux dimensions) constitue un exemple d'espace de Riemann à deux dimensions.

Dont l'élément linéaire est de la forme générale bien connue (cf. le chapitre de Calcul Tensoriel):

$$ds^2 = g_{ik} du^i du^k \quad (22.16)$$

où les du^i sont les composantes contravariantes du vecteur $d\vec{M} = \overline{MM'}$ par rapport au repère naturel $(M, \vec{e}_i)$.

Remarque: L'étude des figures sur des surfaces Riemanniennes fait partie de la géométrie différentielle à laquelle nous consacrons un chapitre entier dans cette section.

Considérons à présent une surface quelconque de coordonnées u^1, u^2 . Les coordonnées cartésiennes x, y, z de l'espace ordinaire où se trouve plongée cette surface s'écrivent de manière générale avec les coordonnées de Gauss:

$$x = x(u^1, u^2) \quad y = y(u^1, u^2) \quad z = z(u^1, u^2) \quad (22.17)$$

Remarquons par ailleurs que l'équation métrique sous forme tensorielle:

$$ds^2 = g_{ik} du^i du^k \quad (22.18)$$

peut s'écrire sous forme développée de la manière suivante (cette relation est démontrée avec une approche géométrique dans le chapitre de Géométrie Différentielle):

$$ds^2 = Edu^2 + 2Fdu^1 du^2 + Gdv^2 \quad (22.19)$$

avec:

$$\begin{aligned} E &= (\partial_1 x)^2 + (\partial_1 y)^2 + (\partial_1 z)^2 \\ G &= (\partial_2 x)^2 + (\partial_2 y)^2 + (\partial_2 z)^2 \\ F &= \partial_1 x \partial_2 x + \partial_1 y \partial_2 y + \partial_1 z \partial_2 z \end{aligned} \quad (22.20)$$

Remarques:

R1. L'expression donnée ci-dessus de l'élément linéaire s'appelle "**forme quadratique fondamentale**" de la surface considérée. Les coefficients E, F, G sont des fonctions des coordonnées curvilignes. De manière générale cette surface, considérée comme un espace à deux dimensions, constituera un exemple d'espace de Riemann, pour des coordonnées curvilignes arbitraires.

R2. Les différents espaces de Riemann constituent ce que nous appelons sous une forme générale (parce qu'il n'y a pas que des espaces de type Riemannien à courbure constante) une "**variété**" munie d'une métrique Riemannienne. Une variété peut être définie (non formellement), par exemple, par un ensemble de points situés dans un espace préexistant. De manière générale une surface donne l'idée d'une variété à deux dimensions. La sphère et le tore sont des variétés à deux dimensions sans frontière. Un cylindre de révolution, un paraboloïde hyperbolique, sont des variétés à deux dimensions ouvertes, avec frontières à l'infini. Mais nous pouvons aussi envisager des variétés abstraites. C'est le cas par exemple d'un espace de configuration. Il s'agit alors d'un espace de points à n dimensions représenté par un ensemble $\mathcal{Q}^i$ (ou noté $\mathcal{U}^i$) de coordonnées généralisées (voir l'introduction au formalisme lagrangien dans la section de mécanique analytique), ces dernières pouvant avoir des valeurs comprises dans un domaine fini ou non.

Nous pouvons maintenant mieux définir ce qu'est un espace de Riemann.

Définition: Un "**espace de Riemann**" est une variété à laquelle nous avons attaché une métrique. Cela signifie que, dans chaque partie de la variété, représentée analytiquement au moyen d'un système de coordonnées $\mathcal{U}^i$, nous nous sommes donné une forme différentielle quadratique:

$$ds^2 = g_{ij} du^i du^j \quad (22.21)$$

qui constitue la métrique de l'espace.

Les coefficients g_{ij} ne sont pas entièrement arbitraires et doivent vérifier, nous l'avons démontré dans le chapitre de Calcul Tensoriel, les conditions suivantes:

C1. Les composantes sont symétriques $g_{ij} = g_{ji}$.

C2. Le déterminant de la matrice $[g_{ij}]$ est différent de zéro.

C3. La forme différentielle de l'élément linéaire, et par conséquent le concept de distance défini par les g_{ij} , est invariante vis-à-vis de tout changement de coordonnées.

C4. Toutes les dérivées partielles d'ordre deux des g_{ij} existent et sont continues donc de classe C^2 .

Un espace de Riemann est donc un espace de points, chacun étant repéré par un système de n coordonnées $\mathcal{U}^i$, doté d'une métrique quelconque telle que la forme différentielle de l'élément linéaire vérifiant les conditions précédentes. Cette métrique est dite dès lors "**métrique Riemannienne**".

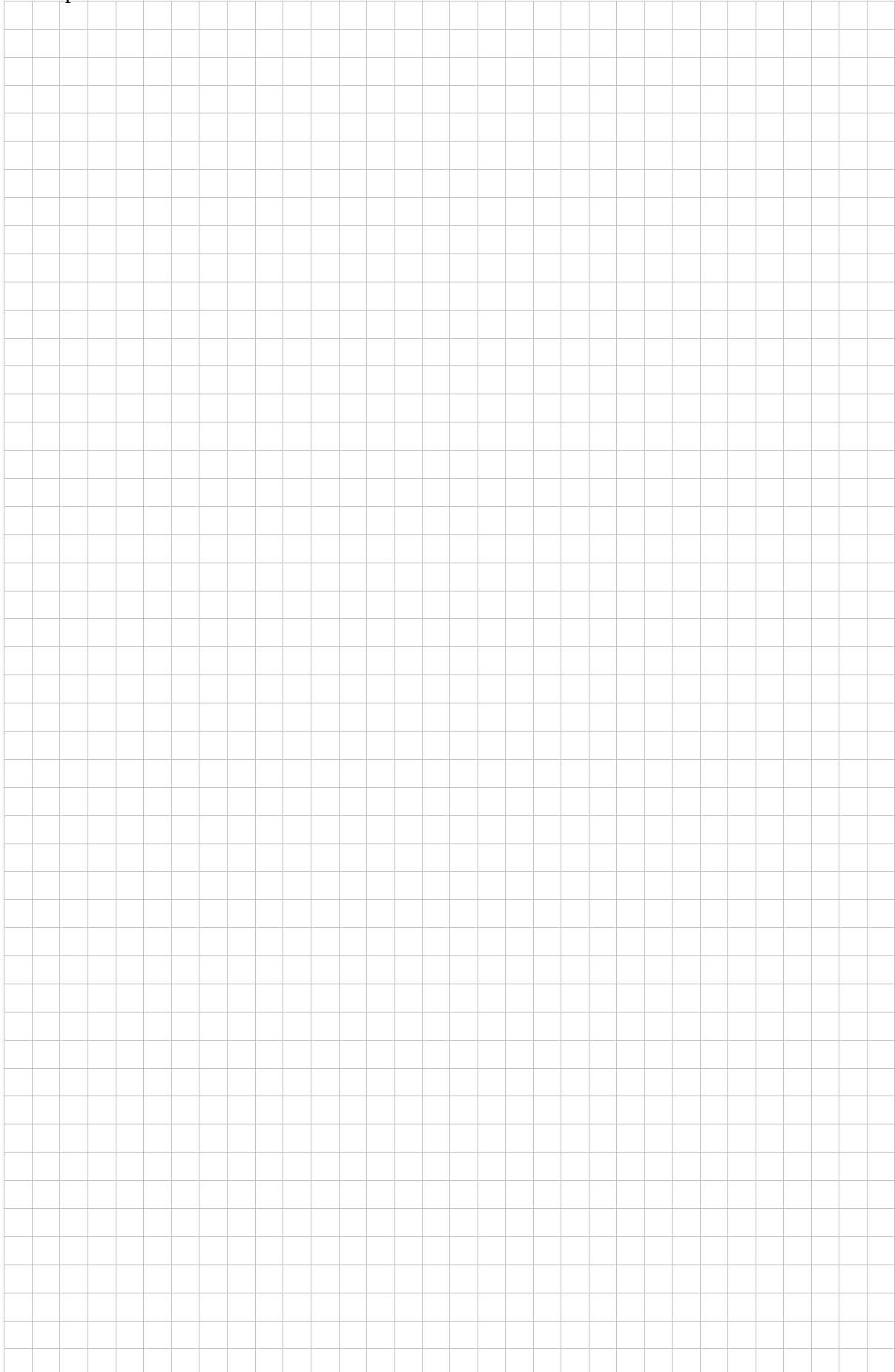
Remarques:

R1. Si la métrique est définie positive, c'est-à-dire si $g_{ij}v^i v^j > 0$ pour tout vecteur $\vec{v}$ non nul, nous disons que l'espace est "**proprement Riemannien**". Dans ce cas, le déterminant de la matrice $[g_{ij}]$ est strictement positif et toutes les valeurs propres de la matrice $[g_{ij}]$ sont strictement positives.

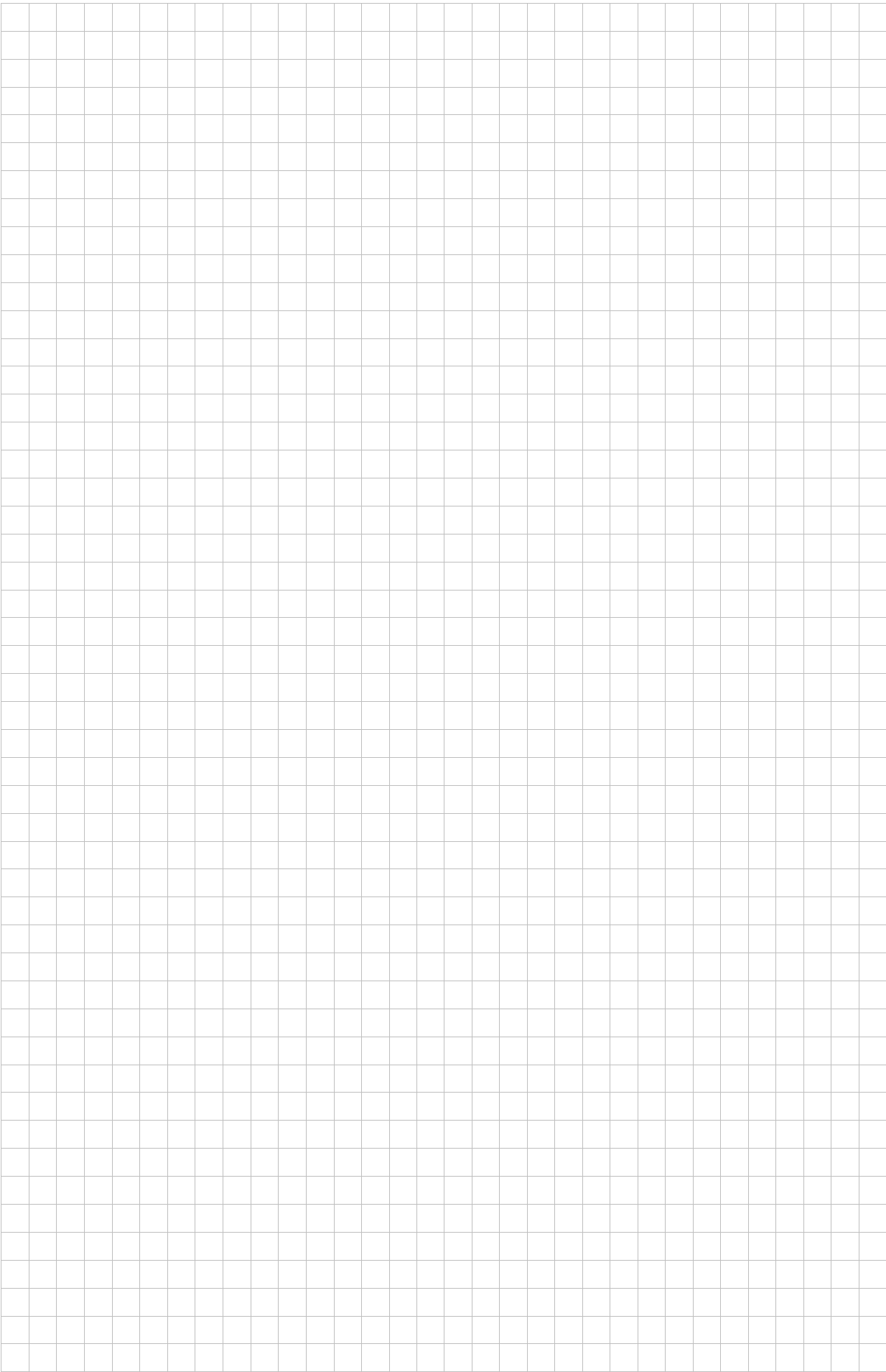
R2. Par définition, nous disons qu'une métrique d'un espace est euclidienne lorsque tout tenseur fondamental de cet espace peut être ramené, par un changement approprié de coordonnées, à une forme telle que (cf. chapitre de Calcul Tensoriel) la base orthonormée canonique: $g_{ij} = \delta_{ij}$.

R3. La définition des espaces Riemanniens montre que l'espace euclidien est un cas très particulier de ces espaces. Il n'existe donc qu'un seul espace euclidien alors que nous pouvons créer une infinité d'espaces Riemanniens.

Notes personnelles:









23. GÉOMÉTRIE PROJECTIVE

Depuis Chasles et jusque dans les années 1930, la géométrie projective était souvent synonyme de "géométrie supérieure". Elle s'opposait à la géométrie euclidienne: élémentaire... et analytique. À l'époque de Monge, de Carnot, de von Staudt, on parlait aussi de "géométrie de position" ou de "situation". Ces géométries étudient les figures au point de vue de leurs positions respectives et des propriétés invariantes qui les lient dans une transformation géométrique (rotation, symétrie, homothétie, etc.) homographique tout particulièrement. Outre la division harmonique, notion de base, elle fait appel au célèbre rapport anharmonique (birapport), à l'inversion, l'involution, la transformation par polaires réciproques, la projection stéréographique, la corrélation, l'homologie, la dualité, les coniques, etc. (voir plus loin dans ce texte pour les définitions).

La géométrie projective très abstraite en dehors de quelques généralités et principes de base exposés ci-dessous. Elle est donc relativement difficile à appréhender et requiert, avant de s'y plonger, une bonne connaissance en géométrie élémentaire voire, dans l'espace, une parfaite maîtrise de l'espace tridimensionnel, dans le cycle observation-représentation-interprétation. Ceci nous a amené à introduire certains concepts qui n'ont normalement pas leur place ici, mais qui nous le pensons, peuvent aider grandement l'intéressé à mieux comprendre cette branche des mathématiques.

Dans un premier lieu, nous allons étudier les concepts élémentaires de la perspective en s'attardant particulièrement sur le concept de présentation "projective" (il existe d'autres méthodes de perspective empiriques: cavalière, isométrique, militaire, ... mais ces dernières n'ont pas de sens mathématique ou réel même si elles représentent assez convenablement des objets volumiques). Ensuite, nous étudierons les représentations mathématiques de quelques objets tridimensionnels dans le cadre d'applications pour enfin passer à l'étude de la géométrie projective dure. Enfin, nous étudierons la mathématique utilisée dans la représentation informatique des formes géométriques (splines, courbes de Bézier, etc...)

Remarque: La "géométrie descriptive" est une forme artistique rigoureuse de la géométrie projective mais non formelle (dans le sens qu'elle ne s'étudie pas mathématiquement).

1. PERSPECTIVE CONIQUE (CENTRALE)

Un des problèmes de l'étude des volumes tridimensionnels et de leur représentation est le concept de "**perspective**". Effectivement, l'être humain ne peut voir les 3 dimensions d'un objet, c'est le cerveau qui interprète les ombres et reflets d'un objet afin que nous l'interprétions comme ayant un volume (il existe des illusions d'optique qui vont dans ce sens...: les trompe l'oeil).

Nous allons nous intéresser dans les paragraphes qui suivent à la "perspective conique", aussi appelée "perspective centrale" ou encore "linéaire".

Remarque: Dans le domaine de la géométrie projective nous ne parlons pas de "perspective conique" mais de "projection conique".

Définition: La "**perspective conique**" ou "**projection perspective**" est par construction la représentation la plus proche de nos perceptions visuelles, elle permet notamment de voir une sphère comme un cercle.

Remarque: La perspective conique est celle des peintres de la Renaissance. C'est aussi celle qui apparaît sur une photo.

La difficulté de la représentation en perspective est de traduire dans un plan (celui de la feuille de papier par exemple ou l'écran d'ordinateur) une construction qui est définie - de manière assez simple, d'ailleurs - dans l'espace à l'aide d'outils mathématiques.

Une projection conique de l'espace usuel à 3 dimensions est donc transformation projective qui envoie tous les points de cet espace sur un même plan de cet espace. Elle nécessite la donnée d'un point O (équivalent à la position de l'oeil de l'observateur) et d'un plan de projection appelé "tableau" ou encore "Vitre de Dürer" (l'équivalent de la rétine).

Nous verrons lors des développements théoriques que contrairement aux projections affines (voir le sous-chapitre suivant), la projection conique ne conserve pas le barycentre (donc les rapports de longueurs sur une droite donnée) mais elle conserve l'alignement et le birapport.

Lorsque nous parlons de perspective conique, nous utilisons quelques plans et droites particuliers dans l'espace tridimensionnel (voir figure plus bas):

- Le "**plan du tableau**" ou "**vitre de Dürer**", noté T , est le plan sur lequel nous faisons le dessin (plan de projection).
- Le "**plan du sol**", noté S , est un plan fixé, perpendiculaire au tableau T .
- Le "**point de vue**" (ou "**centre de projection**"), noté O , est un point hors de T et de S : c'est le point où devra se placer l'oeil pour que le dessin sur le tableau T coïncide avec l'image réelle.
- Le "**plan de l'horizon**", noté H , est le plan parallèle au plan du sol S passant par le point de vue O .
- La "**ligne d'horizon**", notée h , est l'intersection du plan d'horizon H et du tableau T .
- La "**ligne de Terre**", notée LT , est l'intersection du plan de sol S et du tableau T .
- Un plan ou une droite parallèles au plan du sol S sont appelés "**horizontaux**".
- Un plan ou une droite perpendiculaires au plan du sol S sont appelés "**verticaux**".
- Un plan ou une droite parallèles au tableau T sont appelés "**frontaux**".
- Un plan ou une droite perpendiculaires à T sont appelés de "**bout**".

Voici un schéma qui représente ces différentes notions:

L'objectif à partir de ce schéma est de déterminer mathématiquement une représentation d'un objet volumique sur une surface (plan dans un cas simple, quelconque sinon) en connaissant l'équation de la droite entre le point M et le point V (point de vue) afin d'en déterminer les coordonnées d'intersection entre cette droite et le tableau.

Nous pouvons du schéma ci-dessus tirer les relations suivantes selon les théorèmes de Thalès (cf. chapitre de Géométrie Euclidienne) dans les différents triangles:

$$\frac{x}{x'} = \frac{z-h}{z'-h} = \frac{y+\Delta}{y'+\Delta} \quad (23.1)$$

avec $\Delta \geq 0$.

Si nous posons $y' = 0$, conformément à ce que nous montre le schéma ci-dessus, les coordonnées du point m deviennent:

$$x' = x \frac{\Delta}{y+\Delta} \text{ et } z' = h + (z-h) \frac{\Delta}{y+\Delta} \quad (23.2)$$

À partir de ces relations, le problème de la représentation plane d'une forme volumique est complètement résolu, puisque nous pouvons toujours projeter un point (ou la distance entre deux points) sur un tableau à partir des coordonnées de l'original.

Le terme Δ est communément appelé la "**longueur focale**" du point de vue à l'écran et les spécialistes de l'optique le notent habituellement par la lettre f .

Pour mieux comprendre ce dernier résultat, nous pouvons nous mettre dans le contexte d'une étude bidimensionnelle où l'observateur est à la hauteur de la ligne de Terre ($h=0$) et disposé de façon à représenter une personne regardant le tableau (assimilable à un écran télé, ordinateur ou tout autre) dans lequel nous posons les axes conventionnels x et y dans le plan de l'écran et Z perpendiculairement (ainsi, relativement au premier schéma, Y devient Z et inversement).

Ainsi, la relation précédente des rapports:

$$\frac{x}{x'} = \frac{z-h}{z'-h} = \frac{y+\Delta}{y'+\Delta} \quad (23.3)$$

devient avec ce changement d'axes:

$$\frac{x}{x'} = \frac{y-h}{y'-h} = \frac{z+\Delta}{z'+\Delta} \quad (23.4)$$

et comme $h=0$ (ce qui est souvent le cas devant des écrans):

$$y' = \frac{y(z'+\Delta)}{z+\Delta} \quad (23.5)$$

Remarque: Cette relation est une forme particulière de ce que nous appelons les "**transformations homographiques**". Nous reviendrons plus loin sur celles-ci et démontrerons certaines de leurs propriétés.

Si le tableau est posé sur l'axe du référentiel (le tableau de projection est assimilé à l'écran), nous avons alors $z'=0$ ce qui nous donne:

$$y' = \frac{y \cdot \Delta}{z + \Delta} \quad (23.6)$$

et en procédant de même:

$$x' = \frac{x \cdot \Delta}{z + \Delta} \quad (23.7)$$

Remarque: Ces deux dernières relations sont celles que nous utilisons pour faire des animations 3D programmées dans le logiciel Macromedia Flash 6.0 (par exemple).

Nous voyons sur les deux dernières relations un terme identique:

$$P = \frac{\Delta}{z + \Delta} \quad (23.8)$$

ce terme correspond à la "**profondeur**" de la perspective.

Dans certains ouvrages, cette profondeur est notée (simple mise en facteur):

$$P = \frac{1}{1 + \frac{z}{\Delta}} \quad (23.9)$$

Si nous considérons deux points (x_1, x_2 ou y_1, y_2) visibles de la surface d'un volume vu par un observateur et leur distance respective Δ_x ou Δ_y , ces grandeurs se conservent si les deux points se confondent dans le plan du tableau car nous avons alors:

$$P=1 \quad (23.10)$$

puisque $z = 0, \forall \Delta$.

Il est intéressant d'étudier quelle doit être la valeur de la focale pour $z \neq 0$ afin d'avoir $\Delta x' = \Delta x$ ou $\Delta y' = \Delta y$. Ainsi, si nous prenons la limite:

$$\lim_{\Delta \rightarrow \infty} P = \frac{\Delta}{z + \Delta} \quad (23.11)$$

en appliquant la règle de l'Hospital (dérivée au numérateur et dénominateur) et en se rappelant que z est fixé, alors:

$$\lim_{\Delta \rightarrow \infty} P = \frac{\Delta}{z + \Delta} = \frac{\frac{d}{d\Delta} \Delta}{\frac{d}{d\Delta} (1 + \Delta)} = \frac{1}{0 + 1} = 1 \quad (23.12)$$

De ce résultat, nous pouvons conclure la chose suivante:

Pour que toute distance réelle entre deux points non confondus dans le tableau mais se trouvant dans un même plan aient une distance de projection égale, il faut que les équations des deux droites qui déterminent leur intersection avec le tableau soient parallèles. Ce qui implique, puisque l'observateur est un point convergent, qu'il faut éloigner l'observateur à une distance infinie du plan pour conserver les grandeurs ainsi projetées sur le tableau: c'est la "**projection parallèle orthogonale**" aussi parfois appelée "**projection parallèle orthographique**".

Un très bon exemple pour visualiser ces résultats est de programmer en pseudo-3D sur un ordinateur.

Exemple:

Il existe de nombreuses manières de faire de la 3D avec l'informatique. Les plus connues techniquement parlant sont avec OpenGL ou DirectX ou du C++ mais ne sont pas très faciles à aborder... nous allons donc voir comment faire tourner une pseudo-sphère dans l'espace projectif avec Macromedia Flash 6.0 dans le but de montrer comment s'appliquent les différents éléments théoriques présentés plus haut mais aussi de montrer que ce ne sont pas les seuls outils disponibles.

PS: Je regrette d'avoir fait l'exemple avec un logiciel peu accessible à tous. Si j'ai du temps je referai l'exemple avec un logiciel comme Microsoft Office PowerPoint.

Pour cela, ouvrez le logiciel Macromedia Flash 6.0 et enregistrez la nouvelle animation sous le nom Cercle fla:

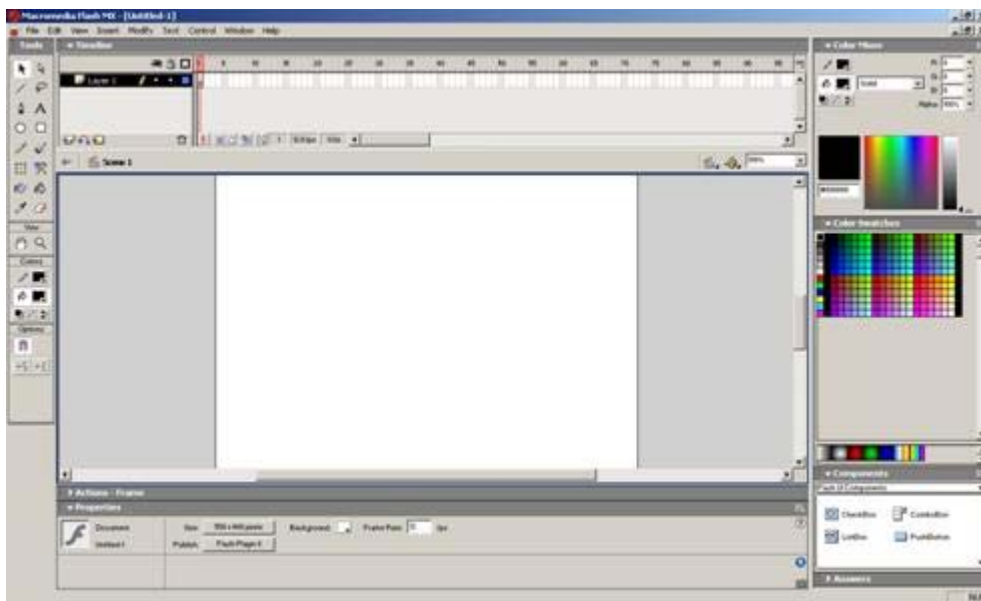


Figure: 23.3 - Interface de Macromedia Flash 6.0

Avec l'outil *Cercle* dans la barre de dessin, dessinez un cercle de dimension respectable dans la zone d'animation:



Figure: 23.4 - Dessin d'un cercle

Ensuite après avoir sélectionné votre cercle, avec l'outil *Remplissage*  choisissez un *Dégradé Radial*:

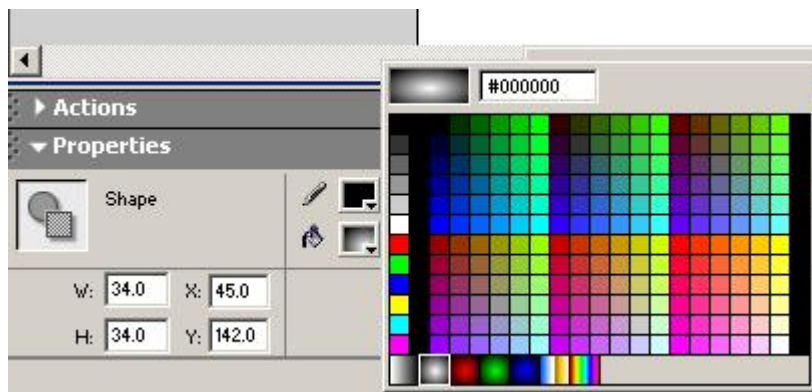


Figure: 23.5 - Palette de couleurs pour remplissage du cercle

Faites un clic-droit sur le cercle et choisissez l'option *Convert to Symbol* et saisissez les informations telles que présentées ci-dessous:



Figure: 23.6 - Boîte de dialogue de conversion d'objet en symbole

Renommez le calque où se trouve votre movie clip avec le nom *3d clip*:



Figure: 23.7 - Renommage du calque d'animation

Faites un double-clic sur votre cercle pour entrer dans votre Movie Clip.

Sélectionnez-y à nouveau le cercle, faites un clic-droit dessus et sélectionnez *Convert to Symbol*:

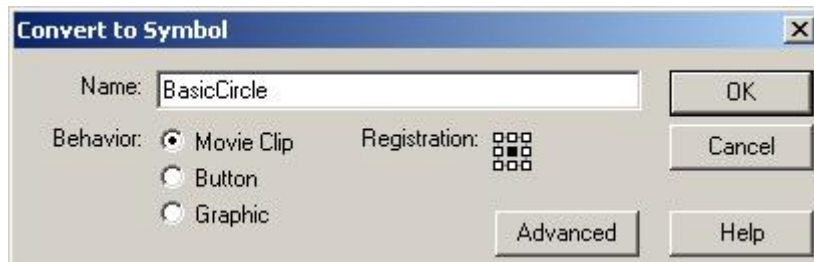


Figure: 23.8 - Définition des plans et droites de la géométrie projective

Ensuite, dans les propriétés du cercle saisissez-y le nom *Point*:

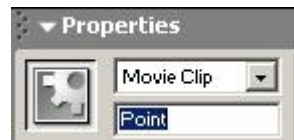


Figure: 23.9 - Double conversion en symbole

Maintenant, dans le Movie Clip nommé *Cercle*, nous allons insérer trois frames: la première pour définir les fonctions mathématiques recalculant les différentes variables, la deuxième qui appelle les fonctions, la troisième qui permet de relancer en boucle et indéfiniment la deuxième.

Afin de faire les choses au propre, nous allons créer un deuxième calque (en renommant celui qui contient notre cercle en *Cercle*) que nous appellerons *Code*:

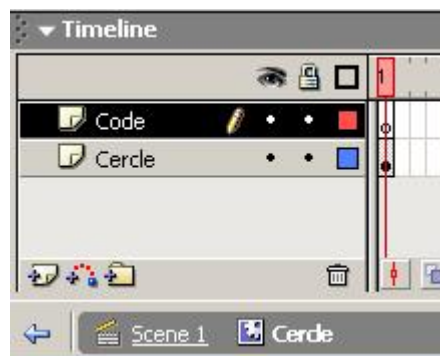


Figure: 23.10 - Selon l'usage on fait un calque de code

Faisons un clic-droit sur la troisième image du calque contenant notre cercle:

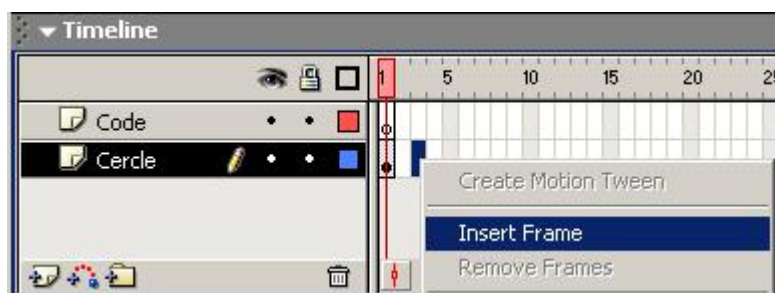


Figure: 23.11 - Préparation d'une animation de base sur 3 images

et choisissons l'option *Insert Frame* et pour le calque *Code* faites presque de même, mais en choisissant l'option *Insert Key Frame*. Vous devriez alors obtenir le visuel suivant:



Figure: 23.12 - Résultat à obtenir

Ensuite en sélectionnant la première image du calque *Code* activez l'affichage des Actions pour y insérer le code suivant:

```
Radius = 100;
degrees = 0;
Xcenter=0;
Ycenter=0;
speed=2;

function setAngle () {
    Angle = degrees * (Math.PI / 180);
    degrees = degrees + speed;
}

function drawPoints () {
    Point._x = Radius * Math.cos(Angle)+Xcenter;
    Point._y = Radius * Math.sin(Angle)+Ycenter;
}
```

Faites de même avec la deuxième image du calque *Code*, mais en mettant:

```
setAngle();
drawPoints();
```

et enfin de même avec la troisième image, mais en y mettant:

```
gotoAndPlay(2);
```

Nous obtenons alors le résultat suivant:

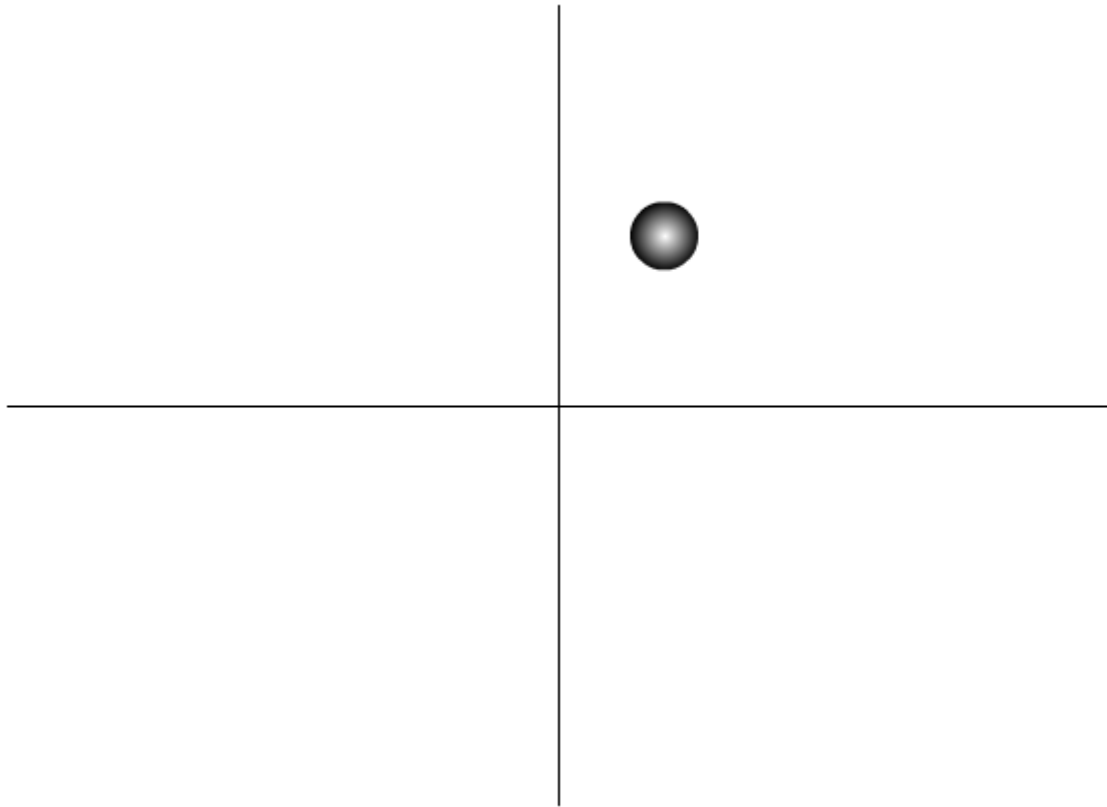


Figure: 23.13 - Animation Flash 6.0 du résultat intermédiaire

Nous allons maintenant faire intervenir l'axe Z en fixant Y (et en faisant donc bouger Z). Bien évidemment, nous ne verrons rien se passer en ce qui concerne Z tant que nous ne définissons pas la projection homographique et parce qu'un ordinateur est incapable de montrer basiquement le concept de profondeur... Le code s'écrit alors:

```
Radius = 100;
degrees = 0;
//coordonnées X,Y,Z dans le repère 3D imaginé
Xcenter=0;
Ycenter=0;
Zcenter=0;
speed=2;

function setAngle () {
    Angle = degrees * (Math.PI / 180);
    degrees = degrees + speed;
}

function drawPoints () {
    Point._x = Radius * Math.cos(Angle)+Xcenter;
    Point._y = Ycenter;
    Zpos = Radius * Math.sin(Angle)+Zcenter;
}
```

Le calcul de *Zpos* va nous permettre plus loin de calculer la profondeur du mouvement de l'objet selon l'axe Z. Et c'est là qu'interviendra la projection homographique.

Nous obtenons alors une pseudo-sphère qui tourne autour d'un axe dans un plan perpendiculaire à l'écran. Raison pour laquelle nous voyons la pseudo-sphère faire des allers-retours gauche/droite (le concept d'éloignement n'est pas encore présent faute de présence d'un facteur profondeur):

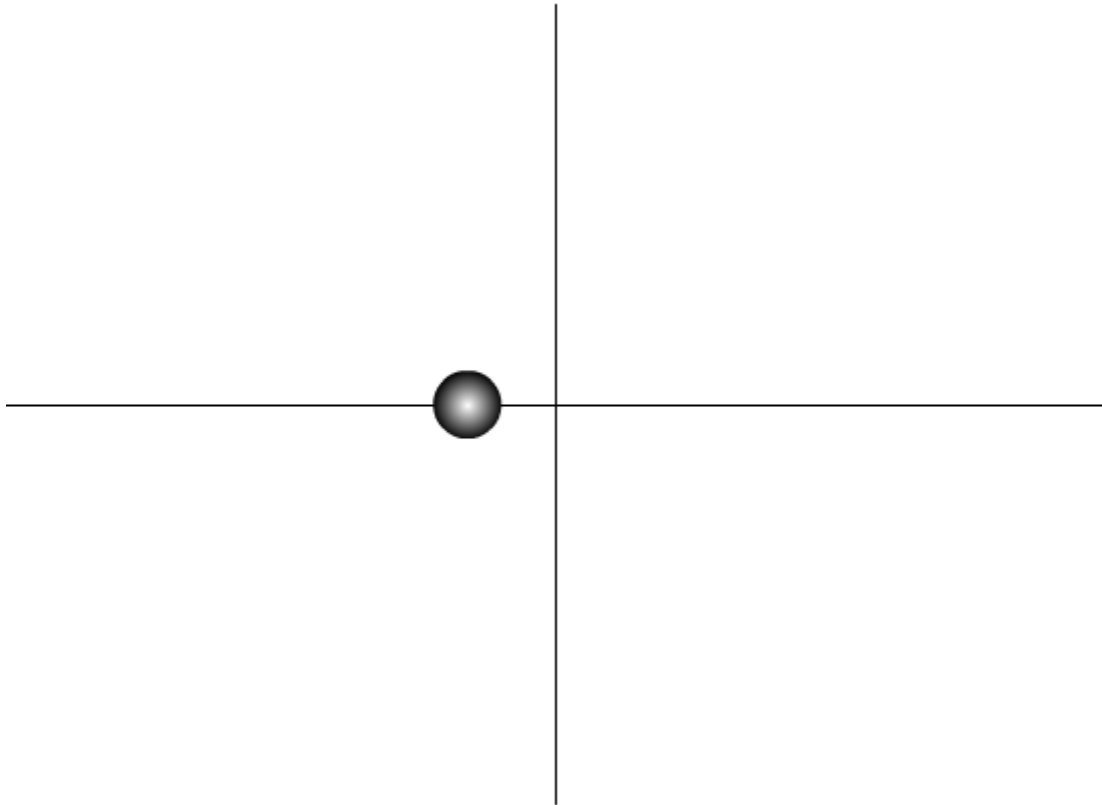


Figure: 23.14 - Animation Flash 6.0 du résultat intermédiaire

Maintenant nous allons utiliser les relations:

$$\begin{aligned}x'(z) &= x \cdot P = x \frac{f}{z + f} \\ y'(z) &= y \cdot P = y \frac{f}{z + f}\end{aligned}\quad (23.13)$$

démontrées plus haut. Si nous cherchons à représenter la profondeur de tout point du tableau de projection c'est le rapport de distance entre deux des points de ce tableau qui vont nous intéresser pour déterminer le changement d'échelle:

$$\begin{aligned}\frac{x'_2(z) - x'_1(z)}{x_2 - x_1} \% &= \frac{f}{z + f} \cdot 100 = P \cdot 100 \\ \frac{y'_2(z) - y'_1(z)}{y_2 - y_1} \% &= \frac{f}{z + f} \cdot 100 = P \cdot 100\end{aligned}\quad (23.14)$$

Il faudra donc appliquer ce résultat comme définissant l'échelle du tableau de projection.

Nous avons alors le code suivant où la profondeur P joue sur la hauteur et la largeur de la surface d'animation de l'instance *Point* de notre pseudo-sphère:

```
Radius = 275;
degrees = 0;
//coordonnées X,Y,Z dans le repère 3D imaginé
Xcenter=0;
Ycenter=100;
Zcenter=0;
speed=2;
f=400;

function setAngle () {
    Angle = degrees * (Math.PI / 180);
    degrees = degrees + speed;
}

function drawPoints () {
    Point._x = Radius * Math.cos(Angle)+Xcenter;
    Point._y = Ycenter;
    Zpos = Radius * Math.sin(Angle)+Zcenter;
    P=f/(f+Zpos);
    //Puisque les échelles des objets sont
    //en % dans Flash
    _xscale = _yscale= P*100;
    //on joue avec la transparence de l'objet
    _alpha= P*100;
}
```

Ce qui nous donne:

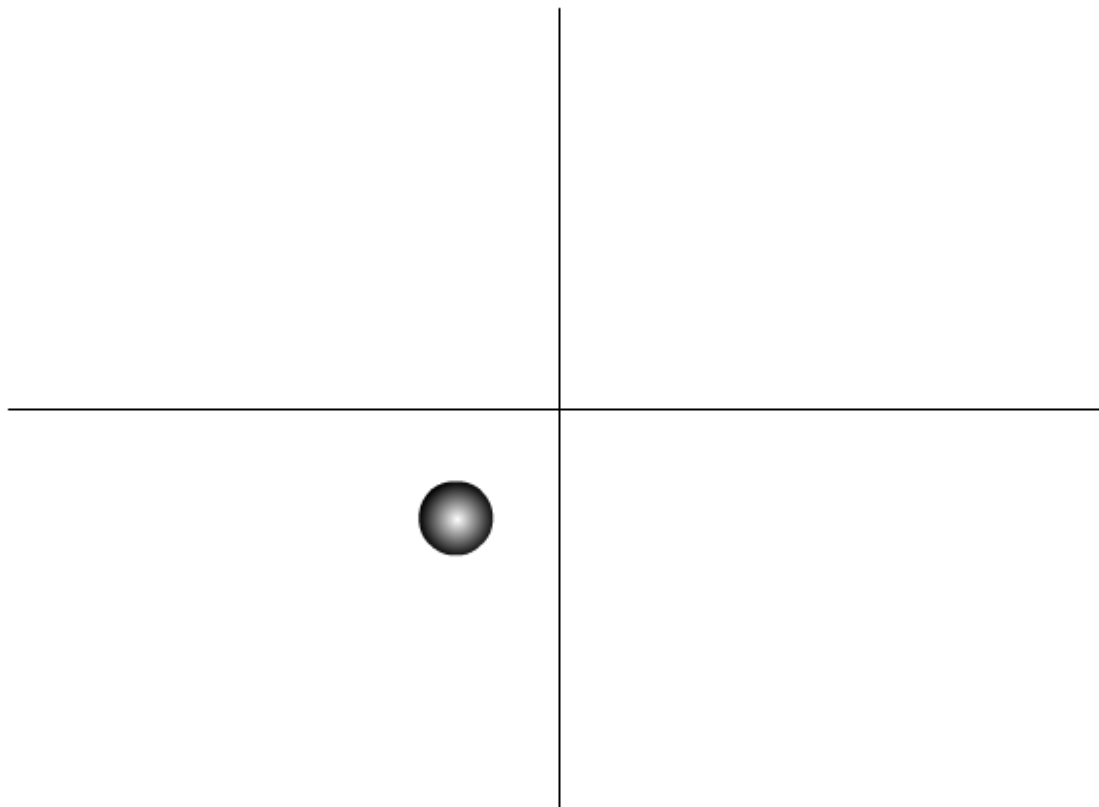


Figure: 23.15 - Animation Flash 6.0 du résultat intermédiaire

et bien évidemment cela fonctionne quelle que soit l'équation paramétrique de la trajectoire! Nous pouvons ensuite copier cette instance d'animation, changer la hauteur, l'angle de départ pour avoir les 4 sommets d'un cube qui tourne dans l'espace.

C'est la prochaine étape:

Effectivement, changeons notre code comme indiqué ci-dessous pour avoir quatre pseudo-sphères tournant autour d'un axe Z imaginaire sortant de l'écran (nous utilisons les matrices de rotation démontrées dans le chapitre de Géométrie Euclidienne):

```
degrees = 0;
speed=1;
Xcenter=0;
Ycenter=0;

// On construit les coordonnées X des 4 pseudo sphères
x = new Array(-50, 50, -50, 50);
// et leurs coordonnées Y respectives dans un tableau à 4 dimensions
y = new Array(-50, -50, 50, 50);
// On compte le nombre d'éléments dans le tableau X (le même qu'en Y)
verticeNum = x.length;

// On duplique les points en les nommant Point0, Point1,... avec une boucle
for (c = 0; c < verticeNum; c++) {
    duplicateMovieClip ("Point", "Point" + c, c);
}
// On masque le Movie Clip ayant permis de générer les 4 pseudo-sphères
Point._visible = 0;

// fonction setAngle
function setAngle () {
    Angle = degrees * (Math.PI / 180);
    degrees = degrees + speed;
}

// fonction déjà connue...
function drawPoints () {
    for (c = 0; c < verticeNum; c++) {
        // On fait tourner autour l'axe Z sortant de l'écran
        // mais en utilisant la matrice de rotation vue plus haut
        Xpos = x[c] * Math.cos (Angle) + y[c] * Math.sin(Angle);
        Ypos = - x[c] * Math.sin (Angle) + y[c] * Math.cos(Angle);

        // et on repositionne les pseudo-sphères à l'aide de leurs coordonnées
        this["Point" + c]._x = Xpos + Xcenter;
        this["Point" + c]._y = Ypos + Ycenter;
    }
}
```

Ce qui nous donne en charme... quatre pseudo-sphères tournant autour d'un centre commun:

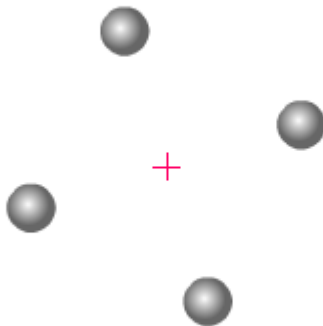


Figure: 23.16 - Animation Flash 6.0 intermédiaire

Maintenant, nous inversons Y et Z à nouveau et appliquons la projection homographique:

```
degrees = 0;
speed=1;
Xcenter=0;
Ycenter=0;
Zcenter=0
f = 150;

//nous avons changé Y en Z sinon rien de particulier
x = new Array(-50, 50, -50, 50);
z = new Array(-50, -50, 50, 50);
verticeNum = x.length;

// rien de nouveau
for (c = 0; c < verticeNum; c++) {
    duplicateMovieClip ("Point", "Point" + c, c);
}
Point._visible = 0;

// rien de nouveau
function setAngle () {
    Yangle = degrees * (Math.PI / 180);
    degrees = degrees + speed;
}
```

```

function drawPoints () {
  for (c = 0; c < verticeNum; c++) {
    // On tourne autour de Y
    Xpos = x[c] * Math.cos (YAngle) - z[c] * Math.sin (YAngle);
    Zpos = x[c] * Math.sin (YAngle) + z[c] * Math.cos (YAngle);

    // La projection habituelle
    P = f / (f+Zpos);

    // et on redessine les points
    this["Point" + c]._x = Xpos * P + Xcenter;
    this["Point" + c]._y = Ycenter
    // on applique l'effet d'échelle à chaque sphère indépendamment
    this["Point" + c]._xscale = this["Point" + c]._yscale = P * 100;

    // On utilise une fonction de Macromedia Flash permettant
    // de dessiner les points sur des couches différentes sinon
    // les pseudo-sphère en se croisant se superposent
    this["Point" + c].swapDepths(P * 500);
  }
}

```

Et nous obtenons:



Figure: 23.17 - Animation Flash 6.0 intermédiaire

Pour la suite, nous allons générer 8 pseudo-sphère et au lieu de les faire tourner toujours autour du même axe, nous allons les faire tourner autour des 3 axes *X*, *Y* ou *Z* en utilisant les variables *XAngle*, *YAngle* ou *ZAngle* et les matrices de rotation autour de chacun de ces axes respectifs:

```

Xdegrees = 0;
Ydegrees = 0;
Zdegrees = 0;
Xcenter = 0;
Ycenter = 0;
Zcenter = 0;
f = 150;

// Nous générons les 8 pseudo-sphères
x = new Array(-50, 50, -50, 50, -50, 50, -50, 50);
y = new Array(-50, -50, 50, 50, -50, -50, 50, 50);
z = new Array(50, 50, 50, 50, -50, -50, -50, -50);
verticeNum = x.length;

// Duplication des pseudo-sphères en Point1,Point2...
for (c = 0; c < verticeNum; c++) {
    duplicateMovieClip ("Point", "Point" + c, c);
}
// Nous masquons le point d'héritage
Point._visible = 0;

// Nous définissons les angles selon chaque axe
function setAngle () {

    Xangle = Xdegrees * (Math.PI / 180);
    Yangle = Ydegrees * (Math.PI / 180);
    Zangle = Zdegrees * (Math.PI / 180);
    //l'angle sera déterminé par la position de la souris
    Xdegrees = this._ymouse;
    Ydegrees = this._xmouse;
    Zdegrees = sqrt(this._xmouse^2+this._ymouse);
}

// Nous calculons les points de chaque sphère
function drawPoints () {
    for (c = 0; c < verticeNum; c++) {
        // On tourne les pseudo-sphères autour de l'axe X en utilisant Xangle
        // et la matrice de rotation
        Ypos_Temp = y[c] * Math.cos (Xangle) + z[c] * Math.sin (Xangle);
        Zpos_Temp = - y[c] * Math.sin (Xangle) + z[c] * Math.cos (Xangle);

        // On tourne les pseudo-sphères autour de l'axe Y en utilisant Yangle
        // et la matrice de rotation
        Xpos_Temp = x[c] * Math.cos (Yangle) - Zpos_Temp * Math.sin (Yangle);
        Zpos = x[c] * Math.sin (Yangle) + Zpos_Temp * Math.cos (Yangle);

        // On tourne les pseudo-sphères autour de l'axe Z en utilisant Xangle
        // et la matrice de rotation
        Xpos = Xpos_Temp * Math.cos (Zangle) + Ypos_Temp * Math.sin (Zangle);
        Ypos = - Xpos_Temp * Math.sin (Zangle) + Ypos_Temp * Math.cos (Zangle);

        // Perspective (j'ai rajouté le Zcenter pour être plus exact!
        P = f/ (f + Zpos + Zcenter);

        // On dessine
        this["Point" + c]._x = Xpos * P + Xcenter;
        this["Point" + c]._y = Ypos * P + Ycenter;
        this["Point" + c]._xscale = this["Point" + c]._yscale = P * 100;

        // Z-Sorting
        this["Point" + c].swapDepths(P * 500);
    }
}

```

Et voici le résultat final:

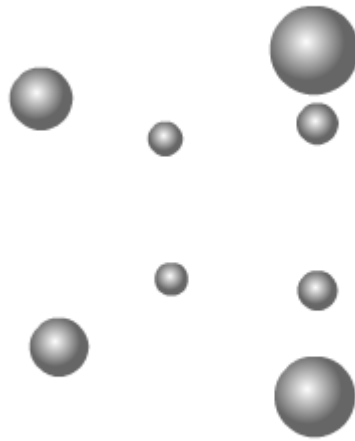


Figure: 23.18 - Animation Flash 6.0 finale

Comme le Flash est mort entre temps... un internaute à refait l'exemple ci-dessous avec du WebGL dont voici le code:

```
<html>
<head>
  <script type="text/javascript">
    window.onload = startDemo;

    function Point3D(x,y,z) {
      this.x = x;
      this.y = y;
      this.z = z;

      this.rotateX = function(angle) {
        var rad, cosa, sina, y, z
        rad = angle * Math.PI / 180
        cosa = Math.cos(rad)
        sina = Math.sin(rad)
        y = this.y * cosa - this.z * sina
        z = this.y * sina + this.z * cosa
        return new Point3D(this.x, y, z)
      }

      this.rotateY = function(angle) {
        var rad, cosa, sina, x, z
        rad = angle * Math.PI / 180
        cosa = Math.cos(rad)
```

```
sina = Math.sin(rad)
z = this.z * cosa - this.x * sina
x = this.z * sina + this.x * cosa
return new Point3D(x, this.y, z)
}

this.rotateZ = function(angle) {
    var rad, cosa, sina, x, y
    rad = angle * Math.PI / 180
    cosa = Math.cos(rad)
    sina = Math.sin(rad)
    x = this.x * cosa - this.y * sina
    y = this.x * sina + this.y * cosa
    return new Point3D(x, y, this.z)
}

this.project = function(viewWidth, viewHeight, fov, viewDistance) {
    var factor, x, y
    factor = fov / (viewDistance + this.z)
    x = this.x * factor + viewWidth / 2
    y = this.y * factor + viewHeight / 2
    return new Point3D(x, y, this.z)
}
}

var vertices = [
    new Point3D(-1,1,-1),
    new Point3D(1,1,-1),
    new Point3D(1,-1,-1),
    new Point3D(-1,-1,-1),
    new Point3D(-1,1,1),
    new Point3D(1,1,1),
    new Point3D(1,-1,1),
    new Point3D(-1,-1,1)
];

// Define the vertices that compose each of the 6 faces. These numbers are
// indices to the vertex list defined above.
var faces = [[0,1,2,3],[1,5,6,2],[5,4,7,6],[4,0,3,7],[0,4,5,1],[3,2,6,7]]

var angle = 0;

function startDemo() {
    canvas = document.getElementById("thecanvas");
    if( canvas && canvas.getContext ) {
        ctx = canvas.getContext("2d");
        setInterval(loop,33);
    }
}
```



```

function loop() {
    var t = new Array();

    ctx.fillStyle = "rgb(255,255,255)";
    ctx.fillRect(0,0,400,200);

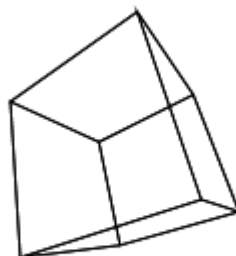
    for( var i = 0; i < vertices.length; i++ ) {
        var v = vertices[i];
        var r = v.rotateX(angle).rotateY(angle).rotateZ(angle);
        var p = r.project(400,200,128,3.5);
        t.push(p)
    }

    ctx.strokeStyle = "rgb(0,0,0)"

    for( var i = 0; i < faces.length; i++ ) {
        var f = faces[i]
        ctx.beginPath()
        ctx.moveTo(t[f[0]].x,t[f[0]].y)
        ctx.lineTo(t[f[1]].x,t[f[1]].y)
        ctx.lineTo(t[f[2]].x,t[f[2]].y)
        ctx.lineTo(t[f[3]].x,t[f[3]].y)
        ctx.closePath()
        ctx.stroke()
    }
    angle += 2
}
</script>
</head>
<body>
    <canvas id="thecanvas" width="400" height="200">
        Your browser does not support the HTML5 canvas element.
    </canvas>
</body>
</html>

```

Ce qui donne:



1.2. IMAGES DE DROITES

Déterminons à partir des résultats précédents, l'image d'une droite parallèle à la ligne de terre (donc à l'axe x) du tableau XY . Dans ce cas, nous avons:

$$z = a, y = b \quad (23.15)$$

où a et b sont des constantes et pour toute valeur de h . Ce qui nous donne d'après les relations obtenues plus haut:

$$\begin{aligned} x' &= \frac{\Delta}{\Delta + a} x \\ y' &= h + (b - h) \frac{\Delta}{\Delta + a} = \frac{ha + \Delta b}{\Delta + a} = c^{te} \end{aligned} \quad (23.16)$$

Donc toute droite parallèle à la ligne de Terre devient en perspective une droite se trouvant à une hauteur y' dans le plan parallèle à XY de notre écran (on pouvait le deviner intuitivement...).

Pour toute droite parallèle à l'axe Z de l'écran (donc dans sa "profondeur"), nous avons:

$$x = a, y = b \quad (23.17)$$

où a et b sont des constantes et pour toute valeur de h . Ce qui nous donne:

$$\left. \begin{aligned} x' &= \frac{\Delta}{\Delta + z} a \\ y' &= h + (b - h) \frac{\Delta}{\Delta + z} \end{aligned} \right\} \frac{\Delta}{\Delta + z} = \frac{x'}{a} \Rightarrow y' = h + (b - h) \frac{x'}{a} \quad (23.18)$$

Les droites d'équation:

$$y' = h + (b - h) \frac{x'}{a} \quad (23.19)$$

passent toutes par le point $P(x' = 0, y' = h)$ lorsque $z \rightarrow +\infty$ qui est le "**point de fuite principal**" et par le point $P(x' = a, y' = b)$ lorsque $z \rightarrow 0$ tel que représenté sur la projection ci-dessous faite dans Adobe Photoshop (les lignes horizontales ont été rajoutées pour donner l'effet de perspective):

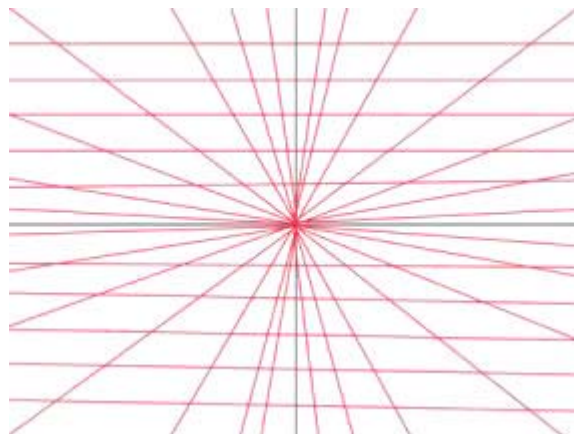


Figure: 23.19 - Exemple d'un point de fuite unique

À partir de la figure ci-dessus, nous pouvons définir le concept "d'angle de fuite" donné par la figure ci-dessous:

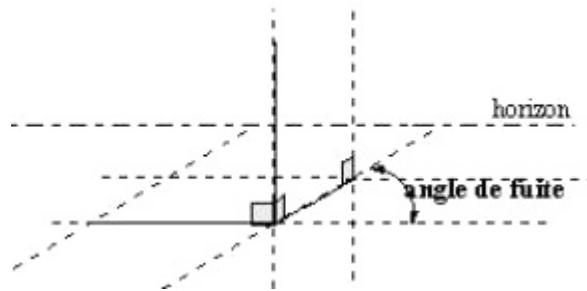


Figure: 23.20 - Exemple d'angle de fuite

Une représentation géométrique autre peut aider à mieux comprendre le résultat. Rappelons donc que le point de fuite d'une droite D est le point d'intersection F du plan du tableau T avec la droite parallèle à D passant par O . Deux droites parallèles D et D' ont donc le même point de fuite. Représenté comme ci-dessous:

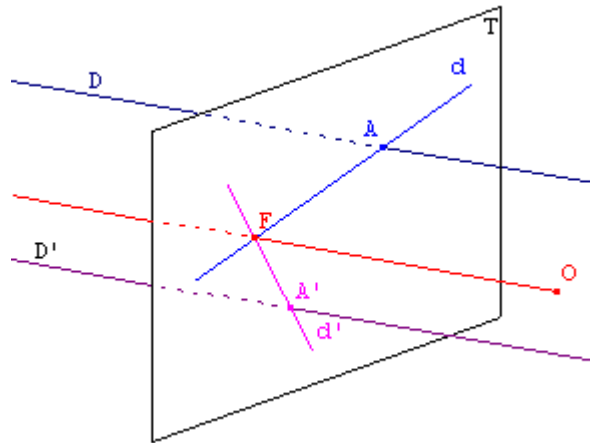


Figure: 23.21 - Autre représentation des points de fuite

Si nous notons A le point d'intersection de D avec T , le dessin en perspective de D est la droite $\overline{AF}$, intersection de T avec le plan contenant O et D . Puisque deux droites parallèles ont le même point de fuite F , elles sont donc représentées par deux droites sécantes en F .

Pour toute droite quelconque se situant dans le plan XZ de l'écran (donc dans sa "profondeur"), nous avons:

$$z = mx + p \text{ et pour } y = 0 \quad (23.20)$$

ce qui nous donne:

$$\begin{aligned} x' &= \frac{\Delta}{\Delta + z} x \Rightarrow x = \frac{\Delta + z}{\Delta} x' \\ y' &= h + (y - h) \frac{\Delta}{\Delta + z} = h - h \frac{\Delta}{\Delta + z} = \frac{hz}{\Delta + z} \end{aligned} \quad (23.21)$$

De cette dernière équation, nous déduisons:

$$z = -\frac{\Delta}{y' - h} y' \quad (23.22)$$

en portant z dans l'expression de x :

$$x = \left(1 + \frac{z}{\Delta}\right) x' = \left(1 - \frac{y'}{y' - h}\right) x' = -\frac{hx'}{y' - h} \quad (23.23)$$

nous remplaçons x et z dans l'équation de la droite $z = mx + p$, et calculs et simplifications faits, nous trouvons:

$$y' = \frac{mh}{\Delta + p} x' + \frac{ph}{\Delta + p} \quad (23.24)$$

Examinons le cas particulier de droites passant par les sommets opposés d'un carrelage, c'est-à-dire inclinées à $\pm\pi/4$ donc avec un coefficient directeur (une pente) de $m = \pm 1$.

L'image de ces droites est alors donnée par:

$$y' = \pm \frac{h}{\Delta + p} x' + \frac{ph}{\Delta + p} \quad (23.25)$$

Si $y' = h$, nous avons selon la relation ci-dessus:

$$x' = \pm\Delta \quad (23.26)$$

cela signifiant que toute projection de droites de coefficient directeur ± 1 se situant dans le plan XY pour $\forall x'$ constitue des points de fuite secondaires situés sur la ligne d'horizon à une distance égale de part et d'autre du point de fuite principal tel que représenté ci-dessous (dans le contexte d'un cours sur Adobe Photoshop, nous y avons rajouté un cube dans cette perspective):

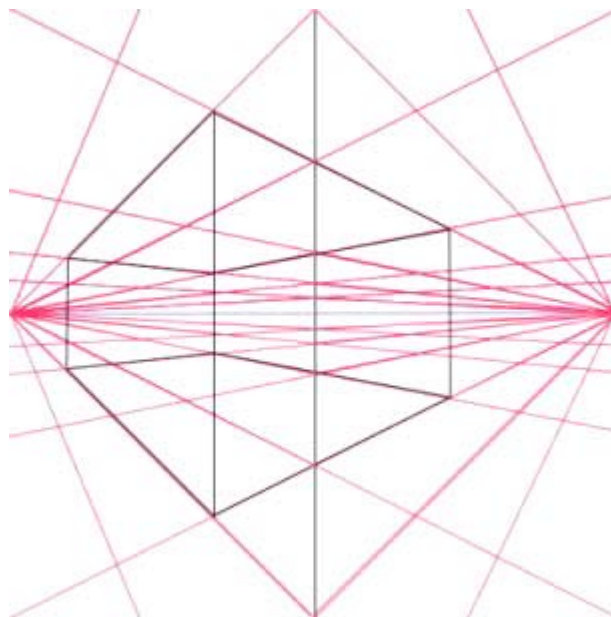


Figure: 23.22 - Exemple de point de fuite secondaire

Nous voyons bien dans la figure ci-dessus la symétrie par rapport à l'axe vertical et les deux points de fuites qui sont l'origine du carrelage.

Considérons maintenant des droites parallèles à l'axe Y de l'écran et leur équation de projection si $x = a$ et $z = b$:

$$\begin{aligned} x' &= \frac{\Delta}{\Delta + b} a = c^* \\ y' &= h + (y - h) \frac{\Delta}{\Delta + b} \end{aligned} \quad (23.27)$$

Ces droites images restent donc des droites parallèles à l'axe y . Autrement dit, les droites images restent parallèles aux droites "objet" tel que représenté ci-dessous (pour différentes positions du point d'observation):

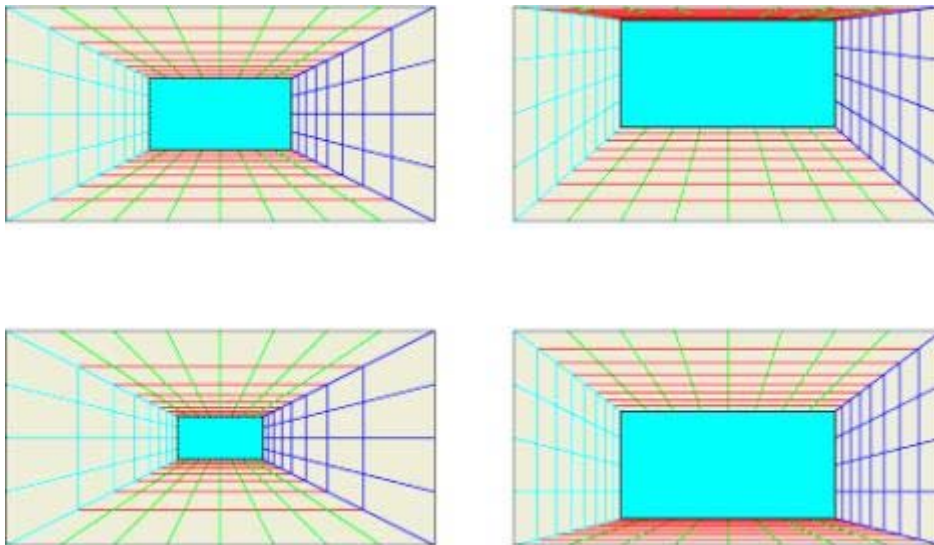


Figure: 23.23 - Différentes perspective en fonction des valeurs des constantes

Prenons pour exemple de ce dernier résultat, des segments de même hauteur H dont le pied est sur le plan horizontal:

$$y = H \quad x = a \quad (23.28)$$

La hauteur du segment image se déduit par:

$$y' = H' = h + (H - h) \frac{\Delta}{\Delta + b} \quad (23.29)$$

Considérons maintenant $x = a$, c'est-à-dire des colonnes verticales de même hauteur H alignées sur la droite $x = a$. Calculons les coordonnées des sommets images:

$$\begin{aligned} x'_s &= \frac{\Delta}{\Delta + b} a \Rightarrow \frac{\Delta}{\Delta + b} = \frac{x'_s}{a} \\ y'_s &= H' = h + (H - h) \frac{\Delta}{\Delta + b} = h + \frac{H - h}{a} x \end{aligned} \quad (23.30)$$

C'est l'équation d'une droite et remarquons que toutes ces droites joignant les sommets passent par le point de coordonnées $(x' = 0, y' = h)$ qui n'est d'autre que le point de fuite principal P .

Une représentation géométrique peut à nouveau aider à la compréhension des résultats précédents. Soit la figure suivante:

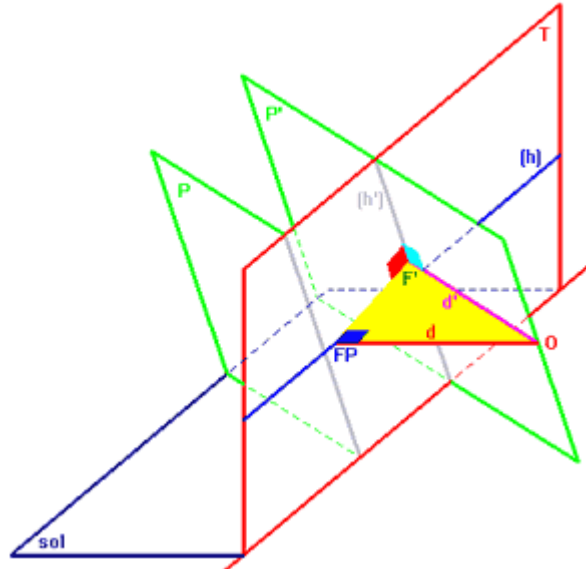


Figure: 23.24 - Représentation géométrique du cas particulier présenté plus haut

La ligne de fuite du plan P est la droite d'intersection h' du plan P' , parallèle à P et passant par O , avec le plan du tableau. Elle contient par conséquent les points de fuite de toutes les droites contenues dans P comme nous l'avons déjà démontré.

Le point de fuite principal du plan P est le projeté orthogonal F' du point O sur la droite h' . C'est le point de fuite des droites de P parallèles à $\overline{OF'}$, donc orthogonales à h' .

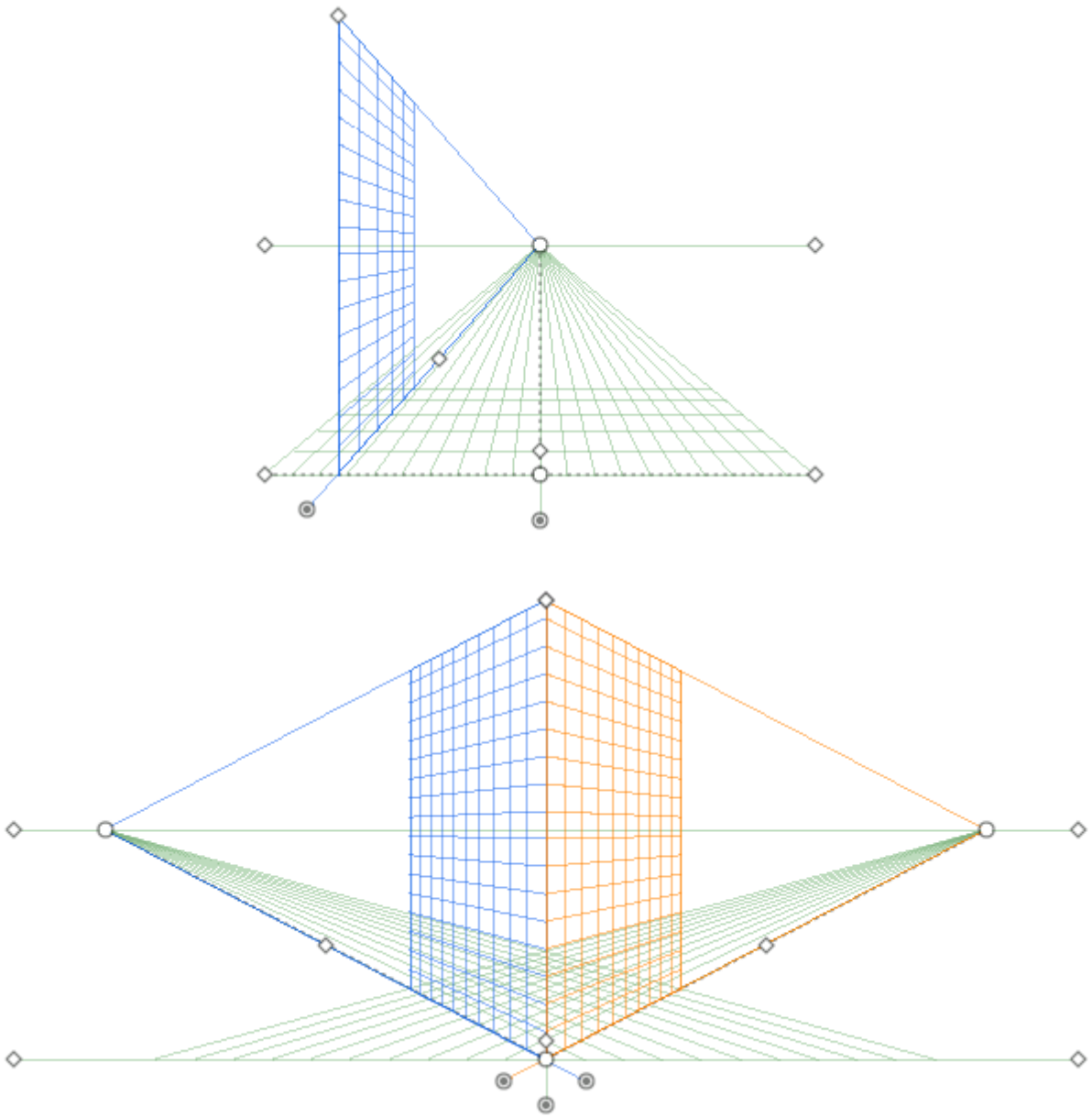
Il résulte de ces définitions que deux plans parallèles ont les mêmes éléments de référence.

À partir de l'exposé précédent, nous avons une nouvelle méthode pour présenter un objet en rotation dans un espace tridimensionnel. Au lieu de faire tourner l'objet autour de différents axes, nous pouvons imaginer à l'aide des équations ci-dessus faire tourner l'observateur autour de l'objet (c'est un point de vue...).

Nous n'avons considéré ici que la perspective projective sur un plan. Au fait, pour travailler sur des méthodes de projection quelconques (sphère sur plan, plan sur sphère, sphère sur sphère, n'importe quoi sur n'importe quoi) il suffit d'étendre l'analyse que nous avons faite ci-dessus dans un système de coordonnées adapté au système étudié (coordonnées polaires, cylindriques, sphérique, ...). C'est probablement ainsi que procèdent les logiciels de simulation 3D pour projeter une image sur une surface réfléchissante et semi-transparente tel qu'un verre ondulé.

Remarque: Des résultats que nous avons obtenus ci-dessus, nous pouvons tirer une conclusion intéressante et intuitive: Pour que l'observateur d'une photo ou d'un tableau voie l'image telle qu'elle était à l'origine, celui-ci doit se placer à des coordonnées déterminées du tableau (de la photo ou du tableau).

Un logiciel comme Adobe Illustrator propose depuis le début des années 2010 un outil pour créer des perspectives à un point de fuite, à deux points de fuite et à trois points de fuite:



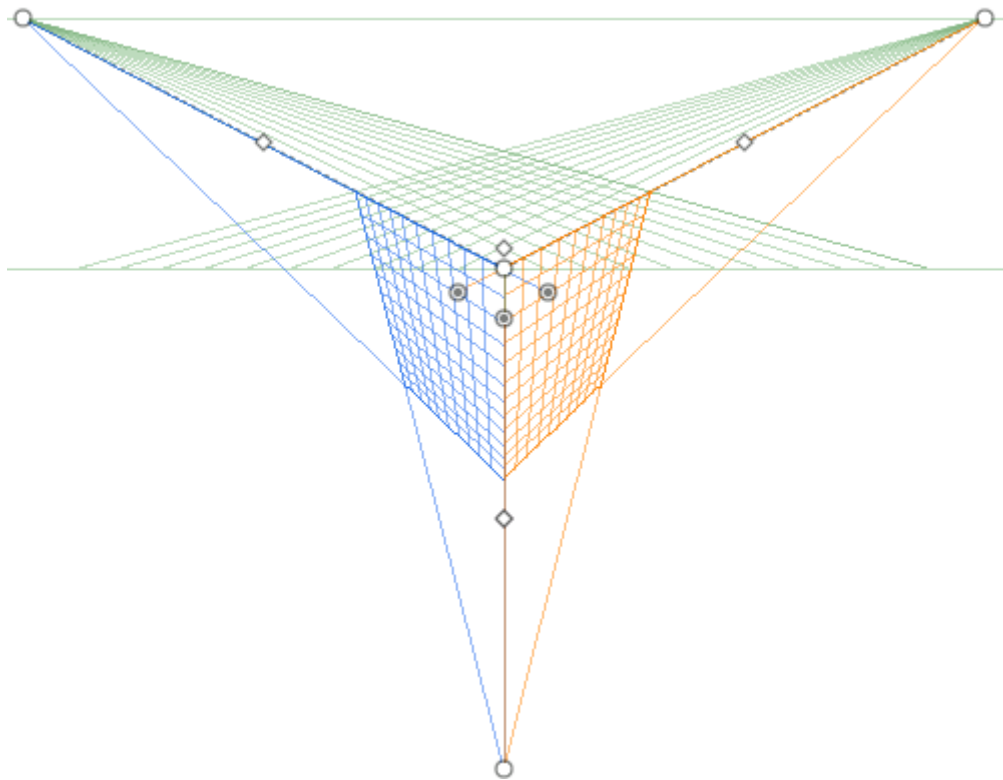


Figure: 23.25 - Perspectives à 1, 2 et 3 points de fuite générées par Illustrator

2. PERSPECTIVES AFFINES

Même si la meilleure méthode de la représentation en perspective est la méthode la perspective conique, nous ne pouvons nous permettre systématiquement de faire de grosses quantités de calcul pour représenter un volume. Ainsi, il est possible de définir des techniques de perspectives qui découlent d'une approximation des résultats mathématiques obtenus précédemment pour arriver à deux techniques (il en existe un plus grand nombre mais ces deux perspectives sont de loin les plus utilisées) que l'on retrouve quotidiennement sur de nombreux supports papiers techniques ou artistiques. Ces deux techniques sont respectivement la "perspective cavalière" et la "perspective isométrique" qui font partie de la famille des "perspectives affines" dites également "projections affines".

Définition: Une "**projection affine**" de l'espace usuel à 3 dimensions est une transformation affine qui envoie tous les points de cet espace sur un même plan de cet espace. Si le point $M(x, y, z)$ n'est pas sur le plan de projection, lui et son image $m(x', y', z')$ forment une droite dont la direction est constante: nous l'appelons "**direction de projection**". La perspective qui en découle est appelée familièrement "**perspective parallèle**" ou encore "**perspective cylindrique**". Comme toutes les transformations affines de l'espace, une projection affine conserve:

- Le parallélisme entre les droites
- Le barycentre, donc toutes les proportions existantes sur une droite donnée

Seuls les longueurs et les angles situés dans un plan parallèle au plan de projection sont conservés.

Sans trop nous attarder sur ces deux techniques, nous les présentons brièvement car elles doivent faire partie de la culture générale du physicien.

2.1. PERSPECTIVE CAVALIÈRE

Prenons une vue (par exemple la vue de face), et nommons les axes x (horizontale) et y (verticale). L'axe z étant l'axe perpendiculaire à la vue.

Dans la perspective cavalière, nous traçons l'axe z avec un certain angle par rapport à l'axe x (par exemple $\pi/6$ ou $\pi/4$), et l'on y reporte les distances en les multipliant par un coefficient inférieur à 1 en appliquant les règles trigonométriques de base tel que:

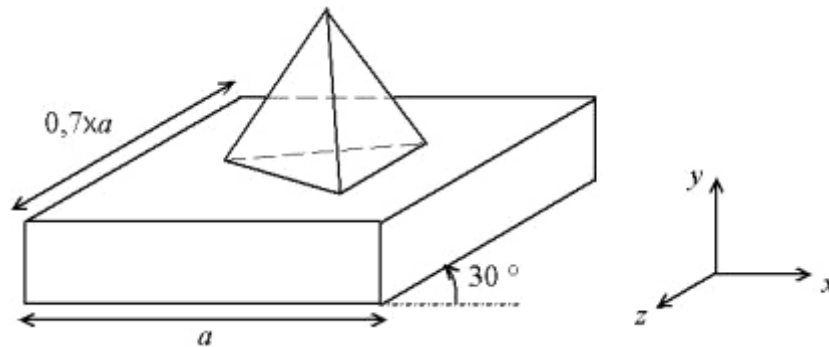


Figure: 23.26 - Exemple de perspective cavalière

C'est en mathématiques, par exemple, que la perspective cavalière est fréquemment choisie pour représenter sur le tableau des classes d'écoles les figures géométriques tridimensionnelles dans une base orthonormée canonique de direction E .

Si la direction de projection n'est pas orthogonale au plan de projection, alors la perspective cavalière transforme une sphère en ellipse. Cette distorsion de la sphère rend la perspective cavalière tout à fait impropre à une utilisation en dessin d'art. Cet inconvénient n'est par contre pas rédhibitoire dans une utilisation en dessin industriel.

2.2. PROJECTION ORTHOGONALE

Si la direction de projection est orthogonale au plan de projection, alors la perspective transforme une sphère en cercle. C'est donc un type de perspective utilisable en dessin comme alternative à la perspective conique (avec laquelle elle va d'ailleurs coïncider quand l'œil de l'observateur est placé infiniment loin du tableau).

La projection orthogonale la plus simple à exprimer est évidemment celle qui envoie l'espace sur un plan parallèle au tableau (il s'agit de la "**projection orthogonale parallèle**"), de cote constante égale à a . Autrement dit, tel que:

$$M(x, y, z) \rightarrow m(x, y, a) \quad (23.31)$$

On obtient alors trivialement les coordonnées dans un repère orthonormé 2D propre à ce plan où $x = x'$ et $y = y'$:

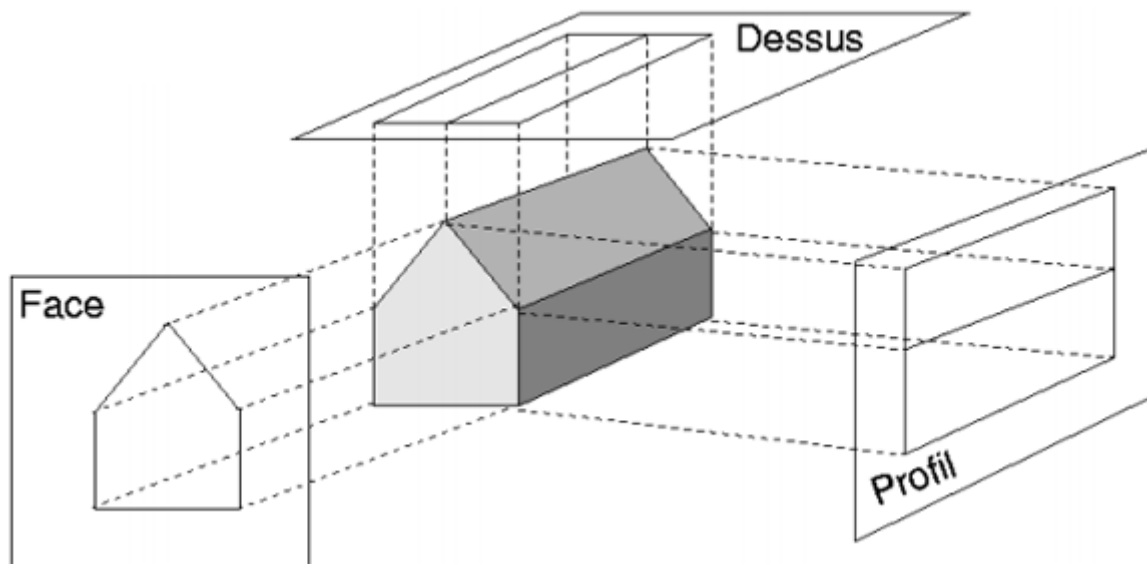


Figure: 23.27 - Projection orthogonale parallèle typique

Bien que ce soit la méthode la plus utilisée dans le monde industriel, elle a cependant un inconvénient: l'effet réel de la profondeur, est entièrement perdu.

Pour y remédier il existe deux solutions équivalentes (généralisables aux autres perspectives): l'une consiste à tourner l'objet par isométrie de l'espace et l'autre à changer le tableau.

Un cas particulier de la projection orthogonale est la "**perspective isométrique**". Elle est très utilisée en dessin industriel, elle projette orthogonalement les points de l'espace sur le plan isotrope, d'équation:

$$x + y + z = 1 \quad (23.32)$$

La direction de projection est donc la normale à ce plan, de vecteur directeur $(1,1,1)$.

Le perspective isométrique consiste à placer les axes x, y, z à $2\pi/3$ (soit 120°) les uns des autres et y reporter les distances telles quelles (d'où son nom).

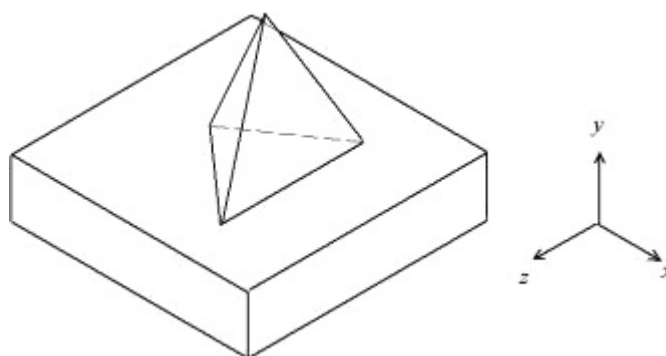


Figure: 23.28 - Exemple de perspective isométrique

À titre de comparaison, nous avons représenté selon les deux perspectives présentées précédemment un cube avec un cercle inscrit dans le carré de chaque face:

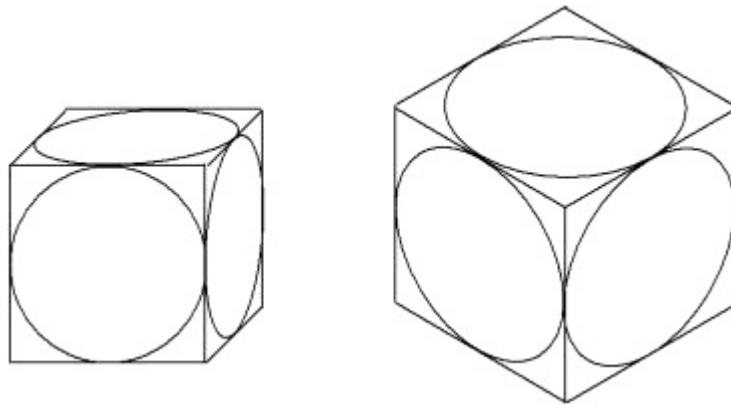


Figure: 23.29 - Perspective cavalière (gauche) et isométrique (droite)

3. COORDONNÉES HOMOGÈNES

En mathématiques, les coordonnées homogènes, introduites par August Ferdinand Möbius, rendent les calculs possibles dans l'espace projectif comme les coordonnées cartésiennes le font dans l'espace euclidien. Les coordonnées homogènes sont largement utilisées en infographie et plus particulièrement pour la représentation de scènes en trois dimensions (3D) car elles sont adaptées à la géométrie projective et elles permettent de caractériser les transformations de l'espace sous une forme algorithmique optimale. La notation sous forme matricielle est plus particulièrement employée dans les bibliothèques de programmation graphique 3D telles que OpenGL et Direct3D.

Nous avons vu lors de notre étude des transformations dans le plan et l'espace (cf. chapitre de Géométrie Euclidienne) que parmi la translation, l'homothétie, la rotation ou la réflexion que la translation ne pouvait être représentée sous forme matricielle sans passer par une astuce qui a consistait à ajouter une dimension supplémentaire factice au vecteur des coordonnées ainsi qu'à la matrice associée à la transformation.

Ainsi, nous avons vu que la translation dans $\mathbb{R}^2$ pouvait s'écrire sous forme matricielle:

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = T \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{bmatrix} 1 & 0 & T_x \\ 0 & 1 & T_y \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} x + T_x \\ y + T_y \\ 1 \end{pmatrix} \quad (23.33)$$

Ainsi, toujours dans $\mathbb{R}^2$, une homothétie de facteur k qui s'écrivait:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} k \cdot x \\ k \cdot y \end{pmatrix} \quad (23.34)$$

devient en coordonnées homogènes:

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = S \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{bmatrix} S_x & 0 & 0 \\ 0 & S_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} S_x \cdot x \\ S_y \cdot y \\ 1 \end{pmatrix} \quad (23.35)$$

Ainsi, toujours dans $\mathbb{R}^2$, une rotation d'angle ϑ de facteur k qui s'écrivait:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \cos \theta - y \sin \theta \\ x \sin \theta + y \cos \theta \end{pmatrix} \quad (23.36)$$

devient:

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = R \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} x \cos \theta - y \sin \theta \\ x \sin \theta + y \cos \theta \\ 1 \end{pmatrix} \quad (23.37)$$

Ainsi, toujours dans $\mathbb{R}^2$, la réflexion qui s'écrivait selon l'axe X :

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -x \\ y \end{pmatrix} \quad (23.38)$$

devient:

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \Delta \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} -x \\ y \\ 1 \end{pmatrix} \quad (23.39)$$

Les mathématiciens disent alors que nous nous plaçons dans l'espace projectif $\mathcal{P}^2$.

In extenso, nous pouvons faire de même avec un point P de $\mathbb{R}^3$ qui sera alors dans $\mathcal{P}^3$ représenté par un vecteur de 4×1 coordonnées:

$$\begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} \quad (23.40)$$

Ainsi, nous avons dans l'espace les matrices de transformations suivantes pour les translations:

$$T = \begin{bmatrix} 1 & 0 & 0 & T_x \\ 0 & 1 & 0 & T_y \\ 0 & 0 & 1 & T_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (23.41)$$

pour les rotations (voir le chapitre de Géométrie Euclidienne pour la démonstration) avec les angles d'Euler:

$$R = R_x R_y R_x = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta & 0 \\ 0 & \sin \theta & \cos \theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos \gamma & 0 & \sin \gamma & 0 \\ 0 & 1 & 0 & 0 \\ -\sin \gamma & 0 & \cos \gamma & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos \theta & -\sin \theta & 0 & 0 \\ \sin \theta & \cos \theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (23.42)$$

sans oublier que les matrices de rotations sont non commutatives au-delà de deux dimensions!

Et pour les homothéties:

$$S = \begin{bmatrix} S_x & 0 & 0 & 0 \\ 0 & S_y & 0 & 0 \\ 0 & 0 & S_z & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (23.43)$$

Nous avons démontré plus haut que dans le cas de la perspective conique, que nous avons les transformations homographiques suivantes:

$$\begin{aligned} x' &= \frac{x \cdot \Delta}{z + \Delta} \\ y' &= \frac{y \cdot \Delta}{z + \Delta} \\ z' &= \Delta \end{aligned} \quad (23.44)$$

où x', y', z' sont les projections du point réel x, y, z sur le tableau avec une distance focale Δ .

Ce qui est traditionnellement noté en posant $z := z + \Delta$ et $f := \Delta$:

$$\begin{aligned} x' &= \frac{x \cdot f}{z} \\ y' &= \frac{y \cdot f}{z} \\ z' &= f \end{aligned} \quad (23.45)$$

où nous voyons donc que si la distance focale f est infinie, l'objet se confond avec le plan XY ou que si l'objet est infiniment loin...

Posons $z/f = s$. Nous avons alors:

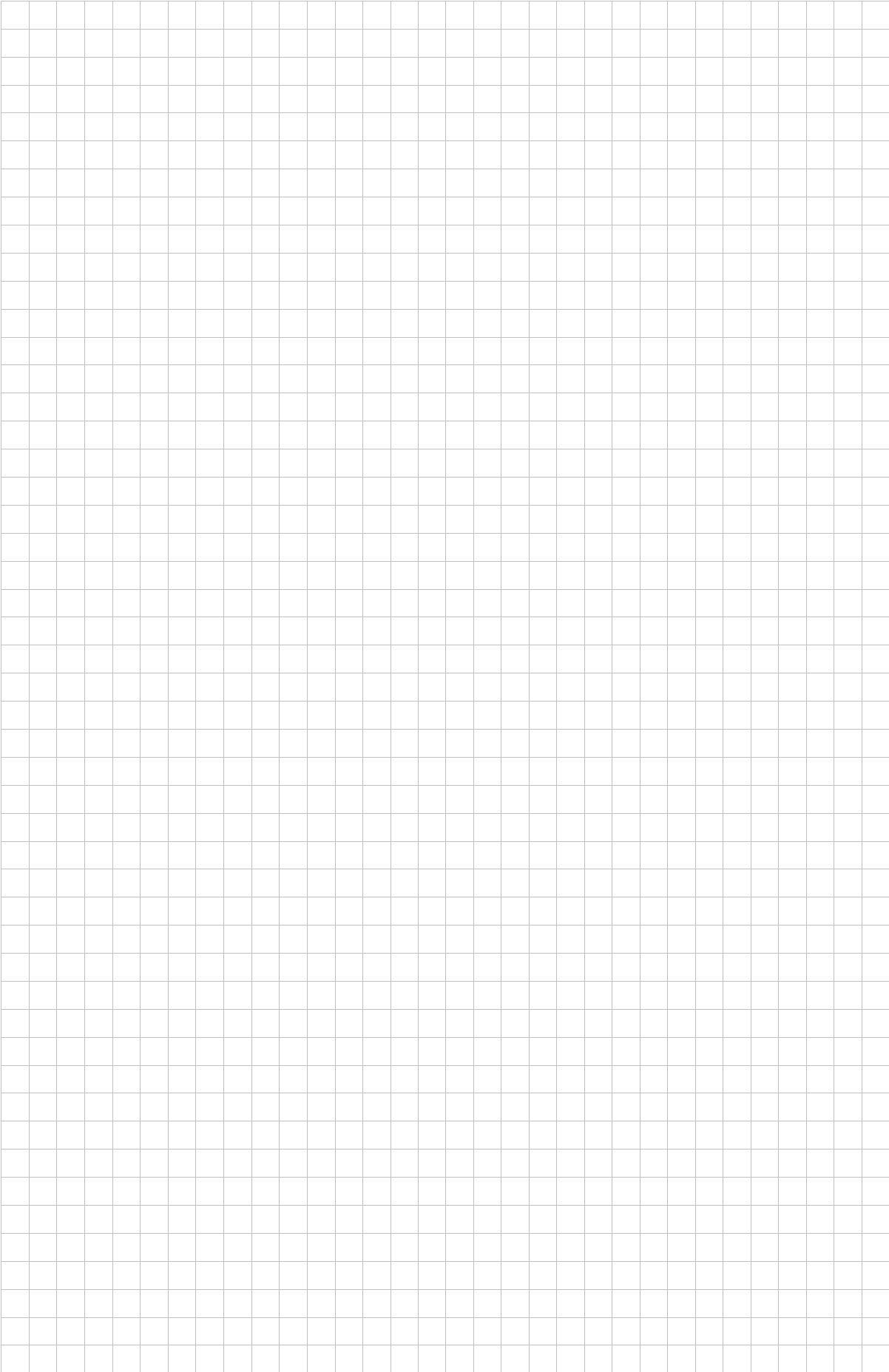
$$\begin{aligned} s x' &= x & s x' &= x \\ s y' &= y & s y' &= y \\ z' &= f & s z' &= z \\ s &= z/f & s &= z/f \end{aligned} \Rightarrow \quad (23.46)$$

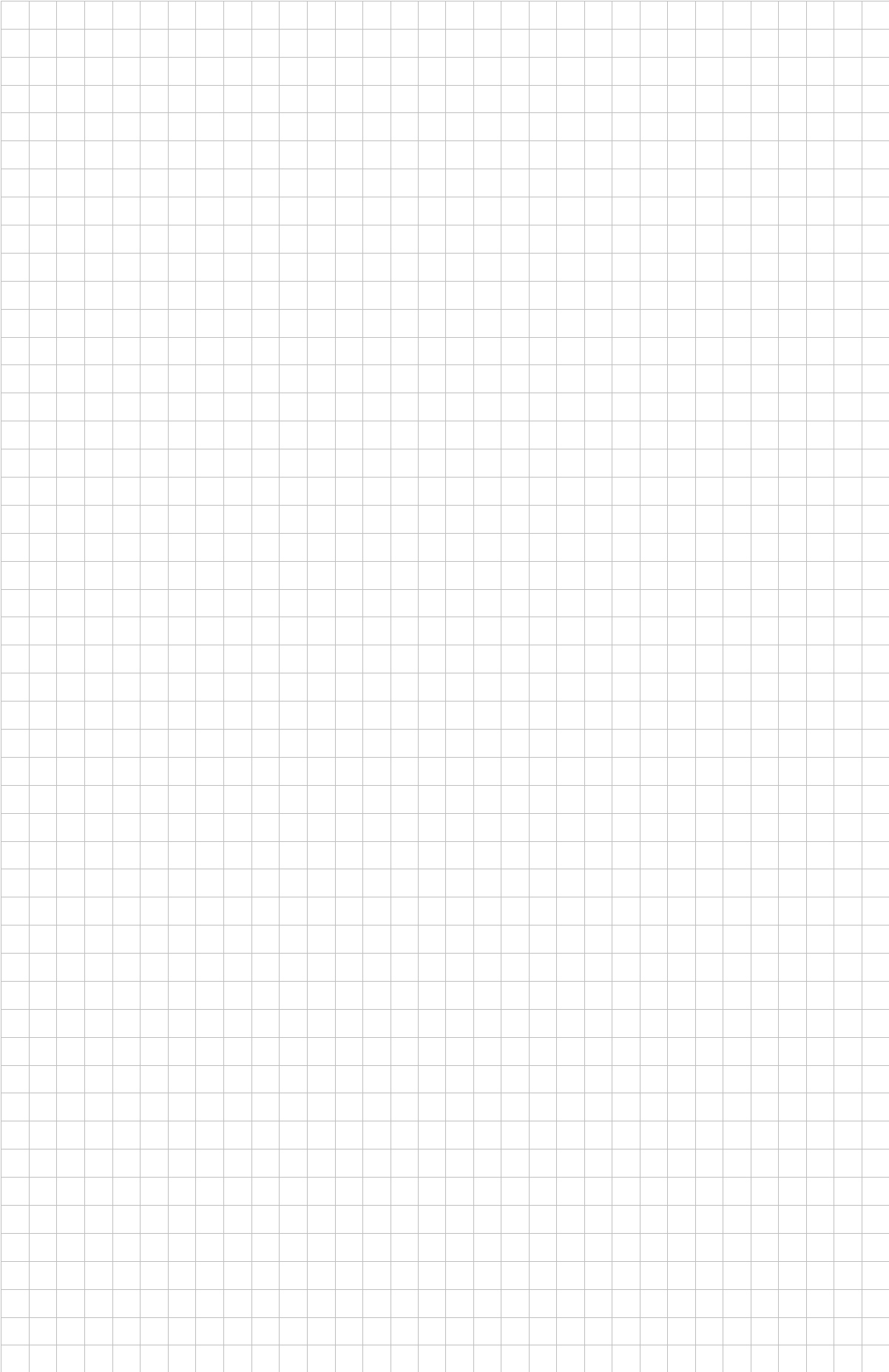
Nous utilisons alors la matrice suivante pour la projection conique:

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1/f & 0 \end{bmatrix} \quad (23.47)$$

et ensuite nous normalisons les coordonnées par le rapport z/f .







24. GÉOMÉTRIE ANALYTIQUE

La "géométrie analytique" est la branche de la géométrie qui s'occupe de l'étude des formes géométriques et de leurs propriétés en utilisant les outils avancés du calcul algébrique tels que l'analyse fonctionnelle, le calcul vectoriel ou l'algèbre linéaire. Sa frontière se situe au niveau des outils utilisés et a pour origine les travaux de René Descartes au début du 17ème siècle. La "géométrie vectorielle" est donc un sous-famille de la géométrie analytique et nous en ferons aussi bien usage dans le présent chapitre que dans celui de Géométrie Différentielle.

Remarque: Lorsque nous faisons usage pour ces mêmes études du calcul différentiel et intégral, alors nous faisons de la "géométrie différentielle" (voir chapitre du même nom).

La géométrie analytique est un très vaste domaine (comme tout le reste) alors... nous n'aborderons ici que les éléments absolument indispensables à l'étude de la physique (et en particulier l'astronomie et la physique quantique corpusculaire) et de l'ingénierie (génie spatial). Ces éléments sont par ailleurs souvent étudiés dans les petites classes et sont (cités dans l'ordre): les coniques, les équations de la droite, du plan, de la sphère, etc... leurs intersections, leurs plans tangents et encore bien d'autres.

1. CONIQUES

Il nous a été très difficile de choisir s'il fallait mettre l'étude des coniques dans la section d'algèbre ou de géométrie. Nous avons finalement décidé de mettre cette étude dans le présent chapitre (donc de géométrie...) qui permet de supposer que le lecteur ayant fait une lecture linéaire du site a déjà parcouru tous les chapitres présentant les outils mathématiques nécessaires à l'étude des coniques. Nous espérons que notre choix s'avérera le meilleur pour le lecteur.

Remarque: L'étude des coniques nous sera très utile dans le chapitre d'Astronomie (par ailleurs c'est à Kepler que l'on doit de nombreux résultats de l'étude des coniques) ainsi que dans le chapitre d'Optique Géométrique. Il convient donc de s'y attarder dans les détails.

1.1. APPROCHE ALGÈBRIQUE

Soit $(O, \vec{i}, \vec{j})$ un repère orthonormé du plan. Les courbes algébriques les plus simples que l'on trouve après les droites dont les équations sont sous forme générale (rappel):

$$ax + by + c = 0 \quad (24.1)$$

sont les courbes du deuxième degré, à savoir par extension:

$$\underbrace{\alpha x^2 + \beta y^2 + \gamma xy}_{\text{terme quadratique}} + \delta x + \varepsilon y + \xi = 0 \quad (24.2)$$

avec α, β, γ non tous nuls. Ces courbes de second degré sont appelées "**coniques**" (appelées également "**quadriques**" de par la présence d'un terme quadratique). Cette dernière relation peut aussi s'écrire sous forme matricielle (écriture très importante dans la pratique des méthodes numériques comme nous le verrons dans le chapitre du même nom):

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} \alpha & \beta/2 \\ \beta/2 & \gamma \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \delta & \varepsilon \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \xi = \vec{x}^T A \vec{x} + B^T \vec{x} + C \quad (24.3)$$

écriture dont il existe plusieurs variantes...

Par exemple, la fonction:

$$f(x, y) = x^2 + y^2 + 2bxy - 5x - 4y + g \quad (24.4)$$

donne pour différentes valeurs de g les tracés suivants:

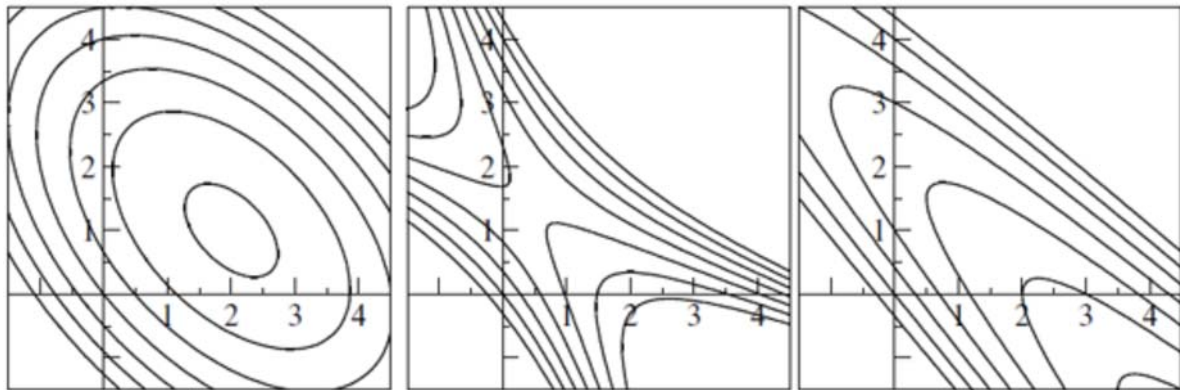


Figure: 24.1 - Coniques avec $b = 0.5, 1.5$ et 1 et plusieurs valeurs de g

Notre première tâche va consister à obtenir, par translation et rotation du repère dans lequel cette relation est exprimée, une équation réduite beaucoup plus simple en éliminant le terme en xy . En effet, choisissons un nouveau repère se déduisant de l'ancien par une rotation d'angle θ . Soit x' et y' les nouvelles coordonnées des points. Nous avons (cf. chapitre de Géométrie Euclidienne):

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix} \quad (24.5)$$

D'où:

$$\begin{aligned} x &= \cos(\theta)x' - \sin(\theta)y' \\ y &= \sin(\theta)x' + \cos(\theta)y' \end{aligned} \quad (24.6)$$

L'équation devient:

$$\begin{aligned} &\alpha(\cos(\theta)x' - \sin(\theta)y')^2 + \beta(\sin(\theta)x' + \cos(\theta)y')^2 \\ &+ \gamma(\cos(\theta)x' - \sin(\theta)y')(\sin(\theta)x' + \cos(\theta)y') + \dots = 0 \end{aligned} \quad (24.7)$$

Nous cherchons donc à ce que les termes en $x'y'$ regroupés soient tels que:

$$-2\alpha \cos \theta \sin \theta + 2\beta \cos \theta \sin \theta + \gamma (\cos^2 \theta - \sin^2 \theta) = 0 \quad (24.8)$$

Puisque (cf. chapitre de Trigonométrie):

$$\cos^2 \theta - \sin^2 \theta = \cos(2\theta) \text{ et } \sin(\alpha + \beta) = \sin(\beta) \cdot \cos(\alpha) + \cos(\beta) \cdot \sin(\alpha) \quad (24.9)$$

par substitution, nous obtenons:

$$(\beta - \alpha) \sin(2\theta) + \gamma \cos(2\theta) = 0 \quad (24.10)$$

Pour avoir que les termes en $x'y'$ se simplifient, il suffit donc de choisir l'angle de rotation θ tel que:

$$\frac{\sin(2\theta)}{\cos(2\theta)} = \frac{\gamma}{(\alpha - \beta)} = \tan(2\theta) \quad (24.11)$$

Nous considérerons alors désormais l'équation:

$$\boxed{\alpha x^2 + \beta y^2 + \delta x + \varepsilon y + \xi = 0} \quad (24.12)$$

1. Si nous posons $\alpha = 0$ et $\beta \neq 0$. Quitte à diviser par β , nous pouvons nous ramener à une équation du type:

$$y^2 + \gamma x + \delta y + \varepsilon = 0 \quad (24.13)$$

où:

- Si $\gamma \neq 0$, nous nous retrouvons avec une équation décrivant la figure d'une "**parabole**" d'axe parallèle à OX .

- Si $\gamma = 0$, il s'agit d'un cas dégénéré

2. Si nous posons $\alpha \neq 0$ et $\beta = 0$ le cas se traite comme précédemment.

3. Si $\alpha \neq 0$ et $\beta \neq 0$, nous pouvons supprimer les termes δx et εy de la façon suivante:

$$\alpha \left(x + \frac{\delta}{2\alpha} \right)^2 + \beta \left(y + \frac{\varepsilon}{2\beta} \right)^2 + \gamma = 0 \quad (24.14)$$

et par un simple changement de repère via des translations, nous arrivons donc à une équation du type:

$$\alpha X^2 + \beta Y^2 + \gamma = 0 \quad (24.15)$$

- Si $\gamma = 0$, alors la relation précédente se réduit à un point dans $\mathbb{R}^2$ si α et β sont de même signe, et à une droite si α et β sont de signe contraire.

- Si $\gamma \neq 0$ posons:

$$\alpha = \operatorname{sgn}(\alpha) \frac{1}{\alpha'^2}, \beta = \operatorname{sgn}(\beta) \frac{1}{\beta'^2} \quad (24.16)$$

où $\operatorname{sgn}(x)$ signifie: 1 multiplié par le signe de x .

Et divisons le tout par $|\gamma|$ tel que:

$$\operatorname{sgn}(\alpha) \frac{X^2}{|\gamma|\alpha'^2} + \operatorname{sgn}(\beta) \frac{Y^2}{|\gamma|\beta'^2} + \operatorname{sgn}(\gamma) = 0 \quad (24.17)$$

Posons:

$$\frac{1}{|\gamma|\alpha'^2} = \frac{1}{a^2}, \frac{1}{|\gamma|\beta'^2} = \frac{1}{b^2} \quad (24.18)$$

Nous obtenons:

$$\operatorname{sgn}(\alpha) \frac{X^2}{a^2} + \operatorname{sgn}(\beta) \frac{Y^2}{b^2} = -\operatorname{sgn}(\gamma) \quad (24.19)$$

Nous avons donc plusieurs situations possibles:

$$\begin{array}{l} +\frac{X^2}{a^2} + \frac{Y^2}{b^2} = 1, \quad -\frac{X^2}{a^2} - \frac{Y^2}{b^2} = -1 \\ +\frac{X^2}{a^2} - \frac{Y^2}{b^2} = 1, \quad -\frac{X^2}{a^2} + \frac{Y^2}{b^2} = -1 \\ -\frac{X^2}{a^2} + \frac{Y^2}{b^2} = +1, \quad \frac{X^2}{a^2} - \frac{Y^2}{b^2} = -1 \\ \cancel{-\frac{X^2}{a^2} - \frac{Y^2}{b^2} = +1}, \quad \cancel{+\frac{X^2}{a^2} + \frac{Y^2}{b^2} = -1} \end{array} \quad (24.20)$$

Deux termes ci-dessus sont impossibles dans $\mathbb{R}^2$, c'est pourquoi nous les avons barrés (la somme de deux nombres positifs ne peut être négative et inversement).

Il y a plusieurs cas de figures intéressants:

- Pour:

$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} = 1 \text{ ou } -\frac{X^2}{a^2} - \frac{Y^2}{b^2} = -1 \quad (24.21)$$

et $a = b = 1$ nous avons un cercle de rayon unité. Le lecteur peut s'amuser à tester cela avec la commande Maple 4.00b suivante:

```
>with(plots):
>a:=4;b:=4;
>implicitplot(x^2/a^2+y^2/b^2 = 1,x=-10..10,y=-10..10);
```

- Pour:

$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} = 1, \quad -\frac{X^2}{a^2} - \frac{Y^2}{b^2} = -1 \quad (24.22)$$

et $\forall a, b \in \mathbb{R}^*$ nous avons une ellipse (qui contient donc le cas particulier qu'est le cercle) de demi-axes respectivement a et b (pour la représentation graphique voir plus bas ou sinon le chapitre de Formes Géométriques).

Le lecteur peut s'amuser à tester cela avec la commande Maple 4.00b suivante:

```
>with(plots):
>a:=4;b:=8;
>implicitplot(x^2/a^2+y^2/b^2 = 1,x=-10..10,y=-10..10);
```

- Pour:

$$\begin{aligned} +\frac{X^2}{a^2} - \frac{Y^2}{b^2} = 1, \quad -\frac{X^2}{a^2} + \frac{Y^2}{b^2} = -1 \\ -\frac{X^2}{a^2} + \frac{Y^2}{b^2} = +1, \quad \frac{X^2}{a^2} - \frac{Y^2}{b^2} = -1 \end{aligned} \quad (24.23)$$

et $\forall a, b \in \mathbb{R}^*$, nous avons des hyperboles dont l'axe de symétrie est soit parallèle à OX soit à OY (pour la représentation graphique voir plus bas). Nous disons que l'hyperbole est "équilatère" lorsque $a = b$.

Le lecteur peut s'amuser à tester cela avec la commande Maple 4.00b suivante:

```
>with(plots):
>a:=4;b:=4;
>implicitplot(x^2/a^2-y^2/b^2 = 1,x=-10..10,y=-10..10);
```

Remarque: Pour voir les figures, utilisez la fonction `implicitplot( )` dans Maple.

Le terme "**conique**" provient du fait que l'une des premières définitions des conques consistait en l'intersection d'un cône et d'un plan.

En effet, soit $z^2 = x^2 + y^2$ l'équation d'un cône ayant un angle de $\pi/4$ au sommet (voir géométrie spatiale)

$\cos \theta \cdot y + \sin \theta \cdot z = h$ l'équation d'un plan de vecteur normal (nous utilisons les cosinus directeurs):

$$(0, \cos \theta_y, \cos \theta_z) = \vec{n}, \theta_i \in [0, \pi/2] \quad (24.24)$$

Posons:

$$\theta_y = \theta_z = \theta \quad (24.25)$$

Explications: nous avons ainsi un vecteur normal dans le plan ZOY et un plan qui n'est jamais en intersection avec l'axe OX . Si le cosinus directeur $\cos \theta_x = 0$, nous avons un plan vertical translaté de h sur l'axe des Y . Si $\cos \theta_y = 0$, nous avons un plan horizontal translaté de h sur l'axe des Z .

Soit la matrice de rotation dans l'espace par rapport à l'axe OZ (cf. chapitre de Géométrie Euclidienne):

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta \\ 0 & \sin \theta & \cos \theta \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \quad (24.26)$$

avec:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} x \\ y \cos \theta - z \sin \theta \\ y \sin \theta + z \cos \theta \end{pmatrix} \quad (24.27)$$

Nous avons donc pour expression de rotation du plan:

$$\cos \theta (y \cos \theta - z \sin \theta) + \sin \theta (y \sin \theta + z \cos \theta) = h \quad (24.28)$$

Après simplification:

$$y = h \quad (24.29)$$

Donc après rotation, nous avons un plan vertical translaté de h selon l'axe des Y .

Identiquement, pour le cône, une rotation correspond selon l'axe des Z (donc il ne se passe pas grand-chose):

$$(y \sin \theta + z \cos \theta)^2 = x^2 + (y \cos \theta - z \sin \theta)^2 \quad (24.30)$$

Après développement et simplification:

$$x^2 - (y^2 - z^2) \cos(2\theta) + yz 2 \sin(2\theta) = 0 \quad (24.31)$$

Equation qui donne un cône horizontal si $\theta = 0$ et un cône vertical si $\theta = \pi/2$.

Ainsi, nous avons le système général:

$$\begin{cases} y = h \\ x^2 - (y^2 - z^2) \cos(2\theta) + yz 2 \sin(2\theta) = 0 \end{cases} \quad (24.32)$$

Nous voyons alors que pour:

- $\theta \in \left[\frac{\pi}{4}, \frac{\pi}{2} \right]$ nous obtenons une intersection entre le plan et le cône donnant une ellipse

- $\theta = \frac{\pi}{4}$ nous obtenons une parabole

- $\theta \in \left[0, \frac{\pi}{4}\right]$ nous obtenons une hyperbole

Voici à peu près ce que cela donne visuellement parlant...:

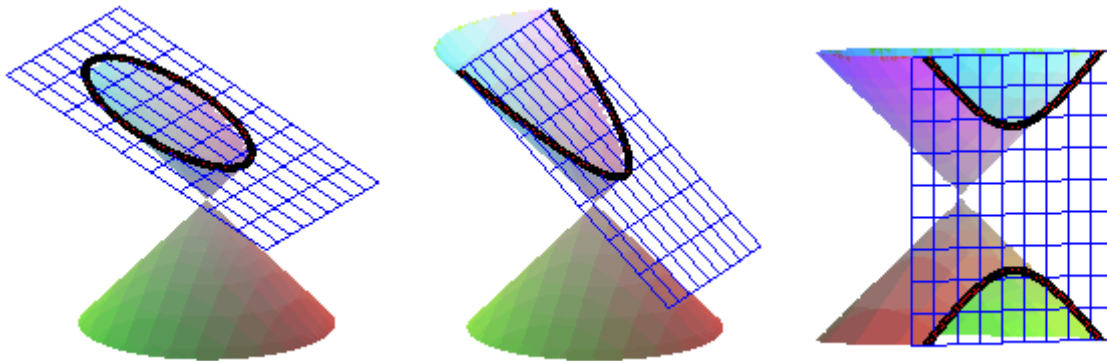


Figure: 24.2 - Différentes intersections entre un cône et un plan

Nous donnons également à la courbe d'équation $xy = 1$ le nom d'hyperbole car, par changement de variable:

$$\begin{aligned} y &= X + Y \\ x &= X - Y \end{aligned} \quad (24.33)$$

Ce qui nous ramène à:

$$X^2 - Y^2 = 1 \quad (24.34)$$

ce qui comme nous l'avons vu précédemment, est bien l'équation d'une hyperbole.

Pour ceux qui ont Maple 4.00b ou ultérieur voici les commandes pour s'amuser à faire une parabole:

```
>restart: with(plots):
>c:=sqrt(x^2+y^2):
>p:=y+3:
> Y:=solve(c=p,y);
> intsect:=subs(y=Y,c);
>intsect := (x + (1/6 x - 3/2) )
> P1:=plot3d(c,x=-5..5,y=-5..5,axes=normal,color=red,numpoints=2000,view=
[-5..5,-5..5,0..5],style=wireframe):
> P2:=plot3d(p,x=-5..5,y=-5..5,axes=normal,color=yellow,numpoints=2000,view=
[-5..5,-5..5,0..5],style=patchnogrid):
> P3:=spacecurve([x,Y,intsect],x=-5..5,color=black,thickness=3):
> display(P1,P2,P3,scaling=constrained, orientation=[-10,75]);
```

ce qui donne:

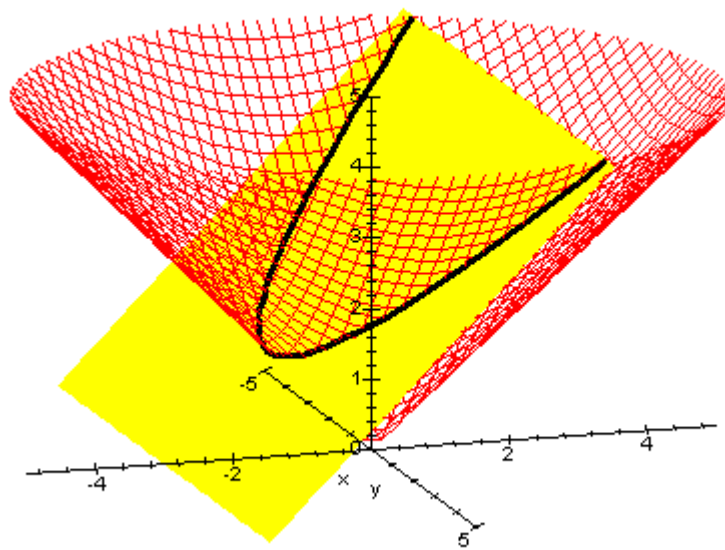


Figure: 24.3 - Plot Maple 4.00b de l'obtention d'une parabole par l'intersection d'un cône et d'un plan

ou pour l'obtention d'une hyperbole:

```
>restart: with(plots):
>c:=sqrt(x^2+y^2):
>p:=4*y+5:
>Y:=solve(c=p,y);
>intsect:=subs(y=Y[1],c);
> P1:=plot3d(c,x=-5..5,y=-5..5,axes=normal,color=red,numpoints=2000,view=
[-5..5,-5..5,0..5],style=wireframe):
>P2:=plot3d(p,x=-5..5,y=-5..5,axes=normal,color=yellow,numpoints=2000,view=
[-5..5,-5..5,0..5],style=patchnogrid):
>P3:=spacecurve([x,Y[1],intsect],x=-5..5,color=black,thickness=3):P3:=spacecurve([x,Y[1],intsect],x=-5..5,color=
>display(P1,P2,P3,scaling=constrained);
```

ce qui donne:

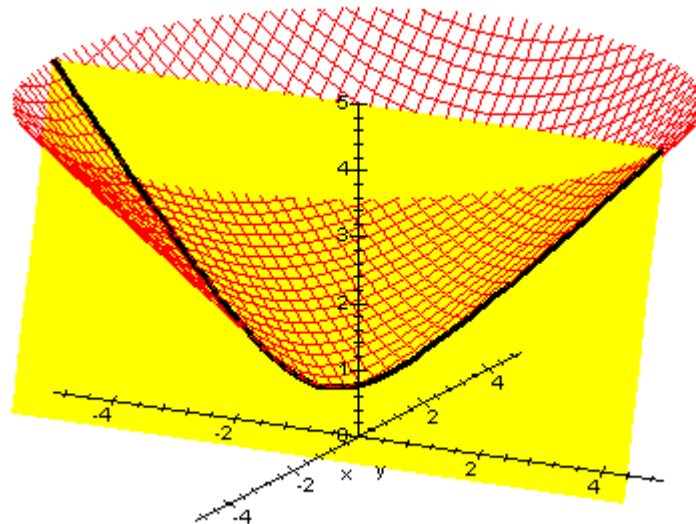


Figure: 24.4 - Plot Maple 4.00b de l'obtention d'une hyperbole par l'intersection d'un cône et d'un plan

ou pour l'obtention d'une ellipse:

```
>restart: with(plots):
>c:=sqrt(x^2+y^2):
>p:=y/3+3:
>Y:=solve(c=p,y);
>E1:=subs(y=Y[1],c); E2:=subs(y=Y[2],c);
>P1:=plot3d(c,x=-5..5,y=-5..5,axes=normal,color=red,numpoints=2000,view=
[-5..5,-5..5,0..5],style=wireframe):
>P2:=plot3d(p,x=-5..5,y=-5..5,axes=normal,color=yellow,numpoints=2000,view=
[-5..5,-5..5,0..5],style=patchnogrid):
>P3:=spacecurve({[x,Y[1],E1],[x,Y[2],E2]},x=-5..5,color=black,thickness=3,numpoints=2000):
>display(P1,P2,P3,scaling=constrained);
```

ce qui donne:

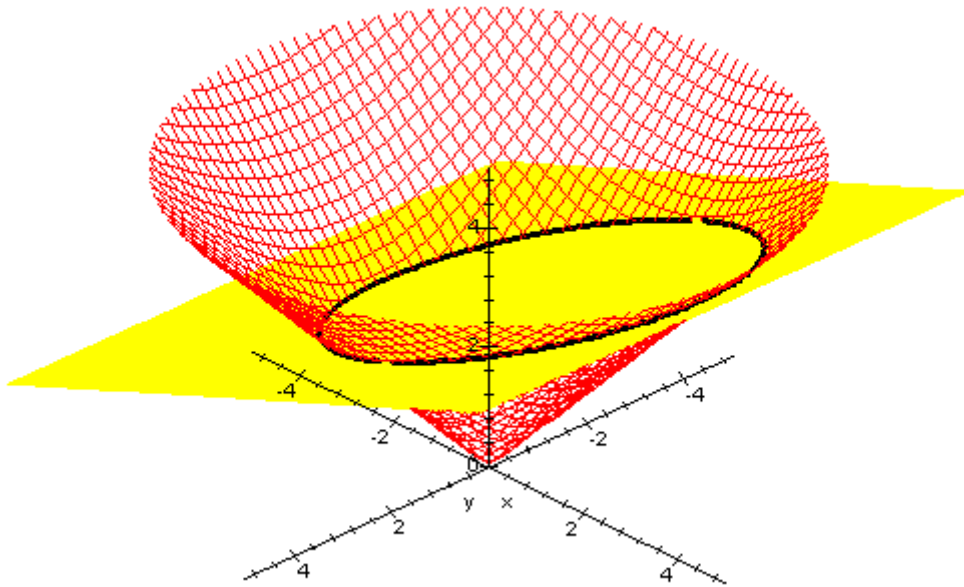


Figure: 24.5 - Plot Maple 4.00b de l'obtention d'une ellipse par l'intersection d'un cône et d'un plan

ou enfin pour l'obtention d'une cercle:

```
>restart: with(plots):
>c:=sqrt(x^2+y^2):
>p:=3:
>Y:=solve(c=p,y);
>circ1:=subs(y=Y[1],c); circ2:=subs(y=Y[2],c);
>P1:=plot3d(c,x=-5..5,y=-5..5,axes=normal,color=red,numpoints=2000,view=
[-5..5,-5..5,0..5],style=wireframe):
>P2:=plot3d(p,x=-5..5,y=-5..5,axes=normal,color=yellow,numpoints=2000,view=
[-5..5,-5..5,0..5],style=patchnogrid):
>P3:=spacecurve([x,Y[1],circ1],[x,Y[2],circ2],x=-5..5,color=black,thickness=3,numpoints=2000):
>display(P1,P2,P3);
```

ce qui donne:

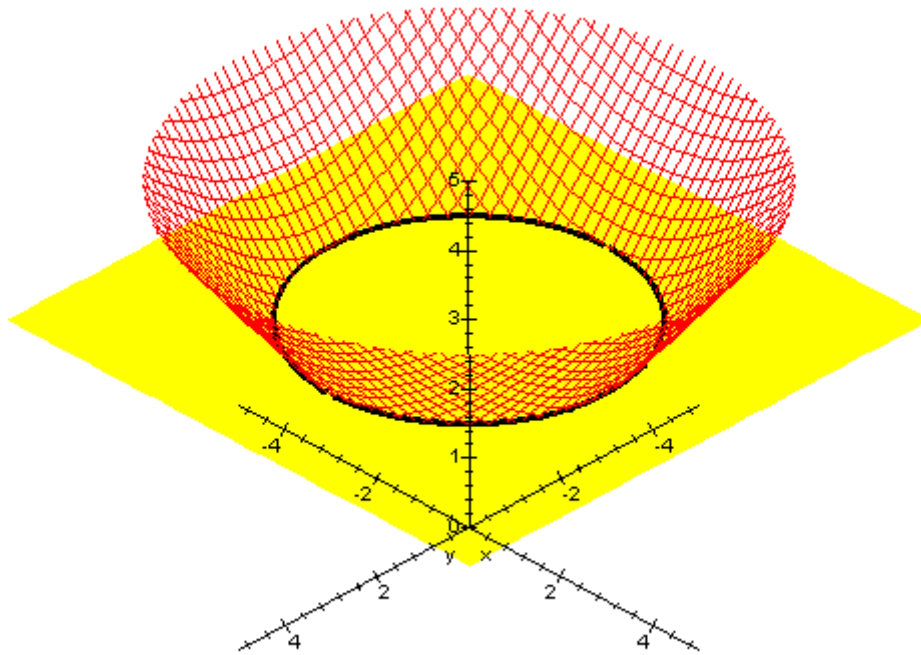


Figure: 24.6 - Plot Maple 4.00b de l'obtention d'un cercle par l'intersection d'un cône et d'un plan

Cependant, les coniques ont aussi une définition géométrique:

1.2. APPROCHE GÉOMÉTRIQUE

Soit F un point du plan, D une droite ne contenant pas F (mais à une distance non nulle quelconque de celui-ci) et e un réel strictement positif. Nous nous intéressons à l'ensemble des points M défini par:

$$\{M \mid e \cdot d(M, D) = d(M, F)\} \quad (24.35)$$

F s'appelant le "foyer", D la "directrice de la conique" et e "l'excentricité":

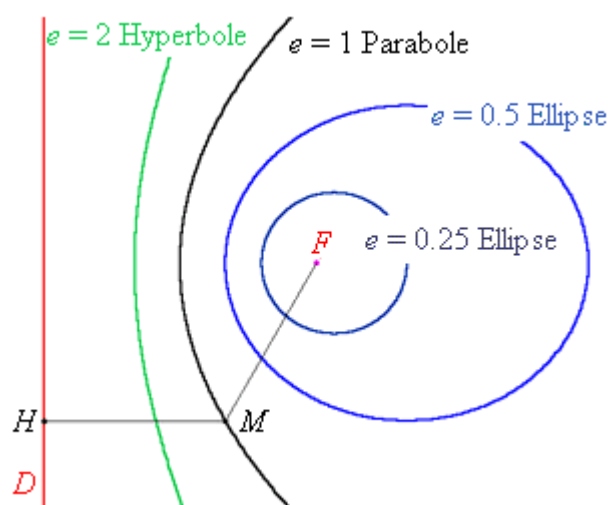


Figure: 24.7 - Définitions du foyer, de la directrice conique et de l'excentricité

Nous nous arrangerons par la suite pour avoir toujours F comme origine du repère des coniques, de façon D aura pour par définition pour équation:

$$x = -h \quad (24.36)$$

avec $h > 0$ et nous notons donc $d(M, D)$ la distance entre le point M et D .

Alors:

$$\begin{aligned} e \cdot d(M, D) &= d(M, F) \\ \Leftrightarrow e \cdot |x + h| &= \sqrt{x^2 + y^2} \\ \Leftrightarrow e^2 (x + h)^2 &= x^2 + y^2 \\ \Leftrightarrow (1 - e^2)x^2 + y^2 - 2e^2 hx - e^2 h^2 &= 0 \end{aligned} \quad (24.37)$$

Nous nous retrouvons bien avec l'équation d'une conique puisque cette dernière relation est un cas particulier de:

$$\alpha x^2 + \beta y^2 + \delta x + \varepsilon y + \xi = 0 \quad (24.38)$$

Nous pouvons considérer maintenant plusieurs cas particuliers:

1. Cas où $e = 1$:

L'équation:

$$(1 - e^2)x^2 + y^2 - 2e^2 hx - e^2 h^2 = 0 \quad (24.39)$$

se réduit alors trivialement à:

$$y^2 - h^2 = 2hx \quad (24.40)$$

Il s'agit d'une parabole d'axe orthogonal à D , dont le sommet Ω est le milieu du segment $\overline{FK}$, où K est la projection de F sur D (voir la figure plus bas)

Si nous récrivons la dernière relation sous la forme:

$$y^2 = 2hx + h^2 = 2h \left(x + \frac{h}{2} \right) \quad (24.41)$$

et redéfinissons l'origine relativement à Ω par une translation de $h/2$, le foyer générateur de la parabole sera donc en $F = (h/2, 0)$ et la dernière équation se réduit alors à:

$$y^2 = 2hx \quad (24.42)$$

où h est appelé "**paramètre de la parabole**" et relativement à Ω , le foyer sera donné par les coordonnées $F = (h/2, 0)$ et la directrice par l'équation $D: x = -h/2$. Comme le montre la figure ci-dessous, la distance de la directrice à Ω est donc imposée par les conditions du modèle:

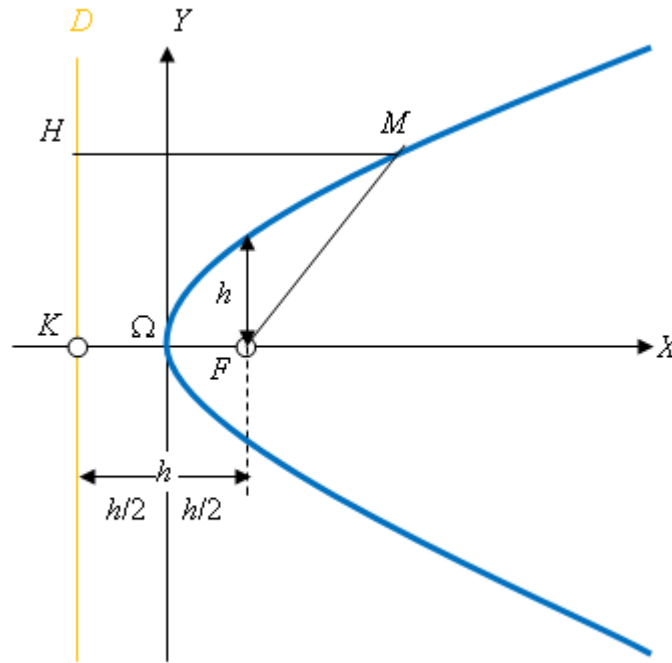


Figure: 24.8 - Représentation des paramètres des coniques pour la parabole

2. Cas où $e < 1$:

Il s'agit d'une ellipse. Effectivement ce n'est pas évident à voir mais en réarrangeant les termes de l'équation:

$$(1 - e^2)x^2 + y^2 - 2e^2hx - e^2h^2 = 0 \quad (24.43)$$

nous avons (ceci indépendamment du fait que $e < 1$ ou non):

$$(1 - e^2) \left(x - \frac{e^2h}{1 - e^2} \right)^2 + y^2 - \frac{e^2h^2}{1 - e^2} = 0 \Leftrightarrow \frac{\left(x - \frac{e^2h}{1 - e^2} \right)^2}{\frac{e^2h^2}{(1 - e^2)^2}} + \frac{y^2}{\frac{e^2h^2}{1 - e^2}} = 1 \quad (24.44)$$

Le dernier terme se retrouvant en effet comme suit après développement:

$$\frac{e^4h^2}{1 - e^2} - \frac{e^2h^2}{1 - e^2} = e^2h^2 \left(\frac{e^2 - 1}{1 - e^2} \right) = -e^2h^2 \quad (24.45)$$

Posons que:

$$F = \frac{e^2h}{1 - e^2} \quad (24.46)$$

est l'origine de l'ellipse en X (donc les x et y doivent maintenant être pris par rapport à cette nouvelle origine!!!). L'équation précédente se simplifie et devient (ceci toujours indépendamment du fait que $e < 1$ ou non):

$$\frac{\frac{x^2}{e^2 h^2}}{(1-e^2)^2} + \frac{\frac{y^2}{e^2 h^2}}{1-e^2} = \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \quad (24.47)$$

Donc si $e < 1$, les dénominateurs sont tous deux positifs et nous retrouvons bien l'équation réduite d'une ellipse puisque $a \neq b$.

Remarque: Il est à noter que cette définition ne peut inclure le cercle parmi les ellipses sinon quoi il y a une singularité car les dénominateurs seraient nuls (l'excentricité étant nulle pour le cercle).

Pour connaître le demi-grand axe de l'ellipse il suffit de poser $y = 0$. Ainsi, il nous reste:

$$\frac{\frac{x^2}{e^2 h^2}}{(1-e^2)^2} = \frac{x^2}{a^2} = 1 \quad (24.48)$$

d'où le demi-grand axe valant:

$$a = \pm \frac{eh}{1-e^2} = \pm \frac{F}{e} \quad (24.49)$$

de la même manière, nous obtenons le demi-petit axe:

$$b = \pm \frac{eh}{\sqrt{1-e^2}} \quad (24.50)$$

en posant $p = eh$ étant le "**paramètre de l'ellipse**" ou "**paramètre focal de l'ellipse**", nous obtenons:

$$|a| = \frac{p}{1-e^2}, |b| = \frac{p}{\sqrt{1-e^2}} \quad (24.51)$$

dont la première relation sera très utile dans le chapitre d'Astronomie et de Relativité Générale.

Il est d'usage de noter la distance du centre Ω de l'ellipse au point F par la lettre c tel que:

$$c = \Omega F \quad (24.52)$$

Nous avons donc:

$$c = \pm a \cdot e \Leftrightarrow e = \pm \frac{c}{a} \quad (24.53)$$

Il existe donc deux foyers à l'ellipse à une distance équivalente mais opposée du centre Ω . Nous définissons dès lors l'excentricité d'une ellipse par le rapport:

$$e = \frac{c}{a} \quad (24.54)$$

Le lecteur aura aussi remarqué que nous avons pour $x = 0$ dans:

$$(1 - e^2)x^2 + y^2 - 2e^2hx - e^2h^2 = 0 \quad (24.55)$$

l'ordonnée qui est donc donnée par le produit eh et que celui-ci peut être exprimé par les paramètres de l'ellipse sous la forme (même si cela peut paraître bizarre au niveau des unités c'est implicitement correct):

$$p = eh = \frac{b^2}{a} \quad (24.56)$$

Nous pouvons alors aussi démontrer une relation que nous retrouvons couramment dans les formulaires:

$$a^2 = b^2 + c^2 \quad (24.57)$$

C'est-à-dire:

$$\frac{e^2h^2}{(1-e^2)^2} = \frac{e^2h^2}{1-e^2} + e^2a^2 = \frac{e^2h^2}{1-e^2} + e^2 \frac{e^2h^2}{(1-e^2)^2} = \frac{e^2h^2(1-e^2) + e^4h^2}{(1-e^2)^2} = \frac{e^2h^2}{(1-e^2)^2} \quad (24.58)$$

L'égalité est donc démontrée et cela nous amène à pouvoir écrire l'excentricité uniquement à partir des paramètres classiques de l'ellipse:

$$e = \frac{c}{a} = \frac{\sqrt{a^2 - b^2}}{a} \quad (24.59)$$

Nous pouvons résumer les résultats obtenus par la figure ci-dessous:

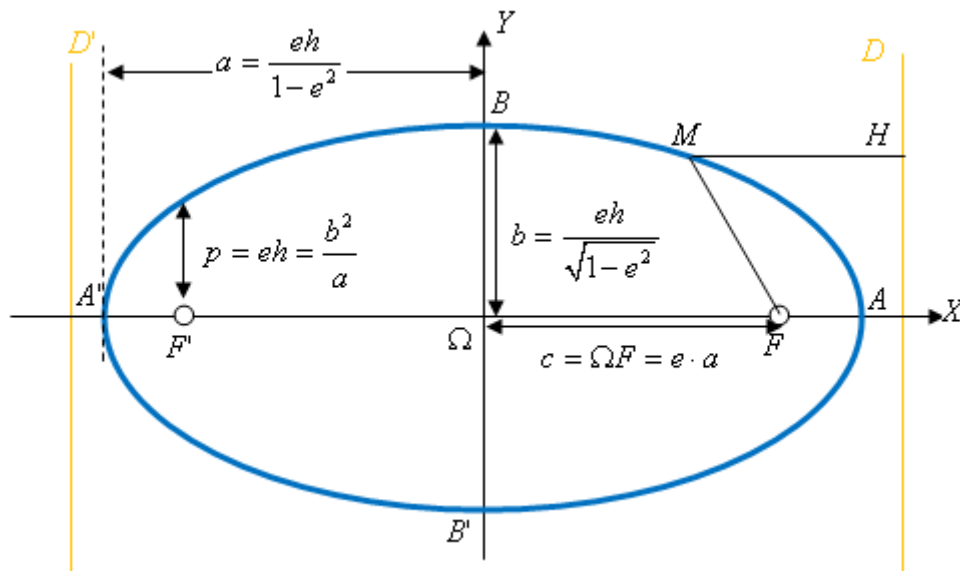


Figure: 24.9 - Représentation des paramètres des coniques pour une ellipse

Nous pouvons aussi déterminer où se trouvent les directrices D de l'ellipse par rapport à la bordure de celle-ci en utilisant la définition de l'excentricité lorsque $y = 0$. Nous avons alors:

$$\begin{aligned}
 e \cdot d(M, D) &= d(M, F) \\
 \Leftrightarrow d(M, D) &= \frac{d(M, F)}{e} = \frac{a - ea}{e} = a \frac{1 - e}{e} \quad (24.60)
 \end{aligned}$$

donc le facteur de a est compris entre 0 et l'infini ce qui signifie que les directrices se trouveront (du moins dans le cas ici présent où l'excentricité est positive et strictement plus petite que l'unité) toujours à l'extérieur de la bordure de l'ellipse et au plus proche tangentes à la bordure (mais dans tous les cas elles se trouveront à l'extérieur de l'ellipse).

Une représentation paramétrique utile et évidente de l'ellipse est:

$$\begin{aligned}
 x &= a \cos(t) \\
 y &= b \sin(t) \quad (24.61)
 \end{aligned}$$

Effectivement si nous considérons l'équation cartésienne de l'ellipse démontrée précédemment:

$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} = 1 \quad (24.62)$$

et en posant $R = X/a$ et $S = Y/b$ alors nous obtenons:

$$R^2 + S^2 = 1 \quad (24.63)$$

Si nous nous souvenons du cercle trigonométrique, cette équation admet les solutions $R = \cos(t)$ et $S = \sin(t)$. Il vient alors:

$$R = \frac{X}{a} = \cos(t) \Rightarrow X = a \cos(t)$$

$$S = \frac{Y}{a} = \sin(t) \Rightarrow Y = b \sin(t)$$
(24.64)

Voilà...

Cependant, il existe une autre forme d'équation de l'ellipse, bien plus importante, que l'on retrouve aussi bien en physique classique, astrophysique et physique quantique corpusculaire.

Rappelons que:

$$e \cdot d(M, D) = d(M, F)$$

$$\Leftrightarrow e \cdot |x + h| = \sqrt{x^2 + y^2}$$
(24.65)

En coordonnées polaires, cela donne:

$$e \cdot d(M, D) = d(M, F)$$

$$\Leftrightarrow e \cdot |r \cos \theta + h| = r$$
(24.66)

Donc:

$$r = e(r \cos \theta - h) \text{ ou } r = -e(r \cos \theta - h)$$
(24.67)

après mise en évidence:

$$r_1 = -\frac{eh}{1 - e \cos \theta} \text{ ou } r_2 = \frac{eh}{1 + e \cos \theta}$$
(24.68)

Nous obtenons deux équations différentes, mais il s'agit en fait de la même courbe qui décrit le rayon de l'ellipse depuis un de ses deux foyers.

Nous remarquerons en effet que:

$$r_1(\theta + \pi) = -r_2(\theta)$$
(24.69)

Etant donné que $p = eh$ est défini comme le paramètre de la conique, l'équation polaire de l'ellipse s'écrit:

$$r = \frac{p}{1 + e \cos \theta}$$
(24.70)

Remarquons les trois valeurs particulières:

$$\begin{aligned}\theta = \frac{\pi}{2} &\Rightarrow r = p = eh : \text{paramètre de l'ellipse} \\ \theta = 0 &\Rightarrow r_0 = \frac{p}{1+e} : \text{"péricentre"} \\ \theta = \pi &\Rightarrow r_\pi = \frac{p}{1-e} : \text{"apocentre"}\end{aligned}\quad (24.71)$$

Le péricentre est plus connu sous le nom de "périgée" en astronomie et de même que l'apocentre qui est plus connu sous le nom "d'apogée" (toujours dans le domaine de l'astronomie).

Dans le cas général, D peut faire un angle quelconque avec l'axe des angles polaires, et l'équation générale est alors:

$$r = \frac{p}{1 + e \cos(\theta - \theta_0)} \quad (24.72)$$

3. Cas où $e > 1$:

Il s'agit d'une hyperbole (même raisonnement que l'ellipse en commençant par considérer n'importe quelle valeur que $e > 1$ ou non):

$$(1 - e^2) \left(x - \frac{e^2 h}{1 - e^2} \right)^2 + y^2 - \frac{e^2 h^2}{1 - e^2} = 0 \Leftrightarrow \frac{\left(x - \frac{e^2 h}{1 - e^2} \right)^2}{\frac{e^2 h^2}{(1 - e^2)^2}} + \frac{y^2}{\frac{e^2 h^2}{1 - e^2}} = 1 \quad (24.73)$$

Nous posons donc à nouveau que:

$$F = \frac{e^2 h}{1 - e^2} \quad (24.74)$$

est l'origine de l'hyperbole. L'équation précédente se simplifie et devient (donc jusque là, nous nous retrouvons avec exactement la même expression que pour l'ellipse):

$$\frac{x^2}{\frac{e^2 h^2}{(1 - e^2)^2}} + \frac{y^2}{\frac{e^2 h^2}{1 - e^2}} = 1 \quad (24.75)$$

Mais comme $e > 1$ le dénominateur du deuxième terme sera négatif et donc nous nous retrouvons bien avec l'équation réduit d'une hyperbole.

Nous avons pour demi-grand axe et demi-petit axe (raisonnement identique à l'ellipse):

$$|a| = \frac{p}{1 - e^2}, |b| = \frac{p}{\sqrt{1 - e^2}} \quad (24.76)$$

et:

$$e = \frac{c}{a} = \frac{\sqrt{a^2 + b^2}}{a} \quad (24.77)$$

et la figure correspondante:

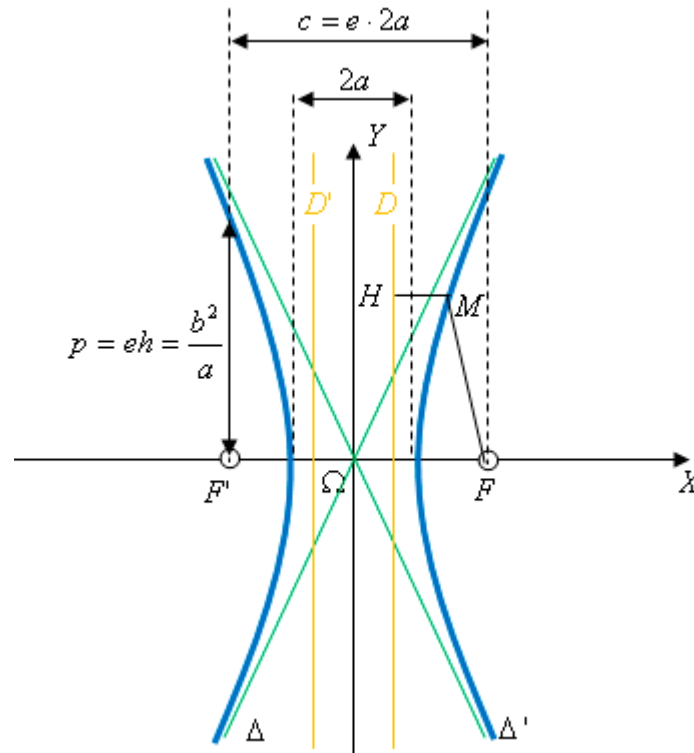


Figure: 24.10 - Représentation des paramètres des coniques pour une hyperbole

où nous avons représenté les deux asymptotes en utilisant une technique simple de passage à la limite:

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \Leftrightarrow y = \sqrt{\frac{b^2}{a^2}x^2 - b^2} \underset{x \rightarrow +\infty}{\cong} \sqrt{\frac{b^2}{a^2}x^2} = \pm \frac{b}{a}x = \Delta \quad (24.78)$$

Si l'hyperbole est équilatère, il est alors évident que les deux asymptotes sont perpendiculaires puisque leur pente est alors respectivement de +1 et de -1. Dans le cas équilatère, nous avons aussi l'excentricité qui est immédiatement donnée par:

$$e = \frac{\sqrt{a^2 + b^2}}{a} = \frac{\sqrt{a^2 + a^2}}{a} = \sqrt{2} \quad (24.79)$$

Nous pouvons aussi déterminer où se trouve les directrices D des hyperboles par rapport à leur bordure en utilisant la définition de l'excentricité lorsque $y = 0$ en procédant exactement de même que pour l'ellipse. Nous avons alors:

$$\begin{aligned} e \cdot d(M, D) &= d(M, F) \\ \Leftrightarrow d(M, D) &= \frac{d(M, F)}{e} = \frac{ea - a}{e} = a \frac{e - 1}{e} \end{aligned} \quad (24.80)$$

donc le facteur de a est compris entre 0 et tends vers 1 quand e est très grand ce qui signifie que les directrices se trouveront (du moins dans le cas ici présent où l'excentricité est positive et strictement plus petite que l'unité) soit tangentes à l'hyperbole soit au plus proche de l'axe des ordonnées (mais dans tous les cas elles se trouveront donc entre les deux hyperboles).

2. PARAMÉTRISATIONS

Pour certaines des formes présentées ci-dessous, il est possible de choisir un autre système de coordonnées que les coordonnées cartésiennes tel que par exemple les coordonnées cylindriques ou sphériques qui sont dans certains cas beaucoup plus simples à mettre en place. Nous tacherons dans la mesure du possible de présenter les plus importantes.

2.1. ÉQUATION DU PLAN

Soit un plan P dont nous connaissons un vecteur normal et unitaire $\vec{n}(a, b, c)$ mais pas l'équation et $A(x_A, y_A, z_A)$ un point de P .

Pour qu'un point M de coordonnées (x, y, z) appartienne au plan P il faut et il suffit que les vecteurs $\overrightarrow{AM}$ et $\vec{n}$ soient orthogonaux. Donc soit le point donné par le vecteur $\overrightarrow{AM}$ étant de coordonnées:

$$\overrightarrow{AM} = \begin{pmatrix} x - x_A \\ y - y_A \\ z - z_A \end{pmatrix} \quad (24.81)$$

Si $\overrightarrow{AM}$ est perpendiculaire à $\vec{n}$ alors le produit scalaire doit être nul tel que:

$$\begin{pmatrix} x - x_A \\ y - y_A \\ z - z_A \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix} = a(x - x_A) + b(y - y_A) + c(z - z_A) = 0 \quad (24.82)$$

Ce qui s'écrit aussi:

$$ax + by + cz + \underbrace{(-ax_A - by_A - cz_A)}_d = 0 \quad (24.83)$$

tel que nous obtenions l'équation cartésienne générale du plan:

$$\boxed{ax + by + cz + d = 0} \quad (24.84)$$

Cette équation où $(a, b, c) \neq (0, 0, 0)$ qui vérifie que les coordonnées d'un point $M(x, y, z)$ quelconque du plan P appartienne à ce plan est donc appelée "**équation cartésienne du plan P** ".

Si nous écrivons l'équation avec les cosinus directeurs de $\vec{n}$ (cf. chapitre de Calcul Vectoriel), nous avons dès lors aussi:

$$\cos(\theta_x)x + \cos(\theta_y)y + \cos(\theta_z)z + d = 0 \quad (24.85)$$

Remarque: Pour obtenir un cube dans l'espace, il suffit d'avoir six plans délimités par des conditions telles que $x \leq \dots, y \leq \dots, z \leq \dots$

Il est relativement aisé de passer au cas par cas de l'équation cartésienne du plan à l'équation paramétrique du plan. Nous pouvons (avec les précautions d'usage...) reprendre l'équation:

$$ax + by + cz + d = 0 \quad (24.86)$$

et la réécrire sous la forme suivante:

$$z = \frac{-d - by - ax}{c} \quad (24.87)$$

et dès lors, l'équation paramétrique du plan dans l'espace de dimension 3, s'écrira:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \\ y \\ \frac{-d - by - ax}{c} \end{pmatrix} \quad (24.88)$$

Exemple:

Avec Maple 4.00b:

```
>a:=3:b:=-2:c:=1:d:=5:
```

```
>plot3d([x,y,(-d-b*y+a*x)/c],x=-2..2,y=-2..2, orientation=[-87,81],style=PATCH,
axes=NORMAL);
```

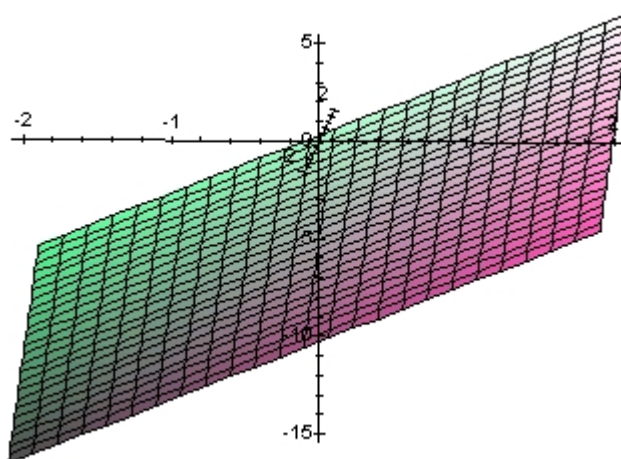


Figure: 24.11 - Représentation paramétrique d'un plan avec Maple 4.00b

2.2. ÉQUATION D'UNE DROITE

Comme nous l'avons vu en analyse fonctionnelle, une droite dans le plan peut être décrite par la fonction:

$$y = f(x) = ax + b \quad (24.89)$$

L'équation cartésienne généralisée de la droite est alors simplement donnée par:

$$\boxed{\alpha x + \beta y + \gamma = 0} \quad (24.90)$$

Effectivement, en simplifiant nous retrouvons "l'équation cartésienne réduite":

$$y = -\frac{\alpha}{\beta}x - \frac{\gamma}{\beta} = ax + b \quad \left. \vphantom{y = -\frac{\alpha}{\beta}x - \frac{\gamma}{\beta} = ax + b} \right\} \begin{array}{l} a = -\frac{\alpha}{\beta} \\ b = -\frac{\gamma}{\beta} \end{array} \quad (24.91)$$

Définition: Nous appelons "**vecteur directeur**" d'une droite D , tout vecteur non nul de même direction que la droite.

Montrons maintenant les deux petits théorèmes sympathiques suivants:

T1. Si une droite a pour équation $y = ax + b$ alors le vecteur:

$$\vec{v}(1, a) \quad (24.92)$$

est directeur de cette droite.

T2. Si une droite a pour équation $ax + by + c = 0$ alors le vecteur:

$$\vec{v}(1, -a/b) \quad (24.93)$$

est directeur de cette droite.

Démonstrations:

DM1. Soient $D: ax + b$ et A, B deux points de cette droite pris tels que $A(0, b), B(1, a + b)$. Comme A, B sont deux points de D alors $\overrightarrow{AB}$ est un vecteur directeur de D alors:

$$\overrightarrow{AB} = B - A = \begin{pmatrix} 1 \\ a + b \end{pmatrix} - \begin{pmatrix} 0 \\ b \end{pmatrix} = \begin{pmatrix} 1 \\ a \end{pmatrix} = \vec{v} \quad (24.94)$$

Un petit corollaire intéressant au passage qui a une application en physique!:

Si une droite D_1 a un vecteur directeur valant:

$$\vec{v}_1 = \begin{pmatrix} 1 \\ a \end{pmatrix} \quad (24.95)$$

et une autre droite D_2 un vecteur directeur valant:

$$\vec{v}_2 = \begin{pmatrix} 1 \\ -\frac{1}{a} \end{pmatrix} \quad (24.96)$$

alors leur produit scalaire (cf. chapitre de Calcul Vectoriel) est nul, ce qui montre que deux droites dont la multiplication des pentes (deuxième coordonnée du vecteur directeur) vaut -1 sont perpendiculaires!

DM2. Soit $D: ax + by + c = 0$ donc $y = -(a/b)x - c/b$ alors le vecteur $\vec{v}(1, -a/b)$ est un vecteur directeur de D ainsi que tout vecteur $\vec{v}' = \alpha \vec{v}, \alpha \in \mathbb{R}$. Ainsi, il existe une infinité de manières de définir la même droite, puisque la droite est composée d'une infinité de points (qui peuvent tous servir de point d'ancrage) et qu'il existe une infinité de multiples du vecteur directeur.

□C.Q.F.D.

Souvent, nous recherchons la distance entre une droite et un point externe à celle-ci. Ainsi, considérons la figure suivante:

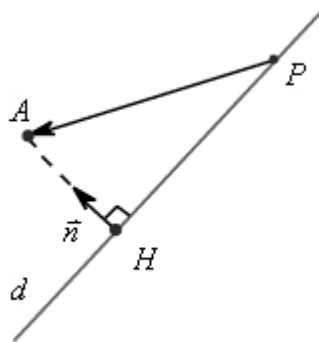


Figure: 24.12 - Représentation de la recherche de la distance d'un point à une droite

avec H la projection orthogonale de A sur la droite d , P un point arbitraire de d et $\vec{n}$ un vecteur orthogonal (normal) à d .

Nous avons (cf. chapitre de Calcul Vectoriel):

$$\overrightarrow{PA} \circ \vec{n} = (\overrightarrow{PH} + \overrightarrow{HA}) \circ \vec{n} = \overrightarrow{PH} \circ \vec{n} + \overrightarrow{HA} \circ \vec{n} = \overrightarrow{HA} \circ \vec{n} = \|\overrightarrow{HA}\| \|\vec{n}\| \cos \alpha = \pm \|\overrightarrow{HA}\| \|\vec{n}\| \quad (24.97)$$

car $\alpha = 0$ ou $\alpha = \pi$. Ainsi:

$$|\overrightarrow{PA} \circ \vec{n}| = \|\overrightarrow{HA}\| \circ \|\vec{n}\| = \delta(A, d) \circ \|\vec{n}\| \quad (24.98)$$

où δ représente la distance (nous ne pouvons noter la distance avec la lettre d comme nous l'avons fait au début de ce chapitre, sinon il y aurait confusion avec le d choisi pour représenter la droite dans le présent développement).

Nous obtenons donc la relation:

$$\delta(A, d) = \frac{|\overrightarrow{PA} \circ \vec{n}|}{\|\vec{n}\|} \quad (24.99)$$

Considérons maintenant le point $A(x_0, y_0)$ et la droite $d : ax + by + c = 0$.

Choisissons un point $P \in d : P(0, -c/b)$ ainsi qu'un vecteur $\vec{n}$, normal à $d : \vec{n} \begin{pmatrix} a \\ b \end{pmatrix}$ (rappelons que $\vec{v}(1, -a/b)$ est vecteur directeur):

Ainsi, en appliquant la relation précédente nous avons:

$$\overrightarrow{PA} \circ \vec{n} = \left(\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} 0 \\ -c/b \end{pmatrix} \right) \circ \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 + c/b \end{pmatrix} \circ \begin{pmatrix} a \\ b \end{pmatrix} = x_0 a + \left(y_0 + \frac{c}{b} \right) b \quad (24.100)$$

et donc:

$$\delta(A, d) = \frac{|x_0 a + (y_0 + c/b)b|}{\sqrt{a^2 + b^2}} = \frac{|ax_0 + by_0 + c|}{\sqrt{a^2 + b^2}} \quad (24.101)$$

Si nous considérons maintenant deux plans non parallèles de l'espace, leur intersection est une droite. Soient deux plans d'équations respectives:

$$\begin{aligned} P &= ax + by + cz + d = 0 \\ P' &= a'x + b'y + c'z + d' = 0 \end{aligned} \quad (24.102)$$

et D leur droite d'intersection.

Un point $M(x, y, z)$ de l'espace appartient à la droite D si et seulement si le point M satisfait le système d'équations:

$$\begin{aligned} ax + by + cz + d &= 0 \\ a'x + b'y + c'z + d' &= 0 \end{aligned} \quad (24.103)$$

Remarque: Alors que dans le plan une droite est caractérisée par une équation du type $ax + by + c = 0$ (cf. chapitre d'Analyse Fonctionnelle), dans l'espace, une seule équation de la forme $ax + by + cz + d = 0$ caractérise un plan. Pour caractériser une droite en dehors des plans des axes, il est nécessaire (équation paramétrique mise à part) d'avoir deux équations.

Il est relativement trivial (mais nous allons quand même le démontrer) que l'équation paramétrique d'une droite est un système d'équations du type:

$$\begin{cases} x = a \cdot t + \alpha \\ y = b \cdot t + \beta \\ z = c \cdot t + \gamma \end{cases} \quad (24.104)$$

Ainsi, chaque composante croît linéairement par rapport à la même variable à une constante et un facteur près. Ceci s'écrit aussi sous forme vectorielle (plus traditionnelle):

$$\begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} t + \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} \quad (24.105)$$

Le vecteur (a, b, c) est appelé "**vecteur directeur**".

Démonstration: Nous avons donc le système d'équations (deux équations à trois inconnues, ainsi une inconnue sera indéterminée):

$$\begin{aligned} ax + by + cz + d &= 0 \\ a'x + b'y + c'z + d' &= 0 \end{aligned} \quad (24.106)$$

Éliminons une des variables (commençons arbitrairement par z):

$$(a - a'\alpha)x + (b - b'\alpha)y + \underbrace{(c - c'\alpha)}_0 z + (d - d'\alpha) = 0 \quad (24.107)$$

où $\alpha = c/c'$ d'où:

$$\begin{aligned} (a - a'\alpha)x + (b - b'\alpha)y + (d - d'\alpha) &= 0 \\ \Rightarrow \\ \left. \begin{aligned} x &= \frac{(b'\alpha - b)}{(a - \alpha a')} y + \frac{(d'\alpha - d)}{(a - \alpha a')} \\ y &= \frac{(a'\alpha - a)}{(b - \alpha b')} x + \frac{(d'\alpha - d)}{(b - \alpha b')} \end{aligned} \right\} x &= \frac{(b'\alpha - b)}{(a - \alpha a')} \left(\frac{(a'\alpha - a)}{(b - \alpha b')} x + \frac{(d'\alpha - d)}{(b - \alpha b')} \right) + \frac{(d'\alpha - d)}{(a - \alpha a')} \end{aligned} \quad (24.108)$$

donc (c'est un peu bête à écrire mais bon...):

$$x = \underbrace{\frac{(b'\alpha - b)}{(a - \alpha a')} \frac{(a'\alpha - a)}{(b - \alpha b')}}_{-1} x + \underbrace{\frac{(b'\alpha - b)}{(a - \alpha a')} \frac{(d'\alpha - d)}{(b - \alpha b')}}_0 + \frac{(d'\alpha - d)}{(a - \alpha a')} = x \quad (24.109)$$

De manière similaire avec y tel $\beta = b/b'$ nous avons:

$$\begin{aligned} (a - a'\beta)x + (c - c'\beta)z + (d - d'\beta) &= 0 \\ \Rightarrow z &= \frac{(a'\beta - a)}{(c - c'\beta)} x + \frac{(d'\beta - d)}{(c - c'\beta)} \end{aligned} \quad (24.110)$$

d'où:

$$z = \frac{(a'\beta - a)}{(c - \beta c')}x + \frac{(d'\beta - d)}{(c - \beta c')} \quad (24.111)$$

Finalement nous avons:

$$\left. \begin{aligned} x &= x \\ y &= \frac{(a'\alpha - a)}{(b - \alpha b')}x + \frac{(d'\alpha - \alpha)}{(b - \alpha b')} \\ z &= \frac{(a'\beta - a)}{(c - \beta c')}x + \frac{(d'\beta - d)}{(c - \beta c')} \end{aligned} \right\} x \equiv t \Rightarrow \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{(a'\alpha - a)}{(b - \alpha b')} \\ \frac{(a'\beta - a)}{(c - \beta c')} \end{pmatrix} t + \begin{pmatrix} 0 \\ \frac{(d'\alpha - \alpha)}{(b - \alpha b')} \\ \frac{(d'\beta - d)}{(c - \beta c')} \end{pmatrix} \quad (24.112)$$

Le vecteur directeur et le vecteur d'ordonnée sont donc des constantes. Ce qui nous permet d'écrire de manière plus générale:

$$\begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} t + \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} \quad (24.113)$$

□C.Q.F.D.

Remarques:

R1. L'équation d'une droite est presque ce qu'il y a de plus important en synthèse d'images 3D car à partir de ces dernières nous pouvons construire des polygones et assembler ces derniers pour construire des formes tridimensionnelles plus complexes.

R2. Pour savoir si une droite est perpendiculaire à un plan il faut déterminer au moins deux droites sécantes dans ce même plan et effectuer le produit vectoriel de leur vecteur directeur et ensuite calculer le produit scalaire entre le résultat du produit vectoriel et la première droite dont nous cherchons l'orthogonalité. Effectivement, une seule droite du plan ne permet pas de déterminer l'orientation de ce dernier; il en faut au moins deux.

2.3. ÉQUATION D'UN CÔNE

Considérons un cône de révolution de sommet O avec un cercle de centre $(0,0,h)$ (où h est un nombre réel positif) et de rayon R :

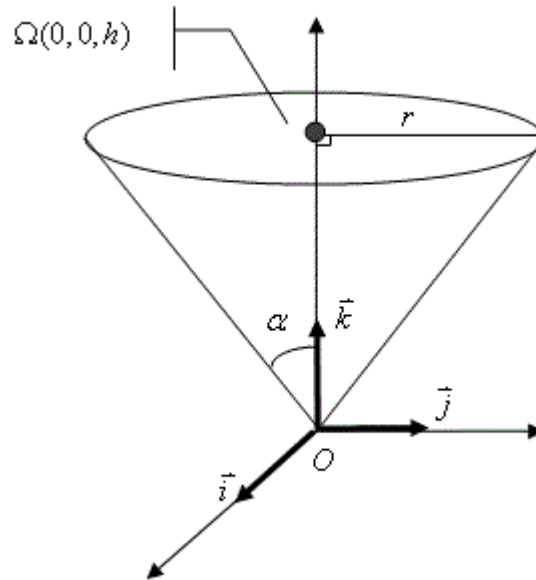


Figure: 24.13 - Représentation d'un cône

Une représentation paramétrique du cercle à la hauteur h est:

$$\begin{cases} x = r \cdot \cos t \\ y = r \cdot \sin t \\ z = h \end{cases} \quad (24.114)$$

où t appartient à l'intervalle $[-\pi, \pi]$.

Par extension, une représentation paramétrique de ce cône est :

$$\begin{pmatrix} x(t,k) \\ y(t,k) \\ z(k) \end{pmatrix} = \begin{pmatrix} k \cdot r \cdot \cos t \\ k \cdot r \cdot \sin t \\ k \cdot h \end{pmatrix} \quad (24.115)$$

En d'autres termes, le cercle se propage linéaire selon tous les directions.

Exemple:

Avec Maple 4.00b:

```
>r:=1:h:=4:
>plot3d([k*r*cos(t),k*r*sin(t),k*h],k=0..10,t=0..2*Pi,orientation=[50,60],style=PATCH,
axes=NORMAL);
```

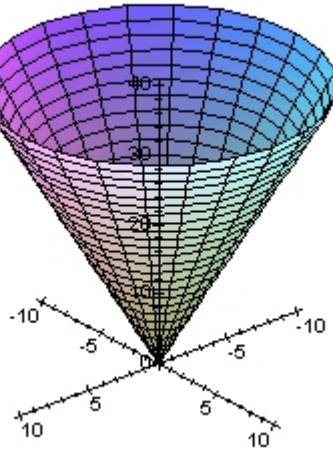


Figure: 24.14 - Représentation paramétrique d'un cône avec Maple 4.00b

On peut en déduire une équation cartésienne de ce cône:

$$X(t)^2 + Y(t)^2 = k^2 \cdot r^2 = \left(\frac{Z(t)}{h} \right)^2 r^2 \quad (24.116)$$

en posant $\alpha = r / h$:

$$X^2 + Y^2 - Z^2 \tan^2 \alpha = 0 \quad (24.117)$$

c'est l'équation cartésienne d'un cône dans l'espace que nous retrouverons en relativité restreinte lors de notre étude des cônes de lumière.

2.4. ÉQUATION D'UNE SPHÈRE

Considérons le repère orthonormé $(\vec{i}, \vec{j}, \vec{k})$, soit S la sphère de centre $\Omega(a, b, c)$ et de rayon r :

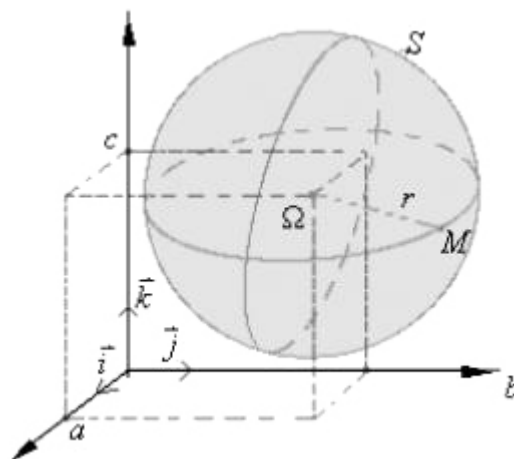


Figure: 24.15 - Représentation d'une sphère

$M(x, y, z)$ appartient à la sphère S de centre Ω et de rayon r si et seulement si:

$$\overline{\Omega M} = r \quad (24.118)$$

c'est-à-dire en appliquant Pythagore :

$$\overline{OM}^2 = r^2 = (x-a)^2 + (y-b)^2 + (z-c)^2 \quad (24.119)$$

D'où "l'équation cartésienne de la sphère" dans le repère $(\vec{i}, \vec{j}, \vec{k})$:

$$(x-a)^2 + (y-b)^2 + (z-c)^2 = r^2 \quad (24.120)$$

Il existe une autre manière de décrire la sphère en utilisant l'équation paramétrique. Effectivement, nous avons vu dans le chapitre de Calcul Vectoriel que le passage des coordonnées cartésiennes aux coordonnées sphériques est donné par les coordonnées curvilignes:

$$x = r \sin \theta \cos \phi, y = r \sin \theta \sin \phi, z = r \cos \theta \quad (24.121)$$

Ainsi, nous avons bien:

$$\begin{aligned} x^2 + y^2 + z^2 &= r^2 \sin^2 \theta \cos^2 \phi + r^2 \sin^2 \theta \sin^2 \phi + r^2 \cos^2 \theta \\ &= r^2 \sin^2 \theta (\cos^2 \phi + \sin^2 \phi) + r^2 \cos^2 \theta = r^2 \sin^2 \theta + r^2 \cos^2 \theta = r^2 (\sin^2 \theta + \cos^2 \theta) \\ &= r^2 \end{aligned} \quad (24.122)$$

Donc l'équation paramétrique de la sphère est bien:

$$\begin{pmatrix} x(\theta, \phi) \\ y(\theta, \phi) \\ z(\theta) \end{pmatrix} = \begin{pmatrix} r \cdot \sin \theta \cos \phi \\ r \cdot \sin \theta \sin \phi \\ r \cdot \cos \theta \end{pmatrix}$$

Nous retrouvons donc bien l'équation cartésienne d'une sphère à une constante de translation près.

Exemple:

Avec Maple 4.00b une sphère de rayon unité:

```
>plot3d([sin(theta)*cos(phi),sin(theta)*sin(phi),cos(theta)],theta=0..Pi,phi=-Pi..Pi,
scaling=CONSTRAINED,orientation=[50,60],style=PATCH,axes=NORMAL);
```

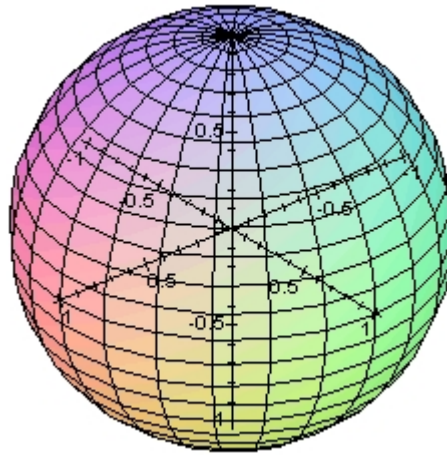


Figure: 24.16 - Représentation paramétrique d'une sphère avec Maple 4.00b

2.5. ÉQUATION D'UN ELLIPSOÏDE

Nous avons vu lors de notre étude des coniques que l'équation d'une ellipse dans le plan était donnée par:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \quad (24.123)$$

avec a, b étant les deux axes de l'ellipse (le petit et le grand).

Ainsi, sans démonstration rigoureuse, nous pouvons vérifier à la main ou à l'aide des ordinateurs que l'équation cartésienne:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \quad (24.124)$$

est un ellipsoïde:

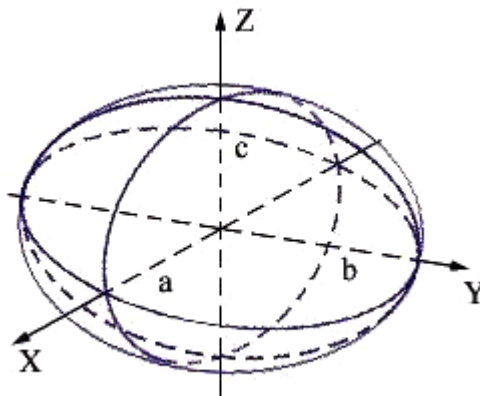


Figure: 24.17 - Représentation d'un ellipsoïde

Cependant, il existe une autre manière de décrire l'ellipsoïde en utilisant les coordonnées curvilignes:

$$x = a \cdot \cos \theta \cdot \cos \lambda, y = b \cdot \cos \theta \cdot \sin \lambda, z = c \cdot \sin \theta \quad (24.125)$$

Donc l'équation paramétrique de l'ellipsoïde sera:

$$\begin{pmatrix} x(\theta, \lambda) \\ y(\theta, \lambda) \\ z(\theta) \end{pmatrix} = \begin{pmatrix} a \cdot \cos \theta \cos \lambda \\ b \cdot \cos \theta \sin \lambda \\ c \cdot \sin \theta \end{pmatrix} \quad (24.126)$$

Nous pouvons constater qu'il s'agit simplement de l'équation paramétrique de la sphère mais avec des rayons différents suivant les axes du repère choisi.

Exemple:

Avec Maple 4.00b:

```
>a:=100;b:=2;c:=20;
```

```
>plot3d([a*cos(theta)*cos(lambda),b*cos(theta)*sin(lambda),c*sin(theta)],lambda=0..Pi,
theta=-Pi..Pi,orientation=[50,60],style=PATCH,axes=NORMAL);
```

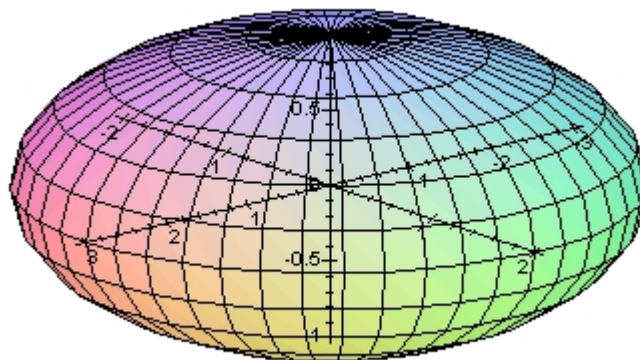


Figure: 24.18 - Représentation paramétrique d'un ellipsoïde avec Maple 4.00b

Nous avons donc:

$$x^2 = a^2 \cdot \cos^2 \theta \cdot \cos^2 \lambda, y^2 = b^2 \cdot \cos^2 \theta \cdot \sin^2 \lambda, z^2 = c^2 \cdot \sin^2 \theta \quad (24.127)$$

d'où:

$$\begin{aligned} \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} &= \cos^2 \theta \cdot \cos^2 \lambda + \cos^2 \theta \cdot \sin^2 \lambda + \sin^2 \theta = \cos^2 \theta (\cos^2 \lambda + \sin^2 \lambda) + \sin^2 \theta \\ &= \cos^2 \theta + \sin^2 \theta = 1 \end{aligned} \quad (24.128)$$

Finalement:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \quad (24.129)$$

2.6. ÉQUATION D'UN CYLINDRE

Il va sans dire que l'équation paramétrique d'un cylindre de rayon r est donnée par:

$$\begin{cases} x = r \cos \phi \\ y = r \sin \phi \\ z = z \end{cases} \quad (24.130)$$

Soit:

$$\begin{pmatrix} x(\phi) \\ y(\phi) \\ z \end{pmatrix} = \begin{pmatrix} r \cdot \cos \phi \\ r \cdot \sin \phi \\ z \end{pmatrix} \quad (24.131)$$

Nous voyons bien que les composantes x, y satisfont l'équation cartésienne d'un cercle puisque:

$$x^2 + y^2 = r^2 \quad (24.132)$$

Exemple:

Avec Maple 4.00b un cylindre de rayon unité:

```
>plot3d([cos(phi),sin(phi),z],phi=-Pi..Pi,z=0..2,orientation=[50,60],style=PATCH,
axes=NORMAL);
```

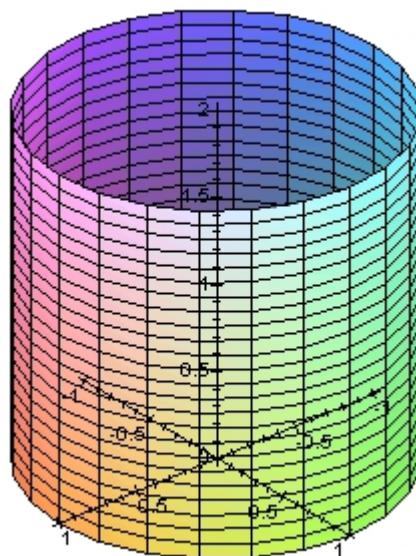


Figure: 24.19 - Représentation paramétrique d'un cylindre avec Maple 4.00b

Au même titre l'équation paramétrique d'un cylindre à base elliptique est donnée par:

$$\begin{cases} x = r_1 \cos \phi \\ y = r_2 \sin \phi \\ z = z \end{cases} \quad (24.133)$$

qui vérifie aussi l'équation paramétrique d'une ellipse dans le plan:

$$\frac{x^2}{r_1^2} + \frac{y^2}{r_2^2} = 1 \quad (24.134)$$

2.7. SURFACES DE RÉVOLUTION

De manière plus générale de nombreuses surfaces (dont certaines que nous avons vues précédemment) peuvent être décrites par révolution d'une forme primaire de dimension inférieure et ensuite par rotation.

Définition: Une "**surface de révolution**" est une surface obtenue en faisant tourner une courbe plane (par exemple $z = f(x)$) autour de l'axe OZ . Ainsi, nous passons alors d'un plan de $\mathbb{R}^2$ à un repère de $\mathbb{R}^3$, l'axe OX engendre dès lors un plan devenu YOZ .

Prenons trois exemples classiques (parmi l'infini):

Exemples:

E1. Soit la parabole particulière d'équation qui tourne autour de l'axe OZ :

$$z = x^2 / 2p \quad (24.135)$$

avec pour rappel:

$$p = e \cdot h$$

Nous avons bien évidemment (en coupant le paraboloïde par un plan ce qui donne un cercle de rayon r) la relation:

$$z = r^2 / 2p \quad (24.136)$$

dite "**équation cylindrique**". Or, nous avons aussi par simple application de Pythagore dans le cercle en question:

$$r^2 = x^2 + y^2 \quad (24.137)$$

d'où:

$$x^2 + y^2 = 2pz \quad (24.138)$$

Nous en déduisons "**l'équation cartésienne du paraboloïde de révolution**":

$$z = \frac{1}{2} \left(\frac{1}{p} x^2 + \frac{1}{p} y^2 \right) \quad (24.139)$$

Nous construisons l'équation paramétrique du paraboloïde exactement de la même manière que le cône, mais à la différence que l'évolution selon l'axe OZ ne se fera non pas linéairement par rapport au paramétrage k mais au carré de ce dernier. Ainsi, nous aurons:

$$\begin{pmatrix} x(t, k) \\ y(t, k) \\ z(k) \end{pmatrix} = \begin{pmatrix} k \cdot r \cdot \cos t \\ k \cdot r \cdot \sin t \\ k^2 \cdot h \end{pmatrix} = \begin{pmatrix} k \cdot \frac{1}{2p} \cdot \cos t \\ k \cdot \frac{1}{2p} \cdot \sin t \\ k^2 \cdot h \end{pmatrix} \quad (24.140)$$

Ce qui donnera avec Maple 4.00b:

```
>p:=1:h:=1:
>plot3d([k*1/(2*p)*cos(t),k*1/(2*p)*sin(t),k^2*h],k=0..10,t=0..2*Pi,orientation=[50,60]
,style=PATCH,axes=NORMAL);
```

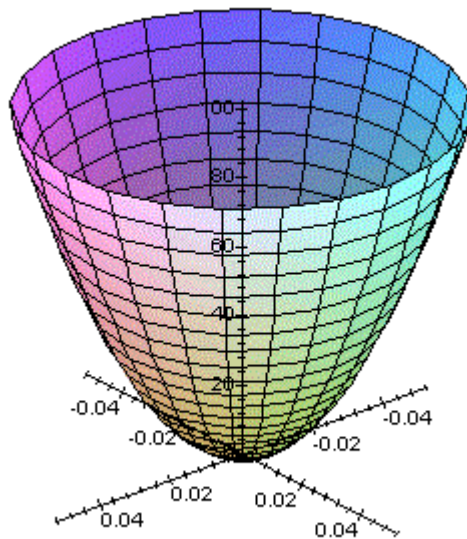


Figure: 24.20 - Représentation paramétrique d'un paraboloïde avec Maple 4.00b

E2. Soit la droite de pente unité:

$$z = x \quad (24.141)$$

que nous faisons tourner autour de Oz. Nous avons:

$$z^2 = x^2 + y^2 \quad (24.142)$$

ce qui nous donne:

$$x^2 + y^2 - z^2 = 0 \quad (24.143)$$

Définition: Toute surface engendrée par une droite est une "surface réglée".

Prenons l'exemple important (cheminée de centrale nucléaire, engrenages, etc.) qu'est l'hyperboloïde à une nappe d'équation:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1 \quad (24.144)$$

Pour simplifier l'exemple prenons $a = b = c = 1$. Nous avons donc:

$$x^2 + y^2 - z^2 = 1 \Rightarrow x^2 - z^2 = 1 - y^2 \quad (24.145)$$

ce qui s'écrit aussi comme le produit de l'équation de deux droites tel que:

$$\begin{aligned} (x-z)(x+z) &= (1-y)(1+y) \\ \Updownarrow \\ (x-z)k(x+z) &= k(1-y)(1+y), k \in \mathbb{R} \\ \Downarrow \\ \overbrace{D1: x-z = k(1-y)} \\ D2: 1+y &= k(x+z) \end{aligned} \quad (24.146)$$

Ainsi, ces deux droites (de pentes opposées) appartiennent à la nappe et tout point appartenant à une de ces deux droites y est contenu. Les figures ci-dessous montrent bien qu'au fait, tout point appartient à ces deux droites.

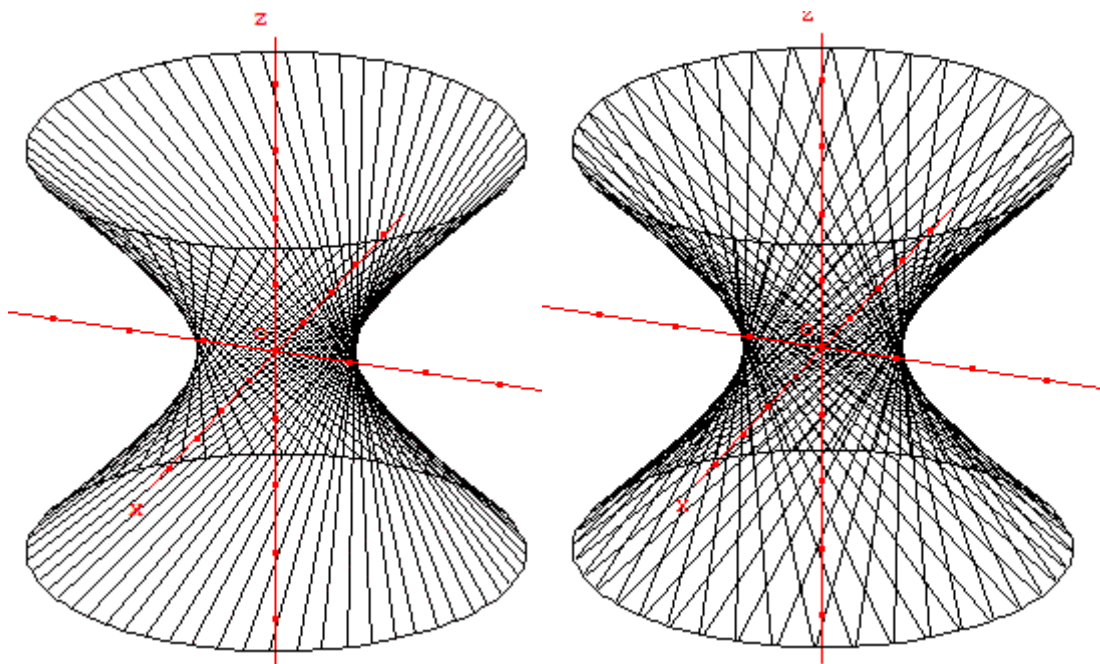


Figure: 24.21 - Construction d'une surface de révolution avec des droites

On pourrait ceci dit très bien décrire par des cercles tels que:

$$x^2 + y^2 - z^2 = 1 \Rightarrow x^2 + y^2 = (1 + z^2) = r^2 \quad (24.147)$$

où $r = f(z)$.

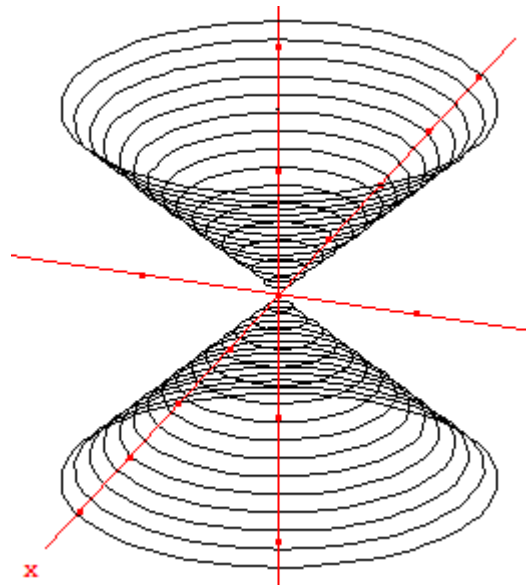


Figure: 24.22 - Construction d'une surface de révolution avec des cercles

E3. Nous savons que pour générer un cercle dans le plan XOY , une équation paramétrique possible est:

$$\begin{aligned} x &= r \cos \phi \\ y &= r \sin \phi \end{aligned} \quad (24.148)$$

Pour décaler ce cercle vers la droite (dans la direction des x positifs), nous n'aurons qu'à ajouter une constante strictement positive au terme x :

$$\begin{aligned} x &= a + r \cos \phi \\ y &= r \sin \phi \end{aligned} \quad (24.149)$$

Dans l'espace, si nous décidons de dessiner un cercle sur la plan XOZ , nous aurons alors:

$$\begin{aligned} x &= a + r \cos \phi \\ z &= r \sin \phi \end{aligned} \quad (24.150)$$

Qu'il est d'usage de noter sous la forme suivante lorsqu'il s'agit de l'étude du tore:

$$\begin{aligned} x &= R + r \cos \phi \\ z &= r \sin \phi \end{aligned} \quad (24.151)$$

Si nous voulons générer un tore, il faudra faire tourner ce cercle autour de l'axe Z en lui faisant suivre à lui-même un cercle suivant le plan XOY . Nous avons alors "**l'équation paramétrique du tore de révolution**":

$$\begin{pmatrix} x(\phi, \theta) \\ y(\phi, \theta) \\ z(\phi) \end{pmatrix} = \begin{pmatrix} (R + r \cos \phi) \cos \theta \\ (R + r \cos \phi) \sin \theta \\ r \sin \phi \end{pmatrix} \quad (24.152)$$

Ce qui donne avec Maple 4.00b:

```
>r:=1:R:=4:  
>plot3d([(R+r*cos(phi))*cos(theta),(R+r*cos(phi))*sin(theta), 1*sin(phi)],  
theta=-Pi..Pi,phi=-Pi..Pi,scaling=CONSTRAINED,orientation=[50,60],  
style=PATCH,axes=NORMAL);
```

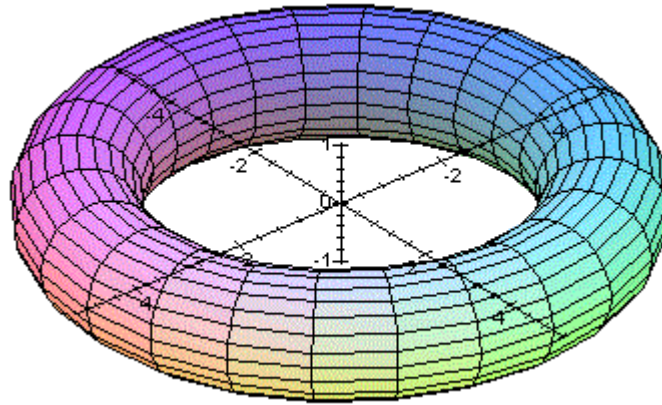
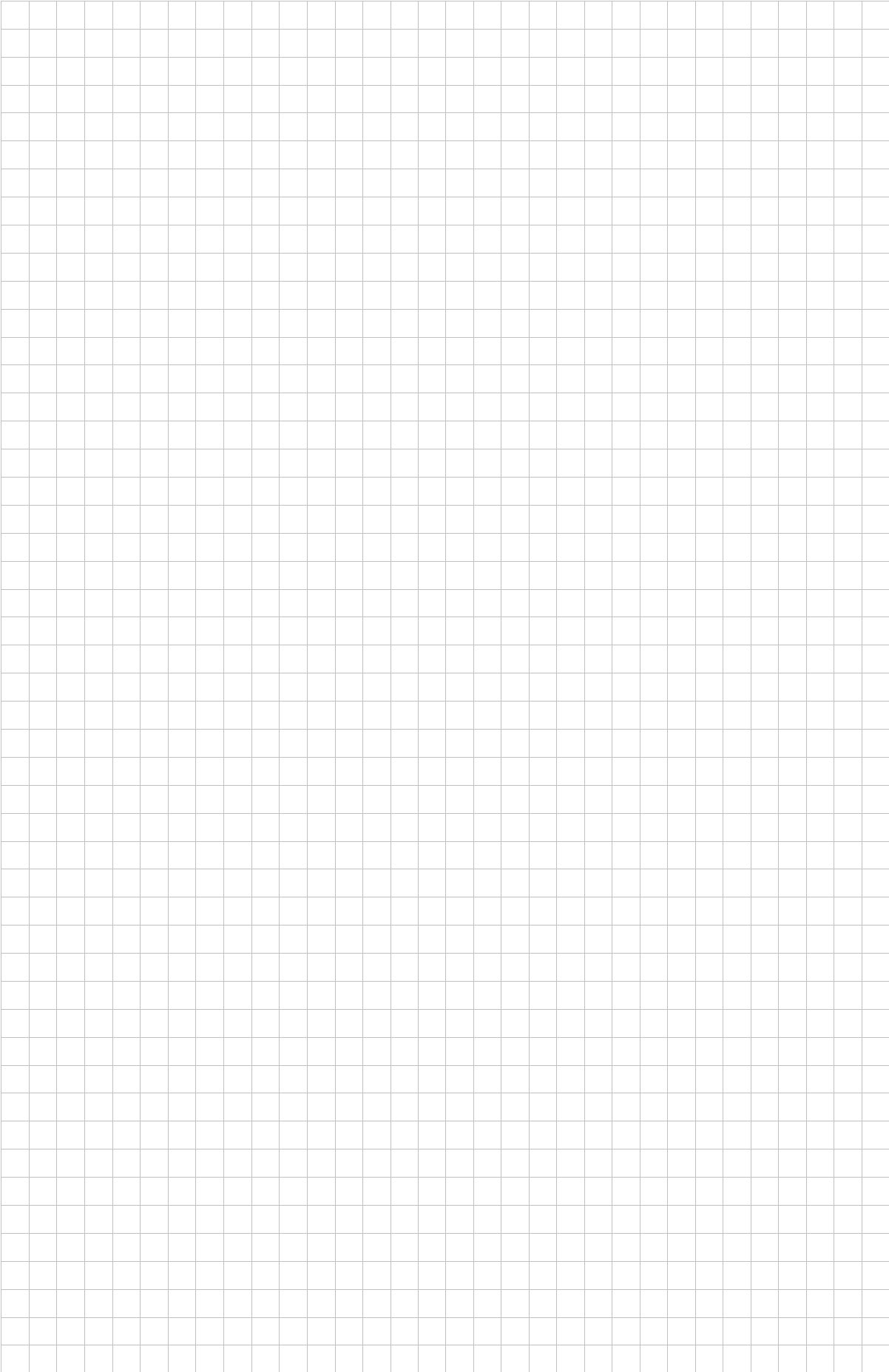
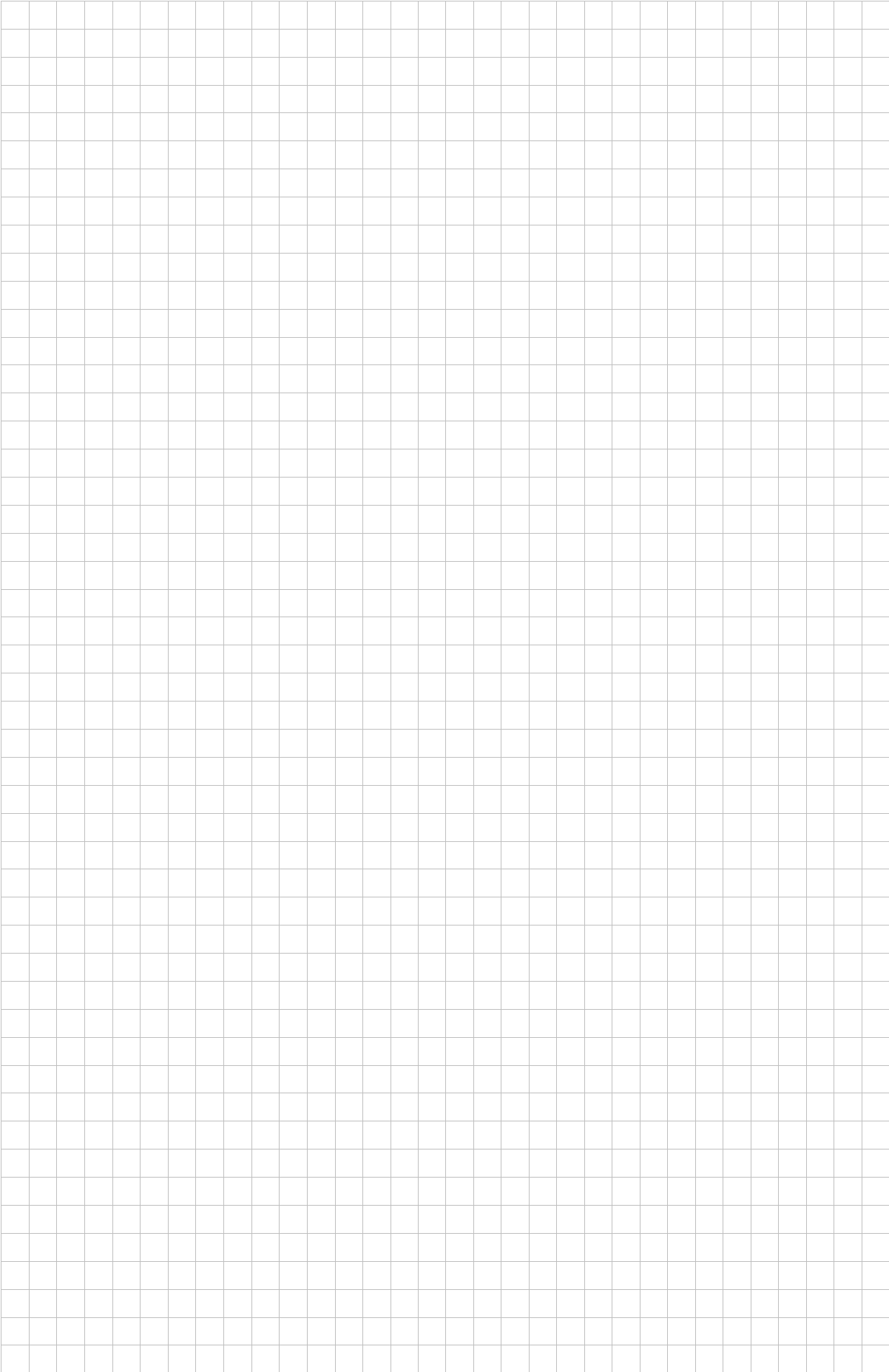


Figure: 24.23 - Représentation paramétrique d'un tore avec Maple 4.00b

Notes personnelles:







25. GÉOMÉTRIE DIFFÉRENTIELLE

Comme nous l'avons déjà mentionné dans le chapitre de Géométrie Non-Euclidienne, la géométrie différentielle est la branche de la géométrie qui vise à étudier les propriétés locales (au voisinage d'un point) et intrinsèques des courbes et des surfaces non-euclidiennes (comme une généralisation des surfaces euclidiennes!).

La géométrie différentielle tient son nom au fait qu'elle est née de la possibilité d'une interprétation cinématique que le calcul infinitésimal apporte à l'étude des courbes. Les points que nous aborderons ici serviront aussi bien dans l'étude de la mécanique classique que de l'analyse complexe appliquée à de nombreux domaines de l'étude des champs.

Remarque: Avant de nous attaquer à la manière très formelle et abstraite d'aborder la géométrie différentielle avec les outils de la topologie (méthode habituelle aux mathématiciens) nous avons choisi dans un premier temps de présenter les éléments essentiels de manière simple et agréable tel que cela se fait dans les écoles d'ingénieurs. Les puristes nous excuseront donc au cas où en attendant mieux...

1. COURBES PARAMÉTRÉES

Définition: Nous assimilerons "l'espace physique" à $\mathbb{R}^3$ et le supposons muni d'un repère $R = (O, \vec{i}, \vec{j}, \vec{k})$ et nous noterons B la base $(\vec{i}, \vec{j}, \vec{k})$

Soient un ensemble $I \subset \mathbb{R}$ et une fonction $\Gamma = f(I)$ telle que:

$$f \left(\begin{array}{l} I \rightarrow \mathbb{R}^3 \\ t \rightarrow f(t) = M \end{array} \right), M(x, y, z)_{\mathbb{R}} \quad (25.1)$$

Remarques:

R1. Si f est continue, alors Γ est une courbe de l'espace appelée "courbe d'un seul tenant".

R2. Une parabole, une sinusoïde sont des courbes appelées "courbes planes". Une ellipse, un cercle sont appelés des "courbes planes fermées". Pour ces exemples, tous les points des courbes considérées sont situés dans un même plan. Inversement, une courbe est appelée "courbe gauche" (gauchir = dévier, tordre) s'il n'en est pas ainsi.

Choisissons $t_0 \in I$ et posons $M_0 = f(t_0)$ que nous noterons par abus de langage $M(t_0)$ nous pouvons alors énoncer la définition suivante: le couple (f, I) où f est une fonction continue est appelé "arc paramétré". Γ est appelée le "support" de (f, I) et t_0 est une "origine" de (f, I) .

Remarques:

R1. Abusivement, nous disons aussi que (f, I) est un "paramétrage" de Γ .

R2. Il est facile de définir d'autres arcs paramétrés admettant aussi Γ comme support. Pour ce faire, il suffit de se donner une fonction φ bijective de I vers $J \subset \mathbb{R}$ et telle que $f(t) = f(\varphi^{-1}(x))$.

Avant de continuer, rappelons qu'en géométrie différentielle, "l'abscisse curviligne" est une sorte de variante algébrique de la longueur d'un arc (c'est donc l'analogue, sur une courbe, de l'abscisse sur une droite orientée).

Considérons maintenant l'abscisse curviligne (cf. chapitre de Calcul Tensoriel):

$$ds^2 = g_{ij} du^i du^j \quad (25.2)$$

nous savons que dans un espace euclidien canonique (dans $\mathbb{R}^3$) l'abscisse curviligne s'écrit alors:

$$ds^2 = \delta_{ij} du^i du^j \quad (25.3)$$

avec $i, j = 1, 2, 3$ et comme nous avons $\delta_{ij} = 0, i \neq j$, il reste:

$$ds^2 = du^1 du^1 + du^2 du^2 + du^3 du^3 = (du^1)^2 + (du^2)^2 + (du^3)^2 \quad (25.4)$$

Dans le système cartésien:

$$u^1 = x, u^2 = y, u^3 = z \quad (25.5)$$

il vient donc que:

$$ds^2 = dx^2 + dy^2 + dz^2 \quad (25.6)$$

qui est donc l'élément différentiel linéaire d'un espace euclidien (le plus court chemin ou encore la "géodésique" ou encore "l'abscisse curviligne différentielle") que nous avons déjà rencontré à maintes reprises dans différents chapitres du site. Ce n'est donc rien de nouveau ni de surprenant!

Si nous nous restreignons au plan, l'abscisse curviligne différentielle d'une courbe plane s'écrit alors:

$$ds^2 = dx^2 + dy^2 \quad (25.7)$$

Nous savons déjà comment utiliser cette dernière équation (nous en avons fait usage dans le chapitre de Mécanique Analytique). Mais comme un rappel n'est jamais mauvais, faisons des exemples avec une droite, une parabole et un demi-cercle (le choix n'est pas innocent...).

Exemples:

E1. Considérons l'équation générale d'une droite dans le plan (ce n'est pas une courbe plane pour rappel mais une droite plane):

$$y = ax + b \quad (25.8)$$

Il vient alors immédiatement:

$$dy = a dx \quad (25.9)$$

Dès lors:

$$ds^2 = dx^2 + dy^2 = dx^2 + (a dx)^2 = (1 + a^2) dx^2 \quad (25.10)$$

E2. Considérons l'équation générale d'une parabole dans le plan:

$$y = ax^2 + b \quad (25.11)$$

Il vient alors immédiatement:

$$dy = 2ax dx \quad (25.12)$$

Dès lors:

$$ds^2 = dx^2 + dy^2 = dx^2 + (2ax dx)^2 = (1 + 4a^2 x^2) dx^2 \quad (25.13)$$

E3. Considérons le réarrangement de l'équation générale d'un cercle de rayon R dans le plan (cf. chapitre de Géométrie Analytique):

$$y = \sqrt{b^2 \left(R - \frac{x^2}{a^2} \right)} = \sqrt{b^2 R - b^2 \frac{x^2}{a^2}} \quad (25.14)$$

Il vient alors immédiatement:

$$dy = - \frac{b^2}{a^2 \sqrt{b^2 R - b^2 \frac{x^2}{a^2}}} x dx \quad (25.15)$$

Dès lors:

$$ds^2 = dx^2 + dy^2 = dx^2 + \left(- \frac{b^2}{a^2 \sqrt{b^2 R - b^2 \frac{x^2}{a^2}}} x dx \right)^2 = \left(1 + \frac{b^2}{(a^4 R - a^2 x^2)} x^2 \right) dx^2 \quad (25.16)$$

Remarquons qu'en faisant une approximation de Maclaurin (lorsque x vaut donc zéro, ce qui correspond à l'étude du pôle du cercle), nous avons (cf. chapitre de Suites Et Séries):

$$y = \sqrt{b^2 \left(R - \frac{x^2}{a^2} \right)} \cong \sqrt{b^2 R} - \frac{1}{2} \frac{\sqrt{b^2 R}}{a^2 R} x^2 \quad (25.17)$$

Suite à la demande d'un lecteur voici les détails du développement du résultat précédent. Rappelons d'abord la série de Taylor (cf. chapitre Suites Et Séries):

$$f(x) \cong f(x_0) + \frac{1}{1!} f'(x_0)(x - x_0) + \frac{1}{2!} f''(x_0)(x - x_0)^2 + \dots \quad (25.18)$$

Si nous posons $x_0 = 0$, nous obtenons la série de Maclaurin:

$$f(x) \cong f(0) + \frac{1}{1!} f'(0)x + \frac{1}{2!} f''(0)x^2 + \dots \quad (25.19)$$

Donc il vient en procédant en deux étapes:

$$\begin{aligned}
y &\cong \frac{d}{dx^0} \sqrt{b^2 \left(R - \frac{x^2}{a^2} \right)} \Big|_{x=x_0} (x-x_0)^0 + \frac{1}{1!} \frac{d}{dx^1} \sqrt{b^2 \left(R - \frac{x^2}{a^2} \right)} \Big|_{x=x_0} (x-x_0)^1 \\
&+ \frac{1}{2!} \frac{d}{dx^2} \sqrt{b^2 \left(R - \frac{x^2}{a^2} \right)} \Big|_{x=x_0} (x-x_0)^2 \\
&= \sqrt{b^2 \left(R - \frac{x_0^2}{a^2} \right)} (x-x_0)^0 + \frac{1}{1} \frac{d}{dx} \sqrt{b^2 \left(R - \frac{x^2}{a^2} \right)} \Big|_{x=x_0} (x-x_0)^1 + \frac{1}{2} \frac{d}{dx^2} \sqrt{b^2 \left(R - \frac{x^2}{a^2} \right)} \Big|_{x=x_0} (x-x_0)^2 \\
&= \sqrt{b^2 \left(R - \frac{x_0^2}{a^2} \right)} (x-x_0)^0 + \frac{-\frac{b^2 x}{a^2}}{\sqrt{b^2 \left(R - \frac{x^2}{a^2} \right)} \Big|_{x=x_0}} (x-x_0)^1 + \frac{1}{2} \frac{d}{dx^2} \sqrt{b^2 \left(R - \frac{x^2}{a^2} \right)} \Big|_{x=x_0} (x-x_0)^2 \\
&= \sqrt{b^2 \left(R - \frac{x_0^2}{a^2} \right)} (x-x_0)^0 + \frac{-\frac{b^2 x}{a^2}}{\sqrt{b^2 \left(R - \frac{x^2}{a^2} \right)} \Big|_{x=x_0}} (x-x_0)^1 + \frac{1}{2} \frac{d}{dx} \frac{-\frac{b^2 x}{a^2}}{\sqrt{b^2 \left(R - \frac{x^2}{a^2} \right)} \Big|_{x=x_0}} (x-x_0)^2 \\
&= \sqrt{b^2 \left(R - \frac{x_0^2}{a^2} \right)} (x-x_0)^0 + \frac{-\frac{2b^2 x_0}{a^2}}{\sqrt{b^2 \left(R - \frac{x_0^2}{a^2} \right)}} (x-x_0)^1 + \frac{1}{2} \left(\frac{-\frac{2b^2}{a^2}}{\sqrt{b^2 \left(R - \frac{x^2}{a^2} \right)} \Big|_{x=x_0}} + -\frac{2b^2}{a^2} x \dots \right) (x-x_0)^2 \\
&= \sqrt{b^2 \left(R - \frac{x_0^2}{a^2} \right)} (x-x_0)^0 + \frac{-\frac{2b^2 x_0}{a^2}}{\sqrt{b^2 \left(R - \frac{x_0^2}{a^2} \right)}} (x-x_0)^1 + \frac{1}{2} \left(\frac{-\frac{2b^2}{a^2}}{\sqrt{b^2 \left(R - \frac{x_0^2}{a^2} \right)}} + -\frac{2b^2}{a^2} x_0 \dots \right) (x-x_0)^2
\end{aligned}$$

(25.20)

En prenant $x_0 = 0$:

$$\begin{aligned}
 y &\cong \sqrt{b^2 R} + 0 \cdot x + \frac{1}{2} \left(\frac{-\frac{b^2}{a^2}}{\sqrt{b^2 \left(R - \frac{0^2}{a^2} \right)}} - \frac{2b^2}{a^2} \cdot 0 \dots \right) x^2 = \sqrt{b^2 R} + \frac{1}{2} \frac{-\frac{b^2}{a^2}}{\sqrt{b^2 R}} x^2 \\
 &= \sqrt{b^2 R} - \frac{1}{2} \frac{b^2}{a^2 \sqrt{b^2 R}} x^2 = \sqrt{b^2 R} - \frac{1}{2} \frac{\sqrt{b^2 R}}{\sqrt{b^2 R}} \frac{b^2}{a^2 \sqrt{b^2 R}} x^2 = \sqrt{b^2 R} - \frac{1}{2} \frac{\sqrt{b^2 R}}{a^2 R} x^2
 \end{aligned} \quad (25.21)$$

Il vient alors immédiatement:

$$dy = -\frac{\sqrt{b^2 R}}{a^2 R} x dx \quad (25.22)$$

Dès lors:

$$ds^2 = dx^2 + dy^2 = dx^2 + \left(-\frac{\sqrt{b^2 R}}{a^2 R} x dx \right)^2 = \left(1 + \frac{b^2}{a^4 R} x^2 \right) dx^2 \quad (25.23)$$

Nous voyons alors que l'abscisse curviligne du cercle dans le plan devient celle d'une parabole lorsque nous faisons un développement en série de Maclaurin de l'équation du cercle au pôle.

Nous pourrions faire de même avec une hyperbole ou une ellipse et nous retrouver avec la même forme d'abscisse curviligne différentielle, généralement notée par tradition:

$$ds^2 = dx^2 + dy^2 \cong \left(1 + k_x^2 x^2 \right) dx^2 \quad (25.24)$$

où k_x est appelé "**paramètre de la parabole osculatrice**".

Ces exemples étant clos, continuons un peu avec la théorie. Nous pouvons bien évidemment réécrire notre abscisse curviligne différentielle en divisant par dt des deux côtés de l'égalité tel que:

$$\left(\frac{ds}{dt} \right)^2 = \left(\frac{dx}{dt} \right)^2 + \left(\frac{dy}{dt} \right)^2 + \left(\frac{dz}{dt} \right)^2 \quad (25.25)$$

Exemple:

Voyons une application de l'abscisse curviligne différentielle paramétrée avec une hélice (les exemples sont jolis en géométrie différentielle et valent donc la peine d'être vus...) qui est un exemple typique de courbe gauche:

Soient $(t, r, h) \in \mathbb{R}^3, t > 0, r > 0, h > 0$ et la fonction:

$$f \left(\begin{array}{l} [0, t] \rightarrow \mathbb{R}^3 \\ t \rightarrow f(t) = M \end{array} \right) \quad (25.26)$$

avec $M(x, y, z)$ et les coordonnées paramétriques:

$$\begin{cases} x = r \cos(t) \\ y = r \sin(t) \\ z = ht \end{cases} \quad (25.27)$$

Nous avons alors avec Maple 4.00b en prenant r et h comme étant égaux à l'unité:

```
>spacecurve([cos(t),sin(t),t,t=-4*Pi..4*Pi,numpoints=1000]);
```

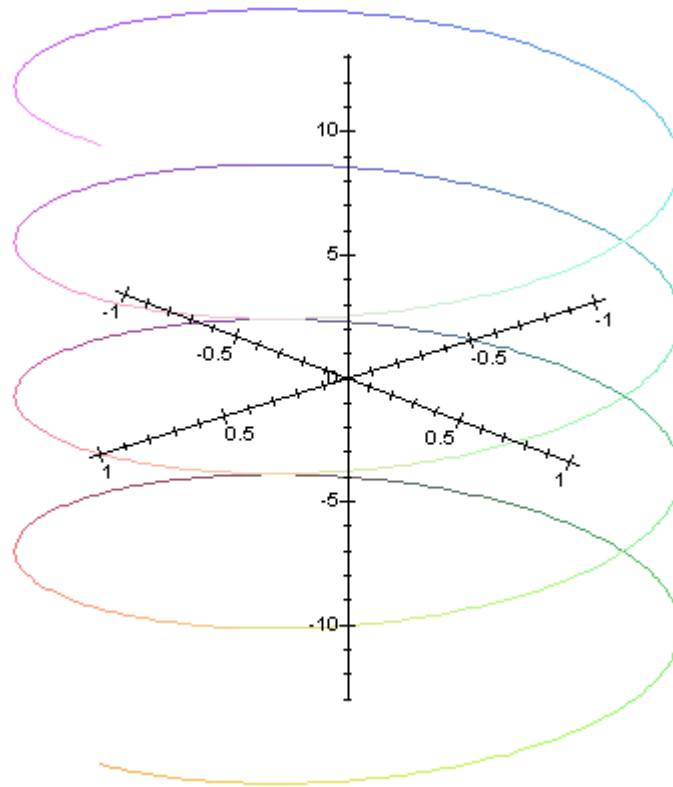


Figure: 25.1 - Représentation paramétrique d'une hélice avec Maple 4.00b

La fonction f est un arc paramétré dont le support est appelé une "**hélice**", r en est le rayon et h le pas. En prenant $t_0 = 0$ comme origine, l'abscisse curviligne de cette hélice (un morceau) est donnée par:

$$\left(\frac{ds}{dt} \right)^2 = \left(\frac{dx}{dt} \right)^2 + \left(\frac{dy}{dt} \right)^2 + \left(\frac{dz}{dt} \right)^2 = r^2 \sin^2(t) + r^2 \cos^2(t) + h^2 = r^2 + h^2 \quad (25.28)$$

Donc:

$$\frac{ds}{dt} = \sqrt{r^2 + h^2} \quad (25.29)$$

et alors:

$$s = \sqrt{r^2 + h^2} t \quad (25.30)$$

2. ISOCLINES

Voyons maintenant un point très important en mathématiques mais en plus dans l'ingénierie médicale, astrophysique, météorologie (parmi encore beaucoup d'autres domaines) que sont les isoclines.

Avant d'aborder le sujet sous forme mathématique, nous proposons au lecteur d'ouvrir Matlab 5.0.0.473 (nous avons déjà fait à peu près le même exemple avec Maple 4.00b dans le chapitre d'Analyse Fonctionnelle) et d'y écrire:

```
EDU» [xx,yy,z]=peaks;  
EDU» figure(1);mesh(xx,yy,z);title('peaks')
```

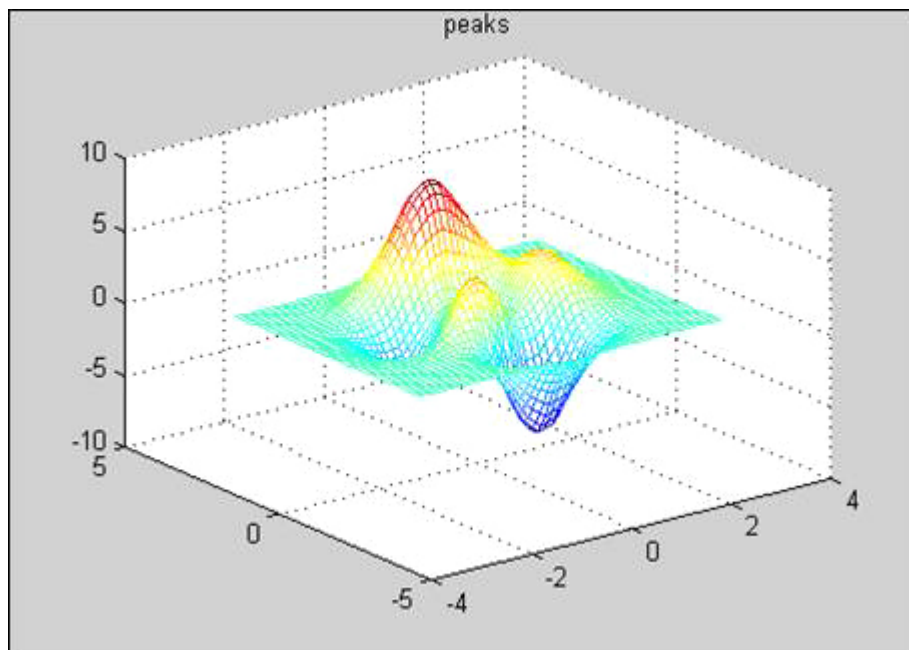


Figure: 25.2 - Tracé initial dans Matlab 5.0.0.473

ensuite pour des raisons esthétiques, d'écrire:

```
EDU» figure(2);surf(xx,yy,z);title('surf')
```

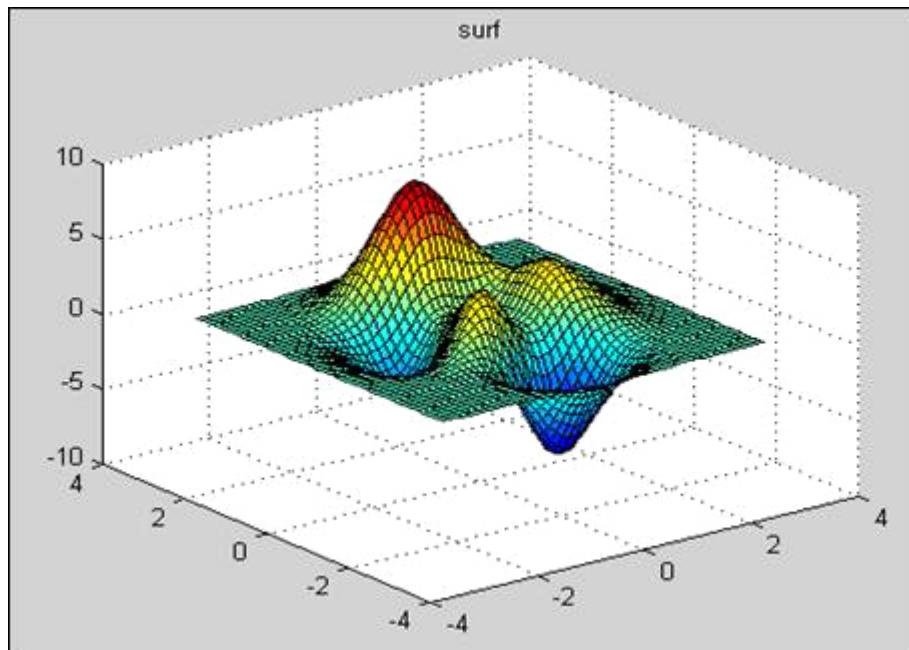


Figure: 25.3 - Amélioration du gradient de couleurs dans Matlab 5.0.0.473

Ensuite nous aimerions que Matlab nous trace quelques courbes de niveau (les points où la valeur de la fonction $f(x,y)$ est constante), appelées par les matheux des "isoclines" ou "courbes d'iso-niveau". Il faut alors écrire:

```
EDU» figure(3);contour3(xx,yy,z);title('contour')
```

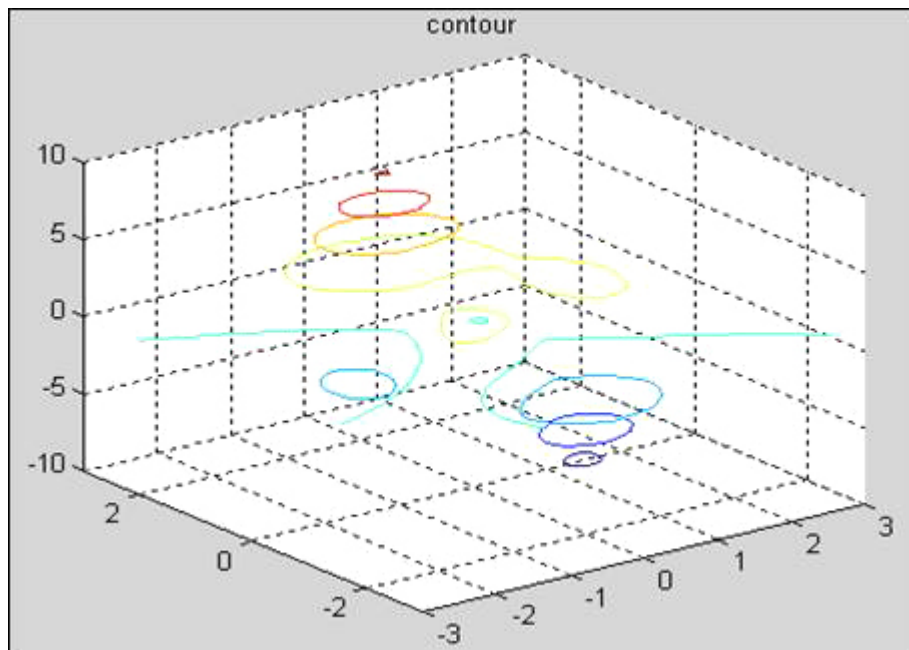


Figure: 17.4 - Affichage des isoclines de la fonction dans Matlab 5.0.0.473

Nous allons ensuite lui demander de les projeter sur le plan X,Y.

Ce qui donne:

```
EDU» figure(3);contour3(xx,yy,z);title('contour')
```

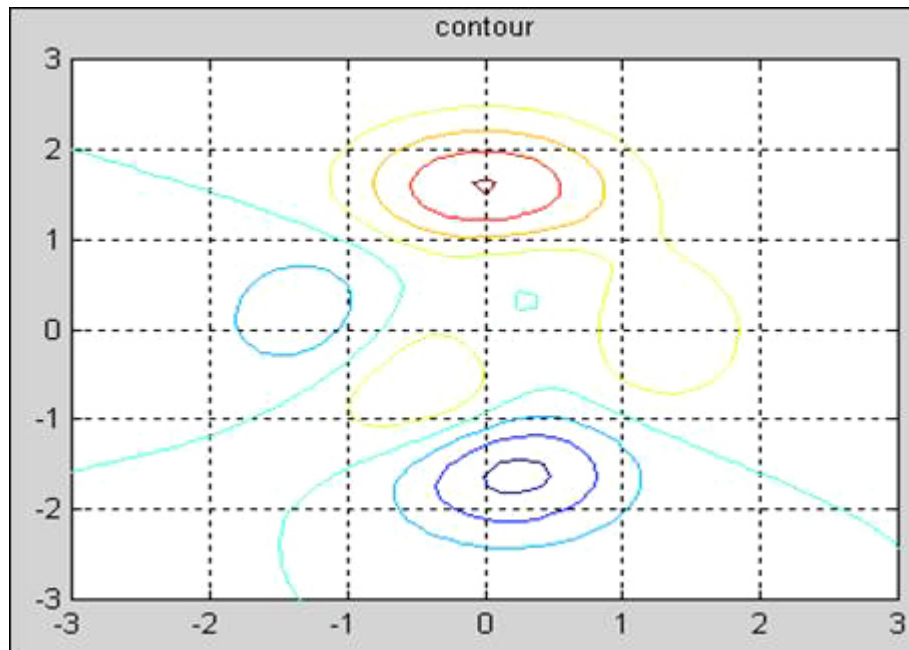


Figure: 25.5 - Projection des isoclines sur un plan dans Matlab 5.0.0.473

Et ce sont ces courbes qui vont nous intéresser. Nous souhaiterions déterminer les équations dans le plan de celles-ci sous forme explicite. Mais avant cela amusons-nous avec Matlab en écrivant encore:

EDU» `figure(4);pcolor(xx,yy,z);title('gradient')`

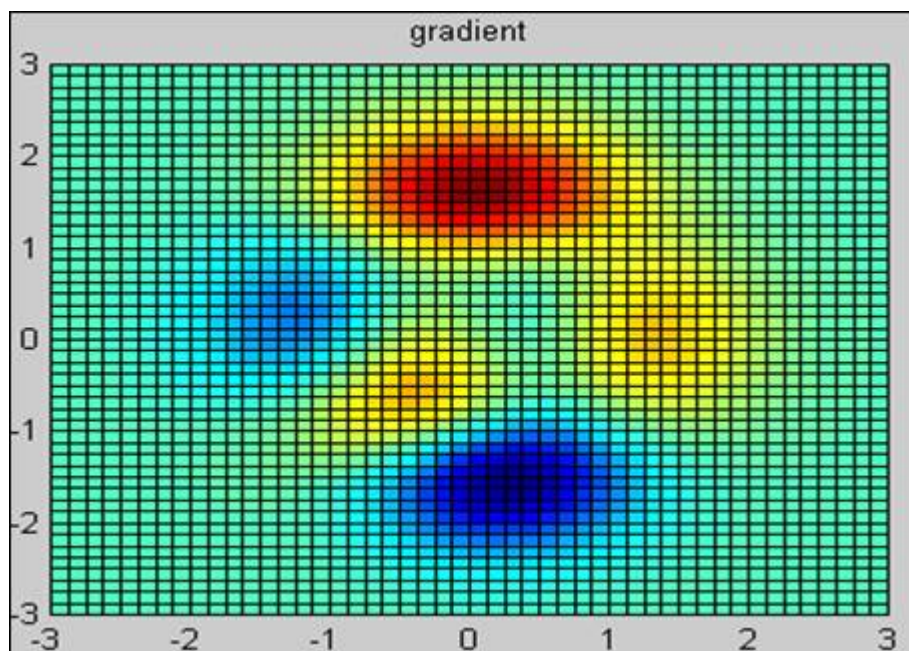


Figure: 25.6 - Représentation plane des isoclines avec gradients colorés dans Matlab 5.0.0.473

mais nous pouvons faire encore mieux en enlevant la grille avec la commande:

EDU» `shading interp`

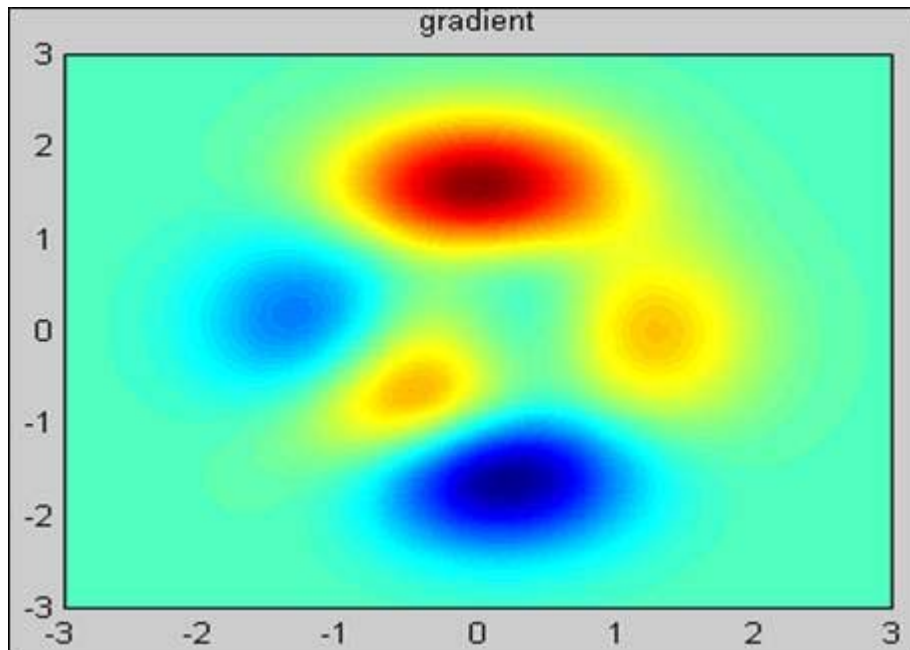


Figure: 25.7 - En retirant la grille...

Ensuite, sans fermer le graphique ci-dessus créé par Matlab 5.0.0.473 rajoutez maintenant la ligne:

EDU» hold on

EDU» contour(xx,yy,z,'k')

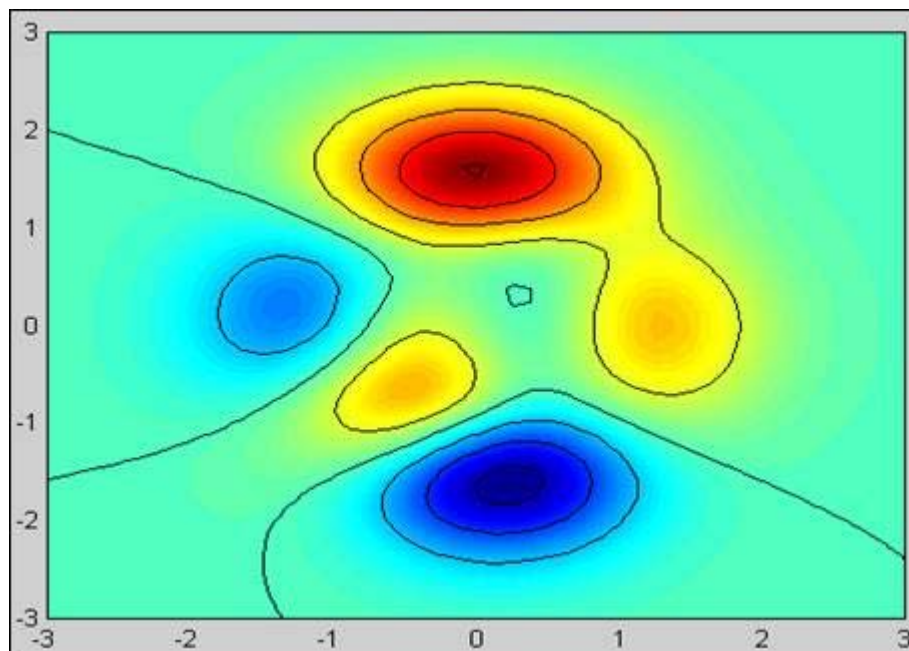


Figure: 25.8 - Association isoclines projetées avec gradients de couleurs

Considérons pour déterminer l'équation des isoclines la fonction $f(x,y)$ de $\mathbb{R}^3$ et que nous imposerons $\mathbb{R}^2$ -différentiable.

La relation:

$$f(x,y) = c^{te} \quad (25.31)$$

définit une courbe Γ plane appelée "isocline". C'est une courbe telle que, lorsque x varie, y ne varie donc pas n'importe comment mais précisément de telle sorte que f reste constante.

Nous avons vu dans le chapitre de Calcul Différentiel Et Intégral que la différentielle de f , pour des variations infinitésimales quelconques de x et y , est:

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy \quad (25.32)$$

Maintenant, si nous voulons que quand x varie de dx , la valeur de la fonction f ne change pas, il faut que dy ne soit pas n'importe quoi mais tel que la variation df soit nulle. Autrement dit:

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy = 0 \quad (25.33)$$

le long de Γ . Mais cette équation ne sert à rien en tant que telle, mais elle nous fixe le rapport de la dérivée de l'isocline dans le plan tel que:

$$\frac{dy}{dx} = - \frac{\frac{\partial f}{\partial x}}{\frac{\partial f}{\partial y}} \quad (25.34)$$

ce qui nous donne la pente de la tangente à Γ et donc après par intégration, la fonction recherchée elle-même!

Il va de soi que le vecteur tangent à la courbe Γ est donc un vecteur parallèle à celui ayant pour composantes (par correspondance avec la relation précédente):

$$(dx, dy) = \left(\frac{\partial f}{\partial y}, -\frac{\partial f}{\partial x} \right) \quad (25.35)$$

que nous noterons:

$$\vec{t} = \left(\frac{\partial f}{\partial y}, -\frac{\partial f}{\partial x} \right) \quad (25.36)$$

De plus rappelons que le gradient est donné par (cf. chapitre de Calcul Vectoriel):

$$\vec{\nabla}(f) = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right) \quad (25.37)$$

Nous remarquerons que ces deux derniers vecteurs sont perpendiculaires (résultat qui nous sera utile dans le chapitre d'Analyse Complexe). Effectivement:

$$\vec{t} \circ \vec{\nabla}(f) = \left(\frac{\partial f}{\partial y}, -\frac{\partial f}{\partial x} \right) \circ \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right) = \frac{\partial f}{\partial x} \frac{\partial f}{\partial y} + \left(-\frac{\partial f}{\partial x} \right) \frac{\partial f}{\partial y} = 0 \quad (25.38)$$

En d'autres termes, le vecteur $\vec{\nabla}(f)$ définit les lignes orthogonales à la courbe Γ .

Exemple:

Prenons l'équation d'une parabole particulière dans $\mathbb{R}^3$:

$$f(z) = \sqrt{x^2 + y^2} \quad (25.39)$$

Nous avons donc les isoclines qui sont données par:

$$f(z) = \sqrt{x^2 + y^2} = c^{te} \quad (25.40)$$

d'où leur équation dans le plan:

$$y = \sqrt{c^{te} - x^2} \quad (25.41)$$

Soient des cercles dans le plan dont le rayon est égal à la racine carrée de la constante choisie correspondante à la hauteur z de la fonction f !

Calculons maintenant la pente de la tangente à Γ :

$$\frac{dy}{dx} = -\frac{\frac{\partial f}{\partial x}}{\frac{\partial f}{\partial y}} = -\frac{\frac{x}{\sqrt{x^2 + y^2}}}{\frac{y}{\sqrt{x^2 + y^2}}} = -\frac{x}{y} = -\frac{x}{\sqrt{c^{te} - x^2}} \quad (25.42)$$

ce qui est conforme à la simple dérivée de:

$$\frac{dy}{dx} = \frac{d}{dx} \sqrt{c^{te} - x^2} = -\frac{x}{\sqrt{c^{te} - x^2}} \quad (25.43)$$

Nous avons aussi:

$$\vec{t} = \left(\frac{y}{\sqrt{x^2 + y^2}}, -\frac{x}{\sqrt{x^2 + y^2}} \right) \quad (25.44)$$

Nous voyons qu'en $x = 0$ ce vecteur est vaut:

$$\vec{t} = (1, 0) \quad (25.45)$$

ce qui est bien conforme au vecteur tangent que nous avons au cercle en ce point de l'axe des abscisses.

3. TRIEDRE DE FRENET

Le repère de Frenet est un outil d'étude du comportement local des courbes. Plus exactement, il s'agit d'un repère local associé à un point décrivant une courbe Γ . Son mode de construction est différent selon que l'espace ambiant est de dimension 2 (courbe plane) ou 3 (courbe gauche).

Le repère de Frenet, et les formules de Frenet (donnant les dérivées des vecteurs de ce repère), permettent de mener de façon systématique des calculs de courbure, de torsion pour les courbes gauches et d'introduire des concepts géométriques intéressants associés aux courbes.

Considérons pour commencer une courbe Γ avec son abscisse curviligne $s(t)$ et $M_0 = s(t_0)$ son origine. Nous notons par définition:

$$\vec{T} = \frac{d\vec{f}(t)}{ds} = \frac{\vec{f}(t+dt) - \vec{f}(t)}{ds} = \frac{d\vec{OM}}{ds} \quad (25.46)$$

la tangente à la courbe Γ de paramètre t au voisinage d'un point M par rapport à un repère posé en O avec ds qui se calcule comme nous l'avons montré précédemment.

Il est intéressant de remarquer que si t s'interprète comme le temps, alors nous avons une vitesse:

$$\vec{v} = \vec{T} \frac{ds}{dt} = \vec{T} \|\vec{v}\| = \vec{T} \cdot v = \frac{d\vec{f}(t)}{dt} \quad (25.47)$$

et donc le vecteur $\vec{T}$ est dirigé dans le sens du mouvement.

De plus, par construction et définition de l'abscisse curviligne, nous avons toujours:

$$\|\vec{T}\| = \frac{\|d\vec{OM}\|}{|ds|} = 1 \quad (25.48)$$

et donc le vecteur tangent $\vec{T}$ au point M est unitaire (et non nul!).

Maintenant, sans savoir exactement à quoi cela va nous servir pour l'instant, intéressons-nous au vecteur:

$$\frac{d\vec{T}}{ds} \quad (25.49)$$

Sachant trivialement de ce qui précède que:

$$\|\vec{T}\| = \vec{T} \circ \vec{T} = 1 \quad (25.50)$$

Alors, nous avons:

$$\frac{d(\vec{T} \circ \vec{T})}{ds} = \frac{d\vec{T}}{ds} \circ \vec{T} + \vec{T} \circ \frac{d\vec{T}}{ds} = 2\vec{T} \circ \frac{d\vec{T}}{ds} = 0 \quad (25.51)$$

donc déjà $\frac{d\vec{T}}{ds}$ n'est a priori pas unitaire et $\vec{T}$ lui est perpendiculaire (résultat qui va nous servir plusieurs fois par la suite donc il faudra s'en souvenir)!

Posons maintenant:

$$\vec{N} = \frac{\frac{d\vec{T}}{ds}}{\left\| \frac{d\vec{T}}{ds} \right\|} = \frac{\frac{d\vec{T}}{ds}}{C} \quad (25.52)$$

Étant donné le résultat précédent, $\vec{N}$ est le vecteur perpendiculaire, appelé "**vecteur de courbure**" unitaire, à $\vec{T}$ en M (nous disons que ce couple de vecteur est "orthonormal direct") et C est par définition la "**courbure**".

Nous pouvons également aborder la courbure C d'une façon plus géométrique plutôt que par une définition tombée du ciel:

Nous savons à ce point de notre discours qu'en un point M_0 d'une courbe (dérivable au moins une fois en tout point...), il existe un vecteur tangent non nul qui est $\vec{T}$.

En tout point voisin M (d'abscisse curviligne s), le vecteur tangent peut s'écrire en approximation:

$$\vec{T}(s) \cong \vec{T} + \frac{d\vec{T}}{ds} ds \quad (25.53)$$

si la courbe se trouve localement dans un même plan (car nous étudions ici la courbure et non la torsion de la courbe)!

Deux normales en M et M_0 se coupant donc en un point Ω , la figure suivante:

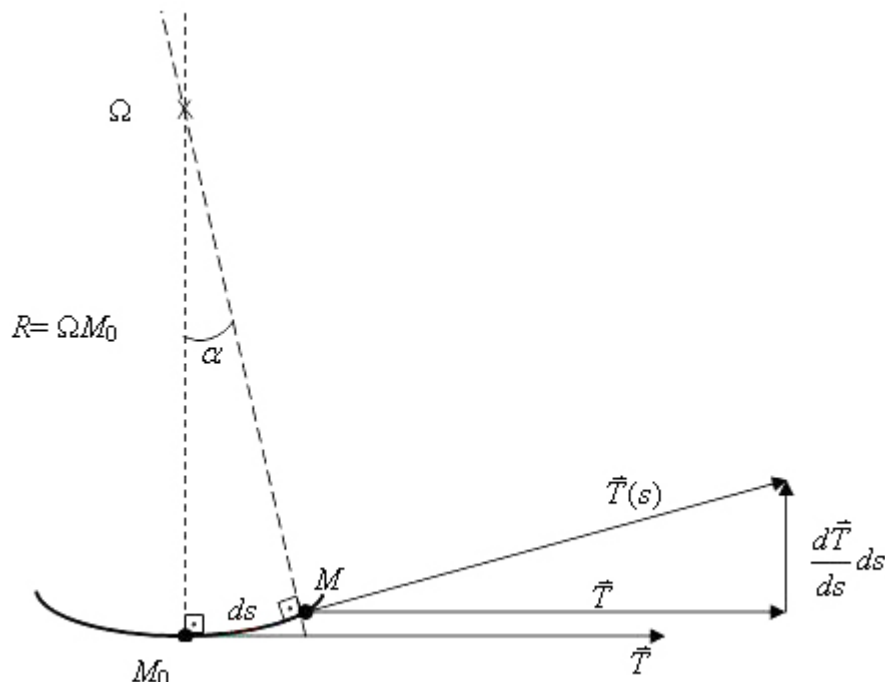


Figure: 25.9 - Décomposition du vecteur tangent de la trajectoire courbe

montre qu'au premier ordre en ds , le point M peut être considéré localement comme déduit du point M_0 par une rotation de centre Ω .

Le cercle ainsi défini, de rayon:

$$\Omega M_0 = R \quad (25.54)$$

est celui qui tangente le mieux la courbe localement au point M_0 . Son rayon se déduit de la figure (deux triangles semblables à la limite):

$$\frac{ds}{R} = \frac{\left\| \frac{d\vec{T}}{ds} \right\| ds}{\left\| \vec{T} \right\|} \quad (25.55)$$

d'où, puisque $\vec{T}$ est unitaire, la définition et la valeur de la courbure:

$$C = \frac{1}{R} = \left\| \frac{d\vec{T}}{ds} \right\| \quad (25.56)$$

et voilà!

Il est possible d'interpréter le concept de courbure comme la vitesse de rotation de la base de Frenet par rapport à une direction fixe.

Le couple de vecteurs $(\vec{T}, \vec{N})$ est appelé "**repère de Frenet**" et ses vecteurs de base les "**vecteurs de Frenet**".

Le repère de Frenet est un repère mobile puisque les éléments de ce repère changent selon le point considéré. En physique, il ne faut pas confondre cette notion avec celle de référentiel: puisque les vecteurs de Frenet se déplacent avec le point!

Remarque: La définition de C telle que ci-dessus est vraie dans le cadre d'un choix d'une courbure positive. C'est un point de vue pris en mécanique mais non nécessaire en mathématiques.

Si $C \neq 0$, alors comme vu précédemment:

$$\vec{N} = \frac{\frac{d\vec{T}}{ds}}{C} = \left(\frac{1}{C} \right) \frac{d\vec{T}}{ds} = R \frac{d\vec{T}}{ds} \quad (25.57)$$

où R est appelé le "**rayon de courbure**".

Quant à la relation:

$$\boxed{\frac{\vec{N}}{R} = \frac{d\vec{T}}{ds}} \quad (25.58)$$

elle est appelée "**1ère formule de Frenet**" et montre que $\frac{d\vec{T}}{ds}$ et $\vec{N}$ sont colinéaires et donc leur produit vectoriel est nul (résultat utilisé plus loin).

Ces relations se justifient par l'analogie avec la mécanique. Effectivement, nous avons démontré plus haut que:

$$\vec{v} = \vec{T} \frac{ds}{dt} = \vec{T} \|\vec{v}\| = \vec{T} \cdot v = \frac{d\vec{f}(t)}{dt} \quad (25.59)$$

Calculons maintenant l'accélération:

$$\begin{aligned} \vec{a} &= \frac{d\vec{v}}{dt} = \frac{d(\vec{T} \cdot v)}{dt} = v \frac{d\vec{T}}{dt} + \vec{T} \frac{dv}{dt} = v \frac{ds}{dt} \frac{d\vec{T}}{ds} + \vec{T} \frac{dv}{dt} = v^2 \frac{d\vec{T}}{ds} + \vec{T} \frac{dv}{dt} \\ &= v^2 \cdot C \cdot \vec{N} + \vec{T} \frac{dv}{dt} = \vec{T} \frac{dv}{dt} + \frac{v^2}{R} \vec{N} = \vec{a}_{tg} + \vec{a}_n \end{aligned} \quad (25.60)$$

nous retrouvons alors le résultat obtenu dans le chapitre de Mécanique Classique lors de notre étude du plan osculateur.

Pour donner une interprétation géométrique plus exacte de la courbure, nous définissons d'abord par Ω le centre du "**cercle osculateur**" (se trouvant dans le plan osculateur) ou "**cercle de courbure**" de rayon R qui tangente le mieux localement Γ tel que dans le repère de Frenet:

$$\vec{\Omega} = \vec{M} + R\vec{N} \quad (25.61)$$

Pour préciser géométriquement ce qu'est le cercle osculateur, prenez une courbe, et un point M sur cette courbe. Tracez ensuite la normale au point de cette courbe localement plane et prenez un point Ω sur la normale. Alors, le cercle de centre O passant par le point M est tangent à la courbe. Mais tous les cercles tangents à la courbe ne sont pas tangents de la même façon! En effet, si Ω est loin de M , le cercle va se situer plutôt à l'extérieur de la courbe (cercle bleu dans la figure ci-dessous). Si Ω est proche de M , le cercle va se situer plutôt à l'intérieur de la courbe (cercle rose dans la figure ci-dessous). Le rayon limite entre être "à l'intérieur de la courbe" et être "à l'extérieur de la courbe" est par convention le "rayon de courbure" que nous avons défini plus haut. Le cercle correspondant à ce rayon est alors le fameux "cercle osculateur".

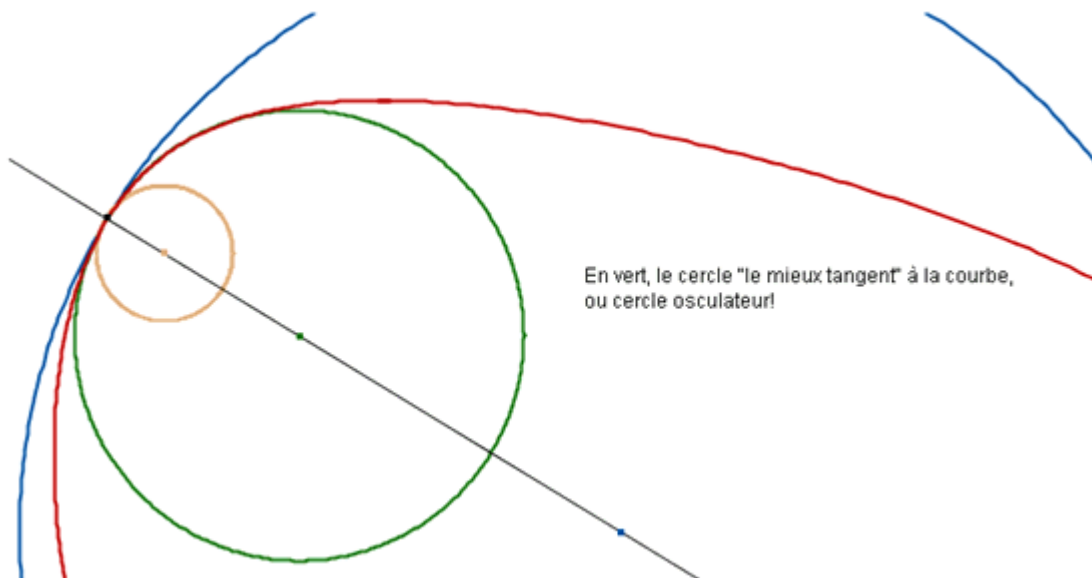


Figure: 25.10 - Représentation du cercle osculateur

Dans le cas particulier où $\vec{T}$ est un vecteur constant:

$$\frac{d\vec{T}}{ds} = \vec{0} \quad (25.62)$$

et donc $C = 0$ ce qui implique que R n'est plus défini. Nous disons quelquefois dans ce cas que le rayon de courbure de Γ est infini (une droite présente alors une courbure nulle en tout point).

Étudions maintenant le vecteur perpendiculaire au plan osculateur défini par:

$$\vec{B} = \vec{T} \times \vec{N} \quad (25.63)$$

Nous pouvons déjà dire, étant donné que $\vec{T}$ et $\vec{N}$ sont unitaires que $\vec{B}$ l'est aussi (ce qui va nous servir plus loin)!

Démontrons que $\frac{d\vec{B}}{ds}$ est orthogonal à $\vec{T}$:

$$\frac{d\vec{B}}{ds} = \frac{d\vec{T}}{ds} \times \vec{N} + \vec{T} \times \frac{d\vec{N}}{ds} = \vec{0} \times \vec{N} + \vec{T} \times \frac{d\vec{N}}{ds} = \vec{T} \times \frac{d\vec{N}}{ds} \quad (25.64)$$

où nous avons pris le cas particulier $\frac{d\vec{T}}{ds} = \vec{0}$ (mais de toute manière en généralité $\frac{d\vec{T}}{ds}$ et $\vec{N}$ sont colinéaires comme nous l'avons démontré donc le produit vectoriel entre ces deux vecteurs est toujours nul).

□C.Q.F.D.

Démontrons maintenant que $\frac{d\vec{B}}{ds}$ est colinéaire à $\vec{N}$:

De la même manière que nous avons démontré plus haut que $\frac{d\vec{T}}{ds}$ est perpendiculaire à $\vec{T}$, nous démontrons que $\frac{d\vec{B}}{ds}$ est perpendiculaire à $\vec{B}$!

Nous avons donc:

$$\vec{B} \circ \frac{d\vec{B}}{ds} = 0 \quad (25.65)$$

Et étant donné que $\frac{d\vec{B}}{ds}$ est aussi perpendiculaire à $\vec{T}$ (démonstration précédente) il est donc colinéaire à $\vec{N}$.

□C.Q.F.D.

Posons maintenant:

$$\boxed{\frac{d\vec{B}}{ds} = \frac{\vec{N}}{R'} = \tau \cdot \vec{N}} \quad (25.66)$$

Cette relation constitue la "**2^{ème} formule de Frenet**" où par définition, $\vec{B}$ est le "**vecteur binormal**" de Γ au point M et τ en est la "**torsion**" et R' le "**rayon de torsion**".

Nous pouvons maintenant établir la "**3^{ème} formule de Frenet**":

$$\boxed{\vec{B} = \vec{T} \times \vec{N} \Leftrightarrow \vec{N} = \vec{B} \times \vec{T}} \quad (25.67)$$

d'où nous tirons:

$$\frac{d\vec{N}}{ds} = \frac{d\vec{B}}{ds} \times \vec{T} + \vec{B} \times \frac{d\vec{T}}{ds} = \frac{\vec{N}}{R'} \times \vec{T} + \vec{B} \times \frac{\vec{N}}{R} \quad (25.68)$$

Or de par les propriétés du produit vectoriel:

$$\vec{N} \times \vec{T} = -\vec{B} \quad \vec{B} \times \vec{N} = -\vec{T} \quad (25.69)$$

d'où la 3^{ème} formule de Frenet:

$$\boxed{\frac{d\vec{N}}{ds} = -\frac{\vec{B}}{R'} - \frac{\vec{T}}{R}} \quad (25.70)$$

Nous appelons "**trièdre de Frenet**" associé à Γ au point M , le repère naturel orthonormal de l'espace $(M, \vec{T}, \vec{N}, \vec{B})$:

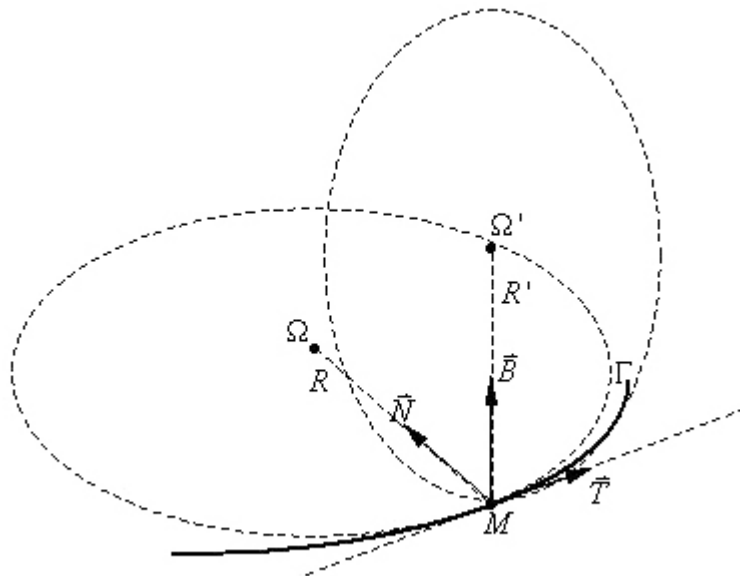


Figure: 25.11 - Représentation du trièdre de Frenet

où, en mécanique, le vecteur $\vec{T}$ est colinéaire à la vitesse et l'accélération tangentielle et $\vec{N}$ est colinéaire à l'accélération normale.

Remarque: Le rayon de courbure R est donc dans le plan osculateur (plan formé par le vecteur tangent et normal à la courbe) qui est le meilleur plan dans lequel est contenue la courbe. Du coup, le rayon de courbure donne en un point (localement) le meilleur ("le plus vrai") rayon de la courbe. La torsion nous donne par contre la tendance qu'a la courbe à sortir du plan osculateur (in extenso si la courbe est contenue dans un plan, la torsion est nulle).

Exemple:

E1. Cherchons le rayon et le centre de courbure en tout M à notre hélice définie plus haut comme exemple pratique. Rappelons que sa fonction paramétrique est donnée par:

$$\overrightarrow{OM} = \begin{pmatrix} x = r \cos(t) \\ y = r \sin(t) \\ z = ht \end{pmatrix} \quad (25.71)$$

et que:

$$\frac{ds}{dt} = \sqrt{r^2 + h^2} \quad (25.72)$$

Nous avons dès lors:

$$\vec{T} = \frac{d\overrightarrow{OM}}{ds} = \frac{d\overrightarrow{OM}}{dt} \frac{dt}{ds} \Rightarrow \vec{T} = \begin{pmatrix} -\frac{r}{\sqrt{r^2 + h^2}} \sin(t) \\ \frac{r}{\sqrt{r^2 + h^2}} \cos(t) \\ \frac{h}{\sqrt{r^2 + h^2}} \end{pmatrix} \quad (25.73)$$

Au passage, vous remarquerez que nous avons bien:

$$\|\vec{T}\| = 1 \quad (25.74)$$

Ainsi, la courbure (l'inverse du rayon de courbure) est donnée par:

$$\frac{d\vec{T}}{ds} = \frac{d\vec{T}}{dt} \frac{dt}{ds} = \begin{pmatrix} -\frac{r}{r^2 + h^2} \cos(t) \\ -\frac{r}{r^2 + h^2} \sin(t) \\ 0 \end{pmatrix} \Rightarrow \left\| \frac{d\vec{T}}{ds} \right\| = \frac{r}{r^2 + h^2} \quad (25.75)$$

Donc, le rayon de courbure vaut:

$$R = \frac{r^2 + h^2}{r} \quad (25.76)$$

Ce qui est conforme à l'intuition puisque, lorsque le pas h de l'hélice est nul, le rayon de courbure vaut r et lorsque le pas h tend vers l'infini le rayon de courbure tend vers l'infini aussi et la courbure vers zéro. Cet exemple est un cas fameux d'ingénierie appliquée aux cheminées d'évacuation de fumées qui sont entourées d'une hélice d'Archimède et dont l'objectif est de faire monter les flux d'air vers le haut (la difficulté étant de déterminer le rayon R de la plaque de métal à couper qui suivra le mieux la courbure voulue... du moins localement en connaissant le rayon de la cheminée et la hauteur h du pas de la spirale):

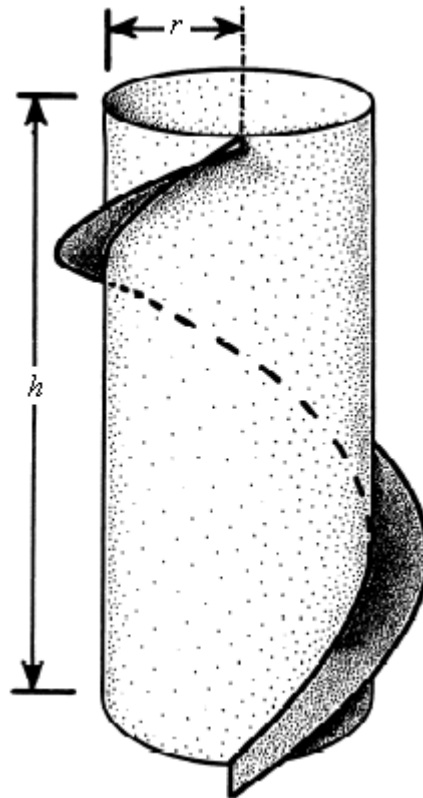


Figure: 25.12 - Principe de base d'une cheminée industrielle avec spirale
(source: Frank Morgan, Riemmanian Geometry)

ou sa version réelle:



Figure: 25.13 - Cheminée industrielle d'évacuation avec spirale

Mais dans ce cas d'ingénierie, la hauteur h doit être obtenue par une rotation complète. Dès lors le rayon de courbure s'écrit:

$$R = \frac{r^2 + \left(\frac{h}{2\pi}\right)^2}{r} \quad (25.77)$$

Bref pour en revenir à notre exemple et le finir, il vient par la première formule de Frenet le vecteur normal suivant:

$$\vec{N} = \begin{pmatrix} -\cos(t) \\ -\sin(t) \\ 0 \end{pmatrix} \quad (25.78)$$

et dont tous les points (extrémités du vecteur) sont confondus avec l'axe Z de notre hélice quel que soit h ! La coordonnée de la composante z de ce vecteur est nulle étant donné que la normale est prise par rapport à un point M de la courbe déjà à une hauteur h implicite.

De par la 3ème formule de Frenet avec le vecteur binormal:

$$\begin{aligned}
\vec{B} = \vec{T} \times \vec{N} &= \begin{pmatrix} -\frac{r}{\sqrt{r^2 + h^2}} \sin(t) \\ \frac{r}{\sqrt{r^2 + h^2}} \cos(t) \\ \frac{h}{\sqrt{r^2 + h^2}} \end{pmatrix} \times \begin{pmatrix} -\cos(t) \\ -\sin(t) \\ 0 \end{pmatrix} \\
&= \frac{1}{\sqrt{r^2 + h^2}} \begin{pmatrix} r \cos(t) \cdot 0 - h(-\sin(t)) \\ h(-\cos(t)) - (-r \sin(t)) \cdot 0 \\ (-r \sin(t))(-\sin(t)) - r \cos(t)(-\cos(t)) \end{pmatrix} \quad (25.79) \\
&= \frac{1}{\sqrt{r^2 + h^2}} \begin{pmatrix} h \sin(t) \\ -h \cos(t) \\ r \end{pmatrix}
\end{aligned}$$

et le rayon de torsion donné par la relation:

$$\frac{d\vec{B}}{ds} = \frac{\vec{N}}{R'} \quad (25.80)$$

nous avons donc:

$$\vec{N} = R' \frac{d\vec{B}}{ds} = R' \frac{d\vec{B}}{dt} \frac{dt}{ds} \quad (25.81)$$

Et comme nous avons obtenu les trois relations suivantes:

$$\vec{N} = \begin{pmatrix} -\cos(t) \\ -\sin(t) \\ 0 \end{pmatrix} \quad \text{et} \quad \frac{d\vec{B}}{dt} = \begin{pmatrix} \frac{h}{\sqrt{r^2 + h^2}} \cos(t) \\ \frac{h}{\sqrt{r^2 + h^2}} \sin(t) \\ 0 \end{pmatrix} \quad \text{et} \quad \frac{dt}{ds} = \frac{1}{\sqrt{r^2 + h^2}} \quad (25.82)$$

Nous en déduisons le rayon de torsion:

$$R' = -\frac{r^2 + h^2}{h} \quad (25.83)$$

E2. Déterminons maintenant le cas important de l'expression explicite du rayon de courbure en coordonnées cartésiennes (résultat utilisé dans le chapitre de Génie Civil et utile dans de nombreux autres domaines de la physique). Considérons pour cela la figure suivante:

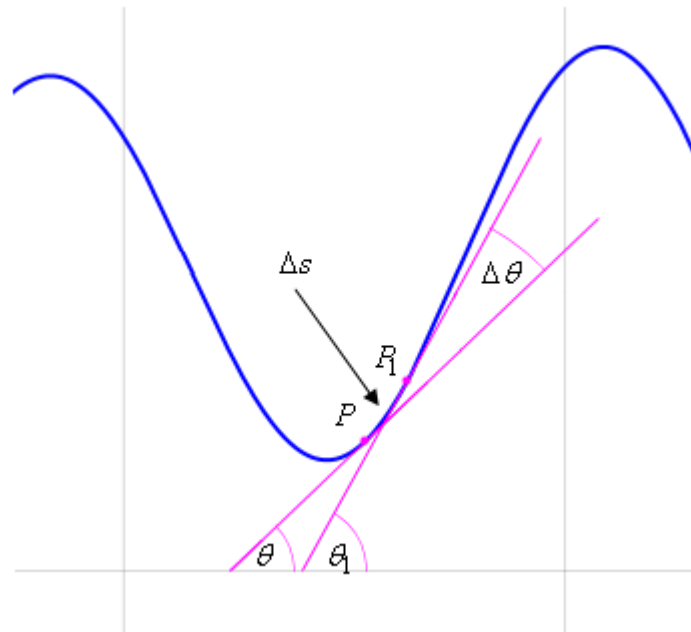


Figure: 25.14 - Illustration de l'approche du rayon de courbure en 1D

Nous avons donc le rayon de courbure qui est donné intuitivement par la relation suivante si nous ne faisons pas usage de l'analyse vectorielle:

$$C = \lim_{\Delta\theta \rightarrow 0} \frac{\Delta\theta}{\Delta s} = \frac{d\theta}{ds} \quad (25.84)$$

Nous avons aussi:

$$\frac{d\theta}{ds} = \frac{d\theta}{dx} \frac{dx}{ds} \quad (25.85)$$

et comme:

$$\tan \theta = \frac{dy}{dx} \quad (25.86)$$

il vient alors:

$$\theta = \arctan\left(\frac{dy}{dx}\right) \quad (25.87)$$

et donc:

$$\frac{d\theta}{dx} = \frac{d}{dx} \left(\arctan\left(\frac{dy}{dx}\right) \right) \quad (25.88)$$

Nous avons démontré dans le chapitre de Calcul Intégral Et Différentiel la dérivée suivante:

$$\frac{d}{dx} \arctan(x) = \frac{1}{1+x^2} \quad (25.89)$$

Nous avons alors immédiatement par les dérivées composées:

$$\frac{d\theta}{dx} = \frac{d}{dx} \left(\arctan \left(\frac{dy}{dx} \right) \right) = \frac{1}{1 + \left(\frac{dy}{dx} \right)^2} \frac{d}{dx} \left(\frac{dy}{dx} \right) = \frac{\frac{d^2 y}{dx^2}}{1 + \left(\frac{dy}{dx} \right)^2} \quad (25.90)$$

Ceci étant fait nous avons aussi besoin de dx/ds . Or, nous avons démontré avec différentes approches dans le chapitre de Mécanique Analytique et Formes Géométriques (entre autres) en utilisant simplement Pythagore que:

$$\frac{dx}{ds} = \frac{1}{\frac{ds}{dx}} = \frac{1}{\sqrt{1 + \left(\frac{dy}{dx} \right)^2}} \quad (25.91)$$

En regroupant le tout nous avons finalement:

$$C = \frac{d\theta}{ds} = \frac{d\theta}{dx} \frac{dx}{ds} = \frac{\frac{d^2 y}{dx^2}}{1 + \left(\frac{dy}{dx} \right)^2} \frac{1}{\sqrt{1 + \left(\frac{dy}{dx} \right)^2}} = \frac{\frac{d^2 y}{dx^2}}{\left(1 + \left(\frac{dy}{dx} \right)^2 \right)^{3/2}} = \frac{y''}{(1+y'^2)^{3/2}} \quad (25.92)$$

Il vient donc alors le rayon de courbure du cercle localement osculateur d'une fonction cartésienne dans le plan (en prenant la valeur absolue de la dérivée seconde pour éviter d'avoir un rayon négatif...):

$$R = \frac{1}{C} = \frac{\left(1 + \left(\frac{dy}{dx} \right)^2 \right)^{3/2}}{\left| \frac{d^2 y}{dx^2} \right|} = \frac{(1+y'^2)^{3/2}}{|y''|} \quad (25.93)$$

4. NAPPES PARAMÉTRÉES

Soient $D \subset \mathbb{R}^2$:

$$g \left(\begin{matrix} D \rightarrow \mathbb{R}^3 \\ (u,v) \rightarrow g(u,v) = M \end{matrix} \right) \text{ avec } \overrightarrow{OM} = \begin{pmatrix} x(u,v) \\ y(u,v) \\ z(u,v) \end{pmatrix} \quad (25.94)$$

Appelons $\Sigma = g(D)$. Si g est continue, alors Σ est une surface de l'espace "**surface d'un seul tenant**". Par définition, dans ce qui suit, le couple (g, D) où g est une fonction supposée continue sera appelé "**nappe paramétrée**", et Σ le "**support**" de la nappe paramétrée. Nous disons encore que (g, D) et (u, v) sont des paramétrages de Σ .

Remarquons que pour une surface Σ (par exemple un disque), il existe plusieurs nappes paramétrées associées (par exemple les coordonnées cartésiennes, polaires, sphériques).

Soit maintenant $I \subset \mathbb{R}$ et:

$$h \begin{pmatrix} I \rightarrow \mathbb{R}^2 \\ t \rightarrow h(t) = (u, v) \end{pmatrix} \quad (25.95)$$

tels que $h(I) \subset D$

Nous pouvons définir $g \circ h$:

$$g \circ h \begin{pmatrix} I \rightarrow \mathbb{R}^3 \\ t \rightarrow (g \circ h)(t) = g(u(t), v(t)) \end{pmatrix} \quad (25.96)$$

Si nous supposons h continue, il est clair que $g \circ h$ est un arc paramétré. Appelons Γ son support, nous avons $\Gamma \subset \Sigma$ et nous disons que Γ est une "**courbe tracée**" ou "**courbe inscrite**" sur Σ définie par les "**coordonnées de Gauss**" u et v (déjà rencontrées dans le chapitre sur les Géométries Non-Euclidiennes).

Remarque: Nous supposons toujours désormais que $D = I \times J$

Soit $M_0 \in \Sigma, M_0 = g(u_0, v_0)$. Intéressons-nous aux deux courbes tracées sur Σ définies par les arcs paramétrés suivants:

$$\begin{aligned} & (g_{v_0}, I) \text{ avec } g_{v_0} \begin{pmatrix} I \rightarrow \mathbb{R}^3 \\ u \rightarrow g_{v_0} = g(u, v_0) \end{pmatrix} \\ & (g_{u_0}, J) \text{ avec } g_{u_0} \begin{pmatrix} J \rightarrow \mathbb{R}^3 \\ v \rightarrow g_{u_0} = g(u_0, v) \end{pmatrix} \end{aligned} \quad (25.97)$$

g_{u_0} et g_{v_0} sont les deux fonctions dites "**fonctions partielles**" de g en (u_0, v_0) .

Les supports de (g_{u_0}, J) et (g_{v_0}, I) sont appelés "**courbes-coordonnées**" de Σ en M_0 relativement au paramétrage (g, D) . Nous les notons respectivement Γ_{u_0} et Γ_{v_0} . Nous appelons aussi Γ_{u_0} "**1ère courbe-coordonnée**" et Γ_{v_0} "**2ème courbe-coordonnée**".

Il est bien sûr évident (cf. chapitre de Calcul Différentiel Et Intégral) que:

$$\frac{dg_{u_0}}{dv} = \frac{\partial g}{\partial v} = \frac{\partial OM}{\partial v} \quad (25.98)$$

est tangent à Γ_{u_0} en M_0 et que $\partial OM / \partial u$ est tangent à Γ_{v_0} en M_0 .

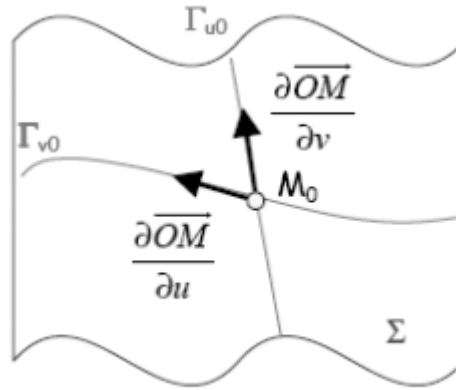


Figure: 25.15 - Représentation d'une nappe paramétrée

4.1. MÉTRIQUE D'UNE SURFACE

Soit:

$$\mathcal{E} \left(\begin{array}{l} D \rightarrow \mathbb{R}^3 \\ (u, v) \rightarrow g(u, v) = M \end{array} \right) \text{ avec } \overrightarrow{OM} = \begin{pmatrix} x(u, v) \\ y(u, v) \\ z(u, v) \end{pmatrix} \quad (25.99)$$

Notons $dg = (dx, dy, dz)$, autrement dit:

$$d\overrightarrow{OM} = \begin{pmatrix} dx \\ dy \\ dz \end{pmatrix} \quad (25.100)$$

Nous avons aussi (cf. chapitre de Calcul Différentiel Et Intégral):

$$\begin{aligned} dx &= \frac{\partial x}{\partial u} du + \frac{\partial x}{\partial v} dv \\ dy &= \frac{\partial y}{\partial u} du + \frac{\partial y}{\partial v} dv \\ dz &= \frac{\partial z}{\partial u} du + \frac{\partial z}{\partial v} dv \end{aligned} \quad (25.101)$$

et nous avons démontré au début de ce chapitre que l'abscisse curviligne dans un espace cartésien (en coordonnées de Riemann) était donnée par:

$$ds^2 = dx^2 + dy^2 + dz^2 \quad (25.102)$$

Nous avons donc après substitution en coordonnées de Gauss:

$$ds^2 = \left[\left(\frac{\partial x}{\partial u} \right)^2 + \left(\frac{\partial y}{\partial u} \right)^2 + \left(\frac{\partial z}{\partial u} \right)^2 \right] du^2 + 2 \left[\frac{\partial x}{\partial u} \frac{\partial x}{\partial v} + \frac{\partial y}{\partial u} \frac{\partial y}{\partial v} + \frac{\partial z}{\partial u} \frac{\partial z}{\partial v} \right] dudv + \left[\left(\frac{\partial x}{\partial v} \right)^2 + \left(\frac{\partial y}{\partial v} \right)^2 + \left(\frac{\partial z}{\partial v} \right)^2 \right] dv^2 \quad (25.103)$$

Ce qui est équivalent à écrire (attention à ne pas y lire qu'on a le carré d'un vecteur mais qu'il s'agit du produit scalaire avec lui-même!):

$$ds^2 = \left(\frac{\partial \overrightarrow{OM}}{\partial u} \right)^2 du^2 + 2 \frac{\partial \overrightarrow{OM}}{\partial u} \frac{\partial \overrightarrow{OM}}{\partial v} dudv + \left(\frac{\partial \overrightarrow{OM}}{\partial v} \right)^2 dv^2 \quad (25.104)$$

De manière plus traditionnelle avec la notation:

$$E = \left(\frac{\partial \overrightarrow{OM}}{\partial u} \right)^2 \quad F = \frac{\partial \overrightarrow{OM}}{\partial u} \frac{\partial \overrightarrow{OM}}{\partial v} \quad G = \left(\frac{\partial \overrightarrow{OM}}{\partial v} \right)^2 \quad (25.105)$$

Nous obtenons une relation appelée la "**première forme quadratique fondamentale**" (nous ne démontrerons pas la deuxième):

$$ds^2 = Edu^2 + 2Fdudv + Gdv^2 \quad (25.106)$$

appelée aussi "**première forme différentielle de Gauss**". Il est intéressant d'écrire cette dernière relation sous la forme:

$$ds^2 = \frac{1}{E} (Edu + Fdv)^2 + \frac{EG - F^2}{E} dv^2 \quad (25.107)$$

et nous voyons que pour que ds^2 soit positif, il faut alors que E ainsi que $EG - F^2$ soient positifs.

Comme nous l'avons déjà démontré en calcul tensoriel, cette expression est indépendante de la nappe paramétrée (g, D) car l'élément de longueur infiniment petit ds est indépendant du paramétrage de Σ . Cette forme quadratique est donc un invariant qui représente la métrique sur Σ . Elle est par ailleurs notée sous la forme suivante par tradition:

$$ds^2 = g_{uu} du^2 + 2g_{uv} dudv + g_{vv} dv^2 \quad (25.108)$$

4.1.1. RÉGULARITÉ D'UNE SURFACE

Définition: Un point M appartenant à la surface Σ est dit (c'est relativement logique...) "**point régulier**", si et seulement si:

$$\frac{\partial \overrightarrow{OM}}{\partial u} \times \frac{\partial \overrightarrow{OM}}{\partial v} \neq \vec{0} \quad (25.109)$$

Une surface Σ est logiquement dite "**surface régulière**", si et seulement si, tous ses points sont réguliers (si le produit vectoriel est nul alors il y a quelque part un "pli" à 90°).

Remarquons que:

$$\begin{aligned}
 \left\| \frac{\partial \vec{OM}}{\partial u} \times \frac{\partial \vec{OM}}{\partial v} \right\|^2 &= \left\| \begin{pmatrix} \frac{\partial x(u,v)}{\partial u} \\ \frac{\partial y(u,v)}{\partial u} \\ \frac{\partial z(u,v)}{\partial u} \end{pmatrix} \times \begin{pmatrix} \frac{\partial x(u,v)}{\partial v} \\ \frac{\partial y(u,v)}{\partial v} \\ \frac{\partial z(u,v)}{\partial v} \end{pmatrix} \right\|^2 = \left\| \begin{pmatrix} \partial_u y \partial_v z - \partial_u z \partial_v y \\ \partial_u z \partial_v x - \partial_u x \partial_v z \\ \partial_u x \partial_v y - \partial_u y \partial_v x \end{pmatrix} \right\|^2 \\
 &= \left\| \sqrt{(\partial_u y \partial_v z - \partial_u z \partial_v y)^2 + (\partial_u z \partial_v x - \partial_u x \partial_v z)^2 + (\partial_u x \partial_v y - \partial_u y \partial_v x)^2} \right\|^2 \quad (25.110) \\
 &= (\partial_u y \partial_v z - \partial_u z \partial_v y)^2 + (\partial_u z \partial_v x - \partial_u x \partial_v z)^2 + (\partial_u x \partial_v y - \partial_u y \partial_v x)^2 \\
 &= (\partial_u y \partial_v z)^2 - 2\partial_u y \partial_v z \partial_u z \partial_v y + (\partial_u z \partial_v y)^2 \\
 &\quad + (\partial_u z \partial_v x)^2 - 2\partial_u z \partial_v x \partial_u x \partial_v z + (\partial_u x \partial_v z)^2 \\
 &\quad + (\partial_u x \partial_v y)^2 - 2\partial_u x \partial_v y \partial_u y \partial_v x + (\partial_u y \partial_v x)^2 \\
 &= \dots \text{je suis fatigué de faire de l'algèbre} \dots = EG - F^2
 \end{aligned}$$

L'angle α entre les deux courbes-coordonnées Γ_{u_o} et Γ_{v_o} à Σ en M_0 est donné par la définition du produit scalaire:

$$\frac{\partial \vec{OM}}{\partial u} \circ \frac{\partial \vec{OM}}{\partial v} = \left\| \frac{\partial \vec{OM}}{\partial u} \right\| \left\| \frac{\partial \vec{OM}}{\partial v} \right\| \cos \alpha = F = 0 \quad (25.111)$$

D'où l'expression:

$$\cos \alpha = \frac{\frac{\partial \vec{OM}}{\partial u} \circ \frac{\partial \vec{OM}}{\partial v}}{\left\| \frac{\partial \vec{OM}}{\partial u} \right\| \left\| \frac{\partial \vec{OM}}{\partial v} \right\|} = \frac{\frac{\partial \vec{OM}}{\partial u} \circ \frac{\partial \vec{OM}}{\partial v}}{\sqrt{\left(\frac{\partial \vec{OM}}{\partial u} \right)^2 \left(\frac{\partial \vec{OM}}{\partial v} \right)^2}} = \frac{F}{\sqrt{EG}} \quad (25.112)$$

Donc une condition nécessaire et suffisant pour que les courbes-coordonnées Γ_{u_o} et Γ_{v_o} soient perpendiculaires à Σ en M_0 et que F soit nul. Dans ce cas particulier, on dit que les coordonnées curvilignes u, v sur la surface sont des coordonnées orthogonales.

Exemples:

E1. Considérons le paramétrage du plan cartésien. Nous avons alors:

$$\overrightarrow{OM} = \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} \Rightarrow \begin{cases} \frac{\partial \overrightarrow{OM}}{\partial x} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \\ \frac{\partial \overrightarrow{OM}}{\partial y} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \end{cases} \quad (25.113)$$

d'où:

$$E = G = 1, F = 0 \quad (25.114)$$

Dès lors:

$$ds^2 = dx^2 + dy^2 \quad (25.115)$$

Nous retrouvons donc la même abscisse curviligne différentielle que celle vue dans le chapitre de Calcul Tensoriel et Relativité Générale avec la métrique diagonale de l'espace plat.

Nous avons aussi:

$$EG - F^2 = 1 \neq 0 \quad (25.116)$$

Donc, la surface est régulière. Nous avons également:

$$\cos \alpha = \frac{F}{\sqrt{EG}} = \frac{0}{\sqrt{1}} = 0 \Rightarrow \pm \frac{\pi}{2} \quad (25.117)$$

Donc, les deux courbes-coordonnées sont perpendiculaires.

E2. Considérons le paramétrage du cylindre. Nous avons alors (cf. chapitre de Calcul Vectoriel):

$$\overrightarrow{OM} = \begin{pmatrix} r \cos \phi \\ r \sin \phi \\ z \end{pmatrix} \Rightarrow \begin{cases} \frac{\partial \overrightarrow{OM}}{\partial r} = \begin{pmatrix} \cos \phi \\ \sin \phi \\ 0 \end{pmatrix} \\ \frac{\partial \overrightarrow{OM}}{\partial \phi} = \begin{pmatrix} -r \sin \phi \\ r \cos \phi \\ 0 \end{pmatrix} \end{cases} \quad (25.118)$$

d'où:

$$E = 1, G = r^2, F = 0 \quad (25.119)$$

Dès lors:

$$ds^2 = dr^2 + r^2 d\phi^2 \quad (25.120)$$

Nous retrouvons donc la même abscisse curviligne différentielle que celle vue dans le chapitre de Relativité Générale avec la métrique diagonale du plan en coordonnées polaires.

Nous avons aussi:

$$EG - F^2 = r^2 \neq 0 \quad (25.121)$$

Donc, la surface est régulière tant que r est non nul. Nous avons également:

$$\cos \alpha = \frac{F}{\sqrt{EG}} = \frac{0}{\sqrt{r^2}} = 0 \Rightarrow \pm \frac{\pi}{2} \quad (25.122)$$

Donc, les deux courbes-coordonnées sont perpendiculaires sur le cylindre.

E3. Considérons le paramétrage la sphère de centre O et de rayon r . Nous avons alors (cf. chapitre de Calcul Vectoriel):

$$\overrightarrow{OM} = \begin{pmatrix} r \sin \theta \cos \phi \\ r \sin \theta \sin \phi \\ r \cos \theta \end{pmatrix} \Rightarrow \begin{cases} \frac{\partial \overrightarrow{OM}}{\partial \theta} = \begin{pmatrix} r \cos \theta \cos \phi \\ r \cos \theta \sin \phi \\ -r \sin \theta \end{pmatrix} \\ \frac{\partial \overrightarrow{OM}}{\partial \phi} = \begin{pmatrix} -r \sin \theta \sin \phi \\ r \sin \theta \cos \phi \\ 0 \end{pmatrix} \end{cases} \quad (25.123)$$

d'où:

$$E = r^2, G = r^2 \sin^2 \theta, F = 0 \quad (25.124)$$

Dès lors:

$$ds^2 = r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (25.125)$$

Nous retrouvons donc la même abscisse curviligne différentielle que celle vue dans le chapitre de Relativité Générale avec la métrique diagonale du plan en coordonnées sphériques.

Nous avons aussi:

$$EG - F^2 = r^4 \sin^2 \theta \neq 0 \quad (25.126)$$

Donc, la surface est régulière tant que r et θ sont non nuls. Nous avons également:

$$\cos \alpha = \frac{F}{\sqrt{EG}} = \frac{0}{\sqrt{r^4 \sin^2 \theta}} = 0 \Rightarrow \pm \frac{\pi}{2} \quad (25.127)$$

Donc, les deux courbes-coordonnées sont perpendiculaires sur la sphère.

E4. Considérons le paramétrage de l'hyperboloïde. Nous avons alors (cf. chapitre de Géométrie Analytique):

$$\overrightarrow{OM} = \begin{pmatrix} a \cdot \sqrt{1+t^2} \cdot \cos(u) \\ a \cdot \sqrt{1+t^2} \cdot \sin(u) \\ b \cdot u \end{pmatrix} \Rightarrow \begin{cases} \frac{\partial \overrightarrow{OM}}{\partial t} = \begin{pmatrix} a \frac{\cos(u) \cdot t}{\sqrt{1+t^2}} \\ a \frac{\sin(u) \cdot t}{\sqrt{1+t^2}} \\ 0 \end{pmatrix} \\ \frac{\partial \overrightarrow{OM}}{\partial u} = \begin{pmatrix} -a \cdot \sqrt{1+t^2} \cdot \sin(u) \\ a \cdot \sqrt{1+t^2} \cdot \cos(u) \\ b \end{pmatrix} \end{cases} \quad (25.128)$$

d'où:

$$E = a^2 \frac{t^2}{1+t^2}, G = a^2(1+t^2) + b^2, F = 0 \quad (25.129)$$

Dès lors:

$$ds^2 = a^2 \frac{t^2}{1+t^2} dt^2 + (a^2(1+t^2) + b^2) du^2 \quad (25.130)$$

Prenons b comme étant nul. Nous avons alors:

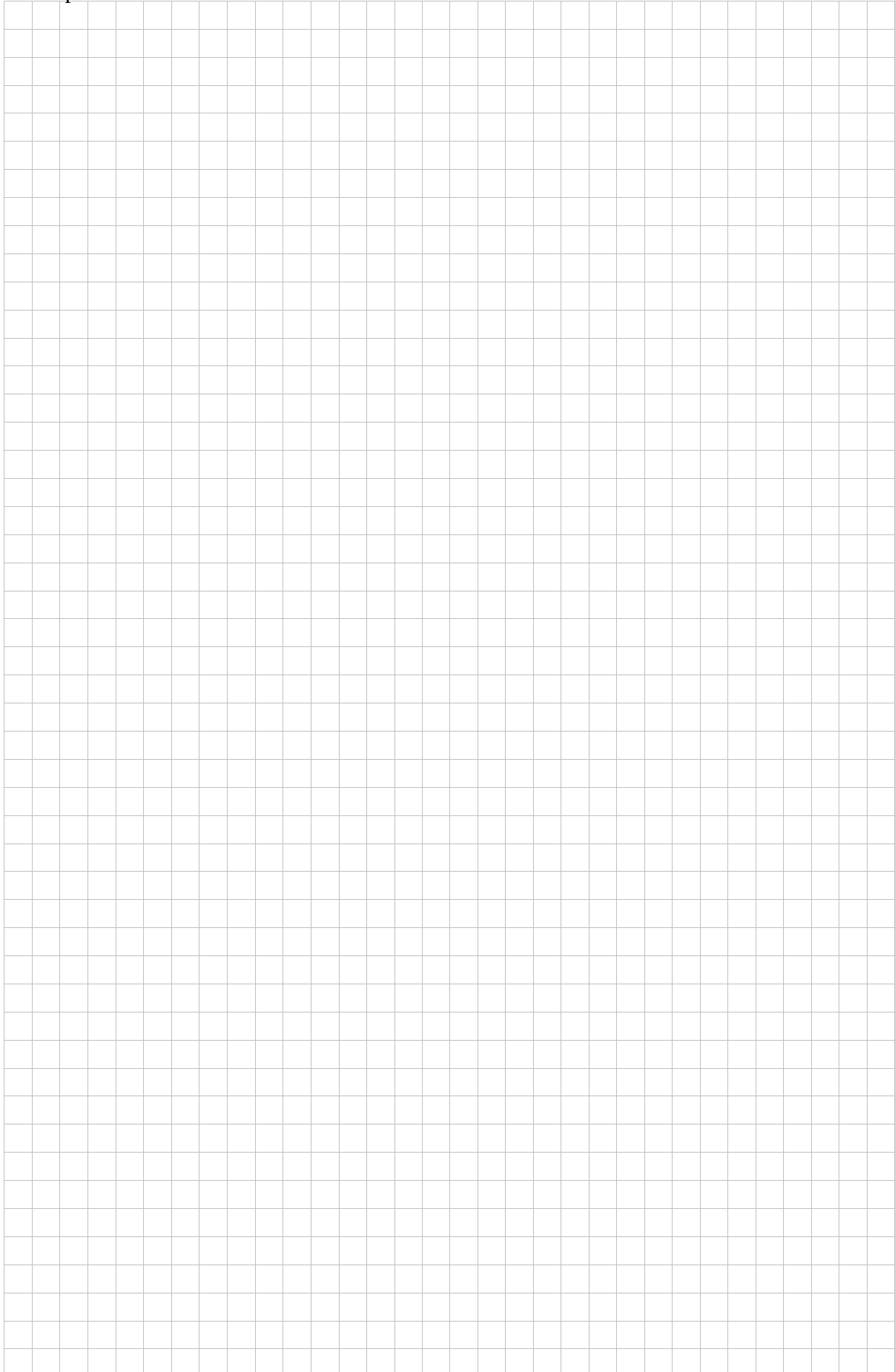
$$EG - F^2 = a^4 t^2 \quad (25.131)$$

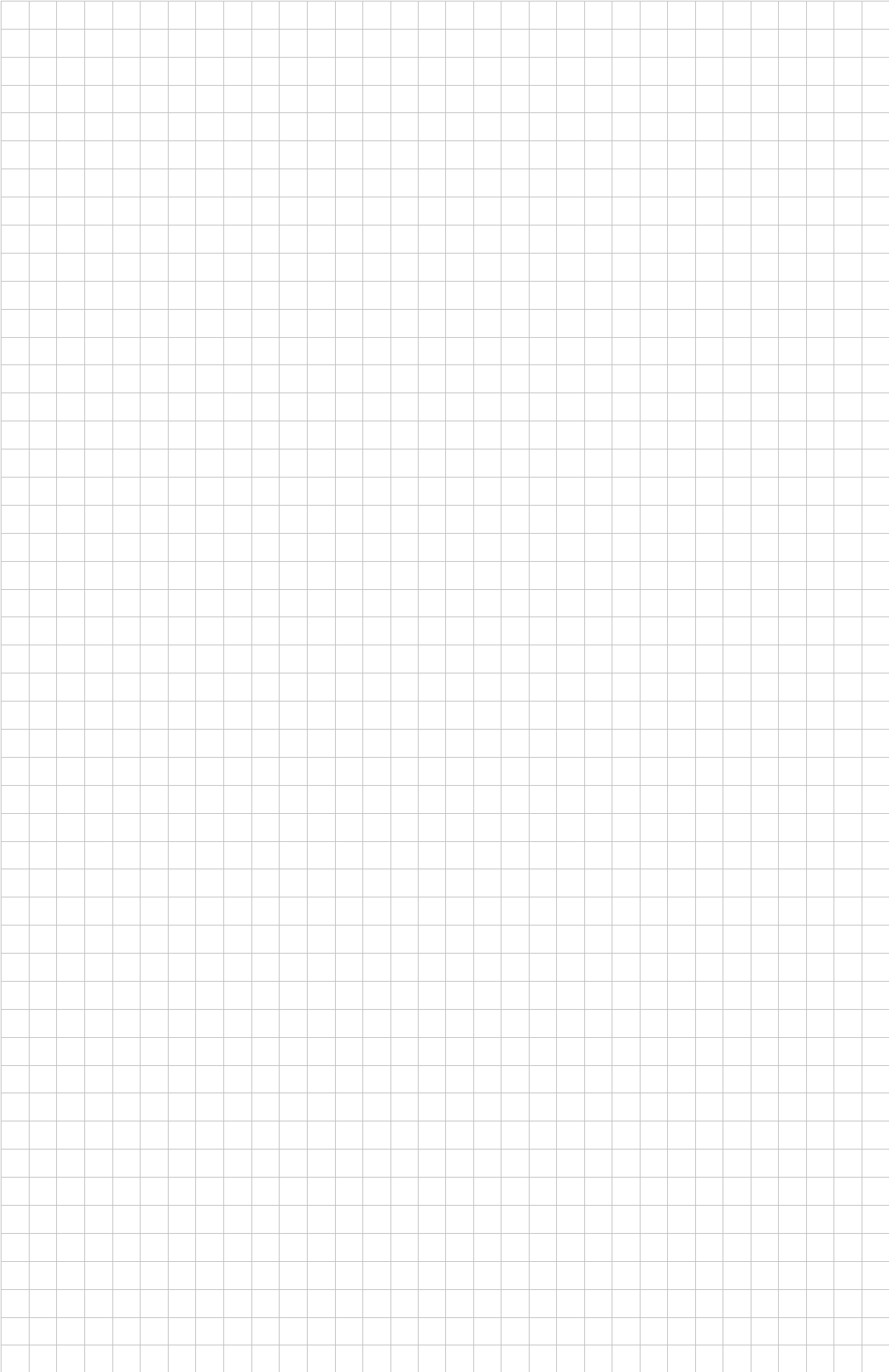
Donc, la surface est régulière tant que a est non nul. Nous avons également:

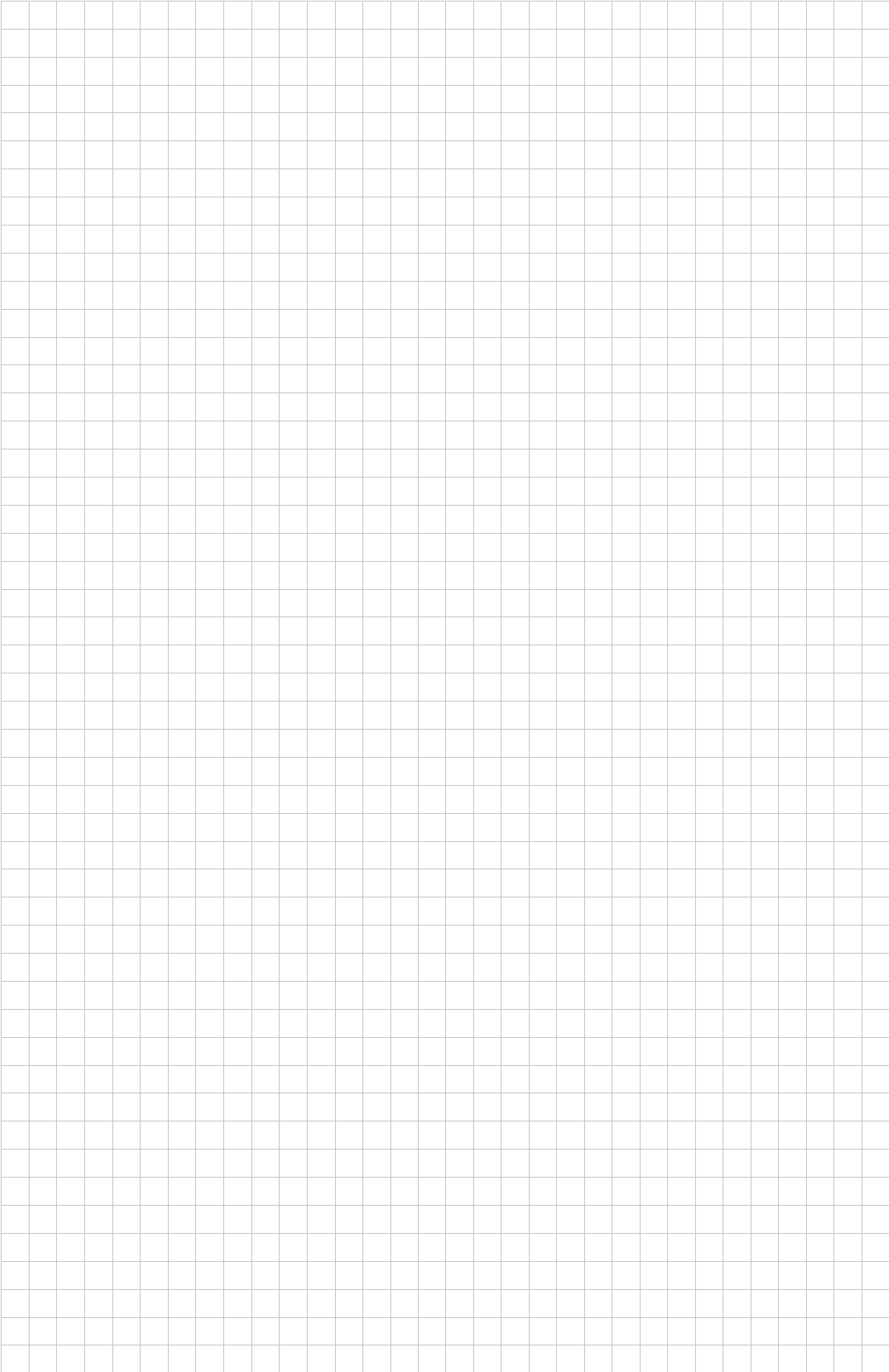
$$\cos \alpha = \frac{F}{\sqrt{EG}} = \frac{0}{a^2 t} = 0 \Rightarrow \pm \frac{\pi}{2} \quad (25.132)$$

Donc, les deux courbes-coordonnées sont perpendiculaires sur l'hyperboloïde.

Notes personnelles:





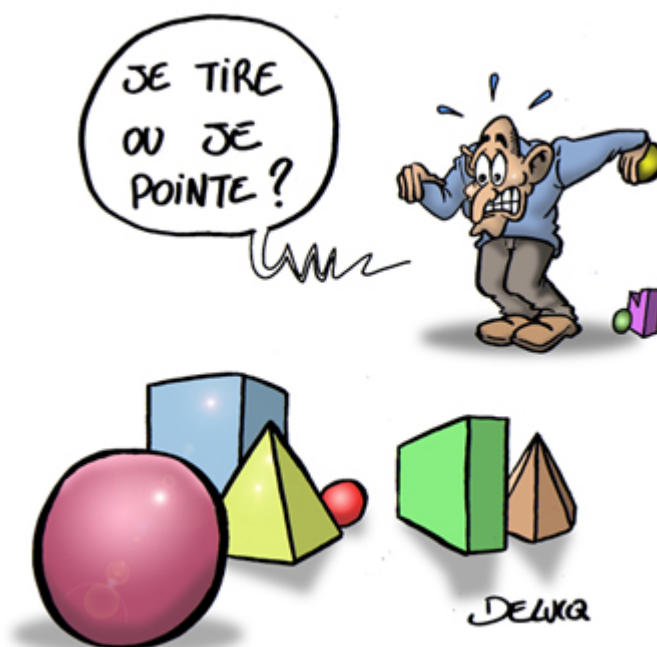




26. FORMES GÉOMÉTRIQUES

Nous avons déjà défini au début du chapitre de Géométrie Euclidienne les concepts de dimensions topologiques, ce qu'étaient un point de dimension nulle et une courbe de dimension unité. Nous ne reviendrons pas sur ces dernières et nous intéresserons aux formes de dimensions supérieures.

Le but du présent chapitre est de répertorier avec démonstrations quelques propriétés mathématiques remarquables des formes et corps géométriques connus (surface, volume, centre de masse, moment d'inertie). Effectivement, il existe nombre de formulaires les répertoriant sans démonstrations mais peu voire pas, d'ouvrages les démontrant toutes (nous n'en avons jamais vu en tout cas...). La liste ci-dessous est à ce jour loin d'être exhaustive (puisque'il existe une infinité de formes géométriques) mais elle sera complétée avec le temps.



Les quelques formes que nous avons souhaité présenter permettent assez facilement de trouver les propriétés remarquables d'un très grand nombre de formes non répertoriées sur cette page par assemblage ou décomposition.

Remarques:

R1. Les relations trigonométriques remarquables dans les formes géométriques ci-dessous ne sont pas démontrées dans ce chapitre. Celles-ci se trouvent déjà toutes dans le chapitre traitant spécifiquement de la Trigonométrie.

R2. Nous entendons par "centre de gravité", le "barycentre" tel que vu dans le chapitre de Géométrie Euclidienne.

1. SURFACES CONNUES

Il existe plusieurs définitions du concept de surface dont une due à Euclide et une autre moderne due à la topologie (voir chapitre du même nom).

Définitions:

D1. Une "**surface plane**" est ce qui a longueur et hauteur.

D2. Une "**surface**" est une variété topologique de dimension 2.

Remarque: Nous nous intéresserons dans un premier temps uniquement aux propriétés (périmètre, surface, centre de gravité, etc.) de surfaces plongées dans des géométries euclidiennes.

1.1. POLYGONES

Définition: Un "**polygone**" est une figure plane limitée par des segments de droites consécutifs (autrement dit: par une polygline fermée).

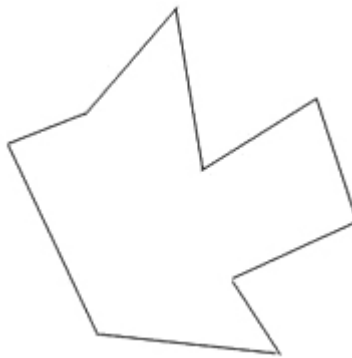


Figure: 26.1 - Exemple de polygone quelconque

Par définition, un "**quadrilatère**", "**pentagone**", "**hexagone**", "**heptagone**" sont des polygones à respectivement quatre, cinq, six, sept... côtés.

Nous distinguons trois grandes familles (mais elles ne sont pas les seules!) de polygones: les polygones croisés, les polygones concaves et les polygones convexes (nous retrouverons ces deux familles dans différents chapitres du site).

Avant d'aller plus loin nous tenons à préciser au lecteur qu'il n'existe de relation mathématique permettant de calculer la surface que pour des polygones simples. Même si dans la pratique nous ne rencontrons quasiment toujours que des polygones non simples, nous avons considéré comme inutile de nous attarder sur la détermination d'une relation qui permettrait de ramener le calcul de la surface d'un polygone quelconque à celui de polygones simples.

Définition: Un polygone est dit "**polygone croisé**" si deux au moins de ses côtés sont sécants, c'est-à-dire si au moins deux de ses côtés se coupent. C'est le cas du pentagone *ABCDE* ci-dessous:

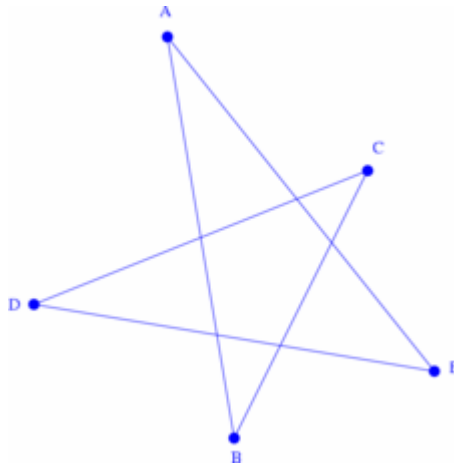


Figure: 26.2 - Exemple de polygone croisé

Remarque: "L'enveloppe" d'un polygone est le polygone obtenu en suivant le contour extérieur de celui-ci. Par exemple, l'enveloppe du pentagone précédent est un décagone dont les sommets sont les cinq sommets du pentagone et les cinq intersections de ses côtés.

Définition: Un polygone est dit "**polygone concave**" s'il n'est pas croisé et si une ou plusieurs de ses diagonales ne sont pas entièrement à l'intérieur de la surface délimitée par le polygone.

Par exemple, le pentagone $ACDBE$ ci-dessous est dit concave car les diagonales BC et CE sont respectivement à l'extérieur et partiellement à l'extérieur de la surface délimitée par le polygone.

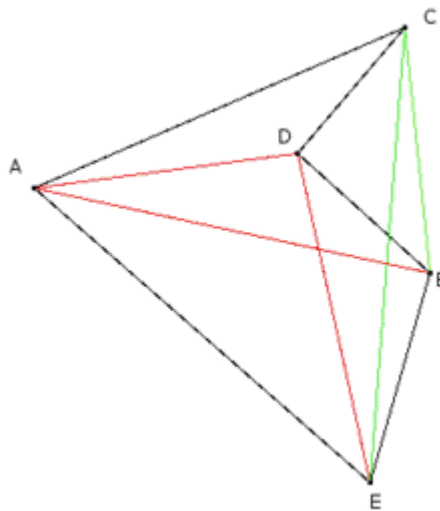


Figure: 26.3 - Exemple de polygone concave

Définition: Un polygone est dit "**polygone convexe**" s'il n'est pas croisé et si toutes ses diagonales sont entièrement à l'intérieur de la surface délimitée par le polygone. Ainsi, l'hexagone $MNOPQR$ ci-dessous est dit convexe:

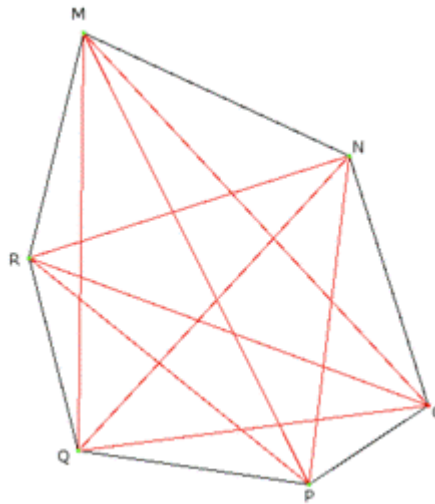


Figure: 26.4 - Exemple de polygone convexe

Relativement aux définitions données précédemment où les diagonales étaient mises en évidence, voyons s'il y a une relation permettant de connaître leur nombre relativement au nombre d'arêtes du polygone.

Partons pour cela d'un polygone de n côtés (notons qu'il a aussi n sommets):



Figure: 26.5 - Point de départ pour la démonstration

Nous définissons le total de segments s comme étant égal à la quantité de côtés (arêtes) n plus la quantité de diagonales d tel que:

$$s = n + d$$

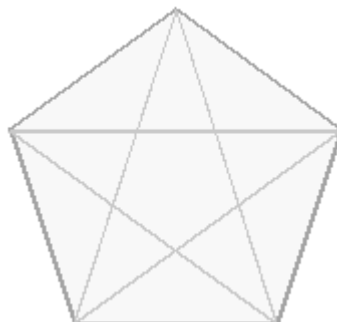


Figure: 26.6 - Représentation des diagonales

Maintenant, prenons le premier point de notre pentagone. Nous voyons que nous pouvons joindre tous

les points n , sauf le point considéré (-1) soit la formation de $n - 1$ segments comme le montre la figure ci-dessous:

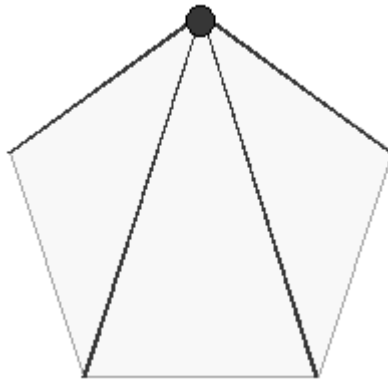


Figure: 26.7 - Exemple des segments partant d'un point

Avec le deuxième point, nous pouvons aussi joindre tous les points n , sauf le point considéré (-1) et le premier point déjà vu (-1) soit la formation de $n - 2$ segments:

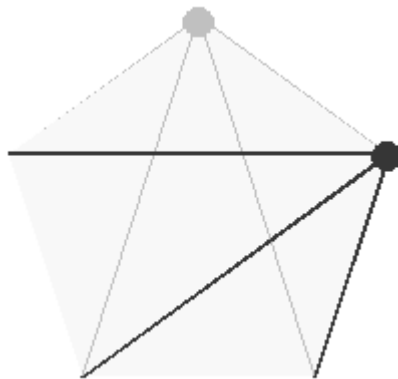


Figure: 26.8 - Démarche avec le 2ème point

Avec le troisième, nous pouvons aussi joindre tous les points n , sauf le point considéré (-1) et sauf les deux points déjà vu (-2) soit la formation de $n - 3$ segments

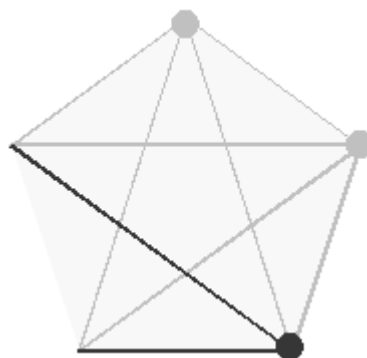


Figure: 26.9 - Démarche avec le troisième point

Nous continuons avec les autres points: le 4ème qui donne $n - 4$ segments, le 5ème qui donne $n - 5$ segments... In extenso, nous voyons donc que le $(n - 2)$ ème point donne donc $n - (n - 2)$ segments, etc.

Nous avons donc finalement pour:

$$\begin{aligned}
 s &= 1 + 2 + \dots + (n-2) + (n-1) \\
 + \frac{s &= (n-1) + (n-2) + \dots + 2 + 1}{2s = n + n + \dots + n + n} \quad (26.1)
 \end{aligned}$$

En simplifiant nous trouvons donc:

$$s = \frac{n(n-1)}{2} \quad (26.2)$$

Nous nous retrouvons donc avec deux relations:

$$s = \frac{n(n-1)}{2} \text{ et } s = n + d \quad (26.3)$$

Dès lors il vient que:

$$d = s - n = \frac{n(n-1)}{2} - n = \frac{n(n-3)}{2} \quad (26.4)$$

1.1.1. RECTANGLE

Définition: Le "rectangle" est un cas particulier du quadrilatère (forme à quatre côtés délimitée par des segments finis telle que: losange, carré, rectangle, trapèze, etc.) dans le sens où ses côtés L et H (notation pour Longueur et Hauteur selon figure ci-dessous) sont égaux deux à deux et à angle droit (en d'autres termes, L n'est pas forcément égal à H).

D'autres définitions possibles consistent à dire qu'un rectangle est un parallélogramme disposant d'un angle droit ou un quadrilatère ayant quatre angles droits.

Remarque: Le rectangle peut être vu comme la composition de deux (ou plus) triangles rectangles (voir plus loin la définition). Pour construire un rectangle, il suffirait d'avoir un seul et unique triangle rectangle et lui faire subir une double réflexion et une rotation par rapport à un axe bien choisi (cf. chapitre de Géométrie Euclidienne).

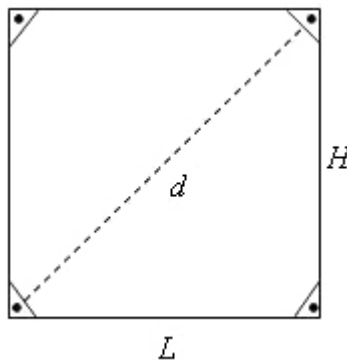


Figure: 26.10 - Exemple de rectangle

De par les axiomes d'Euclide (cf. chapitre de Géométrie Euclidienne), le périmètre d'un rectangle est donné par:

$$P = 2(L + H) \quad (26.5)$$

Et par définition, sa surface par:

$$S = L \cdot H \quad (26.6)$$

et la longueur de sa diagonale par (application du théorème de Pythagore):

$$d = \sqrt{L^2 + H^2} \quad (26.7)$$

La position du centre de gravité du rectangle, si nous posons le repère cartésien dans le coin inférieur gauche de la forme, est trivialement donnée par:

$$G(x,y) = (L/2, H/2) \quad (26.8)$$

Enfin, indiquons que si nous étions des êtres vivants dans un espace à deux dimensions, le rectangle serait ce que nous apercevions si un parallélépipède traversait notre univers parallèlement à ses faces.

1.1.2. CARRÉ

Définition: Le "carré" est un cas particulier du rectangle dans le sens où ses quatre côtés sont égaux tels que $L = H$.

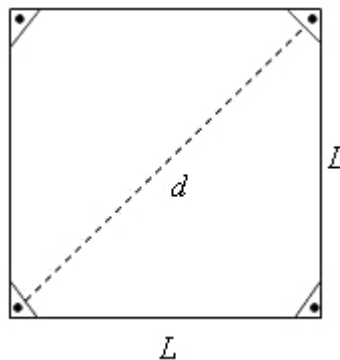


Figure: 26.11 - Exemple de carré

De par les axiomes d'Euclide (cf. chapitre de Géométrie Euclidienne), le périmètre du carré est donné par:

$$P = 4 \cdot L \quad (26.9)$$

Ainsi, il vient pour la surface que:

$$S = L^2 \quad (26.10)$$

et pour sa diagonale:

$$d = \sqrt{L^2 + H^2} = \sqrt{L^2 + L^2} = \sqrt{2L^2} = \sqrt{2} \cdot L \quad (26.11)$$

La position du centre de gravité du carré, si nous posons le repère cartésien dans le coin inférieur gauche de la forme, est trivialement donnée par:

$$G(x,y) = (L/2, L/2) \quad (26.12)$$

Enfin, indiquons que si nous étions des êtres vivants dans un espace à deux dimensions, le carré serait ce que nous apercevions si un cube traversait notre univers parallèlement à ses faces.

1.1.3. TRIANGLE

Définition: Le "**triangle quelconque**" est un polygone à trois côtés et englobe dans les cas particuliers, les triangles: isocèles, équilatéraux et rectangles.

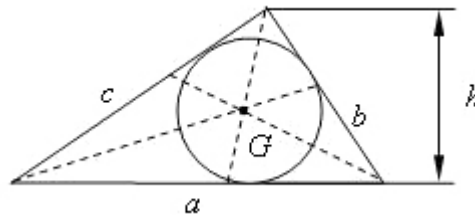


Figure: 26.12 - Exemple de triangle quelconque avec cercle inscrit

De par les axiomes d'Euclide (cf. chapitre de Géométrie Euclidienne), le périmètre d'un triangle quelconque est donné par:

$$P = a + b + c \quad (26.13)$$

Le triangle quelconque est toujours décomposable en deux triangles rectangles. Ainsi, celui de la figure ci-dessus peut se décomposer en deux triangles rectangles de base respective a_1 et a_2 (définis par la projection orthogonale du sommet opposé au segment a) tels que:

$$a = a_1 + a_2 \quad (26.14)$$

La surface de chacun de ces deux triangles rectangles est comme nous l'avons déjà implicitement dit dans notre étude du rectangle, la moitié de la surface d'un rectangle de même longueur et même hauteur. Ainsi:

$$S_{a1} = \frac{a_1 \cdot h}{2} \quad S_{a2} = \frac{a_2 \cdot h}{2} \quad (26.15)$$

Ainsi, la somme de ces surfaces, nous donne la surface du triangle quelconque:

$$S = S_{a1} + S_{a2} = \frac{a_1 \cdot h}{2} + \frac{a_2 \cdot h}{2} = \frac{a_1 + a_2}{2} h = \frac{a \cdot h}{2} \quad (26.16)$$

Nous pouvons dire à partir de cette dernière relation, que la surface de tout triangle quelconque est assimilable à la moitié de la surface d'un rectangle de longueur $L = a$ et hauteur $H = h$.

Remarque: Quelle que soit la base a, b, c et la hauteur respective h_a, h_b, h_c , le raisonnement précédent reste bien évidemment totalement juste.

La détermination du centre de gravité (ou barycentre) G (cf. chapitre de Géométrie Euclidienne) est un

peu moins intuitive que dans le cas du rectangle...

Nous pouvons bien sûr nous servir d'un repère et des outils du calcul vectoriel pour très facilement déterminer ce dernier. Nous allons donc démontrer que le centre de gravité d'un triangle quelconque se situe à l'intersection de toutes les médianes:

Démonstration:

Soit un triangle ABC . Nous appelons A' le milieu du segment BC , B' celui de AC et C' celui de AB :

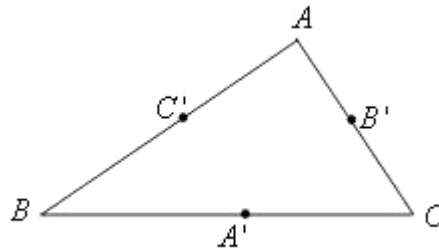


Figure: 26.13 - Point de départ pour la détermination du centre de gravité

Nous allons démontrer que le seul point G vérifiant (cf. le chapitre de Géométrie Euclidienne):

$$\overrightarrow{GA} + \overrightarrow{GB} + \overrightarrow{GC} = \vec{0} \quad (26.17)$$

est le point de concours des trois médianes du triangle ABC . Cette démonstration s'effectuera en deux étapes, en deux propositions. Au terme, nous pourrions conclure par le théorème.

Propositions:

P1. Si ABC est un triangle alors il existe un et un seul centre de gravité G tel que $\overrightarrow{GA} + \overrightarrow{GB} + \overrightarrow{GC} = \vec{0}$

P2. Les trois médianes d'un triangle sont concourantes. Leur point d'intersection est ce point G .

Démonstrations:

DM1. Soit G un point du plan tel que $\overrightarrow{GA} + \overrightarrow{GB} + \overrightarrow{GC} = \vec{0}$. Nous pouvons alors écrire que:

$$\overrightarrow{GA} + \underbrace{\overrightarrow{GA} + \overrightarrow{AB}}_{\overrightarrow{GB}} + \underbrace{\overrightarrow{GA} + \overrightarrow{AC}}_{\overrightarrow{GC}} = \vec{0} \quad (26.18)$$

d'où:

$$\overrightarrow{AG} = \frac{1}{3}(\overrightarrow{AB} + \overrightarrow{AC}) \quad (26.19)$$

Cette relation vectorielle garantit que le point G est unique et que nous pouvons même le placer!

□C.Q.F.D.

DM2. Pour démontrer que les trois médianes sont concourantes, nous allons prouver que G appartient à chacune des trois médianes.

Au point P1., nous avons démontré que G vérifie l'égalité:

$$\overrightarrow{AG} = \frac{1}{3}(\overrightarrow{AB} + \overrightarrow{AC}) \quad (26.20)$$

Comme A' est le milieu du côté BC , nous pouvons alors écrire que:

$$\overrightarrow{AB} + \overrightarrow{AC} = 2\overrightarrow{AA'} \quad (26.21)$$

Il vient alors que:

$$\overrightarrow{AG} = \frac{1}{3}(\overrightarrow{AB} + \overrightarrow{AC}) = \frac{1}{3}(2\overrightarrow{AA'}) = \frac{2}{3}\overrightarrow{AA'} \quad (26.22)$$

Les vecteurs $\overrightarrow{AG}$ et $\overrightarrow{AA'}$ sont donc colinéaires! Donc les points A , G , A' sont alignés. Autrement écrit, le point G fait partie de la médiane AA' du triangle ABC . Nous pouvons même dire qu'il se trouve aux deux tiers du segment AA' à partir du sommet A .

Ce que nous venons de montrer avec la médiane AA' est bien évidemment aussi vrai pour les deux autres médianes. Ainsi:

$$\begin{aligned} \overrightarrow{AG} &= \frac{2}{3}\overrightarrow{AA'} \\ \overrightarrow{BG} &= \frac{2}{3}\overrightarrow{BB'} \\ \overrightarrow{CG} &= \frac{2}{3}\overrightarrow{CC'} \end{aligned} \quad (26.23)$$

En résumé, le point G fait donc partie des trois médianes AA' , BB' et CC' . Ces trois droites sont donc concourantes et le point G en est le point d'intersection. Ce résultat nous sera utile plus tard lors de notre étude des polyèdres.

□C.Q.F.D.

Enfin, indiquons que si nous étions des êtres vivants dans un espace à deux dimensions, le triangle serait ce que nous apercevions si des formes géométriques composées d'au moins trois faces jointes traversaient notre univers par un des sommets.

Nous arrêterons là cette analogie avec un espace à deux dimensions généralisable à chaque forme géométrique que nous allons présenter par la suite (cercle et sphère, ellipse et ellipsoïde, etc.). L'idée était surtout de soumettre la conception que les volumes que nous connaissons dans notre quotidien peuvent aussi être vus comme des formes à 4 dimensions traversant notre espace de 3 dimensions.

1.1.4. TRIANGLE ISOCÈLE

Définition: Un "triangle isocèle" est un cas particulier du triangle quelconque, dans le sens où il a deux côtés égaux (isométriques).

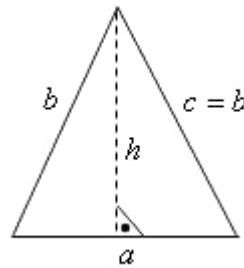


Figure: 26.14 - Exemple de triangle isocèle

Le périmètre d'un tel triangle reste:

$$P = a + b + c \quad (26.24)$$

mais comme il a deux côtés égaux tels que par exemple:

$$P = a + b + b = a + 2b \quad (26.25)$$

La surface comme nous l'avons démontré dans le cas général reste:

$$S = \frac{a \cdot h}{2} \quad (26.26)$$

Et le centre de gravité reste, comme nous l'avons démontré dans le cas général, à la position:

$$\begin{aligned} \overrightarrow{AG} &= \frac{2}{3} \overrightarrow{AA'} \\ \overrightarrow{BG} &= \frac{2}{3} \overrightarrow{BB'} \\ \overrightarrow{CG} &= \frac{2}{3} \overrightarrow{CC'} \end{aligned} \quad (26.27)$$

Propriété remarquable d'un triangle isocèle: la médiatrice et la médiane h du troisième côté non égal aux deux autres sont confondues (cf. chapitre de Géométrie Euclidienne).

1.1.5. TRIANGLE ÉQUILATÉRAL

Définition: Un "triangle équilatéral" est un cas particulier du triangle, dans le sens où il a trois côtés égaux:

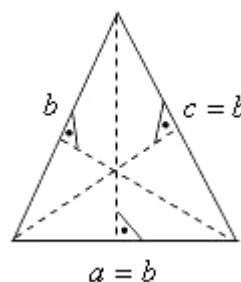


Figure: 26.15 - Exemple de triangle équilatéral

Le périmètre d'un tel triangle reste:

$$P = a + b + c \quad (26.28)$$

mais comme il a trois côtés tel que $a = b = c$:

$$P = 3a \quad (26.29)$$

La surface comme nous l'avons démontré dans le cas général reste:

$$S = \frac{a \cdot h}{2} \quad (26.30)$$

Et le centre de gravité reste comme nous l'avons démontré dans le cas général, à la position:

$$\begin{aligned} \overrightarrow{AG} &= \frac{2}{3} \overrightarrow{AA'} \\ \overrightarrow{BG} &= \frac{2}{3} \overrightarrow{BB'} \\ \overrightarrow{CG} &= \frac{2}{3} \overrightarrow{CC'} \end{aligned} \quad (26.31)$$

Propriété remarquable d'un triangle équilatéral: Médiatrices et médianes sont confondues (cf. chapitre de Géométrie Euclidienne)!

1.1.6. TRIANGLE RECTANGLE

Définition: Un "triangle rectangle" est un cas particulier du triangle, dans le sens que sur un de ses trois angles, il y en a un qui est droit.

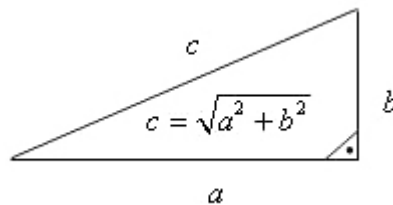


Figure: 26.16 - Exemple de triangle rectangle

Le périmètre d'un tel triangle reste:

$$P = a + b + c \quad (26.32)$$

La surface comme nous l'avons démontré dans le cas général reste (surface de la moitié d'un rectangle de même base et de même hauteur):

$$S = \frac{a \cdot h}{2} \quad (26.33)$$

Et le centre de gravité reste comme nous l'avons démontré dans le cas général, à la position:

$$\begin{aligned}\overrightarrow{AG} &= \frac{2}{3} \overrightarrow{AA'} \\ \overrightarrow{BG} &= \frac{2}{3} \overrightarrow{BB'} \\ \overrightarrow{CG} &= \frac{2}{3} \overrightarrow{CC'}\end{aligned} \quad (26.34)$$

Propriété remarquable d'un triangle rectangle: le triangle rectangle a ceci de particulier, que nous pouvons directement lui appliquer le théorème de Pythagore (cf. chapitre de Géométrie Euclidienne).

1.1.7. TRAPÈZE

Définition: Un "**trapèze**", est un quadrilatère (non croisé) ayant deux côtés (au moins) parallèles.

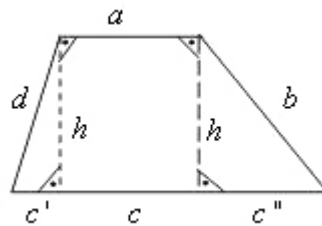


Figure: 26.17 - Exemple de trapèze

Lorsque les deux côtés ont même longueur (ou, sont de même longueur), nous obtenons les cas particuliers du carré, du rectangle, du losange, du parallélogramme (ici, ordre du plus précis au plus général, nous pourrions mettre le losange en n° 2).

Aussi un usage courant consiste à ne retenir qu'une définition plus restrictive, afin de ne pas prendre en compte ces figures particulières. Nous ajoutons dans ce cas que les longueurs des deux côtés parallèles ne sont pas égales (cela permet aux élèves des petites classes d'éviter les confusions résultant de l'existence de deux noms pour le même objet, par exemple losange et trapèze).

Remarque: Il existe un cas particulier de trapèze, le "**trapèze isocèle**", dont les deux côtés non parallèles sont de même longueur. (nous pouvons ajouter: Comme ses deux côtés ne sont pas parallèles, il ne s'agit pas d'un parallélogramme).

1.1.8. PARALLÉLOGRAMME

Définition: Le "**parallélogramme**" est un cas particulier du quadrilatère (et du losange aussi), où les côtés sont parallèles deux à deux:

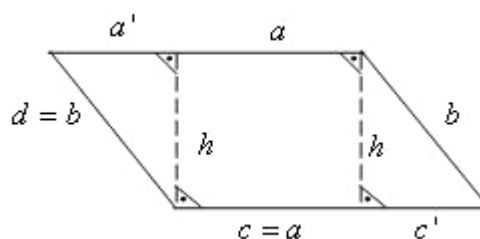


Figure: 26.18 - Exemple de parallélogramme

Remarque: Tous les parallélogrammes sont donc dans la famille des trapèzes.

1.1.9. LOSANGE

Définition: Le "**losange**" est un cas particulier du parallélogramme dans le sens où ses quatre côtés sont égaux.

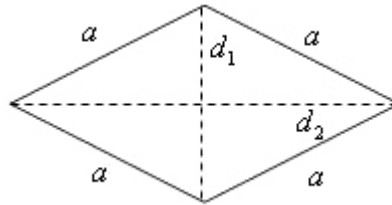


Figure: 26.19 - Exemple de losange

1.1.10. CERCLE

Il existe plusieurs définitions possibles du cercle. Voyons-en au moins deux.

Définitions:

D1. Un "**cercle**" est un cas particulier d'un polygone avec une infinité de côtés.

D2. Un "**cercle**" est une courbe plane dont tous les points sont à égale distance d'un point fixe appelé "**centre**".

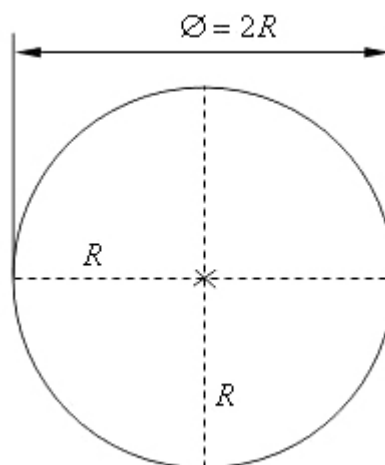


Figure: 26.20 - Exemple de cercle

Nous démontrons dans la section d'Informatique Théorique (cf. chapitre de Méthodes Numériques), que le périmètre d'un cercle de rayon R et donc de diamètre $\varnothing = 2R$ est donné par:

$$P = 2\pi R = (2R)\pi = \varnothing \cdot \pi \quad (26.35)$$

La relation de surface peut être obtenue de deux manières:

1. Par recherche de la primitive du périmètre P ce qui nous donne:

$$S = \int P dR = 2\pi \int R dR = \pi R^2 \quad (26.36)$$

2. La seconde méthode est plus esthétique et fait appel à l'équation paramétrique du cercle, trivialement donnée par les projections orthogonales des coordonnées cartésiennes:

$$\begin{aligned} x &= R \cos \theta \\ y &= R \sin \theta \end{aligned} \quad (26.37)$$

Nous savons que l'aire décrite par une fonction $f(x)$ est donnée par:

$$S = \int_a^b f(x) dx = \int_a^b y dx \quad (26.38)$$

Il nous suffit alors de substituer dans cette intégrale les variables paramétrées:

$$y = R \sin \theta \quad dx = d(R \cos \theta) = R d \cos \theta = -R \sin \theta d\theta \quad (26.39)$$

Ainsi:

$$S = \int_a^b y dx = \int_a^b (R \sin \theta) \cdot (-R \sin \theta d\theta) = -R^2 \int_a^b \sin^2 \theta d\theta \quad (26.40)$$

Les bornes d'intégration étant bien évidemment $\theta = [0, -2\pi]$ nous avons:

$$\begin{aligned} S &= -R^2 \int_0^{-2\pi} \sin^2 \theta d\theta = -2R^2 \int_0^{-\pi} \sin^2 \theta d\theta = -2R^2 \int_0^{-\pi} \frac{1 - \cos(2\theta)}{2} d\theta = 2R^2 \left(\frac{\sin(2\theta)}{4} - \frac{\theta}{2} \right) \Bigg|_0^{-\pi} \\ &= R^2 \theta \Big|_0^{-\pi} - R^2 \frac{\sin(2\theta)}{2} \Big|_0^{-\pi} = R^2 \pi \end{aligned} \quad (26.41)$$

Nous avons donc aussi par cette méthode:

$$S = \pi R^2 = \pi \left(\frac{\varnothing}{2} \right)^2 = \pi \cdot \frac{\varnothing^2}{4} \quad (26.42)$$

La longueur l d'une tranche d'angle d'ouverture α d'un cercle de rayon R est bien évidemment donnée par:

$$l = \frac{2\pi R}{2\pi} \cdot \alpha = R\alpha \quad (26.43)$$

et la surface S_f d'une tranche d'angle d'ouverture α d'un cercle de rayon R de manière identique par:

$$S_f = \frac{\pi R^2}{2\pi} \cdot \alpha = \frac{R^2}{2} \alpha \quad (26.44)$$

Cherchons maintenant à déterminer la surface S_d d'une tranche d'un disque :

Soit connue la relation de calcul de la surface d'un triangle. Nous avons selon la figure ci-dessous (la démonstration tient seulement dans le résultat lui-même):

$$S_d = \frac{R^2}{2} \alpha - \frac{a \cdot (R - h)}{2} = \frac{R^2}{2} \frac{l}{R} - \frac{a \cdot (R - h)}{2} = \frac{Rl - a \cdot (R - h)}{2}$$

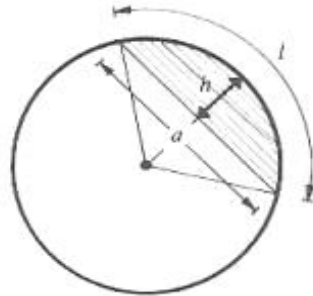


Figure: 26.21 - Calcul de la surface S_d d'une tranche d'un disque

Remarque: Par définition du cercle, il est évident que le centre de gravité du cercle se confond avec le centre de celui-ci.

1.1.11. ELLIPSE

Définition: Une "ellipse" est une courbe fermée dont chaque point est tel que la somme de ses distances à deux points fixes appelés "foyers" est constante (comme nous l'avons vu dans le chapitre d'Algèbre Ensembliste, l'ellipse peut aussi être vue comme une transformation affine du cercle).

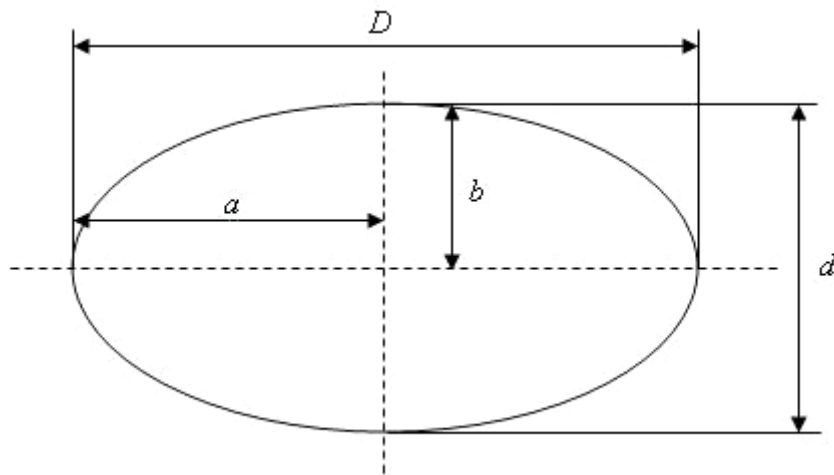


Figure: 26.22 - Exemple d'ellipse

Introduisons pour commencer un petit texte relativement au calcul du périmètre de l'ellipse:

Soit l'équation paramétrique en coordonnées cartésiennes d'une ellipse:

$$\begin{aligned} x &= a \cos \theta \\ y &= b \sin \theta \end{aligned} \quad (26.45)$$

La distance entre le centre de l'ellipse et son périmètre est alors donnée par le théorème de Pythagore:

$$R(\theta) = \sqrt{a^2 \cos^2 \theta + b^2 \sin^2 \theta} \quad (26.46)$$

Un élément d'arc est alors donné par:

$$R(\theta) d\theta = \sqrt{a^2 \cos^2 \theta + b^2 \sin^2 \theta} d\theta \quad (26.47)$$

Le périmètre de l'ellipse est alors donné par l'intégrale:

$$P = \int_0^{2\pi} R(\theta) d\theta = \int_0^{2\pi} \sqrt{a^2 \cos^2 \theta + b^2 \sin^2 \theta} d\theta = 4 \int_0^{\pi/2} \sqrt{a^2 \cos^2 \theta + b^2 \sin^2 \theta} d\theta \quad (26.48)$$

et là ça commence à se corser... Ce genre d'intégrale n'est pas facilement calculable à l'aide des primitives connues, intégration par parties, changements de variable ou autre. Il s'agit de ce que nous appelons une "**intégrale elliptique du second ordre en J**" pour $0 < b < a$ (cf. chapitre de Calcul Différentiel Et Intégral):

$$J(a, b) = \int_0^{\pi/2} \sqrt{a^2 \cos^2 \theta + b^2 \sin^2 \theta} d\theta \quad (26.49)$$

De longs développements que nous présenterons dans quelques années dans le chapitre de Calcul Différentiel Et Intégral donnent pour le périmètre après un calcul en série limitée:

$$P \cong \pi \sqrt{2(a^2 + b^2)} = 2\pi a \sqrt{1 - e^2} \quad (26.50)$$

La relation de surface de l'ellipse peut être obtenue de manière très similaire à celle du cercle et les calculs sont curieusement beaucoup plus simples que ceux du périmètre. Rappelons que l'équation paramétrique l'ellipse est:

$$\begin{aligned} x &= a \cos \theta \\ y &= b \sin \theta \end{aligned} \quad (26.51)$$

Nous savons que l'aire décrite par une fonction $f(x)$ est donnée par:

$$S = \int_a^b f(x) dx = \int_a^b y dx \quad (26.52)$$

Il nous suffit alors de substituer dans cette intégrale les variables paramétrées:

$$y = b \sin \theta \quad dx = d(a \cos \theta) = a d(\cos \theta) = -a \sin \theta d\theta \quad (26.53)$$

Ainsi:

$$S = \int_a^b y dx = \int_a^b (b \sin \theta) \cdot (-a \sin \theta d\theta) = -ab \int_a^b \sin^2 \theta d\theta \quad (26.54)$$

Les bornes d'intégration étant bien évidemment $\theta = [0, -2\pi]$ nous avons:

$$\begin{aligned}
 S &= -ab \int_0^{-2\pi} \sin^2 \theta d\theta = -2ab \int_0^{-\pi} \frac{1 - \cos(2\theta)}{2} d\theta = -2ab \left(\frac{1}{2} \theta - \frac{1}{4} \sin(2\theta) \right) \Big|_0^{-\pi} \\
 &= -2ab \left[\left(\frac{1}{2} (-\pi) - \frac{1}{4} \sin(2 \cdot (-\pi)) \right) - \left(\frac{1}{2} \cdot 0 - \frac{1}{4} \sin(2 \cdot 0) \right) \right] \quad (26.55) \\
 &= -2ab \frac{1}{2} (-\pi) = ab\pi
 \end{aligned}$$

Remarque: Il faut faire attention dans ce genre de calculs à l'ordre des bornes d'intégration. Effectivement, si nous avons pris les bornes allant de $[0, \pi]$ (au lieu de $[0, -\pi]$) il faut imaginer que la fonction intégrée parcourt le périmètre dans le sens négatif de l'axe des abscisses. Donc, l'intégrale serait alors forcément négative.

Nous avons donc aussi par cette méthode:

$$S = \pi ab = \frac{\pi}{4} D \cdot d \quad (26.56)$$

Remarques:

R1. Nous supposons comme évident que le centre de gravité de l'ellipse se confond avec le centre de celle-ci.

R2. Nous renvoyons le lecteur à l'étude des coniques (cf. chapitre de Géométrie Analytique) pour le calcul de la surface d'une ellipse à partir de son "paramètre d'ellipse" et de son "excentricité" (tout y est démontré).

2. VOLUMES CONNUS

Il existe plusieurs définitions du concept de volume (surface qui limite un corps). Une définition due à Euclide et une autre due au domaine de la topologie (voir le chapitre du même nom).

Définitions:

D1. Un "**volume**" est ce qui a longueur, largeur et hauteur.

D2. Un "**volume**" est une variété topologique de dimension 3.

Les surfaces qui limitent un corps peuvent être planes ou courbes:

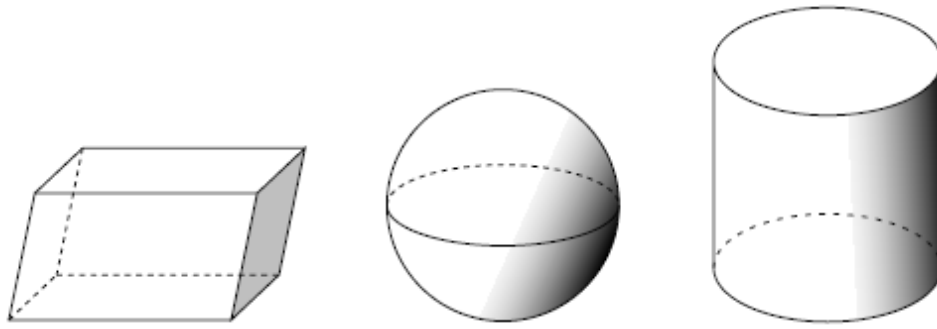


Figure: 26.23 - Exemples de volumes délimités par une surface

À gauche, le corps est limité uniquement par des surfaces planes, au milieu par une et une seule unique surface courbe, et à droite par une surface courbe et deux surfaces planes.

Remarque: Nous nous intéresserons dans un premier temps uniquement aux propriétés (surface, volume, centre de gravité, moment d'inertie...) de volumes plongés dans des géométries euclidiennes.

2.1. POLYÈDRES

L'étude des polyèdres (particulièrement les polyèdres platoniciens) est très importante en physique (pour la cristallographie par exemple) et en mathématiques car elle permet d'avoir une application sympathique des groupes finis (cf. chapitre d'Algèbre Ensembliste). Il convient donc de porter une lecture relativement attentive à ce qui va suivre.

Par ailleurs, l'étude des polyèdres est aussi un moyen très pédagogique et esthétique pour voir la mise en oeuvre de plusieurs théorèmes géométriques, de trigonométrie et d'algèbre vectoriel.

Précisons avant toute chose que les différents polyèdres ne seront délibérément pas présentés sur un pied d'égalité. Ainsi, nous nous concentrerons sur certaines propriétés pour certains et pas pour d'autres.

Définitions:

D1. Un "**polyèdre**" est un solide dont la frontière est formée de plans ou de portions de plan. Les portions de plan, qui comprennent ainsi entre elles le polyèdre, sont les faces; chaque face, étant limitée par intersections (les arêtes) avec les faces voisines, est un polygone. Les côtés de ce polygone sont les arêtes du polyèdre. Nous appelons "sommets" d'un polyèdre tout sommet d'une quelconque de ses faces.

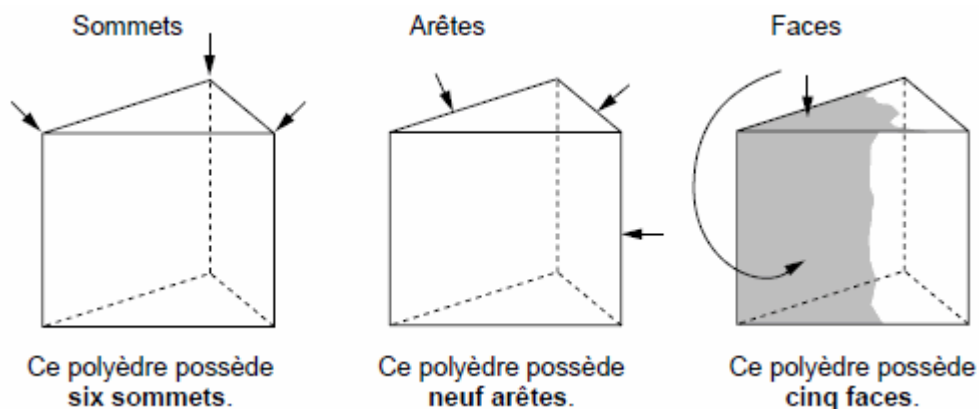


Figure: 26.24 - Exemple de polyèdre

D2. Un "polygone régulier" est un polygone dont les côtés, et tous les angles sont égaux (cette définition nous sera utile pour les polyèdres réguliers).

2.1.1. PARALLÉLÉPIPÈDE

Définition: Le "parallélépipède" est un volume à six faces parallèles deux à deux (donc il n'est pas un polyèdre régulier!).

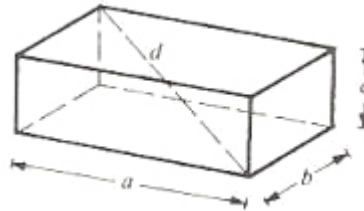


Figure: 26.25 - Exemple de parallélépipède

Son volume est simplement donné par la définition même du volume...:

$$V = a \cdot b \cdot c \quad (26.57)$$

Quant à sa surface, il s'agit simplement de la somme des surfaces des rectangles sans rien de particulier.

Calculons maintenant le moment d'inertie d'une plaque (parallélépipède) d'épaisseur e et de surface transversale S dont l'axe de rotation est y :

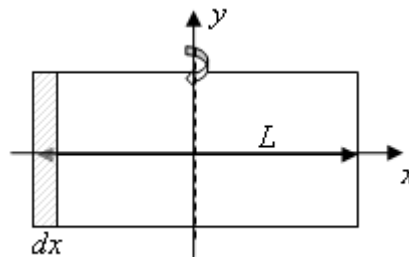


Figure: 26.26 - Recherche du moment d'inertie d'une plaque

Un élément de volume du rectangle (en gris) est donné par:

$$dV = Sdx \Rightarrow dm = \rho Sdx \quad (26.58)$$

et:

$$J_y = \int x^2 dm = \int_{-L/2}^{+L/2} x^2 \rho S dx = \rho S \left\{ \frac{\left(\frac{L}{2}\right)^3}{3} - \frac{\left(-\frac{L}{2}\right)^3}{3} \right\} = \rho \frac{SL^3}{12} = \rho SL \frac{L^2}{12} = M \frac{L^2}{12} \quad (26.59)$$

et occupons-nous maintenant du moment d'inertie de ce rectangle par rapport à l'axe z (perpendiculaire à x et à y donc) et disposons les axes de façon à avoir:

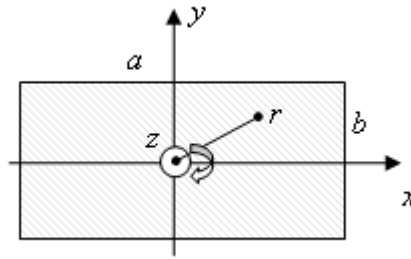


Figure: 26.27 - Recherche du moment d'inertie de la plaque selon un autre axe

Nous avons:

$$J_z = \int r^2 dm = \int (x^2 + y^2) dm = \int x^2 dm + \int y^2 dm = J_y + J_x \quad (26.60)$$

où r est donc dans le plan de x et y .

Avec:

$$\begin{aligned} J_x &= \int y^2 dm = \int_{-b/2}^{+b/2} y^2 \rho e a dy = \rho e a \int_{-b/2}^{+b/2} y^2 dy = \frac{\rho e a}{3} \left\{ \left(\frac{b}{2} \right)^3 - \left(-\frac{b}{2} \right)^3 \right\} = \frac{M}{12} b^2 \\ J_y &= \int x^2 dm = \int_{-a/2}^{+a/2} x^2 \rho e a dx = \rho e a \int_{-a/2}^{+a/2} x^2 dx = \frac{\rho e a}{3} \left\{ \left(\frac{a}{2} \right)^3 - \left(-\frac{a}{2} \right)^3 \right\} = \frac{M}{12} a^2 \end{aligned} \quad (26.61)$$

d'où:

$$J_z = \frac{M}{12} (a^2 + b^2) \quad (26.62)$$

Soit le moment d'inertie d'une plaque rectangulaire:

$$J_z = \frac{m}{12} (a^2 + b^2) \quad (26.63)$$

si la plaque est carrée de côté L :

$$J_z = \frac{M}{12} \underbrace{(a^2 + b^2)}_{2L^2} = \frac{ML^2}{6} \quad (26.64)$$

Nous allons montrer qu'il est dès lors possible de calculer le moment d'inertie du triangle équilatéral et rectangle.

Le moment d'inertie toujours par rapport au même axe, mais pour la moitié du carré, est donné par:

$$J_z = \frac{1}{2} \frac{ML^2}{6} = \frac{ML^2}{12} \quad (26.65)$$

Si le centre de gravité est posé sur le tiers de la médiane partant du centre de gravité du carré et que nous faisons usage du théorème de Steiner (cf. chapitre de Mécanique Classique), il vient:

$$J_z = \frac{ML^2}{12} + \frac{1}{3}\sqrt{2}LM \quad (26.66)$$

qui est donc le moment d'inertie d'un triangle équilatéral.

En procédant exactement de même pour un triangle rectangle de côtés a, b dont l'axe de rotation passe par le centre de masse G , il vient:

$$J_z = \frac{M(a^2 + b^2)}{24} + \frac{1}{3}\sqrt{2}LM \quad (26.67)$$

2.1.2. PYRAMIDE

Définition: La "pyramide" est un polyèdre qui a pour base un polygone et pour faces latérales des triangles réunis en un point appelé "sommet". La pyramide n'est donc pas dans le cas général un polyèdre régulier!

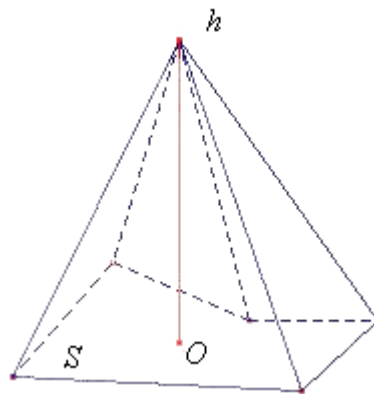


Figure: 26.28 - Exemple de pyramide

Considérons une surface $S(t)$ de la section de la pyramide avec le plan d'équation $z = t$, alors le volume V cherché est égal à:

$$V = \int_0^h S(t) dt \quad (26.68)$$

Nous parlons d'équation de plan alors qu'il n'y a pas de repères défini pour l'instant. Au fait, dans l'intégrale, t varie entre 0 et h . Cela sous-entend que nous prenons un repère centré en H (le pied de la hauteur de la pyramide), d'axe de la droite $\overline{OH}$ (la hauteur de la pyramide) orientée de O vers H (du pied de la hauteur vers le sommet). Les deux autres axes sont choisis quelconques dans le plan de la base de la pyramide.

Il nous faut préciser maintenant ce que vaut $S(t)$ en fonction de t :

Soit S l'aire de la base de la pyramide. La section de la pyramide par le plan d'équation $z = t$ se déduit par l'homothétie de centre O et de rapport t/h . Donc, l'intégrale s'écrit:

$$V = \int_0^h S \cdot \frac{t^2}{h^2} dt \quad (26.69)$$

Le fait d'avoir pris le carré de t/h provient de ce que chaque terme intérieur de S est le produit de deux termes (selon le calcul de la surface d'un triangle) chacun de rapport d'homothétie t/h .

Ainsi, nous avons:

$$V = \frac{S}{h^2} \int_0^h t^2 dt = \frac{S \cdot h}{3} \quad (26.70)$$

2.1.3. PRISME DROIT

Définition: Le "**prisme droit**" est un polyèdre dont les bases sont deux polygones égaux à côtés parallèles (elles ont la même surface!), les faces latérales étant des parallélogrammes. Donc, le prisme droit n'est pas un polyèdre régulier! Les deux faces parallèles et de même forme sont appelées les bases du prisme droit.

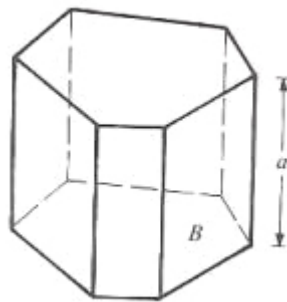


Figure: 26.29 - Exemple de prisme droit

Pour calculer le volume V d'un prisme droit, nous devons tout simplement multiplier l'aire de sa base B par sa hauteur h :

$$V = B \cdot h \quad (26.71)$$

Sa base est un polygone, c'est-à-dire qu'elle peut être un triangle, un quadrilatère, ou un pentagone... Il faut donc savoir calculer ces aires pour calculer le volume du prisme droit.

2.2. POLYÈDRES RÉGULIERS

Définitions:

D1. Un "**polyèdre régulier**" est constitué de faces toutes identiques et régulières.

D2. Un "**polyèdre convexe**" est tel que chaque point d'un segment de droite qui joint deux points quelconques appartient au polyèdre.

Les polyèdres réguliers sont au nombre de neufs, dont cinq sont convexes et étaient connus de Platon. Nous appelons parfois polyèdres réguliers uniquement les solides de Platon et ce sont ceux-ci qui vont nous intéresser ici.

Démontrons d'abord qu'il n'existe que cinq polyèdres réguliers convexes qui sont donc appelés "**les cinq solides platoniciens**" (les autres colonnes du tableau ci-dessous seront démontrées et expliquées un peu plus loin):

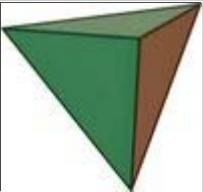
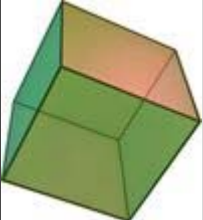
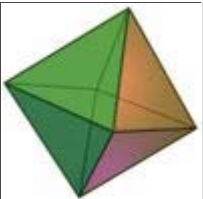
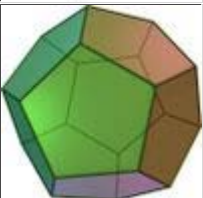
Nom (m,n)	Image	S	A	F	$F - A + S$
Tétraèdre (3,3)		4	6	4	2
Hexaèdre ou cube (4,3)		8	12	6	2
Octaèdre (3,4)		6	12	8	2
Dodécaèdre (5,3)		20	30	12	2
Icosaèdre (3,5)		12	30	20	2

Tableau: 26.1 - Cinq polyèdres réguliers

Démonstration:

Soient m le nombre de côtés de chaque face d'un polyèdre régulier, n le nombre d'arêtes qui se rencontrent en chaque sommet. Nous avons alors que chaque angle d'une face quelconque est donné par:

$$2\beta = \frac{\pi}{2} \left(2 - \frac{4}{m} \right) \quad (26.72)$$

Attention, c'est l'angle 2β qui définit donc l'angle d'une face et non pas β !

Ce qui découle de la figure suivante:

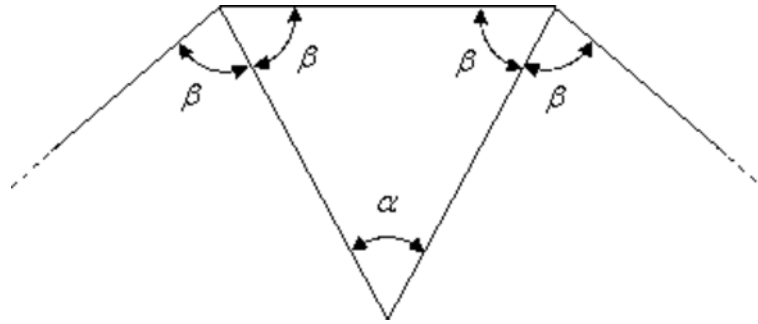


Figure: 26.30 - Angles entre les arêtes d'un polyèdre régulier

où nous avons:

$$\alpha = \frac{2\pi}{m} \quad (26.73)$$

et:

$$2\beta = \pi - \alpha = \pi - \frac{2\pi}{m} = \frac{\pi}{2} \left(2 - \frac{4}{m} \right) \quad (26.74)$$

Mais la somme des n angles groupés autour d'un sommet est plus petite que les n angles qui coupent un plan en parties égales (nous supposons cela intuitif par découpage)! Chacun d'eux est donc inférieur à:

$$\frac{2\pi}{n} \quad (26.75)$$

donc:

$$\frac{\pi}{2} \left(2 - \frac{4}{m} \right) < \frac{2\pi}{n} \quad (26.76)$$

d'où:

$$\frac{1}{m} + \frac{1}{n} > \frac{1}{2} \quad (26.77)$$

Les nombres m et n sont tous deux au moins égaux à 3 (le plus petit polygone étant le triangle). Il en résulte que les seuls cas possibles sont:

$$(m, n) = \{(3, 3), (3, 4), (4, 3), (3, 5), (5, 3)\} \quad (26.78)$$

□C.Q.F.D.

Notons maintenant F le nombre de faces, A le nombre d'arêtes et S le nombre de sommets. Alors, rappelons que nous avons démontré dans le chapitre de Théorie Des Graphes la "formule d'Euler" (ou "théorème de Descartes-Euler") telle que:

$$F - A + S = 2 \quad (26.79)$$

et celle-ci est bien évidemment valable aussi pour l'aplatissement d'un polyèdre dans le plan (et donc in extenso d'un polyèdre).

Remarque: La représentation sous forme d'un graphe de l'aplatissement d'un polyèdre est appelée "**diagramme de Schlegel**".

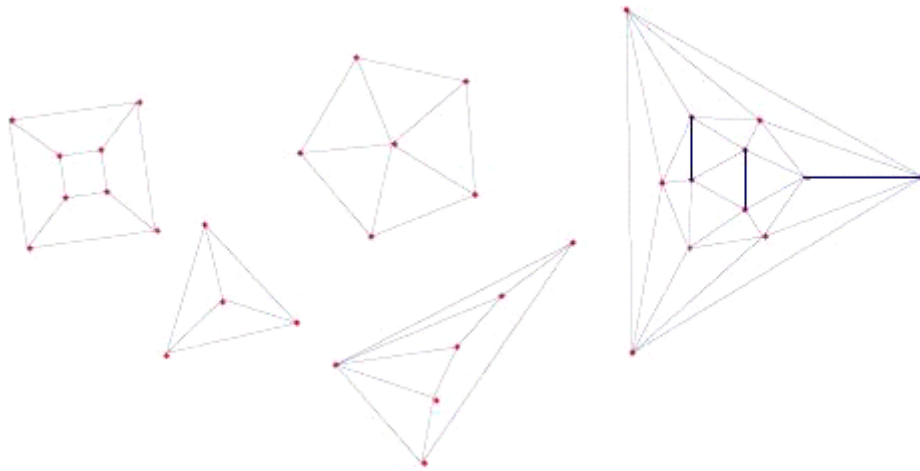


Figure: 26.31 - Exemples de diagrammes de Schlegel

Dans le cas des polyèdres réguliers, chaque face possède m arêtes de sorte que $m \cdot F$ est l'ensemble des arêtes des faces et comme chaque arête rencontre exactement deux faces, nous avons l'égalité (prendre un exemple pour s'en convaincre au cas où!):

$$m \cdot F = 2 \cdot A \quad (26.80)$$

et comme n est le nombre des arêtes qui se rencontrent en chaque sommet, et que chaque arête relie deux sommets, nous avons également:

$$m \cdot F = n \cdot S \quad (26.81)$$

Soit:

$$m \cdot F = n \cdot S = 2 \cdot A \quad (26.82)$$

En injectant dans la formule d'Euler, nous avons alors:

$$\frac{2A}{m} - A + \frac{2A}{n} = 2 \Rightarrow \frac{1}{m} + \frac{1}{n} = \frac{1}{A} + \frac{1}{2} > \frac{1}{2} \quad (26.83)$$

et nous retrouvons l'inégalité du théorème précédent. Reprenons notre calcul:

$$F - A + S = 2 = F - \frac{mF}{2} + \frac{mF}{n} = F \left(1 - \frac{m}{2} + \frac{m}{n} \right) \quad (26.84)$$

d'où nous tirons:

$$F = \frac{2}{1 - \frac{m}{2} + \frac{m}{n}} = \frac{2}{\frac{2n}{2n} - \frac{mn}{2n} + \frac{2m}{2n}} = \frac{4n}{2n - mn + 2m} \quad (26.85)$$

Nous pouvons maintenant entreprendre la classification des polyèdres réguliers.

Le tétraèdre $(m, n) = (3, 3)$:

$$F = \frac{4n}{2n - mn + 2m} = \frac{12}{6 - 9 + 6} = 4 \quad A = \frac{mF}{2} = \frac{12}{2} = 6 \quad S = \frac{mF}{n} = \frac{12}{3} = 4 \quad (26.86)$$

L'octaèdre $(m, n) = (3, 4)$:

$$F = \frac{16}{8 - 12 + 6} = 8 \quad A = \frac{24}{2} = 12 \quad S = \frac{24}{4} = 6 \quad (26.87)$$

L'hexaèdre $(m, n) = (4, 3)$ ou cube:

$$F = \frac{12}{6 - 12 + 8} = 6 \quad A = \frac{24}{2} = 12 \quad S = \frac{24}{3} = 8 \quad (26.88)$$

L'icosaèdre $(m, n) = (3, 5)$:

$$F = \frac{20}{10 - 15 + 6} = 20 \quad A = \frac{60}{2} = 30 \quad S = \frac{60}{5} = 12 \quad (26.89)$$

Le dodécaèdre $(m, n) = (5, 3)$:

$$F = \frac{12}{6 - 15 + 10} = 12 \quad A = \frac{60}{2} = 30 \quad S = \frac{60}{3} = 20 \quad (26.90)$$

ce qui termine notre classification.

2.2.1. TÉTRAÈDRE RÉGULIER

Nous avons montré que pour le tétraèdre $F = 4, m = 3$ et il est relativement aisé de deviner qu'un tel polyèdre est formé de 3 triangles équilatéraux identiques comme le montre la figure ci-dessous:

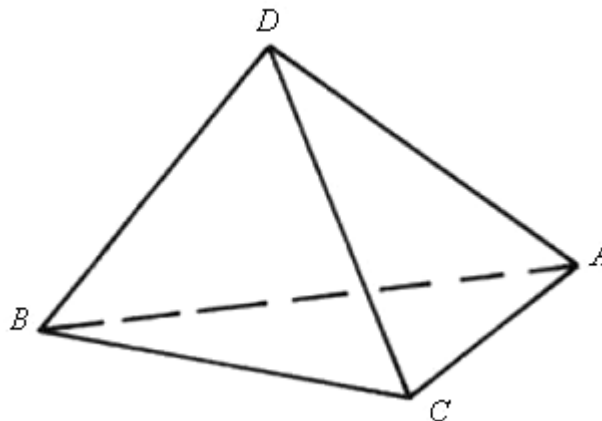


Figure: 26.32 - Exemple de tétraèdre régulier

Pour cela, commençons par étudier le triangle équilatéral suivant:

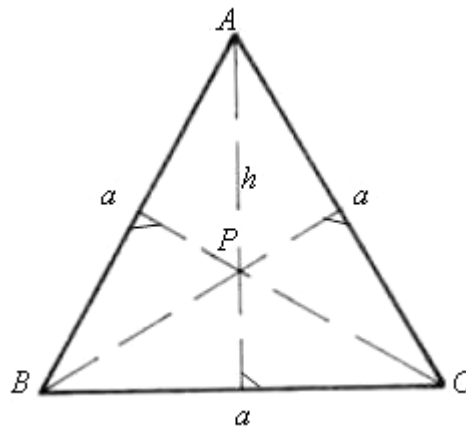


Figure: 26.33 - Triangle équilatéral de départ pour le calcul du volume

Dans ce triangle équilatéral, a est le côté, h la hauteur. Les médiatrices sont h , h' , h'' des côtés respectifs BC , AB , AC .

h et h' se coupent en un point P (barycentre). Par construction du triangle équilatéral, nous avons $PC = PB = PA$ (il suffit d'appliquer Pythagore pour le démontrer).

Nous avons par ailleurs démontré lors de notre étude du triangle, que le barycentre de celui-ci se situe toujours à $2/3$ de la hauteur de la médiane. Comme médianes et médiatrices sont confondues dans le cas du triangle équilatéral, nous avons alors $AP = 2/3 \cdot h$.

Maintenant, nous tirons une droite passant par le point P et perpendiculaire au plan dans lequel se trouve le triangle. Soit D un point sur cette droite, comme $PB = PC = PA = 2/3 \cdot h$ nous aurons bien sûr $BD = AD = CD$ (il suffit d'appliquer Pythagore à nouveau!).

Il ne nous reste donc plus qu'à nous arranger pour que $BD = a$ et nous aurons le tétraèdre régulier que nous voulions. Nous calculons alors:

$$BP^2 + PD^2 = BD^2 = a^2 \quad (26.91)$$

et donc:

$$PD^2 = a^2 - (2/3 \cdot h)^2 = a^2 - \frac{4}{9}h^2 = a^2 - \frac{4}{9}\left(a^2 - \left(\frac{1}{2}a\right)^2\right) = a^2 - \frac{4}{9}\left(\frac{3}{4}a^2\right) = \frac{2}{3}a^2 \quad (26.92)$$

donc:

$$H = \sqrt{\frac{2}{3}}a \quad (26.93)$$

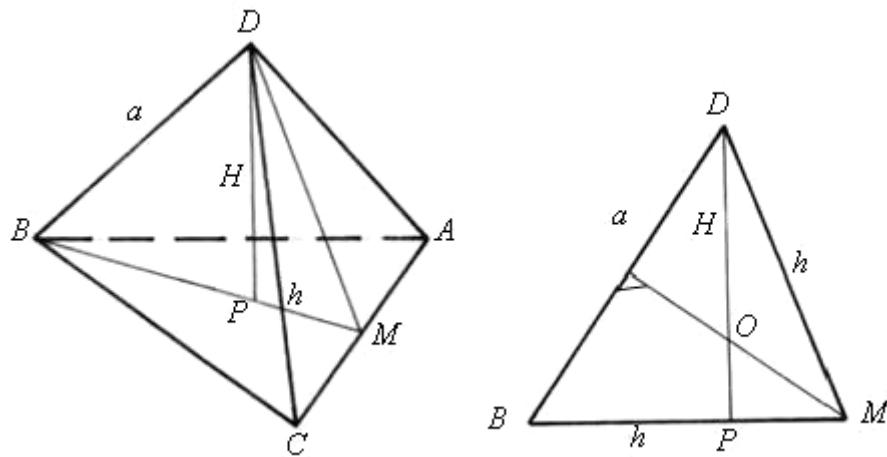


Figure: 26.34 - Synthèse de toutes les variables

La médiatrice de BD passant par M coupe H en un point O , qui n'est rien d'autre que le centre de la sphère circonscrite au tétraèdre. En effet, par construction, nous avons $OB = OC = OA$ et la médiatrice nous donne $OB = OD$.

Thalès nous donne également:

$$\frac{OD}{\frac{1}{2}a} = \frac{a}{PD} \Rightarrow OD = \frac{1}{2}a^2 \frac{1}{PD} = \frac{1}{2}a^2 \frac{1}{\frac{\sqrt{2}}{\sqrt{3}}a} = \frac{\sqrt{3}}{\sqrt{2}} \frac{a}{2} \quad (26.94)$$

et pour les développements qui suivront nous poserons $H = PD, OD = R$.

Calculons maintenant la surface totale. Elle sera nécessairement donnée par la surface d'une seule face multipliée par le nombre de faces, et comme nous avons démontré comment calculer la surface d'un triangle plus haut il vient immédiatement:

$$S = n \cdot \left(\frac{1}{2} \cdot \text{base} \cdot \text{hauteur} \right) = 4 \cdot \frac{1}{2} \cdot a \cdot h = 4 \cdot \frac{1}{2} \cdot a \cdot \frac{\sqrt{3}a}{2} = \sqrt{3}a^2 \quad (26.95)$$

Pour le volume, c'est tout aussi simple puisque nous avons démontré plus haut quel était celui d'une pyramide. Il vient alors immédiatement:

$$V = \frac{1}{3} \cdot \text{base} \cdot \text{hauteur} = \frac{1}{3} \cdot \frac{S}{4} \cdot H = \frac{1}{3} \cdot \frac{\sqrt{3}a^2}{4} \cdot \frac{\sqrt{2}}{\sqrt{3}} \cdot a = \frac{\sqrt{2}}{12} \cdot a^3 \quad (26.96)$$

2.2.2. HEXAÈDRE RÉGULIER (CUBE)

Le cube est le polyèdre régulier qui nous est le plus familier, il compte 6 faces et sa construction ne nécessite probablement pas d'être présentée.

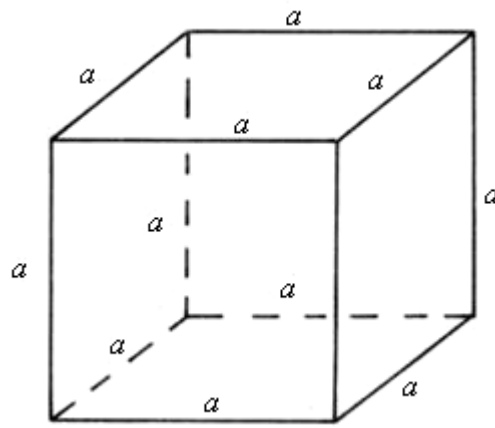


Figure: 26.35 - Exemple d'hexaèdre régulier (cube)

Puisque tous les côtés sont de longueur a , la surface est simplement donnée par la multiplication de la surface des 6 faces. Ainsi:

$$S = 6 \cdot a^2 \quad (26.97)$$

et le volume:

$$V = a^3 \quad (26.98)$$

2.2.3. OCTAÈDRE RÉGULIER

Nous avons montré que pour l'octaèdre $F = 8, m = 4$ et il est relativement aisé de deviner que l'octaèdre régulier est formé (par définition) de 8 triangles équilatéraux identiques.

Pour construire, et montrer qu'il est possible de construire, un tel polyèdre, nous posons comme précédemment que son côté vaut a .

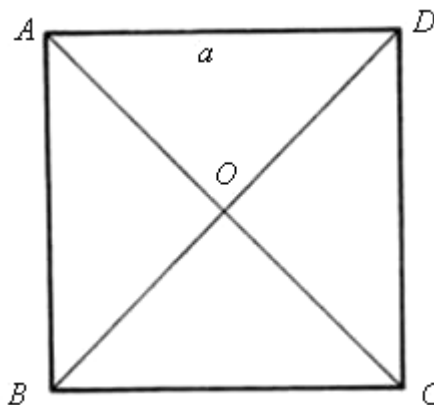


Figure: 26.36 - Exemple d'octaèdre régulier

Ensuite, nous notons O le point d'intersection des deux diagonales. Nous avons alors:

$$AC^2 = a^2 + a^2 = 2a^2 \Rightarrow AC = \sqrt{2}a \quad (26.99)$$

et:

$$OA = \frac{1}{2} AC = \frac{a}{\sqrt{2}} \quad (26.100)$$

Sur la droite perpendiculaire au plan qui contient notre carré, et passant par O , nous ajoutons deux sommets E, F à une distance que nous calculons comme suit :

$$OA^2 + OE^2 = AE^2 = a^2 \quad (26.101)$$

d'où nous tirons:

$$OE^2 = AE^2 - OA^2 = a^2 - \frac{1}{2}a^2 = \frac{1}{2}a^2 \quad (26.102)$$

Donc:

$$OE = \frac{a}{\sqrt{2}} = OA = -OF \quad (26.103)$$

Notre polyèdre est bien formé par huit triangles équilatéraux tous identiques. Chaque sommet compte quatre arêtes et quatre faces, ce qui nous permet d'affirmer qu'il est bien régulier et termine ainsi notre construction.

La surface de l'octaèdre régulier est:

$$S = n \cdot \frac{1}{2} \cdot a \cdot h = 8 \cdot \frac{1}{2} \cdot a \cdot \frac{\sqrt{3}a}{2} = 2\sqrt{3}a^2 \quad (26.104)$$

avec h étant la hauteur du triangle équilatéral de côté a que nous avons déjà calculée plus haut. Pour le volume, c'est encore basé sur celui de la pyramide. Ainsi:

$$V = 2 \cdot \frac{1}{3} \cdot \text{base} \cdot \text{hauteur} = 2 \cdot \frac{1}{3} \cdot a^2 \cdot \frac{a}{\sqrt{2}} = \frac{\sqrt{2}}{3} \cdot a^3 \quad (26.105)$$

Et nous supposons qu'il est évident pour le lecteur que notre octaèdre est inscrit dans une sphère de rayon R dont le centre est le point O . Pour R , nous avons:

$$R = OA = OE = \frac{a}{\sqrt{2}} \quad (26.106)$$

Montrons déjà maintenant que nous pouvons construire l'icosaèdre régulier à partir de l'octaèdre et que ce premier existe et est bien constructible.

Pour cela, nous allons d'abord considérer un repérage vectoriel des points suivants de l'octaèdre avec l'origine O placée au barycentre:

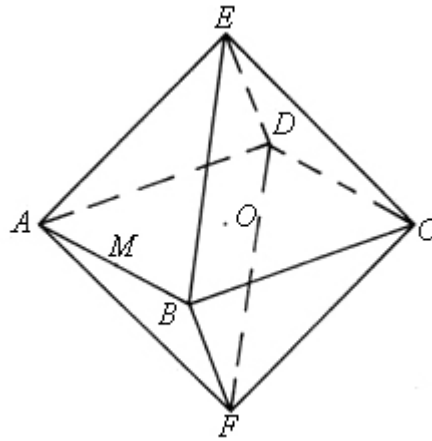


Figure: 26.37 - Exemple d'octaèdre

Nous avons alors:

$$\vec{A} = (R, 0, 0), \vec{B} = (0, R, 0), \vec{C} = (-R, 0, 0), \vec{D} = (0, -R, 0), \vec{E} = (0, 0, R), \vec{F} = (0, 0, -R) \quad (26.107)$$

Une fois ceci posé, considérons la figure suivante:

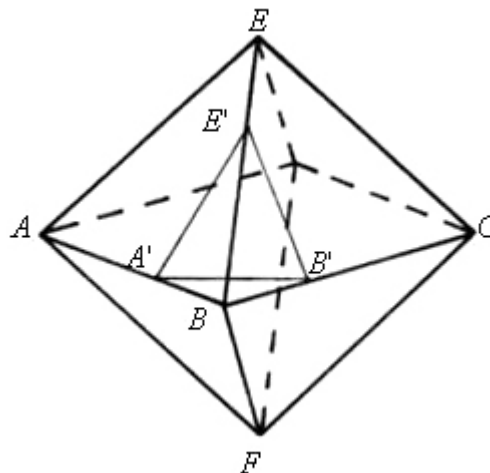


Figure: 26.38 - Recherche de la position du triangle équilatéral

Sur la figure ci-dessus, A' est un point qui part de A et qui arrive en B , et soit B' un point qui part de C et qui arrive en B , et pour finir E' un point qui part de B et arrive en E . Ces trois points partent en même temps et avancent à la même vitesse. Si nous suivons ces trois points, qui forment un triangle $A'B'E'$, nous sentons bien intuitivement qu'il existe un lieu tel que $A'B'E'$ soit un triangle équilatéral.

Déterminons ce lieu:

$$\begin{aligned} \overrightarrow{AB} &= \overrightarrow{OB} - \overrightarrow{OA} = (-R, R, 0) & \overrightarrow{OA'} &= \overrightarrow{OA} + \mu \overrightarrow{AB} = R(1 - \mu, \mu, 0) \\ \overrightarrow{CB} &= \overrightarrow{OB} - \overrightarrow{OC} = (R, R, 0) & \overrightarrow{OB'} &= \overrightarrow{OC} + \mu \overrightarrow{CB} = R(\mu - 1, \mu, 0) \\ \overrightarrow{BE} &= \overrightarrow{OE} - \overrightarrow{OB} = (0, -R, R) & \overrightarrow{OE'} &= \overrightarrow{OB} + \mu \overrightarrow{BE} = R(0, 1 - \mu, \mu) \end{aligned} \quad (26.108)$$

et donc:

$$\begin{aligned} \overrightarrow{A'B'} &= \overrightarrow{OB'} - \overrightarrow{OA'} = R(2\mu - 2, 0, 0) \\ \overrightarrow{A'E'} &= \overrightarrow{OE'} - \overrightarrow{OA'} = R(\mu - 1, 1 - 2\mu, \mu) \end{aligned} \quad (26.109)$$

et nous voulons:

$$\|\overrightarrow{A'B'}\| = \|\overrightarrow{A'E'}\| = \|\overrightarrow{B'E'}\| \quad (26.110)$$

Alors:

$$\|\overrightarrow{A'B'}\|^2 = R^2(2\mu - 2)^2 = \|\overrightarrow{A'E'}\|^2 = R^2[(\mu - 1)^2 + (1 - 2\mu)^2 + \mu^2] \quad (26.111)$$

Soit:

$$4\mu^2 - 8\mu + 4 = 6\mu^2 - 6\mu + 2 \quad (26.112)$$

Ce qui se simplifie en:

$$\mu^2 - \mu - 1 = 0 \quad (26.113)$$

et comme $0 \leq \mu \leq 1$, nous obtenons pour la résolution de ce polynôme du deuxième degré (cf. chapitre de Calcul Algébrique) la seule solution acceptable :

$$\mu = \frac{1}{2}(-1 + \sqrt{5}) \quad (26.114)$$

le lecteur remarquera peut-être qu'il s'agit de l'inverse du nombre d'or.

Selon la figure ci-dessous, si nous posons $\overrightarrow{OI'} = \overrightarrow{OE'} + \mu\overrightarrow{EA'}$ alors nous retrouvons à nouveau la même valeur pour μ (soit le lecteur le vérifiera lui-même, soit nous sur demande nous pouvons faire le détail des calculs) et idem pour tous les autres points:

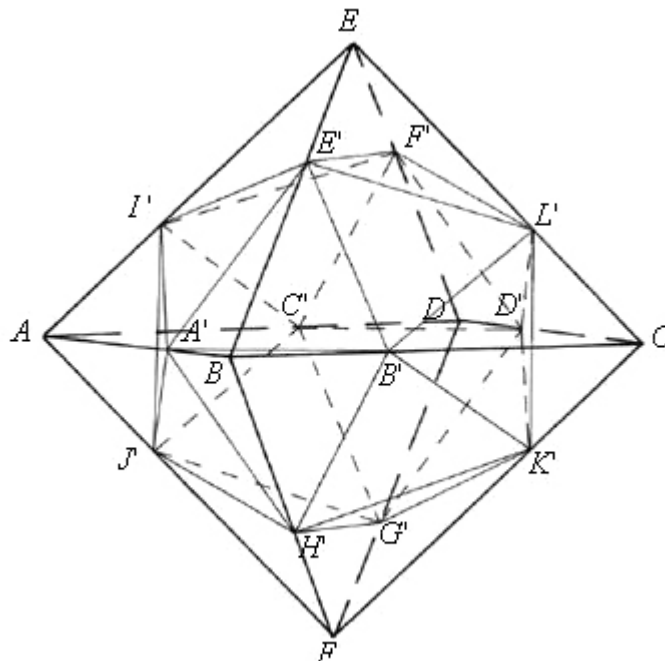


Figure: 26.39 - Construction de l'icosaèdre régulier

Notre nouveau polyèdre comporte donc une face par face de l'octaèdre et une face par arête de l'octaèdre. Nous avons ainsi vingt faces composées de triangles équilatéraux identiques. De plus, cinq

arêtes et cinq faces se rencontrent en chaque sommet. Nous obtenons alors un icosaèdre régulier.

Nous avons pour les coordonnées de chaque sommet (il faut bien observer que les sommets sont opposés par paires en une composante sur la figure):

$$\begin{aligned}
 \overrightarrow{A'} &= R(1-\mu, \mu, 0) & \overrightarrow{D'} &= R(1-\mu, -\mu, 0) \\
 \overrightarrow{B'} &= R(\mu-1, \mu, 0) & \overrightarrow{C'} &= R(1-\mu, -\mu, 0) \\
 \overrightarrow{E'} &= R(0, 1-\mu, \mu) & \overrightarrow{H'} &= R(0, 1-\mu, -\mu) \\
 \overrightarrow{F'} &= R(0, \mu-1, \mu) & \overrightarrow{G'} &= R(0, \mu-1, -\mu) \\
 \overrightarrow{I'} &= R(\mu, 0, 1-\mu) & \overrightarrow{L'} &= R(-\mu, 0, 1-\mu) \\
 \overrightarrow{J'} &= R(\mu, 0, \mu-1) & \overrightarrow{K'} &= R(-\mu, 0, 1-\mu)
 \end{aligned} \quad (26.115)$$

2.2.4. ICOSAÈDRE RÉGULIER

Nous avons vu précédemment comment construire l'icosaèdre régulier. Il existe donc bien.

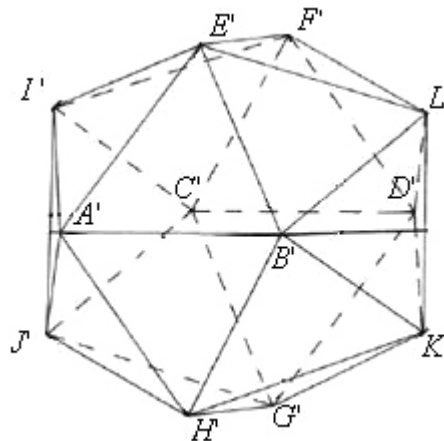


Figure: 26.40 - Représentation de l'icosaèdre régulier

Connaissant les coordonnées des différents sommets, calculons maintenant la surface et le volume de l'icosaèdre régulier.

Le calcul de la surface est simple puisqu'il s'agit de 20 triangles équilatéraux. Nous avons donc:

$$\overrightarrow{A'B'} = R(2\mu-2, 0, 0) \quad (26.116)$$

donc:

$$\|\overrightarrow{A'B'}\| = 2R(\mu-1) = a' \quad (26.117)$$

Donc:

$$S = 20 \cdot \frac{1}{2} a' \cdot h = 20 \cdot \frac{1}{2} a' \cdot \sqrt{a'^2 - \left(\frac{1}{2} a'\right)^2} = 20 \cdot \frac{1}{2} a' \cdot \sqrt{\frac{3}{4} a'^2} = 20 \cdot \frac{1}{2} a' \cdot \frac{\sqrt{3}}{2} a' \quad (26.118)$$

Donc:

$$S = 5 \cdot \sqrt{3} \cdot a^2 \quad (26.119)$$

Le calcul du volume est lui un peu plus subtil...

L'icosaèdre est construit autour du pentagone et de la section d'or comme nous avons pu nous en apercevoir lors de notre étude de l'octaèdre.

Si jamais le lecteur n'est pas convaincu voici une figure supplémentaire où nous voyons bien que chaque arête de l'icosaèdre est une arête d'un pentagone (*AFECB*, *LGHJK*, *DAJKC*, *DEGHA*, *BJILC*, *FELIH*...):

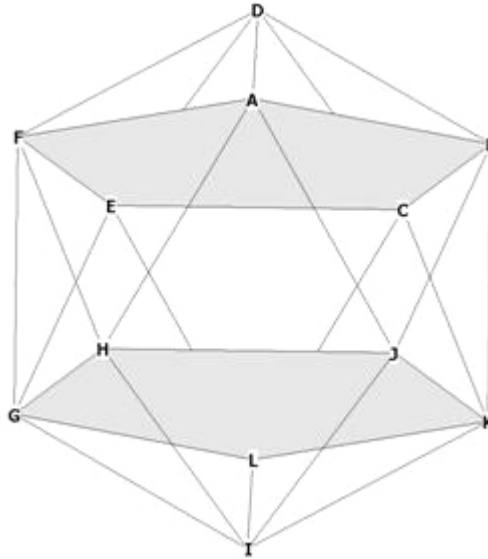


Figure: 26.41 - Pentagones dans l'icosaèdre régulier

Utilisant la méthode des pyramides, nous avons 20 triangles équilatéraux qui servent de base à une pyramide dont la hauteur va jusqu'à l'origine *O* de l'icosaèdre (ou l'origine confondue de la sphère inscrite ou circonscrite).

Prenons pour exemple la base *ABD* avec l'intersection des médiatrices se trouvant au point *M* comme représenté ci-dessous.

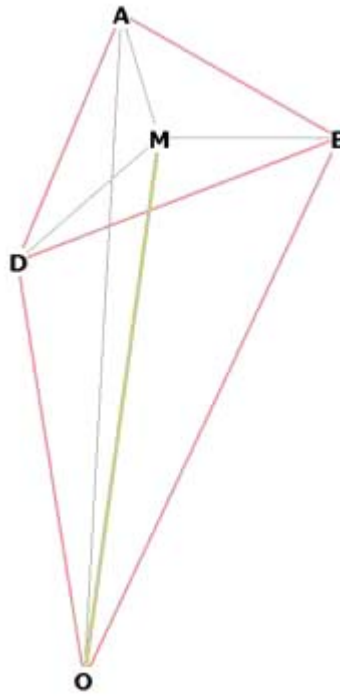


Figure: 26.42 - Représentation d'une des pyramides de l'icosaèdre régulier

Comme nous le savons, le volume de chaque pyramide est:

$$V_p = \frac{1}{3} \cdot b \cdot h \quad (26.120)$$

La surface b est dans notre situation celle du triangle équilatéral ADB et la hauteur h est le segment OM .

Si nous notons a le côté de triangle, alors la surface est donnée par:

$$S = \frac{a \cdot h}{2} = \frac{a \cdot \sqrt{a^2 - \left(\frac{1}{2}a\right)^2}}{2} = \frac{\sqrt{3}}{4} a^2 \quad (26.121)$$

Pour trouver h , nous savons par construction du point M que les triangles OMA , OMB , OMD sont des triangles rectangles.

Travaillons arbitrairement avec le triangle OMD . D'abord, déterminons la longueur DM . Nous avons démontré lors de notre étude des médiatrices de longueur H du triangle équilatéral (cf. chapitre de Géométrie Euclidienne) que DM vaut alors:

$$DM = \frac{2}{3} H \quad (26.122)$$

Or:

$$H = \sqrt{a^2 - \left(\frac{1}{2}a\right)^2} = \frac{\sqrt{3}}{2} a \quad (26.123)$$

Donc finalement:

$$DM = \frac{2}{3} \frac{\sqrt{3}}{2} a = \frac{\sqrt{3}}{3} a = \frac{1}{\sqrt{3}} a \quad (26.124)$$

Pour trouver h nous devons trouver la longueur $OD = r$ en termes de longueur des arêtes a de l'icosaèdre. Pour cela, nous devons reconnaître une des propriétés géométriques élémentaires de l'icosaèdre.

Avant d'aller plus loin, montrons une propriété du pentagone ci-dessous avec ses diagonales d et ses cotés c :

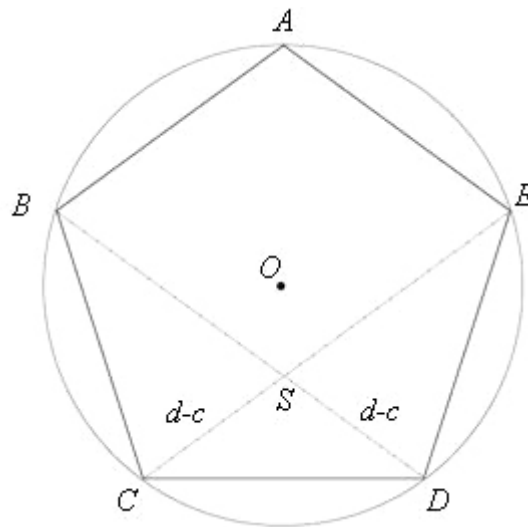


Figure: 26.43 - Parallélogramme dans un pentagone

$BSEA$ est un parallélogramme. Effectivement, la diagonale BD est parallèle au côté AE (par exemple, parce que tous deux sont perpendiculaires à l'axe de symétrie passant par OC). Comme S est sur BD , cela prouve que BS et AE sont parallèles. Nous montrons de la même manière que AB et SE sont parallèles.

Nous en déduisons que:

$$SD = BD - BS = BD - AE = d - c \quad (26.125)$$

et de même pour CS :

$$CS = EC - ES = EC - AB = d - c \quad (26.126)$$

Continuons..., nous avons l'égalité $\widehat{BAE} = \widehat{CSD}$. Comme de plus CD et BE sont parallèles, les triangles SCD et ABE sont semblables. Par conséquent, les rapports de distances entre leurs côtés sont conservés (Thalès):

$$\frac{AE}{BE} = \frac{SD}{CD} \quad (26.127)$$

d'où la relation:

$$\frac{c}{d} = \frac{d-c}{c} \quad (26.128)$$

Après quelques remaniements:

$$\begin{aligned} \frac{d}{c} - \frac{c}{d-c} = 0 &\Leftrightarrow \frac{d(d-c) - c \cdot c}{c(d-c)} = 0 \Leftrightarrow d^2 - dc - c^2 = 0 \\ &\Leftrightarrow \frac{d^2}{c^2} - \frac{d}{c} - 1 = 0 \end{aligned} \quad (26.129)$$

Si μ désigne maintenant le rapport d/c , la relation précédente devient:

$$\mu^2 - \mu - 1 = 0 \quad (26.130)$$

et μ étant strictement positif, nous avons déjà vu lors de notre étude de l'octaèdre que l'unique racine positive est le nombre d'or:

$$\mu = \frac{d}{c} = \frac{1+\sqrt{5}}{2} \quad (26.131)$$

Nous venons donc de montrer qu'une diagonale d'un pentagone est égale au nombre d'or multiplié par la longueur d'une arête de ce même pentagone.

Ainsi, nous avons dans les pentagones $AFECB$ et $LGHIK$ de notre icosaèdre:

$$FB = GK = \mu \cdot a \quad (26.132)$$

Remarquons également le rectangle $FBGK$ dont le barycentre est confondu avec celui de l'icosaèdre. Par ailleurs, FK et BG représentent par construction le diamètre de la sphère circonscrite à l'icosaèdre et donc $OF = OK$ en est le rayon r que nous allons chercher.

Nous avons:

$$FK^2 = FB^2 + BK^2 = \mu^2 a^2 + a^2 = (\mu^2 + 1) a^2 \quad (26.133)$$

Donc:

$$FK = a\sqrt{\mu^2 + 1} = 2 \cdot r \quad (26.134)$$

d'où:

$$r = \frac{a\sqrt{\mu^2 + 1}}{2} \quad (26.135)$$

Maintenant, nous pouvons calculer h :

$$h^2 = OM^2 = r^2 - DM^2 = \frac{a^2(\mu^2 + 1)}{4} - \frac{1}{3}a^2 = \frac{3(\mu^2 + 1) - 4}{12}a^2 = \frac{3\mu^2 - 1}{12}a^2 \quad (26.136)$$

Or:

$$\mu^2 = \left(\frac{1 + \sqrt{5}}{2} \right)^2 = \frac{6 + 2\sqrt{5}}{4} = \mu + 1 \quad (26.137)$$

puisque le nombre d'or est racine de l'équation $x^2 = x + 1$.

Soit:

$$(\mu^2)^2 = (\mu + 1)^2 = \mu^2 + 2\mu + 1 = \mu^2 + 2(\mu^2 - 1) + 1 = 3\mu^2 - 1 \quad (26.138)$$

Donc finalement:

$$h^2 = \frac{3\mu^2 - 1}{12}a^2 = \frac{\mu^4}{12}a^2 \quad (26.139)$$

et:

$$h = \frac{\mu^2}{2\sqrt{3}}a \quad (26.140)$$

Ainsi, le volume d'une pyramide de l'icosaèdre est:

$$V_{\text{pyr}} = \frac{1}{3} \cdot S \cdot h = \frac{1}{3} \cdot \frac{\sqrt{3}}{4}a^2 \cdot \frac{\mu^2}{2\sqrt{3}}a = \frac{\mu^2}{24}a^3 \quad (26.141)$$

Comme il y a 20 pyramides:

$$V = 20 \frac{\mu^2}{24}a^3 = \frac{5}{6}\mu^2a^3 \quad (26.142)$$

2.2.5. DODÉCAÈDRE RÉGULIER

Faute d'avoir trouvé dans la littérature une manière esthétique et simple de faisabilité de construction du dodécaèdre, nous nous en passerons pour l'instant (il est possible de vivre sans...).

Remarquons simplement que le dodécaèdre est composé de 12 pentagones réguliers et son volume est assimilable à celui d'un parallélépipède sur lequel nous avons posé sur chacune des faces une sorte de petit toit ce qui au final va donner les pentagones:

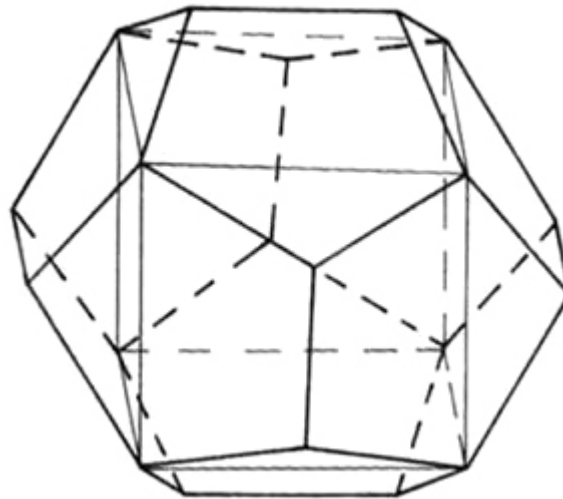


Figure: 26.44 - Dodécaèdre régulier

Pour notre étude du dodécaèdre, nous nous intéresserons uniquement à déterminer sa surface et son volume.

Pour cela, considérons dans un premier temps le pentagone régulier ci-dessous:

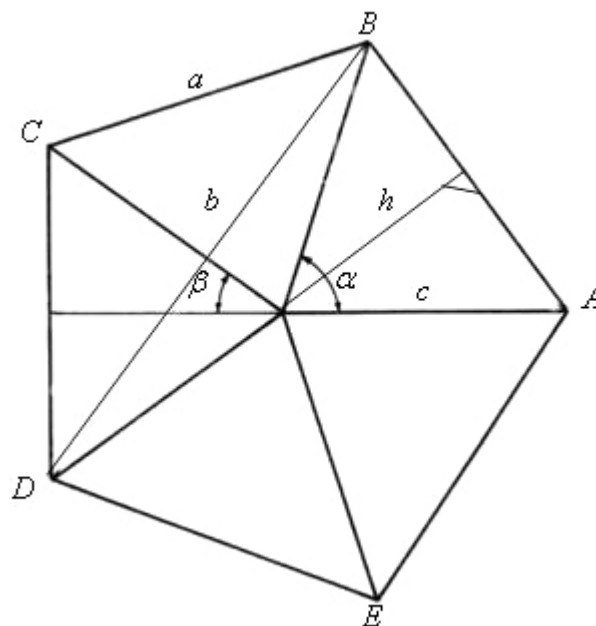


Figure: 26.45 - Pentagone régulier pour la recherche du volume du dodécaèdre régulier

Nous allons d'abord devoir déterminer la longueur de h et de b .

Rappelons d'abord que nous avons lors de notre étude l'icosaèdre déjà démontré que la diagonale d'un pentagone est reliée à la longueur de ses côtés par la relation:

$$b = a \cdot \mu \quad (26.143)$$

où μ est donc le nombre d'Or. Il nous reste alors à déterminer h .

Il est d'abord évident que $\beta = 2\pi/10 = \pi/5 = \alpha/2$ et que:

$$c \cdot \cos(\beta) = h \quad (26.144)$$

Or, deux informations nous manquent ici: l'angle et c . Commençons par déterminer combien vaut le cosinus sans utiliser la calculatrice (vous comprendrez pourquoi...).

Nous avons d'abord selon la relation (cf. chapitre de Trigonométrie) $\cos(2x) = 2\cos^2(x) - 1$:

$$\cos\left(4\frac{\pi}{5}\right) = 2\cos^2\left(2\frac{\pi}{5}\right) - 1 \quad (26.145)$$

Ce qui s'écrit aussi:

$$\cos\left(\pi - \frac{\pi}{5}\right) = 2\cos^2\left(2\frac{\pi}{5}\right) - 1 \quad (26.146)$$

Mais cela s'écrit également en utilisant toujours la même relation trigonométrique remarquable:

$$\cos\left(\pi - \frac{\pi}{5}\right) = 2\left[2\cos^2\left(\frac{\pi}{5}\right) - 1\right]^2 - 1 \quad (26.147)$$

Soit après simplification:

$$-\cos\left(\frac{\pi}{5}\right) = 8\cos^4\left(\frac{\pi}{5}\right) - 8\cos^2\left(\frac{\pi}{5}\right) + 1 \quad (26.148)$$

En faisant un changement de variable et en réarrangeant les différents termes:

$$8X^4 - 8X^2 + X + 1 = 0 \quad (26.149)$$

Nous avons -1 et 1/2 qui sont deux racines évidentes et nous obtenons donc (cf. chapitre de Calcul Algébrique):

$$(X+1)\left(X - \frac{1}{2}\right)\left(X^2 - \frac{1}{2}X - \frac{1}{4}\right) = 0 \quad (26.150)$$

Nous n'avons plus qu'à résoudre une simple équation du deuxième degré dont la solution est triviale en appliquant les méthodes vues dans le chapitre de Calcul Algébrique, et nous obtenons:

$$X = \frac{1 \pm \sqrt{5}}{4} \quad (26.151)$$

Soit en ne prenant que la seule solution admissible, nous avons alors:

$$X = \frac{\mu}{2} = \frac{1 + \sqrt{5}}{4} \quad (26.152)$$

nous retrouvons donc nombre d'Or là aussi! et ceci nous amène directement à écrire que:

$$h = \frac{1}{2} \mu \cdot c \quad (26.153)$$

Il nous reste à déterminer c . Nous avons:

$$\left(\frac{1}{2}a\right)^2 + h^2 = c^2 = \frac{1}{4}a^2 + \frac{1}{4}\mu^2 c^2 \Rightarrow c^2(4 - \mu^2) = a^2 \quad (26.154)$$

et comme $\mu^2 - \mu - 1 = 0$ nous avons:

$$\begin{aligned} (4 - \mu^2)(2 + \mu) &= 8 + 4\mu - 2\mu^2 - \mu^3 = 8 + 4\mu - 2\mu^2 - \mu(\mu + 1) = 8 + 3\mu - 3\mu^2 \\ &= 8 + 3(\mu - \mu^2) = 8 + 3(-1) = 5 \end{aligned} \quad (26.155)$$

et donc:

$$c^2 = \frac{1}{(4 - \mu^2)} a^2 = \frac{2 + \mu}{5} a^2 = \frac{\mu^2 + 1}{5} a^2 \quad (26.156)$$

d'où:

$$c = \frac{\sqrt{\mu^2 + 1}}{\sqrt{5}} a \quad (26.157)$$

Nous avons donc pour le calcul de la surface du pentagone, une surface composée de 12 pentagones dont chacun est composé d'un triangle de base a et de hauteur h .

$$\begin{aligned} S &= 12 \cdot 5 \cdot \frac{1}{2} \cdot a \cdot h = 30 \cdot a \cdot h = 30 \cdot a \cdot \frac{1}{2} \mu \cdot c = 15 \cdot a \cdot \mu \cdot c = 15 \cdot a \cdot \mu \cdot \frac{\sqrt{\mu^2 + 1}}{\sqrt{5}} a \\ &= 15 \cdot \mu \cdot \frac{\sqrt{\mu^2 + 1}}{\sqrt{5}} a^2 = 3 \cdot \mu \cdot a^2 \sqrt{5} \sqrt{\mu^2 + 1} \end{aligned} \quad (26.158)$$

Pour calculer le volume, nous allons faire usage de l'astuce mentionnée au début. C'est-à-dire de découper dans un premier temps le dodécaèdre en un parallélépipède de côté:

$$c = s \cdot \mu = s \cdot \frac{1 + \sqrt{5}}{2} \quad (26.159)$$

puisque le côté du parallélépipède est une diagonale du pentagone de côté s et de 6 petits toits (qui sont bien visibles sur la figure du dodécaèdre donnée précédemment).

Chaque petit toit selon deux vues différentes aura les longueurs suivantes (où nous retrouvons bien évidemment pour certaines arêtes celles des pentagones s ou encore les diagonales c de ceux-ci):

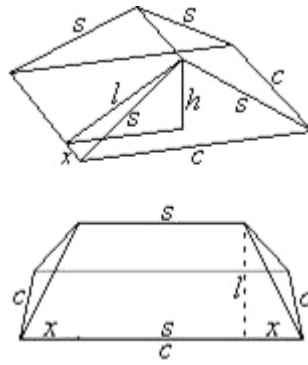


Figure: 26.46 - Autres éléments de volume du dodécaèdre régulier

Pour chaque petit toit, nous traitons à part les extrémités en les séparant et en les réunissant. Finalement, nous avons deux morceaux à traiter: la partie majeure du toit visible à gauche sur la figure ci-dessous et la partie secondaire du toit à droite sur la figure qui n'est autre que la réunion des extrémités du toit:

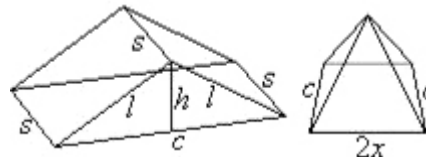


Figure: 26.47 - Décomposition

Il nous faut donc déterminer x et l et h puisque c et s nous sont déjà connus.

D'abord, nous voyons trivialement que:

$$x = \frac{c-s}{2} \quad (26.160)$$

Du théorème de Pythagore, nous avons alors:

$$\begin{aligned} s^2 &= l^2 + x^2 = l^2 + \left(\frac{c-s}{2}\right)^2 \\ l^2 &= h^2 + \left(\frac{c}{2}\right)^2 \end{aligned} \quad (26.161)$$

En combinant ces deux relations, nous avons:

$$s^2 - \left(\frac{c-s}{2}\right)^2 = h^2 + \left(\frac{c}{2}\right)^2 \quad (26.162)$$

Il vient alors:

$$\begin{aligned}
h^2 &= s^2 - \left(\frac{c-s}{2}\right)^2 - \left(\frac{c}{2}\right)^2 = s^2 - \frac{1}{4}(c^2 - 2cs + s^2) - \frac{c^2}{4} = \frac{3}{4}s^2 - \frac{c^2}{2} + \frac{cs}{2} \\
&= \frac{3}{4}s^2 - \frac{\left(\frac{1+\sqrt{5}}{2}s\right)^2}{2} + \frac{\left(\frac{1+\sqrt{5}}{2}s\right)s}{2} = \frac{3}{4}s^2 - \frac{(1+\sqrt{5})^2 s^2}{8} + \frac{(1+\sqrt{5})s^2}{4} \\
&= s^2 \left(\frac{3}{4} - \frac{(1+\sqrt{5})^2}{8} + \frac{1+\sqrt{5}}{4} \right) = s^2 \left(\frac{6-1-2\sqrt{5}-5+2+2\sqrt{5}}{8} \right) = s^2 \left(\frac{2}{8} \right) = \frac{1}{4}s^2
\end{aligned} \tag{26.163}$$

Donc:

$$h = \frac{s}{2} \tag{26.164}$$

Nous pouvons maintenant calculer le volume V_P de l'ensemble des 6 petits toits:

$$\begin{aligned}
V_P &= 6 \left(\frac{1}{3} 2x \cdot c \cdot h + \frac{c \cdot s \cdot h}{2} \right) = 6 \left(\frac{1}{3} 2 \frac{c-s}{2} \cdot c \cdot \frac{s}{2} + \frac{c \cdot s}{2} \frac{s}{2} \right) \\
&= 6 \left(\frac{1}{3} \left(s \cdot \frac{1+\sqrt{5}}{2} - s \right) \cdot s \cdot \frac{1+\sqrt{5}}{2} \cdot \frac{s}{2} + s \cdot \frac{1+\sqrt{5}}{2} \frac{s^2}{4} \right) \\
&= s^3 \left(\left(\frac{1+\sqrt{5}}{2} - 1 \right) \frac{1+\sqrt{5}}{2} + 3 \frac{1+\sqrt{5}}{4} \right) = s^3 \left(\frac{(1+\sqrt{5}-2)(1+\sqrt{5})}{4} + \frac{3+3\sqrt{5}}{4} \right) \\
&= s^3 \left(\frac{(1+\sqrt{5}-2)(1+\sqrt{5}) + 3 + 3\sqrt{5}}{4} \right) = s^3 \left(\frac{(1+\sqrt{5}-2+ \sqrt{5} + 5 - 2\sqrt{5}) + 3 + 3\sqrt{5}}{4} \right) \\
&= s^3 \left(\frac{7+3\sqrt{5}}{4} \right)
\end{aligned} \tag{26.165}$$

Donc, le volume total du dodécaèdre est finalement le volume des 6 petits toits sommé au volume du parallélépipède central:

$$\begin{aligned}
V &= c^3 + s^3 \left(\frac{7+3\sqrt{5}}{4} \right) = \left(\frac{1+\sqrt{5}}{2}s \right)^3 + s^3 \left(\frac{7+3\sqrt{5}}{4} \right) = s^3 \left[\left(\frac{7+3\sqrt{5}}{4} \right) + \left(\frac{1+\sqrt{5}}{2} \right)^3 \right] \\
&= s^3 \left[\left(\frac{7+3\sqrt{5}}{4} \right) + \frac{1}{8}(1+\sqrt{5})^3 \right] = s^3 \left[\frac{7+3\sqrt{5}}{4} + \frac{1}{8}(1+3\sqrt{5}+3 \cdot 5+5\sqrt{5}) \right] \\
&= s^3 \left[\frac{7+3\sqrt{5}}{4} + \frac{1}{8}(16+8\sqrt{5}) \right] = s^3 \left[\frac{7+3\sqrt{5}}{4} + \frac{1}{4}(8+4\sqrt{5}) \right] = s^3 \frac{15+7\sqrt{5}}{4}
\end{aligned} \tag{26.166}$$

2.3. CORPS DE RÉVOLUTIONS

Définition: Un "**corps de révolution**" est un volume que nous obtenons en faisant tourner une courbe 2D autour d'un axe.

Il existe donc autant de corps de révolution que de type de courbe fermée ou non que nous pouvons faire tourner autour d'un axe.

Voyons avant d'aller plus loin la méthode générale permettant de déterminer l'aire d'un corps de révolution. C'est-à-dire la surface du corps engendrée par la rotation d'une courbe de longueur finie autour d'un axe:

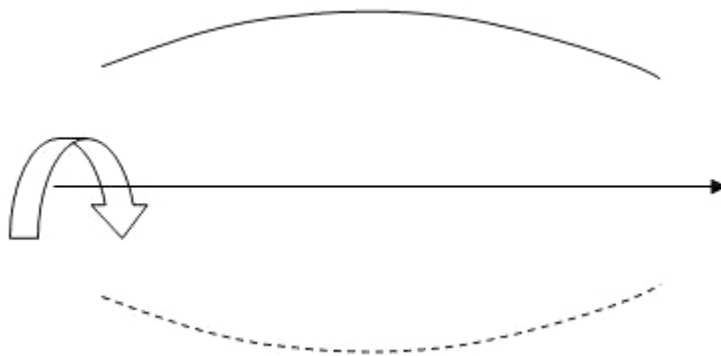


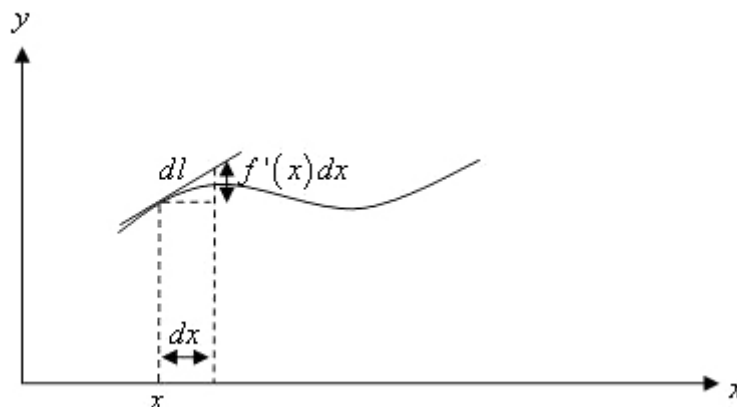
Figure: 26.48 - Exemple de construction d'un corps de révolution

Pour cela, nous remarquons que lorsque la courbe est donnée par une équation $y = f(x)$ nous remarquons par Pythagore (voir figure ci-dessous) que l'élément de longueur dl vérifie (relation que nous avons déjà rencontrée dans d'autres chapitres du site):

$$dl^2 = dx^2 + dy^2 = dx^2 + (f'(x))^2 dx^2 \quad (26.167)$$

donc :

$$dl = \sqrt{1 + (f'(x))^2} dx \quad (26.168)$$



Ainsi, l'élément de surface engendré par la rotation de l'élément de longueur dl est donné par:

$$dS = 2\pi |f(x)| \cdot dl = 2\pi |f(x)| \sqrt{1 + (f'(x))^2} dx \quad (26.169)$$

L'aire de la surface de révolution engendrée par une fonction $f : [a, b] \rightarrow \mathbb{R}$ continûment différentiable est donc donnée par la relation:

$$S = 2\pi \int_a^b |f(x)| \sqrt{1 + (f'(x))^2} dx \quad (26.170)$$

2.3.1. CYLINDRE

Définition: Un "cylindre" est une surface engendrée par une droite qui se déplace parallèlement à une direction fixe en rencontrant une courbe plane fixe, dont le plan coupe la direction donnée.

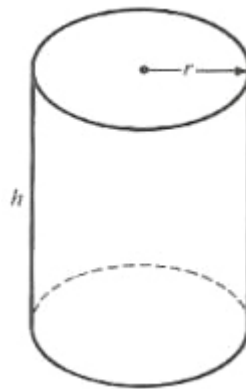


Figure: 26.49 - Exemple de cylindre

Le volume d'un cylindre de révolution de rayon $x = r$ et de hauteur égale à h se calcule par la méthode des disques en sachant que la surface d'un cercle (disque) vaut πr^2 :

$$V = \pi r^2 \int_0^h dy = \pi r^2 y \Big|_0^h = \pi r^2 h \quad (26.171)$$

La surface d'un cylindre étant simplement la somme de la surface des deux disques de base et du sommet et de la surface du rectangle plié de hauteur h et de longueur $2\pi r$:

$$S = 2(\pi r^2) + 2\pi r h \quad (26.172)$$

Calculons maintenant le moment d'inertie d'un cylindre plein par rapport à son axe de symétrie vertical (axe de révolution):

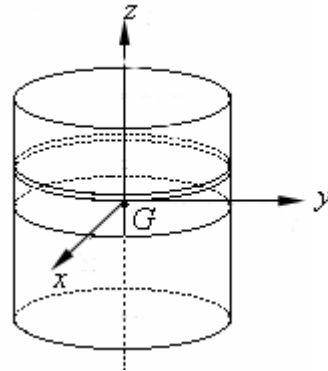


Figure: 26.50 - Recherche d'un des moments d'inertie du cylindre plein

Nous avons:

$$dm = \rho dV = \rho 2\pi r h dr \quad (26.173)$$

Donc:

$$J_z = \int r^2 dm = \int \rho r^2 2\pi r h dr = 2\pi \rho h \int_0^R r^3 dr = 2\pi \rho h \frac{R^4}{4} = \rho \pi R^2 h \cdot \frac{R^2}{2} = M \frac{R^2}{2} \quad (26.174)$$

Soit maintenant G le centre de gravité du cylindre, G_z coïncide avec l'axe de révolution du cylindre.

Les axes G_y et G_x jouent des rôles identiques. Les moments d'inertie J_{G_y} et J_{G_x} par rapport à ces axes sont donc égaux et s'écrivent:

$$J_{G_y} = \int (x^2 + z^2) dm = J_{G_x} = \int (y^2 + z^2) dm = J \quad (26.175)$$

d'où:

$$J_{G_y} + J_{G_x} = 2J_{G_x} = 2J_{G_y} = \int (x^2 + z^2 + y^2 + z^2) dm = \int (x^2 + y^2 + 2z^2) dm \quad (26.176)$$

d'où:

$$J_{G_y} = \frac{1}{2} \int (x^2 + y^2 + 2z^2) dm = \frac{1}{2} \int (x^2 + y^2) dm + \int z^2 dm \quad (26.177)$$

La première intégrale est en fait le moment d'inertie du cylindre par rapport à l'axe G_z et nous savons qu'elle vaut:

$$J_{G_z} = J_z = M \frac{R^2}{2} \quad (26.178)$$

La deuxième intégrale se calcule facilement en découpant le cylindre en tranches d'épaisseur dz perpendiculaires à l'axe G_z . La masse de la tranche élémentaire est $dm = \pi r^2 \rho dz$ soit:

$$\int z^2 dm = 2 \int_0^{h/2} \pi r^2 \rho z^2 dz = 2\pi r^2 \rho \frac{h^3}{24} = M \frac{h^2}{12} \quad (26.179)$$

Le moment d'inertie d'un cylindre par rapport à un axe perpendiculaire à son axe de révolution s'écrit

donc:

$$J_{Gy} = J_y = J_{Gx} = J_x = M \left(\frac{R^2}{4} + \frac{h^2}{12} \right) \quad (26.180)$$

Calculons maintenant le moment d'inertie d'un tube ou d'un cylindre creux d'épaisseur non nulle (toujours donné dans les formulaires techniques): le moment d'inertie d'un tube par rapport à son axe de révolution est un grand classique du traitement du moment d'inertie du cylindre. Ainsi, considérons un tube de rayon extérieur r_e et de rayon intérieur r_i . Comme (cf. chapitre de Mécanique Classique):

$$\sum_{i=1}^n J_i = J \quad (26.181)$$

Il vient dès lors que le moment d'inertie d'un tube peut être vu comme le moment d'inertie du cylindre de rayon égal au rayon externe du tube diminué du moment d'inertie du cylindre de rayon égal au rayon interne du tube. Ainsi:

$$\begin{aligned} J_{tube} &= J_{re} - J_{ri} = \frac{M_e r_e^2}{2} - \frac{M_i r_i^2}{2} = \frac{1}{2} (\rho \pi r_e^2 h r_e^2 - \rho \pi r_i^2 h r_i^2) = \frac{\rho \pi h}{2} (r_e^4 - r_i^4) \\ &= \frac{\rho \pi h}{2} (r_e^2 - r_i^2)(r_e^2 + r_i^2) = \frac{\rho \pi h (r_e^2 - r_i^2)}{2} (r_e^2 + r_i^2) = \frac{M}{2} (r_e^2 + r_i^2) \end{aligned} \quad (26.182)$$

et si $r_e^2 \cong r_i^2 = \bar{r}$, il vient dès lors la relation classique disponible dans nombre de formulaires de physique:

$$J_{tube} = \frac{M 2\bar{r}^2}{2} = M \bar{r}^2 \quad (26.183)$$

2.3.2. CÔNE

Définition: Un "cône" est une surface engendrée par une droite mobile, passant par un point fixe et s'appuyant sur une courbe fixe; solide déterminé par cette surface.

Le volume d'un cône de révolution de rayon à la base r et de hauteur égale à h se calcule également par la méthode des disques.

La droite passant par les points $(r, 0)$ (extrémité de la base du cône) et $(0, h)$ (sommet du cône) est:

$$y = ax + b = \frac{-h}{r} x + h \quad (26.184)$$

La rotation de cette droite par rapport à l'axe des y donne le volume du cône:

$$V = \pi \int_a^b x^2 dy = \pi \int_0^h r^2 \left(1 - \frac{y}{h} \right)^2 dy = \pi r^2 \frac{1}{3} \left[\left(1 - \frac{y}{h} \right)^3 \right]_0^h (-h) = \frac{-\pi r^2 h}{3} (0 - 1) = \frac{\pi r^2 h}{3} \quad (26.185)$$

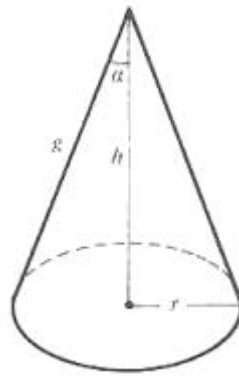


Figure: 26.51 - Exemple de cône

Pour calculer la surface latérale d'un cône, nous allons paramétrer la droite qui part du sommet du cône de $(0,0)$ à (r, h) ce qui est donc une paramétrisation différente que celle pour le volume (cela permet de simplifier un peu les calculs). Nous avons alors:

$$f(x) = \frac{r}{h}x \quad (26.186)$$

et donc:

$$\begin{aligned} S_l &= 2\pi \int_0^h |f(x)| \sqrt{1 + (f'(x))^2} dx = 2\pi \int_0^h \left| \frac{r}{h}x \right| \sqrt{1 + \frac{r^2}{h^2}} dx = 2\pi \frac{r}{h} \sqrt{1 + \frac{r^2}{h^2}} \int_0^h x dx \\ &= 2\pi \frac{r}{h} \sqrt{1 + \frac{r^2}{h^2}} \frac{1}{2} x^2 \Big|_0^h = \pi \frac{r}{h} \sqrt{1 + \frac{r^2}{h^2}} h^2 = \pi r h \sqrt{1 + \frac{r^2}{h^2}} = \pi r \sqrt{h^2 + r^2} = \pi r^2 \sqrt{1 + \frac{h^2}{r^2}} \end{aligned} \quad (26.187)$$

Donc, la surface totale du cône (base + surface latérale) est alors:

$$S = S_b + S_l = \pi r^2 + \pi r^2 \sqrt{1 + \frac{h^2}{r^2}} = \pi r^2 \left(1 + \sqrt{1 + \frac{h^2}{r^2}} \right) \quad (26.188)$$

Calculons maintenant le moment d'inertie d'un cône par rapport à son axe de révolution:

Pour ce calcul, nous allons utiliser la valeur du moment d'inertie du cylindre $J_{z, \text{cyl}}$ et considérer le cône comme un empilement de cylindres infinitésimaux.

Donc:

$$\begin{aligned} J_z &= \int dJ_{z, \text{cyl}} = \int \frac{r^2}{2} dm = \int \frac{r^2}{2} (\rho \pi r^2 dh) = \frac{\pi \rho}{2} \int r^4 dh = \frac{\pi \rho}{2} \frac{H}{R} \int_0^R r^4 dr = \frac{\pi \rho H R^5}{10 R} \\ &= \left(\rho \frac{\pi R^2 H}{3} \right) \frac{3 R^2}{10} = \frac{3}{10} M R^2 \end{aligned} \quad (26.189)$$

2.3.3. SPHÈRE

Définition: La "**sphère**" est le volume engendré par la rotation d'un disque (ou cercle) de rayon r autour de son centre de gravité.

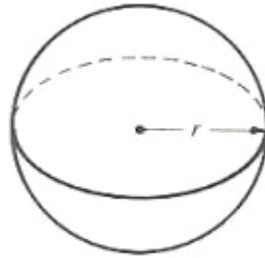


Figure: 26.52 - Exemple de sphère

Nous pouvons voir une sphère de rayon R , comme une surface qui est formée par la rotation d'un demi-cercle autour de son grand axe. La fonction décrivant un demi-cercle étant:

$$y = f(x) = \sqrt{R^2 - x^2} \quad (26.190)$$

La sphère peut donc être disséquée comme une somme de disques d'épaisseur Δx . Les demi-disques étant perpendiculaires à l'axe des abscisses et de largeur Δx à la position x_k (voir figure ci-dessous).

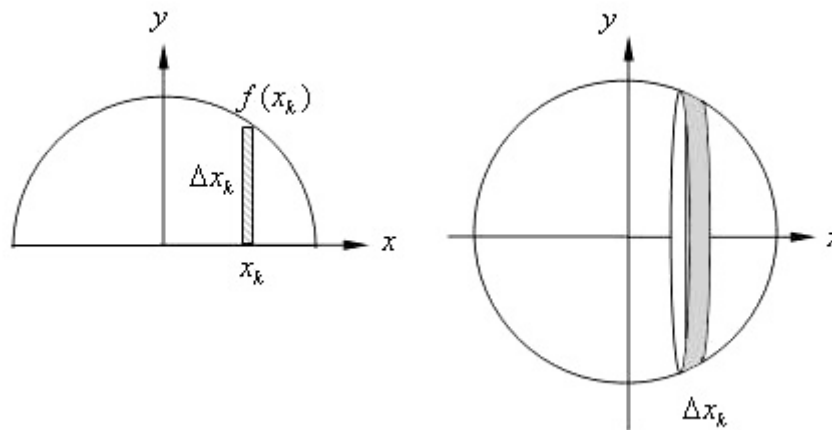


Figure: 26.53 - Calcul du volume de la sphère par décomposition en disques

Nous avons ainsi:

$$x_k = k\Delta x \quad (26.191)$$

Le volume d'un disque (cylindre) étant donné par (en passant à la limite):

$$dV_k = \pi r_k^2 dx \quad (26.192)$$

et le rayon r_k étant donné par la fonction:

$$r_k = f(x_k) = \sqrt{R^2 - x_k^2} \quad (26.193)$$

nous avons alors:

$$dV_k = \pi \left(\sqrt{R^2 - x_k^2} \right)^2 dx = \pi (R^2 - x_k^2) dx \quad (26.194)$$

En intégrant entre $x = [-R, R]$, nous avons alors le volume de la sphère:

$$V = \int_{-R}^R \pi (R^2 - x_k^2) dx = \pi \int_{-R}^R (R^2 - x^2) dx = \pi \left(R^2 x - \frac{x^3}{3} \right) \Big|_{-R}^R \quad (26.195)$$

Nous pouvons également prendre les bornes entre $x = [0, R]$ cela revenant au même à un facteur 2 près:

$$V = 2\pi \left(R^2 x - \frac{x^3}{3} \right) \Big|_0^R = 2\pi \left(R^3 - \frac{R^3}{3} \right) \quad (26.196)$$

Après simplification, nous obtenons pour le volume:

$$V = \frac{4}{3} \pi R^3 \quad (26.197)$$

L'expression de la surface étant donnée par dérivant par rapport à l'élément engendrant la surface, nous obtenons ainsi (c'est un peu limite comme raisonnement mais bon...):

$$S = \frac{dV}{dR} = 4\pi R^2 \quad (26.198)$$

Il existe une autre manière d'aborder ces calculs un peu plus rigoureusement. Effectivement, dans le chapitre de Calcul Différentiel Et Intégral nous avons introduit le concept de Jacobien qui permet de changer les variables d'intégration en fonction du système de coordonnées sur lequel nous travaillons (pour la définition détaillée des termes le lecteur doit se reporter au chapitre de Calcul Différentiel Et Intégral):

$$\iiint_D f(x, y) dx dy = \iiint_D F(u, v, w) |\det(J)| du dv dw \quad (26.199)$$

et nous avons démontré que le jacobien en coordonnées sphériques est :

$$|\det(J)| = r^2 \sin \theta \quad (26.200)$$

Donc au même titre que $dx dy dz$ est un élément différentiel de volume, nous pouvons convertir cet élément en coordonnées sphériques et faire in extenso apparaître un élément différentiel de volume de la sphère de rayon r . Il suffit ensuite d'intégrer correctement pour avoir le volume de toute la sphère.

Dès lors, nous avons :

$$\begin{aligned}
 V &= \iiint_D r^2 \sin \theta dr d\theta d\phi = \int_0^R r^2 dr \int_0^\pi \sin \theta d\theta \int_0^{2\pi} d\phi = \frac{r^3}{3} \Big|_0^R \cdot (-\cos \theta) \Big|_0^\pi \cdot \phi \Big|_0^{2\pi} \\
 &= \frac{R^3}{3} \cdot (-\cos(\pi) - (-\cos(0))) \cdot 2\pi = \frac{2}{3} \pi R^3 (-(-1) - (-1)) = \frac{4}{3} \pi R^3
 \end{aligned}
 \tag{26.201}$$

et pour la surface (pour laquelle le rayon est constant) :

$$\begin{aligned}
 S &= R^2 \iint \sin \theta d\theta d\phi = R^2 \int_0^\pi \sin \theta d\theta \int_0^{2\pi} d\phi = R^2 \cdot (-\cos \theta) \Big|_0^\pi \cdot \phi \Big|_0^{2\pi} \\
 &= R^2 \cdot (-\cos(\pi) - (-\cos(0))) \cdot 2\pi = 2\pi R^2 (-(-1) - (-1)) = 4\pi R^2
 \end{aligned}
 \tag{26.202}$$

Au besoin, on peut trouver l'élément de surface de manière géométrique plutôt que de passer par le jacobien car ce dernier n'est pas très pédagogique dans les petites classes...

Alors en se rappelant que dans le chapitre de Trigonométrie, nous avons démontré que la longueur d'un arc de cercle est donnée par:

$$dL = R d\alpha \tag{26.203}$$

Alors, il devient très aisé de compléter la figure suivante:

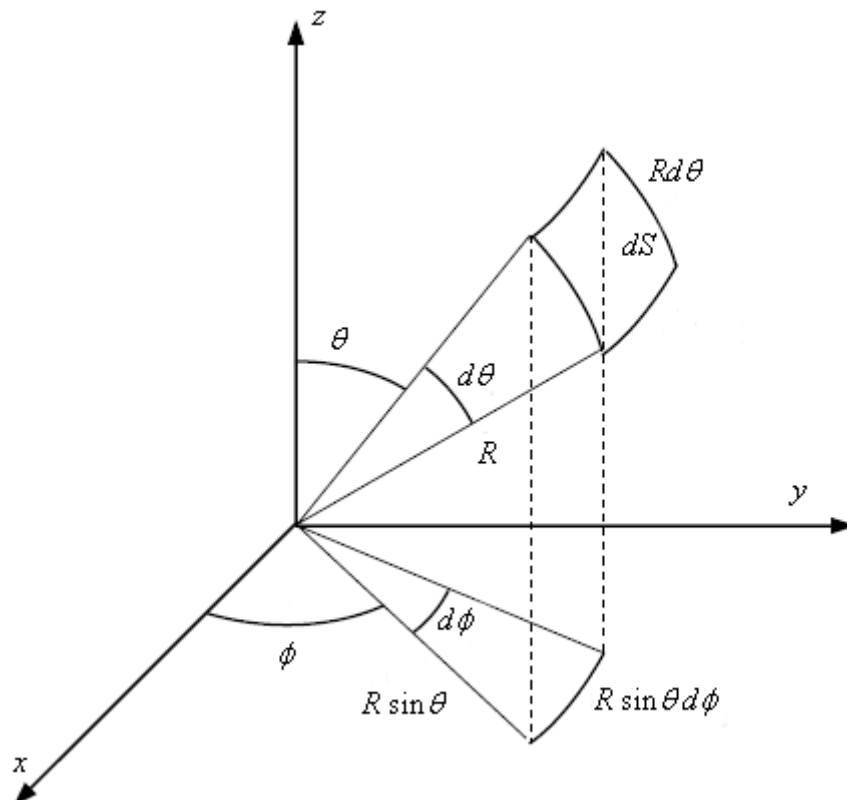


Figure: 26.54 - Représentation d'un élément de surface en coordonnées sphériques

et nous voyons alors immédiatement que:

$$dS = R^2 \sin \theta d\theta d\phi \quad (26.204)$$

ce qui est quand même plus sympa...

Calculons maintenant le moment d'inertie d'une boule pleine homogène de masse M et de masse volumique ρ . Pour cela, la boule présentant une symétrie maximale, il est plus commode de calculer d'abord le moment d'inertie polaire (cf. chapitre de Mécanique Classique), puis de déterminer le moment d'inertie axial à partir de ce premier:

$$J_o = \int r^2 dm = \int r^2 \rho dV = \int r^2 \rho (4\pi r^2 dr) = 4\pi\rho \int_0^R r^4 dr = 4\pi\rho \frac{1}{5} R^5 = \frac{4}{3} \pi \rho \frac{3}{5} R^3 R^2 = \frac{3}{5} MR^2 \quad (26.205)$$

Comme J_x, J_y, J_z sont égaux par symétrie de la boule, il vient:

$$J_o = \frac{1}{2}(3J_x) \Rightarrow J_x = J_y = J_z = \frac{2}{3} J_o = \frac{2}{5} MR^2 \quad (26.206)$$

2.3.4. TORE

Définition: Un "tore" est la surface engendrée par la rotation d'un cercle c de rayon r autour d'une droite située dans son plan, mais ne passant pas par son centre.

Soit l'équation d'un demi-cercle de centre $(0,c)$:

$$x^2 + (y - c)^2 = r^2 \quad (26.207)$$

Afin d'écrire y sous la forme d'une fonction de x , isolons y dans cette équation:

$$x^2 + (y - c)^2 = r^2 \Leftrightarrow (y - c)^2 = r^2 - x^2 \Leftrightarrow y - c = \pm \sqrt{r^2 - x^2} \Leftrightarrow y = c \pm \sqrt{r^2 - x^2} \quad (26.208)$$

Le cercle est alors constitué des graphes des deux fonctions suivantes:

- Demi-cercle supérieur:

$$f_1 : x \mapsto c + \sqrt{r^2 - x^2} \quad (26.209)$$

- Demi-cercle inférieur:

$$f_2 : x \mapsto c - \sqrt{r^2 - x^2} \quad (26.210)$$

Le volume demandé est la différence entre les volumes engendrés par la rotation des surfaces (surfaces définies par l'aire comprise entre la fonction du cercle concerné et l'axe des abscisses compris entre $x = r, x = -r$) dans l'espace autour de l'axe des abscisses.

En appliquant la relation d'intégration des corps de révolution:

$$\begin{aligned}
V &= V_1 - V_2 = \pi \int_{-r}^r \left(c + \sqrt{r^2 - x^2} \right)^2 dx - \pi \int_{-r}^r \left(c - \sqrt{r^2 - x^2} \right)^2 dx \\
&= \pi \int_{-r}^r \left(\left(c + \sqrt{r^2 - x^2} \right)^2 - \left(c - \sqrt{r^2 - x^2} \right)^2 \right) dx \\
&= \pi \int_{-r}^r \left(\left(c^2 + 2c\sqrt{r^2 - x^2} + r^2 - x^2 \right) - \left(c^2 - 2c\sqrt{r^2 - x^2} + r^2 - x^2 \right) \right) dx \quad (26.211) \\
&= \pi \int_{-r}^r \left(c^2 + 2c\sqrt{r^2 - x^2} + r^2 - x^2 - c^2 + 2c\sqrt{r^2 - x^2} - r^2 + x^2 \right) dx \\
&= \pi \int_{-r}^r 4c\sqrt{r^2 - x^2} dx = 4\pi c \int_{-r}^r \sqrt{r^2 - x^2} dx
\end{aligned}$$

Calculons cette dernière intégrale par la substitution classique $x = r \sin t$ donc:

$$dx = (r \sin t)' dt = r \cos t dt \quad (26.212)$$

si $x = -r$:

$$-r = r \sin t \Rightarrow \sin t = -1 \Rightarrow t = -\frac{\pi}{2} \quad (26.213)$$

si $x = r$:

$$r = r \sin t \Rightarrow \sin t = 1 \Rightarrow t = \frac{\pi}{2} \quad (26.214)$$

donc:

$$\begin{aligned}
V &= 4\pi c \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sqrt{r^2 - r^2 \sin^2 t} r \cos t dt = 4\pi c \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sqrt{r^2(1 - \sin^2 t)} r \cos t dt \\
&= 4\pi c \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} r \sqrt{\cos^2 t} r \cos t dt = 4\pi c r^2 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^2 t dt \quad (26.215)
\end{aligned}$$

Linéarisons cette expression en utilisant à nouveau les relations trigonométriques (formule de Carnot):

$$\begin{aligned}
V &= 4\pi c r^2 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{2} (1 + \cos 2t) dt = 2\pi c r^2 \left(\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} dt + \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos 2t dt \right) \\
&= 2\pi c r^2 \left(\left[t \right]_{-\frac{\pi}{2}}^{\frac{\pi}{2}} + \frac{1}{2} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 2 \cos 2t dt \right) = 2\pi c r^2 \left(\frac{\pi}{2} - \left(-\frac{\pi}{2} \right) + \frac{1}{2} \left[\sin 2t \right]_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \right) \quad (26.216) \\
&= 2\pi c r^2 \left(\pi + \frac{1}{2} (\sin \pi - \sin(-\pi)) \right) = 2\pi c r^2 \left(\pi + \frac{1}{2} (0 - 0) \right) = 2c\pi^2 r^2
\end{aligned}$$

Donc, le volume d'un tore de "rayon mineur" r et de "rayon majeur" c est donné par:

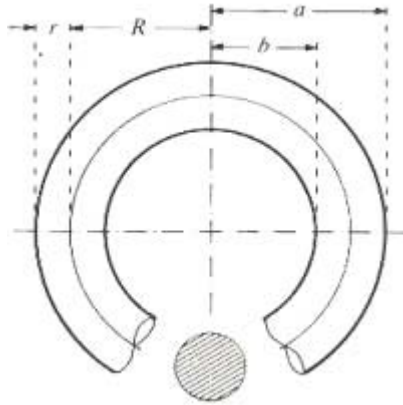
$$V = 2c\pi^2 r^2 \quad (26.217)$$

Adapté à la figure ci-dessous (prise de la littérature):

$$V = 2R\pi^2 r^2 \quad (26.218)$$

et la surface (par dérivation de l'élément génération de surface):

$$S = \frac{dV}{dr} = 4Rr\pi^2 \quad (26.219)$$



Le moment d'inertie du tore relativement à son axe de révolution se calcule de la manière suivante:

Soit le volume du tore (démontré précédemment) noté:

$$V = 2\pi^2 a^2 b \quad (26.220)$$

La densité volumique du tore est donnée par (masse sur volume):

$$\frac{dm}{dV} = \frac{M}{2\pi^2 a^2 b} \Rightarrow dm = \frac{M}{2\pi^2 a^2 b} dV \quad (26.221)$$

En coordonnées cylindriques, nous avons:

$$dV = r dr d\theta dz \quad (26.222)$$

d'où:

$$dm = \frac{M}{2\pi^2 a^2 b} dV = \frac{M}{2\pi^2 a^2 b} r dr d\theta dz \quad (26.223)$$

Le moment d'inertie est alors donné par:

$$\begin{aligned}
 J &= \int r^2 dm = 2 \left(\frac{M}{2\pi a^2 b} \int_0^{2\pi} d\theta \int_{-a}^{+a} r^3 dr \int_0^{+\sqrt{a^2 - (r-b)^2}} dz \right) = \frac{M}{\pi a^2 b} \int_0^{2\pi} d\theta \int_{-a}^{+a} r^3 dr \sqrt{a^2 - (r-b)^2} \\
 &= \frac{M}{\pi a^2 b} \int_0^{2\pi} d\theta \int_{-a}^{+a} r^3 \sqrt{a^2 - (r-b)^2} dr = \frac{2\pi M}{\pi a^2 b} \int_{-a}^{+a} r^3 \sqrt{a^2 - (r-b)^2} dr = \\
 &= \frac{2M}{\pi a^2 b} \int_{-a}^{+a} r^3 \sqrt{a^2 - (r-b)^2} dr
 \end{aligned} \tag{26.224}$$

Posons $s = r - b$, dès lors:

1. Les bornes d'intégration deviennent dès lors $-a$, $+a$ puisque nous ramenons tous les points d'intégration à l'origine en posant $s = r - b$

2. Trivialement, puisque $r = s + b$ nous avons donc $ds = dr$

Ce qui nous donne:

$$\begin{aligned}
 J_z &= \frac{2M}{\pi a^2 b} \int_{-a}^{+a} r^3 \sqrt{a^2 - (r-b)^2} dr = \frac{2M}{\pi a^2 b} \int_{-a}^{+a} (s+b)^3 \sqrt{a^2 - s^2} ds \\
 &= \frac{2M}{\pi a^2 b} \int_{-a}^{+a} (s^3 + 3s^2b + 3sb^2 + b^3) \sqrt{a^2 - s^2} ds
 \end{aligned} \tag{26.225}$$

Comme nous avons démontré dans le chapitre de Calcul Différentiel Et Intégrale que l'intégrale avec deux bornes symétriques d'une fonction impaire (produit d'une fonction paire et impaire) est nulle, les intégrales de:

$$s^3 \sqrt{a^2 - s^2} ds, 3sb^2 \sqrt{a^2 - s^2} ds \tag{26.226}$$

sont nulles.

Nous avons donc à calculer:

$$J_z = \frac{2M}{\pi a^2 b} \int_{-a}^{+a} (3s^2b + b^3) \sqrt{a^2 - s^2} ds \tag{26.227}$$

Posons maintenant $s = a \cdot \sin(t)$ et donc $ds = a \cos(t) dt$. Il vient donc:

$$J_z = \frac{2M}{\pi a^2 b} \int_{-\pi/2}^{+\pi/2} (3a^2 \sin^2(t)b + b^3) \sqrt{a^2 - a^2 \sin^2(t)} a \cos(t) dt \tag{26.228}$$

Or, comme:

$$\cos^2(t) + \sin^2(t) = 1 \Leftrightarrow a^2 \cos^2(t) + a^2 \sin^2(t) = a^2 \Rightarrow \sqrt{a^2 - a^2 \sin^2(t)} = |a \cos(t)| \tag{26.229}$$

Donc:

$$\begin{aligned}
 J_z &= \frac{2M}{\pi a^2} \int_{-\pi/2}^{+\pi/2} (3a^2 \sin^2(t) + b^2) |a \cos(t)| a \cos(t) dt = \frac{2M}{\pi a^2} \int_{-\pi/2}^{+\pi/2} (3a^2 \sin^2(t) + b^2) a^2 \cos^2(t) dt \\
 &= \frac{2M}{\pi} \int_{-\pi/2}^{+\pi/2} (3a^2 \sin^2(t) + b^2) \cos^2(t) dt = \frac{2M}{\pi} \int_{-\pi/2}^{+\pi/2} (3a^2 \sin^2(t) \cos^2(t) + b^2 \cos^2(t)) dt \\
 &= \frac{2M}{\pi} \int_{-\pi/2}^{+\pi/2} (3a^2 (1 - \cos^2(t)) \cos^2(t) + b^2 \cos^2(t)) dt \\
 &= \frac{2M}{\pi} \int_{-\pi/2}^{+\pi/2} (3a^2 \cos^2(t) - 3a^2 \cos^4(t) + b^2 \cos^2(t)) dt
 \end{aligned}
 \tag{26.230}$$

Soit (cf. chapitre de Calcul Différentiel Et Intégral):

$$\begin{aligned}
 \int_{-\pi/2}^{+\pi/2} 3a^2 \cos^2(t) dt &= 3a^2 \int_{-\pi/2}^{+\pi/2} \cos^2(t) dt = \frac{1}{2} 3a^2 (\cos(t) \sin(t) + t) \Big|_{-\pi/2}^{+\pi/2} = 3a^2 \frac{\pi}{2} \\
 \int_{-\pi/2}^{+\pi/2} b^2 \cos^2(t) dt &= b^2 \int_{-\pi/2}^{+\pi/2} \cos^2(t) dt = \frac{b^2}{2} (\cos(t) \sin(t) + t) \Big|_{-\pi/2}^{+\pi/2} = b^2 \frac{\pi}{2} \\
 \int_{-\pi/2}^{+\pi/2} 3a^2 \cos^4(t) dt &= 3a^2 \int_{-\pi/2}^{+\pi/2} \cos^4(t) dt = 3a^2 \left(\frac{1}{4} \cos^3(t) \sin(t) + \frac{3}{8} \cos(t) \sin(t) + \frac{3}{8} t \right) \Big|_{-\pi/2}^{+\pi/2} = 3a^2 \frac{3\pi}{8}
 \end{aligned}
 \tag{26.231}$$

Donc finalement:

$$J_z = \frac{2M}{\pi} \left(-3a^2 \frac{3\pi}{8} + b^2 \frac{\pi}{2} + 3a^2 \frac{\pi}{2} \right) = M \left(\frac{3}{4} a^2 + b^2 \right) \tag{26.232}$$

2.3.5. ELLIPSOÏDE

Définitions:

D1. Un "**ellipsoïde**" est une surface du second degré de l'espace euclidien à trois dimensions. Il fait donc partie des quadriques (cf. chapitre de Géométrie Analytique)

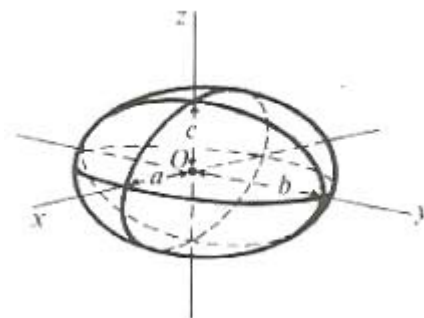


Figure: 26.55 - Exemple d'ellipsoïde

D2. Un "**ellipsoïde de révolution**" est un solide engendré par la révolution d'une ellipse autour de l'un de ses axes.

Pour calculer le volume délimité par l'ellipsoïde, prenons l'équation que nous avons déterminée lors de notre étude des coniques (cf. chapitre de Géométrie Analytique):

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \quad (26.233)$$

Remarque: Dans le cas où seuls deux paramètres aux dénominateurs sont égaux, l'ellipsoïde peut être engendré par la rotation d'une ellipse autour d'un de ses axes. Il s'agit alors de l'ellipsoïde de révolution définie juste précédemment et parfois appelé "sphéroïde".

La section par un plan parallèle au plan Oyz et se trouvant à la distance x de ce dernier, donne l'ellipse:

$$\frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 - \frac{x^2}{a^2} \quad (26.234)$$

ou:

$$\frac{y^2}{b^2 \left(1 - \frac{x^2}{a^2}\right)} + \frac{z^2}{c^2 \left(1 - \frac{x^2}{a^2}\right)} = \frac{y^2}{\left(b \sqrt{1 - \frac{x^2}{a^2}}\right)^2} + \frac{z^2}{\left(c \sqrt{1 - \frac{x^2}{a^2}}\right)^2} = 1 \quad (26.235)$$

avec pour demi-axes:

$$b_1 = b \sqrt{1 - \frac{x^2}{a^2}} \quad c_1 = c \sqrt{1 - \frac{x^2}{a^2}} \quad (26.236)$$

Mais comme nous l'avons démontré, la surface d'une ellipse vaut $\pi b_1 c_1$. Par conséquent:

$$S = \pi b c \left(1 - \frac{x^2}{a^2}\right) \quad (26.237)$$

Le volume de l'ellipsoïde est alors égal à:

$$V = \pi b c \int_{-a}^a \left(1 - \frac{x^2}{a^2}\right) dx = \pi b c \left(x - \frac{x^3}{3a^2}\right) \Big|_{-a}^a = \frac{4}{3} \pi a b c \quad (26.238)$$

Donc:

$$V = \frac{4}{3} \pi a b c \quad (26.239)$$

et si $a = b = c = R$, nous retrouvons l'expression du volume d'une sphère:

$$V = \frac{4}{3} \pi R^3 \quad (26.240)$$

Remarque: Le calcul du moment d'inertie d'un ellipsoïde est très important en astrophysique puisqu'un grand nombre d'étoiles ou de planètes en rotation sur elles-mêmes de par leur déformation à l'équateur à cause de la force centrifuge se voient déformées en première approximation en un tel volume.

Pour un ellipsoïde, définissons C comme étant le moment d'inertie le long de l'axe c , A le moment d'inertie le long de l'axe a et B le moment d'inertie le long de l'axe b .

Pour commencer, considérons le moment d'inertie le long de l'axe c que nous assimilerons à l'axe z . Dès lors, en coordonnées cartésiennes, nous avons:

$$J_z = C = \int r^2 dm = \int \rho r^2 dV = \rho \int r^2 dV = \rho \int (x^2 + y^2) dx dy dz \quad (26.241)$$

En faisant la substitution suivante, nous sous-entendons que l'intégrale précédente est une normalisation d'un ellipsoïde:

$$\begin{aligned} x' &= \frac{x}{a} \\ y' &= \frac{y}{b} \\ z' &= \frac{z}{c} \end{aligned} \quad (26.242)$$

ce qui nous donne pour notre intégrale (nous transformons donc ainsi le volume V de l'ellipsoïde en le volume V' d'une sphère):

$$C = \rho \int \left[(ax')^2 + (by')^2 \right] abcdx' dy' dz' \quad (26.243)$$

Nous pouvons maintenant passer des coordonnées cartésiennes aux coordonnées sphériques (cf. chapitre de Calcul Vectoriel) sans oublier d'utiliser le Jacobien (cf. chapitre de Calcul Différentiel Et Intégral) que nous avons démontré comme valant en coordonnées sphériques:

$$|\det(J)| = r^2 \sin \theta \quad (26.244)$$

Donc (nous utilisons les primitives usuelles démontrées dans le chapitre de Calcul Différentiel Et Intégral):

$$\begin{aligned}
C &= abc \rho \int_0^1 \int_0^{2\pi} \int_0^\pi (a^2 r^2 \cos^2 \phi \sin^2 \theta + b^2 r^2 \sin^2 \phi \sin^2 \theta) r^2 \sin \theta d\theta d\phi dr \\
&= abc \rho \int_0^1 \int_0^{2\pi} \int_0^\pi (a^2 \cos^2 \phi + b^2 \sin^2 \phi) r^4 \sin^3 \theta d\theta d\phi dr \\
&= abc \rho \int_0^1 \int_0^{2\pi} (a^2 \cos^2 \phi + b^2 \sin^2 \phi) r^4 \left(-\frac{1}{3} \cos \theta \sin^2 \theta - \frac{2}{3} \cos \theta \right) \Big|_0^\pi d\phi dr \\
&= \frac{4}{3} abc \rho \int_0^1 \int_0^{2\pi} (a^2 \cos^2 \phi + b^2 \sin^2 \phi) r^4 d\phi dr \\
&= \frac{4}{3} abc \rho \int_0^1 [a^2 \pi + b^2 \pi] r^4 dr = \frac{4}{3} abc \pi \rho (a^2 + b^2) \int_0^1 r^4 dr \\
&= \frac{4}{15} abc \pi \rho (a^2 + b^2) r^5 \Big|_0^1 = \frac{4}{15} \pi \rho abc (a^2 + b^2)
\end{aligned} \tag{26.245}$$

en y insérant pour l'ellipsoïde:

$$\rho = \frac{M}{V} = \frac{M}{\frac{4}{3}\pi abc} \tag{26.246}$$

nous obtenons alors:

$$C = \frac{1}{5} M (a^2 + b^2) \tag{26.247}$$

et par symétrie, nous avons les résultats triviaux suivants:

$$\begin{aligned}
A = J_x &= \frac{1}{5} M (b^2 + c^2) \\
B = J_y &= \frac{1}{5} M (a^2 + c^2) \\
C = J_z &= \frac{1}{5} M (a^2 + b^2)
\end{aligned} \tag{26.248}$$

La matrice d'inertie (cf. chapitre de Mécanique Classique) est alors:

$$(J_o)_{(\mathcal{F}, \mathcal{J}, \mathcal{K})} = \begin{pmatrix} \frac{1}{5} M (b^2 + c^2) & 0 & 0 \\ 0 & \frac{1}{5} M (a^2 + c^2) & 0 \\ 0 & 0 & \frac{1}{5} M (a^2 + b^2) \end{pmatrix} \tag{26.249}$$

2.3.6. PARABOLOÏDE

Définition: Un "paraboloïde" est un solide engendré par la révolution d'une parabole autour de son foyer:

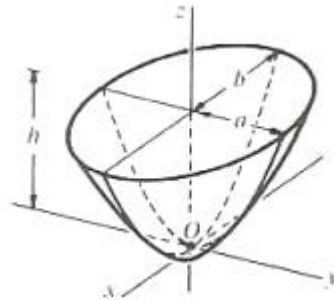


Figure: 26.56 - Exemple de paraboloïde

La méthode de calcul du volume du paraboloïde à base elliptique est exactement la même que celle pour la pyramide à la différence que l'équation d'une parabole est du type $y = \alpha x^2$ et que nous avons aussi $h = \alpha a^2$. Dès lors, nous avons évidemment $x = a(y)$. Le carré de la fonction nous amène à écrire:

$$a(y) = a\sqrt{\frac{y}{h}} \Leftrightarrow a(y) = a\sqrt{\frac{\alpha x^2}{\alpha a^2}} = x \quad (26.250)$$

et idem pour $b(y)$. Dès lors:

$$V = \int_0^h S(y) dy = \int_0^h \pi \cdot a(y) \cdot b(y) dy = \pi \int_0^h a\sqrt{\frac{y}{h}} \cdot b\sqrt{\frac{y}{h}} dy = \pi \frac{ab}{h} \int_0^h y dy = \frac{1}{2} \pi abh \quad (26.251)$$

2.3.7. TONNEAU À SECTION CIRCULAIRE

Maintenant regardons pour le plaisir un volume très connu par les viticulteurs (et pas seulement!):

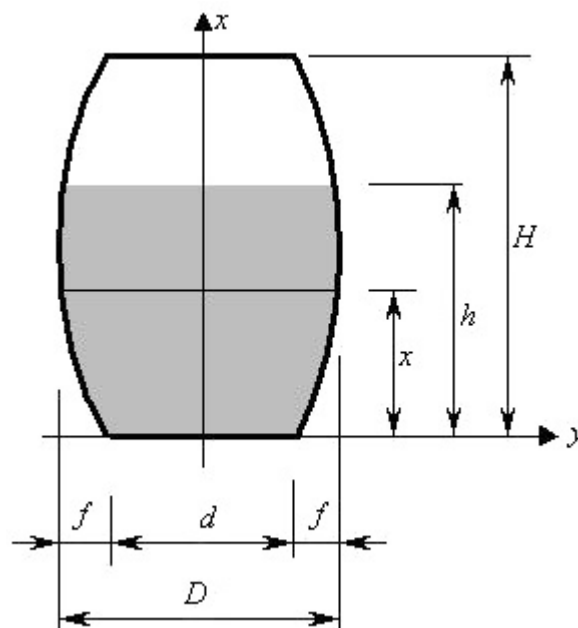


Figure: 26.57 - Exemple de tonneau à section circulaire

Considérons que la courbe latérale du tonneau est une parabole d'équation:

$$y = ax^2 + bx + c \quad (26.252)$$

Posons:

$$f = \frac{D-d}{2} \text{ et } L = \frac{H}{2} \quad (26.253)$$

étant donné la manière dont nous avons disposé les axes x, y il est relativement aisé de déterminer les coefficients du polynôme. Déterminer le coefficient c est le plus simple:

$$c = y(x = 0) = \frac{d}{2} \quad (26.254)$$

Nous avons aussi:

$$y' = 0 = 2ax + b \Rightarrow 2a \frac{H}{2} + b = 2aL + b = 0 \Rightarrow a = -\frac{b}{2L} \quad (26.255)$$

De même que:

$$\begin{aligned} y\left(x = \frac{H}{2}\right) &= a\left(\frac{H}{2}\right)^2 + b \frac{H}{2} + c = -\frac{b}{2L} L^2 + bL + \frac{d}{2} = \frac{D}{2} \\ \Rightarrow b\left(-\frac{L}{2} + L\right) &= \frac{D-d}{2} \Rightarrow b \frac{L}{2} = f \Rightarrow b = \frac{2f}{L} \end{aligned} \quad (26.256)$$

Ainsi, nous avons:

$$a = -\frac{f}{L^2} \quad b = \frac{2f}{L} \quad c = \frac{d}{2} \quad (26.257)$$

Le rayon d'une section horizontale d'ordonnée x est $r = y$ et sa surface est donc:

$$S = \pi r^2 \quad (26.258)$$

ou:

$$S = \pi(ax^2 + bx + c)^2 \quad (26.259)$$

Développons:

$$\begin{aligned} S &= \pi(a^2 x^4 + b^2 x^2 + c^2 + 2ax^2(bx + c) + 2bxc) \\ &= \pi(a^2 x^4 + b^2 x^2 + c^2 + 2abx^3 + 2acx^2 + 2bcx) \end{aligned} \quad (26.260)$$

Le volume de liquide pour une hauteur h sera donc:

$$\begin{aligned}
 V &= \pi \int_0^h \left(a^2 x^4 + b^2 x^2 + c^2 + 2abx^3 + 2acx^2 + 2bcx \right) dx \\
 &= \pi \left[a^2 \frac{x^5}{5} + b^2 \frac{x^3}{3} + c^2 x + 2ab \frac{x^4}{4} + 2ac \frac{x^3}{3} + 2bc \frac{x^2}{2} \right]_0^h \\
 &= \pi \left[a^2 \frac{h^5}{5} + 2ab \frac{h^4}{4} + b^2 \frac{h^3}{3} + 2ac \frac{h^3}{3} + 2bc \frac{h^2}{2} + c^2 h \right] \quad (26.261) \\
 &= \pi \left[a^2 \frac{h^5}{5} + 2ab \frac{h^4}{4} + b^2 \frac{h^3}{3} + 2ac \frac{h^3}{3} + 2bc \frac{h^2}{2} + c^2 h \right] \\
 &= \pi \left[a^2 \frac{h^5}{5} + ab \frac{h^4}{2} + \left(2ac + b^2 \right) \frac{h^3}{3} + bc h^2 + c^2 h \right]
 \end{aligned}$$

Pour calculer la surface intérieure du tonneau, nous considérons la courbe extérieure donnée par un arc de parabole comme représenté ci-dessous:

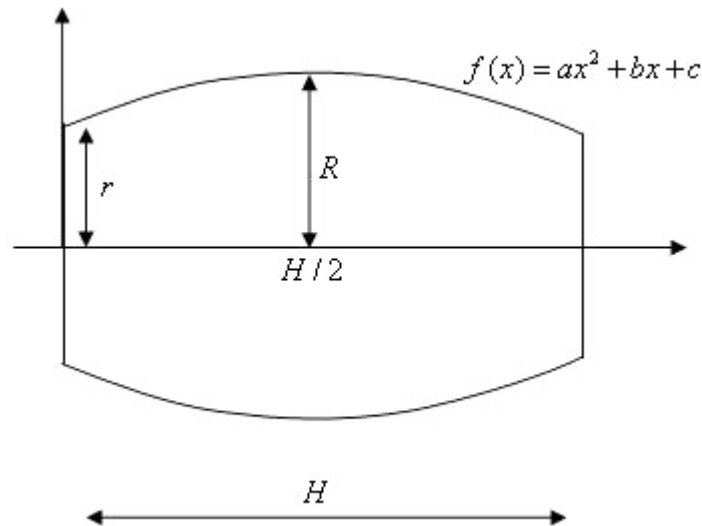


Figure: 26.58 - Tranche verticale du tonneau

Pour calculer l'aire latérale de ce tonneau, nous devons d'abord déterminer l'expression de la parabole ci-dessus.

En regardant la figure, nous obtenons :

$$\begin{aligned}
 f(0) &= r = c \\
 f(H) &= r = aH^2 + bH + c \\
 f(H/2) &= a \frac{H^2}{4} + b \frac{H}{2} + c = R
 \end{aligned} \quad (26.262)$$

qui est un système de trois équations en les inconnues a, b, c .

Après résolution, nous obtenons :

$$\begin{aligned}
 a &= \frac{4(r-R)}{H^2} \\
 b &= \frac{4(R-r)}{H} \\
 c &= r
 \end{aligned} \quad (26.263)$$

La surface latérale du tonneau incluant la surface des deux disques aux extrémités est alors donnée par :

$$S = 2\pi r^2 + 2\pi \int_0^H |f(x)| \sqrt{1 + (f'(x))^2} dx = 2\pi r^2 + 2\pi \int_0^H (ax^2 + bx + c) \sqrt{1 + (2ax + b)^2} dx \quad (26.264)$$

En faisant le changement de variable $u = 2ax + b$ nous obtenons :

$$S = 2\pi r^2 + \frac{\pi}{4a^2} \int_b^{2aH+b} (u^2 - b^2 + 4ac) \sqrt{u^2 + 1} du \quad (26.265)$$

Cette dernière intégrale peut être calculée en utilisant les relations suivantes (dont la deuxième a été démontrée dans le chapitre de Calcul Différentiel Et Intégral):

$$\begin{aligned}
 \int u^2 \sqrt{u^2 + 1} du &= \frac{1}{4} u \left(u^2 + \frac{1}{2} \right) \sqrt{u^2 + 1} - \frac{1}{8} \operatorname{arc\,sinh}(u) \\
 \int \sqrt{u^2 + 1} du &= \frac{1}{2} u \sqrt{u^2 + 1} + \frac{1}{2} \operatorname{arc\,sinh}(u)
 \end{aligned} \quad (26.266)$$

où pour rappel:

$$\operatorname{arc\,sinh}(u) = \ln(u + \sqrt{u^2 + 1}) \quad (26.267)$$

Nous n'irons pas plus loin, car la formule obtenue serait énorme et sans grand intérêt.

Voici néanmoins une application numérique. Supposons:

$$\begin{aligned}
 r &= 20\text{cm} \\
 R &= 30\text{cm} \\
 H &= 60\text{cm}
 \end{aligned} \quad (26.268)$$

Nous calculons :

$$a = -0.0111, \quad b = 0.667, \quad c = 20 \quad (26.269)$$

Donc:

$$S = 2.513 \cdot 10^3 + 6.362 \cdot 10^3 \int_{0.667}^{-0.667} u^2 \sqrt{u^2 + 1} du - 8.483 \cdot 10^3 \int_{0.667}^{-0.667} \sqrt{u^2 + 1} du \quad (26.270)$$

la première des intégrales vaut :

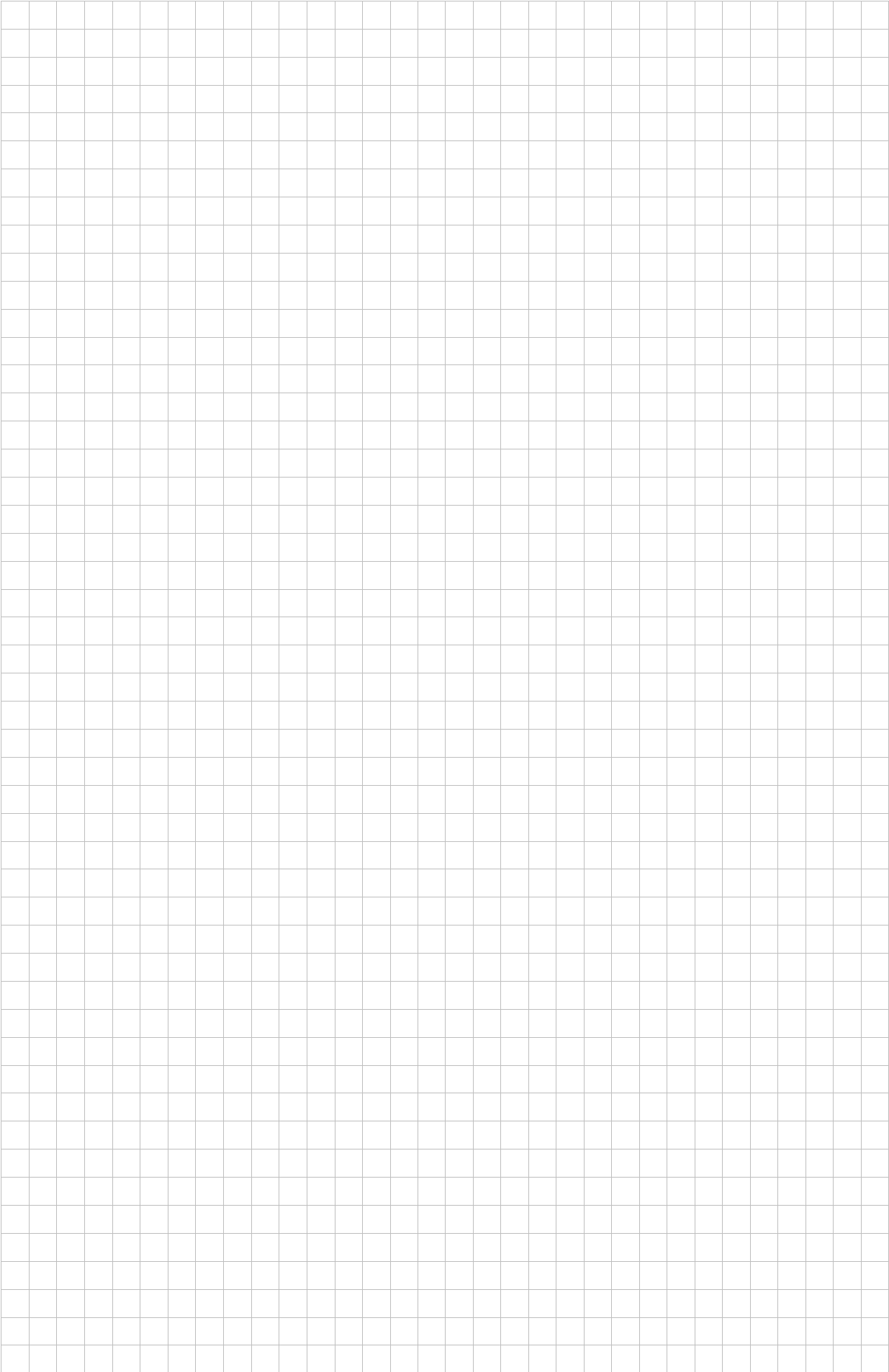
$$\int_{0.667}^{-0.667} u^2 \sqrt{u^2 + 1} \, du = -0.222 \quad (26.271)$$

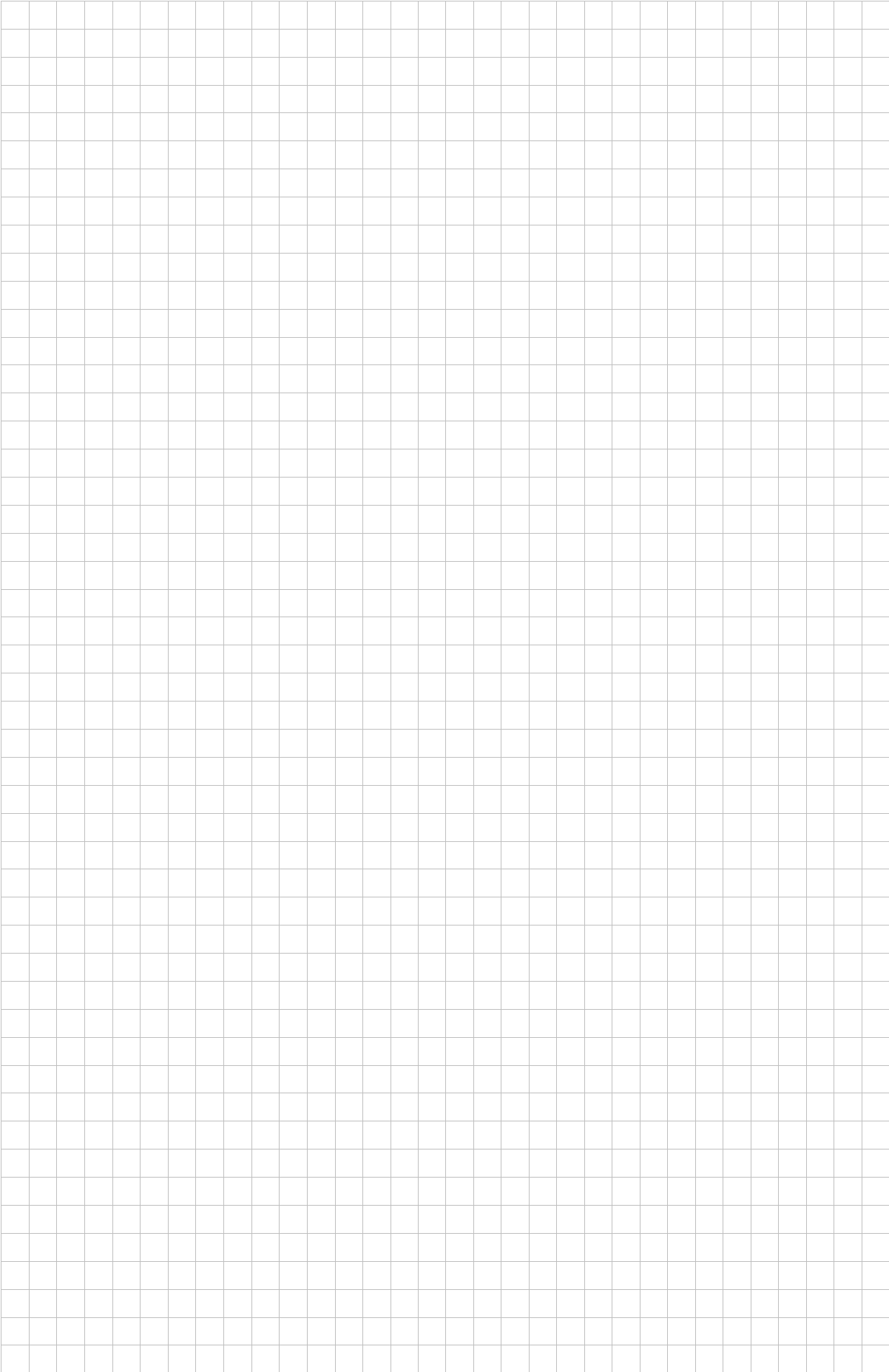
La deuxième vaut :

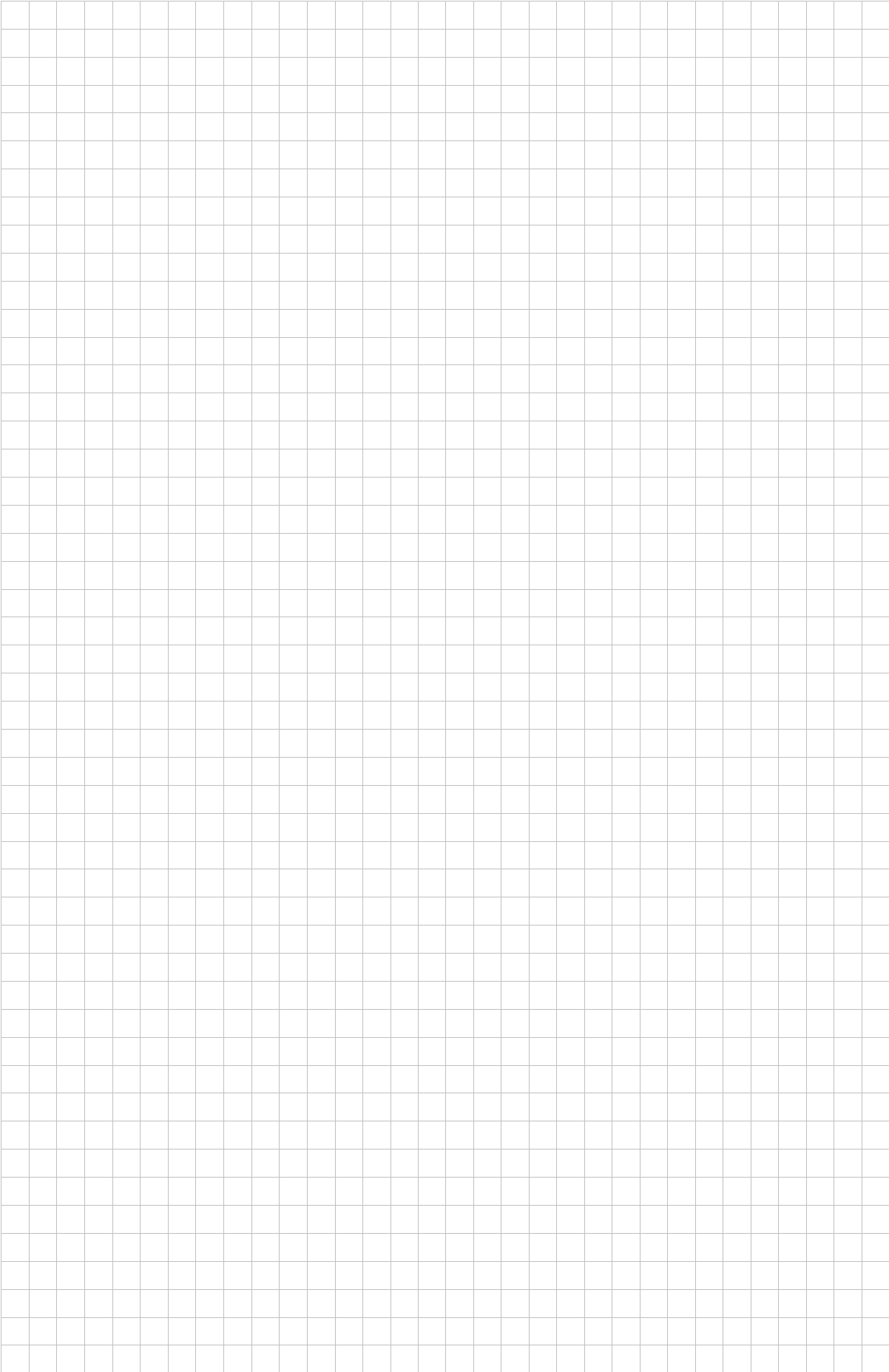
$$\int_{0.667}^{-0.667} \sqrt{u^2 + 1} \, du = -1.427 \quad (26.272)$$

Et finalement:

$$S \cong 1.321 [m^2] \quad (26.273)$$







27. THÉORIE DES GRAPHS

L'histoire de la théorie des graphes (ou des "**complexes cellulaires**") débute peut-être avec les travaux d'Euler au 18ème siècle et trouve son origine dans l'étude de certains problèmes, tels que celui des ponts de Königsberg (les habitants de Königsberg se demandaient s'il était possible, en partant d'un quartier quelconque de la ville, de traverser tous les ponts sans passer deux fois par le même et de revenir à leur point de départ), la marche du cavalier sur l'échiquier ou le problème du coloriage de cartes et du plus court trajet entre deux points.

La théorie des graphes s'est alors développée dans diverses disciplines telles que la chimie (isomères), la biologie, les sciences sociales (réseaux de transports), gestion de projets (C.P.M.), informatique (topologie des réseaux, complexité algorithmique, protocoles de transferts), la physique quantique, etc. Depuis le début du 20ème siècle, elle constitue une branche à part entière des mathématiques, grâce aux travaux de König, Menger, Cayley puis de Berge et d'Erdős.

De manière générale, un graphe permet de représenter la structure, les connexions d'un ensemble complexe en exprimant les relations entre ses éléments: réseau de communication, réseaux routiers, interaction de diverses espèces animales, circuits électriques, ...

Les graphes constituent donc une méthode de pensée qui permet de modéliser une grande variété de problèmes en les ramenant à l'étude de sommets et d'arcs.

Les derniers travaux en théorie des graphes sont souvent effectués par des informaticiens, du fait de l'importance que revêt l'aspect algorithmique dans leur domaine (voir le début du chapitre de Méthodes Numériques pour un petit exemple).

Effectivement, il s'agit essentiellement de modéliser des problèmes. Nous exprimons le problème en termes de graphes de sorte qu'il relève d'un problème de la théorie des graphes que nous savons le plus souvent résoudre car rentrant dans une catégorie de problèmes connus.

Les solutions de problèmes de graphes peuvent être faciles et efficaces (le temps nécessaire pour les traiter informatiquement étant raisonnable car ils dépendent polynomialement du nombre de sommets du graphe) ou difficiles (le temps de traitement étant alors exponentiel) auquel cas nous utilisons une heuristique, c'est-à-dire un processus de recherche d'une solution (pas forcément la meilleure).

Si la théorie des graphes connaît un assez grand engouement ces 30 dernières, peut-être est-ce dû au fait qu'elle ne nécessite pas dans ses concepts élémentaires de bagage mathématique considérable. Effectivement, il suffit d'avoir parcouru les chapitres de Probabilités, de Théorie Des Ensembles et d'Algèbre Linéaire ainsi que de Topologie présentés sur le site pour déjà se sentir à l'aise avec les différentes définitions.

Nous allons introduire le vocabulaire de base de la théorie des graphes. Les termes employés sont ceux du langage commun de la géométrie euclidienne (et malheureusement ils sont aussi en grand nombre...).

Définitions:

D1. Un "**graphe**" (ou "**polygraphe**") G est un couple $G = (X, E)$ constitué d'un ensemble X non vide et fini (les sommets), et d'un ensemble E (les arêtes) de paires d'éléments de X reliés par un segment de droite ou autrement dit (...) d'une partie du produit cartésien X^2 (cf. chapitre de Théorie Des Ensembles).

Remarque: Un graphe est souvent noté en français $G=(S,A)$ où le S est la première lettre du mot "Sommet" et A la première lettre du mot "Arêtes".

L'ensemble des sommets est noté $G(X)$ et l'ensemble des arêtes $G(E)$.

Un graphe est dit "**graphe planaire**" quand nous pouvons le représenter dans un plan sans qu'il y ait intersection d'arêtes.

Maintenant, montrons que si F est le nombre de faces d'un graphe planaire (compte aussi pour une face la face extérieure infinie), A son nombre d'arêtes et S son nombre de sommets, nous avons alors:

$$F - A + S = 2 \quad (27.1)$$

qui est la relation connue sous le nom de "**formule d'Euler**" ou "**théorème de Descartes-Euler**" (démonstration après l'exemple) et qui nous sera utile plusieurs fois sur ce site (dans le présent chapitre et lors de notre étude des polyèdres dans le chapitre sur les Formes Géométriques).

Exemple:

Un graphe à 2 faces (la face en gris clair est la face extérieure infinie), 4 sommets et 4 arêtes:

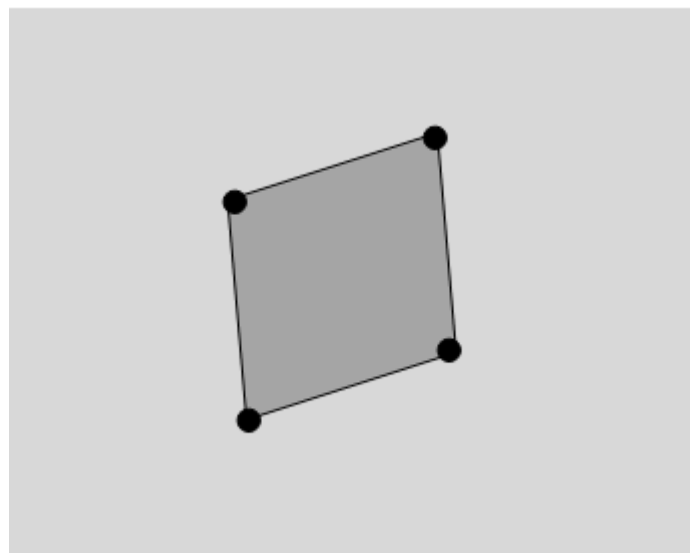


Figure: 27.1 - Graphe à 2 faces, 4 sommets et 4 arêtes

Démonstration:

Nous démontrons cette formule en effectuant une récurrence (cf. chapitre de Théorie De La Démonstration) sur la différence $A - S$:

D'abord, la formule est vraie pour:

$$A - S = -1 \quad (27.2)$$

car, dans ce cas, le graphe est un arbre donc il n'a qu'une seule face (la face extérieure), donc $F = 1$.

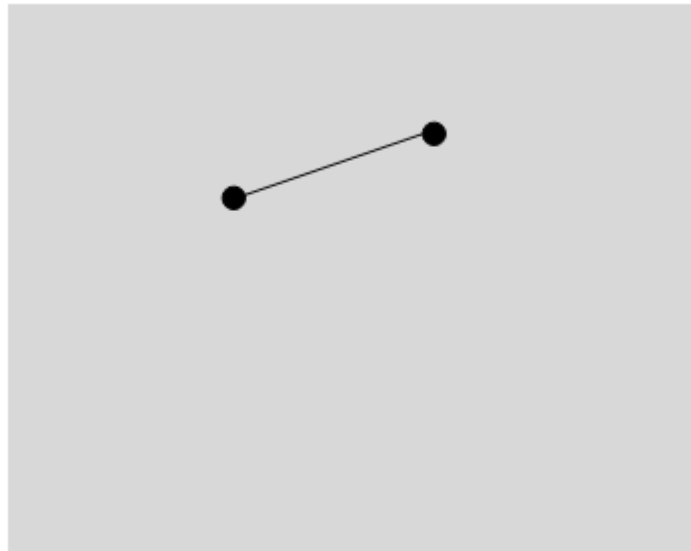


Figure: 27.2 - Arbre avec une arête et 2 sommets

Donc:

$$F - A + S = 1 - (-1) = 2 \quad (27.3)$$

Puis, prenons un graphe connexe (voir définition plus loin) contenant au moins un cycle G (la figure ci-dessous est un exemple de graphe avec 3 cycles):

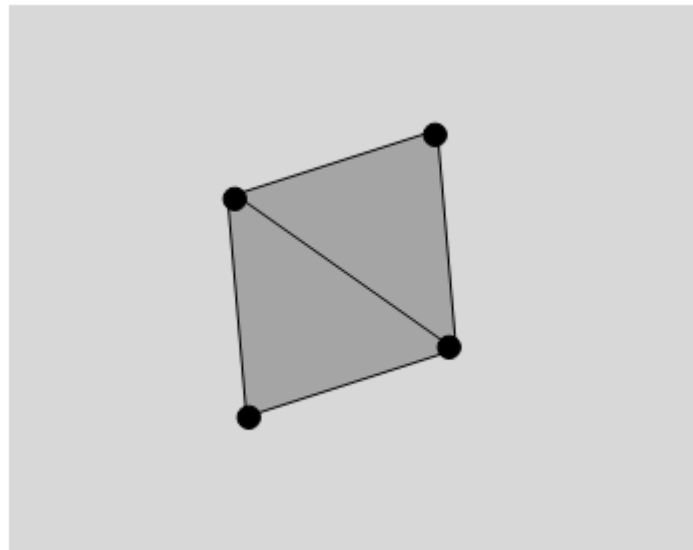


Figure: 27.3 - Graphe avec au moins un cycle

Si nous retirons une arête e à ce cycle, nous devrions pouvoir alors par récurrence appliquer au graphe $G - \{e\}$ la même formule si elle est juste. Effectivement, le graphe amputé de l'arête aura F faces, S sommets et A arêtes et donc la formule:

$$F - A + S = 2 \quad (27.4)$$

si nous lui remettons l'arête alors nous écrivons:

$$(F + 1) - (A + 1) + S = F - A + S = 2 \quad (27.5)$$

□C.Q.F.D.

D2. Les éléments de X sont donc les "**sommets**" du graphe G , ceux de E sont les "**arêtes**" du graphe G (effectivement, une arête est composée de deux sommets reliés par un segment de droite, d'où l'allusion aux paires d'éléments dans la définition précédente).

Remarque: Dans un "**multigraphe**", les deux sommets d'une arête peuvent être identiques (boucle) et deux arêtes distinctes peuvent avoir leurs deux extrémités communes. Un multigraphe ne satisfait plus alors la définition D1.

D3. Soit $e = \{x, y\}$ une arête de G , nous disons que les sommets x, y qui sont les "**extrémités**" de l'arête de G , sont "**adjacents**" ou "**voisins**" dans le graphe G , et que l'arête e est "**incidente**" aux sommets x, y .

D4. Si deux arêtes e et e' ont une extrémité en commun, nous dirons qu'elles sont "**incidentes**", autrement, qu'elles sont indépendantes.

Remarque: Si e est une arête de G , nous noterons $G - e$ le sous-graphe de $G' = (X, E - \{e\})$. Si X' est un sous-ensemble de X , nous noterons $G - X'$ le graphe G privé des sommets de X' .

D5. Ce que nous nommons "**ordre**" du graphe est le nombre de ses sommets.

Soit G un graphe d'ordre n , l'ensemble E doit être par définition choisi comme sous-ensemble de l'ensemble des paires d'éléments de l'ensemble X , donc d'un ensemble de cardinal:

$$\frac{n(n-1)}{2} \quad (27.6)$$

Il s'agit d'un résultat relativement trivial puisque chacun des sommets est lié à tous les autres sommets sauf lui-même (d'où le numérateur) et on divise par deux simplement pour ne pas compter les sommets voisins deux fois (et ils le sont tous lorsque nous les parcourons tous).

En conséquence, il existe (voir le chapitre de Probabilités: arrangements de n éléments non distinguables par couples de deux):

$$2^{\frac{n(n-1)}{2}} \quad (27.7)$$

choix possibles pour E et donc autant de graphes admettant X pour ensemble de sommets. Certains de ces graphes, sont par le fait que nous considérons leurs sommets comme non distinguables "automorphes" (voir la définition de ce terme un peu plus loin dans ce chapitre).

Le résultat obtenu signifie qu'il existe environ 2 millions de graphes à 7 sommets, et quelques $4 \cdot 10^{105}$ graphes à 27 sommets - chiffre à comparer avec le fait que nous estimons à moins de 10^{100} le nombre d'atomes dans l'Univers (...).

D6. Le "**voisinage**" d'un sommet est l'ensemble de ses voisins.

D7. Nous appelons "**degré**" d'un sommet s et notons $D(s)$, le nombre de ses voisins, qui est également le nombre d'arêtes qui lui sont incidentes (un sommet de degré zéro étant appelé un "**sommet isolé**").

Remarque: Un sommet de degré 1 est appelé "**sommet pendant**".

Propriétés (sans démonstration):

P1. La somme des degrés des sommets est égale au double du nombre d'arêtes.

P2. Dans un graphe, le nombre de sommets de degré impair est toujours pair.

P3. Un graphe ayant tous ses sommets de degré pair est d'ordre impair (excepté pour le sommet isolé), c'est-à-dire qu'il compte un nombre impair de sommets.

Remarque: Un "**graphe régulier**" est un graphe dont tous les sommets ont même degré k . Nous disons alors que le graphe est " **k -régulier**".

D8. Nous dirons qu'un graphe $G' = (X', E')$ est un "**sous-graphe**" ou "**sous-graphe induit**" d'un graphe $G = (X, E)$ lorsque $X' \subseteq X$ et $E' \subseteq E$.

D9. Un "**sous-graphe recouvrant**" d'un graphe $G = (X, E)$ est un sous-graphe $G' = (X, E')$, c'est-à-dire un sous-graphe dont sont sommets tous les sommets de G et dont les arêtes sont dans E .

D10. Pour un graphe d'ordre n , il existe deux cas extrêmes pour l'ensemble de ses arêtes: soit le graphe n'a aucune arête, soit toutes les arêtes possibles pouvant relier les sommets deux à deux sont présentes. Dans ce dernier cas le graphe est dit appelé un "**graphe complet**".

Exemple:

Voici quelques graphes complets pour lesquels nous avons bien:

$$\frac{n(n-1)}{2} \quad (27.8)$$

arêtes. Nous remarquons que les quatre premiers graphes sont planaires (effectivement remarquez comment il est possible, par projection d'un sommet dans le plan, de transformer le quatrième K_4 en K_4' de manière à ce qu'il n'y ait plus d'intersections). Le cinquième graphe K_5 est non-planaire (nous ne pouvons trouver des déplacements évitant les croisements).

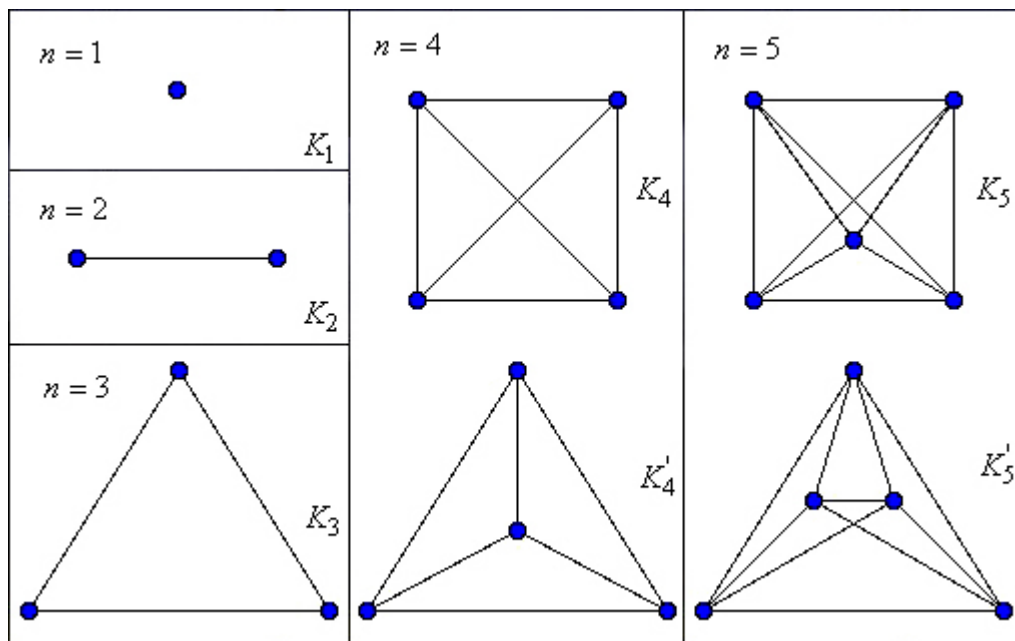


Figure: 27.4 - Exemples de graphes complets

Un graphe complet est donc un graphe où chaque sommet est relié à tous les autres. Le graphe complet d'ordre n est noté K_n . Dans ce graphe chaque sommet est de degré $n-1$.

Ainsi, un cas sympathique à traiter est "l'étoile de David":

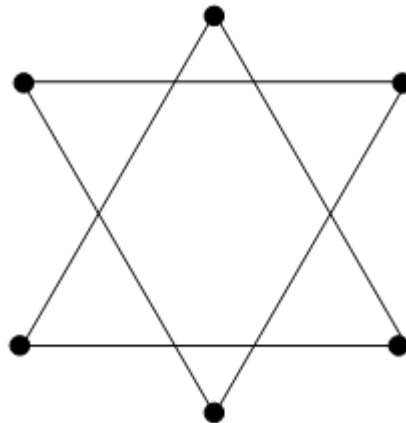


Figure: 27.5 - Étoile de David

qui n'est évidemment pas un graphe complet par définition que si l'on joint tous les sommets entre eux (ainsi nous perdons la géométrie de l'étoile mais obtenons un graphe K_6):

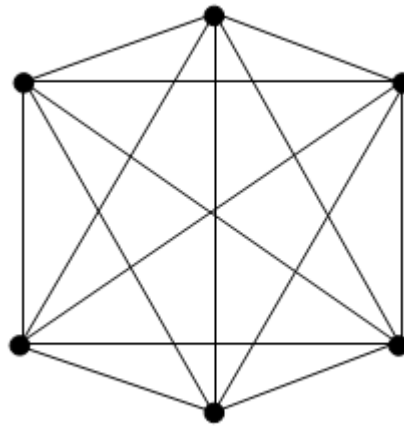


Figure: 27.6 - Étoile de David sous forme de graphe complet

Remarque: Ce résultat est intéressant dans le domaine de la gestion de la communication des projets d'entreprise et de la sécurité informatique (cf. chapitre de Cryptographie). Si par exemple vous gérez un projet avec 10 intervenants (correspondant à n), il y a donc $n(n-1)/2$ canaux de communication (e-mail ou téléphone) possibles, soit 45 (et dans le domaine de la sécurité il y aurait 45 clés de cryptage à système symétrique à générer). D'où l'importance en gestion de projet de définir des règles de communication claires (sous forme d'un graphe) si l'on ne veut pas être noyé par les e-mails ou les téléphones inutilement (et dans le domaine de la sécurité à mettre en place des systèmes asymétriques). Nous retrouverons aussi ce résultat dans le modèle de la goutte liquide du noyau nucléaire dans le chapitre de Physique Nucléaire.

D11. Un "**graphe stable**" est sous-graphe sans arête et une "clique" un sous-graphe complet.

D12. Dans un graphe, il est naturel de vouloir se déplacer de sommet en sommet en suivant les arêtes. Une telle marche passant par n sommets est appelée une "**chaîne**" P_n ou un "**chemin**":

Un chemin ("path" en anglais) est une liste $P_n = P_k = (x_1, \dots, x_k)$ de sommets telle qu'il existe dans le graphe une arête entre chaque paire de sommets successifs: $\forall i = 1, \dots, k-1 \quad (x_i, x_{i+1}) \in E$. La longueur du chemin correspond au nombre d'arêtes parcourues: $k-1$.

Un chemin est dit "**chemin simple**" si chaque arête du chemin est empruntée une seule fois. Voici par exemple un chemin simple avec 5 sommets:

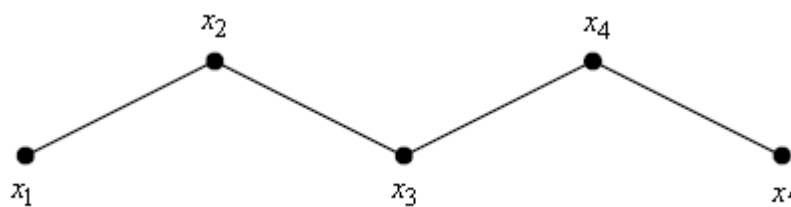


Figure: 27.7 - Exemple de chemin simple

Ainsi, nous définissons aussi un "**cycle**":

$$C_n = C_{k+1} = (x_1, \dots, x_k, x_{k+1}) \quad (27.9)$$

comme étant un chemin simple finissant à son point de départ tel que $x_1 = x_{k+1}$. Ainsi, s'il existe deux chaînes distinctes reliant deux sommets x et y d'un graphe G , alors ce graphe admet un cycle.

D13. Un "**cycle simple**" est un cycle dont toutes les arêtes sont différentes.

D14. Un "**graphe orienté**" est un graphe dont les arêtes ont une direction et un sens et sont dès lors appelées des "**arcs**" (donc à l'opposé du graphe non orienté).

Remarques:

R1. Les termes de "chemin" et de "circuit" s'emploient en propre pour les graphes orientés. Pour les graphes non orientés que nous manipulons principalement ici, nous parlons de "chaîne" et de "cycle". Cependant, la définition formelle est exactement la même dans les deux cas, seule change la structure (graphe orienté ou non) sur laquelle ils sont définis.

R2. Un graphe non orienté n'est qu'un graphe orienté symétrique. Effectivement, si un arc relie le sommet a au sommet b et un autre arc relie le sommet b au sommet a , nous ne traçons alors qu'un trait entre a et b que nous appelons... une arête.

D15. Un chemin $p = (x_1, \dots, x_k)$ est dit "**chemin élémentaire**" si chacun des sommets du parcours est visité une seule fois: $\forall i, j = 1, \dots, k \quad i \neq j, x_i \neq x_j$. Un chemin élémentaire est donc un chemin simple et sans cycle.

Propriétés: Dans un graphe G d'ordre n :

P1. Tout chemin élémentaire est de longueur au plus $n-1$. Effectivement, un chemin élémentaire visitant au plus 1 fois chaque sommet du graphe, sa longueur (nombre d'arêtes) ne peut effectivement excéder $n-1$.

P2. Le nombre de chemins élémentaires dans le graphe est fini. Effectivement, le nombre de chemins de longueur k ($k = 0, \dots, n-1$) est au plus la combinatoire du choix d'une suite de $k+1$ sommets distinguables parmi n . Il y en a donc (cf. chapitre de Probabilités):

$$A_n^p = \frac{n!}{(n-p)!} \quad (27.10)$$

Les chemins élémentaires sont la restriction naturelle que nous recherchons à la notion de chemin. La question qui se pose est de savoir si nous perdons quelque chose en ne considérant que les chemins élémentaires dans un graphe: peut-on toujours remplacer un chemin du graphe par un chemin élémentaire?

Le "**lemme de König**" répond affirmativement à cette question: de tout chemin, nous pouvons extraire un sous-chemin élémentaire.

L1. S'il existe un chemin entre 2 sommets x et y , alors il existe un chemin élémentaire entre x et y .

Démonstration:

L'idée de la preuve est de choisir un chemin particulier entre x et y et de montrer qu'il est élémentaire. Quel chemin choisir? Si un chemin comporte un circuit, ce circuit est un détour sur la route menant de x et y . Un bon candidat à être un chemin élémentaire semble donc être un plus court chemin.

Parmi tous les chemins reliant x à y , choisissons ainsi un chemin:

$$p = (x_1 = x, \dots, x_k = y) \quad (27.11)$$

comportant le moins d'arêtes. Supposons par l'absurde que p n'est pas élémentaire. Il existe alors dans ce chemin un sommet z apparaissant au moins 2 fois le long du chemin p .

Soient i, j les 2 premiers indices tels que $x_i = z$ et $x_j = z$:

$$p = (x_1 = x, \dots, x_i = z, \dots, x_j = z, \dots, x_k = y) \quad (27.12)$$

Pour obtenir une contradiction, il suffit de supprimer le cycle entre $x_i = z$ et $x_j = z$. Alors, nous avons un nouveau chemin:

$$p' = (x_1 = x, \dots, x_{i-1}, \dots, x_{j+1}, \dots, x_k = y) \quad (27.13)$$

est un chemin, reliant x à y . Sa longueur est strictement inférieure à celle de p , ce qui contredit notre choix initial comme étant un plus court chemin.

D16. Un graphe est dit "**graphe connexe**" si et seulement si, il existe au moins un chemin entre chaque paire de sommets (le chemin n'étant donc implicitement pas nécessairement direct - pouvant passer par un ou plusieurs sommets intermédiaires). S'il existe un chemin entre chaque paire de sommets, nous disons que nous avons un "**graphe fortement connexe**".

Remarques: Que se passe-t-il si le graphe G n'est pas connexe? Il apparaît alors comme un ensemble de graphes connexes mis les uns à côté des autres. Chacun de ces graphes est un sous-graphe particulier de G , appelé "**composante connexe**". Il est souvent utile de se placer sur les composantes connexes d'un graphe pour se ramener au cas d'un graphe connexe.

D17. Un "**arbre**" ou "**arbre couvrant**" est un graphe connexe (non orienté), sans cycle simple (acyclique) et sans boucles (il s'agit donc d'une forêt connexe). Dans un arbre le nombre d'arêtes est égal au nombre de sommets - 1.

D18. Un "**arbre valué**" ou "**graphe valué**" est un arbre (respectivement un graphe) où les arêtes ont des valeurs (pondérations) positives. La somme de toutes les valeurs qui sont sur les arêtes parcourues d'un arbre est appelé alors le "**coût d'un arbre valué**" (respectivement "**coût d'un graphe valué**").

Remarque: Les arbres valués sont utilisés dans de très nombreux domaines. Citons les réseaux informatiques dans lesquels on cherche à optimiser le nombre d'interconnexions entre machines pour éviter les redondances d'envois de paquets de données ou la gestion de projets (voir l'exemple ci-dessous).

 **Exemple:**

Un excellent exemple pratique de graphe connexe valué et orienté (abrégé sous le terme de "**digraphe**") est celui utilisé en gestion de projets pour le calcul du chemin critique. Il s'agit d'un graphe qui représente les dépendances entre n tâches intermédiaires nécessaires pour réaliser un projet, communément appelé "**diagramme de Gantt**" ou en encore "**graphe d'ordonnancement**". La durée (poids) de chaque tâche est la valeur des arcs incidents extérieurement au noeud correspondant. Les arcs représentent les contraintes d'enchaînement des tâches. Nous ajoutons toujours un noeud (dans le monde de la gestion de projets on parle plutôt de jalon...) initial et un noeud final. Le premier est relié par un arc de valeur nulle à tous les noeuds sans prédécesseurs, et tous les noeuds sans successeurs sont reliés au noeud final. Le graphe obtenu doit évidemment être acyclique.

Un "**chemin critique**" est un chemin de longueur maximale entre les deux jalons. Il peut éventuellement y en avoir plusieurs, de même longueur. Toute tâche située sur un chemin critique ne peut être retardée sans répercussion sur la durée totale du projet. En d'autres termes, sa "**marge totale**" est nulle (nous disons alors aussi que sa date de fin/début au plus tôt est strictement égale à sa date de fin/début au plus tard). Par ailleurs, nous définissons aussi en gestion de projets, la "**marge libre**" qui indique la durée sur laquelle une tâche peut glisser sans bouger la tâche successeur. La marge libre se calcule comme la différence entre la date de début au plus tôt d'une tâche sommée de sa durée et la date de début au plus tard de la tâche successeur.

Prenons pour l'exemple un projet qui se compose des tâches suivantes:

Tâches	Tâches antérieures	Durée
A	E	3
B	K,C	4
C	-	3
D	E,J	2
E	-	2
F	G,L	3
G	-	4
H	A,M,R	2
J	E	2
K	C	2
L	G	5
M	C	4
N	G	3
R	J	2

Tableau: 27.1 - Ordonnancement de tâches

Le graphe orienté valué connexe correspondant à ce tableau une fois la définition du chemin critique appliqué est le suivant en utilisant les dates de début:

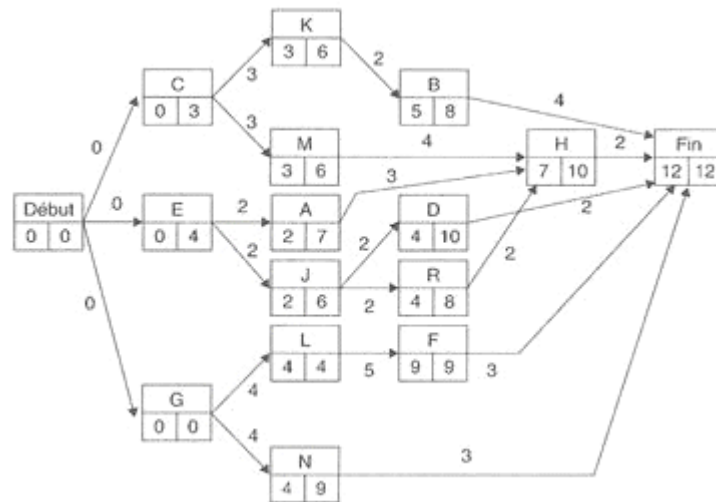


Figure: 27.8 - Exemple de représentation d'un planning sous forme de graphe

Nous voyons dans ce graphe que les tâches $\{\text{Début}, G, L, F, \text{Fin}\}$ sont critiques.

Un excellent outil d'utilisation de tels graphes est MS Project dont le diagramme correspondant à l'exemple ci-dessus est:

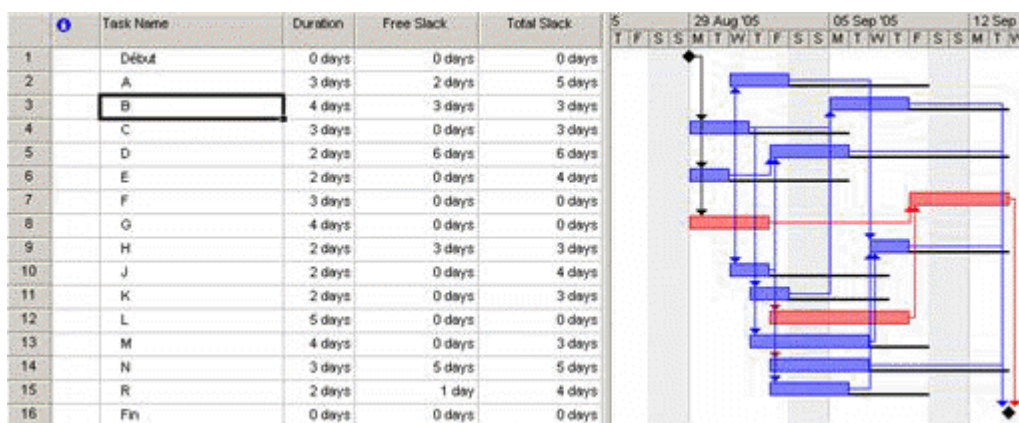


Figure: 27.9 - Même graphe mais vu dans MS Project

D19. Une "**composante connexe**" d'un graphe G est un sous-graphe $G' = (X', E')$ connexe maximal.

Remarque: Un graphe ne possédant qu'une seule composante connexe est simplement un graphe connexe. Un sommet isolé (de degré 0) constitue toujours une composante connexe à lui seul. La relation sur les sommets "il existe un chemin entre ..." est une relation d'équivalence (réflexive, symétrique et transitive). Les composantes connexes d'un graphe correspondent aux classes d'équivalences de cette relation.

Propriété (triviale): Un graphe G d'ordre n connexe comporte au moins $n-1$ arêtes.

D20. Un "**cycle**" est un chemin simple rebouclant sur lui-même. Un graphe dans lequel il n'existe aucun cycle possible est dit "**acyclique**".

Les graphes acycliques non connexes composés d'arbres constituent une classe intéressante de graphes, avec des propriétés remarquables et un nom: les "**forêts**" (terme très souvent utilisé par les informaticiens réseaux).



Figure: 27.10 - Exemple d'un graphe connexe contenant un cycle



Figure: 27.11 - Exemple d'un arbre (graphe connexe ne contenant pas de cycle)

Ci-dessous des exemples d'arbres (chaque arbre est connexe mais l'ensemble forme un graphe acyclique et non-connexe):

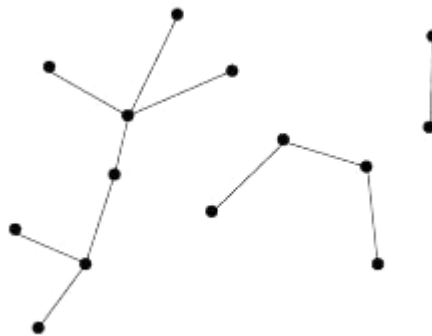


Figure: 27.12 - Exemple d'une forêt (3 composantes connexes)

Nous voyons ainsi qu'une forêt est un graphe dont les composantes sont des arbres. Les sommets de degré 1 sont appelés "**feuilles**" d'un arbre.

Propriétés:

P1. (triviale) Si dans un graphe G tout sommet est de degré supérieur ou égal à 2, alors G possède au moins un cycle.

Remarque: Cette propriété simple implique qu'un graphe sans cycle possède au moins un sommet de degré 0 ou 1. À l'inverse, nous pouvons lier cette fois l'absence de cycle dans un graphe avec le nombre d'arêtes.

P2. (triviale) Un graphe acyclique G à n sommets possède au plus $n-1$ arêtes.

D21. Un "**cycle eulérien**" est un cycle passant une et une seule fois par chaque arête du graphe et revenant au point de départ (nous verrons plus loin les propriétés que doit posséder un tel graphe pour qu'un tel cycle y existe)

D22. Un graphe est dit "**graphe eulérien**" s'il admet un cycle eulérien.

D23. Un "**cycle hamiltonien**" est un cycle simple passant par tous les sommets du graphe une et une seule fois. Pour avoir un cycle hamiltonien, le graphe doit être connexe et il ne doit pas y avoir de sommet pendant.

D24. Un "**graphe hamiltonien**" est un graphe qui possède un cycle hamiltonien.

Il convient d'ouvrir maintenant une parenthèse (pour les paillettes...) sur le problème le plus connu en théorie des graphes: les ponts de Königsberg.

Euler (voir page des biographies), aimait à faire une promenade dans sa bonne ville de Königsberg. Il affectionnait selon la légende tout particulièrement de parcourir les 7 ponts qui enjambent la rivière. L'âge venant (les connaissances mathématiques aussi...), il se demanda si sans sacrifier à sa promenade, il pouvait en raccourcir la longueur en ne traversant chaque pont qu'une seule fois. Ce problème est sans doute l'un des plus anciens en théorie des graphes: celui de l'existence d'une chaîne passant une et une seule fois par chaque arête.



Figure: 27.13 - Carte de la ville de Königsberg

La rivière sépare la ville de Königsberg en quatre parties, a , b , c , d . Chaque pont relie deux de ces parties. Nous pouvons alors représenter notre problème par un graphe avec quatre sommets, où chaque arête représente l'un des sept ponts de Königsberg. Sur cet exemple le graphe n'est pas un graphe simple:

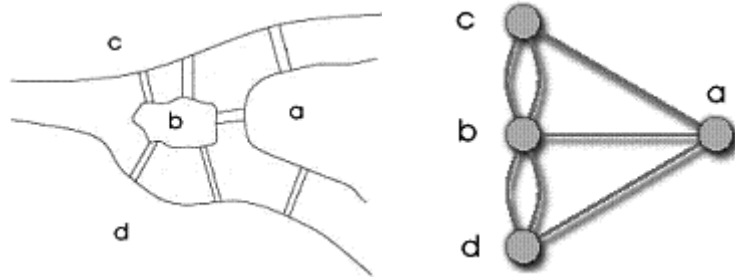


Figure: 27.14 - Représentation des sept ponts de Königsberg

La question est donc ici de savoir si le graphe est eulérien ou non? Si pour notre problème le graphe obtenu est eulérien, il faut pouvoir exhiber un cycle eulérien, ce qui ne semble pas facile. Mais s'il ne l'est pas? Euler a donné une caractérisation très forte des graphes eulériens donnée par l'énoncé suivant:

Théorème d'Euler: Un graphe est eulérien si et seulement s'il est connexe et tous ses sommets sont de degré pair (il y a donc un nombre pair d'arêtes qui arrivent sur chaque sommet dont la moitié d'entre elles servant à arriver sur le sommet, l'autre à en repartir) sauf au plus deux (ces deux exceptions étant les sommets de départ et d'arrivée).

De façon plus précise pour un graphe connexe:

- le graphe n'a pas de sommets impairs, alors il est eulérien (et la chaîne est donc cyclique)
- le graphe ne peut avoir un seul sommet impair de par la propriété (déjà énoncée plus haut) que dans un graphe, le nombre de sommets de degré impair est toujours pair.
- si le graphe a deux sommets impairs, ces sommets sont alors les extrémités de la chaîne eulérienne

Corollaire: un graphe ayant plus de deux sommets impairs ne possède pas de chaîne eulérienne.

Avec cette caractérisation (comme nous allons le démontrer de suite après), les sommets *a*, *b*, *c*, *d* étant de degré impair, on sait immédiatement qu'il est impossible de parcourir tous les ponts de Königsberg seulement une fois au cours d'une promenade.

Démonstration:

1. Supposons qu'un graphe *G* soit eulérien. Il existe alors une chaîne *c* parcourant une et une seule fois chaque arête (jusque-là c'est facile). Bien évidemment, dans le cas d'une chaîne, les sommets se situant aux extrémités de la chaîne sont de degré impair et sont au nombre de deux.
2. Considérons maintenant un sommet *x* et supposons cette fois-ci non pas un graphe eulérien mais un cycle eulérien. Lors du parcours du cycle, à chaque fois que nous passons par le sommet *x*, nous nous retrouvons au point de départ et pour effectuer un nouveau tour chacune des 2 arêtes s'offre à nous (le chemin pouvant être parcouru dans les deux sens puisque le graphe est non orienté). Le sommet *x* est donc de degré pair et peut être défini n'importe où dans le cycle, d'où le fait que l'ensemble des sommets sont de degré pair.
3. Réciproquement considérons un graphe *G* connexe dont tous les sommets sont de degré pair. Nous allons montrer par induction sur le nombre d'arêtes que *G* est alors eulérien:
 - Si *G* se réduit à un unique sommet isolé, il est évidemment eulérien. Sinon tous les sommets de *G* sont de degré supérieur ou égal à 2. Ceci implique qu'il existe un cycle ϕ sur *G*:

- Considérons le graphe partiel H constitué des arêtes en dehors du cycle ϕ . Les sommets de H sont également de degré pair, le cycle contenant un nombre pair d'arêtes incidentes pour chaque sommet. Par induction chaque composante connexe H_i de H est un graphe eulérien, et admet donc un cycle eulérien ϕ_i . Pour reconstruire un cycle eulérien sur G , il nous suffit de fusionner le cycle ϕ avec les différents cycles ϕ_i . Pour cela, nous parcourons le cycle ϕ depuis un sommet arbitraire; lorsque nous rencontrons pour la première fois un sommet x appartenant à H_i , nous lui substituons le cycle ϕ_i . Le cycle obtenu est un cycle eulérien pour G , le cycle ϕ et les cycles ϕ_i formant une partition des arêtes.

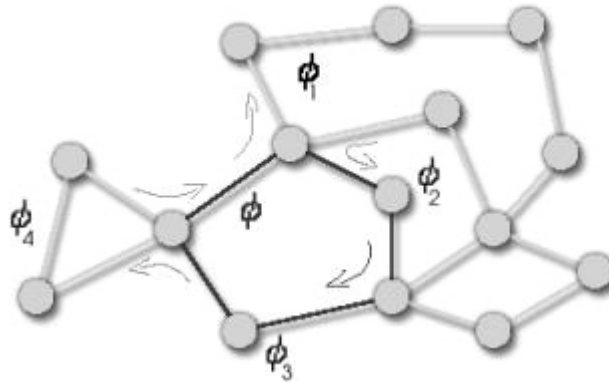


Figure: 27.15 - Graphe Eulérien?

Remarque: Ce principe de décomposer un graphe en graphes connexes et de les sommer permet de construire un algorithme récursif permettant de déterminer si un graphe est eulérien ou non.

D25. Deux graphes G et G' sont "**graphes complémentaires**" lorsqu'ils vérifient les conditions suivantes:

1. Ils ont le même ensemble de sommets
2. Deux sommets x, y sont voisins dans G ne sont pas voisins dans G'

D26. Un "**graphe biparti**" $K_{p,q}$ est un graphe tel que nous puissions partitionner l'ensemble de ses sommets en deux classes respectivement de cardinal p et q de sorte que toute arête ait une extrémité dans chacune des deux classes.

Exemple:

Voici donc une représentation d'un graphe biparti $K_{3,3}$ classique. Il représente le problème fameux de l'approvisionnement de trois maisons au départ de trois usines (eau, électricité, gaz) sans droit d'alignement des services. Nous voyons immédiatement que ce graphe est non-planaire.

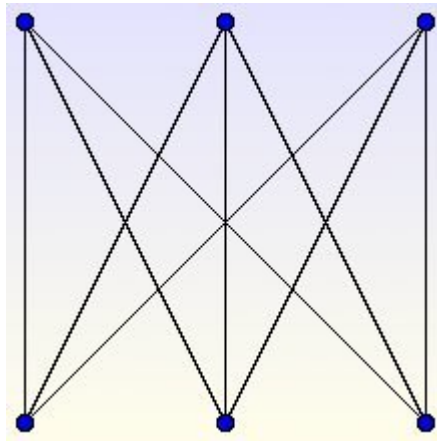


Figure: 27.16 - Exemple de graphe biparti

Remarque: Le graphe biparti complet $K_{p,q}$ est un graphe de sommets de $n = p + q$ sommets configuré de telle sorte que chaque sommet d'une classe soit adjacent à tous les sommets de l'autre et seulement à ceux-ci.

D27. Deux graphes sont isomorphes s'il existe une bijection f de X dans X' telle que, pour tous sommets x et y de G :

$$x \text{ adjacent à } y \text{ dans } G \Leftrightarrow f(x) \text{ adjacent à } f(y) \text{ dans } G'$$

Nous disons aussi que f est un "isomorphisme de G sur G' ".

S'il existe une bijection f de X dans lui-même telle que:

$$x \text{ adjacent à } y \text{ dans } G \Leftrightarrow f(x) \text{ adjacent à } f(y) \text{ dans } G$$

nous disons alors que f est un automorphisme dans G (par permutation des sommets il existe beaucoup d'exemples possibles...).

Remarque: Attention!! Parfois nous parlons de graphes équivalents à "un isomorphisme près". Cela signifie plus clairement, qu'à l'exception d'une et unique violation parmi l'ensemble des arêtes, les graphes sont isomorphes.

Comme l'isomorphisme dans le cas des graphes va d'un ensemble à un autre de même cardinal n , le nombre de bijections possibles différentes est (voir le chapitre de Probabilité sur les arrangements):

$$n! \quad (27.14)$$

Cela signifie qu'il existe au maximum $n!$ graphes qui peuvent se regrouper dans une même classe d'équivalence. En conséquence, il existe au minimum (minorant) le rapport du nombre total de n sommets sur le cardinal majoré de la plus grande classe d'équivalence possible (mais pas nécessairement existante):

$$N = \frac{2^{\frac{n(n-1)}{2}}}{n!} \quad (27.15)$$

En utilisant la majoration grossière $n! < n^n$, nous avons:

$$N > \frac{2^{\frac{n(n-1)}{2}}}{n^n} \quad (27.16)$$

D'où:

$$\log_2 N > \frac{n(n-1)}{2} - n \log_2(n) \quad (27.17)$$

Soit encore:

$$\log_2 N > \frac{n^2}{2} \left(1 - \frac{1}{n} - \frac{2 \log_2(n)}{n} \right) \quad (27.18)$$

Ainsi, quand n tend vers l'infini, $\log_2 N$ admet un minorant de l'ordre de n^2 .

1. MATRICE D'ADJACENCE

Au plan formel, un graphe est aussi un ensemble sur lequel nous avons défini une relation binaire, antiréflexive (aucun élément n'est en relation avec lui-même) et symétrique (si x est en relation avec y , alors y est en relation avec x). La structure de graphe peut alors sembler particulièrement pauvre.

Mais nous pouvons aussi associer à un graphe G de sommets $x_1, \dots, x_n$ une matrice carrée M de dimensions $n \times n$ appelée "**matrice d'adjacence**" du graphe et dont les éléments valent 0 ou 1. En notant m_{ij} le terme située au croisement de la ligne i (représentant le sommet x_i) et de la colonne j (représentant le sommet x_j), M est défini par:

$$m_{ij} = 1 \text{ si et seulement si } x_i, x_j \text{ sont adjacents}$$

De par cette définition et celle du graphe lui-même, il vient que dans une matrice d'adjacence d'un graphe formel que les éléments diagonaux pour lesquels $i = j$ valent tous 0 et que $m_{ij} = m_{ji}$.

Rappelons qu'une telle matrice est dite "symétrique" (cf. chapitre d'Algèbre linéaire).

Remarque: Nous savons aussi que les graphes sont représentés par des matrices dans le cadre de l'étude des chaînes de Markov dans le domaine des probabilités (cf. chapitre de Probabilités).

Voyons un exemple à la fois abstrait mais facilement applicable à de nombreux domaines de l'industrie, de la sociologie et de la biologie. Considérons le graphe orienté suivant et observons qu'il n'est pas antiréflexif ni symétrique:

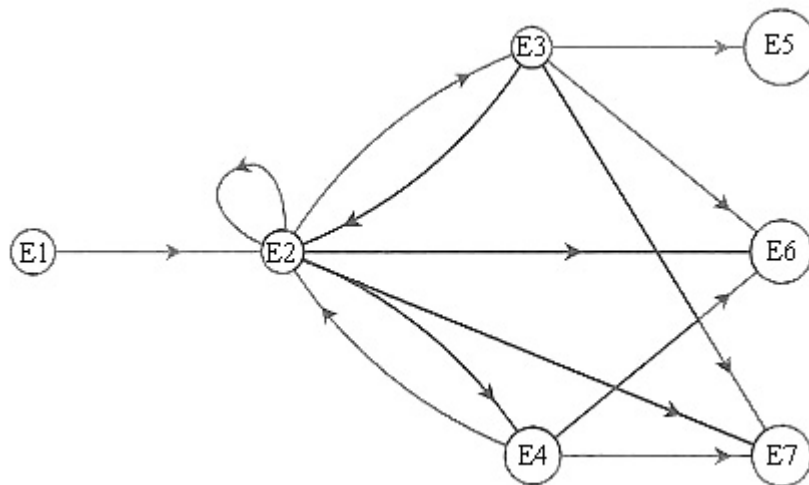


Figure: 27.17 - Exemple de chaîne de Markov (graphe antiréflexif ni symétrique)

Nous pouvons représenter ce diagramme sous forme d'un tableau dans lequel nous noterons "1" quand une transition est possible de l'état mentionné en haut de la colonne de la case vers l'état mentionné au début de la ligne et "0" sinon:

↖	E1	E4	E2	E3	E5	E7	E6
E1	0	0	0	0	0	0	0
E4	0	0	1	0	0	0	0
E2	1	1	1	1	0	0	0
E3	0	0	1	0	0	0	0
E5	0	0	0	1	0	0	0
E7	0	1	1	1	0	0	0
E6	0	1	1	1	0	0	0

Tableau: 27.2 - Matrice d'adjacence

Il est indispensable de comprendre parfaitement la signification des valeurs se trouvant dans ce tableau! Mais à ce niveau du site cela ne devrait pas poser de difficulté majeure.

Maintenant évaluons le nombre de manières permettant d'aller:

1. De E2 vers E2 en deux étapes
2. De E3 vers E4 en deux étapes
3. De E2 vers E7 en deux étapes.

Il est facile dans le cas particulier ci-dessus de dénombrer ces possibilités. Mais dans le cas d'un graphe plus complexe cela devient difficile, voire impossible pour un être humain dans un temps raisonnable.

Nous devons faire appel alors au théorème suivant:

Soit un graphe orienté avec $x_1, x_2, \dots, x_n$ sommets et de matrice d'adjacence:

$$M = (m_{i,j}) \quad (27.19)$$

Pour tout entier naturel k , alors le nombre de chemins de longueur k d'un sommet x_i à un sommet x_j est donné par:

$$M^k = m_{i,j}^{(k)} \quad (27.20)$$

où l'exposant de M dénote la puissance de k de la matrice d'adjacence.

Démonstration:

Effectuons une récurrence sur k :

$$m_{i,j}^{(1)} = m_{i,j} \quad (27.21)$$

désigne bien le nombre de chemins allant de x_i à x_j (où i et j peuvent être égaux). Supposons le résultat vrai pour l'entier $k-1$, avec $k \geq 1$ comme:

$$M^k = M^{k-1}M \quad (27.22)$$

nous avons alors (par construction de la multiplication matricielle):

$$m_{i,j}^{(k)} = \sum_{l=1}^n m_{i,l}^{(k-1)} m_{l,j} \quad (27.23)$$

Par hypothèse de récurrence, $m_{i,l}^{(k-1)}$ est le nombre de chemins de longueur $k-1$ allant de x_i à x_l et $m_{l,j}$ est nous le savons (par construction) le nombre de chemins de longueur 1 allant de x_l à x_j et en particulier il est égal à 1 si (x_l, x_j) est une arête du Graphe et 0 sinon!

Donc, le produit:

$$m_{i,l}^{(k-1)} m_{l,j} \quad (27.24)$$

donne pour une valeur de l donnée le nombre de chemins de longueur k allant de x_i à x_j dont la dernière arête est (x_l, x_j) .

La somme:

$$\sum_{l=1}^n m_{i,l}^{(k-1)} m_{l,j} \quad (27.25)$$

donne donc toutes les possibilités (chemins) de longueur k allant de x_i à x_j quel que soit le point de début de la dernière arête!

□C.Q.F.D.

Ainsi, dans notre exemple, la matrice d'adjacence M est donnée par:

$$M = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix} \quad (27.26)$$

et n'est ni symétrique, ni antiréflexive comme nous en avons déjà fait mention.

Donc si nous la portons à la puissance k , chaque composant $m_{i,j}^{(k)}$ donnera toutes les possibilités (chemins) de longueur k allant de x_i à x_j . Ainsi, nous avons:

$$M^2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 3 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 3 & 1 & 0 & 0 & 0 \\ 1 & 1 & 3 & 1 & 0 & 0 & 0 \end{bmatrix} \quad (27.27)$$

en utilisant par exemple la fonction PRODUITMAT().

Nous avons alors la réponse à nos trois questions en lisant la matrice ci-dessus:

1. De E2 vers E2 en deux étapes nous avons donc 3 possibilités.
2. De E3 vers E4 en deux étapes nous avons donc 1 possibilité.
3. De E2 vers E7 en deux étapes nous avons donc 3 possibilités.

Il est possible de gagner un peu de temps dans ce genre de calculs. Si nous notons C_i , la i -ème colonne de la matrice M , alors:

$$M \cdot C_i = M_i^2 \quad (27.28)$$

donnera la i -ème colonne de la matrice M au carré. Et ainsi de suite:

$$M \cdot M_i^2 = M_i^3 \quad (27.29)$$

Donc nous obtenons systématiquement le nombre de chemins de longueur k d'un point de départ donné correspondant à la colonne i .

Cet exemple a une autre approche très intéressante dans certains domaines étudiant le comportement d'individus dans diverses situations (achats, tourisms, accidents, etc.).

Si au lieu de noter dans la matrice M le nombre de chemins possibles d'un sommet à l'autre, nous notons la probabilité (la partie) du nombre total d'individus qui partent de ce même sommet pour en arriver à un autre alors nous avons par exemple (valeurs imposées par l'expérience!) la matrice:

$$M = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.094 & 0 & 0 & 0 & 0 \\ 1 & 0.954 & 0.662 & 0.75 & 0 & 0 & 0 \\ 0 & 0 & 0.218 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.175 & 1 & 0 & 0 \\ 0 & 0.034 & 0.017 & 0.035 & 0 & 1 & 0 \\ 0 & 0.012 & 0.009 & 0.04 & 0 & 0 & 1 \end{bmatrix} \quad (27.30)$$

qui est déjà la matrice stochastique transposée du graphe visible ci-dessous (conformément à la théorie des chaînes de Markov, nous avons vu qu'il fallait prendre la transposée de la matrice stochastique pour calculer les probabilités d'états).

En considérant que ces probabilités ne changent pas au cours du temps, nous avons alors une chaîne de Markov homogène (sans cycles). Nous voyons alors que:

1. La somme des probabilités de transitions au départ de chaque sommet (état) doit toujours logiquement être égale à 1 (ce que nous avons déjà mentionné dans le chapitre de Probabilités)
2. Tout le monde part de la première étape $E1$
3. Certains stagnent (s'arrêtent) à une certaine étape
4. Ceux qui arrivent à une étape de fin $E5$, $E6$ ou $E7$ y restent et ne reviennent pas sur leurs pas (états absorbants).

Le graphe équivalent devient alors:

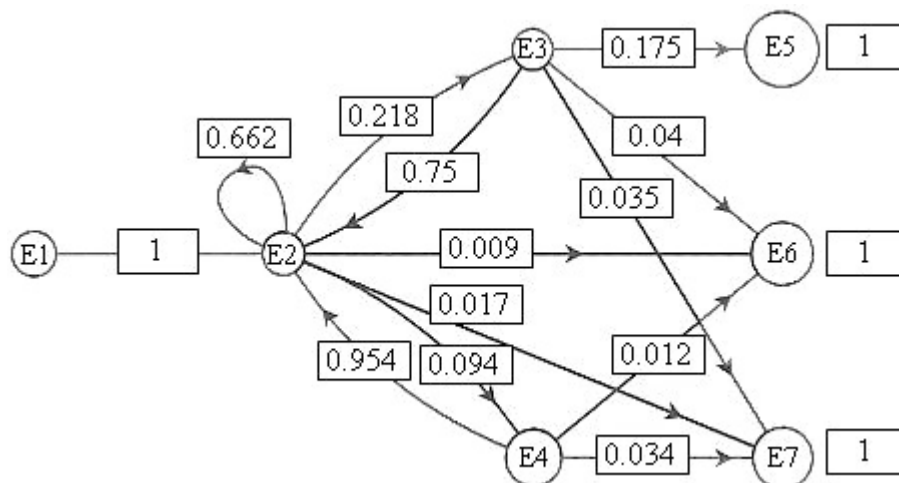


Figure: 27.18 - Graphe équivalent à la matrice

En appelant N (au lieu de M pour ne pas confondre) la matrice construite à partir du graphe ci-dessus nous voyons que si C_i est une des colonnes de la matrice M alors par exemple:

$$N \cdot C_1 = T_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.094 & 0 & 0 & 0 & 0 \\ 1 & 0.954 & 0.662 & 0.75 & 0 & 0 & 0 \\ 0 & 0 & 0.218 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.175 & 1 & 0 & 0 \\ 0 & 0.034 & 0.017 & 0.035 & 0 & 1 & 0 \\ 0 & 0.012 & 0.009 & 0.04 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0.094 \\ 0.662 \\ 0.218 \\ 0 \\ 0.017 \\ 0.009 \end{pmatrix} \quad (27.31)$$

donne la somme des probabilités de transition que ce qui part de $E1$ arrive en 2 étapes en respectivement $E1(0)$, $E2(0.662)$, $E3(0.218)$... (la distribution du vecteur initial peut être quelconque tant que la somme des valeurs de la colonne est égale à 1)!

Si nous multiplions ensuite encore une fois:

$$N \cdot T_1 = T_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.094 & 0 & 0 & 0 & 0 \\ 1 & 0.954 & 0.662 & 0.75 & 0 & 0 & 0 \\ 0 & 0 & 0.218 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.175 & 1 & 0 & 0 \\ 0 & 0.034 & 0.017 & 0.035 & 0 & 1 & 0 \\ 0 & 0.012 & 0.009 & 0.04 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} 0 \\ 0.094 \\ 0.662 \\ 0.218 \\ 0 \\ 0.017 \\ 0.009 \end{pmatrix} = \begin{pmatrix} 0 \\ 0.062 \\ 0.691 \\ 0.144 \\ 0.038 \\ 0.039 \\ 0.025 \end{pmatrix} \quad (27.32)$$

donne la somme des probabilités de transition que ce qui part de $E1$ arrive en 3 étapes en respectivement $E1(0)$, $E2(0.691)$, ... et ainsi de suite. Nous pouvons ainsi savoir qu'elle est la probabilité qu'un individu arrivant à $E2$ puisse arriver à un des sommets terminaux ($E5$, $E6$ ou $E7$) en un nombre d'étapes donné.

Remarque: Se rappeler que la somme des probabilités des colonnes T obtenues est toujours égale à 1 pour la transposée de la matrice stochastique.

En continuant encore longtemps ainsi... nous trouvons que la mesure d'équilibre π qui satisfait (cf. chapitre de Probabilités):

$$P^T \cdot \pi = \pi$$

est:

$$\pi = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0.45 \\ 0.32 \\ 0.23 \end{pmatrix} \quad (27.33)$$

quelle que soit la distribution du vecteur de départ. Propriété que l'on appelle "ergodique" dans le domaine des chaînes de Markov.

Ce qui signifie 45% de probabilité de se trouver en $E5$, 32% de probabilité de se trouver en $E7$ et 23% de probabilité de se trouver en $E6$. Une autre manière de voir les choses est de dire que si une cohorte de 100 individus part de l'étape $E1$ avec les probabilités constantes dans le temps entre les transitions d'étapes, à l'équilibre nous aurons 45 personnes en $E5$, 32 personnes en $E7$ et 23 personnes en $E6$.

2. CATÉGORIES

L'introduction des catégories à travers la "**théorie des catégories**" par Eilenberg et MacLane en 1942 avait pour but de transformer de difficiles problèmes de Topologie en problèmes plus abordables d'algèbre. Plus tard, la théorie des catégories s'est beaucoup développée, à la fois pour elle-même et pour ses applications dans les domaines les plus variés des mathématiques (par exemple en géométrie différentielle). Même si une partie de son développement autonome a parfois été critiquée, les catégories sont maintenant reconnues comme un langage puissant pour développer une sémantique universelle des structures mathématiques. On les utilise aussi en logique et plus récemment en physique, et une collaboration fructueuse semble se développer entre catégoriciens et informaticiens.

Définitions:

D1. Intuitivement une "**catégorie**" est juste un graphe orienté sur lequel nous nous sommes donné une loi pour composer des flèches consécutives, vérifiant certains axiomes.

D2. Un "**graphe orienté**" est formé d'un ensemble d'objets, appelés sommets du graphe, avec des liens entre eux, représentés par des flèches d'un sommet A vers un sommet B , ce que nous notons $f : A \rightarrow B$. Nous disons que A est la "**source**" de la flèche, et B son "**but**". Il peut y avoir plusieurs flèches de même source et de même but (nous les disons "parallèles") et il peut y avoir des flèches "fermées", dont la source et le but sont confondus.

D3. Deux flèches f, g sont dites "**flèches consécutives**" si le but de la première est en même temps la source de la seconde:

$$f : A \rightarrow B, g : B \rightarrow C \quad (27.34)$$

nous disons alors qu'elles forment un chemin de longueur 2 de A vers C .

Une catégorie est donc un graphe dans lequel nous définissons une composition de flèches, associant à tout chemin (f, g) de longueur 2 de A vers C une flèche du graphe de A vers C , dite "composée" du chemin, et notée fg :

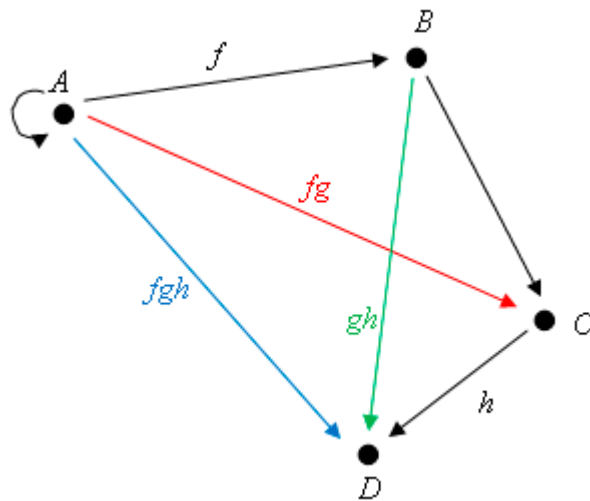


Figure: 27.19 - Exemple de graphe de catégorie

Cette composition vérifie les axiomes suivants:

A1. Associativité : Si fgh est un chemin de longueur 3, les deux composés $f(gh)$ et $(fg)h$ que nous en déduisons sont associatifs. Il s'ensuit qu'à tout chemin de longueur n est aussi associé un seul composé de sommets (invariance de l'itinéraire).

A2. Identités: À tout sommet A est associée une flèche fermée de A vers A , dite "identité" de A et notée Id_A , dont le composé avec une flèche de source ou de but A est égal à cette autre flèche.

Plus généralement, un chemin (de longueur n) de A vers A_n est une suite $(f_1, f_2, \dots, f_n)$ de n flèches consécutives:

$$f_1 : A \rightarrow A_1, f_2 : A_1 \rightarrow A_2, \dots, f_n : A_{n-1} \rightarrow A_n \quad (27.35)$$

Remarques:

R1. Les sommets du graphe sont aussi appelés "**objets**" de la catégorie et ses flèches des "**morphismes**" (ou simplement "**liens**") dans le cadre de la théorie des catégories

R2. Une flèche f est un isomorphisme (cf. chapitre de Théorie Des Ensembles) s'il existe une flèche g (appelée "inverse") telle que les composés fg et gf soient des identités (cet inverse est alors unique).

Ainsi, une catégorie est formée par des objets (les sommets du graphe) et des liens entre eux (les flèches ou morphismes), mais l'idée essentielle est de privilégier les liens sur les objets. En fait, le succès des catégories dans les domaines les plus variés est dû à la richesse des informations sur les objets qui peuvent être déduites de la seule considération des liens et des opérations sur ceux-ci, quelle que soit la nature et l'anatomie de ces objets.

Dans les quelques lignes qui suivent, nous expliquerons comment lire les graphes orientés que nous pouvons rencontrer parfois dans les livres de math. Ceci sera un bon exemple de la théorie des catégories, car nous avons déjà rencontrés de tels graphes sans les décrire dans les chapitres sur les Nombres et la Cryptographie par exemple.

Pour simplifier nous allons expliquer ces diagrammes lorsque les objets de base sont les ensembles (ce qui est le cas le plus courant sur l'ensemble du site de toute manière).

Considérons trois ensembles A , B , et C et trois applications:

$$f: A \rightarrow B, g: A \rightarrow C \text{ et } h: B \rightarrow C \quad (27.36)$$

Nous pouvons considérer les applications f , g et h comme des flèches qui relient les objets (ensembles) A , B , et C pour former un triangle.

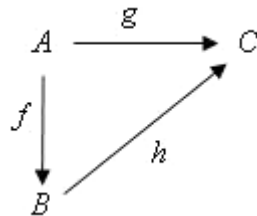


Figure: 27.20 - Exemple de diagramme commutatif

Définition (simpliste): Nous disons que le diagramme fléché ci-dessus est "**un diagramme commutatif**" si tous les chemins que nous empruntons pour aller d'un objet (ensemble) à un autre représentent la même application.

Remarque: Il existe deux façons d'aller de A à C . Nous pouvons y aller directement par g ou bien suivre d'abord f puis h . Ce dernier chemin est représenté par l'application composée $h \circ f$ (cf. chapitre de Théorie Des Ensembles). Ainsi, le diagramme ci-dessus est commutatif si $h \circ f = g$.

Nous pouvons donc introduire la définition plus formelle:

Définition: Le diagramme ci-dessus est commutatif si $h \circ f = g$.

Remarque: Rappelons que ceci veut dire que pour tout élément $a \in A$, $h(f(a)) = g(a)$.

Nous pouvons compliquer à souhait les diagrammes en considérant plus d'ensembles et de flèches (applications) les reliant. Par exemple:

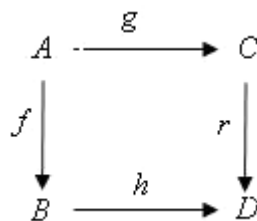


Figure: 27.21 - Autre exemple de diagramme fléché

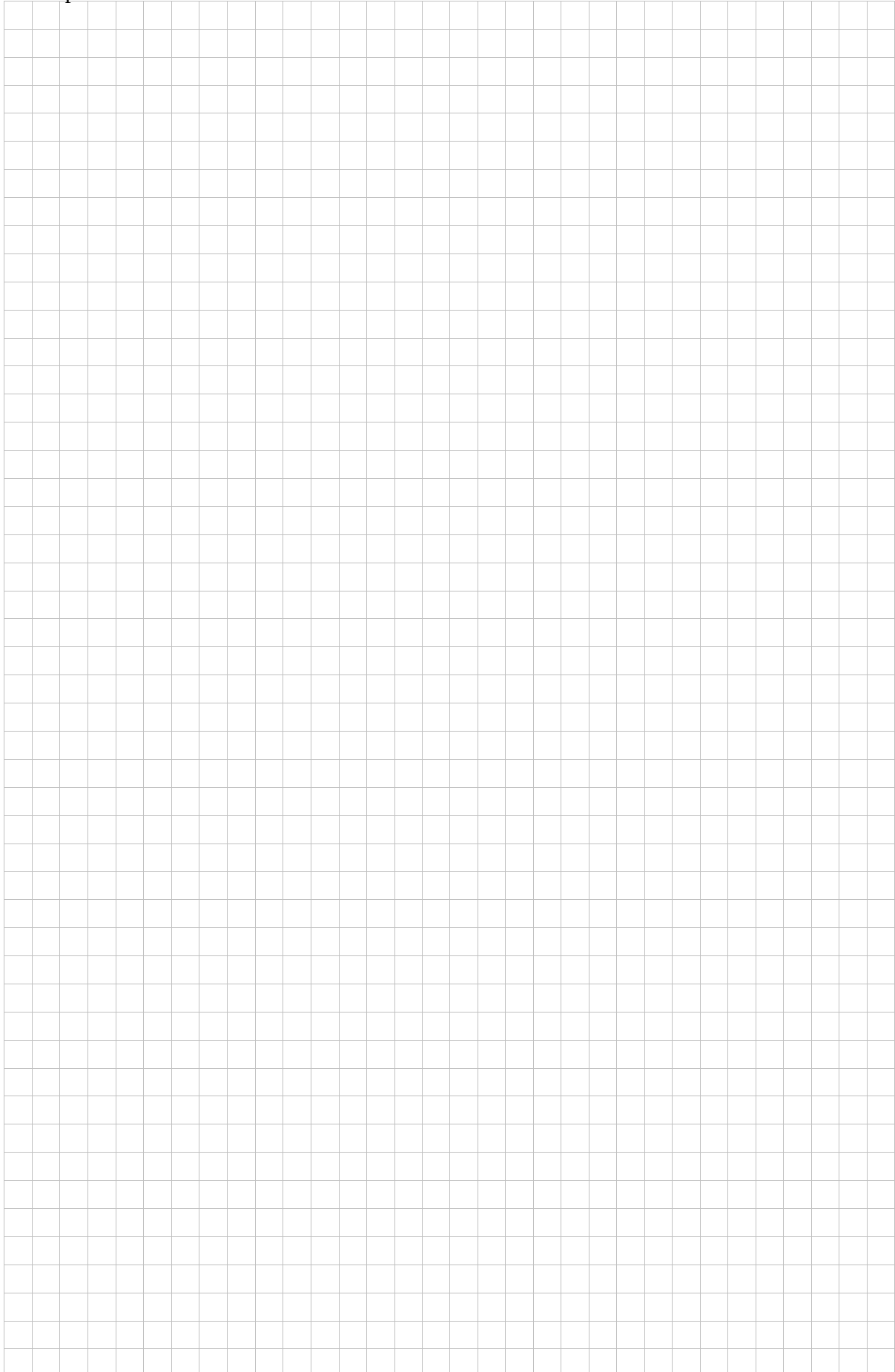
Ce diagramme étant commutatif si et seulement si $h \circ f = r \circ g$.

Remarques:

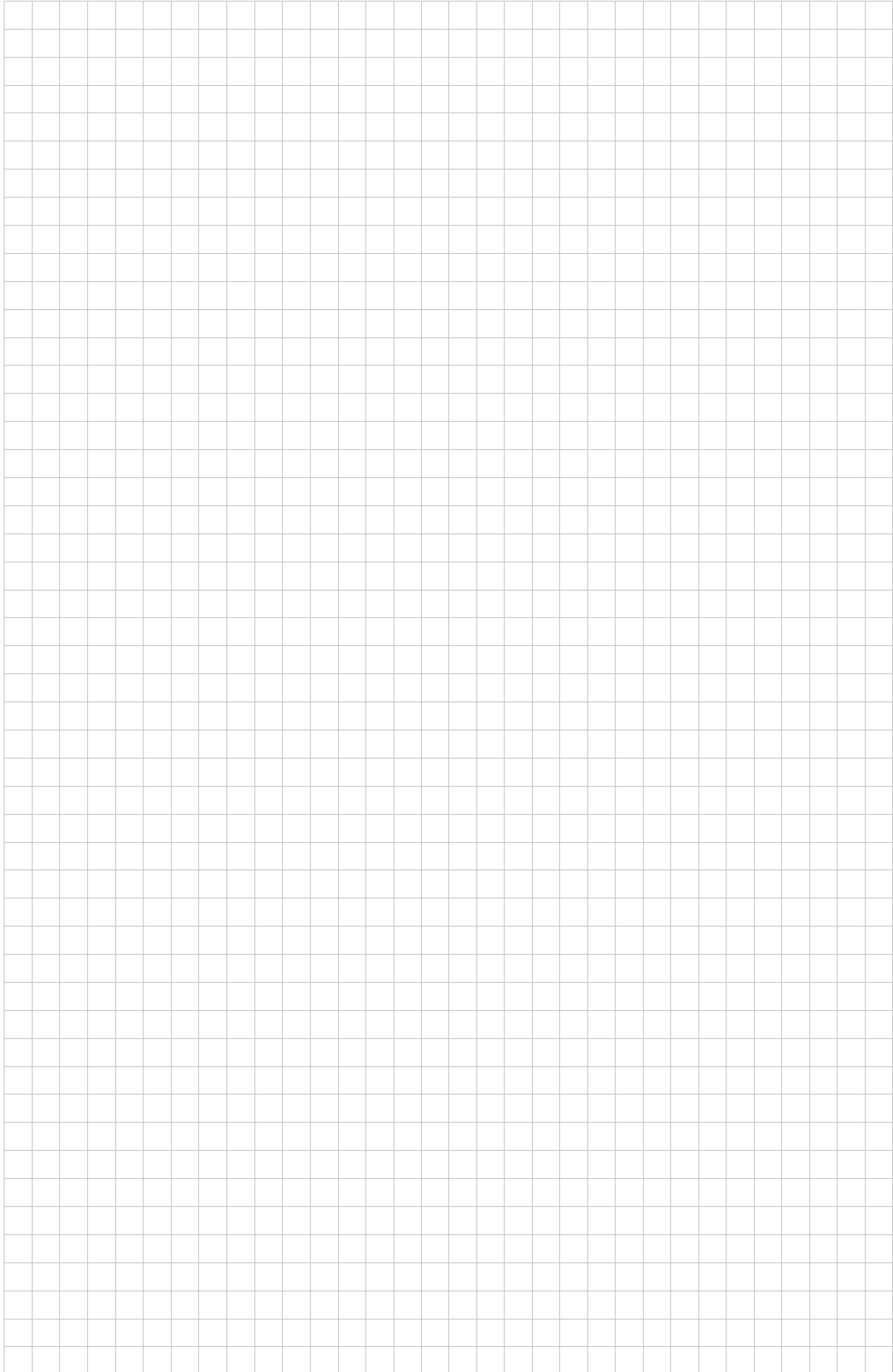
R1. Généralement dans la littérature mathématique de tels diagrammes sont sous-entendus comme étant commutatifs.

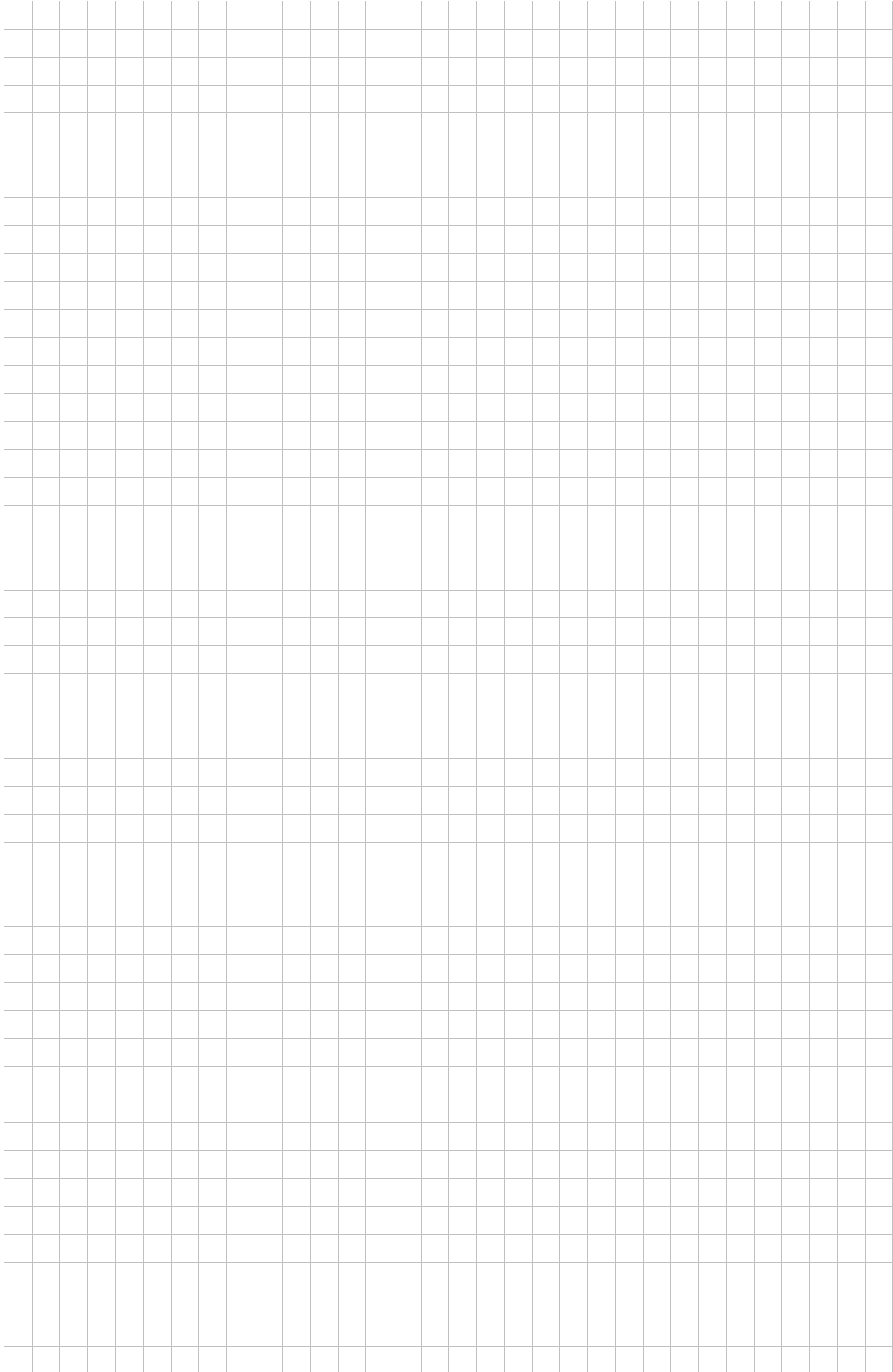
R2. Comme déjà mentionné, les objets de ces diagrammes peuvent plus généralement être des groupes d'anneaux des espaces topologiques, etc. Dans ces cas, les flèches ne sont plus des applications quelconques mais respectivement des homomorphismes de groupes, des homomorphismes d'anneaux, des applications continues, etc.

Notes personnelles:











Mécanique

La mécanique est la branche de la physique qui a pour objet l'étude des forces et de leurs actions sous une forme abstraite (Larousse)

28. PRINCIPES

En introduisant la mécanique nous faisons enfin, après une longue incursion préalable et obligatoire dans la mathématique, nos premiers pas dans le domaine de la physique théorique... version simplifiée...

Comme nous l'avons déjà dit, la physique est donc une science fondamentale qui a une profonde influence sur toutes les autres sciences et sur la société humaine et qui a pour objectif d'expliquer le comment et non le pourquoi (voir l'introduction du site pour plus de détails). Les futurs physiciens et les futurs ingénieurs ne sont ainsi pas les seuls qui doivent avoir parfaitement compris ses idées fondamentales, mais tous ceux qui envisagent une carrière scientifique (y compris les étudiants qui se spécialisent en biologie, en chimie et en mathématiques) doivent avoir acquis la même compréhension.

Le but premier de ce site, nous insistons..., est donc de donner à l'étudiant une vue unifiée de la physique en présentant ce que nous pensons être les idées fondamentales constituant l'essentiel (minimum minimorum) de la physique contemporaine.

Jusqu'à présent, la physique est enseignée comme si elle était une juxtaposition de plusieurs sciences, plus ou moins bien reliées, mais sans aucun réel souci d'unité. Nous avons rejeté ce mode de présentation (pour l'avoir subi pendant nos études) et avons opté pour une présentation logique et unifiée en faisant au besoin à chaque fois référence à un chapitre du site qui contiendrait les démonstrations des outils mathématiques utilisés ou d'une autre théorie physique sous-jacente.

Ce site diffère des supports habituels de physique utilisés à l'université non seulement dans sa conception mais aussi dans son contenu. Nous y avons inclus des sujets fondamentaux que nous ne trouvons pas dans la plupart des cours de physique générale et les avons soigneusement développés et démontrés tout en présentant de la manière la plus pédagogique et rigoureuse possible dans la section de mathématique du site, les outils nécessaires à leurs développements.

Nous insistons sur le fait que tout étudiant devrait connaître les bases de la logique, l'arithmétique, l'algèbre, l'analyse vectorielle, le calcul tensoriel, le calcul différentiel et intégral et la géométrie analytique et différentielle avant toute étude des phénomènes physiques ceci afin de travailler avec rigueur et toute la compréhension nécessaires aux raisonnements mathématiques qui vont être introduits à partir de maintenant (les mathématiques sont les fondations de l'immense édifice de la physique théorique). Effectivement, pas un des outils ou résultats mathématiques présentés jusqu'à maintenant ne sera pas utilisé dans ce qui va suivre.

Rappelons tout de même que la "physique" est donc la "science exacte/déductive" qui s'occupe de modéliser mathématiquement au mieux les phénomènes naturels, artificiels, observables ou non-observables. En de plus brefs termes, nous pourrions parler de description de la "réalité" (quant à savoir s'il s'agit de la réalité sensible ou vraie...).

Lorsque nous voulons prédire ou décrire un phénomène physique concret, nous pouvons généralement passer par un modèle analytique où les différentes grandeurs sont exprimées par des indéterminées (valeurs abstraites) et les lois de la physique par des fonctions, dans la mesure où elles sont connues (le cas échéant, nous pouvons faire une hypothèse et la tester). En mettant en équation un phénomène

physique, nous traduisons la réalité en une expérience mathématique, virtuelle, selon certaines règles. Nous procédons à une simulation de la réalité portant sur des grandeurs exprimées.

Les différentes "lois" sont élaborées historiquement très souvent sur des faits d'abord empiriques et sont vérifiées expérimentalement par la suite (voir la méthode hypothético-déductive dans le chapitre de Théorie De La Démonstration). En admettant que ces lois soient valables dans le contexte, nous pouvons donc nous attendre à ce que l'expérience mathématique soit en adéquation avec les faits expérimentaux attendus (ou inversement). Bien sûr, une expérience virtuelle n'est pas réelle et ne saurait exprimer la réalité dans toute sa subtilité. Ce n'est qu'un modèle ! Il est donc clair que la prédiction d'un phénomène physique peut diverger des faits expérimentaux réels.

Remarque: Il convient peut-être de rappeler (faute d'un abus ou d'une mauvaise compréhension que nous retrouvons trop fréquemment sur les divers forums de l'Internet), qu'une "**expérience scientifique**" est un travail pratique de l'étude d'un phénomène qui est reproductible (par des groupes de chercheurs indépendants) et dont le nombre de reproductions est suffisamment élevé pour s'assurer que les erreurs (écarts-types) sur les mesures deviennent négligeables.

Il convient aussi de préciser que la plupart des modèles théoriques que nous allons exposer sur ce site et qui font usage de l'analyse vectorielle peuvent être réécrits avec les outils de l'analyse tensorielle et basés sur un raisonnement propre au formalisme Lagrangien (voir chapitre de Mécanique Analytique pour ce savoir ce qu'est cela...). Or, ces dernières méthodes ne peuvent être facilement utilisées pour une introduction simple à la physique car elles demandent des efforts supplémentaires de la part du lecteur et beaucoup plus de papier (souvent en tout cas) et de temps pour les mêmes résultats. Cependant, et nous y reviendrons, ces méthodes sont aujourd'hui incontournables et de première importance dans les différents domaines de la physique moderne comme la mécanique des fluides, la relativité générale, la physique quantique des champs, l'analyse de systèmes chaotiques et bien d'autres.

Avant de commencer notre étude des phénomènes physiques, il nous faut définir les concepts sur lesquels se base la physique théorique. Ainsi, nous verrons dans l'ordre que:

- L'être humain a créé un système d'unités de mesures et de dimensions de bases, dont les grandeurs représentatives sont arbitraires à un coefficient près, propres à identifier chaque phénomène physique de façon simple.
- Certains concepts indissociables de la vision de notre environnement nous amènent à poser des hypothèses et des principes (à postuler quelque chose donc..) qui sont relatifs à notre réalité sensible tout en étant transposable à toute autre réalité de ce type.
- La physique fondamentale nous amène à considérer les fondements de la nature en tant que concepts mathématiques abstraits. Ainsi, notre observation commune nous donne une vue concrète de l'Univers alors que la physique théorique nous en donne une vue abstraite.

Nous pouvons alors être amenés à nous poser cependant la question suivante: les faits déterminent-ils quelle théorie est vraie ?

En observant la nature, nous pouvons constater des faits: ce sont des données que nous ne créons pas. Les astronomes par exemple, constatent la position des objets célestes. Nous comprenons un fait dans

la mesure où il apparaît comme la conséquence de l'ordre des choses décrit par une théorie. Mais les théories gardent toujours le statut d'hypothèses: même lorsqu'une théorie s'accorde avec l'ensemble des faits observés, cela ne prouve pas qu'elle soit vraie. En effet, il existe toujours une infinité de théories possibles qui sont toutes compatibles avec tous les faits observés. Nous disons alors que les faits "sous-déterminent" les théories: les faits imposent des contraintes sur les théories, au sens où, seules les théories compatibles avec les faits observés, sont acceptables. Mais ces contraintes seront toujours assez faibles pour laisser le choix parmi une infinité de théories.

Bien entendu, les scientifiques n'envisagent réellement qu'un nombre fini de théories, en fonction de ce qui paraît le plus simple dans un cadre conceptuel donné. A l'époque de Kepler, il existait par exemple trois grandes théories pour expliquer les mouvements des planètes, toutes compatibles avec les faits observés. Selon la théorie ptoléméenne, les orbites des planètes sont circulaires ou situées sur un épicycle autour de la Terre, immobile au centre de l'Univers. Dans la théorie copernicienne, le Soleil occupe le centre, les orbites des planètes et de la Terre étant situées sur des cercles et épicycles. Enfin, dans la théorie képlérienne, les orbites des planètes sont des ellipses dont le Soleil occupe un foyer.

1. SYSTÈMES D'UNITÉS

Définition: Une "**grandeur**" est l'expression nomologique quantitative d'une propriété, d'un effet ou d'une quantité abstraite définie par un modèle que présente l'objet ou le phénomène étudié. Une grandeur ne s'explique pas, elle se décrit par rapport à une définition.

Nous reconnaissons deux types de grandeurs:

- Les constantes: elles possèdent une valeur concrète exprimable numériquement et n'évoluent pas au cours du phénomène étudié. Ce sont des "**grandeurs passives**" (nous y reviendrons plus loin et énumérerons quelques-unes d'entre elles).
- Les variables: elles ne possèdent une valeur concrète que dans un état déterminé, mais pas lorsque nous observons le phénomène physique dans son ensemble. Ce sont des "**grandeurs actives**". Les différentes variables décrivant un phénomène physique sont souvent corrélées entre-elles par le biais de fonctions. Nous disons alors par définition que ces variables ont une "**relation fonctionnelle**" entre elles.

Remarque: Une grandeur n'a de sens que si elle est "**observable**", grandeur à laquelle nous associons un nombre, résultat d'une mesure effectuée à l'aide d'un appareil.

Mesurer une grandeur physique revient à la comparer à une grandeur physique connue, de même nature (nous disons aussi "de même dimension"), pris comme étalon arbitraire. Le résultat de la mesure s'exprime ainsi à l'aide de deux éléments:

- un nombre qui est le rapport de la grandeur mesurée à la grandeur étalon
- un nom identifiant l'étalon choisi

Le "nombre" constitue au fait la valeur mesurée de la grandeur et le "nom" est ce que nous appelons communément "unité physique", ou plus simplement "unité" (l'expression quantitative d'une variable ou d'une fonction). Ces deux éléments sont indissociables, la valeur mesurée n'a de sens que si nous indiquons en même temps l'unité choisie. Elle change si nous changeons d'unité.

Remarque: Le passage d'une unité à une autre pour exprimer une même grandeur est appelé "conversion d'unité".

Définition: Certaines grandeurs peuvent, par souci de simplification d'écriture, s'exprimer à partir d'autres grandeurs. Nous disons alors que la nouvelle grandeur "dérive" des unités de bases. Nous disons également que deux grandeurs physiques sont des "grandeurs homogènes" si elles sont de même nature physique ou si nous pouvons les exprimer toutes les deux dans la (les) même(s) unité(s) de base.

Ainsi, après un longue période de réflexion, et en dernière analyse le monde physique semble pouvoir se ramener aux concepts d'espace, d'énergie et de temps.

Ainsi apparaît donc une autre définition possible de la physique:

Définition: La physique est la science des propriétés et des relations mutuelles dans le temps de la matière et de l'énergie à un facteur de charge près.

Notre rôle consiste donc à donner une description de ces propriétés et relations sous forme de lois ou relations physiques appliquées aux phénomènes observés, dans le cadre d'une théorie fournissant les éléments de prévision.

Les grandeurs physiques ne sont pas toutes indépendantes les unes des autres mais reliées entre elles par certaines lois ou relations. Il serait alors peu raisonnable, quoique possible, de choisir une unité particulière pour chacune des grandeurs physiques sans tenir compte de leurs relations mutuelles.

Constituer un système cohérent d'unités revient donc à déterminer un nombre minimum d'unités qui établissent les règles de construction de ces relations mutuelles. Ce sont les "unités fondamentales". A partir des lois physiques et des relations entre les différentes unités fondamentales, nous déduisons les unités des autres grandeurs qui deviennent alors par souci de simplification d'écriture les "unités dérivées".

Les unités fondamentales sont au nombre de quatre (nous le justifierons plus loin): la longueur (mètres), la masse (kilogrammes), le temps (secondes), la charge électrique (coulombs). Le système ainsi constitué est le système M.K.S.C. (l'auteur du présent site internet assume le choix d'ajouter le Coulomb).

Les unités du système M.K.S.C sont dans le cadre ce site internet :

1. Le mètre $[m]$, pour la longueur L (nous avons déjà défini le concept de longueur dans le chapitre Géométrie mais nous y reviendrons à nouveau plus loin).
2. Le kilo $[kg]$, pour la masse M (nous reviendrons plus loin sur la définition du concept de masse).
3. La seconde $[s]$, pour le temps T (le temps n'est pas mesurable en soi mais l'intervalle de temps Δt est un concept arbitraire tout à fait valable - nous reviendrons également plus loin sur la définition de ce concept).
4. Le coulomb $[C]$ utilisé comme unité élémentaire de charge électrique q (ne dérive d'aucune unité connue à ce jour - nous reviendrons également plus loin sur la définition de ce concept).

Remarques:

R1. Le concept d'angle θ (en radians, degrés ou stéradians - voir les textes traitant de la trigonométrie plane, trigonométrie sphérique et géométrie plane dans la section de géométrie) n'a pas d'unité puisqu'il s'agit par définition d'un rapport de longueurs (pour le radian ou le degré d'angle) ou de surface (pour le stéradian). Il convient donc de l'assimiler à une unité dérivée et non pas comme unité fondamentale. Cependant, en physique, nous avons pris pour habitude d'indiquer sa présence dans les équations dimensionnelles afin d'aider à la relecture de certaines de celles-ci et de savoir que leur résultat est donné par rapport à une unité d'angle (sinon cela pourrait générer des erreurs d'interprétation hasardeuses pour ceux qui utilisent des équations sans en avoir vu la démonstration...).

R2. Le lecteur remarquera que toutes les unités du système M.K.S.C. sont des "**grandeurs extensives**" c'est-à-dire que dans un système sur lequel nous effectuons une mesure, celles-ci sont additives (contrairement aux grandeurs intensives). Nous reviendrons plus en détails sur les grandeurs extensives et intensives en grande partie lors de notre étude de la thermodynamique (voir chapitre du même nom).

R3. C'est une énorme chance d'avoir un système homogène tel que celui que nous avons au 21ème siècle. Effectivement, pour l'anecdote, en 1522 rien que dans la région de Baden-Baden (Allemagne) il y avait 112 unités de mesures différentes de longueur et 92 de surfaces.... c'est dire... le cauchemar! En ce début de 21ème siècle il existe une série de normes ISO 80000 - dont l'accès est malheureusement payant... - qui a pour objectif d'harmoniser les notations, définitions et valeurs des unités dans tous les domaines de la science

Ces précisions étant faites, toute grandeur physique connue à ce jour peut être exprimée à l'aide d'une unité qui s'exprime comme le produit de cinq facteurs dimensionnels et d'un facteur d'échelle arbitraire K :

$$K \cdot [M]^m \cdot [L]^l \cdot [T]^t \cdot [\theta]^\theta \cdot [C]^c \quad (28.1)$$

où les nombres m, l, t, θ, c appelés respectivement "**ordre de masse**", "**ordre de longueur**", "**ordre de temps**", "**ordre d'angle**" et "**ordre de charge**" sont des entiers positifs, négatifs ou nuls.

L'expression précédente s'écrira sous la "**forme canonique**" définie par les étalons:

$$K \cdot [kg]^m \cdot [m]^l \cdot [s]^t \cdot [C]^c \quad (28.2)$$

l'angle n'ayant pas d'unité, nous ne le notons plus (mais il s'y trouve implicitement).

Toute grandeur physique X s'exprime donc comme:

$$X = x \cdot K \cdot [kg]^m \cdot [m]^l \cdot [s]^t \cdot [C]^c \quad (28.3)$$

où x est la valeur de la grandeur physique dans le système d'unité associé au facteur d'échelle K . Il existe plusieurs couples (x, K) possibles, mais nous aurons toujours:

$$x \cdot K = x_0 = c^{te} \quad (28.4)$$

où la constante x_0 est la valeur de la grandeur physique lorsque nous choisissons de l'exprimer dans le système M.K.S.C.

Donc deux grandeurs physiques X_1 et X_2 sont homogènes si et seulement si les quadruplets:

$$(m_1, l_1, t_1, c_1), (m_2, l_2, t_2, c_2) \quad (28.5)$$

qui leur sont attachés sont égaux:

$$m_1 = m_2 \quad l_1 = l_2 \quad t_1 = t_2 \quad c_1 = c_2 \quad (28.6)$$

Il découle, de ce que nous avons dit, que:

- La somme ou la différence d'un nombre quelconque de grandeurs n'a un sens que si ces grandeurs sont homogènes et le résultat aura donc les mêmes unités que les opérandes.
- Le produit ou la division de plusieurs grandeurs a pour unité le produit, respectivement la division des unités des opérandes.

Remarques:

R1. Les unités des différentes grandeurs ont un côté pratique mais pas infallible en physique théorique: elles permettent cependant au physicien de vérifier si une relation démontrée entre deux grandeurs est au moins correcte au niveau des unités. Nous appelons ce genre de démarche une "**analyse dimensionnelle**" (nous vous conseillons d'aller voir la démonstration de la loi de Stokes dans le chapitre de Mécanique Des Milieux Continus pour un très bon exemple d'application).

R2. Le développement des sciences a conduit la conférence générale des poids et mesures à introduire quelques unités supplémentaires pratiques (mais pas nécessaires) telles que: la température exprimée en "Kelvins" (qui dérive de l'énergie moyenne - mouvement brownien), la quantité de matière exprimée en Moles, l'intensité de courant exprimée en Ampères et l'intensité lumineuse exprimée en Candela. Ainsi, le Système International (S.I.) actuel, composé de sept unités de base (centimètre, gramme, seconde, Kelvin, candela, mole et l'Ampère) et de dix-sept unités dérivées suggère-t-il que sept unités sont nécessaires pour décrire toute la physique ? En fait non ! Comme l'analyse de Gauss le suggère, parmi les sept unités de base, quatre - le Kelvin, le Candela, la Mole et l'Ampère - peuvent être dérivées des trois autres. L'introduction de sept unités de base représente un équilibre pragmatique entre des expérimentateurs qui ont besoin d'unités adaptées à leurs mesures, et l'idéalisme des théoriciens, dont le but est de réduire l'arbitraire, la redondance, à son minimum.

1.1. ANALYSE DIMENSIONNELLE

L'analyse dimensionnelle est donc un domaine de la physique qui concerne les unités des grandeurs. Notamment, le fait que les unités soient relativement arbitraires fait que toute équation valable de la physique est homogène: quelque chose qui se mesure en mètres par seconde ne peut pas être égal à quelque chose qui se mesure en kilogrammes par mètre. C'est un moyen très prisé et très efficace de

vérifier ses propres calculs (et celui des autres...).

La puissance prédictive de cette approche valable dans des cas d'études simples a amené certains physiciens à énoncer le "**principe zéro**" de la physique ainsi: Ne jamais faire de calculs avant d'en connaître le résultat.

Cet énoncé, qui peut sembler a priori paradoxal, signifie concrètement: Ne pas se lancer (si possible...) dans un calcul compliqué sans avoir trouvé au préalable la forme qualitative du résultat avec l'analyse dimensionnelle.

Cette forme qualitative est nommée traditionnellement "**l'équation aux dimensions**" et représente donc la formule qui permet de déterminer l'unité dans laquelle doit être exprimé le résultat d'une recherche. C'est une équation de grandeurs, c'est-à-dire dans laquelle on représente les phénomènes mesurés par un symbole d'unité comme ceux que nous avons vus dans les paragraphes plus haut.

Exemple:

Voyons donc un exemple de légende souvent cité dans divers magazines ou livres de vulgarisation:

L'analyse dimensionnelle a permis à Geoffrey Ingram Taylor d'estimer en 1950 l'énergie dégagée par l'explosion d'une bombe atomique, alors que cette information était classée top secret. Il lui aurait suffi pour cela d'observer sur un film d'explosion, imprudemment rendu public par les militaires américains.

Le physicien Taylor suppose pour arriver à ce résultat que le processus d'expansion de la sphère de gaz dépend au minimum des paramètres du temps t , de l'énergie E dégagée par l'explosion et de la masse volumique de l'air ρ .

L'analyse dimensionnelle le conduit alors pour le rayon de la sphère de gaz à l'instant t à :

$$r[m] = k \cdot \left(E[kg \cdot m^2 \cdot s^{-2}] \right)^x \cdot \left(\rho[kg \cdot m^{-3}] \right)^y \cdot (t[s])^z \quad (28.7)$$

où k est une constante sans dimensions.

Et par tâtonnements nous trouvons relativement rapidement $x = 1/5, y = -1/5, z = 2/5$ tels que:

$$r = k \cdot E^{1/5} \cdot \rho^{-1/5} \cdot t^{2/5} \quad (28.8)$$

Effectivement:

$$m = \left(kg^{1/5} \cdot m^{2/5} \cdot s^{-2/5} \right) \cdot \left(kg^{-1/5} \cdot m^{3/5} \right) \cdot \left(s^{2/5} \right) \quad (28.9)$$

Taylor trouve alors la loi de dilatation temporelle du rayon du champignon atomique est proportionnelle à (il est inutile d'indiquer les autres unités puisque uniquement la partie temporelle nous intéresse!):

$$r \propto t^{2/5} \quad (28.10)$$

Si nous connaissons r et t à partir d'un film, et, k étant supposée de l'ordre de l'unité et ρ étant connue, nous obtenons finalement:

$$E \cong \frac{\rho r^5}{t^2} \quad (28.11)$$

ce qui reste une grossière approximation. Mais arriver à un résultat pareil (d'ordre de grandeur) avec l'artillerie lourde de la physique théorique nécessiterait beaucoup plus de temps et de feuilles de calculs.

1.2. NOTATIONS SCIENTIFIQUES

Il est fréquent en physique que les grandeurs manipulées soient très grandes et lourdes à écrire. Par exemple, il est toujours embêtant d'avoir des grandeurs comme 8'000'000'000 ou 0.000'000'000'1.

Alors nous pouvons adopter une convention d'écriture en puissance de dix dite "**notation scientifique**" telle que:

- 8'000'000'000 s'écrit $8 \cdot 10^9$ (neufs zéros après le "8")

- 0.000'000'000'1 s'écrit $1 \cdot 10^{-10}$ ($10^{\text{ème}}$ position après la virgule) ou $0.1 \cdot 10^{-9}$ (neufs zéros après la virgule)

Une écriture encore plus simplifiée consiste à utiliser le tableau ci-dessous mais uniquement si nous avons à travailler avec des grandeurs physiques:

Préfixe	Facteur	Symbole	Préfixe	Facteur	Symbole
exa	10^{18}	E	déci	10^{-1}	d
péta	10^{15}	P	centi	10^{-2}	c
téra	10^{12}	T	milli	10^{-3}	m
giga	10^9	G	micro	10^{-6}	μ
méga	10^6	M	nano	10^{-9}	n
kilo	10^3	k	pico	10^{-12}	p
hecto	10^2	h	femto	10^{-15}	f

déca	10^1	da	atto	10^{-18}	a
------	--------	----	------	------------	---

Tableau: 28.1 - Préfixes des grandeurs d'ordre

Par exemple, 10'000'000 grammes notés conventionnellement:

$$10'000'000 [g] \quad (28.12)$$

sera écrit en notation scientifique:

$$1 \cdot 10^7 [g] \quad (28.13)$$

mais en écriture physique (selon le tableau ci-dessus):

$$1 \cdot 10^4 [Kg] \text{ ou } 10 [Mg] \quad (28.14)$$

Définition: Nous disons que $a \cdot 10^n$ est "**l'écriture scientifique**" d'un nombre positif A si a est un nombre décimal $1 \leq a < 10$ (c'est-à-dire que a s'écrit avec un seul chiffre autre que zéro avant la virgule), n est un nombre entier relatif.

 **Exemple:**

$$1'350 = 1.35 \cdot 10^3 \quad 345.58 \cdot 10^{12} = 3.4558 \cdot 10^{14} \quad 325 \cdot 10^{-4} \Leftrightarrow 3.25 \cdot 10^{-2} \quad (28.15)$$

L'avantage de cette écriture est de donner un ordre de grandeur de A compris entre 2 puissances consécutives de 10 tel que:

$$10^3 < 1.35 \cdot 10^3 < 10^4 \quad 10^{14} < 3.4558 \cdot 10^{14} < 10^{15} \quad 10^{-2} < 3.25 \cdot 10^{-2} < 10^{-1} \quad (28.16)$$

Si de plus, comme il arrive souvent, nous utilisons des unités de physiques de multiple de 1'000 cela permet de placer ces grandeurs entre 2 unités dérivées consécutives.

Remarques:

R1. Si nous avons un chiffre de la forme 154'434'347'786, fréquemment et selon le contexte, nous nous permettons de tronquer ce dernier et nous écrivons alors fréquemment avec une précision de trois chiffres après la virgule ainsi ce dernier nombre devient $154.434 \cdot 10^9$ ce qui est plus simple à écrire mais dangereux à manipuler à cause de l'erreur induite par la troncature. Nous renvoyons à ce sujet le lecteur dans le chapitre de Statistiques à la lecture de la partie traitant des erreurs relatives.

R2. Pour les mathématiciens la notation scientifique n'est qu'une écriture d'un nombre parmi d'autres et le choix de cette écriture est en relation avec le contexte du problème. Evidemment ces "nombres résultats" obtenus peuvent être des nombres purs et durs solution de problèmes abstraits mais aussi de problèmes concrets issus d'expériences, de mesures etc. et là nous nous rejoignons les physiciens.

1.3. TEMPS

Définition: Le "**temps**" est une variable d'état (et non un "mesurable") et donc une notion impalpable mais cependant rigoureusement définie. Il s'agit aussi d'un outil mathématique qui permet de mettre en équation l'observation de phénomènes physiques (observables) et d'en tirer ainsi un certain nombre d'informations. Cet outil existe car il existe des êtres pour observer (et mesurer) la nature et ses changements (principe socratique) et de la matière et du mouvement pour qu'il y ait ces changements.

Remarques:

R1. Le temps (et ses intervalles) étant un concept arbitraire, il est symétrique c'est-à-dire que tout phénomène observé enregistré peut dans l'imaginaire du temps inversé retrouver ces conditions initiales. Nous parlons alors de "**symétrie du temps**" (pour l'instant il n'a jamais été à notre connaissance démontré ou ne serait-ce qu'observé, que le temps peut subir une "brisure de symétrie").

R2. Le temps n'est une grandeur ni extensive ni intensive. On ne peut ni additionner le temps des éléments d'un système physique pour avoir la durée totale de celui-ci (de plus cette question n'a pas de sens) ni la pondérer. Cependant on peut additionner les intervalles de temps qui décrivent l'évolution d'un système!

Nous représentons très souvent en physique le temps (compris dans un intervalle) par une flèche (axe) horizontale représentant le sens du temps. Comme le temps est une notion purement utilitaire, nous pouvons alors définir chaque instant du temps comme étant le temps zéro noté t_0 . Cette notion est très utilisée en physique car souvent la seule chose qui intéresse les physiciens est la différence de temps notée Δt (de par l'utilisation du calcul différentiel et intégral).

Démontrons maintenant que la référence temporelle est indépendante du choix pour un observateur au repos. Soit un temps noté par la lettre t , nous avons alors:

$$\Delta t = (t_1 + t') - (t_0 + t') = t_1 - t_0 \quad (28.17)$$

où t' est la base arbitraire (non nécessaire) lorsque nous comparons une différence temporelle.

L'intervalle de temps est donné par une mesure étalon qui ne peut être qu'un mouvement au mieux parfaitement périodique (qui se répète dans le temps). Ainsi, les premiers moyens de mesure du temps ont été le jour et la nuit, les positions du soleil et de la lune dans le ciel, le mouvement du pendule, la détente de ressorts, la période de dégénérescence du césium 137 ou encore les systèmes binaires d'étoiles massives. Bref, tant qu'un système observable produit un phénomène périodique stable et suffisamment petit pour que toute mesure physique puisse y être réduite, celui-ci peut être utilisé comme étalon d'intervalle temporel.

Définitions:

D1. Un "**événement**" consiste à donner une signification à un point de l'espace-temps.

D2. Deux événements sont dits "**événements simultanés**", s'ils ont même valeur de la coordonnée temporelle.

D3. Nous appelons "**coïncidence**" la simultanéité de deux événements en un même point de l'espace. La coïncidence est un fait absolu, indépendant du choix du référentiel. C'est en fait un cas particulier du principe de conservation de la causalité. Deux événements coïncidant dans un repère peuvent être cause à effet l'un de l'autre (et réciproquement), et cette possibilité est conservée dans le nouveau repère.

1.4. LONGUEUR

Définition: Le concept de "**longueur**" x est donné par l'information qui donne le chemin parcouru par un objet dans un intervalle de temps donné.

Remarques:

R1. S'il n'y avait pas de matière dans l'Univers il n'y aurait pas de notion de mouvement et donc de longueur parcourue et aussi comme nous l'avons déjà fait remarquer, de temps (et encore... c'est sans considérer certains résultats de la physique quantique que nous démontrerons dans le chapitre qui lui est consacré).

R2. La longueur est une grandeur extensive (additive). Effectivement, la longueur totale d'un système est la somme des grandeurs.

Comme pour le temps il n'y a pas d'origine absolue de mesure des longueurs (il n'existe pas de point zéro dans l'Univers comme le postule la théorie de la relativité) et les physiciens s'intéressent de toute façon plus particulièrement aux différences de chemin parcouru Δx par rapport à une origine comme ils le font pour le temps.

Ainsi, de manière identique au temps, nous avons pour un observateur au repos qui observe un point matériel en mouvement:

$$\Delta x = (x_1 + x') - (x_0 + x') = x_1 - x_0 \quad (28.18)$$

où x' est une base arbitraire mathématiquement inutile lorsque nous comparons une différence de position dans une différence de temps d'un point matériel.

Si un point matériel se situe dans un espace à trois dimensions spatiales (cas le plus fréquent en mécanique classique) dont nous avons arbitrairement choisi l'origine O , nous notons la position $\vec{r}$ de ce corps par sa distance en longueur x , largeur y et hauteur z (appelées "coordonnées cartésiennes") par une flèche imaginaire dit "vecteur" (cf. chapitre de Calcul Vectoriel) reliant le point d'origine arbitraire du référentiel spatial au point intéressé de la façon suivante:

$$\vec{r} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (28.19)$$

Remarque: La flèche au-dessus du $\vec{r}$ signifie bien évidemment qu'il s'agit d'un vecteur.

La notation:

$$(x, y, z) \equiv (x_1, x_2, x_3) \quad (28.20)$$

est une notation simplificatrice utilisée fréquemment en physique et qui devrait s'imposer dans les petites classes (attention le fait que les chiffres soient abaissés en indice ne signifie absolument pas ce que sont des composantes covariantes - voir chapitre de Calcul Tensoriel de la section d'Algèbre - il s'agit juste d'une convention simplificatrice d'écriture). Cependant sur le présent site Internet nous passerons de l'une à l'autre des notations en fonction des besoins et des traditions en vigueur (ce sera donc à vous de faire attention à ne pas confondre).

La matrice:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (28.21)$$

est quant à elle le tenseur métrique d'un espace pré-euclidien canonique à signature positive (cf. chapitres de Calcul Vectoriel et Calcul Tensoriel). Ceci constitue un cas particulier en physique théorique mais cependant un cas très fréquent d'étude en mécanique classique (il faut commencer par des espaces simples avant d'aller plus loin...).

Nous reviendrons plus en détails sur ces concepts lors de notre étude des espaces ponctuels plus loin.

1.5. MASSE

Définitions:

D1. La "**masse**" m d'un corps est dans un système fermé une quantité qui se conserve et qui caractérise l'amplitude avec laquelle ce corps interagit avec d'autres corps par le biais de différentes forces (attractives).

Remarques:

- R1. Dans un système isolé, il ne peut pas y avoir création ou destruction spontanée de masse. L'apparition de masse ne peut être due qu'à une action extérieure. Une autre façon de dire la même chose est que la masse totale contenue dans l'Univers est constante.
- R2. La masse est une grandeur extensive (additive). Effectivement, la masse totale d'un système est égale à la somme des masses qui le compose.

En toute rigueur, nous devrions définir également:

D2. La "**masse grave**" (ou "masse de gravitation") qui est l'amplitude avec laquelle un corps matériel interagit avec un champ de potentiel (selon la loi de gravitation de newton - voir chapitre de Mécanique Classique).

D3. La "**masse inerte**" (ou "**masse inertielle**") qui est l'amplitude qui caractérise la résistance avec

laquelle un corps en translation est susceptible de changer de vitesse (c'est-à-dire celle intervenant dans la deuxième loi de Newton - voir chapitre de Mécanique Classique)

Remarques:

R1. Des expériences ont toutefois prouvé que ces deux masses étaient proportionnelles au dix-milliardième près. Cette identité expérimentale appelée "**principe d'équivalence galiléen**" est à la base d'un des postulats de la relativité générale (cf. chapitre de Relativité Restreinte).

R2. Contrairement aux charges électriques (voir plus loin la définition de la "charge"), qui caractérisent l'amplitude d'interaction par la force électrique, il n'existe que des masses positives. En effet, les charges électriques peuvent se repousser aux mêmes titres qu'elles peuvent s'attirer.

De plus, la masse est étant une propriété additive (donc "extensive" comme nous l'avons déjà dit) de la matière: pour un système de n points matériels de masse m_i ($i = 1, \dots, n$), la masse totale est :

$$M_{tot} = \sum_{i=1}^n m_i \quad (28.22)$$

De même, pour une distribution continue (voir plus loin au cas où pour un rappel du concept de distribution continue) en volume de la masse d'un système de volume total V :

$$M_{tot} = \iiint_V \rho_V(A) dV = \iiint_{V=\{A|\vec{r}=\vec{OA}\}} \rho_V(\vec{r}) \circ d^3\vec{r} \quad (28.23)$$

où $\rho_V(A)$ est la "**masse volumique**" ou "**densité volumique**" du système au point A et où $\rho_V(\vec{r})$ est la masse volumique du système au point repéré par $\vec{r} = \vec{OA}$ (c'est ce que signifie l'expression entre accolades en-dessous de la deuxième triple intégrale).

Donc $\rho_V(A)$ est la masse d'un élément de matière, centré autour de A , de dimensions caractéristiques devant celles du système, mais grandes devant les distances interatomiques dans ce système définie par:

$$\rho_V = \frac{dm}{dV} \quad (28.24)$$

Remarques:

R1. L'intégrale est une intégrale triple (sur les trois dimensions de l'espace), mais elle pourra être ramenée à une intégrale simple en exploitant les symétries du système pour choisir judicieusement les volumes élémentaires d'intégration.

R2. Le calcul de la masse lors d'une distribution non continue (discrète) de matière doit être fait avec les composantes vectorielles calculées séparément. Une fois ce travail effectué, il convient d'en prendre la norme.

R3. La masse volume $\rho_v(A)$ est une grandeur intensive. Effectivement, la densité d'un système physique n'est pas égale à la somme de ces densités (c'est du bon sens!). Le lecteur remarquera que cette grandeur intensive qu'est la masse volumique est égale au rapport de deux grandeurs extensives.

Définitions:

D1. Nous disons qu'un système est un "**système homogène**" si sa masse volumique, surfacique, linéique (voir définition ci-dessous) est constante.

D2. Nous disons qu'un système est un "**système isotrope**", si ses propriétés physiques sont identiques en tout point.

Nous définissons aussi parfois la "**masse surfacique**" (ou "**densité surfacique**" de masse) pour des systèmes quasiment sans épaisseur et une "**masse linéique**" (ou "**densité linéique**" de masse) pour des systèmes de section négligeable devant leur longueur. Nous avons alors (S étant une surface et s une abscisse curviligne):

$$M_{\text{tot}} = \iint_S \rho_s(A) dS = \iiint_V \rho_v(\vec{r}) d^3\vec{r} \quad \text{ou} \quad M_{\text{tot}} = \int_L \rho_l(A) ds = \int_L \rho_l(s) ds \quad (28.25)$$

avec dans le cas général:

$$\rho_s = \frac{dm}{dS}, \rho_l = \frac{dm}{dL} \quad (28.26)$$

Remarque: Souvent, dans la littérature, ainsi que dans le présent site internet, la masse volumique est notée simplement ρ , la masse surfacique σ , et la masse linéique γ .

Définition: Avec ce qui précède, nous pouvons définir la "**densité**" comme étant la quantité d'éléments tous identiques et dénombrables par unité de volume, surface ou linéique.

1.6. ÉNERGIE

Nous ne savons pas ce qu'est exactement l'énergie (notée sous sa forme générale par la lettre E dans les petites classes) mais nous en connaissons ses effets. Ce que nous savons cependant, c'est qu'il en existe plusieurs formes dont voici une liste des plus connues:

- "**L'énergie travail**" qui est l'énergie créée par l'application d'une force sur un corps lui donnant une certaine énergie cinétique (cf. chapitre de Mécanique Classique) ou énergie potentielle (qu'elle soit gravifique, électrostatique comme démontré en mécanique classique ou électrodynamique).
- "**L'énergie chaleur**" qui est une forme d'énergie déterminée par le nombre de micro-états d'un système (cf. chapitre de Thermodynamique).
- "**L'énergie de masse**" qui est l'énergie contenue dans une certaine quantité de masse (cf. chapitre de Relativité Restreinte).

De ces trois énergies découlent une grande quantité de familles d'énergies dérivées dont les plus connues sont: l'énergie nucléaire, l'énergie électrique, l'énergie solaire, l'énergie éolienne, l'énergie mécanique, l'énergie gravifique, l'énergie des marées, l'énergie électromagnétique, l'énergie fossile, l'énergie hydraulique, l'énergie corporelle, etc.

Remarques:

R1. La masse et l'énergie sont équivalentes comme nous le verrons lors de notre étude de la relativité restreinte (cf. chapitre de Relativité Restreinte), si nous définissons un système d'unités telles que la vitesse de la lumière vaille $c = 1$ (convention très utilisée par les physiciens dans la recherche de pointe).

R2. L'énergie au même titre que la masse est une grandeur extensive.

Nous pouvons quand même tenter de nous demander ce qu'est l'énergie exactement?

Définition: "**L'énergie**" est l'effet d'une cause d'un changement ou de la conservation des propriétés d'un système. Cette cause étant non nécessairement déterministe et en moyenne nulle et conservative dans un système fermé.

Remarques:

R1. La vitesse, le potentiel, le nombre de micro-états peuvent être considérés comme l'acquisition d'une quantité d'informations sur un système.

R2. Dans un système isolé, il ne peut pas y avoir création ou destruction spontanée d'énergie. L'apparition d'énergie ne peut être due qu'à une action extérieure. Une autre façon de dire la même chose est que l'énergie totale contenue dans l'Univers est constante.

1.7. CHARGE

Il est difficile de dire quelque chose sur la charge électrique (vous pouvez chercher une définition sur l'internet vous verrez...). Cependant si nous nous référons à l'approche de Yukawa (cf. chapitre de Physique Quantique Des Champs) nous pouvons tenter d'en donner la définition suivante:

Définition: Une "**charge électrique**" est une propriété conservative qu'a une particule se situant dans un champ de potentiel à symétrie sphérique à interagir avec la source de ce champ dans le cadre de

l'échange d'un quantum d'interaction (le photon en l'occurrence créé par les fluctuations quantiques du vide en présence d'une masse) définissant un champ vectoriel de type Coulombien.

Remarques:

R1. Dans un système isolé, il ne peut pas y avoir création ou destruction spontanée de charges. L'apparition de charges ne peut être due qu'à une action extérieure. Une autre façon de dire la même chose est que la charge totale contenue dans l'Univers est constante.

R2. La charge est une grandeur extensive. Effectivement, la charge totale d'un système physique est égale à la somme algébrique des charges qui le constitue.

Ou une autre définition similaire à celle de la masse:

Définition: La "**charge électrique**" q d'un corps est dans un système fermé une quantité qui se conserve et qui caractérise l'amplitude avec laquelle ce corps interagit avec d'autres corps par le biais des forces électrostatiques et magnétiques.

Contrairement à la masse, il existe des charges électriques positives et négatives. La charge électrique reste cependant une propriété additive (extensive). Ainsi, pour un système de q particules de charge q_i ($i = 1, \dots, n$), la charge totale est:

$$Q_{\text{tot}} = \sum_{i=1}^n \pm q_i \quad (28.27)$$

et est donc aussi comme la masse, une propriété extensive.

De même, pour une distribution continue en volume de la charge d'un système de volume total V (nous notons les densités de charge de manière identique si l'ambiguïté n'est pas possible de la même manière que pour la masse):

$$Q_{\text{tot}} = \iiint_V \rho_V(A) dV = \iiint_{V=\{A|\vec{r}=\overrightarrow{OA}\}} \rho_V(\vec{r}) \circ d^3\vec{r} \quad (28.28)$$

où $\rho_V(A)$ est la "**densité volumique de charges**" du système au point A , c'est-à-dire la charge d'un élément de matière, centrée autour de A , de dimensions caractéristiques devant celles du système, mais grandes devant les distances interatomiques dans ce système ($\rho_V(\vec{r})$ est la densité volumique de charge du système au point repéré par $\vec{r} = \overrightarrow{OA}$) définie par:

$$\rho_V = \frac{dq}{dV} \quad (28.29)$$

Remarques:

R1. L'intégrale est une intégrale triple, mais elle pourra être ramenée à une intégrale simple en exploitant les symétries du système pour choisir judicieusement les volumes élémentaires d'intégration.

R2. Le calcul de la charge lors d'une distribution non continue (discrète) de matière doit être fait avec les composantes vectorielles calculées séparément. Une fois ce travail effectué, il convient d'en prendre la norme.

R3. La charge volumique $\rho_v(A)$ est une grandeur intensive. Effectivement, la densité de charge d'un système physique n'est pas égale à la somme de ces densités (c'est du bon sens!). Le lecteur remarquera encore une fois que cette grandeur intensive qu'est la masse volumique est égale au rapport de deux grandeurs extensives.

De même que pour la masse, nous pouvons donner les définitions suivantes:

D1. Nous disons qu'un système est "**système homogène**" si sa charge volumique, surfacique, linéique (voir définition ci-dessous) est constante.

D2. Nous disons qu'un système est "**système isotrope**", si ses propriétés physiques sont identiques en tout point

Nous définissons aussi parfois la "**densité surfacique de charge**" (ou "**densité de surface**" de charge) pour des systèmes quasiment sans épaisseur et une "charge linéique" (ou "**densité linéique**" de charge) pour des systèmes de section négligeable devant leur longueur. Nous avons alors (S étant une surface et s une abscisse curviligne):

$$Q_{\text{tot}} = \iint_S \rho_s(A) dS = \iint_S \rho_s(\vec{r}) d^2\vec{r} \quad \text{ou} \quad Q_{\text{tot}} = \int_L \rho_l(A) ds = \int_L \rho_l(s) ds \quad (28.30)$$

avec:

$$\rho_s = \frac{dq}{dS}, \rho_l = \frac{dq}{dL} \quad (28.31)$$

Remarque: Souvent, dans la littérature, ainsi que dans le présent site internet, la densité volumique de charge est notée simplement ρ , la densité surfacique de charge σ , et la masse linéique γ .

1.8. DISTRIBUTIONS

Définitions:

D1. Une masse ou une charge sont dites "**ponctuelles**" si elles occupent un volume dont les dimensions sont très inférieures aux distances d'observations.

Remarque: La charge élémentaire est une excellente approximation d'une charge ponctuelle étant donné sa petite taille dont le rayon classique est de l'ordre du femtomètre, ce qui est bien sûr très petit devant les dimensions d'observation classiques.

D2. Nous considérons N corps de charge ou masse finies dans un volume V . Si ce volume est supposé suffisamment grand pour que la distance moyenne entre les corps soit très supérieure à la dimension de ceux-ci, nous avons affaire à une "**distribution discontinue**" ou "**distributions discrète**" de ces corps (nous parlons parfois aussi de distribution non-uniforme).

Les calculs sont impossibles à faire en partant d'une distribution discrète car, en général, le nombre de corps à prendre en considération est très élevé lorsque le volume est de dimension macroscopique. Dans ce cas, il faut introduire un autre type de distribution.

D3. Nous considérons N corps de charge ou masse finies dans un volume V . Si la répartition des éléments est telle qu'il n'y pas de "trous" entre chacun d'eux (en d'autres termes: chaque élément est serré contre un autre) alors nous avons affaire à une "**distribution continue**". Une distribution continue peut alors être décrite par une fonction qui représente la manière dont les éléments se répartissent dans un volume, surface ou ligne.

Remarques:

R1. Nous pouvons préciser parfois, comme nous en avons déjà fait mention lors des définitions de la masse ou de la charge que les distributions définies précédemment peuvent être de type volumique, surfacique ou linéique. Si cela n'est pas précisé, c'est que l'information est implicitement triviale.

R2. Le terme "continue" dans "distribution continue" provient du fait que nous intégrons la fonction d'où la nécessité qu'elle soit continue (au sens de Riemann ou de Lebesgue suivant les cas... - voir chapitre de Calcul Différentiel Et Intégral).

2. CONSTANTES

La physique à l'opposé des mathématiques est une science exacte dans le sens que sa vérification et sa validité se basent et sont mis constamment à l'épreuve par des faits expérimentaux.

Comme l'être humain a dû choisir arbitrairement un système de mesures, certaines lois établies théoriquement à l'aide de propriétés de la matière ne sont souvent exactes qu'à un facteur multiplicatif constant près de normalisation relativement à ce système de mesure. Apparaissent alors dans les équations de la physique, des constantes dont l'existence n'est due qu'à ce système de mesure (mais cependant ce n'est pas toujours le cas), certaines constantes bien qu'en adaptant le système de mesure n'égaleront (du moins il semblerait) jamais l'unité.

Il existe de nombreuses constantes en physique (une infinité au fait) mais certaines ont un statut particulier dans le sens qu'elles ne peuvent se déduire d'autres constantes. Nous en proposons ici la liste et les valeurs (non exactes) et nous les retrouverons fréquemment lors de nos développements dans ce site.

Remarques:

R1. Les constantes sont données pour certaines au temps auquel le lecteur les lit (...) car selon certaines théories, les valeurs ne sont pas tout à fait fixes.

R2. La série de normes ISO 80000 définit au niveau international les valeurs de nombreuses constantes scientifiques.

2.1. CONSTANTES UNIVERSELLES

Les valeurs listées ci-dessous sont des valeurs dont les scientifiques ont remarqué qu'elles semblaient (...) constantes et indépendantes de tous paramètres utilisés, et que la théorie suppose donc réellement constantes.

Constante gravitationnelle	$G \cong 6.6742 \cdot 10^{-11} [m^3 \cdot s^{-2} \cdot kg^{-1}]$
Température absolue	$T_0 \cong 273.15 [^{\circ}K]$
Vitesse (célérité) de la lumière	$c \cong 299'792'458 [m \cdot s^{-1}]$
Nombre d'Avogadro	$N_{Av.} \cong 6.0221367 \cdot 10^{23} [mol^{-1}]$
Charge de l'électron	$q_e \cong 1.602 \cdot 10^{-19} [C]$
Constante de Planck	$h \cong 6.626 \cdot 10^{-34} [J \cdot s]$
Constante de Boltzmann	$k \cong 1.38 \cdot 10^{-23} [J \cdot K^{-1}]$
Permittivité du vide	$\varepsilon_0 \cong 8.854 \cdot 10^{-12} [C \cdot V^{-1} \cdot m^{-1}]$
Susceptibilité magnétique	$\mu_0 \cong 1.26 \cdot 10^{-6} [V \cdot C^{-1} \cdot s^2 \cdot m^{-1}]$
Pi	$\pi \cong 3.141592654$
Constante de Dirac (utilitaire)	$\hbar = h / 2\pi \cong 1.05 \cdot 10^{-34} [J \cdot s]$
Constante de Coulomb (utilitaire)	$k = 1/4\pi\varepsilon_0 = 8.987 \cdot 10^9 [N \cdot m^2 \cdot C^{-2}]$

Tableau: 28.2 - Constantes universelles

Remarques:

R1. La célérité de la lumière, la permittivité du vide et la susceptibilité magnétique du vide se déduisent les uns des autres par une relation que nous verrons lors de notre étude des équations de Maxwell (cf. chapitre d'Électrodynamique).

R2. La constante de Dirac est aussi parfois appelée "**constante de Planck réduite**".

R3. La constante de Boltzmann peut être calculée comme le rapport de la constante molaire des gaz R (voir plus bas les constantes chimiques) sur le nombre d'Avogadro N (cf. chapitre de Mécanique Des Milieux Continus).

Il existe d'autres constantes d'ordre pratique qui se déterminent théoriquement et dont la valeur sera utile à tout ingénieur ou physicien qui souhaiterait appliquer dans la pratique certaines des relations qui seront démontrées sur ce site:

2.2. CONSTANTES PHYSIQUES

Définition: Une "**constante physique**" est une quantité physique dont la valeur numérique est fixe. Contrairement à une constante mathématique, elle implique directement et toujours une grandeur physiquement mesurable.

Masse de l'électron (au repos)	$m_e \cong 9.11 \cdot 10^{-31} [kg]$
Masse du neutron (au repos)	$m_n \cong 1.674 \cdot 10^{-27} [kg]$
Masse du proton (au repos)	$m_p \cong 1.672 \cdot 10^{-27} [kg]$
Constante de structure fine	$\alpha \cong 7.297 \cdot 10^{-3}$
Quantum de flux magnétique	$\Phi_0 = 2.067 \cdot 10^{-15} [V \cdot s]$
Constante de Stefan (appelée aussi 1ère constante de Stefan-Boltzmann)	$\sigma \cong 5.67 \cdot 10^{-8} [J \cdot s^{-1} \cdot m^{-2} \cdot K^{-4}]$
Constante de Stefan- Boltzmann (appelée aussi 2ème constante de Stefan-Boltzmann)	$\sigma_{s.b} \cong 7.56 \cdot 10^{-16} [J \cdot m^{-3} \cdot K^{-4}]$
Rayon classique de l'électron	$r_e \cong 2.817 \cdot 10^{-15} [m]$
Impédance du vide	$Z_0 \cong 376 [V \cdot C \cdot s^{-1}]$
Magnéton de Bohr	$\mu_B \cong 927.40 \cdot 10^{-24} [J \cdot T^{-1}]$
Constante de Rydberg	$R_y \cong 10.973 \cdot 10^6 [m^{-1}]$

Rayon de Bohr	$a_0 \cong 0.0529 \cdot 10^{-9} [m]$
Electron-Volt	$1[eV] \cong 1.602 \cdot 10^{-19} [J]$
Accélération gravitationnelle terrestre moyenne	$g \cong 9.81 [m \cdot s^{-2}]$
Pression standard	$P_0 = 101'325 [Pa] = 1[atm]$

Tableau: 28.3 - Constantes physiques

2.3. CONSTANTES PHYSICO-CHIMIQUES

Les constantes physico-chimiques sont des constantes physiques que l'on retrouve plus particulièrement dans l'ensemble des domaines ayant trait à la chimie.

Constante molaire des gaz	$R \cong 8.31 [J \cdot mol^{-1} \cdot K^{-1}]$
Constante de Faraday	$F = 96'458 [C \cdot mol^{-1}]$
Volume molaire	$V_M \cong 22.4141 [l \cdot mol^{-1}]$
Unité de masse atomique	$u.m.a \cong 1.6605 \cdot 10^{-27} [kg]$

Tableau: 28.4 - Constantes physico-chimiques

Remarque: Le lecteur intéressé par les propriétés des éléments chimiques peut télécharger le tableau périodique des éléments proposé dans la rubrique de téléchargement du site.

2.4. CONSTANTES ASTROPHYSIQUES

Le tableau suivant contient les valeurs des constantes et paramètres couramment utilisés en astrophysique et aussi plus particulièrement en cosmologie.

Constante de Hubble	$H(t_0) \cong 75 [km \cdot s^{-1} / Mpc]$
Densité critique de l'Univers	$\rho_c = 5 \cdot 10^{-27} [kg \cdot m^{-3}]$
Distance Terre-Soleil (parsec)	$3.856 \cdot 10^{16} [m]$
Rayon Terrestre	$R_T \cong 6.378 \cdot 10^6 [m]$
Rayon Solaire	$R_\odot \cong 6.96 \cdot 10^8 [m]$
Masse Terrestre	$M_T \cong .5972 \cdot 10^{25} [kg]$
Masse Solaire	$M_\odot \cong 1.989 \cdot 10^{30} [kg]$

Tableau: 28.5 - Constantes astrophysiques

2.5. CONSTANTES DE PLANCK

Les constantes de Planck sont principalement des curiosités physiques qui découlent d'un système d'unités particulier et dont les valeurs selon le système S.I. sont données dans le tableau ci-dessous.

Remarque: Le lecteur intéressé par la provenance des différentes constantes de Planck (longueur de Planck, masse de Planck, etc.) devra se rendre dans le chapitre de Physique Quantique Ondulatoire du site où ces constantes sont déterminées avec les détails nécessaires.

Longueur de Planck	$\lambda_p \cong 2 \cdot 10^{-35} [m]$
Temps de Planck	$t_p = 5.391 \cdot 10^{-44} [s]$
Masse de Planck	$M_p \cong 2.176 \cdot 10^{-8} [Kg]$
Température de Planck	$T_p \cong 1.42 \cdot 10^{32} [K]$
Energie de Planck	$E_p \cong 1.956 \cdot 10^9 [J]$
Densité de Planck	$\rho_p \cong 5.1 \cdot 10^{96} [kg \cdot m^{-3}]$
Force de Planck	$F_p \cong 1.21 \cdot 10^{44} [N]$
Puissance de Planck	$P_p \cong 3.629 \cdot 10^{52} [W]$
Pulsation de Planck	$\omega_p \cong 1.855 \cdot 10^{43} [s^{-1}]$ notée aussi parfois avec l'unité des radians explicitement: $\omega_p \cong 1.855 \cdot 10^{43} [Rad \cdot s^{-1}]$
Charge de Planck	$q_p \cong 1.875 \cdot 10^{-18} [C]$
Courant de Planck	$I_p \cong 3.479 \cdot 10^{25} [A]$
Tension de Planck	$U_p \cong 1.0432 \cdot 10^{27} [V]$
Impédance de Planck	$I_p \cong 29.98 [\Omega]$

Tableau: 28.6 - Constantes de Planck

Malgré les exemples donnés combien y-a-t'il de constantes ? Pourquoi jouent-elles un "rôle central" dans les théories physiques ? Ont-elles toutes la même importance ou certaines sont-elles plus fondamentales ? Selon quels critères ? Peut-on alors tester si elles sont vraiment constantes ?

Pour essayer de répondre à certaines de ces questions, remarquons tout d'abord qu'à chaque étape de nos constructions théoriques il subsiste des paramètres constants qui ne sont pas et ne peuvent pas être expliqués en termes de quantités plus fondamentales, simplement parce que ces dernières n'existent pas

dans l'état de nos connaissances. Quand les théories s'affinent, il est en effet possible qu'une constante se trouve expliquée en termes de nouveaux paramètres, plus fondamentaux. Ainsi, la masse du proton est une constante fondamentale de la physique nucléaire, mais doit en principe pouvoir se calculer, dans le cadre de la chromodynamique quantique, en fonction de la masse des quarks et des énergies des liaisons électromagnétique et forte. Ce changement de statut est associé à celui du proton, qui de particule élémentaire devient corps composite.

Nous définirons modestement les constantes comme tous les paramètres non déterminés dans un cadre théorique donné. Cette définition revient à accepter que nous soyons incapables d'écrire une équation d'évolution pour ces constantes qui se révèlent donc comme la limite de ce que les théories où elles apparaissent sont en mesure d'expliquer. Cependant l'hypothèse de constance est implicitement contrôlée par la validation expérimentale de ces théories. Les résultats des expériences doivent être reproductibles à divers moments et dans divers laboratoires. Si c'est le cas, dans la limite des précisions expérimentales, alors il est légitime de considérer que l'hypothèse de constance est valide. Cette définition implique qu'il n'existe pas de liste absolue de constantes, car l'appartenance à une telle liste dépend des cadres théoriques choisis pour décrire la nature et peut donc changer avec les progrès de la connaissance.

Se pose maintenant la question de savoir s'il est possible de caractériser plus précisément le concept de constante et de déterminer si, parmi toutes les constantes, certaines sont plus fondamentales que d'autres. Pour cela, il faut commencer par révéler une relation entre constantes et unités.

Ainsi, Planck découvre en 1900 qu'il était possible d'utiliser les trois constantes physiques fondamentales:

$$\begin{aligned} G &\cong 6.67259 \cdot 10^{-11} [m^3 \cdot s^2 \cdot Kg^{-1}] \\ c &\cong 299'792'458 [m \cdot s^{-1}] \\ \hbar = h/2\pi &\cong 1.05 \cdot 10^{-34} [J \cdot s] \end{aligned}$$

(28.32)

pour définir les trois unités de masse, de temps et de longueur à partir de la masse de Planck, de la longueur de Planck et du temps de Planck (voir le chapitre de Physique Quantique Ondulatoire pour la démarche mathématique qui permet de déterminer leurs valeurs).

Planck baptise ces unités "**Système d'Unités Naturelles**" (SUN) car elles sont indépendantes d'un corps ou d'un matériau et [...] gardent nécessairement leur signification pour tous les temps et toutes les civilisations, même celles qui sont extra-terrestres et non humaines. La signification de ces unités met longtemps à émerger. Elles signalent l'échelle où gravitation et mécanique quantique deviennent de même amplitude. Elles sont donc très adaptées à la cosmologie primordiale et à l'étude des trous noirs ainsi que la mécanique quantique relativiste.

Le choix des unités de Planck comme unités naturelles est lié aux considérations justifiant que G , c , h sont les trois constantes dimensionnées les plus fondamentales (connues à ce jour). Dans ces unités, la valeur numérique de ces trois constantes fondamentales est 1 comme nous l'avons déjà fait déjà remarquer.

Le rôle des constantes dans la structuration des théories physiques peut être assez magnifiquement illustré par le cube magique ou "**cube de Okun**" des théories physiques ci-dessous (dont la validité reste

à vérifier bien sûr). L'idée consiste à "allumer" ou "éteindre" une à une les trois constantes fondamentales afin de voir comment les théories physiques s'articulent les unes par rapport aux autres.

Remarque: Le lecteur comprendra mieux les explications qui vont suivre lorsqu'il aura étudié la relativité générale, la physique quantique des champs ainsi la physique quantique ondulatoire donc si jamais il peut sauter ce texte en attendant.

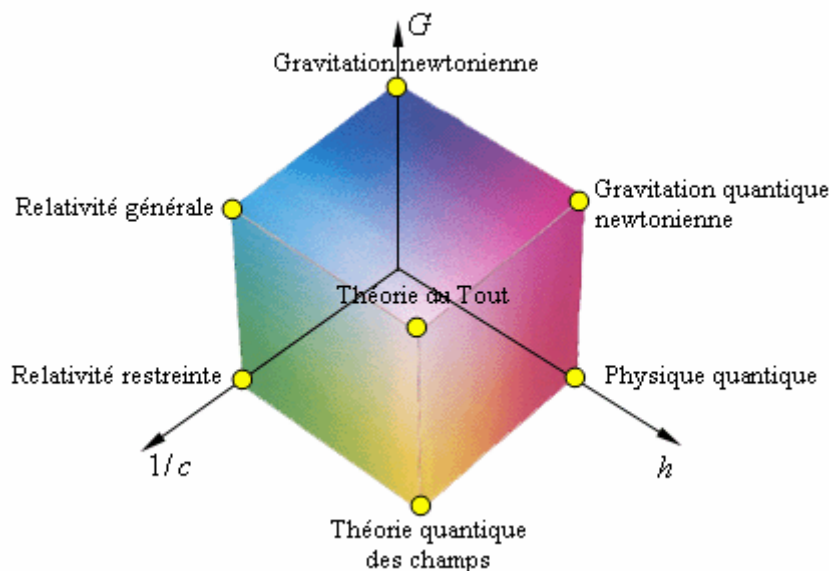


Figure: 28.1 - Cube de Okun

Quand G est mis à 0, cela revient à supprimer toutes les forces gravitationnelles et à découpler la matière de l'espace et du temps. Quand h est mis à 0, nous supprimons le caractère quantique de la nature et nous découplons les natures corpusculaires et ondulatoire (de par la relation de De Broglie), quand $1/c$ est mis à 0, la vitesse de la lumière est infinie et le temps et l'espace se découplent l'un de l'autre (de par les transformations de Lorentz). Pour visualiser cela, nous considérons le cube ci-dessus introduit par le physicien soviétique Mikhaïl Bronshtein qui reprend une idée développée initialement par Lev Landau, Dimitri Ivanenko et Georgi Gamow.

Tout naturellement, au niveau le plus bas, nous trouvons (0,0,0) la mécanique newtonienne, qui ne prend pas en compte les effets relativistes, quantiques et gravitationnels. Au niveau supérieur où nous considérons l'effet d'une constante, nous trouvons les trois théories de la relativité restreinte (1,0,0), de la mécanique quantique en (0,1,0) et de la gravitation newtonienne en (0,0,1), trois théories testées avec une grande précision dans leur domaine de validité.

A un niveau encore supérieur, la théorie quantique des champs en (1,1,0) prend en compte les effets quantiques et relativistes; la relativité générale en (1,0,1) prend en compte les effets gravitationnels et relativistes et la gravité quantique newtonienne en (0,1,1) est censée offrir une description quantique et non relativiste de la gravitation. Seules les deux premières théories sont actuellement fondées expérimentalement et théoriquement.

Au niveau ultime se trouve en (1,1,1) la théorie du Tout (dénomination très prétentieuse et trop commerciale), censée donner une description quantique et relativiste de la gravitation. Sa formulation n'est pas encore connue, bien que les théories des cordes (voir chapitre du même nom), intensivement étudiées de nos jours, semblent des candidats sérieux. Ces théories apparaissent comme des cas limites

d'une théorie plus large et plus profonde mais non encore formulée: la théorie M (le M pour "Mère")

3. PRINCIPES DE LA PHYSIQUE

Les progrès de la science en général et de la physique en particulier étaient fondés il y deux siècles principalement sur l'expérimentation, c'est-à-dire que l'on reproduisait en laboratoire des phénomènes donnés pour les analyser systématiquement (la reproduction d'une observation validant une hypothèse). Cela revenait systématiquement à poser des questions précises à la nature et à décrire et étudier les réactions ainsi provoquées. La répétition à volonté d'un phénomène lors d'une expérience ne serait pas garantie s'il n'existait pas un principe général de causalité.

3.1. PRINCIPE DE CAUSALITÉ

Définition: Nous définissons le "**principe de causalité**" par le fait que dans exactement les mêmes conditions, les mêmes causes conduisent toujours aux mêmes effets. Autrement dit, si certaines conditions initiales sont parfaitement connues, le phénomène se déroulera de façon déterminée, toujours la même.

Au fait, l'expérience n'est pas nécessaire si nous considérons les principes de premier ordre qui sont par définition "les principes logiques que nous pouvons déduire par induction et que nous ne pouvons vérifier expérimentalement avec certitude".

Or, les exigences de la société ont très peu souvent laissé le temps aux grands hommes de science de penser à ces principes du premier ordre par des expériences imaginaires (méthode très usitée par Albert Einstein pour la parenthèse...).

C'est dans un trilemne proposé par le sceptique de l'antiquité Agrippa, selon un argument rapporté par Sextus Empiricus, que la question de la justification de la connaissance a été posée le plus explicitement:

H1. Ou bien la connaissance est fondée en dernière instance sur des principes premiers mais arbitraires

H2. Ou bien nous ne trouvons pas de tels principes et nous avons une régression à l'infini

H3. Ou bien la justification est circulaire

Ce trilemne porte aussi souvent, dans la philosophie contemporaine, notamment chez Karl Popper, le nom de "**trilemne de Fries**" ou "**trilemne de Münchhausen**" et nous ne savons actuellement pas dans quel cas de figure (H1, H2 ou H3) nous nous situons.

Énonçons maintenant trois principes (ou hypothèses) premiers élémentaires:

3.2. PRINCIPE DE CONSERVATION DE L'ÉNERGIE

Le principe premier de conservation de l'énergie s'énonce (basiquement... voir remarques plus bas...) ainsi: L'énergie totale, notée de tout système isolé et inertiel ne varie pas en fonction du temps s'il n'y a pas apport ou retrait d'énergie (ou de masse) ou de chaleur de l'extérieur de ce système.

Ce principe peut être exprimé par la formule :

$$\Delta E_{tot} = \Delta U + \Delta E_c + \Delta E_p$$

où ΔE_{tot} est la variation totale d'énergie du système, ΔU la variation de l'énergie interne du système. C'est à dire son énergie propre correspondant aux énergies cinétiques et potentielles microscopiques, des particules qui le constituent.

ΔE_c est la variation de l'énergie cinétique à l'échelle macroscopique (mouvement du système dans un référentiel donné) et ΔE_p est la variation de l'énergie potentielle à l'échelle macroscopique, du système en interaction avec des champs gravitationnels ou électromagnétiques.

En physique, une loi de conservation (rien ne se perd, rien ne se crée) exprime qu'une propriété mesurable particulière d'un système physique isolé reste constante au cours de l'évolution de ce système. La liste ci-dessous énumère des lois de conservations utiles à l'ingénieur et qui n'ont jamais été prises en défaut à ce jour et qui découlent pour la plupart de la conservation de l'énergie:

- conservation de la quantité de mouvement
- conservation du moment cinétique
- conservation de la charge électrique
- conservation du flux magnétique
- conservation de la masse

Le théorème de Noether que nous verrons un peu plus bas exprime l'équivalence qui existe entre les lois de conservation et l'invariance des lois physiques en ce qui concerne certaines transformations (typiquement appelées symétries). Ce théorème ne s'applique qu'aux systèmes descriptibles par un lagrangien (cf. [chapitre de Mécanique Analytique](#)). Par exemple, l'invariance par translation dans le temps implique que l'énergie est conservée, l'invariance par translation dans l'espace implique que la quantité de mouvement est conservée, et l'invariance par rotation dans l'espace implique que le moment angulaire est conservé.

Cette équivalence est démontrable et découle de l'invariance dans le temps des lois de la physique. Il s'agit du premier principe (théorème) de Noether que nous allons démontrer un peu plus loin.

L'énergie que l'être humain quantifie avec l'unité "**Joules**" ne peut cependant être définie avec exactitude. Répondre à cette question revient à savoir ce qu'est la masse (relation d'équivalence d'Einstein) et donc à connaître l'élément fondamental de l'Univers (nous en avons déjà fait mention plus haut dans le présent texte).

Remarques:

R1. L'énergie peut se trouver sous plusieurs formes (cela ne voulant pas dire qu'il existe plusieurs énergies différentes !!!) comme la chaleur, l'énergie cinétique, potentielle, électrique, magnétique, etc... comme nous en avons déjà fait mention plus haut.

Ainsi, dans les applications grand public, et notamment dans le domaine de la nutrition, nous exprimons fréquemment l'énergie en calories. La calorie étant en toute rigueur l'énergie qu'il faut fournir pour faire chauffer un gramme d'eau de un degré Celsius (cf. chapitre de Thermodynamique), mais les nutritionnistes nomment par simplification "**calorie**" ce que les physiciens nomment (correctement) "**kilocalorie**".

En électricité, nous utilisons aussi le "**Watt-heure**", énergie consommée pendant une heure par un appareil ayant une puissance d'un Watt (joules par secondes).

R2. La violation de ce principe de conservation de l'énergie dans un système isolé n'a encore jamais été observée mais sa validité ne peut être démontrée (d'où le fait que ce soit un "principe premier").

R3. Certains physiciens débattent du fait que ce principe premier découle du théorème de Noether que nous verrons plus loin. Mais cela est tout à fait discutable étant donné que le théorème de Noether considère l'énergie potentielle comme constante dans le temps d'où...

3.3. PRINCIPE DE MOINDRE ACTION

Le principe premier de moindre action (dit également "**principe premier d'économie**" ou "**principe variationnel**") s'énonce ainsi:

Tous les phénomènes naturels s'accordent avec le fait que, la Nature, dans la production de ses effets, agit toujours par les voies les plus simples et les plus directes.

Avec cet énoncé et le principe de conservation de l'énergie, nous pouvons alors établir des outils mathématiques d'une formidable puissance pour l'étude de la physique théorique. Mais nous ne pouvons développer à ce niveau du discours le formalisme mathématique de ce principe car il demande des outils que nous souhaiterions introduire plus loin, dans le chapitre de Mécanique Analytique (lors de l'étude du formalisme lagrangien pour être plus précis).

En attendant voici les deux relations qui le résument:

$$S(q, \dot{q}) = \int_{t_i}^{t_f} L(q, \dot{q}) dt$$

$$\frac{\partial L}{\partial q_i} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} = 0$$

(28.33)

Remarque: La violation de ce principe dans un système isolé n'a encore jamais été observée mais sa validité ne peut être démontrée (d'où le fait que ce soit aussi un "principe premier").

3.4. PRINCIPE DE NOETHER

Le principe premier de Noether (appelé traditionnellement "**théorème de Noether**") associe de façon élégante des quantités physiques conservées aux symétries des lois de la nature. La symétrie de translation dans le temps (phénomène invariant dans le temps) correspond à la conservation de l'énergie, celle de translation dans l'espace à la conservation de l'impulsion (quantité de mouvement), celle de rotation dans l'espace à la conservation du moment cinétique etc.

En d'autres termes, le principe premier de Noether énonce que la physique est:

- Symétrique (invariante) par translation dans le temps (ceci ayant pour conséquence qu'il n'y pas d'origine du temps)
- Symétrique (invariante) par translation dans l'espace (ceci ayant pour conséquence qu'il n'y a pas d'origine à l'espace)
- Symétrique (invariante) par rotation (ceci ayant pour conséquence qu'il n'y a pas de direction privilégiée dans l'espace)

Remarques:

R1. Ce principe implique donc qu'un référentiel galiléen (cf. chapitre de Mécanique Classique) est homogène (pas d'origine de temps ou d'espace privilégiée) et isotrope (pas de direction privilégiée).

R2. Il ne faut pas confondre l'invariance des lois et la non invariance des solutions théoriques auxquelles aboutissent ces lois! Par exemple, la décharge d'un condensateur (cf. chapitre d'Électrocinétique) est invariante par translation dans le temps mais pas la solution.

Ce résultat établi en 1915 par Emmy Noether juste après son arrivée à Göttingen, aurait été qualifié par Albert Einstein de "monument de la pensée mathématique". C'est maintenant un des piliers de la physique théorique.

Aujourd'hui, il est souvent présenté à l'occasion de cours sur la théorie quantique des champs. Cela le fait paraître plus compliqué et mystérieux qu'il n'est, et c'est oublier qu'il s'applique aussi à la mécanique classique.

Remarque: Il est recommandé au lecteur de lire la démonstration du théorème de Noether en parallèle des chapitres de mécanique analytique et de mécanique classique.

Ainsi, les symétries jouent un rôle majeur en physique. Elles permettent d'une part de simplifier les problèmes d'une part et de tirer de nouvelles lois d'autre part. Pour illustrer la première application des symétries il suffit d'évoquer la forme mathématique du potentiel gravitationnel engendré par une masse ponctuelle située à l'origine du référentiel (cf. chapitre de Mécanique Classique). En coordonnées cartésiennes, l'expression du potentiel gravitationnel est relativement complexe

$$\phi = -\frac{G}{\sqrt{x^2 + y^2 + z^2}} \quad (28.34)$$

alors qu'en coordonnées sphériques (système de coordonnées qui tire parti de la symétrie sphérique du potentiel) il prend une forme très simple :

$$\phi = -\frac{G}{r} \quad (28.35)$$

Les propriétés de symétrie d'un problème sont ici exploitées de façon à simplifier le traitement mathématique des lois physiques. Même si ces considérations mathématiques nous renseignent sur les propriétés physiques du système considéré, elles conservent cependant un caractère purement technique.

Les symétries trouvent pourtant une autre application dont la signification physique est beaucoup plus profonde. Le fait, non fortuit, qu'un système possède des symétries doit certainement avoir des implications physiques. Intuitivement, nous pouvons saisir que la présence de symétries dans un système physique se traduit par l'invariance de certaines de ses propriétés physiques sous l'application de transformations spatio-temporelles ou, plus généralement, des transformations géométriques. L'invariance de propriétés physiques doit induire nécessairement des relations d'une nature nouvelle entre les variables du système. De telles relations doivent à leur tour révéler des lois plus profondes qui associent la géométrie du système aux lois de la nature. Ce raisonnement, bien qu'intuitif, nous invite à explorer plus en profondeur les relations qui pourraient exister entre les lois physiques et les propriétés géométriques de l'espace-temps.

Considérons une expérience de mécanique plus ou moins complexe observée simultanément par deux physiciens O et O' situés en des lieux différents tel que chacun d'eux choisit un référentiel dont il est l'origine.

Ils entreprennent de mesurer diverses grandeurs physiques et obtiennent des résultats numériques qui dans l'ensemble diffèrent. Cependant, les lois physiques qu'ils en tirent (à niveau de connaissance égal) sont identiques. Cette conclusion est triviale car nous savons tous que les lois de la nature ne doivent pas dépendre pas de l'emplacement des observateurs.

Mathématiquement, la différence entre les référentiels de O et O' selon le référentiel de l'expérience étudiée est le passage de l'un à l'autre dans un plan par une rotation $R(\phi)$ et/ou une translation $T(l)$ (cf. chapitre de Géométrie Euclidienne).

Le fait que les lois physiques sont indépendantes de la position de l'observateur implique qu'elles ne varient pas après leur avoir appliqué une rotation et/ou une translation. Nous disons alors qu'elles sont "**invariantes par rotation et par translation**" ou encore qu'elles sont "**symétriques par rotation et par translation**".

Rappel: En mathématiques le terme "symétrie" prend un sens plus général qui peut se définir comme suit: "transformation qui ne change ni la forme, ni les dimensions d'une figure". Nous pouvons remarquer que le sens courant du mot "symétrie" correspond à un cas particulier de symétrie au sens géométrique du terme, qui consiste à inverser les objets par rapport à un plan.

Remarque: En physique, la définition d'une symétrie est semblable à celle des mathématiciens mais s'applique aux lois de la nature et non plus aux figures géométriques. Ainsi une symétrie en physique est une transformation des variables du système - qui peuvent être des variables géométriques ou des variables plus abstraites - qui ne change pas la formulation des lois physiques.

Donnons une définition rigoureuse d'une symétrie en physique:

Soit un système S dont l'état évolue au cours du temps. Désignons l'état de S à l'instant t par $S(t)$. A l'instant initial t_0 , S se trouve donc dans l'état $S(t_0)$. Considérons une transformation géométrique T qui agit en chaque point de l'espace et éventuellement du temps. En un instant t , l'action de T sur le système S a pour effet de le transformer en un système $S'(t) = T[S(t)]$ tel qu'à l'instant t_0 , le transformé par T de $S(t_0)$ est $S'(t_0)$.

Définition: La transformation T est appelée une "**symétrie physique**" si la transformée par T du système S (ce qui donne S') évolue de la même façon que S , c'est-à-dire que si nous appliquons les lois de la mécanique sur $S'(t_0)$ pour connaître son état S'' en un instant postérieur t alors $S''(t) = S'(t) = T[S(t)]$.

3.4.1. INVARIANCE PAR TRANSLATION DANS L'ESPACE

Considérons un système isolé constitué de n particules en interaction repérées par les vecteurs position $\vec{r}_i$. L'interaction de deux particules i, j dérive d'un potentiel $\varphi(\vec{r}_i, \vec{r}_j, t)$ (cf. chapitre de Mécanique Classique). Chaque particule est soumise à des forces résultant de l'interaction avec les autres particules. Pour une particule i donnée, la résultante de ces forces s'exprime selon la loi de Newton (voir chapitre de mécanique classique) :

$$\vec{F}_i(\vec{r}, t) = m\vec{\ddot{r}}_i(\vec{r}_i, t) = \sum \vec{F}_{i,j}(\vec{r}_i, \vec{r}_j, t) = \sum -m\vec{\nabla} \varphi(\vec{r}_i, \vec{r}_j, t) \quad (28.36)$$

Appliquons au système la translation dans l'espace suivante :

$$\vec{r}_i \rightarrow \vec{r}'_i = \vec{r}_i + \vec{l} \quad (28.37)$$

où $\vec{l}$ est un vecteur quelconque. Dire que la translation du système est une symétrie signifie que l'accélération et la force qui agit sur chaque particule sont inchangées après la transformation.

$$\vec{F}_i(\vec{r}, t) = \vec{F}_i(\vec{r}_i + \vec{l}, t) \quad (28.38)$$

Ce qui implique :

$$\sum -m\vec{\nabla} \varphi(\vec{r}_i + \vec{l}, \vec{r}_j + \vec{l}, t) = \sum -m\vec{\nabla} \varphi(\vec{r}_i, \vec{r}_j, t) \quad (28.39)$$

Cette égalité doit être vraie quelle que soit la position des particules, donc quels que soient $\vec{r}_i$ et $\vec{r}_j$. Il est clair que la seule manière de vérifier la dernière égalité dans ces conditions est d'égaliser deux à deux les potentiels entre chaque particule j avec la particule i , c'est-à-dire :

$$\varphi(\vec{r}_i + \vec{l}, \vec{r}_j + \vec{l}, t) = \varphi(\vec{r}_i, \vec{r}_j, t) \quad (28.40)$$

Les potentiels sont alors nécessairement (et c'est là, la puissance du théorème de Noether!) des fonctions de $\vec{r}_i - \vec{r}_j$ telles que:

$$\varphi(\vec{r}_i, \vec{r}_j, t) = \varphi(\vec{r}_i - \vec{r}_j, t) \quad (28.41)$$

Dès lors, nous en déduisons que:

$$\vec{F}_{i,j}(\vec{r}_i, \vec{r}_j, t) = -m \vec{\nabla} \cdot \varphi(\vec{r}_i - \vec{r}_j, t) = -(-m \vec{\nabla} \cdot \varphi(\vec{r}_j - \vec{r}_i, t)) = -\vec{F}_{j,i}(\vec{r}_j, \vec{r}_i, t) \quad (28.42)$$

Ce qui entraîne immédiatement que la résultante de toutes les forces appliquées aux particules du système est nulle et que donc la quantité de mouvement totale est conservée:

$$\sum \vec{F}_{i,j}(\vec{r}_i, \vec{r}_j, t) + \sum \vec{F}_{j,i}(\vec{r}_j, \vec{r}_i, t) = \sum \vec{F}_i = 0 \rightarrow \frac{d}{dt} \left(\sum m \dot{\vec{r}}_i \right) = 0 \Leftrightarrow \sum m \dot{\vec{r}}_i = \vec{c}^* \quad (28.43)$$

L'invariance par translation de la loi de Newton entraîne donc nécessairement :

1. Le potentiel entre les particules d'un système isolé est une fonction de leur distance relative (cela se confirmera en astronomie lors de notre étude du champ de potentiel gravitationnel, ainsi qu'en électromagnétisme en ce qui concerne le potentiel électrostatique et les potentiels de Yukawa à symétrie sphérique en théorie quantique des champs).
2. La loi de l'égalité entre l'action et la réaction.
3. La conservation de la quantité de mouvement totale d'un système !

Conséquence du point (3): l'origine de l'espace est inobservable (puisque la conservation de la quantité de mouvement est équivalente à l'invariance par translation dans l'espace)!

3.4.2. INVARIANCE PAR ROTATION DANS L'ESPACE

Imposons maintenant que les rotations autour d'un point fixe soient des symétries. Cette propriété doit être vraie quel que soit le point fixe considéré, notamment, si ce point fixe est précisément la position de l'une des particules du système. Il s'ensuit que le potentiel présente nécessairement une symétrie sphérique. Les forces agissant entre les particules sont donc colinéaires aux vecteurs qui les relient.

Le moment cinétique du système s'exprime comme suit (cf. chapitre de Mécanique Classique) :

$$\vec{b} = \sum \vec{b}_i = \sum m_i \vec{r}_i \times \dot{\vec{r}}_i \quad (28.44)$$

La dérivée par rapport au temps du moment cinétique total donne :

$$\frac{d\vec{b}}{dt} = \sum m_i \frac{d}{dt} (\vec{r}_i \times \dot{\vec{r}}_i) = \sum m_i \frac{d\vec{r}_i}{dt} \times \dot{\vec{r}}_i + \sum m_i \vec{r}_i \times \ddot{\vec{r}}_i = \sum m_i \vec{r}_i \times \ddot{\vec{r}}_i = \sum \vec{r}_i \times m_i \ddot{\vec{r}}_i \quad (28.45)$$

Or le dernier terme du produit vectoriel peut s'écrire :

$$m_i \ddot{\vec{r}}_i = \vec{F}_{ij} \quad (28.46)$$

où les $\vec{F}_{ij}$ sont les forces internes au système des particules j agissant sur la particule i . L'avant-dernière expression devient alors :

$$\frac{d\vec{b}}{dt} = \sum \vec{r}_i \times \vec{F}_{ij} \quad (28.47)$$

Nous pouvons regrouper les termes $\vec{r}_j \times \vec{F}_{ji}$ et $\vec{r}_i \times \vec{F}_{ij}$ deux à deux et de par la propriété du produit vectoriel nous avons nécessairement:

$$\frac{d\vec{b}}{dt} = \sum \vec{r}_i \times \vec{F}_{ij} = \vec{0} \quad (28.48)$$

Donc nous en concluons que le moment cinétique est conservé et la conservation du moment cinétique est donc équivalente à l'invariance par rotation.

Conséquence: il n'y a pas de direction privilégiée dans l'espace!

3.4.3. INVARIANCE PAR TRANSLATION DANS LE TEMPS

L'énergie totale du système est la somme de l'énergie cinétique de toutes les particules et de l'énergie potentielle résultant de l'interaction mutuelle des particules, soit sous la forme de la mécanique classique :

$$E = E_c + E_p = \sum \frac{1}{2} m_i \dot{\vec{r}}_i^2 + V_{i,j}(\vec{r}_i, \vec{r}_j) \quad (28.49)$$

Nous supposons que le potentiel $V_{i,j}(\vec{r}_i, \vec{r}_j)$ ne varie pas avec le temps. Cette hypothèse se justifie de manière empirique par le constat que les potentiels observés dans la nature sont indépendants du temps dans des systèmes fermés à l'équilibre.

Calculons la dérivée de l'énergie par rapport au temps :

$$\begin{aligned} \frac{dE}{dt} &= \frac{d}{dt} \sum \frac{1}{2} m_i \dot{\vec{r}}_i^2 + \frac{d}{dt} V_{i,j}(\vec{r}_i, \vec{r}_j) = \frac{1}{2} \frac{d}{dt} \sum m_i \dot{\vec{r}}_i^2 + 0 = \frac{1}{2} \sum \frac{dm_i}{dt} \dot{\vec{r}}_i^2 + \frac{1}{2} \sum \frac{d\dot{\vec{r}}_i^2}{dt} m_i \\ &= \frac{1}{2} \sum \frac{dm_i}{dt} \dot{\vec{r}}_i^2 + \sum \dot{\vec{r}}_i \cdot \frac{d\dot{\vec{r}}_i}{dt} m_i = \frac{1}{2} \sum \frac{dm_i}{dt} \vec{v}_i^2 + \sum \vec{v}_i \cdot \frac{d\vec{v}_i}{dt} m_i \end{aligned} \quad (28.50)$$

Or, si le système est fermé (pas d'apport de masse de l'extérieur ni apport d'énergie de l'extérieur), le terme dm_i/dt est nul (pas de variation relativiste de la masse non plus car la vitesse de chaque corpuscule ou du système entier est constante ou sa variation est en moyenne nulle). Il en est de même pour le terme $d\dot{\vec{r}}_i/dt$ où si le système est fermé (pas d'apport d'énergie de l'extérieur sous quelle que forme que ce soit) l'accélération moyenne de chaque corpuscule ou de l'ensemble du système par rapport au centre de gravité sera nulle. Donc:

$$\frac{dE}{dt} = \frac{1}{2} \sum \frac{dm_i}{dt} \vec{v}_i^2 + \sum \vec{v}_i \cdot \frac{d\vec{v}_i}{dt} m_i = \frac{1}{2} \sum 0 \cdot \vec{v}_i^2 + \sum \vec{v}_i \cdot \vec{0} \cdot m_i = 0 \quad (28.51)$$

Donc nous en concluons que l'énergie totale du système est une constante!

Quelle est la grandeur mécanique invariante par translation revient donc à se demander quelles sont les grandeurs mathématiques qui sont inchangées lorsque nous leur appliquons une translation. Il en existe deux : les scalaires et les vecteurs.

Intuitivement, un scalaire est assimilé à un nombre réel (cf. [chapitre sur les Nombres](#)). Or, en mécanique, les nombres réels que nous pouvons construire le sont à l'aide de grandeurs vectorielles comme le vecteur position, la vitesse, etc. Pour qu'un tel nombre réel ait le statut de scalaire il doit être indépendant de l'espace. Ainsi, un vecteur position ne peut manifestement être considéré comme un scalaire. L'énergie de la particule est un nombre réel mais n'est pas un scalaire car elle dépend explicitement, dans sa formulation, de la position de la particule dans l'espace au travers de l'énergie potentielle.

De la même façon, un vecteur n'est pas seulement un être mathématique possédant des composantes dans une base. Pour jouir du statut de vecteur, une entité mathématique doit se transformer de la même manière que les vecteurs de base de l'espace vectoriel. Selon cette définition, le moment cinétique n'est pas un vrai vecteur car, étant la composition par produit vectoriel de deux vecteurs, il ne se transforme pas comme les vecteurs de base. D'un point de vue mathématique il s'agit d'un pseudo-vecteur (cf. [chapitre de Calcul Vectoriel](#)).

Le seul vrai vecteur qui reste est la quantité de mouvement car il est construit à l'aide de la dérivée du vecteur position qui est, bien évidemment, un vrai vecteur. Nous en déduisons que la seule grandeur susceptible d'être conservée par translation est la quantité de mouvement totale du système.

Par un raisonnement analogue au précédent, il est possible de supposer quelle grandeur pourrait être invariante par rotation. Sachant que seuls les scalaires et certains pseudo-vecteurs sont effectivement invariants par rotation, nous en concluons que la seule grandeur susceptible d'être conservée lors de rotations est le moment cinétique total du système.

Enfin, toujours par le même raisonnement, l'invariance des lois de la mécanique par déplacement dans le temps, revient à rechercher les grandeurs conservées par une translation dans le temps. Ces grandeurs sont les vrais scalaires et les vecteurs sur la droite du temps. Aucune grandeur mécanique ne peut être assimilée à un vecteur sur la droite du temps. En revanche, l'énergie est bien un scalaire invariant par translation dans le temps puisque l'énergie potentielle est par hypothèse indépendante du temps. L'invariance des lois de la mécanique par déplacement dans le temps laisse donc supposer intuitivement la conservation de l'énergie.

Ces raisonnements ne peuvent évidemment faire office de démonstration mais ils mettent en évidence une relation étroite entre la géométrie et les propriétés d'invariance d'un système.

3.4.4. THÉORÈME DE NOETHER

Soit L le lagrangien (cf. [chapitre de Mécanique Analytique](#)) d'un système représenté par les $2n$ coordonnées généralisées $q_i, \dot{q}_i$ dans l'espace de configuration. Supposons que ce système soit invariant par la transformation infinitésimale h_s suivante :

$$q_i \rightarrow q'_i = h_s(q) \quad (28.52)$$

Où s est un paramètre réel et continu et pour lequel nous avons:

$$h_{s=0}(q) = q \quad (28.53)$$

La fonction $h_s(q)$ agit continûment sur le chemin variationnel selon la démarche intellectuelle qui sera énoncée dans le chapitre de Mécanique Analytique.

Supposons que les fonctions $q_i, \dot{q}_i$ sont solutions des équations de Lagrange (ce que nous démontrerons dans le chapitre de Mécanique Analytique). D'après nos hypothèses les fonctions (définies):

$$\varphi(s, t) = h_s, \quad \dot{\varphi}(s, t) = \frac{d}{dt} h_s \quad (28.54)$$

sont dès lors nécessairement également solutions des équations de Lagrange, ce qui se traduit par (nous omettrons l'indication de la somme par la suite afin d'alléger la lecture!):

$$\sum \left(\frac{\partial L(\varphi(s, t), \dot{\varphi}(s, t))}{\partial q_i} - \frac{d}{dt} \frac{\partial L(\varphi(s, t), \dot{\varphi}(s, t))}{\partial \dot{q}_i} \right) = 0 \quad (28.55)$$

D'autre part, par hypothèse, le lagrangien est invariant pour les transformations du type de celles décrites par $q_i \rightarrow q'_i = h_s(q)$. Il s'ensuit que sa dérivée par rapport au paramètre s est nécessairement nulle:

$$\frac{d}{ds} L(\varphi(s, t), \dot{\varphi}(s, t)) = 0 \quad (28.56)$$

Et nous démontrerons par ailleurs aussi en Mécanique Analytique (sous forme d'intégrale comme étant nulle) la relation:

$$\frac{\partial L}{\partial q_i} \delta q_i + \frac{\partial L}{\partial \dot{q}_i} \delta \dot{q}_i = 0 \quad (28.57)$$

ce qui peut finalement s'écrire:

$$\frac{\partial L}{\partial q_i} \frac{d\varphi_i}{ds} + \frac{\partial L}{\partial \dot{q}_i} \frac{d\dot{\varphi}_i}{ds} = 0 \quad (28.58)$$

mais nous avons aussi de par l'équation d'Euler-Lagrange (cf. chapitre de Mécanique Analytique):

$$\frac{\partial L}{\partial q_i} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} = 0 \Leftrightarrow \frac{\partial L}{\partial q_i} = \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} \quad (28.59)$$

Nous obtenons alors :

$$\left(\frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} \right) \frac{d\varphi_i}{ds} + \frac{\partial L}{\partial \dot{q}_i} \frac{d}{dt} \left(\frac{d\varphi_i}{ds} \right) = \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_i} \frac{d\varphi_i}{ds} \right) = \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_i} \frac{dh_s}{ds} \right) = 0 \quad (28.60)$$

Donc la grandeur $\frac{\partial L}{\partial \dot{q}_i} \frac{dh_s}{ds}$ est bien une constante du mouvement !

Le théorème de Noether s'énonce alors ainsi :

Soit un système ayant un lagrangien $L(q_i, \dot{q}_i, t)$ auquel nous appliquons une transformation infinitésimale $q_i \rightarrow q'_i = h_s(q)$, où s est un paramètre réel et continu. Alors il existe une constante du mouvement notée $C(q'_i, \dot{q}'_i, t)$ dont l'expression est donnée par :

$$C(q'_i, \dot{q}'_i, t) = \left(\frac{\partial L}{\partial \dot{q}'_i} \frac{dh_s}{ds} \right)_{s \rightarrow 0} \quad (28.61)$$

Appliquons le théorème de Noether aux cas étudiés précédemment. Fixons un référentiel arbitraire O cartésien. Notons $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ la base orthonormée de ce référentiel. Considérons un système constitué de n particules repérées dans O par leurs vecteurs position $\vec{r}_i$. Le lagrangien de ce système est alors $L(r_{j,i}, \dot{r}_{j,i}, t)$, où $j = x, y, z$ distingue les composantes spatiales des vecteurs $\vec{r}_i, \dot{\vec{r}}_i$.

Supposons maintenant que le système soit invariant par translation de longueur s le long de l'axe x uniquement. La translation le long de cet axe s'écrit comme suit :

$$r_{x,i}(t) \rightarrow h_s(r_{x,i}(t)) = r_{x,i}(t) + s(\vec{e}_x \circ \vec{e}_j) \quad (28.62)$$

et il s'agit donc d'un scalaire.

La constante du mouvement donnée par application du théorème de Noether est alors (toujours sur l'axe x) :

$$C(r'_{x,i}, \dot{r}'_{x,i}, t) = \left(\frac{\partial L}{\partial \dot{r}_{x,i}} \frac{dh_s}{ds} \right)_{s \rightarrow 0} = \frac{\partial L}{\partial \dot{r}_{x,i}} (\vec{e}_x \circ \vec{e}_j) = \frac{\partial L}{\partial \dot{r}_{x,i}} \quad (28.63)$$

Nous définirons par ailleurs en mécanique analytique $\frac{\partial L}{\partial \dot{r}_{x,i}}$ comme étant le moment conjugué $p_{x,i}$.

Nous en déduisons dès lors que la quantité conservée est : $p_{x,i}$, c'est-à-dire la quantité de mouvement totale du système le long de l'axe x !!!

En procédant de même avec les autres axes, nous démontrerions aisément la conservation de la quantité de mouvement totale le long des axes pour ceux-ci, ce qui nous permet de conclure que dans le cas général d'une translation infinitésimale:

$$d\vec{r} = s(\vec{e}_x + \vec{e}_y + \vec{e}_z) \quad (28.64)$$

la grandeur conservée est la quantité de mouvement totale du système.

Supposons maintenant que le système soit invariant par rotation d'un angle infinitésimal s autour de l'axe z . Cette rotation s'écrit :

$$\vec{r}_i(t) \rightarrow h_s(\vec{r}_i(t)) = \vec{r}_i(t) + s(\vec{e}_z \times \vec{r}_i(t)) \quad (28.65)$$

et il s'agit donc d'un vecteur.

La dérivée de $h_s(\vec{r}_i(t))$ par rapport à s donne :

$$\frac{d}{ds} h_s(\vec{r}_i(t)) = \frac{d}{ds} [\vec{r}_i(t) + s(\vec{e}_z \times \vec{r}_i(t))] = \vec{e}_z \times \vec{r}_i(t) \quad (28.66)$$

En remarquant encore une fois que:

$$\vec{p}_i = \frac{\partial L}{\partial \dot{\vec{r}}_i} \quad (28.67)$$

la grandeur conservée obtenue par application du théorème de Noether est alors :

$$C(\vec{r}'_i, \dot{\vec{r}}'_i, t) = \left(\frac{\partial L}{\partial \dot{\vec{r}}_i} \frac{dh_s}{ds} \right)_{s \rightarrow 0} = \vec{p}_i \circ (\vec{e}_z \times \vec{r}_i(t)) \quad (28.68)$$

et nous avons démontré dans le chapitre de Calcul Vectoriel que:

$$\vec{x} \circ (\vec{y} \times \vec{z}) = -\vec{z} \circ (\vec{y} \times \vec{x}) = -(-\vec{y} \circ (\vec{z} \times \vec{x})) = \vec{y} \circ (\vec{z} \times \vec{x}) \quad (28.69)$$

ce qui nous amène à écrire:

$$C(\vec{r}'_i, \dot{\vec{r}}'_i, t) = \vec{e}_z \circ (\vec{p}_i \times \vec{r}_i(t)) = \vec{e}_z \circ \vec{b}_i = b_{z,i} \quad (28.70)$$

On montrerait de la même façon l'invariance du lagrangien sous les rotations selon les autres axes ce qui conduit à la conservation des composantes suivant ces axes du moment cinétique total du système.

En conclusion, nous avons mis en évidence trois lectures différentes des lois de la physique :

Observation	Loi de conservation	Signification physique
Invariance des lois de la physique par translation	Conservation de la quantité de mouvement	Homogénéité de l'espace : l'espace présente les mêmes propriétés en tous points
Invariance des lois de la physique par rotation	Conservation du moment cinétique	Isotropie de l'espace : l'espace présente les mêmes propriétés dans toutes les directions
Invariance des lois de la physique par déplacement dans le temps	Conservation de l'énergie	Homogénéité du temps : les lois de la nature ne varient pas dans le temps

Tableau: 28.7 - Lois de conservation

Autrement dit, l'Univers serait:

P1. Homogène (pas d'origine de temps, ou d'espace, privilégiée)

P2. Isotrope (pas de direction privilégiée).

3.5. PRINCIPE DE CURIE

Le principe de Curie (que nous devons à Pierre Curie) découle un peu intuitivement du théorème de Noether et s'énonce ainsi:

Si une cause présente une certaine symétrie ou invariance, alors son effet aura la même symétrie (ou la même invariance), ou une symétrie supérieure, à condition toutefois que la solution du problème soit unique.

Remarque: A noter que les éléments de symétrie agissent sur les directions des grandeurs vectorielles, tandis que les invariances agissent sur les variables dont dépendent ces grandeurs.

Exemple:

Conservation de l'énergie/invariance par translation dans le temps, conservation de la quantité de mouvement/invariance par translation dans l'espace, conservation du moment cinétique/invariance par rotation dans l'espace tel que nous l'avons démontré lors de notre étude du théorème de Noether.

Ainsi, dans un espace homogène et isotrope, si nous faisons subir une transformation géométrique à un système physique susceptible de créer certains effets (forces, champs), alors ces effets subissent les mêmes transformations.

Autrement dit, si un système physique S possède un certain degré de symétrie, nous pourrions alors déduire les effets créés par ce système en un point à partir des effets en un autre point.

Voici les six propriétés de symétrie découlant du principe de Curie:

P1. Invariance par translation: si S est invariant dans toute translation parallèle à un axe, les effets sont indépendants des coordonnées de cet axe (l'intérêt étant alors de travailler en coordonnées cartésiennes).

P2. Symétrie axiale: si S est invariant dans toute rotation autour d'un axe donné, alors ses effets exprimés ne dépendent pas de l'angle qui définit la rotation (l'intérêt étant alors de travailler en coordonnées cylindriques).

P3. Symétrie cylindrique: si S est invariant par translation et rotation, alors ses effets ne dépendent que de la distance à l'axe de rotation (l'intérêt étant alors aussi de travailler en coordonnées cylindriques).

P4. Symétrie sphérique: si S est invariant dans toute rotation autour d'un point fixe, alors ses effets ne dépendent que de la distance à ce point fixe (l'intérêt étant alors de travailler en coordonnées sphériques).

P5. Plan de symétrie: si S admet un plan de symétrie, alors en tout point de ce plan:

- un effet à caractère vectoriel est contenu dans le plan
- un effet à caractère pseudo-vectoriel (voir le chapitre de Calcul Vectoriel pour voir la définition d'un pseudo-vecteur) lui est perpendiculaire

P6. Plan d'antisymétrie: si, par symétrie par rapport à un plan, S est transformé en $-S$ alors en tout point

de ce plan:

- un effet à caractère vectoriel est perpendiculaire au plan
- un effet à caractère pseudo-vectoriel est contenu dans ce plan

La symétrie est fondamentale dans les sciences quelles que soient les disciplines. La symétrie est partout. Elle permet de décrire de manière précise de nombreux systèmes, de clarifier et de simplifier l'étude de leurs propriétés. Des résultats très importants peuvent ainsi être prédits de manière rigoureuse sans que l'on ait à faire appel à des théories mathématiques sophistiquées.

4. ESPACES PONCTUELS

L'étude des phénomènes physiques recourt dans un premier temps à leur représentation dans l'espace de la géométrie classique euclidienne à une dimension temporelle et à un nombre quelconque de dimensions spatiales.

Les vecteurs que nous avons étudiés dans le chapitre de Calcul Vectoriel (tenseurs d'ordre 1) et les tenseurs (d'ordre quelconque) que nous avons aussi étudiés dans le chapitre de Calcul Tensoriel peuvent comme nous avons en déjà fait mention, être utilisés pour relier chacun des points de l'espace-temps à un référentiel et former ainsi des champs de vecteurs ou/et de tenseurs. Cet état de fait mathématique, nécessite la définition mathématique d'espaces formés de points ou également appelés "espaces ponctuels".

La définition précise d'espace vectoriel ponctuel que nous allons faire sera construite à partir de la notion d'espace vectoriel que nous avons vue dans le chapitre de Calcul Vectoriel (voir section d'Algèbre)

Voyons tout d'abord l'exemple particulier de l'espace ponctuel formé par des triplets de nombres qui est issu directement de l'espace géométrique classique à trois dimensions.

Ainsi, donnons-nous des triplets de nombres notés:

$$A = (a_1, a_2, a_3) \quad B = (b_1, b_2, b_3) \quad (28.71)$$

etc... Appelons E'_3 l'ensemble de tous les éléments $\{A, B, \dots\}$ formés par des triplets de nombres. À tout couple (A, B) de deux éléments de E'_3 , pris dans cet ordre, nous pouvons faire correspondre un vecteur $\vec{x}$, appartenant à un espace vectoriel E_3 , noté géométriquement $\overrightarrow{AB}$, en définissant celui-ci par un triplet de nombres tel que (nous utilisons la notation indicielle vue dans le chapitre de Calcul Tensoriel):

$$x_i = b_i - a_i \quad (28.72)$$

avec $i = 1, 2, 3$. Nous avons donc:

$$\vec{x} = \overrightarrow{AB} = (x_1, x_2, x_3) \quad (28.73)$$

Si nous définissons par rapport à cet élément l'addition et la multiplication par un scalaire, nous nous retrouvons comme nous l'avons déjà vu en théorie des ensembles (voir chapitre du même nom) avec une structure d'espace vectoriel.

La correspondance que nous établissons ainsi, entre tout couple (A,B) de deux éléments de E'_3 et un vecteur d'un espace vectoriel E_3 , vérifie manifestement les propriétés suivantes:

P1. Antisymétrie: $\overrightarrow{AB} = -\overrightarrow{BA}$

P2. Associativité par rapport à l'addition: $\overrightarrow{AC} = \overrightarrow{AB} + \overrightarrow{BC}$

P3. Si O est un élément arbitraire choisi dans E'_3 , à tout vecteur $\vec{x}$ de E_3 , il correspond un point M et un seul tel que $\overrightarrow{OM} = \vec{x}$.

Lorsque nous avons muni l'ensemble E'_3 de cette loi de correspondance avec E_3 , vérifiant les trois propriétés précédentes, nous disons que l'ensemble des triplets de nombres constitue un "espace ponctuel", noté $\mathcal{E}_3$. Les éléments de $\mathcal{E}_3$ sont alors appelés des "points".

L'espace ponctuel $\mathcal{E}_3$ se confond en tant qu'ensemble d'éléments avec l'ensemble E'_3 mais il s'en distingue en tant qu'espace ponctuel qui constitue un ensemble structuré par la loi de correspondance que nous nous donnons. De même, les espaces E'_3 et $\mathcal{E}_3$ sont distincts par suite de leur structure différente et nous pouvons établir une distinction entre les éléments de chacun des espaces. Nous disons que E'_3 constitue le support des espaces E_3 et $\mathcal{E}_3$.

Nous pouvons bien évidemment généraliser le support à E'_n . Ainsi, E'_n muni de la structure d'espace vectoriel que nous avons définie précédemment constitue un espace ponctuel à n dimensions que nous noterons $\mathcal{E}_n$. Les éléments de $\mathcal{E}_n$ étant appelés des "points".

L'espace vectoriel E_n est appelé "l'espace associé" à $\mathcal{E}_n$. Lorsque l'espace vectoriel associé est un espace pré-euclidien (muni du produit scalaire), nous disons alors que $\mathcal{E}_n$ est un "espace ponctuel pré-euclidien".

Considérons un point O quelconque d'un espace ponctuel pré-euclidien $\mathcal{E}_n$ et une base $(\vec{e}_i)$ de l'espace vectoriel associé E_n .

Définitions:

D1. Nous appelons "repère de l'espace" $\mathcal{E}_n$ l'ensemble constitué par les éléments O (origine) et de la base $(\vec{e}_i)$. Ce genre de repère est noté:

$$(O, \vec{e}_i) \quad (28.74)$$

ou encore simplement:

$$(\vec{e}_i) \quad (28.75)$$

D2. Les "coordonnées" d'un point M d'un espace ponctuel pré-euclidien $\mathcal{E}_n$, par rapport au repère $(\vec{e}_i)$, sont les composantes (contravariantes) x^i du vecteur $\vec{x} = \overrightarrow{OM}$ de l'espace E_n par rapport à la base $(\vec{e}_i)$.

Soient deux points M et M' de $\mathcal{E}_n$ définis par leurs coordonnées respectives x^i et x'^i , nous avons:

$$\overrightarrow{OM} = x^i \vec{e}_i \quad \overrightarrow{OM'} = x'^i \vec{e}_i \quad (28.76)$$

En utilisant les propriétés P1 et P2 données précédemment:

$$\overrightarrow{MM'} = \overrightarrow{MO} + \overrightarrow{OM'} = -\overrightarrow{OM} + \overrightarrow{OM'} = (-x^i + x'^i) \vec{e}_i \quad (28.77)$$

Nous en déduisons que les composantes du vecteur $\overrightarrow{MM'}$, par rapport à la base $(\vec{e}_i)$, sont les n quantités $(x'^i - x^i)$, différences des coordonnées des points M et M' .

Soient $(\vec{e}_i)$ et $(\vec{e}'_j)$ deux repères quelconques de $\mathcal{E}_n$ liés entre eux par les relations générales (cf. chapitres de Calcul Tensoriel et d'Algèbre Linéaire):

$$\vec{e}_i = A^k_i \vec{e}'_k \quad \text{et} \quad \vec{e}'_k = A^i_k \vec{e}_i \quad (28.78)$$

Cherchons les relations entre les coordonnées d'un point M de $\mathcal{E}_n$ par rapport à ces deux repères. Pour cela, exprimons les vecteurs $\overrightarrow{OO'}$ et $\overrightarrow{O'O}$ sur chacune des bases de $\mathcal{E}_n$:

$$\overrightarrow{OO'} = \alpha^j \vec{e}_j \quad \overrightarrow{O'O} = \alpha'^j \vec{e}'_j \quad (28.79)$$

ainsi que les vecteurs $\overrightarrow{OM}$ et $\overrightarrow{O'M}$, soit:

$$\overrightarrow{OM} = x^i \vec{e}_i \quad \overrightarrow{O'M} = x'^i \vec{e}_i \quad (28.80)$$

Nous avons d'autre part:

$$\overrightarrow{OM} = \overrightarrow{OO'} + \overrightarrow{O'M} = \alpha^j \vec{e}_j + x'^j \vec{e}_j = \alpha^j \vec{e}_j + x'^j A^i_j \vec{e}_i = (\alpha^i + A^i_j x'^j) \vec{e}_i \quad (28.81)$$

Identifiant ce résultat par rapport au vecteur $\vec{e}_i$ dans l'expression de $\overrightarrow{OM}$, nous avons:

$$x^i = \alpha^i + A^i_j x'^j \quad (28.82)$$

Et de façon analogue:

$$x'^j = \alpha'^j + A^j_i x^i \quad (28.83)$$

Ces deux relations sont plus que pratiques en physique où nous avons souvent à considérer un référentiel dans un repère (ainsi nous pouvons exprimer la position d'un point depuis l'un ou l'autre repère en usant de ces deux relations).

Considérons maintenant un espace ponctuel pré-euclidien ainsi que M et M' deux points de cet espace. Nous avons démontré lors de notre étude de la topologie (cf. chapitre de Topologie), que la norme du vecteur $\overrightarrow{MM'}$ est une mesure possible de la distance entre M et M' . Nous avons donc:

$$\overline{MM'} = \|\overrightarrow{MM'}\| \quad (28.84)$$

Si les deux points M et M' ont respectivement pour coordonnées x^i et x'^i , par rapport à un repère $(\vec{e}_i)$, nous savons que nous avons:

$$\overrightarrow{MM'} = (-x^i + x'^i) \vec{e}_i \quad (28.85)$$

La norme au carré est donc donnée comme nous l'avons vu lors de notre étude du calcul tensoriel (cf. chapitre de Calcul Tensoriel) par la relation:

$$\|\overrightarrow{MM'}\|^2 = g_{ij} (x'^i - x^i)(x'^j - x^j) \quad (28.86)$$

Si le point M' est infiniment proche du point M , ses coordonnées sont notées $(x^i + dx^i)$ et le vecteur $\overrightarrow{MM'} = d\vec{M}$ a pour composantes les quantiques dx^i .

Si nous notons ds la distance entre les points M et M' . La relation précédente donne l'expression du carré de la distance entre ces points sous la forme:

$$ds^2 = g_{ij} dx^i dx^j \quad (28.87)$$

Rappelons également (cf. chapitre de Calcul Tensoriel) que pour un espace ponctuel pré-euclidien où les vecteurs de base $\vec{e}_i$ sont donc orthonormés, nous avons:

$$g_{ij} = \delta_{ij} \quad (28.88)$$

et l'expression de la distance devient alors:

$$ds^2 = dx^i dx^i \quad (28.89)$$

Nous obtenons ainsi une expression qui généralise, à n dimensions, le carré de la distance élémentaire, par rapport à un repère cartésien orthonormé, dans l'espace de la géométrie classique (euclidienne).

Les vecteurs de la physique sont généralement des fonctions d'une ou plusieurs variables, celles-ci pouvant être des variables d'espace ou du temps. Lorsque à un point M d'un espace ponctuel E_n , nous attachons un tenseur, défini par ses composantes par rapport à un repère $(\vec{e}_i)$, nous dirons que nous nous sommes donnés un "**champ de tenseurs**" (les champs de tenseur d'ordre 1 étant des champs vectoriels).

Pour des vecteurs à n dimensions, la notion de dérivée d'un vecteur à trois dimensions se généralise et nous obtenons toutes les relations classiques relatives aux dérivées.

Considérons ainsi un vecteur $\vec{x}$ appartenant à un espace pré-euclidien E_n dont les composantes, sur une base $(\vec{e}_i)$, sont des fonctions d'un paramètre quelconque α . Nous noterons ce vecteur $\vec{x}(\alpha)$ et nous aurons:

$$\vec{x}(\alpha) = x^i(\alpha) \vec{e}_i \quad (28.90)$$

Par définition, la dérivée du vecteur $\vec{x}$ par rapport à la composante α est un vecteur noté:

$$\vec{x}'(\alpha) \quad (28.91)$$

selon la notation utilisée par les mathématiciens. Ou:

$$\vec{x}(\alpha) \quad (28.92)$$

selon la notation abrégée des physiciens. Ou encore:

$$d_{\alpha} \vec{x}(\alpha) \quad (28.93)$$

selon l'humeur du physicien. Ou encore:

$$\frac{d\vec{x}}{d\alpha} \quad (28.94)$$

si nous respectons les écritures...

Dans ce site, nous basculons d'une notation à l'autre sans préavis en fonction de l'envie de simplifier les écritures (il faudra s'y faire..).

Etant donné que nous faisons actuellement plus de la mathématique que de la physique, nous noterons:

$$\vec{x}'(\alpha) = \frac{dx^i(\alpha)}{d\alpha} \vec{e}_i \quad (28.95)$$

En rappelant (cf. chapitre de Calcul Différentiel Et Intégral) que la différentielle est donnée par:

$$d\vec{x} = \vec{x}'(\alpha) d\alpha \quad (28.96)$$

Les différentes expressions de dérivations des vecteurs à trois dimensions relatives à la somme de vecteurs, au produit par un scalaire de deux vecteurs, sont aisément transposables aux vecteurs à n dimensions.

Si un vecteur $\vec{x}$ de E_n dépend de plusieurs paramètres indépendants, α, β, γ , la dérivée partielle du vecteur $\vec{x}(\alpha, \beta, \gamma)$ par rapport à la variable α , par exemple, est un vecteur noté:

$$\frac{\partial \vec{x}}{\partial \alpha} \text{ ou } \partial_{\alpha} \vec{x} \quad (28.97)$$

dont les composantes sont les dérivées partielles des composantes de $\vec{x}$, soit:

$$\frac{\partial \vec{x}}{\partial \alpha} = \frac{\partial x^i}{\partial \alpha} \vec{e}_i \quad (28.98)$$

La différentielle totale du vecteur $\vec{x}(\alpha, \beta, \gamma)$ s'écrivant (cf. chapitre de Calcul Différentiel Et Intégral):

$$d\vec{x} = \frac{\partial \vec{x}}{\partial \alpha} d\alpha + \frac{\partial \vec{x}}{\partial \beta} d\beta + \frac{\partial \vec{x}}{\partial \gamma} d\gamma \quad (28.99)$$

Considérons maintenant un espace vectoriel pré-euclidien E_n associé à un espace ponctuel E_n . Dans un repère $(\vec{e}_i)$ tout point M de E_n est associé à un vecteur $\vec{x}$ de E_n tel que $\vec{x} = \overrightarrow{OM}$. Si le vecteur $\vec{x}$ dépend d'un paramètre α et admet une dérivée $\vec{x}'(\alpha)$, il en est de même alors pour $\overrightarrow{OM}$.

Montrons que le vecteur dérivé $\vec{x}'(\alpha)$ ne dépend pas du point origine O (statique) mais seulement du point M considéré. En effet, si O' est un autre point origine, nous avons:

$$\overrightarrow{OM} = \overrightarrow{OO'} + \overrightarrow{O'M} \quad (28.100)$$

et puisque le vecteur $\overrightarrow{OO'}$ est fixe et ne dépend pas de α , nous avons:

$$\frac{d\overrightarrow{OO'}}{d\alpha} = \vec{0} \quad (28.101)$$

d'où:

$$\boxed{\frac{d\overrightarrow{OM}}{d\alpha} = \frac{d\overrightarrow{O'M}}{d\alpha} = \vec{x}'(\alpha)} \quad (28.102)$$

Nous pouvons donc noter la dérivée du vecteur $\overrightarrow{OM}$ en mentionnant seulement le point M et nous écrivons:

$$\vec{x}'(\alpha) = \frac{d\overrightarrow{O'M}}{d\alpha} = \frac{d\vec{M}}{d\alpha} = \vec{M}'(\alpha) \quad (28.103)$$

La différentielle de $\overrightarrow{OM}$ s'écrit alors:

$$d\vec{M} = \vec{M}'(\alpha)d\alpha \quad (28.104)$$

Si un point M de $\mathcal{E}_s$ est associé, par rapport à un repère $(\vec{e}_i)$ à un vecteur $\vec{x}(\alpha, \beta, \gamma) = \overrightarrow{OM}$, les dérivées partielles de $\overrightarrow{OM}$ ne dépendront que du point M et nous écrivons, par exemple:

$$\frac{\partial \vec{x}}{\partial \alpha} = \frac{\partial \vec{M}}{\partial \alpha} \quad (28.105)$$

Afin d'alléger les expressions des dérivées partielles totales des fonctions dépendantes de n variables, nous utilisons quand le contexte s'y prête, les notations indicielles suivantes. Si $f(y^1, y^2, \dots, y^*)$ est une fonction des n variables y^i , nous noterons les dérivées partielles sous la forme:

$$\partial_i f = \frac{\partial f}{\partial y^i} \quad (28.106)$$

Les dérivées secondes par rapport aux variables y^i et y^k s'écriront:

$$\partial_{ik} f = \frac{\partial^2 f}{\partial y^i \partial y^k} \quad (28.107)$$

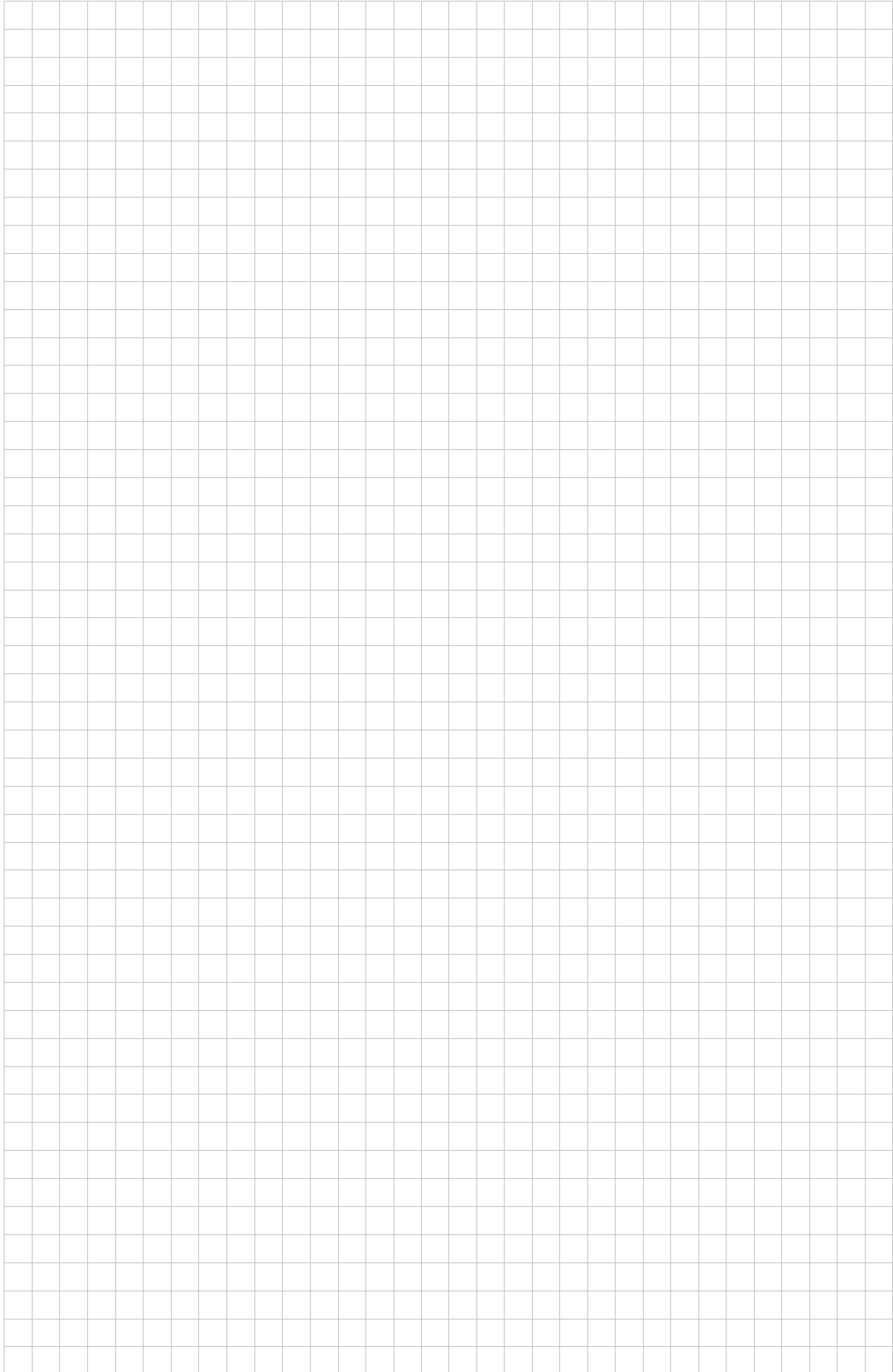
Lorsque $\vec{x}$ est un vecteur tel que $\vec{x} = x^i \vec{e}_i$, dont les composantes sont des fonctions de n variables y^k , soit:

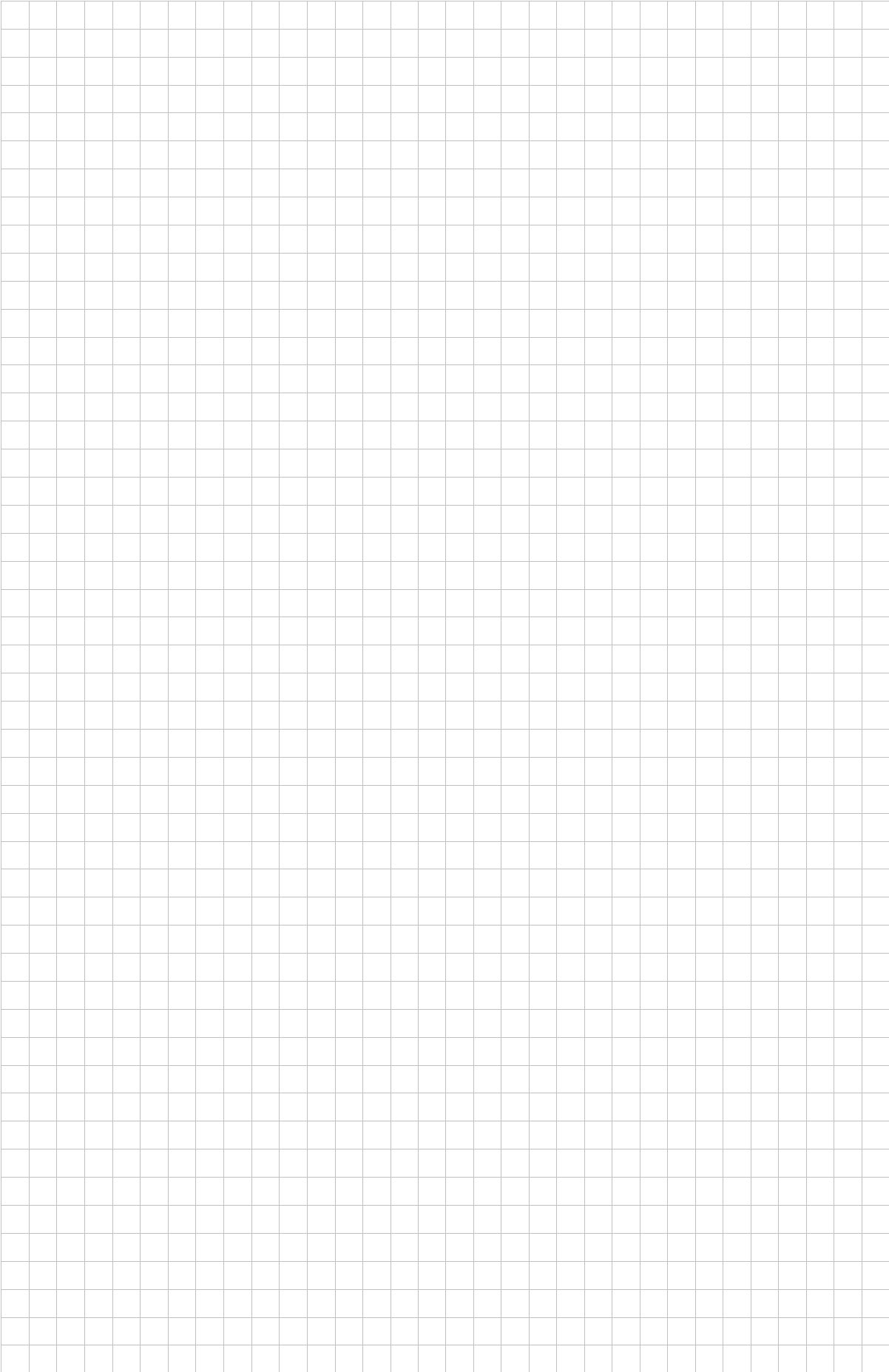
$$x^i = x^i(y^1, y^2, \dots, y^*) \quad (28.108)$$

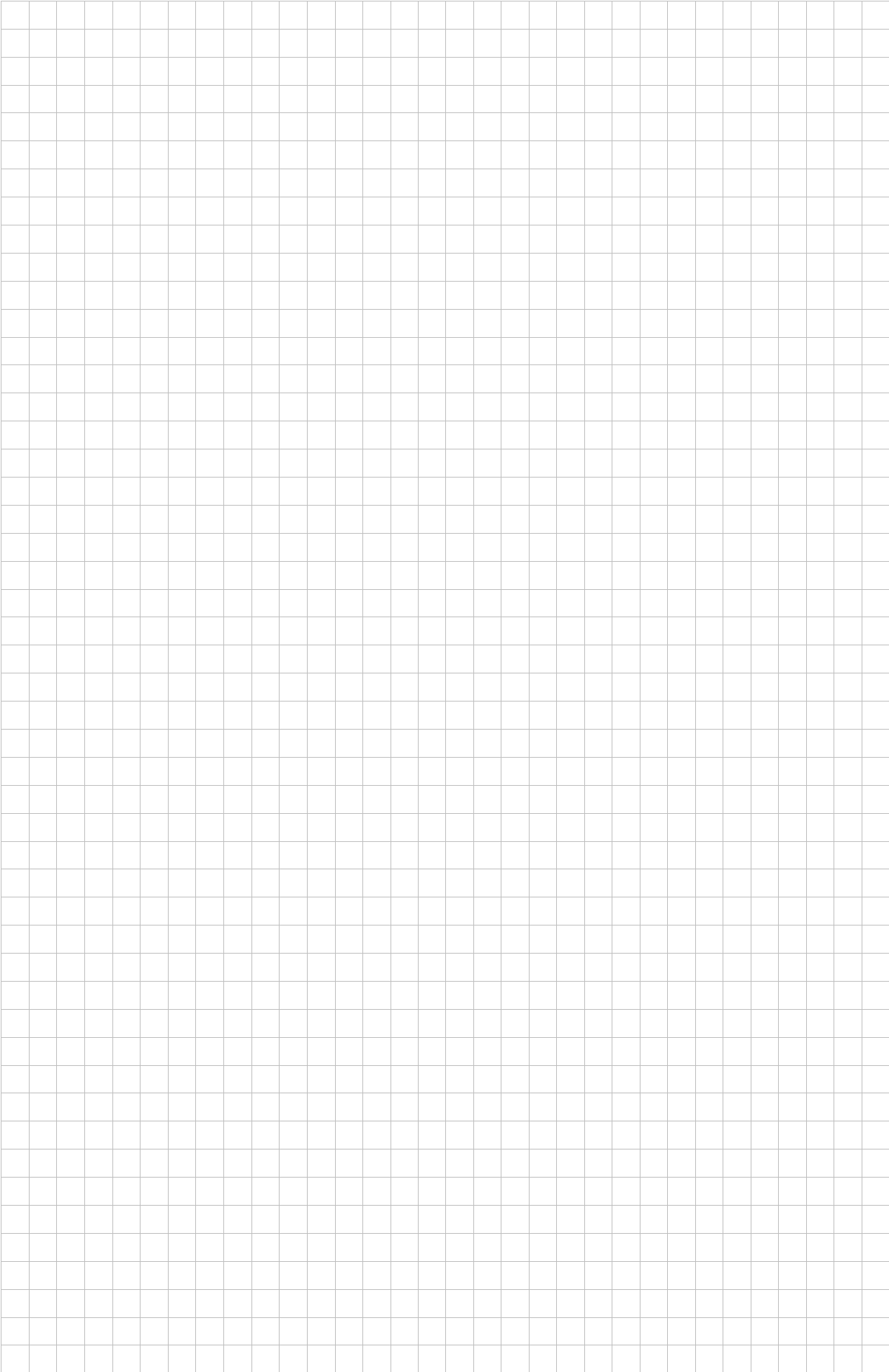
les dérivées partielles du vecteur seront notées, en utilisant la convention de sommation:

$$\frac{\partial \vec{x}}{\partial y^k} = \partial_k \vec{x} = \partial_k (x^j \vec{e}_j) = (\partial_k x^j) \vec{e}_j \quad (28.109)$$

Le concept d'espace ponctuel étant maintenant introduit, nous pouvons maintenant passer à l'étude du formalisme lagrangien et la détermination de la formulation mathématique du principe de moindre action (voir chapitre suivant).







29. MÉCANIQUE ANALYTIQUE

Nous devons la forme actuelle de la mécanique analytique appelée aussi parfois "**mécanique lagrangienne**" aux travaux des frères Bernoulli et particulièrement d'Euler et Lagrange. C'est effectivement en 1696 que commence l'histoire de la vraie physique théorique.

Au fait, l'événement de départ de la mécanique analytique provient de l'observation suivante (énoncée au 17ème siècle): Tout système semble évoluer d'un état à un autre toujours en utilisant les moyens les plus simples et en conservant une grandeur constante entre les deux états.

Remarques:

R1. Les moyens précités peuvent être: le chemin le plus court, le chemin le plus rapide (les trajectoires spatio-temporelles à plus faibles amplitudes en gros...).

R2. Selon le premier principe fondamental de la physique, la grandeur constante est choisie comme étant l'énergie.

Cet énoncé est appelé dans le cadre de la mécanique "**principe de moindre action (de Maupertuis)**" ou dans le cadre de la physique générale "**principe variationnel**" ou encore parfois dans le cadre de l'optique "**principe d'économie**" ou "**principe de Fermat**". Dans le cadre mathématique faisant purement abstraction des concepts physiques, nous parlons de "**principe de Hamilton**".

Plus techniquement, il est aussi formulé de la manière suivante: Un système se meut d'une configuration à une autre de telle façon que la variation de l'action (voir plus loin) entre la trajectoire naturelle effectivement suivie et toute trajectoire virtuelle infiniment voisine ayant les mêmes extrémités dans l'espace et dans le temps soit nulle.

Au fait, bien que cet énoncé puisse paraître comme cohérent, il peut faire douter mais... nous verrons :

1. Qu'en mécanique classique, nous pouvons démontrer la première loi de Newton en admettant ce principe comme vrai et en y superposant le principe de conservation de l'énergie et nous pouvons expliquer le mouvement de nutation de presque tout solide simple.
2. En électromagnétisme, nous retrouverons toutes les équations de Maxwell (in extenso la loi de Biot-Savart, Faraday, force de Lorentz, loi de Laplace, etc.) à partir des propriétés du principe de moindre action et de conservation de l'énergie.
3. En optique, nous démontrerons que le chemin suivi par la lumière est toujours le plus court et cela nous permettra de démontrer le principe de Fermat à la base de toute l'optique géométrique.
4. En physique atomique, les propriétés du principe de moindre action nous permettront de déterminer certaines propriétés mathématiques des atomes et autres particules (les fermions et les bosons en physique quantique des champs).
5. Le principe de moindre action nous permettra également de démontrer que tout corps, avec ou sans

masse, est dévié par un champ d'accélération et... permet donc de déterminer l'équation d'Einstein des champs qui est à la base de tout le chapitre sur la relativité générale.

6. Ce principe s'applique également pour obtenir des résultats puissants en géométrie comme nous allons le voir un peu plus loin. Ainsi, les techniques de la mécanique analytique sont très intimement liées à la mathématique pure.

Il va donc sans dire par ces six petits exemples les applications phénoménales de ce principe!!

Historiquement, il est intéressant de savoir que c'est Pierre-Louis Moreau de Maupertuis qui a énoncé le premier le principe de moindre action sous forme peu scientifique. L'intervention d'Euler et Lagrange dans ce domaine a été de mettre sous forme mathématique ce principe et de démontrer (tenez-vous bien...) qu'il découle d'une simple propriété mathématique des optima des fonctions continues. Il va sans dire, que savoir que cela a permis de redémontrer toutes les lois de la physique classique en a dérangé plus d'un...

Ce principe a eu (et a toujours) des répercussions inimaginables et le problème fut d'appliquer l'expression mathématique de ce dernier à tous les phénomènes physiques qui avaient déjà été démontrés de façon expérimentale et empirique à l'époque. Effectuer cette démonstration revenait ainsi à expliquer pourquoi tel phénomène ou telle loi était ainsi plutôt qu'autrement. Imaginez !

Ainsi, le premier à s'attaquer au problème fût donc le Bâlois (Suisse) Leonhard Euler. Mais nous avons également gardé le nom de Lagrange (d'où l'appellation: "formalisme lagrangien") pour définir toute la méthode et le formalisme mathématique construit autour du principe de moindre action.

1. FORMALISME LAGRANGIEN

La mécanique classique peut être formalisée de différentes manières. La plus courante est la formulation de Newton, qui utilise la notion de force (cf. [chapitre de Mécanique Classique](#)). Elle est de loin la plus simple lorsqu'il s'agit de considérer un problème concret et c'est pourquoi c'est celle qui est enseignée. Mais pour pouvoir traiter des problèmes plus complexes ou plus finement, et pour pouvoir faire des démonstrations rigoureuses, cette formulation n'est pas la plus pratique.

La mécanique analytique, initiée dès le 18ème siècle, regroupe ainsi différentes formulations très mathématisées de la mécanique classique, notamment les mécaniques de Hamilton et de Lagrange (toutes ces formulations sont équivalentes!).

Cette formalisation est assez peu enseignée dans les petites écoles car il faut bien l'avouer le formalisme lagrangien et hamiltonien (contenant donc le principe de moindre action sous forme mathématique) fait appel à un niveau d'abstraction un peu plus élevé que les méthodes normales et malgré qu'il soit souvent d'une aide précieuse dans l'élaboration de théories (physique fondamentale, physique quantique, relativité générale, théorie quantique des champs, théorie des supercordes), il en découle rarement de nouvelles solutions (mais plutôt une réduction et une méthode de validation utile et très puissante).

Commençons donc notre travail:

1.1. COORDONNÉES GÉNÉRALISÉES ET RÉFÉRENTIELS

Un réflexe naturel conduit généralement à référer la position d'un point dans l'espace à la seule connaissance de ses trois coordonnées cartésiennes x, y, z . Cette attitude est d'ailleurs le plus souvent justifiée par la simplicité d'un grand nombre de situations rencontrées dans la pratique, où il n'est pas nécessaire de rechercher de méthodes plus élaborées ou de passer dans d'autres systèmes de coordonnées (cf. chapitre de Calcul Vectoriel).

Pour repérer la position d'un mobile (ou d'un point matériel) en physique il est nécessaire dans un premier temps d'associer un repère au référentiel. Ainsi, un "**repère**" est un système (physique concret) de repérage dans l'espace associé au référentiel.

Les repères conventionnels en mécanique classique constituent majoritairement des bases d'espaces pré-euclidiens canoniques (cf. chapitre de Calcul Vectoriel) orientés et où chaque point, ou vecteur de l'espace, peut-être représenté algébriquement par ses valeurs d'affixes (la valeur à l'ordonnée (projection sur l'axe vertical) et la valeur à l'abscisse (projection sur l'axe horizontal)).

Voici quelques exemples triviaux:

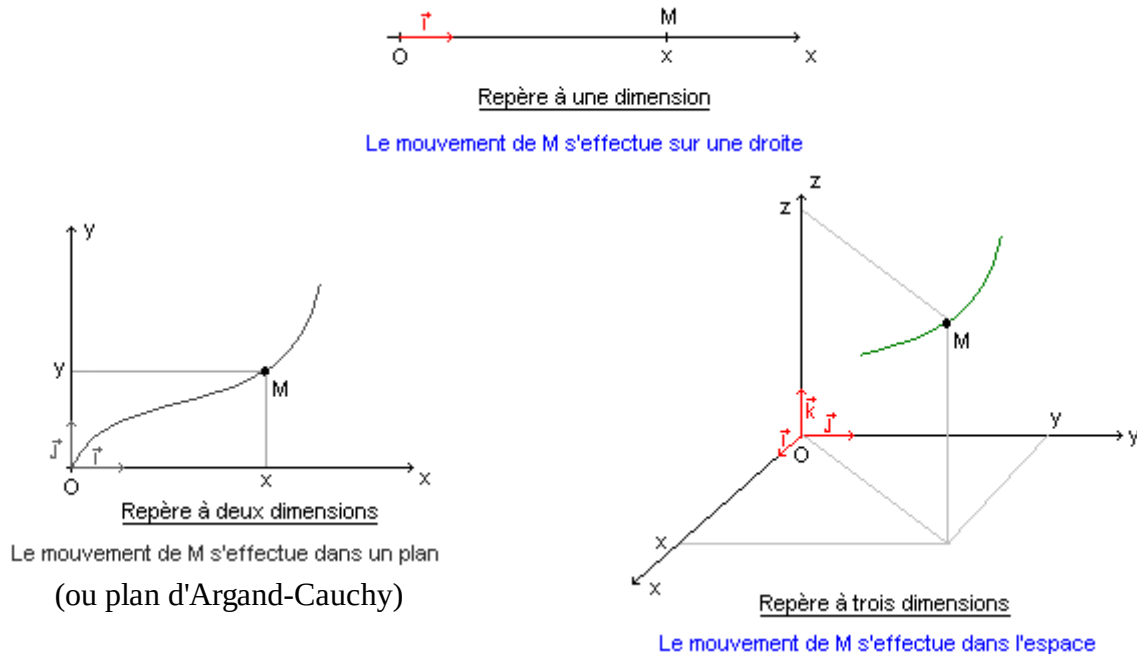


Figure: 29.1 - Mouvements dans des référentiels à 1, 2 ou 3 dimensions

Remarque: Comme nous l'avons vu dans le chapitre de Géométrie Différentielle, la distance entre deux points d'une trajectoire courbe en parcourant la courbe est appelée "**abscisse curviligne**". Sinon, la distance entre deux points d'une trajectoire rectiligne est appelée simplement "**abscisse**".

Définitions:

D1. Un repère, assimilé à un référentiel, est dit "**référentiel Galiléen**" (c'est rare que nous en fassions explicitement mention en physique par manque de rigueur) si:

- Nous pouvons le considérer comme immobile pendant toute l'étude du mouvement du système ou comme étant en translation rectiligne uniforme par rapport à un autre référentiel lui-même immobile.

Donc si nous négligeons le mouvement de rotation du Soleil autour du centre de la galaxie, alors le référentiel héliocentrique peut être considéré comme galiléen. Si nous négligeons le mouvement de rotation de la Terre autour du Soleil, alors le référentiel géocentrique peut être considéré comme galiléen. Si nous négligeons le mouvement de rotation de la Terre sur elle-même, alors le référentiel terrestre peut être considéré comme galiléen. Dans beaucoup d'expériences de mécanique à la surface de la Terre, nous constatons que le référentiel terrestre peut être considéré comme galiléen avec une très bonne précision. Heureusement qu'il y a quand même un tas de phénomènes où il faut tenir compte de la rotation de la Terre (déviations vers l'est, pendule de Foucault...etc.)

- Nous pouvons le considérer comme un système où les lois de Newton sont vérifiées (cf. chapitre de Mécanique Classique)

D2. Un repère, assimilé à un référentiel, est dit "**barycentrique**" (cf. le chapitre de Géométrie Euclidienne) s'il a pour origine le centre de masse (cf. chapitre de Mécanique Classique) du corps étudié.

Ainsi, le "**repère de Copernic**" est assimilé au centre de gravité (d'inertie) du système solaire, le "**repère héliocentrique**" appelé aussi "**repère de Kepler**" au centre d'inertie du Soleil.

D3. Un repère, assimilé à un référentiel, est dit "**référentiel géocentrique**" lorsque nous prenons pour référence un système d'axes placés au centre d'inertie de la Terre. Les axes, parallèles à ceux du référentiel de Copernic, pointent vers trois étoiles fixes. Dans ce référentiel la Terre tourne sur elle-même en 24 [h.].

D4. Un repère, assimilé à un référentiel, est dit "**référentiel Terrestre**" lorsque nous prenons pour référence un système d'axes placés au centre d'inertie de la Terre et qui a un mouvement de rotation uniforme correspondant à la vitesse de rotation de la Terre. Traditionnellement un des axes est dirigé vers l'étoile polaire. C'est le référentiel auquel nous nous référons le plus dans la vie courante il n'est donc pas galiléen en toute rigueur! Ceci va induire des effets particuliers sur les mouvements dans l'atmosphère tels que nous les ressentons.

Remarque: Dire qu'un repère orthonormé $(O, \vec{i}, \vec{j})$ est un "**repère direct**" signifie que l'angle orienté $(\vec{i}, \vec{j})$ a pour mesure principale $\pi/2$ (dans le sens horaire). Dire qu'un repère orthonormé $(O, \vec{i}, \vec{j})$ est un "**repère indirect**" signifie que l'angle orienté a pour mesure principale $-\pi/2$. Dans tout ce qui suit, si nous ne spécifions pas l'orientation, cela sous-entend que $(O, \vec{i}, \vec{j})$ est direct.

Il est bien exact que les trois paramètres x, y, z suffisent parfaitement à repérer un point matériel dans l'espace usuel comme nous en avons déjà fait mention dans notre étude des espaces ponctuels (cf. chapitre sur les Principes), mais il n'en demeure pas moins qu'il est parfois inévitable, ou même tout simplement plus avantageux, d'utiliser un nombre de paramètres supérieur à trois. Nous pouvons évidemment envisager toutes sortes de paramétrages pour atteindre les coordonnées d'un point dans l'espace, de telle sorte que, d'une façon plus généralisée nous serons amenés à prendre en considération des relations du type (nous ne gardons plus la même écriture que celle que nous avons lors de notre étude des espaces ponctuels par cohérence avec les nombreuses références déjà existant sur le sujet):

$$\begin{aligned}
 x &= x^1 = x^1(q^1, q^2, \dots, q^n) \\
 y &= x^2 = x^2(q^1, q^2, \dots, q^n) \\
 z &= x^3 = x^3(q^1, q^2, \dots, q^n)
 \end{aligned}
 \quad (29.1)$$

Les paramètres $q^1, q^2, \dots, q^n$ portent le nom de "**coordonnées généralisées**", paramètres auxquels un problème sera le plus souvent référé. Connaître leur expression en fonction du temps est le problème fondamental de la dynamique. Cela signifie que nous serons parvenus à une solution quand nous disposerons des relations indépendantes:

$$\begin{aligned}
 q^1 &= q^1(t) \\
 q^2 &= q^2(t) \\
 &\dots \\
 q^n &= q^n(t)
 \end{aligned}
 \quad (29.2)$$

Il est donc important de retenir que le nombre de paramètres q^i définissant le repérage d'un point dans l'espace est au moins égal à trois, sans être nécessairement différent de trois. C'est finalement la nature des situations envisagées qui suggèrent le choix du nombre des paramètres à utiliser (coordonnées cartésiennes, cylindriques, sphériques,...).

Dans une vision plus générale, la configuration instantanée d'un système, quelle qu'en soit la nature, sera déterminée par la connaissance, en fonction du temps, de n paramètres, n définissant le nombre de "**degrés de liberté**" du système (cf. chapitre de Mécanique Classique).

Il est tout naturel, mathématiquement, d'associer la manipulation des n paramètres x^α au recours à un hyper-espace à n dimensions, dans lequel les x^α apparaîtraient comme les coordonnées d'un point P représentatif de la configuration d'un système quelconque. Nous donnons à cet espace à n dimensions E_n , le nom "**d'espace de configuration**".

Mais la rigueur de la mathématique-physique, nous amène à disposer d'une description plus précise des phénomènes en ajoutant cette variable importante qu'est le temps, considérée souvent comme variable indépendante, aux x^α . Nous en arriverons donc fatalement à utiliser un autre hyper-espace E_{n+1} auquel nous avons donné le nom "**d'espace des événements**".

Ce dernier espace de référence revêt un intérêt capital pour un grand nombre de problèmes de la science moderne et se trouve particulièrement bien adapté aux raisonnements de nature relativiste. Les variables indépendantes constituant les coordonnées spatiales et temporelles forment alors ce que nous appelons les "**variables d'Euler**".

Dans la mesure où les paramètres x^α sont simplement présentés comme des fonctions explicites du temps, le point P décrit une courbe paramétrée, définie par $x^\alpha = x^\alpha(t)$, avec $\alpha = 1, 2, \dots, n$. Cela revient à exploiter simultanément les équations:

$$\begin{aligned}
 x^1 &= x^1(t) \\
 x^2 &= x^2(t) \\
 &\dots \\
 x^n &= x^n(t)
 \end{aligned}
 \quad (29.3)$$

Il arrivera fréquemment que, pour des raisons d'opportunité, nous souhaitons changer de système de coordonnées généralisées, et utiliser un autre ensemble plus compatible avec les spécificités du problème envisagé. Nous substituerons alors au jeu des x^α un nouveau jeu de coordonnées q^i . Il est alors évident que nous devons, avant toute chose, nous doter des relations de dépendance existant entre les deux ensembles de coordonnées (cf. chapitre de Calcul Vectoriel):

$$\begin{aligned} x^1 &= x^1(q^1, q^2, \dots, q^n) \\ x^2 &= x^2(q^1, q^2, \dots, q^n) \\ &\dots \\ x^n &= x^n(q^1, q^2, \dots, q^n) \end{aligned} \quad (29.4)$$

Les fonctions $x^1, x^2, \dots, x^n$ seront maintenant supposées définies, continues, de classe C^2 (pour travailler avec l'accélération) par rapport aux $q^1, q^2, \dots, q^n$ et devront conduire à un jacobien différent de zéro (cf. chapitre de Calcul Différentiel Et Intégral).

Dans ces conditions, à chaque point P_x de l'espace des configurations des x , noté E_x , correspondra un point P_q de l'espace de configuration des q , noté E_q . Nous avons ainsi effectué une transformation ponctuelle, autrement dit une application de l'espace sur lui-même.

Pour étudier des milieux continus (concept radicalement différent du point matériel), nous aurons cependant deux approches différentes:

1. Méthode de Lagrange: nous cherchons à caractériser le mouvement du milieu décrit par une formulation Lagrangienne consistant donc à le caractériser en se donnant un système d'équations au sens newtonien. Par dérivations, nous avons alors la vitesse et l'accélération du milieu.
2. Méthode d'Euler: Au lieu de suivre le parcours d'un point, nous portons notre attention sur l'évolution des caractéristiques physiques en un point donné comme la vitesse, l'accélération la température, la pression ou autre. Nous parlons alors fréquemment de "**système Eulérien**".

1.2. PRINCIPE VARIATIONNEL

Le "**principe variationnel**" n'est donc que la forme mathématique contemporaine du principe de moindre action qui est, comme nous en avons déjà fait mention, à la base du formalisme lagrangien.

Rappelons que selon l'énoncé du principe variationnel nous devons trouver dans tout phénomène physique, une certaine quantité qui est naturellement optimisée (minimisée ou maximisée) et qui décrit toutes les variables du système étudié et ainsi son issue.

Voici la démarche que nous allons suivre; une fois cette démarche présentée, nous nous attaquerons à sa formalisation mathématique.

Les propositions sont les suivantes:

P1. Nous supposons donc le principe variationnel et le principe de conservation de l'énergie comme justes.

P2. L'énergie totale d'un système fermé est constante et constituée de la sommation de l'énergie cinétique et l'énergie potentielle.

Si nous ne considérons que l'énergie cinétique, alors le système est dit "**système libre**"; si les deux énergies sont considérées, nous disons alors que le système est un "**système généralisé**".

P3. Nous définissons une fonction mathématique (dont les variables sont les coordonnées généralisées) appelée "**Lagrangien**" qui est donnée par la différence entre les deux énergies précitées.

P4. Sur l'évolution d'un système entre deux états, nous cherchons les propriétés de la fonction (du lagrangien) qui donne la minimisation de la variation de la différence des deux énergies sur l'évolution temporelle ou métrique du système.

Enfin, une fois cette propriété déterminée (mise sous la forme que nous appelons "équation d'Euler-Lagrange") nous chercherons toutes les autres propriétés possibles afin d'avoir les outils nécessaires pour la physique théorique et vous allez voir cela marche terriblement bien...

Donc, pour mettre cela sous forme mathématique, nous commençons par poser qu'il existe une fonction réelle de $2n$ variables:

$$(q_1, \dots, q_n, \dot{q}_1, \dots, \dot{q}_n) \mapsto L(q_1, \dots, q_n, \dot{q}_1, \dots, \dot{q}_n) \quad (29.5)$$

que nous appellerons "**Lagrangien généralisé**" du système, dont l'intégrale satisfait à l'énoncé suivant:

Dans un mouvement naturel partant d'un point $A(q_1^A, \dots, q_n^A)$ à l'instant t_A , arrivant au point $B(q_1^B, \dots, q_n^B)$ à l'instant t_B , l'intégrale suivante appelée "**intégrale d'action**" ou simplement "**action**":

$$S(q, \dot{q}) = \int_{t_A}^{t_B} L[q_1(t), \dots, q_n(t), \dot{q}_1(t), \dots, \dot{q}_n(t)] dt \quad (29.6)$$

qui peut aussi être notée dans une écriture plus abrégée:

$$S(q, \dot{q}) = \int_{t_A}^{t_B} L(q, \dot{q}) dt \quad (29.7)$$

doit être un extrémum (en fait, "un minimum" ou "un maximum", puisque nous aurions pu tout aussi bien prendre $-L$ au lieu de $+L$ dans le choix de la définition du Lagrangien généralisé).

L'action S est ce que nous appelons communément en physique une "**fonctionnelle**" et a les unités de l'énergie multipliée par le temps puisque L est une énergie.

1.3. ÉQUATION D'EULER-LAGRANGE

Le principe de moindre action énonce donc que (l'intégrale) S est extrémale si:

$$t \mapsto [q_1(t), \dots, q_n(t), \dot{q}_1(t), \dots, \dot{q}_n(t)] \quad (29.8)$$

est la trajectoire naturelle effectivement suivie par le système physique.

Considérons alors une trajectoire très voisine à la précédente, que nous noterons:

$$(q_1 + \delta q_1, \dots, q_n + \delta q_n, \dot{q}_1 + \delta \dot{q}_1, \dots, \dot{q}_n + \delta \dot{q}_n) \quad (29.9)$$

C'est-à-dire que pour chaque i nous posons:

$$\begin{aligned} q_i(t) &\rightarrow q_i(t) + \delta q_i(t) \\ \dot{q}_i(t) &\rightarrow \dot{q}_i(t) + \delta \dot{q}_i(t) \end{aligned} \quad (29.10)$$

avec:

$$\delta q_i(t_A) = \delta q_i(t_B) = 0 \quad (29.11)$$

pour assurer que nous partons toujours du même point A pour arriver au même point B .

Remarque: Nous avons omis maintenant l'écriture des arguments t des fonctions du temps afin d'alléger les écritures.

Si $S(q, \dot{q})$ est bien l'évolution d'un système évoluant selon le principe de moindre action, alors l'action donnée par la variation:

$$\delta S \equiv S(q + \delta q, \dot{q} + \delta \dot{q}) - S(q, \dot{q}) \quad (29.12)$$

est nulle pour δq et $\delta \dot{q}$ tendant vers zéro (sous-entendu que tout système physique revient à son état initial sans intervention extérieure).

Ce qui nous amène à écrire:

$$\delta S = \int_{t_A}^{t_B} L(q + \delta q, \dot{q} + \delta \dot{q}) dt - \int_{t_A}^{t_B} L(q, \dot{q}) dt = \int_{t_A}^{t_B} [L(q + \delta q, \dot{q} + \delta \dot{q}) - L(q, \dot{q})] dt = 0 \quad (29.13)$$

Ce qui nous permet de justifier la dénomination de "**principe variationnel**" (aussi appelé parfois le "principe de stationnarité de l'action"):

$$\delta S = \delta \int_{t_A}^{t_B} L dt = \int_{t_A}^{t_B} \delta L dt = 0 \quad (29.14)$$

Ce principe stipule donc que la trajectoire d'une particule (ou d'un système de manière plus générale) s'obtient en demandant qu'une certaine fonctionnelle S appelée "**action**" soit stationnaire par rapport à une variation de la trajectoire. En d'autres termes, si nous effectuons une variation infiniment petite de la trajectoire, la variation doit être nulle.

Pour un système mécanique simple, l'action est alors évidemment, de par le principe de conservation de l'énergie égale à l'intégrale sur la trajectoire de (par définition du lagrangien) la différence entre l'énergie cinétique et l'énergie potentielle.

Dès lors, dans une théorie pour laquelle les forces dérivent d'un potentiel V , nous sommes naturellement amenés à définir le "**Lagrangien**" par la relation (il faudra s'en souvenir !):

$$L(q, \dot{q}) = T(q, \dot{q}) - V(q, \dot{q}) \quad (29.15)$$

où T et V sont la notation traditionnelle dans le formalisme Lagrangien de l'énergie cinétique et de l'énergie potentielle données par:

$$T = \frac{1}{2} \dot{q}^2 \text{ et } V = V(q) \quad (29.16)$$

Remarque: Pour l'étude de la relativité générale, nous ne chercherons pas à ce que la variation de la différence des énergies soit minimale tel que c'est le cas pour les systèmes mécaniques, mais bien la variation de la longueur d'un arc ds (non dépendant du temps contrairement à l'exemple précédent) dans un espace quelconque lors d'une trajectoire d'un système libre. Ce qui nous amènera à écrire simplement (rappelez-vous en aussi car ce sera très important) l'action:

$$\delta S = \int \delta(ds) = 0 \quad (29.17)$$

pour une masse unitaire et en prenant les unités naturelles.

Pour revenir à notre application du principe variationnel dans le cas du lagrangien généralisé, nous pouvons alors écrire la différentielle totale exacte (cf. chapitre de Calcul Différentiel Et Intégral) de dL et nous obtenons alors la relation:

$$\delta S = \int_{t_A}^{t_B} \sum_{i=1}^n \left(\frac{\partial L}{\partial q_i} \delta q_i + \frac{\partial L}{\partial \dot{q}_i} \delta \dot{q}_i \right) dt = 0 \quad (29.18)$$

Intégrons par parties (cf. chapitre de Calcul Différentiel Et Intégral) le deuxième terme de la somme de l'intégrale précédente:

$$\int_{t_A}^{t_B} \frac{\partial L}{\partial \dot{q}_i} \delta \dot{q}_i dt = \int_{t_A}^{t_B} \frac{\partial L}{\partial \dot{q}_i} \frac{d}{dt} \delta q_i dt = \frac{\partial L}{\partial \dot{q}_i} \delta q_i \Big|_{t_A}^{t_B} - \int_{t_A}^{t_B} \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} \delta q_i dt \quad (29.19)$$

Le premier terme de la dernière égalité est nul:

$$\frac{\partial L}{\partial \dot{q}_i} \delta q_i \Big|_{t_A}^{t_B} = \frac{\partial L}{\partial \dot{q}_i} (t_B) \delta q_i(t_B) - \frac{\partial L}{\partial \dot{q}_i} (t_A) \delta q_i(t_A) = \frac{\partial L}{\partial \dot{q}_i} (t_B) \cdot 0 - \frac{\partial L}{\partial \dot{q}_i} (t_A) \cdot 0 = 0 \quad (29.20)$$

puisque nous avons déjà mentionné plus haut que par construction il faut que:

$$\delta q_i(t_A) = \delta q_i(t_B) = 0 \quad (29.21)$$

L'expression de l'intégrale de moindre action peut finalement s'écrire :

$$\delta S = \int_{t_A}^{t_B} \sum_{i=1}^n \left(\frac{\partial L}{\partial q_i} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} \right) \delta q_i dt = 0 \quad (29.22)$$

Mais les δq_i et $\delta \dot{q}_i$ tendent vers 0 d'une infinité de manières différentes et nous devons cependant avoir néanmoins $\delta S = 0$. Cela veut dire alors que chaque terme sommé de l'intégrale peut être pris

indépendamment et doit satisfaire:

$$\forall i \in \{1, \dots, n\}, \delta S = \int_{t_a}^{t_b} \left(\frac{\partial L}{\partial q_i} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} \right) \delta q_i dt = 0 \quad (29.23)$$

Mais comme les fonctions δq_i et $\delta \dot{q}_i$ peuvent toujours tendre vers zéro de multiples façons, et que cette intégrale doit être quand même nulle, nous en déduisons que ce sont les intégrandes qui sont nuls:

$$\forall i \in \{1, \dots, n\}, \delta S = \frac{\partial L}{\partial q_i} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} = 0 \quad (29.24)$$

Ces n équations, satisfaites par le lagrangien généralisé du système pour le mouvement effectivement suivi, sont appelées "**équations d'Euler-Lagrange**", ou plus brièvement (mais plus rarement) "**équations de Lagrange**". Ce sont, comme nous allons le voir, les équations du mouvement du système: résolues, elles donnent l'évolution effective du système dans le temps.

$$\frac{\partial L}{\partial q_i} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} = 0 \quad (29.25)$$

ou plus explicitement:

$$\boxed{\frac{\partial L(q_i, \dot{q}_i, t)}{\partial q_i} - \frac{d}{dt} \frac{\partial L(q_i, \dot{q}_i, t)}{\partial \dot{q}_i} = 0} \quad (29.26)$$

Remarque: C'est en étudiant la physique (les chapitres suivants du site) que l'on comprend mieux les applications de cette équation (obtenue quasiment que par des développements purement mathématiques !!!) et qu'il devient alors possible de comprendre sa signification. A notre niveau du discours, il est inutile de dire quoi que ce soit. Il faut faire de la physique, et encore de la physique pour la comprendre et la voir apparaître.

Donc dans l'approche lagrangienne, nous apprenons à raisonner à partir des concepts d'énergie potentielle et cinétique, au lieu des concepts de force. Les deux approches sont évidemment équivalentes physiquement, mais les énergies n'étant pas des quantités vectorielles, elles sont conceptuellement plus faciles à utiliser dans une vaste gamme de problèmes. En physique quantique par exemple, la notion de force n'a aucune signification mais les notions d'énergie demeurent valables. C'est une raison de plus pour se familiariser avec leur utilisation. De plus, la force au sens de Newton est une action instantanée à distance. En relativité, une telle chose est impossible. La notion de force est donc une création purement classique et macroscopique contrairement à notre intuition, son intérêt est limité.

Voyons un exemple d'application particulièrement simple de l'équation d'Euler-Lagrange (les autres exemples seront vus pendant notre étude des lois de Newton, de l'électrodynamique, de la relativité restreinte, de la relativité générale, de la physique quantique des champs, etc.):

Exemple:

Dans un premier temps, posons sous une forme mathématique conventionnelle l'équation d'Euler-Lagrange (la notation des coordonnées généralisées n'est pas identique en mathématiques à celle de la

physique...):

$$\frac{d}{dx} \left(\frac{\partial Z(x, y, y')}{\partial y'} \right) - \frac{\partial Z(x, y, y')}{\partial y} = 0 \quad (29.27)$$

Prenons un exemple mathématique pratique simple mondialement connu et très important (nous réutiliserons les développements effectués ici pour l'étude du pendule de Huygens). L'énoncé du problème est le suivant: déterminer quel est le plus court chemin entre deux points d'un plan (nous devinons que c'est la droite mais il faut le démontrer!).

Ce problème consiste à trouver la courbe paramétrée la plus courte $x(t), t \in [t_0, t_1]$ qui relie deux points (attention la variable t n'a rien à voir avec le temps dans cet exemple!):

$$x(t_0) = x_0, x(t_1) = x_1 \quad (29.28)$$

Ainsi la longueur infinitésimale par application de Pythagore est:

$$\begin{aligned} dL^2 &= dt^2 + dx^2 \\ dL &= \sqrt{dt^2 + dx^2} = \sqrt{1 + \frac{dx^2}{dt^2}} dt = \sqrt{1 + \dot{x}^2(t)} dt \end{aligned} \quad (29.29)$$

Ainsi, la longueur de la courbe paramétrée est donnée par:

$$L = \int_{t_0}^{t_1} \sqrt{1 + \dot{x}^2(t)} dt \quad (29.30)$$

Il s'agit d'une relation que nous retrouverons souvent en physique et en mathématiques!!

Ainsi, ce problème, dont la solution géométrique est très simple, se formule sous forme de problème de calcul variationnel de la manière suivante:

$$\min L(x, \dot{x}, t) = \sqrt{1 + \dot{x}^2} \quad (29.31)$$

Ecrivons l'équation d'Euler-Lagrange que la solution de ce problème, si elle existe, doit vérifier.

Nous avons donc:

$$\begin{aligned} Z(x, y, y') &= \sqrt{1 + y'^2} \\ \frac{\partial Z}{\partial y} &= 0, \frac{\partial Z}{\partial y'} = \frac{y'}{\sqrt{1 + y'^2}} \end{aligned} \quad (29.32)$$

L'équation d'Euler-Lagrange dans ce cas particulier devient alors:

$$\frac{d}{dx} \left(\frac{\partial Z(x, y, y')}{\partial y'} \right) - \frac{\partial Z(x, y, y')}{\partial y} = \frac{d}{dx} \left(\frac{y'}{\sqrt{1 + y'^2}} \right) = 0 \quad (29.33)$$

Donc:

$$\frac{y'}{\sqrt{1+y'^2}} = C, C \in \mathbb{R} \quad (29.34)$$

où C est une constante d'intégration (cf. chapitre de Calcul Différentiel Et Intégral). Cette dernière égalité implique que (nous ajoutons un indice pour la constante qui nous évitera de la confondre avec deux autres constantes un peu plus bas):

$$y' = \frac{dy}{dx} = C_1 \sqrt{1-y'^2} \quad (29.35)$$

En revenant aux notations utilisées au début de l'énoncé du problème:

$$x' = \frac{dx}{dt} = \dot{x} = C_1 \sqrt{1-\dot{x}^2} \quad (29.36)$$

Soit:

$$\dot{x}^2 = C_1^2 - C_1^2 \dot{x}^2 \Leftrightarrow (C_1^2 + 1) \dot{x}^2 = C_1^2 \Leftrightarrow \dot{x} = \sqrt{\frac{C_1^2}{C_1^2 + 1}} \quad (29.37)$$

d'où:

$$dx = C_2 dt \quad (29.38)$$

et par intégration il vient donc:

$$x(t) = C_2 t + C_3 \quad (29.39)$$

ce qui est bien l'équation d'une droite. Autrement écrite:

$$y(x) = ax + b \quad (29.40)$$

1.3.1. IDENTITÉ DE BELTRAMI

Nous allons démontrer ici une relation utile dans certaines situations. Il s'agit d'une relation appelée "identité de Beltrami" qui simplifie l'application de l'équation d'Euler-Lagrange dans certaines situations bien particulières!

Nous rappelons d'abord l'équation d'Euler-Lagrange:

$$\frac{\partial L(q_i, \dot{q}_i, t)}{\partial q_i} - \frac{d}{dt} \frac{\partial L(q_i, \dot{q}_i, t)}{\partial \dot{q}_i} = 0 \quad (29.41)$$

Ecrivons la différentielle totale exacte (cf. chapitre de Calcul Différentiel et Intégral):

$$dL = \frac{\partial L}{\partial t} dt + \frac{\partial L}{\partial q_i} dq_i + \frac{\partial L}{\partial \dot{q}_i} d\dot{q}_i \quad (29.42)$$

Ce que nous écrirons sous la forme:

$$\frac{dL}{dt} = \frac{\partial L}{\partial t} + \frac{\partial L}{\partial q_i} \frac{dq_i}{dt} + \frac{\partial L}{\partial \dot{q}_i} \frac{d\dot{q}_i}{dt} = \frac{\partial L}{\partial t} + \frac{\partial L}{\partial q_i} \dot{q}_i + \frac{\partial L}{\partial \dot{q}_i} \ddot{q}_i \quad (29.43)$$

En réarrangeant:

$$\frac{\partial L}{\partial \dot{q}_i} \ddot{q}_i = \frac{dL}{dt} - \frac{\partial L}{\partial q_i} \dot{q}_i - \frac{\partial L}{\partial t} \quad (29.44)$$

Multiplions avant de continuer l'équation d'Euler-Lagrange par $\dot{q}_i$:

$$\dot{q}_i \frac{\partial L}{\partial \dot{q}_i} - \dot{q}_i \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} = 0 \quad (29.45)$$

et injectons la relation antérieure dans la précédente:

$$\dot{q}_i \frac{\partial L}{\partial \dot{q}_i} - \dot{q}_i \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} = \left(\frac{dL}{dt} - \frac{\partial L}{\partial q_i} \dot{q}_i - \frac{\partial L}{\partial t} \right) - \dot{q}_i \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} = 0 \quad (29.46)$$

Après une petite factorisation nous obtenons:

$$-\frac{\partial L}{\partial t} + \frac{d}{dt} \left(L - \dot{q}_i \frac{\partial L}{\partial \dot{q}_i} \right) = 0 \quad (29.47)$$

Dans les conditions particulières (mais relativement fréquents en physique), que nous appellerons sur ce site la "**condition de Beltrami**", où:

$$\frac{\partial L}{\partial t} = 0 \quad (29.48)$$

nous obtenons "**l'identité de Beltrami**":

$$\frac{d}{dt} \left(L - \dot{q}_i \frac{\partial L}{\partial \dot{q}_i} \right) = 0 \quad (29.49)$$

qui nous amène de suite à avoir:

$$L - \dot{q}_i \frac{\partial L}{\partial \dot{q}_i} = c^{te} \quad (29.50)$$

qui nous sera par exemple utile dans le chapitre de Mécanique Classique lors de notre étude du brachistochrone.

1.3.2. THÉORÈME DU CALCUL VARIATIONNEL

Le théorème du calcul variationnel consiste à montrer qu'en considérant f une fonction continue sur $[a, b]$ à valeurs réelles et H l'ensemble des fonctions continues sur $[a, b]$ indéfiniment dérivables sur $]a, b[$ et qui s'annulent en a et b alors pour toute fonction $h \in H$:

$$\int_a^b f(x) h(x) dx = 0 \quad (29.51)$$

f est nulle sur $[a, b]$.

Pourquoi s'intéresser à ce théorème? Parce que nous le rencontrerons très souvent lors de l'application du principe variationnel ayant une configuration de ce type. Effectivement, rappelons que le principe variationnel amène à avoir:

$$\delta S = \int \delta(ds) = 0 \quad (29.52)$$

et l'expression intégrée est rarement une fonction simple comme le lecteur s'en apercevra au cours de sa lecture des différents chapitres du site. Il est donc important de connaître une propriété qui simplifie parfois l'analyse du problème.

Remarque: Certains penseront que le cas avec $f(x) = 1$ avec $h(x) = \sin(x)$ et $x \in [0, 2\pi]$ contredit l'énoncé du théorème! Au fait ce n'est pas vraiment ça... le théorème se doit d'être valable pour $\forall h \in H$ et non juste pour l'exemple cité. D'où le fait que f devra bien être nul comme nous allons le démontrer.

Démonstration:

Pour simplifier nous prendrons le cas $a = -1$, $b = 1$. A quelques détails techniques près la preuve par l'absurde ci-dessous peut être adaptée au cas a, b quelconques.

Supposons que f ne soit pas nulle sur $[-1, 1]$. Alors il existe $x_0 \in]-1, 1[$ tel que $f(x_0) \neq 0$. Nous pouvons supposer $f(x_0) > 0$ (même raisonnement si $f(x_0) < 0$).

Par l'hypothèse initiale de continuité et de non nullité de f il existe alors un petit intervalle autour de x_0 sur lequel f est strictement positive. C'est-à-dire, qu'il existe $\varepsilon > 0$ tel que $[x_0 - \varepsilon, x_0 + \varepsilon] \subset [-1, 1]$ et $\forall y \in [x_0 - \varepsilon, x_0 + \varepsilon], f(y) > 0$.

Considérons à présent la fonction $\phi: [-1, 1] \rightarrow \mathbb{R}$ définie par

$$\begin{aligned}\phi(x) &= \exp\left(-\frac{1}{1-x^2}\right) \text{ si } |x| < 1 \\ \phi(x) &= 0 \text{ sinon.}\end{aligned} \quad (29.53)$$

Nous vérifions assez facilement que ϕ est continue (positive) sur $[-1, 1]$ et indéfiniment dérivable sur $] -1, 1[$ (cf. chapitres d'Analyse Fonctionnelle et de Calcul Différentiel Et Intégral).

De plus, $\phi(-1) = \phi(1) = 0$. Et donc, $\phi \in H$. Voici une représentation graphique de ϕ :

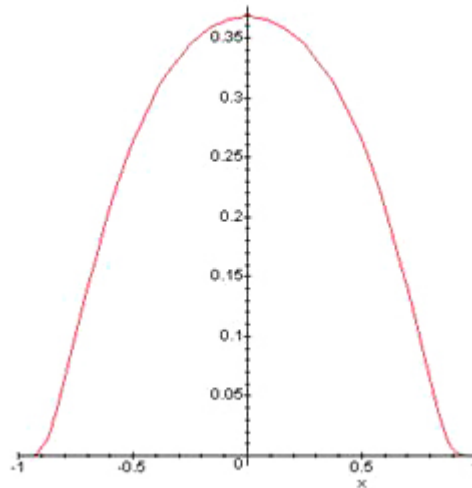


Figure: 29.2 - Exemple de fonction pour la démonstration

A partir de ϕ nous voulons obtenir une fonction continue sur $[-1, 1]$, indéfiniment dérivable sur $] -1, 1[$ positive sur $[x_0 - \varepsilon, x_0 + \varepsilon]$ et nulle en dehors de $[x_0 - \varepsilon, x_0 + \varepsilon]$ afin de montrer l'absurde de l'hypothèse de non nullité de f pour que le théorème soit vérifié (rappelons que nous sommes en train de faire une démonstration par l'absurde!).

Pour ceci, il suffit de centrer ϕ en x_0 et de la contracter.

La fonction $h : [-1, 1] \rightarrow \mathbb{R}$ définie par:

$$\begin{aligned}h(x) &= \phi\left(\frac{x-x_0}{\varepsilon}\right) \text{ si } x \in [x_0 - \varepsilon, x_0 + \varepsilon] \\ h(x) &= 0 \text{ sinon.}\end{aligned} \quad (29.54)$$

répond aux critères exigés. De plus, $h(-1) = h(1) = 0$ et donc, $h \in H$.

Ainsi, la fonction $f \cdot h$ sera continue sur $[-1, 1]$ positive sur $[x_0 - \varepsilon, x_0 + \varepsilon]$ et nulle ailleurs.

Nous avons:

$$\int_{-1}^1 f(x)h(x)dx = 0 \quad (29.55)$$

Or, si une fonction $g : [a, b] \rightarrow \mathbb{R}$ est continue et positive et:

$$\int_a^b g(x) dx = 0 \quad (29.56)$$

cela entraîne forcément (nous supposons cela comme trop intuitif pour avoir besoin d'être démontré) $g = 0$ sur $[a, b]$.

Par conséquent $f(x)h(x) = 0$ sur $[x_0 - \varepsilon, x_0 + \varepsilon]$ or $f(x_0)h(x_0) \neq 0$ selon notre hypothèse absurde initiale, ce qui est contradictoire.

L'hypothèse de départ est donc bien fausse et f doit être nulle sur $[-1, 1]$

□C.Q.F.D.

1.4. FORMALISME CANONIQUE

Le formalisme canonique n'introduit pas une nouvelle physique mais propose une nouvelle gamme d'outils pour étudier les phénomènes physiques. Son élément central, le "**Hamiltonien**", joue un grand rôle en physique quantique.

Comme dans le formalisme de Lagrange nous travaillerons avec des quantités comme l'énergie, T et V plutôt qu'avec des quantités vectorielles comme la force de Newton.

Dans le formalisme de Lagrange, la description d'un système mécanique à n degrés de liberté décrits par les coordonnées générales q_i indépendantes (non contraintes) nous mène à n équations d'Euler-Lagrange:

$$\frac{\partial L}{\partial q_i} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} = 0 \Leftrightarrow \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} - \frac{\partial L}{\partial q_i} = 0 \quad (29.57)$$

qui sont des équations différentielles du 2ème ordre.

Dans le formalisme canonique (ou de Hamilton), un système mécanique à n degrés de liberté toujours décrits par des q_i indépendants nous mènera à $2n$ équations du premier ordre (plus simple à résoudre).

Chez Lagrange nous comparons principalement des trajectoires et par conséquent les q_i et les $\dot{q}_i$ sont tous indépendants. Chez Hamilton nous devons d'abord apprendre à définir les "**moments généralisés**", notés p_i , pour remplacer les coordonnées généralisées $\dot{q}_i$ et q_i qui sont aussi tous indépendants.

Remarque: L'origine des moments conjugués sera triviale dès que nous aurons vu un premier exemple concret.

1.4.1. TRANSFORMATION DE LEGENDRE

Cette transformation est souvent utilisée en thermodynamique où elle permet de relier entre eux les différents potentiels thermodynamiques. En mécanique ou en géométrie elle permet de définir le hamiltonien à partir du lagrangien et inversement. Nous en donnons une description simplifiée et suffisante.

Soit une fonction $f(u,v)$ où u,v sont les deux variables indépendantes dont dépend f .

Définissons:

$$w = \frac{\partial f(u,v)}{\partial v} = w(u,v) \quad (29.58)$$

La transformation de Legendre permet de définir une fonction $g(u,w)$ qui peut remplacer $f(u,v)$:

$$f(u,v) \mapsto g(u,w) = v \cdot w - f \quad (29.59)$$

Soit maintenant la différentielle totale de f (cf. chapitre de Calcul Différentiel Et Intégral):

$$df = \frac{\partial f}{\partial u} du + \frac{\partial f}{\partial v} dv = \frac{\partial f}{\partial u} du + w dv \quad (29.60)$$

De la définition de g nous calculons:

$$\begin{aligned} dg &= w dv + v dw - df \\ &= w dv + v dw - \frac{\partial f}{\partial u} du - w dv \\ &= v dw - \frac{\partial f}{\partial u} du = \frac{\partial g}{\partial w} dw + \frac{\partial g}{\partial u} du \end{aligned} \quad (29.61)$$

et nous avons donc:

$$v = \frac{\partial g}{\partial w}, \quad \frac{\partial g}{\partial u} = -\frac{\partial f}{\partial u} \quad (29.62)$$

1.4.2. HAMILTONIEN

Soit un lagrangien $L(q_i, \dot{q}_i)$ que nous traiterons comme la fonction f ci-dessus avec les q_i jouant le rôle de u et les $\dot{q}_i$ le rôle de v . A la place de w , nous définissons les moments généralisés également appelés "**moments canoniques**":

$$p_i \equiv \frac{\partial L}{\partial \dot{q}_i} = p_i(q_j, \dot{q}_j) \quad (29.63)$$

avec $i, j = 1, 2, \dots, n$.

Avant de continuer voyons ce que nous permet de faire cette définition:

Nous définissons donc, en analogie avec g , une fonction des q_i et des p_i que nous noterons $H(q_i, p_i)$:

$$\begin{array}{ccc}
 g(u, w) & = & vw - f \\
 \Downarrow & & \Downarrow \\
 H(q_i, p_i) & = & \sum_{i=1}^n \dot{q}_i p_i - L(q_i, \dot{q}_i)
 \end{array} \quad (29.64)$$

Attention! La relation obtenue:

$$H(q_i, p_i) = \sum_{i=1}^n \dot{q}_i p_i - L(q_i, \dot{q}_i) \quad (29.65)$$

appelée "**fonction de Hamilton**" ou "**Hamiltonien**" est plus qu'importante (comme tout le reste d'ailleurs). Nous la retrouverons, entre autres, en physique quantique relativiste ou encore en physique quantique des champs. Par ailleurs, un très joli exemple de tout ce que nous avons vu maintenant est donné dans le chapitre de Relativité Restreinte où nous calculons le lagrangien et hamiltonien d'une particule libre. Les résultats sont assez pertinents et leur utilité et justesse en électrodynamique plus que étonnante.

Exemple:

Une autre application importante et très connue de la mécanique analytique est le calcul des surfaces minimales (physique et architecture). Si nous nous intéressons à la détermination d'une telle surface en imposant qu'elle soit une surface de révolution, nous allons voir que nous trouvons une caténoïde (soit la forme que prend un film de savon entre deux anneaux).

Nous nous donnons les rayons R_1 et R_2 de deux cercles et l'écartement l entre les deux cercles. Nous cherchons une fonction y de classe C^1 telle que:

$$y(0) = R_1 \text{ et } y(l) = R_2 \quad (29.66)$$

et que la surface de révolution sous forme paramétrique:

$$\{(x, y(x) \cos \phi, y(x) \sin \phi), x \in [0, l], \phi \in [0, 2\pi]\} \quad (29.67)$$

possède une surface minimale.

Nous savons que la surface d'un volume de révolution peut s'écrire (cf. chapitre Formes Géométriques):

$$S = \int 2\pi f(x) \sqrt{1 + f'(x)^2} dx = 2\pi \int y \sqrt{1 + y'^2} dx = 2\pi \int L(y, y') dx \quad (29.68)$$

Soit en faisant varier la fonction:

$$\delta S = \int \delta L(y, y') dx = \int \left(\frac{\partial L}{\partial y} \delta y + \frac{\partial L}{\partial y'} \delta y' \right) dx \quad (29.69)$$

Puisque $\delta(y') = (\delta y)'$ l'intégration par parties du deuxième terme donne:

$$\int \frac{\partial L}{\partial y'} \delta y' dx = \frac{\partial L}{\partial y'} \delta y - \int \left(\frac{\partial L}{\partial y'} \right)' \delta y dx \quad (29.70)$$

Comme les bornes d'intégration sont fixes, le premier terme sera nul. Il reste alors:

$$\int \frac{\partial L}{\partial y'} \delta y' dx = - \int \left(\frac{\partial L}{\partial y'} \right)' \delta y dx \quad (29.71)$$

et donc:

$$\delta S = \int \delta L(y, y') dx = \int \left(\frac{\partial L}{\partial y} \delta y - \left(\frac{\partial L}{\partial y'} \right)' \delta y \right) dx = \int \left(\frac{\partial L}{\partial y} - \left(\frac{\partial L}{\partial y'} \right)' \right) \delta y dx \quad (29.72)$$

Le minimum cherché correspond à $\delta S = 0$ quel que soit δy ce qui impose la condition:

$$\left(\frac{\partial L}{\partial y'} \right)' - \frac{\partial L}{\partial y} = 0 \quad (29.73)$$

nous retrouvons l'équation d'Euler-Lagrange.

Cette équation peut aussi s'écrire sous une autre forme. En introduisant le moment canonique pour simplifier:

$$p = \frac{\partial L}{\partial y'} \quad (29.74)$$

Nous avons alors immédiatement:

$$\left(\frac{\partial L}{\partial y'} \right)' - \frac{\partial L}{\partial y} = 0 \Rightarrow p' - \frac{\partial L}{\partial y} = 0 \Rightarrow p' = \frac{\partial L}{\partial y} \quad (29.75)$$

Nous obtenons alors:

$$dL = \frac{\partial L}{\partial y} dy + \frac{\partial L}{\partial y'} dy' = p' dy + p dy' \quad (29.76)$$

Ainsi, en posant l'analogie vue plus haut (méthode de Hamilton):

$$H(y, p) = py' - L \quad (29.77)$$

nous aboutissons à:

$$dH = d(y'p - L) = d\left(y' \frac{\partial L}{\partial y'} - L\right) = 0 \quad (29.78)$$

Ainsi en se rappelant qu'au début nous avions:

$$S = 2\pi \int y \sqrt{1 + y'^2} dx = 2\pi \int L(y, y') dx \quad (29.79)$$

Nous aboutissons à:

$$y' \frac{\partial L}{\partial y'} - L = \frac{yy'^2}{\sqrt{1+y'^2}} - y\sqrt{1+y'^2} = \frac{-y}{\sqrt{1+y'^2}} = c^{te} \quad (29.80)$$

Ce que nous pouvons aussi noter (car la constante a un signe indéterminé):

$$\frac{-y}{\sqrt{1+y'^2}} = -c^{te} \quad (29.81)$$

Nous avons alors:

$$\frac{y}{\sqrt{1+y'^2}} = c^{te} \quad (29.82)$$

Nous avons déjà intégré ce type d'équation différentielle en détails dans le chapitre de Génie Civil dans l'étude de la chaînette. Le résultat est:

$$y = \frac{1}{c^{te}} \cosh\left(\frac{x}{c^{te}} + c_1^{te}\right) + c_2^{te} \quad (29.83)$$

la surface de révolution de cette courbe étant une caténoïde:

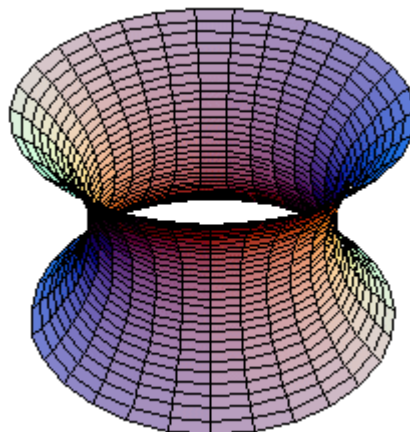


Figure: 29.3 - Tracé de la caténoïde (Source: Wikipedia)

Ce qui est un exemple remarquable qui montre l'intime relation entre la mathématique et la physique!

Cette figure peut être obtenue avec Maple comme suit:

```
>y:=cosh(x);
>plot3d([x,y*cos(phi),y*sin(phi)],x=0..2,phi=0..2*Pi)
```

Maintenant, si L dépend du temps (ce qui est quand même assez souvent le cas...) nous avons comme différentielle totale:

$$\begin{aligned} dL &= \sum_{i=1}^n \frac{\partial L}{\partial q_i} dq_i + \sum_{i=1}^n \frac{\partial L}{\partial \dot{q}_i} d\dot{q}_i + \frac{\partial L}{\partial t} dt \\ &= \sum_{i=1}^n \frac{\partial L}{\partial q_i} dq_i + \sum_{i=1}^n p_i d\dot{q}_i + \frac{\partial L}{\partial t} dt \end{aligned} \quad (29.84)$$

nous calculons aussi la différentielle totale de $H(q_i, p_i)$ et y substituons le résultat obtenu précédemment:

$$\begin{aligned} dH &= \sum_{i=1}^n \dot{q}_i dp_i + \sum_{i=1}^n p_i d\dot{q}_i - \left[\sum_{i=1}^n \frac{\partial L}{\partial q_i} dq_i + \sum_{i=1}^n p_i d\dot{q}_i + \frac{\partial L}{\partial t} dt \right] \\ &= \sum_{i=1}^n \dot{q}_i dp_i - \sum_{i=1}^n \frac{\partial L}{\partial q_i} dq_i - \frac{\partial L}{\partial t} dt \end{aligned} \quad (29.85)$$

ce qui montre bien que $H(q_i, p_i)$ est fonction des q_i (et du temps).

Nous pouvons donc aussi écrire pour sa différentielle totale:

$$dH = \sum_{i=1}^n \frac{\partial H}{\partial p_i} dp_i + \sum_{i=1}^n \frac{\partial H}{\partial q_i} dq_i + \frac{\partial H}{\partial t} dt \quad (29.86)$$

et comme les q_i et p_i sont indépendants nous identifions, en comparant nos deux expressions que:

$$\left. \begin{aligned} \frac{\partial H}{\partial q_i} &= - \frac{\partial L}{\partial q_i} \\ \frac{\partial H}{\partial p_i} &= \dot{q}_i \\ \frac{\partial H}{\partial t} &= - \frac{\partial L}{\partial t} \end{aligned} \right\} \quad (29.87)$$

Ces relations sont extrêmement importantes car nous les retrouverons en magnétostatique, en physique quantique relativiste et aussi en physique quantique des champs sous une forme un peu plus barbare (mais magnifique aussi...).

Considérons maintenant le deuxième terme du premier membre de l'équation d'Euler-Lagrange. Nous avons:

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_i} \right) = \frac{d}{dt} (p_i) = \frac{\partial L}{\partial q_i} = \dot{p}_i \quad (29.88)$$

et ainsi, nous obtenons les $2n$ équations ci-dessous:

$$\left. \begin{aligned} \dot{q}_i &= \frac{\partial H}{\partial p_i}; i = 1, 2, \dots, n \\ \dot{p}_i &= - \frac{\partial H}{\partial q_i}; i = 1, 2, \dots, n \end{aligned} \right\} 2n \quad (29.89)$$

Ces $2n$ équations sont appelées "**équations canoniques du mouvement**" et sont des équations différentielles du premier ordre.

Remarque: L'apparition du signe moins " - " entre les équations pour les q_i et celles pour leurs moments conjugués, s'appelle une "**symétrie symplectique**".

De:

$$dH = \sum_{i=1}^n \frac{\partial H}{\partial p_i} dp_i + \sum_{i=1}^n \frac{\partial H}{\partial q_i} dq_i + \frac{\partial H}{\partial t} dt \quad (29.90)$$

nous pouvons, sur une trajectoire qui obéit aux équations canoniques, calculer:

$$\begin{aligned} \frac{dH}{dt} &= \sum_{i=1}^n \frac{\partial H}{\partial p_i} \dot{p}_i + \sum_{i=1}^n \frac{\partial H}{\partial q_i} \dot{q}_i + \frac{\partial H}{\partial t} \\ &= - \sum_{i=1}^n \dot{p}_i \dot{q}_i + \sum_{i=1}^n \dot{q}_i \dot{p}_i + \frac{\partial H}{\partial t} = \frac{\partial H}{\partial t} = - \frac{\partial L}{\partial t} \end{aligned} \quad (29.91)$$

Remarque: Si H ne dépend pas du temps nous avons alors $\partial H / \partial t = 0$, alors H (ainsi que L), sont une "**constante du mouvement**".

Un exemple s'avère indispensable à ce niveau d'avancement de l'étude du formalisme Lagrangien. Nous allons nous restreindre à un cas particulier d'une particule soumise à une force en une dimension. Mais bien que cet exemple et les développements qui y sont liés soient simples nous retrouverons les résultats obtenus ici dans bien d'autres parties du site. Il est donc important de bien l'étudier et de bien le comprendre (ce qui nécessite malheureusement aussi que le contenu du chapitre de Mécanique Classique soit connu par le lecteur).

Exemple:

Soit une particule de masse m se déplaçant en une dimension (disons x) et soumise à une force dérivant d'un potentiel tel que:

$$F_x = - \frac{\partial V}{\partial x} \Leftrightarrow \ddot{x} = - \frac{1}{m} \frac{\partial V}{\partial x} = - (\nabla V)_x \quad (29.92)$$

Nous savons que son lagrangien est:

$$L = \frac{1}{2} m \dot{x}^2 - V(x) \quad (29.93)$$

Nous n'aurons qu'un seul moment (la quantité de mouvement), noté p , conjugué à x et défini par:

$$p = \frac{\partial L}{\partial \dot{x}} \quad (29.94)$$

équation que nous pouvons (que nous devons !) inverser (de la définition de la quantité de mouvement):

$$\dot{x} = \frac{p}{m} \quad (29.95)$$

Nous pouvons noter en ce point que le moment p correspond (ô hasard !!) à la composante x de la définition élémentaire $p = mv$ (ce qui ne sera pas toujours aussi trivialement le cas).

Selon la définition de l'Hamiltonien il vient alors:

$$H = \dot{x}(p)p - L(x, \dot{x}(p)) = \frac{p}{m}p - \frac{m}{2} \left(\frac{p}{m} \right)^2 + V(x) = \frac{p^2}{2m} + V(x) \quad (29.96)$$

que nous écrivons souvent sous la forme:

$$H = T + V \quad (29.97)$$

où T est donc l'énergie cinétique exprimée en fonction des moments.

1.4.3. CROCHETS DE POISSON

Le crochet de Poisson $\{A, B\}_{q,p}$ est la façon standard de noter une certaine opération qui implique les quantités $A(q_i, p_i)$ et $B(q_i, p_i)$ ainsi que l'ensemble des variables canoniques (q_i, p_i) définie par:

$$\{A, B\}_{q,p} \equiv \sum_{i=1}^n \left[\frac{\partial A}{\partial q_i} \frac{\partial B}{\partial p_i} - \frac{\partial A}{\partial p_i} \frac{\partial B}{\partial q_i} \right] \quad (29.98)$$

qui exprime la manière de parcourir un champ (le crochet étant nul si les deux types de parcours sont égaux).

de cette définition nous pouvons déduire certaines propriétés relativement triviales:

$$P1. \{A, B\} = -\{B, A\}$$

Démonstration:

$$\begin{aligned} \{A, B\} &= \sum_{i=1}^n \left[\frac{\partial A}{\partial q_i} \frac{\partial B}{\partial p_i} - \frac{\partial A}{\partial p_i} \frac{\partial B}{\partial q_i} \right] = \sum_{i=1}^n \left[-\frac{\partial B}{\partial q_i} \frac{\partial A}{\partial p_i} + \frac{\partial B}{\partial p_i} \frac{\partial A}{\partial q_i} \right] \\ &= -\sum_{i=1}^n \left[\frac{\partial B}{\partial p_i} \frac{\partial A}{\partial q_i} - \frac{\partial B}{\partial q_i} \frac{\partial A}{\partial p_i} \right] \\ &= -\{B, A\} \end{aligned} \quad (29.99)$$

□C.Q.F.D.

$$P2. \{A, B + C\} = \{A, B\} + \{A, C\}$$

Démonstration:

$$\begin{aligned}
\{A, B+C\} &= \sum_{i=1}^n \left[\frac{\partial A}{\partial q_i} \frac{\partial (B+C)}{\partial p_i} - \frac{\partial A}{\partial p_i} \frac{\partial (B+C)}{\partial q_i} \right] \\
&= \sum_{i=1}^n \left[\frac{\partial A}{\partial q_i} \left(\frac{\partial B}{\partial p_i} + \frac{\partial C}{\partial p_i} \right) - \frac{\partial A}{\partial p_i} \left(\frac{\partial B}{\partial q_i} + \frac{\partial C}{\partial q_i} \right) \right] \\
&= \sum_{i=1}^n \left[\left(\frac{\partial A}{\partial q_i} \frac{\partial B}{\partial p_i} - \frac{\partial A}{\partial p_i} \frac{\partial B}{\partial q_i} \right) + \left(\frac{\partial A}{\partial q_i} \frac{\partial C}{\partial p_i} - \frac{\partial A}{\partial p_i} \frac{\partial C}{\partial q_i} \right) \right] \quad (29.100) \\
&= \sum_{i=1}^n \left[\frac{\partial A}{\partial q_i} \frac{\partial B}{\partial p_i} - \frac{\partial A}{\partial p_i} \frac{\partial B}{\partial q_i} \right] + \sum_{i=1}^n \left[\frac{\partial A}{\partial q_i} \frac{\partial C}{\partial p_i} - \frac{\partial A}{\partial p_i} \frac{\partial C}{\partial q_i} \right] \\
&= \{A, B\} + \{A, C\}
\end{aligned}$$

□C.Q.F.D.

$$P3. \{A, BC\} = B\{A, C\} + C\{A, B\}$$

Démonstration:

$$\begin{aligned}
\{A, BC\} &= \sum_{i=1}^n \left[\frac{\partial A}{\partial q_i} \frac{\partial (B \cdot C)}{\partial p_i} - \frac{\partial A}{\partial p_i} \frac{\partial (B \cdot C)}{\partial q_i} \right] \\
&= \sum_{i=1}^n \left[\frac{\partial A}{\partial q_i} \left(B \frac{\partial C}{\partial p_i} + C \frac{\partial B}{\partial p_i} \right) - \frac{\partial A}{\partial p_i} \left(B \frac{\partial C}{\partial q_i} + C \frac{\partial B}{\partial q_i} \right) \right] \\
&= \sum_{i=1}^n \left[B \left(\frac{\partial A}{\partial q_i} \frac{\partial C}{\partial p_i} - \frac{\partial A}{\partial p_i} \frac{\partial C}{\partial q_i} \right) + C \left(\frac{\partial A}{\partial q_i} \frac{\partial B}{\partial p_i} - \frac{\partial A}{\partial p_i} \frac{\partial B}{\partial q_i} \right) \right] \quad (29.101) \\
&= B \sum_{i=1}^n \left[\frac{\partial A}{\partial q_i} \frac{\partial C}{\partial p_i} - \frac{\partial A}{\partial p_i} \frac{\partial C}{\partial q_i} \right] + C \sum_{i=1}^n \left[\frac{\partial A}{\partial q_i} \frac{\partial B}{\partial p_i} - \frac{\partial A}{\partial p_i} \frac{\partial B}{\partial q_i} \right] \\
&= B\{A, C\} + C\{A, B\}
\end{aligned}$$

□C.Q.F.D.

$$P4. \{A, \{B, C\}\} + \{C, \{A, B\}\} + \{B, \{C, A\}\} = 0$$

Démonstration (nous allons simplifier la notation pour condenser...):

$$\begin{aligned}
&\{A, \{B, C\}\} + \{C, \{A, B\}\} + \{B, \{C, A\}\} \\
&= \sum_{i=1}^n \left[\partial_q A \partial_p \sum_{i=1}^n [\partial_q B \partial_p C - \partial_p B \partial_q C] - \partial_p A \partial_q \left(\sum_{i=1}^n [\partial_q B \partial_p C - \partial_p B \partial_q C] \right) \right] \\
&+ \sum_{i=1}^n \left[\partial_q C \partial_p \sum_{i=1}^n [\partial_q A \partial_p B - \partial_p A \partial_q B] - \partial_p C \partial_q \left(\sum_{i=1}^n [\partial_q A \partial_p B - \partial_p A \partial_q B] \right) \right] \quad (29.102) \\
&+ \sum_{i=1}^n \left[\partial_q B \partial_p \sum_{i=1}^n [\partial_q C \partial_p A - \partial_p C \partial_q A] - \partial_p B \partial_q \left(\sum_{i=1}^n [\partial_q C \partial_p A - \partial_p C \partial_q A] \right) \right]
\end{aligned}$$

Bon et ici, histoire de pas avoir un truc illisible, long et ennuyeux on va démontrer la propriété pour

$n = 1$ et nous supposons (bien évidemment) qu'elle est valable pour tout n :

$$\begin{aligned}
 & \{A, \{B, C\}\} + \{C, \{A, B\}\} + \{B, \{C, A\}\} \\
 &= \partial_q A \partial_p [\partial_q B \partial_p C - \partial_p B \partial_q C] - \partial_p A \partial_q [\partial_q B \partial_p C - \partial_p B \partial_q C] \\
 &+ \partial_q C \partial_p [\partial_q A \partial_p B - \partial_p A \partial_q B] - \partial_p C \partial_q [\partial_q A \partial_p B - \partial_p A \partial_q B] \\
 &+ \partial_q B \partial_p [\partial_q C \partial_p A - \partial_p C \partial_q A] - \partial_p B \partial_q [\partial_q C \partial_p A - \partial_p C \partial_q A] \\
 &= \partial_q A \partial_{p,q} B \partial_p C + \partial_q A \partial_q B \partial_{p,p} C - \partial_q A \partial_{p,p} B \partial_q C - \partial_q A \partial_p B \partial_{p,q} C \\
 &- \partial_p A \partial_{q,q} B \partial_p C - \partial_p A \partial_q B \partial_{q,p} C + \partial_p A \partial_{q,p} B \partial_q C + \partial_p A \partial_p B \partial_{q,q} C \\
 &+ \partial_q C \partial_{p,q} A \partial_p B + \partial_q C \partial_q A \partial_{p,p} B - \partial_q C \partial_{p,p} A \partial_q B - \partial_q C \partial_p A \partial_{q,q} B \\
 &- \partial_p C \partial_{q,q} A \partial_p B - \partial_p C \partial_q A \partial_{q,p} B + \partial_p C \partial_{q,p} A \partial_q B + \partial_p C \partial_p A \partial_{q,q} B \\
 &+ \partial_q B \partial_{p,q} C \partial_p A + \partial_q B \partial_q C \partial_{p,p} A - \partial_q B \partial_{p,p} C \partial_q A - \partial_q B \partial_p C \partial_{p,q} A \\
 &- \partial_p B \partial_{q,q} C \partial_p A - \partial_p B \partial_q C \partial_{q,p} A + \partial_p B \partial_{q,p} C \partial_q A + \partial_p B \partial_p C \partial_{q,q} A
 \end{aligned} \quad (29.103)$$

Nous avons en plus (cf. chapitre de Calcul Différentiel Et Intégral) sous certaines conditions la propriété $\partial_{p,q} = \partial_{q,p}$. Dès lors l'ensemble des termes s'annulent (c'est de l'algèbre élémentaire) pour avoir finalement :

$$\{A, \{B, C\}\} + \{C, \{A, B\}\} + \{B, \{C, A\}\} = 0 \quad (29.104)$$

où la dernière expression est appelée "**identité de Jacobi**".

□C.Q.F.D.

Au-delà d'une simple notation, le calcul des crochets de Poisson est assez facile et permet d'obtenir nombre de résultats intéressants. D'autre part, ils sont intimement reliés aux "commutateurs" de la physique quantique que nous étudierons dans le détail dans le chapitre concerné.

Considérons maintenant une fonction quelconque $F(q_i, p_i, t)$ dont la dérivée totale par rapport au temps le long d'une trajectoire s'écrit (vous y reconnaîtrez normalement quelque chose que vous connaissez déjà...):

$$\frac{dF}{dt} = \sum_{i=1}^n \left[\frac{\partial F}{\partial q_i} \dot{q}_i + \frac{\partial F}{\partial p_i} \dot{p}_i \right] + \frac{\partial F}{\partial t} \quad (29.105)$$

Si cette trajectoire est une trajectoire physique, elle obéit aux équations canoniques de l'hamiltonien H du système:

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \dot{p}_i = -\frac{\partial H}{\partial q_i} \quad (29.106)$$

et alors:

$$\frac{dF}{dt} = \sum_{i=1}^n \left[\frac{\partial F}{\partial q_i} \frac{\partial H}{\partial p_i} - \frac{\partial F}{\partial p_i} \frac{\partial H}{\partial q_i} \right] + \frac{\partial F}{\partial t} = \{F, H\} + \frac{\partial F}{\partial t} \quad (29.107)$$

En particulier, cette équation permet un calcul facile des constantes du mouvement. En effet, le calcul de $\partial F / \partial t$ est immédiat et le calcul de $\{F, H\}$ un exercice assez simple.

Il existe une famille de résultats intéressants des crochets de Poisson. Parmi les plus importants, calculons certains de ces crochets entre des variables canoniques, coordonnées et moments:

$$\{q_k, q_j\} = \sum_{i=1}^n \left[\frac{\partial q_k}{\partial q_i} \frac{\partial q_j}{\partial p_i} - \frac{\partial q_k}{\partial p_i} \frac{\partial q_j}{\partial q_i} \right] \quad (29.108)$$

puisque par définition, les coordonnées et moments ne sont pas directement dépendants:

$$\frac{\partial q_j}{\partial p_i} = 0, \frac{\partial q_k}{\partial p_i} = 0 \quad (29.109)$$

d'où:

$$\{q_k, q_j\} = \{q_j, q_k\} = 0 \quad (29.110)$$

et de manière identique:

$$\{p_k, p_j\} = \{p_j, p_k\} = 0 \quad (29.111)$$

Mais:

$$\{q_k, p_j\} = \sum_{i=1}^n \left[\frac{\partial q_k}{\partial q_i} \frac{\partial p_j}{\partial p_i} - \frac{\partial q_k}{\partial p_i} \frac{\partial p_j}{\partial q_i} \right] = \sum_{i=1}^n \frac{\partial q_k}{\partial q_i} \frac{\partial p_j}{\partial p_i} = \sum_{i=1}^n \delta_{ki} \delta_{ji} = \delta_{kj} \quad (29.112)$$

où rappelons-le, δ_{ij} est le symbole de Kronecker défini par:

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad (29.113)$$

Attention ! $\{q_k, p_j\}$ n'est pas commutatif. Effectivement, le lecteur contrôlera facilement que:

$$\{q_k, p_j\} \neq \{p_j, q_k\} \quad (29.114)$$

Ce qui implique un résultat assez général que nous retrouverons dans le chapitre de Physique Quantique Ondulatoire :

$$\{q_k, p_j\} - \{p_j, q_k\} \neq 0 \quad (29.115)$$

1.4.4. TRANSFORMATIONS CANONIQUES

Nous disons des q_i, p_i que ce sont des "**variables canoniques généralisées**". Ce n'est pas un euphémisme puisqu'il n'y a pratiquement aucune limite à ce qu'elles peuvent représenter physiquement.

Puisque tel est le cas, il doit exister des transformations entre ces différents choix. Nous noterons Q_i, P_i les nouvelles variables canoniques obtenues suite à une telle transformation.

Nous ne sommes pas surpris par contre de constater que ces transformations sont soumises à des conditions assez sévères. En effet, les q_i, p_i sont généralisés et obéissent à:

$$\{q_k, q_j\} = 0 \quad \{p_k, p_j\} = 0 \quad \{q_k, p_j\} = \delta_{kj} \quad (29.116)$$

et les équations canoniques:

$$\dot{q}_i = \frac{\partial H}{\partial p_i} \quad \dot{p}_i = -\frac{\partial H}{\partial q_i} \quad (29.117)$$

sont invariantes de forme. Ainsi, à la suite d'une transformation des q_i, p_i vers les Q_i, P_i et définissant un nouvel hamiltonien que nous noterons $K(Q_i, P_i)$ nous devons avoir:

$$\{Q_k, Q_j\} = 0 \quad \{P_k, P_j\} = 0 \quad \{Q_k, P_j\} = \delta_{kj} \quad (29.118)$$

et les équations canoniques:

$$\dot{Q}_i = \frac{\partial K}{\partial P_i} \quad \dot{P}_i = -\frac{\partial K}{\partial Q_i} \quad (29.119)$$

Strictement, les équations de transformation peuvent s'écrire:

$$\begin{aligned} Q_i &= Q_i(q_j, p_j, t) \\ P_i &= P_i(q_j, p_j, t) \end{aligned} \quad (29.120)$$

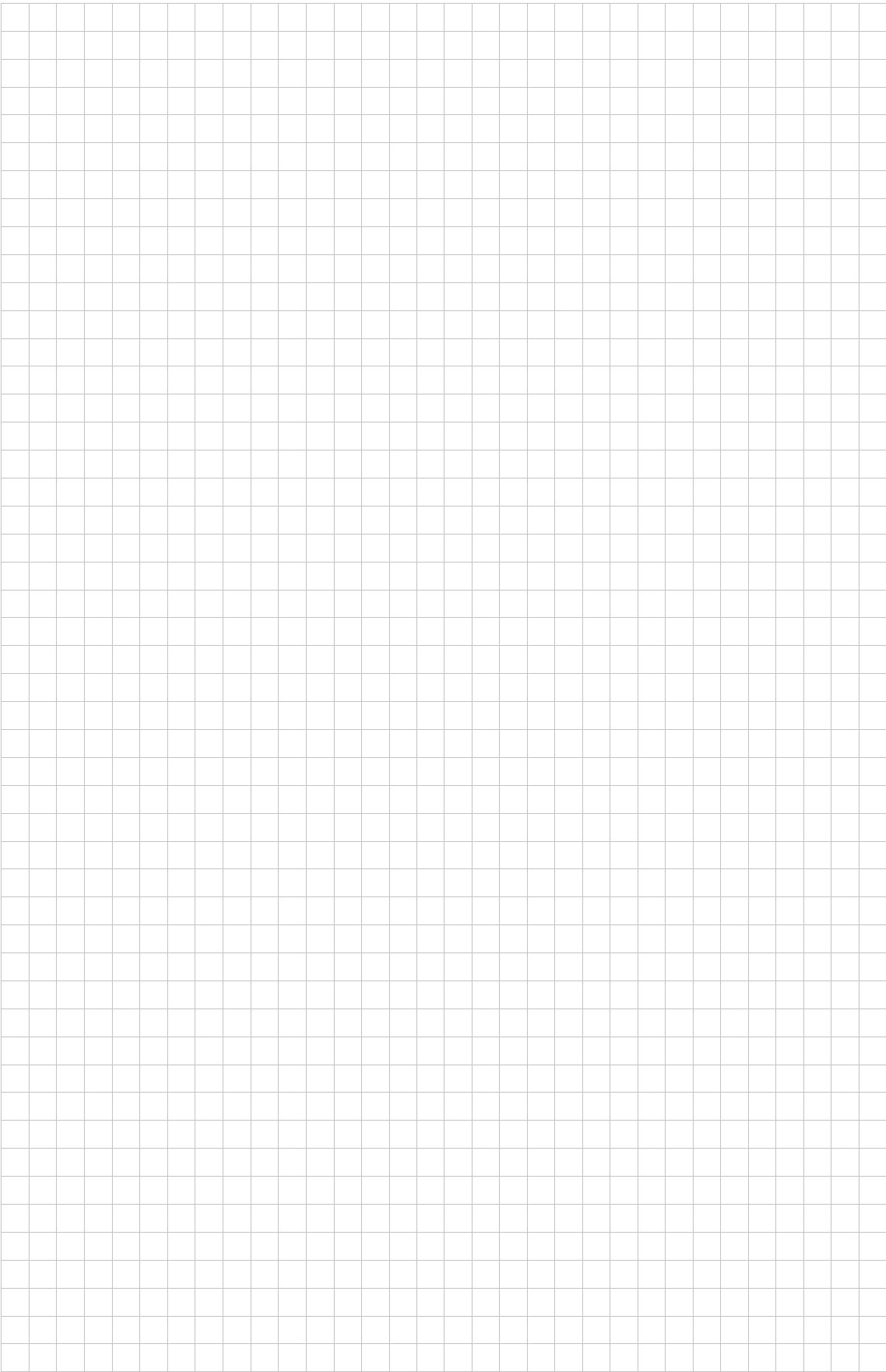
avec $i, j = 1, 2, \dots, n$ et doivent pouvoir s'inverser puisque la physique reste indépendante des variables que nous employons pour la décrire, donc nous pouvons écrire les transformations inverses:

$$\begin{aligned} q_i &= q_i(Q_j, P_j, t) \\ p_i &= p_i(Q_j, P_j, t) \end{aligned} \quad (29.121)$$

avec $i, j = 1, 2, \dots, n$. Les q_i, p_i, Q_i, P_i forment $4n$ variables mais il est évident que seules $2n$ d'entre elles sont indépendantes.

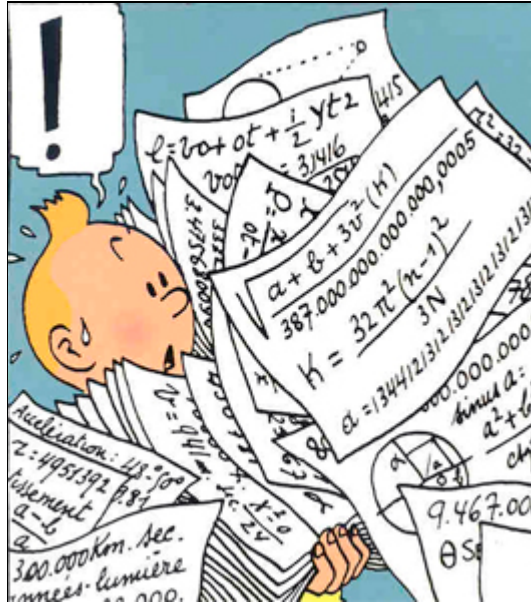






30. MÉCANIQUE CLASSIQUE/RATIONNELLE

Avant d'aborder l'étude des corps solides en mouvement dans le cadre de la mécanique classique (à l'opposé de la mécanique relativiste) appelée également "**mécanique rationnelle**" ou "**mécanique newtonienne**", il peut sembler être dans l'ordre logique des choses de définir et d'étudier les propriétés relativement à leur état statique.

**Définitions:**

D1. Un phénomène est dit "**statique**" ou "**en équilibre**" lorsqu'il ne subit aucune dynamique (accélération ou in extenso: force), du moins apparente. Nous pouvons considérer un équilibre comme un état statique, bien qu'il ne soit qu'apparent car il peut être le résultat de deux dynamiques opposées qui se compensent! Ainsi, les grandeurs qui décrivent un phénomène statique sont des constantes, les valeurs concrètes de ces grandeurs sont calculables. La statique est un cas majeur d'étude du génie mécanique et du génie civil (voir chapitres du même nom).

De manière plus technique cette définition est érigée au rang de principe appelé le "**principe fondamental de la statique**" qui énonce que pour qu'un système soit en équilibre, il faut que la résultante générale et le moment résultant des forces extérieures soit équivalent à zéro par rapport à son centre de masse ou de gravité (la condition est suffisante pour les problèmes de mécanique qui traitent des solides indéformables).

D2. La "**statique**" est l'étude des conditions d'équilibre d'un point matériel soumis à des forces en équilibre.

D3. Toute cause capable d'accélérer (concept défini plus loin) ou de déformer un corps est appelée "**force**" (concept introduit rigoureusement par Newton et sur lequel nous reviendrons en détails plus loin lors de l'énoncé des trois lois de Newton).

Une observation plus approfondie fait apparaître la force comme le résultat macroscopique de phénomènes microscopiques complexes, à savoir des interactions à distance entre particules. Ces

interactions sont au nombre de quatre et je ne désire nullement en parler maintenant car elles font appel à des outils mathématiques qui sont hors contexte dans cette section du site.

Remarque: En mécanique classique nous ne nous posons naturellement pas la question d'une transformation du temps. Les changements envisagés concernent la grandeur "position" et ses dérivées. En effet, en mécanique classique, nous postulons le "**temps de Newton**": le temps s'écoule de façon identique d'un référentiel à l'autre.

D4. Si les lignes d'action de toutes les forces agissant sur un corps sont dans un même plan, le système de forces est dit "**système coplanaire**" et sa représentation schématique un "**dyname**":

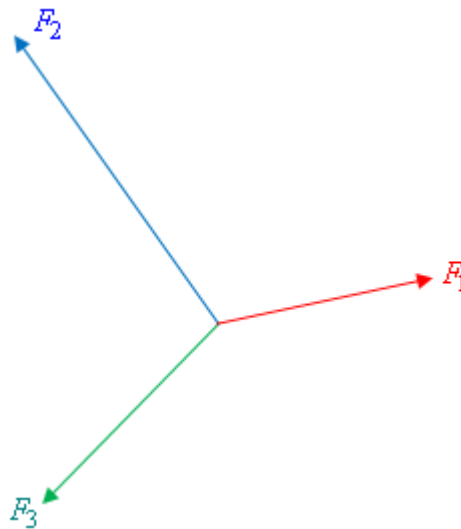


Figure: 30.1 - Exemple particulier de système coplanaire

où l'intersection des forces se nomme un "**noeud de forces**".

Si nous considérons par exemple un ensemble particulier de forces coplanaires dont les intensités (normes) ont été mesurées à l'aide d'un dynamomètre et les angles par rapport à une repère adéquatement choisi avec un rapporteur. Nous aurons alors un schéma du type suivant:

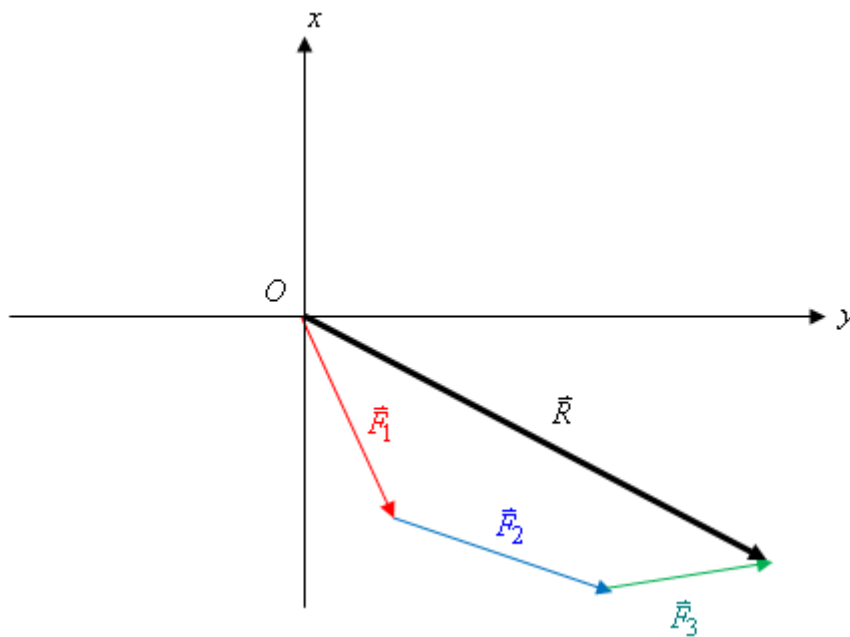


Figure: 30.2 - Exemple particulier de système coplanaire mesuré

Pour calculer les composantes de la résultante (et in extenso de sa norme), il peut être plus aisé de représenter les trois forces sous la forme suivante (après translation):

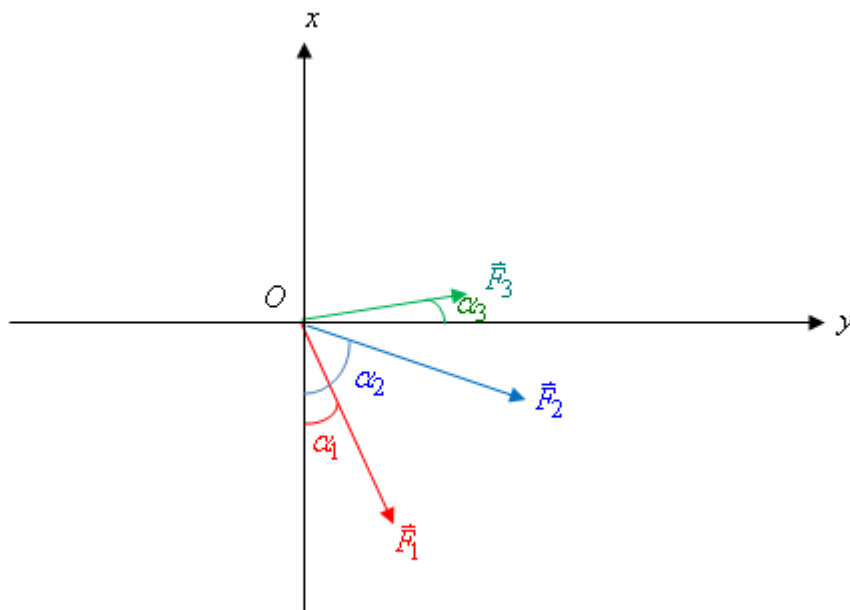


Figure: 30.3 - Représentation simplifiée des forces pour traitement

Ensuite, avec de la trigonométrie élémentaire, connaissant les angles et l'intensité de chacune des trois forces, il est possible de déterminer leurs composantes respectives selon X et Y et leur somme algébrique selon chacun des axes donnera les composantes de la résultante (aujourd'hui ce type de raisonnement à l'air simple mais il a fallu quand même attendre la fin du 16ème siècle pour que ces raisonnements sur la statique des forces émerge).



Figure: 30.4 - Représentation de la résultante des forces d'un pont suspendu

Un petit cas applicatif élémentaire est le portage d'une charge:

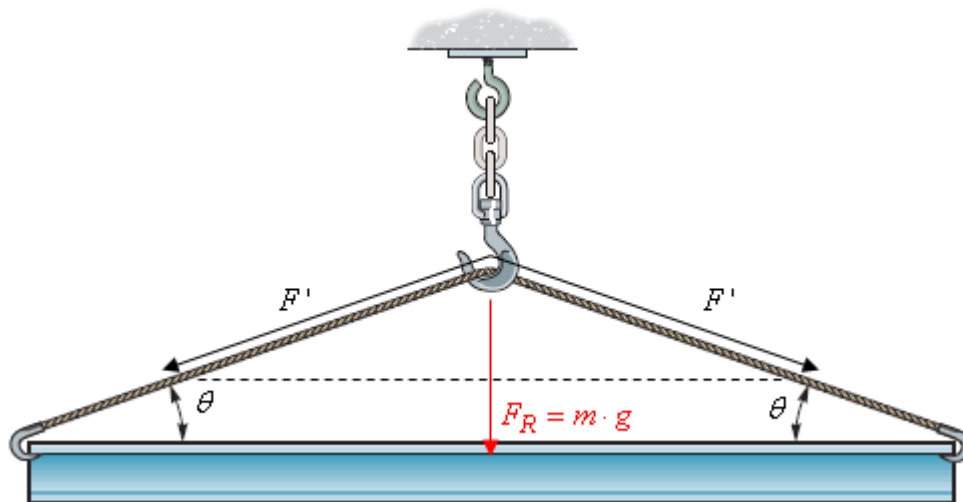


Figure: 30.5 - Port de charge coplanaire symétrique

par symétrie chacune des parties gauche/droite du câble est sous la même tension (force). En appliquant de la trigonométrie élémentaire (cf. chapitre de Trigonométrie), nous avons:

$$\frac{F_R/2}{F'} = \cos\left(\pi - \frac{\pi}{2} - \theta\right) = \cos\left(\frac{\pi}{2} - \theta\right) = \sin \theta \quad (30.1)$$

Soit:

$$F' = \frac{F_R/2}{\sin \theta} \quad (30.2)$$

Nous remarquons alors que nous avons tout intérêt à avoir:

$$\sin \theta > 0.5 \quad (30.3)$$

(soit un angle supérieur à $\sim 30^\circ$) sinon quoi nous aurions:

$$F' > F_R \quad (30.4)$$

Comme nous l'avons déjà mentionné, un système de forces concourantes coplanaires est en équilibre statique, lorsque la résultante de toutes les forces est nulle. Graphiquement, l'extrémité de la dernière force du dynamomètre coïncide avec l'origine de la première force. Mathématiquement cela peut s'écrire:

$$\left. \begin{array}{l} \sum_{i=1}^n F_{xi} = R_x = 0 \\ \sum_{i=1}^n F_{yi} = R_y = 0 \end{array} \right\} \Rightarrow \vec{R} = \vec{0} \quad (30.5)$$

D5. Un système matériel S (ensemble de points matériels $\{P_i, m_i\}$) est dit "**solide indéformable**" (rigide), ou simplement "**solide**", si les distances mutuelles des points matériels le constituant ne varient pas au cours du temps. Ce que notons techniquement sous la forme suivante:

$$\forall t, \forall (i, j) : \|\overrightarrow{P_i P_j}\| = c^{te} \quad (30.6)$$

1. LOIS DE NEWTON

Les trois lois de Newton sont à la base de la mécanique classique. Elles sont à posteriori indémonstrables et non formalisables car elles énoncent des observations et découlent donc de notre expérience quotidienne.

Cependant, les développements de la physique moderne et qui se basent sur les conséquences de ces trois lois sont en tel accord avec les conditions théoriques qu'impose le principe de moindre action et les expériences y relatives, que leur validité pourrait ne plus être mise en doute (...)

1.1. PREMIÈRE LOI (LOI D'INERTIE)

Définition: Tout corps ponctuel ou étendu persévère dans sa forme (géométrie) ou son état de repos ou de mouvement rectiligne uniforme (décrit par le centre de masse), sauf si des "**forces imprimées**" le contraignent d'en changer.

Le corps ponctuel est bien évidemment sans dimension. C'est une création de l'esprit, un modèle, représentant un objet physique qui n'est animé que d'un mouvement de translation (pas de rotation sur lui-même). Nous admettons ici que notre espace physique est à trois dimensions auquel on adjoint le temps qui n'est pas ici une dimension mais un paramètre immuable et indépendant!

Autrement dit: Tout corps au repos ou en mouvement rectiligne uniforme est soit imprimé par un nombre de forces nulles, soit la somme des forces imprimées est nulle (c'est le principe fondamental de la statique appelé aussi "**principe d'inertie**").

Corollaire: Lorsque la trajectoire d'un corps n'est pas une droite ou lorsque la vitesse de ce corps n'est pas constante, on peut en conclure d'après le Principe d'inertie que les forces qui s'exercent sur ce corps ne se compensent pas.

Remarque: Nous avons démontré ce corollaire lors de notre étude du théorème de Noether dans le chapitre traitant des Principes de la physique.

Après la virgule de la première phrase du corollaire, jaillit en pleine lumière le mot "force".
Questionnons donc ce mot: le langage courant regorge de significations différentes: la force du poignet, la force de l'âme... Aussi, la force peut-elle être aveugle ou majeure, selon le cas... Quoi qu'il en soit, elle a le pouvoir de changer le cours (le mouvement) et la forme (géométrie) des choses. Sans ignorer ce halo qui entoure le mot et qui a embarrassé plus d'un physicien avant lui, Newton donne à la force une signification très précise, qui se démarque de l'idée intuitive d'un effort physique.

Propriétés:

P1. La force est une grandeur vectorielle (cf. chapitre de Calcul Vectoriel).

P2. L'effet d'une force, ne change pas si nous faisons glisser la force sur sa droite d'action.

Une force est donc une grandeur physique qui se manifeste par ses effets:

E1. Effet dynamique: une force est une cause capable de produire ou de modifier le mouvement ou la forme (géométrie) d'un corps.

E2. Effet statique: une force est une cause capable de produire une déformation d'un corps.

Toute force peut être représentée par un vecteur dont les quatre propriétés sont:

P1. Direction: droite selon laquelle l'action s'exerce

P2. Sens: sens selon lequel l'action s'exerce sur la droite

P3. Point d'application: point où l'action s'exerce sur le corps

P4. Intensité: la valeur (norme) de la force

Il est possible de ranger la plupart des forces par familles telles que:

F1. Les "**forces de réaction**": chaque corps exerce une force sur un autre corps qui est en contact avec lui. Par exemple, si un objet repose sur une table, cette table exerce une force égale et opposée sur l'objet (afin que ce dernier ne s'enfonce pas dans la table - ce sont des mécanismes quantiques qui sont à l'origine de cette force de réaction). Cette force est toujours à la verticale du point de contact.

F2. Les "**forces de frottement**": la force de frottement existe lorsque deux corps sont en contact. Elle s'oppose toujours au mouvement. La force de frottement qui s'oppose au mouvement n'a pas seulement un effet négatif, elle est indispensable pour assurer aussi le contact entre deux surfaces (par exemple: contact des pneus sur la route, freinage, ...).

F3. Les "**forces de tension**" exercées sur un corps: c'est une force qui tire sur un élément d'un corps comme par exemple, la tension exercée par un fil, par un ressort (cf. chapitre de Génie Mécanique).

F4. Les "**forces à distance**": ce sont les forces qui agissent par l'intermédiaire de champs vectoriels comme par exemple le champ électrique, le champ magnétique, le champ gravitationnel. Ce dernier a

comme particularité s'il est isotrope (nous le démontrerons lors de notre étude de la statique des forces) de pouvoir se réduire à l'étude du centre de gravité du corps.

1.2. DEUXIÈME LOI (PRINCIPE FONDAMENTAL DE LA DYNAMIQUE)

Définition: Le changement de mouvement est proportionnel à la "**force motrice imprimée**", et s'effectue suivant la droite par laquelle cette force est imprimée.

Une force, nous le savons, est dans le langage de Newton ce qui provoque le "changement du mouvement" et pas autre chose... Mais, supplément au programme, les mots "changement du mouvement" de cette loi cachent une signification mathématique, différente de l'intuition "changement de vitesse". Pour Newton, nous avons vu qu'un corps au repos était caractérisé par sa quantité de matière, sa masse. S'inspirant de certains prédécesseurs, Newton pose qu'un corps en mouvement "transporte une certaine quantité", appelée sans fioritures: la "**quantité de mouvement**". C'est en fait cette quantité qui, sous le simple mot "**mouvement**" est contenue dans l'énoncé de la seconde loi. La quantité d'un mouvement est la mesure que nous tirons à la fois de sa vitesse (concept que nous définirons plus loin lors de notre étude de la cinématique) et de sa quantité de matière, autrement dit, par définition, le produit de sa masse par sa vitesse.

$$\vec{p} = m \cdot \vec{v} \quad (30.7)$$

En utilisant les symboles mathématiques modernes, la première partie de cette deuxième loi peut alors se reformuler:

La force est égale à la variation en fonction du temps de la quantité de mouvement, soit dans un cadre non relativiste:

$$\vec{F} = \frac{d\vec{p}}{dt} = \frac{d(m \cdot \vec{v})}{dt} \cong m \cdot \frac{d\vec{v}}{dt} = m \cdot \vec{a} \quad (30.8)$$

Cette relation est donc valable tant que la vitesse est très inférieure à celle de la lumière comme nous le verrons lors de notre étude de la mécanique relativiste bien plus tard, car Newton supposa que la masse ne variait pas (ou ne semblait pas varier...) en fonction de la vitesse. Ainsi, la "**relation fondamentale de la dynamique**" (R.F.D.) est donnée par:

$$\vec{F} = m\vec{a} \quad (30.9)$$

et peut s'énoncer ainsi: Soit un corps de masse m constante, l'accélération subie par un corps dans un référentiel galiléen est proportionnelle à la résultante des forces qu'il subit, et inversement proportionnelle à sa masse m .

Rappel: La "**masse**" est une mesure pour la quantité de matière contenue dans le corps (cf. chapitre sur les Principes De La Mécanique). La masse est une constante indépendante de l'endroit où elle se trouve (unité S.I. kilogramme: [kg]). Le "**poids**", correspond lui à la force (unité S.I. newton: N) qu'un objet exerce sur un autre par l'intermédiaire d'un champ gravitationnel. Il dépend de l'endroit où nous nous trouvons (voir ci-dessous l'équation de la force gravitationnelle de Newton).

Nous verrons (démontrerons) que dans le cadre d'un corps tombant dans un champ gravitationnel à symétrie sphérique, nous avons:

$$\vec{F} = m \cdot \vec{a} = m \cdot \left(\frac{G \cdot M}{R^2} \cdot \frac{\vec{R}}{\|\vec{R}\|} \right) \quad (30.10)$$

dans le cadre de notre bonne vieille Terre, nous avons pour habitude de poser:

$$\|\vec{a}\| = g \quad (30.11)$$

Dans le système Eulérien et en coordonnées cartésiennes, une grandeur donnée φ d'un milieu continu aura une distribution en fonction des quatre variables indépendantes x, y, z, t . Pour de petites variations dx, dy, dz et dt , la variation totale de φ s'exprimant par (cf. chapitre de Calcul Différentiel Et Intégral):

$$d\varphi = \frac{\partial \varphi}{\partial x} dx + \frac{\partial \varphi}{\partial y} dy + \frac{\partial \varphi}{\partial z} dz + \frac{\partial \varphi}{\partial t} dt \quad (30.12)$$

En suivant une particule dans son mouvement, nous observons pendant un temps dt des déplacements dx, dy, dz . Nous pouvons donc exprimer à partir de l'expression précédente la variation totale de φ pendant le temps dt . Nous obtenons ainsi l'expression d'une dérivée très importante en physique théorique dite "**dérivée particulaire**":

$$\frac{d\varphi}{dt} = \frac{\partial \varphi}{\partial x} \frac{dx}{dt} + \frac{\partial \varphi}{\partial y} \frac{dy}{dt} + \frac{\partial \varphi}{\partial z} \frac{dz}{dt} + \frac{\partial \varphi}{\partial t} \quad (30.13)$$

En mécanique nous allons particulièrement travailler avec le champ gravitationnel Newtonien. Dès lors, la relation reliant la force à l'accélération prend une forme plus générale. Voyons comment:

Soit la dérivée particulaire de la vitesse (pour les trois coordonnées spatiales):

$$\frac{d\vec{v}}{dt} = \frac{\partial \vec{v}}{\partial x} \frac{dx}{dt} + \frac{\partial \vec{v}}{\partial y} \frac{dy}{dt} + \frac{\partial \vec{v}}{\partial z} \frac{dz}{dt} + \frac{\partial \vec{v}}{\partial t} \quad (30.14)$$

Ce qui s'écrit aussi:

$$\frac{d\vec{v}}{dt} = \frac{\partial \vec{v}}{\partial t} + v_x \frac{\partial \vec{v}}{\partial x} + v_y \frac{\partial \vec{v}}{\partial y} + v_z \frac{\partial \vec{v}}{\partial z} \quad (30.15)$$

Ce qui peut s'écrire aussi sous forme condensée:

$$\frac{d\vec{v}}{dt} = \frac{\partial \vec{v}}{\partial t} + (\vec{v} \circ \vec{\nabla}) \vec{v} = \left(\frac{\partial}{\partial t} + \vec{v} \circ \vec{\nabla} \right) \vec{v} \quad (30.16)$$

La deuxième loi de Newton s'écrit alors sous forme généralisée:

$$\boxed{\vec{F} = m \left(\frac{\partial}{\partial t} + \vec{v} \circ \vec{\nabla} \right) \vec{v}} \quad (30.17)$$

Cette formulation de la deuxième loi de Newton est de la plus haute importance en physique. Elle rend compte explicitement de la force subie par un point matériel dans un champ vectoriel en fonction de la vitesse et non plus de la position. Nous retrouverons cette formulation dans le chapitre de Mécanique Des Milieux Continus dans notre étude des fluides et plasmas, dans le chapitre d'Électromagnétisme

ainsi que dans celui de Relativité Générale.

1.3. TROISIÈME LOI (LOI D'ACTION ET RÉACTION)

Énoncé: la réaction d'un corps étendu ou ponctuel solide est toujours de sens opposée et d'intensité et de direction égale à la force imprimée.

Cette troisième loi est plus connue sous le nom de: "**principe d'action/réaction**" et découle de la première loi de Newton selon le raisonnement mathématique lors de notre étude du théorème de Noether dans le chapitre traitant des Principes de la physique.

Ainsi, dans la nature, selon ce principe, il n'y pas de force isolée, chaque force à son "contraire", elles agissent par paire.

Nous pouvons également dire encore que deux corps solides ponctuels ou étendus en contact exercent l'un sur l'autre toujours des forces opposées en sens mais égales en intensité et en direction.

2. CONDITIONS D'ÉQUILIBRE

Pour qu'un point matériel, soumis à des forces $\vec{F}_1, \vec{F}_2, \dots, \vec{F}_n$ soit en équilibre statique, il faut que la résultante de ces forces soit nulle. Soit:

$$\sum_{i=1}^n \vec{F}_i = 0 \quad (30.18)$$

La relation précédente, qui définit donc tout corps à l'équilibre, ouvre l'étude à de très nombreux cas pratiques et constitue à elle seule un immense chapitre d'applications pratiques que nous appelons la "**statique des forces**" et que nous développerons après avoir introduit le concept de moment de force plus loin.

Cependant cette condition est suffisante uniquement dans le cas de points matériels. Si nous travaillons avec des objets étendus dans l'espace alors la condition susmentionnée est nécessaire mais pas suffisante. Nous verrons effectivement plus loin, lors de notre étude des "moments de force", qu'il faut rajouter une condition supplémentaire.

3. CENTRE DE MASSE ET MASSE RÉDUITE

Le centre de masse est un cas particulier du barycentre avec toutes ses propriétés que nous avons déjà largement développées dans le chapitre de Géométrie Euclidienne (donc nous conseillons fortement au lecteur de s'y référer) mais rapporté à la physique.

On peut confondre "centre masse" et "centre de gravité" (dit également "barycentre") que si et seulement si la masse du corps étudié est homogène.

Définition: Soit un solide formé de n points de masse $m_1, m_2, m_3, \dots, m_n$ et repérés par leurs vecteurs de position $\vec{r}_1, \vec{r}_2, \dots, \vec{r}_n$ respectifs. Nous appelons "**centre de masse**" (ou "**centre d'inertie**" s'il y a égalité stricte entre masse grave et masse inerte) un point G auquel nous pouvons rattacher toute la masse du système (et donc son analyse!!) et tel que, l'origine étant arbitrairement choisie il soit donné par:

$$\vec{r}_G = \frac{\sum_{i=1}^n \vec{r}_i \cdot m_i}{\sum_{i=1}^n m_i} \quad (30.19)$$

Relation à comparer par celle plus générale vue dans le chapitre de Géométrie Euclidienne:

$$\overrightarrow{OG} = \frac{\sum_{i=1}^n \alpha_i \overrightarrow{OA_i}}{\sum_{i=1}^n \alpha_i} \quad (21.20)$$

De façon identique, nous définissons la masse réduite du système par la relation:

$$\mu = \frac{\prod_{i=1}^n m_i}{\sum_{i=1}^n m_i} \quad (30.21)$$

Si nous considérons le solide comme continu (vrai seulement à l'échelle macroscopique en première approximation) alors il vient:

$$\vec{r}_G = \frac{\int_V \vec{r} \cdot dm}{\int_V dm} \quad (30.22)$$

Où les intégrales sont étendues au volume du solide en entier.

De plus, si le solide est homogène (cas particulier), de masse volumique ρ , alors $dm = \rho dV$, dV étant l'élément de volume. L'équation peut alors s'écrire (la notation de la triple intégrale est réduite à une seule par souci de condensation d'écriture):

$$\vec{r}_G = \frac{\int_V \vec{r} \cdot dV}{\int_V dV} \quad (30.23)$$

Soit en composantes:

$$x_G = \frac{\int x dV}{V}, y_G = \frac{\int y dV}{V}, z_G = \frac{\int z dV}{V} \quad (30.24)$$

Propriétés:

P1. Si le solide possède un axe de symétrie, alors G est sur cet axe.

P2. Si le solide possède un plan de symétrie, alors G est sur ce plan.

P3. Si le solide possède plusieurs axes de symétrie, alors G est à leur intersection.

Remarque: Le centre de masse G peut se trouver hors du solide (exemple: un tabouret, un boomerang, etc.).

Il n'est pas évident de calculer le centre de masse d'un corps donné relativement simple. Ce n'est pas que les outils mathématiques à manipuler soient complexes loin de là (simple intégrale, Pythagore et quelques multiplications et intégrations par parties) mais il faut aborder le problème d'une façon élégante et si nous n'avons pas tout de suite la bonne approche nous nous casserons très vite les dents. Nous conseillons donc aux professeurs qui abordent ce sujet et les exercices y relatifs, de les faire avec les élèves (donc en classe) mais en laissant ces derniers débattre de la façon dont le professeur doit attaquer le problème au tableau noir (cela marche très bien).

3.1. THÉORÈME DU CENTRE DE MASSE

Sous l'action des forces extérieures $\vec{F}_1, \dots, \vec{F}_i, \dots, \vec{F}_n$, agissant en chaque point du solide, chacun de ces points prend l'accélération correspondant à la force appliquée $\vec{F}_i$. En utilisant la deuxième loi de Newton pour chaque point et en sommant les effets nous aurons (dans un cas non relativiste):

$$\vec{F}_{ext} = \sum_{i=1}^n \vec{F}_i = \sum_{i=1}^n m_i \frac{d^2 \vec{r}_i}{dt^2} = \frac{d^2}{dt^2} \sum_{i=1}^n m_i \vec{r}_i \quad (30.25)$$

en vertu de la position du centre de masse donnée par la relation:

$$\vec{r}_G = \frac{\sum_{i=1}^n \vec{r}_i \cdot m_i}{\sum_{i=1}^n m_i} \quad (30.26)$$

il vient si le référentiel est posé sur le centre de masse:

$$\vec{F}_{ext} = \frac{d^2 \vec{r}_G}{dt^2} \sum_{i=1}^n m_i \quad \text{où} \quad \sum_{i=1}^n m_i = M \quad (30.27)$$

soit:

$$\vec{F}_{ext} = M \frac{d^2 \vec{r}_G}{dt^2} = M \cdot \vec{a}_G \quad (30.28)$$

C'est le théorème du centre de masse, que nous pouvons énoncer ainsi:

Le centre de masse d'un solide se meut comme un point matériel de masse égale à celle du solide et auquel serait appliqué la somme des forces extérieures. Un exemple simple est celui d'un projectile explosif décrivant en absence de pesanteur une trajectoire courbe. Si le projectile explose et se fragmente, le centre de masse des éclats continue à décrire la trajectoire courbe qu'il avait entamée.

Remarque: Dans le cas particulier du solide (ensemble de points) soumis au champ de la pesanteur, $\vec{F}_{ext}$ est le poids du solide et G s'appelle alors "**centre de gravité**" (d'où l'origine de cette appellation).

Reprenons l'équation:

$$\vec{r}_G = \frac{\sum_{i=1}^N \vec{r}_i \cdot m_i}{\sum_{i=1}^N m_i} \quad (30.29)$$

donnant la position du centre de masse. Sa vitesse vaut:

$$\vec{v}_G = \frac{d\vec{r}_G}{dt} = \frac{1}{M} \sum_i m_i \frac{d\vec{r}_i}{dt} = \frac{1}{M} \sum_i m_i \vec{v}_i = \frac{1}{M} \sum_i \vec{p}_i \quad (30.30)$$

en posant:

$$\vec{P} = \sum_i \vec{p}_i \quad (30.31)$$

où $\vec{P}$ est la quantité de mouvement du système, il vient:

$$\vec{P} = M \cdot \vec{v}_G \quad (30.32)$$

Cette relation montre que si la somme des forces extérieures est nulle alors:

$$\vec{F}_{ext} = \frac{d\vec{P}}{dt} = 0 \Rightarrow \vec{P} = \vec{c}^{te} \Rightarrow \vec{v}_G = \vec{c}^{te} \quad (30.33)$$

Donc la quantité de mouvement du système entier est conservée et le mouvement du centre de masse du système est inaltéré. Ceci justifie les remarques faites lors de l'étude de la conservation de la quantité de mouvement.

Dans l'étude des interactions entre particules, il est souvent commode d'utiliser un système de référence lié au centre de masse de l'ensemble des particules. Ce centre de masse étant au repos dans ce référentiel, sa vitesse y est nulle ainsi que la quantité de mouvement totale, comme le montrent les équations ci-dessus. Cette propriété constitue le puissant avantage de cette description.

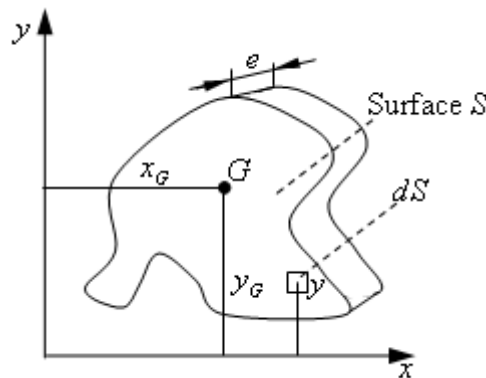
Remarque: En mécanique, l'usage du centre de masse (point matériel) est particulièrement aisé car le système de forces est régi seulement par la loi de Newton. Avec des particules électrisées (charges), il en va tout autrement. Les effets électromagnétiques sont dominants lors de leurs accélérations, ce qui induit des phénomènes ondulatoires interactifs nettement plus complexes. C'est la raison pour laquelle nous ne verrons jamais une étude sur ce site du "centre de charge" lorsque nous aborderons l'électrostatique dans le chapitre d'Électrodynamique...

3.2. THÉORÈMES DE GULDIN

Les théorèmes de Guldin permettent dans certains cas, de simplifier le calcul du centre de masse de certains corps.

Premier théorème: Soit une plaque plane, homogène, d'épaisseur constante e , de masse volumique ρ placée dans un plan cartésien xOy . Nous avons alors par rapport à l'axe y :

$$y_G = \frac{\int y dm}{\int dm} = \frac{\int y dm}{M} = \frac{e \cdot \rho \int y dS}{e \cdot \rho \cdot S} = \frac{\int y dS}{S} \quad (30.34)$$



Envisageons une rotation autour de l'axe x . Le volume décrit par un élément de surface dS lors de cette rotation vaut:

$$dV = 2\pi y dS \quad (30.35)$$

et, par conséquent, le volume total décrit par la surface S complète est:

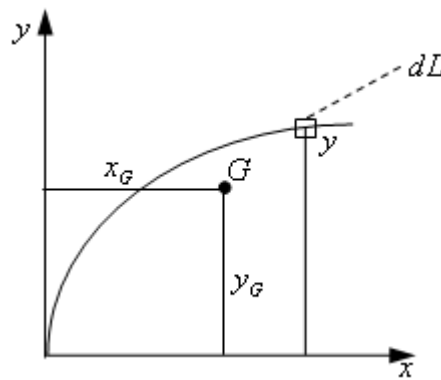
$$V = \int dV = 2\pi \cdot \int y dS = 2\pi \cdot y_G S \quad (30.36)$$

Ainsi, en procédant de même pour x_G , nous obtenons finalement:

$$y_G = \frac{V}{2\pi S} \quad x_G = \frac{V}{2\pi S} \quad (30.37)$$

Deuxième théorème: Soit une tige courbe, homogène, de longueur l , de section constante, de masse linéique ρ_l . Nous avons:

$$y_G = \frac{\int y dm}{\int dm} = \frac{\rho_l \int y dl}{\rho_l \cdot L} = \frac{\int y dl}{L} \quad (30.38)$$



Envisageons une rotation autour de l'axe x . La surface décrite par un élément de longueur dl lors de cette rotation vaut:

$$dV = 2\pi y dl \quad (30.39)$$

et, par conséquent, la surface totale décrite par la tige de longueur L est:

$$S = \int dS = 2\pi \cdot \int y dl = 2\pi \cdot y_G L \quad (30.40)$$

Ainsi, en procédant de même pour x_G , nous obtenons finalement:

$$y_G = \frac{S}{2\pi L} \quad x_G = \frac{S}{2\pi L} \quad (30.41)$$

4. CINÉMATIQUE DU MOUVEMENT RECTILIGNE

Un phénomène est évolutif si, en l'observant, nous constatons un glissement de la valeur concrète d'une ou plusieurs grandeurs. Ces grandeurs ne sont pas des constantes mais des variables. Une évolution implique qu'il y a un début, une infinité d'états intermédiaires et une fin. Un "**état**" est la description d'un instantané d'un phénomène évolutif (pas forcément au sens temporel du terme).

La relation fonctionnelle entre grandeurs pour un état donné peut être décrite par une équation. Pour un phénomène évolutif, il peut y avoir une infinité d'états que nous pouvons décrire par autant d'équations. Sous cette forme, cela n'a pas d'intérêt. Nous cherchons alors à trouver une équation unique qui met en relation les différentes grandeurs vérifiant tous les états que le phénomène évolutif considéré peut admettre. Par cette équation, nous pouvons ensuite calculer n'importe quel état du phénomène évolutif étudié: c'est "**l'équation d'état**" (notion tirée de la thermodynamique).

La "**cinématique**" est donc la partie de la mécanique qui traite des mouvements sans s'occuper de ses causes (c'est Ampère qui a baptisé ce domaine ainsi).

4.1. POSITION

Définition: La position d'un objet est définie par son vecteur position dans le cas particulier d'un espace tridimensionnel:

$$\vec{r} = (x, y, z) \quad (30.42)$$

or chaque coordonnée d'un objet en mouvement peut varier en fonction du temps comme:

$$\vec{r} = (x(t), y(t), z(t)) \quad (30.43)$$

Plutôt que cette notation un peu lourde en parenthèses... les physiciens notent fréquemment le vecteur position (ou vecteur d'espace) sous la forme d'un vecteur de 4 dimensions:

- 3 dimensions spatiales
- 1 dimension temporelle

et nous écrivons alors:

$$\vec{r} = (x, y, z, t) \quad (30.44)$$

et nous appelons alors ce vecteur un "**quadrivecteur d'espace-temps**" dont les composantes sont les coordonnées généralisées du système.

4.2. VITESSE

Définition: La "**vitesse scalaire**", appelée aussi simplement "**vitesse**", notée v , est par définition la distance parcourue par un objet pendant une certaine quantité de temps:

$$v = \frac{d}{t} \quad (30.45)$$

Lorsqu'un corps est en mouvement uniforme rectiligne, c'est-à-dire qu'il parcourt une distance donnée selon une dimension x_i avec $i = \{1, 2, 3\}$ en un temps toujours égal, le rapport précédent est constant dans le temps:

$$v = \frac{x_i}{t} = c^{te} \quad (30.46)$$

La "**vitesse moyenne arithmétique**" est définie comme étant le rapport de la distance parcourue entre un point de départ donné x_{i0} à un instant t_0 et un point d'arrivée x_{if} à un instant t_f :

$$\bar{v} = \frac{x_{if} - x_{i0}}{t_f - t_0} = \frac{\Delta x}{\Delta t} = c^{te} \quad (30.47)$$

Remarque: Il faut prendre garde lors de calculs de vitesses moyennes car il existe plusieurs types de moyennes en mathématique... (cf. chapitre de Statistiques)! Par exemple, il est fréquent d'utiliser la vitesse moyenne harmonique comme nous le montrons dans le chapitre de Statistiques.

Ceci représente donc une moyenne (car nous ne nous intéressons pas de savoir comment le chemin entre x_{if} et x_{i0} a été parcouru) mais nullement la vitesse instantanée du véhicule à un moment donné.

Si nous désirons connaître la vitesse dite "**vitesse scalaire instantanée**" du véhicule en un point de sa trajectoire il faut faire passer le delta du temps Δt à un différentiel (cf. chapitre de Calcul Différentiel Et Intégral) tel que:

$$dt = t_f - t_0 = (t_0 + \varepsilon) - t_0 \quad (30.48)$$

avec ε qui tend vers zéro.

Mathématiquement, nous notons cela correctement de la façon suivante:

$$dt = \lim_{\varepsilon \rightarrow 0} (t_f + \varepsilon) - t_0 \quad (30.49)$$

Ainsi, pendant une différence de temps infiniment petite, la distance parcourue sera également infiniment petite. Nous aurons donc:

$$dx_i = \lim_{\varepsilon \rightarrow 0} (x_{i0} + \varepsilon) - x_{i0} \quad (30.50)$$

et finalement:

$$v = \frac{dx_i}{dt} = \dot{x} \quad (30.51)$$

Si le corps étudié n'est pas en mouvement rectiligne dans un repère cartésien à trois dimensions alors sa position sera donnée par le vecteur (x, y, z) et nous noterons sa vitesse dès lors par:

$$\vec{v} = \left(\frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt} \right) = (v_x, v_y, v_z) = (\dot{x}, \dot{y}, \dot{z}) \quad (30.52)$$

Remarque: Si toutes les parties d'un corps se déplacent à la même vitesse et dans la même direction, nous avons alors un "**mouvement de translation**". Par contre, dans un "**mouvement de rotation**", les vitesses des diverses parties du corps ne sont pas les mêmes, en module et en direction (nous le démontrerons plus loin) et peuvent varier avec le temps.

Attention ! Un mouvement ne peut être décrit que par rapport à un repère fixe: le mouvement absolu n'existe pas. Galilée avait déjà compris que: "Le mouvement est comme rien". Le mouvement n'existe pas en soi, mais relativement à autre chose.

4.3. ACCÉLÉRATION

Définition: "**L'accélération scalaire**", appelée aussi simplement "**accélération**", notée a , est par définition, la variation de la vitesse scalaire pendant une certaine quantité de temps telle que (nous passons directement à la limite):

$$a_i = \frac{dv_i}{dt} = \frac{d^2 x_i}{dt^2} = \ddot{x} \quad (30.53)$$

ou autrement dit: la vitesse avec laquelle évolue la vitesse.

À nouveau, si le corps n'est pas en mouvement uniforme rectiligne nous aurons:

$$\vec{a} = \left(\frac{dv_x}{dt}, \frac{dv_y}{dt}, \frac{dv_z}{dt} \right) = (a_x, a_y, a_z) = (\dot{v}_x, \dot{v}_y, \dot{v}_z) \quad (30.54)$$

Si le corps est en mouvement rectiligne et uniforme (nous pouvons toujours généraliser à un mouvement non rectiligne) nous avons alors:

$$\begin{aligned} v_i &= \frac{dx_i}{dt} \Leftrightarrow v_i \cdot dt = dx_i \\ \Rightarrow \int_{t_0}^t v_i \cdot dt &= \int_{t_0}^t dx_i \Rightarrow v_i \cdot t \Big|_{t_0}^t + c_1^{te} = x_i \Big|_{t_0}^t + c_2^{te} \\ \Rightarrow x_i(t) - x_i(t_0) &= c^{te} + v_i \cdot t - v_i \cdot t_0 \Rightarrow \Delta x_i = c^{te} + v_i \cdot \Delta t \end{aligned} \quad (30.55)$$

La constante est à déterminer en fonction des conditions initiales. Si la position initiale au temps zéro est nulle la constante sera nulle. Dans le cas contraire nous écrivons:

$$\Delta x_i = x_{i0} + v_i \cdot \Delta t \quad (30.56)$$

ce qui nous donne la distance parcourue par un corps pendant un laps de temps donné.

Si le corps est en mouvement rectiligne et accélère constamment nous avons alors:

$$\begin{aligned} a_i &= \frac{dv_i}{dt} \Leftrightarrow a_i \cdot dt = dv_i \\ \Rightarrow \int_{t_0}^t a_i \cdot dt &= \int_{t_0}^t dv_i \Rightarrow a_i \cdot t \Big|_{t_0}^t + c_1^{te} = v_i \Big|_{t_0}^t + c_2^{te} \\ \Rightarrow v_i(t) - v_i(t_0) &= c^{te} + a_i \cdot t - a_i \cdot t_0 \Rightarrow \Delta v_i = c^{te} + a_i \Delta t \end{aligned} \quad (30.57)$$

La constante est à déterminer en fonction des conditions initiales. Si la vitesse initiale parcourue au temps zéro est nulle la constante sera nulle. Dans le cas contraire nous écrivons:

$$\Delta v_i = v_{i0} + a_i \cdot \Delta t \quad (30.58)$$

Nous voyons plus fréquemment cette relation sous la forme:

$$v_i(t) = v_{i0} + a_i(t)$$

ou encore majoritairement sous la forme:

$$v = v_0 + a \cdot t \quad (30.59)$$

mais nous avons:

$$v = \frac{dx}{dt} = v_0 + a \cdot t \Rightarrow dx = v_0 dt + a \cdot t dt \quad (30.60)$$

si nous intégrons cette relation, nous obtenons:

$$\int_{t_0}^t dx = \int_{t_0}^t v_0 dt + \int_{t_0}^t a \cdot t dt \Rightarrow \Delta x = x_0 + v_0 \Delta t + \frac{1}{2} a \cdot \Delta t^2 \quad (30.61)$$

que nous retrouvons dans les écoles le plus fréquemment sous la forme:

$$x = x_0 + v_0 t + \frac{1}{2} a \cdot t^2 \quad (30.62)$$

Cette relation donne la position d'un mobile en mouvement rectiligne et uniformément accéléré (dans la majorité des problèmes en physique on considère l'accélération constante ou en un point précis). De cette dernière, nous déduisons une très quantité de relations qui sont très intéressantes en physique aussi bien en considérant des cas idéaux que des cas réels. Effectivement, en la réarrangeant (alèbre élémentaire), il est possible d'obtenir l'accélération lorsque le temps pour atteindre une vitesse donnée est connu. On peut aussi calculer la distance qu'il faut à un mobile (toujours en mouvement rectiligne) ayant une accélération donnée pour atteindre une certaine vitesse.

Le premier cas que nous considérons comme le plus connu, est la vitesse de chute à accélération constante d'un corps dans un milieu exempt de tout frottement (cas traité plus loin lors de notre étude de la tribologie).

Comme nous l'avons déjà démontré précédemment, nous avons lorsque la vitesse initiale est nulle:

$$v = a \cdot t \Leftrightarrow t = \frac{v}{a} \text{ et } x = \frac{1}{2} a \cdot t^2 \quad (30.63)$$

Les deux relations combinées donnent (conformément à la tradition d'usage dans les écoles, nous avons remplacé le x par un h pour indiquer que la position est souvent assimilée dans la pratique à une hauteur):

$$x = \frac{1}{2} a \cdot \left(\frac{v}{a} \right)^2 = \frac{1}{2} \cdot \frac{v^2}{a} \Rightarrow v = \sqrt{2 \cdot a \cdot h} \quad (30.64)$$

Nous pouvons tirer de cette relation la vitesse de libération d'un astre (relation pratique quand nous étudierons le chapitre d'Astrophysique et intéressante pour comparaison lorsque nous étudierons la relativité générale):

Supposons que vous savez déjà que deux corps s'attirent mutuellement avec une accélération selon le modèle classique de Newton (que nous démontrerons plus loin):

$$a = g = \frac{G \cdot m}{R^2} \quad (30.65)$$

Mis dans la relation de chute d'un corps $v = \sqrt{2 \cdot a \cdot h}$, nous obtenons:

$$v = \sqrt{2 \cdot \frac{G \cdot m}{R^2} \cdot h} \quad (30.66)$$

à la surface du corps attracteur principal nous avons donc la "**vitesse de libération**":

$$v = \sqrt{\frac{2 \cdot G \cdot m}{R^2} \cdot R} = \sqrt{\frac{2 \cdot G \cdot m}{R}} \quad (30.67)$$

Nous pouvons répondre à partir de cette relation, à la question de savoir pourquoi certaines planètes du système solaire ont une atmosphère et d'autres pas (bon normalement il faut prendre en compte l'agitation moléculaire...) comme nous le verrons dans le chapitre d'Astronomie.

Ce qui est aussi intéressant dans cette relation c'est que nous pouvons calculer quel doit être le rayon R d'un corps de masse m pour que sa vitesse de libération soit égale à celle de la lumière (allusion aux Trous Noirs).

Nous avons dès lors:

$$c = \sqrt{\frac{2 \cdot G \cdot m}{R}} \Rightarrow R = \frac{2 \cdot G}{c^2} \cdot m \quad (30.68)$$

Nous verrons dans le chapitre de Relativité Générale qu'après de relativement longs calculs dans un champ gravitationnel isotrope (métrique de Schwarzschild) nous retomberons sur cette relation.

4.3.1. PLAN OSCULATEUR

Les vecteurs vitesse $\vec{v}(t)$ et accélération $\vec{a}(t)$ liés à un point P en mouvement forment, à chaque instant t un plan appelé "**plan osculateur**" de la trajectoire (généralement curviligne sinon quoi le plan se réduit à une droite).

Il est souvent utile de décomposer le vecteur accélération dans le plan osculateur suivant respectivement la tangente et la normale à la trajectoire:

$$\vec{a} = \vec{a}_t + \vec{a}_n \quad (30.69)$$

où le premier terme du membre de droite est un vecteur parallèle à la vitesse et le deuxième un vecteur perpendiculaire à la vitesse et situé du côté concave de la trajectoire.

Exprimons ces deux vecteurs (un exemple plus général est donné dans le chapitre de Géométrie Différentielle):

Nous pouvons écrire que:

$$\vec{v} = \frac{d\vec{r}}{dt} = \frac{d\vec{r}}{ds} \cdot \frac{ds}{dt} = \vec{T} \cdot v \quad (30.70)$$

où ds est un élément courbe (l'abscisse curviligne) de la trajectoire et $\vec{T}$ un vecteur unité tangent à la trajectoire lié au point P .

L'accélération s'écrit alors:

$$\vec{a} = \frac{d\vec{v}}{dt} = \frac{d(\vec{T} \cdot v)}{dt} = \vec{T} \frac{dv}{dt} + v \frac{d\vec{T}}{dt} \quad (30.71)$$

Le premier terme à droite de l'égalité est l'accélération tangentielle quant au second terme, même si la

vitesse est constante ce dernier apparaît dans l'expression de l'accélération pour exprimer le changement de direction de la vitesse.

Décomposons le vecteur $\vec{T}$ dans la base orthonormée euclidienne $\mathbb{R}^2$ générée par la famille de vecteurs $\vec{i}, \vec{j}$:

$$\vec{T} = T_x \cdot \vec{i} + T_y \cdot \vec{j} = \|\vec{T}\| \cdot \cos \alpha \cdot \vec{i} + \|\vec{T}\| \cdot \sin \alpha \cdot \vec{j} = \cos \alpha \cdot \vec{i} + \sin \alpha \cdot \vec{j} \quad (30.72)$$

Ensuite, en dérivant par rapport au temps:

$$\begin{aligned} \frac{d\vec{T}}{dt} &= -\sin \alpha \frac{d\alpha}{dt} \vec{i} + \cos \alpha \frac{d\alpha}{dt} \vec{j} = (-\sin \alpha \cdot \vec{i} + \cos \alpha \cdot \vec{j}) \frac{d\alpha}{dt} \\ &= \left[\cos \left(\alpha + \frac{\pi}{2} \right) \cdot \vec{i} + \sin \left(\alpha + \frac{\pi}{2} \right) \cdot \vec{j} \right] \frac{d\alpha}{dt} \end{aligned} \quad (30.73)$$

En comparant avec l'expression initiale du vecteur $\vec{T}$, nous voyons que les termes entre crochets ci-dessus sont les composantes d'un nouveau vecteur unité $\vec{n}$ perpendiculaire au vecteur $\vec{T}$, donc perpendiculaire à la trajectoire et dirigé vers le centre de courbure.

De plus par la définition du radian, nous avons:

$$\frac{d\alpha}{dt} = \frac{ds}{R} \cdot \frac{1}{dt} = \frac{1}{R} \cdot \frac{ds}{dt} = \frac{v}{R} \quad (30.74)$$

où R est le rayon de courbure de la trajectoire.

L'expression $d\vec{T}/dt$ devient alors:

$$\frac{d\vec{T}}{dt} = \frac{v}{R} \vec{n} \quad (30.75)$$

et le second terme de l'expression générale de l'accélération devient alors:

$$v \frac{d\vec{T}}{dt} = \frac{v^2}{R} \vec{n} \quad (30.76)$$

Nous avons donc finalement (relation démontrée avec une autre approche dans le chapitre de Géométrie Différentielle):

$$\vec{a} = \frac{dv}{dt} \vec{T} + \frac{v^2}{R} \vec{n} \quad (30.77)$$

où "l'accélération tangentielle" donnée par:

$$\vec{a}_{tg} = \frac{dv}{dt} \vec{T} \quad (30.78)$$

est un terme qui exprime la modification de l'intensité de la vitesse sur la trajectoire du point P et où "l'accélération normale":

$$\vec{a}_n = \frac{v^2}{R} \vec{n} \quad (30.79)$$

est un terme qui exprime le changement de direction du point P sans que nécessairement ce dernier change donc de vitesse! Communément cette dernière relation est assimilée à la "**force centrifuge**" (centrifuge signifiant: qui fuit le centre")!!

Remarque: La force centrifuge est considérée en physique comme une force fictive car au fait il ne s'agit pas d'une force qui tend à nous éloigner d'un centre de rotation mais c'est juste qu'il y a une force qui n'est plus suffisante (la force de frottement dans le cadre d'un manège ou gravitationnelle pour des planètes) pour nous empêcher de suivre une trajectoire en ligne droite par simple inertie. Raison pour laquelle lorsque nous sommes éjectés d'un manège nous partons tangentiellement à sa rotation et non pas perpendiculairement à celle-ci.

Nous constatons immédiatement que si $\vec{a}_n = 0$ le mouvement est forcément rectiligne, accéléré ou non, tandis que si $\vec{a}_n \neq 0$ la trajectoire est nécessairement incurvée.

5. PRINCIPE DE RELATIVITÉ GALILÉEN

Définition: Il est impossible pour un observateur animé d'un mouvement uniforme de savoir s'il se meut par rapport à son environnement ou bien à l'inverse si l'environnement se déplace par rapport à lui (nous ne pouvons pas distinguer le repos et le mouvement à vitesse et direction constantes). Dès lors, il ne peut exister de référentiel absolu (ou privilégié) qui puisse être considéré comme fixe vis-à-vis de tous les autres repères galiléens ce qui signifie clairement que tous les repères galiléens doivent jouir du même statut en mécanique puisqu'ils ne peuvent être distingués les uns des autres. Ce principe est nommé le "**principe de relativité galiléen**".

Ce principe, (à ne pas confondre avec le principe de relativité restreinte car les hypothèses de départ diffèrent un tant soit peu...) découle directement de l'étude de ce que nous nommons la "transformation de Galilée".

Définition: Une "**transformation de Galilée**" est une suite d'opérations mathématiques sur une loi physique qui permet de déterminer les propriétés d'une ou plusieurs "**observables**" (vitesse, force, quantité de mouvement, etc.) lorsque nous passons lors de l'étude d'un phénomène physique d'un référentiel à un autre référentiel: l'un supposé au repos, et l'autre en mouvement uniforme.

La question à l'origine historique était de répondre s'il est plus légitime d'étudier un phénomène dans un référentiel ou dans un autre. Plus exactement, nous souhaitons déterminer si la forme des lois physiques gardent les mêmes formes algébrique quelque soient les référentiels dans lequel nous les étudions.

Voyons cela d'un peu plus près:

Soient deux référentiels en mouvements l'un par rapport à l'autre à une vitesse constante $\vec{u}(u_x, u_y, u_z)$. Pour un certain référentiel cartésien $(\vec{x}, \vec{y}, \vec{z})$ au repos (ou supposé tel) nous allons poser le deuxième référentiel $(\vec{x}', \vec{y}', \vec{z}')$ de façon à ce qu'il soit aligné avec l'axe des $\vec{x}$ afin de simplifier les calculs avec $\vec{u}(u_x, 0, 0)$:

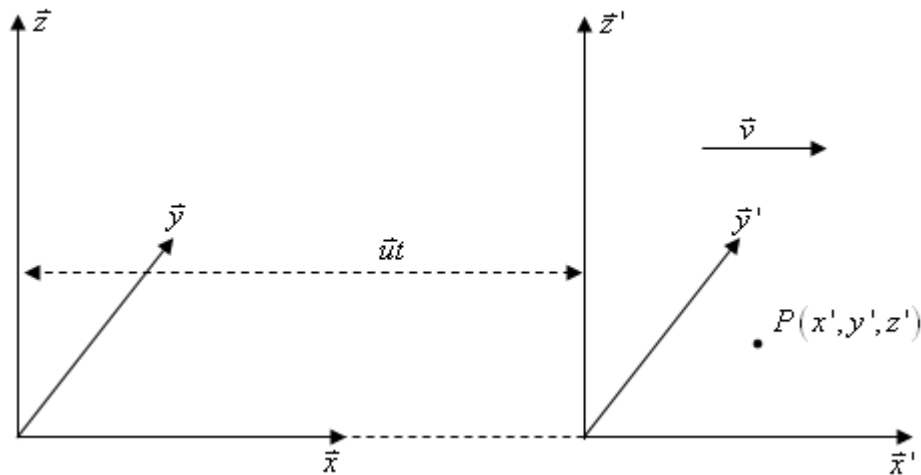


Figure: 30.6 - Exemple de référentiel en mouvement par rapport à un autre

Nous allons également mettre dans le deuxième référentiel en mouvement, un point matériel $\vec{P}$ de coordonnées $\vec{P}(x', y', z')$.

Remarque: Nous supposons connu le concept de "quantité de mouvement" p défini plus loin avec rigueur. Rappelons donc dès lors que la quantité de mouvement du point P animé d'une vitesse v (norme) dans $(\vec{x}', \vec{y}', \vec{z}')$ est alors donnée par:

$$\vec{p}' = m' \vec{v}' \quad (30.80)$$

Nous avons alors en appliquant les relations classiques de la cinématique:

$$\begin{aligned} \begin{cases} x' = x - u_x t, y' = y, z' = z \\ t' = t, m' = m \end{cases} \\ \Rightarrow p'_x = m v'_x = m \frac{dx'}{dt'} = m \frac{d(x - u_x t)}{dt} = m \left(\frac{dx}{dt} - \frac{du_x}{dt} t - \frac{dt}{dt} u_x \right) = m \left(\frac{dx}{dt} - 0 \cdot t - 1 \cdot u_x \right) \\ = m \left(\frac{dx}{dt} - u_x \right) = m(v_x - u_x) \end{aligned} \quad (30.81)$$

d'où $v'_x = v_x - u_x$ et donc (nous supposons connu le concept de "force" défini plus loin avec rigueur):

$$F_x = \frac{dp_x}{dt} = \frac{dm(v_x - u)}{dt} = 0 = F'_x = \frac{dp'_x}{dt'} = \frac{dmv'_x}{dt'} = 0 \quad (30.82)$$

Le résultat obtenu est donc fort intéressant puisque la deuxième loi de Newton garde exactement la même forme, et la même valeur dans les deux référentiels. Le fait que nous nous déplaçons ou pas à vitesse constante ou pas n'a donc aucune influence sur notre vision du monde qui reste exactement la même.

Conséquence: Puisque les forces sont identiques, aucune expérience de mécanique ne peut déterminer

si un référentiel galiléen est le repère absolu (autrement dit deux observateurs, dans deux référentiels galiléens différents, ne peuvent à l'aide d'une expérience de mécanique déterminer lequel se meut par rapport à l'autre).

Donc, en mécanique classique, il n'existe pas de référentiel galiléen absolu!

Toutefois notons bien que ce résultat est obtenu en supposant que:

$$\frac{d(mu)}{dt} = \frac{d(mv_x)}{dt} = \frac{d(mv'_x)}{dt'} = 0 \quad (30.83)$$

c'est-à-dire que nous imposons que la vitesse relative est uniforme (constante) et la masse constante et surtout, que $t = t'$.

Mais au fait, cette transformation est fondamentalement fausse comme nous le verrons plus en détail lors de notre étude de la relativité restreinte (cf. chapitre de Relativité Restreinte). Effectivement, soit un objet se déplaçant le long de l'axe avec une vitesse v mesurée dans le repère primé:

$$x' = vt' \quad (30.84)$$

quel sera alors sa vitesse w dans le repère non primé? Si la transformation de Galilée est fondamentalement vraie, il suffit de remplacer dans la relation précédente x' et t' par leurs expressions en fonction de t :

$$x - ut = vt \quad \text{ou} \quad x = (u + v)t \quad (30.85)$$

soit (loi d'addition des vitesses):

$$w = \frac{dx}{dt} = u + v \quad (30.86)$$

Seul petit hic... une expérience simple impliquant des rayons de lumière fut réalisée au début du siècle, et montra que cette loi était fausse. Cette expérience dite de "**expérience Michelson-Morley**" bouleversa à tout jamais notre vision du monde... et amena Albert Einstein à développer la théorie de la relativité restreinte en imposant que la vitesse de la lumière quel que soit le référentiel est toujours constante (cf. chapitre de Relativité Restreinte):

$$c = c^{te} \quad (30.87)$$

Remarque: Si nous mesurons les vitesses et autres grandeurs vectorielles, nous trouvons que les résultats de mesures des composantes x', y', z', t' ne sont pas identiques à celle obtenues sur x, y, z, t . Elles varient avec le système d'axe. Connaissant ces valeurs dans un repère, nous pouvons passer aux valeurs dans l'autre repère: il s'agit de la "**covariance**" (co-variance: variance avec les coordonnées), ici pour les expressions vectorielles.

Les lois sont des relations entre des observables, relations déduites d'observations nombreuses. La recherche des lois est régie par ce que nous pourrions appeler un "**principe de simplicité**": lois en nombre le plus petit possible, expressions les plus simples possibles entre grandeurs en nombre minimal.

Mais la caractéristique d'une bonne loi est la covariance lors d'un changement de repère. Cette invariance lors d'un changement de repère, cette invariance de la forme (de l'expression littérale) de la loi va permettre d'objectiver au maximum et, en principe totalement, la physique.

La physique (dans le sens de la théorie qui décrit la réalité) ne sera plus liée à l'observateur ni à son espace-temps galiléen associé. Bien sûr cette covariance sera recherchée pour les transformations de référentiels en mouvement les uns par rapport aux autres.

Un contre-exemple simple cependant: la force entre deux charges électriques immobiles dans un référentiel ne fait appel dans ce référentiel qu'à la seule théorie de l'électrostatique. Si ce même système est observé d'un référentiel en mouvement par rapport au premier, il faudra décrire l'ensemble à l'aide de la théorie de l'électromagnétisme.

Par construction même la mécanique classique se trouve être covariante par transformation de Galilée (changement de repères galiléens): le postulat de la dynamique (force) prend en effet la même forme dans les différents référentiels galiléens comme nous venons de le voir.

6. MOMENT CINÉTIQUE

Définition: Le "**moment cinétique**" ou "**moment angulaire**" $\vec{b}$ par rapport à un point O d'une particule de masse m se déplaçant à la vitesse $\vec{v}$ en $\vec{r}$ est défini par:

$$\vec{b} = \vec{r} \times \vec{p} \quad (30.88)$$

avec $\vec{p}$ étant la quantité de mouvement (voir la définition plus loin) donnée par:

$$\vec{p} = m\vec{v} \quad (30.89)$$

Par sa définition, le moment cinétique est un vecteur perpendiculaire au plan contenant les vecteurs $\vec{r}$ et $\vec{p}$ et si la particule se déplace dans un plan, la direction de $\vec{b}$ est constante mais pas nécessairement de même sens.

Un cas particulier mais important en mécanique et astronomie du calcul du moment cinétique est le mouvement circulaire (plan) de rayon r . Dans cette situation, le "**rayon-vecteur**" $\vec{r}$ est alors toujours perpendiculaire à la direction du vecteur-vitesse $\vec{v}$ et donc:

$$\vec{b} = m\vec{r} \times \vec{v} \Rightarrow \|\vec{b}\| = m \cdot v \cdot r = m\omega r^2 \Rightarrow \vec{b} = mr^2 \vec{\omega} \quad (30.90)$$

Nous voyons apparaître ici la définition du "**vecteur rotation**" $\vec{\omega}$ également noté parfois (à tort) $\vec{\Omega}$.

Pour un mouvement plan mais non circulaire (comme une conique par exemple!), nous introduisons les composantes normale et tangentielle de la vitesse:

$$\vec{v} = \vec{v}_n + \vec{v}_{tg} \quad (30.91)$$

pour obtenir (de par les propriétés du produit vectoriel):

$$\vec{b} = m\vec{r} \times (\vec{v}_n + \vec{v}_{tg}) = m\vec{r} \times \vec{v}_{tg} \quad (30.92)$$

et sous forme scalaire:

$$b = mrv_{\text{tg}} = mr^2 \frac{d\theta}{dt} \quad (30.93)$$

où r est dès lors appelé le "**rayon de courbure**" de la trajectoire.

Etudions maintenant la dérivée du moment cinétique:

$$\frac{d\vec{b}}{dt} = \frac{d(\vec{r} \times \vec{p})}{dt} = \frac{d\vec{r}}{dt} \times \vec{p} + \vec{r} \times \frac{d\vec{p}}{dt} \quad (30.94)$$

Dans le membre de droite, nous avons de par la définition du produit vectoriel:

$$\frac{d\vec{r}}{dt} \times \vec{p} = \vec{v} \times m\vec{v} = \vec{0} \quad (30.95)$$

et d'autre part:

$$\vec{r} \times \frac{d\vec{p}}{dt} = \vec{r} \times \vec{F} \quad (30.96)$$

Ce qui nous donne finalement:

$$\boxed{\frac{d\vec{b}}{dt} = \vec{r} \times \vec{F} = \vec{M}} \quad (30.97)$$

La dérivée par rapport au temps du moment cinétique d'un mobile ponctuel est donc égale à ce que nous définissons par le "**moment de force**" $\vec{M}$ sur lequel nous reviendrons plus loin et qui a comme unités celle de l'énergie et est un vecteur perpendiculaire au plan formé par $\vec{r}$ et $\vec{F}$ (par construction du produit vectoriel!).

Remarques:

R1. Cette dernière relation fait que nous appelons parfois le moment cinétique aussi "**moment de la quantité de mouvement**".

R2. Il faut bien sûr prendre garde au fait que (cf. chapitre de Calcul Vectoriel) $\vec{r} \times \vec{F} \neq \vec{F} \times \vec{r}$.

Ce qui est fortement impressionnant dans ce résultat (variation instantanée du moment cinétique), est que tout corps ayant un moment cinétique non nul et soumis à aucun moment de force, conserve l'orientation et la norme de $\vec{b}$ dans l'espace et le temps.

Ce résultat va nous permettre d'étudier la dynamique du gyroscope et de tous les autres corps ayant des propriétés similaires (comme la Terre qui tourne sur elle-même et qui pointe sur l'étoile polaire ce qui est un facteur important de l'origine des saisons!). Nous étudierons plus loin le gyroscope et ses propriétés, car son comportement est fascinant et les résultats théoriques en découlant trouvent des applications en astrophysique, physique atomique et même en philosophie. Effectivement, la conservation de l'orientation du vecteur moment cinétique amène à considérer que même si l'espace était totalement vide de son contenu dans l'ensemble de l'Univers et contenait un objet ayant un

moment cinétique, l'espace vide a toutefois une propriété qui permet à l'objet en question de savoir où il doit s'orienter... Ce qui est assez déroutant! Donc l'espace vide ce n'est pas rien sinon par rapport à quoi tourne l'objet dans le vide? Il tourne par rapport à l'espace lui-même!

Nous avons également:

$$\vec{b} = \int_{t_1}^{t_2} \vec{M} dt \quad (30.98)$$

où l'intégrale s'appelle "**l'impulsion de rotation**" et la relation précédente porte quelquefois le nom de "**théorème du moment cinétique**" (nous verrons une généralisation de ce théorème lors de la démonstration du théorème de König). Il s'énonce ainsi:

L'impulsion de rotation fournie par un moment de force entre les instants t_1 et t_2 est égale à la variation du moment cinétique durant cet intervalle de temps.

En dynamique du solide ce théorème joue un rôle fondamental, analogue à l'équation de Newton $\vec{F} = m\vec{a}$ en dynamique du point.

L'utilisation du moment cinétique permet de montrer facilement la loi des aires (deuxième loi de Kepler), qui joue un rôle important dans la compréhension du mouvement des planètes (cf. chapitre d'Astronomie) ou encore de montrer que dans un système Terre-Lune isolé, le moment cinétique total devant être conservé, si la Terre ralentit sa rotation et la Lune la garde constante, cela oblige la Lune à augmenter sa distance par rapport à la Terre.

Voyons cela:

Imaginons une particule en mouvement sous l'action d'une force $\vec{F}$ constamment parallèle à $\vec{r}$. Nous dirons que cette force est une "**force centrale**" si sa direction passe constamment par un même point fixe, appelé le "**centre de force**". La grandeur de la force ne peut donc plus dépendre que de la distance au centre de la force (dans le cas d'un champ de force).

Dès lors:

$$\frac{d\vec{b}}{dt} = \vec{r} \times \vec{F} = \vec{0} \quad (30.99)$$

Donc le moment cinétique par rapport au centre de force est constant si la force est centrale. La réciproque est aussi vraie: si le moment cinétique est constant, sa dérivée par rapport au temps est nulle et la direction de la force est toujours colinéaire à $\vec{r}$ donc la force est centrale.

Par exemple, dans le cadre du mouvement d'une planète autour du Soleil ou d'un électron autour du noyau de l'atome (dans le cadre du modèle de Bohr) le moment de cette force par rapport au centre est évidemment nul puisque qu'aucun élément extérieur n'agit sur le système, c'est-à-dire en se basant sur le schéma ci-dessous:

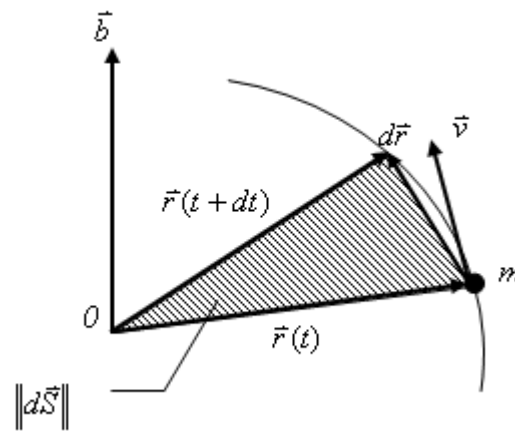


Figure: 30.7 - Illustration du moment cinétique

nous avons alors:

$$\vec{M} = \vec{r} \times \vec{F} = 0 \Rightarrow \vec{M} = \frac{d\vec{b}}{dt} = 0 \quad (30.100)$$

donc:

$$\vec{b} = \vec{r} \times m\vec{v} = \vec{c}^{te} \quad (30.101)$$

D'autre part, l'élément de surface $d\vec{S}$ décrit par le mouvement du rayon $\vec{r}$ vaut (selon la figure ci-dessus et la propriété du produit vectoriel):

$$d\vec{S} = \frac{1}{2}(\vec{r} \times d\vec{r}) \quad (30.102)$$

donc:

$$\frac{d\vec{S}}{dt} = \frac{1}{2} \left(\vec{r} \times \frac{d\vec{r}}{dt} \right) = \frac{1}{2}(\vec{r} \times \vec{v}) \quad (30.103)$$

En utilisant la relation $\vec{b} = \vec{r} \times m\vec{v} = \vec{c}^{te}$ nous obtenons:

$$\boxed{\frac{d\vec{S}}{dt} = \frac{1}{2m}(\vec{r} \times m \cdot \vec{v}) = \vec{c}^{te}} \quad (30.104)$$

Conséquences:

1. La vitesse aréolaire est constante, c'est-à-dire que les aires balayées en des temps égaux sont égales. C'est la loi des aires de Kepler (cf. chapitre d'Astronomie)!
2. Le plan $\vec{r}, \vec{v}$ est fixe car $\vec{r} \times m \cdot \vec{v} = \vec{c}^{te}$. Donc la trajectoire, d'une planète dans un cadre idéal par exemple, est plane.

6.1. MOMENT DE FORCE

Nous venons de voir que le "**moment de force**" se définissait par la relation (variation temporelle du moment cinétique):

$$\vec{M} = \vec{r} \times \vec{F} \quad (30.105)$$

où $\vec{M}$ est donc le moment de la force $\vec{F}$ par rapport au point d'origine du vecteur $\vec{r}$. Il est important de remarquer que le moment de force a les unités d'une énergie et est donc perpendiculaire à $\vec{r}$ et $\vec{F}$ par construction!

Il faut aussi remarquer qu'augmenter le rayon d'application en diminuant ainsi la force pour garder un moment de force constant dans un système mécanique permet certes de diminuer l'effort (la force) mais au final pas l'énergie dépensée puisque la distance parcourue est alors plus grande.

Si nous exprimons le module de $\vec{M}$, de par la définition du produit vectoriel, nous obtenons:

$$\|\vec{M}\| = \|\vec{F}\| \cdot \|\vec{r}\| \cdot \sin \alpha = \|\vec{F}\| \cdot b \quad (30.106)$$

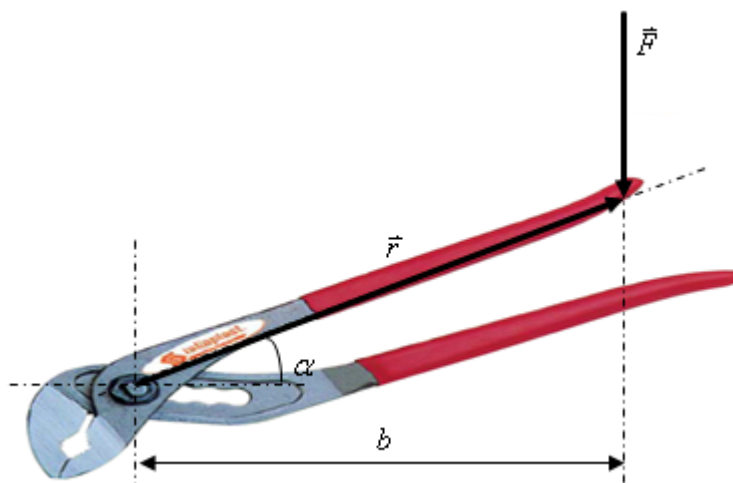
Il apparaît une grandeur:

$$b = \|\vec{r}\| \cdot \sin \alpha \quad (30.107)$$

qui est par définition le "**bras de levier**" de la force $\vec{F}$ et dont l'emplacement est donné par l'axe de rotation du corps due au moment de force résultant (attention à ne pas confondre ce b avec la notation du moment cinétique!).

Attention! Le principe des bras de levier est donc un fantastique démultiplicateur de force mais en aucun cas il démultiplie le travail!

Exemples:



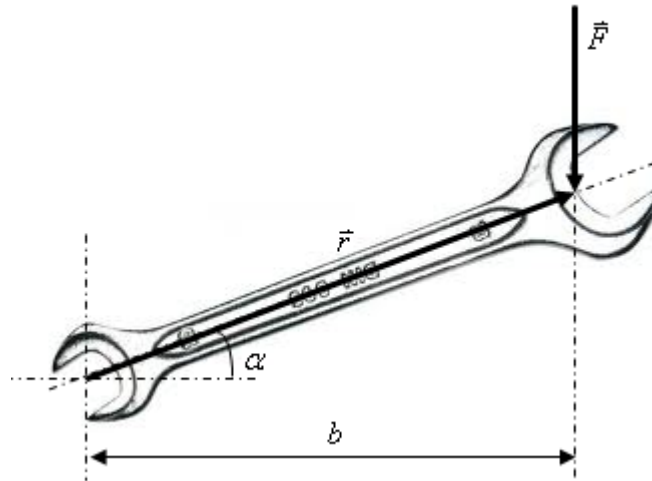


Figure: 30.8 - Exemple de quelques bras de levier de tous les jours

Le cas d'application des changements de roues d'hiver/d'été est très connu par les automobilistes puisqu'il est recommandé par la majeure partie des fabricants d'appliquer un couple de 120 [Nm] pour le serrage des boulons.

Pour qu'un corps étendu, soumis à des forces $\vec{F}_1, \vec{F}_2, \dots, \vec{F}_n$ soit en équilibre total, il ne faut ainsi pas que uniquement la résultante de ces forces soit nulle (pas de translation) mais que la résultante des moments soit nulle aussi (pas de rotation). Soit:

$$\sum_{i=1}^n \vec{F}_i = \vec{0} \text{ et } \sum_{i=1}^n \vec{M}_i = \vec{0} \quad (30.108)$$

Lorsque les composantes d'un système satisfaisant aux deux relations ci-dessus sont connues, nous parlons alors de système "**isostatique**".

Par définition, un "**couple**" est défini comme un ensemble de deux forces de grandeur égale mais de direction opposée, agissant suivant deux droites parallèles sur un même corps étendu. La résultante des forces bien évidemment nulle, indique que le couple ne produit aucun effet de translation. Mais la somme des moments étant non nulle, le corps subit une rotation telle que:

$$\vec{M} = \vec{r}_1 \times \vec{F}_1 + \vec{r}_2 \times \vec{F}_2 = \vec{r}_1 \times \vec{F}_1 - \vec{r}_2 \times \vec{F}_1 = (\vec{r}_1 - \vec{r}_2) \times \vec{F}_1 = \vec{b} \times \vec{F} \quad (30.109)$$

Signalons encore la composition de moments de force avec le cas ultra-classique suivant:

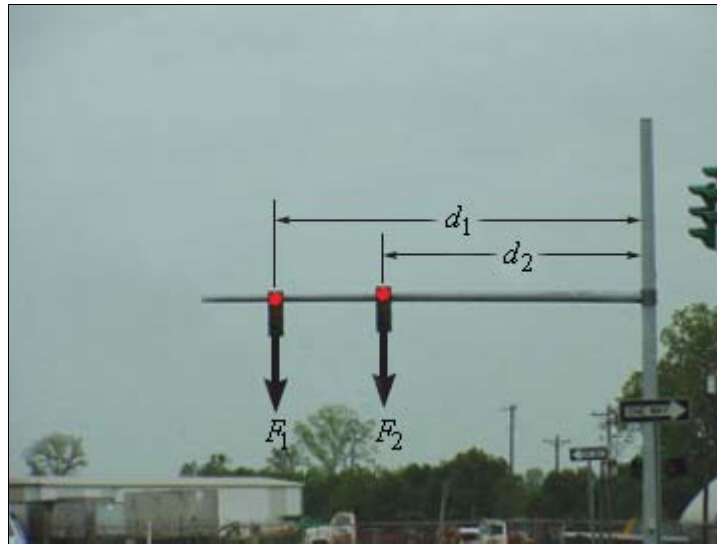


Figure: 30.9 - Moments de force élémentaires

pour lequel il est assez évident que le moment de force résultant a pour composante de force la somme des deux forces élémentaires:

$$F_R = F_1 + F_2 \quad (30.110)$$

Pour le calcul de la distance résultante, nous allons utiliser le théorème du centre de masse qui comme nous l'avons vu plus haut et dans le chapitre de Géométrie Euclidienne est dans le cas général:

$$\overrightarrow{OG} = \frac{\sum_{i=1}^n \alpha_i \overrightarrow{OA_i}}{\sum_{i=1}^n \alpha_i} \quad (21.111)$$

et donc dans la situation ici présente se réduit à:

$$d = \frac{\sum_{i=1}^n F_i \cdot d_i}{\sum_{i=1}^n F_i} = \frac{F_1 \cdot d_1 + F_2 \cdot d_2}{F_1 + F_2} \quad (21.112)$$

Nous avons donc au final le moment résultant qui est:

$$M_R = F_R \cdot d = (F_1 + F_2) \frac{F_1 \cdot d_1 + F_2 \cdot d_2}{F_1 + F_2} = F_1 \cdot d_1 + F_2 \cdot d_2 \quad (21.113)$$

Ainsi, la résultante de moments de forces est la simple somme des moments de force élémentaires.

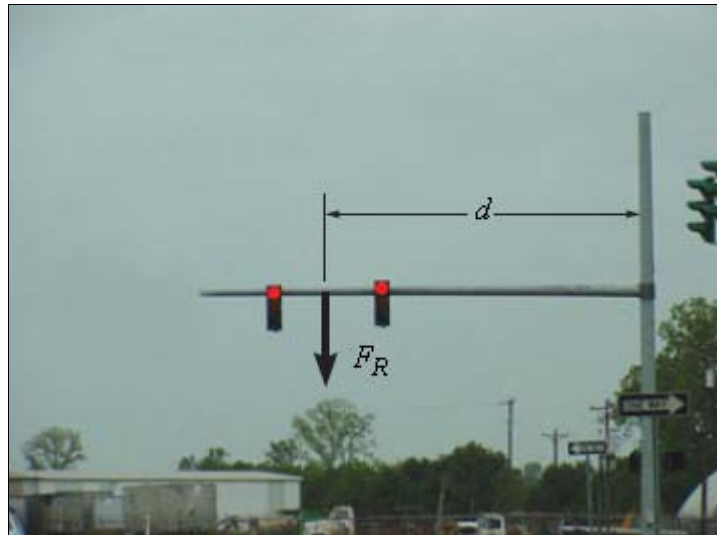


Figure: 30.10 - Moment de force résultant

Maintenant que nous avons convenablement défini ce qu'était une force et un moment de force, nous pouvons de suite aborder l'étude de la statique des forces:

6.2. STATIQUE DES FORCES

La statique des forces est un domaine difficile à généraliser. La plupart des ouvrages se servent de nombreux exemples (comme les systèmes de poulies, les leviers, les équilibres, les frottements, etc.) afin d'amener le lecteur à assimiler la méthode d'analyse qu'il faut pour résoudre les problèmes relatifs à ce domaine de la mécanique classique. Loin d'être contre cette méthode, nous n'avons pas souhaité nous restreindre ou nous étendre (suivant les points de vue) à des exemples particuliers, mais avons préféré proposer une méthode d'analyse qui fonctionnerait à coup sûr.

Définitions:

D1. La "**statique des forces**" est le domaine de la physique qui étudie l'effet de la résultante de forces (ou moments de force) constantes au cours du temps, appliquées sur un corps ponctuel ou étendu.

D2. Quand la somme vectorielle de toutes les forces et moments de force est nulle, il n'y a aucun mouvement. Nous parlons alors d'un "**équilibre statique**" (mais les forces existent tout de même à l'intérieur du système) tel que les forces et moments de forces se compensent mutuellement:

$$\sum_{i=1}^n \vec{F}_i = \vec{0} \text{ ou/et } \sum_{i=1}^n \vec{M}_i = \vec{0} \quad (30.114)$$

Remarque: Les relations précédentes, nous montrent bien que ce n'est pas parce qu'un système est à l'équilibre statique qu'il n'est soumis à aucune force (la somme vectorielle des forces peut s'annuler mais les forces sont non nulles).

Corollaires:

C1. Lors de l'analyse d'un système de statique des forces, il faut toujours (!!!) travailler avec les composantes vectorielles des forces et moments de forces (de par la première loi de Newton).

C2. Il faut donc s'imposer un repère par rapport auquel seront exprimées toutes les composantes de

forces:

- Dans le cas d'un corps ponctuel sur lequel sont appliquées des forces, il faut assimiler l'origine du repère à la position du point.
- Si les lignes de prolongement de toutes les forces sur un corps étendu sont toutes concurrentes en un point donné, le système peut être considéré comme un corps ponctuel ramené à ce point.
- Si le corps est étendu et plongé dans un champ de forces (gravitationnel, électrostatique, magnétique...) isotrope, coplanaire et constant dans le temps, l'ensemble des forces imprimées peut se rapporter au centre de gravité.

Démonstration:

Nous avons vu lors de l'étude du calcul vectoriel (cf. chapitre de Calcul Vectoriel) que la somme des vecteurs d'un même ensemble, mis bout à bout (au niveau de la représentation imagée) ou additionnés algébriquement constitue ce que nous appelons la "résultante" du système de forces ou de moments de force:

$$\vec{R}_{\vec{F}} = \sum_{i=1}^n \vec{F}_i \quad \text{ou/et} \quad \vec{R}_{\vec{M}} = \sum_{i=1}^n \vec{M}_i \quad (30.115)$$

Il est clair qu'un point matériel est donc par définition à l'état statique si la résultante des forces concurrentes est nulle. Ainsi, un corps ponctuel est au repos (vitesse constante nulle) si la grandeur $\vec{R}_{\vec{F}}$ est nulle (voir les lois de Newton plus loin).

Cette condition ne suffit cependant pas pour un corps étendu (non ponctuel): celui-ci peut ne pas se déplacer (pas de mouvement par translation), mais tourner sur lui-même par application de forces en dehors de son centre de gravité (les forces sont alors des moments de forces agissant sur des points du corps en question).

Imaginons maintenant un ensemble de forces $\vec{F}_i$, chacune d'elles appliquée en un point de vecteur-position $\vec{r}_i$ d'un mobile étendu et toutes parallèles à une direction commune donnée, repérée par un vecteur unitaire $\vec{u}$. La résultante de ces forces est alors:

$$\vec{R}_{\vec{F}} = \sum_{i=1}^n \vec{F}_i = \sum_{i=1}^n F_i \vec{u} = \left(\sum_{i=1}^n F_i \right) \vec{u} \quad (30.116)$$

Remarque: La norme de la résultante est donc:

$$\|\vec{R}_{\vec{F}}\| = \sum_{i=1}^n F_i \quad (30.117)$$

De manière analogue, la somme vectorielle des moments parallèles s'écrit:

$$\vec{R}_{\vec{M}} = \sum_{i=1}^n \vec{r}_i \times \vec{F}_i = \sum_{i=1}^n \vec{r}_i \times (F_i \vec{u}) = \left(\sum_{i=1}^n \vec{r}_i F_i \right) \times \vec{u} \quad (30.118)$$

Recherchons maintenant, la position $\vec{r}_C$ d'un point fictif C , appelé le "**centre des forces**" tel que le moment de la résultante $\vec{R}_F$ appliquée au point C soit égal au moment total $\vec{R}_M$. En d'autres termes, $\vec{r}_C$ doit être la solution de l'équation vectorielle:

$$\vec{R}_M = \vec{r}_C \times \vec{R}_F \quad (30.119)$$

S'il est possible de trouver un tel point C , nous ne devons donc plus, en principe, calculer le moment individuel de chaque force et en faire la somme vectorielle. Il suffit plutôt, de déterminer la résultante $\vec{R}_F$ et d'évaluer son moment résultant appliqué au point fictif C .

En combinant les relations précédentes, nous avons:

$$\vec{R}_M = \vec{r}_C \times \vec{R}_F = \vec{r}_C \times \left(\sum_{i=1}^n F_i \vec{u} \right) = \left(\sum_{i=1}^n F_i \right) \vec{r}_C \times \vec{u} \quad (30.120)$$

À son tour, le vecteur $\vec{R}_M$ peut être substitué tel que:

$$\vec{R}_M = \left(\sum_{i=1}^n F_i \right) \vec{r}_C \times \vec{u} = \left(\sum_{i=1}^n \vec{r}_i F_i \right) \times \vec{u} \quad (30.121)$$

d'où nous tirons finalement:

$$\left(\sum_{i=1}^n F_i \right) \vec{r}_C = \left(\sum_{i=1}^n \vec{r}_i F_i \right) \Rightarrow \vec{r}_C = \frac{\sum_{i=1}^n \vec{r}_i F_i}{\sum_{i=1}^n F_i} \quad (30.122)$$

comme $F_i = m_i a_i$ (deuxième loi de Newton) supposons maintenant (cas particulier) que $\forall i, j, a_i = a_j$ nous pouvons alors écrire ce résultat très important:

$$\vec{r}_C = \frac{\cancel{a_i} \sum_{i=1}^n \vec{r}_i m_i}{\cancel{a_i} \sum_{i=1}^n m_i} = \frac{\sum_{i=1}^n \vec{r}_i m_i}{\sum_{i=1}^n m_i} = \vec{r}_G \quad (30.123)$$

□C.Q.F.D.

C3. De par la troisième loi de Newton, tout corps solide rigide en équilibre stable, en contact avec un ensemble de corps solides rigides en équilibre stable eux aussi, subissent tous une force égale identique en chaque point de contact (identiquement répartie) mais opposée par ces derniers (assimilable et passant par leur centre de gravité lorsque c'est un champ de vecteurs isotrope et constant qui est à l'origine du contact). Ainsi:

- les repères des forces d'action/réaction doivent être placés sur les différents points de contact lorsque ce sont une quantité dénombrable de forces qui en sont à l'origine.
- les repères des forces d'action/réaction doivent être placés au centre de masse ou de gravité si les

forces à l'origine du contact (in extenso: de l'accélération) sont à l'origine d'un champ vectoriel gravifique, respectivement électrostatique/magnétique.

7. BALISTIQUE

Le mouvement parabolique est le mouvement d'un mobile animé, dans le champ de la pesanteur, d'une vitesse de translation $\vec{v}_0$ non parallèle à l'accélération de la pesanteur $\vec{g}$. Par exemple un projectile possédant au départ une vitesse $\vec{v}_0$ inclinée d'un angle β par rapport à l'horizontale.

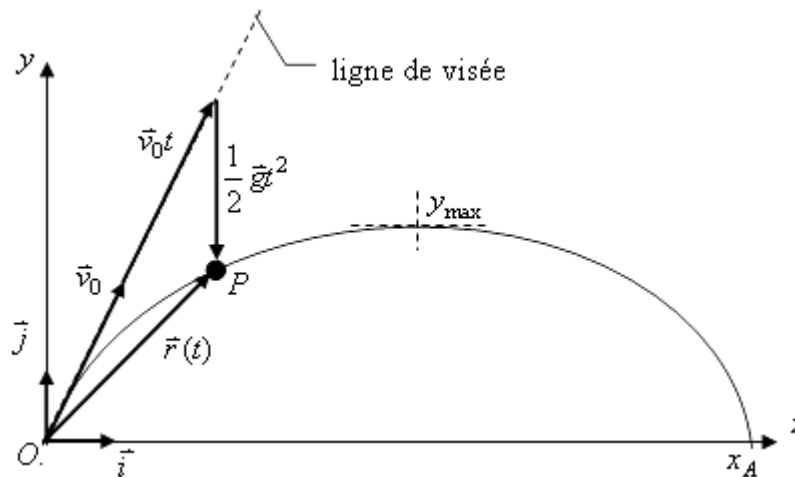


Figure: 30.11 - Exemple de trajectoires balistique

En l'absence de pesanteur et de frottement le mobile P suivrait la ligne de visée indéfiniment. L'action de la pesanteur est de le redescendre, au temps t , de la valeur connue $\frac{1}{2} g t^2$.

Nous posons la projection sur les axes:

$$\vec{r}(t) = v_0 \cos(\beta) \cdot t \cdot \vec{i} + (v_0 \sin(\beta) \cdot t - \frac{1}{2} g t^2) \vec{j} \quad (30.124)$$

combinaison d'un déplacement régulier selon x et d'un mouvement de chute avec vitesse initiale v_{0y} selon y . Ce qui correspond aux équations suivantes:

$$x(t) = v_0 \cos(\beta) t \text{ et } y(t) = v_0 \sin(\beta) t - \frac{1}{2} g t^2 \quad (30.125)$$

en éliminant le temps entre ces deux équations nous obtenons la trajectoire (équation d'une parabole):

$$y = x \cdot \tan(\beta) - \frac{g}{2 v_0^2 \cos^2(\beta)} x^2 \quad (30.126)$$

qui avait déjà été obtenue par Galilée au début du 17ème siècle.

Nous calculons ainsi la portée x_A du projectile en posant $y = 0$ dans l'équation ci-dessus et nous obtenons facilement:

$$x_A = \frac{v_0^2}{g} \sin(2\beta) \quad (30.127)$$

la solution $x = 0$ n'a aucun intérêt.

La hauteur maximale $y_{\max}$ peut être calculée en annulant la dérivée de l'équation de la trajectoire. Ainsi nous obtenons facilement:

$$y_{\max} = \frac{v_0^2}{2g} \sin^2 \beta \quad (30.128)$$

Nous remarquons pour la portée maximale que pour une vitesse initiale donnée et un objectif x_A à atteindre nous avons deux cas typiques dans la pratique:

1. Nous nous sommes donné une portée maximale x_A inaccessible car il n'est pas possible que $\sin(2\beta) > 1$.
2. Une seule valeur $\beta = \pi/4$ donne la portée maximale possible pour une vitesse initiale donnée.

La courbe enveloppant toutes les paraboles, tracée pour une vitesse v_0 donnée dans toutes les directions possibles, est encore une parabole, appelée "**parabole de sûreté**". Sa rotation autour de l'axe y engendre un paraboloides qui circonscrit (contient) la région de l'espace seule accessible aux projectiles.

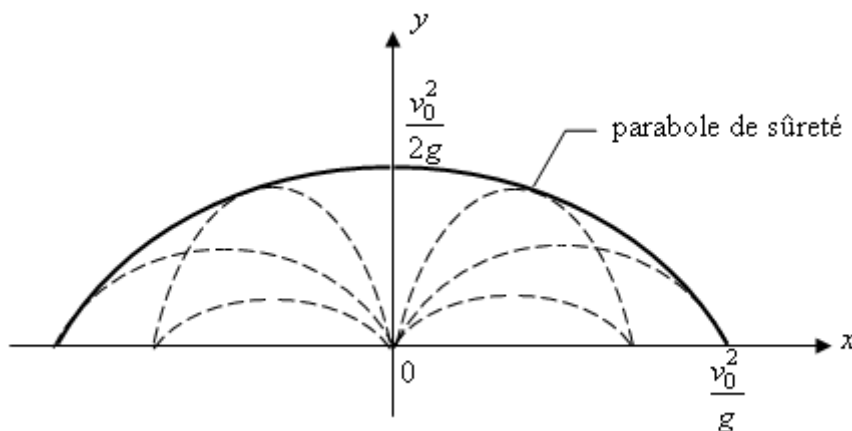


Figure: 30.12 - Illustration de quelques paraboles de sûreté

Ainsi, il n'est pas trop difficile de trouver l'équation de cette parabole de sûreté:

Le tir à la verticale $\beta = \pi/2$ nous est connu et est donné par

$$v_0 = \sqrt{2 \cdot g \cdot y_{\max}} \Leftrightarrow y_{\max} = \frac{v_0^2}{2g} \quad (30.129)$$

La portée maximale est quant à elle donnée par:

$$x_A = \frac{v_0^2}{g} \quad (30.130)$$

Donc quand $x = x_A \Rightarrow y = 0$ tel que:

$$y = \frac{v_0^2}{2g} - \frac{g}{2v_0^2} x^2 \quad (30.131)$$

qui est l'équation de la parabole de sûreté.

8. CINÉMATIQUE DE ROTATION

Les mouvement circulaires, appelés aussi "mouvements de rotation", décrivent donc la rotation d'un objet autour d'un axe (ou d'un point pour faire plus simple). L'usage veut qu'on le définisse par les données suivantes:

- la direction de l'axe du plan de rotation dans l'espace
- le sens de rotation sur le cercle de rayon constant autour de cet axe
- la vitesse de rotation v

tout ceci se résume avec la figure suivante:

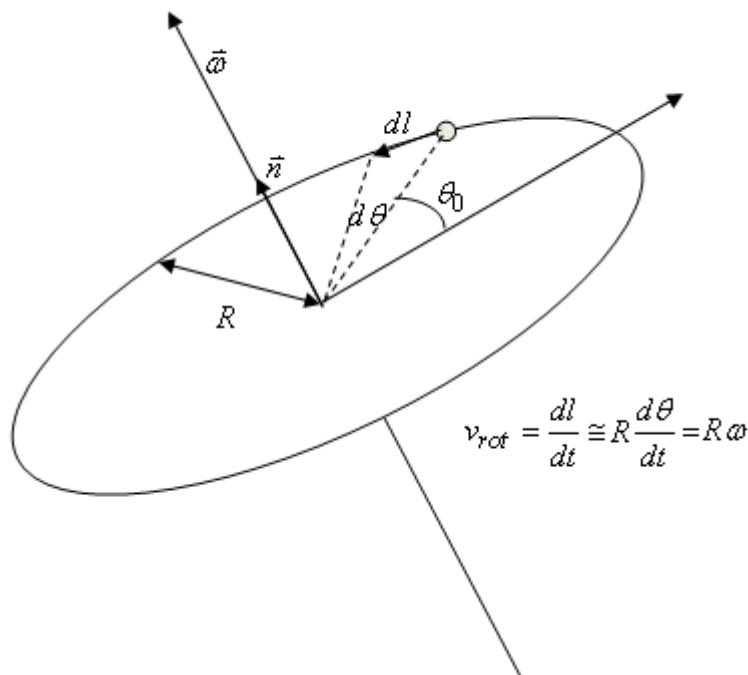


Figure: 30.13 - Illustration de quelques paraboles de sûreté

où nous avons utilisé la relation démontrée dans le chapitre de Trigonométrie:

$$dl \cong R d\theta \quad (30.132)$$

Nous résumons ces trois indications par la donnée d'un vecteur "**vitesse angulaire**" instantanée:

$$\vec{\omega} = \omega \cdot \vec{n} \quad (30.133)$$

Le sens de rotation est dit positif lorsque, le pouce dressé dans la direction du vecteur unitaire $\vec{n}$, nous saisissons l'axe de la main droite et que l'on voit tourner l'objet dans le sens des quatre autres doigts.

La norme de la vitesse angulaire instantanée, représente l'angle parcouru par unité de temps, par l'objet qui se déplace dans le plan perpendiculaire au vecteur unitaire $\vec{n}$:

$$\omega = \frac{d\theta}{dt} \quad (30.134)$$

Bien évidemment, il va de soit que la vitesse angulaire est donnée en radians par seconde et non en tours par minute ou en degrés par seconde. Il faut donc prendre garde à toujours faire la conversion qu'il convient!

Remarques:

R1 Dans le cas général du mouvement circulaire, la vitesse angulaire de l'objet étudié varie au cours du temps: $\omega(t)$.

R2. Lorsque la direction de l'axe change, les composantes du vecteur unitaire $\vec{n}(t)$ sont également des fonctions du temps. C'est le cas d'une roue de moto dans un virage.

Suite à la demande d'un lecteur, avant d'aller plus loin, remontrons que nous avons bien:

$$dl \cong R \cdot d\theta \quad (30.135)$$

En tournant d'un angle $d\theta$, un point de l'objet situé à une distance R de l'axe de rotation décrit un arc de cercle de longueur:

$$l = R \cdot \sin(\theta) = R \cdot \left(\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \dots \right) \quad (30.136)$$

où nous avons utilisé le développement en série de Maclaurin (cf. chapitre de Suites Et Séries).

Donc dans le cas des petits angles:

$$l \cong R \cdot \theta \Rightarrow dl = R \cdot d\theta \quad (30.137)$$

Si dt est le temps nécessaire à ce mouvement, la vitesse curviligne du point est donc bien:

$$v = \frac{dl}{dt} = R \cdot \frac{d\theta}{dt} = \omega \cdot R \quad (30.138)$$

nous retrouvons alors le résultat déjà donné plus haut.

Faisons maintenant de même que lorsque notre étude du mouvement rectiligne uniforme et déterminons la position angulaire en fonction du temps. Nous avons alors:

$$\omega = \frac{d\theta}{dt} \Rightarrow d\theta = \omega dt \quad (30.139)$$

Il vient alors:

$$\int_{\theta_0}^{\theta} d\theta = \int_{t_0}^t \omega dt \Rightarrow \theta - \theta_0 = \omega t - \omega t_0 \quad (30.140)$$

et si $t_0 = 0$, nous avons alors:

$$\theta = \theta_0 + \omega t \quad (30.141)$$

à comparer avec la relation équivalente entre position et vitesse obtenue lors de notre étude de la cinématique rectiligne.

Maintenant, considérons la définition de "**l'accélération angulaire**" (dont la notation traditionnelle un peu malheureuse...):

$$\alpha = \frac{d\omega}{dt} \quad (30.142)$$

Nous avons alors:

$$\alpha = \frac{d\omega}{dt} \Rightarrow d\omega = \alpha dt \quad (30.143)$$

et donc:

$$\int_{\omega_0}^{\omega} d\omega = \int_{t_0}^t \alpha dt \Rightarrow \omega - \omega_0 = \alpha t - \alpha t_0 \quad (30.144)$$

et si $t_0 = 0$, nous avons alors:

$$\omega = \omega_0 + \alpha t \quad (30.145)$$

Ce que nous pouvons noter:

$$\frac{d\theta}{dt} = \omega_0 + \alpha t \quad (30.146)$$

Il vient alors:

$$\int_{t_0}^t d\theta = \int_{t_0}^t \omega_0 dt + \int_{t_0}^t \omega t dt \quad (30.147)$$

soit:

$$\theta(t) - \theta_0 = \omega_0 (t - t_0) + \frac{1}{2} \omega (t^2 - t_0^2) \quad (30.148)$$

et si $t_0 = 0$, nous avons alors:

$$\theta(t) = \theta_0 + \omega_0 t + \frac{1}{2} \omega t^2 \quad (30.149)$$

à comparer avec la relation équivalente entre position, vitesse et accélération obtenue lors de notre étude de la cinématique rectiligne.

Intéressons-nous maintenant à l'aspect vectoriel du mouvement circulaire qui sera extrêmement important un peu plus loin et aussi dans de nombreux autres chapitres. Ainsi, donnons nous un repère euclidien orthonormé tel que:

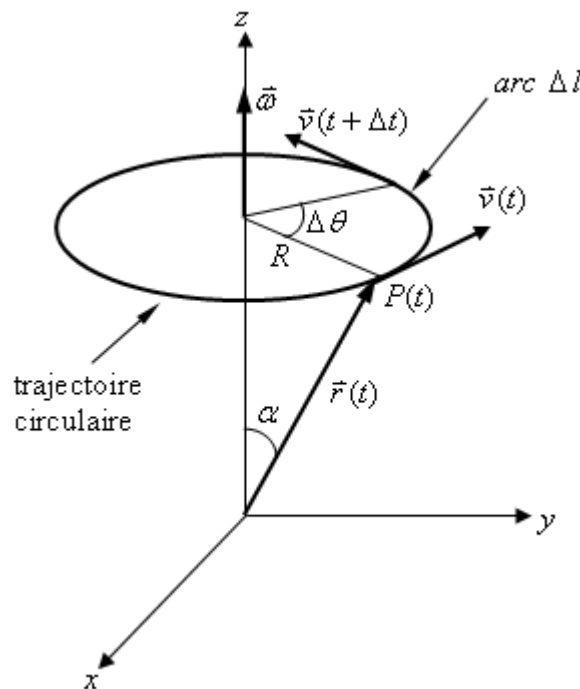


Figure: 30.14 - Exemple de mouvement circulaire autour d'un axe

Nous voyons bien sur cette figure que:

$$\|\vec{r}(t)\| \cdot \sin \alpha = R \quad (30.150)$$

Donc finalement nous avons:

$$v = \omega \cdot R = \omega \cdot \|\vec{r}(t)\| \cdot \sin \alpha = \|\vec{\omega}\| \cdot \|\vec{r}(t)\| \cdot \sin \alpha \quad (30.151)$$

Nous voyons alors que nous avons affaire à un produit vectoriel tel que:

$$\vec{v} = \vec{\omega} \times \vec{r}(t) \quad (30.152)$$

Nous avons donc:

$$\vec{a} = \frac{d\vec{v}}{dt} = \frac{d(\vec{\omega} \times \vec{r})}{dt} = \frac{d\vec{\omega}}{dt} \times \vec{r} + \vec{\omega} \times \frac{d\vec{r}}{dt} \quad (30.153)$$

que nous écrivons également:

$$\vec{a} = a_{tg} + a_n \quad (30.154)$$

L'accélération du mouvement circulaire est formée dans le cas général, de deux termes, le premier étant "l'accélération tangentielle" exprimant toujours la variation de la vitesse sur la trajectoire et le deuxième l'accélération perpendiculaire le long du rayon appelée également "accélération centripète" (centripète signifiant: "qui tend à rapprocher du centre").

Remarque: Si nous exprimons le mouvement circulaire du point P à partir d'un système d'axes situés dans le plan de la trajectoire, pour simplifier, alors, la position du point P est donnée par:

$$\vec{r}(t) = r \cdot \cos(\omega t) \cdot \vec{i} + r \cdot \sin(\omega t) \cdot \vec{j} = r \cdot \sin\left(\omega t + \frac{\pi}{2}\right) \cdot \vec{i} + r \cdot \sin(\omega t) \cdot \vec{j} \quad (30.155)$$

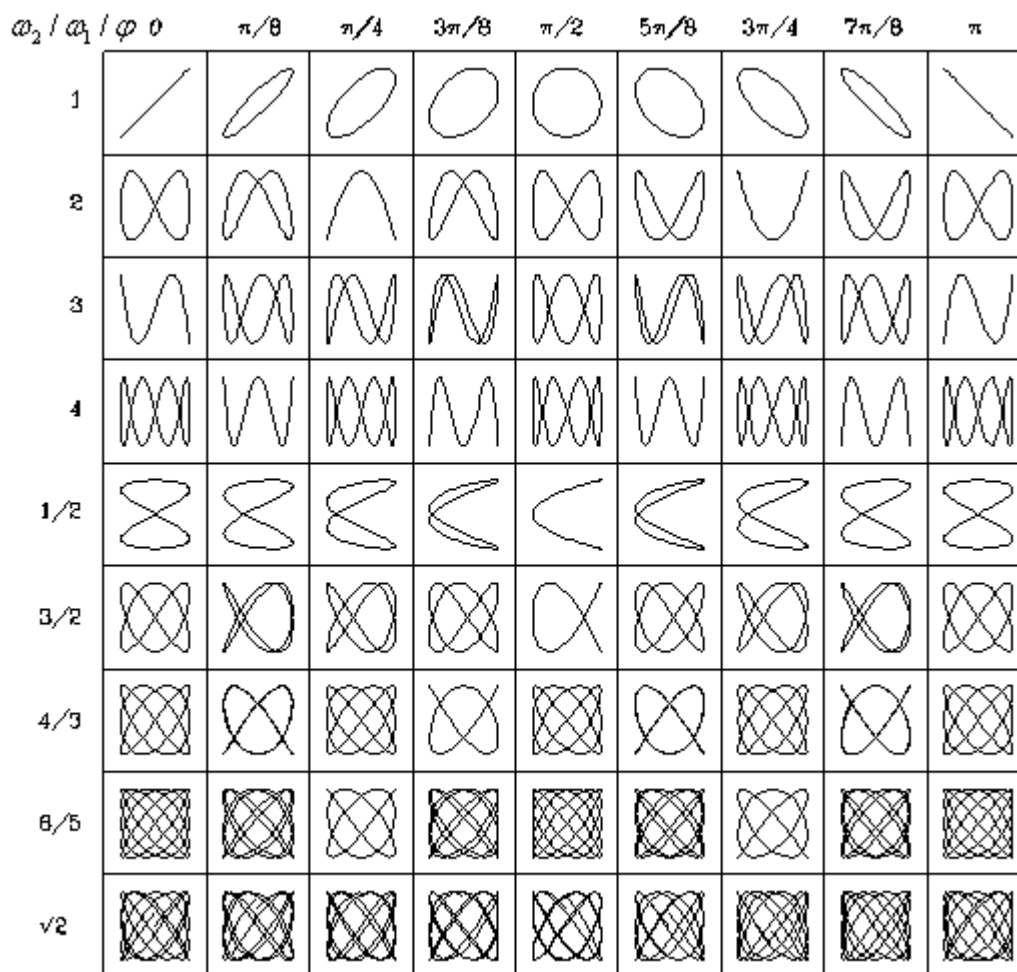
Ce qui montre que le mouvement circulaire peut être considéré comme la superposition de deux mouvements sinusoïdaux déphasés de $\pi/2$. Mais si nous écrivons:

$$\vec{r}(t) = r_1 \cdot \sin\left(\omega t + \frac{\pi}{2}\right) \cdot \vec{i} + r_2 \cdot \sin(\omega t) \cdot \vec{j} \quad (30.156)$$

ce qui est tout à fait envisageable pour une trajectoire imparfaitement circulaire et que que nous regardons les différentes caractéristiques paramétriques:

$$\begin{aligned} x &= A \sin(\omega_x t) \\ y &= B \sin(\omega_y t + \varphi) \end{aligned} \quad (30.157)$$

en faisant varier le déphasage φ et le rapport ω_x / ω_y nous obtenons des courbes que nous appelons des "**figures de Lissajous**":



Le lecteur retrouvera des applications pratiques très importantes de la cinématique du mouvement circulaire pour l'industrie relativement à la mécanique dans le chapitre de Génie Mécanique (section Ingénierie).

9. TRAVAIL ET ÉNERGIE

Si un point de masse m subit un déplacement élémentaire $d\vec{r}$ sous l'effet d'une force $\vec{F}$, cette force effectue un travail élémentaire valant par définition:

$$dW = \vec{F} \circ d\vec{r} \quad (30.158)$$

Si cette masse m est déplacée d'un endroit A à un endroit B , le travail total est:

$$W_{A,B} = \int_A^B \vec{F} \circ d\vec{r} \quad (30.159)$$

Pour les unités, nous avons:

$$[W] = N \cdot m = J$$

où J est la première lettre de "Joules".

De la définition ci-dessous nous pouvons rapidement déduire le travail d'un moment de force (ou autrement dit: le travail d'une force dans un mouvement de rotation) puisque sur un élément infinitésimal de déplacement, nous avons:

$$dW = \vec{F} \times \vec{r} \cdot d\vec{\theta} = \vec{M} \cdot d\vec{\theta} \quad (30.160)$$

Donc dans le cas d'un mouvement circulaire le travail d'un moment de force sera:

$$W_{A,B} = \int_{\theta_A}^{\theta_B} \vec{M} \cdot d\vec{\theta} \quad (30.161)$$

Remarques:

R1. Si W est positif le travail est dit "**travail moteur**". Dans le cas contraire il est dit "**travail résistant**" (exemple: le freinage).

R2. Si la force $\vec{F}$ est constante en grandeur et en direction (cas de la pesanteur au voisinage de la surface terrestre), l'intégrale du calcul de W prend une forme plus simple:

$$W_{A,B} = \int_A^B \vec{F} \circ d\vec{r} = \vec{F} \circ \int_A^B d\vec{r} = \vec{F} \circ (\vec{r}_B - \vec{r}_A) \quad (30.162)$$

Ce résultat montre que le travail ne dépend alors que des positions initiale et finale et pas du chemin parcouru. Le travail de la pesanteur est un cas particulier de ce type.

9.1. ÉNERGIE CINÉTIQUE

La loi de Newton $\vec{F} = m\vec{a}$ est applicable le long du chemin A-B. En l'utilisant dans l'expression du travail il vient:

$$W_{A,B} = \int_A^B \vec{F} \circ d\vec{r} = \int_A^B m \frac{d\vec{v}}{dt} \circ d\vec{r} = \int_A^B m \frac{d\vec{v}}{dt} \circ \vec{v} dt = m \int_A^B \vec{v} \circ d\vec{v} \quad (30.163)$$

et, en développant le produit scalaire au moyen des composantes, nous aurons:

$$W_{A,B} = m \int_A^B (v_x dv_x + v_y dv_y + v_z dv_z) = \frac{1}{2} m (v_x^2 + v_y^2 + v_z^2) \Big|_A^B = \frac{1}{2} m v_B^2 - \frac{1}{2} m v_A^2 = E_c(B) - E_c(A) \quad (30.164)$$

Lorsqu'un corps se déplace sous l'action d'une force résultante $\vec{F}$ quelconque, le travail de cette force d'accélération sur un chemin quelconque A, B est égal à la variation d'énergie cinétique du corps:

Par définition, la relation:

$$E_c = \frac{1}{2} m v^2 \quad (30.165)$$

est appelée "**l'énergie cinétique**" et elle se mesure en "Joules" (ou d'autres unités dérivées exotiques dont les physiciens théoriciens abusent parfois un petit peu trop...) et est toujours positive en mécanique ou n'importe quel autre domaine de la physique.

L'équation:

$$W_{A,B} = E_c(B) - E_c(A) \quad (30.166)$$

porte quelquefois le nom de "**théorème de l'énergie cinétique**".

9.1.1. MOMENT D'INERTIE

Pour un solide rigide tournant autour d'un axe à la vitesse angulaire $\vec{\omega}$, l'énergie cinétique élémentaire d'un point quelconque de masse dm , situé hors de l'axe, vaut:

$$dE_c = \frac{1}{2} v^2 dm = \frac{1}{2} (\vec{\omega} \times \vec{r})^2 dm = \frac{1}{2} (\omega \cdot r \cdot \sin \alpha)^2 dm = \frac{1}{2} (\omega \cdot r)^2 dm \quad (30.167)$$

puisque $\vec{\omega}$ et $\vec{r}$ sont perpendiculaires. L'énergie cinétique totale est alors:

$$E_c = \int dE = \frac{1}{2} \omega^2 \int r^2 dm \quad (30.168)$$

Nous avons pris l'habitude en physique de noter cette dernière relation:

$$E_c = \frac{1}{2} \omega^2 J \quad (30.169)$$

où par définition, le "moment d'inertie" est:

$$J = \int r^2 dm \quad (30.170)$$

Exemples:

E1. Calculons la vitesse finale d'une boule chutant sur un plan incliné de frottement non nul (elle va donc tourner) dans un champ de potentiel gravifique.

La réponse du néophyte en physique sera souvent obtenue en ne considérant que l'énergie cinétique mais pas la vitesse de rotation de la boule. Or, nous devons prendre celle-ci en compte via son moment d'inertie.

Nous avons donc l'énergie cinétique totale étant l'énergie cinétique de translation du centre de masse plus l'énergie de rotation autour de ce même centre de masse:

$$E_c = E_{transl.} + E_{rot.} = \frac{1}{2} m v_G^2 + \frac{1}{2} J_G \omega^2 = \frac{1}{2} v_G^2 \left(m + \frac{J_G}{r^2} \right) \quad (30.171)$$

en égalant cette valeur à l'énergie potentielle gravifique (voir plus loin) et en supposant une vitesse initiale nulle de la chute, nous avons:

$$mgh = \frac{1}{2} v_G^2 \left(m + \frac{J_G}{r^2} \right) \quad (30.172)$$

Soit la vitesse acquise au bas du plan (frottement de roulement non-compris...):

$$v_G = \sqrt{\frac{2gh}{1 + \frac{J_G}{mr^2}}} \quad (30.173)$$

et nous avons démontré dans le chapitre sur les Formes Géométriques que le moment d'inertie d'une boule pleine était:

$$J_G = \frac{2}{5} m r^2 \quad (30.174)$$

Il vient alors:

$$v_G = \sqrt{\frac{2gh}{1 + \frac{2}{5}}} = \sqrt{\frac{5}{7} 2gh} = \sqrt{\frac{10}{7} gh} \quad (30.175)$$

Nous voyons dans le cas particulier de la boule, que la vitesse finale de chute est (sans frottements de l'air ni de roulement) indépendante de sa masse et de son rayon (qu'elle soit creuse ou pleine) ce qui est relativement contre intuitif.

E2. Un deuxième exemple fameux est le calcul de l'énergie cinétique de rotation d'une planète parfaitement sphérique de masse homogène et de période de rotation constante. Nous avons alors:

$$\begin{aligned}
 E_c &= \frac{1}{2} \omega^2 J = \frac{1}{2} \omega^2 \left(\frac{2}{5} MR^2 \right) = \frac{2}{10} \omega^2 MR^2 = \frac{1}{5} \left(\frac{v}{R} \right)^2 MR^2 = \frac{1}{5} v^2 M \\
 &= \frac{1}{5} v^2 (\rho V) = \frac{1}{5} v^2 \left(\rho \frac{4}{3} \pi R^3 \right) = \frac{4\pi}{15} \rho R^3 v^2 = \frac{4\pi}{15} \rho R^3 \left(\frac{2\pi R}{T} \right)^2 \\
 &= \frac{16\pi^3}{15} \rho \frac{R^5}{T^2}
 \end{aligned} \quad (30.176)$$

Remarque: Dans un solide, la répartition de la matière autour d'un axe sera évidemment différente selon l'axe choisi. Le moment d'inertie correspondant sera aussi différent. Il est donc indispensable de préciser l'axe par rapport auquel nous souhaitons déterminer ce moment d'inertie. Nous observons dans la pratique que les ingénieurs placent souvent l'axe de façon à ce qu'il passe par le centre de masse. Dans les tables, nous trouvons fréquemment les expressions des moments d'inerties de formes courantes (selon un axe donné) telles que le cylindre, le cône, la sphère, la barre, le tube (cf. chapitre sur les Formes Géométriques).

Nous avons vu lors de notre étude du moment cinétique que:

$$\vec{b} = m r^2 \vec{\omega} \quad (30.177)$$

et le moment d'inertie étant donné par:

$$J = \int r^2 dm \quad (30.178)$$

Nous avons donc:

$$d\vec{b} = r^2 \vec{\omega} dm \quad (30.179)$$

d'où:

$$\vec{b} = \vec{\omega} \int r^2 dm \quad (30.180)$$

Nous obtenons finalement:

$$\boxed{\vec{b} = J \vec{\omega}} \quad (30.181)$$

c'est l'expression donnant le moment cinétique d'un corps tournant sur lui-même (sur un de ses axes possibles de rotation).

Etant donné que nous avons démontré lors de notre étude du moment cinétique que:

$$\frac{d\vec{b}}{dt} = \vec{r} \times \vec{F} = \vec{M} \quad (30.182)$$

il vient alors dans l'hypothèse que la masse et la géométrie du solide restent constantes... que le moment de force est alors:

$$\vec{M} = \frac{d}{dt}(J \vec{\omega}) = J \frac{d\vec{\omega}}{dt} \quad (30.183)$$

et bien évidemment, si nous étudions un système dans lequel le moment cinétique est conservatif, il va de soi que:

$$\vec{M} = \vec{0} \Rightarrow \frac{d\vec{\omega}}{dt} = \vec{0} \Rightarrow \vec{\omega} = \vec{\omega}^* = J \vec{\omega} \quad (30.184)$$

Cette conservation du moment cinétique trouve une application dans une multitude d'expériences telle que celle connue qui consiste à se faire tourner sur une chaise et à écarter les mains ou les jambes ce qui fera diminuer la vitesse de rotation (et inversement).

Une autre expérience curieuse (mais mathématiquement correcte) consiste à se poser sur un plateau tournant avec une roue en rotation tenue à l'horizontale (le moment cinétique vertical est donc nul) et de mettre celle-ci ensuite à la verticale. Comme le moment cinétique vertical doit rester nul, pour contrecarrer cela, le plateau sur lequel est posé l'expérimentateur se mettra à tourner dans le sens inverse de rotation de la roue.

Les déplacements de masses importantes à la surface de la Terre (icebergs, crues des fleuves, plaques tectoniques, etc.) provoquent des variations du moment d'inertie de la Terre. Il s'ensuit des fluctuations de la vitesse angulaire donc une imperfection de l'étalon astronomique de temps (quelques millièmes par jour).

Revenons maintenant aux méthodes de calcul des moments d'inertie. L'énergie cinétique d'un corps étant la somme de l'énergie cinétique de chaque élément de ce corps, nous avons:

$$E_c = \sum_{i=1}^n E_{c_i} \quad (30.185)$$

Dans le cadre d'un corps solide rigide en rotation autour d'un axe, nous avons:

$$E_c = \sum_{i=1}^n E_{c_i} = \sum_{i=1}^n \frac{1}{2} J_i \omega^2 = \frac{1}{2} \omega^2 \sum_{i=1}^n J_i = \frac{1}{2} \omega^2 J \quad (30.186)$$

Ainsi, pour un corps composé d'un ensemble de corps de géométrie différentes, le moment d'inertie total est la somme des moments d'inertie par rapport à l'axe de rotation tel que:

$$\sum_{i=1}^n J_i = J \quad (30.187)$$

Lorsque nous calculons le moment d'inertie d'un corps par rapport à un axe Δ donné, il peut être intéressant de savoir qu'elle est la distance à l'axe où nous pouvons placer fictivement toute la masse de ce corps pour avoir le même moment d'inertie. Par définition, cette distance notée k et appelée le "**rayon de giration**" est trivialement donnée par:

$$J_{\Delta} = M k^2 \Rightarrow k = \sqrt{\frac{J_{\Delta}}{M}} \quad (30.188)$$

où J_{Δ} est le moment d'inertie connu du corps de masse M par rapport à un axe Δ .

Par définition, le "**moment d'inertie polaire**" (ou également "**moment d'inertie quadratique**") est le moment d'inertie défini par rapport à un point (le pôle) et non plus par rapport à un axe et noté:

$$J_o = \int r^2 dm \quad (30.189)$$

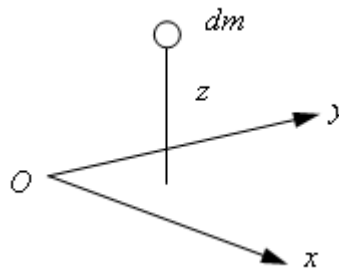
Cette grandeur n'intervient en fait que pour les rotations libres et n'a d'intérêt, pour les rotations autour d'un axe fixe, que parce qu'elle facilite quelquefois le calcul des moments d'inertie axiaux en vertu de la relation suivante (en coordonnées cartésiennes):

$$J_o = \frac{1}{2} (J_x + J_y + J_z) \quad (30.190)$$

Démonstration:

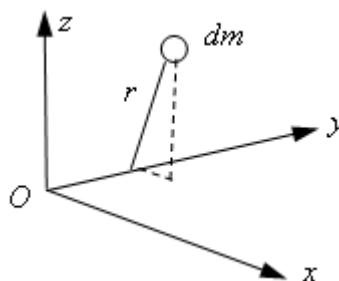
Lemme 1: Le moment d'inertie par rapport à un plan xOy est donné trivialement par:

$$J_{x,y} = \int z^2 dm \quad (30.191)$$



Lemme 2: Le moment d'inertie par rapport à un axe est donné par:

$$\begin{aligned} J_x &= \int r^2 dm = \int (y^2 + z^2) dm \\ J_y &= \int (x^2 + z^2) dm \\ J_z &= \int (x^2 + y^2) dm \end{aligned} \quad (30.192)$$

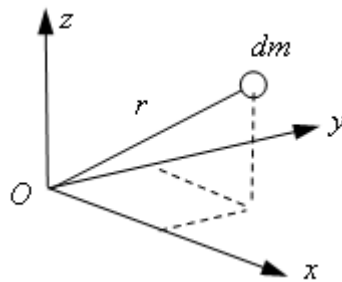


En sommant ces relations, nous en déduisons:

$$J_x + J_y + J_z = 2 \int (x^2 + y^2 + z^2) dm \quad (30.193)$$

Le moment d'inertie polaire est alors donnée par:

$$J_0 = \int r^2 dm = \int (x^2 + y^2 + z^2) \quad (30.194)$$



En comparant avec le lemme 2 il vient:

$$J_0 = \frac{1}{2} (J_x + J_y + J_z) \quad (30.195)$$

□C.Q.F.D.

Si le corps en question à une symétrie sphérique, il vient de suite puisque $J_x = J_y = J_z \equiv J$ que:

$$J_0 = \frac{1}{2} (3J) \Rightarrow J = \frac{3}{2} J_0 \quad (30.196)$$

Un exemple est donné avec la boule (sphère pleine) dans le chapitre traitant des Formes Géométriques dans la section de géométrie.

Supposons maintenant connaître le moment d'inertie J_x d'un corps solide rigide quelconque par rapport à un axe $z \perp x, y$ (cet axe n'étant pas nécessairement uniquement assimilé à l'axe z commun) passant par le centre de masse G. Calculons ensuite le moment d'inertie $J_{z'}$, par rapport à un autre axe z' , parallèle à z et distant de a , et faisons apparaître la liaison existant entre ces deux moments d'inertie différents:

Dans un référentiel cartésien, nous avons pour tout point (x,y):

$$r^2 = x^2 + y^2 \text{ et } r'^2 = x^2 + (y+a)^2 \quad (30.197)$$

Nous avons alors:

$$\begin{aligned} J_z &= \int r^2 dm = \int (x^2 + y^2) dm \\ J_{z'} &= \int r'^2 dm = \int (x^2 + (y+a)^2) dm = \int (x^2 + y^2 + 2ay + a^2) dm \\ &= \int (x^2 + y^2) dm + \int (2ay + a^2) dm = J_z + 2a \int y dm + a^2 \int dm = J_z + 2a \int y dm + a^2 M \end{aligned} \quad (30.198)$$

Le terme:

$$2a \int y dm \quad (30.199)$$

est nul car si le moment d'inertie est calculé par rapport au centre de masse G comme nous l'avons imposé dès le début, alors:

$$y_G = \frac{\int y dV}{\int dV} = 0 \quad (30.200)$$

En définitive, nous obtenons finalement le "**théorème d'Huygens-Steiner**":

$$J_{x'} = J_x + a^2 M \quad (30.201)$$

Comme nous le verrons dans le chapitre des Formes Géométriques dans la section de géométrie du site, il devient alors facile de pouvoir calculer le moment d'inertie d'un triangle équilatéral en connaissant celui d'une plaque carrée et en déplaçant l'axe d'inertie au point où se situe le centre de gravité du triangle (soit au tiers de la médiane située entre le centre du rectangle et un des sommets du rectangle).

Comme il existe autant de moments d'inertie que d'axes de rotation et que ces derniers sont souvent dans les cas d'études assimilés aux axes principaux d'inertie (axe assimilés aux axes de révolutions ou aux plans de symétrie - voir plus loin), il peut être utile d'introduire un être mathématique utile dans le cadre de représentation des moments d'inertie qui n'est autre que la "**matrice d'inertie**" ou appelé encore (formulation plus moderne) "**tenseur d'inertie**".

La démarche pour déterminer rigoureusement l'expression de ce tenseur est la suivante: soit A_i un point donné d'un solide dont nous cherchons à calculer le moment d'inertie et Δ l'axe d'origine O et de vecteur unité $\vec{u}$ par rapport auquel nous souhaitons calculer le moment d'inertie. Tout point A_i du solide peut être projeté (projection orthogonale) sur un point $H \in \Delta$ à partir de la connaissance de l'angle θ entre Δ et $\overrightarrow{OA_i}$ tel que:

$$\overline{HA} = \overline{OA} \sin \theta \Rightarrow \overline{HA}^2 = \overline{OA_i}^2 \sin^2 \theta = \|\overrightarrow{OA_i} \times \vec{u}\|^2 \quad (30.202)$$

Dès lors:

$$J_{\Delta} = \int r^2 dm = \int \|\overrightarrow{OA_i} \times \vec{u}\|^2 dm = \int \|\overrightarrow{OA_i} \times \vec{u}\| \|\overrightarrow{OA_i} \times \vec{u}\| dm = \int (\overrightarrow{OA_i} \times \vec{u}) \circ (\overrightarrow{OA_i} \times \vec{u}) dm \quad (30.203)$$

D'après les propriétés du produit mixte (cf. chapitre de Calcul Vectoriel) et du produit scalaire:

$$(\vec{x} \times \vec{y}) \circ \vec{z} = \vec{x} \circ (\vec{y} \times \vec{z}) \text{ et } (\vec{x} \times \vec{y}) \circ \vec{z} = \vec{z} \circ (\vec{x} \times \vec{y}) \quad (30.204)$$

nous avons:

$$(\overrightarrow{OA_i} \times \vec{u}) \circ (\overrightarrow{OA_i} \times \vec{u}) = \vec{w} \circ (\overrightarrow{OA_i} \times \vec{u}) = (\vec{w} \times \overrightarrow{OA_i}) \circ \vec{u} = \vec{u} \circ (\vec{w} \times \overrightarrow{OA_i}) = \vec{u} \circ ((\overrightarrow{OA_i} \times \vec{u}) \times \overrightarrow{OA_i}) \quad (30.205)$$

et donc:

$$J_{\Delta} = \int \vec{u} \circ ((\overrightarrow{OA_i} \times \vec{u}) \times \overrightarrow{OA_i}) dm \quad (30.206)$$

Comme $\vec{u}$ est un vecteur de direction constante quel que soit le point d'intégration, nous pouvons le sortir de l'intégrale tel que:

$$J_{\Delta} = \vec{u} \circ \int \left((\overrightarrow{OA_i} \times \vec{u}) \times \overrightarrow{OA_i} \right) dm = \vec{u} \circ \vec{v} \quad (30.207)$$

Nous pouvons vérifier que si nous remplaçons $\vec{u}$ par $\vec{u}_1 + \vec{u}_2$, nous obtiendrons un résultat $\vec{v} = \vec{v}_1 + \vec{v}_2$ de par la propriété de linéarité du produit vectoriel (cf. le chapitre de Calcul Vectoriel). Ainsi, l'application qui à $\vec{u}$ associe $\vec{v}$ est donc une application linéaire qui peut être représentée, dans une base B donnée, par une matrice:

$$\vec{u} \rightarrow \vec{v} = (J_o)_{/B} \vec{u} \quad (30.208)$$

La matrice $(J_o)_{/B}$ est le donc "tenseur d'inertie" du système par rapport au point O , dans la base B .

Le moment d'inertie d'un système par rapport à un axe Δ quelconque de vecteur unitaire $\vec{u}$ est donné par:

$$\forall O \in \Delta, J_{\Delta} = \vec{u} \circ \left[(J_o)_{/B} \vec{u} \right] \quad (30.209)$$

Le problème est donc maintenant de pouvoir calculer les éléments du tenseur (J_o) , pour une base B donnée. Soit un repère $(O, \vec{i}, \vec{j}, \vec{k})$ tel que $O \in \Delta$. Nous posons:

$$\vec{u} = \alpha \vec{i} + \beta \vec{j} + \gamma \vec{k} \text{ et } \overrightarrow{OA} = x \vec{i} + y \vec{j} + z \vec{k} \quad (30.210)$$

En utilisant le fait qu'un produit vectoriel puisse être représenté par une matrice antisymétrique (vérifiez c'est facile):

$$\overrightarrow{OA_i} \times \vec{u} = \begin{pmatrix} 0 & -z & y \\ z & 0 & -x \\ -y & x & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} \quad (30.211)$$

nous avons:

$$\overrightarrow{OA_i} \times (\overrightarrow{OA_i} \times \vec{u}) = \begin{pmatrix} 0 & -z & y \\ z & 0 & -x \\ -y & x & 0 \end{pmatrix} \begin{pmatrix} 0 & -z & y \\ z & 0 & -x \\ -y & x & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} -z^2 - y^2 & xy & xz \\ xy & -z^2 - x^2 & yz \\ xz & yz & -y^2 - x^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} \quad (30.212)$$

et donc:

$$\begin{aligned}
(J_o)_{\vec{B}} \vec{u} &= \int ((\overrightarrow{OA_i} \times \vec{u}) \times \overrightarrow{OA_i}) dm = - \int (\overrightarrow{OA_i} \times (\overrightarrow{OA_i} \times \vec{u})) dm \\
&= - \int \begin{pmatrix} -z^2 - y^2 & xy & xz \\ xy & -z^2 - x^2 & yz \\ xz & yz & -y^2 - x^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} dm \\
&= \begin{pmatrix} \int (z^2 + y^2) dm & - \int xy dm & - \int xz dm \\ - \int xy dm & \int (z^2 + x^2) dm & - \int yz dm \\ - \int xz dm & - \int yz dm & \int (y^2 + x^2) dm \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}
\end{aligned} \quad (30.213)$$

Dans l'expression ci-dessus de la matrice d'inertie, nous reconnaissons les éléments diagonaux: il s'agit tout simplement des moments d'inertie du système par rapport aux différents axes de la base. Nous appelons "**produit d'inertie**" les éléments non-diagonaux de la matrice et nous les notons:

$$J_{xy} = - \int xy dm, J_{xz} = - \int xz dm, J_{yz} = - \int yz dm \quad (30.214)$$

Nous avons donc:

$$(J_o)_{(\vec{i}, \vec{j}, \vec{k})} = \begin{pmatrix} J_{Ox} & J_{xy} & J_{xz} \\ J_{xy} & J_{Oy} & J_{yz} \\ J_{xz} & J_{yz} & J_{Oz} \end{pmatrix} \quad (30.215)$$

Si O est assimilé au centre de masse du solide considéré, nous notons simplement:

$$(J_{Ox}, J_{Oy}, J_{Oz}) = (J_x, J_y, J_z) \quad (30.216)$$

Nous pouvons également généraliser le théorème d'Huygens en faisant usage de ce tenseur de symétrie. Pour ce faire, appelons (x', y', z') les coordonnées d'un point A quelconque dans R' et (x, y, z) ses coordonnées dans R . Nous appelons (a, b, c) les coordonnées de l'origine O' de R' dans R :

$$R \Rightarrow \begin{cases} x' = x - a \\ y' = y - b \\ z' = z - c \end{cases} \quad (30.217)$$

puisque:

$$\overrightarrow{OO'} + \overrightarrow{O'A} = \overrightarrow{OA} \Leftrightarrow \overrightarrow{O'A} = \overrightarrow{OA} - \overrightarrow{OO'} \quad (30.218)$$

Nous avons alors:

$$\begin{aligned}
J_{Oz} &= \int (x'^2 + y'^2) dm = \int ((x' + a)^2 + (y' + b)^2) dm \\
&= \int (x'^2 + y'^2) dm + 2a \int x' dm + a^2 \int dm + 2b \int y' dm + b^2 \int dm \\
&= J_{Oz'} + M(a^2 + b^2) + 2a \int x' dm + 2b \int y' dm
\end{aligned} \quad (30.219)$$

Or, si O' coïncide avec le centre de masse G , alors selon la définition du centre de masse:

$$\int \overline{GA} dm = \vec{0} \Rightarrow \begin{cases} \int x' dm = 0 \\ \int y' dm = 0 \\ \int z' dm = 0 \end{cases} \quad (30.220)$$

Nous en déduisons alors:

$$J_{Ox} = J_{Gx} + M(a^2 + b^2) \quad (30.221)$$

et de même:

$$J_{Ox} = J_{Gx} + M(b^2 + c^2), \quad J_{Oy} = J_{Gy} + M(a^2 + c^2) \quad (30.222)$$

avec:

$$\overrightarrow{OG} = a\vec{i} + b\vec{j} + c\vec{k} = x_G\vec{i} + y_G\vec{j} + z_G\vec{k} \quad (30.223)$$

Nous retombons sur le théorème d'Huyghens classique puisque $a^2 + b^2$ n'est autre que la distance au carré entre l'axe Oz et Gz et de même pour $b^2 + c^2$ qui est la distance au carré entre Ox et Gx et $a^2 + c^2$ qui est la distance entre Oy et Gy.

Si nous nous intéressons maintenant aux produits d'inertie, il vient:

$$J_{xy} = -\int xy dm = -\int (x' + a)(y' + b) dm = -\int x'y' dm - a \int y' dm - b \int x' dm - ab \int dm \quad (30.224)$$

d'où, si O' coïncide avec G:

$$J_{xy} = J_{x'y'} - Mab, \quad J_{xz} = J_{x'z'} - Mac, \quad J_{yz} = J_{y'z'} - Mbc \quad (30.225)$$

En résumé, le théorème d'Huygens généralisé, s'écrit:

$$(J_o)_{(\vec{i}, \vec{j}, \vec{k})} = (J_G)_{(\vec{i}, \vec{j}, \vec{k})} + \begin{pmatrix} M(y_G^2 + z_G^2) & -Mx_G y_G & -Mx_G z_G \\ -Mx_G y_G & M(x_G^2 + z_G^2) & -My_G z_G \\ -Mx_G z_G & -My_G z_G & M(x_G^2 + y_G^2) \end{pmatrix} \quad (30.226)$$

Le tenseur d'inertie étant réel et symétrique, nous avons vu dans le chapitre d'Algèbre Linéaire (théorème spectral) qu'il est toujours possible de trouver trois directions perpendiculaires de vecteurs $\vec{e}_1, \vec{e}_2, \vec{e}_3$ telles que le tenseur (matrice) symétrique soit diagonalisable:

$$(J_o)_{\vec{e}_1, \vec{e}_2, \vec{e}_3} = \begin{pmatrix} J_1 & 0 & 0 \\ 0 & J_2 & 0 \\ 0 & 0 & J_3 \end{pmatrix} \quad (30.227)$$

Le trièdre formée par les vecteurs $\vec{e}_1, \vec{e}_2, \vec{e}_3$ est appelé "**trièdre principal d'inertie**" et ses axes sont appelés "**axes principaux d'inertie**". Dans ce repère (J_o) prend le nom de "**tenseur principal d'inertie**". Si de plus O est assimilé à G, nous parlons de "**tenseur central d'inertie**".

En fait, pour trouver les moments d'inertie relativement aux axes principaux il n'est pratiquement jamais nécessaire de diagonaliser le tenseur d'inertie, car il suffit souvent de se laisser guider par la symétrie du système. Nous allons voir avec les théorèmes suivants que s'il existe des axes ou des plans de symétrie pour la distribution de masse, les axes d'inertie sont faciles à trouver. De plus, le système est en général suffisamment simple (ou décomposable en éléments suffisamment simples...) pour que ces axes soient évidents.

Premier théorème: Si le système possède un plan de symétrie matérielle (in extenso: $\rho(A) = \rho(A')$ si A symétrique de A' par rapport au plan) alors tout axe perpendiculaire à ce plan est axe principal d'inertie.

Démonstration:

Choisissons un repère xOy dans le plan par rapport auquel le système a une distribution de masse symétrique et un axe Oz perpendiculaire à ce plan. Pour calculer $\int xz dm$ ou $\int yz dm$, groupons les points par deux, symétriques par rapport à xOy . c'est-à-dire tels que $z_A = -z_{A'}$. Nous aurons alors:

$$x_A z_A + x_{A'} z_{A'} = 0 \quad (30.228)$$

et de même:

$$y_A z_A + y_{A'} z_{A'} = 0 \quad (30.229)$$

c'est-à-dire, puis (symétrie matérielle!):

$$dm(A) = dm(A') \quad (30.230)$$

que toutes les contributions de paires de points symétriques sont nulles, ce qui implique: $J_{xz} = J_{yz} = 0$, c'est-à-dire que l'axe des z est direction principale d'inertie.

□C.Q.F.D.

Deuxième théorème: Choisissons comme axe Oz l'axe de symétrie. De même que ci-dessus, nous avons:

$$\int xz dm = \int yz dm = 0 \quad (30.231)$$

Démonstration:

Effectivement, car si nous groupons les points par paire A' et A symétriques par rapport à Oz , nous avons:

$$x_A = -x_{A'}, y_A = -y_{A'} \quad (30.232)$$

mais $z_A = z_{A'}$ donc toujours:

$$x_A z_A + x_{A'} z_{A'} = 0 \quad (30.233)$$

et de même:

$$y_A z_A + y_{A'} z_{A'} = 0 \quad (30.234)$$

□C.Q.F.D.

Remarque: Lorsque nous avons déterminé deux axes principaux d'inertie grâce aux symétries précédentes, le troisième est tout simplement celui qu'il faut pour compléter un trièdre orthogonal.

Troisième théorème: Si un système admet un axe de révolution pour sa distribution de masse, alors tout trièdre orthogonal incluant l'axe de révolution, est trièdre principal d'inertie. Le système matériel est alors dit "**système cylindrique**" et dans le trièdre principal d'inertie son tenseur prend la forme (en supposant que l'axe de révolution est le 3ème axe du trièdre):

$$(J_0)_{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3} = \begin{pmatrix} J & 0 & 0 \\ 0 & J & 0 \\ 0 & 0 & J' \end{pmatrix} \quad (30.235)$$

Démonstration:

Si Oz est un axe de révolution, tout plan contenant Oz est plan de symétrie et toute droite perpendiculaire à Oz est donc axe principal d'inertie (premier théorème). De plus, toutes ces droites perpendiculaires à Oz sont équivalentes.

□C.Q.F.D.

Définition: Si la matrice d'inertie en O d'un système matériel est du type:

$$(J_0)_{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3} = \begin{pmatrix} J & 0 & 0 \\ 0 & J & 0 \\ 0 & 0 & J \end{pmatrix} \quad (30.236)$$

nous disons alors que le système est un "**système sphérique**" (ou un "**système à symétrie sphérique**").

Remarque: Le choix systématique d'un trièdre principal d'inertie permet de ramener le tenseur d'inertie de 6 à 3 composantes, calculées une fois pour toute. Cependant, ce choix implique l'utilisation d'une base qui sera le plus souvent en mouvement par rapport au référentiel utilisé, ce qui pourra poser des problèmes de dérivations par rapport au temps des vecteurs de la base. Nous pouvons alors, si c'est plus facile, obtenir les composantes du tenseur de symétrie dans une base quelconque à l'aide d'une matrice de passage entre la base principale et la base utilisée pour le calcul du trièdre principal d'inertie.

Lorsque les moments d'inertie d'un solide sont connus dans les directions des axes principaux d'inertie, nous pouvons facilement déterminer le moment d'inertie J par rapport à n'importe quel autre axe passant par le centre de gravité en utilisant que ce nous nommons un "**ellipsoïde d'inertie**" (à ne pas confondre avec le moment d'inertie d'une ellipsoïde - démontré dans le chapitre traitant des Formes Géométriques).

Démonstration:

Soient trois axes, centrés sur G , parallèles aux axes principaux. Dans leurs directions, portons des longueurs proportionnelles à:

$$A = \frac{1}{\sqrt{J_x}}, B = \frac{1}{\sqrt{J_y}}, C = \frac{1}{\sqrt{J_z}}$$

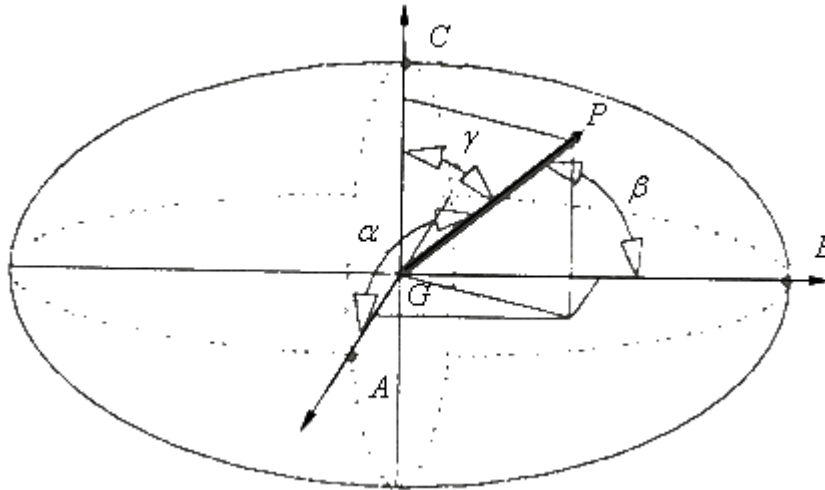


Figure: 30.15 - Illustration de l'ellipsoïde d'inertie

Dans cet espace des phases des moments d'inertie, tout point $P = (x, y, z)$ désigne un moment d'inertie J tel que:

$$\|\overrightarrow{GP}\|^2 = x^2 + y^2 + z^2 = \frac{1}{J} \quad (30.237)$$

Pour déterminer J en fonction des J_x, J_y, J_z , sans devoir calculer x, y, z , nous identifions les cosinus directeurs (cf. chapitre de Calcul Vectoriel) de l'axe de rotation à ceux de la droite $\overrightarrow{GP}$.

Ainsi, nous avons:

$$(\cos \alpha, \cos \beta, \cos \gamma) = \left(\frac{x}{\|\overrightarrow{GP}\|}, \frac{y}{\|\overrightarrow{GP}\|}, \frac{z}{\|\overrightarrow{GP}\|} \right) = \left(\frac{x}{\frac{1}{\sqrt{J}}}, \frac{y}{\frac{1}{\sqrt{J}}}, \frac{z}{\frac{1}{\sqrt{J}}} \right) = \sqrt{J}(x, y, z) \quad (30.238)$$

$$\Leftrightarrow x = \frac{\cos \alpha}{\sqrt{J}}, y = \frac{\cos \beta}{\sqrt{J}}, z = \frac{\cos \gamma}{\sqrt{J}}$$

Soit:

$$J \|\overrightarrow{GP}\|^2 = Jx^2 + Jy^2 + Jz^2 = 1 \quad (30.239)$$

Nous pouvons maintenant calculer les conditions de normalisation de cette relation. Ainsi, si $y, z = 0$ et $x = A$, nous avons:

$$Jx^2 + Jy^2 + Jz^2 = JA^2 = 1 \Rightarrow A^2 = \frac{1}{J} = \frac{1}{J_x} \quad (30.240)$$

Respectivement nous aurons:

$$B^2 = \frac{1}{J} = \frac{1}{J_y}, C^2 = \frac{1}{J} = \frac{1}{J_z} \quad (30.241)$$

Puisque:

$$A = \frac{1}{\sqrt{J_x}} \Rightarrow A^2 = \frac{1}{J_x} \text{ et } B = \frac{1}{\sqrt{J_y}} \Rightarrow B^2 = \frac{1}{J_y}, C = \frac{1}{\sqrt{J_z}} \Rightarrow C^2 = \frac{1}{J_z} \quad (30.242)$$

Ce qui nous amène à écrire:

$$J_x x^2 + J_y y^2 + J_z z^2 = \frac{x^2}{A^2} + \frac{y^2}{B^2} + \frac{z^2}{C^2} = 1 \quad (30.243)$$

Par substitution, nous obtenons:

$$\frac{x^2}{A^2} + \frac{y^2}{B^2} + \frac{z^2}{C^2} = \frac{1}{J} \left(\frac{\cos^2 \alpha}{A^2} + \frac{\cos^2 \beta}{B^2} + \frac{\cos^2 \gamma}{C^2} \right) = \frac{1}{J} (J_x \cos^2 \alpha + J_y \cos^2 \beta + J_z \cos^2 \gamma) \quad (30.244)$$

Donc finalement:

$$J_x \cos^2 \alpha + J_y \cos^2 \beta + J_z \cos^2 \gamma = J \quad (30.245)$$

Ainsi, en connaissant les moments d'inertie d'un corps par rapport à ses axes principaux J_x, J_y, J_z nous pouvons connaître son moment d'inertie par rapport à n'importe quel axe ayant un angle α, β, γ par rapport aux axes principaux.

□C.Q.F.D.

9.1.2. GYROSCOPE

Un solide (de révolution pour simplifier...), pouvant s'orienter librement autour d'un point fixe et tournant rapidement sur lui-même forme par définition un "gyroscope".

Outre leur usage ludique... car ils permettent d'avoir des configurations considérées comme pédagogiquement exceptionnelles... les gyroscopes constituent une part importante des systèmes de navigation par inertie (avant l'apparition des GPS...) dans l'aviation, l'aérospatiale, la marine (stabilisation des bateaux), le cinéma/télévision (stabilisation des caméras) et encore bien d'autres. Les instruments de guidage par inertie de ces systèmes sont constitués de gyroscopes et d'accéléromètres, qui calculent à tout instant la vitesse exacte et la direction de l'appareil en mouvement. Les signaux recueillis sont communiqués à un ordinateur qui les enregistre et qui corrige alors les aberrations éventuelles de la trajectoire.

Les planètes constituent un autre exemple fameux de gyroscopes. L'exemple le plus connu étant notre Terre qui tournant relativement vite autour d'elle-même et étant très massive son moment cinétique fait que son pôle Nord est toujours (à l'échelle du temps d'un humain ...) orientée vers l'étoile Polaire quelle que soit sa position sur son orbite.

La figure suivante est un exemple de gyroscope connu dans les laboratoires des écoles et appelé "gyroscope symétrique pesant". Il s'agit bien évidemment d'un cas particulier et simplifié mais qui permet de comprendre le principe de base du gyroscope:

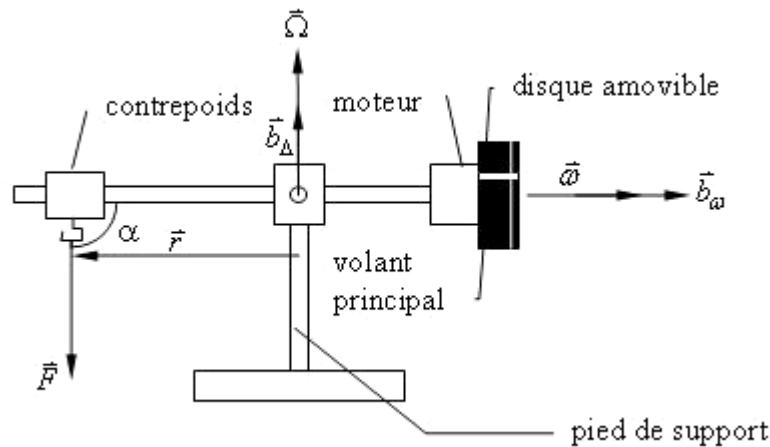


Figure: 30.16 - Exemple de gyroscope symétrique pesant

Se composant d'un moteur électrique dont le rotor, le volant principal, forme la masse principale en rotation angulaire rapide. Le stator du moteur est fixé à une tige sur laquelle est positionné un contrepoids à l'opposé. L'ensemble est posé sur un pied de support à l'extrémité duquel se trouve un cardan monté sur un roulement horizontal qui autorise les orientations du gyroscope presque sans limitations dans toutes les directions.

Dans ce schéma nous avons ω qui est la vitesse angulaire instantanée du disque amovible de rayon R , Ω est la vitesse de précession du gyroscope (rotation autour du pied de support), F est la force de la masse m complémentaire attachée au contrepoids et qui déséquilibre le gyroscope, r est la distance du cardan du gyroscope au contrepoids et finalement α est l'angle d'inclinaison que prend l'axe du gyroscope lorsqu'on le déséquilibre en attachant le poids supplémentaire au contrepoids.

Pour débiter l'étude théorique de ce système, rappelons que nous avons démontré plus haut que le moment cinétique pour un solide ayant un moment d'inertie J s'exprime par la relation suivante:

$$\vec{b}_\omega = J \cdot \vec{\omega} \quad (30.246)$$

et nous avons vu que tout solide en rotation autour d'un axe quelconque a aussi un moment cinétique qu'il est alors d'usage de noter conformément avec ce que nous avons vu plus haut:

$$\vec{b}_\Delta = J_\Delta \vec{\Omega} \quad (30.247)$$

Nous avons aussi démontré plus haut que le rotor, comme toute masse en rotation rapide, produit alors un moment de force donné par:

$$\vec{M} = \frac{d\vec{b}_\omega}{dt} = J \frac{d\vec{\omega}}{dt} \quad (30.248)$$

qui est vectoriellement colinéaire à $\vec{\omega}$ et passe donc par son axe de symétrie. Comme nous le savons déjà, c'est cette dernière relation qui met le mieux en évidence que le gyroscope maintient toujours une direction identique dans l'espace même lorsque nous déplaçons son support.

En d'autres termes un gyroscope libre animé d'une grande vitesse de rotation a pour propriété fondamentale de conserver son axe de rotation selon une orientation fixe par rapport à l'espace absolu. C'est ce que nous appelons la "**première loi gyroscopique**" ou "**loi de fixité**".

Typiquement le "**gyroscope de Foucault**" représenté ci-dessous, excellent exemple pratique de la loi de fixité, garde son orientation quelle que soit la manière dont nous manipulons le socle sur lequel il est posé:



Figure: 30.17 - Gyroscope de Foucault

Si nous posons le gyroscope de Foucault toute une journée sur une table avec un moteur qui maintient la rotation du disque massif central constante, nous observons alors la rotation de la Terre car le gyroscope tourne alors très lentement sur lui-même en 24 heures!

Pour revenir à nos considérations mathématiques... intéressons-nous maintenant au moment de force du contrepoids qui déséquilibre notre gyroscope symétrique alors que le disque est en rotation et qui génère une rotation générale du gyroscope comme le permet de constater l'expérience. Nous avons alors pour le moment de force faisant tourner le gyroscope autour de son axe (tige de soutien):

$$M = \|\vec{M}\| = \|\vec{r} \times \vec{F}\| = \|\vec{r}\| \cdot \|\vec{F}\| \cdot \sin \alpha = r \cdot m \cdot a \cdot \sin \alpha \quad (30.249)$$

Puisque le gyroscope ne précesse pas lorsque le système est équilibré c'est que le moment de force du poids supplémentaire qui déséquilibre le gyroscope génère un moment cinétique selon la relation démontrée plus haut tel que:

$$\frac{db_{\omega}}{dt} = M = r \cdot m \cdot a \cdot \sin \alpha \quad (30.250)$$

Ce qui schématiquement peut être représenté de la manière suivante (il s'agit de notre gyroscope vu d'en haut):

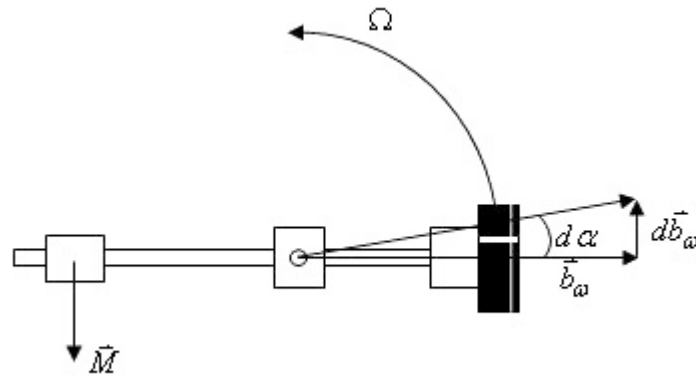


Figure: 30.18 - Gyroscopie symétrique pesant vu de haut

Nous avons alors dès que le gyroscope se met à tourner (dans un mouvement circulaire):

$$\tan(d\alpha) = \frac{\|d\vec{b}_\omega\|}{\|\vec{b}_\omega\|} \cong \frac{db_\omega}{b_\omega} \quad (30.251)$$

En prenant l'approximation de Taylor (cf. chapitre de Suites Et Séries) au premier ordre de la tangente pour les petits angles:

$$d\alpha \cong \frac{db_\omega}{b_\omega} \quad (30.252)$$

Faisons l'hypothèse, pour simplifier l'étude du problème, que la variation du moment cinétique total par rapport à l'axe de rotation du gyroscope (la tige de soutien donc!) peut être assimilée au moment cinétique du rotor seul si ce dernier tourne suffisamment vite et que sa masse est suffisamment grande. C'est-à-dire que:

$$\frac{d\vec{b}_\omega}{dt} \gg \frac{d\vec{b}_\Delta}{dt} \quad (30.253)$$

nous avons alors:

$$\frac{d\vec{b}}{dt} = \frac{d(\vec{b}_\Delta + \vec{b}_\omega)}{dt} \cong \frac{d\vec{b}_\omega}{dt} \quad (30.254)$$

et dès lors puisque de par cette approximation tout le moment de force est assigné à la variation du moment cinétique du rotor seul:

$$d\alpha \cong \frac{Mdt}{b_\omega} \quad (30.255)$$

Il vient enfin:

$$\Omega \cong \frac{d\alpha}{dt} = \frac{M}{b_\omega} = \frac{r \cdot F \cdot \sin \alpha}{J \cdot \omega} \quad (30.256)$$

Donc lorsque le gyroscope symétrique pesant est équilibré (lorsque M est nul au numérateur de la fraction), son moment cinétique garde donc une orientation fixe quelle que soit la valeur du dénominateur puisque Ω sera alors toujours nul.

Le mouvement de rotation résultant d'un déséquilibre du gyroscope est donc dit "**mouvement de précession**" lorsqu'il est provoqué volontairement, et "**dérive**" lorsqu'il est dû à un élément perturbateur.

Indiquons pour finir les gyroscopes ludiques pour petits enfants comme la toupie ci-dessous:



Figure: 30.19 - Gyroscope (toupie) ludique...

que pouvons grossièrement représenter ainsi (vue de côté et vue du dessus) pour en faire une analyse mathématique (faites l'essai avec vos enfants pour voir si cela les intéresse autant que le jouet...):

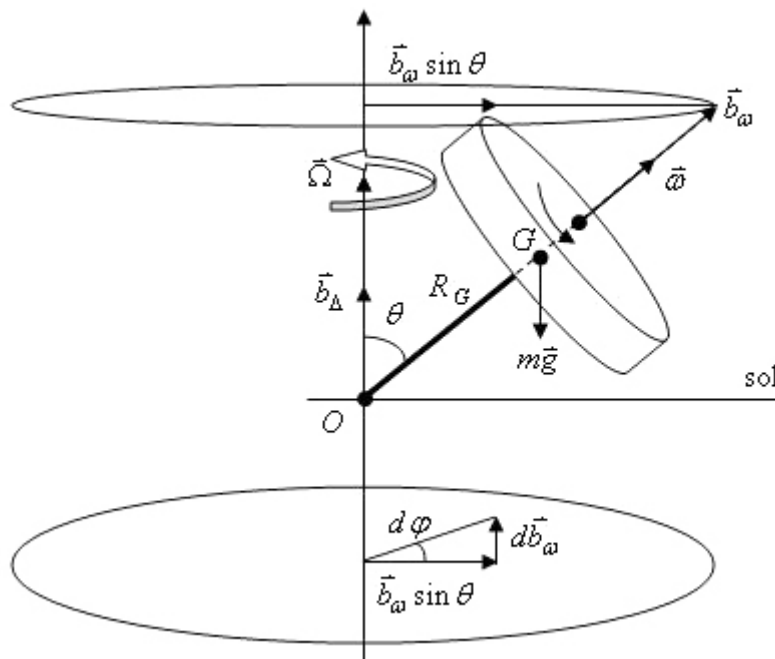


Figure: 30.20 - Illustration technique de la toupie

où nous faisons l'hypothèse que l'extrémité de l'axe de la toupie est posée sur le sol sans possibilité de glissement et que celle-ci a une vitesse angulaire ω constante et suffisamment grande pour ne pas avoir son inclinaison θ qui varie dans le temps.

En utilisant la même technique que pour le gyroscope symétrique pesant nous avons (bon nous aurions pu utiliser plus simplement la relation $L = \alpha R$ vue dans le chapitre de Trigonométrie...):

$$d b_{\omega} = b_{\omega} \sin(\theta) d \varphi \quad (30.257)$$

Nous avons aussi pour le moment de force:

$$M = R_G \cdot m \cdot g \cdot \sin \theta \quad (30.258)$$

Par contre le moment cinétique change! Effectivement, nous avons donc dans ce cas particulier:

$$db_\omega = b_\omega \sin(\theta) d\varphi \quad (30.259)$$

il s'ensuit que sous les mêmes hypothèses que le gyroscope pesant que:

$$\frac{db}{dt} \cong \frac{db_\omega}{dt} = b_\omega \sin(\theta) \frac{d\varphi}{dt} = b_\omega \sin(\theta) \Omega \quad (30.260)$$

d'où sous forme vectorielle:

$$\frac{d\vec{b}_\omega}{dt} = \vec{\Omega} \times \vec{b}_\omega \quad (30.261)$$

et nous savons que cette dernière relation (démonstration faite plus haut) peut être complétée en écrivant:

$$\frac{d\vec{b}_\omega}{dt} = \vec{\Omega} \times \vec{b}_\omega = \vec{M} \quad (30.262)$$

Il vient alors:

$$\Omega \cdot b_\omega \cdot \sin \theta = R_G \cdot m \cdot g \cdot \sin \theta \quad (30.263)$$

Soit:

$$\Omega = \frac{R_G \cdot m \cdot g}{b_\omega} = \frac{R_G \cdot m \cdot g}{J \cdot \omega} \quad (30.264)$$

Nous voyons que la différence avec le gyroscope symétrique pesant est que la vitesse de précession est alors indépendante de l'angle.

Remarques:

R1. Un cycliste roulant en ligne droite est stabilisé (loi de fixité oblige!) par le moment cinétique de ses roues qui est perpendiculaire au sens de roulement.

R2. Sans probablement s'en rendre compte, on se penche en bicyclette dans un virage pour produire une précession dans les roues et tourner plus facilement. Effectivement le mouvement de précession fait pivoter la roue de la bicyclette dans la direction où on se penche sans qu'on ait besoin de tourner le guidon.

9.2. ÉNERGIE POTENTIELLE GRAVIFIQUE

Si le travail de la force $\vec{F}$ entre les points A et B ne dépend pas du chemin suivi, nous disons que cette force dérive d'une énergie potentielle ou bien que le champ de force est un "champ conservatif" (contre-exemple: dans un mouvement avec frottement le travail dépend nécessairement de la voie choisie). Cette indépendance par rapport au chemin suivi implique que:

Soient deux points A et B de l'espace. Il y a plusieurs chemins possibles pour joindre ces deux points. Si nous en choisissons deux au hasard nous avons:

$$\begin{aligned} \text{Sur le 1}^{\text{er}} \text{ chemin: } & \int_A^B \vec{F} \circ d\vec{r}_1 \\ \text{Sur le 2}^{\text{ème}} \text{ chemin: } & \int_A^B \vec{F} \circ d\vec{r}_2 \end{aligned} \quad \begin{array}{c} A \\ \downarrow \\ 1 \quad \downarrow \quad 2 \\ B \end{array}$$

(30.265)

Si le champ est conservatif nous avons:

$$\int_A^B \vec{F} \circ d\vec{r}_1 = \int_A^B \vec{F} \circ d\vec{r}_2 = \int_A^B \vec{F} \circ d\vec{r} = - \int_B^A \vec{F} \circ d\vec{r} \quad (30.266)$$

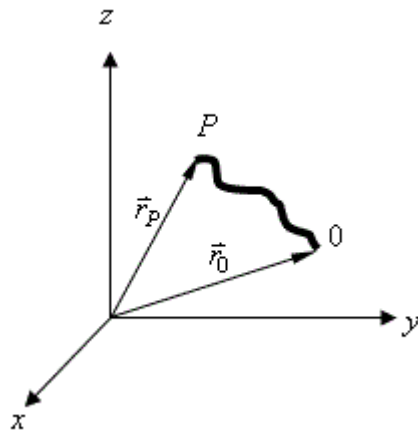
ou encore que le travail total sur un chemin fermé (aller et retour) est nul. Nous notons cela (cf. chapitre de Calcul Différentiel Et Intégral):

$$\oint \vec{F} \circ d\vec{r} = 0 \quad (30.267)$$

Le travail en jeu est donc une fonction du lieu seul (E_P) c'est-à-dire dépendant uniquement du point de départ et du point d'arrivée. En effet, si le travail dépendait du chemin, il serait possible de choisir la voie la plus généreuse quand le système fournit du travail et la voie la plus économique quand nous le ramenons à l'état initial. Ce serait donc un mouvement perpétuel et le principe de conservation de l'énergie l'interdit (cf. chapitre de Thermodynamique).

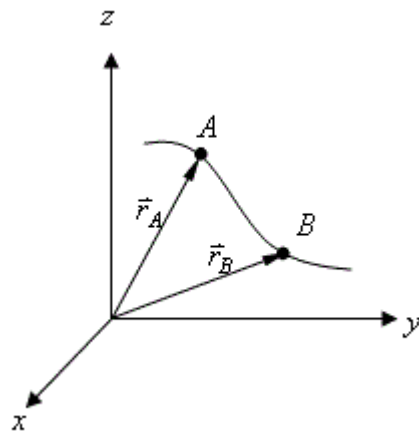
Attachons alors à chaque point P du champ de force une valeur de la fonction E_P (un nombre réel) correspondant au travail effectué par le champ de force lorsque le mobile passe d'un point P à 0, 0 étant un point de référence choisi arbitrairement. Donc par définition:

$$E_P(P) = \int_P^0 \vec{F} \circ d\vec{r} \quad \text{avec} \quad E_P(0) = 0 \quad (30.268)$$



En généralisant cette définition, nous dirons que le travail effectué par une force conservative lorsque le mobile passe de A à B est égal à la diminution d'énergie potentielle entre A et B:

$$W_{A,B} = \int_A^B \vec{F} \circ d\vec{r} = -(E_p(B) - E_p(A)) \quad (30.269)$$



Par définition E_p est l'énergie potentielle et se mesure en Joules.

L'équation précédente s'utilise très souvent sous forme différentielle soit:

$$dW = -dE_p \quad (30.270)$$

Il existe aussi rappelons-le une relation entre l'énergie et le gradient de la force donnée qui découle simplement de la définition du travail:

$$\vec{F} = -\vec{\nabla} \cdot E_p \quad (30.271)$$

Application: Travail de la pesanteur et énergie potentielle gravifique au voisinage de la surface de la Terre. C'est donc un cas particulier où la force est constante...

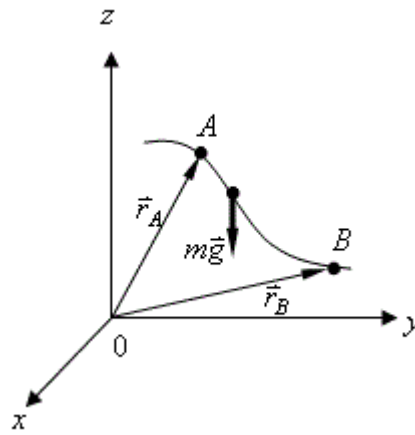


Figure: 30.21 - Exemple avec la force gravifique

Soit un point de masse m se déplaçant selon une trajectoire quelconque AB . Le poids $m\vec{g}$ effectue le travail:

$$W_{A,B} = \int_A^B \vec{F} \circ d\vec{r} = \vec{F} \circ (\vec{r}_B - \vec{r}_A) = m\vec{g} \circ (\vec{r}_B - \vec{r}_A) \quad (30.272)$$

En exprimant les différents vecteurs en composantes:

$$\vec{g} = (0, 0, -g), \quad \vec{r}_A = (x_A, y_A, z_A), \quad \vec{r}_B = (x_B, y_B, z_B) \quad (30.273)$$

et en calculant le produit scalaire au moyen de ces composantes nous obtenons:

$$W_{A,B} = m\vec{g} \circ (\vec{r}_B - \vec{r}_A) = -mg(z_B - z_A) = mg(z_A - z_B) \quad (30.274)$$

La différence $z_A - z_B$ représente la différence d'altitude entre les points A et B . Nous constatons bien que le travail ne dépend pas du chemin suivi mais seulement des points de départ et d'arrivée. Si, en sens inverse, nous voulons faire passer le point de masse de B à A , le travail, fourni alors par un agent extérieur vaut:

$$W_{B,A} = -W_{A,B} = -mg(z_A - z_B) \quad (30.275)$$

ce qui montre bien que le travail total sur un chemin fermé est nul:

$$\oint \vec{F} \circ d\vec{r} = 0 \quad (30.276)$$

En comparant les relations:

$$W_{A,B} = -(E_p(B) - E_p(A)) \text{ et } W_{A,B} = mg(z_A - z_B) \quad (30.277)$$

et en identifiant, nous obtenons ainsi:

$$E_p(z) = m \cdot g \cdot z \quad (30.278)$$

qui est l'énergie potentielle gravifique, z étant l'altitude de la masse m . Nous notons plus simplement la plupart du temps cette relation sous la forme:

$$E_p = m \cdot g \cdot h \quad (30.279)$$

Remarque: Le choix de zéro de l'énergie potentielle est souvent arbitraire; nous le fixons par commodité. Seules les différences d'énergie potentielle sont généralement intéressantes comme nous allons le voir de suite.

La relation précédente est au fait une expression utile à proximité de la surface terrestre. A distance $r \gg R$ où R est le rayon de la terre, la force de gravitation faiblit et l'approximation n'est plus valable (si $r \ll R$ aussi, d'ailleurs...).

Pour déterminer la relation correcte, considérons deux masses m_1, m_2 . La première est supposée au repos et fixe la deuxième est amenée de l'infini à une distance donnée de m_1 (le même raisonnement est applicable pour le champ électrique). Le travail dW de la force gravitationnelle en un point quelconque étant donc:

$$dW = Fdx \quad (30.280)$$

et l'énergie potentielle du système:

$$dE_p = -dW \quad (30.281)$$

Alors:

$$dE_p = -dW = -Fdx = -G \frac{m_1 m_2}{x^2} dx \quad (30.282)$$

d'où simplement après intégration (l'énergie potentielle en un point):

$$E_p = G \frac{m_1 m_2}{x} \quad (30.283)$$

Voyons si cela est cohérent avec $E_p = mgz \dots$

A hauteur nulle de la surface terrestre, $x = R$, nous avons:

$$E_p = -G \frac{m_1 m_2}{R} = E_{0,p} \quad (30.284)$$

où le choix du signe "-" dépend uniquement du référentiel choisi qui est dans le cas présent conforme à ce qui est d'usage de prendre dans les écoles.

Nous élevons l'objet de $h \ll R$:

$$E_p(R+h) = -G \frac{m_1 m_2}{R+h} = -G m_1 m_2 \frac{1}{R+h} = -G m_1 m_2 \frac{1}{R(1+h/R)} = -G \frac{m_1 m_2}{R} \frac{1}{(1+h/R)} \quad (30.285)$$

Nous utilisons l'approximation grossière:

$$\frac{1}{1+x} \cong 1-x \quad (30.286)$$

valable quand $x \ll 1$ d'où:

$$E_p(R+h) = -G \frac{m_1 m_2}{R} (1 - h/R) = -G \frac{m_1 m_2}{R} + G \frac{m_1 m_2}{R^2} h = U_0 - F_G h = U_0 + mgh \quad (30.287)$$

Comme à la surface de la terre nous avons l'habitude de poser en laboratoire $U_0 = 0$, nous obtenons bien finalement:

$$E_p(h) = mgh \quad (30.288)$$

et nous voyons qu'il s'agit effectivement d'une grossière approximation.

Remarque: Nous pourrions appliquer le même développement dans l'étude de la force de Coulomb et du champ électrique mais jusqu'à maintenant nous n'avons jamais mis de laboratoire à la surface d'une charge... (sic!).

9.2.1. ÉNERGIE POTENTIELLE D'UNE SPHÈRE DE MATIÈRE

Nous allons calculer ici l'énergie potentielle d'une sphère de matière. Cet exercice de style va nous être très utile en astrophysique pour déterminer la température interne des étoiles et dans le chapitre de Cosmologie pour le départ du modèle de Friedmann.

L'expression d'une énergie potentielle d'un système de deux masses mises en présence est donnée par:

$$E_p = -\frac{GMm}{r} \quad (30.289)$$

Soit une sphère de masse M , de densité massique ρ et de rayon r et entourée d'un anneau sphérique de rayon intérieur r , de même densité massique ρ et d'épaisseur dr .

L'énergie potentielle de l'anneau sphérique de rayon interne r et d'épaisseur dr se calcule comme suit:

La masse de la sphère de rayon r et de densité massique ρ est:

$$M = \frac{4}{3} \pi r^3 \rho \quad (30.290)$$

La masse de l'anneau entourant la sphère de rayon r , d'épaisseur dr et de densité massique ρ est:

$$m = 4 \pi r^2 dr \rho \quad (30.291)$$

En introduisant les deux dernières expressions dans celle de l'énergie potentielle:

$$dE_p = -\frac{G \left(\frac{4}{3} \pi r^3 \rho \right) (4 \pi r^2 \rho dr)}{r} = -\frac{1}{3} G (4 \pi \rho)^2 r^4 dr \quad (30.292)$$

En intégrant l'expression précédente entre 0 et R , cela revient à ajouter successivement une suite d'anneaux d'épaisseur dr pour obtenir la sphère entière de rayon R et donc l'énergie potentielle de la sphère entière.

$$E_p = -\frac{1}{3}G(4\pi\rho)^2 \int_0^R r^4 dr = -\frac{1}{3}G(4\pi\rho)^2 \cdot \frac{R^5}{5} \quad (30.293)$$

Ce qui s'écrit encore:

$$E_p = -\frac{3}{5}G \underbrace{\left(\frac{4}{3}\pi\rho R^3\right)^2}_{M^2} \cdot \frac{1}{R} \quad (30.294)$$

Soit finalement:

$$E_p = -\frac{3}{5} \frac{GM^2}{R} \quad (30.295)$$

9.3. CONSERVATION DE L'ÉNERGIE MÉCANIQUE TOTALE

Comparons maintenant les équations:

$$W_{A,B} = -(E_p(B) - E_p(A)) = E_c(B) - E_c(A) \quad (30.296)$$

puisque'il s'agit du même travail.

Ce qui entraîne:

$$E_p(A) + E_c(A) = E_c(B) + E_p(A) \quad (30.297)$$

somme des deux formes d'énergie en chaque point ou encore, les lieux A et B étant quelconques, en écrivant l'équation sous une forme générale:

$$E_p + E_c = c^{\text{te}} \quad (30.298)$$

Remarque: Nous nommons souvent l'énergie totale d'un système "**l'hamiltonien du système**" comme nous l'avons déjà mentionné dans le chapitre de Mécanique Analytique.

En l'absence de frottement s'il s'agit d'énergie mécanique, nous écrivons aussi la variation tel que:

$$dE_p + dE_c = 0 \quad (30.299)$$

Une augmentation d'énergie cinétique entraîne donc une diminution d'énergie potentielle (et réciproquement) puisque la somme des deux reste constante.

Contre-exemple: S'il y a frottement, donc dégagement de chaleur, l'énergie mécanique totale n'est plus constante! (l'énergie mécanique seulement).

Par ailleurs reprenons la relation:

$$dW = -dE_p \Leftrightarrow dE_p = -dW \quad (30.300)$$

et donc:

$$dE_p = -dW = -\vec{F} \circ d\vec{r} = -(F_x dx + F_y dy + F_z dz) \quad (30.301)$$

D'autre part, E_p étant une fonction scalaire dépendant des coordonnées d'espace, formons sa différentielle totale:

$$dE_p = \frac{\partial E_p}{\partial x} dx + \frac{\partial E_p}{\partial y} dy + \frac{\partial E_p}{\partial z} dz \quad (30.302)$$

en comparant avec l'équation précédente et en identifiant terme à terme, nous avons:

$$\frac{\partial E_p}{\partial x} = -F_x, \frac{\partial E_p}{\partial y} = -F_y, \frac{\partial E_p}{\partial z} = -F_z \quad (30.303)$$

d'où l'expression affirmant que la force dérive d'une énergie potentielle si le travail en jeu est indépendant du chemin suivi. Si nous exprimons la force $\vec{F}$ en termes de vecteurs-unités, nous obtenons:

$$\vec{F} = F_x \vec{i} + F_y \vec{j} + F_z \vec{k} = - \left(\frac{\partial E_p}{\partial x} \vec{i} + \frac{\partial E_p}{\partial y} \vec{j} + \frac{\partial E_p}{\partial z} \vec{k} \right) = -\overrightarrow{\text{grad}}(E_p) \quad (30.304)$$

En définitive, l'affirmation que la force dérive d'une énergie potentielle E_p peut se résumer ainsi:

$$\boxed{\vec{F} = -\overrightarrow{\text{grad}}(E_p)} \quad (30.305)$$

Dans le cas de la gravitation:

$$\vec{F} = m\vec{g} = -\overrightarrow{\text{grad}}(E_p) \quad (30.306)$$

ce qui s'écrit aussi avec l'opérateur nabla (cf. chapitre de Calcul Vectoriel):

$$\vec{F} = -\overrightarrow{\text{grad}}(E_p) = -\vec{\nabla}(E_p) \quad (30.307)$$

Le champ de gravitation est donc caractérisé par l'ensemble des vecteurs $\vec{g}$.

9.4. CONSERVATION DE LA QUANTITÉ DE MOUVEMENT

Un mobile, lors d'une interaction avec un autre point matériel, peut transmettre tout ou partie de son mouvement (énergie cinétique ou/et potentielle). C'est le cas lors d'un choc, par exemple (ceci dit le calcul de la force d'un choc est extrêmement difficile à effectuer sans de nombreuses simplifications). La grandeur ainsi échangée est la quantité de mouvement $\vec{p}$. Elle vaut par définition (nous l'avons déjà vu lorsque nous avons parlé de la deuxième loi de Newton):

$$\vec{p} = m\vec{v} \quad (30.308)$$

Évidemment, nous avons:

$$m\vec{v}_2 - m\vec{v}_1 = \int_{t_1}^{t_2} \vec{F} dt \quad (30.309)$$

La quantité:

$$\int_{t_1}^{t_2} \vec{F} dt$$

est parfois appelée "**impulsion**", et l'équation précédente porte quelque fois le nom de "**théorème de la quantité de mouvement**".

Il s'énonce ainsi: L'impulsion fournie par une force entre les instants t_1 et t_2 est égale à la variation de la quantité de mouvement durant cet intervalle de temps.

Mais revenons-en à notre conservation de la quantité de mouvement (et donc de l'énergie et réciproquement...). L'intérêt de la grandeur de quantité de mouvement résulte du fait qu'elle est conservée dans les interactions (en première approximation..). En effet, soient deux mobiles en collision, en vertu de l'égalité de l'action et de la réaction (3ème loi de Newton) nous avons:

$$\vec{F}_{2,1} = -\vec{F}_{1,2} \quad (30.310)$$

et en utilisant le théorème de la quantité de mouvement nous pouvons écrire:

$$\frac{d\vec{p}_1}{dt} = \vec{F}_{2,1} \text{ et } \frac{d\vec{p}_2}{dt} = \vec{F}_{1,2} \quad (30.311)$$

En additionnant membre à membre ces deux équations, nous déduisons:

$$\frac{d}{dt}(\vec{p}_1 + \vec{p}_2) = \vec{F}_{2,1} + \vec{F}_{1,2} = 0 \text{ car } \vec{F}_{2,1} = -\vec{F}_{1,2} \quad (30.312)$$

et donc:

$$\boxed{\vec{p}_1 + \vec{p}_2 = c^{te}} \quad (30.313)$$

La quantité de mouvement totale est constante, elle se conserve donc.

9.5. LOI DE NEWTON GENERALISÉE

Revenons maintenant un petit peu à notre principe de moindre action dont nous avons parlé au tout début de cette section:

Prenons le cas d'un objet lancé en l'air et repérons deux points de sa trajectoire en deux instants quelconques. Une infinité de courbes passent entre ces deux points et pourtant la nature n'en choisit qu'une seule. Qu'est-ce qui distingue cette courbe - la trajectoire physique - de toutes les autres? A cette question nous pourrions, très justement, répondre que cette courbe se distingue des autres par le fait qu'elle est solution de l'équation différentielle de la trajectoire ... avec les conditions initiales

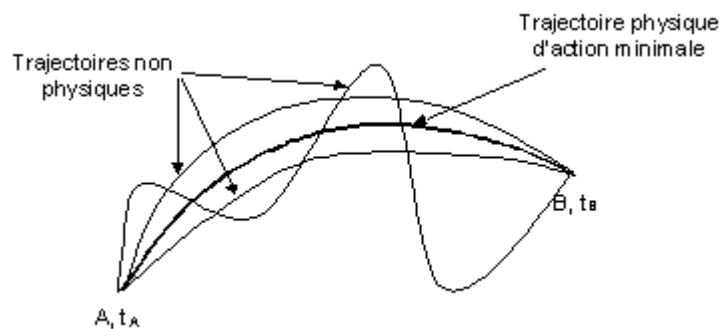
appropriées. Mais dans le cas où nous ignorons les conditions initiales ou lorsque le problème ne peut être ramené à une équation différentielle, par quel moyen pouvons-nous alors distinguer la trajectoire physique de tous les chemins possibles?

Le principe de moindre action s'exprime dans ce contexte par un minimum de vitesse pour un minimum de chemin parcouru.

En fait de vitesse, il convient mieux en mécanique de considérer la quantité de mouvement car cette dernière grandeur est directement liée aux propriétés inertielles des corps. Mathématiquement Maupertuis traduit le principe de moindre action comme suit.

Si nous considérons le mouvement d'un corps entre deux points A en t_A et B en t_B , pour une énergie totale E donnée, la trajectoire sélectionnée par la nature est celle pour laquelle la grandeur A_A^B suivante est minimale:

$$A_A^B = \int_{A, t_A}^{B, t_B} m \vec{v} \circ d\vec{r} \quad (30.314)$$



La trajectoire physique entre deux points A et B aux instants t_A et t_B est celle pour laquelle l'action est minimale.

En sachant que:

$$d\vec{r} = \vec{v} dt \quad (30.315)$$

nous obtenons alors:

$$A_A^B = \int_{t_A}^{t_B} m \vec{v} \circ \vec{v} dt = \int_{t_A}^{t_B} m v^2 dt = \int_{t_A}^{t_B} 2T(\vec{r}, t) dt \quad (30.316)$$

où T est l'énergie cinétique du corps.

Nous le voyons, l'action prend une forme étonnamment simple et s'exprime directement en fonction de l'énergie cinétique. Quelques années plus tard, à partir d'une intuition semblable à celle de Maupertuis, Euler parvint à un énoncé très similaire de l'action mais en partant du constat que les corps tendent à adopter un état où l'énergie potentielle est minimale. L'action d'Euler s'exprimait en fonction de l'énergie potentielle au lieu de l'énergie cinétique. Qui de Maupertuis ou d'Euler avait tort ou raison?

En fait, leurs énoncés respectifs de l'action étaient équivalents. Nous savons que dans un champ conservatif, si nous appelons U l'énergie potentielle alors l'énergie totale E vaut $T + U$ et cette énergie est une constante. Nous en tirons que $T = E - U$ et que donc:

$$2T = T + E - U \quad (30.317)$$

D'où:

$$A_A^B = \int_{t_A}^{t_B} (T(\vec{r}, t) + E - U(\vec{r}, t)) dt \quad (30.318)$$

Cette relation est vraie quel que soit le chemin d'énergie totale initiale E . Nous en concluons que la valeur de la constante E ne permet pas de discriminer les différentes trajectoires et peut donc être éliminée de la formulation de l'action. L'action de Maupertuis peut alors se réduire à une nouvelle grandeur notée S :

$$S_A^B = \int_{t_A}^{t_B} (T(\vec{r}, t) - U(\vec{r}, t)) dt = \int_{t_A}^{t_B} L(\vec{r}, t) dt \quad (30.319)$$

Cette nouvelle formulation de l'action fut donnée par Lagrange en 1788. S s'appelle "**l'action lagrangienne**" ou "**action hamiltonienne**" et la fonction:

$$L(\vec{r}, t) = T(\vec{r}, t) - U(\vec{r}, t) \quad (30.320)$$

porte le nom de "**lagrangien mécanique**". Ainsi formulé, le principe de moindre action devint l'un des outils les plus puissants de la mécanique.

Nous avons déjà vu comment nous exprimons le principe de moindre action mathématiquement. Dans le cas qui nous intéresse, l'action n'est pas une fonction de variables analytiques mais de trajectoires!

Considérons le cas très simple d'un corps de masse m se mouvant sur une seule dimension (que nous représenterons par un axe Ox) d'un point d'abscisse x_1 à l'instant t_1 à un point de coordonnée x_2 à l'instant t_2 . Supposons qu'il est soumis à un potentiel U qui ne varie pas avec le temps c'est-à-dire $U = U(x)$. L'action de ce corps sur un chemin C quelconque menant de (x_1, t_1) à (x_2, t_2) est alors:

$$S(C) = \int_{t_1}^{t_2} [T(x, t) - U(x)] dt = \int_{t_1}^{t_2} \left[\frac{1}{2} m \left(\frac{dx}{dt} \right)^2 - U(x) \right] dt \quad (30.321)$$

Soit C_0 le chemin physique et S_0 l'action sur ce chemin. Notons par $\bar{x}$ les valeurs de la position x sur le chemin physique. Considérons maintenant un chemin C très proche de C_0 tel que les positions le long de C aient les valeurs $x(t) = \bar{x}(t) + \varepsilon(t)$ que nous écrirons, pour alléger les écritures $x = \bar{x} + \varepsilon$.

Calculons l'action pour ce chemin:

$$S(C) = \int_{t_1}^{t_2} \left[\frac{1}{2} m \left(\frac{d(\bar{x} + \varepsilon)}{dt} \right)^2 - U(\bar{x} + \varepsilon) \right] dt \quad (30.322)$$

Comme ε est infiniment petit, il est possible de développer le potentiel en développement limité:

$$U(\bar{x} + \varepsilon) = U(\bar{x}) + \varepsilon \frac{dU}{dx} \quad (30.323)$$

Quant au premier terme, il se ramène à:

$$\left(\frac{d(\bar{x} + \varepsilon)}{dt} \right)^2 = \left(\frac{d\bar{x}}{dt} \right)^2 + 2 \frac{d\bar{x}}{dt} \frac{d\varepsilon}{dt} + \left(\frac{d\varepsilon}{dt} \right)^2 \quad (30.324)$$

Comme nous ne considérons que les variations du premier ordre, le dernier terme peut être négligé, ce qui donne pour l'action sur le chemin C :

$$S(C) = \int_{t_1}^{t_2} \left[\frac{1}{2} m \left(\left(\frac{d\bar{x}}{dt} \right)^2 + 2 \frac{d\bar{x}}{dt} \frac{d\varepsilon}{dt} \right) - U(\bar{x}) - \varepsilon \frac{dU}{dx} \right] dt \quad (30.325)$$

Posons maintenant que la variation δ de l'action entre le chemin physique C_0 et C est nulle:

$$\delta S = S(C) - S(C_0) = 0 \quad (30.326)$$

et ainsi:

$$\begin{aligned} \delta S &= \int_{t_1}^{t_2} \left[\frac{1}{2} m \left(\left(\frac{d\bar{x}}{dt} \right)^2 + 2 \frac{d\bar{x}}{dt} \frac{d\varepsilon}{dt} \right) - U(\bar{x}) - \varepsilon \frac{dU}{dx} \right] dt - \int_{t_1}^{t_2} \left[\frac{1}{2} m \left(\frac{d\bar{x}}{dt} \right)^2 - U(\bar{x}) \right] dt \\ &= \int_{t_1}^{t_2} \left[m \frac{d\bar{x}}{dt} \frac{d\varepsilon}{dt} - \varepsilon \frac{dU}{dx} \right] dt = 0 \end{aligned} \quad (30.327)$$

Le premier terme dans la dernière intégrale peut s'intégrer par parties comme suit:

$$\int_{t_1}^{t_2} m \frac{d\bar{x}}{dt} \frac{d\varepsilon}{dt} dt = m \frac{d\bar{x}(t)}{dt} \varepsilon(t) \Big|_{t_1}^{t_2} - \int_{t_1}^{t_2} m \varepsilon \frac{d^2 \bar{x}}{dt^2} dt \quad (30.328)$$

Or, tous les chemins partent de x_1 à l'instant t_1 et arrivent à x_2 à l'instant t_2 . Ceci implique qu'en t_1 et t_2 la variation ε est nulle ce que nous écrivons $\varepsilon(x_1, t_1) = \varepsilon(x_2, t_2) = 0$. Donc le premier terme de l'intégration par parties est nul. La variation de l'action prend alors la forme:

$$\delta S = \int_{t_1}^{t_2} \left[m \varepsilon \frac{d^2 \bar{x}}{dt^2} - \varepsilon \frac{dU}{dx} \right] dt = \int_{t_1}^{t_2} \left[m \frac{d^2 \bar{x}}{dt^2} - \frac{dU}{dx} \right] \varepsilon dt = 0 \quad (30.329)$$

Cette intégrale doit être nulle pour tous les chemins très proches du chemin physique C_0 , donc quelle que soit la valeur de ε . Pour qu'une telle condition soit remplie il faut que le terme devant ε soit nul, c'est-à-dire:

$$m \frac{d^2 \bar{x}}{dt^2} + \frac{dU}{dx} = 0 \quad (30.330)$$

Or nous connaissons au fait cette équation: le premier terme n'est rien d'autre que $m \cdot a$ où a est l'accélération du corps, et le second - l'opposé du gradient du potentiel - est l'intensité de la force en un point donné. Celle-ci se réduit donc à l'équation:

$$\boxed{F = m \frac{d^2 \bar{x}}{dt^2} + \frac{dU}{dx}} \quad (30.331)$$

qui n'est autre que la deuxième loi de Newton sous forme généralisée que nous avons obtenu plus haut sous la forme suivante:

$$\vec{F} = m \left(\frac{\partial}{\partial t} + \vec{v} \circ \vec{\nabla} \right) \vec{v} \quad (30.332)$$

Le principe de moindre action contient donc implicitement la mécanique newtonienne. Ainsi, il est possible de reconstruire toute la mécanique de Newton avec le seul principe de moindre action!!!

Cet échafaudage de calculs peut paraître bien compliqué pour aboutir à un résultat que nous connaissions déjà mais tout l'intérêt du principe de moindre action réside dans le fait qu'il permet de tirer des lois fondamentales à partir de la seule connaissance du lagrangien d'un système.

Les théories les plus récentes comme la théorie quantique des champs, les théories de jauge ou la théorie des supercordes ont toutes pour point de départ l'expression de l'action du système. Les physiciens en dégagent ensuite des lois fondamentales qui régissent le comportement des particules élémentaires.

10. PUISSANCE

Définition: La puissance est le taux instantané de variation du travail (énergie sous forme quelconque). Nous avons donc la "**puissance instantanée**" donnée par:

$$\boxed{P = \frac{dW}{dt}} \quad (30.333)$$

Si le travail est fourni de façon régulière (constante), nous avons alors la "**puissance moyenne**":

$$P = \frac{W}{t} \quad (30.334)$$

Avec cette définition, le lecteur pourrait penser qu'un véhicule qui roule à une vitesse constante fournit donc une puissance nulle puisque son énergie cinétique ne varie pas. En réalité il n'en est rien, car la voiture doit constamment vaincre le frottement des pneus avec la route (voir plus loin l'étude de la tribologie), le frottement visqueux avec l'air (cf. chapitre de Mécanique Des Milieux Continus), et la perte d'énergie due aux vibrations et frottements des ses propres composants comme les essieux, les roulements à bille, les ressorts, etc. Ainsi, un véhicule doit à chaque seconde fournir l'énergie qu'il a perdu dans ces différents frottements. Nous avons alors la "**puissance d'une force**":

$$P = \frac{dW}{dt} = \frac{d(F_T \cdot r)}{dt} = F_T \frac{d(r)}{dt} = F_T v \quad (30.335)$$

où nous avons utilisé la définition de l'énergie (force sur une distance) et où $\vec{F}_T$ représente la somme des diverses forces.

Remarques:

R1. L'unité de la puissance est le "**Watt**" et se note $[W]$ mais en technique, certains utilisent encore souvent le "**cheval**" $[ch]$ défini comme suit étant égal à 736 $[W]$ (car un cheval pouvait à l'époque soulever 75 kilos à 1 mètre en 1 seconde sous la gravité terrestre).

R2. En exprimant le travail (énergie) à partir de l'équation $\mathcal{W} = P \cdot t$, où la puissance est donnée en $[kW]$ et le temps en heures, il apparaît alors l'unité d'énergie $[kWh]$ (kilowattheure), très utilisée en pratique.

10.1. PUISSANCE D'UNE MACHINE TOURNANTE

Le travail élémentaire dW effectué par la force $\vec{F}$ faisant tourner une solide (un cylindre dans le cas présenté de suite) autour de son axe d'un angle θ vaut:

$$dW = Fdr = F \left(\frac{2\pi r}{2\pi} d\theta \right) = Frd\theta \quad (30.336)$$

La puissance instantanée est alors:

$$P = \frac{dW}{dt} = F \cdot r \frac{d\theta}{dt} = F \cdot r \cdot \omega \quad (30.337)$$

Or, comme nous l'avons démontré plus haut (il s'agit en fait plutôt d'une définition...):

$$F \cdot r = \|\vec{M}\| \quad (30.338)$$

La puissance d'un couple est alors donnée par:

$$P = M \cdot \omega \quad (30.339)$$

Il s'agit d'une relation très prisée par les passionnés de véhicules à moteur. Effectivement, connaissant le couple moteur (le moment de force) et le régime moteur (qu'il faut convertir dans les bonnes unités), nous obtenons facilement une approximation de la puissance développée par le moteur. Si nous divisons le résultat par 736, le lecteur obtiendra la mesure de la puissance en "chevaux".

10.1.1. RENDEMENT

A cause des frottements, la puissance restituée par une machine appelée aussi "**puissance utile**", est toujours inférieure à la puissance absorbée. Nous tenons compte de cet effet au moyen du rendement défini par:

$$\eta = \frac{P_{\text{utile}}}{P_{\text{absorbée}}} < 1 \quad (30.340)$$

Nous y reviendrons beaucoup plus en détail lors de notre étude de la thermodynamique (cf. chapitre de Thermodynamique).

11. MOUVEMENTS RELATIFS ET FORCES D'INERTIES

Voyons maintenant des développements qui vont nous permettre d'introduire un élément très important et utile en mécanique des fluides (cf. chapitre de Mécanique Statistique) et en météorologie (cf. chapitre de Génie Marin & Météo).

Considérons un référentiel fixe X, Y, Z et un référentiel mobile x, y, z . Ils sont donc en mouvement relatif et nous envisageons une rotation possible du référentiel mobile. Il s'agit d'exprimer, la vitesse, l'accélération d'un point P de l'espace au moyen des coordonnées du référentiel fixe (coordonnées absolues) à partir de celles attachées au référentiel mobile (coordonnées relatives) et du mouvement d'entraînement du référentiel mobile.

Nous définissons dans notre étude:

$\vec{r} = (x(t), y(t), z(t))$ vecteur de position de P par rapport au référentiel mobile

$\vec{r}_0 = (x_0(t), y_0(t), z_0(t))$ vecteur de position de P par rapport au référentiel fixe

$\vec{\rho} = (...)$ vecteur de position de O par rapport à l'origine du référentiel fixe

$\vec{v}_a = (...)$ vitesse absolue de P par rapport au référentiel fixe (supposé inconnue)

$\vec{a}_a = (...)$ accélération absolue de P par rapport au référentiel fixe (supposé inconnue)

$\vec{v}_r = (...)$ vitesse relative de P par rapport au référentiel mobile (supposé connue)

$\vec{a}_r = (...)$ accélération relative de P par rapport au référentiel mobile (supposé connue)

$\vec{v}_e = (...)$ vitesse d'entraînement au point O du mouvement relatif du référentiel mobile par rapport au référentiel fixe (supposé connue)

$\vec{a}_e = (...)$ accélération d'entraînement du référentiel mobile (supposé connue)

$\vec{\omega} = (...)$ vitesse angulaire du référentiel mobile

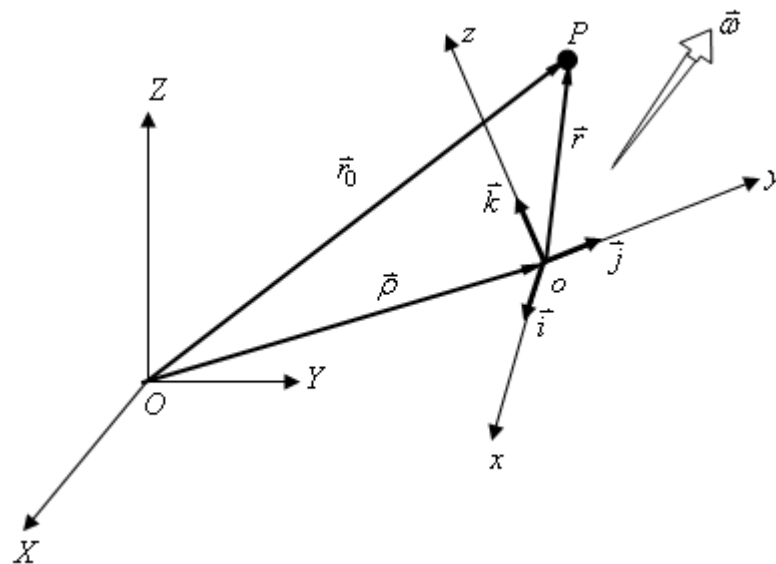


Figure: 30.22 - Exemple de référentiel en mouvement et en rotation par rapport à un référentiel fixe

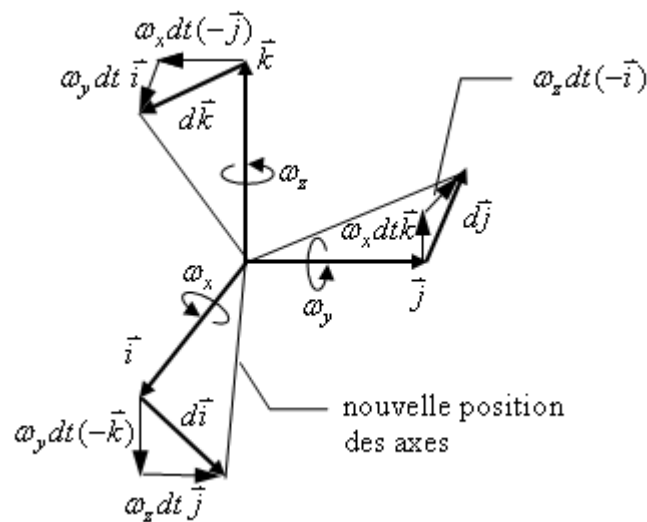
la position du point P est donc donnée par la "**relation de composition des positions**":

$$\vec{r}_0 = \vec{\rho} + \vec{r} \quad (30.341)$$

La vitesse absolue se calcule comme suit:

$$\begin{aligned} \vec{v}_a &= \frac{d\vec{r}_0}{dt} \\ &= \frac{d\vec{\rho}}{dt} + \frac{d\vec{r}}{dt} = \frac{d\vec{\rho}}{dt} + \frac{d}{dt}(x \cdot \vec{i} + y \cdot \vec{j} + z \cdot \vec{k}) \\ &= \vec{v}_e + \left(\frac{dx}{dt} \cdot \vec{i} + \frac{dy}{dt} \cdot \vec{j} + \frac{dz}{dt} \cdot \vec{k} \right) + \left(x \frac{d\vec{i}}{dt} + y \frac{d\vec{j}}{dt} + z \frac{d\vec{k}}{dt} \right) = \vec{v}_e + \vec{v}_r + \left(x \frac{d\vec{i}}{dt} + y \frac{d\vec{j}}{dt} + z \frac{d\vec{k}}{dt} \right) \end{aligned} \quad (30.342)$$

Le dernier terme est la contribution due à la rotation du référentiel mobile. Il s'agit maintenant d'exprimer la valeur de cette contribution en envisageant des rotations d'angle $d\vartheta_i = \omega_i dt$ autour de chacun des axes, successivement:



$$d\vec{i} = (\vec{j}\omega_z - \vec{k}\omega_y)dt, d\vec{j} = (\vec{k}\omega_x - \vec{i}\omega_z)dt, d\vec{k} = (\vec{i}\omega_y - \vec{j}\omega_x)dt \quad (30.343)$$

Nous obtenons ainsi les vecteurs élémentaires $d\vec{i}, d\vec{j}, d\vec{k}$ figurant les déplacements des extrémités des vecteurs-unités $\vec{i}, \vec{j}, \vec{k}$. Nous les introduisons dans l'expression ci-dessous qui devient, après réarrangement des termes:

$$x \frac{d\vec{i}}{dt} + y \frac{d\vec{j}}{dt} + z \frac{d\vec{k}}{dt} = (\omega_z y - \omega_y z) \cdot \vec{i} + (\omega_x z - \omega_z x) \cdot \vec{j} + (\omega_x y - \omega_y x) \cdot \vec{k} = \vec{\omega} \times \vec{r} \quad (30.344)$$

par définition du produit vectoriel. La vitesse absolue du point P s'exprime donc selon la "**loi de composition des vitesses**":

$$\vec{v}_a = \vec{v}_e + \vec{v}_r + \vec{\omega} \times \vec{r} \quad (30.345)$$

nous constatons que dans le cas particulier où le référentiel mobile ne subit qu'une translation ($\vec{\omega} = 0$), nous trouvons l'équation $\vec{v}_a = \vec{v}_e + \vec{v}_r$ caractéristique de la transformation de Galilée et nous disons alors que les référentiels sont en "**translations relatives**".

Remarque: Si nous nous concentrons uniquement sur les termes de vitesse d'entraînement et de rotation du référentiel mobile nous obtenons alors ce que nous appelons la "**formule de Bour**".

En procédant de la même façon que pour la recherche de la vitesse absolue il vient, en dérivant la relation précédente:

$$\vec{a}_a = \vec{a}_e + \frac{d\vec{v}_r}{dt} + \frac{d}{dt}(\vec{\omega} \times \vec{r}) \quad (30.346)$$

avec:

$$\begin{aligned}
\frac{d(\vec{\omega} \times \vec{r})}{dt} &= \frac{d[(\omega_y z - y \omega_z) \cdot \vec{i} + (\omega_z x - z \omega_x) \cdot \vec{j} + (\omega_x y - x \omega_y) \cdot \vec{k}]}{dt} \\
&= \frac{d\omega_y}{dt} z \cdot \vec{i} + \omega_y \frac{dz}{dt} \cdot \vec{i} + \omega_y z \frac{d\vec{i}}{dt} - \frac{d\omega_z}{dt} y \cdot \vec{i} - \omega_z \frac{dy}{dt} \cdot \vec{i} - \omega_z y \frac{d\vec{i}}{dt} \\
&+ \frac{d\omega_z}{dt} x \cdot \vec{j} + \omega_z \frac{dx}{dt} \cdot \vec{j} + \omega_z x \frac{d\vec{j}}{dt} - \frac{d\omega_x}{dt} z \cdot \vec{j} - \omega_x \frac{dz}{dt} \cdot \vec{j} - \omega_x z \frac{d\vec{j}}{dt} \\
&+ \frac{d\omega_x}{dt} y \cdot \vec{k} + \omega_x \frac{dy}{dt} \cdot \vec{k} + \omega_x y \frac{d\vec{k}}{dt} - \frac{d\omega_y}{dt} x \cdot \vec{k} - \omega_y \frac{dx}{dt} \cdot \vec{k} - \omega_y x \frac{d\vec{k}}{dt}
\end{aligned} \quad (30.347)$$

Si nous regroupons les termes:

$$\left(\omega_y \frac{dz}{dt} \cdot \vec{i} - \omega_z \frac{dy}{dt} \cdot \vec{i} \right) + \left(\omega_z \frac{dx}{dt} \cdot \vec{j} - \omega_x \frac{dz}{dt} \cdot \vec{j} \right) + \left(\omega_x \frac{dy}{dt} \cdot \vec{k} - \omega_y \frac{dx}{dt} \cdot \vec{k} \right) = \vec{\omega} \times \vec{v}_r \quad (30.348)$$

Si nous regroupons les termes:

$$\left(\frac{d\omega_y}{dt} z \cdot \vec{i} - \frac{d\omega_z}{dt} y \cdot \vec{i} \right) + \left(\frac{d\omega_z}{dt} x \cdot \vec{j} - \frac{d\omega_x}{dt} z \cdot \vec{j} \right) + \left(\frac{d\omega_x}{dt} y \cdot \vec{k} - \frac{d\omega_y}{dt} x \cdot \vec{k} \right) = \frac{d\vec{\omega}}{dt} \times \vec{r} \quad (30.349)$$

Si nous regroupons les termes:

$$\begin{aligned}
&\left(\omega_y z \frac{d\vec{i}}{dt} - \omega_z y \frac{d\vec{i}}{dt} \right) + \left(\omega_z x \frac{d\vec{j}}{dt} - \omega_x z \frac{d\vec{j}}{dt} \right) + \left(\omega_x y \frac{d\vec{k}}{dt} - \omega_y x \frac{d\vec{k}}{dt} \right) \\
&= (\omega_y z - \omega_z y) \frac{d\vec{i}}{dt} + (\omega_z x - \omega_x z) \frac{d\vec{j}}{dt} + (\omega_x y - \omega_y x) \frac{d\vec{k}}{dt}
\end{aligned} \quad (30.350)$$

et nous avons:

$$\frac{d\vec{i}}{dt} = (\vec{j} \omega_z - \vec{k} \omega_y), \frac{d\vec{j}}{dt} = (\vec{k} \omega_x - \vec{i} \omega_z), \frac{d\vec{k}}{dt} = (\vec{i} \omega_y - \vec{j} \omega_x) \quad (30.351)$$

Donc:

$$\begin{aligned}
&(\omega_y z - \omega_z y)(\vec{j} \omega_z - \vec{k} \omega_y) + (\omega_z x - \omega_x z)(\vec{k} \omega_x - \vec{i} \omega_z) + (\omega_x y - \omega_y x)(\vec{i} \omega_y - \vec{j} \omega_x) \\
&= (\omega_z x - \omega_x z)(-\vec{i} \omega_z) + (\omega_x y - \omega_y x)(\vec{i} \omega_y) \\
&+ (\omega_y z - \omega_z y)(\vec{j} \omega_z) + (\omega_x y - \omega_y x)(-\vec{j} \omega_x) \\
&+ (\omega_y z - \omega_z y)(-\vec{k} \omega_y) + (\omega_z x - \omega_x z)(\vec{k} \omega_x) \\
&= [\omega_y (\omega_x y - \omega_y x) - \omega_z (\omega_z x - \omega_x z)] \vec{i} + [\omega_z (\omega_y z - \omega_z y) - \omega_x (\omega_x y - \omega_y x)] \vec{j} \\
&+ [\omega_x (\omega_z x - \omega_x z) - \omega_y (\omega_y z - \omega_z y)] \vec{k} \\
&= \vec{\omega} \times (\vec{\omega} \times \vec{r})
\end{aligned} \quad (30.352)$$

Finalement:

$$\vec{a}_a = \vec{a}_e + \frac{d\vec{v}_r}{dt} + \vec{\omega} \times \vec{v}_r + \frac{d\vec{\omega}}{dt} \times \vec{r} + \vec{\omega} \times (\vec{\omega} \times \vec{r}) \quad (30.353)$$

mais (!) rappelons-nous que:

$$\vec{v}_r = \left(\frac{dx}{dt} \cdot \vec{i} + \frac{dy}{dt} \cdot \vec{j} + \frac{dz}{dt} \cdot \vec{k} \right) \quad (30.354)$$

ainsi:

$$\begin{aligned} \frac{d\vec{v}_r}{dt} &= \frac{d}{dt} \left(\frac{dx}{dt} \cdot \vec{i} + \frac{dy}{dt} \cdot \vec{j} + \frac{dz}{dt} \cdot \vec{k} \right) \\ &= \left(\frac{d^2x}{dt^2} \cdot \vec{i} + \frac{d^2y}{dt^2} \cdot \vec{j} + \frac{d^2z}{dt^2} \cdot \vec{k} \right) + \left(\frac{dx}{dt} \cdot \frac{d\vec{i}}{dt} + \frac{dy}{dt} \cdot \frac{d\vec{j}}{dt} + \frac{dz}{dt} \cdot \frac{d\vec{k}}{dt} \right) \quad (30.355) \\ &= \vec{a}_r + \vec{\omega} \times \vec{v}_r \end{aligned}$$

L'accélération absolue ou la "**loi de composition des accélérations**" s'exprime alors comme:

$$\vec{a}_a = \vec{a}_e + \vec{a}_r + 2\vec{\omega} \times \vec{v}_r + \frac{d\vec{\omega}}{dt} \times \vec{r} + \vec{\omega} \times (\vec{\omega} \times \vec{r}) \quad (30.356)$$

Le terme:

$$\vec{a}_c = 2\vec{\omega} \times \vec{v} \quad (30.357)$$

est appelé "**accélération de Coriolis**" (~1820) et le terme $\vec{\omega} \times (\vec{\omega} \times \vec{r})$ est simplement l'expression de l'accélération centripète dans ce cas particulier.

La loi de Newton $\vec{F} = m \cdot \vec{a}$ doit comporter tous les termes contenus dans l'équation générale ci-dessus. Pour un observateur situé dans le système fixe, cette loi s'écrit alors:

$$\vec{F} = m\vec{a}_e + m\vec{a}_r + 2m\vec{\omega} \times \vec{v}_r + m \frac{d\vec{\omega}}{dt} \times \vec{r} + m\vec{\omega} \times (\vec{\omega} \times \vec{r}) \quad (30.358)$$

où l'on a en premier terme à droite de l'égalité la force d'entraînement, en troisième la force de Coriolis et en dernier la force centripète.

Si le point P est lié rigidement au référentiel mobile, un observateur dans ce système ne perçoit aucun mouvement, par conséquent aucune accélération $\vec{a}_r$. Nous avons donc affaire à un système de forces en équilibre. Le problème de dynamique est alors ramené à un problème de statique. C'est le "**principe d'Alembert**".

Exemple:

Etant donné que pratiquement toutes nos observations sont faites sur Terre, c'est-à-dire dans un référentiel mobile dans l'Univers, la force de Coriolis peut y être mise en évidence.

L'étude du mouvement d'un corps par rapport à la Terre est l'une des applications les plus intéressantes de l'équation démontrée précédemment. La Terre a une vitesse angulaire (supposée constante!) dont la

direction est celle de l'axe de rotation de la Terre. Appelons $\vec{g}_0$ l'accélération de la pesanteur mesurée en un point A à la surface de la Terre si celle-ci ne tournait pas. $\vec{g}_0$ correspond alors à $\vec{a}_a$. En tirant l'accélération d'entraînement et relative nous obtenons l'accélération mesurée par un observateur en mouvement avec la Terre:

$$\vec{a}_e + \vec{a}_r = \vec{a}' = \vec{a}_a - 2\vec{\omega} \times \vec{v}_r - \frac{d\vec{\omega}}{dt} \times \vec{r} - \vec{\omega} \times (\vec{\omega} \times \vec{r}) \quad (30.359)$$

où $\frac{d\vec{\omega}}{dt} \times \vec{r} \cong 0$ est négligé dans le cas de la rotation de la Terre.

Nous considérons d'abord le cas d'un corps initialement au repos, ou se déplaçant très lentement de sorte que le terme de Coriolis est nul ou négligeable comparé au dernier terme. L'accélération que nous mesurons dans ce cas est appelée "**accélération effective**" de la pesanteur, et nous la désignons par $\vec{g}$.

Par suite:

$$\vec{g} = \vec{g}_0 - \vec{\omega} \times (\vec{\omega} \times \vec{r}) \quad (30.360)$$

En supposant que la Terre est une sphère (en fait sa forme s'en écarte légèrement) et qu'il n'y a pas d'anomalies locales, nous pouvons estimer que $\vec{g}_0$ est dirigé vers le centre de la Terre. Le deuxième terme $-\vec{\omega}(\vec{\omega} \times \vec{r})$ étant l'accélération centrifuge elle est dirigée vers l'extérieur.

Puisque $\vec{g}$ est la somme de $\vec{g}_0$ et de l'accélération centrifuge, la direction de $\vec{g}$, appelée la "**direction verticale**", s'écarte en réalité légèrement de la direction radiale; elle est expérimentalement déterminée par un fil à plomb. Les liquides se maintiennent toujours en équilibre avec leur surface perpendiculaire à $\vec{g}$.

L'ordre de grandeur de l'accélération centrifuge est:

$$\vec{\omega} \times (\vec{\omega} \times \vec{r}) \cong \omega^2 \cdot r = 3.39 \cdot 10^{-2} [m \cdot s^{-2}] \quad (30.361)$$

où r est le rayon de la Terre. L'accélération centrifuge décroît de l'équateur aux pôles car le rayon de la Terre n'est pas constant (la Terre est aplatie aux pôles). Cette variation de l'accélération est toujours très petite quand nous la comparons avec la pesanteur $\vec{g}_0 \cong 9.80 [m \cdot s^{-2}]$ mais elle explique cependant la plupart des variations observées de la valeur de la pesanteur avec la latitude.

Le gradient de l'accélération centrifuge a pour effet de déplacer légèrement la direction radiale d'un corps qui tombe en chute libre: le déplacement est vers le Sud dans l'hémisphère Nord et vers le Nord dans l'hémisphère Sud.

Considérons ensuite le terme de Coriolis. Dans le cas de la chute d'un corps, la vitesse $\vec{v}_r$ est dirigée vers le bas. D'autre part, comme $\vec{\omega}$ se trouve le long de l'axe de la Terre, $\vec{\omega} \times \vec{v}_r$ est dirigé vers l'Ouest. Le terme de Coriolis $-2m\vec{\omega} \times \vec{v}_r$ est donc dirigé vers l'Est; le corps qui tombe sera dévié dans cette direction.

Pour un corps tombant dans un plan parallèle et tangent à la surface de la Terre, nous avons:

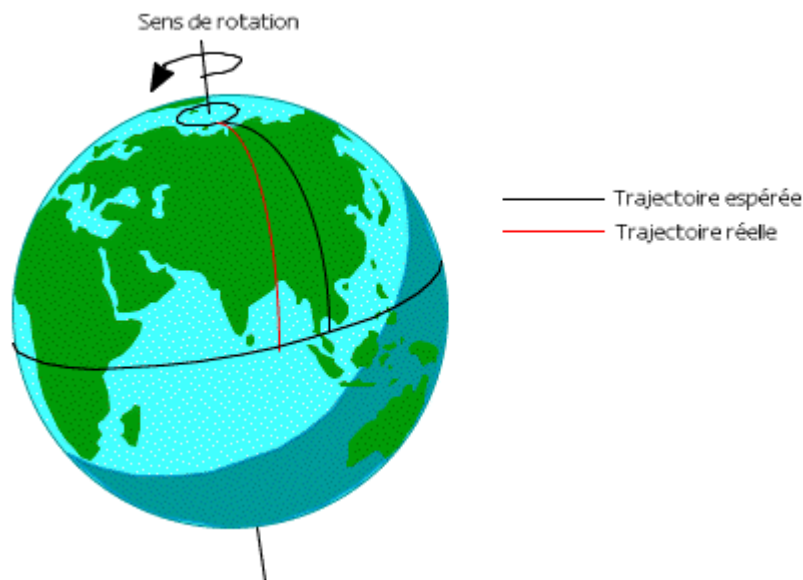


Figure: 30.23 - Illustration de l'effet de la force de Coriolis avec un projectile

C'est exactement ce phénomène que l'on observe dans le cas des cyclones (nous y reviendrons plus en détail dans notre étude de la météorologie dans le chapitre de Génie Météo). Une zone atmosphérique dépressionnaire (de faible pression relative) donnerait des courants atmosphériques (vents) convergents vers la dépression si la Terre ne tournait pas autour de son axe.

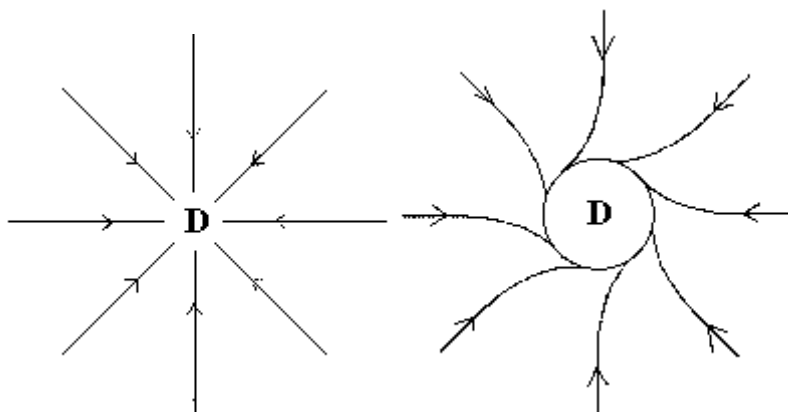


Figure: 30.24 - Génération des cyclons de par la force de Coriolis

La force de Coriolis due à la rotation de la Terre dévie donc les vents Nord-Sud en direction de l'Ouest et les vents Sud-Nord vers l'Est pour un observateur se situant au Pôle Nord. Nous observons dès lors la formation de cyclones tournants dans le sens contraire des aiguilles d'une montre dans l'hémisphère Nord et inversement dans l'hémisphère Sud (à cause de la direction du vecteur $\vec{\omega}$ dans cette partie de l'hémisphère).

Comme second exemple, considérons les oscillations d'un pendule. Pour des oscillations de faible amplitude, nous pouvons supposer que le mouvement du pendule se fait selon une trajectoire horizontale. Si l'on fait osciller le pendule initialement dans la direction Nord-Sud, la force de Coriolis va dévier le mouvement du pendule vers la droite pour un observateur situé au Pôle Nord. En d'autres termes, le pendule tourne dans le sens des aiguilles d'une montre dans l'hémisphère Nord et dans le sens contraire dans l'hémisphère Sud. Cet effet observable est nul à l'équateur (parallélisme parfait entre $\vec{v}_r$ et $\vec{\omega}$) et maximale aux Pôles.

Cet effet fut démontré de façon spectaculaire par le physicien français Jean Léon Foucault, quand en

1851 il suspendit un pendule de 67 mètres de long à l'intérieur du Dôme des Invalides. A chaque oscillation, le pendule faisait tomber du sable sur un cercle, ce qui démontrait expérimentalement que son plan d'oscillation est de $11^{\circ}15'$ par heure. L'expérience de Foucault est une preuve frappante de la rotation de la Terre. Même si la Terre était toujours couverte de nuages, cette expérience aurait montré aux physiciens que la Terre tournait.

Comme troisième exemple parlons des tourbillons que l'on peut observer dans la baignoire ou le lavabo. Ce n'est qu'une légende que ce dernier tourne différemment en fonction des hémisphères. Car la vitesse et la masse mises en jeu sont beaucoup trop faibles pour être observables dans de tels objets. Au fait, le sens de rotation est dû aux imperfections (aspérités) du siphon. Par contre, si vous allez en équateur, il y a des étudiants qui se font un plaisir de vous montrer que l'effet existe avec une petite expérience mise en place avec une allumette. En se déplaçant de dix mètres, ils vous montreront que le sens de rotation du siphon change en fonction de l'hémisphère dans lequel on se trouve!

12. THÉORÈMES DE KÖNIG

Nous avons vu jusqu'à maintenant, comment calculer le moment cinétique ou l'énergie cinétique d'un système dynamique par rapport à un unique référentiel (soit galiléen, soit barycentrique)

Les théorèmes de König donnent eux les moments cinétiques et l'énergie cinétique totale d'un système dynamique par rapport à un référentiel galiléen $R_{gal.}$ et barycentrique $R_{bary.}$

12.1. PREMIER THÉORÈME DE KÖNIG

Utilisons pour démontrer ce théorème le moment cinétique d'un corps de masse M (l'exemple étant toujours facilement extensible à un système dynamique discret ou continu de matière).

Exprimons le moment cinétique d'un élément m_i du corps solide par rapport à l'origine O du référentiel galiléen $R_{gal.}$ (noté: $/O, R$ par la suite):

$$\vec{b}_{/O,R} = \sum_i m_i \overrightarrow{OA_i} \times \vec{v}_{i/R} \quad (30.362)$$

Exprimons le moment cinétique dans $R_{bary.}$ par rapport à son centre de masse G (noté: $/G, R'$):

$$\vec{b}_{/G,R'} = \sum_i m_i \overrightarrow{GA_i} \times \vec{v}_{i/R'} \quad (30.363)$$

Le référentiel $R_{bary.}$ étant en translation par rapport à $R_{gal.}$, nous avons:

$$\vec{v}_{i/R} = \vec{v}_{i/R'} + \vec{v}_{G/R} \quad (30.364)$$

Sans oublier que:

$$\overrightarrow{OA_i} = \overrightarrow{OG} + \overrightarrow{GA_i} \quad (30.365)$$

que nous insérons dans l'expression du moment cinétique:

$$\vec{b}_{/O,R} = \sum_i m_i \overrightarrow{OA_i} \times \vec{v}_{i/R} = \sum_i m_i (\overrightarrow{OG} + \overrightarrow{GA_i}) \times (\vec{v}_{i/R'} + \vec{v}_{G/R}) \quad (30.366)$$

De par la propriété du produit vectoriel, nous avons:

$$\vec{b}_{/O,R} = \sum_i m_i \vec{OG} \times \vec{v}_{i/R'} + \sum_i m_i \vec{GA_i} \times \vec{v}_{i/R'} + \sum_i m_i \vec{OG} \times \vec{v}_{G/R} + \sum_i m_i \vec{GA_i} \times \vec{v}_{G/R} \quad (30.367)$$

Étudions maintenant la valeur que prend chacun des quatre termes de la relation précédente. Nous savons par la définition du centre de masse que (dans un cadre non relativiste):

$$\frac{\sum_i m_i \vec{GA_i}}{\sum_i m_i} = \vec{0} \Rightarrow \sum_i m_i \vec{GA_i} = \vec{0} \Rightarrow \frac{d}{dt} \sum_i m_i \vec{GA_i} = \sum_i m_i \left(\frac{d}{dt} \vec{GA_i} \right) = \sum_i m_i \vec{v}_{i/R'} = \vec{0} \quad (30.368)$$

d'où:

$$\sum_i m_i \vec{OG} \times \vec{v}_{i/R'} = \vec{OG} \times \left(\sum_i m_i \vec{v}_{i/R'} \right) = \vec{0} \quad (30.369)$$

et également:

$$\sum_i m_i \vec{GA_i} \times \vec{v}_{G/R} = \left(\sum_i m_i \vec{GA_i} \right) \times \vec{v}_{G/R} = \vec{0} \quad (30.370)$$

Finalement, il vient:

$$\vec{b}_{/O,R} = \sum_i m_i \vec{GA_i} \times \vec{v}_{i/R'} + \sum_i m_i \vec{OG} \times \vec{v}_{G/R} = \vec{b}_{/G,R'} + \vec{OG} \times \vec{v}_{G/R} \sum_i m_i = \vec{b}_{/G,R'} + M \vec{OG} \times \vec{v}_{G/R} \quad (30.371)$$

Donc finalement:

$$\vec{b}_{/O,R} = \vec{b}_{/G,R'} + M \vec{OG} \times \vec{v}_{G/R} \quad (30.372)$$

Ce théorème qui se rapporte à un point fixe permet l'application plus aisée du théorème du moment cinétique.

12.2. DEUXIÈME THÉORÈME DE KÖNIG

Utilisons pour démontrer ce théorème l'énergie cinétique d'un corps de masse M (l'exemple étant toujours facilement extensible à un système dynamique discret ou continu de matière).

Exprimons l'énergie cinétique d'un élément m_i du corps solide par rapport à l'origine O du référentiel galiléen R_{gal} . (noté: $/O, R$ par la suite):

$$E_{c/O,R} = \frac{1}{2} \sum_i m_i v_{i/R}^2 \quad (30.373)$$

Exprimons l'énergie cinétique dans R_{mov} par rapport à son centre de masse G (noté: $/G, R'$):

$$E_{c/G,R'} = \frac{1}{2} \sum_i m_i v_{i/R'}^2 \quad (30.374)$$

Avec de même que précédemment:

$$\vec{v}_{i/R} = \vec{v}_{i/R'} + \vec{v}_{G/R} \quad (30.375)$$

Il vient dès lors:

$$E_{c/O,R} = \frac{1}{2} \sum_i m_i v_{i/R}^2 = \frac{1}{2} \sum_i m_i (\vec{v}_{i/R'} + \vec{v}_{G/R})^2 \quad (30.376)$$

et donc:

$$E_{c/O,R} = \frac{1}{2} \sum_i m_i \vec{v}_{i/R'}^2 + \frac{1}{2} \sum_i m_i \vec{v}_{G/R}^2 + \frac{1}{2} \sum_i 2m_i \vec{v}_{i/R'} \vec{v}_{G/R} = E_{c/G,R'} + \frac{1}{2} M \vec{v}_{G/R}^2 + \vec{v}_{G/R} \sum_i m_i \vec{v}_{i/R'} \quad (30.377)$$

et comme pour le moment cinétique, de par la définition du centre de masse, nous avons:

$$\sum_i m_i \vec{v}_{i/R'} = \vec{0} \quad (30.378)$$

d'où le deuxième théorème de König:

$$E_{c/O,R} = E_{c/G,R'} + \frac{1}{2} M \vec{v}_{G/R}^2 \quad (30.379)$$

13. MOUVEMENTS OSCILLANTS

Le mouvement oscillatoire est le mouvement d'un corps qui va et vient de part et d'autre de sa position d'équilibre. Il existe une quantité incroyable de phénomènes physiques de ce genre. Nous allons traiter dans cette section les grands classiques à partir desquels les développements sur des phénomènes plus complexes s'inspirent.

Nous étudierons dans l'ordre les pendules des plus simples aux plus complexes et utiliserons souvent des résultats antérieurs pour en déterminer de nouveaux.

Nous retreindrons notre étude des mouvements oscillatoires aux pendules. Les autres viendront au fur et à mesure dans leurs chapitres respectifs.

Il existe neufs pendules très connus qui sont les suivants (ordre dans lequel nous les étudierons): pendule de Newton, pendule simple, pendule physique, pendule élastique, pendule conique, pendule de torsion, pendule de Foucault, pendule de Huygens.

13.1. PENDULE DE NEWTON

Nous n'allons pas trop nous étendre à décrire le pendule de Newton. Une photo suffira:

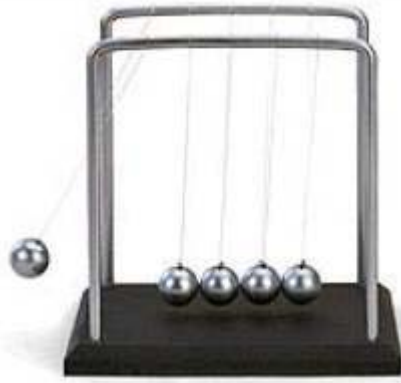


Figure: 30.25 - Pendule de Newton

Le principe de fonctionnement est le suivant:

Si vous lancez une bille, à l'extrémité une seule et unique bille se déplacera. Cela semble logique et cohérent d'après la conservation de la quantité de mouvement qui découle de la conservation de l'énergie comme nous l'avons déjà vu.

Un peu plus curieux, lorsque vous lancez initialement deux billes, ce sont deux billes qui se déplacent à l'autre extrémité!

La démonstration est simple et le fonctionnement se base sur une condition très simple que nous allons déterminer pour le cas particulier de deux billes (c'est toujours le même principe pour un nombre de billes supérieur):

Soient p_{1i}, p_{2i} les quantités de mouvement des deux billes initiales et p_{1f}, p_{2f} celles des deux billes se situant à l'autre extrémité. Nous avons donc:

$$m_1 v_{1i} + m_2 v_{2i} = m_1 v_{1f} + m_2 v_{2f} \quad (30.380)$$

Nous avons pour l'énergie cinétique:

$$\frac{1}{2} m_1 v_{1i}^2 + \frac{1}{2} m_2 v_{2i}^2 = \frac{1}{2} m_1 v_{1f}^2 + \frac{1}{2} m_2 v_{2f}^2 \quad (30.381)$$

après regroupement et simplification de chacune des deux relations précédentes:

$$\begin{aligned} m_1 (v_{1i} - v_{1f}) &= m_2 (v_{2f} - v_{2i}) \\ m_1 (v_{1i}^2 - v_{1f}^2) &= m_2 (v_{2f}^2 - v_{2i}^2) \end{aligned} \quad (30.382)$$

De la deuxième relation ci-dessus nous avons:

$$m_1 (v_{1i} - v_{1f})(v_{1i} + v_{1f}) = m_2 (v_{2f} - v_{2i})(v_{2f} + v_{2i}) \quad (30.383)$$

En divisant par la première il reste:

$$(v_{1i} + v_{1f}) = (v_{2f} + v_{2i}) \quad (30.384)$$

Nous en tirons:

$$\begin{aligned} v_{1f} &= v_{2f} + v_{2i} - v_{1i} \\ v_{2f} &= v_{1i} + v_{1f} - v_{2i} \end{aligned} \quad (30.385)$$

Injectons la première de ces deux relations dans:

$$m_1 (v_{1i} - v_{1f}) = m_2 (v_{2f} - v_{2i}) \quad (30.386)$$

Nous avons alors:

$$m_1 (v_{1i} - (v_{2f} + v_{2i} - v_{1i})) = m_2 (v_{2f} - v_{2i}) \quad (30.387)$$

Ce qui nous donne en réarrangeant:

$$v_{2f} = \frac{v_{2i}(m_2 - m_1) + 2m_1 v_{1i}}{m_1 + m_2} \quad (30.388)$$

Au final en procédant de même pour l'autre vitesse finale, nous déduisons l'expression des deux vitesses après le choc élastique:

$$v_{1f} = \frac{v_{1i}(m_1 - m_2) + 2m_2 v_{2i}}{m_1 + m_2}, \quad v_{2f} = \frac{v_{2i}(m_2 - m_1) + 2m_1 v_{1i}}{m_1 + m_2} \quad (30.389)$$

Hypothèse: Supposons maintenant qu'en prenant une seule des billes avec $v_{2i} = 0$, il y en ait deux qui partent à l'autre extrémité tel que:

$$v_{1f} = \frac{v_{1i}(m_1 - m_2)}{m_1 + m_2}, \quad v_{2f} = \frac{2m_1 v_{1i}}{m_1 + m_2} \quad (30.390)$$

et dans cette dernière situation considérons le cas où toutes les billes du pendule de Newton ont la même masse (cas correspondant à celui que l'on trouve dans le commerce). Alors:

$$m_1 = m_2 \Rightarrow v_{1f} = 0, \quad v_{2f} = v_{1i} \quad (30.391)$$

Nous voyons que notre hypothèse initiale est fautive: si à masses égales, une seule bille est lancée alors, à l'autre extrémité, une seule bille partira de par la conservation de la quantité de mouvement (hypothèse des "**chocs élastiques**")!

Par contre, si nous lançons deux billes dans un pendule de Newton composé de masses identiques nous avons après simplification des équations:

$$\alpha = \frac{x}{L} \quad (30.397)$$

Si nous dérivons cette expression par rapport au temps... Nous obtenons alors:

$$\frac{d\alpha}{dt} = \frac{d\left(\frac{x}{L}\right)}{dt} = \frac{1}{L} \cdot \frac{dx}{dt} = \frac{v}{L} \quad (30.398)$$

Si nous revenons à:

$$v^2 = 2 \cdot g \cdot L \cdot (\cos \alpha_0 - \cos \alpha) \quad (30.399)$$

et que nous le dérivons, nous obtenons:

$$2 \cdot v \cdot \frac{dv}{dt} = 2 \cdot g \cdot L \cdot \sin \alpha \cdot \frac{d\alpha}{dt} \Rightarrow 2 \cdot v \cdot \alpha = 2 \cdot g \cdot L \cdot \frac{v}{L} \cdot \left(\sin \left(\frac{x}{L} \right) \right) \Rightarrow \alpha = g \cdot \sin \left(\frac{x}{L} \right) \quad (30.400)$$

Si l'angle α est petit, nous pouvons remplacer avec l'aide de la série de Taylor (cf. chapitre de Suites Et Séries) et sans erreur trop grave, $\sin \alpha$ par le premier terme de son développement en série:

$$\sin \alpha \cong \alpha - \frac{\alpha^3}{3!} + \frac{\alpha^5}{5!} - \dots \quad (30.401)$$

et comme $\alpha = \frac{x}{L}$ nous obtenons:

$$\alpha = g \cdot \frac{x}{L} \quad (30.402)$$

Etant donné que dans un cadre périodique:

$$\alpha_{\max} = \omega^2 \cdot x \quad (30.403)$$

nous pouvons alors écrire que:

$$\omega^2 = \frac{g}{L} \quad (30.404)$$

d'où:

$$\omega = \sqrt{\frac{g}{L}} \quad (30.405)$$

Comme $\omega = \frac{2\pi}{T}$ nous pouvons poser:

$$T \cong 2\pi \cdot \sqrt{\frac{L}{g}} \quad (30.406)$$

Donc la période de balancement est indépendante de l'amplitude ce qui explique pourquoi le nombre de balancements par minute d'un pendule simple est constant, quelle que soit l'ardeur que nous mettions à le faire balancer... Nous parlons alors "**d'isochronisme**".

Si:

$$L = l_0 + N \cdot P \quad (30.407)$$

où l_0 est la position du centre de masse de l'objet et N le nombre de maillons éventuels que l'on aurait pris pour la longueur L de la chaîne et P étant le pas de la chaîne.

Ce qui nous donne finalement:

$$T \cong 2\pi \sqrt{\frac{l_0 + N \cdot P}{g}} \quad (30.408)$$

13.3. PENDULE PHYSIQUE

Nous appelons "**pendule physique**" un solide quelconque pouvant osciller librement dans la pesanteur, autour d'un axe A , avec une petite amplitude ($\theta < 5^\circ$).

Son mouvement est déterminé par l'équation suivante:

$$M = J_A \cdot \frac{d^2 \theta}{dt^2} \quad (30.409)$$

où M est le moment de rappel et J_A le moment d'inertie du pendule par rapport à son axe d'appui A .

En faisant une analyse des forces sur notre pendule nous obtenons une autre relation pour M :

$$M = -m \cdot g \cdot r = -m \cdot g \cdot d \cdot \sin \theta = -m \cdot g \cdot d \cdot \theta \quad (30.410)$$

pour $-\theta_0 \leq \theta \leq \theta_0$ et où d est la distance de l'axe d'appui du pendule à son centre de masse. Le terme négatif apparaît ici pour exprimer le fait que la période diminue avec le temps. Comme l'angle θ est petit, nous avons remplacé $\sin \theta$ et sans erreur trop grave par le premier terme de son développement en série de Taylor:

$$\sin \theta \cong \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \dots \quad (30.411)$$

Donc le moment de rappel peut s'écrire:

$$M = J_A \cdot \frac{d^2 \theta}{dt^2} = -m \cdot g \cdot d \cdot \theta \quad (30.412)$$

d'où l'équation différentielle du mouvement:

$$J_A \frac{d^2 \theta}{dt^2} + m \cdot g \cdot d \cdot \theta = 0 \Leftrightarrow \frac{d^2 \theta}{dt^2} + \left(\frac{m \cdot g \cdot d}{J_A} \right) \cdot \theta = 0 \quad (30.413)$$

Nous avons vu dans les mouvements harmoniques oscillants que nous obtenions la position angulaire d'une masse par la relation:

$$\theta = \theta_0 \sin(\omega \cdot t) \quad (30.414)$$

ce qui nous permet d'écrire:

$$\frac{d^2 (\theta_0 \sin(\omega \cdot t))}{dt^2} + \left(\frac{m \cdot g \cdot d}{J_A} \right) \cdot \theta_0 \sin(\omega \cdot t) = -\theta_0 \omega^2 \sin(\omega \cdot t) + \left(\frac{m \cdot g \cdot d}{J_A} \right) \theta_0 \sin(\omega \cdot t) = 0 \quad (30.415)$$

et par simplification nous obtenons:

$$\omega = \sqrt{\frac{m \cdot g \cdot d}{J_A}} \quad (30.416)$$

d'où:

$$T = 2\pi \sqrt{\frac{J_A}{m \cdot g \cdot d}} \quad (30.417)$$

En exprimant le moment d'inertie J_A au moyen du théorème de Steiner en déduisons que:

$$J_A = J_G + m \cdot d^2 \quad (30.418)$$

et en introduisant encore le rayon de giration k :

$$J_A = m \cdot k^2 + m \cdot d^2 \quad (30.419)$$

d'où:

$$J_A = m \cdot d \cdot \left(\frac{k^2}{d} + d \right) \quad (30.420)$$

Soit x la position de l'axe de rotation A mesurée par rapport à une origine quelconque et a la position du centre de gravité par rapport à la même origine nous avons:

$$d = |x - a| \quad (30.421)$$

tel que représenté ci-dessous:

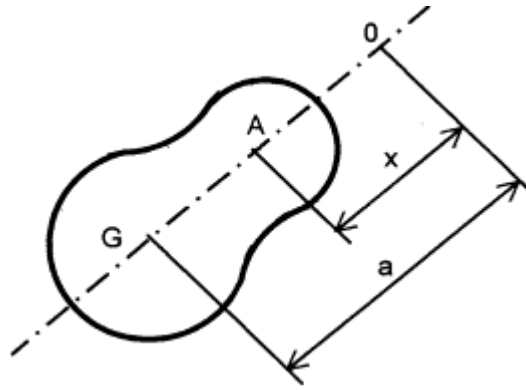


Figure: 30.27 - Pendule physique dit "pendule d'inertie"

d'où:

$$J_A = m \cdot |x - a| \cdot \left(\frac{k^2}{|x - a|} + |x - a| \right) \quad (30.422)$$

ce qui nous donne aussi pour la période:

$$T = 2\pi \sqrt{\frac{1}{g} \cdot \left(\frac{k^2}{|x - a|} + |x - a| \right)} \quad (30.423)$$

comme la racine nous gêne nous élevons le tout au carré, ce qui nous donne finalement

$$T^2 = \frac{4\pi^2}{g} \cdot \left(\frac{k^2}{|x - a|} + |x - a| \right) \quad (30.424)$$

Comme nous connaissons x et T , cette relation nous permettrait à partir du tracé d'un graphique de déterminer la position de G et k .

Ainsi, en portant sur un graphique T^2 en fonction de x :

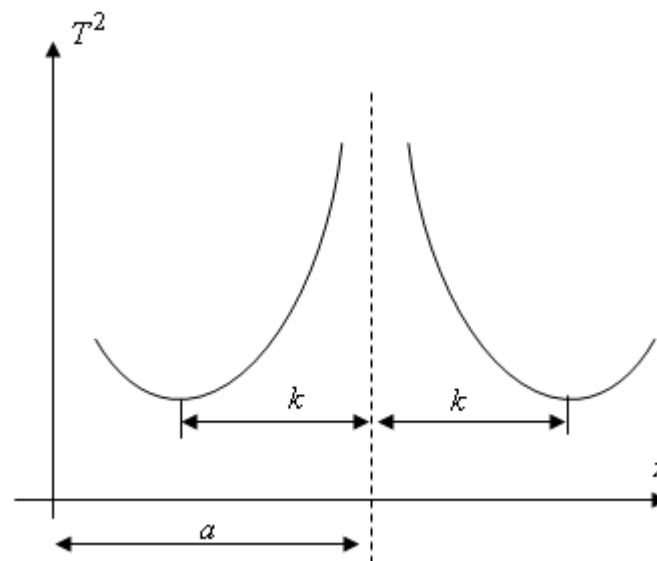


Figure: 30.28 - Illustration de l'asymptote vertical pour la détermination du centre d'inertie

La courbe obtenue présente une asymptote verticale ($T = \infty$) pour $x = a$ et deux minima.

En dérivant T^2 par rapport à x et en annulant les dérivées, nous trouvons la position des minima:

$$x = a \pm k \quad (30.425)$$

13.4. PENDULE ÉLASTIQUE

Étudions maintenant les oscillations propres d'un solide suspendu à un ressort élastique tel qu'il oscille. Après l'écart du solide de la position d'équilibre, il accomplira des oscillations harmoniques dans le sens vertical, si le ressort élastique subit des déformations proportionnelles à l'allongement du ressort.

Nous aurons souvent dans ce site à faire avec de petits mouvements autour d'une position d'équilibre. Ce type de mouvement caractéristique de ce que nous appelons un "**oscillateur harmonique**" est très fréquent. Il se généralise à toutes sortes de situations physiques, telles que les circuits RLC (cf. chapitre de Génie Électrique), le modèle quantique corpusculaire et ondulatoire de l'atome, les résonateurs à quartz ou toute autre structure vibrante faiblement autour de son point d'équilibre.

Nous savons que la force de rappel d'un ressort est proportionnelle et opposée à la déformation telle que (voir le chapitre de Génie Civil pour la démonstration):

$$F = ma = -k \cdot \Delta L \quad (30.426)$$

L'équation différentielle de l'oscillateur harmonique peut donc s'écrire:

$$\frac{d^2x}{dt^2} = -\frac{k}{m}x \quad (30.427)$$

Nous prendrons la démarche très simple qui consiste à essayer une solution, en l'occurrence:

$$x = \cos(\omega t) \quad (30.428)$$

C'est une solution, car en effet:

$$\begin{aligned} \frac{dx}{dt} &= -\omega \sin(\omega t) \\ \frac{d^2x}{dt^2} &= -\omega^2 \cos(\omega t) \end{aligned} \quad (30.429)$$

pour autant que nous prenions la fréquence propre:

$$\omega = \sqrt{\frac{k}{m}} \quad (30.430)$$

Nous avons aussi le "**mode propre**":

$$x = \sin(\omega t) \quad (30.431)$$

comme solution.

Une solution générale est donc:

$$x(t) = A \sin(\omega t) + B \cos(\omega t) \quad (30.432)$$

Pour trouver A et B , il faut spécifier les conditions initiales. Prenons par exemple:

$$x = \Delta l \quad v = v_0 \quad (30.433)$$

à $t = 0$.

Nous avons alors:

$$\begin{aligned} x(t=0) &= B = \Delta l \\ v(t) &= A\omega \cos(\omega t) - B\omega \sin(\omega t) \\ v(0) &= A\omega = v_0 \end{aligned} \quad (30.434)$$

Calculons maintenant le travail (énergie) nécessaire pour déformer l'oscillateur harmonique. Nous avons:

$$W = \int_{x_1}^{x_2} F dx = \int_{x_1}^{x_2} kx dx = k \int_{x_1}^{x_2} x dx = \frac{1}{2} k (x_2^2 - x_1^2) = \frac{1}{2} k \Delta x^2 \quad (30.435)$$

Ainsi, l'énergie potentielle élastique dans un ressort de constante k , ayant subi une déformation x est donc donnée par:

$$E_{\text{élastique}} = \frac{1}{2} kx^2 \quad (30.436)$$

Pour une description plus réaliste, une meilleure modélisation, nous allons supposer que l'oscillateur est soumis à une force supplémentaire représentant les frottements. Il arrive souvent que l'approximation par laquelle la force de frottement est proportionnelle à la vitesse, et opposée à la vitesse, soit une bonne approximation. Ce n'est pas la seule possible, et ce n'est pas toujours la meilleure. Nous parlerons des forces de frottement plus tard.

Ainsi nous considérons une force de friction de la forme (ne pas confondre avec la notation du moment cinétique qui n'a absolument aucun rapport):

$$\vec{F}_f = -b \vec{v} \quad (30.437)$$

Pour notre système de coordonnées:

$$F_f = -b \frac{dx}{dt} \quad (30.438)$$

La deuxième loi de Newton impose:

$$m \frac{d^2 x}{dt^2} = -kx - b \frac{dx}{dt} \quad (30.439)$$

Pour se conformer à une notation usuelle dans le cadre de l'oscillateur, nous notons:

$$\frac{k}{m} = \omega_o^2 \quad \frac{b}{2m} = \gamma \quad (30.440)$$

d'où l'équation différentielle:

$$\frac{d^2 x}{dt^2} + 2\gamma \frac{dx}{dt} + \omega_o^2 x = 0 \quad (30.441)$$

Nous prenons la fonction d'essai:

$$x = e^{\lambda t} \quad (30.442)$$

En substituant, nous trouvons:

$$e^{\lambda t} (\lambda^2 + 2\gamma\lambda + \omega_o^2) = 0 \quad (30.443)$$

Comme nous cherchons des solutions non nulles ($e^{\lambda t} \neq 0$) il faut que:

$$(\lambda^2 + 2\gamma\lambda + \omega_o^2) = 0 \quad (30.444)$$

d'où:

$$\begin{aligned} \lambda_1 &= -\gamma + \sqrt{\gamma^2 - \omega_o^2} \\ \lambda_2 &= -\gamma - \sqrt{\gamma^2 - \omega_o^2} \end{aligned} \quad (30.445)$$

et la solution générale est:

$$x(t) = A_1 e^{\lambda_1 t} + A_2 e^{\lambda_2 t} = e^{-\gamma t} \left(A_1 e^{\sqrt{\gamma^2 - \omega_o^2} t} + A_2 e^{-\sqrt{\gamma^2 - \omega_o^2} t} \right) \quad (30.446)$$

où deux constantes sont déterminées par les conditions initiales.

Nous verrons qu'il correspond à un amortissement faible. En effet, nous pouvons écrire avec des racines carrées réelles:

$$\lambda_{1,2} = -\gamma \pm \sqrt{\gamma^2 - \omega_o^2} = -\gamma \pm i\sqrt{\omega_o^2 - \gamma^2} \quad (30.447)$$

et la solution générale peut alors s'écrire:

$$x(t) = e^{-\gamma t} \left(A_1 e^{i\sqrt{\omega_o^2 - \gamma^2} t} + A_2 e^{-i\sqrt{\omega_o^2 - \gamma^2} t} \right) \quad (30.448)$$

En utilisant les propriétés complexes des exponentielles et en particulier la "formule d'Euler" (cf. chapitre sur les Nombres):

$$x(t) = e^{-\gamma t} \left(A_1 e^{+i\sqrt{\omega_0^2 - \gamma^2}t} + A_2 e^{-i\sqrt{\omega_0^2 - \gamma^2}t} \right) = e^{-\gamma t} \begin{pmatrix} A_1 \cos(\sqrt{\omega_0^2 - \gamma^2}t) + iA_1 \sin(\sqrt{\omega_0^2 - \gamma^2}t) \\ + A_2 \cos(-\sqrt{\omega_0^2 - \gamma^2}t) + iA_2 \sin(-\sqrt{\omega_0^2 - \gamma^2}t) \end{pmatrix} \quad (30.449)$$

Choisissons $A_1 = A_2$ et rappelons que $\sin(-\alpha) = -\sin(\alpha)$ (cf. chapitre de Trigonométrie). Ainsi:

$$x(t) = e^{-\gamma t} \left(A_2 \cos(\sqrt{\omega_0^2 - \gamma^2}t) + A_2 \cos(-\sqrt{\omega_0^2 - \gamma^2}t) \right) \quad (30.450)$$

et comme nous avons aussi $\cos(-\alpha) = \cos(\alpha)$. Alors:

$$x(t) = e^{-\gamma t} 2A_2 \cos(\sqrt{\omega_0^2 - \gamma^2}t) = e^{-\gamma t} C \cos(\sqrt{\omega_0^2 - \gamma^2}t) \quad (30.451)$$

posons $\omega_1 = \sqrt{\omega_0^2 - \gamma^2}$ et comme la fonction trigonométrique est périodique à $\phi = \pm 2k\pi$ avec $k \in \mathbb{N}$ alors:

$$x(t) = e^{-\gamma t} C \cos(\omega_1 t + \phi) \quad (30.452)$$

L'allure générale de la $x(t)$ normalisée à l'unité est la suivante:

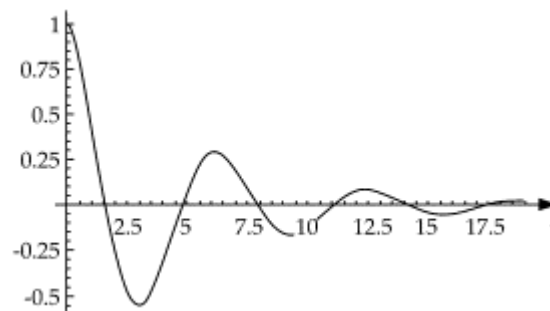


Figure: 30.29 - Illustration de l'amortissement de l'amplitude d'une pendule élastique

Quand $\omega_0 = \gamma$ nous disons qu'il y a "**amortissement critique**", quand $\omega_0 < \gamma$, qu'il y a "**amortissement sur-critique**".

Le rapport:

$$\frac{\omega_0}{2\gamma} \quad (30.453)$$

est quant à lui appelé "**facteur de qualité**".

13.5. PENDULE CONIQUE

Le pendule conique consiste à prendre une masse m considérée comme ponctuelle et suspendue en A d'un fil $OA = L$ fixé en O.

La masse étant écartée d'un angle α de la verticale, l'objectif de ce pendule est fréquemment (car c'est le cas le plus simple) de déterminer la dépendance entre l'angle et la vitesse si l'on considère que les

trajectoires sont circulaires.

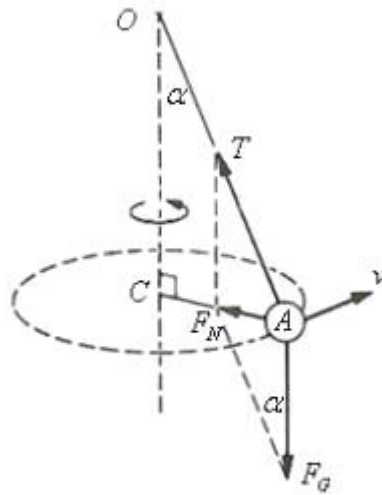


Figure: 30.30 - Illustration du pendule conique

La masse m se déplace autour de la verticale OC , en décrivant un cercle de rayon:

$$r = CA = OA \sin \alpha = L \sin \alpha \quad (30.454)$$

Les forces suivantes agissent sur la masse m :

$$\begin{aligned} F_g &= mg \\ F_N &= m \frac{v^2}{R} = m \frac{(\omega R)^2}{R} = m \omega^2 R = m \omega^2 L \sin \alpha \end{aligned} \quad (30.455)$$

D'après la figure, nous voyons que:

$$\tan \alpha = \frac{F_N}{F_g} = \frac{\omega^2 L \sin \alpha}{g} \quad (30.456)$$

ou, comme:

$$\tan \alpha = \frac{\sin \alpha}{\cos \alpha} \quad (30.457)$$

alors:

$$\cos \alpha = \frac{g}{\omega^2 L} \quad (30.458)$$

L'angle α est donc d'autant plus grand que la vitesse angulaire ω est élevée, ce que confirme l'expérience. Pour cette raison, le pendule conique fut longtemps utilisé comme régulateur de vitesse sur les machines à vapeur (il ferme l'arrivée de vapeur quand la vitesse dépasse une limite fixée à l'avance et l'ouvre quand elle tombe au-dessous de cette valeur).

Nous avons aussi:

$$\cos \alpha = \frac{g}{\left(\frac{v}{R}\right)^2 L} \Rightarrow v = \sqrt{\frac{R^2 g}{\cos \alpha L}} = \sqrt{\frac{g}{\cos \alpha L}} R = \sqrt{\frac{g}{\cos \alpha L}} L \sin \alpha \quad (30.459)$$

d'où après simplification:

$$v = \sqrt{\frac{gL}{\cos \alpha}} \sin \alpha \quad (30.460)$$

13.6. PENDULE DE TORSION

Le pendule de torsion est un système qui fut utilisé par Coulomb pour la mesure de la charge électrique élémentaire et par Cavendish pour la mesure de la constante gravitationnelle G .

Le pendule de torsion consiste en un solide rigide suspendu à fil de torsion vertical. Lors des oscillations le fil exerce un moment de rappel que l'on supposera proportionnel à l'angle de torsion θ :

$$M = -k\theta \quad (30.461)$$

où k est la "**constante de torsion**" de ce fil particulier (cf. chapitre de Génie Mécanique).

Nous avons donc:

$$M = J_z \cdot \frac{d^2 \theta}{dt^2} = -k\theta \quad (30.462)$$

soit l'équation différentielle:

$$\frac{d^2 \theta}{dt^2} + \frac{k}{J_z} \theta = 0 \quad (30.463)$$

Par analogie avec le pendule physique où nous avons une équation différentielle identique à un facteur près, il vient:

$$T = 2\pi \sqrt{\frac{J_z}{k}} \quad (30.464)$$

13.7. PENDULE DE FOUCAULT

Le pendule de Foucault est une expérience formidable pour rendre compte de la rotation de la Terre. Il existe plusieurs méthodes mathématiques pour analyser le comportement du pendule de Foucault. Nous avons choisi de présenter la plus simple qui ne nécessite que peu de pages de calcul.

D'abord un petit texte explicatif peut s'avérer pertinent tellement cette expérience est importante.

L'expérience de Foucault a pour but de démontrer que la Terre tourne sur elle-même. Vous lancez un balancier (une bille de plomb au bout d'un fil). Il a un mouvement de va-et-vient régulier dans la même direction. Si vous l'emportez dans une voiture et que vous ne tournez pas trop brusquement, le pendule se moque des virages: il continue à battre dans la même direction. C'est qu'un pendule reste toujours

dans le même plan, malgré les mouvements de son support.

C'est pourquoi le physicien français Léon Foucault eut l'idée d'attacher un lourd balancier de 67 mètres de long sous le dôme du Panthéon, en présence de Napoléon III et de quelques savants. A chacune de ses allées et venues, le pendule venait écorner un tas de sable où il laissait une marque. Or, la trace n'était jamais à la même place: il y avait 3 à 4 millimètres d'écart entre un balancement et le suivant, 16 secondes plus tard. Le pendule restait dans le même plan, mais le Panthéon, Paris, la Terre tournaient!

Soit la figure ci-dessous:

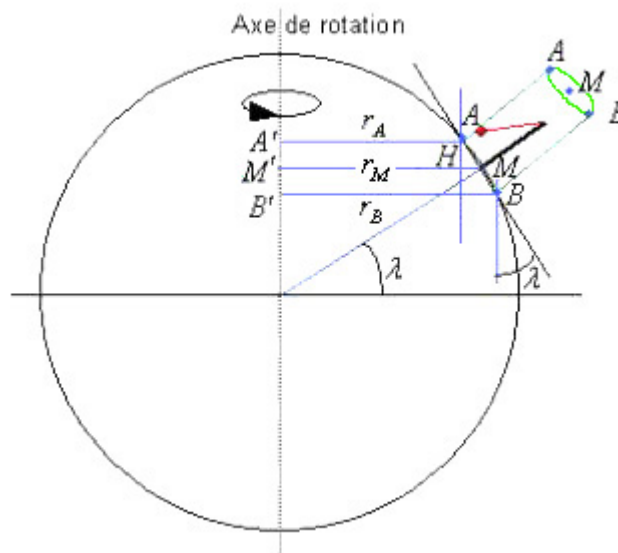


Figure: 30.31 - Illustration élémentaire du pendule de Foucault

Nous considérons que c'est la vue d'un référentiel géocentrique (la Terre) vu en coupe selon un plan qui contient l'axe de rotation.

La taille du pendule est bien évidemment exagérée sur la figure. Il oscille cependant quand même dans un plan méridien, entre A et B (un observateur terrestre voit la droite AB tourner par rapport au sol terrestre selon le cercle vert, vu en perspective, dans le sens rétrograde).

Soit T la période de rotation de A (ou B). La vitesse de A (v_a), sur ce cercle, est due au fait que, dans le référentiel géocentrique, le point M, à la verticale du point de suspension, et le point du sol terrestre coïncidant avec A à un instant donné, n'ont pas la même vitesse dans le référentiel géocentrique: le point M étant plus éloigné de l'axe de rotation terrestre il va plus vite que A (de même la vitesse de B étant supérieure à la vitesse de M).

La différence de ces vitesses se calcule aisément en supposant que la rotation terrestre est uniforme en raisonnant sur une période d'une journée (sidérale) T_0 .

Nous savons que:

$$v = \omega \cdot r = \frac{2\pi}{T} \quad (30.465)$$

De ceci il découle facilement que:

$$\Delta v = \frac{2\pi}{T_0} (r_M - r_A) \quad (30.466)$$

étant donné que dans le triangle AHM :

$$r_M - r_A = HM = AM \cdot \sin \lambda \quad (30.467)$$

alors:

$$\Delta v = \frac{2\pi}{T_0} AM \cdot \sin \lambda \quad (30.468)$$

Or, Δv n'est autre que:

$$v_a = \frac{2\pi}{T} AM \quad (30.469)$$

Donc:

$$T = \frac{T_0}{\sin \lambda} \quad (30.470)$$

Nous avons donc obtenu l'expression de la période du pendule de Foucault.

Exemple:

La période du pendule du Panthéon (aller et retour) est de 16.5 secondes, l'amplitude maximale de 6 mètres et le temps d'amortissement de 6 heures. Nous pouvons ainsi observer un déplacement de plusieurs millimètres par aller et retour du pendule.

Remarque: Le sens de la rotation est celui des aiguilles d'une montre, pour un observateur placé au-dessus du pendule, dans l'hémisphère Nord ; et dans le sens contraire du sens de rotation des aiguilles d'une montre dans l'hémisphère Sud.

Aux pôles (où l'angle est de 90° et le sinus unitaire), la période du pendule égale celle de la Terre et est donc de 24h. A l'équateur (où l'angle est de 0° et le sinus nul), la période de rotation du plan d'oscillation est infinie: le plan d'oscillation est fixe par rapport à la Terre. A Paris (où l'angle est de $48^\circ 52'$ et le sinus 0.75), la période de rotation est de 31 heures et 57 minutes.

Cependant, l'importance du pendule de Foucault est autre...

Le plan d'oscillation du pendule est en réalité fixe et c'est la rotation de la Terre sur elle-même qui donne lieu à une rotation apparente. Mais finalement... quel est le système de référence ?

En effet, tout mouvement est relatif. Si la Terre est en rotation, elle l'est par rapport à quelque chose. Nous ne pouvons pas parler d'un mouvement sans définir un cadre de référence. La question qui se pose donc est de savoir par rapport à quel système de référence le plan d'oscillation du pendule est fixe.

La première idée qui vient à l'esprit consiste à dire que le plan du pendule est fixe par rapport au Soleil.

Mais, si Foucault avait réussi à construire un pendule capable d'osciller suffisamment longtemps, disons pendant un mois, il se serait aperçu que le plan d'oscillation dérivait également par rapport à la position du Soleil. Notre étoile ne fait donc pas partie du système de référence en question.

Peut-être faut-il alors considérer les étoiles proches du Soleil ? Mais là aussi, si l'expérience pouvait durer suffisamment longtemps, elle montrerait que le plan des oscillations se déplace nettement par rapport aux étoiles après quelques années. Quel objet choisir dans ce cas ? Le centre galactique, la galaxie d'Andromède, le Groupe Local, le superamas local ? Chacun de ces objets donnerait l'illusion d'être fixe par rapport au plan des oscillations, mais finirait, après un temps de plus en plus long, par révéler une dérive.

Finalement, en dernier recours, nous pouvons considérer les objets les plus lointains, les galaxies ou quasars situés à des milliards d'années-lumière. Avec ce système de référence, et si l'expérience de Foucault était réalisable, le plan des oscillations serait enfin fixe et il n'y aurait plus de dérive. Ce n'est donc qu'en considérant les objets les plus lointains, en fait l'Univers observable dans son ensemble, que nous pouvons obtenir un cadre par rapport auquel le plan des oscillations se stabilise.

Le pendule de Foucault se moque donc de la présence de la Terre, du Soleil ou de la Galaxie. Son mouvement lui est directement dicté par l'Univers dans son ensemble. Cette expérience met en évidence une sorte de lien mystérieux entre chaque point et l'Univers tout entier. Jusqu'à nouvel ordre, la nature de ce lien reste inconnue.

Une conclusion similaire fut tirée par le physicien autrichien Ernst Mach à la fin du XIX^e siècle (nous retrouverons le "principe de Mach" dans le chapitre de Relativité Restreinte).

D'après la physique de Newton, le produit de la masse d'un corps par son accélération est égal à la force qui s'exerce sur lui. Par conséquent, pour une force donnée, plus un objet est massif, plus son accélération est faible. De ce point de vue, la masse est donc une mesure de l'inertie du corps, c'est-à-dire de sa faculté à résister à une force.

Supposons maintenant que toute la matière de l'Univers disparaisse, excepté pour ce corps. Ce dernier est alors complètement isolé et plus aucune force ne s'exerce sur lui. Cela signifie, d'après la physique de Newton, que le produit de sa masse par son accélération est égal à zéro. Or l'accélération ne peut pas être nulle. En effet, comme toute la matière de l'Univers a disparu, il n'y a plus de système de référence par rapport auquel on pourrait définir la vitesse ou l'accélération. Cette dernière est donc indéfinie et non pas nulle. D'un point de vue mathématique, il ne reste qu'une seule possibilité, que la masse du corps soit nulle.

Ce raisonnement montre que la masse et l'inertie d'un corps ne sont pas vraiment des propriétés de l'objet lui-même, mais plutôt le résultat d'une interaction avec le reste de l'Univers. Tout comme le pendule de Foucault, le principe de Mach nous montre qu'il doit exister une sorte de connexion entre les propriétés locales d'un corps et les propriétés globales de l'Univers. Comme dans le cas précédent, la nature de cette connexion mystérieuse reste à déterminer.

13.8. PENDULE DE HUYGENS

Nous cherchons à construire un pendule dont la période soit indépendante de l'amplitude (et non pas juste en approximation comme nous l'avons vu plus haut!), ce qu'il est d'usage d'appeler "**isochronisme rigoureux**". Pour cela nous disposons deux lamelles de forme cycloïdale à des positions symétriques et déterminées telles que représentées sur la figure ci-dessous:

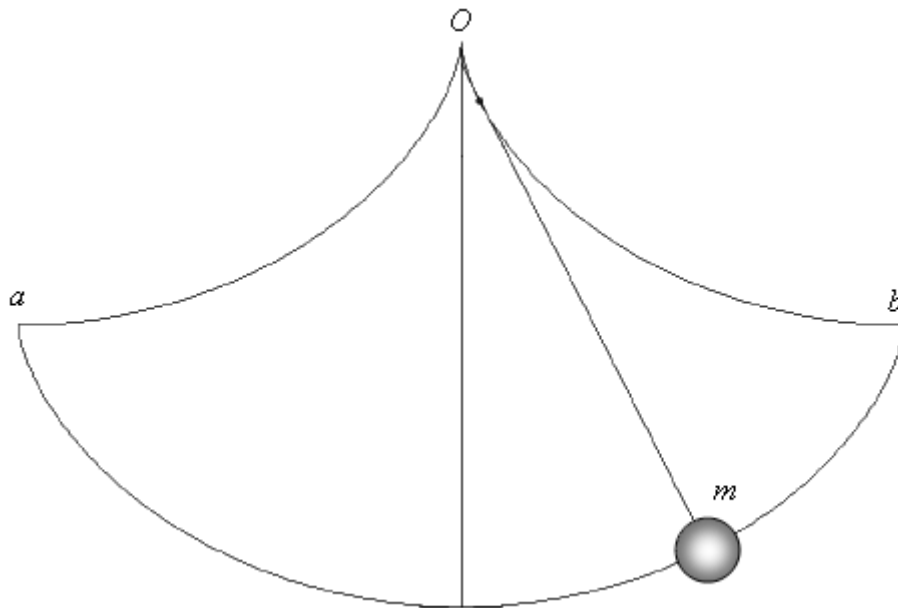


Figure: 30.32 - Principe du pendule de Huyghens

Le choix de la cycloïde est dû au fait qu'il s'agit d'une courbe "**brachistochrone**" (voir la définition plus bas) et depuis les travaux de Huygens en 1659, nous savons aussi qu'il s'agit d'une courbe "**tautochrone**" (les balanciers dans les montres modernes ont par tradition cette forme). C'est-à-dire que les corps qui tombent dans une cycloïde renversée arrivent au point le plus bas dans le même temps, de quelque hauteur qu'ils commencent à tomber.

Donc contrairement à une idée reçue, le chemin le plus rapide pour un corps en mouvement non horizontal tombant sur un support solide n'est pas la ligne droite.

En effet, l'un des problèmes les plus connus de l'histoire des mathématiques est le problème du brachistochrone qui consiste donc à trouver la courbe le long de laquelle une particule glisserait d'un point à un autre en un minimum de temps en étant soumis à un champ uniforme de pesanteur. Ce problème a été posé par Jean Bernoulli en 1696 comme un challenge pour les mathématiciens de son époque (et s'en fut un !!!). La solution fut trouvée par Jean Bernoulli lui-même ainsi que par son frère Jacques Bernoulli, Newton, Leibniz et le marquis de l'Hospital. Le problème brachistochrone est important dans le développement des mathématiques et s'avère être une des applications principales de la méthode du calcul des variations.

Nous considérons dans le champ de la pesanteur deux points a et b et un point matériel m se déplaçant sans frottement sur une courbe d'extrémités a et b . Déterminer la courbe, appelée brachistochrone, pour laquelle le temps de parcours est minimal lorsque le point m part du point a avec une vitesse nulle.

Considérons le schéma ci-dessous:

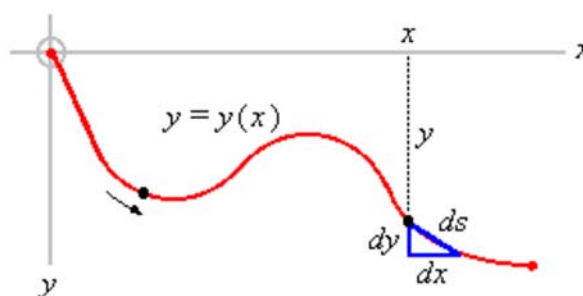


Figure: 30.33 - Chemin quelconque pour attaquer le problème sans biais

A l'abscisse x sur le graphe, l'énergie potentielle perdue est $E_p = mgy$, équivalente à l'énergie cinétique acquise par le point matériel depuis le départ telle que:

$$\frac{1}{2}mv^2 = mgh \quad (30.471)$$

D'où sans trop de surprises:

$$v = \sqrt{2gh} \quad (30.472)$$

La vitesse v est mesurée le long de la courbe si bien que nous devons réécrire l'expression en composantes horizontales et verticales:

Nous allons poser que s représente l'abscisse curviligne et ds l'accroissement de cette distance le long de la courbe. dx et dy représentent les composantes horizontale et verticale de ds .

Ainsi:

- ds/dt représente la vitesse le long de la courbe

- dx/dt représente la composante x de la vitesse

- dx et dy sont données par le théorème de Pythagore exactement de la même façon que nous l'avions fait dans le cadre de notre étude du formalisme lagrangien.

$$v = \frac{ds}{dt} = \frac{\sqrt{dx^2 + dy^2}}{dt} = \sqrt{1 + \left(\frac{dy}{dx}\right)^2} \cdot \frac{dx}{dt} = \sqrt{1 + (y'(x))^2} \cdot \frac{dx}{dt} \quad (30.473)$$

En insérant l'équation obtenue d'après les principes de la dynamique:

$$\sqrt{2gy(x)} = \sqrt{1 + (y'(x))^2} \cdot \frac{dx}{dt} \quad (30.474)$$

Une simple intégration nous donne alors l'expression de t à minimiser:

$$t = \int \sqrt{\frac{1 + (y'(x))^2}{2gy(x)}} \cdot dx \quad (30.475)$$

Nous nous retrouvons avec une fonction similaire à celle que nous avons lors de notre étude d'un cas pratique du formalisme lagrangien.

Il s'agit maintenant de trouver le minimum atteint par t parmi toutes les fonctions $y(x)$ satisfaisant:

$$y(x=0) = 0, \quad y(x=a) = b \quad (30.476)$$

Le problème fondamental dit du "**calcul des variations**" consiste à chercher, parmi les fonctions $y = f(x)$ continûment dérivables sur un intervalle donné $[a,b]$ et pour lesquelles les fonctions $f(a)$ et

$f(b)$ sont des valeurs données, celles qui rendent maximum ou minimum l'intégrale précédente.

Pour appliquer cette méthode, nous partons de l'équation d'Euler-Lagrange (cf. chapitre de Mécanique Analytique):

$$\frac{d}{dx} \left(\frac{\partial Z(x, y, y')}{\partial y'} \right) - \frac{\partial Z(x, y, y')}{\partial y} = 0 \quad (30.477)$$

qui donne les extremums de l'intégrale.

Identiquement à un exemple que nous avons vu le cadre de notre étude du formalisme lagrangien (cf. chapitre de Mécanique Analytique) nous posons dans notre cas:

$$Z(x, y, y') = \sqrt{\frac{1 + (y'(x))^2}{2gy(x)}} \quad (30.478)$$

Donc:

$$\begin{aligned} \frac{\partial Z}{\partial y} &= -\frac{1}{4} \frac{\sqrt{2}(1+y'^2)}{\sqrt{\frac{1+y'^2}{gy}} gy^2} = -\frac{1}{4} \frac{\sqrt{2}(1+y'^2)}{y\sqrt{gy}} \\ \frac{\partial Z}{\partial y'} &= \frac{\frac{2y'}{2gy}}{2\sqrt{\frac{1+y'^2}{2gy}}} = \frac{y'}{2gy\sqrt{1+y'^2}} = \frac{y'}{\sqrt{1+y'^2}\sqrt{2gy}} \end{aligned} \quad (30.479)$$

Nous souhaitons injecter des deux dernières relations dans l'équation d'Euler-Lagrange:

$$\frac{d}{dx} \left(\frac{\partial Z(x, y, y')}{\partial y'} \right) - \frac{\partial Z(x, y, y')}{\partial y} = 0 \quad (30.480)$$

mais nous anticipons relativement facilement que la dérivée par rapport à x va nous donner un monstre indigeste à simplifier (j'ai essayé... mais j'ai l'excuse de ne pas être doué pour les maths).

Nous allons alors utiliser l'identité de Beltrami démontrée dans le chapitre de Mécanique Analytique qui s'écrit avec la notation choisie ci-dessus:

$$Z - y' \frac{\partial Z}{\partial y'} = c^{te} \quad (30.481)$$

ce que nous avons le droit de faire puisque la condition de Beltrami est ici satisfaite:

$$\frac{\partial Z}{\partial t} = 0 \quad (30.482)$$

Ce qui nous donne:

$$\sqrt{\frac{1+y'^2}{2gy}} - y' \frac{y'}{\sqrt{1+y'^2} \sqrt{2gy}} = c^{te} \quad (30.483)$$

Soit après réarrangement:

$$\frac{\sqrt{1+y'^2} \sqrt{1+y'^2}}{\sqrt{2gy} \sqrt{1+y'^2}} - y' \frac{y'}{\sqrt{1+y'^2} \sqrt{2gy}} = \frac{\sqrt{1+y'^2} \sqrt{1+y'^2} - y'^2}{\sqrt{1+y'^2} \sqrt{2gy}} = \frac{1}{\sqrt{1+y'^2} \sqrt{2gy}} = c^{te} \quad (30.484)$$

donc au final:

$$\frac{1}{\sqrt{1+y'^2} \sqrt{2gy}} = c^{te} \quad (30.485)$$

ou autrement écrit:

$$y(1+y'^2) = \frac{1}{2gc^{te2}} = K \quad (30.486)$$

où la constante n'est autre que $-D$, l'altitude minimale atteinte par le point mobile. Plus explicitement:

$$y(1+y'^2) = \frac{1}{2gc^{te2}} = K \quad (30.487)$$

Il faut donc résoudre cette équation différentielle pour trouver la fonction qui donne le chemin le plus rapide. Mettons-la sous la forme:

$$y \left(1 + \left(\frac{dy}{dx} \right)^2 \right) = K \quad (30.488)$$

et rappelons que nous avons:

$$\frac{dy}{dx} = \tan \theta \quad (30.489)$$

Nous avons alors (cf. chapitre de Trigonométrie):

$$y(1 + \tan^2 \theta) = \frac{y}{\cos^2 \theta} = K \quad (30.490)$$

D'où:

$$y = K \cos^2 \theta = \frac{K}{2} (1 + \cos(2\theta)) \quad (30.491)$$

Remarquons que nous avons (toujours en utilisant des relations trigonométriques remarquables):

$$\begin{aligned}\frac{dx}{d\theta} &= \frac{dx}{dy} \frac{dy}{d\theta} = -\frac{1}{\tan \theta} K \sin(2\theta) = -\frac{1}{\tan \theta} K 2 \tan \theta \cos^2 \theta = -2K \cos^2 \theta \\ &= -2K \frac{(1 + \cos(2\theta))}{2} = -K(1 + \cos(2\theta))\end{aligned}\quad (30.492)$$

Donc:

$$\frac{dx}{d\theta} = -K(1 + \cos(2\theta)) \quad (30.493)$$

Ce qui est facile à intégrer (cf. chapitre de Calcul Différentiel Et Intégral):

$$\begin{aligned}x &= -K \left(\theta + \frac{1}{2} \sin(2\theta) \right) + c^{te} = -\frac{2}{2} K \left(\theta + \frac{1}{2} \sin(2\theta) \right) + c^{te} \\ &= -\frac{K}{2} (2\theta + \sin(2\theta)) + c^{te}\end{aligned}\quad (30.494)$$

Nous avons donc à ce point (pour rappel):

$$\begin{aligned}x &= -\frac{K}{2} (2\theta + \sin(2\theta)) + c^{te} \\ y &= \frac{K}{2} (1 + \cos(2\theta))\end{aligned}\quad (30.495)$$

Posons pour simplifier l'écriture $\alpha = 2\theta$, nous avons alors:

$$\begin{aligned}x &= -\frac{K}{2} (\alpha + \sin \alpha) + c^{te} \\ y &= \frac{K}{2} (1 + \cos \alpha)\end{aligned}\quad (30.496)$$

Rappelons maintenant que nous avons la condition initiale:

$$y(x=0) = 0 \quad (30.497)$$

et comme K est non nul, cela impose que:

$$\cos(\alpha_{x=0}) = -1 \quad (30.498)$$

et donc que:

$$\alpha_{x=0} = -\pi \quad (30.499)$$

Dès lors, nous devons avoir:

$$x = -\frac{K}{2}(-\pi + \sin(-\pi)) + c^{te} = 0 \quad (30.500)$$

et nous en déduisons immédiatement que:

$$c^{te} = -\frac{K\pi}{2} \quad (30.501)$$

Dès lors:

$$\begin{aligned} x &= -\frac{K}{2}(\alpha + \sin \alpha) - \frac{K\pi}{2} \\ y &= \frac{K}{2}(1 + \cos \alpha) \end{aligned} \quad (30.502)$$

Pour clore, faisons le changement de variable traditionnel:

$$\begin{aligned} \alpha &= \varphi - \pi \\ a &= \frac{K}{2} \end{aligned} \quad (30.503)$$

Il vient alors:

$$\begin{aligned} x &= -a((\varphi - \pi) + \sin(\varphi - \pi)) - a\pi = -a(\varphi + \sin(\varphi - \pi)) = -a(\varphi - \sin(\varphi)) \\ y &= a(1 + \cos \alpha) = a(1 + \cos(\varphi - \pi)) = a(1 - \cos(\varphi)) \end{aligned} \quad (30.504)$$

Donc au final:

$$\begin{aligned} x &= -a(\varphi - \sin(\varphi)) \\ y &= a(1 - \cos(\varphi)) \end{aligned} \quad (30.505)$$

Comme le signe du coefficient a de x n'a que pour effet une translation sur l'axe des X , il est d'usage de représentation cette paire d'équations sous la forme:

$$\begin{aligned} x &= a(\theta - \sin \theta) \\ y &= a(1 - \cos \theta) \end{aligned} \quad (30.506)$$

En posant $a = 1$ nous avons dans Maple 4.00b:

```
>plot([theta-sin(theta),1-cos(theta),theta=0..6*Pi]);
```

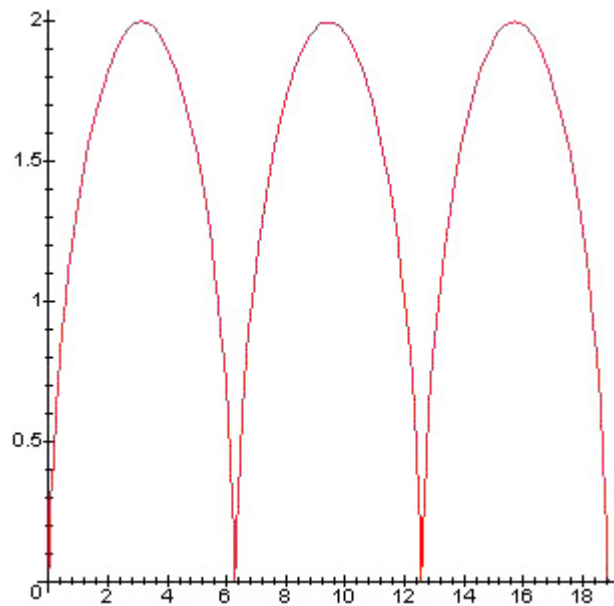


Figure: 30.34 - Tracé de la cycloïde avec Maple 4.00b

Donc à comparer par rapport aux réponses intuitives de la courbe la plus rapide que donnent souvent les humaines par intuition (en rouge la courbe brachistochrone):

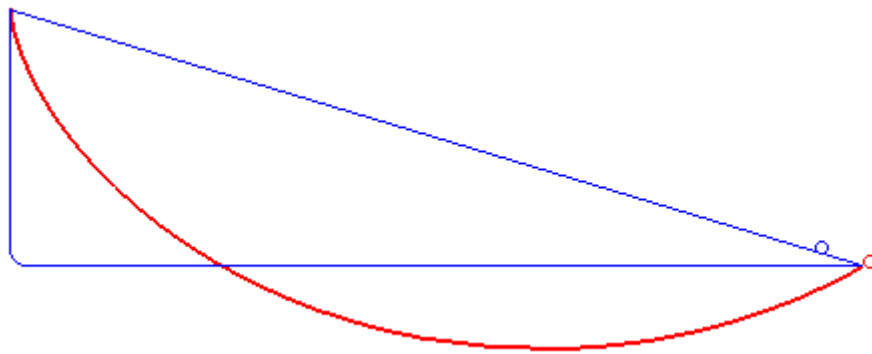


Figure: 30.35 - Comparaisons de la brachistochrone avec les autres chemins intuitifs (source: Wikipédia)

Soit les dérivées:

$$\begin{aligned}\frac{dx}{d\theta} &= a(1 - \cos \theta) \\ \frac{dy}{d\theta} &= a \sin \theta\end{aligned}\quad (30.507)$$

Ainsi:

$$(d_{\theta}x)^2 + (d_{\theta}y)^2 = a^2 \left[(1 - 2\cos \theta + \cos^2 \theta) + \sin^2 \theta \right] = 2a^2(1 - \cos \theta) \quad (30.508)$$

La conservation de l'énergie:

$$\frac{1}{2}mv^2 = mgy \quad (30.509)$$

s'écrit donc:

$$v = \frac{ds}{dt} = \sqrt{2gy} \quad (30.510)$$

d'où:

$$dt = \frac{ds}{\sqrt{2gy}} = \frac{\sqrt{(d_\theta x)^2 + (d_\theta y)^2}}{\sqrt{2gy}} = \frac{a\sqrt{2(1-\cos\theta)}d\theta}{\sqrt{2ga(1-\cos\theta)}} = \sqrt{\frac{a}{g}}d\theta \quad (30.511)$$

Donc le temps requis pour aller du haut au bas de la cycloïde que décrit le pendule de Huygens est:

$$T = \int_0^\pi \sqrt{\frac{a}{g}} d\theta = \pi \sqrt{\frac{a}{g}} \quad (30.512)$$

Cette durée ne dépend donc que de paramètres fixes.

L'énoncé en 1696 du problème brachistochrone peut être considéré comme l'authentique acte de naissance du calcul des variations, car c'est ce problème qui suscita la recherche de méthodes générales progressivement élaborées au cours d'une véritable compétition.

Remarque: Une ligne brachistochrone d'une surface est une courbe sur laquelle doit glisser sans frottement un point matériel pesant placé dans un champ de pesanteur uniforme de sorte que le temps de parcours soit minimal parmi toutes les courbes joignant deux points fixés. Autrement dit, ce sont les lignes les plus courtes en temps, alors que les géodésiques (cf. chapitre de Relativité Restreinte) sont les lignes les plus courtes en distance.

13.9. PENDULE DOUBLE

Le problème du pendule double est un exemple classique d'application du formalisme lagrangien et de la théorie du chaos et donc in extenso un joli exemple scolaire de physique non linéaire.

Voici comme ce dernier est souvent représenté (les coordonnées x et y sont positives sur les deux axes représentés ci-dessous!):

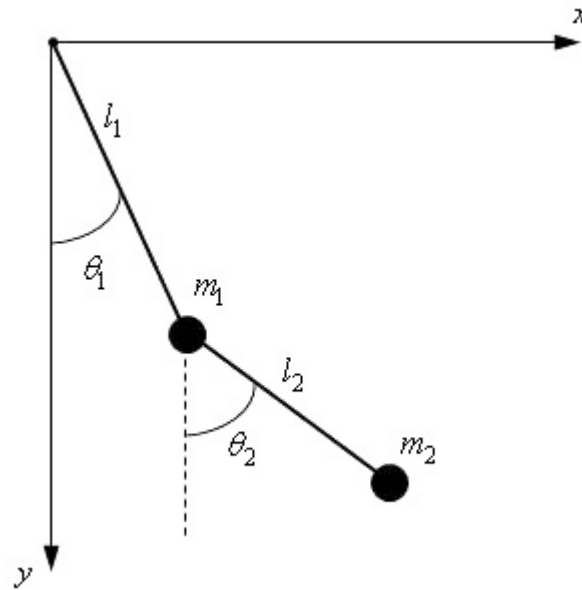


Figure: 30.36 - Illustration du pendule double

La position de la masse m_1 sera donnée par les coordonnées (x_1, y_1) et celle de la masse m_2 - solidaire de m_1 et donc non indépendante - sera donnée par les coordonnées (x_2, y_2) .

Le mieux est de réduire l'étude du système en passant par les coordonnées généralisées (θ_1, θ_2) via les transformations:

$$x_1 = l_1 \sin \theta_1 \quad y_1 = l_1 \cos \theta_1 \quad (30.513)$$

et:

$$x_2 = x_1 + l_2 \sin \theta_2 \quad y_2 = y_1 + l_2 \cos \theta_2 \quad (30.514)$$

Donc en coordonnées cartésiennes et en utilisant la notation traditionnelle du formalisme lagrangien, l'énergie cinétique est alors:

$$T = \frac{1}{2} m_1 v_1^2 + \frac{1}{2} m_2 v_2^2 \quad (30.515)$$

et comme:

$$\begin{aligned} v_1 &= \left\| \frac{d\vec{r}_1}{dt} \right\| = \left\| \frac{d}{dt} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} \right\| = \left\| \frac{d}{dt} \begin{pmatrix} l_1 \sin \theta_1 \\ l_1 \cos \theta_1 \end{pmatrix} \right\| = \left\| l_1 \begin{pmatrix} \dot{\theta}_1 \cos \theta_1 \\ -\dot{\theta}_1 \sin \theta_1 \end{pmatrix} \right\| \\ &= \sqrt{(l_1 \dot{\theta}_1 \cos \theta_1)^2 + (-l_1 \dot{\theta}_1 \sin \theta_1)^2} = \sqrt{l_1^2 \dot{\theta}_1^2 \cos^2 \theta_1 + l_1^2 \dot{\theta}_1^2 \sin^2 \theta_1} = \dot{\theta}_1 \cdot l_1 \end{aligned} \quad (30.516)$$

et:

$$\begin{aligned}
v_2 &= \left\| \frac{d\vec{r}_2}{dt} \right\| = \left\| \frac{d}{dt} \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \right\| = \left\| \frac{d}{dt} \begin{pmatrix} x_1 + l_2 \sin \theta_2 \\ y_1 + l_2 \cos \theta_2 \end{pmatrix} \right\| = \left\| \frac{d}{dt} \begin{pmatrix} l_1 \sin \theta_1 + l_2 \sin \theta_2 \\ l_1 \cos \theta_1 + l_2 \cos \theta_2 \end{pmatrix} \right\| \\
&= \left\| \begin{pmatrix} l_1 \dot{\theta}_1 \cos \theta_1 + l_2 \dot{\theta}_2 \cos \theta_2 \\ -l_1 \dot{\theta}_1 \sin \theta_1 - l_2 \dot{\theta}_2 \sin \theta_2 \end{pmatrix} \right\| = \sqrt{(l_1 \dot{\theta}_1 \cos \theta_1 + l_2 \dot{\theta}_2 \cos \theta_2)^2 + (-l_1 \dot{\theta}_1 \sin \theta_1 - l_2 \dot{\theta}_2 \sin \theta_2)^2} \\
&= \sqrt{l_1^2 \dot{\theta}_1^2 \cos^2 \theta_1 + 2l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 \cos \theta_1 \cos \theta_2 + l_2^2 \dot{\theta}_2^2 \cos^2 \theta_2} \\
&\quad + \sqrt{l_1^2 \dot{\theta}_1^2 \sin^2 \theta_1 + 2l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 \sin \theta_1 \sin \theta_2 + l_2^2 \dot{\theta}_2^2 \sin^2 \theta_2} \\
&= \sqrt{l_1^2 \dot{\theta}_1^2 + l_2^2 \dot{\theta}_2^2 + 2l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 (\cos \theta_1 \cos \theta_2 + \sin \theta_1 \sin \theta_2)} \\
&= \sqrt{l_1^2 \dot{\theta}_1^2 + l_2^2 \dot{\theta}_2^2 + 2l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 \cos(\theta_1 - \theta_2)}
\end{aligned} \tag{30.517}$$

Nous avons alors:

$$\begin{aligned}
T &= \frac{1}{2} m_1 v_1^2 + \frac{1}{2} m_2 v_2^2 = \frac{1}{2} m_1 (\dot{\theta}_1 \cdot l_1)^2 + \frac{1}{2} m_2 \left(\sqrt{l_1^2 \dot{\theta}_1^2 + l_2^2 \dot{\theta}_2^2 + 2l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 \cos(\theta_1 - \theta_2)} \right)^2 \\
&= \frac{1}{2} m_1 \dot{\theta}_1^2 l_1^2 + \frac{1}{2} m_2 (l_1^2 \dot{\theta}_1^2 + l_2^2 \dot{\theta}_2^2 + 2l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 \cos(\theta_1 - \theta_2)) \\
&= \frac{1}{2} (m_1 + m_2) \dot{\theta}_1^2 l_1^2 + \frac{1}{2} m_2 (l_2^2 \dot{\theta}_2^2 + 2l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 \cos(\theta_1 - \theta_2))
\end{aligned} \tag{30.518}$$

Pour l'énergie potentielle nous avons (le signe est négatif car les masses sont en-dessous du point 0):

$$V = -m_1 g y_1 - m_2 g y_2 = -m_1 g l_1 \cos \theta_1 - m_2 g (l_1 \cos \theta_1 + l_2 \cos \theta_2) \tag{30.519}$$

Le lagrangien (cf. chapitre de Mécanique Analytique) vaut donc:

$$\begin{aligned}
L = T - V &= \frac{1}{2} (m_1 + m_2) \dot{\theta}_1^2 l_1^2 + \frac{1}{2} m_2 l_2^2 \dot{\theta}_2^2 + m_2 l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 \cos(\theta_1 - \theta_2) \\
&\quad + (m_1 + m_2) g l_1 \cos \theta_1 + m_2 g l_2 \cos \theta_2
\end{aligned} \tag{30.520}$$

Nous utilisons maintenant l'équation d'Euler-Lagrange avec les coordonnées généralisées choisies:

$$\begin{aligned}
\frac{\partial L}{\partial \theta_1} - \frac{d}{dt} \frac{\partial L}{\partial \dot{\theta}_1} &= 0 \\
\frac{\partial L}{\partial \theta_2} - \frac{d}{dt} \frac{\partial L}{\partial \dot{\theta}_2} &= 0
\end{aligned} \tag{30.521}$$

Nous avons alors pour θ_1 :

$$\begin{aligned}
\frac{\partial L}{\partial \dot{\theta}_1} &= -m_2 l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 \sin(\theta_1 - \theta_2) - (m_1 + m_2) g l_1 \sin \theta_1 \\
\frac{\partial L}{\partial \dot{\theta}_1} &= (m_1 + m_2) \dot{\theta}_1 l_1^2 + m_2 l_1 l_2 \dot{\theta}_2 \cos(\theta_1 - \theta_2) \\
\frac{d}{dt} \frac{\partial L}{\partial \dot{\theta}_1} &= (m_1 + m_2) \ddot{\theta}_1 l_1^2 + m_2 l_1 l_2 \ddot{\theta}_2 \cos(\theta_1 - \theta_2) + m_2 l_1 l_2 \dot{\theta}_2 \frac{d}{dt} \cos(\theta_1 - \theta_2) \\
&= (m_1 + m_2) \ddot{\theta}_1 l_1^2 + m_2 l_1 l_2 \ddot{\theta}_2 \cos(\theta_1 - \theta_2) + m_2 l_1 l_2 \dot{\theta}_2 (-\dot{\theta}_1 \sin(\theta_1 - \theta_2) + \dot{\theta}_2 \sin(\theta_1 - \theta_2)) \\
&= (m_1 + m_2) \ddot{\theta}_1 l_1^2 + m_2 l_1 l_2 \ddot{\theta}_2 \cos(\theta_1 - \theta_2) + m_2 l_1 l_2 \dot{\theta}_2 (\dot{\theta}_2 - \dot{\theta}_1) \sin(\theta_1 - \theta_2)
\end{aligned} \tag{30.522}$$

Donc nous avons au final pour le lagrangien avec la coordonnée généralisée θ_1 :

$$\begin{aligned}
&-m_2 l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 \sin(\theta_1 - \theta_2) - (m_1 + m_2) g l_1 \sin \theta_1 \\
&-(m_1 + m_2) \ddot{\theta}_1 l_1^2 - m_2 l_1 l_2 \ddot{\theta}_2 \cos(\theta_1 - \theta_2) - m_2 l_1 l_2 \dot{\theta}_2 (\dot{\theta}_2 - \dot{\theta}_1) \sin(\theta_1 - \theta_2) \\
&= -(m_1 + m_2) g l_1 \sin \theta_1 - (m_1 + m_2) \ddot{\theta}_1 l_1^2 - m_2 l_1 l_2 \ddot{\theta}_2 \cos(\theta_1 - \theta_2) - m_2 l_1 l_2 \dot{\theta}_2^2 \sin(\theta_1 - \theta_2) \\
&= 0
\end{aligned} \tag{30.523}$$

Nous avons de même pour θ_2 :

$$\begin{aligned}
\frac{\partial L}{\partial \dot{\theta}_2} &= -m_2 l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 \sin(\theta_1 - \theta_2) - m_2 g l_2 \sin \theta_2 \\
\frac{\partial L}{\partial \dot{\theta}_2} &= m_2 l_2^2 \dot{\theta}_2 + m_2 l_1 l_2 \dot{\theta}_1 \cos(\theta_1 - \theta_2) \\
\frac{d}{dt} \frac{\partial L}{\partial \dot{\theta}_2} &= m_2 l_2^2 \ddot{\theta}_2 + m_2 l_1 l_2 \ddot{\theta}_1 \cos(\theta_1 - \theta_2) + m_2 l_1 l_2 \dot{\theta}_1 (\dot{\theta}_2 - \dot{\theta}_1) \sin(\theta_1 - \theta_2)
\end{aligned} \tag{30.524}$$

Donc nous avons au final pour le lagrangien avec la coordonnée généralisée θ_2 :

$$\begin{aligned}
&m_2 l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 \sin(\theta_1 - \theta_2) - m_2 g l_2 \sin \theta_2 \\
&-m_2 l_2^2 \ddot{\theta}_2 - m_2 l_1 l_2 \ddot{\theta}_1 \cos(\theta_1 - \theta_2) - m_2 l_1 l_2 \dot{\theta}_1 (\dot{\theta}_2 - \dot{\theta}_1) \sin(\theta_1 - \theta_2) \\
&= -m_2 g l_2 \sin \theta_2 - m_2 l_2^2 \ddot{\theta}_2 - m_2 l_1 l_2 \ddot{\theta}_1 \cos(\theta_1 - \theta_2) + m_2 l_1 l_2 \dot{\theta}_1^2 \sin(\theta_1 - \theta_2) = 0
\end{aligned} \tag{30.525}$$

Ce qui nous donne au final le système d'équations différentielles suivant:

$$\begin{aligned}
(m_1 + m_2) g l_1 \sin \theta_1 + (m_1 + m_2) \ddot{\theta}_1 l_1^2 + m_2 l_1 l_2 \ddot{\theta}_2 \cos(\theta_1 - \theta_2) + m_2 l_1 l_2 \dot{\theta}_2^2 \sin(\theta_1 - \theta_2) &= 0 \\
m_2 g l_2 \sin \theta_2 + m_2 l_2^2 \ddot{\theta}_2 + m_2 l_1 l_2 \ddot{\theta}_1 \cos(\theta_1 - \theta_2) + m_2 l_1 l_2 \dot{\theta}_1^2 \sin(\theta_1 - \theta_2) &= 0
\end{aligned} \tag{30.526}$$

Avec Maple 4.00b cela donne pour la représentation des coordonnées de m_2 dans le temps:

>with(plots):

> with(plottools):

>

```
Eq1:=(m1+m2)*l1^2*diff(diff(theta1(t),t),t)+m2*l1*l2*cos(theta1(t)-theta2(t))*diff(diff(theta2(t),t),t)
+m2*l1*l2*sin(theta1(t)-theta2(t))*diff(theta2(t),t)^2+(m1+m2)*g*l1*sin(theta1(t))=0;
```

```

> Eq2:=m2*l1*l2*cos(theta1(t)-theta2(t))*diff(diff(theta1(t),t),t)+m2*l2^2*diff(diff(theta2(t),t),t)
-m2*l1*l2*sin(theta1(t)-theta2(t))*diff(theta1(t),t)^2+m2*g*l2*sin(theta2(t))=0;
> m1:=2;m2:=3;l1:=6;l2:=4;g:=9.81;
> ff:=dsolve({Eq1,Eq2,theta1(0)=0.5,D(theta1)(0)=4,theta2(0)=1,D(theta2)(0)=
-2},{theta1(t),theta2(t)},type=numeric,output=listprocedure);
> Theta1:=subs(ff,theta1(t));Theta2:=subs(ff,theta2(t));
> X1:=t->l1*sin(Theta1(t));
> Y1:=t->l1*cos(Theta1(t));
> X2:=t->l1*sin(Theta1(t))+l2*sin(Theta2(t));
> Y2:=t->l1*cos(Theta1(t))+l2*cos(Theta2(t));
> plot([Y2,X2,0..100],numpoints=100);

```

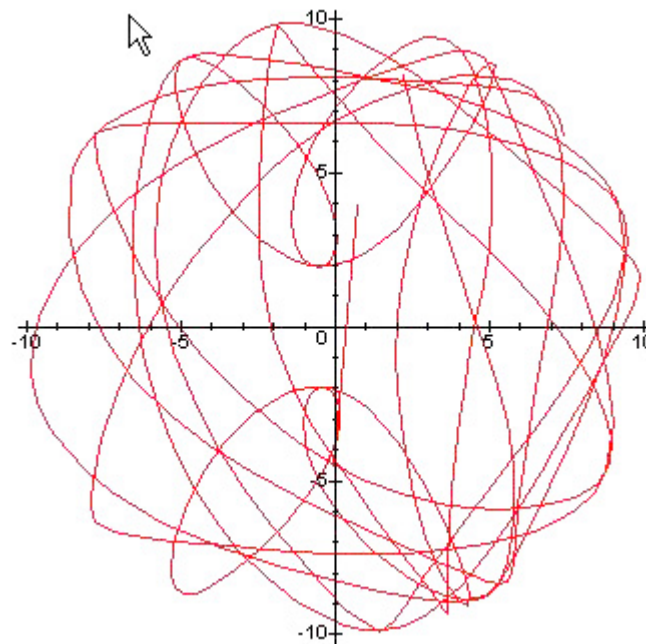


Figure: 30.37 - Tracé avec Maple 4.00b des coordonnées d'une des masses

14. TRIBOLOGIE

Définition: La "**tribologie**" est la science des frottements (notion très intuitive à tout un chacun car nous pouvons ressentir ses effets dans la vie quotidienne) qui interviennent lorsque deux surfaces en contact sont mises en mouvement l'une par rapport à l'autre, produisant une force qui s'oppose au mouvement.

La plupart de ces phénomènes relatifs aux frottements peuvent se comprendre en première approximation sur la base des lois phénoménologiques du frottement énoncées dès le 18ème siècle par Amontons et Coulomb (mais déjà mises en évidence par Léonard de Vinci 200 ans auparavant), à partir de la notion de coefficient de frottement.

Ceux-ci observèrent déjà deux types de frottements a priori distincts:

1. Le "**frottement statique**" est celui qui oppose une résistance lorsqu'un objet posé sur un plan est à la limite du glissement alors qu'on lui impose une force de traction $\vec{F}_T$ (tangentielle au plan). Cette opposition à la force de traction est par ailleurs expérimentalement proportionnelle au poids $\vec{P}$ de

l'objet.

Mais intervient une valeur limite de la force tangentielle de traction à partir de laquelle l'objet commence à glisser. C'est ce que nous notons:

$$\|\vec{F}_{T,\text{lim}}\| = \mu_s \cdot \|\vec{P}\| \quad (30.527)$$

où $\vec{F}_{T,\text{lim}}$ est donc la force limite de traction permettant de faire bouger l'objet initialement statique, μ_s est le "**coefficient de frottement statique**" sans dimensions et exprime la proportionnalité de la force limite de frottement avec le poids $\vec{P}$ de l'objet ou de sa forme normale au plan d'application. Raison pour laquelle cette dernière relation est aussi souvent notée:

$$\|\vec{F}_{T,\text{lim}}\| = \mu_s \|\vec{N}\| \quad (30.528)$$

Remarque: Dans la pratique, il est infiniment facile de déterminer ce coefficient avec un simple dynamomètre pour connaître la force limite et une balance pour connaître le poids de l'objet étudié.

Nous observons expérimentalement que contrairement à l'intuition, la force limite de traction est en première approximation indépendante de la surface de contact entre l'objet et le sol (dans les limites des cas physiques courants évidemment car plus la surface est petite, plus la pression est grande et alors la surface de contact peut devenir plastique aux hautes pressions).

Autrement dit, si un kilo de sucre est posé sur une table. Pour déplacer cet objet, de poids $\vec{P}$ (la masse multipliée par la constante de gravité), il faut exercer une force $\vec{F}_T$ parallèlement à la surface de la table. Mais l'expérience montre que cet objet ne se déplacera pas tant que la force $\vec{F}_T$ est inférieure à une force minimale $\vec{F}_{T,\text{lim}}$. Et Amontons et Coulomb ont montré que cette force minimale est directement proportionnelle via un coefficient de frottement statique au poids.

Nous pouvons détailler l'approche de la relation précédente en s'imaginant deux surfaces présentant des rugosités en dents de scie d'un angle α et imbriquées:

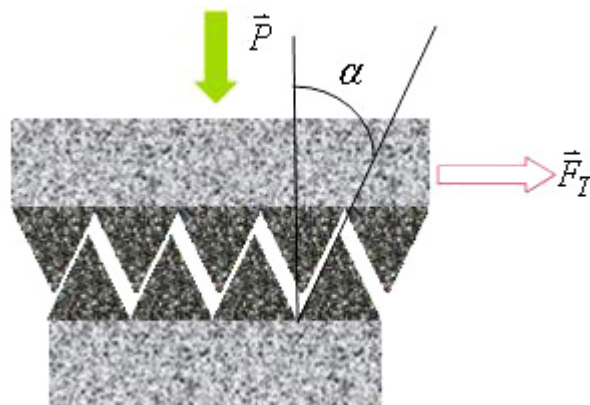


Figure: 30.38 - Zoom sur la surface de contact

Si nous appliquons une force normale correspondante au poids $\vec{P}$ et une force horizontale $\vec{F}_T$ nous

avons à cause des dents de scie dans le cas limite la situation suivante:

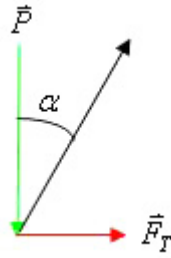


Figure: 30.39 - Simplification du problème avec les vecteurs

nous voyons alors bien que la pièce mobile ne commencera à bouger que quand il y aura début de glissement soit lorsque:

$$\|\vec{F}_T\| = \tan \alpha \cdot \|\vec{P}\| \quad (30.529)$$

Pour simpliste qu'elle soit, cette approche permet de lier le frottement (statique) aux caractéristiques de la rugosité. De plus les valeurs expérimentales typiques des coefficients de frottement statique, de l'ordre de 0.3, correspondent à des pentes de la rugosité de surface de l'ordre de 15-20 degrés, ce qui est tout à fait compatible avec les caractéristiques typiques que l'on peut mesurer pour les rugosités de surfaces!

Cet argument repose cependant sur une hypothèse implicite: l'emboîtement parfait entre les rugosités des deux surfaces. Nous parlons dans ce cas de "**surfaces commensurables**". Ce n'est bien sûr pas le cas en général dans la nature: même à l'échelle atomique, deux surfaces idéales, présentent des légères différences de distance interatomique qui empêchent l'emboîtement. Une légère disparité suffit à rendre très irrégulière la répartition des points de contact entre les deux surfaces contrairement au cas commensurable. Nous parlons alors de "**surfaces incommensurables**".

Autrement dit, nous aboutissons très vite à la conclusion que le frottement entre deux surfaces incommensurables en tout point est non-nul, tandis qu'il s'annule exactement si ces deux surfaces sont en tout point commensurables.

2. Le "**frottement dynamique**" est celui qui oppose une résistance lorsqu'un objet posé sur un plan est déjà en glissement. Cette opposition à la traction est par ailleurs expérimentalement proportionnelle encore une fois au poids de l'objet tel que:

$$\|\vec{F}_T\| = \mu_D \cdot \|\vec{P}\| \quad (30.530)$$

mais avec le "**coefficient de frottement dynamique**" (qui existe en plusieurs sous-familles: coefficient de roulement, de glissement,) qui est en général beaucoup plus petit que le coefficient de frottement statique:

$$\mu_S \ll \mu_D \quad (30.531)$$

Donc le frottement n'est pas le même au départ de notre objet que lors de son glissement. Cela correspond très bien à notre expérience quotidienne du frottement (lors de déplacements de meubles dans nos habitations par exemples).

A nouveau, nous remarquons expérimentalement que contrairement à l'intuition, la force de traction est indépendante de la surface de contact entre l'objet et le sol.

Ainsi, que l'on pose le kilo de sucre bien à plat ou sur la tranche, la force de frottement est la même (si la qualité de surface est la même de tous les côtés du paquet de sucre)!

Un autre fait étonnant concerne la valeur typique de ces coefficients de frottement, qui s'écarte assez peu de $\mu_{D,S} \cong 0.3$, pour des surfaces très différentes les unes des autres. La technologie permet toutefois de concevoir des surfaces avec des coefficients de frottement soit bien plus petits ($\mu_{D,S} \cong 0.001$) soit plus grand ($\mu_{D,S} \gg 1$).

Ces deux lois, sont appelées "**lois de Coulomb**":

$$\|\vec{F}_{T,\text{lim}}\| = \mu_S \cdot \|\vec{N}\| \quad \|\vec{F}_T\| = \mu_D \cdot \|\vec{N}\| \quad (30.532)$$

L'origine simpliste du frottement entre deux solides est donc dû au fait que:

- Tout solide n'est jamais lisse mais possède des aspérités qui rendent la surface de contact rugueuse (les aspérités s'imbriquent partiellement ou non et provoquent plus ou moins de frottement).
- Les impuretés entre les deux surfaces de contact sont souvent plus importantes au niveau des sources de frottement que les imbrications des aspérités de surface.
- Le frottement est faiblement dépendant de la surface car la rugosité à l'échelle atomique est telle que seulement un très faible pourcentage de la surface totale des deux objets sont réellement en contact (surface de contact réelle est donc beaucoup plus petite que la surface de contact apparente) ce qui explique que la force de traction tangentielle soit proportionnelle au poids car cela force la surface de contact réelle à augmenter.

La complexité sous-jacente du frottement est donc extrême. L'origine du frottement fait dans la réalité intervenir une multitude d'ingrédients, couvrant un spectre très large de phénomènes physiques: rugosité des surfaces, élasticité, plasticité, adhésion, lubrification, thermique, usure, chimie des surfaces, humidité, etc.

Nous allons ici faire une analyse scolaire des quelques frottements courants dans les cas d'études simples de la physique cinétique (du mouvement). Il faut bien prendre garde que ces modèles sont simplifiés à l'extrême afin de montrer seulement la démarche intellectuelle.

14.1. FROTTEMENT EXPONENTIEL

Il est d'usage d'appeler par "frottement exponentiel" le frottement généré par des courroies non-crantées (cas très important dans la pratique) sur un support fixe!

Pour étudier ce cas de frottement considérons la figure ci-dessous où β est l'angle d'enroulement de la courroie autour de la poulie où on néglige l'épaisseur et la masse de la courroie par rapport aux autres éléments et où on se place à la limite du glissement et en mouvement uniforme.

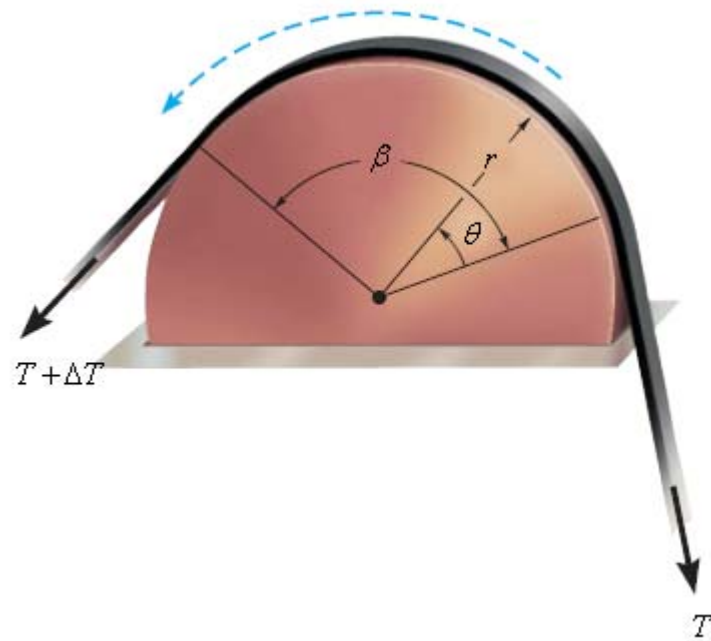


Figure: 30.40 - Schéma macro du contact poulie/courroie

De par l'existence de frottement il y a évidemment un différentiel de tension dT sur un point de contact mais cela étant dû uniquement au frottement puisque nous supposons la rotation uniforme. Voyons cela de plus près isolant une portion et en appliquant le principe fondamental de la statique (avec subtilité...):

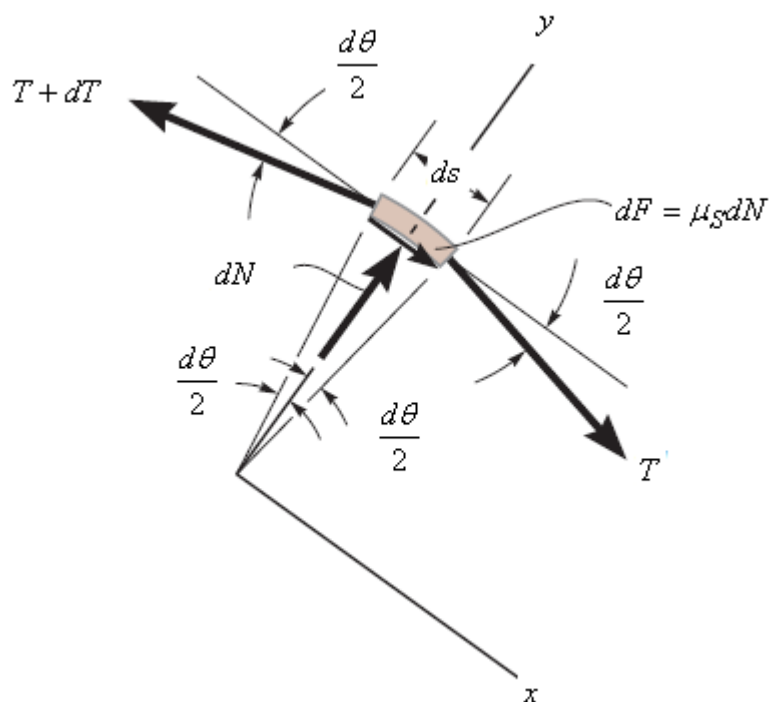


Figure: 30.41 - Zoom sur un élément différentiel de la courroie

Nous avons alors:

$$\begin{aligned}\sum F_x = 0: & T \cos\left(\frac{d\theta}{2}\right) + \mu_S dN - (T + dT) \cos\left(\frac{d\theta}{2}\right) = 0 \\ \sum F_y = 0: & dN - (T + dT) \sin\left(\frac{d\theta}{2}\right) - T \sin\left(\frac{d\theta}{2}\right) = 0\end{aligned}\quad (30.533)$$

Soit après simplification:

$$\begin{aligned}\sum F_x = 0: & \mu_S dN - dT \cos\left(\frac{d\theta}{2}\right) = 0 \\ \sum F_y = 0: & dN - dT \sin\left(\frac{d\theta}{2}\right) - 2T \sin\left(\frac{d\theta}{2}\right) = 0\end{aligned}\quad (30.534)$$

et comme l'angle est très petit, nous avons:

$$\cos\left(\frac{d\theta}{2}\right) \cong \cos(0) = 1 \quad (30.535)$$

et pour le sinus nous utiliserons le développement de MacLaurin (cf. chapitre Suites Et Séries) pour les petits angles:

$$\sin\left(\frac{d\theta}{2}\right) \cong \frac{d\theta}{2} \quad (30.536)$$

Nous avons alors:

$$\begin{aligned}\sum F_x = 0: & \mu_S dN - dT = 0 \\ \sum F_y = 0: & dN - dT \frac{d\theta}{2} - 2T \frac{d\theta}{2} = 0\end{aligned}\quad (30.537)$$

En négligeant la multiplication d'éléments infinitésimaux et en simplifiant il reste:

$$\begin{aligned}\sum F_x = 0: & \mu_S dN = dT \\ \sum F_y = 0: & dN = T d\theta\end{aligned}\quad (30.538)$$

où la première équation nous est déjà connu puisqu'elle correspondant au frottement statistique horizontal habituel déjà vu plus haut. Nous pouvons maintenant combiner ces deux relations pour obtenir:

$$\mu_S d\theta = \frac{dT}{T} \quad (30.539)$$

En intégrant il vient alors:

$$\int_0^{\beta} \mu_S d\theta = \int_{T_1}^{T_2} \frac{dT}{T} \quad (30.540)$$

Nous avons alors:

$$\ln\left(\frac{T_2}{T_1}\right) = \mu_S \beta \quad (30.541)$$

et au final:

$$T_2 = T_1 e^{\mu_S \beta} \quad (30.542)$$

ce qui est très important! Par exemple dans la pratique avec une tension de 500 [N] en tirant, cela ne produit une tension réelle de l'autre côté du support circulaire fixe que de 250 [N] parce que le frottement est exponentiel. Raison pour laquelle il vaut mieux utiliser des poulies que des supports circulaires fixes!

Évidemment à cause de la friction on aura:

$$T_2 > T_1 \quad (30.543)$$

ce qui implique évidemment que:

$$e^{\mu_S \beta} > 1 \quad (30.544)$$

et donc:

$$\mu_S \beta > 0 \quad (30.545)$$

14.2. FROTTEMENT VISQUEUX HORIZONTAL

Nous avons donc vu qu'en première approximation, la force de frottement dans le cas de glissement est proportionnelle au poids d'un corps par un coefficient de frottement dont la valeur dépend de la nature et de l'état des surfaces de contact mais indépendant de l'aire de contact. Cependant nous n'avons pas dit que l'expérience montre que dans des cas réels typiques la force de frottement est aussi indépendante de la vitesse communiquée au corps.

En lubrifiant par un fluide visqueux les surfaces en contact, la force de frottement est réduite et dépend de la vitesse (c'est typiquement le cas des pneus de voiture qui sont visqueux).

Remarque: Pour rappel, le terme "visqueux" ne signifie pas forcément que ça coule et que ça bave. Cela signifie que la loi de comportement dépend de la vitesse de déformation (cf. chapitre de Mécanique Des Milieux Continus).

Considérons alors un mobile en contact avec un sol plan via un fluide ou matériau visqueux. Nous

savons qu'il y aura frottement et supposons que celui-ci soit proportionnel à la vitesse:

$$\boxed{F_f = -kv} \quad (30.546)$$

où k est le "coefficient de frottement visqueux".

Nous avons alors en appliquant la première loi de Newton:

$$F = m \frac{dv}{dt} = -kv \quad (30.547)$$

Dès lors il vient:

$$\frac{1}{v} dv = -\frac{k}{m} dt \quad (30.548)$$

En intégrant il vient:

$$\ln(v) - \ln(v_0) = -\frac{k}{m} t \quad (30.549)$$

Soit:

$$\ln(v) = \ln(v_0) - \frac{k}{m} t \quad (30.550)$$

En prenant l'exponentielle:

$$v = e^{\ln(v_0)} e^{-\frac{k}{m} t} = v_0 e^{-\frac{k}{m} t} \quad (30.551)$$

Ainsi, la vitesse décroît exponentiellement d'une vitesse initiale jusqu'à une valeur nulle asymptotique sous l'hypothèse de proportionnalité du frottement avec la vitesse.

C'est une relation très souvent utilisée dans les animations faites par ordinateur représentant des objets qui semblent s'arrêter de manière naturelle. Il faut simplement bien choisir la valeur de k .

Nous observons une chose intéressante c'est qu'un corps mobile lourd décélère moins vite à cause des forces de frottement qu'un corps mobile léger!

Montrons comment nous calculons la puissance perdue par frottement. Nous savons que:

$$P = \frac{dW}{dt} = \frac{d(F \cdot x)}{dt} = \frac{dF \cdot x}{dt} + \frac{F \cdot dx}{dt} \quad (30.552)$$

si la variation de la force est négligeable par rapport à la variation de vitesse nous avons alors:

$$P \cong \frac{F \cdot dx}{dt} = F \cdot v \quad (30.553)$$

et donc dans le cas du frottement (en valeur absolue):

$$P = |F_f| \cdot v = |-kv| \cdot v = k \cdot v^2 \quad (30.554)$$

Ainsi, la puissance dissipée lors d'un mouvement est proportionnelle à la vitesse en cas de frottement coulombien et proportionnelle au carré de la vitesse en cas de frottement visqueux.

14.3. FROTTEMENT VISQUEUX VERTICAL

Soit un solide indéformable chutant à la verticale dans un champ de gravité. Nous assumons que la force de la résistance de l'air est proportionnelle à la vitesse (comportement visqueux aux faibles vitesses):

$$F_f = -kv \quad (30.555)$$

et utilisant le principe fondamental de la dynamique:

$$mg - kv = m \frac{dv}{dt} \quad (30.556)$$

Soit autrement écrit (plus traditionnel):

$$m \frac{dv}{dt} + kv = mg \quad (30.557)$$

La solution de cette équation différentielle linéaire du 1er ordre est (cf. chapitre de Calcul Différentiel Et Intégral):

$$v = \frac{mg}{k} + C e^{-\frac{k}{m}t} \quad (30.558)$$

que nous pouvons détailler si besoin (sur demande).

En posant qu'à l'instant nul nous avons une vitesse initiale donnée il vient:

$$v(t=0) = v_0 = \frac{mg}{k} + C \Rightarrow C = v_0 - \frac{mg}{k} \quad (30.559)$$

Ainsi:

$$v = \frac{mg}{k} + \left(v_0 - \frac{mg}{k} \right) e^{-\frac{k}{m}t} = \frac{mg}{k} \left(1 - e^{-\frac{k}{m}t} \right) + v_0 e^{-\frac{k}{m}t} \quad (30.560)$$

Ainsi, nous voyons que lorsque le temps tend vers l'infini (suffisamment grand quoi...) alors la vitesse tend vers:

$$v_{t \rightarrow +\infty} = \frac{mg}{k} \quad (30.561)$$

donc k peut être déterminé expérimentalement!

C'est une relation très souvent utilisée dans les animations faites par ordinateur représentant des objets

qui semblent freiner jusqu'à une vitesse constante de manière naturelle.

14.4. FROTTEMENT VISQUEUX DE STOKES VERTICAL

C'est typiquement le cas du parachutiste effectuant une chute libre. Nous avons vu dans le chapitre de Mécanique Des Milieux Continus que la force visqueuse de Stokes était donnée par (comportement visqueux aux vitesses moyennes et élevées):

$$F_f = -\rho v^2 L^2 C(R_e) = -\rho S C v^2 = -k v^2 \quad (30.562)$$

lorsque la vitesse est subsonique (modeste en d'autres termes...).

L'équation différentielle est la même qu'avant dans le cas de la présence du champ de gravitation à la différence que la vitesse est cette fois-ci au carré:

$$mg - k v^2 = m \frac{dv}{dt} \quad (30.563)$$

Soit:

$$\frac{k}{m} v^2 + \frac{dv}{dt} - g = 0 \quad (30.564)$$

mais nous n'allons pas chercher à la résoudre, seulement à déterminer la valeur limite de la vitesse et justement vitesse limite est atteinte lorsque celle-ci.... ne varie plus (ben oui forcément...). Donc à ce moment:

$$\frac{dv}{dt} = 0 \quad (30.565)$$

et l'équation différentielle devient:

$$\frac{k}{m} v^2 - g = 0 \Rightarrow v_{\text{lim}}^2 = \frac{gm}{k} = \frac{gm}{\rho S C} \quad (30.566)$$

Ainsi, il est possible de changer sa vitesse de chute limite en fonction de son facteur de forme, et de sa surface d'exposition apparente et de sa masse (dans le cas d'étude ci-dessus nous négligeons la force d'Archimède qui s'applique sur la parachutiste et qui freine aussi sa chute).

Exemple:

Considérons une sphère de rayon R , de masse volumique ρ lâchée sans vitesse dans un liquide de masse volumique ρ_{liq} , de viscosité μ . La sphère est soumise à son propre poids, à une force de frottement visqueux et à la poussée d'Archimède.

Nous avons donc globalement selon la première loi de Newton:

$$m \frac{dv}{dt} = \underbrace{\frac{4}{3} \pi R^3 \rho g}_{\text{force d'attraction}} - \underbrace{6 \pi \mu R v}_{\text{force visqueuse de Stokes}} - \underbrace{\frac{4}{3} \pi R^3 \rho_{\text{liq}} g}_{\text{force d'Archimède}} \quad (30.567)$$

où les deux derniers termes (force visqueuse de Stokes aux faibles vitesses et force d'Archimède) ont été démontrés dans le chapitre de Mécanique Des Milieux Continus.

En réarrangeant, nous avons:

$$m \frac{dv}{dt} + 6\pi\mu R v = \frac{4}{3} \pi R^3 g (\rho - \rho_{liq}) = \frac{4}{3} \pi R^3 g \rho \left(1 - \frac{\rho_{liq}}{\rho}\right) \quad (30.568)$$

et encore:

$$\begin{aligned} \frac{m}{\frac{4}{3} \pi R^3 \rho} \frac{dv}{dt} + \frac{6\pi\mu R}{\frac{4}{3} \pi R^3 \rho} v &= \frac{m}{\frac{4}{3} \pi R^3 \rho} \frac{1}{\rho} \frac{dv}{dt} + \frac{18\mu}{4R^2 \rho} v = \rho \frac{1}{\rho} \frac{dv}{dt} + \frac{9\mu}{2R^2 \rho} v \\ &= \frac{dv}{dt} + \frac{9\mu}{2R^2 \rho} v = \left(1 - \frac{\rho_{liq}}{\rho}\right) g \end{aligned} \quad (30.569)$$

Il nous reste donc:

$$\frac{dv}{dt} + \frac{9\mu}{2R^2 \rho} v = \left(1 - \frac{\rho_{liq}}{\rho}\right) g \quad (30.570)$$

Nous posons:

$$\tau = \frac{2R^2 \rho}{9\mu} \quad (30.571)$$

qui sera assimilée à une constante de temps. Nous avons alors:

$$\tau \frac{dv}{dt} + v = \tau \left(1 - \frac{\rho_{liq}}{\rho}\right) g \quad (30.572)$$

Or lorsque la vitesse de chute deviendra constante, nous aurons:

$$\frac{dv}{dt} = 0 \quad (30.573)$$

ce qui donne:

$$v_{lim} = \tau \left(1 - \frac{\rho_{liq}}{\rho}\right) g \quad (30.574)$$

Réolvons ceci dit l'équation différentielle en commençant par celle sans second membre (cf. chapitre de Calcul Différentiel et Intégral):

$$\frac{dv}{dt} + \frac{1}{\tau} v = 0 \quad (30.575)$$

Nous avons alors vu dans le chapitre de Calcul Différentiel Et Intégral que la solution homogène était

alors donnée par:

$$v_h = C e^{-\frac{1}{\tau}t} \quad (30.576)$$

Nous pouvons ajouter la solution particulière qui est logiquement lorsque t tend vers l'infini:

$$v_p = v_{\lim} = \tau \left(1 - \frac{\rho_{\text{liq}}}{\rho} \right) g \quad (30.577)$$

Nous avons alors:

$$v = v_p + v_h = v_{\lim} + v_h = \tau \left(1 - \frac{\rho_{\text{liq}}}{\rho} \right) g + C e^{-\frac{1}{\tau}t} \quad (30.578)$$

Il nous reste à déterminer C qui s'obtient lors t tend vers 0 car nous avons alors:

$$v(t=0) = 0 \quad (30.579)$$

Donc:

$$v(t=0) = 0 = \tau \left(1 - \frac{\rho_{\text{liq}}}{\rho} \right) g + C e^{-\frac{1}{\tau}0} = \tau \left(1 - \frac{\rho_{\text{liq}}}{\rho} \right) g + C \Rightarrow C = -\tau \left(1 - \frac{\rho_{\text{liq}}}{\rho} \right) g \quad (30.580)$$

Nous avons alors:

$$v = \tau \left(1 - \frac{\rho_{\text{liq}}}{\rho} \right) g - \tau \left(1 - \frac{\rho_{\text{liq}}}{\rho} \right) g e^{-\frac{1}{\tau}t} = \tau \left(1 - \frac{\rho_{\text{liq}}}{\rho} \right) g \left(1 - e^{-\frac{1}{\tau}t} \right) \quad (30.581)$$

14.5. FROTTEMENT VISQUEUX DE STOKES HORIZONTAL

C'est une première approximation où l'on s'intéresse par exemple à la distance d'arrêt sans freinage d'un mobile sans prendre en compte le coefficient de frottement avec le sol mais seulement avec l'air ambiant (vitesse subsonique toujours...).

Nous avons alors selon la première loi de Newton:

$$m \frac{dv}{dt} = -k v^2 \quad (30.582)$$

Supposons que nous souhaitions savoir en quel temps T le mobile qui avait une vitesse initiale v_0 aura décéléré à une vitesse donnée v_f .

Nous avons alors:

$$dt = -\frac{m}{k} \frac{dv}{v^2} \quad (30.583)$$

Donc:

$$T = \int_0^T dt = -\frac{m}{k} \int_{v_0}^{v_f} \frac{dv}{v^2} = -\frac{m}{k} \left(-\frac{1}{v} \right) \Big|_{v_0}^{v_f} = -\frac{m}{k} \left[-\frac{1}{v_f} - \left(-\frac{1}{v_0} \right) \right] = \frac{m}{k} \left(\frac{1}{v_f} - \frac{1}{v_0} \right) \quad (30.584)$$

où nous observons déjà un premier problème avec ce modèle c'est que à l'arrêt, la vitesse finale étant nulle, il faudra un temps infini pour y arriver... mais continuons, nous reviendrons plus loin sur ce constat.

La loi d'évolution de la vitesse se détermine de façon analogue puisque:

$$t = -\frac{m}{k} \left[-\frac{1}{v_f} - \left(-\frac{1}{v_0} \right) \right] = \frac{m}{k} \left(\frac{1}{v_f} - \frac{1}{v_0} \right) \quad (30.585)$$

Alors:

$$t = \frac{m}{k} \frac{1}{v_f} - \frac{m}{k} \frac{1}{v_0} \Rightarrow \frac{m}{k} \frac{1}{v_f} = \frac{m}{k} \frac{1}{v_0} + t \Rightarrow \frac{1}{v_f} = \frac{1}{v_0} + \frac{k}{m} t = \frac{1}{v_0} \left(1 + \frac{v_0 k}{m} t \right) \quad (30.586)$$

Notons la constante de temps:

$$\tau = \frac{m}{v_0 k} \quad (30.587)$$

Alors:

$$\frac{1}{v_f} = \frac{1}{v_0} \left(1 + \frac{t}{\tau} \right) \quad (30.588)$$

soit:

$$v_f(t) = \frac{v_0}{1 + \frac{t}{\tau}} \quad (30.589)$$

La distance parcourue à l'instant t en ne laissant le mobile ralentir que par les forces de frottement est donc:

$$x = \int_0^t v_f(t) dt = \int_0^t \frac{v_0}{1 + \frac{t}{\tau}} dt = v_0 \int_0^t \frac{1}{1 + \frac{t}{\tau}} dt = v_0 \tau \ln \left(1 + \frac{t}{\tau} \right) \quad (30.590)$$

Le résultat est joli mais on se rend bien compte que ce n'est pas vraiment juste car à un temps infini, la voiture aura parcouru une distance infinie ce qui est manifestement irréaliste. Cela provient du modèle qui est trop simpliste donc améliorons-le.

L'idéal, objectivement parlant, serait de prendre en compte le frottement visqueux pneu/sol plus le frottement de l'air. Nous aurions alors:

$$m \frac{dv}{dt} = -kv - k'v^2 \quad (30.591)$$

mais le problème avec cette équation différentielle, c'est qu'elle va nous amener à une singularité si nous continuons les calculs. Elle n'est donc pas exploitable...

Nous essayons alors avec la forme suivante:

$$m \frac{dv}{dt} = -m\alpha - m\beta v^2 \quad (30.592)$$

qui exprime donc qu'il y a une force de frottement proportionnelle au poids du véhicule ce qui est simplement la forme suivante de la deuxième loi de Coulomb:

$$F_{T,lm} = \mu_D \cdot P = \mu_D \cdot m \cdot g = m \cdot (g \cdot \mu_D) = m \cdot \alpha \quad (30.593)$$

et le deuxième terme étant le frottement visqueux de Stokes dont nous avons sorti le terme de masse compris dans la densité se trouvant implicitement dans la constante k de kv^2 .

Nous avons alors en simplifiant les termes de masse:

$$\frac{dv}{dt} = -\alpha - \beta v^2 \quad (30.594)$$

Donc on voit déjà que dans ce modèle nous allons perdre l'effet de la masse qui a normalement pour implication de rallonger le trajet d'arrêt (dans la réalité!). Mais continuons quand même...

Nous avons donc à intégrer:

$$-\int_0^t dt = \int_{v_0}^0 \frac{dv}{\alpha + \beta v^2} = \frac{1}{\beta} \int_{v_0}^0 \frac{\frac{\alpha}{\beta}}{\frac{\alpha}{\beta} + v^2} \quad (30.595)$$

Nous avons démontré dans le chapitre de Calcul Différentiel Et Intégral que:

$$\int \frac{1}{a^2 + x^2} = \frac{1}{a} \arctan\left(\frac{x}{a}\right) \quad (30.596)$$

Donc en posant:

$$a^2 = \frac{\alpha}{\beta} \Rightarrow a = \sqrt{\frac{\alpha}{\beta}} \quad (30.597)$$

Soit:

$$-\int_0^t dt = -t = \frac{1}{\beta} \int_{v_0}^0 \frac{dv}{\frac{\alpha}{\beta} + v^2} = -\frac{1}{\beta} \frac{1}{\sqrt{\frac{\alpha}{\beta}}} \arctan \left(\frac{v_0}{\sqrt{\frac{\alpha}{\beta}}} \right) \quad (30.598)$$

Un peu réarrangé cela donne:

$$t = \frac{1}{\sqrt{\alpha\beta}} \arctan \left(\sqrt{\frac{\beta}{\alpha}} v_0 \right) \quad (30.599)$$

Donc on tombe maintenant déjà sur un temps fini... ce qui est plus rassurant comme résultat.

Cherchons maintenant la distance d'arrêt. Nous avons en utilisant le fait que:

$$v = \frac{dx}{dt} \Rightarrow dt = dx / v \quad (30.600)$$

la possibilité d'écrire:

$$\frac{dv}{dt} = \frac{dv}{dx / v} = \frac{v dv}{dx} = -\alpha - \beta v^2 \quad (30.601)$$

Soit:

$$-dx = \frac{v dv}{\alpha + \beta v^2} \quad (30.602)$$

Ce qui nous amène à:

$$-\int_0^L dx = \int_{v_0}^0 \frac{v dv}{\alpha + \beta v^2} \quad (30.603)$$

Ce qui donne déjà:

$$-L = \int_{v_0}^0 \frac{v dv}{\alpha + \beta v^2} = \frac{1}{2\beta} \int_{v_0}^0 \frac{2\beta v dv}{\alpha + \beta v^2} = \frac{1}{2\beta} \ln(\alpha + \beta v^2) \Big|_{v_0}^0 = \frac{1}{2\beta} \left[\ln(\alpha) - \ln(\alpha + \beta v_0^2) \right] \quad (30.604)$$

Nous avons donc:

$$-L = \frac{1}{2\beta} \left[\ln(\alpha) - \ln(\alpha + \beta v_0^2) \right] \quad (30.605)$$

Soit:

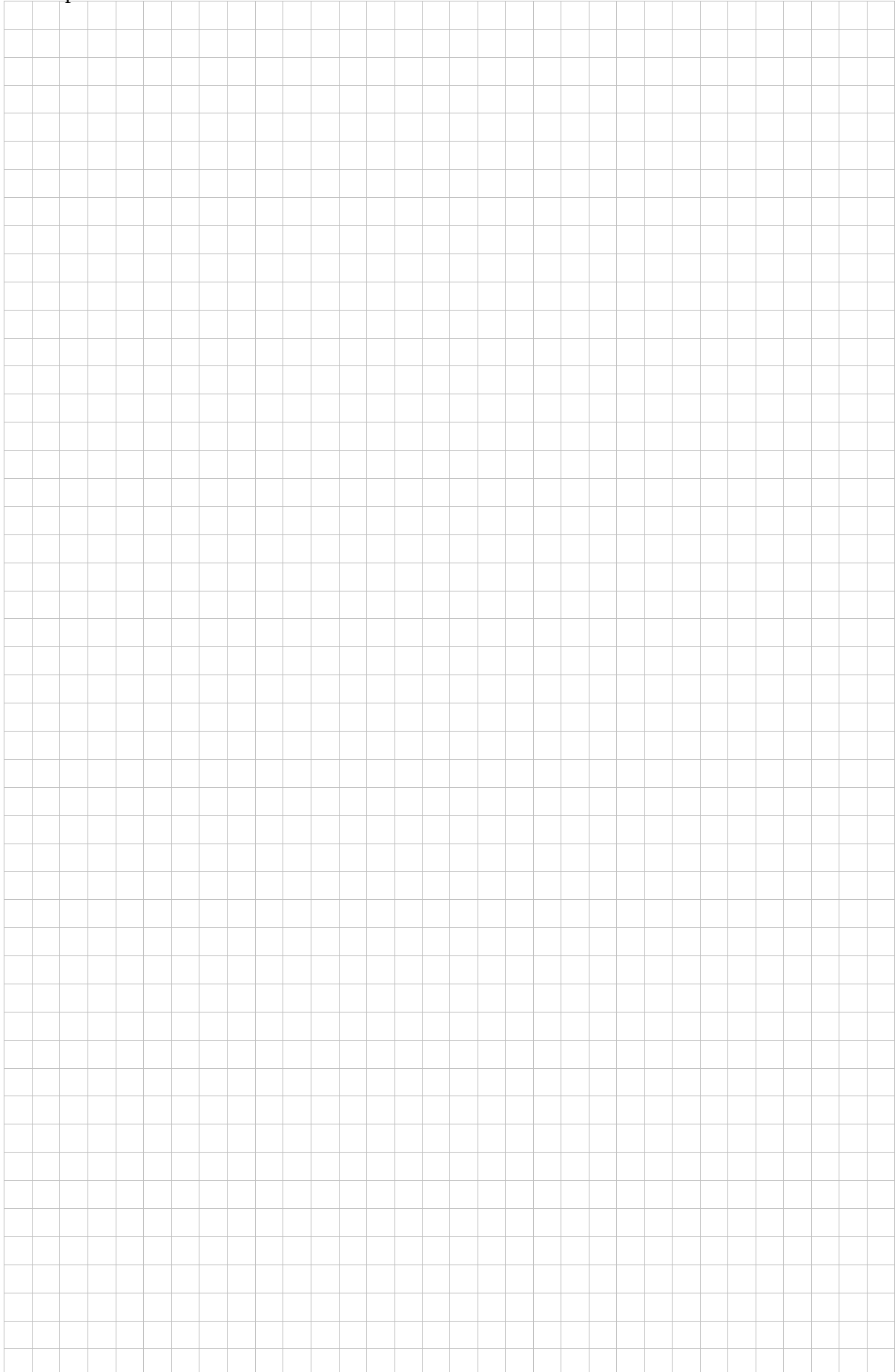
$$L = \frac{1}{2\beta} \left[\ln(\alpha + \beta v_0^2) - \ln(\alpha) \right] = \frac{1}{2\beta} \ln \left(1 + \frac{\beta}{\alpha} v_0^2 \right) \quad (30.606)$$

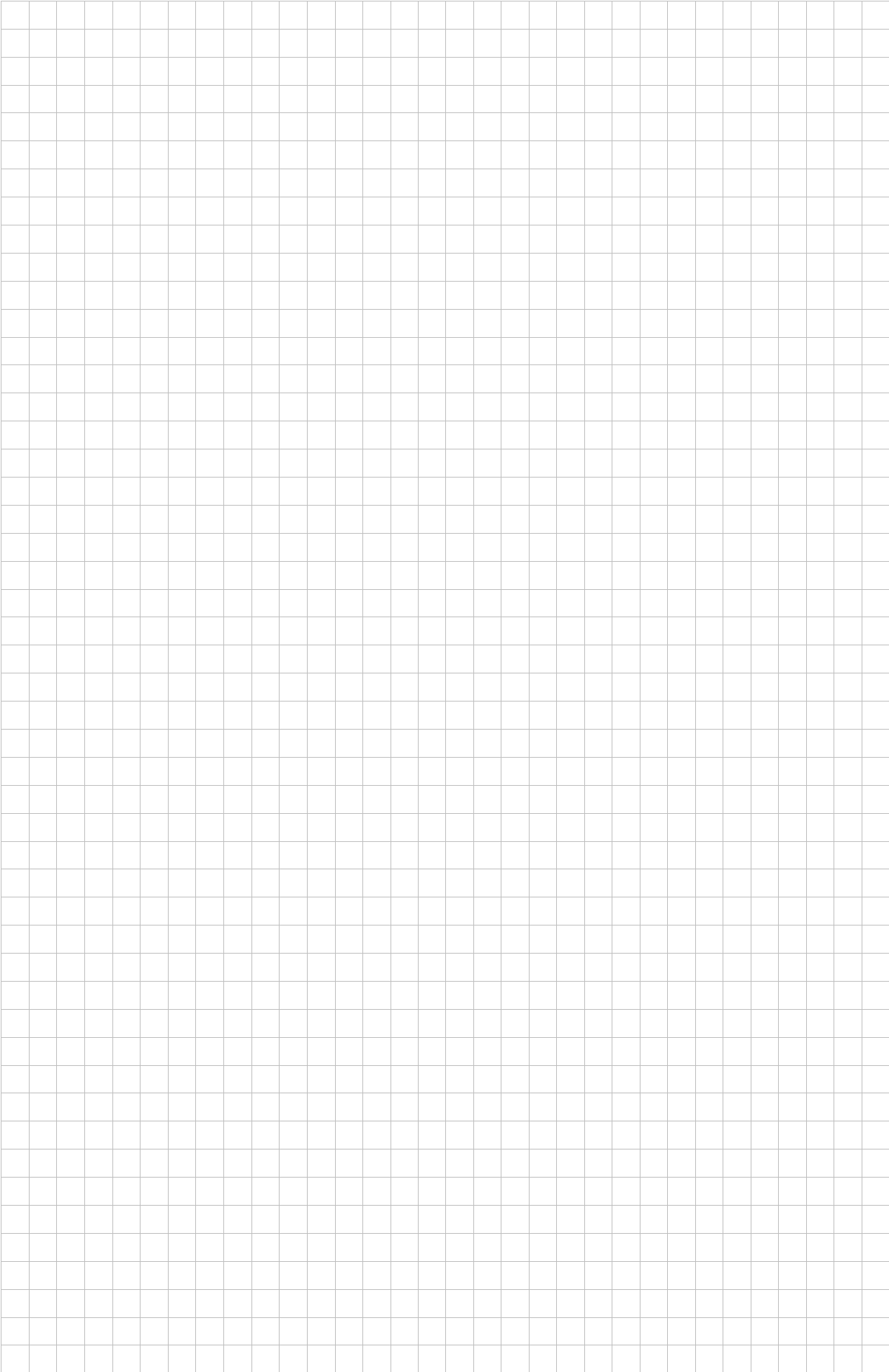
Nous avons donc notre résultat final. Meilleur que le précédent mais indépendant de la masse... mais en

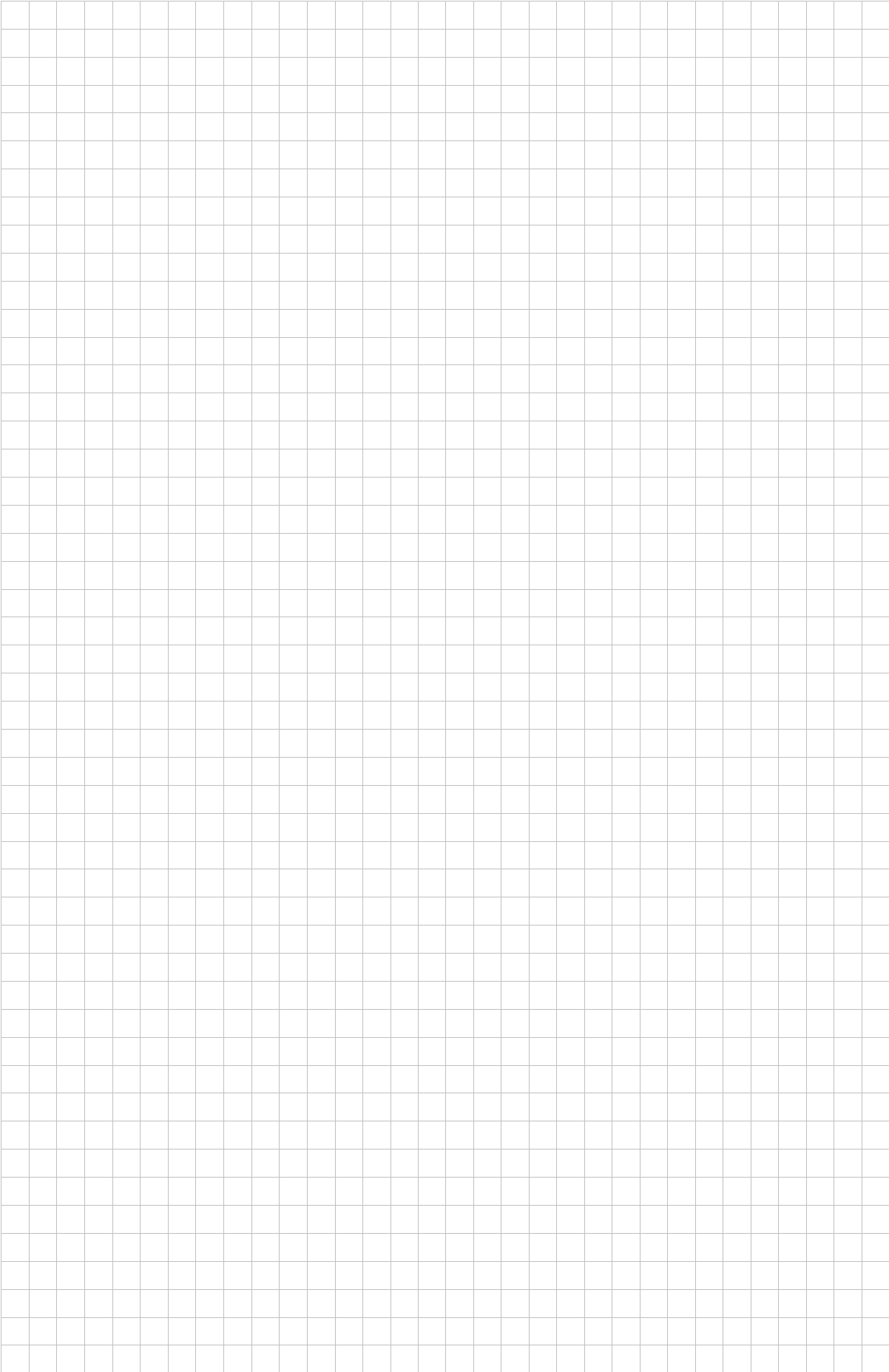
attendant c'est mieux que rien...

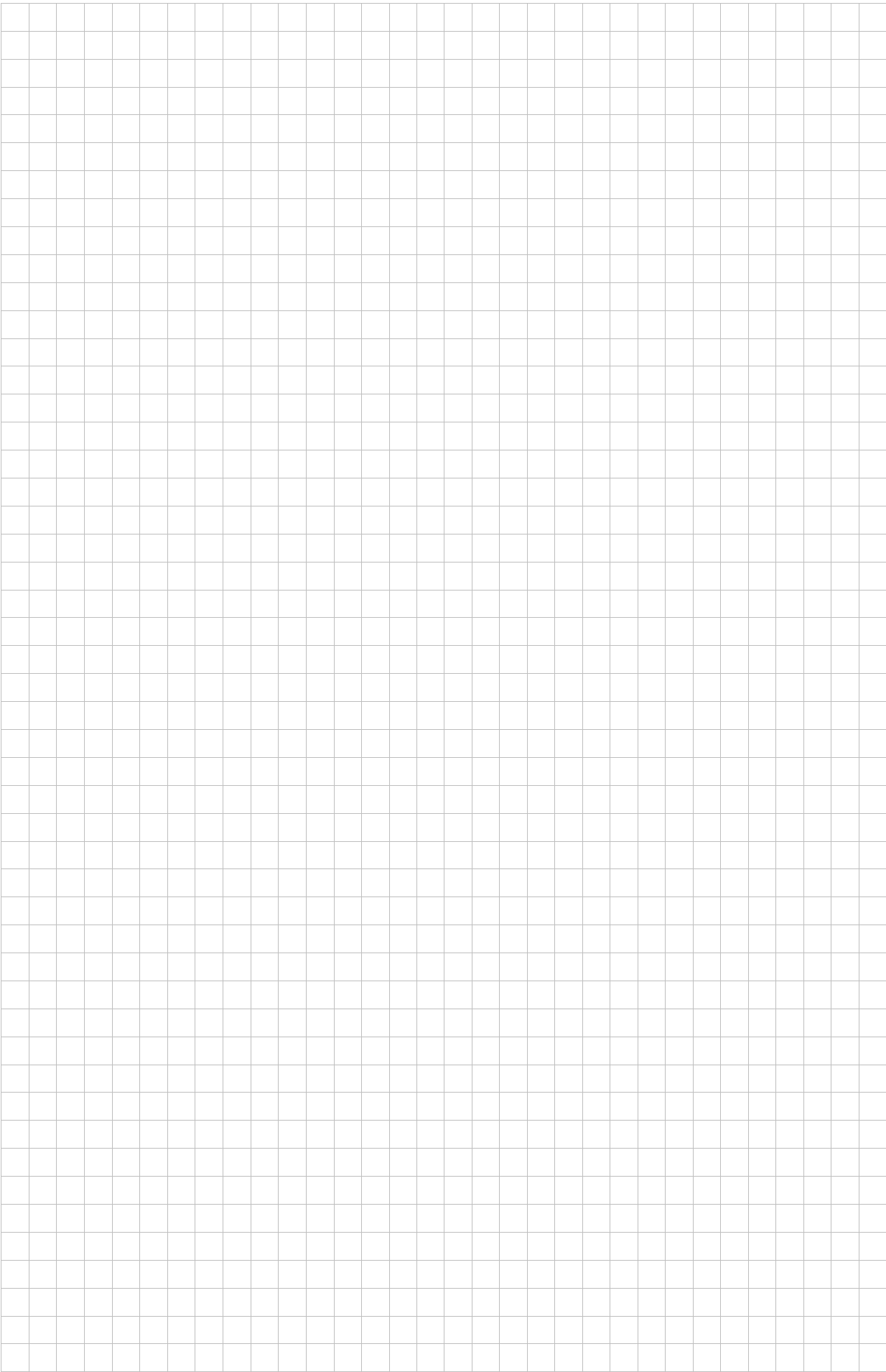
Remarque: Avec un freinage sec l'essieu avant d'une petite voiture (Mercedes Classe A) a une force de freinage typiquement de l'ordre de 2.8 [kN].

Notes personnelles:









31. MÉCANIQUE ONDULATOIRE

Nous nous intéresserons ici à l'étude des propriétés mathématiques des cordes vibrantes que nous pouvons également par extension et dans un souci de généralisation à des cas immatériels assimiler au concept des "ondes". Cette étude est très importante car nous aurons besoin de certains des résultats obtenus ici en thermodynamique, physique quantique, astrophysique, électrodynamique et théorie des cordes (pour ne citer que les plus importants).

Définitions:

D1. Une "onde" est un transport d'énergie sans transport de matière. Elle est concrétisée par la propagation d'une perturbation d'un milieu d'où l'appellation "d'ondes progressives". La vitesse avec laquelle l'onde progresse dépend des propriétés physiques du milieu.

D2. Dans le cas où la perturbation du milieu (déformation de l'onde) se fait de façon perpendiculaire à la direction de propagation de l'onde, nous parlons "d'onde transversale" ou de "perturbation transversale" (typique des ondes dans les cordes par exemple).

D3. Dans le cas où la perturbation du milieu se fait parallèlement à la direction de propagation de l'onde, nous parlons "d'onde longitudinale" ou de "perturbation longitudinale" (typique des ressorts).

1. FONCTION D'ONDE

Soit une perturbation $y = f(x)$ définie dans une région donnée de l'espace. Si nous remplaçons x par $x-b$, nous définissons dans cette même région, une perturbation $f(x-b)$ identique à $f(x)$ mais translatée d'une distance b dans la direction des X positifs (à droite donc si l'on adopte le système de représentation conventionnel vu en analyse fonctionnelle).

Si t représente un temps et si l'on pose $b = v \cdot t$, alors v peut désigner la vitesse de translation de la perturbation.

Ainsi, nous appelons "fonction d'onde", la relation mathématique:

$$f(x \pm vt) \quad (31.1)$$

qui décrit la progression d'une perturbation $y(x,t)$ dans l'espace:

- $y(x,t) = f(x - vt)$ décrivant une onde qui progresse vers $+X$
- $y(x,t) = f(x + vt)$ décrivant une onde qui progresse vers $-X$

v est par définition appelée "vitesse de phase de l'onde". Elle est constante dans un milieu homogène. "L'amplitude de l'onde" est la valeur maximale de la perturbation:

$$y_{\max} = \hat{y} \quad (31.2)$$

En l'absence d'amortissement, elle conservera la même valeur en chacun des points x où l'onde passe.

2. EQUATION D'ONDE

Sans aller dans des considérations trop techniques, nous dirons que toute fonction f dont l'argument est $x \pm vt$ jouit de la propriété:

$$\frac{\partial}{\partial x} f(x \pm vt) = \pm \frac{1}{v} \frac{\partial}{\partial t} f(x \pm vt) \quad (31.3)$$

Démonstration:

$$\begin{aligned} \frac{\partial}{\partial x} f(x - vt) &= 1 \cdot f'(x - vt) \Rightarrow f'(x - vt) = \frac{\partial}{\partial x} f(x - vt) \\ \frac{\partial}{\partial t} f(x - vt) &= -v \cdot f'(x - vt) \Rightarrow f'(x - vt) = -\frac{1}{v} \frac{\partial}{\partial t} f(x - vt) \end{aligned} \quad (31.4)$$

et donc l'égalité s'ensuit immédiatement (il s'agit simplement de l'application des dérivées composées tel que démontré dans le chapitre de Calcul Différentiel et Intégral).

Suite à la demande d'un internaute qui n'a pas trouvé cela très clair, faisons un exemple. Considérons la fonction suivante:

$$f(x - vt) = (x - vt)^2 \quad (31.5)$$

et dès lors:

$$\begin{aligned} \frac{\partial}{\partial x} f(x - vt) &= \frac{\partial}{\partial x} (x - vt)^2 = \underbrace{1}_{=\frac{\partial}{\partial x}(x-vt)} \cdot \underbrace{2(x-vt)}_{=f'} = 2x - vt \\ \frac{\partial}{\partial t} f(x - vt) &= \frac{\partial}{\partial t} (x - vt)^2 = \underbrace{-v}_{=\frac{\partial}{\partial t}(x-vt)} \cdot \underbrace{2(x-vt)}_{=f'} = -2xv + v^2 t \end{aligned} \quad (31.6)$$

Pour en revenir au cas général, dérivons une seconde fois, nous obtenons alors une autre forme de l'équation d'onde que nous retrouverons aussi fréquemment:

$$\begin{aligned} \frac{\partial^2}{\partial x^2} f(x - vt) &= 1 \cdot \frac{\partial}{\partial x} f'(x - vt) = 1 \cdot 1 \cdot f''(x - vt) \Rightarrow f''(x - vt) = \frac{\partial^2}{\partial x^2} f(x - vt) \\ \frac{\partial^2}{\partial t^2} f(x - vt) &= -v \cdot \frac{\partial}{\partial t} f'(x - vt) = v^2 \cdot f''(x - vt) \Rightarrow f''(x - vt) = \frac{1}{v^2} \frac{\partial^2}{\partial t^2} f(x - vt) \end{aligned} \quad (31.7)$$

Ce qui nous amène à écrire l'une des relations les plus importantes en physique appelée "**équation de propagation**", "**équation d'onde**" ou encore "**équation de d'Alembert**" et que nous retrouverons dans de nombreux autres chapitres du site (Électrodynamique, Physique Quantique Ondulatoire, Relativité Générale, Acoustique):

$$\frac{\partial^2}{\partial x^2} f(x \pm vt) = \frac{1}{v^2} \frac{\partial^2}{\partial t^2} f(x \pm vt) \quad (31.8)$$

Remarques:

R1. Ne pas oublier (cf. chapitre de Calcul Différentiel Et Intégral) que la somme des solutions à une équation différentielle est aussi solution de l'équation différentielle. Ainsi, la solution générale de l'équation d'Alembert est la superposition de deux ondes progressives arbitraires allant dans des directions opposées.

R2. Lorsque deux ou plusieurs ondes se propagent dans un milieu, la fonction d'onde qui en résulte est la somme algébrique des fonctions d'onde de chaque onde. Nous disons alors que les ondes "**interfèrent**" et nommons cette considération le "**principe de superposition des ondes**".

Considérons maintenant une corde de longueur L attachée par l'une de ses extrémités à une terminaison fixe. Supposons maintenant qu'une perturbation se propage sur cette corde. Lorsque la perturbation arrive à la terminaison, nous observons que celle-ci change de signe en même temps que sa vitesse de propagation s'inverse: l'onde subit ainsi une réflexion avec inversion.

Pour décrire le phénomène, il faut imposer:

$$y(L, t) = 0 \quad \forall t \quad (31.9)$$

Une fonction d'onde $f(x - vt)$ quelconque, qui progresse vers la terminaison, ne peut pas vérifier la condition:

$$f(x - vt)_{x=L} = f(L - vt) = 0 \quad (31.10)$$

pour toutes les valeurs du temps !

L'astuce consiste à la remplacer par une autre fonction d'onde $y(x, t)$ dont la forme est semblable à f à grande distance de l'origine de la perturbation, et qui s'annule au point de terminaison pour toutes les valeurs du temps. Pour cela, nous pouvons imaginer au point de la terminaison, un miroir qui donne de la corde une image de même longueur dans laquelle nous inventons une onde virtuelle:

$$f(2L - (x + vt)) = f((2L - x) - vt) \quad (31.11)$$

symétrique de $f(x - vt)$, mais de signe opposé.

Nous décidons ainsi:

- que les deux ondes progressent l'une vers l'autre pour s'annuler au point de terminaison
- toute partie de l'onde réelle qui dépasse le point de terminaison devient virtuelle
- toute partie de l'onde virtuelle qui pénètre dans la corde devient réelle

A leur intersection, les deux ondes réalisent une interférence destructive au point de terminaison. La somme algébrique de ces deux fonctions d'onde est aussi une fonction d'onde:

$$y(x, t) = f(x - vt) - f(2L - (x + vt)) \quad (31.12)$$

qui a la propriété voulue en $x=L$:

$$y(L, t) = f(L - vt) - f(2L - (L + vt)) = 0 \quad \forall t \quad (31.13)$$

Si nous considérons maintenant une terminaison libre, sans frottements avec son support d'attache, nous nous retrouvons dans un cas similaire au précédent à la différence que l'interférence est constructive au point de terminaison plutôt que destructive telle que la fonction d'onde s'écrive:

$$y(x, t) = f(x - vt) + f(2L - (x + vt)) \quad (31.14)$$

Remarques:

R1. Lorsque l'onde arrive sur une terminaison libre ou fixe, l'énergie transportée est intégralement renvoyée en arrière.

R2. Lorsqu'une terminaison n'est pas exactement adaptée, seule une partie de l'énergie est absorbée par le point Q, le reste est réfléchi.

3. TYPE D'ONDES

En physique théorique (et dans la pratique), nous restreignons fréquemment l'étude de certains phénomènes à des cas particuliers d'ondes. Principalement, nous en distinguons trois que nous allons brièvement mais soigneusement développer:

3.1. ONDES PÉRIODIQUES

Si un événement produit une onde qui ne transporte qu'une seule perturbation produite en un point donné, il existe de nombreuses perturbations qui sont capables d'exciter un milieu de manière répétitive.

Le point spatial de la source subit alors périodiquement la même perturbation. La durée d'un cycle complet est appelée identiquement à l'étude des pendules: "**la période**" T .

Si la perturbation peut se propager sous forme d'onde, à vitesse v , elle est décrite par la fonction d'onde que nous connaissons:

$$y = f(x, t) = f(x - vt) \quad (31.15)$$

En chaque point du milieu perturbé, l'onde périodique impose une "**périodicité temporelle**" de la perturbation qui nous impose d'écrire:

$$y(x = c^{te}) = f(x, t + n_1 \cdot T) = f(x, t) \quad \forall x \quad n_1 \in \mathbb{Z} \quad (31.16)$$

Après plusieurs cycles d'excitations de la source, plusieurs perturbations sont distribuées dans l'espace. La distance entre deux perturbations successives est appelée "**longueur d'onde**" λ .

La "**périodicité spatiale**" impose ainsi aussi:

$$y(t = c^{\#}) = f(x + n_2 \lambda, t) = f(x, t) \quad \forall t \quad n_2 \in \mathbb{Z} \quad (31.17)$$

λ est le chemin parcouru par l'onde pendant le temps T :

$$\lambda = v \cdot T \Leftrightarrow v = \frac{\lambda}{T} \quad (31.18)$$

Si une fonction d'onde est périodique dans le temps, elle l'est aussi dans l'espace, pour autant que l'impulsion ne se déforme pas lors de sa progression.

Démonstration:

$$\begin{aligned} f(x, t) &= f(x, t + n_1 \cdot T) = f(x - v(t + n_1 \cdot T)) = f(x - vt - n_1 vT) = f(x - n_1 \lambda - vt) \\ &= f(x - n_1 \lambda, t) \end{aligned} \quad (31.19)$$

En posant $n_2 = -n_1$, nous avons bien:

$$f(x, t) = f(x + n_2 \lambda, t) \quad (31.20)$$

□C.Q.F.D.

3.2. ONDES HARMONIQUES

Pour ces ondes, la fonction d'onde solution de l'équation d'Alembert est une fonction trigonométrique de type sinus ou cosinus (ou une somme):

$$y(x, t) = f(x - vt) = \hat{y} \cdot \sin(k(x - vt)) \quad \text{ou} \quad y(x, t) = f(x - vt) = \hat{y} \cdot \cos(k(x - vt)) \quad (31.21)$$

La présence de k appelé "**nombre d'onde**" est exigée pour 2 raisons:

- k s'exprime en $[1/m]$ pour la cohérence des unités des fonctions trigonométriques
- la valeur de k doit assurer la périodicité de la fonction d'onde:

1. périodicité angulaire de la fonction mathématique: $\sin(k(x - vt) \pm 2\pi) = \sin(k(x - vt))$

2. périodicité spatiale de la fonction d'onde: $\sin(k(x \pm \lambda) - vt) = \sin(k(x - vt))$

En égalant ces deux expressions:

$$k(x - vt) \pm 2\pi = k((x \pm \lambda) - vt) = k(x - vt) \pm k\lambda \quad (31.22)$$

ce qui implique:

$$k = \frac{2\pi}{\lambda} \quad (31.23)$$

Introduisons:

$$\lambda = v \cdot T \quad (31.24)$$

dans l'expression de k :

$$k \cdot v = \frac{2\pi}{T} = 2\pi f = \omega \quad (31.25)$$

d'où autre relation importante:

$$v = \frac{\omega}{k} \quad (31.26)$$

La fonction d'onde de l'onde harmonique peut alors s'écrire sous la forme:

$$y(x, t) = \hat{y} \cdot \sin(kx - \omega t) \text{ ou } y(x, t) = \hat{y} \cdot \sin \left[2\pi \left(\frac{x}{\lambda} - \frac{t}{T} \right) \right] \text{ progressive selon } +X$$

$$y(x, t) = \hat{y} \cdot \sin(kx + \omega t) \text{ ou } y(x, t) = \hat{y} \cdot \sin \left[2\pi \left(\frac{x}{\lambda} + \frac{t}{T} \right) \right] \text{ progressive selon } -X$$

(31.27)

3.3. ONDES STATIONNAIRES

Imaginons une corde excitée de manière harmonique. Au lieu d'adapter sa terminaison pour extraire de l'énergie de la corde, imposons que cette terminaison soit fixe. L'onde est alors réfléchiée.

Une nouvelle fonction d'onde doit être définie pour décrire la superposition de l'onde incidente:

$$y_1(x, t) = \hat{y} \cdot \sin(kx - \omega t) \quad (31.28)$$

et de l'onde réfléchiée (symétrique et de signe opposé):

$$y_2(x, t) = \hat{y} \cdot \sin(k(2L - x) - \omega t) \quad (31.29)$$

en analogie avec le résultat que nous avons trouvé lors de notre étude des terminaisons:

$$y(x, t) = f((2L - x) - vt) \quad (31.30)$$

La relation trigonométrique:

$$\sin(p) - \sin(q) = 2 \sin \left(\frac{p - q}{2} \right) \cos \left(\frac{p + q}{2} \right) \quad (31.31)$$

nous donne:

$$y(x, t) = y_1(x, t) + y_2(x, t) = 2\hat{y} \sin(k(x - L)) \cdot \cos(kL - \omega t) \quad (31.32)$$

Ce n'est plus une onde progressive car x et t ne se combinent plus comme $(x - vt)$. Certains points de la corde ne bougent jamais. Ils satisfont:

$$\sin(k(x - L)) = 0 \Rightarrow k(x - L) = \pm n\pi \Rightarrow \frac{2\pi}{\lambda}(x - L) = \pm n\pi \quad (31.33)$$

et sont situés en:

$$x = L \pm n \cdot \frac{\lambda}{2} \quad (31.34)$$

Pour des raisons évidentes, nous ne conservons que les valeurs de n pour lesquelles $x < L$.

Remarque: Chacun de ces points est appelé "**noeud de la vibration**".

Nous observons dans un tel système, des fuseaux de vibration, de longueur $\lambda/2$, dans lesquels la corde vibre transversalement dans une zone de hauteur $4\hat{y}$ (deux vers le haut, deux vers le bas).

Remarque: Les points où l'amplitude de vibration est maximale sont des "**ventres de la vibration**".

Puisque $\sin(k(x-L)) = 1$, les ventres sont distants de $\lambda/2$ et situés à $\lambda/4$ des noeuds.

Si nous imposons maintenant une terminaison fixe aux deux extrémités d'une corde en vibration, nous nous retrouvons avec une "**mise en résonance**".

Le plus souvent, nous n'observons pas grand-chose jusqu'à ce que nous trouvions la fréquence d'excitation qui place les noeuds de vibrations sur les deux points de terminaison fixe.

Dès lors pour une corde de longueur L :

$$x = 0 = L \pm n \cdot \frac{\lambda}{2} \quad (31.35)$$

implique:

$$L = n \cdot \frac{\lambda}{2} \Leftrightarrow \lambda = \frac{2L}{n} \quad (31.36)$$

La corde est alors le siège d'une onde stationnaire dont l'amplitude de vibration est considérablement plus grande que l'amplitude d'excitation (quatre fois).

Nous disons alors que la corde est rentrée en "**résonance**" avec l'excitateur.

La relation:

$$\lambda = \frac{2L}{n} \quad (31.37)$$

montre qu'il y a plusieurs longueurs d'onde possibles dont la plus grande correspondant à $n=1$ est appelée "**longueur d'onde fondamentale**" et vaut bien évidemment $\lambda = 2L$.

3.4. MODES DE VIBRATION TRANSVERSALE DANS UN FIL TENDU

Nous avons vu comment une onde peut progresser dans une corde. Montrons maintenant pourquoi c'est possible et établissons la relation $y(x,t)$, donnant la forme de la corde en fonction du temps.

Soit un fil de diamètre $\varnothing$, longueur L et masse m , la densité linéique du fil (supposée constante le long de celui-ci) est alors:

$$\rho_l = \frac{m}{L} = \frac{\Delta m}{\Delta L} = \frac{\Delta m}{\Delta x} \quad (31.38)$$

Par un léger choc, créons une petite perturbation (afin de ne pas déformer le câble et maintenir constant sa densité linéique) transversale. Isolons, dans la zone perturbée, un élément de fil, de longueur ΔL .

Approximations:

A1. Chaque élément de la corde peut être découpé de façon infinitésimale de façon à être presque parallèle à l'axe x . Les angles θ_P, θ_Q sont donc considérés comme petits.

A2. La corde est considérée comme déformable mais non allongeable donc la norme des forces dans la corde est constante en tout point quelle que soit la déformation.

Pour la suite du raisonnement, nous nous servons de la figure ci-dessous:

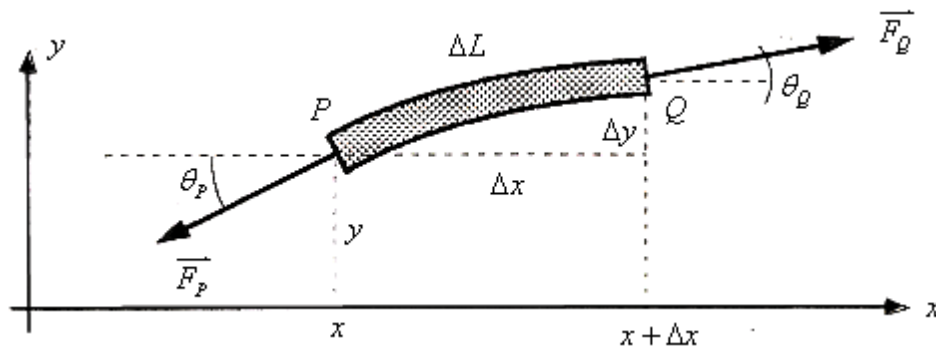


Figure: 31.1 - Illustration d'un élément de corde

Si les angles sont petits, le bilan des forces donne:

$$\sum F_x = F \cos \theta_Q - (F \cos \theta_P) = F(\cos \theta_Q - \cos \theta_P) \cong 0 \quad (31.39)$$

ce qui signifie qu'il n'y pas de déplacements selon x :

$$\sum F_y = F \sin \theta_Q - (F \sin \theta_P) = F(\sin \theta_Q - \sin \theta_P) \quad (31.40)$$

Si les angles sont vraiment petits, nous avons le premier terme du développement qui donne:

$$\sin x \cong \tan x \quad (31.41)$$

Donc:

$$\sum F_y = F(\sin \theta_Q - \sin \theta_P) \cong F(\tan \theta_Q - \tan \theta_P) \quad (31.42)$$

accélération selon y.

La loi de Newton appliquée à la masse $\Delta m = \rho_l \Delta x$ donne (nous considérons que chaque point de masse se déplace seulement selon y car il n'y a pas allongement):

$$\sum F_y = \Delta m \cdot a_y = \rho_l \Delta x \frac{\partial^2 y}{\partial t^2} \quad (31.43)$$

Les tangentes sont données par les dérivées partielles de la fonction y(x):

$$\sum F_y = F(\tan \theta_Q - \tan \theta_P) = F \left(\left. \frac{\partial y}{\partial x} \right|_Q - \left. \frac{\partial y}{\partial x} \right|_P \right) \quad (31.44)$$

Qui s'égale avec l'avant-dernière relation:

$$\rho_l \Delta x \frac{\partial^2 y}{\partial t^2} = F \left(\left. \frac{\partial y}{\partial x} \right|_Q - \left. \frac{\partial y}{\partial x} \right|_P \right) \quad (31.45)$$

et donc:

$$\frac{\partial^2 y}{\partial t^2} = \frac{F}{\rho_l} \frac{\left(\left. \frac{\partial y}{\partial x} \right|_Q - \left. \frac{\partial y}{\partial x} \right|_P \right)}{\Delta x} \quad (31.46)$$

Si $\Delta x \rightarrow 0$, les deux tangentes tendent vers la même valeur, mais la fraction du membre de droite tend vers une valeur finie:

$$\lim_{\Delta x \rightarrow 0} \frac{\left(\left. \frac{\partial y}{\partial x} \right|_Q - \left. \frac{\partial y}{\partial x} \right|_P \right)}{\Delta x} = \frac{\partial^2 y}{\partial x^2} \quad (31.47)$$

Il en résulte l'équation différentielle:

$$\frac{\partial^2 y}{\partial t^2} = \frac{F}{\rho_l} \frac{\partial^2 y}{\partial x^2} \quad (31.48)$$

Cette dernière relation s'écrit plus souvent sous la forme suivante:

$$\frac{\partial^2 y}{\partial x^2} - \frac{\rho_l}{F} \frac{\partial^2 y}{\partial t^2} = 0 \quad (31.49)$$

et se nomme "équation des cordes vibrantes".

Remarque: Dans certains ouvrages, la densité linéique est notée μ_0 et la force de tension dans la corde T_0 ce qui donne:

$$\frac{\partial^2 y}{\partial x^2} - \frac{\mu_0}{T_0} \frac{\partial^2 y}{\partial t^2} = 0 \quad (31.50)$$

Si nous vérifions que les unités de F / ρ_l sont celles du carré d'une vitesse $([m \cdot s^{-1}])^2$, comme l'exige l'analyse dimensionnelle. Pour simplifier l'écriture, nous posons:

$$v = \sqrt{\frac{F}{\rho_l}} \quad (31.51)$$

Nous allons maintenant considérer un cas particulier très intéressant dans le cadre de la musicologie qui est celui de la corde tendue (la plupart des instruments à corde fonctionnant ainsi).

3.4.1. CONDITIONS DE DIRICHLET

L'objectif est dans le cadre de l'équation différentielle obtenue précédemment (petites déformations dans les cadres des instruments de musique) de trouver une fonction $y(x,t)$ solution de cette dernière avec les conditions initiales, typiques à un instrument de musique suivantes:

C.I.1. $y(0,t) = y(L,t) = 0$ (les extrémités A et B sont fixes - il s'agit des "**conditions de Dirichlet**")

C.I.2. $y(x,0) = f(x)$ (forme initiale du fil à l'excitation)

C.I.3. $\frac{\partial y(x,0)}{\partial t} = 0$ (vitesse initiale nulle en tout point)

Les deux dernières conditions sont appelées "**conditions de Cauchy**".

Pour résoudre cette équation différentielle linéaire, nous allons faire usage de la méthode de séparation de variables en posant:

$$y(x,t) = X(x)T(t) \quad (31.52)$$

L'équation différentielle:

$$\frac{\partial^2 y}{\partial x^2} - \frac{\rho_l}{F} \frac{\partial^2 y}{\partial t^2} = 0 \quad (31.53)$$

devient dès lors:

$$\begin{aligned}\frac{\partial^2 y}{\partial t^2} &= \frac{\partial^2 (X(x)T(t))}{\partial t^2} = \frac{\partial}{\partial t} \left(\underbrace{\frac{\partial X(x)}{\partial t}}_{=0} T(t) + X(x) \frac{\partial T(t)}{\partial t} \right) = X(x) \frac{\partial^2 T(t)}{\partial t^2} = \frac{F}{\rho_l} \frac{\partial^2 y}{\partial x^2} \\ \frac{\partial^2 y}{\partial x^2} &= \frac{\partial^2 (X(x)T(t))}{\partial x^2} = \frac{\partial}{\partial x} \left(\frac{\partial X(x)}{\partial x} T(t) + X(x) \underbrace{\frac{\partial T(t)}{\partial x}}_{=0} \right) = T(t) \frac{\partial^2 X(x)}{\partial x^2}\end{aligned}\quad (31.54)$$

Nous avons donc:

$$\frac{\frac{\partial^2 X}{\partial x^2}}{X} = \frac{\rho_l}{F} \frac{\frac{\partial^2 T}{\partial t^2}}{T} \quad (31.55)$$

En mettant la deuxième relation dans la première, nous en tirons:

$$\begin{aligned}X(x) \frac{\partial^2 T(t)}{\partial t^2} &= \frac{F}{\rho_l} \frac{\partial^2 y}{\partial x^2} \\ \frac{\partial^2 y}{\partial x^2} &= T(t) \frac{\partial^2 X(x)}{\partial x^2}\end{aligned}\quad (31.56)$$

Le membre de gauche de la dernière relation ne contient pas la variable t et celui de droite ne contient pas la variable x . La seule et unique façon d'égaliser ces deux expressions est de les considérer chacune comme constante, que nous noterons $-\lambda^2$:

$$\frac{\frac{\partial^2 X}{\partial x^2}}{X} = \frac{\rho_l}{F} \frac{\frac{\partial^2 T}{\partial t^2}}{T} = c^* = -\lambda^2 \quad (31.57)$$

Ainsi, nous avons deux équations différentielles:

$$\frac{\partial^2 X}{\partial x^2} + \lambda^2 X = 0 \text{ et } \frac{\partial^2 T}{\partial t^2} + (\nu\lambda)^2 T = 0 \quad (31.58)$$

Ces deux équations étant similaires, résolvons-les de manière générale (cf. chapitre de Calcul Différentiel Et Intégral):

$$\frac{\partial^2 \alpha(t)}{\partial t^2} + \beta \alpha(t) = 0 \text{ avec } \beta \in \mathbb{R}^+ \quad (31.59)$$

L'équation caractéristique est donc:

$$r^2 + \beta = 0 \Rightarrow \Delta = b^2 - 4ac = -4\beta \quad (31.60)$$

d'où:

$$r_{1,2} = \frac{\pm \sqrt{-4\beta}}{2} = \pm i\sqrt{\beta} = u \pm iv = 0 \pm iv \quad (31.61)$$

Nous savons que la solution générale si les racines de l'équation caractéristique sont complexes, est de la forme:

$$\alpha(t) = C_1 e^{ut} \cos(vt) + C_2 e^{ut} \sin(vt) = C_1 \cos(\sqrt{\beta}t) + C_2 \sin(\sqrt{\beta}t) \quad (31.62)$$

Pour nos deux équations différentielles, nous avons donc par similitude:

$$X(x) = A_1 \cos(\lambda x) + B_1 \sin(\lambda x) \text{ et } T(t) = A_2 \cos(\lambda vt) + B_2 \sin(\lambda vt) \quad (31.63)$$

Cela donne pour la solution de notre équation d'onde:

$$y(x, t) = X(x)T(t) = (A_1 \cos(\lambda x) + B_1 \sin(\lambda x))(A_2 \cos(\lambda vt) + B_2 \sin(\lambda vt)) \quad (31.64)$$

Déterminons les constantes A_i, B_i en tenant compte des conditions initiales.

$$y(0, t) = 0 \Rightarrow 0 = A_1 (A_2 \cos(\lambda vt) + B_2 \sin(\lambda vt)), \forall t \Rightarrow A_1 = 0 \quad (31.65)$$

Il ne reste que:

$$y(x, t) = (B_1 \sin(\lambda x))(A_2 \cos(\lambda vt) + B_2 \sin(\lambda vt)) = \sin(\lambda x)(B_1 A_2 \cos(\lambda vt) + B_1 B_2 \sin(\lambda vt)) \quad (31.66)$$

Posons:

$$B_1 B_2 = a, B_1 A_2 = b \Rightarrow y(x, t) = \sin(\lambda x)(b \cos(\lambda vt) + a \sin(\lambda vt)) \quad (31.67)$$

La condition initiale $y(L, t) = 0, \forall t$ impose:

$$\begin{aligned} \sin(\lambda L) &= 0 \Rightarrow \lambda \cdot L = n \cdot \pi \Rightarrow \lambda = \frac{n \cdot \pi}{L}, n \in \mathbb{N} \\ \Rightarrow y(x, t) &= \sin\left(\frac{n\pi x}{L}\right) \left[b \cos\left(\frac{n\pi v}{L}t\right) + a \sin\left(\frac{n\pi v}{L}t\right) \right] \end{aligned} \quad (31.68)$$

Pour tenir compte de la vitesse initiale nulle, dérivons $y(x, t)$ par rapport au temps:

$$\begin{aligned} \frac{\partial y(x, t)}{\partial t} &= \sin\left(\frac{n\pi x}{L}\right) \left[-b \frac{n\pi v}{L} \sin\left(\frac{n\pi v}{L}t\right) + a \frac{n\pi v}{L} \cos\left(\frac{n\pi v}{L}t\right) \right] \\ \Rightarrow 0 &= \frac{\partial y(x, 0)}{\partial t} = \sin\left(\frac{n\pi x}{L}\right) \left[a \frac{n\pi v}{L} \cos(0) \right] \Rightarrow a = 0 \end{aligned} \quad (31.69)$$

Il ne reste que:

$$y(x, t) = b \sin\left(\frac{n\pi x}{L}\right) \cos\left(\frac{n\pi v}{L}t\right), y(x, t) \in [-b, +b] \quad (31.70)$$

La constante b représente donc l'amplitude du déplacement transversal du fil. Cette amplitude ne

pouvant être la même partout en un temps donné et une position donnée pour tout type d'excitation satisfaisant les conditions initiales, il doit exister autant de valeurs b_n que nous choisissons de valeurs n dans $n\pi x / \lambda l$.

Le principe de superposition des solutions des équations différentielles linéaires (cf. chapitre de Calcul Différentiel Et Intégral) permet d'écrire que la combinaison linéaire de toutes les solutions pour la corde est finalement:

$$y(x, t) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right) \cos\left(\frac{n\pi v}{L} t\right) \quad (31.71)$$

Les b_n doivent être choisis de manière à satisfaire la condition initiale qui donne la forme de la perturbation:

$$f(x) = y(x, 0) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right) \cos(0) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right) \quad (31.72)$$

Cette expression pour $f(x)$ suggère de la comparer au développement en série de Fourier (cf. chapitre des Suites Et Séries):

$$f(x) = \sum_{n=0}^{\infty} (a_n \cos(nkx) + b_n \sin(nkx)) \quad (31.73)$$

Dans laquelle $k = \pi / L$ et $a_n = 0, \forall n$. Le théorème de Fourier impose alors que les b_n sont donnés par:

$$b_0 = 0, b_n = \frac{2}{L} \int_0^L f(x) \sin(nkx) \quad (31.74)$$

Imaginons maintenant une corde de longueur L fixée en ses extrémités et tendue. Choisissons la perturbation la plus simple possible: nous grattons la corde en son milieu de manière très sèche, pour l'écarter d'une petite distance H de sa position d'équilibre.

La perturbation initiale $y(x, 0)$ est alors:

$$f(x) = \frac{2Hx}{L} \text{ pour } 0 \leq x \leq L/2 \text{ et } f(x) = \frac{2H(L-x)}{L} \text{ pour } L/2 \leq x \leq L \quad (31.75)$$

Calculons les coefficients de Fourier:

$$b_n = \frac{2}{L} \int_0^L f(x) \sin\left(\frac{n\pi x}{L}\right) dx = \frac{2}{L} \left\{ \int_0^{L/2} \frac{2Hx}{L} \sin\left(\frac{n\pi x}{L}\right) dx + \int_{L/2}^L \frac{2H(L-x)}{L} \sin\left(\frac{n\pi x}{L}\right) dx \right\} \quad (31.76)$$

L'intégration par parties (cf. chapitre de Calcul Différentiel Et Intégral) donne:

$$b_n = \frac{8H}{\pi^2} \frac{\sin\left(\frac{n\pi}{2}\right)}{n^2} \quad (31.77)$$

La fonction d'onde devient:

$$y(x,t) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right) \cos\left(\frac{n\pi v}{L}t\right) = \frac{8H}{\pi^2} \sum_{n=1}^{\infty} \frac{\sin\left(n\frac{\pi}{2}\right)}{n^2} \sin\left(\frac{n\pi x}{L}\right) \cos\left(\frac{n\pi v}{L}t\right) \quad (31.78)$$

À cause du $\sin(n\pi/2)$, les termes pour lesquels n est pair sont tous nuls. Il reste:

$$y(x,t) = \frac{8H}{\pi^2} \left[\frac{1}{1^2} \sin\left(\frac{\pi x}{L}\right) \cos\left(\frac{\pi v}{L}t\right) - \frac{1}{3^2} \sin\left(\frac{3\pi x}{L}\right) \cos\left(\frac{3\pi v}{L}t\right) + \frac{1}{5^2} \sin\left(\frac{\pi x}{L}\right) \cos\left(\frac{\pi v}{L}t\right) - \dots \right] \quad (31.79)$$

Si nous ne retenons que le terme en $n=1$, nous aurions:

$$y(x,t) = \frac{8H}{\pi^2} \sin\left(\frac{\pi x}{L}\right) \cos\left(\frac{\pi v}{L}t\right) = \frac{8H}{\pi^2} \sin\left(\frac{2\pi}{2L}x\right) \cos\left(\frac{2\pi}{2L}vt\right) = \frac{8H}{\pi^2} \sin(k_1 x) \cos(k_1 vt) \quad (31.80)$$

Nous avons:

$$k_1 = \frac{2\pi}{2L} \quad (31.81)$$

qui est le nombre d'onde correspondant à une longueur d'onde:

$$\lambda_1 = \frac{2\pi}{k_1} = 2L \quad (31.82)$$

et:

$$k_1 v = \omega_1 = 2\pi f_1 \Rightarrow f_1 = \frac{k_1 v}{2\pi} = \frac{v}{2L} = \frac{1}{2L} \sqrt{\frac{F}{\rho_i}} \quad (31.83)$$

qui serait la fréquence de vibration du fil de la première harmonique.

Ainsi, pour une valeur n quelconque, il est facile de démontrer que le n -ème "mode propre" est donné par:

$$y_n(x,t) = \frac{\hat{y}}{n^2} \sin\left(\frac{2\pi}{\lambda_n}x\right) \cos\left(\frac{2\pi}{f_n}t\right) \quad (31.84)$$

avec:

$$\lambda_n = \frac{2L}{n}, f_n = \frac{n}{2L} \sqrt{\frac{F}{\rho_i}} \quad (31.85)$$

relations appelées "lois de Mersenne" (1644-1648).

où le mode de plus basse fréquence avec n valant 1 est appelé le "mode fondamental" associé à sa "fréquence fondamentale".

Ainsi, après avoir été gratté sec au milieu de sa longueur L , un fil maintenu rigide à ses deux extrémités peut osciller suivant plusieurs modes. Le mode fondamental $n = 1$ (harmonique fondamentale) correspond à la plus petite fréquence possible. Il lui correspond la longueur d'onde $\lambda = 2L$.

Les fréquences d'ordre n supérieures sont appelées "**fréquences harmoniques**". Pour un même déplacement initial H , l'amplitude maximale de la vibration diminue selon n^{-2} comme nous le voyons dans l'expression de notre fonction.

Une autre manière d'exciter la corde est de la faire osciller de manière sinusoïdale, ce qui signifie dès lors que $y(x,t)$ est de la forme:

$$y(x,t) = y(x) \sin(\omega t - \phi) \quad (31.86)$$

En substituant cette relation dans l'équation d'onde de la corde, nous obtenons:

$$\frac{d^2 y(x)}{dx^2} + \omega^2 \frac{\rho_l}{F} y(x) = 0 \quad (31.87)$$

La solution se réduit alors à:

$$y(x) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right) \text{ ou } y_n(x) = b_n \sin\left(\frac{n\pi x}{L}\right), n \in \mathbb{N}^* \quad (31.88)$$

La valeur $n=0$ ne peut pas être incluse car elle donne une corde sans excitation. En mettant cette fonction dans l'équation d'onde précédente et en simplifiant, nous obtenons trivialement:

$$\omega_n = \sqrt{\frac{F}{\rho_l}} \left(\frac{n\pi}{L} \right) \quad (31.89)$$

Ce sont les "**fréquences d'oscillations de Dirichlet**" pour une corde. Les cordes d'un violon par exemple sont des cordes de Dirichlet.

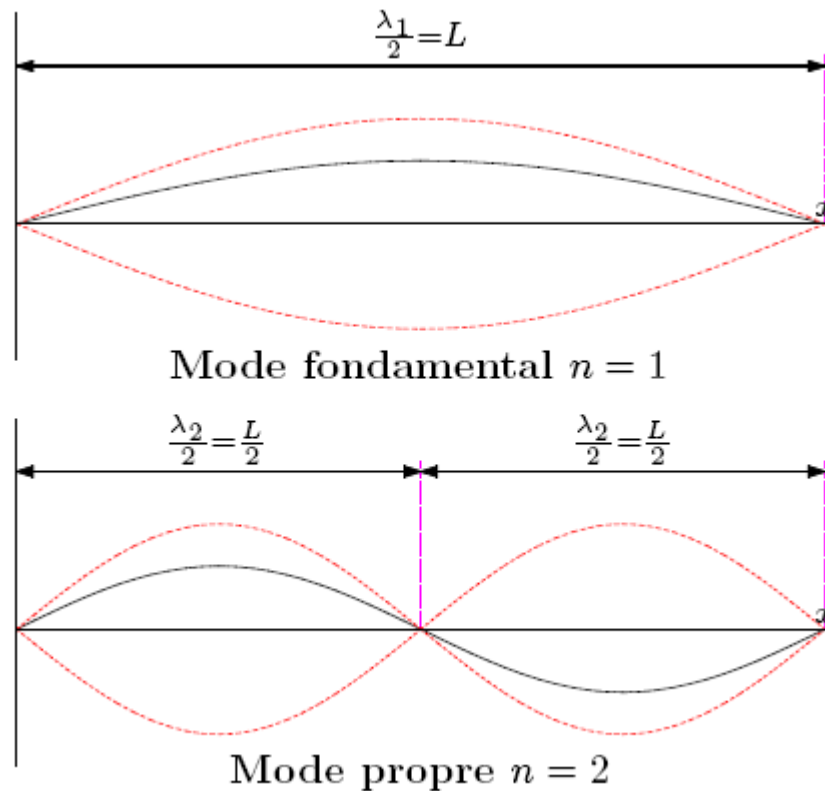


Figure: 31.2 - Quelques modes fondamentaux d'une corde fixée

Les mêmes analogies, raisonnements et développements pourront être faits avec les conditions de Neumann ci-dessous.

Remarques:

R1. La théorie prédit que la vibration peut être une combinaison linéaire de plusieurs modes. Ce phénomène porte le nom de "**vibration simultanée**". Il se produit abondamment dans un piano.

R2. Les instruments de musique sont conçus pour émettre des sons à des fréquences conventionnelles, étant admis que la hauteur d'une note perçue par l'oreille est définie par la fréquence fondamentale, par exemple le: Do (264 Hertz), La (440 Hertz).

R3. Lors de la construction de l'instrument, nous décidons de la valeur de ρ_i (en choisissant le diamètre et de la nature de la corde) et nous déterminons la longueur L en cherchant le compromis entre l'intensité sonore que nous voulons émettre et la résistance mécanique de l'instrument qui doit supporter les forces F de tension.

3.4.2. CONDITIONS DE NEUMANN

Alternativement aux conditions de Dirichlet où les extrémités sont fixes et à hauteur égales, les conditions de Neumann supposent que les extrémités sont de petites boucles autorisées à glisser le long de deux barres sans frottements.

Pour notre corde, les conditions de Neumann spécifient les valeurs $\partial y / \partial x$ aux extrémités. Mais tant que les boucles sont supposées sans masse et sans frottements, la dérivée $\partial y / \partial x$ doit s'annuler aux

extrémités $x = 0, L$. Si tel n'était pas le cas, alors de par la nullité de la masse de l'extrémité, le changement de vitesse sera dû à une accélération infinie, ce qui ne peut être autorisé ! C'est ainsi que nous imposons au lieu de la condition de Dirichlet, la condition de Neumann définie par:

$$\text{C.I.1. } \frac{\partial y}{\partial x}(0, t) = \frac{\partial y}{\partial x}(L, t) = 0$$

les conditions C.I.2. et C.I.3. restant identiques.

Ce changement de condition n'empêche pas que la méthode de résolution par séparation de variables est la même que précédemment et que nous tomberons identiquement sur la relation suivante à laquelle il faudra appliquer la nouvelle condition initiale:

$$y(x, t) = X(x)T(t) = (A_1 \cos(\lambda x) + B_1 \sin(\lambda x))(A_2 \cos(\lambda vt) + B_2 \sin(\lambda vt)) \quad (31.90)$$

sur laquelle nous appliquons donc la condition de Neumann:

$$\begin{aligned} \frac{\partial y}{\partial x}(0, t) &= 0 \\ \frac{\partial y}{\partial x} &= \lambda (-A_1 \sin(\lambda x) + B_1 \cos(\lambda x))(A_2 \cos(\lambda vt) + B_2 \sin(\lambda vt)) \\ \frac{\partial y}{\partial x}(0, t) &= 0 \Rightarrow \lambda (B_1)(A_2 \cos(\lambda vt) + B_2 \sin(\lambda vt)) = 0, \forall t \Rightarrow B_1 = 0 \end{aligned} \quad (31.91)$$

Il reste donc:

$$y(x, t) = \cos(\lambda x)(A_1 A_2 \cos(\lambda vt) + A_1 B_2 \sin(\lambda vt)) \quad (31.92)$$

en posant $b = A_1 A_2, a = A_1 B_2$ la fonction se simplifie en:

$$y(x, t) = \cos(\lambda x)(b \cos(\lambda vt) + a \sin(\lambda vt)) \quad (31.93)$$

La condition initiale, $\frac{\partial y}{\partial x}(L, t) = 0, \forall t$ impose:

$$\begin{aligned} -\sin(\lambda L) &= 0 \Rightarrow \lambda \cdot L = n \cdot \pi \Rightarrow \lambda = \frac{n \cdot \pi}{L}, n \in \mathbb{N} \\ \Rightarrow y(x, t) &= \cos\left(\frac{n\pi x}{L}\right) \left[b \cos\left(\frac{n\pi v}{L} t\right) + a \sin\left(\frac{n\pi v}{L} t\right) \right] \end{aligned} \quad (31.94)$$

Les mêmes développements pour la C.I.2. que nous avons faits avec la C.I.1. de Dirichlet s'appliquent ensuite de manière identique:

$$y(x, t) = b \cos\left(\frac{n\pi x}{L}\right) \cos\left(\frac{n\pi v}{L} t\right) \quad (31.95)$$

ensuite, l'analogie avec les séries de Fourier s'applique de manière similaire mais avec les cosinus au lieu des sinus.

Les fréquences de Neumann d'une corde sont les mêmes que pour celle de Dirichlet soit:

$$\omega_n = \sqrt{\frac{F}{\rho_l}} \left(\frac{n\pi}{L} \right) \quad (31.96)$$

La particularité réside cependant dans la valeur de la fonction spatiale qui vaut cette fois trivialement:

$$y(x) = \sum_{n=0}^{\infty} b_n \cos\left(\frac{n\pi x}{L}\right) \text{ ou } y_n(x) = b_n \cos\left(\frac{n\pi x}{L}\right), n \in \mathbb{N} \quad (31.97)$$

Effectivement, pour $n=0$ nous avons cette fois une amplitude b_n identique qui est transmise tout le long de la corde sans que celle-ci ne vibre cependant !

Par ailleurs, faisons remarquer, que la fonction $y(x,t) = at + b$ satisfait aussi pleinement les trois conditions initiales incluant celle de Neumann.

Effectivement, nous avons bien:

$$\begin{aligned} \text{C.I.1. } \frac{\partial y}{\partial x}(0,t) &= \frac{\partial y}{\partial x}(L,t) = 0 \\ \text{C.I.2. } y(x,0) &= f(x) = b \\ \text{C.I.3. } \frac{\partial y(x,0)}{\partial t} &= 0 \end{aligned} \quad (31.98)$$

et de plus, $y(x,t) = at + b$ vérifie aussi l'équation d'onde:

$$\frac{\partial^2 y}{\partial x^2} - \frac{1}{v^2} \frac{\partial^2 y}{\partial t^2} = 0 \quad (31.99)$$

3.4.3. LAGRANGIEN D'UNE CORDE

Nous allons maintenant déterminer le lagrangien d'une corde, calcul qui nous sera en partie utile lors de l'étude de la théorie des cordes.

Nous gardons donc notre corde ayant une densité linéique et tension constante dont les extrémités sont situées en $x=0, x=L$ et dont la vitesse de la perturbation transversale est non relativiste.

L'énergie cinétique est alors simplement la somme des énergies cinétiques de chaque élément infinitésimal de la corde. Nous pouvons alors écrire en notation Lagrangienne:

$$T = \int_0^L \frac{1}{2} (\rho_l dx) \left(\frac{\partial y}{\partial t} \right)^2 \quad (31.100)$$

L'énergie potentielle intervient dans l'élongation de la corde dont une portion infinitésimale peut être vue comme variant de $(x,0)$ à $(x+dx,0)$ quand la corde est à l'équilibre. Quand une corde est momentanément mise sous tension de (x,y) à $(x+dx,y+dy)$ alors la variation de la longueur dl d'un élément infinitésimal de la corde est donnée trivialement par:

$$dl = \sqrt{(dx)^2 + (dy)^2} - dx = dx \left(\sqrt{1 + \left(\frac{\partial y}{\partial x} \right)^2} - 1 \right) \cong dx \frac{1}{2} \left(\frac{\partial y}{\partial x} \right)^2 \quad (31.101)$$

Nous avons utilisé ci-dessus pour approximation le développement limité au deuxième terme en série de Taylor (cf. chapitre sur les Suites Et Séries), qui nous donne:

$$\sqrt{1+x} = 1 + \frac{1}{2}x - \frac{1}{2 \cdot 4}x^2 + \frac{1 \cdot 3}{2 \cdot 4 \cdot 6}x^3 - \dots, |x| \leq 1 \quad (31.102)$$

Le travail effectué pour étirer chaque élément infinitésimal étant $F \Delta l$, l'énergie potentielle totale est alors exprimée par:

$$V = \int_0^L F dl \cong \int_0^L \frac{1}{2} F \left(\frac{\partial y}{\partial x} \right)^2 dx \quad (31.103)$$

La lagrangien étant défini par $L = T - V$ (cf. chapitre de Mécanique Analytique), nous avons alors:

$$L = \int_0^L \left[\frac{1}{2} \rho_t \left(\frac{\partial y}{\partial t} \right)^2 - \frac{1}{2} F \left(\frac{\partial y}{\partial x} \right)^2 \right] dx = \int_0^L \mathcal{L} dx \quad (31.104)$$

où $\mathcal{L}$ est défini, très justement, comme étant la "**densité lagrangienne**":

$$\mathcal{L} = \frac{1}{2} \rho_t \left(\frac{\partial y}{\partial t} \right)^2 - \frac{1}{2} F \left(\frac{\partial y}{\partial x} \right)^2 \quad (31.105)$$

L'action pour notre corde est alors:

$$S = \int_{t_i}^{t_f} L dt = \int_{t_i}^{t_f} dt \int_0^L dx \left[\frac{1}{2} \rho_t \left(\frac{\partial y}{\partial t} \right)^2 - \frac{1}{2} F \left(\frac{\partial y}{\partial x} \right)^2 \right] \quad (31.106)$$

Dans cette action, le chemin d'action est la fonction $y(x,t)$. Pour trouver les équations du mouvement, nous devons examiner la variation de l'action quand nous varions:

$$y(x,t) \rightarrow y(x,t) + \delta y(x,t) \quad (31.107)$$

Ce qui donne:

$$\begin{aligned}
\delta S &= \int_{t_i}^{t_f} dt \int_0^L dx \left[\frac{1}{2} \rho_t \left(\frac{\partial (y(x,t) + \delta y(x,t))}{\partial t} \right)^2 - \frac{1}{2} F \left(\frac{\partial (y(x,t) + \delta y(x,t))}{\partial x} \right)^2 \right] \\
&= \int_{t_i}^{t_f} dt \int_0^L dx \left[\frac{1}{2} \rho_t \left(\frac{\partial y(x,t)^2}{\partial t^2} + \frac{2 \partial y(x,t) \partial (\delta y(x,t))}{\partial t^2} + \frac{\partial (\delta y(x,t))^2}{\partial t^2} \right) \right. \\
&\quad \left. - \frac{1}{2} F \left(\frac{\partial y(x,t)^2}{\partial x^2} + \frac{2 \partial y(x,t) \partial (\delta y(x,t))}{\partial x^2} + \frac{\partial (\delta y(x,t))^2}{\partial x^2} \right) \right] \quad (31.108) \\
&\cong \int_{t_i}^{t_f} dt \int_0^L dx \left[\rho_t \frac{\partial y(x,t)}{\partial t} \frac{\partial (\delta y(x,t))}{\partial t} - F \frac{\partial y(x,t)}{\partial x} \frac{\partial (\delta y(x,t))}{\partial x} \right] \\
&= \int_{t_i}^{t_f} dt \int_0^L dx \left[\rho_t \frac{\partial y}{\partial t} \frac{\partial (\delta y)}{\partial t} - F \frac{\partial y}{\partial x} \frac{\partial (\delta y)}{\partial x} \right]
\end{aligned}$$

Car:

$$\left(\frac{\partial (\delta y(x,t))}{\partial t} \right)^2 \ll \left(\frac{\partial y(x,t)}{\partial t} \right)^2 \ll \frac{2 \partial y(x,t) \partial (\delta y(x,t))}{\partial t^2} \quad (31.109)$$

et ce identiquement pour le second terme.

Nous ne devons pas avoir de dérivées temporelles agissant sur les variations. Alors en utilisant la relation triviale suivante sur le premier terme:

$$\frac{\partial}{\partial t} \rho_t \left(\frac{\partial y}{\partial t} \delta y \right) = \rho_t \frac{\partial^2 y}{\partial t^2} \delta y + \rho_t \frac{\partial y}{\partial t} \frac{\partial (\delta y)}{\partial t} \Rightarrow \rho_t \frac{\partial y}{\partial t} \frac{\partial (\delta y)}{\partial t} = \frac{\partial}{\partial t} \rho_t \left(\frac{\partial y}{\partial t} \delta y \right) - \rho_t \frac{\partial^2 y}{\partial t^2} \delta y \quad (31.110)$$

et identiquement sur le deuxième, nous pouvons réécrire l'action:

$$\begin{aligned}
\delta S &= \int_{t_i}^{t_f} dt \int_0^L dx \left[\rho_t \frac{\partial y}{\partial t} \frac{\partial (\delta y)}{\partial t} - F \frac{\partial y}{\partial x} \frac{\partial (\delta y)}{\partial x} \right] \\
&= \int_{t_i}^{t_f} dt \int_0^L dx \left[\frac{\partial}{\partial t} \left(\rho_t \frac{\partial y}{\partial t} \delta y \right) - \rho_t \frac{\partial^2 y}{\partial t^2} \delta y + \frac{\partial}{\partial x} \left(-F \frac{\partial y}{\partial x} \delta y \right) + F \frac{\partial^2 y}{\partial x^2} \delta y \right] \quad (31.111) \\
&= \int_{t_i}^{t_f} \partial \left(\rho_t \frac{\partial y}{\partial t} \delta y \right) \Big|_0^L - \int_{t_i}^{t_f} dt \int_0^L dx \rho_t \frac{\partial^2 y}{\partial t^2} \delta y + \int_0^L \partial \left(-F \frac{\partial y}{\partial x} \delta y \right) \Big|_{t_i}^{t_f} + \int_{t_i}^{t_f} dt \int_0^L dx F \frac{\partial^2 y}{\partial x^2} \delta y
\end{aligned}$$

Comme nous l'avons vu en mécanique analytique, le bon chemin est donné par $\delta S = 0$. Dès lors, nous devons avoir:

$$\delta S = \int_0^L \left[\rho_t \frac{\partial y}{\partial t} \delta y \right]_{t_i}^{t_f} dx + \int_{t_i}^{t_f} \left[-F \frac{\partial y}{\partial x} \delta y \right]_0^L dt - \int_{t_i}^{t_f} dt \int_0^L dx \left(\rho_t \frac{\partial^2 y}{\partial t^2} - F \frac{\partial^2 y}{\partial x^2} \right) \delta y = 0 \quad (31.112)$$

Ainsi, notre expression contient trois termes. Chacun de ces trois termes doit s'annuler

indépendamment comme nous allons le voir:

1. L'annulation du troisième terme se fait selon une condition triviale qui nous est déjà bien connue (heureusement...):

$$\int_{t_i}^{t_f} dt \int_0^L dx \left(\rho_i \frac{\partial^2 y}{\partial t^2} - F \frac{\partial^2 y}{\partial x^2} \right) \delta y = 0 \Rightarrow \rho_i \frac{\partial^2 y}{\partial t^2} - F \frac{\partial^2 y}{\partial x^2} = 0 \quad (31.113)$$

et donc:

$$\frac{\partial^2 y}{\partial x^2} - \frac{\rho_i}{F} \frac{\partial^2 y}{\partial t^2} = 0 \quad (31.114)$$

nous retrouvons donc l'équation différentielle d'une onde transversale telle que nous l'avions démontrée plus haut. Notre hypothèse sur le troisième terme ne peut donc être que juste ainsi que l'expression de notre action.

2. Le premier terme est déterminé par la configuration de la corde aux temps t_i, t_f :

$$\int_0^L \left[\rho_i \frac{\partial y}{\partial t} \delta y \right]_{t_i}^{t_f} dx \quad (31.115)$$

Or, si nous imposons la connaissance de ces configurations dans le temps, nous aurons par définition:

$$\delta y(x, t_f) = \delta y(x, t_i) = 0 \quad (31.116)$$

(connaissance totale du chemin d'action car connaissance des conditions initiales, donc pas de variation). Cela valide encore une fois l'expression de notre action et la valeur nulle du terme comme attendu.

3. Le second terme est un peu plus intéressant:

$$\int_{t_i}^{t_f} \left[-F \frac{\partial y}{\partial x} \delta y \right]_0^L dt \quad (31.117)$$

D'abord, ce n'est que parce que nous connaissons les positions des extrémités de la corde que nous pouvons connaître ces modes de vibrations, nous le savons bien! Il nous faut donc savoir comment se comportent les extrémités. Pour cela nous allons revenir sur des choses qui nous sont connues: les conditions de Dirichlet et de Neumann d'une corde.

Supposons que nous imposons les conditions de Dirichlet (voir plus haut), les extrémités sont alors fixes et nous aurons forcément à ces mêmes extrémités:

$$\delta y(0, t) = \delta y(L, t) = 0 \quad (31.118)$$

et donc le deuxième terme disparaît bien (ouf!).

Si, à l'opposé, nous choisissons que les extrémités se meuvent librement, alors les variations:

$$\delta y(0,t) = \delta y(L,t) \quad (31.119)$$

sont non contraintes et dès lors, seulement les conditions de Neumann:

$$\frac{\partial y}{\partial x}(t,0) = \frac{\partial y}{\partial x}(t,L) = 0$$

(voir plus haut pour plus de détails) nous permettront d'avoir le deuxième terme de l'action nul.

Pour prendre pleinement conscience de l'importance de ces conditions initiales, considérons la quantité de mouvement p_y portée par la corde (il n'y pas d'autres composantes du mouvement car nous avons supposé implicitement une excitation transversale dès le début seulement dans cette direction y).

La quantité de mouvement est simplement la somme des quantités de mouvement de chaque élément infinitésimal le long de la corde:

$$p_y = \int_0^L (\rho_l dx) \frac{\partial y}{\partial t} \quad (31.120)$$

Vérifions juste par curiosité (c'est une curiosité anticipée...) si la quantité de mouvement est bien conservée:

$$\frac{dp_y}{dt} = \int_0^L (\rho_l dx) \frac{\partial^2 y}{\partial t^2} = \int_0^L F \frac{\partial^2 y}{\partial x^2} dx = F \frac{\partial y}{\partial x} \Big|_{x=0}^{x=L} \quad (31.121)$$

où nous avons utilisé l'équation d'onde transversale pour la substitution.

Nous voyons par le résultat de ce petit calcul que la quantité de mouvement est trivialement conservée si nous imposons les conditions de Neumann, alors que pour les conditions de Dirichlet, la plupart du temps la conservation n'est pas respectée! Effectivement, c'est trivial (il n'y pas besoin de calculs pour s'en rendre compte), lorsque les extrémités sont attachées au mur, le mur exerce constamment une force sur la corde.

3.5. MODES DE VIBRATION DANS UNE MEMBRANE TENDUE

Nous dérivons le phénomène de la même manière que la vibration transversale de la corde. Toutefois, la densité linéique ρ_l du fil doit être remplacée par la masse surfacique ρ_s de la membrane.

De plus, nous remplaçons la force F de tension unidirectionnelle du fil par une force de tension appliquée sur le pourtour de la membrane. Cette force s'exerce dans toutes les directions du plan et se décrit par unité de longueur:

$$\tau = \frac{dF}{dl} \quad (31.122)$$

Nous avons (analyse dimensionnelle):

$$dF_z = \tau dx dy \cdot f_{x,y}(x,y,t) = \sigma dx dy f_t(x,y,t) \quad (31.123)$$

Il est d'abord évident que:

$$f_t(x, y, t) = \frac{\partial^2 \psi(x, y, t)}{\partial t^2} \Rightarrow dF_z = \sigma dx dy \frac{\partial^2 \psi(x, y, t)}{\partial t^2} \quad (31.124)$$

et comme:

$$dF_z = dF_x + dF_y = \tau dx dy \cdot f_x(x, y, t) + \tau dx dy \cdot f_y(x, y, t) = \tau dx dy (f_x(x, y, t) + f_y(x, y, t)) \quad (31.125)$$

L'analyse dimensionnelle (eh oui ... à nouveau...) donne:

$$dF_z = \tau dx dy \left(\frac{\partial^2 \psi(x, y, t)}{\partial x^2} + \frac{\partial^2 \psi(x, y, t)}{\partial y^2} \right) \quad (31.126)$$

Nous avons donc:

$$\tau dx dy \left(\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} \right) = \sigma dx dy \frac{\partial^2 \psi}{\partial t^2} \Rightarrow \frac{\partial^2 \psi}{\partial t^2} = \frac{\tau}{\sigma} \left(\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} \right) \quad (31.127)$$

L'analyse dimensionnelle donne:

$$v = \sqrt{\frac{\tau}{\sigma}} \quad (31.128)$$

Donc finalement nous obtenons pour équation d'onde en coordonnées cartésiennes (exprimé avec le laplacien):

$$\left(\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} \right) = \vec{\nabla}^2 \psi = \frac{1}{v^2} \frac{\partial^2 \psi}{\partial t^2} \quad (31.129)$$

Nous cherchons la solution particulière de cette équation qui vérifie les conditions suivantes:

C.I.1. La membrane est fixée sur son pourtour R (conditions aux limites)

C.I.2. La position et la vitesse initiales sont données (conditions initiales)

La symétrie du problème suggère d'utiliser le laplacien en coordonnées polaires (cf. chapitre de Calcul Vectoriel):

$$\frac{\partial^2 y}{\partial r^2} + \frac{1}{r} \frac{\partial y}{\partial r} + \frac{1}{r^2} \frac{\partial^2 y}{\partial \phi^2} = \frac{1}{v^2} \frac{\partial^2 y}{\partial t^2} \quad (31.130)$$

Remarque: Nous avons changé de notation en posant $y := \psi$

Et les conditions fixées:

C1. $\forall t \geq 0, \forall \theta \in [0, 2\pi], y(R, \theta, t) = 0$ (conditions aux limites)

$$C2. \forall \theta \in [0, 2\pi], \forall r \in [0, R] \begin{cases} f(r, \theta, 0) = f(r, \theta) \\ \frac{\partial y}{\partial t}(r, \theta, 0) = \varphi(r, \theta) \end{cases} \quad (\text{conditions initiales})$$

où les fonctions f, φ sont données.

À nouveau, pour chercher la solution, nous allons utiliser la méthode de séparation des variables telle que:

$$y(r, \theta, t) = X(r, \theta)T(t) \quad (31.131)$$

et de même que pour la corde:

$$\frac{\frac{\partial^2 X}{\partial x^2}}{x} = \frac{1}{v^2} \frac{\partial^2 T}{T} \quad (31.132)$$

et identiquement que pour la corde, nous obtenons pour T une solution du type:

$$T(t) = A \cos(\lambda v t) + B \sin(\lambda v t) \quad (31.133)$$

Pour $X(r, \theta)$ la méthode change car nous avons maintenant une équation différentielle à deux variables telle que:

$$\frac{\partial^2 X(r, \theta)}{\partial x(r, \theta)^2} + \lambda^2 x(r, \theta) = 0 \quad (31.134)$$

Pour intégrer cette équation, nous cherchons les solutions de la forme $X(r, \theta) = R(r)\Theta(\theta)$, nous obtenons en reportant:

$$R''\Theta + \frac{1}{r}R'\Theta + \frac{1}{r^2}R\Theta'' + \lambda^2 R\Theta = 0 \quad (31.135)$$

En se rappelant qu'en coordonnées polaires:

$$\vec{\nabla}^2 f = \frac{\partial^2 f}{\partial r^2} + \frac{1}{r} \frac{\partial f}{\partial r} + \frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2} \quad (31.136)$$

D'où, en séparant les variables:

$$\frac{\Theta''}{\Theta}(\theta) = -\frac{1}{R(r)}(r^2 R''(r) + r R'(r) + \lambda^2 r^2 R(r)) \quad (31.137)$$

Le membre de gauche de la dernière relation ne contient pas la variable r et celui de droite ne contient pas la variable θ . La seule et unique façon d'égaliser ces deux expressions est de les considérer chacune comme constante, que nous noterons μ . Les équations différentielles vérifiées par R et θ sont alors:

$$r^2 R''(r) + r R'(r) + (\lambda^2 r^2 + \mu) R(r) = 0 \quad \Theta''(\theta) - \mu \Theta(\theta) = 0 \quad (31.138)$$

La fonction Θ est périodique de période 2π , il existe donc un entier naturel n tel que $\mu = -n^2$ et donc manière identique à la corde, nous obtenons:

$$\Theta(\theta) = C \cos(n\theta) + D \sin(n\theta) \quad (31.139)$$

Dans la première équation différentielle:

$$r^2 R''(r) + r R'(r) + (\lambda^2 r^2 - n^2) R(r) = 0 \quad (31.140)$$

Pour simplifier, nous effectuons le changement de variable $x = \lambda r$. L'équation différentielle devient:

$$x^2 \frac{d^2 R}{dx^2} + x \frac{dR}{dx} + (x^2 - n^2) R = 0 \quad (31.141)$$

Nous reconnaissons ici l'équation différentielle de Bessel d'ordre n telle que nous l'avons avec sa solution présentée dans le chapitre des Suites Et Séries. Dès lors, la solution générale est du type:

$$R(r) = J_n(x) = J_n(\lambda r) \quad (31.142)$$

Ce qui nous donne finalement:

$$y(r, \theta, t) = R(r)\Theta(\theta)T(t) = J_n(\lambda r) (C \cos(n\theta) + D \sin(n\theta)) (A \cos(\lambda vt) + B \sin(\lambda vt)) \quad (31.143)$$

Parmi les solutions à cette équation, cherchons celles qui vérifient les conditions aux limites en posant $R_{\text{membrane}} = L$:

$$\forall \theta \in [0, 2\pi], \forall t \geq 0, R(L)\Theta(\theta)T(t) = 0 \quad (31.144)$$

À moins que Θ ou T soit la fonction nulle, ce qui donne pour solution la position d'équilibre... (qui ne vérifie sans doute pas les conditions initiales), nous devons avoir $R(L) = 0$, c'est-à-dire:

$$J_n(bl) = 0 \quad (31.145)$$

La fonction Bessel d'ordre J_n a une infinité de zéros positifs $\left(\frac{\xi_m^n}{L}\right)_{m \in \mathbb{N}}$ (il suffit de tracer cette fonction avec un ordinateur pour le voir tel qu'avec Maple en mettant la commande: `plot(BesselJ(2,x),x=0...100)` où vous pouvez changer la valeur 2 par une autre valeur) qui fournissent une infinité de valeurs convenables de b telle que:

$$\lambda_m^n = \frac{\xi_m^n}{L} \quad (31.146)$$

Ce qui correspond finalement à une infinité de solutions de l'équation différentielle initiale que nous pouvons écrire:

$$y_m^n(r, \theta, t) = J_n(\lambda_m^n r) \left\{ \left(\alpha_m^n \cos(\omega_m^n t) + \beta_m^n \sin(\omega_m^n t) \right) \cos(n\theta) + \left(\gamma_m^n \cos(\omega_m^n t) + \delta_m^n \sin(\omega_m^n t) \right) \sin(n\theta) \right\} \quad (31.147)$$

En ayant modifié le nom des constantes d'intégration et en ayant posé $\omega_m^n = \lambda_m^n \nu$ (ce qui vérifie l'analyse dimensionnelle). Maintenant que cette solution satisfait les conditions aux limites, nous devons nous attaquer aux conditions initiales.

D'abord pour les mêmes raisons que la corde, la solution finale est la superposition linéaire des solutions telle que:

$$y(r, \theta, t) = \sum_{n=0}^{\infty} \sum_{m=1}^{\infty} y_m^n(r, \theta, t) \quad (31.148)$$

Nous allons déterminer les coefficients $\alpha_m^n, \beta_m^n, \gamma_m^n, \delta_m^n$ de façon à ce que la solution y donnée précédemment vérifie également les conditions initiales, à savoir:

$$\begin{aligned} f(r, \theta, 0) &= \sum_{n=0}^{\infty} \sum_{m=1}^{\infty} J_n(\lambda_m^n r) (\alpha_m^n \cos(n\theta) + \gamma_m^n \sin(n\theta)) \\ \varphi(r, \theta, 0) &= \sum_{n=0}^{\infty} \sum_{m=1}^{\infty} \omega_m^n J_n(\lambda_m^n r) (\alpha_m^n \cos(n\theta) + \gamma_m^n \sin(n\theta)) \end{aligned} \quad (31.149)$$

Ces deux relations sont similaires, étudions la première. Elle peut s'écrire:

$$f(r, \theta, 0) = \sum_{n=0}^{\infty} \left\{ \left(\sum_{m=1}^{\infty} \alpha_m^n J_n(\lambda_m^n r) \right) \cos(n\theta) + \left(\sum_{m=1}^{\infty} \gamma_m^n J_n(\lambda_m^n r) \right) \sin(n\theta) \right\} \quad (31.150)$$

qui est le développement en série de Fourier de la fonction $f(r, \theta)$. Nous avons donc (cf. chapitre sur les Suites Et Séries):

$$\begin{aligned} \sum_{m=1}^{\infty} \alpha_m^0 J_n(\lambda_m^0 r) &= \frac{1}{2\pi} \int_{-\pi}^{+\pi} f(r, \theta) d\theta = f_0(r) \\ \sum_{m=1}^{\infty} \alpha_m^n J_n(\lambda_m^n r) &= \frac{1}{\pi} \int_{-\pi}^{+\pi} f(r, \theta) \cos(n\theta) d\theta = f_n(r), n \in \mathbb{N}^* \\ \sum_{m=1}^{\infty} \gamma_m^n J_n(\lambda_m^n r) &= \frac{1}{\pi} \int_{-\pi}^{+\pi} f(r, \theta) \sin(n\theta) d\theta = g_n(r), n \in \mathbb{N}^* \end{aligned} \quad (31.151)$$

En utilisant l'orthogonalité des fonctions de Bessel nous pouvons déduire de ces relations les coefficients $\alpha_m^0, \alpha_m^n, \gamma_m^n$ (et de même pour les autres).

Pour cela, supposons n fixé et posons $Z_m(r) = J_n(\lambda_m^n r)$. Montrons $m \neq p \Rightarrow Z_m \circ Z_p = 0$ où le produit scalaire est défini par:

$$Z_m \circ Z_p = \int_0^L r Z_m(r) Z_p(r) dr \quad (31.152)$$

Puisque Z_m, Z_p vérifient l'équation différentielle en $R(r)$, nous avons:

$$\begin{cases} r^2 Z_m''(r) + r Z_m'(r) + ((\lambda_m^n r)^2 - n^2) Z_m(r) = 0 \\ r^2 Z_p''(r) + r Z_p'(r) + ((\lambda_p^n r)^2 - n^2) Z_p(r) = 0 \end{cases} \quad (31.153)$$

En combinant ces deux relations nous obtenons:

$$((\lambda_p^n)^2 - (\lambda_m^n)^2) r Z_m Z_p = r (Z_m'' Z_p - Z_p'' Z_m) + (Z_m' Z_p - Z_p' Z_m) = (r (Z_m' Z_p - Z_p' Z_m))' \quad (31.154)$$

En intégrant membre à membre entre 0 et L et en tenant compte de:

$$Z_p(L) = J_n(\lambda_p^n L) = 0 \text{ et } Z_m(L) = J_n(\lambda_m^n L) = 0 \quad (31.155)$$

Nous obtenons:

$$((\lambda_p^n)^2 - (\lambda_m^n)^2) \int_0^L r Z_m(r) Z_p(r) dr = (r (Z_m'(r) Z_p(r) - Z_p'(r) Z_m(r))) \Big|_0^L = 0 \quad (31.156)$$

D'où le résultat énoncé puisque $m \neq p \Rightarrow \lambda_m^n \neq \lambda_p^n$.

La relation:

$$\sum_{m=1}^{\infty} \alpha_m^n J_n(\lambda_m^n r) = \frac{1}{\pi} \int_{-\pi}^{+\pi} f(r, \theta) \cos(n\theta) d\theta = f_n(r) \quad (31.157)$$

Peut donc s'écrire:

$$f_n = \sum_{m=1}^{\infty} \alpha_m^n Z_m \quad (31.158)$$

Utilisant l'orthogonalité de Z_m, Z_p pour $m \neq p$ nous en déduisons:

$$f_n \circ Z_p = \left(\sum_{m=1}^{\infty} \alpha_m^n Z_m \right) \circ Z_p = \sum_{m=1}^{\infty} \alpha_m^n (Z_m \circ Z_p) = \alpha_p^n (Z_m \circ Z_p) \quad (31.159)$$

Les coefficients α_p^n sont donc donnés par:

$$\alpha_p^n = \frac{f_n \circ Z_p}{Z_p \circ Z_p} = \frac{\int_0^L r f_n(r) J_n(\lambda_p^n r) dr}{\int_0^L r (J_n(\lambda_p^n r))^2 dr} \quad (31.160)$$

Ce qui n'est pas aisé à calculer à la main....

Nous procédons de la même façon pour les autres coefficients.

4. PHASEURS

Il existe plusieurs façons d'exprimer les fonctions d'ondes que nous avons vues précédemment. Les physiciens (ainsi que les électrotechniciens) utilisent une formulation, appelée "**phaseur**" ou "**représentation de Fresnel**", permettant d'économiser avantageusement le poids des écritures et ainsi de simplifier considérablement l'étude des problèmes complexes (ou simples). Les phaseurs font usage des propriétés des nombres complexes pour exprimer les fonctions d'onde trigonométriques sous une forme simplifiée dans tous les phénomènes où apparaissent des oscillations.

Ce que nous appelons "phaseur", est une fonction f dont la valeur est complexe et qui, dans un espace à 1 dimension, s'écrit:

$$f(x, t) = e^{i(kx - \omega t)} \quad (31.161)$$

Dans toutes les applications en physique, t est la variable du temps.

Comme cette fonction est complexe, elle a une partie réelle que nous appelons ici g et une partie imaginaire que nous appelons h . Leur identification est facile puisque comme nous l'avons déjà démontré lors de notre étude des nombres complexes (cf. chapitre sur les Nombres):

$$f(x, t) = \cos(kx - \omega t) + i \sin(kx - \omega t) \quad (31.162)$$

Ainsi, les parties réelles et imaginaires sont simplement:

$$\begin{aligned} g(x, t) &= \Re(f) = \cos(kx - \omega t) \\ h(x, t) &= \Im(f) = \sin(kx - \omega t) \end{aligned} \quad (31.163)$$

Le module de f se calcule aisément en calculant:

$$|f| = f \cdot \bar{f} = \cos^2(kx - \omega t) + \sin^2(kx - \omega t) = 1 \quad (31.164)$$

Les parties réelles et imaginaires varient lorsque la position ou le temps varient. Le module ne change donc pas, il est toujours égal à 1. Le changement se manifeste simplement par la simple variation d'angle que fait le vecteur représentant f dans son plan complexe. C'est là une raison suffisante pour parler de phaseur, puisque la variation de f peut être visualisée comme un simple changement d'angle ou de phase.

Si nous sommes dans un espace physique à plus d'une dimension, disons 3, alors l'expression pour f devient:

$$f(\vec{r}, t) = e^{i(\vec{k} \circ \vec{r} - \omega t)} = \cos(\vec{k} \circ \vec{r} - \omega t) + i \sin(\vec{k} \circ \vec{r} - \omega t) \quad (31.165)$$

La situation est un peu plus difficile à visualiser (...). Elle est la même qu'en une dimension, mais là tout se passe le long d'une direction définie dans l'espace 3D par le vecteur $\vec{k}$. Plus précisément, nous aurons:

$$\begin{aligned} \cos(\vec{k} \circ \vec{r} - \omega t) &= \cos(k_x x + k_y y + k_z z - \omega t) \\ \sin(\vec{k} \circ \vec{r} - \omega t) &= \sin(k_x x + k_y y + k_z z - \omega t) \end{aligned} \quad (31.166)$$

La variation dans le temps reste la même qu'en une dimension spatiale mais un déplacement dans l'espace est un peu plus compliqué qu'en une dimension. Ici tout déplacement spatial dans une direction qui n'est pas orthonormale à $\vec{k}$ fera que le produit scalaire changera de telle sorte que l'argument f variera. Ici ce n'est pas seulement la grandeur ou la norme du vecteur $\vec{k}$ qui donne le taux de variation de f sous un déplacement spatial mais aussi l'angle que fait ce déplacement par rapport à la direction de $\vec{k}$ puisque nous avons un argument qui varie comme $\vec{k} \circ \vec{r}$ et donc qui dépend de cet angle noté ici χ . En effet, nous avons:

$$\vec{k} \circ \vec{r} = \|\vec{k}\| \cdot \|\vec{r}\| \cos(\chi) \quad (31.167)$$

La quantité $\vec{k}$ est souvent appelée le "**vecteur d'onde**". Physiquement il est relié le plus souvent à l'équivalent du moment (linéaire) de l'onde pour laquelle il est évident que la définition usuelle de la quantité de mouvement ($p = mv$) n'a plus de sens.

Une partie importante des systèmes étudiés en physique ne peuvent être caractérisés par un point et donc décrits par une trajectoire. Une vague, une onde, une bande élastique qui oscillent n'ont pas une unique position définie, ce sont des "**milieux continus**" sur un certain intervalle. La question que nous nous posons dans notre tentative de les décrire est plutôt la suivante: comment décrire le déplacement de ce milieu dans l'espace et dans le temps. Par exemple, pour une vague, si nous figeons le temps, comment l'amplitude A , de cette vague varie-t-elle d'un endroit à l'autre de l'espace? Nous pouvons aussi figer l'espace en regardant un seul endroit et demander comment l'amplitude varie avec le temps?

Les coordonnées et le temps jouent maintenant un rôle similaire de paramètres indépendants. Nous mesurons l'amplitude du phénomène en tout temps et en tout lieu. Nous chercherons donc à obtenir une expression du type:

$$A = A(x, t) \text{ ou } A = A(\vec{r}, t) \quad (31.168)$$

en une ou trois dimensions.

Le point important est que nous cherchons à exprimer A , dont le nom correct est un "**champ**", comme une fonction des coordonnées du temps. Par exemple, si la vague est très régulière peut-être est-elle décrite adéquatement par $A \sim \cos(\omega t)$ si je regarde à un seul endroit et par $A \sim \cos(kx)$ lors d'une fixation imaginaire du temps ?

La quantité ω va caractériser la fréquence de la variation du même phénomène.

Le nom de "**fréquence angulaire**" est facile à comprendre puisque la fonction cosinus ou sinus fait un cycle si son argument change de 2π sur un temps d'une période $\Delta t = T$.

Rappelons que:

$$T\omega = 2\pi \Rightarrow \omega = \frac{2\pi}{T} = 2\pi\nu \quad (31.169)$$

Dans la description des systèmes harmoniques, la notation phaseur peut être très utile comme nous l'avons déjà dit. L'équation la plus souvent rencontrée est l'équation d'onde (que nous avons démontrée au début de ce chapitre). En une dimension, elle s'écrit:

$$\frac{\partial^2}{\partial x^2} A(x,t) - \frac{1}{v^2} \frac{\partial^2}{\partial t^2} A(x,t) = 0 \quad (31.170)$$

Nous vérifions par simple substitution que la solution est du type:

$$A(x,t) = A_0 \sin(kx - \omega t) \text{ ou } A(x,t) = A_0 \cos(kx - \omega t) \quad (31.171)$$

ou une combinaison linéaire dont la forme la plus générale est:

$$A(x,t) = A_0 \cos(kx - \omega t) + A_0 \sin(kx - \omega t) \quad (31.172)$$

Une manière rapide et efficace d'écrire toutes ces relations de façon condensée et utilement imagée, est d'écrire:

$$A(x,t) = A_0 e^{i(kx - \omega t)} \quad (31.173)$$

Dans les trois cas, la substitution permet de le vérifier. Prenons, par exemple, la solution sinus. Alors:

$$\begin{aligned} A(x,t) &= A_0 \sin(kx - \omega t) \\ \Rightarrow \frac{\partial A}{\partial x} &= k A_0 \cos(kx - \omega t) \Rightarrow \frac{\partial^2 A}{\partial x^2} = -k^2 A_0 \sin(kx - \omega t) \\ \Rightarrow \frac{\partial A}{\partial t} &= -\omega A_0 \cos(kx - \omega t) \Rightarrow \frac{\partial^2 A}{\partial t^2} = -\omega^2 A_0 \sin(kx - \omega t) \end{aligned} \quad (31.174)$$

Le remplacement dans l'équation donne:

$$\left(-k^2 + \frac{\omega^2}{v^2} \right) A_0 \sin(kx - \omega t) = 0 \quad (31.175)$$

qui sera satisfait si et seulement si:

$$k^2 = \frac{\omega^2}{v^2} \Leftrightarrow \omega^2 = k^2 v^2 \quad (31.176)$$

La substitution du phaseur comme solution de l'équation d'onde transforme cette dernière équation différentielle en une simple équation algébrique $\omega = \omega(k)$ appelée "**relation de dispersion**".

Elle est évidemment caractéristique de l'équation qui la génère. Celle qui apparaît ci-dessus est particulièrement simple et caractérise une onde libre dans un milieu non-dispersif, tel que décrit par l'équation d'onde que nous avons écrite.

La solution phaseur:

$$A(x,t) = A_0 e^{i(kx - \omega t)} \quad (31.177)$$

satisfait donc aussi l'équation, avec la même relation de dispersion. Est-elle donc aussi la description d'une onde? La réponse est oui, et même deux fois plutôt qu'une, comme nous allons le voir ci-dessous.

Une onde physique n'est évidemment pas complexe. La solution phaseur est complexe et a donc une

partie réelle et une partie imaginaire. Nous montrons ici que chacune des deux parties peut représenter une onde réelle générale. Utilisons même un point de départ un peu plus général. Imaginons que l'amplitude A_0 est elle-même complexe. Nous pouvons donc l'écrire:

$$A_0 = |A_0| e^{i\delta} \quad (31.178)$$

Nous avons donc:

$$A(x,t) = A_0 e^{i(kx - \omega t + \delta)} = A_0 (\cos(kx - \omega t + \delta) + i \sin(kx - \omega t + \delta)) \quad (31.179)$$

Nous étudions d'abord la partie réelle de cette expression:

$$\begin{aligned} \Re(A(x,t)) &= A_0 \cos(kx - \omega t + \delta) = A_0 [\cos(kx - \omega t) \cos(\delta) - \sin(kx - \omega t) \sin \delta] \\ &= A_0 \cos \delta \cos(kx - \omega t) - A_0 \sin \delta \sin(kx - \omega t) \end{aligned} \quad (31.180)$$

Cette partie réelle est donc de la forme la plus générale (et réelle) de la solution monochromatique pour l'équation d'onde, soit:

$$A(x,t) = A_1 \cos(kx - \omega t) + A_2 \sin(kx - \omega t) \quad (31.181)$$

Clairement, il suffit d'identifier:

$$\left. \begin{aligned} A_0 \cos \delta &= A_1 \\ -A_0 \sin \delta &= A_2 \end{aligned} \right\} \Rightarrow \begin{cases} A_0 = \sqrt{A_1^2 + A_2^2} \\ \delta = -\arctan(A_2 / A_1) \end{cases} \quad (31.182)$$

qui relie deux paramètres à deux autres. La partie réelle du phaseur est donc suffisante pour décrire entièrement l'onde monochromatique.

Nous pouvons refaire exactement la même chose avec la partie imaginaire du phaseur et démontrer de façon identique qu'elle est suffisante pour décrire entièrement l'onde monochromatique.

Conclusion: il est donc possible d'utiliser le phaseur pour faire toutes les manipulations mathématiques demandées par le problème physique et à la fin, ne garder que la partie réelle ou imaginaire, selon ce qui a été convenu dès le début.

Comme nous l'avons déjà dit, la forme réelle la plus générale de la solution est:

$$A(x,t) = A_1 \sin(kx - \omega t) + A_2 \cos(kx - \omega t) \quad (31.183)$$

Cette fonction a le comportement d'un sinus (ou d'un cosinus) dont l'amplitude est donnée par:

$$A_0^2 = A_1^2 + A_2^2 \quad (31.184)$$

De plus les conditions initiales ajustent A_2 de telle sorte qu'à $x = 0$ et $t = 0$ (initialement), le champ $A(x,t)$ a la valeur A_2 donc. Ces deux conditions fixent ces deux paramètres. Nous aurions pu utiliser la forme toute aussi générale:

$$A(x,t) = A_0 \sin(kx - \omega t + \delta) \quad (31.185)$$

Ici l'amplitude connue est A_0 et les conditions initiales sont telles qu'à $x = 0$ et $t = 0$, le champ a la valeur:

$$A(x, t) = A_0 \sin(\vartheta) \quad (31.186)$$

Encore une fois, deux conditions fixent deux paramètres.

L'amplitude est une chose évidente, la phase un peu moins. Pour qu'une fonction du type:

$$A(x, t) = A_0 \sin(\arg + \vartheta) \quad (31.187)$$

ait n'importe quelle valeur que l'on veut lorsque son $\arg$ (argument) est, disons nul, il suffit d'ajuster la valeur de la phase ϑ . C'est comme faire glisser la fonction sinusoïdale le long de l'axe, de façon à satisfaire des conditions initiales physiques imposées par le système étudié.

De deux fonctions de type sinus ou cosinus qui ne commencent pas au même point de leur cycle, nous disons qu'elles sont "**déphasées**". Ceci devient vital lorsqu'il y a plus d'une onde en présence. Imaginons le cas le plus simple de deux ondes de même amplitude:

$$\begin{aligned} A_1(x, t) &= A_0 \sin(\vartheta + \delta_1) \\ A_2(x, t) &= A_0 \sin(\vartheta + \delta_2) \end{aligned} \quad (31.188)$$

Ici l'argument est une variable et les phases des paramètres constants. Nous considérons le résultat physique de l'onde résultant de l'addition de ces deux ondes.

Si $\delta_2 - \delta_1 = 2k\pi, k \in \mathbb{Z}$ l'onde résultante sera une onde sinusoïdale d'amplitude 2 fois A_0 . Cependant, si $\delta_2 - \delta_1 = k\pi, k \in \mathbb{Z}$, l'onde résultante sera identiquement nulle partout. La différence est donc considérable et nous trouvons toutes les situations intermédiaires. Il est donc important de garder en tête la phase à l'origine de l'onde ou mieux, sa phase relative par rapport à d'autres ondes de notre système physique.

Dans le phaseur, soit la partie réelle, soit la partie imaginaire, est suffisante pour donner une description générale de l'onde (toujours monochromatique jusqu'à maintenant). Elles sont respectivement composées d'un cosinus et d'un sinus et donc déphasées de $\pi/2$ l'une par rapport à l'autre !

Il est souhaitable de revenir sur la "**relation de dispersion**" que nous avons obtenue. Nous avons vu que, pour l'onde monochromatique libre, celle qui est solution de l'équation homogène, cette relation s'écrit pour que la phase de la solution corresponde à cette réalité (attention il s'agit de ne pas confondre le symbole de la vitesse et de la fréquence!):

$$\omega^2 = k^2 v^2 \Rightarrow \omega = \omega(k) \Leftrightarrow k = k(\omega) \quad (31.189)$$

Toujours dans le cas libre, nous pouvons avoir une situation physique qui correspond à l'addition de plusieurs ondes monochromatiques libres. Le résultat n'est pas monochromatique et s'écrit évidemment comme une somme:

$$A(x, t) = \sum_n A_n(x, t) = \sum_n A_n e^{i(k_n x - \omega_n t + \delta_n)} \quad (31.190)$$

Nous lui donnons souvent le nom de "**paquet d'ondes**" pour des raisons évidentes. Puisque tout est libre, chaque composante satisfera:

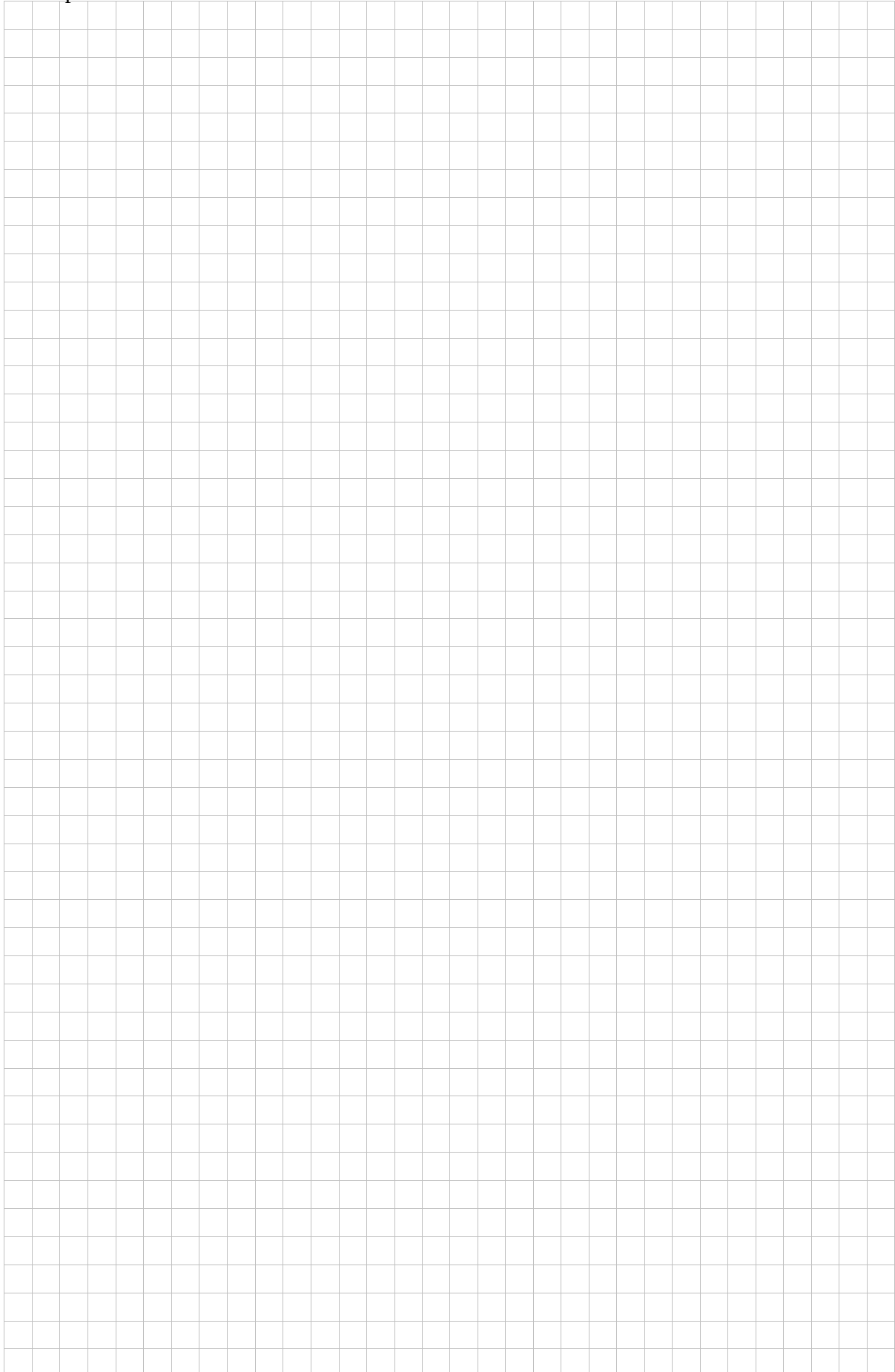
$$\omega_x^2 = k_x^2 v_x^2 \quad (31.191)$$

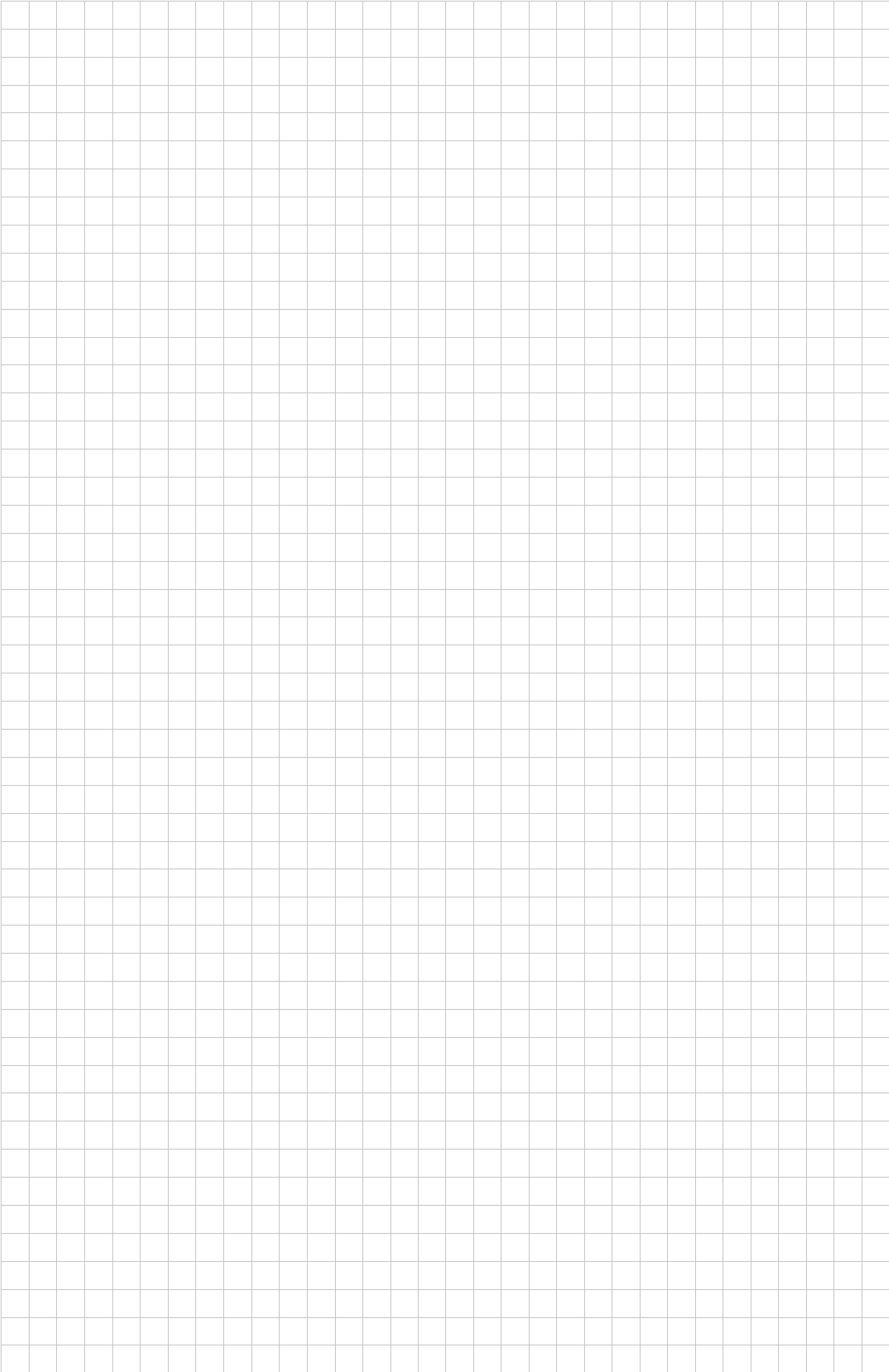
Nous noterons ici deux choses pour l'équation des ondes libres:

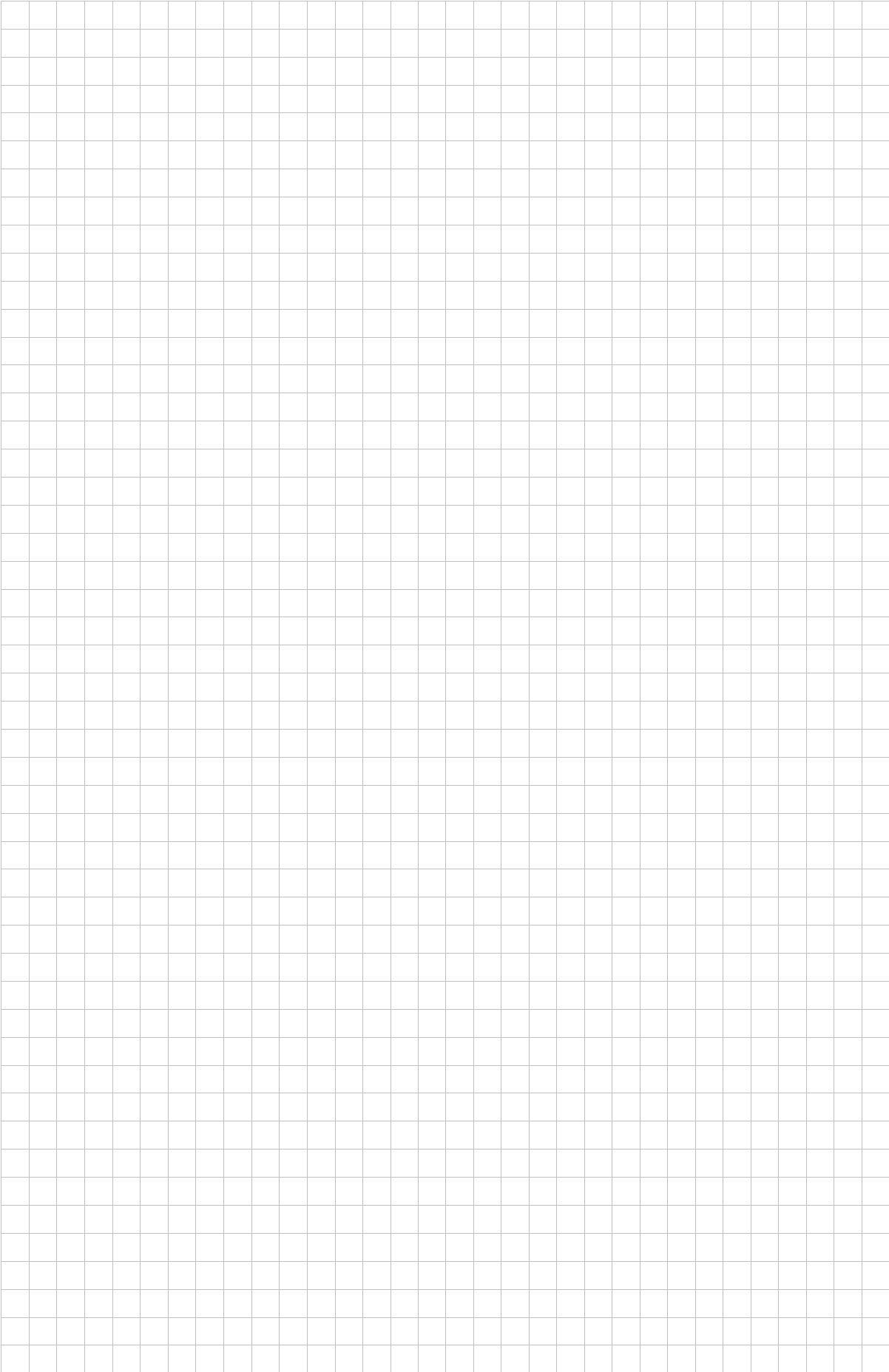
- D'abord, même pour des ondes libres, elle est quadratique, nous pouvons donc changer le signe du vecteur d'onde et/ou de la pulsation sans affecter l'équation. Nous observons trivialement que pour une ou plusieurs composantes du vecteur d'onde positives l'onde se propage vers les x croissants pour ces composantes positives et qu'inversement, pour une ou plusieurs composantes négatives l'onde se propage vers les x décroissants. De même, une pulsation négative ou positive signifie que le temps varie vers le passé, respectivement le futur.

- D'autre part, certains types d'ondes n'obéissent pas à une équation aussi simple que l'équation homogène. C'est le cas des vagues, par exemple, tant du fait de la nature du liquide dans lequel elles se propagent que de la force de rappel gravitationnel. Parfois aussi, une onde qui serait totalement libre, ou à peu près, cherche à se propager dans un milieu où les conditions de propagation sont sérieusement affectées. Par exemple une onde sonore qui tente de se propager dans le mastic ou une onde électromagnétique qui cherche à se propager dans un conducteur (un métal). Dans ce cas, une partie importante de la différence entre onde libre et onde modifiée par le milieu peut se décrire par un changement de la dispersion.

Notes personnelles:









32. MÉCANIQUE STATISTIQUE

La mécanique statistique, appelée aussi "**thermodynamique statistique**" ou encore plus généralement "**physique statistique**", a pour but d'expliquer le comportement des systèmes macroscopiques (constitués d'un grand nombre d'objets en interaction) à partir de leurs caractéristiques microscopiques. C'est de façon beaucoup plus générale, la physique quantique qui décrit les propriétés et l'évolution des systèmes physiques à l'échelle microscopique. La mécanique statistique est donc construite sur cette description quantique comme nous le verrons sur les développements mathématiques qui suivront.

La démarche présentée ici est d'aborder la mécanique statistique élémentaire pour en déduire ensuite la thermodynamique. La mécanique statistique constitue en effet, avec la physique quantique et la relativité, l'un des piliers de la physique moderne dans l'explication de phénomènes à partir de leurs constituants. Il est important de la percevoir d'emblée comme une théorie fondamentale, et non pas comme une simple tentative pour justifier à posteriori la thermodynamique. La thermodynamique elle-même y gagne en retour compréhension plus juste et plus profonde de ses principes et de ses méthodes.

1. THÉORIE STATISTIQUE DE L'INFORMATION

Le mot "**information**" est utilisé dans des contextes très variés, dans des sens totalement différents suivants les disciplines scientifiques: nous pouvons à titre d'exemple citer la thermodynamique avec le concept d'entropie, la physique appliquée avec la théorie du signal, la biologie avec la théorie du génome et la physique quantique avec la probabilité d'obtenir de l'information.

Se pose alors la question, s'il est possible de construire une théorie de l'information et si elle est unique? Notre démarche ici, ne vise pas l'information en tant que telle, mais la quantité d'information. Lorsque nous parlons de quantité et de mesure, nous pensons à la notion de contenu ou de valeur de l'information. La science de l'information de par son objet doit se sentir concernée par ce questionnement. Si nous définissons "**l'infométrie**" comme l'ensemble des techniques de mesures mathématiques et statistiques de l'information, nous souhaiterions avoir une définition suffisamment claire du concept de quantité d'informations qui puisse nous amener à définir une mesure, c'est-à-dire un ensemble d'opérations parfaitement définies, nous amenant à des axiomes clairs et dont le résultat est un nombre. La synthèse que nous développons ici n'est pas ambitieuse.

Nous nous intéressons donc ici aux fondements de la théorie statistique de l'information connue également sous le nom de "**théorie de Shannon**". La formule de Shannon qui en ressort est certainement l'un des concepts fondamentaux de toute la physique puisqu'elle touche la brique irréductible de la physique: l'information !!

Nous montrerons (plus bas) qu'un système physique isolé a pour état le plus probable, celui qui contient le plus d'états et qui est donc à fortiori le plus imprévisible. L'état le plus improbable est donc celui qui est le plus prévisible. Dès lors, puisque l'imprévisibilité apparaît comme un attribut essentiel de l'information, nous identifions la mesure quantitative de l'information à son improbabilité.

Ainsi, la quantité d'information $h(x)$ apportée par la réalisation d'un événement x de probabilité $p(x)$ sera une fonction croissante f de son improbabilité $1/p(x)$:

$$h(x) = f\left(\frac{1}{p(x)}\right) \quad (32.1)$$

De plus, la réalisation de deux événements indépendants x et y apporte intuitivement une quantité d'information qui est la somme de leurs quantités d'informations respectives, soit:

$$h(x, y) = f\left(\frac{1}{p(x, y)}\right) = f\left(\frac{1}{p(x)p(y)}\right) = f\left(\frac{1}{p(x)}\right) + f\left(\frac{1}{p(y)}\right) = h(x) + h(y) \quad (32.2)$$

La fonction logarithme est donc par ses propriétés une candidate naturelle pour f telle que:

$$f = \lambda \log_{\alpha}\left(\frac{1}{p(\dots)}\right) \quad (32.3)$$

où λ est bien évidemment un nombre positif.

Remarque: Le choix de la base du logarithme définit l'unité d'information qui est complètement arbitraire. Par la suite, et sauf précision contraire, "log" désignera le logarithme en base α .

Ainsi, la "**quantité d'information intrinsèque**" d'un événement x est donc de par les propriétés du logarithme:

$$h(x) = -\lambda \log(p(x)) \quad (32.4)$$

Elle peut être considérée, comme nous l'avons fait, comme une mesure d'incertitude sur l'événement, ou comme celle de l'information nécessaire pour résoudre cette incertitude. Voyons un exemple qui motive cette définition.

Exemple:

Convenons que la connaissance de la réalisation d'un événement qui avait pour probabilité $1/2$ fournit 1 bit d'information. Lançons maintenant n pièces de monnaie et prenons connaissance du résultat de l'expérience : nous acquérons n bits d'information. Or l'événement en question avait pour probabilité $1/2^n$. Donc le nombre de bits d'information est ici:

$$-\log_2(p) \quad (32.5)$$

où p est la probabilité de l'événement. Ce qui motive la définition!

Définitions:

D1. Nous définissons "**l'information intrinsèque par paire**" de deux événements x et y de probabilité conjointe $p(x, y)$ (cf. chapitre de Probabilités) par:

$$h(x, y) = -\lambda \log(p(x, y)) \quad (32.6)$$

D2. Nous définissons de même "**l'information conditionnelle**" de x sachant y par (cf. chapitre de

Probabilités):

$$h(x/y) = -\lambda \log(p(x/y)) \quad (32.7)$$

qui peut donc aussi s'écrire (ce n'est qu'une manière de noter différente que nous avons déjà présentée dans le chapitre de Probabilités):

$$h(x|y) = -\lambda \log(p(x|y)) \quad (32.8)$$

Il s'agit de la quantité d'information restant sur x après l'observation de y . La formule de Bayes (cf. chapitre de Probabilités) nous permet de remarquer immédiatement que si x et y sont indépendants:

$$h(x/y) = h(x) \quad (32.9)$$

ce qui concorde avec le sens commun.

Nous souhaitons aussi mesurer la quantité d'information que la donnée d'une variable, par exemple y , apporte sur l'autre, x . C'est le cas en particulier lorsque nous identifions x au choix d'un signal appliqué à l'entrée d'un canal et y au signal correspondant observé en sa sortie. $p(x)$ est alors la probabilité à fortiori que x soit émis et $p(x/y)$ la probabilité à fortiori que x ait été émis, sachant que y a été reçu.

Une mesure de cette quantité d'information, nommée "**information mutuelle**" est:

$$i(x,y) = \lambda \log \left(\frac{p(x/y)}{p(x)} \right) \quad (32.10)$$

il s'agit de la mesure logarithmique de l'accroissement de la probabilité de x (donc de la baisse de sa quantité d'information) dû à son conditionnement sur y . Si la donnée de y est équivalente à celle de x (cas d'un canal parfait), elle est égale à l'information intrinsèque $h(x)$. Elle est nulle si, à l'inverse, x et y sont indépendants.

Nous avons bien évidemment:

$$i(x,y) = h(x) - h(x/y) \quad (32.11)$$

et de par les propriétés des logarithmes:

$$i(x,y) = \lambda \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \quad (32.12)$$

cette dernière égalité justifiant le terme "mutuelle". Alors que les informations intrinsèques étaient positives, l'information mutuelle peut être négative. Nous verrons que sa moyenne, beaucoup plus importante dans la pratique, ne peut l'être.

Les événements individuels étant généralement moins importants que les moyennes, nous les considérerons par la suite comme provenant d'une source aléatoire, discrète, finie, stationnaire, et blanche (i.e. de réalisations successives indépendantes). Les événements sont donc interprétés comme le choix d'un symbole dans l'alphabet de la source. Soit n la taille de cet alphabet, et $x_1, \dots, x_n$ ses symboles. La source est donc décrite par la variable aléatoire x , qui prend ses valeurs dans l'alphabet,

avec des probabilités respectives $p_1, \dots, p_n$, telles que:

$$\sum_{i=1}^n p_i = 1 \quad (32.13)$$

La quantité d'information moyenne de cette source est l'espérance de l'information intrinsèque de chaque symbole de l'alphabet de la source (cf. chapitre de Statistiques). Elle est appelée "**entropie**" (par la notation $S(X)$) de X et est donnée par la relation:

$$S(X) = E\{h(x)\} = -\lambda \sum_{i=1}^n p_i \log(p_i) \quad (32.14)$$

appelée "**formule de Shannon**" et il s'agit bien d'une espérance car son expression générale pour une variable aléatoire discrète est comme nous l'avons vu dans le chapitre de Statistiques donnée par:

$$E(X) = \sum_{i=1}^n p_i x_i \quad (32.15)$$

avec dans le cas présent:

$$x_i = -\lambda \log(p_i) \quad (32.16)$$

Cette dernière écriture constitue cependant un abus de notation: en effet, l'espérance mathématique a un sens si $h(x)$ est une fonction de x . Or $h(x)$ ne dépend pas des valeurs de x , mais seulement des probabilités associées. Nous noterons parfois plus rigoureusement l'entropie d'une distribution:

$$S(p_1, \dots, p_n) \quad (32.17)$$

Les "**entropies conjointes et conditionnelles**" sont définies de manières similaires avec les notations idoines:

$$S(X, Y) = E\{h(x, y)\} = -\lambda \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log(p(x_i, y_j)) \quad (32.18)$$

et (nous adoptons cette fois-ci le symbole "|" et non plus le symbole "/" pour indiquer la relation conditionnelle):

$$S(Y|X) = E\{h(y|x)\} = -\lambda \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log(p(y_j|x_i)) \quad (32.19)$$

Il faut noter dans la dernière expression que l'espérance est effectuée dans l'espace produit, et que donc le coefficient est la probabilité conjointe.

"**L'information mutuelle moyenne**", appelée par abus de langage "information mutuelle" se définit elle aussi de manière directe:

$$\begin{aligned}
 I(X, Y) &= E\{i(x, y)\} = \lambda \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log \left(\frac{p(x_i | y_j)}{p(x_i)} \right) \\
 &= \lambda \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log \left(\frac{p(x_i, y_j)}{p(x_i) p(y_j)} \right)
 \end{aligned} \quad (32.20)$$

Remarque: Il est à noter que la définition de la quantité d'information, par une mesure logarithmique peut paraître arbitraire, quoique raisonnable, compte tenu des propriétés attendues d'une telle mesure. Shannon, et plus tard Khintchine ont montré que compte tenues de certaines propriétés posées en axiomes, la fonction logarithmique est la seule à convenir.

Exemple:

Soit une variable aléatoire binaire, valant 1 avec une probabilité p (et donc 0 avec une probabilité $1-p$). Son entropie vaut:

$$\begin{aligned}
 S_2(p) &= S_2(1-p) \Leftrightarrow S_2(p) - S_2(1-p) = 0 \\
 \Rightarrow -p \lambda \log(p) - (1-p) \lambda \log(1-p) &= 0
 \end{aligned} \quad (32.21)$$

avec $0 < p < 1$ et avec un logarithme en base 2 tel que pour un événement à deux états équiprobables, l'entropie d'obtention d'un des deux états soit égale à l'unité. Ceci dit, il vient naturellement que $\lambda = 1$.

Elle est représentée à la figure ci-dessous, en "Shannon" (unité correspondant à l'utilisation du logarithme à base 2). Nous remarquerons sa symétrie par rapport à $\frac{1}{2}$, valeur pour laquelle elle atteint son maximum, égal à 1.

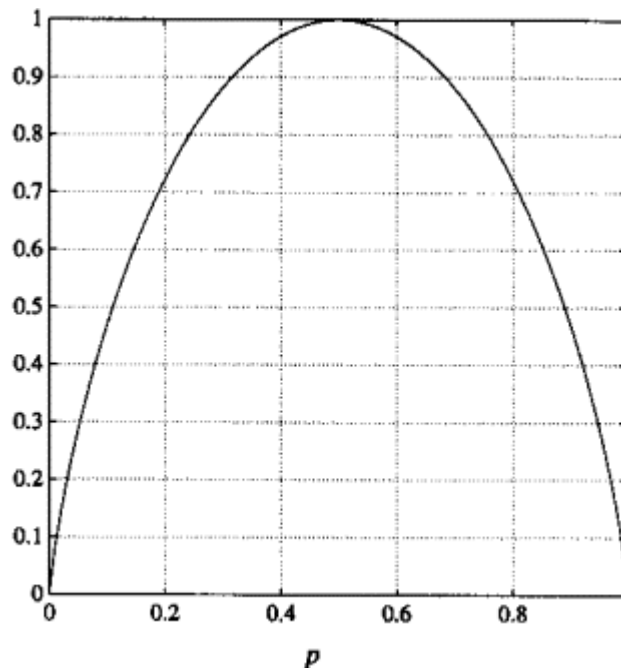


Figure: 32.1 - Entropie d'une variable binaire

Il convient maintenant de faire la liaison entre la théorie statistique de l'information et la mécanique statistique:

2. LOI DE BOLTZMANN

Nous allons d'abord démontrer par l'intermédiaire d'un cas simple, que pour tout système, l'état le plus probable est l'état d'équilibre !

Considérons un système isolé (un système est dit "isolé" lorsqu'il est imperméable à tout flux - chaleur (adiabatique), matière, champs, ...) peuplé de N particules discernables. Ce système est partagé en deux compartiments (ou niveaux) identiques et séparés d'une paroi imperméable. Chaque compartiment est supposé contenir un nombre N_1, N_2 de particules.

Pour une configuration donnée du système, nous parlons de "macro-état" dans le sens où il est possible de par la quantité de particules de mesurer une grandeur dite macroscopique telle que l'énergie, la masse, la pression, etc.

Si nous fixons ce système particulier, il est bien sûr possible pour un nombre N de particules de concevoir un nombre donné de macro-états. Tel que:

- 1 particule: 2 macro-états (2 configurations possibles soit 1 configuration par macro-état)
- 2 particules: 3 macro-états (4 configurations possibles par permutations des compartiments)
- 3 particules: 4 macro-états (8 configurations possibles par permutations des compartiments)
- 4 particules: 5 macro-états (16 configurations possibles par permutations des compartiments)

etc.

Définition: Nous appelons "**micro-état**", une configuration de permutations du macro-état.

Remarque: Parfois au lieu de "micro-état" nous trouvons dans la littérature "**probabilité thermodynamique**" ou "**complexions**".

Déterminons maintenant à l'aide de l'analyse combinatoire (cf. chapitre de Probabilités) le nombre de micro-états Ω possibles pour chaque macro-état. Par analogie, ceci correspond à s'imaginer que le système est une tige sur laquelle sont enfilées des boules (particules) et que la tige est séparée par une frontière imaginaire en un de ses points (boulier chinois). Pour une telle situation, nous avons:

$$\Omega = \frac{N!}{N_1! N_2!} \quad (32.22)$$

Ceci nous donne tous les arrangements possibles des "**particules gauches**" avec les "**particules droites**" (de la frontière) pour un macro-état donné (le nombre de manières dont les particules peuvent se partager entre les deux compartiments). Mais nous avons aussi dans ce cas particulier:

$$\Omega = \frac{N!}{N_1! N_2!} = \frac{N!}{N_1! (N - N_1)!} = \frac{N!}{(N - N_2)! N_2!} \quad (32.23)$$

Or cela correspond à la combinatoire tel que:

$$C_{N1}^N = \frac{N!}{N_1!(N-N_1)!}, C_{N2}^N = \frac{N!}{(N-N_2)!N_2!} \quad (32.24)$$

et donc:

$$\Omega = \frac{N!}{N_1!N_2!} = C_{N1}^N = C_{N2}^N \quad (32.25)$$

Nous avons finalement pour tous les macro-états d'un système de N particules, un total de:

$$\sum_{N1=0}^N C_{N1}^N = \sum_{N2=0}^N C_{N2}^N \quad (32.26)$$

micro-états (configurations) possibles. Or, nous avons bien vu dans l'exemple initial que:

$$\sum_{N1=0}^N C_{N1}^N = \sum_{N2=0}^N C_{N2}^N = 2^N \quad (32.27)$$

Ainsi, la probabilité d'existence d'un micro-état donné est de 2^{-N} et elle est équiprobable !!

Nous pouvons maintenant énoncer le premier postulat de la mécanique statistique (postulat de Gibbs): tous les micro-états discernables et accessibles d'un système isolé sont équiprobables.

Revenons-en maintenant à notre question initiale sur l'équilibre:

La notion d'équilibre associée à un macro-état nous est fournie par la thermodynamique classique. Nous y voyons qu'un système est dit à l'équilibre lorsque son état est caractérisé par l'indépendance temporelle des grandeurs macroscopiques (masse, énergie, pression, ...) et de la constance des potentiels thermodynamiques (énergie interne, enthalpie, énergie de Gibbs, ...).

Pour savoir pourquoi l'équilibre est l'état le plus probable, il nous suffit de chercher quel est le couple (N_1, N_2) qui maximise:

$$\Omega_{\max} = \Omega = \frac{N!}{N_1!N_2!} \quad (32.28)$$

puisque tous les micro-états sont de toute façon équiprobables. Il est facile de contrôler que ce maximum est donné pour:

$$N_1 = N_2 \quad (32.29)$$

Nous pouvons dès lors énoncer le deuxième postulat de la mécanique statistique: l'état d'équilibre est l'état qui correspond au plus grand nombre de configurations (micro-états) et est l'état le plus probable!!

Ou en d'autres termes: Un système atteint l'équilibre lorsque son entropie devient maximale!!

Soit maintenant à considérer le système suivant:

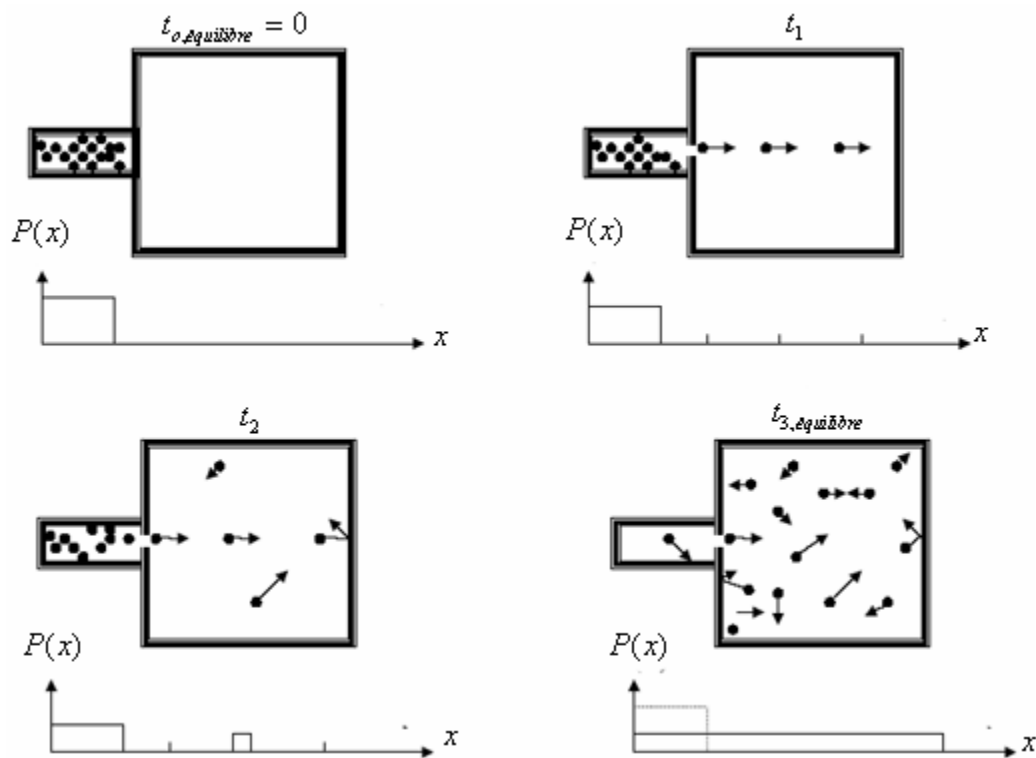


Figure: 32.2 - Exemple d'évolution de l'entropie d'un système

La fonction de distribution $P(x)$ qui décrit la position des particules selon l'axe x à l'équilibre ($t = 0$) va évoluer vers une autre fonction de distribution correspondant au nouvel équilibre t_3 . À l'équilibre $P(x)$ est constante. Mais entre les deux équilibres, elle évolue, et devient de plus en plus large. Nous perdons donc de l'information sur la position des particules. Nous pouvons donc ré-énoncer le deuxième postulat en disant qu'un système hors d'équilibre évolue toujours dans le sens d'une perte d'informations (d'un élargissement de la fonction de distribution caractéristique).

Parallèlement, le deuxième principe de la thermodynamique classique nous indique que toute évolution naturelle doit nécessairement correspondre à un accroissement d'entropie ($dS > 0$). Il doit donc exister un lien étroit entre l'information que nous possédons sur l'état de chacune des particules et l'entropie du système.

Le cas que nous venons de décrire montre clairement que les paramètres ou concepts: nombre de configurations, désordre, équilibre, quantité d'information et entropie d'un système isolé servent à représenter l'état d'un système. Ces paramètres jouent le même rôle. Des relations mathématiques doivent donc les relier les uns aux autres.

Rappelons que nous avons démontré que l'entropie statistique infométrique d'un système est donnée par:

$$S(X) = E\{h(x)\} = -\lambda \sum_{i=1}^n p_i \log(p_i) \quad (32.30)$$

Si nous appliquons cette relation au cas d'un système physique en équilibre pour lequel nous souhaitons calculer l'entropie, nous avons démontré:

$$p_1, p_2, \dots, p_n = c^{\frac{1}{n}} \quad (32.31)$$

Il nous faut encore savoir à quoi correspond cette probabilité constante. Nous avons démontré précédemment qu'à l'équilibre, nous avons:

$$\Omega = \Omega_{\text{max}} \quad (32.32)$$

qui est donc le nombre de micro-états à l'équilibre. Ainsi, la probabilité de tirer un micro-état parmi tous est de:

$$p = \frac{1}{\Omega_{\text{max}}} \quad (32.33)$$

que nous notons simplement (un peu dangereusement...) par tradition :

$$p = \frac{1}{\Omega} \quad (32.34)$$

Nous avons ainsi:

$$\begin{aligned} S &= -\lambda \sum_{i=1}^n p_i \log(p_i) = -\lambda \left(\frac{1}{\Omega} \log\left(\frac{1}{\Omega}\right) + \frac{1}{\Omega} \log\left(\frac{1}{\Omega}\right) + \dots + \frac{1}{\Omega} \log\left(\frac{1}{\Omega}\right) \right) \\ &= -\lambda \left(-\frac{1}{\Omega} \log(\Omega) - \frac{1}{\Omega} \log(\Omega) - \dots - \frac{1}{\Omega} \log(\Omega) \right) = \lambda \log(\Omega) \left(\frac{1}{\Omega} + \frac{1}{\Omega} + \dots + \frac{1}{\Omega} \right) \end{aligned} \quad (32.35)$$

Comme les probabilités des micro-états sont équiprobables et que nous sommes sur l'ensemble de ces derniers, il vient:

$$\left(\frac{1}{\Omega} + \frac{1}{\Omega} + \dots + \frac{1}{\Omega} \right) = 1 \quad (32.36)$$

et donc (sans oublier que dans ce cas particulier les probabilités ont toutes la même valeur!):

$$S = -\lambda \sum_{i=1}^n p_i \log(p_i) = \lambda \log(\Omega) = c^{\text{te}} \log(\Omega) \quad (32.37)$$

Puisque l'équilibre est lié au désordre maximum, et que le désordre est lié à l'information manquante, il paraît raisonnable de relier l'entropie statistique de l'information à l'entropie statistique thermodynamique en physique. Pour cela, il faut que la constante nous permette d'obtenir les bonnes unités et il vient naturellement de choisir cette constante telle qu'elle soit égale à la constante de Boltzmann k qui a les mêmes unités que l'entropie thermodynamique. Ainsi:

$$S = k \log(\Omega) \quad (32.38)$$

Il nous faut encore choisir la base du logarithme. L'expérience montre qu'il faut choisir le logarithme népérien qui permet de retrouver des résultats de la mécanique classique après développements.

Ainsi, nous obtenons finalement la "**loi de Boltzmann**":

$$S = k \ln(\Omega) \quad (32.39)$$

qui nous donne l'entropie thermodynamique d'un système à l'équilibre!

De par les propriétés mathématiques de l'espérance (cf. chapitre de Statistiques), en particulier la multiplication par une constante, nous avons pour un ensemble N de sous-systèmes:

$$S(N \cdot \Omega) = N \cdot S(\Omega) = Nk \ln(\Omega) = -kN \sum_{i=1}^N p_i \log(p_i) \quad (32.40)$$

puisque comme nous l'avons signalé plus haut lors de notre présentation de la formule de Shannon, l'entropie S est une espérance (sans oublier que dans ce cas particulier les probabilités ont toutes la même valeur!).

3. DISTRIBUTIONS STATISTIQUES PHYSIQUES

Nous distinguerons quatre différentes statistiques qui proviennent ou non d'effets quantiques et qui conduiront à quatre distributions distinctes connues. Ce sont les distributions de Maxwell, Maxwell-Boltzmann, Fermi-Dirac et de Bose-Einstein. Elles trouvent de nombreuses applications en physique comme le rayonnement du corps noir qui sera démontré dans le chapitre de Thermodynamique.

3.1. DISTRIBUTION DE MAXWELL (DISTRIBUTION DES VITESSES)

Pour un gaz en état d'équilibre, posons-nous la question suivante: Quelle est la probabilité qu'une molécule ait ses composantes de vitesse comprises entre v_x et $v_x + dv_x$, v_y et $v_y + dv_y$, v_z et $v_z + dv_z$ dans un repère cartésien habituel ?

Cette probabilité d^3P dépend de $\vec{v}$ (c'est-à-dire: (v_x, v_y, v_z)) et de dv_x, dv_y, dv_z . Elle ne dépend pas de la position de la molécule puisque le gaz est supposé à l'équilibre par rapport à son centre de masse.

Nous postulons dans un premier temps que d^3P est proportionnelle à chacun des intervalles dv_x, dv_y, dv_z tel que:

$$d^3P = f(\vec{v}) dv_x dv_y dv_z \quad (32.41)$$

et qu'il n'y a pas de directions privilégiées. Nous pouvons faire une rotation circulaire des axes cartésiens, la probabilité sera inchangée (isotropie de l'espace).

L'isotropie amène que la fonction $f(\vec{v})$ ne dépend pas du vecteur $\vec{v}$, au mieux elle dépend de la norme de cette vitesse, tel que:

$$d^3P = f(v) dv_x dv_y dv_z \quad (32.42)$$

Soit dP_x (noté aussi souvent $dP(v_x)$ en fonction du contexte) la probabilité pour une molécule d'avoir sa composante suivant l'axe Ox comprise entre v_x et $v_x + dv_x$ alors:

$$dP_x = dP(v_x) = \varphi_x(v_x) dv_x \quad (32.43)$$

de même:

$$dP_y = \varphi_y(v_y)dv_y \quad dP_z = \varphi_z(v_z)dv_z \quad (32.44)$$

L'isotropie de l'espace impose:

$$\varphi(v_x) = \varphi(v_y) = \varphi(v_z) = \varphi \quad (32.45)$$

La loi des probabilités composées implique:

$$d^3P = dP_x dP_y dP_z \Rightarrow f(v) = \varphi(v_x)\varphi(v_y)\varphi(v_z) \quad (32.46)$$

Déterminer les fonctions f et φ relève de la méthodologie mathématique suivante:

$$\begin{aligned} \frac{\partial f}{\partial v_x} &= \frac{\partial f}{\partial v} \frac{\partial v}{\partial v_x} = \frac{df}{dv} \frac{\partial v}{\partial v_x} = \frac{\partial}{\partial v_x} [\varphi(v_x)\varphi(v_y)\varphi(v_z)] = \varphi(v_y)\varphi(v_z) \frac{d\varphi(v_x)}{dv_x} \\ &= \frac{f}{\varphi(v_x)} \frac{d\varphi(v_x)}{dv_x} \end{aligned} \quad (32.47)$$

Puisque:

$$v = \sqrt{v_x^2 + v_y^2 + v_z^2} \Rightarrow \frac{\partial v}{\partial v_x} = \frac{v_x}{\sqrt{v_x^2 + v_y^2 + v_z^2}} = \frac{v_x}{v} \quad (32.48)$$

Ce qui nous donne finalement:

$$\frac{df}{dv} \frac{\partial v}{\partial v_x} = \frac{v_x}{v} \frac{df}{dv} = \frac{f}{\varphi(v_x)} \frac{d\varphi(v_x)}{dv_x} \Leftrightarrow \frac{1}{v} \frac{df}{dv} = \frac{f}{v_x \varphi(v_x)} \frac{d\varphi(v_x)}{dv_x} \quad (32.49)$$

Le membre de gauche de la dernière égalité dépend uniquement de v , celui de droite uniquement de v_x . Le résultat ne peut être qu'une constante que nous noterons α . Il suit que:

$$\frac{f}{v_x \varphi(v_x)} \frac{d\varphi(v_x)}{dv_x} = \alpha \Leftrightarrow \frac{1}{\varphi(v_x)} d\varphi(v_x) = \alpha v_x dv_x \quad (32.50)$$

En intégrant:

$$\int \frac{1}{\varphi(v_x)} d\varphi(v_x) = \ln(\varphi(v_x)) + c_1^{te} = \int \alpha v_x dv_x = \frac{\alpha}{2} v_x^2 + c_2^{te} \quad (32.51)$$

Donc:

$$\ln(\varphi(v_x)) = \frac{\alpha}{2} v_x^2 + (c_2^{te} - c_1^{te}) = \frac{\alpha}{2} v_x^2 + c^{te} \quad (32.52)$$

et posons $c^{te} = \ln(A)$. Il vient donc:

$$\ln(\varphi(v_x)) = \frac{\alpha}{2} v_x^2 + \ln(A) \Rightarrow \ln\left(\frac{\varphi(v_x)}{A}\right) = \frac{\alpha}{2} v_x^2 \Rightarrow \varphi(v_x) = Ae^{\frac{\alpha}{2} v_x^2} \quad (32.53)$$

Donc identique:

$$\varphi(v_x) = Ae^{\frac{\alpha}{2}v_x^2} \quad \varphi(v_y) = Ae^{\frac{\alpha}{2}v_y^2} \quad \varphi(v_z) = Ae^{\frac{\alpha}{2}v_z^2} \quad (32.54)$$

La loi des probabilités composées impliquant:

$$f(v) = \varphi(v_x)\varphi(v_y)\varphi(v_z) \quad (32.55)$$

Nous avons finalement:

$$f(v) = A^3 e^{\frac{\alpha}{2}(v_x^2+v_y^2+v_z^2)} = A^3 e^{\frac{\alpha v^2}{2}} \quad (32.56)$$

Remarquons que α est nécessairement négatif sinon la probabilité pour une molécule d'avoir une composante de vitesse infinie serait infinie ce qui voudrait dire que toutes les molécules seraient à vitesse infinie et que l'énergie serait infinie !!!

Nous posons:

$$\frac{\alpha}{2} = -B \text{ avec } B > 0 \quad (32.57)$$

Finalement, il vient que:

$$d^3P = dP_x dP_y dP_z = A^3 e^{-Bv^2} dv_x dv_y dv_z \quad (32.58)$$

Il nous reste à normaliser A . La probabilité pour une molécule d'avoir une composante de vitesse comprise entre $]-\infty, +\infty[$ ou une intensité de vitesse comprise entre $[0, +\infty[$ est égale à 1 (100% de chances). Ainsi en utilisant exactement la même méthode calcul que celle vue dans le chapitre de Statistiques pour la loi Normale, nous avons:

$$\int_{-\infty}^{+\infty} dP(v_x) = \int_{-\infty}^{+\infty} A e^{-Bv_x^2} dv_x = A \int_{-\infty}^{+\infty} e^{-Bv_x^2} dv_x = 1 = A \sqrt{\frac{\pi}{B}} \Rightarrow A = \sqrt{\frac{B}{\pi}} \quad (32.59)$$

Nous avons jusqu'ici parlé en terme de probabilités. Un langage équivalent consiste à chercher, dans une enceinte contenant N molécules, le nombre dN de molécules ayant certaines caractéristiques, à savoir par exemple, le nombre de molécules $dN(v_x)$ ayant une composante de vitesse comprise entre v_x et $v_x + dv_x$. Ce nombre étant bien évidemment égal dans une dimension à:

$$dN(v_x) = NdP_x = NA e^{-Bv_x^2} dv_x \quad (32.60)$$

Plus généralement:

$$dN(v) = Nd^3P = N \left(\frac{B}{\pi} \right)^{3/2} e^{-Bv^2} dv_x dv_y dv_z \quad (32.61)$$

Pour obtenir dP nous nous plaçons dans l'espace des vitesses c'est-à-dire un repère cartésien de

coordonnées (v_x, v_y, v_z) . Les composantes v_x, v_y, v_z ne sont pas indépendantes puisque liées par la relation:

$$v = v_x^2 + v_y^2 + v_z^2 \quad (32.62)$$

L'extrémité des vecteurs vitesse $\vec{v}$ ayant une intensité de vitesse v , c'est-à-dire de composantes de vitesse liées par la relation ci-dessus, se trouve dans l'espace des vitesses sur la sphère de rayon v . Il en sera de même pour l'extrémité des vecteurs vitesse ayant une intensité $v + dv$.

Dans l'espace des vitesses, nous délimitons une portion d'espace comprise entre la sphère de rayon v et la sphère de rayon $v + dv$, de volume égal à $4\pi v^2 dv$.

La probabilité d^3P est proportionnelle à $dv_x dv_y dv_z$ c'est-à-dire au volume élémentaire dans l'espace des vitesses. Pour obtenir $dP(v)$, nous devons intégrer d^3P à tous les vecteurs vitesses possibles c'est à dire ayant leur extrémité entre les deux sphères. Cette intégration est particulièrement simple puisque l'intensité v de la vitesse dans cet espace (inter-volume) est constante. Nous obtenons donc:

$$dP(v) = \left(\frac{B}{\pi}\right)^{3/2} e^{-Bv^2} 4\pi v^2 dv \quad (32.63)$$

et ainsi, dans une enceinte contenant N molécules, le nombre de molécules $dN(v)$ ayant un module de vitesse compris entre v et $v + dv$ est:

$$dN(v) = NdP(v) \quad (32.64)$$

Rappelons maintenant que conformément à ce que nous avons longuement étudié dans le chapitre de Statistiques, la valeur moyenne $\langle G \rangle$ d'une grandeur G est le produit de G pondérée par la probabilité d'obtenir G intégré à toutes les valeurs possibles de cette grandeur telle qu'en une dimension nous avons (selon x):

$$\langle v_x^2 \rangle = \int_0^{+\infty} v_x^2 dP(v_x) = \left(\frac{B}{\pi}\right)^{1/2} \int_0^{+\infty} v_x^2 \exp(-Bv_x^2) dv_x = \left(\frac{B}{\pi}\right)^{1/2} \frac{\sqrt{\pi}}{4B^{3/2}} = \frac{1}{2B} \quad (32.65)$$

où l'intégrale a été calculée en faisant un changement de variable $v_x^2 = x$ et où nous avons ensuite utilisé une intégration par parties.

Nous avons respectivement de manière générale:

$$\langle v^2 \rangle = \int_0^{+\infty} v^2 P(v) dv = 4\pi \left(\frac{B}{\pi}\right)^{3/2} \int_0^{+\infty} v^4 \exp(-Bv^2) dv = 4\pi \left(\frac{B}{\pi}\right)^{3/2} \frac{3\sqrt{\pi}}{8B^{5/2}} = \frac{3}{2B} \quad (32.66)$$

où l'intégrale a été calculée de la même manière.

Donc:

$$\langle v_x^2 \rangle = \langle v_y^2 \rangle = \langle v_z^2 \rangle = \frac{1}{3} \mu^2 \text{ ou } \langle v_x^2 \rangle + \langle v_y^2 \rangle + \langle v_z^2 \rangle = \frac{3}{2B} \quad (32.67)$$

La thermodynamique, ou la mécanique des fluides (voir théorème du Viriel) nous donne pour un gaz parfait monoatomique (cf. chapitres de Mécanique Des Milieux Continus et de Thermodynamique), de capacité calorifique et volume constant:

$$E_c = \frac{3}{2} NkT \quad (32.68)$$

Si nous formulons l'hypothèse que la partie de l'énergie liée à la température est due à l'agitation cinétique des molécules, nous pouvons écrire:

$$E_c = \frac{3}{2} NkT = N \frac{1}{2} m \langle v^2 \rangle \Rightarrow \langle v^2 \rangle = \frac{3kT}{m} \quad (32.69)$$

Nous définissons alors la vitesse thermique moyenne par:

$$\langle v_{th} \rangle = \sqrt{\langle v^2 \rangle} = \sqrt{\frac{3kT}{m}} \quad (32.70)$$

Une application pour l'électron dans le cas des semi-conducteurs (où nous approximations certaines relations par une distribution de Maxwell-Boltzmann) donne une vitesse thermique moyenne à température ambiante de 120'000 [m/s].

Mais pour en revenir à notre distribution.... Nous avons donc:

$$\langle v^2 \rangle = \frac{3}{2B} = \frac{3kT}{m} \Rightarrow B = \frac{m}{2kT} \quad (32.71)$$

Ainsi, nous avons finalement:

$$dP(v) = 4\pi v^2 \left(\frac{m}{2\pi kT} \right)^{3/2} \exp\left(-\frac{m}{2kT} v^2 \right) dv \quad (32.72)$$

qui est donc la distribution des vitesses dans un gaz monoatomique dont voici un exemple de tracé (les unités des axes en ordonnées sont arbitraires):

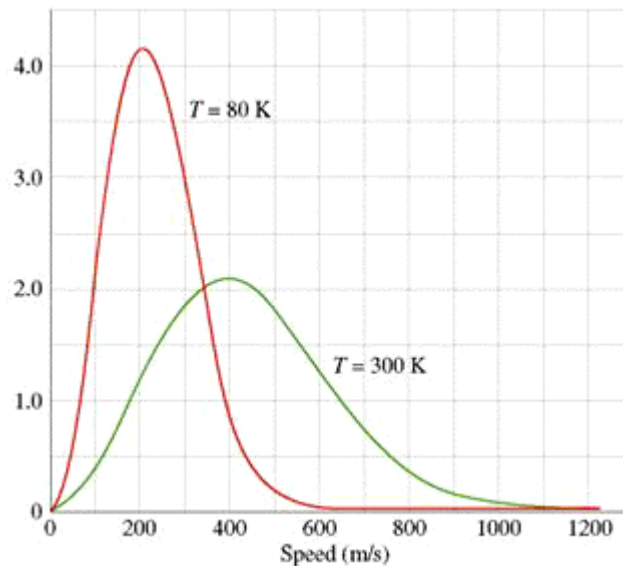


Figure: 32.3 - Distribution des vitesses dans un gaz monoatomique

La relation précédente donne donc la proportion des molécules de gaz ayant à un instant t donné une vitesse v . Nous voyons que la densité de probabilité s'étend jusqu'à l'infini et nous trouvons une probabilité non nulle (tendant vers zéro) que la particule ait une vitesse extrêmement élevée. En fait, si nous avons un système de particules identiques à la même température T , alors nous en retrouverons un certain nombre avec ces vitesses très élevées. Au cœur des étoiles ce sont ces rares particules très rapides qui sont capables de vaincre la répulsion coulombienne entre les noyaux et de faire qu'ils s'approchent l'un et l'autre pour initier le mécanisme de fusion nucléaire qui est la source d'énergie des étoiles.

Nous avons alors pour une unique direction spatiale et selon une notation courante dans la littérature anglo-saxonne (en adoptant les développements précédents on tombe relativement facilement sur ce résultat) la fonction de distribution suivante:

$$f(v) = 4\pi v^2 \left(\frac{m}{2\pi kT} \right)^{1/2} \exp\left(-\frac{mv^2}{2kT} \right) \quad (32.73)$$

et pour les trois dimensions spatiales la notation suivante aussi courante dans la littérature anglo-saxonne (certains termes sont simplement mis sous la racine):

$$f(v) = \sqrt{\frac{2}{\pi} \left(\frac{m}{kT} \right)^3} v^2 \exp\left(-\frac{mv^2}{2kT} \right) \quad (32.74)$$

Cette fonction de distribution permet donc de définir la vitesse la plus probable (la valeur "modale" comme l'on dit en statistiques) notée parfois v_m , qui correspond au maximum de la courbe $f(v)$, soit là où la dérivée première est nulle. Ainsi, pour une unique dimension spatiale nous avons en adoptant la notation anglo-saxonne pour la proportion de molécules de gaz:

$$\begin{aligned}
\frac{\partial f(v)}{\partial v} &= \frac{\partial}{\partial v} 4\pi v^2 \left(\frac{m}{2\pi kT} \right)^{1/2} \exp\left(-\frac{mv^2}{2kT}\right) \\
&= 8\pi v \left(\frac{m}{2\pi kT} \right)^{1/2} \exp\left(-\frac{mv^2}{2kT}\right) - 8 \frac{m}{2kT} \pi v^3 \left(\frac{m}{2\pi kT} \right)^{1/2} \exp\left(-\frac{mv^2}{2kT}\right) \quad (32.75) \\
&= 8 \cdot \pi \cdot v \cdot \left[1 - \frac{m}{2kT} v^2 \right] \left(\frac{m}{2\pi kT} \right)^{1/2} \exp\left(-\frac{mv^2}{2kT}\right) = 0
\end{aligned}$$

Dès lors (vous remarquerez que le résultat aurait été le même si nous avions pris toutes les dimensions spatiales... ce qui est cohérent!):

$$\begin{aligned}
1 - \frac{m}{2kT} v^2 &= 0 \\
\Rightarrow \frac{m}{2kT} v^2 &= 1 \quad (32.76) \\
\Rightarrow v^2 &= \frac{2kT}{m}
\end{aligned}$$

d'où la vitesse la plus probable (attention! ce n'est pas la vitesse maximale puisque nous savons que la distribution s'étend jusqu'à la vitesse infinie):

$$v = \sqrt{\frac{2kT}{m}} \quad (32.77)$$

Pour la vitesse moyenne (l'espérance) et en utilisant directement toutes les dimensions spatiales nous avons:

$$\begin{aligned}
\bar{v} = E(v) &= \int_0^{+\infty} v f(v) dv = \int_0^{+\infty} v 4\pi v^2 \left(\frac{m}{2\pi kT} \right)^{3/2} \exp\left(-\frac{mv^2}{2kT}\right) dv \\
&= 4\pi \left(\frac{m}{2\pi kT} \right)^{3/2} \int_0^{+\infty} v^3 \exp\left(-\frac{m}{2kT} v^2\right) dv \quad (32.78)
\end{aligned}$$

Nous avons donc une intégrale du type:

$$\int_0^{+\infty} x^3 \exp(-ax^2) dx \quad (32.79)$$

Le mieux est de décomposer $x^3 e^{-ax^2}$ en $x^2 \cdot x e^{-ax^2}$ et on intègre par parties. Nous avons alors:

$$\int_0^{+\infty} x^2 \cdot (x e^{-ax^2}) dx = \underbrace{-x^2 \cdot \frac{1}{2a} e^{-ax^2}}_0 \Big|_0^{+\infty} + \int_0^{+\infty} 2x \cdot \frac{1}{2a} e^{-ax^2} dx \quad (32.80)$$

La dernière intégrale est facilement calculable. C'est la même que dans le chapitre de Statistiques pour la loi Normale. Donc:

$$\int_0^{+\infty} \frac{1}{a} x \cdot e^{-ax^2} dx = \frac{1}{a} \left(-\frac{1}{2a} e^{-ax^2} \right) \Big|_0^{+\infty} = \frac{1}{2a^2} \quad (32.81)$$

Il vient finalement:

$$\begin{aligned} \bar{v} &= 4\pi \left(\frac{m}{2\pi kT} \right)^{3/2} \int_0^{+\infty} v^3 \exp\left(-\frac{m}{2kT} v^2\right) dv = 4\pi \left(\frac{m}{2\pi kT} \right)^{3/2} \cdot \frac{1}{2 \left(\frac{m}{2kT} \right)^2} \\ &= \left(\frac{m}{2\pi kT} \right)^{3/2} \cdot \frac{8\pi k^2 T^2}{m^2} = \left(\frac{8kT}{\pi m} \right)^{1/2} \end{aligned} \quad (32.82)$$

et là par contre le résultat est dépendant du nombre de dimensions spatiales que nous prenons!

3.2. DISTRIBUTION DE MAXWELL-BOLTZMANN

En physique quantique corpusculaire, nous apprenons que l'énergie d'une particule est quantifiée, c'est à dire que les valeurs possibles pour l'énergie forment un spectre discret. Même si, dans un certain nombre de situations courantes, pour une particule et encore plus pour un système constitué d'un grand nombre de particules, les niveaux d'énergie sont si serrés que nous pouvons traiter, sur le plan mathématique, ce spectre comme continu (approximation du continuum), il n'empêche qu'en toute rigueur ils sont quantifiés. Cette approche quantifiée des distributions physiques au niveau corpusculaire de la matière est souvent désignée sous le nom de "**statistique quantique**".

Une particule, ayant un niveau d'énergie E_i , peut être dans différents sous-états.

Nous savons (nous n'avons pas utilisé l'équation de Schrödinger dans le chapitre de Physique Quantique Corpusculaire pour le montrer strictement) que, pour décrire un atome, nous introduisons quatre nombres quantiques, à savoir:

- le nombre quantique principal qui quantifie l'énergie
- le nombre quantique secondaire qui quantifie le moment cinétique
- le nombre quantique magnétique qui quantifie le moment magnétique
- le spin qui quantifie la rotation propre des électrons de l'atome

Ainsi pour une même énergie (pour une valeur particulière du nombre quantique principal), un atome ou un électron peuvent posséder différentes valeurs des nombres quantiques secondaire, magnétique ou de spin.

Pour qualifier la possibilité de sous-états correspondant à une même énergie, nous employons l'expression "**dégénérescence**" et nous traduisons par la variable g_i le nombre de dégénérescences correspondant à un même niveau d'énergie E_i .

Nous allons considérer un système composé de N particules qui se placent sur K différents niveaux d'énergie E_i . Nous trouvons n_i particules sur le niveau d'énergie E_i . Nous avons alors toujours les relations suivantes pour l'énergie totale (que nous notons par la même lettre que celle utilisée en

et en utilisant la formule de Stirling (cf. chapitre de Méthodes Numériques):

$$n! = e^{n \ln(n)} e^{-n} \Rightarrow \ln(n!) = n \ln(n) - n \quad (32.87)$$

en se rappelant que celle-ci est une approximation grossière pour les plus petites valeurs de n mais devient une approximation acceptable pour n supérieur à 1000.

Nous avons:

$$\ln(\Omega) = N \ln(N) - N + \sum_{i=1}^K (n_i \ln(g_i) - n_i \ln(n_i) + n_i) = N \ln(N) + \sum_{i=1}^K n_i \ln\left(\frac{g_i}{n_i}\right) \quad (32.88)$$

Nous recherchons maintenant la distribution la plus probable, c'est-à-dire celle qui maximise Ω (et donc qui maximise implicitement l'entropie qui ne peut selon l'expérience qu'augmenter pour un système isolé). Pour trouver l'extremum, nous allons différencier cette expression tel que :

$$\frac{\partial \ln(\Omega)}{\partial n_i} = 0 \quad (32.89)$$

donc tous les autres paramètres autres que n_i sont fixés.

Or, comme:

$$\begin{aligned} \sum_{i=1}^K n_i E_i = U &\Leftrightarrow U - \sum_{i=1}^K n_i E_i = 0 \Leftrightarrow \beta \left(U - \sum_{i=1}^K n_i E_i \right) = 0 \\ \sum_{i=1}^K n_i = N &\Leftrightarrow N - \sum_{i=1}^K n_i = 0 \Leftrightarrow \alpha \left(N - \sum_{i=1}^K n_i \right) = 0 \end{aligned} \quad (32.90)$$

où α, β sont des constantes qui permettent de s'assurer de la cohérence de l'analyse dimensionnelle (des unités quoi...).

Il est donc équivalent et nécessaire pour prendre en compte également ces paramètres intrinsèques (le fait d'ajouter ainsi de manière astucieuse des termes nuls est nommé "**méthode des multiplicateurs de Lagrange**") d'écrire:

$$\frac{\partial \ln(\Omega)}{\partial n_i} = 0 \quad (32.91)$$

sous la forme plus complète suivante:

$$\frac{\partial}{\partial n_i} \left(\ln(\Omega) + \alpha \left(N - \sum_{i=1}^K n_i \right) + \beta \left(U - \sum_{i=1}^K n_i E_i \right) \right) = 0 \quad (32.92)$$

Donc après réarrangement:

$$\frac{\partial \ln(\Omega)}{\partial n_i} = - \frac{\partial}{\partial n_i} \left[\alpha \left(N - \sum_{i=1}^K n_i \right) + \beta \left(U - \sum_{i=1}^K n_i E_i \right) \right] \quad (32.93)$$

ce qui donne après une première simplification (élimination de la dérivée des constantes):

$$\frac{\partial \ln(\Omega)}{\partial n_i} = \frac{\partial}{\partial n_i} \left[\alpha \sum_{i=1}^K n_i + \beta \sum_{i=1}^K n_i E_i \right] = \frac{\partial}{\partial n_i} \sum_{i=1}^K [\alpha n_i + \beta n_i E_i] \quad (32.94)$$

Ce qui nous amène à écrire (puisque tous les termes de la somme indexés sur une valeur i différente de celle de la dérivée partielle sont nuls):

$$\frac{\partial \ln(\Omega)}{\partial n_i} = \frac{\partial}{\partial n_i} (\alpha n_i + \beta n_i E_i) = \alpha + \beta E_i \quad (32.95)$$

Ce qui nous donne finalement:

$$\frac{\partial \ln(\Omega)}{\partial n_i} = \alpha + \beta E_i \quad (32.96)$$

qui devra donc aussi être nul. Ainsi, nous avons astucieusement fait apparaître un résultat qui n'en est pas moins très puissant. En imposant de maximiser l'entropie nous avons fait apparaître deux termes (incluant chacun une constante qu'il faudra déterminer plus tard!) et dont la somme doit être nulle pour assurer la validité expérimentale du modèle théorique lorsque l'entropie du système isolé est maximale.

Mais cette dernière relation équivaut aussi à rechercher (ce qui va nous être très utile plus loin!):

$$d \ln(\Omega) = d \left(N \ln(N) + \sum_{i=1}^K n_i \ln \left(\frac{g_i}{n_i} \right) \right) = 0 \quad (32.97)$$

Effectivement voyons cela en utilisant à nouveau la méthode astucieuse des multiplicateurs de Lagrange (ajout de termes nuls en supposant les niveaux d'énergie E_i et le nombre de dégénérescences g_i fixes) et en se rappelant (cf. chapitre de Calcul Différentiel Et Intégral) que:

$$\frac{d \left(x \cdot \ln \left(\frac{a}{x} \right) \right)}{dx} = \ln \left(\frac{a}{x} \right) - 1 \Leftrightarrow d \left(x \cdot \ln \left(\frac{a}{x} \right) \right) = \ln \left(\frac{a}{x} \right) dx - dx \quad (32.98)$$

Nous avons alors:

$$\begin{aligned}
& d \left(N \ln(N) + \sum_{i=1}^K n_i \ln \left(\frac{g_i}{n_i} \right) + \alpha \left(N - \sum_{i=1}^K n_i \right) + \beta \left(U - \sum_{i=1}^K n_i E_i \right) \right) = 0 \\
& \Rightarrow \sum_{i=1}^K dn_i \ln \left(\frac{g_i}{n_i} \right) - \underbrace{\sum_{i=1}^K dn_i}_{\rightarrow 0} - \alpha \sum_{i=1}^K dn_i - \beta \sum_{i=1}^K E_i dn_i = \sum_{i=1}^K \left[\left(\ln \left(\frac{g_i}{n_i} \right) - \alpha - \beta E_i \right) dn_i \right] = 0 \\
& \Rightarrow \left(\ln \left(\frac{g_i}{n_i} \right) - \alpha - \beta E_i \right) dn_i = 0 \quad (32.99) \\
& \Rightarrow \ln \left(\frac{g_i}{n_i} \right) = \alpha + \beta E_i \\
& \Rightarrow \ln \left(\frac{n_i}{g_i} \right) = -\alpha - \beta E_i
\end{aligned}$$

d'où:

$$n_i = g_i e^{-\alpha - \beta E_i} \quad (32.100)$$

Rappelons que nous avons pour l'entropie de ce système:

$$S = k \ln(\Omega) \quad (32.101)$$

et donc:

$$dS = k d(\ln(\Omega)) = k \sum_i \frac{\partial \ln(\Omega)}{\partial n_i} dn_i \quad (32.102)$$

et comme:

$$\frac{\partial \ln(\Omega)}{\partial n_i} = \alpha + \beta E_i \quad (32.103)$$

Nous avons donc:

$$\frac{dS}{k} = \sum_{i=1}^K (\alpha + \beta E_i) dn_i = \alpha \sum_{i=1}^K dn_i + \beta \sum_{i=1}^K E_i dn_i = \alpha dN + \beta dU \quad (32.104)$$

La thermodynamique nous conduit à (cf. chapitre de Thermodynamique):

$$dU = TdS + \sum_{i=1} \mu_i dN_i \quad (32.105)$$

où nous voyons donc apparaître le potentiel chimique qui a une influence lorsque le nombre de particules considérées varie.

Si toutes les particules (ou groupes de particules) sont identiques $\mu_i = c^{te} = \mu$ donc:

$$dU = TdS + \mu \sum_{i=1} dN_i = TdS + \mu dN \quad (32.106)$$

Ce qui finalement conduit à:

$$\frac{dS}{k} = \frac{dU}{kT} - \frac{\mu}{kT} dN \quad (32.107)$$

et nous amène à identifier:

$$\frac{dS}{k} = \frac{dU}{kT} - \frac{\mu}{kT} dN = \beta dU + \alpha dN \quad (32.108)$$

et donc:

$$\alpha = -\frac{\mu}{kT} \text{ et } \beta = \frac{1}{kT} \quad (32.109)$$

Nous pouvons donc réécrire la relation suivante:

$$n_i = g_i e^{-\alpha - \beta E_i}$$

sous la forme finale et traditionnelle de la "**fonction de Maxwell-Boltzmann**":

$$n_i = g_i \exp\left(\frac{-E_i + \mu}{kT}\right) = g_i \exp\left(\frac{\mu}{kT}\right) \exp\left(\frac{-E_i}{kT}\right) \quad (32.110)$$

Que nous retrouvons plus fréquemment dans la littérature sous la forme suivante:

$$n_i = g_i \exp\left(\frac{\mu - E_i}{kT}\right) \quad (32.111)$$

et particulièrement sous cette forme:

$$f_{MB}(E, T, \mu) = \exp\left(\frac{\mu - E_i}{kT}\right) \quad (32.112)$$

Nous retrouverons cette relation dans le domaine de la théorie des semi-conducteurs (cf. chapitre d'Électrocinétique) ainsi que dans notre étude des plasmas (cf. chapitre de Mécanique Des Milieux Continus).

Comme $N = \sum_{i=1}^K n_i$, nous avons:

$$N = g_i \exp\left(\frac{\mu}{kT}\right) \sum_{i=1}^K \exp\left(\frac{-E_i}{kT}\right) \quad (32.113)$$

Nous pouvons alors calculer n_i / N et nous obtenons ainsi la formulation discrète de la "**statistique de**

Maxwell-Boltzmann":

$$P = \frac{g_i \exp\left(\frac{-E_i}{kT}\right)}{\sum_{i=1}^K g_i \exp\left(\frac{-E_i}{kT}\right)} \quad (32.114)$$

Cette relation donne donc le rapport entre le nombre de particules qui par hypothèse n'interagissent pas entre elles (sinon le modèle est trop complexe) dans un état d'énergie donné E_i et le nombre de particules qui n'interagissent pas entre elles et pouvant prendre les différents états d'énergie discrets E_i . Ainsi, connaissant N , il est possible à l'aide de cette relation de connaître la proportion de particules dans un état d'énergie particulier.

Remarque: La statistique de Maxwell-Boltzmann s'applique en l'absence d'interaction entre particules et est donc valable pour un gaz parfait mais ne s'applique pas, par exemple, à un liquide. De plus, elle s'applique uniquement aux hautes températures lorsque les effets quantiques sont négligeables. À basse température nous utilisons la statistique de Bose-Einstein pour les bosons (en référence au nom de Bose...) et la statistique de Fermi-Dirac pour les fermions (en référence au nom de Fermi) que nous étudierons plus loin.

Nous appelons le terme au dénominateur, la "**fonction de partition canonique**". Elle est le plus souvent notée Z_C tel que:

$$P = \frac{g_i \exp\left(\frac{-E_i}{kT}\right)}{Z_C} \quad (32.115)$$

Les ensembles microcanonique, canonique et grand canonique correspondent à des ensembles de systèmes soumis à des contraintes différentes qui sont respectivement :

- Microcanonique: énergie fixée, nombre de particules fixé, volume fixé
- Canonique: énergie moyenne fixée = température imposée, nombre de particules fixé, volume fixé
- Grand canonique: énergie moyenne fixée = température imposée, nombre de particules moyen fixé = potentiel chimique imposé, volume fixé)

Ceci est vrai en quantique comme en classique.

3.2.1. DISTRIBUTION DE BOLTZMANN

Voyons maintenant une première application intéressante de la fonction de distribution de Maxwell-Boltzmann.

Considérons le cas intéressant où:

- Le volume ainsi que le nombre de particules sont fixés (système microcanonique)

- Le fait que le nombre de particules soit fixé, le potentiel chimique est nul
- Une température suffisamment élevée pour que les états d'énergie soient en quantité infinie

Il va relativement de soi que si les états d'énergie sont en quantité infinie, le facteur de dégénérescence g_i sera égal à l'unité.

Nous avons alors la distribution de Maxwell-Boltzmann qui devient sous ces conditions la "**distribution de Boltzmann**":

$$P = \frac{\exp\left(\frac{-E}{kT}\right)}{Z_{mc}} = \frac{1}{Z_{mc}} \exp\left(\frac{-E}{kT}\right) \quad (32.116)$$

que nous retrouvons plus fréquemment sous la forme suivante:

$$f_B(E, T) = \exp\left(-\frac{E}{kT}\right) \quad (32.117) .$$

Si nous considérons que dans un gaz globalement neutre, l'énergie est définie uniquement par l'énergie cinétique (cas du gaz parfait). Donc avec le passage au continu, nous avons:

$$Z_{mc} = \int_{-\infty}^{+\infty} \exp\left(\frac{-E}{kT}\right) dv \quad (32.118)$$

Dès lors, nous pouvons intégrer la fonction de Boltzmann ci-dessus sur toutes les vitesses et nous devrions avoir:

$$P(-\infty < v < +\infty) = \int_{-\infty}^{+\infty} \frac{1}{Z_{mc}} \exp\left(\frac{-E}{kT}\right) dv = \frac{1}{Z_{mc}} \int_{-\infty}^{+\infty} \exp\left(\frac{-\frac{1}{2}mv^2}{kT}\right) dv = 1 \quad (32.119)$$

Avec la fonction de partition microcanonique:

$$Z_{mc} = \int_{-\infty}^{+\infty} \exp\left(-\frac{mv^2}{2kT}\right) dv \quad (32.120)$$

pour que la condition de normalisation soit satisfaite (et puis de toute façon cela correspond aux développements effectués plus haut!).

Nous reconnaissons ici une intégrale qui nous est familière et démontrée dans le chapitre de Statistiques lors de notre étude de la loi Normale. Nous avons alors:

$$Z_{mc} = \sqrt{\frac{2\pi kT}{m}} \quad (32.121)$$

Dès lors, la fonction de densité de probabilité s'écrit:

$$P(-\infty < v < +\infty) = \frac{1}{\sqrt{\frac{2\pi kT}{m}}} \int_{-\infty}^{+\infty} \exp\left(-\frac{mv^2}{2kT}\right) dv \quad (32.122)$$

Maintenant, puisque la vitesse est traitée comme une variable aléatoire, intéressons-nous au calcul de la valeur moyenne du carré de la vitesse qui se note traditionnellement de la manière suivante (la notation n'a aucun rapport avec les ket-bra de la physique quantique ou des produits scalaires fonctionnels!):

$$\langle v^2 \rangle = \frac{1}{\sqrt{\frac{2\pi kT}{m}}} \int_{-\infty}^{+\infty} v^2 \exp\left(-\frac{mv^2}{2kT}\right) dv \quad (32.123)$$

En statistiques, nous noterions cela:

$$E(v^2) = \frac{1}{\sqrt{\frac{2\pi kT}{m}}} \int_{-\infty}^{+\infty} v^2 \exp\left(-\frac{mv^2}{2kT}\right) dv \quad (32.124)$$

Concentrons-nous sur:

$$\int_{-\infty}^{+\infty} v^2 \exp\left(-\frac{mv^2}{2kT}\right) dv \quad (32.125)$$

Or, si nous posons:

$$\sigma = \sqrt{\frac{2kT}{m}} \quad (32.126)$$

Nous avons alors:

$$\int_{-\infty}^{+\infty} v^2 e^{-\left(\frac{v}{\sigma}\right)^2} dv \quad (32.127)$$

Soit avec un changement de variable:

$$\sigma^3 \int_{-\infty}^{+\infty} y^2 e^{-y^2} dy \quad (32.128)$$

et il s'agit exactement de l'intégrale de la loi Normale Centrée démontrée dans le chapitre de Statistique comme valant:

$$\int_{-\infty}^{+\infty} v^2 e^{-\left(\frac{v}{\sigma}\right)^2} dv = \sigma^3 \frac{\sqrt{\pi}}{2} \quad (32.129)$$

Dès lors:

$$\begin{aligned}
 E(v^2) &= \frac{1}{\sqrt{\frac{2\pi kT}{m}}} \int_{-\infty}^{+\infty} v^2 \exp\left(-\frac{mv^2}{2kT}\right) dv = \frac{1}{\sqrt{\frac{2\pi kT}{m}}} \left(\frac{2kT}{m}\right)^{3/2} \frac{\sqrt{\pi}}{2} \\
 &= \frac{1}{\sqrt{\frac{2kT}{m}}} \left(\frac{2kT}{m}\right) \left(\frac{2kT}{m}\right)^{1/2} \frac{1}{2} = \left(\frac{2kT}{m}\right) \frac{1}{2} = \frac{kT}{m}
 \end{aligned}
 \tag{32.130}$$

et au final:

$$\langle E \rangle = \frac{1}{2} m \langle v^2 \rangle = \frac{1}{2} m \frac{kT}{m} = \frac{1}{2} kT \tag{32.131}$$

pour une particule pour une dimension spatiale. Nous retrouvons au fait le même résultat qu'avec la distribution de Maxwell des vitesses démontrée, avec une démarche plus rigoureuse, mais nous arrivons aux mêmes conclusions.

3.3. DISTRIBUTION DE FERMI-DIRAC

Le principe d'indiscernabilité peut avoir des conséquences très importantes sur la statistique. Nous distinguons deux types de particules indiscernables: les bosons (comme le photon) et les fermions (comme l'électron).

Rappels:

R1. Les premiers correspondent à des particules dont la fonction d'onde représentative est toujours symétrique alors que celle des fermions est antisymétrique.

R2. Le principe d'exclusion de Pauli impose que 2 fermions ne peuvent pas se trouver dans le même état quantique. Les bosons, eux le peuvent !

Leurs propriétés respectives ont pour conséquence importante que l'énergie minimale d'un ensemble de N bosons est égale à N fois l'énergie minimale de chaque boson. Alors que pour un ensemble de fermions, l'énergie minimale est égale à la somme des N énergies les plus faibles.

Ces deux types de particules entraînent deux types de statistiques: la statistique de Fermi-Dirac pour les fermions (que nous allons démontrer en premier) et la statistique de Bose-Einstein pour les bosons (qui suivra).

Il n'existe donc qu'une seule manière de répartir N fermions sur les états d'énergies accessibles (au lieu des $N!$ pour les particules discernables). Il ne peut pas y avoir plus de particules n_i dans un niveau d'énergie E_i qu'il existe de dégénérescences g_i . Donc:

$$0 \leq n_i \leq g_i \tag{32.132}$$

Le nombre de combinaisons possibles pour un niveau E_i , dégénéré g_i fois et comportant n_i particules est donc la combinatoire $C_{n_i}^{E_i}$. Le nombre total de configurations est donc:

$$\Omega = \prod_i C_{n_i}^{g_i} = \prod_i \frac{g_i!}{n_i! (g_i - n_i)!} \quad (32.133)$$

Ce qui peut se vérifier avec la figure ci-dessous:

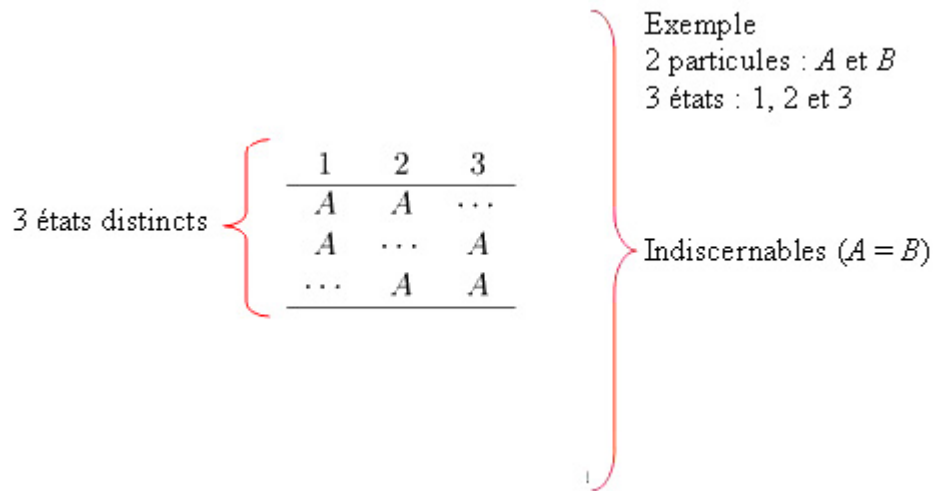


Figure: 32.5 - États dans la distribution de Fermi-Dirac

La statistique est donc bien différente du cas classique de Maxwell-Boltzmann. En prenant le logarithme du nombre de micro-états et en faisant usage de l'approximation par la formule de Stirling comme plus haut, il vient:

$$\ln(\Omega) = \sum_i \left[(g_i \ln(g_i) - g_i) - (n_i \ln(n_i) - n_i) - ((g_i - n_i) \ln(g_i - n_i) - (g_i - n_i)) \right] \quad (32.134)$$

Que nous pouvons déjà simplifier une première fois:

$$\ln(\Omega) = \sum_i \left[g_i \ln(g_i) - n_i \ln(n_i) - (g_i - n_i) \ln(g_i - n_i) \right] \quad (32.135)$$

et en différenciant cette expression pour trouver le maximum, nous obtenons:

$$d \ln(\Omega) = d \sum_i \left[g_i \ln(g_i) - n_i \ln(n_i) - (g_i - n_i) \ln(g_i - n_i) \right] = 0 \quad (32.136)$$

Terme à terme:

$$\begin{aligned} 1. \quad d(g_i \ln(g_i)) &= 0 \\ 2. \quad d(n_i \ln(n_i)) &= dn_i \ln(n_i) + n_i d(\ln(n_i)) = dn_i \ln(n_i) + n_i \frac{1}{n_i} dn_i = dn_i \ln(n_i) + dn_i \\ 3. \quad d((g_i - n_i) \ln(g_i - n_i)) &= d(g_i \ln(g_i - n_i)) - d(n_i \ln(g_i - n_i)) \\ &= \underbrace{d g_i}_{=0} \ln(g_i - n_i) - g_i d \ln(g_i - n_i) - dn_i \ln(g_i - n_i) - n_i d \ln(g_i - n_i) \\ &= g_i \left(\frac{1}{g_i - n_i} \right) - dn_i \ln(g_i - n_i) + n_i \left(\frac{1}{g_i - n_i} \right) = -dn_i \ln(g_i - n_i) + \frac{g_i + n_i}{g_i - n_i} \end{aligned} \quad (32.137)$$

Or, nous avons par conservation et par symétrie:

$$\sum_i dn_i = 0, \sum_i \frac{g_i + n_i}{g_i - n_i} = 0 \quad (32.138)$$

Donc finalement:

$$-d \ln(\Omega) = \sum_i [\ln(n_i) - \ln(g_i - n_i)] dn_i = 0 \quad (32.139)$$

Pour respecter les contraintes sur l'énergie et le nombre de particules, nous utilisons encore une fois la méthode des multiplicateurs de Lagrange:

$$\ln(n_i) - \ln(g_i - n_i) + \alpha + \beta E_i = 0 \quad (32.140)$$

Ce qui nous amène à la distribution de Fermi-Dirac:

$$n_i = \frac{g_i}{e^{\alpha + \beta E_i} + 1} \quad (32.141)$$

Les paramètres β et α jouent le même rôle que dans la distribution de Maxwell-Boltzmann. Nous avons ainsi:

$$\beta = \frac{1}{kT} \text{ et } \alpha = -\frac{\mu}{kT} \quad (32.142)$$

Nous avons alors:

$$n_i = \frac{g_i}{\exp\left(\frac{E_i - \mu}{kT}\right) + 1} \quad (32.143)$$

Ainsi, en physique quantique, la statistique de Fermi-Dirac désigne le nombre de fermions indiscernables qui n'interagissent pas entre eux (sinon le modèle est trop complexe) sur les états d'énergie d'un système à l'équilibre thermodynamique dégénéré (donc le nombre de fermions occupant le niveau d'énergie donné!).

Pour les systèmes macroscopiques, les niveaux d'énergie sont si serrés (ou tellement nombreux) que nous pouvons considérer le spectre d'énergies comme continu (approximation du continuum).

Nous raisonnerons donc dans ce contexte de continuum, ce qui nous permet d'écrire en normant aux nombres de particules mises en jeu (tout ce que l'on demande à la fonction c'est de nous dire comment sont réparties les N particules):

$$f_{FD}(E, T, \mu) = \frac{1}{\exp\left(\frac{E - \mu}{kT}\right) + 1} \quad (32.144)$$

ou:

$$f_{FD}(\varepsilon, T, \mu) = \frac{1}{e^{\beta(\varepsilon - \mu)} + 1} \quad (32.145)$$

pour la fonction de Fermi-Dirac avec le tracé correspondant plus bas de la distribution.

Cette relation est très importante par exemple dans la théorie des semi-conducteurs (cf. chapitre d'Electrocinétique) qui est à la base de l'électronique des 20ème et 21ème siècles.

De manière numérique, nous pouvons simuler l'évolution de l'allure de la distribution (qui n'est pas une distribution dans le sens mathématique du terme) en fonction de l'énergie et de la température. Pour cela, nous supposons pour simplifier que la constante de Boltzmann vaut 1 et que le potentiel chimique μ vaut 2 (nous notons que celui-ci est en toute rigueur fonction de la température). Ainsi, nous pouvons écrire le petit programme suivant sous Matlab 5.0.0.473:

```
clear all; kb=1; % Constante de Boltzmann
mu=2; % Potentiel Chimique
T=0.001:0.1:1; % " Gradient " de température pour le programme
for j=1:length(T)
    beta(j)=1/(kb*T(j));
    epsilon=0.1:0.1:4; % L'énergie avec le pas
    for i=1:length(epsilon)
        Nf(i,j)=1/(exp(beta(j)*(epsilon(i)-mu))+1); % Nb(epsilon,beta) moyen de fermions au
    end % niveau d'énergie epsilon
    hold on %
    plot(epsilon,Nf(:,j));
    pause(0)
end
```

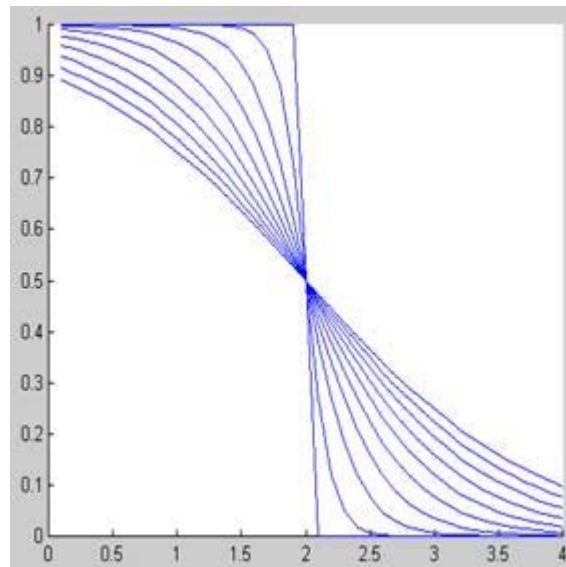


Figure: 32.6 - Tracé avec Matlab 5.0.0.473 de la distribution de Fermi-Dirac

Au zéro absolu $T = 0$ nous voyons que nous avons une marche. A cette température, les niveaux d'énergie dégénérés sont occupés à bloc en partant du niveau d'énergie le plus bas jusqu'à un certain niveau représentant la chute de la marche. Nous disons alors que le gaz de fermions est "**complètement dégénéré**".

La fonction de Fermi-Dirac, au zéro absolu, vaut donc 1 si E est inférieur à μ et 0 pour les valeurs supérieures (le système choisit son état d'énergie minimale où les N particules occupent les N états de plus basse énergie).

Évidemment, dans les hautes énergies d'excitation (ou hautes températures) la probabilité (rapport) d'occupation d'un état est très faible et donc il n'est pas important d'appliquer le principe de Pauli (la probabilité que deux électrons veuillent occuper le même niveau d'énergie est très faible). Pour ces raisons, ce régime limite entre le comportement quantique et classique est appelé parfois "**régime classique**" et la haute température correspondante à l'excitation énergétique est appelée "**température de Fermi**".

Le potentiel chimique est quant à lui par définition le dernier niveau énergétique occupé au zéro absolu (la fameuse marche d'escalier!). Comme nous le verrons dans le chapitre de Physique Quantique Ondulatoire, celui-ci est noté E_f (et à l'opposé μ en chimie...) et nous l'appelons "**niveau de Fermi**" ou "**énergie de Fermi**". Nous voyons alors immédiatement que quelle que soit la température:

$$f_{FD}(E_f) = \frac{1}{\exp\left(\frac{E_f - \mu}{kT}\right) + 1} = \frac{1}{\exp\left(\frac{\mu - \mu}{kT}\right) + 1} = \frac{1}{1+1} = \frac{1}{2} \quad (32.146)$$

la probabilité d'occupation par les électrons est donc de 1/2 dans ce niveau. La définition du niveau de Fermi peut alors être donnée par:

$$E_f - \mu = 0 \quad (32.147)$$

La principale application aux solides de cette statistique est la modélisation des phénomènes de transport électronique (gaz d'électrons): théorie des métaux, des semi-conducteurs, population des niveaux d'énergie et propriétés de conduction (cf. chapitre d'Electrocinétique).

3.4. DISTRIBUTION DE BOSE-EINSTEIN

Les bosons sont d'autres particules quantiques qui peuvent indistinctement se placer sur tous les niveaux d'énergie. Le principe de Pauli ne s'y applique donc pas! Dans ce cas, le nombre d'objet à permuter est $n_i + g_i - 1$ (les n_i particules et les $g_i - 1$ intervalles entre les niveaux). Parce que les particules sont indiscernables et les niveaux et sous niveaux permutable, il faut diviser par $n_i!$ puis aussi par $(g_i - 1)!$ Le nombre de configurations sur un niveau E_i , g_i fois dégénéré qui contient n_i particules indiscernables est ainsi égal à:

$$\Omega = \prod_i \frac{(n_i + g_i - 1)!}{n_i! (g_i - 1)!} \quad (32.148)$$

Ce qui peut se vérifier avec la figure ci-dessous:

6 états distincts	1	2	3	Exemple 2 particules : A et B 3 états : 1, 2 et 3 Indiscernables (A = B)
	AA	...	...	
	...	AA	...	
	...	...	AA	
	A	A	...	
	A	...	A	
	...	A	A	

Figure: 32.7 - États dans la distribution de Bose-Einstein

Son logarithme avec usage de la formule de Stirling et après simplification est donné par:

$$\ln(\Omega) = \sum_i [(n_i + g_i - 1) \ln(n_i + g_i - 1) - n_i \ln(n_i) - (g_i - 1) \ln(g_i - 1)] \quad (32.149)$$

Le maximum de $\ln(\Omega)$ correspond à annuler $d(\ln(\Omega))$, soit:

$$-d(\ln(\Omega)) = \sum_i [-\ln(n_i + g_i - 1) + \ln(n_i)] dn_i = 0 \quad (32.150)$$

Pour respecter les contraintes sur l'énergie et le nombre de particules, nous utilisons encore une fois la méthode des multiplicateurs de Lagrange:

$$-\ln(n_i + g_i - 1) + \ln(n_i) + \alpha + \beta E_i = 0 \quad (32.151)$$

et donc:

$$n_i = \frac{g_i}{e^{\alpha + \beta E_i} - 1} \quad (32.152)$$

qui est la distribution statistique de Bose-Einstein. Cette distribution diverge lorsque:

$$\alpha + \beta E_i = 0 \quad (32.153)$$

C'est la "**condensation de Bose-Einstein**". Dans cet état, tous les bosons se retrouvent dans le même état.

Les paramètres β et α jouent le même rôle que dans la distribution de Maxwell-Boltzmann et Fermi-Dirac. Nous avons ainsi:

$$\beta = \frac{1}{kT} \text{ et } \alpha = -\frac{\mu}{kT} \quad (32.154)$$

Le nombre (moyen) de bosons est alors donné par:

$$n_i = \frac{\xi_i}{\exp\left(\frac{E_i - \mu}{kT}\right) - 1} \quad (32.155)$$

à comparer avec le nombre (moyen) de fermions au même niveau d'énergie (fonction de Fermi-Dirac):

$$n_i = \frac{\xi_i}{\exp\left(\frac{E_i - \mu}{kT}\right) + 1} \quad (32.156)$$

Ainsi, en mécanique quantique, la statistique de Bose-Einstein désigne la distribution statistique de bosons indiscernables (tous similaires) et n'interagissant pas entre eux (sinon le modèle est trop complexe) sur les états d'énergie d'un système à l'équilibre thermodynamique.

Pour les systèmes macroscopiques, les niveaux d'énergie sont si serrés (ou tellement nombreux) que nous pouvons considérer le spectre d'énergies comme continu (approximation du continuum).

Nous raisonnerons donc dans ce contexte, ce qui nous permet d'écrire en normant aux nombres de particules mises en jeu (tout ce que l'on demande à la fonction c'est de nous dire comment sont réparties les N particules):

$$f_{BE}(E, T, \mu) = \frac{1}{\exp\left(\frac{E - \mu}{kT}\right) - 1} \quad (32.157)$$

pour la fonction de Bose-Einstein. Elle n'est donc définie que pour les énergies supérieures au potentiel chimique (sinon quoi elle est négative!).

De manière numérique, nous pouvons simuler l'évolution de l'allure de la distribution en fonction de l'énergie et de la température. Pour cela, nous supposons pour simplifier que la constante de Boltzmann vaut 1 et que le potentiel chimique $\mu=2$ (nous notons que celui-ci est dans l'absolu fonction de la température). Ainsi, nous pouvons écrire le petit programme suivant sous Matlab 5.0.0.473:

```
clear all; kb=1; % Constante de Boltzmann
mu=2; % Potentiel Chimique
T=0.001:0.1:1; % " Gradient " de température pour le programme
for j=1:length(T)
    beta(j)=1/(kb*T(j));
    epsilon=0.1:0.1:4;
    for i=1:length(epsilon)
        Nf(i,j)=1/(exp(beta(j)*(epsilon(i)-mu))-1); % Nb(epsilon,beta) moyen de bosons au
    end % niveau d'énergie epsilon
    hold on %
    plot(epsilon,Nf(:,j));
    pause(0)
end
```

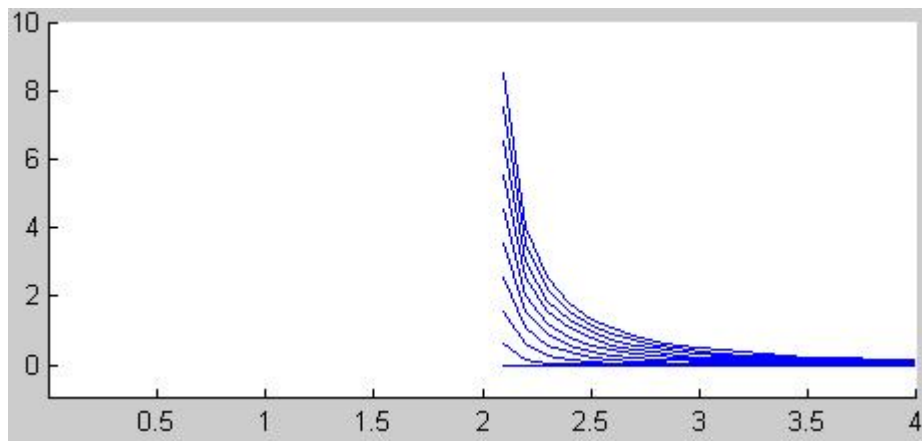


Figure: 32.8 - Tracé avec Matlab 5.0.0.473 de la distribution de Bose-Einstein

À haute température, lorsque les effets quantiques ne se font plus sentir, la statistique de Bose-Einstein, comme la statistique de Fermi-Dirac qui régit les fermions, tend vers la statistique de Maxwell-Boltzmann:

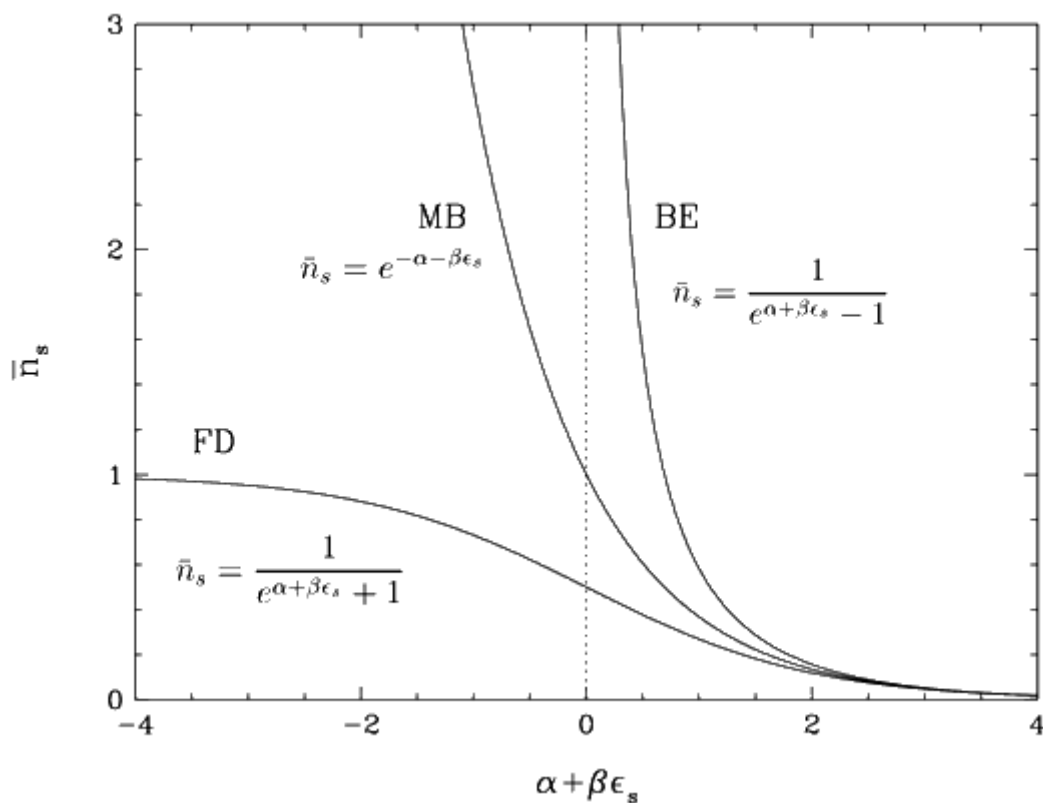


Figure: 32.9 - Exemple de tracé des trois distributions

Aux basses températures, cependant, les statistiques de Bose-Einstein et de Fermi-Dirac diffèrent entre elles. Nous nous plaçons, par exemple, à température nulle: dans la première, nous attendons alors que le niveau de plus basse énergie contienne tous les bosons, tandis que dans la seconde, les niveaux de plus basse énergie contiennent g_i fermions.

Par ailleurs, à température nulle (-273.15 [°C]), la statistique de Bose-Einstein montre de manière évidente que toutes les particules doivent occuper le même état quantique: celui de plus basse énergie. Ce phénomène est observable à l'échelle macroscopique et constitue un "**condensat de Bose-Einstein**".

La statistique de Bose-Einstein est utile à la compréhension des phénomènes électromagnétiques

ondulatoires car les photons sont des Bosons (rayonnement du corps noir, interaction matière/rayonnement). Elle est très largement utile à l'étude des phénomènes vibrationnels dans les solides (les phonons suivent la statistique de Bose-Einstein). Elle a aussi été utilisée pour expliquer les transitions de phase dans l'Hélium (phénomène à très basse température).

Remarques:

R1. La statistique de Bose-Einstein a été introduite par Satyendranath Bose en 1920 pour les photons et généralisée aux atomes par Albert Einstein en 1924.

R2. Un résultat mathématique appelé "théorème spin-statistique" relie le spin d'une particule et le type de statistique qu'elle suit. Il stipule que les particules de spin entier sont des bosons, alors que les particules de spin demi-entier sont des fermions. La démonstration de ce théorème ne se trouve pas encore sur le présent site Internet à ce jour.

Pour terminer, voici un résumé simplifié des choses qui peut éventuellement aider à une meilleure compréhension:

Particules	Statistiques	Particularités
Bosons (photons, gluons...)	Statistique de Bose-Einstein $n_i = \frac{g_i}{\exp\left(\frac{E_i - \mu}{kT}\right) - 1}$	Particules indiscernables qui n'interagissent pas. Aucune contrainte sur le nombre de particules par état
Fermions (électrons, protons, neutrinos...)	Statistique de Fermi-Dirac $n_i = \frac{g_i}{\exp\left(\frac{E_i - \mu}{kT}\right) + 1}$	Particules indiscernables qui n'interagissent pas. Le nombre de particules par état est égal à 0 ou 1
Classiques	Statistique de Maxwell-Boltzmann $n_i = g_i \exp\left(\frac{\mu - E_i}{kT}\right)$	Particules discernables qui n'interagissent pas. Aucune contrainte sur le nombre de particules par état

Tableau: 32.1 - Similitudes des différentes distributions quantiques

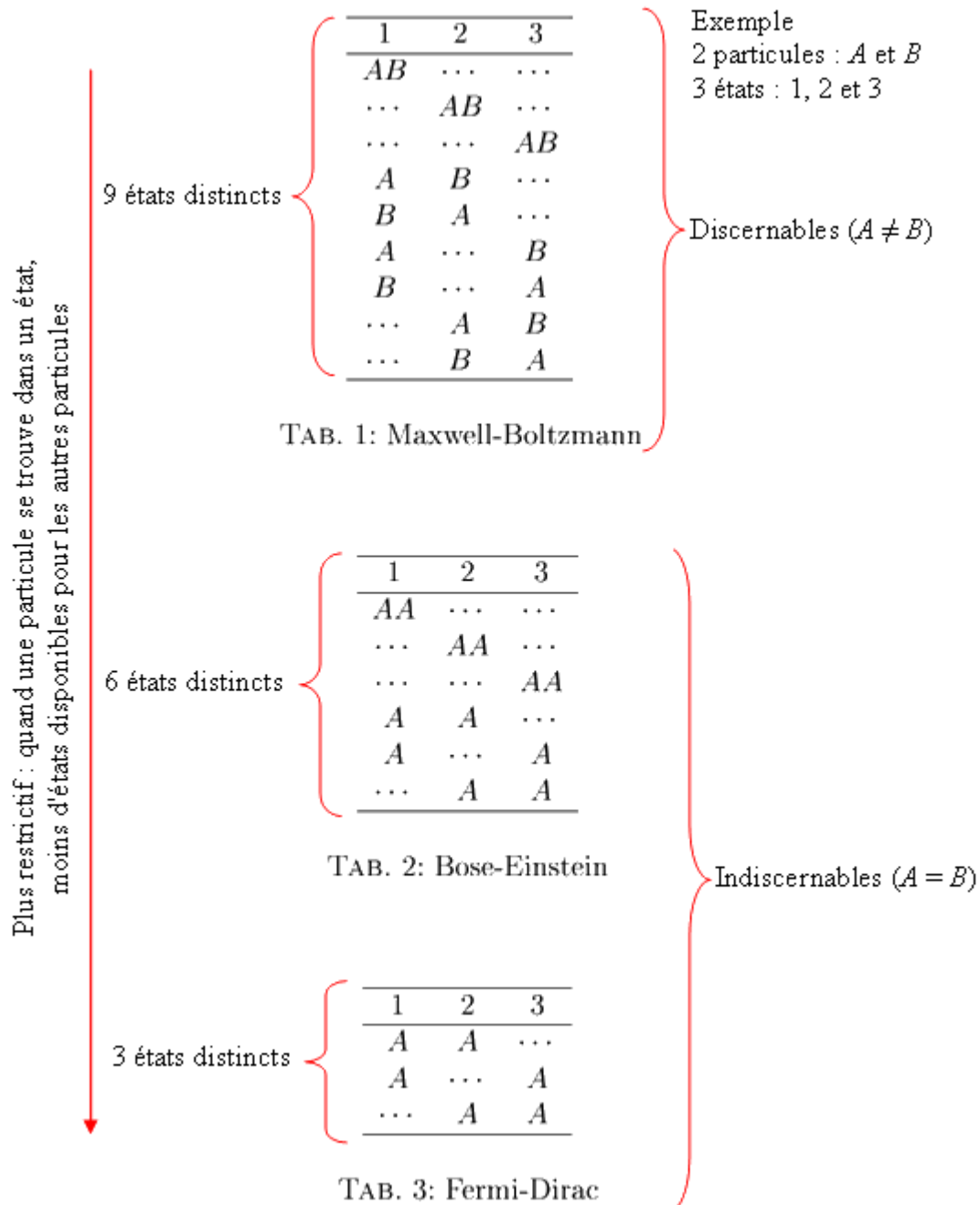


Figure: 32.10 - Résumé des états dans les trois distributions

4. LOI DE FICK

Nous avons vu dans le chapitre de Thermodynamique la démonstration de l'équation de propagation de la chaleur proposée par Fourier en 1822 obtenue à partir de l'équation de continuité. Nous avons obtenu (il est très recommandé au lecteur de s'y référer à nouveau ne serait-ce que pour lire les remarques relatives à la démonstration):

$$\frac{\partial T(x,t)}{\partial t} - D_T \frac{\partial^2 T(x,t)}{\partial x^2} = 0 \quad (32.158)$$

En se basant sur les mêmes hypothèses que Fourier, Fick proposa en 1855 qu'un flux de particules pourrait se diffuser à travers un matériau selon une loi similaire, la "**deuxième loi de Fick**", de la forme:

$$\frac{\partial \rho(x,t)}{\partial t} - D_p \frac{\partial^2 \rho(x,t)}{\partial x^2} = 0 \quad (32.159)$$

où la constante de proportionnalité est le "**coefficient de diffusion de la matière**" et ρ la densité de particules par unité de volume (et non la densité de masse!).

Remarque: En pratique, la diffusion joue un rôle essentiel dans la fabrication de céramique, de semi-conducteurs (dopage), de cellules-solaires et dans la solidification des métaux (traitement au carbone et à la chaleur). Car lorsque deux matériaux chauffés sont mis en contact, leurs atomes diffusent l'un dans l'autre.

Il faut comprendre que tout diffuse dans tout! Donc pensez aux pesticides sur les fruits et légumes, à la pollution dans les nappes phréatiques, au PET dans les boissons...

Dès lors, la relation du flux surfacique de chaleur que nous avons utilisée en thermodynamique (voir chapitre du même nom) pour obtenir la loi de Fourier et qui était:

$$\vec{\Phi}_q = -\kappa \vec{\nabla} \cdot T \quad (32.160)$$

peut alors s'écrire certainement aussi (nous allons le démontrer) dans le cas de la masse sous la forme d'un flux surfacique de particules appelé "**première loi de Fick**":

$$\vec{\Phi}_p = -D \vec{\nabla} \cdot \rho \quad (32.161)$$

où D est le coefficient de transport de la matière (à déterminer...). Nous retrouvons dans la littérature cette dernière relation fréquemment sous forme unidimensionnelle. Il vient alors:

$$\Phi_p = -D \frac{\partial \rho}{\partial x} \quad (32.162)$$

Soit sous forme discrète:

$$\Phi_p = -D \frac{\Delta \rho}{\Delta x} \quad (32.163)$$

Si la variation de distance est mis en correspondant avec une longueur L parcourue par la diffusion, nous avons alors la notation encore plus simplifiée:

$$\Phi_p = -D \frac{\Delta \rho}{L} \quad (32.164)$$

Pour retrouver la relation connue par les petites classes de chimie il faut savoir que les chimiste on pour habitude de noter la densité par la lettre C pour indiquer qu'il s'agit d'une Concentration. Dès lors, nous avons:

$$\Phi_p = -D \frac{\Delta C}{L} \quad (32.165)$$

Mais ce n'est pas le seul changement de notation fait par les chimistes. Effectivement il faut se rappeler que le flux surfacique de particules (masse) est défini par la quantité de masse passant par une surface constante par unité de temps. Nous avons alors:

$$\frac{\frac{\Delta m}{\Delta t}}{S} = -D \frac{\Delta C}{L} \quad (32.166)$$

Comme l'origine du temps est souvent pris comme étant zéro et que la masse initiale est nulle il vient alors après réarrangement la relation classique finale connue par les chimistes et biologistes (au signe près puisque ce n'est qu'une question de convention):

$$m = - \frac{DS\Delta Ct}{L} \quad (32.167)$$

Exemple:

Les plantes absorbent de l'eau pour la photosynthèse. Nous savons par mesures expérimentales que la coefficient de diffusion de l'eau dans l'air est de $D \cong 2.4 \cdot 10^{-5} \left[m^2 \cdot s^{-1} \right]$. Les pores d'absorption des plantes ont typiquement une section de l'ordre de $S \cong 8.0 \cdot 10^{-11} \left[m^2 \right]$. La distance de diffusion est de l'ordre de $L \cong 2.5 \cdot 10^{-5} \left[m \right]$. La densité de vapeur d'eau à l'intérieur de la plante est de l'ordre de $C_2 \cong 0.022 \left[kg \cdot m^{-3} \right]$ et à l'extérieur $C_1 \cong 0.011 \left[kg \cdot m^{-3} \right]$. Dès lors, la masse d'eau absorbée en une heure est d'environ:

$$m = \frac{DS\Delta Ct}{L} \cong \frac{2.4 \cdot 10^{-5} \cdot 8 \cdot 10^{-11} \cdot (0.022 - 0.011) \cdot 3'600}{2.5 \cdot 10^{-5}} \cong 3 \cdot 10^{-9} \left[kg \right] \quad (32.168)$$

Résultat qui convient d'être multiplié par les millions de pores qu'a une plante.

Remarque: Au fait, Fick démontra d'abord la première loi et en procédant en tous points de manière identique à l'équation de la chaleur il obtint la deuxième loi qui porte son nom.

Les coefficients κ, D sont appelés globalement "**coefficients de transport**" et respectivement "**coefficient de diffusion thermique**" dans le domaine de la chaleur et "**coefficient de diffusion**" dans le domaine de la matière.

Nous pouvons estimer les valeurs de ces coefficients à l'aide d'un modèle microscopique simple.

Considérons pour cela une tranche de fluide (flux de chaleur et flux de masse sont considérés comme un fluide) perpendiculaire à l'axe des x et d'épaisseur $2\lambda_x$ où λ_x correspond au libre parcours moyen

(projeté selon x), dans laquelle existe un gradient de concentration dirigé selon l'axe x . Déterminons le courant de ce gradient à travers la section S d'abscisse x .

Pour faire simple, nous pouvons considérer que, parmi toutes les particules se trouvant entre l'abscisse $x - \lambda_x$ et x , un tiers, ont leur vitesse dirigée selon x (les deux autres tiers étant sur y et z), et parmi ces dernières, la moitié a une vitesse positive (finalement nous devons considérer le $1/6$ par direction).

Comme λ_x est le libre parcours moyen, ces dernières particules franchiront la section S sans avoir subi de collision: elles participeront donc au courant de diffusion.

Notant $\rho(x - \lambda_x)$ la concentration volumique à l'abscisse $x - \lambda_x$ (et considérant que cette concentration est constante entre $x - \lambda_x$ et x , ce qui, vu l'ordre de grandeur de λ_x , est à peu près vérifié), le nombre de particules se trouvant entre les abscisses $x - \lambda_x$ et x et traversant effectivement la section S vaut alors:

$$N = \frac{\rho(x - \lambda_x) S \lambda_x}{6} \quad (32.169)$$

Cette traversée prend un temps égal à $\lambda_x / \bar{v}$, où $\bar{v}$ est la vitesse moyenne d'agitation thermique. Par conséquent, la densité de courant circulant de la gauche vers la droite vaut:

$$\Phi_{1p} = \frac{\rho(x - \lambda_x) \lambda_x}{6} \frac{\bar{v}}{\lambda_x} = \frac{\rho(x - \lambda_x) \bar{v}}{6} \quad (32.170)$$

En procédant de la même manière pour les particules se trouvant à droite de x , nous obtenons pour la densité de courant circulant de droite à gauche:

$$\Phi_{2p} = \frac{\rho(x + \lambda_x) \bar{v}}{6} \quad (32.171)$$

La densité de courant totale circulant à travers S vaut donc:

$$\Phi_p = \frac{\rho(x - \lambda_x) \bar{v}}{6} - \frac{\rho(x + \lambda_x) \bar{v}}{6} = \frac{\rho(x - \lambda_x) - \rho(x + \lambda_x)}{6} \bar{v} \quad (32.172)$$

Or, nous pouvons aussi écrire cela sous la forme:

$$\Phi_p = \frac{2(\rho(x - \lambda_x) - \rho(x))}{6} \bar{v} = \frac{\rho(x - \lambda_x) - \rho(x)}{3} \bar{v} \quad (32.173)$$

Si λ_x est très petit, nous pouvons écrire:

$$\Phi_p = \frac{\rho(x - \lambda_x) - \rho(x)}{3} \bar{v} \cong -\frac{1}{3} \lambda_x \bar{v} \frac{\partial \rho}{\partial x} \quad (32.174)$$

Vu les simplifications apportées au modèle, le facteur 3 a toutes les chances d'être peu réaliste. En revanche, la relation de proportionnalité entre gradient de concentration et courant de diffusion est tout à fait crédible, Nous écrirons finalement en généralisant à l'espace:

$$\vec{\Phi}_p = -D \vec{\nabla} \cdot \rho \quad (32.175)$$

où D alors donnée par:

$$D \cong \lambda \bar{v} \quad (32.176)$$

est la constante de diffusion massique. Comme D est positive, nous constatons que le mouvement de diffusion des particules a lieu dans le sens opposé au gradient, ce qui tend bien à homogénéiser les concentrations.

Remarque: Si nous souhaitons obtenir le flux de charge, il suffit de multiplier la relation obtenue à gauche et à droite par la charge élémentaire.

Nous pouvons également estimer le flux d'énergie thermique transportée par ces mêmes particules selon x . En effet, dans chaque tranche de fluide, n particules transportent chacune une énergie E correspondant à une quantité de chaleur Q donnée (selon la loi de Joule). Nous avons donc un flux surfacique d'énergie Φ_q dont la première composante est donnée par le même type de bilan que les développements précédents:

$$\Phi_q \cong -\frac{1}{3} \frac{\partial(nQ)}{\partial x} \lambda_x \bar{v} = -\frac{1}{3} \lambda_x \bar{v} \frac{\partial(nQ)}{\partial T} \frac{\partial T}{\partial x} \quad (32.177)$$

Nous y trouvons immédiatement la définition de la capacité calorifique (si nous divisons par la masse nous aurions la capacité calorifique massique selon ce que nous avons vu dans le chapitre de Thermodynamique). Ainsi, dans le cas unidimensionnel:

$$\kappa = \frac{1}{3} \lambda_x \bar{v} C = C \cdot D \quad (32.178)$$

Il y a donc un simple rapport de proportionnalité entre κ et C .

Remarque: Selon les auteurs le flux est noté avec le symbole de la densité de courant, soit $\vec{j}$.

Faisons un petit tableau récapitulatif pour les quelques lois de diffusion démontrées jusqu'à maintenant sur ce site (dans leurs chapitres respectifs) en utilisant la notation la plus courante en physique (et non celle des thermodynamiciens...):

Loi de Fourier	Loi de Fick	Loi d'Ohm
Thermodynamique	Mécanique Statistique	Électrocinétique
$\vec{j}_q = -\kappa \vec{\nabla} \cdot T$	$\vec{j}_p = -D \vec{\nabla} \cdot \rho$	$\vec{j} = \sigma \vec{E} = -\sigma \vec{\nabla} \cdot U$
Densité de courant thermique T : température	Densité de courant particulaire ρ : concentration	Densité de courant électrique U : potentiel électrique
κ : conductivité thermique	D : coefficient de diffusion	σ : conductivité électrique
flux thermique: $\Phi_q = \iint_S \vec{j}_q \circ d\vec{S}$	flux de particules: $\Phi_p = \iint_S \vec{j}_p \circ d\vec{S}$	flux de courant électrique: $I = \iint_S \vec{j} \circ d\vec{S}$

Tableau: 32.2 - Similitudes des différentes lois de diffusion en physique

5. MOUVEMENT BROWNIEN

Le mouvement brownien est une description mathématique du mouvement aléatoire d'une petite impureté organique ou non immergée dans un fluide et qui n'est soumise à aucune autre interaction que des chocs avec les petites molécules du fluide environnant. Il en résulte un mouvement très irrégulier de la grosse particule, qui a été décrit pour la première fois en 1827 par le botaniste Robert Brown en observant des mouvements de particules à l'intérieur de grains de pollen:

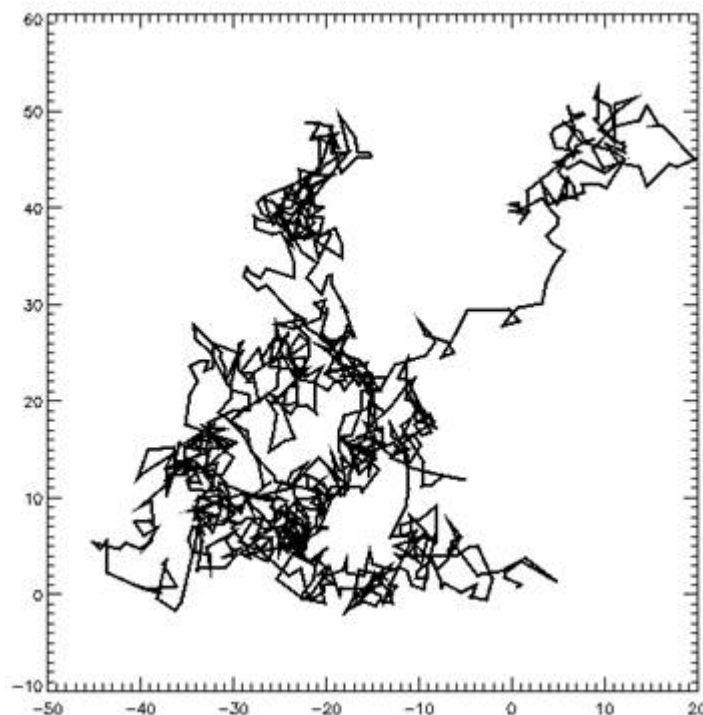


Figure: 32.11 -Exemple de mouvement Brownien (source: Wikipédia)

L'origine du mouvement dû aux molécules n'était pas du tout évidente au début du 19ème siècle car:

1. Il n'était pas encore communément admis que la matière était simplement discontinue et donc composée de molécules.

2. On ne comprenait pas même en admettant l'aspect moléculaire de la matière que quelques molécules soient capables de déplacer des impuretés plusieurs millions de fois plus grandes que celles des molécules du fluide.

3. Même si on admettait l'aspect moléculaire, on ne comprenait pas qu'une très grande quantité de chocs de milliards de molécules ne s'annulent pas entre eux et que l'impureté soit finalement immobile (on verra que cela est dû au fait que la matière n'est pas continue).

Le traitement mathématique du problème utilisant l'aspect moléculaire et statistique de la matière a été une affirmation pour les scientifiques partisans de l'aspect atomique de la matière et ouvrit encore une fois la porte au nouveau domaine de la mécanique statistique déjà un peu utilisée à l'époque par Maxwell dans le cadre des gaz.

Comme toujours il existe plusieurs modèles mathématiques et suivant la tradition du présent site Internet nous avons choisi la plus simple et... ce n'est pas celle d'Einstein mais celle de Langevin (effectuée deux ans après dans un souci de simplification).

Le point de départ est le théorème d'équipartition de l'énergie cinétique (cf. chapitre de Mécanique des Milieux Continus):

$$E_c = N \frac{1}{2} m \overline{v^2} = \frac{3}{2} \frac{R}{N_A} N T \quad (32.179)$$

soit pour une particule:

$$E_c = \frac{1}{2} m \overline{v^2} = \frac{3}{2} \frac{RT}{N_A} \quad (32.180)$$

avec pour rappel v qui est la vitesse moyenne! Si nous simplifions l'analyse à un seul axe de translation possible la relation devient (nous spécifions maintenant qu'il s'agit de la moyenne afin de ne pas nous mélanger les pinceaux avec la suite):

$$\langle E_c \rangle = \frac{1}{2} m \langle v_x^2 \rangle = \frac{1}{2} \frac{RT}{N_A} \quad (32.181)$$

En écrivant ceci nous stipulons donc qu'une particule en suspension dans un fluide en équilibre thermique possède, dans la direction x par exemple, une énergie cinétique moyenne égale à celle d'une molécule gazeuse de nature quelconque, dans une direction donnée, à la même température. Il s'agit donc à nouveau d'une identité forte entre les solutions diluées et les gaz parfaits.

Une particule comme celle que nous considérons, grande par rapport aux molécules du liquide, et se mouvant à la vitesse v par rapport à celui-ci subit comme nous l'avons pseudo-démontré dans le chapitre de Mécanique Des Milieux Continus (loi de Stokes) une résistance de:

$$F_{st.} = -6\pi\mu Rv = -\xi v \quad (32.182)$$

où ξ sera appelé "**coefficient de friction**" (correspondant donc à la viscosité) et où R est le rayon de la sphère.

Maintenant, écrivons selon l'équation de la dynamique de Newton (on ne spécifie plus que la vitesse est

selon x dans la notation qui suit):

$$m \frac{dv}{dt} = m \frac{d^2 x}{dt^2} = -6\pi\eta R v + F_{alea.} \quad (32.183)$$

La force complémentaire $F_{alea.}$, introduite par Langevin, est aléatoire (stochastique). On sait a priori peu de choses d'elle, si ce n'est qu'elle semble a priori indifféremment positive ou négative, et que sa grandeur est telle qu'elle maintient l'agitation de la particule, qui, sans elle, finirait par s'arrêter sous l'effet de la résistance visqueuse.

L'équation précédente multipliée par x , peut s'écrire encore:

$$mx \frac{dv}{dt} = -\xi xv + xF_{alea.} \quad (32.184)$$

Elle est alors appelée "**équation de Langevin**" ou "**équation stochastique de Langevin**".

Or, rappelons que (cf. chapitre de Calcul Différentiel Et Intégral):

$$d(xv) = xdv + vdx \Leftrightarrow \frac{d(xv)}{dt} = x \frac{dv}{dt} + v \frac{dx}{dt} \quad (32.185)$$

d'où:

$$x \frac{dv}{dt} = \frac{d(xv)}{dt} - v \frac{dx}{dt} = \frac{d(xv)}{dt} - v^2 = \frac{d\left(x \frac{dx}{dt}\right)}{dt} - v^2 = \frac{d\left(\frac{1}{2} \frac{dx^2}{dt}\right)}{dt} - v^2 = \frac{1}{2} \frac{d^2 x^2}{dt^2} - v^2 \quad (32.186)$$

En établissant cette dernière relation, nous avons aussi montré que:

$$xv = \frac{1}{2} \frac{dx^2}{dt} \quad (32.187)$$

Nous pouvons alors écrire:

$$\frac{1}{2} m \frac{d^2 x^2}{dt^2} - mv^2 = -\xi xv + xF_{alea.} \quad (32.188)$$

et aussi:

$$\frac{1}{2} m \frac{d^2 x^2}{dt^2} - mv^2 = -\xi \frac{1}{2} \frac{dx^2}{dt} + xF_{alea.} \quad (32.189)$$

Si nous prenons la moyenne (nous changeons de notation pour la moyenne car la barre n'est pas très adaptée esthétiquement parlant lorsqu'il y a des puissances):

$$\frac{1}{2} m \frac{d^2 \langle x^2 \rangle}{dt^2} - m \langle v^2 \rangle = -\xi \frac{1}{2} \frac{d \langle x^2 \rangle}{dt} + \langle xF_{alea.} \rangle \quad (32.190)$$

Nous posons maintenant que:

$$\langle xF_{alea.} \rangle = 0 \quad (32.191)$$

en d'autres termes nous pouvons voir cela comme le travail moyen de la force aléatoire qui est nul.

Il reste alors:

$$\frac{1}{2}m \frac{d^2 \langle x^2 \rangle}{dt^2} - m \langle v^2 \rangle = -\xi \frac{1}{2} \frac{d \langle x^2 \rangle}{dt} \quad (32.192)$$

Posons:

$$u = \frac{1}{2} \frac{d \langle x^2 \rangle}{dt} \quad (32.193)$$

Nous avons alors:

$$m \frac{du}{dt} - m \langle v^2 \rangle = -\xi u \quad (32.194)$$

Et comme:

$$m \langle v^2 \rangle = \frac{RT}{N_A} \quad (32.195)$$

Donc:

$$m \frac{du}{dt} - \frac{RT}{N_A} = -\xi u \quad (32.196)$$

Soit autrement écrit:

$$m \frac{du}{dt} + \xi u = \frac{RT}{N_A} \quad (32.197)$$

Il s'agit donc d'une équation différentielle à coefficients constants d'ordre 1. Nous avons démontré dans le chapitre de Calcul Différentiel Et Intégral que la solution homogène était:

$$u_h = C e^{-\frac{\xi}{m}t} \quad (32.198)$$

où la valeur de C nous importe peu car ce terme va de toute façon disparaître dans les développements qui vont suivre. La solution particulière peut être déterminée simplement par la contrainte au temps zéro:

$$m \frac{d \left(C e^{-\frac{\xi}{m} t} + u_p \right)}{dt} + \xi \left(C e^{-\frac{\xi}{m} t} + u_p \right) = \frac{RT}{N_A} \quad (32.199)$$

Soit après dérivation et avoir posé $t = 0$ donne:

$$m \frac{du_p}{dt} + \xi u_p = \frac{RT}{N_A} \quad (32.200)$$

Une solution triviale simple est alors:

$$u_p = c^{te} = \frac{RT}{\xi N_A} \quad (32.201)$$

Nous avons alors:

$$u = \frac{1}{2} \frac{d \langle x^2 \rangle}{dt} = u_h + u_p = \frac{RT}{\xi N_A} + C e^{-\frac{\xi}{m} t} \quad (32.202)$$

Or après très peu de temps le terme en exponentielle devient quasiment nul. Nous avons donc en régime permanent:

$$\frac{1}{2} \frac{d \langle x^2 \rangle}{dt} = \frac{RT}{\xi N_A} \quad (32.203)$$

Soit:

$$\boxed{\langle x^2 \rangle = \frac{2RT}{\xi N_A} t = 2Dt} \quad (32.204)$$

où D est appelé "**coefficient de diffusion**" (comme en thermodynamique). Nous remarquons dès lors que la position moyenne de la distance x parcourue est in extenso proportionnelle à la racine carrée du temps. Einstein avait appliqué un exemple numérique avec des particules en suspensions d'un rayon d'un millièème de millimètre avec une eau à 290 [K] et trouva ainsi un déplacement moyen de six millièmes de millimètre en une minute.

La relation (où il convient de ne pas confondre le R du rayon de la particule en suspensions au dénominateur avec le R de la constante des gaz parfaits au numérateur):

$$\boxed{D = \frac{RT}{\xi N_A} = \frac{RT}{N_A} \frac{1}{6\pi\eta R}} \quad (32.205)$$

est parfois appelée "**relation de Sutherland-Einstein**" (car William Sutherland l'a découverte en Australie pratiquement au moment où Einstein écrivait cette même relation dans sa thèse) et que l'on retrouve aussi fréquemment sous la forme suivante (où k est pour rappel la constante de Boltzmann):

$$D = \frac{RT}{\zeta N_A} = \frac{kT}{\zeta} \quad (32.206)$$

Einstein avait ainsi trouvé une relation qui offrait la possibilité de calculer le rayon R des atomes avec un thermomètre, un microscope et un chronomètre pour seuls instruments. Le physicien Jean Perrin reçut le prix Nobel pour la vérification expérimentale de ce résultat.

Le lecteur remarquera, que si nous avions faire les mêmes développements que précédemment, mais en divisant à gauche et à droite l'équation différentielle:

$$m \frac{du_p}{dt} + \xi u_p = \frac{RT}{N_A} \quad (32.207)$$

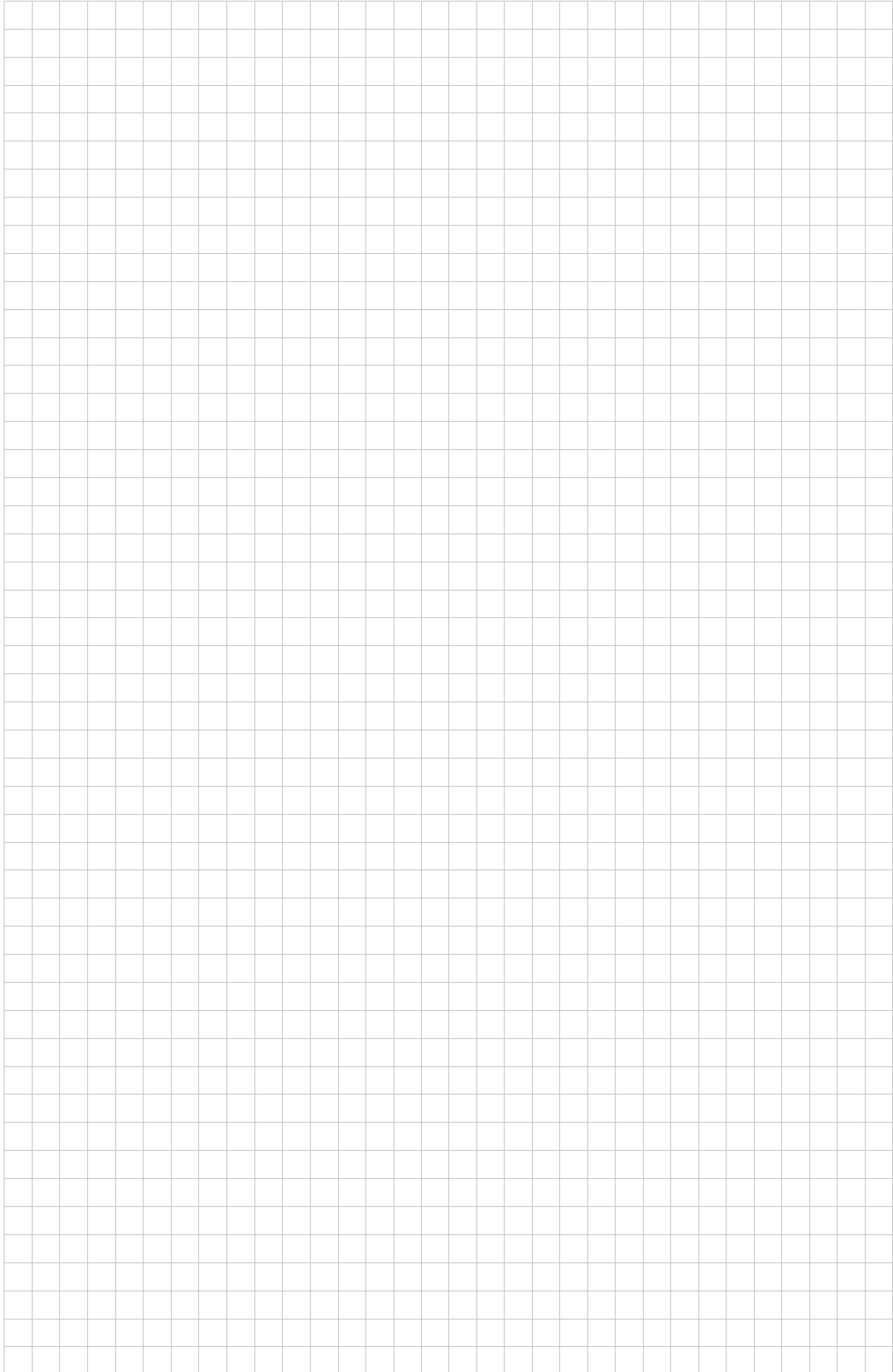
par m nous aurions eu comme résultat (considéré à juste titre comme plus général):

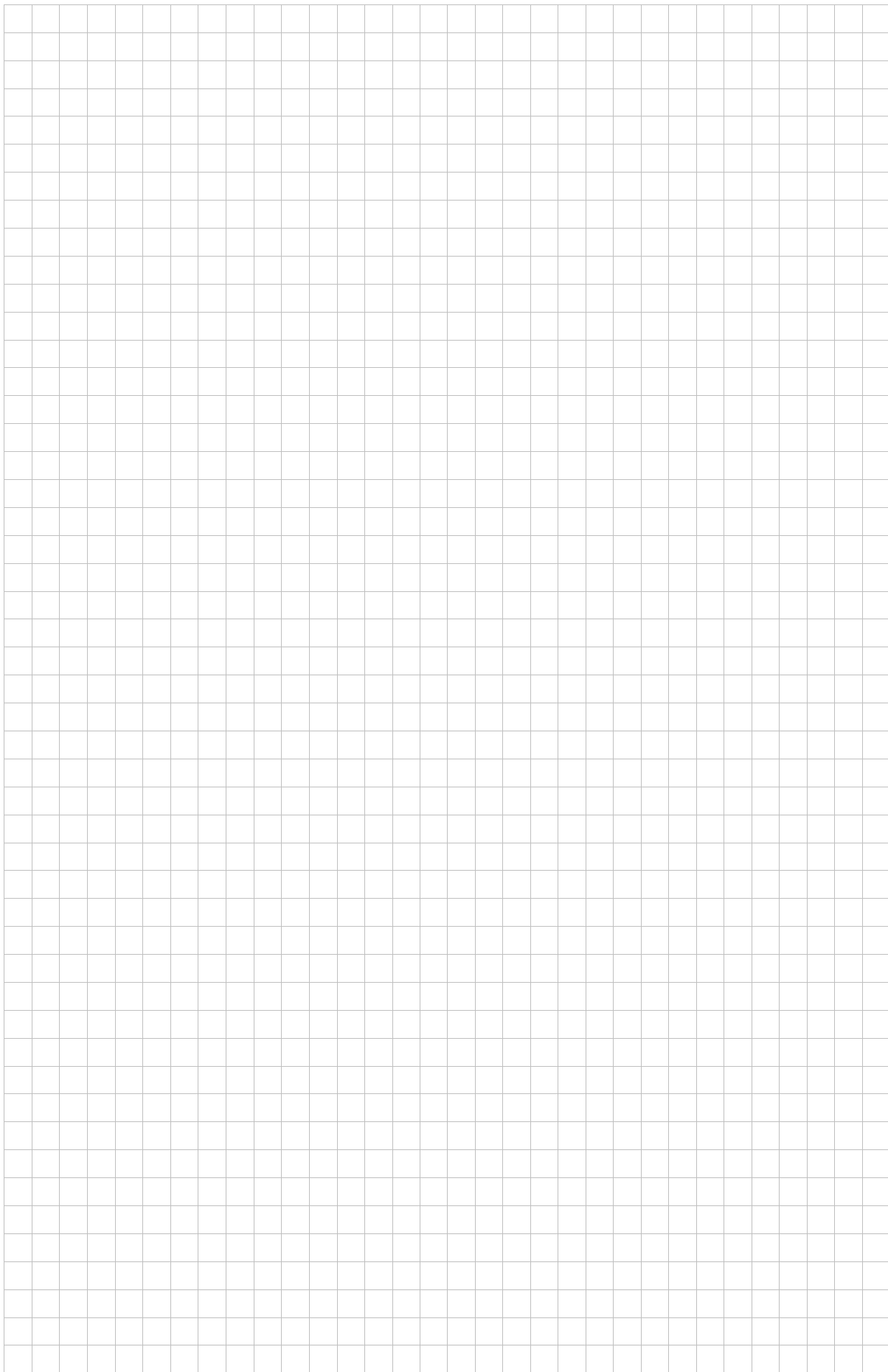
$$D = \frac{kT}{\zeta m} \quad (32.208)$$

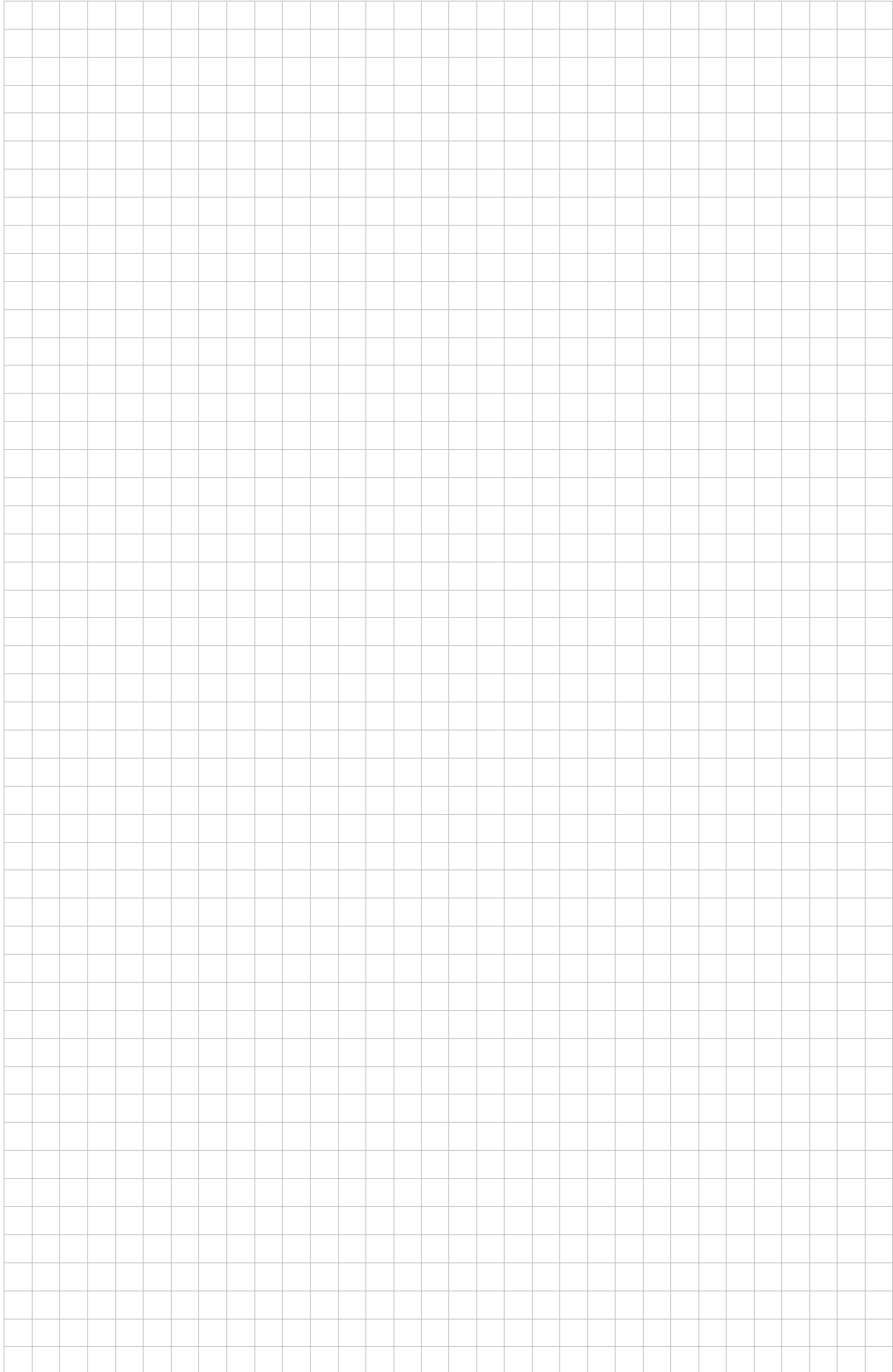
La méthode de Langevin redonne le même résultat que le premier modèle proposé par Einstein deux ans plus tôt et que le deuxième modèle deux ans plus tard.

On peut se demander jusqu'à quel point la relation de Sutherland-Einstein prouve l'existence de molécules. Autrement dit, que serait la limite du coefficient de diffusion D si la nature était continue, c'est-à-dire si le nombre d'Avogadro était infini?

Nous pressentons bien évidemment alors que D s'annulerait, et que le déplacement de diffusion brownien disparaîtrait tout simplement dans cette limite. Autrement dit, le mouvement brownien cesserait immédiatement si la Nature était continue! Ce qui est un résultat remarquable de preuve de l'aspect discret de la nature!







33. THERMODYNAMIQUE

La "thermodynamique" est la partie de la physique qui traite des relations permettant de déterminer formellement les échanges (variations) d'énergie sous forme de travail mécanique et de chaleur dans le cadre de l'étude des transformations des 4 états de la matière (mais principalement des gaz parfaits dans le cadre scolaire) sous la base d'hypothèses simplificatrices entre un système (isolé, ouvert ou fermé) et son environnement extérieur.

Les objectifs principaux de la thermodynamique sont:

1. Avec un minimum de variables de pouvoir déterminer l'état et les échanges énergétiques d'un système sous des contraintes prédéfinies et souvent considérées comme idéales... (il y a à peu près une différence de 5% entre les valeurs théoriques et celles qui sont mesurées) et entre états d'équilibre et donc sans faire intervenir le temps
2. De trouver les "variables d'état" (définissant l'état du système étudié à l'équilibre thermodynamique), telles que ces différentes informations puissent être obtenues en ne connaissant dans l'idéal que l'état final et initial du système.
3. De se débrouiller à ramener les équations toujours à une forme mettant en évidence des variables (variations) facilement mesurables dans la pratique.

Nous verrons plus loin que tout système peut au point de vue énergétique être décrit par:

- Son volume, sa masse, sa pression, sa température,...
- Son énergie potentielle, son énergie cinétique, son potentiel chimique,...
- Ses propriétés physiques comme la capacité à absorber la chaleur, à irradier, ...

Attention!! Bien que la puissance prédictive théorique des équations de la thermodynamique soit passionnante à mettre en pratique, cette théorie est au final malheureusement (et forcément) une jungle d'équations à cause du nombre de combinaisons possibles des quelques paramètres que nous pouvons choisir à loisir comme fixes ou variables dans la pratique. Alors courage et n'oubliez pas que c'est en forgeant qu'on devient forgeron!

Remarque: Il est très difficile de présenter la thermodynamique d'une manière pédagogique purement linéaire au contraire des autres domaines de la physique. Le plan de ce chapitre est donc actuellement un vrai méli-mélo de notions qui seront parfois définies préalablement et d'autres qui ne le sont que bien plus tard dans le texte. Le lecteur aura donc compris que ce chapitre nécessite dans sa forme actuelle une restructuration majeure pour arriver à maturité et cela sera fait dans un avenir plus ou moins longtemps...

1. VARIABLES THERMODYNAMIQUES

En thermodynamique, il existe plusieurs concepts majeurs très utiles suivant le type de système maintenu ou laissé en libre évolution étudié et les moyens de mesure à disposition.

Nous souhaitons ici donner les définitions les plus importantes et nous démontrerons si possible leur provenance plus tard sachant qu'il est très difficile d'avoir une présentation purement linéaire de la thermodynamique (comme déjà mentionné!).

Définitions:

D1. Un "système" est un corps ou un groupe de corps subissant ou non des évolutions et qui constitue un ensemble bien défini et bien délimité dans l'espace et entouré par le milieu extérieur. L'ensemble système/milieu extérieur est dénommé "univers".

D2. Le "**travail**", noté W , est l'énergie associée à la valeur ou au changement de la dynamique d'un système par l'agissement de forces mécaniques diverses (d'où le choix du terme "travail...").

D3. La "**chaleur**" ou "**énergie calorifique**", notée Q , est l'énergie associée à la valeur ou au changement de la dynamique d'un système dû à l'agitation moyenne des molécules (énergie cinétique moyenne).

D4. "**L'énergie totale**" d'un système, notée E , est la somme de toutes les énergies qui spécifient ce système par rapport à son centre de masse (moment d'inertie, de masse, énergie cinétique interne,...) et aussi (!) par rapport à un référentiel extérieur (énergie cinétique, énergie potentielle, rayonnement entrant).

D5. "L'énergie externe" qui est liée à la position du système comme l'énergie cinétique ou l'énergie potentielle (gravitationnelle/électrostatique).

D6. "**L'énergie interne**" d'un système, notée U , est la somme de tous les types d'énergies internes qui le distinguent uniquement par rapport à son centre de masse tels que le travail W échangé avec l'extérieur, la chaleur Q échangée (énergie cinétique moyenne interne), l'énergie de masse (relativiste, nucléaire), le moment d'inertie...

De nombreux ouvrage de vulgarisation montrent qu'à ce point de notre étude de la physique nous avons typiquement comme énergie totale d'un corps:

$$E_{\text{tot}} = \underbrace{\frac{1}{2}mv^2}_{\text{énergie cinétique}} + \underbrace{\frac{1}{2}J\omega^2}_{\text{énergie de rotation}} + \underbrace{mgh}_{\text{énergie gravitationnelle}} + \underbrace{\frac{1}{2}kx^2}_{\text{énergie potentielle élastique}} + \underbrace{m_0c^2}_{\text{énergie de masse}} + \dots \quad (33.1)$$

D7. Les "**énergies échangées**" qui sont donc les énergies qui sont échangées avec l'extérieur. Les plus courantes sont le travail mécanique W et la chaleur Q précédemment définies.

D8. "**L'entropie**", notée S , permet de quantifier la qualité et le sens d'évolution de l'énergie d'un système. Nous démontrerons que l'entropie d'un système isolé ne peut que croître.

D9. "**L'enthalpie**", notée H , est la quantité de chaleur, reçue par un système qui évolue à pression constante (isobare). En chimie ce concept est très utile, car dans une transformation chimique une partie de l'énergie injectée aura juste servi à repousser l'atmosphère ambiante lors du changement de volume dans la transformation. Ainsi, l'enthalpie rajoute à l'énergie interne U un terme correctif prenant en compte l'énergie emmagasinée/perdue par la pression environnante qui compresse le système (s'il est compressible...).

D10. "**L'énergie libre**" ou "**énergie de Helmholtz**", notée F , caractérise la fraction d'énergie interne utilisable sous forme de travail. Elle est simplement la différence entre l'énergie interne et l'énergie calorifique dissipée à cause de l'entropie à une température donnée.

D11. "**L'enthalpie libre**" ou "**énergie libre de Gibbs**", notée G , caractérise la fraction d'enthalpie disponible sous forme de travail. Elle est simplement la différence entre l'enthalpie et l'énergie calorifique dissipée à cause de l'entropie à une température donnée.

Pour résumer, dans l'ordre d'apparition et d'importance nous avons les types d'énergie suivant:

Symbole	Légende
W	travail
Q	chaleur
P	pression
T	température absolue
V	volume
E	énergie totale
U	énergie interne
S	entropie
H	enthalpie
F	énergie libre
G	enthalpie libre

Tableau: 33.1 - Différentes grandeurs thermodynamiques

Indiquons que nous ne reviendrons pas sur le concept de température T qui est comme nous l'avons vu en Mécanique Des Milieux Continus (théorème du Viriel) et en Mécanique Statistique (rayonnement du corps noir), un paramètre qui permet de relier avantageusement le mouvement moyen de différents corps avec leur énergie cinétique moyenne (sous-entendu leur "excitation" ou "désordre") ou respectivement le rayonnement de certains corps avec leur énergie d'émission. Mais rappelons quand même l'origine du zéro Kelvin, car c'est une question redondante sur le web:

Nous avons démontré dans le chapitre de Mécanique des Milieux Continus qu'à pression P constante (système "isobare"), le volume d'une quantité fixée de gaz parfait est proportionnel à la température absolue. C'est la "**loi de Gay-Lussac**" (dans le cas des gaz parfaits...!):

$$V = c^t T \quad (33.2)$$

et nous y avons alors mentionné que les mesures expérimentales donnent alors une droite, qui extrapolée (un peu brutalement...) dans les valeurs négatives de la température en Celsius donne un volume nul pour un gaz parfait (en négligeant les aspects quantiques à cette frontière...) à une température systématiquement de $-273.15\text{ }^{\circ}\text{C}$:

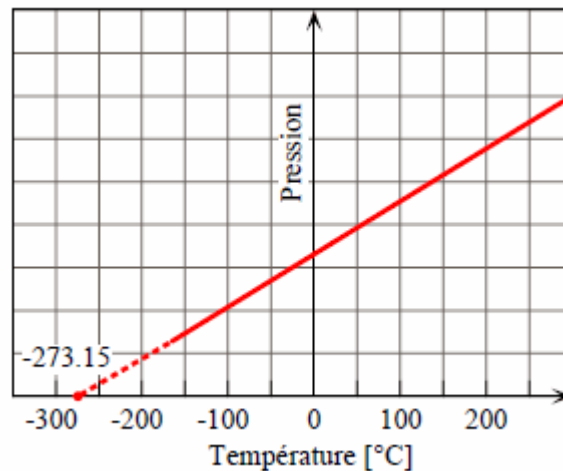


Figure: 33.1 - Représentation de la loi de Gay-Lussac avec la pression

d'où le choix confortable de poser le 0 [K] à cette valeur et de redéfinir une nouvelle échelle de températures.

Par ailleurs, une difficulté est de parler de variation de température, lorsque par exemple nous étudions la sensibilité de l'eau à passer de la phase solide à la phase liquide. Nous observons effectivement expérimentalement qu'il suffit dans des conditions idéales de laboratoire de passer de $-0.1\text{ }^{\circ}\text{C}$ à $+0.1\text{ }^{\circ}\text{C}$ pour observer le passage de changement qualitatif de l'eau de l'état solide à liquide. Parler alors de sensibilité en % de température est difficile avec l'échelle traditionnelle dans le cas présent. Mais si nous parlons en échelle absolue alors cela correspond au passage d'une température de 273.05 [K] à 273.25 [K] , soit une sensibilité de transition de phase pour l'eau de 0.1% en température autour de son point de fusion.

2. SYSTÈMES THERMODYNAMIQUES

D'une manière générale, un système thermodynamique est l'ensemble des corps situés à l'intérieur d'une surface fermée imaginaire et souvent considérée comme extensible sans déperdition d'énergie que nous appelons "**frontière**".

La frontière peut aussi sous indication de ses caractéristiques être matérielle!

Signalons que la frontière peut se limiter à une surface élémentaire dS associée avec son vecteur normal enveloppant une particule fluide (voir le chapitre de Génie Météo pour un bon exemple!). Nous l'appelons dans ce cas "**frontière particulaire**".

Il est souvent intéressant en thermodynamique, de faire le bilan des énergies qui sont transférées entre le système thermodynamique et le milieu extérieur, c'est-à-dire de considérer tout ce qui traverse la frontière.

Les principaux transferts (mais pas les seuls!) susceptibles d'être opérés sont:

1. Le "**transfert-travail**" W : Travail (mécanique) macroscopique ordonné effectué par une force sur une distance. Quand aucun transfert-travail (énergie) n'est opéré à l'échelle macroscopique, le système est dit "**système sans travail**".

2. Le "**transfert-chaleur**" Q : Énergie provenant de la variation du nombre de micro-états à l'échelle microscopique. Quand aucun transfert-chaleur (énergie) n'est opéré à l'échelle microscopique, le système est dit "**système adiabatique**", dans le cas contraire il est dit "**diatherme**".

3. Le "**transfert de masse**" M : Masse injectée dans le système. Quand aucun transfert de masse n'est opéré, le système est dit "**système fermé**".

Définitions:

D1. Un "**système isolé**" ne peut échanger ni travail, ni chaleur, ni masse avec le milieu extérieur.

D2. Un "**système ouvert**" peut échanger du travail, de la chaleur et de la masse avec le milieu extérieur.

D3. Un "**système fermé**" peut échanger du travail et de la chaleur mais pas de matière avec le milieu extérieur.

Remarques:

R1. Certains systèmes comptabilisent un bilan des actions extérieures qui leur sont appliquées nul. Ils sont alors dits "**pseudo-isolés**".

R2. Nous parlons de "**système homogène**" si la nature de ses constituants est égale en tout point, alors qu'il est un "**système uniforme**" ou "**système isotrope**" si ses caractéristiques sont égales en tout point.

3. TRANSFORMATIONS THERMODYNAMIQUES

Une "**transformation thermodynamique**" est l'opération au cours de laquelle l'état du système se modifie en passant d'un état initial à un état final.

Nous en distinguons au moins de deux types:

1. La "**transformation thermodynamique quasi-statique**" qui amène un système d'un état initial à un état final à travers une succession d'états qui sont exclusivement des états d'équilibre.

2.. Les transformations où toutes les variables d'état changent simultanément appelées "**transformations polytropiques**".

Il est possible que certaines variables restent constantes lors d'une transformation thermodynamique. Dans ce cas, nous utilisons une dénomination bien spécifique:

Symbole	Légende	Dénomination
P	pression	isobare
T	température	isotherme
V	volume	isochore
U	énergie interne	isoénergétique
S	entropie	isentropique
H	enthalpie	isenthalpique
Q	chaleur	adiabatique
G	enthalpie libre (énergie de Gibbs)	extensive

Tableau: 33.2 - Dénomination des grandeurs constantes

Remarques:

R1. Les études isobares sont d'un intérêt pratique important puisque tous les systèmes en contact avec l'atmosphère sont souvent, à l'équilibre, naturellement ou de manière forcée à pression constante (c'est-à-dire à la même pression que l'atmosphère environnante!) dans les laboratoires.

R2. Il est difficile de construire réellement des transformations isothermes utiles. Le maintien de la température exige un excellent contact thermique et un temps de réaction assez long pour garder la température uniforme. Ce type de transformation sera donc, dans le monde réel, assez lent. Attention! Une transformation adiabatique peut être réversible ou irréversible, alors qu'une transformation isotherme est forcément réversible pour la gaz parfaits.

R3. Beaucoup de transformations réelles sont considérées comme adiabatiques. Il suffit pour cela que le contenant soit très bien isolé ou même que la transformation soit assez rapide pour que les échanges thermiques soient négligeables.

Nous distinguons également deux cycles de transformations thermodynamiques principaux qui sont:

1. Le "**cycle thermodynamique fermé**": le système décrit une suite de transformations telles que l'état final et l'état initial de la transformation est identique et que la quantité et les propriétés des éléments participants au cycle sont toujours les mêmes.
2. Le "**cycle thermodynamique ouvert**": le système décrit une suite de transformations telles que l'état final et l'état initial de la transformation est identique et que la quantité et les propriétés des éléments participants au cycle ne sont pas toujours les mêmes

4. VARIABLES D'ÉTAT

Définition: "**L'état thermodynamique**" d'un système est l'ensemble des propriétés qui le caractérisent, indépendamment de la forme de sa frontière. Les variables qui décrivent l'état du système en ne connaissant que l'état final et initial de celui-ci sont appelées principalement "**fonctions d'état**" ou fréquemment "**variables d'état**" et encore parfois "**grandeurs d'état**"...

Remarque: Certaines fonctions d'état jouent un rôle particulier dans la définition des états d'équilibre d'un système. Ce sont des grandeurs accessibles, à l'échelle macroscopique, directement ou indirectement grâce à des instruments de mesure. Ces fonctions d'état particulières (comme la pression, la température, le volume, etc.) sont appelées "**variables d'état d'équilibre d'un système thermodynamique**".

La définition précédente suppose implicitement l'existence d'un état et des variables d'état, c'est-à-dire que les grandeurs caractéristiques du système sont définies (ou théoriquement accessibles dans le sens de la mesure) à tout instant et en tout point du système. Ceci est loin d'être évident si nous considérons des évolutions rapides telles que des explosions.

Cette difficulté peut être éludée en se retranchant derrière "**l'hypothèse de l'état local**": Nous supposons qu'à tout instant, les grandeurs caractéristiques ont localement les mêmes expressions que dans une configuration stationnaire, ce qui sous-entend que les temps nécessaires aux changements d'état sont négligeables devant les durées caractéristiques de l'évolution.

Le choix des variables d'état dépend de la nature du problème traité. Nous pouvons néanmoins séparer l'ensemble de ces variables d'état en:

1. Des variables d'état dites "**grandeurs extensives**", proportionnelles à la quantité de matière et donc au nombre d'atomes/molécules du système servant à les définir (donc elles sont additives). C'est typiquement le cas de la masse, le volume, l'entropie,...
2. Des variables d'état dites "**grandeurs intensives**", indépendantes de la quantité de matière (donc par extension, non additives). C'est typiquement le cas de la pression, la température, l'énergie ...

D'une manière générale, une variable intensive dépend du point envisagé dans le système étudié (température, concentrations peuvent varier d'un point à l'autre) alors que la grandeur extensive est définie sur la globalité du système.

Montrons maintenant avec un exemple particulier que le rapport de deux variables extensives est une grandeur intensive. Pour cela, rappelons la loi des gaz parfaits (cf. chapitre de Mécanique Des Milieux Continus) qui est donc une équation d'état (nous reviendrons sur les équations d'état en détails plus tard):

$$P \cdot V = N_{Av} \cdot n \cdot k \cdot T = n \cdot R \cdot T \quad (33.3)$$

qui comprend donc trois variables thermodynamiques P , V et T (ce qui laisse à choix deux variables de contrôle et indépendantes et une variable dépendante).

Nous pouvons calculer la concentration C (à ne pas confondre avec la notation de la capacité calorifique que nous verrons plus loin!), qui est donc une grandeur intensive, par le rapport de deux grandeurs extensives n , V :

$$C = \frac{n}{V} = \frac{P}{RT} \quad (33.4)$$

Les variables d'état possèdent par ailleurs une propriété particulière: leurs variations ne dépendent pas de la nature de la transformation qui affecte le système mais uniquement de l'état final et initial du système à l'équilibre (ce qui est très utile en pratique...!). Il s'agit du concept d'intégrale de chemin que nous avons déjà traité en détail dans le chapitre de Mécanique Classique et de Calcul Différentiel Et Intégral.

Citons les grandeurs extensives et intensives les plus courantes en thermodynamique:

Symbole	Légende	Grandeur
P	pression	intensive
T	température	intensive
V	volume	extensive
E	énergie totale	extensive
U	énergie interne	extensive
S	entropie	extensive
H	enthalpie	extensive
F	énergie libre	extensive
G	enthalpie libre	extensive

Tableau: 33.3 - Grandeurs extensives et intensives courantes

À l'opposé, le travail W et la chaleur Q ne sont pas des variables d'état, car elles dépendent de la nature de la transformation. Néanmoins il existe des cas particuliers où la chaleur et le travail ne dépendent plus du chemin suivi lorsque les transformations s'effectuent soit à pression constante, soit à volume constant (nous verrons cela plus loin).

4.1. PHASES

Définition: Un système dans lequel les différentes grandeurs intensives varient de façon continue constitue une "**phase**".

Nous pouvons donc considérer que toute grandeur intensive dépend des coordonnées du point envisagé: le système est constitué par une seule phase si la grandeur intensive est continue dans tout le système. C'est le cas des gaz, des liquides et de certains solides constituant des solutions solides.

Si la grandeur intensive présente une discontinuité (ou plusieurs), le système est dit "**polyphasique**". Cependant, si les grandeurs intensives ont même valeur en tout point du système, la phase est dite "**phase uniforme**": le système a alors même température, pression, composition en chacun de ses points.

Pour un système homogène, il peut être commode de ramener les variables extensives à l'unité de masse. Nous parlons alors de "**grandeurs massiques**" (ou "spécifiques"), généralement notées en minuscules.

Nous utilisons si possible pour éviter toute confusion les règles de notation suivantes:

- Toute grandeur non massique est représentée par une lettre latine majuscule
- Toute grandeur massique (très utilisée en chimie!) est représentée par une lettre latine minuscule

Dans le cas contraire, nous spécifierons de quel type de variable il s'agit.

5. ÉQUATIONS D'ÉTAT

En physique, et plus particulièrement en thermodynamique, "**l'équation d'état**" d'un système à l'équilibre thermodynamique est une relation entre différents paramètres physiques, appelés donc les "variables d'état" déjà traitées au début de ce chapitre, qui déterminent son état. Il peut s'agir par exemple d'une relation entre sa température, sa pression et son volume. À partir de l'équation d'état caractéristique d'un système physique, il est possible de déterminer la totalité des quantités thermodynamiques décrivant ce système et par suite de prédire ses propriétés.

Les équations d'état sont généralement restreintes à un type de comportement ou de phénomènes physiques donnés. Un même corps peut donc avoir plusieurs équations d'état, concernant par exemple son état magnétique ou son état thermodynamique.

Pour qu'un corps puisse être caractérisé par une équation d'état à un instant donné, il faut que l'état de ce corps dépende uniquement des valeurs prises par les paramètres à cet instant. Les corps présentant un phénomène d'hystérésis ne peuvent donc pas être caractérisés par une équation d'état.

5.1. ÉQUATION D'ÉTAT D'UN GAZ PARFAIT

Nous avons démontré dans le chapitre de Mécanique Des Milieux Continus que l'équation d'état d'un gaz parfait était (relation déjà introduite une peu fortuitement plus haut):

$$PV = nRT \quad (33.5)$$

Soit:

$$V(P, T) = \frac{nRT}{P} \quad (33.6)$$

et l'équation d'état des gaz réels (aussi démontrée dans le chapitre de Mécanique Des Milieux Continus):

$$\left(P + a \left(\frac{N}{V} \right)^2 \right) (V - N \cdot b) \cong kNT \quad (33.7)$$

Pour déterminer des équations d'état particulières et idéalisées des solides et liquides, il faudra d'abord introduire plusieurs coefficients permettent de caractériser certaines propriétés globales de milieux continus.

Ces coefficients doivent être définis avec précaution, dépendant de la façon dont on les mesure, c'est-à-dire dépendant de ce qui est réellement mesuré.

Prenons, par exemple, le concept de compressibilité (volumique) d'un échantillon. Il est raisonnable de dire que l'idée sous-jacente est la mesure du changement de volume sous un changement de pression. Si la pression augmente, le volume diminuera, il apparaît donc raisonnable de définir un coefficient de compressibilité par:

$$\chi = - \frac{\partial V}{\partial P} \quad (33.8)$$

où le signe "-" sert à garantir un χ positif dans le cas d'une compression.

Mais cette définition ne veut rien dire. Nous n'y avons pas spécifié les conditions dans lesquelles les mesures sont prises. Clairement, nous augmentons (ou diminuons) la pression mesurée à l'aide d'un manomètre par exemple et nous mesurons avec une règle (!) le changement de volume qui en résulte. Cela semble simple, mais c'est incomplet. Si nous augmentons la pression assez vite, le volume va diminuer mais la température va augmenter. Mesure-t-on le changement de volume immédiatement ou attend-on que la température soit redevenue celle qu'elle était en plaçant l'échantillon en contact thermique avec un thermostat avant de mesurer le changement de volume. Dans ce dernier cas, la mesure aura été faite à température constante.

En d'autres termes, mesure-t-on le changement de volume d'un réceptacle en contact thermique parfait avec son environnement extérieur (nous lui laissons le temps d'aller à l'équilibre) ou isolé thermiquement de cet environnement. C'est très différent!

De plus, V étant une variable extensive, il y aura avantage à définir la compressibilité comme une quantité intensive, c'est-à-dire indépendante de la quantité de gaz considéré (ici isothermique), ce qui rend plus facile la construction des tables de valeur. Nous avons alors intuitivement la définition suivante du "coefficient de compressibilité (relative) isothermique":

$$\chi_T = -\frac{1}{V} \left(\frac{\partial V}{\partial P} \right)_T \quad (33.9)$$

sur lequel nous reviendrons un peu plus tard.

De la même manière intuitive, nous pouvons définir le "coefficient de dilatation (relative) volumique isobare" (appelé plus rarement "coefficient de compression (relative) volumique isobare"):

$$\alpha_{V,P} = \frac{1}{V} \left(\frac{\partial V}{\partial T} \right)_P \quad (33.10)$$

sur lequel nous reviendrons aussi un peu plus tard.

Ainsi, par exemple, pour un gaz parfait il vient:

$$\alpha_{V,P} = \frac{1}{V} \left(\frac{\partial \left(\frac{nRT}{P} \right)}{\partial T} \right)_P = \frac{1}{V} \frac{nR}{P} = \frac{nR}{VP} = \frac{nR}{nRT} = \frac{1}{T} \quad (33.11)$$

Nous verrons un peu plus tard une autre approche pour retrouver cette égalité.

5.2. ÉQUATION D'ÉTAT D'UN LIQUIDE

Maintenant les variables d'état (macroscopiques) d'un liquide sont assez trivialement son volume V , sa pression P et sa température T . Son équation d'état peut alors par exemple se mettre aussi sous la forme:

$$V = f(T, P) \quad (33.12)$$

Sous forme différentielle cela donne:

$$dV = \left(\frac{\partial V}{\partial T} \right)_P dT + \left(\frac{\partial V}{\partial P} \right)_T dP \quad (33.13)$$

En introduisant les coefficients thermoélastiques comme nous l'avons déjà fait plus haut, il vient:

$$\frac{dV}{V} = \alpha_{V,P} dT - \chi_T dP \quad (33.14)$$

L'expérience montre que pour les liquides, les coefficients thermoélastiques varient très peu avec la température et la pression (oui... l'expérience c'est quand même utile parfois...). Nous pouvons donc les supposer constants, pourvu que les variations de la température et de la pression restent modérées. Comme par ailleurs les variations de volume d'un liquide sont très faibles, nous pouvons faire la même approximation que celle démontrée dans le chapitre de Mécanique Des Milieux Continus avec la déformation vraie longitudinale. C'est-à-dire que:

$$\frac{dV}{V} \cong \frac{dV}{V_0} \quad (33.15)$$

où V_0 est le volume de référence constant. Compte tenu de ces remarques, nous obtenons trivialement par intégration l'équation d'état générale d'un liquide:

$$V(T, P) = V_0 \left[1 + \alpha_{V,P} (T - T_0) - \chi_T (P - P_0) \right] \quad (33.16)$$

Remarquons que l'approximation d'un fluide incompressible, très utilisée en mécanique des fluides, consiste à supposer que:

$$\alpha_{V,P} = 0, \chi_T = 0 \quad (33.17)$$

L'équation d'état du liquide devient alors:

$$V(T, P) = V_0 = c^{te} \quad (33.18)$$

Indiquons aussi que dans les petites classes, les solides sont considérés comme des liquides incompressibles. Nous avons alors la fameuse relation connue de nombreux écoliers:

$$V = V_0 \left[1 + \alpha_{V,P} (T - T_0) \right] \quad (33.19)$$

ou autrement écrit:

$$\Delta V = \alpha_{V,P} V_0 \Delta T \quad (33.20)$$

Ces deux dernières relations sont exactement les mêmes que la formulation technique de loi Gay-Lussac pour les gaz (cf. chapitre de Mécanique Des Milieux Continus).

Nous pouvons refaire le même raisonnement avec un élément de longueur mais en définissant un autre coefficient thermoélastique linéique. Dès lors, il vient une autre fameuse relation connue de nombreux écoliers:

$$\Delta L = \alpha_{L,P} L_0 \Delta T \quad (33.21)$$

Nous pouvons alors déterminer la relation entre $\alpha_{L,P}$ et $\alpha_{V,P}$ (qui sont tous deux de beaucoup inférieur à l'unité). Effectivement pour un cube de côté L , nous avons:

$$\begin{aligned} \Delta V &= (L_0 + \Delta L)^3 - L_0^3 = (L_0 + \alpha_{L,P} L_0 \Delta T)^3 - L_0^3 = L_0^3 (1 + \alpha_{L,P} \Delta T)^3 - L_0^3 \\ &= L_0^3 \left(1 + 3\alpha_{L,P} \Delta T + \underbrace{3\alpha_{L,P}^2 \Delta T^2}_{\rightarrow 0} + \underbrace{\alpha_{L,P}^3 \Delta T^3}_{\rightarrow 0} \right) - L_0^3 \approx 3\alpha_{L,P} L_0^3 \Delta T \end{aligned} \quad (33.22)$$

et donc il vient:

$$3\alpha_{L,P} \cong \alpha_{V,P} \quad (33.23)$$

Nous avons par exemple pour l'eau à température ambiante:

$$\alpha_{V,P} \cong 2.6 \cdot 10^{-4} \left[K^{-1} \right], \quad \chi_V \cong 4.5 \cdot 10^{-7} \left[Pa^{-1} \right] \quad (33.24)$$

Exemple:

Un réservoir de volume V , aux parois supposées indéformables, est complètement rempli d'eau initialement à l'équilibre thermodynamique à la pression initiale P_0 de 1 [bar] et à la température initiale T_0 de 298 [K]. Après avoir reçu une certaine quantité de chaleur, le liquide atteint un nouvel état d'équilibre thermodynamique, à la température de 398 [K]. Nous avons alors puisque $V = V_0$, l'équation d'état du liquide qui se réduit à:

$$\frac{V(T,P)}{V_0} = 1 = \left[1 + \alpha_{V,P} (T - T_0) - \chi_V (P - P_0) \right] \quad (33.25)$$

Soit après réarrangement:

$$P = \frac{\alpha_{V,P}}{\chi_V} (T - T_0) + P_0 \cong 1.58 \text{ [bar]} \quad (33.26)$$

5.3. ÉQUATION D'ÉTAT D'UN SOLIDE

Considérons maintenant le cas particulier d'une poutre (ou d'un fil), composée d'un solide élastique. Les variables d'état typiques du système sont alors sa longueur L , sa température T et la force de tension F (positive dans le cas de la traction et négative dans le cas de la compression). L'équation d'état se met donc sous la forme différentielle suivante:

$$dL = \left(\frac{\partial L}{\partial T} \right)_F dT + \left(\frac{\partial L}{\partial F} \right)_T dF \quad (33.27)$$

Par analogie avec le coefficient de dilatation volumique isobare $\alpha_{V,P}$, nous définissons le "**coefficient de dilatation linéique**":

$$\alpha_{L,F} = \frac{1}{L} \left(\frac{\partial L}{\partial T} \right)_F \quad (33.28)$$

Par ailleurs, nous avons démontré dans le chapitre de Mécanique Des Milieux Continus, que pour un essai de traction (sans composante tangentielle) isotherme effectué sur un solide élastique nous avons (implicitement c'est la loi de Hooke pour rappel):

$$\frac{|\Delta L|}{L_0} = \frac{F}{E \cdot S} \quad (33.29)$$

Mais nous pouvons aussi utiliser la déformation vraie longitudinale après quoi il faut évidemment adapter le terme de droite en conséquence:

$$\frac{dL}{L} = \frac{dF}{E \cdot S} \quad (33.30)$$

Il vient alors:

$$\frac{1}{E \cdot S} = \frac{1}{L} \left(\frac{\partial L}{\partial F} \right)_T \quad (33.31)$$

L'équation d'état du solide peut alors s'écrire en y injectant cette dernière relation:

$$\frac{dL}{L} = \alpha_{L,F} dT + \frac{dF}{E \cdot S} \quad (33.32)$$

Cette relation nous donne donc la variation d'allongement relative correspondant à une variation infinitésimale de la force de tension et de la température, connaissant le coefficient de dilatation linéique.

L'équation d'état ci-dessus s'intègre facilement en supposant que le coefficient de dilation linéique, le module de Young et la surface de section sont constants:

$$L = L_0 e^{\alpha_L \Delta T + \frac{\Delta F}{E \cdot S}} = L_0 e^{\alpha_L \Delta T} e^{\frac{\Delta F}{E \cdot S}} \quad (33.33)$$

Si les déformations sont très faibles, nous avons démontré dans le chapitre de Mécanique Des Milieux Continus que la variation relative d'allongement pouvait s'écrire:

$$\frac{dL}{L} \cong \frac{dL}{L_0} \quad (33.34)$$

Nous avons alors l'équation d'état qui s'écrit:

$$\frac{dL}{L_0} = \alpha_{L,F} dT + \frac{dF}{E \cdot S} \quad (33.35)$$

Donc après intégration:

$$\frac{\Delta L}{L_0} = \alpha_{L,F} \Delta T + \frac{\Delta F}{E \cdot S} \quad (33.36)$$

Exemple:

Une poutre cylindrique en acier de rayon de 15 [mm] est initialement dans un état d'équilibre thermodynamique caractérisé par une longueur initiale de 1 [m] à la température initiale de 300 [K] et une force de tension initiale nulle. La poutre est chauffée jusqu'à un nouvel état d'équilibre pour lequel sa température finale est de 400 [K]. Pour calculer sa nouvelle longueur sous l'hypothèse particulière que la poutre est libre de tout mouvement et que la déformation est considérée comme faible:

$$\frac{L - L_0}{L_0} = \alpha_L (T - T_0) + \frac{\frac{\Delta F}{=0}}{E \cdot S} \quad (33.37)$$

Soit:

$$L = L_0 \alpha_L (T - T_0) + L_0 = 1 \cdot 12 \cdot 10^{-6} \cdot 100 + 1 = 1.0012 \text{ [m]} \quad (33.38)$$

Si la poutre est par contre encastree des deux côtés, nous pouvons alors calculer la force de tension interne générée sous l'hypothèse des petites déformations. Nous utilisons alors:

$$\frac{\frac{\Delta L}{=0}}{L_0} = \alpha_L (T - T_0) + \frac{\Delta F}{E \cdot S} \quad (33.39)$$

Soit:

$$F = -\alpha_L E \cdot S (T - T_0) + F_0 = -\alpha_L E \cdot S (T - T_0) \quad (33.40)$$

$$= -12 \cdot 10^{-6} \cdot 2 \cdot 10^{11} \cdot \pi \cdot 0.015^2 \cdot 100 \cong 170 \text{ [kN]}$$

Nous voyons là un phénomène bien connu en construction mécanique: la déformation d'origine thermique d'un solide est généralement très faible, mais l'empêcher peut générer des efforts énormes, susceptibles de détruire un mécanisme.

6. PRINCIPES DE LA THERMODYNAMIQUE

Les principes de la thermodynamique sont les briques de la mécanique énergétique ou thermique. Chaque principe implique une grande quantité de concepts que nous essaierons de présenter et définir au mieux. C'est la partie "sensible" de ce domaine de la physique.

La thermodynamique est basée sur quatre principes fondamentaux (dont certains sont démontrables):

P0. Le "**principe zéro de la thermodynamique**" ou "**principe de l'équilibre thermique**" est défini par le fait que si deux systèmes thermodynamiques 1 et 2 sont en équilibre thermique avec un troisième 3, ils sont eux-mêmes en équilibre thermique (il s'agit d'une "assertion" dans le langage de la théorie de la démonstration).

Donc si deux corps en contact sont en équilibre thermique, ils ont la même température et donc que l'état reste stationnaire. Deux corps à même température en contact sont en équilibre thermique.

Remarque: Nous avons utilisé ce principe implicitement (lorsque nous avons énoncé que l'état d'équilibre était le plus probable) dans le chapitre de Mécanique Statistique pour démontrer la loi de Boltzmann.

P1. Le "**premier principe de la thermodynamique**" ou "**principe de conservation**" concerne le caractère conservatif de l'énergie et énonce qu'au cours d'une transformation quelconque d'un système, la variation de son énergie totale est égale à la somme des variations de tous les types d'énergie le définissant:

$$\Delta E_{tot} = \Delta U + \Delta E_c + \Delta E_p \quad (33.41)$$

Si le système est isolé la variation d'énergie totale sera bien évidemment toujours nulle!

Bien évidemment, lorsque le système étudié (le conteneur et son contenu) est dans le même référentiel que l'observateur, il ne nous reste plus que le terme de variation d'énergie interne:

$$\Delta E_{tot} = \Delta U \quad (33.42)$$

ou en termes différentiels:

$$dE_{tot} = dU \quad (33.43)$$

Reste à déterminer l'expression exacte de la variation d'énergie interne U et des variables dont elle dépend (développements que nous ferons un peu plus loin).

Remarque: Ce principe est démontrable si nous acceptons le théorème de Noether (cf. chapitre Principes) d'invariance dans le temps des lois de la physique comme un principe supérieur.

P2. Le "**deuxième principe de la thermodynamique**" appelé aussi "**principe de Carnot-Clausius**" ou encore "**principe d'évolution**" concerne le caractère d'irréversibilité et est associé au concept d'entropie. Ce principe énonce que la chaleur ne peut passer d'elle-même que d'un corps chaud à un corps moins chaud (ou qu'un corps va inexorablement refroidir). L'opération inverse nécessitant l'apport de travail mécanique pris sur le système extérieur ce qui est donné par la relation (démontrée un peu plus loin):

$$dS = \frac{\delta Q}{T} \geq 0 \quad (33.44)$$

Le terme de gauche de l'égalité nous dit que l'entropie ne peut qu'augmenter et que pour cela, l'échange de quantité chaleur avec le système extérieur est toujours positif (numérateur du terme de droite de l'égalité). Donc le corps fournit de la chaleur au système extérieur et refroidit inexorablement.

La notation particulière δQ est ici pour indiquer que la chaleur est une différentielle totale inexacte. Il s'agit d'un concept sur lequel nous reviendrons plus loin.

Remarque: Nous démontrerons grossièrement... cette relation plus loin en utilisant la loi de Boltzmann sur l'accès à l'information dans un système telle qu'étudiée dans le chapitre de Mécanique Statistique.

P3. Le "**troisième principe de la thermodynamique**" ou "**principe de Nernst**" concerne les propriétés de la matière dans le voisinage du zéro absolu et énonce qu'à la limite du zéro absolu, température qui ne saurait être atteinte (cf. chapitre de Physique Quantique Ondulatoire), l'entropie d'équilibre d'un système tend vers une constante indépendante des autres paramètres intensifs, constante qui est prise nulle.

Démonstration: Il s'agit du "**théorème de Nernst**" dont le résultat formel est simplement basé sur le deuxième principe donnant l'entropie (cf. chapitre de Mécanique Statistique):

$$S = k \ln(\Omega) \quad (33.45)$$

Nous voyons que si nous considérons un cristal à la température du zéro absolu, la position des particules, les unes par rapport aux autres est parfaitement définie de façon unique. Par conséquent, il n'y a qu'une seule complexion possible pour un cristal au zéro absolu. Le nombre de complexions Ω est donc égal à 1. Ce qui entraîne que:

$$S = k \ln(1) = 0 \quad (33.46) .$$

Rappelons que le "**zéro absolu**" est la température la plus basse accessible théoriquement en physique. À cette température, une substance ne contient plus à l'échelle macroscopique, l'énergie thermique (ou chaleur) nécessaire à l'occupation de plusieurs niveaux énergétiques microscopiques (cf. chapitre de Mécanique Statistique). Les particules qui la composent (atomes, molécules) sont toutes dans le même état d'énergie minimale (état fondamental). Cela se traduit par une entropie nulle due à l'indiscernabilité de ces particules dans ce même niveau d'énergie fondamentale (selon le troisième principe de la thermodynamique) et par une totale immobilité au sens classique selon le théorème du Viriel.

Cependant, au sens quantique, nous avons vu dans le chapitre de Physique Quantique Ondulatoire que les particules possèdent toujours une quantité de mouvement non nulle d'après les relations d'incertitudes de Heisenberg et que selon les distributions de Fermi-Dirac et de Bose-Einstein (cf. chapitre de Mécanique Statistique), qu'il y a quand même différents niveaux d'énergie lorsque nous prenons en compte le principe d'exclusion de Pauli. D'où l'émergence de la thermodynamique quantique.

7. CAPACITÉS CALORIFIQUES

Dans une transformation donnée, un apport ΔQ de chaleur se traduit en général par une élévation ΔT de la température du système.

Définition: Nous appelons "**capacité calorifique**" (ou "**capacité calorique**", ou encore "**capacité thermique**" ou enfin "**chaleur spécifique**"...) C du système, la quantité:

$$C = \frac{\Delta Q}{\Delta T} \quad (33.47)$$

La capacité calorifique est donc de par ses unités, l'énergie qu'il faut apporter à un corps pour augmenter sa température de 1 Kelvin. C'est une grandeur extensive (plus la quantité de matière est importante, plus la capacité calorifique est grande). Sous forme infinitésimale, nous avons:

$$C = \frac{\delta Q}{\partial T} \quad (33.48)$$

Nous voyons que l'avantage pédagogique d'avoir une écriture utilisant des variations discrètes plutôt qu'infinitésimales permet de ne pas avoir se poser la question si nous avons une différentielle totale exacte ou inexacte... (voir plus loin la distinction).

Dans le cas d'un gaz parfait, nous admettrons que la totalité de la chaleur échangée (énergie) est mise sous forme d'énergie cinétique des atomes ou molécules (énergie thermique), donc nous pouvons alors écrire:

$$C = \frac{\Delta Q}{\Delta T} = \frac{\Delta U}{\Delta T} = \frac{\Delta E_c}{\Delta T} \quad (33.49)$$

et nous démontrerons plus loin comment déterminer exactement la valeur de cette capacité pour de tels gaz.

Voyons cependant déjà comment déterminer cette capacité calorifique pour des solides cristallins (que nous supposons à volume constant). Pour cela, rappelons que nous avons démontré dans le chapitre d'Électrocinétique que dans un cristal, la densité d'états dans un volume sous certaines conditions bien précises était:

$$D(\omega) = \frac{dN}{dk} \frac{dk}{d\omega} = \frac{V k^2}{2\pi^2} \frac{dk}{d\omega} \quad (33.50)$$

et que le nombre d'états relatifs était donné:

$$N = \frac{k^3}{6\pi^2} V \quad (33.51)$$

Or la relation antécédente est indépendante du fait que dans le domaine de l'électrocinétique nous considérons des électrons ou toute autre particule vecteur d'une propriété quelconque (charge, chaleur ou autre). Ainsi dans le cadre de la thermodynamique des solides cristallins, un modèle consiste à supposer le vecteur de chaleur comme une particule ayant des propriétés ondulatoires similaires à l'électron qu'il est d'usage d'appeler le "**phonon**".

En supposant que les phonons sont identiques dans les trois directions de l'espace et en se rappelant la relation de Planck-Einstein (cf. chapitre de Physique Quantique Corpusculaire) valable pour toute particule de par le principe de dualité onde-corpuscule:

$$E = h\nu \quad (33.52)$$

et (cf. chapitre de Mécanique Statistique) de la statistique de Bose-Einstein pour les bosons (particules non soumises au principe d'exclusion de Pauli pour rappel...):

$$f_{BE}(E, T, \mu) = \frac{1}{\exp\left(\frac{E - \mu}{kT}\right) - 1} \quad (33.53)$$

que nous noterons dans le cas présent sous la forme suivante (puisque le potentiel chimique est nul dans le cas qui nous intéresse):

$$n(\omega) = \frac{1}{\exp\left(\frac{E}{kT}\right) - 1} = \frac{1}{e^{h\nu/kT} - 1} = \frac{1}{e^{\hbar\omega/kT} - 1} \quad (33.54)$$

nous avons en combinant ces trois grands résultats:

$$\begin{aligned} U &= 3 \int_{\omega} h\nu n(\omega) D(\omega) d\omega = 3 \int_{\omega} h \frac{\omega}{2\pi} n(\omega) D(\omega) d\omega = 3 \int_{\omega} \hbar\omega n(\omega) D(\omega) d\omega \\ &= 3 \int_{\omega} \frac{\hbar\omega}{e^{\hbar\omega/kT} - 1} D(\omega) d\omega \end{aligned} \quad (33.55)$$

Nous avons alors (en écrivant abusivement que la chaleur est :

$$\begin{aligned}
C_V &= \left(\frac{\delta Q}{\delta T} \right)_V = \left(\frac{dU - \delta W}{\delta T} \right)_V = \left(\frac{dU - P \underbrace{dV}_{=0}}{\delta T} \right)_V = \left(\frac{dU}{\delta T} \right)_V \\
&= 3 \int_{\omega} \hbar \omega \frac{\partial}{\partial T} \left(\frac{1}{e^{\hbar \omega / kT} - 1} \right) D(\omega) d\omega = 3 \int_{\omega} \hbar \omega \frac{\frac{\hbar \omega}{kT} \frac{e^{\hbar \omega / kT}}{(e^{\hbar \omega / kT} - 1)^2}}{D(\omega) d\omega} \\
&= 3 \int_{\omega} \hbar \omega \left[-\frac{1}{(e^{\hbar \omega / kT} - 1)^2} \cdot \left(-\frac{\hbar \omega}{kT^2} e^{\hbar \omega / kT} \right) \right] D(\omega) d\omega \\
&= 3k \int_{\omega} \left(\frac{\hbar \omega}{kT} \right)^2 \frac{e^{\hbar \omega / kT}}{(e^{\hbar \omega / kT} - 1)^2} D(\omega) d\omega
\end{aligned} \tag{33.56}$$

À partir de là il existe deux principaux modèles de prévision de la capacité calorifique des cristaux, le modèle de Debye et le modèle d'Einstein, chacun faisant des hypothèses sur la densité de modes des phonons.

Dans le modèle d'Einstein un seul mode de fréquence (respectivement pulsation) est supposé possible, de sorte la vibration du solide est schématisée par N oscillateurs harmoniques à cette fréquence. Cette pulsation particulière est souvent associée à ce que nous appelons la "**température d'Einstein**":

$$\Theta_E = \frac{\hbar \omega_E}{k} \tag{33.57}$$

La densité de modes est alors symbolisée par N distributions de Dirac (cf. chapitre de Calcul Différentiel Et Intégral) autour de cette valeur tel que:

$$\begin{aligned}
C_V &= 3k \int_{\omega} \left(\frac{\hbar \omega}{kT} \right)^2 \frac{e^{\hbar \omega / kT}}{(e^{\hbar \omega / kT} - 1)^2} N \delta(\omega_E) d\omega = 3kN \left(\frac{\hbar \omega_E}{kT} \right)^2 \frac{e^{\hbar \omega_E / kT}}{(e^{\hbar \omega_E / kT} - 1)^2} \\
&= 3kN \left(\frac{\Theta}{T} \right)^2 \frac{e^{\Theta/T}}{(e^{\Theta/T} - 1)^2}
\end{aligned} \tag{33.58}$$

Ce modèle prévoit une valeur limite de la capacité calorifique aux hautes températures (nous utilisons un développement de Taylor):

$$\begin{aligned}
C_{V,T \rightarrow \infty} &\cong 3kN \left(\left(\frac{\hbar \omega}{kT^2} \right)^2 \frac{(1 + \hbar \omega / kT)}{(1 + \hbar \omega / kT - 1)^2} \right) = 3kN \left(\left(\frac{\hbar \omega}{kT^2} \right)^2 \frac{(1 + \hbar \omega / kT)}{(\hbar \omega / kT)^2} \right) \\
&= 3kN(1 + \hbar \omega / kT) \cong 3kN
\end{aligned} \tag{33.59}$$

Il est d'usage lorsqu'on présente cette relation de considérer une quantité de matière correspondant à une mole et de faire apparaître ainsi la constante des gaz parfaits:

$$C_{V,T \rightarrow +\infty} \cong 3kN_A = 3R \quad (33.60)$$

Cette valeur avait été donnée initialement par Dulong et Petit et elle est en assez bon accord avec les résultats expérimentaux. Nous parlons souvent de la "**loi de Dulong et Petit**".

En revanche aux faibles températures, il faut faire appel au modèle de Debye qui consiste à supposer la relation de dispersion comme linéaire avec la vitesse de propagation des phonons (vitesse du son dans le cristal) supposée comme indépendant de la température:

$$\omega(k) = vk \quad (33.61)$$

La densité d'états s'écrit alors:

$$D(\omega) = \frac{Vk^2}{2\pi^2} \frac{dk}{d\omega} = \frac{Vk^2}{2\pi^2} \frac{d\left(\frac{\omega}{v}\right)}{d\omega} = \frac{Vk^2}{2\pi^2 v} = \frac{V\left(\frac{\omega}{v}\right)^2}{2\pi^2 v} = \frac{V\omega^2}{2\pi^2 v^3} \quad (33.62)$$

Nous avons alors pour la capacité calorifique:

$$U = 3 \int_0^{\omega_p} \hbar \omega \frac{1}{e^{\hbar \omega / kT} - 1} \frac{V\omega^2}{2\pi^2 v^3} d\omega = \frac{3V\hbar}{2\pi^2 v^3} \int_0^{\omega_p} \frac{\omega^3}{e^{\hbar \omega / kT} - 1} d\omega \quad (33.63)$$

Dans le modèle de Debye, seule la branche acoustique des phonons est considérée, ce qui revient à schématiser le cristal comme un réseau monoatomique formé des atomes les plus lourds. Dès lors, les modes de résonances sont limités à une certaine fréquence (pulsation), avec le volume correspondant, dite de coupure telle que (nous changeons la notation de la constante de Boltzmann afin de ne pas la confondre avec le nombre d'onde):

$$\begin{aligned} U &= \frac{3V\hbar}{2\pi^2 v^3} \int_0^{\omega_p} \frac{\omega^3}{e^{\hbar \omega / kT} - 1} d\omega = \frac{3V\hbar}{2\pi^2 v^3} \int_0^{x_p} \frac{\left(\frac{k_B T}{\hbar} x\right)^3}{e^x - 1} \frac{k_B T}{\hbar} dx \\ &= \frac{3V\hbar}{2\pi^2 v^3} \left(\frac{k_B T}{\hbar}\right)^4 \int_0^{x_p} \frac{x^3}{e^x - 1} dx = \frac{3V\hbar}{2\pi^2} \frac{k_B T}{\hbar} \left(\frac{k_B T}{v\hbar}\right)^3 \int_0^{x_p} \frac{x^3}{e^x - 1} dx \\ &= \frac{3V}{2\pi^2} k_B T \left(\frac{k_B T}{v\hbar}\right)^3 \int_0^{x_p} \frac{x^3}{e^x - 1} dx = \frac{3\left(\frac{N6\pi^2}{k^3}\right)}{2\pi^2} k_B T \left(\frac{k_B T}{v\hbar}\right)^3 \int_0^{x_p} \frac{x^3}{e^x - 1} dx \quad (33.64) \\ &= 9Nk_B T \left(T \frac{k_B}{v\hbar} \sqrt[3]{\frac{1}{k^3}}\right)^3 \int_0^{x_p} \frac{x^3}{e^x - 1} dx = 9Nk_B T \left(T \frac{k_B}{v\hbar} \sqrt[3]{\frac{V}{6\pi^2 N}}\right)^3 \int_0^{x_p} \frac{x^3}{e^x - 1} dx \\ &= 9Nk_B T \left(\frac{T}{T_D}\right)^3 \int_0^{x_p} \frac{x^3}{e^x - 1} dx \end{aligned}$$

où (nous avons mis les deux formes d'écritures les plus courantes):

$$T_D = \frac{v\hbar}{k_B} \left(6\pi^2 \frac{N}{V} \right)^{1/3} = \frac{v_{son}\hbar}{2k_B L} \left(\frac{6N}{\pi} \right)^{1/3} \quad (33.65)$$

Nous avons alors pour la capacité calorifique:

$$\begin{aligned} C_V &= \left(\frac{\partial U}{\partial T} \right)_V = \frac{3V\hbar}{2\pi^2 v^3} \frac{\partial}{\partial T} \int_0^{\omega_p} \frac{\omega^3}{e^{\hbar\omega/k_B T} - 1} d\omega = \frac{3V\hbar}{2\pi^2 v^3} \int_0^{\omega_p} \frac{\partial}{\partial T} \frac{\omega^3}{e^{\hbar\omega/k_B T} - 1} d\omega \\ &= \frac{3V\hbar}{2\pi^2 v^3} \int_0^{\omega_p} \left(-\frac{\omega^3}{(e^{\hbar\omega/k_B T} - 1)^2} \cdot \left(-\frac{\hbar\omega}{k_B T^2} e^{\hbar\omega/k_B T} \right) \right) d\omega = \frac{3V\hbar}{2\pi^2 v^3} \frac{\hbar}{k_B T^2} \int_0^{\omega_p} \frac{\omega^4 e^{\hbar\omega/k_B T}}{(e^{\hbar\omega/k_B T} - 1)^2} d\omega \\ &= \frac{3V\hbar^2}{2\pi^2 v^3 k_B T^2} \int_0^{x_p} \frac{\left(\frac{k_B T}{\hbar} x \right)^4 e^x}{(e^x - 1)^2} \frac{k_B T}{\hbar} dx = \frac{3V\hbar}{2\pi^2 v^3 T} \left(\frac{k_B T}{\hbar} \right)^4 \int_0^{x_p} \frac{x^4 e^x}{(e^x - 1)^2} dx \\ &= 9Nk_B \left(\frac{T}{T_D} \right)^3 \int_0^{x_p} \frac{x^4 e^x}{(e^x - 1)^2} dx \end{aligned} \quad (33.66)$$

Si la température est petite, nous avons alors:

$$C_{V,T \rightarrow 0} = 9Nk_B \left(\frac{T}{T_D} \right)^3 \int_0^{+\infty} \frac{x^4 e^x}{(e^x - 1)^2} dx \quad (33.67)$$

Ce qui est difficile à calculer. Alors nous passons plutôt par le cheminement suivant:

$$C_{V,T \rightarrow 0} = \left(\frac{\partial U}{\partial T} \right)_{V,T \rightarrow 0} = 9Nk_B \frac{\partial}{\partial T} \left[T \left(\frac{T}{T_D} \right)^3 \int_0^{+\infty} \frac{x^3}{e^x - 1} dx \right] \quad (33.68)$$

Et cette dernière intégrale nous est connue. Effectivement, nous la démontrerons en détail plus bas en utilisant la fonction zêta de Riemann lors de démonstration de la constante de Stefan-Boltzmann:

$$\begin{aligned} C_{V,T \rightarrow 0} &= 9Nk_B \frac{\partial}{\partial T} \left[T \left(\frac{T}{T_D} \right)^3 \int_0^{+\infty} \frac{x^3}{e^x - 1} dx \right] = 9Nk_B \frac{\partial}{\partial T} \left[T \left(\frac{T}{T_D} \right)^3 \frac{\pi^2}{15} \right] \\ &= 9Nk_B \frac{\pi^2}{15T_D^3} \frac{\partial}{\partial T} T^4 = 3Nk_B \frac{\pi^2}{5T_D^3} 4T^3 = \frac{12Nk_B \pi^2}{5} \left(\frac{T}{T_D} \right)^3 \end{aligned} \quad (33.69)$$

Soit:

$$C_{V,T \rightarrow 0} = \frac{12Nk_B\pi^2}{5} \left(\frac{T}{T_D} \right)^3 \quad (33.70)$$

et les relevés expérimentaux correspondent bien à ce comportement!

Remarque: Cette capacité n'est pas constante. Elle dépend bien évidemment elle-même de la température et du matériau/fluide considéré défini par le nombre de degrés de liberté qui change en fonction des atomes/molécules (ce qui explique la différence de capacité calorifique entre les gaz monoatomiques et les autres).

En thermodynamique, il faut aussi toujours préciser quelle est la transformation considérée car ΔQ dépend de la nature de la transformation. Nous distinguerons en particulier la capacité calorifique à volume constant (isochore) notée:

$$C_V = \left(\frac{\delta Q}{\partial T} \right)_V \quad (33.71)$$

et la capacité calorifique à pression constante (isobare) souvent utilisée en chimie et notée:

$$C_P = \left(\frac{\delta Q}{\partial T} \right)_P \quad (33.72)$$

La différence entre la chaleur spécifique à pression constante et la chaleur spécifique à volume constant tient évidemment au travail qui doit être fourni pour dilater le corps en présence d'une pression externe. Ces deux expressions mettent aussi en évidence une différence fondamentale dans les montages expérimentaux de mesure, c'est-à-dire des situations physiquement différentes. En pratique, pour utiliser ce concept de capacité calorifique, nous devons identifier pour notre expérience quel C utiliser selon la façon dont le montage/machine aura été constitué.

Il est alors assez intuitif que la chaleur spécifique à pression constante est toujours strictement plus grande que la chaleur spécifique à volume constant. Nous démontrerons d'ailleurs plus bas que nous avons même précisément pour un gaz parfait:

$$C_P - C_V = nR = Nk \quad (33.73)$$

Si au lieu de considérer le système entier, nous rapportons les mesures à une unité de masse (ce qui est beaucoup plus utile en pratique!), nous définissons alors la "**capacité calorifique massique**" à volume constant (isochore):

$$c_V = \frac{1}{M} \left(\frac{\delta Q}{\partial T} \right)_V \quad (33.74)$$

et la capacité calorifique massique à pression constante (isobare):

$$c_P = \frac{1}{M} \left(\frac{\delta Q}{\partial T} \right)_P \quad (33.75)$$

Si au lieu de considérer le système entier, nous rapportons les mesures à une mole du corps considéré (ce qui est beaucoup plus utile en chimie!), nous définissons le "**capacité calorifique molaire**" à volume constant (isochore):

$$c_V^M = \frac{1}{n} \left(\frac{\partial Q}{\partial T} \right)_V \quad (33.76)$$

et la capacité calorifique molaire à pression constante (isobare):

$$c_P^M = \frac{1}{n} \left(\frac{\partial Q}{\partial T} \right)_P \quad (33.77)$$

où n est le nombre de moles du système.

Nous avons donc bien évidemment par extension, en passant au cas discret, la relation très utile en pratique:

$$\Delta Q = m \cdot c \cdot \Delta T \quad (33.78)$$

où c est la capacité calorifique polytropique massique.

Il s'agit d'une relation très importante dans la pratique et dont les applications au quotidien sont légion car elle permet dans des cas idéalisés de calculer la chaleur utilisée ou fournie lorsqu'une certaine quantité de matière subit un changement de température ou de calculer la température finale de deux corps mis en contact.

Pensez-y lorsque vous prenez votre douche ou que vous chauffez une casserole d'eau. Vous connaissez alors ΔT ainsi que la quantité d'eau M utilisée pour votre douche/bain ou casserole. Cela vous donnera l'énergie totale utilisée ΔQ et en connaissant le coût de l'énergie dans votre ville vous en déduirez rapidement le coût d'une douche, d'un bain ou d'une casserole d'eau bouillante!

Une autre application sympathique est de calculer l'augmentation de température (dans un cas idéal) d'un objet lâché à une certaine hauteur sous l'hypothèse que toute son énergie potentielle est convertie en chaleur après avoir touché le sol (le seul problème étant de trouver la valeur de capacité calorifique polytropique massique dans les tables):

$$mgh = mc_j \Delta T \quad (33.79)$$

où c_j est la capacité énergétique polytropique massique (il s'agit simplement de c mais donné en joules par kilo et par kelvin plutôt qu'en calories par kilo et par kelvin). Donc après réarrangement nous avons:

$$\Delta T = \frac{mgh}{mc_j} \quad (33.80)$$

Le cas important de la température d'équilibre finale de deux corps mis en contact est simple à obtenir. Effectivement, comme la chaleur perdue par l'un des deux corps sera gagnée par l'autre, nous avons en choisissant convenablement ΔT pour chacun des deux corps afin que ΔQ soit de même signe:

$$c_1 m_1 (T_1 - T_f) = c_2 m_2 (T_f - T_2) \quad (33.81)$$

Soit après un réarrangement élémentaire:

$$T_f = \frac{c_1 m_1 T_1 + c_2 m_2 T_2}{c_1 m_1 + c_2 m_2} \quad (33.82)$$

Exemple:

Nous avons donc dans un système sans échange de travail mécanique mais uniquement de chaleur la relation suivant:

$$\Delta U = C \Delta T = C (T_2 - T_1) \quad (33.83)$$

Or, la puissance est l'énergie divisée par le temps (cf. chapitre de Mécanique Classique). Il vient alors (attention à ne pas confondre la notation de la puissance avec celle de la pression!):

$$P = \frac{C (T_2 - T_1)}{t} \quad (33.84)$$

et si la capacité calorifique est donnée sous forme massique à pression constante:

$$P = \frac{m \cdot c_p (T_2 - T_1)}{t} = \frac{m}{t} c_p (T_2 - T_1) \quad (33.85)$$

Ce qui permet déjà de calculer la puissance à fournir pour chauffer une quantité de matière donnée à une température donnée et un temps donné sans considérer de changement d'état sinon quoi il faut prendre en compte la chaleur latente.

La fraction précédente peut aussi s'exprimer comme le débit volumique d'un fluide $\dot{V}$ en $[m^3 s^{-1}]$ multiplié par la densité du fluide:

$$P = \dot{V} \cdot \rho \cdot c_p (T_2 - T_1) \quad (33.86)$$

ce qui permet de connaître la puissance fournie par un débit d'eau entrant dans un radiateur à une certaine température et sortant de celui-ci avec le même débit mais avec une autre température.

Nous retrouvons aussi cette relation parfois dans la littérature sous la forme suivante:

$$E = m c_p (T_2 - T_1) = \rho V c_p (T_2 - T_1) \quad (33.87)$$

après avoir bien évidemment multiplié les deux côtés de l'égalité par le temps.

Dans certaines transformations, nous avons un apport ou un retrait de chaleur alors que la température du système reste constante (par exemple dans les transformations de changement de phase d'un corps). Cela provient d'un autre phénomène:

Définition: "L'**enthalpie de changement d'état**" ou "**chaleur latente**", notée L et donnée en joules par kilogramme, correspond à la quantité de chaleur nécessaire à l'unité de quantité de matière ou de masse d'un corps pour qu'il change d'état. C'est une valeur très importante en pratique pour effectuer des calculs.

Nous avons donc bien évidemment par extension la relation très utile en pratique:

$$Q = m \cdot L \quad (33.88)$$

qui donne donc la quantité de chaleur à fournir pour faire changer de phase une certaine quantité de masse m (le signe de Q est alors par convention négatif). Inversement, si la quantité de matière revient à un niveau d'énergie plus faible, cette quantité de chaleur latente sera donnée à l'environnement et non absorbée et le signe de Q est alors positif!

Remarque: Pour le passage de l'état liquide à l'état de vapeur, nous parlerons "d'enthalpie de vaporisation".

L'enthalpie échangée lors du changement d'état résulte de la modification (rupture ou établissement) de liaisons interatomiques ou intermoléculaires. Il existe trois états physiques principaux pour tout corps pur: l'état solide, l'état liquide et l'état gazeux. Les liaisons sont plus fortes dans l'état solide que dans l'état liquide et ces liaisons sont quasi absentes dans l'état gazeux. Il existe comme nous le savons déjà un quatrième état obtenu à très haute température où la matière se trouve sous la forme d'un plasma d'ions et d'électrons (cf. chapitre de Mécanique des Milieux Continus).

Nous différencions trois modes d'échange de chaleur:

1. La "**convection**": fluide venant alternativement au contact d'un corps chaud et d'un corps froid (phénomène typique des mouvements de l'atmosphère terrestre).
2. Le "**rayonnement**": un corps à température donnée émet un rayonnement électromagnétique susceptible d'être absorbé par un autre corps.
3. La "**conduction**": le transfert d'énergie se produit des zones chaudes vers les zones froides par collision des particules les plus excitées (zone chaude) avec les particules voisines moins excitées et ainsi de suite de proche en proche.

8. ÉNERGIE INTERNE

Dans le chapitre de Mécanique Classique, nous avons vu (deuxième théorème de König) que l'énergie totale d'un corps par rapport à un référentiel galiléen extérieur R' au centre de masse était donnée par la somme de l'énergie cinétique et potentielle par rapport à ce même référentiel, additionnée de l'énergie cinétique et potentielle de ce même corps par rapport au référentiel assimilé à son centre de masse R (référentiel barycentrique).

Ce que nous écrivons sous la forme:

$$E_{tot} = (E_{c,G/R} + E_{p,G/R}) + (E_{c,G/R'} + E_{p,G/R'}) = (E_{c,G/R} + E_{p,G/R}) + U \quad (33.89)$$

où U est la grandeur énergétique (énergie interne) propre au centre de masse du corps concerné.

Sous une forme plus simplifiée, la relation précédente s'écrit traditionnellement en thermodynamique:

$$E_{tot} = E_c + E_p + U \quad (33.90)$$

Le premier principe de la thermodynamique s'écrit alors:

$$Q + W = E_{tot} = E_c + E_p + U \quad (33.91)$$

Donc l'énergie totale est la somme des énergies mécaniques macroscopiques (cinétiques et potentielles) et microscopiques (énergie interne).

En règle générale, les systèmes étudiés en thermodynamique sont globalement au repos ($v_G = 0$) par rapport à l'expérimentateur et donc $\Delta E_c = 0$. De même, nous avons en général un champ de potentiel constant et isotrope dans la chambre d'expérimentation ce qui implique $\Delta E_p = 0$.

En conséquence de ces simplifications la loi de conservation de l'énergie, ou premier principe de la thermodynamique, se réduit comme nous l'avons déjà mentionné à l'énergie interne tel que:

$$Q + W = U \quad (33.92)$$

Remarque: L'énergie interne U n'est souvent donnée qu'à une constante additive près, c'est la raison pour laquelle certains l'appellent à juste titre "**surénergie interne**". Comme nous l'avons démontré dans le chapitre de Mécanique Classique, l'énergie totale d'un système, est la somme des énergies élémentaires de celui-ci. Ainsi, l'énergie interne est une grandeur extensive.

Considérons maintenant un système décrit par les variables d'état thermodynamiques:

$$x_1, x_2, \dots, x_n \quad (33.93)$$

Dans une transformation élémentaire $dx_1, dx_2, \dots, dx_n$ le travail élémentaire dW (si nous ne nous intéressons qu'à cette forme d'énergie) se mettra sous la forme (cf. chapitre de Calcul Différentiel Et Intégral) d'une différentielle totale exacte:

$$dW = A_1 dx_1 + A_2 dx_2 + \dots + A_n dx_n \quad (33.94)$$

si nous avons la relation suivante, démontrée dans le chapitre de Calcul Différentiel Et Intégral, qui est satisfaite (qui sont les coefficients des dx):

$$\frac{\partial A_i}{\partial x_j} = \frac{\partial A_j}{\partial x_i} \quad (33.95)$$

Sinon, si cette relation n'est pas satisfaite, ce qui impliquerait:

$$\frac{\partial A_i}{\partial x_j} \neq \frac{\partial A_j}{\partial x_i} \quad (33.96)$$

alors dans ce dernier cas, nous avons une différentielle totale inexacte (la différentielle dépend donc du chemin parcouru!) ce qui nous amènerait à écrire:

$$\delta W = A_1 dx_1 + A_2 dx_2 + \dots + A_n dx_n \quad (33.97)$$

Puisque le travail W est d'une manière ou une autre une force sur une distance, nous avons déjà démontré dans le chapitre de Mécanique Classique que celle-ci dépend du chemin parcouru et pas seulement du point de départ et final.

Dans la même transformation, la quantité de chaleur dQ se met sous une forme similaire:

$$dQ = B_1 dx_1 + B_2 dx_2 + \dots + B_n dx_n \quad (33.98)$$

si encore une fois la relation suivante est vérifiée pour les variables d'état:

$$\frac{\partial B_i}{\partial x_j} = \frac{\partial B_j}{\partial x_i} \quad (33.99)$$

ou sinon dans le cas contraire:

$$\delta Q = B_1 dx_1 + B_2 dx_2 + \dots + B_n dx_n \quad (33.100)$$

Les grandeurs caractéristiques d'une transformation du système mais dont la valeur dépend non seulement de l'état initial et de l'état final, mais aussi du chemin suivi sont appelées "**grandeurs de transfert**".

Rappelons qu'au vu de ce qui a déjà été étudié dans le chapitre de Calcul Différentiel Et Intégral, nous avons alors pour un cycle thermodynamique:

$$\Delta U = \oint dU = 0 \quad \Delta Q = \oint \delta Q \neq 0 \quad \Delta W = \oint \delta W \neq 0 \quad (33.101)$$

et c'est le fait que l'énergie interne soit une différentielle totale exacte qui impose logiquement que la chaleur Q soit une différentielle totale inexacte de par la propriété de linéarité de l'intégrale (sinon quoi elle n'annulerait pas par sommation la variation de travail W sur un cycle fermé). Effectivement, selon le premier principe de la thermodynamique, si l'énergie est conservative alors nous avons pour l'énergie interne d'un système fermé et isolé (sous forme discrète):

$$\Delta U = 0 \quad (33.102)$$

Ce qui implique que quelle que soit la transformation, les variations de travail et de chaleur sont données à l'intérieur de ce système conservatif par (sous forme discrète):

$$\boxed{\Delta Q + \Delta W = 0} \quad (33.103)$$

ou plus techniquement:

$$\oint dU = \oint (\delta Q + \delta W) = \underbrace{\oint \delta Q}_{\neq 0} + \underbrace{\oint \delta W}_{\neq 0} = 0 \Rightarrow \oint \delta Q = -\oint \delta W \quad (33.104)$$

Donc quelles que soient les façons dont les transformations se font à l'intérieur du système, l'état final et initial sont identiques en termes énergétiques. Ce qui nous impose que les transformations thermodynamiques soient indépendantes de la manière dont les phénomènes ont lieu à l'intérieur du système. ΔU est alors une différentielle totale exacte (cf. chapitre de Calcul Différentiel Et Intégral) que nous écrirons dU .

8.1. TRAVAIL DES FORCES MÉCANIQUES

Rappelons maintenant que nous avons vu dans le chapitre de Mécanique Classique que par définition, le travail est donné par une force sur une distance (quelle que soit l'origine de cette force: mécanique, rayonnante, nucléaire, etc.). Il s'ensuit donc la relation très importante en thermodynamique:

$$dW = \vec{F} \cdot d\vec{l} = P \cdot S \cdot dx = P \cdot dV \quad (33.105)$$

qui exprime donc, à pression constante (isobare), la variation de l'énergie due au travail des forces extérieures de pression sur un système (généralement un gaz en thermodynamique...) dont le volume a varié (sans être restreint par une frontière rigide!).

Bien évidemment, si la variation de volume est nulle ou la pression est nulle... la variation de l'énergie due au travail des forces de pression sera alors nulle...

A l'équilibre, la pression P est prise comme étant celle du système considéré (pression interne) soit celle de l'atmosphère environnante puisque la pression est alors égale (sinon il n'y aurait pas équilibre...).

De même, la variation de volume est prise comme étant soit celle du système considéré (volume propre) soit la variation de l'atmosphère environnante puisque de toute manière la variation de volume sera la même pour les deux!

De plus, nous voyons que le chemin intervient dans l'expression précédente du travail et donc que celui-ci est une grandeur qui dépend du chemin parcouru (cf. chapitre de Calcul Différentiel Et Intégral). Ceci implique que nous devons écrire δW au lieu de dW . Il existe bien sûr quantité d'autres manières d'exprimer le travail, mais par la définition de celui-ci même, il s'agira toujours d'une différentielle inexacte.

Ce résultat a une implication directe sur l'expression de la variation de chaleur pour laquelle le théorème du Viriel (cf. chapitre de Mécanique des milieux continus), que celle-ci est donnée par son agitation thermique. Il convient de rappeler que cette agitation est donnée par l'énergie cinétique moyenne et que l'énergie cinétique existe de par l'application d'une force sur une distance pour chaque particule. Ainsi, la chaleur est elle, aussi une différentielle inexacte δQ !

Finalement, nous avons dans le cas d'un fluide (liquide ou gazeux) ayant une variation d'énergie interne polytropicque:

$$dU = \delta Q + \delta W \quad (33.106)$$

Au cours de son évolution, le volume V du système peut donc varier. Si nous considérons une évolution infiniment petite au cours de laquelle le volume varie de dV et si nous notons P_e la pression extérieure subie par le système, nous écrirons dans un cadre plus général (le signe négatif devant la pression est une convention!):

$$\delta W = -P_e dV + dT \quad (33.107)$$

où $-P_e dV$ est le travail dit de "**refoulement extérieur**" (si $dV > 0$, l'augmentation du volume du système exprime ainsi la fourniture d'un travail à l'environnement extérieur, ce qui explique le signe - indiquant que l'énergie interne du système diminue) et où dT (à ne pas confondre avec une variation de température!!) est une autre forme d'énergie (si le système est réactif au niveau chimique par exemple, il peut fournir une éventuelle énergie chimique aussi).

Sauf systèmes particuliers (piles, accumulateurs,...) la seule énergie échangée est le travail des forces de pression de sorte que nous n'ayons plus que:

$$\delta W = -P_e dV \quad (33.108)$$

et alors nous pouvons à loisir écrire dW au lieu d'utiliser la différentielle inexacte puisque ce travail n'est dépendant que d'une variable d'état (la notation dW étant souvent utilisée en physique dans ce cas)!

Par ailleurs, la plupart des systèmes étudiés sont par hypothèse, à l'équilibre, à la même pression interne que la pression extérieure (atmosphère environnante). Ce qui nous autorise dans ce cas particulier à écrire:

$$P_e = P \quad (33.109)$$

où P est donc la pression dans et en dehors du système (le système est alors dit "isobare" comme nous le savons déjà et plus rarement "monobare").

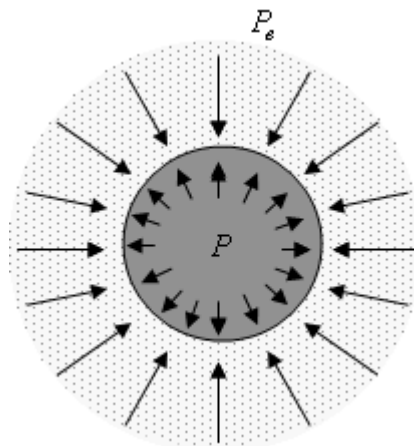


Figure: 33.2 - Exemple d'une sphère à l'équilibre avec l'environnement

Dès lors, nous avons dans ce cas particulier:

$$\begin{aligned} \delta W &= -P dV \\ \Delta W &= -\int_1^2 P dV \end{aligned} \quad (33.110)$$

La quantité de chaleur Q_V mise en jeu dans une transformation isochore (à volume constant) se réduit alors bien évidemment à la variation de l'énergie interne telle que (attention à la notation traditionnelle Q_V qui peut faire oublier que nous avons affaire à une variation de chaleur!):

$$\delta Q \equiv Q_V = U_2 - U_1 \quad (33.111)$$

Si nous considérons un système évoluant d'un état 1 vers un état 2 dans une atmosphère environnante isobare (à pression constante) et donc que sa propre pression interne est à l'équilibre avec cette atmosphère, nous avons alors le "**travail des forces mécaniques à pression constante**":

$$\Delta W = - \int_1^2 P dV = -P(V_2 - V_1) \quad (33.112)$$

Ou si le système évolue d'un état 1 vers un état 2 dans un environnement isotherme, nous avons alors le "**travail des forces mécaniques à température constante**":

$$\Delta W = - \int_1^2 P dV = - \int_1^2 \left(\frac{nRT}{V} \right) dV = -nRT \int_1^2 \frac{dV}{V} = -nRT \ln \left(\frac{V_2}{V_1} \right) \quad (33.113)$$

Si on veut produire une quantité raisonnable de travail, il semble à priori que le volume devra atteindre des dimensions totalement irréalistes. Aussi nous construisons la majorité des machines thermiques avec un retour cyclique.

8.2. ENTHALPIE

Considérons le cas isobare très courant dans les laboratoires de chimie puisque les béchers sont ouverts à l'atmosphère ambiante. Si nous considérons une telle situation, nous avons alors la variation d'énergie interne:

$$U_2 - U_1 = \Delta Q - P(V_2 - V_1) \rightarrow \Delta Q = U_2 - U_1 + P(V_2 - V_1) = (U_2 + PV_2) - (U_1 + PV_1) \quad (33.114)$$

où donc la pression P et le volume V sont les variables d'état internes au système étudié! Il est important de remarquer à nouveau que si la variation de chaleur ΔQ est nulle et que le volume interne augmente la variation d'énergie interne U est alors négative, ce qui sera interprété par l'expérimentateur comme un emprunt d'énergie au système extérieur et donc il est courant de parler de perte ou de réaction endothermique!

La quantité de chaleur Q_p mise en jeu dans une transformation isobare (l'indice de Q_p est indiqué dans ce sens... peut faire oublier que nous avons affaire à une variation de chaleur) est égale à la variation de deux termes à forme identique:

$$\Delta Q \equiv Q_p = (U_2 + PV_2) - (U_1 + PV_1) = H_2 - H_1 = \Delta H \quad (33.115)$$

où nous définissons donc une nouvelle fonction d'état commode, "**l'enthalpie**" H (grandeur extensive) dans une transformation isobare comme donnée par:

$$H = U + PV = U + nRT \quad (33.116)$$

où n est le nombre de moles internes au gaz parfait subissant le changement de volume! Nous voyons par ailleurs que si la pression environnante est nulle, l'enthalpie est égale à l'énergie interne. Le fait qu'il y ait des forces de pressions extérieures (ou intérieures) rajoute une énergie au système qui définit donc le concept d'enthalpie!

La relation antérieure exprime donc le fait que lorsqu'un système évolue à pression constante, la chaleur reçue (ou échangée par le système avec le milieu extérieur) est égale à sa variation d'enthalpie.

Il vient également avec cette définition une nouvelle écriture possible pour la capacité calorifique à pression constante très fréquemment utilisée en chimie:

$$C_P = \left(\frac{\delta Q}{\delta T} \right)_P = \frac{\partial H}{\partial T} \quad (33.117)$$

De manière plus explicite, nous avons encore pour la relation antérieure en utilisant la loi des gaz parfaits:

$$H = U + PV = \frac{ddl}{2} nRT + nRT = \frac{ddl + 2}{2} nRT \quad (33.118)$$

où pour rappel (cf. chapitre de Mécanique de Milieux Continus) le nombre de degrés de liberté ddl pour un gaz parfait monoatomique est de 3. Le dernier terme de cette relation n'est évidemment pas valable pour les fluides!

Parfois, nous écrivons évidemment également la définition de l'enthalpie sous la forme d'une variation telle que:

$$\Delta H = \Delta U + P\Delta V = nR\Delta T \quad (33.119)$$

dans le cas particulier d'application aux domaines de validité des gaz parfaits (cf. chapitre de Mécanique Des Milieux Continus) en ce qui concerne le dernier terme.

Il est important de bien mémoriser que le terme PV dans les relations précédentes représente le travail des forces de pression de l'atmosphère environnante sur le système ou, par équivalence, respectivement du système sur l'atmosphère environnante qui l'entoure. Mais la variable n est toujours le nombre de moles du gaz parfait du système étudié et non pas de l'atmosphère environnante.

L'enthalpie est un concept énormément utilisé en chimie thermique (voir chapitre du même nom) et nous l'utiliserons sans cesse lors de son étude.

Dans la pratique il est cependant difficile (voire impossible) de connaître l'énergie interne. Nous calculons alors plutôt la relation suivante pour une mole (n valant alors 1) de gaz parfait:

$$H - U = RT = PV \quad (33.120)$$

qui est donc une valeur strictement positive donnant l'énergie (ou plutôt le surplus d'énergie) disponible dans le gaz parfait à cause de la pression environnante (il n'y a qu'à poser la pression comme étant nulle pour le voir!).

Exemple:

Prenons une unité de volume molaire de gaz parfait aux conditions normales de température et de pression (quel que soit ce gaz parfait, son volume molaire sera dans ces conditions toujours de 24 litres selon la loi des gaz parfaits!).

Le travail des forces de pression qui ont été nécessaires pour amener ce gaz parfait aux conditions susmentionnées est alors de:

$$H - U = PV \cong 1.015 \cdot 10^5 \cdot \frac{24}{1'000} = 2'436 [J] \quad (33.121)$$

ou:

$$H - U = R \cdot T \cong 293.15 \cdot 8.31 = 2'436 [J] \quad (33.122)$$

qui est donc l'énergie sous forme du travail des forces (mécaniques) de pression que nous pourrions récupérer d'un volume molaire d'air.

L'énergie interne sous forme de chaleur ET de travail mécanique que nous pourrions en théorie récupérer (le problème est de trouver comment...) de ce volume molaire est donné, pour un gaz monoatomique (voir plus bas pour la démonstration), par:

$$H = Q = c_p^M RT = U + PV = \frac{ddl + 2}{2} RT = \frac{5}{2} RT \cong 6 [KJ] \quad (33.123)$$

et nous pourrions calculer aussi l'énergie interne de liaison des électrons, et des noyaux des atomes... et l'équivalence masse et énergie, mais nous sortirions des cas industriels courants...

Pour l'eau, où nous ne pouvons utiliser que la première relation, car la deuxième n'est valable que pour les gaz parfaits, le volume molaire étant 1'000 fois plus petit, nous ne pouvons en tirer qu'une énergie que 1'000 fois plus petite (environ 2 [J]). Raison pour laquelle dans le cas des fluides, nous considérons qu'il n'y a aucune différence entre l'enthalpie et l'énergie interne.

Voyons maintenant d'autres implications théoriques des quelques éléments vus précédemment qui nous seront très utiles aussi bien en acoustique (cf. chapitre de Musique Mathématique) ou en mécanique des milieux continus (voir chapitre du même nom).

8.3. LOI DE LAPLACE

Nous avons démontré dans le chapitre de Mécanique Des Milieux Continus avec le théorème du Viriel que l'énergie interne (énergie cinétique) d'un gaz parfait monoatomique était donnée par:

$$U = \frac{3}{2} kNT \quad (33.124)$$

où nous avons repris la notation du chapitre de Mécanique Des Milieux Continus (un grand N au lieu d'un petit n).

Nous avons donc:

$$dU = \frac{3}{2} kNdT \quad (33.125)$$

Si le processus est à volume constant, nous supposons qu'il n'y a aucun travail mécanique (force sur une distance) fourni (collisions inélastiques sur les parois) et alors:

$$dU = \delta Q + \delta W = \delta Q_V = \frac{3}{2} k N dT \quad (33.126)$$

donc où dW est nul!

Il vient alors la possibilité de calculer la capacité calorifique à volume constant:

$$C_V = \left(\frac{\delta Q}{dT} \right)_V = \frac{3}{2} k N = \frac{3}{2} \frac{R}{N_A} N = \frac{3}{2} R \frac{N}{N_A} \quad (33.127)$$

où l'on observe dès lors, que pour un gaz parfait, la capacité calorifique volumique est indépendante du volume!! Nous avons in extenso pour une mole:

$$c_V^M = \frac{C_V}{N} = \frac{C_V}{N} N_A = \frac{3}{2} R \quad (33.128)$$

de sorte que nous pouvons écrire pour un gaz monoatomique parfait à volume constant l'énergie:

$$U = \frac{3}{2} R \frac{N}{N_A} T = C_V T \quad (33.129)$$

ce qui permet donc de connaître dans la pratique, et à partir de tables, la variation d'énergie interne d'un gaz en fonction de sa variation de température puisque:

$$\Delta U = C_V \Delta T \quad (33.130)$$

ce qui correspond évidemment à un travail fourni ou absorbé. Sous forme différentielle nous l'appelons la "première loi de Joule" (l'énergie interne d'un gaz parfait ne dépend que de la température):

$$dU = n C_V dT$$

Un cas intéressant de la relation antécédente est de se rappeler que comme nous sommes à volume constant nous avons alors:

$$\left. \begin{array}{l} P_1 V_1 = n R T_1 \\ P_2 V_2 = n R T_2 \end{array} \right\} \frac{P_1}{T_1} = \frac{P_2}{T_2} \Rightarrow T_2 = \frac{P_2}{P_1} T_1 \quad (33.131)$$

ce qui nous permet d'écrire:

$$\Delta U = C_V \Delta T = C_V (T_2 - T_1) = C_V \left(\frac{P_2 - P_1}{P_1} \right) T_1 \quad (33.132)$$

Donc s'il y a décompression, cette dernière relation est négative. Dans ce cas le gaz perd de la chaleur au profit de l'environnement raison pour laquelle lors d'une détente gazeuse le récipient refroidit!

Si le processus a lieu à pression constante (énergie cinétique constante des atomes du gaz) alors nous avons (voir théorème du Viriel):

$$\delta W = -PdV = -kNdT \quad (33.133)$$

(les collisions qui repoussent la paroi du volume font perdre de l'énergie au système d'où le signe "-").

Ainsi:

$$dU = \delta Q_P + \delta W \Rightarrow \delta Q_P = dU - \delta W = \frac{3}{2}kNdT + kNdT = \frac{5}{2}kNdT \quad (33.134)$$

Ainsi nous pouvons aussi calculer la capacité calorifique à pression constante:

$$C_P = \left(\frac{\delta Q}{dT} \right)_P = \frac{5}{2}kN = \frac{5}{2} \frac{R}{N_A} N = \frac{5}{2} R \frac{N}{N_A} \quad (33.135)$$

où l'on observe dès lors, que pour un gaz parfait, la capacité calorifique à pression constante est aussi indépendante de la pression!! Nous avons in extenso pour une mole:

$$c_P^M = \frac{C_P}{N} = \frac{C_P}{N} N_A = \frac{5}{2} R \quad (33.136)$$

Des deux résultats encadrés précédents, nous obtenons pour un gaz parfait monoatomique:

$$\gamma = \frac{C_P}{C_V} = \frac{5}{3} \quad (33.137)$$

avec la "relation de Mayer":

$$C_P - C_V = \frac{5}{2} R \frac{N}{N_A} - \frac{3}{2} R \frac{N}{N_A} = nR \quad (33.138)$$

qui nous donne par la même occasion puisque R est positif:

$$C_P > C_V \quad (33.139)$$

Si le gaz parfait est diatomique, il y a 5 degrés de liberté (3 pour la position du premier atome +3 pour la position du deuxième -1 pour la contrainte que la distance entre les deux est fixée) et nous avons alors:

$$dU \cong \frac{5}{2} kNdT \quad (33.140)$$

En faisant les mêmes développements, nous obtenons (valeur que nous utiliserons dans le chapitre de Musique Mathématique mais qui est utile dans de nombreux autres domaines):

$$\gamma = \frac{C_p}{C_v} = \frac{7}{5} \quad (33.141)$$

Quand un système isolé de gaz parfait subit une transformation adiabatique à pression constante, la variation d'énergie interne du système sera soustraite par la variation de travail interne. Ce qui traditionnellement se note par un signe négatif tel que (en utilisant le résultat obtenu plus haut):

$$dU = -\delta W = -PdV = C_v dT \quad (33.142)$$

Remarque: Attention!!! Rappelons que le choix du signe pour le travail W n'est qu'une convention de signe!! Ainsi, dans le présent cas d'étude il est pour tradition de mettre un "-" au lieu d'un "+". Mais cela ne change rien aux résultats qui vont suivre!!!

Soit:

$$PdV + C_v dT = 0 \quad (33.143)$$

Prenons maintenant l'équation des gaz parfaits $PV = nRT$ (sans collisions) et différencions. Nous obtenons:

$$PdV + VdP = Nk dT \Rightarrow dT = \frac{PdV + VdP}{Nk} \quad (33.144)$$

Soit en éliminant dT entre les deux dernières relations, nous obtenons:

$$PdV + C_v \frac{PdV + VdP}{Nk} = 0 \quad (33.145)$$

Soit après simplification et réarrangement des termes:

$$(Nk + C_v)PdV + C_v VdP = 0 \quad (33.146)$$

ce qui rapporté aux quantités de moles s'écrit (selon l'habitude des chimistes):

$$(R + c_v^M)pdV + c_v^M Vdp = 0 \quad (33.147)$$

bref... et en nous rappelant que:

$$C_p - C_v = nR = Nk \quad (33.148)$$

Nous avons:

$$C_p PdV + C_v VdP = 0 \quad (33.149)$$

En utilisant la définition du "**coefficient de Laplace**", appelé aussi "**coefficient adiabatique**" déjà rencontré plus haut:

$$\gamma = \frac{C_P}{C_V} \quad (33.150)$$

nous avons l'expression:

$$\gamma PdV + VdP = 0 \quad (33.151)$$

En remaniant:

$$\gamma \frac{dV}{V} + \frac{dP}{P} = 0 \quad (33.152)$$

Nous obtenons par intégration:

$$\gamma \int \frac{dV}{V} + \int \frac{dP}{P} = c^{te} \quad (33.153)$$

soit:

$$\ln(P) + \gamma \ln(V) = c^{te} \quad (33.154)$$

qui est équivalent en utilisant les propriétés des logarithmes à:

$$PV^\gamma = c^{te} \quad (33.155)$$

Il s'agit de la "**loi de Laplace**" qui donne la relation entre pression et volume dans une transformation adiabatique d'un gaz (ce qui ne signifie pas que la température est constante rappelons-le mais seulement que l'échange de chaleur avec le système extérieur est nul ou négligeable!). Nous retrouverons d'ailleurs cette relation dans le chapitre de Génie Météo.

Ainsi, nous avons aussi l'information qui peut être utile dans l'industrie:

$$\begin{aligned} W &= \int_{V_1}^{V_2} PdV = c^{te} \int_{V_1}^{V_2} V^{-\gamma} dV = c^{te} \left[\frac{V^{1-\gamma}}{1-\gamma} \right]_{V_1}^{V_2} = c^{te} \frac{V_2^{1-\gamma} - V_1^{1-\gamma}}{1-\gamma} = \frac{c^{te} V_2^{-\gamma} V_2 - c^{te} V_1^{-\gamma} V_1}{1-\gamma} \\ &= \frac{P_2 V_2 - P_1 V_1}{1-\gamma} = \frac{Nk(T_2 - T_1)}{1-\gamma} \end{aligned} \quad (33.156)$$

8.4. COEFFICIENTS THERMOÉLASTIQUES

Si nous différencions $V(P,T)$ nous avons (cf. chapitre de Calcul Différentiel Et Intégral):

$$dV = \frac{\partial V}{\partial T} dT + \frac{\partial V}{\partial P} dP \quad (33.157)$$

ou autrement écrit:

$$dV = \left(\frac{\partial V}{\partial T} \right)_P dT + \left(\frac{\partial V}{\partial P} \right)_T dP \quad (33.158)$$

Nous avons introduit de manière naturelle dans le chapitre de Musique Mathématique un coefficient nommé "**coefficient de compressibilité isotherme**" (l'indice T indique que la température est maintenue constante):

$$\chi_T = -\frac{1}{V} \left(\frac{\partial V}{\partial P} \right)_T \quad (33.159)$$

que nous pouvons réintroduire ici:

$$dV = \left(\frac{\partial V}{\partial T} \right)_P dT - \chi_T V dP \quad (33.160)$$

De même il serait intéressant d'avoir un autre coefficient pour le premier terme qu'il suffirait de définir par analogie (l'indice V et là pour indiquer qu'il s'agit d'un coefficient volumique et le P pour indiquer que nous sommes à pression constante):

$$\alpha_{V,P} = \frac{1}{V} \left(\frac{\partial V}{\partial T} \right)_P \quad (33.161)$$

appelé "**coefficient de dilatations volumique isobare**".

Ainsi, nous avons:

$$dV = \alpha_{V,P} V dT - \chi_T V dP \quad (33.162)$$

Soit nous avons le travail mécanique (la différentielle totale est inexacte car nous avons plus d'une variable d'état) en multipliant par la pression pour avoir les bonnes unités:

$$\delta W = \alpha_{V,P} P V dT - \chi_T P V dP \quad (33.163)$$

Pour une transformation isotherme le premier terme est nul, et pour une transformation isobare, c'est le second qui est nul.

Les données des coefficients thermoélastiques (mesurés expérimentalement) doivent permettre de remonter à l'équation d'état des gaz parfaits par intégration de $V(P,T)$, ce qui est licite, puisque V est une fonction d'état. Dans le cas du gaz parfait par exemple, nous pouvons écrire par intuition des dimensions des constantes (nous présenterons plus loin de notre étude des équations d'état le cheminement mathématique inverse pour des raisons pédagogiques):

$$\alpha = \frac{1}{T} \quad \chi = \frac{1}{P} \quad (33.164)$$

Nous avons donc:

$$dV = \frac{V}{T} dT - \frac{V}{P} dP \quad (33.165)$$

et donc en multipliant par la pression:

$$PdV = \frac{PV}{T} dT - VdP \Rightarrow PdV + VdP = PV \frac{dT}{T} \quad (33.166)$$

Ce qui conduit à:

$$d(PV) = PV \frac{dT}{T} \quad (33.167)$$

Soit:

$$\frac{d(PV)}{PV} = \frac{dT}{T} \quad (33.168)$$

Ce qui donne immédiatement après intégration:

$$\ln(PV) = \ln(T) + c^{te} \Rightarrow \ln\left(\frac{PV}{T}\right) = c^{te} \quad (33.169)$$

Soit:

$$PV = c^{te} T \quad (33.170)$$

Cela dit, nous avons différencié V pour obtenir deux coefficients tels que:

$$dV = \alpha V dT - \chi V dP \quad (33.171)$$

Nous pourrions faire de même pour la pression et la température et nous avons alors au total trois relations:

$$\begin{aligned} dV &= \left(\frac{\partial V}{\partial P}\right)_T dP + \left(\frac{\partial V}{\partial T}\right)_P dT \\ dP &= \left(\frac{\partial P}{\partial T}\right)_V dT + \left(\frac{\partial P}{\partial V}\right)_T dV \\ dT &= \left(\frac{\partial T}{\partial P}\right)_V dP + \left(\frac{\partial T}{\partial V}\right)_P dV \end{aligned} \quad (33.172)$$

mais parmi les 6 facteurs que nous voyons dans ces trois relations, quatre sont déjà définis (certains sont l'inverse des coefficients définis plus haut). Il manque par contre la définition d'un seul coefficient pour les deux facteurs manquants. Nous choisissons celui qui dans la pratique est le plus souvent utilisé en analogie avec les autres coefficients:

$$\beta = \frac{1}{P} \left(\frac{\partial P}{\partial T}\right)_V \quad (33.173)$$

appelé "**coefficient d'augmentation de pression isochore**".

Nous avons ainsi les trois coefficients très utilisés dans la pratique:

$$\chi_T = -\frac{1}{V} \left(\frac{\partial V}{\partial P} \right)_T \quad \alpha_{V,P} = \frac{1}{V} \left(\frac{\partial V}{\partial T} \right)_P \quad \beta = \frac{1}{P} \left(\frac{\partial P}{\partial T} \right)_V \quad (33.174)$$

respectivement et dans l'ordre:

1. Coefficient de compressibilité isotherme.
2. Coefficient de dilatation isobare
3. Coefficient d'augmentation de pression isochore

Nous retrouverons ces coefficients lors de notre étude des mouvements de convections dans le chapitre de Génie Marin & Météo.

9. CHALEUR

Avant de continuer, il va à tout prix nous falloir éliminer une des deux plus grandes difficultés dans la bonne compréhension de la thermodynamique (mise à part celle de la différenciation entre les différentielles totales exactes et inexactes qui a déjà été réglée dans le chapitre de Calcul Différentiel Et Intégral):

- La différence entre la chaleur et la température
- La différence entre l'énergie-travail et l'énergie-chaleur

Comme nous l'avons vu dans le chapitre de Mécanique Des Milieux Continus, la température caractérise un état d'équilibre thermodynamique et traduit l'existence d'une agitation thermique (théorème du Viriel) et elle peut varier lorsque l'extérieur fournit un travail $W > 0$. Cependant, l'expérience nous montre que c'est en "chauffant" un système que nous augmentons le plus aisément sa température. Mais qu'est-ce donc la chaleur?:

Considérons un système thermodynamique à l'équilibre, et écrivons son énergie interne totale E comme étant la somme des produits de l'énergie E d'un micro-état i par la taille de la population P de ce même micro-état i :

$$E = \sum_i E_i P_i \quad (33.175)$$

Sa variation au cours d'une transformation infinitésimale (que nous supposons à nombre de particules constantes) est (différentielle totale):

$$dE = \left(\frac{\partial E_i}{\partial P_i} \right)_{E_i} dP_i + \left(\frac{\partial E_i}{\partial E_i} \right)_{P_i} dE_i = \sum_i E_i dP_i + \sum_i P_i dE_i \quad (33.176)$$

où dE_i est le déplacement de l'énergie E_i du micro-état i provoqué par la transformation, et dP_i la variation de la population P_i de ce même état i .

Nous avons vu plus haut que sous certaines conditions, si le système au cours de cette transformation infinitésimale subit une variation de travail δW et de chaleur δQ , son énergie interne varie de:

$$dE = dU = \delta W + \delta Q \quad (33.177)$$

Nous pouvons maintenant comparer cette dernière relation (exprimant le premier principe de la thermodynamique) avec celle qui la précède (découlant de la mécanique statistique) et identifier les termes de la manière suivante:

$$dU = \sum_i E_i dP_i + \sum_i P_i dE_i = \delta Q + \delta W \quad (33.178)$$

Examinons le cas d'une transformation dans laquelle le système reçoit seulement de la chaleur:

$$\delta W = 0 \quad (33.179)$$

Dans cette situation aucun des paramètres extérieurs au système ne varie en général dans la transformation de sorte que $dE_i = 0$. Nous en déduisons:

$$\delta Q = \sum_i E_i dP_i \quad (33.180)$$

Ainsi, lorsqu'un système reçoit seulement de la chaleur, son énergie varie par modification des populations de ses états microscopiques: si la quantité de chaleur reçue est positive, la probabilité des états d'énergie élevée augmente, au détriment de celle des états de plus basse énergie. Enfin, si nous tenons compte que le système doit être à l'équilibre (l'état le plus probable comme nous l'avons démontré dans le chapitre de Mécanique Statistique) dans les états initial et final de la transformation, nous constatons, comme nous pouvions nous y attendre, que la température du système varie: elle a augmenté si la quantité de chaleur reçue est positive (cela se démontre en choisissant une distribution canonique pour décrire les états d'équilibre macroscopiques du système).

Cela explique la confusion fréquente entre ces deux concepts très différents que sont température et chaleur. Cette confusion est accentuée par le décalage entre le langage quotidien et la terminologie scientifique. Dans le langage quotidien, lorsque nous parlons de chaleur d'un corps, nous affirmons en réalité que sa température est élevée. La confusion est regrettable parce que la notion de chaleur est bien présente en physique mais sa signification est autre.

Ainsi, chauffer un système, c'est lui fournir de la chaleur, c'est augmenter son énergie interne (le nombre de micro-états de haute énergie) par des moyens qui ne sont pas purement mécaniques. La chaleur est donc une forme d'énergie particulière!

9.1. ENTROPIE

Un système macroscopique isolé tend vers l'équilibre. Il l'atteint en un temps fini (qui peut être extrêmement grand). L'état d'équilibre est unique: les exceptions à cette affirmation sont trop spéciales pour mériter une digression.

L'existence même d'un état d'équilibre est fondamentale pour la thermodynamique. Cependant, le processus de marche à l'équilibre ne résulte pas d'un dogme: il ne doit pas en exister en physique! Comme toute autre loi, il est soumis à vérification et doit être analysé. Une question, notamment, se pose: quelle est la contrepartie microscopique de la marche à l'équilibre, processus macroscopique.

Nous avons vu dans le chapitre de Mécanique Statistique que par définition: l'état d'équilibre est l'état qui correspond au plus grand nombre de configurations (micro-états) et est l'état le plus probable.

Ce qui nous avait amenés à la relation suivante:

$$S = \lambda \ln(\Omega) \quad (33.181)$$

où S représente l'espérance statistique de l'information sur les micro-états et que nous avons appelée "entropie". Il est évident que S a les unités de λ qui comme nous l'avions montré est une constante.

La question qui se pose alors en thermodynamique est: quelle est la constante qui permet de caractériser pour un gaz, fluide ou solide l'espérance mathématique du nombre des états.

Il vient alors en regardant toutes les relations qui existent en thermodynamique qu'une seule constante apparaît systématiquement dès qu'il s'agit de caractériser un état thermodynamique. Il s'agit de la constante de Boltzmann:

$$\lambda = k = 1.3806 \cdot 10^{-23} \left[J \cdot K^{-1} \right] \quad (33.182)$$

Donc S a les unités correspondant au rapport J/K qui permet donc de mesurer le degré de désordre d'un système au niveau microscopique. Intuitivement: plus l'entropie du système est élevée, moins ses éléments sont ordonnés et capables de produire des effets mécaniques, et plus grande est la part de l'énergie inutilisée ou utilisée de façon incohérente.

L'entropie est une grandeur extensive. Effectivement, nous avons montré que le choix du logarithme dans la loi de Boltzmann venait de ce que l'entropie d'un macro-état était l'espérance de l'ensemble des micro-états:

$$\begin{aligned} S &= -\lambda \sum_{i=1}^n p_i \log(p_i) = -\lambda \left(\frac{1}{\Omega} \log\left(\frac{1}{\Omega}\right) + \frac{1}{\Omega} \log\left(\frac{1}{\Omega}\right) + \dots + \frac{1}{\Omega} \log\left(\frac{1}{\Omega}\right) \right) \\ &= -\lambda \left(-\frac{1}{\Omega} \log(\Omega) - \frac{1}{\Omega} \log(\Omega) - \dots - \frac{1}{\Omega} \log(\Omega) \right) = \lambda \log(\Omega) \left(\frac{1}{\Omega} + \frac{1}{\Omega} + \dots + \frac{1}{\Omega} \right) \end{aligned} \quad (33.183)$$

ce qui nous avait amenés à:

$$S = \lambda \ln(\Omega) \quad (33.184)$$

et montre bien que l'entropie est une grandeur extensive car sommable sur les micro-états (complexions).

Ce qui reste difficile maintenant c'est de savoir si l'énergie dans les unités de l'entropie provient du travail W , de la chaleur Q ou des deux? Au fait la réponse est simple car dans notre développement de la loi de Boltzmann, à aucun moment le système (idéal) étudié n'a fourni un travail (mécanique). Donc la seule énergie mise en cause est celle de la chaleur.

Ainsi:

$$S = \frac{Q}{T} \quad (33.185)$$

Or, l'entropie ne peut pas être donnée en toute rigueur par:

$$S = \frac{\Delta Q}{\Delta T} \quad (33.186)$$

car c'est la définition de la chaleur spécifique. Par ailleurs, si le lecteur se rappelle de nos développements en Mécanique Statistique, l'étude se faisait dans un système isolé avec deux cavités. Donc si le passage d'une cavité à l'autre se fait très lentement (de façon à ce qu'il n'y ait pas une détente du gaz) la température restera constante (détente isotherme). Ce qui implique donc la définition (nous retrouvons donc la forme discrète du deuxième principe de la thermodynamique déjà introduit au début de ce chapitre9:

$$\Delta S = \frac{\Delta Q}{T} \quad (33.187)$$

Pour passer à la forme différentielle, il convient de se rappeler que la chaleur Q est une différentielle inexacte. Donc pour une transformation réversible:

$$dS = \frac{\delta Q}{T} \quad (33.188)$$

Donc l'entropie est une différentielle totale exacte (la chaleur dépend de la manière dont se fait la transformation, mais S ne dépend que de l'état final et initial de la chaleur Q) et comme:

$$S = k \ln(\Omega) \geq 0 \quad (33.189)$$

Nous avons finalement:

$$\boxed{dS = \frac{\delta Q}{T} \geq 0} \quad (33.190)$$

Si le système est dans une transformation adiabatique (sans échange de chaleur et de travail avec l'extérieur) alors le numérateur est nul et l'entropie est in extenso nulle aussi (raison pour laquelle un système adiabatique et forcément isentropique!). Sinon, le système prend de l'entropie à l'Univers dans une évolution naturelle.

Ce qui signifie que l'entropie (espérance de l'information intrinsèque) dans un système en contact avec l'extérieur ne peut qu'augmenter ou rester constante.

Ce qui est important c'est que tout processus (non adiabatique) convertissant de l'énergie d'une forme en une autre dans un système isolé en perd obligatoirement une partie sous forme de chaleur.

En ce qui concerne l'Univers... toute la question est de savoir s'il s'agit d'un système thermodynamique isolé ou non... Mais comme par hypothèse le Big Bang a engendré le temps et l'espace (puisque la relativité montre que espace et temps sont indissociables), alors rien n'existait avant lui puisque le temps n'existait pas avant... et il n'y avait pas d'espace puisque c'est le Big Bang est à l'origine. Donc il n'y rien autour (par hypothèse...).

Nous avons alors la relation très utile en mécanique des fluides (et en cosmologie):

$$\boxed{dU = \delta Q + \delta W = TdS - PdV} \quad (33.191)$$

qui se nomme "**identité thermodynamique**" relative à l'énergie interne U ou encore "**fonction caractéristique d'un fluide à l'équilibre**" ou plus souvent "**équation fondamentale de Gibbs**" (on remarquera au passage qu'il s'agit d'une somme de produits de grandeurs intensives et extensives). Si la quantité de matière n'est pas constante (matière insérée ou extraite du système), cette dernière relation s'écrit alors pour une variation de dn moles:

$$dU = TdS - PdV + \mu dn \quad (33.192)$$

où μ est un coefficient de proportionnalité qui porte le nom de "**potentiel chimique**" et qui peut être donc mesuré de par le fait que la relation précédente amène à la définition que:

$$\mu = \left(\frac{\partial U}{\partial n} \right)_{S,V} \quad (33.193)$$

qui dit que le potentiel chimique est l'énergie dont varie le gaz lorsqu'on lui injecte ou retire dn mole de gaz, tout en gardant constant le volume de l'enceinte et en interdisant les échanges de chaleur avec l'environnement.

On retrouve souvent dans la pratique la relation antécédente sous la forme suivante après réarrangement (et en supposant la quantité de matière constante):

$$dS = \frac{dU}{T} + \frac{P}{T}dV \quad (33.194)$$

Cette dernière relation est souvent assimilée au premier principe de la thermodynamique pour des systèmes fermés dont la variation d'énergie potentielle et cinétique globale est constante.

Remarquons qu'en utilisant la relation qui lie la variation d'énergie interne d'un gaz parfait monoatomique en fonction de sa variation de température à volume constant démontrée plus haut (première loi de Joule):

$$dU = C_V dT \quad (33.195)$$

et l'équation des gaz parfaits (cf. chapitre de Mécanique Des Fluides):

$$PV = nRT \quad (33.196)$$

Nous avons alors en utilisant la capacité calorifique molaire à volume constant (se rappeler que nous ne pouvons l'injecter dans cette relation que pour un gaz parfait car elle est alors indépendante du volume!):

$$dS = nc_V \frac{dT}{T} + nR \frac{dV}{V} \quad (33.197)$$

Et donc en intégrant entre deux états, il vient:

$$\begin{aligned}
 \Delta S &= \int_1^2 dS = \int_1^2 nc_V^M \frac{dT}{T} + nR \frac{dV}{V} = \int_1^2 nc_V^M \frac{dT}{T} + \int_1^2 nR \frac{dV}{V} \\
 &= nc_V^M \int_1^2 \frac{dT}{T} + nR \int_1^2 \frac{dV}{V} = nc_V^M \ln(T) \Big|_1^2 + nR \ln(V) \Big|_1^2 = nc_V^M \ln\left(\frac{T_2}{T_1}\right) + nR \ln\left(\frac{V_2}{V_1}\right)
 \end{aligned} \quad (33.198)$$

Soit la variation d'entropie d'un gaz parfait à pression constante:

$$\Delta S = nc_V \ln\left(\frac{T_2}{T_1}\right) + nR \ln\left(\frac{V_2}{V_1}\right) \quad (33.199)$$

Et comme nous avons démontré aussi plus haut que:

$$c_V^M = \frac{3}{2} R \quad (33.200)$$

Il vient alors en utilisant aussi les propriétés algébriques du logarithme:

$$\begin{aligned}
 \Delta S &= \frac{3nR}{2} \ln\left(\frac{T_2}{T_1}\right) + nR \ln\left(\frac{V_2}{V_1}\right) = nR \left[\frac{3}{2} \ln\left(\frac{T_2}{T_1}\right) + \ln\left(\frac{V_2}{V_1}\right) \right] \\
 &= nR \left[\ln\left(\left(\frac{T_2}{T_1}\right)^{\frac{3}{2}}\right) + \ln\left(\frac{V_2}{V_1}\right) \right] = nR \ln\left(\left(\frac{T_2}{T_1}\right)^{\frac{3}{2}} \cdot \frac{V_2}{V_1}\right)
 \end{aligned} \quad (33.201)$$

Et comme dans la pratique il est plus facile de mesurer une pression qu'un volume, nous avons en utilisant la loi des gaz parfaits où à température constante:

$$P_1 V_1 = P_2 V_2 \quad (33.202)$$

le résultat suivant très important dans la pratique considéré comme la variation d'entropie d'un gaz parfait à volume constant:

$$\Delta S = nR \ln\left(\left(\frac{T_2}{T_1}\right)^{\frac{3}{2}} \cdot \frac{P_2}{P_1}\right) \quad (33.203)$$

Pour l'une et l'autre des deux relations, il vient bien évidemment dans le cas isotherme que:

$$\Delta S = nR \ln\left(\left(\frac{T_2}{T_1}\right)^{\frac{3}{2}} \cdot \frac{P_2}{P_1}\right) \quad (33.204)$$

puisque le rapport des deux températures est égal à l'unité.

9.1.1. ÉCOULEMENT DE LA CHALEUR

Maintenant que nous avons introduit les notions élémentaires de la thermodynamique, faisons de suite un exemple d'application très important des premier et deuxième principes de la thermodynamique.

Considérons deux systèmes fermés 1 et 2 étant en contact thermique et formant un système total qui est supposé isolé thermiquement de l'environnement..

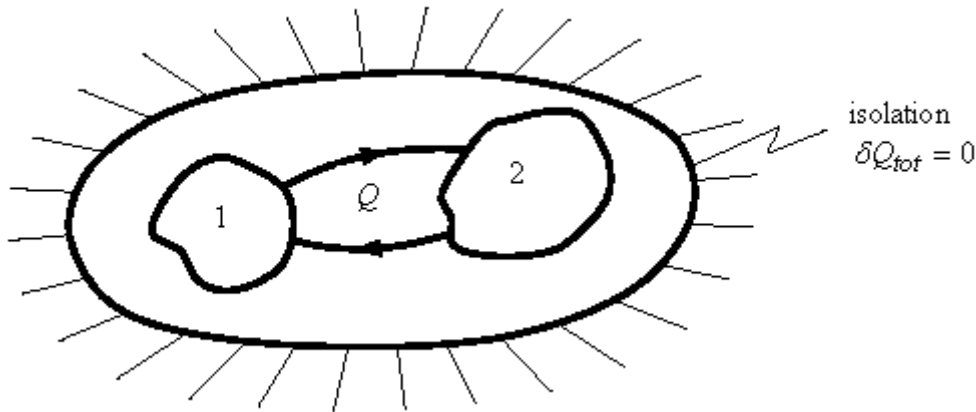


Figure: 33.3 - Système total isolé avec deux sous-systèmes en contact

Dans les deux systèmes, il n'y a pas d'éléments mécaniques, comme des pistons ou autres, qui peuvent travailler. Donc:

$$\delta W_1 = \delta W_2 = \delta W_{tot} = 0 \quad (33.205)$$

Les conditions et les hypothèses étant posées, nous utilisons les ingrédients de la thermodynamique pour faire notre recette. Il vient alors par application du premier principe tel que démontré plus haut dans un premier temps pour l'ensemble:

$$\begin{aligned} dU_{tot} &= dU_1 + dU_2 \\ dU_{tot} &= \underbrace{\delta Q_{tot}}_{=0} + \underbrace{\delta W_{tot}}_{=0} = 0 \end{aligned} \quad (33.206)$$

Dont nous déduisons que:

$$dU_1 + dU_2 = 0 \quad (33.207)$$

Nous avons aussi pour chaque sous-système:

$$\begin{aligned} \delta W_1 &= 0 = dU_1 - \delta Q_1 \\ \delta W_2 &= 0 = dU_2 - \delta Q_2 \end{aligned} \quad (33.208)$$

D'où nous déduisons de ce dernier système d'équations en sommant:

$$dU_1 + dU_2 = \delta Q_1 + \delta Q_2 \quad (33.209)$$

et comme nous avons démontré juste avant que:

$$dU_1 + dU_2 = 0 \quad (33.210)$$

Il vient:

$$dU_1 + dU_2 = 0 = \delta Q_1 + \delta Q_2 \quad (33.211)$$

et donc:

$$\delta Q_1 = -\delta Q_2 \quad (33.212)$$

Appliquons maintenant le deuxième principe (principe d'évolution). Nous avons:

$$dS_{tot} = dS_1 + dS_2 = \frac{\delta Q_1}{T_1} + \frac{\delta Q_2}{T_2} = \frac{\delta Q_{tot}}{T} \quad (33.213)$$

Or avec la relation antérieure, nous avons:

$$dS_{tot} = \frac{\delta Q_1}{T_1} + \frac{-\delta Q_1}{T_2} = \delta Q_1 \left(\frac{1}{T_1} - \frac{1}{T_2} \right) \quad (33.214)$$

Nous avons maintenant deux scénarios possibles, pour le premier, nous avons:

$$T_2 > T_1 \Rightarrow \left(\frac{1}{T_1} - \frac{1}{T_2} \right) > 0 \quad (33.215)$$

et lorsque:

$$T_2 > T_1 \Rightarrow \delta Q_1 > 0 \quad (33.216)$$

Donc:

$$dS_{tot} > 0 \quad (33.217)$$

Pour le deuxième scénario, nous avons:

$$T_1 > T_2 \Rightarrow \left(\frac{1}{T_1} - \frac{1}{T_2} \right) < 0 \quad (33.218)$$

et lorsque:

$$T_1 > T_2 \Rightarrow \delta Q_1 < 0 \quad (33.219)$$

Donc:

$$dS_{tot} > 0 \quad (33.220)$$

Nous concluons que quoi qu'il arrive, l'entropie augmente pour un système isolé. Ce résultat est donc conforme à ce que nous avons communiqué lors de notre présentation du deuxième principe. Donc ce résultat démontre l'impossibilité d'une machine thermique ditherme isolée fournissant un travail perpétuel puisque l'entropie ne peut faire qu'augmenter dans tous les cas!

Nous voyons aussi comme nous l'avons déjà mentionné que l'entropie définit la flèche du temps dans l'Univers (ou inversement).

9.2. CYCLE DE CARNOT

Le cycle de Carnot a pour objectif principal de calculer le rendement thermique d'un cycle thermodynamique d'une machine ditherme très idéalisée... (donc le modèle est très éloigné de la pratique!) basée sur une source de chaleur à laquelle nous prenons de la chaleur pour la transformer en travail et ensuite pour laquelle nous utilisons du travail afin de lui fournir de la chaleur et ainsi de suite en boucle (penser à une locomotive par exemple!). Nous parlons alors de "cycle réversible".

L'idée est assez simple... mais comme toujours il fallait y penser! Pour information nous avons choisi une approche mathématique du cycle de Carnot qui permet de se soustraire de toute intégrale est qui est plus proche du raisonnement fait à l'époque de Carnot par Carnot lui-même. Cette approche peut donc faire hérisser les cheveux à certains lecteurs...

Du point de vue de l'énergie interne, sur un cycle thermodynamique, il n'y aura pas eu de variations (puisque à chaque début de cycle nous devons nous retrouver dans la situation du cycle précédent). Donc:

$$\Delta U = 0 \quad (33.221)$$

Il s'agit jusqu'ici simplement du rappel du principe de conservation de l'énergie (premier principe de la thermodynamique).

Ce qui implique que sur un cycle le bilan de la chaleur Q_e empruntée à la machine pour produire un travail (mécanique) W_p et de la chaleur Q_i réinjectée ayant nécessité un travail (mécanique) W_i est nul tel que sur un cycle:

$$Q_e + Q_i + W_p + W_i = 0 \quad (33.222)$$

ou évidemment ce qui est emprunté aura, comme il est d'usage, un signe positif et ce qui est injecté un signe négatif!!!! Il est d'usage de noter cette dernière relation sous la forme condensée:

$$Q_e + Q_i + W = 0 \Leftrightarrow W = -(Q_e + Q_i) \quad (33.223)$$

Le rendement thermique η est par définition le rapport de ce que la machine peut donner réellement donner sous forme de travail (donc le différentiel entre ce qu'elle a donné W_p et ce qu'on doit lui donner pour qu'elle fonctionne W_i) sur la chaleur que l'on lui a injecté (à l'aide du travail W_i). Donc:

$$\eta = \frac{|W_p + W_i|}{Q_i} = \frac{|W|}{Q_i} = \frac{|-(Q_e + Q_i)|}{Q_i} = \frac{Q_e + Q_i}{Q_i} = 1 + \frac{Q_e}{Q_i} \quad (33.224)$$

ce qui est toujours inférieur à l'unité (et supérieur à 0) puisque la quantité de chaleur extraite (signe positif par convention) est obligatoirement inférieure ou égale à la quantité de chaleur injectée (signe négatif par convention).

Or, nous avons démontré que dans la condition d'une détente isotherme ou compression isotherme (donc le cycle est supposé comme étant très lent dans le modèle de Cournot), nous avons:

$$\Delta Q = T \Delta S \quad (33.225)$$

Or, dans notre cas Q_e, Q_i sont en réalité un abus de notation courant en thermodynamique car il s'agit de variation de chaleur (nous devrions noter objectivement $\Delta Q_e, \Delta Q_i$). Nous avons alors pour le rendement thermique réversible à compression ou détente isotherme:

$$\eta = 1 + \frac{T_e \Delta S_e}{T_i \Delta S_i} \quad (33.226)$$

étant donné que dans une transformation thermodynamique quasi-statique la variation d'entropie ne dépend que des valeurs initiales et finales de température (les variations intermédiaires étant considérées comme nulles puisque infiniment petites car infiniment lentes) elle tend alors à être une transformation réversible... Il vient alors:

$$\Delta S_e = -\Delta S_i \quad (33.227)$$

et donc nous avons pour le rendement thermique réversible à compression ou détente isotherme:

$$\eta = 1 - \frac{T_e}{T_i} \quad (33.228)$$

La transformation étant quasi-statique, la relation obtenue est bien évidemment l'efficacité maximale théorique pour un moteur fonctionnant entre ces deux températures. Elle n'est donc jamais atteinte dans un cycle réel. Nous disons alors que le "**rendement de Carnot**" est un maximum (l'entropie injectée étant dans un cas réel toujours plus grande que l'entropie extraite par construction!).

Dans le cas d'une machine à vapeur d'eau, l'efficacité théorique maximale serait donc (ne pas oublier que la température est en Kelvin et donc toujours positive!):

$$\eta = 1 - \frac{T_e}{T_i} = 1 - \frac{298}{373} \cong 0.2 = 20\% \quad (33.229)$$

Par exemple dans le cas d'une central nucléaire à eau sous pression, nous trouvons le fameux rendement souvent mentionné dans la presse:

$$\eta = 1 - \frac{T_e}{T_i} \cong 1 - \frac{(35 + 273)}{(285 + 273)} \cong 45\% \quad (33.230)$$

Le rendement réel de la conversion de la chaleur des fissions en électricité est de l'ordre de 33% seulement car le rendement réel est toujours inférieur au rendement théorique maximum.

Remarque: Nous pouvons noter en passant l'intérêt du condenseur. Si on laissait s'échapper à l'atmosphère la vapeur sortant des turbines, d'une part on perdrait une eau purifiée au lieu de la recycler, d'autre part on aurait une température de source froide plus élevée: 100 [°C] au lieu de 35 [°C] on perdrait alors environ 12% sur le rendement de Carnot.

Nous retrouvons enfin implicitement également l'un des énoncés historiques du second principe de la thermodynamique en envisageant le cas où $T_e = T_c$. Dans ce cas, l'efficacité est nulle et le moteur ne fournit donc aucun travail.

10. RELATIONS DE MAXWELL

Revenons dans un premier temps à ce que nous avons déjà rappelé un peu plus haut mais en nous restreignant à deux variables. C'est-à-dire à la différentielle totale exacte:

$$dz = Mdx + Ndy = \frac{\partial z}{\partial x} dx + \frac{\partial z}{\partial y} dy = \left(\frac{\partial z}{\partial x} \right)_y dx + \left(\frac{\partial z}{\partial y} \right)_x dy \quad (33.231)$$

Nous avons donc aussi:

$$\begin{aligned} dx &= \left(\frac{\partial x}{\partial y} \right)_z dy + \left(\frac{\partial x}{\partial z} \right)_y dz \\ dy &= \left(\frac{\partial y}{\partial x} \right)_z dx + \left(\frac{\partial y}{\partial z} \right)_x dz \\ dz &= \left(\frac{\partial z}{\partial x} \right)_y dx + \left(\frac{\partial z}{\partial y} \right)_x dy \end{aligned} \quad (33.232)$$

En insérant dy dans dx :

$$\begin{aligned} dx &= \left(\frac{\partial x}{\partial y} \right)_z \left(\left(\frac{\partial y}{\partial x} \right)_z dx + \left(\frac{\partial y}{\partial z} \right)_x dz \right) + \left(\frac{\partial x}{\partial z} \right)_y dz \\ &= \left(\frac{\partial x}{\partial y} \right)_z \left(\frac{\partial y}{\partial x} \right)_z dx + \left(\left(\frac{\partial x}{\partial y} \right)_z \left(\frac{\partial y}{\partial z} \right)_x + \left(\frac{\partial x}{\partial z} \right)_y \right) dz \end{aligned} \quad (33.233)$$

ou encore:

$$\left[\left(\frac{\partial x}{\partial y} \right)_z \left(\frac{\partial y}{\partial x} \right)_z - 1 \right] dx + \left(\left(\frac{\partial x}{\partial y} \right)_z \left(\frac{\partial y}{\partial z} \right)_x + \left(\frac{\partial x}{\partial z} \right)_y \right) dz = 0 \quad (33.234)$$

comme les termes entre parenthèses sont des fonctions et que dx et dz sont par contre arbitraires, la seule solution à cette relation est:

$$\left(\frac{\partial x}{\partial y}\right)_z \left(\frac{\partial y}{\partial x}\right)_z - 1 = 0 \quad \left(\frac{\partial x}{\partial y}\right)_z \left(\frac{\partial y}{\partial z}\right)_x + \left(\frac{\partial x}{\partial z}\right)_y = 0 \quad (33.235)$$

Multipliant la deuxième relation par $(\partial z / \partial x)_y$:

$$\left(\frac{\partial z}{\partial x}\right)_y \left(\frac{\partial x}{\partial y}\right)_z \left(\frac{\partial y}{\partial z}\right)_x + \left(\frac{\partial z}{\partial x}\right)_y \left(\frac{\partial x}{\partial z}\right)_y = 0 \quad (33.236)$$

Nous avons alors:

$$\left(\frac{\partial z}{\partial x}\right)_y \left(\frac{\partial x}{\partial y}\right)_z \left(\frac{\partial y}{\partial z}\right)_x + 1 = 0 \quad (33.237)$$

Venons-en maintenant aux faits. Rappelons l'équation fondamentale de Gibbs:

$$dU = TdS - PdV \quad (33.238)$$

relation très utile dans les fluides où la pression est constante et la variation de chaleur se fait par celle de l'entropie. Ainsi que la relation définissant l'enthalpie:

$$H = U + PV \quad (33.239)$$

dont nous allons modifier la différentielle:

$$dH = dU + VdP + PdV \quad (33.240)$$

et y injectant (premier principe):

$$dU = \delta Q - PdV \quad (33.241)$$

nous avons (remarquez que si la transformation est isobare la variation d'enthalpie représente alors uniquement la quantité de chaleur reçue par le système fermé):

$$dH = \delta Q + VdP \quad (33.242)$$

et en y injectant l'entropie (deuxième principe) nous obtenons:

$$dH = TdS + VdP \quad (33.243)$$

Ainsi, en ayant juste rajouté PV à l'énergie interne U nous avons bien une fonction d'état (H) où nous maîtrisons la pression et l'entropie comme variables indépendantes, alors que pour U seule, nous maîtrisons l'entropie et le volume en tant que variables indépendantes. Nous

Nous allons donc utiliser les deux relations suivantes qui vont nous être utiles:

$$\begin{aligned}dU &= TdS - PdV \\dH &= TdS + VdP\end{aligned}\quad (33.244)$$

Nous introduisons maintenant une nouvelle quantité que nous appelons "**énergie libre**" (celle qui est réellement disponible dans le système) et qui sera donnée par la "**relation de Gibbs-Helmholtz**":

$$F = U - TS \quad (33.245)$$

et donne simplement la différence entre l'énergie interne et l'énergie calorifique dissipée à cause de l'entropie à une température donnée. L'idée étant cette fois de contrôler le volume et la température comme variables indépendantes (voir juste un peu plus bas lorsque nous en prenons la différentielle), ce qui est bien évidemment une situation des plus importantes en chimie!

Nous introduisons également une autre nouvelle quantité que nous appelons "**enthalpie libre**" (celle qui est réellement disponible dans le système) et qui sera donnée identiquement par:

$$G = H - TS \quad (33.246)$$

qui est simplement la différence entre l'enthalpie et l'énergie calorifique dissipée à cause de l'entropie à une température donnée.

Nous avons donc pour l'énergie libre la forme différentielle:

$$dF = dU - SdT - TdS \quad (33.247)$$

en y injectant le premier principe et deuxième principe:

$$dF = (\delta Q - PdV) - SdT - TdS = (TdS - PdV) - SdT - TdS = -PdV - SdT \quad (33.248)$$

Ainsi, nous voyons que si la transformation est isotherme dF se réduit au travail reçu par le système fermé (puisque dT est alors nul).

De même pour l'enthalpie libre:

$$dG = dH - SdT - TdS \quad (33.249)$$

en y injectant:

$$dG = (TdS + VdP) - SdT - TdS = VdP - SdT \quad (33.250)$$

Par conséquent, dans le cas d'une transformation isotherme et isobare la variation d'enthalpie libre est nulle.

Nous avons donc quatre relations:

$$\begin{aligned}dU &= TdS - PdV \\dH &= TdS + VdP \\dF &= -PdV - SdT \\dG &= VdP - SdT\end{aligned}\quad (33.251)$$

appelées "équations de Gibbs".

Nous remarquons que ces équations sont toutes de la forme:

$$dz = Mdx + Ndy \quad (33.252)$$

Or, rappelons que selon le théorème de Schwarz (cf. chapitre de Calcul Différentiel Et Intégral) si dz est bien une différentielle totale exacte, nous avons alors:

$$\left(\frac{\partial M}{\partial y}\right)_x = \left(\frac{\partial N}{\partial x}\right)_y \quad (33.253)$$

Ce qui nous donne les quatre relations:

$$\begin{aligned} \left(\frac{\partial T}{\partial V}\right)_S &= -\left(\frac{\partial P}{\partial S}\right)_V \\ \left(\frac{\partial T}{\partial P}\right)_S &= \left(\frac{\partial V}{\partial S}\right)_P \\ \left(\frac{\partial P}{\partial T}\right)_V &= \left(\frac{\partial S}{\partial V}\right)_T \\ \left(\frac{\partial V}{\partial T}\right)_P &= -\left(\frac{\partial S}{\partial P}\right)_T \end{aligned} \quad (33.254)$$

Par ailleurs, par la définition même des dérivées partielles et des quatre relations:

$$\begin{aligned} dU &= TdS - PdV \\ dH &= TdS + VdP \\ dF &= -PdV - SdT \\ dG &= VdP - SdT \end{aligned} \quad (33.255)$$

nous avons:

$$\begin{aligned} \left(\frac{\partial U}{\partial S}\right)_V &= T & \left(\frac{\partial U}{\partial V}\right)_S &= -P \\ \left(\frac{\partial H}{\partial S}\right)_P &= T & \left(\frac{\partial H}{\partial P}\right)_S &= V \\ \left(\frac{\partial F}{\partial T}\right)_V &= -S & \left(\frac{\partial F}{\partial V}\right)_T &= -P \\ \left(\frac{\partial G}{\partial T}\right)_P &= -S & \left(\frac{\partial G}{\partial P}\right)_T &= V \end{aligned} \quad (33.256)$$

Toutes ces relations sont mises à profit pour calculer les variables thermodynamiques non directement mesurables à partir des données expérimentales.

Pour clore les définitions de ces variables thermodynamiques, indiquons le fait que d'avoir posé:

$$\begin{aligned}
 H &= U + PV \\
 F &= U - TS \\
 G &= H - TS
 \end{aligned} \quad (33.257)$$

ou autrement dit, le fait que pour tout couple X, Y de variables conjuguées et une fonction thermodynamique f nous posions:

$$g = f - XY \quad (33.258)$$

s'appelle une "**transformation de Legendre**" (voir pour les détails mathématiques de la transformation le chapitre de Mécanique Analytique). Ce type de transformation permet donc de modifier l'ensemble des variables indépendantes pour avoir un ensemble de variables mieux adaptées au problème considéré.

Maintenant voyons une relation qui nous sera utile en météorologie!:

Nous savons que la chaleur spécifique est donnée par définition à pression constante par:

$$C_p = \frac{\Delta Q}{\Delta T} \quad (33.259)$$

Or à pression constante, la variation de chaleur peut s'écrire par définition avec la variation d'enthalpie:

$$C_p = \left(\frac{\partial H}{\partial T} \right)_p \quad (33.260)$$

Maintenant rappelons que l'enthalpie s'écrit:

$$dH = TdS + VdP \quad (33.261)$$

comme dS est une différentielle exacte, nous pouvons l'écrire en fonction des paramètres de température et de pression seuls:

$$dS = \left(\frac{\partial S}{\partial T} \right)_p dT + \left(\frac{\partial S}{\partial P} \right)_T dP \quad (33.262)$$

Nous avons donc:

$$dH = T \left[\left(\frac{\partial S}{\partial T} \right)_p dT + \left(\frac{\partial S}{\partial P} \right)_T dP \right] + VdP = T \left(\frac{\partial S}{\partial T} \right)_p dT + \left[T \left(\frac{\partial S}{\partial P} \right)_T + V \right] dP \quad (33.263)$$

Comme par ailleurs dH est une différentielle exacte, nous pouvons aussi l'écrire en fonction des paramètres de température et de pression seuls:

$$dH = \left(\frac{\partial H}{\partial T} \right)_p dT + \left(\frac{\partial H}{\partial P} \right)_T dP = C_p dT + \left(\frac{\partial H}{\partial P} \right)_T dP \quad (33.264)$$

Nous avons alors les deux relations à identifier:

$$dH = T \left(\frac{\partial S}{\partial T} \right)_P dT + \left[T \left(\frac{\partial S}{\partial P} \right)_T + V \right] dP \quad (33.265)$$

$$dH = C_P dT + \left(\frac{\partial H}{\partial P} \right)_T dP$$

Il vient alors:

$$T \left(\frac{\partial S}{\partial T} \right)_P = C_P \quad T \left(\frac{\partial S}{\partial P} \right)_T + V = \left(\frac{\partial H}{\partial P} \right)_T \quad (33.266)$$

11. ÉQUATION DE CONTINUITÉ

Considérons de manière générale un système ouvert, limité par une frontière Ω quelconque (déformable ou non) et animé d'un mouvement quelconque (en déplacement ou immobile) par rapport à un référentiel considéré comme fixe.

Ce système, qui est représenté sur la figure ci-dessous, est susceptible de transférer de l'énergie (ou de la masse) entre lui-même et l'extérieur. Ce système peut être inertielle ou non.

Soit une grandeur extensive A (comme la masse ou la charge). La grandeur quantitative correspondante est a (elle peut exprimer par exemple l'isotropie ou l'anisotropie du système).

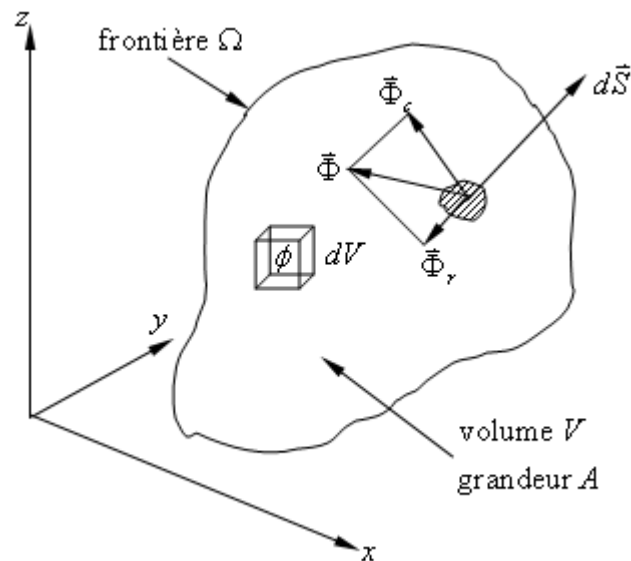


Figure: 33.4 - Système ouvert en mouvement ou non dans un référentiel et transférant de l'énergie

D'une façon générale, la valeur de A à l'intérieur du système est, à un instant quelconque:

$$A = \iiint_V \rho \cdot a \cdot dV \quad (33.267)$$

ρ étant la densité de la grandeur extensive A .

Le taux de variation spatial de A est donné par la dérivée dA/dt . Les causes de variations de A peuvent être liées à deux phénomènes différents: les flux et les sources ou puits.

En comptant positivement ce qui entre dans le système, le flux de A à travers la frontière Ω est donné par l'intégrale de surface:

$$\iint_{\Omega} \vec{\Phi} \cdot d\vec{S} \quad (33.268)$$

dans laquelle nous définissons:

- $\vec{\Phi}$ comme le vecteur flux surfacique (ou le vecteur densité de courant) total relatif à A
- $d\vec{S}$ comme l'élément de frontière, exprimé par un vecteur normal à la surface et dirigé vers l'extérieur

Remarquons que, contrairement à l'acceptation usuelle en physique, le concept de flux contient déjà la dérivation par rapport au temps. Par ailleurs, afin d'alléger le texte, l'expression vecteur flux surfacique est réduite au terme flux dans tout ce qui suit.

Ce flux peut être décomposé en plusieurs flux, selon la relation:

$$\vec{\Phi} = \vec{\Phi}_a - \vec{\Phi}_d + \vec{\Phi}_c = \vec{\Phi}_r + \vec{\Phi}_c \quad (33.269)$$

Le terme $\vec{\Phi}_a$ est un flux par déplacement absolu, caractérisant un flux lié à un écoulement fluide. Nous avons la relation:

$$\vec{\Phi}_a = \rho \cdot a \cdot \vec{v}_a \quad (33.270)$$

où $\vec{v}_a$ est la vitesse absolue d'une particule fluide par rapport au référentiel fixe.

Le terme $\vec{\Phi}_d$ est un flux par déplacement apparent, mis en jeu seulement lorsque la frontière Ω se déplace (par exemple si le volume V est en révolution). Nous avons la relation:

$$\vec{\Phi}_d = \rho \cdot a \cdot \vec{v}_d \quad (33.271)$$

où $\vec{v}_d$ est la vitesse apparente de déplacement (dans le sens déformation!) d'un point de la frontière Ω , par rapport au référentiel fixe.

Le terme $\vec{\Phi}_c$ est le flux total par conduction, caractérisant un flux lié à un phénomène de transfert de proche en proche, sans déplacement fluide (par exemple: conduction thermique, conduction électrique, travail mécanique).

Le terme $\vec{\Phi}_r$ est un flux par déplacement relatif, résultant à la fois du déplacement du fluide et de celui de la frontière Ω . Nous avons la relation:

$$\vec{\Phi}_r = \rho \cdot a \cdot \vec{v}_r \quad (33.272)$$

où $\vec{v}_r$ est la vitesse relative d'une particule fluide par rapport à un point bien défini de la frontière Ω . En vertu du principe de composition des vitesses (vitesse absolue est la somme de la vitesse relative et de sa vitesse apparente), nous avons:

$$\vec{v}_r = \vec{v}_a - \vec{v}_d$$

$$\vec{\Phi}_r = \vec{\Phi}_a - \vec{\Phi}_d \quad (33.273)$$

Lorsque la frontière Ω est traversée par un fluide, le débit-masse élémentaire (c'est la masse qui nous intéresse le plus souvent en physique donc ρ sera une densité massique) est:

$$\frac{dM}{dt} = \rho \cdot \vec{v}_r \circ d\vec{S} \quad (33.274)$$

Le flux de A correspondant est alors donné par:

$$\iint_{S_r} a \frac{dM}{dt} = \iint_{S_r} \vec{\Phi}_r \circ d\vec{S} \quad (33.275)$$

où S_r désigne la portion de frontière Ω traversée par le débit masse (ou fluide).

Si nous comptons positivement l'effet d'une source, le taux d'augmentation de A est donné par:

$$\iiint_V \phi dV \quad (33.276)$$

où ϕ est le flux volumique d'une source de A.

En tenant compte à la fois des flux et des sources, nous avons le taux de variation spatial de A:

$$\frac{dA}{dt} = \iint_S \vec{\Phi} \circ d\vec{S} + \iiint_V \phi dV \quad (33.277)$$

Le bilan spatial de A est finalement exprimé par la relation:

$$\frac{dA}{dt} = \iint_{S_r} a dM + \iint_S \vec{\Phi}_e \circ d\vec{S} + \iiint_V \phi dV \quad (33.278)$$

Dans le cas particulier d'un système en régime permanent (par exemple dans le cas d'un fluide qui s'écoule ou d'un solide qui est le siège de conduction thermique, de conduction électrique, de réaction nucléaire,...), toutes les grandeurs locales sont constantes en tout point du système. Si, de plus, nous choisissons une frontière Ω indéformable, il est possible de raisonner par rapport à un référentiel lié au système. Nous avons alors, en tout point fixe du système par rapport à ce référentiel:

$$\rho = c^{\text{te}}; a = c^{\text{te}}; V = c^{\text{te}} \quad (33.279)$$

Il en résulte pour l'ensemble du système:

$$A = c^{\text{te}} \Leftrightarrow \frac{dA}{dt} = 0 \quad (33.280)$$

Donc dans le cas particulier d'un système en régime permanent, avec une frontière indéformable Ω , liée au système, le taux de variation spatial de toute grandeur extensive scalaire est nul.

Le taux de variation uniquement spatial de A est:

$$\frac{dA}{dt} = \frac{d}{dt} \iiint_V \rho \cdot \alpha \cdot dV = \iiint_V \frac{\partial(\rho \cdot \alpha)}{\partial t} dV + \iiint_V \rho \cdot \alpha \cdot \frac{\partial}{\partial t}(dV) \quad (33.281)$$

La variation élémentaire du volume V est due au déplacement (au sens de la déformation!) de la frontière Ω , de sorte que

$$\partial(dV) = d(dV) = (d\vec{S} \circ \vec{v}_a) dt \quad (33.282)$$

Dès lors nous pouvons écrire que:

$$\iiint_V \rho \cdot \alpha \cdot \frac{\partial}{\partial t}(dV) = \iiint_V \rho \cdot \alpha \cdot \frac{\partial}{\partial t}(d\vec{S} \circ \vec{v}_a) dt = \iiint_V \vec{\Phi}_a \circ d\vec{S} \quad (33.283)$$

nous avons donc:

$$\frac{dA}{dt} = \iiint_V \frac{\partial(\rho \cdot \alpha)}{\partial t} dV + \iiint_V \vec{\Phi}_a \circ d\vec{S} \quad (33.284)$$

En prenant en compte les flux des sources et des puits, nous avons:

$$\iiint_V \frac{\partial(\rho \cdot \alpha)}{\partial t} dV = \iint_S \vec{\Phi} \circ d\vec{S} + \iiint_V \phi dV = \iint_S (\vec{\Phi}_a + \vec{\Phi}_e) \circ d\vec{S} + \iiint_V \phi \cdot dV \quad (33.285)$$

Le théorème de Gauss-Ostrogradsky (cf. chapitre de Calcul Vectoriel) va nous permettre d'écrire l'intégrale de surface en une intégrale de volume, et en groupant tous les termes sous le même signe intégrale, nous obtenons:

$$\iiint_V \left[\frac{\partial(\rho \cdot \alpha)}{\partial t} - \vec{\nabla} \circ (\vec{\Phi}_a + \vec{\Phi}_e) - \phi \right] dV = 0 \quad (33.286)$$

Comme les limites d'intégration (frontière Ω) sont arbitraires, l'expression entre crochets est identiquement nulle.

$$\frac{\partial(\rho \cdot \alpha)}{\partial t} - \vec{\nabla} \circ (\vec{\Phi}_a + \vec{\Phi}_e) - \phi = 0 \quad (33.287)$$

Considérons maintenant que la grandeur extensive scalaire soit la masse M , nous avons alors:

$$A = M = \iiint_V \rho \cdot dV \quad (33.288)$$

avec $\alpha = 1$.

Comme la masse n'est pas susceptible d'être transférée par un phénomène de conduction (dans un cas classique (non quantique)), nous avons $\vec{\Phi}_e$ qui est nul. Comme la masse est conservative, il n'y a ni source, ni puits de masse de sorte que ϕ est également nul.

Nous avons dès lors:

$$\frac{\partial(\rho \cdot a)}{\partial t} - \vec{\nabla} \circ (\vec{\Phi}_a + \vec{\Phi}_c) - \phi = \frac{\partial \rho}{\partial t} - \vec{\nabla} \circ (\rho \cdot \vec{v}_a) = 0 \quad (33.289)$$

La relation:

$$\boxed{\frac{\partial \rho}{\partial t} - \vec{\nabla} \circ (\rho \cdot \vec{v}_a) = 0} \quad (33.290)$$

est appelée "**équation de continuité**" ou encore "**équation de conservation** (de la masse)".

Le signe "-" est ici car nous avons défini le flux entrant comme étant positif. Il est possible que dans la littérature ainsi que sur ce site, vous trouviez un "+" à la place de ce signe.

Il y a une autre forme beaucoup plus fréquente sous laquelle nous trouvons l'équation de continuité. Le lecteur aura remarqué que le terme $\rho \cdot \vec{v}_a$ a les unités d'une densité de surface de courant massique ($\text{Kg} \cdot \text{m}^{-2} \text{s}^{-1}$) ce qui nous amène en analogie avec l'électronique (cf. chapitre d'Électrocinétique) à noter:

$$(\rho \cdot \vec{v}_a) = \vec{j} \quad (33.291)$$

Ce qui ramène l'équation de continuité à:

$$\boxed{\frac{\partial \rho}{\partial t} - \vec{\nabla} \circ \vec{j} = 0} \quad (33.292)$$

11.1. ÉQUATION DE LA CHALEUR

Appliquons maintenant ce résultat à la diffusion de la chaleur.

Comme pour l'équation de conservation de la masse, nous pouvons écrire pour la chaleur dans le cas d'absence de sources:

$$\boxed{\frac{\partial q}{\partial t} + \vec{\nabla} \circ \vec{\Phi}_q = 0} \quad (33.293)$$

où q est la quantité de chaleur par unité de volume (ne pas l'oublier sinon nous aurions pris un Q majuscule!) et $\vec{\Phi}_q$ le flux de chaleur dont la quantité entrante a été définie comme négative.

Une variation de température entraînant une variation de la quantité de chaleur est définie en première approximation par la loi physique suivante (cela découle de la définition de la chaleur spécifique massique aussi...):

$$dq = \rho \cdot c_p \cdot dT \quad (33.294)$$

où ρ est la densité de matière et c_p est la capacité calorifique massique. Ou de manière équivalente (puisque en thermodynamique, comme nous l'avons déjà précisé les minuscules sont rapportées à la masse):

$$dQ = \rho C dt$$

Le flux de chaleur étant trivialement induit par une différence spatiale de température, nous obtenons alors la "**loi de Fourier**" qui exprime le flux de chaleur proportionnellement au gradient spatial de température:

$$\vec{\Phi}_q = -\kappa \cdot \vec{\nabla} \cdot T \quad (33.295)$$

Le signe "-" étant simplement dû au fait que le flux de chaleur va du plus chaud au plus froid et κ est le "**coefficient de transport de la chaleur**" exprimant la "**conductivité thermique**" du matériau dépendant des propriétés atomiques de la matière (cf. chapitre de Mécanique Statistique).

En insérant les deux précédentes relations dans l'équation de conservation de la chaleur, nous avons:

$$\begin{aligned} \frac{\rho \cdot c_p \cdot dT}{dt} + \vec{\nabla} \circ (-\kappa \cdot \vec{\nabla} \cdot T) &= 0 \\ \frac{\partial T}{\partial t} - \frac{\kappa}{\rho \cdot c_p} \vec{\nabla} \circ (\vec{\nabla} \cdot T) &= 0 \end{aligned} \quad (33.296)$$

De façon plus esthétique et générale, nous la retrouvons sous la forme condensée de "**l'équation de diffusion de la chaleur**" ou appelée plus sobrement "**équation de la chaleur**":

$$\frac{\partial T}{\partial t} - D_T \frac{\partial^2 T}{\partial x^2} = 0 \quad (33.297)$$

où le coefficient de proportionnalité est appelé dans le cadre de la chaleur: "**coefficient de diffusion thermique**":

$$D_T = \frac{\kappa}{c_p \rho} \quad (33.298)$$

Il est possible de démontrer son origine microscopique comme nous l'avons fait dans le chapitre de Mécanique Statistique.

Il faut cependant toujours faire attention aux unités de κ suivant que nous travaillons avec la capacité calorifique massique c_p ou la capacité calorifique C au dénominateur!

Donc sous forme totalement explicite, nous avons en une dimension:

$$\frac{\partial T(x,t)}{\partial t} - \frac{\kappa}{\rho c_p} \frac{\partial^2 T(x,t)}{\partial x^2} = 0 \quad (33.299)$$

Remarques:

R1. Nous retrouverons cette équation dans le chapitre de Méthodes Numériques pour introduire le lecteur au concept de résolution d'équations différentielles par la méthode des éléments finis avec un exemple de résolution avec MS Excel.

R2. C'est en étudiant cette équation que Fourier a introduit les séries et la transformée qui portent son nom, et qui sont devenues si importantes dans l'étude des phénomènes de propagation/diffusion.

R3. L'équation de diffusion se retrouve dans de nombreux domaines (thermodynamique, fluides, finance,...) et il existe une littérature considérable sur les différentes solutions de cette équation différentielle du second ordre. Dans le chapitre d'Analyse Fonctionnelle nous montrerons comment résoudre cette équation avec la transformée de Fourier et la transformée de Laplace.

Insistons sur le fait que toutes les relations du type:

$$\frac{\partial u}{\partial t} - D \frac{\partial^2 u}{\partial x^2} = 0 \quad (33.300)$$

sont appelées "**équations de diffusion**" du paramètre physique D . Nous allons tout de suite voir comment la résoudre en prenant comme exemple l'équation de diffusion de la chaleur et cela nous permettra aussi de comprendre pourquoi Fourier a introduit la fameuse transformée qui porte son nom. Mais rappelons au lecteur que nous la retrouvons dans de multiples contextes (cf. chapitre de Mécanique Statistique).

Remarque: Cette équation (du moins sa forme et donc l'étude de sa résolution!) se retrouve dans des domaines inattendus comme dans la diffraction en physique ondulatoire, dans l'équation de Schrödinger en physique quantique, en finance dans l'équation de Black & Scholes, en électrocinétique dans le domaine des résistances, dans l'étude de la propagation des champs électromagnétiques dans la matière, dans l'étude des réactions en chimie, en neutronique nucléaire, etc.

Réolvons donc la forme générale de l'équation de diffusion::

$$\boxed{\frac{\partial u}{\partial t} = D \frac{\partial^2 u}{\partial x^2}} \quad (33.301)$$

Pour résoudre cette équation différentielle du second ordre, nous allons utiliser la méthode de séparation des variables (cf. chapitre de Calcul Différentiel Et Intégral) plutôt qu'en attaquant directement par les transformées de Fourier (qui nécessite la démonstration préalable du théorème de convergence dominée qui me fait horreur).

Nous supposons dès lors en utilisant la méthode de séparation des variables que:

$$u(x, t) = X(x)T(t) \quad (33.302)$$

où $T(t)$ par le principe du second principe de la thermodynamique doit diminuer lorsque t croît.

Nous avons:

$$\begin{aligned}\frac{\partial u}{\partial t} &= \frac{\partial X(x)T(t)}{\partial t} = \frac{\partial X(x)}{\partial t} T(t) + X(x) \frac{\partial T(t)}{\partial t} = X \frac{\partial T}{\partial t} \\ \frac{\partial^2 u}{\partial x^2} &= \frac{\partial^2 X(x)T(t)}{\partial x^2} = \frac{\partial^2 X(x)}{\partial x^2} T(t) + X(x) \frac{\partial^2 T(t)}{\partial x^2} = T \frac{\partial^2 X}{\partial x^2}\end{aligned}\quad (33.303)$$

d'où l'équation de diffusion:

$$\frac{\partial u}{\partial t} = D \frac{\partial^2 u}{\partial x^2} \Rightarrow X \frac{\partial T}{\partial t} = DT \frac{\partial^2 X}{\partial x^2} \quad (33.304)$$

ce qui remanié et condensé s'écrit aussi:

$$\frac{1}{D} \frac{T'}{T} = \frac{X''}{X} \quad (33.305)$$

Ce qui peut s'écrire en posant que chacun des expressions à gauche et à droite de l'égalité sont assimilables à des fonctions:

$$G(t) = F(x) \quad (33.306)$$

donc pour que l'égalité soit vraie pour tout t et x les fonctions G et F doivent être constantes. Donc nous avons le droit d'écrire:

$$\frac{1}{D} \frac{T'}{T} = \frac{X''}{X} = c^{ts} = -\lambda^2 \quad (33.307)$$

Le fait d'écrire la constante négative et au carré est une simple anticipation du résultat historiquement déjà connu... Mais pour comprendre pourquoi la constante est obligatoirement négative, il suffit de penser que que T' est obligatoirement négatif (la température d'un système isolé ne va pas augmenter tout seul mais naturellement diminuer selon le principe de l'entropie) et comme D et T sont positifs alors in extenso...

Ce qui nous donne le système de deux équations différentielles indépendantes du deuxième ordre:

$$\begin{aligned}X'' + \lambda^2 X &= 0 \\ T' + \lambda^2 DT &= 0\end{aligned}\quad (33.308)$$

que les américains écrivent sous sous la forme suivante:

$$\begin{aligned}X_{xx} + \lambda^2 X &= 0 \\ T_t + \lambda^2 DT &= 0\end{aligned}\quad (33.309)$$

Nous résolvons la deuxième équation différentielle:

$$\frac{1}{T(t)} \frac{\partial T(t)}{\partial t} = -\lambda^2 D \Rightarrow \ln(T(t)) = -\lambda^2 Dt + c^{ts} \quad (33.310)$$

Donc:

$$T(t) = e^{-\lambda^2 Dt + c^{te}} = e^{c^{te}} e^{-\lambda^2 Dt} = A(\lambda) e^{-\lambda^2 Dt} \quad (33.311)$$

Pour la première équation différentielle:

$$\frac{\partial X(x)^2}{\partial x^2} + \lambda^2 X(x) = 0 \quad (33.312)$$

Nous avons alors le polynôme caractéristique:

$$r^2 + \lambda^2 = 0 \quad (33.313)$$

Soient les racines:

$$r_{1,2} = \pm \sqrt{-\lambda^2} = \pm i\lambda \quad (33.314)$$

Donc comme le discriminant est négatif (cf. chapitre de Calcul Différentiel Et Intégral):

$$X(x) = A'(\lambda) e^{i(\lambda x + c^{te})} + B'(\lambda) e^{-i(\lambda x + c^{te})} \quad (33.315)$$

Soit:

$$\begin{aligned} u(x, t) &= T(t) X(x) = A(\lambda) e^{-\lambda^2 Dt} \left(A'(\lambda) e^{i(\lambda x + c^{te})} + B'(\lambda) e^{-i(\lambda x + c^{te})} \right) \\ &= \left(C_1(\lambda) e^{i(\lambda x + c^{te})} + C_2(\lambda) e^{-i(\lambda x + c^{te})} \right) e^{-\lambda^2 Dt} \end{aligned} \quad (33.316)$$

Si nous mettons l'exponentielle des constantes en exposant dans les facteurs tel que:

$$\begin{aligned} C_1(\lambda) e^{ic^{te}} &= C'_1(\lambda) \\ C_2(\lambda) e^{-ic^{te}} &= C'_2(\lambda) \end{aligned} \quad (33.317)$$

cela simplifie l'expression:

$$u(x, t) = \left(C'_1(\lambda) e^{i\lambda x} + C'_2(\lambda) e^{-i\lambda x} \right) e^{-\lambda^2 Dt} \quad (33.318)$$

dont les coefficients seront déterminés au cas par cas par les conditions initiales dont les cas scolaires les plus connus et faisant l'objet d'une littérature abondante sont:

- Conditions de Dirichlet:

$$x = (0, L) \quad u(0, t) = 0 = u(L, t) \quad (33.319)$$

- Conditions de Neumann:

$$x = (0, L) \quad \frac{du}{dx}(0, t) = 0 = \frac{du}{dx}(L, t) \quad (33.320)$$

Puisque pour chaque valeur de λ possible nous obtenons une solution, il apparaît donc qu'en faisant la somme de toutes ces solutions, nous obtenons la solution générale. Nous avons donc:

$$u(x, t) = \sum_{\lambda} \left(C_1'(\lambda) e^{i\lambda x} + C_2'(\lambda) e^{-i\lambda x} \right) e^{-\lambda^2 Dt} \quad (33.321)$$

à cause de la présence au carré du λ dans l'exponentielle temporelle en facteur de la parenthèse, si nous sommions sur tous les λ de $-\infty, +\infty$ alors la somme de tous les termes de la parenthèse s'annule pour tout couple λ positif et négatif de même valeur. Il nous faut alors sommer que sur la demi-droite infinie (soit à gauche soit à droite cela n'a pas d'importance) si nous ne voulons pas que le résultat soit nul. Si nous choisissons arbitrairement la demi-droite infinie positive, nous avons alors:

$$u(x, t) = \sum_{\lambda=0}^{+\infty} \left(C_1'(\lambda) e^{i\lambda x} + C_2'(\lambda) e^{-i\lambda x} \right) e^{-\lambda^2 Dt} \quad (33.322)$$

Rien ne nous interdit maintenant de sortir un facteur λ des constantes $C_1'(\lambda), C_2'(\lambda)$, le seul effet que cela aura soit de changer leur normalisation donnée par les conditions initiales.. Nous avons alors:

$$\begin{aligned} u(x, t) &= \sum_{\lambda=0}^{+\infty} \left(\lambda C_1''(\lambda) e^{i\lambda x} + \lambda C_2''(\lambda) e^{-i\lambda x} \right) e^{-\lambda^2 Dt} \\ &= \sum_{\lambda=0}^{+\infty} \left(C_1''(\lambda) e^{i\lambda x} + C_2''(\lambda) e^{-i\lambda x} \right) e^{-\lambda^2 Dt} \lambda \end{aligned}$$

Comme λ est un paramètre réel continu (et accessoirement positif), en le faisant tendre vers zéro la somme peut être changée une intégrale (en faisant cela à la physicienne) et nous avons alors:

$$u(x, t) = \int_0^{+\infty} C_1''(\lambda) e^{-\lambda^2 Dt + i\lambda x} d\lambda + \int_0^{+\infty} C_2''(\lambda) e^{-\lambda^2 Dt - i\lambda x} d\lambda \quad (33.323)$$

Et en faisant un changement de variable pour la deuxième intégrale, nous avons:

$$\begin{aligned} u(x, t) &= \int_0^{+\infty} C_1''(\lambda) e^{-\lambda^2 Dt + i(\lambda x + c^{te})} d\lambda - \int_0^{+\infty} C_2''(-\lambda) e^{-\lambda^2 Dt - i(-\lambda x + c^{te})} d\lambda \\ &= \int_0^{+\infty} C_1''(\lambda) e^{-\lambda^2 Dt + i(\lambda x + c^{te})} d\lambda + \int_{-\infty}^0 C_2''(-\lambda) e^{-\lambda^2 Dt + i(\lambda x - c^{te})} d\lambda \end{aligned} \quad (33.324)$$

En regroupant les constantes, nous avons:

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F(\lambda) e^{-\lambda^2 Dt + i\lambda x} d\lambda \quad (33.325)$$

avec:

$$\begin{aligned} F(\lambda) &= \sqrt{2\pi} C_1''(\lambda) \quad \text{pour } \lambda > 0 \\ F(\lambda) &= \sqrt{2\pi} C_2''(\lambda) \quad \text{pour } \lambda < 0 \end{aligned} \quad (33.326)$$

Nous avons maintenant une chose intéressante! Au temps t nul, l'intégrale précédente s'écrit:

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F(\lambda) e^{i\lambda x} d\lambda \quad (33.327)$$

et en changeant la notation pour le paramètre λ ainsi que la notation pour x , nous avons:

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F(\omega) e^{i\omega t} d\omega \quad (33.328)$$

Et alors direz-vous? Eh bien comme nous l'avons vu lors de notre étude des transformées de Fourier (cf. chapitre Suites Et Séries) il s'agit d'une transformée de Fourier inverse qui est une somme infinie de fonctions trigonométriques réelles! Ce qui signifie que toute fonction décrivant la distribution de température au temps t nul pour une barre de longueur infinie et dont les extrémités au même moment tendent vers zéro (donc implicitement il s'agit d'une fonction de période infinie) telle que:

$$\begin{aligned} u(x,0) &= f(x) \\ f(\pm\infty) &\rightarrow 0 \end{aligned} \quad (33.329)$$

satisfait l'équation de la Chaleur en tant que série trigonométrique infinie. Revenons à notre notation initiale:

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F(\lambda) e^{-\lambda^2 Dt + i\lambda x} d\lambda \quad (33.330)$$

Pour rappel du chapitre de Suites Et Séries (avec la notation adaptée ici présente) nous avons en appliquant la transformer de Fourier inverse:

$$u(x,0) = \mathcal{F}^{-1}(F(\lambda))(x) = f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F(\lambda) e^{-i\lambda x} dx \quad (33.331)$$

Donc $F(\lambda)$ est la transformée de Fourier de la fonction recherchée telle que:

$$F(\lambda) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x) e^{-i\lambda x} dx \quad (33.332)$$

Ce qui donne alors en faisant en injectant la transformée de Fourier avec un changement de variable qui s'impose:

$$\begin{aligned} u(x, t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F(\lambda) e^{-\lambda^2 Dt + i\lambda x} d\lambda = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(v) e^{-i\lambda v} dv \right) e^{-\lambda^2 Dt + i\lambda x} d\lambda \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} f(v) e^{-i\lambda v} dv \right) e^{-\lambda^2 Dt + i\lambda x} d\lambda \end{aligned} \quad (33.333)$$

Toujours en gardant en tête que si nous posons t comme étant nu ci-dessous, nous retrouvons alors bien:

$$u(x, 0) = f(x) \quad (33.334)$$

Puisque dans la dernière égalité le terme $e^{-\lambda^2 Dt + i\lambda x}$ est indépendant de la variable v , nous pouvons le mettre dans la deuxième intégrale tel que:

$$u(x, t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} f(v) e^{-i\lambda v} dv \right) e^{-\lambda^2 Dt + i\lambda x} d\lambda = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(v) e^{-i\lambda v} e^{-\lambda^2 Dt + i\lambda x} dv d\lambda \quad (33.335)$$

Nous y reconnaissons alors une double intégrale dont nous pouvons changer l'ordre d'intégration par l'application du théorème de Fubini (cf. chapitre de Calcul Différentiel Et Intégral):

$$\begin{aligned} u(x, t) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(v) e^{-i\lambda v} e^{-\lambda^2 Dt + i\lambda x} d\lambda dv = \frac{1}{2\pi} \int_{-\infty}^{+\infty} f(v) \int_{-\infty}^{+\infty} e^{-i\lambda v} e^{-\lambda^2 Dt + i\lambda x} d\lambda dv \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} f(v) \int_{-\infty}^{+\infty} e^{-\lambda^2 Dt - i\lambda(v-x)} d\lambda dv \end{aligned} \quad (33.336)$$

et rappelons que dans la chapitre de Suites Et Séries nous avons démontré que la transformée de Fourier d'une Gaussienne est une autre Gaussienne tel que:

$$\mathcal{F}(f(t)) = \mathcal{F}(e^{-at^2}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-ax^2} e^{-i\omega t} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-ax^2 - i\omega t} dt = \frac{1}{\sqrt{2a}} e^{-\frac{\omega^2}{4a}} \quad (33.337)$$

Il vient alors dans le cas qui nous intéresse en faisant une correspondance terme par terme:

$$\begin{aligned}
 u(x,t) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} f(v) \int_{-\infty}^{+\infty} e^{-Dt\lambda^2 - i(v-x)\lambda} d\lambda dv = \frac{1}{2\pi} \int_{-\infty}^{+\infty} f(v) \left(\sqrt{2\pi} \frac{1}{\sqrt{2Dt}} e^{-\frac{(v-x)^2}{4Dt}} \right) dv \\
 &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} f(v) \frac{\sqrt{\pi}}{\sqrt{Dt}} e^{-\frac{(v-x)^2}{4Dt}} dv
 \end{aligned} \tag{33.338}$$

Soit:

$$u(x,t) = \frac{1}{2\sqrt{\pi Dt}} \int_{-\infty}^{+\infty} f(v) e^{-\frac{(v-x)^2}{4Dt}} dv \tag{33.339}$$

Donc toute fonction $f(v)$ injectée dans la relation ci-dessus satisfera alors aux conditions initiales que nous nous sommes imposés:

$$\begin{aligned}
 u(x,0) &= f(x) \\
 f(\pm\infty) &\rightarrow 0
 \end{aligned} \tag{33.340}$$

Pour information, de nombreux physiciens nomment dans la solution précédente, l'expression:

$$K(t, x, v) = \frac{1}{\sqrt{4\pi t}} e^{-\frac{(v-x)^2}{4t}} \tag{33.341}$$

le "**noyau de chaleur**". Donc le résultat final n'est en réalité qu'une convolution de la $f(v)$ avec ce "noyau".

Dans la pratique cette intégrale n'est pas calculable analytique puisqu'il s'agit d'une convolution avec une Gaussienne (cf. chapitre de Statistiques). Raisons pour laquelle on résout presque toujours l'équation de la chaleur aux conditions initiales susmentionnées avec des méthodes numériques (cf. chapitre de Méthodes Numériques).

Mais résolvons un cas simplistes mais réel:

Exemples:

E1. Lorsque deux extrémités d'un système de taille L sont maintenues à deux températures différentes T_1 et T_2 , la solution de l'équation de la chaleur est stationnaire (indépendante du temps). Nous avons alors:

$$\frac{\partial^2 T(x)}{\partial x^2} = 0 \tag{33.342}$$

La solution à cette équation différentielle est très simple et ne nécessite pas l'utilisation du résultat précédent (c'est du calcul intégral de base avec conditions initiales connues):

$$T(x) = T_1 + (T_2 - T_1) \frac{x}{L} \quad (33.343)$$

C'est une situation que nous retrouvons dans la vie de tous les jours...

E2. Considérons le cas où nous mettons en contact deux barreaux de longueur infinie (puisque avec la transformée de Fourier il faut toujours qu'une des variables ballade l'ensemble des réels) et de température opposée en signe (sur l'échelle des Celsius donc) tel que:

$$u(x, 0) = \begin{cases} T_0 & x > 0 \\ -T_0 & x < 0 \end{cases} \quad (33.344)$$

Nous avons alors:

$$\begin{aligned} u(x, t) &= \frac{1}{2\sqrt{\pi Dt}} \int_{-\infty}^{+\infty} f(v) e^{-\frac{(v-x)^2}{4Dt}} dv = \frac{1}{2\sqrt{\pi Dt}} \left(\int_{-\infty}^0 -T_0 e^{-\frac{(v-x)^2}{4Dt}} dv + \int_0^{+\infty} T_0 e^{-\frac{(v-x)^2}{4Dt}} dv \right) \\ &= \frac{T_0}{2\sqrt{\pi Dt}} \left(\int_0^{+\infty} e^{-\frac{(v-x)^2}{4Dt}} dv - \int_{-\infty}^0 e^{-\frac{(v-x)^2}{4Dt}} dv \right) \end{aligned} \quad (33.345)$$

et faisons le changement de variable:

$$\frac{v-x}{\sqrt{4Dt}} = \frac{v-x}{2\sqrt{Dt}} = \tau \quad (33.346)$$

Soit:

$$dv = -2\sqrt{Dt} d\tau \quad (33.347)$$

Nous avons alors:

$$\begin{aligned} u(x, t) &= \frac{T_0}{2\sqrt{\pi Dt}} \left(\int_{-x/2\sqrt{Dt}}^{+\infty} -2\sqrt{Dt} e^{-\tau^2} d\tau - \int_{-\infty}^{-x/2\sqrt{Dt}} -2\sqrt{Dt} e^{-\tau^2} d\tau \right) \\ &= \frac{T_0}{\sqrt{\pi}} \left(\int_{-x/2\sqrt{Dt}}^{+\infty} -e^{-\tau^2} d\tau - \int_{-\infty}^{-x/2\sqrt{Dt}} -e^{-\tau^2} d\tau \right) = \frac{T_0}{\sqrt{\pi}} \left(\int_{-x/2\sqrt{Dt}}^{+\infty} -e^{-\tau^2} d\tau + \int_{-\infty}^{-x/2\sqrt{Dt}} e^{-\tau^2} d\tau \right) \\ &= \frac{T_0}{\sqrt{\pi}} \left(\int_{-\infty}^{-x/2\sqrt{Dt}} e^{-\tau^2} d\tau - \int_{-x/2\sqrt{Dt}}^{+\infty} e^{-\tau^2} d\tau \right) \end{aligned} \quad (33.348)$$

Comme la fonction intégrée est une fonction paire (cf. chapitre d'Analyse Fonctionnelle) nous avons:

$$\int_{-\infty}^{-x/2\sqrt{Dt}} e^{-\tau^2} d\tau = \int_{x/2\sqrt{Dt}}^{+\infty} e^{-\tau^2} d\tau \quad (33.349)$$

Soit:

$$\begin{aligned} u(x,t) &= \frac{T_0}{\sqrt{\pi}} \left(\int_{x/2\sqrt{Dt}}^{+\infty} e^{-\tau^2} d\tau - \int_{-x/2\sqrt{Dt}}^{+\infty} e^{-\tau^2} d\tau \right) = \frac{T_0}{\sqrt{\pi}} \left(\int_{-x/2\sqrt{Dt}}^{x/2\sqrt{Dt}} e^{-\tau^2} d\tau \right) \\ &= \frac{2T_0}{\sqrt{\pi}} \left(\int_0^{x/2\sqrt{Dt}} e^{-\tau^2} d\tau \right) \end{aligned} \quad (33.350)$$

Ce qu'il est d'usage de noter:

$$u(x,t) = T_0 \operatorname{erf} \left(\frac{x}{2\sqrt{Dt}} \right) \quad (33.351)$$

où "erf" est appelée la "**fonction d'erreur de Gauss**" et n'est pas calculable de façon exacte analytiquement (il faut passer par des développements en série limitée). On la retrouve cependant dans des tableurs comme Microsoft Excel sous le nom ERF().

12. RAYONNEMENT THERMIQUE

L'étude du corps noir est à la base de la célèbre théorie de la physique quantique ondulatoire, un des piliers de la physique moderne. En effet, certains résultats expérimentaux ne pouvaient pas être expliqués sans l'introduction de la fameuse constante de Planck, de l'utilisation de la quantification de l'énergie (le quanta), de l'acceptation du modèle atomique et de la théorie statistique de Boltzmann (implicitement le deuxième principe de la thermodynamique dont Planck était un spécialiste).

Avant de commencer nous allons faire exceptionnellement (pour ce site) faire un crochet historique qui s'avère très utile pour comprendre pourquoi l'étude du rayonnement thermique est si importante en physique.

Le fait que tous les objets chauffés émettent une lumière d'une même couleur à la même température est une curiosité qui était connue depuis de nombreux siècles dans certains métiers, longtemps avant que Gustav Krichhoff commence ses investigations scientifiques et théoriques sur la nature de cette étrange corrélation. Afin de simplifier son analyse, il élaborait le concept d'un objet parfaitement absorbant et parfaitement émissif. Ainsi, en tant que parfait émetteur, il serait tout sauf noir si sa température était assez élevée pour qu'il puisse rayonner dans des longueurs d'onde de la partie visible du spectre. Donc relativement au retour d'expérience, le rayonnement de cet objet aurait une intensité et une étendue spectrale indépendante du type de matériau dont il est constitué. L'objectif était alors de mesurer la répartition spectrale de l'énergie rayonnée pour chaque longueur d'onde et pour chaque palier de température et en dériver une équation dépendante donc que de deux variables (longueur d'onde et température) permettant de reproduire cette répartition et donc ce pour tout matériau mises dans les conditions d'un corps parfaitement émissif et réceptif.

Définition: Un "corps noir" (ou "récepteur intégral") est défini comme un corps ayant un "coefficient d'absorption énergétique" α et un "coefficient d'émissivité" ε égaux à l'unité (cf. chapitre d'Optique Géométrique)

Le premier principe de la thermodynamique établit une équivalence entre le travail et chaleur comme modes de transfert d'énergie entre un système et son environnement (et en fait le bilan au niveau de l'énergie interne). Nous nous intéressons ici à la chaleur, que nous pouvons définir comme "l'énergie qu'un corps communique à un autre à cause de leur différence de température".

La chaleur se communique d'un endroit à un autre de trois manières différentes comme nous en avons déjà fait mention plus haut:

1. Par conduction: c'est un transfert de chaleur dans un ensemble de points matériels en contact qui se fait sans mouvements macroscopiques, sous l'influence d'un gradient de température. La conduction est donc le résultat de collisions moléculaires. Nous l'observons principalement dans les solides: dans les métaux, elle fait intervenir les électrons libres qui les rendent bons conducteurs de chaleur. En revanche, dans les isolants, la conduction se fait mal. De là la forte correspondance entre les propriétés thermiques et électriques des solides.
2. Par convection: la convection implique le transport de la chaleur par une partie d'un fluide qui se mélange avec une autre particule. Elle prend sa source dans un transport macroscopique de matière et ne concerne donc pas les solides.
3. Par rayonnement: la conduction et la convection supposent la présence de matière. Le rayonnement, lui, permet un transfert d'énergie qui peut s'effectuer à travers le vide. Il s'agit ici de rayonnement électromagnétique. Soulignons que le rayonnement n'est pas un mode de transfert de chaleur mais d'énergie, celle-ci pouvant se transformer en chaleur au contact d'un corps.

Le rayonnement thermique émis par un corps porté à une certaine température résulte d'une conversion de l'énergie interne du corps en rayonnement. Inversement, l'absorption est la transformation de l'énergie incidente en énergie interne.

Lorsqu'une surface est soumise à un rayonnement absorbé, nous effectuons le bilan d'énergie selon la loi de Kirchhoff vue en photométrie:

$$\alpha + \rho + \tau = 1 \quad (33.352)$$

où rappelons-le quand même, α est la fraction du rayonnement absorbée, ρ est la partie réfléchie (diffusée) et τ la partie transmise (qui traverse la surface). Ce bilan résulte du principe de la conservation de l'énergie.

Nous allons maintenant nous pencher sur les mécanismes d'absorption et d'émission et établir un lien entre chaleur et énergie rayonnante avant de nous intéresser directement au corps noir:

12.1. LOI DE STEFAN-BOLTZMANN

Nous avons défini lors de notre étude de la photométrie (cf. chapitre d'Optique Géométrique) le concept d'émittance (énergie irradiée par un corps non ponctuel par unité de surface) pour l'ensemble du spectre.

Ce que nous avons omis de préciser cependant, c'est que pour qu'un corps rayonne (outre le fait qu'il puisse être lui-même éclairé par un autre corps) il faut qu'il soit chauffé (que l'on fournisse une énergie d'excitation aux constituants du corps en question - sous-entendu aux électrons).

Donc nous devrions pouvoir établir une relation entre la température d'un corps et son émittance.

En 1879, le physicien autrichien Stefan a pu établir expérimentalement que l'émittance totale $M(T)$ du corps noir (ou "**exitance énergétique**" du corps noir) à une température T augmentait proportionnellement à la quatrième puissance de la température telle que:

$$M(T) = \sigma \cdot T^4 \quad (33.353)$$

où $M(T)$ est l'intégration sur toutes les longueurs d'onde (ou les fréquences... peu importe) de $M(\lambda, T)$:

$$M(T) = \int_0^{\infty} M(\lambda, T) d\lambda \quad (33.354)$$

avec $M(\lambda, T)$ donné par la loi de Planck que nous déterminerons plus tard.

Rappelons également que (ceci sera démontré lors de notre démonstration de la loi de Planck):

$$\sigma = 5.67 \cdot 10^{-8} [W \cdot m^{-2} \cdot K^{-4}] \quad (33.355)$$

est la "**constante de Stefan**".

En 1884, Boltzmann a démontré indirectement la loi de Stefan en se basant sur l'étude du corps noir à l'équilibre thermique (où nous considérons que les bords de la paroi du corps noir définissent les terminaisons des ondes électromagnétiques) à partir de la théorie de l'électromagnétisme et d'un raisonnement thermodynamique.

Dans un premier temps, Boltzmann a déterminé quelle était la pression de radiation du rayonnement dans une telle enceinte (ou dans un tel corps).

Voici les développements qui l'ont mené à déterminer la pression de radiation $P(T)$ à la température d'équilibre thermodynamique T pour la densité interne d'énergie $\rho(T)$ correspondante:

Rappelons l'expression de la "relation d'Einstein" que nous avons démontrée lors de notre étude de la relativité restreinte:

$$p = \frac{h\nu}{c} \quad (33.356)$$

Considérons maintenant une enceinte de volume V dont les parois sont réfléchissantes pour les photons (cas du corps noir). Nous étudions la variation de la quantité de mouvement avant et après la collision sur une surface infiniment petite ds (ce qui permet de considérer les trajectoires avant et après le choc comme rectilignes et symétriques par rapport à l'axe orienté OX perpendiculaire à la surface du corps noir coïncidant avec la surface ds).

Ainsi, nous avons avant collision pour la quantité de mouvement:

$$\vec{p} = p_x \vec{u}_x + p_y \vec{u}_y \quad (33.357)$$

et après collision:

$$\vec{p}' = -p'_x \vec{u}_x + p'_y \vec{u}_y \quad (33.358)$$

Si la collision est élastique (ce qui est confortant relativement au photon...):

$$|p'_x| = |p_x| \text{ et } |p'_y| = |p_y| \quad (33.359)$$

Nous avons alors:

$$\vec{p}' = -p_x \vec{u}_x + p_y \vec{u}_y \quad (33.360)$$

La variation de la quantité de mouvement est alors:

$$d\vec{p} = \vec{p}' - \vec{p} = (-p_x \vec{u}_x + p_y \vec{u}_y) - (p_x \vec{u}_x + p_y \vec{u}_y) = -2p_x \vec{u}_x \quad (33.361)$$

Comme:

$$p_x = p \cos \alpha \quad (33.362)$$

nous avons alors:

$$d\vec{p} = -2p \vec{u}_x \cos \alpha \quad (33.363)$$

En ne considérant que la norme de l'expression et qu'il s'agit d'un unique photon:

$$dp = 2 \frac{h\nu}{c} \cos \alpha \quad (33.364)$$

Remarque: Nous supposons qu'après son rebond, le photon conserve sa fréquence (ce qui nous amène à supposer que le corps noir comporte des ondes stationnaires à l'équilibre thermodynamique).

Jusqu'à présent, nous avons raisonné sur un unique photon, mais l'enceinte contient en réalité un gaz de photons. L'énergie interne volumique u du rayonnement contient une densité volumique n de photons de fréquence identique ν . Dès lors, la quantité n de photons par unité de volume dans l'enceinte est:

$$n = \frac{u}{h\nu} [m^{-3}] \quad (33.365)$$

Remarque: Nous précisons les unités, car nous avons remarqué que la suite posait parfois quelques problèmes de compréhension.

Nous considérons que pendant un intervalle de temps dt , le nombre de photons pouvant potentiellement frapper la surface ds sous un angle d'indice α est contenu dans un cylindre de génératrice $c dt$ dont l'axe est incliné nécessairement d'un angle α et ayant comme surface de base ds . Le volume de ce cylindre est de par la projection de la surface de base:

$$dV = c dt \cos \alpha ds \quad (33.366)$$

Le nombre de photons dn_c pouvant potentiellement heurter la paroi ds par unité de temps est:

$$dn_c = n dV = \frac{u}{h\nu} dV = \frac{uc}{h\nu} \cos \alpha ds dt \quad (33.367)$$

Dans cette dernière expression, nous avons supposé que tous les photons de dV avaient une quantité de mouvement dans la direction sous-tendue par α . En réalité, les photons arrivants réellement sur ds sont contenus dans un angle solide $d\Omega$ entre deux cônes de demi-angle au sommet α et $\alpha + d\alpha$ (pour des raisons de géométrie de l'expérience du corps noir qui était, sauf erreur, sphérique et par ailleurs cette symétrie sphérique facilite les calculs...).

La relation entre $d\Omega$ et α est comme nous l'avons vue dans le chapitre de Trigonométrie:

$$d\Omega = d(2\pi(1 - \cos \alpha)) = 2\pi \sin \alpha d\alpha \quad (33.368)$$

Sachant que dans le volume entier (rappel), l'angle solide vaut:

$$\Omega = \frac{4\pi r^2}{r^2} = 4\pi \quad (33.369)$$

Le nombre dn compris dans l'angle solide élémentaire $d\alpha$ qui parvient sur la surface ds sous un angle d'incidence compris entre α et $d\alpha$ est alors:

$$dn = \frac{d\Omega}{\Omega} dn_c = \frac{2\pi \sin \alpha d\alpha}{4\pi} \frac{uc \cos \alpha}{h\nu} ds dt = \frac{uc \sin \alpha \cos \alpha d\alpha ds dt}{2h\nu} \quad (33.370)$$

Soit maintenant la définition de la pression P :

$$P = \frac{dF}{ds} = \frac{F dn}{ds} = \left(\frac{dp}{dt} \right) \frac{dn}{ds} \quad (33.371)$$

en substituant ce qu'il convient:

$$P = \frac{\frac{dp}{dt} dn}{ds} = \frac{\frac{2 \frac{h\nu}{c} \cos \alpha}{dt} \frac{uc \sin \alpha \cos \alpha}{2h\nu} d\alpha ds}{ds} \quad (33.372)$$

Ce qui donne après simplification:

$$P = u \sin \alpha \cos^2 \alpha d\alpha = u \cos^2 \alpha d(-\cos \alpha) \quad (33.373)$$

La pression totale de radiation dans ce cas particulier étant donnée par:

$$P = n\hbar\nu \int_0^{\pi/2} \cos^2 \alpha d(-\cos \alpha) = -n\hbar\nu \frac{\cos^3 \alpha}{3} \Big|_0^{\pi/2} = \frac{n\hbar\nu}{3} \quad (33.374)$$

Ce qui est équivalent à écrire (à l'équilibre thermodynamique pour une température donnée):

$$P(T) = \frac{\rho(T)}{3} \quad (33.375)$$

relation trrrrrès importante que nous utiliserons dans le chapitre de Cosmologie lors de notre étude du modèle d'Univers de Friedmann dominé par la matière.

L'énergie totale est la densité d'énergie multipliée par le volume considéré:

$$E = \rho(T)V \quad (33.376)$$

Supposons que ce volume puisse varier. Le travail de la pression de radiation lors d'une dilatation dV du volume est:

$$dW = -P(T)dV = -\frac{\rho(T)}{3}dV \quad (33.377)$$

La variation d'énergie interne du système en vertu du premier principe de la thermodynamique est:

$$dE = dQ + dW \quad (33.378)$$

Or d'après $E = \rho(T)V$, nous avons:

$$dE = d\rho(T)V + \rho(T)dV = d\rho(T)V \frac{dT}{dT} + \rho(T)dV \quad (33.379)$$

d'où:

$$dQ = dE - dW = \rho(T)dV + \frac{d\rho(T)VdT}{dT} - \left(-\frac{\rho(T)}{3}dV \right) = \frac{4}{3}\rho(T)dV + \frac{d\rho(T)VdT}{dT} \quad (33.380)$$

et selon le deuxième principe de la thermodynamique (ne pas confondre la notation avec la surface...):

$$dS = \frac{dQ}{T} \quad (33.381)$$

Nous avons:

$$dS = \frac{4}{3} \frac{\rho(T)dV}{T} + \frac{d\rho(T)VdT}{TdT} \quad (33.382)$$

Autrement écrit, cela correspond à:

$$dS = \frac{4}{3} \frac{\rho(T)}{T} dV + \frac{d\rho(T)V}{T dT} dT = \frac{\partial S}{\partial V} dV + \frac{\partial S}{\partial T} dT \quad (33.383)$$

Comme dS est une différentielle totale, nous avons démontré dans le chapitre de Calcul Différentiel et Intégral que S satisfait alors le théorème de Schwarz:

$$\frac{\partial^2 S}{\partial T \partial V} = \frac{\partial^2 S}{\partial V \partial T} \quad (33.384)$$

nous avons dans ce cas:

$$\frac{d}{dT} \left(\frac{4}{3} \frac{\rho(T)}{T} \right) = \frac{d}{dV} \left[\frac{d\rho(T)V}{T dT} \right] = \frac{d\rho(T)}{T dT} \quad (33.385)$$

Ce qui nous amène à écrire:

$$\frac{d\rho(T)}{T dT} = \frac{4}{3} \frac{d}{dT} \left(\frac{\rho(T)}{T} \right) \quad (33.386)$$

En calculant la dérivée du membre de droite:

$$\frac{d\rho(T)}{T dT} = \frac{4}{3} \frac{\frac{d\rho(T)}{dT} T - \rho(T) \frac{dT}{dT}}{T^2} \quad (33.387)$$

En simplifiant:

$$\frac{d\rho(T)}{dT} = \frac{4}{3} \frac{\rho(T)}{dT} - \frac{4}{3} \frac{\rho(T)}{T} \quad (33.388)$$

ce qui s'écrit encore:

$$\frac{d\rho(T)}{dT} \left(1 - \frac{4}{3} \right) = -\frac{4}{3} \frac{\rho(T)}{T} \quad (33.389)$$

Soit:

$$-\frac{1}{3} \frac{d\rho(T)}{dT} = -\frac{4}{3} \frac{\rho(T)}{T} \quad (33.390)$$

qui devient l'équation:

$$\frac{d\rho(T)}{\rho(T)} = 4 \frac{dT}{T} \quad (33.391)$$

Ce qui donne après intégration:

$$\ln(\rho(T)) = 4\ln(T) + c^{te} \quad (33.392)$$

Finalement:

$$\rho(T) = c^{te} \cdot T^4 = \sigma_{s.B.} \cdot T^4 \quad (33.393)$$

avec:

$$\sigma_{s.B.} = 7,56 \cdot 10^{-16} [J \cdot m^{-3} \cdot K^{-4}] \quad (33.394)$$

étant la constante de Stefan-Boltzmann dont la valeur avait été donnée à l'époque dans un premier temps expérimentalement.

Nous voyons ci-dessus la correspondance qu'il y a entre la relation que nous avons posée au début et celle que nous venons d'obtenir:

$$\rho(T) = \sigma_{s.B.} T^4 \text{ et } M(T) = \sigma T^4 \quad (33.395)$$

Comme nous n'avons pas encore, à ce point, démontré la loi de Planck, nous pouvons faire un raisonnement osé mais que nous justifierons par la suite avec démonstration à l'appui.

Remarque: Les deux dernières relations nous donnent une information fondamentale comme quoi tous les corps qui ne sont pas à zéro kelvin (au zéro absolu) rayonnent!

$M(T)$ et $\rho(T)$ sont différenciées au niveau des unités par les dimensions d'une vitesse. Or, intuitivement et grossièrement (...), la vitesse qui peut tout de suite nous apparaître comme triviale dans ce cas d'étude est la vitesse de la lumière c . Ainsi, nous remarquons que:

$$\frac{\sigma}{\sigma_{s.B.}} = \frac{c}{4} \quad (33.396)$$

Ce qui nous donne:

$$M(T) = \frac{c}{4} \rho(T) \quad (33.397)$$

Curieux n'est-ce pas... mais nous le démontrerons plus loin car notre philosophie sur ce site est de ne jamais (ou le moins possible) laisser place à l'intuition.

Remarque: Lorsque nous étudierons la loi de Planck, $\rho(T)$ sera notée $R(T)$ afin de ne pas confondre une densité d'énergie avec la radiance (car la notation peut malheureusement porter à confusion).

Considérons maintenant une chambre ou cavité isolée (comme une fournaise) en équilibre thermique à une certaine température T . Cette cavité sera sûrement remplie de rayonnements électromagnétiques de différentes longueurs d'onde. Supposons qu'il existe une fonction de distribution $M(T)$ dépendant uniquement de la température.

Logiquement, la quantité totale d'énergie électromagnétique, à toutes les longueurs d'onde, absorbée par les murs de la cavité doit être égale à celle émise par les murs autrement le corps formant la cavité verrait sa température changer. Kirchhoff raisonna que si le corps formant la cavité est fait de différents matériaux (se comportant donc de façons différentes avec la température), l'équilibre entre radiation émise et radiation absorbée doit s'appliquer alors pour chaque longueur d'onde (ou domaine de longueur d'onde).

Nous voyons ainsi que $M(T)$ est une fonction universelle, la même pour toutes les cavités sans égard à leur composition, leur géométrie ou la couleur de leurs parois. Kirchhoff ne donna pas cette fonction, mais il fit remarquer qu'un corps parfaitement absorbant, c'est-à-dire un corps pour lequel $\alpha = 1$ apparaîtra (façon de dire...) noir.

Il vient alors que le rayonnement emmagasiné en équilibre dans une cavité isolée en équilibre thermodynamique (comme le sont les étoiles) est à tous égards le même que celui émis par un corps parfaitement noir à la même température.

Évidemment, si la cavité est fermée, nous ne pouvons pas mesurer le courant d'énergie qui s'en échappe. Mais pratiquons un tout petit trou dans cette cavité (suffisamment petit pour ne pas perturber l'équilibre du rayonnement électromagnétique à l'intérieur), alors l'énergie électromagnétique s'échappant de ce petit trou est la même que celle émise par un corps parfaitement noir.

Cependant, aucun objet n'est réellement un corps noir. Le noir de charbon a un coefficient d'absorption très près de 1 mais seulement pour certaines fréquences (incluant, bien sûr, le visible). Son coefficient d'absorption est beaucoup plus petit dans l'infrarouge lointain. Tout de même, la plupart des objets s'en approchent dans certaines gammes de fréquences. Le corps humain, par exemple, est presque un corps noir dans l'infrarouge (d'où les lunettes de nuit militaires...). Pour traiter les différents corps, appelés "**corps gris**", nous introduisons un facteur appelé "**émissivité totale**", ε , qui relie l'émittance émise par le corps à celle émise par un corps noir parfait pour lequel $\varepsilon = 1$. Nous avons donc:

$$M(T) = \varepsilon \sigma T^4 \quad (33.398)$$

Remarque: La relation de Stefan-Boltzmann nous donne la puissance émise par un corps par unité de surface en l'exprimant de façon proportionnelle à la quatrième puissance de la température. Cet exposant nous donne la raison pour laquelle il devient de plus en plus difficile d'augmenter la température d'un corps en le chauffant, celui-ci perdant de plus en plus rapidement l'énergie que nous fournissons pour son échauffement.

12.2. LOI DE PLANCK

Pour établir la loi de Planck, nous n'allons pas faire le développement original de Max Planck bien que passionnant du point de vue intuitif il n'était qu'une succession de bricolages (admis par Max Planck lui-même). Nous allons introduire cette loi avec une approche moderne et la plus simple possible en matière de formalisme mathématique.

Pour cela, nous considérons maintenant le corps noir comme un système isolé à l'équilibre thermique, dans lequel le rayonnement est à l'état stationnaire et réfléchi totalement par les parois. Les photons peuvent être dès lors considérés comme des particules n'interagissant pas entre elles dans un puits de potentiel à parois rectilignes.

Ainsi, identiquement à ce que nous avons vu dans le chapitre de Physique Quantique Ondulatoire, la résolution du problème est celle d'un puits de potentiel à parois rectilignes pour lequel nous avons obtenu pour fonction d'onde:

$$\Psi_x = \Psi_{x,0} \sin(\varpi x) = \Psi_{x,0} \sin \left(x \cdot \sqrt{\frac{8\pi^2 m}{h^2} (E_{tot} - E_{pot})} \right) \quad (33.399)$$

fonction à laquelle il convient d'appliquer les conditions aux limites.

Les conditions que nous avons imposées lors de notre étude de ce cas en physique quantique ondulatoire étaient trop restrictives (c'est la raison pour laquelle elles sont appelées "conditions aux limites strictes"). Effectivement, les atomes de la paroi absorbent et émettent le rayonnement quelle que soit la manière dont le rayonnement est incident. Mais l'équilibre impose au moins que les conditions aux limites soient périodiques de par la définition même de l'équilibre. C'est la raison pour laquelle nous imposons ce que nous appelons les "**conditions aux limites périodiques**":

- pour $x = 0$ et $x = L_x$, nous avons: $\Psi(0) = \Psi(L_x) = c^{\pm}$
- la fonction d'onde Ψ doit présenter un nombre entier de demi-longueurs d'onde sur la longueur L_x
- dans le corps noir, $U = 0$ donc $E_{pot} = 0$
- si aux extrémités ($x = 0$ et $x = L_x$) nous avons $\Psi = c^{\pm}$ l'argument du sinus a la même valeur $n2\pi$ (à un facteur multiplicatif réel près) en 0 et en L_x .

Donc nous devons avoir:

$$L_x \cdot \sqrt{\frac{8\pi^2 m}{h^2} (E_{tot} - E_{pot})} = n_x 2\pi \quad (33.400)$$

et comme $E_{pot} = 0$, après quelques simplifications élémentaires, nous avons:

$$\frac{8\pi^2 m}{h^2} E_{tot} = \frac{n_x^2}{L_x^2} 4\pi^2 \Rightarrow E_{tot} = \frac{n_x^2}{L_x^2} 4\pi^2 \frac{h^2}{8\pi^2 m} = \frac{n_x^2}{L_x^2} \frac{h^2}{2m} \quad (33.401)$$

où E_{tot} est l'énergie totale du niveau quantique n selon x (cf. chapitre de Physique Quantique Ondulatoire).

L'énergie totale de la particule présente donc une suite discrète de valeurs, les seules permises. La valeur de L est quant à elle déterminée à l'aide du modèle de Bohr ou de Sommerfeld en fonction des cas.

Puisque les fonctions d'onde correspondantes dans le puits sont $U = 0 \equiv E_{pot} = 0$, nous avons donc:

$$\begin{aligned} \Psi_x &= \Psi_{x,0} \sin \left(x \cdot \sqrt{\frac{8\pi^2 m}{h^2} E_{tot}} \right) = \Psi_{x,0} \sin \left(x \cdot \sqrt{\frac{8\pi^2 m}{h^2} \frac{n_x^2}{L_x^2} \frac{h^2}{2m}} \right) = \Psi_{x,0} \sin \left(x \cdot \sqrt{\frac{4\pi^2 n_x^2}{L_x^2}} \right) \\ &= \Psi_{x,0} \sin \left(\frac{2\pi n_x}{L_x} \cdot x \right) = \Psi_{x,0} \sin(k_x \cdot x) \end{aligned} \quad (33.402)$$

Ainsi, l'énergie totale peut s'écrire:

$$E_{nx} = \frac{n_x^2}{L_x^2} \frac{2\pi^2 \hbar^2}{m} = \frac{\hbar^2}{2m} \frac{4\pi^2 n_x^2}{L_x^2} = \frac{\hbar^2}{2m} \left(\frac{2\pi n_x}{L_x} \right)^2 = \frac{\hbar^2}{2m} k_x^2 \quad (33.403)$$

Ainsi, étant donné que la fonction d'onde est une probabilité conditionnelle, nous avons sous forme de phaseur:

$$\Psi = \Psi_x \Psi_y \Psi_z = c^{\frac{1}{2}} e^{ik_x x} e^{ik_y y} e^{ik_z z} = c^{\frac{1}{2}} e^{i\vec{K} \cdot \vec{r}} \quad (33.404)$$

et les énergies discrètes associées sont alors:

$$E_n = \frac{\hbar^2}{2m} (k_x^2 + k_y^2 + k_z^2) = \frac{\hbar^2}{2m} \|\vec{K}\|^2 \quad (33.405)$$

Le vecteur $\vec{K}$ étant donc défini par:

$$\vec{K} = \left(n_x \frac{2\pi}{L_x}, n_y \frac{2\pi}{L_y}, n_z \frac{2\pi}{L_z} \right) \quad (33.406)$$

Remarque: Nous constatons facilement que les écarts d'énergie entre niveaux consécutifs sont d'autant plus faibles que les dimensions du corps noir (assimilé à une boîte) L_x, L_y, L_z sont plus grandes; pour des dimensions macroscopiques, ces écarts sont alors totalement inappréciables. Ce constat nous permettra un peu plus loin de faire une petite approximation.

Explication: Pour un électron ($m \cong 10^{-30}$ [Kg]) enfermé dans une boîte cubique de côté $L = 1$ [m], l'écart entre deux niveaux consécutifs est:

$$E_{n_x+1, n_y, n_z} - E_{n_x, n_y, n_z} = (2n_x + 1) \frac{2\pi^2 \hbar^2}{m L_x^2} \quad (33.407)$$

donc environ 10^{-35} [J] ...

Les vecteurs $\vec{K}$ qui nous intéressent (puisque'ils représentent respectivement chacun un micro-état possible), plongés dans l'espace des phases des nombres d'ondes, ont leur extrémité située en l'un des noeuds d'un réseau tridimensionnel constitué de mailles élémentaires dont les arêtes sont parallèles aux axes et qui mesurent respectivement $2\pi/L_x, 2\pi/L_y, 2\pi/L_z$. Nous voulons évaluer le nombre de vecteurs pour lesquels cette extrémité tombe dans l'intervalle entre les deux sphères centrées à l'origine et de rayons de norme K et $K + dK$. Le volume de la coquille sphérique comprise entre les deux sphères est donc trivialement donné par:

$$dV = 4\pi K^2 dK \quad (33.408)$$

Le nombre de mailles élémentaires (de micro-états) incluses dans cette région de l'espace des $\vec{K}$ est, à peu de chose près, égal au nombre de fois que son volume contient celui de la maille élémentaire, qui vaut:

$$\frac{2\pi}{L_x} \cdot \frac{2\pi}{L_y} \cdot \frac{2\pi}{L_z} = \frac{8\pi^3}{V} \quad (33.409)$$

Nous obtenons ainsi le nombre de micro-états dans le volume dV (donc la densité de micro-états):

$$\frac{4\pi K^2 dK}{\frac{8\pi^3}{V}} = \frac{VK^2 dK}{2\pi^2} \quad (33.410)$$

Or, il ne faut pas oublier les relations suivantes (cf. chapitres de Mécanique Ondulatoire, Physique Quantique Corpusculaire et Relativité Restreinte):

$$K = \frac{2\pi}{\lambda} = \frac{\omega}{c} \quad (33.411)$$

Donc comme:

$$K^2 = \frac{\omega^2}{c^2}, dK = \frac{1}{c} d\omega \quad (33.412)$$

et (rappel):

$$\omega = 2\pi\nu \quad (33.413)$$

il vient alors:

$$\frac{VK^2 dK}{2\pi^2} = \frac{V\omega^2}{2\pi^2 c^3} d\omega = \frac{V4\pi^2 \nu^2}{2\pi^2 c^3} 2\pi d\nu = \frac{V4\pi^3 \nu^2}{\pi^2 c^3} d\nu = \frac{V4\pi \nu^2}{c^3} d\nu \quad (33.414)$$

Mais quand les physiciens avaient développé ce modèle théorique (il paraît que c'est Bose qui aurait été le premier à le faire), ils avaient remarqué que le résultat final ne correspondait pas à l'expérience à facteur 2 près. Dès lors ils multiplièrent à l'époque empiriquement la densité de micro-états par un facteur 2 tel que:

$$2 \cdot \frac{V4\pi \nu^2}{c^3} d\nu = \frac{V8\pi \nu^2}{c^3} d\nu \quad (33.415)$$

Ce facteur 2 s'expliquerait aujourd'hui (je n'en ai jamais vu la démonstration personnellement...) par le spin 1 du photon. Ce qui autorise a priori trois valeurs pour sa projection: -1, 0, 1. La valeur 0 serait interdite par la théorie quantique des champs (si quelqu'un possède la démonstration, je suis preneur!).

Dans un corps noir à l'équilibre thermodynamique, les photons (qui par ailleurs sont des bosons) forment un gaz dont les constituants n'interagissent pas entre eux chimiquement. Ce type de situation est typiquement décrit par la distribution de Bose-Einstein que nous avons démontrée dans le chapitre de Mécanique Statistique. Ainsi, puisque $\mu = 0$, nous avons dans le cas d'un spectre discret d'états d'énergie (c'est donc ici qu'apparaît le concept de quantification que Planck avait introduit à l'origine en bricolant):

$$n_i = \frac{g_i}{\exp\left(\frac{E_i}{kT}\right) - 1} = \frac{g_i(\nu)}{\exp\left(\frac{h\nu_i}{kT}\right) - 1} \quad (33.416)$$

et dans un cas que nous considérons comme continu:

$$dn = \frac{g(\nu)}{\exp\left(\frac{h\nu}{kT}\right) - 1} = \frac{\frac{V8\pi\nu^2}{c^3} d\nu}{\exp\left(\frac{h\nu}{kT}\right) - 1} \quad (33.417)$$

Avant de continuer il est important que le lecteur se rende compte que l'analyse du rayonnement du corps noir vient donc suite aux deux relations précédentes à imposer:

- Premièrement une quantification des niveaux d'énergie par paquets $h\nu$ alors qu'à l'époque de la découverte du phénomène la nature était considérée comme continue dans tous ses phénomènes.
- Deuxièmement à considérer que la théorie de l'entropie de Boltzmann et la distribution statistique de Bose-Einstein qui s'ensuit est un pilier de l'étude des systèmes physiques.

Dans le corps noir, nous avons pour énergie interne:

$$U = \sum_i E_i n_i = \int_0^\infty h\nu \frac{\frac{V8\pi\nu^2}{c^3} d\nu}{\exp\left(\frac{h\nu}{kT}\right) - 1} = V \int_0^\infty \frac{8\pi h}{c^3} \frac{\nu^3}{\exp\left(\frac{h\nu}{kT}\right) - 1} d\nu \quad (33.418)$$

$$\Rightarrow \frac{U}{V} = \int_0^\infty \rho(\nu, T) d\nu \quad [J \cdot m^{-3}]$$

La radiation d'un corps noir (sa "**brillance monochromatique**" comme disent certains...) est donc donnée par la "**loi de Planck**":

$$R(\nu, T) = \frac{8\pi h \nu^3}{c^3} \frac{1}{\exp\left(\frac{h\nu}{kT}\right) - 1} = \frac{2\hbar \omega^3}{\pi c^3} \frac{1}{\exp\left(\frac{\hbar \omega}{kT}\right) - 1} \quad [J \cdot s \cdot m^{-3}] \quad (33.419)$$

La première égalité est souvent écrite sous la forme suivante qui permet une interprétation du résultat:

$$R(\nu, T) = \frac{8\pi\nu^2}{c^3} h\nu \frac{1}{\exp\left(\frac{h\nu}{kT}\right) - 1} \quad (33.420)$$

où $\frac{8\pi\nu^2}{c^3}$ correspond à la densité volumique de mode de rayonnement du corps noir, $h\nu$ à l'énergie

moyenne par quantum d'énergie et $\frac{1}{\exp\left(\frac{h\nu}{kT}\right) - 1}$ représentent la population des modes.

Enfin, voyons une autre manière courante d'écrire la loi de Planck. Puisque:

$$R(\nu, T) d\nu = R(\lambda, T) d\lambda \text{ et } \nu = \frac{c}{\lambda} \Rightarrow \frac{d\nu}{d\lambda} = \frac{-c}{\lambda^2} \quad (33.421)$$

donc:

$$R(\nu, T) \frac{d\nu}{d\lambda} = \frac{-c}{\lambda^2} R(\nu, T) = R(\lambda, T) \quad (33.422)$$

Or, comme $R(\lambda, T) \geq 0, R(\nu, T) \geq 0$, il convient de prendre la valeur absolue telle que:

$$R(\nu, T) \frac{d\nu}{d\lambda} = \left| -\frac{c}{\lambda^2} \right| R(\nu, T) = \frac{c}{\lambda^2} R(\nu, T) \quad (33.423)$$

Enfin, nous obtenons encore une autre forme de la loi de Planck qui exprime la densité de flux d'énergie pour une longueur d'onde précise donnée par:

$$R(\lambda, T) = \frac{8\pi ch}{\lambda^5} \frac{1}{\exp\left(\frac{hc}{\lambda kT}\right) - 1} \quad (33.424)$$

Remarque: Planck a donc proposé cette loi par une successions de bricolages théoriques et d'analogies en 1900 sans connaître la distribution statistique de Bose-Einstein ce qui est remarquable expérimentalement parlant!

Si $h\nu \ll kT$ (donc dans le domaine des grandes longueurs d'ondes), le développement de Taylor (cf. chapitre Suites Et Séries) de e^x pour x petit donne donc:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots \cong 1 + x \quad (33.425)$$

Ce qui nous donne:

$$\exp\left(\frac{hc}{\lambda kT}\right) \cong 1 + \frac{hc}{\lambda kT} \quad (33.426)$$

et la loi de Planck devient donc la "**loi de Rayleigh-Jeans**" (qui avait été découverte avant la loi de Planck):

$$R(\lambda, T) = \frac{8\pi h}{\lambda^3} \frac{1}{\frac{hc}{\lambda kT}} = \frac{8\pi h}{\lambda^3} \frac{\lambda kT}{hc} = \frac{8\pi kT}{\lambda^2 c} \quad (33.427)$$

Que nous retrouvons aussi parfois dans la littérature spécialisée sous la forme:

$$R(\nu, T) = \frac{8\pi}{\left(\frac{c}{\nu}\right)^2} \frac{kT}{c} = \frac{8\pi \nu^2}{c^3} kT \quad (33.428)$$

A l'inverse, $h\nu \gg kT$ nous avons:

$$\exp\left(\frac{hc}{\lambda kT}\right) - 1 \cong \exp\left(\frac{hc}{\lambda kT}\right) \quad (33.429)$$

et la loi de Planck devient donc:

$$R(\lambda, T) = \frac{8\pi h}{\lambda^3} \frac{1}{\exp\left(\frac{hc}{\lambda kT}\right)} = \frac{8\pi h}{\lambda^3} \exp\left(-\frac{hc}{\lambda kT}\right) \quad (33.430)$$

qui n'est rien d'autre que la "**première loi de Wien**" (qui avait été découverte avant la loi de Rayleigh-Jeans). Cette loi décrit effectivement la présence d'un maximum de rayonnement, mais, contrairement à la loi de Rayleigh-Jeans, elle fournit des valeurs fausses pour les grandes longueurs d'onde (donc les petites fréquences). En outre, elle implique que l'intensité de rayonnement tende vers zéro avec l'augmentation de la température, ce que contredit également l'expérience.

Voici un schéma présentant les différences entre les trois lois:

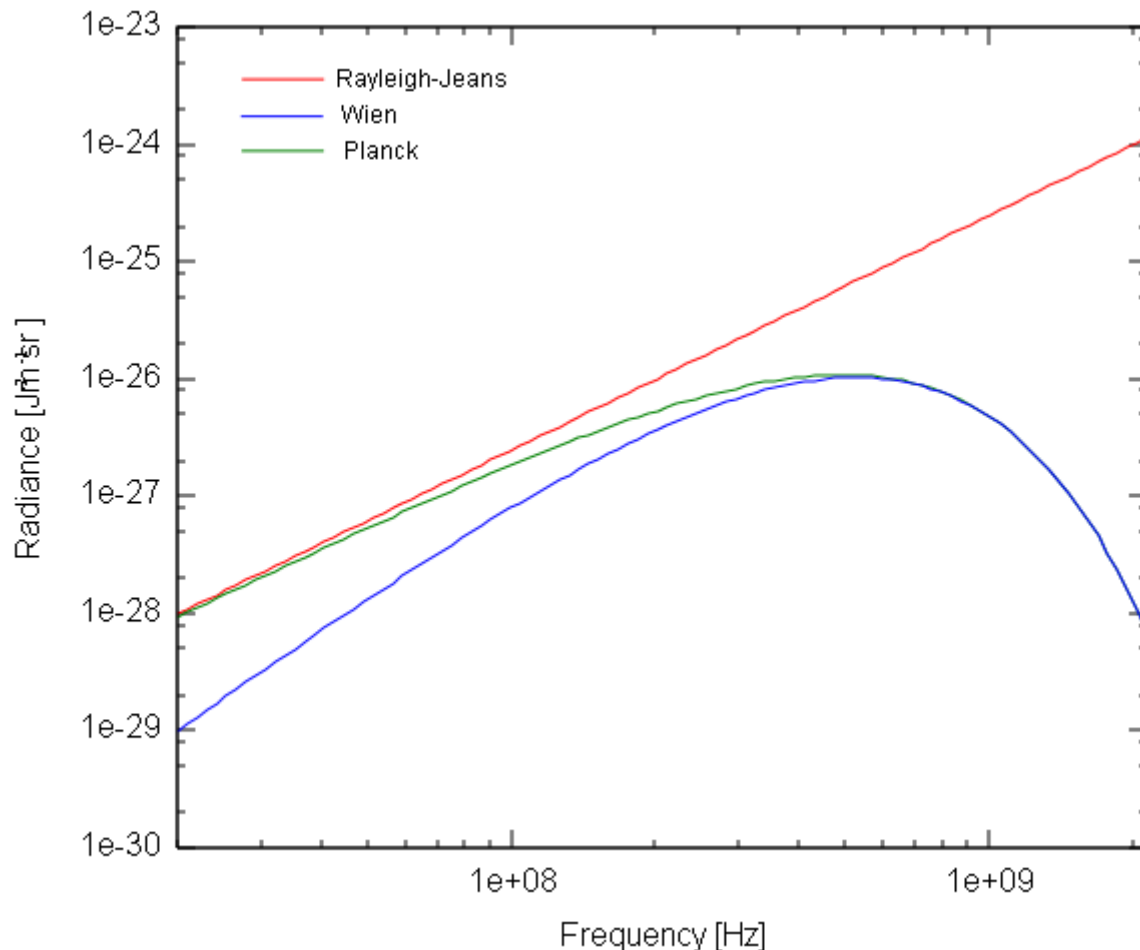


Figure: 33.5 - Tracé des trois lois dans une échelle log-log pour comparaison (source: Wikipédia)

Nous voyons ci-dessus que la loi de Jeans était bonne aux basses fréquences alors que celle de Wien l'était aux hautes. Historiquement il a fallu de nombreux essais et pas mal d'années aux physiciens pour trouver la modèle adéquat qui amena finalement Max Planck à considérer la quantification des états d'énergie (ce dont il ne renda compte lui même que bien après avoir élaboré son modèle).

Nous pouvons également redémontrer la loi de Stefan (nous l'avons déjà fait plus haut mais avec une autre démarche) mais cette fois-ci en explicitant la provenance de la constante de Stefan-Boltzmann $\sigma_{s.B}$.

Rappelons d'abord que le flux énergétique (cf. chapitre d'Optique Géométrique) est entre autres donné par:

$$d\Phi = L(\theta) dS d\Omega \cos \theta \quad (33.431)$$

Comme la luminance dépend de la fréquence et donc de la température du corps émetteur, nous pouvons ajouter:

$$d\Phi(\nu, T) d\nu = L(\theta, \nu, T) dS d\Omega \cos \theta d\nu \quad (33.432)$$

L'énergie rayonnée à travers une surface élémentaire $dS \cos \theta$ donnée est donc dès lors:

$$d\Phi(\nu, T) d\nu dt = L(\theta, \nu, T) dS d\Omega \cos \theta d\nu dt \quad (33.433)$$

Si le volume d'émission est considéré comme un volume élémentaire assimilé à un cylindre de hauteur cdt et de sommet ayant pour surface $dS \cos \theta$ (cf. chapitre d'Optique Géométrique) la densité d'énergie par unité de fréquence et par stéradian est alors donnée par:

$$dU(\nu, T) = \frac{L(\theta, \nu, T) dS d\Omega \cos \theta d\nu dt}{cdt dS \cos \theta} = \frac{1}{c} L(\theta, \nu, T) d\Omega d\nu \quad (33.434)$$

Compte tenu de l'isotropie du corps noir à l'équilibre, nous avons en intégrant sur l'ensemble de l'angle solide la densité d'énergie par unité de fréquence seule:

$$dU(\nu, T) = \frac{4\pi}{c} L(\theta, \nu, T) d\nu \quad (33.435)$$

L'analyse dimensionnelle nous donne:

$$\frac{4\pi}{c} L(\theta, \nu, T) = R(\nu, T) \Rightarrow L(\theta, \nu, T) = \frac{c}{4\pi} R(\nu, T) \quad (33.436)$$

Enfin, il est utile de considérer la puissance totale émise par unité de surface (donc l'émittance):

$$dM(\nu, T) = L(\theta) d\Omega \cos \theta \quad (33.437)$$

Si nous intégrons sur la demi-surface d'une sphère (par rapport au point de surface de l'émetteur):

$$M(\nu, T) = \int_0^{2\pi} \int_0^{\pi/2} L(\theta) d\Omega \cos \theta = \int_0^{\pi/2} L(\theta) \cos \theta (2\pi \sin \theta d\theta) \quad (33.438)$$

Effectivement pour une sphère (cf. chapitre de Trigonométrie):

$$d\Omega = 2\pi \sin \theta d\theta \quad (33.439)$$

Comme la luminance est indépendante de θ (isotropie du rayonnement du corps noir), l'intégration est élémentaire, et nous trouvons:

$$M(\nu, T) = \pi L = \pi \frac{c}{4\pi} R(\nu, T) = \frac{c}{4} R(\nu, T) \quad [J \cdot m^{-2}] \quad (33.440)$$

L'émittance totale est alors donnée par:

$$M(T) = \frac{c}{4} \int_0^\infty R(\nu, T) d\nu = \frac{c}{4} \int_0^\infty 8\pi \frac{h\nu^3}{c^3} \frac{1}{\exp\left(\frac{h\nu}{kT}\right) - 1} d\nu = 2\pi \frac{h}{c^2} \int_0^\infty \frac{\nu^3}{\exp\left(\frac{h\nu}{kT}\right) - 1} d\nu \quad (33.441)$$

En posant $x = h\nu/kT$, nous pouvons simplifier l'intégrande de sorte que:

$$M(T) = \frac{2\pi h}{c^2} \frac{(kT)^4}{h^4} \int_0^\infty \frac{x^3}{e^x - 1} dx = \frac{2\pi h}{c^2} \frac{(kT)^4}{h^4} \left(\frac{\pi^4}{15} \right) = \frac{2\pi^5 k^4}{15c^2 h^3} T^4 \quad (33.442)$$

Démontrons, car ce n'est de loin pas évident, que:

$$\int_0^{\infty} \frac{x^3}{e^x - 1} dx = \frac{\pi^4}{15} \quad (33.443)$$

Écrivons l'intégrande sous la forme suivante:

$$x^3 (e^x - 1)^{-1} = x^3 e^{-x} (e^x - 1)^{-1} e^x = x^3 e^{-x} (1 - e^{-x})^{-1} \quad (33.444)$$

Pour le terme:

$$\frac{1}{1 - e^{-x}} \quad (33.445)$$

Posons $t = e^{-x}$, nous avons alors:

$$\frac{1}{1 - t} \quad (33.446)$$

et nous avons démontré dans le chapitre d'Analyse Fonctionnelle, en utilisant des résultats du chapitre de Suites Et Série, que sous la condition $|t| < 1$, la dernière fraction peu s'écrire:

$$\frac{1}{1 - t} = \sum_{n=0}^{\infty} t^n = 1 + t + t^2 + t^3 + \dots \quad (33.447)$$

Dès lors:

$$x^3 e^{-x} (1 - e^{-x})^{-1} = x^3 e^{-x} (1 + e^{-x} + e^{-2x} + e^{-3x} + \dots) = x^3 e^{-x} + x^3 e^{-2x} + x^3 e^{-3x} + \dots \quad (33.448)$$

D'où:

$$\int_0^{\infty} \frac{x^3}{e^x - 1} dx = \int_0^{\infty} (x^3 e^{-x} + x^3 e^{-2x} + x^3 e^{-3x} + \dots) dx \quad (33.449)$$

Nous pourrions donc remplacer notre intégrale par une somme d'intégrales définies. Nous voyons qu'en faisant des intégrations par parties successives, nous obtenons:

$$\begin{aligned} \int_0^{\infty} x^3 e^{-nx} dx &= -\frac{1}{n} \left[x^3 e^{-nx} \right]_0^{\infty} + \frac{1}{n} \int_0^{\infty} 3x^2 e^{-nx} dx = \frac{3}{n} \int_0^{\infty} x^2 e^{-nx} dx \\ &= -\frac{3}{n^2} \left[x^2 e^{-nx} \right]_0^{\infty} + \frac{3}{n^2} \int_0^{\infty} 2x e^{-nx} dx = \frac{3}{n^2} \int_0^{\infty} 2x e^{-nx} dx \\ &= -\frac{6}{n^3} \left[x e^{-nx} \right]_0^{\infty} + \frac{6}{n^3} \int_0^{\infty} e^{-nx} dx = \frac{6}{n^3} \int_0^{\infty} e^{-nx} dx = \frac{6}{n^3} \left[-\frac{1}{n} e^{-nx} \right]_0^{\infty} \\ &= -\frac{6}{n^4} \left[e^{-nx} \right]_0^{\infty} = -\frac{6}{n^4} (e^{-\infty} - e^{-0}) = -\frac{6}{n^4} (0 - 1) = \frac{6}{n^4} \end{aligned} \quad (33.450)$$

Donc:

$$\int_0^{\infty} \frac{x^3}{e^x - 1} dx = 6 \left(1 + \frac{1}{2^4} + \frac{1}{3^4} + \frac{1}{4^4} + \dots \right) = 6 \sum_{n=1}^{\infty} \frac{1}{n^4} \quad (33.451)$$

Or, nous avons démontré dans le chapitre de Suite Et Séries lors de notre étude de la fonction de zêta de Riemann que:

$$\zeta(4) = \sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90} \quad (33.452)$$

Donc:

$$\int_0^{\infty} \frac{x^3}{e^x - 1} dx = 6 \sum_{n=1}^{\infty} \frac{1}{n^4} = 6\zeta(4) = \frac{6\pi^4}{90} = \frac{\pi^4}{15} \quad (33.453)$$

Au final, la constante de Stefan-Boltzmann peut donc s'exprimer analytiquement sous la forme:

$$\sigma = \frac{2\pi^5 k^4}{15c^2 h^3} = 5.67051 \cdot 10^{-8} \left[W \cdot m^{-2} \cdot K^4 \right] \quad (33.454)$$

Franchement..., il était difficile de le deviner...

Déterminons pour quelle fréquence, nous avons le maximum de densité d'énergie. En d'autres termes, cela revient à chercher où la dérivée:

$$\frac{dR(\nu, T)}{d\nu} = \frac{d}{d\nu} \left(8\pi \frac{h\nu^3}{c^3} \frac{1}{\exp\left(\frac{h\nu}{kT}\right) - 1} \right) \quad (33.455)$$

s'annule. Donc:

$$\frac{dR(\nu, T)}{d\nu} = \frac{8 \cdot 3\pi h\nu^2}{c^3} \frac{1}{\exp\left(\frac{h\nu}{kT}\right) - 1} - 8\pi \frac{h\nu^3}{c^3} \frac{h}{kT} \frac{\exp\left(\frac{h\nu}{kT}\right)}{\left(\exp\left(\frac{h\nu}{kT}\right) - 1\right)^2} \quad (33.456)$$

Divisons par $R(\nu, T)$:

$$\begin{aligned} \frac{1}{R(\nu, T)} \frac{dR(\nu, T)}{d\nu} &= \frac{3}{\nu} - \frac{h}{kT} \frac{\exp\left(\frac{h\nu}{kT}\right)}{\exp\left(\frac{h\nu}{kT}\right) - 1} = 0 \Rightarrow \frac{3}{\nu} \exp\left(\frac{h\nu}{kT}\right) - \frac{3}{\nu} - \frac{h}{kT} \exp\left(\frac{h\nu}{kT}\right) = 0 \\ \Rightarrow \frac{3}{\nu} - \frac{3}{\nu} \exp\left(-\frac{h\nu}{kT}\right) - \frac{h}{kT} &= 0 \Leftrightarrow \exp\left(-\frac{h\nu}{kT}\right) = 1 - \frac{1}{3} \frac{h\nu}{kT} \end{aligned} \quad (33.457)$$

La dernière relation admet une seule racine positive que nous pouvons déterminer avec Maple 4.00b (en utilisant la commande: >evalf(solve(exp(-x)-1+1/3*x=0,x));):

$$\frac{h\nu}{kT} = 2.82144 \quad (33.458)$$

Ce qui nous donne la "**deuxième loi de Wien**" ou "**loi de déplacement de Wien**" qui dit donc qu'à mesure qu'on augmente la température d'un corps noir, la longueur d'onde à laquelle l'intensité du rayonnement est la plus forte devient de plus en plus courte:

$$\nu_{\max} = 2.82144 \frac{kT}{h} = a \cdot \frac{kT}{h} \quad (33.459)$$

où a est appelée "**constante de Wien**". La figure ci-dessous représente bien ce résultat à l'aide des échelles logarithmiques:

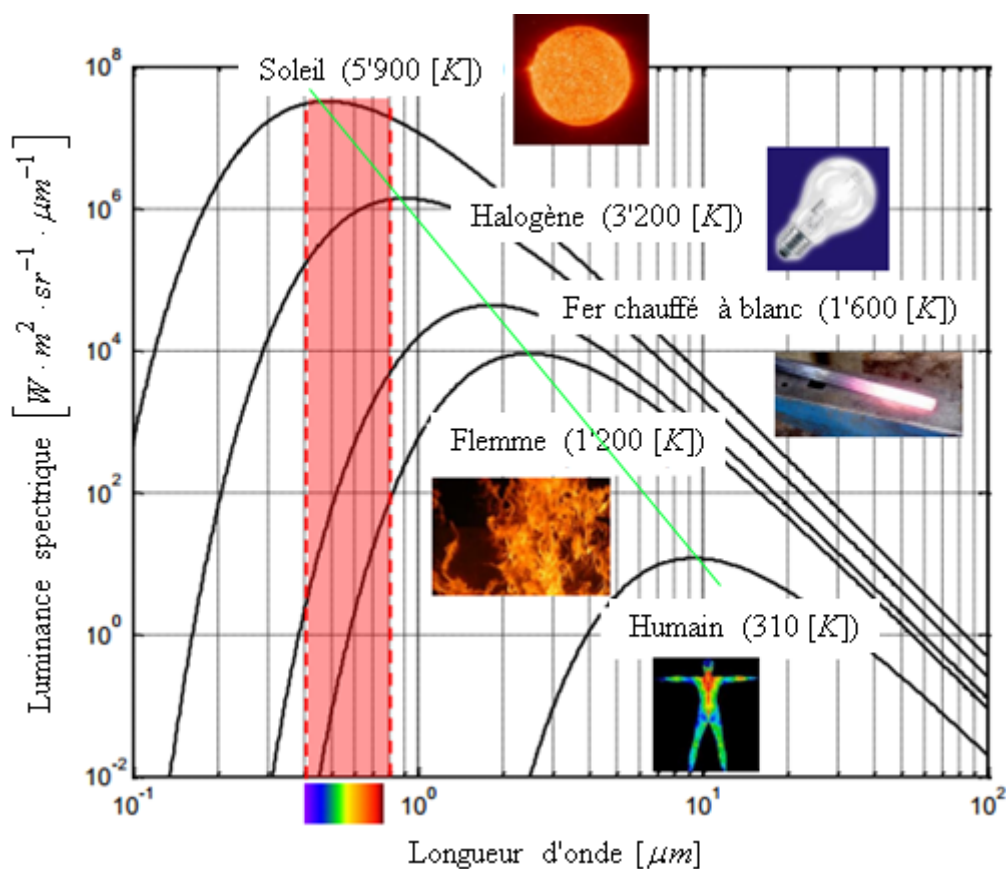


Figure: 33.6 - Représentation de la loi (distribution) de Planck avec le déplacement de Wien

Ainsi, non seulement l'augmentation de la température entraîne un accroissement de la quantité totale d'énergie rayonnée mais la loi de déplacement de Wien nous dit que la longueur d'onde à laquelle la quantité maximale de rayonnement est émise multipliée par la température du corps noir est toujours une constante (ou autrement vu: la longueur d'onde maximale est inversement proportionnelle à la température comme le montre le trait vert dans la figure ci-dessus)! Ce qui est un résultat extrêmement simple qui nous dit deux choses importantes:

- Si la température double, alors la longueur d'onde maximale sera la moitié de la longueur d'onde précédente.

- Si nous connaissons la constante (ce qui est notre cas), alors nous pouvons calculer la longueur d'onde maximale pour toute température à laquelle se trouve un corps noir.

Il est alors aisé de comprendre maintenant pourquoi tout matériau à des températures basses émet principalement des radiations de grande longueur d'onde dans la partie infrarouge du spectre et qu'à mesure que la température s'élève, il y a plus d'énergie rayonnée dans chaque région du spectre et la longueur d'onde de pointe diminue en se déplaçant vers les longueurs d'onde plus courtes. Il en résulte que la couleur de la lumière émise passe du rouge à l'orange, puis au jaune et enfin au blanc bleuâtre (et plus loin dans l'ultraviolet).

Remarque: Bien évidemment la deuxième loi de Wien est aussi parfois donnée dans la littérature non pas par rapport à la fréquence, mais à la longueur d'onde...

Insistons sur le fait que la loi de Planck n'est valable que dans les cas où le rayonnement est à l'équilibre thermique. Cette restriction est importante dans la pratique, car les phénomènes d'émission ou d'absorption de rayonnement par la matière se produisent le plus souvent dans des conditions hors de l'équilibre: dans le cas par exemple de l'éclairage par une lampe électrique ou du chauffage électrique par rayonnement infrarouge, il y a transformation irréversible (et donc hors d'équilibre) d'énergie électrique en énergie de rayonnement; de même, le rayonnement solaire est produit par les réactions nucléaires qui ont lieu à l'intérieur du soleil et qui consomment peu à peu sa substance; au niveau microscopique également, l'émission d'un photon par un atome excité est très souvent un retour irréversible de l'atome à son état fondamental (émission spontanée hors d'équilibre). Dans le cas du corps noir, au contraire, le rayonnement est confiné à l'intérieur d'une enceinte fermée (nous laissons éventuellement une fraction négligeable de ce rayonnement s'échapper à l'extérieur pour y être soumise aux mesures) et nous pouvons ainsi parvenir à l'équilibre thermique avec les parois.

La loi de Planck que nous avons démontrée précédemment est parfaitement vérifiée par l'expérience dans tout le domaine des températures accessibles à ce jour:

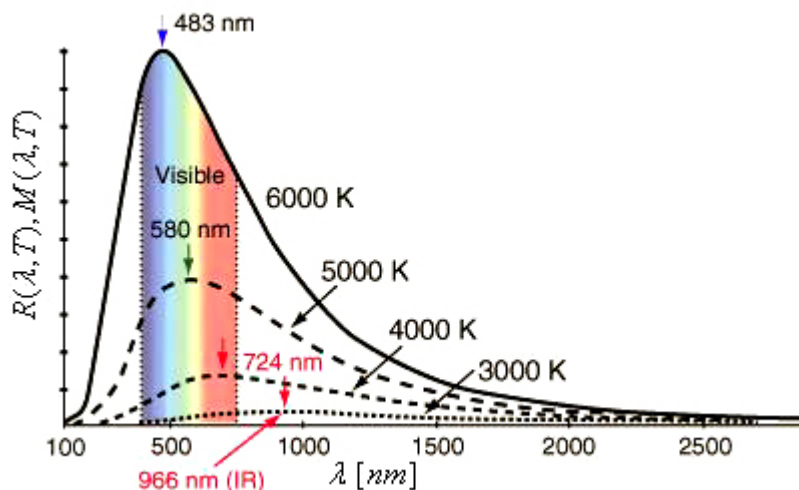
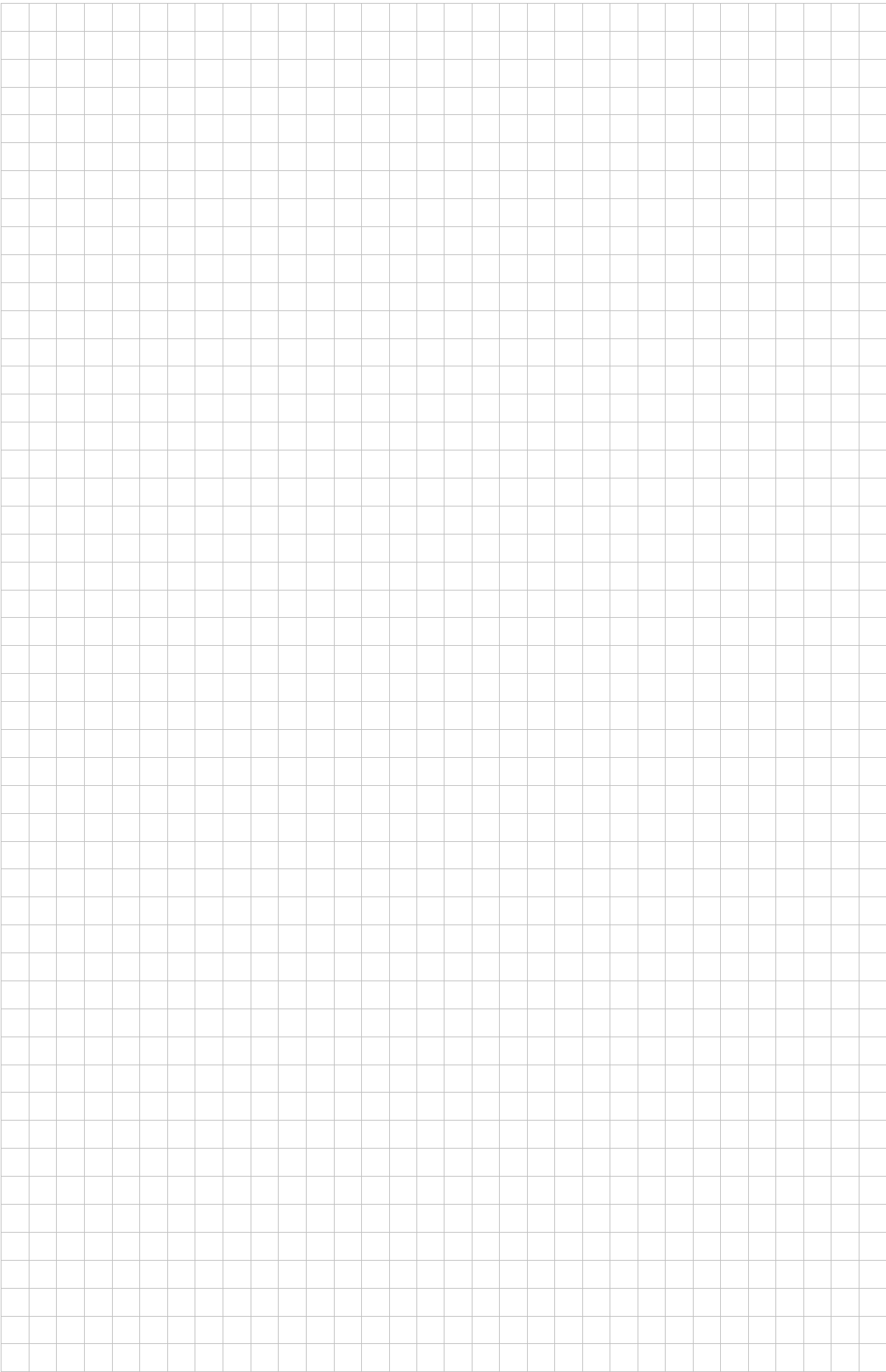


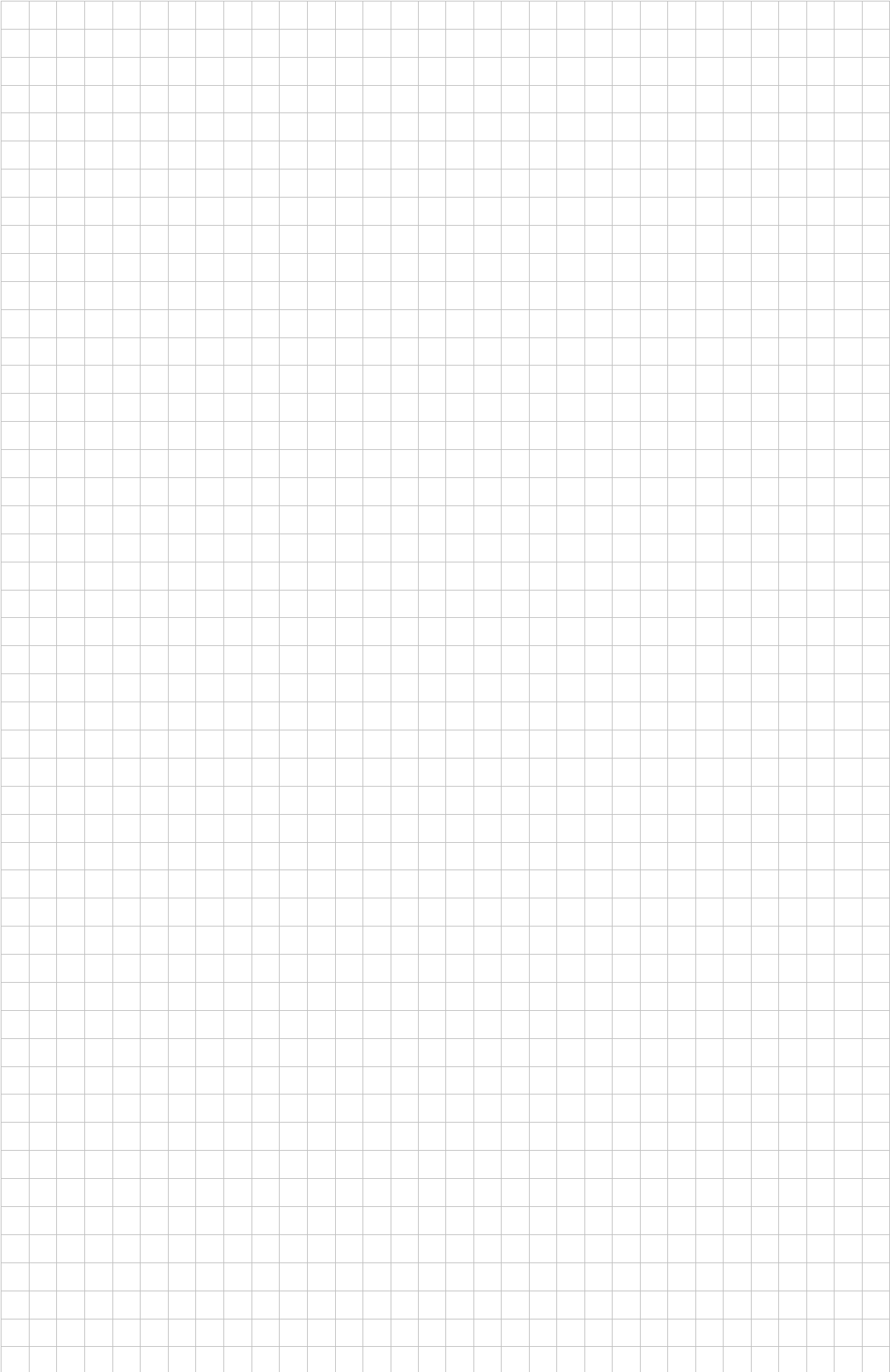
Figure: 33.7 - Représentation de la loi (distribution) de Planck

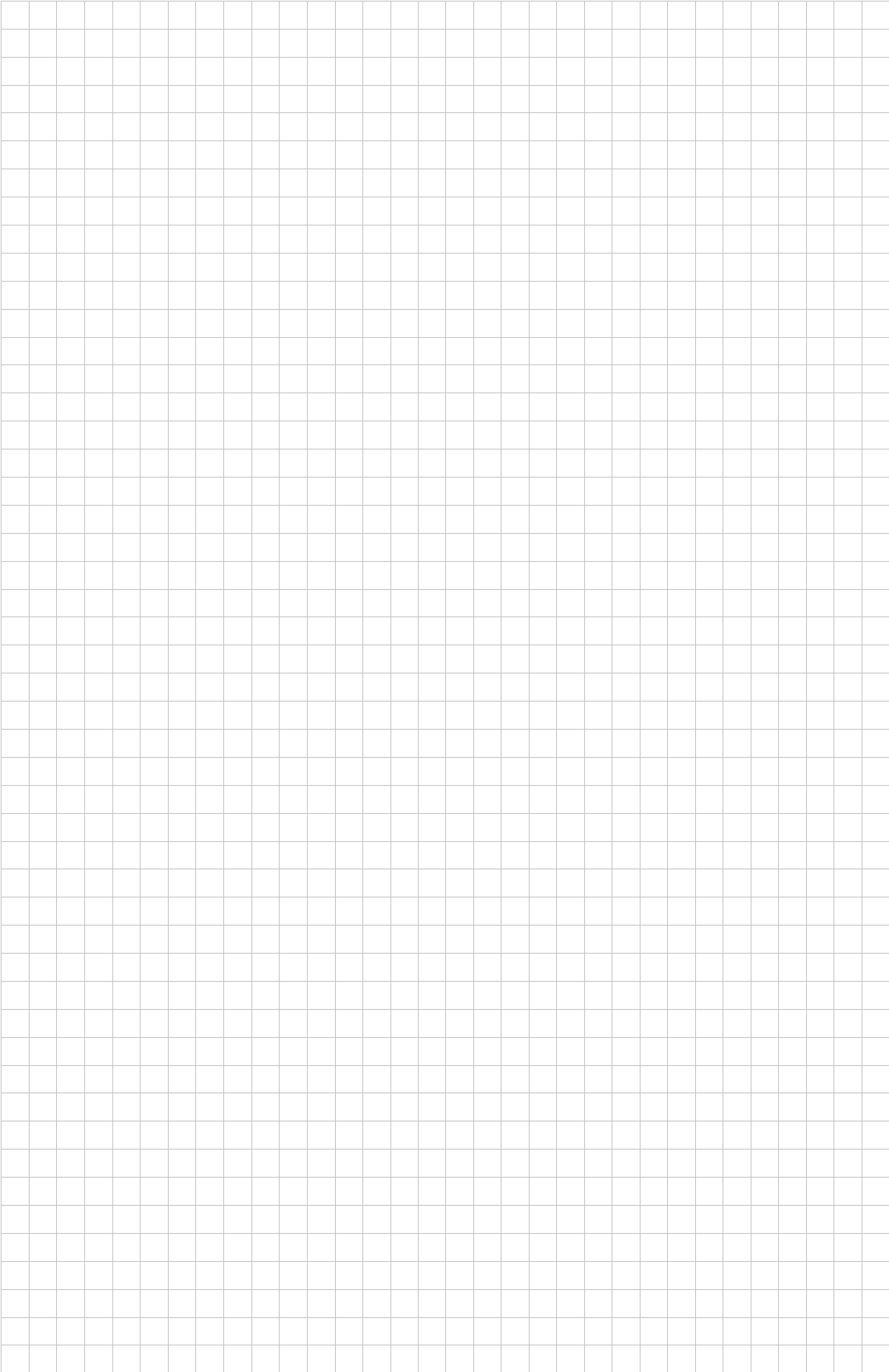
Nous remarquons à la lecture du graphique ci-dessus, qu'un corps chauffé entre 5'000 et 6'000 [K] a un pic d'émission au milieu du spectre visible. Dans le domaine de la colorimétrie, nous associons une température à une couleur en cherchant la température du corps noir pour laquelle le pic de radiation a son maximum dans la longueur d'onde de la couleur donnée.

Il est à noter que beaucoup de sources lumineuses émettent un flux lumineux qui ne suit pas la loi du corps noir (un filament d'ampoule, par exemple) et que la loi de Wien ne s'applique pas à eux. En revanche, il reste avéré qu'ils émettent à une longueur d'onde d'autant plus courte qu'ils sont chauds.

Il faut également garder à l'esprit que le flux lumineux provenant d'un objet n'est pas forcément de nature thermique ; autrement dit sa couleur ne renseigne pas toujours sur sa température. Par exemple, la couleur du ciel provient de lumière solaire bleue diffusée par l'air et non d'une hypothétique température de 15'000 [K]. De même un arbre est vert, non pas parce qu'il est à 8'000 [K], mais parce qu'il réfléchit la lumière verte qui compose la lumière du jour.







34. MÉCANIQUE DES MILIEUX CONTINUS

Au sens strict du terme, la mécanique des milieux continus (abrégée M.M.C.) est la branche de la mécanique qui a comme propos l'étude des mouvements, des déformations, des champs de contraintes au sein de milieux continus.

Définition:

D1. Nous désignons par "**milieu**", tout fluide (solide, liquide, gaz ou plasma selon ce que nous avons vu en thermodynamique), déformable ou non, quand nous le considérons d'un point de vue macroscopique, par opposition à une description corpusculaire.

D2. Nous désignons par "**milieu continu**", un milieu tel que si M et M' appartiennent à un milieu et si M' appartient au voisinage M , alors quelle que soit la déformation subie par ce milieu, dM' appartiendra au voisinage de dM .

Cette branche apparaît souvent comme la science de l'ingénieur qui permet de comprendre et de décrire le monde matériel qui nous entoure et les phénomènes courants qui s'y déroulent: mouvements de liquides, de gaz, vol des avions, hélicoptères, fusées, satellites, navigation des bateaux, déformations des corps solides, structure interne des étoiles, etc. Par ses attaches à la mécanique thermique (thermodynamique), elle s'étend jusqu'à la thermique, l'énergétique, l'acoustique.

Prenant en compte les comportements des milieux continus, elle englobe l'hydrodynamique, la dynamique des gaz, l'élasticité, l'acoustique, la plasticité et d'autres comportements. Elle est la clé de ce que nous appelons aujourd'hui la "modélisation", qui n'est autre que l'art d'analyser un phénomène physique et de le décrire en termes mathématiques, ce qui permet de l'étudier avec la rigueur propre à cette discipline.

Cette section du site est divisée en 4 parties principales: solides, liquides, gaz et plasmas (dont certaines notions ont délibérément été développées dans le chapitre de Musique Mathématique du site). Dans chaque partie, nous introduirons les outils mathématiques spécifiques à l'étude de tel ou tel milieu continu avec une complexité (toute relative) croissante. Cependant, par choix, il a été décidé d'exposer les théorèmes avec les outils mathématiques les plus simples possibles mais tout en arrivant aux mêmes résultats. Ainsi, par exemple, la démonstration de l'équation de Navier-Stokes qui prendrait 150 pages de développements mathématiques rigoureux n'en prend plus que 27. Il y a donc un avantage non négligeable aussi bien pour l'auteur que pour le lecteur à procéder ainsi.

Remarque: Concernant les équations de Navier-Stokes, nous donnerons aussi des exemples pratiques de celles-ci lors de notre étude de la météorologie (cf. chapitre de Génie Marin & Météo).

1. SOLIDES

Des atomes d'un même élément ou d'éléments différents s'assemblent en des édifices spécifiques. Cela conditionne la force de leurs interactions électriques, qui définissent la structure finale de la substance. Dans les conditions normales sur notre planète, la matière existe à l'état solide, liquide, gaz ou plasma. Si les forces interatomiques sont assez intenses, la collection de particules conserve sa forme et son volume.

Cette propriété de conserver la forme et le volume, ainsi que des propriétés élastiques distinguent les solides.

1.1. PRESSIONS

Les notions de "compression" et "contrainte" (que nous pouvons englober abusivement dans le terme de "pression") sont de première importance en mécanique des fluides (solides inclus donc!). Il convient donc de définir ces différents types de pression avec un minimum de rigueur!

Définitions:

D1. Nous appelons "**pression de compression**", notée traditionnellement P , le rapport exprimé par la force F qui s'exerce (s'appuie) sur un élément de surface S à la perpendiculaire de celle-ci. Ainsi, sous forme scalaire:

$$P = \frac{F}{S} \quad (34.1)$$

Remarque: Si une force agit sur une surface finie, nous parlons alors aussi de "**force répartie**".

D2. Nous appelons "**pression de contrainte**" le rapport exprimé par la force F qui tire sur un élément de surface S non nécessairement à la perpendiculaire, force qui peut dès lors être décomposée en deux vecteurs respectivement tangent et normal. Ainsi, sous forme vectorielle:

$$\vec{\sigma} + \vec{\tau} = \frac{d\vec{F}}{dS}$$

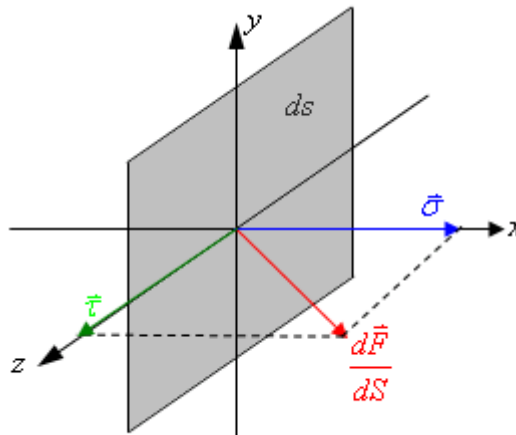


Figure: 34.1 - Illustration des contraintes tangentielle et normale

où $\vec{\sigma}$ et $\vec{\tau}$ sont respectivement la "**contrainte normale**" et la "**contrainte tangentielle**" (parfois indiquées avec un s en indice pour indiquer que c'est par rapport à une surface).

Nous pourrions très bien englober les deux définitions ci-dessus en une seule et travailler avec les signes des forces. Mais par souci de cohérence avec ce qui est enseigné dans les écoles, nous garderons ces deux définitions qui s'identifient par définition par le fait que leurs forces sont opposées par rapport à un élément de surface S .

1.2. ÉLASTICITÉ DES SOLIDES

D'une manière ou d'une autre, une contrainte de compression ou de traction peut déformer le triplet hauteur, largeur, épaisseur d'un corps. S'attaquer directement à l'étude d'un cas qui déforme ces trois paramètres est un peu long et sera abordé plus bas dans la partie traitant de la détermination de l'expression du module de Young de cisaillement.

Mais il est utile, ne serait-ce que du point de vue du vocabulaire de donner un exemple à partir du cas le plus simpliste qui puisse être. Si nous imaginons un corps élastique à une dimension (n'ayant ni hauteur, ni largeur mais juste une longueur) sous l'application de deux forces de contraintes parfaitement colinéaires mais antagonistes, nous pouvons imaginer que le corps en considération s'allonge d'un certain facteur.

Définition: La "**déformation normale**" sous des forces axiales et antagonistes est donnée par le rapport entre la variation de longueur du corps sur sa longueur initiale (soit: l'allongement relatif) tel que:

$$\varepsilon = \frac{|\Delta L|}{L_0} \quad (34.2)$$

Cette relation est une forme extrêmement simplifiée de tous les types de déformations qui peuvent exister et que nous verrons plus loin en détails.

Il y a nécessairement une relation entre forces de compression et de traction et la variation de dimension d'un corps. Cette relation est dépendante de la structure atomique du matériau et devrait rigoureusement faire appel à la physique quantique pour être déterminée (nous nous en abstenons cependant dans cette section du site). Nous observons cependant suivant les matériaux des caractéristiques diverses qui intéressent au plus haut point les ingénieurs:

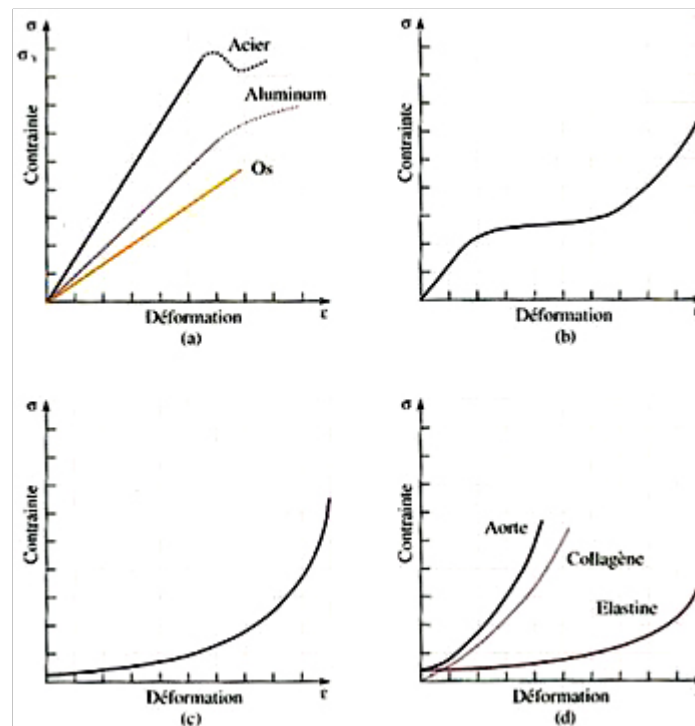


Figure: 34.2 - Comportement sous contrainte/compression pour certains matériaux

Les figures ci-dessus représentent la variation de la contrainte de compression en fonction de la

déformation pour certains matériaux (habituellement nous représentons ces caractéristiques en inversant les axes).

- Les matériaux ductiles comme l'acier doux (a), cessent d'être linéaires à la limite d'élasticité notée σ_y ci-dessus.
- Sous traction les polymères (b) caoutchouteux s'allongent d'abord en dépliant leurs molécules (cf. chapitre de Génie Des Matériaux) puis en tirant sur les liaisons chimiques (cf. chapitre de Chimie Quantique).
- La plupart des matériaux biologiques (c) sont sous contrainte, même lorsqu'ils ne sont pas déformés. La peau, par exemple, est comme un gant de caoutchouc enveloppant le corps.
- L'élastine (d) est habituellement renforcée de collagène dans les systèmes biologiques comme les artères. Un tendon est fait principalement de collagène.

Dans un cas plus général, les ingénieurs ont pour habitude de définir les points représentés ci-dessous dans leurs mesures d'essais de traction:

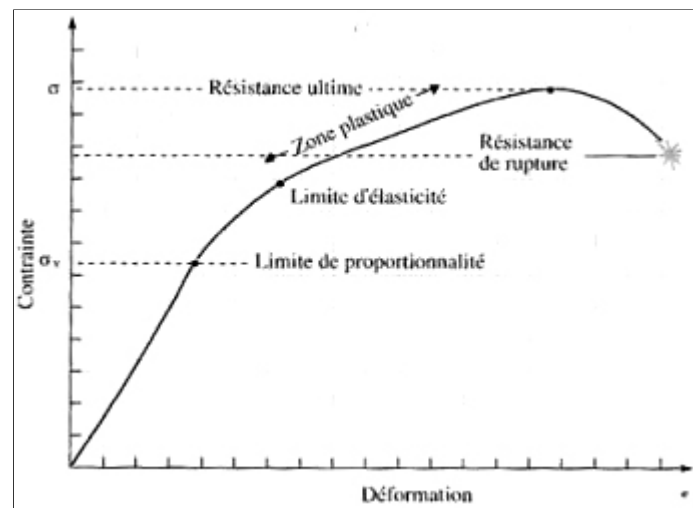


Figure: 34.3 - Définitions de termes importants pour l'étude des déformations

La caractéristique ci-dessus comporte une partie linéaire comme c'est le cas d'une certaine classe de matériaux. Cela signifie que la pente de la caractéristique est une constante, qui reflète la déformation élastique du matériau sous l'effet de la contrainte croissante. Cette contrainte élastique par unité de déformation définit le "**module de Young**" (il n'y a pas de composante tangentielle dans ce cas d'étude!):

$$E = \frac{\sigma}{\varepsilon} \quad (34.3)$$

cette relation étant valable aussi bien en contraintes de compression qu'en traction. Nous reviendrons sur cette relation dans les paragraphes suivants.

Remarques:

R1. La "**rhéologie**" est une partie de la mécanique qui étudie la plasticité, l'élasticité, la viscosité et la fluidité caractéristiques des corps déformables. C'est une branche très importante de l'ingénierie industrielle.

R2. Attention les calculs qui vont suivre sont relativement longs et difficiles et ce même si nous avons essayé de les simplifier au maximum. Cependant, tous les résultats nous seront infiniment utiles que ce soit pour déterminer l'équation de Navier-Stokes ou pour l'étude de la résistance des matériaux (cf. chapitre de Génie Mécanique)!

1.2.1. LOI DE HOOKE

Étant donnés les définitions données précédemment, nous obtenons la relation:

$$\frac{|\Delta L|}{L_0} = \frac{F}{E \cdot S} \quad (34.4)$$

qui est par définition la "**loi linéaire de Hooke**" en contrainte normale uniquement!

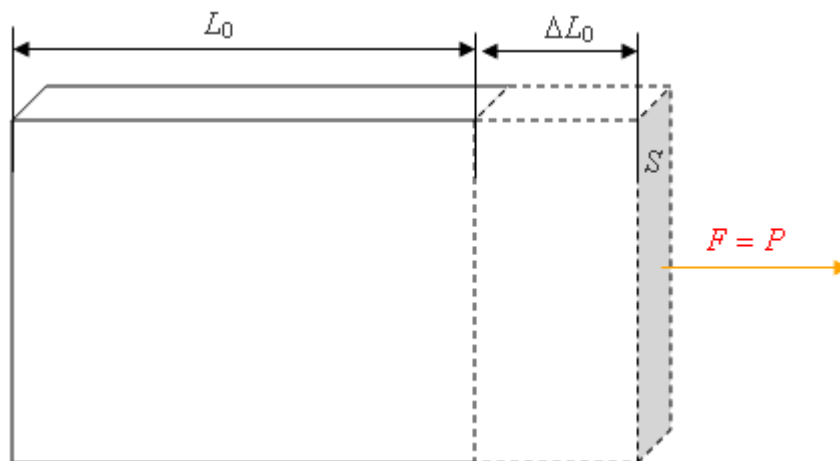


Figure: 34.4 - Illustration de l'effet d'une contrainte normale

Il est assez intuitif de supposer que plus la force de liaison des atomes constituant le matériau étudié est grande, plus grande est la force à appliquer pour éloigner les atomes, donc pour étirer le corps. Les solides, qui ont des grandes forces de liaisons, ont une haute température de fusion (cela est approfondi dans le chapitre traitant de la Chimie Quantique).

Si nous notons:

$$\frac{E \cdot S}{L_0} = k \quad (34.5)$$

Nous nous retrouvons avec la loi que nous connaissons:

$$F = k |\Delta L| \quad (34.6)$$

qui est la force de rappel des ressorts (cf. chapitre de Mécanique Classique et Génie Mécanique).

Mais il existe plusieurs types de contraintes avec leurs modules respectifs. Ainsi voici les définitions des plus importantes dans la partie linéaire de leur caractéristique avec le schéma explicatif associé:

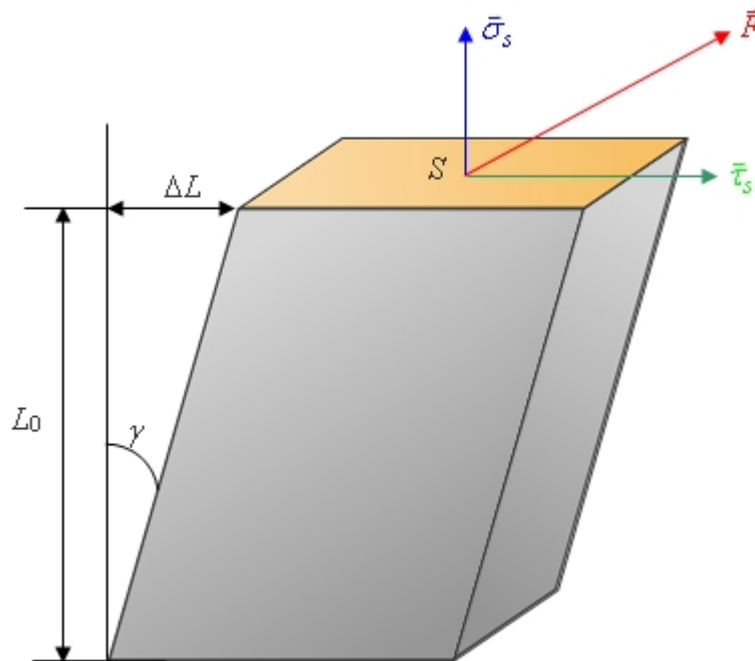


Figure: 34.5 - Illustration de l'effet d'une contrainte de cisaillement

D1. Nous définissons le "**module de cisaillement**" ou "**module de rigidité**" par le rapport de la composante normale de la force (pression de compression) à la déformation de cisaillement:

$$G = \frac{\sigma_s}{\varepsilon_s} = \frac{F_N / S}{\gamma} \quad (34.7)$$

où le numérateur est appelé "**contrainte de cisaillement**" et où γ est "**l'angle de déformation**". Généralement cet angle étant petit, nous avons l'approximation:

$$\varepsilon_s = \gamma \cong \tan(\gamma) = \frac{|\Delta L|}{L_0}$$

S est la surface de la face supérieure ou inférieure du corps déformé représenté ci-dessus.

D2. Nous définissons le "**module d'élasticité de glissement**", appelé également "**module de glissement**" ou encore "**module de Coulomb**" par le rapport de la composante tangentielle de la force (pression de contrainte) à la déformation de cisaillement:

$$G = \frac{\tau_s}{\varepsilon_s} = \frac{F_T / S}{\gamma} = \frac{E}{2(1+\eta)} \quad (34.8)$$

où η est le "**coefficient de Poisson**" dont nous démontrerons l'origine un peu plus bas dans le présent texte.

Remarquez que bien que le numérateur de la définition précédente soit une force divisée par une

surface, il ne s'agit pas d'une pression car la force est tangentielle (d'où le T en indice de F) à la surface.

C'est parce que toute force peut être décomposée en une force normale et tangentielle (voir la définition plus haut de la pression de compression et de la pression de contrainte) que nous avons les deux définitions distinctes ci-dessus. Dans la grande majorité des cas de laboratoires, nous nous arrangeons pour avoir une force purement tangentielle (d'où le T en indice de F) ou purement normale (d'où le N en indice de F) à la surface S .

Dans la pratique il n'est souvent fait usage que de la deuxième définition et ce à un point tel que cette dernière est souvent assimilée au "module de rigidité" aussi...

Exemple:

Une chose intéressante (pour la parenthèse...) si nous considérons que les plaques tectoniques sont en cisaillement entre elles nous avons alors d'après le module de glissement:

$$F_T = \gamma \cdot G \cdot S = \frac{|\Delta L|}{L_0} \cdot G \cdot S \quad (34.9)$$

Or pour une plaque tectonique en frottement de longueur L_0 sur une hauteur H :

$$F_T = \frac{|\Delta L|}{L_0} \cdot G \cdot (L_0 \cdot H) = G \cdot \left(\frac{|\Delta L|}{L_0} \right) \cdot (L_0 \cdot H) \quad (34.10)$$

et puisque l'énergie est une force multipliée par une distance, il vient:

$$E = F_T \cdot \Delta L = \left[G \cdot \left(\frac{|\Delta L|}{L_0} \right) \cdot (L_0 \cdot H) \right] \Delta L = G \cdot \Delta L^2 \cdot H \quad (34.11)$$

qui est typiquement l'énergie dégagée par le cisaillement de la friction de deux plaques tectoniques dont les surfaces de contact ont une hauteur moyenne H , une longueur initiale L_0 et qui subissent une déformation de ΔL .

Typiquement pour un tremblement de terre du type Sumatra (2004), nous avons:

$$\Delta L \cong 10 \text{ [m]}, H \cong 10'000 \text{ [m]}, L_0 \cong 10'000 \text{ [m]}, G \cong 10^{10} \text{ [Pa]} \quad (34.12)$$

Dès lors il vient:

$$E \cong 10^{16} \text{ [J]} \quad (34.13)$$

en d'autres termes... mille fois l'énergie de la bombe nucléaire d'Hiroshima.

Soit en notant M la magnitude sur l'échelle de Richter:

$$E \cong 10^{5+1.44M} = 10^{16} \Rightarrow M \cong 7.6 \quad (34.14)$$

alors que les estimations donnent un intervalle de 6.2 à 8.5... donc, nous ne sommes pas trop mauvais dans l'approche théorique.

Voilà pour un exemple non appliqué à l'industrie...

D3. Nous définissons le "**module de compressibilité omnidirectionnel**", comme le rapport de la contrainte volumique à la déformation volumique (nous démontrerons plus loin les développements mathématiques qui amènent au dernier terme de la relation):

$$\kappa = \frac{\varepsilon_v}{\sigma} = \frac{F/S}{\Delta V/V_0} = \frac{3(1-2\nu)}{E} \quad (34.15)$$

Nous pourrions encore définir beaucoup de modules tels que le module de flexion, de flexion pure, de flexion composée, de torsion... Nous étudierons certains d'entre eux plus loin.

Pour chacune des différentes définitions de modules que nous pouvons envisager, nous pouvons définir une loi de Hooke qui lui est adapté. Cependant, tout cela peut paraître assez arbitraire, mais au fait il n'en est rien car toutes les définitions de modules que nous avons vues précédemment sont un cas particulier d'une relation mathématique généralisée qui sera démontrée sur ce site dans un proche avenir.

1.2.2. MODULE DE GLISSEMENT

La condition nécessaire pour qu'un solide rigide soit en équilibre statique est comme nous l'avons vue dans le chapitre de Mécanique Classique, que la résultante des forces que l'extérieur exerce sur le corps soit nulle:

$$\vec{F}_{\text{ext}} = \vec{0} \quad (34.16)$$

Cependant, quand un solide subit des contraintes et qu'il peut en subir, il peut y avoir déformation qui peut être suivie d'une rupture ou d'une modification similaire. Plus, précisément, il y a "**déformation**" d'un corps (non nécessairement solide) quand les distances entre certains points du corps ont changé.

Lorsque dans l'étude théorique de l'élasticité, nous excluons les modifications du corps étudié telles que les ruptures, nous disons que nous nous restreignons aux "**déformations élastiques**".

La géométrie et la physique des déformations peuvent être complexes. Leur description se déduit de celle d'un certain nombre de déformations élémentaires dont nous préciserons plus loin les caractéristiques.

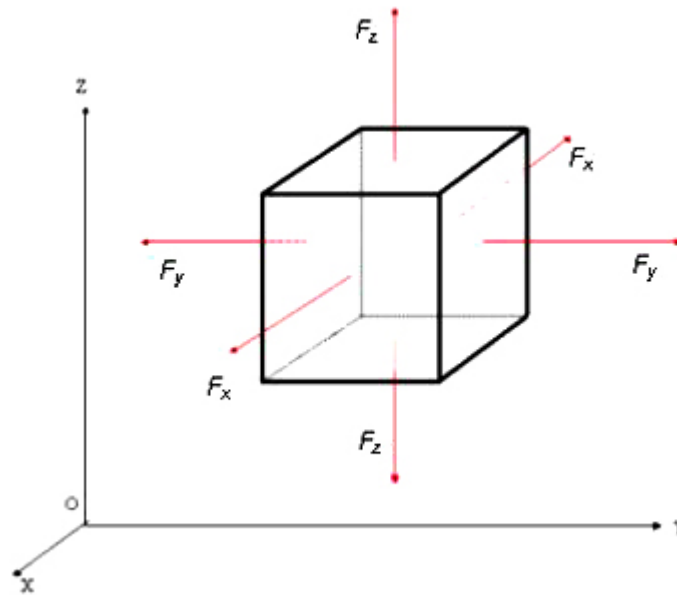


Figure: 34.6 - Cube sous contraintes normales

Les forces scalaires de contraintes de traction F_x, F_y, F_z engendrent sur leurs faces respectives des tensions "normales" (perpendiculaires donc!):

$$\sigma_x = \frac{F_x}{S_x} \quad \sigma_y = \frac{F_y}{S_y} \quad \sigma_z = \frac{F_z}{S_z} \quad (34.17)$$

En admettant que la force F_x agit seule, la déformation unitaire est par définition:

$$\varepsilon'_x = \frac{\sigma_x}{E} \quad (34.18)$$

Lorsqu'un parallélépipède est soumis à un effort de traction F_y , il y a intuitivement contraction des dimensions dans la direction x . Contraction observable de façon tout aussi intuitive pour F_z .

Nous avons alors si F_y agit seule:

$$\varepsilon''_x = -\eta \frac{\sigma_y}{E} \quad (34.19)$$

où le signe "-" indique une contraction et où η est un coefficient appelé "**coefficient de Poisson**".

Si F_z agit seule:

$$\varepsilon'''_x = -\eta \frac{\sigma_z}{E} \quad (34.20)$$

En acceptant le principe de superposition des forces, l'effet produit par plusieurs forces agissant simultanément est égal à la somme des effets produits par chacune des forces superposées agissant séparément. Dès lors:

$$\varepsilon_x = \varepsilon'_x + \varepsilon''_x + \varepsilon'''_x \quad (34.21)$$

Ceci est admissible, étant donné la linéarité des équations unissant la déformation unitaire et la tension normale. Nous obtenons alors:

$$\begin{aligned}\varepsilon_x &= \frac{1}{E} \left[\sigma_x - \eta (\sigma_y + \sigma_z) \right] \\ \varepsilon_y &= \frac{1}{E} \left[\sigma_y - \eta (\sigma_x + \sigma_z) \right] \\ \varepsilon_z &= \frac{1}{E} \left[\sigma_z - \eta (\sigma_x + \sigma_y) \right]\end{aligned} \quad (34.22)$$

En ayant procédé de manière identique pour les deux autres directions OY et OZ.

À partir des relations précédentes, il est aisé de trouver les équations unissant σ à ε :

$$\begin{aligned}\sigma_x &= \frac{E}{1+\eta} \left[\varepsilon_x + \frac{\eta}{1-2\eta} (\varepsilon_x + \varepsilon_y + \varepsilon_z) \right] \\ \sigma_y &= \frac{E}{1+\eta} \left[\varepsilon_y + \frac{\eta}{1-2\eta} (\varepsilon_x + \varepsilon_y + \varepsilon_z) \right] \\ \sigma_z &= \frac{E}{1+\eta} \left[\varepsilon_z + \frac{\eta}{1-2\eta} (\varepsilon_x + \varepsilon_y + \varepsilon_z) \right]\end{aligned} \quad (34.23)$$

Soit un matériau soumis à des contraintes diverses. À l'intérieur de celui-ci, nous opérons, par la pensée, l'extraction d'un parallélépipède rectangle. Les faces de celui-ci sont sollicitées par des contraintes normales σ et tangentiels τ (sur le schéma ci-dessous le solide est en équilibre statique).

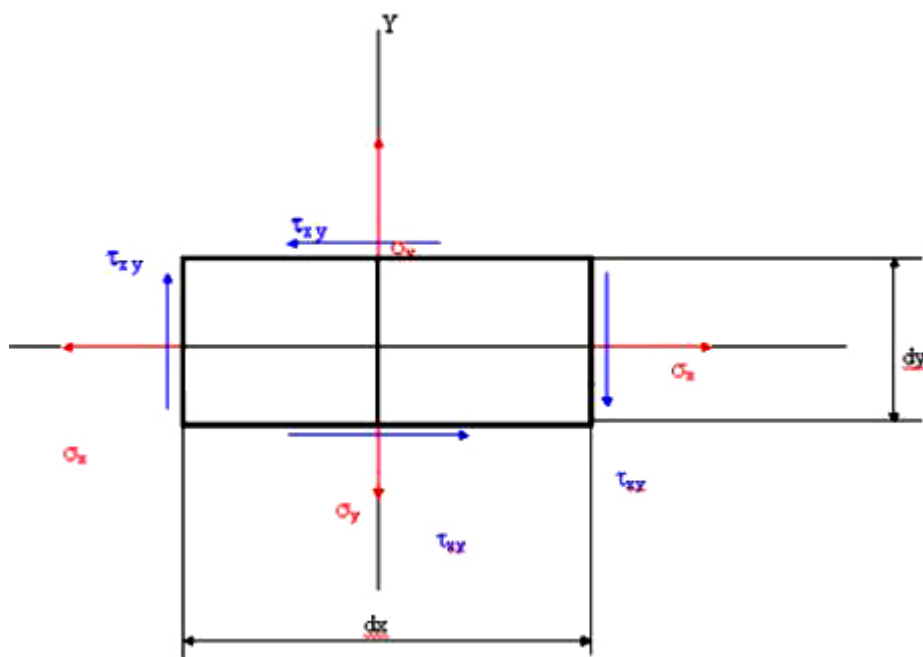


Figure: 34.7 - Illustration générique d'un matériau sous contraintes normales et tangentiels

Les contraintes normales σ et de tangentiels τ représentent les actions du parallélépipède de

matériau ôté mentalement sur les faces de l'élément examiné.

Il est intéressant (dans le sens que cela facilite l'analyse) de rechercher les contraintes qui existent dans un plan faisant un angle θ avec l'axe des x . Pour ce faire, nous imaginons un triangle de matière ayant un angle θ au sommet enlevé hors de la matière mentalement. Nous négligerons l'effet de la pesanteur.

Soit:

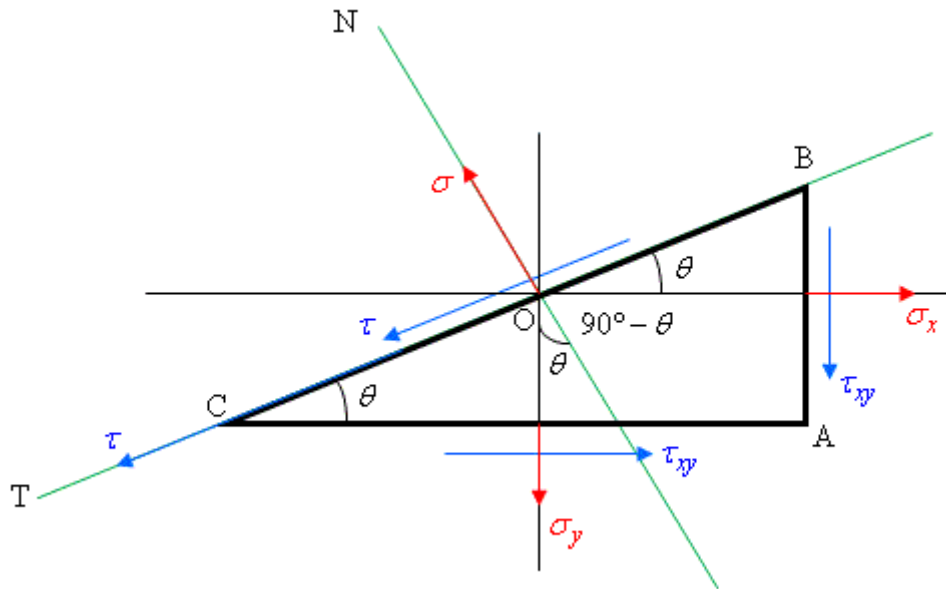


Figure: 34.8 - Recherche des expressions des contraintes dans un plan oblique

Posons:

$$\overline{CB} = ds, \overline{CA} = dx, \overline{AB} = dy \quad (34.24)$$

et dz étant l'épaisseur du solide (non représenté sur le schéma ci-dessus).

Sur la longueur ds , des contraintes apparaissent et se décomposent en contraintes normales σ et tangentielles τ (ces dernières étant aussi appelées "**contraintes de cisaillement**" ou "**contraintes de flexion**").

Le problème consiste à établir les relations entre σ, τ et σ_x, σ_y et τ_{xy} .

Les conventions de signes sont:

- Les contraintes σ exerçant une traction sont positives alors que les tensions σ exerçant une compression sont négatives.
- Les contraintes τ ayant tendance à faire tourner le parallélépipède dans le sens des aiguilles d'une montre, sont positives. Dans le sens antihoraire, elles seront négatives.

L'équation d'équilibre de projection sur la direction ON est:

$$\sigma_x dz ds - \left[\sigma_y dz dx \cos \theta + \sigma_x dz dy \cos (90^\circ - \theta) + \tau_{xy} dz dy \cos \theta + \tau_{xy} dz dx \cos (90^\circ - \theta) \right] = 0 \quad (34.25)$$

Rappelons que:

$$\sigma_i = \frac{F_{iN}}{S} \Leftrightarrow F_{iN} = S\sigma_i \quad (34.26)$$

Comme $dx = ds \cos \theta$ et $dy = ds \sin \theta$ nous avons:

$$\begin{aligned} \sigma ds &= \sigma_x ds \sin^2 \theta + \sigma_y ds \cos^2 \theta + \tau_{xy} ds \sin \theta \cos \theta + \tau_{xy} ds \sin \theta \cos \theta \\ \sigma &= \sigma_x \sin^2 \theta + \sigma_y \cos^2 \theta + \tau_{xy} \sin 2\theta \end{aligned} \quad (34.27)$$

comme:

$$\sin^2 \theta = \frac{1 - \cos 2\theta}{2} \quad \text{et} \quad \cos^2 \theta = \frac{1 + \cos 2\theta}{2} \quad (34.28)$$

alors:

$$\sigma = \frac{\sigma_x}{2} (1 - \cos 2\theta) + \frac{\sigma_y}{2} (1 + \cos 2\theta) + \tau_{xy} \sin 2\theta \quad (34.29)$$

Finalement:

$$\sigma = \frac{1}{2} (\sigma_x + \sigma_y) - \frac{1}{2} (\sigma_x - \sigma_y) \cos 2\theta + \tau_{xy} \sin 2\theta \quad (34.30)$$

Conclusion: En fonction de σ_x , σ_y et τ_{xy} , il est possible de calculer la tension normale qui existe sur une surface plane quelconque d'angle θ .

L'équation d'équilibre de projection sur la direction de OT est:

$$\begin{aligned} \tau dz ds - \sigma_x dz dy \cos \theta + \tau_{xy} dz dy \cos(90^\circ - \theta) + \sigma_y dz dx \cos(90^\circ - \theta) - \tau_{xy} dz dx \cos \theta &= 0 \\ \tau ds &= \sigma_x ds \sin \theta \cos \theta - \sigma_y ds \sin \theta \cos \theta - \tau_{xy} ds \sin^2 \theta + \tau_{xy} ds \cos^2 \theta = 0 \\ \tau &= \frac{1}{2} (\sigma_x - \sigma_y) \sin 2\theta + \tau_{xy} (\cos^2 \theta - \sin^2 \theta) \end{aligned} \quad (34.31)$$

comme $\cos^2 \theta - \sin^2 \theta = \cos 2\theta$ alors finalement:

$$\tau = \frac{1}{2} (\sigma_x - \sigma_y) \sin 2\theta + \tau_{xy} \cos 2\theta \quad (34.32)$$

Conclusion: En fonction de σ_x , σ_y et τ_{xy} , il est possible de calculer la tension τ tangentielle qui existe sur une surface plane quelconque d'angle θ .

Soit, à présent, la situation suivante:

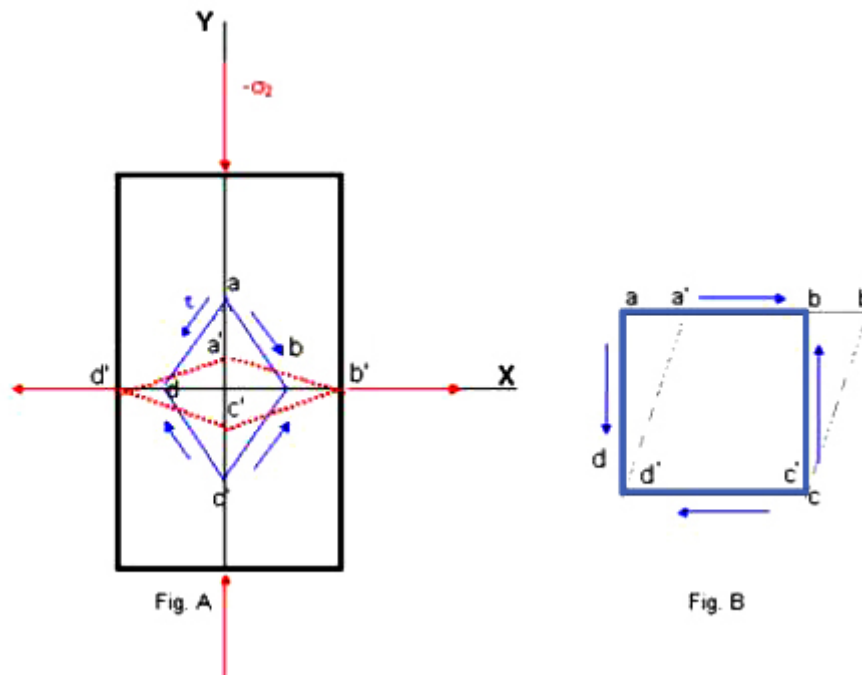


Figure: 34.9 - Mise en situation pour revenir au cas tridimensionnel

Il s'agit à gauche d'un bloc de matière dont l'on extrait virtuellement un petit plan de forme carrée (en bleu sur la figure de gauche) que l'on va étudier en ne prenant en premier lieu qu'un des triangles rectangles le composant pour ensuite étudier l'ensemble.

Avant la sollicitation, nous considérons donc le losange $abcd$ qui est en fait initialement un carré à $\pi/4$ suivant la direction OX (schéma perspective suite à la demande d'un internaute):

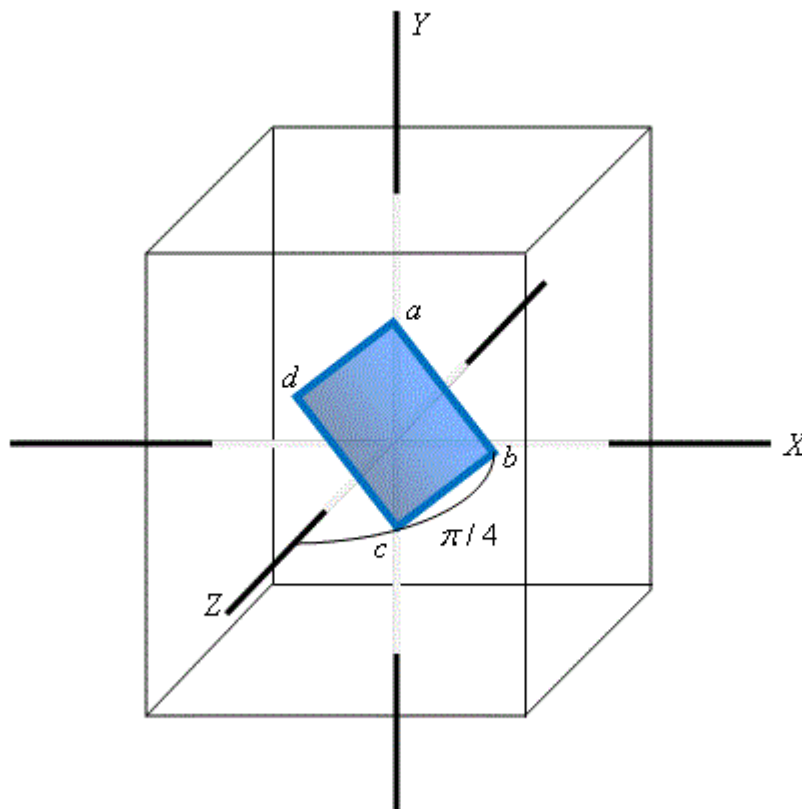


Figure: 34.10 - Situation initiale

Pendant la sollicitation, ce losange se déforme sous l'action des contraintes tangentielles décomposées en contraintes de cisaillement pures et devient le losange $a'b'c'd'$ (schéma perspective suite à la demande d'un internaute):

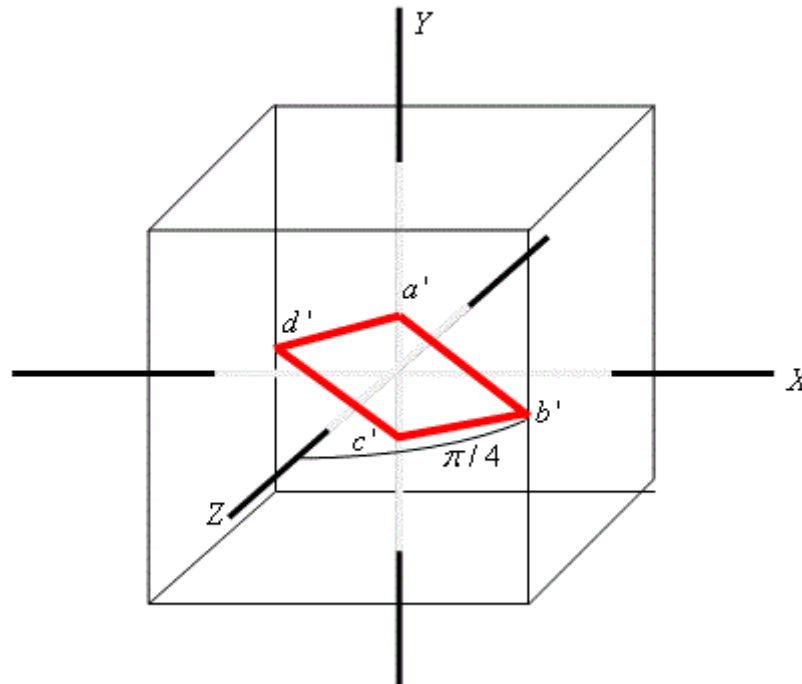


Figure: 34.11 - Situation finale

La diagonale bd est alors étendue et la diagonale ac est comprimée. L'angle en a qui valait $\pi/2$ vaut après déformation $\pi/2 + \gamma$ (en a'). De même, l'angle en b qui valait $\pi/2$ vaut à présent $\pi/2 - \gamma$ (Figure A).

Remarque: L'angle γ est appelé "**angle de glissement**" et nous le considérerons comme faible.

Nous pouvons nous rendre compte de l'effet de la déformation en isolant le losange et en lui faisant subir une rotation de $\pi/4$. Après déformation, nous avons la forme indiquée par les lignes en pointillés (Figure B).

L'angle de glissement étant petit, nous avons:

$$\frac{aa'}{ad} = \tan \gamma \approx \gamma \quad (34.33)$$

Donc γ représente le glissement du côté ab par rapport à dc divisé par la distance entre les deux plans ab et dc . L'analyse qui vient d'être effectuée reste valable quel que soit le corps solide ou liquide considéré.

Soit, à présent, le cas d'un solide élastique obéissant à la loi de Hooke. Le problème va consister à établir la relation entre l'angle de glissement γ et les contraintes tangentielles τ agissant sur les côtés du losange.

Soit le triangle rectangle oab . L'allongement du côté ob et le raccourcissement du côté oa pendant la

déformation s'obtiennent à partir des équations suivantes:

$$\begin{aligned}\varepsilon_x &= \frac{1}{E} \left[\sigma_x - \eta (\sigma_y + \sigma_z) \right] \\ \varepsilon_y &= \frac{1}{E} \left[\sigma_y - \eta (\sigma_x + \sigma_z) \right] \\ \varepsilon_z &= \frac{1}{E} \left[\sigma_z - \eta (\sigma_x + \sigma_y) \right]\end{aligned} \quad (34.34)$$

Comme:

$$\sigma_x = \sigma_1 = \tau \quad \sigma_y = \sigma_2 = -\tau \quad \sigma_z = 0 \quad (34.35)$$

Nous avons:

$$\varepsilon_x = \frac{1}{E} [\tau + \eta \tau] = \frac{(1+\eta)}{E} \tau \quad \text{et} \quad \varepsilon_y = \frac{1}{E} [-\tau - \eta \tau] = -\frac{(1+\eta)}{E} \tau \quad (34.36)$$

Donc:

$$oa' = oa + oa\varepsilon_y = oa(1 + \varepsilon_y) = oa \left[1 - \frac{(1+\eta)}{E} \tau \right] \quad (34.37)$$

alors la longueur oa' diminue si τ augmente .

$$ob' = ob + ob\varepsilon_x = ob(1 + \varepsilon_x) = ob \left[1 + \frac{(1+\eta)}{E} \tau \right] \quad (34.38)$$

donc ob' augmente si τ augmente.

Pour l'angle triangle rectangle $oa'b'$, nous avons:

$$\tan(oa'b') = \tan \left[\frac{1}{2} \left(\frac{\pi}{2} + \gamma \right) \right] = \operatorname{tg} \left(\frac{\pi}{4} + \frac{\gamma}{2} \right) = \frac{ob'}{oa'} = \frac{ob \left[1 + \frac{(1+\eta)}{E} \tau \right]}{oa \left[1 - \frac{(1+\eta)}{E} \tau \right]} = \frac{1 + \frac{(1+\eta)}{E} \tau}{1 - \frac{(1+\eta)}{E} \tau} \quad (34.39)$$

Or:

$$\tan \left(\frac{\pi}{4} + \frac{\gamma}{2} \right) = \frac{\tan \left(\frac{\pi}{4} \right) + \tan \left(\frac{\gamma}{2} \right)}{1 - \tan \left(\frac{\pi}{4} \right) \tan \left(\frac{\gamma}{2} \right)} = \frac{1 + \tan \left(\frac{\gamma}{2} \right)}{1 - \tan \left(\frac{\gamma}{2} \right)} \quad (34.40)$$

Comme $\tan \frac{\gamma}{2} \cong \frac{\gamma}{2}$ (γ est petit) nous avons:

$$\frac{1 + \frac{\gamma}{2}}{1 - \frac{\gamma}{2}} = \frac{1 + \left(\frac{1 + \eta}{E} \right) \tau}{1 - \left(\frac{1 + \eta}{E} \right) \tau} \quad (34.41)$$

Soit:

$$\frac{\gamma}{2} = \frac{1 + \eta}{E} \tau \quad (34.42)$$

Finalement, nous avons la relation donnant le "**module de glissement**", ou "**module de Coulomb**", que nous avons donné plus haut sans démonstration:

$$\frac{\tau}{\gamma} = G = \frac{E}{2(1 + \eta)} \quad (34.43)$$

1.2.3. MODULE DE COMPRESSIBILITÉ

Nous reste encore à voir la provenance mathématique de l'expression d'un autre module tout aussi important que le module en cisaillement: le module de compressibilité κ .

Soient les équations déterminées dans l'étude précédente:

$$\begin{aligned} \varepsilon_x &= \frac{1}{E} \left[\sigma_x - \eta (\sigma_y + \sigma_z) \right] \\ \varepsilon_y &= \frac{1}{E} \left[\sigma_y - \eta (\sigma_x + \sigma_z) \right] \\ \varepsilon_z &= \frac{1}{E} \left[\sigma_z - \eta (\sigma_x + \sigma_y) \right] \end{aligned} \quad (34.44)$$

Si les forces appliquées sur le cube sont égales en intensité, nous avons:

$$\sigma_x = \sigma_y = \sigma_z = \sigma \quad (34.45)$$

Ce qui nous donne:

$$\begin{aligned} \varepsilon_x &= \frac{1}{E} \left[\sigma - \eta (\sigma + \sigma) \right] = \frac{\sigma}{E} (1 - 2\eta) \\ \varepsilon_y &= \frac{1}{E} \left[\sigma - \eta (\sigma + \sigma) \right] = \frac{\sigma}{E} (1 - 2\eta) \\ \varepsilon_z &= \frac{1}{E} \left[\sigma - \eta (\sigma + \sigma) \right] = \frac{\sigma}{E} (1 - 2\eta) \end{aligned} \quad (34.46)$$

En sommant les termes selon le principe de superposition linéaire des forces:

$$\varepsilon_x + \varepsilon_y + \varepsilon_z = \frac{3\sigma}{E} (1 - 2\eta) \quad (34.47)$$

Or:

$$\begin{aligned}
\varepsilon_x + \varepsilon_y + \varepsilon_z &= \frac{|\Delta L_x|}{L_{x0}} + \frac{|\Delta L_y|}{L_{y0}} + \frac{|\Delta L_z|}{L_{z0}} \\
\varepsilon_x L_{x0} L_{y0} L_{z0} + \varepsilon_y L_{x0} L_{y0} L_{z0} + \varepsilon_z L_{x0} L_{y0} L_{z0} &= |\Delta L_x| L_{y0} L_{z0} + |\Delta L_y| L_{x0} L_{z0} + |\Delta L_z| L_{x0} L_{y0} \quad (34.48) \\
\Rightarrow V_0 (\varepsilon_x + \varepsilon_y + \varepsilon_z) &= \Delta V \\
\Rightarrow \varepsilon_x + \varepsilon_y + \varepsilon_z &= \frac{\Delta V}{V_0}
\end{aligned}$$

Finalement:

$$\frac{\Delta V}{V_0} = \sigma \frac{3(1-2\eta)}{E} \quad (34.49)$$

ce que nous notons également:

$$\varepsilon_v = \sigma \frac{3(1-2\eta)}{E} = \sigma \cdot \kappa \quad (34.50)$$

ou encore:

$$\kappa = \frac{\varepsilon_v}{\sigma} \quad (34.51)$$

avec κ étant par définition le "**coefficient de compressibilité**".

1.2.4. MODULE DE FLEXION

Pour l'étude du module de flexion considérons la situation ci-dessous:

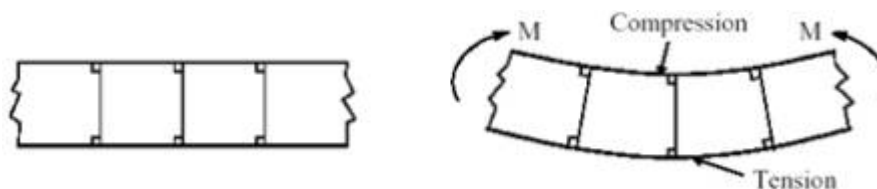


Figure: 34.12 - Exemple d'une barre en flexion

La figure de gauche ci-dessus représente un matériau à l'état statique. La figure de droite représente le même matériau mais soumis à un moment de force couplé M .

Comme le matériau subit à sa surface à la fois une compression et à l'opposé une tension, il doit donc exister une frontière (une ligne ou un plan) où aucune contrainte n'existe. Cette ligne ou ce plan (c'est rare que nous ayons affaire à un matériau ayant uniquement deux dimensions...) est appelé "**plan neutre**". Ce plan neutre va nous servir de référence pour définir la contrainte de flexion.

Maintenant que ce plan est défini, considérons les figures ci-dessous:

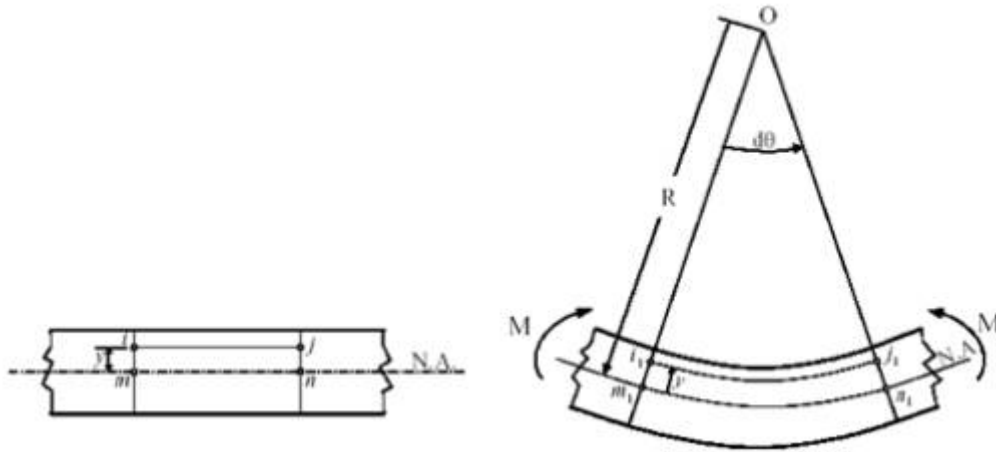


Figure: 34.13 - Illustration du plan de flexion pour déterminer le module de flexion

Soit R le rayon de courbure de la barre (cylindre, plaque, parallélépipède, ...). La déformation sur le segment i_1j_1 est définie par la relation:

$$\varepsilon_{NP} = \frac{i_1j_1 - ij}{ij} \quad (34.52)$$

Les longueurs mn et ij sont définies par:

$$ij = mn = m_1n_1 = R d\theta \quad (34.53)$$

et la longueur i_1j_1 par:

$$i_1j_1 = (R - y) d\theta \quad (34.54)$$

ainsi l'expression de la déformation devient:

$$\varepsilon_{NP} = \frac{(R - y) d\theta - R d\theta}{R d\theta} = -\frac{y}{R} \quad (34.55)$$

ce qui indique que la déformation varie de façon linéaire avec y .

Nous pouvons définir le module de flexion par:

$$E = \frac{\sigma}{\varepsilon_{NP}} = \frac{-\sigma R}{y} \quad (34.56)$$

Considérons l'état statique de la barre. La somme des contraintes de tractions et compressions sont alors nulles. Effectivement, nous le voyons bien si nous considérons le schéma ci-dessous:

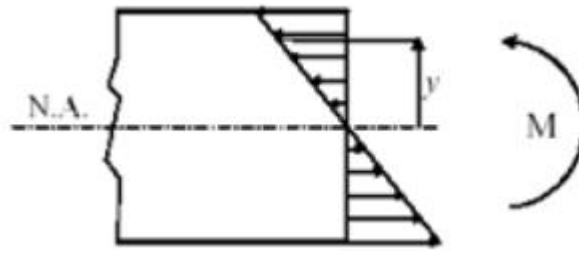


Figure: 34.14 - Agrandissement sur le plan de flexion

Considérons σdS la force agissante sur un élément de surface dS . Nous pouvons considérer l'équilibre des forces à l'état statique tel que:

$$\sum F = \int \sigma dS = 0 \quad (34.57)$$

En substituant l'expression de la contrainte obtenue précédemment:

$$\int -\frac{E y}{R} dS = -\frac{E}{R} \int y dS = 0 \quad (34.58)$$

En supposant linéaire la caractéristique de contrainte en première approximation donc $E = c^{te}$.

En simplifiant un tant soit peu:

$$\int y dS = 0 \quad (34.59)$$

Si nous multiplions l'intégrale par σ alors la relation doit être égale au moment de force M appliqué tel que:

$$\int \sigma y dS = M \quad (34.60)$$

En substituant par l'expression de la contrainte obtenue précédemment:

$$M = \int \sigma y dS = \int -\frac{E y^2}{R} dS = -\frac{E}{R} \int y^2 dS \quad (34.61)$$

Ce qui nous amène à définir le terme:

$$I = \int y^2 dS \quad (34.62)$$

que les ingénieurs nomment le "**moment d'inertie de la barre par rapport au plan neutral**" ou encore "**moment d'inertie statique**". Ce terme représente une mesure de la rigidité de la section transversale de la barre d'un point de vue géométrique, sans considérations des propriétés matérielles.

Substituant cette relation dans l'équation de contrainte de flexion, nous obtenons le "**module de flexion**":

$$E = \frac{MR}{I} = -\frac{\sigma R}{y} \quad (34.63)$$

La difficulté pour l'ingénieur consiste souvent à localiser mathématiquement le plan neutre...

1.3. ONDES TRANSVERSALES DANS LES SOLIDES

Les ondes sonores transversales ou "**ondes S**" (ondes de cisaillement) ne se produisent que dans les solides. Les couches successives du milieu se déplacent latéralement sans qu'il y ait de changement de volume, de densité ou de pression:

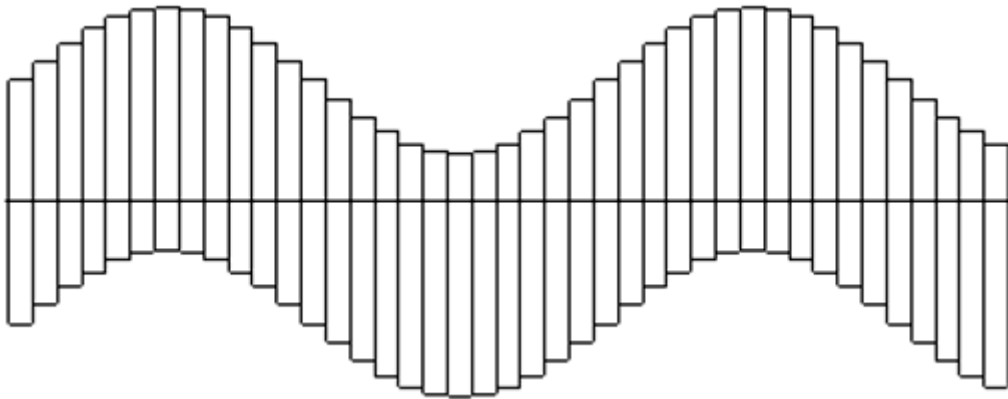


Figure: 34.15 - Exemple d'onde de cisaillement

Le milieu se déforme de la même manière que vous pouvez déformer un livre ou une rame de papier posé à plat en poussant le haut horizontalement. Ni le livre, ni la rame ne changent de volume.

L'obtention de l'équation d'onde pour des ondes transversales est presque la même que pour une corde (cf. chapitre de Mécanique Ondulatoire). Prenons trois minces couches planes contiguës du milieu (voir figure ci-dessous):

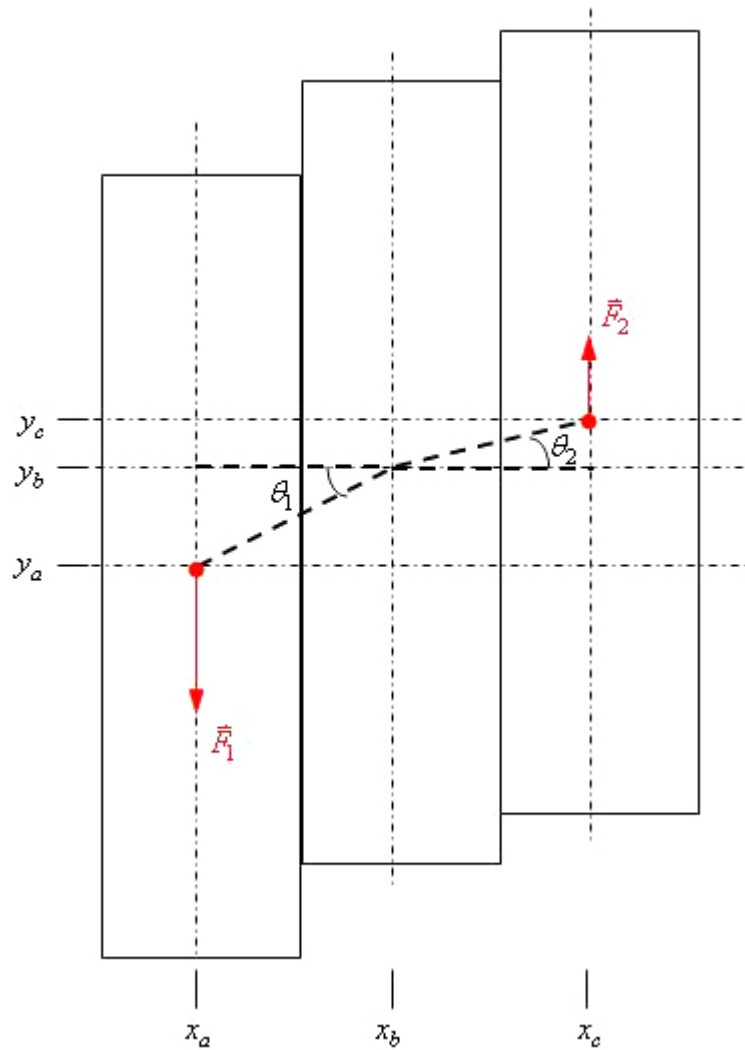


Figure: 34.16 - Agrandissement sur trois couches d'une onde de cisaillement

Les centres des couches se situent en x_a, x_b, x_c avec:

$$x_c - x_b = x_b - x_a = \Delta x \quad (34.64)$$

Le déplacement transversal des trois couches adjacentes est y_a, y_b, y_c . Les angles de déformation respectivement entre la couche b et la couche a , et, entre la couche c et la couche b sont au premier ordre en approximation de Taylor (cf. chapitre sur les Suites et Séries):

$$\begin{aligned} \theta_1 &= \arctan \left(\frac{y_b - y_a}{x_b - x_a} \right) \cong \frac{y_b - y_a}{x_b - x_a} \\ \theta_2 &= \arctan \left(\frac{y_c - y_b}{x_c - x_b} \right) \cong \frac{y_c - y_b}{x_c - x_b} \end{aligned} \quad (34.65)$$

Si nous calculons les forces entre les couches pour un morceau de couche de surface S , nous obtenons:

$$\begin{aligned} \frac{F_1}{S} &= G \frac{y_b - y_a}{dx} \\ \frac{F_2}{S} &= G \frac{y_c - y_b}{dx} \end{aligned} \quad (34.66)$$

où G est le module de glissement du milieu. La résultante des forces est alors:

$$F_b = F_2 - F_1 = SG \left(\frac{y_c - y_b}{dx} - \frac{y_b - y_a}{dx} \right) \quad (34.67)$$

La force de la tranche F_b sera égale à tout moment au produit de la masse du morceau de couche b , d'épaisseur dx , surface S et densité ρ , multipliée par l'accélération de la couche:

$$\left(\frac{\partial^2 y}{\partial t^2} \right)_b \quad (34.68)$$

Nous avons alors:

$$F_b = ma_b = \rho S dx \left(\frac{\partial^2 y}{\partial t^2} \right)_b = SG \left(\frac{y_c - y_b}{dx} - \frac{y_b - y_a}{dx} \right) \quad (34.69)$$

Ce qui donne:

$$\left(\frac{\partial^2 y}{\partial t^2} \right)_b = \frac{G}{\rho} \frac{\left(\frac{y_c - y_b}{dx} - \frac{y_b - y_a}{dx} \right)}{dx} = \frac{G}{\rho} \left(\frac{\partial^2 y}{\partial x^2} \right)_b \quad (34.70)$$

Ce que nous venons de déduire pour une valeur quelconque x_b , est aussi vrai pour n'importe quelle coordonnée:

$$\frac{\partial^2 y}{\partial t^2} = \frac{G}{\rho} \frac{\partial^2 y}{\partial x^2} \quad (34.71)$$

et la vitesse de propagation des ondes transversales est donc:

$$\frac{\partial^2 y}{\partial t^2} = \frac{G}{\rho} \frac{\partial^2 y}{\partial x^2} \quad (34.72)$$

Le rapport G/ρ a les unités du carré d'une vitesse:

$$\frac{G}{\rho} = \frac{\frac{N}{m^2}}{\frac{kg}{m^3}} = \frac{\frac{kg \cdot m}{m^2 \cdot s^2}}{\frac{kg}{m^3}} = \frac{kg \cdot m \cdot m^3}{m^2 \cdot s^2 \cdot kg} = \frac{m^2}{s^2} \quad (34.73)$$

Il s'agit donc d'une équation d'onde de la forme (rappel) d'une équation de Poisson (plus particulièrement il s'agit d'une équation de d'Alembert):

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{v^2} \frac{\partial^2 y}{\partial t^2} \quad (34.74)$$

avec:

$$v = \sqrt{\frac{G}{\rho}} \quad (34.75)$$

Les ondes transversales ne se propagent que dans les solides et de ce fait, nous ne pouvons pas les entendre à moins de les transformer en ondes longitudinales par des moyens mécaniques ou électriques. Les ondes transversales peuvent se transmettre le long d'une barre ou d'une tige quelconque ou même d'un fil métallique, et ceci sans besoin que ce dernier soit sous tension. Même si le fil métallique est sous tension, la vitesse des ondes de cisaillement ne dépend pas de la tension. C'est le module de cisaillement élevé de l'acier qui donne aux guitares électriques ce bruit caractéristique.

Un autre cas remarquable d'ondes transversales (de cisaillement) est celui des ondes sismiques. On y trouve des ondes sismiques de cisaillement et aussi des ondes longitudinales ou de pression. Les ondes de cisaillement se propagent dans la croûte terrestre à $3'600 \left[m \cdot s^{-1} \right]$ et les ondes de pression à $6'000 \left[m \cdot s^{-1} \right]$. Lors d'un séisme ou d'une explosion atomique, les deux types d'onde seront produits, mais comme les ondes se propagent à des vitesses différentes, elles n'arriveront pas en même temps à des stations de détection lointaines. C'est à partir de cette différence des temps d'arrivée que l'on détermine la distance à l'épicentre. La direction est obtenue à partir de la direction des oscillations. Seules les stations suffisamment éloignées pour recevoir les deux types d'onde séparément peuvent faire la détermination de l'épicentre.

Pour résumer, nous avons pour les ondes longitudinales dans un solide (cf. chapitre de Musique Mathématique):

$$\frac{\partial^2 \xi}{\partial t^2} = \frac{E}{\rho} \frac{\partial^2 \xi}{\partial x^2} \quad (34.76)$$

et pour les ondes transversales:

$$\frac{\partial^2 y}{\partial t^2} = \frac{G}{\rho} \frac{\partial^2 y}{\partial x^2} \quad (34.77)$$

Pour les détails des développements mathématiques concernant les gaz et les solides, le lecteur devra se rendre dans le chapitre de Musique Mathématique (Acoustique).

2. LIQUIDES

Les fluides usuels sont de deux types: les liquides et les gaz (les solides sont aussi parfois considérés comme des fluides... ce n'est qu'une question d'opinion...). Étymologiquement, un fluide est susceptible de s'écouler. Le liquide adopte la forme du récipient qui le contient tout en conservant un volume propre à peu près invariable. Le gaz n'a pas de volume propre: il envahit uniformément (mécanique statistique de Boltzmann) le récipient dans lequel il est maintenu. Une atmosphère en constitue un cas spécial, du fait qu'elle est maintenue par la gravité à la périphérie d'un astre, ce qui exclut l'uniformité de la densité ou pression.

La distinction entre liquide et gaz est subtile. Nous pouvons cependant dire que le volume propre des liquides manifeste l'existence d'une cohésion liée à une densité assez grande (liaisons de Van der Waals); cette cohésion disparaît avec le volume propre chez les gaz.

Si nous comparons les fluides avec les solides, la première remarque qui s'impose concerne l'isotropie (les propriétés sont les mêmes dans toutes les directions spatiales) des fluides usuels qui est toujours réalisée (si nous n'agissons pas sur le fluide en tout cas!).

Nous allons aborder la théorie de la mécanique des fluides en difficulté croissante et par redondance. D'abord, il va être démontré que les propriétés d'un fluide statique sont isotropes (théorème de Pascal). A l'aide de ce résultat, il va être plus simple de comprendre le théorème de Bernoulli qui va nous permettre, entre autres, de définir le concept de "pression hydrostatique". Ensuite, nous construirons un modèle très important de la dynamique des fluides, connu sous le nom de "équations de Navier-Stokes", que l'on retrouve dans tous les domaines possibles (astrophysique, mécanique quantique, météorologie,...). Ce modèle de dynamique des fluides est conséquent en développements théoriques et résultats expérimentaux et peut être considéré comme un terrain difficile. Cependant, pour faciliter la lecture, nous avons choisi de ne pas aborder celui-ci directement par usage du calcul tensoriel. Nous avons ainsi fait en sorte que les variables tensorielles apparaissent d'elles-mêmes d'écoulant des résultats simples de l'analyse vectorielle que nous obtiendrons. Une fois les équations de Navier-Stokes déterminées et démontrées, nous verrons que nous pouvons retrouver l'expression du théorème de Bernoulli à partir de ces mêmes équations.

La dynamique des fluides, ou "hydrodynamique", est de loin, le domaine de la mécanique classique le moins aisé en ce qui concerne la description et la prédiction. C'est pourquoi le théorème de Bernoulli s'utilise fréquemment, non pour expliquer en détail le comportement d'un fluide, mais pour en faire une description qualitative.

2.1. THÉORÈME DE PASCAL

Le résultat qui va suivre est de la plus haute importance pour comprendre l'ensemble de la mécanique des fluides. Il faut prendre le temps de comprendre!

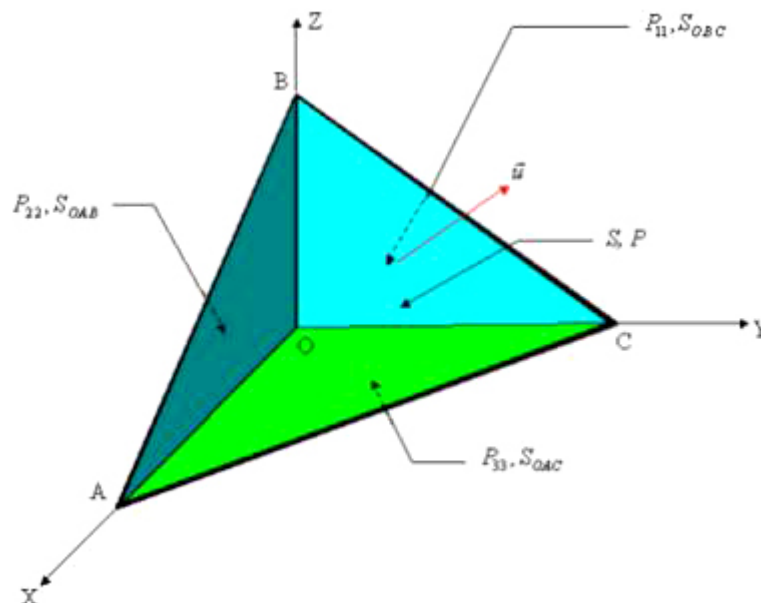


Figure: 34.17 - Tétraèdre régulier élémentaire pour faire émerger le théorème de Pascal

Si nous considérons les forces s'exerçant, en l'absence de mouvement, sur un tétraèdre élémentaire OABC de volume élémentaire V , il est toujours possible d'adopter un volume suffisamment petit pour avoir une pression uniforme s'exerçant sur les faces du tétraèdre.

Soient $P_{11}, P_{22}, P_{33}, P$, les pressions de réaction du fluide dues aux contraintes extérieures sollicitant les faces respectives OBC , OAB , OAC et ABC de surface $S_{OBC}, S_{OAB}, S_{OAC}, S$. Soient également les cosinus directeurs (cf. chapitre de Calcul Vectoriel) du vecteur unitaire $\vec{u}$ normal à la surface ABC .

Le système étant en équilibre, la résultante $\vec{R}$ des forces de réaction du système est nulle. Nous avons donc les équations suivantes résultant de la projection suivant les trois axes de coordonnées:

$$\begin{aligned} R_x &= P_{11} \cdot S_{OBC} - P \cdot S_{OBC} = 0 \\ R_y &= P_{22} \cdot S_{OAB} - P \cdot S_{OAB} = 0 \\ R_z &= P_{33} \cdot S_{OAC} - P \cdot S_{OAC} = 0 \end{aligned} \quad (34.78)$$

Par simplification élémentaire, il vient:

$$\begin{aligned} P - P_{11} &= 0 \\ P - P_{22} &= 0 \\ P - P_{33} &= 0 \end{aligned} \quad (34.79)$$

Nous obtenons alors la relation suivante:

$$P = p_{11} = p_{22} = p_{33} \quad (34.80)$$

Conclusion importante: en un point quelconque d'un fluide, la pression est indépendante de la direction de la normale $\vec{n}$ à la surface élémentaire sur laquelle elle s'exerce.

Par le principe de l'action et de réaction de Newton, nous sommes amenés à énoncer le "**théorème de Pascal**":

Les fluides incompressibles transmettent intégralement et dans toutes les directions, les pressions qui leur sont appliquées.

Ce théorème est fondamental aussi bien en mécanique des fluides qu'en mécanique des gaz et les implications pratiques sont énormes (ce théorème explique entre autres, que la pression est indépendante de la géométrie du contenant du liquide)!

2.2. VISCOSITÉ

En mécanique des fluides, il est utile de considérer plusieurs types fluides ayant des caractéristiques qui les différencient. Ceci s'avère particulièrement pratique pour les simulations tout en restant conforme à l'observation expérimentale (cf. chapitre de Génie Météo Et Marin).

Nous définissons la "**viscosité**" μ par les forces internes s'opposant au déplacement des diverses couches composant le fluide. Nous distinguons la "**viscosité dynamique**" μ_d et la "**viscosité cinématique**" μ_c .

1. La viscosité dynamique:

$$\mu_d = \frac{dF}{\partial_x v dS} \quad (34.81)$$

avec μ_d étant le coefficient de viscosité dynamique (l'unité étant le Poiseuille [PI]), dF variation de la force de frottement entre deux couches infiniment voisines, $\partial_x v$ variation de la vitesse par la distance entre deux couches infiniment voisines [s^{-1}] et dS étant la surface considérée [m^2].

Conclusion: le Poiseuille est la viscosité d'un fluide nécessitant 1 Newton pour faire glisser à la vitesse de 1 mètre par seconde, deux couches fluides de 1 mètre carré distantes de 1 mètre.

Remarque: Anciennement, l'unité employée était la "**poise**": $1 [P] = 10^{-1} [PI]$

2. La viscosité cinématique est définie par:

$$\mu_c = \frac{\mu_d}{\rho} \quad (34.82)$$

Une transformation de la définition de la viscosité dynamique donne (il faut se rappeler de cette relation pour plus tard !!):

$$\tau = \frac{dF}{dS} = \mu_d \partial_x v \quad (34.83)$$

Soit:

$$\partial_x v = \frac{1}{\mu_d} \tau \quad (34.84)$$

Par définition les fluides ayant les caractéristiques suivantes:

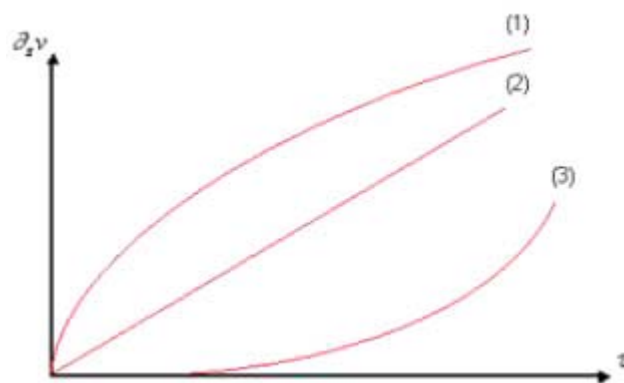


Figure: 34.18 - Caractéristiques de viscosité de différents fluides

sont nommés respectivement:

- (1) Fluides pseudo-plastiques
- (2) Fluides newtoniens (contraintes de cisaillement proportionnelles au gradient de vitesse)
- (3) Fluides dilatants

Il existe encore 3 autres types de fluides non représentés sur la figure et dont la viscosité est supposée nulle (cf. chapitre de Thermodynamique):

- (4) Fluides parfaits
- (5) Fluides semi-parfaits
- (6) Fluides réels

Remarques:

R1. Le comportement d'un fluide parfait est très différent de celui d'un fluide réel aussi petite soit la viscosité de ce dernier. En effet, le fluide parfait, parce qu'il n'a pas de viscosité, ne dissipe jamais l'énergie cinétique. Alors qu'un fluide réel très peu visqueux la dissipe efficacement grâce à la turbulence, et au phénomène de cascade qui l'accompagne.

R2. Nous reviendrons sur les propriétés de la viscosité dynamique et cinématique lors de la démonstration des équations de Navier-Stokes-(Reynolds).

R3. Les fluides qui ne sont pas newtoniens sont appelés en toute généralité dans la littérature "**fluides non-newtoniens**"... et nous ne traiterons pas les ferrofluides ici car trop complexes théoriquement à analyser.

Les fluides non-newtoniens ont donc une déformation qui dépend de la force que nous leur appliquons. Le meilleur exemple est celui du sable mouillé en bord de mer: quand nous frappons le sable, il a la viscosité élevée d'un solide, alors que lorsque nous appuyons doucement dessus, il se comporte comme une pâte. Par ailleurs, certains fluides non-newtoniens ont des propriétés telles qu'il est possible pour un individu de courir dessus sans couler ou de couler en restant en position...

2.2.1. LOI DE POISEUILLE

En 1835 un médecin français, Jean Léonard Marie Poiseuille fit une série d'expériences soignées, pour déterminer comment un fluide visqueux s'écoule dans un tuyau étroit. Son but était de comprendre la dynamique de la circulation sanguine chez l'homme. Le plasma du sang se comporte comme un fluide newtonien, tandis que le sang entier ne l'est pas. Presque la moitié du volume normal du sang est faite de cellules assez grandes pour perturber l'écoulement laminaire, surtout quand elles entrent en contact avec les parois des vaisseaux, un phénomène qui prend de l'importance dans les capillaires très étroits. Néanmoins, l'analyse de Poiseuille s'applique à l'écoulement dans les veines et les plus grosses artères et elle a une grande valeur, bien qu'elle soit un peu simpliste.

Le résultat de Poiseuille peut être établi en considérant le fluide dans un tuyau comme formé de couches cylindriques orientées selon un axe x de rayon r concentriques qui se déplacent à des vitesses qui vont en décroissant à partir du centre (symétrie circulaire supposée).

Alors la relation définissant la viscosité s'écrit:

$$F = \mu_a S \frac{dv_x}{dr} \quad (34.85)$$

Ce qui nous donne la force de viscosité sur le cylindre. La surface de contact de chaque couche cylindrique de longueur l est donnée par $S = 2\pi r l$ et donc:

$$F = \mu_a 2\pi r l \frac{dv_x}{dr} \quad (34.86)$$

L'origine de l'accélération (in extenso de la force) ne peut se faire que par une différence de pression telle que:

$$\Delta P \cdot S = \Delta P \cdot \pi r^2 = F \quad (34.87)$$

ce qui nous amène à écrire:

$$\Delta P \pi r^2 = \mu_a 2\pi r l \frac{dv_x}{dr} \Rightarrow dv_x = \frac{\Delta P}{2\mu_a l} r dr \quad (34.88)$$

En intégrant membre à membre, nous obtenons:

$$\int_0^{v_x} dv_x = \frac{\Delta P}{2\mu_a l} \int_r^R r dr \quad (34.89)$$

Soit:

$$v_x = \frac{\Delta P}{4\mu_a l} (R^2 - r^2) \quad (34.90)$$

La courbe représentative de la vitesse en fonction de r est une parabole dont le sommet se situe sur l'axe centre du cylindre ($r = 0$). Le débit volumique transporté par une couche cylindrique entre r et $r + dr$ est $dJ = v_x dS = v_x (2\pi r dr)$. Ainsi, le débit total est:

$$J = \int_0^R \frac{\Delta P}{4\mu_a l} (R^2 - r^2) 2\pi r dr \quad (34.91)$$

et nous obtenons la "**loi de Poiseuille**" pour le débit laminaire visqueux:

$$J = \frac{\pi R^4}{8\mu_a} \frac{\Delta P}{l} \quad (34.92)$$

Nous trouvons donc le résultat logique que le débit augmente avec le gradient de pression $\Delta P/l$ et le rayon du tube, et diminue avec la viscosité.

Nous trouvons par ailleurs une relation analogue à la loi d'Ohm (cf. chapitre d'Électrocinétique) où la différence de potentiel est donnée par la résistance multipliée par le courant alors que la différence de pression est donnée par la résistance visqueuse multipliée par le débit.

2.3. THÉORÈME DE BERNOULLI

Quand nous discutons du mouvement d'un fluide, l'équation de continuité (cf. chapitre Thermodynamique), qui exprime la conservation de la masse (volumique) du fluide est une notion importante.

$$\frac{\partial \rho}{\partial t} - \vec{\nabla} \circ (\rho \cdot \vec{v}_a) = 0 \quad (34.93)$$

Considérons cette équation dans le cas particulier qui nous intéresse ici un fluide non visqueux en écoulement laminaire se déplaçant à l'intérieur d'un tube de lignes de courants parallèles (le mouvement du fluide est de type irrotationnel - voir chapitre de Calcul Vectoriel), délimité par la surface A_1 :

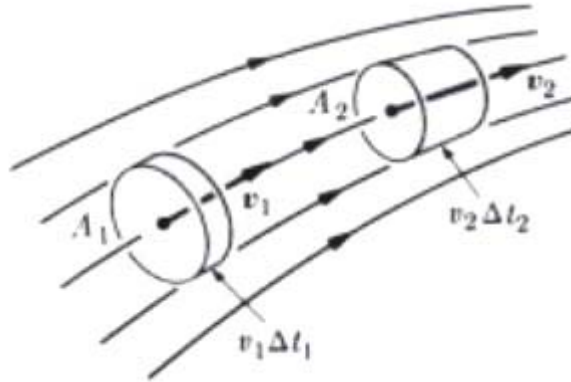


Figure: 34.19 - Fluide non visqueux en écoulement laminaire d'un tube de lignes de courants parallèles

Nous sommes en régime stationnaire (l'aspect du mouvement est indépendant du temps) et la masse n'est ni apportée par une source ni enlevée par un puits à l'intérieur de la région considérée. Le volume de fluide qui traverse A_1 dans l'intervalle Δt correspond à un cylindre de base A_1 , de longueur $v_1 \cdot \Delta t$ et donc de volume $A_1 \cdot v_1 \cdot \Delta t$. La masse de fluide qui a traversé A_1 pendant le temps Δt est donc:

$$\rho_1 \cdot A_1 \cdot v_1 \cdot \Delta t \quad (34.94)$$

De même:

$$\rho_2 \cdot A_2 \cdot v_2 \cdot \Delta t \quad (34.95)$$

est la masse de fluide qui a traversé A_2 pendant le même intervalle de temps. Avec les hypothèses faites, l'équation de conservation de la masse exige que les deux masses soient les mêmes, ou exprimé autrement:

$$\rho_1 \cdot A_1 \cdot v_1 \cdot \Delta t = \rho_2 \cdot A_2 \cdot v_2 \cdot \Delta t \quad (34.96)$$

D'où:

$$\rho_1 \cdot A_1 \cdot v_1 = \rho_2 \cdot A_2 \cdot v_2 \quad (34.97)$$

Ceci est la forme de l'équation de continuité dans le contexte qui nous intéresse. De plus, si le fluide est incompressible, la densité est partout la même et l'équation précédente se réduit à:

$$A_1 \cdot v_1 = A_2 \cdot v_2 \quad (34.98)$$

Considérons maintenant une région dans un fluide où il y a un flux stationnaire comme l'indique la figure ci-dessous:

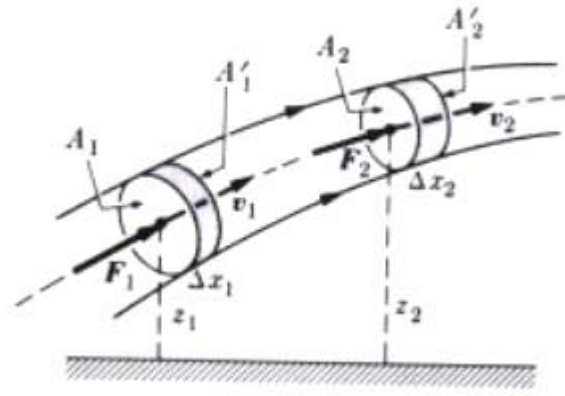


Figure: 34.20 - Agrandissement sur une région du fluide

Pendant un court intervalle de temps Δt , le fluide qui, initialement, traversait A_1 a progressé jusqu'à une surface A'_1 à la distance $\Delta x_1 = v_1 \cdot \Delta t$ tandis que le fluide qui traversait A_2 se retrouve en A'_2 à une distance Δx_2 . Puisque le reste du volume entre les surfaces A_1 et A_2 reste inchangé, nous allons porter notre attention sur les deux volumes (égaux) hachurés sur la figure.

Ces deux volumes sont égaux, car le fluide est incompressible et l'équation de continuité est valable. Soient F_1 et F_2 les forces exercées sur les surfaces A_1 et A_2 en raison de la pression existant dans le fluide. À cause de ces forces, le fluide produit ou reçoit du travail en déplaçant les deux volumes. En A_1 , la surface est poussée par le fluide et le travail exercé sur le fluide est $F_1 \cdot \Delta x$ alors qu'en A_2 le fluide pousse la surface et le travail effectué par le fluide est $F_2 \cdot \Delta x$. Le travail total exercé sur le volume de fluide situé entre A_1 et A_2 est donc:

$$W = F_1 \cdot \Delta x_1 - F_2 \cdot \Delta x_2 = P_1 \cdot A_1 \cdot \Delta x_1 - P_2 \cdot A_2 \cdot \Delta x_2 \quad (34.99)$$

en appelant P_1 et P_2 les pressions respectives en A_1 et A_2 et en écrivant:

$$F = P \cdot A \quad (34.100)$$

d'après la définition de la pression. Comme:

$$A_1 \cdot \Delta x_1 = A_2 \cdot \Delta x_2 \quad (34.101)$$

d'après l'équation de continuité et l'hypothèse d'incompressibilité, nous pouvons écrire que:

$$W = (P_1 - P_2) \cdot A_1 \cdot \Delta x_1 \quad (34.102)$$

Le travail extérieur exercé sur le système change son énergie propre comme l'établit la thermodynamique ($W = \Delta U$). Pour le volume de fluide considéré, l'énergie propre des volumes mise en évidence comprend l'énergie cinétique et l'énergie potentielle de gravitation. Le fluide entre A_1 et A_2 gagne de l'énergie dans le volume $A_2 \cdot \Delta x_2$. Supposons que les deux volumes aient une masse égale m , de nouveau à cause de l'équation de continuité. Alors le gain net d'énergie est:

$$\Delta U = U_2 - U_1 = \left(\frac{1}{2} \cdot m \cdot v_2^2 + m \cdot g \cdot z_2 \right) - \left(\frac{1}{2} \cdot m \cdot v_1^2 + m \cdot g \cdot z_1 \right) \quad (34.103)$$

Puisque nous avons déjà supposé le fluide incompressible, la densité ρ est la même partout et m peut être remplacé par $\rho \cdot A_1 \cdot \Delta x_1$ aux deux extrémités. D'où:

$$\Delta U = \left[\left(\frac{1}{2} \cdot \rho \cdot v_2^2 + \rho \cdot g \cdot z_2 \right) - \left(\frac{1}{2} \cdot \rho \cdot v_1^2 + \rho \cdot g \cdot z_1 \right) \right] \cdot A_1 \cdot \Delta x_1 \quad (34.104)$$

En combinant cette relation avec $W = (P_1 - P_2) \cdot A_1 \cdot \Delta x_1$ nous obtenons:

$$P_1 - P_2 = \left(\frac{1}{2} \cdot \rho \cdot v_2^2 + \rho \cdot g \cdot z_2 \right) - \left(\frac{1}{2} \cdot \rho \cdot v_1^2 + \rho \cdot g \cdot z_1 \right) \quad (34.105)$$

ou:

$$\left[\frac{1}{2} \cdot \rho \cdot v_1^2 + \rho \cdot g \cdot z_1 + P_1 = \frac{1}{2} \cdot \rho \cdot v_2^2 + \rho \cdot g \cdot z_2 + P_2 \right] \quad (34.106)$$

Comme l'équation ci-dessus concerne des grandeurs prises en deux points arbitraires le long d'une ligne de courant, nous pouvons généraliser et écrire:

$$\boxed{\frac{1}{2} \cdot \rho \cdot v^2 + \rho \cdot g \cdot z + P = c^{te}} \quad (34.107)$$

Ce résultat, connu sous le nom de "**théorème de Bernoulli**", exprime la constance de la pression le long d'une ligne de courant dans un fluide incompressible, irrotationnel et non visqueux et où les forces volumiques extérieures dérivent d'une énergie potentielle (nous reviendrons là-dessus après avoir déterminé les équations de Navier-Stokes).

Signalons aussi une manière élégante et simple de retrouver cette relation. La conservation de l'énergie nous donne le long d'une ligne de courant:

$$E_e + E_p + E_P = c^{te} \quad (34.108)$$

avec respectivement et dans l'ordre la somme de l'énergie cinétique, de l'énergie potentielle et de l'énergie de pression. Soit:

$$\frac{1}{2} m v^2 + m g z + F \cdot d = c^{te} \quad (34.109)$$

et si nous divisons tout cela par le volume, nous obtenons alors:

$$\frac{1}{2} \rho v^2 + \rho g z + P = c^{te} \quad (34.110)$$

voilà....

Remarques:

R1. D'une ligne de courant à l'autre, c'est la valeur de la constante qui change. De plus, l'utilisation du théorème de Bernoulli exige de connaître la forme des lignes de courant.

R2. La conservation de la quantité de gauche exprime la conservation de l'énergie le long d'une ligne de courant et nous y trouvons respectivement l'énergie cinétique volumique, l'énergie potentielle volumique de pesanteur et la pression.

Considérons maintenant deux applications importantes du théorème de Bernoulli.

Si le fluide se déplace dans un plan horizontal, l'énergie potentielle de gravitation reste constante et l'équation de Bernoulli se réduit alors à:

$$\frac{1}{2} \cdot \rho \cdot \left(\frac{dx}{dt} \right)^2 = c^{te} \quad (34.111)$$

Donc, dans un tuyau horizontal, la vitesse est d'autant plus grande que la pression est plus faible et réciproquement. Nous utilisons aussi cet effet pour participer à la poussée d'un avion (attention ce paramètre est mineur, car ce n'est pas ce qui contribue le plus au vol d'un avion, c'est l'effet Magnus dont la démonstration sera donnée plus loin un jour...).

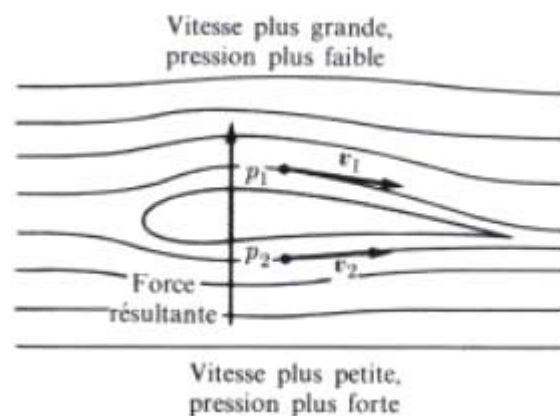


Figure: 34.21 - Illustration du profil d'une aile avec les pressions correspondantes

Le profil d'une aile est construit de telle sorte que l'air ait une vitesse plus grande au-dessus de la surface de l'aile qu'au-dessous, ce qui produit une pression plus forte au-dessous qu'au-dessus. Il en résulte donc une force résultante vers le haut.

Autrement dit, une spécialiste dans l'aérodynamique (pour les avions) ou en hydrodynamique (pour les stabilisateurs de roulis des gros bateaux) dirait:

- À l'extrados: Par effet de courbure, les particules d'air (d'eau) sont contraintes de parcourir une distance plus grande. Leur vitesse va donc d'abord s'accroître fortement pour diminuer ensuite afin de retrouver au bord de fuite la vitesse initiale de l'écoulement. Tout l'extrados est donc le siège d'une dépression locale généralisée. La couche limite, d'abord laminaire, devient peu à peu turbulente, voire tourbillonnaire lorsqu'on s'approche du bord de fuite.

- À l'intrados: le profil constituant un obstacle à l'écoulement, l'air (l'eau) va se trouver freiné: nous voyons donc apparaître une surpression localisée sur l'intrados. En fait, avec la forme des ailes d'avion actuelle, en position horizontale, l'effet Bernoulli serait négligeable. Pour qu'un avion décolle, il faudrait que l'extrados ait une surface beaucoup plus grande.

C'est bien mieux ainsi non ?

Autre chose encore, si le fluide n'est pas en mouvement, nous avons l'équation de Bernoulli qui s'écrit:

$$\rho \cdot g \cdot z + P = c^{te} \quad (34.112)$$

Il s'agit de "**équation de Laplace**" en hydrostatique (utilisée dans les vases communicants).

2.3.1. THÉORÈME DE TORRICELLI

Le théorème de Torricelli permet de déterminer la vitesse d'écoulement d'un liquide. C'est un cas classique d'étude dans les petites écoles.

Considérons un volume fermé contenant un liquide de masse volumique ρ et muni d'un orifice de surface S_2 , duquel le liquide coule vers l'extérieur. Nous voulons déterminer la vitesse v_2 d'écoulement du liquide de cet orifice. Le volume est supposé être assez grand pour que ni le niveau du liquide, ni la pression P au-dessus de sa surface S_1 ne varient de façon appréciable pendant l'écoulement. Comme le tube d'échappement de liquide va de la région de la surface du liquide à l'orifice ouvert à l'air libre, nous avons $P_1 = P$. Un liquide coulant à l'air libre est à la pression atmosphérique, $P_2 = P_A$, car le liquide est entouré d'air libre et rien ne peut maintenir une différence de pression. D'après l'équation de Bernoulli, avec $z_1 - z_2 = h$, nous trouvons sur une ligne de courant:

$$P + \frac{1}{2} \rho v_1^2 + \rho g h = P_A + \frac{1}{2} \rho v_2^2 \quad (34.113)$$

d'où:

$$v_2^2 = v_1^2 + \frac{2(P - P_A)}{\rho} + 2gh \quad (34.114)$$

De l'équation de continuité ($v_1 S_1 = v_2 S_2$), nous déduisons que si $S_1 \gg S_2$ alors $v_2 \gg v_1$ et v_1^2 est alors négligeable devant v_2^2 . Dans le cas particulier, mais fréquent, où le réservoir est ouvert à l'air libre ($P = P_A$), la densité d'énergie de pression disparaît. Le fluide coule sous l'effet de la gravité, sans être poussé par une différence de pression. Nous trouvons alors (en multipliant par la surface de l'orifice, nous obtenons le débit):

$$v_2 = \sqrt{2gh} \quad (34.115)$$

Cette relation constitue le "**théorème de Torricelli**". Chose curieuse, nous avons déjà vu cette relation en mécanique classique pour la vitesse de chute libre d'un corps. Il en retourne l'observation faite par Torricelli: si le jet est dirigé directement vers le haut, il atteint presque le niveau de la surface du liquide dans le volume. La raison pour laquelle le jet n'atteint pas effectivement ce niveau est une certaine perte d'énergie à cause du frottement.

2.3.2. EFFET VENTURI

Certaines applications pratiques de la mécanique des fluides résultent de l'interdépendance de la pression et la vitesse. Il y a une catégorie de situations dans lesquelles la variation d'énergie potentielle gravitationnelle est négligeable. L'équation de Bernoulli relie alors la différence de pression à la différence d'énergie cinétique donc la variation du carré de la vitesse.

Nous considérons un fluide incompressible (!), non visqueux et de masse volumique ρ . Le fluide s'écoule en régime permanent dans une canalisation cylindrique de rayon R_1 et de section S_1 suivie par un tube cylindrique de rayon R_2 et de section S_2 . Le raccordement est fait par une canalisation conique assez longue pour que l'on reste en régime laminaire.

Nous savons (équation de continuité) que:

$$S_1 v_1 = S_2 v_2 \quad (34.116)$$

qui veut dire, comme nous l'avons vu, qu'une diminution de la section traversée par le fluide se traduit par une augmentation de sa vitesse.

Dans toute situation où le flux entrant est environ au même niveau que le rétrécissement $z_1 \cong z_2$, l'équation de Bernoulli s'emploie pour exprimer la différence de pression:

$$P_1 + \frac{1}{2} \rho v_1^2 + \rho g y_1 = P_2 + \frac{1}{2} \rho v_2^2 + \rho g y_2 \quad (34.117)$$

devient:

$$P_1 + \frac{1}{2} \rho v_1^2 = P_2 + \frac{1}{2} \rho v_2^2 \quad (34.118)$$

Utilisant l'équation de continuité, pour éliminer v_1 , nous obtenons:

$$P_1 - P_2 = \frac{1}{2} \rho v_2^2 \left(1 - \frac{S_2^2}{S_1^2} \right) \quad (34.119)$$

Comme $S_1 > S_2$ le second membre de la relation est positif et $P_1 > P_2$: il y a donc une chute de pression dans la région étroite. En arrivant à la région divergente à nouveau en $S_3 = S_2$, la pression du fluide augmente de nouveau et la vitesse reprend sa valeur initiale. Cette diminution de la pression qui accompagne l'augmentation de la vitesse est appelée "effet Bernoulli" ou "effet Venturi".

Ainsi, la vitesse du fluide augmente dans un goulot d'étranglement pour satisfaire l'équation de continuité (conservation du flux/masse) et le fait qu'il soit incompressible (sinon il y aurait une sorte de bouchon...).

Remarque: Paradoxalement l'effet Venturi se produit aussi lors du franchissement d'un sommet ou d'une crête par l'air atmosphérique ou également dans les rues des villes. En effet, l'air qui arrive sur la montagne ou la crête a tendance à "s'écraser" dessus. La section d'écoulement de l'air au sommet est donc plus faible qu'à la base. Il se produit donc également un effet Venturi: la vitesse du vent est plus élevée sur les sommets et les crêtes qu'en bas (les professionnels du planeur en savent quelque chose...).

2.3.3. TUBE DE PITOT

Le tube de Pitot permet la mesure de la vitesse d'écoulement d'un gaz subsonique. Le tube de Pitot consiste à pratiquer dans un tube, un orifice de prise de pression en A et en B:

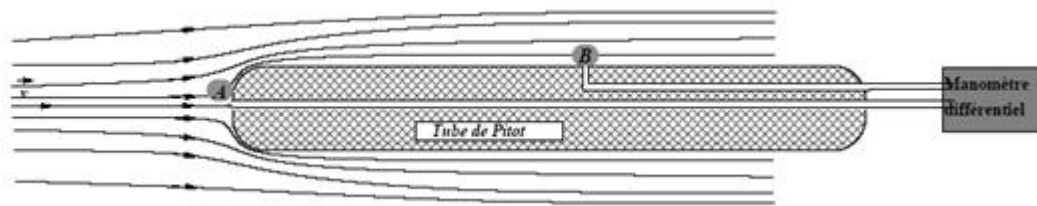


Figure: 34.22 - Tube de Pitot

Le point A est un point d'arrêt, car la vitesse y est nulle (il n'y a pas d'écoulement dans l'orifice, c'est juste une prise de pression). Loin de l'obstacle (le tube de Pitot) l'écoulement est supposé uniforme de vitesse v et de pression P_B .

En A (point d'arrêt), en utilisant la relation de Bernoulli le long de la ligne de courant et en considérant la variation de hauteur entre A et B négligeable, la pression vaut:

$$P_A = P_B + \frac{1}{2} \rho v^2 \quad (34.120)$$

Nous avons donc:

$$v = \sqrt{\frac{2(P_A - P_B)}{\rho}} \quad (34.121)$$

Donc pour les avions à partir de la différence d'une mesure de pression et de la connaissance de la densité du gaz, il est possible de connaître la vitesse!

Remarque: En aéronautique, la pression dynamique s'ajoute à la pression statique pour donner la pression totale qui peut être mesurée au point de vitesse nulle du tube Pitot. En enlevant la pression statique, on trouve la "**pression dynamique**".

2.3.4. PERTE DE CHARGE (PRESSION)

Lorsque, dans un écoulement d'un fluide parfait, il n'y a aucune machine (ni pompe ni turbine) entre les points (1) et (2) d'une même ligne de courant, la relation de Bernoulli peut s'écrire sous la forme suivante :

$$\frac{1}{2}\rho(v_2^2 - v_1^2) + \rho g(z_2 - z_1) + (P_2 - P_1) = 0 \quad (34.122)$$

Lorsque le fluide traverse une machine hydraulique, il échange de l'énergie avec cette machine sous forme de travail pendant une durée donnée. La puissance P échangée est alors (cf. chapitre de Mécanique Classique):

$$P = \frac{\Delta W}{\Delta t} \quad (34.123)$$

où par convention, si $P > 0$ l'énergie est reçue par le fluide (pompe) sinon, si $P < 0$ l'énergie est fournie par le fluide (turbine).

Si le débit-volume est Q_V , la relation de Bernoulli s'écrit alors logiquement:

$$\frac{1}{2}\rho(v_2^2 - v_1^2) + \rho g(z_2 - z_1) + (P_2 - P_1) = \frac{P}{Q_V} \quad (34.124)$$

où:

$$w = \frac{P}{Q_V} \quad [J \cdot m^{-3}] \quad (34.125)$$

Un fluide parfait n'existe pas. Lors d'un écoulement dans une conduite, les forces de frottement dissipent une partie de l'énergie cinétique et potentielle ce qui se traduit par l'existence de pertes de charges dont il s'agit de tenir compte.

Considérons un écoulement cylindrique horizontal stationnaire et incompressible. Si nous appliquons la relation de Bernoulli entre l'entrée et la sortie, nous obtenons:

$$P_1 = P_2 \quad (34.126)$$

Or, expérimentalement, nous observons qu'il faut imposer une pression plus importante en entrée pour entretenir le régime permanent. En effet, les forces de viscosité résistent à l'écoulement. Il faut donc imposer une suppression ΔP_f que nous appelons "perte de charge en pression" et qui est due à l'existence de forces de frottements (viscosité) ou de pertes singulières (géométrie des circuits de distribution).

L'équation de Bernoulli généralisée s'écrit alors dans ce cas d'étude qui fait partie de l'ingénierie des procédés:

$$\frac{1}{2}\rho v_1^2 + \rho g z_1 + P_1 + \frac{P}{Q_V} = \frac{1}{2}\rho v_2^2 + \rho g z_2 + P_2 + \Delta P_f \quad (34.127)$$

Cette relation est souvent utilisée dans l'étude théorique (...) des problèmes de conduite.

2.4. ÉQUATIONS DE NAVIER-STOKES

Soit un parallélépipède élémentaire extrait d'un fluide statique à l'équilibre de dimensions dx , dy , dz représenté à la figure ci-dessous. La matière à l'équilibre composant le parallélépipède est en général soumise à des forces de volume dans toutes les directions (théorème de Pascal) dont les composantes sur les trois axes orthogonaux sont représentées sur la figure ci-dessous (ces forces peuvent être de nature gravitationnelles, électromagnétiques ou inertielles...).

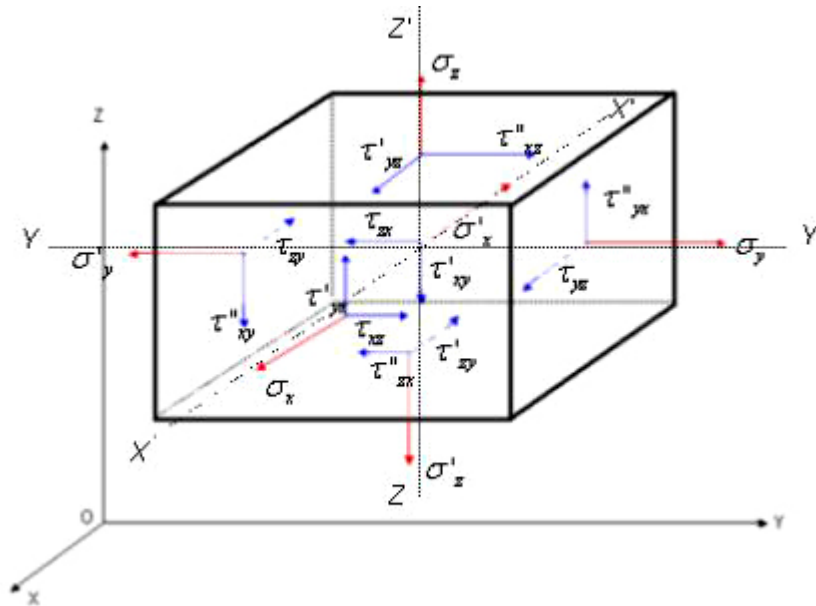


Figure: 34.23 - Parallélépipède élémentaire extrait d'un fluide statique à l'équilibre

Remarques:

R1. Il est important de remarquer que les composantes de tous les vecteurs visibles sur la figure ci-dessus sont exprimées en newton par unité de surface, soit en d'autres termes par unité de pression (qui est l'unité de la contrainte pour rappel...).

R2. Il est important d'être attentif au plus haut point à ce qui va suivre car certains des résultats que nous obtiendrons ici seront réutilisés dans le chapitre de Relativité Générale pour comprendre le tenseur d'énergie-impulsion!!

Nous pouvons, comme nous l'avons représenté ci-dessus, décomposer et traduire l'ensemble des forces auxquelles est soumis le parallélépipède aux centres des faces de ce dernier. Nous représentons bien évidemment chacune des contraintes sur chacune des faces comme la somme des contraintes normales et tangentielles telles que nous l'avons fait pour l'étude des solides sous contrainte (selon les trois axes toujours, d'où la somme de trois composantes!).

Au total, nous nous retrouvons avec 18 composantes de contraintes normales et tangentielles:

$$\begin{aligned}
\bar{\sigma}_x + \bar{\tau}'_{yx} + \bar{\tau}_{xz} &= \frac{d\bar{F}_{+x}}{dS}, \bar{\sigma}'_x + \bar{\tau}'_{xy} + \bar{\tau}_{zx} = \frac{d\bar{F}_{-x}}{dS} \\
\bar{\sigma}_y + \bar{\tau}_{yz} + \bar{\tau}''_{yx} &= \frac{d\bar{F}_{+y}}{dS}, \bar{\sigma}'_y + \bar{\tau}_{xy} + \bar{\tau}''_{xy} = \frac{d\bar{F}_{-y}}{dS} \\
\bar{\sigma}_z + \bar{\tau}''_{xz} + \bar{\tau}'_{yz} &= \frac{d\bar{F}_{+z}}{dS}, \bar{\sigma}'_z + \bar{\tau}''_{zx} + \bar{\tau}'_{zy} = \frac{d\bar{F}_{-z}}{dS}
\end{aligned} \quad (34.128)$$

Nous cherchons à minimiser le nombre de composantes normales afin de déterminer quelles sont les contraintes suffisantes sur chacun des axes. Ainsi, nous poserons:

$$\begin{aligned}
\bar{\sigma}_x + \bar{\sigma}'_x &= \bar{\sigma}_X \\
\bar{\sigma}_y + \bar{\sigma}'_y &= \bar{\sigma}_Y \\
\bar{\sigma}_z + \bar{\sigma}'_z &= \bar{\sigma}_Z
\end{aligned} \quad (34.129)$$

Donc trois composantes suffisent pour connaître les forces de contraintes normales aux surfaces selon chaque axe.

Si nous effectuons la somme des moments de forces par rapport aux centres de gravité pour chaque axe de symétrie du parallélépipède (XX',YY',ZZ') il est évident que sur les 12 composantes tangentielle, 6 suffisent pour décrire l'ensemble du système.

Ainsi pour le plan XOY passant par le centre de gravité nous avons:

$$\begin{aligned}
& \left(\bar{\tau}_{xy} dz dx \right) \cdot \frac{dy}{2} + \left(\bar{\tau}_{xz} dy dz \right) \cdot \frac{dx}{2} + \left(\bar{\tau}_{yx} dx dz \right) \cdot \frac{dy}{2} + \left(\bar{\tau}_{zx} dz dy \right) \cdot \frac{dx}{2} \\
&= \frac{1}{2} dV \left[(\bar{\tau}_{xy} + \bar{\tau}_{yx}) + (\bar{\tau}_{xz} + \bar{\tau}_{zx}) \right] = \frac{1}{2} dV \left[\bar{\tau}_{zy} + \bar{\tau}_{zx} \right] = \frac{1}{2} dV \bar{\tau}_{xoy}
\end{aligned} \quad (34.130)$$

Pour le plan XOZ:

$$\begin{aligned}
& \left(\bar{\tau}'_{yx} dy dz \right) \cdot \frac{dx}{2} + \left(\bar{\tau}'_{yz} dx dy \right) \cdot \frac{dz}{2} + \left(\bar{\tau}'_{xy} dz dy \right) \cdot \frac{dx}{2} + \left(\bar{\tau}'_{zy} dy dx \right) \cdot \frac{dz}{2} \\
&= \frac{1}{2} dV \left[(\bar{\tau}'_{yx} + \bar{\tau}'_{xy}) + (\bar{\tau}'_{yz} + \bar{\tau}'_{zy}) \right] = \frac{1}{2} dV \left[\bar{\tau}_{yx} + \bar{\tau}_{yz} \right] = \frac{1}{2} dV \bar{\tau}_{xoz}
\end{aligned} \quad (34.131)$$

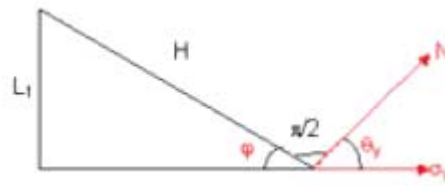
Pour le plan ZOY:

$$\begin{aligned}
& \left(\bar{\tau}''_{xz} dx dy \right) \cdot \frac{dz}{2} + \left(\bar{\tau}''_{yx} dx dz \right) \cdot \frac{dy}{2} + \left(\bar{\tau}''_{zx} dy dx \right) \cdot \frac{dz}{2} + \left(\bar{\tau}''_{xy} dz dx \right) \cdot \frac{dy}{2} \\
&= \frac{1}{2} dV \left[(\bar{\tau}''_{xz} + \bar{\tau}''_{zx}) + (\bar{\tau}''_{yx} + \bar{\tau}''_{xy}) \right] = \frac{1}{2} dV \left[\bar{\tau}_{zx} + \bar{\tau}_{xy} \right] = \frac{1}{2} dV \bar{\tau}_{zoy}
\end{aligned} \quad (34.132)$$

Donc pour chaque plan (XOY, ZOY, ZOX), une composante suffit pour décrire l'ensemble de moments de forces.

Ainsi, par souci de simplification d'écriture, nous poserons (il est plus conforme de faire les développements avec des indices en minuscules):

$$\varphi = \pi - \pi/2 - \theta_y = \pi/2 - \theta_y \quad (34.136)$$



et donc:

$$\sin(\varphi) = \frac{L_1}{H} = \sin(\pi/2 - \theta_y) = \cos(\theta_y) \quad (34.137)$$

Finalement:

$$S_y = \frac{H \cos \theta_y \cdot L_1}{2} \quad (34.138)$$

Le rapport:

$$\frac{S_y}{S} = \frac{\frac{H \cos \theta_y \cdot L_1}{2}}{H \cdot L_1 \frac{\sqrt{2}}{2}} = \frac{\cos \theta_y}{\sqrt{2}} \quad (34.139)$$

d'où:

$$S_y = \frac{S}{\sqrt{2}} \cos \theta_y \quad (34.140)$$

Le principe d'analyse étant le même pour toutes les autres surfaces telles que:

$$S_x = \frac{S}{\sqrt{2}} \cos \theta_x, S_y = \frac{S}{\sqrt{2}} \cos \theta_y, S_z = \frac{S}{\sqrt{2}} \cos \theta_z \quad (34.141)$$

Nous écrirons donc:

$$\frac{\cos \theta_x}{\sqrt{2}} = l, \frac{\cos \theta_y}{\sqrt{2}} = m, \frac{\cos \theta_z}{\sqrt{2}} = n \quad (34.142)$$

tel que:

$$S_x = S \cdot l, S_y = S \cdot m, S_z = S \cdot n \quad (34.143)$$

Remarque: Nous pouvons facilement connaître les valeurs des θ_i à l'aide de l'analyse vectorielle. Effectivement, le plan ABC étant d'équation:

$$ax + by + cz + d = 0, a = b = c = L_1 \quad (34.144)$$

en simplifiant par L_1 :

$$1 \cdot x + 1 \cdot y + 1 \cdot z + d' = 0 \quad (34.145)$$

Le vecteur normal au plan étant bien:

$$\vec{N} = (1, 1, 1) \quad (34.146)$$

pour connaître les cosinus de l'angle du vecteur normal avec les $\vec{\sigma}_i$, il suffit d'assimiler ces derniers aux vecteurs de base $\vec{i}, \vec{j}, \vec{k}$ tel que (trigonométrie élémentaire):

$$\cos \theta_i = \frac{\|\vec{i}\|}{\|\vec{N}\|} = \frac{\|(1, 0, 0)\|}{\|(1, 1, 1)\|} = \frac{1}{\sqrt{3}} \quad (34.147)$$

et en procédant de même pour tous les autres θ .

L'équilibre des forces nous donne:

$$N_x \cdot S - \sigma_x S \cdot l - \tau_{xy} S \cdot m - \tau_{xz} S \cdot n = 0 \quad (34.148)$$

Après simplification:

$$N_x = \sigma_x l + \tau_{xy} m + \tau_{xz} n \quad (34.149)$$

Suivant les autres axes:

$$\begin{aligned} N_y &= \tau_{yx} l + \sigma_y m + \tau_{yz} n \\ N_z &= \tau_{zx} l + \tau_{zy} m + \sigma_z n \end{aligned} \quad (34.150)$$

Soit en résumé:

$$\begin{aligned} N_x &= \sigma_x l + \tau_{xy} m + \tau_{xz} n \\ N_y &= \tau_{yx} l + \sigma_y m + \tau_{yz} n \\ N_z &= \tau_{zx} l + \tau_{zy} m + \sigma_z n \end{aligned} \quad (34.151)$$

En utilisant la représentation matricielle, nous obtenons:

$$\begin{bmatrix} N_x \\ N_y \\ N_z \end{bmatrix} = \begin{bmatrix} \sigma_{xx} & \tau_{xy} & \tau_{xz} \\ \tau_{yx} & \sigma_{yy} & \tau_{yz} \\ \tau_{zx} & \tau_{zy} & \sigma_{zz} \end{bmatrix} \begin{bmatrix} l \\ m \\ n \end{bmatrix} \quad (34.152)$$

Soit en notation indicielle les contraintes normales et tangentielles sont données par la relation traditionnelle suivante (où nous ne distinguons plus ce qui est tangentiel de ce qui est normal donc il y a une perte de clarté):

$$N_i = \sigma_{ij} n_j = \sum_j \sigma_{ij} n_j \quad (34.153) .$$

Nous voyons apparaître une grandeur mathématique σ_{ij} ayant 9 composantes, alors qu'un vecteur dans le même espace $\mathbb{R}^3$ en possède 3. Nous connaissons ce genre d'être mathématique que nous avons déjà étudié en algèbre dans le chapitre de Calcul Tensoriel. La grandeur σ_{ij} est appelée "**tenseur des contraintes du second ordre**". En outre, certaines composantes peuvent être égales ($\sigma_{ij} = \sigma_{ji}$, si $i \neq j$), ce qui le rendrait symétrique. Il ne possède alors plus que les 6 composantes distinctes, relativement aux nombres de composantes suffisantes pour décrire totalement un système à l'équilibre.

Pour étudier les déformations d'un milieu continu tel qu'un fluide, nous considérerons d'abord le cas de très faibles déformations. Les petits déplacements $\vec{D}$ d'un point seront représentés par u, v, w parallèles aux axes d'un référentiel $OXYZ$. Nous admettons que ces composantes sont des quantités très faibles variant d'une façon continue dans le volume du corps considéré.

Soit un segment linéaire OP situé dans un solide avant déformation. Dans un référentiel $OXYZ$, nous noterons x_0, y_0, z_0 et x_p, y_p, z_p les coordonnées de O et P .

Pendant la déformation, la ligne OP devient $O'P'$ tel que représenté ci-dessous:

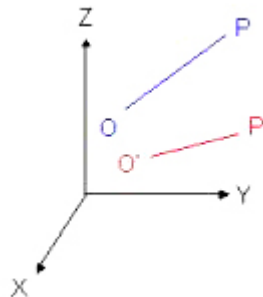


Figure: 34.25 - Segment linéaire dans un solide avant et après déformation

Soient u_0, v_0, w_0 les déplacements du point O parallèlement aux axes OX, OY, OZ et

u_p, v_p, w_p les déplacements du point P parallèlement aux mêmes axes.

Les coordonnées des points O' et P' sont alors:

$$(x_0 + u_0, y_0 + v_0, z_0 + w_0) \text{ et } (x_p + u_p, y_p + v_p, z_p + w_p) \quad (34.154)$$

Avant déformation, soit L la longueur OP :

$$L^2 = (x_p - x_0)^2 + (y_p - y_0)^2 + (z_p - z_0)^2 \quad (34.155)$$

Après déformation, nous avons une longueur L' valant:

$$L'^2 = \left[(x_p + u_p) - (x_0 + u_0) \right]^2 + \left[(y_p + v_p) - (y_0 + v_0) \right]^2 + \left[(z_p + w_p) - (z_0 + w_0) \right]^2 \quad (34.156)$$

Si ΔL est l'allongement de l'élément OP pendant la déformation, nous avons:

$$L'^2 = (L + \Delta L)^2 = L^2 + 2 L \Delta L + \Delta L^2 \quad (34.157)$$

En effectuant les quelques transformations suivantes:

$$\begin{aligned} L'^2 &= L^2 + 2 L \Delta L + \Delta L^2 = \\ &\left[(x_p + u_p)^2 + (x_0 + u_0)^2 - 2(x_p + u_p)(x_0 + u_0) \right] + \\ &\left[(y_p + v_p)^2 + (y_0 + v_0)^2 - 2(y_p + v_p)(y_0 + v_0) \right] + \\ &\left[(z_p + w_p)^2 + (z_0 + w_0)^2 - 2(z_p + w_p)(z_0 + w_0) \right] \end{aligned} \quad (34.158)$$

En développant:

$$\begin{aligned} &\left[x_p^2 + u_p^2 + 2x_p u_p + x_0^2 + u_0^2 + 2x_0 u_0 - 2x_p x_0 - 2x_p u_0 - 2x_0 u_p - 2u_p u_0 \right] + \\ &\left[y_p^2 + v_p^2 + 2y_p v_p + y_0^2 + v_0^2 + 2y_0 v_0 - 2y_p y_0 - 2y_p v_0 - 2y_0 v_p - 2v_p v_0 \right] + \\ &\left[z_p^2 + w_p^2 + 2z_p w_p + z_0^2 + w_0^2 + 2z_0 w_0 - 2z_p z_0 - 2z_p w_0 - 2z_0 w_p - 2w_p w_0 \right] \end{aligned} \quad (34.159)$$

Soit:

$$\begin{aligned} L^2 + 2 L \Delta L + \Delta L^2 &= \\ &\left[(x_p - x_0)^2 + (y_p - y_0)^2 + (z_p - z_0)^2 \right] + \\ &2 \left[(x_p - x_0)(u_p - u_0) + (y_p - y_0)(v_p - v_0) + (z_p - z_0)(w_p - w_0) \right] + \\ &\left[(u_p - u_0)^2 + (v_p - v_0)^2 + (w_p - w_0)^2 \right] \end{aligned} \quad (34.160)$$

En négligeant les termes de déplacement d'ordre supérieur et en tenant compte de la relation:

$$L^2 = (x_p - x_0)^2 + (y_p - y_0)^2 + (z_p - z_0)^2 \quad (34.161)$$

il vient que ΔL^2 disparaît avec $(x_p - x_0)^2 + (y_p - y_0)^2 + (z_p - z_0)^2$ ainsi que les termes au carré, nous avons:

$$\frac{\Delta L}{L} = \frac{1}{L^2} \left[(x_p - x_0)(u_p - u_0) + (y_p - y_0)(v_p - v_0) + (z_p - z_0)(w_p - w_0) \right] \quad (34.162)$$

Or, la géométrie analytique (trigonométrie élémentaire; rapport des côtés opposés et adjacents à l'hypoténuse) donne les relations suivantes:

$$\frac{x_p - x_0}{L} = l, \quad \frac{y_p - y_0}{L} = m, \quad \frac{z_p - z_0}{L} = n \quad (34.163)$$

qui sont les cosinus directeurs de la droite L .

Nous pouvons alors écrire:

$$\varepsilon = \frac{\Delta L}{L} = \frac{l}{L}(u_p - u_0) + \frac{m}{L}(v_p - v_0) + \frac{n}{L}(w_p - w_0) \quad (34.164)$$

La variation u_p étant un déplacement faible, nous avons recours à un développement en série de Taylor (cf. chapitres Suites Et Séries) dont nous négligeons les termes d'ordres supérieurs (linéarisation des équations):

$$u_p = u + x_p \partial_x u + y_p \partial_y u + z_p \partial_z u \quad (34.165)$$

Nous avons également:

$$u_0 = u + x_0 \partial_x u + y_0 \partial_y u + z_0 \partial_z u \quad (34.166)$$

La différence donne:

$$u_p - u_0 = (x_p - x_0) \partial_x u + (y_p - y_0) \partial_y u + (z_p - z_0) \partial_z u \quad (34.167)$$

Donc, nous pouvons maintenant écrire:

$$\begin{aligned} u_p - u_0 &= l L \partial_x u + m L \partial_y u + n L \partial_z u \\ v_p - v_0 &= l L \partial_x v + m L \partial_y v + n L \partial_z v \\ w_p - w_0 &= l L \partial_x w + m L \partial_y w + n L \partial_z w \end{aligned} \quad (34.168)$$

Finalement:

$$\begin{aligned} \varepsilon &= \frac{l}{L} (l L \partial_x u + m L \partial_y u + n L \partial_z u) + \frac{m}{L} (l L \partial_x v + m L \partial_y v + n L \partial_z v) \\ &+ \frac{n}{L} (l L \partial_x w + m L \partial_y w + n L \partial_z w) \\ &= l^2 \partial_x u + m l \partial_y u + n l \partial_z u + m l \partial_x v + m^2 \partial_y v + m n \partial_z v + n l \partial_x w + m n \partial_y w + n^2 \partial_z w \end{aligned} \quad (34.169)$$

En groupant, nous avons:

$$\varepsilon = l^2 \partial_x u + m^2 \partial_y v + n^2 \partial_z w + m n (\partial_z v + \partial_y w) + n l (\partial_z u + \partial_x w) + m l (\partial_y u + \partial_x v) \quad (34.170)$$

Cette expression permet en un point quelconque le calcul de la déformation ε dans une direction ayant comme cosinus directeur l, m, n en fonction des déplacements u, v, w en ce point !

Soit le cas où la ligne L coïncide avec l'axe OX , nous avons $l = 1, m = n = 0$, l'équation précédente devient alors:

$$\varepsilon_x = \partial_x u \quad (34.171)$$

Nous avons, si L coïncide avec l'axe OY ($m = 1, l = n = 0$) ou avec l'axe OZ ($n = 1, l = m = 0$):

$$\varepsilon_y = \partial_y v \text{ et } \varepsilon_x = \partial_x w \quad (34.172)$$

Les grandeurs $\varepsilon_x, \varepsilon_y, \varepsilon_z$ sont appelées "**déformations normales**" et n'ont pas d'unités.

Pour l'interprétation des termes $(\partial_x v + \partial_y w), (\partial_x u + \partial_y w), (\partial_y u + \partial_x v)$, nous nous référerons à la figure suivante:

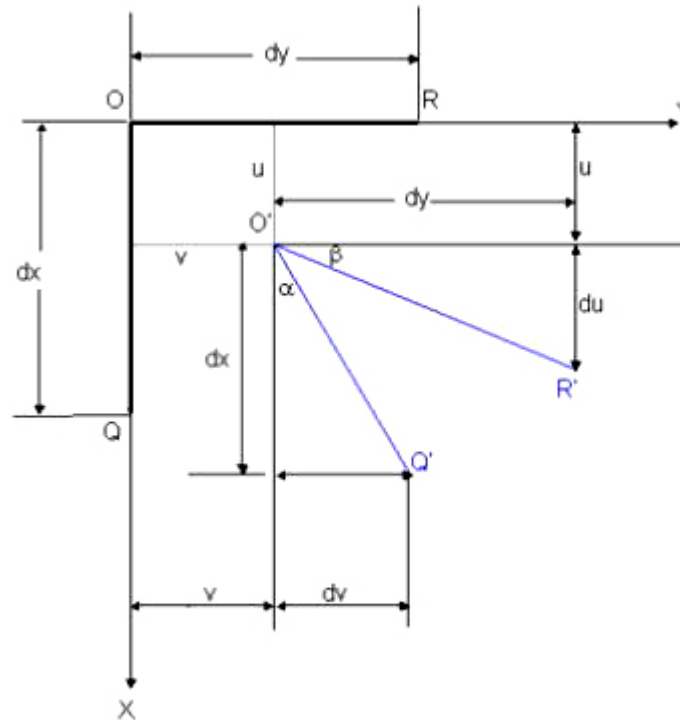


Figure: 34.26 - Situation permettant de faire émerger les tensions de cisaillement

Soient deux segments de droite OR et OQ situés dans le plan XOY . Avant déformation OR et OQ coïncidaient avec le référentiel orthonormé YOX . Après déformation, ils peuvent prendre la position $O'R'$ et $O'Q'$. Les composantes du déplacement de O sont u, v .

- La composante du déplacement de R' est calculée comme suit:

$$du = dy \tan \beta \text{ avec } \beta \cong \tan \beta = \frac{du}{dy} \quad (34.173)$$

car l'angle est faible .

En toute généralité comme $u = f(x, y, z)$, nous écrirons:

$$\beta = \partial_y u \quad (34.174)$$

- La composante du déplacement de Q' est elle:

$$\alpha = \partial_x v \quad (34.175)$$

Comme avant déformation, l'angle QOR est de $\pi/2$, après déformation, l'angle droit est réduit de $\alpha + \beta$. Cette réduction $\alpha + \beta$ est appelée "**déformation de cisaillement**" ou "**déformation tangentielle**".

et est notée par γ_{xy} .

Nous procéderons de la même façon pour les autres termes, d'où:

$$\begin{aligned}\gamma_{xy} &= \partial_y u + \partial_x v \\ \gamma_{xz} &= \partial_z u + \partial_x w \\ \gamma_{yz} &= \partial_z v + \partial_y w\end{aligned}\quad (34.176)$$

Compte tenu du quadruplet de groupes d'équations démontrées précédemment dans cette section (voir les déformations des solides):

$$\begin{aligned}\sigma_x &= \frac{E}{1+\eta} \left[\varepsilon_x + \frac{\eta}{1-2\eta} (\varepsilon_x + \varepsilon_y + \varepsilon_z) \right] \\ \sigma_y &= \frac{E}{1+\eta} \left[\varepsilon_y + \frac{\eta}{1-2\eta} (\varepsilon_x + \varepsilon_y + \varepsilon_z) \right] \\ \sigma_z &= \frac{E}{1+\eta} \left[\varepsilon_z + \frac{\eta}{1-2\eta} (\varepsilon_x + \varepsilon_y + \varepsilon_z) \right] \\ \gamma_{xy} &= \partial_y u + \partial_x v, \gamma_{xz} = \partial_z u + \partial_x w, \gamma_{yz} = \partial_z v + \partial_y w \\ \begin{bmatrix} N_x \\ N_y \\ N_z \end{bmatrix} &= \begin{bmatrix} \sigma_{xx} & \tau_{xy} & \tau_{xz} \\ \tau_{yx} & \sigma_{yy} & \tau_{yz} \\ \tau_{zx} & \tau_{zy} & \sigma_{zz} \end{bmatrix} \begin{bmatrix} l \\ m \\ n \end{bmatrix} \\ \tau &= \frac{E}{2(1+\eta)} \gamma \\ &\quad (34.177)\end{aligned}$$

Nous pouvons résumer:

$$\begin{aligned}\sigma_{xx} &= \frac{E}{1+\eta} \left[\varepsilon_{xx} + \frac{\eta}{1-2\eta} (\varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz}) \right] \\ \sigma_{yy} &= \frac{E}{1+\eta} \left[\varepsilon_{yy} + \frac{\eta}{1-2\eta} (\varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz}) \right] \\ \sigma_{zz} &= \frac{E}{1+\eta} \left[\varepsilon_{zz} + \frac{\eta}{1-2\eta} (\varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz}) \right] \\ \tau_{xy} &= \tau_{yx} = \frac{E}{2(1+\eta)} \gamma_{xy} = \frac{E}{2(1+\eta)} (\partial_y u + \partial_x v) \\ \tau_{yz} &= \tau_{zy} = \frac{E}{2(1+\eta)} \gamma_{yz} = \frac{E}{2(1+\eta)} (\partial_z v + \partial_y w) \\ \tau_{xz} &= \tau_{zx} = \frac{E}{2(1+\eta)} \gamma_{xz} = \frac{E}{2(1+\eta)} (\partial_z u + \partial_x w)\end{aligned}\quad (34.178)$$

Généralement, nous posons pour simplifier les notations (il faut cependant ne pas croire que la déformation en cisaillement devient une déformation normale ! ce n'est qu'une convention d'écriture dont le physicien doit se rappeler !):

$$\gamma_{xy} = 2\varepsilon_{xy}, \gamma_{xz} = 2\varepsilon_{xz}, \gamma_{yz} = 2\varepsilon_{yz} \quad (34.179)$$

De même, nous posons:

$$\varepsilon = \varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz} \quad (34.180)$$

Soit finalement:

$$\begin{aligned} \sigma_{xx} &= \frac{E}{1+\eta} \left(\varepsilon_{xx} + \frac{\eta}{1-2\eta} \varepsilon \right) & \tau_{xy} &= \tau_{yx} = \frac{E}{2(1+\eta)} 2\varepsilon_{xy} \\ \sigma_{yy} &= \frac{E}{1+\eta} \left(\varepsilon_{yy} + \frac{\eta}{1-2\eta} \varepsilon \right) & \tau_{yz} &= \tau_{zy} = \frac{E}{2(1+\eta)} 2\varepsilon_{yz} \\ \sigma_{zz} &= \frac{E}{1+\eta} \left(\varepsilon_{zz} + \frac{\eta}{1-2\eta} \varepsilon \right) & \tau_{xz} &= \tau_{zx} = \frac{E}{2(1+\eta)} 2\varepsilon_{xz} \end{aligned} \quad (34.181)$$

En tenant compte que:

$$G = \frac{E}{2(1+\eta)} \quad (34.182)$$

Nous obtenons les tensions de cisaillement comme suit:

$$\begin{aligned} \sigma_{xx} &= 2G \left(\varepsilon_{xx} + \frac{\eta}{1-2\eta} \varepsilon \right) & \tau_{xy} &= \tau_{yx} = 2G \varepsilon_{xy} \\ \sigma_{yy} &= 2G \left(\varepsilon_{yy} + \frac{\eta}{1-2\eta} \varepsilon \right) & \tau_{yz} &= \tau_{zy} = 2G \varepsilon_{yz} \\ \sigma_{zz} &= 2G \left(\varepsilon_{zz} + \frac{\eta}{1-2\eta} \varepsilon \right) & \tau_{xz} &= \tau_{zx} = 2G \varepsilon_{xz} \end{aligned} \quad (34.183)$$

Considérons maintenant, pour exemple, un fluide circulant dans la direction de OY avec un gradient de vitesse dans la direction de x:

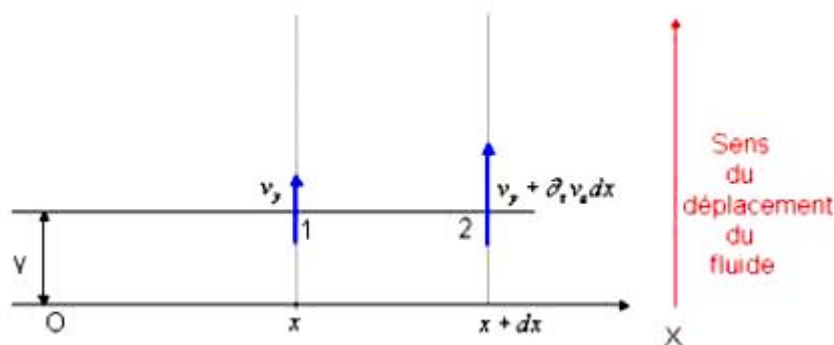


Figure: 34.27 - Situation permettant de faire émerger les contraintes normales

En se plaçant au niveau de y et au point 1 d'abscisse x , nous avons une vitesse v_y et au point 2 d'abscisse $x+dx$, une vitesse:

$$v_y + \partial_x v_y dx \text{ (avec } \partial_x v_y \neq 0 \text{)} \quad (34.184)$$

Dans la direction de x , il n'y a pas de composante de vitesse donc:

$$v_x = 0 \text{ (avec } \partial_y v_x = 0 \text{)} \quad (34.185)$$

Nous supposons maintenant que les tensions de cisaillement sont proportionnelles à $\partial_x v_y$ à un facteur près tel que:

$$\tau_{xy} = \mu_d \partial_x v_y \quad (34.186)$$

avec pour rappel:

$$\mu_d = \frac{dF}{\partial_x v dS} \quad (34.187)$$

Il est donc possible de considérer des déplacements par unité de temps en posant:

$$u \rightarrow v_x \quad v \rightarrow v_y \quad w \rightarrow v_z \quad (34.188)$$

En rapprochant cette dernière relation de:

$$\begin{aligned} \tau_{xy} &= G (\partial_y u + \partial_x v) = \mu_d (\partial_y v_x + \partial_x v_y) \\ \tau_{yz} &= G (\partial_z v + \partial_y w) = \mu_d (\partial_z v_y + \partial_y v_z) \\ \tau_{zx} &= G (\partial_x w + \partial_z u) = \mu_d (\partial_x v_z + \partial_z v_x) \end{aligned} \quad (34.189)$$

nous pouvons dire alors que G initialement valable dans un milieu élastique solide considéré par ses déplacements est l'analogie de μ_d dans le cas d'un fluide visqueux considéré par ses déplacements par unité de temps. Ainsi, nous voyons que les unités sont conservées.

En considérant également les déformations par unité de temps pour les contraintes normales (nous y reviendrons en détail un peu plus loin), nous avons alors le système d'équations:

$$\begin{aligned} \sigma_{xx} &= 2\mu_d \left(\dot{\epsilon}_{xx} + \frac{\eta}{1-2\eta} \dot{\epsilon} \right) = 2\mu_d \theta_{xx} + 2\mu_d \frac{\eta}{1-2\eta} \dot{\epsilon} = 2\mu_d \theta_{xx} + \alpha(\theta_{xx} + \theta_{yy} + \theta_{zz}) \\ \sigma_{yy} &= 2\mu_d \left(\dot{\epsilon}_{yy} + \frac{\eta}{1-2\eta} \dot{\epsilon} \right) = 2\mu_d \theta_{yy} + 2\mu_d \frac{\eta}{1-2\eta} \dot{\epsilon} = 2\mu_d \theta_{yy} + \alpha(\theta_{xx} + \theta_{yy} + \theta_{zz}) \\ \sigma_{zz} &= 2\mu_d \left(\dot{\epsilon}_{zz} + \frac{\eta}{1-2\eta} \dot{\epsilon} \right) = 2\mu_d \theta_{zz} + 2\mu_d \frac{\eta}{1-2\eta} \dot{\epsilon} = 2\mu_d \theta_{zz} + \alpha(\theta_{xx} + \theta_{yy} + \theta_{zz}) \end{aligned}$$

$$\begin{aligned}\tau_{xy} &= \tau_{yx} = 2\mu_d \dot{\epsilon}_{xy} = 2\mu_d \theta_{xy} \\ \tau_{yz} &= \tau_{zy} = 2\mu_d \dot{\epsilon}_{yz} = 2\mu_d \theta_{yz} \\ \tau_{xz} &= \tau_{zx} = 2\mu_d \dot{\epsilon}_{xz} = 2\mu_d \theta_{xz}\end{aligned}$$

(34.190)

Ainsi, nous obtenons une écriture condensée:

$$P_{ij} = 2\mu_d \theta_{ij} + \alpha (\theta_{11} + \theta_{22} + \theta_{33}) \delta_{ij} \quad (34.191)$$

où δ_{ij} est le symbole de Kronecker:

$$\delta_{ij} = \begin{cases} \delta_{ij}=1 & , i = j \\ \delta_{ij}=0 & , i \neq j \end{cases} \quad (34.192)$$

Le tenseur P_{ij} décrit ainsi en partie l'ensemble des contraintes d'un fluide visqueux dans lequel nous avons supposé dans le cadre de l'hypothèse d'un fluide newtonien qu'il y a des relations linéaires entre les tensions et les déformations normales.

Nous posons maintenant la somme des contraintes dynamiques sous une forme générale que nous allons justifier:

$$P_{ij} = 2\mu_d \theta_{ij} + \alpha (\theta_{11} + \theta_{22} + \theta_{33}) \delta_{ij} - p \delta_{ij} \quad (34.193)$$

où le terme $-p \delta_{ij}$ se justifie par le fait que dans le cas statique, une pression dynamique constante p existe toujours en un point d'un fluide ce que l'on n'a pas dans le cas d'un solide. Pour justifier le signe négatif, nous observerons que dans l'expression de P_{ij} , les deux premiers termes du membre de droite correspondent, dans l'étude précédente, à des contraintes d'extension, alors que la pression p correspond à une compression du fluide.

Il nous reste à présent, à déterminer le coefficient α . Soit $i = j$, nous avons alors $\delta_{ij} = 1$. Il vient successivement et par addition:

$$\begin{aligned}P_{11} &= 2\mu_d \theta_{11} + \alpha (\theta_{11} + \theta_{22} + \theta_{33}) - p \\ P_{22} &= 2\mu_d \theta_{22} + \alpha (\theta_{11} + \theta_{22} + \theta_{33}) - p \\ P_{33} &= 2\mu_d \theta_{33} + \alpha (\theta_{11} + \theta_{22} + \theta_{33}) - p \\ \hline P_{11} + P_{22} + P_{33} &= 2\mu_d (\theta_{11} + \theta_{22} + \theta_{33}) + 3\alpha (\theta_{11} + \theta_{22} + \theta_{33}) - 3p\end{aligned} \quad (34.194)$$

Cette expression doit répondre à un fluide qui est également dans une situation statique (au repos) telle que:

$$P_{11} + P_{22} + P_{33} = -3p \quad (34.195)$$

Il vient alors que dans le cas statique:

$$2\mu_d(\theta_{11} + \theta_{22} + \theta_{33}) + 3\alpha(\theta_{11} + \theta_{22} + \theta_{33}) = 0 \quad (34.196)$$

Puisque:

$$\theta_{11} + \theta_{22} + \theta_{33} \neq 0 \quad (34.197)$$

Nous avons alors:

$$2\mu_d + 3\alpha = 0 \Rightarrow \alpha = -\frac{2}{3}\mu_d \quad (34.198)$$

L'expression générale des contraintes s'écrit alors:

$$P_{ij} = 2\mu_d \theta_{ij} - \frac{2}{3}\mu_d (\theta_{11} + \theta_{22} + \theta_{33}) \delta_{ij} - p \delta_{ij} \quad (34.199)$$

Présentement, nous allons introduire les opérateurs de l'analyse vectorielle afin de disposer d'une expression plus générale. De cette façon, nous pourrions adapter la formulation à n'importe quel système de coordonnées (cartésiennes, cylindriques, sphériques,...) ce qui facilitera la résolution de problèmes pratiques.

Nous avons vu que pour un solide, nous avons:

$$\begin{aligned} \gamma_{xy} &= 2\varepsilon_{xy} = \partial_y u + \partial_x v \\ \gamma_{xz} &= 2\varepsilon_{xz} = \partial_z u + \partial_x w \\ \gamma_{yz} &= 2\varepsilon_{yz} = \partial_z v + \partial_y w \end{aligned} \quad (34.200)$$

Nous allons déterminer ces équations sous une forme indicielle en considérant toujours des déplacements par unité de temps (vitesses).

$$\dot{\varepsilon}_{ij} = \frac{1}{2}(\partial_{x^i} v_j + \partial_{x^j} v_i) \quad (34.201)$$

tel que $i, j = \{1, 2, 3\}$ et que $1 \rightarrow x, x^1 = x$; $2 \rightarrow y, x^2 = y$; $3 \rightarrow z, x^3 = z$

Pour $i = j = 1$ nous avons ainsi:

$$\dot{\varepsilon}_{11} = \frac{1}{2}(\partial_{x^1} v_1 + \partial_{x^1} v_1) = \partial_{x^1} v_1 \text{ ou } \dot{\varepsilon}_{xx} = \partial_x v_x \quad (34.202)$$

Pour $i = 1, j = 2$ nous avons:

$$\dot{\varepsilon}_{12} = \frac{1}{2}(\partial_{x^1} v_2 + \partial_{x^2} v_1) \text{ ou } \dot{\varepsilon}_{xy} = \frac{1}{2}(\partial_y v_x + \partial_x v_y) \quad (34.203)$$

Nous pouvons dès lors écrire:

$$\begin{aligned}\dot{\varepsilon}_{xx} &= \partial_x v_x, \quad \dot{\varepsilon}_{yy} = \partial_y v_y, \quad \dot{\varepsilon}_{zz} = \partial_z v_z \\ \dot{\varepsilon}_{xy} &= \frac{1}{2}(\partial_y v_x + \partial_x v_y), \quad \dot{\varepsilon}_{xz} = \frac{1}{2}(\partial_z v_x + \partial_x v_z), \quad \dot{\varepsilon}_{yz} = \frac{1}{2}(\partial_z v_y + \partial_y v_z)\end{aligned}\quad (34.204)$$

En effectuant la somme des termes de:

$$\dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy} + \dot{\varepsilon}_{zz} = \partial_x v_x + \partial_y v_y + \partial_z v_z \quad (34.205)$$

Or, les outils de l'analyse vectorielle nous permettent d'écrire:

$$\dot{\varepsilon}_{xx} + \dot{\varepsilon}_{yy} + \dot{\varepsilon}_{zz} = \partial_x v_x + \partial_y v_y + \partial_z v_z = \vec{\nabla} \circ \vec{v} \quad (34.206)$$

Pour le fluide, nous aurons ainsi:

$$\theta_{11} + \theta_{22} + \theta_{33} = \vec{\nabla} \circ \vec{v} \quad (34.207)$$

L'équation générale dynamique des contraintes s'écrira alors sous la forme suivante pour un fluide newtonien:

$$P_{ij} = 2\mu_d \theta_{ij} - \frac{2}{3}\mu_d (\vec{\nabla} \circ \vec{v}) \delta_{ij} - p \delta_{ij} \quad (34.208)$$

Tenseurs des contraintes que certains auteurs notent (l'écriture est un peu dangereuse mais elle a une justification dans un cadre d'étude plus approfondi des fluides!):

$$\sigma = -pI + \lambda (\vec{\nabla} \circ \vec{v}) I + 2\mu D \quad (34.209)$$

ou encore pour différencier vecteur et tenseur:

$$\underline{\underline{\sigma}} = -p\underline{\underline{I}} + \lambda (\underline{\underline{\nabla}} \circ \underline{\underline{v}}) \underline{\underline{I}} + 2\mu \underline{\underline{D}} \quad (34.210)$$

Si les contraintes normales (fluide incompressible) sont négligeables le deuxième terme se simplifie et nous avons alors (relation que nous retrouverons dans le chapitre de Génie Marin Et Météo):

$$\underline{\underline{\sigma}} = -p\underline{\underline{I}} + 2\mu \underline{\underline{D}} \quad (34.211)$$

Il est, à présent, utile de repasser sous une forme développée pour l'équation précédente, en se rappelant que (voir plus haut):

$$\begin{aligned}P_{xx} &= -p - \frac{2}{3}\mu_d (\vec{\nabla} \circ \vec{v}) + 2\mu_d \partial_x v_x \\ P_{yy} &= -p - \frac{2}{3}\mu_d (\vec{\nabla} \circ \vec{v}) + 2\mu_d \partial_y v_y \\ P_{zz} &= -p - \frac{2}{3}\mu_d (\vec{\nabla} \circ \vec{v}) + 2\mu_d \partial_z v_z\end{aligned}$$

$$\begin{aligned}
 P_{xy} &= \mu_d (\partial_x v_x + \partial_x v_y) \\
 P_{xz} &= \mu_d (\partial_x v_x + \partial_x v_z) \\
 P_{yz} &= \mu_d (\partial_x v_y + \partial_y v_z)
 \end{aligned}
 \quad (34.212)$$

Écrivons maintenant le système d'équations de Newton (sommes des contraintes dynamiques internes et externes à un élément de volume d'un fluide) qui est:

$$\frac{F_T}{V} = \frac{F_{int}}{V} + \frac{F_{ext}}{V} = \rho d_t v_i = \rho f_i + \frac{\partial}{\partial x_j} \sum_{j=1}^3 P_{ij} \quad (34.213)$$

où:

- $\frac{\partial}{\partial x_j} \sum_{j=1}^3 P_{ij}$ est la somme des forces externes par unité de volume

- f_i est la notation traditionnelle (malheureuse...) de l'accélération massique en $m \cdot s^{-2}$

- ρ est la densité du fluide

et qui peut s'écrire sous forme condensée:

$$\rho d_t \vec{v} = \vec{\nabla}(\vec{\sigma}) + \rho \vec{f} \quad (34.214)$$

avec:

$$(\text{div} \vec{\sigma})_i = \vec{\nabla}(\vec{\sigma})_i = \frac{\partial \sigma_{ij}}{\partial x_j} = \frac{\partial P_{ij}}{\partial x_j} \quad (34.215)$$

Nous avons:

$$\begin{aligned}
 \rho d_t v_x &= \partial_x P_{xx} + \partial_y P_{xy} + \partial_z P_{xz} + \rho f_x \\
 \rho d_t v_y &= \partial_x P_{yx} + \partial_y P_{yy} + \partial_z P_{yz} + \rho f_y \\
 \rho d_t v_z &= \partial_x P_{zx} + \partial_y P_{zy} + \partial_z P_{zz} + \rho f_z
 \end{aligned}
 \quad (34.216)$$

En introduisant les expressions de P_{ij} obtenues dans la relation ci-dessus, nous aboutissons aux équations:

$$\begin{aligned}
 \rho d_t v_x &= -\partial_x p - \partial_x \left(\frac{2}{3} \mu_d (\vec{\nabla} \circ \vec{v}) - 2 \mu_d \partial_x v_x \right) + \partial_y [\mu_d (\partial_y v_x + \partial_x v_y)] + \partial_z [\mu_d (\partial_x v_x + \partial_x v_z)] + \rho f_x \\
 \rho d_t v_y &= \partial_x [\mu_d (\partial_y v_x + \partial_x v_y)] - \partial_y p - \partial_y \left(\frac{2}{3} \mu_d (\vec{\nabla} \circ \vec{v}) - 2 \mu_d \partial_y v_y \right) + \partial_z [\mu_d (\partial_x v_y + \partial_y v_z)] + \rho f_y \\
 \rho d_t v_z &= \partial_x [\mu_d (\partial_x v_x + \partial_x v_z)] + \partial_y [\mu_d (\partial_x v_y + \partial_y v_z)] - \partial_z p - \partial_z \left(\frac{2}{3} \mu_d (\vec{\nabla} \circ \vec{v}) - 2 \mu_d \partial_z v_z \right) + \rho f_z
 \end{aligned}
 \quad (34.217)$$

Ce sont les "équations de Navier-Stokes de la dynamique des fluides newtoniens". Il en existe deux formes condensées que nous allons de suite déterminer:

En reprenant la première équation de Navier-Stokes et en la développant, il vient:

$$\begin{aligned}
 \rho d_t v_x &= -\partial_x p - \partial_x \left(\frac{2}{3} \mu_d (\vec{\nabla} \circ \vec{v}) - 2\mu_d \partial_x v_x \right) + \partial_y \left[\mu_d (\partial_y v_x + \partial_x v_y) \right] + \partial_z \left[\mu_d (\partial_z v_x + \partial_x v_z) \right] + \rho f_x \\
 &= -\partial_x p - \frac{2}{3} \mu_d \partial_x (\vec{\nabla} \circ \vec{v}) + 2\mu_d \partial_x^2 v_x + \mu_d \partial_y^2 v_x + \mu_d \partial_{yx}^2 v_y + \mu_d \partial_z^2 v_x + \mu_d \partial_{zx}^2 v_z + \rho f_x \\
 &= -\partial_x p - \frac{2}{3} \mu_d \partial_x (\vec{\nabla} \circ \vec{v}) + \overbrace{\mu_d \partial_x^2 v_x}^{2\mu_d \partial_x^2 v_x} + \mu_d \partial_y^2 v_x + \mu_d \partial_{yx}^2 v_y + \mu_d \partial_z^2 v_x + \mu_d \partial_{zx}^2 v_z + \rho f_x \\
 &= -\partial_x p - \frac{2}{3} \mu_d \partial_x (\vec{\nabla} \circ \vec{v}) + \mu_d (\partial_x^2 v_x + \partial_{yx}^2 v_y + \partial_{zx}^2 v_z) + \mu_d (\partial_x^2 v_x + \partial_y^2 v_x + \partial_z^2 v_x) + \rho f_x
 \end{aligned}
 \tag{34.218}$$

Comme:

$$\partial_x^2 v_x + \partial_{yx}^2 v_y + \partial_{zx}^2 v_z = \partial_x (\partial_x v_x + \partial_y v_y + \partial_z v_z) = \partial_x (\vec{\nabla} \circ \vec{v}) \tag{34.219}$$

et que:

$$\partial_x^2 v_x + \partial_y^2 v_x + \partial_z^2 v_x = \vec{\nabla} \circ (\vec{\nabla}(v_x)) \tag{34.220}$$

Nous obtenons:

$$\rho d_t v_x = -\partial_x p - \frac{2}{3} \mu_d \partial_x (\vec{\nabla} \circ \vec{v}) + \mu_d \partial_x (\vec{\nabla} \circ \vec{v}) + \mu_d \vec{\nabla} \circ (\vec{\nabla}(v_x)) + \rho f_x \tag{34.221}$$

En simplifiant, il vient finalement:

$$\rho d_t v_x = -\partial_x p + \frac{1}{3} \mu_d \partial_x (\vec{\nabla} \circ \vec{v}) + \mu_d \vec{\nabla} \circ (\vec{\nabla}(v_x)) + \rho f_x \tag{34.222}$$

En opérant de la même manière pour les deux autres composantes, nous pouvons réduire le système d'équations de Navier-Stokes à une seule équation vectorielle:

$$\rho d_t \vec{v} = -\vec{\nabla} \cdot p + \frac{1}{3} \mu_d \vec{\nabla} (\vec{\nabla} \circ \vec{v}) + \mu_d \vec{\nabla} \circ (\vec{\nabla} \cdot \vec{v}) + \rho \vec{f} \tag{34.223}$$

Comme (cf. chapitre de Calcul Vectoriel):

$$\Delta \cdot v = \vec{\nabla}^2 v = \vec{\nabla} \circ (\vec{\nabla} \cdot \vec{v}) = \vec{\nabla} (\vec{\nabla} \circ \vec{v}) - \vec{\nabla} \times (\vec{\nabla} \times \vec{v}) \tag{34.224}$$

Nous avons:

$$\rho d_t \vec{v} = -\vec{\nabla} \cdot p + \frac{1}{3} \mu_d \vec{\nabla} (\vec{\nabla} \circ \vec{v}) + \mu_d [\vec{\nabla} (\vec{\nabla} \circ \vec{v}) - \vec{\nabla} \times (\vec{\nabla} \times \vec{v})] + \rho \vec{f} \tag{34.225}$$

Soit en final:

$$\rho d_t \vec{v} = -\vec{\nabla} \cdot p + \frac{4}{3} \mu_a \vec{\nabla} (\vec{\nabla} \circ \vec{v}) - \mu_a \vec{\nabla} \times (\vec{\nabla} \times \vec{v}) + \rho \vec{f} \quad (34.226)$$

Remarque: Nous trouvons également parfois dans la littérature, une équation contenant une seconde viscosité μ'_a , alors que μ_a ne se manifeste rigoureusement que lors du cisaillement pur selon nos hypothèses, μ'_a apparaît lors d'une compression omnidirectionnelle s'accompagnant d'une variation de densité.

L'équation précédente s'écrit alors:

$$\rho d_t \vec{v} = -\vec{\nabla} \cdot p + \frac{4}{3} (\mu_a + \mu'_a) \vec{\nabla} (\vec{\nabla} \circ \vec{v}) - \mu_a \vec{\nabla} \times (\vec{\nabla} \times \vec{v}) + \rho \vec{f} \quad (34.227)$$

C'est "l'équation de Navier-Stokes" ou aussi appelée "équation de mouvement pour un fluide newtonien".

2.4.1. FLUIDE INCOMPRESSIBLE

Dans un fluide incompressible, nous avons par définition $\rho = c^*$. L'équation de conservation qui est (cf. chapitre de Thermodynamique):

$$\vec{\nabla} \circ (\rho \cdot \vec{v}) + \partial_t \rho = 0 \quad (34.228)$$

s'écrit alors:

$$\rho \vec{\nabla} \circ \vec{v} = 0 \quad (34.229)$$

soit:

$$\vec{\nabla} \circ \vec{v} = 0 \quad (34.230)$$

L'équation de Navier-Stokes sous la forme:

$$\rho d_t \vec{v} = -\vec{\nabla} \cdot p + \frac{1}{3} \mu_a \vec{\nabla} (\vec{\nabla} \circ \vec{v}) + \mu_a \vec{\nabla} \circ (\vec{\nabla} \cdot \vec{v}) + \rho \vec{f} \quad (34.231)$$

s'écrit alors:

$$\rho d_t \vec{v} = -\vec{\nabla} \cdot p + \mu_a \vec{\nabla} \circ (\vec{\nabla} \cdot \vec{v}) + \rho \vec{f} \quad (34.232)$$

ou autrement:

$$\rho d_t \vec{v} = -\vec{\nabla} \cdot p + \mu_a \vec{\nabla}^2 v + \rho \vec{f} \quad (34.233)$$

Si de plus la viscosité μ_a est négligeable, nous avons donc pour un fluide parfait:

$$\rho d_t \vec{v} = -\vec{\nabla} \cdot p + \rho \vec{f} \quad (34.234)$$

Cette équation est appelée "**équation d'Euler de 1ère forme**" ou encore "**équation locale du bilan de conservation de la quantité de mouvement**". Nous réutiliserons cette relation dans le cadre de notre étude des ondes de gravité (vagues) dans le chapitre de Génie Météo Et Marin.

Il existe une deuxième forme de l'équation d'Euler dans le cadre d'un fluide incompressible et à viscosité négligeable que nous allons de suite déterminer (souvent utilisée dans l'industrie):

Si $\vec{v} = f(x, y, z, t)$, nous pouvons écrire:

$$d_t \vec{v} = \frac{\partial \vec{v}}{\partial x} \frac{dx}{dt} + \frac{\partial \vec{v}}{\partial y} \frac{dy}{dt} + \frac{\partial \vec{v}}{\partial z} \frac{dz}{dt} + \frac{\partial \vec{v}}{\partial t} \quad (34.235)$$

Ce qui peut aussi s'écrire:

$$\begin{aligned} d_t \vec{v} &= (\partial_x \vec{v}) v_x + (\partial_y \vec{v}) v_y + (\partial_z \vec{v}) v_z + \partial_t \vec{v} \\ &= \left[\partial_x (v_x \vec{i} + v_y \vec{j} + v_z \vec{k}) \right] v_x + \left[\partial_y (v_x \vec{i} + v_y \vec{j} + v_z \vec{k}) \right] v_y + \left[\partial_z (v_x \vec{i} + v_y \vec{j} + v_z \vec{k}) \right] v_z + \partial_t \vec{v} \\ &= [v_x \partial_x v_x] \vec{i} + [v_x \partial_x v_y] \vec{j} + [v_x \partial_x v_z] \vec{k} + [v_y \partial_y v_x] \vec{i} + [v_y \partial_y v_y] \vec{j} + [v_y \partial_y v_z] \vec{k} + [v_z \partial_z v_x] \vec{i} + \\ &\quad [v_z \partial_z v_y] \vec{j} + [v_z \partial_z v_z] \vec{k} + \partial_t \vec{v} \end{aligned} \quad (34.236)$$

Ce qui s'écrit encore:

$$(v_x \partial_x + v_y \partial_y + v_z \partial_z) (v_x \vec{i} + v_y \vec{j} + v_z \vec{k}) + \partial_t \vec{v} \quad (34.237)$$

Le premier facteur peut être considéré comme le produit scalaire suivant:

$$(v_x \partial_x + v_y \partial_y + v_z \partial_z) = \left[\underbrace{(v_x \vec{i} + v_y \vec{j} + v_z \vec{k})}_{\vec{v}} \underbrace{(\partial_x \vec{i} + \partial_y \vec{j} + \partial_z \vec{k})}_{\vec{\nabla}} \right] = \vec{v} \circ \vec{\nabla} \quad (34.238)$$

Soit:

$$d_t \vec{v} = (v_x \partial_x + v_y \partial_y + v_z \partial_z) (v_x \vec{i} + v_y \vec{j} + v_z \vec{k}) + \partial_t \vec{v} = (\vec{v} \circ \vec{\nabla}) \vec{v} + \partial_t \vec{v} \quad (34.239)$$

La "**dérivée particulaire**" peut alors prendre la forme condensée suivante:

$$d_t \vec{v} = \partial_t \vec{v} + (\vec{v} \circ \vec{\nabla}) \vec{v} \quad (34.240)$$

Remarque: La composante en x de la dérivée particulaire est donc (nous retrouverons cela dans le chapitre de Génie Marin Et Météo!) :

$$d_t v_x = \partial_t v_x + v_x \partial_x v_x + v_y \partial_y v_x + v_z \partial_z v_x \quad (34.241)$$

ce que les spécialistes du domaine notent de manière générale pour toute composante :

$$\frac{D}{Dt} = \partial_t + v_x \partial_x + v_y \partial_y + v_z \partial_z = \partial_t + \vec{v} \circ \vec{\nabla} \quad (34.242)$$

L'équation d'Euler de 1^{ère} forme:

$$\rho d_t \vec{v} = -\vec{\nabla} \cdot p + \rho \vec{f} \quad (34.243)$$

devient compte tenu de la dérivée particulaire:

$$\rho \left[\partial_t \vec{v} + (\vec{v} \circ \vec{\nabla}) \vec{v} \right] = \rho \vec{f} - \vec{\nabla} \cdot p \quad (34.244)$$

ou encore (forme courante dans la littérature):

$$\partial_t \vec{v} + (\vec{v} \circ \vec{\nabla}) \vec{v} = -\frac{1}{\rho} \vec{\nabla} \cdot p + \vec{f} \quad (34.245)$$

Nous avons vu dans le chapitre de Calcul Vectoriel que:

$$\vec{\nabla} \cdot (\vec{u} \circ \vec{v}) = \vec{u} \times (\vec{\nabla} \times \vec{v}) + \vec{v} \times (\vec{\nabla} \times \vec{u}) + (\vec{u} \circ \vec{\nabla}) \vec{v} + (\vec{v} \circ \vec{\nabla}) \vec{u} \quad (34.246)$$

Si nous posons $\vec{u} = \vec{v}$, nous avons:

$$\vec{\nabla} \cdot \|\vec{v}\|^2 = \vec{\nabla} \cdot v^2 = 2\vec{v} \times (\vec{\nabla} \times \vec{v}) + 2(\vec{v} \circ \vec{\nabla}) \vec{v} \quad (34.247)$$

Soit:

$$(\vec{v} \circ \vec{\nabla}) \vec{v} = \frac{1}{2} \vec{\nabla} \cdot v^2 - \vec{v} \times (\vec{\nabla} \times \vec{v}) \quad (34.248)$$

Finalement, nous obtenons une nouvelle équation appelée "**équation d'Euler de 2ème forme**" et qui s'écrit:

$$\left| \rho d_t \vec{v} = \rho \vec{f} - \vec{\nabla} \cdot p - \frac{\rho}{2} \vec{\nabla} \cdot v^2 + \rho \vec{v} \times (\vec{\nabla} \times \vec{v}) \right| \quad (34.249)$$

Bien que les deux équations d'Euler soient très importantes, il en existe une forme variée très utile en météorologie que nous allons de suite déterminer.

Nous nous basons toujours sur l'écoulement d'un fluide incompressible et non visqueux, mais dont les

forces de volume dérivent cette fois-ci d'un potentiel $\vec{f} = -\vec{\nabla} \cdot U$ (U étant un potentiel).

Dans ce cas, nous recourons à l'équation d'Euler sous sa 1^{ère} forme:

$$\rho d_t \vec{v} = -\vec{\nabla} \cdot p + \rho \vec{f} \quad (34.250)$$

Puisque les forces volumiques $\vec{f}$ dérivent d'un potentiel U , nous avons:

$$\rho d_t \vec{v} = -\vec{\nabla} \cdot (p + \rho U) \quad (34.251)$$

Nous rappelons la relation:

$$(\vec{v} \circ \vec{\nabla}) \vec{v} = \frac{1}{2} \vec{\nabla} \cdot v^2 - \vec{v} \times (\vec{\nabla} \times \vec{v}) \quad (34.252)$$

Soit $\vec{\nabla} \times \vec{v}$ un vecteur $\vec{A}$, il vient:

$$-\vec{v} \times \vec{A} = \vec{A} \times \vec{v} \quad (34.253)$$

donc:

$$-\vec{v} \times (\vec{\nabla} \times \vec{v}) = (\vec{\nabla} \times \vec{v}) \times \vec{v} \quad (34.254)$$

donc nous pouvons aussi écrire:

$$(\vec{v} \circ \vec{\nabla}) \vec{v} = \frac{1}{2} \vec{\nabla} \cdot v^2 + (\vec{\nabla} \times \vec{v}) \times \vec{v} \quad (34.255)$$

En reprenant la relation:

$$d_t \vec{v} = \partial_t \vec{v} + (\vec{v} \circ \vec{\nabla}) \vec{v} \quad (34.256)$$

l'équation:

$$\rho d_t \vec{v} = -\vec{\nabla} \cdot (p + \rho U) \quad (34.257)$$

devient alors:

$$\rho d_t \vec{v} = \rho \left[\partial_t \vec{v} + (\vec{v} \circ \vec{\nabla}) \vec{v} \right] = -\vec{\nabla} \cdot (p + \rho U) \quad (34.258)$$

et en utilisant:

$$(\vec{v} \circ \vec{\nabla}) \vec{v} = \frac{1}{2} \vec{\nabla} \cdot v^2 + (\vec{\nabla} \times \vec{v}) \times \vec{v} \quad (34.259)$$

cette dernière devient:

$$\rho \partial_t \vec{v} + \rho \left[\frac{1}{2} \vec{\nabla} \cdot v^2 + (\vec{\nabla} \times \vec{v}) \times \vec{v} \right] = -\vec{\nabla} \cdot (p + \rho U) \quad (34.260)$$

et puisque:

$$\frac{1}{2} \vec{\nabla} \cdot \mathbf{v}^2 = \vec{\nabla} \cdot \left(\frac{\mathbf{v}^2}{2} \right) \quad (34.261)$$

nous pouvons finalement écrire:

$$\begin{aligned} \rho \partial_t \vec{v} + \rho \left[\frac{1}{2} \vec{\nabla} \cdot \mathbf{v}^2 + (\vec{\nabla} \times \vec{v}) \times \vec{v} \right] + \vec{\nabla} \cdot (p + \rho U) &= \vec{0} \\ \partial_t \vec{v} + \vec{\nabla} \cdot \left[\left(\frac{\mathbf{v}^2}{2} \right) + \frac{p}{\rho} + U \right] + (\vec{\nabla} \times \vec{v}) \times \vec{v} &= \vec{0} \end{aligned} \quad (34.262)$$

Généralisons cette dernière relation en faisant apparaître d'éventuelles rotations. Pour cela, nous savons que $\vec{v} = \vec{\omega} \times \vec{r}$ donc:

$$\vec{\nabla} \times \vec{v} = \vec{\nabla} \times (\vec{\omega} \times \vec{r}) \quad (34.263)$$

En écrivant le produit vectoriel $\vec{\omega} \times \vec{r}$ sous forme développée, nous avons:

$$\vec{\nabla} \times \left[(\omega_y z - \omega_z y) \vec{i} + (\omega_z x - \omega_x z) \vec{j} + (\omega_x y - \omega_y x) \vec{k} \right] \quad (34.264)$$

Ce qui donne:

$$\begin{aligned} \vec{\nabla} \times \vec{v} &= \partial_y (\omega_x y - \omega_y x) \vec{i} + \partial_x (\omega_z x - \omega_x z) \vec{k} + \partial_z (\omega_y z - \omega_z y) \vec{j} \\ &\quad - \partial_y (\omega_y z - \omega_z y) \vec{k} - \partial_z (\omega_z x - \omega_x z) \vec{i} - \partial_x (\omega_x y - \omega_y x) \vec{j} \end{aligned} \quad (34.265)$$

Supposons que $\vec{\omega}$ soit un vecteur vitesse angulaire constant, nous avons alors:

$$\vec{\nabla} \times \vec{v} = \omega_x \vec{i} + \omega_z \vec{k} + \omega_y \vec{j} + \omega_z \vec{k} + \omega_x \vec{i} + \omega_y \vec{j} = 2(\omega_x \vec{i} + \omega_y \vec{j} + \omega_z \vec{k}) = 2\vec{\omega} \quad (34.266)$$

Définition: Nous disons qu'un "écoulement est tourbillonnaire" si:

$$\vec{\nabla} \times \vec{v} \neq \vec{0} \quad (34.267)$$

partout ou en certains points. Nous définissons aussi de la relation antérieure la "vorticité" par:

$$\vec{\omega} = \frac{1}{2} \vec{\nabla} \times \vec{v} \quad (34.268)$$

Exemple d'écoulement partiellement tourbillonnaire (en certains points):

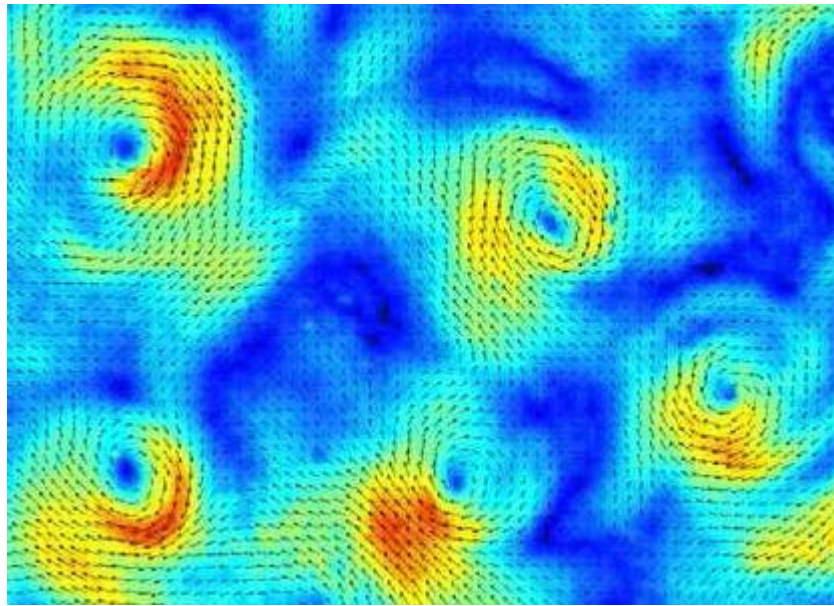


Figure: 34.28 - Exemple d'un écoulement tourbillonnaire

L'équation:

$$\partial_t \vec{v} + \vec{\nabla} \cdot \left[\frac{v^2}{2} + \frac{p}{\rho} + U \right] + (\vec{\nabla} \times \vec{v}) \times \vec{v} = \vec{0} \quad (34.269)$$

s'écrit alors:

$$\partial_t \vec{v} + \vec{\nabla} \cdot \left[\frac{v^2}{2} + \frac{p}{\rho} + U \right] + 2\vec{\omega} \times \vec{v} = \vec{0} \quad (34.270)$$

Nous retrouvons dans cette équation, utilisée en météorologie, l'accélération de Coriolis que nous avons déterminée dans le chapitre de Mécanique Classique.

Si l'écoulement s'effectue à vitesse constante $\partial_t \vec{v} = \vec{0}$ et n'est pas rotationnel (non turbulent) $\vec{\omega} \times \vec{v} = \vec{0}$, alors l'équation précédente se réduit à:

$$\vec{\nabla} \cdot \left[\frac{v^2}{2} + \frac{p}{\rho} + U \right] = \vec{0} \quad (34.271)$$

En dynamique classique du point matériel rigide, nous avons montré que dans le cas d'un potentiel gravitationnel Terrestre:

$$U = g(z - z_0) \quad (34.272)$$

z étant l'altitude d'un point du fluide par rapport à un niveau de référence z_0 . Si nous prenons z_0 pour le niveau du sol, l'avant-dernière relation devient donc dans le cas d'un écoulement dit alors "écoulement potentiel":

$$\vec{\nabla} \cdot \left[\frac{v^2}{2} + \frac{p}{\rho} + gz \right] = \vec{0} \quad (34.273)$$

Le terme entre crochets pour satisfaire cette relation doit être tel que:

$$\frac{1}{2} \rho v^2 + \rho g z + p = c^* \quad (34.274)$$

Nous retrouvons donc bien le théorème de Bernoulli, ce qui conforte notre modèle des fluides newtoniens selon le modèle de Navier-Stokes.

2.4.2. FLUIDE COMPRESSIBLE

Dans ce cas ρ_0 est une fonction de la pression p (cas des "fluides barotropes"). Nous considérons également que la viscosité est négligeable. Il vient alors:

$$\frac{\vec{\nabla} \cdot \vec{p}}{\rho} = \vec{\nabla} \cdot \int \frac{d\vec{p}}{\rho} \quad (34.275)$$

L'équation:

$$\partial_t \vec{v} + \vec{\nabla} \cdot \left[\frac{v^2}{2} + \frac{p}{\rho} + U \right] + 2\vec{\omega} \times \vec{v} = \vec{0} \quad (34.276)$$

s'écrit alors:

$$\partial_t \vec{v} + \vec{\nabla} \cdot \left[\frac{v^2}{2} + \int \frac{d\vec{p}}{\rho} + U \right] + 2\vec{\omega} \times \vec{v} = \vec{0} \quad (34.277)$$

2.4.3. FLUIDE STATIQUE

Dans le cas statique $\vec{v} = \vec{0}$ et $\mu_D = \vec{0}$ l'équation:

$$\rho d_t \vec{v} = -\vec{\nabla} \cdot \vec{p} + \frac{4}{3}(\mu_s + \mu'_s)\vec{\nabla}(\vec{\nabla} \cdot \vec{v}) - \mu_s \vec{\nabla} \times (\vec{\nabla} \times \vec{v}) + \rho \vec{f} \quad (34.278)$$

devient simplement:

$$\vec{\nabla} \cdot \vec{p} - \rho \vec{f} = 0 \quad (34.279)$$

qui est "l'équation de la statique des fluides" ou la "loi fondamentale de l'hydrostatique".

Remarque: Les viscosités disparaissant, la statique des fluides est la même pour les fluides visqueux ou non visqueux.

2.4.4. NOMBRE DE REYNOLDS

Considérons d'abord, pour simplifier, le cas incompressible.

L'équation de continuité, ou de conservation de la masse, (cf. chapitre de Thermodynamique):

$$\frac{\partial \rho}{\partial t} - \vec{\nabla} \circ (\rho \cdot \vec{v}_a) = 0 \quad (34.280)$$

s'écrit alors dans ce cas particulier:

$$\rho \vec{\nabla} \circ \vec{v}_a = 0 \Rightarrow \vec{\nabla} \circ \vec{v}_a = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z} = 0 \quad (34.281)$$

Nous choisissons maintenant plusieurs grandeurs de références sans dimensions notées par un indice r telles que:

$$\begin{aligned} x' = \frac{x}{L_r}, y' = \frac{y}{L_r}, z' = \frac{z}{L_r} \quad \text{et} \quad \rho' = \frac{\rho}{\rho_r}, t' = \frac{t}{t_r} \\ v_x' = \frac{v_x}{v_r}, v_y' = \frac{v_y}{v_r}, v_z' = \frac{v_z}{v_r} \quad f_i' = \frac{f_i}{f_r}, p' = \frac{p}{p_r}, \mu_d' = \frac{\mu_d}{\mu_r} \end{aligned} \quad (34.282)$$

De par ces définitions, nous avons par exemple:

$$\frac{\partial (v_r v_x')}{\partial (L_r x')} = \frac{v_r}{L_r} \frac{\partial v_x'}{\partial x'} \quad (34.283)$$

donc l'équation des déformations par unité de temps devient:

$$\rho' \vec{\nabla}' \circ \vec{v}' = \frac{v_r}{L_r} \left(\frac{\partial v_x'}{\partial x'} + \frac{\partial v_y'}{\partial y'} + \frac{\partial v_z'}{\partial z'} \right) = 0 \quad (34.284)$$

Mais nous avons également:

$$\frac{d\vec{v}}{dt} = \frac{\partial \vec{v}}{\partial x} \frac{dx}{dt} + \frac{\partial \vec{v}}{\partial y} \frac{dy}{dt} + \frac{\partial \vec{v}}{\partial z} \frac{dz}{dt} + \frac{\partial \vec{v}}{\partial t} \quad (34.285)$$

Restreignons-nous à l'étude d'une composante seulement:

$$\frac{dv_x}{dt} = \frac{\partial v_x}{\partial x} \frac{dx}{dt} + \frac{\partial v_x}{\partial y} \frac{dy}{dt} + \frac{\partial v_x}{\partial z} \frac{dz}{dt} + \frac{\partial v_x}{\partial t} \quad (34.286)$$

En multipliant cette dernière relation par la densité ρ et par définition de la vitesse:

$$\rho \frac{dv_x}{dt} = \rho \left[\frac{\partial v_x}{\partial x} \frac{dx}{dt} + \frac{\partial v_x}{\partial y} \frac{dy}{dt} + \frac{\partial v_x}{\partial z} \frac{dz}{dt} + \frac{\partial v_x}{\partial t} \right] = \rho \left[\frac{\partial v_x}{\partial x} v_x + \frac{\partial v_x}{\partial y} v_y + \frac{\partial v_x}{\partial z} v_z + \frac{\partial v_x}{\partial t} \right] \quad (34.287)$$

Reprenons maintenant une des formulations possible de l'équation de Navier-Stokes démontrée plus haut:

$$\rho d_t v_x = -\partial_x p + \frac{1}{3} \mu_d \partial_x (\vec{\nabla} \circ \vec{v}) + \mu_d \vec{\nabla} \circ (\vec{\nabla} (v_x)) + \rho f_x \quad (34.288)$$

En n'oubliant pas que pour un fluide incompressible nous avons:

$$\vec{\nabla} \circ \vec{v} = 0 \quad (34.289)$$

l'équation de Navier-Stokes précédente, se réduit à:

$$\rho \frac{dv_x}{dt} = -\frac{\partial p}{\partial x} + \mu_d \left(\frac{\partial}{\partial x} \frac{\partial}{\partial x} v_x + \frac{\partial}{\partial y} \frac{\partial}{\partial y} v_x + \frac{\partial}{\partial z} \frac{\partial}{\partial z} v_x \right) + \rho f_x \quad (34.290)$$

Or, pour un fluide nous avons supposé plus haut que les tensions de cisaillement étaient données par:

$$\tau_{xy} = \mu_d \partial_x v_y \quad (34.291)$$

Les termes où apparaissent les coefficients de viscosité peuvent être réécrits tels que:

$$\rho \frac{dv_x}{dt} = -\frac{\partial p}{\partial x} + \left(\frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{yx}}{\partial y} + \frac{\partial \tau_{zx}}{\partial z} \right) + \rho f_x \quad (34.292)$$

Ainsi par correspondance:

$$\rho \left[\frac{\partial v_x}{\partial x} v_x + \frac{\partial v_x}{\partial y} v_y + \frac{\partial v_x}{\partial z} v_z + \frac{\partial v_x}{\partial t} \right] = -\frac{\partial p}{\partial x} + \left(\frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{yx}}{\partial y} + \frac{\partial \tau_{zx}}{\partial z} \right) + \rho f_x \quad (34.293)$$

que nous pouvons écrire sous forme encore plus condensée en utilisant un peu abusivement la notation tensorielle:

$$\rho \frac{dv_x}{dt} = -\frac{\partial p}{\partial x} + \frac{\partial \tau_{ij}}{\partial x_i} + \rho f_x \quad (34.294)$$

En introduisant les variables adimensionnelles:

$$\rho' \rho_r \left[\frac{v_r^2}{L_r} \left(\frac{\partial v'_x}{\partial x'} v'_x + \frac{\partial v'_y}{\partial y'} v'_y + \frac{\partial v'_z}{\partial z'} v'_z \right) + \frac{v_r}{t_r} \frac{\partial v'_x}{\partial t'} \right] = -\frac{p_r}{L_r} \frac{\partial p'}{\partial x'} + \frac{v_r \mu_r}{L_r} \frac{1}{L_r} \frac{\partial \tau'_{ij}}{\partial x'_i} + \rho_r \rho' f'_r f'_x \quad (34.295)$$

Maintenant, multiplions cette dernière relation par $\frac{L_r}{v_r^2}$ et divisons-la par ρ_r des deux côtés de

l'égalité tel qu'elle devienne:

$$\rho' \left[\left(\frac{\partial v'_x}{\partial x'} v'_x + \frac{\partial v'_y}{\partial y'} v'_y + \frac{\partial v'_z}{\partial z'} v'_z \right) + \frac{L_r}{t_r v_r} \frac{\partial v'_x}{\partial t'} \right] = -\frac{p_r}{\rho_r v_r^2} \frac{\partial p'}{\partial x'} + \frac{\mu_r}{\rho_r L_r v_r} \frac{\partial \tau'_{ij}}{\partial x'_i} + \rho' \frac{L_r}{v_r^2} f'_r f'_x \quad (34.296)$$

Au niveau dimensionnel, remarquons que nous avons:

$$t_r = \frac{L_r}{v_r} \quad \frac{p_r}{\rho_r v_r^2} = 1 \quad \frac{L_r}{v_r^2} f'_r = 1 \quad (34.297)$$

Finalement:

$$\rho' \left[\left(\frac{\partial v'_x}{\partial x'} v'_x + \frac{\partial v'_y}{\partial y'} v'_y + \frac{\partial v'_z}{\partial z'} v'_z \right) + \frac{\partial v'_x}{\partial t'} \right] = - \frac{\partial p'}{\partial x'} + \frac{1}{R_e} \frac{\partial \tau'_{xy}}{\partial x'_i} + \rho' f'_x \quad (34.298)$$

Cette équation différentielle exprimée en variables relatives et sans dimensions est appelée "**équation de Navier-Stokes-Reynolds adimensionnelle**"

Le terme R_e , appelé "**nombre de Reynolds**", représente au niveau symbolique le rapport des forces d'inerties sur les forces visqueuses:

$$R_e = \frac{\rho_r L_r v_r}{\mu_r} = \frac{L_r v_r}{\mu_{re}} \quad (34.299)$$

où μ_{re} est la "**viscosité cinématique relative**".

La viscosité dynamique est donc un terme inversement proportionnel à la valeur du nombre de Reynolds.

2.4.5. APPROXIMATION DE BOUSSINESQ

Soit la relation déjà démontrée précédemment:

$$\partial_t \vec{v} + \vec{\nabla} \cdot \left[\frac{v^2}{2} + \frac{p}{\rho} + U \right] + 2\vec{\omega} \times \vec{v} = \vec{0} \quad (34.300)$$

En y remettant le terme contenant la viscosité:

$$\partial_t \vec{v} + \vec{\nabla} \cdot \left[\frac{v^2}{2} + \frac{p}{\rho} + U \right] + 2\vec{\omega} \times \vec{v} - \mu_d \vec{\nabla}^2 \vec{v} = \vec{0} \quad (34.301)$$

sans oublier qu'au niveau des notations (nous savons...c'est un peu embêtant):

$$\Delta \cdot \vec{v} = \vec{\nabla}^2 \vec{v} = \vec{\nabla} \circ (\vec{\nabla} \cdot \vec{v}) = \vec{\nabla} (\vec{\nabla} \circ \vec{v}) - \vec{\nabla} \times (\vec{\nabla} \times \vec{v}) \quad (34.302)$$

Si le potentiel est de type gravitationnel, il va de soi que:

$$\vec{\nabla} \cdot U = \vec{\nabla} \cdot (gz) = \vec{g} \quad (34.303)$$

Donc:

$$\partial_t \vec{v} + \vec{\nabla} \cdot \frac{v^2}{2} + \frac{1}{\rho} \vec{\nabla} \cdot p + \vec{g} + 2\vec{\omega} \times \vec{v} - \mu_d \vec{\nabla}^2 \vec{v} = \vec{0} \quad (34.304)$$

Si l'on peut considérer le contexte de l'expérience telle que la densité volumique est inférieure ou égale à celle de l'eau et que les vitesses sont petites, alors nous pouvons éliminer les termes de second degré, tel que la relation précédente s'écrit:

$$\partial_t \vec{v} + 2\vec{\omega} \times \vec{v} + \frac{1}{\rho} \vec{\nabla} \cdot p + \vec{g} - \mu_t \vec{\nabla}^2 \vec{v} = \vec{0} \quad (34.305)$$

Nous nous plaçons dans le cadre d'un fluide faiblement turbulent, dans lequel la pression et la densité s'écrivent:

$$p = p_0 + p', \rho = \rho_0 + \rho', \frac{\rho'}{\rho_0} \ll 1 \quad (34.306)$$

où p', ρ' représentent le terme d'accroissement turbulent par rapport aux valeurs statiques du fluide.

Nous négligeons également les frottements sur les bords et donc la viscosité en supposant que l'effet des turbulences devient vite prépondérant sur la valeur du frottement.

Donc, nous avons le système d'équations:

$$\partial_t \vec{v} + 2\vec{\omega} \times \vec{v} + \frac{1}{\rho} \vec{\nabla} \cdot p + \vec{g} = \vec{0} \quad (34.307)$$

qui peut s'écrire:

$$\rho \partial_t \vec{v} + \rho 2\vec{\omega} \times \vec{v} + \vec{\nabla} \cdot p + \rho \vec{g} = \vec{0} \quad (34.308)$$

et encore:

$$\rho \partial_t \vec{v} + \rho 2\vec{\omega} \times \vec{v} + \vec{\nabla} \cdot (p_0 + p') + (\rho_0 + \rho') \vec{g} = \vec{0} \quad (34.309)$$

ce qui s'écrit aussi:

$$\rho \partial_t \vec{v} + \rho 2\vec{\omega} \times \vec{v} + (\vec{\nabla} \cdot p' + \rho' \vec{g}) + (\vec{\nabla} \cdot p_0 + \rho_0 \vec{g}) = \vec{0} \quad (34.310)$$

Mais dans le cas statique:

$$\vec{\nabla} \cdot p_0 + \rho_0 \vec{g} = \vec{0} \quad (34.311)$$

Il nous reste donc:

$$\rho \partial_t \vec{v} + \rho 2\vec{\omega} \times \vec{v} + \vec{\nabla} \cdot p' + \rho' \vec{g} = \vec{0} \quad (34.312)$$

En divisant le tout par ρ :

$$\partial_t \vec{v} + 2\vec{\omega} \times \vec{v} + \frac{1}{\rho} \vec{\nabla} \cdot p' + \frac{\rho'}{\rho} \vec{g} = \vec{0} \quad (34.313)$$

mais encore une fois:

$$\partial_t \vec{v} + 2\vec{\omega} \times \vec{v} + \frac{1}{\rho' + \rho_0} \vec{\nabla} \cdot p' + \frac{\rho'}{\rho' + \rho_0} \vec{g} = \vec{0} \quad (34.314)$$

L'approximation de Boussinesq consistant à supposer que le fluide est incompressible et que le système

est à température constante et peu turbulent, nous avons:

$$\rho_0 \gg \rho' \quad (34.315)$$

Ce qui nous donne:

$$\partial_t \vec{v} + 2\vec{\omega} \times \vec{v} + \frac{1}{\rho_0} \vec{\nabla} \cdot p' + \frac{\rho'}{\rho_0} \vec{g} = \vec{0} \quad (34.316)$$

Cette équation s'appelle "**équation de Boussinesq**" et va nous permettre d'introduire la théorie du chaos dans le domaine de la météorologie et des fluides dans le cas particulier des cellules de convection.

2.4.6. LOI DE STOKES

La complexité de l'hydrodynamique est un terrain tout désigné pour l'application de l'analyse dimensionnelle dont nous avons parlé au tout début de notre étude de la mécanique analytique. L'exemple analysé ici montre clairement les possibilités, mais aussi les limites de la méthode.

Nous envisageons un solide de forme quelconque plongé dans un fluide incompressible animé d'une vitesse uniforme à grande distance (le problème est équivalent à celui d'un solide qui se déplace à vitesse constante dans un fluide au repos). Nous cherchons à exprimer la force F qu'exerce le fluide sur l'obstacle, supposé immobile (et notamment dépourvu de tout mouvement de rotation).

La solution analytique est trop complexe pour perdre son temps à résoudre ce genre de problème pratique. Il convient de recourir à l'analyse dimensionnelle.

Les paramètres pertinents sont dans notre étude:

- L la dimension linéaire de l'obstacle
- v la vitesse du fluide à grande distance
- ρ la masse du fluide
- μ le coefficient de viscosité du fluide

Comme il se doit, tous ces paramètres sont des constantes, bien que la vitesse varie en direction et en norme au voisinage de l'obstacle: à grande distance, elle est uniforme et sa valeur v est bien un paramètre pertinent.

Nous pourrions nous demander si la pression ne devrait pas compter au nombre de ces paramètres. Ce n'est pas le cas. La pression est conditionnée par la valeur de la vitesse et par celles des paramètres constants comme nous l'avons vu dans le théorème de Bernoulli. Inutile donc de rajouter un terme redondant.

Sans chercher l'unique combinaison sans dimension des quatre premières, nous appliquons la démarche systématique. Nous voulons déterminer A, B, C, D , tels que:

$$F = L^A v^B \rho^C \mu^D \quad (34.317)$$

Comme:

$$\begin{aligned} F &= \text{kg} \cdot \text{m} \cdot \text{s}^{-2} & v &= \text{m} \cdot \text{s}^{-1} \\ \rho &= \text{kg} \cdot \text{m}^{-3} & \mu &= \text{kg} \cdot \text{m}^{-1} \cdot \text{s}^{-1} \end{aligned} \quad (34.318)$$

Il vient:

$$\text{kg} \cdot \text{m} \cdot \text{s}^{-2} = \text{kg}^{C+D} \cdot \text{m}^{A+B-3C-D} \cdot \text{s}^{-B-D} \quad (34.319)$$

Le système de dimensionnalité s'écrit:

$$\begin{aligned} C + D &= 1 \Rightarrow C = 1 - D \\ A + B - 3C - D &= 1 \\ B + D &= 2 \Rightarrow B = 2 - D \end{aligned} \quad (34.320)$$

Ainsi:

$$A = 1 - B + 3C + D = 2 - D \quad (34.321)$$

Dès lors:

$$F = \rho v^2 L^2 \left(\frac{Lv\rho}{\mu} \right)^{-D} \quad (34.322)$$

et curieusement nous retrouvons ici ce que nous avons vu dans notre développement de l'approximation de Boussinesq:

$$R_e = \frac{\rho_r L_r v_r}{\mu_{rd}} = \frac{L_r v_r}{\mu_{rc}} \quad (34.323)$$

Donc la force exercée par le fluide s'écrit:

$$F = \rho v^2 L^2 R_e^{-D} \quad (34.324)$$

Dans la littérature, nous trouvons la notation:

$$\boxed{F = \rho v^2 L^2 C(R_e)} \quad (34.325)$$

où C dépend de R_e .

Les limites de la méthode analytique dimensionnelle (et même analytique tout court...) apparaît lorsque l'on confronte ce modèle à l'expérience (évidemment nous pourrions faire des modèles numériques de l'équation de Navier-Stokes-Reynolds pour l'ordinateur et ainsi l'honneur serait sauf):

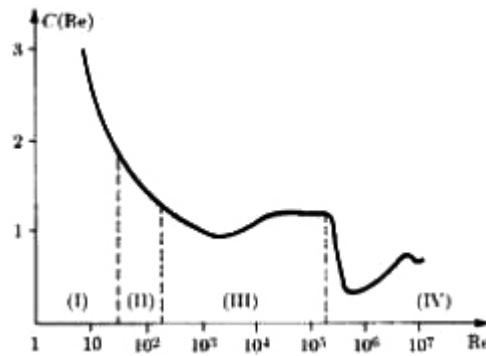


Figure: 34.29 - Illustration du paramètre C pour un cylindre dans différents régimes

Ce graphique correspond à l'écoulement autour d'un cylindre; la vitesse $\vec{v}$ étant perpendiculaire à l'axe du cylindre. Les régimes sont signalés en chiffres romains: stationnaire (I), périodique laminaire (II), turbulent avec superposition d'état périodique (III), turbulent (IV).

La courbe a deux caractéristiques remarquables:

1. Elle a été obtenue en modifiant de manière indépendante les valeurs des quatre paramètres. Nous constatons que C ne dépend que du seul nombre sans dimension Re : c'est un succès de l'analyse dimensionnelle.
2. Il est vain d'espérer trouver une fonction analytique simple qui reproduise la courbe expérimentale. Il faut donc aller voir de plus près les divers régimes correspondants à cette courbe complexe.

La figure ci-dessous schématise l'écoulement d'un fluide visqueux autour d'un cylindre pour différentes valeurs du nombre de Reynolds:

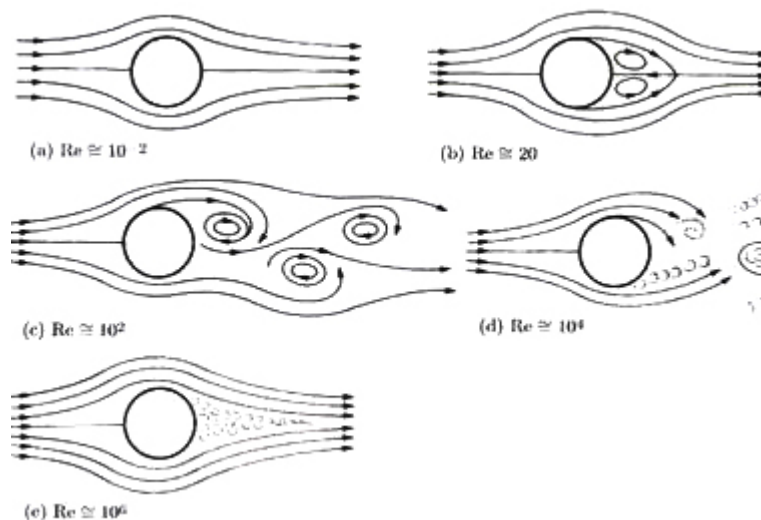


Figure: 34.30 - Écoulements autour d'un cylindre en fonction de différentes valeur du nombre de Reynolds

Le régime correspondant à la figure (a) est dit "**régime stationnaire**". Nous pouvons parler d'un déplacement "quasi-statique" de la part du fluide où en chaque point l'accélération est négligeable. Nous devons donc nous attendre à ce que l'inertie du fluide n'intervienne pas dans l'expression de la force. Pour cela, il faut et il suffit que:

$$\lim_{Re \rightarrow 0} C(Re) \rightarrow C' \frac{1}{Re} = C' \frac{\mu}{L v \rho} \quad (34.326)$$

où C est indépendant de R_e .

Nous avons donc:

$$F = \rho v^2 L^2 C' \frac{\mu}{L v \rho} = C' v L \mu \quad (34.327)$$

Le paramètre C' sans dimensions ne peut dépendre que de la géométrie de l'obstacle. Dans le cas où l'obstacle est sphérique (cas très important en physique avec $L=R$), C' a été déterminé expérimentalement comme valant 6π tel que:

$$F = 6\pi\mu R v \quad (34.328)$$

connue sous le nom de "**loi de Stokes**" ou "**formule de Stokes**". Attention.... cette loi ne s'applique bien que pour les petites vitesses et des petites sphères.

Dans le régime décrit par (b), deux tourbillons s'installent symétriquement derrière le cylindre. Quand R_e augmente au-delà de 40, nous distinguons l'allée de "**tourbillons de von Kármán**".

2.5. PRESSION HYDROSTATIQUE

Nous avons précédemment démontré sans mal que:

$$P = -z \rho g - \frac{1}{2} \rho v^2 - c^{\text{te}} \quad (34.329)$$

Si la vitesse du fluide est nulle:

$$P = -z \rho g - c^{\text{te}} \quad (34.330)$$

Ce qui donne sous forme différentielle:

$$\frac{dP}{dz} = -\rho g \quad (34.331)$$

Si nous mesurons la pression du liquide à partir de sa face supérieure z_0 :

$$\int_{z_0}^z dP = P = - \int_{z_0}^z \rho g dz = -(z \rho g - z_0 \rho g) + c^{\text{te}} \quad (34.332)$$

Si nous prenons z_0 comme référence, nous pouvons poser que:

$$h = z_0 - z \quad (34.333)$$

d'où:

$$P = c^{\text{te}} + \rho g h \quad (34.334)$$

Si nous nous trouvons dans le cas d'un récipient rempli d'un fluide en contact avec l'atmosphère, pour calculer la pression dans ce fluide à une hauteur h donnée, il faudrait prendre en considération la

pression atmosphérique P_0 qui "s'appuie" également sur le fluide. Ainsi, la "pression hydrostatique" est donnée par:

$$P = P_0 + h \cdot \rho \cdot g \quad (34.335)$$

Conséquence: dans un liquide au repos, homogène, les équipotentielles gravifiques sont confondues avec les surface isobares. Sans quoi, il y aurait mouvement transversal.

2.6. POUSSÉE D'ARCHIMÈDE

La poussée d'Archimède, phénomène mondialement connu..., est souvent rebelle à l'intuition première. Au fait, nous avons trop tendance dans les écoles à poser la poussée d'Archimède comme un "principe" et ce à tort puisqu'une simple analyse mathématique suffit à la démontrer.

Si nous isolons une portion ΔV arbitraire d'un fluide en équilibre statique, les conditions de cet équilibre s'écrivent nécessairement (sinon quoi le volume se dissocie et n'est plus en équilibre statique):

$$\vec{P} + \oint (\vec{n} \cdot p) dS = \vec{0} \quad (34.336)$$

$\vec{P}$ désigne le poids ($\vec{P} = m \cdot \vec{a}$ en première approximation...) de ΔV alors que le terme $\oint (\vec{n} \cdot p) dS$ décrit la résultante des forces de pression exercée sur la surface de ΔV .

Chaque élément de surface dS subit donc une force:

$$\vec{F}_{\rightarrow dS} = (\vec{n} \cdot p) dS \quad (34.337)$$

où p est la pression qui s'exerce localement sur dS . Quant à $\vec{n}$, il s'agit d'un vecteur unité dirigé normalement (à la perpendiculaire) à dS et vers l'intérieur de ΔV . La résultante de toutes ces forces se note historiquement de la façon suivante:

$$\vec{F}_A = \oint (\vec{n} \cdot p) dS \quad (34.338)$$

qui exprime donc, comme vous le devinez, la fameuse "poussée d'Archimède" que le reste du fluide exerce sur l'élément. L'intégrale porte sur toute la surface (cette surface est fermée, d'où l'intégrale curviligne correspondante) de l'élément ΔV .

La condition d'équilibre impose donc que:

$$\vec{F}_A = -\vec{P} \quad (34.339)$$

Nous comprenons aisément que $\vec{F}_A$ soit dirigé vers le haut: sous l'effet du champ gravitationnel et donc p augmente avec la profondeur.

Si nous remplaçons le fluide contenu dans le volume par un objet fluide ou solide quelconque mais qui occupe le même volume, la poussée d'Archimède n'est pas modifiée. À cause de la relation $\vec{F}_A = -\vec{P}$ nous avons coutume de dire qu'elle est équivalente au poids du fluide déplacé.

Dans le cas où la direction et l'intensité dans le temps de $\vec{g}$ sont uniformes et constantes nous pouvons

écrire:

$$F_A = -P = \int_V \rho \cdot g dV = \rho \cdot g \cdot V \quad (34.340)$$

et nous retrouvons la relation de la "**loi d'Archimède**" bien connue de tous les écoliers:

$$F_A = \rho \cdot g \cdot V \quad (34.341)$$

Il existe une autre possibilité pour arriver à cette démonstration qui demande moins d'outils mathématiques et qui est donc plus abordable:

Considérons un cylindre de volume V plongé dans un liquide à la verticale. Les composantes horizontales des forces de pression s'annulent, mais la composante verticale au sommet du cylindre P_1 (proche de la surface) est inférieure en intensité (sauf cause extérieure) à celle se trouvant à sa base P_2 . Nous pouvons donc écrire:

$$F_A = (P_2 - P_1) \cdot S = (h_2 \cdot \rho \cdot g - h_1 \cdot \rho \cdot g) \cdot S = \Delta h \cdot S \cdot \rho \cdot g = V \cdot \rho \cdot g \quad (34.342)$$

C'est un peu plus simple et ça tient en une ligne sans intégrales...

Il convient de se rappeler que la poussée d'Archimède est une force qui s'applique à des fluides et donc aussi à des gaz. C'est ainsi grâce à la poussée d'Archimède qu'une montgolfière ou un dirigeable peuvent s'élever dans les airs (dans les deux cas, un gaz de masse volumique plus faible que l'air est utilisé, que ce soit de l'air chauffé ou de l'hélium).

Il est aussi amusant, après démonstration de la loi des gaz parfaits (voir plus loin), de déterminer la pression que devrait avoir notre atmosphère pour avoir la même densité que l'eau et qu'un humain puisse ensuite flotter dans l'air...

2.7. VITESSE DU SON DANS UN LIQUIDE

Intéressons-nous un petit moment au calcul de la vitesse du son dans un liquide. Nous avons démontré dans le cas de notre étude des ondes sonores longitudinales du chapitre de Musique Mathématique que:

$$\frac{dP}{d\rho} = -\gamma \frac{P}{\rho_0} \quad (34.343)$$

où pour rappel, γ est le "**coefficient de Laplace**", appelé aussi "**coefficient adiabatique**" défini par:

$$\gamma = \frac{C_P}{C_V} \quad (33.344)$$

et nous avons démontré que la vitesse de l'onde sonore était donnée par:

$$v = \sqrt{\frac{\gamma P}{\rho_0}} \quad (34.345)$$

En combinant il vient:

$$v = \sqrt{\frac{1}{-\frac{d\rho}{dP}}} = \sqrt{\frac{1}{d\left(\frac{m}{V}\right)}} = \sqrt{\frac{1}{-\frac{m}{V^2}dV}} = \sqrt{\frac{1}{\frac{m}{V} \frac{\Delta V}{\Delta P}}} = \sqrt{\frac{1}{\rho \frac{V}{\Delta P}}} = \sqrt{\frac{B}{\rho}} \quad (34.346)$$

La fraction:

$$B = - \frac{\Delta P}{\frac{\Delta V}{V}}$$

c'est-à-dire le rapport entre une variation de pression et la variation relative de volume qu'elle entraîne reçoit le nom de "**module d'élasticité volumique**". Remarquez qu'il faut le signe - pour que B soit positif: quand la pression augmente, le volume diminue.

Nous avons alors par exemple pour l'eau:

$$v = \sqrt{\frac{1}{\rho \frac{V}{\Delta P}}} \cong \sqrt{\frac{1}{\frac{1'000}{2 \cdot 10^9}}} = 1'414 \left[m \cdot s^{-1} \right] \quad (34.347)$$

La valeur mesurée étant de $1'441 \left[m \cdot s^{-1} \right]$. Il peut paraître surprenant que la vitesse du son dans un liquide, qui est beaucoup plus difficile à comprimer qu'un gaz soit seulement 5 fois plus grande que dans un gaz. La raison est que la densité d'un liquide est environ mille fois plus élevée que celle d'un gaz. L'une dans l'autre, les deux propriétés se compensent partiellement.

3. GAZ

Les solides ont une forme bien définie et sont difficiles à comprimer. Les liquides peuvent s'écouler librement et leur écoulement est limité par des surfaces autoformées. Les gaz se dilatent librement pour occuper le volume du récipient qui les contient, et ont une densité environ mille fois inférieure à celle des liquides et des solides. Ils conduisent peu la chaleur et l'électricité, sauf si nous les ionisons (formation d'un plasma). Les molécules d'un gaz neutre se déplacent suivant des trajectoires rectilignes qui changent de direction à chaque collision avec une autre molécule. Contrairement aux solides et aux liquides, les interactions entre molécules restent faibles. Les propriétés macroscopiques d'un gaz se déduisent donc directement des propriétés des molécules qui le composent (ou des atomes dans le cas d'un gaz monoatomique).

3.1. TYPES DE GAZ

En théorie des gaz (nous parlons souvent de "théorie cinétique des gaz") nous considérons toujours deux types de gaz neutres:

3.1.1. GAZ PARFAIT

Il s'agit d'un modèle dans lequel nous négligeons les interactions moléculaires du gaz, à l'exception des collisions, et dont le volume propre est négligeable devant le volume du récipient.

Lorsqu'un gaz est à faible pression, les interactions entre ses molécules sont faibles. Ainsi, les propriétés d'un gaz réel à basse pression se rapprochent de celles d'un gaz parfait. Nous pouvons alors décrire le comportement du gaz par "**l'équation d'état des gaz parfaits**" que nous démontrons plus bas lors de notre étude du théorème du Viriel:

$$PV = nRT \quad (34.348)$$

avec n le nombre de moles de gaz, P la pression du gaz, V le volume occupé par les n moles et T la température absolue du gaz. La constante R étant la constante des gaz parfaits.

Cette équation montre que trivialement:

- À température T constante (système "isotherme"), le volume d'une quantité fixée de gaz est inversement proportionnel à sa pression. C'est la "**loi de Boyle-Mariotte**":

$$V = \frac{c^{te}}{P}$$

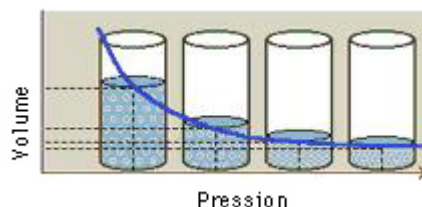


Figure: 34.31 - Illustration de la loi de Boyle-Mariotte

- À pression P constante (système "isobare"), le volume d'une quantité fixée de gaz est proportionnel à la température absolue. C'est la "**loi de Gay-Lussac**" (dans le cas des gaz parfaits...):

$$V = c^{te} T$$

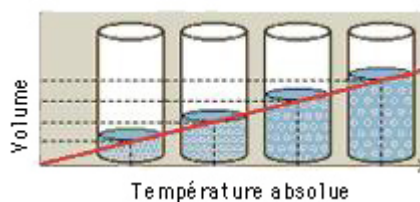


Figure: 34.32 - Illustration de la loi de Gay-Lussac

C'est cette relation qui est souvent utilisée dans les labos des petites classes pour montrer qu'avec une extrapolation de la droite mesurée, le volume devient théoriquement... nul à une température de -273.15 [°C]. Non sérieusement à partir d'une certaine température, il faut utiliser des modèles quantiques et de plus cette relation n'est valable vraiment que pour les gaz (ainsi, lorsque la vapeur d'eau devient liquide... ce n'est plus valable).

- À volume V constant (système "isochore"), la pression d'une quantité fixée de gaz est proportionnelle à sa température absolue. C'est la "**loi de Charles**":

$$P = c^{te} T$$

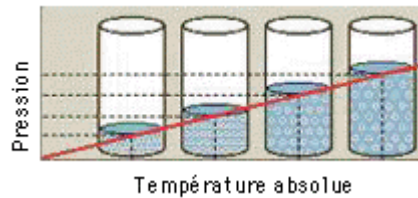


Figure: 34.33 - Illustration de la loi de Charles

Par la suite, Amedeo Avogadro, à qui l'on doit le vocabulaire de "molécule" affirme le concept moléculaire des gaz et conclut que des volumes égaux de gaz différents, pris dans les mêmes conditions de température et de pression, contiennent le même nombre de molécules.

3.1.2. GAZ RÉEL

L'équation d'état des gaz parfaits est approximative. Par exemple, un gaz parfait ne pourrait ni se liquéfier ni se solidifier, quels que soient le refroidissement et la compression auxquels il est soumis. Les gaz réels, surtout dans des conditions de pression et de température proches de la transition à l'état liquide, peuvent présenter des écarts considérables avec la loi des gaz parfaits!

Il faut donc l'adapter aux cas réels. Pour cela, nous utilisons l'équation d'état de Van der Waals qui est particulièrement utile et bien connue. Elle peut être obtenue de manière qualitative une fois l'équation des gaz parfaits démontrée et est alors donnée par (voir plus bas comment nous l'obtenons):

$$\left(P + \frac{a}{V^2} \right) (V - b) = RT \quad (34.349)$$

pour une mole, a et b étant des paramètres adaptables déterminés par des mesures expérimentales effectuées sur le gaz concerné. Ce sont des paramètres qui varient d'un gaz à un autre.

L'équation de Van der Waals peut également être interprétée au niveau microscopique. Les molécules interagissent les unes avec les autres. Cette interaction est fortement répulsive pour les molécules proches les unes des autres, devient légèrement attractive pour un éloignement moyen et disparaît lorsque l'éloignement est important. À pression élevée, la loi des gaz parfaits doit être rectifiée pour prendre en compte les forces attractives ou répulsives.

3.2. THÉORÈME DU VIRIEL

Nous allons ici aborder une étude des gaz parfaits via une méthode particulière. Elle permet d'obtenir un résultat intéressant et particulièrement pour l'astrophysique (cf. chapitre d'Astrophysique). Le théorème du Viriel permet également d'obtenir d'autres résultats très intéressants mais qui pédagogiquement sont un peu difficiles d'accès. Le lecteur qui serait intéressé par cette deuxième partie de résultats pourra directement se reporter un peu plus loin où les concepts de pression et de température cinétique sont traités.

Par définition, "l'expression du Viriel" V_{vir} d'un point matériel est le scalaire:

$$V_{\text{vir}} = \vec{F} \circ \vec{r} \quad (34.350)$$

Par définition, le "Viriel" V_{vir} d'un système composé de N points matériels est:

$$V_{ir.} = \sum \vec{F}_i \circ \vec{r}_i \quad (34.351)$$

Soumis à une force centrale, le Viriel s'écrit (par les propriétés du produit scalaire):

$$V_{ir.} = \sum F_i r_i \quad (34.352)$$

Le "**théorème du Viriel**" s'énonce ainsi: Pour un système en équilibre (!), l'énergie interne est égale à l'opposé de son demi-Viriel total lorsque toutes les particules sont repérées par rapport à son centre de masse.

Démonstration:

Soit la relation mathématique:

$$d_t r_i^2 = 2 r_i d_t r_i \quad (34.353)$$

Sa dérivée seconde:

$$d_t^2 r_i^2 = d_t (2 r_i d_t r_i) = 2 d_t r_i d_t r_i + 2 r_i d_t^2 r_i = 2 (d_t r_i)^2 + 2 r_i d_t^2 r_i \quad (34.354)$$

En multipliant par m_i et en sommant sur i :

$$d_t^2 \sum m_i r_i^2 = 2 \sum [m_i (d_t r_i)^2 + \vec{r}_i m_i d_t^2 r_i] \quad (34.355)$$

Or:

$$m_i d_t^2 r_i = F_i \quad (34.356)$$

et:

$$d_t r_i = v \Rightarrow (d_t r_i)^2 = v_i^2 \quad (34.357)$$

Donc:

$$d_t^2 \sum m_i r_i^2 = 2 \left[\sum m_i v_i^2 + \sum r_i F_i \right] \quad (34.358)$$

Cette dernière expression est valable quelle que soit la position d'un système de coordonnées adopté. Cependant, il est intéressant de placer son origine au centre de masse du système car nous ne sommes plus dépendants de son mouvement.

Si le système est en équilibre, les quantités macroscopiques qui la caractérisent ne sont pas dépendantes du temps. Nous en concluons alors que la somme de n'importe quelle quantité attachée à n'importe quel point matériel du système est en fait une quantité dudit système.

Ainsi, $\sum m_i r_i$ est une quantité macroscopique indépendante du temps. Cela implique que:

$$\sum m_i v_i^2 + \sum r_i F_i = 0 \quad (34.359)$$

Ce qui s'écrit encore (nous multiplions par 1/2 des deux côtés):

$$\frac{1}{2} \sum m_i v_i^2 = -\frac{1}{2} \sum r_i F_i \quad (34.360)$$

Nous avons donc finalement:

$$E_c = -\frac{1}{2} V_{ir}. \quad (34.361)$$

Cette expression de l'énergie cinétique est connue sous le nom de "**théorème de Viriel**" et le membre de droite est donc appelé le "**Viriel du système**".

□C.Q.F.D.

Nous noterons que:

$$E_c = N \cdot \langle E_c \rangle \quad (34.362)$$

où E_c est l'énergie cinétique totale associée à l'ensemble des points matériels du système. Nous l'appelons "**l'énergie interne du système**" et les thermodynamiciens la note souvent avec la lettre U . $\langle E_c \rangle$ est l'énergie d'un point matériel quelconque du système.

Il est possible de retrouver l'expression du Viriel à partir d'un système de particules (nuage en accrétion). Strictement, l'équilibre n'existe pas dans un tel cas. Néanmoins, nous pouvons admettre que si la contraction gravitationnelle est suffisamment lente alors ses différentes phases peuvent être considérées comme une succession d'états d'équilibre.

Dans le cas d'une force centrale et dérivant d'un potentiel, nous pouvons écrire:

$$F = -\frac{\partial E_p}{\partial r} \quad (34.363)$$

et donc:

$$F \cdot r = -r \frac{\partial E_p}{\partial r} \quad (34.364)$$

Si l'énergie potentielle est de la forme k/r (ce qui est le cas pour le potentiel électrique et gravitationnel) alors il vient:

$$F \cdot r = E_p \quad (34.365)$$

et il reste:

$$E_c = -\frac{1}{2} E_p \quad (34.366)$$

En résumé, le théorème du Viriel nous donne une relation entre les énergies cinétiques et potentielles totales. Pour être valable, le mobile doit décrire une trajectoire autour du centre de force central et rester indéfiniment dans un volume fini (état lié). Ce type de raisonnement est applicable à un très

grand nombre de phénomènes, depuis la structure de certaines galaxies jusqu'au dégagement d'énergie dans les explosions nucléaires en passant par l'étude du Soleil et le comportement des gaz réels. Il s'agit du premier résultat qui nous intéressait.

Dans un système gazeux, l'énergie potentielle peut s'écrire comme la somme de l'énergie des forces agissant de l'extérieur plus celles qui sont internes même au gaz. Tel que:

$$E_p = \sum \vec{F}_i \circ \vec{r}_i = \sum \vec{F}_{i,int} \circ \vec{r}_i + \sum \vec{F}_{i,ext} \circ \vec{r}_i \quad (34.367)$$

Or les forces internes peuvent s'écrire comme:

$$\vec{F}_{i,int} = \sum \vec{F}_{ij} \quad (34.368)$$

Il ne faut pas dans cette somme prendre la force qu'exerce chacune des particules sur elle-même. Tel que:

$$\vec{F}_{i,int} = \sum_{j \neq i} \vec{F}_{ij} \quad (34.369)$$

Ce qui nous donne:

$$E_p = \sum \vec{F}_i \circ \vec{r}_i = \sum \sum_{j \neq i} \vec{F}_{ij} \circ \vec{r}_i + \sum \vec{F}_{i,ext} \circ \vec{r}_i \quad (34.370)$$

Dans la double somme, nous pouvons regrouper les termes deux à deux et utiliser le principe d'action-réaction tel que:

$$\vec{F}_{ij} = -\vec{F}_{ji} \quad (34.371)$$

Pour obtenir:

$$\vec{F}_{ij} \circ \vec{r}_i + \vec{F}_{ji} \circ \vec{r}_j = \vec{F}_{ij} (\vec{r}_i - \vec{r}_j) \stackrel{def}{=} \vec{F}_{ij} \circ \vec{r}_{ij} \quad (34.372)$$

Ce qui finalement nous donne:

$$E_p = \sum \sum_{j \neq i} \vec{F}_{ij} \circ \vec{r}_{ij} + \sum \vec{F}_{i,ext} \circ \vec{r}_i \quad (34.373)$$

Et:

$$E_c = -\frac{1}{2} \left(\sum \sum_{j \neq i} \vec{F}_{ij} \circ \vec{r}_{ij} + \sum \vec{F}_{i,ext} \circ \vec{r}_i \right) \quad (34.374)$$

Le premier terme de droite fait intervenir les forces intérieures (interactions) entre les (paires de) particules et le deuxième terme de droite fait intervenir les forces extérieures.

Considérons maintenant un gaz contenu dans un récipient. Ses molécules ne sont sujettes à des forces extérieures que lorsqu'elles heurtent une paroi et nous imaginons qu'en moyenne cette force est perpendiculaire à la paroi (chocs élastiques).

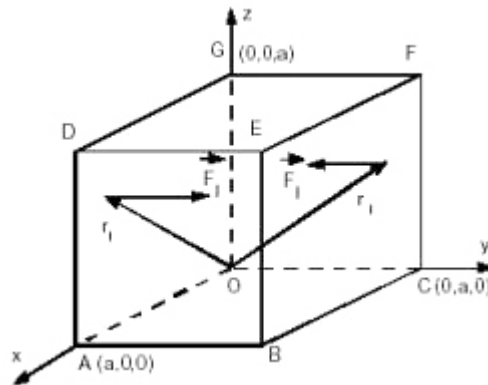


Figure: 34.34 - Gaz contenu dans un récipient

Pour toutes les faces contenues dans les plans définis par les axes, nous avons toujours:

$$\vec{F}_i \circ \vec{F}_i^\dagger = 0 \quad (34.375)$$

Puisqu'en moyenne $\vec{F}_2$ est toujours perpendiculaire à $\vec{r}_2$.

Pour les autres faces ($BCFE$ par exemple) nous avons $F_i = -F_i \vec{\mu}_y$, et donc:

$$\vec{r}_i \circ \vec{F}_i = r_i F_i \cos \alpha = -a F_i \quad (34.376)$$

où nous appelons a la coordonnée selon Oy de l'extrémité de $\vec{r}_1$ (ne pas confondre avec l'accélération!).

Dès lors pour chaque face:

$$\sum \vec{F}_i \circ \vec{r}_i = -a \sum F_i = -aF = -Pa^3 = -PV \quad (34.377)$$

Effectivement car la pression interne du système sur les parois est définie par:

$$P = \frac{F}{S} = \frac{F}{a^2} \quad (34.378)$$

Par le théorème du Viriel, en ajoutant les contributions non nulles des faces $BCFE$, $DEFG$ et $ABED$, il vient:

$$E_c = \frac{3}{2}PV - \frac{1}{2} \sum_{\vec{a}, \vec{b}} \sum_{\vec{c}} \vec{F}_{\vec{a}, \vec{b}}^{\vec{c}} \circ \vec{F}_{\vec{a}, \vec{b}}^{\vec{c}} \quad (34.379)$$

Si l'énergie cinétique moyenne d'une molécule est:

$$E_c = \frac{1}{2}mv^2 = \frac{3}{2}kT \quad (34.380)$$

L'énergie cinétique moyenne totale pour N molécules est alors (nous reviendrons sur cette relation plus loin mais avec une autre approche):

$$E_c = \frac{N}{2} m v^2 = \frac{3}{2} k N T = \frac{3}{2} \frac{R}{N_A} N T \quad (34.381)$$

ce que les thermodynamiciens notent souvent:

$$U = n \cdot \frac{ddl}{2} RT \quad (34.382)$$

où U est donc l'énergie interne du gaz et ddl le nombre de degrés de liberté des constituants.

La relation antécédente, appelée "**théorème d'équipartition de l'énergie**", est importante car elle permet:

1. Une interprétation microscopique de température T et de déterminer l'énergie interne d'un gaz parfait monoatomique (et par extension d'autres gaz avec des degrés de liberté autres).
2. De constater que le système de température en Celsius n'est pas adapté à la réalité physique. Effectivement, dans le système utilisant les Celsius, à 0 [°C] tout devrait être immobile (énergie cinétique nulle) or il est évident que ce n'est pas le cas pour toutes les substances. Donc il faut introduire une nouvelle température qui met en adéquation l'énergie cinétique mesurée et la température traditionnelle. Il s'agit de la "**température absolue**" mesurée en Kelvin [K] dont l'équivalence énergie cinétique/température mesurée est telle que le 0 [°C] correspond à 273.15 [K].

L'énergie interne est une contribution à l'énergie qui n'apparaît pas en mécanique classique. Du point de vue macroscopique, un récipient immobile qui contient un fluide ne possède pas d'énergie cinétique, alors que son énergie potentielle est constante. Nous pouvons l'ignorer en donnant la valeur zéro à cette constante.

Du point de vue microscopique, les choses changent cependant! Effectivement, les atomes ou molécules du fluide sont en mouvement et interagissent. Il faut leur associer une énergie (l'énergie interne) qui est la somme des contributions relatives à chaque atome.

Dès lors:

$$PV = \frac{1}{3} Nmv^2 + \frac{1}{3} \sum_{j=1}^N \sum_{i=1}^N \vec{F}_{ij} \cdot \vec{r}_{ij} = NkT + \frac{1}{3} \sum_{j=1}^N \sum_{i=1}^N \vec{F}_{ij} \cdot \vec{r}_{ij} \quad (34.383)$$

C'est "**l'équation générale d'état d'un gaz réel**", c'est-à-dire l'équation d'état qui tient compte des interactions entre molécules. Il est intéressant de remarquer que finalement cette relation peut nous permettre de calculer l'énergie du gaz même s'il n'y pas de parois!

Si le gaz est parfait, il n'y a pas d'interactions entre les N molécules (par hypothèse) et alors nous avons "**l'équation des gaz parfaits**" suivante:

$$PV = NkT \quad (34.384)$$

Que nous retrouvons beaucoup plus fréquemment sous la forme:

$$PV = nRT \quad (34.385)$$

si n est exprimé en moles avec R la constante des gaz parfaits.

Si la température est constante, nous retrouvons la "**loi de Boyle-Mariotte**":

$$PV = c^{te} \quad (34.386)$$

Pour arriver à l'équation des gaz parfaits, il est utile de rappeler que nous avons fait trois hypothèses:

H1. Les molécules sont assimilées à des sphères dures dont le diamètre est négligeable devant la distance moyenne qui les sépare. C'est que ce que nous appelons "**l'hypothèse structurale**".

H2. À la limite, et c'est ce que nous avons retenu, si nous considérons les molécules comme ponctuelles, la possibilité d'interaction entre les particules s'annule. Les seules interactions qui subsistent seront les chocs sur les parois du récipient qui contient le gaz. Ces chocs sont parfaitement élastiques de sorte que nous puissions appliquer les lois de conservation de la quantité de mouvement de l'énergie cinétique. C'est ce que nous appelons "**l'hypothèse interactive limite**".

H3. Le gaz est étudié dans un état d'équilibre thermodynamique ce qui se traduit par l'homogénéité des variables intensives et extensives. C'est que ce que nous appelons "**l'hypothèse du chaos moléculaire**".

Dans un cas particulier, si les interactions dérivent d'un potentiel central:

$$\vec{F}_{ij} = - \frac{\partial E_p(r_{ij})}{\partial r_{ij}} \vec{u}_{ij} \quad (34.387)$$

Il vient ainsi:

$$F_{ij}r_{ij} = -r_{ij} \frac{\partial E_p(r_{ij})}{\partial r_{ij}} \quad (34.388)$$

Si en outre l'énergie potentielle est du type (attention de ne pas confondre le paramètre k avec la constante de Boltzmann notée de la même manière!):

$$E_p(r_{ij}) = \frac{k}{r_{ij}} \quad (34.389)$$

nous avons:

$$F_{ij}r_{ij} = -nE_p(r_{ij}) \quad (34.390)$$

et dès lors:

$$PV = NkT + \frac{1}{3} E_p \quad (34.391)$$

où E_p est l'énergie totale moyenne du système.

Au fait, il faut bien prendre garde au fait que nous n'avons pas rigoureusement démontré l'équation des gaz parfaits. Effectivement, lorsque nous avons posé plus haut:

$$E_c = \frac{1}{2} m \bar{v}^2 = \frac{3}{2} kT \quad (34.392)$$

Cela supposait implicitement que l'équation des gaz parfaits était déjà connue (...).

Cette approche de la cinétique des gaz est intéressante car utile en astrophysique. Cependant, ce n'est de loin pas la plus simple dans un cadre scolaire et pédagogiquement. Nous nous proposerons de revenir sur ces mêmes résultats via les concepts de pression et de température cinétique une fois l'équation de Van der Waals déterminée.

En 1875, le savant néerlandais J.D. Van der Waals (1837-1923) essaya donc de remplacer l'équation des gaz parfaits par une relation qui tiendrait compte des forces intermoléculaires et de la taille des molécules. La première correction, et la plus évidente, à l'équation des gaz parfaits:

$$PV = kNT \quad (34.393)$$

est de soustraire le volume des molécules de gaz du volume V . Nous pouvons le faire logiquement en remplaçant V par $V - Nb$ où b est une constante très faible représentant bien évidemment le volume moyen par molécule (il existe des tables pour cela). Donc:

$$P = \frac{kNT}{V - N \cdot b} \quad (34.394)$$

où le terme $N \cdot b$ est communément appelé le "**covolume**".

Pour tenir compte des forces intermoléculaires que nous avons négligées précédemment, nous pouvons tenter une approche approximative en sachant déjà que la force d'attraction de chaque molécule se fera sur $N-1$ molécules. Par conséquent, le numérateur de la force d'attraction contiendra (par la somme de tous les termes) trivialement si le gaz est isotrope et homogène un terme du type $N(N-1)$ pour l'influence de toutes les molécules entre elles ce qui si N est très grand peut être approximé par N^2 .

Nous savons également qu'au numérateur, il y aura un terme de masse pour chaque particule. Si nous connaissons N/V alors il ne reste plus qu'à connaître la densité massique du gaz (mais ce n'est pas une variable extensive donc nous éviterons de la faire apparaître explicitement). Ainsi, le terme N^2 peut s'écrire directement $(N/V)^2$

En suivant ce raisonnement, Van der Waals ajouta au terme de droite de l'équation ci-dessus un terme négatif proportionnel à la quantité $(N/V)^2$. La présence de ce terme se traduit par un abaissement de la pression au fur et à mesure que croît la densité du gaz. La relation modifiée est ainsi:

$$P \cong \frac{kNT}{V - N \cdot b} - a \left(\frac{N}{V} \right)^2 \quad (34.395)$$

où a est une constante de proportionnalité. Nous pouvons réécrire cette relation sous la forme:

$$\left(P + a \left(\frac{N}{V} \right)^2 \right) (V - N \cdot b) \cong kNT \quad (34.396)$$

qui est appelée "**équation d'état de Van der Waals**" ou encore "**équation de gaz réels de Van der Waals**" (on retrouve dans la littérature plusieurs manières équivalentes d'écrire cette dernière relation). Elle est une excellente description de l'équation d'état dans un large domaine des variables P, V, T , les valeurs a et b étant caractéristiques de chaque gaz. Les constantes a et b sont déterminées expérimentalement comme nous en avons déjà fait mention.

Nous pouvons représenter un l'équation de Van der Waals dans un diagramme P - V (donc à température constante). Dont voici un exemple avec le CO_2 (gaz carbonique/dioxyde de carbone):

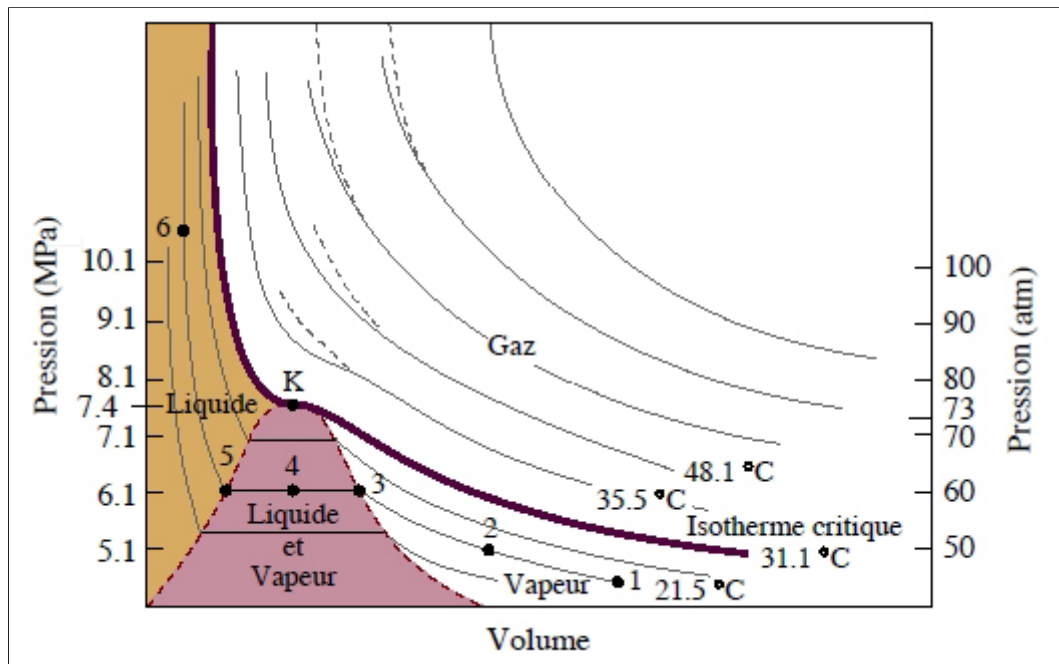


Figure: 34.35 - Diagramme P - V de Van der Waals

Ce qui correspond à:

$$P \cong \frac{kNT}{V - N \cdot b} - a \frac{N^2}{V^2} \quad (34.397)$$

La figure du bas montre le case général. On peut passer de l'état F (vapeur) à l'état I (liquide) en suivant un chemin passant par les états A et E (palier de liquéfaction), on observe alors une transition de phase. Si on suit un autre chemin, par exemple en utilisant l'isotherme GH ($T > T_k$, chemin rouge), il n'y a pas de palier de liquéfaction, on passe de manière continue de l'état gazeux à l'état liquide:

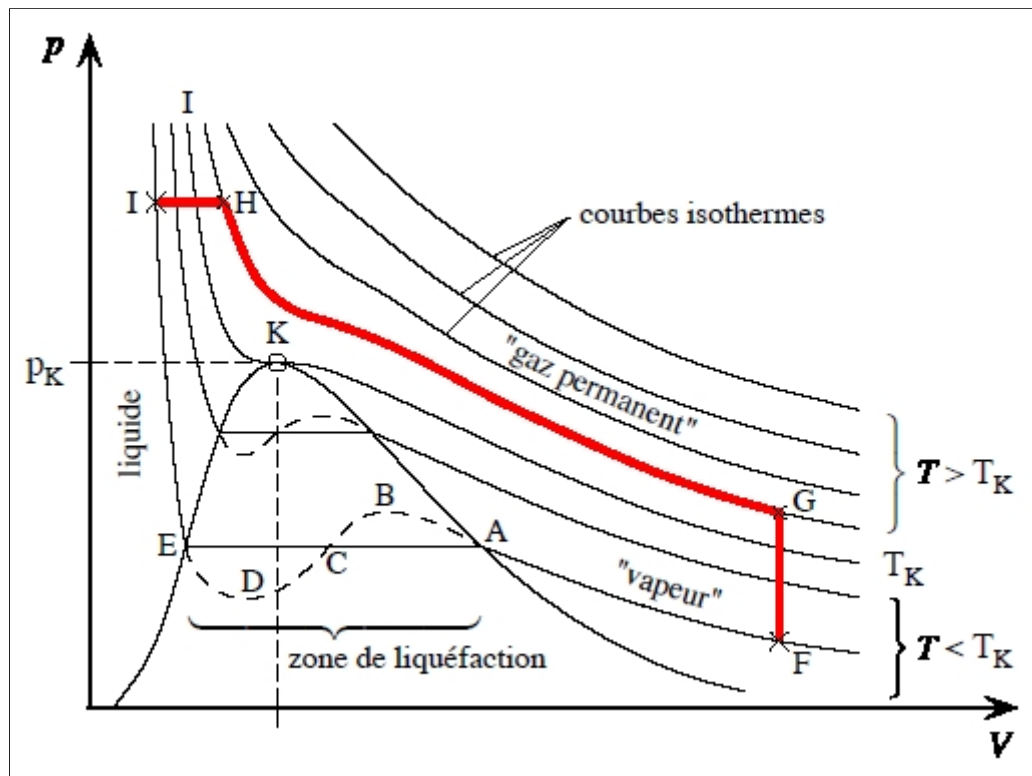


Figure: 34.36 - Diagramme P-V de Van der Waals générique

L'équation de Van der Waals peut être mise sous forme d'un développement du type Viriel en utilisant un développement de Taylor (Maclaurin):

$$\begin{aligned}
 P &\cong \frac{kNT}{V - Nb} - a \left(\frac{N}{V} \right)^2 = \frac{kNT}{V(1 - Nb/V)} - a \left(\frac{N}{V} \right)^2 = \frac{kNT}{V(1 - Nb/V)} - \frac{akNT}{V} \left(\frac{N}{kTV} \right) \\
 &= \frac{kNT}{V} \left(1 + \frac{Nb}{V} + \frac{N^2 b^2}{V^2} + \dots \right) - \frac{akNT}{V} \left(\frac{N}{kTV} \right) \\
 &= \frac{kNT}{V} \left(1 + b \frac{N}{V} + b^2 \frac{N^2}{V^2} + \dots \right) - \frac{akNT}{V} \left(\frac{N}{kTV} \right)
 \end{aligned} \quad (34.398)$$

ou en réorganisant les termes:

$$P \cong \frac{kNT}{V} \left(1 + \left(b - \frac{a}{kT} \right) \frac{N}{V} + b^2 \frac{N^2}{V^2} + \dots \right) \quad (34.399)$$

Remarquons qu'il existe une température pour laquelle $\left(b - \frac{a}{kT} \right)$ est nulle. Nous l'appelons

"température de Boyle" du gaz: c'est la température à laquelle le gaz réel ressemble à un gaz parfait. Il est très délicat d'obtenir ces paramètres expérimentalement.

3.3. PRESSION CINÉTIQUE

Recherchons le nombre de molécules d'un gaz parfait toutes supposées animées d'une vitesse égale v qui viennent frapper une surface S pendant une durée dt .

Si le gaz étudié est dans un état d'équilibre thermodynamique, cela se traduira par l'homogénéité des variables intensives (hypothèse du chaos moléculaire).

Il s'ensuit que la densité des particules est constante:

$$N_v = \frac{N}{V} = c^{te} \quad (34.400)$$

Si nous admettons que les molécules sont en mouvement, nous supposons qu'il y a isotropie des vitesses. En d'autres termes, puisque les vitesses peuvent vectoriellement être décrites dans un système de 3 axes orthogonaux (trois dimensions spatiales), il y a 6 directions possibles primaires au total (2 directions par axe: en avant, en arrière...).

Cela se traduit par l'équivalence entre les différentes directions. Il y a:

$$\frac{N_v}{6} \quad (34.401)$$

particules ayant une vitesse v selon l'une des directions primaires.

Donc pendant une durée dt , la surface S de la paroi n'est percutée que par une partie des molécules contenues dans le volume $S \cdot v \cdot dt$. En effet, seul 1/6 des molécules contenues dans ce volume se dirigent effectivement vers la surface S .

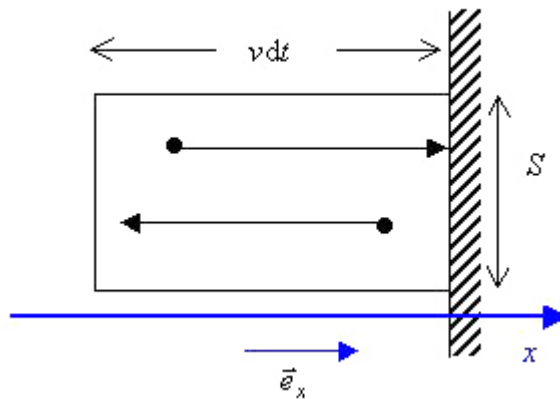


Figure: 34.37 - Une des parois de l'enceinte

Le nombre de molécules qui viennent heurter la paroi pendant la durée dt est donc :

$$\frac{1}{6} N_v \cdot S \cdot v \cdot dt \quad (34.402)$$

Étudions maintenant la dynamique du choc d'une particule sur la paroi :

La particule de masse m qui arrive sur la paroi avec la vitesse $\vec{v}_x$ repart avec une vitesse $-\vec{v}_x$ si le choc est parfaitement élastique.

La variation de la quantité de mouvement de la particule est donc :

$$\Delta \vec{p}_{\text{molécule}} = m(\vec{v}_f - \vec{v}_i) = m((-v_x) - v_x) = -2mv_x \quad (34.403)$$

En vertu de la conservation de la quantité de mouvement, cette quantité de mouvement est transférée à la paroi : la variation de la quantité de mouvement subie par la paroi est l'opposé de celle de la particule :

$$\Delta \vec{p}_{\text{paroi}} = -m(\vec{v}_f - \vec{v}_i) = 2mv_x \quad (34.404)$$

La variation totale de la quantité de mouvement de toutes les particules qui viennent frapper la paroi pendant la durée dt vaut alors en module :

$$dp = (2mv) \frac{1}{6} N_v \cdot S \cdot v \cdot dt \quad (34.405)$$

En appliquant le principe fondamental de la dynamique, nous pouvons passer à la force subie par la paroi pendant la percussion de ces molécules :

$$F = \frac{dp}{dt} \quad (34.406)$$

donc :

$$F = \frac{1}{3} N_v \cdot S \cdot mv^2 \quad (34.407)$$

Or par définition de la pression et en prenant le module :

$$P = \frac{F}{S} \quad (34.408)$$

nous avons alors la "**pression cinétique**" donnée par :

$$P = \frac{1}{3} N_v mv^2 \quad (34.409)$$

La pression cinétique est donc la traduction de la fréquence des chocs sur la paroi. Plus les molécules sont nombreuses (terme en N_v) et rapides (terme en v), plus le nombre de chocs augmente. Cela est conforme à l'expérience et à l'intuition.

3.4. TEMPÉRATURE CINÉTIQUE

En remplaçant N_v par le quotient N/V , l'équation précédente peut se mettre sous la forme :

$$PV = \frac{1}{3} Nmv^2 \quad (34.410)$$

Si nous l'identifions avec l'équation d'état des gaz parfaits vue plus haut :

$$PV = nRT \quad (34.411)$$

il vient :

$$\frac{1}{3} N m v^2 = n R T \quad (34.412)$$

Or la quantité de matière n est égale au rapport du nombre de particules sur le nombre d'Avogadro:

$$n = N / N_{Av}. \quad (34.413)$$

ce qui permet de définir la "**température cinétique**" par :

$$T_C = \frac{1}{3} \frac{N}{n} \frac{m v^2}{R} = \frac{1}{3} \frac{N_{Av.} m v^2}{R} \quad (34.414)$$

Nous pouvons alors introduire une constante qui nous est déjà connue appelée par Max Planck "**constante de Boltzmann**" et telle que:

$$k = \frac{R}{N_{Av.}} = 1.38 \cdot 10^{-23} [J \cdot K^{-1}] \quad (34.415)$$

Remarque: Nous avons déjà fait mention de cette relation lors de notre présentation des constantes universelles dans le chapitre traitant des Principes de la mécanique.

L'expression de température cinétique prend alors la forme :

$$T = \frac{m v^2}{3k} \quad (34.416)$$

Cette relation montre que la température cinétique est le reflet de l'énergie cinétique des particules. D'une façon imagée, c'est l'image de la violence des chocs.

Ainsi, l'énergie d'un gaz parfait se réduit à la somme des énergies cinétiques des particules qui le constituent:

$$U = \sum_i \frac{1}{2} m v_i^2 = \frac{N}{2} m v^2 = \frac{3}{2} N k T = \frac{3}{2} n R T \quad (34.417)$$

Les atomes d'un gaz parfait monoatomique sont assimilables à des points matériels. Leur énergie cinétique est une énergie cinétique de translation dont la valeur moyenne, par atome, s'écrit :

$$E_c = \frac{1}{2} m v^2 = \frac{3}{2} k T \quad (34.418)$$

Dans l'espace des vitesses, toutes les directions sont équivalentes : il y a isotropie de la distribution des vitesses. En coordonnées cartésiennes, il vient :

$$v^2 = v_x^2 + v_y^2 + v_z^2 \quad (34.419)$$

et par suite de l'isotropie:

$$v_x^2 = v_y^2 = v_z^2 = \frac{1}{3}v^2 \quad (34.420)$$

d'où :

$$E_{e,x} = E_{e,y} = E_{e,z} = \frac{1}{2}kT \quad (34.421)$$

ainsi, l'énergie cinétique moyenne par degré de liberté de translation est égale à :

$$\frac{1}{2}kT \quad (34.422)$$

Le lecteur remarquera bien évidemment que nous obtenons ici les mêmes résultats et conclusions que lors de notre étude du théorème du Viriel à la différence que l'approche est ici plus simple et donc plus didactique.

3.5. LIBRE PARCOURS MOYEN

Nous allons voir maintenant un cas d'études très intéressant des gaz qui permet de mettre au clair beaucoup d'incompréhensions dans la vie de tous les jours (fumée dans les restaurants, chaleurs près d'un radiateur, ...). Cependant, les phénomènes sont en réalité plus complexes il faut aussi prendre en compte la diffusion, la convection, etc.

Considérons une molécule, qui se déplace à la vitesse moyenne $\bar{v}$. Sa sphère d'influence V_i balaie alors, pendant l'unité de temps de son déplacement, un volume donné par :

$$V_i = \pi r^2 \bar{v} \quad (34.423)$$

Remarque: Dans le cas où l'on considère un atome ou une molécule comme une sphère d'influence, on parle souvent de "**volume de Van der Waals**" et du "**rayon de Van der Waals**" associé.

Si l'unité de volume renferme n molécules (n a donc les unités d'une densité volumique), le nombre de chocs pendant l'unité de temps dans cette même unité de volume sera alors de :

$$C = n\pi r^2 \bar{v} \quad (34.424)$$

si les autres sphères d'influence étaient immobiles... Donc pour tenir compte du mouvement des autres sphères d'influence, considérons la figure ci-dessous avec trois scénarios simplistes :

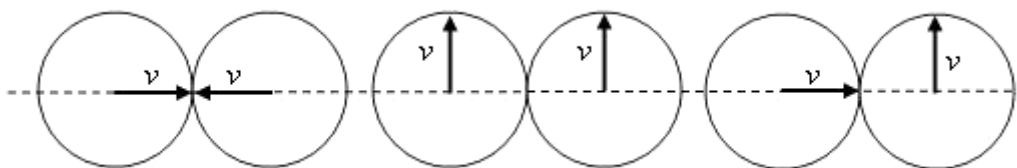


Figure: 34.38 - 3 scénarios de collisions

De gauche à droite nous avons :

1. La vitesse relative des molécules est $2 \cdot \bar{v}$

2. La vitesse relative est nulle

3. La vitesse relative vaut (Pythagore) $\sqrt{2} \cdot \bar{v}$

Les deux premiers cas sont quand même un peu extrêmes... Le troisième sera considéré comme représentant une moyenne et peut servir de nouvelle base au calcul précédent.

Ainsi, en utilisant la vitesse relative du dernier scénario, le nombre de chocs pendant l'unité de temps dans l'unité de volume devient :

$$C = n\pi r^2 \sqrt{2} \bar{v} \quad (34.425)$$

Ainsi, entre deux chocs, une molécule parcourt une distance moyenne :

$$\lambda = \frac{\bar{v}}{C} = \frac{1}{n\pi r^2 \sqrt{2}} \quad (34.426)$$

qu'il est d'usage de noter:

$$\lambda = \frac{\bar{v}}{C} = \frac{1}{n\pi r^2 \sqrt{2}} \quad (34.427)$$

C'est donc l'expression du libre parcours moyen en fonction de la densité moléculaire n (et non du nombre de moles!!!) et du rayon r de la sphère d'influence, paramètre intrinsèque du gaz considéré. Compte tenu de la relation:

$$PV = NkT \Rightarrow P = nkT \Rightarrow n = \frac{P}{kT} \quad (34.428)$$

où n est toujours la densité moléculaire (et non le nombre de moles!!!) nous obtenons:

$$\lambda = \frac{k}{\pi r^2 \sqrt{2}} \frac{T}{P} \quad (34.429)$$

Une application numérique donne pour l'élément aux conditions normales de température et de pression un libre parcours moyen qui vaut pour la majorité des molécules courantes plusieurs milliers de fois le diamètre d'une molécule de taille standard (le libre parcours moyen est donc de l'ordre du micromètre).

En utilisant la relation de la vitesse moyenne la plus probable dans le cadre de l'hypothèse d'une distribution de Maxwell des vitesses (cf. chapitre de Mécanique Statistique) nous avons:

$$\bar{v} = \sqrt{\frac{2kT}{m}} = \sqrt{\frac{2RT}{M_m}} \cong 389 \left[m \cdot s^{-1} \right] \quad (34.430)$$

pour une molécule de masse molaire de 30 [g] à température normale. Ce qui donne un nombre de collisions pour une mole de molécules:

$$C = n\pi r^2 \sqrt{2v} \cong 15 \cdot 10^6 \left[\text{collisions} \cdot \text{s}^{-1} \right] \quad (34.431)$$

Cette fréquence élevée de chocs explique, dans des conditions normales de température et de pression, la rapidité avec laquelle l'équilibre statistique s'établit au sein d'un gaz.

À la température ambiante et pour un vide de $7.5 \cdot 10^{-10} \text{ [Pa]}$, nous obtenons avec les mêmes valeurs : $\lambda \cong 1 \text{ [km]}$.

Les dimensions des récipients contenant les gaz étant toujours inférieures à cet ordre de grandeur, il apparaît que lorsque le vide est réalisé dans une enceinte, les chocs intermoléculaires sont négligeables vis-à-vis des chocs molécules-parois.

4. PLASMAS

Nous définissons le "**plasma**" comme un état de la matière dans lequel certaines liaisons électroniques ont été rompues, provoquant l'apparition d'électrons libres, chargés négativement et d'ions, chargés positivement. Les gaz faiblement ionisés appelés "plasmas" par abus de langage, possèdent les mêmes propriétés mécaniques (écoulements, ondes acoustiques, etc.) que les gaz neutres, en revanche leurs propriétés électromagnétiques (conductivité électrique, indice de réfraction) en diffèrent par suite de la présence d'électrons libres en leur sein.

Remarque: Le plasma est aussi nommé "quatrième état de la matière" (après les états solide, liquide et gazeux et avant le cinquième état de la matière: le condensat de Bose-Einstein).

Dans leur état normal, les gaz sont des isolants électriques. Cela tient au fait qu'ils ne contiennent pas de particules chargées libres, mais seulement des molécules neutres. Cependant, si nous leur appliquons des champs électriques assez intenses, ils deviennent conducteurs. Les phénomènes complexes qui se produisent alors portent le nom de décharges dans les gaz et sont dus à l'apparition d'électrons et d'ions libres.

Le résultat d'une décharge dans un gaz est donc la production d'un gaz ionisé contenant par exemple une densité moyenne de n_e électrons, n_i ions positifs et n_0 neutres (atomes ou molécules). En général, le gaz est macroscopiquement neutre. Nous avons alors:

$$n_e = n_i = n \quad (34.432)$$

ou autrement exprimée, la neutralité s'écrit aussi:

$$q_i n_i + q_e n_e = 0 \quad (34.433)$$

Cette neutralité est la conséquence des forces électrostatiques très intenses qui apparaissent dès que l'on a $n_e \neq n_i$. La densité de particules est donc la première grandeur fondamentale.

Remarque: La neutralité n'est que globale. À une échelle plus fine, les électrons, plus mobiles que les ions, forment un nuage autour de chaque ion.

Le "**degré d'ionisation**" d'un gaz est défini par le rapport:

$$\alpha = \frac{n}{n_0 + n} \quad (34.434)$$

où n_0 est la densité de particules neutres et n celle des électrons (ou des ions positifs). La valeur du degré d'ionisation dans les divers types de gaz ionisés varie en pratique depuis des valeurs très faibles, de l'ordre de 10^{-10} , par exemple, jusqu'à 1.

La deuxième grandeur fondamentale est la température. Lorsque l'on chauffe un gaz à une température T suffisamment élevée (de l'ordre de $10^4 K$), l'énergie moyenne (voir théorème du Viriel):

$$\frac{3}{2} kT \quad (34.435)$$

de translation de ses molécules peut devenir du même ordre que leur énergie d'ionisation E_i . Dans ces conditions, lorsque deux molécules entrent en collision, il peut y avoir ionisation de l'une d'entre elles.

Si le gaz est en équilibre thermodynamique, l'ionisation par collision est contrebalancée par des processus de recombinaison entre électrons et ions et il en résulte que les trois variables α, n et T ne sont pas indépendantes: l'ionisation est déterminée par la pression et la température, nous disons alors que le gaz est en "**état d'équilibre d'ionisation thermique**".

À des températures plus élevées, les atomes du gaz peuvent d'ailleurs s'ioniser plusieurs fois. Dans de nombreux cas, l'ionisation est due à un champ électrique extérieur, et le gaz n'est pas en équilibre thermodynamique. Il atteindra souvent un état stationnaire que l'on pourra caractériser par les paramètres α, n, T_e (température des électrons), T_i (température des ions) et T_0 (température des molécules).

Les trois températures ainsi introduites sont définies par la condition que $\frac{3}{2} kT_a$ représente l'énergie cinétique moyenne des particules d'espèce a , dans un repère où elles ont une vitesse moyenne nulle. L'écart entre T_e, T_i et T_0 peut être important: par exemple, dans un tube à décharges typique, nous pourrions avoir: $T_0 \approx T_i \approx 300 K$ et $T_e \approx 3 \cdot 10^4 K$. La forte valeur de T_e est due à l'action du champ électrique sur les électrons, et l'ionisation est alors produite par les collisions de ces électrons chauds sur les molécules neutres du gaz.

En conclusion il n'y a que deux grandeurs de base permettant de caractériser un plasma: la densité et la température électronique. Nous allons maintenant nous pencher sur deux autres grandeurs importantes mais non fondamentales dans le sens où elles s'expriment à partir de la densité et de la température.

4.1. FRÉQUENCE PLASMA

Si dans un plasma initialement neutre, nous produisons une perturbation locale sous la forme d'un excès de charge électrique positive ou négative, celui-ci va tendre à revenir vers l'état d'équilibre de neutralité. Cependant, nous pouvons voir facilement que la perturbation initiale engendre en général

une oscillation pendulaire non amortie du plasma autour de son état d'équilibre. Considérons par exemple la situation représentée sur la figure ci-dessous.

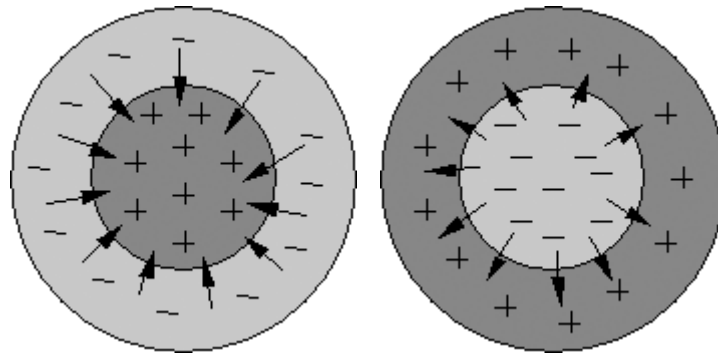


Figure: 34.39 - Oscillation électronique du plasma

À l'instant initial la région au centre contient un déficit d'électrons et la région tout autour un excès d'électrons. Cela produit un champ électrique tendant à créer un mouvement des électrons dans le sens des flèches. Dans ce mouvement, ceux-ci acquerront une certaine énergie cinétique et ils pourront, au bout d'un certain temps, dépasser la position d'équilibre. Un trop grand nombre d'électrons ayant quitté la région externe, il y aura un défaut d'électrons dans cette région et un champ électrique tendant à les ramener vers elle. Au bout d'un certain temps, la situation initiale est reconstituée et le cycle recommence. Les vibrations ainsi produites sont appelées oscillations de plasma électroniques.

Nous pouvons étudier quantitativement ce problème en posant les équations générales d'une oscillation de charge électronique et moyennant les hypothèses simplificatrices suivantes:

H1. Les ions sont supposés immobiles étant donné qu'ils sont beaucoup plus lourds que les électrons, et leur quantité égale à n_{i0}

H2. L'agitation thermique est négligeable

H3. Les collisions sont négligeables

H4. Les oscillations sont de faible amplitude

H5. Il n'y a pas de champ électrique ou magnétique imposé par des sources extérieures

Maintenant, rappelons que nous avons démontré dans le chapitre d'Électrodynamique que (équation de conservation de la charge):

$$\frac{\partial \rho_g}{\partial t} + \vec{\nabla} \cdot \vec{j} = 0 \quad (34.436)$$

et dans le chapitre d'Électrocinétique que:

$$\vec{j} = \rho_g \cdot \vec{v} \quad (34.437)$$

Dès lors, en adoptant les notations susmentionnées, il vient localement:

$$\frac{\partial n_e}{\partial t} + \vec{\nabla} \circ (n_e \cdot \vec{v}_e) = 0 \quad (34.438)$$

relation qui constitue "l'équation hydrodynamique des électrons".

Remarque: Un plasma est certes en théorie globalement neutre, mais nous pouvons avoir en théorie localement un volume non neutre. C'est cette hypothèse qui nous permet de poser que la divergence du courant n'est pas nulle!

Rappelons maintenant la force de Coulomb (cf. chapitre d'Électrostatique):

$$\vec{F}_e = m_e \vec{a}_e = m_e \frac{d\vec{v}_e}{dt} = q_e \vec{E} \quad (34.439)$$

où nous avons donc clairement négligé le terme de pression cinétique et le terme de collision (hypothèses 2 et 3) et négligé le champ magnétique lié à l'oscillation.

Nous pouvons simplifier ces équations en utilisant l'hypothèse 4 sous la forme:

$$n_e(\vec{r}, t) = n_{e0} + n_{e1}(\vec{r}, t) \quad (34.440)$$

où $n_{e1}(\vec{r}, t)$ est une petite perturbation.

Supposons de plus que les quantités variables varient à la pulsation ω , nous pouvons donc écrire:

$$\begin{aligned} E(\vec{r}, t) &= E(\vec{r}) \cdot e^{-i\omega t} \\ n_{e1}(\vec{r}, t) &= n_{e1}(\vec{r}) \cdot e^{-i\omega t} \end{aligned} \quad (34.441)$$

Dès lors, l'équation hydrodynamique des électrons devient:

$$\begin{aligned} \frac{\partial n_e}{\partial t} + \vec{\nabla} \circ (n_e \cdot \vec{v}_e) &= \frac{\partial (n_{e0} + n_{e1})}{\partial t} + \vec{\nabla} \circ ((n_{e0} + n_{e1}) \cdot \vec{v}_e) \\ &= \frac{\partial n_{e1}}{\partial t} + \vec{\nabla} \circ (n_{e0} \cdot \vec{v}_e) + \underbrace{\vec{\nabla} \circ (n_{e1} \cdot \vec{v}_e)}_{\cong 0} = \frac{\partial n_{e1}}{\partial t} + \vec{\nabla} \circ (n_{e0} \cdot \vec{v}_e) \\ &= \frac{\partial n_{e1}}{\partial t} + n_{e0} \vec{\nabla} \circ \vec{v}_e = \frac{\partial n_{e1}(\vec{r}) e^{-i\omega t}}{\partial t} + n_{e0} \vec{\nabla} \circ \vec{v}_e = -i\omega n_{e1}(\vec{r}) e^{-i\omega t} + n_{e0} \vec{\nabla} \circ \vec{v}_e \\ &= -i\omega n_{e1} + n_{e0} \vec{\nabla} \circ \vec{v}_e \end{aligned} \quad (34.442)$$

Soit au final:

$$-i\omega n_{e1} + n_{e0} \vec{\nabla} \circ \vec{v}_e = 0 \quad (34.443)$$

De l'expression de la force de Coulomb nous déduisons également:

$$d\vec{v}_e = \frac{q_e}{m_e} \vec{E} dt \quad (34.444)$$

d'où nous tirons:

$$\vec{v}_e = \frac{q_e}{m_e} \int \vec{E} dt = \frac{q_e}{m_e} \int E(\vec{r}) e^{-i\omega t} dt = \frac{q_e}{m_e} E(\vec{r}) \int e^{-i\omega t} dt = -\frac{q_e}{i\omega m_e} E(\vec{r}) \quad (34.445)$$

Mais nous avons d'autre part la loi de Gauss (cf. chapitre d'Électrodynamique):

$$\vec{\nabla} \circ \vec{E} = \frac{(n_i q_i + n_{e0} q_e + n_{e1} q_e)}{\epsilon_0} = \frac{n_{e1} q_e}{\epsilon_0} \quad (34.446)$$

en effet compte tenu de la condition de neutralité du plasma non perturbé, nous avons:

$$(n_{i0} q_i + n_{e0} q_e) = 0 \quad (34.447)$$

Donc:

$$\vec{\nabla} \circ \vec{E} - \frac{n_{e1} q_e}{\epsilon_0} = 0 \quad (34.448)$$

De la relation démontrée précédemment:

$$\vec{v}_e = \frac{q_e}{-i\omega m_e} \vec{E} \quad (34.449)$$

en remplaçant dans la relation suivante (aussi démontrée précédemment):

$$-i\omega n_{e1} + n_{e0} \vec{\nabla} \circ \vec{v}_e = 0 \quad (34.450)$$

nous avons:

$$n_{e1} = \frac{n_{e0} q_e}{\omega^2 m_e} \cdot \vec{\nabla} \circ \vec{E} \quad (34.451)$$

Finalement en remplaçant cette dernière expression dans la loi de Gauss, nous tirons:

$$\vec{\nabla} \circ \vec{E} - \frac{n_{e0} q_e^2}{\omega^2 m_e \epsilon_0} \vec{\nabla} \circ \vec{E} = \vec{\nabla} \circ \vec{E} \left(1 - \frac{n_{e0} q_e^2}{\omega^2 m_e \epsilon_0} \right) = 0 \quad (34.452)$$

Mais dans les oscillations de charges d'espace nous avons par définition :

$$\vec{\nabla} \circ \vec{E} \neq 0 \quad (34.453)$$

Ce qui nous amène à:

$$\frac{n_{e0} q_e^2}{\omega^2 m_e \epsilon_0} = 1 \Leftrightarrow \omega = \sqrt{\frac{n_{e0} q_e^2}{m_e \epsilon_0}} \quad (34.454)$$

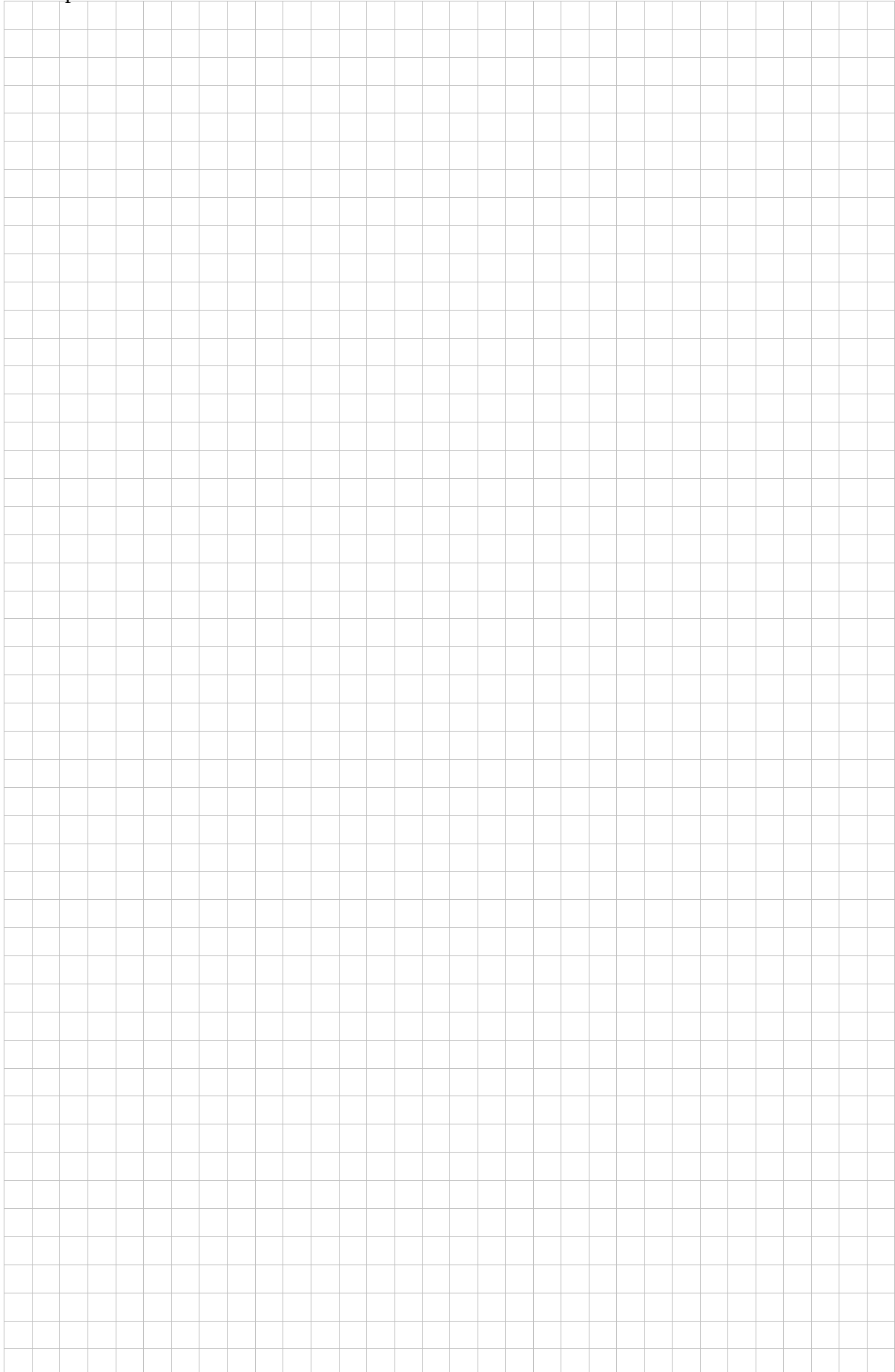
Finalement, nous avons la "**fréquence plasma**" ou encore "**fréquence de Langmuir**" donnée par:

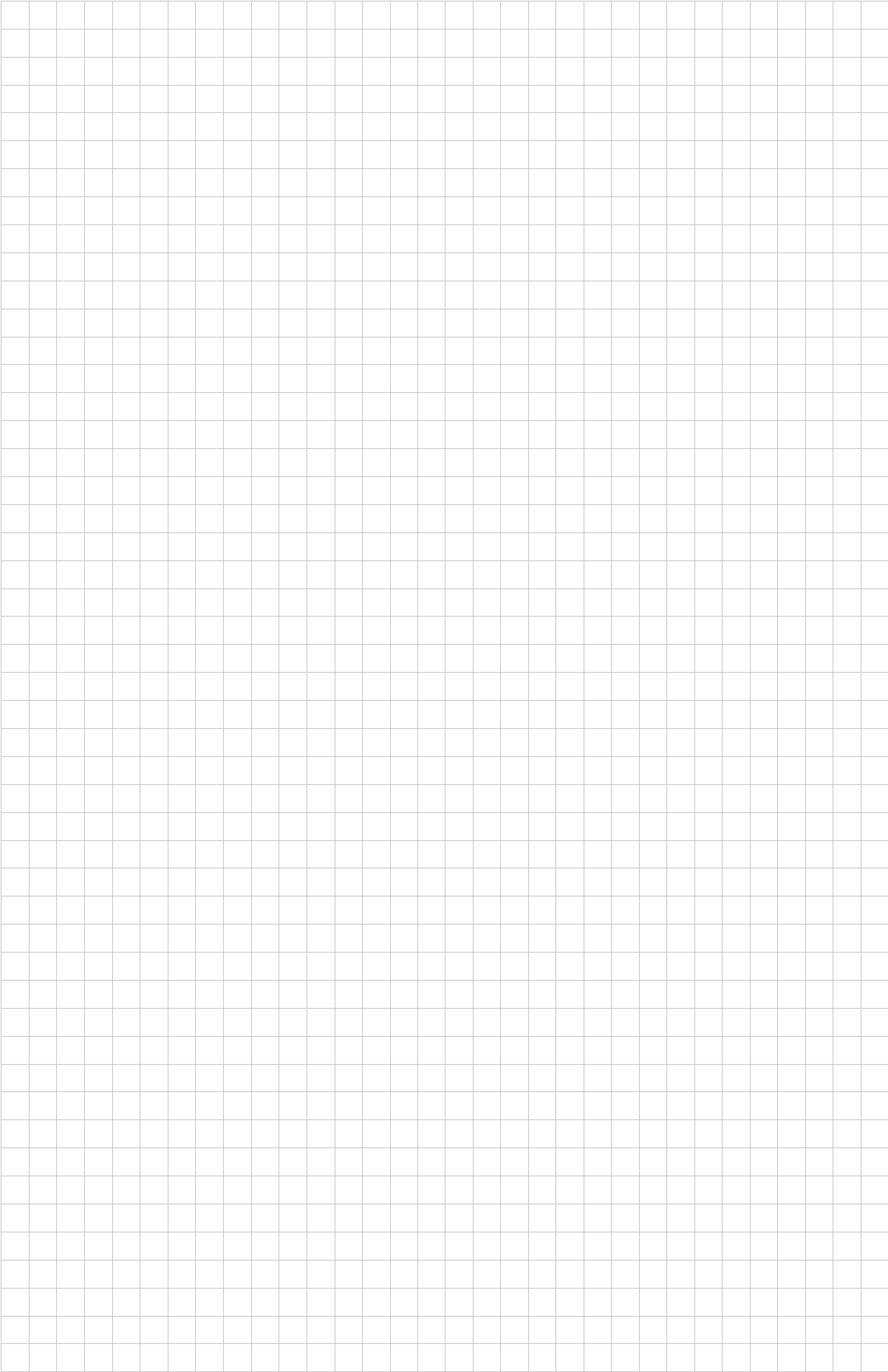
$$f_{pe} = \sqrt{\frac{n_{e0} q_e^2}{4 \pi^2 m_e \epsilon_0}} \quad (34.455)$$

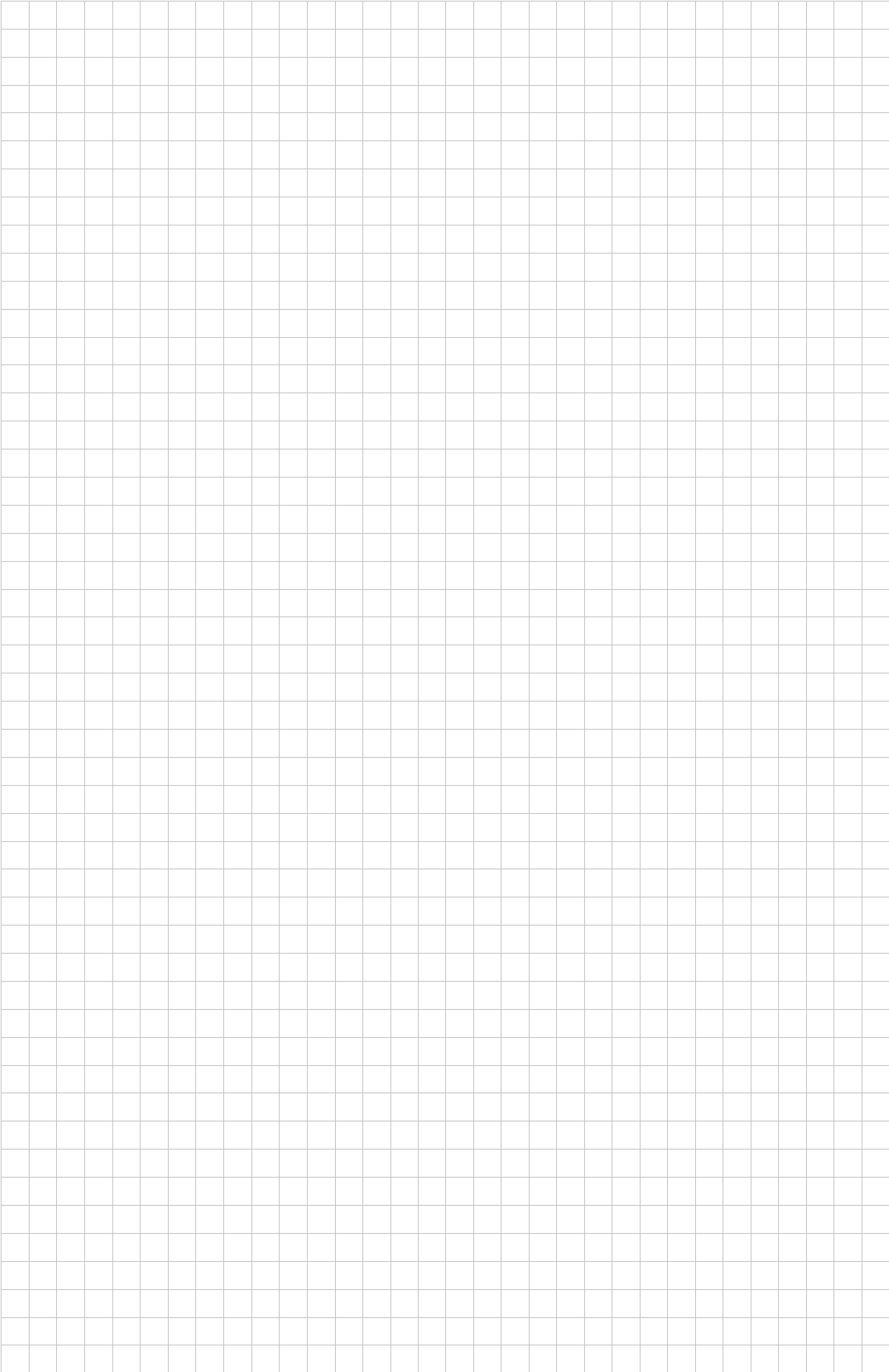
En physique, la fréquence plasma est ainsi la fréquence caractéristique des ondes de plasma, c'est-à-dire des oscillations des charges électriques présentes dans les milieux conducteurs, comme le métal ou les plasmas. À l'image de l'onde électromagnétique qui, quantifiée, est décrite par des photons, cette onde de plasma est quantifiée en "**plasmons**".

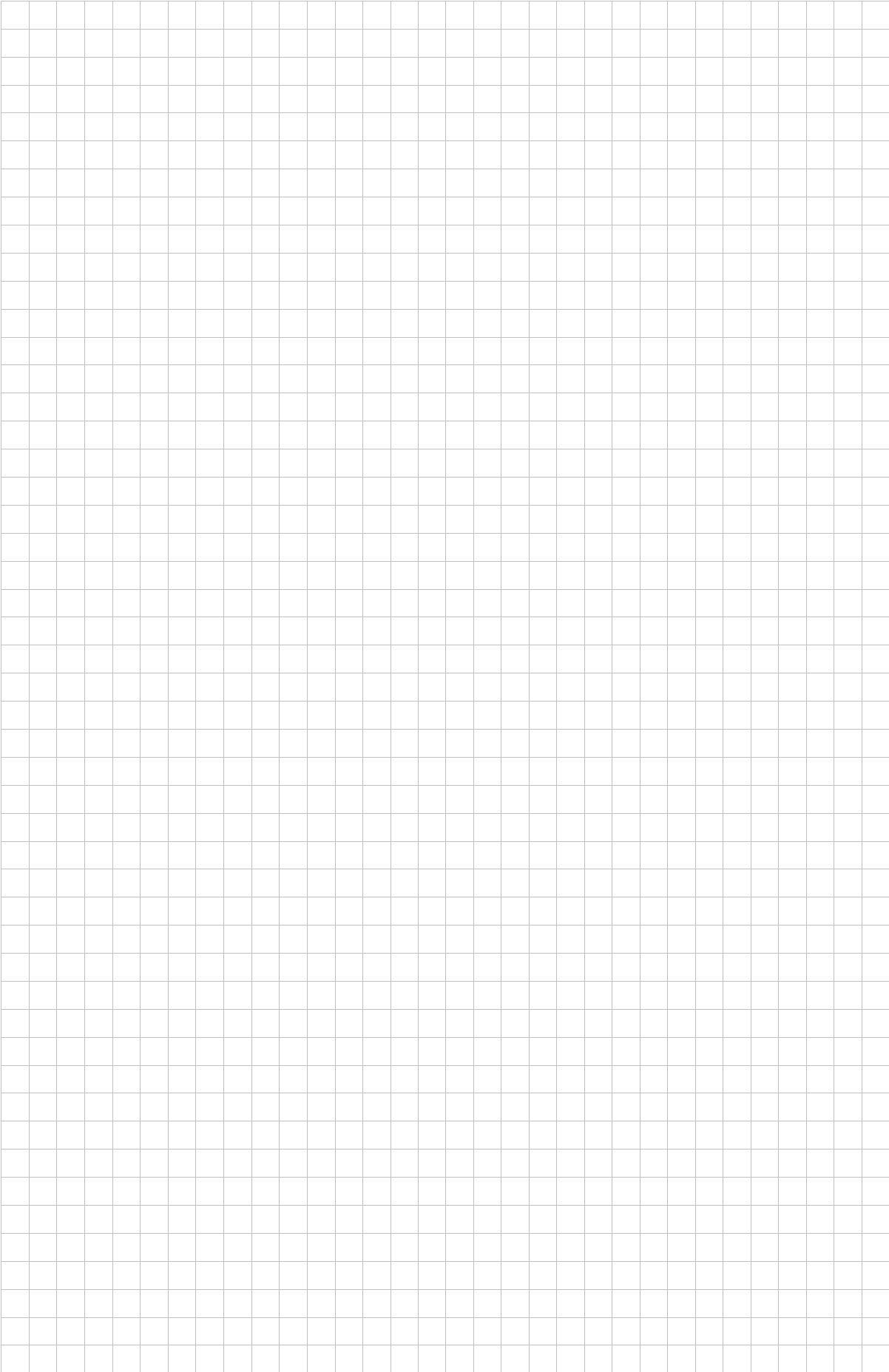
Les oscillations des charges électriques peuvent être comprises grâce au raisonnement suivant: si les électrons d'une zone du plasma sont déplacés, alors les ions de cette zone, n'ayant que peu bougé du fait de leur masse importante, vont exercer sur ces électrons une force de Coulomb attractive. Ceux-ci vont donc revenir vers leur position initiale, et ainsi de suite...

Notes personnelles:











Électrodynamique

35. ÉLECTROSTATIQUE

Jusqu'ici nous nous sommes concentrés sur l'interaction gravitationnelle et la grandeur caractéristique de la matière, appelée "masse", qui lui est associée. Nous avons évoqué l'interaction électromagnétique, en analysant des phénomènes macroscopiques, comme le frottement, la cohésion, l'élasticité, les forces de contact, etc. Maintenant nous nous penchons sur les forces électroniques et la caractéristique de la matière, appelée "charge", qui leur est associée. L'interaction électromagnétique lie la matière, sous toutes ses formes observables. C'est elle qui fait tenir les électrons au noyau dans l'atome, qui fait tenir ensemble les atomes dans les molécules, les molécules dans les objets et même votre nez à votre visage (eh oui... nous tenons pas à grand-chose.. lol).

La "charge" produit la "force électrique" ou "force de Coulomb" et nous commençons seulement à comprendre cette force. La charge est une notion fondamentale, qui ne peut pas être décrite en termes de concepts plus simples et plus fondamentaux. Nous la connaissons par ses effets et malheureusement pas par ce qu'elle est (c'est idem pour la masse rappelons-le aussi).

L'expérience a montré aussi que bien que la charge est comme la masse une propriété additive, elle comporte cependant aussi des valeurs négatives (et non exclusivement positive comme l'est a priori la masse). Ainsi, dans le langage actuel et comme l'expérience le confirme, deux charges identiques se repoussent et deux charges opposées s'attirent.

Voyons maintenant la force qui est associée à la charge:

1. FORCE ÉLECTRIQUE

Il a expérimentalement été établi par Coulomb qu'une particule témoin subit une force d'une intensité $\vec{F}$ proportionnelle à sa charge q , lorsqu'elle est placée au voisinage d'une ou plusieurs charges électriques Q_i , dans un milieu de permittivité ϵ (permittivité au champ électrique bien sûr...) donnée par (sous forme vectorielle et non relativiste):

$$\vec{F} = q \left(\frac{1}{4\pi\epsilon} \sum_i \frac{Q_i}{\|\vec{r} - \vec{r}_i\|^3} (\vec{r} - \vec{r}_i) \right) \quad (35.1)$$

où $\vec{r}_i$ est le vecteur position d'une charge témoin.

En d'autres termes, deux corps chargés ponctuels s'attirent ou se repoussent selon une force directement proportionnelle à leur charge et inversement proportionnelle au carré de la distance qui les sépare.

Dans le cas d'un système à deux particules séparées par une distance r , nous avons la même relation simplifiée et nous retrouvons la forme plus commune de la force électrique ou "**force de Coulomb**" telle qu'elle est donnée dans la plupart des ouvrages (sous forme scalaire et non relativiste):

$$F = \frac{1}{4\pi\epsilon} \frac{qQ}{r^2} \quad (35.2)$$

Remarques:

R1. Fréquemment, cette dernière relation est définie sous le nom de "**loi de Coulomb**" dans la plupart des écoles et admise comme non démontrable. Au fait, il n'en est rien ! Cette relation peut se démontrer comme nous le verrons lors de l'étude de la physique quantique des champs (cf. chapitre de Physique Quantique Des Champs) en utilisant l'équation de Klein-Gordon dans le contexte d'un champ de potentiel à symétrie sphérique (démonstration effectuée par Yukawa).

R2. Pour la forme relativiste de la loi de Coulomb, le lecteur se reportera au chapitre de Relativité Restreinte où il est démontré que (forme vectorielle):

$$\vec{F} = \frac{q \cdot (1 - v^2 / c^2)}{\left[1 - (v^2 / c^2) \sin^2 \theta\right]^{3/2}} \frac{Q\vec{r}}{4\pi\epsilon r^3} \quad (35.3)$$

La valeur de permittivité dans le vide est quant à elle donnée expérimentalement par:

$$\epsilon_0 = 8.8541878 \cdot 10^{-12} \left[C^2 N^{-1} m^{-2} \right] \quad (35.4)$$

et relativement au milieu considéré, nous définissons une permittivité relative ϵ_r qui permet plus facilement de déterminer les propriétés d'un matériau par rapport au champ électrique tel que:

$$\epsilon = \epsilon_0 \cdot \epsilon_r \quad (35.5)$$

Il convient d'indiquer que certains auteurs définissent la permittivité du vide à partir de la vitesse de la lumière et de la perméabilité magnétique du vide (cf. chapitre de Magnétostatique). Dès lors, la valeur de la permittivité du vide est bien évidemment exacte par définition. Mais cela n'a de sens qu'une fois la théorie de Maxwell connue et celle-ci sera présentée et démontrée que plus tard dans le chapitre d'Électrodynamique (nous suivons les démarches dans l'ordre historique des découvertes scientifiques).

Nous définissons également le rapport:

$$K_e = \epsilon / \epsilon_r \quad (35.6)$$

appelé "**constante diélectrique**".

Le facteur entre parenthèses dans:

$$\vec{F} = q \left(\frac{1}{4\pi\epsilon} \sum_i \frac{Q_i}{\|\vec{r} - \vec{r}_i\|^3} (\vec{r} - \vec{r}_i) \right) \quad (35.7)$$

ne dépend que de la distribution des charges Q_i dans l'espace et de la permittivité ϵ du milieu considéré. Puisque sa valeur varie d'un endroit à l'autre et dépend du vecteur position $\vec{r}$ de la charge témoin, il forme un ensemble de vecteurs, dont la propriété est celle d'une multitude de lignes de champs électriques d'où l'utilisation du terme "**champ électrique**".

L'ensemble de ces vecteurs porte donc le nom de "champ électrique" $\vec{E}$, au point $\vec{r}$, dans la distribution de charges Q_i :

$$\vec{E}(\vec{r}) = \frac{1}{4\pi\epsilon} \sum_i \frac{Q_i}{\|\vec{r} - \vec{r}_i\|^3} (\vec{r} - \vec{r}_i) \quad (35.8)$$

Les ingénieurs utilisent souvent une autre notation qui permet de caractériser uniquement la géométrie du champ et ce indépendamment du milieu et introduisent le concept de "champ de déplacement":

$$\vec{D} = \epsilon \vec{E}$$

nous retrouverons ce vecteur dans le chapitre d'Électrodynamique lors de notre synthèse des équations de Maxwell.

La force Coulombienne, agissant sur la charge témoin q , s'écrit alors de façon conventionnelle:

$$\vec{F} = q \cdot \vec{E}(\vec{r}) \quad (35.9)$$

2. POTENTIEL ÉLECTRIQUE

Soient deux points A et B dans une région de l'espace où il existe un champ électrique $\vec{E}(x,y,z)$ et soit un chemin Γ reliant ces deux points. Alors, dans le cas particulier où la source d'un champ $\vec{E}$ est une sphère ou un corps ponctuel et que nous posons une charge à son voisinage, nous avons pour le travail effectué par la force pour déplacer la charge du point A au point B :

$$\begin{aligned} dW_{A,B} &= \vec{F} \circ d\vec{r} \Rightarrow \int_A^B \vec{F} \circ d\vec{r} = q \int_A^B \vec{E} \circ d\vec{r} = q \int_A^B E \cos \alpha dr = q \int_A^B E dr = q \frac{Q}{4\pi\epsilon} \int_A^B \frac{1}{r^2} dr \\ &= q \frac{Q}{4\pi\epsilon} \left(-\frac{1}{r_B} + \frac{1}{r_A} \right) = q \frac{Q}{4\pi\epsilon} \left(\frac{1}{r_A} - \frac{1}{r_B} \right) \end{aligned} \quad (35.10)$$

Par ailleurs, ce travail est comme nous le verrons plus loin, assimilable à l'énergie potentielle. Nous définissons ainsi la "différence de potentiel" ou simplement le "potentiel" par la relation:

$$-\Delta U = \frac{Q}{4\pi\epsilon} \left(\frac{1}{r_B} - \frac{1}{r_A} \right) = U_B - U_A \quad (35.11)$$

et donc:

$$\Delta U = - \int \vec{E} \circ d\vec{r} = - \int E dr \quad (35.12)$$

Remarques:

R1. Le potentiel est souvent appelé "**tension**" par les électriciens, électrotechniciens ou autres ingénieurs. Parfois par abus de la langue anglophone le terme "**voltage**" est ensuite utilisé par référence à l'unité de mesure du potentiel qui est le "**Volt**" noté [V].

R2. La différence de potentiel peut aussi bien se faire entre deux bornes chargées de manières opposées (+,-) qu'entre deux bornes (+,neutre) ou encore (-,neutre). Ces deux derniers cas représentent typiquement la configuration utilisée par les trains, trams, l'orage et presque tous les appareils électroménagers.

Démontrons maintenant dans le cadre le plus général qui soit que le champ vectoriel stationnaire $\vec{E}$ dérive d'un champ de potentiel:

Soit une charge Q repérée par rapport à un référentiel par le vecteur $\vec{r}_Q$. Alors en chaque point de l'espace, il existe un champ $\vec{E}$ tel que:

$$\vec{E}(\vec{r}) = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{\|\vec{r} - \vec{r}_Q\|^3} \cdot (\vec{r} - \vec{r}_Q) \quad (35.13)$$

développons cette expression:

$$\vec{E}(x,y,z) = \frac{K \cdot Q}{[(x - r_{Qx})^2 + (y - r_{Qy})^2 + (z - r_{Qz})^2]^{\frac{3}{2}}} \cdot \begin{pmatrix} x - r_{Qx} \\ y - r_{Qy} \\ z - r_{Qz} \end{pmatrix} \quad (35.14)$$

Si $\vec{E}$ est un champ de potentiel stationnaire alors, il doit exister un potentiel $\Phi(x,y,z)$ de ce champ qui satisfasse:

$$\frac{\partial \Phi}{\partial x}(x,y,z) = E_x(x,y,z); \frac{\partial \Phi}{\partial y}(x,y,z) = E_y(x,y,z); \frac{\partial \Phi}{\partial z}(x,y,z) = E_z(x,y,z) \quad (35.15)$$

Regardons si le potentiel Φ existe pour un champ de Coulomb. Nous devons alors avoir pour le champ en x:

$$\frac{\partial \Phi}{\partial x} = E_x = \frac{K \cdot Q}{[(x - r_{Qx})^2 + (y - r_{Qy})^2 + (z - r_{Qz})^2]^{\frac{3}{2}}} \cdot (x - r_{Qx}) \quad (35.16)$$

d'où:

$$\Phi(x,y,z) = -K \cdot Q \cdot [(x - r_{Qx})^2 + (y - r_{Qy})^2 + (z - r_{Qz})^2]^{-\frac{1}{2}} + C \quad (35.17)$$

et si nous effectuons le même développement pour chaque composante, nous obtenons également le même résultat. Donc le potentiel électrique est un champ scalaire et non vectoriel (le champ électrique est donc lui un champ vectoriel)!

Le pontentiel $\Phi(x,y,z)$ est appelé dans le cas d'un champ de Coulomb "**potentiel coulombien**" et est par convention choisi tel que:

$$\Phi(x,y,z) = -U \quad (35.18)$$

Comme nous pouvons le constater par l'expression de $\Phi(x,y,z)$ antéprécédente, C est une constante arbitraire, qui impose dans le cas d'absence de charges que:

$$U = 0 \quad (35.19)$$

Ce qui nous donne finalement:

$$E_x = -\frac{\partial U}{\partial x}, E_y = -\frac{\partial U}{\partial y}, E_z = -\frac{\partial U}{\partial z} \quad (35.20)$$

Ce qui donne pour toutes les composantes:

$$\vec{E} = - \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \cdot U \quad (35.21)$$

que nous notons plus brièvement:

$$\vec{E} = -\vec{\nabla} \cdot U = -\overrightarrow{grad}(U) \quad (35.22)$$

Remarque: Les mêmes développements et résultats (et ceux qui vont suivre) sont applicables en ce qui concerne le champ de potentiel gravitationnel. Cependant, il est rare qu'ils soient effectués dans la littérature ou les écoles car l'être humain ne contrôle pas le champ gravitationnel avec une facilité et une intensité équivalente à celle du champ électrique...

2.1. INDÉPENDANCE DU CHEMIN

Démontrons maintenant que la différence de potentiel entre deux points A et B ne dépend pas du chemin Γ parcouru tel que nous l'avons fait pour le champ de potentiel gravitationnel dans le chapitre de Mécanique Classique.

Soit Γ un chemin reliant deux points A et B et un champ $\vec{E}$ et faisons en sorte d'exprimer le champ en x, y et z par rapport à une seule variable t (qui n'a rien avoir avec le temps...) qui rendrait compte de sa variation lors d'un déplacement quelconque entre ces deux points:

$$\vec{E} \circ d\vec{r} = -(\vec{\nabla} \cdot U(r(t))) \circ d\vec{r} \quad (35.23)$$

avec donc A qui correspondrait à la valeur t_1 du paramétrage et B à la valeur t_2 .

Or, nous savons que (cf. chapitre de Calcul Différentiel et Intégral et Calcul Vectoriel):

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \frac{\partial f}{\partial z} dz = \partial_x f dx + \partial_y f dy + \partial_z f dz = (\vec{\nabla} \cdot f) \circ d\vec{r} \quad (35.24)$$

Il vient alors:

$$\vec{E} \circ d\vec{r} = -dU(r(t)) \quad (35.25)$$

Dès lors:

$$\underbrace{-\int_{\Gamma} \vec{E} \circ d\vec{r}}_{=\Delta U} = \int_{t_1}^{t_2} dU(r(t)) = U(r(t)) \Big|_{t_1}^{t_2} = U(r(t_2)) - U(r(t_1)) \quad (35.26)$$

Cette dernière expression montre bien que U est indépendant du chemin Γ quelle que soit la manière dont nous paramétrons celui-ci.

Le champ de Coulomb est donc un "**champ conservatif**". En effet, si nous considérons un chemin fermé Γ et soient A et B deux points confondus du chemin alors la différence de potentiel sera nulle.

Signalons que parfois nous disons aussi que le champ gradient du potentiel est conservatif.

2.2. ÉQUIPOTENTIELLES ET LIGNES DE CHAMP

Nous pouvons maintenant à partir de ce que nous avons établi, définir les "équipotentielles" et les "lignes de champ".

Soit un champ de Coulomb défini par rapport à un référentiel. Alors à chaque point (x,y,z) de l'espace, nous pouvons associer un vecteur champ électrique $\vec{E}(x,y,z)$ ainsi qu'un potentiel électrique.

Définition: Nous définissons les "**lignes de champ**" comme étant une famille de courbes pour lesquelles le vecteur $\vec{E}(x,y,z)$ est tangent et constant en chaque point et les "**équipotentielles**" comme étant des lignes pour lesquelles le potentiel $U(x,y,z)$ est aussi constant.

Dans ce cas, et c'est ce que nous allons démontrer, toutes les lignes de champ sont perpendiculaires à toutes les équipotentielles.

Démonstration:

Utilisons la propriété suivante de conservation du champ de Coulomb pour la démonstration:

$$\Delta U = - \oint_{\vec{r}} \vec{E} \circ d\vec{r} = U(B) - U(A) = 0 \quad (35.27)$$

Comme nous sommes en présence d'un champ électrique, celui-ci dérive donc d'un potentiel comme nous le savons. Ceci implique que si le champ n'est pas nul le potentiel ne l'est également pas. Donc, dans l'intégrale curviligne:

$$\Delta U = - \oint_{\vec{r}} \vec{E} \circ d\vec{r} \quad (35.28)$$

un des termes est nul ! Ce n'est pas le champ électrique $\vec{E}$ puisqu'on est présence d'un, ce qui discrédite le potentiel U et comme la charge se déplace $d\vec{r}$ n'est pas nul non plus. Écrivons alors l'intégrale curviligne d'une autre manière:

$$\Delta U = - \oint_{\vec{r}} \vec{E} \circ d\vec{r} = - \oint_{\vec{r}} E \cdot dr \cdot \cos \alpha = 0 \Leftrightarrow \frac{dU}{dr} = -E \cdot \cos \alpha = 0 \Leftrightarrow - \frac{\left(\frac{dU}{dr} \right)}{E} = \cos \alpha = 0 \quad (35.29)$$

d'où:

$$\alpha = \frac{\pi}{2} \quad (35.30)$$

nous pouvons donc conclure que les équipotentiels sont bien perpendiculaires aux lignes de champ électrique et inversement. C'est ce qu'il fallait démontrer.

Voici des exemples de lignes de niveaux comprenant lignes de champ et lignes de potentiel obtenus à l'aide de Maple 4.00b (nous montrerons lors de notre étude des équations différentielles comment obtenir les fonctions mathématiques des lignes de champ):

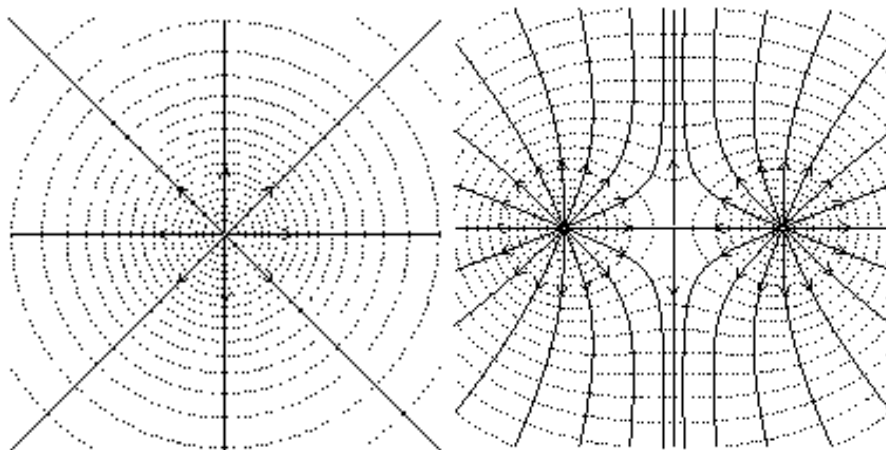


Figure: 35.1 - À gauche: une seule charge - À droite: deux charges de même signe

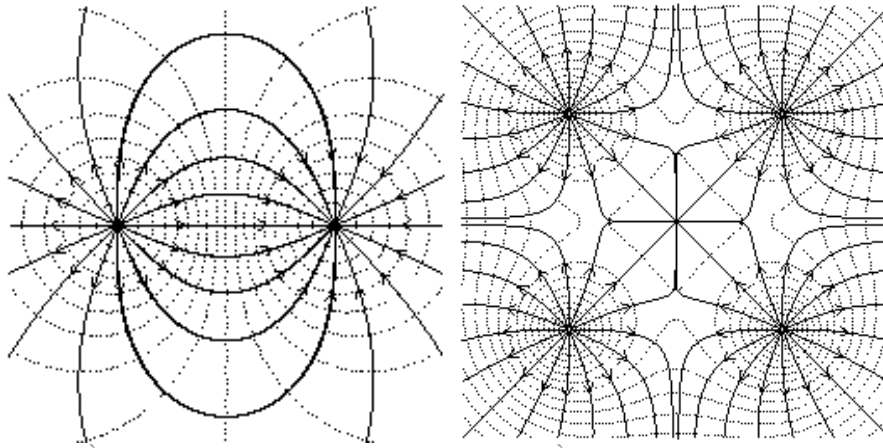


Figure: 35.2 - À gauche: deux charges de signes opposés - À droite: quatre charges de même signe

Remarque: Mises à part les charges opposées, nous rappelons que les mêmes résultats sont applicables pour les masses avec le champ gravitationnel.

Deux applications de ces résultats sont très importantes (pour lesquelles nous nous limiterons à l'étude des propriétés les plus importantes):

1. La détermination des lignes de champ et des lignes équipotentiels pour un fil rectiligne infini tel que nous pouvons en approximation en considérer dans les circuits électriques ou les lignes haute tension aériennes (afin de déterminer l'influence des champs des fils avec leur environnement - cette étude fait partie du domaine de l'électrodynamique de l'ingénieur que nous appelons la CEM pour "Compatibilité Électromagnétique"). Les résultats pourront aussi être utilisés pour déterminer la "**tension de pas**" pour certains systèmes rectilignes qui détermine pour une distance donnée, le potentiel par mètre pour lequel un mammifère peut être tué par électrochoc à proximité d'un tel fil. Une extension (sur laquelle je ne souhaite pas trop m'attarder bien que le sujet soit passionnant mais très chaud) est aussi l'influence d'un tel type de potentiel sur le fonctionnement du cerveau humain dans le cas de l'usage des téléphones portables (antennes émettrices d'un potentiel) ou d'habitations proches de lignes haute tension....

Remarque: Nous déterminerons dans le chapitre de Magnétostatique la loi de Biot et Savart qui donne le champ magnétique pour un tel fil parcouru par une intensité de courant donnée.

2. La détermination des lignes de champ et équipotentiels du dipôle électrique a une énorme importance en chimie. Nous verrons également quelle est la dynamique de celui-ci lorsqu'il est plongé dans un champ électrique uniforme et l'énergie d'interaction entre dipôles (comme c'est souvent le cas en chimie).

2.2.1. FIL RECTILIGNE INFINI

Soit:

$$\vec{u}_r = \vec{r} / \|\vec{r}\| \quad (35.31)$$

Nous avons:

$$d\vec{E} = \frac{1}{4\pi\epsilon} \frac{dq}{r^2} \vec{u}_r \quad (35.32)$$

en faisant usage du concept de densité linéique de charges telle que nous l'avons définie dans le chapitre Principes de la section de Mécanique, nous avons:

$$\vec{E} = \frac{1}{4\pi\epsilon} \int \frac{\gamma dl}{r^2} \vec{u}_r \quad (35.33)$$

Considérons une ligne infinie de section négligeable, et portant une charge linéique continue γ . Le but est donc de calculer le champ électrique et le potentiel en tout point M de l'espace extérieur à cette ligne afin de connaître les influences des charges de cette ligne sur son environnement en ne considérant que l'influence du champ électrique (si les charges étaient en mouvement il faudrait également prendre en compte l'influence du champ magnétique, ce que nous ferons dans le chapitre de Magnétostatique).

Pour cela, la méthode consiste à découper la ligne en de petits éléments de ligne dl , chacun de ces éléments portant une charge dq . Le champ créé par la charge en P au point M situé à distance x et de projection orthogonale H sur la ligne est:

$$d\vec{E}_P(M) = \frac{1}{4\pi\epsilon} \frac{\gamma dl}{x^2} \vec{u}_x \quad (35.34)$$

L'astuce consiste maintenant à prendre le symétrique P' de P par rapport à H (la projection orthogonale de M sur le fil):

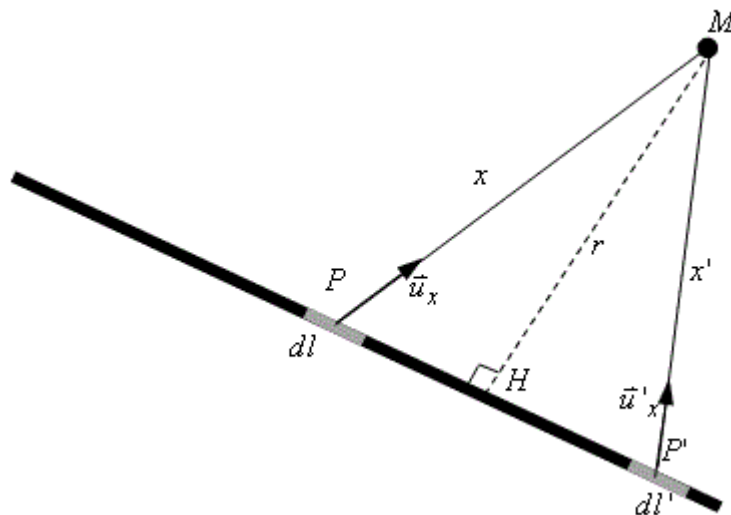


Figure: 35.3 - Configuration de l'analyse du fil rectiligne infini

pour lequel nous avons identiquement:

$$d\vec{E}_{P'}(M) = \frac{1}{4\pi\epsilon} \frac{\gamma dl'}{x'^2} \vec{u}_{x'} \quad (35.35)$$

Le champ total est donc:

$$d\vec{E}(M) = d\vec{E}_P(M) + d\vec{E}_{P'}(M) = \frac{1}{4\pi\epsilon} \frac{\gamma dl}{x^2} \vec{u}_x + \frac{1}{4\pi\epsilon} \frac{\gamma dl'}{x'^2} \vec{u}_{x'} \quad (35.36)$$

Or, nous avons:

$$x = x', dl = dl', \vec{u}_x = \frac{\overrightarrow{PM}}{PM}, \vec{u}_{x'} = \frac{\overrightarrow{P'M}}{P'M} \quad (35.37)$$

Donc:

$$d\vec{E}(M) = \frac{1}{4\pi\epsilon} \frac{\gamma dl}{x^2} \left(\frac{\overrightarrow{PM}}{PM} + \frac{\overrightarrow{P'M}}{P'M} \right) = \frac{1}{4\pi\epsilon} \frac{\gamma dl}{x^3} (\overrightarrow{PM} + \overrightarrow{P'M}) = \frac{2}{4\pi\epsilon} \frac{\gamma dl}{x^3} \overrightarrow{HM} \quad (35.38)$$

Comme nous pouvons nous en douter, cette dernière relation montre bien que le champ est orthogonal à la ligne (au fil...).

La norme de $d\vec{E}(M)$ est:

$$\|d\vec{E}(M)\| = \frac{2}{4\pi\epsilon} \frac{\gamma dl}{x^3} \|\overrightarrow{HM}\| = \frac{2}{4\pi\epsilon} \frac{\gamma dl}{x^3} r \quad (35.39)$$

Cette relation comporte 3 variables dépendantes r, dl, x . La norme du champ total en un point est donc la somme des normes sur l'ensemble de la longueur du fil puisque tous les vecteurs $d\vec{E}(M)$ ont même direction.

Pour effectuer ce calcul, nous allons effectuer un changement de variable, et mettre r, dl, x en fonction de l'angle α entre la ligne et le vecteur $\overrightarrow{PM}$. Dans le triangle rectangle HMP :

$$\cos \alpha = \frac{\overrightarrow{HP}}{\overrightarrow{PM}} = \frac{z}{x} \quad (35.40)$$

si nous prenons l'origine des z en H . Nous avons aussi:

$$\sin \alpha = \frac{\overrightarrow{HM}}{\overrightarrow{PM}} = \frac{r}{x} \quad \tan \alpha = \frac{\overrightarrow{HM}}{\overrightarrow{HP}} = \frac{r}{z} \quad (35.41)$$

et:

$$dl = dz = -r \frac{(1 + \tan^2 \alpha)}{\tan^2 \alpha} d\alpha = -r \frac{(1 + \frac{\sin^2 \alpha}{\cos^2 \alpha})}{\frac{\sin^2 \alpha}{\cos^2 \alpha}} d\alpha = -r \frac{(\sin^2 \alpha + \cos^2 \alpha)}{\sin^2 \alpha} d\alpha = \frac{-r}{\sin^2 \alpha} d\alpha \quad (35.42)$$

d'où:

$$dE = -\frac{\gamma}{2\pi\epsilon} \frac{r}{\sin^2 \alpha} d\alpha \frac{\sin^3 \alpha}{r^3} r = -\frac{\gamma}{2\pi\epsilon} \sin \alpha d\alpha \quad (35.43)$$

L'intégration est facile, mais il faut faire attention aux bornes. Nous devons intégrer sur une moitié de ligne, donc entre 0 et $\pi/2$:

$$E = \frac{\gamma}{2\pi\epsilon r} \cos \alpha \Big|_0^{\pi/2} \quad (35.44)$$

et alors:

$$\vec{E} = -\frac{\gamma}{2\pi\epsilon r} \vec{u}_r \quad (35.45)$$

Le potentiel se déduit aisément en prenant la primitive de E puisque:

$$\vec{E} = - \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \cdot U \quad (35.46)$$

Nous avons alors:

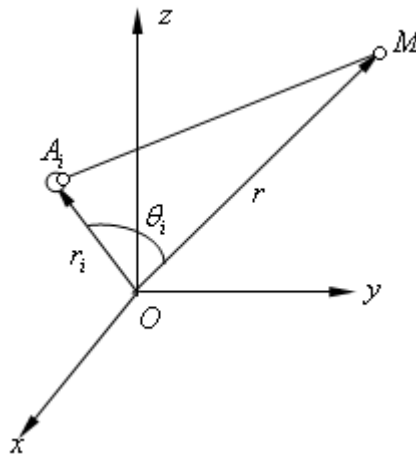
$$U = \frac{\gamma}{2\pi\epsilon} \ln(r) + c^{te} \quad (35.47)$$

La constante est indéterminée puisque lorsque r tend vers l'infini, U tendant vers zéro conduit à une constante infinie. Cette indétermination est due essentiellement à l'approximation de la ligne infinie.

2.3. DIPÔLE ÉLECTRIQUE RIGIDE

Une disposition des charges très intéressante est celle constituant un "**dipôle**" électrique appelé rigoureusement "**dipôle électrique rigide**" ou "**dipôle électrostatique**". Elle consiste en deux charges égales et opposées $+q, -q$ séparées par une très petite distance. Nous allons chercher à déterminer le potentiel et le champ électrique en un point M de l'environnement du dipôle.

Pour déterminer cela, considérons une charge q_i quelconque en un point A_i et un point M très éloigné de A_i . Prenons un repère quelconque centré en O :

Figure: 35.4 - Champ électrique en un point M très éloigné du dipôle

Le potentiel créé au point M par la charge q_i s'écrit:

$$U_i(M) = \frac{1}{4\pi\epsilon} \frac{q_i}{\overline{A_iM}} \quad (35.48)$$

Dans le triangle OA_iM , la distance $\overline{A_iM}$ peut être écrite selon le théorème du cosinus:

$$\overline{A_iM} = \sqrt{r_i^2 + r^2 - 2r_i r \cos \theta_i} \quad (35.49)$$

Le potentiel devient:

$$U_i(M) = \frac{1}{4\pi\epsilon} \frac{q_i}{\sqrt{r_i^2 + r^2 - 2r_i r \cos \theta_i}} \quad (35.50)$$

ou encore:

$$U_i(M) = \frac{1}{4\pi\epsilon} \frac{q_i}{r \sqrt{1 + \frac{r_i^2}{r^2} - \frac{2r_i \cos \theta_i}{r}}} \quad (35.51)$$

À très grande distance, r devient très supérieur à r_i , la quantité:

$$u = \frac{r_i^2}{r^2} - \frac{2r_i \cos \theta_i}{r} \quad (35.52)$$

tend alors vers zéro. Nous pouvons donc effectuer un développement de Maclaurin (cf. chapitre sur les Suites Et Séries) de $(1+u)^{-1/2}$ au voisinage de $u = 0$. Pour ne pas alourdir le calcul, nous nous limiterons à l'ordre deux en r :

$$(1+u)^{-1/2} = 1 - \frac{1}{2}u - \frac{1}{2} \left(-\frac{3}{2}u^2 \right) \frac{1}{2!} = 1 - \frac{1}{2}u + \frac{3}{8}u^2 \quad (35.53)$$

donc:

$$I = \left(1 + \frac{r_i^2}{r^2} - \frac{2r_i \cos \theta_i}{r} \right)^{-1/2} \cong 1 - \frac{1}{2} \left(\frac{r_i^2}{r^2} - \frac{2r_i \cos \theta_i}{r} \right) + \frac{3}{8} \left(\frac{r_i^2}{r^2} - \frac{2r_i \cos \theta_i}{r} \right)^2 \quad (35.54)$$

En ne gardant que les termes du second ordre en r :

$$I = 1 - \frac{1}{2} \frac{r_i^2}{r^2} + \frac{r_i \cos \theta_i}{r} + \frac{3}{2} \frac{r_i^2 \cos^2 \theta_i}{r^2} \quad (35.55)$$

Le potentiel devient:

$$\begin{aligned} U_i(M) &= \frac{1}{4\pi\epsilon} \frac{q_i}{r} \left(1 - \frac{1}{2} \frac{r_i^2}{r^2} + \frac{r_i \cos \theta_i}{r} + \frac{3}{2} \frac{r_i^2 \cos^2 \theta_i}{r^2} \right) \\ &= \frac{1}{4\pi\epsilon} \frac{q_i}{r} + \frac{1}{4\pi\epsilon} \frac{q_i r_i \cos \theta_i}{r^2} + \frac{1}{4\pi\epsilon} \frac{q_i r_i^2}{r^3} \left(-\frac{1}{2} + \frac{3}{2} \cos^2 \theta_i \right) = U_i^0 + U_i^1 + U_i^2 \end{aligned} \quad (35.56)$$

Nous avons gardé dans l'expression du potentiel trois termes. Le terme U_i^0 est le potentiel créé par une charge qui se trouverait en O . Autrement dit, à l'ordre zéro, le potentiel créé par une charge située en un point proche de O est identique au potentiel créé par une charge qui se trouverait en O . Les termes U_i^1, U_i^2 sont des termes correctifs, à l'ordre un et à l'ordre deux respectivement. Nous remarquons que ces deux termes varient en $1/r^2, 1/r^3$, donc décroissants plus vite que le premier. Ces deux termes sont donc plus efficaces à plus petite distance.

Nous voyons que les termes U_i^1, U_i^2 font intervenir la quantité $q_i r_i$. Cette quantité est ce que nous définissons comme étant le "**moment dipolaire**" du dipôle électrostatique:

$$\vec{p}_i = q_i \vec{r}_i \quad (35.57)$$

Remarque: Le moment dipolaire est exprimé en Coulomb par mètre, mais par mesure de commodité (...) il est exprimé en Debye [D] par certains ingénieurs.

Le potentiel créé à grande distance par une distribution discrète de charges s'obtient en sommant toutes les contributions individuelles:

$$U(M) = \sum_i U_i(M) = \frac{1}{4\pi\epsilon} \frac{1}{r} \sum_i q_i + \frac{1}{4\pi\epsilon} \frac{1}{r^2} \sum_i q_i r_i \cos \theta_i + \frac{1}{4\pi\epsilon} \sum_i \frac{q_i r_i^2}{r^3} \left(-\frac{1}{2} + \frac{3}{2} \cos^2 \theta_i \right) \quad (35.58)$$

Ce qui peut aussi s'écrire:

$$U(M) = U_0(M) + U_1(M) + U_2(M) \quad (35.59)$$

Par définition, $U_0(M)$ est le terme unipolaire ou monopolaire, $U_1(M)$ le terme dipolaire, $U_2(M)$ quadripolaire. Si la distribution de charge est au total nulle, comme c'est le cas d'un atome ou d'une molécule non ionisée, seuls subsistent les contributions multipolaires.

Revenons au cas particulier du dipôle:

Le terme monopolaire y est nul, puisque la somme des charges est nulle. Si nous négligeons les termes d'ordre supérieur au premier, il reste la contribution dipolaire.

Les angles ϑ_1 et ϑ_2 du dipôle sont complémentaires, donc $\cos \vartheta_1 = -\cos \vartheta_2$. Mais, comme $q_1 = -q_2 = q$, le produit $q_i \cos \vartheta_i$ est constant.

Les deux charges du dipôle sont à une distance constante l'une de l'autre et à équidistance de l'origine O. Nous poserons que $r_1 = r_2 = d$.

Le potentiel se réduit alors à:

$$U(M) = \frac{1}{4\pi\epsilon} \frac{2qd \cos \theta}{r^2} = \frac{1}{4\pi\epsilon} \frac{qa \cos \theta}{r^2} \quad (35.60)$$

où a est simplement la distance constante entre les deux charges.

Il est d'usage dans le cas de l'étude du dipôle électrique d'écrire la relation précédente sous la forme:

$$U(M) = \frac{1}{4\pi\epsilon} \frac{\vec{p} \cdot \vec{u}_r}{r^2} \quad (35.61)$$

où $\vec{p}$ est la définition de moment dipolaire et:

$$\vec{u}_r = \overrightarrow{OM} / \|\overrightarrow{OM}\| \quad (35.62)$$

Rappelons maintenant que nous avons démontré au début de ce chapitre que:

$$\vec{E} = -\vec{\nabla} \cdot U \quad (35.63)$$

et comme nous l'avons vu dans le chapitre de Calcul Vectoriel, le gradient en coordonnées sphériques nous amène à écrire:

$$\vec{E} = -\vec{\nabla} \cdot U = - \begin{pmatrix} \frac{\partial U}{\partial r} \\ \frac{1}{r} \frac{\partial U}{\partial \theta} \\ \frac{1}{r \sin \theta} \frac{\partial U}{\partial \phi} \end{pmatrix} \quad (35.64)$$

d'où:

$$E_r = \frac{p \cos \theta}{2\pi\epsilon r^3}, E_\theta = \frac{p \sin \theta}{4\pi\epsilon r^3}, E_\phi = 0 \quad (35.65)$$

Pour déterminer l'équation des équipotentiels, rappelons que ces lignes (ou "surfaces" dans l'espace) s'obtiennent par la contrainte:

$$U = c^{\text{te}} = U_0 \quad (35.66)$$

d'où:

$$U = U_0 \Leftrightarrow \frac{p \cos \theta}{4\pi\epsilon r^2} = U_0 \Leftrightarrow r = \sqrt{\frac{p \cos \theta}{4\pi\epsilon U_0}} = r_0 \sqrt{\cos \theta} \quad (35.67)$$

avec:

$$r_0 = \sqrt{p / 4\pi\epsilon U_0} \quad (35.68)$$

Le champ électrique doit être par définition tangent aux lignes de champ, donc parallèle au déplacement élémentaire.

$$\vec{E} \times d\vec{r} = 0 \Leftrightarrow \begin{pmatrix} E_r \\ E_\theta \\ E_\phi \end{pmatrix} \times \begin{pmatrix} dr \\ r d\theta \\ r \sin \theta d\phi \end{pmatrix} = \begin{pmatrix} E_\theta r \sin \theta d\phi - E_\phi r d\theta \\ E_\phi dr - E_r r \sin \theta d\phi \\ E_r r d\theta - E_\theta dr \end{pmatrix} = 0 \quad (35.69)$$

Puisque $E_\phi = 0$, nous avons:

$$\vec{E} \times d\vec{r} = 0 \Leftrightarrow \begin{pmatrix} E_r \\ E_\theta \\ E_\phi \end{pmatrix} \times \begin{pmatrix} dr \\ r d\theta \\ r \sin \theta d\phi \end{pmatrix} = \begin{pmatrix} E_\theta r \sin \theta d\phi \\ -E_r r \sin \theta d\phi \\ E_r r d\theta - E_\theta dr \end{pmatrix} = 0 \quad (35.70)$$

Donc finalement il ne reste plus que:

$$E_r r d\theta - E_\theta dr = 0 \Leftrightarrow E_r r d\theta = E_\theta dr \quad (35.71)$$

Qui est donc une équation différentielle qui s'intègre facilement:

$$\frac{dr}{r} = \frac{E_r}{E_\theta} d\theta = 2 \frac{\cos \theta d\theta}{\sin \theta} = 2 \frac{d \sin \theta}{\sin \theta} \Leftrightarrow \ln(r) = \ln(\sin^2 \theta) + c^{\#} \quad (35.72)$$

Ce qui équivaut à écrire:

$$r = c^{\#} \cdot \sin^2 \theta \quad (35.73)$$

Le tracé des lignes de champ et des équipotentiels donne alors en coordonnées sphériques (ne pas oublier que la composante verticale est nulle par symétrie):

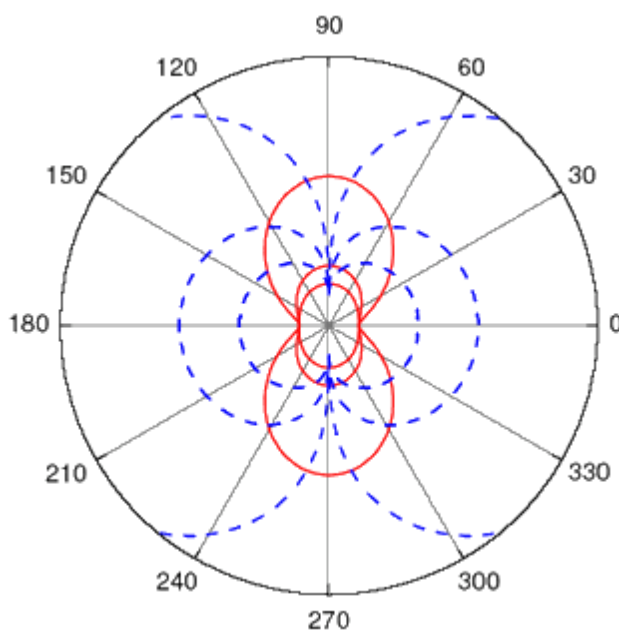


Figure: 35.5 - Tracé des lignes de champ d'un dipôle électrique

Bien que dans un dipôle électrique les deux charges soient égales et opposées, donnant une charge résultante nulle, le fait qu'elles soient légèrement déplacées est suffisant pour produire un champ électrique non identiquement nul. Dans les atomes, le centre de masse des électrons coïncide avec le noyau, et par conséquent le moment électrique dipolaire moyen de l'atome est nul. Mais si un champ extérieur est appliqué, le mouvement des électrons est distordu et le centre de masse des électrons est déplacé d'une distance x par rapport au noyau. L'atome est alors polarisé et devient un dipôle électrique de moment p . Ce moment étant proportionnel au champ extérieur $\vec{E}$.

Remarque: Les molécules par ailleurs peuvent avoir un moment électrique permanent. De telles molécules sont dites "**molécules polaires**". Par exemple, dans la molécule HCl l'électron de l'atome d'hydrogène passe plus de temps à se déplacer autour de l'atome de chlore qu'autour de l'atome d'hydrogène. Aussi, le centre des charges négatives ne coïncide-t-il pas avec le centre des charges positives et la molécule possède un moment dipolaire. Par contre, dans la molécule CO_2 , tous les atomes sont alignés, et le moment électrique dipolaire résultant est nul par raison de symétrie.

Quand un dipôle électrique est placé dans un champ électrique, une force s'exerce sur chacune des charges du dipôle. La force résultante est:

$$\vec{F} = q\vec{E}_+ - q\vec{E}_- = q(\vec{E}_+ - \vec{E}_-) \quad (35.74)$$

Considérons le cas particulier où le champ électrique est dirigé le long de l'axe des X et où le dipôle est orienté parallèlement à ce champ. Si nous considérons seulement les grandeurs:

$$E_+ - E_- = \left(\frac{dE}{dx} \right) a \quad (35.75)$$

avec a étant la distance entre les deux charges, et par conséquent:

$$F = q \frac{dE}{dx} a = p \frac{dE}{dx} \quad (35.76)$$

Ce résultat montre qu'un dipôle électrique orienté parallèlement au champ tend à se déplacer dans la direction dans laquelle le champ s'accroît (selon le gradient de celui-ci). Nous remarquons que si le champ électrique est uniforme, la force résultante sur le dipôle est nulle.

L'énergie potentielle du dipôle est:

$$E_p = qU_+ - qU_- = q(U_+ - U_-) = -qa \left(-\frac{U_+ - U_-}{a} \right) \quad (35.77)$$

Si nous utilisons la relation:

$$\vec{E} = -\vec{\nabla} \cdot U \quad (35.78)$$

pour décrire le champ électrique uniforme et si θ est l'angle entre le dipôle et le champ électrique, le dernier facteur $-(U_+ - U_-)/a$ est juste la composante $E_a = E \cos \theta$ du champ $\vec{E}$ parallèle à a . Donc:

$$E_p = -qaE_a \quad (35.79)$$

ou

$$E_p = -pE \cos \theta = -\vec{p} \cdot \vec{E} \quad (35.80)$$

L'énergie potentielle est minimale pour $\theta = 0$, ce qui montre que le dipôle est en équilibre quand il est orienté parallèlement au champ.

Ces configurations d'un dipôle placé dans un champ électrique ont des applications très importantes. Par exemple, le champ électrique d'un ion en solution polarise les molécules du solvant qui entoure les ions et elles s'orientent comme sur la figure ci-dessous:

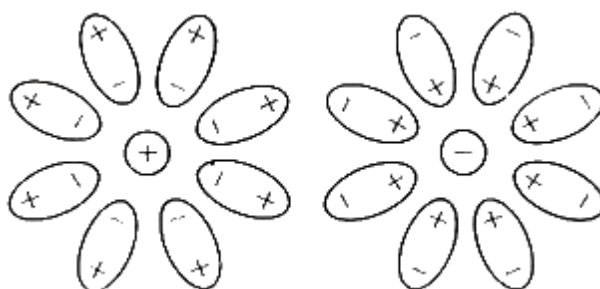


Figure: 35.6 - Exemple de ce qui se passe dans une solution avec un ion

Dans un solvant à molécules polaires tel que l'eau, les ions d'un électrolyte en solution s'entourent d'un certain nombre de ces molécules en raison de l'interaction charge-dipôle. Ce phénomène est appelé la "**solvation**" de l'ion, précisément "**hydratation**" si le solvant est de l'eau.

Ces molécules orientées deviennent plus ou moins solidaires de l'ion, augmentant sa masse effective et diminuant sa charge effective, qui est partiellement masquée par les molécules. L'effet net est que la mobilité de l'ion dans un champ extérieur est réduite. De même, lorsqu'un gaz ou un liquide, dont les molécules sont des dipôles permanents est placé dans un champ électrique, les molécules à la suite des couples dus au champ électrique, tendent à s'aligner avec leurs dipôles parallèles. Nous disons alors que la substance a été "**polarisée**".

Il peut donc être intéressant de déterminer le champ électrique vectoriel produit par un dipôle plutôt que le potentiel. Le champ électrostatique créé en un point M par le doublet s'obtient en effectuant la somme vectorielle des champs créés en ce point par des charges positive P et négative N , d'où:

$$\begin{aligned}\vec{E} &= \vec{E}_P + \vec{E}_N = \frac{q}{4\pi\epsilon} \left(\frac{\overrightarrow{PM}}{\|\overrightarrow{PM}\|^3} - \frac{\overrightarrow{NM}}{\|\overrightarrow{NM}\|^3} \right) = \frac{q}{4\pi\epsilon} \left(\frac{\vec{r} - a\vec{e}_z/2}{\|\vec{r} - a\vec{e}_z/2\|^3} - \frac{\vec{r} + a\vec{e}_z/2}{\|\vec{r} + a\vec{e}_z/2\|^3} \right) \\ &= \frac{q}{4\pi\epsilon} \left(\left(\vec{r} - \frac{a}{2}\vec{e}_z \right) \|\vec{r} - a\vec{e}_z/2\|^{-3} - \left(\vec{r} + \frac{a}{2}\vec{e}_z \right) \|\vec{r} + a\vec{e}_z/2\|^{-3} \right)\end{aligned}\quad (35.81)$$

La distribution des charges étant invariante par rotation autour de l'axe Oz du doublet, la topographie est indépendante de l'angle azimutal ϕ des coordonnées sphériques. Nous pouvons la représenter dans un plan méridien quelconque passant par l'axe NP . Le champ $\vec{E}$ est donc donné par:

$$\vec{E} = \frac{q}{4\pi\epsilon} \left(\left(\vec{r} - \frac{a}{2}\vec{e}_z \right) \|\vec{r} - a\vec{e}_z/2\|^{-3} - \left(\vec{r} + \frac{a}{2}\vec{e}_z \right) \|\vec{r} + a\vec{e}_z/2\|^{-3} \right) \quad (35.82)$$

Ayant:

$$\left\| \vec{r} + \frac{a}{2}\vec{e}_z \right\|^3 = \left(\left\| \vec{r} + \frac{a}{2}\vec{e}_z \right\|^2 \right)^{3/2} \quad (35.83)$$

vectoriellement, nous avons:

$$\vec{r} + \frac{a}{2}\vec{e}_z = \begin{pmatrix} r_x \\ r_y \\ r_z + \frac{a}{2} \end{pmatrix} \Rightarrow \left\| \vec{r} + \frac{a}{2}\vec{e}_z \right\|^2 = \underbrace{r_x^2 + r_y^2 + r_z^2}_{=r^2} + 2r_z \frac{a}{2} + \frac{a^2}{4} = r^2 + ar_z + \frac{a^2}{4} \quad (35.84)$$

Le produit scalaire étant la multiplication des composantes une à une, nous avons:

$$\vec{r} \circ \vec{e}_z = r_z \quad (35.85)$$

d'où:

$$\left\| \vec{r} + \frac{a}{2}\vec{e}_z \right\|^2 = r^2 + a\vec{r} \circ \vec{e}_z + \frac{a^2}{4} \quad (35.86)$$

Finalement:

$$\left(\left\| \vec{r} + \frac{a}{2}\vec{e}_z \right\|^2 \right)^{3/2} = \left(r^2 + a\vec{r} \circ \vec{e}_z + \frac{a^2}{4} \right)^{3/2} = r^3 \left(1 + a \frac{\vec{r} \circ \vec{e}_z}{r^2} + \frac{a^2}{4r^2} \right)^{3/2} \quad (35.87)$$

Donc par un développement limité en série de Maclaurin comme nous l'avons fait au début:

$$\begin{aligned}
\vec{E} &= \frac{q}{4\pi\epsilon r^3} \left(\left(\vec{r} - \frac{a}{2} \vec{e}_z \right) \left[1 - a \frac{\vec{r} \circ \vec{e}_z}{r^2} + \frac{a^2}{4r^2} \right]^{-3/2} - \left(\vec{r} + \frac{a}{2} \vec{e}_z \right) \left[1 + a \frac{\vec{r} \circ \vec{e}_z}{r^2} + \frac{a^2}{4r^2} \right]^{-3/2} \right) \\
&\cong \frac{q}{4\pi\epsilon r^3} \left(\left(\vec{r} - \frac{a}{2} \vec{e}_z \right) \left[1 + \frac{3}{2} a \frac{\vec{r} \circ \vec{e}_z}{r^2} + \dots \right] - \left(\vec{r} + \frac{a}{2} \vec{e}_z \right) \left[1 - \frac{3}{2} a \frac{\vec{r} \circ \vec{e}_z}{r^2} + \dots \right] \right) \\
&= -\frac{q}{4\pi\epsilon r^3} a \vec{e}_z + \frac{3q\vec{r}}{4\pi\epsilon r^5} a \vec{r} \circ \vec{e}_z + \dots \cong \frac{1}{4\pi\epsilon} \left[\frac{3(\vec{p} \circ \vec{r}) \cdot \vec{r}}{r^5} - \frac{\vec{p}}{r^3} \right]
\end{aligned} \quad (35.88)$$

soit en introduisant $\vec{u}_r = \vec{r} / \|\vec{r}\| = \vec{r} / r$:

$$\vec{E} \cong \frac{[3(\vec{p} \circ \vec{u}_r) \cdot \vec{u}_r - \vec{p}]}{4\pi\epsilon r^3} \quad (35.89)$$

Il peut être pertinent aussi de calculer l'énergie d'interaction entre deux dipôles électriques. Si nous appelons $\vec{p}_1$ le moment dipolaire, nous pouvons écrire:

$$\vec{E}_1 \cong \frac{[3(\vec{p}_1 \circ \vec{u}_r) \cdot \vec{u}_r - \vec{p}_1]}{4\pi\epsilon r^3} \quad (35.90)$$

Si nous désignons par $\vec{p}_2$ le moment du second dipôle et si nous utilisons la relation:

$$E_p = -\vec{p} \circ \vec{E} \quad (35.91)$$

nous trouvons que l'énergie d'interaction entre les deux dipôles est:

$$E_{p_1,2} = -\vec{p}_2 \circ \vec{E}_1 = -\frac{[3((\vec{p}_1 \circ \vec{u}_r) \cdot \vec{u}_r) \circ \vec{p}_2 - \vec{p}_1 \circ \vec{p}_2]}{4\pi\epsilon r^3} \quad (35.92)$$

Nous pouvons tirer plusieurs conclusions importantes de ce résultat. L'énergie d'interaction $E_{p_1,2}$ est symétrique par rapport aux deux dipôles, car la permutation de $\vec{p}_1$ et $\vec{p}_2$ la laisse inchangée. C'est un résultat prévu. L'interaction entre deux dipôles n'est pas centrale car elle dépend des angles que le vecteur de position ou le vecteur unitaire $\vec{u}_r$ fait avec $\vec{p}_1$ et $\vec{p}_2$.

Un atome, une molécule ou un ion, dont le moment dipolaire est nul à l'état fondamental, acquièrent un moment dipolaire sous l'action du champ électrique appliqué comme nous l'avons vu puisque les charges de signes opposées sont sollicitées dans des sens opposés. Les barycentres des charges positives et négatives ne coïncident plus, il apparaît un "**moment dipolaire induit**". Dans une approximation expérimentale linéaire valable pour des champs excitateurs faibles, ce moment dipolaire induit est proportionnel au champ appliqué $\vec{E}$, ce que nous traduisons par (il s'agit au fait d'une approximation de la relation de Langevin-Debye que nous démontrerons plus tard):

$$\vec{p} = \alpha \epsilon \vec{E} \quad (35.93)$$

La quantité α , dont la dimension physique est celle d'un volume, est la "**polarisabilité**" de l'édifice. L'interaction électrostatique dipôle-dipôle a été introduite par J.D. Van der Waals en 1873, dans le cas des molécules, afin d'interpréter les écarts réels par rapport aux gaz parfaits.

Les forces de Van der Waals sont répulsives lorsque la distance entre les molécules est très faible car elles s'opposent à l'interpénétration des nuages électroniques, ce que nous exprimons en introduisant leur volume (covolume).

En revanche, elles sont attractives lorsque cette distance est suffisante. Nous attribuons cette attraction à trois types d'interaction mettant en cause des dipôles rigides ou induits:

1. Les forces entre molécules polaires (dipôles rigides), dites de W. Keesom.
2. Les forces entre une molécule polaire (dipôle rigide), et une molécule polarisable (dipôle induit) dites de Debye.
3. Les forces moyennes entre les dipôles induits instantanés qui apparaissent même lorsque les molécules ne sont pas polaires, dites de F. London.

Dans ces trois cas, l'énergie électrostatique est négative (attraction) et varie comme r^{-6} . Pour le montrer, calculons l'énergie d'interaction entre deux dipôles rigides, de moments dipolaires $\vec{p}_1$ et $\vec{p}_2$:

$$E_{p,2,1} = -\vec{p}_1 \circ \vec{E}_2 \quad (35.94)$$

avec:

$$\vec{E}_2 \cong \frac{[3(\vec{p}_2 \circ \vec{u}_r) \cdot \vec{u}_r - \vec{p}_2]}{4\pi\epsilon r^3} \quad (35.95)$$

et:

$$\vec{p}_1 = \alpha\epsilon \vec{E}_2 \quad (35.96)$$

Par conséquent:

$$E_{p,2,1} = -\alpha\epsilon \vec{E}_2 \circ \vec{E}_2 = -\alpha\epsilon E_2^2 \cong -\alpha\epsilon \left(\frac{2p_2}{4\pi\epsilon r^3} \right)^2 = -\frac{c^2}{r^6} \quad (35.97)$$

d'où:

$$F = -\vec{\nabla} \cdot E_p = \frac{c^2}{r^7} \quad (35.98)$$

Ainsi, la dépendance radiale de la force est en r^{-7} . Cette décroissance très rapide de la force de Van der Waals avec la distance permet d'expliquer sa très courte portée et par conséquent son influence lorsque le milieu est suffisamment dense.

Remarque: L'interaction entre molécules polaires, de type Keesom, est rendue très importante par la présence de l'atome d'hydrogène, car ce dernier, en raison de sa petite taille, interagit aussi avec les atomes des autres molécules. C'est elle qui est à l'origine de la "**liaison hydrogène**".

3. FLUX DU CHAMP ÉLECTRIQUE

Soit $\vec{V}$ un champ vectoriel et S une surface appelée "**surface de Gauss**" dans l'espace. Si nous divisons cette surface en un nombre N de petites surfaces dS chacune traversée par un champ $\vec{V}_i$ et ayant un vecteur unitaire $\vec{n}_i$ perpendiculaire (cas particulier) à leur surface, nous pouvons alors former la somme:

$$\sum_{i=1}^N \vec{V}_i \circ \vec{n}_i dS \quad (35.99)$$

Lorsque N tend vers l'infini et tous les dS vers zéro, nous obtenons pour cette somme:

$$\Phi_V = \int_S \vec{V} \circ d\vec{S} \quad (35.100)$$

La valeur de cette intégrale donne donc le flux Φ_V du champ $\vec{V}$ à travers la surface S délimitée par un domaine Γ et où:

$$d\vec{S} = \vec{n} \cdot dS \quad (35.101)$$

Dans le cas du champ électrostatique, nous écrivons:

$$\Phi_E = \int_S \vec{E} \circ d\vec{S} \quad (35.102)$$

Cette expression définit le "**flux électrique**".

La question inévitable qui se pose alors est: quelle est sa signification physique ? Le flux d'un fluide est la quantité de fluide (notamment le volume) qui traverse une surface par seconde; il y alors écoulement de quelque chose. Quant au flux électrique, du point de vue classique, rien ne s'écoule, le champ électrique est déjà établi et il est statique, mais il traverse la surface. La valeur du champ électrique en tout point de l'espace est l'intensité du champ en ce point, tandis que le flux peut être considéré comme la quantité de champ qui traverse la surface S . Il y a une centaine d'années, les physiciens identifiaient le flux avec le nombre des lignes de champ traversant la surface. Mais le moins que nous puissions dire est que la vision simpliste que les lignes de champ ont une réalité distincte et que nous pouvons les compter est trompeuse. Nous verrons en mécanique quantique des champs que celle-ci soutient qu'un courant de photons virtuels est la nature même des interactions électromagnétiques. Malgré cela, les physiciens ne se sont pas pressés d'associer le flux des photons virtuels du 20ème siècle à l'image des lignes de champ continues du 19ème siècle. Quelle que soit sa nature, la notion de flux est puissante et de grande utilité pratique, aussi bien en électricité qu'en magnétisme.

Comme nous le démontrerons dans le cadre des équations de Maxwell (cf. chapitre d'Électrodynamique), la résolution de cette intégrale est (c'est la "loi de Gauss" ou également dit "théorème de Gauss"):

$$\Phi_E = \oint_S \vec{E} \circ d\vec{S} = \frac{Q}{\epsilon} = \frac{\sum_i q_i}{\epsilon} \quad (35.103)$$

3.1. CAPACITÉS

Comme application directe du théorème de Gauss, très utile en électronique et pour les ingénieurs, considérons une grande feuille mince et plane, portant une charge surfacique homogène σ et baignant dans un milieu de permittivité ε . Dans la région proche de son centre, le champ résultant de tous les champs des charges est normal, uniforme, constant et s'éloigne de la feuille. Considérons une surface de Gauss en forme d'un cylindre limité par les bases $S_1 = S_2 = S$ et sa surface tubulaire S_3 et symétrique par rapport à la feuille. Elle enferme donc une charge σS . Il en résulte que:

$$\oint \vec{E} \cdot d\vec{S} = E_1 \cdot S_1 + E_2 \cdot S_2 + E_3 \cdot S_3 = \frac{\sum_i q_i}{\varepsilon} \quad (35.104)$$

et comme $E = E_1 = E_2$ et $E_3 = 0$, nous trouvons:

$$E \cdot S + E \cdot S = 2E \cdot S = \frac{Q}{\varepsilon} \Rightarrow E = \frac{Q/S}{2\varepsilon} = \frac{Q/S}{2\varepsilon} = \frac{\sigma}{2\varepsilon} \quad (35.105)$$

Finalement, le champ électrique d'une grande feuille chargée plane et mince est:

$$E = \frac{\sigma}{2\varepsilon} \quad (35.106)$$

Si nous mettons face à face deux plaques identiques mais avec des charges opposées, la somme algébrique donnera bien évidemment:

$$E = \frac{\sigma}{\varepsilon} \quad (35.107)$$

À l'exception des extrémités, où l'effet de bord est important, le champ global est partout la somme vectorielle des champs uniformes produits par les deux couches minces opposées. Nous appelons un tel système un "**condensateur plan et parallèle**".

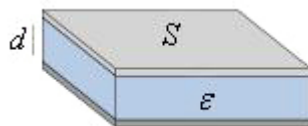


Figure: 35.7 - Exemple de condensateur plan et parallèle

Le résultat est aussi remarquable, car il est indépendant de la distance d entre les plans. Le calcul du potentiel électrique y est donc simplifié. Soit:

$$\Delta U = - \int_0^d E dr = - \frac{\sigma}{\varepsilon} \cdot d \Rightarrow U = |\Delta U| = \frac{\sigma \cdot d}{\varepsilon} = \frac{Q \cdot d}{\varepsilon \cdot S} \quad (35.108)$$

Ainsi, la capacité du condensateur plan et parallèle vaut donc:

$$C = \frac{Q}{U} = \frac{\varepsilon \cdot S}{d} \quad (35.109)$$

Voyons un deuxième exemple scolaire qu'est le "**condensateur cylindrique**":

Les armatures d'un condensateur cylindrique sont deux cylindres infinis (ou très grands relativement à leur diamètre) coaxiaux de rayon R_1 et R_2 . Il s'agit donc du cas très important du câble coaxial (dont le diélectrique est souvent du polyéthylène) que l'on retrouve dans de nombreux laboratoires:

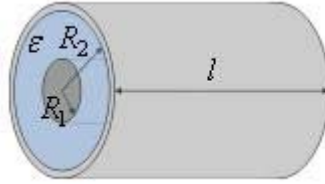


Figure: 35.8 - Exemple de condensateur cylindrique

Par le théorème de Gauss, nous savons que:

$$\vec{E} \circ \vec{S} = \frac{Q}{\epsilon} \quad (35.110)$$

Et puisque le champ est colinéaire en tout point à la surface, il vient immédiatement en connaissant l'expression de la surface du cylindre (cf. chapitre Formes Géométriques):

$$E = \frac{Q}{2\pi \cdot r \cdot l \cdot \epsilon} \quad (35.111)$$

Or:

$$\begin{aligned} \Delta U &= - \int_{R_1}^{R_2} E dr = - \int_{R_1}^{R_2} \frac{Q}{2\pi \cdot l \cdot \epsilon} \cdot \frac{1}{r} \cdot dr = - \frac{Q}{2\pi \cdot \epsilon \cdot l} \cdot [\ln(r)]_{R_1}^{R_2} \\ &= - \frac{Q}{2\pi \cdot \epsilon \cdot l} \cdot [\ln(R_2) - \ln(R_1)] = - \frac{Q}{2\pi \cdot \epsilon \cdot l} \cdot \ln\left(\frac{R_2}{R_1}\right) \end{aligned} \quad (35.112)$$

Et donc,

$$C = \frac{Q}{U} = \frac{Q}{|\Delta U|} = 2\pi \cdot \epsilon \cdot l \cdot \frac{1}{\ln\left(\frac{R_2}{R_1}\right)} \quad (35.113)$$

Calculons aussi la capacité d'un condensateur sphérique qui correspond en première approximation à certains générateurs de Van Der Graaf que nous avons dans les labos de quelques écoles, de musées ou même de centres de recherche:

Un "**condensateur sphérique**" est constitué de deux sphères concentriques de rayon R_1 et R_2 avec $R_1 < R_2$.

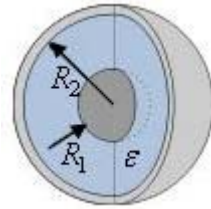


Figure: 35.9 - Exemple de condensateur sphérique

Nous avons maintenant immédiatement:

$$\vec{E} \circ \vec{S} = \frac{Q}{\epsilon} \quad (35.114)$$

Et donc puisque le champ est colinéaire en tout point à la surface il vient immédiatement en connaissant l'expression de la surface d'une sphère:

$$E = \frac{Q}{4\pi \cdot r^2 \cdot \epsilon} \quad (35.115)$$

Or:

$$\Delta U = - \int_{R_1}^{R_2} E dr = - \int_{R_1}^{R_2} \frac{Q}{4\pi \cdot \epsilon} \cdot \frac{1}{r^2} \cdot dr = - \frac{Q}{4\pi \cdot \epsilon} \left[\frac{1}{r} \right]_{R_1}^{R_2} = - \frac{Q}{4\pi \cdot \epsilon} \cdot \left(\frac{1}{R_2} - \frac{1}{R_1} \right) \quad (35.116)$$

Nous avons alors:

$$C = \frac{Q}{U} = \frac{Q}{|\Delta U|} = 4\pi \cdot \epsilon \cdot \left(\frac{1}{R_2} - \frac{1}{R_1} \right)^{-1} \quad (35.117)$$

Donc, nous avons alors:

$$C = 4\pi \cdot \epsilon \cdot \frac{R_1 \cdot R_2}{R_1 - R_2} \quad (35.118)$$

Voilà pour les exemples classiques....

Nous venons donc de voir que la capacité était définie par:

$$Q = C \cdot U \quad (35.119)$$

soit en régime non continu (cf. chapitre d'Électrocinétique):

$$q = C \cdot u \Rightarrow \frac{dq}{dt} = i = C \frac{du}{dt} \quad (35.120)$$

Nous avons alors pour la puissance instantanée (cf. chapitre d'Électrocinétique):

$$p = \Delta u \cdot i = u \cdot i = (r \cdot i - e) \cdot i = r \cdot i^2 - e \cdot i \quad (35.121)$$

En supposant un condensateur idéal (qui ne dissipe pas d'énergie par effet Joule) il vient:

$$p = u \cdot i = u \cdot C \frac{du}{dt} = Cu \frac{du}{dt} = \frac{d\left(\frac{1}{2}Cu^2\right)}{dt} \quad (35.122)$$

Donc par intégration dans un intervalle de temps donné de 0 à t nous avons:

$$E_C = \frac{1}{2} Cu^2 \quad (35.123)$$

Cette énergie est donc toujours positive et est stockée sous forme électrostatique dans le condensateur.

Remarque: Des expériences scientifiques nécessitant d'énormes énergies utilisent des milliers de condensateurs géants chargés sur le long terme pour accélérer des particules ou faire fonctionner des LASER mégajoules. Cependant on ne peut pas stocker le surplus d'énergie électrique des certaines centrales électriques, raison pour laquelle on utilise ce surplus pour remonter l'eau dans les barrages hydrauliques qui peuvent utiliser leur bassin comme réserve d'énergie potentielle pour produire un complément d'électricité au moment d'une pointe de consommation (transformation inverses).

Dans le cadre d'un régime sinusoïdal, la puissance moyenne sera nulle. Nous pouvons généraliser ceci en admettant qu'un condensateur parfait ne dissipe aucune puissance par effet Joule.

3.1.1. RIGIDITÉ DIÉLECTRIQUE

La "**rigidité diélectrique**" $\mathfrak{R}$ d'un milieu isolant représente la valeur maximum du champ en $[V \cdot m^{-1}]$ que le milieu peut supporter avant le déclenchement d'un arc électrique (donc d'un court-circuit). Pour un condensateur utilisé en électronique, si nous dépassons cette valeur, nous observons la destruction de l'élément. Cette valeur maximale de la tension appliquée aux bornes, est appelée "**tension de claquage**" U_c du condensateur. Nous pouvons définir la rigidité du milieu comme étant:

$$\mathfrak{R} = \frac{U_c}{d} \quad (35.124)$$

Exemple:

Pour l'air, on trouve dans les tables la valeur:

$$\mathfrak{R}_{\text{air}} = 30 \left[kV \cdot cm^{-1} \right] \quad (35.125)$$

Lorsque nous parlons de rigidité diélectrique, nous parlons aussi du diélectrique qui est un isolant ou une substance qui ne conduit pas l'électricité et qui est polarisable par un champ électrique. Dans la plupart des cas, les propriétés du diélectrique sont dues à la polarisation de la substance. Lorsque le diélectrique (dans notre cas, l'air est le diélectrique) est placé dans un champ électrique, les électrons et les protons de ses atomes se réorientent et, dans certains cas, à l'échelle moléculaire, une polarisation est induite (comme nous l'avons vu lors de notre étude des dipôles). Cette polarisation engendre une différence de potentiel, ou tension, entre les deux bornes du diélectrique; celui-ci emmagasine alors de

l'énergie qui devient disponible lorsque le champ électrique est supprimé. L'efficacité d'un diélectrique est sa capacité relative à emmagasiner de l'énergie comparée à celle du vide. Elle s'exprime par la permittivité relative, déterminée par rapport à celle du vide. La force diélectrique est la capacité d'un diélectrique à résister aux champs électriques sans perdre ses propriétés isolantes. Un diélectrique efficace libère une grande partie de l'énergie qu'il a emmagasinée lorsque le champ électrique est inversé.

4. ÉNERGIE POTENTIELLE ÉLECTROSTATIQUE

Considérons deux charges q_1, q_2 . La première est supposée au repos et fixe; la deuxième est amenée de l'infini à une distance a de q_1 (le même raisonnement a été appliqué pour le champ gravitationnel dans le chapitre de Mécanique Classique). Supposons que les deux charges soient de même signe. Comme q_1, q_2 ont tendance à se repousser mutuellement, il faut fournir une énergie potentielle E_p pour approcher q_2 (infiniment lentement) de q_1 . Le travail dW fourni par la force électrostatique en un point quelconque est par définition:

$$dW = Fdx \quad (35.126)$$

L'énergie potentielle du système est:

$$dE_p = -dW \quad (35.127)$$

car F est résistant (d'où l'origine du signe "-").

Donc:

$$dE_p = -dW = -Fdx = -\frac{q_1 q_2}{4\pi\epsilon} \frac{1}{x^2} dx \quad (35.128)$$

Nous obtenons alors simplement l'énergie potentielle en un point (donc le x au numérateur se simplifie avec un des x au dénominateur) au signe près:

$$E_p = \frac{q_1 q_2}{4\pi\epsilon} \frac{1}{x} \quad (35.129)$$

Cette énergie potentielle peut donc être négative ou positive.

Cela n'empêche pas que pour avoir la variation d'énergie potentielle, il faut intégrer la relation antérieure.

Ainsi que la relation:

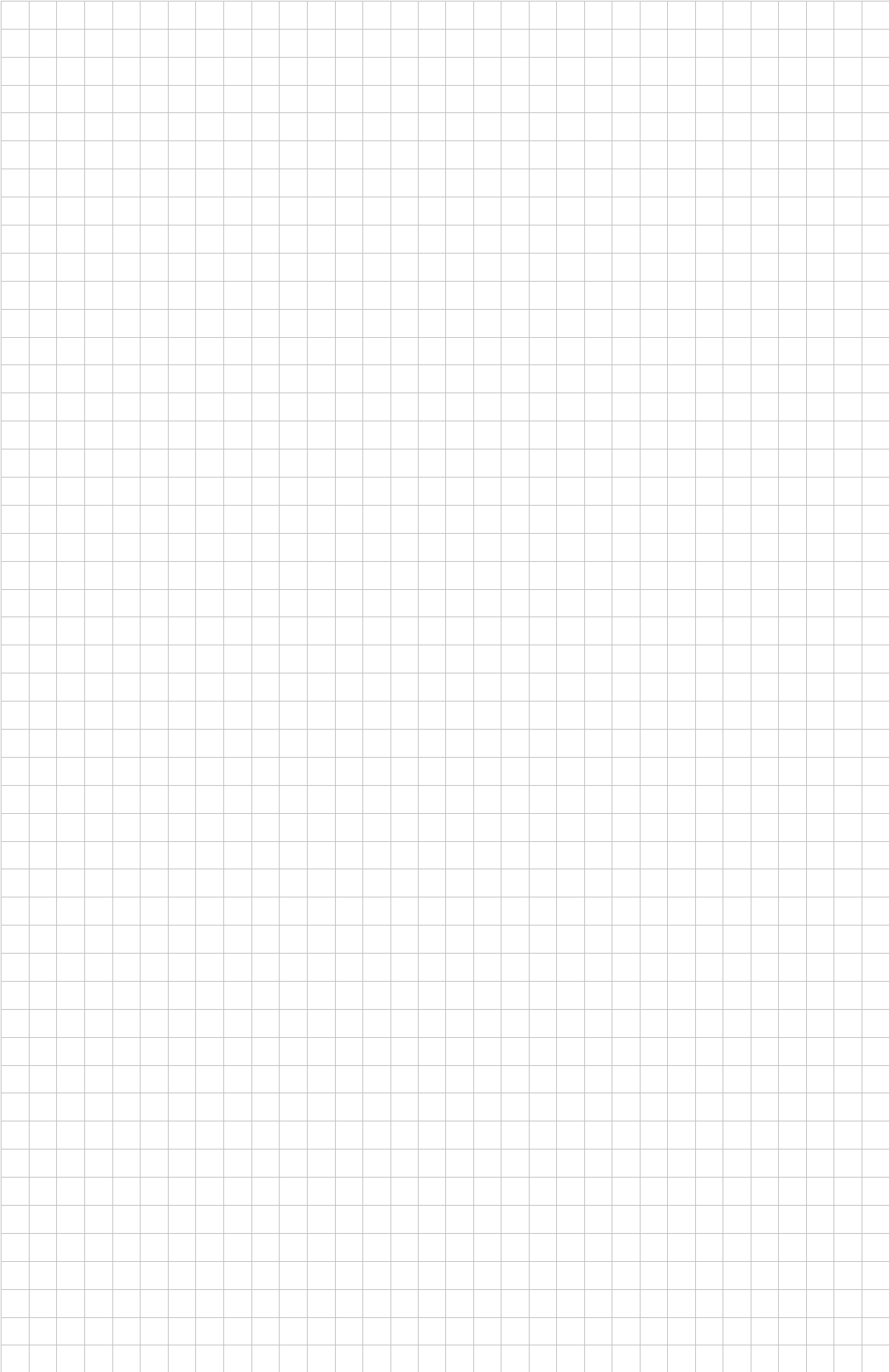
$$\vec{F} = -\vec{\nabla} \cdot E_p \quad (35.130)$$

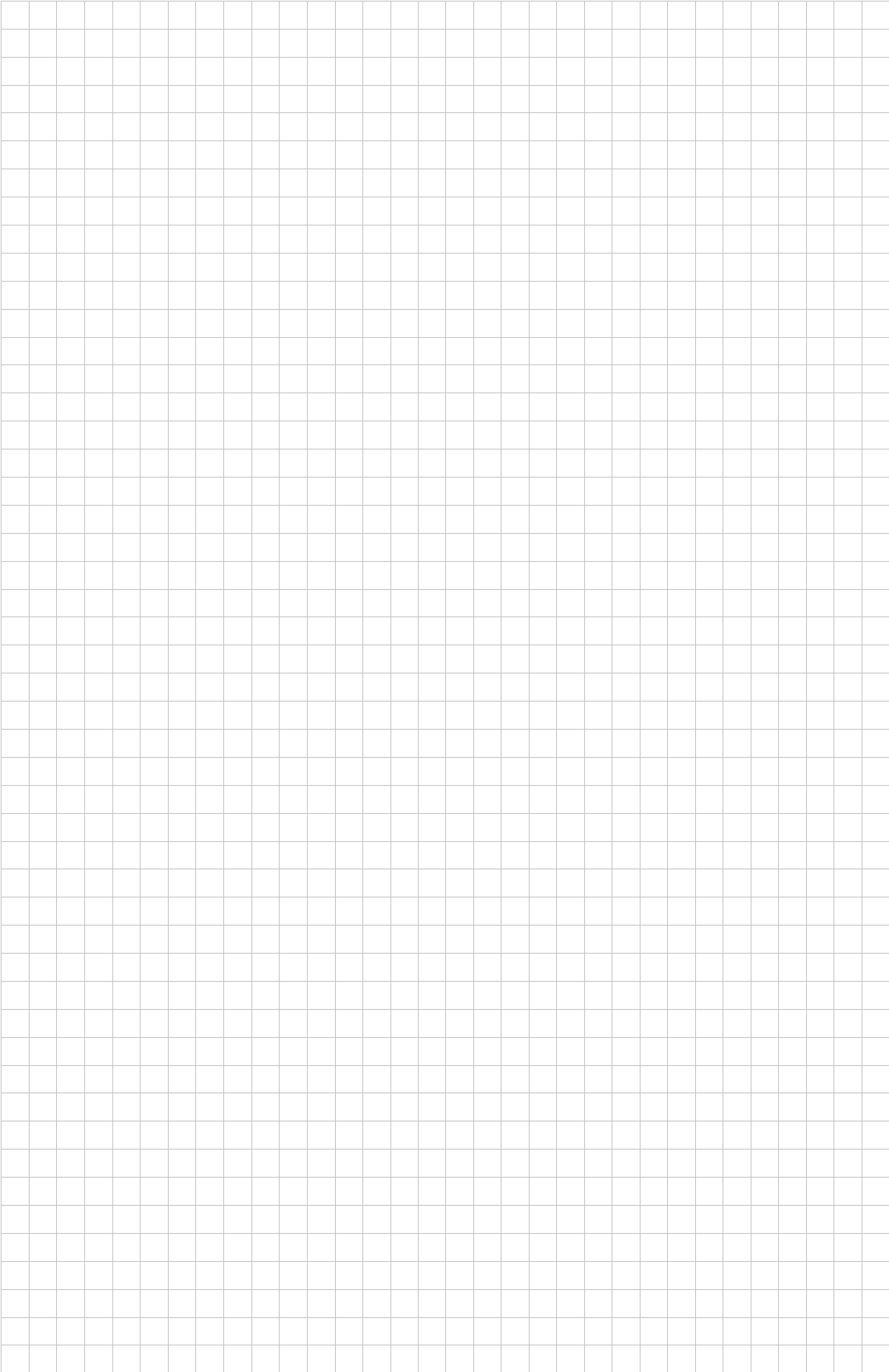
Remarquons que l'avant-dernière relation peut aussi se mettre sous la forme:

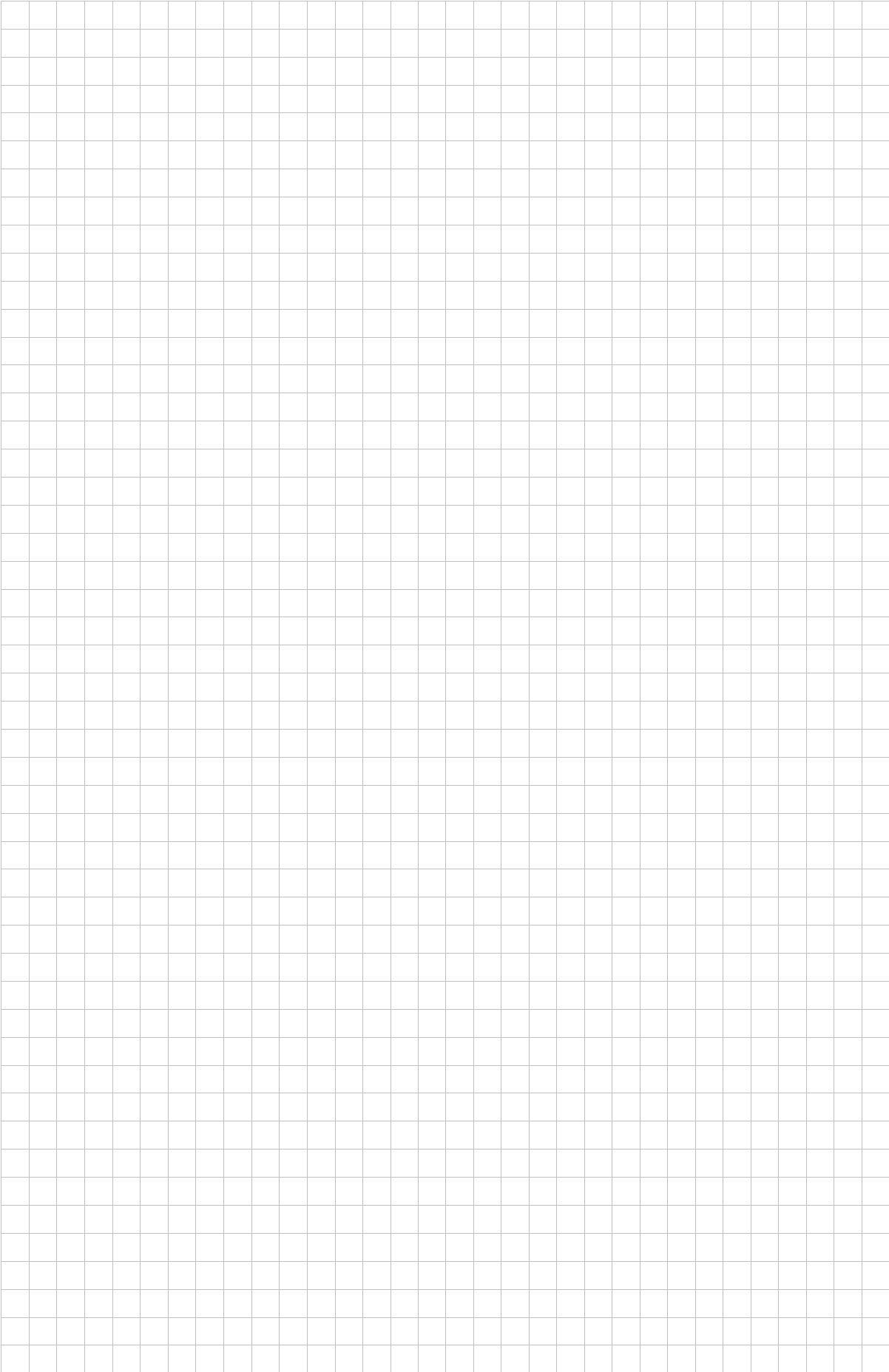
$$E_p = q_1 U_2 = q_2 U_1 = \frac{1}{2} (q_1 U_2 + q_2 U_1) \quad (35.131)$$

Remarque: Attention! Quand on fait de la physique, il faut voir de quelle énergie potentielle on parle. Là est tout le problème! Si vous prenez par exemple l'énergie potentielle due à la force de gravitation, elle peut prendre n'importe quelle valeur en fonction du point de référence. Si la référence est le niveau de la mer, un point situé sous le niveau de la mer aura une énergie potentielle négative, par contre si la référence est le centre de la Terre, il n'y aura que des énergies potentielles positives. C'est pour cela que nous écrivons plutôt l'énergie potentielle sous forme de différence de hauteur par rapport à une référence en mécanique. Pour l'énergie potentielle de l'électron, il faut savoir avec quoi il interagit. Si c'est avec une charge négative, le produit des charges est positif et donc l'énergie potentielle d'interaction sera positive, s'il interagit avec une charge positive, le produit des charges est négatif et l'énergie potentielle d'interaction électrostatique est négative. Bref, il faut bien savoir de quoi on parle. Les mots ont leur importance en physique aussi.

En général, si l'énergie potentielle diminue avec la distance, la force est répulsive, si elle augmente avec la distance, la force est attractive.







36. MAGNÉTOSTATIQUE

Les aimants sont connus depuis l'Antiquité (sans pour autant que l'on sache à l'époque qu'elle était l'origine de leurs propriétés) sous le nom de "**magnétites**", pierres noires trouvées à proximité de la ville de Magnesia (Turquie). C'est de cette pierre par ailleurs que provient le nom actuel de champ magnétique. Les Chinois furent les premiers à utiliser les propriétés des aimants différentes de celles des particules chargées, il y a plus de 1'000 ans, pour faire des boussoles. Elles étaient constituées d'une aiguille de magnétite posée sur de la paille flottant sur de l'eau contenue dans un récipient gradué.

Au même titre que le champ électrique, une bonne/meilleure compréhension de l'origine de ce champ ne peut se faire que par l'intermédiaire de théories modernes comme la physique quantique ondulatoire ou la physique quantique des champs. Le lecteur débutant devra donc prendre son mal en patience avant d'avoir les connaissances nécessaires pour étudier ces théories.

L'étude quantitative des interactions entre aimants et courants fut faite par les physiciens Biot et Savart à partir de 1820 seulement. Ils mesurèrent l'amplitude des oscillations d'une aiguille aimantée en fonction de sa distance à un courant rectiligne. Ils trouvèrent que la force agissant sur un pôle est dirigée perpendiculairement à la direction reliant ce pôle au conducteur et qu'elle varie en raison inverse de la distance. C'est le premier cas que nous allons étudier:

Soit un déplacement de charges électriques produisant dans l'espace un champ vectoriel dont les effets sont mesurables et dont les propriétés diffèrent de celles du champ électrostatique. Nous en déduisons l'existence d'un nouveau champ vectoriel que nous appelons (temporairement) "**champ magnétique**" et que nous noterons $\vec{B}$.

Les unités physiques du champ magnétique découleront naturellement du moment où nous arrivons à relier ce champ magnétique à quelque chose de connu comme une Force (ce que nous verrons plus loin). C'est ce que nous verrons lors de notre étude la "Force de Laplace".

Le cas d'étude le plus simple consistant en un fil rectiligne indéfini (exemple que nous pouvons aussi assimiler à un simple déplacement de charges sans nécessairement avoir un fil comme support) parcouru par un courant I (cf. chapitre d'Électrocinétique) montre que les lignes de champ magnétique sont des cercles ayant le fil pour axe.

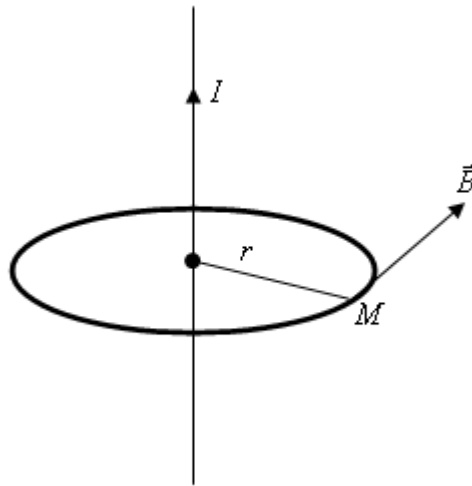


Figure: 36.1 - Champ magnétique autour d'un fil rectiligne infini

Remarque: Le sens de $\vec{B}$ se définit habituellement par l'intermédiaire de "l'observateur d'Ampère", c'est-à-dire un observateur qui serait placé le long du fil, de façon que le courant aille de ses pieds vers sa tête et qui regarderait le point M où nous évaluons le champ magnétique. $\vec{B}$ est dirigé de la droite vers la gauche de cet observateur.

Il aurait été expérimentalement établi par Biot et Savart en 1820 que la norme du champ magnétique B à la distance r du fil est proportionnelle au courant I qui le parcourt et inversement proportionnel à r :

$$B = k \cdot I \cdot r^{-1} \quad (36.1)$$

Cette relation constitue traditionnellement la base de l'étude théorique du champ magnétique.

Le coefficient de proportionnalité k dépend comme toujours des unités choisies. Pour l'ensemble de ces conséquences, il est avantageux d'écrire l'expression précédente sous une forme qui fasse apparaître la longueur du cercle de rayon r . Nous posons donc:

$$k = \frac{\mu_0}{2 \cdot \pi} \quad (36.2)$$

et obtenons ainsi la valeur du champ magnétique à une distance r d'un fil conducteur parcouru par un courant constant:

$$B = \frac{\mu_0 \cdot I}{2 \cdot \pi \cdot r} \quad (36.3)$$

où μ_0 est une nouvelle constante que nous appelons "**perméabilité magnétique du vide**" (à nouveau au même titre que pour la permittivité électrique, il existe une "**perméabilité magnétique relative**") et dont la valeur sont données comme à l'habitude sur ce site avec les autres constantes dans le chapitre Principes de la section Mécanique.

Les unités de cette constante, bien que données dans le chapitre Principes de la section Mécanique, se déduiront aussi automatiquement dès le moment où nous aurons réussi à relier le champ magnétique avec un concept connu comme la Force (voir plus loin). C'est ce que nous verrons aussi lors de notre étude la "Force de Laplace".

1. THÉORÈME D'AMPÈRE

Il est intéressant de calculer la "**circulation du champ magnétique**" $\vec{B}$ dans le vide le long d'un contour Γ qui tourne une fois dans le sens positif autour du fil orienté dans le sens du courant (observateur d'Ampère):

$$C_B = \oint_{\Gamma} \vec{B} \circ d\vec{l} = \oint_{\Gamma} B \cdot dl = \int_0^{2\pi} B \cdot r \cdot d\theta = \frac{\mu_0 \cdot I}{2 \cdot \pi} \int_0^{2\pi} d\theta = \mu_0 \cdot I \quad (36.4)$$

Remarque: Le champ est colinéaire le long du contour comme nous l'avons vu précédemment d'où le fait que le produit scalaire puisse s'écrire comme simple produit de normes.

Nous obtenons ainsi par définition la "**loi d'Ampère**" (ou appelée à tort "**théorème d'Ampère**" car ce résultat n'est pas démontrable... du moins à ma connaissance):

$$C_B = \oint_{\Gamma} \vec{B} \circ d\vec{l} = \mu_0 \cdot I \quad (36.5)$$

où le courant I dans un système à forte symétrie peut être assimilé à une simple somme algébrique des courants enlacés par le chemin tel que:

$$C_B = \oint_{\Gamma} \vec{B} \circ d\vec{l} = \mu_0 \sum_k I_k \quad (36.6)$$

Attention!!! Ce n'est pas parce que la circulation du champ magnétique est nulle dans une région de l'espace que le champ magnétique y est nul en tout point!

Remarques:

R1. La loi d'Ampère permet de déterminer la quatrième équation de Maxwell que nous démontrerons dans le chapitre d'Électrodynamique.

R2. La relation antérieure est parfois appelée à tort "**théorème d'Ampère**" alors qu'en réalité ce résultat n'est pas démontrable. Certains physiciens utilisent cependant la quatrième équation de Maxwell pour démontrer la relation antérieure mais alors c'est le serpent qui se mord la queue...

L'expression que nous avons obtenue peut encore être simplifiée si nous introduisons un nouvel être physique appelé "**intensité du champ magnétique**" ou encore plus couramment "**excitation magnétique**" et qui est notée par la lettre $\vec{H}$ (qui est intrinséquement indépendant du milieu de propagation).

Si nous considérons que nous sommes toujours dans le vide où il n'y a aucun dipôle magnétique alors nous le définissons dans le vide par:

$$\vec{B} = \mu_0 \cdot \vec{H} \quad (36.7)$$

Dès lors, nous sommes souvent amenés à parler de "**induction magnétique**" pour $\vec{B}$ et de "**champ magnétique**" pour $\vec{H}$. Mais les deux sont allègrement confondus suivant les auteurs et surtout les contextes (de même que ce sera le cas dans ce site Internet). Lorsque nous avons affaire à des aimants qui ont une magnétisation intrinsèque de par les propriétés du matériau qui les compose, nous notons de manière distincte le champ magnétique extérieur par:

$$\vec{B} = \mu_0 (\vec{H} + \vec{M}) \quad (36.8)$$

qui est donc une forme plus générale de la relation précédente. Il est alors d'usage de définir la "**susceptibilité magnétique**" comme étant le rapport sans dimensions:

$$\chi_m = \frac{M}{H} \quad (36.9)$$

Ainsi, la susceptibilité magnétique indique l'amplitude avec laquelle un matériau répond magnétiquement à la présence d'une excitation magnétique. Nous avons alors de par cette définition la relation entre perméabilité magnétique relative et susceptibilité magnétique:

$$\vec{B} = \mu_0 (\vec{H} + \vec{M}) = \mu_0 (\vec{H} + \chi_m \vec{H}) = \mu_0 (1 + \chi_m) \vec{H} = \mu_0 \mu_r \vec{H} = \mu \vec{H} \quad (36.10)$$

Soit:

$$\mu_r = \frac{\mu}{\mu_0} = 1 + \chi_m \quad (36.11)$$

où μ est appelé la "**perméabilité magnétique absolue**".

Il est d'usage d'appeler les matériaux qui ont une susceptibilité magnétique positive de "**matériaux paramagnétiques**" (contribuant au augmente le champ magnétique) et ceux qui ont une susceptibilité magnétique négative de "**matériaux diamagnétiques**" (de s'opposer au champ magnétique). Nous verrons plus loin les modèles théorique de Langevin permettant d'expliquer quantitativement avec une relativement bonne approximation des deux phénomènes (dans les deux cas la susceptibilité magnétique a une valeur qui est très faible).

Alors, finalement nous pouvons écrire la loi d'Ampère sous la forme:

$$\oint_{\Gamma} \vec{H} \circ d\vec{l} = I \quad (36.12)$$

L'intérêt de la loi d'Ampère ainsi que du concept de circulation du champ magnétique paraît (peut paraître) ainsi plus évident.

Cette dernière relation a bien évidemment une grande utilité en physique théorique car elle nous permettra de déterminer d'autres résultats forts importants. Sinon, au niveau de la pratique, le physicien de laboratoire ou l'électricien/électrotechnicien sera souvent confronté à devoir utiliser pour de petites et moyennes expériences des électro-aimants, dont il pourrait souhaiter recalibrer les valeurs nominales, ou encore des solénoïdes.

1.1. BOBINE SOLÉNOÏDALE INFINIE

Une application aussi particulièrement importante en électronique et électrotechnique est celle du calcul du champ d'induction dans une bobine de fil parcourue par un courant que nous considérerons comme constant dans un premier temps. Il s'agit ni plus ni moins d'une bobine d'induction plus techniquement appelée une "**inductance**". Voyons de quoi il s'agit:

Un solénoïde est une bobine formée par un fil conducteur enroulé en hélice et parcouru par un courant d'intensité I . Dans ce qui suit, nous supposons que le champ d'induction $\vec{B}$ d'un solénoïde est nul entre les spires et parallèle à l'axe du solénoïde.

Considérons la figure suivante et intéressons-nous en approximation qu'à la partie interne du solénoïde en admettant que le champ extérieur est nul par la longueur infinie de celui-ci et la parfaite jointure des bobines...:

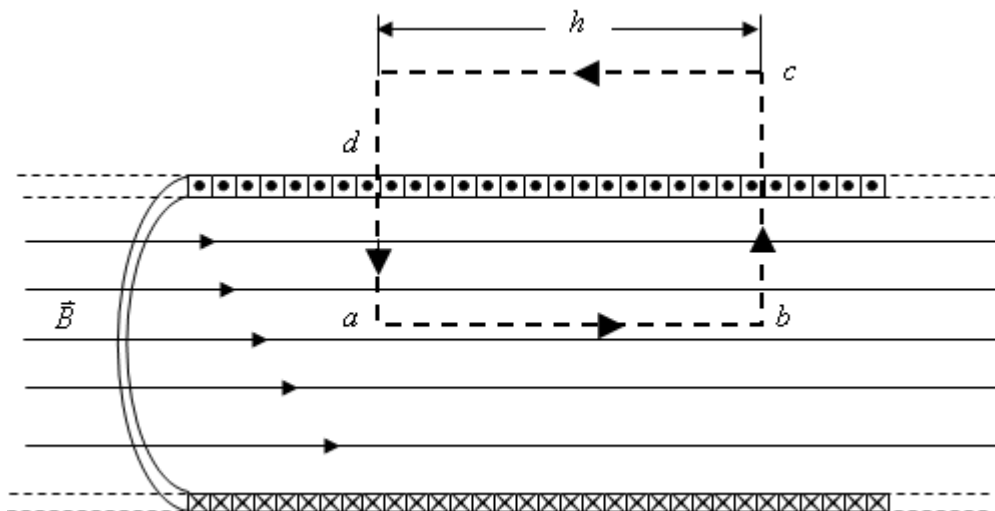


Figure: 36.2 - Solénoïde infini

Appliquons la loi d'ampère au trajet rectangulaire $abcd$. Ainsi:

$$\oint \vec{B} \circ d\vec{l} = \int_a^b \vec{B} \circ d\vec{l} + \int_b^c \vec{B} \circ d\vec{l} + \int_c^d \vec{B} \circ d\vec{l} + \int_d^a \vec{B} \circ d\vec{l} \quad (36.13)$$

La première intégrale du membre de droite donne $B \cdot h$ où B est la grandeur de $\vec{B}$ à l'intérieur du solénoïde et h , la longueur du segment ab . Nous pouvons remarquer que le segment ab , même s'il est parallèle à l'axe du solénoïde, ne doit pas nécessairement coïncider avec lui.

La deuxième et la quatrième intégrale sont nulles car, pour ces deux segments $\vec{B}$ et $d\vec{l}$ sont partout perpendiculaires: étant donné que $\vec{B} \circ d\vec{l}$ est nul partout, les deux intégrales sont nulles. La troisième intégrale est également nulle puisque le segment calculé se trouve à l'extérieur du solénoïde où nous avons supposé que le champ magnétique de la bobine était idéal.

Ainsi, l'intégrale $\oint \vec{B} \cdot d\vec{l}$ pour tout le trajet rectangulaire est $B \cdot h$ tel que:

$$\oint \vec{B} \cdot d\vec{l} = B \cdot h = \mu_0 \cdot I \quad (36.14)$$

mais le courant I est la somme des courants I_0 passant dans chacune des N spires contenues dans le chemin d'intégration. Mais en électronique, nous avons l'habitude de travailler avec la valeur n (nous choisissons la lettre minuscule par analogie avec la thermodynamique où les minuscules représentent des densités) qui est le nombre de spires par unité de longueur:

$$n = N / h \quad (36.15)$$

Ainsi, nous avons:

$$B \cdot h = \mu_0 \cdot I = \mu_0 \cdot I_0 \cdot N \Rightarrow B = \mu_0 I_0 n \quad (36.16)$$

Bien que cette relation ait été établie pour un solénoïde idéal infini, elle donne une grandeur assez précise (sans être exacte!) du champ d'induction magnétique pour les points d'intérieur situés près du centre d'un solénoïde réel. Cette relation révèle par ailleurs que le champ magnétique est en approximation indépendant du diamètre du solénoïde et qu'il est uniforme à travers la section de celui-ci. En laboratoire, un solénoïde est un dispositif pratique pour produire un champ d'induction uniforme de la même façon que le condensateur plan est utilisé pour produire un champ électrique uniforme.

1.2. BOBINE TOROÏDALE

La bobine toroïdale est un autre exemple important de l'application de la loi d'Ampère. Effectivement, nous retrouvons particulièrement ce type de configuration dans l'électronique de petite puissance (ordinateurs par exemple) où les inductances sont pour la plupart toroïdales ou dans la production d'énergie avec les fameux Tokomak qui de façon schématisée (très...) se réduisent à des bobines toroïdales.

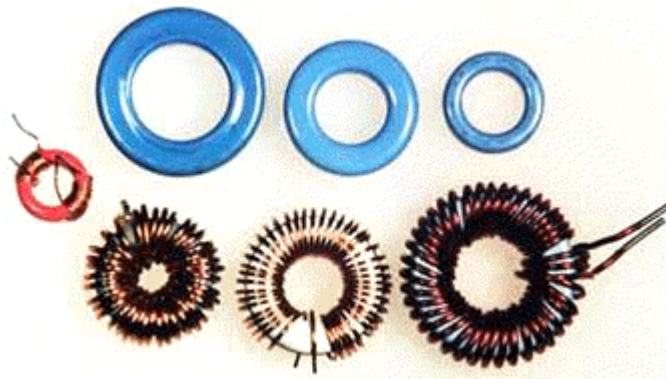


Figure: 36.3 - Photo de quelques bobines toroïdales

Pour des raisons de symétrie, il est clair que les lignes d'induction magnétique forment des cercles concentriques à l'intérieur de la bobine. Appliquons la loi d'Ampère au trajet d'intégration circulaire de rayon r :

$$\oint \vec{B} \cdot d\vec{l} = \mu_0 I \quad (36.17)$$

C'est-à-dire:

$$B \cdot 2\pi r = \mu_0 I = \mu_0 I_0 N \quad (36.18)$$

Il s'ensuit que:

$$B = \frac{\mu_0}{2\pi} \frac{I_0 N}{r} \quad (36.19)$$

Ainsi, contrairement à B l'intérieur d'un solénoïde, B n'est pas constant à l'intérieur de la bobine toroïdale.

1.3. ÉLECTRO-AIMANT

Déterminons donc par exemple (important et intéressant) le champ magnétique dans l'entrefer de longueur L_a et de section S_a d'un électro-aimant d'une longueur L_{Fe} et de section S_{Fe} comme représenté ci-dessous:

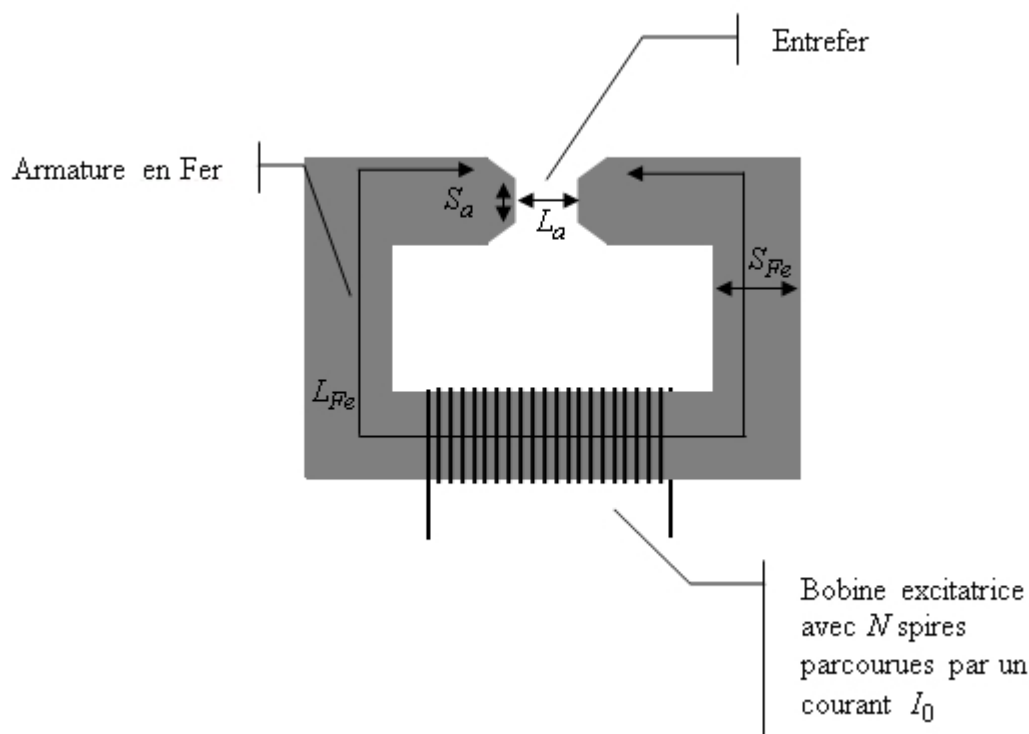


Figure: 36.4 - Exemple d'électro-aimant de laboratoire d'école

La loi d'Ampère nous donne dans le vide:

$$\oint_{\Sigma} \vec{B} \cdot d\vec{l} = \mu_0 \sum_k I_k \quad (36.20)$$

dans le cas de l'électro-aimant, nous pouvons écrire que la circulation du champ est la somme de la circulation du champ de l'entrefer et de l'aimant lui-même:

$$\oint_{\Sigma} \vec{H} \cdot d\vec{l} = H_{Fe} \cdot L_{Fe} + H_a \cdot L_a = \sum_k I_k = N \cdot I_0 \quad (36.21)$$

où N correspond au nombre de boucles de courant entourant l'aimant et qui permet la production du champ magnétique.

Nous avons par définition:

$$H_{Fe} = \frac{B_{Fe}}{\mu_{Fe}} \text{ et } H_a = \frac{B_a}{\mu_a} \quad (36.22)$$

d'où:

$$\oint_{\Sigma} \vec{H} \cdot d\vec{l} = H_{Fe} \cdot L_{Fe} + H_a \cdot L_a = \frac{B_{Fe}}{\mu_{Fe}} \cdot L_{Fe} + \frac{B_a}{\mu_a} \cdot L_a = N \cdot I_0 \quad (36.23)$$

Si l'entrefer n'est pas trop grand $L_{Fe} \gg L_a$ nous pouvons écrire:

$$B_a \approx B_{Fe} \quad (36.24)$$

alors:

$$N \cdot I_0 = B_a \cdot \left(\frac{L_{Fe}}{\mu_{Fe}} + \frac{L_a}{\mu_a} \right) \Leftrightarrow B_a = \frac{N \cdot I_0}{\frac{L_{Fe}}{\mu_{Fe}} + \frac{L_a}{\mu_a}} = \frac{N \cdot I_0}{\frac{L_{Fe}}{\mu_{Fe} \cdot \mu_0} + \frac{L_a}{\mu_a \cdot \mu_0}} \quad (36.25)$$

d'où:

$$B_a \cong \frac{\mu_0 \cdot N \cdot I_0}{\frac{L_{Fe}}{\mu_{Fe}} + L_a} \quad (36.26)$$

La relation est la même pour un électro-aimant ayant deux bobines! Le lecteur remarquera au passage que cette relation peut aussi servir expérimentalement dans le cas où nous cherchons à déterminer la valeur de la perméabilité magnétique relative du Fer quand tous les autres paramètres sont connus.

1.3.1. FORCE D'UN AIMANT OU ÉLECTRO-AIMANT

Si nous avons connaissance de la norme du champ magnétique B produit par un aimant à sa surface, nous pouvons calculer avec une certaine approximation la force nécessaire pour le décoller d'une surface en Fer.

Pour cela, nous noterons F la force nécessaire pour faire décoller l'aimant à une distance d d'une surface de Fer. Nous supposons la distance d suffisamment petite pour que l'on puisse accepter que dans tout le volume situé entre l'aimant et le Fer, le champ magnétique est constant.

Ainsi, le travail fourni par la force F est (cf. chapitre de Mécanique Classique):

$$W = F \cdot d \quad (36.27)$$

Ce travail s'est transformé en énergie du champ magnétique dans le volume créé entre l'aimant et le Fer. La densité volumique d'énergie due au champ magnétique dans l'air étant (cf. chapitre d'Électrodynamique):

$$u_B = \frac{1}{2\mu} B^2 \quad (36.28)$$

Le volume de l'espace créé entre l'aimant et le Fer étant égal à $V = S \cdot d$ où S est la surface de l'aimant qui était collée au Fer. Nous avons alors l'équivalence dimensionnelle suivante:

$$F \cdot d = \left(\frac{1}{2\mu_0} B^2 \right) \cdot S \cdot d \quad (36.29)$$

Nous en déduisons la force de contact pour de petites valeurs de d :

$$F = \left(\frac{1}{2\mu_0} B^2 \right) \cdot S \quad (36.30)$$

où B est la valeur limite du champ magnétique qui amène notre matériau à se coller à l'aimant (de façon à ce qu'en soulevant l'aimant, le matériau associé suive).

Si nous regardons un électro-aimant d'élévation de rayon 0.75 [m] capable de soulever 200 [kg]:



Figure: 36.5 - Électro-aimant d'élévation

Nous avons alors:

$$B = \sqrt{\frac{2 \cdot \mu_0 \cdot m \cdot g}{S}} = \sqrt{\frac{2 \cdot \mu_0 \cdot m \cdot g}{\pi R^2}} \cong 0.05 [T] \quad (36.31)$$

Il est possible d'utiliser aussi grossièrement le même calcul pour déterminer le champ magnétique de l'électro-aimant du jouet ludique suivant mondialement connu par les passionnés de physique:



Figure: 36.6 - Électro-aimant d'élévation ludique

2. RELATION DE MAXWELL-AMPÈRE

Soit $\vec{J}$ la densité de courant en un point quelconque de l'espace dans le cas d'une distribution à trois dimensions et soit S une surface fermée qui s'appuie sur un contour Γ quelconque. Le courant I qui traverse Γ est bien évidemment donné par:

$$I = \oint_S \vec{J} \circ d\vec{S} \quad (36.32)$$

D'après la loi d'Ampère, la circulation du champ magnétique le long de Γ est égale à cette intégrale. Elle peut donc prendre ici, selon le choix du contour Γ , une infinité de valeurs variables de façon continue. D'autre part, le théorème de Stokes (cf. chapitre de Calcul Vectoriel) fournit que:

$$\oint_S \vec{B} \circ d\vec{l} = \oint_S \vec{\nabla} \times \vec{B} \circ d\vec{S} \quad (36.33)$$

d'où:

$$\oint_S \vec{B} \circ d\vec{l} = \mu_0 I = \mu_0 \oint_S \vec{J} \circ d\vec{S} = \oint_S \vec{\nabla} \times \vec{B} \circ d\vec{S} \quad (36.34)$$

et nous en ressortons finalement que:

$$\boxed{\vec{\nabla} \times \vec{B} = \mu_0 \vec{J}} \quad (36.35)$$

Nous pouvons faire une comparaison osée de ce résultat avec la relation ci-dessous (démontrée dans le chapitre d'Électrodynamique), par extension de la charge statique et de la charge dynamique:

$$\vec{\nabla} \circ \vec{E} = \frac{\rho}{\epsilon} \quad (36.36)$$

qui n'est autre que la première équation de Maxwell (cf. chapitre d'Électrodynamique). Dès lors, comme nous l'avons vu dans le chapitre d'Électrostatique, nous avons:

$$\vec{E} = \frac{1}{4\pi\epsilon} \iiint_V \frac{\rho(\vec{r}) dV}{r^2} \quad (36.37)$$

Par analogie, l'idée est de poser (cette hypothèse se vérifie un peu plus bas par les résultats remarquables obtenus):

$$\vec{B} = \frac{\mu_0}{4\pi} \oint_V \frac{J(\vec{r}) dV}{r^2} \quad (36.38)$$

relation que nous pouvons écrire de manière plus élégante en supposant le courant non dépendant de la position de l'observateur dans l'espace et colinéaire au vecteur perpendiculaire à la surface traversée:

$$\vec{B} = \frac{\mu_0}{4\pi} \oint_V \frac{J(\vec{r}) dV}{r^2} = \frac{\mu_0}{4\pi} \oint_V \frac{\frac{I(\vec{r})}{S} dV}{r^2} = \frac{\mu_0}{4\pi} \oint_{\Gamma} \frac{I(\vec{r}) dl}{r^2} = \frac{\mu_0}{4\pi} I \oint_{\Gamma} \frac{d\vec{l}}{r^2} \quad (36.39)$$

où Γ représente le périmètre du fil dans lequel le courant I circule.

3. LOI DE BIOT-SAVART

Du dernier développement, nous tirons donc:

$$d\vec{B} = \frac{\mu_0}{4\pi} \frac{I \cdot d\vec{l}}{r^2} \quad (36.40)$$

Rappelez-vous qu'à la dernière étape de notre développement précédent (nous l'avons précisé implicitement) le chemin d'intégration est perpendiculaire au courant! Mais le champ magnétique ne peut pas être nul en tout point de la ligne du courant. Dès lors, nous sommes amenés à écrire ce qui est caché:

$$d\vec{B} = \frac{\mu_0 \cdot I \cdot d\vec{l}}{4 \cdot \pi \cdot r^2} \cdot 1 = \frac{\mu_0 \cdot I \cdot d\vec{l}}{4 \cdot \pi \cdot r^2} \cdot \sin\left(\frac{\pi}{2}\right) \quad (36.41)$$

La relation ci-dessus nous permet donc, par extension, d'écrire sous une forme plus générale:

$$\boxed{\vec{B} = \frac{\mu_0 \cdot I \cdot \sin(\alpha)}{4 \cdot \pi \cdot r^2} \cdot \vec{l} = \frac{\mu_0 \cdot I \cdot \vec{l} \times \vec{r}}{4 \cdot \pi \cdot r^3}} \quad (36.42)$$

qui n'est autre que la "**loi de Biot-Savart**" souvent présentée en premier dans les classes scolaires comme début d'étude du magnétisme (à l'origine elle a été déterminée expérimentalement par Biot et Savart avec l'aide mathématique de Laplace).

Cette dernière relation peut tout aussi bien s'écrire (forme très importante):

$$d\vec{B} = \frac{\mu_0}{4\pi r^3} I \cdot d\vec{l} \times \vec{r} = \frac{\mu_0}{4\pi r^3} \frac{dQ}{dt} \cdot d\vec{l} \times \vec{r} = \frac{\mu_0}{4\pi r^3} dQ \cdot \frac{d\vec{l}}{dt} \times \vec{r} = \frac{\mu_0 dQ}{4\pi r^3} \vec{v} \times \vec{r} \quad (36.43)$$

Donc:

$$\vec{B} = \frac{\mu_0 Q}{4\pi r^3} \vec{v} \times \vec{r} \quad (36.44)$$

Nous retrouvons ici l'approximation non relativiste du champ magnétique tel que nous l'avons déterminé lors de notre étude de la relativité restreinte (cf. chapitre de Relativité Restreinte), où nous avons démontré que:

$$\vec{B} = \frac{Q}{c^2 \epsilon_0 4\pi r^3} \frac{(1 - v^2/c^2)}{[1 - (v^2/c^2) \sin^2 \theta]^{3/2}} \cdot \vec{v} \times \vec{r} \quad (36.45)$$

Une autre forme importante de l'expression du champ magnétique est:

$$d\vec{B} = \frac{\mu_0}{4\pi} \frac{I \cdot d\vec{l} \times \vec{r}}{r^3} = \frac{\mu_0}{4\pi} \frac{(J \cdot S) \cdot d\vec{l} \times \vec{r}}{r^3} \quad (36.46)$$

Comme la densité de courant $\vec{J}$ est colinéaire à $d\vec{l}$, nous pouvons écrire:

$$d\vec{B} = \frac{\mu_0}{4\pi} \frac{(Sdl) \cdot \vec{J} \times \vec{r}}{r^3} = \frac{\mu_0}{4\pi} \frac{dV \cdot \vec{J} \times \vec{r}}{r^3} = \frac{\mu_0}{4\pi} \frac{\vec{J} \times \vec{r}}{r^3} dV \quad (36.47)$$

Donc:

$$\vec{B} = \frac{\mu_0}{4\pi} \oint \frac{\vec{J} \times \vec{r}}{r^3} dV \quad (36.48)$$

Une remarque importante s'impose à notre niveau du discours: dans le cadre des études scolaires pré-universitaires, les formulations mathématiques des champs magnétique et électrique sont considérées comme des lois indémontrables d'où l'on tire plus tard les équations de Maxwell (de plus les développements ne sont pas des plus esthétiques et rigoureux). L'aspect totalement expérimental de relations aussi importantes peut donner une image négative de la physique théorique aux étudiants. Il convient dès lors de préciser que lors des études universitaires, nous avons une approche juste un peu moins pragmatique.

Effectivement, nous postulons l'équation de Schrödinger (cf. chapitre de Physique Quantique Ondulatoire) dont nous nous servons pour démontrer la formulation non relativiste de la loi de Coulomb à l'aide de la théorie de Yukawa (cf. chapitre de Physique Quantique Des Champs). Ensuite, pendant l'étude de la relativité restreinte (cf. chapitre de Relativité Restreinte), nous déterminons la forme relativiste de la loi de Coulomb. Ensuite, nous admettons l'existence du champ magnétique dont l'expression est donnée expérimentalement par la force de Lorentz (voir plus bas dans ce chapitre) et de par les propriétés des transformations de Lorentz et de la connaissance de l'expression relativiste de la loi de Coulomb, nous déterminons l'expression relativiste du champ magnétique. Ensuite, par approximation non relativiste, nous tombons sur la loi de Biot-Savart. Cette manière de procéder est beaucoup mieux accueillie par les étudiants mais pas nécessairement accessible à tous les niveaux.

Revenons maintenant sur la loi de Biot-Savart. Un exemple important en astrophysique de la loi de Biot-Savart dans le cadre des jets de plasmas des disques d'accrétion sont les boucles de courant circulaires uniques (il faut y rajouter aussi la force de Laplace dans le cadre relativiste pour comprendre la dynamique de ces jets).

3.1. CHAMP MAGNÉTIQUE POUR UNE BOUCLE DE COURANT

La figure ci-dessous en représente un bon exemple:

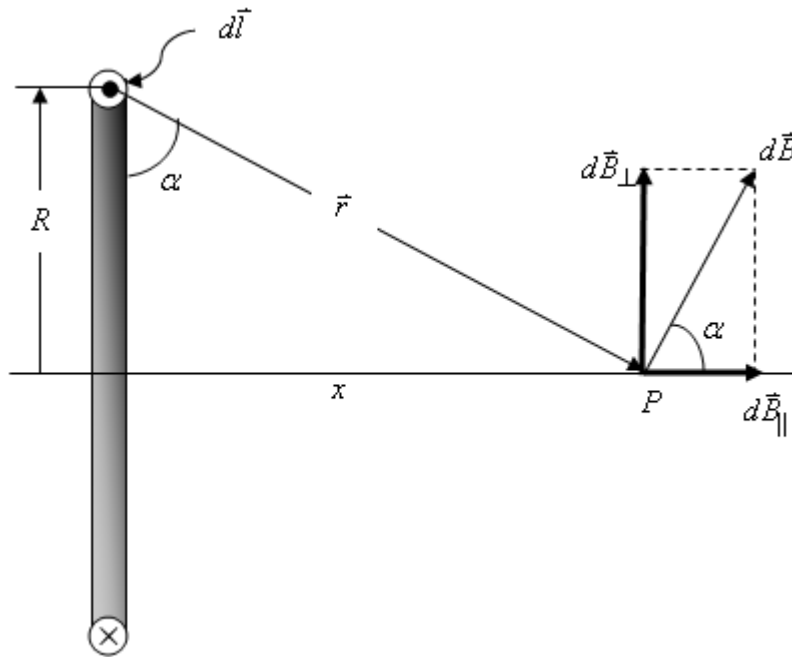


Figure: 36.7 - Champ magnétique pour une boucle de courant

Nous avons donc une boucle circulaire de rayon R parcourue par un courant d'intensité I . L'objectif étant de calculer $\vec{B}$ en un point P de l'axe de cette boucle.

Le vecteur $d\vec{l}$ correspondant à un courant élémentaire au sommet de la boucle sort perpendiculairement du plan de la page. L'angle θ entre ce vecteur et $\vec{r}$ est donc de $\pi/2$. Le plan formé par $d\vec{l}$ et $\vec{r}$ est normal à la figure. Le vecteur $d\vec{B}$ produit par ce courant élémentaire est normal à ce plan de par la forme de la loi de Biot-Savart. Il est donc dans le plan de la figure et à angle droit avec le vecteur $\vec{r}$ comme indiqué sur la figure.

Décomposons $d\vec{B}$ en deux parties: la première, $d\vec{B}_\parallel$ est le long de l'axe de la boucle et la seconde, $d\vec{B}_\perp$ est perpendiculaire à cet axe. Seule la composante $d\vec{B}_\parallel$ contribue à l'induction magnétique totale au point P . Il en est ainsi du fait que les composantes $d\vec{B}_\parallel$ de tous les courants élémentaires sont sur l'axe et qu'elles s'additionnent directement. Quant aux composantes $d\vec{B}_\perp$, elles sont dirigées dans différentes directions perpendiculairement à cet axe de sorte que, par symétrie, leur contribution est nulle sur cet axe (prenez vraiment garde à ce cas particulier!).

Nous obtenons:

$$B = \int d\vec{B}_\parallel \quad (36.49)$$

C'est une intégrale scalaire effectuée sur tous les courants élémentaires. Nous obtenons d'après la loi de Biot-Savart:

$$dB = \frac{\mu_0 I}{4\pi} \frac{dl \sin(\pi/2)}{r^2} \quad (36.50)$$

De plus, nous avons selon le schéma:

$$dB_{\parallel} = dB \cos \alpha \quad (36.51)$$

En combinant ces relations, nous obtenons:

$$dB_{\parallel} = \frac{\mu_0 I}{4\pi} \frac{\cos \alpha dl}{r^2} \quad (36.52)$$

La figure révèle que r et α ne sont pas des variables indépendantes. Nous pouvons les exprimer en fonction de la nouvelle variable x , la distance entre le centre de la boucle et le point P . Les relations entre ces variables sont:

$$\begin{aligned} r &= \sqrt{R^2 + x^2} \\ \cos \alpha &= \frac{R}{r} = \frac{R}{\sqrt{R^2 + x^2}} \end{aligned} \quad (36.53)$$

En substituant ces valeurs dans l'expression de $dB_{\parallel}$, nous obtenons:

$$dB_{\parallel} = \frac{\mu_0 IR}{4\pi (R^2 + x^2)^{3/2}} dl \quad (36.54)$$

Nous remarquons que, pour tous les courants élémentaires, I, R, x ont respectivement les mêmes valeurs. L'intégration de cette différentielle donne:

$$B = \int dB_{\parallel} = \frac{\mu_0 IR}{4\pi (R^2 + x^2)^{3/2}} \oint dl = \frac{\mu_0 IR}{4\pi (R^2 + x^2)^{3/2}} 2\pi R = \frac{\mu_0 IR^2}{2 (R^2 + x^2)^{3/2}} \quad (36.55)$$

Un point important de cette relation est en $x = 0$ où nous obtenons donc:

$$B = \frac{\mu_0 I}{2R} \quad (36.56)$$

Un autre cas d'application important de la loi de Biot-Savart consiste à reprendre l'exemple précédent, mais pour une forme continue plane quelconque et considérée comme ponctuelle et dont nous aimerions connaître la valeur du champ ailleurs que sur l'axe de symétrie. Les résultats seront très utiles lorsque nous étudierons la physique quantique corpusculaire et donc les propriétés magnétiques des métaux.

3.2. CHAMP MAGNÉTIQUE POUR UN FIL INFINI

Montrons aussi (c'est un exemple intéressant!) qu'à partir de la loi de Biot-Savart:

$$dB = \frac{\mu_0}{4\pi} \frac{Idl \sin \theta}{r^2} \quad (36.57)$$

nous pouvons aussi obtenir pour un fil rectiligne infini la relation:

$$B = \frac{\mu_0 I}{2\pi r} \quad (36.58)$$

que nous avons obtenue avec le théorème d'Ampère (ce qui montre l'équivalence entre les deux manières de calculer!).

Choisissons pour le fil rectiligne infini ci-dessous x comme variable:

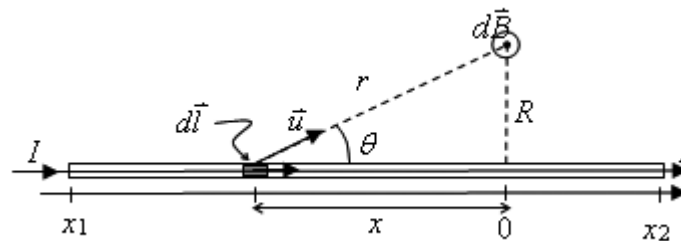


Figure: 36.8 - Fil rectiligne infini

Nous avons alors à partir de la figure ci-dessus:

$$\begin{aligned} dl &= dx \\ r &= \sqrt{x^2 + R^2} \\ \sin \theta &= \frac{R}{r} = \frac{R}{\sqrt{x^2 + R^2}} \end{aligned} \quad (36.59)$$

d'où:

$$dB = \frac{\mu_0}{4\pi} \frac{I dx \sin \theta}{r^2} = \frac{\mu_0}{4\pi} \frac{I dx}{r^2} \frac{R}{r} = \frac{\mu_0}{4\pi} IR \frac{dx}{(x^2 + R^2)^{\frac{3}{2}}} \quad (36.60)$$

En intégrant:

$$\begin{aligned} B &= \frac{\mu_0}{4\pi} IR \int_{x_1}^{x_2} \frac{1}{(x^2 + R^2)^{\frac{3}{2}}} dx = \frac{\mu_0}{4\pi} IR \left[\frac{x}{R^2 (x^2 + R^2)^{\frac{1}{2}}} \right]_{x_1}^{x_2} \\ &= \frac{\mu_0 I}{4\pi R} \left[\frac{x_2}{(x_2^2 + R^2)^{\frac{1}{2}}} - \frac{x_1}{(x_1^2 + R^2)^{\frac{1}{2}}} \right] \end{aligned} \quad (36.61)$$

Pour la suite, l'astuce consiste à utiliser la configuration suivante:

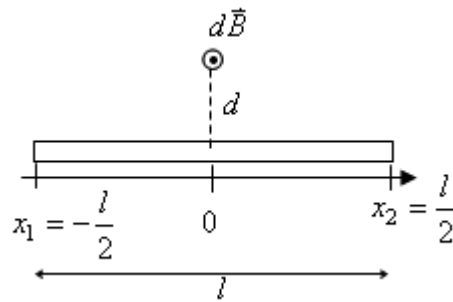


Figure: 36.9 - Morceau du fil

Donc:

$$x_2 = \frac{l}{2} \quad x_1 = -\frac{l}{2} \quad R = d \quad (36.62)$$

Ce qui nous donne:

$$B = \frac{\mu_0 I}{4\pi d} \left[\frac{l/2}{\left(\frac{l^2}{4} + d^2\right)^{\frac{1}{2}}} - \frac{-l/2}{\left(\frac{l^2}{4} + d^2\right)^{\frac{1}{2}}} \right] = \frac{\mu_0 I}{4\pi d} \left[\frac{l}{\left(\frac{l^2}{4} + d^2\right)^{\frac{1}{2}}} \right] \quad (36.63)$$

Après simplification:

$$B = \frac{\mu_0 I l}{2\pi d \left(l^2 + 4d^2\right)^{\frac{1}{2}}} = \frac{\mu_0 I l}{2\pi d l \left(1 + \frac{4d^2}{l^2}\right)^{\frac{1}{2}}} = \frac{\mu_0 I}{2\pi d \left(1 + \frac{4d^2}{l^2}\right)^{\frac{1}{2}}} \quad (36.64)$$

et donc quand la longueur du fil tend vers l'infini, nous avons alors:

$$B = \frac{\mu_0 I l}{2\pi d l \left(1 + \underbrace{\frac{4d^2}{l^2}}_{\rightarrow 0}\right)^{\frac{1}{2}}} \cong \frac{\mu_0 I}{2\pi d} \quad (36.65)$$

3.3. DIPÔLE MAGNÉTIQUE

Le dipôle magnétique a tout comme son homologue en électrostatique, une énorme importance dans l'étude des propriétés magnétiques des matériaux pour lesquelles il permet d'élaborer de bons modèles théoriques.

Avant de lire ce qui va suivre, nous conseillerions au lecteur (c'est même plus qu'un conseil) de lire absolument tout le développement du dipôle électrostatique rigide dans le chapitre d'Électrostatique. Effectivement, la plupart des calculs qui vont suivre comportent les mêmes raisonnements, développements et approximations mathématiques à quelques infimes nuances près. Nous n'avons dès lors pas souhaité refaire les mêmes calculs intermédiaires déjà présents lors du calcul du dipôle électrostatique (cependant, si vraiment il y a difficulté de la part du lecteur, nous sommes prêts à compléter... mais bon...)

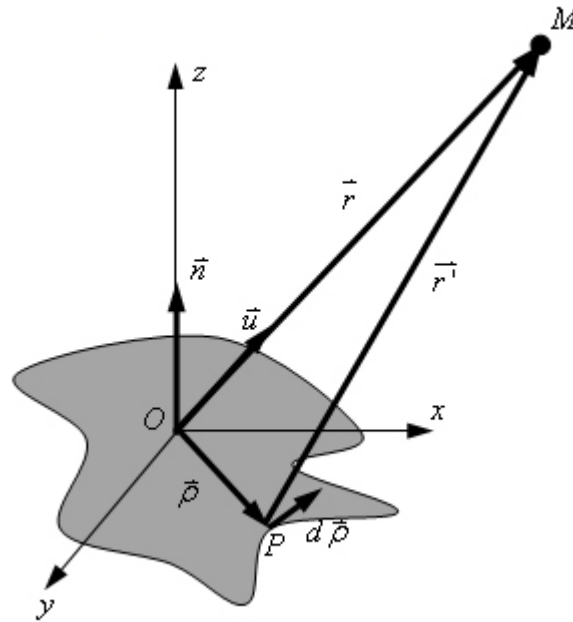
Le dipôle magnétique a une différence non négligeable relativement au cas pratique que nous nous imposons comme cadre d'étude... il n'y a pas 2 charges ! Effectivement, des charges au repos émettent en première approximation (c'est expérimental et... théorique) un champ magnétique intrinsèque beaucoup trop faible pour être considéré comme intéressant dans le cadre de l'étude des propriétés magnétiques des matériaux. Il convient cependant de préciser quelque chose d'intéressant (de sympa), les charges coulombiennes élémentaires sont parfois modélisées (à tort!) par les physiciens comme en rotation sur elles-mêmes (le "spin") et sont représentées comme une superposition de spires circulaires (tiens... une spire...) en infiniment petites ce qui fait qu'un observateur dans un référentiel au repos (au centre de la charge) peut interpréter la charge coulombienne globale comme étant un courant en déplacement dans les différentes spires, induisant ainsi un champ magnétique intrinsèque (joli non !?).

Bref, considérons une spire plane (tiens... encore une spire...), de forme quelconque, de centre O , parcourue par un courant permanent et constant I dont un des points de la spire est noté par P . Nous allons calculer le champ magnétique créé par cette spire en tout point M de l'espace, situé à grande distance r de la spire (précisément, à des distances grandes comparées à la taille de la spire).

Remarque: Personnellement il y a certaines étapes du calcul que je trouve... comment dire... de très loin pas convaincantes... mais bon... il y a tellement d'approximations que l'on est plus à ça près... hummm....

Nous posons:

$$\vec{r} = \overrightarrow{OM} \quad \vec{r}' = \overrightarrow{PM} \quad \vec{\rho} = \overrightarrow{OP} = \vec{r} - \vec{r}' \quad \vec{u} = \frac{\vec{r}}{r} \quad (36.66)$$



Nous allons donc utiliser la loi de Biot-Savart:

$$\vec{B} = \frac{\mu_0}{4\pi} \oint \frac{\vec{J} \times \vec{r}}{r^3} dV \quad (36.67)$$

sous l'hypothèse que le point M est donc situé à très grande distance de la spire. Ce qui nous donne le droit d'écrire:

$$\vec{B}(M) \cong \frac{\mu_0 I}{4\pi} \oint \frac{d\vec{l} \times \vec{r}'}{r^3} \quad (36.68)$$

Mais $d\vec{l} = d\vec{\rho}$ donc:

$$\vec{B}(M) \cong \frac{\mu_0 I}{4\pi} \oint \frac{d\vec{\rho} \times \vec{r}'}{r^3} \quad (36.69)$$

Évaluons le terme $\vec{r}'/r^3$ pour des points M situés à grande distance de la spire (au dénominateur nous avons utilisé le théorème du cosinus comme lors de notre étude dipôle électrostatique rigide dans le chapitre d'Électrostatique):

$$\begin{aligned} \frac{\vec{r}'}{r^3} &= \frac{\vec{r} - \vec{\rho}}{(r^2 + \rho^2 - 2\vec{r} \cdot \vec{\rho})^{3/2}} \cong \frac{\vec{r} - \vec{\rho}}{r^3 \left(1 - 2\frac{\vec{r} \cdot \vec{\rho}}{r^2}\right)^{3/2}} \cong \frac{\vec{r} - \vec{\rho}}{r^3} \left(1 + 3\frac{\vec{r} \cdot \vec{\rho}}{r^2}\right) \\ &\cong \frac{\vec{u}}{r^2} - \frac{\vec{\rho}}{r^3} + 3\frac{\vec{u} \cdot \vec{\rho}}{r^3} \vec{u} \end{aligned} \quad (36.70)$$

où nous avons fait comme pour le dipôle électrostatique rigide un développement limité à l'ordre 1.

Remarque: La dernière approximation est très grossière dans le sens qu'il s'agit d'un choix astucieux des termes à négliger pour arriver à un résultat esthétique visuellement et permettant de définir le moment magnétique dipolaire (voir un peu plus loin)...

En reportant cette expression dans la loi de Biot-Savart, nous obtenons:

$$\vec{B}(M) \cong \frac{\mu_0 I}{4\pi r^2} \left[\oint d\vec{\rho} \times \vec{u} - \frac{1}{r} \oint d\vec{\rho} \times \vec{\rho} + \frac{3}{r} \oint d\vec{\rho} \times \vec{u} (\vec{u} \circ \vec{\rho}) \right] \quad (36.71)$$

Évaluons séparément chaque terme intervenant dans la parenthèse:

$$1. \oint d\vec{\rho} \times \vec{u} = \left(\oint d\vec{\rho} \right) \times \vec{u} = (\vec{\rho}(P_0) - \vec{\rho}(P_0)) \times \vec{u} = \vec{0} \times \vec{u} = \vec{0}$$

puisque le vecteur $\vec{u}$ est indépendant du point P sur la spire et que nous faisons une intégration curviligne sur toute la spire, en revenant au point de départ.

$$2. -\frac{1}{r} \oint d\vec{\rho} \times \vec{\rho}$$

De par les propriétés du produit vectoriel:

$$-\frac{1}{r} \oint d\vec{\rho} \times \vec{\rho} = \frac{1}{r} \oint \vec{\rho} \times d\vec{\rho} \quad (36.72)$$

Or puisque $d\vec{\rho}$ et $\vec{\rho}$ sont perpendiculaires et dans un même plan, nous avons $\vec{\rho} \times d\vec{\rho}$ qui est la surface infinitésimale dS' d'un rectangle et cela ne représente rien étant donné que l'abscisse est curviligne par rapport à O . Effectivement:

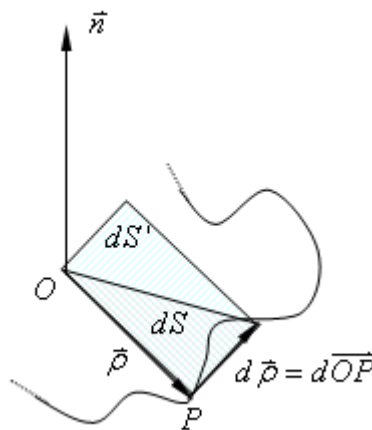


Figure: 36.10 - Représentation de la perpendicularité des deux vecteurs

Donc, nous pouvons écrire:

$$\vec{\rho} \times d\vec{\rho} = dS' \vec{n} = 2dS \vec{n} \Leftrightarrow \vec{\rho} \times d\vec{\rho} = 2dS \vec{n} \quad (36.73)$$

où $\vec{n}$ est le vecteur normal au plan de la spire (vecteur de base de l'axe Z). Ce résultat est général, valable quelle que soit la surface.

D'où:

$$-\frac{1}{r} \oint d\vec{\rho} \times \vec{\rho} = \frac{1}{r} \oint \vec{\rho} \times d\vec{\rho} = \frac{1}{r} \oint 2dS\vec{n} = \frac{1}{r} \vec{n} 2 \oint dS = \frac{2S}{r} \vec{n} \quad (36.74)$$

$$3. \oint d\vec{\rho} \times \vec{u}(\vec{u} \circ \vec{\rho}) = -\vec{u} \times \oint d\vec{\rho}(\vec{u} \circ \vec{\rho})$$

de par les propriétés du produit vectoriel (cf. chapitre de Calcul Vectoriel).

Nous allons utiliser ces relations pour calculer l'intégrale inconnue du début. Si nous décomposons les vecteurs $\vec{\rho}$ et $\vec{u}$ dans la base $(\vec{e}_1, \vec{e}_2)$ engendrant le plan de la spire, nous obtenons:

$$d\vec{\rho}(\vec{u} \circ \vec{\rho}) = d\rho_1(\rho_1 u_1 + \rho_2 u_2)\vec{e}_1 + d\rho_2(\rho_1 u_1 + \rho_2 u_2)\vec{e}_2 \quad (36.75)$$

puisque $d\rho_3 = 0$ et $\rho_3 = 0$.

Nous avons aussi:

$$\begin{aligned} \oint \rho_1 u_1 d\rho_1 &= \frac{u_1}{2} (\rho_1^2(P_0) - \rho_1^2(P_0)) = 0 \\ \oint \rho_2 u_2 d\rho_2 &= \frac{u_2}{2} (\rho_2^2(P_0) - \rho_2^2(P_0)) = 0 \end{aligned} \quad (36.76)$$

D'où:

$$\oint d\vec{\rho}(\vec{u} \circ \vec{\rho}) = u_2 \oint \rho_2 d\rho_1 \vec{e}_1 + u_1 \oint \rho_1 d\rho_2 \vec{e}_2 = u_2 \vec{e}_1 \oint \rho_2 d\rho_1 + u_1 \vec{e}_2 \oint \rho_1 d\rho_2 \quad (36.77)$$

Rappelons que:

$$\vec{\rho} \times d\vec{\rho} = dS' \vec{n} = 2dS\vec{n} \Leftrightarrow \frac{\vec{\rho} \times d\vec{\rho}}{2} = dS\vec{n} \quad (36.78)$$

Sous forme de composantes (seulement la troisième est non nulle puisque $d\rho_3 = 0$ et $\rho_3 = 0$), nous avons:

$$\frac{1}{2}(\rho_1 d\rho_2 - \rho_2 d\rho_1) = dS \quad (36.79)$$

d'où:

$$\rho_1 d\rho_2 - \rho_2 d\rho_1 = 2dS \Leftrightarrow \rho_1 d\rho_2 = dS, \rho_2 d\rho_1 = -dS \quad (36.80)$$

Ce qui nous amène à écrire:

$$\oint d\vec{\rho}(\vec{u} \circ \vec{\rho}) = u_2 \vec{e}_1 \oint \rho_2 d\rho_1 + u_1 \vec{e}_2 \oint \rho_1 d\rho_2 = -u_2 \vec{e}_1 \oint dS + u_1 \vec{e}_2 \oint dS \quad (36.81)$$

Soit:

$$-(u_2 \vec{e}_1 S - u_1 \vec{e}_2 S) = -S(u_2 \vec{e}_1 - u_1 \vec{e}_2) = -S \left(u_2 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - u_1 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right) = S \begin{pmatrix} -u_2 \\ u_1 \\ 0 \end{pmatrix} \quad (36.82)$$

Remarquons que cette dernière relation est égale à:

$$S \cdot \vec{n} \times \vec{u} = S \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \times \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = S \begin{pmatrix} -1 \cdot u_2 \\ 1 \cdot u_1 \\ 0 \end{pmatrix} = S \begin{pmatrix} -u_2 \\ u_1 \\ 0 \end{pmatrix} \quad (36.83)$$

Donc au final:

$$\oint d\vec{\rho} (\vec{u} \circ \vec{\rho}) = S \cdot \vec{n} \times \vec{u} \quad (36.84)$$

En rassemblant ces résultats, nous obtenons pour le champ magnétique:

$$\vec{B}(M) \cong \frac{\mu_0 I}{4\pi r^2} \left[\frac{2S}{r} \vec{n} - \frac{3\vec{u}}{r} \times (S \vec{n} \times \vec{u}) \right] \quad (36.85)$$

Nous voyons donc apparaître une grandeur importante car décrivant complètement la spire vue depuis une grande distance, à savoir le "**moment magnétique dipolaire local**" ou plus simplement "**moment dipolaire magnétique**":

$$\boxed{\vec{\mu}_I = I \cdot S \cdot \vec{n}} \quad (36.86)$$

souvent noté aussi par un M stylisé par certains auteurs. Nous avons alors la relation antérieure qui peut s'écrire:

$$\vec{B}(M) \cong \frac{\mu_0}{4\pi r^2} \left[\frac{2\vec{\mu}_I}{r} - \frac{3\vec{u}}{r} \times (\vec{\mu}_I \times \vec{u}) \right] \cong \frac{\mu_0}{4\pi} \left[\frac{2\vec{\mu}_I}{r^3} - \frac{3\vec{u}}{r^3} \times (\vec{\mu}_I \times \vec{u}) \right] \quad (36.87)$$

En faisant usage de la propriété suivante du double produit vectoriel (cf. chapitre de Calcul Vectoriel):

$$\vec{u} \times (\vec{\mu}_I \times \vec{u}) = (\vec{u} \circ \vec{u}) \vec{\mu}_I - (\vec{u} \circ \vec{\mu}_I) \vec{u} = \vec{\mu}_I \cdot (\vec{u} \circ \vec{u}) - \vec{u} \cdot (\vec{u} \circ \vec{\mu}_I) \quad (36.88)$$

Nous obtenons alors une autre forme de l'expression du champ magnétique approximatif créé par un dipôle:

$$\begin{aligned}\vec{B}(M) &\cong \frac{\mu_0}{4\pi} \left[\frac{2\vec{\mu}_l}{r^3} - \frac{3\vec{u}}{r^3} \times (\vec{\mu}_l \times \vec{u}) \right] \cong \frac{\mu_0}{4\pi r^3} \left[2\vec{\mu}_l - 3(\vec{\mu}_l \cdot (\vec{u} \circ \vec{u}) - \vec{u} \cdot (\vec{u} \circ \vec{\mu}_l)) \right] \\ &\cong \frac{\mu_0}{4\pi r^3} \left[2\vec{\mu}_l - 3\vec{\mu}_l \cdot \underbrace{(\vec{u} \circ \vec{u})}_{=1} + 3\vec{u} \cdot (\vec{u} \circ \vec{\mu}_l) \right] \cong \frac{\mu_0}{4\pi r^3} \left[3\vec{u} \cdot (\vec{u} \circ \vec{\mu}_l) - \vec{\mu}_l \right]\end{aligned}\quad (36.89)$$

C'est sous cette dernière forme que l'on retrouve le plus souvent l'expression du moment magnétique dans la littérature.

À comparer (pour le fun) avec l'expression du champ électrique pour un dipôle électrique rigide:

$$\begin{aligned}\vec{B} &\cong \frac{\mu_0}{4\pi r^3} \left[3\vec{u} \cdot (\vec{u} \circ \vec{\mu}_l) - \vec{\mu}_l \right] \\ \vec{E} &\cong \frac{1}{4\pi \epsilon r^3} \left[3\vec{u}_r \cdot (\vec{u}_r \circ \vec{p}) - \vec{p} \right]\end{aligned}\quad (36.90)$$

et donc nous voyons qu'il y a correspondance parfaite.

Nous sommes quand même arrivés à mettre cela sous une forme assez identique et esthétique après quelques approximations...

Nous avons aussi:

$$\vec{\mu}_l \circ \vec{u} = \|\vec{\mu}_l\| \|\vec{u}\| \cos \theta = \mu_l \cos \theta \quad (36.91)$$

d'où:

$$\vec{B}(M) \cong \frac{\mu_0}{4\pi r^3} \left[3\vec{u} \mu_l \cos \theta - \vec{\mu}_l \right] \quad (36.92)$$

L'origine du champ magnétique d'un matériau quelconque doit être microscopique. En utilisant le modèle de Bohr de l'atome (cf. chapitre de Physique Quantique Corpusculaire), nous pouvons nous convaincre que les atomes (du moins certains) ont un moment magnétique dipolaire intrinsèque. Effectivement, le modèle de Bohr de l'atome d'Hydrogène consiste en un électron de charge $q = -e$ en mouvement (circulaire) autour d'un noyau centre (un proton) avec une période $T = 2\pi\omega$.

Si nous regardons sur des échelles de temps longues par rapport à T , tout se passe comme s'il y avait un courant:

$$I = \frac{q}{T} = \frac{q\omega}{2\pi} \quad (36.93)$$

Nous avons donc une sorte de spire circulaire, de rayon moyen la distance moyenne au proton, c'est à dire le rayon de Bohr r_0 . L'atome d'Hydrogène aurait donc un moment magnétique intrinsèque:

$$\vec{\mu}_l = IS\vec{n} = \frac{q\omega}{2} r_0^2 \vec{n} = \frac{q}{2m} (m\omega r_0^2 \vec{n}) = \frac{q}{2m} \vec{b} \quad (36.94)$$

où $\vec{L}$ est le moment cinétique de l'électron et $q/2m$ le "facteur gyromagnétique" (ce résultat est très important pour le modèle de Langevin du diamagnétisme). Ce raisonnement peut se généraliser aux autres atomes. En effet, un ensemble de charges en rotation autour d'un axe vont produire un moment magnétique proportionnel au moment cinétique total. Cela se produit même si la charge totale est nulle (matériau ou atome neutre): ce qui compte c'est l'existence (scalaire) d'un courant.

Du coup, nous pouvons expliquer qualitativement les propriétés magnétiques des matériaux en fonction de l'orientation des moments magnétiques des atomes qui les composent:

- Matériaux amagnétiques: ce sont les matériaux où les moments sont distribués aléatoirement, il n'y a pas de champ magnétique intrinsèque.
- Matériaux diamagnétiques: ce sont les matériaux qui soumis à un champ magnétique, ont leur moments qui s'opposent à celui-ci et sont donc repoussés (très faiblement) par les aimants. Ils induisent donc un moment magnétique opposé à la direction du champ magnétique.
- Matériaux paramagnétiques: ce sont les matériaux pour lesquels les moments peuvent s'orienter dans la direction d'un champ magnétique extérieur et pouvant donc être ainsi aimantés (attirés) momentanément. Ils induisent donc un moment magnétique dans la direction du champ magnétique.
- Matériaux ferromagnétiques: ce sont les matériaux dont les moments sont déjà orientés dans une direction particulière, de façon permanente (aimants naturels).

Remarque: La Terre est connue pour avoir un champ magnétique dipolaire, où le pôle Nord magnétique correspond au pôle Sud géographique (à un angle près). Au niveau macroscopique, l'explication de l'existence du champ magnétique observé sur les étoiles est encore aujourd'hui loin d'être satisfaisante. La théorie de "l'effet dynamo" essaie de rendre compte des champs observés par la présence de courants, essentiellement azimutaux, dans le coeur des astres. Plusieurs faits connus restent partiellement non éclaircis:

- Les cycles magnétiques: le Soleil a un champ magnétique à grande échelle qui ressemble à celui de la Terre, approximativement dipolaire. Cependant, il y a une inversion de polarité tous les 11 ans (sur 11 ans). Pour la Terre, on a pu mettre en évidence qu'il y avait eu une inversion il y a environ 700'000 ans.
- Non-alignement avec le moment cinétique de l'astre: s'il est de l'ordre d'une dizaine de degrés pour la Terre, il est perpendiculaire pour Neptune!

4. LOI DE LORENTZ

En électrostatique, nous avons calculé la force exercée par une ou un ensemble de charges au repos sur une charge immobile ou en mouvement. La force exercée s'écrivait alors de la manière suivante:

$$\vec{F} = q \cdot \vec{E}(\vec{r}) \quad (36.95)$$

Dans le cas le plus général, où les charges agissantes sont en mouvement, la force qu'elles exercent sur une charge ponctuelle q placée en un point de l'espace est la somme de deux termes: l'un qui est indépendant de la vitesse $\vec{v}$ de cette charge, l'autre qui en dépend. Voici comment s'écrit cette relation:

$$\vec{F} = q \left(\vec{E} + \vec{v} \times \vec{B} \right) \quad (36.96)$$

qui n'est autre que la "**loi de Lorentz**" ou "**force de Lorentz**".

Pour démontrer cette relation, nous allons poser deux hypothèses, mais avant il est important d'informer le lecteur que cette démonstration nécessite des outils mathématiques non nécessairement évidents (il faut avoir lu le chapitre de Mécanique Analytique et de Physique Quantique Ondulatoire pour comprendre):

H1. Soit une particule ponctuelle non-relativiste de masse m , de position $x_i (i = 1, 2, 3)$ et de vitesse $\dot{x}_i (i = 1, 2, 3)$; nous supposons qu'elle est soumise à une force $\vec{F}$ et qu'elle satisfait les équations de Newton:

$$m\ddot{\vec{x}} = \vec{F}(\vec{x}, \vec{v}, t) \quad (36.97)$$

avec les relations de commutations suivantes (cf. chapitre de Physique Quantique Ondulatoire):

$$\begin{aligned} [x_i, x_j] &= 0 \\ m[x_i, v_j] &= i\hbar\delta_{ij} \end{aligned} \quad (36.98)$$

Il faut bien voir que la dernière relation est une hypothèse et qu'elle n'est pas équivalente aux règles de commutation que nous avons vues en physique quantique entre positions et impulsions!

H2. Il existe des champs $\vec{E}(\vec{x}, t)$ et $\vec{B}(\vec{x}, t)$, ne dépendant pas des vitesses, tels que:

$$\vec{F} = q \left(\vec{E} + \vec{v} \times \vec{B} \right) \quad (36.99)$$

et qui vérifient les équations de Maxwell (cf. chapitre d'Électrodynamique):

$$\begin{aligned} \vec{\nabla} \times \vec{E} &= -\frac{\partial \vec{B}}{\partial t} \\ \vec{\nabla} \cdot \vec{B} &= 0 \end{aligned} \quad (36.100)$$

À un niveau classique, nous exprimons les hypothèses de commutation en utilisant la correspondance commutateurs-crochets de Poisson (cf. chapitre de Mécanique Analytique), soit:

$$\begin{aligned} \{x_i, x_j\} &= 0 \\ m\{x_i, \dot{x}_j\} &= \delta_{ij} \end{aligned} \quad (36.101)$$

avec (rappel):

$$\{A, B\} = \sum_{i=1}^3 \left\{ \frac{\partial A}{\partial x_i} \frac{\partial B}{\partial p_i} - \frac{\partial A}{\partial p_i} \frac{\partial B}{\partial x_i} \right\} \quad (36.102)$$

Maintenant, nous définissons un potentiel vecteur $\vec{A}$ (cf. chapitre d'Électrodynamique) tel que:

$$p_j = m\dot{x}_j + qA_j \quad (36.103)$$

alors l'hypothèse $(\{x_i, p_j\} = \delta_{ij})$ de commutation peut s'écrire pour $i \neq j$:

$$\{x_i, p_j\} = x_i(m\dot{x}_j + qA_j) - (m\dot{x}_j + qA_j)x_i = 0 = \{x_i, A_j\} \quad (36.104)$$

donc nous pouvons dire que $\vec{A}$ ne dépend que de $\vec{x}$ et t puisqu'il commute identiquement à p_j .

De plus, nous savons que la mécanique classique admet une formulation lagrangienne (équivalente aux équations de Newton) pour laquelle les équations de la mécanique deviennent (cf. chapitre de Mécanique Analytique):

$$\begin{aligned} p_j &= \frac{\partial L}{\partial \dot{x}_j} \\ \dot{p}_j &= \frac{\partial L}{\partial x_j} \end{aligned} \quad (36.105)$$

où L désigne le lagrangien du système. Dès lors, avec:

$$p_j = m\dot{x}_j + qA_j(\vec{x}, t) \quad (36.106)$$

nous pouvons intégrer la relation:

$$p_j = \frac{\partial L}{\partial \dot{x}_j} \quad (36.107)$$

et nous obtenons:

$$L = \frac{1}{2} m \dot{x}_j^2 + q(\dot{x}_j A(\vec{x}, t) - \phi(\vec{x}, t)) \quad (36.108)$$

Le signe "-" de la constante d'intégration du potentiel vecteur se justifie pour être en cohérence avec ce que nous avons vu en théorie de Jauge (cf. chapitre d'Électrodynamique).

La seconde équation de Lagrange $\dot{p}_j = m\ddot{x}_j + q\dot{A}_j(\vec{x}, t)$ nous donne alors:

$$\dot{p}_j = m\ddot{x}_j + q\dot{A}_j = \frac{\partial L}{\partial x_j} \quad (36.109)$$

En développant un peu:

$$m\ddot{x}_j = \dot{p}_j - q\dot{A}_j \text{ et } \dot{p}_j = \frac{\partial L}{\partial x_j} = \frac{\partial}{\partial x_j} \left(\frac{1}{2} m \dot{x}_j^2 + q(\dot{x}_j A(\vec{x}, t) - \phi(\vec{x}, t)) \right) \quad (36.110)$$

Pour l'ensemble des coordonnées, cela donne sous forme condensée et en utilisant les outils de l'analyse vectorielle:

$$m\ddot{\vec{x}} = q \left(-\vec{\nabla} \cdot \phi - \frac{\partial \vec{A}}{\partial t} + \vec{x} \times (\vec{\nabla} \times \vec{A}) \right) \quad (36.111)$$

Donc:

$$m\vec{\ddot{x}} = \vec{F} = q \left(\vec{E} + \vec{\dot{x}} \times \vec{B} \right) \quad (36.112)$$

ou autrement écrit:

$$\boxed{\vec{F} = q \left(\vec{E} + \vec{v} \times \vec{B} \right)} \quad (36.113)$$

Nous retrouvons donc bien l'expression de la force de Lorentz où $\vec{E}$ et $\vec{B}$ sont donnés par:

$$\begin{aligned} \vec{E} &= -\vec{\nabla} \cdot \phi - \frac{\partial \vec{A}}{\partial t} \\ \vec{B} &= \vec{\nabla} \times \vec{A} \end{aligned} \quad (36.114)$$

comme nous l'avons vu en théorie de Jauges. Certes la démonstration est loin d'être évidente, mais elle est possible.

Arrêtons-nous un instant sur l'expression de la force de Lorentz. Nous voyons avec cette relation, qu'une charge immobile (ou non) dans un champ électrique subira une force qui lui donnera l'impulsion nécessaire à faire varier son énergie cinétique (nulle ou non nulle au départ). Cette constatation n'est cependant pas valable pour le champ magnétique. Effectivement, lorsque nous plaçons une charge immobile dans un champ magnétique, cette dernière ne subira aucune force du champ magnétique et donc ne verra pas son énergie cinétique varier. Si la particule chargée a une vitesse initiale non nulle, il s'ensuit que le champ magnétique va changer les composantes du vecteur vitesse mais pas la norme. Ainsi, nous avons pour habitude de dire que: "le champ magnétique ne travaille pas" (dans le sens que le champ magnétique ne va pas mettre en mouvement une particule chargée au repos ni changer la norme de sa vitesse).

Voyons mathématiquement comment nous pouvons montrer que le champ magnétique ne travaille pas.

Démonstration:

Nous savons que pour une particule chargée plongée dans un champ magnétique, nous avons:

$$\vec{F} = q\vec{v} \times \vec{B} = m\vec{a} = m \frac{d\vec{v}}{dt} \quad (36.115)$$

d'où:

$$\frac{d\vec{v}}{dt} = \frac{q}{m} \cdot \vec{v} \times \vec{B} \quad (36.116)$$

Et exprimons la variation temporelle de l'énergie cinétique:

$$\frac{d}{dt} E_c = \frac{d}{dt} \left(\frac{1}{2} m v^2 \right) = \frac{d}{dt} \left(\frac{1}{2} m \vec{v} \circ \vec{v} \right) = m \vec{v} \circ \frac{d\vec{v}}{dt} \quad (36.117)$$

et en substituant la dérivée de la vitesse par la relation antérieure, il vient:

$$\frac{d}{dt} E_c = m\vec{v} \circ \frac{d\vec{v}}{dt} = m\vec{v} \circ \left(\frac{q}{m} \vec{v} \times \vec{B} \right) = q \underbrace{\vec{v} \circ (\vec{v} \times \vec{B})}_{\vec{v} \times \vec{B} \perp \vec{v}} = q \cdot 0 = 0 \quad (36.118)$$

L'énergie cinétique de la particule ne change donc effectivement pas à cause du champ magnétique.

Maintenant, si nous nous intéressons uniquement au second terme de cette relation, nous pouvons arriver à démontrer la loi de Laplace:

Nous avons:

$$d\vec{F} = dq \cdot \vec{v} \times \vec{B} = \rho \cdot dV \cdot \vec{v} \times \vec{B} = \rho \cdot d\vec{l} \cdot dS \cdot \vec{v} \times \vec{B} \quad (36.119)$$

où ρ est la densité volumique de charge. Si $\vec{v}$ et $d\vec{l}$ sont supposés parallèles nous pouvons écrire que:

$$d\vec{F} = \rho \cdot v \cdot dS \cdot d\vec{l} \times \vec{B} \quad (36.120)$$

Une densité de courant nous permet de calculer la vitesse d'entraînement des porteurs de charges dans un conducteur. Le nombre d'électrons de conduction dans un fil est égal à:

$$n \cdot A \cdot l = \rho \cdot A \cdot l \quad (36.121)$$

où n est le nombre d'électrons de conduction par unité de volume et $A \cdot l$ le volume du fil (et donc A est l'aire de la section du conducteur)..

Une quantité de charges $q = \rho \cdot A \cdot l$ traverse un fil en un temps t donné par:

$$t = \frac{l}{v} \quad (36.122)$$

L'intensité I du courant étant définie par:

$$I = \frac{q}{t} = \frac{\rho \cdot A \cdot l}{l/v} = \rho \cdot A \cdot v \quad (36.123)$$

nous obtenons que:

$$v = \frac{J}{\rho} \quad (36.124)$$

De:

$$d\vec{F} = \rho \cdot v \cdot dS \cdot d\vec{l} \times \vec{B} \quad (36.125)$$

Nous pouvons maintenant tirer que:

$$d\vec{F} = J \cdot dS \cdot d\vec{l} \times \vec{B} \quad (36.126)$$

Enfin, nous trouvons que:

$$\boxed{\vec{F} = I \cdot \vec{l} \times \vec{B}} \quad (36.127)$$

qui est la "**loi de Laplace**" ou "**force de Laplace**" et qui dérive donc de la loi de Lorentz. Nous en déduisons dans les unités du champ magnétique:

$$\left[N \cdot m^{-1} \cdot A^{-1} \right] = \left[kg \cdot s^{-2} \cdot A^{-1} \right] = \left[kg \cdot s^{-2} \cdot s \cdot C^{-1} \right] = \left[kg \cdot s^{-1} \cdot C^{-1} \right] = [T] \quad (36.128)$$

où T est communément une unité tolérée dans l'usage appelée le "**Tesla**" (qui contient implicitement donc l'unité du Coulomb qui est à l'origine du champ magnétique!). Connaissant maintenant enfin explicitement l'unité du champ magnétique, nous pouvons déterminer les unités de la constante de perméabilité magnétique en repartant de:

$$B = \frac{\mu_0 \cdot I}{2 \cdot \pi \cdot r} \quad (36.129)$$

Il vient alors pour les unités de la constante de perméabilité magnétique:

$$\left[T \cdot m \cdot A^{-1} \right] = \left[N \cdot m^{-1} \cdot A^{-1} \cdot m \cdot A^{-1} \right] = \left[N \cdot A^{-2} \right] = \left[kg \cdot m \cdot s^{-2} \cdot A^{-2} \right] \quad (36.130)$$

Ceci étant fait, voyons quelques cas importants d'application de la loi de Lorentz:

4.1. EFFET HALL CLASSIQUE

Précédemment, nous avons étudié l'action d'une induction magnétique sur un circuit filiforme en ayant pour but de trouver l'expression des forces magnétiques appliquées à la matière même de ce circuit.

Portons maintenant notre attention sur les électrons de conductivité eux-mêmes, en nous plaçant dans le cas de la figure ci-dessous:

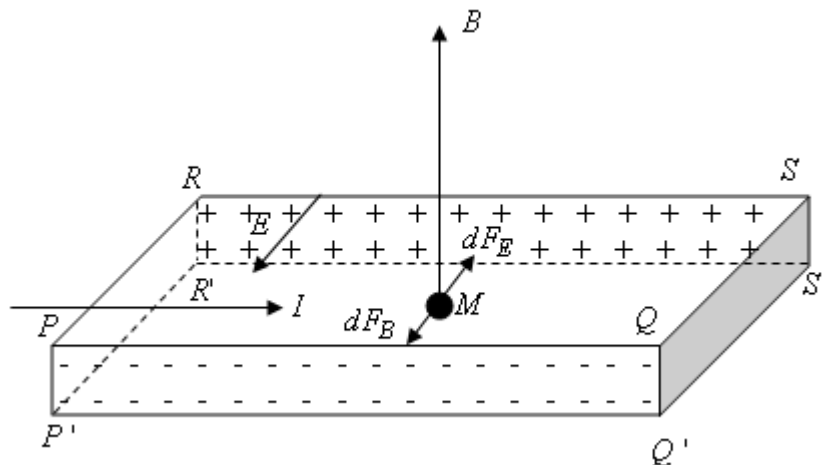


Figure: 36.11 - Ruban métallique parcouru par un courant continu

où un ruban métallique est parcouru par un courant continu $\vec{I}$. Le vecteur densité de courant $\vec{j}$ est constant et parallèle aux grands côtés PQ ou RS du ruban.

Imaginons alors que le ruban soit plongé dans un champ magnétique uniforme perpendiculaire aux plans PQ et RS (selon l'axe Z). Les charges mobiles de densité volumique ρ contenues dans un élément de volume dV sont donc soumises à la force magnétique:

$$d\vec{F} = \rho \cdot dV \cdot \vec{v} \times \vec{B} = \vec{J} \times \vec{B} \cdot dV \quad (36.131)$$

Cette force modifie les trajectoires des électrons mobiles et, au cours d'un régime transitoire, provoque leur accumulation sur le bord avant du ruban tandis qu'un excès de charges positives apparaît sur le bord arrière.

Ce phénomène produit un champ électrique supplémentaire parallèle à RP qui exerce sur les charges mobiles du volume dV une force électrique:

$$d\vec{F} = \rho \cdot dV \cdot \vec{E} \quad (36.132)$$

Les deux forces s'opposent donc l'une à l'autre et la force coulombienne tend à ramener les trajectoires électroniques dans leur position initiale. Un régime permanent s'établit peu à peu.

Remarque: En fait, à chaque fois que nous parlons de régime permanent en physique, nous mentons un peu. Il s'agit au fait juste d'un équilibre stable et en général, le système oscille autour de sa position d'équilibre. Au bout d'un certain temps, un système comme le conducteur impliqué dans notre exemple montre des oscillations négligeables. La physique c'est aussi parfois qu'une question d'approximations...

Quand ce régime est atteint, la densité de courant est à nouveau parallèle à PQ et les forces électriques et magnétiques ci-dessus sont vectoriellement opposées. Nous avons donc:

$$\vec{E} = -K \cdot (\vec{J} \times \vec{B}) \quad (36.133)$$

avec:

$$K = \rho^{-1} \quad (36.134)$$

Dans certains ouvrages, le produit vectoriel est explicité sous forme de ses composantes tel que:

$$E_y = \frac{J_x}{\rho} \cdot B_z \quad (36.135)$$

car les autres composantes sont nulles (la densité de courant est parallèle au ruban et le champ magnétique perpendiculaire).

Or, comme nous l'avons démontré dans le chapitre d'Électrocinétique:

$$J_x / \rho = v_x \quad (36.136)$$

dès lors:

$$E_y = v_x \cdot B_z \quad (36.137)$$

Nous définissons alors le "coefficient de Hall" par:

$$R_H = \frac{E_y}{B_z J_x} \quad (36.138)$$

R_H peut être aussi bien utilisé à l'équilibre pour la mesure de J_x que par extension si nous supposons $v_x = c^{te}$ alors $R_H \propto \rho^{-1}$ donc à la mesure de la densité de porteurs dans l'échantillon.

Remarque: Nous parlons également de "**résistance de Hall**". Il s'agit simplement du rapport de la tension de Hall sur le courant circulant dans l'échantillon. Il ne faut cependant pas confondre la résistance de Hall avec R_H . Notons que la résistance de Hall varie linéairement avec le champ magnétique.

Dans un semi-conducteur à deux dimensions, l'effet Hall est également mesurable. Par contre, à suffisamment basse température, nous observons une série de plateaux pour la résistance Hall en fonction du champ magnétique. Ces plateaux apparaissent à des valeurs précises de résistance, et ce, indépendamment de l'échantillon utilisé. Ceci fait l'objet de "**l'effet Hall quantique**" que nous n'étudierons pas dans ce chapitre.

Sous forme scalaire la relation de "**l'effet Hall**", s'écrit:

$$E = -K \cdot J \cdot B \cdot \sin \alpha \quad (36.139)$$

Nous pouvons aussi l'exprimer en explicitant la différence de potentiel qui correspond par définition au champ électrique.

Si l est la largeur du ruban, nous avons:

$$U = E \cdot l \quad (36.140)$$

Si e est son épaisseur, le courant I qui le parcourt est:

$$I = e \cdot l \cdot J \quad (36.141)$$

Compte tenu des positions relatives des divers vecteurs, la relation exprimant l'effet Hall équivaut donc à:

$$\frac{U}{l} = -K \cdot \frac{I}{e \cdot l} \cdot B \cdot \sin \alpha \quad (36.142)$$

Plus esthétiquement et sous une forme traditionnelle, la tension de l'effet Hall est donnée par:

$$U_H = k_H \cdot I \cdot B \cdot \sin \alpha \quad (36.143)$$

avec:

$$k_H = -\frac{1}{\rho \cdot e} \quad (36.144)$$

qui est la "**constante de Hall**". Elle est inversement proportionnelle à la densité des porteurs libres et dans le cadre des métaux, elle est négative.

Dans d'autres domaines d'étude comme celui des semi-conducteurs, nous écrivons la tension de Hall sous la forme traditionnelle:

$$U_H = -\frac{1}{\rho \cdot e} \cdot I \cdot B \cdot \sin \alpha = -\frac{1}{q \cdot n} \frac{I \cdot B}{e} \cdot \sin \alpha = k_H \frac{I \cdot B}{e} \cdot \sin \alpha \quad (36.145)$$

où q est la charge de l'électron et n la notation traditionnelle (sic!) de la densité de porteurs dans le cadre de l'étude des semi-conducteurs.

Nous avons alors dans ce dernier domaine la constante de Hall qui est définie par:

$$k_H = -\frac{1}{q \cdot n} \quad (36.146)$$

Ce qui a fait cependant la renommée de l'effet Hall, outre le fait que ce résultat est énormément utilisé pour fabriquer des sondes de champs magnétiques de tous genres (car les sondes à effet Hall fonctionnent sans contact physique avec les aimants), c'est que pour certains types de semi-conducteurs cette constante de Hall est positive!!!! Ce qui signifierait avec les modèles standards que nous avons à notre disposition jusqu'à maintenant, qu'il y aurait des charges positives qui feraient office de courant... et à l'époque de la mise en place de cette expérience pour les semi-conducteurs, ceci était inexplicable. À l'époque de Hall cette expérience servait à vérifier si c'était des charges positives ou négatives qui se déplaçaient et Hall conclua en testant cela sur des métaux conducteurs que seulement l'électricité négative circule dans les fils conducteurs.

Remarque: La sonde à effet Hall la plus connue par la population est le compteur de vitesse sur les vélos (odmètres) qui fonctionne sur la base de l'accrochage d'un petit aimant sur un des rayons d'une des roues et dont le passage devant la sonde à effet Hall produit un signal traité par l'électronique de l'odomètre.

Or nous verrons plus tard qu'en utilisant la théorie quantique dans le cadre des semi-conducteurs (cf. chapitre d'Électrocinétique) des charges positives peuvent pourtant sous certaines conditions apparaître et être à l'origine d'un courant!

4.2. RAYON DE LARMOR

Un cas très intéressant d'étude de laboratoire est le mouvement d'une charge dans un champ magnétique uniforme. Pour cette étude, considérons une particule de masse m et de charge q placée dans un champ magnétique uniforme avec une vitesse initiale $\vec{v}_0$.

Nous avons selon la loi de Lorentz:

$$\frac{d\vec{v}}{dt} = \frac{q}{m} \vec{v} \times \vec{B} \quad (36.147)$$

Nous allons tirer parti du fait que la force magnétique est nulle dans la direction du champ magnétique.

Nous allons donc décomposer la vitesse en deux composantes, l'une parallèle et l'autre perpendiculaire au champ magnétique tel que:

$$\vec{v} = \vec{v}_g + \vec{v}_\perp \quad (36.148)$$

L'équation du mouvement s'écrit alors:

$$\frac{d\vec{v}_{tg}}{dt} = \vec{0} \quad \frac{d\vec{v}_x}{dt} = \frac{q}{m} \vec{v}_x \times \vec{B} \quad (36.149)$$

La trajectoire reste donc rectiligne uniforme dans la direction du champ magnétique! En d'autres termes, si la vitesse de la particule chargée était nulle initialement dans la direction du champ alors elle restera nulle!

Prenons maintenant un repère cartésien dont l'axe Z est donné par la direction du champ magnétique tel que $dv_z/dt = 0$. L'équation du mouvement ne s'écrit dès lors plus que sur deux composantes puisque:

$$\begin{pmatrix} \frac{dv_x}{dt} \\ \frac{dv_y}{dt} \\ \frac{dv_z}{dt} \end{pmatrix} = \begin{pmatrix} \frac{dv_x}{dt} \\ \frac{dv_y}{dt} \\ 0 \end{pmatrix} = \frac{q}{m} \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} \times \begin{pmatrix} B_x \\ B_y \\ B_z \end{pmatrix} = \frac{q}{m} \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} \times \begin{pmatrix} 0 \\ 0 \\ B \end{pmatrix} = \frac{q}{m} \begin{pmatrix} v_y B \\ -v_x B \\ 0 \end{pmatrix} \quad (36.150)$$

d'où:

$$\frac{dv_x}{dt} = \frac{qB}{m} v_y = \omega_g v_y \quad \frac{dv_y}{dt} = -\frac{qB}{m} v_x = -\omega_g v_x \quad (36.151)$$

Une solution très simple à ces deux équations différentielles est dans un cadre non relativiste:

$$v_x = v_0 \cos(\omega t) \quad v_y = -v_0 \sin(\omega t) \quad (36.152)$$

où nous avons donc choisi une vitesse initiale suivant X. En intégrant, nous obtenons:

$$x = \frac{v_0}{\omega} \sin(\omega t) \quad y = \frac{v_0}{\omega} \cos(\omega t) \quad (36.153)$$

où les constantes d'intégration ont été choisies nulles (choix arbitraire). La trajectoire est donc un cercle de rayon:

$$R_L = \frac{mv_0}{qB} \quad (36.154)$$

perpendiculaire au champ magnétique et appelé "**rayon de Larmor**", décrit avec la pulsation:

$$\omega_g = \frac{qB}{m} \quad (36.155)$$

dite "**pulsation gyro-synchrotron**". Ce cercle est parcouru dans le sens conventionnel positif pour des charges négatives.

Remarque: Le mouvement n'est circulaire que si l'a particule, au départ, n'a donc pas de vitesse dans la direction du champ magnétique. Si elle en a une, elle la garde (le champ magnétique n'a pas d'action dans cette direction).

Le problème d'une telle configuration pour construire un accélérateur, c'est que si nous augmentons l'énergie de la particule (en ajoutant un champ électrique synchronisé sur la pulsation gyro-synchrotron et colinéaire au mouvement), sa vitesse augmente mais le rayon de Larmor aussi. Or, le "**cyclotron**" qui est basé sur ce système a un rayon limité puisqu'il est difficile de maintenir un champ magnétique constant sur une grande surface.

Plus difficile encore, dans le cas relativiste, la pulsation s'écrit alors avec le facteur de Fitzgerald-Lorentz (cf. chapitre de Relativité Restreinte):

$$\omega_{gs} = \frac{qB}{\gamma m} \quad (36.156)$$

Nous voyons alors qu'il faut ajuster la pulsation du champ électrique à la pulsation de rotation lorsque la vitesse augmente: l'accélérateur est maintenant un "**synchrocyclotron**".

Pour résoudre le problème de l'augmentation du rayon, nous utilisons alors un "**synchrotron**" constitué d'un tube à vide unique comportant de sections droites contenant des cavités accélératrices et des sections courbes équipées d'aimants créant à chaque instant le champ magnétique adapté à la vitesse des particules. Cette technique, dont il est facile de parler mais très difficile à mettre en pratique, est la plus utilisée de nos jours. Le LHC du CERN fait partie de la famille des synchrotrons

À partir de cette relation, il est inversement aisé d'avoir l'énergie cinétique de la particule:

$$E_c = \frac{1}{2} m v_0^2 = \frac{1}{2} m \left(\frac{q}{m} B R_L \right)^2 = \frac{1}{2} \frac{q^2}{m} (B R_L)^2 \quad (36.157)$$

C'est sur la base de cette relation que fonctionnent les "**spectromètres de masse de Dempster**". C'est en utilisant cette technique que les chercheurs ont découvert dans les années 1920 que les atomes d'un même élément chimique n'ont pas nécessairement la même masse. Les différentes variétés d'atomes d'un même élément chimique, variétés qui diffèrent par leur masse, sont les isotopes (cf. chapitre de Physique Nucléaire).

Le rayon de Larmor correspond à la distance la plus grande que peut parcourir une particule dans la direction transverse avant d'être déviée de sa trajectoire. Cela correspond donc à une sorte de distance de piégeage. À moins de recevoir de l'énergie cinétique supplémentaire, une particule chargée est ainsi piégée dans un champ magnétique.

Il est intéressant de noter que plus l'énergie cinétique transverse d'une particule est élevée (grande masse ou grande vitesse transverse) et plus le rayon de Larmor est grand. Inversement, plus le champ magnétique est élevé et plus ce rayon est petit.

Nous reviendrons sur ces notions dans le chapitre d'Électrodynamique, ou après avoir étudié les équations de Maxwell, nous ferons quelques développements pour les accélérateurs de type Bêatron.

Remarque: Le confinement du plasma dans un tokamak est basé sur cette propriété qu'ont les particules chargées de décrire une trajectoire en hélice autour d'une ligne de champ magnétique. D'où l'intérêt d'utiliser un tore.

4.3. ÉNERGIE D'UN DIPOLE MAGNÉTIQUE

Grâce au fait que nous ayons maintenant les unités du champ magnétique et celles de la constante de perméabilité magnétique, nous allons pouvoir déterminer par l'analyse dimensionnelle et par l'intuition l'énergie total d'un dipôle magnétique statique (donc orienté!) ce qui va nous être très utile pour la théorie du paramagnétisme. Considérons pour cela un aimant rigide sous forme de cylindre de longueur L et de rayon négligeable pouvant être considéré comme un dipôle Nord/Sud (un "aimant droit") plongé dans un champ magnétique constant et homogène dans le plan perpendiculaire à l'axe de rotation du dipôle:

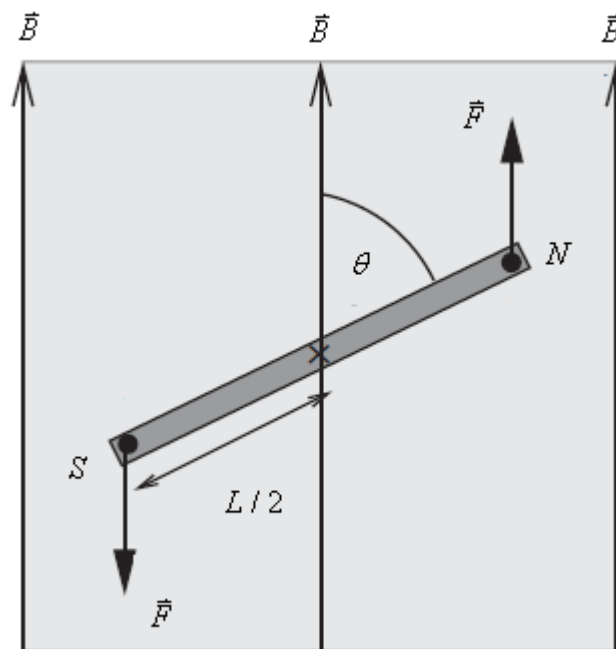


Figure: 36.12 - Dipôle simple

L'expérience montre que lorsque le dipôle est colinéaire avec le champ magnétique, celui-ci ne bouge plus. Il s'ensuit que la force sur une des extrémités dépend de façon proportionnelle au sinus de l'angle et du champ magnétique tel que:

$$\vec{F} \propto \vec{B} \sin \theta \quad (36.158)$$

Au niveau des unités, cela donne donc pour l'instant:

$$[kg \cdot m \cdot s^{-2}] = [kg \cdot s^{-2} \cdot A^{-1}] \quad (36.159)$$

Il nous faudrait donc nous débarrasser des ampères en faisant déjà au minimum intervenir ce qui caractérise un dipôle magnétique et que nous avons déjà déterminé plus: son moment magnétique μ ! Donc les unités sont pour rappel:

$$\left[A \cdot m^2 \right] \quad (36.160)$$

Il semble alors assez naturel d'écrire pour aller un peu plus loin:

$$\vec{F} \propto \mu_l \vec{B} \sin \theta \quad (36.161)$$

Ce qui donne maintenant au niveau des unités:

$$\left[kg \cdot m \cdot s^{-2} \right] = \left[kg \cdot s^{-2} \cdot m^2 \right] \quad (36.162)$$

Nous avons alors une unité de longueur en trop. Il semble alors assez naturel d'introduire la longueur du dipôle tel que (la force devant logiquement être égale en tout point du dipôle donc il n'y a aucune raison d'introduire ici la moitié de la longueur!):

$$\vec{F} = \mu_l \vec{B} \sin \theta = \frac{\mu_l}{L} \vec{B} \sin \theta \quad (36.163)$$

Maintenant pour en revenir à l'énergie du dipôle magnétique, nous considérons que celle-ci est nulle lorsque le dipôle est initialement dans une position perpendiculaire à celle du champ magnétique. En utilisant l'approximation habituelle comme quoi un déplacement infinitésimal d'une des deux extrémités est donné par:

$$dl = \frac{L}{2} d\theta \quad (36.164)$$

Ainsi, le travail (énergie) élémentaire pour faire tourner le dipôle sera (nous multiplions par deux car il faut sommer les deux forces qui agissent sur chacun des pôles) en notant B la norme du champ magnétique:

$$dE = 2 \left(\frac{\mu_l}{L} B \sin \theta \right) \frac{L}{2} d\theta = \mu_l B \sin \theta d\theta \quad (36.165)$$

où nous voyons que la longueur du dipôle n'intervient plus. En réalité il ne faut pas oublier que c'est ans le moment magnétique dipolaire μ_l que nous avons la surface équivalente du dipole.

Il vient alors par intégration pour un angle final donné:

$$E = \mu_l B \int_{\pi/2}^{\theta} \sin \theta d\theta = -\mu_l B \cos \theta \Big|_{\pi/2}^{\theta} = -\mu_l B \cos \theta \quad (36.166)$$

Soit:

$$E = -\vec{\mu}_l \cdot \vec{B} \quad (36.167)$$

Voyons maintenant un résultat classique et scolaire que nous pouvons obtenir de ce résultat et que nous retrouverons lors de notre étude du spin dans le chapitre de Physique Quantique Ondulatoire.

Nous avons donc démontré plus haut qu'une particule chargée est déviée par une force donnée par la relation de la loi de Lorentz:

$$\vec{F} = q\vec{v} \times \vec{B} \quad (36.168)$$

Il s'ensuit que si le champ a seulement une composante constante en Z et la vitesse une composante seulement en X, cela va provoquer un mouvement hélicoïdal dans le plan perpendiculaire au champ comme nous l'avons déjà démontré plus haut lors de notre étude du rayon de Larmor.

Considérons maintenant une particule chargée lancée à vitesse uniforme selon un axe X entre deux pôles d'aimants opposés qui génèrent un champ magnétique vertical hétérogène et intéressons-nous uniquement à la déflexion en Z de la trajectoire de la particule.

Du point de vue de l'axe Z la particule peut être considérée comme en mouvement rectiligne uniformément accéléré (cf. chapitre de Mécanique Classique):

$$z = z_0 + v_{0,z}t + \frac{1}{2}a_z t^2 \quad (36.169)$$

Puisque la position initiale en Z de la particule se situe à mi-distance entre les deux pôles et que sa vitesse initiale en Z est nulle:

$$\begin{aligned} z_0 &= 0 \\ v_{0,z} &= 0 \end{aligned} \quad (36.170)$$

Nous avons alors:

$$z = \frac{1}{2}a_z t^2 \quad (36.171)$$

Puisque le champ magnétique ne travaille pas et que cela implique que l'énergie cinétique reste constante, nous pouvons écrire que le temps est simplement le rapport de la distance parcourue par la particule sur le module de sa vitesse:

$$z = \frac{1}{2}a_z \left(\frac{L}{v}\right)^2 = \frac{1}{2}a_z L^2 \frac{1}{v^2} = \frac{1}{2}a_z L^2 \left(\frac{m}{2E_c}\right) = \frac{mL^2}{4E_c} a_z \quad (36.172)$$

Rappelons que nous venons de démontrer plus haut que l'énergie potentielle d'un dipôle magnétique était dans un champ constant et homogène:

$$E = -\vec{\mu}_l \cdot \vec{B} \quad (36.173)$$

Il vient alors dans notre cas que:

$$E = -\mu_{lz} B_z \quad (36.174)$$

Et comme nous pouvons associer une force à une énergie potentielle, il vient en supposant que le moment magnétique dipolaire reste constante selon Z et que le champ magnétique est quand même un peu inhomogène (vive le mélange des hypothèses... mais on fait de l'ingénierie physique ici!):

$$F_z = -\frac{\partial E}{\partial z} \cong \mu_{lz} \frac{\partial B_z}{\partial z} = m \cdot a_z \quad (36.175)$$

Et donc:

$$z = \frac{mL^2}{4E_c} a_z = \frac{mL^2}{4E_c} \left(\frac{\mu_{lz}}{m} \frac{\partial B_z}{\partial z} \right) = \frac{\mu_{lz} L^2}{4E_c} \frac{\partial B_z}{\partial z} \quad (36.176)$$

Donc nous voyons dans tout les cas que z peut prendre un spectre de valeur continu qui dépend du moment magnétique dipolaire de la particulaire. Or, comme nous le verrons dans le chapitre de Physique Quantique Ondulatoire lors de notre étude des opérateurs quantiques du moment cinétique orbital et de spin qu'une expérience dite de "Stern-Gerlach" a montré que tel n'était pas le cas pour des particules ou certains atomes pour lesquels la valeur z est clairement discrète, ce que la physique classique semble incapable d'expliquer.

5. MODÈLE DE LANGEVIN DU DIAMAGNÉTISME

Le but de ce modèle est de rendre compte d'un magnétisme négatif qui s'oppose donc à l'excitation magnétique. Ce modèle est grossier par rapport au modèle quantique mais il est intéressant pour deux raisons majeures: la première c'est que cela donne déjà de quoi débiter au lecteur qui n'est pas encore familier avec la théorie quantique, la deuxième étant que c'est un modèle formateur (dans le sens scolaire du terme) car il montre comment des bricolages et approximations successives peuvent amener à quelque chose de relativement acceptable sur le point de vue pratique.

Remarque: Cette théorie a bien sa place dans le chapitre de Magnétostatique car le champ d'excitation utilisé dans le modèle est supposé constant!

Pour cela, nous considérerons le modèle classique de Langevin (le modèle quantique donnant le même résultat) où l'électron est considéré comme parcourant une orbite circulaire r et est alors assimilable à un courant électrique dans une boucle produisant une force électromotrice (cf. chapitre d'Électrocinétique):

$$\mathcal{E} = -\frac{d\Phi}{dt} = -\frac{d(B \cdot S)}{dt} = -\frac{d(\mu_0 H S)}{dt} = -\mu_0 S \frac{dH}{dt} = -\mu_0 \pi r^2 \frac{dH}{dt} \quad (36.177)$$

que nous pouvons assimiler à un champ électromoteur tel que (cf. chapitre d'Électrocinétique):

$$e = -\frac{d\Phi}{dt} = -\frac{d(B \cdot S)}{dt} = -\frac{d(\mu_0 HS)}{dt} = -\mu_0 S \frac{dH}{dt} = -\mu_0 \pi r^2 \frac{dH}{dt} \quad (36.178)$$

Il vient alors:

$$E_m = -\frac{\mu_0 r}{2} \frac{dH}{dt} \quad (36.179)$$

Nous avons en notant m_e la masse au repos de l'électron et q_e sa charge électrique:

$$\frac{dv}{dt} = a = \frac{q_e E_m}{m_e} = -\frac{q_e \mu_0 r}{m_e} \frac{dH}{dt} \quad (36.180)$$

d'où:

$$\Delta v = -\frac{q_e \mu_0 r}{2m_e} \Delta H = -\frac{q_e \mu_0 r}{2m_e} \left(H - \underbrace{H_0}_{=0} \right) = -\frac{q_e \mu_0 r}{2m_e} H \quad (36.181)$$

L'application d'une excitation magnétique extérieure aura pour effet de changer le moment magnétique dipolaire μ_l d'une quantité $\Delta \mu_l$. Or, nous avons démontré plus haut que le moment magnétique dipolaire était donné par:

$$\mu_l = \frac{q \omega}{2} r^2 = \frac{q \frac{v}{r}}{2} r^2 = \frac{qr}{2} v \quad (36.182)$$

Il vient alors pour l'électron:

$$\Delta \mu_l = \frac{q_e r}{2} \Delta v = \frac{q_e r}{2} \left(-\frac{q_e \mu_0 r}{2m_e} H \right) = -\frac{q_e^2 \mu_0 r^2}{4m_e} H \quad (36.183)$$

Ensuite, l'idée très astucieuse et grossière (dans le sens approximatif du terme par rapport à l'expérience et au modèle quantique élaboré bien des années plus tard) est de prendre en considération le fait que l'électron sous forme classique peut être considéré comme un objet ponctuel pouvant se mouvoir dans toute une sphère de rayon R donné dans le cas d'un atome monoélectronique et non pas seulement dans un plan circulaire de rayon r perpendiculaire à la direction du champ d'excitation magnétique.

Dans ce cas de figure, nous avons alors bien évidemment:

$$x^2 + y^2 + z^2 = R^2 \quad (36.184)$$

et nous allons considérer que les trois coordonnées sont des variables aléatoires indépendantes et identiquement distribuées (donc déjà là le modèle théorique est mis à mal mais c'est déjà mieux que rien...). Dès lors, il vient que leur espérance est égale tel que:

$$E(x^2 + y^2 + z^2) = E(R^2) \quad (36.185)$$

En utilisant la propriété de linéarité de l'espérance (cf. chapitre de Statistiques) et la notation à la physicien, cela s'écrit alors:

$$\langle x^2 \rangle + \langle y^2 \rangle + \langle z^2 \rangle = \langle R^2 \rangle \quad (36.186)$$

et comme les coordonnées sont considérées comme des variables aléatoires identiquement distribuées, nous avons aussi:

$$\langle x^2 \rangle = \langle y^2 \rangle = \langle z^2 \rangle \quad (36.187)$$

Il s'ensuit immédiatement que:

$$\langle x^2 \rangle = \langle y^2 \rangle = \langle z^2 \rangle = \frac{\langle R^2 \rangle}{3} \quad (36.188)$$

Et donc si nous nous intéressons uniquement au rayon moyen du disque contenant toutes les orbites perpendiculaires à la direction du champ d'excitation magnétique dirigé selon l'axe Z, il vient alors:

$$\langle x^2 \rangle + \langle y^2 \rangle = \frac{2}{3} \langle R^2 \rangle \quad (36.189)$$

Donc au final pour un électron dans toutes les orbites possibles d'une sphère limitée:

$$\langle \Delta \mu_l \rangle = -\frac{q_e^2 \mu_0 \langle r^2 \rangle}{4m_e} H = -\frac{q_e^2 \mu_0 \frac{2}{3} \langle R^2 \rangle}{4m_e} H = -\frac{q_e^2 \mu_0 \langle R^2 \rangle}{6m_e} H \quad (36.190)$$

où $\langle R^2 \rangle$ peut être calculé explicitement avec le modèle quantique (des fonctions d'onde pour être plus précis!).

Pour un atome contenant Z électrons, nous ferons l'hypothèse grossière qu'une simple somme des effets est valable...:

$$\langle \Delta \mu_l \rangle = \sum_{i=1}^Z -\frac{q_e^2 \mu_0 \langle R_i^2 \rangle}{6m_e} H = -\frac{q_e^2 \mu_0 H}{6m_e} \sum_{i=1}^Z \langle R_i^2 \rangle \quad (36.191)$$

D'un point de vue macroscopique, le nombre d'atomes contenus dans une unité de volume sera le rapport de la masse volumique du matériau divisé par la masse atomique de l'élément considéré multiplié par le nombre d'Avogadro:

$$n = N_A \frac{\rho}{m_A} \quad (36.192)$$

Il vient alors par unité de volume:

$$\langle \Delta \mu_i \rangle = -n \frac{q_e^2 \mu_0 H}{6m_e} \sum_{i=1}^Z \langle R_i^2 \rangle = -H \frac{n q_e^2 \mu_0}{6m_e} \sum_{i=1}^Z \langle R_i^2 \rangle = -H \chi_d \quad (36.193)$$

et c'est ce résultat qui est assimilé à la susceptibilité magnétique tel qu'on l'écrit sous la forme de la "relation de la susceptibilité diamagnétique de Langevin":

$$\chi_d = -\frac{q_e^2 n \mu_0}{6m_e} \sum_{i=1}^Z \langle R_i^2 \rangle \quad (36.194)$$

Au niveau des unités, nous avons bien:

$$\begin{aligned} \chi_d &= \frac{[C^2][m^{-3}][kg \cdot m \cdot s^{-2} \cdot A^{-2}]}{[kg]} [m^2] = [C^2][s^{-2} \cdot A^{-2}] \\ &= [C^2][s^{-2} \cdot s^2 \cdot C^{-2}] = [-] \end{aligned} \quad (36.195)$$

C'est donc bien un coefficient sans unités.

et c'est parce que cette valeur est négative que l'on assimile ce modèle au diamagnétisme (par définition). Le lecteur remarquera que si l'excitation magnétique est nulle, la susceptibilité l'est aussi... ce qui est un minimum attendu du modèle théorique. Par contre celui ne dépend pas de la température (l'influence de celle-ci est de toute façon presque négligeable).

L'accord expérience/théorie est excellent pour les gaz nobles (à symétrie sphérique) de l'ordre de plus ou moins 10% d'erreur. Pour les éléments non sphériques l'erreur atteint souvent les 50% par rapport à la l'expérience.

6. MODÈLE DE LANGEVIN DU PARAMAGNÉTISME

Langevin tenta (avec plus ou moins de succès là aussi) d'expliquer le paramagnétisme avec les mêmes idées sous-jacentes mais toutefois en devant opter pour une approche mathématique totalement différente pour s'assurer d'un résultat final positif.... (bricolage quand tu nous tiens...). Ce que Langevin savait aussi c'est que le paramagnétisme dépendait fortement de la température selon les études expérimentales des matériaux ferromagnétiques, il fallait donc choisir une approche faisant ressortir la température et à l'époque il n'y avait pas 10'000 façons de faire cela! Il s'ensuit que ce modèle ouvre aussi la porte à la théorie du ferromagnétisme!

Comme point de départ à l'époque, on prenait naturellement la distribution de (Maxwell-)Boltzmann démontrée dans le chapitre de Mécanique Statistique qui décrit pour rappel la distribution des particules discernables qui n'interagissent pas avec aucune contrainte sur le nombre de particules par état... (à l'époque de Langevin il n'y avait que ce modèle à disposition...).

$$f_B(E, T) = \exp\left(-\frac{E}{kT}\right) \quad (36.196)$$

cette dernière écriture faisant, comme nous l'avons vu dans le chapitre de Mécanique Statistique, abstraction de la constante de normalisation pour en faire une vraie fonction de densité de probabilité (mais nous allons bien évidemment calculer cette constante un peu plus loin).

Nous avons aussi démontré plus que l'énergie potentielle magnétique d'un dipôle était donnée par:

$$E = -\vec{\mu}_l \circ \vec{B} = -\mu_l B \cos \theta \quad (36.197)$$

Maintenant, rappelons que nous avons démontré dans le chapitre de Formes Géométriques qu'un élément de surface d'une sphère était donnée par:

$$dS = R^2 \sin \theta d\theta d\phi \quad (36.198)$$

Il vient alors (le lecteur peut se référer au schéma du chapitre susmentionné) que pour une couronne de la sphère définie par deux plans parallèles dont le milieu contient l'origine de la sphère est alors donné par (au besoin nous pouvons refaire un schéma):

$$dS = R^2 \sin \theta \frac{d\theta}{2} 2\pi = R^2 \sin \theta d\theta \quad (36.199)$$

Pourquoi parlons-nous de cela? Eh bien parce qu'une partie du nombre total de dipôles magnétiques compris dans un intervalle d'angle $d\theta$ est comme nous le voyons ci-dessus proportionnel à un élément de surface puisque:

$$d\theta = \frac{dS}{R^2 \sin \theta} \quad (36.200)$$

Nous avons alors ce nombre qui est donné par (à un facteur constant inconnu près):

$$dn \propto KR^2 \sin \theta d\theta e^{-E/kT} = KR^2 \sin \theta d\theta e^{\mu_l B \cos \theta / kT} \quad (36.201)$$

La proportion correspondante (donc c'est aussi une probabilité) par rapport à l'ensemble des angles, pour un angle donné, est alors donnée après normalisation par:

$$p(\theta) = \frac{KR^2 \sin \theta e^{\mu_l B \cos \theta / kT}}{\int_0^\pi KR^2 \sin \theta e^{\mu_l B \cos \theta / kT} d\theta} = \frac{\sin \theta e^{\mu_l B \cos \theta / kT}}{\int_0^\pi \sin \theta e^{\mu_l B \cos \theta / kT} d\theta} \quad (36.202)$$

Nous avons vu lors de notre étude du moment dipolaire magnétique que $\vec{\mu}_i$ contribue donc au champ magnétique. Si nous avons une densité volumique de n dipôles magnétiques, nous aurons alors une contribution de l'ordre de $n\vec{\mu}_i$ s'ils sont tous orientés dans la même direction. Mais si nous projetons le vecteur du moment magnétique sur la direction du champ magnétique, la contribution des dipôles s'écrira alors:

$$n\mu_i \cos \theta \quad (36.203)$$

mais comme il y en a qui sont dans de nombreuses directions différentes et faisons des angles divers et variés avec le champ, il nous faut alors prendre la moyenne telle que la contribution au champ magnétique soit proportionnelle à:

$$n\mu_i \langle \cos \theta \rangle \quad (36.204)$$

Et en utilisant simplement les propriétés des fonctions de densité statistiques, l'espérance de la contribution est alors donnée par:

$$n\mu_i \langle \cos \theta \rangle = n\mu_i \int_0^\pi \cos \theta p(\theta) d\theta = n\mu_i \int_0^\pi \cos \theta \frac{\sin \theta e^{\mu_i B \cos \theta / kT}}{\int_0^\pi \sin \theta e^{\mu_i B \cos \theta / kT} d\theta} d\theta \quad (36.205)$$

et comme le dénominateur est juste une constante de normalisation, nous pouvons le sortir de la première intégrale:

$$n\mu_i \langle \cos \theta \rangle = n\mu_i \frac{\int_0^\pi \cos \theta \sin \theta e^{\mu_i B \cos \theta / kT} d\theta}{\int_0^\pi \sin \theta e^{\mu_i B \cos \theta / kT} d\theta} \quad (36.206)$$

Pour intégrer, faisons la petite simplification d'écriture:

$$\gamma = \frac{\mu_i B}{kT} \quad (36.207)$$

Cela nous donne:

$$n\mu_i \langle \cos \theta \rangle = n\mu_i \frac{\int_0^\pi \cos \theta \sin \theta e^{\gamma \cos \theta} d\theta}{\int_0^\pi \sin \theta e^{\gamma \cos \theta} d\theta} \quad (36.208)$$

et posons:

$$x = \cos \theta \quad (36.209)$$

Nous avons alors:

$$\frac{dx}{d\theta} = -\sin \theta \Rightarrow d\theta = -\frac{dx}{\sin \theta} \quad (36.210)$$

Il vient alors:

$$\mu_B \langle x \rangle = \frac{\int_{-1}^1 x e^{\gamma x} dx}{\int_{-1}^1 e^{\gamma x} dx} \quad (36.211)$$

La primitive au numérateur nous est connue car elle fait partie des primitives usuelles démontrées en détail dans le chapitre de Calcul Différentiel Et Intégral! L'intégrale au dénominateur est elle triviale:

$$\begin{aligned} \mu_B \langle x \rangle &= \mu_B \frac{e^{\gamma x} \left(\frac{1}{\gamma} x - \frac{1}{\gamma^2} \right) \Big|_{-1}^1}{\frac{1}{\gamma} e^{\gamma x} \Big|_{-1}^1} = \mu_B \frac{e^{-\gamma} \left(-1 - \frac{1}{\gamma} \right) - e^{\gamma} \left(1 - \frac{1}{\gamma} \right)}{e^{-\gamma} - e^{\gamma}} = \\ &= \mu_B \frac{e^{\gamma} \left(1 - \frac{1}{\gamma} \right) - e^{-\gamma} \left(-1 - \frac{1}{\gamma} \right)}{e^{\gamma} - e^{-\gamma}} = \mu_B \frac{e^{\gamma} + e^{-\gamma} - e^{\gamma} \frac{1}{\gamma} + e^{-\gamma} \frac{1}{\gamma}}{e^{\gamma} - e^{-\gamma}} \quad (36.212) \\ &= \mu_B \left[\frac{e^{\gamma} + e^{-\gamma}}{e^{\gamma} - e^{-\gamma}} - \frac{1}{\gamma} \frac{e^{\gamma} - e^{-\gamma}}{e^{\gamma} - e^{-\gamma}} \right] = \mu_B \left(\frac{e^{\gamma} + e^{-\gamma}}{e^{\gamma} - e^{-\gamma}} - \frac{1}{\gamma} \right) = \mu_B \left(\coth \gamma - \frac{1}{\gamma} \right) \end{aligned}$$

La fonction:

$$L(\gamma) = \coth \gamma - \frac{1}{\gamma} \quad (36.213)$$

est souvent appelée "**fonction de Langevin**" avec pour rappel:

$$\gamma = \frac{\mu_B B}{kT} \quad (36.214)$$

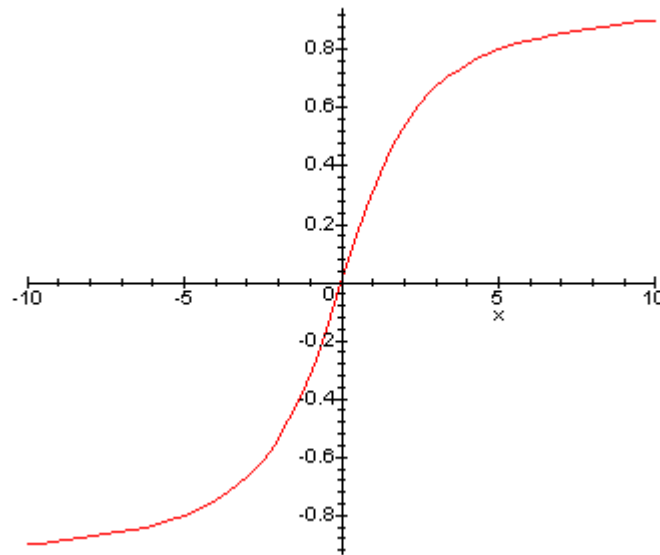


Figure: 36.13 - Tracé de la fonction de Langevin avec MapleV 4.00b

La fonction de Langevin vaut 0 quand son paramètre vaut 0 et tend vers 1 quand son paramètre tend vers l'infini. Donc le système finit par saturer quand le champ magnétique augmente, ce qui correspond bien au comportement expérimental des matériaux paramagnétiques. Par contre l'augmentation de la température fait diminuer fait tendre donc la fonction de Langevin vers 0 et a pour effet d'annuler l'alignement des dipôles.

Pour de petites valeurs du paramètre, la fonction peut être considérée comme linéaire comme nous le voyons sur la tracé ci-dessus.

Pour simplifier l'expression, nous allons en calculer l'approximation de Taylor de la cotangente hyperbolique en utilisant détaillée dans le chapitre de Suites Et Séries lorsque l'argument de cotangente hyperbolique est pour rappel strictement inférieur à 1 en valeur absolue:

$$\coth(\gamma) \cong \frac{1}{\gamma} + \frac{\gamma}{3} \quad (36.215)$$

Nous avons alors:

$$L(\gamma) \cong \frac{\gamma}{3} = \frac{\mu_l B}{3kT} \quad (36.216)$$

Nous avons donc le champ magnétique qui est proportionnel à:

$$n\mu_l \langle x \rangle = n\mu_l L(\gamma) \cong n\mu_l \frac{\mu_l B}{3kT} = \frac{n\mu_l^2 B}{3kT} = \frac{n\mu_l^2 \mu_0 H}{3kT} = \frac{n\mu_l^2 \mu_0}{3kT} H \quad (36.217)$$

Nous voyons que le facteur:

$$\frac{n\mu_l^2 \mu_0}{3kT} \quad (36.218)$$

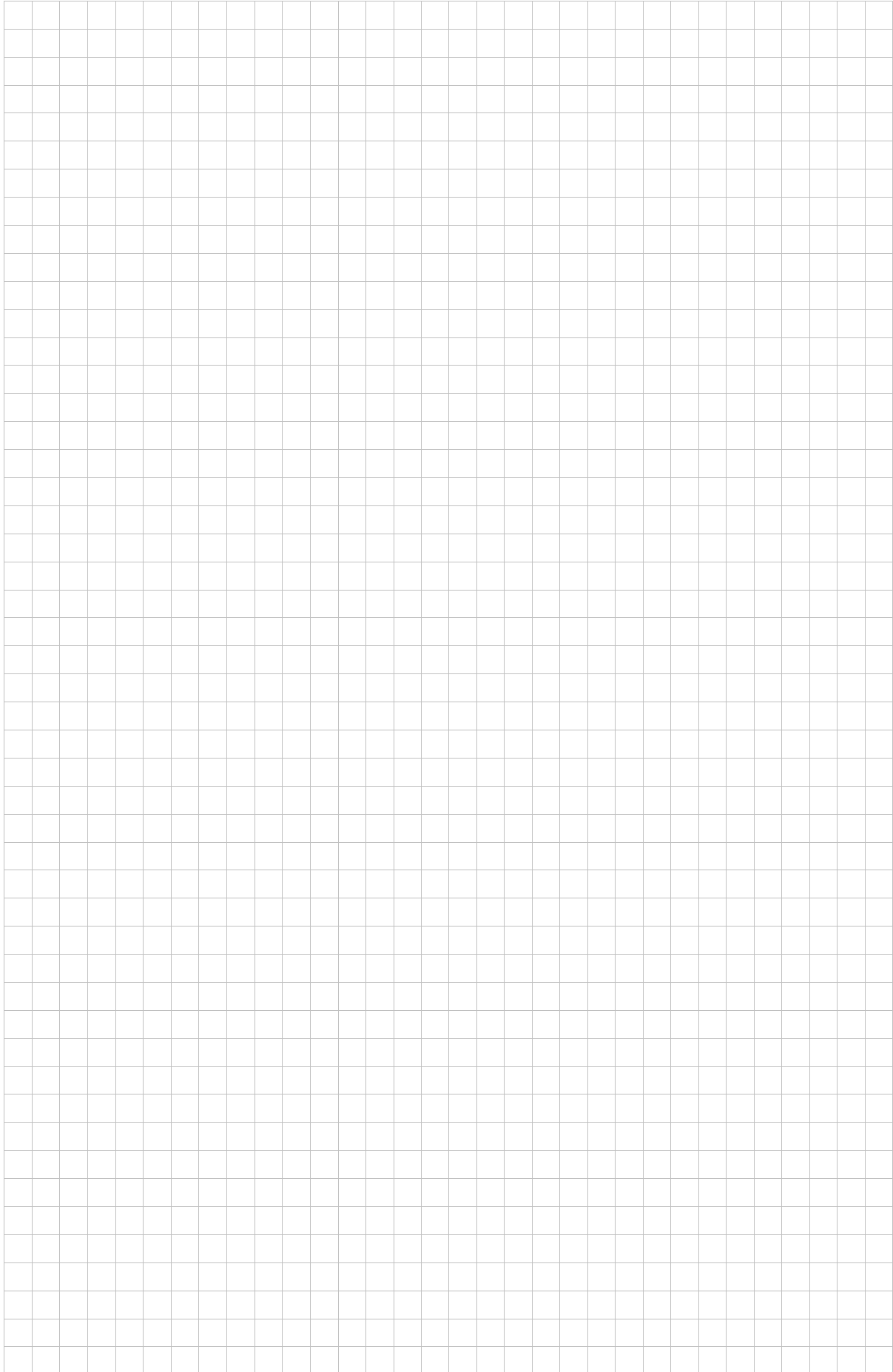
est sans dimensions. Effectivement:

$$\frac{[m^{-3}][A \cdot m^2]^2 [kg \cdot m \cdot s^{-2} \cdot A^{-2}]}{[m^2 \cdot kg \cdot s^{-2} \cdot K^{-1}][K]} = \frac{[m^{-3}][A^2 \cdot m^4][m \cdot A^{-2}]}{[m^2]} = \frac{[m^2]}{[m^2]} = [-] \quad (36.219)$$

Nous pouvons donc considérer qu'il s'agit de la susceptibilité paramagnétique et noter la "**relation de la susceptibilité paramagnétique de Langevin**"::

$$\chi_p \cong \frac{N \mu_1^2 \mu_0}{3kT} = \frac{N \mu_1^2 \mu_0}{3k} \frac{1}{T} = \frac{C}{T} \quad (36.220)$$

plus connue sous le nom de "loi de Curie" et qui montre que la susceptibilité magnétique est inversement proportionnelle à la température (mais bon évidemment cette loi devient fautive aux basses températures et il faut alors dériver empirique la loi de Curie-Weiss).







37. ÉLECTRODYNAMIQUE

Nous allons dans ce chapitre dégager un ensemble d'équations qui peuvent résumer à elles seules l'ensemble de nos connaissances sur l'électrostatique et la magnétostatique. Ces équations, au nombre de quatre, se nomment "**équations de Maxwell-Heaviside**" (dénomination que nous abrègerons par abus de langage comme de nombreux autres ouvrages "**équations de Maxwell**") et vont nous permettre d'aborder la branche de la physique appelée "**électrodynamique**" et donc des ondes électromagnétiques.

L'électrodynamique est un pilier de la révolution électronique! Sans cette théorie: pas de radio, pas de téléphones ou téléphones portables, pas d'ordinateurs, pas de satellites, pas d'électroménager, pas de moteur électrique, bref nous serions encore à l'état technologique de la fin du 19ème siècle.

Remarque: Il est très important de bien comprendre ce qui va suivre! Certains des développements seront réutilisés dans les chapitres de Relativité Restreinte, de Physique Quantique Des Champs, etc. Par ailleurs, il faudrait que le lecteur lise en parallèle le chapitre de Relativité Restreinte pour mieux comprendre les tenants et aboutissants de certains résultats et la provenance de quelques outils mathématiques.

Nous supposerons avant de nous attaquer aux modèles mathématiques que tout un chacun admet en ce début de 3ème millénaire que les rayons gamma, les ondes radio, les micro-ondes, la lumière visible (et non visible) sont des ondes électromagnétiques (E.M.) de fréquences différentes:

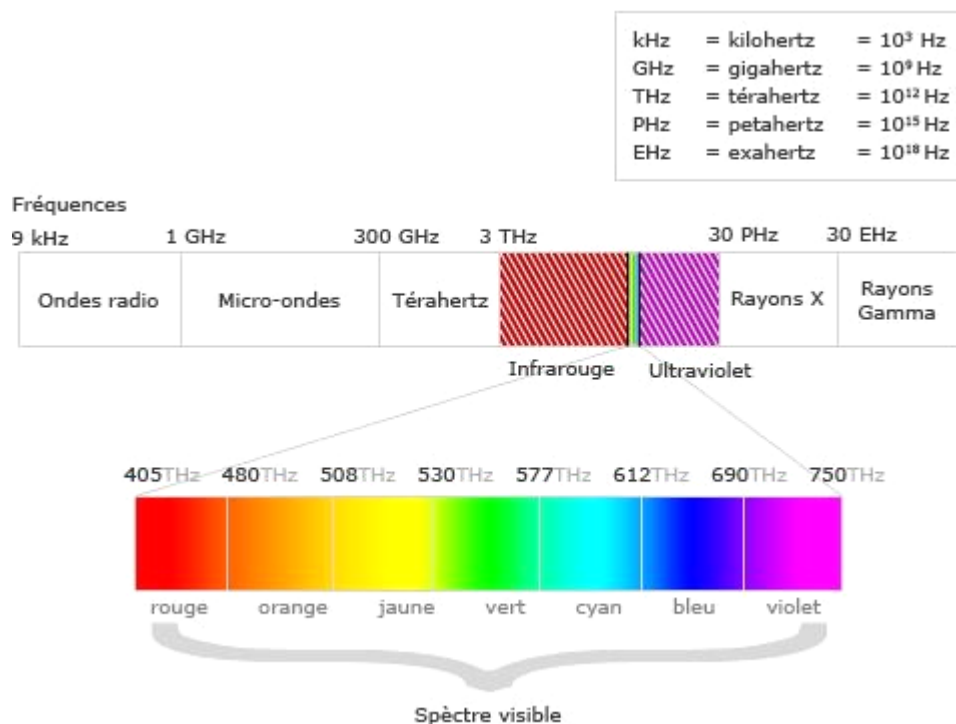


Figure: 37.1 - Spectre électromagnétique (source: Le Figaro.fr)

Avec un bon résumé des applications courantes des fréquences en ce début de 21ème siècle:

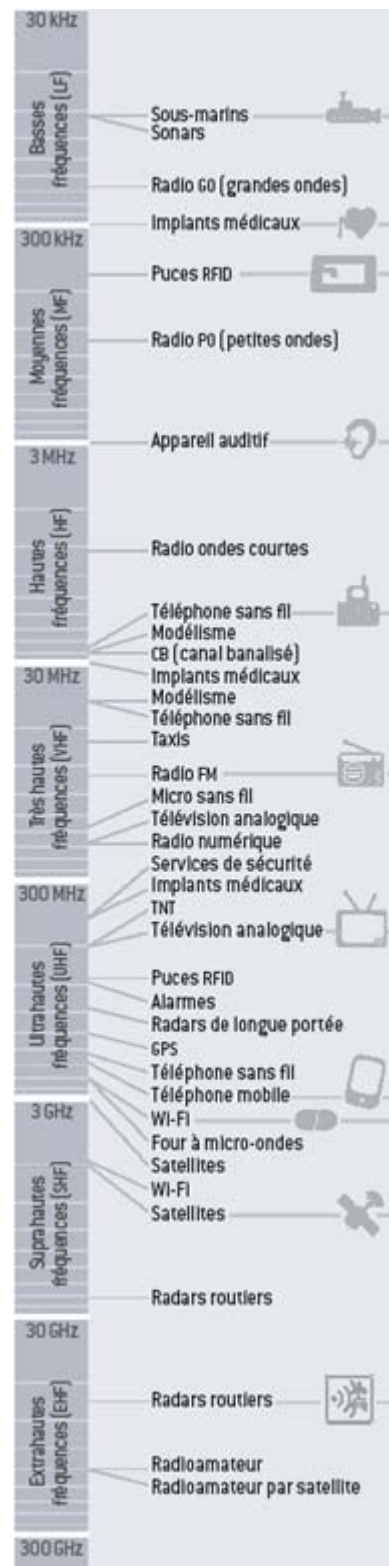


Figure: 37.2 - Applications économiques courantes (source: Pour la Science)

L'attribution des fréquences à l'industrie économique, civile et militaire étant le rôle de l'Union Internationale des Télécommunications.

1. PREMIÈRE ÉQUATION DE MAXWELL

Soit défini un champ de vecteurs $\vec{E}$ dans l'espace. Considérons une surface S fermée dans ce champ. Alors à chaque point (x, y, z) appartenant à la surface correspond un vecteur du champ.

Dans ce cas le théorème d'Ostrogradsky (cf. chapitre de Calcul Vectoriel) donne:

$$\Phi_{\mathcal{E}} = \oiint_S (\vec{\mathcal{E}} \circ \vec{n}) dS = \iiint_V (\vec{\nabla} \circ \vec{\mathcal{E}}) dV \quad (37.1)$$

avec V étant le volume délimité par la surface (dite pour rappel: "**surface de Gauss**") fermée.

Remarque: Le théorème d'Ostrogradsky est vérifié à condition qu'il n'existe pas de singularités dans le volume V .

Rappelons avant de continuer que dans le cas du théorème d'Ostrogradsky le vecteur $\vec{n}$ est conventionnellement dirigé vers l'extérieur de la surface.

Dans le cas particulier d'un champ électrique, nous obtenons des résultats très intéressants. En effet soit une charge Q repérée par rapport à un référentiel par le vecteur $\vec{r}_Q$.

Alors, nous avons vu dans le chapitre d'Électrostatique qu'en chaque point de l'espace, il existe un champ $\vec{E}$ tel que:

$$\vec{E}(\vec{r}) = \frac{K \cdot Q}{\|\vec{r} - \vec{r}_Q\|^3} \cdot (\vec{r} - \vec{r}_Q) \quad (37.2)$$

d'où:

$$\vec{E}(x, y, z) = \frac{K \cdot Q}{[(x - r_{Qx})^2 + (y - r_{Qy})^2 + (z - r_{Qz})^2]^{\frac{3}{2}}} \cdot \begin{pmatrix} x - r_{Qx} \\ y - r_{Qy} \\ z - r_{Qz} \end{pmatrix} \quad (37.3)$$

Comme nous pouvons le constater, le champ $\vec{E}$ possède une singularité en (r_{Qx}, r_{Qy}, r_{Qz}) . Considérons une surface de Gauss telle que la charge Q se trouve à l'extérieur de cette surface. À l'intérieur du volume V délimité par la surface S le champ $\vec{E}$ ne possède alors pas de singularité. Nous pouvons donc calculer la divergence de $\vec{E}$:

$$\begin{aligned} \frac{\partial E_x}{\partial x}(x, y, z) &= \frac{K \cdot Q}{[(x - r_{Qx})^2 + (y - r_{Qy})^2 + (z - r_{Qz})^2]^{\frac{3}{2}}} - \frac{3 \cdot K \cdot Q \cdot (x - r_{Qx})^2}{[(x - r_{Qx})^2 + (y - r_{Qy})^2 + (z - r_{Qz})^2]^{\frac{5}{2}}} \\ \frac{\partial E_y}{\partial y}(x, y, z) &= \frac{K \cdot Q}{[(x - r_{Qx})^2 + (y - r_{Qy})^2 + (z - r_{Qz})^2]^{\frac{3}{2}}} - \frac{3 \cdot K \cdot Q \cdot (y - r_{Qy})^2}{[(x - r_{Qx})^2 + (y - r_{Qy})^2 + (z - r_{Qz})^2]^{\frac{5}{2}}} \\ \frac{\partial E_z}{\partial z}(x, y, z) &= \frac{K \cdot Q}{[(x - r_{Qx})^2 + (y - r_{Qy})^2 + (z - r_{Qz})^2]^{\frac{3}{2}}} - \frac{3 \cdot K \cdot Q \cdot (z - r_{Qz})^2}{[(x - r_{Qx})^2 + (y - r_{Qy})^2 + (z - r_{Qz})^2]^{\frac{5}{2}}} \end{aligned}$$

Il vient alors:

$$\operatorname{div}(\vec{E}) = \vec{\nabla} \circ \vec{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = \frac{(3 \cdot K \cdot Q - 3 \cdot K \cdot Q) \cdot [(x-r_{\mathcal{D}x})^2 + (y-r_{\mathcal{D}y})^2 + (z-r_{\mathcal{D}z})^2]}{[(x-r_{\mathcal{D}x})^2 + (y-r_{\mathcal{D}y})^2 + (z-r_{\mathcal{D}z})^2]^{\frac{5}{2}}} = 0 \quad (37.4)$$

Donc si nous calculons le flux à travers cette surface nous trouvons (voir le chapitre de Calcul Vectoriel pour la description détaillée de l'opérateur del représenté par le symbole Nabla):

$$\Phi_{\mathcal{E}} = \iiint_V (\vec{\nabla} \circ \vec{E}) \cdot dV = 0 \quad (37.5)$$

Le flux est nul!

Dans le cas où la charge Q se trouve à l'intérieur de la surface de Gauss $\vec{\nabla} \circ \vec{E}$ n'est plus définie en $(r_{\mathcal{D}x}, r_{\mathcal{D}y}, r_{\mathcal{D}z})$ nous avons alors:

$$\Phi_{\mathcal{E}} = \iiint_{V-B} (\vec{\nabla} \circ \vec{E}) \cdot dV + \Phi_B \quad (37.6)$$

Avec Φ_B étant le flux de $\vec{E}$ sur une petite boule B entourant la charge ponctuelle Q .

Dans ce cas:

$$\Phi_{\mathcal{E}} = \iiint_{V-B} (\vec{\nabla} \circ \vec{E}) \cdot dV = 0 \quad (37.7)$$

car la divergence est définie partout sur $V-B$. Il nous reste donc:

$$\Phi_{\mathcal{E}} = \Phi_B = \oiint_S \vec{E} \circ d\vec{S} \quad (37.8)$$

Mais dans le cas d'une sphère, il est relativement facile de calculer:

$$\oiint_S \vec{E} \circ d\vec{S} \quad (37.9)$$

Nous avons:

$$\Phi_{\mathcal{E}} = \oiint_S \vec{E} \circ d\vec{S} = \oiint_S E \cdot d\vec{S} = E \cdot \oiint_S d\vec{S} = E \cdot S = 4 \cdot \pi \cdot r^2 \cdot \left(\frac{1}{4\pi\epsilon_0} \frac{Q}{r^2} \right) = \frac{Q}{\epsilon_0} \quad (37.10)$$

d'où la "**première équation de Maxwell**" ou "**loi de Gauss**" pour le champ électrique (ou "théorème de Gauss") avec une notation un peu condensée:

$$\boxed{\Phi_{\mathcal{E}} = \oiint_S \vec{E} \circ d\vec{S} = \int_V (\vec{\nabla} \circ \vec{E}) dV = Q \cdot \epsilon_0^{-1} \Leftrightarrow \vec{\nabla} \circ \vec{E} = \rho_{\mathcal{D}} \cdot \epsilon_0^{-1}} \quad (37.11)$$

où $\rho_{\mathcal{D}}$ est la densité de chage exprimée en coulombs par m^3 . À gauche nous avons donc la forme intégrale de l'équation de Maxwell et à droite sa forme différentielle.

Cette équation suggère donc que le flux du champ électrique traversant une surface close (d'où le cercle sur l'intégrale) est égal, à un facteur dimensionnel près, à la charge totale enfermée dans cette surface.

Remarque: L'intégrale de la dernière relation est une intégrale curviligne (donc évaluée sur une courbe). Dans le domaine de l'électrodynamique les intégrales curvilignes s'appliquent très souvent sur des chemins ou surfaces fermées d'où l'indication d'un cercle superposé au symbole de l'intégrale portant alors le nom de "**circulation du champ de vecteurs**".

Si nous exprimons maintenant cette équation en fonction du potentiel électrique pour lequel nous avons démontré dans le chapitre d'Électrostatique que:

$$\vec{E} = -\vec{\nabla} \cdot U = -\overrightarrow{grad}(U) \quad (37.12)$$

nous obtenons:

$$\vec{\nabla} \circ (\vec{\nabla} \cdot U) = -\frac{\rho_E}{\epsilon_0} \quad (37.13)$$

Nous pouvons noter la relation ci-dessus de façon plus esthétique en utilisant le laplacien scalaire (cf. chapitre de Calcul Vectoriel), tel que nous obtenions la relation:

$$\Delta \cdot \varphi = -\frac{\rho_E}{\epsilon_0} \quad (37.14)$$

appelée "**équation de Maxwell-Poisson**".

2. DEUXIÈME ÉQUATION DE MAXWELL

Dans le cas particulier d'un champ magnétique, nous obtenons également des résultats très intéressants.

En effet soit un courant I repéré par rapport à un référentiel par le vecteur $\vec{r}_I$. Alors en chaque point $\vec{r}$ de l'espace, nous avons vu dans le chapitre de Magnétostatique qu'il existe un champ $\vec{B}$ tel que:

$$\vec{B}(\vec{r}) = \underbrace{\frac{\mu_0}{4\pi}}_{=K} \frac{I \cdot \vec{l} \times (\vec{r} - \vec{r}_I)}{\|\vec{r} - \vec{r}_I\|^3} \quad (37.15)$$

d'où:

$$\vec{B}(x,y,z) = \frac{K \cdot I}{[(x-r_E)^2 + (y-r_F)^2 + (z-r_E)^2]^{\frac{3}{2}}} \cdot \begin{pmatrix} l_y(z-r_E) - l_x(y-r_F) \\ l_x(x-r_E) - l_z(z-r_E) \\ l_x(y-r_F) - l_y(x-r_E) \end{pmatrix} \quad (37.16)$$

Comme nous pouvons le constater, le champ $\vec{B}$ possède une singularité en (r_E, r_F, r_E) . Considérons alors une surface de Gauss telle que le courant I se trouve à l'extérieur de cette surface.

À l'intérieur du volume V délimité par surface S le champ $\vec{B}$ ne possède alors pas de singularité. Nous pouvons donc calculer la divergence de $\vec{B}$ (cf. chapitre de Calcul Vectoriel):

$$\begin{aligned}\frac{\partial B_x}{\partial x}(x,y,z) &= - \frac{3 \cdot K \cdot I \cdot (x-r_{\vec{B}}) \cdot (l_y(z-r_{\vec{B}}) - l_z(y-r_{\vec{B}}))}{[(x-r_{\vec{B}})^2 + (y-r_{\vec{B}})^2 + (z-r_{\vec{B}})^2]^{\frac{5}{2}}} \\ \frac{\partial B_y}{\partial y}(x,y,z) &= - \frac{3 \cdot K \cdot I \cdot (y-r_{\vec{B}}) \cdot (l_z(x-r_{\vec{B}}) - l_x(z-r_{\vec{B}}))}{[(x-r_{\vec{B}})^2 + (y-r_{\vec{B}})^2 + (z-r_{\vec{B}})^2]^{\frac{5}{2}}} \\ \frac{\partial B_z}{\partial z}(x,y,z) &= - \frac{3 \cdot K \cdot I \cdot (z-r_{\vec{B}}) \cdot (l_x(y-r_{\vec{B}}) - l_y(x-r_{\vec{B}}))}{[(x-r_{\vec{B}})^2 + (y-r_{\vec{B}})^2 + (z-r_{\vec{B}})^2]^{\frac{5}{2}}}\end{aligned}\quad (37.17)$$

D'où:

$$\text{div}(\vec{B}) = \vec{\nabla} \circ \vec{B} = \frac{\partial B_x}{\partial x} + \frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z} = 0 \quad (37.18)$$

Si nous calculons le flux à travers cette surface, nous trouvons alors:

$$\Phi_B = \iiint_V (\vec{\nabla} \circ \vec{B}) \cdot dV = 0 \quad (37.19)$$

Le flux est nul!

Dans le cas où le courant I se trouve à l'intérieur de la surface de Gauss $\vec{\nabla} \circ \vec{B}$ n'est plus défini en $(r_{\vec{B}}, r_{\vec{B}}, r_{\vec{B}})$ nous avons alors:

$$\Phi_B = \iiint_{V-B'} (\vec{\nabla} \circ \vec{B}) \cdot dV + \Phi_{B'} \quad (37.20)$$

Avec $\Phi_{B'}$ étant le flux de $\vec{B}$ sur une petite boule B' entourant partiellement le conducteur rectiligne transportant le courant I . Dans ce cas:

$$\iiint_{V-B'} (\vec{\nabla} \circ \vec{B}) \cdot dV = 0 \quad (37.21)$$

car la divergence est définie partout sur $V-B'$. Il nous reste donc:

$$\Phi_B = \Phi_{B'} = \oiint_S \vec{B} \circ d\vec{S} \quad (37.22)$$

Mais dans le cas d'une sphère, il est facile de calculer:

$$\oiint_S \vec{B} \circ d\vec{S} \quad (37.23)$$

Nous avons alors la loi de Gauss pour le champ magnétique:

$$\Phi_B = \oiint_S \vec{B} \circ d\vec{S} = \oiint_S B \cdot d\vec{S} \cdot \cos(\varphi) = 0 \quad (37.24)$$

En effet, dans le cas du champ magnétique, $d\vec{S}$ et $\vec{B}$ sont perpendiculaires donc:

$$\varphi = \frac{\pi}{2} \Rightarrow \cos(\varphi) = 0 \quad (37.25)$$

Remarque: D'où nous pouvons aussi déduire que $\vec{B} \perp \vec{E}$!

Donc, soit donnée une surface de Gauss dans un champ magnétique, alors le flux du champ magnétique à travers cette surface vaut:

$$\Phi_B = \oiint_S \vec{B} \circ d\vec{S} = \iiint_V (\vec{\nabla} \circ \vec{B}) dV = 0 \Leftrightarrow \vec{\nabla} \circ \vec{B} = 0 \quad (37.26)$$

relation qui constitue la "**deuxième équation de Maxwell**". À gauche nous avons donc la forme intégrale de l'équation de Maxwell et à droite sa forme différentielle.

Cette deuxième équation revient donc à dire qu'il n'existe aucun "**monopôle magnétique**" dans la nature, c'est-à-dire, qu'à tout pôle positif, nous devons retrouver un pôle négatif (à partir d'un aimant, les lignes du champ ne divergent pas). La deuxième équation vient toutefois rajouter l'idée (démontrée par Dirac) que s'il était possible de retrouver un monopôle dans la nature, il serait le point de source du champ magnétique. Nous verrons cela un peu plus loin dans le détail.

3. TROISIÈME ÉQUATION DE MAXWELL

Nous démontrerons dans le chapitre d'Électrocinétique (car il faut des notions que nous n'avons pas encore rencontrées), que la variation du flux du champ magnétique dans le temps à travers une boucle conductrice induit une tension dans cette boucle donnée par la "**loi de Faraday**" ou "**loi de Lenz-Faraday**":

$$U(t) = - \frac{d\Phi_B}{dt} \quad (37.27)$$

et nous avons déjà démontré dans le chapitre d'Électrostatique que:

$$\Delta U = - \int \vec{E} \circ d\vec{r} = - \int E dr \quad (37.28)$$

où la dernière égalité n'est valable que dans le cas particulier si le chemin parcouru est colinéaire au champ électrique.

Remarque: Nous verrons dans le chapitre d'Électrocinétique qu'il n'est pas tout à fait correct de noter le potentiel U comme ci-dessus car au fait, la loi de Faraday exprime la force électromotrice (potentiel électromoteur) e et ce potentiel est non conservatif contrairement au potentiel électrostatique de Coulomb (pour lequel l'intégrale sur un chemin fermé est nulle comme nous l'avons démontré dans le chapitre d'Électrostatique).

Pour un élément $d\vec{l}$ d'un circuit Γ il vient:

$$U(t) = \int_{\Gamma} \vec{E} \circ d\vec{l} \quad (37.29)$$

le changement de signe étant ici justifié par la loi de Lenz, selon laquelle le courant induit (et le flux magnétique qui lui est associé) a une orientation telle qu'il s'oppose à la variation de flux à travers le circuit.

Si nous développons cette relation, en utilisant le théorème de Stokes (cf. chapitre de Calcul Vectoriel) qui est pour rappel:

$$\oint_{\Gamma} \vec{f} \circ d\vec{r} = \int_S (\vec{\nabla} \times \vec{f}) \circ d\vec{S} \quad (37.30)$$

Nous avons alors:

$$U(t) = -\frac{d\Phi}{dt} = \oint_{\Gamma} \vec{E} \circ d\vec{r} = \int_S (\vec{\nabla} \times \vec{E}) \circ d\vec{S} \quad (37.31)$$

Où, comme nous le verrons dans le chapitre d'Électrocinétique, le champ électrique ci-dessus n'est pas le simple champ coulombien mais la somme d'un champ coulombien et d'un champ électromoteur (généré implicitement par la force de Biot-Savart).

Nous avons alors:

$$-\frac{d\Phi}{dt} = -\frac{d \int_S \vec{E} \circ d\vec{S}}{dt} = \int_S (\vec{\nabla} \times \vec{E}) \circ d\vec{S} \quad (37.32)$$

Et si l'élément de surface ne bouge pas dans l'espace et que seul le champ magnétique varie dans le temps, nous avons alors:

$$-\frac{d}{dt} \int_S \vec{E} \circ d\vec{S} = -\int_S \frac{\partial \vec{E}}{\partial t} \circ d\vec{S} = \int_S (\vec{\nabla} \times \vec{E}) \circ d\vec{S} \quad (37.33)$$

Soit:

$$\int_S \left(\vec{\nabla} \times \vec{E} + \frac{\partial \vec{B}}{\partial t} \right) \circ d\vec{S} = 0 \quad (37.34)$$

Une solution triviale est alors de dire que:

$$\vec{\nabla} \times \vec{E} + \frac{\partial \vec{B}}{\partial t} = \vec{0} \quad (37.35)$$

Nous obtenons alors au final:

$$U(t) = -\frac{d\Phi_B}{dt} = \oint_{\Gamma} \vec{E} \circ d\vec{l} \Leftrightarrow \vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \quad (37.36)$$

Ceci est la "**troisième équation de Maxwell**" ou "**loi de Maxwell-Faraday**" dite parfois encore "**loi d'induction**". À gauche nous avons donc la forme intégrale de l'équation de Maxwell et à droite sa forme différentielle.

La troisième équation affirme donc qu'une variation du champ magnétique produit un champ électrique dans une boucle conductrice. Nous disons alors que le terme avec la dérivée partielle du champ magnétique est le "**terme de couplage magnétique**". Cette équation est donc basée sur la théorie de Faraday.

Souvent dans la littérature scientifique, le potentiel $U(t)$ peut être simplement noté par un u minuscule.

La loi de Faraday de l'induction est typiquement utilisée par de petits appareils portatifs comme le PEG ci-dessous (Personal Energy Generator) pour recharger des appareils électroniques portables:



Figure: 37.3 - Photo d'un PEG (droite) avec un mobile

3.1. BÊTATRON

Parmi les nombreux exemples d'application de la troisième loi que nous verrons dans d'autres chapitres du site, il y en est un particulièrement sympathique car il fait penser à la physique moderne à grande échelle (même si dans la réalité on en est très loin).

À l'aide des équations de Maxwell et des relations démontrées dans le chapitre de Magnétostatique, nous pouvons faire une petite étude théorique non exhaustive du principe physique à la base d'un des plus vieux accélérateurs de particules non-linéaires.

Une des premières méthodes non-linéaires qui vient à l'esprit consiste à accélérer une particule chargée via une induction magnétique. Ce type d'accélérateur est appelé un "**bêtatron**" (dans l'idée qu'il accélère les électrons aussi vite que lors de la radioactivité bêta...) et a été conceptualisé dans les années 1930.

Le bêtatron est un accélérateur de particules qui consiste à injecter des électrons dans un tore sous vide (en blanc sur la photo ci-dessous) soumis à un champ magnétique qui sera considéré ici comme homogène entre les deux aimants (en rouges sur la photo ci-dessous) afin d'obtenir des rayonnements X ou gamma intenses utiles à certaines activités professionnelles (médecine, analyse de structures, etc.). Cet accélérateur est donc limité par l'intensité du champ magnétique qu'il peut produire ou supporter.

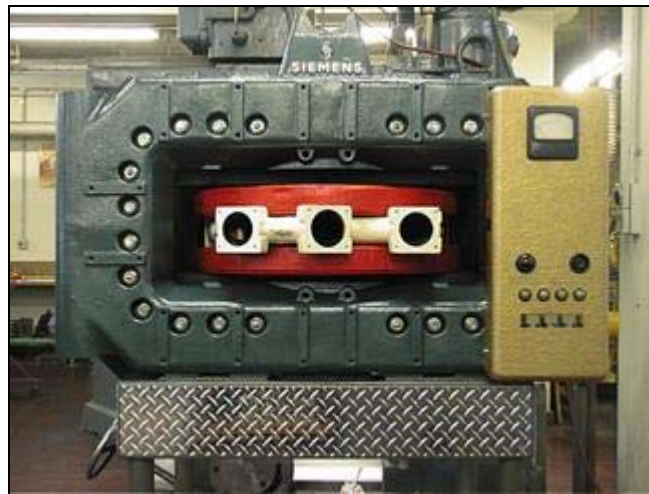


Figure: 37.4 - Photo d'un bêtatron

Pour cette étude théorique, nous allons d'abord utiliser le résultat démontré dans le chapitre de Magnétostatique lorsque nous avons abordé le rayon de Larmor: un électron en mouvement dans un champ magnétique aura une trajectoire circulaire qui sera perpendiculaire au champ magnétique.

Ensuite, nous allons aussi avoir besoin de la troisième équation de Maxwell sous forme d'intégrale:

$$U(t) = -\frac{d\Phi_B}{dt} = \int_{\Gamma} \vec{E} \circ d\vec{l} \Leftrightarrow \vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \quad (37.37)$$

qui - pour rappel - dit qu'une variation du champ magnétique produit un champ électrique dans une boucle conductrice (ou un mouvement de particules chargées qui peut être assimilé à une boucle conductrice!).

Nous avons:

$$\int_{\Gamma} \vec{E} \circ d\vec{l} = -\frac{d\Phi_B}{dt} = -\frac{d}{dt} \left(\oint_S \vec{B} \circ d\vec{S} \right) \quad (37.38)$$

et comme la trajectoire est donc circulaire dans le bêtatron comme nous l'avons démontré lors de notre étude du rayon de Larmor dans le chapitre de Magnétostatique, nous avons:

$$\int_{\Gamma} \vec{E} \circ d\vec{l} = \int_{\Gamma} \|\vec{E}\| \cos \alpha dl \quad (37.39)$$

Or comme le champ électrique est tangent à la trajectoire circulaire des électrons et que ces derniers vont dans le sens inverse de ce même champ (sens qui est donc constant en tout point de la trajectoire), nous avons puisque les électrons parcourent des cercles:

$$\int_{\Gamma} \vec{E} \circ d\vec{l} = \int_{\Gamma} \|\vec{E}\| \cos \alpha dl = \|\vec{E}\| \int_{\Gamma} \cos(\pi) dl = -\|\vec{E}\| \int_{\Gamma} dl = -2\pi R_L \|\vec{E}\| \quad (37.40)$$

Mais nous avons aussi:

$$-\frac{d}{dt} \left(\oint_S \vec{B} \circ d\vec{S} \right) = -\pi R_L^2 \frac{d}{dt} \|\vec{B}\| \quad (37.41)$$

Il vient alors:

$$-2\pi R_L \|\vec{E}\| = -\pi R_L^2 \frac{d}{dt} \|\vec{B}\| \quad (37.42)$$

d'où:

$$\|\vec{E}\| = \frac{R_L}{2} \frac{d}{dt} \|\vec{B}\| \quad (37.43)$$

Nous souhaiterions calculer l'énergie cinétique que la particule chargée négativement acquiert après plusieurs tours. Celle-ci est alors égale au travail fourni par le champ électrique pour déplacer la charge sur la trajectoire circulaire (pour rappel le champ magnétique ne "travaille" pas).

Comme nous l'avons montré dans le chapitre d'Électrostatique, nous avons dans le long d'un ligne de champ champ électrique (constant):

$$\Delta W = q \cdot E \cdot \Delta x \quad (37.44)$$

Dès lors il vient lorsque la charge parcourt N fois la circonférence du bêtatron:

$$\begin{aligned} \Delta E_c = \Delta W &= q \cdot E \cdot \Delta x = N \cdot 2\pi \cdot R_L \cdot q \cdot \|\vec{E}\| = N \cdot 2\pi \cdot R_L \cdot q \cdot \frac{R_L}{2} \frac{d}{dt} \|\vec{B}\| \\ &= \pi N q R_L^2 \frac{d}{dt} \|\vec{B}\| \end{aligned} \quad (37.45)$$

Exemple:

Considérons un champ magnétique sinusoïdal d'amplitude $B_0 = 1 [T]$ à une fréquence de $f = 50 [Hz]$, soit une période T de $20 [ms]$. Ce qui signifie qu'en $5 [ms]$ le champ magnétique passe d'un maximum à une valeur nulle. Considérons que nous avons un bêtatron avec une trajectoire circulaire de $1 [m]$ et que l'électron peut rester environ 480'000 tours sur cette trajectoire avec ce rayon précis sans trop dévier (soit l'équivalent d'à peu près 3'000 [km] parcourus). L'électron est injecté avec une énergie de $2 [MeV]$ dans le tore sous vide (ce qui est déjà très proche de la vitesse de la lumière!).

Calculons d'abord le rayon initial pour la trajectoire selon le rayon de Larmor relativiste. Pour cela, il nous faut d'abord la vitesse correspondant à l'énergie de 2 [MeV]:

$$E_c = mc^2 - m_0c^2 = (\gamma - 1)m_0c^2 = \left(\frac{1}{\sqrt{1 - \frac{v_0^2}{c^2}}} - 1 \right) m_0c^2 \quad (37.46)$$

Après quelques opérations algébriques élémentaires, nous trouvons:

$$v_0 = 299'407'509 \cong 99.87\% \cdot c \quad (37.47)$$

Dès lors le rayon de Larmor initial vaut:

$$R_L = \frac{\gamma \cdot m_0 \cdot v_0}{q \cdot B_0} = 0.007 [m] \quad (37.48)$$

Nous avons alors pendant toute la durée de l'accélération qui va amener l'électron à un rayon de Larmor de 1 [m] un gain d'énergie cinétique de:

$$\Delta E_c = \pi N q R_L^2 \frac{\Delta \|\vec{B}\|}{\Delta t} \cong \pi 480'000 \cdot 1.6 \cdot 10^{-19} \cdot 1^2 \cdot \frac{(1-0)}{(0.005-0)} = 4.82548632 \cdot 10^{-11} [J] \quad (37.49)$$

Ce qui correspond en électrons-volts à:

$$\Delta E_c \cong \frac{4.82548632 \cdot 10^{-11}}{1.6 \cdot 10^{-19}} \cong 300 [MeV] \quad (37.50)$$

soit l'énergie qui était mesurée expérimentalement à l'époque. Cette énergie correspond aussi à une vitesse qui est très très proche de celle de la lumière. Ainsi, avec le même calcul que précédemment, nous obtenons:

$$v_0 = 299'407'509 \cong 99.87\% \cdot c \quad (37.51)$$

vitesse atteinte en quelques centièmes de seconde seulement!

Remarque: Donc dans la réalité, la force centrifuge augmente au fur et à mesure que l'électron acquiert de l'énergie cinétique (et donc de la vitesse). Il faut compenser cette force en augmentant la force de Lorentz d'autant.

4. QUATRIÈME ÉQUATION DE MAXWELL

La 4ème équation de Maxwell est probablement la plus importante. Elle est une généralisation de la loi d'Ampère qui a déjà été présentée dans le chapitre de Magnétostatique et pour laquelle nous avons obtenu la circulation C_B du champ magnétique:

$$C_B = \oint_{\Gamma} \vec{B} \cdot d\vec{l} = \mu_0 \cdot I \quad (37.52)$$

La troisième équation de Maxwell nous dit que la variation d'un champ magnétique donne lieu à un champ électrique. Nous pouvons donc supposer que la réciproque est vraie.

Un endroit typique où l'on peut observer une variation d'un champ électrique est par exemple le condensateur (cf. chapitre d'Électrocinétique).

Nous savons que:

$$I = \frac{dQ}{dt} \quad (37.53)$$

et que le champ électrique entre deux plans parallèles, de surface S , portant des charges $Q_+ = -Q_-$, uniformément distribuées est donné par (cf. chapitre d'Électrostatique):

$$E(x) = \frac{\sigma}{\epsilon_0} = \frac{Q}{\epsilon_0 \cdot S} \quad (37.54)$$

où σ est la densité de charge surfacique.

Ce résultat est indépendant de la distance D entre les plans. La première équation de Maxwell donne:

$$\Delta U = - \int_0^d E(x) dx = - \frac{Q}{\epsilon_0 \cdot S} \cdot D \Rightarrow U = |\Delta U| = \frac{Q \cdot D}{\epsilon_0 \cdot S} \quad (37.55)$$

La capacité d'un condensateur étant définie par (cf. chapitre d'Électrostatique):

$$C = \frac{Q}{U} \quad (37.56)$$

nous avons obtenu dans le cas particulier d'un condensateur plan et parallèle que la capacité vaut:

$$C = \frac{\epsilon_0 \cdot S}{D} \quad (37.57)$$

Donc il vient:

$$I = \frac{dQ}{dt} = \frac{d(C \cdot u)}{dt} = C \frac{du}{dt} \quad (37.58)$$

et en utilisant le fait que le potentiel électrostatique est le champ électrique multipliée par une distance qui sera prise dans le cas présence comme la distance D entre les deux plans du condensateur, nous avons:

$$I = C \cdot \frac{d(D \cdot E)}{dt} = C \cdot D \cdot \frac{dE}{dt} = \epsilon_0 S \frac{dE}{dt} \quad (37.59)$$

Comme le champ électrique est variable, il est souvent d'usage de mettre un i minuscule pour le courant variable (c'est une tradition que nous retrouverons dans le chapitre d'Électrocinétique) et comme entre les deux plaques du condensateur il n'y a que du vide, nous parlons alors de "**courant de déplacement**", raison pour laquelle cette dernière relation est souvent notée sous la forme suivante:

$$I_D = i_D = \varepsilon_0 S \frac{dE}{dt} \quad (37.60)$$

En exprimant l'expression ci-dessus en utilisant la densité superficielle de courant, il vient:

$$\vec{j}_D = \varepsilon_0 \frac{d\vec{E}}{dt} \quad (37.61)$$

Si le champ électrique n'est pas homogène dans l'espace et dépend donc des coordonnées spatiales, nous devons utiliser les dérivées partielles tel que:

$$\vec{j}_D = \varepsilon_0 \frac{\partial \vec{E}}{\partial t} \quad (37.62)$$

Le courant de déplacement engendre un champ magnétique calculable au moyen de la loi d'Ampère:

$$C_B = \oint_{\Gamma} \vec{B} \circ d\vec{l} = \mu_0 \cdot I_D = \mu_0 \iint_{\Sigma} \vec{j}_D \circ d\vec{S} = \mu_0 \cdot \varepsilon_0 \iint_{\Sigma} \frac{d\vec{E}}{dt} \circ d\vec{S} = \mu_0 \cdot \varepsilon_0 \frac{d}{dt} \iint_{\Sigma} \vec{E} \circ d\vec{S} \quad (37.63)$$

Dans tout phénomène où nous observons un déplacement de charge, nous pouvons supposer qu'il y a création d'un courant de déplacement qui se superpose au courant de conduction à cause des effets capacitifs dans la matière. Nous écrivons dès lors:

$$C_B = \oint_{\Gamma} \vec{B} \circ d\vec{l} = \mu_0 \cdot I_{Tot} = \mu_0 \cdot (I + I_D) = \mu_0 \iint_{\Sigma} \vec{j} \circ d\vec{S} + \mu_0 \varepsilon_0 \frac{d}{dt} \iint_{\Sigma} \vec{E} \circ d\vec{S} \quad (37.64)$$

où nous avons (rappel du chapitre d'Électrostatique et de Magnétostatique):

$$\varepsilon = \varepsilon_0 \cdot \varepsilon_r \text{ et } \mu = \mu_0 \cdot \mu_r \quad (37.65)$$

D'autre part, le théorème de Stokes fournit que:

$$\int_{\Gamma} \vec{H} \circ d\vec{l} = \iint_{\Sigma} (\vec{\nabla} \times \vec{H}) \circ d\vec{S} \quad (37.66)$$

d'où:

$$\iint_{\Sigma} \left(\vec{j} + \varepsilon \frac{d\vec{E}}{dt} \right) \circ d\vec{S} = \iint_{\Sigma} (\vec{\nabla} \times \vec{H}) \circ d\vec{S} \quad (37.67)$$

et nous en ressortons finalement que:

$$\oint_{\Gamma} \vec{B} \circ d\vec{l} = \mu \iint_{\Sigma} \vec{j} \circ d\vec{S} + \mu \cdot \varepsilon \frac{d}{dt} \iint_{\Sigma} \vec{E} \circ d\vec{S} \Leftrightarrow \vec{\nabla} \times \vec{H} = \vec{j} + \varepsilon \frac{\partial \vec{E}}{\partial t} \quad (37.68)$$

Ceci est la "**quatrième équation de Maxwell**" ou "**équation de Maxwell-Ampère**". À gauche nous avons donc la forme intégrale de l'équation de Maxwell et à droite sa forme différentielle.

Explication: La quatrième et dernière équation de Maxwell associe la création d'un champ magnétique à toute variation d'un champ électrique et/ou à la présence d'un courant électrique (la présence d'un courant électrique étant une condition suffisante mais pas nécessaire au vu du deuxième terme). Nous disons alors que le terme avec le dérivée partielle du champ électrique est le "**terme de couplage électrique**".

Résumé:

Nous avons donc les quatre équations de Maxwell suivantes appelées "**formes locales des équations de Maxwell**" sous forme de différentielles (lorsque les intégrales ne sont pas indiquées):

$$\begin{array}{ll} \vec{\nabla} \circ \vec{E} = \frac{\rho_e}{\varepsilon} & \vec{\nabla} \circ \vec{B} = 0 \\ \vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} & \vec{\nabla} \times \vec{B} = \mu \vec{j} + \mu \varepsilon \frac{\partial \vec{E}}{\partial t} \end{array} \quad (37.69)$$

Dans le cas où $\mu_r, \varepsilon_r \neq 1$, c'est-à-dire dans le cas où nous ne travaillons pas dans le vide mais dans la matière, nous notons les équations locales de Maxwell sous la forme suivante:

$$\begin{array}{ll} \vec{\nabla} \circ \vec{D} = \rho_e & \vec{\nabla} \circ \vec{B} = 0 \\ \vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} & \vec{\nabla} \times \vec{H} = \vec{j} + \frac{\partial \vec{D}}{\partial t} \\ \vec{D} = \varepsilon \vec{E} & \vec{B} = \mu \vec{H} \end{array} \quad (37.70)$$

où $\vec{D}$ est (rappel) appelé "**champ de déplacement**" ou encore "**induction électrique**" et (rappel) $\vec{H}$ "**excitation magnétique**".

Remarque: Attention! $\vec{E}$ est une réaction du vide au champ $\vec{D}$. Cela s'explique par la constante de permittivité du vide mise dans l'intégrale (du moins c'est une façon de voir la chose...).

Mais dans le vide et dans le cas où nous considérons une absence de charges, nous obtenons:

$$\begin{array}{ll} \vec{\nabla} \circ \vec{E} = 0 & \vec{\nabla} \circ \vec{B} = 0 \\ \vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} & \vec{\nabla} \times \vec{B} = \varepsilon_0 \mu_0 \frac{\partial \vec{E}}{\partial t} \end{array} \quad (37.71)$$

Ce résultat est important, car il exprime la propagation possible d'un champ électrique et magnétique et ce même en l'absence de sources. Nous utiliserons ces équations pour déterminer les équations d'ondes électromagnétiques plus loin.

Remarque: Il est possible d'exprimer les équations de Maxwell sous forme relativiste (la relativité restreinte) mais... en réalité, comme nous l'avons déjà fait remarquer, les équations sont inchangées! En effet, les équations de Maxwell sont déjà relativistes. Ceci n'a rien d'étonnant, car les vecteurs des champs électrique et magnétique, les photons (cf. chapitre de Physique Quantique Des Champs), se propagent à la vitesse de la lumière. À cette vitesse, la relativité est reine et une théorie correcte ne pouvait être que relativiste. On peut toutefois exprimer les équations à l'aide des notations mathématiques tensorielles (voir plus loin notre démonstration du tenseur du champ électromagnétique). Sous cette forme les quatre équations deviennent incroyablement simples et compactes (une seule équation extrêmement courte). Formulés de cette manière, les champs électriques et magnétiques s'écrivent comme un champ unique appelé bien évidemment "**champ électromagnétique**". C'est un champ tensoriel comme nous le verrons plus loin.

Signalons enfin aussi la forme intégrale des quatre équations de Maxwell dans le vide que nous avons obtenus:

$$\begin{aligned}\Phi_E &= \oint_{\Sigma} \vec{E} \cdot d\vec{S} = \oint_V (\nabla \cdot \vec{E}) dV & \Phi_B &= \oint_{\Sigma} \vec{B} \cdot d\vec{S} = \oint_V (\nabla \cdot \vec{B}) dV \\ U(t) &= -\frac{d\Phi_B}{dt} = \oint_{\Gamma} \vec{E} \circ d\vec{l} & \oint_{\Gamma} \vec{B} \circ d\vec{l} &= \mu \oint_{\Sigma} \vec{j} \cdot d\vec{S} + \mu \cdot \varepsilon \frac{d}{dt} \oint_{\Sigma} \vec{E} \cdot d\vec{S}\end{aligned} \quad (37.72)$$

4.1. MONOPÔLES MAGNÉTIQUES

Remarquons qu'en optant pour le système de mesure naturel où $c = 1$, nous avons alors pour les équations de Maxwell dans le vide:

$$\begin{aligned}\nabla \cdot \vec{E} &= 0 & \nabla \times \vec{B} - \frac{\partial \vec{E}}{\partial t} &= \vec{0} \\ \nabla \cdot \vec{B} &= 0 & \nabla \times \vec{E} + \frac{\partial \vec{B}}{\partial t} &= \vec{0}\end{aligned} \quad (37.73)$$

puisque comme nous le démontrerons plus loin, dans le vide:

$$\varepsilon_0 \mu_0 = \frac{1}{c^2} \quad (37.74)$$

Alors la transformation:

$$\vec{B} \mapsto \vec{E} \quad \vec{E} \mapsto -\vec{B} \quad (37.75)$$

ramène la seconde paire d'équations précédentes à la première! Cette symétrie des équations de Maxwell est appelée "**dualité**" et c'est un indice qui tend à montrer que les champs électrique et magnétique ne sont que les parties unifiées d'un tout.

De plus, si nous introduisons le champ complexe suivant:

$$\vec{\mathcal{E}} = \vec{E} + i\vec{B} \quad (37.76)$$

la dualité (en prenant la partie réelle seulement), s'écrit alors:

$$\left. \begin{array}{l} \vec{B} \mapsto \vec{E} \\ \vec{E} \mapsto -\vec{B} \end{array} \right\} \vec{E} \mapsto -i\vec{E} \quad (37.77)$$

la paire d'équations de Maxwell indiquée précédemment se réduit alors à (nous utilisons la propriété de linéarité du produit vectoriel) une seule paire d'équations dont il ne faut pas oublier de ne prendre que la partie réelle:

$$\vec{\nabla} \circ \vec{E} = 0 \quad \vec{\nabla} \times \vec{E} = i \frac{\partial \vec{E}}{\partial t} \quad (37.78)$$

Cependant, cette symétrie ne s'étend pas aux équations de Maxwell avec sources exprimées dans le système naturel par:

$$\begin{array}{ll} \vec{\nabla} \circ \vec{E} = \rho_e & \vec{\nabla} \times \vec{B} - \frac{\partial \vec{E}}{\partial t} = \vec{j} \\ \vec{\nabla} \circ \vec{B} = 0 & \vec{\nabla} \times \vec{E} + \frac{\partial \vec{B}}{\partial t} = \vec{0} \end{array} \quad (37.79)$$

car cela se traduirait au mieux (n'oubliez pas de ne prendre que les coefficients réels pour le champ intéressé):

$$\vec{\nabla} \circ \vec{E} = \rho_e \quad \vec{\nabla} \times \vec{E} = i \left(\frac{\partial \vec{E}}{\partial t} + \vec{j} \right) \quad (37.80)$$

mais une fois sur deux cela ne marche pas (faites la substitution de $\vec{E}$ vous verrez que vous obtenez toujours une des équations sur la paire qui est conforme et l'autre pas). L'astuce consiste alors à séparer les deux densités en leur partie imaginaire et réelle respectives:

$$\begin{array}{l} \rho_e = \rho_e + i \cdot \rho_m \\ \vec{j} = \vec{j}_e + i \cdot \vec{j}_m \end{array} \quad (37.81)$$

Nous obtenons alors (toujours sans oublier de ne prendre les parties réelles et sans oublier que nous sommes en unités naturelles):

$$\begin{array}{ll} \vec{\nabla} \circ \vec{E} = \rho_e & \vec{\nabla} \times \vec{B} - \frac{\partial \vec{E}}{\partial t} = \vec{j}_e \\ \vec{\nabla} \circ \vec{B} = \rho_m & \vec{\nabla} \times \vec{E} + \frac{\partial \vec{B}}{\partial t} = \vec{j}_m \end{array} \quad (37.82)$$

il suffit alors de poser $\rho_m = \vec{j}_m = 0$. Ces équations sont certes, charmantes mais leur généralisation n'apporte rien de nouveau cependant car aucune charge magnétique exprimée par:

$$\vec{\nabla} \circ \vec{B} = \rho_m \quad (37.83)$$

appelée "**monopôle magnétique**" n'a été observée à ce jour. Dans un cadre expérimental, nous disons alors que $\rho_e, \vec{j}$ sont réels tel que nous ayons bien:

$$\rho_m = \vec{j}_m = 0 \quad (37.84)$$

Suite à la proposition d'un lecteur, indiquons que Dirac aurait proposé une autre manière beaucoup plus élégante pour compléter empiriquement les 4 équations de Maxwell avec sources démontrées plus haut:

$$\begin{aligned} \vec{\nabla} \circ \vec{E} &= 0 & \vec{\nabla} \circ \vec{B} &= 0 \\ \vec{\nabla} \times \vec{E} &= -\frac{\partial \vec{B}}{\partial t} & \vec{\nabla} \times \vec{B} &= \varepsilon_0 \mu_0 \frac{\partial \vec{E}}{\partial t} \end{aligned} \quad (37.85)$$

sous la forme suivante appelée "**équations de Dirac-Maxwell symétrisées**":

$$\begin{aligned} \vec{\nabla} \circ \vec{E} &= \frac{\rho_e}{\varepsilon} & \vec{\nabla} \circ \vec{B} &= \mu \rho_m \\ \vec{\nabla} \times \vec{E} &= -\frac{\partial \vec{B}}{\partial t} - \mu \vec{j}_m & \vec{\nabla} \times \vec{B} &= \mu \vec{j}_e + \mu \varepsilon \frac{\partial \vec{E}}{\partial t} \end{aligned} \quad (37.86)$$

Alors évidemment quand l'on voit cela on se demande qu'elle est l'argumentation qui permet d'arriver à supposer que la symétrie doit être ainsi écrite?! En réalité c'est très simple et très astucieux (comme souvent...). Effectivement, si nous prenons la divergence du rotationnel du champ électrique, nous avons de par le fait que (cf. chapitre de Calcul Vectoriel) celle-ci est toujours nulle quelle que soit la fonction considérée:

$$\vec{\nabla} \circ (\vec{\nabla} \times \vec{f}) = 0 \quad (37.87)$$

le résultat suivant:

$$\vec{\nabla} \circ (\vec{\nabla} \times \vec{E}) = \vec{\nabla} \circ \left(-\frac{\partial \vec{B}}{\partial t} - \mu \vec{j}_m \right) = 0 \quad (37.88)$$

Il en découle alors:

$$-\frac{\partial (\vec{\nabla} \circ \vec{B})}{\partial t} - \mu (\vec{\nabla} \circ \vec{j}_m) = 0 \quad (37.89)$$

et donc:

$$-\frac{\partial (\mu \rho_m)}{\partial t} - \mu \vec{\nabla} \circ \vec{j}_m = 0 \quad (37.90)$$

Soit après simplification et réarrangement:

$$\vec{\nabla} \circ \vec{j}_m + \frac{\partial \rho_m}{\partial t} = 0 \quad (37.91)$$

Donc c'est le fait de retomber sur une équation de continuité (de forme identique à l'équation de continuité en thermodynamique, en mécanique des fluides, de celle du champ électrique, de probabilité en mécanique quantique, etc.) qui aurait donc amené Dirac à compléter les 4 équations de Maxwell que nous avons écrites juste plus haut.

5. ÉQUATION DE CONSERVATION DE LA CHARGE

Nous avons donc démontré les quatre équations de Maxwell qui sont les fondements de l'électrodynamique classique.

Les équations de Maxwell peuvent être divisées en deux groupes:

- des "équations sans source":

$$\vec{\nabla} \circ \vec{B} = 0 \text{ et } \vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \quad (37.92)$$

- des "équations avec sources" (dans le vide):

$$\vec{\nabla} \circ \vec{E} = \frac{\rho_E}{\epsilon_0} \text{ et } \vec{\nabla} \times \vec{B} = \mu_0 \vec{j} + \frac{1}{c^2} \frac{\partial \vec{E}}{\partial t} \quad (37.93)$$

Dérivant la première équation avec sources par rapport au temps:

$$\frac{d(\vec{\nabla} \circ \vec{E})}{dt} = \vec{\nabla} \circ \frac{d\vec{E}}{dt} = \frac{1}{\epsilon_0} \frac{d\rho_E}{dt} \quad (37.94)$$

et prenant la divergence de la seconde, nous obtenons:

$$\vec{\nabla} \circ (\vec{\nabla} \times \vec{B}) = \mu_0 \vec{\nabla} \circ \vec{j} + \frac{1}{c^2} \vec{\nabla} \circ \frac{\partial \vec{E}}{\partial t} \quad (37.95)$$

La divergence d'un rotationnel est toujours nulle comme nous l'avons démontré dans le chapitre de Calcul Vectoriel et donc la dernière expression est nulle. Mais étant donné qu'un lecteur nous l'a demandé, nous détaillons ce résultat de façon plus explicite en simplifiant un peu:

$$(\vec{\nabla} \circ \vec{\nabla}) \times (\vec{\nabla} \circ \vec{B}) = \mu_0 \vec{\nabla} \circ \vec{j} + \frac{1}{c^2 \epsilon_0} \frac{\partial \rho_E}{\partial t} \quad (37.96)$$

or, $\vec{\nabla} \circ \vec{B} = 0$ et donc:

$$\mu_0 \vec{\nabla} \circ \vec{j} + \frac{1}{c^2 \epsilon_0} \frac{\partial \rho_E}{\partial t} = 0 \quad (37.97)$$

Après simplification via l'introduction des unités naturelles, nous obtenons:

$$\boxed{\vec{\nabla} \circ \vec{j} + \frac{\partial \rho_E}{\partial t} = 0} \quad (37.98)$$

qui est appelée "**équation de conservation de la charge**" ou "**équation de continuité**" et qui dit qu'en deux instants voisins $t + dt$, la variation dQ de la charge contenue dans une surface fermée délimitant un système ne peut être attribuée exclusivement qu'à un échange de charges avec l'extérieur.

Cette équation est très importante, car elle implique lors de l'étude de la relativité restreinte, que la charge est une quantité invariante par translation.

6. THÉORIE DE JAUGES

Avant de commencer à lire ce sous-chapitre, il est de première importance pour le lecteur d'aller faire un petit tour dans la section d'Algèbre du site, dans laquelle se trouve un chapitre de Calcul Vectoriel où nous faisons un rappel des différents opérateurs vectoriels indispensables en physique et de leurs propriétés.

Ce qui va suivre est très important car outre le fait que nous allons faire apparaître naturellement un nouveau champ (le potentiel-vecteur) qui est indispensable dans certaines équations de la physique quantique relativistes (voir chapitre du même nom) nous reprendrons cette démarche de jauges dans le chapitre de physique quantique ondulatoire où les conséquences sont beaucoup plus vastes!

Soit la relation connue:

$$\vec{\nabla} \circ \vec{B} = 0 \quad (37.99)$$

Il existe de par les propriétés des opérateurs rotationnel et divergence (cf. chapitre de Calcul Vectoriel) un "**potentiel vecteur**" $\vec{A}(\vec{r}, t)$ tel que:

$$\vec{B} = \vec{\nabla} \times \vec{A} \quad (37.100)$$

qui satisfait donc (la divergence du rotationnel d'un champ est toujours nulle car il s'agit d'une propriété mathématique):

$$\vec{\nabla} \circ (\vec{\nabla} \times \vec{A}) = 0 \quad (37.101)$$

Remarque: Le potentiel-vecteur est donc... un potentiel et un vecteur! De même que nous pouvons définir un potentiel U dont dérive $\vec{E}$, nous pouvons définir un potentiel $\vec{A}(\vec{r}, t)$ pour le champ $\vec{B}$. Mais pour des raisons techniques (provenant de l'expression des rotationnels de $\vec{E}$ et de $\vec{B}$ dans les équations de Maxwell), le potentiel $\vec{A}(\vec{r}, t)$ n'est pas aussi simple que U et ne peut pas s'exprimer comme un simple scalaire: il faut utiliser un potentiel-vecteur.

Si nous portons la relation $\vec{B} = \vec{\nabla} \times \vec{A}$ dans l'équation de Maxwell $\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$ nous obtenons:

$$\vec{\nabla} \times \left(\vec{E} + \frac{\partial \vec{A}}{\partial t} \right) = \vec{0} \quad (37.102)$$

Nous posons maintenant (la notation F n'a aucun rapport avec la force newtonienne!):

$$\vec{\nabla} \times \left(\vec{E} + \frac{\partial \vec{A}}{\partial t} \right) = \vec{\nabla} \times \vec{F} = \vec{0} \quad (37.103)$$

et nous utilisons les propriétés mathématiques des opérateurs rotationnel et gradient pour écrire une nouvelle relation (le signe "-" est là par anticipation de ce qui suivra):

$$\vec{F} = -\vec{\nabla} \phi \quad (37.104)$$

où dès lors:

$$\vec{E} = -\vec{\nabla} \phi - \frac{\partial \vec{A}}{\partial t} \quad (37.105)$$

où $\phi(\vec{r}, t)$ est un "potentiel scalaire".

Remarques:

R1. Le champ $\vec{F}$ semble obéir aux mêmes propriétés que le champ gravitationnel (loi de Newton-Poisson) mais ce n'est qu'une curiosité (les unités et les autres propriétés mathématiques n'étant pas équivalentes).

R2. Le lecteur voit sans peine que si le potentiel vecteur est nul, nous retrouvons alors (cf. chapitre d'Électrostatique):

$$\vec{E} = -\vec{\nabla} \phi \equiv -\vec{\nabla} U \quad (37.106)$$

ce qui renforce les hypothèses des développements précédents (et ce n'est pas tout...)

De plus, les champs $\vec{E}$ et $\vec{B}$ restent inchangés si nous effectuons dans les relations précédentes les remplacements suivants (les termes s'annulent trivialement):

$$\phi \rightarrow \phi' = \phi - \frac{\partial \psi}{\partial t}, \vec{A} \rightarrow \vec{A}' = \vec{A} + \vec{\nabla} \psi \quad (37.107)$$

où ψ est une fonction arbitraire de $\vec{r}$ et t .

Nous appelons une telle transformation un "changement de jauge". La liberté sur le choix des potentiels permet de leur imposer une contrainte que nous appelons la "contrainte de Jauge".

Il existe plusieurs manières de forger cette contrainte parmi lesquelles nous en distinguons deux:

Ainsi nous utiliserons soit la "jauge de Lorenz" en imposant:

$$\vec{\nabla} \cdot \vec{A} + \frac{1}{c^2} \frac{\partial \phi}{\partial t} = 0 \quad (37.108)$$

ou soit la "jauge de Coulomb" en imposant:

$$\boxed{\vec{\nabla} \circ \vec{A} = 0} \quad (37.109)$$

Remarque: Nous trouvons souvent dans la littérature, la dénomination "jauge de Lorentz" à la place de "jauge de Lorenz", car comme nous l'avons déjà démontré dans le chapitre de Relativité Restreinte, la jauge de Lorenz est invariante dans les transformations de Lorentz (ce qui constitue un avantage certain par rapport à la jauge de Coulomb!).

Montrons qu'il est toujours possible d'imposer la jauge de Coulomb. Pour cela, étant donnés $\vec{A}'$ et $\vec{\phi}'$, il suffit de trouver ψ dans les équations:

$$\phi \rightarrow \phi' = \phi - \frac{\partial \psi}{\partial t}, \vec{A} \rightarrow \vec{A}' = \vec{A} + \vec{\nabla} \psi \quad (37.110)$$

tel que la relation (jauge de Coulomb):

$$\vec{\nabla} \circ \vec{A} = 0 \quad (37.111)$$

soit vérifiée. Ainsi, ψ doit vérifier:

$$\begin{aligned} \vec{\nabla} \circ \vec{A}' &= \vec{\nabla} \circ \vec{A} + \vec{\nabla} \circ \vec{\nabla} \psi \Rightarrow \vec{\nabla} \circ \vec{A}' - \vec{\nabla} \circ \vec{\nabla} \psi = \vec{\nabla} \circ \vec{A} = 0 \\ \Rightarrow \vec{\nabla} \circ \vec{A}' &= \vec{\nabla} \circ \vec{\nabla} \psi = \Delta \psi = f(\vec{r}, t) \end{aligned} \quad (37.112)$$

La relation:

$$\boxed{\vec{\nabla} \circ \vec{A}' = \Delta \psi} \quad (37.113)$$

est appelée "**équation de Poisson du potentiel-vecteur**".

De même, pour montrer qu'il est toujours possible d'imposer la condition de Lorenz, il suffit de trouver ψ dans les équations précitées:

$$\phi \rightarrow \phi' = \phi - \frac{\partial \psi}{\partial t}, \vec{A} \rightarrow \vec{A}' = \vec{A} + \vec{\nabla} \psi \quad (37.114)$$

tel que la relation (jauge de Lorenz):

$$\boxed{\vec{\nabla} \circ \vec{A}' + \frac{1}{c^2} \frac{\partial \phi'}{\partial t} = 0} \quad (37.115)$$

soit vérifiée. Ainsi, ψ doit vérifier:

$$\begin{aligned} \vec{\nabla} \circ \vec{A}' + \frac{1}{c^2} \frac{\partial \phi'}{\partial t} &= \vec{\nabla} \circ \vec{A} + \vec{\nabla} \circ \vec{\nabla} \psi + \frac{1}{c^2} \frac{\partial \phi}{\partial t} - \frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2} = \vec{\nabla} \circ \vec{A} + \Delta \psi + \frac{1}{c^2} \frac{\partial \phi}{\partial t} - \frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2} \\ \Rightarrow \vec{\nabla} \circ \vec{A}' + \frac{1}{c^2} \frac{\partial \phi'}{\partial t} &= \Delta \psi - \frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2} \end{aligned} \quad (37.116)$$

Soit en d'autres termes et de façon plus condensée:

$$\square \psi = \frac{1}{c^2} \frac{\partial \phi'}{\partial t} + \vec{\nabla} \circ \vec{A}' = g(\vec{r}, t) \quad (37.117)$$

où l'opérateur:

$$\square = \frac{1}{c^2} \frac{\partial^2}{\partial t^2} - \Delta \quad (37.118)$$

est par définition appelé le "**d'Alembertien**" (nous le retrouverons souvent ce terme à partir de maintenant aussi bien en électrodynamique qu'en physique quantique) qui est donc aussi invariant par transformation Lorentz comme nous le verrons lors de notre étude de la relativité restreinte (cf. chapitre de Relativité Restreinte).

En reportant les équations:

$$\vec{B} = \vec{\nabla} \times \vec{A} \text{ et } \vec{E} = -\vec{\nabla} \phi - \frac{\partial \vec{A}}{\partial t} \quad (37.119)$$

dans les deux autres équations de Maxwell dans le vide:

$$\vec{\nabla} \circ \vec{E} = \frac{\rho_g}{\epsilon_0} \text{ et } \vec{\nabla} \times \vec{B} = \mu_0 \vec{j} + \frac{1}{c^2} \frac{\partial \vec{E}}{\partial t} \quad (37.120)$$

nous obtenons, en faisant apparaître le laplacien d'un champ vectoriel $\Delta \vec{A}$ par une des propriétés des opérateurs vectoriels rotationnel, gradient et divergence (cf. chapitre de Calcul Vectoriel):

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{A}) = \vec{\nabla} (\vec{\nabla} \circ \vec{A}) - \Delta \vec{A} \quad (37.121)$$

les relations suivantes:

$$\begin{aligned} -\Delta \phi &= \frac{\rho_g}{\epsilon_0} + \vec{\nabla} \circ \frac{\partial \vec{A}}{\partial t} \\ \square \vec{A} &= \mu_0 \vec{j} - \vec{\nabla} \left(\vec{\nabla} \circ \vec{A} + \frac{1}{c^2} \frac{\partial \phi}{\partial t} \right) \end{aligned} \quad (37.122)$$

la dernière relation étant appelée "**jauge arbitraire**".

Pour la jauge de Lorenz, ces deux dernières équations se simplifient en (n'hésitez pas à nous contacter si vous ne voyez pas comment):

$$\begin{aligned} \square \phi &= \frac{\rho_g}{\epsilon_0} \\ \square \vec{A} &= \mu_0 \vec{j} \end{aligned} \quad (37.123)$$

que nous appelons "**équations d'onde des potentiels électromagnétiques**" en analogie avec les équations d'onde des champs électrique et magnétique que nous déterminerons plus loin.

Pour la jauge de Coulomb, les mêmes équations se simplifient en:

$$\begin{aligned} \Delta\phi &= -\frac{\rho_B}{\varepsilon_0} \\ \square \vec{A} &= \mu_0 \vec{j} - \frac{1}{c^2} \vec{\nabla} \frac{\partial \phi}{\partial t} \end{aligned} \quad (37.124)$$

Sachant que $c^2 = 1/\varepsilon_0 \mu_0$ nous pouvons aussi écrire les deux équations d'onde des potentiels électromagnétiques sous la forme:

$$\begin{aligned} \square \phi &= \frac{\rho_B}{\varepsilon_0} = \mu_0 c^2 \rho_B \\ \square \vec{A} &= \mu_0 \vec{j} \end{aligned} \quad (37.125)$$

Posons maintenant $A^0 = \phi/c$ (afin d'homogénéiser les unités) tel que nous définissions un "**quadrivecteur potentiel**" qui nous permet d'écrire vectoriellement les deux relations ci-dessus de manière unifiée:

$$\square A^\alpha = (\phi/c, \mu_0 \vec{j}) = (\mu_0 c \rho_B, \mu_0 \vec{j}) = \mu_0 (c \rho_B, \vec{j}) \quad (37.126)$$

Remarque: Le fait que le d'Alembertien du quadrivecteur potentiel s'exprime à partir du quadrivecteur courant qui est contravariant (cf. chapitre de Relativité Restreinte) nous amène à poser que le quadrivecteur potentiel est lui-même contravariant!

Relation que nous noterons sous une forme condensée de la manière suivante:

$$\square A^\alpha = j^\alpha \quad (37.127)$$

où j^α sera appelé "**quadrivecteur courant**".

Remarque: Nous retrouverons ce quadrivecteur lors de notre détermination du tenseur du champ électromagnétique plus loin (à la différence que nous serons en unités naturelles mais cela ne change pas le fond...).

Le quadrivecteur potentiel tel que défini nous amène à pouvoir écrire la (quadrivergence) jauge de Lorenz en faisant usage de la notation tensorielle:

$$\begin{aligned} \vec{\nabla} \circ \vec{A} + \frac{1}{c^2} \frac{\partial \phi}{\partial t} &= \frac{1}{c^2} \frac{\partial \phi}{\partial t} + \partial_x A_x + \partial_y A_y + \partial_z A_z \\ &= \frac{1}{c} \frac{\partial \phi}{\partial t} + \partial_1 A^1 + \partial_2 A^2 + \partial_3 A^3 = \partial_0 A^0 + \partial_1 A^1 + \partial_2 A^2 + \partial_3 A^3 = 0 \end{aligned} \quad (37.128)$$

Ce qui permet finalement d'écrire la jauge de Lorenz sous forme covariante:

$$\partial_\alpha A^\alpha = 0 \quad (37.129)$$

Il s'agit donc d'une équation de la forme de celle de Klein-Gordon pour une particule de masse nulle (cf. chapitre de Physique Quantique Relativiste). Donc, nous pouvons dire dans un sens que l'invariance de jauge électromagnétique est reliée au fait que la masse du photon est nulle!

Remarque: Il est utile de noter que le fait de poser $\frac{1}{c} \frac{\partial}{\partial t} = \partial_0$ (avec ou sans les unités naturelles où $c = 1$) est une notation qui sera également adoptée lors de notre étude de l'équation de Dirac (cf. chapitre de Physique Quantique Relativiste) ou encore en physique quantique de champs (mis à part qu'il y aura une partie imaginaire).

Ces notations nous amènent enfin à pouvoir écrire:

$$\begin{aligned}\partial_\alpha (\Box A^\alpha) &= \partial_\alpha (\mu_0 j^\alpha) \\ \partial_\alpha A^\alpha &= \mu_0 \partial_\alpha j^\alpha = 0\end{aligned}\quad (37.130)$$

Nous obtenons ainsi l'équation de continuité:

$$\boxed{\partial_\alpha j^\alpha = 0} \quad (37.131)$$

équivalent (sous forme) tensoriel de (voir la démonstration juste plus haut dans le texte):

$$\vec{\nabla} \circ \vec{j} + \frac{\partial \rho_e}{\partial t} = 0 \quad (37.132)$$

Pour résumer en gros...:

Un certain nombre d'effets physiques se modélisent, selon les cas, par des champs qui peuvent être scalaires, vectoriels, spinoriels ou encore tensoriels que nous appelons donc des jauges. Un certain nombre de phénomènes physiques s'avèrent respecter des conditions dites de symétrie, vis-à-vis de ces jauges. Cette symétrie s'exprime par ce que nous appelons donc une invariance de jauge.

Par exemple, le champ qui permet de modéliser le champ électromagnétique est comme nous l'avons vu, un champ de quadrivecteurs formé d'un potentiel scalaire ϕ (dont le gradient est le champ électrique $\vec{E}$) et d'un potentiel-vecteur $\vec{A}$ (dont le rotationnel est le champ magnétique $\vec{B}$). Ce champ quadrivectoriel qui permet de modéliser le champ électromagnétique est appelé une jauge.

Il s'avère que nous obtenions donc exactement les mêmes effets physiques sur un système de particules chargées si nous remplaçons cette jauge par une autre jauge en lui rajoutant une contrainte de jauge (exemple typique entre la jauge de Lorenz ou de Coulomb vues plus haut). L'invariance des lois de la physique lors du passage d'une jauge à une autre étant une invariance de jauge. Dans le cas du champ électromagnétique, cette invariance de jauge s'avère exprimer la conservation de la charge électrique (comme nous l'avons montré).

Mathématiquement, de tels changements de jauges s'avèrent être le résultat de l'action d'un groupe de symétrie de dimension infinie (transformant ces jauges les unes en les autres) que nous appelons le "groupe de jauge" de l'interaction considérée (ici l'interaction électromagnétique).

Pour le champ gravitationnel par exemple (cf. chapitre de Relativité Restreinte), l'interaction gravitationnelle se modélise par un champ de tenseurs symétriques de rang 2 et avec une signature donnée. Ce champ de métrique est distribué sur une variété 4D modélisant l'espace-temps. C'est la jauge de l'interaction gravitationnelle. D'après la relativité générale (principe d'équivalence) nous ne changeons rien à l'interaction gravitationnelle si nous changeons le système de coordonnées spatio-temporelles dans lequel nous exprimons la métrique. Le passage d'une expression de la métrique à une autre en changeant de système de coordonnées est aussi un changement de jauge. L'invariance de jauge de la relativité générale exprime alors la possibilité de passer d'une jauge à une autre sans changer pour autant les géodésiques suivies par des particules tests tombant en chute libre dans le champ gravitationnel modélisé par le champ de métrique.

L'invariance de jauge de la relativité générale est ce que nous appelons l'invariance par difféomorphisme (changement de système de coordonnées bijectif présentant un certain degré de régularité) et le groupe de jauge de la relativité générale est donc le groupe des difféomorphismes de $\mathbb{R}^4$ (appelé le "groupe souple").

Il convient de préciser aussi que le potentiel-vecteur $\vec{A}$ n'est peut-être pas si virtuel que ça. En effet, il est possible de modifier les trajectoires de particules chargées passant à l'extérieur du volume cylindrique où règne un champ magnétique $\vec{B}$ induit par un courant électrique (circulant dans l'enroulement d'un solénoïde où ce champ $\vec{B}$ est "emprisonné"). Il est donc possible d'influer sur la trajectoire de particules circulant dans une zone où le champ magnétique $\vec{B}$ est nul mais où son potentiel-vecteur $\vec{A}$ ne l'est pas.

Par ailleurs, nous utiliserons les résultats obtenus ici lors de notre étude de la théorie de Yang-Mills dans la voie de l'unification électrofaible (voir le modèle standard dans le chapitre de Physique Quantique Des Champs).

Remarque: L'expérience connue qui fait intervenir le potentiel-vecteur est celle d'Aharonov-Bohm (cf. chapitre de Physique Quantique Ondulatoire).

6.1. TENSEUR DU CHAMP ÉLECTROMAGNÉTIQUE

Afin de déterminer le tenseur du champ électromagnétique supposons dans un premier temps que l'action (cf. chapitre de Mécanique Analytique) d'une particule chargée dans un champ électromagnétique serait donnée par (choix a priori empirique mais... vous verrez un peu plus loin):

$$S_1 = \int (q\vec{A} \circ \vec{v} - q\phi) dt = \int L_1 dt \quad (37.133)$$

Remarque: La notation S_0 reste réservée à l'action d'une particule libre (cf. chapitre de Relativité Restreinte).

Le lagrangien pour une particule chargée dans un champ électromagnétique est donc la somme du lagrangien de la particule en interaction avec le champ électromagnétique L_1 additionné au lagrangien de la particule libre L_0 (cf. chapitre de Relativité Restreinte):

$$L = L_0 + L_1 = -m_0 c^2 \sqrt{1 - v^2 / c^2} + q\vec{A} \circ \vec{v} - q\phi \quad (37.134)$$

Remarque: Il s'agit donc du lagrangien de l'interaction de la particule avec le champ additionné au lagrangien de la masse de la particule. Dès lors on voit qu'il manque encore le lagrangien du champ électromagnétique lui-même en l'absence de charges (appelé: lagrangien du champ libre) mais nous verrons cela plus loin.

Ceci est donc (a priori) le lagrangien d'une particule chargée dans un champ électromagnétique.

Nous allons démontrer que ce lagrangien est correct:

Le moment généralisé est donc (cf. chapitre de Mécanique Analytique et de Relativité Restreinte):

$$\vec{p} = \frac{\partial L}{\partial \vec{v}} = \frac{m_0 \vec{v}}{\sqrt{1 - v^2 / c^2}} + q \vec{A} \quad (37.135)$$

Pour vérifier que nous avons fait le bon choix de lagrangien au départ, nous allons obtenir les équations du mouvement et s'assurer qu'elles coïncident avec la force de Lorentz. Les équations de Lagrange sont, dans ce cas:

$$\begin{aligned} \frac{d}{dt} \frac{\partial L}{\partial \vec{v}} - \frac{\partial L}{\partial \vec{r}} &= 0 = \frac{d}{dt} \left(\frac{m_0 \vec{v}}{\sqrt{1 - v^2 / c^2}} + q \vec{A} \right) - \frac{\partial}{\partial \vec{r}} \left(-m_0 c \sqrt{1 - v^2 / c^2} + q \vec{A} \circ \vec{v} - q \phi \right) \\ &= \frac{d}{dt} \left(\frac{m_0 \vec{v}}{\sqrt{1 - v^2 / c^2}} \right) + q \dot{\vec{A}} - q \vec{\nabla} \cdot (\vec{A} \circ \vec{v}) + q \vec{\nabla} \cdot \phi \end{aligned} \quad (37.136)$$

Or nous avons:

$$\vec{F} = \frac{d}{dt} \left(\frac{m_0 \vec{v}}{\sqrt{1 - v^2 / c^2}} \right) \quad (37.137)$$

et donc:

$$\vec{F} = -q \dot{\vec{A}} + q \vec{\nabla} \cdot (\vec{A} \circ \vec{v}) - q \vec{\nabla} \cdot \phi \quad (37.138)$$

Mais nous avons fait remarquer lors de la définition du potentiel scalaire que $\vec{E} = -\vec{\nabla} \cdot \phi$ d'où:

$$\vec{F} = q \vec{E} + q \left(\vec{\nabla} \cdot (\vec{A} \circ \vec{v}) - \dot{\vec{A}} \right) \quad (37.139)$$

Nous devrions donc nécessairement avoir par analogie avec la force de Lorentz:

$$\vec{\nabla} \cdot (\vec{A} \circ \vec{v}) - \dot{\vec{A}} = \vec{v} \times \vec{B} = \vec{v} \times (\vec{\nabla} \times \vec{A}) \quad (37.140)$$

Il nous faut donc avant de poursuivre, vérifier que:

$$\vec{\nabla} \cdot (\vec{A} \circ \vec{v}) - \frac{d\vec{A}}{dt} = \vec{v} \times (\vec{\nabla} \times \vec{A}) \quad (37.141)$$

Avec:

$$d\vec{A} = \frac{\partial \vec{A}}{\partial x} dx + \frac{\partial \vec{A}}{\partial y} dy + \frac{\partial \vec{A}}{\partial z} dz \Rightarrow \frac{d\vec{A}}{dt} = \frac{\partial \vec{A}}{\partial x} \frac{dx}{dt} + \frac{\partial \vec{A}}{\partial y} \frac{dy}{dt} + \frac{\partial \vec{A}}{\partial z} \frac{dz}{dt} = \frac{\partial \vec{A}}{\partial x} v_x + \frac{\partial \vec{A}}{\partial y} v_y + \frac{\partial \vec{A}}{\partial z} v_z \quad (37.142)$$

En composantes:

$$\frac{d\vec{A}}{dt} = \begin{pmatrix} v_x \partial_x A_x + v_y \partial_y A_x + v_z \partial_z A_x \\ v_x \partial_x A_y + v_y \partial_y A_y + v_z \partial_z A_y \\ v_x \partial_x A_z + v_y \partial_y A_z + v_z \partial_z A_z \end{pmatrix} \quad (37.143)$$

Donc:

$$\begin{pmatrix} \frac{\partial}{\partial x} A_x v_x + \frac{\partial}{\partial x} A_y v_y + \frac{\partial}{\partial x} A_z v_z \\ \frac{\partial}{\partial y} A_x v_x + \frac{\partial}{\partial y} A_y v_y + \frac{\partial}{\partial y} A_z v_z \\ \frac{\partial}{\partial z} A_x v_x + \frac{\partial}{\partial z} A_y v_y + \frac{\partial}{\partial z} A_z v_z \end{pmatrix} - \dot{\vec{A}} = \vec{v} \times \begin{pmatrix} \partial_y A_z - \partial_z A_y \\ \partial_z A_x - \partial_x A_z \\ \partial_x A_y - \partial_y A_x \end{pmatrix}$$

$$\begin{pmatrix} v_x \partial_x A_x + v_y \partial_x A_y + v_z \partial_x A_z \\ v_x \partial_y A_x + v_y \partial_y A_y + v_z \partial_y A_z \\ v_x \partial_z A_x + v_y \partial_z A_y + v_z \partial_z A_z \end{pmatrix} = \begin{pmatrix} v_y (\partial_x A_y - \partial_y A_x) - v_z (\partial_z A_x - \partial_x A_z) \\ v_z (\partial_y A_z - \partial_z A_y) - v_x (\partial_x A_y - \partial_y A_x) \\ v_x (\partial_z A_x - \partial_x A_z) - v_y (\partial_y A_z - \partial_z A_y) \end{pmatrix} + \dot{\vec{A}} \quad (37.144)$$

$$\begin{pmatrix} v_x \partial_x A_x + v_y \partial_x A_y + v_z \partial_x A_z \\ v_x \partial_y A_x + v_y \partial_y A_y + v_z \partial_y A_z \\ v_x \partial_z A_x + v_y \partial_z A_y + v_z \partial_z A_z \end{pmatrix} = \begin{pmatrix} v_y \partial_x A_y - v_y \partial_y A_x - v_z \partial_z A_x + v_z \partial_x A_z \\ v_z \partial_y A_z - v_z \partial_z A_y - v_x \partial_x A_y + v_x \partial_y A_x \\ v_x \partial_z A_x - v_x \partial_x A_z - v_y \partial_y A_z + v_y \partial_z A_y \end{pmatrix} + \dot{\vec{A}}$$

et comme:

$$\begin{pmatrix} v_y \partial_x A_y - v_y \partial_y A_x - v_z \partial_z A_x + v_z \partial_x A_z \\ v_z \partial_y A_z - v_z \partial_z A_y - v_x \partial_x A_y + v_x \partial_y A_x \\ v_x \partial_z A_x - v_x \partial_x A_z - v_y \partial_y A_z + v_y \partial_z A_y \end{pmatrix} + \dot{\vec{A}}$$

$$= \begin{pmatrix} v_y \partial_x A_y - \cancel{v_y \partial_y A_x} - \cancel{v_z \partial_z A_x} + v_z \partial_x A_z \\ v_z \partial_y A_z - \cancel{v_z \partial_z A_y} - \cancel{v_x \partial_x A_y} + v_x \partial_y A_x \\ v_x \partial_z A_x - \cancel{v_x \partial_x A_z} - \cancel{v_y \partial_y A_z} + v_y \partial_z A_y \end{pmatrix} + \begin{pmatrix} \cancel{v_x \partial_x A_x} + \cancel{v_y \partial_y A_y} + \cancel{v_z \partial_z A_z} \\ \cancel{v_x \partial_x A_y} + v_y \partial_y A_y + \cancel{v_z \partial_z A_y} \\ \cancel{v_x \partial_x A_z} + \cancel{v_y \partial_y A_z} + v_z \partial_z A_z \end{pmatrix} \quad (37.145)$$

$$= \begin{pmatrix} v_y \partial_x A_y + v_z \partial_x A_z + v_x \partial_x A_x \\ v_z \partial_y A_z + v_x \partial_y A_x + v_y \partial_y A_y \\ v_x \partial_z A_x + v_y \partial_z A_y + v_z \partial_z A_z \end{pmatrix} = \begin{pmatrix} v_x \partial_x A_x + v_y \partial_x A_y + v_z \partial_x A_z \\ v_x \partial_y A_x + v_y \partial_y A_y + v_z \partial_y A_z \\ v_x \partial_z A_x + v_y \partial_z A_y + v_z \partial_z A_z \end{pmatrix}$$

Nous avons donc bien l'égalité:

$$\begin{pmatrix} v_x \partial_x A_x + v_y \partial_x A_y + v_z \partial_x A_z \\ v_x \partial_y A_x + v_y \partial_y A_y + v_z \partial_y A_z \\ v_x \partial_z A_x + v_y \partial_z A_y + v_z \partial_z A_z \end{pmatrix} = \begin{pmatrix} v_x \partial_x A_x + v_y \partial_x A_y + v_z \partial_x A_z \\ v_x \partial_y A_x + v_y \partial_y A_y + v_z \partial_y A_z \\ v_x \partial_z A_x + v_y \partial_z A_y + v_z \partial_z A_z \end{pmatrix} \quad (37.146)$$

Ces développements confirment donc notre hypothèse initiale comme quoi l'action du champ peut s'écrire:

$$S_1 = \int (q \vec{A} \circ \vec{v} - q \phi) dt = \int L_1 dt \quad (37.147)$$

et qu'elle exprime l'interaction d'une particule chargée avec un champ (car on y retrouve la force de Lorentz!).

Nous avons donc maintenant démontré que le "**lagrangien de l'interaction courants-champs**":

$$L = -m_0 c^2 \sqrt{1 - v^2 / c^2} + q \vec{A} \circ \vec{v} - q \phi \quad (37.148)$$

dont nous avons supposé empiriquement la forme au début est donc finalement bien correcte!

L'intégrale d'action s'écrivant alors:

$$S = \int_{t_1}^{t_2} \left[-m_0 c^2 \sqrt{1 - v^2 / c^2} + q \vec{A} \circ \vec{v} - q \phi \right] dt \quad (37.149)$$

Introduisons la vitesse $\vec{v}$ de la particule sous la forme $\vec{v} = d\vec{r} / dt$ et l'intégrale s'écrit:

$$S = \int_{t_1}^{t_2} \left(-m_0 c^2 \sqrt{1 - v^2 / c^2} dt + q \vec{A} \circ d\vec{r} - q \phi dt \right) \quad (37.150)$$

Nous avons vu en relativité restreinte que:

$$ds^2 = c^2 dt^2 - (dx^2 + dy^2 + dz^2) \quad (37.151)$$

et de même:

$$ds'^2 = c^2 dt'^2 - (dx'^2 + dy'^2 + dz'^2) \quad (37.152)$$

Les intervalles d'espace-temps sont des invariants tels que (cf. chapitre de Relativité Restreinte):

$$c^2 dt'^2 - (dx'^2 + dy'^2 + dz'^2) = c^2 dt^2 - (dx^2 + dy^2 + dz^2) \quad (37.153)$$

Si le référentiel O' n'est pas en mouvement ($dx' = dy' = dz' = 0$), nous avons:

$$ds^2 = c^2 dt'^2 = c^2 dt^2 - (dx^2 + dy^2 + dz^2) \quad (37.154)$$

d'où:

$$dt' = \frac{ds}{c} = dt \sqrt{1 - \frac{dx^2 + dy^2 + dz^2}{c^2 dt^2}} \quad (37.155)$$

ce qui s'écrit aussi:

$$\frac{ds}{c} = dt \sqrt{1 - \frac{v^2}{c^2}} \quad (37.156)$$

Dès lors:

$$S = \int_{t_1}^{t_2} \left(-m_0 c^2 \sqrt{1 - v^2 / c^2} dt + q \vec{A} \circ d\vec{r} - q \phi dt \right) = \int_{\alpha}^{\beta} \left(-m_0 c ds + q \vec{A} \circ d\vec{r} - q \phi dt \right) \quad (37.157)$$

Faisons usage du quadrivecteur potentiel contravariant (voir plus haut):

$$A^i = \left(\frac{\phi}{c}, A_1, A_2, A_3 \right) = \left(\frac{\phi}{c}, \vec{A} \right) \quad (37.158)$$

et du quadrivecteur déplacement contravariant (cf. chapitre de Relativité Restreinte):

$$dx^i = (cdt, dx_1, dx_2, dx_3) = (cdt, d\vec{x}) \quad (37.159)$$

L'expression de l'action d'une particule chargée dans un champ électromagnétique et dans une métrique de Minkowski η_{ij} (cf. chapitre de Relativité Restreinte et Relativité Générale) se réduit finalement à l'expression condensée:

$$S = \int_{\alpha}^{\beta} -m_0 c ds - q A^i (\eta_{ij} dx^j) = \int_{\alpha}^{\beta} -m_0 c ds - q A^i dx_i \quad (37.160)$$

avec donc:

$$dx_i = (cdt, -dx, -dy, -dz) \quad (37.161)$$

sans oublier que nous utilisons ici la métrique +,-,-,- (cf. chapitre de Relativité Restreinte et Relativité Générale).

Remarquons que l'intégrale d'action en l'absence de champ magnétique et électrique s'écrit:

$$S = -m_0 c \int_{\alpha}^{\beta} ds \quad (37.162)$$

ce qui correspond bien à ce que nous avons obtenu en relativité restreinte pour une particule libre !

D'après le principe de moindre action, l'intégrale d'action a une variation nulle pour le mouvement effectif de la particule, soit:

$$\delta S = - \left(\int_{\alpha}^{\beta} m_0 c \delta ds + q \delta \left(A^i dx_i \right) \right) = 0 \quad (37.163)$$

Remarque: De par l'égalité avec zéro, nous pouvons éliminer le signe moins devant l'intégrale.

Utilisant l'expression de l'abscisse curviligne (cf. chapitres de Calcul tensoriel et de Relativité Générale):

$$ds^2 = g_{ij} dx^i dx^j \quad (37.164)$$

Pour la métrique de Minkowski, nous pouvons écrire (rappelons que dans la métrique euclidienne seuls les termes de la diagonale où $i = j$ sont non nuls):

$$ds^2 = \eta_{ij} dx^i dx^j = dx_i dx^i \quad (37.165)$$

Ainsi:

$$ds = \sqrt{dx_i dx^i} \Leftrightarrow \delta ds = \frac{1}{2} \frac{\delta dx_i dx^i}{\sqrt{dx_i dx^i}} + \frac{1}{2} \frac{dx_i \delta dx^i}{\sqrt{dx_i dx^i}} = \frac{dx^i \delta dx_i}{ds} \quad (37.166)$$

l'intégrale précédente s'écrit alors:

$$\delta S = \int_{\alpha}^{\beta} m_0 c \delta ds + q \delta \left(A^i dx_i \right) = \int_{\alpha}^{\beta} m_0 c \frac{dx^i \delta dx_i}{ds} + q A^i \delta dx_i + q \delta A^i dx_i \quad (37.167)$$

Cela donne en utilisant les composantes curvilignes (cf. chapitre de Calcul Tensoriel):

$$\delta S = \int_{\alpha}^{\beta} \left(m_0 c u^i + q A^i \right) \delta dx_i + \int_{\alpha}^{\beta} q \delta A^i dx_i \quad (37.168)$$

Intégrons par parties (cf. chapitre de Calcul Intégral Et Différentiel) la première intégrale:

$$\delta S = \left(m_0 c u^i + q A^i \right) dx_i \Big|_{\alpha}^{\beta} - \int_{\alpha}^{\beta} \left(m_0 c du^i \delta x_i + q dA^i \delta x_i \right) + \int_{\alpha}^{\beta} q \delta A^i dx_i \quad (37.169)$$

Or, comme:

$$\delta x_i \Big|_{\alpha} = 0 \Rightarrow dx_i \Big|_{\alpha} = 0 \text{ et } \delta x_i \Big|_{\beta} = 0 \Rightarrow dx_i \Big|_{\beta} = 0 \quad (37.170)$$

Alors:

$$\delta S = \int_{\alpha}^{\beta} \left(-m_0 c du^i \delta x_i - q dA^i \delta x_i + q \delta A^i dx_i \right) \quad (37.171)$$

avec:

$$\delta A^i = \frac{\partial A^i}{\partial x^k} \delta x^k = \partial^k A^i \delta x^k \text{ et } dA^i = \frac{\partial A^i}{\partial x^k} dx^k = \partial^k A^i dx^k \quad (37.172)$$

devient:

$$\begin{aligned} \delta S &= \int_{\alpha}^{\beta} \left(-m_0 c du^i \delta x_i - q \partial^k A^i dx^k \delta x_i + q \partial^k A^i \delta x^k dx_i \right) \\ &= \int_{\alpha}^{\beta} \left(-m_0 c \left(\frac{du^i}{ds} ds \right) \delta x_i - q \partial^k A^i (u^k ds) \delta x_i + q \partial^k A^i \delta x^k (u_i ds) \right) \\ &= \int_{\alpha}^{\beta} \left(-m_0 c \left(\frac{du^i}{ds} ds \right) \delta x_i - q \partial^k A^i (u^k ds) \delta x_i + q \partial^i A^k \delta x^i (u_k ds) \right) \quad (37.173) \\ &= \int_{\alpha}^{\beta} \left(-m_0 c \left(\frac{du^i}{ds} \right) - q \partial^k A^i (u^k) + q \partial^i A^k (u_k) \right) \delta x_i ds \\ &= \int_{\alpha}^{\beta} \left(-m_0 c \left(\frac{du^i}{ds} \right) + q (\partial^i A^k - \partial^k A^i) u^k \right) \delta x_i ds \end{aligned}$$

Les quantités δx_i étant arbitraires, l'expression entre crochets est nulle:

$$-m_0 c \left(\frac{du^i}{ds} \right) + q (\partial^i A^k - \partial^k A^i) u^k = 0 \quad (37.174)$$

Notons:

$$\boxed{F^{ik} = \partial^i A^k - \partial^k A^i} \quad (37.175)$$

Les quantités F^{ik} contravariantes forment les composantes contravariantes de ce que nous appelons le **"tenseur du champ électromagnétique"** ou le **"tenseur de Faraday"** (d'où le $F...$) ou plus couramment le **"tenseur de Maxwell"**. Nous disons alors que F^{ik} est le **"rotationnel du potentiel"**.

Les **"équations du mouvement d'une particule dans un champ électromagnétique"** prennent ainsi la forme:

$$\boxed{m_0 c \frac{du^i}{ds} - q F^{ik} u^k = 0} \quad (37.176)$$

que certains physiciens appellent **"géodésique corrigée par une force de Lorentz"**.

Remarque: Le tenseur du champ électromagnétique est invariant sous les transformations:

$$A^i \rightarrow A^{i'} = A^i - \partial^i \alpha \quad (37.177)$$

Effectivement:

$$\begin{aligned} F^{ik} &= \partial^i (A^k - \partial^k \alpha) - \partial^k (A^i - \partial^i \alpha) \\ &= F^{ik} - \partial^i \partial^k \alpha + \partial^k \partial^i \alpha = F^{ik} \end{aligned} \quad (37.178)$$

Dans une métrique de Minkowski η_{ij} (nous allons avoir besoin du tenseur du champ électromagnétique dans le chapitre de Relativité Restreinte, d'où le choix de cette métrique), nous avons cependant:

$$m_0 c \frac{\eta_{ij} du^i}{ds} - q \eta_{ij} (\partial^i A^k - \partial^k A^i) u^k = 0 \quad (37.179)$$

Ce qui donne:

$$m_0 c \frac{\eta_{ij} du^i}{ds} - q (\partial_j A^k - \partial^k A_j) u^k = 0 \quad (37.180)$$

Le terme $\partial_j A^k - \partial^k A_j$ est souvent noté F^{ik} (même s'il n'est pas plus totalement contravariant).

Il nous reste à déterminer les composantes du tenseur F^{ik} contravariant (tenseur qui a la propriété d'être antisymétrique tel que $F^{ik} = -F^{ki}$).

Commençons par le plus simple. Nous supposons comme évident que:

$$\begin{aligned} F^{00} &= \partial_0 A^0 - \partial^0 A_0 = 0 \\ F^{11} &= \partial_1 A^1 - \partial^1 A_1 = 0 \\ F^{22} &= \partial_2 A^2 - \partial^2 A_2 = 0 \\ F^{33} &= \partial_3 A^3 - \partial^3 A_3 = 0 \end{aligned} \quad (37.181)$$

Ensuite, en se rappelant que $\vec{B} = \vec{\nabla} \times \vec{A}$:

$$\begin{pmatrix} B_x \\ B_y \\ B_z \end{pmatrix} = \begin{pmatrix} \partial_y A_z - \partial_z A_y \\ \partial_z A_x - \partial_x A_z \\ \partial_x A_y - \partial_y A_x \end{pmatrix} = \begin{pmatrix} \partial_2 A_3 - \partial_3 A_2 \\ \partial_3 A_1 - \partial_1 A_3 \\ \partial_1 A_2 - \partial_2 A_1 \end{pmatrix} \quad (37.182)$$

D'où (en choisissant la métrique Minkowski avec la signature +, -, -, -):

$$\begin{aligned}
F^{21} &= -F^{12} = \partial_2 A^1 - \partial^1 A_2 = \frac{\partial}{\partial y} A^1 - \frac{\partial}{\partial x} (-A_2) = \frac{\partial}{\partial x} A_2 - \frac{\partial}{\partial y} A^1 = B_z \\
F^{32} &= -F^{23} = \partial_3 A^2 - \partial^2 A_3 = \frac{\partial}{\partial z} A^2 - \frac{\partial}{\partial y} (-A_3) = \frac{\partial}{\partial y} A_3 - \frac{\partial}{\partial z} A^2 = B_x \\
F^{13} &= -F^{31} = \partial_1 A^3 - \partial^3 A_1 = \frac{\partial}{\partial x} A^3 - \frac{\partial}{\partial z} (-A_1) = \frac{\partial}{\partial z} A_1 - \frac{\partial}{\partial x} A^3 = B_y
\end{aligned} \quad (37.183)$$

Ce qui nous donne pour l'instant:

$$F^{ik} = \begin{bmatrix} F^{00} & F^{01} & F^{02} & F^{03} \\ F^{10} & F^{11} & F^{12} & F^{13} \\ F^{20} & F^{21} & F^{22} & F^{23} \\ F^{30} & F^{31} & F^{32} & F^{33} \end{bmatrix} = \begin{bmatrix} 0 & ? & ? & ? \\ ? & 0 & -B_z & B_y \\ ? & B_z & 0 & -B_x \\ ? & -B_y & B_x & 0 \end{bmatrix} \quad (37.184)$$

Remarque: En toute rigueur pour ne pas confondre le tenseur de Faraday $F^{ik} = \partial^i A^k - \partial^k A^i$ avec sa forme matricielle, nous devrions mettre le premier terme de l'égalité ci-dessus entre crochets aussi comme nous l'avons déjà précisé dans le chapitre de Calcul Tensoriel!

Maintenant, étant connu que $x_0 = ct$ et $A_0 = \phi/c$ les autres composantes du tenseur F^{ik} s'écrivent compte tenu de:

$$\vec{E} = -\vec{\nabla} \cdot \phi - \frac{\partial \vec{A}}{\partial t} \quad (37.185)$$

et donc:

$$\frac{1}{c} \left(\frac{\partial \vec{A}}{\partial t} + \vec{\nabla} \cdot \phi \right) = -\frac{\vec{E}}{c} \quad (37.186)$$

ainsi, avec les dérivées partielles contravariantes selon la métrique de Minkowski:

$$\begin{aligned}
F^{01} &= -F^{10} = \partial_0 A^1 - \partial^1 A_0 = \frac{1}{c} \left(\frac{\partial A_1}{\partial t} - \frac{\partial \phi}{\partial x} \right) = \frac{1}{c} \left(\frac{\partial A_1}{\partial t} + \frac{\partial \phi}{\partial x} \right) = -\frac{E_x}{c} \\
F^{02} &= -F^{20} = \partial_0 A^2 - \partial^2 A_0 = \frac{1}{c} \left(\frac{\partial A_2}{\partial t} - \frac{\partial \phi}{\partial y} \right) = \frac{1}{c} \left(\frac{\partial A_2}{\partial t} + \frac{\partial \phi}{\partial y} \right) = -\frac{E_y}{c} \\
F^{03} &= -F^{30} = \partial_0 A^3 - \partial^3 A_0 = \frac{1}{c} \left(\frac{\partial A_3}{\partial t} - \frac{\partial \phi}{\partial z} \right) = \frac{1}{c} \left(\frac{\partial A_3}{\partial t} + \frac{\partial \phi}{\partial z} \right) = -\frac{E_z}{c}
\end{aligned} \quad (37.187)$$

Ainsi, nous avons pour le tenseur du champ électromagnétique en composantes contravariantes avec et toujours avec la métrique de Minkowski de signature +, -, -, - :

$$F^{ik} = \begin{bmatrix} F^{00} & F^{01} & F^{02} & F^{03} \\ F^{10} & F^{11} & F^{12} & F^{13} \\ F^{20} & F^{21} & F^{22} & F^{23} \\ F^{30} & F^{31} & F^{32} & F^{33} \end{bmatrix} = \begin{bmatrix} 0 & -\frac{E_x}{c} & -\frac{E_y}{c} & -\frac{E_z}{c} \\ \frac{E_x}{c} & 0 & -B_z & B_y \\ \frac{E_y}{c} & B_z & 0 & -B_x \\ \frac{E_z}{c} & -B_y & B_x & 0 \end{bmatrix} \quad (37.188)$$

Ce qui fait que l'équation du mouvement est finalement:

$$m_0 c^2 \frac{du^i}{ds} - q F^{ik} u^k = 0 \quad (37.189)$$

Mais comme nous le verrons dans le chapitre de Relativité Restreinte, le vrai tenseur du champ électromagnétique est défini par (toujours dans la métrique +, -, -, -):

$$F^{\mu\nu} = c F^{ik} = \begin{bmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & -cB_z & cB_y \\ E_y & cB_z & 0 & -cB_x \\ E_z & -cB_y & cB_x & 0 \end{bmatrix} \quad (37.190)$$

afin que les transformées de Lorentz soient conformes.

L'expression sous forme tensorielle du champ électromagnétique met bien en évidence l'unité du champ électromagnétique alors que généralement les champs électrique et magnétique sont considérés séparément en théorie classique.

Mais comme en physique théorique nous travaillons souvent en unités naturelles (c'est un peu la norme...), nous avons alors:

$$F^{\mu\nu} = \begin{bmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & -B_z & B_y \\ E_y & B_z & 0 & -B_x \\ E_z & -B_y & B_x & 0 \end{bmatrix} \quad (37.191)$$

et donc l'équation du mouvement:

$$m_0 \frac{du^\mu}{ds} - q F^{\mu\nu} u^\nu = 0 \quad (37.192)$$

En notant maintenant les composantes de 1 à 4 au lieu de 0 à 3 (c'est plus facile pour les élèves de se repérer dans la matrice) et sans oublier que les dérivées partielles sont covariantes et en adoptant, à nouveau, les unités naturelles telles que $\mu = \varepsilon = 1$ (in extenso $c = 1$), les deux équations de Maxwell avec sources s'écrivent:

$$\begin{aligned}\vec{\nabla} \circ \vec{E} &= \frac{\rho_Q}{\varepsilon} \xrightarrow{\text{unités naturelles}} \vec{\nabla} \circ \vec{E} = \vec{\nabla} \circ \vec{E} = \rho_Q \\ \vec{\nabla} \times \vec{B} - \mu\varepsilon \frac{d\vec{E}}{dt} &= \mu\vec{j} \xrightarrow{\text{unités naturelles}} \vec{\nabla} \times \vec{B} - \frac{d\vec{E}}{dt} = \vec{j}\end{aligned}\quad (37.193)$$

En utilisant le tenseur du champ électromagnétique, il apparaît alors remarquablement que ces deux équations peuvent être écrites sous la forme de l'équation tensorielle condensée suivante:

$$\partial_\mu F^{\nu\mu} = j^\nu \quad (37.194)$$

où j^ν est le "**quadrivecteur courant**" défini par (en unités naturelles!):

$$j^\nu \equiv (j^0, \vec{j}) = (\rho_Q, \vec{j}) \quad (37.195)$$

En utilisant la première définition du tenseur de Faraday (celle où les composantes du champ sont divisées par c) et en prenant pour connu (nous le démontrerons plus tard) que $c = 1/\sqrt{\varepsilon_0\mu_0}$ nous avons dans le système SI:

$$\partial_\mu F^{\nu\mu} = \mu_0 j^\nu \text{ avec } j^\nu = (c\rho_Q, \vec{j}) \quad (37.196)$$

Comme nous allons de suite le voir, la partie temporelle de cette équation donne la divergence du champ électrique et la partie spatiale le rotationnel du champ magnétique.

Remarque: Nous avons déjà rencontré (défini) ce quadrivecteur lors de notre étude de la jauge de Coulomb plus haut ainsi que lors de notre étude de la relativité restreinte (cf. chapitre de Relativité Restreinte).

Effectivement:

$$\begin{aligned}
\partial_{\mu} F^{1\mu} &= \partial_1 F^{11} + \partial_2 F^{12} + \partial_3 F^{13} + \partial_4 F^{14} = \partial_t F^{11} + \partial_x F^{12} + \partial_y F^{13} + \partial_z F^{14} \\
&= \partial_t 0 + \partial_x E_x + \partial_y E_y + \partial_z E_z = \partial_x E_x + \partial_y E_y + \partial_z E_z \\
&\quad \underbrace{\partial_{\mu} F^{1\mu}}_{\vec{\nabla} \circ \vec{E}} \\
\partial_{\mu} F^{2\mu} &= \partial_1 F^{21} + \partial_2 F^{22} + \partial_3 F^{23} + \partial_4 F^{24} = \partial_t F^{21} + \partial_x F^{22} + \partial_y F^{23} + \partial_z F^{24} \\
&= \partial_t (-E_x) + \partial_x 0 + \partial_y (-B_z) + \partial_z (B_y) = \partial_z B_y - \partial_y B_z - \partial_t E_x \\
\partial_{\mu} F^{3\mu} &= \partial_1 F^{31} + \partial_2 F^{32} + \partial_3 F^{33} + \partial_4 F^{34} = \partial_t F^{31} + \partial_x F^{32} + \partial_y F^{33} + \partial_z F^{34} \quad (37.197) \\
&= \partial_t (-E_y) + \partial_x B_z + \partial_y 0 + \partial_z (-B_x) = \partial_x B_z - \partial_z B_x - \partial_t E_y \\
\partial_{\mu} F^{4\mu} &= \partial_1 F^{41} + \partial_2 F^{42} + \partial_3 F^{43} + \partial_4 F^{44} = \partial_t F^{41} + \partial_x F^{42} + \partial_y F^{43} + \partial_z F^{44} \\
&= \partial_t (-E_z) + \partial_x (-B_y) + \partial_y (B_x) + \partial_z 0 = \partial_y B_x - \partial_x B_y - \partial_t E_z \\
&\quad \underbrace{\begin{pmatrix} \partial_{\mu} F^{2\mu} \\ \partial_{\mu} F^{3\mu} \\ \partial_{\mu} F^{4\mu} \end{pmatrix}}_{\vec{\nabla} \times \vec{B} - \partial_t \vec{E}}
\end{aligned}$$

De même, les deux équations de Maxwell:

$$\vec{\nabla} \circ \vec{B} = 0, \vec{\nabla} \times \vec{E} + \frac{\partial \vec{B}}{\partial t} = \vec{0} \quad (37.198)$$

peuvent s'écrire sous la forme condensée tensorielle:

$$\partial_{\mu} F^{\nu\lambda} + \partial_{\nu} F^{\lambda\mu} + \partial_{\lambda} F^{\mu\nu} = 0 \quad (37.199)$$

Effectivement:

$$\begin{aligned}
\partial_2 F^{43} + \partial_4 F^{32} + \partial_3 F^{24} &= \partial_x F^{43} + \partial_z F^{32} + \partial_y F^{24} = \partial_x B_x + \partial_z B_z + \partial_y B_y \\
&= \partial_x B_x + \partial_y B_y + \partial_z B_z \\
&\quad \underbrace{\partial_{\mu} F^{4\mu}}_{\vec{\nabla} \circ \vec{B}} \\
\partial_1 F^{43} + \partial_4 F^{31} + \partial_3 F^{14} &= \partial_t B_x + \partial_z (-E_y) + \partial_y (E_z) = \partial_y E_z - \partial_z E_y + \partial_t B_x \\
\partial_1 F^{24} + \partial_2 F^{41} + \partial_4 F^{12} &= \partial_t B_y + \partial_x (-E_z) + \partial_z (E_x) = \partial_z E_x - \partial_x E_z + \partial_t B_y \quad (37.200) \\
\partial_1 F^{32} + \partial_3 F^{21} + \partial_2 F^{13} &= \partial_t B_z + \partial_y (-E_x) + \partial_x (E_y) = \partial_x E_y - \partial_y E_x + \partial_t B_z \\
&\quad \underbrace{\begin{pmatrix} \partial_1 F^{43} + \partial_4 F^{31} + \partial_3 F^{14} \\ \partial_1 F^{24} + \partial_2 F^{41} + \partial_4 F^{12} \\ \partial_1 F^{32} + \partial_3 F^{21} + \partial_2 F^{13} \end{pmatrix}}_{\vec{\nabla} \times \vec{E} + \partial_t \vec{B}}
\end{aligned}$$

Finalement, toutes les équations de Maxwell, en adoptant les unités naturelles, se résument à:

$$\begin{aligned} \partial_\mu F^{\nu\mu} &= j^\nu \\ \partial_\mu F^{\nu\lambda} + \partial_\nu F^{\lambda\mu} + \partial_\lambda F^{\mu\nu} &= 0 \end{aligned} \quad (37.201)$$

Nous pouvons aussi utiliser un pseudo-tenseur antisymétrique de rang 4 qui peut être vu comme une généralisation du tenseur de Levi-Civita (cf. chapitre de Calcul Tensoriel) tel que nous puissions écrire:

$$\begin{aligned} \partial_\mu F^{\nu\mu} &= j^\nu \\ \varepsilon^{\mu\nu\rho\sigma} \partial_\nu F_{\rho\sigma} &= 0 \end{aligned} \quad (37.202)$$

avec:

$$\varepsilon^{\mu\nu\rho\sigma} = \begin{cases} 1 & \text{si } (\mu\nu\rho\sigma) \text{ est une permutation paire de } (1234) \\ -1 & \text{si } (\mu\nu\rho\sigma) \text{ est une permutation impaire de } (1234) \\ 0 & \text{sinon} \end{cases}$$

Le lagrangien que nous avons déterminé plus haut n'est cependant pas complet. Effectivement, lorsque nous appliquons le principe variationnel, nous avons déjà vu de nombreuses fois dans les différents chapitres de ce site (mécanique classique, mécanique ondulatoire, magnétostatique, relativité restreinte, relativité générale, etc.) que nous pouvions obtenir les équations du mouvement (trajectoires) des sujets (corps) étudiés. Les équations obtenues contenaient aussi des paramètres qui expliquaient la source de ce mouvement (propriétés de la matière, vitesse, champ, etc.) comme cela a été le cas avant!

Précédemment, nous avons appliqué le principe variationnel sur le lagrangien d'interaction charge-champ (magnétique + électrostatique) et avons obtenu l'équation du mouvement corrigée par la force de Laplace.

Lorsque nous avons déterminé les équations du mouvement de la particule chargée à partir du principe de moindre action, nous avons fixé le champ électromagnétique (le champ est connu) et nous avons fait varier la trajectoire. Le principe variationnel, doit alors également nous permettre d'obtenir les équations du champ à partir de la démarche inverse: nous fixons la trajectoire de la particule (trajectoire connue) et nous faisons varier le champ électromagnétique (potentiel et tenseur).

Nous devrions alors obtenir les équations de Maxwell qui, au même titre que l'on obtient ce qui fait le mouvement de la particule lorsque l'on fixe le champ dans le principe variationnel, nous donne l'information sur ce qui est la source du champ électrique et magnétique lorsque l'on fixe la trajectoire dans le principe variationnel (j'espère que vous avez suivi...).

L'envie est alors très grande de reprendre simplement l'expression de l'action obtenue plus haut:

$$S = \int -m_0 c ds - q A^i dx_i \quad (37.203)$$

et de lui appliquer une variation sur le champ après un petit changement dans la manière de l'écrire:

Nous savons que les charges électriques bien qu'elles soient ponctuelles, sont considérées généralement comme une charge transportée par un courant réparti de façon continue dans l'espace. Soit ρ cette densité de charge, nous avons alors $dq = \rho dV$ tel que:

$$S = \int -m_0 c ds - A^i (\rho dV) dx_i \quad (37.204)$$

Considérons des charges électriques se déplaçant à la vitesse v et écrivons la quantité suivante (ne pas oublier que nous continuons à travailler en unités naturelles telles que $dx_0 = c dt = dt$!):

$$dq dx_i = (\rho dV) dx_i = \rho dV dx_0 \frac{dx_i}{dx_0} = \rho d\Omega \frac{dx_i}{dt} = \left(\rho \frac{dx_i}{dt} \right) d\Omega = j_i d\Omega \quad (37.205)$$

avec en unités naturelles: $d\Omega = dx_1 dx_2 dx_3 dx_4 = dt dx dy dz$

Ainsi, nous avons:

$$S = \int -m_0 ds - A^i j_i d\Omega \quad (37.206)$$

Si nous appliquons le principe variationnel seulement sur le champ (constant en amplitude donc la source du champ est constante telle que $\delta j_i = 0$) et que nous considérons donc le mouvement des charges connus, il est immédiat que le premier terme ci-dessus est nul. Nous avons alors:

$$\delta S = \int \delta (A^i j_i) d\Omega = \int \delta A^i j_i d\Omega + A^i \delta j_i d\Omega = \int \delta A^i j_i d\Omega = \int j_i \delta A^i d\Omega = 0 \quad (37.207)$$

pour que cette intégrale soit nulle, il faudrait que j_i soit nul... ce qui est plutôt gênant si nous souhaitons déterminer les caractéristiques d'une source qui alors n'existerait pas... Dès lors, nous remarquons qu'il manque quelque chose à notre lagrangien!

L'idée est alors la suivante: nous connaissons une équation tensorielle qui fait intervenir la densité de courant qui est $\partial_\mu F^{\nu\mu} = j^\nu$ et qui implicitement contient les deux seules équations de Maxwell qui donnent des informations sur la source des champs électrique et magnétique respectifs (les deux autres donnant des propriétés des champs et non pas des sources) soient (toujours en unités naturelles):

$$\begin{aligned} \vec{\nabla} \circ \vec{E} &= \rho \\ \vec{\nabla} \times \vec{B} &= \frac{d\vec{E}}{dt} + \vec{j} \end{aligned} \quad (37.208)$$

Il est donc suffisant d'obtenir ces deux équations (donc l'équation tensorielle y relative) suite au principe variationnel pour avoir les propriétés de la source du champ.

Ce qui signifie simplement que dans l'idéal, nous devrions (et nous attendons à) avoir :

$$\delta S = \int (j^\nu - \partial_\mu F^{\nu\mu}) \delta A^\nu d\Omega = 0 \quad (37.209)$$

où l'intégrale s'annule exactement lorsque $\partial_\mu F^{\nu\mu} = j^\nu$!

Il est alors tentant d'écrire quelque chose de la forme (remarquez que nous avons abaissé l'indice du potentiel A et monté celui de la densité de courant j dans la seconde intégrale ce qui ne change rien mathématiquement parlant au résultat)

$$S = - \int m_0 ds - \int A_\nu j^\nu d\Omega + \int ??? d\Omega \quad (37.210)$$

Nous pouvons nous aider de la propriété suivante des quantités du lagrangien pour déterminer l'expression "???" manquante: elles sont toutes invariantes. En d'autres termes et pour rappel, leur pseudo-norme (scalaire) est égale par changement de référentiel Galiléen (cf. chapitre de Relativité Restreinte) telle que:

$$\begin{aligned} ds &= ds' \\ A_\nu j^\nu &= A'_\nu j'^\nu \end{aligned} \quad (37.211)$$

La première relation est évidente, nous l'avons déjà démontrée de nombreuses fois. La deuxième l'est peut-être moins alors donnons une petite indication (non générale) pour vérifier qu'elle soit correcte: $A_\nu j^\nu$ est le produit scalaire de j et de A . Si nous faisons subir la même (quadri)rotation aux deux vecteurs, puisque les transformations de Lorentz sont des rotations (cf. chapitre de Relativité Restreinte), l'angle entre j et A reste inchangé et donc le produit scalaire.

Il nous faut donc ceci dit, trouver la quantité "???" comme étant un scalaire invariant faisant intervenir le tenseur de Faraday d'une manière ou d'une autre.

Nous pouvons alors essayer directement avec la quantité suivante (sachant d'avance, grâce à nos précurseurs que c'est la bonne hypothèse):

$$F_{\mu\nu} F^{\mu\nu} = \left(g_{\mu\sigma} g_{\nu\rho} F^{\sigma\rho} \right) F^{\mu\nu} = \left(F^{\mu\nu} \right)^2 \quad (37.212)$$

faisant intervenir le tenseur covariant $F_{\mu\nu}$ et contravariant $F^{\mu\nu}$ de Faraday car nous savons que:

1. C'est un scalaire invariant! Effectivement, écrivons $F_{\mu\nu} F^{\mu\nu}$ en termes de champs électrique et magnétique pour en comprendre la signification physique (en unités naturelles):

$$\begin{aligned} & \left(g_{\mu\sigma} g_{\nu\rho} F^{\sigma\rho} \right) F^{\mu\nu} = F_{\mu\nu} F^{\mu\nu} \\ &= \begin{bmatrix} 0 & E_x & E_y & E_z \\ -E_x & 0 & -B_z & B_y \\ -E_y & B_z & 0 & -B_x \\ -E_z & -B_y & B_x & 0 \end{bmatrix} \begin{bmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & -B_z & B_y \\ E_y & B_z & 0 & -B_x \\ E_z & -B_y & B_x & 0 \end{bmatrix} \\ &= F_{1\nu} F^{1\nu} + F_{2\nu} F^{2\nu} + F_{3\nu} F^{3\nu} + F_{4\nu} F^{4\nu} \\ &= -E_x^2 - E_y^2 - E_z^2 - E_x^2 + B_z^2 + B_y^2 - E_y^2 + B_z^2 + B_x^2 - E_z^2 + B_y^2 + B_x^2 \\ &= 2 \left(\vec{B} \circ \vec{B} - \vec{E} \circ \vec{E} \right) = 2 \left(\|\vec{B}\|^2 - \|\vec{E}\|^2 \right) \end{aligned} \quad (37.213)$$

Remarque: Si nous n'étions pas en unités naturelles, le résultat du calcul serait de la forme:

$$F_{\mu\nu} F^{\mu\nu} = \frac{2}{c^2} \left(c^2 \|\vec{B}\|^2 - \|\vec{E}\|^2 \right) \quad (37.214)$$

La quantité $c^2 \|\vec{B}\|^2 - \|\vec{E}\|^2$ (ou $\|\vec{B}\|^2 - \|\vec{E}\|^2$ en unités naturelles) est donc un invariant du champ.

Exemple:

Dans un référentiel O , considérons une onde électromagnétique plane. Les modules du champ électrique et du champ magnétique sont reliés par $E = cB$ (voir plus loin la démonstration). L'invariant du champ considéré est donc nul. Dans un autre référentiel, avec la même structure du champ, nous aurons alors aussi $E' = cB'$.

2. Parce qu'une variationnelle sur ce terme donne:

$$\delta \left(F^{\mu\nu} \right)^2 = \delta \left(F_{\mu\nu} F^{\mu\nu} \right) = F_{\mu\nu} \delta F^{\mu\nu} + \delta F_{\mu\nu} F^{\mu\nu} = 2 \delta F_{\mu\nu} F^{\mu\nu} \quad (37.215)$$

où l'on devine... qu'en creusant un peu, $\delta F_{\mu\nu}$ contient implicitement le terme δA^ν . Nous voyons aussi qu'un facteur 2 apparaît tel qu'il nous faudra introduire une constante de normalisation α , ne serait-ce déjà aussi que pour l'homogénéité des unités de l'expression de l'action.

Donc finalement essayons avec quelque chose du genre:

$$S = - \int m_0 ds - \int A_\mu j^\mu d\Omega + \alpha \int \left(F^{\mu\nu} \right)^2 d\Omega \quad (37.216)$$

À présent, pour chercher les équations du champ électromagnétique, nous considérons que les mouvements des charges sont connus et nous utilisons le principe de moindre action en faisant varier seulement les composantes du potentiel-vecteur et celles du tenseur du champ électromagnétique.

Il en résulte que la variation de la première intégrale est nulle et qu'il reste:

$$\begin{aligned} \delta S &= - \int \delta A_\mu j^\mu d\Omega + \alpha \int \delta \left(F^{\mu\nu} \right)^2 d\Omega \\ &= - \int \delta A_\mu j^\mu d\Omega + \alpha \int \delta \left(F_{\mu\nu} F^{\mu\nu} \right) d\Omega = - \int \delta A_\mu j^\mu d\Omega + 2\alpha \int F^{\mu\nu} \delta F_{\mu\nu} d\Omega \end{aligned} \quad (37.217)$$

Substituons dans la seconde intégrale, les composantes $F_{\mu\nu}$ par leur expression implicite

$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, il vient:

$$F^{\mu\nu} \delta F_{\mu\nu} = F^{\mu\nu} \partial_\mu \delta A_\nu - F^{\mu\nu} \partial_\nu \delta A_\mu \quad (37.218)$$

Or nous savons que $F_{\mu\nu}$ est égal à $-F_{\nu\mu}$ puisque le tenseur de Faraday est antisymétrique:

$$F^{\mu\nu} \delta F_{\mu\nu} = -F^{\mu\nu} \partial_\mu \delta A_\nu - F^{\mu\nu} \partial_\nu \delta A_\mu \quad (37.219)$$

Rien ne nous empêche de permuter les indices μ, ν dans le premier membre à droite de l'égalité:

$$F^{\mu\nu} \delta F_{\mu\nu} = -F^{\mu\nu} \partial_\nu \delta A_\mu - F^{\mu\nu} \partial_\mu \delta A_\nu = -2F^{\mu\nu} \partial_\nu \delta A_\mu \quad (37.220)$$

Donc finalement:

$$\delta S = -\int \delta A_\mu j^\mu d\Omega - 4\alpha \int F^{\mu\nu} \partial_\nu \delta A_\mu d\Omega = 0 \quad (37.221)$$

Intéressons-nous à la seconde intégrale:

$$\int F^{\mu\nu} \partial_\nu \delta A_\mu d\Omega = \int F^{\mu\nu} \frac{\partial}{\partial x^\nu} (\delta A_\mu) dx_0 dx_1 dx_2 dx_3 \quad (37.222)$$

En appliquant le théorème de Fubini (cf. chapitre de Calcul Différentiel Et Intégral) qui dit que l'on peut intégrer selon n'importe quel ordre les variables d'intégration (sous certaines conditions) on peut alors appliquer l'intégration par parties (cf. chapitre de Calcul Différentiel Et Intégral) de manière à écrire:

$$\int F^{\mu\nu} \partial_\nu \delta A_\mu d\Omega = \int F^{\mu\nu} \delta A_\mu dS - \int \partial_\nu F^{\mu\nu} \delta A_\mu d\Omega \quad (37.223)$$

où dS représente la frontière-surface de l'hyper-volume $d\Omega$ sur lequel on intégrait initialement et qui omet la variable prise en considération par le choix de l'indice supérieur ν .

Maintenant selon l'indice supérieur ν concerné, les bornes du premier terme de l'égalité:

$$\int F^{\mu\nu} \delta A_\mu dS \quad (37.224)$$

seront sur les composantes de temps ou les composantes d'espace. Si nous nous concentrons sur les bornes temporelles d'intégration, il s'agit des moments initiaux et finaux de l'action sur laquelle nous appliquons ce variationnel.

Or aux extrémités temporelles, le variationnel du potentiel-vecteur δA_μ est nul (par définition) donc l'intégrale sur la composante de temps sera nulle.

Maintenant sur les composantes spatiales, les bornes (spatiales) sont celles qui permettent d'intégrer la surface-frontière de l'hyper-volume au temps final. Si celui-ci est pris comme l'infini, le rayon de la surface-frontière sera infini et en tout point de cette surface, l'énergie transportée par le champ ainsi que l'amplitude des composantes du champ sera nulle (voir démonstration plus bas).

Donc le variationnel de l'action s'écrit finalement:

$$\delta S = -\int \delta A_\mu j^\mu d\Omega + 4\alpha \int \partial_\nu F^{\mu\nu} \delta A_\mu d\Omega = \int (4\alpha \partial_\nu F^{\mu\nu} - j^\mu) \delta A_\mu d\Omega = 0 \quad (37.225)$$

Les variations du potentiel-vecteur étant arbitraires, l'intégrale précédente sera nulle si l'intégrande elle l'est, d'où la relation

$$4\alpha\partial_\nu F^{\mu\nu} - j^\mu = 0 \quad (37.226)$$

ce qui nous amène à:

$$4\alpha\partial_\nu F^{\mu\nu} = j^\mu \quad (37.227)$$

nous retrouvons donc les deux équations de Maxwell exprimant la source si et seulement si (en unités naturelles):

$$\alpha = 1/4 \quad (37.228)$$

Nous avons donc alors:

$$\partial_\nu F^{\mu\nu} = j^\mu \quad (37.229)$$

Avec finalement pour "**lagrangien total de l'interaction charge-champ**" en unités naturelles:

$$L = -m_0 ds - j^\nu A_\nu + \frac{1}{4} F_{\mu\nu} F^{\mu\nu} \quad (37.230)$$

ou avec le système SI:

$$L = -m_0 c ds - j^\nu A_\nu + \frac{1}{4\mu_0} F_{\mu\nu} F^{\mu\nu} \quad (37.231)$$

Remarque: Nous reviendrons sur ce lagrangien avec une autre approche (très intéressante) dans le chapitre de Physique Quantique Des Champs.

7. ÉQUATIONS D'ONDE ÉLECTROMAGNÉTIQUE

Maxwell supposa que l'onde électromagnétique était une combinaison des phénomènes qu'explicitent les troisième et quatrième équations. Si une onde électromagnétique est éloignée de sa source, nous pouvons alors négliger la densité superficielle de courant de la source comme ayant une influence nulle sur l'onde (nous disons alors que ce sont les équations de Maxwell sans source dont nous avons déjà fait mention plus haut). Alors, les troisième et quatrième équations de Maxwell s'écrivent:

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \quad \text{et} \quad \vec{\nabla} \times \vec{H} = \epsilon \frac{d\vec{E}}{dt} \quad (37.232)$$

Les champs d'excitation magnétique $\vec{H}$ et électrique $\vec{E}$ étant perpendiculaires, plaçons-les de façon commode dans un système d'axes orthogonaux $(\vec{i}, \vec{j}, \vec{k})$ unitaires et euclidiens appartenant à $\mathbb{R}^3$ en choisissant que:

$$\vec{H} = (0, 0, H) = \frac{B}{\mu} \cdot \vec{k} \quad \text{et} \quad \vec{E} = (0, E, 0) \quad (37.233)$$

Remarque: Attention! Il faut bien se rappeler que dans ce qui suit, H est la composante en z de $\vec{H}$ et E la composante en y de $\vec{E}$.

Les calculs (simples) de $\vec{\nabla} \times \vec{E}$ et $\vec{\nabla} \times \vec{H}$ donnent, après simplification:

$$\vec{\nabla} \times \vec{E} = \frac{\partial \vec{E}}{\partial x} = -\frac{\partial \vec{B}}{\partial t} \text{ et } \vec{\nabla} \times \vec{H} = -\frac{\partial H}{\partial x} = \varepsilon \frac{\partial \vec{E}}{\partial t} \quad (37.234)$$

Avant d'aller plus loin, un lecteur nous a demandé de développer les détails qui permettent d'arriver à l'égalité de gauche. Nous partons donc de:

$$\vec{\nabla} \times \vec{E} = \begin{pmatrix} \partial_x \\ \partial_y \\ \partial_z \end{pmatrix} \times \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} = \begin{pmatrix} \partial_x E_z - \partial_z E_y \\ \partial_z E_x - \partial_x E_z \\ \partial_x E_y - \partial_y E_x \end{pmatrix} = \begin{pmatrix} -\partial_z E_y \\ 0 \\ \partial_x E_y \end{pmatrix} = -\frac{\partial \vec{B}}{\partial t} \quad (37.235)$$

Or:

$$\partial_z E_y = 0 \quad (37.236)$$

car l'onde est plane et la composante du champ électrique étant en y , elle ne varie pas selon z . Nous avons alors:

$$\vec{\nabla} \times \vec{E} = \begin{pmatrix} -\partial_z E_y \\ 0 \\ \partial_x E_y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \partial_x E_y \end{pmatrix} = \partial_x \vec{E} = -\frac{\partial \vec{B}}{\partial t} \quad (37.237)$$

Ceci étant fait, si nous continuons, nous avons donc:

$$\frac{\partial E}{\partial x} = -\frac{\partial B}{\partial t} \Rightarrow \frac{\partial^2 E}{\partial x^2} = -\mu \frac{\partial^2 H}{\partial x \partial t} \text{ et } -\frac{\partial H}{\partial x} = \varepsilon \frac{\partial E}{\partial t} \Rightarrow -\frac{\partial^2 H}{\partial x \partial t} = \varepsilon \frac{\partial^2 E}{\partial t^2} \quad (37.238)$$

En identifiant les termes semblables, nous obtenons "l'équation de propagation" du champ électrique:

$$\boxed{\frac{\partial^2 E}{\partial t^2} = \frac{1}{\mu \cdot \varepsilon} \cdot \frac{\partial^2 E}{\partial x^2}} \quad (37.239)$$

et procédant de manière identique:

$$\boxed{\frac{\partial^2 B}{\partial t^2} = \frac{1}{\mu \cdot \varepsilon} \frac{\partial^2 B}{\partial x^2}} \quad (37.240)$$

relations qui sont toutes deux de la forme d'une équation d'onde (cf. chapitre de Mécanique Ondulatoire) du type (rappel) équation de Poisson (plus particulièrement il s'agit d'une équation de d'Alembert):

$$\frac{\partial^2 y}{\partial t^2} = v^2 \cdot \frac{\partial^2 y}{\partial x^2} \quad (37.241)$$

où nous avons:

$$\varepsilon = \varepsilon_0 \cdot \varepsilon_r \left[\text{A} \cdot \text{s} \cdot \text{V}^{-1} \cdot \text{m}^{-1} \right] \text{ et } \mu = \mu_0 \cdot \mu_r \left[\text{V} \cdot \text{s} \cdot \text{A}^{-1} \cdot \text{m}^{-1} \right] \quad (37.242)$$

La vitesse de propagation de l'onde électromagnétique dans le vide est donc:

$$c = \frac{1}{\sqrt{\mu_0 \cdot \varepsilon_0}} = \frac{1}{\sqrt{4\pi \cdot 10^{-7} \cdot 8.8542 \cdot 10^{-12}}} = 2.9979210^8 \left[\text{m} \cdot \text{s}^{-1} \right] \quad (37.243)$$

les unités ainsi que les valeurs numériques concordent...

La vitesse de propagation de l'onde électromagnétique dans la matière est donc:

$$v = \frac{1}{\sqrt{\mu_0 \cdot \mu_r \cdot \varepsilon_0 \cdot \varepsilon_r}} \quad (37.244)$$

$v < c$ car l'expérience montre que nous ne pouvons dépasser la vitesse de la lumière, ce qui est un des postulats de la relativité restreinte et générale.

Donc, nous pouvons finalement écrire:

$$\begin{aligned} \frac{\partial^2 E}{\partial t^2} &= c^2 \frac{\partial^2 E}{\partial x^2} \Leftrightarrow \left(\frac{1}{c^2} \frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2} \right) E = 0 \\ \frac{\partial^2 B}{\partial t^2} &= c^2 \frac{\partial^2 B}{\partial x^2} \Leftrightarrow \left(\frac{1}{c^2} \frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2} \right) B = 0 \end{aligned} \quad (37.245)$$

soit en utilisant le d'Alembertien en une dimension:

$$\begin{cases} \square E = 0 \\ \square B = 0 \end{cases} \quad (37.246)$$

À défaut d'avoir trouvé l'expression directe de $E(x,t)$ et $B(x,t)$, nous venons d'obtenir des équations différentielles ne contenant qu'un seul de ces champs. Nous appelons ces équations respectivement "équation d'onde pour le champ électrique" et "équation d'onde pour le champ d'induction magnétique".

Elles ont la même forme et admettent une solution du même type. Une solution évidente et particulière (nous laissons le soin au lecteur de faire cette vérification) de ces équations différentielles est la fonction trigonométrique sinus:

$$\begin{cases} E(x,t) = \hat{E} \sin(kx - \omega t) \\ B(x,t) = \hat{B} \sin(kx - \omega t) \end{cases} \quad (37.247)$$

en n'oubliant pas la relation entre la pulsation ω , la vitesse de propagation c et le nombre d'onde k que nous avons démontrée dans le chapitre de Mécanique Ondulatoire!

Une solution plus générale est la somme des solutions triviales (cf. chapitre de Calcul Différentiel Et Intégral):

$$\begin{aligned} E(x, t) &= \hat{E} \sin(kx - \omega t) + \hat{E} \cos(kx - \omega t) \\ B(x, t) &= \hat{B} \sin(kx - \omega t) + \hat{B} \cos(kx - \omega t) \end{aligned} \quad (37.248)$$

Mais nous avons vu lors de notre étude des phaseurs (cf. chapitre de Mécanique Ondulatoire) que cette solution réelle n'est qu'un cas particulier d'une solution plus générale et se trouvant dans le corps des complexes. Donc finalement, nous pouvons écrire:

$$\begin{aligned} E(x, t) &= \hat{E} e^{i(kx - \omega t)} \\ B(x, t) &= \hat{B} e^{i(kx - \omega t)} \end{aligned} \quad (37.249)$$

ce qui constitue l'onde plane monochromatique qui est le type d'onde le plus simple à manipuler en physique.

En trois dimensions, la solution est par extension:

$$\begin{aligned} \vec{E}(\vec{r}, t) &= \vec{E}_0 e^{i(\vec{k} \cdot \vec{r} - \omega(\vec{k})t)} \\ \vec{B}(\vec{r}, t) &= \vec{B}_0 e^{i(\vec{k} \cdot \vec{r} - \omega(\vec{k})t)} \end{aligned} \quad (37.250)$$

Remarque: L'onde monochromatique ne peut pas représenter une réalité physique. En effet, si nous calculons l'énergie électrique associée à tout l'espace, nous obtenons pour celle-ci une énergie infinie (car elle n'a ni début, ni fin!) ce qui n'est pas réaliste.

Or, l'équation des ondes est linéaire (solution est toujours la somme d'autres solutions). Donc ceci implique qu'une superposition d'ondes de fréquences différentes (nombre d'onde et pulsation aussi alors!) est également solution. Ainsi, en variant le vecteur d'onde (et implicitement via sa norme, la pulsation, la fréquence et la période) nous balayons également l'ensemble des directions de propagation possibles.

Écrit mathématiquement cela donne, pour le champ électrique:

$$\vec{E}(\vec{r}, t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \vec{E}_0(\vec{k}) e^{i(\vec{k} \cdot \vec{r} - \omega(\vec{k})t)} d^3k \quad (37.251)$$

et rien ne nous empêche de sortir un coefficient de l'amplitude initiale du champ tel que:

$$\vec{E}(\vec{r}, t) = \frac{1}{(2\pi)^{3/2}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \vec{E}_0(\vec{k}) e^{i(\vec{k} \cdot \vec{r} - \omega(\vec{k})t)} d^3k \quad (37.252)$$

et nous retrouvons donc ici une relation très similaire à celle d'une transformée de Fourier inverse (cf. chapitre sur les Suites Et Séries) ce qui est remarquable! Alors l'astuce consiste maintenant à poser $t = 0$ car la relation précédente n'est alors pas qu'une simple analogie avec la transformée de Fourier, c'est une transformée de Fourier!

Nous pouvons donc relier le champ réel $\vec{E}(\vec{r}, 0)$ au champ $\vec{E}_0(\vec{k})$:

$$\begin{aligned}\vec{E}(\vec{r}, 0) &= \frac{1}{(2\pi)^{3/2}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \vec{E}_0(\vec{k}) e^{i\vec{k} \cdot \vec{r}} d^3k \\ \vec{E}_0(\vec{k}) &= \frac{1}{(2\pi)^{3/2}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \vec{E}(\vec{r}, 0) e^{-i\vec{k} \cdot \vec{r}} d^3r\end{aligned}\quad (37.253)$$

Ces deux relations étant souvent condensées sous la forme:

$$\begin{aligned}\vec{E}(\vec{r}, 0) &= \frac{1}{(2\pi)^{3/2}} \int \vec{E}_0(\vec{k}) e^{i\vec{k} \cdot \vec{r}} d^3k \\ \vec{E}_0(\vec{k}) &= \frac{1}{(2\pi)^{3/2}} \int \vec{E}(\vec{r}, 0) e^{-i\vec{k} \cdot \vec{r}} d^3r\end{aligned}\quad (37.254)$$

Le champ réel est donc à l'instant initial la transformée de Fourier inverse du champ $\vec{E}_0(\vec{k})$. Le terme $\vec{E}_0(\vec{k})$ représente donc la composante spectrale liée au vecteur d'onde particulier $\vec{k}$ du champ réel. Cette solution générale de l'équation des ondes s'appelle un **"paquet d'ondes"**

Rappels:

R1. Identiquement à la mécanique ondulatoire (cf. chapitre de Mécanique Ondulatoire), les coefficients ω (pulsation) et k (nombre d'onde) sont exigés pour exprimer la variation du sinus par des radians et pour lui donner une direction et une pulsation.

R2. La périodicité dans le temps de la fonction sinus impose:

$$\sin(kx - \omega(t + T)) = \sin(kx - \omega t - 2\pi) \quad (37.255)$$

d'où la définition de la période de l'onde:

$$T = \frac{2\pi}{\omega} \quad (37.256)$$

R3. La périodicité dans l'espace permet de définir de façon identique la longueur d'onde de la fonction comme:

$$\lambda = \frac{2\pi}{k} \quad (37.257)$$

Nous constatons donc que l'onde plane se déplace selon x en parcourant une distance λ en un temps T . La vitesse de l'onde électromagnétique est alors:

$$v = \frac{\lambda}{T} = c \quad (37.258)$$

En introduisant:

$$\begin{aligned} E(x,t) &= \hat{E} \sin(kx - \omega t) = \hat{E} \sin[k(x - ct)] \\ B(x,t) &= \hat{B} \sin(kx - \omega t) = \hat{B} \sin[k(x - ct)] \end{aligned} \quad (37.259)$$

dans $\frac{\partial \vec{E}}{\partial x} = -\frac{\partial \vec{B}}{\partial t}$ nous obtenons le résultat remarquable pour l'onde plane oscillatoire:

$$\frac{\hat{E}}{\hat{B}} = c \quad (37.260)$$

7.1. ÉQUATION DE HELMHOLTZ

Maintenant, examinons en détail une autre solution de la forme:

$$E_y(x,t) = E_0 \cos(\omega t + \varphi) \quad (37.261)$$

où cette fois-ci, nous faisons explicitement mention des coordonnées afin d'éviter toute confusion.

Remarque: La solution particulière avec le cosinus est plus appréciée par les enseignants que celle avec le sinus, car elle permet comme nous allons le voir, une écriture condensée avec les phaseurs (cf. chapitre de Mécanique Ondulatoire).

Si nous utilisons la notion de phaseur, nous pouvons réécrire cette solution sous la forme:

$$E_y(x,t) = E_0 \cos(\omega t + \varphi) = \Re(E_0 e^{i(\omega t + \varphi)}) = \Re(E_0 e^{i\varphi} e^{i\omega t}) = \Re(\tilde{E}_y e^{i\omega t}) \quad (37.262)$$

Donc:

$$\frac{\partial^2 E_y(x,t)}{\partial t^2} = \frac{\partial^2 \Re(\tilde{E}_y e^{i\omega t})}{\partial t^2} = -\omega^2 \Re(\tilde{E}_y e^{i\omega t}) \quad (37.263)$$

dans l'équation d'onde:

$$\frac{\partial^2 E_y(x,t)}{\partial t^2} = \frac{1}{\mu \cdot \varepsilon} \frac{\partial^2 E_y(x,t)}{\partial x^2} \quad (37.264)$$

nous obtenons:

$$\frac{\partial^2 E_y(x,t)}{\partial x^2} = -\omega^2 \mu \cdot \varepsilon \cdot \Re(\tilde{E}_y e^{i\omega t}) \quad (37.265)$$

qui n'est autre que "**l'équation de Helmholtz**" (pour l'électrodynamique) à une dimension. Il s'agit bêtement de l'équation d'onde écrite d'une manière traditionnelle particulière que nous retrouvons dans de nombreux autres domaines de la physique.

7.2. ÉNERGIE VÉHICULÉE

Il est relativement intuitif que toute onde électromagnétique transporte donc de l'énergie. Exprimons la valeur de cette énergie.

La direction de propagation d'une onde électromagnétique étant celle du vecteur $\vec{E} \times \vec{B}$, nous définissons alors le vecteur de Poynting $\vec{S}$ comme:

$$\vec{S} = \frac{1}{\mu_0} \vec{E} \times \vec{B} \quad (37.266)$$

dont la valeur s'exprime en joules par seconde et par unité de surface: $\left[\frac{J}{s \cdot m^2} \right]$

La norme du vecteur de Poynting représente donc la puissance instantanée qui est transportée par l'onde électromagnétique à travers une surface unitaire, perpendiculaire (nous insistons sur le "perpendiculaire") à sa direction de propagation. Dès lors, nous pouvons aussi écrire le vecteur de Poynting sous la forme (attention à ne pas confondre l'énergie et le champ électrique qui sont représentés par la même lettre):

$$P = \frac{dE}{dt} = \frac{1}{\mu_0} \oint_S \|\vec{E} \times \vec{B}\| dS = \frac{1}{\mu_0} \oint_S (\vec{E} \times \vec{B}) \cdot \vec{n} dS = c^2 \epsilon_0 \oint_S (\vec{E} \times \vec{B}) \cdot \vec{n} dS \quad (37.267)$$

où $\vec{n}$ est comme à l'habitude le vecteur unitaire perpendiculaire à dS (cette dernière relation nous sera utile pour étudier une petite propriété du rayonnement synchrotron).

Pour une onde électromagnétique plane, la norme du vecteur de Poynting vaut:

$$S(x,t) = \frac{E(x,t) \cdot B(x,t) \cdot \sin(\pi/2)}{\mu_0} = \frac{\hat{E} \cdot \hat{B}}{\mu_0} \cdot \sin^2(kx - \omega t) = \frac{\hat{E}^2}{\mu_0 \cdot c} \cdot \sin^2(kx - \omega t) \quad (37.268)$$

Cette grandeur varie en fonction du temps et du lieu. En un endroit donné, sa valeur moyenne est la valeur moyenne du $\sin^2(\cdot)$ pendant une période T :

Rappel:

$$\frac{1}{T} \int_0^T \sin^2(\omega \cdot t) dt = \frac{1}{T} \left[\frac{1}{2} \int_0^T dt - \frac{1}{2} \int_0^T \cos(2\omega \cdot t) dt \right] = \frac{1}{T} \left[\frac{T}{2} - \frac{1}{2} \left[\frac{1}{2\omega} \sin(2\omega \cdot t) \right]_0^T \right] = \frac{1}{2} \quad (37.269)$$

Donc:

$$\forall x \Rightarrow \bar{S} = \frac{\hat{E} \cdot \hat{B}}{2\mu_0} = \frac{\hat{E}^2}{2\mu_0 c} = \frac{c \hat{B}^2}{2\mu_0} \quad (37.270)$$

La valeur moyenne du vecteur de Poynting d'une onde électromagnétique plane est une constante... qui ne dépend ni de la position et du temps.

Remarque: Nous pouvons faire une analogie osée et amusante avec l'électronique en faisant une analyse dimensionnelle du produit $\mu_0 \cdot c$ ci-dessus. Nous avons:

$$\mu_0 \cdot c = \sqrt{\frac{\mu_0}{\epsilon_0}} \cong 377 \text{ } [\Omega] \quad (37.271)$$

...pour démontrer l'énergie contenue dans une unité de volume les physiciens pragmatiques feraient une analyse dimensionnelle. Évitions cela et intéressons-nous toujours au cas particulier de l'onde plane:

Basons-nous sur l'énergie électrique d'une capacité plane idéale productrice d'ondes électromagnétiques planes avec un rendement de 100%:

$$\begin{aligned} E_E &= \frac{1}{2} C \cdot U = \frac{1}{2} \cdot \left(\frac{S \cdot \epsilon}{L} \right) \cdot (\hat{E}^2 \cdot L^2) = \frac{1}{2} \cdot S \cdot L \cdot \epsilon \cdot \hat{E}^2 = \frac{1}{2} \cdot V \cdot \epsilon \cdot \hat{E}^2 \\ \Rightarrow \frac{E_E}{V} &= \frac{1}{2} \cdot \epsilon \cdot \hat{E}^2 \end{aligned} \quad (37.272)$$

et notons la densité volumique d'énergie:

$$u_E = \frac{1}{2} \cdot \epsilon \cdot \hat{E}^2 \quad (37.273)$$

d'où nous tirons que:

$$u_E = \frac{1}{2\mu} \hat{B}^2 \quad (37.274)$$

et l'énergie totale transportée par l'onde électromagnétique dans ce cas particulier est donc:

$$u_{\text{tot}} = \frac{1}{2} \epsilon \hat{E}^2 + \frac{1}{2\mu} \hat{B}^2 = \frac{1}{2} \epsilon \hat{E}^2 + \frac{1}{2\mu c^2} \hat{E}^2 = \epsilon \hat{E}^2 = \frac{1}{\mu} \hat{B}^2 \quad (37.275)$$

Donc la densité d'énergie électrique d'une onde électromagnétique est égale à sa densité d'énergie magnétique.

De par ce résultat, nous sommes amenés à définir "**l'intensité I (moyenne) d'une onde électromagnétique**" par la valeur moyenne de son vecteur de Poynting:

$$\boxed{I = \bar{S}} \quad (37.276)$$

C'est donc la puissance moyenne que transporte l'onde par unité de surface. Or, nous avons démontré plus haut l'expression moyenne du vecteur de Poynting, ce qui nous amène à écrire:

$$I = \bar{S} = \frac{\hat{E} \cdot \hat{B}}{2\mu_0} \quad (37.277)$$

Maintenant, utilisant la relation entre énergie et quantité de mouvement (cf. chapitre de Physique Quantique Ondulatoire):

$$p = \frac{E_{tot}}{c} \quad (37.278)$$

nous obtenons la densité de quantité de mouvement de l'onde électromagnétique:

$$p = \frac{u_{tot}}{c} = \frac{\varepsilon \hat{E}^2}{c} = \frac{\hat{B}^2 c}{\mu} \quad (37.279)$$

Or si la direction de $\vec{E} \times \vec{B}$ est perpendiculaire au front d'onde et est donc confondue avec la direction de propagation de l'onde son module est:

$$\|\vec{E} \times \vec{B}\| = \hat{E} \cdot \hat{B} = \frac{1}{c} \hat{E}^2 = c \hat{B}^2 \quad (37.280)$$

Nous avons donc pour la densité de quantité de mouvement:

$$p = \varepsilon \|\vec{E} \times \vec{B}\| = \frac{1}{\mu c^2} \|\vec{E} \times \vec{B}\| \quad (37.281)$$

Comme la quantité de mouvement doit avoir la direction de la propagation, nous pouvons écrire sous forme vectorielle:

$$\vec{p} = \varepsilon \vec{E} \times \vec{B} = \frac{1}{\mu c^2} \vec{E} \times \vec{B} \quad (37.282)$$

Si une onde électromagnétique possède de la quantité de mouvement, elle possède aussi une densité de moment cinétique. Le moment cinétique par unité de volume est alors:

$$\vec{b} = \vec{r} \times \vec{p} = \varepsilon \vec{r} \times (\vec{E} \times \vec{B}) = \frac{1}{\mu c^2} \vec{r} \times (\vec{E} \times \vec{B}) \quad (37.283)$$

Ainsi, une onde électromagnétique transporte de la quantité de mouvement et du moment cinétique aussi bien que de l'énergie!!!

Ce résultat n'est pas surprenant. Une interaction électromagnétique entre deux charges électriques implique un échange d'énergie et de quantité de mouvement entre les charges. Cela s'effectue par l'intermédiaire du champ électromagnétique qui transporte une densité d'énergie et de quantité de mouvement échangées.

7.3. ÉMISSIONS

Pour prévoir la forme et les propriétés du rayonnement émis par des antennes ou autres sources, il faudrait rigoureusement faire appel à des ordinateurs et aux modèles numériques correspondants au problème à étudier. Formellement, la résolution des équations de Maxwell dans des systèmes macroscopiques est assez difficile et prend du temps. De plus, ceci est plutôt le travail de l'ingénieur qui cherche une exploitation pratique à partir de théories fondamentales. Le physicien théoricien s'intéresse aux fondements de l'Univers et aux systèmes isolés et parfaits.

Cependant, nous souhaiterions exposer la théorie de la diffraction et pour cela, nous devons faire un crochet théorique via une approximation des propriétés du rayonnement d'une source ponctuelle sphérique dans le vide.

L'onde dans le cas d'une source ponctuelle sphérique se propage sphériquement dans l'espace (nous parlons alors "**d'onde sphérique**") et le vecteur de Poynting est radial.

Les vecteurs $\vec{E}(r,t)$ et $\vec{B}(r,t)$ sont localement contenus dans le plan tangent à la sphère de rayon r (c'est logique!) comme le montre la figure ci-dessous:

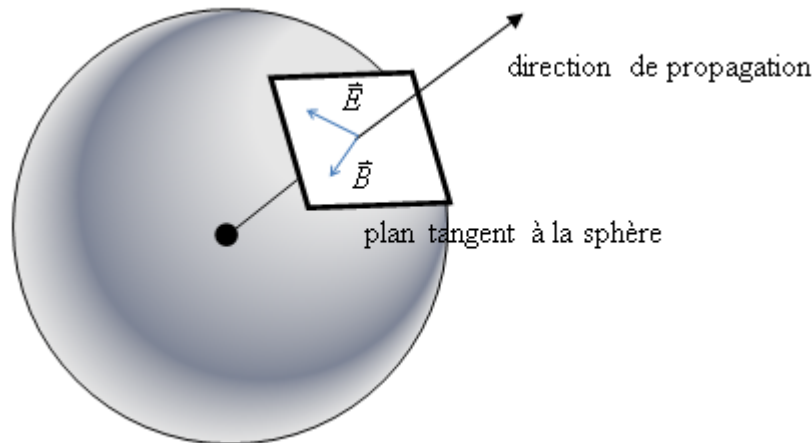


Figure: 37.5 - Représentation de la propagation par rapport au plan tangent à la sphère

Pour que le flux d'énergie soit constant, l'intensité de l'onde doit diminuer avec la distance. En effet, la conservation de l'énergie impose qu'à travers une sphère de rayon r_1 l'énergie E_1 rayonnée par unité de temps (écrite avec un "E" droit afin de ne pas confondre avec la notation du champ électrique) soit égale à celle qui traverse la sphère de rayon r_2 :

$$\frac{dE_1}{dt} = I_1 \cdot 4\pi r_1^2 = \frac{dE_2}{dt} = I_2 \cdot 4\pi r_2^2 \quad (37.284)$$

Ceci implique naturellement:

$$\frac{I_2}{I_1} = \left(\frac{r_1}{r_2} \right)^2 \quad (37.285)$$

Mais en utilisant la relation démontrée plus haut:

$$I = \vec{S} = \frac{\hat{E} \cdot \hat{B}}{2\mu_0} \quad (37.286)$$

et en utilisant la propriété de perpendicularité du champ électrique et magnétique pour une onde plane:

$$I = \|\vec{S}\| = \frac{1}{2\mu_0} \|\vec{E} \times \vec{B}\| = \frac{1}{2} \frac{c \cdot \hat{B}^2}{\mu_0} = \frac{1}{2c} \mu_0 \hat{E}^2 \Rightarrow \frac{I_2}{I_1} = \left(\frac{\hat{B}_2}{\hat{B}_1} \right)^2 = \left(\frac{\hat{E}_2}{\hat{E}_1} \right)^2 = \left(\frac{r_1}{r_2} \right)^2 \quad (37.287)$$

ce qui implique:

$$\hat{E}(r_2) \cdot r_2 = \hat{E}(r_1) \cdot r_1 = c^{te} \Rightarrow \hat{E}(r) \cdot r = c^{te} \Rightarrow \hat{E}(r) = \frac{c^{te}}{r} \quad (37.288)$$

Nous pouvons faire de même pour la composante du champ magnétique.

Donc l'intensité I d'une onde électromagnétique sphérique se propageant dans le vide diminue en $1/r^2$ puisque:

$$\begin{aligned} I &= \frac{1}{2} \frac{c}{\mu_0} \hat{B}(r)^2 = \frac{1}{2} \frac{c}{\mu_0} \frac{c^{te}}{r^2} = \frac{1}{2} \frac{c}{\mu_0} \frac{\hat{B}(r=0)^2}{r^2} = \frac{1}{2} \frac{c}{\mu_0} \frac{\hat{B}_0^2}{r^2} \\ I &= \frac{1}{2} c \mu_0 \hat{E}(r)^2 = \frac{1}{2} c \mu_0 \frac{c^{te}}{r^2} = \frac{1}{2} c \mu_0 \frac{\hat{E}(r=0)^2}{r^2} = \frac{1}{2} c \mu_0 \frac{\hat{E}_0^2}{r^2} \end{aligned} \quad (37.289)$$

et l'amplitude des champs électrique et magnétique diminue en $1/r$. Par extension (information importante pour les téléphones portables), au vu des résultats démontrés précédemment, l'énergie transportée diminue donc en $1/r^2$ puisque:

$$u_E = \frac{1}{2} \cdot \varepsilon \cdot \hat{E}^2 \quad \text{et} \quad u_B = \frac{1}{2\mu} \hat{B}^2 \quad (37.290)$$

Il est facilement compréhensible maintenant d'appréhender pourquoi les physiciens utilisent systématiquement la fréquence pour caractériser une onde, car l'amplitude n'est pas constante dans le vide alors que la fréquence est une sorte de signature de l'émetteur qui ne se perd pas à travers l'espace vide!!!

8. RAYONNEMENT SYNCHROTRON

Considérons une charge en mouvement uniforme rectiligne. Les champs électrique et magnétique d'une telle charge ont été étudiés dans les chapitres précédents. Nous avons également démontré plus haut que le champ magnétique est dans cette configuration, toujours perpendiculaire au champ électrique. La première conséquence est que le champ électrique est radial et le champ magnétique transversal.

Donc si nous entourons la particule en mouvement d'une surface sphérique fermée imaginaire, nous avons alors trivialement (voir la définition du vecteur de Poynting):

$$\frac{dE}{dt} = c^2 \varepsilon_0 \oint_S (\vec{E} \times \vec{B}) \cdot \vec{n} dS = 0 \quad (37.291)$$

puisque effectivement, en tout point de la surface, $\vec{E}$ en est perpendiculaire, $\vec{B}$ tangent, donc $\vec{E} \times \vec{B}$ tangent aussi et donc l'angle entre $\vec{E} \times \vec{B}$ et $\vec{n}$ est égal à un angle droit donc le produit scalaire est nul.

Donc en conclusion le flux total d'énergie rayonnée est nul pour une charge en mouvement rectiligne uniforme. Autrement dit, une charge en mouvement rectiligne uniforme, ne rayonne pas d'énergie électromagnétique mais transporte avec elle l'énergie du champ électromagnétique (nous voilà rassuré!). Ceci est confirmé par les observations expérimentales.

Cependant, la situation est très différente pour une charge en mouvement accéléré. Le champ électrique d'une charge accélérée n'est plus radial et ne possède plus la symétrie par rapport à la charge qu'il possède lorsque le mouvement est uniforme (nous allons le démontrer). Conséquence... une charge électrique accélérée rayonne de l'énergie électromagnétique et donc voit son énergie cinétique diminuer !

Une conclusion importante est qu'il faut, pour maintenir une charge en mouvement accéléré, fournir de l'énergie pour compenser celle perdue par rayonnement. Si la particule au lieu d'être accélérée est décélérée (c'est typiquement ce que nous cherchons à faire en radioprotection) à nouveau la particule va émettre de la même manière le même rayonnement (nous allons aussi le démontrer). C'est ce qui se produit, par exemple, lorsqu'une charge, telle qu'un électron ou un proton, heurte une cible à grande vitesse. Une fraction substantielle de son énergie totale s'en va sous forme d'un rayonnement appelé "**rayonnement de freinage**" ou plus communément "**bremsstrahlung**" (de l'allemand Bremsung: freinage; et Strahlung: rayonnement).

Les équations que nous allons déterminer restent valables pour n'importe quel type de mouvement accéléré relativiste ou non. Par exemple, une particule chargée se déplaçant sur une orbite circulaire est soumise à une accélération centripète et émet donc du rayonnement. Par conséquent, lorsqu'un ion est accéléré dans un accélérateur cyclique, comme un cyclotron, un bêtatron ou un synchrotron, une fraction de l'énergie qui lui est fournie est perdue sous forme de rayonnement électromagnétique, cet effet étant relativement plus important dans les accélérateurs cycliques que dans les accélérateurs linéaires.

Quand les charges atteignent des énergies très élevées, comme cela se produit dans les synchrotrons où l'accélération est grande (heureusement pour nous car cela va nous permettre de faire une petite approximation fort utile...), les pertes dues au rayonnement, appelé "**rayonnement synchrotron**", deviennent importantes et constituent une limitation sérieuse dans la construction d'accélérateurs cycliques de très haute énergie mais restent cependant infiniment utiles à l'industrie de pointe.

Une autre considération importante se rapporte à la structure atomique. Selon le modèle atomique de Rutherford (cf. chapitre de Physique Quantique Corpusculaire), nous imaginons l'atome comme formé d'un noyau central chargé positivement, les électrons chargés négativement décrivant autour de lui des orbites fermées. Mais ceci implique, que les électrons se déplacent suivant un mouvement ayant une accélération et, si nous appliquons les idées développées jusqu'à maintenant, tous les atomes devraient rayonner continuellement de l'énergie (même en l'absence de source d'énergie extérieure comme le Soleil). Par suite de cette perte d'énergie, les orbites électroniques devraient se contracter, amenant à une réduction correspondante de la taille de tous les corps. Heureusement pour nous, cela ne s'observe pas (la matière ne s'effondre pas sur elle-même) mais cela nous amène donc à supposer dans le cadre du modèle de Rutherford que les mouvements des électrons dans les atomes est gouverné par certains principes supplémentaires que nous n'avons pas encore envisagés. C'est ce qui nous amènera à créer le modèle de Bohr de l'atome (cf. chapitre de Physique Quantique Corpusculaire) mais qui aura, lui aussi comme nous le verrons, d'autres défauts.

Pour déterminer l'énergie émise par une charge en mouvement accéléré nous allons devoir faire usage d'outils mathématiques qui ne sont plus du même niveau que ceux utilisés précédemment. Il est donc conseillé que le lecteur ait un bon bagage mathématique. Par ailleurs, exceptionnellement nous ferons usage de logiciels de calculs pour certains points du développement.

Considérons tout d'abord la figure suivante:

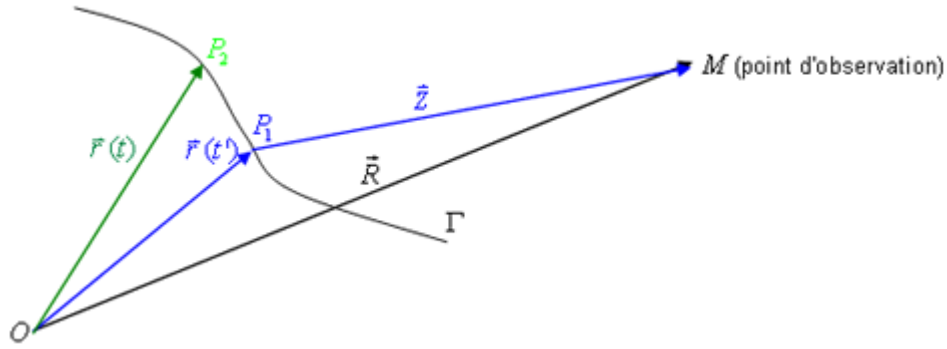


Figure: 37.6 - Scénario à considérer pour l'étude du rayonnement synchrotron

Lorsque la distribution de charges $\rho(\vec{r}, t')$ et la distribution de courant $\vec{J}(\vec{r}, t')$ se trouvent au point P_2 , le point M reçoit l'onde électromagnétique émise par les charges et le courant lorsqu'ils étaient au point P_1 c'est-à-dire à l'instant t' (à cause de la vitesse limite de la propagation du champ dans l'espace). Le retard temporel est la durée de propagation depuis le point P_1 vers le point M , soit:

$$\Delta t = \frac{\|\vec{R} - \vec{r}(t')\|}{c} \quad (37.292)$$

Donc:

$$t = t' + \Delta t \rightarrow t' = t - \Delta t = t - \frac{\|\vec{R} - \vec{r}(t')\|}{c} \quad (37.293)$$

Soit:

$$t' = t - \frac{\|\vec{R} - \vec{r}(t')\|}{c} \quad (37.294)$$

Les potentiels scalaires et vectoriels associés respectivement au champ électrique et magnétique au point de coordonnée vectorielle $\vec{R}$ au temps t ont au vu des résultats obtenus dans les deux chapitres précédents les expressions suivantes:

$$\begin{aligned} U(\vec{R}, t) &= \frac{1}{4\pi\epsilon_0} \oint_V \frac{\rho(\vec{r}, t') dV}{\|\vec{R} - \vec{r}\|} \\ \vec{A}(\vec{R}, t) &= \frac{\mu_0}{4\pi} \oint_V \frac{\vec{J}(\vec{r}, t') dV}{\|\vec{R} - \vec{r}\|} \end{aligned} \quad (37.295)$$

où nous devons par contre de suite démontrer en détail que le potentiel vecteur associé au champ magnétique s'exprime bien ainsi!

Remarque: Nous allons faire usage de ces deux relations de potentiel dans notre étude du champ rayonné car leur forme mathématique similaire nous permettra, du moins nous l'espérons..., de simplifier les développements.

Ces deux relations nous sont déjà partiellement familières, la première qui exprime le potentiel électrique (retardé) a été démontrée dans le chapitre d'Électrostatique dans le cadre non relativiste (donc nos calculs risquent de ne pas être corrects si nous tombons sur un résultat qui dépend de la vitesse ! ... nous verrons bien).

Concernant la deuxième relation qui exprime le potentiel-vecteur retardé, nous avons vu plus haut que $\vec{B} = \vec{\nabla} \times \vec{A}$ était toujours juste au gradient d'une fonction additive près pour $\vec{A}$ (de par les propriétés des opérateurs vectoriels différentiels) tel que:

$$\vec{B} = \vec{\nabla} \times \vec{A} = \vec{\nabla} \times (\vec{A}' + \vec{\nabla} \cdot \psi) = \vec{\nabla} \times \vec{A}' + \underbrace{\vec{\nabla} \times \vec{\nabla} \cdot \psi}_{=0} \quad (37.296)$$

et que $\vec{B}$ soit sous forme relativiste ou non, nous avons:

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{J} + \mu_0 \varepsilon \frac{d\vec{E}}{dt} \quad (37.297)$$

Rappelons aussi (cf. chapitre de Magnétostatique) que:

$$d\vec{B} = \frac{\mu_0}{4\pi} \frac{\vec{J} \times \vec{r}}{r^3} dV = \frac{\mu_0}{4\pi} \frac{\vec{J} \times \vec{r}}{\|\vec{r}\|^3} dV \quad (37.298)$$

Il s'ensuit que si nous posons:

$$d\vec{A} = \frac{\mu_0}{4\pi} \frac{\vec{J} dV}{\|\vec{r}\|} = f(\|\vec{r}\|) \cdot \vec{J} \quad (37.299)$$

que nous retrouvons la loi de Biot-Savart puisque si et seulement si $\vec{J}$ ne dépend pas de r alors (trivial):

$$\vec{\nabla} \times (\vec{J} \cdot f(\|\vec{r}\|)) = (\vec{\nabla} \cdot f(\|\vec{r}\|)) \times \vec{J} \quad (37.300)$$

Nous obtenons donc bien:

$$\begin{aligned} d\vec{B} = \vec{\nabla} \times d\vec{A} &= \vec{\nabla} \times \left(\frac{\mu_0}{4\pi} \frac{\vec{J} dV}{\|\vec{r}\|} \right) = \vec{\nabla} \times (\vec{J} \cdot f(\|\vec{r}\|)) = (\vec{\nabla} \cdot f(\|\vec{r}\|)) \times \vec{J} = \\ &= \frac{\mu_0}{4\pi} \begin{pmatrix} -\frac{dV_x}{\|\vec{r}\|^3} \\ -\frac{dV_y}{\|\vec{r}\|^3} \\ -\frac{dV_z}{\|\vec{r}\|^3} \end{pmatrix} \times \vec{J} = \frac{\mu_0}{4\pi} (-) \vec{J} \times \begin{pmatrix} -\frac{dV_x}{\|\vec{r}\|^3} \\ -\frac{dV_y}{\|\vec{r}\|^3} \\ -\frac{dV_z}{\|\vec{r}\|^3} \end{pmatrix} = \frac{\mu_0}{4\pi} \frac{\vec{J} \times \vec{r}}{\|\vec{r}\|^3} dV \end{aligned} \quad (37.301)$$

Bien que cette forme du potentiel vecteur ne donne que la loi de Biot-Savart sous forme non relativiste, comme elle satisfait toujours:

$$\vec{\nabla} \times \vec{B} = \mu \vec{j} + \mu \varepsilon \frac{d\vec{E}}{dt} \quad (37.302)$$

elle est quand même valable dans le cadre relativiste car cette équation de Maxwell ne dépend pas de la vitesse. De plus, si nos résultats dans l'étude du rayonnement synchrotron nous donnent à la fin une expression indépendante de la vitesse, nous aurons encore une fois confirmé cet état de fait.

8.1. POTENTIELS DE LIÉNARD-WIECHERT

Soit le cas où une particule de masse m et de charge q parcourt une trajectoire Γ . Par rapport à un point origine O , sa coordonnée vectorielle est $\vec{OP} = \vec{h}(t)$, son vecteur vitesse sera noté:

$$\vec{v} = d_t \vec{h}(t) \quad (37.303)$$

et son accélération:

$$\vec{a} = d_t^2 \vec{h}(t) \quad (37.304)$$

Si la charge ponctuelle q se situe à l'origine O , nous avons vu dans le chapitre de Calcul Différentiel Et Intégral que la fonction de Dirac nous donne:

$$\int_{-\infty}^{+\infty} q \delta(x) dx = q \quad (37.305)$$

ainsi que si la charge ponctuelle q se situe à une abscisse x_0 , nous avons:

$$\int_{-\infty}^{+\infty} q \delta(x - x_0) dx = \begin{cases} q, & x = x_0 \\ 0, & x \neq x_0 \end{cases} \quad (37.306)$$

Ce qui vient d'être dit pour un espace à une dimension peut aussi être appliqué à un espace à trois dimensions comme nous l'avons vu et nous écrivons alors:

$$q(\vec{r}, t) = q \delta(\vec{r} - \vec{h}(t)) \quad (37.307)$$

Si nous choisissons pour unités pour la fonction de Dirac des $[m^{-3}]$, alors nous pouvons écrire:

$$\rho(\vec{r}, t) = q \delta(\vec{r} - \vec{h}(t)) \quad (37.308)$$

où q est alors la charge totale au point $\vec{r}$.

Pour la distribution de la densité de courant, nous avons de même toujours en choisissant les mêmes unités que pour la fonction de Dirac:

$$\vec{J}(\vec{r}, t) = q \delta(\vec{r} - \vec{h}(t)) d_t \vec{h}(t) = q \delta(\vec{r} - \vec{h}(t)) \vec{v} \quad (37.309)$$

Dès lors au point M , les potentiels au temps t ont pour expression:

$$\begin{aligned}
 U(\vec{R}, t) &= \frac{1}{4\pi\epsilon_0} \oint \frac{q\delta(\vec{R} - \vec{h}(t'))}{\|\vec{R} - \vec{r}\|} dV \\
 \vec{A}(\vec{R}, t) &= \frac{\mu_0}{4\pi} \oint \frac{q\delta(\vec{R} - \vec{h}(t'))\vec{v}'}{\|\vec{R} - \vec{r}\|} dV
 \end{aligned} \quad (37.310)$$

C'est une formulation bien utile (un détour) qui va nous permettre de résoudre notre problème.

Pour cela, lorsque la charge se trouve au point P_1 au temps t' , nous posons:

$$\vec{h}' = \vec{h}(t') \quad \vec{v}' = \vec{v}(t') \quad \vec{a}' = \vec{a}(t') \quad (37.311)$$

Nous allons utiliser un long artifice afin de résoudre l'intégrale du potentiel électrique (qui est donc une intégrale multiple en coordonnées cartésiennes)!

Celui-ci commence en multipliant le facteur sous l'intégrande de $U(\vec{R}, t)$ par:

$$\delta(t' - t + \|\vec{R} - \vec{r}\|c^{-1}) \quad (37.312)$$

cela ne modifie pas l'intégrale puisque:

$$t' = t - \|\vec{R} - \vec{r}\|c^{-1} \rightarrow t' - t + \|\vec{R} - \vec{r}\|c^{-1} = 0 \quad (37.313)$$

et que (cf. chapitre de Calcul Différentiel Et Intégral):

$$\delta(0) = 1 \quad (37.314)$$

Nous disposons alors de l'expression suivante dans laquelle apparaît le temps t' :

$$\begin{aligned}
 U(\vec{R}, t) &= \frac{1}{4\pi\epsilon_0} \int \delta\left(t' - t + \frac{\|\vec{R} - \vec{r}(t')\|}{c}\right) dt' \int \frac{q\delta(\vec{R} - \vec{h}(t'))}{\|\vec{R} - \vec{r}(t')\|} dV \\
 &= \frac{1}{4\pi\epsilon_0} \int \frac{q\delta(\vec{R} - \vec{h}(t'))}{\|\vec{R} - \vec{r}(t')\|} \delta\left(t' - t + \frac{\|\vec{R} - \vec{r}(t')\|}{c}\right) dx dy dz dt' \\
 &= \frac{1}{4\pi\epsilon_0} \int \frac{q}{\|\vec{R} - \vec{r}(t')\|} \delta(\vec{R} - \vec{h}(t')) \delta\left(t' - t + \frac{\|\vec{R} - \vec{r}(t')\|}{c}\right) dx dy dz dt'
 \end{aligned} \quad (37.315)$$

ce que nous avons le droit d'écrire car la deuxième intégrale ne dépend pas explicitement de t' .

Bon maintenant si nous essayons de résoudre cette intégrale, nous allons y passer notre vie... pour rien. Il va falloir être astucieux.

Avant de rechercher une solution de cette intégrale, nous devons d'abord traiter le cas plus général de l'intégrale suivante:

$$I = \int g[f_1, f_2, f_3, f_4] \delta[f_1(x, y, z, t')] \delta[f_2(x, y, z, t')] \delta[f_3(x, y, z, t')] \delta[f_4(x, y, z, t')] dx dy dz dt' \quad (37.316)$$

Soit écrit de manière plus condensée:

$$I = \int g[f_1, f_2, f_3, f_4] \delta[f_1] \delta[f_2] \delta[f_3] \delta[f_4] dx dy dz dt' \quad (37.317)$$

qu'il est facile de rapprocher avec l'intégrale antérieure:

$$\begin{aligned} f_1 &= \|\vec{R} - \vec{r}(t')\| \\ g[f_1, f_2, f_3, f_4] &= \frac{1}{4\pi\epsilon_0} \frac{q}{\|\vec{R} - \vec{r}(t')\|} = \frac{1}{4\pi\epsilon_0} \frac{q}{f_1} \\ \delta[f_1] &= \delta(R_x - h_x(t')) \\ \delta[f_2] &= \delta(R_y - h_y(t')) \\ \delta[f_3] &= \delta(R_z - h_z(t')) \\ \delta[f_4] &= \delta\left(t' - t + \frac{\|\vec{R} - \vec{r}(t')\|}{c}\right) \end{aligned} \quad (37.318)$$

où nous nous sommes donc arrangés pour que f_1, f_2, f_3, f_4 ne dépendent respectivement (explicitement) que de x, y, z et t' .

Nous souhaitons maintenant faire le changement de variables:

$$x \rightarrow f_1 ; y \rightarrow f_2 ; z \rightarrow f_3 ; t' \rightarrow f_4 \quad (37.319)$$

Nous rappelons que dans des changements de variables dans les intégrales multiples (voir le Jacobien dans le chapitre de Calcul Différentiel Et Intégral), nous avons, en passant des coordonnées cartésiennes aux coordonnées curvilignes les relations suivantes:

$$I = \int_D F(x, y, z) dx dy dz = \int_{D'} G(r_1, r_2, r_3) \left| \frac{\partial(x, y, z)}{\partial(r_1, r_2, r_3)} \right| dr_1 dr_2 dr_3 \quad (37.320)$$

où pour rappel:

$$r_1 = r_1(x, y, z), r_2 = r_2(x, y, z), r_3 = r_3(x, y, z) \quad (37.321)$$

et où:

$$\left| \frac{\partial(x, y, z)}{\partial(r_1, r_2, r_3)} \right| \quad (37.322)$$

n'est pas une valeur absolue mais le déterminant d'une matrice!

Or, dans notre cas traité, rappelons que nous avons tous les f_i qui sont nuls et donc:

$$\delta[f_i]_{f_i=0} = 1 \quad (37.323)$$

et au cas où pendant les développements un des f_i ne le serait plus pour des raisons encore non déterminées, nous aurions:

$$\delta[f_i]_{f_i \neq 0} = 0 \quad (37.324)$$

L'intégrale multiple devient alors:

$$I = \int g[f_1, f_2, f_3, f_4] \left\{ \delta[f_1] \delta[f_2] \delta[f_3] \delta[f_4] \left| \frac{\partial(x, y, z, t')}{\partial(f_1, f_2, f_3, f_4)} \right| \right\}_{f_i=0} df_1 df_2 df_3 df_4 \quad (37.325)$$

où le terme entre accolades est pris à $f_i = 0$ par nécessité de la construction des développements précédents préparant l'artifice mathématique!

Et rappelons encore une fois (!!) la propriété des fonctions de Dirac:

$$\int \varphi(x) \delta(x) dx = [\varphi(x)]_{x=0} \quad (37.326)$$

Nous avons alors immédiatement la simplification:

$$I = g[f_1, f_2, f_3, f_4] \left| \frac{\partial(x, y, z, t')}{\partial(f_1, f_2, f_3, f_4)} \right|_{f_i=0} \quad (37.327)$$

où:

$$\left| \frac{\partial(x, y, z, t')}{\partial(f_1, f_2, f_3, f_4)} \right| \quad (37.328)$$

est donc le Jacobien de la transformation de l'artifice...

Il est évident que par construction du Jacobien, nous avons:

$$\left| \frac{\partial(x, y, z)}{\partial(r_1, r_2, r_3)} \right| \cdot \left| \frac{\partial(r_1, r_2, r_3)}{\partial(x, y, z)} \right| = |\det(J)| \cdot |\det(J)|^{-1} = 1 \quad (37.329)$$

Dès lors il vient:

$$\left| \frac{\partial(x, y, z)}{\partial(r_1, r_2, r_3)} \right| = \left| \frac{\partial(r_1, r_2, r_3)}{\partial(x, y, z)} \right|^{-1} \quad (37.330)$$

Pour l'intégrale I nous avons alors:

$$I = g[f_1, f_2, f_3, f_4] \left| \frac{\partial(f_1, f_2, f_3, f_4)}{\partial(x, y, z, t')} \right|_{f_i=0}^1 \quad (37.331)$$

Calculons donc maintenant notre Jacobien...:

$$|\det(J)|^{-1} = \left| \det \begin{bmatrix} \partial_x f_1 & \partial_x f_2 & \partial_x f_3 & \partial_x f_4 \\ \partial_y f_1 & \partial_y f_2 & \partial_y f_3 & \partial_y f_4 \\ \partial_z f_1 & \partial_z f_2 & \partial_z f_3 & \partial_z f_4 \\ \partial_{t'} f_1 & \partial_{t'} f_2 & \partial_{t'} f_3 & \partial_{t'} f_4 \end{bmatrix} \right|^{-1} \quad (37.332)$$

En revenant au cas traité, $\vec{R} - \vec{h}(t')$ a donc pour composantes:

$$\begin{aligned} f_1 &= R_x - h_x(t') = x - h'_x \\ f_2 &= R_y - h_y(t') = y - h'_y \\ f_3 &= R_z - h_z(t') = z - h'_z \\ f_4 &= t' - t + \frac{\|\vec{R} - \vec{r}(t')\|}{c} = t' - t + \frac{\sqrt{(X-x)^2 + (Y-y)^2 + (Z-z)^2}}{c} \end{aligned} \quad (37.333)$$

Ainsi, nous avons le calcul des éléments de l'inverse du Jacobien:

$$|\det(J)|^{-1} = \left| \det \begin{bmatrix} \partial_x f_1 = 1 & \partial_x f_2 = 0 & \partial_x f_3 = 0 & \partial_x f_4 = -\frac{X-x}{c\|\vec{R} - \vec{r}(t')\|} \\ \partial_y f_1 = 0 & \partial_y f_2 = 1 & \partial_y f_3 = 0 & \partial_y f_4 = -\frac{Y-y}{c\|\vec{R} - \vec{r}(t')\|} \\ \partial_z f_1 = 0 & \partial_z f_2 = 0 & \partial_z f_3 = 1 & \partial_z f_4 = -\frac{Z-z}{c\|\vec{R} - \vec{r}(t')\|} \\ \partial_{t'} f_1 = -h'_x & \partial_{t'} f_2 = -h'_y & \partial_{t'} f_3 = -h'_z & \partial_{t'} f_4 = 1 \end{bmatrix} \right|^{-1} \quad (37.334)$$

Bon maintenant que nous avons les composantes de la matrice Jacobienne, il ne nous reste qu'à calculer son déterminant. Donc soit nous utilisons la relation générale du calcul de déterminant démontrée dans le chapitre d'Algèbre Linéaire, soit nous utilisons Maple... Alors histoire de gagner un peu de temps faisons-le avec Maple 4.00b:

>with(linalg):

> A:= matrix(4,4,[1,0,0,a,0,1,0,b,0,0,1,c,d,e,f,1]);

où:

$$A = \begin{bmatrix} 1 & 0 & 0 & a \\ 0 & 1 & 0 & b \\ 0 & 0 & 1 & c \\ d & e & f & 1 \end{bmatrix} \quad (37.335)$$

avec:

$$\begin{aligned} a &= -\frac{X-x}{c \|\vec{R}-\vec{r}(t')\|} & d &= -\dot{h}_x \\ b &= -\frac{Y-y}{c \|\vec{R}-\vec{r}(t')\|} & e &= -\dot{h}_y \\ c &= -\frac{Z-z}{c \|\vec{R}-\vec{r}(t')\|} & f &= -\dot{h}_z \end{aligned} \quad \text{et} \quad (37.336)$$

Continuons avec Maple 4.00b:

>det (A);

Ce qui donne:

$$1 - cf - eb - da = 1 - (fc + eb + da) \quad (37.337)$$

L'inverse du Jacobien a alors pour expression:

$$\begin{aligned} |\det(J)|^{-1} &= \left| 1 - \left\{ (-\dot{h}_x) \left[-\frac{X-x}{c \|\vec{R}-\vec{r}(t')\|} \right] + (-\dot{h}_y) \left[-\frac{Y-y}{c \|\vec{R}-\vec{r}(t')\|} \right] + (-\dot{h}_z) \left[-\frac{Z-z}{c \|\vec{R}-\vec{r}(t')\|} \right] \right\} \right|^{-1} \\ &= \left| 1 - \left\{ \frac{\dot{h}_x (X-x) + \dot{h}_y (Y-y) + \dot{h}_z (Z-z)}{c \|\vec{R}-\vec{r}(t')\|} \right\} \right|^{-1} = \left| 1 - \left\{ \frac{\vec{h}' \circ (\vec{R}-\vec{r}(t'))}{c \|\vec{R}-\vec{r}(t')\|} \right\} \right|^{-1} = \left| 1 - \left\{ \frac{\vec{v}' \circ (\vec{R}-\vec{r}(t'))}{c \|\vec{R}-\vec{r}(t')\|} \right\} \right|^{-1} \end{aligned} \quad (37.338)$$

où nous avons utilisé le produit scalaire dans la dernière relation afin de condenser l'expression.

Soit:

$$|\det(J)|^{-1} = \frac{1}{\left| 1 - \frac{\vec{v}' \circ (\vec{R}-\vec{r}(t'))}{c \|\vec{R}-\vec{r}(t')\|} \right|} \quad (37.339)$$

L'intégrale multiple:

$$I = g[f_1, f_2, f_3, f_4] \left| \frac{\partial(f_1, f_2, f_3, f_4)}{\partial(x, y, z, t')} \right|_{f_i=0}^{-1} \quad (37.340)$$

où pour rappel:

$$g[f_1, f_2, f_3, f_4] = \frac{1}{4\pi\epsilon_0} \frac{q}{\|\vec{R}-\vec{r}(t')\|} \quad (37.341)$$

soit autrement écrit:

$$g[f_1, f_2, f_3, f_4] = \frac{1}{4\pi\epsilon_0} \frac{q}{\left\| \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} - \begin{pmatrix} x \\ y \\ z \end{pmatrix} \right\|} \quad (37.342)$$

mais suite à notre changement de système de coordonnées nous avons pour rappel:

$$\begin{aligned} & g[f_1, f_2, f_3, f_4] \\ & \rightarrow g\left[\left(f_1 = x - h'_x\right), \left(f_2 = y - h'_y\right), \left(f_3 = z - h'_z\right), \left(f_4 = t' - t + \frac{\|\vec{R} - \vec{r}(t')\|}{c}\right)\right] \end{aligned} \quad (37.343)$$

Or, rappelons encore une fois que:

$$\int \varphi(x) \delta(x) dx = [\varphi(x)]_{x=0} \quad (37.344)$$

Donc il faut prendre g en $f_1 = f_2 = f_3 = f_4 = 0$! Il vient:

$$g\left[\left(x = h'_x\right), \left(y = h'_y\right), \left(z = h'_z\right), \left(t' = t - \frac{\|\vec{R} - \vec{r}(t')\|}{c}\right)\right] \quad (37.345)$$

Ce qui permet d'écrire:

$$g[f_1, f_2, f_3, f_4] = \frac{1}{4\pi\epsilon_0} \frac{q}{\left\| \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} - \begin{pmatrix} h'_x \\ h'_y \\ h'_z \end{pmatrix} \right\|} = \frac{1}{4\pi\epsilon_0} \frac{q}{\|\vec{R} - \vec{h}'\|} \quad (37.346)$$

Il en est de même pour:

$$|\det(J)|^{-1} = \frac{1}{\left|1 - \frac{\vec{v}'}{c} \cdot \frac{(\vec{R} - \vec{r}(t'))}{\|\vec{R} - \vec{r}(t')\|}\right|} \quad (37.347)$$

qui s'écrit alors:

$$|\det(J)|^{-1} = \frac{1}{\left|1 - \frac{\vec{v}'}{c} \cdot \frac{(\vec{R} - \vec{h}')}{\|\vec{R} - \vec{h}'\|}\right|} \quad (37.348)$$

Finalement la résolution de l'intégrale I s'écrit:

$$I = g[f_1, f_2, f_3, f_4] \left| \frac{\partial(f_1, f_2, f_3, f_4)}{\partial(x, y, z, t')} \right|_{f_i=0}^{-1} = g[f_1, f_2, f_3, f_4] \frac{1}{\left| 1 - \frac{\vec{v}'}{c} \cdot \frac{(\vec{R} - \vec{h}')}{\|\vec{R} - \vec{h}'\|} \right|} \quad (37.349)$$

$$= \frac{1}{4\pi\epsilon_o} \frac{q}{\|\vec{R} - \vec{h}'\|} \frac{1}{\left| 1 - \frac{\vec{v}'}{c} \cdot \frac{(\vec{R} - \vec{h}')}{\|\vec{R} - \vec{h}'\|} \right|} = \frac{1}{4\pi\epsilon_o} \frac{q}{\|\vec{R} - \vec{h}'\|} \frac{1}{\left| 1 - \frac{\vec{v}'}{c} \cdot \frac{(\vec{R} - \vec{h}')}{\|\vec{R} - \vec{h}'\|} \right|}$$

On accède ainsi enfin aux expressions des potentiels.

- Le potentiel scalaire s'écrit:

$$U(\vec{R}, t) = \frac{q}{4\pi\epsilon_o} \frac{1}{\|\vec{R} - \vec{h}'\| \left| 1 - \frac{\vec{v}'}{c} \cdot \frac{(\vec{R} - \vec{h}')}{\|\vec{R} - \vec{h}'\|} \right|} \quad (37.350)$$

- Le potentiel-vecteur s'écrit:

$$A(\vec{R}, t) = \frac{\mu_o}{4\pi} \oint \frac{q \delta(\vec{R} - \vec{h}(t')) \vec{v}' dV}{\|\vec{R} - \vec{r}(t')\|} = \frac{\mu_o}{4\pi} \oint \frac{q \delta(\vec{R} - \vec{h}') \vec{v}'}{\|\vec{R} - \vec{r}(t')\|} dx dy dz \quad (37.351)$$

Compte tenu de l'intégrale qui est quasiment la même que pour le potentiel scalaire excepté le terme $\vec{v}'$, nous arrivons en faisant les mêmes développements que précédemment à l'expression:

$$A(\vec{R}, t) = \frac{\mu_o}{4\pi} \frac{q \vec{v}'}{\|\vec{R} - \vec{h}'\| \left| 1 - \frac{\vec{v}'}{c} \cdot \frac{(\vec{R} - \vec{h}')}{\|\vec{R} - \vec{h}'\|} \right|} \quad (37.352)$$

En résumé, les potentiels pris à l'instant (retard temporel de propagation):

$$t' = t - \frac{\|\vec{R} - \vec{r}(t')\|}{c} \quad (37.353)$$

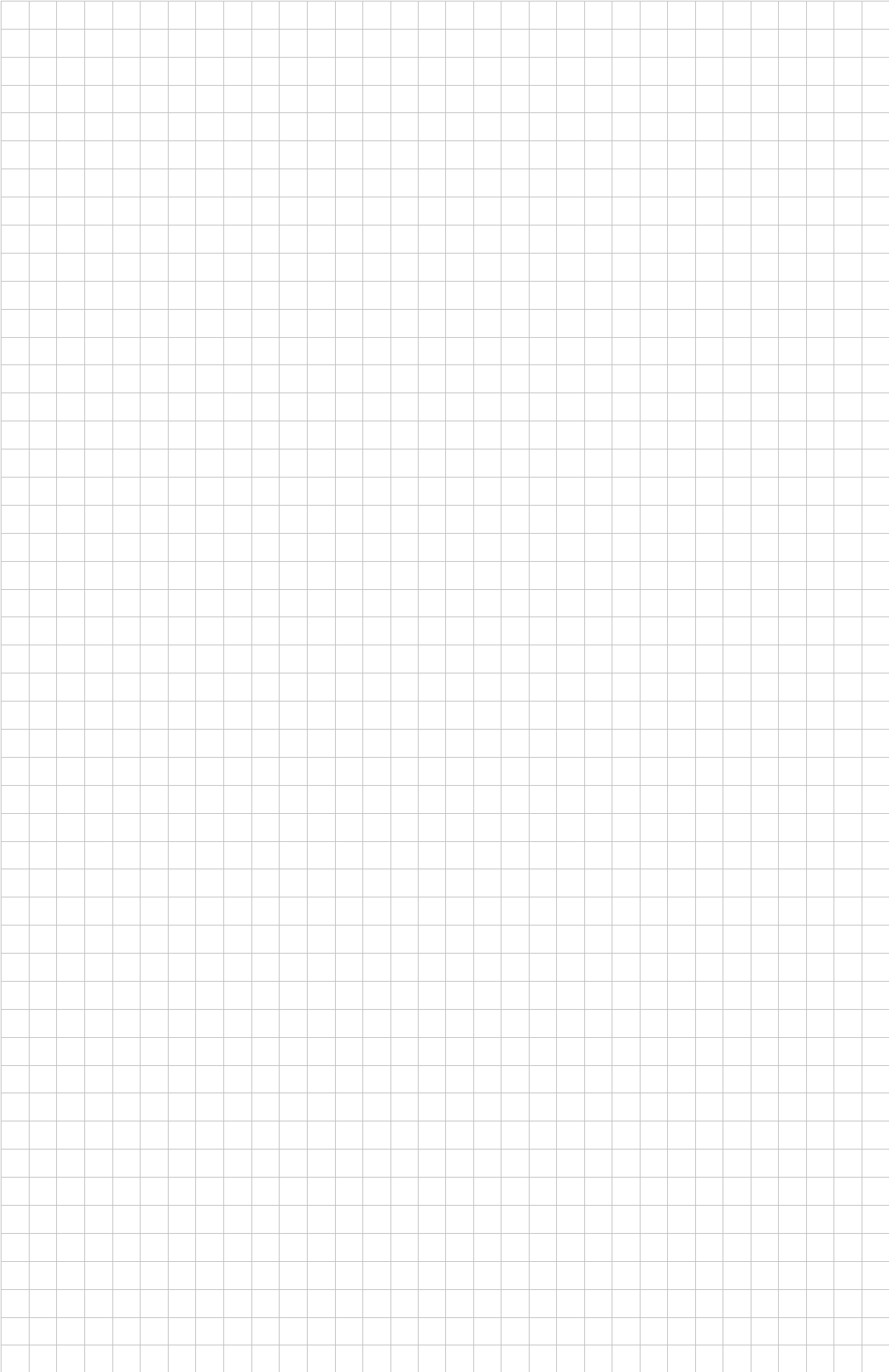
ont pour expressions:

$$U(\vec{R}, t) = \frac{q}{4\pi\epsilon_o} \frac{1}{s} \quad (37.354)$$

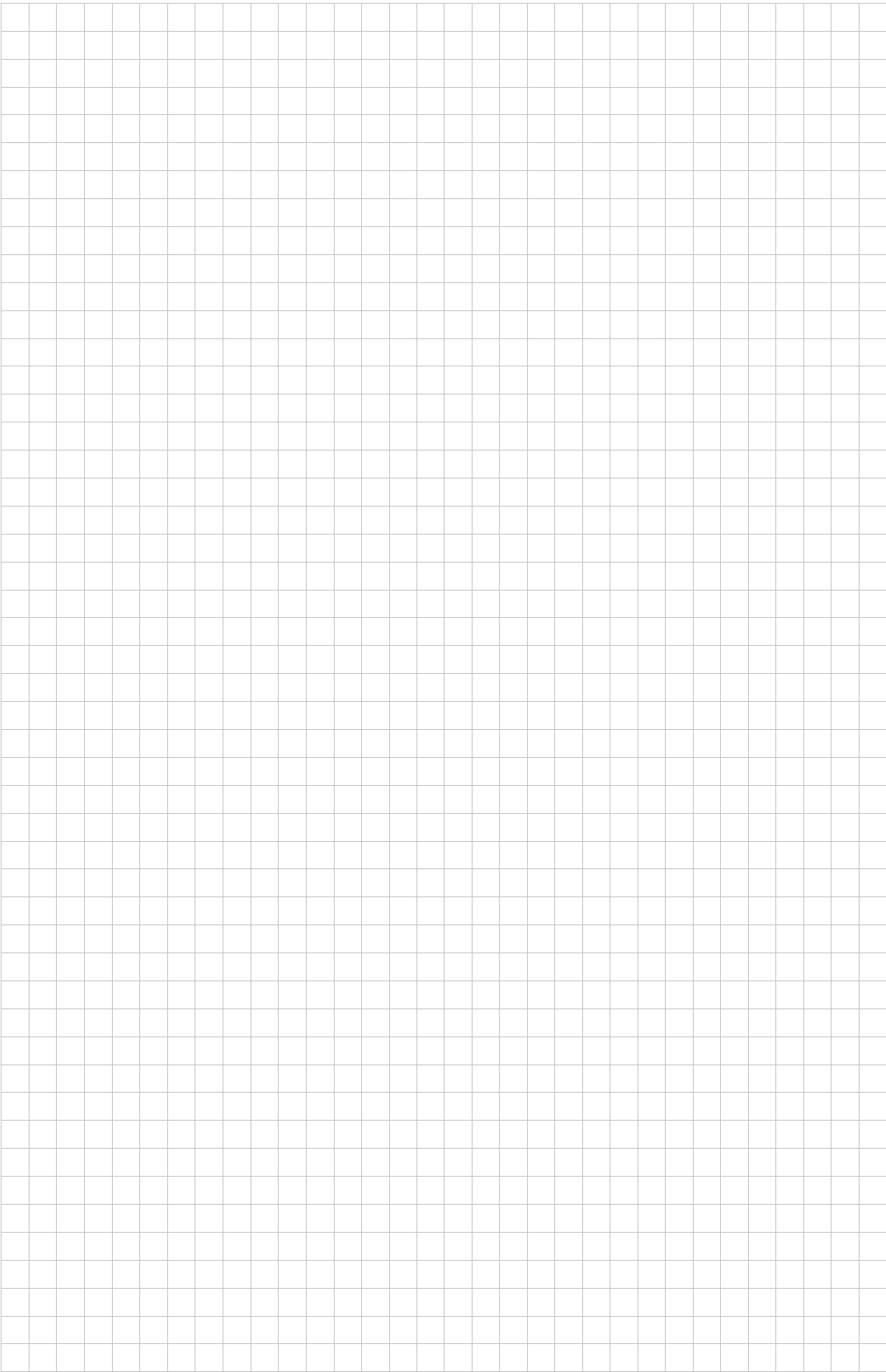
$$A(\vec{R}, t) = \frac{\mu_o}{4\pi} \frac{\vec{v}'}{s}$$

ces potentiels sont appelés "**potentiels de Liénard-Wiechert**" avec:

$$s = \left(\left\| \vec{R} - \vec{h} \right\| \cdot \left| 1 - \frac{\vec{v}'}{c} \cdot \frac{(\vec{R} - \vec{h}')}{\left\| \vec{R} - \vec{h}' \right\|} \right| \right)_{t \mapsto t - \frac{\left\| \vec{R} - \vec{r}(t') \right\|}{c}} \quad (37.355)$$







38. ÉLECTRODYNAMIQUE

Le développement de l'électrodynamique a permis à une grande partie de l'humanité de modifier considérablement sa qualité de vie. Nous savons à peu près tous aujourd'hui ce que nous lui devons: lumière, frigo, radio, télévision, ordinateurs, voitures, trams, trains, avions, robots, et d'autres choses merveilleuses et parfois moins aussi...

Avant de commencer à étudier l'électrocinétique (les ingénieurs parlent "d'électronique" ou "d'électrotechnique") nous allons définir les deux lois (le terme est mal choisi puisque la première est démontrée dans le chapitre d'Électrostatique et la seconde dans le chapitre d'Électrodynamique mais bon... conformons-nous à la tradition...) fondamentales de l'étude de l'électrocinétique et la terminologie de base des circuits ou installations électriques (les cas pratiques étant étudiés dans le chapitre de Génie Électrique). Même si certains éléments au début ne seront peut-être pas compris de suite par le lecteur, ceux-ci deviendront triviaux au fur et à mesure de l'avancement de sa lecture.

Définitions:

D1. Un circuit électrique est constitué d'un ensemble de dispositifs appelés "**dipôles**", reliés entre eux par un fil conducteur.

D2. Un "**noeud**" d'un circuit est une interconnexion où arrivent 3 fils ou plus.

D3. Une "**branche**" est un tronçon de circuit situé entre deux noeuds.

D4. Enfin, une "**maille**" est un ensemble de branches formant une boucle fermée.

Remarque: Un dipôle s'insère dans un circuit par l'intermédiaire de deux pôles, l'un par où s'effectue l'entrée du courant (borne +), l'autre la sortie (borne moins) selon la convention des physiciens (celle des électriciens est l'inverse...).

Le dipôle est caractérisé par la réponse du courant I à une différence de potentiel U entre ses bornes: c'est à dire par la courbe caractéristique:

$$I = f(U) \quad (38.1)$$

Nous verrons que dans tout conducteur, la présence d'une résistivité (voir plus loin) entraîne une chute de tension et, en toute rigueur, il en va de même pour les fils. Mais ceux-ci étant mis en série avec d'autres dipôles, nous négligeons en général dans les petits circuits la résistance des fils devant celle des dipôles présents. Donc, les fils situés entre deux dipôles d'un circuit seront supposés équipotentiels (le potentiel est le même sur les deux bornes).

1. LOIS DE KIRCHHOFF

Les lois de Kirchhoff en électrocinétique (à ne pas confondre avec celles de la thermodynamique et de l'optique) expriment les propriétés physiques de la charge et du champ électrique et sont donc au nombre de deux (une loi pour chaque).

Elles vont nous permettre sans faire appel à l'artillerie mathématique implicitement cachée derrière d'obtenir simplement des résultats forts pertinents.

1.1. LOI DES MAILLES

La loi des mailles (implicitement il s'agit simplement de la conservation de l'énergie) exprime le fait que lorsqu'une charge parcourt un circuit fermé (chemin fermé), l'énergie qu'elle perd en traversant une partie du circuit est égale à l'énergie qu'elle gagne dans l'autre partie. Ainsi, la somme algébrique des potentiels le long d'une maille est nulle telle que:

$$\sum_{\text{maille}} U = 0 \quad (38.2)$$

Pour cela, il faut choisir arbitrairement un sens de parcours de la maille et convenir que les tensions dont la flèche pointe dans le sens du parcours sont comptées comme positives et les autres comme négatives.

Remarque: Cette loi exprime tout simplement le fait que le champ électrique (Coulombien) est un champ conservatif comme nous l'avons vu dans le chapitre d'Électrostatique.

1.2. LOI DES NOEUDS

La loi des noeuds (implicitement il s'agit simplement de la conservation du courant) exprime la conservation de la charge qui signifie que la somme des courants sortant d'un noeud (un noeud peut être vu comme un séparateur de lignes de champ - in extenso des volumes rattachés par une même surface) est égale à la somme des courants entrants. Autrement dit, la somme algébrique des courants est nulle en tout noeud d'un circuit tel que:

$$\sum_{\text{noeud}} I = 0 \quad (38.3)$$

Pour cela, il faut choisir un signe pour les courants entrants et le signe contraire pour les courants sortants (comme nous le faisons en thermodynamique avec la masse).

Remarque: Cette loi exprime tout simplement l'équation de conservation de la charge (ou de continuité de la charge) que nous avons démontrée aussi dans le chapitre d'Électrodynamique.

2. MODÈLE DE DRUDE

Le modèle de Drude de la conduction électrique va nous permettre d'introduire les concepts élémentaires de l'électrocinétique. Dans un premier temps, nous allons définir dans ce qui va suivre les concepts de courant, de densité de courant et ensuite de résistance.

Un conducteur électrique (nous ne parlons pas de semi-conducteurs ou supraconducteurs à ce niveau du discours) peut être vu de manière très simplifiée comme un tuyau de section $\vec{S}$ contenant un gaz d'électrons formé de n charges élémentaires q par unité de volume.

En l'absence de champ électrique, chaque électron possède une vitesse moyenne vectorielle nulle car il reste au voisinage de l'atome. Sous l'action d'un champ électrique $\vec{E}$ homogène et constant (cas du courant continu donc!), certains électrons sont déplacés dans une direction privilégiée, jusqu'à ce qu'ils entrent en collision avec un autre atome (aspect classique) où ils reprennent une vitesse moyenne de dérive nulle et ainsi de suite.

C'est le modèle le plus ancien et le plus élémentaire du courant électrique. Les bases en furent jetées par Drude en 1902, peu après la découverte de l'électron par Thomson (1897). D'où le nom de "**modèle de Drude**".

Insuffisant pour concevoir et a fortiori développer les composants qui forment depuis la fin du 20ème siècle l'essentiel des éléments actifs utilisés en électronique, le modèle des boules de billard présente néanmoins des intérêts considérables:

- C'est un auxiliaire utile pour donner à notre esprit une image de phénomènes dont nous n'avons en fait aucune perception directe, puisqu'ils se déroulent dans l'infiniment petit.
- Les résultats, pour l'ingénieur, de théories plus exactes, comme la théorie des bandes d'énergie en particulier, se laissent formuler au moyen des mêmes concepts que ceux qui apparaissent dans le modèle Boules de billard. Citons parmi ceux-ci le nombre volumique et la mobilité des électrons.
- Tout primitif qu'il soit, ce modèle conduit à une interprétation phénoménologique intéressante des lois fondamentales telles que la loi d'Ohm ou la loi de Joule. Il lie les phénomènes microscopiques à certaines grandeurs observables.

Son nom l'indique, ce modèle assimile les électrons à de minuscules boules de billard. Ces particules sont donc des objets classiques, simplement régis par la loi de Newton et les lois de Maxwell. Cette conception corpusculaire de l'électron n'est d'ailleurs pas totalement opposée aux résultats de la mécanique quantique, dans laquelle un paquet d'ondes, peut toujours être interprété comme une particule, avec sa masse et sa vitesse (voir le théorème d'Ehrenfest dans le chapitre de Physique Quantique Ondulatoire).

Dans un millimètre cube de cuivre, nous admettons que le nombre d'électrons est tellement élevé qu'il n'est donc alors pas question de les traiter individuellement, ce qui serait d'ailleurs sans intérêt. C'est le comportement moyen des électrons qu'il convient d'étudier. Deux types d'interactions conditionnent ce comportement, ce sont:

- l'interaction des électrons avec la matière dans laquelle ils évoluent, et dont ils font partie;
- l'interaction des électrons avec les champs électromagnétiques appliqués de l'extérieur.

La distance λ parcourue par un électron est appelée "**libre parcours moyen de l'électron de conduction**" et si τ est l'intervalle de temps entre deux collisions successives alors nous avons trivialement:

$$\lambda = v \cdot \tau \quad (38.4)$$

Le temps de collision est une variable aléatoire. Tous paramètres physiques restants constants, cette variable aléatoire est stationnaire, sa valeur moyenne porte le nom de "**temps de collision moyen**".

Nous supposons que:

$$v = a \cdot \tau \quad (38.5)$$

la vitesse moyenne, est créée par l'accélération du champ électrique:

$$a = \frac{F}{m} = \frac{(q \cdot E)}{m} \quad (38.6)$$

Nous obtenons alors la "vitesse moyenne de dérive" ou "vitesse d'entraînement" des électrons (drift velocity) donnée par:

$$v = \frac{(q \cdot E)}{m} \cdot \tau \quad (38.7)$$

Cette relation est nommée ainsi, car leur vitesse initiale est due à l'agitation thermique entretenue de l'environnement extérieur et correspond à la vitesse thermique dont nous avons déterminé l'expression lors de notre étude de la distribution de Maxwell-Boltzmann dans le chapitre de Mécanique Statistique (nous en calculerons les valeurs un peu plus bas dans le présent texte).

Nous admettons donc, dans le cadre du modèle Boules de billard, que les électrons se comportent comme les atomes d'un gaz parfait. C'est une hypothèse grossière mais suffisante pour l'instant!

La vitesse moyenne est supposée identique pour tous les électrons libres lorsque le champ électrique appliqué est supposé uniforme, stationnaire, et dirigé selon un seul axe. Elle permet de définir "l'intensité" I du courant électrique dans le conducteur.

Définition: Le "courant" ou "intensité" I mesure la charge $dQ = n \cdot q$ qui traverse la section droite S d'un conducteur par unité de temps dt et est donc donné selon ce qui a été montré juste avant par:

$$I = \frac{dQ}{dt} \quad (38.8)$$

Une tranche de conducteur, de volume $dV = S \cdot dL$ contient donc la charge:

$$dQ = \frac{n \cdot q}{V} \cdot dV = \rho_q S dL \quad (38.9)$$

Elle traverse la section S en un temps dt , tel que:

$$dL = v dt \quad (38.10)$$

Le courant s'écrit alors:

$$I = \frac{dQ}{dt} = \rho_q \cdot S \cdot v \quad (38.11)$$

Si I est vu comme le flux d'une "densité de courant" J à travers la surface S , nous avons alors:

$$I = \oint_S J \cdot dS = J \cdot S \quad (38.12)$$

la densité de courant étant supposée constante sur chaque point de la surface.

Nous avons donc:

$$I = \rho_q \cdot S \cdot v = J \cdot S \quad (38.13)$$

et après simplification:

$$J = \rho_q \cdot v \quad (38.14)$$

qui est donc l'expression de la "densité de courant" dans le conducteur.

Comme nous connaissons l'expression de la vitesse, nous pouvons écrire:

$$J = \frac{\rho_q \cdot q \cdot \tau}{m} \cdot E \quad (38.15)$$

En nous définissons la "conductivité" par:

$$\sigma = \frac{\rho_q \cdot q \cdot \tau}{m} = \frac{n \cdot q^2 \cdot \tau}{m} \quad (38.16)$$

où cette fois n désigne non pas le nombre d'électrons, mais le nombre volumique d'électrons! Par définition, la "résistance" est l'inverse de la conductivité.

Nous remarquons que la conductivité contient le produit du nombre volumique des électrons par leur mobilité. Il faut par conséquent que l'une au moins de ces grandeurs ait une valeur élevée pour qu'un matériau présente une haute conductivité.

La mobilité est plus grande dans les semi-conducteurs que dans les métaux. Cette caractéristique est cependant complètement masquée par le rapport des nombres volumiques des électrons: n est 1'000'000 à 100'000'000 fois plus faible dans les semi-conducteurs que dans les métaux, ce qui explique la conductivité supérieure de ces derniers.

Selon la relation:

$$\sigma = \frac{\rho_q \cdot q \cdot \tau}{m} = \frac{n \cdot q^2 \cdot \tau}{m} \quad (38.17)$$

démontrée juste plus haut, la conductivité dépendrait du champ électrique, par l'intermédiaire du temps de collision. En effet, plus le champ électrique croît, plus la vitesse des électrons augmente. La distance entre les points de chocs possibles restant la même, le temps de collision, et par conséquent la conductivité, devraient diminuer (et donc la résistance augmenter!).

Or, l'indépendance de la conductivité (et respectivement de la résistance) avec le champ électrique est un fait expérimental établi avec précision dans tous les conducteurs habituels dans des conditions normales d'utilisations civiles.

L'origine de cette contradiction réside dans la différence considérable des ordres de grandeur de la vitesse thermique donnée par la distribution de Maxwell-Boltzmann (cf. chapitre de Mécanique Statistique):

$$\langle v_{th} \rangle = \sqrt{\langle v^2 \rangle} = \sqrt{\frac{3kT}{m}} \quad (38.18)$$

et la vitesse moyenne de dérive vue plus haut:

$$v = \frac{(q \cdot E)}{m} \cdot \tau \quad (38.19)$$

avec le temps de libre parcours moyen qui sera obtenu à l'aide de l'expression:

$$\sigma = \frac{\rho_q \cdot q \cdot \tau}{m} = \frac{n \cdot q^2 \cdot \tau}{m} \quad (38.20)$$

Nous avons vu dans le chapitre de Mécanique Statistique que pour un électron à température ambiante:

$$\langle v_{th} \rangle \cong 1.2 \cdot 10^5 \left[\frac{m}{s} \right] \quad (38.21)$$

Et calculons la vitesse de dérive pour le cuivre avec dans ce métal particulier les valeurs suivantes:

$$\sigma = 5.81 \cdot 10^7 \left[\Omega^{-1} \right] \text{ et } n = 1.16 \cdot 10^{29} \left[m^{-3} \right] \quad (38.22)$$

Ce qui nous permet d'obtenir la valeur:

$$\tau = \frac{\sigma \cdot m}{n \cdot q^2} = 1.79 \cdot 10^{-14} \left[s \right] \quad (38.23)$$

et donc:

$$v = \frac{(q \cdot E)}{m} \cdot \tau = \frac{(q \cdot E)}{m} \cdot \frac{\sigma \cdot m}{n \cdot q^2} = \frac{(q \cdot E)}{m} \cdot 1.79 \cdot 10^{-14} \quad (38.24)$$

En prenant $E = 0.32 \left[V \cdot m^{-1} \right]$, ce qui est à considérer comme une valeur élevée puisque ce champ produit une densité de courant de:

$$J = \frac{n \cdot q^2 \cdot \tau}{m} \cdot E = 1.86 \cdot 10^7 \left[A \cdot m^{-2} \right] = 18.6 \left[A \cdot mm^{-2} \right] \quad (38.25)$$

nous avons finalement:

$$v = \frac{(q \cdot E)}{m} \cdot 1.79 \cdot 10^{-14} = 1 \cdot 10^{-3} \left[m \cdot s^{-1} \right] \quad (38.26)$$

Par conséquent, même dans un fort champ électrique industriel, la vitesse de dérive est négligeable par rapport à la vitesse thermique.

Comme la vitesse thermique ne dépend que très peu du champ électrique, il s'avère qu'en pratique la vitesse des électrons est indépendante du champ électrique. En d'autres termes, l'établissement d'un courant, même intense, n'a qu'une incidence absolument négligeable sur la vitesse des électrons!

Remarque: Dans la très grande majorité des cas, les dimensions des conducteurs sont grandes, comparées à la distance moyenne parcourue par un électron entre deux chocs consécutifs. Le comportement des électrons à la surface du conducteur revêt alors une importance secondaire. C'est la raison pour laquelle le milieu conducteur est souvent, implicitement, considéré comme infini. Les transistors FET et MOST constituent à cet égard une exception importante. Le courant y circule dans une couche suffisamment mince pour que la mobilité des électrons soit affectée par la diffusion des électrons aux surfaces délimitant cette couche.

Cependant, un point important à constater est le calcul du libre parcours moyen des électrons dans le modèle classique de Drude. Nous avons effectivement:

$$l \cong v_{th} \cdot \tau \cong 2 \text{ [nm]} \quad (38.27)$$

qui est donc très supérieur, d'au moins un ordre de grandeur (facteur 10), aux distances interatomiques. Il en résulte que les collisions successives sur les atomes du réseau ne sont pas responsables de la loi d'Ohm (que nous allons voir maintenant) contrairement à une des hypothèses de départ du modèle de Drude mais que ce sont les impuretés et les défauts du matériau qui en sont responsables! Nous verrons aussi un peu plus loin qu'avec le modèle théorique des bandes d'énergie le libre parcours moyen est au fait nettement plus grand encore!

Attention!!! Cette relation peut faire penser que puisque le libre parcours moyen est proportionnel à la vitesse thermique et donc proportionnel à la racine carrée de la température, alors la résistance diminue avec la température. Mais en fait il n'en est rien! Le modèle de Drude est trop simpliste car en réalité c'est l'inverse qui a lieu pour les conducteurs (la résistance augmente avec la température parce que l'intervalle de temps τ entre deux collisions diminue plus vite que la vitesse augmente). Et puis il y a aussi le problème inverse... à température presque nulle le libre parcours moyen serait presque nul or les supraconducteurs nous montrent bien qu'il n'en est rien! Bref, sans relation explicite en fonction de la température nous sommes dans l'obscurité la plus totale!

La seule chose que nous savons faire c'est admettre qu'à un α facteur constant près (positif ou négatif), une variation de la température implique une variation relative de la résistance selon:

$$\frac{\Delta R}{R_0} = \alpha \Delta T \quad (38.28)$$

soit:

$$\frac{R - R_0}{R_0} = \alpha (T - T_0) \quad (38.29)$$

d'où la relation connue dans les petites classes:

$$R = R_0 [1 + \alpha (T - T_0)] \quad (38.30)$$

Enfin, précisons que la quatrième équation de Maxwell (cf. chapitre Électrodynamique) peut alors s'écrire au vu des résultats obtenus ci-dessus:

$$\vec{\nabla} \times \vec{H} = \vec{j} + \frac{d\vec{D}}{dt} = \sigma \vec{E} + \frac{d\vec{D}}{dt} \quad (38.31)$$

qui fait alors apparaître explicitement le coefficient de conductivité.

3. LOI D'OHM

À partir de la relation démontrée précédemment:

$$J = \frac{\rho_q \cdot q \cdot \tau}{m} \cdot E \quad (38.32)$$

et en prenant la définition de la "**conductivité**" par:

$$\sigma = \frac{\rho_q \cdot q \cdot \tau}{m} = \frac{n \cdot q^2 \cdot \tau}{m} \quad (38.33)$$

Il vient finalement:

$$\boxed{J = \sigma \cdot E} \quad (38.34)$$

qui est la "**loi locale d'Ohm**". Nous la retrouverons sous forme différentielle dans le chapitre de Mécanique Statistique et nous verrons qu'elle appartient au fait à la famille des lois de diffusion!

Remarque: Puisque la conductivité est nécessairement un scalaire, l'écriture vectorielle de la loi d'Ohm implique que les lignes de champ électrostatiques indiquent également le chemin pris par les charges électriques. Par ailleurs, comme la conductivité est un scalaire nécessairement positif dans le modèle classique, ceci implique que le courant a la même direction que le champ électrique.

Si nous multiplions l'égalité sous forme scalaire à droite et à gauche par L nous obtenons:

$$J \cdot L = \sigma \cdot E \cdot L \Leftrightarrow I = \sigma \cdot \frac{S}{L} \cdot U = G \cdot U \quad (38.35)$$

Donc nous avons:

$$I = G \cdot U \text{ ou } U = \frac{1}{G} \cdot I \quad (38.36)$$

Nous définissons l'inverse de la conductivité comme la "**résistance électrique**" définie par:

$$\boxed{R = \frac{1}{G} = \frac{1}{\left(\sigma \cdot \frac{S}{L}\right)}} \quad (38.37)$$

Remarque: Il est important de remarquer que la résistance électrique est proportionnelle à la longueur de l'élément résistif et inversement proportionnel à sa surface de section. Par exemple dans les câbles hautes tensions la résistance est donnée en Ohm par kilomètre, ce qui permet ensuite de calculer la puissance perdue par kilomètre et donc aussi l'argent perdu par perte Joule.

Dès lors, nous pouvons écrire la loi d'Ohm sous sa forme la plus communément connue:

$$U = R \cdot I \quad (38.38)$$

où donc (attention!!!) le potentiel U représente la différence de potentiel sur la longueur de l'élément résistif (appelé également "**dipôle résistif**") comme nous le voyons dans les développements et non pas le potentiel total extérieur!

Remarque: Cette relation n'est valable que pour des conducteurs idéaux dans des conditions normales de températures et de pression et pour lesquels le modèle de Drude s'applique. Donc les semi-conducteurs et supraconducteurs en sont exclus.

Puisque U est le potentiel de l'élément résistif, nous faisons alors souvent référence dans le domaine de l'électrotechnique à la "**chute de potentiel**" (effectivement, au-delà de l'élément résistif le potentiel n'est plus le même qu'au point qui précède ce même élément résistif).

Pour les câbles en cuivre typiques d'usage non industriel, il existe une table américaine très utile dans la pratique donnant avec une relativement bonne tolérance la résistivité en fonction du diamètre et le courant maximal admissible. Voici un échantillon de cette table:

AWG	Diamètre du fil en mm (avec isolant)	Résistance en Ω par mètre	Courant max. théoriquement admissible à l'air libre en Ampères	Courant max. théoriquement admissible en Ampères
1	7.35	0.0040	211	119
2	6.54	0.0051	181	94
...	...	...	...	...
12	2.05	0.00521	41	9.3
13	1.83	0.00657	35	7.4
14	1.63	0.00829	32	5.9
15	1.45	0.0104	28	4.7
16	1.29	0.0132	22	3.7
...	...	...	...	...

Tableau: 38.1 - Codes AWG (source: Wikipédia)

où AWG signifie "**American Wire Gauge**" et correspond à une petite jauge qu'on peut acheter pour rapidement déterminer le diamètre d'un câble à l'aide de la table ci-dessus sans avoir un pied à coulisse:



Figure: 38.1 - Jauge AWG (source: Wikipédia)

3.1. RÉSISTANCES ÉQUIVALENTES

Nous pouvons maintenant nous intéresser sur toute la longueur d'une ligne de champ électrique parcourue colinéairement par un courant I supposé constant en tout point (c'est une approximation donc...) à la résistance totale si n éléments résistifs R_i sont mis les uns à côtés des autres linéairement.

La réponse est relativement simple puisque si nous notons U_{n-1} le potentiel à la première extrémité de l'élément résistif et U_n l'autre extrémité, nous avons alors (le lecteur remarquera que l'usage de la loi des mailles dans la relation suivante se fait logiquement sans même avoir nécessairement connaissance de celle-ci):

$$U = (U_0 - U_1) + (U_1 - U_2) + \dots + (U_{n-1} - U_n) = R_1 I + R_2 I + \dots + R_n I \quad (38.39)$$

c'est-à-dire un résultat analogue à celui obtenu par une résistance unique dont la valeur est donnée approximativement par (si le courant est constant sur toute la ligne) la "**résistance équivalente de résistances en série**":

$$R_e \cong \sum_{i=1}^n R_i \quad (38.40)$$

qui est la somme arithmétique des résistances individuelles.

Considérons maintenant n résistances en parallèles toutes sous une tension U (de par la loi des mailles) et alimentées par un courant I . Le courant se sépare alors en n courants:

$$I_i = \frac{U}{R_i} \quad (38.41)$$

Dans chacune des n branches. En vertu de la loi des noeuds, nous avons:

$$I_e = \sum_{i=1}^n I_i = \sum_{i=1}^n \frac{U}{R_i} = \frac{U}{R_e} \quad (38.42)$$

c'est-à-dire que l'ensemble des résistances mises en parallèle est analogue à une "**résistance équivalente de résistances en parallèle**":

$$\frac{1}{R_e} = \sum_{i=1}^n \frac{1}{R_i} \Leftrightarrow R_e = \frac{1}{\sum_{i=1}^n \frac{1}{R_i}} \quad (38.43)$$

donnée donc par la moyenne harmonique (cf. chapitre de Statistiques)!

Le fait de brancher des appareils en parallèle permet donc d'avoir toujours la même tension aux bornes de ceux-ci. C'est ainsi que sont disposées par ailleurs les prises électriques dans une installation domestique!

3.2. CAPACITÉS ÉQUIVALENTES

Nous pouvons de même, appliquer le même type de raisonnement aux capacités. Rappelons que nous avons défini dans le chapitre d'Électrostatique, la capacité comme étant donnée par:

$$C = \frac{Q}{U} \quad (38.44)$$

Considérons, au même titre que les résistances, n condensateurs de capacités C_i mis en série les uns derrière les autres. Nous portons aux potentiels U_0 et U_n les deux extrémités de la chaîne et nous apportons la charge Q sur l'ensemble du système. Le potentiel (tension) total aux bornes de la chaîne de condensateurs s'écrit alors simplement:

$$U = U_0 - U_n = (U_0 - U_1) + (U_1 - U_2) + \dots + (U_{n-1} - U_n) = \frac{Q_1}{C_1} + \frac{Q_2}{C_2} + \dots + \frac{Q_n}{C_n} = Q \sum_{i=1}^n \frac{1}{C_i} \quad (38.45)$$

et correspond donc à celle d'une capacité unique C de "**capacité équivalente de capacités en série**":

$$\frac{1}{C_e} = \sum_{i=1}^n \frac{1}{C_i} \Leftrightarrow C_e = \frac{1}{\sum_{i=1}^n \frac{1}{C_i}} \quad (38.46)$$

où nous retrouvons une moyenne harmonique.

Considérons maintenant n condensateurs de capacités C_i mis en parallèle avec le même potentiel U . La charge électrique de chacun d'entre eux est alors imposée (de par la loi des mailles) par la relation:

$$Q_i = C_i U \quad (38.47)$$

La charge électrique totale est simplement:

$$Q = \sum_{i=1}^n Q_i = U \sum_{i=1}^n C_i \quad (38.48)$$

ce qui correspond à une "**capacité équivalente de capacités en parallèle**":

$$C_e = \sum_{i=1}^n C_i \quad (38.49)$$

qui est la somme arithmétique des capacités individuelles.

4. FORCE ÉLECTROMOTRICE

Soit une portion AB d'un circuit, parcourue par un courant permanent I allant de A vers B . L'existence de ce courant implique que le potentiel en A est supérieur (différent) en valeur absolue à celui en B (en valeur absolue). Cette différence de potentiel se traduit par l'existence du champ électrostatique $\vec{E}$ produisant une force de Coulomb:

$$\vec{F} = q\vec{E} \quad (38.50)$$

capable d'accélérer une charge q .

Ainsi, soit:

$$P_q = F \cdot v \quad (38.51)$$

la puissance nécessaire pour communiquer une vitesse v à une particule de charge q quelconque. Sachant que dans ce conducteur il y a ρ_q porteurs de charge par unité de volume, la puissance totale P mise en jeu dans le brin AB parcouru par un courant I est:

$$\begin{aligned} P &= \int_V n_q P_q dV = \int_A^B dl \int_S n_q P_q dS = \int_A^B dl \int_S n_q F \cdot v dS = \int_A^B dl \int_S (n_q \cdot q \cdot E \cdot v) dS \\ &= \int_A^B \int_S (\rho_q v dS) E dl = \int_S J dS \int_A^B E dl = I \cdot \int_A^B E dl = I \cdot (U_A - U_B) = I \cdot \Delta U = I \cdot U \end{aligned} \quad (38.52)$$

c'est-à-dire:

$$\boxed{P = U \cdot I} \quad (38.53)$$

où:

$$U = U_A - U_B > 0 \quad (38.54)$$

Cette puissance est donc la "**puissance électrique**" disponible entre A et B , du simple fait qu'il y circule un courant I .

Si nous considérons dans ce circuit AB une partie résistive pour laquelle nous mesurons une différence de potentiel:

$$U = R \cdot I \quad (38.55)$$

alors la puissance disponible à l'intérieur de celui-ci est donnée par la "**puissance joule**":

$$\boxed{P_J = R \cdot I^2} \quad (38.56)$$

Ainsi, parmi cette puissance disponible, une certaine partie est dissipée sous forme de chaleur (effet Joule) dans un dipôle passif tel que la résistance. Évidemment c'est cette puissance que nous facture notre compagnie d'électricité et connaître l'énergie consommée, il suffit de multiplier la puissance de l'appareil que l'on utilise par la durée de fonctionnement.

Cependant, quelque chose cloche dans nos développements précédents si nous y regardons de plus près. Effectivement, si nous appliquons le raisonnement à un circuit fermé, c'est-à-dire si nous regardons la puissance totale fournie entre A et A par la force de Coulomb, nous obtenons (bien évidemment puisque le champ électrostatique coulombien est conservatif):

$$P = I \cdot \int_A^A \vec{E} d\vec{l} = I \cdot (U_A - U_A) = I \cdot 0 = 0 \quad (38.57)$$

c'est-à-dire une puissance nulle?! Eh oui! Cela signifie qu'il ne peut y avoir de courant en régime permanent dans une boucle fermée et lorsque qu'il y a un courant, alors cela implique que la force de Coulomb n'est pas responsable du mouvement global des porteurs de charge dans un conducteur!!

Dès lors, le courant dans un conducteur peut être compris avec l'analogie de la rivière circulant dans son lit. Pour qu'il y ait un écoulement, il faut que l'eau s'écoule d'une région plus élevée vers une région plus basse (d'un potentiel gravitationnel plus haut vers un autre plus bas). Ainsi, le mouvement de l'eau d'un point élevé vers un point plus bas est bien dû à la simple force de gravitation. Mais si nous voulons constituer un circuit fermé, alors il faut fournir de l'énergie (grâce à une pompe) pour amener l'eau à une plus grande hauteur et le cycle peut alors recommencer.

C'est exactement ce qui se passe dans un circuit électrique. Si nous voulons qu'un courant permanent circule, il faut qu'une autre force que la force électrostatique permette aux charges de fermer le chemin (c'est un raisonnement purement mathématique) ! C'est à ce titre que nous devons faire intervenir une source d'énergie "artificielle" externe tel que le "générateur électrique" qui est alors l'équivalent de la pompe hydraulique pour l'eau.

Le générateur doit alors nous imposer comme propriété physique que lorsque son circuit est ouvert (courant I étant alors nul) une "différence de potentiel" D.D.P. se maintienne entre ses bornes impliquant nécessairement la présence d'une autre force compensant l'attraction coulombienne du conducteur. Ainsi, la force totale s'exerçant sur une charge q s'écrit dès lors:

$$\vec{F} = q (\vec{E}_s + \vec{E}_M) \quad (38.58)$$

avec $\vec{E}_s$ étant le champ électrostatique et $\vec{E}_M$ le "champ électromoteur". À l'équilibre et en l'absence de courant, nous devons avoir:

$$\vec{E}_s + \vec{E}_M = \vec{0} \quad (38.59)$$

Cela signifie que la D.D.P. aux bornes d'un générateur ouvert vaut alors:

$$\vec{E}_s + \vec{E}_M = \vec{0} \Leftrightarrow \vec{E}_s = -\vec{E}_M \Leftrightarrow \int_A^B \vec{E}_s \circ d\vec{l} = - \int_A^B \vec{E}_M \circ d\vec{l} = U_A - U_B \quad (38.60)$$

Nous appelons et notons:

$$\mathcal{E} = \int_A^B \vec{E}_M \circ d\vec{l} = U_B - U_A > 0 \quad (38.61)$$

(un peu maladroitement) la "force électromotrice" FEM propre du générateur.

Puisque, à l'intérieur du générateur, nous avons:

$$\vec{E}_S = -\vec{E}_M \neq \vec{0} \quad (38.62)$$

à circuit ouvert, cela signifie qu'un générateur est un conducteur non-équipotentiel (ou à "champ non conservatif").

À l'équilibre, mais en présence d'un courant I (générateur branché dans un circuit fermé), les porteurs de charge responsables de ce courant subissent une force supplémentaire, due aux collisions se produisant à l'intérieur du conducteur. Pour un générateur idéal, ces collisions sont négligeables et nous obtenons:

$$U_A - U_B = -e \quad (38.63)$$

En revanche, pour un générateur non idéal, de telles collisions se produisent et se traduisent par l'existence d'une résistance interne r (très faible pour les générateurs à l'état neuf!). Ainsi, la vraie force électromotrice est donnée par:

$$(U_A - U_B) + e = r \cdot I \Rightarrow U_A - U_B = r \cdot I - e \quad (38.64)$$

La résistance interne du générateur introduit donc une chute de tension proportionnelle au courant fourni, ce qui fait qu'il délivre un potentiel inférieur à celui donné par sa FEM.

Cette dernière relation est parfois notée sous la forme suivante:

$$U_B - U_A = e - r \cdot I \quad (38.65)$$

et souvent avec l'écriture:

$$U = \mathcal{E} - r \cdot I \quad (38.66)$$

ce que l'on mesure avec un voltmètre est cependant la FEM puisque les générateurs ont une résistance interne admise comme infinie et impliquent donc un courant I quasi nul.

Les générateurs diffèrent selon la source d'énergie utilisée et la méthode de conversion de celle-ci en énergie électrique (autrement dit, selon la nature de $\vec{E}_M$). Nous pouvons ainsi produire de l'énergie électrique à partir d'une pile (énergie chimique), d'un générateur électrostatique (énergie mécanique), d'une dynamo (énergie mécanique), d'une pile solaire (énergie du rayonnement) ou d'un thermocouple (énergie chaleur).

Reprenons le calcul fait précédemment mais appliquons-le cette fois-ci à l'ensemble du circuit. Soit alors V le volume total occupé par le conducteur formant le circuit et $\vec{F}$ la force s'exerçant sur les charges mobiles q et donc responsable de leur mouvement.

La puissance totale P qui doit être fournie en régime permanent est alors:

$$\begin{aligned} P &= \int_V n_q P_q dV = \oint_{\text{circuit}} dl \int_S n_q P_q dS = \oint_{\text{circuit}} dl \int_S n_q \vec{F} \cdot \vec{v} dS = \oint_{\text{circuit}} dl \int_S (n_q \cdot q \cdot \vec{v} \cdot d\vec{S}) \frac{F}{q} \\ &= \oint_{\text{circuit}} \frac{F \cdot dl}{q} \int_S J dS = I \cdot \oint_{\text{circuit}} \frac{F \cdot dl}{q} = I \cdot e \end{aligned} \quad (38.67)$$

où:

$$e = \oint_{\text{circuit}} \frac{F}{q} \cdot dl = \oint_{\text{circuit}} E_M \cdot dl \quad (38.68)$$

est la FEM totale du circuit. L'intégrale portant sur l'ensemble du circuit, la FEM totale est donc la somme des FEM présentes le long du circuit (s'il y en a). Si celles-ci sont localisées dans des dipôles, l'expression devient:

$$e = \sum_k e_k \quad (38.69)$$

où les e_k sont les valeurs algébriques des différentes FEM:

1. $e_k > 0$ correspond à un "générateur" (production d'énergie électrique)
2. $e_k < 0$ correspond à un "récepteur" (consommation d'énergie électrique)

Nous avons aussi pour la puissance électrique:

$$P = I \cdot e = \frac{e}{R+r} \cdot e = \frac{e^2}{R+r} \quad (38.70)$$

et la puissance joule:

$$P_J = R \cdot I^2 = R \cdot \left(\frac{e}{R+r} \right)^2 \quad (38.71)$$

Un moteur convertit de l'énergie électrique en énergie mécanique et correspond donc à un récepteur de FEM: nous disons également, qu'il possède une "force contre-électromotrice" ou FCEM.

4.1. LOI DE FARADAY

Maintenant que nous avons démontré la nécessité de la force électromotrice, nous allons pouvoir démontrer la provenance de la "loi de Faraday" ainsi que la "loi de Lenz" dont nous avons fait usage en électrodynamique pour démontrer la troisième équation de Maxwell. La détermination de la loi de Faraday va également nous permettre de définir le concept d'inductance et d'étudier ses propriétés.

Faisons la même démarche que Faraday et posons-nous la question suivante: Comment crée-t-on un courant?

Un courant est un déplacement de charges dans un matériau conducteur. Ces charges sont mises en mouvement grâce à une D.D.P. qui est maintenue par une FEM. Ainsi, une pile, en convertissant son énergie chimique pendant un instant dt fournit donc une puissance P modifiant l'énergie cinétique des dQ porteurs de charge produisant ainsi un courant I .

Soit P_q la puissance nécessaire pour communiquer une vitesse $\vec{v}$ à une particule de charge q . Sachant que dans un conducteur, il y a n porteurs de charge par unité de volume, la puissance totale P que doit fournir le générateur (idéal) est alors (voir plus haut):

$$P = I \cdot e \quad (38.72)$$

Nous posons donc que la FEM idéale d'un circuit est:

$$e = \frac{P}{I} = \oint_{\text{circuit}} \frac{\vec{F} \cdot d\vec{l}}{q} = \oint_{\text{circuit}} \frac{\vec{F} \circ d\vec{l}}{q} \quad (38.73)$$

Or, la force de Coulomb est incapable de produire une FEM comme nous l'avons démontré tout à l'heure. Pour créer un courant continu dans un circuit fermé, il faut donc un champ électromoteur dont la circulation le long du circuit ne soit pas nulle. L'expérience de Faraday montre donc que c'est l'existence du champ magnétique qui permet l'apparition du courant (!!!). Cela signifie que la force de Lorentz doit être responsable de l'apparition d'une FEM, c'est-à-dire:

$$e = \oint_{\text{circuit}} \frac{q (\vec{E} + \vec{v} \times \vec{B}) \circ d\vec{l}}{q} = \oint_{\text{circuit}} (\vec{E} + \vec{v} \times \vec{B}) \circ d\vec{l} \quad (38.74)$$

Donc:

$$e = \oint_{\text{circuit}} (\vec{E} + \vec{v} \times \vec{B}) \cdot d\vec{l} = \oint_{\text{circuit}} \vec{E} \circ d\vec{l} + \oint_{\text{circuit}} (\vec{v} \times \vec{B}) \circ d\vec{l} = 0 + \oint_{\text{circuit}} (\vec{v} \times \vec{B}) \circ d\vec{l} \quad (38.75)$$

Les propriétés du produit vectoriel (cf. chapitre de Calcul Vectoriel) nous donnant:

$$(\vec{x} \times \vec{y}) \circ \vec{z} = -(\vec{x} \times \vec{z}) \circ \vec{y} \quad (38.76)$$

nous pouvons écrire:

$$e = \oint_{\text{circuit}} (\vec{v} \times \vec{B}) \circ d\vec{l} = - \oint_{\text{circuit}} (\vec{v} \times d\vec{l}) \circ \vec{B} = - \frac{dt}{dt} \oint_{\text{circuit}} (\vec{v} \times d\vec{l}) \circ \vec{B} = - \frac{1}{dt} \oint_{\text{circuit}} (\vec{v} dt \times d\vec{l}) \circ \vec{B} \quad (38.77)$$

Une petite remarque s'impose à ce niveau du discours. Si $\vec{v}$ est bien le vecteur vitesse des charges q il ne peut être celui qui est colinéaire à $d\vec{l}$ car sinon nous aurions:

$$\vec{v} dt \times d\vec{l} = \vec{0} \quad (38.78)$$

et donc e serait nul et ceci n'est pas possible car contredirait tous les développements faits jusqu'à présent ! Au fait, $\vec{v}$ est la vitesse de l'ensemble du circuit qui entraîne avec lui l'ensemble des charges à la même vitesse !

Ainsi, pendant un temps dt , le circuit se déplace d'une distance:

$$d\vec{r} = \vec{v} dt \quad (38.79)$$

vecteur qui est perpendiculaire à $d\vec{l}$. Dès lors:

$$d\vec{r} \times d\vec{l} \quad (38.80)$$

est la surface (voir les propriétés du produit vectoriel dans le chapitre de Calcul Vectoriel) décrite par le déplacement de l'élément $d\vec{l}$ sur la distance $d\vec{r}$ tel que:

$$d\vec{r} \times d\vec{l} = d^2\vec{S} \quad (38.81)$$

Nous avons alors:

$$e = -\frac{1}{dt} \oint_{\text{circuit}} (\vec{v} dt \times d\vec{l}) \circ \vec{B} = -\frac{1}{dt} \oint_{\text{circuit}} \vec{B} \circ d^2\vec{S} \quad (38.82)$$

Nous reconnaissons l'expression du flux (dit "flux coupé") à travers la surface élémentaire $d^2\vec{S}$. Ce qui nous amène à écrire (il y a un petit peu d'intuition - bon sens - avec la manipulation des différentielles mais bon c'est aussi ça la physique...):

$$e = -\frac{1}{dt} \oint_{\text{circuit}} \vec{B} \circ d^2\vec{S} = -\frac{1}{dt} \oint_{\text{circuit}} d^2\Phi_c = -\frac{1}{dt} d \oint_{\text{circuit}} d\Phi_c = -\frac{d\Phi}{dt} \quad (38.83)$$

Nous venons de démontrer la "**loi de Faraday**" dans le cas d'un circuit rigide plongé dans un champ magnétique variable. Nous avons vu apparaître naturellement l'expression du flux coupé. En fait, la seule chose qui compte, c'est l'existence d'un mouvement d'ensemble du tout ou d'une partie du circuit (revoir la démonstration pour s'en convaincre). Ainsi, l'expression de la FEM induite:

$$e = -\frac{d\Phi}{dt} \quad (38.84)$$

reste valable pour un circuit déformé et/ou déplacé dans un champ magnétique statique. Cette démonstration s'est faite à partir de la force de Lorentz et est donc a priori indépendante du référentiel choisi!

4.1.1. LOI DE LENZ

L'énoncé de la loi de Lenz est le suivant: L'induction produit des effets qui s'opposent aux causes qui lui ont donné naissance.

Cette loi est, comme la règle du flux maximum, déjà contenue dans les équations et n'apporte rien de plus, hormis une intuition des phénomènes physiques. En l'occurrence, la loi de Lenz n'est que l'expression du signe "-" contenu dans la loi de Faraday.

Exemple:

Si nous approchons un circuit du pôle nord d'un aimant, le flux augmente et donc la FEM induite est négative. Le courant induit sera alors négatif et produira lui-même un champ magnétique induit opposé à celui de l'aimant. Deux conséquences:

1. L'augmentation du flux à travers le circuit est amoindrie.
2. Il apparaît une force de Laplace (cf. chapitre de Magnétostatique) $\vec{F} = I \cdot d\vec{l} \times \vec{B}$ négative, s'opposant à l'approche de l'aimant.

Ce signe "-" dans la loi de Faraday (la loi de Lenz) décrit le fait que dans des conditions normales, il n'y a pas d'emballlement possible (exemple: courant ne faisant qu'augmenter).

C'est la raison pour laquelle la loi de Lorentz est souvent appelée "**loi de Lenz-Faraday**".

4.2. INDUCTANCE

Nous avons donc:

$$\Phi = \int_S \vec{B} \circ d\vec{S} \quad (38.85)$$

Or la loi de Biot-Savart nous donne (cf. chapitre de Magnétostatique):

$$d\vec{B} = \frac{\mu_0}{4\pi} I \frac{d\vec{l} \times \vec{r}}{r^3} \Leftrightarrow \oint_r d\vec{B} = \frac{\mu_0}{4\pi} I \oint_r \frac{d\vec{l} \times \vec{r}}{r^3} = \vec{B} \quad (38.86)$$

Dès lors:

$$\Phi = \int_S \left(\frac{\mu_0}{4\pi} I \oint_r \frac{d\vec{l} \times \vec{r}}{r^3} \right) \circ d\vec{S} = I \left[\frac{\mu_0}{4\pi} \int_S \left(\oint_r \frac{d\vec{l} \times \vec{r}}{r^3} \right) \circ d\vec{S} \right] \quad (38.87)$$

que nous écrivons historiquement sous forme condensée de la manière suivante:

$$\boxed{\Phi = L \cdot I} \quad (38.88)$$

où L est le "**coefficient d'auto-induction**" ou "**auto-inductance**" (ou "**self**"), exprimé en "Henry" [H]. Il ne dépend que des propriétés géométriques du circuit et est nécessairement positif.

Avec les lois que nous avons énoncées jusqu'à présent, nous sommes en mesure d'étudier certains régimes variables. En effet, tous les raisonnements basés sur la notion d'un champ (électrique ou magnétique) constant au cours du temps peuvent aisément être appliqués à des systèmes physiques variables (champs dépendant du temps), pourvu que cette variabilité s'effectue sur des échelles de temps longues par rapport au temps caractéristique d'ajustement du champ. Voici tout de suite un exemple concret:

La plupart des lois de la magnétostatique supposent un courant permanent, c'est-à-dire le même dans tout le circuit. Lorsque nous fermons un interrupteur, un signal électromagnétique se propage dans tout le circuit et c'est ainsi que peut s'établir un courant permanent: cela prend un temps de l'ordre de l/c où l est la taille du circuit et c la vitesse de la lumière. Si nous avons maintenant un générateur de tension sinusoïdale de période T (c'est juste un exemple... pris au hasard...), alors nous pourrions malgré tout utiliser les relations déduites de la magnétostatique si:

$$T \gg l/c \quad (38.89)$$

Ainsi, bien que le courant soit variable, la création d'un champ magnétique obéira à la loi de Biot-Savart tant que le critère ci-dessus reste satisfait. Ce type de régime variable est appelé "**régime quasi statique**" dans le sens qu'il est transitoire.

Donc, puisque nous avons:

$$\mathcal{E} = -\frac{d\Phi}{dt} \text{ et } \Phi = L \cdot I \quad (38.90)$$

Nous avons alors si et seulement si le courant est variable dans le circuit:

$$e = -\frac{d\Phi}{dt} = -L \frac{dI(t)}{dt} = -L \frac{di}{dt} \quad (38.91)$$

L étant constant pour un circuit rigide. La self ("inductance" en français) crée donc une force électromotrice inverse de celle générée par le courant à ses bornes. Cette force électromotrice a donc un sens inverse à celle du générateur électrique.

Remarque: Nous voyons bien dans la relation obtenue, qu'en régime stationnaire, si le courant est constant, alors la force électromotrice est nulle et la self se comporte alors comme une simple équipotentielle!

Il convient de donner maintenant un exemple important et simple à la fois de la loi de Lenz en l'appliquant au calcul de l'inductance d'un solénoïde de rayon r (l'inductance d'un solénoïde torique à section circulaire ayant déjà été mise en évidence dans le chapitre de Magnétostatique). Nous avons vu dans le chapitre de Magnétostatique que le champ magnétique dans un solénoïde était donné par:

$$B = \mu_0 \cdot \frac{I \cdot N}{l} \quad (38.92)$$

où pour rappel N est le nombre de spires et l est la longueur du solénoïde. Nous avons vu plus haut que la loi de Faraday était donnée par:

$$e = -\frac{1}{dt} d \oint_{\text{circuit}} d\Phi_c = -\frac{d\Phi}{dt} = -L \frac{di}{dt} \quad (38.93)$$

et dans le cas d'une spire nous allons parcourir N fois le chemin de l'intégrale. Il vient alors:

$$e = -N \frac{d\Phi}{dt} = -L \frac{di}{dt} \quad (38.94)$$

Nous avons vu plus haut que le flux du champ magnétique était donné par (si le champ est perpendiculaire à la surface traversée):

$$\Phi = B \cdot S \quad (38.95)$$

Dès lors:

$$\Phi = B \cdot S = \left(\mu_0 \cdot \frac{I \cdot N}{l} \right) \cdot S = \frac{\mu_0 \cdot S \cdot N}{l} \cdot I \quad (38.96)$$

Remarque: Attention!! Le flux dans un solénoïde n'est pas égal au flux dans une spire multiplié par le nombre de spires.

Le taux de variation du flux magnétique se trouve par dérivation, soit:

$$N \frac{d\Phi}{dt} = N \frac{d}{dt} \left(\frac{\mu_0 \cdot S \cdot N}{l} \cdot I \right) = \frac{\mu_0 \cdot S \cdot N^2}{l} \frac{dI(t)}{dt} = \frac{\mu_0 \cdot S \cdot N^2}{l} \frac{di}{dt} \quad (38.97)$$

Soit dans le cas de spires circulaires:

$$N \frac{d\Phi}{dt} = \frac{\mu_0 \cdot S \cdot N^2}{l} \frac{di}{dt} = \frac{\mu_0 \cdot 2\pi r^2 \cdot N^2}{l} \frac{di}{dt} \quad (38.98)$$

La force électromotrice engendrée est ainsi:

$$\mathcal{e} = -N \frac{d\Phi}{dt} = -L \frac{di}{dt} = - \left(\frac{\mu_0 \cdot 2\pi r^2 \cdot N^2}{l} \right) \frac{di}{dt} \quad (38.99)$$

et donc par correspondance:

$$L = \frac{\mu_0 2\pi r^2 N^2}{l} \cong 4\pi^2 \cdot 10^{-7} \frac{N^2 r^2}{l} \quad (38.100)$$

Calculons maintenant la puissance reçue par une bobine. Nous avons démontré plus haut que nous avons toujours dans notre cas d'étude et si nous modélisons l'inductance comme un dipôle non idéal:

$$p = \Delta u \cdot i = u \cdot i = (r \cdot i - \mathcal{e}) \cdot i = r \cdot i^2 - \mathcal{e} \cdot i \quad (38.101)$$

où les lettres en minuscules indiquent que nous sommes en régime non constant:

$$p = r \cdot i^2 - \mathcal{e} \cdot i = r \cdot i^2 - \left(-L \frac{di}{dt} \right) \cdot i = r \cdot i^2 + L \frac{di}{dt} i = P_J + \frac{d \left(\frac{1}{2} L i^2 \right)}{dt} \quad (38.102)$$

Contrairement au développement que nous avons fait dans le chapitre d'Électrostatique pour le même calcul en ce qui concerne la capacité, nous n'avons pas négligé ici la dissipation d'énergie par effet Joule. Mais il faut savoir que dans la majorité des cas ce terme est aussi négligé!

Donc par intégration dans un intervalle de temps donné de 0 à t nous avons pour le deuxième terme:

$$E_L = \frac{1}{2} L i^2 \quad (38.103)$$

Soit explicitement dans le cas d'un solénoïde circulaire:

$$E_L = \frac{1}{2} L i^2 \cong \frac{1}{2} \left(4\pi \cdot 10^{-7} \frac{N^2 r^2}{l} \right) i^2 \quad (38.104)$$

Lorsque i décroît, la bobine restitue cette énergie. Nous ne pouvons donc pas stocker de l'énergie dans une bobine isolée contrairement à un condensateur.

Dans le cadre d'un régime sinusoïdal, la puissance moyenne sera nulle. Nous pouvons généraliser ceci en admettant qu'une inductance parfaite ne dissipe aucune puissance par effet Joule.

5. EFFET DE PEAU

L'effet de peau ou effet pelliculaire (ou plus rarement effet Kelvin) est un phénomène électromagnétique qui fait que, à fréquence élevée, le courant a tendance à ne circuler qu'en surface des conducteurs. Ce phénomène d'origine électromagnétique existe pour tous les conducteurs parcourus par des courants alternatifs. Il provoque la décroissance de la densité de courant à mesure que l'on s'éloigne de la périphérie du conducteur. Il en résulte une augmentation de la résistance du conducteur.

Cet effet peut être utilisé pour alléger le poids des lignes de transmission à haute fréquence en utilisant des conducteurs tubulaires, ou même des tuyaux, sans perte de courant. Il sert aussi dans le blindage électromagnétique des fils coaxiaux en les entourant d'un mince étui métallique qui garde les courants induits par les hautes fréquences ambiantes sur l'extérieur du câble.

Ce que nous souhaiterions déterminer maintenant, c'est l'atténuation du champ électrique (ou un coefficient d'atténuation) dans la matière d'un câble conducteur cylindrique plein en fonction de la distance de son axe de symétrie longitudinal à sa surface extérieure..

Pour cela, nous reprenons la quatrième équation de Maxwell sous la forme donnée précédemment:

$$\vec{\nabla} \times \vec{H} = \vec{j} + \frac{d\vec{D}}{dt} = \sigma \vec{E} + \frac{d\vec{D}}{dt} \quad (38.105)$$

et que nous supposons travailler avec un conducteur n'ayant pas d'effet capacitif (donc pas de courant de déplacement) à l'opposé du cas général démontré dans le chapitre d'Électrodynamique, cette dernière se réduit alors à:

$$\vec{\nabla} \times \vec{H} = \frac{1}{\mu} \vec{\nabla} \times \vec{B} = \sigma \vec{E} \quad (38.106)$$

et si nous l'associons à la troisième équation de Maxwell (cf. chapitre Électrodynamique) qui est pour rappel:

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \quad (38.107)$$

Nous avons alors en utilisant le théorème de Schwarz (cf. chapitre de Calcul Différentiel Et Intégral):

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{E}) = -\frac{\partial}{\partial t} (\vec{\nabla} \times \vec{B}) = -\mu \sigma \frac{\partial \vec{E}}{\partial t} \quad (38.108)$$

Or nous avons démontré dans le chapitre de Calcul Vectoriel qu'en toute généralité le rotationnel du rotationnel d'un champ vectoriel est égal au gradient de la divergence de ce champ moins son laplacien vectoriel:

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{f}) = \vec{\nabla} (\vec{\nabla} \cdot \vec{f}) - \Delta \vec{f} \quad (38.109)$$

Or comme la divergence du champ électrique est nulle à travers la section d'un conducteur nous avons:

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{E}) = \vec{\nabla} (\vec{\nabla} \cdot \vec{E}) - \Delta \vec{E} = -\frac{\partial}{\partial t} (\vec{\nabla} \times \vec{B}) = -\mu\sigma \frac{\partial \vec{E}}{\partial t} \quad (38.110)$$

d'où:

$$\Delta \vec{E} = \mu\sigma \frac{\partial \vec{E}}{\partial t} \quad (38.111)$$

Ce qui donne explicitement pour une composante (par exemple en z):

$$\frac{\partial^2 E_z}{\partial z^2} = \mu\sigma \frac{\partial E_z}{\partial t} \quad (38.112)$$

Plaçons-nous dans le cas important d'un régime harmonique:

$$E_z(z, t) = E_0(z) \cos(kz - \omega t) \quad (38.113)$$

et utilisons temporairement la notation en phaseurs:

$$E_z(z, t) = E_0 e^{i(kz - \omega t)} = E_0 e^{ikz} e^{-i\omega t} \quad (38.114)$$

Nous avons alors

$$\begin{aligned} \frac{\partial^2 E_z}{\partial z^2} &= \frac{\partial (E_0 e^{ikz})}{\partial z^2} e^{-i\omega t} \\ \frac{\partial E_z}{\partial t} &= -i\omega E_0 e^{ikz} e^{-i\omega t} \end{aligned} \quad (38.115)$$

Nous obtenons alors l'équation différentielle suivante à une variable:

$$\frac{\partial^2 e^{ikz}}{\partial z^2} = -i\omega\sigma\mu e^{ikz} \quad (38.116)$$

Soit:

$$-k^2 = -i\omega\sigma\mu \quad (38.117)$$

et donc:

$$k = \pm \sqrt{i\omega\sigma\mu} \quad (38.118)$$

en se rappelant (cf. chapitre Nombres) que:

$$\sqrt{i} = \sqrt{e^{i\frac{\pi}{2}}} = \left(e^{i\frac{\pi}{2}} \right)^{\frac{1}{2}} = e^{i\frac{\pi}{4}} = \cos\left(\frac{\pi}{4}\right) - i \sin\left(\frac{\pi}{4}\right) = \frac{1-i}{\sqrt{2}} \quad (38.119)$$

il vient:

$$k = \pm \frac{1-i}{\sqrt{2}} \sqrt{\omega \sigma \mu} = \pm (1-i) \sqrt{\frac{\omega \sigma \mu}{2}} \quad (38.120)$$

Dès lors:

$$E_x(z, t) = E_0(z) e^{ikz} e^{-i\omega t} = E_0 e^{\pm(1-i) \sqrt{\frac{\omega \sigma \mu}{2}} z} e^{-i\omega t} = E_0 e^{\pm(i+1) \sqrt{\frac{\omega \sigma \mu}{2}} z} e^{-i\omega t} \quad (38.121)$$

Donc:

$$E_x(z, t) = E_0 e^{+\sqrt{\frac{\omega \sigma \mu}{2}} z} e^{i\left(\pm \sqrt{\frac{\omega \sigma \mu}{2}} z - \omega t\right)} \quad (38.122)$$

Nous devons rejeter pour des raisons physiques (conservation de l'énergie) la solution:

$$k = +\sqrt{i\omega\sigma\mu} \quad (38.123)$$

Il nous reste donc:

$$E_x(z, t) = E_0 e^{-\sqrt{\frac{\omega \sigma \mu}{2}} z} e^{-i\left(\sqrt{\frac{\omega \sigma \mu}{2}} z + \omega t\right)} \quad (38.124)$$

que les physiciens notent:

$$E_x(z, t) = E_0 e^{-\frac{z}{\delta}} e^{i\left(-\frac{z}{\delta} - \omega t\right)} \quad (38.125)$$

car δ se mesure en mètres (coefficient qui est nul si le champ électrique est constant) et est assimilé au "coefficient d'atténuation" que nous nous étions fixé de déterminer au début:

$$\delta = \sqrt{\frac{2}{\omega \sigma \mu}} = \sqrt{\frac{2\rho}{\omega \mu}} \quad (38.126)$$

Pour un conducteur en cuivre, nous avons selon Wikipédia les valeurs ci-dessous.

Fréquence	δ
50 Hz	9.38 mm
60 Hz	8.57 mm
10 kHz	0.66 mm
100 kHz	0.21 mm
1 MHz	66 μ m
10 MHz	21 μ m

Tableau: 38.2 - Coefficient d'atténuation

6. SEMI-CONDUCTEURS

Le défaut principal du modèle de Drude vu précédemment est de considérer l'électron comme une particule classique. Un ensemble de telles particules n'est évidemment pas soumis aux distributions quantiques et donc à une relation explicite de la température.

De plus, si nous observons notre modèle de Drude, il est difficile de dire quoi que ce soit à propos de la résistivité en fonction de la température.

Au fait, nous retenons en général quatre dates à la source du développement de cette théorie des semi-conducteurs:

- En 1833, Michael Faraday fait état de la conductivité d'un matériau qui augmente avec la température.
- En 1839, Antoine Becquerel découvre que sous illumination une tension électrique apparaît à la jonction de certains matériaux (et liquides). C'est l'effet photovoltaïque, qui donnera naissance beaucoup plus tard (vers 1950) aux cellules solaires.
- En 1873, Willoughby Smith montre que la conductivité de certaines substances augmente lorsque qu'on les illumine. C'est la photoconductivité.
- Enfin, en 1874 Karl Ferdinand Braun découvre le phénomène de redressement électrique lorsqu'une pointe métallique est déposée sur certains conducteurs, c'est-à-dire que le courant électrique passe dans un sens lorsque le potentiel électrique appliqué à la pointe est positif mais non lorsqu'il est négatif!

Bien que ces découvertes fussent totalement incomprises et surtout non reconnues comme étant les différentes expressions d'un même phénomène physique (la semi-conductivité), les applications pratiques furent immédiates et menèrent à la deuxième révolution industrielle qui est celle de la microélectronique!

Ce type de difficulté (parmi de nombreuses autres...) s'efface en grande partie avec le modèle de l'électron libre dans un puits de potentiel, imaginé par Sommerfeld en 1928. Dans ce modèle les électrons, soumis au principe de Pauli, suivent la distribution en énergie de Fermi-Dirac (cf. chapitre de Mécanique Statistique), alors que dans le modèle de Drude ils suivaient la loi de Maxwell-Boltzmann.

Il en découle deux résultats importants:

- Seule une fraction des électrons est susceptible de voir son énergie varier sous l'effet d'une action extérieure (température, champ électrique, etc.)
- Même au zéro absolu, l'énergie cinétique des électrons n'est pas nulle.

Le modèle de Sommerfeld fournit une base pour l'édification de théories plus spécifiques et est à la base du domaine de la "**physique du solide**" selon certaines sources. Ce n'est donc pas un modèle achevé traitant d'un problème précis comme la conduction électrique ou l'émission thermoélectronique. Cette base est la distribution en énergie des électrons, obtenue par le produit de deux fonctions: la densité des états et la distribution de Fermi-Dirac.

Malgré les améliorations qu'il apporte, ce modèle ne donne cependant pas une description satisfaisante des propriétés électroniques des solides dans tous les cas. Ses limitations proviennent du fait qu'il ne tient pas compte implicitement de la structure réelle des matériaux et des interactions entre électrons. Ce modèle ne permettra donc jamais d'expliquer objectivement pourquoi tel cristal est conducteur, et tel autre isolant ou semi-conducteur (par exemple le diamant et le silicium ont la même structure cristalline et configuration électronique mais à partir d'une certaine température l'un devient conducteur et l'autre reste isolant!).

La théorie des semi-conducteurs, appelée plus souvent "**théorie des bandes**" pour des raisons que nous verrons plus loin, est aussi un exemple fameux de l'application des résultats de la physique quantique ondulatoire (voir chapitre du même nom) et de la statistique quantique (cf. chapitre de Mécanique Statistique).

Pour son étude, nous nous concentrerons ici sur le modèle scolaire qualitatif le plus simple qui est celui basé sur un semi-conducteur cristallin avec un réseau parfaitement périodique et à bandes paraboliques (nous préciserons cela à nouveau plus loin).

Le lecteur un peu critique verra que les développements qui vont suivre ne sont cependant pas purement quantiques (il y a même des développements utilisant la mécanique classique qui sont limites acceptables suivant le point de vue)... donc l'approche est un peu grossière mais elle permet d'avoir une idée qualitative des phénomènes dans les semi-conducteurs. C'est une des raisons pour laquelle ce modèle est appelé "**modèle semi-classique des bandes paraboliques**".

Au fur et à mesure des années nous compléterons les développements qui vont suivre pour au final tenter d'avoir toute la démarche détaillée. D'ici là il faudra être patient...

Nous ferons abstraction des concepts qui ne sont pas absolument nécessaires à l'introduction du modèle pour présenter ici uniquement l'essentiel qui suffit à l'ingénieur dans son travail quotidien.

Pour commencer la partie mathématique de l'étude des semi-conducteurs, nous considérerons un cristal soumis à une différence de potentiel. Un électron de conduction du cristal sera donc soumis d'une part à une force interne $\vec{F}_i$ résultant du champ cristallin, et d'autre part à une force d'origine externe $\vec{F}_e$ résultant du champ électrique appliqué au cristal.

Les hypothèses du modèle sont:

H1. Il existe à la surface des métaux une barrière de potentiel empêchant les électrons de quitter la matière.

H2. À l'intérieur de la matière, les électrons sont soumis à un potentiel constant!

H3. Les électrons sont indépendants (pas d'interactions entre eux).

H4. Les électrons obéissent aux lois de la mécanique quantique et classique.

H5. Les électrons obéissent aux lois de l'électrodynamique de Maxwell.

H6. Les bandes d'énergie forme un spectre continu de niveaux d'énergie.

La première hypothèse repose sur l'observation élémentaire suivante: les électrons se déplaçant dans un métal ne franchissent pas, à température ambiante tout au moins, les surfaces limitant l'échantillon.

La deuxième hypothèse paraît assez brutale. C'est elle qui bannit du modèle la notion de structure de la matière. Elle sera remplacée dans le modèle des bandes d'énergie par un potentiel périodique rendant compte de l'influence des noyaux chargés positivement. Cette hypothèse traduit le fait que les électrons sont considérés comme libres dans le puits de potentiel.

La barrière de potentiel possède une largeur finie, c'est-à-dire que le passage du potentiel régnant à l'intérieur de la matière au potentiel régnant à l'extérieur se fait sur quelques distances interatomiques. Mais les dimensions de l'échantillon étant en pratique toujours très grandes vis-à-vis d'une distance interatomique, on peut considérer la barrière de potentiel comme infiniment abrupte, ce qui simplifie les calculs.

Remarque: Nous admettrons pour simplifier les calculs que les électrons se déplacent dans une seule direction (celle du champ électrique) ce qui évitera de se balader avec des vecteurs.

L'équation de la dynamique s'écrit alors naturellement pour cet électron:

$$m_0 \frac{d\vec{v}}{dt} = \vec{F}_i + \vec{F}_e \quad (38.127)$$

Nous écrivons alors (rien ne nous interdit de le faire) que l'électron dans le cristal répond à la sollicitation de la force externe $\vec{F}_e$ comme une quasi-particule de masse m_e dans le vide:

$$m_0 \frac{d\vec{v}}{dt} = \vec{F}_i + \vec{F}_e = \vec{F}_i + m_e \frac{d\vec{v}}{dt} \quad (38.128)$$

C'est l'étude de ce dernier terme qui va nous intéresser. Pour cela rappelons que dans le cadre de l'étude détaillée de la propagation de l'électron libre dans le vide, où nous négligeons les effets de son spin, nous avons démontré qu'il doit obligatoirement être décrit selon l'équation de Schrödinger par un paquet d'onde (cf. chapitre de Physique Quantique Ondulatoire) centré sur un état k_0 sinon quoi son énergie serait infinie.

On peut cependant se poser la question... de ce qui nous amène à le considérer comme libre.... Eh bien c'est l'expérience qui montre que lorsque nous appliquons un certain potentiel seuil, un courant commence à apparaître dans les semi-conducteurs.

Nous avons démontré (toujours dans le cadre de la propagation de la particule libre sans spin dans le chapitre de Physique Quantique Ondulatoire) que le paquet d'onde peut alors être vu dans sa solution mathématique comme une onde plane (libre) se déplaçant à la vitesse de phase:

$$v_\varphi = \frac{\omega_0}{k_0} \quad (38.129)$$

que nous noterons pour la suite afin de simplifier les notations:

$$v = \frac{\omega}{k} \quad (38.130)$$

Or, dans le réseau cristallin, la vitesse de phase peut varier, en fonction de l'endroit du réseau où se trouve l'électron à cause de la forme géométrique du potentiel dans le cristal. Il nous faut donc utiliser la vitesse de phase instantanée:

$$\vec{v} = \frac{d\varphi}{d\vec{k}} = \begin{pmatrix} \frac{d\varphi}{dk_x} \\ \frac{d\varphi}{dk_y} \\ \frac{d\varphi}{dk_z} \end{pmatrix} \quad (38.131)$$

Rappelons que nous avons aussi en toute généralité l'énergie totale donnée par:

$$E = \hbar \nu = \hbar \frac{\omega}{2\pi} = \hbar \omega \quad (38.132)$$

Il vient alors:

$$\nu = \frac{d\varphi}{dk} = \frac{1}{\hbar} \frac{dE}{d\vec{k}} \quad (38.133)$$

Le terme:

$$\frac{dE}{d\vec{k}} \quad (38.134)$$

n'est de loin pas simple dans le cas d'un cristal (c'est même un cauchemar...).

Évidemment pour une particule libre (cf. chapitre de Physique Quantique Ondulatoire), rappelons qu'il s'agit de:

$$\frac{dE}{dk} = \frac{d}{dk} \left(\frac{\hbar^2 k^2}{2m} \right) = \frac{\hbar^2 k}{m} \quad (38.135)$$

Mais pour une particule dans un champ de potentiel ayant une géométrie complexe l'énergie E commence à avoir une expression dépendante de k en fonction des zones qui peut devenir très complexe (voir les exemples du chapitre de Physique Quantique Ondulatoire). D'où la justification de l'utilisation de la dérivée.

L'accélération au sens classique de cet électron est alors donnée par:

$$\vec{a} = \frac{d\vec{v}}{dt} = \frac{1}{\hbar} \frac{d}{dt} \left(\frac{dE}{d\vec{k}} \right) = \frac{1}{\hbar} \frac{d}{d\vec{k}} \frac{dE}{dt} \quad (38.136)$$

Nous avons aussi (cf. chapitre de Mécanique Classique):

$$dE = \vec{F} \circ d\vec{x} = \vec{F} \circ \vec{v} dt \quad (38.137)$$

donc:

$$\frac{dE}{dt} = \vec{F} \circ \vec{v} \quad (38.138)$$

d'où:

$$\vec{a} = \frac{d\vec{v}}{dt} = \frac{1}{\hbar} \frac{d}{dk} \frac{dE}{dt} = \frac{1}{\hbar} \frac{d}{dk} (\vec{F} \circ \vec{v}) = \frac{1}{\hbar} \frac{d}{dk} \left(\vec{F} \circ \frac{1}{\hbar} \frac{dE}{dk} \right) \quad (38.139)$$

La dérivée de $\vec{F}$ par rapport à $\vec{k}$ dans la relation précédente s'annulera car la force découle du potentiel appliqué sur le semi-conducteur seulement et non pas du vecteur d'onde de l'électron lui-même! Il nous reste alors:

$$\vec{a} = \vec{F} \frac{1}{\hbar^2} \frac{d^2 E}{dk^2} \quad (38.140)$$

Puisque ici E est uniquement l'énergie totale provenant du potentiel soumis de l'extérieur, alors la force $\vec{F}$ est la force externe $\vec{F}_e$ générée par l'application de ce même potentiel. Nous avons alors:

$$\vec{F}_e = m_e \frac{d\vec{v}}{dt} = m_e \cdot \vec{a} \quad (38.141)$$

et:

$$\vec{F}_e = \frac{1}{\frac{1}{\hbar^2} \frac{d^2 E}{dk^2}} \vec{a} \quad (38.142)$$

Il vient alors par égalisation:

$$m_e = \frac{\hbar^2}{\frac{d^2 E}{dk^2}} \quad (38.143)$$

Puisque l'énergie de l'électron peut avoir une forme mathématique compliquée conformément aux cas applicatifs vus dans le chapitre de Physique Quantique Ondulatoire, exprimons $E(\vec{k})$ sous forme de développement limité de Taylor (cf. chapitre de Suites et Séries) d'une fonction à trois variables au deuxième ordre en laissant tomber les termes d'interactions et en ne prenant pas les termes de premier degré:

$$E(\vec{k}) = E(k_x, k_y, k_z) \cong E(k_{x0}, k_{y0}, k_{z0}) + \frac{1}{2!} \frac{\partial^2 E}{\partial k_x^2} (k_x - k_{x0})^2 + \frac{1}{2!} \frac{\partial^2 E}{\partial k_y^2} (k_y - k_{y0})^2 + \frac{1}{2!} \frac{\partial^2 E}{\partial k_z^2} (k_z - k_{z0})^2 \quad (38.144)$$

Au fait cette approximation grossière mais toutefois acceptable dans pas mal de cas pratiques tient au fait que l'expérience montre que les surfaces d'énergie en fonction de k ont en approximation une forme parabolique dans certains cristaux semi-conducteurs.

Dans les conducteurs, l'approximation de la relation précédente n'est prise qu'au premier terme.

Une autre manière de le voir est que pour un électron libre, nous avons pour rappel, en une dimension, la courbe de dispersion (cf. chapitre de Physique Quantique Ondulatoire):

$$E(k) = \frac{\hbar^2}{2m} k^2 \quad (38.145)$$

qui est bien une parabole en fonction de k . Effectivement, si nous prenons notre développement de Taylor en une dimension il nous reste:

$$E(k) = E(k) \cong E(k_0) + \frac{1}{2!} \frac{\partial^2 E}{\partial k^2} (k - k_0)^2 \quad (38.146)$$

et comme nous avons déterminé avant que:

$$m_e = \frac{\hbar^2}{\frac{d^2 E}{dk^2}} \quad (38.147)$$

Il vient:

$$E(k) \cong E(k_0) + \frac{1}{2!} \frac{\partial^2 E}{\partial k^2} (k - k_0)^2 = E(k_0) + \frac{1}{2!} \frac{\hbar^2}{m_e} (k - k_0)^2 \quad (38.148)$$

Si l'électron est libre la courbe de dispersion nous impose d'avoir (sans présence de potentiel):

$$E(k_0) = 0 \quad (38.149)$$

qui est alors considérée comme "**l'énergie du minimum**" $E_{\min}$. Il nous reste alors:

$$E(k) = \frac{\hbar^2}{2m_e} (k_x - k_{x0})^2 \quad (38.150)$$

et en prenant $k_{x0} = 0$ nous retombons sur courbe de dispersion d'une particule libre (ce qui justifie donc le fait d'avoir posé $E(k_0) = 0$ pour un électron libre):

$$E(k) = \frac{\hbar^2}{2m_e} k_x^2 \quad (38.151)$$

Ce qui montre que l'approximation n'est pas trop fausse... et justifie le fait que dans certains ouvrages la relation précédente (série de Taylor) décrit une particule dite "**quasi-libre**".

Mais revenons-en à:

$$E(k) = E(k_0) + \frac{\hbar^2}{2m_e} (k - k_0)^2 \quad (38.152)$$

Et puisque le paquet d'onde est centré autour de k_0 normalisons-la comme valant 0 (ce qui équivaut à centrer les valeurs du vecteur d'onde). Nous avons alors:

$$E(k) = E(0) + \frac{\hbar^2}{2m_e} k^2 = E_0 + \frac{\hbar^2}{2m_e} k^2 \quad (38.153)$$

Ce qui est intéressant avec ces développements, c'est que nous sommes partis d'un électron libre sous forme de paquet d'onde et grâce au développement de Taylor nous nous retrouvons avec une expression extrêmement simple de l'énergie d'un électron quasi-libre.

Il en sort que pour un électron quasi-libre, sans interactions et sans prendre en compte les effets de spin nous avons:

$$k \cong \sqrt{(E - E_0) \frac{2m_e}{\hbar^2}} \quad (38.154)$$

Nous remarquons alors une chose fort sympathique! C'est que notre électron quasi-libre a un nombre d'onde qui ressemble en tout point à celui d'une particule coincée dans un puits de potentiel à parois rectilignes (voir la démonstration dans le chapitre de Physique Quantique Ondulatoire).

Nous souhaiterons maintenant calculer à l'aide de l'expression de k (n'ayant pas directement celle de E car trop complexe) la densité d'états (in extenso d'électrons) dans le volume donné par le puits rectangulaire correspondant.

Nous avons démontré dans le chapitre de Physique Quantique Ondulatoire que pour le puits de potentiel à barrières rectangulaires:

$$k = n \frac{\pi}{L} \quad (38.155)$$

si nous imposons un nombre entier de demi-longueur d'onde. Si nous imposons un nombre entier de longueur d'onde (conditions de Born-von Karman afin qu'après une translation du réseau périodique du cristal nous retrouvions les mêmes propriétés) pour que la solution soit physiquement acceptable, nous avons alors:

$$k = n \frac{2\pi}{L} \quad (38.156)$$

ce qui implique bien évidemment deux fois moins d'états.

Par extension, pour l'espace, nous avons alors dans le cas tridimensionnel:

$$k_x = n_x \frac{2\pi}{L_x}, k_y = n_y \frac{2\pi}{L_y}, k_z = n_z \frac{2\pi}{L_z} \quad (38.157)$$

avec:

$$E_{n_x, n_y, n_z} = \frac{\pi^2 \hbar^2}{2mL^2} (n_x^2 + n_y^2 + n_z^2) = E_0 (n_x^2 + n_y^2 + n_z^2) \quad (38.158)$$

et où $n_{x,y,z} = 1, 2, 3, \dots$

Le résultat est très similaire à celui du puits de potentiel rectangulaire à une dimension mais nous avons maintenant des conditions aux bords particulières afin d'avoir une correspondance avec l'expérience et trois nombres quantiques principaux au lieu d'un seul. De plus, chaque combinaison de ces trois nombres correspond à une fonction d'onde (état) différente. De plus, ces nombres sont indépendants (aucune condition imposée).

Nous avons alors le premier niveau où tous les n sont unitaires:

$$k_x k_y k_z = \frac{(2\pi)^3}{L_x L_y L_z} \quad (38.159)$$

Si nous acceptons pour simplifier que le puits a des arêtes de longueurs égales (semi-conducteur à réseau cristallin cubique), nous avons alors:

$$k_x k_y k_z = \frac{(2\pi)^3}{L^3} = (2\pi/L)^3 \quad (38.160)$$

Représentons l'espace des k pour un tel réseau cubique et pour différents multiples de n_x, n_y, n_z :

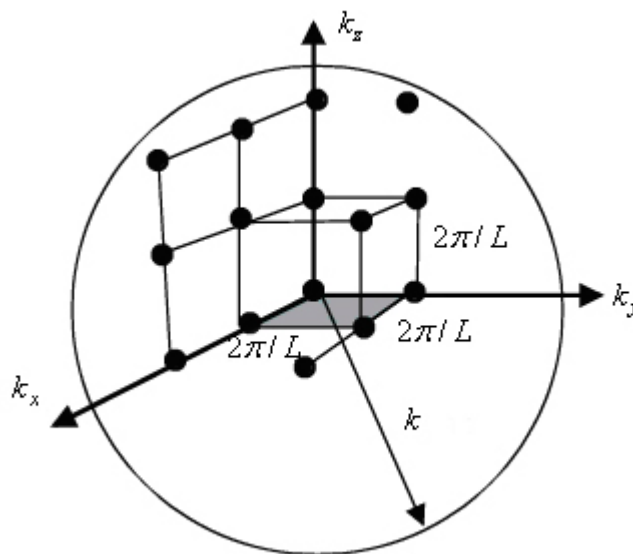


Figure: 38.2 - Espace des k pour un réseau cristallin cubique

Donc tous les états quantifiés ne peuvent prendre que des valeurs espacées de $2\pi/L$ dans l'espace des k ce qui signifie que par volume élémentaire il n'y a qu'un seul vecteur d'onde possible et donc qu'un seul état associé. Effectivement, faites un dessin par-dessus la figure ci-dessus si vous voulez et vous verrez (!) mais ne vous fiez pas aux gros points noirs qui sont là uniquement pour montrer les extrémités des volumes élémentaires et qui ne correspondent pas tous à des états possibles!

Ainsi, dans un volume sphérique de rayon k de l'espace des k . Nous avons un nombre précis (limite supérieure) de volume élémentaires (états):

$$N(k) = \frac{\frac{4}{3} \pi k^3}{k_x k_y k_z} = \frac{\frac{4}{3} \pi k^3}{(2\pi/L)^3} = \frac{k^3}{6\pi^2} V \quad (38.161)$$

où dans la littérature il est d'usage (tradition) de ne conserver que la forme de la deuxième égalité dans les développements. Cette relation nous a été pour rappel utile dans le chapitre de Thermodynamique pour déterminer le modèle de Debye-Einstein de la capacité thermique à volume constant des solides cristallins!

La densité de modes dans un volume V sera alors donné par (relation utilisée dans le chapitre de Thermodynamique pour exprimer la capacité calorifique à volume constant des solides):

$$D(\omega) = \frac{dN}{dk} \frac{dk}{d\omega} = \frac{Vk^2}{2\pi^2} \frac{dk}{d\omega} \quad (38.162)$$

Remarque: La sphère de rayon k , contenant les niveaux à un électron qui sont occupés est appelée parfois "**sphère de Fermi**". La valeur du rayon est alors notée k_F et appelée "**vecteur d'onde de Fermi**". La surface de la sphère de Fermi, qui sépare les niveaux occupés de ceux qui ne le sont pas comme nous le verrons plus tard est appelée "**surface de Fermi**".

En considérant maintenant le spin (ben oui tant qu'on y est...) nous multiplions par 2 puisqu'il y a deux états de spin possible par état:

$$N(k) = 2 \cdot \frac{4}{3} \frac{\pi k^3}{(2\pi)^3} L^3 = \frac{k^3}{3\pi^2} V \quad (38.163)$$

(relation que nous retrouverons dans le chapitre de Physique Nucléaire lors de notre étude du modèle de noyau nucléaire sous forme de goutte liquide) et en y injectant:

$$k \cong \sqrt{(E - E_0) \frac{2m_e}{\hbar^2}} \quad (38.164)$$

nous avons alors:

$$\begin{aligned} N(E) &\cong 2 \cdot \frac{\frac{4}{3} \pi}{(2\pi)^3} \left((E - E_0) \frac{2m_e}{\hbar^2} \right)^{3/2} L^3 = \frac{\frac{4}{3}}{(2\pi)^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} (E - E_0)^{3/2} L^3 \\ &= \frac{4}{(2\pi)^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} (E - E_0)^{3/2} \cdot V \end{aligned} \quad (38.165)$$

La densité volumique d'états (quasi-)libres sera obtenue en dérivant cette dernière relation par le volume:

$$\rho(E) = \frac{dN(E)}{dV} = \frac{\frac{4}{3}}{(2\pi)^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} (E - E_0)^{3/2} \quad (38.166)$$

Et si nous souhaitons la densité d'états (quasi-)libres (de vibration) par unité d'énergie et de volume il va nous falloir en plus dériver par rapport à l'énergie:

$$\frac{d\rho(E)}{dE} \cong \frac{\frac{4}{3}}{(2\pi)^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \frac{d}{dE} (E - E_0)^{3/2} = \frac{\frac{4}{3}}{(2\pi)^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \frac{3}{2} (E - E_0)^{1/2} \quad (38.167)$$

Ce qui donne:

$$D(E) \cong \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} (E - E_0)^{1/2} \quad (38.168)$$

Ce résultat ne dépendant pas du volume, il est inchangé lorsque celui-ci tend vers l'infini! Donc il est valable pour tout point du cristal semi-conducteur si celui-ci est parfait...

Ce que nous trouvons également parfois sous les formes (un peu malheureuses...) suivantes dans certains ouvrages:

$$\begin{aligned} D(E) &\cong \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \sqrt{E} \\ D(E) &\cong \frac{4}{h^3 \sqrt{\pi}} (2\pi m_e)^{3/2} (E - E_0)^{1/2} \end{aligned} \quad (38.169)$$

et il y aussi ceux qui ne prennent en compte le spin que plus tard... ce qui donne une forme identique à celle des trois relations précédentes mais à diviser par 2.

6.1. DENSITÉ STATISTIQUE NON-DÉGÉNÉRÉE DES PORTEURS NÉGATIFS

Bref, cependant cette relation a un défaut (encore un...)! Effectivement, nous avons vu dans le chapitre de Mécanique Statistique lors de notre étude de la statistique quantique que dans un système où même le spectre d'énergie est considéré comme continu, il est impossible de ne pas prendre en compte la dégénérescence des différents niveaux d'énergie. Nous avons démontré alors que pour une population de fermions, à une énergie (ou température) donnée le pourcentage de niveaux dégénérés occupés est donné par la fonction de Fermi-Dirac:

$$f_{FD}(E, T, \mu) = \frac{1}{\exp\left(\frac{E - \mu}{kT}\right) + 1} \quad (38.170)$$

et que la fonction retournait donc une valeur comprise entre 0 au minimum et 1 au maximum.

Cette fonction donne donc pour une température T fixée la probabilité qu'un électron occupe un état d'énergie E .

Ce qui fait que notre relation $D(E)$ surestime la valeur réelle de densité d'états (quasi-)libres occupés pour une énergie (ou température) donnée. Ce qui fait que pour avoir une meilleure approximation nous écrivions en toute logique la densité volumique d'états (quasi-)libres par unité d'énergie:

$$D(E) \cong \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} (E - E_0)^{1/2} \cdot \frac{1}{\exp\left(\frac{E - \mu}{kT}\right) + 1} \quad (38.171)$$

Cependant, dans la pratique, nous allons chercher à calculer la densité volumique d'états (quasi-)libres dans un spectre (intervalle) d'énergie. Il vient alors avec la correction ajoutée précédemment:

$$\frac{d\rho(E)}{dE} \cong \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \frac{(E - E_0)^{1/2}}{\exp\left(\frac{E - \mu}{kT}\right) + 1} \quad (38.172)$$

Soit:

$$d\rho(E) \cong \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \frac{(E - E_0)^{1/2}}{\exp\left(\frac{E - \mu}{kT}\right) + 1} dE \quad (38.173)$$

Il vient alors immédiatement que la densité volumique d'états (d'électrons) (quasi-)libres à une température donnée (conditions normales de température pour les applications civiles) en prenant en compte tous les états (niveaux continus) d'énergie possibles est alors:

$$\rho(E) \cong \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \int_{E_0}^{+\infty} \frac{(E - E_0)^{1/2}}{\exp\left(\frac{E - \mu}{kT}\right) + 1} dE \quad (38.174)$$

Prendre E_0 comme borne inférieure nous évite, comme nous allons le voir explicitement un peu plus bas, de nous retrouver avec une racine négative... ce qui serait fort gênant!

De plus, nous pouvons sans erreur appréciable reporter la limite de l'intégrale à l'infini car $f_{FD} \rightarrow 0$ quand E est importante.

Malheureusement, cette intégrale n'est en général pas soluble analytiquement. Il va donc falloir recourir à des approximations.

Nous allons commencer par faire l'hypothèse que nous sommes dans le régime classique du gaz d'électrons. C'est-à-dire que le terme:

$$\exp\left(\frac{E - \mu}{kT}\right) \gg 1 \quad (38.175)$$

ce qui implique:

$$E - \mu \gg kT \quad (38.176)$$

Dès lors, nous avons aussi l'approximation:

$$\frac{1}{\exp\left(\frac{E - \mu}{kT}\right) + 1} \cong \frac{1}{\exp\left(\frac{E - \mu}{kT}\right)} = \exp\left(-\frac{E - \mu}{kT}\right) \quad (38.177)$$

En d'autres termes l'énergie E doit être bien supérieure au potentiel chimique μ (assimilé souvent malheureusement et à ma connaissance à tort dans la littérature sur les semi-conducteurs au niveau de Fermi E_F). Les physiciens notent alors cette énergie E_C pour la distinguer et l'appellent "**énergie minimale de la bande de conduction**" (qui correspond à l'énergie minimale d'un électron quasi-libre pour satisfaire cette condition).

Dès lors, nous changeons aussi la notation pour la densité de charge:

$$\rho(E) \rightarrow \rho_C(E) \quad (38.178)$$

Remarque: Malheureusement, comme précisé dans le paragraphe précédent (!) dans beaucoup d'ouvrages de qualité sur les semi-conducteurs, le potentiel chimique μ , qui est pour rappel une notion purement thermodynamique impliquant une hypothèse d'interactions, est remplacée par le concept d'énergie de Fermi E_F et pourtant ce n'est pas la même chose! Les deux énergies ne coïncident que dans le cas où la température T est nulle!

Donc nous devons considérer le terme de "**niveau de Fermi**", comme n'étant rien d'autre qu'un synonyme de "**potentiel chimique**" dans le contexte des semi-conducteurs.

Nous avons alors:

$$\begin{aligned} \rho_C(E) &\cong \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \int_{E_C}^{+\infty} (E - E_C)^{1/2} \exp\left(-\frac{E - \mu}{kT}\right) dE \\ &= \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \int_{E_C}^{+\infty} (E - E_C)^{1/2} f_{MB} dE \\ &= \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \exp\left(\frac{\mu}{kT}\right) \int_{E_C}^{+\infty} (E - E_C)^{1/2} \exp\left(-\frac{E}{kT}\right) dE \end{aligned} \quad (38.179)$$

Où f_{MB} est la fonction de Maxwell-Boltzmann (cf. chapitre de Mécanique Statistique) donnée pour rappel par:

$$f_{MB}(E, T, \mu) = \exp\left(-\frac{E - \mu}{kT}\right) \quad (38.180)$$

Et correspond donc bien à un comportement non quantique (soit un gaz d'électrons non-dégénéré!) car lorsque:

$$E - \mu \gg kT \quad (38.181)$$

nous avons:

$$f_{MB}(E, T, \mu) \ll 1 \quad (38.182)$$

et donc les états d'énergie ne sont de loin pas tous occupés par les électrons (il n'y a donc pas dégénérescence).

Nous sommes donc bien dans une situation où la physique classique prédomine sur la physique quantique. C'est la raison pour laquelle dans cette approximation (de Maxwell-Boltzmann) nous disons alors que nous avons affaire à un "**semi-conducteur non-dégénéré**" car les électrons ne sont pas entassés dans les niveaux les plus bas disponibles.

Pour pouvoir continuer, nous faisons un changement de variable en posant:

$$x = \frac{E - E_C}{kT} \quad (38.183)$$

d'où:

$$dx = \frac{1}{kT} dE \text{ et } E = x \cdot kT + E_C \quad (38.184)$$

Il vient alors:

$$\begin{aligned} \int_{E_C}^{+\infty} \sqrt{E - E_C} \exp\left(-\frac{E}{kT}\right) dE &= \sqrt{kT} \int_0^{+\infty} x^{1/2} \exp\left(-\frac{x \cdot kT + E_C}{kT}\right) dx kT \\ &= (kT)^{3/2} \exp\left(-\frac{E_C}{kT}\right) \int_0^{+\infty} x^{1/2} e^{-x} dx \end{aligned} \quad (38.185)$$

Nous faisons une intégration par parties:

$$\int_0^{+\infty} x^{1/2} e^{-x} dx = -\left[x^{1/2} e^{-x}\right]_0^{+\infty} + \frac{1}{2} \int_0^{+\infty} \frac{e^{-x}}{\sqrt{x}} dx = \frac{1}{2} \int_0^{+\infty} \frac{e^{-x}}{\sqrt{x}} dx \quad (38.186)$$

nous faisons ensuite un changement de variable en posant:

$$u = \sqrt{x} \Rightarrow \frac{du}{dx} = \frac{1}{2\sqrt{x}} \Rightarrow dx = 2\sqrt{x} du = 2u du \quad (38.187)$$

Ce qui donne:

$$\frac{1}{2} \int_0^{+\infty} \frac{e^{-x}}{\sqrt{x}} dx = \frac{1}{2} \int_0^{+\infty} \frac{e^{-u^2}}{u} 2u du = \int_0^{+\infty} e^{-u^2} du \quad (38.188)$$

Nous avons déjà calculé cette intégrale dans le chapitre de Statistiques. Il vient:

$$\int_0^{+\infty} x^{1/2} e^{-x} dx = \int_0^{+\infty} e^{-u^2} du = \frac{\sqrt{\pi}}{2} \quad (38.189)$$

Nous avons alors finalement:

$$\begin{aligned} \rho_C(E) &\cong \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \exp\left(\frac{\mu}{kT}\right) (kT)^{3/2} \exp\left(-\frac{E_C}{kT}\right) \frac{\sqrt{\pi}}{2} \\ &= \frac{1}{2\pi^2} \left(\frac{2m_e kT}{\hbar^2} \right)^{3/2} \exp\left(\frac{\mu - E_C}{kT}\right) \frac{\sqrt{\pi}}{2} = \frac{1}{2^2 \pi^{3/2}} \left(\frac{2m_e kT}{\hbar^2} \right)^{3/2} \exp\left(\frac{\mu - E_C}{kT}\right) \quad (38.190) \\ &= 2 \left(\frac{m_e kT}{2\pi\hbar^2} \right)^{3/2} \exp\left(\frac{\mu - E_C}{kT}\right) \end{aligned}$$

Où pour rappel, m_e est la masse de la quasi-particule (et non la masse de l'électron pour rappel!). Donc après intégration tout se passe comme si tous les électrons étaient concentrés sur le niveau d'énergie E_C avec un nombre de places disponibles correspondant à:

$$2 \left(\frac{m_e kT}{2\pi\hbar^2} \right)^{3/2} \quad (38.191)$$

Ce que nous notons traditionnellement (et de manière un peu malheureuse... car il n'est pas évident de se rappeler qu'il s'agit d'une densité):

$$N_n = N_{n,T} \exp\left(\frac{\mu - E_C}{kT}\right) \quad (38.192)$$

Ou encore:

$$\boxed{N_n = N_C \exp\left(\frac{\mu - E_C}{kT}\right)} \quad (38.193)$$

Où nous avons environ à température ambiante (c'est le paramètre de la masse effective qui varie entre les deux) les valeurs suivantes d'états (quasi-)libres respectives pour le Silicium:

$$N_C = 2 \left(\frac{m_e kT}{2\pi\hbar^2} \right)^{3/2} \cong 2.7 \cdot 10^{19} \text{ [cm}^{-3}\text{]} \quad (38.194)$$

Et pour le Germanium:

$$N_C = 2 \left(\frac{m_e kT}{2\pi\hbar^2} \right)^{3/2} \cong 1 \cdot 10^{19} \left[cm^{-3} \right] \quad (38.195)$$

Alors qu'il y a environ une densité d'atomes de $4.5 \cdot 10^{22} \left[cm^{-3} \right]$ et environ $10^{24} \left[cm^{-3} \right]$ électrons pour ces deux éléments.

Cela signifie qu'il y a donc un rapport de 1/100000 entre la densité d'électrons totale et le nombre d'électrons quasi-libres.

Nous remarquons cependant également que ce modèle théorique ne prend pas en compte la structure électronique (numéro atomique) du matériau étudié.

Ainsi, nous voyons que les variations des densités d'électrons quasi-libres en fonction de la température (dans la gamme de validité de la température...) sont essentiellement de type exponentiel croissant ou décroissant.

À partir de la densité des électrons libres (attention il faut bien se rappeler que ce sont uniquement les électrons quasi-libres qui se baladent dans nos équations mathématiques jusqu'à maintenant) dans le cristal semi-conducteur, nous pouvons en déduire l'énergie du niveau de Fermi (plus rigoureusement il s'agit du potentiel chimique!):

$$N_n = N_C \exp\left(\frac{\mu - E_C}{kT}\right) \Rightarrow \frac{N_n}{N_C} = \exp\left(\frac{\mu - E_C}{kT}\right) \Rightarrow \ln\left(\frac{N_n}{N_C}\right) = \frac{\mu - E_C}{kT} \quad (38.196)$$

d'où:

$$\mu = E_C + kT \ln\left(\frac{N_n}{N_C}\right) \quad (38.197)$$

Et puisque $N_T \geq N_n$ nous avons toujours à cause du logarithme qui est négatif, l'énergie de Fermi (plus rigoureusement il s'agit du potentiel chimique!) qui est inférieure ou égale à l'énergie des électrons quasi-libres:

$$\mu \leq E_C \quad (38.198)$$

Ou en d'autres termes, les électrons (quasi-)libres ont une énergie supérieure à l'énergie de Fermi (potentiel chimique...) ce qui est conforme à l'approximation du gaz non dégénéré faite plus haut. Cela donne une condition d'importance capitale pour que les porteurs négatifs puissent être les générateurs de la conduction dans le matériau.

Ainsi, lorsque nous nous plaçons à une température différente du zéro absolu, les états électroniques ne sont pas tous dégénérés: il y a étalement des états occupés au voisinage de ce qui constitue par définition l'énergie de Fermi (cf. chapitre de Mécanique Statistique et Physique Quantique Ondulatoire), effet d'autant plus accentué que la température est élevée.

6.2. DENSITÉ STATISTIQUE NON-DÉGÉNÉRÉE DES PORTEURS POSITIFS

Avant toute chose, il faut savoir que dans l'état actuel de nos connaissances les "trous" n'émergent pas des équations mais sont une construction empirique qui permet de faire correspondre la théorie et l'expérience (charges positives de l'effet Hall par exemple). Il s'agit donc d'un artifice pour faire une théorie simple d'une question intraitable rigoureusement à notre époque par la physique quantique.

Personnellement, je considère les trous de la même manière que les points de Lagrange en astronomie: Même s'il n'y a aucun corps en ces points de Lagrange cela n'empêche pas un satellite de se mettre en orbite (quasi-stable) autour de ceux-ci (possibilité que nous n'avons pas démontrée dans le chapitre d'Astronomie) comme s'il y avait une masse! Par ailleurs des expériences auraient montré au début des années 2000 que des points de Lagrange apparaissent au niveau de l'atome dans certaines conditions idéales et simplifiées!

Ceci dit, il faut se rappeler qu'un trou n'est pas un électron qui manque! C'est une idiotie (selon moi...) que l'on voit dans certains ouvrages spécialisés.

Au risque de se répéter un peu souvent, rappelons que pour une température T fixée la probabilité qu'un électron occupe un état d'énergie E est donnée par:

$$f_{FD}(E, T, \mu) = \frac{1}{\exp\left(\frac{E - \mu}{kT}\right) + 1} \quad (38.199)$$

Ce qui fait que pour avoir une meilleure approximation, nous écrivons en toute logique la densité volumique d'états occupés par unité d'énergie:

$$D(E) \cong \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2}\right)^{3/2} (E - E_0)^{1/2} \cdot \frac{1}{\exp\left(\frac{E - \mu}{kT}\right) + 1} \quad (38.200)$$

Ce qui nous a amené finalement à la relation suivante de la densité volumique d'états de charges négatives où la présence d'une masse dans la relation indique que les états occupés le sont par des quasi-particules telle que:

$$N_n = 2 \left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2} \exp\left(\frac{\mu - E_C}{kT}\right) \quad (38.201)$$

Mais qu'en est-il de la probabilité qu'un électron n'occupe pas pour une température T fixée un état d'énergie E et trivialement donnée par la différence:

$$f_p(E, T, \mu) = 1 - f_n(E, T, \mu) = 1 - f_{FD}(E, T, \mu) \quad (38.202)$$

où le n en indice est là pour indiquer que la distribution concerne les porteurs "négatifs" (distribution donnée comme nous l'avons démontré juste précédemment par la distribution de Maxwell-Boltzmann qui découle d'une approximation de la loi de Fermi-Dirac).

Eh bien nous allons constater que les équations nous conduisent à la possibilité d'associer aussi à ces états non occupés une densité volumique d'états avec une masse effective donnée. Nous verrons aussi plus tard qu'il sera possible d'associer à ces états non occupés une charge électrique positive et égale et à celle de l'électron, d'où le p en indice dans la relation précédente et signifiant "positif".

Nous avons donc pour ces porteurs positifs:

$$f_p(E, T, \mu) = 1 - \frac{1}{\exp\left(\frac{E - \mu}{kT}\right) + 1} = \frac{\exp\left(\frac{E - \mu}{kT}\right) + 1}{\exp\left(\frac{E - \mu}{kT}\right) + 1} - \frac{1}{\exp\left(\frac{E - \mu}{kT}\right) + 1} \quad (38.203)$$

$$= \frac{\exp\left(\frac{E - \mu}{kT}\right)}{\exp\left(\frac{E - \mu}{kT}\right) + 1} = \frac{\exp\left(\frac{E - \mu}{kT}\right)}{\exp\left(\frac{E - \mu}{kT}\right) + 1} \frac{\exp\left(\frac{\mu - E}{kT}\right)}{\exp\left(\frac{\mu - E}{kT}\right)} = \frac{1}{\exp\left(\frac{\mu - E}{kT}\right) + 1}$$

faisons maintenant une approximation similaire à celle utilisée pour les porteurs négatifs, c'est-à-dire que:

$$f_p(E, T, \mu) \ll 1 \quad (38.204)$$

pour imposer un régime semi-classique et donc les états d'énergie ne sont de loin pas tous occupés par les trous (il n'y a donc pas dégénérescence).

Cette restriction impose:

$$\exp\left(\frac{\mu - E}{kT}\right) \gg 1 \quad (38.205)$$

Soit écrit de la même manière que pour les porteurs négatifs:

$$\exp\left(-\frac{E - \mu}{kT}\right) \gg 1 \quad (38.206)$$

Soit contrairement aux porteurs négatifs cela impose:

$$E - \mu \ll kT \quad (38.207)$$

en d'autres termes l'énergie doit être bien inférieure au niveau de Fermi (potentiel chimique). Les physiciens notent alors cette énergie E_V pour la distinguer et l'appellent "**énergie maximale de la bande de valence**" (qui correspond à l'énergie maximale d'un trou quasi-libre pour satisfaire cette condition).

Remarque: Le lecteur pourra observer que les conditions susmentionnées imposent aussi que E est soit très petit en valeur absolue, soit négatif. Ce qui nous donne déjà une piste pour les bornes d'intégration à venir...

Nous avons alors:

$$f_p(E, T, \mu) = \frac{1}{\exp\left(\frac{\mu - E}{kT}\right) + 1} \cong \frac{1}{\exp\left(\frac{\mu - E}{kT}\right)} = \exp\left(-\frac{\mu - E}{kT}\right) \quad (38.208)$$

Dès lors, nous avons aussi l'approximation:

$$f_p(E, T, \mu) \cong \exp\left(-\frac{\mu - E}{kT}\right) \quad (38.209)$$

Nous sommes donc bien dans une situation où la physique classique prédomine sur la physique quantique. C'est la raison pour laquelle dans cette approximation nous disons que nous avons alors affaire à un "semi-conducteur non-dégénéré" car les trous ne sont pas entassés dans les niveaux les plus hauts disponibles.

Nous avons alors:

$$\begin{aligned} \rho_V(E) &\cong \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2}\right)^{3/2} \int_{-\infty}^{E_V} (E_V - E)^{1/2} \exp\left(-\frac{\mu - E}{kT}\right) dE \\ &= \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2}\right)^{3/2} \exp\left(\frac{-\mu}{kT}\right) \int_{-\infty}^{E_V} (E_V - E)^{1/2} \exp\left(\frac{E}{kT}\right) dE \end{aligned} \quad (38.210)$$

où le lecteur aura pu observé que les bornes d'intégration ont été choisies conformément à la remarque que nous avons faite juste précédemment et que les termes dans la racine carrée ont été permuté afin de ne pas y avoir de valeur négative.

Pour pouvoir continuer, nous faisons un changement de variable en posant:

$$x = \frac{E_V - E}{kT} \quad (38.211)$$

d'où:

$$dx = -\frac{1}{kT} dE \text{ et } E = -x \cdot kT + E_V \quad (38.212)$$

Il vient alors:

$$\begin{aligned} \int_{-\infty}^{E_V} \sqrt{E_V - E} \exp\left(\frac{E}{kT}\right) dE &= -\sqrt{kT} \int_{+\infty}^0 x^{1/2} \exp\left(\frac{-x \cdot kT + E_V}{kT}\right) dx kT \\ &= -(kT)^{3/2} \exp\left(\frac{E_V}{kT}\right) \int_{+\infty}^0 x^{1/2} e^{-x} dx \end{aligned} \quad (38.213)$$

Nous faisons une intégration par parties:

$$\int_{+\infty}^0 x^{1/2} e^{-x} dx = -\left[x^{1/2} e^{-x}\right]_{+\infty}^0 + \frac{1}{2} \int_{+\infty}^0 \frac{e^{-x}}{\sqrt{x}} dx = \frac{1}{2} \int_{+\infty}^0 \frac{e^{-x}}{\sqrt{x}} dx \quad (38.214)$$

nous faisons ensuite un changement de variable en posant:

$$u = \sqrt{x} \Rightarrow \frac{du}{dx} = \frac{1}{2\sqrt{x}} \Rightarrow dx = 2\sqrt{x}du = 2udu \quad (38.215)$$

Ce qui donne:

$$\frac{1}{2} \int_{+\infty}^0 \frac{e^{-x}}{\sqrt{x}} dx = \frac{1}{2} \int_{+\infty}^0 \frac{e^{-u^2}}{u} 2udu = \int_{+\infty}^0 e^{-u^2} du \quad (38.216)$$

Nous avons déjà calculé cette intégrale dans le chapitre de Statistiques. Il vient (puisque la fonction est paire nous utilisons la propriété démontrée dans le chapitre de Calcul Différentiel Et Intégral):

$$\int_{+\infty}^0 x^{1/2} e^{-x} dx = - \int_0^{+\infty} e^{-u^2} du = - \frac{\sqrt{\pi}}{2} \quad (38.217)$$

Nous avons alors finalement:

$$\begin{aligned} \rho_V(E) &\cong \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \exp\left(-\frac{\mu}{kT}\right) (kT)^{3/2} \exp\left(\frac{E_V}{kT}\right) \frac{\sqrt{\pi}}{2} \\ &= \frac{1}{2\pi^2} \left(\frac{2m_e kT}{\hbar^2} \right)^{3/2} \exp\left(\frac{E_V - \mu}{kT}\right) \frac{\sqrt{\pi}}{2} = \frac{1}{2^2 \pi^{3/2}} \left(\frac{2m_e kT}{\hbar^2} \right)^{3/2} \exp\left(\frac{E_V - \mu}{kT}\right) \quad (38.218) \\ &= 2 \left(\frac{m_e kT}{2\pi\hbar^2} \right)^{3/2} \exp\left(\frac{E_V - \mu}{kT}\right) \end{aligned}$$

où pour rappel, m_e est la masse de la quasi-particule (et non la masse du trou pour rappel!). Donc après intégration tout se passe comme si tous les trous étaient concentrés sur le niveau d'énergie E_V avec un nombre de places disponibles correspondant à:

$$2 \left(\frac{m_e kT}{2\pi\hbar^2} \right)^{3/2} \quad (38.219)$$

Ce que nous notons traditionnellement (et de manière un peu malheureuse... car il n'est pas évident de se rappeler qu'il s'agit d'une densité):

$$N_p = N_{p,T} \exp\left(\frac{E_V - \mu}{kT}\right) \quad (38.220)$$

ou encore:

$$N_p = N_V \exp\left(\frac{E_V - \mu}{kT}\right) \quad (38.221)$$

où nous avons environ à température ambiante (c'est le paramètre de la masse effective qui varie entre les deux) les valeurs suivantes d'états (quasi-)libres respective pour le Silicium:

$$N_V = 2 \left(\frac{m_e kT}{2\pi \hbar^2} \right)^{3/2} \cong 1.1 \cdot 10^{19} \left[cm^{-3} \right] \quad (38.222)$$

et pour le Germanium:

$$N_V = 2 \left(\frac{m_e kT}{2\pi \hbar^2} \right)^{3/2} \cong 0.5 \cdot 10^{19} \left[cm^{-3} \right] \quad (38.223)$$

6.3. BANDES D'ÉNERGIE

Les développements précédents pour les porteurs négatifs et positifs nous ont montré que dans le cadre de l'approximation d'un gaz de fermions non dégénéré, l'énergie des porteurs négatifs doit se trouver bien au-dessus du niveau de Fermi (potentiel chimique) et l'énergie des porteurs positifs bien en-dessous.

C'est donc comme s'il y avait un intervalle d'énergie interdit ou ni électrons, ni trous n'ont droit de se situer! Cet intervalle d'énergie est traditionnellement appelé "**bande d'énergie interdite**" ou plus simplement "**bande interdite**" et abrégée B.I.

L'intervalle d'énergie interdit est quant à lui souvent appelé "**gap**" et est noté E_g .

Voyons ceci sous forme schématique grossière en prenant garde au fait que ce schéma est donc quelque peu trompeur car il donne l'impression que la bande de conduction ou de valence occupe tout un bloc, alors qu'en réalité la bande de valence est constituée par la dernière couche complètement remplie, la bande d'énergie permise qui la suit étant appelée "bande de conduction".

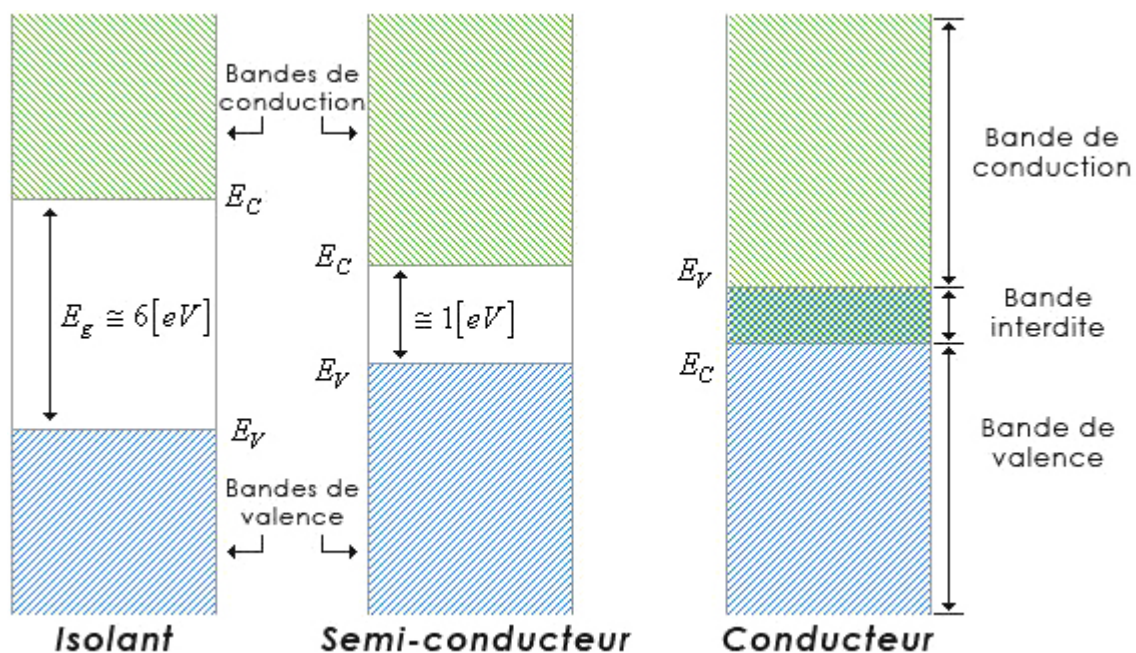


Figure: 38.3 - Représentation des structures de bandes dans différents matériaux

De plus, sachant que la chimie moléculaire permet de démontrer que des structures sont composées de multiples bandes (en fonction du premier et deuxième nombre quantique) il vient alors les définitions rigoureuses suivantes:

D1. La "**bande de conduction**" (notée BC) d'une structure solide est la bande de plus basse énergie partiellement occupée ou vide (sachant que d'autres bandes se situent au-dessus en termes énergétiques mais ne se rempliront que sous des températures élevées et n'existent que par une description théorique lorsqu'elles sont vides).

D2. La "**bande de valence**" (notée BV) d'une structure solide est la bande de plus haute énergie saturée, c'est-à-dire dont tous les états sont occupés (sachant qu'il peut y avoir en-dessous de la couche supérieure de la BV de multiples bandes en termes énergétiques et toutes saturées).

Nous avons également l'association schématique traditionnelle des bandes de conduction et de valence avec la fonction de Fermi-Dirac (qui comme déjà mentionné en toute rigueur devrait être le potentiel chimique à température non nulle!) représentée sous forme simplifiée par:

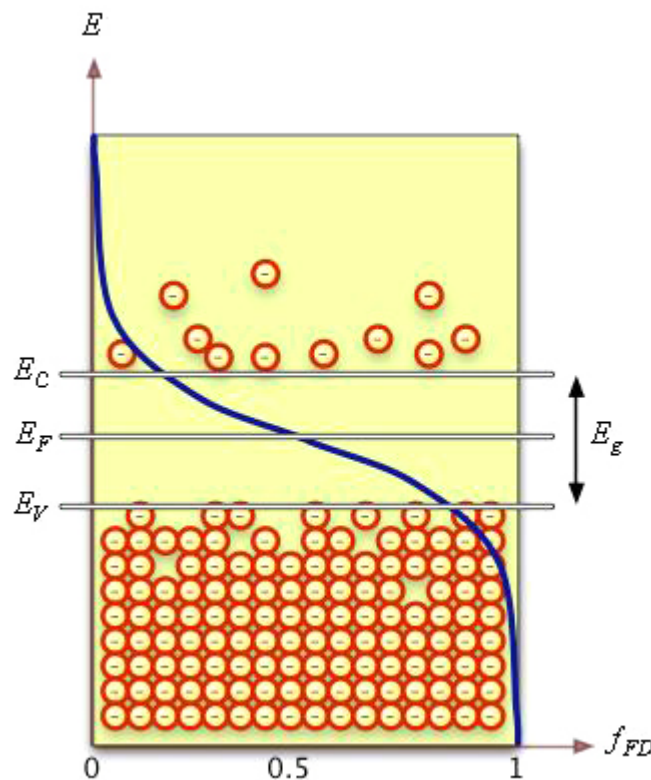


Figure: 38.4 - Association structure de bande avec fonction de Fermi-Dirac

Mais au fait cette représentation, que nous retrouvons un peu partout dans certains ouvrages est relativement erronée... puisqu'en faisant une approximation semi-classique par la loi de Maxwell-Boltzmann il n'est plus question en toute rigueur de représenter la distribution sous forme de loi de Fermi-Dirac comme l'aura remarqué le lecteur attentif! Comme quoi il faut faire attention car la représentation traditionnelle de f_{FD} dans le modèle semi-classique indiquerait qu'il y aurait des états occupés dans la bande interdite alors que si nous représentions la fonction de Maxwell-Boltzmann, nous verrions deux fonctions distinctes au-dessus et en-dessous de la bande interdite!!!

Et il faut se rappeler (!) que la figure ci-dessus (même si elle est assez fautive) représente conceptuellement un semi-conducteur non dégénéré suite aux approximations semi-classiques que nous avons faites dans nos développements en utilisant le modèle d'un gaz non dégénéré (approximation de Maxwell-Boltzmann) et qui imposait théoriquement:

$$E_C \gg \mu \gg E_V \quad (38.224)$$

que de nombreux auteurs écrivent (à nouveau c'est malheureux mais c'est ainsi...):

$$E_C \gg E_F \gg E_V \quad (38.225)$$

Donc il vient une autre définition possible du semi-conducteur non dégénéré: c'est celui où le niveau de Fermi (le potentiel chimique!) se situe dans la bande interdite et ce cas correspond au fonctionnement de la majorité des composants microélectroniques.

Remarque: Rappelons (cf. chapitre de Mécanique Statistique) que la statistique de Maxwell-Boltzmann a été bâtie en supposant l'absence d'interaction entre les particules concernées. De plus, cette statistique est construite dans le cadre de la mécanique classique et ne s'applique donc que lorsque les effets quantiques sont négligeables, par exemple à des températures suffisamment hautes!

Voici quelques valeurs expérimentales pour des semi-conducteurs courants:

E_g [eV]	300 [K]	0 [K]
C	5.47	5.51
Ge	0.66	0.75
Si	1.12	1.16
GaAs	1.43	1.53

Tableau: 38.3 - Valeurs de quelques gaps

nous comprenons alors de suite au vu de ces chiffres pourquoi le diamant, à structure cristalline et atomique quasi-identique, est isolant alors que le Silicium devient lui conducteur!

Ce qui est intéressant pour les chercheurs, c'est de combiner des matériaux afin de jouer avec la largeur de E_g en fonction des besoins!

Par ailleurs, nous pouvons aussi conclure hâtivement... que ce qui différencie isolants et semi-conducteurs c'est la largeur de leur bande interdite.

Remarquons aussi que l'énergie nécessaire à un électron pour passer de la bande de valence à la bande de conduction peut lui être fournie par un rayonnement. Dans le cas d'une absorption de lumière, l'énergie $h\nu$ d'un photon peut être suffisante pour cela tant que:

$$h\nu \geq E_g \quad (38.226)$$

À basse température, un tel processus est capable de rendre le matériau conducteur (technologie des télescopes spatiaux à basse température). Cette propriété est appelée la "**photoconductivité**".

Enfin, rappelons les deux relations obtenus plus haut:

$$\begin{aligned} \bar{n} &= \rho_C(E) = 2 \left(\frac{m_e kT}{2\pi\hbar^2} \right)^{3/2} \exp\left(\frac{\mu - E_C}{kT} \right) \\ \bar{p} &= \rho_V(E) = 2 \left(\frac{m_e kT}{2\pi\hbar^2} \right)^{3/2} \exp\left(\frac{E_V - \mu}{kT} \right) \end{aligned} \quad (38.227)$$

Le produit de ces deux densités possède une propriété très intéressante. Nous pouvons en effet remarquer qu'il est indépendant de la position du niveau de Fermi et appelé "**densité intrinsèque**":

$$\begin{aligned}\bar{n} \cdot \bar{p} &= 2 \left(\frac{m_e kT}{2\pi\hbar^2} \right)^{3/2} \exp\left(\frac{\mu - E_C}{kT}\right) \cdot 2 \left(\frac{m_e kT}{2\pi\hbar^2} \right)^{3/2} \exp\left(\frac{E_V - \mu}{kT}\right) \\ &= \left[2 \left(\frac{m_e kT}{2\pi\hbar^2} \right)^{3/2} \right]^2 \exp\left(\frac{E_V - E_C}{kT}\right) = N^2 \exp\left(\frac{E_V - E_C}{kT}\right) = N^2 \exp\left(\frac{E_g}{kT}\right)\end{aligned}\quad (38.228)$$

Par exemple, quelques valeurs de la racine carrée de la densité intrinsèque à 300 [K] sont données dans la table ci-dessous:

Ge	$2.5 \cdot 10^{13} \text{ [cm}^{-3}\text{]}$
Si	$1.6 \cdot 10^{10} \text{ [cm}^{-3}\text{]}$
GaAs	$1.1 \cdot 10^7 \text{ [cm}^{-3}\text{]}$

Tableau: 38.4 - Valeurs de densités intrinsèques

Nous remarquons aussi que la densité critique est fortement dépendant de la température. Ces valeurs de densités sont bien évidemment idéalisées, dans la réalité ces valeurs sont bien inférieures à cause des imperfections (impuretés résiduelles, défauts de cristallisation,...) qui perturbent localement la périodicité du potentiel et, de ce fait, introduisent des niveaux énergétiques qui peuvent être accessibles aux électrons. Par opposition avec les niveaux correspondant au matériau pur, nous parlerons de "**niveaux extrinsèques**".

6.4. LOI D'OHM

Nous avons démontré dans le cadre du modèle de Drude que la conductivité était donnée par:

$$\sigma = \frac{n \cdot q^2 \cdot \tau}{m} \quad (38.229)$$

où n est pour rappel la densité de porteurs dans le matériau. Nous avons également démontré que le courant est inversement proportionnel à la conductivité selon la relation:

$$U = RI = \left(\frac{1}{\sigma \frac{S}{L}} \right) I \quad (38.230)$$

Dans le cadre des développements faits plus haut nous avons vu que la densité n des porteurs était donnée respectivement par les relations suivantes à un potentiel constant (hypothèse du modèle):

$$N_n = 2 \left(\frac{m_e kT}{2\pi\hbar^2} \right)^{3/2} \exp\left(\frac{\mu - E_C}{kT}\right) \text{ et } N_p = 2 \left(\frac{m_e kT}{2\pi\hbar^2} \right)^{3/2} \exp\left(\frac{E_V - \mu}{kT}\right) \quad (38.231)$$

où les masses relatives m_e des quasi-particules (porteur négatif ou porteur positif) ne sont pas nécessairement égales! Ainsi, nous avons donc la résistance qui peut être approchée par une relation de la forme:

$$R = A \exp\left(\frac{B}{T}\right) \quad (38.232)$$

et nous vérifions aisément cette dépendance en représentant graphiquement:

$$\ln(R) = \ln(A) + \frac{B}{T} \quad (38.233)$$

soit $\ln(R)$ en fonction de $1/T$ (la résistance ne dépend donc que de la température en théorie... à tension constante).

La vraie complexité tient au fait que beaucoup de termes sont dépendants de la température (le niveau de Fermi, le temps de libre parcours moyen, etc.) et du potentiel appliqué ce qui fait que dans la réalité les courbes obtenues ne sont de loin pas conformes à la théorie....!

Une application numérique montre que les densités de porteurs N_n et N_p augmentent donc très rapidement déjà à partir de la température ambiante! Ce qui est conforme à l'expérience avec les semi-conducteurs non-dégénérés car nous aurons alors la conductivité qui augmente tout aussi fortement ce qui implique une baisse rapide de la résistance!

La grande sensibilité de la conductivité de certains solides aux variations de température est à l'origine de nombreuses applications, tant pour les métaux conducteurs que pour les semi-conducteurs. C'est ce que nous appelons des "thermistances".

Enfin, indiquons que dans le cas du Silicium, nous avons $E_g = 1.12 \text{ [eV]}$ alors que l'énergie cinétique dû à l'agitation thermique (cf. chapitre de Mécanique des Milieux Continus) est donnée à température ambiante par:

$$E_{c,th} = \frac{3kT}{2} = \frac{1}{2} m_e v_{th}^2 \cong 0.037 \text{ [eV]} \quad (38.234)$$

Or, nous avons vu précédemment que seuls les électrons dont l'énergie était voisine de celle du niveau de Fermi pouvaient participer à la conduction. Leur énergie cinétique valant alors:

$$E_{c,F} = \frac{1}{2} m_e v_F^2 \cong 1.12 \text{ [eV]} \quad (38.235)$$

où v_F est la "vitesse de Fermi".

En égalisant les deux dernières relations:

$$\frac{1}{2} m_e v_F^2 = \frac{1}{2} m_e v_{th}^2 \cong 0.037 \text{ [eV]} \quad (38.236)$$

Il y a donc un rapport d'un facteur de 30 entre les deux énergies, soit en prenant la racine carrée, un rapport 5 entre les vitesses. Nous avons donc:

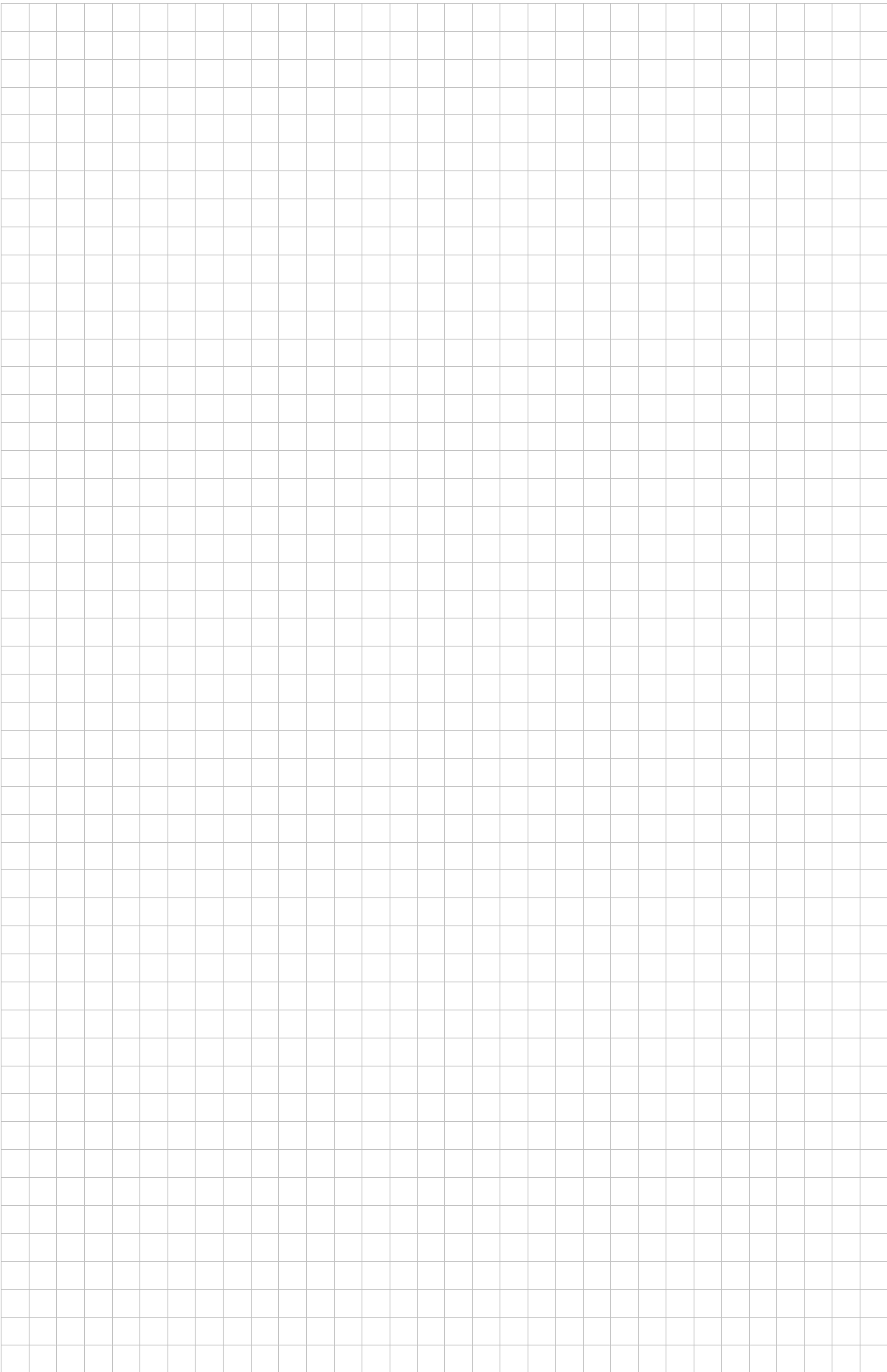
$$v_F \gg v_{th} \quad (38.237)$$

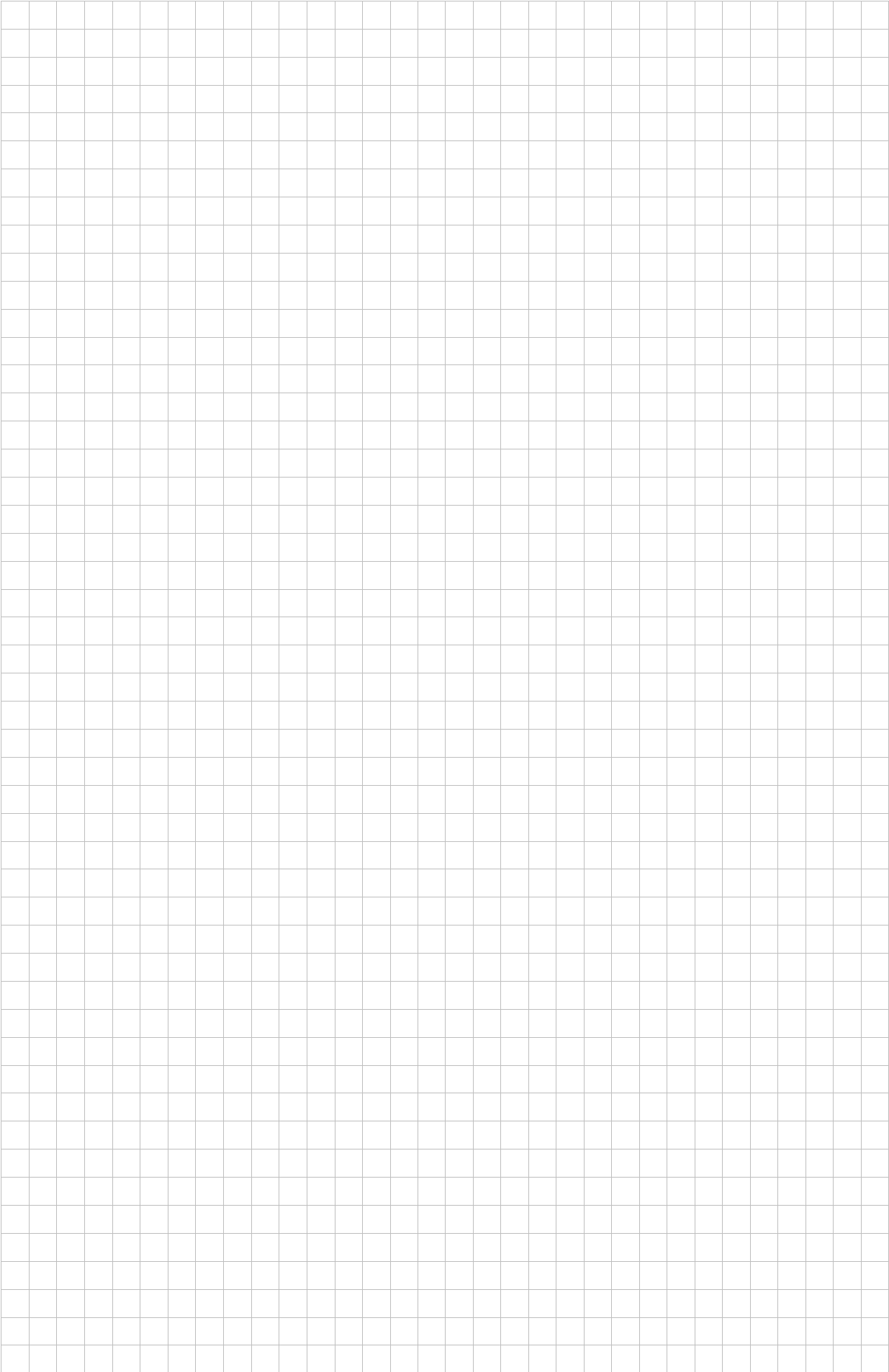
Or, nous avons déjà vu lors de notre étude du modèle de Drude que la vitesse thermique nous amenait à un libre parcours moyen supérieur d'un ordre de grandeur (facteur 10) des distances interatomiques. Et ici nous avons donc un facteur 5 en plus!!!! Soit plus de 50 distances interatomiques! Le libre parcours moyen l d'un électron de conduction est donc beaucoup plus grand que celui que nous avons déterminé à partir du modèle classique de Drude. Ainsi, le libre parcours moyen ne semble pas dû aux collisions avec les ions du réseau mais elle est imputable aux imperfections du réseau: défauts de structure, atomes étrangers...

Un semi-conducteur parfait (pur), soit sans imperfections, tel que nous l'avons traité théoriquement jusqu'à maintenant est appelé un "**semi-conducteur intrinsèque**": il ne comporte donc aucune impureté et son comportement électrique ne dépend que de la structure du matériau. Ce comportement correspond à un semi-conducteur parfait, c'est-à-dire sans défaut structurel ou impureté chimique. Un semi-conducteur réel n'est jamais parfaitement intrinsèque, mais peut parfois en être proche comme le silicium monocristallin pur.

Dans un semi-conducteur intrinsèque, les porteurs de charge ne sont créés que par excitation thermique. Le nombre d'électrons dans la bande de conduction est alors égal au nombre de trous dans la bande de valence comme nous l'a montré notre modèle théorique.

Il faut savoir qu'en réalité ces semi-conducteurs ne conduisent pas, ou très peu, le courant, excepté si nous les portons à haute température.







39. OPTIQUE GÉOMÉTRIQUE

L'optique est l'étude de la fraction de l'énergie rayonnante sensible à la rétine, c'est-à-dire la "lumière" ou dit de manière plus générale: les "ondes électromagnétiques" (cf. chapitre d'Électrodynamique) et ce dans une large bande de fréquence qui ne se limite pas (suivant les cas étudiés) à la lumière visible!

Nous avons choisi sur ce site de scinder l'étude de l'optique en trois parties: la photométrie (voir plus bas), l'optique géométrique (le présent chapitre) et l'optique ondulatoire (prochain chapitre).

1. La "photométrie" s'occupe de la partie des définitions des grandeurs relatives aux propriétés énergétiques des ondes électromagnétiques relativement à la sensibilité visuelle.
2. "L'optique géométrique" où nous décrivons la propagation de la lumière dans les milieux transparents sans faire intervenir la nature même de la lumière. Il s'agit d'une partie de la physique présentant l'avantage de ne pas demander d'outils mathématiques compliqués, mais de beaucoup de bon sens géométrique...
3. "L'optique ondulatoire" où les phénomènes lumineux sont interprétés en tenant compte de la nature de la lumière. Celle-ci est considérée comme une onde électromagnétique d'une longueur d'onde donnée définissant sa couleur (grandeur subjective comme nous le verrons plus loin).

Dans certaines expériences, nous devons cependant considérer la lumière comme un phénomène corpusculaire (cf. chapitre de Physique Quantique Ondulatoire) nous la supposons alors constituée de particules, les "photons", dont l'énergie est proportionnelle à la fréquence lumineuse selon la loi de Planck (pas celle de la thermodynamique... l'autre).

Pour des raisons de cohérence, comme nous en avons déjà fait mention, nous avons choisi de mettre la photométrie dans le chapitre d'Optique Géométrique (ici même donc...).

Avant de commencer à étudier l'aspect mathématique de l'optique géométrique, il nous a semblé judicieux d'éclaircir certaines zones floues du domaine de l'optique qui sont rarement bien précisées voire même pas traitées du tout dans les ouvrages sur le sujet. Ainsi, nous allons d'abord présenter ce qu'est une source ou une absence de lumière et ensuite comment les couleurs sont vues et traitées par l'être humain.

1. SOURCES ET OMBRES

L'expérience nous enseigne que dans un milieu homogène et transparent la lumière se propage en ligne droite et que celle-ci provient toujours de "sources lumineuses":

Certains objets sont lumineux par eux-mêmes (Soleil, flammes). Les autres objets ne sont généralement pas visibles dans l'obscurité (absence de lumière) mais s'ils sont éclairés ils renvoient tout ou partie de la lumière dans toutes les directions (voir le chapitre d'Électrodynamique et de physique quantique corpusculaire) et se comportent donc dès lors comme des sources lumineuses.

Nous définissons:

D1. Une "source ponctuelle" comme étant un seul "point lumineux"

D2. Une "source étendue" comme un ensemble de sources ponctuelles

D3. Un "rayon lumineux" comme toute droite suivant laquelle se propage la lumière

D4. Un "faisceau lumineux" comme un ensemble de rayons lumineux

D5. Le "diamètre apparent" comme étant l'angle, généralement petit, sous lequel nous voyons une des dimensions de l'objet (angle exprimé en radians).

La lumière traverse le vide sans subir d'altération. C'est ainsi que la lumière du Soleil, avant d'atteindre la limite de l'atmosphère terrestre, traverse d'immenses espaces vides sans subir de transformations.

Sur Terre, entre un objet lumineux et l'oeil qui voit cet objet, la lumière traverse une certaine épaisseur d'air. L'objet demeure visible dans d'autres gaz, ou bien à travers une lame de verre, de mica, de cellophane..., ou bien encore à travers une couche d'eau, d'alcool, de glycérine... de tels corps constituent des "milieux transparents".

La plupart des corps ne se laissent pas traverser par la lumière. Placés entre l'oeil et un objet lumineux, ils suppriment la vision de cet objet: nous disons alors qu'ils sont "corps opaques".

En fait, aucune substance n'est parfaitement transparente et la propagation dans un milieu transparent s'accompagne toujours d'un affaiblissement. Ce phénomène d'absorption dépend de la nature du milieu et augmente avec l'épaisseur de substance traversée. C'est ainsi que l'eau, même très pure, est opaque sous une épaisseur d'une centaine de mètres. Aussi les grands fonds marins ne reçoivent-ils jamais de lumière solaire.

Il arrive que certains corps, dits "corps translucides", laissent filtrer de la lumière sans permettre à l'oeil d'identifier l'objet lumineux qui l'émet. Tels sont le verre dépoli, le verre strié, la porcelaine mince, le papier huilé...

Dans un espace sombre, l'oeil situé hors du trajet de la lumière, aperçoit ce trajet grâce aux fines particules solides (poussières, fumée de tabac, brouillard,...) en suspension dans l'air. Ces particules éclairées diffusent la lumière qu'elles reçoivent, devenant autant de points lumineux qui matérialisent le volume traversé par la lumière. L'observation familière montre que ces volumes lumineux paraissent toujours limités par des lignes droites.

Nous pouvons dès lors appliquer le théorème des rapports de Thalès à certains phénomènes lumineux. Ainsi, imaginons l'expérience suivante:

Nous réalisons des sources de dimensions assez faibles pour que nous puissions les considérer comme des sources ponctuelles (c'est-à-dire des points lumineux).

Soit S une telle source ponctuelle de lumière. Considérons le volume que la source S illumine à travers une ouverture dans un diaphragme se situant dans la trajectoire de la lumière à la distance d . Si nous notons AB le diamètre circulaire de cette ouverture du diaphragme K et que nous coupons la trajectoire lumineuse par un écran E , parallèle à K et à distance D de la source, nous observerions que la partie éclairée se limite à un cercle $A'B'$.

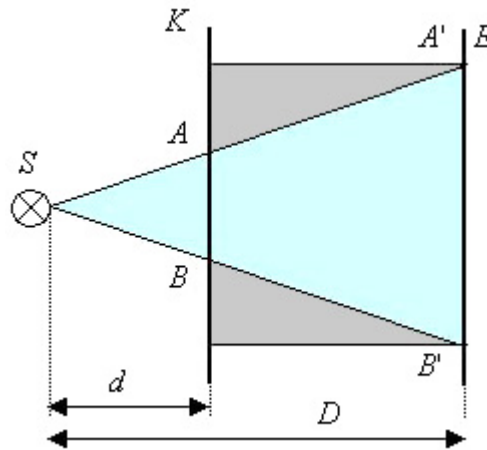


Figure: 39.1 - Application du théorème de Thalès sur des sources ponctuelles

Si nous pouvions mesurer les diamètres AB et $A'B'$ des deux cercles, ainsi que leurs distances d et D à la source, nous trouverions qu'ils satisfont au théorème des rapports de Thalès et ainsi que:

$$\frac{A'B'}{AB} = \frac{D}{d} \quad (39.1)$$

C'est également la preuve que le volume lumineux est effectivement limité par des droites issues de S et s'appuyant sur le bord de l'ouverture du diaphragme.

Ces faits d'observation et d'expérience élémentaires suggèrent l'hypothèse suivante:

Dans un milieu transparent homogène (rappelons qu'un milieu est homogène quand tous ses éléments de volume possèdent les mêmes propriétés), la lumière provenant d'un point lumineux se propage suivant des lignes droites issues de ce point. Ces droites sont appelées des "**rayons lumineux**".

Si nous revenons à la figure précédente, l'ensemble des rayons lumineux contenus dans le cône défini par la source S et le diaphragme K constitue un "**faisceau lumineux**".

1. La lumière se propageant ici à partir de S , nous disons que les rayons "**divergent**" ou encore que le faisceau est un "**faisceau divergent**".
2. Quand une source ponctuelle est à l'infini (comme l'est pratiquement une étoile, par exemple), les rayons qui en partent sont parallèles et les faisceaux qu'ils forment sont appelés "**faisceaux parallèles**", ou encore "**faisceaux cylindriques**".
3. À l'aide d'une lentille convergente (une loupe, par exemple), nous verrons qu'il est possible de changer les directions de rayons issus d'une source ponctuelle et de les faire concourir en un point S' . Un tel ensemble de rayons constitue alors un "**faisceau convergent**".

Un faisceau lumineux très étroit prend le nom de "**pinceau lumineux**". Par exemple, les rayons allant d'un point lumineux à l'œil forment toujours un pinceau lumineux très délié, parce que la distance du point observé à l'œil est nécessairement grande, comparée au diamètre de la pupille.

Si nous revenons à notre expérience avec le diaphragme: si nous diminuons l'ouverture de ce dernier qui limite un pinceau de rayons lumineux, nous observons (lorsque le diamètre est réduit à moins de quelques dixièmes de millimètre) que la trace du pinceau sur un écran E , au lieu de s'amenuiser,

s'agrandit au contraire, preuve que la lumière parvient maintenant en des points situés hors du cône $SA'B'$.

Tout se passe comme si la très petite ouverture AB devenait elle-même une source ponctuelle: nous disons que la lumière se "**diffracte**". Nous reviendrons plus tard sur cette propriété de la lumière car il s'agit d'une étude mathématique assez élaborée (cf. chapitre d'Optique Ondulatoire) et donc complexe à manipuler mais cependant fort intéressante.

Considérons maintenant une source ponctuelle de lumière. Entre la source et un écran E , interposons un corps opaque de forme quelconque. Conformément à l'hypothèse de la propagation rectiligne, nous observons un "**cône d'ombre**" limité par les rayons qui s'appuient sur le contour du corps interposé.

La région non éclairée du corps opaque est "**l'ombre propre**", celle qui correspond sur l'écran est "**l'ombre portée**".

Si la source de lumière est étendue, l'ombre portée et l'ombre propre n'ont plus leurs contours nettement délimités. Leurs bords s'entourent d'une zone intermédiaire que l'on appelle la "**pénombre**".

2. COULEUR

Définition: Nous nommons "**couleur**" la perception d'une excitation lumineuse suite à un processus neurophotochimique par l'oeil d'une ou plusieurs fréquences d'ondes lumineuses avec une (ou des) amplitude(s) donnée(s).

Remarque: Il importe de ne jamais confondre "couleur", notion perceptive, et "longueur d'onde", notion physique. Ainsi, l'oeil humain est le plus souvent incapable de distinguer un jaune monochromatique théorique (une seule longueur d'onde) d'une composition correspondante de vert et de rouge. Cette illusion permet d'afficher du jaune sur nos écrans d'ordinateur, et, plus généralement n'importe quelle couleur.

De par le fait que la partie sensible de la rétine de l'oeil humain est composée d'éléments appelés "**cônes**" sensibles chacun à un petit intervalle correspondant respectivement au rouge (via la molécule d'erythrolabe), au vert (via la molécule de chlorolabe) et au bleu (via la molécule de cyanolabe), nous pouvons créer n'importe quelle couleur en additionnant ces trois couleurs de base appelées "**couleurs fondamentales additives**" (ou "**couleurs primaires additives**"). Cela s'appelle la "**synthèse additive**" des couleurs.

L'association française de normalisation (AFNOR) a défini au 20ème siècle le principe de trivariance visuelle de la manière suivante: Un rayonnement de couleurs quelconques peut être produit visuellement à l'identique par le mélange algébrique, en proportions définies de manière unique, des flux lumineux de trois rayonnements qui peuvent être arbitrairement choisis, sous réserve qu'aucun d'entre eux ne puisse être reproduit par un mélange des deux autres.

Dans ce qui suit, nous noterons le rouge (R), le vert (V), le bleu (B), le blanc (W), le noir (N).

Couleur	Longueur d'onde [nm]	Fréquence [THz]
rouge	~ 625-740	~ 480-405
orange	~ 590-625	~ 510-480
jaune	~ 565-590	~ 530-510
vert	~ 520-565	~ 580-530
cyan	~ 500-520	~ 600-580
bleu	~ 446-500	~ 690-600
violet	~ 380-446	~ 790-690

Tableau: 39.1 - Valeurs de quelques longueurs d'onde et fréquences

Il est clair que vu les fréquences du spectre visible ce ne sera pas demain qu'avec les matériaux connus au début du 21ème siècle que nous allons construire des antennes ou paraboles capables d'émettre à de telles fréquences! Déjà que 120 [GHz] c'est un exploit alors 500 [THz] demain...

Il faut savoir que jusqu'en 1800 on ne savait pas si les couleurs se limitaient ou non à celles visibles par l'oeil humain. Ce fut avec l'apparition des thermomètres à mercure suffisamment sensibles et précis, que l'astronome Herschel en plaça un devant un spectre lumineux et trouva qu'en le promenant d'une bande de couleur à l'autre, du violet au rouge, la température s'élevait. À sa grande surprise, elle continua de s'élever lorsqu'il laissa accidentellement le thermomètre à un deux centimètres au-delà de la zone de la lumière rouge. Herschel avait détecté une lumière invisible à l'oeil humain, qualifiée plus tard de rayonnement infrarouge.

Un exemple magistral et pédagogique de ce que l'on peut voir dans les différents spectres est la nébuleuse du Crabe qui est un rémanent de supernova résultant de l'explosion d'une supernova historique (SN 1054) observée par plusieurs astronomes chinois de la dynastie Song de juillet 1054 à avril 1056:

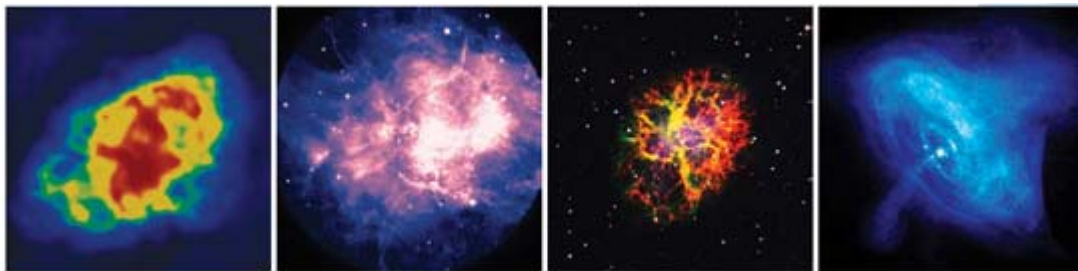


Figure: 39.2 - SN 1054 respectivement (de gauche à droite) en ondes radio, infrarouge, visible et rayons X

Remarque: Les cônes *L* de la rétine sont sensibles aux ondes longues (700 [nm]), donc les rouges. Les cônes *M*, sensibles aux ondes moyennes (545 [nm]), donc les verts. Les cônes *S*, sensibles aux ondes courtes (440 [nm]), donc les bleus. Quant au choix de cette gamme précise du spectre électromagnétique par la Nature, il suffit de regarder le spectre d'absorption de l'eau pour voir que ça tombe pile dans une fenêtre où l'eau absorbe très peu. Du coup, nous pouvons voir loin même par temps humide.

En pointant trois faisceaux lumineux (*R*, *V* et *B*) au même endroit, nous pouvons obtenir (au fait il serait plus rigoureux de dire "percevoir" car ceci est propre seulement à certains mammifères trichromates) de la lumière blanche. Nous disons alors que le blanc (dans le sens humain du terme) est la somme des trois couleurs fondamentales additives (rappelons qu'au fait le blanc est rigoureusement la somme de toutes les couleurs du spectre - donc que le blanc est constitué d'un spectre lumineux

continu). Toutes les couleurs imaginables sont obtenues en variant l'intensité de chacun des trois faisceaux. Le noir est obtenu quand nous n'envoyons aucune lumière du tout.

Par exemple, si nous additionnons (dans le sens théorique du terme: avec des composants de couleurs infiniment petits et transparents...) juste du rouge et du vert, nous obtenons du jaune (J), si nous additionnons du rouge et du bleu, nous obtenons du Magenta (M), si nous additionnons du vert et du bleu, on obtient du Cyan (C). Nous pouvons donc résumer cela par les équations suivantes:

$$\begin{aligned}RV &= R + V = J \\RB &= R + B = M \quad (39.2) \\VB &= V + B = C\end{aligned}$$

Ces trois couleurs (J , M , C) obtenues en additionnant deux couleurs fondamentales additives sont appelées "**couleurs secondaires additives**".

Schéma de la synthèse additive:

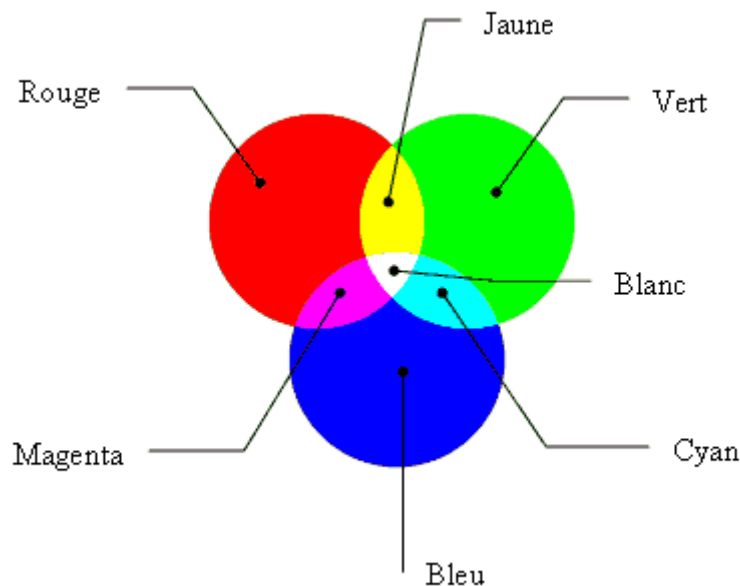


Figure: 39.3 - Représentation de la synthèse additive

L'existence de ces trois types de pigments dans les photorécepteurs des cônes sert de base physiologique au "**modèle trichromatique**" ou de "**trivariance visuelle**".

Définition: Une couleur est dite "**couleur complémentaire**" d'une autre si elles donnent du blanc quand on les additionne. Par exemple, le jaune est la couleur complémentaire du bleu:

$$J + B = (R + V) + B = W \quad (39.3)$$

À l'opposé de la synthèse additive, il existe la "**synthèse soustractive des couleurs**": c'est celle dont nous parlons quand nous enlevons de la couleur à une couleur de base. C'est par exemple le cas de l'encre ou des filtres colorés (dans le sens où il y a un support de base dont il faut traiter la couleur).

Pour comprendre de quoi il s'agit, posons un filtre rouge sur un rétroprojecteur. La lumière projetée sera rouge. Nous remarquons donc que le filtre a enlevé de la couleur à la lumière blanche: W est devenu R

mais comme $W = RVB$, cela veut dire que le filtre rouge a enlevé les couleurs VB à la lumière blanche du rétroprojecteur. Avec le même raisonnement, nous comprenons qu'un filtre V soustrait les couleurs RB et un filtre B soustrait RV .

Si nous empilons deux filtres de couleurs fondamentales différentes: par exemple, un filtre R et un filtre V , nous n'obtiendrons rien du tout, autrement dit, du N . En effet, le filtre R ne laisse passer que la lumière rouge et le filtre V soustrait cette couleur (ainsi que le B). Il ne reste donc plus aucune couleur, autrement dit du N .

Nous remarquons donc que les filtres R , V et B ne permettent pas de synthétiser différentes couleurs par soustraction puisque nous obtenons du noir dès que nous en superposons deux différents. Ce qui est très embêtant lorsque le support concerné est du papier et que l'objectif est d'imprimer quelque chose de coloré.

Il est donc plus utile d'utiliser les filtres jaunes, magenta et cyan (J , M , et C) des couleurs additives secondaires. En effet, un filtre J laisse passer du jaune, c'est-à-dire RV . Il ne soustrait donc que le B à la lumière blanche d'origine. Selon le même principe, un filtre M soustrait V et un filtre C soustrait R .

Nous remarquons alors que la superposition de deux filtres de ces couleurs secondaires donne une nouvelle couleur sur un support existant. Nous pouvons ainsi synthétiser n'importe quelle couleur en variant l'intensité de chacun des trois filtres (J , M et C) que nous superposons (sur le rétroprojecteur ou le papier par exemple). Nous appelons ces trois couleurs les "**couleurs fondamentales soustractives**".

Schéma de la synthèse soustractive:

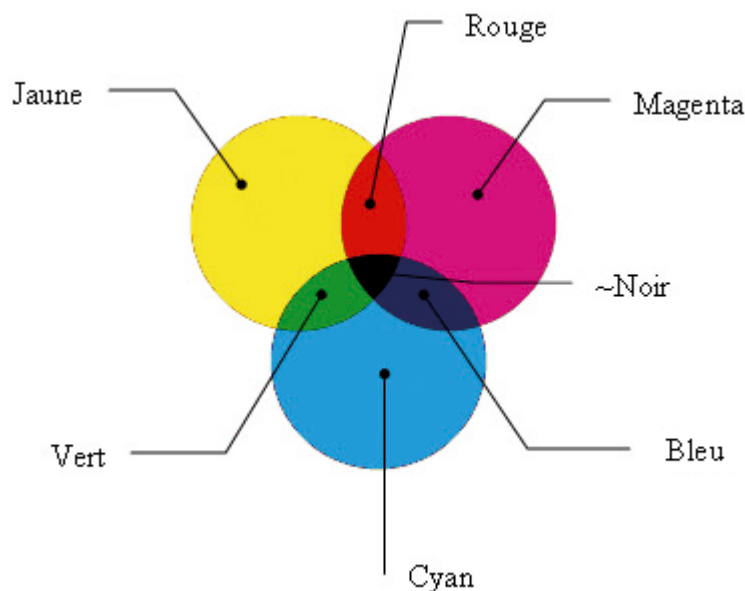


Figure: 39.4 - Représentation de la synthèse soustractive

Exemples:

E1. Un écran de télévision ou d'ordinateur fonctionne sur le principe de la synthèse additive des couleurs. En effet, en regardant l'écran à la loupe, on peut se rendre compte qu'il est rempli de petits groupes de trois luminophores (zone brillante quand on l'excite) R , V et B . Ces luminophores sont

tellement proches que quand ils s'allument ensemble, ils donnent l'impression de se confondre et on perçoit uniquement la synthèse additive des trois pixels. Par exemple, sur un écran de télévision entièrement rouge, seuls les luminophores rouges brillent. Par contre, si l'écran vire au jaune, cela veut dire que les luminophores verts brillent en même temps que les rouges.

E2. À l'opposé de la télévision, nous trouvons les procédés d'imprimerie qui fonctionnent en synthèse soustractive. En effet, la feuille est blanche et il faut lui enlever des couleurs pour obtenir celle que nous désirons. La technique est la même que celle des filtres: les encres contiennent des pigments qui filtrent certaines couleurs. En utilisant des encres J , M et C , nous pouvons obtenir toutes les couleurs du spectre visible. Toutefois, les pigments ne sont pas parfaits et le noir est très difficile à obtenir (surcharge d'encre et teinte plutôt brun foncée). Nous avons donc recours au noir comme quatrième couleur. Ce système s'appelle "**l'impression en quadrichromie**". Il est utilisé par exemple par la plupart des imprimantes couleurs et dans les rotatives de journaux.

Il est intéressant maintenant de s'intéresser aux phénomènes qui superposent les deux concepts (si nous pouvons dire...). Ainsi, un système qui projette de la couleur selon le système RVB additif ou soustractif peut lui-même être éclairé par un système équivalent. Il en résulte ainsi une superposition d'effets.

Ainsi, quand nous parlons de la couleur des objets, nous nous référons normalement à l'aspect qu'ils ont quand ils sont éclairés par de la lumière blanche.

Exemple:

Une tomate rouge, absorbe une partie de la lumière blanche W (VB) et diffuse le reste (R). C'est pour cela qu'elle nous apparaît rouge quand on l'éclaire avec de la lumière blanche. Un citron, lui, apparaît jaune car il absorbe le bleu de la lumière blanche W et diffuse le reste (RV).... Mais qu'en est-il d'une tomate éclairée par une lumière bleue? À quoi ressemble le citron si nous l'éclairons en rouge?

Nous pouvons répondre en raisonnant comme suit: comme la tomate absorbe VB et donc intrinsèquement le bleu (B), il ne reste donc rien. Elle apparaît alors noire. Quant au citron, comme il absorbe le bleu (B) et diffuse la lumière $R+V$ alors si nous l'éclairons seulement avec du rouge R il ne diffusera que du rouge et apparaîtra donc rouge.

3. PHOTOMÉTRIE

La matière est capable d'émettre, de transmettre et/ou d'absorber de l'énergie électromagnétique. Plusieurs facteurs caractérisent ce rayonnement tels que: sa gamme spectrale, son intensité, sa direction ainsi que certaines propriétés intrinsèques à la matière. La photométrie se propose de rechercher les grandeurs qui lui sont spécifiques ainsi que les lois qui les régissent.

Nous reconnaissons deux types de photométrie: la "**photométrie énergétique**" et la "**photométrie visuelle**". Dans ce qui va suivre, nous nous en tiendrons principalement à la photométrie énergétique.

Au préalable, nous devons spécifier les conditions dans lesquelles nous allons définir les nouvelles grandeurs. Nous admettrons donc les hypothèses suivantes:

H1. Le rayonnement se propage dans un milieu transparent pour toutes les intensités, les longueurs d'onde et leur polarisation.

H2. La propagation s'effectue suivant des angles solides (cf. chapitre de Trigonométrie). Nous écartons

ainsi la propagation selon des rayons parallèles.

H3. La surface élémentaire dS d'étude est suffisamment petite pour que les rayonnements de ses points soient identiques mais pas trop petits pour éviter des phénomènes comme la diffraction.

3.1. FLUX ÉNERGÉTIQUE

Définition: Le "**flux énergétique**" d'une source de rayonnement est la puissance qu'elle rayonne. Le flux Φ se mesure en Watts [W] (soit des joules par seconde [J/s]) et il découle dès lors que pour une source qui rayonne une énergie (non nécessairement constante), nous avons:

$$\Phi = \frac{dE_r(t)}{dt} \quad (39.4)$$

Dans certains domaines professionnels le flux énergétique s'exprime en unités photométriques comme étant le "**Lumen**" noté [lm] ou en unités photoniques comme un nombre de photons par seconde: [s^{-1}]. Raison pour laquelle, lorsque vous achetez des écrans ou lampes dans certains magasins, les unités ne sont pas les mêmes d'une marque à l'autre.

3.1.1. LOI DE BEER-LAMBERT

Si l'absorption et la diffusion d'un milieu peuvent être considérées comme proportionnelles à l'épaisseur dz de matière traversée, la variation de flux pourra s'écrire:

$$d\Phi = -\mu\Phi_0 dz \quad (39.5)$$

dans cette expression Φ_0 est le flux incident et μ [m^{-1}] est le "**coefficient d'atténuation linéique**" qui est fonction de la fréquence du rayonnement.

Nous aurons donc une simple équation différentielle (cf. chapitre de Calcul Différentiel Et Intégral):

$$\frac{d\Phi}{\Phi_0} = -\mu dz \Rightarrow \Phi(z) = \Phi_0 e^{-\mu z} \quad (39.6)$$

qui est la "**loi de Beer-Lambert**" (qui peut aussi s'exprimer à partir de l'intensité lumineuse que nous définirons de suite après).

Ordres de grandeur: Atmosphère $\mu_{jaune} = 10^{-5}$ [m^{-1}]; $\mu_{violet} = 4 \cdot 10^{-5}$ [m^{-1}], Verre (BK7)

$\mu_{0,55 \mu m} = 0,2$ [m^{-1}], ...

Remarque: La variation du coefficient d'absorption atmosphérique avec la longueur d'onde permet notamment d'expliquer la couleur bleue du ciel.

Il existe de nombreuses autres formulations de la loi de Beer-Lambert dont une assez utilisée en physique nucléaire (voir chapitre du même nom) dans le cadre de la radioprotection. Voyons de quoi il s'agit:

Considérons un flux Φ de particules frappant perpendiculairement la surface d'un matériau d'épaisseur

dz et de densité atomique N ($[atomes/m^3]$). Si nous considérons les particules frappant une surface S , ces dernières peuvent théoriquement rencontrer $N \cdot S \cdot dz$ atomes cibles dans cette couche. Le nombre de particules interagissant sera proportionnel à l'intensité fois ce nombre, et nous avons:

$$d\Phi \cdot S = -\sigma \cdot \Phi_0 \cdot N \cdot S \cdot dz \quad (39.7)$$

Remarques:

R1. σ est la constante de proportionnalité et est nommée "**section efficace microscopique**". Ces unités sont souvent exprimées en "**barn**" ($1 [barn] = 10^{-26} [m]$).

R2. La densité atomique N est égale à $(\rho \cdot N_{Av}) / M_{mol.}$, où ρ est la densité en kg/m^3 , $N_{Av.}$ le nombre d'Avogadro $6.022 \cdot 10^{23} [atomes/mole]$ et $M_{mol.}$ la masse molaire de la cible exprimée en $[Kg/mole]$.

Si nous admettons maintenant que les centres de diffusion sont les électrons et non pas les atomes cibles, alors il faut remplacer N par N_e où $N_e = N \cdot Z$ avec Z étant le nombre d'électrons interagissant par atome cible. D'où:

$$\Phi = \Phi_0 \exp(-\sigma N_e \cdot z) = \Phi_0 \exp\left(-\sigma \cdot Z \cdot \frac{\rho \cdot N_{Av.}}{M_{mol.}} \cdot z\right) \quad (39.8)$$

En identifiant avec la première formulation de la loi de Beer-Lambert, nous voyons que μ joue le même rôle que:

$$\mu = \sigma \cdot Z \cdot \frac{\rho \cdot N_{Av.}}{M_{mol.}} \text{ et } \sigma = \frac{\mu}{\rho} \cdot \frac{M_{mol.}}{Z \cdot N_{Av.}} \quad (39.9)$$

Et dans l'hypothèse où l'électron constitue une "**sphère d'action**" présentant une surface frontale πr_e^2 , r_e étant le rayon de cette sphère, alors:

$$\sigma = \pi r_e^2 = \frac{\mu}{\rho} \cdot \frac{M_{mol.}}{Z \cdot N_{Av.}} \quad (39.10)$$

et nous avons pour le rayon de la sphère d'action de l'électron:

$$r_e = \sqrt{\frac{\mu}{\rho} \cdot \frac{M_{mol.}}{Z \cdot N_{Av.} \cdot \pi}} \quad (39.11)$$

3.2. INTENSITÉ LUMINEUSE

Pour décrire le flux énergétique Φ d'une source, il faut commencer par le mesurer. Le capteur utilisé (thermocouple, bolomètre, cellule photoélectrique, oeil ou autres) ne peut recevoir qu'une partie: celle qui arrive dans l'angle solide $d\Omega$ défini par sa section.

Définition: "**L'intensité lumineuse**" ou "**intensité énergétique**" I d'une source ponctuelle est le flux rayonné Φ dans l'unité d'angle solide Ω centré autour d'une direction Δ d'émission:

$$I = \frac{d\Phi}{d\Omega} \quad (39.12)$$

L'intensité lumineuse est exprimée dans certains domaines professionnels en unités photométriques en "**Candela**" [Cd] ou en unités photoniques en [s^{-1}] (rappelons que les stéradians n'ont pas d'unité au même titre que les radians). Raison pour laquelle, lorsque vous achetez des écrans ou lampes dans certains magasins, les unités ne sont pas les mêmes d'une marque à l'autre.

Remarques: Une source est dite "**source anisotrope**" ou "**source directionnelle**" si son intensité varie avec la direction d'observation.

Par comparaison (car cela aide), une unité de Candela est équivalent à l'intensité d'une source dans une direction donnée, qui émet un rayonnement monochromatique de fréquence 540.1012 [Hz] (ce qui correspond approximativement à la fréquence à laquelle l'oeil est le plus sensible), et dont le flux lumineux (ou intensité) dans cette direction est 1/683 [W] par stéradian.

3.3. ÉMITTANCE ÉNERGÉTIQUE

Définition: "**L'émittance énergétique**", "**excitance**" ou encore "**éclairage**" M d'une source est le flux énergétique rayonné (puissance) par unité de surface dS en [W/m^2] dans toutes les directions de l'espace extérieur à la source et dépend des propriétés physico-chimiques de la surface émettrice:

$$M = \frac{d\Phi_{\text{em.}}}{dS} \quad (39.13)$$

Elle est souvent assimilée dans le vocabulaire courant à la "luminosité" d'une source de lumière ce qui porte parfois à confusion avec le concept d'intensité lumineuse.

L'émittance énergétique est exprimée dans de nombreux domaines professionnels en unités photométriques appelée "**Lux**" [lx] ou encore en unités photoniques [$m^{-2}s^{-2}$] ou pire encore... en [lm/m^2]. Par exemple quand vous achetez une voiture, les feux de croisement sont indiqués comme valant environ ~ 20 [lx].

Attention à ne pas confondre l'émittance énergétique avec le flux énergétique!!!

Si la source est ponctuelle et son rayonnement isotrope, sa direction n'est pas à prendre en considération. Dans le cas de ladite sphère de rayon r , l'émittance a alors pour expression:

$$M = \frac{\Phi}{4\pi r^2} \quad (39.14)$$

Dans le cas précédent de la sphère, un élément dS de la surface sphérique reçoit perpendiculairement le rayonnement. En toute généralité, une surface élémentaire peut être inclinée par rapport à la direction du rayonnement avec un angle θ . Ainsi, nous devons projeter la surface sur la perpendiculaire du rayonnement en utilisant les raisonnements élémentaires de la trigonométrie:

$$d\Phi = MdS \cos \theta \quad (39.15)$$

C'est cette projection qui explique les saisons sur la Terre: la surface balayée par l'émittance à peu près constante et isotrope du soleil (considéré comme une source ponctuelle) est maximale à l'équateur (surface perpendiculaire) et donc implique un flux supérieur par rapport à ce que reçoit une latitude supérieure ou inférieure pour laquelle la projection perpendiculaire de la surface concernée est plus petite que celle à l'équateur pour une émittance identique.

Remarques:

- R1. L'émittance énergétique n'est calculée que dans le demi-espace extérieur avant (celui d'où nous regardons la source), car seule la moitié de l'énergie échangée par les points de la surface dS est émise sous forme de rayonnement. L'autre moitié est échangée avec les atomes situés dans le corps.
- R2. L'émittance est habituellement aussi parfois notée F ou encore E . Il faudra prendre garde cependant à ne pas confondre l'émittance M avec la magnitude (notée de la même manière) que nous définissons en astrophysique.

3.4. LUMINANCE ÉNERGÉTIQUE

Soit une source non ponctuelle dont l'émittance énergétique M est connue en tout point. Un élément dS de la surface de ce genre de source sera par définition de l'intensité pas nécessairement isotrope et donc plus lumineux (puissant) lorsque l'on l'observe colinéairement au vecteur $d\vec{S}$.

L'intensité énergétique I qu'il rayonne dans une direction, formant un angle θ , avec la normale à la surface d'émission est toujours inférieure à celle rayonnée dans la direction du vecteur $d\vec{S}$. Ainsi par simple application des règles trigonométriques, nous obtenons la définition de la "**luminance**" (ou "**radiance**"):

$$L(\theta) = \frac{dI}{dS \cos \theta} = \frac{d^2\Phi}{dS d\Omega \cos \theta} \quad (39.16)$$

exprimée dans certains domaines, en unités photométriques en "**Nits**" [$\text{Cd} \cdot \text{m}^{-2}$] ou en unités photoniques en [$\text{m}^{-2} \text{s}^{-1}$]

Remarque: Lorsque nous ne nous préoccupons que de la lumière visible, la luminance d'une source est quelquefois appelée "**brillance**" ou "**éclat**" (attention ceci n'est pas le cas lorsque l'on traite de l'éclat comme il est vu en astrophysique).

Nous pouvons aussi écrire:

$$I(\theta) = \int_L L(\theta) \cos \theta dS \quad (39.17)$$

qui nous donne l'intensité énergétique que rayonne une source de luminance L dans une direction donnée.

Jean-Henri Lambert (1728-1777) a observé que l'intensité énergétique de certaines sources (parmi toutes les types de sources imaginables...) anisotropes diminue comme le cosinus de l'angle θ , autour de la direction perpendiculaire à la surface de la source:

$$I(\theta) = I_0 \cos \theta \quad (39.18)$$

Cette variation de l'intensité est observée lorsque nous mesurons l'énergie thermique rayonnée par un orifice percé dans un four (ce qui nous ramène au corps noir), isolé thermiquement et dont la température interne est supérieure à la température externe. Dans ce contexte, l'orifice est appelé un "émetteur Lambert" et ne balaye un espace que de 2π stéradians.

Remarque: Une source qui obéit à cette loi est dite "source orthotrope".

3.4.1. LOI DE LAMBERT

Une source obéit à la loi de Lambert si sa luminance énergétique est la même dans toutes les directions, c'est-à-dire que son intensité est isotrope et donc indépendante de l'angle θ .

Nous avons alors:

$$I = \frac{d\Phi}{d\Omega} = I_0 \quad (39.19)$$

Calculons l'émittance d'un émetteur Lambert:

Nous avons donc par définition même de la propriété d'un émetteur Lambert:

$$L(\theta) = \frac{dI}{dS \cos(\theta)} = \frac{d(I_0 \cos \theta)}{dS \cos \theta} = \frac{I_0}{S} \quad (39.20)$$

et nous avons:

$$\Phi = \int_{\Omega} I \cdot d\Omega = \int_{\Omega} I_0 \cos \theta \cdot d\Omega \quad (39.21)$$

Or nous avons démontré dans le chapitre de Trigonométrie, qu'un angle solide élémentaire était donné par:

$$d\Omega = 2\pi \sin \theta d\theta \quad (39.22)$$

Ce qui nous amène à écrire en utilisant les relations démontrées dans le chapitre de Trigonométrie:

$$\begin{aligned} \Phi &= \int_0^{\pi/2} I_0 \cos \theta d\Omega = \int_0^{\pi/2} I_0 \cos \theta \cdot 2\pi \sin \theta d\theta = I_0 2\pi \int_0^{\pi/2} \cos \theta \sin \theta d\theta = I_0 2\pi \int_0^{\pi/2} \frac{\sin 2\theta}{2} d\theta \\ &= I_0 \pi \int_0^{\pi/2} \sin 2\theta d\theta = I_0 \pi \left(-\frac{1}{2} \cos 2\theta \right) \Big|_0^{\pi/2} = I_0 \pi \left(\frac{1}{2} - \left(-\frac{1}{2} \right) \right) = I_0 \pi \end{aligned} \quad (39.23)$$

L'émittance valant:

$$M = \frac{\Phi}{S} \Rightarrow M = \frac{I_0 \pi}{S} = \frac{I_0}{S} \pi = L \cdot \pi \quad (39.24)$$

Ce résultat est important pour l'étude du rayonnement du corps noir, puisque la valeur de la luminance mesurée par un capteur permet de déduire l'émittance M , donc le flux énergétique de la source:

$$\Phi = S \cdot L \cdot \pi \quad (39.25)$$

Remarque: Nous parlons de la "**luminance**" d'une source et de "**l'éclairement**" d'un objet (par une source).

3.5. LOI DE KIRCHHOFF

Tout corps irradié par une source énergétique voit le flux énergétique incident se répartir selon trois termes intuitifs:

$$\Phi = \Phi_r + \Phi_t + \Phi_a \quad (39.26)$$

où:

- $\Phi_r = \rho \cdot \Phi$ est le flux énergétique géométrique réfléchi ou diffusé
- $\Phi_t = \tau \cdot \Phi$ est le flux énergétique qui traverse le corps sans interactions (transparence intégrale)
- $\Phi_a = \alpha \cdot \Phi$ est absorbé et transformé sous d'autres formes d'énergie

Les trois coefficients appelés respectivement "**facteur de réflexion**" ρ , "**facteur de transmission**" τ et "**facteur d'absorption**" α , dépendent de la longueur d'onde λ de la lumière incidente et de la température du corps récepteur.

Pour chaque objet, nous avons bien évidemment:

$$\rho + \tau + \alpha = 1 \quad (39.27)$$

qui est l'expression de la "**loi de Kirchhoff simple**" (contrairement à la version différentielle) en photométrie.

Remarque: En physique, nous retrouvons souvent des énoncés de conservation sous la dénomination "loi de Kirchhoff" comme en électrocinétique par exemple.

3.6. DÉCOMPOSITION SPECTRALE

De ce qui vient d'être dit, il découle que toutes les grandeurs définies précédemment peuvent être rapportées à leur décomposition spectrale en longueur d'onde. Ceci résulte du principe de superposition: tout rayonnement peut être traité comme la superposition de rayonnements monochromatiques.

Ainsi, nous définissons:

$$\Phi = \int_{\lambda} \Phi(\lambda) d\lambda \quad I = \int_{\lambda} I(\lambda) d\lambda \quad L = \int_{\lambda} L(\lambda) d\lambda \quad M = \int_{\lambda} M(\lambda) d\lambda \quad (39.28)$$

et de même:

$$\alpha = \int_{\lambda} \alpha(\lambda) d\lambda \quad \rho = \int_{\lambda} \rho(\lambda) d\lambda \quad \tau = \int_{\lambda} \tau(\lambda) d\lambda \quad (39.29)$$

Remarque: Les unités du flux spectral (ou "décomposé"), intensité spectrale (ou "décomposée"), luminance spectrale (ou "décomposée") ou émittance spectrale (ou "décomposée") ainsi que les facteurs d'absorption spectrale (ou "décomposée"), de réflexion spectrale (ou "décomposée") et de transmission spectrale (ou "décomposée") ne sont bien sûr pas équivalentes à leur expression intégrée au niveau dimensionnel.

Nous aurons un grand besoin de la densité de l'émittance lors de l'étude du corps noir dans le chapitre de Thermodynamique de la section de Mécanique. Rappelez-vous uniquement que nous avons en unités S.I. sur le principe de décomposition (et inversement superposition) spectrale:

$$M \left[\frac{W}{m^2} \right] \quad M(\lambda) \left[\frac{W}{m^3} \right] \quad (39.30)$$

Remarque: Nous avons vu en thermodynamique que les paramètres définis ci-dessus, étant dépendants de la longueur d'onde, sont également dépendants de la température qui émet ces mêmes ondes.

4. LOI DE RÉFRACTION

Pierre de Fermat proposa que les rayons lumineux (ondes électromagnétiques) répondaient à un principe très général selon lequel le chemin emprunté par la lumière pour se rendre d'un point donné à un autre était celui pour lequel le temps de parcours était minimum (en fait un extremum qui peut être un minimum ou un maximum). Cette proposition, appelée "**principe de Fermat**", à la base de l'optique géométrique s'appuie sur le principe de moindre action (principe que nous avons déjà introduit dans le chapitre de Mécanique Analytique) ce que nous démontrerons plus loin.

Avant de commencer les développements, donnons quelques définitions importantes donc certaines sont basées sur la figure ci-dessous:

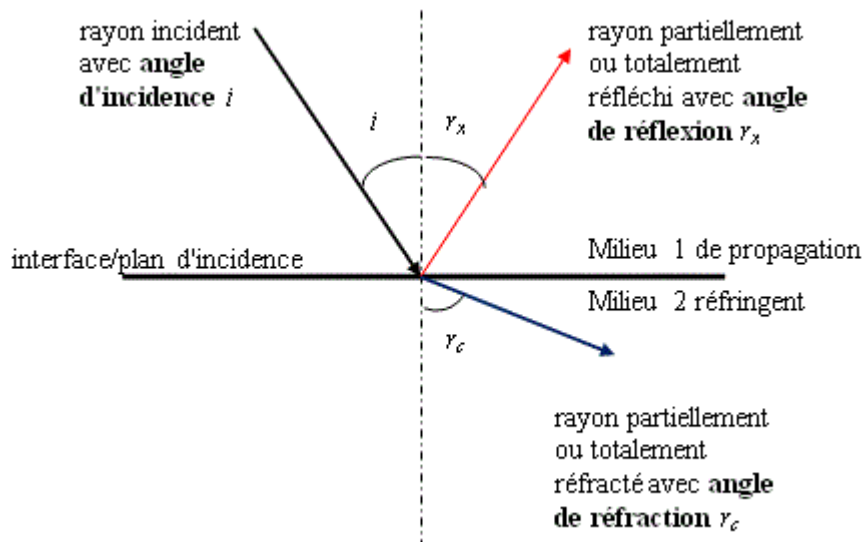


Figure: 39.5 - Vocabulaire pour l'étude de l'optique géométrique

Définitions:

D1. Un "**milieu réfringent**" est un milieu qui cause la déviation d'un rayon lumineux incident.

D2. Le "**rayon incident**" est le rayon lumineux qui se propageant dans un milieu 1, passe totalement ou partiellement dans un milieu réfringent 2, le reste étant absorbé ou partiellement réfléchi.

D3. "**L'angle d'incidence**", parfois noté i , est l'angle par lequel le rayon incident pénètre dans le milieu réfringent.

D4. Le "**rayon partiellement ou totalement réfléchi**" est la partie du rayon lumineux qui ayant rencontré l'interface séparant le milieu de propagation du milieu réfringent, continue son parcours dans le milieu de propagation.

D5. "**L'angle de réflexion**", parfois noté r_x ou simplement r s'il n'y a pas de confusion possible, est l'angle par lequel le rayon est réfléchi par rapport au plan représentant l'interface entre le milieu de propagation et lui-même. Nous démontrerons que les angles incidents et réfléchis sont égaux en valeurs absolues.

D6. Le "**rayon partiellement ou totalement réfracté**" est la partie du rayon lumineux qui ayant rencontré l'interface séparant le milieu de propagation du milieu réfringent, continue son parcours dans le milieu réfringent.

D7. "**L'angle de réfraction**", parfois noté r_c ou simplement r s'il n'y a pas de confusion possible, est l'angle par lequel le rayon est réfracté par rapport au plan représentant l'interface entre le milieu de propagation et le milieu réfringent. Les angles incidents et réfractés sont liés par une relation que nous démontrerons plus loin.

D8. "**L'indice de réfraction absolu**" d'un milieu à une longueur d'onde λ donnée (et donc de fréquence ν) mesure le facteur de réduction de la vitesse de phase de la lumière dans le milieu par rapport au vide (la plus grande qui soit) et est donné en toute généralité par la "**loi de Cauchy**" (la seule démonstration mathématique que j'ai eu entre les mains à ce jour partait des équations de Maxwell et tenait sur

environ 3 pages A4, mais elle était basée sur tellement de bricolages successifs que l'on va passer outre et admettre qu'elle ne peut être établie qu'expérimentalement):

$$n_{\lambda} = A + \frac{B}{\lambda^2} = A + \frac{Bv^2}{c^2} \quad (39.31)$$

où A et B sont des constantes établies expérimentalement. Nous pouvons remarquer à travers la loi de Cauchy que l'indice de réfraction absolu diminue lorsque la longueur d'onde augmente (in extenso lorsque la fréquence diminue).

Tous les matériaux possèdent un indice de réfraction absolu, d'une valeur positive et supérieure à 1. Plus un milieu est dense, plus la vitesse de phase de la lumière est ralentie, plus l'indice de réfraction absolu est élevé.

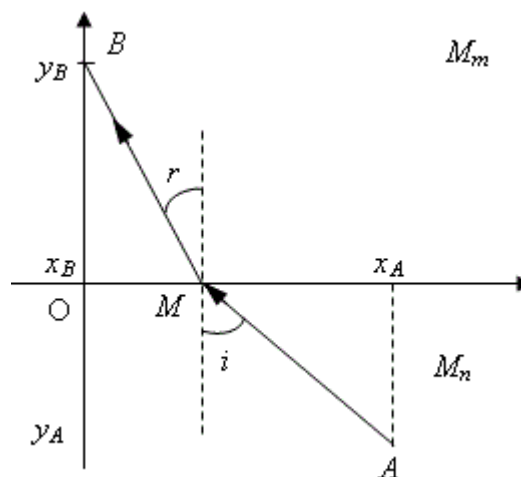
Considérons (voir figure ci-dessous) maintenant deux milieux M_n et M_m d'indices de réfraction respectifs n et m (implicitement dépendant de la longueur d'onde) et dont la surface de contact est plane. Prenons deux points A et B situés respectivement dans le milieu d'indice n (le point A) et dans le milieu d'indice m (le point B).

Considérons le chemin de la lumière allant de A à B . Le principe de Fermat nous enseigne que le chemin emprunté par la lumière est tel que le temps mis pour le parcourir est minimum. Nous nous proposons dans un premier temps d'appliquer une méthode classique pour calculer le chemin du rayon lumineux et dans un second temps, nous montrerons que le principe de Fermat peut être énoncé comme un principe variationnel.

Choisissons un repère qui simplifie le problème: faisons passer l'axe des abscisses par le plan de contact des deux milieux et l'axe des ordonnées par le point B . Dans un tel repère, les points A et B ont les coordonnées suivantes: $A(x_A, y_A)$, $B(0, y_B)$.

Appelons $M(x, 0)$, le point où le rayon lumineux traverse la surface de contact entre les deux milieux. Le temps T mis par la lumière pour aller de A à B est alors:

$$T = \frac{AM}{v_n} + \frac{MB}{v_m} = \frac{n \cdot AM}{c} + \frac{m \cdot MB}{c} \quad (39.32)$$



où:

$$v_n = \frac{c}{n} \text{ et } v_m = \frac{c}{m} \quad (39.33)$$

sont les vitesses de phase de la lumière dans les milieux M_n et M_m .

Nous pouvons observer sur la figure ci-dessus que les rayons incidents sont réfractés de l'autre côté de l'axe perpendiculaire à l'interface. Ceci est une caractéristique type des matériaux ayant un indice de réfraction absolu positif. Mais il est possible physiquement de construire depuis les années 1990 des "métamatériaux" composites artificiels à indice de réfraction absolu négatif.

L'écriture des deux relations précédentes:

$$v_n = \frac{c}{n} \text{ et } v_m = \frac{c}{m} \quad (39.34)$$

se justifie par le fait que nous pouvons nous permettre de faire l'hypothèse que la vitesse de phase de la lumière ne croît pas en traversant un corps dense mais se voit divisée par un facteur donné dépendant du milieu qu'elle traverse. Pour s'en convaincre, il suffit d'imaginer un cas absurde où la lumière traverserait sans perte de vitesse un corps de densité infinie!

En développant les valeurs de AM et MB nous obtenons la dépendance suivante de T en fonction de la position x de M :

$$T(x) = \frac{n \cdot \sqrt{(x_A - x)^2 + y_A^2}}{c} + \frac{m \cdot \sqrt{x^2 + y_B^2}}{c} \quad (39.35)$$

Selon le principe de Fermat, le chemin emprunté par la lumière est celui pour lequel T est minimum. L'extremum de $T(x)$ est atteint lorsque sa dérivée par rapport à x est nulle.

$$\frac{dT}{dx} = \frac{n}{c} \cdot \frac{(x - x_A)}{\sqrt{(x_A - x)^2 + y_A^2}} + \frac{m}{c} \cdot \frac{x}{\sqrt{x^2 + y_B^2}} \quad (39.36)$$

Notons que:

$$\frac{(x - x_A)}{\sqrt{(x_A - x)^2 + y_A^2}} = \frac{(x - x_A)}{AM} = -\sin(i) \text{ et } \frac{x}{\sqrt{x^2 + y_B^2}} = \frac{x}{MB} = \sin(r) \quad (39.37)$$

où pour rappel, r est "l'angle de réfraction" (à ne pas confondre avec "l'angle de réflexion"!) et i "l'angle d'incidence"

de la lumière allant de A à B .

La condition d'un temps extremum mis par la lumière s'exprime alors:

$$\frac{dT}{dx} = 0 \Leftrightarrow -\frac{n}{c} \sin(i) + \frac{m}{c} \sin(r) = 0 \quad (39.38)$$

D'où nous tirons la relation, connue sous le nom de "**loi de Snell-Descartes**" (qui n'est plus une loi puisque démontrée):

$$n \cdot \sin(i) = m \cdot \sin(r) \quad (39.39)$$

Il suffit que les angles d'incidence et de réfraction remplissent cette condition pour que le chemin parcouru par la lumière soit effectivement celui qui prend le moins de temps.

Nous notons plus fréquemment la loi de Snell-Descartes en physique de la manière suivante:

$$n_{12} = \frac{n}{m} = \frac{n_1}{n_2} = \frac{\sin(r)}{\sin(i)} \quad (39.40)$$

où n_{12} est "**l'indice de réfraction relatif**" du milieu 2 par rapport au milieu 1 qui ont respectivement leur propre "**indice de réfraction absolu**" n_1, n_2 . Ainsi, nous voyons bien à travers cette relation que l'angle incident est en valeur absolue obligatoirement égal à l'angle réfléchi.

Remarques:

R1. Nous verrons lors de notre étude de l'optique ondulatoire que nous pouvons retrouver (démontrer) cette même relation mais sans les hypothèses de bases de l'optique géométrique. Dès lors, cette dernière relation est appelée "relation de Descartes-Snellius" ou plus simplement "loi de Snell".

R2. Quand nous parlons de l'indice de réfraction relatif d'un milieu m sans faire référence à un autre milieu, le milieu implicite est le vide.

R3. Certains matériaux n'ont pas un indice de réfraction absolu isotrope: il dépend alors de la direction de propagation et l'état de polarisation de la lumière. Cette propriété porte le nom de "**biréfringence**".

Étudions maintenant la relation entre l'indice de réfraction relatif et la vitesse de phase de la lumière dans les différents milieux qu'elle traverse:

Un rayon lumineux relie deux points A_1 et A_2 situés de part et d'autre de S . Ce rayon n'est pas représenté dans la figure. Ne sont tracés que trajets situés de part et d'autre du rayon qui réalise l'extremum (nous nous basons sur l'étude du trajet maximum maintenant). Par hypothèse, ils sont extrêmement proches, si bien que la distance $\overline{PQ}$ est très faible:

$$\overline{PQ} \ll \overline{A_1 A_2} \quad (39.41)$$

Nous admettons qu'ils correspondent au même temps de parcours.

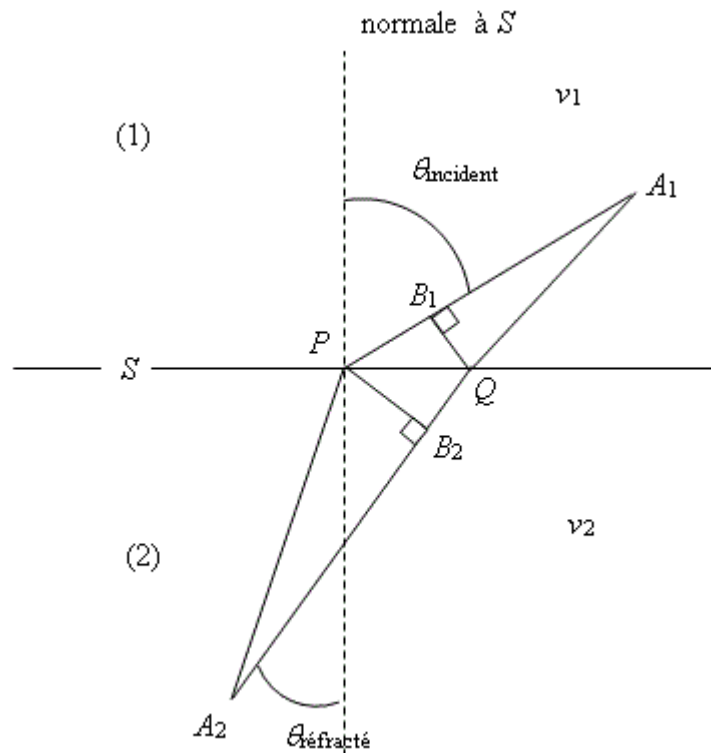


Figure: 39.6 - Figure permettant de mettre en relation vitesse de phase et indice de réfraction

Puisque les deux trajets sont très proches, nous pouvons admettre l'égalité des distances $\overline{A_1B_1}$ et $\overline{A_1Q}$ d'une part, $\overline{PA_2}$ et $\overline{B_2A_2}$ de l'autre. Ainsi, par hypothèse:

$$\frac{\overline{PB_1}}{v_1} = \frac{\overline{QB_2}}{v_2} \quad (39.42)$$

Mais, sous la même hypothèse:

$$\begin{aligned} \overline{PB_1} &= \overline{PQ} \cos(\pi/2 - \theta_i) = \overline{PQ} \sin(\theta_i) \\ \overline{QB_2} &= \overline{PQ} \cos(\pi/2 - \theta_r) = \overline{PQ} \sin(\theta_r) \end{aligned} \quad (39.43)$$

si bien que:

$$\frac{\sin \theta_r}{\sin \theta_i} = \frac{\sin(r)}{\sin(i)} = \frac{v_2}{v_1} \quad (39.44)$$

La "loi de la réfraction" s'énonce finalement en général:

$$n_{12} = \frac{v_2}{v_1} = \frac{n_1}{n_2} = \frac{\sin(r)}{\sin(i)} \quad (39.45)$$

Quant à l'angle de réflexion, comme nous l'avons déjà précisé, celui-ci reste donc égal à l'angle d'incidence si la surface de réflexion est parfaitement régulière et plate.

Remarque: Si nous considérons l'écriture suivante:

$$\sin(i) = \frac{n_2}{n_1} \sin(r) \quad (39.46)$$

et le cas où $n_1 > n_2$ (par exemple passage de l'eau vers l'air). Alors, pour des valeurs proches de 1, c'est-à-dire pour des incidences rasantes (rayon incident proche de la surface), la loi de Snell-Descartes donne une valeur supérieure à 1. Nous sortons alors du domaine de validité de la loi. Cela correspond à des situations où il n'y a pas de réfraction mais uniquement de la réflexion, nous parlons alors de "réflexion totale".

Le principe de Fermat présente donc d'évidentes similitudes avec le principe de moindre action en cela qu'il consiste en un principe du minimum. Bien qu'une description rigoureuse de la lumière nécessite l'introduction de la physique quantique, il est toutefois possible de l'appréhender par le biais de la mécanique analytique et de lui appliquer, sous certaines conditions, le principe de moindre action. Nous allons montrer que nous retrouvons ainsi le principe de Fermat.

Les calculs que nous allons présenter, introduisent de nombreuses hypothèses hasardeuses mais en tout état de cause, ce procédé doit être considéré comme une approximation. À noter que le principe de Fermat procède lui aussi d'une même approximation que nous pouvons qualifier de "limite classique".

Imaginons que la lumière est composée de "grains" matériels. Il faut alors admettre que ces grains obéissent à des propriétés physiques plutôt singulières: leur masse est nulle puisque selon la description classique, les rayons lumineux ne sont pas déviés par le champ gravitationnel. Cette absence de masse les rend donc insensibles au champ gravitationnel terrestre (attention ! nous sommes dans une description "classique").

Écrivons l'action pour l'un de ces grains de lumière:

$$S = \int (T - V) dt \quad (39.47)$$

Or, en supposant que le seul champ de potentiel V présent est celui qui dérive du champ gravitationnel et que nous admettons la lumière comme y étant insensible (nous savons en relativité générale que cela est faux mais nous avons précisé tout à l'heure que nous ferions des approximations!), il s'ensuit que l'action de la lumière peut s'écrire:

$$S = \int T dt \quad (39.48)$$

Or, aucune force ne s'applique sur la lumière, par conséquent l'énergie cinétique T est une constante du mouvement. Appliquons le principe variationnel de moindre action:

$$\delta S = 0 = \delta \left(\int T dt \right) = T \cdot \delta \int dt \quad (39.49)$$

D'où nous tirons:

$$\delta \int dt = 0 \quad (39.50)$$

Cette équation signifie que le temps mis par la lumière le long de sa trajectoire est minimum (ou plus généralement, est un extremum). Nous retrouvons le principe de Fermat. Nous avons donc montré, qu'à la limite classique et sous certaines hypothèses, le principe de Fermat découle directement du principe de moindre action.

4.1. EFFET TCHERENKOV (CERENKOV)

Nous avons vu dans les paragraphes précédents l'hypothèse (relativement intuitive) que la vitesse de phase de propagation de la lumière dans un milieu d'indice de réfraction absolu n n'était pas égale à c mais toujours inférieure en écrivant cela:

$$v_n = \frac{c}{n} \quad (39.51)$$

L'effet Tcherenkov est (basiquement) un phénomène similaire à une onde de choc (en acoustique), produisant un flash de lumière, et qui a lieu sur le trajet d'une particule chargée se déplaçant dans un milieu avec une vitesse de phase supérieure à la vitesse de la lumière du milieu (l'explication rigoureuse sort du cadre d'étude de ce site de par sa complexité de traitement!).

Effectivement, rappelons d'abord que nous avons vu dans le chapitre d'Électrodynamique que toute particule chargée en mouvement émettait une radiation électromagnétique. Ensuite, nous avons vu dans les paragraphes précédents que la vitesse de la lumière dans un milieu donné dépendait de l'indice de réfraction absolu de ce milieu (hypothèse qui se vérifie par la justesse expérimentale des développements théoriques qui en découlent).

Remarques:

R1. C'est cet effet qui provoque la luminosité bleue de l'eau entourant le coeur d'un réacteur nucléaire.

R2. Parfois certains se demandent pourquoi les particules chargées peuvent aller plus vite que la lumière dans un milieu autre que le vide. C'est simple au fait: même si les deux particules rencontrent à peu près les mêmes obstacles et difficultés à se propager le photon ne peut être accéléré par une impulsion alors qu'une particule chargée peut se voir être accélérée par un phénomène donné dans un milieu donné.

Nous avons donc deux données de bases. La vitesse de la particule chargée qui peut s'écrire sous la forme suivante avec les notations relativistes:

$$v = \beta c \quad (39.52)$$

et la vitesse de phase de la lumière dans un milieu avec un indice de réfraction absolu donné:

$$v_e = \frac{c}{n} \quad (39.53)$$

Il est facile de voir que pour obtenir $v > v_c$ il faut avoir:

$$\beta c > \frac{c}{n} \quad (39.54)$$

Soit:

$$\beta > 1/n \quad (39.55)$$

Certains auteurs préfèrent comparer la distance parcourue par la lumière par rapport à celle parcourue par la particule. Il vient ainsi:

$$\frac{d_l}{d_p} = \frac{\left(\frac{c}{n}\right) \cdot \Delta t}{\beta c \cdot \Delta t} = \frac{1}{\beta \cdot n} \quad (39.56)$$

Et donc pour que la particule parcoure des distances égales à celles de la lumière dans le même temps il faut que $\beta = 1/n$. Au-delà, apparaît l'effet Tcherenkov.

5. FORMULES DE DESCARTES

Nous avons discuté précédemment certains phénomènes qui se produisent lorsqu'un front d'onde passe d'un milieu à un autre dans lequel la propagation est différente. Non seulement nous avons analysé ce que devient le front d'onde, mais encore nous avons introduit le concept de "rayon" qui est particulièrement utile pour les constructions géométriques. Nous nous proposons maintenant d'approfondir les phénomènes de réfraction et de réflexion d'un point de vue géométrique en utilisant le concept de rayon comme l'outil permettant de décrire les processus qui prennent place aux surfaces de discontinuité de la propagation. Nous admettrons également que les processus se limitent à des réflexions et réfractions, aucune autre modification n'affectant les surfaces d'onde.

Ce traitement géométrique est correct tant que les surfaces et les discontinuités rencontrées par l'onde au cours de sa propagation sont très grandes devant la longueur d'onde. Tant que cette condition est remplie, le traitement s'applique aussi bien aux ondes lumineuses, acoustiques (en particulier ultrasonores - très hautes fréquences), sismiques, etc.

Nous commençons par considérer la réflexion des ondes sur une surface sphérique. Nous devons d'abord établir certaines définitions. Le centre de courbure C (cf. chapitre de Géométrie Différentielle) est le centre de la surface sphérique de la figure ci-dessous et le sommet O est le pôle de la calotte sphérique.

Définition: La droite passant par O et C est appelée "**axe optique**".

Si nous prenons O pour origine des coordonnées, toutes les quantités mesurées à droite de O seront prises comme positives, toutes celles à gauche comme négatives!!!

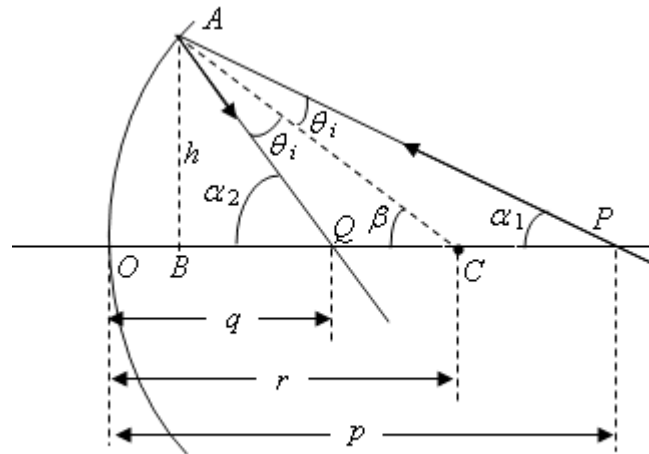


Figure: 39.7 - Représentation du concept d'axe optique

Supposons que le point P soit une source d'ondes sphériques. Le rayon $\overline{PA}$ donne par réflexion le rayon $\overline{AQ}$ et, comme les angles d'incidence et de réflexion sont égaux par rapport à la perpendiculaire AC de la surface (comme nous l'avons déjà fait remarquer lors de notre étude de la réfraction), nous voyons sur la figure que:

$$\beta = \theta_i + \alpha_1 \text{ et } \alpha_2 = \beta + \theta_i \quad (39.57)$$

d'où:

$$\alpha_1 + \alpha_2 = 2\beta \quad (39.58)$$

En admettant que les angles α_1, α_2 et β sont très petits, c'est-à-dire que les rayons sont "para-axiaux" et que la source est très distante ou que le détecteur est très petit par rapport à la source, nous pouvons écrire avec une bonne approximation avec un développement de Maclaurin (cf. chapitre sur les Suites Et Séries) pour de petits angles:

$$\begin{aligned} \alpha_1 &\cong \tan(\alpha_1) = \frac{\overline{AB}}{\overline{BP}} \cong \frac{h}{p} \\ \alpha_2 &\cong \tan(\alpha_2) = \frac{\overline{AB}}{\overline{BQ}} \cong \frac{h}{q} \\ \beta &\cong \tan(\beta) = \frac{\overline{AB}}{\overline{BC}} \cong \frac{h}{r} \end{aligned} \quad (39.59)$$

En substituant ces valeurs approximatives de α_1, α_2 et β dans $\alpha_1 + \alpha_2 = 2\beta$, nous obtenons:

$$\boxed{\frac{1}{p} + \frac{1}{q} = \frac{2}{r}} \quad (39.60)$$

qui est la "formule de Descartes pour la réflexion sur une surface sphérique concave". Elle implique, dans l'approximation utilisée pour l'établir, que pour tous les rayons incidents passant par P passeront par Q après réflexion sur la surface. Nous pouvons alors dire que Q est "l'image de l'objet" P .

Dans le cas particulier où le rayon incident est parallèle à l'axe optique, ce qui équivaut à placer l'objet à une très grande distance de la lentille, nous avons $p = \infty$. La formule de Descartes pour la réflexion sur une surface sphérique concave devient alors:

$$\frac{1}{q} = \frac{2}{r} \quad (39.61)$$

et l'image se forme au point F appelé "**foyer**", et sa distance de la lentille donnée par:

$$f = q = \frac{r}{2} \quad (39.62)$$

est appelée "**distance focale**". Nous obtenons aussi le rapport $r / 2$ si nous faisons tendre q vers l'infini.

La relation obtenue précédemment est également valable pour une surface convexe. Effectivement, il suffit de tirer les traits représentant les rayons lumineux au-delà de la surface concave pour voir que l'objet d'étude est le même à une symétrie près:

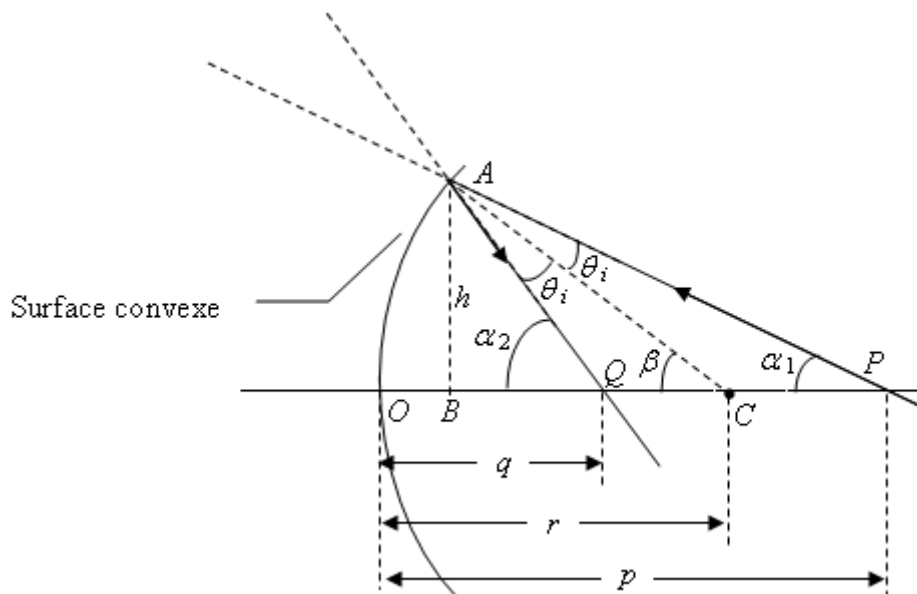


Figure: 39.8 - Principe de l'extension à une surface convexe

La seule différence entre la surface concave et convexe tient au fait que dans le cas de la surface convexe, l'image de l'objet réfléchi apparaît comme s'il semblait être derrière la surface (à l'équivalent du point P). Ceci nous amène à définir la terminologie suivante:

Définitions:

D1. Une "**image virtuelle**" est un terme utilisé en optique pour désigner toute image qui se forme avant la face de sortie d'un instrument d'optique (dans le sens de parcours de la lumière) et ne peut donc pas être visualisée sur un écran. Pour une lentille mince convergente un objet placé entre le foyer objet et le centre optique de la lentille donnera une image virtuelle. C'est notamment le cas d'un système optique utilisé comme loupe, qui permet d'obtenir une image agrandie de l'objet observé à travers la lentille.

D2. Une "**image réelle**" est un terme utilisé en optique pour désigner toute image qui se forme après la face de sortie d'un instrument d'optique (dans le sens de parcours de la lumière). Pour une lentille mince

convergente un objet placé après le foyer objet de la lentille donnera une image réelle.

Remarque: Si l'ouverture du miroir est grande, de telle sorte qu'il reçoive des rayons fortement inclinés la formule de Descartes que nous avons précédemment déterminée n'est plus, nous le savons, une bonne approximation. Il n'y a plus dans ce cas une image ponctuelle bien définie d'un "point objet", mais un nombre infini d'entre elles: en conséquence l'image d'un objet de grandes dimensions apparaît floue puisque les images se superposent. Cet effet porte le nom "**d'aberration de sphéricité**" et la partie de l'axe optique qui contient l'ensemble des images réfléchies s'appelle alors la "**caustique par réflexion**". L'aberration de sphéricité ne peut pas être éliminée, mais un dessin approprié de la surface permet de la supprimer pour certaines positions sur l'axe optique appelées "**stigmatiques**". Par exemple, dans notre cas d'étude précédent, il est évident (par construction géométrique) que si nous posons P en C , alors le point C devient alors le point stigmatique. Nous disons alors qu'il est le point "**rigoureusement stigmatique**".

Par contre, pour le miroir parabolique tous les rayons convergent vers le foyer du miroir où est concentrée l'énergie lumineuse reçue par le miroir. Réciproquement, nous plaçons le filament d'une lampe au foyer d'un miroir parabolique pour obtenir des projecteurs de grande portée (typique des phares ne possédant pas de lentille de Fresnel). Nous donnons aussi une forme parabolique aux antennes de réception des ondes hertziennes. Pour la télévision diffusée par des satellites comme on travaille en ondes centimétriques (fréquence de quelques GHz) une distance focale de l'ordre du mètre est convenable pour l'antenne (in extenso cela s'applique aux télescopes et radiotélescopes).

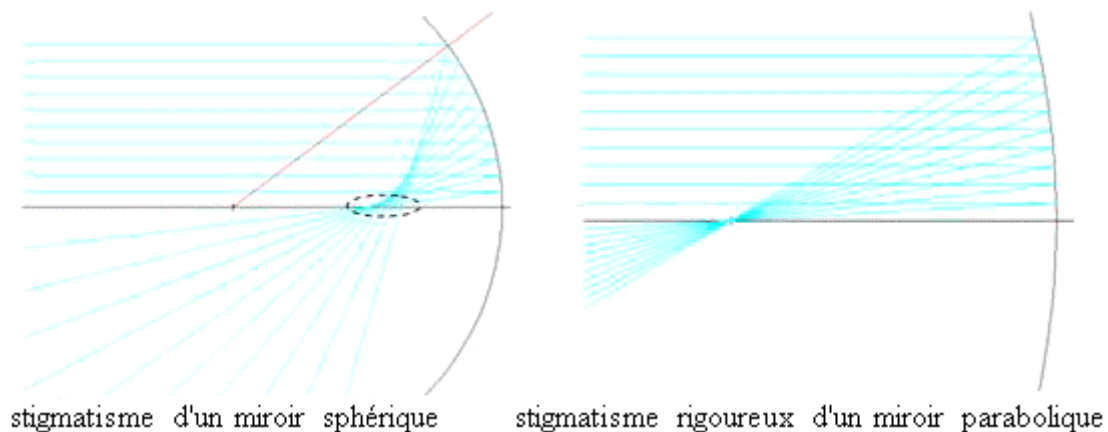


Figure: 39.9 - Représentation du concept de stigmatisme

L'idée pour démontrer que le foyer de la parabole est le point stigmatique rigoureux est la suivante:

Reprenons le schéma que nous avons utilisé lors de notre étude des coniques dans le chapitre de Géométrie Analytique:

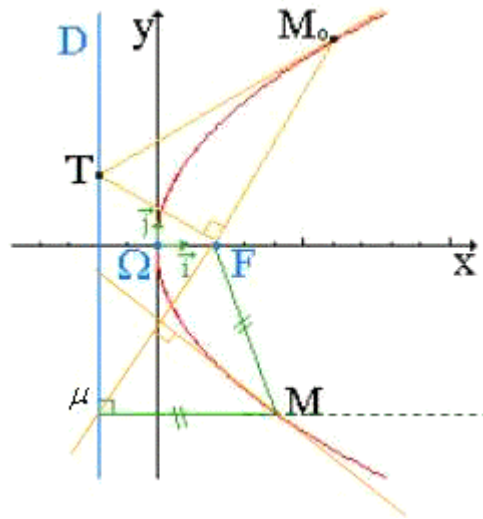


Figure: 39.10 - Schéma général des propriétés de la parabole

Nous y avons rajouté le point μ qui est la projection orthogonale du point M (point d'incidence du rayon lumineux) ainsi que la tangente à la parabole au point M . Si nous arrivons à démontrer que la tangente à M est la médiatrice du segment $\overline{F\mu}$, alors nous démontrons également que l'angle d'incidence et de réflexion sont bien égaux.

Prenons l'équation:

$$y^2 - 2hx = 0 \quad (39.63)$$

d'une parabole de paramètre h (cf. chapitre de Géométrie Analytique) rapportée à un repère principal $(\vec{i}, \vec{j})$. Le foyer a donc pour coordonnées $F(h/2, 0)$ et la directrice a pour équation:

$$x + h/2 = 0 \quad (39.64)$$

Nous obtenons l'équation de la tangente en M_0 par la dérivée en ce même point (attention... rappelez-vous de l'orientation particulière de la parabole!):

$$\frac{(x - x_0)}{(y - y_0)} = \left(\frac{y^2}{2h} \right)'_{y_0} = \frac{2y_0}{2h} = \frac{y_0}{h} \Rightarrow (x - x_0) - (y - y_0) \frac{y_0}{h} = 0 \quad (39.65)$$

Ce qui s'écrit encore:

$$hx - hx_0 - y_0 y + y_0^2 = 0 \Leftrightarrow hx_0 - hx - y_0^2 + y_0 y = 0 \quad (39.66)$$

et en sachant que:

$$y_0^2 = 2hx_0 \quad (39.67)$$

nous obtenons donc l'équation de la tangente:

$$hx_0 - hx - y_0^2 + y_0 y = hx_0 - hx - 2hx_0 + y_0 y = y_0 y - h(x + x_0) = 0 \quad (39.68)$$

Un des vecteurs directeurs de la tangente est donc alors (cf. chapitre de Géométrie Analytique):

$$\vec{v} = y_0 \vec{i} + p \vec{j} \quad (39.69)$$

où dans le cas d'une parabole, p est égal à h .

D'autre part, nous avons (cela se vérifie facilement en posant $x_0 = 0$):

$$\overrightarrow{MF} = \left(\frac{h}{2} - x_0 \right) \vec{i} - y_0 \vec{j} \text{ et } \overrightarrow{M\mu} = - \left(\frac{h}{2} + x_0 \right) \vec{i} \quad (39.70)$$

Nous avons donc le produit scalaire:

$$\vec{v} \circ \overrightarrow{MF} = -y_0 \left(x_0 + \frac{h}{2} \right) = \vec{v} \circ \overrightarrow{M\mu} \quad (39.71)$$

comme les vecteurs $\overrightarrow{MF}$ et $\overrightarrow{M\mu}$ ont même norme d'après la définition de la parabole, nous en déduisons que le vecteur $\vec{v}$ (directeur de la tangente) dirige la bissectrice de l'angle des vecteurs $\overrightarrow{MF}$ et $\overrightarrow{M\mu}$ et donc par extension que la tangente à M est bien la médiatrice de $F\mu$.

Avant d'étudier le grandissement des lentilles convexes sphériques, intéressons nous de manière générale à la définition de ce qu'est un grandissement. Considérons d'abord la figure suivante:

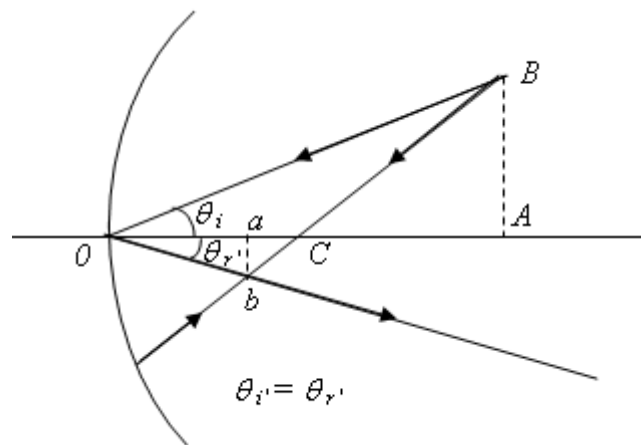


Figure: 39.11 - Principe de base du grandissement

où pour rappel le centre de courbure C (cf. chapitre de Géométrie Différentielle) est le centre de courbure de la surface sphérique de la figure ci-dessous et le sommet O est le pôle de la calotte sphérique.

Ainsi, le "**grandissement**" M d'un système optique quelconque est défini comme le rapport de la grandeur de l'image ab à celle de l'objet réel AB , c'est-à-dire:

$$M = \frac{ab}{AB} \quad (39.72)$$

Nous voyons d'après la figure ci-dessus que:

$$\tan \theta_i = \frac{\overline{AB}}{\overline{OA}} = \frac{AB}{p} \quad (39.73)$$

$$\tan \theta_r' = \frac{ab}{Oa} = \frac{ab}{q}$$

Nous avons donc, en tant compte de ce que $\theta_i = \theta_r'$:

$$\frac{\tan \theta_r'}{\tan \theta_i} = 1 = \frac{ab}{q} \frac{p}{AB} \quad (39.74)$$

d'où:

$$M = \frac{ab}{AB} = \frac{q}{p} \quad (39.75)$$

Faisons maintenant une étude équivalente à celle effectuée précédemment, ayant les mêmes propriétés de symétrie et les défauts, mais sur les "**dioptres sphériques**" (résultats intéressants pour ce qui est de l'étude de l'oeil). Les résultats vont être utiles avant d'aborder la lentille convexe sphérique (la loupe traditionnelle).

Nous allons donc considérer la réfraction au passage d'une surface sphérique séparant deux milieux d'indices de réfraction absolus n_1 et n_2 (voir figure ci-dessous).

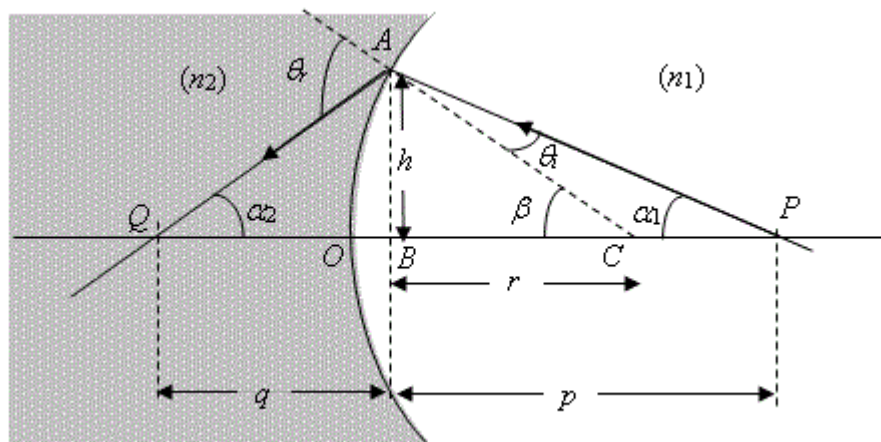


Figure: 39.12 - Concept de dioptre sphérique

où pour rappel le centre de courbure C (cf. chapitre de Géométrie Différentielle) est le centre de la surface sphérique de la figure ci-dessous et le sommet O est le pôle de la calotte sphérique.

Les éléments géométriques fondamentaux sont les mêmes que ceux définis pour les surfaces sphériques. Nous considérons donc dans un premier temps un dioptre concave et observant que la "**distance objet**" est située à l'opposé des autres points, nous devons opter pour une convention de signe pour mettre cette observation en évidence dans les équations. Ainsi, q sera défini comme une valeur négative.

Un rayon incident tel que PA est réfracté suivant AQ et coupe donc l'axe optique en Q . Nous observons sur la figure que:

$$\beta = \theta_i + \alpha_1 \text{ et } \beta = \theta_r - \alpha_2 \quad (39.76)$$

Nous avons d'après la loi de Snell-Descartes:

$$\frac{n_1}{\sin \theta_r} = \frac{n_2}{\sin \theta_i}$$

et nous admettrons comme pour les surfaces sphériques que les rayons sont peu inclinés. Dans ces conditions les angles $\theta_i, \theta_r, \alpha_1, \alpha_2$ et β sont très petits et nous pouvons écrire à l'aide des développements en série de Maclaurin (cf. chapitre Suites et Séries):

$$\sin \theta_i \cong \theta_i \text{ et } \sin \theta_r \cong \theta_r \quad (39.77)$$

de sorte que la loi de Snell-Descartes s'écrit:

$$n_1 (\beta - \alpha_1) = n_2 (\beta + \alpha_2) \quad (39.78)$$

D'après la figure, nous pouvons faire les approximations:

$$\begin{aligned} \tan \alpha_1 &\cong \alpha_1 \cong \frac{h}{p} \\ \tan \alpha_2 &\cong \alpha_2 \cong \frac{h}{q} \\ \tan \beta &\cong \beta \cong \frac{h}{r} \end{aligned} \quad (39.79)$$

de sorte qu'en substituant dans l'approximation de la loi de Snell-Descartes nous trouvons après simplification élémentaire:

$$n_1 \left(\frac{h}{r} - \frac{h}{p} \right) = n_2 \left(\frac{h}{r} + \frac{h}{q} \right) \Leftrightarrow n_1 \left(\frac{1}{r} - \frac{1}{p} \right) = n_2 \left(\frac{1}{r} + \frac{1}{q} \right) \quad (39.80)$$

d'où pour une surface concave:

$$\boxed{\frac{n_1}{p} + \frac{n_2}{q} = \frac{n_1 - n_2}{r}} \quad (39.81)$$

qui constitue la "**formule de Descartes pour la réfraction au passage d'une surface sphérique**" où q est donc un nombre négatif (puisque à gauche de l'origine O).

Bien que la dernière relation ait été démontrée dans le cas d'une surface concave, elle reste valable pour une surface convexe en tenant compte alors de ce que r est négatif à son tour et dès lors:

$$\boxed{\frac{n_1}{p} + \frac{n_2}{q} = \frac{n_2 - n_1}{r}} \quad (39.82)$$

où q est toujours négatif (puisque à gauche de l'origine O).

Le "**foyer objet**" F_0 appelé également "**premier point focal**" d'une surface sphérique réfringente est la position d'un point objet de l'axe optique tel que les rayons réfractés soient parallèles à l'axe optique, ce qui revient à former l'image du point à l'infini, où $q = \infty$.

La distance de l'objet à la surface sphérique est appelée alors "**distance focale objet**", et nous la désignons par f_0 . En posant $p = f_0$ et $q = \infty$. Nous avons alors pour le cas concave:

$$\frac{n_1}{f_0} = \frac{n_1 - n_2}{r} \Rightarrow f_0 = \left(\frac{n_1}{n_1 - n_2} \right) r \quad (39.83)$$

La distance focale f_0 est positive et le système dit "**convergent**" quand le foyer objet est réel, placé devant la surface sphérique. Quand le foyer objet est virtuel la distance focale f_0 est négative et le système est dit "**divergent**".

De même, si les rayons incidents sont parallèles à l'axe optique, ce qui revient à avoir un objet très éloigné de la surface sphérique $p = \infty$, les rayons réfractés passent par un point F_i de l'axe optique appelé "**foyer image**" ou "**second point focal**" (avec à nouveau les mêmes problèmes de stigmatisme).

Dans ce cas la distance de la surface sphérique à l'image est appelée "**distance focale image**" et nous la désignons par f_i . En posant $p = \infty$ et $q = f_i$ nous avons alors pour le cas concave:

$$\frac{n_2}{f_i} = \frac{n_1 - n_2}{r} \Rightarrow f_i = \left(\frac{n_2}{n_1 - n_2} \right) r \quad (39.84)$$

Maintenant intéressons-nous au type de surfaces réfléchissantes et réfractantes que nous attendons: les lentilles!

Une lentille est donc par définition un milieu transparent limité par deux surfaces courbes (généralement sphériques), bien que l'une des faces d'une lentille puisse être plane. Une onde incidente subit donc deux réfractions à la traversée de la lentille. Admettons pour simplifier que les milieux de part et d'autre de la lentille soient identiques et leur indice de réfraction absolu égal à 1 (l'air ou le vide par exemple). Nous ne considérerons également que des lentilles minces, c'est-à-dire dont l'épaisseur est très petite devant les rayons de courbure:

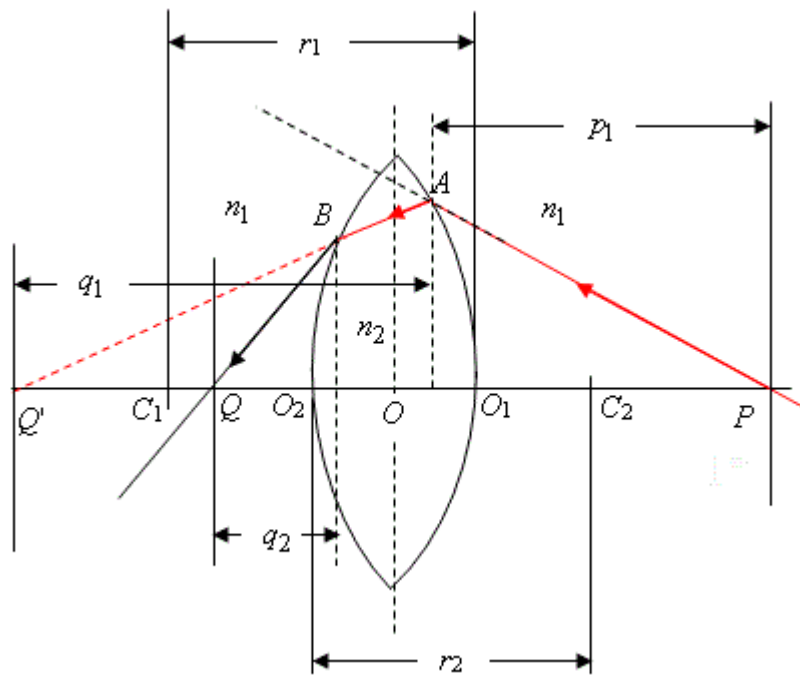


Figure: 39.13 - Représentation d'une lentille

L'axe optique est la droite déterminée par les deux centres C_1, C_2 . Nous cherchons à déterminer une relation qui lie la position de P et Q à partir de paramètres physiques facilement mesurables!

Nous considérerons pour l'analyse que l'image formée après réfraction sur la première surface est l'objet pour la réfraction sur la seconde surface.

Considérons le rayon incident PA passant par P . Au passage de la première surface, le rayon incident est réfracté suivant le rayon AB et continue virtuellement jusqu'à Q' conformément au comportement d'une surface sphérique convexe. Il nous faut donc appliquer la relation démontrée plus haut pour le dioptré convexe:

$$\frac{n_1}{p_1} + \frac{n_2}{q_1} = \frac{n_2 - n_1}{r_1} \quad (39.85)$$

En B le rayon subit une deuxième réfraction et devient le rayon BQ conformément au comportement d'une surface sphérique concave. Nous pouvons imaginer que le rayon incident en B provient d'un point P' virtuel (non représentable sur la figure ci-dessus) se trouvant sur l'axe optique et plongé dans le matériau virtuellement étendu vers la droite de la lentille (c'est le côté difficile de cette démonstration... il faut se l'imaginer!).

Il nous faut donc appliquer la relation démontrée plus haut pour le dioptré concave mais en prenant garde cette fois-ci à l'ordre des indices de réfraction absolus (piège subtil!) et en imaginant que le point:

$$\frac{n_2}{p_2} + \frac{n_1}{q_2} = \frac{n_1 - n_2}{r_2} \quad (39.86)$$

Comme nous considérons que la lentille est entourée d'air (indice de réfraction unitaire), nous avons alors:

$$\frac{1}{p_1} + \frac{n}{q_1} = \frac{n-1}{r_1} \quad \frac{n}{p_2} + \frac{1}{q_2} = \frac{1-n}{r_2} \quad (39.87)$$

Nous allons faire maintenant l'hypothèse que l'épaisseur de la lentille tend vers zéro. En d'autres termes que ses deux rayons de courbure tendent vers l'infini. Nous avons alors:

$$\left. \begin{array}{l} \frac{1}{p_1} + \frac{n}{q_1} = \frac{n-1}{r_1} \underset{r_1 \rightarrow +\infty}{=} 0 \\ \frac{n}{p_2} + \frac{1}{q_2} = \frac{1-n}{r_2} \underset{r_2 \rightarrow +\infty}{=} 0 \end{array} \right\} \frac{1}{p_1} + \frac{n}{q_1} = \frac{n}{p_2} + \frac{1}{q_2} \quad (39.88)$$

En identifiant terme à terme et en se rappelant que ce qui est à gauche de l'origine est négatif, nous avons alors:

$$p_2 \cong -q_1 \quad (39.89)$$

tout en faisant subtilement abstraction des deux autres termes... (on comprend aisément pourquoi cette démonstration est souvent omise dans la littérature...).

Dès lors, les deux relations antérieures deviennent:

$$\frac{1}{p_1} + \frac{n}{q_1} = \frac{n-1}{r_1} \quad \frac{n}{-q_1} + \frac{1}{q_2} \cong \frac{1-n}{r_2} \quad (39.90)$$

En sommant il vient au final:

$$\frac{1}{p_1} + \frac{1}{q_2} \cong (n-1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right) \quad (39.91)$$

et qui est souvent notée sous la forme suivante, appelée "**première formule de Descartes pour les lentilles minces**" ou "**équation des lentilles minces**":

$$\boxed{\frac{1}{p} + \frac{1}{q} \cong (n-1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right)} \quad (39.92)$$

Tout en prenant bien garde à réadapter la schématique:

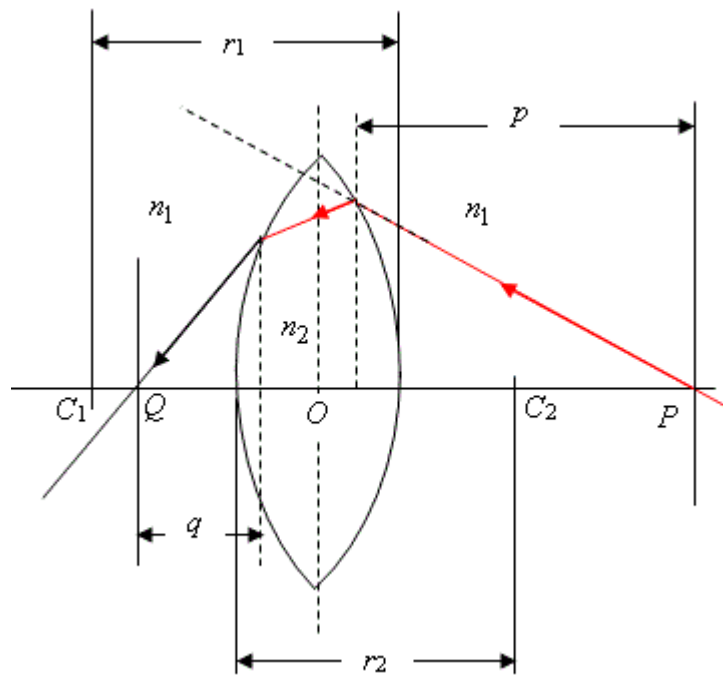


Figure: 39.14 - Lentille mince épurée avec la notation traditionnelle

En écrivant cette équation, il convient d'appliquer à r_1, r_2 la convention des signes que nous avons fixée, c'est-à-dire que les rayons sont positifs pour une surface concave et négatifs pour une surface convexe, vue du côté sur lequel la lumière vient frapper la lentille. Ainsi, si les deux rayons sont les mêmes, nous avons:

$$\frac{1}{p} + \frac{1}{q} \cong (n-1) \left(\frac{1}{r} - \frac{1}{(-r)} \right) = (n-1) \frac{2}{r} \quad (39.93)$$

Le terme à droite de l'équation des lentilles minces est une constante propre uniquement aux caractéristiques physiques de la lentille qu'il d'usage d'appeler "**puissance dioptrique**" et dont l'unité est le "**dioptre**" et de noter:

$$\boxed{\frac{1}{p} + \frac{1}{q} \cong \mathcal{D}} \quad (39.94)$$

Le point O dans la figure précédente, est choisi de façon à coïncider avec le "**centre optique**" de la lentille. Le centre optique a pour propriété d'être un point tel que tout rayon passant par lui sort parallèlement à la direction du rayon incident!! C'est une propriété importante car tout point d'un objet se situant d'un côté de la lentille (peu importe lequel par symétrie) va émettre de la lumière dont certains rayons vont passer par le centre optique. Ce qui permet donc d'avoir des triangles semblables à gauche et à droite de l'axe de symétrie de la lentille et d'appliquer Thalès (cf. chapitre de Géométrie Euclidienne) pour calculer le grandissement.

Pour montrer qu'un tel point existe, considérons, dans la lentille ci-dessous (à symétrie horizontale et verticale):

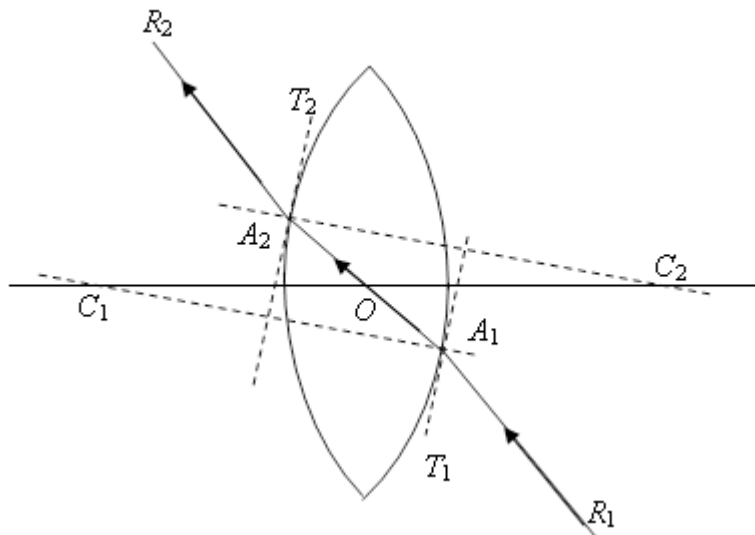


Figure: 39.15 - Figure représentant la propriété du centre optique d'une lentille

Considérons les deux rayons de courbure parallèles $\overline{C_1A_1} = r_1$, $\overline{A_2C_2} = -r_2$ générateurs des dioptries (éléments de la lentille sphérique mince) choisis tels que les plans tangents correspondants T_1 et T_2 sont par construction aussi parallèles.

Pour le rayon $\overline{R_1A_1}$, dont la direction est telle qu'il se réfracte suivant $\overline{A_1A_2}$, le rayon émergent est $\overline{A_2R_2}$ et parallèle à $\overline{A_1R_1}$ de par la symétrie horizontale de la lentille. Ainsi, les triangles C_1A_1O et C_2A_2O étant semblables quels que soient les "rayons générateurs", nous voyons ainsi que la position du centre optique O est satisfaite par la relation:

$$\frac{\overline{C_1O}}{\overline{OC_2}} = \frac{\overline{C_1A_1}}{\overline{A_2C_2}} = -\frac{r_1}{r_2} \quad (39.95)$$

et existe donc indépendamment des rayons générateurs.

Comme dans le cas d'un simple dioptré, le "foyer objet" F_0 , ou "premier point focal d'une lentille" est la position de l'objet pour laquelle les rayons émergent parallèlement à l'axe optique ($q = \infty$) après avoir traversé la lentille. La distance de la lentille au foyer objet est alors appelée "distance focale objet" nous la désignons dans la pratique souvent par la lettre f .

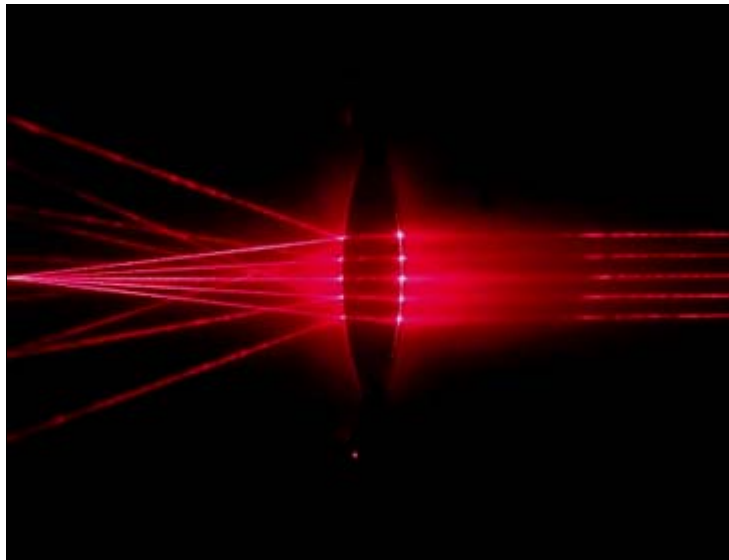


Figure: 39.16 - Premier point focal (source) est ici à gauche de la lentille

En posant alors $p = f$ et $q = \infty$ dans l'équation des lentilles minces sphériques:

$$\frac{1}{p} + \frac{1}{q} = (n - 1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right)$$

nous obtenons la distance focale objet sous la forme:

$$\frac{1}{f} = (n - 1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right) \quad (39.96)$$

De même dans le cas d'un simple dioptré, le "**foyer image**" F_i , ou "**deuxième point focal d'une lentille**" est l'endroit où convergent les rayons lumineux après avoir traversé la lentille mais qui étaient avant la lentille parallèles entre eux et avec l'axe optique ($p = \infty$). Ainsi, étant donné la symétrie centrale des lentilles minces sphériques il suffit de façon imaginaire d'inverser la photo précédente pour visualiser le concept:

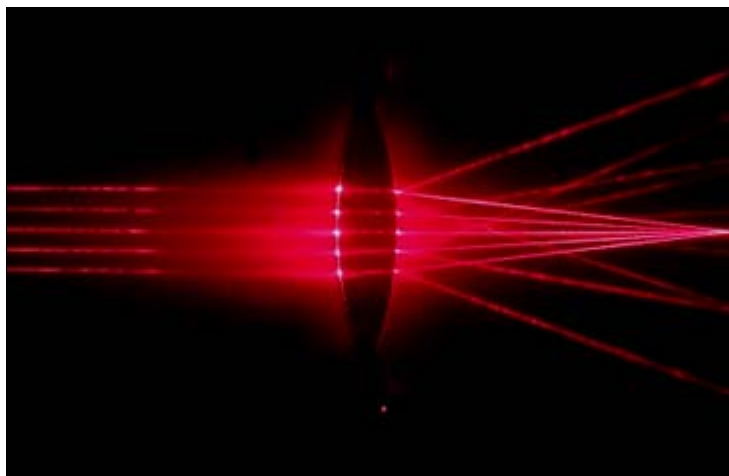


Figure: 39.17 - Deuxième point focal (cible) est ici à droite de la lentille

La distance de la lentille au foyer image est alors appelée "**distance focale image**" nous la désignons dans la pratique par la même lettre f car par symétrie de la lentille mince, de la lentille, en posant

$q = f$ et $p = \infty$ nous avons:

$$\frac{1}{f} = (n-1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right) \quad (39.97)$$

Par conséquent, dans une lentille mince les deux foyers sont placés symétriquement de chaque côté!

Donc comme dans les deux l'inverse de la focale est égale à la puissance dioptrique (indépendant de p et q) rien ne nous empêche alors d'écrire l'équation des lentilles minces comme on l'a retrouve souvent dans les livres scolaires:

$$\boxed{\frac{1}{p} + \frac{1}{q} = \frac{1}{f} = (n-1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right) = \mathcal{D}} \quad (39.98)$$

Sous cette forme on l'appelle alors "**l'équation des opticiens**" ou encore "**équation des lunetiers**". Sous la forme simple suivante qui ne fait pas apparaître les propriétés physique de la lentille (et qui est souvent la relation vue dans les petites classes raison pour laquelle il semblerait qu'elle ait un nom différent):

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{f} \quad (39.99)$$

on l'appelle "**équation de conjugaison**" ou "**deuxième formule de Descartes pour les lentilles minces**".

Par ailleurs, si la distance focale est positive, et donc respectivement la puissance dioptrique aussi, alors la lentille est dite "**lentille convergente**":

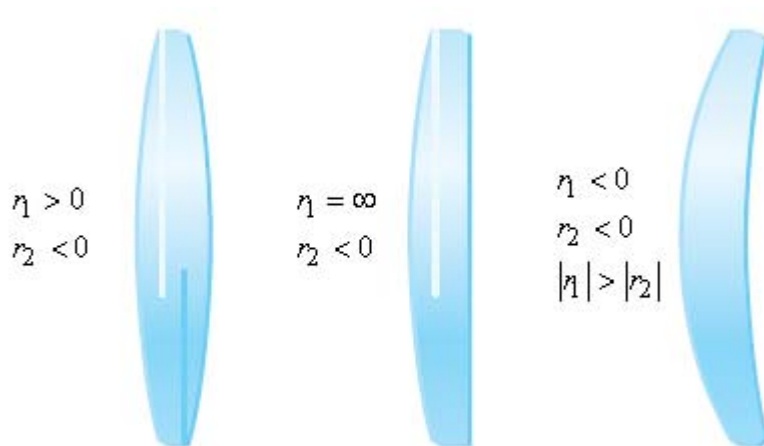


Figure: 39.18 - Exemples de lentilles convergentes (biconvexe, plan-convexe, convexe ménisque)

si la distance focale est négative, et donc respectivement la puissance dioptrique aussi, elle est dite "**lentille divergente**":

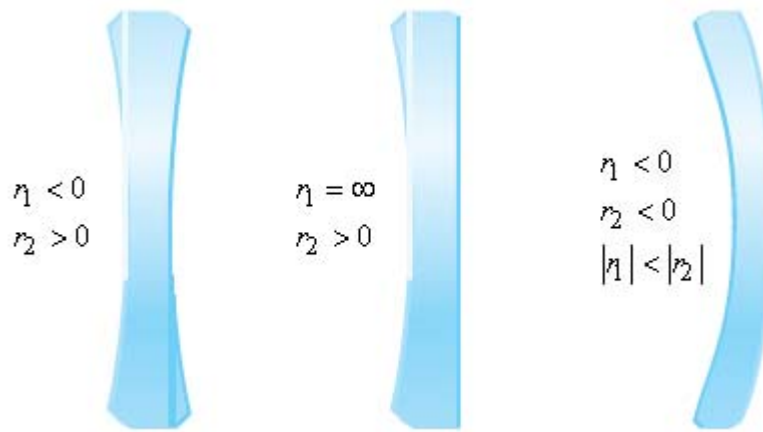


Figure: 39.19 - Exemples de lentilles divergentes (biconcave, plan-concave, concave ménisque)

Revenons maintenant sur la définition du grandissement pour tout système optique qui était naturellement pour rappel le rapport de la grandeur de l'image (réelle ou virtuelle) ab à celle de l'objet réel AB , c'est-à-dire:

$$M = \frac{ab}{AB} \quad (39.100)$$

Donc dans le cas d'une lentille sphérique mince symétrique dont les rayons passent par le centre optique nous obtenons et qui décrivent des triangles semblables de chaque côté de l'axe de symétrie de la lentille mince sphérique nous avons alors en appliquant Thalès:

$$M = \frac{ab}{AB} = \frac{q}{p} \quad (39.101)$$

Soit le même résultat que celui obtenu déjà plus haut.

Exemple:

Les deux faces d'une lentille biconvexe ont un rayon de 3 [cm]. L'indice du matériau la lentille est de 1.52. Un objet de 1.80 [m] de hauteur est placé à 14 [m] de la lentille (peut importe que cela soit à gauche ou à droite de la lentille puisqu'elle est supposée biconvexe et donc symétrique). Alors la puissance dioptrique de la lentille est d'abord:

$$\mathcal{D} = (n - 1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right) = \frac{1}{f} = (1.52 - 1) \frac{2}{r} = 0.52 \frac{2}{0.3} \cong 3.47 \left[m^{-1} \right] \quad (39.102)$$

Ce qui est déjà bien une valeur positive et donne donc une focale (image ou objet peu importe de par la symétrie de la lentille mince sphérique!) d'environ 28.84 centimètres (donc il vaut mieux avoir un appareil photo avec un téléobjectif dans le cas présent). Nous remarquons, au vu de la valeur de la focale (foyer), que l'objet se trouve au-delà de la focale. La position de l'image sera donnée par:

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{14} + \frac{1}{q} \cong 3.47 \quad (39.103)$$

Soit:

$$q \cong 0.2942 \quad (39.104)$$

Donc l'image se trouve à environ 29.42 centimètres elle est aussi au-delà de la focale et sera par définition appelée "image réelle" inversée. Le grandissement sera lui de:

$$M = \frac{q}{p} = \frac{0.2942}{14} \cong 0.021 \quad (39.105)$$

La grandeur de l'image réelle inversée sera donc en q de:

$$h \cong 0.021 \cdot 1.8 = 0.0378 \quad (39.106)$$

Soit environ 3.78 centimètres.

Remarque: À nouveau, les problèmes d'aberrations sont aussi existants pour les lentilles.

À toutes fins utiles, indiquons la figure suivante:

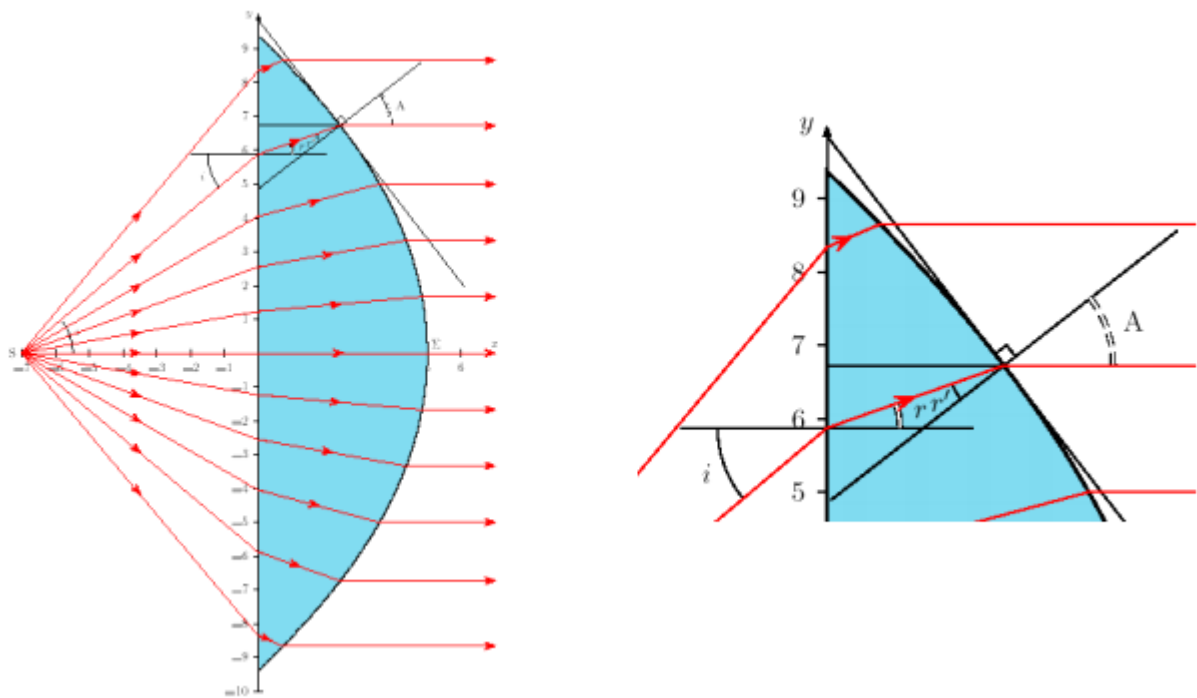


Figure: 39.20 - Lentille plan-courbe pleine

et donc une lentille plan-courbe pleine, comme un miroir parabolique, a la propriété de rendre parallèles les rayons partis de son foyer; elle produit par réfraction l'effet que le miroir parabolique produit par réflexion.

Fresnel inventa une lentille que l'on voit dans de nombreux phares et qui permet d'obtenir le même résultat avec moins de matière:

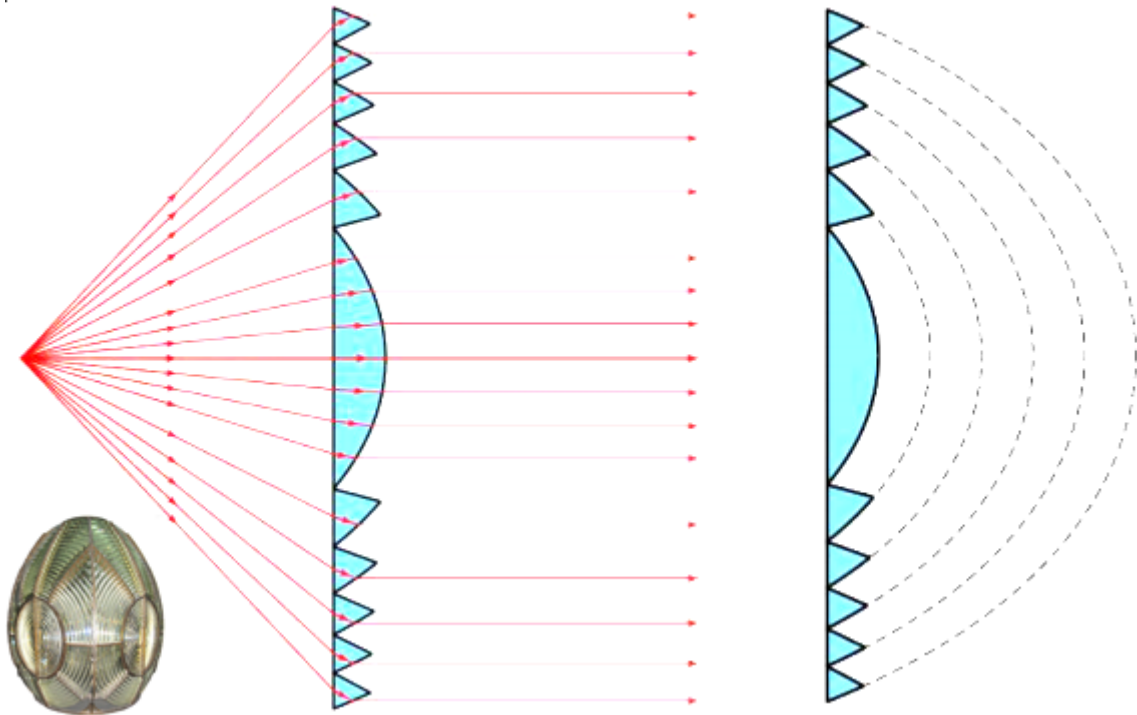


Figure: 39.21 - Lentille de Fresnel (réelle dans le coin en bas à gauche et schémas de principe à droite)

Faisons un peu de biologie pour clore....:

Le cristallin de l'oeil pouvant se déformer sous l'effet de certains muscles, constitue une lentille à focale variable permettant d'accommoder la vision des objets à distance variable. La distance du centre optique à la rétine étant fixe, le seul moyen de voir clairement des objets situés à des distances différentes est de modifier la distance focale. Dans son état ordinaire, le cristallin a une configuration assez plate, avec un grand rayon de courbure (il a alors une grande distance focale).

L'oeil a pour rôle de focaliser la lumière provenant d'un objet à l'infini (environ 25 centimètres pour un humain moyen...) sur la rétine. Mais tous les yeux ne font pas cela correctement et le "**punctum remotum**" (distance maximale de vision distincte sans accommodation) est parfois à une distance finie, même parfois inférieure à cinq mètres (entraînant probablement une fatigue des yeux).

Si l'objet s'approche, les muscles se contractent, le cristallin gonfle et sa distance focale diminue de façon que l'image se forme toujours sur sa rétine. Le point le plus proche qui peut être vu clairement avec le maximum d'accommodation est appelé le "**punctum proximum**". Cette distance évolue beaucoup avec l'âge: elle est de dix centimètres pour un enfant de dix ans, de cent centimètres pour une personne de soixante ans (c'est la presbytie).

6. PRISME

En optique, le prisme est un des composants les plus importants. On le retrouve en chimie, en physique de la matière condensée, en astrophysique, en optoélectronique et encore dans beaucoup d'autres appareils courants de la vie de tous les jours (comme les lentilles). Il s'agit probablement du premier outil façonné par l'homme pour faire de la "spectroscopie" (analyse du spectre) après l'arc-en-ciel (qui lui est un phénomène naturel de la spectroscopie).

Nous allons dans les paragraphes qui suivent déterminer les relations les plus importantes à connaître relativement aux prismes et utiles à l'ingénieur et au physicien.

Nous nous intéressons aux rayons lumineux entrant par une face et sortant par une autre ayant subi deux réfractions (nous n'étudierons pas les réflexions).

Voici la représentation type d'un prisme en optique géométrique avec le rayon incident S et sortant S' et les deux normales N, N' aux arêtes du sommet d'ouverture α . Plus les divers angles d'incidence et de réfraction:

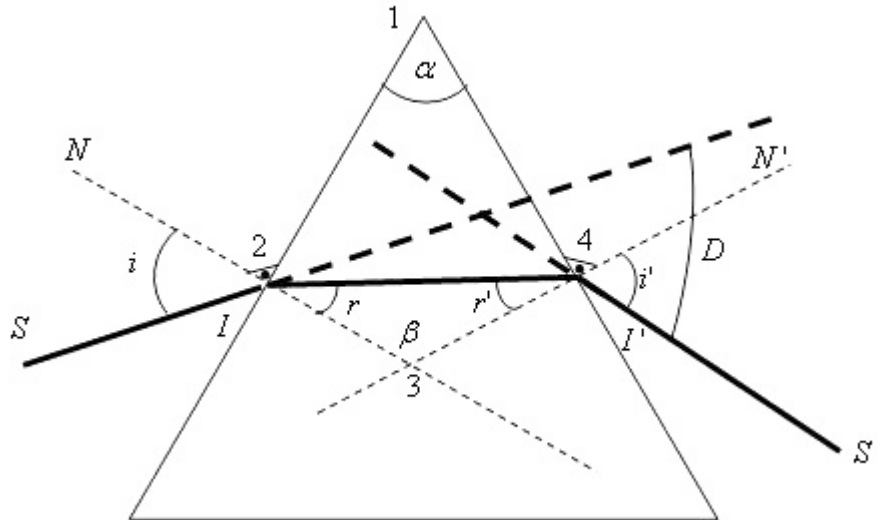


Figure: 39.22 - Représentation générique du prisme

Nous savons que la somme des angles d'un quadrilatère (toujours décomposable en deux triangles dont la somme des angles est π) vaut 2π . Donc dans le quadrilatère délimité par les sommets 1234. Nous avons la somme:

$$\alpha + 2 \cdot \frac{\pi}{2} + \beta = 2\pi \Rightarrow \beta = \pi - \alpha \quad (39.107)$$

Maintenant que la situation est posée passons à la partie optique...

Nous avons quatre relations fondamentales à démontrer pour le prisme.

D'abord, nous avons au point d'incidence I et I' la loi de Descartes qui nous permet d'écrire:

$$\sin(i) = \frac{m}{n} \sin(r) = n_{21} \sin(r) \quad (39.108)$$

Comme l'indice de réfraction absolu de l'air est de 1 alors nous avons simplement en I :

$$\sin(i) = m \sin(r) \quad (39.109)$$

Dans la même idée en I' nous avons:

$$\sin(r') = \frac{n}{m} \sin(i') = \frac{1}{m} \sin(i') \quad (39.110)$$

Donc:

$$\sin(i') = m \sin(r') \quad (39.111)$$

Nous avons aussi la relation:

$$\pi = r + r' + \beta = r + r' + \pi - \alpha \quad (39.112)$$

Soit:

$$\alpha = r + r' \quad (39.113)$$

L'angle de déviation D est facile à déterminer. Il suffit de prendre le quadrilatère central:

$$2\pi = i + i' + \beta + (\pi - D) \Rightarrow D = -\pi + i + i' + \beta = -\pi + i + i' + (\pi - \alpha) \quad (39.114)$$

Donc:

$$D = i + i' - \alpha \quad (39.115)$$

Nous avons donc les 4 relations fondamentales du prisme:

$$\begin{cases} \sin(i) = m \sin(r) \\ \sin(i') = m \sin(r') \\ \alpha = r + r' \\ D = i + i' - \alpha \end{cases} \quad (39.116)$$

Connaissant i et i' et l'indice de réfraction relatif m nous pouvons alors déterminer tous les paramètres.

L'idéal serait encore de pouvoir se débarrasser de la connaissance expérimentale de i' .

Nous avons donc:

$$i' = \arcsin(m \sin(r')) \quad (39.117)$$

Or:

$$r' = \pi - \beta - r = \pi - (\pi - \alpha) - r = \alpha - r = \alpha - \arcsin\left(\frac{\sin(i)}{m}\right) \quad (39.118)$$

Ainsi, il vient:

$$i' = \arcsin(m \sin(r')) = \arcsin\left\{m \sin\left[\alpha - \arcsin\left(\frac{\sin(i)}{m}\right)\right]\right\} \quad (39.119)$$

Donc:

$$D = i + \arcsin\left\{m \sin\left[\alpha - \arcsin\left(\frac{\sin(i)}{m}\right)\right]\right\} - \alpha \quad (39.120)$$

Puisqu'il est avéré que l'indice m d'un milieu varie avec la longueur d'onde suivant la loi de Cauchy, on comprend aisément que le prisme est capable de disperser la lumière blanche.

Enfin, si i est petit :

$$D \cong \arcsin \left\{ m \sin \left[\alpha - \arcsin \left(\frac{\sin(i)}{m} \right) \right] \right\} - \alpha \quad (39.121)$$

et si i et α sont petits, nous avons au premier ordre en développement de Maclaurin:

$$\begin{aligned} D &\cong \arcsin \left\{ m \sin \left[\alpha - \frac{i}{m} \right] \right\} - \alpha = \arcsin(m\alpha - i) - \alpha \cong \arcsin(m\alpha) - \alpha \\ &\cong m\alpha - \alpha = \alpha(m-1) \end{aligned} \quad (39.122)$$

Donc :

$$D \cong \alpha(m-1) \quad (39.123)$$

soit en introduisant explicitement la loi de Cauchy:

$$D \cong \alpha(m_\lambda - 1) = \alpha \left(\left(A + \frac{B}{\lambda^2} \right) - 1 \right) \quad (39.124)$$

7. ARC-EN-CIEL

Un arc-en-ciel est un phénomène optique et météorologique qui rend visible le spectre continu de la lumière quand le soleil brille pendant la pluie et que l'observateur contemple le ciel dans une direction opposée à celle du soleil. C'est un arc coloré avec le rouge à l'extérieur et le violet à l'intérieur.

L'arc-en-ciel est provoqué par la dispersion de la lumière du soleil par des gouttes de pluie approximativement sphériques. La lumière est d'abord réfractée en pénétrant la surface de la goutte, subit ensuite une réflexion partielle à l'arrière de cette goutte et est réfractée à nouveau en sortant. L'effet global est que la lumière entrante est principalement réfractée vers l'arrière sous un angle d'environ 40-42°, indépendamment de la taille de la goutte. La valeur précise de l'angle de réfraction dépend de la longueur d'onde (la couleur) des composantes de la lumière. Dans le cas de l'entrée dans un milieu plus réfringent, l'angle de réfraction de la lumière bleue est inférieur à celui de la lumière rouge (phénomène mis en évidence dans les prismes). Ainsi, après réflexion à l'interface eau-air, la lumière bleue sort d'une goutte au-dessus de la lumière rouge (voir figure ci-contre). L'observateur étant fixe, il voit la lumière issue de différentes gouttes d'eau avec des angles différents par rapport à la lumière du soleil. Le rouge apparaît donc plus haut dans le ciel que le bleu.

Parfois, un second arc-en-ciel moins lumineux peut être aperçu au-dessus de l'arc primaire. Il est provoqué par une double réflexion de la lumière du soleil à l'intérieur des gouttes de pluie et apparaît sous un angle de 50-53° dans la direction opposée au Soleil. En raison de la réflexion supplémentaire, les couleurs de ce second arc sont inversées par rapport à l'arc primaire, avec le bleu à l'extérieur et le rouge à l'intérieur, et l'arc est moins lumineux. C'est la raison pour laquelle il est plus difficile à observer. Un troisième arc-en-ciel peut être présent au voisinage du second, et inversé par rapport à celui-ci (donc identique au premier). Il est cependant nettement moins lumineux et observable uniquement dans des conditions exceptionnelles. En pratique, il n'est pas très facile à distinguer des arcs surnuméraires associés à l'arc secondaire. Il correspond aux rayons lumineux ayant subi cinq réflexions dans les gouttes d'eau. Deux arcs inversés l'un par rapport à l'autre peuvent également être observés dans la direction opposée, à environ 45° du Soleil (donc dans la direction de celui-ci), mais ceci est particulièrement difficile du fait de la proximité du Soleil. Les rares observations de ces deux arcs font

mention de morceaux d'arcs visibles par intermittence. Ces deux arcs correspondent aux rayons lumineux ayant subi trois et quatre réflexions dans les gouttes d'eau. Comme ils sont situés face au Soleil, ce ne sont pas les mêmes gouttes d'eau qui y contribuent. En pratique, les configurations favorables à leur observation sont nettement moins nombreuses que celles qui sont favorables à l'observation de l'arc secondaire, en particulier en raison de leur proximité du Soleil.

Pour étudier le phénomène considérons d'abord la goutte sphérique ci-dessous avec un rayon de lumière incident (une réfraction, deux réflexions) dont nous avons représenté la composante rouge et violet (les angles y sont approximatifs) ainsi que les indices de réfraction de l'eau et de l'air:

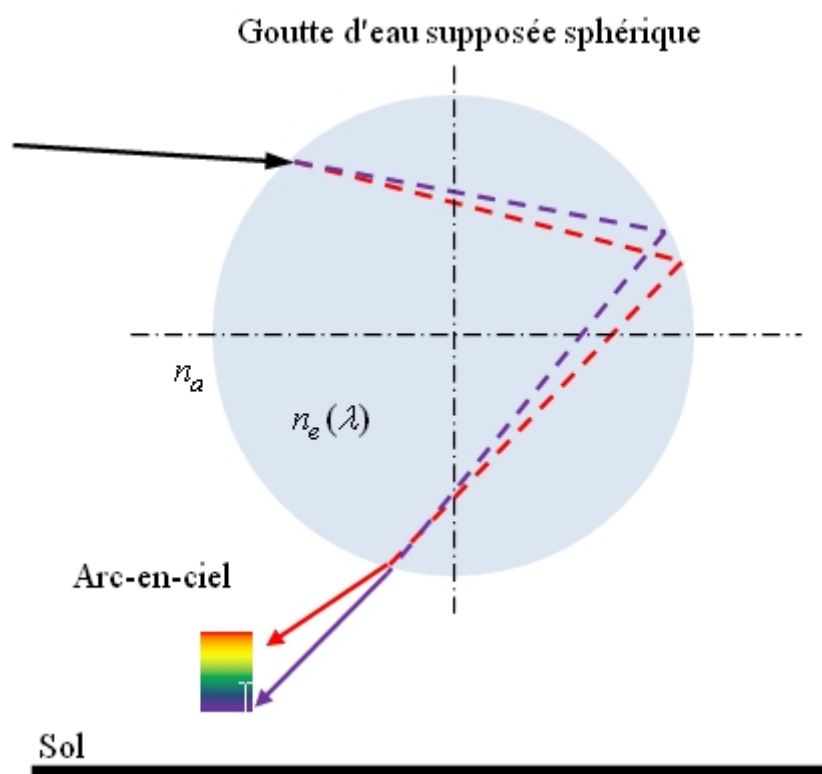


Figure: 39.23 - Goutte d'eau sphérique génératrice de l'arc-en-ciel

Nous cherchons à déterminer l'angle entre le rayon lumineux entrant (faisceau considéré comme contenant toutes les composantes de la lumière visible) supposé monochromatique et le rayon lumineux sortant (à l'opposé du Soleil: antisolaire). Ainsi, la différence d'angle pour deux couleurs, nous donnera l'angle par lequel nous devons changer notre regard pour observer deux couleurs différentes dans l'arc-en-ciel.

Remarque: Il n'y a pas de sens selon moi de calculer l'angle que doit faire le regard avec le sol (supposé plan) pour observer un arc-en-ciel comme le font certains ouvrages. Effectivement, de toute manière si nous dirigeons notre regard vers un arc-en-ciel nous le verrons de toute façon sur une grande étendue d'angle par rapport au sol. La seule chose qui a vraiment du sens, c'est donc la différence d'angle entre deux couleurs monochromatiques.

Pour cette étude, nous allons considérer la figure approximative suivante:

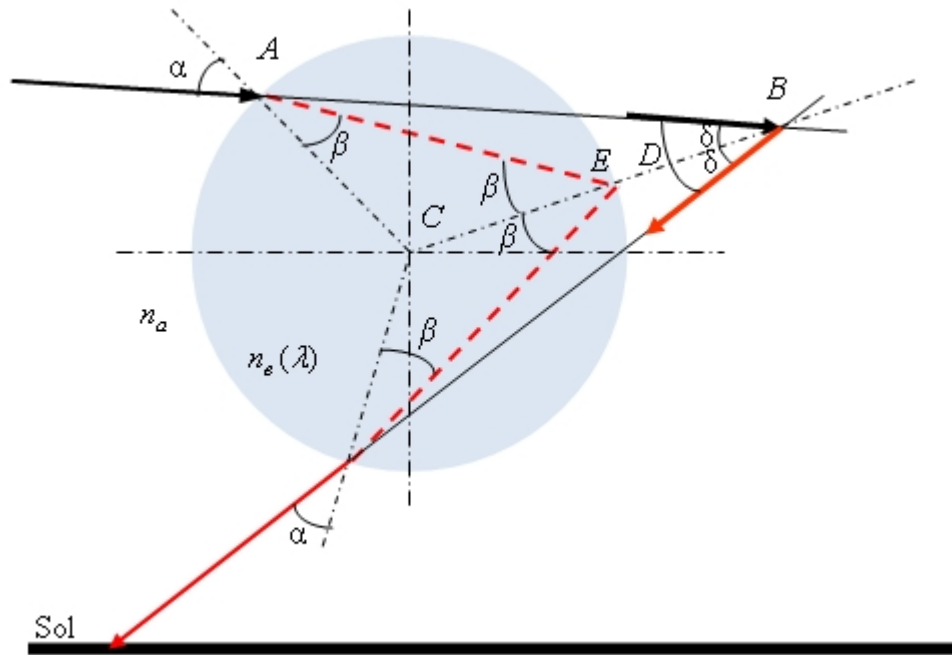


Figure: 39.24 - Chemin parcouru par la lumière dans la goutte d'eau

Avec la loi de Snell-Descartes nous avons dans un premier temps:

$$\frac{\sin(\alpha)}{\sin(\beta)} = \frac{n_e}{n_a} = n(\lambda) \quad (39.125)$$

Ce qui nous intéresse ici est donc l'angle de réflexion apparent 2δ , que nous noterons D (attention! certains auteurs choisissent la convention $D = \pi - 2\delta$). Pour le déterminer, nous partons de la relation suivante du triangle ABE :

$$(\alpha - \beta) + \delta + (\pi - \beta) = \pi \quad (39.126)$$

Il vient alors:

$$\delta = 2\beta - \alpha \quad (39.127)$$

D'où:

$$2\delta = 4\beta - 2\alpha = D \quad (39.128)$$

Ce que nous noterons finalement:

$$D = 4\beta - 2\alpha \quad (39.129)$$

et comme:

$$\sin(\beta) = \sin(\alpha) \frac{n_a}{n_e} \rightarrow \beta = \arcsin\left(\frac{n_a}{n_e} \sin(\alpha)\right) \quad (39.130)$$

Il vient:

$$D(n_a, n_e, \alpha) = 4 \arcsin\left(\frac{n_a}{n_e} \sin(\alpha)\right) - 2\alpha \quad (39.131)$$

ce qui est parfois noté un peu abusivement:

$$D(n_a, n_e, \alpha) = 4 \arcsin(n \sin(\alpha)) - 2\alpha \quad (39.132)$$

Si nous faisons une application pratique, nous avons pour le rouge 750 [nm] avec par exemple un angle d'incidence de 30° dans la goutte d'eau:

$$D(n_a, n_e, \alpha) = D\left(1.00, 1.32, \frac{\pi}{6}\right) = 4 \arcsin\left(\frac{1.00}{1.32} \sin\left(\frac{\pi}{6}\right)\right) - 2\left(\frac{\pi}{6}\right) \cong 0.506 = 29.03^\circ \quad (39.133)$$

et pour le violet 400 [nm] avec le même angle d'incidence de 30° dans la goutte d'eau:

$$D(n_a, n_e, \alpha) = D\left(1.00, 1.34, \frac{\pi}{6}\right) = 4 \arcsin\left(\frac{1.00}{1.34} \sin\left(\frac{\pi}{6}\right)\right) - 2\left(\frac{\pi}{6}\right) \cong 0.482 = 27.63^\circ \quad (39.134)$$

Soit une différence d'angle d'environ 2.4°.

Enfin, nous pourrions nous intéresser à l'angle α pour lequel l'angle D est maximum (ce qui correspond à l'arc-en-ciel le plus visible en termes de taille dans la réalité). Nous partons alors de:

$$f'(\alpha) = -2\alpha + 4 \arcsin(n \sin(\alpha)) \quad (39.135)$$

et nous cherchons les solutions de:

$$f'(\alpha) = 0 \quad (39.136)$$

avec $\alpha \in [0, \pi/2]$ et $n \in]0, 1[$.

Rappelons que nous avons démontré dans le chapitre de Calcul Différentiel Et Intégral que:

$$\arcsin'(x) = \frac{1}{\sqrt{1-x^2}} \quad (39.137)$$

Nous avons alors:

$$f'(\alpha) = -2 + \frac{4n \cos(\alpha)}{\sqrt{1-n^2 \sin^2(\alpha)}} = 0 \quad (39.138)$$

d'où:

$$\begin{aligned} \frac{2n \cos(\alpha)}{\sqrt{1-n^2 \sin^2(\alpha)}} = 1 &\Leftrightarrow \frac{4n^2 \cos^2(\alpha)}{1-n^2 \sin^2(\alpha)} = 1 \Leftrightarrow \frac{4n^2 \cos^2(\alpha)}{1-n^2(1-\cos^2(\alpha))} = 1 \\ &\Leftrightarrow 4n^2 \cos^2(\alpha) = 1-n^2+n^2 \cos^2(\alpha) \Leftrightarrow \cos^2 \alpha = \frac{1-n^2}{3n^2} \end{aligned} \quad (39.139)$$

De là nous tirons:

$$|\cos \alpha| = \frac{\sqrt{1-n^2}}{\sqrt{3}n} \quad (39.140)$$

c'est-à-dire:

$$\cos \alpha = \frac{\sqrt{1-n^2}}{\sqrt{3}n} \quad (39.141)$$

car nous cherchons les solutions avec $\alpha \in [0, \pi/2]$ et que le cosinus est positif sur cet intervalle.

Ainsi:

$$\alpha_{D_{\max}} = \arccos\left(\frac{\sqrt{1-n^2}}{\sqrt{3}n}\right) \quad (39.142)$$

Il vient alors pour n valant ~ 0.746 :

$$\alpha_{D_{\max}} = \arccos\left(\frac{\sqrt{1-n^2}}{\sqrt{3}n}\right) \cong 1.029 \cong 60^\circ \quad (39.143)$$

ce qui correspond relativement bien à la réalité (l'angle avec lequel nous levons la tête pour voir l'arc-en-ciel le plus visible/large).

La déviation correspondante est alors de:

$$D = 4 \arcsin\left(\frac{n_a}{n_e} \sin(\alpha)\right) - 2\alpha = 4 \arcsin\left(0.746 \sin\left(\frac{\pi}{3}\right)\right) - 2\frac{\pi}{3} \cong 0.715 \cong 40.97^\circ \quad (39.144)$$

et s'appelle "**angle de l'arc-en-ciel**".

Ainsi, les rayons lumineux perçus par l'observateur et dans lesquels le rouge (bord externe de l'arc-en-ciel) domine correspondent à l'ensemble des rayons issus du mur de pluie et faisant un angle d'environ 40° avec la direction des rayons solaires (voir figure ci-dessous). Les rayons lumineux constituant chaque couleur de l'arc-en-ciel forment alors des cônes de sommet les yeux de l'observateur et d'axe le rayon solaire passant par les yeux de l'observateur:

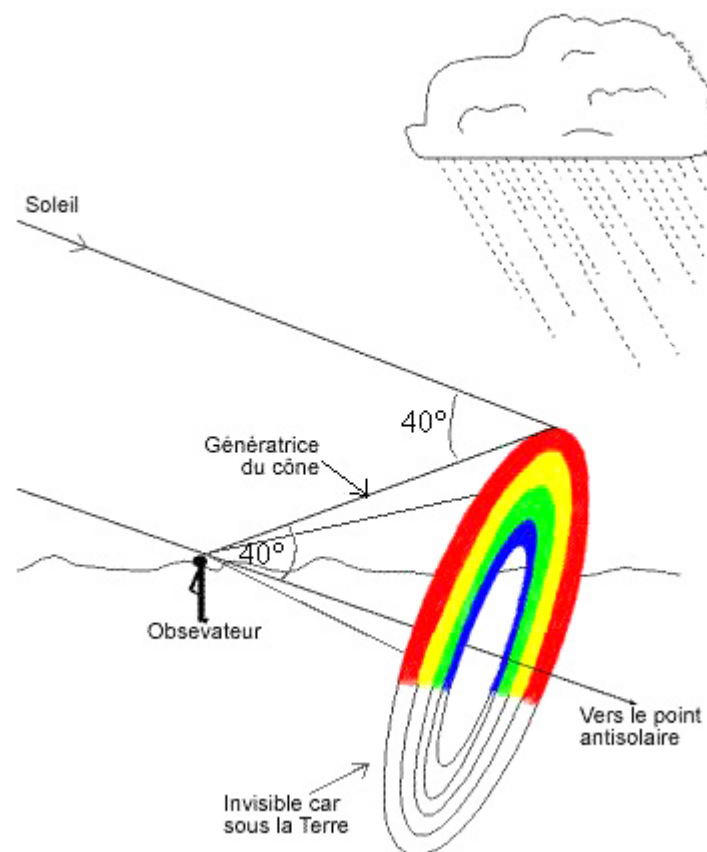
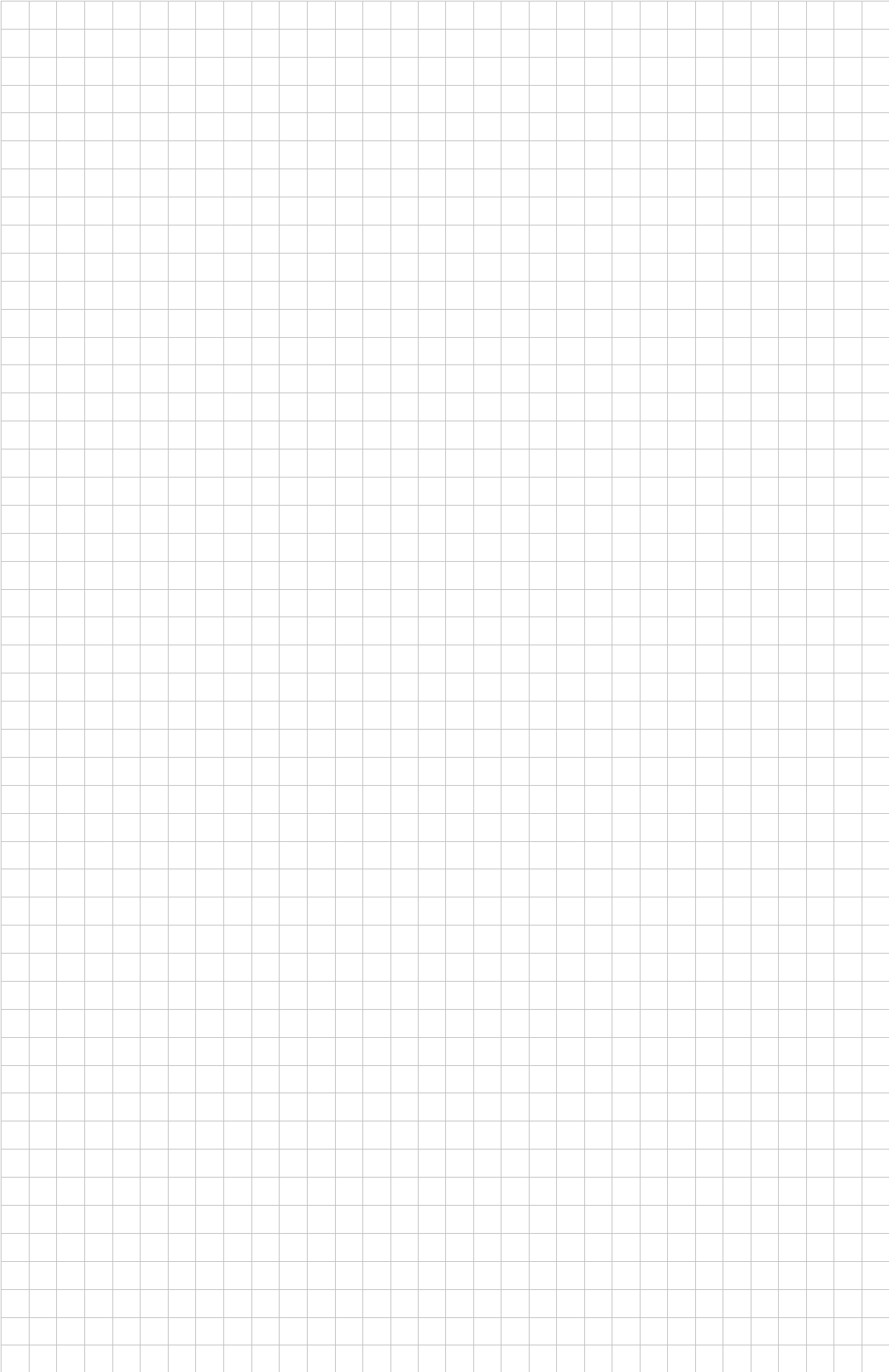
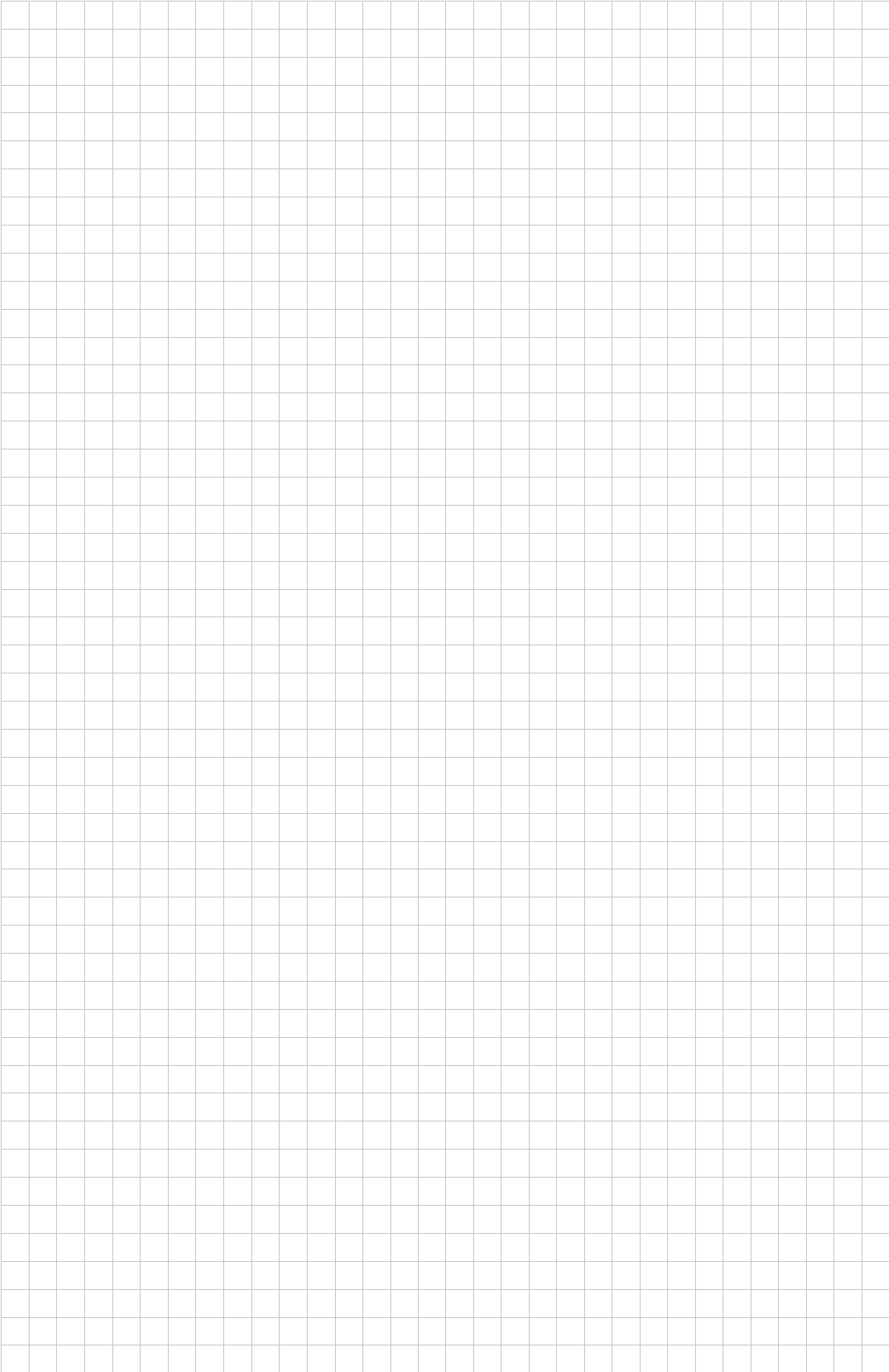
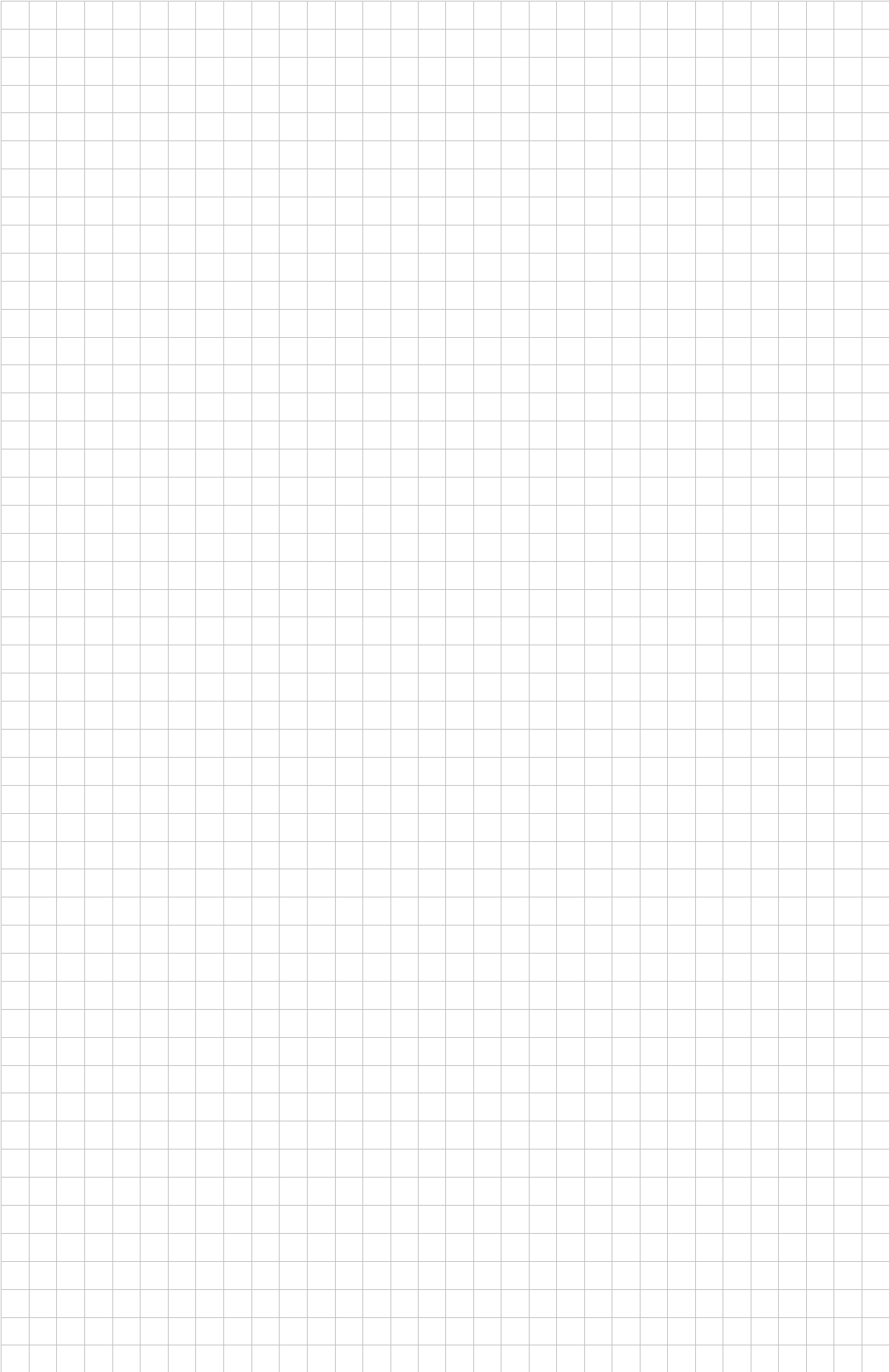


Figure: 39.25 - Figure représentant la génération de l'arc-en-ciel (source: ENS Culture-Sciences)







40. OPTIQUE ONDULATOIRE

Dans ce chapitre seront dégagés certains éléments qui ont conduit au développement de la mécanique quantique. Effectivement, la mécanique quantique est née, en premier lieu, d'une étude attentive de la nature de la lumière. Bien que cette science nouvelle se soit développée au début du 20ème siècle, les considérations qui l'ont guidée alors sont incontestablement le résultat de 25 siècles de maturation. Au fond, c'est à une longue histoire de la lumière pleine de controverses à laquelle la mécanique quantique apporte enfin au 20ème siècle une magistrale conclusion.

1. PRINCIPE D'HUYGENS

Huygens visualisait la propagation de la lumière comme résultant d'un processus de génération d'ondelettes sphériques en chaque point atteint par un front d'onde, ondelettes dont la somme donnait le champ en propagation. En traçant la tangente aux fronts d'onde des ondelettes à un instant donné, on obtenait le front d'onde de l'onde totale à ce même instant.

Nous rappelons qu'une surface d'onde ou "**front d'onde**" (cf. chapitre de Mécanique Ondulatoire) est le lieu des points du milieu atteints par le mouvement ondulatoire au même instant. La perturbation a donc même phase en tout point d'une surface d'onde. Pour une onde plane, par exemple, la perturbation s'exprime par (nous l'avons démontré dans le chapitre de Mécanique Ondulatoire):

$$f(\vec{k} \cdot \vec{x} - v \cdot t) \quad (40.1)$$

ou dans une formulation plus générale:

$$f(\vec{k} \cdot \vec{r} - vt) \quad (40.2)$$

qui donne donc l'expression de la propagation de la perturbation pour laquelle la "**surface d'onde**" est le lieu des points où la phase $\vec{k} \cdot \vec{x} - vt$ a même valeur à un instant donné. La surface d'onde est donnée en conséquence par l'équation:

$$\vec{k} \cdot \vec{r} - vt = c^*t \quad (40.3)$$

Huygens, a donné une méthode imagée de représentation du passage d'une surface d'onde à une autre dans le cas où l'onde est supposée résulter du mouvement des particules constituant le milieu matériel. Ainsi, si nous considérons la surface d'onde S ci-dessous:

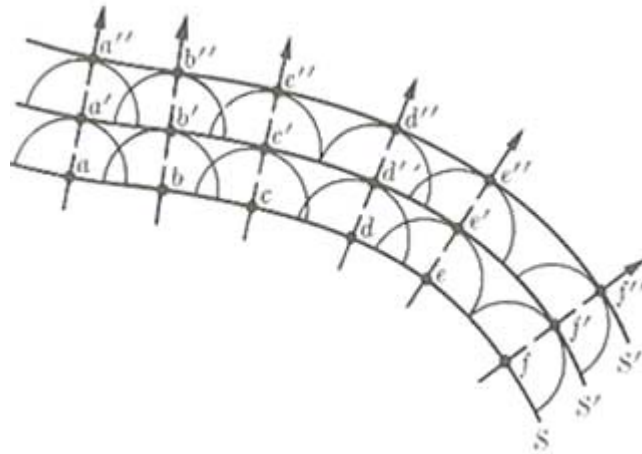


Figure: 40.1 - Représentation d'une surface d'onde selon Huygens

Quand le mouvement ondulatoire atteint cette surface, chaque particule $a, b, c, \dots$ de la surface devient à son tour une source d'ondes, émettant des ondes secondaires (indiquées par les petits demi-cercles) qui atteignent la couche suivante de particules du milieu. Ces particules sont mises en mouvement et forment la nouvelle surface d'onde S' et ainsi de suite... Ainsi, Huygens avait une conception ondulatoire de la lumière, mais il ne considérait pas la nature périodique de l'onde, ce qui ne lui permettait pas d'introduire la notion de couleur de la lumière; de plus, selon son principe, une onde se propageant en sens inverse à celui de l'onde incidente devrait aussi se manifester, ce qui n'est pas le cas dans un matériau homogène.

L'intuition d'Huygens est cependant proche de la réalité, comme le montrera Fresnel dans sa théorie de la diffraction. Il faudra cependant attendre Kirchhoff, qui introduira un facteur d'inclinaison (obliquité) dans la théorie, pour obtenir une explication de l'absence d'onde se propageant vers l'arrière (le temps venu nous rédigerons les développements y relatifs).

1.1. LOI DE MALUS

Comme tous les "**points correspondants**" $\{a, b, c, \dots\}, \{a', b', c', \dots\}, \dots$ sont équidistants, par le principe d'Huygens, la "**loi de Malus**" (la première donc et pas celle obtenue lors de l'étude de la polarisation de la lumière comme nous le verrons plus loin) affirme que l'intervalle de temps entre les points correspondants de deux surfaces d'onde est le même pour tout couple de points correspondants.

Conséquences (se référer en même temps à la figure ci-dessous):

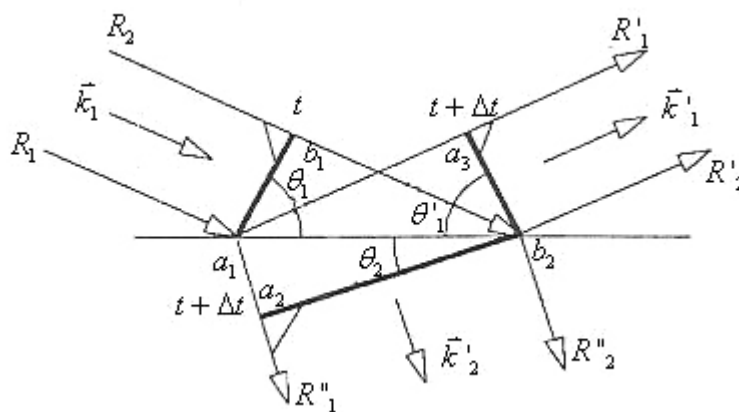


Figure: 40.2 - Configuration d'expérience de pensée pour l'étude de la loi de Malus

Si nous notons les vitesses de propagation des rayons incidents R_1, R_2 par v_1 et respectivement v_2 , nous avons

$$\overline{b_1 b_2} = v_1 \Delta t \quad \overline{a_1 a_3} = v_1 \Delta t \quad \overline{a_1 a_2} = v_2 \Delta t \quad (40.4)$$

- Lorsque l'onde se propage dans un milieu homogène, les rayons lumineux doivent être rectilignes et les surfaces d'onde rester parallèles.
- Lorsque l'onde change de milieu, les distances entre deux paires de points correspondants varient d'un milieu à l'autre, si les vitesses de propagation sont différentes.

Cette loi permet de retrouver la loi de Descartes-Snellius que nous avons déjà démontrée dans le chapitre d'Optique Géométrique, ce qui assure a priori que le principe d'Huygens reste valide dans le cadre de l'optique géométrique.

Démonstration:

Selon la figure ci-dessus, nous avons:

$$\Delta t = \frac{\overline{b_1 b_2}}{v_1} = \frac{\overline{a_1 a_3}}{v_1} = \frac{\overline{a_1 a_2}}{v_2} \quad (40.5)$$

en divisant chaque terme par $\overline{a_1 b_2}$, nous obtenons:

$$\frac{\sin \theta_1}{v_1} = \frac{\sin \theta'_1}{v_1} = \frac{\sin \theta_2}{v_2} \quad (40.6)$$

et nous retrouvons donc bien la loi de Descartes-Snellius telle que nous l'avions obtenue dans le chapitre d'Optique Géométrique:

$$\frac{\sin \theta_1}{\sin \theta_2} = \frac{v_1}{v_2} = n_{12} \quad (40.7)$$

en notant au passage que sur le schéma, nous avons aussi $\theta_1 = \theta'_1$.

2. DIFFRACTION DE FRAUNHOFER

Du point de vue de l'optique géométrique, un faisceau lumineux est un cylindre de section Σ qui rassemble un grand nombre de rayons parallèles. Il est donc supposé rectiligne lorsqu'il est défini dans un milieu homogène.

L'émittance énergétique $M [Wm^{-2}]$ du faisceau ne varie que si une lentille (ou un autre dispositif) fait varier sa section Σ ou si le milieu absorbe de l'énergie.

Le faisceau lumineux "éclate" quand un obstacle ne laisse passer qu'une partie Σ de sa section.

Le principe d'Huygens montre que ce sont les bords de l'obstacle qui engendrent cette diffraction.

Le phénomène est général mais n'est bien observable que si le rapport L/Σ est très grand. L étant la

longueur des bords. Cette condition est nécessaire pour que l'intensité de la partie non diffractée du faisceau ne masque pas l'effet.

Définitions:

D1. Nous parlons de "**diffraction de Fraunhofer**" lorsque, comme supposé précédemment, les rayons lumineux incidents sont parallèles et le phénomène observé à relativement grande distance de l'écran.

D2. Nous parlons de "**diffraction de Fresnel**" lorsque les rayons incidents forment un faisceau divergent, en provenance d'une source ponctuelle ou si nous observons le phénomène à faible distance.

Considérons un cas générique et le plus répandu dans les laboratoires de physique qui est la diffraction par une fente rectangulaire étroite:

Pour cela, nous considérons que le faisceau incident, perpendiculaire à la fente, présente un front d'onde électromagnétique plane et périodique, et donné par:

$$f(x, t) = \hat{f} \sin(kx - \omega t) \quad (40.8)$$

où pour rappel, sa longueur d'onde est donnée par:

$$\lambda = 2\pi / k \quad (40.9)$$

2.1. CAS D'UNE FENTE RECTANGULAIRE

La largeur e de la fente est orientée selon l'axe y , sa hauteur h (paramètre que l'on ne peut pas représenter dans la figure puisqu'il s'agit d'une vue du dessus) est supposée très grande afin de pouvoir négliger l'effet des extrémités.

Suivant le principe d'Huygens, le front de l'onde plane, délimité par la fente, constitue une multitude de sources $df(x, t)$, de largeur dy , qui émettent, en phase, des ondelettes sphériques décrites par leur vecteur champ associé:

$$df(x, t) = \frac{\hat{f}}{e} dy \sin(kx - \omega t) \quad (40.10)$$

Considérons maintenant un point d'observation P , à une distance R de la source (assimilée à la fente). Nous avons vu lors de l'étude des sources d'émission de type sphérique (cf. chapitre d'Électrodynamique) que leur amplitude diminuait de manière inversement proportionnelle à la distance telle que:

$$df(x, t) = \frac{\hat{f}}{e \cdot R} dy \sin(kx - \omega t) \quad (40.11)$$

Or, les ondelettes, chacune suivant le point de la fente auquel elle est assimilée, ne vont pas toutes parcourir la même distance R mais une distance propre r . Cependant, si R est suffisamment éloigné de la fente, nous nous permettrons d'approximer:

$$\frac{\hat{f}}{e \cdot R} \approx \frac{\hat{f}}{e \cdot r} \quad (40.12)$$

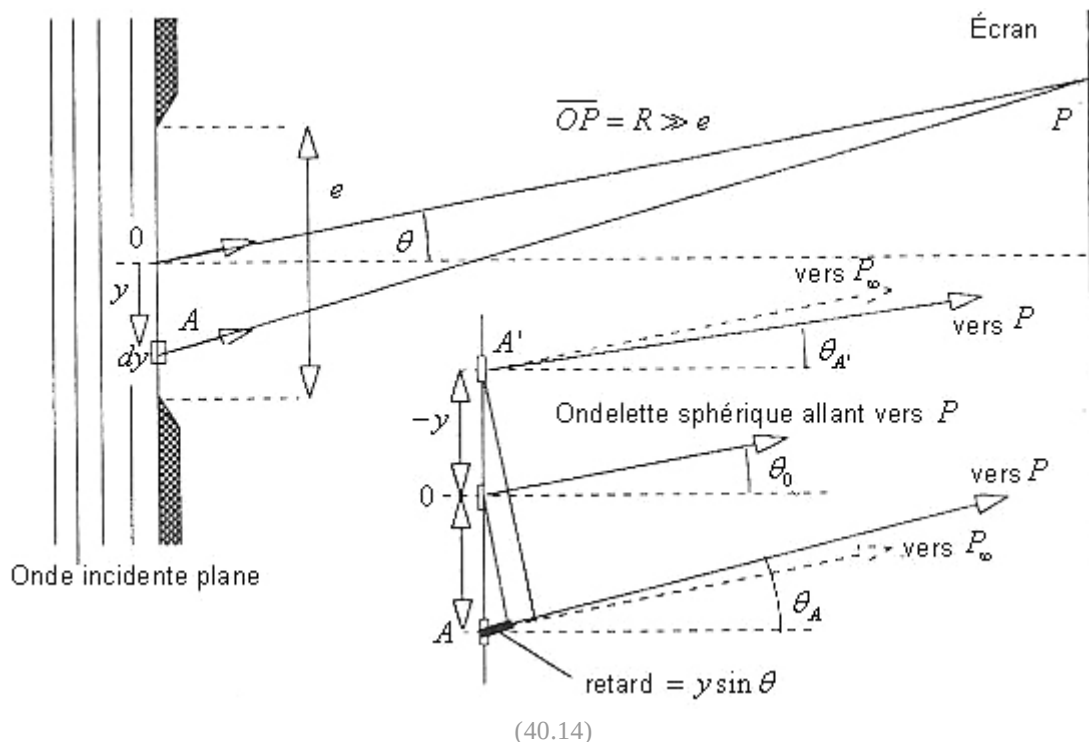
reste encore le terme périodique $\sin(kx - \omega t)$ où nous posons $x = r$. Or, nous avons pour valeurs extrêmes:

$$R - \frac{e}{2} \sin \theta \leq r \leq R + \frac{e}{2} \sin \theta \quad (40.13)$$

Ces valeurs extrêmes correspondant respectivement, à l'avance et au retard des fonctions d'onde décrivant la propagation des ondelettes aux extrémités de la fente.

Effectivement, il suffit de voir la figure ci-dessous, en considérant donc $R \gg e$ et ainsi:

$$\theta := \theta_{A'} \cong \theta_0 \cong \theta_A$$



Ainsi, en plaçant l'origine de la coordonnée y au milieu de la fente, nous avons:

$$df(x, t) = \frac{\hat{f}}{e \cdot R} dy \sin(kr - \omega t) = \frac{\hat{f}}{e \cdot R} dy \sin(k(R + y \sin \theta) - \omega t) \quad (40.15)$$

Donc les différentes ondelettes sont déphasées et produisent ainsi des interférences.

Définition: En mécanique ondulatoire, on parle "**d'interférences**" lorsque deux ondes de même type se chevauchent. Ce phénomène se rencontre souvent en optique avec les ondes lumineuses, mais il apparaît également avec les ondes sonores.

L'onde diffractée dans la direction de θ , est alors donnée par la somme de toutes les contributions:

$$f(x, t) = \frac{\hat{f}}{e \cdot R} \int_{-e/2}^{e/2} \sin(k(R + y \sin \theta) - \omega t) dy = \frac{\hat{f}}{e \cdot R} \int_{-e/2}^{e/2} \sin((kR - \omega t) + ky \sin \theta) dy \quad (40.16)$$

Sachant que (relations trigonométriques explicitées suite à la demande d'un lecteur):

$$\begin{aligned}
 \int_{-e/2}^{e/2} \sin(a + by) dy &= \int_{-e/2}^{e/2} (\cos(a) \sin(by) + \sin(a) \cos(by)) dy \\
 &= \cos(a) \int_{-e/2}^{e/2} \sin(by) dy + \sin(a) \int_{-e/2}^{e/2} \cos(by) dy = -\frac{\cos(a) \cos(by)}{b} + \frac{\sin(a) \sin(by)}{b} \Big|_{-e/2}^{e/2} \\
 &= \frac{2 \sin(a) \sin\left(b \frac{e}{2}\right)}{b}
 \end{aligned} \quad (40.17)$$

Nous avons donc:

$$\begin{aligned}
 f(x, t) &= \frac{\hat{f}}{e \cdot R} \int_{-e/2}^{e/2} \sin((kR - \omega t) + ky \sin \theta) dy = \frac{\hat{f}}{e \cdot R} \frac{2 \sin(kR - \omega t) \sin\left(k \sin \theta \frac{e}{2}\right)}{k \sin \theta} \\
 &= \frac{\hat{f}}{R} \frac{\sin\left(k \sin \theta \frac{e}{2}\right)}{\frac{ek \sin \theta}{2}} \sin(kR - \omega t) = \frac{\hat{f}}{R} \frac{\sin\left(\frac{\pi e}{\lambda} \sin \theta\right)}{\frac{\pi e}{\lambda} \sin \theta} \sin(kR - \omega t)
 \end{aligned} \quad (40.18)$$

Nous avons démontré dans le chapitre d'Électrodynamique que l'énergie (in extenso l'intensité) d'une onde électromagnétique était donnée (dans le vide) par la valeur scalaire moyenne du vecteur de Poynting:

$$\bar{S} = \frac{\hat{E} \hat{B}}{2\mu_0} \quad (40.19)$$

Nous avons donc en considérant que le champ magnétique et électrique sont proportionnels au terme:

$$\frac{\sin\left(\frac{\pi e}{\lambda} \sin \theta\right)}{\frac{\pi e}{\lambda} \sin \theta} \quad (40.20)$$

le résultat suivant:

$$\bar{S} \propto \bar{S}_0 \left(\frac{\sin\left(\frac{\pi e}{\lambda} \sin \theta\right)}{\frac{\pi e}{\lambda} \sin \theta} \right)^2 = \bar{S}_0 \left(\frac{\sin(\pi \Gamma \sin \theta)}{\pi \Gamma \sin \theta} \right)^2 \quad (40.21)$$

qui est l'émittance lumineuse émise dans la direction θ et où nous avons posé:

$$\Gamma = \frac{e}{\lambda} \quad (40.22)$$

Si nous introduisons le sinus cardinal que nous avons déjà rencontré lors de notre étude des transformées de Fourier dans le chapitre sur les Suites et Séries nous avons alors l'écriture de la relation

précédente qui se trouve nettement condensée:

$$\bar{S} \propto \bar{S}_0 \cdot \text{sinc}^2(\pi \Gamma \sin \theta) \quad (40.23)$$

Dont nous pouvons avoir une forme générique tracée avec Maple en utilisant:

```
>Gamma:=3;plot((sin(Gamma*x)/(Gamma*x))^2, x=-Pi..Pi);
```

Donc nous pouvons obtenir le même résultat en prenant le module au carré de la transformée de Fourier d'un signal monochromatique au travers d'une fenêtre rectangulaire. Ainsi, il semble possible d'étudier les phénomènes de diffraction en utilisant la transformée de Fourier et ce domaine se nomme "l'optique de Fourier".

Voici une représentation graphique du rapport $\bar{S}/\bar{S}_0$ pour différentes valeurs du rapport $\Gamma = e/\lambda$:

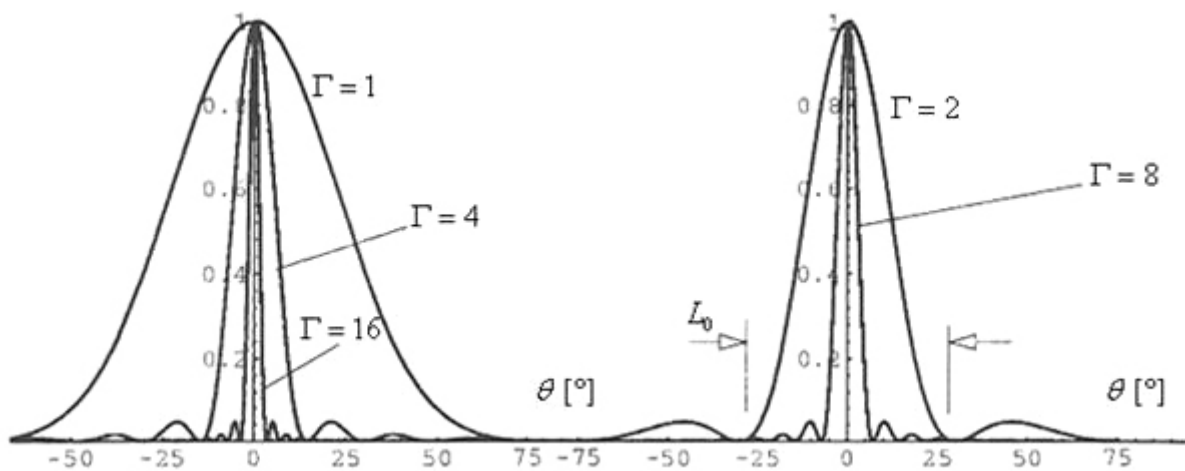
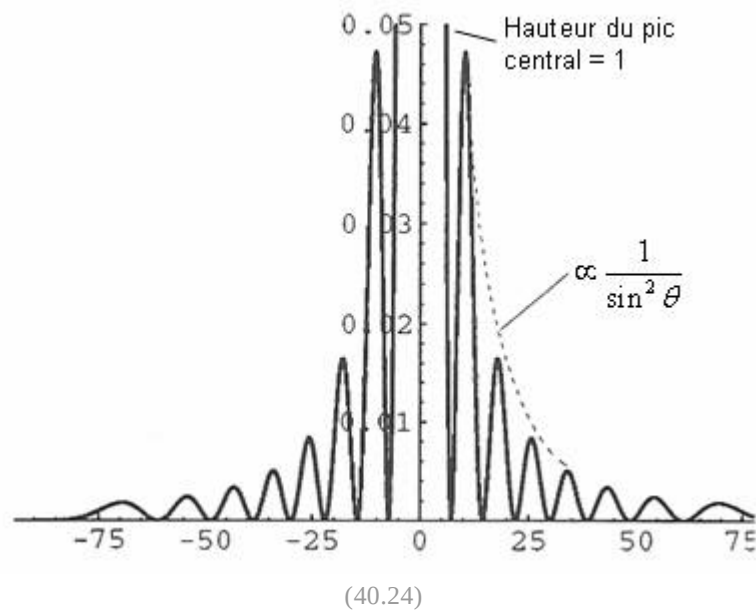


Figure: 40.3 - Représentation des franges de diffraction

De part et d'autre de la frange centrale, il y en a d'autres, plus étroites et disposées symétriquement. Leur intensité diminue très rapidement selon le terme prépondérant au dénominateur:

$$\frac{\bar{S}_{\max}}{\bar{S}_0} \propto \frac{1}{(\pi \Gamma \sin \theta)^2} = \frac{\lambda^2}{\pi^2 e^2} \frac{1}{\sin^2 \theta}$$



Dont voici une image réelle:

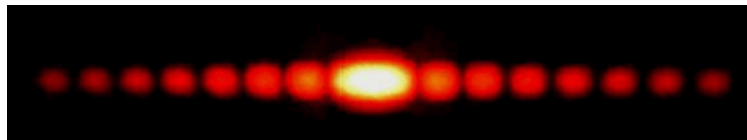


Figure: 40.4 - Photo d'une frange de diffraction réelle

Entre les franges, se trouvent des zones d'obscurité qui sont le siège d'interférences destructives. Leur position est donnée par la condition:

$$\sin\left(\frac{\pi e}{\lambda} \sin \theta\right) = 0 \Rightarrow \frac{\pi e}{\lambda} \sin \theta = n\pi \quad (40.25)$$

sauf pour $n = 0 \Rightarrow \theta = 0$ où l'on observe un maximum !

Nous observons donc des franges sombres dans les directions:

$$\theta_n = \arcsin\left(\frac{n\lambda}{e}\right) \quad \forall n \in \mathbb{N}^* \quad (40.26)$$

Ainsi, la largeur angulaire de la frange centrale est le double de la valeur angulaire obtenue pour le premier minimum:

$$L_0 = 2R \arcsin\left(\frac{\lambda}{e}\right) \quad (40.27)$$

Nous obtenons la largeur des pics suivants, comme suit:

Deux minima successifs satisfont donc les conditions:

$$\sin \theta_n = \frac{n\lambda}{e} \quad \sin \theta_{n+1} = \frac{(n+1)\lambda}{e} \quad (40.28)$$

Ainsi:

$$\sin \theta_n - \sin \theta_{n+1} = \frac{\lambda}{e} = 2 \sin \left(\frac{\theta_{n+1} - \theta_n}{2} \right) \cos \left(\frac{\theta_{n+1} + \theta_n}{2} \right) = 2 \sin \left(\frac{\Delta \theta}{2} \right) \cos \left(\frac{\theta_{n+1} + \theta_n}{2} \right) \quad (40.29)$$

En posant:

$$(\theta_{n+1} + \theta_n) / 2 \approx \theta \quad (40.30)$$

il vient dès lors:

$$\Delta \theta = 2 \arcsin \left(\frac{\lambda}{2e \cos \theta} \right) \quad (40.31)$$

Puisque l'émittance énergétique diminue très rapidement, seules les premières franges (pour lesquelles $\cos \theta \approx 1$) sont observables. Il reste:

$$L_n = 2R \arcsin \left(\frac{\lambda}{2e} \right) \cong \frac{L_0}{2} \quad (40.32)$$

Les positions des maxima sont quant à elles données par la condition:

$$\begin{aligned} \frac{dM}{d\theta} = 0 &= \frac{d}{d\theta} \left(\frac{\sin \left(\frac{\pi e}{\lambda} \sin \theta \right)}{\frac{\pi e}{\lambda} \sin \theta} \right) = \frac{\lambda}{\pi e} \frac{d}{d\theta} \left(\frac{\sin \left(\frac{\pi e}{\lambda} \sin \theta \right)}{\sin \theta} \right) \\ &= \frac{\lambda}{\pi e} \frac{1}{\sin^2 \theta} \left[\sin \theta \frac{d}{d\theta} \sin \left(\frac{\pi e}{\lambda} \sin \theta \right) - \sin \left(\frac{\pi e}{\lambda} \sin \theta \right) \frac{d}{d\theta} \sin \theta \right] \\ &= \frac{\lambda}{\pi e} \frac{1}{\sin^2 \theta} \left[\sin \theta \cos \left(\frac{\pi e}{\lambda} \sin \theta \right) \frac{\pi e}{\lambda} \cos \theta - \sin \left(\frac{\pi e}{\lambda} \sin \theta \right) \cos \theta \right] \\ &= \cos \left(\frac{\pi e}{\lambda} \sin \theta \right) \cot \theta - \frac{\lambda}{\pi e} \sin \left(\frac{\pi e}{\lambda} \sin \theta \right) \frac{\cot \theta}{\sin \theta} \\ \Rightarrow \cos \left(\frac{\pi e}{\lambda} \sin \theta \right) \cot \theta &= \frac{\lambda}{\pi e} \sin \left(\frac{\pi e}{\lambda} \sin \theta \right) \frac{\cot \theta}{\sin \theta} \Rightarrow \tan \left(\frac{\pi e}{\lambda} \sin \theta \right) = \frac{\pi e}{\lambda} \sin \theta \end{aligned} \quad (40.33)$$

Posons:

$$\phi = \frac{\pi e}{\lambda} \sin \theta \quad (40.34)$$

La résolution numérique de:

$$\tan \phi = \phi \quad (40.35)$$

donne (en radians):

$$\phi = \pm \{0, 4.49, 7.72, 10.9, 14.1, \dots\} \quad (40.36)$$

Les positions des maxima successifs sont alors:

$$\begin{aligned}\phi = 0 &\Rightarrow \theta = 0 \\ \phi = 4.49 &\Rightarrow \theta = \arcsin\left(\frac{4.49}{\pi} \frac{\lambda}{e}\right) \cong \arcsin\left(1.5 \frac{\lambda}{e}\right) \quad (40.37)\end{aligned}$$

etc...

Nous aurions facilement pu obtenir une approximation convenable de ce résultat, en considérant que l'intensité est maximale lorsque:

$$\left. \frac{\sin\left(\frac{\pi e}{\lambda} \sin \theta\right)}{\frac{\pi e}{\lambda} \sin \theta} \right|_{\max} \cong \sin\left(\frac{\pi e}{\lambda} \sin \theta\right) \Big|_{\max} \Rightarrow \sin\left(\frac{\pi e}{\lambda} \sin \theta\right) = 1 \Rightarrow \frac{\pi e}{\lambda} \sin \theta = (2n+1) \frac{\pi}{2} \quad (40.38)$$

Ce qui nous amène à écrire:

$$\theta_{\max, n} = \arcsin\left(\frac{(2n+1)\lambda}{2e}\right) \text{ avec } n \in \mathbb{N}^* \quad (40.39)$$

Remarque: Un résultat remarquable de l'expérience de Fraunhofer est qu'elle remet en question la vision corpusculaire de la lumière telle que nous l'avions au 19ème siècle.

Effectivement, beaucoup d'expériences telle que la projection de l'ombre d'un objet sur un mur semblait bien montrer que la lumière était tel un corpuscule ne traversant pas la matière et étant stoppée net par tout obstacle que ce soit en son centre ou en ses bords (il faut attirer votre attention sur les "bords" en particulier).

Or, l'expérience de Fraunhofer ainsi qu'en particulier celle de Fresnel en ce qui concerne les bords (nous la verrons plus loin car elle est mathématiquement plus délicate à aborder), montrent bien que la lumière semble pouvoir se comporter non pas comme un simple corpuscule mais bien comme une onde (à partir du principe de d'Huygens que nous avons utilisé pour nos développements) tel que nous l'ont montré les développements précédents qui expliquent parfaitement bien les résultats expérimentaux des diffractions de Fraunhofer.

Mais alors pourquoi garder le modèle corpusculaire de la lumière? Tout simplement pour d'autres résultats expérimentaux et théoriques parmi lesquels les plus connus sont l'effet photo-électrique ou la diffraction Compton (cf. chapitre de Physique Nucléaire) qui s'expliquent théoriquement à merveille si ce n'est parfaitement avec un modèle corpusculaire de la lumière (et certaines autres particules de dimension, charge, spin, etc. donné).

Au fait, comme nous le verrons dans le chapitre de Physique Quantique Ondulatoire, c'est le physicien De Broglie qui va mettre définitivement un terme à cette dualité paradoxale en reliant à l'aide des outils de la mécanique relativiste et physique quantique ondulatoire les deux aspects mathématiquement.

2.1.1. POUVOIR DE RÉOLUTION

Selon le critère du physicien anglais Lord Rayleigh: le "**pouvoir de résolution**" (ou "**pouvoir séparateur angulaire**") d'une fente, est l'angle $\alpha_{\min}$ entre deux rayons lumineux de longueur d'onde λ , issus de deux sources ponctuelles S_1, S_2 , éloignées, dont les figures de diffractions sont séparées telles que le premier zéro de la figure de diffraction se trouve à la place du maximum de l'autre:

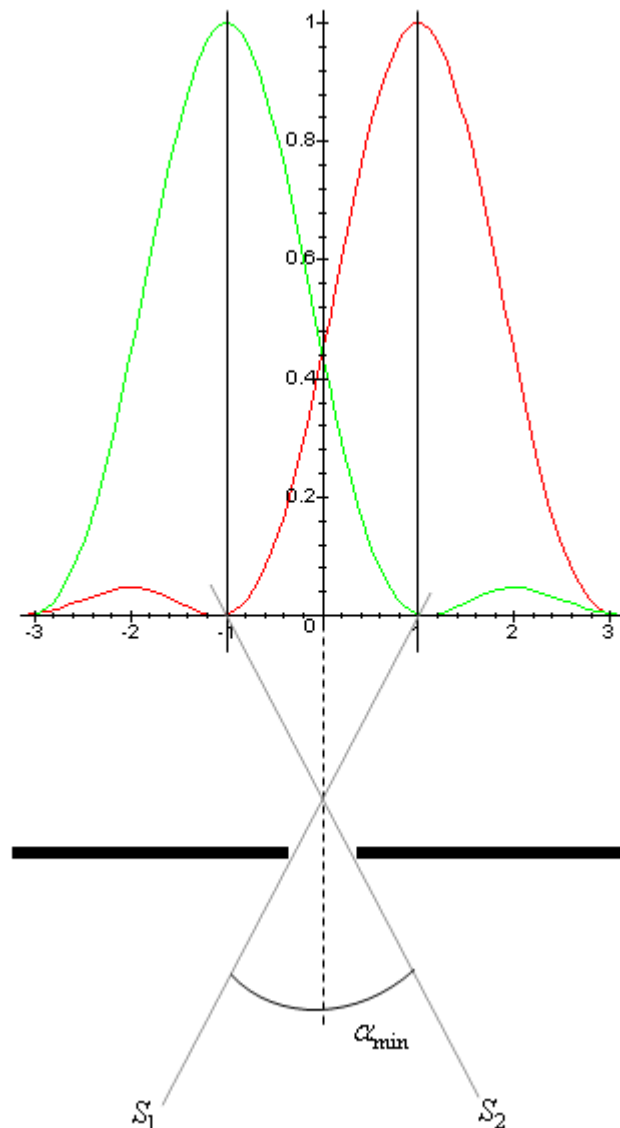


Figure: 40.5 - Illustration du principe de pouvoir de résolution

Ce concept est énormément utilisé en photographie, astronomie, radioastronomie, etc. Il convient donc d'y porter une attention toute particulière!

Or, nous avons vu que les minimas étaient donnés par:

$$\sin \theta_n = \frac{n\lambda}{e} \quad (40.40)$$

Et si nous prenons le cas où $n = 1$, nous retrouvons la relation disponible dans de nombreux ouvrages sans démonstration:

$$\theta \cong \sin \theta_x = \frac{\lambda}{e} = \alpha_{\text{min}} \quad (40.41)$$

Donc le pouvoir séparateur angulaire est proportionnel au rapport de la longueur d'onde à l'épaisseur de la fente. Évidemment dans la pratique le but est d'avoir une valeur d'angle α_{min} la plus grande possible (sinon les objets se confondent et l'image est floue).

Pour augmenter le pouvoir de résolution, il faut donc soit travailler à une longueur d'onde plus courte soit augmenter l'épaisseur de la fente de l'instrument et comme la longueur d'onde est souvent imposée par le sujet d'expérience, il est naturel de vouloir faire varier e .

Si la lumière qui passe à travers une fente forme une image sur un écran, et que l'image est observée au microscope par exemple, il est impossible, quel que soit le grandissement du microscope, d'observer plus de détails dans l'image qu'il n'est permis par le pouvoir de résolution de la fente. Il faut tenir compte de ces considérations dans la conception des instruments d'optique.

2.2. CAS D'UN RÉSEAU DE FENTES RECTANGULAIRES

Considérons maintenant un réseau de N fentes étroites de largeur $e \cong \lambda$, de hauteur $H \gg e$ et distantes de d . Un unique faisceau incident éclaire toutes les fentes.

Remarque: L'étude de ce modèle va nous permettre de comprendre en partie comment fonctionne le prisme et le fonctionnement des goniomètres utilisés en astronomie pour l'analyse du spectre ainsi que la diffraction par rayons X par un réseau d'atomes (donc l'importance est non négligeable).

Soit le schéma suivant:

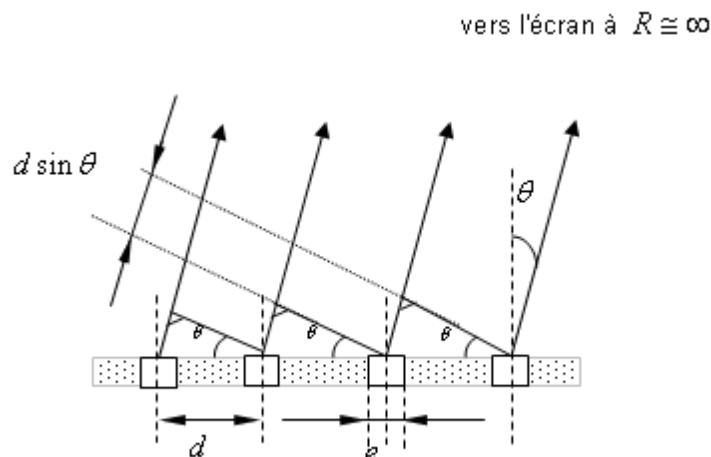


Figure: 40.6 - Principe de base du réseau de diffraction à fentes rectangulaires

Nous voyons sur le schéma ci-dessus que pour certaines directions θ , la distance $d \sin \theta$ est telle que des interférences constructives ou destructives se réalisent.

Posons que le réseau est placé dans le plan YZ et que la direction du faisceau se fait selon l'axe X . Plaçons-nous en un point d'observation P situé dans le plan XY . Selon les propriétés des ondes électromagnétiques (cf. chapitre d'Électrodynamique), le vecteur champ électrique $\vec{E}_i$ de l'onde émise par la i -ème fente est perpendiculaire à la direction d'observation et peut s'exprimer par:

$$E_i(\theta, R, t) = \frac{\hat{E}_i(\theta)}{R} \sin(k \cdot R - \omega \cdot t) \quad (40.42)$$

et nous avons aussi vu plus haut que:

$$f(R, t) = \frac{\hat{f}}{R} \frac{\sin\left(\frac{\pi e}{\lambda} \sin \theta\right)}{\frac{\pi e}{\lambda} \sin \theta} \sin(kR - \omega t) \quad (40.43)$$

d'où par analogie entre ces deux relations, nous identifions que:

$$\hat{E}_i(\theta) = \hat{f} \frac{\sin\left(\frac{\pi e}{\lambda} \sin \theta\right)}{\frac{\pi e}{\lambda} \sin \theta} \quad (40.44)$$

et il vient alors:

$$\begin{aligned} \frac{\hat{E}_i(\theta)}{R} \sin(k \cdot R - \omega \cdot t) &= \frac{\hat{E}_{i, \theta=0}}{R} \frac{\sin\left(\frac{\pi e}{\lambda} \sin \theta\right)}{\frac{\pi e}{\lambda} \sin \theta} \sin(kR - \omega t) \\ \Rightarrow \frac{\hat{E}_i(\theta)}{R} &= \frac{\hat{E}_{i, \theta=0}}{R} \frac{\sin\left(\frac{\pi e}{\lambda} \sin \theta\right)}{\frac{\pi e}{\lambda} \sin \theta} \Rightarrow \hat{E}_i(\theta) = \hat{E}_{i, \theta=0} \frac{\sin\left(\frac{\pi e}{\lambda} \sin \theta\right)}{\frac{\pi e}{\lambda} \sin \theta} \end{aligned} \quad (40.45)$$

Dans une direction θ quelconque, les ondes issues de deux fentes adjacentes sont déphasées de $d \sin \theta$ (où d est la valeur du pas entre 2 fentes) et au point P d'observation, le champ électrique résultant est donné par la somme des contributions de chaque fente avec son $i^{\text{ème}}$ décalage propre. D'où:

$$\begin{aligned} E(\theta, R, t) &= \sum_{i=1}^N E_i(\theta, R, t) = \sum_{i=1}^N \hat{E}_i(\theta) \sin(k(R + (i-1)d \sin \theta) - \omega t) \\ &= \hat{E} \sum_{i=1}^N \sin((kR - \omega t) + k(i-1)d \sin \theta) \end{aligned} \quad (40.46)$$

Nous voyons donc que chaque onde est déphasée de:

$$\Delta\varphi = kd \sin \theta = \frac{2\pi}{\lambda} d \sin \theta = c^{\text{te}} \quad (40.47)$$

Nous pouvons maintenant représenter $E(\theta, R, t)$ en utilisant les phaseurs (cf. chapitre de Mécanique Ondulatoire) dans l'espace des phases tel que:

$$E(\theta, R, t) = \hat{E} \sum_{j=1}^N e^{i[(kR - \omega t) + k(j-1)d \sin \theta]} = \hat{E} \sum_{j=1}^N e^{i(kR - \omega t)} e^{ik(j-1)d \sin \theta} \quad (40.48)$$

Ce qui donne graphiquement pour le deuxième terme contenant la variable de sommation j pour une distance R fixe:

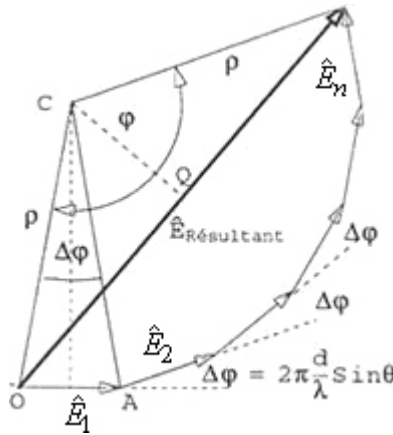


Figure: 40.7 - Représentation de la sommation des termes

Nous voyons que les $\vec{E}_i$ mis bout à bout forment un polygone régulier, inscrit dans un cercle de rayon:

$$\rho = \frac{\frac{\hat{E}_i}{2}}{\sin\left(\frac{\Delta\varphi}{2}\right)} \quad (40.49)$$

La norme du champ électrique résultant étant égale à la corde définie par l'angle:

$$\varphi = \sum_{i=1}^n \Delta\varphi \quad (40.50)$$

nous aurons:

$$\hat{E}_{\text{résultant}} = 2\rho \sin\left(\frac{\varphi}{2}\right) = 2 \cdot \frac{\hat{E}_i}{2} \frac{\sin\left(\frac{n \cdot \Delta\varphi}{2}\right)}{\sin\left(\frac{\Delta\varphi}{2}\right)} = \hat{E}_i \frac{\sin\left(n \frac{\pi}{\lambda} d \sin\theta\right)}{\sin\left(\frac{\pi}{\lambda} d \sin\theta\right)} \quad (40.51)$$

L'énergie lumineuse (in extenso l'intensité) émise dans la direction θ étant proportionnelle au carré du champ électrique (cf. chapitre d'Électrodynamique), nous avons alors pour les interférences destructives ou constructives:

$$\bar{S}(\theta, N, d) = \bar{S}_i(\theta) \left(\frac{\sin\left(n \frac{\pi}{\lambda} d \sin\theta\right)}{\sin\left(\frac{\pi}{\lambda} d \sin\theta\right)} \right)^2 \quad (40.52)$$

Nous substituons maintenant $\bar{S}_i(\theta)$ par l'expression trouvée lors de notre étude plus haut de la diffraction par une seule fente:

$$\bar{S} \propto \bar{S}_0 \left(\frac{\sin \left(\frac{\pi e}{\lambda} \sin \theta \right)}{\frac{\pi e}{\lambda} \sin \theta} \right)^2 = \bar{S}_0 \left(\frac{\sin (\pi \Gamma \sin \theta)}{\pi \Gamma \sin \theta} \right)^2 \quad (40.53)$$

Ainsi, nous obtenons pour l'addition des effets d'interférences et de diffraction:

$$\bar{S}(\theta, N, d) = \bar{S}_i(\theta) \left(\frac{\sin \left(n \frac{\pi}{\lambda} d \sin \theta \right)}{\sin \left(\frac{\pi}{\lambda} d \sin \theta \right)} \right)^2 = \bar{S}_0 \left(\frac{\sin \left(\frac{\pi e}{\lambda} \sin \theta \right)}{\frac{\pi e}{\lambda} \sin \theta} \right)^2 \left(\frac{\sin \left(n \frac{\pi}{\lambda} d \sin \theta \right)}{\sin \left(\frac{\pi}{\lambda} d \sin \theta \right)} \right)^2 \quad (40.54)$$

Bien que cette relation semble compliquée, tous ses paramètres n'ont pas la même importance pratique.

En effet considérons la fonction:

$$\psi = \left(\frac{\sin(Nx)}{\sin(x)} \right)^2 = \frac{A}{B} \quad (40.55)$$

Le terme A présente des maxima lorsque:

$$Nx_{\max} = \pi/2 + m\pi \Leftrightarrow x_{\max} = \pi/2N + m\pi/N \quad (40.56)$$

et des valeurs nulles si:

$$Nx_{\text{nul}} = m\pi \Leftrightarrow x_{\text{nul}} = m\pi/N \text{ avec } m \in \mathbb{N} \quad (40.57)$$

Bien que le terme B fasse diverger la relation pour $x = m\pi$, la règle de l'Hospital (cf. chapitre de Calcul Différentiel Et Intégral) nous donne que:

$$\lim_{x \rightarrow m\pi} \left(\frac{\sin(Nx)}{\sin(x)} \right)^2 = \left(\lim_{x \rightarrow m\pi} \frac{\sin(Nx)}{\sin(x)} \right)^2 = \left(\lim_{x \rightarrow m\pi} \frac{d_x \sin(Nx)}{d_x \sin(x)} \right)^2 = \left(\lim_{x \rightarrow m\pi} \frac{N \cos(Nx)}{\cos(x)} \right)^2 = (N)^2 \quad (40.58)$$

Il en résulte que pour $x = m\pi$ et donc des valeurs nulles de A et de B, la fonction ψ présente des énormes pics de hauteur $\psi = N^2$.

Vu leur grande amplitude, les pics principaux sont ceux que l'on observe expérimentalement le plus facilement. Ainsi, la position angulaire des maxima de la fonction ψ est donnée par:

$$B = \sin(x_{\text{pics principaux}}) = 0 \Rightarrow x_{\text{pics principaux}} = n\pi, n \in \mathbb{Z} \quad (40.59)$$

La valeur de n , qualifie le "**numéro d'ordre du maximum d'interférence**".

Appliquons ses résultats à la relation d'interférence:

$$\bar{S}(\theta, N, d) = \bar{S}_1(\theta) \left(\frac{\sin \left(n \frac{\pi}{\lambda} d \sin \theta \right)}{\sin \left(\frac{\pi}{\lambda} d \sin \theta \right)} \right)^2 \quad (40.60)$$

Le pic d'ordre n est centré sur la valeur équivalente θ qui annule le numérateur et le dénominateur de cette fraction tel que:

$$\frac{\pi}{\lambda} d \sin \theta = m \cdot \pi \Rightarrow \sin \theta = m \frac{\lambda}{d}, m \in \mathbb{Z} \quad (40.61)$$

d'où:

$$\sin \theta_m = m \frac{\lambda}{d} \quad (40.62)$$

Ainsi, un réseau dont nous connaissons la valeur d du pas peut être utilisé pour mesurer la longueur d'onde λ d'une lumière incidente inconnue.

Cependant, si la lumière incidente est polychromatique (typiquement pour les observations astronomiques), la relation précédente nous donne pour une longueur d'onde donnée la position des franges d'interférences. Ainsi, un astronome faisant passer de la lumière polychromatique de son télescope par un réseau diffraction peut faire une analyse spectroscopique de la lumière.

La relation nous donne également que pour des valeurs fixes de m et d , plus λ est grand, plus l'angle θ l'est aussi dans un intervalle compris entre $[0, \pi/2]$. Ainsi, les raies spectrales résultant de l'incidence d'un faisceau polychromatique montrent un spectre allant du violet (faible longueur d'onde donc petit angle) au rouge (grande longueur d'onde donc grand angle).

Au moyen d'un goniomètre, nous mesurons les angles θ_m des pics principaux d'ordre m , pour le plus grand nombre possible de valeurs de m . Nous en déduisons λ de la pente du graphique:

$$\lambda = d \frac{\Delta \sin \theta_m}{\Delta n} \quad (40.63)$$

Le pied du pic est situé à $\theta_m + \Delta \theta_m$ en un endroit où le numérateur $\sin \left(n \frac{\pi}{\lambda} d \sin \theta \right)$ s'annule pour la première fois après le passage du pic.

Puisque l'argument de cette fonction augmente de $m \pi$ entre deux pics successifs (parmi tous les pics principaux et secondaires), il vaut $m \cdot (n \cdot \pi)$ à l'endroit du pic d'ordre m (pic principal donc) et doit parcourir π radians supplémentaires pour atteindre le pied du pic.

Le numérateur vaut donc:

$$n \frac{\pi}{\lambda} d \sin (\theta_m + \Delta \theta_m) = m \cdot n \cdot \pi + \pi = (m \cdot n + 1) \pi \Rightarrow \sin (\theta_m + \Delta \theta_m) = \left(m + \frac{1}{n} \right) \frac{\lambda}{d} \quad (40.64)$$

La distance angulaire $\Delta \theta_m$ entre le sommet et le pied du pic principal est donc donnée par:

$$\sin(\theta_m + \Delta\theta_m) - \sin(\theta_m) = \left(m + \frac{1}{n}\right) \frac{\lambda}{d} - m \frac{\lambda}{d} = \frac{1}{n} \frac{\lambda}{d} \quad (40.65)$$

Mais dès le premier ordre, nous avons $\theta_m \ll \Delta\theta_m$. La différence des deux sinus donne (cf. chapitre de Trigonométrie):

$$\begin{aligned} \sin(\theta_m + \Delta\theta_m) - \sin(\theta_m) &= 2 \sin\left(\frac{\theta_m + \Delta\theta_m - \theta_m}{2}\right) \cos\left(\frac{\theta_m + \Delta\theta_m + \theta_m}{2}\right) \\ &= 2 \sin\left(\frac{\Delta\theta_m}{2}\right) \cos\left(\frac{2\theta_m + \Delta\theta_m}{2}\right) \cong 2 \sin\left(\frac{\Delta\theta_m}{2}\right) \cos(\theta_m) \end{aligned} \quad (40.66)$$

Un développement de Maclaurin (cf. chapitre des Suites Et Séries) de $\sin(\alpha)$ donne lorsque l'on prend le premier terme du développement $\sin(\alpha) \cong \alpha$:

$$2 \sin\left(\frac{\Delta\theta_m}{2}\right) \cos(\theta_m) \cong \Delta\theta_m \cos(\theta_m) \quad (40.67)$$

mais nous avons aussi la relation remarquable $\cos(\theta_m) = \sqrt{1 - \sin^2(\theta_m)}$. D'où la largeur angulaire d'un pic d'ordre m :

$$\Theta_m = 2\Delta\theta_m = \frac{2}{n} \frac{\lambda}{d} \frac{1}{\cos(\theta_m)} = \frac{1}{nd} \frac{2\lambda}{\sqrt{1 - \sin^2(\theta_m)}} = \frac{1}{n} \frac{2\lambda}{\sqrt{d^2 - d^2 \sin^2(\theta_m)}} \quad (40.68)$$

Or:

$$\sin \theta_m = m \frac{\lambda}{d} \Leftrightarrow d \sin \theta_m = m \lambda \quad (40.69)$$

Donc:

$$\Theta_m = \frac{1}{n} \frac{2\lambda}{\sqrt{d^2 - (m\lambda)^2}} \quad (40.70)$$

Il est clair que deux raies superposées seront vues comme distinctes si elles sont séparées d'une distance angulaire égale à leur largeur angulaire. L'expression:

$$\sin \theta_m = m \frac{\lambda}{d} \quad (40.71)$$

établit qu'à deux positions angulaires correspondent deux longueurs d'onde. Nous pouvons donc donner la séparation de deux raies par $\Delta\lambda = \lambda_2 - \lambda_1$ au lieu de $\Delta\theta = \theta_2 - \theta_1$.

Ainsi de:

$$\Theta_m = 2\Delta\theta_m = \frac{2}{n} \frac{\lambda}{d} \frac{1}{\cos(\theta_m)} \quad (40.72)$$

nous tirons:

$$\cos(\theta_m) \Delta\theta_m = \frac{1}{n} \frac{\lambda}{d} \quad (40.73)$$

Mais:

$$\sin \theta_m = m \frac{\lambda}{d} \Rightarrow m \frac{\Delta\lambda}{d} = \Delta \sin(\theta_m) \quad (40.74)$$

Lorsque $\Delta\theta_m$ et θ_m sont petits, nous avons:

$$\Delta \sin(\theta_m) \cong \cos(\theta_m) \Delta\theta_m \quad (40.75)$$

Ce qui nous amène à écrire par substitution:

$$m \frac{\Delta\lambda}{d} \cong \frac{1}{n} \frac{\lambda}{d} \Rightarrow \frac{\Delta\lambda}{\lambda} \cong \frac{1}{m \cdot n} \quad (40.76)$$

Le pouvoir de résolution R d'un réseau représente sa capacité de séparer deux raies spectrales de longueurs d'onde λ et $\lambda + \Delta\lambda$ voisines tel que:

$$R = \frac{\lambda}{\Delta\lambda} \cong m \cdot n \quad (40.77)$$

Nous voyons que le pouvoir de résolution augmente proportionnellement à l'ordre de diffraction.

2.3. FENTES DE YOUNG

Selon le principe de la dualité onde-corpuscule, la lumière se comporte à la fois comme une onde et comme un corpuscule (particule matérielle). C'est la résolution de problèmes comme ceux du corps noir (cf. chapitre de Thermodynamique), de l'effet photoélectrique (cf. chapitre de Physique Nucléaire) ou encore celui de l'effet Compton (cf. chapitre de Physique Nucléaire) qui a révélé l'existence de cette dualité.

Mais nous allons nous maintenant étudier la manière la plus flagrante mettant en évidence l'aspect ondulatoire de la matière à l'échelle atomique à l'aide de l'expérience des fentes de Young. Nous allons aborder celle-ci de manière simplifiée comme un cas particulier du réseau de fentes rectangulaires mais ayant l'avantage de mettre expérimentalement en évidence de manière aisée le comportement dual et probabiliste de la matière à l'échelle atomique.

Soit une source de lumière S , qui rayonne une onde monochromatique Ψ de longueur d'onde λ à travers deux fentes F_1 et F_2 percées dans un obstacle opaque à la lumière, comme le montre la figure ci-dessous:

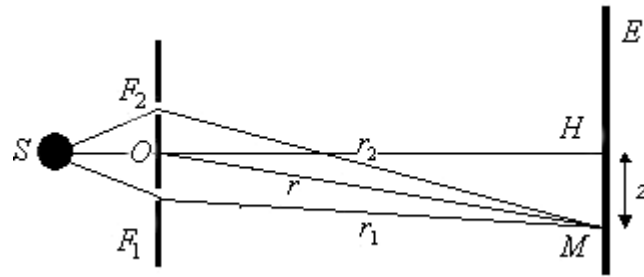


Figure: 40.8 - Mise en place de l'expérience des fentes de Young

Remarque: L'intérêt du dispositif est qu'il permet de produire deux sources de lumières cohérentes. C'est-à-dire deux sources dont la différence de phase est constante tout au long de l'expérience.

Nous disposons un écran d'observation E en un point H tel que la distance:

$$\overline{OH} = D \gg \overline{F_1 F_2} = a \quad (40.78)$$

où a serait typiquement de l'ordre du millimètre et D du mètre.

L'onde Ψ donnera après son passage à travers les fentes F_1 et F_2 , comme nous l'avons déjà vu, naissance à deux ondes "filles" $\Psi_1(r_1, t)$ et $\Psi_2(r_2, t)$ de même pulsation ω qui emprunteront respectivement les chemins r_1 et r_2 et qui iront interférer au point M de l'écran E .

Si l'interférence en M est constructive, ce point sera alors situé sur une frange brillante et si l'interférence en M est destructive, il sera sur une frange obscure. Pour observer cela, écrivons d'abord l'onde résultante au point M :

$$\Psi_M = \Psi_1(r_1, t) + \Psi_2(r_2, t) \quad (40.79)$$

dans laquelle nous avons en termes de phaseurs (cf. chapitre de Mécanique Ondulatoire):

$$\Psi_1(r_1, t) = A e^{i(kr_1 - \omega t)} \text{ et } \Psi_2(r_2, t) = A e^{i(kr_2 - \omega t)} \quad (40.80)$$

où A est l'amplitude, k est le vecteur d'onde et t représente la variable temps comme nous l'avons déjà étudié en détail dans le chapitre de Mécanique Ondulatoire.

Maintenant, faisons un changement de variable (histoire de ne pas avoir à trimbalier de longues exponentielles):

$$\Psi_1(r_1, t) = A e^{ikD_1} \text{ et } \Psi_2(r_2, t) = A e^{ikD_2} \quad (40.81)$$

Remarque: Nous verrons plus loin qu'au fait $D_1 = r_1$ et $D_2 = r_2$

Pour le calcul de l'intensité au point M , nous allons prendre la norme complexe (module) de Ψ_M ce qui s'écrit donc comme le produit du complexe et de son conjugué:

$$I = |\Psi_M|^2 = \Psi_M \overline{\Psi_M} = (\Psi_1 + \Psi_2) \overline{(\Psi_1 + \Psi_2)} \quad (40.82)$$

Remarque: Ce calcul est très important, car l'analogie avec la physique quantique ondulatoire est très forte à ce niveau et similaire au calcul de l'amplitude de probabilité (cf. chapitre de Physique Quantique Ondulatoire).

Donc:

$$\begin{aligned} I &= A^2 \left(e^{ikD_1} + e^{ikD_2} \right) \left(e^{-ikD_1} + e^{-ikD_2} \right) = A^2 \left[1 + e^{ik(D_1-D_2)} + e^{ik(D_2-D_1)} + 1 \right] \\ &= A^2 \left[2 + e^{ik(D_1-D_2)} + e^{ik(D_2-D_1)} \right] = A^2 \left[2 + \cos(k(D_1-D_2)) + \cos(k(D_2-D_1)) \right] \quad (40.83) \\ &= A^2 \left[2 + 2\cos(k(D_1-D_2)) \right] \end{aligned}$$

L'intensité est donc maximale si et seulement si:

$$\cos(k(D_1-D_2)) = 1 \quad (40.84)$$

Donc que:

$$k(D_1-D_2) = 2n\pi \Rightarrow D_1-D_2 = n\lambda \quad (40.85)$$

avec $n \in \mathbb{N}$. Ce qui donne:

$$D_1-D_2 = (r_1 - \omega t) - (r_2 - \omega t) = r_1 - r_2 = n\lambda \quad (40.86)$$

Remarque: C'est ici que trivialement nous voyons que $D_1 = r_1$ et $D_2 = r_2$

L'intensité est donc nulle si et seulement si:

$$\cos(k(D_1-D_2)) = -1 \quad (40.87)$$

Donc que:

$$k(D_1-D_2) = 2\left(n + \frac{1}{2}\right)\pi \Rightarrow D_1-D_2 = \left(n + \frac{1}{2}\right)\lambda \quad (40.88)$$

avec $n \in \mathbb{N}$. Ce qui donne:

$$r_1 - r_2 = \left(n + \frac{1}{2}\right)\lambda \quad (40.89)$$

Maintenant, il nous faut calculer $r_1 - r_2$ en fonction de z pour savoir ce que nous observons sur l'écran E .

Considérons pour cela le schéma suivant:

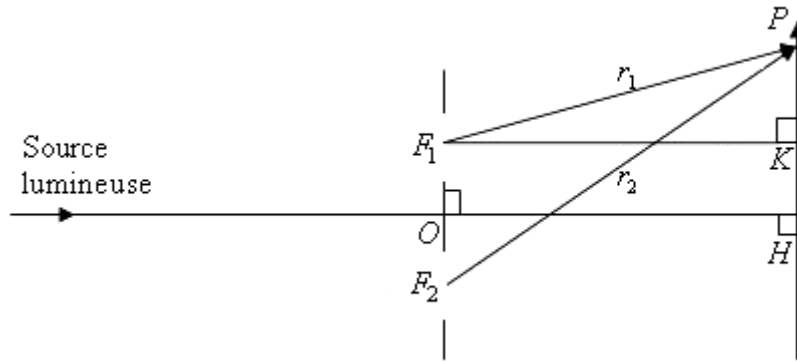


Figure: 40.9 - Agrandissement et situation particulière de l'expérience de Young

où $\overline{F_1P} = r_1$ et $\overline{F_2P} = r_2$.

Nous avons sur notre schéma:

$$r_1^2 = D^2 + \overline{PK}^2 \quad (40.90)$$

Or, $\overline{PK} = z - a/2$ donc nous avons:

$$r_1^2 = D^2 + \left(z - \frac{a}{2}\right)^2 \Rightarrow r_1 = D \sqrt{1 + \frac{1}{D^2} \left(z - \frac{a}{2}\right)^2} \quad (40.91)$$

Comme z et a sont petits devant D et en utilisant l'approximation:

$$\sqrt{1 + \varepsilon} \cong 1 + \frac{\varepsilon}{2} \quad (40.92)$$

si ε est petit devant 1. Nous avons alors:

$$r_1 = D \sqrt{1 + \frac{1}{D^2} \left(z - \frac{a}{2}\right)^2} \cong D \left(1 + \frac{1}{2D^2} \left(z - \frac{a}{2}\right)^2\right) = D + \frac{1}{2D} \left(z - \frac{a}{2}\right)^2 \quad (40.93)$$

De même:

$$r_2 = D \sqrt{1 + \frac{1}{D^2} \left(z + \frac{a}{2}\right)^2} \cong D \left(1 + \frac{1}{2D^2} \left(z + \frac{a}{2}\right)^2\right) = D + \frac{1}{2D} \left(z + \frac{a}{2}\right)^2 \quad (40.94)$$

Donc en soustrayant ces deux relations:

$$r_1 - r_2 = -z \frac{a}{D} \quad (40.95)$$

Donc finalement en utilisant la relation:

$$r_1 - r_2 = n\lambda \quad (40.96)$$

il vient:

$$z = -\frac{n\lambda D}{a} \quad (40.97)$$

Ainsi, la distance entre deux maximums consécutifs est:

$$i_{Max} = \frac{\lambda D}{a} \quad (40.98)$$

et est appelée "**interfrange**".

Pour les franges d'intensité nulle il vient immédiatement:

$$z = -\left(n + \frac{1}{2}\right) \frac{\lambda D}{a} \quad (40.99)$$

Cette relation révèle que l'intensité I présente des minima (franges obscures) et maxima (franges brillantes) distribués selon la direction z de manière périodique. Cela ne nous étonne pas plus que cela pour l'instant car cela découle du cas plus général étudié plus haut.

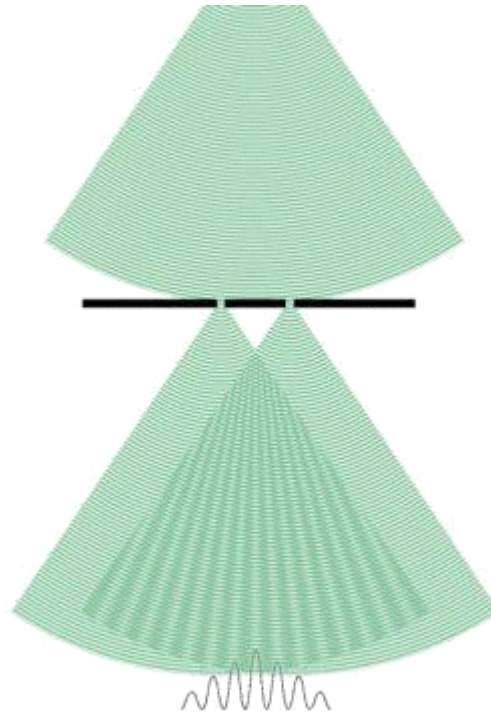


Figure: 40.10 - Représentation imagée du résultat de l'expérience de Young

Il convient cependant de préciser que les calculs précédents montrent que l'intensité des franges est partout égale. Or nous observons expérimentalement (voir la figure ci-dessus) que leur intensité diminue lorsqu'on s'éloigne du centre de l'écran. Comme nous l'avons déjà vu, deux phénomènes sont à l'origine de cette observation:

Premièrement, les fentes ont une certaine largeur, ce qui implique un phénomène de diffraction. En effet, une lumière envoyée sur un petit trou n'en ressort pas de façon isotrope. Cela se traduit par le fait que la lumière est majoritairement dirigée vers l'avant. Cet effet se répercute sur la figure observée après les fentes de Young: l'intensité des franges décroît au fur et à mesure que l'on s'éloigne du centre.

Le second phénomène à prendre en compte est le fait que les ondes émises en F_1 et F_2 sont des ondes

sphériques, c'est-à-dire que leur amplitude décroît au fur-et-à-mesure qu'elles avancent. Ainsi l'amplitude de E_1 et E_2 ne sera pas la même au point M .

Donc nos calculs restent approximatifs par rapport à l'étude que nous avons faite du réseau de fentes rectangulaires mais c'est ainsi que l'expérience des fentes de Young est présentée dans les écoles et cela suffit à mettre en évidence le résultat principal.

L'expérience originelle de Thomas Young peut donc être interprétée en utilisant les simples lois de Fresnel comme nous l'avons fait avec le réseau de fentes. Ce qui met en évidence le caractère ondulatoire de la lumière. Mais cette expérience a par la suite été raffinée, notamment faisant en sorte que la source S émette un quantum à la fois. Par exemple, on peut à l'heure actuelle émettre des photons ou des électrons ou encore des atomes un par un. Ceux-ci sont détectés un par un sur l'écran placé après les fentes de Young. Nous observons alors que ces impacts forment petit à petit la figure d'interférences. Selon des lois classiques concernant les trajectoires de ces corpuscules, il est impossible d'interpréter ce phénomène!!! D'où l'intérêt de l'étude théorique et expérimentale des fentes de Young.

De gauche à droite et de haut en bas, voici les motifs obtenus en accumulant 10, 300, 2'000 et 6'000 électrons avec un flux de 10 électrons/seconde. L'accumulation des électrons finit par constituer des franges d'interférence ce qui est assez déroutant a priori!

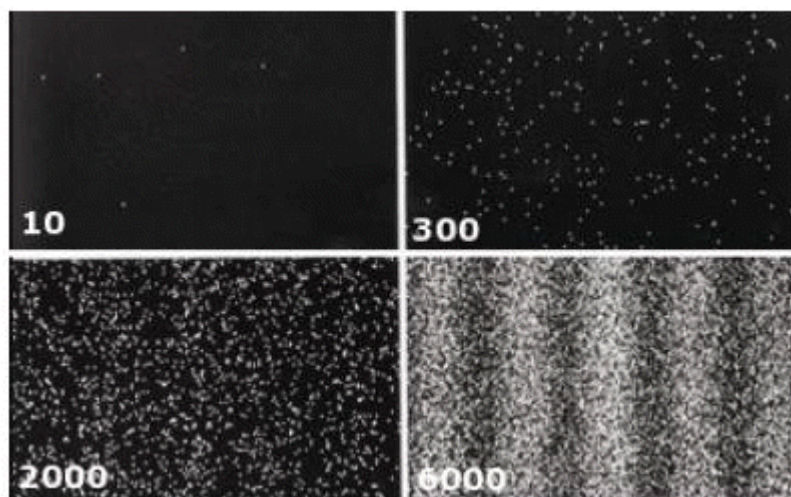


Figure: 40.11 - Photos des motifs obtenus dans une expérience de Young réelle

Nous reviendrons sur ce phénomène crucial dans le chapitre de Physique Quantique Ondulatoire pour en dire un peu plus.

3. POLARISATION DE LA LUMIÈRE

Ce n'est qu'au 19ème siècle que l'on découvrit la polarisation de la lumière (nous allons de suite expliquer de quoi il s'agit). Cependant, à l'époque de Newton, on connaissait déjà un phénomène dû à la polarisation: l'existence de cristaux dits "**cristaux biréfringents**" (tel le spath d'Islande) qui ont la propriété de réfracter un seul rayon en deux rayons distincts (aujourd'hui nous savons que les deux rayons réfractés par un tel cristal sont polarisés).

Pour comprendre ce qu'est la "**polarisation de la lumière**", revenons au cas d'une onde se propageant sur une corde (cf. chapitre de Mécanique Ondulatoire). Une telle onde peut le faire dans un plan vertical (droite) aussi bien que dans un plan horizontal (gauche) ou dans tous les plans intermédiaires:

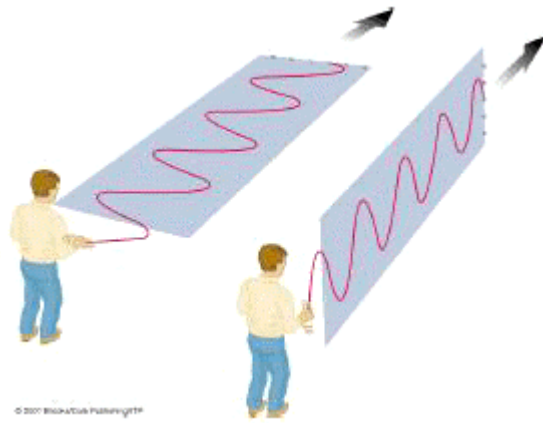


Figure: 40.12 - Représentation imagée du concept de polarisation de la lumière

Dans les deux cas, nous disons que l'onde est "**polarisée linéairement**", ce qui signifie que les oscillations se font uniquement et toujours dans le même plan, appelé "**plan de polarisation**". Une telle onde peut passer à travers une fente verticale si elle est polarisée verticalement, une onde polarisée horizontalement ne le pourra pas.

Rappel: nous avons vu dans le chapitre d'Électrodynamique que pour les ondes électromagnétiques, le champ électrique $\vec{E}$ oscille (du moins pour la solution standard des équations de Maxwell) et est orthogonal à la direction de propagation.

Le vecteur champ électrique $\vec{E}$ d'une onde peut être décomposé en deux composantes perpendiculaires l'une à l'autre, $(\vec{E}_x, \vec{E}_y)$ si l'onde se propage dans la direction z et transportant chacune la moitié de l'intensité de l'onde. Ces deux composantes changent à tout moment lorsque $\vec{E}$ varie. Le résultat à tout instant est un champ horizontal total et un champ vertical total.

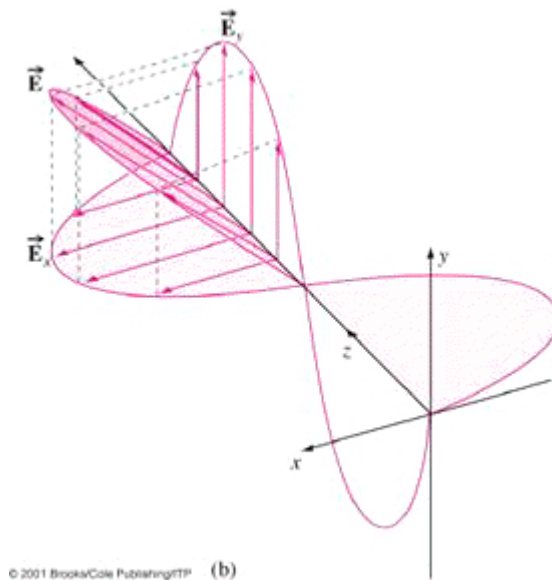


Figure: 40.13 - Illustration de la décomposition du champ

Si $\vec{E}$ tourne autour de la direction de propagation avec son extrémité décrivant un cercle, nous disons alors que l'onde est "**polarisée circulairement**":

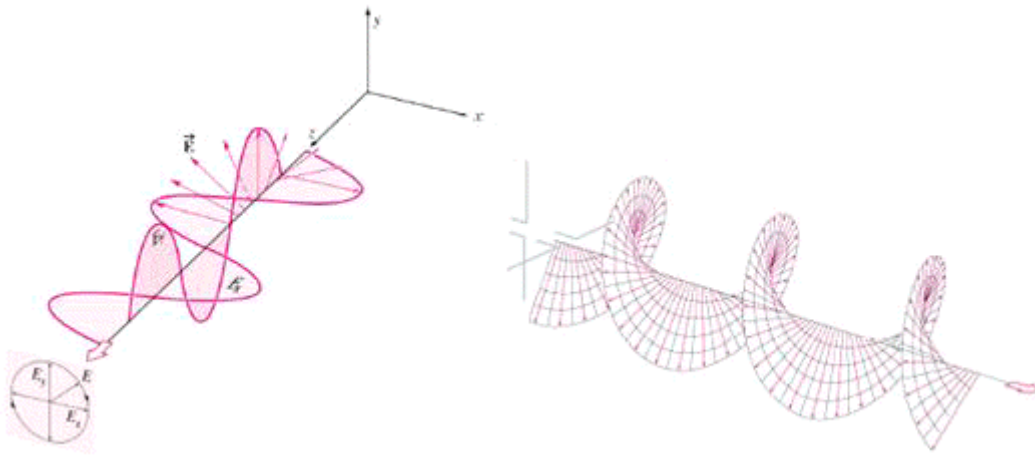


Figure: 40.14 - Représentation imagée d'une onde polarisée

$\vec{E}$ reste alors constant en module mais tourne tout en progressant, effectuant un tour complet pour chaque parcours égal à une longueur d'onde.

Remarque: La lumière n'est pas forcément polarisée ! Chaque atome émet un train d'ondes qui dure moins d'un cent-millionième de seconde (ces trains d'ondes sont parfaitement expliqués par la propagation de la particule libre en physique quantique avec les transformées de Fourier), et toutes ces ondes n'ont aucune corrélation de phase ou d'orientation. Le champ résultant en une position donnée de l'espace, est la somme géométrique de tous ces trains d'ondes: il change constamment.

Ainsi, la lumière naturelle est un mélange aléatoire et très rapidement variable d'ondes linéairement polarisées dans toutes les directions. En regardant vers la source, nous observons un champ $\vec{E}$, résultant qui oscille dans une certaine direction durant une fraction de période puis saute brusquement à une nouvelle direction aléatoire tout en restant perpendiculaire à la direction de propagation:

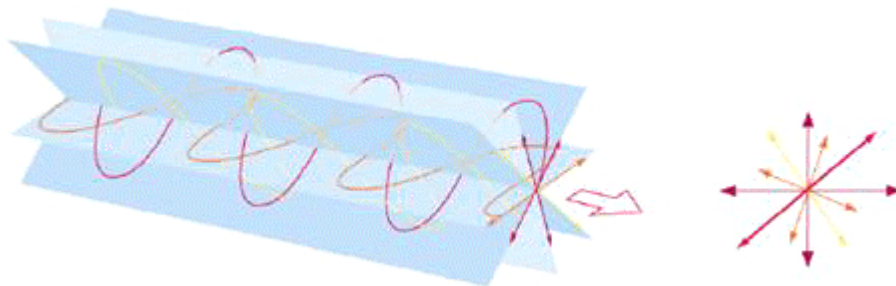


Figure: 40.15 - Représentation imagée d'une onde de la lumière naturelle

Cette introduction ayant été faite, passons à quelque chose d'un peu plus formel:

Nous avons donc vu en électrodynamique qu'une onde plane progressive monochromatique (même si physiquement elle n'existe pas...) se propageant dans le vide était composée d'un champ $\vec{E}$ et d'un champ magnétique $\vec{B}$ et était caractérisée par sa pulsation ω , son amplitude en champ électrique $\hat{E}$ et en champ magnétique $\hat{B}$ et sa direction de propagation donnée par un vecteur unitaire $\vec{u}_x, \vec{u}_y, \vec{u}_z$ au choix selon l'orientation du repère choisi.

Nous avons vu également que ces ondes possèdent des propriétés structurales remarquables, en particulier :

- $\vec{E}$ et $\vec{B}$ sont transverses, c'est-à-dire que leur direction est en tout point et à tout instant orthogonale à la direction de propagation (théorème de Malus). Ceci, permettant de définir un plan d'onde, plan généré par les deux directions de $\vec{E}$ et $\vec{B}$.
- Les normes de ces deux vecteurs sont reliées par $\hat{B} = \hat{E}/c$, où c est la vitesse de la lumière dans le vide (c'est ce rapport immense entre le champ magnétique et le champ électrique d'une onde électromagnétique qui fait que les développements présentés plus loin se font de préférence par rapport à la composante $\vec{E}$ de l'onde).
- Enfin, ces deux vecteurs sont orthogonaux entre eux, et le trièdre $(\vec{E}, \vec{B}, \vec{u}_z)$ est un trièdre orthogonal direct.

Ces trois propriétés se résument par la relation:

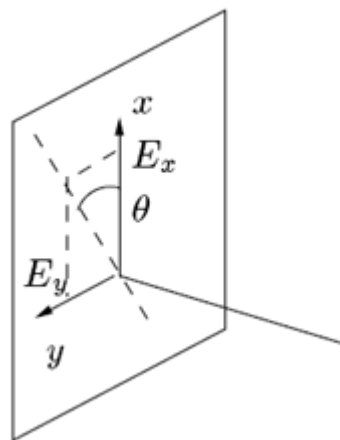
$$\vec{B} = \frac{\vec{u}_z \times \vec{E}}{c} \quad (40.100)$$

où nous avons choisi le repère tel que l'onde se propage selon la direction $\vec{u}_z$. De plus, nous avons montré que le champ électrique est une fonction d'onde trigonométrique donnée à l'arbitraire de phase près par:

$$\|\vec{E}\| = \hat{E} \cos(\omega t - kz + \varphi_1) \text{ ou } \|\vec{E}\| = \hat{E} \sin(\omega t - kz + \varphi_2) \quad (40.101)$$

Plaçons-nous maintenant dans une base (x, y, z) . L'expression la plus générale du champ électrique d'une onde plane progressive monochromatique se propageant selon $\vec{u}_z$ peut être décomposée selon deux composantes:

$$\vec{E} = \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} = \begin{pmatrix} \hat{E} \cos(\theta) \cos(\omega t - kz + \varphi_x) \\ \hat{E} \sin(\theta) \cos(\omega t - kz + \varphi_y) \\ 0 \end{pmatrix} = \begin{pmatrix} \hat{E}_{0x} \cos(\omega t - kz + \varphi_x) \\ \hat{E}_{0y} \cos(\omega t - kz + \varphi_y) \\ 0 \end{pmatrix} \quad (40.102)$$



La norme du champ étant dès lors donnée par:

$$\|\vec{E}(t)\| = \sqrt{\hat{E}^2 \cos^2(\theta) \cos^2(\omega t - kz + \varphi_x) + \hat{E}^2 \sin^2(\theta) \cos^2(\omega t - kz + \varphi_y)} \quad (40.103)$$

Si $\varphi_x = \varphi_y$ (ce qui est le cas le plus souvent) nous avons alors:

$$\|\vec{E}(t)\| = \hat{E} \cos(\omega t - kz + \varphi) \leq \hat{E} \quad (40.104)$$

En choisissant une autre origine des temps, nous pouvons toujours nous ramener à écrire:

$$\vec{E} = \begin{pmatrix} \hat{E}_{0x} \cos(\omega t - kz) \\ \hat{E}_{0y} \cos(\omega t - kz - \varphi) \\ 0 \end{pmatrix} \quad (40.105)$$

avec $\varphi = \varphi_x - \varphi_y$.

Remarque: Le choix d'écrire $-\varphi$ plutôt que $+\varphi$ nous sera utile plus tard pour l'utilisation des relations trigonométriques remarquables et nous permettra de trouver l'équation d'une ellipse (patience... ce n'est plus très loin).

En utilisant les phaseurs (cf. chapitre de Mécanique Ondulatoire) ces dernières relations peuvent se ramener à:

$$\vec{E} = \Re \begin{pmatrix} \hat{E}_{0x} e^{i(\omega t - kz)} \\ \hat{E}_{0y} e^{i(\omega t - kz - \varphi)} \\ 0 \end{pmatrix} \quad (40.106)$$

Mais la polarisation la plus générale est décrite par un vecteur complexe normalisé à l'unité dans un espace à deux dimensions de composantes:

$$\lambda = \cos \theta e^{i\varphi_x} \quad \mu = \sin \theta e^{i\varphi_y} \quad (40.107)$$

avec $|\lambda|^2 + |\mu|^2 = 1$.

Cependant, pour décrire ce champ, et donc l'ensemble de l'onde, il est commode de se placer dans le plan $z = 0$ et de décrire l'évolution du vecteur $\vec{E}$ dans ce plan. C'est ce que nous allons faire par la suite. Ceci revient en fait à choisir une origine des coordonnées selon z . Dans ce cas, nous pouvons écrire:

$$\vec{E} = \Re \begin{pmatrix} \hat{E}_{0x} e^{i(\omega t)} \\ \hat{E}_{0y} e^{i(\omega t - \varphi)} \end{pmatrix} \quad (40.108)$$

3.1. POLARISATION LINÉAIRE

Définition: Nous disons qu'une onde est "**polarisée linéairement**" lorsque $\varphi = 0$ ou $\varphi = \pi$.

Dans le premier cas ($\varphi = 0 \Rightarrow \varphi_x = \varphi_y$), nous avons:

$$\vec{E} = \Re \begin{pmatrix} \hat{E}_{0x} e^{i\omega t} \\ \hat{E}_{0y} e^{i\omega t} \end{pmatrix} \quad (40.109)$$

Dès lors, nous avons $\hat{E}_{0x}, \hat{E}_{0y}$ qui ont des valeurs comprises respectivement entre $[-\hat{E}_{0x}, +\hat{E}_{0x}]$, $[-\hat{E}_{0y}, +\hat{E}_{0y}]$.

Remarque: Relativement à un diagramme que nous verrons plus loin il convient de prendre en compte que lorsqu'une composante est positive l'autre l'est aussi et inversement.

Nous avons dès lors à chaque instant:

$$\left. \begin{aligned} E_x &= E_{0x} \cos(\omega t) \\ E_y &= E_{0y} \cos(\omega t) \end{aligned} \right\} \frac{E_x}{E_{0x}} = \frac{E_y}{E_{0y}} \quad (40.110)$$

ce qui signifie que le champ garde une direction fixe. D'où le fait que nous parlions d'onde polarisée linéairement.

Si $\varphi = \pi$ nous avons alors:

$$\vec{E} = \begin{pmatrix} \hat{E}_{0x} \cos(\omega t - kz) \\ \hat{E}_{0y} \cos(\omega t - kz - \pi) \\ 0 \end{pmatrix} = \begin{pmatrix} \hat{E}_{0x} \cos(\omega t - kz) \\ -\hat{E}_{0y} \cos(\omega t - kz) \\ 0 \end{pmatrix} = \Re \begin{pmatrix} \hat{E}_{0x} e^{i\omega t} \\ -\hat{E}_{0y} e^{i\omega t} \end{pmatrix} \quad (40.111)$$

Dès lors, nous avons $\hat{E}_{0x}, \hat{E}_{0y}$ qui ont des valeurs comprises aussi entre $[-\hat{E}_{0x}, +\hat{E}_{0x}]$, $[-\hat{E}_{0y}, +\hat{E}_{0y}]$.

Remarque: Relativement à un diagramme que nous verrons plus loin il convient de prendre en compte que lorsqu'une composante est positive l'autre est négative et inversement.

Nous avons dès lors à chaque instant:

$$\left. \begin{aligned} E_x &= \hat{E}_{0x} \cos(\omega t) \\ E_y &= -\hat{E}_{0y} \cos(\omega t) \end{aligned} \right\} \frac{E_x}{\hat{E}_{0x}} = -\frac{E_y}{\hat{E}_{0y}} \quad (40.112)$$

ce qui signifie aussi que le champ garde une direction fixe. D'où le fait que nous parlions également d'onde polarisée linéairement.

3.2. POLARISATION ELLIPTIQUE

Si φ est quelconque, et en nous plaçant en $z = 0$, nous avons:

$$\frac{E_x}{\hat{E}_{0x}} = \cos(\omega t) \text{ et } \frac{E_y}{\hat{E}_{0y}} = \cos(\omega t) \cos \varphi + \sin(\omega t) \sin \varphi \quad (40.113)$$

d'où:

$$\sin(\omega t) \sin \varphi = \frac{E_y}{\hat{E}_{0y}} - \cos(\omega t) \cos \varphi = \frac{E_y}{\hat{E}_{0y}} - \frac{E_x}{\hat{E}_{0x}} \cos \varphi \quad (40.114)$$

De plus, nous pouvons écrire:

$$\frac{E_x}{\hat{E}_{0x}} = \cos(\omega t) \Leftrightarrow \frac{E_x}{\hat{E}_{0x}} \sin \varphi = \cos(\omega t) \sin \varphi \quad (40.115)$$

En portant chacune des 2 relations précédentes au carré:

$$\begin{aligned} \sin^2(\omega t) \sin^2 \varphi &= \left(\frac{E_y}{\hat{E}_{0y}} \right)^2 - 2 \frac{E_y}{\hat{E}_{0y}} \frac{E_x}{\hat{E}_{0x}} \cos \varphi + \left(\frac{E_x}{\hat{E}_{0x}} \right)^2 \cos^2 \varphi \\ \left(\frac{E_x}{\hat{E}_{0x}} \right)^2 \sin^2 \varphi &= \cos^2(\omega t) \sin^2 \varphi \end{aligned} \quad (40.116)$$

et en sommant, nous éliminons le temps et obtenons:

$$\sin^2 \varphi = \left(\frac{E_y}{\hat{E}_{0y}} \right)^2 - 2 \frac{E_y}{\hat{E}_{0y}} \frac{E_x}{\hat{E}_{0x}} \cos \varphi + \left(\frac{E_x}{\hat{E}_{0x}} \right)^2 \quad (40.117)$$

Nous remarquons que si $\varphi = 0, \varphi = \pi$ nous retrouvons:

$$\left(\frac{E_y}{\hat{E}_{0y}} \right)^2 - 2 \frac{E_y}{\hat{E}_{0y}} \frac{E_x}{\hat{E}_{0x}} + \left(\frac{E_x}{\hat{E}_{0x}} \right)^2 = \left(\frac{E_y}{\hat{E}_{0y}} - \frac{E_x}{\hat{E}_{0x}} \right)^2 = 0 \Rightarrow \frac{E_y}{\hat{E}_{0y}} = \frac{E_x}{\hat{E}_{0x}} \quad (40.118)$$

Ceci dit, ceci est l'équation d'une ellipse:

$$y^2 + x^2 + (-2 \cos \varphi) xy + (-\sin^2 \varphi) = 0 \quad (40.119)$$

En tout point similaire à la forme générale des coniques que nous avons dans le chapitre de Géométrie Analytique:

$$\alpha y^2 + \beta x^2 + \gamma xy + \delta x + \varepsilon y + \xi = 0 \quad (40.120)$$

Dans ce cas, l'extrémité de $\vec{E}$ décrit donc une ellipse et nous parlons dès lors naturellement de "**polarisation elliptique**".

Suivant la valeur de φ , cette ellipse peut être parcourue dans un sens ou dans l'autre. Pour déterminer ce sens, dérivons l'expression du champ et plaçons-nous à $t = 0$ toujours dans le même plan d'onde en $z = 0$:

$$d_t \vec{E} = \begin{pmatrix} -\hat{E}_{0x} \omega \sin(\omega t - kz) \\ \hat{E}_{0y} \omega \sin(\omega t - kz - \varphi) \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ -\hat{E}_{0y} \omega \sin \varphi \\ 0 \end{pmatrix} \quad (40.121)$$

Ainsi:

- Si $0 < \varphi < \pi/2$ l'ellipse est parcourue dans le sens direct (inverse des aiguilles d'une montre) comme le montre la figure plus loin. Nous disons alors que la polarisation est "**elliptique gauche directe**".
- Si $\pi/2 < \varphi < \pi$ l'ellipse est parcourue dans le sens direct aussi (inverse des aiguilles d'une montre) comme le montre la figure plus loin. Nous disons alors que la polarisation est "**elliptique droite directe**".
- Si $\pi < \varphi < 3\pi/2$ l'ellipse est parcourue dans le sens horaire (sens des aiguilles d'une montre). Nous disons alors que la polarisation est "**elliptique droite indirecte**".
- Si $3\pi/2 < \varphi < 2\pi$ l'ellipse est parcourue dans le sens horaire (sens des aiguilles d'une montre) comme le montre la figure plus loin. Nous disons alors que la polarisation est "**elliptique gauche indirecte**".

3.3. POLARISATION CIRCULAIRE

Si:

$$\varphi = \pi/2, \varphi = 3\pi/2 \quad (40.122)$$

et:

$$\hat{E}_{0x} = \hat{E}_{0y} = E_0 \quad (40.123)$$

nous avons alors l'équation de l'ellipse qui se réduit à:

$$E_x^2 + E_y^2 = E_{0x,y}^2 \quad (40.124)$$

qui est l'équation d'un cercle de rayon $E_{0x,y}$, le sens étant toujours donné par le signe du sinus:

- Si $\varphi = \pi/2$ il s'agit d'une polarisation circulaire gauche
- Si $\varphi = 3\pi/2$ il s'agit d'une polarisation circulaire droite

.... voir la figure plus bas pour un schéma.

3.4. POLARISATION NATURELLE

Nous pouvons considérer l'émission d'une source comme une succession d'ondes planes progressives monochromatiques dont l'expression sera donc:

$$\vec{E} = \begin{pmatrix} \hat{E}_{0x} \cos(\omega t - kz) \\ \hat{E}_{0y} \cos(\omega t - kz - \varphi) \\ 0 \end{pmatrix} \quad (40.125)$$

Ces trains d'ondes sont donc dans un état de polarisation particulier. Cependant, cet état varie aléatoirement d'un train d'onde à l'autre, et ceci en un temps très court par rapport au temps d'intégration des détecteurs. Ceux-ci ne verront donc pas de polarisation particulière, et le champ $\vec{E}$ n'aura pas de direction particulière.

Nous parlons dès lors de "**lumière non polarisée**". Si nous superposons cette lumière à une onde polarisée, nous obtenons ce que nous appelons une "**polarisation partielle**".

Finalement, nous pouvons résumer tout ce que nous avons vu jusqu'à maintenant par la figure suivante où nous avons:

- La polarisation linéaire $\varphi = 0, \varphi = \pi, \varphi = 2\pi$
- La polarisation linéaire partielle (n'est pas représentée)
- La polarisation elliptique directe gauche $0 < \varphi < \pi/2$ ou droite $\pi/2 < \varphi < \pi$
- La polarisation elliptique indirecte droite $\pi < \varphi < 3\pi/2$ ou gauche $3\pi/2 < \varphi < 2\pi$
- La polarisation elliptique partielle (n'est pas représentée)
- La polarisation circulaire gauche $\varphi = \pi/2$ ou droite $\varphi = 3\pi/2$
- La polarisation circulaire partielle (n'est pas représentée)

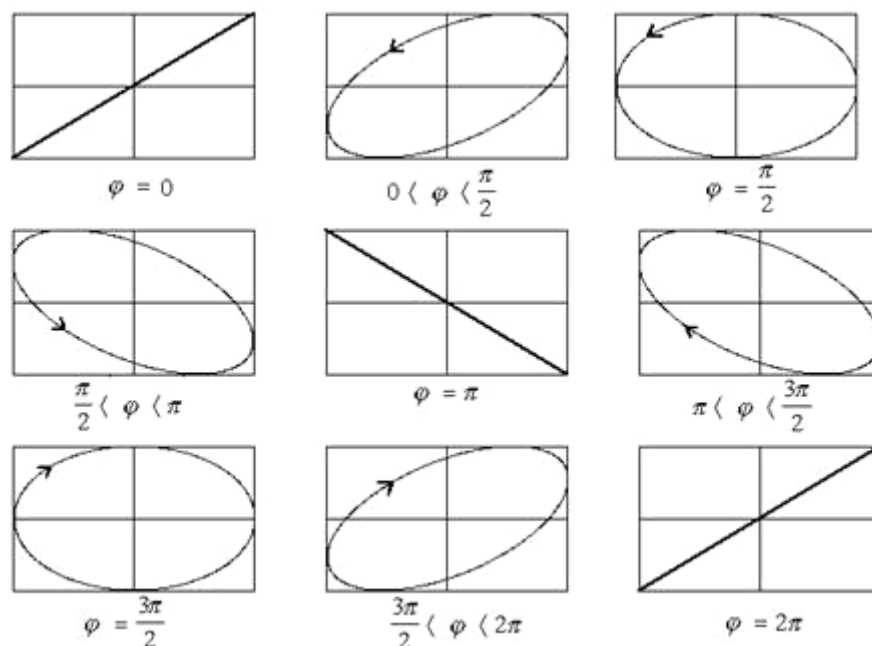


Figure: 40.16 - Représentations des différentes polarisations

Nous pouvons représenter cela de manière animée avec Maple (nous n'avons pas mis le *.gif ci-dessous

afin de ne pas trop charger le document...) et les commandes suivantes:

```
> restart;
> with (plots):
> Ex:=1;Ey:=1;phi:=Pi/4;k:=1;omega:=1;
> animate3d([x,a*Ex*cos(omega*t-k*x),a*Ey*cos(omega*t-
k*x-phi)],a=0..1,x=-10..10,t=0..2*Pi,frames=15,grid=[35,35],style=patchnogrid,axes=boxed);
```

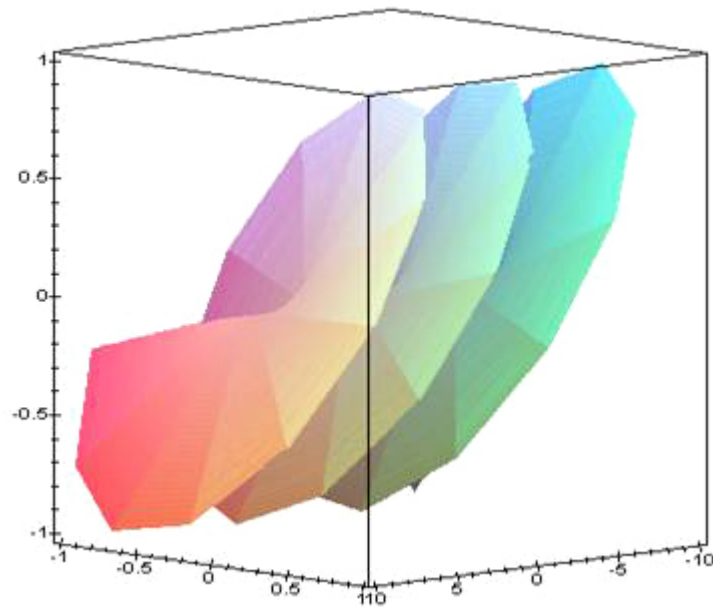


Figure: 40.17 - Animation d'une onde polarisée avec Maple

Il est bien entendu possible de modifier les paramètres. Par exemple, $\phi = \pi/2$ donne une polarisation circulaire, $\phi = \pi$ donne une polarisation rectiligne comme nous l'avons montré plus haut.

3.5. LOI DE MALUS

Pour polariser de la lumière, le physicien fera usage de polariseurs. Nous n'entrerons pas ici (car ce n'est pas dans le cadre de l'optique ondulatoire) dans les détails des propriétés atomiques ou moléculaires de la matière qui sont la cause de la polarisation de la lumière transmise.

Pour nos besoins, nous allons nous restreindre à un polariseur qui polarise une lumière incidente de manière linéaire selon l'axe x (la composante E_y étant dès lors nulle). Il vient dès lors:

$$\vec{E} = \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} = \Re \left(\begin{pmatrix} \hat{E} \cos(\theta) e^{i(\omega t - kx)} \\ \hat{E} \sin(\theta) e^{i(\omega t - kx - \psi)} \\ 0 \end{pmatrix} \right) = \Re \left(\begin{pmatrix} \hat{E} \cos(\theta) e^{i(\omega t - kx)} \\ 0 \\ 0 \end{pmatrix} \right) \quad (40.126)$$

Or, nous avons vu dans le chapitre traitant des équations de Maxwell (chapitre d'Électrodynamique) que:

$$I = \bar{S} = \frac{\hat{E}\hat{B}}{2\mu_0} = \frac{\hat{E}^2}{2\mu_0 c} \quad (40.127)$$

Dès lors, il vient pour l'intensité maximale (telle que $e^{i(\omega t - kx)} = 1$):

$$I = \frac{\hat{E}^2 \cos^2(\theta)}{2\mu_0 c} = \frac{\hat{E}^2}{2\mu_0 c} \cos^2(\theta) = I_0 \cos^2(\theta) \quad (40.128)$$

relation qui constitue la non moins fameuse "**loi de Malus**".

Pour étudier de façon quantitative la polarisation, nous allons nous servir d'un ensemble polariseur/analyseur. Nous faisons d'abord passer la lumière dans un polariseur dont l'axe fait un angle θ avec l'axe x , puis dans un second polariseur, appelé "**analyseur**", dont l'axe fait un angle α avec le même axe (voir figure ci-dessous) avec:

$$\vec{n} = (\cos \alpha, \sin \alpha) \quad (40.129)$$

dont la norme est égale à l'unité !

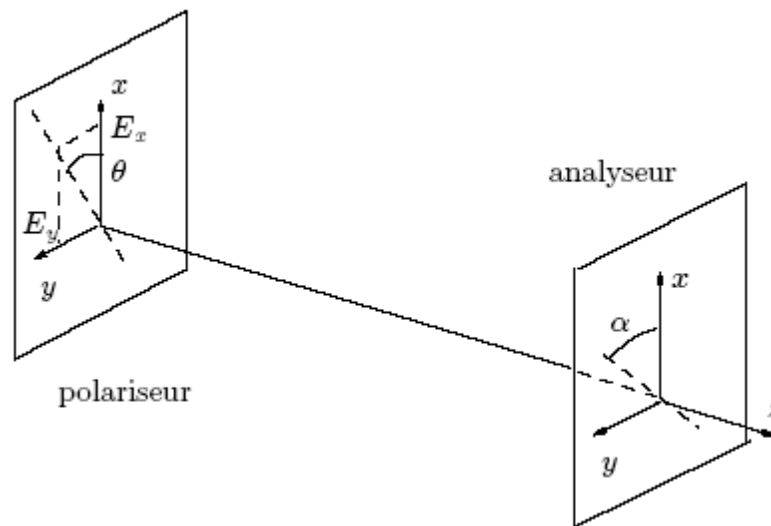


Figure: 40.18 - Exemple simplifié d'un analyseur

À la sortie de l'analyseur, le champ électrique $\vec{E}'$ s'obtient en projetant la lumière polarisée linéairement ($\varphi = 0$) obtenue à la sortie du polaroïd:

$$\vec{E} = \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} = \begin{pmatrix} \hat{E} \cos(\theta) \cos(\omega t) \\ \hat{E} \sin(\theta) \cos(\omega t) \\ 0 \end{pmatrix} = \hat{E} \cdot \vec{p} \cos(\omega t) \text{ avec } \vec{p} = (\cos \theta, \sin \theta) \quad (40.130)$$

sur $\vec{n}$ (ce qui signifie: projection=produit scalaire, pour obtenir un vecteur on multiplie par le vecteur sur lequel on projette):

$$\begin{aligned} \vec{E}' &= (\vec{E} \cdot \vec{n}) \vec{n} = (E_0 \cos(\omega t) (\vec{p} \cdot \vec{n})) \cdot \vec{n} = E_0 \cos(\omega t) (\cos \theta \cos \alpha + \sin \theta \sin \alpha) \vec{n} \\ &= [E_0 \cos(\omega t) \cos(\theta - \alpha)] \cdot \vec{n} \end{aligned} \quad (40.131)$$

Nous en déduisons la loi de Malus pour l'intensité:

$$I' = I \cos^2(\theta - \alpha) \quad (40.132)$$

dans le cas particulier de la polarisation linéaire bien sûr. Nous réutiliserons ce résultat en cryptographie quantique (cf. chapitre de Cryptographie).

4. COHÉRENCE ET INTERFÉRENCE

Nous allons maintenant voir quelles sont les conditions nécessaires à ce que des ondes planes interfèrent entre elles. Ces développements permettent de comprendre bien des choses sur la vision du monde qui nous entoure via notre oeil (surtout pourquoi l'ensemble des ondes reçues par nos rétines ne se mélangent pas et donc les couleurs non plus!).

Considérons deux ondes planes Σ_1 et Σ_2 de pulsations ω_1 et ω_2 , de vecteurs d'onde $\vec{k}_1$ et $\vec{k}_2$ se propageant toutes deux parallèlement à l'axe Oz .

Nous notons ψ_1 et ψ_2 les amplitudes complexes des deux ondes et nous nous intéressons à l'intensité moyenne observée en un point O pris comme origine des coordonnées:

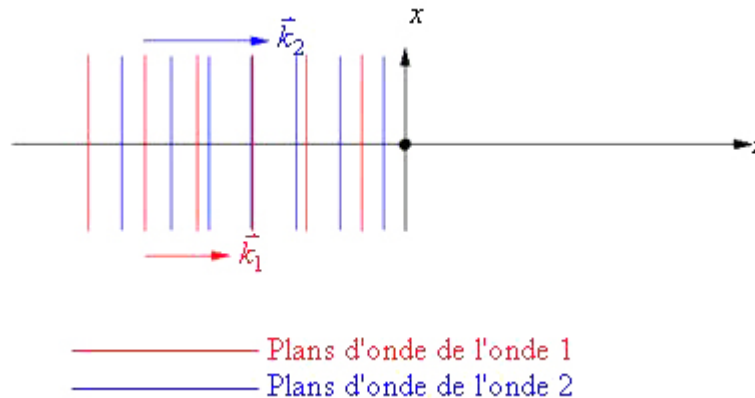


Figure: 40.19 - Représentation des ondes planes

Nous posons:

$$\omega_0 = \frac{\omega_1 + \omega_2}{2} \quad (40.133)$$

et nous supposons:

$$\delta\omega = \omega_2 - \omega_1 \ll 2\omega_0 \quad (40.134)$$

Au point O les amplitudes complexes s'écrivent

$$\begin{aligned} \psi_1 &= \psi_{01} e^{-i\omega_1 t} = |\psi_1| e^{i\theta_1} e^{-i\omega_1 t} \\ \psi_2 &= \psi_{02} e^{-i\omega_2 t} = |\psi_2| e^{i\theta_2} e^{-i\omega_2 t} \end{aligned} \quad (40.135)$$

où θ_1 et θ_2 représentent les phases de ψ_1 et ψ_2 .

Calculons maintenant l'intensité instantanée au point O qui sera notée $J(t)$. Comme l'intensité moyenne I est proportionnelle au carré de l'amplitude, nous supposons qu'il en sera de même pour l'intensité

instantanée. Ce qui nous amène à calculer la somme des parties réelles des amplitudes des deux ondes:

$$J(t) = (\mathcal{R}(\psi_1) + \mathcal{R}(\psi_2))^2 = \mathcal{R}(\psi_1)^2 + \mathcal{R}(\psi_2)^2 + 2\mathcal{R}(\psi_1)\mathcal{R}(\psi_2) \quad (40.136)$$

Ce qui s'écrit en se rappelant que (cf. chapitre sur les Nombres):

$$\mathcal{R}(\psi) = \frac{1}{2}(\psi + \bar{\psi}) \quad (40.137)$$

Soit:

$$\begin{aligned} \mathcal{R}(\psi_1)\mathcal{R}(\psi_2) &= \frac{1}{2}(\psi_1 + \bar{\psi}_1) \frac{1}{2}(\psi_2 + \bar{\psi}_2) = \frac{1}{4}(\psi_1 + \bar{\psi}_1)(\psi_2 + \bar{\psi}_2) \\ &= \frac{1}{4}(\psi_1\psi_2 + \psi_1\bar{\psi}_2 + \bar{\psi}_1\psi_2 + \bar{\psi}_1\bar{\psi}_2) = \frac{1}{4}(\psi_1\psi_2 + \bar{\psi}_1\bar{\psi}_2) + \frac{1}{4}(\bar{\psi}_1\psi_2 + \psi_1\bar{\psi}_2) \quad (40.138) \\ &= \frac{1}{2}\mathcal{R}(\psi_1\psi_2) + \frac{1}{2}\mathcal{R}(\psi_1\bar{\psi}_2) \end{aligned}$$

Et nous avons alors:

$$J(t) = \mathcal{R}(\psi_1)^2 + \mathcal{R}(\psi_2)^2 + 2\mathcal{R}(\psi_1)\mathcal{R}(\psi_2) = \mathcal{R}(\psi_1)^2 + \mathcal{R}(\psi_2)^2 + \mathcal{R}(\psi_1\psi_1) + \mathcal{R}(\psi_1\bar{\psi}_2) \quad (40.139)$$

Il en découle la somme de quatre termes:

$$\begin{aligned} J(t) &= |\psi_1|^2 \cos^2(\omega_1 t - \theta_1) \\ &+ |\psi_2|^2 \cos^2(\omega_2 t - \theta_2) \\ &+ |\psi_1\psi_2| \cos[(\omega_2 + \omega_1)t - (\theta_1 + \theta_2)] \\ &+ |\psi_1\psi_2| \cos[(\omega_2 - \omega_1)t + (\theta_1 - \theta_2)] \end{aligned} \quad (40.140)$$

Pour calculer l'intensité moyenne, nous allons choisir une approche expérimentale. L'intensité moyenne sur le temps de pose τ du détecteur (électronique ou biologique) sera donc donnée par:

$$I = \frac{1}{\tau} \int_0^\tau J(t) dt = \langle J(t) \rangle_\tau \quad (40.141)$$

I est donc la somme des moyennes des quatre termes intervenant dans $J(t)$. En lumière visible (cas de notre oeil), les fréquences sont de l'ordre de 10^{15} [Hz] et les temps de pose des détecteurs varient entre la milliseconde et la seconde. τ contient alors typiquement 10^{12} périodes de ψ_1 et ψ_2 !!

Examinons l'effet sur la valeur moyenne de chacun des termes de $J(t)$ en rappelant d'abord la relation (cf. chapitre de Mécanique Ondulatoire):

$$\omega = 2\pi f = \frac{2\pi}{T} \quad (40.142)$$

1. Nous avons en utilisant les intégrales usuelles démontrées dans le chapitre de Calcul Différentiel Et

Intégral, la valorisation suivante de l'intégrande sur un grand nombre de périodes:

$$\begin{aligned}
 I_1 &= \frac{1}{nT_1} \int_0^{nT_1} |\psi_1|^2 \cos^2(\omega_1 t - \theta_1) dt = \frac{|\psi_1|^2}{nT_1 \omega_1} \int_{-\theta_1}^{n2\pi - \theta_1} \cos^2(x) dx = \frac{|\psi_1|^2}{nT_1 2\pi f_1} \int_{-\theta_1}^{n2\pi - \theta_1} \cos^2(x) dx \\
 &= \frac{|\psi_1|^2}{n2\pi} \int_{-\theta_1}^{n2\pi - \theta_1} \cos^2(x) dx = \frac{|\psi_1|^2}{2\pi n} \left[\frac{1}{2} \left(\int_{-\theta_1}^{n2\pi - \theta_1} dx + \sin(x) \cos(x) \right) \right]_{-\theta_1}^{n2\pi - \theta_1} \\
 &= \frac{|\psi_1|^2}{2\pi n} \left[\frac{1}{2} \left([(n2\pi - \theta_1) - (-\theta_1)] + (\sin(n2\pi - \theta_1) \cos(n2\pi - \theta_1) - \sin(-\theta_1) \cos(-\theta_1)) \right) \right] \\
 &= \frac{|\psi_1|^2}{2\pi n} \left[\frac{1}{2} (n2\pi + (\sin(-\theta_1) \cos(-\theta_1) - \sin(-\theta_1) \cos(-\theta_1))) \right] = \frac{|\psi_1|^2}{2\pi n} \left[\frac{1}{2} (n2\pi) \right] = \frac{|\psi_1|^2}{2}
 \end{aligned}
 \tag{40.143}$$

Calcul que l'on note traditionnellement et très abusivement sous la forme condensée suivante:

$$I_1 = \frac{1}{2\pi} \int_0^{2\pi} |\psi_1|^2 \cos^2(\omega_1 t - \theta_1) dt = \frac{|\psi_1|^2}{2\pi} \int_0^{2\pi} \cos^2(\omega_1 t - \theta_1) dt = \frac{|\psi_1|^2}{2\pi} \pi = \frac{|\psi_1|^2}{2} \tag{40.144}$$

Nous pouvons estimer que sur un grand nombre de périodes (temps d'ouverture du détecteur), c'est cette moyenne qui sera mesurée (en l'occurrence c'est celle-ci!).

2. Nous avons de même:

$$I_2 = \frac{1}{2\pi} \int_0^{2\pi} |\psi_2|^2 \cos^2(\omega_2 t - \theta_2) dt = \frac{|\psi_2|^2}{2\pi} \int_0^{2\pi} \cos^2(\omega_2 t - \theta_2) dt = \frac{|\psi_2|^2}{2\pi} \pi = \frac{|\psi_2|^2}{2} \tag{40.145}$$

avec la même remarque que précédemment en ce qui concerne le détecteur!

3. Pour le troisième terme c'est un peu différent:

$$\begin{aligned}
 I_3 &= \frac{1}{2\pi} \int_0^{2\pi} |\psi_1 \psi_2| \cos[(\omega_2 + \omega_1)t - (\theta_1 + \theta_2)] dt = \frac{|\psi_1 \psi_2|}{2\pi} \int_0^{2\pi} \cos[(\omega_2 + \omega_1)t - (\theta_1 + \theta_2)] dt \\
 &= \frac{|\psi_1 \psi_2|}{2\pi} \left[\int_0^{2\pi} \cos((\omega_2 + \omega_1)t) \cos(\theta_1 + \theta_2) dt + \int_0^{2\pi} \sin((\omega_2 + \omega_1)t) \sin(\theta_1 + \theta_2) dt \right] \\
 &= \frac{|\psi_1 \psi_2|}{2\pi} \left[\cos(\theta_1 + \theta_2) \int_0^{2\pi} \cos((\omega_2 + \omega_1)t) dt + \sin(\theta_1 + \theta_2) \int_0^{2\pi} \sin((\omega_2 + \omega_1)t) dt \right] \\
 &= \frac{|\psi_1 \psi_2|}{2\pi} \left[\cos(\theta_1 + \theta_2) \int_0^{2\pi} \cos(2\omega_0 t) dt + \sin(\theta_1 + \theta_2) \int_0^{2\pi} \sin(2\omega_0 t) dt \right]
 \end{aligned}
 \tag{40.146}$$

Or, la moyenne d'un cosinus et d'un sinus sur une période est nulle. Donc si le détecteur fait une mesure sur un temps d'exposition supérieur à $1/2\omega_0$, soit sur un grand nombre de périodes, nous aurons:

$$I_3 = \frac{|\psi_1 \psi_2|}{2\pi} \left[\cos(\theta_1 + \theta_2) \int_0^{2\pi} \cos(2\omega_0 t) dt + \sin(\theta_1 + \theta_2) \int_0^{2\pi} \sin(2\omega_0 t) dt \right] \quad (40.147)$$

$$= \frac{|\psi_1 \psi_2|}{2\pi} [\cos(\theta_1 + \theta_2) \cdot 0 + \sin(\theta_1 + \theta_2) \cdot 0] = 0$$

4. Pour le quatrième terme c'est encore différent dans l'approximation expérimentale. Effectivement:

$$I_4 = \frac{1}{2\pi} \int_0^{2\pi} |\psi_1 \psi_2| \cos[(\omega_2 - \omega_1)t + (\theta_1 - \theta_2)] dt = \frac{|\psi_1 \psi_2|}{2\pi} \int_0^{2\pi} \cos[(\omega_2 - \omega_1)t + (\theta_1 - \theta_2)] dt$$

$$= \frac{|\psi_1 \psi_2|}{2\pi} \left[\int_0^{2\pi} \cos((\omega_2 - \omega_1)t) \cos(\theta_1 - \theta_2) dt - \int_0^{2\pi} \sin((\omega_2 - \omega_1)t) \sin(\theta_1 - \theta_2) dt \right] \quad (40.148)$$

$$= \frac{|\psi_1 \psi_2|}{2\pi} \left[\cos(\theta_1 - \theta_2) \int_0^{2\pi} \cos((\omega_2 - \omega_1)t) dt - \sin(\theta_1 - \theta_2) \int_0^{2\pi} \sin((\omega_2 - \omega_1)t) dt \right]$$

$$= \frac{|\psi_1 \psi_2|}{2\pi} \left[\cos(\theta_1 + \theta_2) \int_0^{2\pi} \cos(\delta\omega t) dt - \sin(\theta_1 + \theta_2) \int_0^{2\pi} \sin(\delta\omega t) dt \right]$$

Or, $\delta\omega \ll 2\omega_0$. Donc le détecteur n'a pas le temps de mesurer l'intensité moyenne sur une période entière en première approximation puisque:

$$T = \frac{2\pi}{\delta\omega} \quad (40.149)$$

et que cette valeur est beaucoup beaucoup plus grande dans le spectre du visible que le temps d'ouverture/échantillonnage de l'oeil qui est lui de 0.1 [s].

Ainsi, nous noterons la moyenne du quatrième terme par:

$$I_4 = \frac{|\psi_1 \psi_2|}{2\pi} \int_0^{2\pi} \cos[(\omega_2 - \omega_1)t + (\theta_1 - \theta_2)] dt = |\psi_1 \psi_2| \langle \cos[\delta\omega t + (\theta_1 - \theta_2)] \rangle_\tau \quad (40.150)$$

$$= f(\tau) |\psi_1 \psi_2|$$

L'intensité moyenne vaut donc dans un cadre expérimental:

$$I = \frac{|\psi_1|^2}{2} + \frac{|\psi_2|^2}{2} + f(\tau) |\psi_1 \psi_2| \quad (40.151)$$

ou:

$$I = I_1 + I_2 + 2f(\tau) \sqrt{I_1 I_2} \quad (40.152)$$

Si les pulsations ω_1, ω_2 sont égales (ou pratiquement égales), c'est alors l'interférence entre deux ondes planes monochromatiques. L'intensité moyenne s'écrit alors:

$$\begin{aligned}
 I &= I_1 + I_2 + 2\sqrt{I_1 I_2} f(\tau) = I_1 + I_2 + 2\sqrt{I_1 I_2} \langle \cos(\theta_1 - \theta_2) \rangle_\tau \\
 &= I_1 + I_2 + 2\sqrt{I_1 I_2} \frac{1}{\tau} \int_0^\tau \cos(\theta_1 - \theta_2) dt = I_1 + I_2 + 2\sqrt{I_1 I_2} \cos(\theta_1 - \theta_2) \quad (40.153)
 \end{aligned}$$

L'intensité mutuelle est non nulle et nous disons alors qu'il y a cohérence. Dans le cas contraire, si les deux pulsations sont très différentes, la moyenne $f(\tau)$ est nulle et nous avons alors:

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2} f(\tau) = I_1 + I_2 \quad (40.154)$$

Le terme d'interférences a disparu, l'intensité moyenne est la somme des intensités moyennes des deux ondes. Nous disons dans ce cas que les deux ondes sont incohérentes entre elles.

Quand nous savons que l'oeil interprète l'intensité pour former les perceptions des objets nous comprenons pourquoi deux objets de deux couleurs différentes ne forment pas une perception correspondant à un mélange des deux couleurs car même si dans le spectre du visible, les pulsations sont presque égales, leur déphasage en un point donné de l'espace est rarement nul tel que:

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2} \cos(0) = I_1 + I_2 + 2\sqrt{I_1 I_2} \quad (40.155)$$

Il n'y a donc pas interférence et nous avons en réalité:

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2} \cos(\theta_2 - \theta_1) \cong I_1 + I_2 \quad (40.156)$$

et ce d'autant plus que le déphasage n'est pas constant dans le temps et que la moyenne de déphasages fait que le troisième terme s'annule. On ne peut donc pas interférer de manière simple des ondes planes de sources différentes. Par contre, lorsque la source est identique nous retrouvons ce que font nos écrans avec les trois couleurs primaires RVB.

Lorsque $\theta_2 - \theta_1$ est un multiple de 2π , I est maximale (interférence constructive). Lorsque $\theta_2 - \theta_1$ est de la forme $(2k+1)\pi$, I est minimale. Nous avons alors une interférence destructive.

Remarque: Lors de la composition de plusieurs ondes, nous pouvons toujours considérer qu'il y a interférence. Toutefois, nous appelons "conditions d'interférences" des conditions d'observation de ces interférences, in extenso des conditions pour que le résultat de leur composition soit suffisamment stable pour être observé. Il est d'usage de parler de visibilité, ce qui restreint à la seule observation par l'oeil (humain).

Nous avons vu pour l'oeil que la fréquence temps d'échantillonnage est de $10 \text{ [s}^{-1}\text{]}$. Sachant que la lumière visible à une fréquence de $f \cong 10^{14} \text{ [s}^{-1}\text{]}$, la fréquence doit donc être stabilisée par la source pendant:

$$\frac{10}{10^{14}} = 10^{-13} \quad (40.157)$$

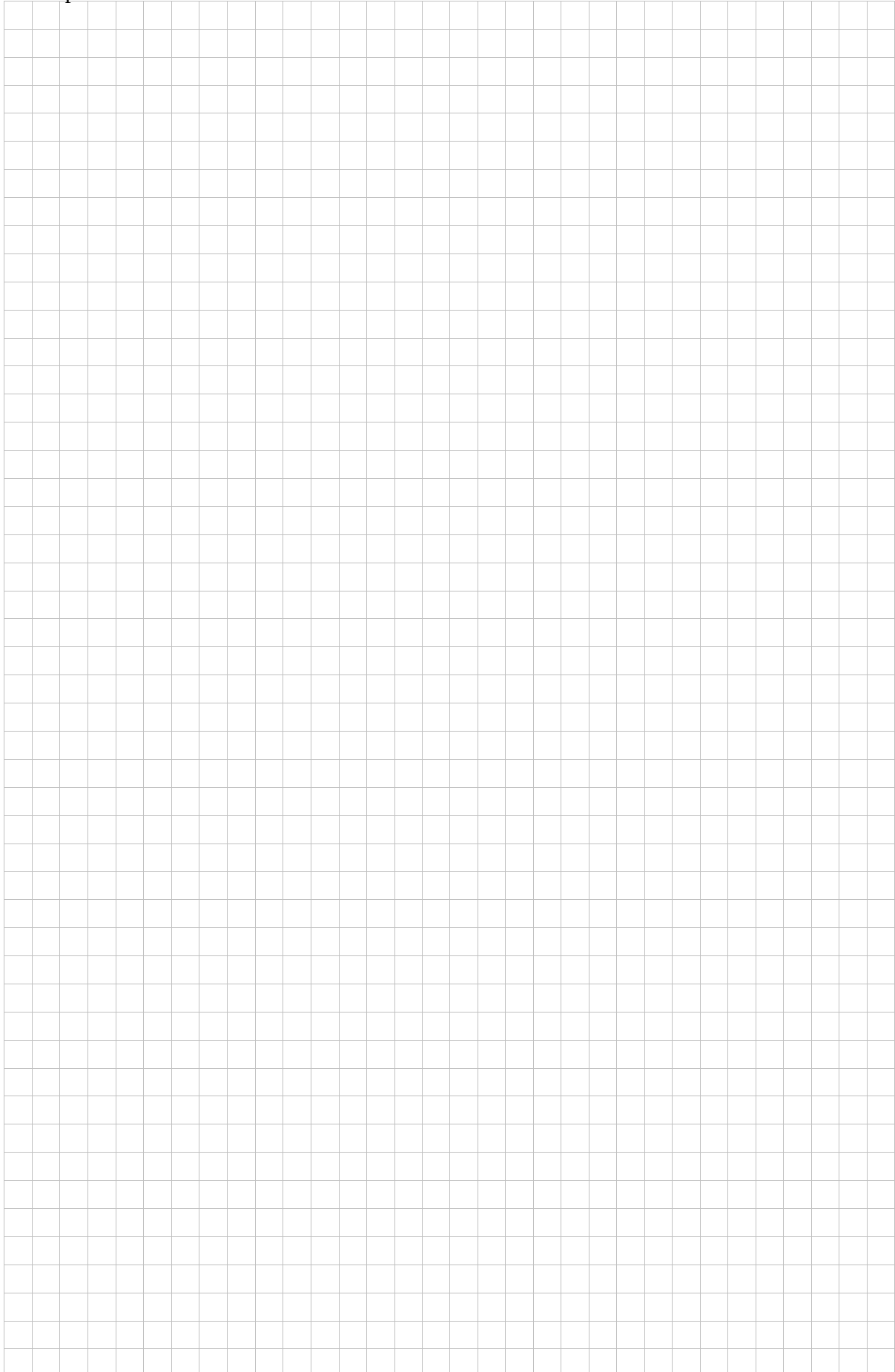
ce qui matériellement est impossible sauf à ce que la source soit la même. Nous en déduisons que pour que des interférences soient visibles à l'oeil, les sources doivent être synchrones à mieux que 10^{-13} ce

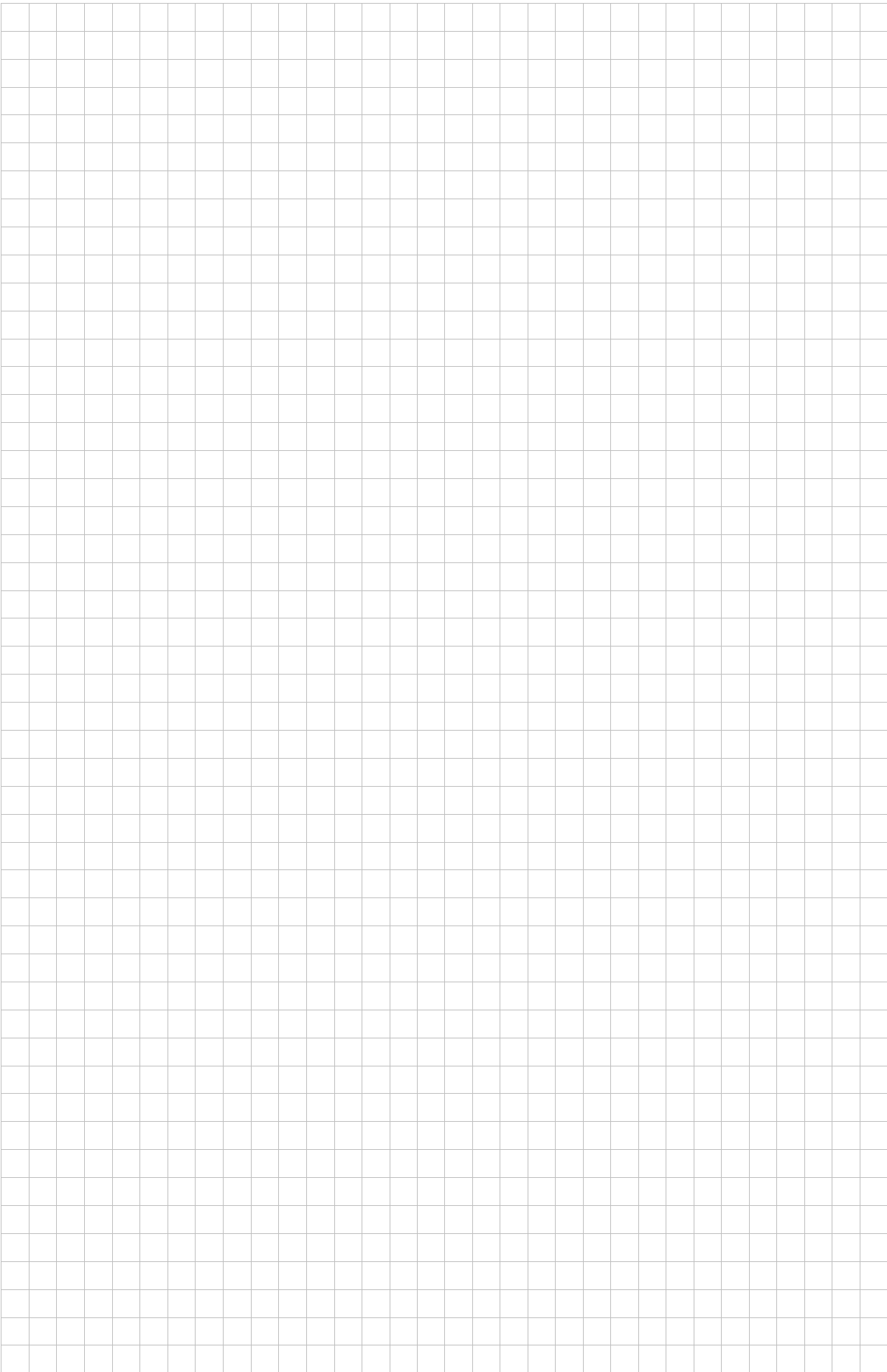
qui en pratique amène à ne considérer que des sources absolument synchronisées sur une source unique.

Dans le modèle précédent, nous avons par ailleurs négligé le fait qu'une onde réelle est limitée dans le temps. Un photon est représenté par un paquet d'onde limité. Soit T sa durée, il aura une longueur $l_c = c \cdot T$ dans le vide ou dans l'air que nous appelons "**longueur de cohérence temporelle**".

Un rayonnement donné est donc une superposition d'une succession de trains d'ondes dont la longueur moyenne est l_c , les trains d'ondes successifs n'ont pas de relation de phases entre eux: ils ne peuvent pas interférer.

Notes personnelles:











Atomistique

La physique atomique est la partie de la physique qui s'occupe des états quantifiés d'énergie de la matière corpusculaire et ondulatoire et des échanges d'énergies au sein de l'atome (Larousse).

41. PHYSIQUE QUANTIQUE CORPUSCULAIRE

Voici venu le moment de nous plonger dans les eaux obscures et impénétrables de la physique atomique.

Il va de soi que ne nous parcourons les théories de la physique atomique que dans les grandes lignes. Nous passerons ainsi sur beaucoup de détails mathématiques qui auront déjà été démontrés et vérifiés dans d'autres chapitres du site.

La physique atomique comme vous le savez déjà certainement est le monde de l'infiniment petit (points de dimension nulle). C'est un monde, vous le verrez, assez particulier où les lois classiques, celles qui gouvernent notre quotidien macroscopique, ne s'appliquent pas.

Ainsi, au début du 20ème siècle nous savions uniquement que les atomes étaient formés au plus simple par un noyau central et des électrons en orbite.

L'électron, la première particule subatomique (plus petite que l'atome) à nous être révélée, fut mis en évidence par des expériences sur les courants électriques dans les solides, les liquides et les gaz. Au 19ème siècle, les physiciens n'avaient aucune idée de ce qu'était la charge, si elle était continue ou particulaire. Aujourd'hui, nous savons que la charge est une propriété de la matière et que la charge totale dans un système est toujours un multiple d'une charge élémentaire correspondant à la charge d'un électron (ou d'un proton).

Michael Faraday suggéra par des expériences d'électrolyse que l'électricité était composée de particules de charge e et qu'une mole de ces charges (voir la section de Chimie pour la définition de la mole) était équivalente à une charge de 1 Faraday soit 96'485 [C]. Comme le nombre d'Avogadro n'était pas connu à l'époque, il n'était pas possible de déterminer e . Cependant, une mole d'une substance monovalente pouvant transporter 1 [F] de charge, il devait s'ensuivre qu'une demi-mole de la même substance devait transporter 1/2 [F] et ainsi de suite jusqu'à la plus petite unité de charge e , qui devait être transportée par la plus petite unité de masse m , correspondant à la masse d'un seul atome de cette substance. En 1881, Helmholtz affirma que si on acceptait l'hypothèse que les substances élémentaires étaient composées d'atomes, nous devions logiquement en déduire que l'électricité, tant positive que négative, devait être divisée en portions finies qui devaient se comporter comme des atomes d'électricité. Stoney nomma cette unité fondamentale de charge "**électron**". La valeur élémentaire de charge se nomme aujourd'hui prosaïquement le "**quantum de charge**".

Toutes les charges subatomiques connues aujourd'hui qu'elles soient positives ou négatives, transportent une charge nette qui est un multiple entier de e . Les quarks ont eux une charge fractionnaire mais ils n'apparaissent pas comme entités isolées. Il existe également des charges fractionnaires dans l'effet Hall quantique mais cela est une toute autre histoire...

Encore aujourd'hui, les meilleurs physiciens disent ne pas vraiment savoir ce qu'est un électron et même un atome. Au fait, on ne sait toujours pas ce qu'est vraiment la matière...

Les scientifiques ont tenté l'élaboration de plusieurs modèles pour expliquer les observations obtenues de résultats expérimentaux du monde microscopique. Ainsi, il y a eu dans l'ordre les modèles de Dalton, Thomson, Rutherford, Bohr, Sommerfeld et Schrödinger (ce dernier incluant les contributions majeures de Heisenberg, De Broglie, Pauli, Dirac et Einstein pour les plus fameux).

On peut situer la naissance de la physique quantique corpusculaire ou "**physique des quanta**" ("quantum" voulant dire "quantité fixe") en 1900, année où Max Planck présentant son célèbre article sur le rayonnement du corps noir (cf. chapitre de Thermodynamique) à une réunion de la société allemande de physique et l'incapacité de la physique classique (mécanique, thermodynamique, électromagnétisme) tenta d'expliquer certains comportement de la matière au niveau microscopique, c'est-à-dire certains phénomènes où interviennent des particules de faibles masses localisées dans de très petites régions de l'espace.



Pour parvenir à donner une interprétation cohérente de ces expériences, il a été nécessaire d'introduire des concepts radicalement différents de ceux de la physique classique. Par exemple, on a dû abandonner la notion de trajectoire, la quantification de l'énergie (loi de Planck) et considérer que les particules microscopiques ont parfois un comportement semblable à une onde. L'ensemble de ces nouveaux concepts a donné naissance à une nouvelle physique, la "physique quantique", qui s'est développée rapidement puisqu'en 1927, déjà, les fondements de la théorie sont achevés. Par son abandon des concepts-clés de la mécanique classique, on peut dire que la physique quantique constitue une véritable révolution (on l'appelle par ailleurs la "2ème révolution", la première étant la théorie de la relativité) dans notre façon d'interpréter les mesures expérimentales. Avec la relativité introduite par Einstein, la physique quantique est un des piliers de l'édifice théorique de la physique contemporaine du 21ème siècle.

Tout comme la relativité contient la mécanique classique comme cas limite (les lois relativistes approchent les lois classiques lorsque la vitesse d'une particule est suffisamment faible par rapport à celle de la lumière), la nouvelle physique quantique contient comme cas limites les lois classiques de la mécanique statistique voire même de l'électromagnétisme.

Remarque: Nous verrons que la constante fondamentale qui caractérise la physique quantique (comme la vitesse de la lumière caractérise la relativité) est la constante de Planck.

1. MODÈLE DE DALTON

En 1803, John Dalton fit l'hypothèse que la matière est composée d'atomes de différentes masses et qui se combinent en respectant des proportions massiques simples (cependant l'idée d'atome n'était pas nouvelle, elle datait de bien bien plus tôt!). C'est cette théorie que Dalton proposa qui est la pierre d'angle de la science physique moderne. En 1808, l'oeuvre de Dalton intitulée "Un nouveau système de philosophie chimique" fut publié. Dans ce livre, il dressa la liste des masses atomiques d'un certain nombre d'éléments connus par rapport à la masse de l'hydrogène. Ses masses "**U.M.A**" (cf. chapitre de Physique Nucléaire) n'étaient pas entièrement correctes, mais elles forment la base de la table périodique moderne des éléments. Dalton arriva à sa théorie atomique par une étude des propriétés physiques de l'air atmosphérique et des autres gaz.

Dalton supposa que l'atome était une sphère:



Figure: 41.1 - L'approche idéale de Dalton

Ainsi, il put faire une première estimation de la taille des atomes:

En effet, soit ρ la densité typique, m_A la masse atomique et R le rayon (valeur inconnue) d'un élément dont nous cherchons à déterminer la dimension de l'atome. Nous avons alors très simplement:

$$\rho \cdot \frac{4}{3} \pi R^3 = m_A \Leftrightarrow R = \sqrt[3]{\frac{3}{4} \cdot \frac{m_A}{\rho \cdot \pi}} \quad (41.1)$$

Connaissant m_A et ρ , nous obtenons $R \cong 1 \cdot 10^{-10} [m]$.

2. MODÈLE DE THOMSON

Thomson est à l'origine de la découverte de l'électron par ses expérimentations sur les flux de particules (électrons) créés par des rayons cathodiques. Théoricien et expérimentateur, Thomson avança en 1898 la "**théorie du pain aux raisins**" sur la structure atomique, dans laquelle les électrons sont considérés comme des raisins négatifs enfoncés dans un pain de matière positive. Son modèle de l'atome est représenté par la figure ci-dessous:

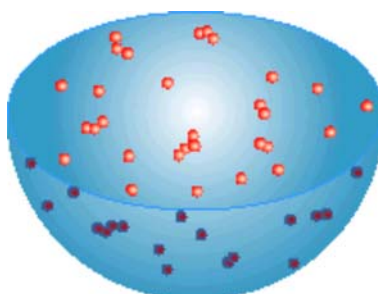


Figure: 41.2 - L'approche gourmande de Thomson

Or, nous savons (les physiciens du 19ème le savaient aussi) qu'aucun arrangement de charges

électriques statiques n'est stable si ces charges sont sous l'influence de la force de Coulomb:

$$\vec{F} = q \cdot \left(\frac{1}{4\pi\epsilon} \sum_i \frac{Q_i}{\|\vec{r} - \vec{r}_i\|^3} \cdot (\vec{r} - \vec{r}_i) \right) \quad (41.2)$$

que nous avons étudiée en détails dans le chapitre d'Électrostatique. Il faut donc que les particules qui constituent l'atome soient en mouvement ce qui nous amène à mettre en place un autre modèle: le "modèle de Rutherford" suivant:

3. MODÈLE DE RUTHERFORD

Rutherford assimila donc intuitivement par cette observation théorique, peu d'années après la découverte de Thomson, l'atome à un système planétaire. Représenté comme ci-dessous:

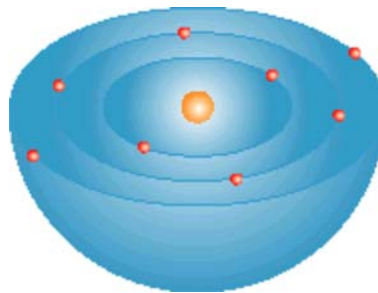


Figure: 41.3 - L'approche planétaire de Rutherford

Il appliqua les résultats que nous avons obtenus en astronomie (cf. chapitre d'Astronomie) lors de l'étude des orbites képlériennes à l'atome et obtint donc des trajectoires coniques pour la rotation de l'électron autour du noyau tel que:

$$r = \frac{p}{1 + e \cos \theta} \quad (41.3)$$

où e est l'excentricité (rapport du petit axe $e = c/a < 1$) et p le paramètre focal ($p = b^2/a$) d'une ellipse (cf. chapitre de Géométrie Analytique) et où:

$$p = \frac{K^2}{\mu} \text{ et } e = \sqrt{1 + \frac{2EK^2}{\mu^2}} \quad (41.4)$$

Remarques:

R1. Il faudra se rappeler lorsque nous aborderons plus loin le modèle de Bohr que dans le modèle de Rutherford, r peut prendre n'importe quelle valeur théoriquement!

R2. Nous verrons lors de notre étude de la diffusion de Rutherford (cf. chapitre de Physique Nucléaire) que Rutherford détermina la taille de l'atome d'or comme valant $R \cong 3 \cdot 10^{-14} [m]$. Nous avons donc un facteur 10'000 avec le modèle de Dalton (c'est dire...).

Or, nous avons vu en électromagnétisme que les équations de déplacement de Maxwell (cf. chapitre d'Électrodynamique):

$$U(t) = -\frac{d\Phi_B}{dt} = \int_V \vec{E} \circ d\vec{l} \Leftrightarrow \vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \quad (41.5)$$

et:

$$\oint_V \vec{B} \circ d\vec{l} = \mu \iint_S \vec{j} \circ d\vec{S} + \mu \cdot \varepsilon \frac{d}{dt} \iint_S \vec{E} \circ d\vec{S} \Leftrightarrow \vec{\nabla} \times \vec{H} = \vec{j} + \varepsilon \frac{d\vec{E}}{dt} \quad (41.6)$$

décrivent qu'un électron en mouvement (accélération) émet de l'énergie sous forme de rayonnement électromagnétique que nous appelons en physique le "bremstrahlung" expliqué par les potentiels de Liénard-Wiechert (cf. chapitre d'Électrodynamique).

Rutherford et Thomson se trouvèrent donc confronté au dilemme suivant:

Si l'électron émet de l'énergie sous forme de rayonnement électromagnétique, il perd donc de l'énergie cinétique (vitesse) et finira donc nécessairement un jour ou l'autre (sauf intervention extérieure) par tomber sur le noyau (illustration du phénomène dans la figure ci-dessous). Or la matière nous environnant est stable.

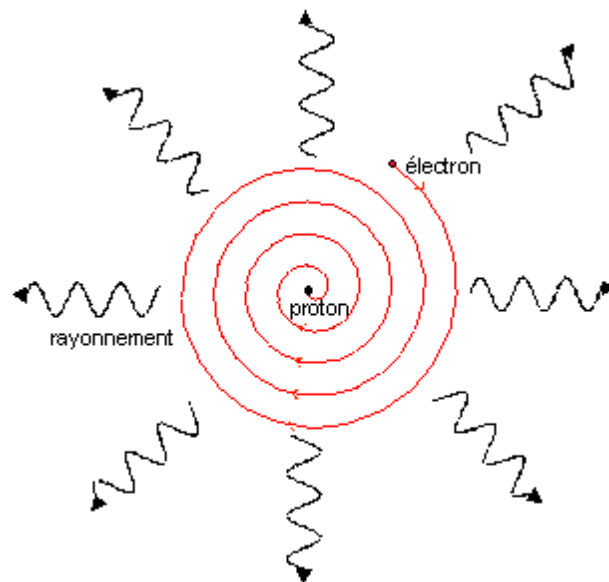


Figure: 41.4 - Illustration simpliste du bremstrahlung

Ils rejetèrent donc leur modèle et Bohr intervint à ce moment-là...

4. MODÈLE DE BOHR

En 1913, Niels Bohr, qui a participé aux travaux de Rutherford sur la diffusion des particules α (noyaux de 2 protons, 2 neutrons libres d'électrons), reprend le modèle de Rutherford mais y inclut trois postulats fondamentaux:

4.1. POSTULATS DE BOHR

P1. L'électron n'émet pas de rayonnement lorsqu'il se trouve sur certaines orbites dites "**orbites stationnaires**". Cette affirmation est contraire aux théories de l'électrodynamique. Donc ceci implique que toutes les orbites ne sont pas autorisées et constitue une véritable révolution dans l'approche de la physique.

P2. Sur toute orbite stable la quantité de mouvement p intégrée sur le chemin r est un multiple entier de la constante de Planck h (postulat découlant du premier) conformément à la quantification des échanges d'énergie établis par la relation de Planck. Ce postulat est parfois appelé "**hypothèse quantique de Planck**".

P3. La relation expérimentale (loi) de Planck:

$$\boxed{E = h\nu} \quad (41.7)$$

est valable pour l'émission ou l'absorption d'une radiation lors de la transition d'un électron d'un état énergétique E_1 vers un état E_2 (postulat qui solidifie le premier postulat).

Au fait, nous trouvons ici un concept révolutionnaire et indémontrable (aujourd'hui et à notre connaissance) qui consiste à quantifier certaines propriétés de la physique.

Continuons donc notre analyse:

4.2. QUANTIFICATION

Soit M la masse du noyau central de charge électrique $+e$ et m la masse de l'électron en "orbite". Nous faisons l'hypothèse que $M \gg m$ et que la masse centrale est immobile (ce qui est évidemment faux dans la réalité).

Nous assimilons le mouvement circulaire de l'électron autour du noyau à celui d'un oscillateur harmonique (masse reliée à un ressort exerçant une force opposée proportionnelle à une constante de rappel k_r afin de retenir l'objet lié).

Si l'oscillation a lieu dans un plan, son équation différentielle est (cf. chapitre de Mécanique Classique):

$$m \frac{d^2 x}{dt^2} + k_r x = 0 \Leftrightarrow \frac{d^2 x}{dt^2} + \omega_0^2 x = 0 \quad (41.8)$$

Une solution (particulière) de cette équation (cf. chapitre de Calcul Différentiel Et Intégral) est:

$$x = A \sin(\omega_0 t) \quad (41.9)$$

L'énergie cinétique du système étant donnée dès lors par:

$$E_c = \frac{1}{2} m \left(\frac{dx}{dt} \right)^2 = \frac{1}{2} m A^2 \omega_0^2 \cos^2(\omega_0 t) \quad (41.10)$$

et l'énergie potentielle du système par (cf. chapitre de Mécanique Ondulatoire):

$$E_p = \frac{1}{2} k_r x^2 = \frac{1}{2} m \omega_0^2 A^2 \sin^2(\omega_0 t) \quad (41.11)$$

Si nous notons ν la fréquence d'oscillation du mouvement oscillatoire, nous avons alors bien évidemment (cf. chapitre de Mécanique Ondulatoire):

$$\omega_0 = 2\pi\nu \quad (41.12)$$

L'énergie totale du système s'écrit finalement après sommation et simplification (trigonométrie élémentaire):

$$E = 2\pi^2 mA^2 \nu^2 \quad (41.13)$$

Nous admettons maintenant que l'électron lié ne peut occuper que certains niveaux d'énergie (premier postulat) selon la loi de Planck:

$$E = n\varepsilon = n\hbar\nu \quad (41.14)$$

Ce qui nous donne lorsque nous incluons la loi de Planck dans l'avant-dernière relation:

$$A^2 = n \frac{\hbar}{2\pi^2 m \nu} \quad (41.15)$$

Nous remarquons ici que puisque l'énergie de l'électron est quantifiée l'amplitude de son mouvement l'est également.

Soit à présent l'intégrale de chemin suivante (attention la notation ambiguë entre la fréquence et la vitesse peut porter à confusion) dite également "**intégrale d'action**" (il s'agit au fait du moment cinétique):

$$\oint p dr = \oint m v dr = \oint m v \frac{dr}{dt} dt = \oint m v^2 dt \quad (41.16)$$

et compte tenu de l'expression de la vitesse obtenue auparavant:

$$\oint p dr = \oint m A^2 \omega_0^2 \cos^2(\omega_0 t) dt = m A^2 \omega_0^2 \oint \cos^2(\omega_0 t) dt \quad (41.17)$$

Sur une période de révolution, nous avons:

$$\oint \cos^2(\omega_0 t) dt = \int_0^T \cos^2(\omega_0 t) dt \quad (41.18)$$

Étant donné que (cf. chapitre de Trigonométrie):

$$\cos^2(\omega_0 t) = \frac{1 + \cos(2\omega_0 t)}{2} \quad (41.19)$$

L'intégration devient:

$$\int_0^T \cos^2(\omega_0 t) dt = \frac{1}{2} \int_0^T [1 + \cos(2\omega_0 t)] dt = \frac{1}{2} \left(T + \frac{1}{2\omega_0} \sin(2\omega_0 t) \Big|_0^T \right) \quad (41.20)$$

comme $\omega = 2\pi/T$ (cf. chapitre de Mécanique Ondulatoire) nous avons:

$$\int_0^T \cos^2(\omega_0 t) dt = \frac{1}{2} \left(T + \frac{1}{2\omega_0} \left[\sin\left(2 \frac{2\pi}{T} T\right) - \sin(0) \right] \right) = \frac{1}{2} T \quad (41.21)$$

Nous obtenons donc finalement:

$$\oint p dr = \frac{1}{2} m A^2 \omega_0^2 T \quad (41.22)$$

Compte tenu que $A^2 = n \frac{h}{2\pi^2 m v}$ et $\omega_0 = 2\pi\nu$ ainsi que $\nu = 1/T$:

$$\oint p dr = \frac{1}{2} m A^2 \omega_0^2 T = \frac{1}{2} m \left(\frac{n h}{2\pi^2 m v} \right) 4\pi^2 \nu^2 \frac{1}{\nu} = n h \quad (41.23)$$

Finalement:

$$\boxed{\oint p dr = n h} \quad (41.24)$$

Cette condition imposée par Bohr (2ème postulat) résulte de la quantification des échanges d'énergie (loi de Planck). Ce qui a pour conséquence d'imposer des niveaux stationnaires d'énergie que l'électron peut occuper autour du noyau.

Pour une orbite circulaire (rappelez-vous bien que nous considérons pour l'instant une orbite circulaire!) de rayon r le moment cinétique (oui l'intégrale d'action n'est au fait que le moment cinétique) sur la longueur de l'orbitale est donc:

$$\oint p dl = m v \cdot 2\pi r = n h \Rightarrow m v r = n \frac{h}{2\pi} = n \hbar \quad (41.25)$$

ou bien en utilisant la notation traditionnelle du moment cinétique:

$$\boxed{b = m v r = n \hbar} \quad (41.26)$$

Le moment cinétique est donc quantifié!

4.3. MODÈLE DES ATOMES HYDROGÉNOÏDES SANS ENTRAÎNEMENT

Nous entendons par l'étude des "atomes hydrogénéoïdes sans entraînement" le fait de considérer des atomes avec un unique électron de masse m en rotation autour d'un noyau central de charge $Z \cdot e$ et de masse M tel que $M \gg m$ (donc le noyau est supposé fixe).

Calculons les rayons des orbites stationnaires:

Sur son orbite stationnaire, l'électron est en équilibre car il y a un antagonisme exact entre la force coulombienne et la force centrifuge. Ceci doit se traduire par l'égalité des forces suivante:

$$\frac{Z e^2}{4\pi\epsilon_0 r^2} = m \frac{v^2}{r} = m \frac{(\omega \cdot r)^2}{r} = m \omega^2 r \quad (41.27)$$

Nous posons à partir de maintenant (afin d'alléger l'écriture) que:

$$k = \frac{1}{4\pi\epsilon_0} \quad (41.28)$$

Ce qui nous permet d'écrire la relation:

$$\omega^2 r^3 = k \frac{Ze^2}{m} \quad (41.29)$$

En recourant à la condition de quantification de Bohr et en élevant au carré:

$$m\omega r^2 = n\hbar = n \frac{h}{2\pi} \Leftrightarrow \omega^2 r^4 = n^2 \frac{h^2}{4\pi^2 m^2} \quad (41.30)$$

En divisant les deux dernières relations l'une par l'autre:

$$\frac{\omega^2 r^4}{\omega^2 \cdot r^3} = \frac{\frac{n^2 h^2}{4\pi^2 m^2}}{k \frac{Ze^2}{m}} \quad (41.31)$$

nous obtenons:

$$r = \frac{1}{k} \frac{n^2 h^2}{4\pi^2 m Ze^2} = n^2 \frac{\varepsilon_0 h^2}{\pi m Ze^2} \quad (41.32)$$

compte tenu de l'expression de k .

Le rayon des orbites autorisées pour l'électron est donc:

$$r_n = n^2 \frac{\varepsilon_0 h^2}{\pi m Ze^2} \quad (41.33)$$

avec $n \in \mathbb{N}^*$ et cette relation est communément appelée le "**rayon de Bohr**" pour $n = Z = 1$.

Les orbites d'un atome selon ce modèle ressemblent donc à:

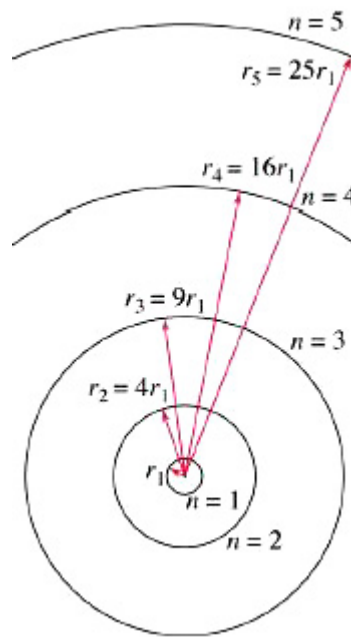


Figure: 41.5 - Le modèle planaire de Bohr

L'énergie de l'atome hydrogénoïde sans entraînement est donnée par la mécanique classique (cas d'une force centrale), somme de l'énergie cinétique et potentielle électrostatique:

$$E_{tot} = \frac{1}{2} m \omega^2 r^2 - k \frac{Ze^2}{r} \quad (41.34)$$

Avec:

$$\frac{Ze^2}{4\pi\epsilon_0 r^2} = m \omega^2 r \quad (41.35)$$

il vient:

$$E_{tot} = \frac{1}{2} m \omega^2 r^2 - k \frac{Ze^2}{r} = \frac{1}{2} k \frac{Ze^2}{r} - k \frac{Ze^2}{r} = -\frac{1}{2} k \frac{Ze^2}{r} \quad (41.36)$$

En y introduisant l'expression du rayon quantifié obtenu précédemment:

$$E_{tot} = -\frac{1}{2} k \frac{Ze^2}{r_n} = -\frac{1}{2} k \frac{Ze^2}{n^2 \frac{\epsilon_0 h^2}{\pi m Z e^2}} = -\frac{m Z^2 e^4}{8 \epsilon_0^2 n^2 h^2} \quad (41.37)$$

Nous trouvons donc que l'énergie totale de l'atome considéré est quantifiée et négative (ce qui correspond à des états stables car il faut un apport de l'énergie pour les défaire) telle que:

$$E_{tot}^n = -\frac{1}{n^2} \frac{m Z^2 e^4}{8 \epsilon_0^2 h^2} \quad (41.38)$$

Entre deux niveaux, le passage d'un électron du niveau n_2 vers un niveau n_1 (nous précisons comment lors de l'étude de l'effet photoélectrique plus loin) se traduit par l'émission d'une raie de fréquence donnée par l'expression de l'hypothèse de quantification de Planck:

$$E_{tot}^{n_2} - E_{tot}^{n_1} = h\nu_{n_2,n_1} \quad (41.39)$$

En fait, si nous admettons avec Bohr que les énergies d'un électron sur son orbite sont données par l'inverse du carré du nombre entier, la différence d'énergie entre deux orbites caractérisées par de grandes valeurs de ces nombres entiers tend vers zéro lorsque les nombres entiers tendent vers l'infini. Nous retrouvons alors un semblant de variation continue pour les énergies échangées par un atome avec le champ électromagnétique et la notion de trajectoire d'un électron prend alors à nouveau du sens.

En faisant appel à l'expression complète de l'énergie totale, nous trouvons alors la fréquence correspondante à la raie émise:

$$\nu_{n_2,n_1} = Z^2 \frac{me^4}{8\varepsilon_0^2 h^3} \left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right) \quad (41.40)$$

la longueur d'onde émise s'en déduit aisément:

$$\frac{1}{\lambda} = Z^2 \frac{me^4}{8\varepsilon_0^2 h^3 c} \left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right) = Z^2 R_H \left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right) \quad (41.41)$$

La constante R_H (notée aussi R_y selon les situations) est appelée la "**constante de Rydberg**".

Un électron qui occupe une orbite n est dans un "**état stationnaire**" si son énergie ne varie pas. En revanche, une transition directe $n_2 \rightarrow n_1$ ($n_1 < n_2$) s'accompagne de l'émission d'un photon dont l'énergie est donnée par le calcul de la fréquence comme nous allons le démontrer.

"**L'énergie d'ionisation**" est l'énergie qu'il faut fournir pour éloigner l'électron à l'infini de son orbite. Ainsi pour l'état fondamental de l'hydrogène, il faudrait poser $n_1 = 1$ et $n_2 = +\infty$.

Le résultat obtenu par Bohr pour l'expression de la fréquence en fonction des niveaux d'énergie de l'électron est un résultat formidable car le chimiste Balmer avait en 1885 (28 ans auparavant) découvert expérimentalement que le spectre des raies de l'hydrogène suivait aussi cette loi.

Balmer avait remarqué que les raies spectrales étaient extrêmement fines. Cela laissait supposer que l'énergie n'était pas émise par les atomes d'une manière continue mais seulement à certaines fréquences bien précises. En outre, cette finesse des raies explique la précision avec laquelle il avait pu déterminer la constante de Rydberg.

Les chimistes avaient également constaté que chaque élément atomique possédait son propre spectre. Il était dès lors clair que toute théorie atomique devrait rendre compte de ces 2 caractéristiques et c'est ce que fit brillamment le modèle de Bohr à l'aide des postulats des niveaux d'énergies.

Nous définissons les séries suivantes du spectre de l'atome d'hydrogène:

- Pour la série avec $n_1 = 1$ et $n_2 = 2, 3, 4, \dots$ on obtient le résultat des mesures effectuées (le spectre) par Lyman en 1906 dans l'UV.
- Pour la série avec $n_1 = 2$ et $n_2 = 3, 4, 5, \dots$ on obtient le résultat des mesures effectuées (le spectre) par Balmer en 1885 dans le visible.
- Pour la série avec $n_1 = 3$ et $n_2 = 4, 5, 6, \dots$ on obtient le résultat des mesures effectuées (le spectre) par Paschen en 1908 dans l'infrarouge.
- Pour la série avec $n_1 = 4$ et $n_2 = 5, 6, 7, \dots$ on obtient le résultat des mesures effectuées (le spectre) par Brackett en 1928 dans l'infrarouge.
- Pour la série avec $n_1 = 5$ et $n_2 = 6, 7, 8, \dots$ on obtient le résultat des mesures effectuées (le spectre) par Pfund en 1924 dans l'infrarouge.

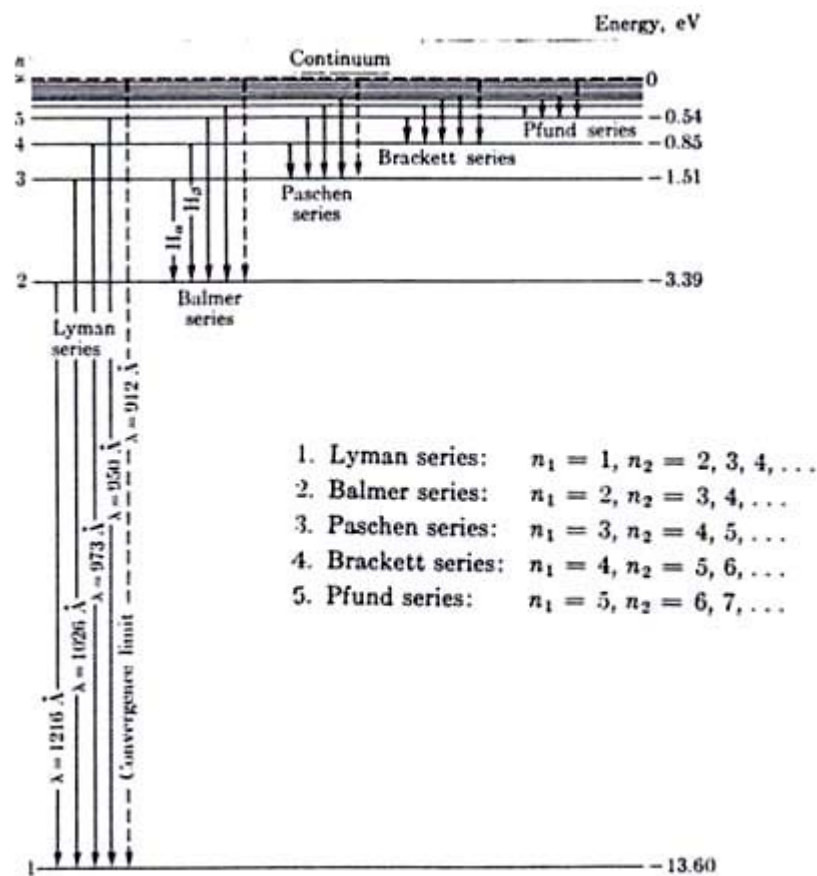


Figure: 41.6 - Quelques séries spectrales

Les quatre raies principales de la "**série de Balmer**" (visible) sont les plus connues:

$$\lambda [\text{\AA}]: H_{\alpha} \cong 6563, H_{\beta} \cong 4861, H_{\gamma} \cong 4341, H_{\delta} \cong 4102 \quad (41.42)$$

Cependant une petite différence subsistait entre la constante de Rydberg théorique et pratique (connue avec très grande précision). Ceci va conduire à complexifier le modèle:

4.4. MODÈLE DES ATOMES HYDROGÉNOÏDES AVEC ENTRAÎNEMENT

Le noyau de l'atome possède une masse M que nous avons supposée immobile par simplification. En réalité l'ensemble noyau (M) et électron (m) tourne autour d'un centre de masse commun (évidemment!).

Hypothèses:

H1. L'atome hydrogénoïde est considéré comme un système isolé.

H2. Le noyau et l'électron gravitent chacun sur une orbite circulaire autour d'un centre commun: le "centre de masse" (cf. chapitre de Mécanique Classique).

H3. Ils ont même vitesse angulaire.

L'atome hydrogénoïde étant un système isolé, le mouvement du centre de masse est soit en mouvement rectiligne et uniforme soit au repos. Il est donc licite d'y placer un système de repère inertiel.

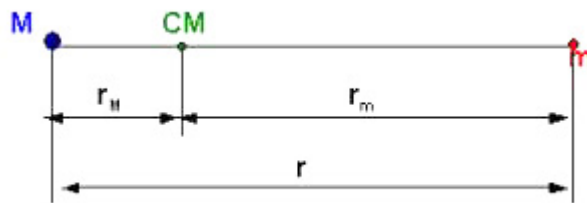


Figure: 41.7 - Mise en situation de l'atome hydrogénoïde

La définition du centre de masse dans un système de laboratoire est donnée par le théorème du centre de masse (cf. chapitre de Mécanique Classique):

$$\vec{r} = \frac{m_1 \vec{r}_1 + m_2 \vec{r}_2}{m_1 + m_2} \quad (41.43)$$

L'étude présente sera effectuée par rapport au centre de masse, la relation précédente devient donc (cf. chapitre de Mécanique Classique):

$$M \vec{r}_M + m \vec{r}_m = \vec{0} \quad (41.44)$$

De la relation précédente, en prenant la norme et la valeur absolue, il vient que:

$$r_M = \frac{m}{M} r_m \quad (41.45)$$

La distance entre le noyau et l'électron demeurant constante et égalant $r = r_M + r_m$ nous écrivons:

$$r = \frac{m}{M} r_m + r_m = \left(1 + \frac{m}{M}\right) r_m = \left(\frac{M + m}{M}\right) r_m \quad (41.46)$$

Nous en déduisons trivialement que:

$$r_m = \frac{M}{M+m} r \text{ et } r_M = \frac{m}{M+m} r \quad (41.47)$$

En appliquant la loi de la dynamique, nous écrivons que la somme des forces sollicitantes (électrostatique et centrifuge) de l'électron (uniquement) s'équilibre telle que:

$$k \frac{Ze^2}{r^2} = m\omega^2 r_m \quad (41.48)$$

que nous pouvons écrire en isolant ω^2 :

$$\omega^2 = \frac{kZe^2}{mr^2 \left(\frac{M}{m+M} \right) r} = \frac{kZe^2}{r^3} \frac{1}{\frac{mM}{m+M}} = \frac{kZe^2}{\mu r^3} \quad (41.49)$$

Nous retrouvons l'expression de la masse réduite bien connue dans un système à deux corps:

$$\mu = \frac{mM}{m+M} = \frac{m}{1 + \frac{m}{M}} \quad (41.50)$$

Attaquons-nous maintenant à la détermination de l'énergie totale de l'atome:

L'énergie cinétique de l'atome est la somme des énergies cinétiques du noyau (N) et de l'électron (e) telle que:

$$E_c = \frac{1}{2} m v_e^2 + \frac{1}{2} M v_N^2 \quad (41.51)$$

Comme $v = \omega r$ avec comme hypothèse que la pulsation est identique pour le noyau et l'électron:

$$E_c = \frac{\omega^2}{2} (m r_m^2 + M r_M^2) \quad (41.52)$$

Avec les relations des différents rayons déterminées précédemment:

$$\begin{aligned} E_c &= \frac{\omega^2 r^2}{2} \left[m \frac{M^2}{(m+M)^2} + M \frac{m^2}{(m+M)^2} \right] = \frac{\omega^2 r^2}{2} \left[\frac{mM^2 + Mm^2}{(m+M)^2} \right] \\ &= \frac{\omega^2 r^2}{2} \left[\frac{mM}{(m+M)^2} (m+M) \right] = \mu \frac{\omega^2 r^2}{2} \end{aligned} \quad (41.53)$$

et connaissant l'expression du carré de la pulsation:

$$E_c = \mu \frac{\left(k \frac{Ze^2}{\mu r^3} \right) r^2}{2} = \frac{k}{2} \frac{Ze^2}{r} \quad (41.54)$$

Par ailleurs, de l'avant-dernier développement nous tirons une relation dont nous allons faire usage plus loin:

$$E_c = \frac{\omega^2}{2} (mr_m^2 + Mr_M^2) = \mu \frac{\omega^2 r^2}{2} \Rightarrow (mr_m^2 + Mr_M^2) = \mu r^2 \quad (41.55)$$

L'énergie potentielle de l'électron par rapport au centre de masse étant donnée par (cf. chapitre d'Électrostatique):

$$E_{pot} = -\frac{kZe^2}{r} \quad (41.56)$$

L'énergie totale de l'atome hydrogénoïde est alors:

$$E_{tot} = \frac{kZe^2}{2r} - \frac{kZe^2}{r} = -\frac{kZe^2}{2r} \quad (41.57)$$

Par rapport au centre de masse, le moment cinétique total est la somme des moments cinétiques de l'électron b_m et du noyau b_M (rappelons que le moment cinétique est aussi souvent noté par la lettre L).

$$b_{tot} = b_m + b_M = mv_m r_m + Mv_M r_M = \omega(mr_m^2 + Mr_M^2) \quad (41.58)$$

La parenthèse de la dernière égalité a déjà fait l'objet d'un calcul précédemment et nous avons donc:

$$b_{tot} = \mu \omega r^2 \quad (41.59)$$

c'est ici que Bohr introduit sa condition de quantification:

$$b_{tot} = \mu \omega r^2 = n\hbar \quad (41.60)$$

or, nous connaissons l'expression détaillée du carré de la pulsation:

$$b_{tot}^2 = \mu^2 \omega^2 r^4 = \mu^2 \left(\frac{kZe^2}{\mu r^3} \right) r^4 = n^2 \hbar^2 \quad (41.61)$$

Le rayon quantifié a donc pour expression:

$$r_n = n^2 \frac{\varepsilon_0 \hbar^2}{\pi \mu Z e^2} \quad (41.62)$$

L'énergie totale de l'atome devient finalement:

$$E_{tot}^n = -\frac{kZe^2}{2 \frac{n^2 \varepsilon_0 \hbar^2}{\pi \mu Z e^2}} = -\frac{1}{n^2} \frac{\mu Z^2 e^4}{8 \varepsilon_0^2 \hbar^2} = -\frac{1}{n^2} \frac{m Z^2 e^4}{8 \varepsilon_0^2 \hbar^2} \left(\frac{1}{1 + \frac{m}{M}} \right) \quad (41.63)$$

Soit de manière condensée:

$$E_{tot}^n = -\frac{1}{n^2} h c Z^2 R_H \left(\frac{1}{1 + \frac{m}{M}} \right) \quad (41.64)$$

À partir de cette dernière relation, nous pouvons déterminer facilement l'expression (comme nous l'avons déjà fait) des longueurs d'ondes émises par une désexcitation de l'électron d'une orbite n_2 à n_1 . Calculons au même titre que nous l'avons fait pour le modèle sans entrainement, l'expression de la longueur d'onde émise lors du passage d'un niveau à l'autre. Nous faisons alors des développements identiques:

$$E_{tot}^{n_2} - E_{tot}^{n_1} = h \nu_{n_2, n_1} \quad (41.65)$$

Il vient alors:

$$\nu_{n_2, n_1} = \left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right) c Z^2 R_H \left(\frac{1}{1 + \frac{m}{M}} \right) \quad (41.66)$$

la longueur d'onde émise s'en déduit aisément:

$$\frac{1}{\lambda} = Z^2 R_H \left(\frac{1}{1 + \frac{m}{M}} \right) \left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right) \quad (41.67)$$

Remarque: Il convient bien évidemment de rendre compte que ce modèle est plus précis que le précédent.

4.5. HYPOTHÈSE DU NEUTRON

Les résultats de spectroscopie sont connus avec très grande précision, par conséquent les constantes de Rydberg également (car dépendante de la masse de l'élément atomique étudié).

Les deux raies bleues mesurées de la série de Balmer de l'hydrogène noté H ($\lambda_H \cong 4861.3 [\text{\AA}]$ composé d'un proton et d'un électron) et du deutérium D (isotope de l'hydrogène composé d'un neutron en plus) présentent une différence de longueur d'onde de $\Delta\lambda \cong 1.32 [\text{\AA}]$ Angström.

La longueur d'onde appartenant à la série de Balmer s'exprime dès lors (avec la correction du centre de masse vue précédemment) comme:

$$f = R_y \left(\frac{1}{2^2} - \frac{1}{n^2} \right) \Rightarrow \lambda = \frac{1}{R_y \left(\frac{1}{2^2} - \frac{1}{n^2} \right)} \cdot \left(1 + \frac{m}{M} \right) \quad (41.68)$$

Cette dernière expression écrite successivement pour l'hydrogène et le deutérium mène à:

$$\lambda_H = [\dots] \cdot \left(1 + \frac{m}{M_H}\right) \text{ et } \lambda_D = [\dots] \cdot \left(1 + \frac{m}{M_D}\right) \quad (41.69)$$

où nous rappelons que la masse de l'électron nous est connue! Ce qui est intéressant c'est que ces deux éléments ont des propriétés chimiques identiques (hydrogène et deutérium) mais des raies différentes. Les scientifiques de l'époque se demandaient pourquoi et après que le modèle de Bohr de l'atome hydrogénoïde avec entraînement soit à leur disposition ils ont pu conclure que cette différence dans les raies venait de la différence de la masse du noyau de l'atome.

Encore fallait-il déterminer cette différence de masse et expliquer sa provenance!

Nous avons donc:

$$\begin{aligned} \frac{\Delta\lambda}{\lambda_H} &= \frac{\lambda_H - \lambda_D}{\lambda_H} = \frac{[\dots] \left\{ \left(1 + \frac{m}{M_H}\right) - \left(1 + \frac{m}{M_D}\right) \right\}}{[\dots] \left(1 + \frac{m}{M_H}\right)} = \dots \cong \frac{1 - \frac{M_H}{M_D}}{1 + 1836.5} \\ &\Rightarrow \frac{\Delta\lambda}{\lambda_H} = \frac{1 - \frac{M_H}{M_D}}{1 + 1836.5} \Rightarrow \frac{1.32}{4861.3} 1837.5 = 1 - \frac{M_H}{M_D} \Rightarrow \frac{M_H}{M_D} \cong 0.501 \end{aligned} \quad (41.70)$$

ce qui montra aux scientifiques de l'époque que le noyau de deutérium est formé de 2 particules de masse équivalente à celle du proton. Donc par déduction logique, ce noyau se doit d'être composé d'un proton (ce que l'on sait évidemment!) et d'une particule neutre.

Cette hypothèse est celle du "**neutron**", qui fut découvert ultérieurement de manière expérimentale en 1932 par Chadwick.

5. MODÈLE DE SOMMERFELD ET WILSON

Pour élaborer leur modèle, Sommerfeld et Wilson firent appel à la dynamique classique pour généraliser le modèle de Bohr à des orbites de type képlérien (donc non uniquement circulaires mais elliptiques dans le cas général).

Comme nous l'avons vu plus haut, dans le cas d'un système à deux corps sollicités par une force centrale, l'énergie totale du système est (nous négligeons l'énergie potentielle gravitationnelle):

$$E_{tot} = \frac{1}{2}mv^2 - k \frac{Ze^2}{r} = m \left(\frac{1}{2}v^2 - k \frac{Ze^2}{mr} \right) = mE \quad (41.71)$$

Pour trouver l'expression de la trajectoire de la masse m , nous allons procéder exactement de la même manière que celle utilisée en astronomie (cf. chapitre d'Astronomie) pour déterminer les orbites képlériennes.

Ainsi, nous avons démontré dans le chapitre d'Astronomie que:

$$r(\theta) = \frac{p}{1 + e \cos \theta} \quad (41.72)$$

avec:

$$p = \frac{K^2}{\mu} = \frac{\left(r^2 \frac{d\theta}{dt}\right)^2}{Gm} \quad \text{et} \quad e = \sqrt{1 + \frac{2EK^2}{\mu^2}} = \sqrt{1 + \frac{2E_{\text{tot}} \left(r^2 \frac{d\theta}{dt}\right)^2}{m(Gm)^2}} \quad (41.73)$$

Il va sans dire que dans notre cas, il ne s'agit plus d'un potentiel gravitationnel mais électrique. Ce qui nous amène à écrire pour notre problème:

$$\mu = \frac{kZe^2}{m} \quad (41.74)$$

Encore nous reste-t-il à trouver l'expression de K sous forme quantifiée (selon les postulats de Bohr).

Attaquons-nous d'abord à déterminer l'expression du paramètre focal p de la trajectoire:

Dans notre problème actuel, l'énergie cinétique et potentielle exprimées en coordonnées polaires donnent (cf. chapitre de Calcul Vectoriel):

$$E_{\text{cin}} = \frac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2) \quad \text{et} \quad E_{\text{pot}} = -k\frac{Ze^2}{r} \quad (41.75)$$

L'énergie totale de l'atome est donc donnée par:

$$E_{\text{tot}} = \frac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2) - k\frac{Ze^2}{r} \quad (41.76)$$

De façon identique à celle de Bohr, Sommerfeld et Wilson appliquèrent la même forme de quantification pour le rayon-vecteur et l'étendirent à la quantification pour l'angle azimutal.

Soit les moments cinétiques:

$$b_r = \oint p_r dr = n_r h \quad \text{et} \quad b_\theta = \oint p_\theta d\theta = n_\theta h \quad (41.77)$$

Les quantités de mouvement s'obtiennent par dérivation du lagrangien par rapport aux coordonnées généralisées puisque (cf. chapitre de Mécanique Analytique):

$$\begin{aligned} p_r &= \frac{\partial L}{\partial \dot{r}} = \frac{\partial}{\partial \dot{r}} \left[\frac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2) + k\frac{Ze^2}{r} \right] = m\dot{r} \\ p_\theta &= \frac{\partial L}{\partial \dot{\theta}} = \frac{\partial}{\partial \dot{\theta}} \left[\frac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2) + k\frac{Ze^2}{r} \right] = mr^2\dot{\theta} \end{aligned} \quad (41.78)$$

La quantification sur l'angle est immédiate, puisque p_θ est une constante du mouvement. Effectivement, le lagrangien L étant indépendant de θ (mais pas de $\dot{\theta}$), l'invariance du moment cinétique se traduit par l'équation de Lagrange:

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{\theta}} - 0 = 0 \Leftrightarrow \dot{p}_\theta = 0 \Leftrightarrow p_\theta = c^{\text{ste}} \quad (41.79)$$

Ce qui nous donne:

$$\oint p_\theta d\theta = p_\theta \oint d\theta = p_\theta 2\pi = n_\theta h \Rightarrow p_\theta = n_\theta \hbar \quad (41.80)$$

avec $n_\theta \in \mathbb{N}^*$ étant le "**nombre quantique azimutal**", pour rappeler qu'il est lié à la quantification de l'angle polaire.

De cette dernière relation nous obtenons aussi:

$$p_\theta = n_\theta \hbar = mr^2 \dot{\theta} \Leftrightarrow \frac{d\theta}{dt} = \frac{n_\theta \hbar}{mr^2} \quad (41.81)$$

Revenons maintenant à:

$$p = \frac{K^2}{\mu} = \frac{\left(r^2 \frac{d\theta}{dt}\right)^2}{\frac{kZe^2}{m}} \quad (41.82)$$

ce qui nous donne:

$$p = \frac{\left(\frac{n_\theta \hbar}{m}\right)^2}{\frac{kZe^2}{m}} = n_\theta^2 \frac{\hbar^2}{mkZe^2} \quad (41.83)$$

Attaquons-nous maintenant à déterminer l'excentricité e de la trajectoire (à ne pas confondre avec la notation de la charge électrique si possible!).

Ce qui nous donne:

$$e = \sqrt{1 + \frac{2}{\mu^2} EK^2} = \sqrt{1 + \frac{2}{\left(\frac{k^2 Z^2 e^4}{m^2}\right)} \frac{E_{tot} n_\theta^2 \hbar^2}{m m^2}} = \sqrt{1 + \frac{2E_{tot} n_\theta^2 \hbar^2}{mk^2 Z^2 e^4}} \quad (41.84)$$

Pour déterminer la quantification du moment cinétique par rapport à la variable radiale, nous allons nous servir d'une substitution:

$$\oint p_r dr = \oint m \dot{r} dr = n_r \hbar \quad (41.85)$$

En notant simplement r' la dérivée $dr/d\theta$, l'intégrale s'écrit:

$$\oint m \dot{r} dr = \int_0^{2\pi} m r'^2 \dot{\theta} d\theta = p_\theta \int_0^{2\pi} \frac{r'^2}{r^2} d\theta \quad (41.86)$$

où nous avons utilisé $\dot{\theta} = p_\theta / (mr^2)$ comme nous l'avons déjà démontré.

En reportant:

$$r(\theta) = \frac{p}{1 + e \cos \theta} \quad (41.87)$$

dans l'intégrale du moment cinétique radial, nous obtenons (simple à obtenir):

$$b_r = n_r \hbar = 2\pi \left(\frac{1}{\sqrt{1-e^2}} - 1 \right) p_\theta \quad (41.88)$$

d'où nous déduisons compte tenu de $p_\theta = n_\theta \hbar$ que:

$$n_r \hbar = 2\pi \left(\frac{1}{\sqrt{1-e^2}} - 1 \right) n_\theta \hbar \quad (41.89)$$

ce qui nous amène à:

$$e = \sqrt{1 - \frac{1}{\left(\frac{n_\theta + n_r}{n_\theta} \right)^2}} \quad (41.90)$$

et donc:

$$1 - \frac{1}{\left(\frac{n_\theta + n_r}{n_\theta} \right)^2} = 1 + \frac{2E_{\text{tot}} n_\theta^2 \hbar^2}{m k^2 Z^2 e^4} \quad (41.91)$$

Après quelques simplifications élémentaires nous obtenons finalement:

$$E_{\text{tot}}^n = \frac{-mZ^2 e^4}{8\varepsilon_0^2 \hbar^2 (n_\theta + n_r)^2} \quad (41.92)$$

où n_r , appelé également "**nombre quantique radial**" peut lui être nul! Car c'est le cas si $\dot{r} = 0$, c'est-à-dire si la trajectoire est un cercle (cas particulier de Bohr).

Nous introduisons alors un entier n appelé "**nombre quantique principal**" tel que:

$$n = n_r + n_\theta \quad (41.93)$$

avec $n \in \mathbb{N}^*$.

Sommerfeld et Wilson montrent par-là que les orbitales du modèle de Bohr doivent pouvoir être déterminées par ces deux nouveaux nombres quantiques:

Exemple:

Pour $n = 2$ nous avons deux sous-orbitales possibles:

$$n = 2 = n_\theta + n_\gamma = \begin{cases} 2 = 1 + 1 \\ 2 = 2 + 0 \end{cases} \quad (41.94)$$

La valeur $n_\theta = 0$ est impossible par définition car cela signifierait que le petit axe est nul (ellipse dégénérée en une droite) et l'électron ne peut traverser le noyau (dans le modèle classique en tout cas). Donc la plus petite valeur entière de n_θ possible est 1.

Il y a donc alors n orbites donnant le même terme spectral. Autrement dit, il y a n fois la même quantification d'énergie. Nous disons également que le niveau d'énergie (total) E_n est " **n fois dégénéré**".

L'idée de Sommerfeld était de rendre compte de la richesse des spectres observés. De ce point de vue, les résultats sont décevants: la quantification de tous les degrés de liberté fait bien apparaître plus d'états (il faut maintenant deux nombres quantiques pour spécifier complètement l'état, alors que le modèle de Bohr n'en considère qu'un) mais le degré supplémentaire ne fait qu'introduire une dégénérescence en énergie.

Pour résumer ce modèle, il y a donc exactement le même nombre de niveaux d'énergie et donc le même nombre de transitions d'états énergétiques possibles que celui de Bohr. Du point de vue spectral, la théorie de Sommerfeld-Wilson n'apporte rien de plus que celle de Bohr mis à part que les orbites sont elliptiques et n'explique donc pas l'étendue des spectres observés.

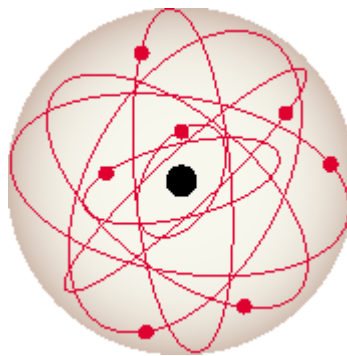


Figure: 41.8 - L'approche excentrique... de Sommerfeld

Au fait, l'idée à partir de maintenant va être de reprendre le même modèle en y ajoutant les corrections relativistes. Le travail va nécessairement être plus long mais ô combien fructueux!

6. MODÈLE RELATIVISTE DE SOMMERFELD

Cependant, le modèle de Sommerfeld et Wilson peut être considéré comme incomplet si nous ne prenons pas en compte les variations de paramètres qu'engendrent les résultats de la théorie de la relativité restreinte (cf. chapitre de Mécanique Relativiste).

Remarque: Le chapitre Physique Quantique Relativiste est réservé uniquement à l'étude de la physique quantique probabiliste basée sur la version relativiste de l'équation de Schrödinger (donc en toute rigueur le chapitre de Physique Quantique Relativiste devrait se nommer "Physique Quantique Relativiste Ondulatoire"). Raison pour laquelle il nous a semblé plus judicieux de mettre un modèle corpusculaire, déterministe et relativiste (à l'opposé d'un modèle ondulatoire, probabiliste et relativiste) comme celui que nous allons étudier maintenant dans le chapitre de Physique Quantique Corpusculaire.

Effectivement, comme nous l'avons démontré dans le développement du modèle de Bohr, l'énergie cinétique de l'électron est donnée par:

$$E_{\text{cin}}^* = \frac{1}{2}mv^2 = \frac{Z^2 e^4 m}{8\varepsilon_0^2 n^2 \hbar^2} \quad (41.95)$$

ce qui nous donne:

$$v = \frac{Ze^2}{2\varepsilon_0 n \hbar} \quad (41.96)$$

Pour l'hydrogène et le niveau $n = 1$, nous trouvons $v = 2.19 \cdot 10^6 [m \cdot s^{-1}]$ et comme facteur de Michelson-Morley (cf. chapitre de Relativité Restreinte):

$$\left(\sqrt{1-\beta^2}\right)^{-1} \cong 1.00002667 \quad (41.97)$$

avec:

$$\beta = v/c \quad (41.98)$$

Ce que le lecteur pourra vérifier avec la version française de MS Excel:

$$=1/RACINE(1-(1*(1.60217656E-19)^2/(2*8.854187E-12*1*6.62068E-34))^2/(299792458)^2)$$

Certes, la variation est faible mais les valeurs de spectrométrie étaient tellement précises qu'il fallait introduire la relativité restreinte pour prendre en compte ces infimes variations et ainsi valider la théorie par l'expérience.

Remarque: Comme nous pouvons le voir facilement, la relation montre que plus la particule est éloignée du noyau (n grand) plus sa vitesse est faible. Ce résultat a été confirmé expérimentalement en remplaçant l'électron artificiellement par un muon et les scientifiques ont ainsi remarqué que la durée de vie de ce dernier augmentait faiblement en fonction de la valeur de n .

Déterminons dans l'ordre des choses, l'expression des conditions de quantification avec les facteurs relativistes. Avant de commencer, il est important de comprendre que nous considérons le noyau comme fixe et comme référentiel de notre système. Ainsi, par rapport à ce référentiel la masse de l'électron subit une variation relativiste mais non le potentiel électrique (il faudrait prendre en compte la variation de ce dernier si et seulement si le référentiel était l'électron lui-même).

En dynamique relativiste (cf. chapitre de Relativité Restreinte), nous avons démontré que l'énergie cinétique (sous forme de notation Lagrangienne avec " T " au lieu de E_c) s'exprime sous la forme:

$$T = mc^2 - m_0 c^2 = mc^2 \left(1 - \sqrt{1-\beta^2}\right) \quad (41.99)$$

L'énergie potentielle (sous forme de notation Lagrangienne avec " V " au lieu de E_p) ne subissant pas de variation relativiste, nous avons toujours:

$$V = -k \frac{Ze^2}{r} \quad (41.100)$$

Le lagrangien est donc:

$$L = T - V = mc^2 \left(1 - \sqrt{1 - \beta^2}\right) + k \frac{Ze^2}{r} \quad (41.101)$$

En travaillant en coordonnées polaires, dans lesquelles la vitesse a pour expression:

$$v^2 = \dot{r}^2 + r^2 \dot{\theta}^2 \quad (41.102)$$

Dès lors:

$$\beta^2 = (v/c)^2 = \frac{\dot{r}^2 + r^2 \dot{\theta}^2}{c^2} \quad (41.103)$$

Les conditions de quantification de Sommerfeld étant:

$$b_r = \oint p_r dr = n_r h$$

$$b_\theta = \oint p_\theta d\theta = n_\theta h \quad (41.104)$$

À présent, nous devons rechercher des expressions relativistes pour p_r et p_θ .

Commençons par p_r :

$$p_r = \frac{\partial L}{\partial \dot{r}} = \frac{\partial}{\partial \dot{r}} \left[mc^2 \left(1 - \sqrt{1 - \beta^2}\right) + k \frac{Ze^2}{r} \right] \quad (41.105)$$

avec:

$$\beta^2 = \frac{1}{c^2} (\dot{r}^2 + r^2 \dot{\theta}^2) \quad (41.106)$$

Soit:

$$\begin{aligned} m_e c^2 \frac{\partial}{\partial \dot{r}} \left(1 - \sqrt{1 - \beta^2}\right)^{\frac{1}{2}} &= -m_e c^2 \left[\frac{1}{2} (1 - \beta^2)^{-\frac{1}{2}} \right] \cdot \frac{\partial}{\partial \dot{r}} (1 - \beta^2) \\ \frac{\partial}{\partial \dot{r}} (1 - \beta^2) &= -\frac{\partial}{\partial \dot{r}} \beta^2 = -\frac{1}{c^2} \frac{\partial}{\partial \dot{r}} v^2 = -\frac{1}{c^2} \frac{\partial}{\partial \dot{r}} (\dot{r}^2 + r^2 \dot{\theta}^2) = -\frac{2\dot{r}}{c^2} \end{aligned} \quad (41.107)$$

Ce qui donne:

$$-m_e c^2 \left[\frac{1}{2} (1 - \beta^2)^{-\frac{1}{2}} \right] \cdot \left(-\frac{2\dot{r}}{c^2} \right) = \frac{m_e c^2 2\dot{r}}{2 (1 - \beta^2)^{\frac{1}{2}} c^2} = \frac{m \dot{r}}{(1 - \beta^2)^{\frac{1}{2}}} \quad (41.108)$$

Comme:

$$\frac{\partial}{\partial \dot{r}} k \frac{Z e^2}{r} = 0 \quad (41.109)$$

nous avons finalement:

$$p_r = \frac{m \dot{r}}{(1-\beta^2)^{\frac{1}{2}}} \quad (41.110)$$

La première condition de quantification s'écrit donc:

$$b_r = \oint p_r dr = \oint \frac{m \dot{r}}{(1-\beta^2)^{\frac{1}{2}}} dr = n_r h \quad (41.111)$$

Pour p_θ :

$$p_\theta = \frac{\partial L}{\partial \dot{\theta}} = \frac{\partial}{\partial \dot{\theta}} \left[mc^2 (1 - \sqrt{1-\beta^2}) + k \frac{Z e^2}{r} \right] \quad (41.112)$$

toujours avec:

$$\beta^2 = \frac{1}{c^2} (\dot{r}^2 + r^2 \dot{\theta}^2) \quad (41.113)$$

Soit:

$$\begin{aligned} mc^2 \frac{\partial}{\partial \dot{\theta}} (1 - \sqrt{1-\beta^2}) &= -mc^2 \left[\frac{1}{2} (1-\beta^2)^{-\frac{1}{2}} \right] \frac{\partial}{\partial \dot{\theta}} (1-\beta^2) \\ \frac{\partial}{\partial \dot{\theta}} (1-\beta^2) &= -\frac{\partial}{\partial \dot{\theta}} \beta^2 = -\frac{1}{c^2} \frac{\partial}{\partial \dot{\theta}} (\dot{r}^2 + r^2 \dot{\theta}^2) = -\frac{2r^2 \dot{\theta}}{c^2} \end{aligned} \quad (41.114)$$

Ce qui donne:

$$-mc^2 \left[\frac{1}{2} (1-\beta^2)^{-\frac{1}{2}} \times -\frac{2r^2 \dot{\theta}}{c^2} \right] = \frac{mr^2 \dot{\theta}}{(1-\beta^2)^{\frac{1}{2}}} \quad (41.115)$$

Comme:

$$\frac{\partial}{\partial \dot{\theta}} k \frac{Z e^2}{r} = 0 \quad (41.116)$$

nous avons finalement:

$$p_\theta = \frac{mr^2 \dot{\theta}}{(1-\beta^2)^{\frac{1}{2}}} \quad (41.117)$$

La seconde condition de quantification s'écrit donc:

$$b_\theta = \oint p_\theta d\theta = \oint \frac{mr^2 \dot{\theta}}{(1-\beta^2)^{\frac{1}{2}}} d\theta = n_\theta h \quad (41.118)$$

En résumé, les conditions de quantification de l'atome relativiste de Sommerfeld sont:

$$b_r = \oint \frac{m\dot{r}}{(1-\beta^2)^{\frac{1}{2}}} dr = n_r h \text{ et } b_\theta = \oint \frac{mr^2 \dot{\theta}}{(1-\beta^2)^{\frac{1}{2}}} d\theta = n_\theta h \quad (41.119)$$

Nous pourrions, en voyant les deux résultats ci-dessus, conclure un peu trop rapidement en pensant qu'il aurait suffi finalement de multiplier les deux conditions de quantification par le facteur de Michelson-Morley relativement à la transformation relativiste de la masse. Or, un tel raccourci est complètement faux et tout sauf rigoureux ! Effectivement, si vous appliquez un tel raisonnement, il suffirait alors de prendre l'expression de l'énergie totale du modèle non relativiste de Sommerfeld-Wilson et d'introduire partout où la masse se situe le facteur de Michelson-Morley. Pourtant, le résultat final n'a absolument rien de commun avec le résultat que nous allons obtenir plus loin. Il faut donc toujours être prudent et travailler comme le mathématicien sans brûler les étapes !

L'énergie totale relativiste de l'atome (somme de l'énergie cinétique, de l'énergie de masse et de l'énergie potentielle du champ électrique pour l'ensemble de l'atome) est donnée par:

$$E_{tot} = \frac{mc^2}{(1-\beta^2)^{\frac{1}{2}}} - k \frac{Ze^2}{r} \quad (41.120)$$

Effectivement, en utilisant la notation du chapitre de Relativité Restreinte:

$$E_{tot} = E_c + E_m + E_{pot} = \underbrace{E_c + m_0 c^2}_{=mc^2} + \left(-k \frac{Ze^2}{r} \right) = mc^2 + \left(-k \frac{Ze^2}{r} \right) = \frac{m_0 c^2}{\sqrt{1-\beta^2}} + \left(-k \frac{Ze^2}{r} \right) \quad (41.121)$$

Or, dans le cadre de l'étude de la physique quantique corpusculaire, il est (malheureusement) d'usage de noter la masse au repos avec le symbole de la masse relativiste tel que:

$$m_0 = m \quad (41.122)$$

Donc:

$$E_{tot} = \frac{mc^2}{\sqrt{1-\beta^2}} + \left(-k \frac{Ze^2}{r} \right) \quad (41.123)$$

Il nous faut exprimer cette énergie totale en fonction des conditions de quantification. Il y a un long travail mathématique à effectuer mais indispensable pour arriver au résultat de notre étude.

Soit le calcul de l'expression:

$$\frac{p_{\theta}^2}{r^2} + p_r^2 \quad (41.124)$$

avec:

$$p_{\theta} = \frac{mr^2 \dot{\theta}}{(1-\beta^2)^{\frac{1}{2}}} \text{ et } p_r = \frac{m\dot{r}}{(1-\beta^2)^{\frac{1}{2}}} \quad (41.125)$$

En élevant au carré:

$$p_{\theta}^2 = \frac{m^2 r^2 \dot{\theta}^2}{(1-\beta^2)} \rightarrow \frac{p_{\theta}^2}{r^2} = \frac{m^2 \dot{\theta}^2}{1-\beta^2} \text{ et } p_r^2 = \frac{m^2 \dot{r}^2}{1-\beta^2} \quad (41.126)$$

Donc:

$$\frac{p_{\theta}^2}{r^2} + p_r^2 = \frac{m^2 r^2 \dot{\theta}^2}{1-\beta^2} + \frac{m^2 \dot{r}^2}{1-\beta^2} = \frac{m^2 (\dot{r}^2 + r^2 \dot{\theta}^2)}{1-\beta^2} = \frac{m^2 v^2}{1-\beta^2} \quad (41.127)$$

Nous ajoutons des deux côtés de l'égalité $m^2 c^2$ (dans l'idée d'y inclure l'énergie de masse comme vous allez le voir quelques lignes plus loin), ce qui donne:

$$\begin{aligned} m^2 c^2 + \frac{p_{\theta}^2}{r^2} + p_r^2 &= m^2 c^2 + \frac{m^2 v^2}{1-\beta^2} = m^2 c^2 + \frac{m^2 \frac{v^2}{c^2} c^2}{1-\beta^2} \\ &= \frac{m^2 c^2 (1-\beta^2) + m^2 \beta^2 c^2}{1-\beta^2} = \frac{m^2 c^2}{1-\beta^2} \end{aligned} \quad (41.128)$$

En multipliant des deux côtés par c^2 il vient:

$$c^2 \left(m^2 c^2 + \frac{p_{\theta}^2}{r^2} + p_r^2 \right) = \frac{m^2 c^4}{1-\beta^2} \quad (41.129)$$

En extrayant la racine carrée:

$$\sqrt{c^2 \left(m^2 c^2 + \frac{p_{\theta}^2}{r^2} + p_r^2 \right)} = \frac{m c^2}{\sqrt{1-\beta^2}} \quad (41.130)$$

Si nous introduisons cette dernière relation dans l'expression de l'énergie totale, nous obtenons:

$$E_{tot} = c \sqrt{m^2 c^2 + p_r^2 + \frac{p_{\theta}^2}{r^2}} - k \frac{Ze^2}{r} \quad (41.131)$$

Maintenant, il nous reste à déterminer les expressions de n_r et n_{θ} en fonction de p_r et p_{θ} .

L'intégrale de quantification de l'angle azimutal est immédiate:

$$b_{\theta} = \oint \frac{m r^2 \dot{\theta}}{(1-\beta^2)^{\frac{1}{2}}} d\theta = \int_0^{2\pi} \frac{m r^2 \dot{\theta}}{(1-\beta^2)^{\frac{1}{2}}} d\theta = \frac{m r^2 \dot{\theta}}{(1-\beta^2)^{\frac{1}{2}}} \theta \Big|_0^{2\pi} = \frac{2\pi m r^2 \dot{\theta}}{(1-\beta^2)^{\frac{1}{2}}} = n_{\theta} h \quad (41.132)$$

Soit:

$$\boxed{\frac{m r^2 \dot{\theta}}{(1-\beta^2)^{\frac{1}{2}}} = p_{\theta} = n_{\theta} h} \quad (41.133)$$

L'intégrale de quantification du rayon-vecteur nécessite un développement plus conséquent:

$$b_r = \oint \frac{m \dot{r}}{(1-\beta^2)^{\frac{1}{2}}} dr = n_r h \quad (41.134)$$

Ensuite, viennent de longs et joyeux développements mathématiques:

En reprenant l'expression de l'énergie totale:

$$E_{tot} = c \sqrt{\left(m^2 c^2 + p_r^2 + \frac{p_{\theta}^2}{r^2} \right)} - k \frac{Z e^2}{r} \quad (41.135)$$

Nous obtenons:

$$\left(E_{tot} + k \frac{Z e^2}{r} \right) = c \sqrt{\left(m^2 c^2 + p_r^2 + \frac{p_{\theta}^2}{r^2} \right)} \quad (41.136)$$

En élevant au carré et en faisant quelques transformations:

$$\begin{aligned} \left(E_{tot} + \frac{k Z e^2}{r} \right)^2 &= c^2 \left(m^2 c^2 + p_r^2 + \frac{p_{\theta}^2}{r^2} \right) \\ \frac{1}{c^2} \left(E_{tot} + \frac{k Z e^2}{r} \right)^2 &= m^2 c^2 + p_r^2 + \frac{p_{\theta}^2}{r^2} \\ p_r^2 &= \frac{1}{c^2} \left(E_{tot} + \frac{k Z e^2}{r} \right)^2 - m^2 c^2 - \frac{p_{\theta}^2}{r^2} \\ &= \frac{1}{c^2} \left(E_{tot}^2 + \frac{k^2 Z^2 e^4}{r^2} + 2 E_{tot} \frac{k Z e^2}{r} \right) - m^2 c^2 - \frac{p_{\theta}^2}{r^2} \\ &= \frac{E_{tot}^2}{c^2} + \frac{k^2 Z^2 e^4}{c^2 r^2} + 2 \frac{E_{tot}}{c^2} \frac{k Z e^2}{r} - m^2 c^2 - \frac{p_{\theta}^2}{r^2} \\ &= \left(\frac{E_{tot}^2}{c^2} - m^2 c^2 \right) + \frac{k^2 Z^2 e^4}{c^2 r^2} + 2 \frac{E_{tot}}{c^2} \frac{k Z e^2}{r} - \frac{p_{\theta}^2}{r^2} \end{aligned} \quad (41.137)$$

En travaillant sur le terme entre parenthèses, on le posera égal à A tel que:

$$A = \frac{E_{\text{tot}}^2}{c^2} - m^2 c^2 \quad (41.138)$$

En ajoutant et en retranchant $2 E_{\text{tot}} m$ et en décomposant le terme $-m^2 c^2$ en $m^2 c^2 - 2 m^2 c^2$ et ensuite en les regroupant:

$$\begin{aligned} A &= \left[\frac{E_{\text{tot}}^2}{c^2} + 2 E_{\text{tot}} m - 2 E_{\text{tot}} m \right] + \left[m^2 c^2 - 2 m^2 c^2 \right] = \\ &= \left[\frac{E_{\text{tot}}^2}{c^2} - 2 E_{\text{tot}} m + m^2 c^2 \right] + \left[2 E_{\text{tot}} m - 2 m^2 c^2 \right] = \\ &= \frac{E_{\text{tot}}^2 - 2 E_{\text{tot}} m c^2 + m^2 c^4}{c^2} + \left[2 E_{\text{tot}} m - 2 m^2 c^2 \right] = \\ &= \frac{(E_{\text{tot}} - m c^2)^2}{c^2} + 2 m (E_{\text{tot}} - m c^2) \end{aligned} \quad (41.139)$$

Nous posons en vue de la simplification des calculs (pour alléger le nombre de termes à manipuler):

$$E' = E_{\text{tot}} - m c^2 \quad (41.140)$$

Nous obtenons ainsi:

$$A = \frac{E'^2}{c^2} + 2 m E' \quad (41.141)$$

En mettant $m^2 c^2$ en évidence, nous avons:

$$A = m^2 c^2 \left(\frac{E'^2}{m^2 c^4} + 2 \frac{E'}{m c^2} \right) \quad (41.142)$$

En ajoutant et en retranchant 1 dans la parenthèse:

$$A = m^2 c^2 \left[\left(\frac{E'^2}{m^2 c^4} + 2 \frac{E'}{m c^2} + 1 \right) - 1 \right] = m^2 c^2 \left[\left(1 + \frac{E'}{m c^2} \right)^2 - 1 \right] \quad (41.143)$$

En travaillant, à présent, sur les trois derniers termes:

$$\frac{k^2 Z^2 e^4}{c^2 r^2} + 2 \frac{E_{\text{tot}}}{c^2} \frac{k Z e^2}{r} - \frac{p_e^2}{r^2} \quad (41.144)$$

Comme $E_{\text{tot}} = E' + m c^2$ nous avons:

$$2 \frac{k Z e^2 (E' + m c^2)}{r c^2} + \frac{k^2 Z^2 e^4}{c^2 r^2} - \frac{p_e^2}{r^2} = \frac{2}{r} \frac{k Z e^2 (E' + m c^2)}{c^2} + \frac{k^2 Z^2 e^4}{c^2 r^2} - \frac{p_e^2}{r^2} \quad (41.145)$$

En posant:

$$B = k \frac{Z e^2 E'}{c^2} + k m Z e^2 \quad (41.146)$$

Et en posant également:

$$C = \frac{k^2 Z^2 e^4}{c^2} - n_\theta^2 \hbar^2 \quad (41.147)$$

puisque $p_\theta = n_\theta \hbar$.

Sommerfeld introduit alors ce qu'il appelle une "**constante de structure fine**" α définie par la relation:

$$\alpha = \frac{e^2}{\hbar c} \quad (41.148)$$

valant:

$$\alpha \cong 1/137 \quad (41.149)$$

Remarque: La constante de structure fine est une des constantes les plus importantes de la physique. D'abord parce qu'elle est sans dimensions, et secundo parce qu'elle est à ce jour la mieux connue (au niveau de la précision) de toutes les constantes et tertio, parce qu'elle ne dépend que de termes qui semblent être des constantes fondamentales. Les physiciens et astrophysiciens cherchent donc à observer si la valeur de cette constante varie au cours du temps, ce qui impliquerait immédiatement qu'une au moins des constantes implicites n'est pas atemporelle.

Compte tenu de la constante de structure fine, nous écrivons:

$$C = k^2 Z^2 \left(\frac{e^4}{c^2 \hbar^2} \right) \hbar^2 - n_\theta^2 \hbar^2 = k^2 Z^2 \alpha^2 \hbar^2 - n_\theta^2 \hbar^2 = -n_\theta^2 \hbar^2 \left(1 - \frac{k^2 Z^2 \alpha^2}{n_\theta^2} \right) \quad (41.150)$$

En résumé:

$$p_r = \pm \sqrt{A + \frac{2B}{r} + \frac{C}{r^2}} \quad (41.151)$$

Avec:

$$\begin{aligned} A &= \frac{E'^2}{c^2} + 2mE' = m^2 c^2 \left[\left(1 + \frac{E'}{mc^2} \right)^2 - 1 \right] \\ B &= \frac{k Z e^2 E'}{c^2} + k m Z e^2 \\ C &= -n_\theta^2 \hbar^2 \left(1 - \frac{k^2 Z^2 \alpha^2}{n_\theta^2} \right) \end{aligned} \quad (41.152)$$

Nous aboutissons donc à l'intégrale suivante:

$$\oint \mathcal{P}_r dr = \oint \frac{m \dot{r}}{(1-\beta^2)^{\frac{1}{2}}} \dot{r} dr = \oint \sqrt{A + \frac{2B}{r} + \frac{C}{r^2}} dr = n_r h \quad (41.153)$$

Le théorème des résidus (cf. chapitre d'Analyse Complexe) appliqué à l'intégrale précédente donne pour expression:

$$\oint \sqrt{A + \frac{2B}{r} + \frac{C}{r^2}} dr = 2\pi i \sum r_i \quad (41.154)$$

Nous voyons trivialement qu'il y a un pôle à l'origine $r = 0$.

Nous allons calculer le résidu en ce point en passant à la limite pour $r = 0$. Nous posons pour cela:

$$f(r) = \sqrt{A + \frac{2B}{r} + \frac{C}{r^2}} \quad (41.155)$$

En passant à la limite construite sur la base du théorème des résidus:

$$\lim_{r \rightarrow 0} r f(r) = \lim_{r \rightarrow 0} \sqrt{A r^2 + 2B r + C} = \sqrt{C} \quad (41.156)$$

Le résidu correspondant au pôle $r = 0$ est donc:

$$r_1 = \sqrt{C} \quad (41.157)$$

Nous voyons également qu'il y a un second résidu à l'infini $r = \infty$ et pour le calculer, nous effectuons à nouveau un changement de variable. Nous posons (conformément à la méthode que nous avons vue dans le chapitre d'Analyse Complexe):

$$z = \frac{1}{r} \Rightarrow dz = -\frac{dr}{r^2} = -z^2 dr \Rightarrow dr = -\frac{dz}{z^2} \quad (41.158)$$

L'intégrale s'écrit alors:

$$\oint \sqrt{A + \frac{2B}{r} + \frac{C}{r^2}} dr = \oint \sqrt{A + 2Bz + Cz^2} \left(-\frac{dz}{z^2} \right) = \oint -\frac{\sqrt{A}}{z^2} \sqrt{1 + \frac{2B}{A}z + \frac{C}{A}z^2} dz \quad (41.159)$$

Pour trouver le résidu, nous allons faire un développement en série de Laurent de:

$$-\frac{\sqrt{A}}{z^2} \sqrt{1 + \frac{2B}{A}z + \frac{C}{A}z^2} \quad (41.160)$$

autour de ce pôle de valeur nulle. Pour ce faire, nous posons:

$$u = \frac{2B}{A}z + \frac{C}{A}z^2 \quad (41.161)$$

Nous connaissons le développement de Taylor de l'expression résultante de ce changement de variable:

$$\sqrt{1+u} = 1 + \frac{1}{2}u - \frac{1}{8}u^2 + \dots \quad (41.162)$$

Appliqué au radical, nous obtenons:

$$\sqrt{1 + \frac{2B}{A}z + \frac{C}{A}z^2} = 1 + \frac{1}{2}\left(\frac{2B}{A}z + \frac{C}{A}z^2\right) - \frac{1}{8}\left(\frac{2B}{A}z + \frac{C}{A}z^2\right)^2 + \dots \quad (41.163)$$

Il vient alors automatiquement la série de Laurent (chouette!):

$$\begin{aligned} -\frac{\sqrt{A}}{z^2} \sqrt{1 + \frac{2B}{A}z + \frac{C}{A}z^2} &= -\frac{\sqrt{A}}{z^2} - \frac{\sqrt{A}}{2z^2} \left(\frac{2B}{A}z + \frac{C}{A}z^2 \right) + \frac{\sqrt{A}}{8z^2} \left(\frac{2B}{A}z + \frac{C}{A}z^2 \right)^2 - \dots = \\ &= -\frac{\sqrt{A}}{z^2} - \underbrace{\frac{\sqrt{A}B}{A}}_{\text{résidu}} \frac{1}{z} - \frac{\sqrt{A}C}{2A} + \frac{\sqrt{A}}{8z^2} \left(\frac{4B^2z^2}{A^2} + \frac{C^2}{A}z^4 + \frac{2BC}{A^2}z^3 \right) - \dots \end{aligned} \quad (41.164)$$

où nous voyons immédiatement que le pôle est d'ordre 2.

Le second résidu est le coefficient en $\frac{1}{z}$:

$$r_2 = -\frac{B}{\sqrt{A}} \quad (41.165)$$

Effectivement, nous avons simplement appliqué la relation démontrée dans le chapitre d'Analyse Complexe:

$$\text{Res}[g(z); 0] = -\lim_{t \rightarrow 0} \frac{1}{(k-1)!} \frac{d^{k-1}}{dt^{k-1}} \left[z^k \frac{1}{z^2} f\left(\frac{1}{z}\right) \right] \quad (41.166)$$

pour déterminer le résidu se trouvant dans la série de Laurent avec l'ordre du pôle k valant donc 2.

En final, nous aboutissons à:

$$\oint \sqrt{A + \frac{2B}{r} + \frac{C}{r^2}} dr = 2\pi i \sum r_i = 2\pi i \left(\sqrt{C} - \frac{B}{\sqrt{A}} \right) = n_r \hbar \quad (41.167)$$

Avec:

$$\sqrt{C} = \sqrt{-n_\theta^2 \hbar^2 \left(1 - \frac{k^2 Z^2 \alpha^2}{n_\theta^2} \right)} = i n_\theta \hbar \sqrt{1 - \frac{k^2 Z^2 \alpha^2}{n_\theta^2}} \quad (41.168)$$

Pour le calcul de $\frac{B}{\sqrt{A}}$ nous avons:

$$\begin{aligned}
 \frac{B}{\sqrt{A}} &= \frac{\frac{kZe^2 E'}{c^2} + kmZe^2}{\sqrt{m^2 c^2 \left[\left(1 + \frac{E'}{mc^2} \right)^2 - 1 \right]}} = \frac{\frac{kZe^2 E'}{mc^3} + \frac{kZe^2}{c}}{\sqrt{\left[\left(1 + \frac{E'}{mc^2} \right)^2 - 1 \right]}} = \frac{\frac{kZe^2 E'}{mc^3} + \frac{kZe^2}{c}}{i \sqrt{1 - \left(1 + \frac{E'}{mc^2} \right)^2}} = \\
 &= -i \frac{\frac{kZe^2 E'}{mc^3} + \frac{kZe^2}{c}}{\sqrt{1 - \left(1 + \frac{E'}{mc^2} \right)^2}}
 \end{aligned} \quad (41.169)$$

Dès lors, l'intégrale curviligne a pour expression:

$$2\pi i \left(i n_\theta \hbar \sqrt{1 - \frac{k^2 Z^2 \alpha^2}{n_\theta^2}} \right) - 2\pi i \left[-i \frac{\frac{kZe^2 E'}{mc^3} + \frac{kZe^2}{c}}{\sqrt{1 - \left(1 + \frac{E'}{mc^2} \right)^2}} \right] = n_r \hbar \quad (41.170)$$

Après simplification:

$$\begin{aligned}
& -n_{\theta} h \sqrt{1 - \frac{k^2 Z^2 \alpha^2}{n_{\theta}^2}} - \frac{2\pi \left(\frac{k Z e^2 E'}{mc^3} + \frac{k Z e^2}{c} \right)}{\sqrt{1 - \left(1 + \frac{E'}{mc^2} \right)^2}} = n_r h \\
& - \frac{\frac{2\pi k Z e^2 E'}{mc^3} + \frac{2\pi k Z e^2}{c}}{\sqrt{1 - \left(1 + \frac{E'}{mc^2} \right)^2}} = n_{\theta} h \sqrt{1 - \frac{k^2 Z^2 \alpha^2}{n_{\theta}^2}} + n_r h \\
& - \frac{\frac{2\pi k Z e^2 E'}{mc^3 h} + \frac{2\pi k Z e^2}{c h}}{\sqrt{1 - \left(1 + \frac{E'}{mc^2} \right)^2}} = n_{\theta} \sqrt{1 - \frac{k^2 Z^2 \alpha^2}{n_{\theta}^2}} + n_r = n_r + \sqrt{n_{\theta}^2 - k^2 Z^2 \alpha^2} \\
& - \frac{\sqrt{1 - \left(1 + \frac{E'}{mc^2} \right)^2}}{\frac{2\pi k Z e^2 E'}{mc^3 h} + \frac{2\pi k Z e^2}{c h}} = \frac{1}{n_r + \sqrt{n_{\theta}^2 - k^2 Z^2 \alpha^2}} \\
& - \frac{\sqrt{1 - \left(1 + \frac{E'}{m_e c^2} \right)^2}}{k Z \underbrace{\frac{2\pi e^2}{c h}}_{\alpha} \left(\frac{E'}{mc^2} + 1 \right)} = \frac{1}{n_r + \sqrt{n_{\theta}^2 - k^2 Z^2 \alpha^2}} \\
& - \frac{\sqrt{1 - \left(1 + \frac{E'}{mc^2} \right)^2}}{k Z \left(\frac{E'}{mc^2} + 1 \right)} = \frac{\alpha}{n_r + \sqrt{n_{\theta}^2 - k^2 Z^2 \alpha^2}}
\end{aligned} \tag{41.171}$$

Nous élevons au carré:

$$\begin{aligned}
\frac{1 - \left(1 + \frac{E'}{m_e c^2}\right)^2}{k^2 Z^2 \left(\frac{E'}{mc^2} + 1\right)^2} &= \frac{\alpha^2}{\left(n_r + \sqrt{n_\theta^2 - k^2 Z^2 \alpha^2}\right)^2} \\
\frac{1 - \left(1 + \frac{E'}{mc^2}\right)^2}{\left(\frac{E'}{mc^2} + 1\right)^2} &= \frac{k^2 Z^2 \alpha^2}{\left(n_r + \sqrt{n_\theta^2 - k^2 Z^2 \alpha^2}\right)^2} \\
\frac{1}{\left(1 + \frac{E'}{mc^2}\right)^2} - 1 &= \frac{k^2 Z^2 \alpha^2}{\left(n_r + \sqrt{n_\theta^2 - k^2 Z^2 \alpha^2}\right)^2} \quad (41.172) \\
\frac{1}{\left(1 + \frac{E'}{mc^2}\right)^2} &= 1 + \frac{k^2 Z^2 \alpha^2}{\left(n_r + \sqrt{n_\theta^2 - k^2 Z^2 \alpha^2}\right)^2} \\
\left(1 + \frac{E'}{mc^2}\right)^2 &= \left[1 + \frac{k^2 Z^2 \alpha^2}{\left(n_r + \sqrt{n_\theta^2 - k^2 Z^2 \alpha^2}\right)^2}\right]^{-1} \\
1 + \frac{E'}{mc^2} &= \left[1 + \frac{k^2 Z^2 \alpha^2}{\left(n_r + \sqrt{n_\theta^2 - k^2 Z^2 \alpha^2}\right)^2}\right]^{-\frac{1}{2}}
\end{aligned}$$

Donc:

$$\frac{E'}{mc^2} = -1 + \left[1 + \frac{k^2 Z^2 \alpha^2}{\left(n_r + \sqrt{n_\theta^2 - k^2 Z^2 \alpha^2}\right)^2}\right]^{-\frac{1}{2}} \quad (41.173)$$

d'où:

$$E' = mc^2 \left\{ -1 + \left[1 + \frac{k^2 Z^2 \alpha^2}{\left(n_r + \sqrt{n_\theta^2 - k^2 Z^2 \alpha^2}\right)^2}\right]^{-\frac{1}{2}} \right\} = -mc^2 \left\{ 1 - \left[1 + \frac{k^2 Z^2 \alpha^2}{\left(n_r + \sqrt{n_\theta^2 - k^2 Z^2 \alpha^2}\right)^2}\right]^{-\frac{1}{2}} \right\} \quad (41.174)$$

Nous posons $n = n_r + n_\theta$:

En travaillant sur le dénominateur $\left(n_r + \sqrt{n_\theta^2 - k^2 Z^2 \alpha^2}\right)$:

$$n_r + n_\theta \sqrt{1 - \left(\frac{kZ\alpha}{n_\theta} \right)^2} \quad (41.175)$$

En ajoutant et en retranchant n_θ :

$$\begin{aligned} & (n_r + n_\theta) + n_\theta \left(\sqrt{1 - \left(\frac{kZ\alpha}{n_\theta} \right)^2} - 1 \right) \\ &= n + n_\theta \left(\sqrt{1 - \left(\frac{kZ\alpha}{n_\theta} \right)^2} - 1 \right) \end{aligned} \quad (41.176)$$

Donc:

$$E' = -mc^2 \left\{ 1 - \left[1 + \frac{k^2 Z^2 \alpha^2}{\left[n + n_\theta \left(\sqrt{1 - \left(\frac{kZ\alpha}{n_\theta} \right)^2} - 1 \right) \right]^2} \right]^{\frac{1}{2}} \right\} \quad (41.177)$$

ou encore:

$$E' = -mc^2 \left\{ 1 - \left[1 + \frac{k^2 Z^2 \left(\frac{\alpha}{n} \right)^2}{\left[1 + \frac{n_\theta}{n} \left(\sqrt{1 - \left(\frac{kZ\alpha}{n_\theta} \right)^2} - 1 \right) \right]^2} \right]^{\frac{1}{2}} \right\} \quad (41.178)$$

Ou encore:

$$E' = -mc^2 \left\{ 1 - \left\{ 1 + k^2 Z^2 \left(\frac{\alpha}{n} \right)^2 \left[1 + \frac{n_\theta}{n} \left(\sqrt{1 - \left(\frac{kZ\alpha}{n_\theta} \right)^2} - 1 \right) \right]^{-2} \right\}^{-\frac{1}{2}} \right\} \quad (41.179)$$

Nous considérons dans le terme $\left(\sqrt{1 - \left(\frac{kZ\alpha}{n_\theta} \right)^2} - 1 \right)$ le radical qui s'écrit encore:

$$\sqrt[1-\left(\frac{kZ\alpha}{n_\theta}\right)^2]} = \left\{1 + \left[-\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]\right\}^{\frac{1}{2}} \quad (41.180)$$

Soit le développement en série $(1+x)^{\frac{1}{2}} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \frac{3}{48}x^3 - \dots$ (cf. chapitre sur les Suites Et Séries) alors:

$$\sqrt[1-\left(\frac{kZ\alpha}{n_\theta}\right)^2]} = \left\{1 + \left[-\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]\right\}^{\frac{1}{2}} = 1 + \frac{1}{2}\left[-\left(\frac{kZ\alpha}{n_\theta}\right)^2\right] - \frac{1}{8}\left[-\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]^2 + \dots \quad (41.181)$$

Donc:

$$\left\{1 + \left[-\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]\right\}^{\frac{1}{2}} = 1 - \frac{1}{2}\left(\frac{kZ\alpha}{n_\theta}\right)^2 - \frac{1}{8}\left(\frac{kZ\alpha}{n_\theta}\right)^4 + \dots \quad (41.182)$$

Comme $\frac{\alpha}{n_\theta} \ll 1$, nous pouvons négliger les termes au-delà de l'ordre 2 tel que:

$$\left\{1 + \left[-\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]\right\}^{\frac{1}{2}} = 1 - \frac{1}{2}\left(\frac{kZ\alpha}{n_\theta}\right)^2 \quad (41.183)$$

Le terme suivant s'écrit alors:

$$\sqrt[1-\left(\frac{kZ\alpha}{n_\theta}\right)^2]} - 1 = 1 - \frac{1}{2}\left(\frac{kZ\alpha}{n_\theta}\right)^2 - 1 = -\frac{1}{2}\left(\frac{kZ\alpha}{n_\theta}\right)^2 \quad (41.184)$$

En travaillant maintenant sur le terme entre les crochets et en considérant uniquement le carré sans tenir compte de son signe négatif (!):

$$\left[1 + \frac{n_\theta}{n} \underbrace{\sqrt[1-\left(\frac{kZ\alpha}{n_\theta}\right)^2]}_{-\frac{1}{2}\left(\frac{kZ\alpha}{n_\theta}\right)^2} - 1\right]^{-2} = \left\{1 - \frac{1}{2} \frac{n_\theta}{n} \left(\frac{kZ\alpha}{n_\theta}\right)^2\right\}^{-2} \quad (41.185)$$

Soit le développement en série de Taylor de $(1+x)^{-2} = 1 - 2x + 3x^2 - \dots$ (cf. chapitre sur les Suites Et Séries) alors

$$\left\{ 1 + \left[-\frac{1}{2} \frac{n_\theta}{n} \left(\frac{kZ\alpha}{n_\theta} \right)^2 \right] \right\}^{-2} = 1 - 2 \left[-\frac{1}{2} \frac{n_\theta}{n} \left(\frac{kZ\alpha}{n_\theta} \right)^2 \right] + 3 \left[-\frac{1}{2} \frac{n_\theta}{n} \left(\frac{kZ\alpha}{n_\theta} \right)^2 \right]^2 - \dots \quad (41.186)$$

En négligeant les termes au-delà de l'ordre 2:

$$\left[1 + \frac{n_\theta}{n} \sqrt{1 - \left(\frac{kZ\alpha}{n_\theta} \right)^2} - 1 \right]^{-2} = 1 + \frac{n_\theta}{n} \left(\frac{kZ\alpha}{n_\theta} \right)^2 \quad (41.187)$$

Le terme entre les accolades s'écrit:

$$\left\{ 1 + \left(\frac{kZ\alpha}{n} \right)^2 \left[1 + \frac{n_\theta}{n} \left(\sqrt{1 - \left(\frac{kZ\alpha}{n_\theta} \right)^2} - 1 \right) \right] \right\}^{-\frac{1}{2}} = \left\{ 1 + \left(\frac{kZ\alpha}{n} \right)^2 \left[1 + \frac{n_\theta}{n} \left(\frac{kZ\alpha}{n_\theta} \right)^2 \right] \right\}^{-\frac{1}{2}} \quad (41.188)$$

Nous entreprenons le développement en série de Taylor du terme entre les accolades:

$$(1+x)^{-\frac{1}{2}} = 1 - \frac{1}{2}x + \frac{3}{8}x^2 - \frac{15}{48}x^3 + \dots \quad (41.189)$$

En négligeant les termes au-delà de l'ordre 2:

$$\left\{ 1 + \underbrace{\left(\frac{kZ\alpha}{n} \right)^2 \left[1 + \frac{n_\theta}{n} \left(\frac{kZ\alpha}{n_\theta} \right)^2 \right]}_x \right\}^{-\frac{1}{2}} \cong 1 - \frac{1}{2} \left\{ \left(\frac{kZ\alpha}{n} \right)^2 \left[1 + \frac{n_\theta}{n} \left(\frac{kZ\alpha}{n_\theta} \right)^2 \right] \right\} + \frac{3}{8} \left\{ \left(\frac{kZ\alpha}{n} \right)^2 \left[1 + \frac{n_\theta}{n} \left(\frac{kZ\alpha}{n_\theta} \right)^2 \right] \right\}^2 \quad (41.190)$$

En développant le carré du troisième terme, il vient:

$$\left\{ \left(\frac{kZ\alpha}{n} \right)^2 \left[1 + \frac{n_\theta}{n} \left(\frac{kZ\alpha}{n_\theta} \right)^2 \right] \right\}^2 = \left\{ \left(\frac{kZ\alpha}{n} \right)^4 \left[1 + \underbrace{\left(\frac{n_\theta}{n} \right)^2 \left(\frac{kZ\alpha}{n_\theta} \right)^4}_{\ll 1} + \underbrace{2 \frac{n_\theta}{n} \left(\frac{kZ\alpha}{n_\theta} \right)^2}_{\ll 1} \right] \right\} \quad (41.191)$$

Soit:

$$\left\{ 1 + \underbrace{\left(\frac{kZ\alpha}{n} \right)^2 \left[1 + \frac{n_\theta}{n} \left(\frac{kZ\alpha}{n_\theta} \right)^2 \right]}_x \right\}^{-\frac{1}{2}} \cong 1 - \frac{1}{2} \left(\frac{kZ\alpha}{n} \right)^2 \left[1 + \frac{n_\theta}{n} \left(\frac{kZ\alpha}{n_\theta} \right)^2 \right] + \frac{3}{8} \left[\left(\frac{kZ\alpha}{n} \right)^2 \right]^2 \quad (41.192)$$

L'énergie totale de l'atome devient:

$$\begin{aligned}
E'_{n,n_\theta} &= -mc^2 \left(1 - \left\{ 1 + k^2 Z^2 \left(\frac{\alpha}{n} \right)^2 \left[1 + \frac{n_\theta}{n} \left(\sqrt{1 - \left(\frac{kZ\alpha}{n_\theta} \right)^2} - 1 \right) \right]^{-2} \right\}^{\frac{1}{2}} \right) \\
&\cong -mc^2 \left(1 - \left\{ 1 - \frac{k^2 Z^2}{2} \left(\frac{\alpha}{n} \right)^2 \left[1 + \frac{n_\theta}{n} \left(\frac{kZ\alpha}{n_\theta} \right)^2 \right] + \frac{3}{8} \left[\left(\frac{kZ\alpha}{n} \right)^2 \right]^2 \right\} \right) \\
&= -mc^2 \left\{ \frac{1}{2} \left(\frac{kZ\alpha}{n} \right)^2 \left[1 + \frac{n_\theta}{n} \left(\frac{kZ\alpha}{n_\theta} \right)^2 \right] - \frac{3}{8} \left[\left(\frac{kZ\alpha}{n} \right)^2 \right]^2 \right\} \quad (41.193) \\
&= -\frac{mc^2 k^2 Z^2 \alpha^2}{2n^2} \left[1 + \frac{n_\theta}{n} \left(\frac{kZ\alpha}{n_\theta} \right)^2 - \frac{3}{4} \left[\left(\frac{kZ\alpha}{n} \right)^2 \right] \right] \\
&= -\frac{mc^2 k^2 Z^2 \alpha^2}{2n^2} \left[1 + \frac{n_\theta n}{n^2} \left(\frac{kZ\alpha}{n_\theta} \right)^2 - \frac{3}{4} \left[\left(\frac{kZ\alpha}{n} \right)^2 \right] \right] \\
&= -\frac{mc^2 k^2 Z^2 \alpha^2}{2n^2} \left[1 + \frac{n}{n_\theta} \left(\frac{kZ\alpha}{n} \right)^2 - \frac{3}{4} \left[\left(\frac{kZ\alpha}{n} \right)^2 \right] \right]
\end{aligned}$$

Finalement, nous obtenons pour l'expression de l'énergie:

$$E'_{n,n_\theta} \cong -\frac{mc^2 k^2 Z^2 \alpha^2}{2n^2} \left\{ 1 + \left(\frac{kZ\alpha}{n} \right)^2 \left[\frac{n}{n_\theta} - \frac{3}{4} \right] \right\} \quad (41.194)$$

Nous pouvons donner une autre expression pour l'énergie de l'atome hydrogénoïde puisque:

$$k = (4\pi\epsilon_0)^{-1} \quad \text{et} \quad \alpha = \frac{e^2}{\hbar c} = \frac{2\pi e^2}{hc} \quad (41.195)$$

L'expression de l'énergie totale de l'atome hydrogénoïde devient:

$$E'_{n,n_\theta} \cong -\frac{mc^2 \left(\frac{1}{16\pi^2 \epsilon_0^2} \right) Z^2 \left(\frac{4\pi^2 e^4}{h^2 c^2} \right)}{2n^2} \left\{ 1 + \left[\frac{\frac{1}{4\pi\epsilon_0} Z \frac{2\pi e^2}{hc}}{n} \right]^2 \left[\frac{n}{n_\theta} - \frac{3}{4} \right] \right\} \quad (41.196)$$

Soit:

$$E'_{n,n_\theta} \cong -\frac{mZ^2 e^4}{8\epsilon_0^2 h^2 n^2} \left\{ 1 + \frac{Z^2 e^4}{4\epsilon_0^2 h^2 c^2 n^2} \left[\frac{n}{n_\theta} - \frac{3}{4} \right] \right\} \quad (41.197)$$

Dans la littérature, nous trouvons d'autres expressions pour l'énergie totale qui sont plus intéressantes que les précédentes (car plus traditionnelles). Ainsi, en considérant que $n = n_\gamma + n_\theta$, il vient:

$$E'_{n_r, n_\theta} \cong -\frac{mc^2 k^2 Z^2 \alpha^2}{2n^2} \left\{ 1 + \left(\frac{kZ\alpha}{n} \right)^2 \left[\frac{n_r + n_\theta}{n_\theta} - \frac{3}{4} \right] \right\} = -\frac{mc^2 k^2 Z^2 \alpha^2}{2n^2} \left\{ 1 + \left(\frac{kZ\alpha}{n} \right)^2 \left[\frac{1}{4} + \frac{n_r}{n_\theta} \right] \right\} \quad (41.198)$$

Si nous cherchons une expression en fonction de la constante de Rydberg R_H (voir plus haut):

$$E'_{n_r, n_\theta} \cong -\frac{\frac{me^4}{8\varepsilon_0^2 \hbar^3 c}}{R_H} Z^2 \frac{\hbar c}{n^2} \left\{ 1 + \frac{Z^2 e^2}{4\varepsilon_0^2 \hbar^2 c^2 n^2} \left[\frac{n}{n_\theta} - \frac{3}{4} \right] \right\} = -\frac{\hbar c}{n^2} Z^2 R_H \left\{ 1 + \frac{Z^2 e^2}{4\varepsilon_0^2 \hbar^2 c^2 n^2} \left[\frac{n}{n_\theta} - \frac{3}{4} \right] \right\} \quad (41.199)$$

Donc l'expression de l'énergie totale relativiste de l'atome hydrogénoïde la plus condensée que nous puissions trouver dans la littérature et que nous adopterons dans le présent site est:

$$E'_{n_r, n_\theta} \cong -\frac{mc^2 k^2 Z^2 \alpha^2}{2n^2} \left\{ 1 + \left(\frac{kZ\alpha}{n} \right)^2 \left[\frac{1}{4} + \frac{n_r}{n_\theta} \right] \right\} \quad (41.200)$$

La relation ci-dessus révèle bien l'existence d'une structure fine puisque les caractéristiques n_r et n_θ de l'orbite de l'électron apparaissent séparément dans un rapport et non plus uniquement sous la forme d'une somme comme dans le premier modèle de Sommerfeld et Wilson.

Mais en toute rigueur, nous devrions du fait de l'entraînement du noyau:

$$E'_{n_r, n_\theta} \cong -\frac{\mu Z^2 e^4}{8\varepsilon_0^2 \hbar^2 n^2} \left\{ 1 + \frac{Z^2 e^2}{4\varepsilon_0^2 \hbar^2 c^2 n^2} \left[\frac{n}{n_\theta} - \frac{3}{4} \right] \right\} \quad (41.201)$$

ou:

$$E'_{n_r, n_\theta} \cong -\frac{\hbar c}{n^2} Z^2 R'_H \left\{ 1 + \frac{Z^2 e^2}{4\varepsilon_0^2 \hbar^2 c^2 n^2} \left[\frac{1}{4} + \frac{n_r}{n_\theta} \right] \right\} \quad (41.202)$$

Dans laquelle la constante de Rydberg a pour expression:

$$R'_H = \frac{\mu e^4}{8\varepsilon_0^2 \hbar^3 c} \quad (41.203)$$

Cependant comme la masse du noyau est 1840 fois plus lourde que celle de l'électron, nous pouvons admettre en première approximation que:

$$\mu = \frac{m_e m_n}{m_e + m_n} = m_e \quad (41.204)$$

6.1. MOMENT MAGNÉTIQUE DIPOLAIRE QUANTIQUE

À l'époque du développement du modèle de Sommerfeld, certains physiciens s'attachent à étudier une autre propriété de l'atome. Ils observèrent que sous l'application du champ magnétique, les raies se doubleraient. Pour expliquer cela, ils eurent l'idée géniale et extrêmement simple d'expliquer ce phénomène par le moment magnétique de l'électron.

Remarque: Nous verrons en physique quantique ondulatoire, qu'au fait, même en l'absence de champ magnétique une mesure très fine des raies montre qu'elles sont toutes doubles et ce à cause du couplage spin-orbite. Dès lors, une interprétation correcte est de dire qu'il y a doublement du dédoublement des raies sous l'application du champ magnétique.

Ainsi, soit l'expression de la norme du moment magnétique dipolaire (cf. chapitre de Magnétostatique):

$$\mu_l = S \cdot I \quad (41.205)$$

le moment magnétique est donc égal à la surface entourée par l'orbite de l'électron multipliée par le courant de l'électron (perpendiculaire au vecteur unitaire de la surface) sur sa ligne d'orbite soit:

$$\mu_l = \pi \cdot r_n^2 \cdot \frac{e}{T_n} \quad (41.206)$$

où:

$$T_n = \frac{2\pi r_n}{v_n} \quad (41.207)$$

est la période du mouvement.

Nous avons vu que la somme des moments cinétiques étant égale à:

$$b = n \cdot \hbar = m \cdot r_n \cdot v_n \quad (41.208)$$

donc le rapport moment magnétique/moment cinétique donne:

$$\frac{\mu_l}{b} = \frac{e}{2m} \Leftrightarrow \mu_l = \frac{e}{2m} \cdot b = \frac{e \cdot \hbar}{2m} \cdot n = \mu_B \cdot n \quad (41.209)$$

Le rapport $e/2m$ est appelé le "**rapport gyromagnétique orbital**" et la quantité:

$$\mu_B = 9.2740154 \cdot 10^{-24} [J \cdot T^{-1}] \quad (41.210)$$

est appelée "**magnéton de Bohr**".

Remarque: Il est important de se souvenir des quelques développements et définitions qui viennent d'être faits lorsque nous développerons l'équation de Pauli en Physique Quantique Relativiste.

Fréquemment nous notons la relation ci-dessus ainsi:

$$\mu_z = \frac{e \cdot n}{2 \cdot m} \cdot \hbar = m_z \cdot \hbar \quad (41.211)$$

où m_z est appelé "**nombre quantique magnétique**".

Sachant que le nombre quantique principal est décomposé par les nombres quantiques radiaux et azimutaux, il y a alors autant de moments magnétiques qu'il y a de géométries différentes d'orbites pour une valeur donnée du nombre quantique principal. Au fait, il y en a même le double si nous considérons que l'électron peut tourner dans le même sens ou dans le sens inverse des aiguilles d'une montre (le moment magnétique étant une grandeur vectorielle).

Maintenant, prenons les deux exemples:

$$n = 1 = n_\theta + n_r = 1 + 0 \text{ et } n = 2 = n_\theta + n_r = \begin{cases} 2 = 1 + 1 \\ 2 = 2 + 0 \end{cases} \quad (41.212)$$

pour lequel nous posons maintenant $n_r = l$, nombre que nous appelons "**nombre quantique de moment cinétique orbital**" et ayant des valeurs comprises entre: $0 \leq l \leq n - 1$.

Qu'avons-nous finalement ?

1. Lorsque $n = 1$, nous avons $l = 0$ et comme $n = 1$ n'a qu'une seule sous-couche, alors lors de l'application d'un champ magnétique nous n'avons toujours qu'une et une seule raie de visible.

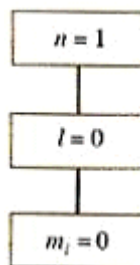


Figure: 41.9 - Décomposition d'une orbitale très basse...

2. Lorsque $n = 2$, nous avons $l = 0$ et $l = 1$ et comme $n = 2$ a deux sous-couches. Lorsque aucun champ magnétique n'est appliqué, les raies des deux sous-couches sont superposées donc indiscernables (on n'en voit qu'une seule). Mais lorsqu'un champ magnétique est appliqué les deux sous-couches se distinguent de par le moment magnétique et dès lors nous avons deux raies mais au total il en existe théoriquement 3 (une sans champ, et deux avec champ).

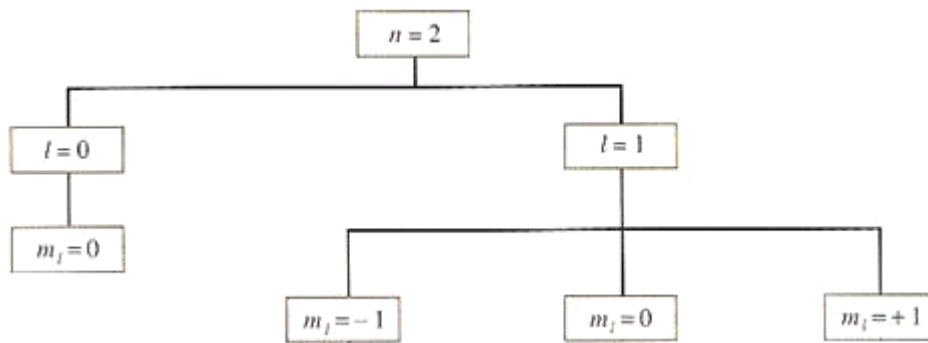


Figure: 41.10 - Décomposition de l'orbitale supérieure

Ainsi, nous avons:

$$-l \leq m_l \leq +l \quad (41.213)$$

où:

$$0 \leq l \leq (n-1) \quad (41.214)$$

L'énergie potentielle d'un moment magnétique μ_l placé dans un champ magnétique B vaut:

$$E = \mu_l \cdot B \cdot \cos \alpha = m_l \cdot \hbar \cdot B \cdot \cos \alpha \quad (41.215)$$

Donc finalement pour chaque orbitale d'électron soumise à un champ magnétique nous avons:

$$E = E_n + m_l \cdot \hbar \cdot B \cdot \cos \alpha \quad (41.216)$$

toujours avec $-l \leq m_l \leq +l$.

L'observation du spectre d'un atome dans un champ magnétique a pour effet d'ajouter des raies de par l'énergie potentielle du moment magnétique. C'est ce que nous appelons "**l'effet Zeeman**" car c'est ce dernier qui a mesuré ces raies pour la première fois (avant la théorie).

6.2. SPIN

Diverses constatations expérimentales ont conduit à attribuer à l'électron un moment cinétique et magnétique propre (dédoublage des raies Zeeman elles-mêmes !!!).

Il a effectivement été expérimentalement mesuré que le moment magnétique résultant était juste égal à la valeur du magnéton de Bohr. Il est alors tentant d'attribuer ce moment magnétique à l'électron et émettre l'hypothèse que ce dernier viendrait peut-être du fait qu'il tourne sur lui-même (moment cinétique intrinsèque): il posséderait donc un "**spin**" égal au magnéton de Bohr et ce dernier pouvant prendre des valeurs négatives ou positives. Nous parlons alors de "**nombre quantique de spin**" et ce dernier donne le nombre de différentes valeurs que peut prendre le spin.

Cependant, cette vision classique d'une rotation propre (moment cinétique intrinsèque) de la particule est en fait trop naïve et par là même erronée.

En effet, dans un premier temps, si la particule est ponctuelle, la notion de rotation propre autour de son axe est tout simplement dénuée de sens physique. Rappelons que puisque par définition, l'axe de

rotation d'un objet est le lieu de points de cet objet qui restent immobiles, alors si la particule est ponctuelle, son axe propre est sur la particule, donc celle-ci est immobile.

Dans un deuxième temps, si la particule n'est pas ponctuelle, alors la notion possède un sens, mais on se heurte dans ce cas à une autre difficulté. Supposons par exemple que la particule soit un électron, modélisé comme étant un corps sphérique de rayon a . Nous obtenons une estimation du rayon a en écrivant que l'énergie de masse de l'électron est de l'ordre de grandeur de son énergie potentielle électrostatique (cf. chapitre d'Électrostatique), soit :

$$mc^2 \cong \frac{e^2}{4\pi\epsilon_0 a} \Rightarrow a \cong \frac{e^2}{4\pi\epsilon_0 mc^2} \quad (41.217)$$

La valeur numérique de ce "**rayon classique de l'électron**" est $a = 10^{-15}$ [m] en prenant sa masse au repos.

Si nous attribuons alors à cet électron un moment cinétique égal à $\hbar/2$ (qui a les unités d'un moment cinétique), nous obtenons pour un point de l'équateur une vitesse v vérifiant :

$$mav = \frac{\hbar}{2} \Rightarrow v = \frac{\hbar}{2ma} \cong \frac{2\pi\epsilon_0 \hbar c^2}{e^2} \quad (41.218)$$

La valeur numérique de la vitesse vaut alors $v \cong 6 \cdot 10^{10}$ [m/s] ... donc la vitesse de rotation propre serait supérieure à la vitesse de la lumière dans le vide, ce qui pose bien évidemment des problèmes avec la théorie de la relativité restreinte (cf. chapitre de Relativité Restreinte).

Nous ne pouvons donc avec les outils mathématiques de la physique quantique corpusculaire formaliser rigoureusement la notion de spin, mais nous y reviendrons dans le chapitre de physique quantique relativiste (équation de Pauli) et nous montrerons que le spin est au fait quelque chose de beaucoup plus subtil qu'une simple rotation.

Mais revenons à notre vision classique en attendant. Donc, lorsque nous observons un dédoublement des raies de Zeeman, nous supposons que cela est dû au spin s de l'électron qui peut prendre deux orientations (sens vectoriels) différentes.

Il a donc été mesuré que le moment magnétique propre de l'électron est égal à la valeur du magnéton de Bohr soit:

$$\mu_i = \mu_B = \frac{e}{2m} \cdot \hbar \quad (41.219)$$

Si nous posons (ce que les physiciens aiment bien faire) $e = 1, m = 1$ nous avons:

$$\mu_i = \frac{1}{2} \hbar = s \cdot \hbar \quad (41.220)$$

(ceci juste afin d'obtenir une similitude avec $\hbar = n \cdot \hbar$...)

Cette valeur est constante mais peut être négative ou positive en fonction du sens de rotation propre de l'électron relativement à l'observateur (le moment cinétique ayant une orientation vectorielle). Ainsi:

$$s = \pm \frac{1}{2} \quad (41.221)$$

Ce résultat, de la plus haute importance, nous amène aussi à la conclusion que chaque nombre quantique magnétique est dégénéré deux fois par le nombre quantique de spin! Ainsi, comme nous le verrons un peu plus loin dans des exemples concrets (avec schémas à l'appui), chaque nombre quantique principal n est dit "dégénéré" un nombre $2n^2$ de fois:

6.3. PRINCIPE D'EXCLUSION DE PAULI

Suite au fait que l'état d'un électron atomique peut être caractérisé avec au moins les 4 nombres quantiques suivants dont nous avons démontré la provenance:

$$n, l, m_l, m_s \quad (41.222)$$

ou sous forme étendue suivante:

$$n, n_\theta, n_\varphi, l, m_l, s \quad (41.223)$$

Wolfgang Pauli, a alors posé pour expliquer certaines régularités dans les propriétés atomiques un principe d'exclusion nommé aujourd'hui "**principe d'exclusion de Pauli**" et qui s'énonce de la manière suivante: Dans un atome, deux électrons ne peuvent avoir le même quadruplet n, l, m_l, m_s ordonné de nombres quantiques.

Remarques:

R1. Nous notons parfois selon les situations $m_s = s$ (pour ce que cela change...).

R2. Nous savons par la physique quantique ondulatoire que le principe d'exclusion s'applique aux particules qui sont des "**fermions**". Ce sont les particules (élémentaires ou composées) qui ont un spin demi-entier, comme le proton, le neutron et le neutrino. Ce principe ne s'applique pas au groupe de particules dites "**bosons**", qui ont un spin nul ou entier.

Il est possible à partir de ce principe, d'établir une sorte de catalogue des éléments atomiques à partir des possibilités de remplissage des orbitales, supposées disposées en couches, améliorant ainsi la classification de Mendeleïev.

Les étudiants les voient fréquemment pour la première fois dans les écoles lors de leurs cours de chimie. Ils les utilisent la plupart du temps, sans savoir ce qu'ils représentent vraiment.

7. COUCHES ÉLECTRONIQUES

Au cours des années 1920, Bohr, Stoner et d'autres conçurent un modèle de la structure électronique des atomes qui permet de comprendre le tableau périodique des éléments. Le travail de Moseley a permis de déterminer le nombre de protons dans le noyau et, comme l'atome est neutre, c'est aussi le nombre des électrons orbitaux. Il n'est pas simple de déterminer la structure atomique et dans cette analyse, les physiciens ont été aidés par les expériences menées par les chimistes.

Ainsi, selon les chimistes les électrons occupent des couches et des sous-couches autour du noyau par ordre d'énergie croissante selon des règles associées à leurs nombres quantiques que nous avons déterminés précédemment. Ainsi, la "**configuration électronique**" est l'arrangement des électrons dans un atome, une molécule ou un autre corps. Précisément, c'est la position des électrons dans une orbitale atomique, moléculaire ou d'autres formes d'orbitales électroniques.

Remarque: Rigoureusement ce concept de "couche" comme nous pouvons nous l'imaginer visuellement n'a aucun sens si l'on se réfère aux résultats de la mécanique quantique (cf. chapitre de Physique Quantique Ondulatoire). C'est la raison pour laquelle le débat qui consiste à savoir comment remplir les couches est stérile car il n'existe pas rigoureusement sans une grossière approximation de règle générale.

Chaque couche correspond à une valeur spécifique du "**nombre quantique principal**" n et traditionnellement les couches sont désignées (cette tradition devrait être abandonnée... mais comme toutes les traditions elle a la peau dure...) par les lettres majuscules $K, L, M, N, O, \dots$ correspondant aux nombres 1, 2, 3, 4, 5... que peut prendre le nombre quantique principal.

Le "**nombre quantique secondaire/azimutal**" noté conventionnellement par la lettre l correspond aux états de dégénérescence que peuvent prendre les couches pour une valeur donnée de n tel que:

Pour la couche K ($n = 1$) nous avons une unique sous-couche comme:

$$[n_\theta = 1, n_\gamma = 0] \equiv s \quad (41.224)$$

et ainsi pour la couche L ($n = 2$) nous avons deux sous-couches:

$$[n_\theta = 1, n_\gamma = 1] \equiv p, [n_\theta = 2, n_\gamma = 0] \equiv s \quad (41.225)$$

et ainsi pour M ($n = 3$) avons trois sous-couches:

$$[n_\theta = 1, n_\gamma = 2] \equiv d, [n_\theta = 2, n_\gamma = 1] \equiv p, [n_\theta = 3, n_\gamma = 0] \equiv s \quad (41.226)$$

et ainsi de suite...

Les chimistes ont pour habitude de noter les premières sous-couches par les lettres latines:

s (sharp), p (principal), d (diffuse), f (fondamental)....

Qui sont l'équivalent alphabétique du nombre quantique secondaire l .

Le "**nombre quantique magnétique**" m_l désigne donc la position de l'orbitale dans l'espace (au fait, il représente le vecteur directeur perpendiculaire à la surface de l'orbite décrite par l'électron).

Ce dernier nombre prend, nous l'avons démontré, le double de la valeur qu'il y a de sous-couches par couche (effet Zeeman: doublement des raies) et puisque:

$$-l \leq m_l \leq +l \quad (41.227)$$

Le "**nombre quantique de spin**" m_s désigne le nombre de différentes valeurs que peut prendre le

moment magnétique de l'électron sur une orbitale donnée. Évidemment, il ne peut prendre que deux valeurs qui correspondent au sens de rotation propre de l'électron par rapport à l'observateur et produit un dédoublement du doublement des raies (dédoublement Zeeman):

$$s = \pm \frac{1}{2} \quad (41.228)$$

Définitions:

D1. Une "**couche électronique**" est un groupe d'états qui ont le même nombre quantique principal n .

D2. Une "**sous-couche**" est un groupe plus petit d'états qui sont caractérisés par les nombres quantiques de n et l .

D3. Une "**orbitale**" est précisée par les trois nombres quantique n, l, m_l et elle peut contenir deux électrons l'un de spin haut et l'autre de spin bas.

D4. Un "**état**" est défini par les quatre nombres quantiques n, l, m_l, m_s et contient un seul électron comme l'exige le principe d'exclusion.

Résumons sous forme de schémas à ce que nous avons vu jusqu'à maintenant:

Prenons l'exemple:

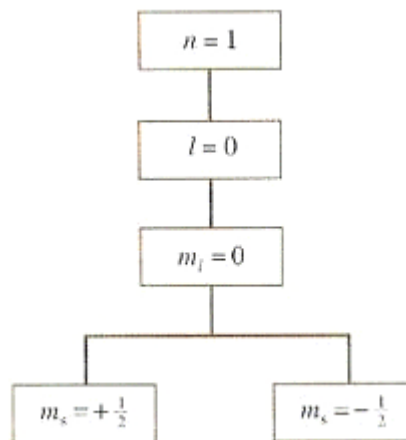


Figure: 41.11 - Décomposition d'une orbitale basse avec spin

Ainsi, le principe d'exclusion de Pauli permet qu'il y ait deux électrons sur la couche K.

Prenons l'exemple:

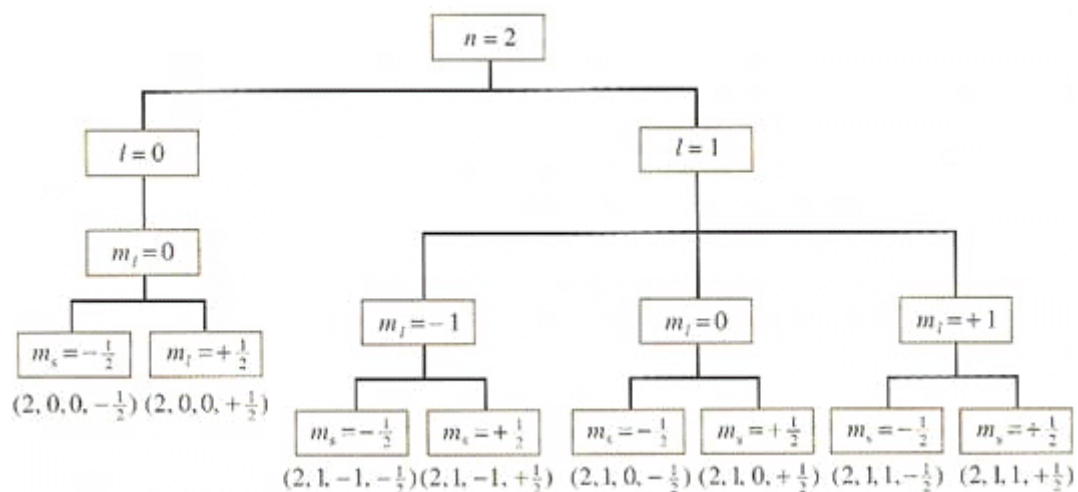


Figure: 41.12 - Décomposition d'une orbitale directement au-dessus

Ainsi, le principe d'exclusion de Pauli permet qu'il y ait 8 électrons sur la couche L .

Et ainsi de suite...

Sous forme d'atomes schématisés selon le modèle de Bohr, cela donne:

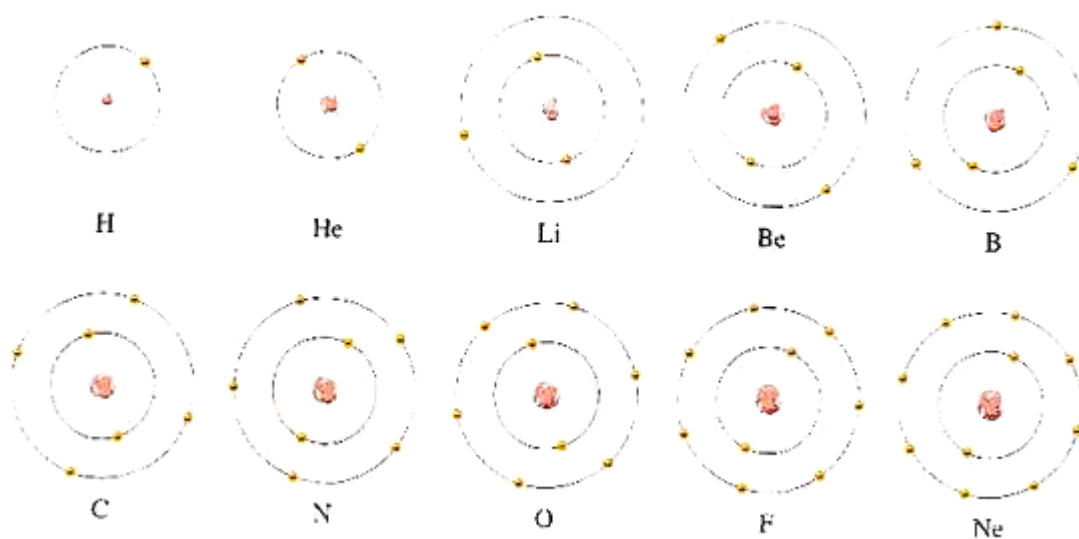


Figure: 41.13 - Forme courante dans les livres de chimie du modèle relativiste de Sommerfeld avec spin

Sous une notation conforme à celle des chimistes, les configurations fondamentales de quelques éléments s'écrivent:

Elément	Symbole	Numéro atomique	Configuration électronique
Hydrogène	H	1	$1s$
Hélium	He	2	$1s^2$
Lithium	Li	3	$1s^2.2s$
Béryllium	Be	4	$1s^2.2s^2$
Bore	B	5	$1s^2.2s^2.2p$
Carbone	C	6	$1s^2.2s^2.2p^2$
Azote	N	7	$1s^2.2s^2.2p^3$
Oxygène	O	8	$1s^2.2s^2.2p^4$
Fluor	F	9	$1s^2.2s^2.2p^5$
Néon	Ne	10	$1s^2.2s^2.2p^6$
Sodium	Na	11	$1s^2.2s^2.2p^6.3s$
Magnésium	Mg	12	$1s^2.2s^2.2p^6.3s^2$
Aluminium	Al	13	$1s^2.2s^2.2p^6.3s^2.3p$
Silicium	Si	14	$1s^2.2s^2.2p^6.3s^2.3p^2$
Phosphore	P	15	$1s^2.2s^2.2p^6.3s^2.3p^3$
Soufre	S	16	$1s^2.2s^2.2p^6.3s^2.3p^4$
Chlore	Cl	17	$1s^2.2s^2.2p^6.3s^2.3p^5$
Argon	Ar	18	$1s^2.2s^2.2p^6.3s^2.3p^6$

Figure: 41.14 -Forme tabulée de quelques éléments selon la notation des chimistes

Qui est une forme condensée du tableau équivalent suivant:

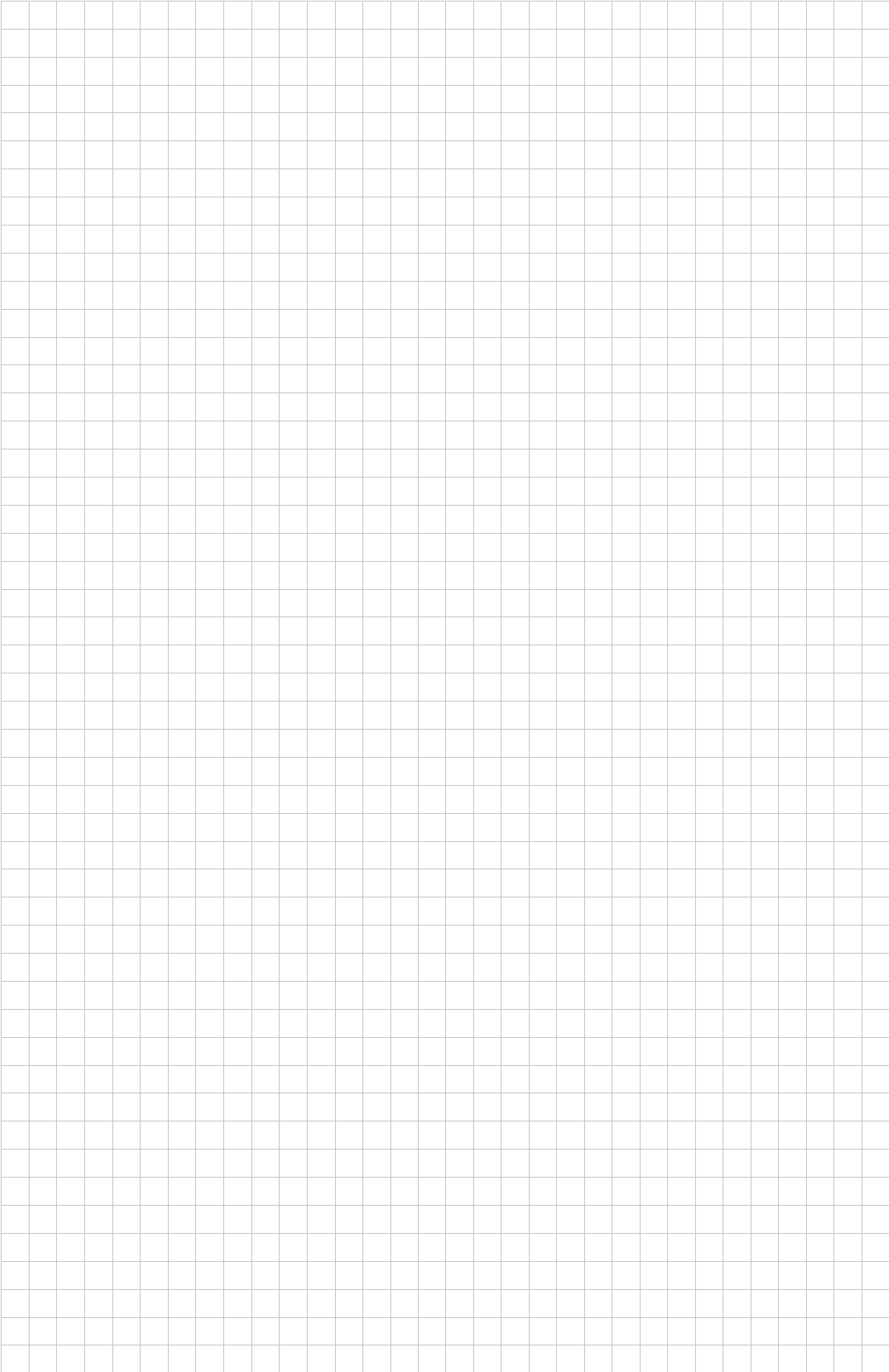
Z	Elément	n : l :	K	L		M			N				
			1	2	3	4	s	p	d	f			
			s	s	p	s					p	d	
1	Hydrogène	H	1										
2	Hélium	He	2										
3	Lithium	Li	2	1									
4	Béryllium	Be	2	2									
5	Bore	B	2	2	1								
6	Carbone	C	2	2	2								
7	Azote	N	2	2	3								
8	Oxygène	O	2	2	4								
9	Fluor	F	2	2	5								
10	Néon	Ne	2	2	6								
11	Sodium	Na	2	2	6		1						
12	Magnésium	Mg	2	2	6		2						
13	Aluminium	Al	2	2	6		2	1					
14	Silicium	Si	2	2	6		2	2					
15	Phosphore	P	2	2	6		2	3					
16	Soufre	S	2	2	6		2	4					
17	Chlore	Cl	2	2	6		2	5					
18	Argon	Ar	2	2	6		2	6					
19	Potassium	K	2	2	6		2	6	.		1		
20	Calcium	Ca	2	2	6		2	6	.		2		
21	Scandium	Sc	2	2	6		2	6	1		2		
22	Titanium	Ti	2	2	6		2	6	2		2		
23	Vanadium	V	2	2	6		2	6	3		2		
24	Chrome	Cr	2	2	6		2	6	4		1		
25	Manganèse	Mn	2	2	2		6	2	5		2		
26	Fer	Fe	2	2	6		2	6	6		2		
27	Cobalt	Co	2	2	6		2	6	7		2		
28	Nickel	Ni	2	2	6		2	6	8		2		
29	Cuivre	Cu	2	2	6		2	6	10		1		
30	Zinc	Zn	2	2	6		2	6	10		2		
31	Gallium	Ga	2	2	6		2	6	10		2	1	
32	Germanium	Ge	2	2	6		2	6	10		2	2	
33	Arsenic	As	2	2	6		2	6	10		2	3	

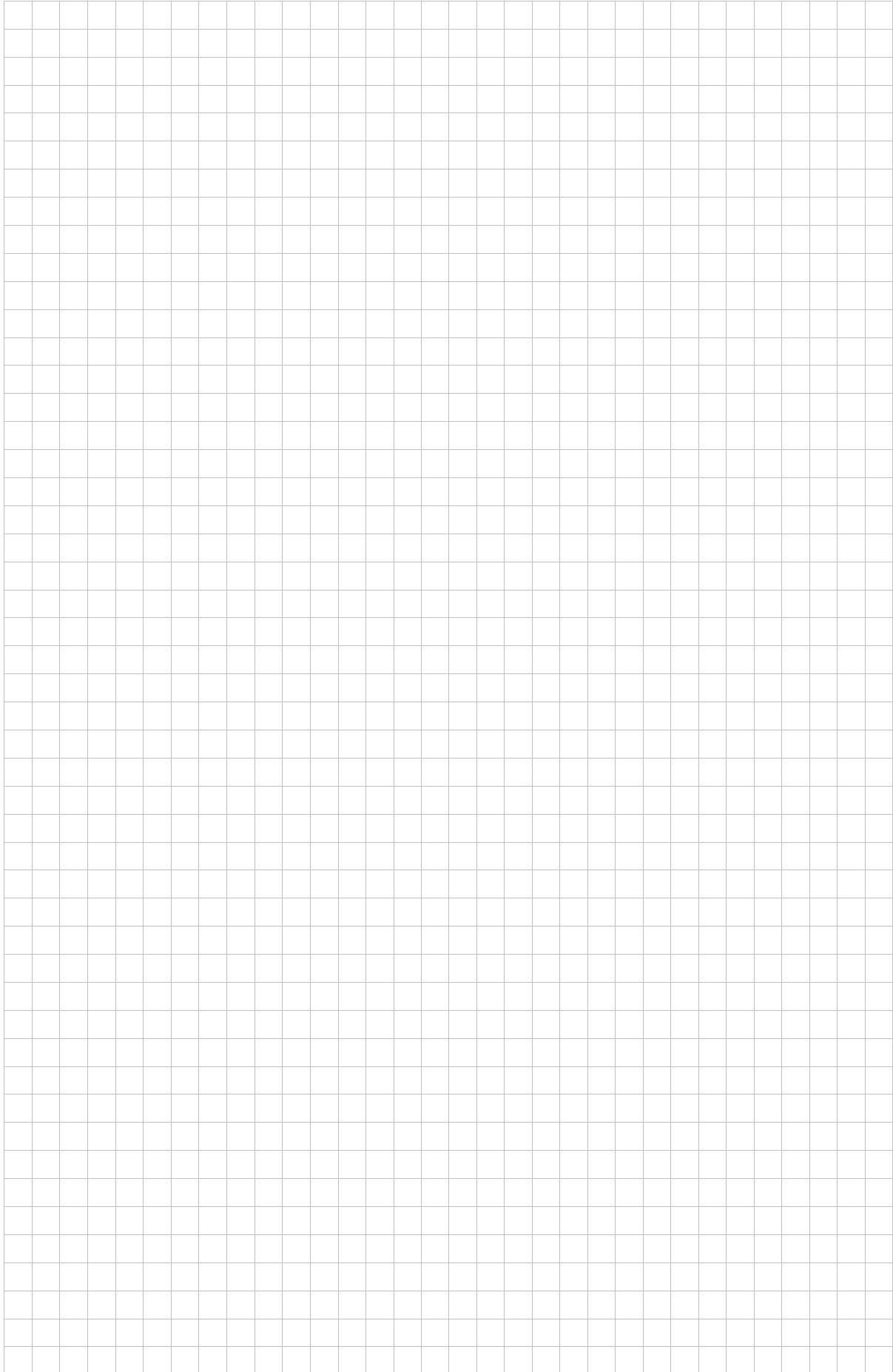
Figure: 41.15 -Forme extensive... du tableau précédent

Cependant, bien que le modèle relativiste de Sommerfeld soit d'une précision et d'une cohérence redoutable par rapport aux observations expérimentales, il n'explique pas certains phénomènes importants que nous observons à l'échelle de l'atome. Ainsi, ce modèle est dans l'incapacité d'expliquer la désintégration des éléments, le comportement dual (complémentaire) de la matière entre onde et corpuscule, l'annihilation entre matière et antimatière et encore bien d'autres.

Ce sont des développements beaucoup plus complexes et à la fois compatibles avec ce que nous avons vu qui vont être développés dans le chapitre suivant traitant de la Physique Quantique Ondulatoire permettant d'expliquer de manière parfaitement satisfaisante nombre de phénomènes qui étaient inexpliqués à l'échelle du nanomètre.

Notes personnelles:







42. PHYSIQUE QUANTIQUE ONDULATOIRE

Fille de l'ancienne théorie des quanta (cf. chapitre de Physique Quantique Corpusculaire), la physique quantique ondulatoire appelée aussi "**mécanique quantique**" constitue le pilier d'un ensemble de théories physiques que nous regroupons sous l'appellation générale de "**physique quantique**".

Cette dénomination s'oppose à celle de la physique classique, celle-ci échouant dans sa description du monde microscopique (atomes et particules) ainsi que dans celle de certaines propriétés du rayonnement électromagnétique (voir typiquement les expériences des fentes de Young dans le chapitre d'Optique Ondulatoire) ou des semi-conducteurs (voir typiquement l'Effet Hall dans le chapitre d'Électrocinétique).

Remarque: L'extension relativiste pertinente de la physique quantique est la physique quantique relativiste (voir chapitre du même nom).

La physique quantique a repris et développé l'idée de dualité onde-corpuscule introduite par De Broglie consistant à considérer toutes les particules (et même les atomes) de matière non pas seulement comme des corpuscules ponctuels, mais aussi comme des ondes, possédant une certaine étendue spatiale (cf. chapitre de Physique Quantique Corpusculaire). Ces deux aspects onde/corpuscule des particules ("**quanton**"), mutuellement exclusifs, ne peuvent être observés simultanément. Si nous observons une propriété ondulatoire, l'aspect corpusculaire disparaît et réciproquement.

À ce jour, aucune contradiction n'a pu être décelée entre les prédictions de la physique quantique et les tests expérimentaux associés. Ce succès a hélas un prix: la théorie repose sur un formalisme mathématique assez abstrait, qui rend son abord assez difficile (heureusement tant qu'on ne traite pas du spin on peut se passer du formalisme matriciel de Heisenberg pour se concentrer sur le formalisme de l'équation d'onde de Schrödinger qui est beaucoup plus simple et qui jusqu'à un certain niveau permet de se faire une représentation mentale de ce qui se passe).

Ce qui est plus difficile encore c'est qu'il est très compliqué, voire impossible, de présenter ce domaine de la physique de manière pédagogique linéaire... Ceci a pour conséquence que bon nombre d'ouvrages à son sujet (dont le présent texte ne saurait être exclu), qu'ils s'adressent à des spécialistes ou non, voient leur explications ou textes soumis à de nombreuses critiques d'interprétations, de relecture et de compléments.

Pour en sortir il est favorable de prendre pour base le "**principe d'objectivité**" dû à Werner Heisenberg qui est à la base de la "**physique quantique standard**": existe seulement ce qui est expérimentalement observable. Il faut aussi s'aligner sur la conviction de Max Born de reconstruire la physique quantique corpusculaire qui est objectivement un pot-pourri de règles quantiques et de physique classique pour céder la place à une nouvelle théorie cohérente basée uniquement sur quelques postulats (Max Born s'inspirait de l'élégance des théories d'Albert Einstein qui ne faisaient usage que de quelques postulats bien définis). Wolfgang Pauli lui aussi estimait qu'il était indispensable de cesser d'émettre des hypothèses ad hoc arbitraires chaque fois que des expériences produisaient des données en désaccord avec la théorie.

Le principe de dualité est admis par la majorité des physiciens, mais non la totalité. Un électron est-il présent à plusieurs endroits? Pour que cela soit recevable, il faut une expérience qui le trouve à plusieurs endroits, ce qui est impossible donc nous ne sommes pas tenu de répondre à la question! Dire qu'il est à plusieurs endroits avant que nous l'observions n'est pas recevable en physique: principe d'objectivité. D'une manière générale, nous allons donc aussi renoncer à la notion de trajectoire et de mouvement, ce qui va permettre, de lever la contradiction du freinage par rayonnement (cf. chapitre d'Électrodynamique): car il n'y a plus de mouvement au sens classique. Les notions de vitesse et d'accélération perdent tout sens à cette échelle!

Une minorité de physiciens nie ce principe et ont fondé une physique quantique non-standard avec des grandeurs classiques ce qui explique que l'on puisse trouver surtout dans les revues de vulgarisation des exposés qui s'écartent de la physique quantique standard (celle de la majorité des physiciens). Cette version non-standard donne les mêmes prévisions pour toute expérience réalisable, c'est donc un modèle possible.

En conclusion la physique quantique est une théorie considérée par la majorité actuelle des physiciens comme inachevée et dans laquelle beaucoup de points demeurent encore assez obscurs.

Avant de nous attaquer à la partie mathématique, nous tenons à souligner que nous allons nous limiter uniquement aux développements théoriques effectués entre 1910 et environ 1935 (au-delà la complexité des théories nécessite trop de pages pour un site Internet généraliste comme Sciences.ch).

1. POSTULATS

Contrairement à la majorité des ouvrages sur le sujet, nous sommes pédagogiquement (et non pas techniquement!) très peu convaincus quant à l'impact de la présentation des postulats de la physique quantique au début de son étude dans les classes. Nous nous permettons d'exposer nos raisons (expérience faite):

1. Ils peuvent se déduire de raisonnements mathématiques simples et logiques (algèbre élémentaire et probabilités) fondés sur les postulats de la physique quantique corpusculaire et du principe de complémentarité et découlent donc d'une évolution de cette dernière. Même si rigoureusement la démarche est fautive au moins, elle est pédagogique!
2. Ces postulats sont indigestes, voire incompréhensibles si la physique quantique (son formalisme et son vocabulaire) n'a pas été d'abord appréhendée par un certain nombre d'exercices ou d'un usage régulier.

Nous pouvons alors considérer que les seuls éléments non démontrables théoriquement (à notre connaissance) qui auraient leur place au rang de postulat seraient: le principe de complémentarité de De Broglie (nous en parlerons plus tard), la relation de Planck-Einstein (déjà vue dans un chapitre précédent) et la mesure d'une observable.

Néanmoins..., dans l'objectif de respecter la tradition, et surtout de respecter la méthodologie scientifique, nous avons choisi de quand même présenter ces postulats en début de ce chapitre mais sans trop insister dessus. Nous conseillons cependant vivement au lecteur non averti, de lire ceux-ci sans trop chercher à les comprendre mais simplement de penser à y revenir régulièrement pendant la lecture du chapitre. Dès lors, tout deviendra probablement plus limpide et la lumière sera...

Remarques: Nous verrons des cas pratiques, dans ce chapitre même, de la théorie quantique pour un usage ultérieur en physique quantique des champs et physique nucléaire. Nous conseillons cependant au lecteur de lire en même temps les chapitres d'Informatique Quantique, de Chimie Quantique et de Chimie Moléculaire qui semblerait-il aident plus que grandement à la compréhension de certains passages un peu trop théoriques présentés ici.

1.1. 1ER POSTULAT: ÉTAT QUANTIQUE

L'état d'un système quantique classique est spécifié par les coordonnées généralisées (cf. chapitre de Mécanique Analytique) $(q_1, q_2, \dots)$ et est complètement décrit par une fonction différentiable et finie partout notée en toute généralité:

$$\Psi(q_1, q_2, \dots, t) \quad (42.1)$$

dite "**fonction d'état**" ou "**fonction d'onde**", dont le module au carré (multiplication de la fonction par sa conjuguée) doit donner la densité de probabilité de trouver instantanément le système dans la configuration $(q_1, q_2, \dots)$ au temps t (si le système est dépendant du temps):

$$P(q_1, q_2, \dots, t) = |\Psi(q_1, q_2, \dots, t)|^2 dq_1 dq_2 \dots dq_n \quad (42.2)$$

ce que nous justifierons plus loin!

Remarques:

R1. Le fait que nous parlions "d'onde" au lieu de "particule" vient du postulat génial et ma foi assez logique de De Broglie que nous appelons "postulat de complémentarité" (que nous détaillerons plus loin aussi) et qui associe à toute particule de matière, une onde et réciproquement.

R2. Le fait que nous traitons des probabilités et que celles-ci soient proportionnelles au carré du module de la fonction d'onde vient des principes d'incertitudes de Heisenberg que nous démontrerons plus loin et principalement de l'expérience des fentes de Young avec des électrons (cf. chapitre d'Optique Ondulatoire).

En corollaire, la particule étant nécessairement située quelque part dans l'espace entier, nous avons la condition de normalisation que l'intégrale sur tout l'espace vaut:

$$\int |\Psi|^2 dV = 1 \quad (42.3)$$

à un facteur de phase près. En d'autres termes Ψ doit être normée, ce que nous appelons traditionnellement la "**condition de normalisation de De Broglie**" (bien qu'a posteriori le concept provienne de Max Born).

Remarques:

R1. Notons que même normée, Ψ est déterminée à un facteur de phase près. De plus, il est préférable que Ψ soit différentiable, car des opérateurs différentiels agissent sur elle pour obtenir des prévisions théoriques sur des propriétés mesurables, et finie pour qu'elle soit normalisable...

R2. Lorsque l'intégrale donnée plus haut permet d'obtenir une quantité finie, nous disons qu'elle est de "**carré sommable**". Dans le cas contraire, il faut la normaliser pour que le modèle théorique corresponde à la réalité! Nous y reviendrons aussi plus en détails (avec démonstrations!).

Rapelons qu'un "**facteur de phase**" est un facteur complexe constant de module unitaire. Nous pouvons l'écrire (selon ce que nous avons étudié dans le chapitre des Nombres lors de notre étude des nombres complexes) $e^{i\delta}$, où δ est un angle quelconque, appelé la "**phase**" (cf. chapitre de Mécanique Ondulatoire). Nous démontrerons aussi plus loin en toute rigueur pourquoi celui-ci n'a aucune influence.

Nous pouvons exprimer ce postulat de manière un peu plus formelle, car comme nous le verrons dans plusieurs exemples, la fonction d'onde est souvent un polynôme complexe qui peut dès lors s'exprimer dans l'espace de Hilbert des polynômes. Cela donne dès lors dans le langage du formalisme bra-ket de Dirac (voir plus loin les détails) la définition suivante:

Le vecteur d'état "ket" représenté par $|\Psi(t)\rangle$ appartenant à l'espace vectoriel $\mathcal{H}$ (espace de Hilbert) définit l'état du système quantique à l'instant t . Ce vecteur d'état possède toutes les propriétés mathématiques requises par la physique quantique et en particulier le produit scalaire du vecteur $|\Psi(t)\rangle$ par son vecteur dual (conjugué complexe) "bra" $\langle\Psi(t)|$ doit satisfaire le produit scalaire fonctionnel:

$$\langle\Psi(t)|\Psi(t)\rangle = 1 \quad (42.4)$$

Remarque: La notation bra-ket a été introduite par Paul Dirac pour faciliter (du moins c'est censé...) l'écriture des équations de la physique quantique, mais aussi pour souligner l'aspect vectoriel possible de l'objet représentant un état quantique. Ce qui est donc spécifique à la physique quantique est que les vecteurs ne sont pas dessinés avec des flèches mais avec des ket et des bras (cela vaut mieux qu'un pied....), mais cela n'est qu'une question de notation et n'apporte aucune nouveauté mathématique. Par ailleurs, il ne faut pas imaginer que nous écrivions explicitement dans les calculs ces vecteurs sous forme de colonnes (pensez aux nombres complexes.... il est rare que nous les écrivions sous forme vectorielle)!

Pour résumer ces derniers paragraphes, les deux relations:

$$\langle\Psi(t)|\Psi(t)\rangle = 1 \quad (42.5)$$

et:

$$\int |\Psi|^2 dV = 1 \quad (42.6)$$

sont donc équivalentes!

1.2. 2ÈME POSTULAT: ÉVOLUTION TEMPORELLE D'UN ÉTAT QUANTIQUE

Si le système n'est pas perturbé, l'évolution (supposée non relativiste!) de son état est gouvernée par l'équation d'évolution de Schrödinger (dépendante du temps donc):

$$i\hbar \frac{\partial \Psi}{\partial t} = H \Psi \quad (42.7)$$

Cette relation signifie simplement que c'est l'opérateur "énergie totale" ou "hamiltonien" H du système, qui est responsable de l'évolution du système dans le temps. En effet, la forme de l'équation montre qu'en appliquant l'hamiltonien à la fonction d'onde Ψ du système, nous obtenons sa dérivée par rapport au temps c'est-à-dire comment elle varie dans le temps.

Remarque: Nous démontrerons plus loin cette relation (ce ne sera pas trivial malheureusement mais c'est possible et donc cela élimine le besoin de la définir en tant que postulat).

Dans cette dernière relation, H est l'opérateur, l'hamiltonien (énergie totale) du système que nous démontrerons comme valant dans un cas particulier et simple:

$$H = E_{cin} + E_{pot} = -\frac{\hbar^2}{2m} \cdot \vec{\nabla}^2 + \vec{V}(q_1, q_2, \dots, t) = \sum_{k=1}^n \left(-\frac{\hbar^2}{2m_k} \cdot \frac{\partial}{\partial q_k^2} \right) + \vec{V}(q_1, q_2, \dots, t) \quad (42.8)$$

Dans le cas où le potentiel $V(q_1, q_2, \dots)$ est indépendant du temps (correspondant à un système conservatif en mécanique classique), il existe (nous le verrons dans des exemples) un ensemble de solutions particulières indépendantes du temps et satisfaisant (relation dont nous démontrerons la provenance):

$$H \psi_k(q_1, q_2, \dots) = E_k \psi_k(q_1, q_2, \dots) \quad (42.9)$$

où $\psi_k(q_1, q_2, \dots)$ est appelée une "**fonction propre**" (en analogie avec les vecteurs propres vus dans le chapitre d'Algèbre Linéaire) de l'hamiltonien/opérateur H avec valeur propre/observable E_k .

Ces solutions particulières décrivent alors des états spéciaux appelés "**états stationnaires**" (puisque indépendants du temps...), dont nous démontrerons plus tard les propriétés et l'origine du nom, et qui forment une base orthogonale.

L'équation aux valeurs propres précédente est souvent appelée "**équation de Schrödinger indépendante du temps**". Elle définit les états stationnaires et n'a un sens bien évidemment que si le système est conservatif.

C'est surtout l'équation de Schrödinger indépendante du temps qui concerne la chimie quantique et la chimie moléculaire (sujets que nous traitons dans la section de Chimie du site en détails). Nous cherchons en effet à obtenir les fonctions d'onde décrivant les états stationnaires, et surtout l'état de la plus basse énergie, "**l'état fondamental**", des atomes et des molécules. Les transitions observées en spectroscopie s'effectuant entre ces états stationnaires (nous le démontrerons plus loin), leur détermination est donc un prérequis pour l'étude de la spectroscopie. Cependant, il faut bien se rappeler que c'est l'équation d'évolution de Schrödinger, qui est (dans un premier temps...) l'équation fondamentale de la physique quantique ondulatoire non relativiste: elle joue le même rôle que l'équation de Newton en mécanique classique, soit celui d'une équation du mouvement (voir la démonstration du théorème d'Ehrenfest plus bas).

Remarque: Au fait, nous verrons (cf. chapitre de Physique Quantique Relativiste) que l'équation d'évolution de Schrödinger n'est qu'un cas particulier de ce que nous appelons "l'équation de Klein-Gordon libre" qui elle-même est un cas particulier de l'équation de "Klein-Gordon généralisée", elle-même étant un modèle limité par rapport à "l'équation de Dirac linéarisée" ... bref on n'a pas fini...

1.3. 3ÈME POSTULAT: OBSERVABLES ET OPÉRATEURS

À chaque propriété physique mesurable (une observable) d'un système notée par exemple:

$$\mathcal{O}(q_k, p_k) \quad (42.10)$$

où q_k sont les coordonnées généralisées et p_k les moments généralisés conformément aux notations adoptées dans le chapitre de Mécanique Analytique, correspond un opérateur linéaire (donc cela peut être aussi une matrice!), appelé "**opérateur hermitique**", noté fréquemment avec un chapeau tel que pour l'exemple choisi celui-ci sera noté:

$$\hat{\mathcal{O}}(q_k, p_k) \quad (42.11)$$

qui intervient toujours dans le calcul théorique d'une propriété physiquement mesurable (voir le chapitre d'Algèbre Linéaire pour avoir un rappel de ce qu'est une matrice hermitique).

Autrement dit (mais ceci est plus en relation avec le postula qui va suivre): Une valeur observable est représentée par un opérateur quantique qui opère sur la fonction d'onde pour prédire la valeur de la mesure de l'observable.

Pour faire simple..., un opérateur hermitique en physique quantique est une expression mathématique telle que si on prend son conjugué complexe (ou sa matrice adjointe si l'expression mathématique est une matrice) alors le calcul théorique de la valeur mesurable est toujours donné par la même expression.

Exemples:

Voici les plus connus dont nous démontrerons l'origine dans le présent chapitre et celui de Physique Quantique Relativiste:

E1. Coordonnées:

$$q_k \rightarrow \hat{q}_k \quad (42.12)$$

dont nous verrons un exemple pratique avec le théorème d'Ehrenfest dans le présent chapitre.

E2. Quantité de mouvement:

$$p_k \rightarrow \hat{p}_k \equiv -i\hbar\partial_k \quad (42.13)$$

dont nous verrons aussi plusieurs exemples pratiques (dont un avec le théorème d'Ehrenfest).

Ces deux derniers exemples sont parfois référés sous le nom de "**principe de correspondance**".

E3. Moment cinétique:

$$\hat{L}_z \rightarrow \hat{L}_z \equiv x\hat{p}_y - \hat{p}_x y \quad (42.14)$$

E4. Les matrices de Pauli (qui correspondent comme nous le verrons bien plus loin aux opérateurs de spin):

$$\sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \sigma_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (42.15)$$

E4. L'opérateur d'énergie d'évolution d'un état quantique:

$$E \rightarrow \hat{E} \equiv i\hbar\partial_t \quad (42.16)$$

Remarques:

R1. Cela peut sembler tomber du ciel..., mais nous verrons que cela vient tout seul lorsque nous ferons les développements plus loin de quelques exemples bien concrets ou lors de la lecture du chapitre d'Informatique Quantique.

R2. Dans le cadre de ce site, nous notons indifféremment, les opérateurs et les observables sans circonflexes (c'est au lecteur de savoir sur quoi nous travaillons sans se mélanger les pinceaux...).

Nous verrons par ailleurs que certains opérateurs ne sont pas commutatifs et qu'ils obéissent à ce que nous appelons des "**relations d'anticommutation**" (qui sont à l'origine des principes d'incertitudes de Heisenberg).

 **Exemple** (que nous démontrerons plus loin!):

$$[q_k, p_l] = [q_k, p_l]_- = q_k p_l - p_l q_k = i\hbar\delta_{kl} \quad (42.17)$$

nous disons alors que les composantes des deux observables position et quantité de mouvement sont "**conjuguées canoniquement**".

Nous verrons par ailleurs à l'aide d'un cas pratique que deux observables A, B dont les opérateurs respectifs commutent tel que:

$$[\hat{A}, \hat{B}] = 0 \quad (42.18)$$

possèdent une base de vecteurs propres commune. Nous disons alors qu'ils sont simultanément mesurables (la détermination précise de l'une n'empêche pas celle de l'autre) avec précision (dans le cas contraire nous avons une incertitude... de Heisenberg). Les deux grandeurs A, B peuvent alors être appelées "**observables compatibles**" (O.C). Ceci signifie alors que les deux opérateurs doivent avoir les mêmes fonctions propres (donc ils ne sont compatibles que si les opérateurs associés admettent des fonctions propres communes).

L'ensemble des O.C. attachées à un système physique constituent un "**ensemble complet d'observables compatibles**" (ECOC).

1.4. 4ÈME POSTULAT: MESURE D'UNE PROPRIÉTÉ

La conséquence du postulat précédent est que la mesure de $\hat{O}$ donne donc toujours une valeur propre de l'opérateur hermitique associé, $\hat{O}$. En d'autres termes, les seules valeurs observables de la propriété $\hat{O}$ sont les valeurs propres (notées par exemple: o) de l'opérateur $\hat{O}$!

$$\begin{array}{ccccccc} \hat{O} & & \varphi_k & = & o_k & & \varphi_k \\ \text{opérateur} & & k\text{-ème fonction propre} & & k\text{-ème valeur propre} & & k\text{-ème fonction propre} \end{array} \quad (42.19)$$

Les vecteurs propres et les valeurs propres d'un opérateur ont une signification spéciale: les valeurs propres sont les valeurs pouvant résulter d'une mesure idéale de cette propriété, les vecteurs propres étant les états quantiques du système lors de cette mesure.

C'est à cause de ce postulat qu'il est important de s'assurer que toute propriété physique soit représentée par un opérateur hermitique. En d'autres termes, l'hermiticité de $\hat{O}$ assure que ses valeurs propres (donc notées par exemple: o) sont réelles.

Faisons la démonstration d'abord avec les notations algébriques habituelles et ensuite avec une autre approche en utilisant la notation de Dirac.

Démonstration:

Puisqu'un opérateur est associé à une fonction propre et une valeur propre par la relation (cf. chapitre d'Algèbre Linéaire):

$$\hat{O}\varphi_k = o_k \varphi_k \quad (42.20)$$

En utilisant la propriété démontrée lors de notre étude des nombres complexes dans le chapitre Nombres comme quoi le conjugué complexe du produit de deux nombres complexes est égal au produit des deux nombres conjugués, nous avons:

$$\overline{\hat{O}\varphi_k} = \overline{o_k \varphi_k} \quad (42.21)$$

Si nous multiplions la dernière relation par le conjugué complexe de φ_k et intégrons sur tout l'espace, nous avons:

$$\int \overline{\varphi_k} \hat{O} \varphi_k dV = o_k \int \overline{\varphi_k} \varphi_k dV \quad (42.22)$$

et de même avec la relation antérieure:

$$\int \varphi_k \hat{O} \bar{\varphi}_k dV = \bar{o}_k \int \varphi_k \bar{\varphi}_k dV \quad (42.23)$$

Comme par définition $\hat{O}$ est hermitique (donc égal à son propre conjugué complexe), les deux côtés gauche des deux relations précédentes sont égaux. Nous avons alors:

$$o_k \int \bar{\varphi}_k \varphi_k dV = \bar{o}_k \int \varphi_k \bar{\varphi}_k dV \quad (42.24)$$

Il vient alors:

$$(o_k - \bar{o}_k) \int \bar{\varphi}_k \varphi_k dV = 0 \quad (42.25)$$

et comme l'intégrale est physiquement de carré sommable et donc non nulle, cela impose que:

$$o_k - \bar{o}_k = 0 \quad (42.26)$$

et donc:

$$o_k = \bar{o}_k \quad (42.27)$$

et cela n'est possible que si la valeur propre est réelle.

Maintenant, avec la notation de Dirac et avec une approche un tout petit peu différente:

Puisqu'un opérateur est associé à une fonction propre et une valeur propre par la relation (cf. chapitre d'Algèbre Linéaire):

$$\hat{O} |\varphi_k\rangle = o_k |\varphi_k\rangle \quad (42.28)$$

Nous avons alors si la fonction propre est normée en notation de Dirac:

$$o_k = \langle \varphi_k | \hat{O} | \varphi_k \rangle \quad (42.29)$$

Et si l'opérateur $\hat{O}$ est bien un opérateur hermitique (donc égal à son conjugué complexe s'il s'agit d'une fonction et égal à la transposée complexe s'il s'agit d'une matrice), nous avons:

$$\bar{o}_k = \overline{\langle \varphi_k | \hat{O} | \varphi_k \rangle} = \langle \varphi_k | \hat{O}^\dagger | \varphi_k \rangle = \langle \varphi_k | \hat{O} | \varphi_k \rangle = o_k \quad (42.30)$$

Donc si l'opérateur est bien hermitique nous avons:

$$\bar{o}_k = o_k \quad (42.31)$$

et ceci ne peut être satisfait que si les valeurs propres sont réelles.

Ainsi, les valeurs propres d'opérateur hermitiques sont toujours des nombres réels (heureusement...).

1.5. 5ÈME POSTULAT: MOYENNE D'UNE PROPRIÉTÉ

Ce postulat est le moins intuitif et le plus difficile à démontrer (nous le démontrerons par l'exemple lors de l'étude du théorème d'Ehrenfest). Son énoncé est le suivant:

La valeur moyenne (espérance) d'une propriété physique $\mathcal{O}$, quand le système se trouve dans l'état décrit par la fonction normalisée Ψ est donnée par (attention à ne pas confondre la notation de la moyenne d'un opérateur avec celle du conjugué complexe d'une valeur propre!!!):

$$\langle \mathcal{O} \rangle = \bar{\mathcal{O}} = \int \bar{\Psi} \hat{\mathcal{O}} \Psi dV = \langle \Psi | \hat{\mathcal{O}} | \Psi \rangle \quad (42.32)$$

Une expression équivalente et que je trouve compliquée est la suivante: la probabilité de trouver la valeur propre o_k (de l'opérateur $\hat{\mathcal{O}}$ majoritairement hermitique), lors d'une mesure de la propriété $\mathcal{O}$ effectuée au temps t sur un système quantique préparé dans l'état décrit par la fonction Ψ , est donnée par le carré du module de la projection de la fonction Ψ sur la fonction propre φ_k associée à la valeur propre o_k (et son opérateur):

$$P(o_k, t | \Psi) = |\langle \varphi_k | \Psi \rangle|^2 \quad (42.33)$$

où la "projection" (ou "représentative") est définie par:

$$\langle \varphi_k | \Psi \rangle = \int \bar{\varphi}_k \Psi dV \quad (42.34)$$

l'indice k étant ici pour indiquer qu'il peut y avoir pour certains opérateurs plusieurs valeurs et vecteurs propres.

Remarque: Nous reviendrons sur ce formalisme et ces relations plus tard. Cependant, plusieurs exemples pratiques sont proposés dans le présent chapitre et celui d'Informatique Quantique ainsi qu'un seul et bel exemple à la fin du chapitre de Chimie Quantique (pour le rayon moyen et le moment cinétique moyen).

Par exemple, à une dimension et pour un système dépendant du temps, nous aurons en utilisant les opérateurs présentés dans le postulat précédent (mais sur lesques nous reviendrons en détails):

$$\begin{aligned} \langle x \rangle &= \int \bar{\Psi} \hat{x} \Psi dx = \int \bar{\Psi} x \Psi dx = \int x \bar{\Psi} \Psi dx \\ \langle p \rangle &= \int \bar{\Psi} \hat{p} \Psi dx = \int \bar{\Psi} (-i\hbar \partial_x) \Psi dx = -i\hbar \int \bar{\Psi} \frac{\partial \Psi}{\partial x} dx \\ \langle E \rangle &= \int \bar{\Psi} \hat{E} \Psi dx = \int \bar{\Psi} (i\hbar \partial_t) \Psi dx = i\hbar \int \bar{\Psi} \frac{\partial \Psi}{\partial t} dx \end{aligned} \quad (42.35)$$

Dont nous verrons des cas pratiques dans les chapitres déjà mentionnés!

2. PRINCIPES D'INCERTITUDES CLASSIQUES

Avant de s'attaquer directement à la physique quantique et à ses outils mathématiques (et pseudo-démonstrations des cinq postulats), nous devons d'abord introduire un exemple classique simple dans lequel apparaît un type particulier de phénomènes: la présence intrinsèque de l'incertitude dans toute mesure.

Cette étude sous forme classique et fondamentalement pas très rigoureuse, nous aidera à mieux appréhender l'incertitude quantique (nous l'espérons) que nous étudierons et déterminerons plus tard et qui elle n'est pas d'origine expérimentale! Cependant il s'agit vraiment d'une des approches utilisées par Heisenberg lui-même!

Imaginons que nous souhaiterions mesurer au moyen d'un microscope l'abscisse x d'une particule et les composantes de sa quantité de mouvement $\vec{p}$. Pour que la mesure de x soit possible, il faut qu'un faisceau de lumière monochromatique (pour simplifier) parallèle à O_x vienne éclairer la particule, et donc qu'au moins un photon vienne frapper la particule et parvienne à l'oeil de l'observateur:

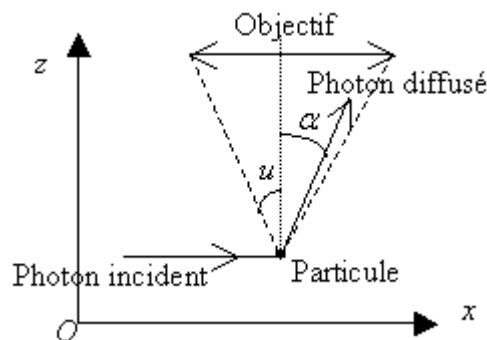


Figure: 42.1 - Microscope de Heisenberg

Une fois x mesurée, nous pouvons imaginer n'importe quel procédé pour mesurer la quantité de mouvement.

Un certain nombre de physiciens appellent cette configuration qui permet d'introduire de manière pédagogique les incertitudes le "**microscope de Heisenberg**".

Posons α comme étant l'angle que fait la direction du photon après le choc, avec O_x . Supposons pour alléger les calculs que la particule ait une masse assez élevée pour que nous puissions négliger le changement d'énergie du photon. Nous voyons qu'après le choc, les composantes de la quantité de mouvement du photon diffusé selon O_x et O_z sont:

$$\frac{h\nu}{c} \sin \alpha, \frac{h\nu}{c} \cos \alpha \quad (42.36)$$

Effectivement, rappelons que les relations entre les ondes électromagnétiques, l'équivalence masse-énergie et la quantité de mouvement (cf. chapitre de Relativité Restreinte) sont les suivantes:

$$E = mc^2 = h\nu \Rightarrow m = \frac{h\nu}{c^2} \Rightarrow p = mv = mc = \frac{h\nu}{c^2} c = \frac{h\nu}{c} \quad (42.37)$$

Il s'ensuit que la particule, elle verra sa quantité de mouvement altérée. Les composantes de sa variation sont alors (ne pas oublier qu'initialement elle était nulle en z) celles de la variation du photon selon:

$$\frac{h\nu}{c}(1 - \sin \alpha), -\frac{h\nu}{c} \cos \alpha \quad (42.38)$$

entre sa quantité de mouvement initiale et finale.

La seule information que nous possédons sur l'angle α , c'est que ce dernier est, en valeur absolue, inférieur ou égal à l'angle d'ouverture u de l'objectif du microscope (restriction technique).

Donc cela implique que:

$$|\sin \alpha| \leq \sin u \quad \cos u \leq |\cos \alpha| \leq \sin u \quad (42.39)$$

2.1. PREMIÈRE RELATION D'INCERTITUDE CLASSIQUE

Quand nous aurons mesuré la quantité de mouvement $\vec{p}$ à la fin de l'expérience, nous avons vu qu'il faudra donc effectuer les corrections:

$$\frac{h\nu}{c}(1 - \sin \alpha), -\frac{h\nu}{c} \cos \alpha \quad (42.40)$$

de la quantité de mouvement du photon pour savoir la vraie valeur de $\vec{p}$ de la particule juste avant le début de la mesure.

Dans ces corrections, il y a une partie inconnue qui correspond à des erreurs de mesure sur p_x et p_z du photon. Il est possible d'établir que l'erreur maximale de Δp_x et Δp_z sur la quantité de mouvement initiale est donnée par la composante x de la "**première relation d'incertitude d'impulsion classique**":

$$\boxed{\Delta p_x = \frac{h\nu}{c} \sin u} \quad (42.41)$$

puisque nous avons $|\sin \alpha| < \sin u$. Il s'agit donc de l'erreur au pire en quelque sorte!

2.2. DEUXIÈME RELATION D'INCERTITUDE CLASSIQUE

Voyons maintenant ce qu'il en est de la mesure de la position de la particule.

Rappelons maintenant que (cf. chapitre d'Optique Ondulatoire) pour une fente rectangulaire nous avons en posant $n = 1$:

$$\sin \theta = \frac{\lambda}{e} \quad (42.42)$$

où θ (en optique ondulatoire) est l'angle permettant de distinguer clairement deux minima de diffraction (et donc clairement un objet émettant un rayonnement identique entre deux points). Inversement, du point de vue de la diffraction, l'ouverture e est donc donnée par:

$$e = \frac{\lambda}{\sin \theta} \quad (42.43)$$

La valeur de e peut aussi être vue comme le champ de vision (projection orthogonale de la fente sur l'axe X) de largeur $x = e$ de la particule. Dès lors:

$$x = \frac{\lambda}{\sin \theta} \quad (42.44)$$

Au même titre que l'erreur maximale sur la quantité de mouvement est donnée par la condition $|\sin \alpha| < \sin u$, nous pouvons aussi écrire $|\sin \theta| < \sin u$, ce qui nous amène à écrire que:

$$\Delta x = \frac{\lambda}{\sin u} \quad (42.45)$$

Si nous multiplions:

$$\Delta p_x = \frac{h\nu}{c} \sin u = \frac{h}{\lambda} \sin u$$

et :

$$\Delta x = \frac{\lambda}{\sin u} \quad (42.46)$$

nous obtenons la "**deuxième relation d'incertitude classique**" également appelée "**l'incertitude spatiale classique**":

$$\boxed{\Delta x \Delta p_x = h} \quad (42.47)$$

qui représente donc l'erreur maximale expérimentale d'un microscope à faible ouverture rectangulaire (que de conditions!). Beaucoup d'ouvrages de physique quantique montrent que nous retrouvons exactement la même expression dans beaucoup de situations.

Remarque: Le lecteur vérifiera sans peine que cette relation appliquée pour un objet macroscopique (de l'ordre du centimètre) dont la position serait mesurable avec une précision de l'ordre du micromètre donne une incertitude ridiculement faible sur la quantité de mouvement et donc la vitesse. Par contre, la même relation appliquée pour la masse d'une particule telle que l'électron avec une précision de mesure de la position supposée du dixième de nanomètre donnera une incertitude sur la vitesse de l'ordre 1'000 [m/s]...!!

Ainsi, si nous essayons de situer une particule avec une précision de plus en plus grande, sa quantité de mouvement atteint des valeurs extrêmes. À un certain point, la quantité de mouvement peut être si grande que l'énergie correspondante est suffisante pour produire une paire de particule-antiparticule. En d'autres termes, si nous essayons de confiner une particule dans une boîte de plus en plus petite, d'une part, nous connaissons de moins en moins sa quantité de mouvement et à partir d'un certain seuil, nous ne savons même pas combien de particules il y a dans la boîte!

Cependant (!), nous verrons lors de l'étude des commutateurs appliqués à la théorie de la physique quantique, que la vraie relation d'incertitude (dont la valeur diffère de celle ci-dessus) apparaît tout naturellement uniquement à partir de propriétés mathématiques et de la définition de la quantité de mouvement.

Plus généralement, pour une particule dans un volume à dimensions (x, y, z) , un état classique est caractérisé par les 6 quantités (x, y, z, p_x, p_y, p_z) dans l'espace de phases (espace de phases qui est donc de dimension 6) et l'état quantique occupe le "cube" de volume:

$$(\Delta x \Delta p_x)(\Delta y \Delta p_y)(\Delta z \Delta p_z) = h^3 \quad (42.48)$$

Ce qui est remarquable dans cette approche simpliste c'est que la constante de Planck émerge naturellement comme l'unité de mesure universelle minimale d'incertitude de la physique expérimentale au travers du dualisme onde-corpuscule! Heisenberg écrivit que ce résultat établit l'échec final de la causalité en physique quantique.

2.3. TROISIÈME RELATION D'INCERTITUDE CLASSIQUE

En relativité restreinte, nous avons vu que x, y, z, ct constituent les composantes d'un quadrivecteur d'espace-temps ainsi que $p_x, p_y, p_z, E/c$ celles d'un vecteur d'énergie-impulsion.

Il est donc naturel de compléter les trois relations spatiales du type $\Delta x \Delta p_x = h$ par extension:

$$\begin{aligned} x &\Rightarrow p_x \rightarrow \Delta x \Delta p_x = h \\ y &\Rightarrow p_y \rightarrow \Delta y \Delta p_y = h \\ z &\Rightarrow p_z \rightarrow \Delta z \Delta p_z = h \\ \underbrace{ct \Rightarrow \frac{E}{c} \rightarrow \Delta(ct) \Delta\left(\frac{E}{c}\right)}_{(42.49)} &= h \end{aligned}$$

Nous obtenons ainsi grossièrement la "**troisième relation d'incertitude classique**" appelée également "**incertitude temporelle classique**":

$$\boxed{\Delta t \cdot \Delta E = h} \quad (42.50)$$

Cependant (!), nous verrons lors de l'étude des commutateurs appliqués à la théorie de la physique quantique, que cette relation d'incertitude (dont la valeur diffère de celle ci-dessus) apparaît aussi tout naturellement à partir des seules propriétés mathématiques et de la définition de la quantité de mouvement.

Remarque: Nous reviendrons plus tard sur les implications de cette incertitude temporelle dont les implications sont à la base de la cosmologie quantique (et de la création de notre Univers) et de la théorie quantique des champs en particulier en ce qui concerne le potentiel de Yukawa (cf. chapitre de Physique Quantique Des Champs).

Les incertitudes classiques établies vont nous permettre de mieux comprendre les incertitudes sous leur forme quantique réelle. Pour cela, parmi d'autres, il va nous falloir faire usage de l'artillerie mathématique nécessaire. Cependant, dans un souci de clarté, nous avons souhaité présenter la physique quantique ondulatoire de la manière la plus simple et la moins formelle possible. Cette présentation peut porter le lecteur à de nombreux contresens et il doit donc rester prudent tant qu'il n'en a pas vu la démonstration rigoureuse!

3. ALGÈBRE QUANTIQUE

Sous ce terme peu courant et non officiel "d'algèbre quantique" (donc à ne pas en abuser!) nous souhaitons introduire et rappeler au lecteur des outils " mathématiques qui vont nous être très utiles pour résoudre certaines équations de la physique quantique. Il est donc de première importance de comprendre (ou d'avoir compris, en ce qui concerne les rappels) au mieux ce qui va suivre!

Remarque: Les puristes risquent de grimper aux rideaux en lisant ...

3.1. OPÉRATEURS LINÉAIRES FONCTIONNELS

Définition: Les "**opérateurs linéaires**" sont des êtres mathématiques agissant sur des fonctions ou vecteurs (cf. chapitre de Calcul Vectoriel).

Les fonctions sur lesquelles peuvent opérer ces opérateurs peuvent être des fonctions d'une seule variable x , soit $f(x)$, ou des trois coordonnées d'un point x, y, z soit $f(x, y, z)$ ou écrites encore plus brièvement $f(\mathbf{x})$.

Nous serons amenés à écrire des intégrales de ces fonctions, qui sont le plus souvent étendues à tout l'espace. Dans le cas d'une fonction des trois coordonnées spatiales d'un point, nous adopterons la notation suivante:

$$\int f(\mathbf{x}) d^3\mathbf{x} = \iiint f(x, y, z) dx dy dz \quad (42.51)$$

Ces notations, indispensables pour l'allègement des expressions que nous rencontrerons en physique quantique étant établies, nous en revenons à nos opérateurs.

Partant d'une fonction f , si nous savons lui associer une fonction g de même nature, c'est-à-dire dépendante des mêmes variables, nous pouvons dire que g est le résultat de l'action d'un opérateur α sur f et écrire cela symboliquement comme un produit simple:

$$g = \alpha \cdot f \quad (42.52)$$

Mais nous introduisons tout de suite une restriction fondamentale: seuls nous intéressent les opérateurs linéaires (comme en algèbre linéaire quoi...), c'est-à-dire par exemple tels que:

$$\alpha(\lambda_1 \cdot f_1 + \lambda_2 \cdot f_2) = \lambda_1 \cdot (\alpha \cdot f_1) + \lambda_2 \cdot (\alpha \cdot f_2) \quad (42.53)$$

quels que soient les coefficients λ_1 et λ_2 .

Une catégorie très simple d'opérateurs est constituée par les nombres (scalaires réels ou complexes). Ainsi dans la relation :

$$\chi = \lambda \cdot f \quad (42.54)$$

χ est une fonction qui dépend linéairement de f , au travers d'un opérateur linéaire que nous écrivons λ , λ étant un nombre (typiquement l'opérateur de position en physique quantique).

Il y a deux cas particuliers importants:

1. Opérateur zéro: $\lambda = 0$ où $\chi = 0 \cdot f = 0$ sera une fonction bien évidemment nulle partout...
2. Opérateur unité (ou identité): $\lambda = 1$ où $\chi = 1 \cdot f = f$ (ce qui est tout aussi simple...)

Remarque: L'opérateur "Nabla" est également un opérateur linéaire fonctionnel (nous le verrons un peu plus loin) qui en physique quantique se retrouve dans l'opérateur d'énergie.

Nous vérifions sans peine pour les opérateurs fonctionnels que ces derniers sont (voir les chapitres de Théorie des Ensembles et Algèbre Linéaire au besoin):

- commutatifs par rapport à l'addition
- associatifs par rapport à l'addition et la multiplication
- distributifs par rapport à l'addition à gauche et à droite

Jusqu'à présent, rien ne distingue l'algèbre des opérateurs de celle des nombres. Mais il y a cependant deux propriétés qu'il faut toujours avoir en tête pour ne pas commettre des erreurs quand nous faisons du calcul d'opérateurs:

1. Deux opérateurs ne commutent pas en général par rapport à la multiplication (comme en algèbre linéaire...), c'est-à-dire qu'en général soient deux opérateurs fonctionnels α et β :

$$\alpha \cdot \beta \neq \beta \cdot \alpha \quad (42.55)$$

2. Si nous rencontrons une expression telle que $\alpha_1 \cdot \beta + \beta \cdot \alpha_2$, nous n'avons donc pas le droit d'effectuer en général, la mise en facteur (il s'agit donc d'une structure particulière de groupe qui est non-commutatif)!

Exemple:

Un exemple simple et important, car utile pour la suite (très proche d'un cas pratique que nous verrons plus loin), de deux opérateurs qui ne commutent pas avec une fonction d'une seule variable est le suivant (où f est quelconque). Considérons l'opérateur d/dx agissant sur $xf(x)$:

$$\frac{d}{dx}(x \cdot f) = 1 \cdot f + x \frac{d}{dx}f \quad (42.56)$$

en simplifiant par f :

$$\frac{d}{dx}x = 1 + x \frac{d}{dx} \quad (42.57)$$

Donc nous avons ci-dessus un exemple de deux opérateurs qui ne commutent pas puisque:

$$\frac{d}{dx} x \neq x \frac{d}{dx} \quad (42.58)$$

Remarques:

R1. Si un opérateur peut commuter n'importe comment avec un autre opérateur, c'est que ce dernier est un nombre (cela rejoint le concept de mesure dont nous avons fait mention dans les postulats).

R2. Lorsqu'un état (une fonction mathématique au sens formel) est inchangé par un opérateur, l'état est alors appelé "**état propre**" ou "**vecteur propre**" du système (nous verrons des exemples pratiques plus loin). L'état est alors parfaitement mesurable et est assimilé à l'observable classique.

Exemple (d'opérateur):

Partons de l'équation de Schrödinger tridimensionnelle (que nous démontrerons plus loin) à admettre pour l'instant:

$$\frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} + \frac{\partial^2 \Psi}{\partial z^2} + \frac{8\pi^2 m}{h^2} \cdot (E_{tot} - E_{pot}) \Psi = 0 \quad (42.59)$$

ou bien écrite autrement (c'est plus esthétique...) avec le laplacien d'un champ scalaire $\nabla^2 = \Delta$ (cf. chapitre de Calcul Vectoriel):

$$\Delta \Psi + \frac{2m}{h^2} (E_{tot} - E_{pot}) \Psi = \Delta \Psi + \frac{2m}{h^2} (H - V) \Psi = 0 \quad (42.60)$$

ou encore:

$$\frac{2m}{h^2} (E_{tot} - E_{pot}) \Psi = -\Delta \Psi \quad (42.61)$$

autrement encore...:

$$\frac{2m}{h^2} (E_{tot} - E_{pot}) = -\Delta \quad (42.62)$$

Alors l'opérateur énergie totale (l'hamiltonien H en d'autres termes...) s'exprime comme:

$$E_{tot} = -\frac{h^2}{2m} \Delta + E_{pot} \quad (42.63)$$

ou en notation lagrangienne:

$$H = -\frac{h^2}{2m} \Delta + V \quad (42.64)$$

Remarque: Nous retrouvons ici naturellement la deuxième expression donnée dans le deuxième postulat mais la notation V pour l'énergie potentielle peut porter à confusion avec le potentiel électrique.

D'autre part, nous savons que:

$$E_{tot} = E_{cin} + E_{pot} = \frac{p^2}{2m} + E_{pot} \quad (42.65)$$

Les deux dernières expressions doivent être identiques. La seule possibilité pour satisfaire à ces égalités est de poser:

$$p_x = \frac{1}{i} \hbar \frac{\partial}{\partial x}, p_y = \frac{1}{i} \hbar \frac{\partial}{\partial y}, p_z = \frac{1}{i} \hbar \frac{\partial}{\partial z} \quad (42.66)$$

qui sont les "**opérateurs hermitiques de la quantité de mouvement**" en physique quantique (exprimés en coordonnées cartésiennes) et dont il faudra se rappeler tout au long de ce chapitre!

Remarque: Nous retrouvons ici naturellement un des opérateurs cités dans le troisième postulat.

Nous pouvons vérifier la légitimité de ces opérateurs en les réinjectant dans l'expression de l'énergie cinétique:

$$\begin{aligned} E_{cin} &= \frac{p^2}{2m} = \frac{p_x^2 + p_y^2 + p_z^2}{2m} \stackrel{?}{=} \frac{1}{2m} \left[\left(-i\hbar \cdot \frac{\partial}{\partial x} \right)^2 + \left(-i\hbar \cdot \frac{\partial}{\partial y} \right)^2 + \left(-i\hbar \cdot \frac{\partial}{\partial z} \right)^2 \right] \\ &\stackrel{?}{=} \frac{1}{2m} \left[+i^2 \hbar^2 \cdot \frac{\partial^2}{\partial x^2} + i^2 \hbar^2 \cdot \frac{\partial^2}{\partial y^2} + i^2 \hbar^2 \cdot \frac{\partial^2}{\partial z^2} \right] = -\frac{1}{2m} \hbar^2 \left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right] \quad (42.67) \\ &= -\frac{1}{2m} \hbar^2 \cdot \Delta \end{aligned}$$

Par ailleurs, il est aisé de vérifier que ce développement reste juste si nous prenons le conjugué complexe de l'opérateur de la quantité de mouvement.

Ainsi, l'opérateur d'énergie totale (l'hamiltonien) est, lui aussi, bien hermitique! Ce résultat est très important pour vérifier par exemple des calculs en utilisant la propriété d'orthogonalité des fonctions propres que nous verrons plus loin.

3.1.1. OPÉRATEURS ADJOINTS ET HERMITIQUES

Remarque: La lecture des lignes qui vont suivre pourrait s'avérer assez abstraite. Cependant, si vous ne comprenez pas grand-chose ce n'est pas bien grave car souvent tout devient évident pendant l'étude et les développements d'exemples concrets qui seront donnés plus loin.

Considérons les deux intégrales étendues à tout l'espace (à l'intérieur de l'intégrale il s'agit d'une multiplication de fonctions et d'opérateurs) sans chercher à comprendre leur utilité pour l'instant:

$$Q = \int \bar{f}(\alpha \cdot \chi) d\mathbf{x} \text{ et } \bar{Q} = \int \bar{\chi}(\beta \cdot f) d\mathbf{x} \quad (42.68)$$

où rappelons que la notation $\bar{z}$ est le conjugué complexe de z . Il faut savoir que dans ces deux intégrales, $\bar{\epsilon}$ et β représentent des opérateurs.

Nous constatons dans les développements de la physique quantique que ces deux intégrales sont égales (donc donnant des valeurs réelles) et qu'il y a entre les opérateurs $\bar{\epsilon}$ et β une correspondance biunivoque, nous disons que β est "l'adjoint" de $\bar{\epsilon}$ (la transposée de la conjuguée) ou qu'il est "hermitique" (les deux termes sont d'usage) et nous écrivons:

$$\beta = \alpha^\dagger \quad (42.69)$$

si les deux intégrales précédentes sont respectées.

De cette définition, nous déduisons l'identité importante suivante:

$$\int \bar{f}(\alpha \cdot \chi) d\mathbf{x} = \left[\int \bar{\chi}(\alpha^\dagger \cdot f) d\mathbf{x} \right] \quad (42.70)$$

Exemple:

Considérons:

$$\alpha = \frac{d}{dx} \quad (42.71)$$

Alors par intégration par parties:

$$\begin{aligned} \int_{-\infty}^{+\infty} \bar{f}(\alpha \chi) dx &= \int_{-\infty}^{+\infty} \bar{f} \left(\frac{d\chi}{dx} \right) dx = (\bar{f} \chi(+\infty) - \bar{f} \chi(-\infty)) - \int_{-\infty}^{+\infty} \left(\frac{d\bar{f}}{dx} \right) \chi dx \\ &= - \int_{-\infty}^{+\infty} \chi \left(\frac{d\bar{f}}{dx} \right) dx = \int_{-\infty}^{+\infty} \chi \left(-\frac{d\bar{f}}{dx} \right) dx = \int_{-\infty}^{+\infty} \chi (\alpha^\dagger \bar{f}) dx = \int_{-\infty}^{+\infty} \bar{\chi} (\alpha^\dagger f) dx = \int_{-\infty}^{+\infty} \bar{\chi} (\alpha^\dagger f) dx \end{aligned} \quad (42.72)$$

Remarque: Nous démontrerons, plus loin, la relation ci-dessus dans un exemple concret mais particulier de la physique quantique des champs (chapitre suivant) et nous y reviendrons de manière plus rigoureuse dans notre présentation du formalisme de Dirac dans le chapitre de Physique Quantique Relativiste.

L'opérateur adjoint a plusieurs propriétés, dont les seules qui vont nous intéresser dans ce chapitre sont:

P1. $(\alpha^\dagger)^\dagger = \alpha$ qu'il est inutile de démontrer, car cette relation découle de la définition même de l'opérateur adjoint.

P2. λ étant envisagé comme un nombre complexe (opérateur particulier) nous avons alors $\lambda^\dagger = \bar{\lambda}$. Ce que nous avons vérifié juste avant par l'exemple avec l'opérateur de quantité de mouvement!

Une catégorie extrêmement importante d'opérateurs est donc constituée par les "**opérateurs hermitiques self-adjoints**", ou plus simplement "**opérateurs hermitiques**" égaux par définition à leurs adjoints:

$$\alpha = \alpha^\dagger \quad (42.73)$$

puisque ce sont les seuls qui émergent dans les développements de la physique quantique ondulatoire.

Nous remarquons aussi que si nous prenons un opérateur hermitique (comme celui de la quantité de mouvement pour faire simple par exemple...) et que nous multiplions celui par le nombre imaginaire unitaire pur i alors il devient antihermitique, c'est-à-dire: non-hermitique.

Remarque: Les termes "hermitique" ou "hermitien" sont équivalents et rappelez-vous que ces opérateurs peuvent être aussi des matrices!

Un opérateur quelconque, soit ξ , peut se décomposer d'une façon unique en parties hermitique et antihermitique, c'est-à-dire que nous pouvons écrire:

$$\xi = \alpha + i\beta \quad (42.74)$$

où α, β sont donc hermitiques .

Démonstration:

Si:

$$\xi^\dagger = \alpha - i\beta \quad (42.75)$$

car il s'agit d'un simple nombre complexe, alors:

$$\alpha = \frac{1}{2}(\xi + \xi^\dagger), \beta = \frac{1}{2}(\xi - \xi^\dagger) \quad (42.76)$$

La somme de l'opérateur et de son adjoint est donc un opérateur hermitique (la somme ou la soustraction entre opérateurs hermitiques, reste donc hermitique).

En général, il est trivial que le produit de deux opérateurs hermitiques $(\alpha \cdot \beta)^\dagger$ n'est pas nécessairement un opérateur hermitique, car nous vérifions que la condition pour laquelle le produit de deux opérateurs hermitiques soit lui-même hermitique, est que les deux opérateurs "commutent" (voir ce qui suit).

3.1.2. COMMUTATEURS ET ANTICOMMUTATEURS

Définitions:

D1. Le "**commutateur**" de deux opérateurs α et β , s'écrit:

$$[\alpha, \beta] = \alpha\beta - \beta\alpha = [\alpha, \beta] \quad (42.77)$$

D2. "**L'anticommutateur**" de deux opérateurs α et β , s'écrit:

$$[\alpha, \beta]_+ = \alpha\beta + \beta\alpha \quad (42.78)$$

qui s'écrit généralement aussi sous la forme suivante:

$$\{\alpha, \beta\} = \alpha\beta + \beta\alpha \quad (42.79)$$

Remarques:

R1. Comme le commutateur est beaucoup plus fréquent dans les développements que l'anticommutateur, s'il n'y a pas de confusion possible, nous le notons donc simplement $[\alpha, \beta]$.

R2. Des exemples concrets et triviaux de ces commutateurs dans le cadre de notre étude de la physique quantique ondulatoire seront présentés dans le texte qui suit.

Citons quelques propriétés évidentes des commutateurs (celles que nous utiliserons le plus):

$$\begin{aligned} [\alpha, \beta] &= -[\beta, \alpha] \\ [\lambda_1 \alpha_1 + \lambda_2 \alpha_2, \beta] &= \lambda_1 [\alpha_1, \beta] + \lambda_2 [\alpha_2, \beta] \quad (42.80) \\ [\alpha, \lambda_1 \beta_1 + \lambda_2 \beta_2] &= \lambda_1 [\alpha, \beta_1] + \lambda_2 [\alpha, \beta_2] \end{aligned}$$

où λ_1, λ_2 sont des nombre quelconques (les démonstrations sont faites - au besoin - pendant le développement d'exemples pratiques).

Cherchons l'adjoint de $[\alpha, \beta]$:

$$[\alpha, \beta]^\dagger = (\alpha\beta - \beta\alpha)^\dagger = (\beta^\dagger \alpha^\dagger - \alpha^\dagger \beta^\dagger) \quad (42.81)$$

d'où un résultat très simple:

$$[\alpha, \beta]^\dagger = [\beta^\dagger, \alpha^\dagger] \quad (42.82)$$

ce qui pourra se vérifier aisément avec l'exemple pratique que nous ferons juste quelques lignes en-dessous.

La relation suivante est très utile dans la pratique (triviale, mais comme d'habitude au besoin nous pouvons rajouter la démonstration):

$$[\alpha_1 \alpha_2, \beta] = \alpha_1 [\alpha_2, \beta] + [\alpha_1, \beta] \alpha_2 \quad (42.83)$$

nous avons de même:

$$[\alpha, \beta_1 \beta_2] = \beta_1 [\alpha, \beta_2] + [\alpha, \beta_1] \beta_2 \quad (42.84)$$

Nous démontrerons plus loin dans un cas concret, que si deux opérateurs ne commutent pas, alors il est impossible d'avoir un état ayant une valeur précise et unique pour les deux opérateurs à la fois (en physique quantique il existe une configuration d'expérience où le premier opérateur représente la quantité de mouvement et le second la coordonnée spatiale). Ce résultat implique que les opérateurs sont souvent nommés des "**observables**".

Attardons-nous un moment sur un exemple concret des commutateurs et dont un des résultats est fondamental!

Nous avons démontré plus haut les relations:

$$p_x = \frac{1}{i}\hbar \frac{\partial}{\partial x}, p_y = \frac{1}{i}\hbar \frac{\partial}{\partial y}, p_z = \frac{1}{i}\hbar \frac{\partial}{\partial z} \quad (42.85)$$

Considérons la relation (simple différentielle mathématique habituelle):

$$\frac{\partial}{\partial x}[x\psi(x)] = x \frac{\partial}{\partial x}\psi(x) + \psi(x) \quad (42.86)$$

Si nous divisons par $\psi(x)$ des deux côtés de l'égalité et qu'ensuite nous multiplions par $\hbar/i$, nous obtenons:

$$\frac{\hbar}{i} \frac{\partial}{\partial x} x = x \frac{\hbar}{i} \frac{\partial}{\partial x} + \frac{\hbar}{i} \quad (42.87)$$

ce qui nous donne:

$$p_x x = x p_x - i\hbar \Leftrightarrow x p_x - p_x x = i\hbar \quad (42.88)$$

donc il vient que le commutateur de x et p_x est égal à $i\hbar$ et donc que les quantités ne commutent pas. Nous avons donc la "**relation d'anticommutation**" suivante:

$$\boxed{[x, p_x] = i\hbar} \quad (42.89)$$

(cycl.)

Ce que nous retrouvons souvent dans la littérature sous la forme (avec le symbole de Kronecker):

$$[x_\mu, p_\nu] = i\hbar \delta_{\mu\nu} \quad (42.90)$$

Ainsi (en nous basant sur le deuxième postulat), les deux observables x et p_x , dont les opérateurs ne commutent pas, ne possèdent pas une base de vecteurs propres commune. Ils ne sont donc pas simultanément mesurables avec précision et constituent donc une incertitude d'Heisenberg!

Remarques:

R1. L'abréviation (cycl.) signifiant que l'on peut permuter circulairement les lettres (x, y, z) et que le résultat reste le même.

R2. Bien que ce résultat puisse paraître étonnant, il n'en est pas moins extrêmement correct puisque découlant d'un raisonnement mathématique. Nous ne pouvons pas faire plus simple et rigoureux.

Considérons donc maintenant aussi la relation:

$$\frac{\partial}{\partial x} [y \psi(x)] = y \frac{\partial}{\partial x} \psi(x) \quad (42.91)$$

et en procédant de la même manière que précédemment, nous obtenons:

$$\boxed{[y, p_x] = 0} \quad (42.92)$$

(cycl.)

Les deux relations:

$$[x, p_x] = i\hbar \text{ et } [y, p_x] = 0 \quad (42.93)$$

peuvent se résumer à:

$$\boxed{[q_k, p_l] = [q_k, p_l]_- = q_k p_l - p_l q_k = i\hbar \delta_{kl}} \quad (42.94)$$

en utilisant les coordonnées et moments généralisés et sont remarquables sous plusieurs angles:

- Premièrement, parce qu'à partir de considérations purement théoriques et mathématiques nous retrouvons également en physique quantique une incertitude équivalente (mais pas égale!) à celle obtenue lors de notre étude des principes d'incertitudes de Heisenberg (qui rappelons-le avaient été obtenues à partir d'un cas pratique classique).

Effectivement, si nous prenons le module du commutateur de gauche, nous avons alors la "**relation d'incertitude spatiale de Heisenberg**":

$$\boxed{|[x, p_x]| = \hbar} \quad (42.95)$$

qui rappelons-le, peut également s'écrire sous la forme:

$$|[x, v_x]| = \frac{\hbar}{m} \quad (42.96)$$

La constante de Planck étant extrêmement petite, cela explique que cet effet est impossible à détecter à notre échelle macroscopique. Par contre, la masse des électrons étant extrêmement petite aussi, la fraction ci-dessus devient notable pour un électron et l'effet de cette incertitude est important!

Enfin, par commutation des composantes du quadrivecteur impulsion (cf. chapitre de Relativité Restreinte), nous avons la "**relation d'incertitude temporelle de Heisenberg**":

$$\boxed{|[t, E]| = \hbar} \quad (42.97)$$

Une conséquence fantastique découle de l'incertitude sur le temps et l'énergie et de la relativité. Imaginons-nous le vide le plus total (vide quantique) et supposons que nous regardions ce qui se passe en un point de l'espace donné pendant un temps très court. Alors le principe d'incertitude temporelle nous dit que l'énergie de cet état (le vide!) est très imprécise. Or la relativité dit que l'énergie, c'est aussi de la masse (et aussi un champ), donc des particules. Donc, pendant ce temps très court des particules peuvent apparaître spontanément du vide ! Nous les appelons des "**particules virtuelles**" car elles disparaissent très vite et sont engendrées par les "**fluctuations quantiques du vide**".

Cette variation est suffisamment faible pour que nous puissions la mesurer aujourd'hui avec nos instruments. Cependant, nous en observons les effets seulement dans les grands collisionneurs de particules de la planète.

- Deuxièmement, ces relations sont remarquables parce que l'incertitude est une valeur complexe. Ce qui amène à considérer que le corps des complexes est inhérent à la structure réaliste de notre environnement (espace-temps) au niveau du monde quantique. Le monde quantique est donc un monde d'incertitude complexe. Et cette probabilité ne semble pas être une conséquence de notre imprécision ou de notre ignorance mais semble bien être une propriété intrinsèque de la nature.

Remarque: Les relations et propriétés de commutation et d'anticommutation seront indispensables pour développer la théorie quantifiée du moment cinétique et du spin.

3.2. REPRÉSENTATIVES

Introduisons maintenant les notations quantiques contemporaines, que nous considérons pour l'instant comme des abréviations d'intégrales portant sur des fonctions d'ondes, nous écrirons (dans le but futur de calculer des densités de probabilités):

$$\langle \psi | \chi \rangle = \int \bar{\psi} \cdot \chi d\mathbf{x} \text{ et } \langle \psi | \alpha | \chi \rangle = \int \bar{\psi}(\alpha \cdot \chi) d\mathbf{x} \quad (42.98)$$

car il s'agit d'un produit scalaire fonctionnel complexe (cf. chapitres d'Analyse Fonctionnelle et de Calcul Vectoriel).

Avec cette notation, la relation que nous avons présentée plus haut lors de notre étude des opérateurs:

$$\int \bar{\psi} \cdot (\alpha \cdot \chi) d\mathbf{x} = \overline{\left[\int \bar{\chi} \cdot (\alpha^\dagger \cdot \psi) d\mathbf{x} \right]} \quad (42.99)$$

devient (c'est plus léger déjà... mais moins pédagogique):

$$\langle \psi | \alpha | \chi \rangle = \overline{\langle \chi | \alpha^\dagger | \psi \rangle} \quad (42.100)$$

Cela dit, l'ensemble E des fonctions ψ qui nous intéressent en physique quantique ondulatoire constitue un espace linéaire fonctionnel. Effectivement, en physique quantique, les équations différentielles que nous devons résoudre (équation de Schrödinger) pour décrire le comportement d'une particule, sont telles que la solution générale peut être très souvent décomposée en la somme des solutions particulières (nous démontrerons cela!). En mathématiques, nous disons alors que les états sont linéaires, c'est-à-dire que toute combinaison d'états est encore un état.

Ainsi, l'état d'une particule est, comme nous le démontrerons plus tard, représenté par un "état quantique" ou un "vecteur d'état" noté $|\Psi\rangle$ qui correspond aussi à une fonction mathématique la décrivant complètement.

Par exemple, si $|\Psi_1\rangle$ et $|\Psi_2\rangle$ sont deux états possibles, alors:

$$|\Psi_1\rangle + |\Psi_2\rangle \quad (42.101)$$

est également un état possible pour le système (de par la propriété des espaces linéaires fonctionnels).

Revenons maintenant à notre espace linéaire fonctionnel (ou "**espace linéaire des états**"). Le fait que l'ensemble E des fonctions ψ qui nous intéressent constitue un espace linéaire fonctionnel signifie que si $\psi \in E$, $\chi \in E$, nous avons aussi:

$$(\lambda\psi + \mu\chi) \in E \quad (42.102)$$

quels que soient les coefficients λ et μ (cf. chapitre de Calcul Vectoriel).

Si les fonctions ψ constituent un espace, il est alors naturel de chercher à les rapporter à une base orthonormée. Ainsi, une suite de fonctions (qui sont les fonctions propres) $\varphi_i \in E$ constituera une base orthonormée si nous avons (forme de relation démontrée en calcul tensoriel):

$$\langle \varphi_i | \varphi_j \rangle = \delta_{ij} \quad (42.103)$$

où nous le rappelons, δ_{ij} est le symbole de Kronecker (cf. chapitre de Calcul Tensoriel).

Définition: La base est dite "**base complète**" si bien évidemment toute fonction $\psi \in E$ peut se développer en série des fonctions propres φ_i telle que:

$$\psi = \sum_i c_i \varphi_i \quad (42.104)$$

où c_i est un nombre quelconque (c'est en partie ici qu'il faut revenir aux quatrième et cinquième postulats de la physique quantique ondulatoire).

Calculons maintenant le produit scalaire fonctionnel (cf. chapitre d'Analyse Fonctionnelle):

$$\int \bar{\varphi}_j \cdot \psi dx = \langle \varphi_j | \psi \rangle = \int \bar{\varphi}_j \cdot \sum_i c_i \varphi_i dx = \sum_i c_i \int \bar{\varphi}_j \varphi_i dx = \sum_i c_i \delta_{ji} = c_j \quad (42.105)$$

Cette dernière relation montre que nous avons identiquement (nous changeons la notation des indices):

$$c_i = \langle \varphi_i | \psi \rangle \quad (42.106)$$

Ainsi, dans une base orthonormée complète φ_i , une fonction ψ sera bien décrite par la donnée des coefficients c_i . Nous aurons souvent intérêt à les mettre sous le format de la matrice représentative de ψ dans la base φ_i :

$$\begin{bmatrix} c_1 \\ c_2 \\ . \\ . \\ c_i \end{bmatrix} \quad (42.107)$$

Considérons maintenant un opérateur α tel que:

$$\chi = \alpha \cdot \psi = \alpha \cdot \sum_i c_i \varphi_i \quad (42.108)$$

Mais nous pouvons également écrire (remarquez l'apostrophe dans la relation!):

$$\chi = \sum_i \alpha \cdot c_i \cdot \varphi_i = \sum_i c_i' \cdot \varphi_i \quad (42.109)$$

Multiplions cette dernière relation par $\bar{\varphi}_j$ et calculons le produit scalaire fonctionnel:

$$\int \bar{\varphi}_j \cdot \chi dx = \int \bar{\varphi}_j \cdot \sum_i \alpha \cdot c_i \cdot \varphi_i dx = \sum_i c_i' \int \bar{\varphi}_j \cdot \alpha \cdot \varphi_i dx = \sum_i c_i' \langle \varphi_j | \alpha | \varphi_i \rangle = \sum_i c_i' A_{ji} \quad (42.110)$$

À comparer avec (obtenu plus haut):

$$\langle \varphi_j | \psi \rangle = \sum_i c_i \delta_{ji} \quad (42.111)$$

En notant A_{ji} , la "**matrice représentative**" de α dans la base φ_i , nous pouvons d'après la relation:

$$\langle \psi | \alpha | \chi \rangle = \overline{\langle \chi | \alpha^\dagger | \psi \rangle} \quad (42.112)$$

écrire finalement:

$$\langle \varphi_j | \alpha | \varphi_i \rangle = A_{ji} = \overline{\langle \varphi_i | \alpha^\dagger | \varphi_j \rangle} = \bar{A}_{ij} \quad (42.113)$$

3.3. VALEURS ET FONCTIONS PROPRES

Soit un opérateur α (hermitique ou non). Le nombre a est dit "**valeur propre de l'opérateur**" de α , s'il existe une fonction φ non identiquement nulle telle que (pour un rappel de notions similaires voir le chapitre d'Algèbre Linéaire):

$$\alpha \cdot \varphi = a \cdot \varphi \quad (42.114)$$

φ est alors une "**fonction propre**" (en analogie avec les "vecteurs propres") de α , associée à la valeur propre de a . Notons que a peut très bien être nul (vous comprendrez mieux cela au moment où nous passerons à l'étude de cas concrets).

En des termes plus physiques, cela revient à dire que lorsqu'un état (une fonction mathématique au sens formel telle que φ) est inchangé par un opérateur, l'état est alors appelé "**état propre**" ou "**vecteur propre**" du système.

Soit l'ensemble E_a des fonctions propres associées à a et un espace linéaire fonctionnel, que nous nommerons le "**sous-espace propre associé**" à a . Le nombre de dimensions de E_a s'appelle "**multiplicité**" (ou "**ordre de dégénérescence**") de la valeur propre a , et nous le notons g .

Soit maintenant a une valeur propre simple, ou non dégénérée, $g = 1$. Cela veut dire qu'il y a une seule fonction propre associée à a , à un coefficient multiplicatif non nul près.

Si $g = 2$ (valeur propre double), nous pouvons trouver deux fonctions propres non proportionnelles (non liées) associées à a , etc.

 **Exemple:**

Voyons un exemple particulier d'une fonction propre avec une valeur propre autre que le cas classique de l'énergie.

Soit:

$$\alpha \cdot \Psi = a \cdot \Psi \quad (42.115)$$

avec $\alpha = -i\hbar \frac{\partial}{\partial x}$ (opérateur que nous avons déjà vu précédemment) et a une valeur propre.

L'équation devient:

$$-i\hbar \frac{\partial \Psi}{\partial x} = a \Psi \quad (42.116)$$

qui se vérifie aisément si:

$$\boxed{\Psi = e^{\frac{iax}{\hbar}}} \quad (42.117)$$

qui est bien une fonction propre de l'opérateur susmentionné et qui nous sera des plus utiles dans ce qui va suivre.

3.3.1. ORTHOGONALITÉ DES FONCTIONS PROPRES

Deux fonctions (vecteurs) propres φ et φ' associées à deux valeurs propres différentes d'un même opérateur hermitique α sont orthogonales, c'est-à-dire que:

$$\langle \varphi | \varphi' \rangle = 0 \quad (42.118)$$

Démonstration:

Partons cette fois-ci d'abord avec la notation de Dirac avec deux fonctions propres et deux valeurs propres associées:

$$\alpha \cdot \varphi = a \cdot \varphi$$

$$\alpha \cdot \varphi' = a' \cdot \varphi' \quad (42.119)$$

avec $a' \neq a$.

Nous multiplions respectivement les deux relations précédentes par $\bar{\varphi}'$ et $\bar{\varphi}$, et nous intégrons pour obtenir le produit scalaire fonctionnel:

$$(1) \quad \langle \varphi' | \alpha | \varphi \rangle = a \langle \varphi' | \varphi \rangle$$

$$(2) \quad \langle \varphi | \alpha | \varphi' \rangle = a' \langle \varphi | \varphi' \rangle \quad (42.120)$$

Rappelons pour continuer que nous avons démontré que:

$$\langle \psi | \alpha | \chi \rangle = \overline{\langle \chi | \alpha^\dagger | \psi \rangle} \quad (42.121)$$

donc si l'opérateur α est autoadjoint (ce qui est le cas de l'hamiltonien comme nous l'avons montré), c'est-à-dire que $\alpha = \alpha^\dagger$, nous avons:

$$\langle \psi | \alpha | \chi \rangle = \overline{\langle \chi | \alpha | \psi \rangle} \quad (42.122)$$

Dès lors, en retranchant de la relation (1) le complexe conjugué de la relation (2), a étant supposé réel (ou un entier...), nous avons:

$$\begin{aligned} \langle \varphi' | \alpha | \varphi \rangle - \overline{\langle \varphi | \alpha | \varphi' \rangle} &= a \langle \varphi' | \varphi \rangle - a' \overline{\langle \varphi | \varphi' \rangle} \\ \langle \varphi' | \alpha | \varphi \rangle - \langle \varphi' | \alpha | \varphi \rangle &= 0 = a \langle \varphi' | \varphi \rangle - a' \langle \varphi' | \varphi \rangle = (a - a') \langle \varphi' | \varphi \rangle \end{aligned} \quad (42.123)$$

ce qui montre bien que:

$$\langle \varphi | \varphi' \rangle = 0 \quad (42.124)$$

puisque $a' \neq a$.

Faisons maintenant la même démonstration avec l'écriture algébrique classique et une approche un peu différente. Nous partons de:

$$\begin{aligned} (1) \quad \int \bar{\varphi}' \alpha \varphi dV &= a \int \bar{\varphi}' \varphi dV \\ (2) \quad \int \bar{\varphi} \alpha \varphi' dV &= a' \int \bar{\varphi} \varphi' dV \end{aligned} \quad (42.125)$$

Et comme nous avons démontré plus haut que:

$$\overline{\int \bar{\varphi}' \alpha \varphi dV} = \int \bar{\varphi} \alpha^\dagger \varphi' dV \quad (42.126)$$

donc si l'opérateur α est autoadjoint (ce qui est le cas de l'hamiltonien comme nous l'avons montré), c'est-à-dire que $\alpha = \alpha^\dagger$, nous avons:

$$\overline{\int \bar{\varphi}' \alpha \varphi dV} = \int \bar{\varphi} \alpha \varphi' dV \quad (42.127)$$

Il vient alors que les deux relations (1) et (2) sont égales (ne pas oublier que la valeur propre est réelle!). Tel que nous puissions écrire:

$$a \int \bar{\varphi}' \varphi dV = a' \int \bar{\varphi} \varphi' dV \quad (42.128)$$

Nous pouvons alors écrire:

$$(a - a') \int \bar{\varphi}' \varphi dV = 0 \quad (42.129)$$

Et donc si les valeurs propres ne sont pas égales il est forcé que ce soit l'intégrale qui soit nulle (le produit scalaire fonctionnel) et donc que les deux fonctions propres sont orthogonales.

□C.Q.F.D

La même démonstration, mais avec la notation traditionnelle et plus pédagogique donne:

$$\begin{aligned} H \Psi_1 &= E_1 \Psi_1 \\ H \Psi_2 &= E_2 \Psi_2 \end{aligned} \quad (42.130)$$

Si nous multiplions la première équation à gauche par $\bar{\Psi}_2$, et la seconde équation par $\bar{\Psi}_1$, et que nous intégrons sur la totalité de l'espace, nous obtenons les deux expressions suivantes (correspondant donc aux valeurs moyennes):

$$\begin{aligned} \int_{-\infty}^{+\infty} \bar{\Psi}_2 H \Psi_1 dx dy dz &= \int_{-\infty}^{+\infty} \bar{\Psi}_2 E_1 \Psi_1 dx dy dz = E_1 \int_{-\infty}^{+\infty} \bar{\Psi}_2 \Psi_1 dx dy dz \\ \int_{-\infty}^{+\infty} \bar{\Psi}_1 H \Psi_2 dx dy dz &= \int_{-\infty}^{+\infty} \bar{\Psi}_1 E_2 \Psi_2 dx dy dz = E_2 \int_{-\infty}^{+\infty} \bar{\Psi}_1 \Psi_2 dx dy dz \end{aligned} \quad (42.131)$$

Si nous prenons le cas de fonctions réelles, nous pouvons écrire:

$$\begin{aligned} \int_{-\infty}^{+\infty} \Psi_2 H \Psi_1 dx dy dz &= E_1 \int_{-\infty}^{+\infty} \Psi_2 \Psi_1 dx dy dz \\ \int_{-\infty}^{+\infty} \Psi_1 H \Psi_2 dx dy dz &= E_2 \int_{-\infty}^{+\infty} \Psi_1 \Psi_2 dx dy dz \end{aligned} \quad (42.132)$$

L'opérateur H étant hermitique (autoadjoint) comme nous l'avons démontré plus haut, nous avons:

$$\overline{\int_{-\infty}^{+\infty} \Psi_1 H \Psi_2 dx dy dz} = \int_{-\infty}^{+\infty} \Psi_2 H \Psi_1 dx dy dz \quad (42.133)$$

et comme E_2, Ψ_1, Ψ_2 sont admis comme étant des fonctions réelles, nous avons aussi:

$$\overline{E_2 \int_{-\infty}^{+\infty} \Psi_1 \Psi_2 dx dy dz} = E_2 \int_{-\infty}^{+\infty} \Psi_1 \Psi_2 dx dy dz = E_2 \int_{-\infty}^{+\infty} \Psi_2 \Psi_1 dx dy dz \quad (42.134)$$

Donc:

$$\int_{-\infty}^{+\infty} \Psi_2 H \Psi_1 dx dy dz - \overline{\int_{-\infty}^{+\infty} \Psi_1 H \Psi_2 dx dy dz} = E_1 \int_{-\infty}^{+\infty} \Psi_2 \Psi_1 dx dy dz - E_2 \int_{-\infty}^{+\infty} \Psi_1 \Psi_2 dx dy dz \quad (42.135)$$

S'écrit:

$$\int_{-\infty}^{+\infty} \Psi_2 H \Psi_1 dx dy dz - \int_{-\infty}^{+\infty} \Psi_2 H \Psi_1 dx dy dz = E_1 \int_{-\infty}^{+\infty} \Psi_2 \Psi_1 dx dy dz - E_2 \int_{-\infty}^{+\infty} \Psi_2 \Psi_1 dx dy dz \quad (42.136)$$

Il vient alors:

$$(E_1 - E_2) \int_{-\infty}^{+\infty} \Psi_2 \Psi_1 dx dy dz = 0 \quad (42.137)$$

ce qui montre bien que Ψ_1, Ψ_2 sont orthogonales selon la définition du produit scalaire fonctionnel.

3.4. FORMALISME DE DIRAC

Dirac a conçu un formalisme général très pratique, mondialement utilisé par les physiciens, dont nous allons donner les éléments essentiels. Les notations utilisées ont d'ailleurs été déjà partiellement introduites dans ce qui a précédé.

Nous utiliserons le formalisme de Dirac pour deux points, le premier étant de mieux comprendre ce qui a été vu jusqu'à maintenant lors de l'introduction aux opérateurs fonctionnels, le second étant d'introduire une notation et une méthode de résolution que l'on retrouve dans certains ouvrages. Par ailleurs, sur ce site par simplification d'écriture, nous utiliserons parfois ce formalisme.

3.4.1. KETS ET BRAS

Nous considérons un espace vectoriel E_n à n dimensions où n peut très bien être infini (espace de Hilbert). Un vecteur est défini par n composantes x_i ($i = 1, 2, \dots, n$) que nous pouvons ranger en colonne pour former une matrice-colonne:

$$\begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \quad (42.138)$$

Nous dirons que cette matrice décrit le "**vecteur droit**" ou le "**ket**" $|x\rangle$ (cela doit vous rappeler les "représentatives"). Il est possible d'associer à la matrice colonne la matrice adjointe (transposée conjuguée):

$$\overline{x_1} \ \overline{x_2} \ \dots \ \overline{x_n} \quad (42.139)$$

où les $\overline{x_i}$ sont les complexes conjugués des x_i . Nous dirons que la matrice ligne adjointe décrit le "**vecteur gauche**" ou le "**bra**" $\langle x|$ (cela doit également vous rappeler les "représentatives").

L'addition et la multiplication par un nombre λ vont de soi. Notons que si $|y\rangle = \lambda |x\rangle$, nous avons trivialement $\langle y| = \overline{\lambda} \langle x|$.

Avec deux vecteurs de composantes x_i et y_i , nous pouvons former la quantité suivante, dite "**produit scalaire hermitique**":

$$\overline{\begin{pmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_n \end{pmatrix}} \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} = \sum_i \bar{x}_i y_i \quad (42.140)$$

nous convenons de l'écrire $\langle x|y \rangle$. Notons que:

$$\langle x|y \rangle = \overline{\langle y|x \rangle} \quad (42.141)$$

le produit scalaire hermitique n'est donc pas simplement commutatif!

Le produit $\langle x|y \rangle$ dépend linéairement de $\langle x|$ et de $|y \rangle$. Réciproquement si un nombre Q dépend linéairement d'un ket $|x \rangle$, il existe un bra $\langle a|$ tel que:

$$Q = \langle a|x \rangle \quad (42.142)$$

En physique quantique, $\langle x|y \rangle$ est appelé "**l'amplitude**" d'être dans l'état x si le système est dans l'état y . Ce produit scalaire hermitique sera interprété comme la probabilité que le système physique se projette dans l'état x s'il est dans l'état y .

Une base orthonormée de l'espace étudié est constituée par n vecteurs $|i \rangle$ tels que:

$$\langle i|j \rangle = \delta_{ij} \quad (42.143)$$

où rappelons-le, δ_{ij} est le symbole de Kronecker (cf. chapitre de Calcul Tensoriel).

Tout vecteur $|x \rangle$ de E_n peut se développer sur cette base selon (cf. chapitre de Calcul Vectoriel):

$$|x \rangle = \sum_i x_i |i \rangle \quad (42.144)$$

où les x_i sont les composantes de $|x \rangle$ dans la base choisie. Nous vérifions vraiment aisément que (déjà vu maintes fois dans le chapitre de Calcul Vectoriel):

$$x_i = \langle i|x \rangle \quad (42.145)$$

Si un ket $|y \rangle$ dépend linéairement d'un ket $|x \rangle$, nous écrivons symboliquement:

$$|y \rangle = \alpha |x \rangle \quad (42.146)$$

où α est un opérateur linéaire. Soit donc un opérateur linéaire défini par la relation précédente et un bra $\langle u|$, le produit scalaire hermitique:

$$\langle u|y \rangle = \langle u|[\alpha |x \rangle] = Q \quad (42.147)$$

est un nombre Q qui dépend linéairement de $|x\rangle$. D'après ce qui a été vu plus haut, il existe un bra $\langle v|$ tel que $Q = \langle v|x\rangle$. $\langle v|$ dépend visiblement de $\langle u|$ de manière linéaire. Nous convenons de poser:

$$\langle v| = \langle u|\alpha \quad (42.148)$$

À l'aide de cette convention, nous pouvons écrire:

$$\langle u|[\alpha|x]\rangle = [\langle u|\alpha]|x\rangle = \langle u|\alpha|x\rangle \quad (42.149)$$

Si $|y\rangle = \alpha|x\rangle$, $\langle y|$ dépend linéairement de $\langle x|$. Par définition, nous écrivons:

$$\langle y| = \langle x|\alpha^\dagger \quad (42.150)$$

où $\alpha^\dagger$ est l'opérateur adjoint de α .

Formons avec un bra $\langle u|$ le produit scalaire hermitique:

$$\langle u|y\rangle = \langle u|\alpha|x\rangle \quad (42.151)$$

et nous pouvons écrire (nous l'avons démontré précédemment):

$$\overline{\langle u|y\rangle} = \langle y|u\rangle = \langle x|\alpha^\dagger|u\rangle \quad (42.152)$$

d'où la relation de première importance que nous avons déjà rencontrée plusieurs fois sans en avoir expliqué vraiment l'origine:

$$\boxed{\langle x|\alpha^\dagger|u\rangle = \overline{\langle u|\alpha|x\rangle}} \quad (42.153)$$

Nous rappelons simplement avec cette relation qu'un opérateur hermitique est un opérateur égal à son adjoint.

Grâce au formalisme de Dirac, ce qui était avant des définitions abstraites est devenu maintenant des évidences démontrées.

Pour résumer:

Formalisme classique	Formalisme de Dirac
$\int \bar{\varphi}_i \varphi_j dV$	$\langle \varphi_i \varphi_j \rangle$
$\int \bar{\varphi}_i \alpha \varphi_j dV$	$\langle \varphi_i \alpha \varphi_j \rangle$
$\int \bar{\varphi}_i \varphi_j dV = \int \bar{\varphi}_j \varphi_i dV$	$\langle \varphi_i \varphi_j \rangle = \langle \varphi_j \varphi_i \rangle$
$\int \bar{\varphi}_i \alpha \varphi_j dV = \int \bar{\varphi}_j \alpha^\dagger \varphi_i dV = \int \bar{\varphi}_j \alpha \varphi_i dV$	$\langle \varphi_i \alpha \varphi_j \rangle = \langle \varphi_j \alpha \varphi_i \rangle$
$\Psi = \sum_i c_i \varphi_i$	$ \Psi\rangle = \sum_i c_i \varphi_i\rangle$
$\sum_i c_i' \int \bar{\varphi}_j \cdot \alpha \cdot \varphi_i dx$	$\sum_i c_i' \langle \varphi_j \alpha \varphi_i \rangle$
$c_i = \int \bar{\varphi}_i \Psi dV$	$c_i = \langle \varphi_i \Psi \rangle$

Tableau: 42.1 - Comparaison formalisme classique/Dirac avec ce qui a été vu jusqu'à maintenant

Remarque: À nouveau, un excellent exemple pratique d'application du formalisme de Dirac est proposé dans le chapitre d'Informatique Quantique (voir section d'Informatique Théorique).

4. MODÈLE DE SCHRÖDINGER

Des expériences (effet Compton, effet photoélectrique, fentes de Young, optique géométrique/ondulatoire, etc.) ont montré que les ondes pouvaient, dans certaines situations être traitées comme des corpuscules (et inversement). Ce sont ces observations qui amenèrent Niels Bohr à énoncer son "principe de complémentarité" qui dit que suivant les expériences effectuées, il faut considérer la matière soit comme une onde, soit comme des corpuscules. Ces deux aspects se complétant l'un et l'autre.

4.1. ONDE ASSOCIÉE DE DE BROGLIE

Le physicien français Louis Victor De Broglie suggère, en 1924, que réciproquement, les particules (électrons, protons, et autres ainsi que même les atomes et les molécules) pourraient aussi, dans certains cas, montrer des propriétés d'ondes au même titre que la lumière! De Broglie émit alors l'idée qu'il existait entre la longueur d'onde associée fictive d'une particule de matière et sa quantité de mouvement, une relation similaire à celle d'un photon, soit (ν est la notation pour la fréquence pour rappel...):

$$E_{tot} = h \cdot \nu = h \cdot \frac{\omega}{2\pi} = \hbar \cdot \omega = m \cdot c^2 \quad (42.154)$$

donc nous pouvons écrire en utilisant la relation établie dans la chapitre de Mécanique Ondulatoire:

$$p = m \cdot c = \frac{E_{tot}}{c} = \frac{h \cdot \nu}{c} = \frac{\hbar \cdot \omega}{c} = \frac{\hbar \cdot k \cdot c}{c} = \hbar \cdot k = \frac{h}{\lambda} \quad (42.155)$$

où le rapport:

$$\lambda = \frac{h}{p} \quad (42.156)$$

De Broglie émet donc dès lors l'hypothèse suivante: Pour un corpuscule de masse m et de vitesse v nous avons:

$$\lambda = \frac{h}{mv} \quad (42.157)$$

où λ est appelé "**longueur d'onde associée de De Broglie**".

La matière en mouvement aurait donc une longueur d'onde associée. C'est une longueur d'onde extrêmement petite pour des masses de l'ordre du kilogramme. Si la vitesse est par exemple de l'ordre de $1 [m \cdot s^{-1}]$ alors $\lambda \cong 10^{-34} [m]$.

Comme nous l'avons vu, les phénomènes d'interférence et de diffraction sont importants seulement lorsque la taille des objets ou fentes n'est pas beaucoup plus grande que la longueur d'onde. Il est donc impossible de détecter les propriétés ondulatoires des objets de tous les jours. Il n'en est pas de même pour les particules élémentaires, les électrons en particulier.

Les électrons peuvent donc avoir des longueurs d'onde de l'ordre de $10^{-10} [m]$ ce qui correspond à l'espacement des atomes d'un cristal. C.J. Davisson et L.H. Germer exécutèrent une expérience cruciale: ils diffusèrent des électrons sur la surface d'un cristal et au début 1927 observèrent que les électrons éjectés étaient distribués en pics réguliers. Lorsqu'ils interprétèrent ces pics comme des pics de diffraction, ils trouvèrent que la longueur d'onde de l'électron diffracté était exactement celle prédite par De Broglie.

Mais alors qu'est-ce qu'un électron?? Les illustrations qui montrent un électron comme une minuscule sphère chargée négativement ne sont que des images commodes, mais inexactes. En fait, nous devons utiliser le modèle corpusculaire ou ondulatoire, celui qui fonctionne le mieux selon la situation de façon à pouvoir comprendre ce qui se produit. Mais il ne faut pas en conclure qu'un électron est une onde ou une particule. Nous devrions plutôt dire qu'un électron est "l'ensemble de ses propriétés mesurables". Certains physiciens emploient encore l'expression "**quanton**" pour décrire tout système se comportant soit comme une onde soit comme une particule.

De Broglie put alors suggérer que chaque orbite électronique quantifiée (selon le postulat de quantification de Bohr) est alors une onde stationnaire et qu'un électron ne pouvait occuper que les orbites pouvant accepter un nombre entier de longueurs d'onde de son onde associée fictive. S'il n'y avait pas de coïncidence exacte, il ne pouvait y avoir d'onde stationnaire, et donc pas d'orbite stationnaire non plus.

Dès lors, comme pour les modes résonnants d'une corde (ondes stationnaires), seules les ondes dont la circonférence de l'orbite circulaire contient un nombre entier de λ existent, soit (l'amplitude sera calculée par des techniques particulières que nous verrons plus loin):

$$2\pi r_n = n\lambda \text{ avec } n \in \mathbb{N}^*$$

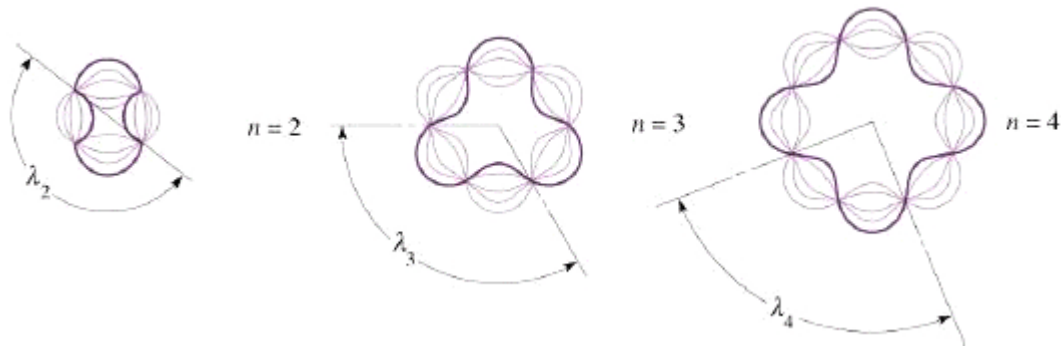


Figure: 42.2 - Approche de l'aspect ondulatoire des orbites par De Broglie

En remplaçant λ par h/mv , nous obtenons:

$$2\pi r_n = \frac{nh}{mv} \Leftrightarrow mvr_n = \frac{nh}{2\pi} \quad (42.158)$$

Ce qui est bien la condition quantique proposée par Bohr. Les orbites et les états d'énergie quantifiés du modèle de Bohr, sont dus à la nature ondulatoire de l'électron et au fait que seules des ondes stationnaires résonantes persistent. Ceci suppose que la dualité onde-corpuscule est à la base de la structure de l'atome.

La notion ondulatoire de la particule permet ensuite au physicien Erwin Schrödinger de développer une équation dite "**équation d'onde**" pour décrire les propriétés ondulatoires des particules.

Petit interlude sympathique... puisque connue l'onde associée de De Broglie et étant donné le résultat vu lors de notre étude du théorème du Viriel dans le chapitre de Mécanique Des Milieux Continus, nous pouvons mettre en relation:

$$\lambda = \frac{h}{p} = \frac{h}{mv} \quad E_c = \frac{1}{2}mv^2 = \frac{1}{2}pv = \frac{3}{2}kT \Rightarrow v = \sqrt{\frac{3kT}{m}} \quad (42.159)$$

Ainsi, nous pouvons pour un fluide (liquide), obtenir la valeur de "**l'onde thermique associée de De Broglie**". Ce qui nous donne:

$$\lambda = \frac{h}{mv} = \frac{h}{m\sqrt{\frac{3kT}{m}}} = \frac{h}{\sqrt{3mkT}} \quad (42.160)$$

Nous reviendrons sur cette relation lors de notre étude des superfluides dans le chapitre de Mécanique Des Milieux Continus.

4.2. ÉQUATION CLASSIQUE DE SCHRÖDINGER

Le physicien Peter Debye trouva le modèle de De Broglie tiré par les cheveux. Il argumenta que la physique des ondes, n'importe lesquelles, depuis les ondes sonores jusqu'aux ondes électromagnétiques, même les ondes se propageant sur une corde exige une équation qui les décrive. Or, il n'y avait pas d'équation d'onde la modèle atomique de De Broglie (pour la simple raison que ce dernier n'avait jamais essayé et Einstein non plus). Le physicien Erwin Schrödinger prit alors sur lui de trouver l'équation qui manquait et le fit avec brio.

Rappelons la forme unidimensionnelle de l'équation d'onde (cf. chapitre de Mécanique Ondulatoire):

$$(1) \quad \frac{\partial^2 f(x,t)}{\partial x^2} = \frac{1}{v^2} \frac{\partial^2 f(x,t)}{\partial t^2} \quad (42.161)$$

Pour simplifier, cherchons une solution particulière de la forme (voir le chapitre de Mécanique Ondulatoire ou le chapitre d'Électrodynamique pour l'analogie):

$$(2) \quad f(x,t) = \Psi(x) \cdot \sin(\omega \cdot t) \quad (42.162)$$

$\Psi(x)$ est l'amplitude du champ associé à la particule. Il est important de remarquer que la partie périodique ne contient pas de paramètres de déplacement (comme c'est le cas en électrodynamique par exemple) car la fonction se doit de décrire des solutions "statiques" (attention à ne pas prendre ce terme à la lettre).

Pour des raisons historiques cette amplitude est couramment appelée "**fonction d'onde**" bien que cette appellation soit trompeuse. Il serait peut-être meilleur de l'appeler simplement "**amplitude du champ associé à la matière**".

C'est la recherche de l'expression de cette fonction qui va nous amener lors de l'étude d'un cas particulier (bien plus loin dans le texte) à l'expression bien connue de l'énergie d'ionisation d'un électron de nombre quantique n donné et pour son atome de numéro atomique N donné.

Si nous introduisons (2) dans (1), nous obtenons:

$$\begin{aligned} \cancel{\sin(\omega t)} \frac{\partial^2 \Psi(x)}{\partial x^2} &= -\frac{\omega^2}{v^2} \Psi(x) \cancel{\sin(\omega t)} \\ \Rightarrow \frac{\partial^2 \Psi(x)}{\partial x^2} &= -\frac{\omega^2}{c^2} \Psi(x) \Rightarrow \frac{\partial^2 \Psi(x)}{\partial x^2} = -\frac{(2\pi\nu)^2}{c^2} \Psi(x) \\ (3) \quad \Rightarrow \frac{\partial^2 \Psi(x)}{\partial x^2} &= -\left(\frac{2\pi c}{\lambda}\right)^2 \Psi(x) \Rightarrow \frac{\partial^2 \Psi(x)}{\partial x^2} = -\frac{4\pi^2}{\lambda^2} \Psi(x) \\ \Rightarrow \frac{\partial^2 \Psi(x)}{\partial x^2} + \frac{4\pi^2}{\lambda^2} \Psi(x) &= 0 \end{aligned} \quad (42.163)$$

Nous avons aussi:

$$E_{cin} = \frac{1}{2} m \cdot v^2 = \frac{1}{2 \cdot m} \cdot p^2 = \frac{1}{2 \cdot m} \cdot \left(\frac{h}{\lambda}\right)^2 = E_{tot} - E_{pot} \quad (42.164)$$

d'où:

$$(4) \quad \frac{1}{\lambda^2} = \frac{2m \cdot E_{cin}}{h^2} = \frac{2m \cdot (E_{tot} - E_{pot})}{h^2} \quad (42.165)$$

si nous introduisons (4) dans (3) nous obtenons alors "**l'équation de Schrödinger unidimensionnelle classique**" (en l'absence de champ magnétique...):

$$\frac{\partial^2 \Psi}{\partial x^2} + \frac{8\pi^2 \cdot m(E_{\text{tot}} - E_{\text{pot}})}{h^2} \Psi(x) = 0 \quad (42.166)$$

Remarque: L'énergie potentielle pourrait aussi bien être gravitationnelle, qu'électrique ou les deux combinées (donc de nature quelconque). Mais la gravitation est tellement faible à cette échelle par rapport aux forces électrostatiques qu'elle est négligée.

Nous pouvons réécrire l'équation précédente en la généralisant à un système à trois dimensions. Ce qui nous donne finalement:

$$\Delta \Psi + \frac{2m}{h^2} (E_{\text{tot}} - E_{\text{pot}}) \Psi = 0 \quad (42.167)$$

où Δ n'est pour rappel que le laplacien d'un champ scalaire (cf. chapitre de Calcul Vectoriel):

$$\Delta = \frac{\partial^2}{\partial^2 x} + \frac{\partial^2}{\partial^2 y} + \frac{\partial^2}{\partial^2 z} \quad (42.168)$$

Remarques:

R1. Cette équation n'est pas un invariant de Lorentz étant donné qu'elle a été établie à partir de l'expression classique de l'énergie (et non relativiste).

R2. La fonction d'onde plane que nous avons prise au départ n'a pas une signification physique étant donné qu'elle transporte une énergie infinie. Une meilleure solution est de considérer un paquet d'ondes. Toutefois, les paquets d'onde généralement employés sont constitués d'une superposition d'ondes planes. Dès lors, en étudiant ses effets sur une des ondes planes, nous pouvons accepter les conclusions physiques que nous pouvons en déduire.

Si simple que cela puisse paraître, il fallut toute l'habileté et l'expérience de Schrödinger pour être le premier à écrire cette équation d'onde et ce fut le fondement sur lequel il construisit l'édifice mathématique de la physique quantique ondulatoire dans les mois qui suivirent. Mais d'abord il lui fallait prouver que c'était bien l'équation d'onde recherchée en l'appliquant à l'atome d'hydrogène (cf. chapitre de Chimie Quantique). Son modèle élimina tous les bricolages successifs et empiriques de la physique quantique corpusculaire.

Certains physiciens jugèrent le modèle de Schrödinger comme de la folie pure (en particulier Sommerfeld) tout en changeant d'avis peu de temps après... tellement le modèle était efficace et remplaçait avantageusement le modèle matriciel horriblement abstrait de Heisenberg (à grand malheur de ce dernier) sur lequel même Pauli avait échoué pour les résultats expérimentaux de l'atome d'hydrogène. Peu de temps après Max Born décrivit le modèle ondulatoire comme la forme la plus profonde des lois quantiques ce qui évidemment ne fit pas tellement plaisir à son proche ami Heisenberg...

Cependant le modèle de Heisenberg décrivait des particules alors que Schrödinger décrit des ondes. L'avantage d'avoir deux formalismes différents - mais équivalents - en physique quantique devint vite évident. Pour la plupart des problèmes que rencontrent les physiciens, la mécanique ondulatoire offre la voie la plus facile vers la solution. Toutefois, pour d'autres, comme ceux qui impliquent le spin, c'est la démarche matricielle de Heisenberg qui prouve sa valeur.

4.2.1. HAMILTONIEN DE SCHRÖDINGER

L'équation de Schrödinger peut également s'écrire sous la forme (après quelques petites mises en facteurs élémentaires) suivante:

$$\left(E_{pot} - \frac{\hbar^2}{2m} \Delta \right) \Psi = E_{tot} \Psi \quad (42.169)$$

Nous écrivons cela en physique quantique sous la forme:

$$\boxed{H \cdot \Psi = E_{tot} \cdot \Psi} \quad (42.170)$$

où H est donc l'hamiltonien du système (ou énergie totale) et constitue un opérateur fonctionnel et l'énergie totale, la valeur propre.

L'équation de Schrödinger est donc une équation aux dérivées partielles du second ordre, linéaire homogène. Quelle que soit l'énergie totale, elle admet des solutions (ouf!), mais nous montrons qu'en général ces solutions croissent très rapidement (croissance de type exponentiel) quand nous nous éloignons à l'infini dans certaines directions et sont donc physiquement inacceptables. Il n'y a que des valeurs particulières de l'énergie totale qui donnent lieu à des solutions physiquement acceptables et en général, l'ensemble de ces valeurs comprend des valeurs discrètes (fonctions trigonométriques à la source) qui sont les "**niveaux liés**" du système (parce que leur fonction propre décroît rapidement à l'infini) et un continuum de valeurs qui sont les "**niveaux non liés**" (leur fonction propre restant finie à l'infini). Plus précisément, si W est la borne inférieure des valeurs de l'énergie potentielle à l'infini, les niveaux liés se situent au-dessous de W , alors que les valeurs supérieures à W constituent le continuum des niveaux non liés.

Par exemple, dans l'étude de l'oscillateur harmonique (un des cas pratiques les plus difficiles au niveau du formalisme) que nous ferons plus loin, nous avons:

$$E_{pot} = \frac{Cx^2}{2} \quad (42.171)$$

avec $W = +\infty$. Il n'existe donc que des niveaux liés.

Dans l'atome d'hydrogène:

$$E_{pot} = \frac{-e^2}{4\pi\epsilon_0 r} \quad (42.172)$$

avec $W = 0$. Les niveaux liés seront négatifs, et toutes les valeurs positives de l'énergie seront des niveaux non liés.

Ceci ayant été dit, voyons également comme exemple (très important) la manière de déterminer l'hamiltonien H de l'équation de Schrödinger d'une particule chargée non relativiste dans un champ électromagnétique.

Nous avons vu dans le chapitre de Mécanique Analytique que le lagrangien était défini par la soustraction de l'énergie cinétique et potentielle selon la relation:

$$L = T - V \quad (42.173)$$

Nous avons dans le chapitre d'Électrodynamique que le lagrangien de l'interaction champ-courant relativiste était donné par:

$$L = L_0 + L_1 = -m_0 c^2 \sqrt{1 - v^2 / c^2} + q \vec{A} \circ \vec{v} - q \phi \quad (42.174)$$

où pour rappel ϕ est le potentiel vecteur du champ électrique (donc dont le gradient est le champ électrique $\vec{E}$) et $\vec{A}$ le potentiel vecteur du champ magnétique (dont le rotationnel est le champ magnétique $\vec{B}$).

Si nous rajoutons un champ électrique (et donc un potentiel électrostatique U) en plus du champ électromagnétique le lagrangien s'écrit alors (puisque le potentiel se soustrait selon la définition du lagrangien!):

$$L = L_0 + L_1 = -m_0 c^2 \sqrt{1 - v^2 / c^2} - q \cdot U + q \vec{A} \circ \vec{v} - q \phi \quad (42.175)$$

Dans l'approximation classique (non relativiste) nous savons que nous avons (cf. chapitre de Relativité Restreinte):

$$L = -m_0 c^2 + \frac{1}{2} m_0 v^2 - q \cdot U + q \vec{A} \circ \vec{v} - q \phi \quad (42.176)$$

Comme nous nous restreignons au cas non relativiste, nous pouvons éliminer le terme constant d'énergie de la masse au repos tel que:

$$L = \frac{1}{2} m_0 v^2 - q \cdot U + q \vec{A} \circ \vec{v} - q \phi \quad (42.177)$$

Toujours dans le chapitre de Mécanique Analytique, nous avons démontré que l'hamiltonien était donné par:

$$H = \sum_i \dot{x}_i p_i - L \quad (42.178)$$

Nous avons donc:

$$H = \sum_i \dot{x}_i p_i - \frac{1}{2} m_0 v^2 + q \cdot U - q \vec{A} \circ \vec{v} + q \phi \quad (42.179)$$

De plus, nous avons vu dans le chapitre de Mécanique Analytique que:

$$p_i = \frac{\partial L}{\partial \dot{x}_i} \quad (42.180)$$

Il vient donc que:

$$p_i = \frac{\partial}{\partial \dot{x}_i} \left(\frac{1}{2} m_0 v_i^2 - q \cdot U + q \vec{A} \circ \vec{v} - q \phi \right) = m_0 v_i + q A_i \quad (42.181)$$

Finalement:

$$H = m_0 v^2 + q \vec{A} \circ \vec{v} - \frac{1}{2} m_0 v^2 + q \cdot U - q \vec{A} \circ \vec{v} + q \phi \quad (42.182)$$

Soit après simplification:

$$H = \frac{1}{2} m_0 v^2 + q \cdot U + q \phi = \frac{1}{2} m_0 v^2 + q \cdot U + V \quad (42.183)$$

H contient donc l'énergie cinétique et l'énergie potentielle totale. Il n'y a pas de terme magnétique car la force de Laplace, comme nous l'avons démontré dans le chapitre de Magnétostatique, ne travaille pas (y'en a qui ont de la chance...). H est bien l'énergie totale du système classique, cependant la relation précédente n'est pas vraiment adaptée au formalisme de Hamilton car les moments conjugués n'apparaissent pas. Mais il est très simple de les introduire à partir du résultat obtenu précédemment qui était:

$$p_i = m_0 v_i + q A_i \Rightarrow v_i = \frac{p_i - q A_i}{m_0} \quad (42.184)$$

donc:

$$H = \frac{1}{2} m_0 v^2 + q \cdot U + V = \frac{1}{2} m_0 \left(\frac{p_i - q A_i}{m_0} \right)^2 + q \cdot U + V = \frac{(p_i - q A_i)^2}{2 m_0} + q \cdot U + V \quad (42.185)$$

Si nous passons en physique quantique, nous devons remplacer les p_i par leurs opérateurs respectifs:

$$p_i = -i \hbar \partial_i \quad (42.186)$$

dont nous avons démontré l'origine plus haut. Ainsi, nous avons:

$$H = \frac{(p_i - q A_i)^2}{2 m_0} + q \cdot U + V \quad (42.187)$$

qui doit s'écrire dans le cas général (comme nous ne savons pas si le potentiel vecteur commute avec la quantité de mouvement, pour ne pas perdre en généralité nous allons supposer qu'il anticommute):

$$H_i = \frac{p_i^2}{2 m_0} - \frac{q (p_i A_i + A_i p_i)}{2 m_0} + \frac{q^2 A_i^2}{2 m_0} + q \cdot U + V \quad (42.188)$$

Ce que l'on note traditionnellement sous la forme (sic!):

$$H = \frac{p^2}{2m_0} - q \frac{(pA + Ap)}{2m_0} + \frac{q^2 A^2}{2m_0} + qU + V \quad (42.189)$$

On retrouve très souvent cette dernière relation sous la forme simplifiée suivante dans la littérature en absence de potentiel U et en explicitant V :

$$H = \frac{1}{2m_0} (p - qA)^2 + q\phi \quad (42.190)$$

Remarque: Dans le chapitre de Physique Quantique Relativiste nous démontrerons la forme relativiste de cet hamiltonien associée à l'équation de Klein-Gordon généralisée ou encore celui de Dirac qui inclut le spin.

Heureusement, nous ne traiterons pas d'exemples où il faudra trouver des solutions à l'équation de Schrödinger avec un tel hamiltonien sur ce site...

Dans les cas scolaires quand on aborde toutefois des prémisses d'application..., on annule soit le potentiel vecteur (donc la particule n'est pas plongée dans un champ magnétique), soit le potentiel scalaire (donc la particule n'est pas plongée dans un champ électrique). De plus quand on choisit l'un des champs comme étant non nul, on s'arrange à prendre le cas où il est constant et en une seule dimension... Ainsi, si on veut un champ électrique constant selon un seul axe (par exemple l'axe X), on prendra comme potentiel scalaire:

$$\phi(x,t) = E_0 \cdot x \quad (42.191)$$

puisque son gradient donnera bien une constante selon X . Dans le cas d'un champ magnétique constant selon un seul axe (par exemple l'axe Z), on s'arrangera pour prendre le potentiel vecteur:

$$A = \frac{B_0}{2} \begin{pmatrix} -y \\ x \\ 0 \end{pmatrix} \quad (42.192)$$

dont le rotationnel donne bien un champ magnétique constant selon Z .

4.2.2. CONDITION DE NORMALISATION DE DE BROGLIE

En général, dans un état dynamique donné, la particule (s'il s'agit d'un système à une particule) décrite par la résolution de l'équation de Schrödinger pour des paramètres bien définis est mal localisée, car x, y, z sont mal déterminés de par même le principe d'incertitude de Heisenberg. Il y a donc lieu de définir une probabilité dP de trouver la particule dans l'élément de volume $dx dy dz$ entourant un point (x, y, z) , d'où l'existence d'une fonction de distribution des coordonnées $\rho(x, y, z)$ telle que:

$$dP = \rho(x, y, z) dx dy dz \quad (42.193)$$

où $\rho(x, y, z)$ est donc une quantité essentiellement positive ou nulle (probabilités obligent!) qui doit s'exprimer à l'aide de la fonction de Schrödinger $\Psi(x, y, z)$. Nous avons d'ailleurs de tels exemples très détaillés dans le présent chapitre et à la fin de celui de Chimie Quantique.

Des analogies avec la physique ondulatoire classique, plus précisément avec l'électrodynamique, ont conduit à admettre que comme la densité volumique d'énergie d'une onde électromagnétique est proportionnelle au carré de son amplitude (cf. chapitre d'Électrodynamique), la densité volumique de probabilité devait être proportionnelle au carré de l'intensité du champ associé tel que:

$$\rho(x, y, z) = c^{te} \cdot |\Psi(x, y, z)|^2 \quad (42.194)$$

où nous utilisons le module de la fonction de Schrödinger comme analogie de l'amplitude et où la constante est un nombre réel. Dans le cadre de la physique quantique, il est beaucoup plus fréquent de trouver cette dernière relation sous la forme évidente suivante:

$$\rho(x, y, z) = \frac{|\Psi(x, y, z)|^2}{\int |\Psi(x, y, z)|^2 dx dy dz} = \left| \frac{\Psi(x, y, z)}{\sqrt{\int |\Psi(x, y, z)|^2 dx dy dz}} \right|^2 = c^{te} |\Psi(x, y, z)|^2 \quad (42.195)$$

qui met mieux en évidence la normalisation nécessaire de la fonction de Schrödinger.

Enfin, il est important de noter que les physiciens ont pour habitude pendant les développements de garder la même notation pour la fonction de Schrödinger non normalisée que celle normalisée (ce qui peut prêter à confusion) telle que:

$$|\Psi(x, y, z)|^2 = \left| \frac{\Psi(x, y, z)}{\sqrt{\int |\Psi(x, y, z)|^2 dx dy dz}} \right|^2 \quad (42.196)$$

où $|\Psi(x, y, z)|^2$ représente alors la probabilité de trouver la particule en un certain point de l'espace.

Il est évident alors qu'avec cette manière de noter les choses nous avons alors sur tout l'espace:

$$\boxed{\iiint \rho(x, y, z) dx dy dz = \iiint |\Psi(x, y, z)|^2 dx dy dz = 1} \quad (42.197)$$

comme déjà mentionné lors de notre présentation du premier postulat.

Nous pouvons maintenant considérer la signification physique qui peut-être attachée à l'intensité du champ associé à la matière. Comme ce champ décrit le mouvement d'une particule, nous pouvons dire que les régions de l'espace dans lesquelles la particule a le plus de chance de se trouver sont celles dans lesquelles $|\Psi(x, y, z)|^2$ est maximum.

Indiquons aussi que la relation antérieure s'écrit en utilisant la notation ket-bra, que nous introduirons un peu plus tard, de la manière très raffinée (et très courante...) suivante:

$$\|\Psi\rangle^2 = \left| \frac{|\Psi\rangle}{\sqrt{\langle\P|\Psi\rangle}} \right|^2 = \frac{\|\Psi\|^2}{\langle\P|\Psi\rangle} \quad (42.198)$$

ou suivante (c'est simplement la racine carrée de la précédente):

$$|\Psi\rangle = \frac{|\Psi\rangle}{\sqrt{\langle\Psi|\Psi\rangle}} \quad (42.199)$$

où le module au dénominateur disparaît puisque, pour rappel, l'intégrale est un nombre réel. Il ne faut par ailleurs jamais oublier que les physiciens, pour la grande majorité, notent de manière identique la fonction de Schrödinger non normalisée et normalisée comme nous le rappelle cette dernière relation.

Comme nous l'avons déjà dit, nous verrons de nombreux exemples détaillés de cette normalisation dans ce chapitre avec des espaces unidimensionnels et dans le cadre de volumes dans le chapitre de Chimie Quantique.

Indiquons encore une chose en ce qui concerne la normalisation. Si vous observez l'expression de l'hamiltonien des équations de Schrödinger vues jusqu'à présent, alors si λ est une constante réelle ou complexe nous avons toujours:

$$H(\lambda\Psi) = \lambda H(\Psi) \quad (42.200)$$

Si nous posons que Ψ est une solution de l'équation de Schrödinger, nous voyons alors que $\lambda\Psi$ est aussi solution de l'équation. En effet, nous obtenons:

$$H(\lambda\Psi) = \lambda H(\Psi) = \lambda E\Psi = E(\lambda\Psi) \quad (42.201)$$

En prenant en compte le fait que la fonction $\lambda\Psi$ est normalisée, nous obtenons alors:

$$\int |\lambda\Psi|^2 dx = 1 \Rightarrow |\lambda|^2 \int |\Psi|^2 dx = 1 \Rightarrow |\lambda|^2 = 1 \quad (42.202)$$

d'où (certains livres se restreignent à ces solutions pour des raisons pédagogiques):

$$\lambda = \pm 1 \text{ ou } \lambda = \pm i \quad (42.203)$$

ou rigoureusement nous avons plus généralement:

$$\lambda = e^{i\vartheta} \quad (42.204)$$

où ϑ est un nombre réel. C'est ce que nous appelons "**l'arbitraire de phase**" dont nous avons déjà fait mention au début de ce chapitre sans démonstration.

Ces solutions sont normalisées et correspondent à la même valeur d'énergie E ainsi qu'à la même densité de probabilité. Ceci montre qu'il n'est pas utile de chercher la signification d'une valeur négative de Ψ (si nous prenons le cas particulier pédagogique $\lambda = -1$), car $|\Psi|^2$ est réel et n'est pas négatif. Seul le carré d'une fonction d'onde, qui correspond à la densité de probabilité, est significatif d'un point de vue physique.

4.2.3. ÉTATS LIÉS ET NON LIÉS

Supposons que $\Psi(x, y, z)$ décroisse assez rapidement à l'infini, de telle sorte que l'intégrale:

$$\int_{-\infty}^{+\infty} |\Psi(x, y, z)|^2 dx dy dz \quad (42.205)$$

converge. Il est alors possible de profiter de l'arbitraire régnant sur la fonction d'onde (le fait que Ψ et $\lambda\Psi$ décrivent le même état) pour rendre cette intégrale égale à l'unité. Nous disons alors que Ψ est une "fonction d'état de champ normée":

$$\int_{-\infty}^{+\infty} |\Psi(x, y, z)|^2 dx dy dz = 1 \quad (42.206)$$

Notons qu'il règne encore un arbitraire sur Ψ par un nombre complexe de module 1, $e^{i\varphi}$, sans que la condition de normalisation soit altérée. Nous appelons cela "l'arbitraire de phase" et en verrons un exemple plus tard.

Un tel état dynamique est dit "état lié" ou "niveau lié", parce que la particule se manifeste dans une région limitée de l'espace à cause d'un potentiel. Lorsque, par exemple, l'atome d'hydrogène est situé sur un niveau fondamental, il est dans un état lié. Nous savons qu'il n'y a aucune chance de trouver l'électron à plus de quelques angströms du proton, traité comme infiniment lourd et placé à l'origine comme nous l'avons vu lors de l'étude du modèle de Bohr. Voici une bonne vision schématique de la chose (état lié):

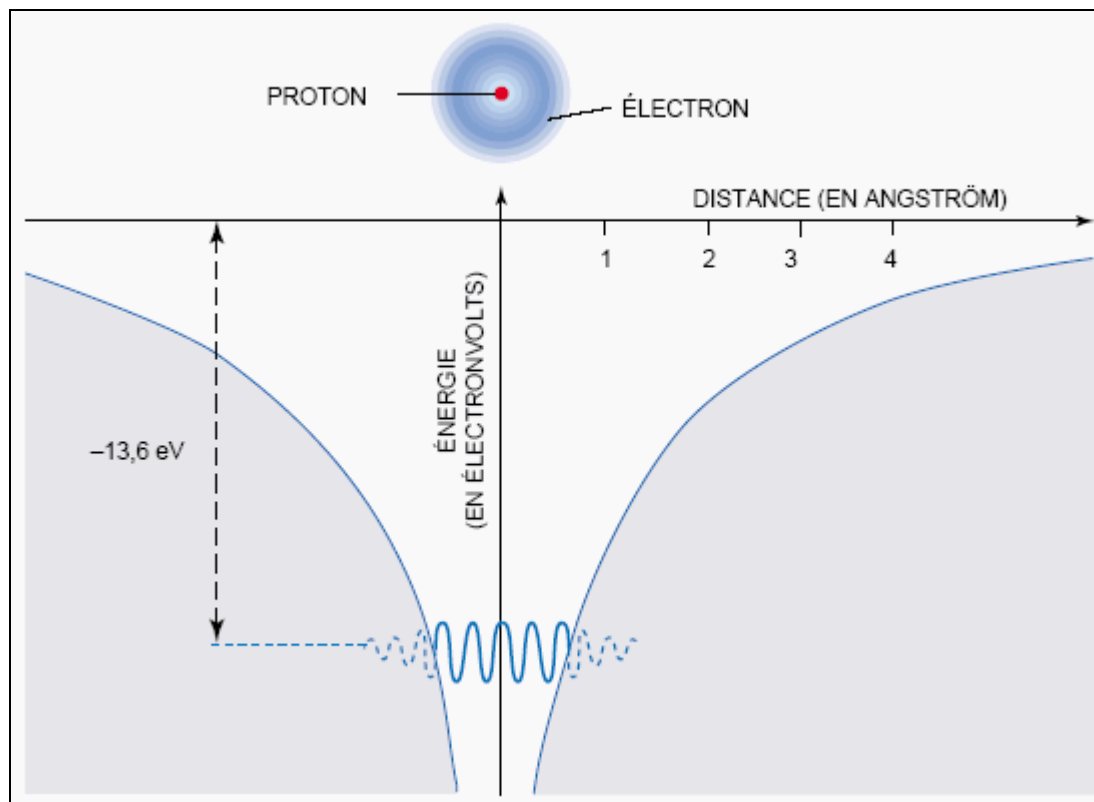


Figure: 42.3 - Représentation imagée d'un état lié (source: Pour la Science)

Un exemple d'état par défaut non lié est la particule libre qui peut se propager indéfiniment dans toutes les directions de l'espace (au fait pour ce dernier exemple c'est un peu plus compliqué... mais nous le traiterons plus loin).

Remarque: Il est bon de noter que ces concepts d'états liés ont des analogues classiques. Ainsi, les niveaux liés de l'atome d'hydrogène correspondent aux orbites elliptiques, les niveaux non liés (énergie positive) correspondent aux orbites hyperboliques.

4.3. ÉQUATION D'ÉVOLUTION CLASSIQUE DE SCHRÖDINGER

Nous savons qu'en mécanique classique l'état dynamique d'un système évolue, en général, dans le temps. Cela veut dire que la position et la quantité de mouvement (par exemple) sont fonction du temps. Pour un système d'hamiltonien donné, la connaissance de l'état dynamique initial permet de prévoir exactement l'évolution ultérieure de ce système du fait des propriétés bien connues des équations de Hamilton.

En physique quantique, les états dynamiques évolueront, en général, dans le temps. La fonction d'onde décrivant un état dynamique ne sera alors pas seulement fonction des coordonnées des particules constituant le système, mais elle dépendra donc aussi du temps et s'écrira:

$$\Psi(x, y, z, t) \quad (42.207)$$

Il est tout naturel d'admettre, ne serait-ce que par analogie avec la mécanique classique, que pour un système donné, d'hamiltonien connu, la connaissance de l'état dynamique initial à l'instant t_0 , permet de prévoir quel sera l'état dynamique du système à un instant ultérieur $t > t_0$.

Notons en passant que cela revient à dire qu'un ensemble initialement "pur" reste un ensemble pur au cours de l'évolution ultérieure des systèmes qui le constituent sans action extérieure. Cela cesserait donc d'être vrai si tous les systèmes de l'ensemble n'avaient pas exactement le même hamiltonien.

Indiquons qu'il existe deux approches possibles pour déterminer les fonctions dépendantes du temps:

- La première, courante dans de nombreux domaines d'application de la physique quantique, consiste à utiliser un "opérateur d'évolution" et permet de faire apparaître de manière explicite l'équation d'évolution de Schrödinger. Nous commencerons par celle-ci même si c'est la plus compliquée ou abstraite.
- La deuxième, très utilisée à des fins pédagogiques, permet d'obtenir les fonctions dépendantes du temps par l'intermédiaire de la technique de séparation des variables des équations différentielles mais nécessite d'admettre l'équation d'évolution de Schrödinger comme un postulat.

4.3.1. OPÉRATEUR D'ÉVOLUTION

Soit $\Psi(t)$ la fonction d'onde normée décrivant l'état dynamique du système à l'instant t (nous n'écrivons pas les autres variables dont dépend Ψ par souci de simplification, à savoir les coordonnées spatiales des particules du système). D'après ce qui précède, si $\Psi(t_0)$ est connue, $\Psi(t)$ l'est aussi. Nous avons une correspondance:

$$\Psi(t) \rightarrow \Psi(t_0) \quad (42.208)$$

et nous admettrons qu'elle est linéaire! Il existe donc un opérateur $T(t, t_0)$, appelé "**opérateur d'évolution**", tel que:

$$\Psi(t) = T(t, t_0)\Psi(t_0) \quad (42.209)$$

La fonction $\Psi(t + dt)$ dépend linéairement de $\Psi(t)$. Il en est alors de même de:

$$\frac{\Psi(t + dt) - \Psi(t)}{dt} \quad (42.210)$$

Il existe donc un opérateur linéaire K , tel que:

$$\frac{\partial \Psi}{\partial t} = iK\Psi \quad (42.211)$$

le nombre complexe i venant simplement du fait que nous devinons intuitivement que le résultat sera une fonction d'onde complexe.

Ce qui a aussi amené les physiciens à poser cette dernière égalité ainsi étaient les résultats connus de l'équation d'onde décrivant un état dynamique d'après l'idée de De Broglie. Nous allons donc tout de suite montrer que poser l'égalité ainsi est justifié.

Nous devons déterminer K puisque la connaissance de l'hamiltonien H commande l'évolution du système, K doit donc dépendre de H . Pour préciser la loi qui lie K à H , nous examinerons un cas particulier, celui de la particule libre (dont nous ferons une étude détaillée plus loin). Dans ce cas, H s'identifie à l'énergie cinétique uniquement.

D'après les idées de De Broglie, il est naturel d'admettre que la fonction d'onde décrivant un état dynamique dans lequel la quantité de mouvement est bien déterminée, soit $\hbar k$ (relation démontrée pendant l'étude de la particule libre), et où l'énergie totale est donc également bien déterminée, soit:

$$E = \hbar\omega \quad (42.212)$$

est une onde plane de la forme classique:

$$\Psi(x, y, z, t) = e^{i(\mathbf{k} \cdot \mathbf{x} + \omega t)} \quad (42.213)$$

où $\mathbf{k}$ est le vecteur d'onde de l'onde et $\mathbf{x} = (x, y, z)$ ses coordonnées spatiales.

Nous voyons alors très bien à l'arbitraire de phase près (pris comme étant négatif) que:

$$\frac{\partial \Psi}{\partial t} = -i\omega\Psi = \frac{E}{\hbar i}\Psi \quad (42.214)$$

Mais nous avons la relation entre opérateur hamiltonien et valeur propre suivante:

$$H\Psi = E\Psi \quad (42.215)$$

Les deux équations précédentes conduisent à écrire:

$$\frac{\partial \Psi}{\partial t} = \frac{H}{\hbar i}\Psi \quad (42.216)$$

En comparant cette dernière relation avec:

$$\frac{\partial \Psi}{\partial t} = iK\Psi \quad (42.217)$$

nous sommes amenés à poser:

$$K = -\frac{H}{\hbar} \quad (42.218)$$

Les physiciens supposent que cette relation entre K et H est générale. Alors, l'équation:

$$\frac{\partial \Psi}{\partial t} = iK\Psi \quad (42.219)$$

dans laquelle K est remplacé par son expression:

$$K = -\frac{H}{\hbar} \quad (42.220)$$

devient alors:

$$\boxed{i\hbar \frac{\partial \Psi}{\partial t} = H\Psi} \quad (42.221)$$

Cette équation constitue "**l'équation d'évolution classique de Schrödinger**" qui permet d'étudier des systèmes évoluant avec le temps, notamment les processus impliquant l'absorption et l'émission de rayonnement et la diffusion du rayonnement par les atomes.

En particulier, pour une particule sans spin soumise à une énergie potentielle E_{pot} , en maintenant toujours que la relation entre K et H est générale, l'équation d'évolution s'écrit alors:

$$i\hbar \frac{\partial \Psi}{\partial t} = \left(E_{pot} - \frac{\hbar^2}{2m} \Delta \right) \Psi \quad (42.222)$$

où les termes entre parenthèses correspondent donc à l'expression de l'hamiltonien.

Il convient maintenant de résoudre l'équation différentielle d'évolution de Schrödinger. Pour cela, nous allons nous servir de la condition de normalisation de De Broglie.

Rappelons que cette condition s'écrit:

$$\int_{-\infty}^{+\infty} |\Psi(x, y, z)|^2 dx dy dz = 1 \quad (42.223)$$

et généralisons à une étude multidimensionnelle et temporelle de cette condition telle que (selon les propriétés des complexes):

$$\int |\Psi(t, q)|^2 dq = \int \bar{\Psi}(t, q) \Psi(t, q) dq = 1 \quad (42.224)$$

Cette intégrale sur tout le domaine de définition n'est certainement pas égale à l'unité si nous n'introduisons pas une fonction de normalisation assimilée à une observable que nous noterons X et telle que nous ayons bien sur tout le domaine de définition:

$$\int \bar{\Psi}(t, q) X \Psi(t, q) dq = 1 \quad (42.225)$$

D'après cette condition, cette intégrale doit nécessairement rester constante en fonction du temps et de fait égale à l'unité.

Calculons d'abord la dérivée par rapport au temps de l'intégrale de normalisation et X . Nous avons donc nécessairement:

$$\begin{aligned} \frac{d}{dt} \int \bar{\Psi}(t, q) \Psi(t, q) dq &= \int \left[\frac{d}{dt} \bar{\Psi}(t, q) \right] \Psi(t, q) dq + \int \bar{\Psi}(t, q) \left[\frac{d}{dt} \Psi(t, q) \right] dq \\ &= \int \dot{\bar{\Psi}}(t, q) \Psi(t, q) dq + \int \bar{\Psi}(t, q) \dot{\Psi}(t, q) dq = 0 \end{aligned} \quad (42.226)$$

et utilisons l'équation d'évolution de Schrödinger:

$$i\hbar \dot{\Psi} = H \Psi \Rightarrow \begin{cases} \dot{\Psi} = -\frac{i}{\hbar} H \Psi \\ \dot{\bar{\Psi}} = \frac{i}{\hbar} \bar{H} \bar{\Psi} \end{cases} \quad (42.227)$$

ce qui nous donne pour notre intégrale après substitution:

$$\begin{aligned} \int \left[-\frac{i}{\hbar} H \bar{\Psi}(t, q) \right] \Psi(t, q) dq + \int \bar{\Psi}(t, q) \left[\frac{i}{\hbar} \bar{H} \Psi(t, q) \right] dq &= 0 \\ \frac{i}{\hbar} \left[\int \bar{\Psi}(t, q) \bar{H} \Psi(t, q) dq - \int H \bar{\Psi} \Psi(t, q) dq \right] &= 0 \\ \int \bar{\Psi}(t, q) \bar{H} \Psi(t, q) dq - \int H \bar{\Psi}(t, q) \Psi(t, q) dq &= 0 \end{aligned} \quad (42.228)$$

Démontrons maintenant que nous pouvons écrire:

$$\int \bar{\Psi}(t, q) \bar{H} \Psi(t, q) dq - \int H \bar{\Psi}(t, q) \Psi(t, q) dq = \int \bar{\Psi}(t, q) \bar{H} \Psi(t, q) dq - \int \bar{\Psi}(t, q) H \Psi(t, q) dq \quad (42.229)$$

Cela revient à démontrer que H peut agir identiquement "en arrière" tel que:

$$\int H \bar{\Psi}(t, q) \Psi(t, q) dq = \int \bar{\Psi}(t, q) H \Psi(t, q) dq \quad (42.230)$$

H pouvant être (ou contenir si vous préférez) un opérateur (différentiel par exemple).

Cette relation est démontrable si et seulement si Ψ est une fonction décroissante vers l'infini et dont la dérivée tende vers zéro vers l'infini!

Prouvons cela sur un cas particulier (mais fréquent en physique) et pour voir comment cela peut se faire, considérons dans H , un terme particulier de la forme suivante :

$$H = f(q) \cdot \frac{\partial}{\partial q} \cdot g(q) = f \frac{\partial}{\partial q} g \quad (42.231)$$

ce qui nous conduit à écrire:

$$\int \bar{\Psi}(t,q) H \Psi(t,q) dq = \int \bar{\Psi}(t,q) \left[f \frac{\partial}{\partial q} g \Psi(t,q) \right] dq \quad (42.232)$$

Par intégration par parties (cf. chapitre de Calcul Différentiel Et Intégral) sur le terme l'opérateur différentiel:

$$\int_{-\infty}^{+\infty} \bar{\Psi}(t,q) f \cdot \frac{\partial}{\partial q} (g \Psi(t,q)) dq = \bar{\Psi}(t,q) f \cdot g \Psi(t,q) \Big|_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} \frac{\partial}{\partial q} (\bar{\Psi}(t,q) f) \cdot g \Psi(t,q) dq \quad (42.233)$$

Or, puisque Ψ est une fonction décroissante vers l'infini par hypothèse (physiquement nécessaire!), nous aurons le premier terme qui sera toujours nul. Il nous reste ainsi:

$$\int_{-\infty}^{+\infty} \bar{\Psi} f \cdot \frac{\partial}{\partial q} (g \Psi) dq = - \int_{-\infty}^{+\infty} \frac{\partial}{\partial q} (\bar{\Psi} f) \cdot g \Psi dq = \int_{-\infty}^{+\infty} \left(- \frac{\partial}{\partial q} f \bar{\Psi} \right) \cdot g \Psi dq \quad (42.234)$$

Donc cela ne fait aucune différence de considérer que l'opérateur différencie tout ce qui est à droite ou tout ce qui est à gauche, dans la mesure où il est bien entendu que ce dernier cas implique un changement de signe. Il est d'usage d'appeler ce résultat parfois la "**condition d'hermité**"....

Donc nous pouvons bien nous permettre d'écrire:

$$\int \bar{\Psi}(t,q) \bar{H} \Psi(t,q) dq - \int \bar{\Psi}(t,q) H \Psi(t,q) dq = 0 \quad (42.235)$$

ce qui nous amène également à écrire:

$$\int \bar{\Psi}(t,q) (\bar{H} - H) \Psi(t,q) dq = 0 \quad (42.236)$$

Ceci ne peut être satisfait uniquement que si :

$$\bar{H} = H \quad (42.237)$$

et dans le domaine mathématique traitant des opérateurs nous avons vu que nous devons noter cette égalité:

$$H^\dagger = H \quad (42.238)$$

Ce qui nous amène à:

$$\int \bar{\Psi}(t,q) (H^\dagger - H) \Psi(t,q) dq = 0 \quad (42.239)$$

soit en utilisant la notation des représentatives (ket-bra):

$$\langle \Psi | H^\dagger - H | \Psi \rangle = 0 \quad (42.240)$$

Pour revenir à la résolution de:

$$i\hbar \frac{\partial \Psi}{\partial t} = H\Psi \quad (42.241)$$

il est évident qu'une solution possible est alors:

$$\Psi(t, q) = \Psi(0, q) e^{-\frac{it}{\hbar} H} \quad (42.242)$$

qui est donc constituée d'une partie purement spatiale (indépendante du temps) et d'une exponentielle complexe dépendante du temps. Vérifions:

$$\begin{aligned} i\hbar \frac{\partial}{\partial t} \Psi(t, q) &= i\hbar \frac{\partial}{\partial t} \left(\Psi(0, q) e^{-\frac{it}{\hbar} H} \right) = i\hbar \left(\frac{\partial \Psi(0, q)}{\partial t} e^{-\frac{it}{\hbar} H} + \Psi(0, q) \frac{\partial}{\partial t} e^{-\frac{it}{\hbar} H} \right) \\ &= i\hbar \Psi(0, q) \left(-\frac{i}{\hbar} \right) H e^{-\frac{it}{\hbar} H} = \Psi(0, q) H e^{-\frac{it}{\hbar} H} = H \Psi(t, q) \end{aligned} \quad (42.243)$$

C'est ce qu'il fallait démontrer (...).

Remarquons également qu'une fois les solutions purement spatiales déterminées, les solutions dépendantes du temps et de l'espace s'obtiennent aisément.

De même, grâce à la relation $H^\dagger = H$ que nous avons démontrée avant, nous pouvons écrire:

$$\bar{\Psi}(t, q) = \bar{\Psi}(0, q) e^{\frac{it}{\hbar} H} \quad (42.244)$$

Finalement, la relation:

$$\int \bar{\Psi}(t, q) X \Psi(t, q) dq = 1 \quad (42.245)$$

devient:

$$\int \bar{\Psi}(0, q) e^{-\frac{it}{\hbar} H} X \Psi(0, q) e^{\frac{it}{\hbar} H} dq = \int \bar{\Psi}(0, q) X(t) \Psi(0, q) dq = 1 \quad (42.246)$$

avec "l'opérateur de Heisenberg" défini par:

$$X(t) = e^{-\frac{it}{\hbar} H} X e^{\frac{it}{\hbar} H} = e^{\frac{it}{\hbar} H} X e^{-\frac{it}{\hbar} H} \quad (42.247)$$

Remarque: Il se peut très bien que X soit parfois une simple constante (nous en verrons un exemple plus bas).

4.3.2. SÉPARATION DES VARIABLES

Voyons également une manipulation mathématique intéressante et un peu similaire à la précédente de l'équation d'évolution de Schrödinger. Cette manipulation va nous permettre de voir que la séparation des variables fonctionne très bien avec l'équation d'évolution et qu'elle va nous permettre de retomber sur un résultat obtenu précédemment (c'est toujours bien pédagogiquement de voir plusieurs approches).

Nous avons donc dans un cas particulier:

$$i\hbar \frac{\partial \Psi}{\partial t} = \left(E_{pot} - \frac{\hbar^2}{2m} \Delta \right) \Psi \quad (42.248)$$

Réécrite sous forme traditionnelle (selon la littérature) et à une dimension, pour un potentiel constant dans le temps, cette relation s'écrit alors:

$$-\frac{\hbar}{i} \frac{\partial \Psi(x,t)}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \Psi(x,t)}{\partial x^2} + U(x) \Psi(x,t) \quad (42.249)$$

Supposons maintenant que la fonction d'onde puisse se séparer en deux fonctions dont elle est le produit telles que:

$$\Psi(x,t) = f(t) \Psi(x) \quad (42.250)$$

Nous aurions alors:

$$\frac{\partial \Psi(x,t)}{\partial t} = \frac{\partial f(t)}{\partial t} \Psi(x) \text{ et } \frac{\partial^2 \Psi(x,t)}{\partial x^2} = f(t) \frac{\partial^2 \Psi(x)}{\partial x^2} \quad (42.251)$$

Ce qui injecté dans l'équation d'évolution unidimensionnelle donne:

$$-\frac{\hbar}{i} \frac{\partial f(t)}{\partial t} \Psi(x) = -\frac{\hbar^2}{2m} \frac{\partial^2 \Psi(x)}{\partial x^2} f(t) + U(x) \Psi(x) f(t) \quad (42.252)$$

ce qui donne après simplification:

$$-\frac{\hbar}{i} \frac{1}{f(t)} \frac{\partial f(t)}{\partial t} = -\frac{\hbar^2}{2m} \frac{1}{\Psi(x)} \frac{\partial^2 \Psi(x)}{\partial x^2} + U(x) \quad (42.253)$$

Le terme de gauche ne dépend que de t , celui de droite que de x . Puisqu'ils sont égaux, ils sont nécessairement égaux aussi à une constante près qui a la dimension d'une énergie ($U(x)$ est une énergie potentielle pour rappel).

Donc pour le terme de gauche:

$$-\frac{\hbar}{i} \frac{1}{f(t)} \frac{\partial f(t)}{\partial t} = H \Rightarrow \ln(f(t)) = -\frac{iHt}{\hbar} + c^{te} \quad (42.254)$$

alors:

$$f(t) = e^{cte} e^{-i\frac{Ht}{\hbar}} = Ae^{-i\frac{Ht}{\hbar}} \quad (42.255)$$

et pour le terme de droite:

$$-\frac{\hbar^2}{2m} \frac{1}{\Psi(x)} \frac{\partial^2 \Psi(x)}{\partial^2 x} + U(x) \quad (42.256)$$

qui peut s'écrire:

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \Psi(x)}{\partial^2 x} + U(x) \Psi(x) = H \Psi(x) \quad (42.257)$$

après factorisation:

$$\frac{\partial^2 \Psi(x)}{\partial^2 x} + \frac{2m}{\hbar^2} [H - U(x)] \Psi(x) = 0 \quad (42.258)$$

Soit avec les notations du site:

$$\left(E_{pot} - \frac{\hbar^2}{2m} \frac{\partial^2}{\partial^2 x} \right) \Psi = E_{tot} \Psi \quad (42.259)$$

nous retrouvons donc l'équation de Schrödinger classique unidimensionnelle ce qui est pas mal du tout comme résultat!

Maintenant, puisque nous avons posé:

$$\Psi(x, t) = f(t) \Psi(x) \quad (42.260)$$

alors nous avons finalement:

$$\Psi(x, t) = Ae^{-i\frac{Ht}{\hbar}} \Psi(x) \quad (42.261)$$

ce que nous pouvons écrire sous les notations des paragraphes précédents:

$$\Psi(t, q) = \Psi(0, q) e^{-i\frac{Ht}{\hbar}} \quad (42.262)$$

Nous trouvons également cette dernière relation sous plusieurs formes différentes dans la littérature dont voici trois échantillons:

$$\begin{aligned} \Psi(x, t) &= \Psi(x) e^{-i\frac{Ht}{\hbar}} \\ \Psi(x, t) &= \varphi(x) e^{-i\frac{Ht}{\hbar}} \\ \Psi(x, t) &= \varphi(x) e^{-i\frac{Et}{\hbar}} \end{aligned} \quad (42.263)$$

4.3.3. COMBINAISON LINÉAIRE DES ÉTATS

Il faut remarquer avant que nous passions à un autre sujet quelque chose de très important que nous avons juste mentionné dans le deuxième postulat!

Effectivement, toute équation de la forme suivante vue précédemment:

$$\Psi(x)e^{-i\frac{Et}{\hbar}} \quad (42.264)$$

est donc solution de l'équation évolutive de Schrödinger et comme dans les systèmes quantiques l'hamiltonien peut prendre (ou être associé à) plusieurs valeurs propres discrètes notées traditionnellement E_n nous avons alors, comme mentionné au début de ce chapitre, par le principe de combinaison linéaire des équations différentielles la solution générale suivante:

$$\Psi(x,t) = \sum_{n=1}^{+\infty} \Psi_n(x) e^{-i\frac{E_n t}{\hbar}} \quad (42.265)$$

dont nous aurons plusieurs exemples pratiques (de la discrétisation des états d'énergie et que ceux-ci sont en nombre infini) dans le présent chapitre et celui de Chimie Quantique.

Si nous écrivons la constante de normalisation de $\Psi_n(x)$ de la relation précédente, nous avons alors:

$$\Psi(x,t) = \sum_{n=1}^{+\infty} A_n \Psi_n(x) e^{-i\frac{E_n t}{\hbar}} \quad (42.266)$$

Cette dernière relation s'écrirait sous la forme ket-bra traditionnelle suivante:

$$|\Psi\rangle = \sum_{n=1}^{+\infty} A_n \Psi_n(x) e^{-i\frac{E_n t}{\hbar}} = \sum_{n=1}^{+\infty} c_n |\psi_n\rangle \quad (42.267)$$

où le coefficient constant c_n est assimilé à A_n (avouez que c'est plus simple non?).

Nous disons alors que l'état $|\Psi\rangle$ est une combinaison linéaire d'états élémentaires. $|\Psi\rangle$ représente donc aussi une particule d'onde comme étant simultanément en plusieurs sous-états différents.

Il est intéressant de remarquer que chaque solution:

$$A_n \Psi_n(x) e^{-i\frac{E_n t}{\hbar}} \quad (42.268)$$

décrit un "état stationnaire". Voyons (enfin!) rigoureusement de quoi il s'agit.

En effet, nous avons:

$$\left| A_n \Psi_n(x) e^{-i\frac{E_n t}{\hbar}} \right|^2 = |A_n \Psi_n(x)|^2 \left| e^{-i\frac{E_n t}{\hbar}} \right|^2 = |A_n \Psi_n(x)|^2 \quad (42.269)$$

qui est donc indépendant du temps d'où l'origine du nom "état stationnaire" (nous avons promis d'en définir l'origine en début de chapitre... donc voilà qui est fait!).

Les fonctions étant normalisées nous avons donc:

$$\int |\Psi|^2 dx = \int \bar{\Psi} \Psi dx = \int \left(\sum_k \bar{A}_k \bar{\Psi}_k \right) \left(\sum_n A_n \Psi_n \right) dx = \sum_k \sum_n \bar{A}_k A_n \int \bar{\Psi}_k \Psi_n dx \quad (42.270)$$

Les calculs nous ont montré plus haut (nous avons fait la démonstration de deux manières différentes) que les fonctions propres ont les propriétés suivantes:

$$\int \bar{\Psi}_k \Psi_n dx = 1 \quad (42.271)$$

quand $k = n$ et:

$$\int \bar{\Psi}_k \Psi_n dx = 0 \quad (42.272)$$

quand $k \neq n$. C'est cette propriété qui nous avait amené dans le troisième postulat à parler de "**base orthogonale des fonctions propres stationnaires**".

Continuons notre calcul qui peut s'écrire en utilisant le symbole de Kronecker (cf. chapitre de Calcul Tensoriel):

$$\int |\Psi|^2 dx = \sum_k \sum_n \bar{A}_k A_n \int \bar{\Psi}_k \Psi_n dx = \sum_k \sum_n \bar{A}_k A_n \delta_{kn} = \sum_k \bar{A}_k A_k = \sum_k |A_k|^2 = 1 \quad (42.273)$$

Nous pouvons alors interpréter le terme $|A_k|^2$ comme le poids de la fonction propre Ψ_k dans l'état quantique k , la probabilité d'être en fait dans l'état propre Ψ_k vaut alors $|A_k|^2$ et la normalisation impose alors:

$$\sum_k |A_k|^2 = 1 \quad (42.274)$$

Retenons donc qu'un état quantique quelconque peut toujours être interprété comme étant une combinaison linéaire d'états propres. Le coefficient A_k d'une fonction/état propre Ψ_k est alors associé à une probabilité $|A_k|^2$.

C'est ce résultat mathématique, super important!, qui est à l'origine du paradoxe du chat de Schrödinger (parmi d'autres...) et de nombreux débats.

Pour clore ce petit sujet, remarquons une chose:

Si les coefficients A_n ne sont pas les coefficients déjà normalisés, mais non normalisés, les physiciens notent alors leur normalisation ainsi:

$$|A_n|^2 = \frac{|A_n|^2}{\sum_n |A_n|^2} \quad (42.275)$$

car très souvent ils utilisent la même notation pour le coefficient normalisé et le non normalisé dans leurs développements...

L'écriture de la dernière relation se justifie aisément car rappelons que nous devons avoir:

$$\sum_n |A_n|^2 = 1 \quad (42.276)$$

et nous avons effectivement après réarrangement:

$$|A_n|^2 = \frac{|A_n|^2}{\sum_n |A_n|^2} \Rightarrow \sum_n |A_n|^2 = \frac{|A_n|^2}{|A_n|^2} = 1 \quad (42.277)$$

Notons enfin qu'avec la notation ket-bra traditionnelle, la relation:

$$|A_n|^2 = \frac{|A_n|^2}{\sum_n |A_n|^2} \quad (42.278)$$

se note souvent dans certains ouvrages spécialisés:

$$p(x_n) = \frac{|c_n|^2}{\sum_n |c_n|^2} \quad (42.279)$$

qui donne donc toujours la probabilité de trouver l'état n à la position x .

4.3.4. ÉQUATION DE CONTINUITÉ

Considérons maintenant l'exemple important de l'équation d'évolution pour une particule libre, c'est-à-dire avec $E_{pot} = 0$. Nous avons donc:

$$-\frac{\hbar}{i} \frac{\partial \Psi}{\partial t} + \frac{\hbar^2}{2m} \Delta \Psi = 0 \quad (42.280)$$

La probabilité de trouver la particule dans un volume V est comme nous l'avons vu, donnée par:

$$\int_V \rho(\vec{r}, t) d\vec{r} = \int_V |\psi(\vec{r}, t)|^2 d\vec{r} \quad (42.281)$$

d'où:

$$\frac{\partial}{\partial t} \int_V \rho(\vec{r}, t) d\vec{r} = \int_V \left(\bar{\Psi}(\vec{r}, t) \frac{\partial \Psi(\vec{r}, t)}{\partial t} + \frac{\partial \bar{\Psi}(\vec{r}, t)}{\partial t} \Psi(\vec{r}, t) \right) d\vec{r} \quad (42.282)$$

En tenant compte de l'équation d'évolution de la particule libre, le second terme de l'égalité s'écrit:

$$\frac{\partial}{\partial t} \int_V \rho(\vec{r}, t) d\vec{r} = \frac{i\hbar}{2m} \int_V (\bar{\Psi} \nabla^2 \Psi - \nabla^2 \bar{\Psi} \Psi) d\vec{r} = - \int_V \vec{\nabla} \circ \vec{j} d\vec{r} \quad (42.283)$$

où nous avons posé:

$$\vec{j} = \frac{\hbar}{2mi} (\bar{\Psi} (\vec{\nabla} \cdot \Psi) - (\vec{\nabla} \cdot \bar{\Psi}) \Psi) \quad (42.284)$$

D'après le théorème d'Ostrogradsky (cf. chapitre de Calcul Vectoriel), il vient donc:

$$\frac{\partial}{\partial t} \int_V \rho(\vec{r}, t) d\vec{r} = - \int_V \vec{j} \circ d\vec{S} \quad (42.285)$$

où l'intégrale de droite est effectuée sur la surface S qui limite le volume V . La relation précédente exprime donc bien que la variation par unité de temps de la probabilité de trouver la particule dans V est égale au flux traversant la surface S et le vecteur $\vec{j}$ peut être interprété comme une densité de courant de probabilité qui satisfait l'équation de continuité telle que nous l'avons déterminée en thermodynamique:

$$\frac{\partial}{\partial t} \int_V \rho(\vec{r}, t) d\vec{r} = - \int_V \vec{\nabla} \circ \vec{j} d\vec{r} \Leftrightarrow \int_V \left[\frac{\partial}{\partial t} \rho(\vec{r}, t) \right] d\vec{r} = \int_V [-\vec{\nabla} \circ \vec{j}] d\vec{r} \quad (42.286)$$

d'où:

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \circ \vec{j} = 0 \quad (42.287)$$

En physique quantique, il y aurait donc conservation du flux de particules: Il n'y a ni création ni disparition de particule, alors que dans la nature (les observations expérimentales) nous observons pourtant de tels phénomènes... il y a donc contradiction entre l'expérience et la théorie ce qui invalide nos développements.

Par contre, cette équation exprime la conservation de la probabilité aussi! Donc de la propriété d'existence de la particule et des caractéristiques qu'elle transporte. Par exemple, si nous multiplions cette dernière relation par la charge de la particule, nous exprimons alors la continuité du courant.

4.4. IMPLICATIONS ET APPLICATIONS

Les différentes définitions et outils qui ont été vus précédemment, vont nous permettre d'étudier certains cas fondamentaux qui débouchent sur des résultats splendides.

Dans un premier temps, nous allons voir comment traiter le cas de la particule libre (état non lié) et quels sont les problèmes que pose cette configuration simple.

Ensuite, nous allons résoudre l'équation de Schrödinger avec une particule sans spin dans un puits de potentiel à parois rectilignes et montrer que nous retrouverons avec le formalisme de la physique quantique les mêmes résultats que le modèle de Bohr (plus généralisé même!).

Après quoi, nous allons introduire l'étude de l'oscillateur harmonique en repassant au préalable brièvement sur la résolution de l'équation de Schrödinger d'une particule libre. Cet exemple constitue une forme d'introduction quantique à l'étude théorique de systèmes atomiques. C'est dans cet exemple, que nous utiliserons toute la puissance des opérateurs linéaires fonctionnels. Il sera donc important de ne pas brûler les étapes lors de sa lecture.

Il nous faudra également étudier un autre phénomène fameux, l'effet tunnel! Évidemment, nous avons décidé de faire une introduction d'un cas particulier afin que le lecteur puisse voir le raisonnement qui a amené à la découverte de ce phénomène épatant (mais logique). Encore une fois, cet exemple appuiera la validité de la théorie quantique et démontrant la valeur des constantes de désintégration des isotopes nucléaires!

En ce qui concerne les cas relativistes, avec ou sans spin nous renvoyons le lecteur au chapitre de Physique Quantique Relativiste et en ce qui concerne le modèle atomique simple, nous le renvoyons au chapitre de Chimie Quantique.

Enjoy!

4.4.1. PARTICULE LIBRE

Curieusement la résolution de l'équation de Schrödinger pour une particule libre (où le potentiel est nul) est le cas simple... le plus complexe... mathématiquement parlant car les bornes d'intégration de la normalisation sont infinies.

Voyons cela:

Rappelons d'abord que nous avons démontré de manière simplifiée dans le chapitre de Suites et Séries que la transformée de Fourier d'une fonction f et son inverse étaient données par:

$$\mathcal{F}(\vec{k}) = \frac{1}{(2\pi)^{3/2}} \iiint f(\vec{r}) e^{-i\vec{k} \cdot \vec{r}} d^3r \quad f(\vec{r}) = \frac{1}{(2\pi)^{3/2}} \iiint \mathcal{F}(\vec{k}) e^{i\vec{k} \cdot \vec{r}} d^3k \quad (42.288)$$

Soit sous forme unidimensionnelle:

$$\mathcal{F}(k) = \frac{1}{\sqrt{2\pi}} \int f(x) e^{-ikx} dx \quad f(x) = \frac{1}{\sqrt{2\pi}} \int \mathcal{F}(k) e^{ikx} dk \quad (42.289)$$

Procédons maintenant au changement de variable qui relie le nombre d'onde k à la quantité de mouvement (relation introduite au début de ce chapitre):

$$k = \frac{p}{\hbar} \quad (42.290)$$

Ce qui nous donne:

$$\mathcal{F}(k) = \frac{1}{\sqrt{2\pi}} \int f(x) e^{-i\frac{p}{\hbar}x} dx \quad f(x) = \frac{1}{\sqrt{2\pi}} \int F(k) e^{i\frac{p}{\hbar}x} dp \quad (42.291)$$

Revenons maintenant à l'équation de Schrödinger d'évolution:

$$i\hbar \frac{\partial \Psi}{\partial t} = \left(E_{pot} - \frac{\hbar^2}{2m} \Delta \right) \Psi \quad (42.292)$$

Si la particule est libre, il n'y a pas de potentiel et à une dimension nous avons alors:

$$i \frac{\partial \Psi}{\partial t} = - \frac{\hbar}{2m} \frac{\partial^2}{\partial x^2} \Psi \quad (42.293)$$

Cette équation différentielle admet des solutions en ondes planes monochromatiques du type (cf. chapitre d'Électrodynamique):

$$\Psi(x,t) = A e^{i(kx - \omega t)} \quad (42.294)$$

avec bien évidemment la petite nuance que nous avons à utiliser la relation (sinon ça ne joue pas par contre!):

$$\omega = \frac{\hbar k^2}{2m} = \frac{p^2}{2m\hbar} \quad (42.295)$$

Sans oublier que (cela nous sera utile par la suite):

$$E = \frac{p^2}{2m} = \frac{\hbar^2 k^2}{2m} \quad (42.296)$$

La courbe de l'énergie E en fonction du vecteur d'onde k est parfois appelée "**courbe de dispersion**" et c'est une parabole (puisque k est au carré) pour une particule libre!

Bien évidemment, la densité de probabilité de cette solution vaut:

$$\langle \Psi | \Psi \rangle = \bar{\Psi} \Psi = A e^{i(kx - \omega t)} \bar{A} e^{-i(kx - \omega t)} = |A|^2 \quad (42.297)$$

mais cela ne peut pas correspondre à la réalité car nous ne pouvons pas normaliser la probabilité sur des distances infinies! Une onde plane monochromatique de module constant dans tout l'espace n'étant pas de carré sommable: elle ne peut donc pas représenter un état physique d'une particule libre.

Au fait la solution vient du fait que la vraie solution utilise le principe de superposition de toutes les ondes monochromatiques de toutes les fréquences tel que:

$$\Psi(x,t) = \int_{-\infty}^{+\infty} A(k) e^{i(kx - \omega(k)t)} dk \quad (42.298)$$

et nous retrouvons donc ici une relation très similaire à une transformée de Fourier inverse (cf. chapitre Suites Et Séries). Une telle superposition d'ondes planes est appelée: "**paquet d'ondes unidimensionnel**".

Ce que nous pouvons réécrire:

$$\Psi(x,t) = \int_{-\infty}^{+\infty} A(k) e^{ikx} e^{-i\frac{\hbar k^2}{2m}t} dk = \int_{-\infty}^{+\infty} A(p) e^{i\frac{p}{\hbar}x} e^{-i\frac{p^2}{2m\hbar}t} dp = \int_{-\infty}^{+\infty} A(p) e^{i\frac{p}{\hbar}x} e^{-i\frac{E}{\hbar}t} dp \quad (42.299)$$

Or, nous voyons de suite que nous ne pourrions pas non plus normaliser suivant:

$$\int_{-\infty}^{+\infty} |\Psi|^2 dx = \int_{-\infty}^{+\infty} \bar{\Psi}(x,t) \Psi(x,t) dx \quad (42.300)$$

Dès lors, il n'y a plus de solution générale. Il faut donner une enveloppe porteuse aux ondes imposant une normalisation possible. Cette enveloppe porteuse peut être un Dirac ou une Gaussienne ou d'autres fonctions de distributions plus ou moins complexes. Ensuite les physiciens doivent utiliser une propriété des transformées de Fourier qui font naturellement apparaître les incertitudes de Heisenberg. Ainsi, ces dernières sont une condition à la normalisation des particules libres utilisant les transformées de Fourier.

À ce jour, nous n'avons pas de démonstration pédagogique et simple à proposer sur ce dernier point. Cela viendra peut-être plus tard.

Par contre, nous pouvons prendre comme solution triviale les modes propres de la particule tel que:

$$\Psi_x = e^{\pm i(kx - \omega t)} \quad (42.301)$$

Effectivement:

$$\begin{aligned} i \frac{\partial \Psi}{\partial t} &= -\frac{\hbar}{2m} \frac{\partial^2}{\partial x^2} \Psi \Rightarrow i \frac{\partial e^{i(kx - \omega t)}}{\partial t} = -\frac{\hbar}{2m} \frac{\partial^2 e^{i(kx - \omega t)}}{\partial x^2} \\ \Rightarrow -i^2 \omega e^{i(kx - \omega t)} &= -\frac{\hbar}{2m} i^2 k^2 e^{i(kx - \omega t)} \Rightarrow -\omega = -\frac{\hbar}{2m} k^2 \quad (42.302) \\ \Rightarrow \omega &= \frac{\hbar}{2m} k^2 = \frac{p^2}{2m\hbar} \end{aligned}$$

C'est ce que nous utiliserons comme situation lors de notre étude plus bas de l'oscillateur harmonique.

Avant d'étudier le cas particulier du paquet d'ondes quasi-monochromatiques, nous allons rappeler quelques résultats concernant la somme de deux ondes planes.

Commençons par sommer deux ondes planes monochromatiques de fréquences voisines:

$$\Psi_1(x,t) = A e^{i(k_1 x - \omega_1 t)} \text{ et } \Psi_2(x,t) = A e^{i(k_2 x - \omega_2 t)} \quad (42.303)$$

avec:

$$\omega_1 = \omega_0 - \Delta \omega \text{ et } \omega_2 = \omega_0 + \Delta \omega \quad (42.304)$$

et:

$$k_1 = k_0 - \Delta k \text{ et } k_2 = k_0 + \Delta k \quad (42.305)$$

À noter que nous imposons donc:

$$\frac{\Delta\omega}{\omega_0} \ll 1 \quad \text{et} \quad \frac{\Delta k}{k_0} \ll 1 \quad (42.306)$$

L'onde résultante a pour expression:

$$\begin{aligned} \Psi(x,t) &= \Psi_1(x,t) + \Psi_2(x,t) = A \left(e^{i(k_1 x - \omega_1 t)} + e^{i(k_2 x - \omega_2 t)} \right) \\ &= A \left(\cos(k_1 x - \omega_1 t) + i \sin(k_1 x - \omega_1 t) + \cos(k_2 x - \omega_2 t) + i \sin(k_2 x - \omega_2 t) \right) \\ (42.307) \quad &= A \left(\begin{aligned} &\cos((k_0 - \Delta k)x - (\omega_0 - \Delta\omega)t) + i \sin((k_0 - \Delta k)x - (\omega_0 - \Delta\omega)t) + \\ &\cos((k_0 + \Delta k)x - (\omega_0 + \Delta\omega)t) + i \sin((k_0 + \Delta k)x - (\omega_0 + \Delta\omega)t) \end{aligned} \right) \\ &= A \left(\begin{aligned} &\cos((k_0 - \Delta k)x - (\omega_0 - \Delta\omega)t) + \cos((k_0 + \Delta k)x - (\omega_0 + \Delta\omega)t) + \\ &i \sin((k_0 - \Delta k)x - (\omega_0 - \Delta\omega)t) + i \sin((k_0 + \Delta k)x - (\omega_0 + \Delta\omega)t) \end{aligned} \right) \end{aligned}$$

Soit en utilisant les relations trigonométriques remarquables (cf. chapitre de Trigonométrie):

$$\begin{aligned} \Psi(x,t) &= A \left(2 \cos(k_0 x - \omega_0 t) \cos(\Delta k x - \Delta\omega t) + 2i \sin(k_0 x - \omega_0 t) \cos(\Delta k x - \Delta\omega t) \right) \\ &= 2A \cos(\Delta k x - \Delta\omega t) \left(\cos(k_0 x - \omega_0 t) + i \sin(k_0 x - \omega_0 t) \right) \quad (42.308) \\ &= 2A \cos(\Delta k x - \Delta\omega t) e^{i(k_0 x - \omega_0 t)} = \Psi_0 e^{i(k_0 x - \omega_0 t)} \end{aligned}$$

qui est une onde plane se propageant selon x avec la pulsation ω_0 et le vecteur d'onde moyen k_0 , et donc à la vitesse de phase:

$$v_\varphi = \frac{\omega_0}{k_0} = \frac{\frac{\hbar k_0^2}{2m}}{k_0} = \frac{\hbar k_0}{2m} = \frac{\hbar \frac{\sqrt{E 2m}}{\hbar}}{2m} = \frac{\sqrt{E 2m}}{2m} = \sqrt{\frac{E}{2m}} \quad (42.309)$$

Le terme en cosinus s'interprète alors comme l'amplitude lentement variable de cette onde plane.

Remarquons un point assez important!: La vitesse de phase n'est pas conforme à la vitesse que nous obtenons en utilisant l'énergie cinétique d'une particule libre. Effectivement:

$$E_c = \frac{1}{2} m v^2 \Leftrightarrow v = \sqrt{\frac{2E_c}{m}} \quad (42.310)$$

Dès lors la vitesse de phase ne représente pas la vitesse dans le sens classique habituel mais se l'onde se déplaçant à la vitesse de groupe:

$$v_g = \frac{\Delta\omega}{\Delta k} \underset{\Delta \rightarrow 0}{=} \frac{d\omega}{dk} = \frac{d}{dk} \left(\frac{\hbar k^2}{2m} \right) = \frac{\hbar 2k}{2m} = \frac{\hbar k}{m} = \frac{\hbar \frac{\sqrt{E 2m}}{\hbar}}{m} = \sqrt{\frac{2E}{m}} \quad (42.311)$$

où nous retrouvons donc la formulation classique de la vitesse à partir de l'énergie cinétique (pas mal...)!

Nous pouvons représenter aisément tout cela avec Maple 4.00b:

```
>restart:with(plots):
>lambda[0]:=1; T[0]:=1; k[0]:=2*Pi/lambda[0]; w[0]:=2*Pi/T[0];
> delta_k:=k[0]/8; k[1]:=k[0]-delta_k; k[2]:=k[0]+delta_k;
delta_w:=w[0]/10; w[1]:=w[0]-delta_w; w[2]:=w[0]+delta_w;
> P1:=animate(cos(k[1]*x-w[1]*t)+cos(k[2]*x-w[2]*t), x=0..1*2*Pi/delta_k, t=0..2*Pi/delta_w,
numpoints=200, frames=15, color=red):
> P2:=animate({2*cos(-1/2*k[1]*x+1/2*w[1]*t+1/2*k[2]*x-1/2*w[2]*t), -2*cos(-1/2*k[1]*x+1/2*w[1]*t
+1/2*k[2]*x-1/2*w[2]*t)}, x=0..1*2*Pi/delta_k, t=0..2*Pi/delta_w, numpoints=100, frames=15,
color=blue):
> display(P1,P2);
```

Ce qui donne:

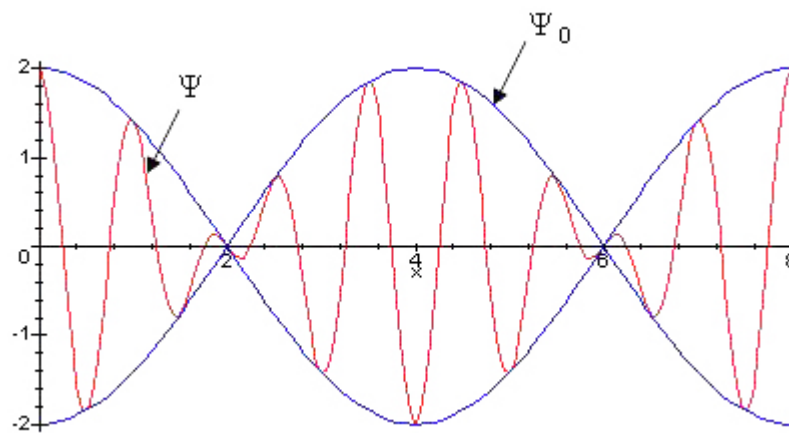


Figure: 42.4 - Représentation du concept de vitesse de phase et vitesse de groupe

À la différence de l'onde plane harmonique, cette onde n'a pas un module constant: son module est nul dans certaines zones. Par contre, elle s'étend toujours sur une distance infinie, donc a une norme (somme de la probabilité sur tout l'espace) infinie. Elle ne possède donc pas de sens physique.

L'étude précédente peut être étendue en sommant un nombre N de plus en plus grand d'ondes planes au voisinage de k_0 et ω_0 . Une telle superposition conduit à une fonction de plus en plus localisée dans certaines zones de l'espace (en particulier vers $x = 0$ par exemple pour $t = 0$), la distance entre ces zones augmentant proportionnellement avec N . A la limite $N \rightarrow +\infty$, alors seule la zone vers $x = 0$ demeure, les autres étant rejetées à l'infini. Le passage à cette limite $N \rightarrow +\infty$ s'effectue en remplaçant la somme discrète sur les ondes planes par une sommation continue c'est-à-dire par une intégrale de la forme:

$$\Psi(x, t) = \int_{k_0 - \delta k/2}^{k_0 + \delta k/2} F(k) e^{i(kx - \omega t)} dk \quad (42.312)$$

avec:

$$k \in \left[k_0 - \frac{\delta k}{2}, k_0 + \frac{\delta k}{2} \right] \quad (42.313)$$

avec donc:

$$F(k) = A \quad \text{et} \quad \frac{\delta k}{k_0} \ll 1 \quad (42.314)$$

Un tel paquet est donc appelé "**paquet d'ondes quasi-monochromatiques**".

Cette expression peut se réécrire:

$$\Psi(x, t) = A \int_{k_0 - \delta k/2}^{k_0 + \delta k/2} e^{i(kx - \omega t)} dk \quad (42.315)$$

Il importe de comprendre que ω est une fonction de k , donnée par l'équation de dispersion. Nous allons faire le calcul de cette expression en utilisant le fait que $\delta k / k_0 \ll 1$.

$\delta k / k_0 \ll 1$ implique que $\delta \omega / \omega_0 \ll 1$. Il est possible d'effectuer un développement limité au voisinage de ω_0 :

$$\omega = \omega_0 + (k - k_0) \left(\frac{\partial \omega}{\partial k} \right)_{k=k_0} + \dots = \omega_0 + v_g (k - k_0) \quad (42.316)$$

où v_g est la vitesse de groupe. Alors:

$$kx - \omega t = (k - k_0 + k_0)x - [\omega_0 + v_g (k - k_0)]t = k_0 x - \omega_0 t + (x - v_g t)(k - k_0) \quad (42.317)$$

Posons $y = k - k_0$:

$$\Psi(x, t) = A e^{i(k_0 x - \omega_0 t)} \int_{-\delta k/2}^{+\delta k/2} e^{i(x - v_g t)y} dy \quad (42.318)$$

Calculons l'intégrale:

$$\Psi(x, t) = A e^{i(k_0 x - \omega_0 t)} \int_{-\delta k/2}^{+\delta k/2} e^{i(x - v_g t)y} dy = \frac{A e^{i(k_0 x - \omega_0 t)}}{i(x - v_g t)} e^{i\alpha y} \Bigg|_{-\delta k/2}^{+\delta k/2} \quad (42.319)$$

avec:

$$\alpha = x - v_g t \quad (42.320)$$

Soit:

$$\begin{aligned}
\Psi(x,t) &= \frac{A}{i(x-v_g t)} \left(e^{i\alpha \frac{\delta k}{2}} - e^{-i\alpha \frac{\delta k}{2}} \right) e^{i(k_0 x - \omega_0 t)} \\
&= \frac{A}{i(x-v_g t)} \left(\cos\left(\alpha \frac{\delta k}{2}\right) + i \sin\left(\alpha \frac{\delta k}{2}\right) - \cos\left(\alpha \frac{\delta k}{2}\right) + i \sin\left(\alpha \frac{\delta k}{2}\right) \right) e^{i(k_0 x - \omega_0 t)} \\
&= \frac{A}{i(x-v_g t)} \left(i \sin\left(\alpha \frac{\delta k}{2}\right) + i \sin\left(\alpha \frac{\delta k}{2}\right) \right) e^{i(k_0 x - \omega_0 t)} \\
&= 2A \frac{\sin\left((x-v_g t) \frac{\delta k}{2}\right)}{(x-v_g t)} e^{i(k_0 x - \omega_0 t)} = 2A \frac{\delta k}{2} \frac{\sin\left((x-v_g t) \frac{\delta k}{2}\right)}{(x-v_g t) \frac{\delta k}{2}} e^{i(k_0 x - \omega_0 t)} \quad (42.321) \\
&= A \delta k \frac{\sin\left((x-v_g t) \frac{\delta k}{2}\right)}{(x-v_g t) \frac{\delta k}{2}} e^{i(k_0 x - \omega_0 t)} = A \delta k \operatorname{sinc}\left((x-v_g t) \frac{\delta k}{2}\right) e^{i(k_0 x - \omega_0 t)} \\
&= \Psi_0 e^{i(k_0 x - \omega_0 t)}
\end{aligned}$$

Le dernier terme s'interprète à nouveau comme une onde plane se déplaçant à la vitesse de phase:

$$v_\varphi = \frac{\omega_0}{k_0} \quad (42.322)$$

L'amplitude de cette onde plane est donnée par une fonction de type sinus cardinal. A $t = 0$, cette fonction sinc n'a des valeurs importantes que dans la zone:

$$\delta x = \frac{2\pi}{\delta k} \quad (42.323)$$

Il s'agit donc d'une fonction bien localisée. En conséquence, Ψ est une fonction de carré sommable. Le calcul donne:

$$\int_{-\infty}^{+\infty} |\Psi|^2 dx = |A|^2 \delta k \quad (42.324)$$

La fonction peut donc être normalisée en posant donc:

$$A = \frac{1}{\sqrt{\delta k}} \quad (42.325)$$

Nous avons donc réussi à obtenir une fonction satisfaisant à la fois l'équation de Schrödinger et la condition de normalisation, grâce à l'emploi d'une somme infinie d'ondes harmoniques. L'exemple que nous avons traité n'est qu'un cas particulier. D'autres types de paquets d'ondes peuvent être obtenus en prenant d'autres distributions pour les amplitudes des ondes planes qui composent le paquet (nous avons supposé ici qu'elles avaient toutes la même amplitude). Dès lors, la vitesse de groupe est associée classiquement à la vitesse de la particule de masse m et d'impulsion p .

Ainsi, le paquet d'ondes se déplace globalement à la vitesse de groupe, qui s'identifie à la vitesse donnée par la mécanique classique.

Les relations d'incertitude ont déjà été introduites au début de ce chapitre de deux manières différentes. Mais dans l'exemple du paquet d'ondes étudié au paragraphe précédent, nous avons vu que la fonction est localisée dans une zone d'extension (largeur à mi-hauteur):

$$\delta x = \frac{2\pi}{\delta k} \quad (42.326)$$

Nous avons donc la relation:

$$\delta x \cdot \delta k = 2\pi \quad (42.327)$$

Nous retrouvons ici une expression de type incertitude. Le coefficient numérique pourrait être légèrement différent suivant la définition choisie pour δx et δk , ou le type de paquet. Il pourrait en particulier être nettement plus grand dans certains cas. Nous avons donc en fait une inégalité du type:

$$\delta x \cdot \delta k \geq 2\pi \quad (42.328)$$

En physique quantique, ces inégalités s'expriment en fonction de l'impulsion p , reliées à k par $p = \hbar k$. Nous avons donc:

$$\delta x \cdot \delta p_x \geq \hbar \quad (42.329)$$

Ainsi, plus la quantité de mouvement (implicitement la longueur d'onde) d'un paquet d'ondes est définie avec précision, moins il a de composantes et plus il est étalé, ce qui augmente d'autant l'incertitude relative à sa position et respectivement plus sa position est bien connue moins sa longueur d'onde le sera.

Il ne s'agit donc pas d'incertitudes au sens de la mesure, et qui seraient limitées par les appareils de mesure, mais d'une propriété fondamentale intrinsèque, liée à la représentation quantique d'une particule selon le modèle mathématique proposé. Le modèle de l'atome de Bohr est donc à rejeter pour les niveaux d'énergie qui sont proches de cette égalité.

4.4.2. Puits DE POTENTIEL A PAROIS RECTILIGNES

Prenons pour premier exemple, très important pour le chapitre de Physique Nucléaire et pour les spécialistes des semi-conducteurs, la résolution sous forme classique du puits de potentiel à parois rectilignes, également appelé "**puits rectangulaire**" (cet exemple est vraiment très important, prenez vraiment votre temps afin de le comprendre et de le maîtriser au mieux).

C'est l'exemple le plus simple d'une fonction $E_{pot}(x)$, nulle à l'intérieur du puits et infiniment grande sur les parois, distantes d'une longueur L .

Remarque: Lorsque $U = \infty$ nous disons que les parois sont parfaitement réfléchissantes.

Nous supposons une particule piégée dans ce puits. Elle ne peut s'en échapper puisque les parois (c'est-à-dire le potentiel U) ont une hauteur infinie. Mais à l'intérieur, elle est libre de se déplacer sans faire d'interaction avec les parois.

Cette configuration se traduit par les conditions aux limites où l'énergie potentielle électrostatique est notée U :

$$\begin{aligned} U &= 0 \text{ si } 0 \leq x \leq L \\ U &= \infty \text{ si } x < 0 \text{ ou } x > L \end{aligned} \quad (42.330)$$

Il existe deux manières d'aborder problème. Voyons les deux types de traitements car le premier permet d'avoir une approche simpliste alors que le deuxième permet d'avoir une approche plus générale qui nous sera utile par la suite lors de notre étude de l'effet Tunnel:

1ère approche

L'équation de Schrödinger (classique):

$$\frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} + \frac{\partial^2 \Psi}{\partial z^2} + \frac{8\pi^2 m}{h^2} \cdot (E_{tot} - E_{pot}) \Psi = 0 \quad (42.331)$$

a donc une solution simple respectant les conditions initiales en une dimension, du type:

$$\Psi = \Psi_0 \sin(kx) \quad (42.332)$$

dont la dérivée seconde est:

$$\frac{\partial^2 \Psi}{\partial x^2} = -k^2 \Psi \quad (42.333)$$

Introduits dans l'équation de Schrödinger, nous obtenons après quelques simplifications d'algèbre élémentaire:

$$k^2 = \frac{8\pi^2 m}{h^2} (E_{tot} - E_{pot}) = \frac{2m}{\hbar^2} (E_{tot} - E_{pot}) \quad (42.334)$$

Donc finalement la solution s'écrit:

$$\Psi = \Psi_0 \sin \left(\frac{\sqrt{2m(E_{tot} - E_{pot})}}{\hbar} \cdot x \right) \quad (42.335)$$

à propos de laquelle il faut appliquer les conditions aux limites (la solution en cosinus est en tout point similaire).

Si nous voulons pouvoir, par la suite, faire un parallèle avec un (ou des) électron(s) piégé(s) dans le puits du potentiel du noyau de l'atome (qui n'est pas rectangulaire lui!), nous sommes amenés aux considérations suivantes:

La stabilité des atomes suggère l'existence d'une onde stationnaire électronique dans le puits. De plus, l'observation montre que seuls certains niveaux d'énergie semblent autorisés dans ce dernier.

Si nous faisons une similitude avec les cordes vibrantes, la fonction d'onde de l'électron doit être telle que:

1. Pour $x = 0$ et $x = L$ il doit y avoir un noeud de vibration. Donc: $\Psi(0) = \Psi(L) = 0$
2. La fonction d'onde Ψ doit présenter un nombre entier de demi-longueur d'onde sur la longueur L
3. Dans la boîte $U = 0$ donc $E_{pot} = 0$
4. Si aux extrémités ($x = 0$ et $x = L$) $\Psi = 0$ alors l'argument du sinus vaut $n \cdot \pi \forall n \in \mathbb{N}^*$

Donc nous devons avoir:

$$x = L \Rightarrow \sqrt{\frac{8\pi^2 m}{h^2} (E_{tot} - E_{pot})} \cdot L = n \cdot \pi \quad (42.336)$$

d'où puisque l'énergie potentielle est nulle:

$$E_n = \frac{\pi^2 \hbar^2}{2mL^2 \pi^2 4} n^2 = \frac{\pi^2 \hbar^2}{2mL^2} n^2 \quad (42.337)$$

L'énergie totale de la particule s'échelonne donc sur une suite discrète de valeurs, les seules permises. La valeur de L est quant à elle déterminée à l'aide du modèle de Bohr ou de Sommerfeld en fonction des cas (cf. chapitre Physique Quantique Corpusculaire).

L'énergie totale de la particule ci-dessus est composée des "**valeurs propres**" de l'énergie dans le puits de potentiel.

Donc l'équation de Schrödinger permet de faire abstraction du 3ème postulat de Bohr dans le sens où elle explicite directement la notion de quantification des niveaux par des valeurs entières (discrètes) solutions des conditions aux limites d'un puits de potentiel considéré comme parfait.

Les fonctions d'onde correspondantes dans le puits où $U = 0 \equiv E_{pot} = 0$ sont donc:

$$\Psi = \Psi_0 \sin \left[\sqrt{\frac{8\pi^2 m}{h^2} E_{tot}} \cdot x \right] = \Psi_0 \sin \left[\sqrt{\frac{8\pi^2 m}{h^2} \frac{\hbar^2}{8 \cdot m \cdot L^2} n^2} \cdot x \right] \quad (42.338)$$

Soit après simplification:

$$\Psi_n = \Psi_0 \sin \left(\frac{n\pi}{L} \cdot x \right) \quad (42.339)$$

C'est l'expression d'une des solutions de l'équation pour le puits de potentiel rectangulaire idéal. Ainsi, il existe une suite discrète de fonctions d'onde solutions. Ce sont les "**fonctions propres**" de la particule.

La constante Ψ_0 dans cette expression est déterminée par la normalisation de De Broglie (dont nous avons parlé au début de ce chapitre), c'est-à-dire par la condition:

$$\int_0^L |\Psi_n|^2 dx = \Psi_0^2 \int_0^L \sin^2 \left(\frac{n\pi x}{L} \right) dx = 1 \quad (42.340)$$

Nous trouvons alors (calcul d'intégration normalement élémentaire):

$$\Psi_0 = \sqrt{\frac{2}{L}} \quad (42.341)$$

et l'expression finale de la fonction d'onde associée à la valeur propre E_n se lit donc:

$$\Psi_n(x) = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) \quad (42.342)$$

Certains physiciens ont pour habitude de noter cela sous forme complexe en ne considérant bien évidemment que la partie réelle de l'expression suivante (nous utilisons la "formule d'Euler" vue lors de l'introduction aux complexes dans le chapitre des Nombres):

$$\Psi_n(x) = \sqrt{\frac{2}{L}} e^{ik_n x} \quad (42.343)$$

avec:

$$k_n = \frac{n\pi}{L_n} \quad (42.344)$$

Nous disons alors que nous avons des "**conditions de quantification**" sur k imposées par les conditions aux limites.

Cette notation est parfois utile et nous l'utiliserons lors de l'étude de l'effet tunnel dans le chapitre de Physique Nucléaire.

Nous pouvons déduire de l'expression obtenue, les propriétés principales des fonctions d'onde décrivant les états stationnaires de la particule dans une boîte:

1. La figure ci-dessous représente des fonctions Ψ_n et des densités de probabilités $|\Psi_n|^2$ pour les premiers niveaux d'énergie E_n :

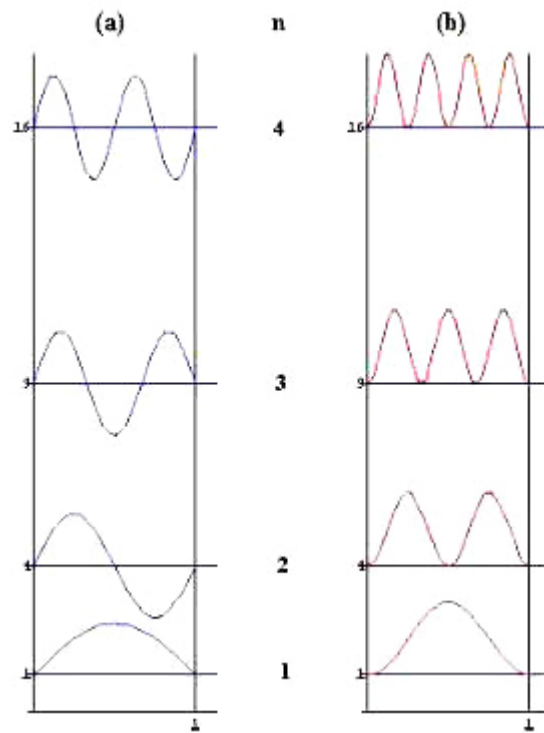


Figure: 42.5 - Représentation des fonctions d'onde et densité pour quelques niveaux d'énergie

Nous remarquons que (évidemment nous pourrions analyser ceci de façon analytique et non graphique si nous le désirions), en plus des points $x = 0$ et $x = L$, Ψ_n a $(n-1)$ zéros situés en:

$$x = x_m = mL/n \text{ avec } m = 1, 2, \dots, n-1 \quad (42.345)$$

Ces points, où la fonction d'onde et la densité de probabilité sont nulles, sont appelés "**points nodaux**" ou simplement "**noeuds**" de la fonction d'onde. Le nombre de noeuds augmente quand n augmente, c'est-à-dire quand l'on passe à des états de plus en plus excités. La fonction d'onde Ψ_1 de l'état fondamental à $n = 1$ et donc avec:

$$E = E_1 = \frac{h^2}{8mL^2} \quad (42.346)$$

n'a pas de noeud, celle du premier état excité Ψ_2 d'énergie:

$$E = E_2 = 4E_1 \quad (42.347)$$

a un point nodal, celle du deuxième état excité Ψ_3 a deux points nodaux, etc...

La variation des propriétés nodales des fonctions d'onde quand n varie traduit l'orthogonalité des états stationnaires d'énergie différente. En effet, nous vérifions aisément que $\langle \Psi_n | \Psi_m \rangle$ est nul quand $m \neq n$:

$$\begin{aligned}
\langle \Psi_n | \Psi_m \rangle &= \int_0^L \Psi_n(x) \Psi_m(x) dx = \frac{2}{L} \int_0^L \sin\left(\frac{n\pi x}{L}\right) \sin\left(\frac{m\pi x}{L}\right) dx \\
&= \frac{1}{L} \int_0^L \left\{ \cos\left[\frac{(n-m)\pi x}{L}\right] - \cos\left[\frac{(n+m)\pi x}{L}\right] \right\} dx \\
&= \frac{1}{L} \left\{ \frac{L}{(n-m)\pi} \sin\left[\frac{(n-m)\pi x}{L}\right] + \frac{L}{(n+m)\pi} \sin\left[\frac{(n+m)\pi x}{L}\right] \right\} \Bigg|_0^L = 0
\end{aligned} \tag{42.348}$$

où nous avons utilisé une des relations trigonométriques remarquables démontrées dans le chapitre de Trigonométrie.

2. Comme nous pouvons le voir sur la figure précédente, la densité de probabilité associée à tout état stationnaire de la particule est symétrique par rapport au point médian $x = L/2$

Nous anticipons donc que la valeur moyenne de x sera exactement égale à $L/2$ dans un tel état. En effet nous avons vu dans le chapitre de Statistiques que l'espérance (moyenne) d'un événement de probabilité $P(x)$ est définie par:

$$E(x) = \int_{-\infty}^{+\infty} x P(x) dx \tag{42.349}$$

où x , $E(x)$ et $P(x)$ n'ont pas d'unités (attention nous allons faire une analyse dimensionnelle).

Or, en physique quantique $E(x)$ et x sont des grandeurs dimensionnelles identiques. Ce qui signifie que les dimensions de $P(x)$ doivent annuler celles de dx . Ainsi, nous devinons suite à l'étude des conditions de normalisation de De Broglie que:

$$P(x) \equiv |\Psi_n(x)|^2 \tag{42.350}$$

est une probabilité linéique de présence de la particule.

Le domaine d'intégration étant $[0; L]$ nous avons finalement:

$$\begin{aligned}
E(x) &= \int_0^L x |\Psi_n(x)|^2 dx = \Psi_0^2 \int_0^L x \sin^2\left(\frac{n\pi x}{L}\right) dx = \frac{2}{L} \int_0^L x \sin^2\left(\frac{n\pi x}{L}\right) dx \\
&= \frac{1}{L} \int_0^L \left\{ 1 - \cos\left(\frac{2n\pi x}{L}\right) \right\} x dx = \frac{1}{L} \left\{ \int_0^L x dx - \int_0^L \cos\left(\frac{2n\pi x}{L}\right) x dx \right\} \\
&= \frac{1}{L} \left\{ \frac{1}{2} x^2 \Big|_0^L - \left(\frac{L}{2n\pi} \right) \left[\sin\left(\frac{2n\pi x}{L}\right) x \Big|_0^L - \int_0^L \sin\left(\frac{2n\pi x}{L}\right) dx \right] \right\} \\
&= \frac{1}{L} \left\{ \frac{L^2}{2} - \left(\frac{L}{2n\pi} \right) \left[0 + \frac{L}{2n\pi} \cos\left(\frac{2n\pi x}{L}\right) \Big|_0^L \right] \right\} \\
&= \frac{1}{L} \left\{ \frac{L^2}{2} - \left(\frac{L}{2n\pi} \right) \left[0 + \frac{L}{2n\pi} (1 - 1) \right] \right\} = \frac{1}{L} \frac{L^2}{2} = \frac{L}{2}
\end{aligned} \tag{42.351}$$

3. Également sans démonstration car ce résultat est trop évident (si jamais il ne l'est pas pour vous dites-le et nous ajouterons le développement comme pour tout autre chose dans ce site d'ailleurs), la quantité de mouvement le long de x est nulle: $p_x = 0$

4. Nous pouvons par ailleurs vérifier sans trop de peine que ce que nous avons vu lors de l'énoncé du 2ème postulat se vérifie bien dans cet exemple. C'est-à-dire que les fonctions propres de l'onde sont reliées à l'opérateur hamiltonien via les valeurs propres de l'énergie:

$$H\Psi = E_{tot}\Psi \quad (42.352)$$

Effectivement, dans notre exemple, cela donne:

$$\begin{aligned} \left(\underbrace{E_{pot}}_{=0} - \frac{\hbar^2}{2m} \Delta \right) \Psi &= E_{tot} \Psi \\ \Rightarrow -\frac{\hbar^2}{2m} \Delta \Psi &= \frac{\pi^2 \hbar^2 n^2}{2mL^2} \Psi \\ \Rightarrow -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \left(\sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) \right) &= \frac{\pi^2 \hbar^2 n^2}{2mL^2} \left(\sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) \right) \quad (42.353) \\ \Rightarrow -\frac{\partial^2}{\partial x^2} \left(\sin\left(\frac{n\pi x}{L}\right) \right) &= \frac{\pi^2 n^2}{L^2} \left(\sin\left(\frac{n\pi x}{L}\right) \right) \\ \Rightarrow -\frac{\pi^2 n^2}{L^2} \left(-\sin\left(\frac{n\pi x}{L}\right) \right) &= \frac{\pi^2 n^2}{L^2} \left(\sin\left(\frac{n\pi x}{L}\right) \right) \\ \Rightarrow 1 &= 1 \end{aligned}$$

voilà... pour la première approche du problème. Voyons maintenant la deuxième:

2ème approche

Nous avons donc l'équation de Schrödinger dans le cas unidimensionnel:

$$\frac{\partial^2 \Psi}{\partial x^2} + \frac{8\pi^2 m}{h^2} (E_{tot} - E_{pot}) \Psi = 0 \quad (42.354)$$

Dans les régions situées en dehors de la boîte où le potentiel est infini, nous avons:

$$\frac{\partial^2 \Psi}{\partial x^2} + \frac{8\pi^2 m}{h^2} (E_{tot} - \infty) \Psi = 0 \quad (42.355)$$

Soit:

$$\frac{\partial^2 \Psi}{\partial x^2} = \infty \Psi \quad (42.356)$$

ce qui donne:

$$\Psi = \frac{1}{\infty} \frac{\partial^2 \Psi}{\partial x^2} = 0 \quad (42.357)$$

Ainsi, les fonctions d'onde sont nulles dans les régions où le potentiel est infini.

Considérons maintenant le cas du puits où puisque le potentiel électrostatique est nul, l'équation de Schrödinger se réduit à:

$$\frac{\partial^2 \Psi}{\partial x^2} + \frac{8\pi^2 m}{h^2} E_{tot} \Psi = 0 \quad (42.358)$$

C'est donc une équation différentielle linéaire d'ordre 2 avec des coefficients constants, équation qu'il est relativement aisé de résoudre dans le cas général (cf. chapitre de Calcul Différentiel Et Intégral). Soit l'équation:

$$y'' + py' + qy = 0 \quad (42.359)$$

En nous aidant des résultats obtenus lors du traitement de la solution particulière, supposons que la fonction y qui satisfait cette équation différentielle soit de la forme $y = e^{Kx}$. Nous avons alors:

$$K^2 e^{Kx} + pK e^{Kx} + q e^{Kx} = 0 \text{ ou } K^2 + pK + q = 0 \quad (42.360)$$

pourvu, bien sûr, que $e^{Kx} \neq 0$. Cette dernière relation est donc l'équation quadratique auxiliaire de l'équation différentielle (polynôme caractéristique). Elle a deux solutions/racines (c'est une simple résolution d'un polynôme du deuxième degré) que nous noterons dans le cas général K_1, K_2 . Ce qui signifie que:

$$K_1^2 e^{K_1 x} + pK_1 e^{K_1 x} + q e^{K_1 x} = 0 \text{ et } K_2^2 e^{K_2 x} + pK_2 e^{K_2 x} + q e^{K_2 x} = 0 \quad (42.361)$$

est satisfait pour les deux racines. Si nous faisons la somme puisque les deux sont égales à la même constante:

$$(K_1^2 e^{K_1 x} + K_2^2 e^{K_2 x}) + p(K_1 e^{K_1 x} + K_2 e^{K_2 x}) + q(e^{K_1 x} + e^{K_2 x}) = 0 \quad (42.362)$$

Ainsi, il est immédiat que la solution générale de y est du type:

$$y = A e^{K_1 x} + B e^{K_2 x} \quad (42.363)$$

où le lecteur devrait normalement sans peine pouvoir vérifier que l'ajout des constantes A et B ne change en rien les développements des paragraphes précédents.

Dans le cas qui nous occupe:

$$\Psi'' + \frac{8\pi^2 m}{h^2} E_{tot} \Psi = 0 \quad (42.364)$$

L'équation quadratique est:

$$k^2 + \frac{8\pi^2 m}{h^2} E_{\text{tot}} = 0 \Rightarrow k^2 = -2mE_{\text{tot}} \hbar^{-2} \quad (42.365)$$

soit:

$$k = \pm \frac{\sqrt{-2mE_{\text{tot}}}}{\hbar} = \pm i \frac{\sqrt{2mE_{\text{tot}}}}{\hbar} \quad (42.366)$$

Donc finalement la solution générale est de la forme:

$$\Psi = A \exp\left(i \frac{\sqrt{2mE_{\text{tot}}}}{\hbar} x\right) + B \exp\left(-i \frac{\sqrt{2mE_{\text{tot}}}}{\hbar} x\right) \quad (42.367)$$

Posons maintenant:

$$\theta = \frac{\sqrt{2mE_{\text{tot}}}}{\hbar} x \quad (42.368)$$

Nous avons alors:

$$\begin{aligned} \Psi &= A e^{i\theta} + B e^{-i\theta} = A \cos \theta + iA \sin \theta + B \cos \theta - iB \sin \theta \\ &= A' \cos \theta + B' \sin \theta \end{aligned} \quad (42.369)$$

avec:

$$A' = A + B \text{ et } B' = i(A - B) \quad (42.370)$$

Il faut maintenant déterminer A' et B' en utilisant les conditions aux limites. Ainsi, en $x=0$ et $x=L$ nous devrions avoir $\Psi = 0$ et nous avons pour $x=0$:

$$\Psi = A' \cos\left(\frac{\sqrt{2mE_{\text{tot}}}}{\hbar} x\right) + B' \sin\left(\frac{\sqrt{2mE_{\text{tot}}}}{\hbar} x\right) = A' \quad (42.371)$$

Le coefficient A' doit donc être nul. Et en $x=L$ nous devrions avoir:

$$\begin{aligned} \Psi &= A' \cos\left(\frac{\sqrt{2mE_{\text{tot}}}}{\hbar} x\right) + B' \sin\left(\frac{\sqrt{2mE_{\text{tot}}}}{\hbar} x\right) \\ &= 0 \cdot \cos\left(\frac{\sqrt{2mE_{\text{tot}}}}{\hbar} x\right) + B' \sin\left(\frac{\sqrt{2mE_{\text{tot}}}}{\hbar} L\right) = B' \sin\left(\frac{\sqrt{2mE_{\text{tot}}}}{\hbar} L\right) = 0 \end{aligned} \quad (42.372)$$

Mais dans ce cas, B' doit être différent de zéro. En effet, s'il était nul, la fonction d'onde serait nulle dans tout le puits ce qui est contraire à la réalité physique du problème. Il faut donc que ce soit le sinus qui soit nul, ou encore que son argument soit égal à un multiple d'un nombre entier non nul d'angle π tel que:

$$\frac{\sqrt{2mE_{\text{tot}}}}{\hbar} L = \pm n\pi \quad (42.373)$$

Donc:

$$E_{tot} = \frac{\pi^2 \hbar^2}{2mL^2} n^2 \quad (42.374)$$

Nous retrouvons donc exactement le même résultat que la méthode précédente.

Il reste à déterminer B et la méthode est exactement identique à la première méthode de résolution que nous avons vue plus haut. Ainsi, nous avons bien:

$$\Psi = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) \quad (42.375)$$

Ce qui est important surtout dans cette méthode, c'est de se souvenir pour plus tard de la forme générale de la solution:

$$\Psi = Ae^{ikx} + Be^{-ikx} \quad (42.376)$$

4.4.3. ÉNERGIE DE FERMI

L'énergie de Fermi, E_f , est un concept en physique quantique qui désigne l'énergie du plus haut état quantique occupé dans un système idéalisé où toutes les couches se remplissent successivement et sans discontinuités, c'est-à-dire en pratique lorsqu'un système est au zéro absolu: 0 [K] (cf. chapitre de Mécanique Statistiques).

Le mur de potentiel rectiligne est un excellent moyen pédagogique d'introduire pratiquement le niveau d'énergie de Fermi pour ensuite l'étendre à d'autres cas particuliers.

Rappelons donc que nous venons d'obtenir de deux manières différentes:

$$E_n = \frac{\pi^2 \hbar^2}{2mL^2 \pi^2 4} n^2 = \frac{\pi^2 \hbar^2}{2mL^2} n^2 \quad (42.377)$$

Alors l'énergie totale de tout un système composé de N particules pouvant prendre toutes le même état fondamental $n = 1$ et violant donc le principe de Pauli sera:

$$E_{tot} = \left(\frac{\pi^2 \hbar^2}{2mL^2} \right) N \quad (42.378)$$

Mais si nous appliquons le principe d'exclusion de Pauli des électrons (fermions), chaque niveau ne peut alors dans ce modèle (qui ne contient pas de sous-couches ou autres subtilités) prendre que deux états (oppositions de spin). Dès lors, chaque le niveau fondamental ne peut être occupé que par deux états et il en est de même de chaque niveau. L'énergie totale est alors:

$$E_{tot} = 2 \sum_{n=1}^{N/2} E_n = 2 \sum_{n=1}^{N/2} \frac{\pi^2 \hbar^2}{2mL^2} n^2 = \frac{\pi^2 \hbar^2}{mL^2} \sum_{n=1}^{N/2} n^2 \quad (42.379)$$

En utilisant la somme des carrés démontrée dans le chapitre de Suites Et Séries et en notant comme il est d'usage:

$$n_{\max} = \frac{N}{2} \quad (42.380)$$

Nous avons alors:

$$E_{\text{tot}} = \frac{\pi^2 \hbar^2}{mL^2} \sum_{n=1}^{N/2} n^2 = \frac{\pi^2 \hbar^2}{mL^2} \left(\frac{n_{\max} (n_{\max} + 1) (2n_{\max} + 1)}{6} \right) \cong \frac{\pi^2 \hbar^2}{mL^2} \frac{n_{\max}^3}{3} = \frac{\pi^2 \hbar^2}{24mL^2} N^3 \quad (42.381)$$

L'énergie moyenne par particule est alors:

$$\bar{E} = \frac{E_{\text{tot}}}{N} \cong \frac{\pi^2 \hbar^2}{24mL^2} N^2 \quad (42.382)$$

Par définition, l'énergie du dernier niveau à être occupé est celui de Fermi et donc donné par:

$$E_n = \frac{\pi^2 \hbar^2}{2mL^2} n^2 \Rightarrow E_{N/2} = E_f = \frac{\pi^2 \hbar^2}{8mL^2} N^2 \quad (42.383)$$

Nous remarquons que nous avons quel que soit N :

$$\bar{E} = \frac{1}{3} E_f \quad (42.384)$$

4.4.4. OSCILLATEUR HARMONIQUE

L'étude de l'oscillateur harmonique correspondant à celle d'une fonction d'onde coincée dans un puits de potentiel parabolique. Ce qui est assimilable grosso modo aux atomes où les parois du puits de potentiel ne sont naturellement pas rectangulaires et infinies... L'étude qui va suivre est donc ce qui est le plus proche de ce qui est disponible dans la Nature au niveau atomique.

Dans le cas d'une particule libre en déplacement rectiligne, nous avons vu que l'énergie potentielle est nulle $E_{\text{pot}} = 0$ et l'équation de Schrödinger devient alors:

$$\frac{\partial^2 \Psi}{\partial x^2} + \frac{2 \cdot m \cdot E_{\text{tot}}}{\hbar^2} \cdot \Psi = 0 \quad (42.385)$$

Cependant, pour une particule libre (en l'absence de champ de potentiel) l'énergie totale est donc égale à l'énergie cinétique:

$$E_{\text{tot}} = \frac{1}{2} m \cdot v^2 = \frac{p^2}{2 \cdot m} \quad (42.386)$$

Mais nous avons:

$$E_{tot} = h \cdot \nu = h \cdot \frac{\omega}{2\pi} = \hbar \cdot \omega = m \cdot c^2 \Leftrightarrow p = m \cdot c = \frac{E_{tot}}{c} = \frac{h \cdot \nu}{c} = h \cdot \lambda \quad (42.387)$$

Le rapport:

$$\lambda = \frac{h}{p} \quad (42.388)$$

étant la longueur d'onde associée de De Broglie. En introduisant le nombre d'onde $k = 2\pi/\lambda$ (cf. chapitre de Mécanique Ondulatoire), nous avons:

$$\boxed{p = \hbar \cdot k} \quad (42.389)$$

appelée "**relation de De Broglie**". Finalement:

$$E_{tot} = \frac{p^2}{2 \cdot m} = \frac{(\hbar \cdot k)^2}{2 \cdot m} \quad (42.390)$$

Dès lors, l'équation de Schrödinger peut s'écrire:

$$\frac{\partial^2 \Psi}{\partial x^2} + k^2 \Psi = 0 \quad (42.391)$$

Nous voyons par substitution directe que cette équation différentielle admet pour solutions les fonctions d'onde:

$$\Psi_1(x) = A e^{ikx} \text{ et } \Psi_2(x) = B e^{-ikx} \quad (42.392)$$

Ces deux différentes solutions représentent le déplacement d'une même particule une fois dans la direction $+x$ et l'autre dans $-x$. Si $A = B = 1$ nous avons:

$$|\Psi(x)|^2 = \bar{\Psi}(x) \cdot \Psi(x) = e^{-ikx} \cdot e^{ikx} = 1 \quad (42.393)$$

Le fait que ce résultat soit égal à l'unité, signifie que la probabilité de trouver la particule est la même en tout point. En d'autres termes, $\Psi(x) = e^{\pm ikx}$ décrit une situation dans laquelle l'incertitude sur la position est totale. Ce résultat est en accord avec le principe d'incertitude puisque $\Psi(x) = e^{\pm ikx}$ décrit une particule dont nous connaissons avec précision la quantité mouvement $p = \hbar \cdot k$: c'est-à-dire que $\Delta p = 0$, ce qui implique $\Delta x \rightarrow \infty$.

Dans le chapitre de Calcul Différentiel Et Intégral nous avons montré que la solution la plus générale d'une équation différentielle est la somme de ces solutions. Autrement dit dans notre exemple:

$$\Psi(x) = A e^{ikx} + B e^{-ikx} \quad (42.394)$$

avec:

$$k = \frac{\sqrt{2 \cdot m \cdot E_{tot}}}{\hbar} \quad (42.395)$$

Au fait, nous pouvons remarquer que si $E_{pot} \neq 0$ alors le résultat est le même à la différence près que nous aurons:

$$k = \frac{\sqrt{2 \cdot m \cdot (E_{tot} - E_{pot})}}{\hbar} \quad (42.396)$$

Lorsque la particule qui nous intéresse se trouve dans un puits de potentiel décrit par la fonction (parabole) :

$$E_{pot}(x) = \frac{1}{2} \cdot c^{te} \cdot x^2 \quad (42.397)$$

nous parlons alors "**d'oscillateur harmonique**".

Ce système est très important car l'hamiltonien de l'équation intervient dans tous les problèmes mettant en jeu des oscillations telles que vibrations moléculaires et cristallines (cf. chapitre de Chimie Quantique).

Prenons d'abord comme exemple l'oscillateur harmonique classique qui consiste en un corps assujéti à se déplacer le long d'un axe et soumis à une force de rappel proportionnelle à la distance à un point situé sur cet axe.

L'équation de ce corps est régie par l'équation de la dynamique:

$$m \cdot \frac{d^2 x}{dt^2} + k \cdot x = 0 \quad (42.398)$$

Nous avons vu en mécanique classique que la solution générale de cette équation est:

$$x = x_0 \cdot \cos(\omega \cdot t + \varphi) \quad (42.399)$$

avec comme pulsation:

$$\omega = \sqrt{\frac{k}{m}} \quad (42.400)$$

L'énergie totale du système étant l'hamiltonien classique, nous écrivons:

$$H = E_{cin} + E_{pot} = \frac{1}{2} m \cdot \left(\frac{dx}{dt} \right)^2 + \frac{1}{2} k \cdot x(t)^2 = \frac{p^2}{2m} + \frac{1}{2} k \cdot x(t)^2 \quad (42.401)$$

Il s'ensuit au vu de l'expression de l'énergie potentielle que la constante que nous avons dans la relation de départ qui définissait la puits de potentiel parabolique:

$$E_{pot}(x) = \frac{1}{2} \cdot c^{te} \cdot x^2 \quad (42.402)$$

est simplement k et dès lors:

$$E_{pot}(x) = \frac{1}{2} \cdot \omega^2 \cdot m \cdot x^2 \quad (42.403)$$

mais que nous noterons pour la suite et par tradition sous la forme suivante:

$$E_{pot} = \frac{1}{2} \cdot \omega_0^2 \cdot m \cdot x^2 \quad (42.404)$$

Maintenant revenons à notre cadre quantique. De ce point de vue nous avons pour hamiltonien (ou énergie totale):

$$H = E_{cin} + E_{pot} = \frac{p^2}{2m} + \frac{c^{te}}{2} \cdot x^2 \quad (42.405)$$

où donc pour rappel suite à ce que nous venons de voir:

$$c^{te} = \omega_0^2 \cdot m \quad (42.406)$$

En utilisant ce que nous définissons comme une "**écriture réduite**", nous écrivons:

$$H = \frac{\hbar \cdot \omega_0}{2} (P^2 + Q^2) \quad (42.407)$$

où les opérateurs quantité de mouvements et positions réduits sont respectivement:

$$P = \frac{p}{\sqrt{m \cdot \hbar \cdot \omega_0}} \text{ et } Q = \sqrt{\frac{c^{te}}{\hbar \cdot \omega_0}} x \quad (42.408)$$

et où nous avons remplacé la constante par $\omega_0 = \sqrt{\frac{c^{te}}{m}}$ identiquement à l'oscillateur harmonique classique (cf. chapitre de Mécanique Classique).

Il est plus ou moins facile d'obtenir la relation de commutation:

$$[Q, P] = \sqrt{\frac{c^{te}}{m}} \frac{1}{\hbar \omega_0} [x, p] \quad (42.409)$$

Démonstration:

Rappelez-vous de la relation ci-dessous que nous avons vue lors de notre étude des opérateurs linéaires fonctionnels au début de ce chapitre:

$$p_x = \frac{1}{i} \hbar \frac{\partial}{\partial x}, p_y = \frac{1}{i} \hbar \frac{\partial}{\partial y}, p_z = \frac{1}{i} \hbar \frac{\partial}{\partial z} \quad (42.410)$$

Étudions les propriétés des commutateurs avec la quantité de mouvement. Nous avons démontré également plus haut la relation ci-dessous:

$$\frac{d}{dx}x = 1 + x \frac{d}{dx} \quad (42.411)$$

En multipliant cette dernière par $\hbar/i$, il vient:

$$\frac{\partial}{\partial x} \frac{\hbar}{i} x = x \frac{\partial}{\partial x} \frac{\hbar}{i} + 1 \frac{\hbar}{i} \quad (42.412)$$

que nous pouvons également écrire:

$$p_x \cdot x = x \cdot p_x - \hbar \cdot i \quad (42.413)$$

Si vous vous rappelez de la définition des commutateurs ($[\alpha, \beta] = \alpha\beta - \beta\alpha$), nous avons:

$$[x, p_x] = x \cdot p_x - p_x \cdot x = \hbar \cdot i \quad (42.414)$$

Nous avons donc pour notre oscillateur:

$$P = \frac{p}{\sqrt{m \cdot \hbar \cdot \omega_0}} = \frac{1}{\sqrt{m \cdot \hbar \cdot \omega_0}} \frac{\partial}{\partial x} i\hbar \text{ et } Q = \sqrt{\frac{c^2}{\hbar \cdot \omega_0}} x \quad (42.415)$$

écrivons la définition du commutateur:

$$[Q, P] = QP - PQ \quad (42.416)$$

Donc:

$$\begin{aligned} [Q, P] &= \sqrt{\frac{c^2}{\hbar \cdot \omega_0}} x \frac{1}{\sqrt{m \cdot \hbar \cdot \omega_0}} \frac{\partial}{\partial x} i\hbar - \frac{1}{\sqrt{m \cdot \hbar \cdot \omega_0}} \frac{\partial}{\partial x} i\hbar \cdot \sqrt{\frac{c^2}{\hbar \cdot \omega_0}} x \\ &= \sqrt{\frac{c^2}{\hbar \cdot \omega_0}} \frac{1}{\sqrt{m \cdot \hbar \cdot \omega_0}} \cdot \left(x \cdot \frac{\partial}{\partial x} i\hbar - \frac{\partial}{\partial x} i\hbar \cdot x \right) = \sqrt{\frac{c^2}{m \cdot \hbar \cdot \omega_0}} \cdot \left(x \cdot \frac{\partial}{\partial x} i\hbar - \frac{\partial}{\partial x} i\hbar \cdot x \right) \\ &= \sqrt{\frac{c^2}{m \cdot \hbar \cdot \omega_0}} \cdot (x \cdot p_x - p_x \cdot x) = \sqrt{\frac{c^2}{m \cdot \hbar \cdot \omega_0}} \cdot [x, p_x] = \omega_0 \frac{1}{\hbar \cdot \omega_0} \cdot [x, p_x] = \frac{1}{\hbar} [x, p_x] = i \end{aligned} \quad (42.417)$$

c'est ce qu'il fallait démontrer...

Nous avons maintenant intérêt pour résoudre l'équation différentielle d'utiliser les opérateurs non hermitiques a et $a^\dagger$ définis (c'est une définition donc ne cherchez pas trop loin):

$$\begin{aligned} \sqrt{2}a &= Q + iP \\ \sqrt{2}a^\dagger &= Q - iP \end{aligned} \quad (42.418)$$

Ce qui nous définit donc les opérateurs (en posant temporairement $m = 1$):

$$a = \frac{Q + iP}{\sqrt{2}} = \frac{\omega x + ip}{\sqrt{2\hbar\omega_0}} \quad a^\dagger = \frac{Q - iP}{\sqrt{2}} = \frac{\omega x - ip}{\sqrt{2\hbar\omega_0}} \quad (42.419)$$

qui sont des des combinaisons linéaires des opérateurs position et quantité de mouvement et ne sont trivialement pas hermitiques.

Nous retrouvons ces deux opérateurs très fréquemment en physique quantique et les physiciens parlent alors de "l'opérateur de création" $a^\dagger$ et de "l'opérateur de destruction" a (appelé aussi "opérateur d'annihilation").

Compte tenu de la relation de commutation, nous vérifions:

$$2aa^\dagger = P^2 + Q^2 + 1 \text{ et } 2a^\dagger a = P^2 + Q^2 - 1 \quad (42.420)$$

Démonstration:

$$\begin{aligned} 2aa^\dagger &= (Q + iP)(Q - iP) = Q^2 + iPQ - iQP + P^2 \\ &= Q^2 - i[QP - PQ] + P^2 = Q^2 - i[Q, P] + P^2 \\ &= P^2 + Q^2 - i \cdot i = P^2 + Q^2 + 1 \end{aligned} \quad (42.421)$$

et:

$$\begin{aligned} 2a^\dagger a &= (Q - iP)(Q + iP) = Q^2 - iPQ + iQP + P^2 \\ &= Q^2 + i[QP - PQ] + P^2 = Q^2 + i[Q, P] + P^2 \\ &= P^2 + Q^2 + i \cdot i = P^2 + Q^2 - 1 \end{aligned} \quad (42.422)$$

et d'autre part:

$$[a, a^\dagger]_- = aa^\dagger - a^\dagger a = 1 \quad (42.423)$$

Démonstration:

$$2aa^\dagger - 2a^\dagger a = (P^2 + Q^2) - (P^2 + Q^2) + 1 - (-1) = 2 \quad (42.424)$$

et donc en divisant par 2 des deux côtés de l'égalité, nous avons:

$$aa^\dagger - a^\dagger a = 1 \quad (42.425)$$

Revenons à la relation:

$$H = \frac{\hbar \cdot \omega_0}{2} (P^2 + Q^2) \quad (42.426)$$

Utilisons:

$$a^\dagger a = \frac{P^2 + Q^2 - 1}{2} \Rightarrow \frac{P^2 + Q^2}{2} = a^\dagger a + \frac{1}{2} = N + \frac{1}{2} \quad (42.427)$$

où il apparaît que N peut être vu comme un opérateur, appelé "opérateur compteur" (hermitique) puisque:

$$N = a^\dagger a \quad (42.428)$$

Il suffit alors de quantifier l'opérateur compteur pour les connaître les valeurs propres et les états propres de H puisque ces deux sont liés alors par:

$$H = \hbar \cdot \omega_0 \left(\frac{P^2 + Q^2}{2} \right) = \hbar \cdot \omega_0 \left(N + \frac{1}{2} \right) \quad (42.429)$$

Nous faisons maintenant l'hypothèse que Ψ est une fonction propre de N associée à la valeur propre n , telle que:

$$\begin{aligned} H\Psi &= E\Psi \\ \Rightarrow \hbar \omega_0 \left(N + \frac{1}{2} \right) \Psi_n &= \hbar \omega_0 \left(n + \frac{1}{2} \right) \Psi_n \quad (42.430) \\ \Rightarrow N\Psi_n &= n\Psi_n \end{aligned}$$

Cette hypothèse est très importante car nous allons nous en servir comme principe d'induction pour trouver toutes les fonctions propres à partir de la fondamentale!

Établissons maintenant des relations de commutation entre N et les opérateurs a ou $a^\dagger$. Pour cela, multiplions d'abord $aa^\dagger - a^\dagger a = 1$ le tout par $a^\dagger$, nous obtenons:

$$Na^\dagger - a^\dagger N = a^\dagger \quad (42.431)$$

De même en multipliant $aa^\dagger - a^\dagger a = 1$ par a , nous obtenons:

$$aN - Na = a \quad (42.432)$$

Puisque selon notre hypothèse Ψ et n sont respectivement fonction et valeur propre de N , nous pouvons écrire:

$$N\Psi = n\Psi \quad (42.433)$$

Or, nous avons:

$$Na = aN - a \quad (42.434)$$

qui multipliée à droite par la fonction d'onde donne la relation:

$$Na\Psi = (n-1)a\Psi = na\Psi - a\Psi \quad (42.435)$$

Cette équation entraîne les conséquences suivantes:

- Ou bien $a\Psi = 0$ tel que $Na\Psi = na\Psi$
- Ou bien $a\Psi$ est fonction propre de N pour la valeur propre $n-1$!!

Le même raisonnement établirait que $a^\dagger\Psi$ est fonction propre de N pour la valeur propre $n+1$, si elle n'est pas nulle (nous verrons plus loin que $a^\dagger\Psi$ n'est jamais nulle):

$$N(a^\dagger\Psi) = (n+1)a^\dagger\Psi \quad (42.436)$$

Cette relation est importante car si $a^\dagger \Psi$ n'est pas nulle pour une fonction propre donnée elle ne le sera pas non plus pour les autres fonctions propres de valeur propre $n+1$!!

Nous savons qu'il existe une valeur propre n_0 plus petite que toutes les autres correspondant au niveau fondamental (d'après le modèle de Bohr-Sommerfeld cette valeur propre existe toujours).

Nécessairement, sa fonction propre Ψ_0 obéit à la relation (le lecteur pourra vérifier avec les résultats plus loin):

$$a \Psi_0 = 0 \quad (42.437)$$

sinon quoi $n_0 - 1$ serait valeur propre et il y aurait contradiction.

En multipliant cette dernière relation par $a^\dagger$ nous obtenons:

$$N \Psi_0 = 0 \quad (42.438)$$

ce qui montre que la valeur propre minimale n_0 est nulle. Nous connaissons donc le niveau fondamental de l'oscillateur:

$$E_0 = \hbar \omega_0 \left(n_0 + \frac{1}{2} \right) = \frac{1}{2} \hbar \omega_0 \quad (42.439)$$

Remarque: Il faut noter que l'oscillateur n'est jamais dans un état de repos (mettre $n = 0$ dans l'expression de l'énergie plus haut) ce qui veut aussi dire que le zéro absolu ne peut pas être accessible puisque la température "chiffre" l'agitation atomique, or le repos n'existe pas!

Pour obtenir la fonction propre correspondante, nous avons besoin de l'expression explicite de a . D'après:

$$P = \frac{p}{\sqrt{m \cdot \hbar \cdot \omega_0}} \quad \text{et} \quad Q = \sqrt{\frac{c^{te}}{\hbar \cdot \omega_0}} x \quad (42.440)$$

nous avons:

$$P = \frac{1}{\sqrt{m \cdot \hbar \cdot \omega_0}} \frac{\hbar}{i} \frac{d}{dx} \quad \text{et} \quad dQ = \sqrt{\frac{c^{te}}{\hbar \cdot \omega_0}} dx \Rightarrow \frac{1}{dx} = \sqrt{\frac{c^{te}}{\hbar \cdot \omega_0}} \frac{1}{dQ} \quad (42.441)$$

ce qui nous donne:

$$P = \frac{1}{\sqrt{m \cdot \hbar \cdot \omega_0}} \frac{\hbar}{i} \sqrt{\frac{c^{te}}{\hbar \cdot \omega_0}} \frac{d}{dQ} = \frac{1}{i} \frac{d}{dQ} \quad (42.442)$$

car rappelons-le:

$$\omega_0 = \sqrt{\frac{c^{te}}{m}} \quad (42.443)$$

d'où:

$$\begin{aligned}\sqrt{2}a &= Q + \frac{d}{dQ} \\ \sqrt{2}a^\dagger &= Q - \frac{d}{dQ}\end{aligned}\quad (42.444)$$

Mais d'après $a\Psi_0 = 0$:

$$\left(Q + \frac{d}{dQ}\right)\Psi_0 = 0 \quad (42.445)$$

d'où:

$$\frac{1}{\Psi_0} \frac{d}{dQ} \Psi_0 = -Q \quad (42.446)$$

soit (résolution d'une simple équation différentielle):

$$\Psi_0(Q) = C^{te} e^{-Q^2/2} \quad (42.447)$$

Nous devons envisager, en réalité, Ψ_0 comme fonction de x par le biais de la coordonnée réduite Q .

D'après:

$$Q = \sqrt{\frac{c^{te}}{\hbar \cdot \omega_0}} x \quad (42.448)$$

en introduisant la longueur A :

$$x = A \cdot Q \Rightarrow dx = A \cdot dQ \quad (42.449)$$

avec:

$$A = \sqrt{\frac{\hbar \omega_0}{c^{te}}} \quad (42.450)$$

et comme:

$$c^{te} = \omega_0^2 \cdot m \quad (42.451)$$

Il vient alors:

$$A = \sqrt{\frac{\hbar}{\omega_0 m}} \quad (42.452) \quad \text{et} \quad Q = \sqrt{\frac{\omega_0 m}{\hbar}} x \quad (42.453)$$

Nous allons vérifier si nous satisfaisons la condition de normalisation de De Broglie, sinon quoi il sera nécessaire d'introduire une constante multiplicative. Nous avons alors:

$$\int_{-\infty}^{+\infty} |\Psi_0|^2 dx = 1 = \int_{-\infty}^{+\infty} \left(C^{te} e^{-Q^2/2} \right)^2 A dQ = A |C^{te}|^2 \int_{-\infty}^{+\infty} e^{-Q^2} dQ = A |C^{te}|^2 \sqrt{\pi} \quad (42.454)$$

et donc nous devons avoir:

$$1 = A |C^{te}|^2 \sqrt{\pi} \Rightarrow C^{te} = A^{-1/2} \pi^{-1/4} \quad (42.455)$$

Il est loisible de choisir la constante réelle et positive, nous avons finalement:

$$\Psi_0 = C^{te} e^{-Q^2/2} = A^{-1/2} \pi^{-1/4} e^{-Q^2/2} \quad (42.456)$$

Corollaire...: D'après ce que nous avons vu précédemment, en faisant agir $\alpha^\dagger$ sur Ψ_0 (explicitement nous faisons référence au résultat $N(\alpha^\dagger \Psi) = (n+1)\alpha^\dagger \Psi$), nous obtenons les fonctions propres de N pour les valeurs propres entières 1, 2, etc. Nous vérifierons plus loin que nous épuisons ainsi toutes les valeurs propres de N .

Il reste à construire les autres fonctions propres et à les normer. En effet, si Ψ_n est fonction propre normée associée au niveau E_n , nous avons vu plus haut que $\alpha^\dagger \Psi_n$ est fonction propre associée au niveau $n+1$ (l'opérateur de création $\alpha^\dagger$ fait augmenter l'énergie d'où son nom...), mais il n'y a pas de raison de la normer à nouveau puisqu'elle est justement associée à une fonction propre déjà normée.

Nous pouvons écrire:

$$\alpha^\dagger \Psi_n = \alpha_n \Psi_{n+1} \quad (42.457)$$

α_n étant un coefficient à déterminer. Exprimons le fait que Ψ_{n+1} est déjà normée:

$$|\alpha_n|^2 = \int_{-\infty}^{+\infty} (\alpha^\dagger \Psi_n) \cdot (\alpha^\dagger \Psi_n) dx = \int_{-\infty}^{+\infty} \bar{\Psi}_n \cdot (a \alpha^\dagger \Psi_n) dx \quad (42.458)$$

Soit en tenant compte de la relation $a \alpha^\dagger - \alpha^\dagger a = 1$ nous avons:

$$|\alpha_n|^2 = \int_{-\infty}^{+\infty} \bar{\Psi}_n [(1 + \alpha^\dagger a) \Psi_n] dx \quad (42.459)$$

Rappelons que $N(\alpha^\dagger \Psi) = (n+1)(\alpha^\dagger \Psi)$ donc:

$$|\alpha_n|^2 = \int_{-\infty}^{+\infty} \bar{\Psi}_n [(1 + \alpha^\dagger a) \Psi_n] dx = n+1 \quad (42.460)$$

Nous venons de vérifier au passage que $\alpha^\dagger \Psi$ n'est jamais nul (fait que nous avons supposé plus haut).

Toutes les fonctions Ψ_n (sauf Ψ_0 déjà fixée) ont un facteur de phase arbitraire (notion que nous avons vues lors de la définition des états liés et non liés), indépendamment les uns des autres, l'argument de α_n reste donc à notre disposition et nous choisirons α_n réel positif. Cela fixe toutes les Ψ_n :

$$\alpha^\dagger \Psi_n = \sqrt{n+1} \Psi_{n+1} \quad (42.461)$$

En itérant cette relation sur la fonction d'onde, nous obtenons aisément (algèbre élémentaire):

$$\Psi_n = \frac{1}{\sqrt{n!}} (\alpha^\dagger)^n \Psi_0 \quad (42.462)$$

soit en tenant compte des relations suivantes (que nous avons déjà démontrées précédemment):

$$\begin{aligned} \sqrt{2}a &= Q + \frac{d}{dQ} \\ \sqrt{2}a^\dagger &= Q - \frac{d}{dQ} \end{aligned} \quad \text{et} \quad \Psi_0 = A^{-1/2} \pi^{-1/4} e^{-Q^2/2} \quad (42.463)$$

Nous avons alors:

$$\Psi_n = \left(A \pi^{-1/2} n! 2^n \right)^{-1/2} \left(Q - \frac{d}{dQ} \right)^n e^{-Q^2/2} \quad (42.464)$$

Cette équation prend une forme plus simple, en s'appuyant sur la relation:

$$\left(Q - \frac{d}{dQ} \right) f(Q) = -e^{Q^2/2} \frac{d}{dQ} e^{-Q^2/2} f(Q) \quad (42.465)$$

Vérification:

$$\begin{aligned} -e^{Q^2/2} \frac{d}{dQ} e^{-Q^2/2} f(Q) &= -e^{Q^2/2} \left(-Q e^{-Q^2/2} f(Q) + e^{-Q^2/2} \frac{d}{dQ} f(Q) \right) \\ &= -e^{Q^2/2} e^{-Q^2/2} \left(-Q f(Q) + \frac{d}{dQ} f(Q) \right) = \left(Q - \frac{d}{dQ} \right) f(Q) \end{aligned} \quad (42.466)$$

soit, en langage d'opérateurs:

$$Q - \frac{d}{dQ} = -e^{Q^2/2} \frac{d}{dQ} e^{-Q^2/2} \quad (42.467)$$

Ainsi:

$$\left(Q - \frac{d}{dQ} \right)^n = (-)^n e^{Q^2/2} \left(\frac{d}{dQ} \right)^n e^{-Q^2/2} \quad (42.468)$$

Nous obtenons ainsi l'expression de Ψ_n :

$$\Psi_n = (-)^n \left(A\pi^{-1/2} n! 2^n \right)^{-1/2} e^{\varrho^2/2} \frac{d^n}{dQ^n} e^{-\varrho^2} \quad (42.469)$$

Par ailleurs, dans la théorie mathématique des familles de polynômes orthogonaux, nous rencontrons les "polynômes d'Hermite" H_n définis par:

$$H_n(Q) = (-)^n e^{\varrho^2} \frac{d^n}{dQ^n} e^{-\varrho^2} \quad (42.470)$$

Ce sont des polynômes de degré n , pairs ou impairs ($H_0 = 1, H_1 = 2Q, H_2 = 4Q^2 - 2$). En les employant, nous allégeons la relation précédente qui devient:

$$\Psi_n = \left(A\pi^{-1/2} n! 2^n \right)^{-1/2} e^{-\varrho^2/2} H_n(Q) \quad (42.471)$$

Soit explicitement:

$$\Psi_n = (-)^n \left(\sqrt{\frac{\hbar}{\omega_0 m \pi}} n! 2^n \right)^{-1/2} e^{\frac{\left(\sqrt{\frac{\omega_0 m}{\hbar}} x \right)^2}{2}} \frac{d^n}{d \left(\sqrt{\frac{\omega_0 m}{\hbar}} x \right)^n} e^{-\left(\sqrt{\frac{\omega_0 m}{\hbar}} x \right)^2} \quad (42.472)$$

Ces polynômes constituent donc une base orthonormée de l'état quantique global et apparaissent donc naturellement dans l'expression générale des fonctions/états propres.

Finalement nous avons:

n	E_n	Ψ_n
0	$\frac{1}{2} \hbar \omega_0$	$\left(\sqrt{\frac{m \omega_0 \pi}{\hbar}} \right)^{1/2} e^{-\frac{\omega_0 m x^2}{\hbar 2}}$
1	$\frac{3}{2} \hbar \omega_0$	$\left(\frac{1}{2} \sqrt{\frac{m \omega_0 \pi}{\hbar}} \right)^{1/2} 2 \sqrt{\frac{\omega_0 m}{\hbar}} x e^{-\frac{\omega_0 m x^2}{\hbar 2}}$
2	$\frac{5}{2} \hbar \omega_0$	$\left(\frac{1}{8} \sqrt{\frac{m \omega_0 \pi}{\hbar}} \right)^{1/2} \left(4 \frac{\omega_0 m}{\hbar} x^2 - 2 \right) e^{-\frac{\omega_0 m x^2}{\hbar 2}}$
3	$\frac{7}{2} \hbar \omega_0$	$\left(\frac{1}{48} \sqrt{\frac{m \omega_0 \pi}{\hbar}} \right)^{1/2} \left(8 \left(\frac{\omega_0 m}{\hbar} \right)^{2/3} x^3 - 12 \sqrt{\frac{\omega_0 m}{\hbar}} x \right) e^{-\frac{\omega_0 m x^2}{\hbar 2}}$
..	...	...

Tableau: 42.2 - Fonctions et énergies propres de l'oscillateur harmonique pour $n=1..3$

Avec la non moins fameuse représentation graphique avec à gauche les fonctions propres associées Ψ_n et à droite la probabilité de présence:

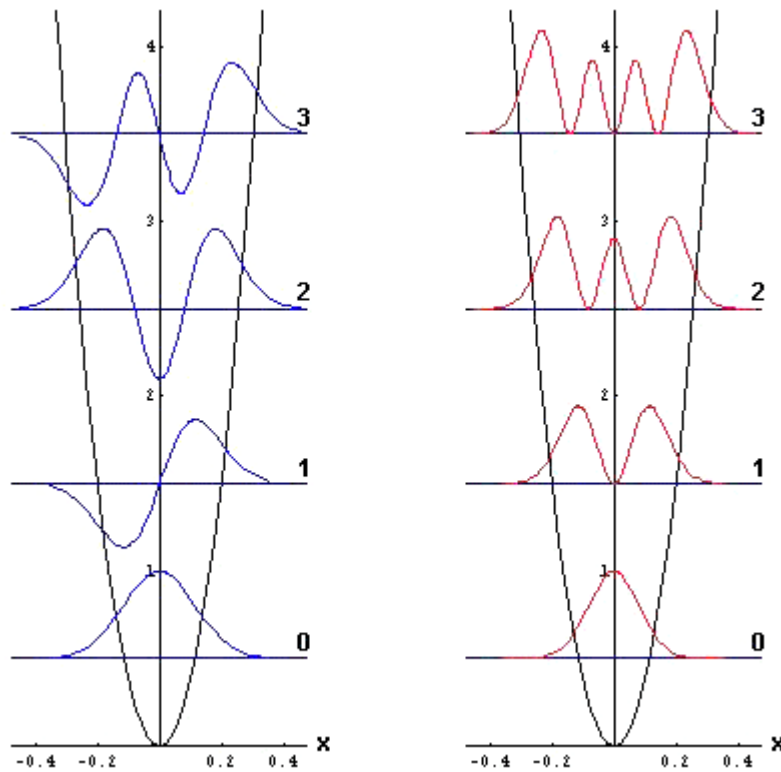


Figure: 42.6 - Représentation des fonctions propres et fonctions de densité de quelques niveaux d'énergie

On peut facilement obtenir la partie de gauche avec Maple 4.00b (l'exemple est non normalisé d'où l'abscisse qui est un peu spéciale par rapport à la figure ci-dessus!):

```
>m:=1;omega:=1;h:=1;
>plot([(sqrt(m*omega*Pi/h))^0.5*exp(-m*omega*x^2/(2*h)),(1/2*sqrt(m*omega*Pi/h))^0.5*sqrt
(m*omega/h)*x*exp(-m*omega*x^2/(2*h)),(1/8*sqrt(m*omega*Pi/h))^0.5*(4*m*omega/h*x^2-2)*exp(-
m*omega*x^2/(2*h)),(1/48*sqrt(m*omega*Pi/h))^0.5*(8*(m*omega/h)^(2/3)*x^3-12*sqrt(m*omega/h)
*x)*exp(-m*omega*x^2/(2*h))],x=-6..6);
```

En analysant ces fonctions d'ondes, nous retrouvons de nombreux résultats classiques: la particule dans le puits de potentiel a une probabilité de présence plus élargie si elle a une énergie plus haute (une bille au fond d'un puits va monter plus haut sur les bords si elle a plus d'énergie), la particule a plus de chance se retrouver sur ces positions éloignées du centre du puits (la bille a une vitesse d'autant plus petite qu'elle est haut dans le puits: elle va donc passer beaucoup plus de temps en hauteur qu'au fond du puits).

Pour tous les calculs où des particules sont dans un puits de potentiel, l'approximation harmonique est très intéressante. Par exemple, si nous souhaitons étudier un "**piège harmonique**" à deux dimensions, soit condensat de Bose-Einstein 2D (cf. chapitre de Mécanique Statistique) nous pourrions poser l'hamiltonien suivant pour débiter l'étude (en analogie avec celui à une dimension utilisé plus haut):

$$H = \frac{p_x^2 + p_y^2}{2m} + \frac{1}{2} m \omega_0^2 (x^2 + y^2) \quad (42.473)$$

4.4.5. EFFET TUNNEL

L'effet tunnel désigne la propriété que possède un objet quantique de franchir une barrière de potentiel, franchissement impossible selon la mécanique classique. Généralement, la fonction d'onde d'une particule, dont le carré du module représente l'amplitude de sa probabilité de présence, ne s'annule pas au niveau de la barrière, mais s'atténue à l'intérieur de la barrière, pratiquement exponentiellement pour une barrière assez large comme nous le démontrerons. Si, à la sortie de la barrière de potentiel, la particule possède une probabilité de présence non nulle, elle peut donc traverser cette barrière.

L'étude théorique de ce phénomène est d'une importance cruciale dans la théorie des semi-conducteurs et de la désintégration en physique nucléaire. Il convient donc d'y accorder une attention bien particulière!

La barrière quantique de largeur L sépare dans les cas simples l'espace en trois, dont les parties gauche et droite sont considérées comme ayant des potentiels constants jusqu'à l'infini. La partie intermédiaire constitue la barrière, qui peut être compliquée, révélant un profil doux, ou au contraire formé de barrières rectangulaires, ou autres éventuellement en séries.

Étudions maintenant le cas de systèmes où l'énergie potentielle E_{pot} (implicitement le potentiel y relatif) tend vers des limites finies, non forcément égales quand $x \rightarrow \pm\infty$. Il s'agit donc d'un problème d'états non liés.

D'abord, nous définissons une région I loin à gauche où $E_{pot}(x)$ sera notée:

$$E_{pot}(-\infty) = E_{pot}^I \quad (42.474)$$

une région III loin à droite où $E_{pot}(x)$ sera notée:

$$E_{pot}(+\infty) = E_{pot}^{III} \quad (42.475)$$

En se bornant aux situations les plus simples, il y a trois possibilités relativement aux relations données précédemment: puits de potentiel (a), marche de potentiel (b), barrière de potentiel (c) comme représentés dans l'ordre énoncé sur la figure ci-dessous:

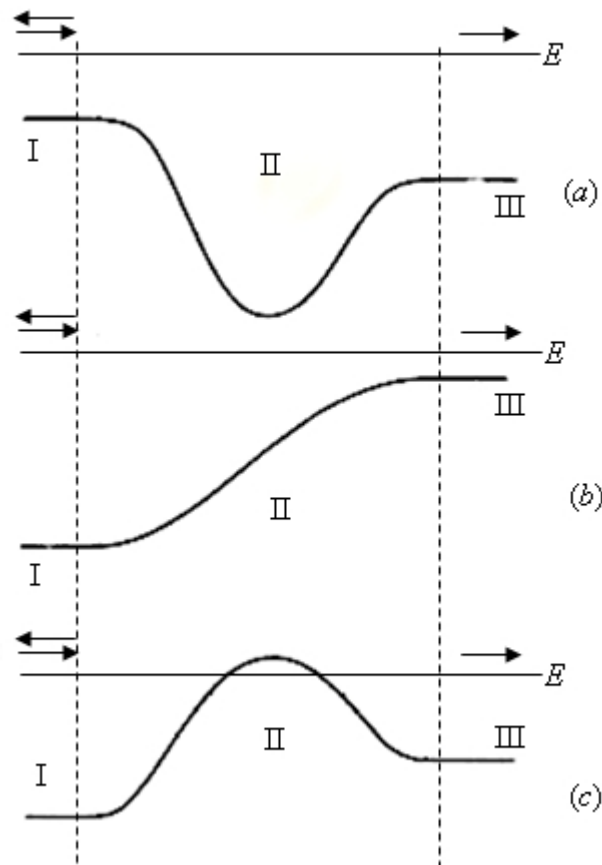


Figure: 42.7 - Types classiques de variation de potentiel

Maintenant, écrivons l'équation de Schrödinger:

$$\Delta\Psi + \frac{2m}{\hbar^2}(E_{\text{tot}} - E_{\text{pot}})\Psi = 0 \quad (42.476)$$

Dans les régions I et III de la barrière de potentiel, l'idée est que $E_{\text{tot}} - E_{\text{pot}}$ est constant et positif donc l'équation différentielle peut s'écrire en une dimension:

$$\frac{\partial^2\Psi}{\partial x^2} + \frac{2m}{\hbar^2}(E_{\text{tot}} - E_{\text{pot}})\Psi = \frac{\partial^2\Psi}{\partial x^2} + k^2\Psi = 0 \quad (42.477)$$

nous obtenons ainsi très simplement l'expression analytique de Ψ dans ces régions sous forme générale:

$$\begin{aligned} \Psi_I &= Ae^{ik_I x} + Be^{-ik_I x} \\ \Psi_{III} &= A' e^{ik_{III} x} + B' e^{-ik_{III} x} \end{aligned} \quad (42.478)$$

Nous trouvons ces deux expressions de façon identique que lors de notre étude du puits de potentiel à parois rectangulaires, à la différence que nous avons écrit ci-dessus les solutions générales de l'équation différentielle (cf. chapitre de Calcul Différentiel Et Intégral) sans en déterminer les coefficients (car nous nous intéressons ici à une généralisation).

Ainsi, dans l'étude du puits à parois rectangulaires plus haut nous avons déjà déterminé que:

$$k_I = \sqrt{\frac{8\pi^2 m}{\hbar^2} (E_{tot} - E_{pot}^I)} \quad \text{et} \quad k_{III} = \sqrt{\frac{8\pi^2 m}{\hbar^2} (E_{tot} - E_{pot}^{III})} \quad (42.479)$$

Remarques:

R1. Nous voyons que les nombres d'ondes k sont donc proportionnels à la racine de l'énergie cinétique. Et comme l'énergie cinétique est proportionnelle à la vitesse au carré des particules, il vient alors que la vitesse est proportionnelle au nombre d'onde (et réciproquement)!

R2. Dans certains ouvrages, pour simplifier les notations, le potentiel dans les régions I et III est posé comme référence et donc égalisé à 0. Il disparaît donc des deux expressions précédentes et cela a pour effet d'égaliser les deux nombres d'ondes qui sont alors notés simplement k .

Dans la région II, l'idée est que $E_{tot} - E_{pot}$ est négatif et constant donc l'équation différentielle peut s'écrire en une dimension:

$$\frac{\partial^2 \Psi}{\partial x^2} + \frac{2m}{\hbar^2} (E_{tot} - E_{pot}) \Psi = \frac{\partial^2 \Psi}{\partial x^2} - \frac{2m}{\hbar^2} (E_{pot} - E_{tot}) \Psi = \frac{\partial^2 \Psi}{\partial x^2} - K^2 \Psi = 0 \quad (42.480)$$

et comme nous l'avons vu lors de notre étude du puits de potentiel rectangulaire infini selon la 2ème approche, la solution est alors de la forme:

$$\Psi_{II} = C e^{Kx} + D e^{-Kx} \quad (42.481)$$

avec:

$$K = \frac{\sqrt{2m(E_{pot} - E_{tot})}}{\hbar} \quad (42.482)$$

Remarque: La parenthèse sous la racine de la relation précédente doit donc être positive. Or cela signifierait que l'énergie cinétique de la particule est négative... Pour pallier à ce problème dans le cadre de ce modèle simplifié, on dit que la particule n'a pas le droit d'exister dans la barrière et qu'elle empreinte de l'énergie au vide. Mais il y a d'autres modèles plus complexes qui ne nécessitent pas ce genre de fantaisies.

Nous obtenons ainsi très simplement l'expression analytique de Ψ dans les trois régions sous forme générale:

$$\begin{aligned} \Psi_I &= A e^{ik_I x} + B e^{-ik_I x} \\ \Psi_{II} &= C e^{Kx} + D e^{-Kx} \\ \Psi_{III} &= A' e^{ik_{III} x} + B' e^{-ik_{III} x} \end{aligned} \quad (42.483)$$

Supposons maintenant que nous ayons à $-\infty$ (région I), une source de particules (qui les envoie vers la droite), avec une énergie cinétique valant évidemment $E_{tot} - E_{pot}^I$.

Ainsi, ces particules ont une énergie E_{tot} et la fonction d'onde qui les décrit obéit à l'équation de Schrödinger. Dans la région III, il sera supposé qu'il ne peut exister que des particules allant vers la droite (pas de source à $+\infty$, par hypothèse).

La région III, comme du reste la région I, est d'étendue infinie, donc le principe d'incertitude nous permet de parler en théorie d'une quantité de mouvement parfaitement déterminée que nous noterons p' .

Nous savons que (c'est de la mécanique classique!) dans la région III nous avons alors:

$$\frac{p'^2}{2m} = E_{tot} - E_{pot}^{III} \quad (42.484)$$

Si $E_{tot} > E_{pot}^{III}$ alors p' est positif, donc grâce à la relation précédente et à la relation de De Broglie nous avons:

$$p' = \sqrt{2m(E_{tot} - E_{pot}^{III})} = \frac{\hbar k_{III}}{2\pi} \quad (42.485)$$

Soit:

$$k_{III} = \frac{\sqrt{2m(E_{tot} - E_{pot}^{III})}}{\hbar} \quad k_I = \frac{\sqrt{2m(E_{tot} - E_{pot}^I)}}{\hbar} \quad (42.486)$$

Les nombres d'onde étant maintenant connus formellement revenons à l'interprétation de la solution III:

$$\Psi_{III} = A' e^{ik_{III}x} + B' e^{-ik_{III}x} \quad (42.487)$$

L'hypothèse comme quoi les particules viennent de la gauche nous impose $B' = 0$ pour que la solution décrive uniquement des particules qui vont vers la droite. Ensuite, il est loisible, pour celles venant de la gauche, de prendre $A' = 1$. La région III est donc relativement simple d'analyse...

Remarque: Les conditions et hypothèses utilisées précédemment sont souvent appelées "**conditions de scattering**".

Les constantes A et B de la région I vont être elles complètement déterminées en effectuant le raccord des solutions d'une région à l'autre.

Intéressons-nous donc maintenant à l'interprétation de l'équation dans la région I:

$$\Psi_I = A e^{ik_I x} + B e^{-ik_I x} \quad (42.488)$$

Il est évident que $A e^{ik_I x}$ décrit des particules qui, dans la région I, se dirigent vers la droite alors $B e^{-ik_I x}$ décrit des particules qui, dans cette même région, se dirigent vers la gauche. Comme nous le savons, les premières sont les particules incidentes, les secondes sont les particules réfléchies.

Ce que nous demandons à la physique quantique apparaît maintenant d'une façon claire: une particule arrivant de la gauche (incidente) peut soit:

1. Continuer vers la droite, c'est-à-dire franchir la région II et devenir une particule transmise
2. Retourner vers la gauche et devenir une particule réfléchie.

Nous sommes amenés à définir un "**coefficient de transmission**" T assimilé à la probabilité qu'a la particule incidente de franchir la région II et un "**coefficient de réflexion**" R , probabilité qu'a la particule incidente d'être réfléchie. Nous devons avoir:

$$R + T = 1 \quad (42.489)$$

Dans le cas d'une barrière de potentiel, T est également appelé la "**transparence de la barrière**".

Pour calculer R et T , nous définirons les flux courants des diverses catégories de particules (incidentes, transmises, réfléchies).

Par exemple, puisque les particules incidentes sont décrites par $Ae^{ik_I x}$, le nombre moyen de ces particules, par unité de longueur dans la région I, doit certainement être proportionnel à un facteur près à $|A|^2$.

Soit v_I leur vitesse, nous voyons que le courant des particules incidentes j_i , est alors proportionnel à un facteur près à $|A|^2 v_I$ (analyse dimensionnelle). Ainsi, le coefficient de proportionnalité étant de même nature pour les trois catégories de particules (incidentes i , réfléchies j , transmises t) et du fait que v_I et v_{III} sont proportionnels à k_I et k_{III} , il s'ensuit que j_i, j_r (courants incident et réfléchi) et j_t (courant transmis) sont respectivement proportionnels (donc toujours à un facteur dimensionnel près!) à $k_I |A|^2$, $k_I |B|^2$ et k_{III} (puisque rappelons que pour la région III nous avons trouvé $A'=1$ et $B'=0$).

Nous déduisons de là très simplement, par un simple rapport, les expressions des coefficients de réflexion R et de transmission T :

$$R = \frac{|B|^2}{|A|^2} = \left| \frac{B}{A} \right|^2 \quad T = \frac{|A'|^2}{|A|^2} = \frac{k_{III}}{k_I} \frac{1}{|A|^2} \quad (42.490)$$

et comme dans notre cas particulier $k_I = k_{III}$ il vient:

$$T = \frac{1}{|A|^2} \quad (42.491)$$

Une autre façon d'écrire les choses est de dire que puisque l'onde incidente se résume à:

$$\Psi^I = A e^{ik_I x} \quad (42.492)$$

et l'onde transmise à:

$$\Psi^T = A' e^{ik_{III} x} \quad (42.493)$$

alors:

$$T = \frac{|\Psi^T|^2}{|\Psi^I|^2} \quad (42.494)$$

Dans toutes ces situations, la théorie quantique conduit, en général, à des valeurs de R et T petites, mais pas nulles !

Exemples:

Déterminons l'expression explicite de la transparence pour notre exemple de barrière rectangulaire.

Pour cela, nous savons que nous devons imposer la continuité de Ψ en $x = 0$ et $x = L$, ainsi que la continuité de $d\Psi/dx$ en $x = 0$ et $x = L$.

Donc rappelons d'abord que nous avons les trois relations (en mettant la référence du potentiel à 0):

$$\begin{aligned} \Psi_I &= Ae^{ikx} + Be^{-ikx} \\ \Psi_{II} &= Ce^{Kx} + De^{-Kx} \\ \Psi_{III} &= A'e^{ikx} + B'e^{-ikx} \end{aligned} \quad (42.495)$$

avec donc:

$$k = \frac{\sqrt{2mE_{tot}}}{\hbar} \quad \text{et} \quad K = \frac{\sqrt{2m(E_{pot} - E_{tot})}}{\hbar} \quad (42.496)$$

Nous avons alors pour la continuité de Ψ en $x = 0$ et $x = L$:

$$\begin{aligned} A + B &= C + D \\ Ce^{KL} + De^{-KL} &= A'e^{ikL} + B'e^{-ikL} \end{aligned} \quad (42.497)$$

ainsi que la continuité de $d\Psi/dx$ en $x = 0$ et $x = L$:

$$\begin{aligned} ikA - ikB &= KC - KD \\ KCe^{KL} - KDE^{-KL} &= ikA'e^{ikL} - ikB'e^{-ikL} \end{aligned} \quad (42.498)$$

Puisque B' est nul nous avons un système de 4 équations à 5 inconnues:

$$\begin{cases} A + B = C + D \\ ikA - ikB = KC - KD \\ Ce^{KL} + De^{-KL} = A'e^{ikL} \\ KCe^{KL} - KDE^{-KL} = ikA'e^{ikL} \end{cases} \Rightarrow \begin{cases} -B + C + D = A \\ ikB + KC - KD = ikA \\ Ce^{KL} + De^{-KL} - A'e^{ikL} = 0 \\ KCe^{KL} - KDE^{-KL} - ikA'e^{ikL} = 0 \end{cases} \quad (42.499)$$

Nous allons choisir d'exprimer toutes les constantes à partir de A . Pour cela nous multiplions la première ligne par ik et la sommons à la deuxième ligne. Nous avons alors:

$$\begin{aligned}
 & -ikB + ikC + ikD = ikA \\
 & + \quad ikB + KC - KD = ikA \quad (42.500) \\
 \hline
 & (K + ik)C - (K - ik)D = 2ikA
 \end{aligned}$$

et ensuite nous multiplions la troisième ligne par $-ik$ et la sommons à la quatrième ligne. Nous avons alors:

$$\begin{aligned}
 & -ikCe^{KL} - ikDe^{-KL} + ikA'e^{ikL} = 0 \\
 & + \quad KCe^{KL} - KDe^{-KL} - ikA'e^{ikL} = 0 \quad (42.501) \\
 \hline
 & (K - ik)e^{KL}C - (K + ik)e^{-KL}D = 0
 \end{aligned}$$

Nous avons donc les deux relations:

$$\begin{aligned}
 & (K + ik)C - (K - ik)D = 2ikA \\
 & (K - ik)e^{KL}C - (K + ik)e^{-KL}D = 0 \quad (42.502)
 \end{aligned}$$

ou en posant $\alpha = k / K$:

$$\begin{aligned}
 & (1 + i\alpha)C - (1 - i\alpha)D = 2i\alpha A \\
 & (1 - i\alpha)e^{KL}C - (1 + i\alpha)e^{-KL}D = 0 \quad (42.503)
 \end{aligned}$$

De la deuxième relation, il vient:

$$D = \frac{(1 - i\alpha)e^{KL}C}{(1 + i\alpha)e^{-KL}} \quad (42.504)$$

et injecté dans la première:

$$(1 + i\alpha)C - (1 - i\alpha) \frac{(1 - i\alpha)e^{KL}C}{(1 + i\alpha)e^{-KL}} = 2i\alpha A \quad (42.505)$$

Soit:

$$C \frac{(1 + i\alpha)^2 e^{-KL} - (1 - i\alpha)^2 e^{KL}}{(1 + i\alpha)e^{-KL}} = 2i\alpha A \quad (42.506)$$

Nous avons alors:

$$C = \frac{2i\alpha(1 + i\alpha)e^{-KL}}{(1 + i\alpha)^2 e^{-KL} - (1 - i\alpha)^2 e^{KL}} A \quad (42.507)$$

et si nous notons:

$$F = \frac{1}{2} \left((1 + i\alpha)^2 e^{-KL} - (1 - i\alpha)^2 e^{KL} \right) \quad (42.508)$$

Il vient alors:

$$C = \frac{i\alpha(1+i\alpha)e^{-KL}}{F} A \quad (42.509)$$

De même en repartant de:

$$\begin{aligned} (1+i\alpha)C - (1-i\alpha)D &= 2i\alpha A \\ (1-i\alpha)e^{KL}C - (1+i\alpha)e^{-KL}D &= 0 \end{aligned} \quad (42.510)$$

De la deuxième relation, il vient:

$$C = \frac{(1+i\alpha)e^{-KL}D}{(1-i\alpha)e^{KL}} \quad (42.511)$$

et injecté dans la première:

$$(1+i\alpha) \frac{(1+i\alpha)e^{-KL}D}{(1-i\alpha)e^{KL}} - (1-i\alpha)D = 2i\alpha A \quad (42.512)$$

Soit:

$$D = \frac{(1+i\alpha)^2 e^{-KL} - (1-i\alpha)^2 e^{KL}}{(1-i\alpha)e^{KL}} = 2i\alpha A \quad (42.513)$$

Nous avons alors:

$$D = \frac{2i\alpha(1-i\alpha)e^{KL}}{(1+i\alpha)^2 e^{-KL} - (1-i\alpha)^2 e^{KL}} A \quad (42.514)$$

et notons toujours:

$$F = \frac{1}{2} \left((1+i\alpha)^2 e^{-KL} - (1-i\alpha)^2 e^{KL} \right) \quad (42.515)$$

Il vient alors:

$$D = \frac{i\alpha(1-i\alpha)e^{KL}}{F} A \quad (42.516)$$

Notez que nous avons aussi:

$$\begin{aligned}
F &= 2i\alpha \cosh(KL) - (1 - \alpha^2) \sinh(KL) = 2i\alpha \frac{e^{KL} + e^{-KL}}{2} - (1 - \alpha^2) \frac{e^{KL} - e^{-KL}}{2} \\
&= \frac{2i\alpha e^{KL} + 2i\alpha e^{-KL} - (1 - \alpha^2)(e^{KL} - e^{-KL})}{2} \\
&= \frac{1}{2} (2i\alpha e^{KL} + 2i\alpha e^{-KL} - e^{KL} + e^{-KL} + \alpha^2 e^{KL} - \alpha^2 e^{-KL}) \\
&= \frac{1}{2} (\alpha^2 e^{KL} + 2i\alpha e^{KL} - e^{KL} + 2i\alpha e^{-KL} - \alpha^2 e^{-KL} + e^{-KL}) \\
&= \frac{1}{2} [(\alpha^2 + 2i\alpha - 1)e^{KL} + (2i\alpha - \alpha^2 + 1)e^{-KL}] \tag{42.517} \\
&= \frac{1}{2} [(2i\alpha - \alpha^2 + 1)e^{-KL} + (\alpha^2 + 2i\alpha - 1)e^{KL}] \\
&= \frac{1}{2} [(2i\alpha - \alpha^2 + 1)e^{-KL} - (-\alpha^2 - 2i\alpha + 1)e^{KL}] \\
&= \frac{1}{2} [(1 + 2i\alpha - \alpha^2)e^{-KL} - (1 - 2i\alpha - \alpha^2)e^{KL}] \\
&= \frac{1}{2} [(1 + i\alpha)^2 e^{-KL} - (1 - i\alpha)^2 e^{KL}]
\end{aligned}$$

Nous pouvons maintenant exprimer les constantes A' et B en fonction de A à l'aide des relations précédentes:

$$\begin{aligned}
B &= C + D - A = \frac{i\alpha(1 + i\alpha)e^{-KL}}{F} A + \frac{i\alpha(1 - i\alpha)e^{KL}}{F} A - A \\
&= A \frac{i\alpha(1 + i\alpha)e^{-KL} + i\alpha(1 - i\alpha)e^{KL} - F}{F} \\
&= A \frac{i\alpha e^{-KL} - \alpha^2 e^{-KL} + i\alpha e^{KL} + \alpha^2 e^{KL} - F}{F} \\
&= A \frac{i\alpha(e^{KL} + e^{-KL}) + \alpha^2(e^{KL} - e^{-KL}) - F}{F} = A \frac{i\alpha 2 \cosh(KL) + \alpha^2 2 \sinh(KL) - F}{F} \tag{42.518} \\
&= A \frac{i\alpha 2 \cosh(KL) + \alpha^2 2 \sinh(KL) - 2i\alpha \cosh(KL) + (1 - \alpha^2) \sinh(KL)}{F} \\
&= A \frac{\alpha^2 2 \sinh(KL) + (1 - \alpha^2) \sinh(KL)}{F} = A \frac{(1 + \alpha^2) \sinh(KL)}{F}
\end{aligned}$$

et:

$$\begin{aligned}
C e^{KL} + D e^{-KL} - A' e^{ikL} &= 0 \Rightarrow A' = e^{-ikL} (C e^{KL} + D e^{-KL}) \\
&= e^{-ikL} \left(\frac{i\alpha(1 + i\alpha)e^{-KL}}{F} A e^{KL} + \frac{i\alpha(1 - i\alpha)e^{KL}}{F} A e^{-KL} \right) \tag{42.519} \\
&= \frac{i\alpha A e^{-ikL}}{F} ((1 + i\alpha) + (1 - i\alpha)) = \frac{i\alpha A e^{-ikL}}{F} 2 = \frac{2i\alpha A e^{-ikL}}{F}
\end{aligned}$$

Donc finalement nous avons:

$$\begin{aligned}
 \Psi_I &= Ae^{ikx} + Be^{-ikx} = Ae^{ikx} + A \frac{(1 + \alpha^2) \sinh(KL)}{F} e^{-ikx} \\
 &= A \left(e^{ikx} + \frac{(1 + \alpha^2) \sinh(KL)}{F} e^{-ikx} \right) \\
 \Psi_{II} &= Ce^{Kx} + De^{-Kx} = \frac{i\alpha(1+i\alpha)e^{-KL}}{F} Ae^{Kx} + \frac{i\alpha(1-i\alpha)e^{KL}}{F} Ae^{-Kx} \quad (42.520) \\
 &= A \left(\frac{i\alpha(1+i\alpha)e^{K(x-L)} + i\alpha(1-i\alpha)e^{-K(x-L)}}{F} \right) \\
 \Psi_{III} &= A' e^{ikx} + B' e^{-ikx} = A' e^{ikx} = \frac{2i\alpha Ae^{-ikL}}{F} e^{ikx} = \frac{2i\alpha Ae^{ik(x-L)}}{F}
 \end{aligned}$$

Et donc alors:

$$T = \frac{|\Psi_I|^2}{|\Psi_I|^2} = \frac{\left| \frac{2i\alpha Ae^{ik(x-L)}}{F} \right|^2}{|Ae^{ikx}|^2} \quad (42.521)$$

en utilisant les propriétés du module complexe (cf. chapitre Nombres):

$$T = \frac{\left| \frac{2i\alpha Ae^{-ikL}}{F} \right|^2 |Ae^{ikx}|^2}{|Ae^{ikx}|^2} = \left| \frac{2i\alpha e^{-ikL}}{F} \right|^2 = |2i\alpha|^2 \left| \frac{e^{-ikL}}{F} \right|^2 = 4\alpha^2 \frac{|e^{-ikL}|^2}{|F|^2} = 4\alpha^2 \frac{1}{|F|^2} \quad (42.522)$$

Il ne nous reste donc plus qu'à calculer:

$$\begin{aligned}
 |F|^2 &= F \cdot \bar{F} = |2i\alpha \cosh(KL) - (1 - \alpha^2) \sinh(KL)| \\
 &= (2i\alpha \cosh(KL) - (1 - \alpha^2) \sinh(KL)) (-2i\alpha \cosh(KL) - (1 - \alpha^2) \sinh(KL)) \\
 &= 4\alpha^2 \cosh^2(KL) - 2i\alpha(1 - \alpha^2) \sinh(KL) \cosh(KL) + 2i\alpha \cosh(KL)(1 - \alpha^2) \sinh(KL) \\
 &\quad + (1 - \alpha^2)^2 \sinh^2(KL) \\
 &= 4\alpha^2 \cosh^2(KL) + (1 - \alpha^2)^2 \sinh^2(KL) \\
 &= 4\alpha^2 \cosh^2(KL) + (1 - 2\alpha^2 + \alpha^4) \sinh^2(KL) \quad (42.523) \\
 &= 4\alpha^2 \cosh^2(KL) + \sinh^2(KL) - 2\alpha^2 \sinh^2(KL) + \alpha^4 \sinh^2(KL) \\
 &= 2\alpha^2 \cosh^2(KL) + \sinh^2(KL) + 2\alpha^2 + \alpha^4 \sinh^2(KL) \\
 &= 2\alpha^2 (1 + \cosh^2(KL)) + (1 + \alpha^4) \sinh^2(KL) = 2\alpha^2 (2 + \sinh^2(KL)) + (1 + \alpha^4) \sinh^2(KL) \\
 &= 4\alpha^2 + 2\alpha^2 \sinh^2(KL) + \sinh^2(KL) + \alpha^4 \sinh^2(KL) \\
 &= 4\alpha^2 + (2\alpha^2 + 1 + \alpha^4) \sinh^2(KL) = 4\alpha^2 + (1 + \alpha^2)^2 \sinh^2(KL)
 \end{aligned}$$

Donc:

$$\begin{aligned}
 T &= 4\alpha^2 \frac{1}{4\alpha^2 + (1 + \alpha^2)^2 \sinh^2(KL)} = \frac{1}{1 + \frac{(1 + \alpha^2)^2}{4\alpha^2} \sinh^2(KL)} \\
 &= \frac{1}{1 + \frac{\left(1 + \frac{k^2}{K^2}\right)^2}{4 \frac{k^2}{K^2}} \sinh^2(KL)} = \frac{1}{1 + \frac{\left(1 + \frac{E_{tot}}{E_{pot} - E_{tot}}\right)^2}{4 \frac{E_{tot}}{E_{pot} - E_{tot}}} \sinh^2(KL)} \\
 &= \frac{1}{1 + \frac{E_{pot} - E_{tot}}{4E_{tot}} \left(1 + \frac{E_{tot}}{E_{pot} - E_{tot}}\right)^2 \sinh^2(KL)} \\
 &= \frac{1}{1 + \frac{E_{pot} - E_{tot}}{4E_{tot}} \left(1 + 2 \frac{E_{tot}}{E_{pot} - E_{tot}} + \frac{E_{tot}^2}{(E_{pot} - E_{tot})^2}\right) \sinh^2(KL)} \quad (42.524) \\
 &= \frac{1}{1 + \frac{1}{4} \left(\frac{E_{pot} - E_{tot}}{E_{tot}} + 2 + \frac{E_{tot}}{E_{pot} - E_{tot}}\right) \sinh^2(KL)} \\
 &= \frac{1}{1 + \frac{1}{4} \frac{(E_{pot} - E_{tot})^2 + 2 E_{tot} (E_{pot} - E_{tot}) + E_{tot}^2}{E_{tot} (E_{pot} - E_{tot})} \sinh^2(KL)} \\
 &= \frac{1}{1 + \frac{E_{pot}^2}{4E_{tot} (E_{pot} - E_{tot})} \sinh^2(KL)}
 \end{aligned}$$

Nous avons donc:

$$T = \frac{1}{1 + \frac{E_{pot}^2}{4E_{tot} (E_{pot} - E_{tot})} \sinh^2(KL)} \quad (42.525)$$

Or, comme:

$$\sinh(KL) = \frac{e^{KL} - e^{-KL}}{2} \quad (42.526)$$

si $KL \gg 1$ (donc à l'échelle atomique c'est plutôt K qui est immense relativement à L) nous avons:

$$\sinh(KL) \cong \frac{e^{KL}}{2} \quad (42.527)$$

Donc:

$$T \cong \frac{1}{1 + \frac{E_{pot}^2 e^{2KL}}{16E_{tot}(E_{pot} - E_{tot})}} \cong \frac{1}{\frac{E_{pot}^2 e^{2KL}}{16E_{tot}(E_{pot} - E_{tot})}} = \frac{16E_{tot}(E_{pot} - E_{tot})}{E_{pot}^2} e^{-2KL} \quad (42.528)$$

relation qu'on retrouve très souvent (sans démonstration détaillée) dans de nombreux ouvrages. Ci-dessous nous avons tracé T :

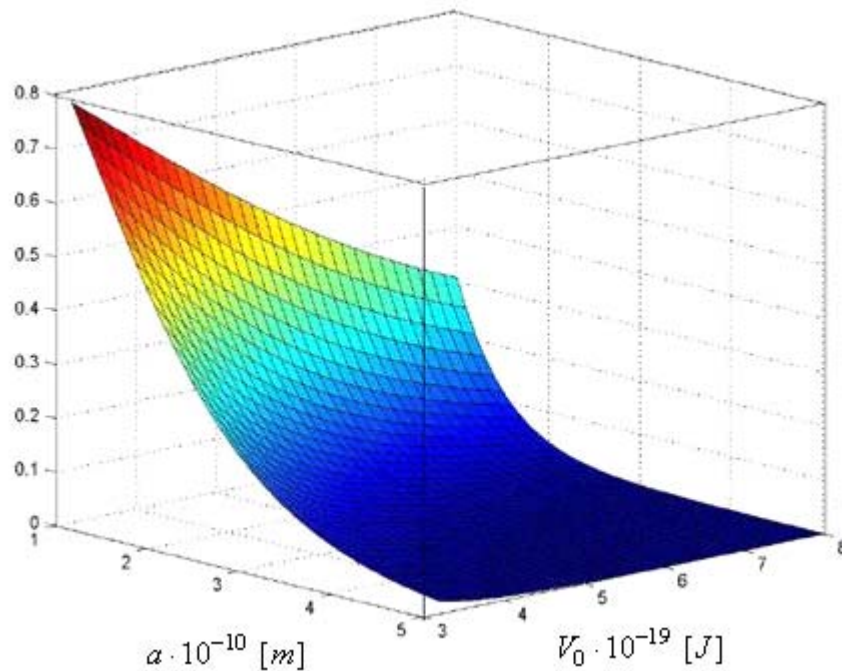


Figure: 42.8 - Représentation graphique du coefficient T avec Matlab 5.0.0.473

selon la relation:

$$T \cong \frac{16E_{tot}(V_0 - E_{tot})}{V_0^2} e^{-2Ka} \quad (42.529)$$

Nous constatons que le coefficient T est très sensible (exponentiellement) à une faible variation la largeur de la barrière, a , lorsque le potentiel de cette barrière est faible. Nous pourrions donc visualiser des sites atomiques, par exemple dans du silicium, en utilisant une pointe très proche du matériau à observer. C'est le principe du microscope à effet tunnel où en approchant une pointe conductrice taillée très finement (quelques atomes seulement) à une proximité d'environ 5 Angströms d'une surface conductrice, et en imposant une différence de potentiel de quelques mV , on mesure un courant de quelques nanoampères. Le nombre d'électrons qui passent à travers la barrière de potentiel (ici c'est le vide entre les deux électrodes conductrices) diminue de manière exponentielle avec la largeur de la barrière. En analysant le signal d'erreur d'un asservissement sur le courant passant dans le circuit, on peut avoir accès à une cartographie très précise de la surface mesurée de l'ordre de 0.1 Angströms en vertical.

Nous remarquons également selon la relation obtenue que les particules légères comme les électrons ont une probabilité plus grande de faire un effet tunnel que les particules plus lourdes à cause du terme de masse.

En utilisant la relation obtenue précédemment, on peut assez simplement calculer la probabilité qu'un être humain de masse m de traverser un mur avec une hauteur h (donc facile de calculer l'énergie potentielle) et une épaisseur a . La probabilité est de l'ordre de $10^{-41} 10^{30} \dots$

Ceci dit, l'exemple le plus célèbre d'effet tunnel pouvant être traité est celui de l'émission de particules α par des noyaux lourds radioactifs dont l'explication a été donnée par le physicien russe G. Gamov en 1928.

La démonstration est relativement simple mais comme elle constitue un cas pratique particulier, nous ne souhaitons pas la détailler dans ce chapitre mais dans celui de Physique Nucléaire. Cependant, pour résoudre ce problème il faut utiliser une méthode d'approximation connue sous le nom de méthode W.K.B. du nom des physiciens Wentzel, Kramers et Brillouin.

Les résultats donnent dès lors un facteur de transmission T pour la particule α de:

$$T \cong 2 \cdot 10^{-39} \quad (42.530)$$

pour l'atome d'Uranium $^{238}_{92}\text{U}$. Par ailleurs, dans l'approximation semi-classique, la particule α a, dans le puits, une vitesse de l'ordre de $10^7 \text{ [ms}^{-1}\text{]}$ et elle effectue des allers-retours dans un noyau dont le rayon est de l'ordre de 10^{-14} [m] . Elle effectue donc environ 10^{21} oscillations par seconde où chaque fois elle a une probabilité T de franchir la barrière de potentiel. Cette probabilité par unité de temps est ainsi déterminée par:

$$\lambda \cong 10^{21} \cdot T \cong 2 \cdot 10^{-18} \text{ [s}^{-1}\text{]} \quad (42.531)$$

Expérimentalement, nous trouvons:

$$\lambda \cong 5 \cdot 10^{-18} \text{ [s}^{-1}\text{]} \quad (42.532)$$

le modèle présenté donne donc des résultats assez satisfaisants.

Outre cet exemple technique, nous rencontrons le phénomène d'effet tunnel aussi dans un cas beaucoup plus accessible et très pédagogique. Ainsi, lorsque sous condition de réflexion totale d'un faisceau de lumière, nous approchons un autre prisme (sur la face du prisme où aucun rayon de lumière ne sort ni ne rentre) de manière à produire une lame d'air suffisamment mince, un faible rayon transmis est observé.

4.5. PRINCIPE DE SUPERPOSITION

La notion d'état dynamique d'un système classique joue un rôle capital dans la dynamique analytique classique.

Est-il possible de retrouver cette notion lorsque nous avons affaire à un système quantique, c'est-à-dire un système tel qu'un atome, un noyau ou une molécule, bref un système de la microphysique?

À première vue non, car nous savons que l'on définit l'état dynamique d'un système classique par la donnée des coordonnées généralisées q_i et des moments conjugués p_i à un instant donné (cf. chapitre de *Mécanique Analytique*). Or, le principe d'incertitude s'oppose à cette procédure dès que nous sommes dans le domaine de la microphysique, vu l'impossibilité de mesurer avec précision les q_i et p_i . Cela est particulièrement clair lorsque le système se réduit à une seule particule que nous décrivons par ses coordonnées cartésiennes $q_i (x, y, z)$ et les composantes de sa quantité de mouvement $\vec{p} (p_x, p_y, p_z)$.

Fort heureusement, il existe une autre définition de l'état dynamique d'un système qui s'applique indifféremment aux systèmes classiques et quantiques et qui, dans le cas des premiers, s'identifie avec la définition habituelle. Nous allons donner cette définition en nous appuyant sur une brève théorie des ensembles de systèmes identiques.

Si nous avons un ensemble (E) d'un très grand nombre de systèmes identiques, nous ferons une enquête statistique pour caractériser cet ensemble de la façon suivante: nous prenons un système de l'ensemble, nous mesurons une variable dynamique (coordonnée, composante de quantité de mouvement, énergie cinétique, etc.) et nous rejettons le système (qui perturbé par la mesure, ne doit pas être réincorporé à l'ensemble). Nous dressons ainsi un bilan qui se traduit par des fonctions de distribution de toutes les variables dynamiques possibles. Cela permet de définir sans ambiguïté la notion d'identité:

Définition: Deux ensembles sont identiques, si les bilans des résultats des mesures sont les mêmes pour les deux.

Considérons maintenant un ensemble unique (E). Est-il possible de le réaliser par juxtaposition de deux ensembles non identiques (E_1) et (E_2)? Ce qui permettrait d'écrire:

$$(E) = (E_1) + (E_2) \quad (42.533)$$

Si oui, nous dirons que (E) est un mélange. Inversement, au moyen d'un tri convenable, un mélange peut être décomposé en deux sous-ensembles différents. Si non, nous dirons que (E) est un ensemble pur. Tout tri décomposera l'ensemble pur en deux sous-ensembles identiques entre eux et nécessairement avec (E) ! Nous convenons alors de dire que tous les systèmes d'un ensemble pur sont dans le même état dynamique et que deux ensembles purs différents donnent lieu à des états dynamiques différents. Il va de soi que les systèmes constituant un mélange seront eux dans des états dynamiques différents.

Supposons maintenant que les systèmes étudiés obéissent aux lois de la mécanique classique. Si les systèmes d'un ensemble présentent des jeux (q_i, p_i) différents, nous les trions en les groupant par systèmes ayant tous un même jeu (q_i, p_i) . Nous vérifions bien que la nouvelle définition de l'état dynamique coïncide avec la définition habituelle. Notons ce fait évident, mais important (par opposition avec les systèmes quantiques): dans un ensemble pur de systèmes classiques, c'est-à-dire pour un état dynamique donné, toute variable dynamique est bien déterminée. En effet, en mécanique analytique classique, une telle variable est une fonction des q_i et p_i et, de ce fait, présente une valeur unique.

Passons aux systèmes quantiques. Il est maintenant possible de définir pour ceux-ci un état dynamique, mais tout de suite nous voyons une distinction fondamentale avec la mécanique classique. En effet, dans un ensemble pur de systèmes quantiques, c'est-à-dire pour un état dynamique donné, une variable dynamique n'est pas, en général, bien déterminée. Quand nous la mesurons sur des systèmes extraits de l'ensemble pur, on ne trouve généralement pas comme résultat, une valeur unique, mais une distribution de valeurs.

L'indétermination qui règne sur la valeur d'une variable dynamique dans un état dynamique donné est donc de nature purement quantique et il convient de bien la distinguer de l'indétermination d'origine statistique qui se manifeste dans un mélange, qu'il s'agisse de systèmes classiques ou quantiques.

Le formalisme de la physique quantique ne peut s'édifier que si nous savons décrire mathématiquement les états dynamiques et les variables dynamiques. Nous avons vu que nous ne pouvons attendre de ce formalisme une prédiction précise comme en mécanique classique, mais, simplement les probabilités d'obtenir telle ou telle valeur, lorsque nous mesurons une variable dynamique sur un système dont l'état dynamique est donné.

Toute la théorie que nous avons vue jusqu'ici nous permet de conclure jusqu'ici que les états dynamiques d'un système d'une particule sans spin sont décrits par des fonctions d'onde complexes, non nulles partout.

Si nous appliquons cette condition aux systèmes dynamiques, nous avons alors le postulat suivant:

Soient deux états dynamiques différents, décrits par des fonctions d'onde Ψ_1 et Ψ_2 , nécessairement non proportionnelles. λ_1, λ_2 étant des nombres complexes non simultanément nuls, nous construisons la combinaison linéaire:

$$\Psi = \lambda_1 \Psi_1 + \lambda_2 \Psi_2 \quad (42.534)$$

Ψ est alors une fonction d'onde décrivant un état dynamique possible du système. Ce qui s'écrit aussi souvent avec le formalisme de Dirac sous la forme suivante:

$$|\Psi\rangle = \lambda_1 |\Psi_1\rangle + \lambda_2 |\Psi_2\rangle \quad (42.535)$$

Ce postulat paraît assez naturel du fait de l'aspect ondulatoire que présente la physique des microsystèmes. En effet, dans les phénomènes ondulatoires de la physique classique les équations d'onde sont, le plus souvent, linéaires homogènes et il s'ensuit que l'on peut superposer les ondes. Or, le grand intérêt de ce postulat est qu'il contient en germe l'explication de ce fait capital qu'est l'indétermination quantique (appelée aussi parfois "cohérence quantique").

Voyons-le sur un cas très simple où nous supposons qu'une variable dynamique A , a une valeur bien définie a_1 dans l'état dynamique (Ψ_1) , et une valeur bien définie a_2 dans l'état dynamique (Ψ_2) avec $a_2 \neq a_1$. Cela signifie que si nous répétons la mesure de A sur des systèmes tous dans l'état dynamique décrit par (Ψ_1) , nous trouvons chaque fois comme résultat a_1 , de même pour Ψ_2 et a_2 .

Une question vient naturellement à l'esprit: si nous mesurons A sur des systèmes tous dans l'état dynamique (Ψ) qu'allons nous obtenir? Une idée naïve serait de croire que A prendra une valeur bien définie intermédiaire entre a_1 et a_2 .

Ces deux hypothèses sont fausses et nous le savons bien. Premièrement, A n'est pas bien déterminée en physique quantique (incertitude) et n'est mathématiquement pas nécessairement située entre a_1 et a_2 . L'interprétation correcte est la suivante:

Si nous mesurons A sur le système dans l'état dynamique (Ψ) , nous trouvons comme résultat de mesure, tantôt a_1 , avec une probabilité ϖ_1 , tantôt a_2 , avec une probabilité $\varpi_2 = 1 - \varpi_1$. Bien entendu, ϖ_1 et ϖ_2 devront pouvoir être calculés en fonction de λ_1 et λ_2 .

Remarque: Il ne faut surtout pas confondre l'ensemble pur des systèmes décrits par (Ψ) , avec le mélange que nous obtiendrions en juxtaposant deux ensembles purs de systèmes respectivement (Ψ_1) et (Ψ_2) .

Il convient donc de mettre en garde le lecteur contre cette confusion, d'autant que dans la littérature courante utilisant la physique quantique, on dit souvent que la fonction d'onde Ψ est un mélange de Ψ_1 et Ψ_2 . C'est par exemple dans ce sens que nous parlons de "**mélange de configurations**" pour traduire le fait que la fonction d'onde d'un atome à plusieurs électrons est une combinaison linéaire de fonctions d'onde appartenant à diverses configurations. Cette terminologie ne doit pas cacher le fait que les systèmes décrits par Ψ constituent un ensemble pur et non un mélange.

Exemple:

Considérons le cas scolaire suivante de superposition:

$$|\Psi\rangle = \frac{\sqrt{3}}{3}|\Psi_1\rangle + \frac{2}{3}|\Psi_2\rangle + \frac{\sqrt{2}}{3}|\Psi_3\rangle \quad (42.536)$$

Nous pouvons déjà aisément vérifier que le système est normalisé sous l'hypothèse $\langle\Psi_i|\Psi_j\rangle = \delta_{ij}$ donc de normalité:

$$|\langle\Psi|\Psi\rangle|^2 = \left| \frac{\sqrt{3}}{3}\langle\Psi_1|\Psi_1\rangle + \frac{2}{3}\langle\Psi_2|\Psi_2\rangle + \frac{\sqrt{2}}{3}\langle\Psi_3|\Psi_3\rangle \right|^2 = \left| \frac{1}{3} + \frac{4}{9} + \frac{2}{9} \right|^2 = |1|^2 = 1 \quad (42.537)$$

Comme la superposition est normalisée, il vient alors la probabilité de trouver le système dans l'état Ψ_1 qui est alors:

$$|\langle\Psi_1|\Psi\rangle|^2 = \left| \frac{\sqrt{3}}{3} \underbrace{\langle\Psi_1|\Psi_1\rangle}_{=1} + \frac{2}{3} \underbrace{\langle\Psi_1|\Psi_2\rangle}_{=0} + \frac{\sqrt{2}}{3} \underbrace{\langle\Psi_1|\Psi_3\rangle}_{=0} \right|^2 = \left| \frac{\sqrt{3}}{3} \right|^2 = \frac{1}{3} \quad (42.538)$$

et si nous faisons de même pour chacun des deux autres états, la somme des probabilités sera toujours égal à 1. Il convient de signaler que cette probabilité est aussi la proportion d'états qui seront mesurés dans l'état Ψ_1 si le système est constitué de N composants identiques.

En fait, l'interprétation que donne la théorie de De Broglie (associer une fonction d'onde à une particule) aux principes d'incertitudes est l'exemple le plus frappant et le plus connu de la physique quantique au niveau des superpositions d'états (chat de Schrödinger mis à part):

Considérons une onde de De Broglie se propageant dans le sens de l'axe X , mais limitée à un intervalle $(-a, +a)$ à un instant donné ($t = 0$ si nous voulons). Donc à $t = 0$ l'onde s'écrit, en laissant tomber la constante multiplicative:

$$\begin{aligned}\Psi(x) &= e^{ik_0 x} \text{ pour } |x| < a \\ \Psi(x) &= 0 \text{ pour } |x| > a\end{aligned} \quad (42.539)$$

Si nous mesurons la coordonnée de la particule, nous devons la trouver là nécessairement où Ψ n'est pas nulle (sinon nous ne pourrions rien mesurer). Nous pouvons dire que $x = 0$ avec une incertitude $\Delta x = a$ (l'intervalle où nous sommes sûrs de trouver la particule par rapport à l'ordonnée à l'origine divisé par deux)

Si nous mesurons p , que trouvons-nous ? Nous ne devons pas trouver $\hbar k_0$ (relation que nous avons déjà démontrée plus haut), car ceci serait vrai pour une onde plane indéfinie, ce qui n'est pas le cas ici. Alors, nous allons décomposer l'onde en ondes planes au moyen de la transformation de Fourier (cf. chapitre de Suites et Séries):

$$\Psi(x) = \int_{-\infty}^{+\infty} a(k) e^{ikx} dk \quad (42.540)$$

Comment interpréter cette relation? Une des ondes planes élémentaires (que nous pouvons aussi interpréter comme un état), $a(k) e^{ikx} dk$, dont la somme redonne $\Psi(x)$, conduit à une valeur $p = \hbar k$ de la quantité de mouvement. Or, les valeurs de k forment un continuum. Nous sommes conduits à dire que les valeurs possibles de p forment dès lors aussi un continuum et qu'il y a donc une incertitude sur la valeur de p . Pour aller plus loin, il faut évaluer $a(k)$ (qui doit être considéré comme variable de la probabilité de présence de chaque onde plane provenant de la décomposition de $\Psi(x)$) au moyen de la relation (selon les propriétés des transformations de Fourier):

$$a(k) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Psi(x) e^{-ikx} dx \quad (42.541)$$

qui donne ici:

$$a(k) = \frac{1}{2\pi} \int_{-a}^{+a} e^{-i(k_0 - k)x} dx \quad (42.542)$$

Posons $k_0 - k = \Delta k$, l'intégrale devient alors:

$$a(k) = \frac{1}{2\pi} \int_{-a}^{+a} e^{-i\Delta k x} dx = \frac{1}{2\pi} \left[\frac{1}{-i\Delta k} e^{-i\Delta k x} \right]_{-a}^{+a} = \frac{1}{-i\pi\Delta k} \left[\frac{e^{i(-a\Delta k)} - e^{i(a\Delta k)}}{2i} \right] = \frac{a}{\pi} \frac{\sin(a\Delta k)}{a\Delta k} \quad (42.543)$$

Le graphique de la fonction $\sin(u)/u$ montre que $a(k)$ prend des valeurs qui peuvent être considérées comme négligeables pour $|u| > \pi$.

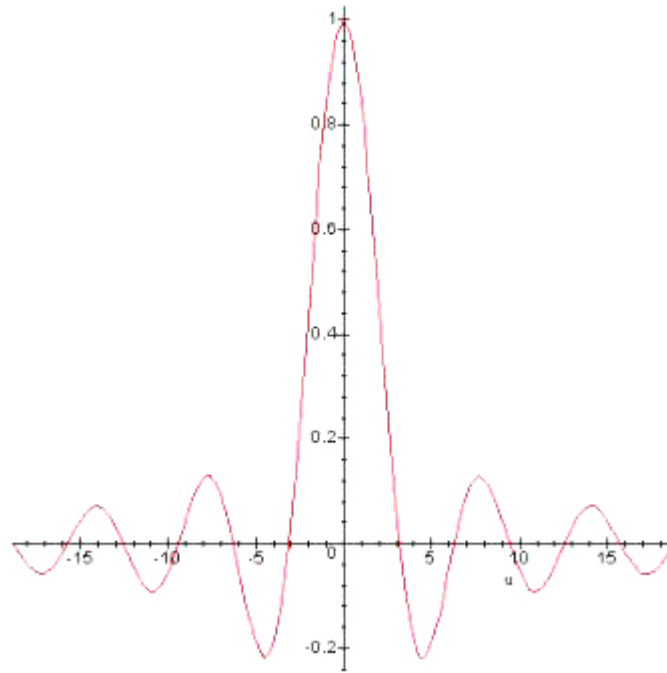


Figure: 42.9 - Tracé de la fonction sinus cardinal

Il s'ensuit que dans l'intégrale:

$$\Psi(x) = \int_{-\infty}^{+\infty} a(k) e^{ikx} dk \quad (42.544)$$

ce sont les k voisins de k_0 qui sont effectifs, et plus précisément les k tels que:

$$|k - k_0| a < \pi \quad (42.545)$$

puisque:

$$a(k) = \frac{a}{\pi} \frac{\sin(a\Delta k)}{a\Delta k} = \frac{a}{\pi} \frac{\sin(a(k - k_0))}{a\Delta k} \quad (42.546)$$

Il s'ensuit que les valeurs à retenir de p sont celles voisines de $\hbar k_0$ aussi, plus précisément nous avons:

$$|p - \hbar k_0| a < \pi \hbar \quad (42.547)$$

Cette relation montre que les incertitudes Δx et Δp obéissent à la relation:

$$\Delta p \Delta x = \pi \hbar = \frac{h}{2} \quad (42.548)$$

De manière similaire, si nous nous proposons de déterminer la coordonnée x d'un électron en le faisant passer à travers une fente de largeur $2b$ percée dans un écran:

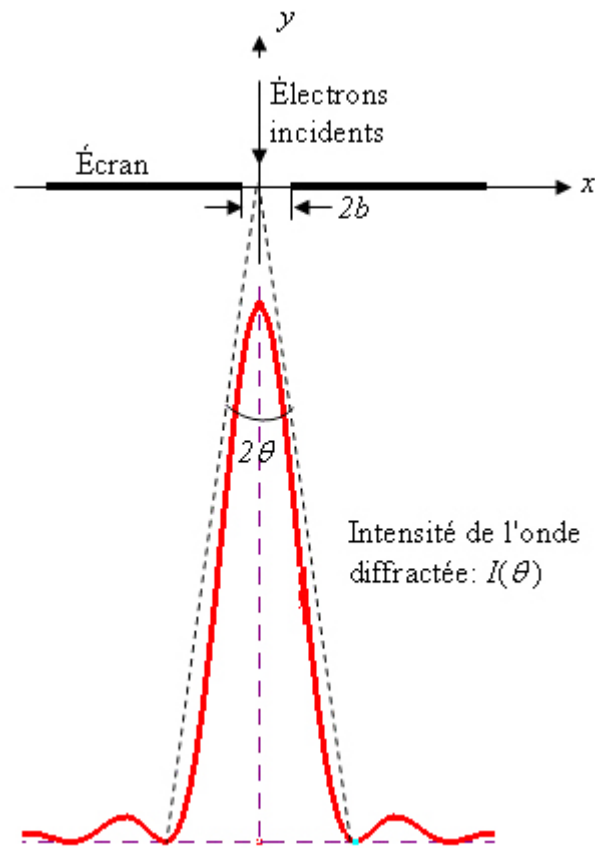


Figure: 42.10 - Configuration d'étude de l'électron passant par une fente rectangulaire

La précision avec laquelle nous connaissons la position de cet électron est limitée par la taille de la fente, soit $\Delta x = b$. D'autre part, la fente perturbe l'onde associée. Il en résulte une modification du mouvement de l'électron qui se traduit par le diagramme de diffraction de l'onde (qui est en fait une représentation de la superposition linéaire de ses états intrinsèques).

L'incertitude sur la composante dynamique p_x de la quantité mouvement de l'électron est déterminée par l'angle θ correspondant au maximum central de la figure de diffraction. D'après la théorie de la diffraction (cf. chapitre d'Optique Ondulatoire) produite par une fente rectangulaire, nous avons $\sin \theta = \lambda / 2b$ puisque l'intensité $I(\theta)$ s'écrit:

$$I(\theta) = I(0) \sin^2 \left(\frac{1}{\left[\frac{2\pi b}{\lambda} \sin \theta \right]} \right) \quad (42.549)$$

Donc p_x est compris entre $p \sin \theta$ et $-p \sin \theta$, p étant l'impulsion de l'électron incident. Ainsi l'incertitude Δp_x est de:

$$\Delta p_x = p \sin \theta = \frac{h}{\lambda} \frac{\lambda}{2b} = \frac{h}{2} \cdot \frac{1}{b} \quad (42.550)$$

Ce résultat simple est assez extraordinaire si nous le mettons en relation, en ordre de grandeur, avec le résultat que nous avons obtenu juste plus haut:

$$\Delta p \Delta x = \pi \hbar = \frac{h}{2} \quad (42.551)$$

Nous pouvons en tirer plusieurs conclusions de la première importance:

1. L'onde associée de De Broglie est étroitement liée au principe d'incertitude et la physique quantique doit tenir compte simultanément de ces deux propriétés.
2. Si nous tenons compte que la répartition de l'intensité est obtenue à partir du comptage des électrons (ou particules) en fonction de l'angle et que nous obtenons la même répartition quelle que soit l'intensité du faisceau d'électrons monocinétiques qui arrive sur la fente et ce, même si les électrons sont envoyés un par un. Nous observons alors que le mouvement des particules n'est plus déterministe mais probabiliste. Ainsi, la fonction d'onde Ψ de l'électron peut être considérée comme une superposition linéaire des états définis chacun comme nous l'avons fait précédemment, par sa décomposition spectrale possible par la transformée de Fourier.

Que pouvons-nous conclure de tout ce que nous avons vu jusqu'ici:

1. Les équations de la physique quantique nous donnent une densité de probabilité de trouver une particule dans un certain volume de l'espace-temps.
2. La superposition linéaire des états peut s'interpréter comme le fait qu'il est possible de trouver une particule en plusieurs points de l'espace-temps à un instant donné, et avec pour chacun de ces points une certaine probabilité de l'y trouver (par décomposition possible de l'équation d'onde).

Si le point (1) a été largement étudié jusqu'à maintenant sur ce site, le point (2) est quant à lui nouveau et découle d'une simple opération mathématique de décomposition ou de superposition.

Mais dès lors, que se passe-t-il si nous cherchons à mesurer l'énergie d'un atome qui se trouve dans une superposition d'états d'énergie? Nous ne détecterons jamais cette superposition, mais seulement l'une des énergies qui la constituent, l'action de mesurer fait disparaître la superposition des états au profit d'un seul - nous parlons alors de "**décohérence quantique**" (il s'agit de l'interprétation de Copenhague dont nous avons fait mention implicitement au tout début de ce chapitre et sur lequel nous reviendrons). Mais lequel? La physique quantique ne peut tout bonnement répondre à cette question. Le choix s'effectue au hasard! En revanche, à défaut de prédire l'état précis qui sera mesuré parmi tous ceux qui constituaient la superposition, la théorie quantique peut donner la probabilité qu'on a de mesurer chaque état (ce que l'on a déjà fait maintes fois jusqu'ici). Si l'on effectue de nombreuses mesures, on trouve finalement les proportions prédites par la théorie (même si chaque mesure est imprévisible).

Erwin Schrödinger, avait souligné l'absurdité (selon lui) de ces superpositions en ayant recours à une expérience de pensée devenue célèbre: Imaginez un chat enfermé dans une boîte hermétique. Dans la boîte se trouve aussi un atome radioactif et un dispositif capable de répandre du poison. Quand l'atome radioactif se désintègre, il déclenche le dispositif mortel: le poison se répand dans la boîte et le chat meurt.

Mais la désintégration radioactive est un phénomène quantique: tant que nous ne l'avons pas détecté, l'atome est dans une superposition d'états "désintégré et pas désintégré". Dans la boîte, le système chat-dispositif à poison-atome doit donc lui aussi, se trouver dans une superposition des deux états "atome désintégré-chat mort" et "atome intact-chat vivant". Bref, si nous prenons la physique quantique au pied de la lettre, le chat est à la fois mort et vivant tant que la mesure n'a pas été effectuée.

L'absurdité de cette expérience est manifeste... mais difficile à démontrer, du moins tant que nous n'avons pas compris ce qui distingue un chat d'une particule. Toujours le problème de la frontière quantique-classique...

Il faudra attendre les années 80 pour que la situation progresse enfin, à la fois sur le front de l'expérience et sur celui de la théorie. En 1982, Wojciech Zurek, chercheur au laboratoire national de Los Alamos (Nouveau-Mexique), reprend une idée fort simple mais géniale: dans une mesure, ce qui produit la décohérence, c'est l'interaction du système avec son environnement. Plus généralement, les objets quantiques ne sont jamais complètement isolés de leur environnement - nous entendons par là tout ce qui interagit avec le système: un appareil, des molécules d'air, des photons lumineux. Si bien qu'en réalité les lois quantiques doivent s'appliquer à l'ensemble constitué de l'objet et de tout ce qui l'entoure. Or, Zurek démontre que les multiples interactions avec l'environnement entraînent une destruction très rapide des de la cohérence quantique des superpositions d'états (appelée également "interférence quantique" puisque mathématiquement l'on traite des fonctions d'onde). En détruisant les interférences, l'environnement supprime les superpositions d'états et le comportement quantique du système, de sorte qu'il ne reste plus que des états simples et qu'on retrouve le comportement classique.

Dans un objet macroscopique - un chat par exemple... - chacun des atomes est environné de nombreux autres atomes qui interagissent avec lui. Toutes ces interactions provoquent spontanément un brouillage des interférences quantiques qui disparaissent très vite. Voilà donc pourquoi la physique quantique ne s'applique pas à notre échelle: les systèmes ne sont jamais isolés!

La vitesse de la décohérence augmente avec la taille du système: un chat qui compte 10^{27} particules, "décohère" en 10^{-23} secondes, ce qui explique pourquoi on n'a jamais vu de chats morts-vivants jusqu'à aujourd'hui!

La physique quantique est donc une théorie:

- non-déterministe (probabiliste) d'où le fait qu'elle soit considérée comme une théorie de l'information
- non-locale: les objets quantiques peuvent avoir simultanément plusieurs positions
- non-séparable: plusieurs objets quantiques peuvent être superposés au point de ne pouvoir être considérés séparément.

Un autre excellent exemple de la superposition linéaire des états est une application remarquable au principe de moindre action.

Considérons une particule quantique allant d'un point M_1 à l'instant t_1 au point M_2 à l'instant t_2 . Nous savons que la probabilité de trouver une particule en un point et en un instant donnés est reliée au carré du module de la fonction d'onde qui lui est associée. Plaçons-nous dans le cas le plus simple où la fonction d'onde de la particule est une onde plane $\Psi(x,t)$ donnée par la fonction solution de l'équation d'évolution de Schrödinger:

$$\Psi(x,t) = e^{2\pi i \left(\frac{x}{\lambda} - \nu t \right)} \quad (42.552)$$

où λ et ν sont respectivement la longueur d'onde et la fréquence de l'onde associée à la particule.

La particule peut emprunter une infinité de chemins pour se rendre de (M_1, t_1) à (M_2, t_2) . Choisissons l'un quelconque de ces chemins que nous appellerons C . Nous pouvons découper le chemin C en un nombre entier de tronçons de durée Δt .

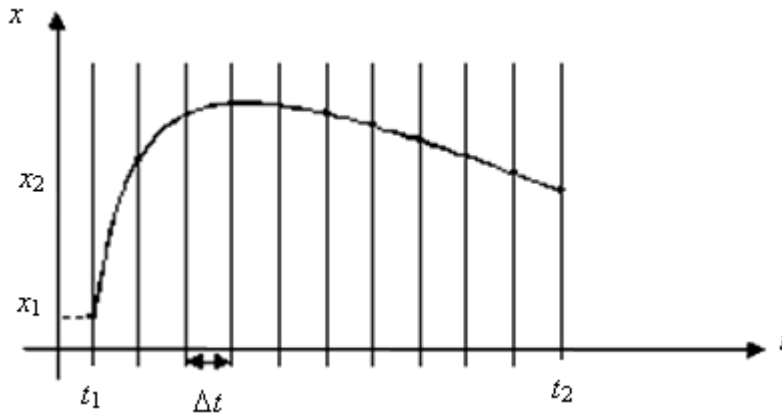


Figure: 42.11 - Principe de découpage du chemin

Après le parcours du premier tronçon, la fonction d'onde a la valeur suivante:

$$\Psi(x_1 + \Delta x, t_1 + \Delta t) = e^{2\pi i \left(\frac{x_1 + \Delta x}{\lambda} - \nu(t_1 + \Delta t) \right)} = e^{2\pi i \left(\frac{x_1}{\lambda} - \nu t_1 \right)} e^{2\pi i \left(\frac{\Delta x}{\lambda} - \nu \Delta t \right)} \quad (42.553)$$

D'où nous tirons que:

$$\Psi(x_1 + \Delta x, t_1 + \Delta t) = \Psi(x_1, t_1) \cdot e^{2\pi i \left(\frac{\Delta x}{\lambda} - \nu \Delta t \right)} \quad (42.554)$$

Or, Planck et De Broglie ont établi (postulé) les relations suivantes comme nous l'avons montré:

$$E = h \cdot \nu \text{ et } \lambda = \frac{h}{p} \quad (42.555)$$

d'où, en remplaçant λ et ν dans la relation précédente, nous obtenons:

$$\Psi(x_1 + \Delta x, t_1 + \Delta t) = \Psi(x_1, t_1) \cdot e^{\frac{2\pi i}{h}(p\Delta x - E\Delta t)} = \Psi(x_1, t_1) \cdot e^{\frac{2\pi i}{h} \left[\left(p \frac{\Delta x}{\Delta t} - E \right) \Delta t \right]} \quad (42.556)$$

En appliquant la même technique pour le tronçon suivant nous obtenons:

$$\begin{aligned} \Psi(x_1 + 2\Delta x, t_1 + 2\Delta t) &= \Psi(x_1 + \Delta x, t_1 + \Delta t) \cdot e^{\frac{2\pi i}{h} \left[\left(p \frac{\Delta x}{\Delta t} - E \right) \Delta t \right]} \\ &= \Psi(x_1, t_1) \cdot e^{\frac{2\pi i}{h} \left[\left(p \frac{\Delta x}{\Delta t} - E \right) \Delta t \right]_{\text{tronçon 1}}} \cdot e^{\frac{2\pi i}{h} \left[\left(p \frac{\Delta x}{\Delta t} - E \right) \Delta t \right]_{\text{tronçon 2}}} \end{aligned} \quad (42.557)$$

Procédant ainsi de tronçon en tronçon, tout le long du chemin C nous obtenons alors la valeur de la fonction d'onde en (M_2, t_2) pour la particule venant de (M_1, t_1) en suivant le chemin C :

$$\Psi_C(x_2, t_2) = \Psi(x_1, t_1) \cdot \prod_{t_1}^{t_2} e^{\frac{2\pi i}{h} \left[\left(p \frac{\Delta x}{\Delta t} - E \right) \Delta t \right]} = \Psi(x_1, t_1) \cdot e^{\frac{2\pi i}{h} \left[\sum_{t_1}^{t_2} \left(p \frac{\Delta x}{\Delta t} - E \right) \Delta t \right]} \quad (42.558)$$

Maintenant, faisons tendre la durée dt de chaque tronçon de trajectoire vers zéro. La quantité $\Delta x / \Delta t$ tend alors vers la vitesse instantanée de la particule que nous noterons $\dot{x}$. La relation précédente devient alors:

$$\Psi_C(x_2, t_2) = \Psi(x_1, t_1) \cdot e^{\frac{2\pi i}{\hbar} \left[\int_{t_1}^{t_2} (p\dot{x} - E) dt \right]} \quad (42.559)$$

Dans le chapitre de Mécanique Analytique, nous avons montré que la quantité $p \cdot \dot{x} - E$ est égale au lagrangien. En substituant le lagrangien dans la relation précédente, nous obtenons:

$$\Psi_C(x_2, t_2) = \Psi(x_1, t_1) \cdot e^{\frac{2\pi i}{\hbar} \left[\int_{t_1}^{t_2} L(t) dt \right]} = \Psi(x_1, t_1) \cdot e^{\frac{2\pi i}{\hbar} S_C} \quad (42.560)$$

où S_C est l'action de la particule ayant parcouru le chemin C .

Notons (sans démonstration) que le module de $\Psi_C(x_2, t_2)$ prend la même valeur pour:

$$S_C = 2n\pi\hbar = n\hbar \quad (42.561)$$

pour tout n . La constante de Planck trouve alors une signification physique directement liée à l'action de la particule !

Rappelons la condition de normalisation de De Broglie:

$$|\Psi_C(x_2, t_2)|^2 \quad (42.562)$$

qui donne donc la probabilité pour que la particule, partant de x_1 à l'instant t_1 , se trouve en x_2 à l'instant t_2 en ayant emprunté le chemin C .

La probabilité totale est donc:

$$|\Psi_{totale}(x_1, t_1 \rightarrow x_2, t_2)|^2 \quad (42.563)$$

pour trouver la particule partie de x_1 à l'instant t_1 en x_2 à l'instant t_2 nécessite de calculer la somme des contributions de chaque chemin soit (en appliquant le principe de superposition linéaire puisque nous effectuons une somme des fonctions d'onde):

$$\Psi_{totale}(x_1, t_1 \rightarrow x_2, t_2) = \int_{\text{tous les chemins}} \Psi_C(x_2, t_2) = \int_{\text{tous les chemins}} \Psi_C(x_1, t_1) \cdot e^{\frac{2\pi i}{\hbar} S_C} \quad (42.564)$$

Cette intégrale fut découverte par Richard Feynman. En première analyse elle semble diverger dans la mesure où il existe une infinité de chemins possibles entre deux points. Regardons de plus près ce qui se passe. Plaçons-nous dans le cas où la trajectoire est macroscopique. La valeur de l'action S_C est alors beaucoup plus grande que $\hbar$ et varie beaucoup d'un chemin à un autre, sauf pour les chemins proches du chemin physique classique pour lesquels la variation est quasiment nulle (application de l'énoncé variationnel du principe de moindre action).

Comme les actions des chemins interviennent comme une phase dans l'intégrale de chemin, leurs contributions sont destructives et donc tendent à s'annuler, sauf dans le cas des chemins proches du chemin physique classique où les contributions s'ajoutent. Il s'ensuit que l'intégrale de chemin prend la valeur de l'action classique, indiquant que la physique quantique permet de retrouver les lois de la mécanique classique à l'échelle macroscopique.

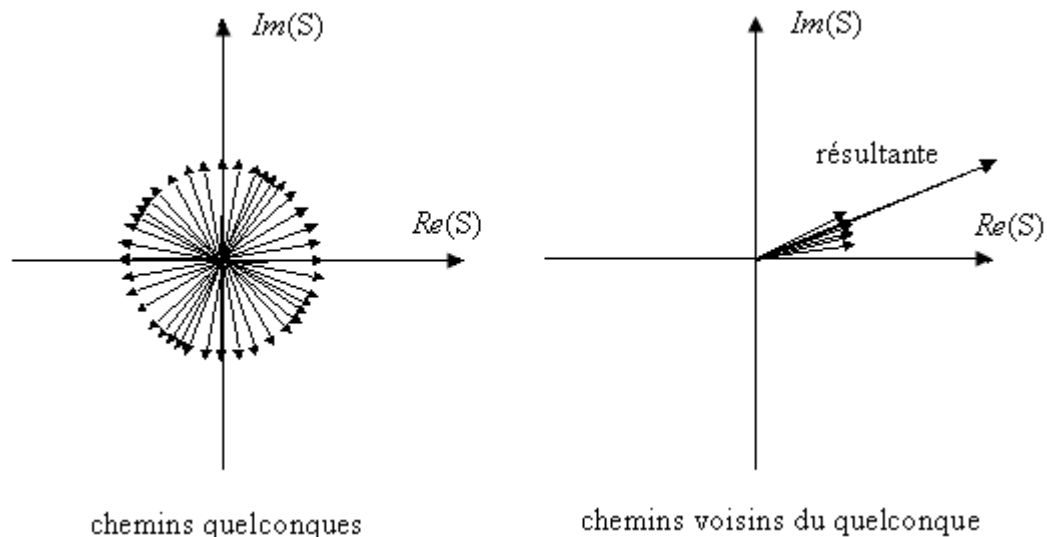


Figure: 42.12 - Représentation des phases en fonction du type de chemin

La situation devient très différente à l'échelle quantique, c'est-à-dire pour des valeurs de l'action dont l'ordre de grandeur est celui de la constante $\hbar$. Une infinité de chemins apporte alors des contributions non destructives. Feynman a pu montrer que l'intégrale de chemin convergeait mais d'un autre côté, il n'est plus possible de prédire quel chemin la particule va emprunter au point que la notion même de chemin s'évanouit. Ainsi à l'échelle quantique la particule semble chercher son chemin parmi tous ceux qui sont possibles mais à l'échelle macroscopique, ce tâtonnement quantique semble avoir permis à la particule de trouver le "bon chemin".

Le formalisme de l'intégrale de chemin constitue une façon très originale d'aborder et d'interpréter la physique quantique qui s'est ajoutée à celles qui avaient été développées par Schrödinger.

4.6. THÉORÈME D'EHRENFEST

Ce théorème permet de connecter la mécanique classique de Newton à la physique quantique en établissant des relations similaires en ce qui concerne la quantité de mouvement et la force.

Pour cela, nous partons l'exemple particulier d'une particule massive se déplaçant à une vitesse non relativiste dans un potentiel. Nous avons alors l'équation de Schrödinger d'évolution à une dimension:

$$i\hbar \frac{\partial \Psi}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \Psi}{\partial x^2} + E_{pot}(x)\Psi \quad (42.565)$$

d'où nous tirons (utile pour plus loin):

$$\frac{\partial \Psi}{\partial t} = -\frac{i}{\hbar} \left(-\frac{\hbar^2}{2m} \frac{\partial^2 \Psi}{\partial x^2} + E_{pot}(x)\Psi \right) \quad (42.566)$$

Si nous prenons en toute généralité le conjugué complexe des deux côtés de l'égalité et en multipliant les 2 membres par $\hbar$:

$$-i\hbar \frac{\partial \bar{\Psi}}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \bar{\Psi}}{\partial x^2} + E_{pot}(x) \bar{\Psi} \quad (42.567)$$

d'où nous tirons (utile aussi pour plus loin):

$$\frac{\partial \bar{\Psi}}{\partial t} = \frac{i}{\hbar} \left(-\frac{\hbar^2}{2m} \frac{\partial^2 \bar{\Psi}}{\partial x^2} + E_{pot}(x) \bar{\Psi} \right) \quad (42.568)$$

Prenons la variation temporelle de la position moyenne de la particule (5ème postulat):

$$\begin{aligned} \frac{d}{dt} \langle x \rangle &= \frac{d}{dt} \int_{-\infty}^{+\infty} \bar{\Psi} x \Psi dx = \int_{-\infty}^{+\infty} \frac{d}{dt} (\bar{\Psi} x \Psi) dx = \int_{-\infty}^{+\infty} \left(\frac{\partial \bar{\Psi}}{\partial t} x \Psi + \bar{\Psi} x \frac{\partial \Psi}{\partial t} \right) dx \\ &= \int_{-\infty}^{+\infty} \frac{\partial \bar{\Psi}}{\partial t} x \Psi dx + \int_{-\infty}^{+\infty} \bar{\Psi} x \frac{\partial \Psi}{\partial t} dx = \int_{-\infty}^{+\infty} \left(\frac{\partial \bar{\Psi}}{\partial t} \right) x \Psi dx + \int_{-\infty}^{+\infty} \bar{\Psi} x \left(\frac{\partial \Psi}{\partial t} \right) dx \\ &= \frac{i}{\hbar} \left[\int_{-\infty}^{+\infty} \left(-\frac{\hbar^2}{2m} \frac{\partial^2 \bar{\Psi}}{\partial x^2} + E_{pot}(x) \bar{\Psi} \right) x \Psi dx - \int_{-\infty}^{+\infty} \bar{\Psi} x \left(-\frac{\hbar^2}{2m} \frac{\partial^2 \Psi}{\partial x^2} + E_{pot}(x) \Psi \right) dx \right] \\ &= \frac{i\hbar}{2m} \left[\int_{-\infty}^{+\infty} \left(-\frac{\partial^2 \bar{\Psi}}{\partial x^2} \right) x \Psi dx - \int_{-\infty}^{+\infty} \bar{\Psi} x \left(-\frac{\partial^2 \Psi}{\partial x^2} \right) dx \right] = \frac{i\hbar}{2m} \int_{-\infty}^{+\infty} \left(\bar{\Psi} x \frac{\partial^2 \Psi}{\partial x^2} - \frac{\partial^2 \bar{\Psi}}{\partial x^2} x \Psi \right) dx \end{aligned} \quad (42.569)$$

Nous avons:

$$\frac{\partial^2 (x\Psi)}{\partial x^2} = \frac{\partial}{\partial x} \left(\frac{\partial x}{\partial x} \Psi + x \frac{\partial \Psi}{\partial x} \right) = \left(\frac{\partial^2 x}{\partial x^2} \Psi + \frac{\partial x}{\partial x} \frac{\partial \Psi}{\partial x} + \frac{\partial x}{\partial x} \frac{\partial \Psi}{\partial x} + x \frac{\partial^2 \Psi}{\partial x^2} \right) \quad (42.570)$$

d'où:

$$\frac{\partial^2 (x\Psi)}{\partial x^2} - 2 \frac{\partial \Psi}{\partial x} = x \frac{\partial^2 \Psi}{\partial x^2} \quad (42.571)$$

Utilisons cette dernière relation:

$$\begin{aligned} \frac{d}{dt} \langle x \rangle &= \frac{i\hbar}{2m} \int_{-\infty}^{+\infty} \left(\bar{\Psi} x \frac{\partial^2 \Psi}{\partial x^2} - \frac{\partial^2 \bar{\Psi}}{\partial x^2} x \Psi \right) dx = \frac{i\hbar}{2m} \int_{-\infty}^{+\infty} \left(\bar{\Psi} \left(\frac{\partial^2 (x\Psi)}{\partial x^2} - 2 \frac{\partial \Psi}{\partial x} \right) - \frac{\partial^2 \bar{\Psi}}{\partial x^2} x \Psi \right) dx \\ &= \frac{i\hbar}{2m} \int_{-\infty}^{+\infty} \left(\bar{\Psi} \frac{\partial^2 (x\Psi)}{\partial x^2} - \frac{\partial^2 \bar{\Psi}}{\partial x^2} (x\Psi) \right) dx - \frac{i\hbar}{m} \int_{-\infty}^{+\infty} \bar{\Psi} \frac{\partial \Psi}{\partial x} dx \end{aligned} \quad (42.572)$$

Utilisons maintenant la relation:

$$\bar{\Psi} \frac{\partial^2 (x\Psi)}{\partial x^2} - \frac{\partial^2 \bar{\Psi}}{\partial x^2} (x\Psi) = \frac{\partial}{\partial x} \left(\bar{\Psi} \frac{\partial (x\Psi)}{\partial x} - \frac{\partial \bar{\Psi}}{\partial x} (x\Psi) \right) \quad (42.573)$$

et injectons-la dans la relation antérieure:

$$\frac{d}{dt}\langle x \rangle = \frac{i\hbar}{2m} \int_{-\infty}^{+\infty} \frac{\partial}{\partial x} \left(\bar{\Psi} \frac{\partial(x\Psi)}{\partial x} - \frac{\partial \bar{\Psi}}{\partial x} (x\Psi) \right) dx + \frac{1}{m} \int_{-\infty}^{+\infty} \bar{\Psi} \left(-i\hbar \frac{\partial \Psi}{\partial x} \right) dx \quad (42.574)$$

Le premier terme à droite de l'égalité est facile à intégrer... (puisque'il n'y a pas besoin de l'intégrer):

$$\int_{-\infty}^{+\infty} \frac{\partial}{\partial x} \left(\bar{\Psi} \frac{\partial(x\Psi)}{\partial x} - \frac{\partial \bar{\Psi}}{\partial x} (x\Psi) \right) dx = \left[\bar{\Psi} \frac{\partial(x\Psi)}{\partial x} - \frac{\partial \bar{\Psi}}{\partial x} (x\Psi) \right] \Bigg|_{-\infty}^{+\infty} \quad (42.575)$$

et comme la fonction d'onde doit valoir 0 à $x = \pm\infty$ (sinon l'énergie est infinie) alors cette dernière relation est nulle. Il nous reste alors:

$$\frac{d}{dt}\langle x \rangle = \frac{1}{m} \int_{-\infty}^{+\infty} \bar{\Psi} \left(-i\hbar \frac{\partial \Psi}{\partial x} \right) dx \quad (42.576)$$

Soit:

$$m \frac{d\langle x \rangle}{dt} = \int_{-\infty}^{+\infty} \bar{\Psi} \left(-i\hbar \frac{\partial}{\partial x} \right) \Psi dx \quad (42.577)$$

et finalement:

$$\boxed{m \frac{d\langle x \rangle}{dt} = \int_{-\infty}^{+\infty} \bar{\Psi} p_x \Psi dx = \langle p_x \rangle} \quad (42.578)$$

ce qui est l'équivalent en mécanique classique de:

$$m \frac{dx}{dt} = mv = p \quad (42.579)$$

et qui reconfirme l'existence de l'être mathématique:

$$p_x = -i\hbar \frac{\partial}{\partial x} = \frac{\hbar}{i} \frac{\partial}{\partial x} \quad (42.580)$$

comme étant l'opérateur de quantité de mouvement et que nous avons déterminé plus haut en retrouvant la deuxième loi de Newton.

Mais nous pouvons faire un peu mieux au niveau de l'analogie classique/quantique en dérivant:

$$\langle p_x \rangle = \int_{-\infty}^{+\infty} \bar{\Psi} p_x \Psi dx \quad (42.581)$$

Ce qui donne:

$$\frac{d\langle p_x \rangle}{dt} = \frac{d}{dt} \int_{-\infty}^{+\infty} \bar{\Psi} p_x \Psi dx = -i\hbar \int_{-\infty}^{+\infty} \frac{d}{dt} \left(\bar{\Psi} \frac{\partial \Psi}{\partial x} \right) dx = -i\hbar \int_{-\infty}^{+\infty} \left(\frac{\partial \bar{\Psi}}{\partial t} \frac{\partial \Psi}{\partial x} + \bar{\Psi} \frac{\partial}{\partial t} \frac{\partial \Psi}{\partial x} \right) dx \quad (42.582)$$

d'où:

$$-\frac{1}{i\hbar} \frac{d\langle p_x \rangle}{dt} = \int_{-\infty}^{+\infty} \frac{\partial \bar{\Psi}}{\partial t} \frac{\partial \Psi}{\partial x} dx + \int_{-\infty}^{+\infty} \bar{\Psi} \frac{\partial}{\partial t} \frac{\partial \Psi}{\partial x} dx = \int_{-\infty}^{+\infty} \frac{\partial \bar{\Psi}}{\partial t} \frac{\partial \Psi}{\partial x} dx + \int_{-\infty}^{+\infty} \bar{\Psi} \frac{\partial}{\partial x} \frac{\partial \Psi}{\partial t} dx \quad (42.583)$$

En utilisant:

$$i\hbar \frac{\partial \Psi}{\partial t} = H\Psi \Leftrightarrow i\hbar \frac{\partial \Psi}{\partial t} = \left(-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + E_{pot}(x) \right) \Psi \quad (42.584)$$

Il vient:

$$\begin{aligned} -\frac{1}{i\hbar} \frac{d\langle p_x \rangle}{dt} &= -\frac{1}{i\hbar} \int_{-\infty}^{+\infty} H \bar{\Psi} \frac{\partial \Psi}{\partial x} dx + \frac{1}{i\hbar} \int_{-\infty}^{+\infty} \bar{\Psi} \frac{\partial (H\Psi)}{\partial x} dx \\ &= -\frac{1}{i\hbar} \int_{-\infty}^{+\infty} \left(-\frac{\hbar^2}{2m} \frac{\partial^2 \bar{\Psi}}{\partial x^2} + E_{pot}(x) \bar{\Psi} \right) \frac{\partial \Psi}{\partial x} dx + \frac{1}{i\hbar} \int_{-\infty}^{+\infty} \bar{\Psi} \frac{\partial \left(-\frac{\hbar^2}{2m} \frac{\partial^2 \Psi}{\partial x^2} + E_{pot}(x) \Psi \right)}{\partial x} dx \\ &= \frac{1}{i\hbar} \left[-\int_{-\infty}^{+\infty} \left(-\frac{\hbar^2}{2m} \frac{\partial^2 \bar{\Psi}}{\partial x^2} + E_{pot}(x) \bar{\Psi} \right) \frac{\partial \Psi}{\partial x} dx + \int_{-\infty}^{+\infty} \bar{\Psi} \frac{\partial \left(-\frac{\hbar^2}{2m} \frac{\partial^2 \Psi}{\partial x^2} + E_{pot}(x) \Psi \right)}{\partial x} dx \right] \quad (42.585) \\ &= \frac{1}{i\hbar} \left[\frac{\hbar^2}{2m} \int_{-\infty}^{+\infty} \left(\frac{\partial^2 \bar{\Psi}}{\partial x^2} \frac{\partial \Psi}{\partial x} - \bar{\Psi} \frac{\partial}{\partial x} \frac{\partial^2 \Psi}{\partial x^2} \right) dx + \int_{-\infty}^{+\infty} \left(\bar{\Psi} \frac{\partial (E_{pot}(x) \Psi)}{\partial x} - E_{pot}(x) \bar{\Psi} \frac{\partial \Psi}{\partial x} \right) dx \right] \end{aligned}$$

Concentrons-nous sur:

$$\int_{-\infty}^{+\infty} \left(\frac{\partial^2 \bar{\Psi}}{\partial x^2} \frac{\partial \Psi}{\partial x} - \bar{\Psi} \frac{\partial}{\partial x} \frac{\partial^2 \Psi}{\partial x^2} \right) dx \quad (42.586)$$

Intégrons par parties le premier terme deux fois selon la relation démontrée dans le chapitre de Calcul Différentiel Et Intégral:

$$\int_a^b f(t) \cdot g'(t) dt = f(t) \cdot g(t) \Big|_a^b - \int_a^b f'(t) \cdot g(t) dt \quad (42.587)$$

Nous avons alors (toujours en considérant Ψ comme étant une fonction décroissante vers l'infini):

$$\int_{-\infty}^{+\infty} \frac{\partial^2 \bar{\Psi}}{\partial x^2} \frac{\partial \Psi}{\partial x} dx = \int_{-\infty}^{+\infty} \frac{\partial \Psi}{\partial x} \frac{\partial^2 \bar{\Psi}}{\partial x^2} dx = \frac{\partial \Psi}{\partial x} \frac{\partial \bar{\Psi}}{\partial x} \Big|_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} \frac{\partial}{\partial x} \frac{\partial \Psi}{\partial x} \frac{\partial \bar{\Psi}}{\partial x} dx = - \int_{-\infty}^{+\infty} \frac{\partial^2 \Psi}{\partial x^2} \frac{\partial \bar{\Psi}}{\partial x} dx \quad (42.588)$$

et encore une fois:

$$-\int_{-\infty}^{+\infty} \frac{\partial^2 \Psi}{\partial x^2} \frac{\partial \bar{\Psi}}{\partial x} dx = -\frac{\partial^2 \Psi}{\partial x^2} \bar{\Psi} \Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} \frac{\partial}{\partial x} \frac{\partial^2 \Psi}{\partial x^2} \bar{\Psi} dx = \int_{-\infty}^{+\infty} \frac{\partial}{\partial x} \frac{\partial^2 \Psi}{\partial x^2} \bar{\Psi} dx \quad (42.589)$$

Donc finalement:

$$\int_{-\infty}^{+\infty} \left(\frac{\partial^2 \bar{\Psi}}{\partial x^2} \frac{\partial \Psi}{\partial x} - \bar{\Psi} \frac{\partial}{\partial x} \frac{\partial^2 \Psi}{\partial x^2} \right) dx = \int_{-\infty}^{+\infty} \left(\frac{\partial}{\partial x} \frac{\partial^2 \Psi}{\partial x^2} \bar{\Psi} - \bar{\Psi} \frac{\partial}{\partial x} \frac{\partial^2 \Psi}{\partial x^2} \right) dx = 0 \quad (42.590)$$

Il nous reste alors:

$$\begin{aligned} \frac{d \langle p_x \rangle}{dt} &= - \left[\int_{-\infty}^{+\infty} \left(\bar{\Psi} \frac{\partial (E_{pot}(x) \Psi)}{\partial x} - E_{pot}(x) \bar{\Psi} \frac{\partial \Psi}{\partial x} \right) dx \right] \\ &= - \left[\int_{-\infty}^{+\infty} \left(\bar{\Psi} \frac{\partial (E_{pot}(x))}{\partial x} \Psi + \cancel{\bar{\Psi} E_{pot}(x) \frac{\partial \Psi}{\partial x}} - \cancel{E_{pot}(x) \bar{\Psi} \frac{\partial \Psi}{\partial x}} \right) dx \right] = - \int_{-\infty}^{+\infty} \bar{\Psi} \frac{\partial E_{pot}(x)}{\partial x} \Psi \quad (42.591) \\ &= \int_{-\infty}^{+\infty} \bar{\Psi} \left(- \frac{dE_{pot}}{dx} \right) \Psi \end{aligned}$$

Or, nous avons démontré dans le chapitre de Mécanique Classique que:

$$\vec{F} = -\vec{\nabla} E_p \quad (42.592)$$

Il vient donc que:

$$\boxed{\frac{d \langle p_x \rangle}{dt} = \int_{-\infty}^{+\infty} \bar{\Psi} F_x \Psi = \langle F_x \rangle} \quad (42.593)$$

Ce résultat extraordinairement simple constitue le "**théorème d'Ehrenfest**". Nous retrouvons donc la loi fondamentale de la dynamique classique au sens des valeurs moyennes de position et de la force, calculées à l'aide de la probabilité de présence!

5. MOMENT CINÉTIQUE ET SPIN

Tout comme l'oscillateur harmonique, la notion de moment cinétique (ou moment angulaire) est d'une importance capitale en théorie quantique et possède de nombreuses applications dans tous les domaines de la physique: physique atomique et moléculaire, physique nucléaire et subnucléaire, physique de l'état condensé, etc. Ainsi, il joue un rôle essentiel dans l'étude du mouvement d'une particule dans un potentiel à symétrie sphérique, comme nous le verrons en chimie quantique (qui en est un excellent exemple pratique). Le moment cinétique est également à la base du groupe des rotations qui satisfait à l'algèbre des opérateurs de moment cinétique (cf. chapitre d'Algèbre Ensembliste). De ce fait, il permet non seulement de construire la fonction d'onde d'un système quantique de symétrie donnée, mais aussi de prédire si une transition optique est permise et d'en déterminer son intensité (par exemple, lors de l'étude des transitions optiques entre états d'impureté (en état solide), états moléculaires (chimie quantique), en physique nucléaire, etc.).

Enfin, nous verrons que la méthode algébrique appliquée à l'étude du moment cinétique nous permettra d'introduire tout naturellement la notion de moment cinétique intrinsèque d'une particule, le "spin", qui n'a pas d'équivalent classique.

Les développements qui vont suivre peuvent paraître assez déconcertants dans le sens qu'il ne faut plus du tout se fier à l'intuition mais uniquement aux propriétés et résultats des mathématiques. Comme d'habitude, si vous avez besoin de compléments d'informations, n'hésitez pas à nous contacter.

Ainsi, rappelons que le moment cinétique d'une particule par rapport à l'origine est donné par (cf. chapitre de Mécanique Classique):

$$\vec{b} = m\vec{r} \times \vec{v} = \vec{r} \times \vec{p} \quad (42.594)$$

La quantité de mouvement étant quantifiée (c'est une valeur propre rattachée à l'énergie d'une façon ou d'une autre), le moment cinétique l'est nécessairement aussi (le moment cinétique est donc aussi une valeur propre) et l'expérience a appuyé ce résultat (Stern-Gerlach).

Soit la composante en z du produit vectoriel résultant:

$$\boxed{b_z = x p_y - y p_x} \quad (42.595)$$

(cycl.)

Cette relation étant cyclique, nous pouvons changer les indices pour obtenir les autres coordonnées.

Comme x et y commutent (dans le sens que leur commutateur est nul) et que nous avons démontré:

$$p_x = -i\hbar \frac{\partial}{\partial x}, p_y = -i\hbar \frac{\partial}{\partial y} \quad (42.596)$$

nous avons alors:

$$b_z = \frac{\hbar}{i} x \frac{\partial}{\partial y} - \frac{\hbar}{i} y \frac{\partial}{\partial x} = \frac{\hbar}{i} \left(x \frac{\partial}{\partial y} - y \frac{\partial}{\partial x} \right) \quad (42.597)$$

Ce qui donne:

$$b_z = \frac{\hbar}{i} \left(x \frac{\partial}{\partial y} - y \frac{\partial}{\partial x} \right) \quad (42.598)$$

(cycl.)

En utilisant le gradient (nous retrouverons cette relation dans le chapitre de Physique Quantique Relativiste lors de notre étude de l'équation de Pauli!!):

$$\boxed{\vec{b} = \frac{\hbar}{i} \vec{r} \times \vec{\nabla} = -i\hbar \vec{r} \times \vec{\nabla}} \quad (42.599)$$

et en posant pour "l'opérateur du moment cinétique orbital":

$$\frac{1}{i} \vec{r} \times \vec{\nabla} = \frac{1}{\hbar} \vec{r} \times \vec{p} = \vec{l} \quad (42.600)$$

Ce qui nous amène à écrire:

$$\vec{b} = \hbar \vec{l} \quad (42.601)$$

Avec:

$$\begin{pmatrix} l_x \\ l_y \\ l_z \end{pmatrix} = \frac{1}{\hbar} \begin{pmatrix} y p_z - z p_y \\ z p_x - x p_z \\ x p_y - y p_x \end{pmatrix} \quad (42.602)$$

Remarque: Le plus souvent dans la littérature le moment cinétique orbital est noté $\vec{L}$ (nous avons déjà fait cette remarque dans le chapitre de Mécanique Classique) mais nous avons évité cette notation ici afin de différencier le moment cinétique orbital et le moment cinétique orbital total.

Nous allons établir certaines relations de commutation concernant $\vec{l}$ qui joueront un rôle essentiel dans l'étude du spin. En faisant usage des relations de commutation suivantes (démontrées lors de notre étude des principes d'incertitudes):

$$[x, p_x] = i\hbar \text{ (cycl.)} \quad (42.603)$$

et:

$$[y, p_x] = 0 \text{ (cycl.)} \quad (42.604)$$

Nous avons la relation (il est de tradition de faire l'analyse sur la composante de la projection de $\vec{l}$ en z):

$$[\hbar l_z, x] = [x p_y, x] - [y p_x, x] = -y [p_x, x] = i\hbar y \quad (42.605)$$

Donc:

$$[l_z, x] = iy \text{ (cycl.)} \quad (42.606)$$

et en procédant de la même manière:

$$[l_z, y] = -ix \text{ (cycl.) et } [l_z, z] = 0 \text{ (cycl.)} \quad (42.607)$$

Remarque: Nous trouvons des relations analogues avec la quantité de mouvement:

$$\begin{aligned} [l_x, p_x] &= i p_y \\ [l_x, p_y] &= -i p_x \text{ (cycl.)} \\ [l_x, p_z] &= 0 \end{aligned} \quad (42.608)$$

Évaluons maintenant la quantité (suite à la demande d'un lecteur, nous avons mis tous les détails):

$$\begin{aligned}
[l_x, \hbar l_y] &= [l_x, z p_x - x p_z] = [l_x, z p_x] - [l_x, x p_z] = \left[\frac{1}{\hbar} (y p_z - z p_y), z p_x \right] - \left[\frac{1}{\hbar} (y p_z - z p_y), x p_z \right] \\
&= \left[\frac{1}{\hbar} y p_z, z p_x \right] - \left[\frac{1}{\hbar} z p_y, z p_x \right] - \left[\frac{1}{\hbar} y p_z, x p_z \right] + \left[\frac{1}{\hbar} z p_y, x p_z \right] \\
&= \left(\frac{1}{\hbar} y p_z z p_x - \frac{1}{\hbar} z \underbrace{p_x y}_{=-i\hbar \frac{\partial}{\partial x} y=0} p_z \right) - \left(\frac{1}{\hbar} z \underbrace{p_y z}_{=-i\hbar \frac{\partial}{\partial y} z=0} p_x - \frac{1}{\hbar} z \underbrace{p_x z}_{=-i\hbar \frac{\partial}{\partial x} z=0} p_y \right) \\
&\quad - \left(\frac{1}{\hbar} y \underbrace{p_z x}_{=-i\hbar \frac{\partial}{\partial z} x=0} p_z - \frac{1}{\hbar} x \underbrace{p_z y}_{=-i\hbar \frac{\partial}{\partial z} y=0} p_z \right) + \left(\frac{1}{\hbar} z \underbrace{p_y x}_{=-i\hbar \frac{\partial}{\partial y} x=0} p_z - \frac{1}{\hbar} x p_z z p_y \right) \\
&= \frac{1}{\hbar} y p_z z p_x - \frac{1}{\hbar} x p_z z p_y = \frac{1}{\hbar} y \left(-i\hbar \frac{\partial}{\partial z} z \right) p_x - \frac{1}{\hbar} x \left(-i\hbar \frac{\partial}{\partial z} z \right) p_y = \frac{1}{\hbar} y (-i\hbar) p_x - \frac{1}{\hbar} x (-i\hbar) p_y \\
&= -i y p_x + i x p_y = i (x p_y - y p_x) = i \hbar l_z
\end{aligned} \tag{42.609}$$

Soit après simplification (c'est assez embêtant pour l'expérience que cela ne commute pas):

$$\boxed{[l_x, l_y] = i \hbar l_z} \tag{42.610}$$

(cycl.)

par ailleurs, à ce stade, si le lecteur a déjà parcouru au préalable le chapitre de Calcul Spinoriel, il remarquera que les matrices de Pauli satisfont aux relations précédentes si nous nous mettons en unités naturelles (la constante de Planck réduite valant alors 1):

$$\begin{aligned}
\sigma_x \sigma_y - \sigma_y \sigma_x &= 2i \sigma_z \\
\sigma_y \sigma_z - \sigma_z \sigma_y &= 2i \sigma_x \\
\sigma_z \sigma_x - \sigma_x \sigma_z &= 2i \sigma_y
\end{aligned}$$

Ce constat sera utile pour notre étude de la physique quantique relativiste (voir chapitre du même nom). Effectivement, nous savons de par notre étude du calcul spinoriel (cf. chapitre de Calcul Spinoriel) que les matrices 2 par 2 complexes unitaires de déterminant 1 forment le groupe des rotations dans l'espace $SU(2)$, dont les matrices Pauli sont les génératrices. Fondamentalement, l'origine du spin vient du lien qui existe entre $SU(2)$ et le groupe des rotations de notre espace ordinaire, $SO(3)$ (cf. chapitre d'Algèbre Ensembliste).

Maintenant, considérons la norme:

$$\|\vec{l}\|^2 = l_x^2 + l_y^2 + l_z^2 \tag{42.611}$$

où il faut considérer le carré d'un de ces opérateurs sous la forme suivante:

$$l_z^2 = \left(\frac{1}{i} (x \partial_y - y \partial_x) \right)^2 = - (x \partial_y - y \partial_x)^2 \quad (42.612)$$

Étudions son commutateur avec une composante (sans avoir à expliciter la chose!):

$$\begin{aligned} [l_z, \|\vec{l}\|^2] &= [l_z, l_x^2] + [l_z, l_y^2] + \underbrace{[l_z, l_z^2]}_{=0} = l_x [l_z, l_x] + [l_z, l_x] l_x + l_y [l_z, l_y] + [l_z, l_y] l_y \\ &= l_x [l_z, l_x] + [l_z, l_x] l_x - l_y [l_y, l_z] - [l_y, l_z] l_y \end{aligned} \quad (42.613)$$

en utilisant la relation cyclique $[l_x, l_y] = i l_z$ il vient:

$$[l_z, \|\vec{l}\|^2] = l_x [l_z, l_x] + [l_z, l_x] l_x - l_y [l_y, l_z] - [l_y, l_z] l_y = i(l_x l_y + l_y l_x) - i(l_x l_y + l_y l_x) = 0 \quad (42.614)$$

Donc la norme du moment cinétique orbital commute avec ses composantes:

$$\left[l_z, \|\vec{l}\|^2 \right] = 0 \quad (42.615)$$

(cycl.)

Conclusions des résultats obtenus jusqu'à maintenant: Comme le commutateur est nul (les quantités commutent) il est donc possible de mesurer simultanément avec précision une composante ainsi que le carré du moment cinétique (sa norme au carré), mais il est impossible de faire la même chose pour deux composantes!

Notons enfin que la relation que $[l_x, l_y] = i l_z$ peut s'écrire:

$$[l_x, l_y] = l_x l_y - l_y l_x = i l_z \quad (42.616)$$

et donc d'une façon un peu curieuse:

$$\boxed{\vec{l} \times \vec{l} = i \vec{l}} \quad (42.617)$$

Si nous avons un système de particules numérotées par l'indice k , chacune a un moment cinétique individuel $\hbar \vec{l}^{(k)}$ et le moment cinétique orbital total du système $\hbar \vec{L}$ (ne pas confondre la notation avec le Lagrangien!!!), est défini par (en unités naturelles $\hbar$):

$$\vec{L} = \sum_k \vec{l}^{(k)} \quad (42.618)$$

Mais $\vec{L}$ n'est pas encore vraiment le moment cinétique total du système! Effectivement, une particule peut posséder un moment cinétique intrinsèque, ou "spin". Nous pouvons donner une image simple du spin en disant qu'il traduit une rotation infinitésimale de la particule sur elle-même (attention !!! ce n'est qu'une image car au fait la particule ne tourne pas sur elle-même !). Comme nous l'avons vu dans le chapitre de Calcul Spinoriel, cela correspond mathématiquement au développement limité de la matrice de rotations au voisinage de la matrice identité.

Nous noterons $\vec{s}^{(k)}$ le moment cinétique de spin de la k -ème particule (en unités naturelles $\hbar$) et la relation:

$$\vec{S} = \sum_k \vec{s}^{(k)} \quad (42.619)$$

sera le spin total et enfin:

$$\boxed{\vec{J} = \vec{L} + \vec{S}} \quad (42.620)$$

sera le "**moment cinétique total**" du système (ne pas confondre la notation J avec le moment cinétique orbital ou la densité de courant!!!) et nous démontrerons lors de notre étude du couplage spin-orbite que ce moment cinétique est une constante du mouvement en présence de ce couplage.

Nous allons supposer (mais c'est relativement facile à démontrer une fois, entre autres, les spineurs connus) que chaque $\vec{s}^{(k)}$ et $\vec{J}^{(k)}$ obéit aussi aux lois de commutation vues précédemment:

$$[s_x^{(k)}, s_y^{(k)}] = i s_z^{(k)} \text{ (cycl.) et } [J_x^{(k)}, J_y^{(k)}] = i J_z^{(k)} \text{ (cycl.)} \quad (42.621)$$

Ce qui s'écrit sous forme tensorielle en utilisant le symbole de Levi-Civita (cf. chapitre de Calcul Tensoriel):

$$\boxed{[J_i, J_j] = i \varepsilon_{ijk} J_k} \text{ et } \boxed{[S_i, S_j] = i \varepsilon_{ijk} S_k} \quad (42.622)$$

Nous remarquons que dans la représentation matricielle de Heisenberg il y a aussi des composantes de matrices qui satisfont ces deux relations. Par exemple les composantes des matrices hermitiques (dont la transposée conjuguée est égale à elle-même pour rappel....) et de trace nulle suivantes:

$$\underbrace{J_z = \begin{bmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, J_x = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{bmatrix}, J_y = \begin{bmatrix} 0 & 0 & i \\ 0 & 0 & 0 \\ -i & 0 & 0 \end{bmatrix}}_{(J^i)_{mn} = -i \varepsilon_{imn}} \quad (42.623)$$

Les deux relation antérieures entraînent (aussi) au même titre que pour le moment cinétique orbital:

$$\boxed{[J_z, \|\vec{J}\|^2] = 0} \quad (42.624)$$

(cycl.)

avec bien évidemment la relation:

$$\|\vec{J}\|^2 = J_x^2 + J_y^2 + J_z^2$$

appelée par les mathématiciens "**élément de Casimir**" ou encore "**opérateur de Casimir**" (un simple développement parfaitement similaire à celui obtenu plus haut suffit à la démontrer).

Définissons maintenant de façon purement formelle les deux opérateurs non hermitiques dits "opérateurs d'échelle" (les matrices de Pauli satisfont toujours à ces relations!):

$$\begin{aligned} J_{\pm} &= J_x \pm iJ_y \quad (42.625) \\ (J_+)^{\dagger} &= J_- \end{aligned}$$

où respectivement J_+ est appelé "opérateur élévateur" et J_- "opérateur abaisseur".

Les $J_{\pm}$ commutent avec $\|\vec{J}\|^2$, puisque celui-ci commute avec J_x et J_y . Ce qui nous permet d'écrire le produit:

$$J_+ \cdot \|\vec{J}\|^2 = \|\vec{J}\|^2 \cdot J_- \quad (42.626)$$

Par ailleurs:

$$[J_z, J_+] = [J_z, J_x] + i[J_z, J_y] = iJ_y + J_x \quad (42.627)$$

Donc:

$$J_z J_+ - J_+ J_z = J_+ \quad (42.628)$$

De même:

$$J_z J_- - J_- J_z = -J_- \quad (42.629)$$

Enfin, évaluons les produits $J_+ J_-$ et $J_- J_+$:

$$\begin{aligned} J_- J_+ &= (J_x - iJ_y)(J_x + iJ_y) = J_x^2 + J_y^2 + i(J_x J_y - J_y J_x) \\ &= \|\vec{J}\|^2 - J_z^2 - J_z \end{aligned} \quad (42.630)$$

De même:

$$J_+ J_- = \|\vec{J}\|^2 - J_z^2 + J_z \quad (42.631)$$

Puisque les deux opérateurs hermitiques $\|\vec{J}\|^2$ et J_z commutent ils ont donc des états et valeurs propres communes et, plus précisément, ils ont une base propre complète commune. Lorsque des observables commutent et ont une base propre commune, rappelons que nous avons pour habitude de parler d'un "ECOC" (Ensemble Complet d'Opérateurs qui Commutent).

Pour étudier leurs valeurs propres posons:

$$\begin{aligned} \|\vec{J}\|^2 |\Psi\rangle &= K |\Psi\rangle \\ J_z |\Psi\rangle &= m |\Psi\rangle \end{aligned} \quad (42.632)$$

Système qui est parfois noté sous la forme suivante dans la littérature spécialisée:

$$\begin{aligned}\|\vec{J}\|^2 |K, m\rangle &= K |K, m\rangle \\ J_z |K, m\rangle &= m |K, m\rangle\end{aligned}\quad (42.633)$$

Car elle met en évidence que les états propres associés seront définis au moins en partie par les paramètres K, m .

Pour commencer, nous savons que les valeurs propres K et m ne sont pas indépendantes puisque nous avons:

$$\|\vec{J}\|^2 - J_z^2 = J_x^2 + J_y^2 \quad (42.634)$$

La moyenne étant notée par les crochets $\langle \rangle$, nous avons par linéarité de l'espérance (cf. chapitre de Statistiques):

$$\langle \|\vec{J}\|^2 \rangle - \langle J_z^2 \rangle = \langle J_x^2 \rangle + \langle J_y^2 \rangle \quad (42.635)$$

Ce qui peut s'écrire:

$$\langle \|\vec{J}\|^2 - J_z^2 \rangle = \langle J_x^2 \rangle + \langle J_y^2 \rangle \quad (42.636)$$

Nous voyons que le membre de gauche de la relation ci-dessus est donc égal par définition à:

$$K - m^2 \quad (42.637)$$

Comme l'opérateur du moment cinétique orbital total au carré est de toute façon hermitique (il n'a pas de composante complexe dans $\mathbb{C}$), nous avons alors par construction des postulats de la physique quantique:

$$\langle J_x^2 \rangle = \langle \Psi | J_x^2 | \Psi \rangle \geq 0 \quad (42.638)$$

Il vient alors que:

$$\left. \begin{aligned} \langle J_x^2 \rangle &\geq 0 \\ \langle J_y^2 \rangle &\geq 0 \end{aligned} \right\} \langle \|\vec{J}\|^2 - J_z^2 \rangle \geq 0 \quad (42.639)$$

Cette dernière relation implique donc que:

$$K \geq m^2 \quad (42.640)$$

Ce qui nous apporte jusqu'ici les informations suivantes:

$$\boxed{K \geq 0, |m| \leq \sqrt{K}} \quad (42.641)$$

À partir de $|\Psi\rangle$, nous bâtissons l'état $J_+|\Psi\rangle$, nous allons montrer que si cet état n'est pas identiquement nul, il est état propre de $\|\vec{J}\|^2$ et de J_z . De la relation:

$$J_+ \cdot \|\vec{J}\|^2 = \|\vec{J}\|^2 \cdot J_- \quad (42.642)$$

déjà démontrée précédemment, nous posons:

$$\|\vec{J}\|^2 J_+|\Psi\rangle = J_- \|\vec{J}\|^2 |\Psi\rangle = K J_+|\Psi\rangle \quad (42.643)$$

$J_{\pm}$ commutent avec $\|\vec{J}\|^2$, puisque celui-ci commute avec J_x et J_y . Ce qui nous donne que la relation précédente est nulle telle que:

$$K J_+|\Psi\rangle = 0 \Leftrightarrow J_+|\Psi\rangle = 0 \quad (42.644)$$

De la relation $J_z J_+ - J_+ J_z = J_+$ nous posons de façon identique:

$$J_z J_+|\Psi\rangle = J_+ J_z|\Psi\rangle + J_+|\Psi\rangle = (m+1) J_+|\Psi\rangle \quad (42.645)$$

Toujours avec:

$$J_+|\Psi\rangle = 0 \quad (42.646)$$

Nous avons finalement le paquet de relations:

$$\boxed{J_{\pm}|\Psi\rangle = 0, \quad \|\vec{J}\|^2 J_{\pm}|\Psi\rangle = K J_{\pm}|\Psi\rangle, \quad J_z J_{\pm}|\Psi\rangle = (m \pm 1) J_{\pm}|\Psi\rangle} \quad (42.647)$$

Donc $J_+|\Psi\rangle$ et $J_-|\Psi\rangle$ sont identiquement nuls et $J_+|\Psi\rangle$ et $J_-|\Psi\rangle$ sont des états propres de l'opérateur $\|\vec{J}\|^2$ pour la valeur propre K , et de J_z pour la valeur propre $m \pm 1$.

Puisque le moment cinétique est quantifié, ses valeurs propres doivent donc avoir un minimum et un maximum avec pour chacune la fonction propre associée.

Posons pour la suite que m' et $|\Psi'\rangle$ sont la valeur et état propre associé maximal et m'' et $|\Psi''\rangle$ la valeur et état propre minimal.

Étant données les trois relations démontrées jusqu'ici:

$$\begin{aligned} J_+ J_- &= \|\vec{J}\|^2 - J_z^2 + J_z \\ J_- J_+ &= \|\vec{J}\|^2 - J_z^2 - J_z \\ J_z |\Psi\rangle &= 0 \end{aligned} \quad (42.648)$$

Nous écrivons:

$$J_- J_+ |\Psi'\rangle = (K - m'^2 - m'') |\Psi'\rangle = 0$$

$$J_+ J_- |\Psi''\rangle = (K - m'^2 - m'') |\Psi'\rangle = 0$$

(42.649)

Ce qui intuitivement n'est pas évident à poser mais qui mathématiquement est tout à fait justifiable.

À partir des deux dernières relations ci-dessus, nous pouvons écrire en soustrayant la première à la deuxième:

$$m'^2 - m''^2 + m' + m'' = 0 \quad (42.650)$$

soit:

$$(m' + m'')(m' - m'' + 1) = 0 \quad (42.651)$$

m' étant le maximum, m'' le minimum d'un même ensemble, nous avons:

$$m' - m'' \geq 0 \quad (42.652)$$

Ce qui nous donne après simplification de la deuxième parenthèse:

$$(m' + m'') = 0 \Leftrightarrow m'' = -m' \quad (42.653)$$

Notons J la valeur m' (qui correspond donc à la valeur propre maximale de la quantité J_x) puisque $m'' = -m'$ nous avons:

$$m'' = -J \quad (42.654)$$

(où souvent dans la littérature nous retrouvons un j minuscule afin de ne pas avoir de confusion possible avec l'opérateur associé) donc:

$$m' - m'' = J - (-J) = 2J \quad (42.655)$$

Comme la différence à gauche de l'égalité est obligatoirement un nombre entier (en nous inspirant des résultats connus de la physique quantique corpusculaire), cela impose que $2J$ est un nombre entier positif ou nul mais cela implique aussi directement que J ne peut être qu'un nombre entier, demi-entier ou nul tel que:

$$J = 0, \frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, \dots \quad (42.656)$$

Donc, si nous nous fixons un J , puisque par construction $m' - m'' = 2J$, il vient logiquement que:

$$-J \leq m \leq J \quad (42.657)$$

et donc puisque m ne peut être qu'entier, il ne peut prendre que les valeurs:

$$m = -J, -J+1, -J+2, \dots, 0, \dots, J-2, J-1, J \quad (42.658)$$

Enfin, comme:

$$(K - m'^2 - m'^2) = 0 \Leftrightarrow K = m'^2 + m'^2 \text{ et } (K - m''^2 - m''^2) = 0 \Leftrightarrow K = m''^2 + m''^2 \quad (42.659)$$

nous avons alors:

$$K = J^2 + J = J(J+1) \text{ et } K = (-J)^2 - (-J) = J(J+1) \quad (42.660)$$

et finalement cela nous donne la valeur propre:

$$K = J(J+1) \quad (42.661)$$

Et puisque nous avons posé que m' est égal à J et que nous avons la relation:

$$\sqrt{K} \geq |m|$$

Il vient alors:

$$J \leq \sqrt{J(J+1)} \quad (42.662)$$

Sous forme plus explicite et moins confuse (attention à ne pas confondre les valeurs propres avec les opérateurs!):

$$J \leq \sqrt{J_z(J_z + 1)} \quad (42.663)$$

et en définitive, en multipliant à gauche et à droite par $\hbar$ pour revenir en unités du système international (S.I.), nous avons pour la composante verticale du moment cinétique orbital total, la valeur propre:

$$\boxed{b \leq \sqrt{J_z(J_z + 1)}\hbar} \quad (42.664)$$

Il vient sinon au final:

$$\boxed{\|\vec{J}\|^2 |\Psi\rangle = \hbar^2 J(J+1) |\Psi\rangle} \quad (42.665)$$

Comme nous avons vu plus haut que:

$$J = L + S \quad (42.666)$$

(composant par composante de leur vecteur respectif) et si la particule n'a pas de spin ($S = 0$) alors nous avons la valeur propre du moment cinétique orbital total qui se réduit à la valeur propre du moment cinétique:

$$b \leq \sqrt{J(J+1)}\hbar = \sqrt{L(L+1)}\hbar \quad (42.667)$$

où nous n'indiquons plus les indices des composantes (inutile!)

Si nous n'avons qu'une seule particule alors:

$$L = \sum_{i=1}^n l_i = l_1 = l \quad (42.668)$$

Donc le moment cinétique orbitale s'écrit en se rappelant (cf. chapitre de Physique Quantique Corpusculaire) que l est quantifié:

$$b \leq \sqrt{l(l+1)}\hbar \quad (42.669)$$

Si nous avons $l \gg 1$, alors dans ce cas:

$$b \cong l\hbar \quad (42.670)$$

Nous retrouvons donc le résultat obtenu au début de notre étude du moment cinétique.

Grossièrement, si nous posons maintenant $l \cong n$, nous retrouvons à partir du modèle ondulatoire l'hypothèse de quantification du moment cinétique postulée par Bohr vue dans le chapitre de Physique Quantique Corpusculaire. Raison pour laquelle il est d'usage de ne prendre que les valeurs entières de l !

Remarque: Rappelons que réellement $0 \leq l \leq n-1$ et donc qu'à la différence du modèle corpusculaire de Bohr le moment cinétique peut être nul dans le modèle ondulatoire...! Une autre manière d'accepter les valeurs prises par n outre le faire de se reporter au modèle de Bohr dans le chapitre de Physique Quantique Corpusculaire est de regarder les valeurs que peuvent prendre l dans le modèle quantique de l'atome hydrogénoïde du chapitre de Chimie Quantique sinon quoi les polynômes associés de Legendre ne sont plus définis!

Cette constatation justifie maintenant physiquement l'utilisation du nombre quantique l dans l'utilisation du tableau périodique des éléments tel que nous l'avions vu et défini (sans aucune justification réelle) dans le chapitre précédent.

Enfin, indiquons qu'exactement le même raisonnement amène aux valeurs possibles suivantes du moment cinétique de spin:

$$S = 0, \frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, \dots \quad (42.671)$$

où l'expérience nous montre (pour ne citer que les plus connus) que le spin 0 est caractéristique du boson de Higgs ou de certains atomes, le spin 1/2 est une caractéristique de l'électron/positron, le spin 1 est une caractéristique du photon, le spin 2 serait une caractéristique encore théorique du graviton. Au jour où nous écrivons ces lignes, aucune particule de spin 3/2 ou 5/2 n'est connue.

La valeur entière ou demi-entière du spin détermine une propriété cruciale de la particule : si son spin est entier, c'est un boson, si son spin est demi-entier, c'est un fermion.

Le moment cinétique total vaut donc approximativement:

$$\|\vec{J}\| = \|\vec{L} + \vec{S}\| \cong (l+s)\hbar = \sqrt{J(J+1)}\hbar \quad (42.672)$$

Par analogie (c'est vraiment une analogie douteuse...), nous écrivons pour J suffisamment grand...:

$$\|\vec{J}\| \cong (l + s) \hbar = j \hbar \quad (42.673)$$

Mais comme le spin peut avoir que deux orientations possibles, les valeurs de j seront dans le cas d'une particule de spin 1/2:

$$\begin{aligned} j &= l + \frac{1}{2} \text{ spins parallèles} \\ j &= l - \frac{1}{2} \text{ spins anti-parallèles} \end{aligned} \quad (42.674)$$

D'où une classification possible des électrons atomiques tenant compte de leur spin:

Type d'orbitale	s	p	d	f
l	0	1	2	3
j	$\frac{1}{2}$	$\frac{1}{2}, \frac{3}{2}$	$\frac{3}{2}, \frac{5}{2}$	$\frac{5}{2}, \frac{7}{2}$
notation	$S_{1/2}$	$P_{1/2}, P_{3/2}$	$D_{3/2}, D_{5/2}$	$F_{5/2}, F_{7/2}$

Tableau: 42.3 - Types d'orbitales et spin

etc... Soit sous forme schématique avec les niveaux d'énergie correspondants:

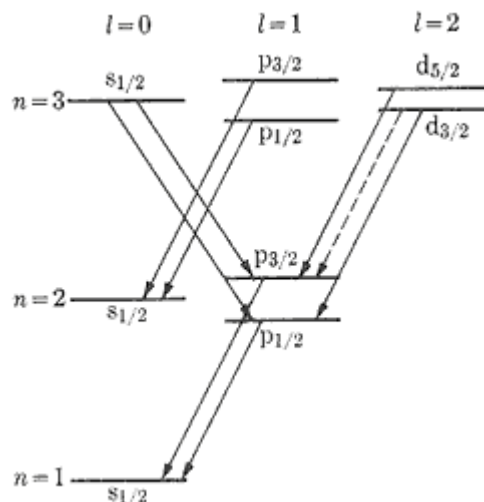


Figure: 42.13 - Forme schématique des orbitales et spin associés

Ce tableau nous amène à constater que nous avons finalement:

$$|l - s| \leq j \leq |l + s| \quad (42.675)$$

Pour revenir à des considérations plus pratiques... nous avons finalement obtenu pour la norme du moment cinétique total (dans le cas d'une particule seule et sans spin):

$$J \leq \sqrt{l(l+1)} \hbar \quad (42.676)$$

où l est un entier. Nous savons également de par le chapitre de Physique Quantique Corpusculaire que le moment magnétique est lui donné par:

$$\mu_l = m_l \hbar \quad (42.677)$$

et que le nombre quantique secondaire l et le nombre quantique magnétique m_l sont d'une certaine manière indissociables.

De la même manière nous obtenons:

$$S \leq \sqrt{s(s+1)} \hbar \quad (42.678)$$

où nous s ne peut prendre pour une particule comme l'électron que les valeurs:

$$s = \pm \frac{1}{2} \quad (42.679)$$

qui correspondent simplement aux deux valeurs propres de la matrice:

$$\frac{\hbar}{2} \sigma_i \quad (42.680)$$

qui lie l'opérateur de spin aux matrices de Pauli de par l'équation de Dirac comme nous le démontrerons dans le chapitre la chapitre de Physique Quantique Relativiste:

$$S_i = \frac{\hbar}{2} s_i = \frac{\hbar}{2} \sigma_i \quad (42.681)$$

Maintenant, ce que nous savons de nos résultats obtenus dans le chapitre de Physique Quantique Corpusculaire c'est que lorsque l vaut 1 nous avons le moment magnétique qui peut prendre trois valeurs différentes suivant qu'un champ magnétique est appliqué ou non:

$$m_l = -1, 0, +1 \quad (42.682)$$

À ce moment, bien que la norme du moment cinétique total reste constante (car conservative), ses composantes doivent forcément changer. Comme nous ne pouvons connaître qu'une seule des composantes du moment cinétique en connaissant sa norme (opérateurs qui commutent) nous choisissons de nous intéresser par convention pédagogique à J_x .

Nous choisissons un référentiel tel qu'une des composantes spatiales soit nulle (c'est toujours possible). Il suffit ensuite par exemple dans le référentiel plan X, Z choisi (donc la composante Y sera nulle) d'avoir la norme de J qui vaut pour $l = 1$:

$$J \leq \sqrt{1(1+1)} \hbar = \sqrt{2} \hbar \quad (42.683)$$

et idem avec S en imposant que la norme vaille pour $s = 1/2$:

$$S \leq \sqrt{\frac{1}{2} \left(\frac{1}{2} + 1 \right)} \hbar = \frac{\sqrt{3}}{2} \hbar \quad (42.684)$$

Il y a alors trois possibilités pour arriver au même résultat en appliquant simplement la norme euclidienne si une des composantes est toujours imposée comme nulle! C'est que nous ayons:

$$\left. \begin{array}{l} J_x = \hbar, J_z = \hbar \\ J_x = \sqrt{2}\hbar, J_z = 0 \\ J_x = \hbar, J_z = -\hbar \end{array} \right\} \|\vec{J}\| = \sqrt{2}\hbar \quad (42.685)$$

Ce que nous pouvons aussi écrire en introduisant le nombre quantique de projection orbital (qui quantifie donc la projection du moment cinétique orbital selon Z et est en multiplicité $2l + 1$):

$$\left. \begin{array}{ll} J_x = \hbar, J_z = m_l \hbar & m_l = +1 \\ J_x = \sqrt{2}\hbar, J_z = m_l \hbar & m_l = 0 \\ J_x = \hbar, J_z = m_l \hbar & m_l = -1 \end{array} \right\} \|\vec{J}\| = \sqrt{2}\hbar \quad (42.686)$$

Ce que les physiciens aiment bien représenter de manière très simplifiée par le schéma suivant:

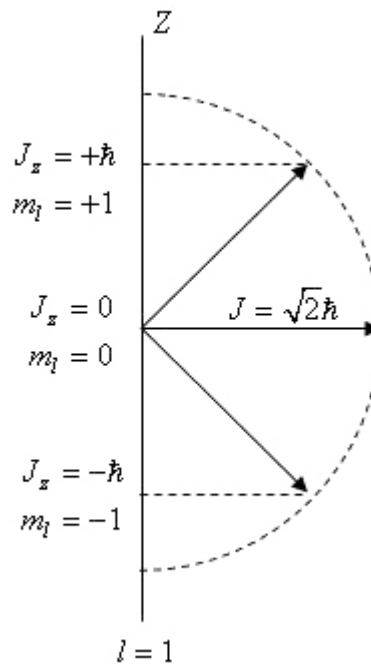


Figure: 42.14 - Représentation schématique simplifiée de la quantification du moment cinétique total

Mais qui en réalité (de par le carré des composantes de la norme) devrait se dessiner sous la forme suivante:

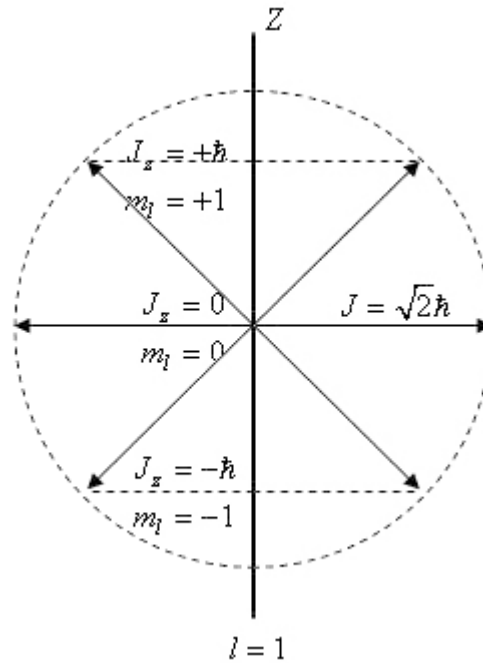


Figure: 42.15 - Représentation schématique complète de la quantification du moment cinétique total

Ce qui nous permet de constater au passage par symétrique que:

$$\langle J_x \rangle = \langle J_y \rangle = 0 \quad (42.687)$$

Remarquons qu'avec de la trigonométrie élémentaire, que nous avons:

$$\theta = \arccos\left(\frac{J_z}{\|\vec{J}\|}\right) = \arccos\left(\frac{m_l \hbar}{\hbar \sqrt{J_x^2 + J_z^2}}\right) = \arccos\left(\frac{m_l}{\sqrt{l(l+1)}}\right) \quad (42.688)$$

Et donc les angles prennent les valeurs suivantes:

$$\arccos\left(\frac{-l}{\sqrt{l(l+1)}}\right), \arccos\left(\frac{-l+1}{\sqrt{l(l+1)}}\right), \dots, \frac{\pi}{2}, \dots, \arccos\left(\frac{l-1}{\sqrt{l(l+1)}}\right), \arccos\left(\frac{l}{\sqrt{l(l+1)}}\right) \quad (42.689)$$

Enfin, indiquons que nous avons alors dans ce cas particulier où $l = 1, m_l = +1$, le système "abstrait" que nous avons utilisé plus haut:

$$\begin{aligned} \|\vec{J}\|^2 |K, m\rangle &= K |K, m\rangle \\ J_z |K, m\rangle &= m |K, m\rangle \end{aligned} \quad (42.690)$$

devient alors dans le cas de notre exemple particulier:

$$\begin{aligned}\|\vec{L}\|^2 |l, m_l\rangle &= \hbar^2 l(l+1) |l, m_l\rangle \Rightarrow \|\vec{L}\|^2 |1, +1\rangle = 2\hbar^2 |1, +1\rangle \\ L_z |l, m_l\rangle &= \hbar l |l, m_l\rangle \Rightarrow L_z |1, +1\rangle = \hbar |1, +1\rangle\end{aligned}\quad (42.691)$$

Ce qui peut être noté de façon condensée sous la forme suivante:

$$\begin{aligned}\hat{L}^2 |1, +1\rangle &= 2\hbar^2 |1, +1\rangle \\ \hat{L}_z |1, +1\rangle &= \hbar |1, +1\rangle\end{aligned}\quad (42.692)$$

De la même façon avec le spin 1/2 nous avons en introduisant le nombre quantique de spin (qui quantifie donc la projection du moment cinétique de spin selon Z et peut prendre autant de valeurs qu'il y a entre -s et +s mais par pas de 1 comme l'impose les résultats expérimentaux, raison pour lesquels il n'y pas de composante nulle en Z ci-dessous):

$$\left. \begin{aligned} S_x &= \frac{\hbar}{\sqrt{2}}, S_z = m_s \hbar & m_s &= +\frac{1}{2} \\ S_x &= \frac{\hbar}{\sqrt{2}}, S_z = m_s \hbar & m_s &= -\frac{1}{2} \end{aligned} \right\} \|\vec{S}\| = \frac{\sqrt{3}}{2} \hbar \quad (42.693)$$

Le lecteur pourra aisément vérifier que le nombre quantique de projection de spin est aussi de multiplicité $2s + 1$.

Ce que les physiciens aiment aussi bien représenter de manière très simplifiée par le schéma suivant:

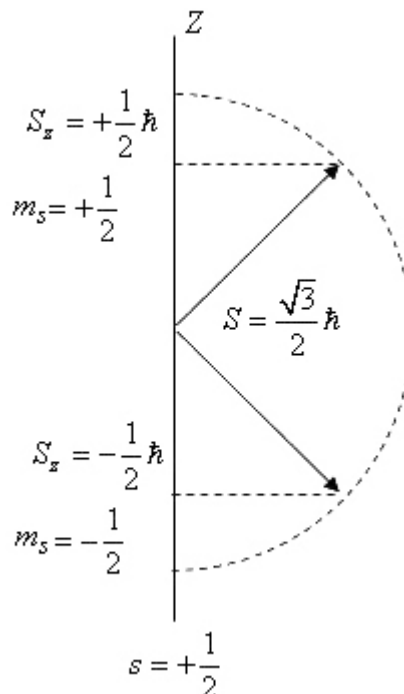


Figure: 42.16 - Représentation schématique de la quantification du spin

avec pour angle:

$$\arccos\left(\frac{m_s}{\sqrt{s(s+1)}}\right) = \arccos\left(\frac{\pm\frac{1}{2}}{\sqrt{\frac{1}{2}\left(\frac{1}{2}+1\right)}}\right) = \arccos\left(\frac{\pm\frac{1}{2}}{\frac{\sqrt{3}}{2}}\right) = \arccos\left(\pm\frac{1}{\sqrt{3}}\right) \cong \pm 54.73^\circ \quad (42.694)$$

De même qu'avant, nous avons dans ce cas particulier où $s = \frac{1}{2}, m_s = \pm\frac{1}{2}$:

$$\begin{aligned} \|\hat{S}\|^2 |s, m_s\rangle &= \hbar^2 s(s+1) |s, m_s\rangle \Rightarrow \|\hat{S}\|^2 \left|\frac{1}{2}, \pm\frac{1}{2}\right\rangle = \frac{3}{4} \hbar^2 \left|\frac{1}{2}, \pm\frac{1}{2}\right\rangle \\ S_z |s, m_s\rangle &= \hbar s |s, m_s\rangle \Rightarrow S_z \left|\frac{1}{2}, \pm\frac{1}{2}\right\rangle = \frac{\hbar}{2} \left|\frac{1}{2}, \pm\frac{1}{2}\right\rangle \end{aligned} \quad (42.695)$$

Ce qui peut être noté de façon condensée sous la forme suivante:

$$\begin{aligned} \hat{S}^2 \left|\frac{1}{2}, \pm\frac{1}{2}\right\rangle &= \frac{3}{4} \hbar^2 \left|\frac{1}{2}, \pm\frac{1}{2}\right\rangle \\ \hat{S}_z \left|\frac{1}{2}, \pm\frac{1}{2}\right\rangle &= \frac{\hbar}{2} \left|\frac{1}{2}, \pm\frac{1}{2}\right\rangle \end{aligned} \quad (42.696)$$

Nous avons donc les seuls éléments variables mesurables expérimentalement qui sont:

$$\boxed{L_z = m_l \hbar} \quad \text{et} \quad \boxed{S_z = s \hbar} \quad (42.697)$$

qui sont donc des observables discrètes (bivaluées en ce qui concerne le spin).

Avec une vue d'artiste du concept pour le plaisir des yeux:

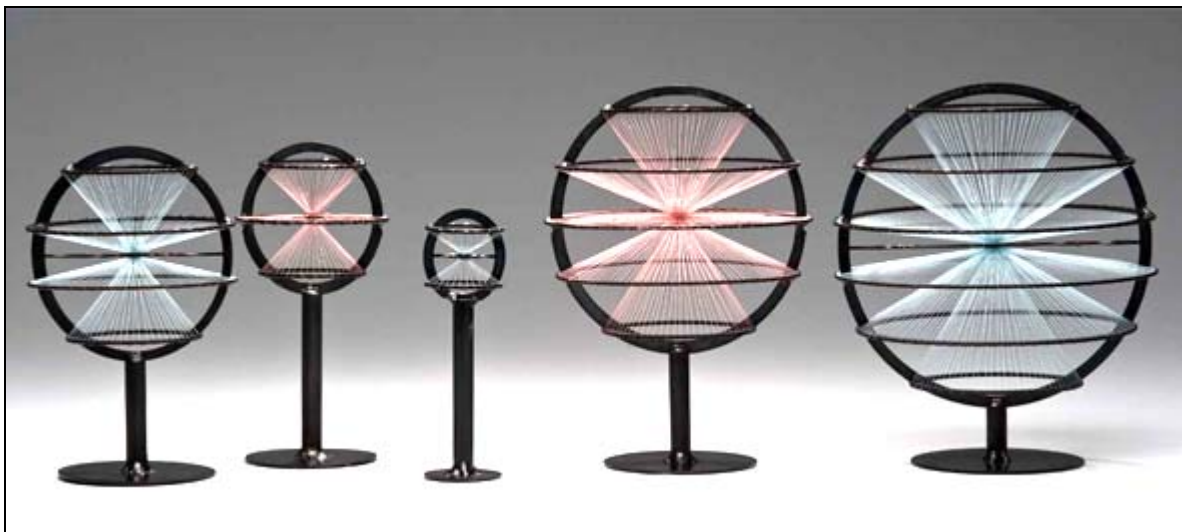


Figure: 42.17 - Représentation schématique de diverses quantifications du moment cinétique par le physicien et sculpteur Julian Voss-Andreae

Donc en appliquant un champ magnétique, l'hamiltonien de Pauli (cf. chapitre de Physique Quantique Relativiste) effectuera des sauts équivalents à la relation:

$$E_p = -\frac{q\hbar}{2m_0} B (m_l + 2s) \quad (42.698)$$

Ce résultat signifie que les niveaux d'énergie pour une énergie donnée (couche n) sont séparés en plusieurs niveaux distants de $\frac{q\hbar}{2m_0} B$ quand l'atome est placé dans un champ magnétique. Ce résultat est l'effet Zeeman dont nous avons parlé plusieurs fois.

Tout cela permet de mieux comprendre l'origine mathématique des 4 nombres quantiques (nombre quantique principal, nombre quantique secondaire ou azimutal, nombre quantique magnétique, spin):

$$n, l, m_l, m_s \quad (42.699)$$

notés aussi (puisque dans le cas particulier des particules étudiées sur ce site le nombre quantique magnétique de projection de spin à la même valeur que le spin puisque nous traitons majoritairement de l'électron):

$$n, l, m_l, s \quad (42.700)$$

Avec pour résumer un peu tout cela...:

$$\begin{aligned} J &= L + S \\ n &= 1, 2, 3, 4, \dots \\ 0 &\leq l \leq n-1 \\ j &= 0, \frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, \dots \\ s &= 0, \frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, \dots \\ |l-s| &\leq j \leq |l+s| \\ m_l &= -l, -l+1, -l+2, \dots, 0, \dots, l-2, l-1, l \\ -l &\leq m_l \leq +l \\ m_s &= -s, -s+1, -s+2, \dots, 0, \dots, s-2, s-1, s \\ -s &\leq m_s \leq +s \end{aligned} \quad (42.701)$$

5.1. COUPLAGE SPIN-ORBITE

Nous avons fait remarquer dans le chapitre de physique quantique corpusculaire que quand nous analysons à haute résolution les raies spectrales de l'hydrogène en l'absence d'un quelconque champ extérieur, nous voyons qu'elles sont en fait constituées de doublets très serrés, séparés de $0.016[nm]$. Ce phénomène étant dû à un soi-disant couplage spin-orbite. Il est temps maintenant de voir d'où cela vient. Rappelons que nous avons obtenu précédemment:

$$\vec{J} = \vec{L} + \vec{S} \quad (42.702)$$

Dès lors, la norme (ce qui est mesuré) nous amène à écrire:

$$\|\vec{J}\|^2 = \|\vec{L} + \vec{S}\|^2 = (L_x + S_x)^2 + (L_y + S_y)^2 + (L_z + S_z)^2 \quad (42.703)$$

ce qui nous donne après regroupement:

$$\|\vec{J}\|^2 = \|\vec{L} + \vec{S}\|^2 = \|\vec{L}\|^2 + \|\vec{S}\|^2 + 2\vec{L} \circ \vec{S} \quad (42.704)$$

Le terme $\vec{L} \circ \vec{S}$ est appelé "**couplage spin-orbite**". C'est lui qui lors des mesures très précises fait apparaître un dédoublement des raies dû au couplage entre le spin de l'électron et le moment cinétique orbital (ce n'est pas $\|\vec{S}\|^2$ car ce terme est toujours positif).

Remarque: Lorsque nous avons deux corps en interaction le moment cinétique total est une constante du mouvement. Il peut donc y avoir un transfert de moment cinétique entre ces deux corps (c'est le couplage spin-orbite). L'un perd du moment l'autre en gagne. À noter qu'un corps étendu possède un moment cinétique de rotation autour d'un point et un moment cinétique de rotation sur lui-même. C'est ce dernier que nous appelons par une analogie abusive: le spin.

L'écart mesuré est donc attribué à l'interaction du spin de l'électron avec son moment orbital. L'électron tourne autour du noyau, mais si nous nous plaçons sur l'électron, nous voyons le noyau tourner (sur la Terre le soleil tourne autour de la Terre!). Tout se passe comme si le noyau créait un champ magnétique au niveau de l'électron, et ce champ interagit avec le moment magnétique de l'électron, le spin, et ceci différemment selon que le spin est dans le sens du champ ou opposé, c'est cette différence qui ajoute ou retranche un peu d'énergie au niveau.

Voici un schéma qui résume le tout:

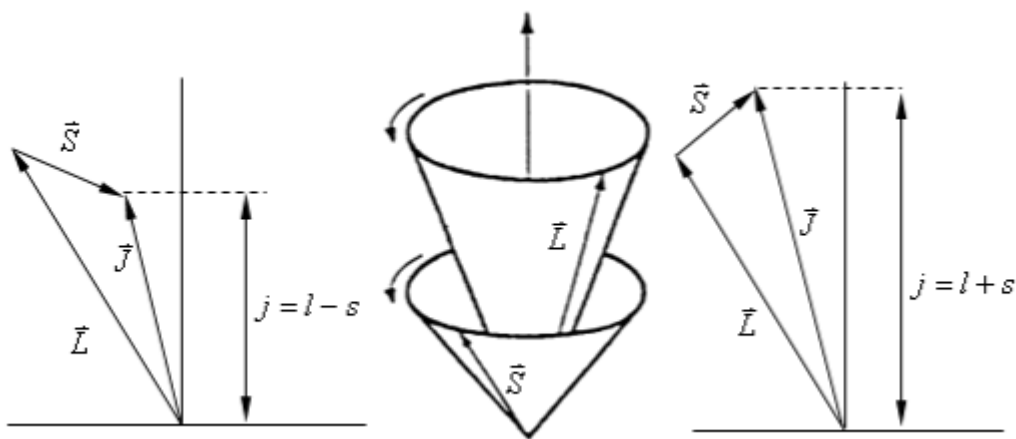


Figure: 42.18 - Représentation imagée de l'interaction spin-orbite

Montrons en effet que $\vec{J}$ tel que défini, est une constante du mouvement. Nous avons (inutile de préciser qu'en mettant au carré, il s'agit des composantes du vecteur que nous mettons au carré et non le vecteur lui-même!):

$$\vec{L} \circ \vec{S} = \frac{1}{2} \left[(\vec{L} + \vec{S})^2 - \vec{L}^2 - \vec{S}^2 \right] = \frac{1}{2} \left[\vec{J}^2 - \vec{L}^2 - \vec{S}^2 \right] \quad (42.705)$$

d'où:

$$[\vec{J}, \vec{L} \circ \vec{S}] = \frac{1}{2} [\vec{J}, \vec{J}^2 - \vec{L}^2 - \vec{S}^2] \quad (42.706)$$

Faisons le développement pour une composante:

$$\begin{aligned} \frac{1}{2} [J_x, J_x^2 - L_x^2 - S_x^2] &= \frac{1}{2} [J_x (J_x^2 - L_x^2 - S_x^2) - (J_x^2 - L_x^2 - S_x^2) J_x] \\ &= \frac{1}{2} [\cancel{J_x^3} - J_x L_x^2 - J_x S_x^2 - \cancel{J_x^3} + L_x^2 J_x + S_x^2 J_x] = -\frac{1}{2} [J_x L_x^2 + J_x S_x^2 - L_x^2 J_x - S_x^2 J_x] \quad (42.707) \\ &= -\frac{1}{2} [J_x, L_x^2 + S_x^2] \end{aligned}$$

Or, par définition (de notation) $J_x = L_x + S_x$ donc:

$$\begin{aligned} [\vec{J}, \vec{L} \circ \vec{S}]_x &= -\frac{1}{2} [J_x, L_x^2 + S_x^2] = -\frac{1}{2} [L_x + S_x, L_x^2 + S_x^2] \\ &= -\frac{1}{2} [(L_x L_x^2 + L_x S_x^2 + S_x L_x^2 + S_x S_x^2) - (L_x^2 L_x + S_x^2 L_x + L_x^2 S_x + S_x^2 S_x)] \quad (42.708) \\ &= -\frac{1}{2} ([L_x, L_x^2] + [S_x, S_x^2] + [L_x, S_x^2] + [S_x, L_x^2]) \end{aligned}$$

Or, nous savons que $[L_x, L_x^2] = 0$ (car un opérateur commute toujours avec lui-même) et en ce qui concerne $[S_x, S_x^2]$, nous en avons fait mention dans le chapitre de Calcul Spinoriel et nous le démontrerons dans le cadre de l'étude de l'équation de Dirac libre classique (cf. chapitre de Physique Quantique Relativiste), que le spin est totalement décrit par les matrices de Pauli qui sont des opérateurs linéaires. Écrivons alors à un facteur constant près:

$$S_x = c^{te} \cdot \sigma_x = c^{te} \cdot \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad (42.709)$$

et nous verrons que cela est bien conforme à l'équation de Pauli que nous verrons dans le chapitre de Physique Quantique Relativiste (et inversement)!!!

Donc en faisant abstraction de la constante multiplicative:

$$[S_x, S_x^2] = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \right)^2 - \left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \right)^2 \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = 0 \quad (42.710)$$

ce qui était de toute façon 100% prévisible puisque de toute façon, encore une fois, un même opérateur commute toujours avec lui-même.

Donc finalement:

$$[\vec{J}, \vec{L} \circ \vec{S}]_x = -\frac{1}{2} ([L_x, S_x^2] + [S_x, L_x^2]) \quad (42.711)$$

Dès lors:

$$\begin{aligned} [L_x, S_x^2] &= L_x \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}^2 - \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}^2 L_x = L_x \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} L_x = 0 \\ [S_x, L_x^2] &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \cdot L_x^2 - L_x^2 \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = 0 \end{aligned} \quad (42.712)$$

d'où finalement:

$$[\vec{J}, \vec{L} \circ \vec{S}] = 0 \quad (42.713)$$

$\vec{J}$ est bien le moment cinétique total qui, même en présence d'interaction spin-orbite, est une constante du mouvement (une obligation pour un système isolé).

Remarque: Une autre manière de lire la chose consiste à dire que la mesure sur un des éléments du commutateur précédent adapte l'autre immédiatement pour que leur commutation soit nulle donc par extension le moment cinétique total est une constante du mouvement.

Revenons maintenant sur la relation démontrée plus haut:

$$\begin{aligned} \hat{S}^2 \left| \frac{1}{2}, \pm \frac{1}{2} \right\rangle &= \frac{3}{4} \hbar^2 \left| \frac{1}{2}, \pm \frac{1}{2} \right\rangle \\ \hat{S}_z \left| \frac{1}{2}, \pm \frac{1}{2} \right\rangle &= \frac{\hbar}{2} \left| \frac{1}{2}, \pm \frac{1}{2} \right\rangle \end{aligned} \quad (42.714)$$

Nous avons certes obtenues les valeurs propres. Mais il serait judicieux de déterminer l'expression des opérateurs de Spin. Nous avons vu dans le chapitre de Calcul Spinoriel que:

$$\sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (42.715)$$

et nous avons démontré que pour la valeur propre +1 les vecteurs propres associés étaient:

$$\pm \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (42.716)$$

Nous avons alors:

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (42.717)$$

En multipliant à gauche et droite par un terme familier:

$$\frac{\hbar}{2} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{\hbar}{2} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (42.718)$$

Et par analogie du fait que les matrices de Pauli sont des matrices de rotations particulières, nous posons:

$$\hat{S}_z = \frac{\hbar}{2} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (42.719)$$

Et évidemment nous en déduisons:

$$\hat{S}^2 = \|\vec{\hat{S}}\|^2 = \left(\frac{\hbar}{2}\sigma_x\right)^2 + \left(\frac{\hbar}{2}\sigma_y\right)^2 + \left(\frac{\hbar}{2}\sigma_z\right)^2 = \frac{\hbar^2}{4}(\sigma_x^2 + \sigma_y^2 + \sigma_z^2) = \frac{3\hbar^2}{4} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (42.720)$$

6. DIMENSIONS DE PLANCK

Il convient d'ouvrir une petite parenthèse pour finir sur la constante de Planck (car beaucoup d'ouvrages font mention de ce que nous allons voir sans les précautions de rigueur). Nous venons de voir que la mesure des objets dépend du principe d'indétermination de Heisenberg. Cette précision joue tant sur les mesures du temps que sur la trajectoire des particules ou la densité d'énergie de l'Univers. Voyons que cela à, par extension... d'autres éventuelles implications.

Nous avons démontré précédemment au début de ce chapitre qu'une des relations d'incertitudes est donnée, en prenant le module, par (de l'ordre de la constante de Planck donc à un facteur près):

$$[x, p_x] = \hbar \quad (42.721)$$

Grossièrement, nous pouvons donc dire qu'à une fluctuation λ de l'espace (à ne pas confondre avec la notation de la longueur d'onde), nous pouvons associer la quantité de mouvement:

$$p_x = \frac{\hbar}{\lambda} \quad (42.722)$$

À celle-ci correspond, d'après nos résultats du chapitre de Relativité Restreinte, la relation de l'énergie $E = Mc^2 = p \cdot c$, ou la masse équivalente (en divisant par c^2) p/c . En désignant par M cette masse associée à la perturbation λ , nous avons donc:

$$M = \frac{\hbar}{c \cdot \lambda} \quad (42.723)$$

La gravitation due à cette masse est caractérisée par une longueur R que nous déterminerons en ordre de grandeur en écrivant que l'énergie potentielle qui lui est associée (cela suppose que la gravitation classique et quantique sont régies par les mêmes lois...), GM^2/R (cf. chapitre de Mécanique Classique), est égale à la masse-énergie Mc^2 . Cela donne:

$$R = \frac{GM}{c^2} \quad (42.724)$$

ou, en remplaçant M par son expression précédente:

$$R = \frac{G\hbar}{c^3 \lambda} \quad (42.725)$$

Pour qu'il n'y ait pas auto-amplification (et donc divergence) du phénomène de fluctuation quantique du vide, nous devons avoir de préférence $R = \lambda$. En écrivant l'égalité entre ces deux grandeurs, nous aboutissons donc à une quantité qui représente la dimension minimale (en ordre de grandeur) que puisse concevoir la physique. C'est la fameuse "**longueur de Planck**":

$$\lambda_p = \sqrt{\frac{G\hbar}{c^3}} \cong 1.616 \cdot 10^{-35} [m] \quad (42.726)$$

pour laquelle il correspond la période ou "**temps de Planck**" $t = \lambda / c$ d'où:

$$t_p = \sqrt{\frac{G\hbar}{c^5}} = 5.391 \cdot 10^{-44} [s] \quad (42.727)$$

Nous pouvons maintenant revenir à une autre expression plus intéressante de la masse fluctuante. Puisque:

$$M = \frac{\hbar}{c \cdot \lambda} \Rightarrow \lambda = \frac{\hbar}{c \cdot M} \text{ et } \lambda = \sqrt{\frac{G\hbar}{c^3}} \quad (42.728)$$

nous avons dès lors la "**masse de Planck**":

$$M_p = \sqrt{\frac{\hbar c}{G}} \cong 2.176 \cdot 10^{-8} [Kg] \quad (42.729)$$

L'analyse dimensionnelle nous donne à une constante près et selon le théorème du Viriel (cf. chapitre de Mécanique Des Milieux Continus):

$$E = kT \quad (42.730)$$

et donc:

$$E = M_p c^2 = pc = kT \Rightarrow \frac{p}{c} = \frac{kT}{c^2} = M_p \Rightarrow \sqrt{\frac{\hbar c}{G}} = \frac{kT}{c^2} \quad (42.731)$$

d'où la "**température de Planck**":

$$T_p = \sqrt{\frac{\hbar c^5}{Gk}} \cong 1.42 \cdot 10^{32} [K] \quad (42.732)$$

et encore "**l'énergie de Planck**":

$$E_p = M_p c^2 = \sqrt{\frac{\hbar c}{G}} c^2 = \sqrt{\frac{\hbar c^5}{G}} = 1.956 \cdot 10^9 [J] \quad (42.733)$$

Après tout cela, nous obtenons facilement la "**densité de Planck**":

$$\rho_P = \frac{M_P}{\lambda_P^3} = \frac{\sqrt{\frac{\hbar c}{G}}}{\sqrt{\frac{G\hbar}{c^3}}^3} = \frac{\sqrt{\frac{\hbar c}{G}} \sqrt{\frac{c^3}{G\hbar}}}{\frac{G\hbar}{c^3}} = \sqrt{\frac{c^4}{G^2}} \frac{c^3}{G\hbar} = \sqrt{\frac{c^{10}}{G^4 \hbar^2}} = \frac{c^5}{G^2 \hbar} \cong 5.1 \cdot 10^{96} \text{ [kg} \cdot \text{m}^{-3}] \quad (42.734)$$

Nous pouvons nous amuser à obtenir encore d'autres valeurs de Planck encore, mais qui ne veulent plus dire grand-chose à force (et nous pourrions continuer ainsi longtemps avec énormément d'autres grandeurs):

La "force de Planck":

$$F_P = M_P \frac{\lambda_P}{t_P^2} = \sqrt{\frac{\hbar c}{G}} \frac{\sqrt{\frac{G\hbar}{c^3}}}{\left(\sqrt{\frac{G\hbar}{c^5}}\right)^2} = \frac{c^4}{G} \cong 1.21 \cdot 10^{44} \text{ [N]} \quad (42.735)$$

La "puissance de Planck":

$$P_P = \frac{E_P}{t_P} = \frac{c^5}{G} = 3.629 \cdot 10^{52} \text{ [W]} \quad (42.736)$$

La "pulsation de Planck":

$$\omega_P = 2\pi f_P = \frac{1}{t_P} = \sqrt{\frac{c^5}{\hbar G}} = 1.855 \cdot 10^{43} \text{ [rad} \cdot \text{s}^{-1}] \quad (42.737)$$

En procédant avec le même raisonnement initial fait avec la masse, mais en utilisant l'énergie potentielle électrostatique au lieu de l'énergie potentielle gravitationnelle nous pouvons obtenir la "charge de Planck":

$$q_P = \sqrt{c \hbar 4 \pi \epsilon_0} = 1.875 \cdot 10^{-18} \text{ [C]} \quad (42.738)$$

Dès lors nous pouvons calculer un "courant de Planck":

$$I_P = \frac{Q_P}{T_P} = \sqrt{\frac{c^6 4 \pi \epsilon_0}{G}} = 3.479 \cdot 10^{25} \text{ [A]} \quad (42.739)$$

ainsi que la "tension de Planck":

$$U_P = \frac{E_P}{q_P} = \sqrt{\frac{c^4}{G 4 \pi \epsilon_0}} = 1.0432 \cdot 10^{27} \text{ [V]} \quad (42.740)$$

et "l'impédance de Planck" (...):

$$Z_P = \frac{U_P}{I_P} = \frac{1}{4 \pi \epsilon_0 c} = \frac{Z_0}{4 \pi} = 29.98 \text{ [\Omega]} \quad (42.741)$$

Remarque: Certains physiciens se sont servis (et se servent toujours) des résultats ci-dessus pour des raisonnements farfelus et dangereux qui ne sont qu'interprétation. Il convient donc de prendre avec des pincettes toutes les informations relatives aux dimensions de Planck que vous pourriez trouver (même si celles-ci paraissent fort sympathiques). L'exemple le plus connu est donné par la longueur d'onde de Compton λ_c (cf. chapitre de Physique Nucléaire) qui dépend de la masse-énergie du photon. Si cette longueur d'onde est égale au rayon de Schwarzschild classique pour la même masse-énergie (cf. chapitre d'Astrophysique), alors dans ce cas sa valeur est celle de la longueur de Planck et sa masse est égale à la masse de Planck. Il est alors tentant de dire que la particule forme alors un trou noir. Mais il s'agit d'une analogie car dans ce cas, rien ne nous dit que l'expression du rayon de Schwarzschild s'applique à la physique quantique...

7. INTERPRÉTATION DE COPENHAGUE

En 1930, l'interprétation probabiliste de l'amplitude de l'onde d'une particule et le principe d'incertitude d'Heisenberg constituent les éléments de l'interprétation "standard" non déterministe de la physique quantique comme nous en avons déjà fait mention au début de ce chapitre. Cette interprétation est souvent appelée "**interprétation de Copenhague**", car Niels Bohr qui y contribua largement y dirigeait un institut de physique renommé à cette époque. Pourtant de nombreux physiciens tels Einstein et Schrödinger, qui acceptaient la formulation mathématique de la physique quantique, n'étaient pas à l'aise avec l'interprétation de Copenhague et la critiquaient. Et jusqu'à nos jours, la question de l'interprétation correcte de la formulation mathématique reste un problème.

En effet, nous pouvons nous poser la question suivante: Où se trouve la réalité? Y a-t-il une réalité? Niels Bohr répond non: il n'y a rien au niveau quantique, la réalité n'existe ou n'apparaît que lors d'une mesure. Cette vision partagée par la plupart des physiciens (interprétation de Copenhague), implique que la mesure "crée" la position de l'électron (voir le sous-chapitre traitant du principe de superposition linéaire des états). Autrement dit: Aucun phénomène élémentaire n'est un phénomène réel avant d'être un phénomène observé.

Einstein pensait que la physique quantique, bien que très efficace et très impressionnante, n'est pas complète et ne donne qu'une image imparfaite du monde quantique. Pour lui, il y aurait autre chose, au-delà, qui clarifierait et affinerait notre présente vision (au même titre que la théorie des gaz pour laquelle il avait fallu attendre les modèles statistiques, Einstein pensait qu'il restait à découvrir des variables cachées)

Ainsi, dans l'interprétation de Copenhague de la mécanique quantique le principe d'incertitude signifie qu'à un niveau élémentaire, l'univers physique n'existe plus de manière déterministe, mais plutôt comme une série de probabilités ou de potentiels. Par exemple, le motif produit par des millions de photons passant à travers une fente de diffraction peut être calculé à l'aide de la mécanique quantique, mais le chemin de chaque photon ne peut être prédit par aucune méthode connue. L'interprétation de Copenhague dit qu'il ne pourra être calculé par aucune méthode. C'est cette interprétation qu'Einstein mettait en doute lorsqu'il disait: "je ne peux pas croire que Dieu joue aux dés avec l'Univers". D'un point de vue physique autant que philosophique, le principe d'incertitude implique la réfutation du déterminisme universel défendu par Laplace au début du 19ème siècle.

Une réduction instantanée de tous les états possible se produit dès l'observation du système selon l'interprétation de Copenhague. Cette décision aléatoire de l'état observé respecte les probabilités, correspondant au carré des amplitudes des états. De surcroît, l'interprétation de Copenhague stipule que, lors d'une mesure, un processus de réduction, originaire de l'objet macroscopique, élimine les superpositions d'états quantiques.

L'interprétation de l'école de Copenhague conduit donc au problème de la mesure, l'expérience de pensée du chat de Schrödinger stipulant que lorsqu'on mesure une quantité, telle que la position ou l'impulsion, nous intervenons dans le processus de mesure en provoquant un changement radical de l'état quantique, de la fonction d'onde. Nous modifions les quantités mesurées de façon imprévisible et cet état ne peut être décrit par l'équation déterminée de Schrödinger. Les physiciens et les philosophes ont réagi de plusieurs manières à cette interprétation:

- Soit nous considérons comme Bohr et Heisenberg que ce principe fait loi et qu'il est préférable de ne pas rechercher l'interprétation ultime. C'est une attitude qui est admise par la plupart des physiciens.
- Soit nous considérons que la physique quantique est une théorie incomplète et certains, tel Einstein, Eugène Wigner ou David Bohm n'ont pas hésité à rechercher d'autres solutions, stériles jusqu'à présent.
- Enfin, Hugh Everett et bien d'autres prennent l'équation de Schrödinger très au sérieux, la considérant comme une représentation de la réalité. Ils considèrent que l'interprétation de l'école de Copenhague représente réellement l'évolution de la fonction d'onde. Les différents termes de l'équation correspondraient aux différents niveaux d'énergie dans lesquels se trouve le système. La réduction du paquet d'ondes s'interpréterait comme une division totale de l'objet et de l'instrument de mesure dans des univers parallèles.

Aujourd'hui le débat reste ouvert, mais plusieurs expériences réalisées depuis les années 1930 nous permettent, pas à pas, de dissiper l'épais brouillard qui recouvre le fond de la réalité et de répondre à quelques questions. Cela dit, toutes ces expériences confirment néanmoins que l'époque des certitudes est bien révolue. L'expérience la plus fameuse restant le paradoxe EPR suite à la publication d'un article de Einstein, Podolsky et Rosen ayant pour seul objectif de mettre à mal l'interprétation de Copenhague.

L'article d'origine étant un peu difficile, nous allons prendre la version scolaire d'usage simplifiée mais qui est celle utilisée dans les laboratoires, proposée à l'origine par David Bohm. Alors qu'à l'origine le paradoxe était présenté avec le couple {position, quantité de mouvement}, Bohm proposa d'utiliser le spin qui est une propriété à priori purement quantique.

Un électron, ne pouvant alors avoir que deux états de spin "en haut" ou "en bas", l'expérience EPR proposée par Bohm consiste alors à prendre une particule de spin nul qui se désintègre, produisant ainsi deux électrons *A* et *B*. Puisque leur spin combiné doit demeurer égal à zéro, l'un des électrons doit avoir son spin en haut et l'autre en bas. Les électrons foncent dans des directions opposées jusqu'à ce que la distance les séparant soit assez grande pour éliminer toute interaction physique entre eux, et on mesure le spin de chaque électron exactement au même instant à l'aide d'un détecteur de spin.

Selon Bohr, tant qu'aucune mesure n'a été effectuée, ni l'électron *A*, ni l'électron *B* ne possèdent un spin pré-existant dans aucune direction. Au lieu de quoi, avant d'être observés, les électrons existent dans une superposition d'états, si bien qu'ils sont en haut et en bas en même temps. Puisque les deux électrons sont intriqués, l'information concernant l'état de leur spin est donnée par une fonction d'onde du type:

$$\Psi = (A \text{ spin } \uparrow \text{ et } B \text{ spin } \downarrow) \text{ ou } (A \text{ spin } \downarrow \text{ et } B \text{ spin } \uparrow)$$

L'électron A n'a pas de composante x de spin avant qu'une mesure effectuée pour la déterminer fasse s'effondrer la fonction d'onde du système A et B , après quoi elle est soit "en haut", soit "en bas". À cet instant précis, son partenaire B acquiert le spin opposé dans la même direction, même s'il est à l'autre bout de l'Univers. L'interprétation de Copenhague est alors dite "**interprétation non-locale**" alors qu'Einstein croyait au réalisme local: c'est-à-dire qu'une particule ne peut être instantanément influencée par un événement lointain et que ses propriétés existent indépendamment de toute mesure.

L'approche de Bohm a cependant une faille qu'Einstein aurait probablement utilisé comme argument: les corrélations s'expliqueraient en avançant que les deux électrons possèdent chacun des valeurs de spin définies sur chacun des trois axes x , y , z . qu'elles soient mesurées ou non. Donc à nouveau, selon Einstein, le fait que les états de spin pré-existants du couple d'électrons ne puissent être pris en compte par la physique quantique ondulatoire aurait encore été une preuve de son incomplétude.

Un physicien John Bell eut cependant l'idée d'un moyen expérimental et théorique pour sortir de l'impasse du paradoxe EPR en changeant l'orientation relative des deux détecteurs de spin.

Ainsi, si les détecteurs mesurant le spin des électrons A et B sont alignés de façon à être parallèles, alors il y a une corrélation de 100% entre les deux ensembles de mesures chaque fois que le spin en haut est mesuré par un détecteur, le spin en bas est enregistré par l'autre détecteur, et vice versa. Si l'on fait tourner légèrement un des détecteurs, ils ne sont plus alignés. À présent, si on mesure l'état de spin de nombreux couples d'électrons intriqués, lorsqu'on trouve "en haut" pour l'électron A , la mesure correspondante pour B donnera parfois "en haut" elle aussi. Augmenter l'angle entre les axes des deux détecteurs conduit donc à une réduction du degré de corrélation. Si les détecteurs sont à angle droit l'un de l'autre et que l'expérience est à nouveau répétée de nombreuses fois, ce n'est que dans la moitié des cas qu'on détectera un spin en bas chez B lorsqu'on détecte un spin en haut chez A sur l'axe x . Si les détecteurs sont orientés à 180 degrés l'un de l'autre, le couple d'électrons sera totalement anticorrélé. Si la mesure donne "en haut" pour l'état de spin de A , alors le spin de B sera "en bas".

Bien qu'il s'agisse d'une expérience imaginaire, il est possible de calculer le degré exact de corrélation du spin pour une orientation donnée des détecteurs, tel qu'il est prédit par la théorie quantique. Il n'est cependant pas possible d'effectuer un calcul similaire en se servant d'une théorie à variables cachées archétypique et qui conserve la localité. La seule chose que pareille théorie pourrait prédire serait un couplage imparfait entre les états de spin de A et de B . Cependant, en toute rigueur, c'est insuffisant pour choisir entre la théorie quantique et une théorie locale à variables cachées. Bell fit alors une découverte étonnante. Il était possible de décider entre les prédictions de la mécanique quantique et celles de toute théorie à variables cachées en mesurant les corrélations de couples d'électrons pour une configuration donnée des détecteurs et en répétant ensuite l'expérience avec une orientation différente. Ce qui permit à Bell de calculer la corrélation totale pour les deux configurations d'orientation en termes de résultats individuels prédits par toute théorie locale à variables cachées. Puisque, dans toute théorie de cette sorte, le résultat d'une mesure effectuée par un détecteur ne peut être affecté par ce qui est mesuré avec l'autre, il est possible de distinguer entre les variables cachées et la mécanique quantique.

Bell réussit à calculer les limites du degré de corrélation de spin entre couples d'électrons intriqués dans une expérience EPR modifiée par Bohm. Il trouva qu'au royaume éthéré des quanta il y a un plus grand degré de corrélation si la mécanique quantique règne en maîtresse absolue que dans tout univers qui dépend de variables cachées et de la localité. Le théorème de Bell disait qu'aucune théorie locale à variables cachées ne pouvait reproduire le même ensemble de corrélations que la mécanique quantique. Toute théorie locale à variables cachées conduirait à des corrélations de spin générant des nombres, appelés coefficients de corrélation, entre -2 et +2. Or, pour certaines orientations des détecteurs de spin, la mécanique quantique prédit des coefficients de corrélation qui se trouvent à l'extérieur de la plage, appelée "**inégalités de Bell**", allant de -2 à +2.

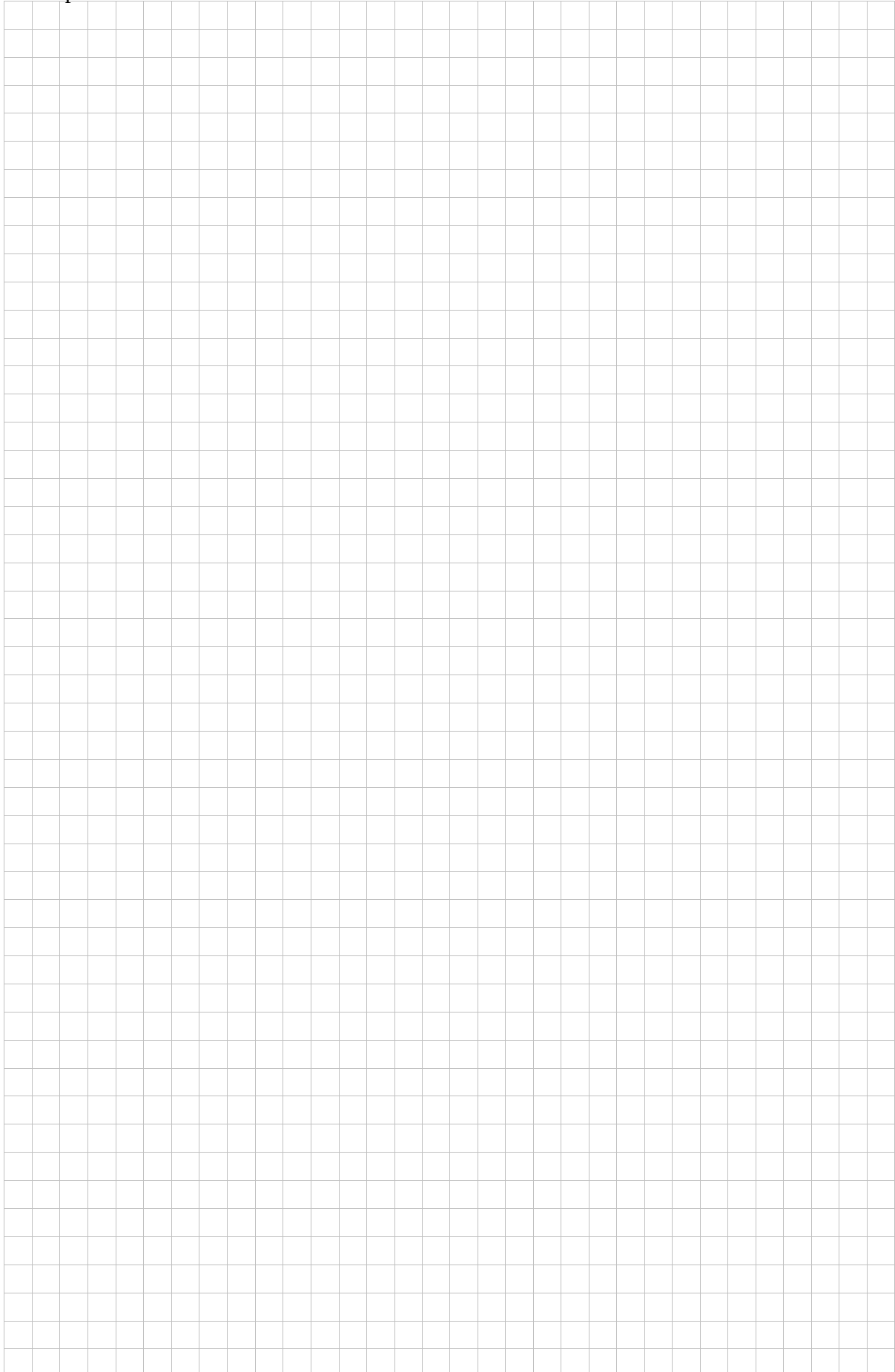
Le théorème de Bell permet donc de tester en face de l'interprétation de Copenhague soutenue par Bohr la réalité locale préconisée par Einstein, à savoir que l'univers quantique existe indépendamment de l'observation et que les effets physiques ne peuvent se transmettre à une vitesse supérieure à celle de la lumière. Bell avait transporté le débat Einstein-Bohr dans une arène nouvelle, la philosophie expérimentale. Si l'inégalité de Bell résistait, alors l'affirmation d'Einstein que la mécanique quantique était incomplète serait exacte. Si toutefois cette inégalité venait à être violée, ce serait Bohr qui triompherait. Plus d'expériences de pensée! Ce serait maintenant Einstein contre Bohr au laboratoire.

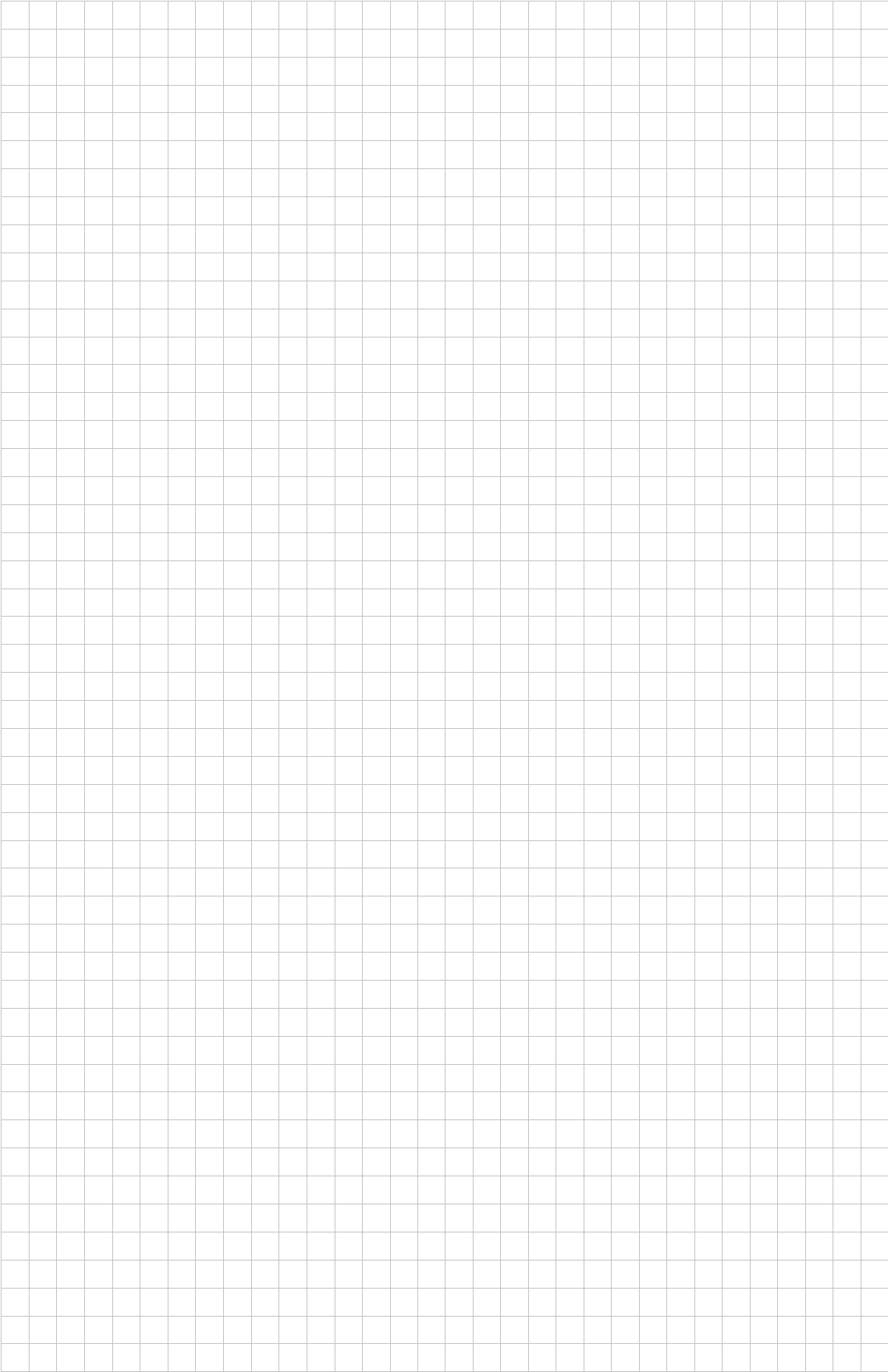
La première expérience qui testa les inégalités de Bell utilisa des couples de photons au lieu de couples d'électrons. Ce changement était possible parce que les photons possèdent le propriété de polarisation, qui, pour les besoins du test, jouait le rôle du spin quantique (de plus les photons sont plus simples à manipuler). C'est certes une simplification, mais on peut considérer un photon comme étant polarisé "en haut" ou "en bas". À l'instar du spin de l'électron, si la polarisation d'un des photons sur l'axe x est mesurée comme étant "en haut", alors la mesure de l'autre donnera "en bas", puisque les polarisations combinées des deux photons doivent aboutir à zéro.

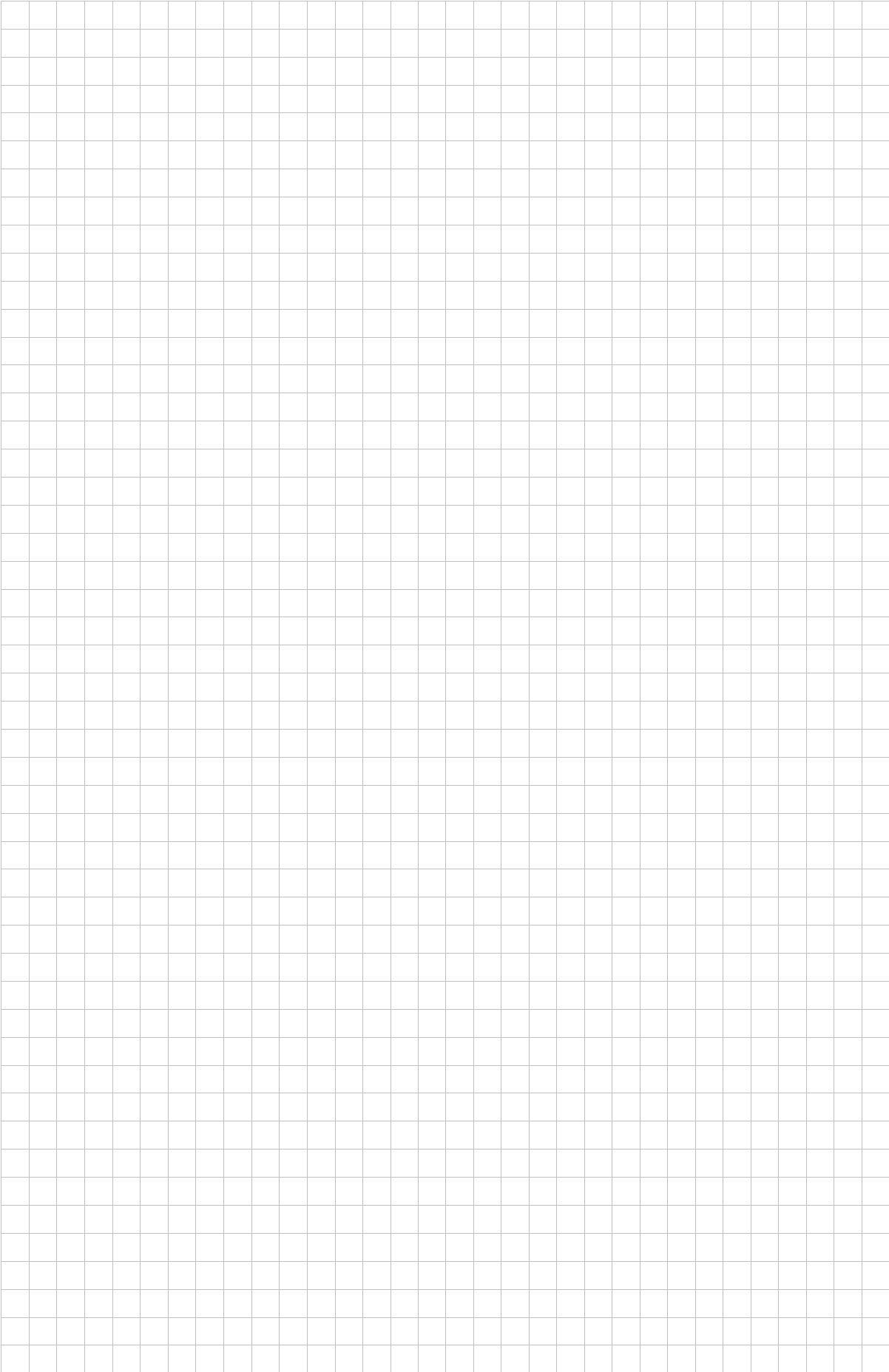
Les résultats violèrent les inégalités de Bell ce qui était en faveur de l'interprétation de Copenhague non locale soutenue par Bohr et contre la réalité locale soutenue par Einstein.

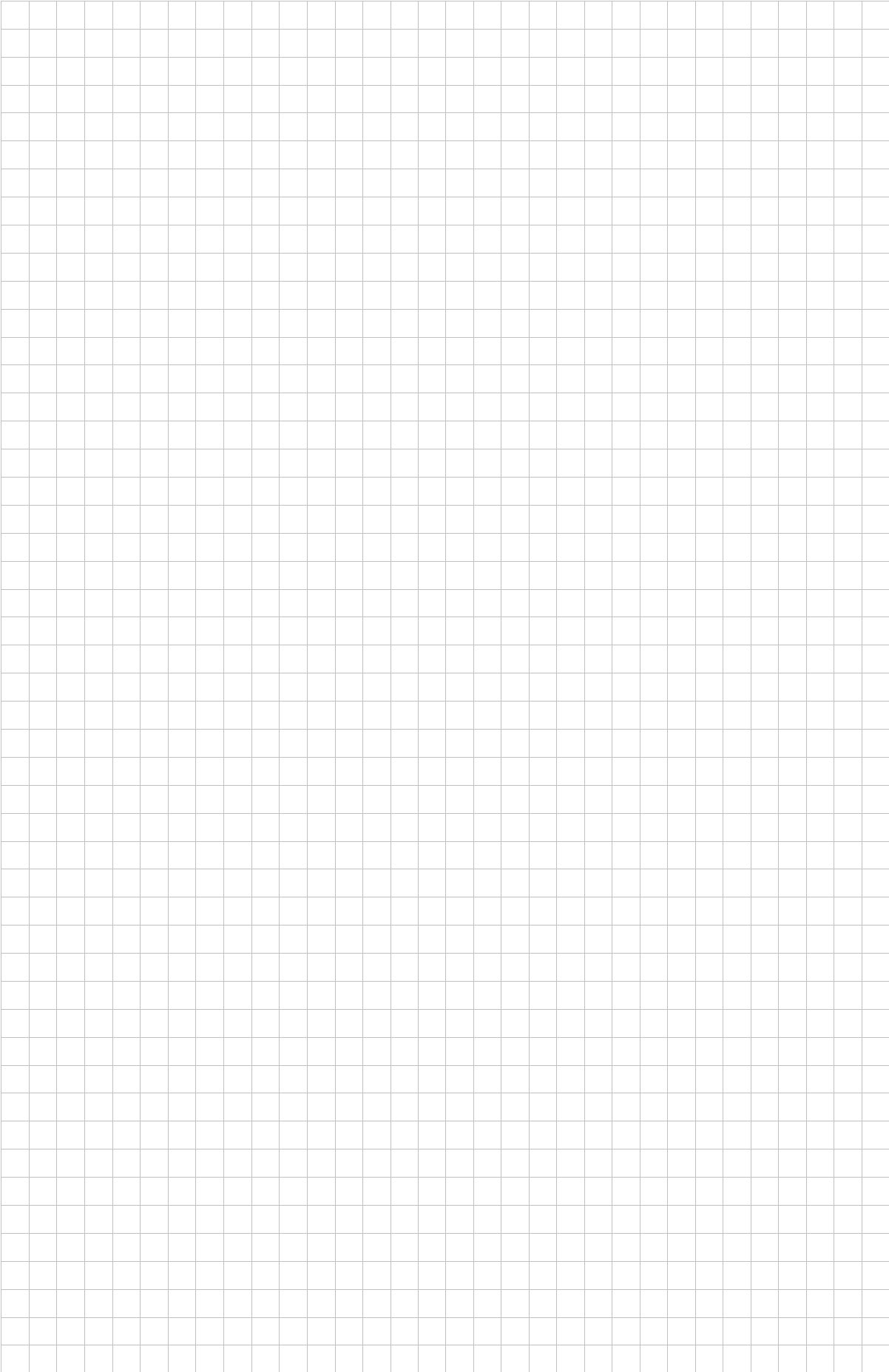
Bell dérivait cette inégalité de deux suppositions. Primo, il existe une réalité indépendante de l'observateur. Ce qui se traduit par le fait qu'une particule possède une propriété bien définie comme le spin avant d'être mesurée. Secundo, la localité est conservée. Il n'y a pas d'influence supraluminique, si bien que ce qui se produit ici ne peut affecter instantanément ce qui se produit ailleurs. Les résultats expérimentaux signifient qu'il faut abandonner l'une de ces deux suppositions, mais laquelle? Bell était disposé à abandonner la localité.

Notes personnelles:









43. PHYSIQUE QUANTIQUE RELATIVISTE

Le lecteur attentif aura noté que la mécanique quantique (physique quantique ondulatoire) est une théorie non relativiste: elle n'incorpore pas les principes de la relativité restreinte d'Einstein (cf. [chapitre de Relativité Restreinte](#)). Nous allons donc nous efforcer à combler ce manque en étudiant maintenant la version relativiste de la physique quantique ondulatoire (probabiliste).

Avant de nous attaquer à la partie mathématique, nous tenons à souligner et à rappeler que nous allons nous limiter uniquement aux développements théoriques effectués entre 1910 et environ 1935 (au-delà la complexité des théories nécessite trop de pages pour un site Internet généraliste).

1. ÉQUATION D'ÉVOLUTION RELATIVISTE DE SCHRÖDINGER

La physique des particules ne peut être correctement et totalement décrite dans le cadre de la mécanique quantique. Comme les énergies sont généralement supérieures aux masses des particules, il est nécessaire, en plus, de travailler dans le contexte de la théorie de la relativité restreinte. Voyons comment inclure celle-ci par une première approche basique.

L'énergie-impulsion d'une particule libre de masse m , satisfait comme nous l'avons démontré dans le chapitre de Relativité Restreinte à la relation:

$$E^2 = c^2 p^2 + m_0^2 c^4 \quad (43.1)$$

Nous cherchons à quantifier cette équation. Pour cela, nous allons revenir à des relations que nous avons démontrées lors de l'étude des opérateurs linéaires fonctionnels et de l'équation évolutive de Schrödinger.

Rappelons que la quantité de mouvement est décrite par la relation (utilisant l'opérateur de divergence):

$$\vec{p} = -i\hbar \vec{\nabla} \quad (43.2)$$

et l'énergie totale par:

$$E = i\hbar \frac{\partial}{\partial t} \quad (43.3)$$

Ces deux relations ayant été démontrées dans le chapitre de Physique Quantique Ondulatoire!

Les substitutions des deux relations précédentes appliquées à la relation:

$$E^2 = c^2 p^2 + m_0^2 c^4 \quad (43.4)$$

et multipliée par la fonction d'onde (cf. [chapitre de Physique Quantique Ondulatoire](#)) des deux côtés de l'égalité conduisent au développement:

$$\begin{aligned}
 E^2 &= c^2 p^2 + m_0^2 c^4 \Leftrightarrow (i\hbar\partial_t)^2 = c^2 (-i\hbar\vec{\nabla})^2 + m_0^2 c^4 \\
 \Rightarrow (-\hbar^2\partial_t^2 + \hbar^2\Delta c^2)\Psi &= m_0^2 c^4 \Psi \Leftrightarrow -\hbar^2\left(\frac{1}{c^2}\partial_t^2 - \Delta\right)\Psi = m_0^2 c^2 \Psi \quad (43.5) \\
 \left(\hbar^2\left(\frac{1}{c^2}\partial_t^2 - \Delta\right) + m_0^2 c^4\right)\Psi &= 0
 \end{aligned}$$

En utilisant le d'Alembertien (cf. chapitre d'Électrodynamique), nous pouvons écrire cette dernière relation sous la forme condensée finale suivante appelée "**équation d'évolution relativiste de Schrödinger**" ou plus fréquemment "**équation de Klein-Gordon libre**" (en l'absence de champ magnétique!):

$$\boxed{(\hbar^2\Box + m_0^2 c^2)\Psi = 0} \quad (43.6)$$

avec donc la signature +, -, -, - pour la dérivée partielle:

$$\Box = \frac{\partial^2}{c^2\partial t^2} - \frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2} - \frac{\partial^2}{\partial z^2} \quad (43.7)$$

Certain auteurs préfèrent la signature -, +, +, +, dès lors, nous avons (c'est par ailleurs le choix que nous ferons dans le chapitre de Physique Quantique Des Champs puisque c'est la tradition...):

$$(\hbar^2\Box - m_0^2 c^2)\Psi = 0 \quad (43.8)$$

avec:

$$\Box = -\frac{\partial^2}{c^2\partial t^2} + \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \quad (43.9)$$

Remarque: En physique des particules élémentaires, cette équation est nommée "**équation relativiste covariante des bosons**" quelle que soit la signature choisie.

L'équation de Klein-Gordon libre est aussi souvent donnée sous la forme suivante (plus esthétique):

$$\left(\Box + \left(\frac{m_0 c}{\hbar}\right)^2\right)\Psi = 0 \quad (43.10)$$

Il est important de remarquer que l'équation de Klein-Gordon fait intervenir des scalaires et caractérise donc des particules de spin zéro.

Remarques:

R1. Nous pouvons vérifier que les ondes planes de la forme:

$$\Psi = \Psi_0 e^{\frac{i}{\hbar}(\vec{p} \cdot \vec{r} - E_{\text{tot}} t)} \quad (43.11)$$

sont des solutions de l'équation de Klein-Gordon libre (nous y reviendrons plus en détail dans le chapitre de Physique Des Particules Élémentaires).

R2. Nous reviendrons lors de notre étude de l'équation de Dirac et du spin des fermions sur l'équation de Klein-Gordon libre (afin de la généraliser).

1.1. ANTIMATIÈRE

Lors de la démonstration de l'équation de Klein-Gordon libre, nous avons laissé exprès de côté un cas très intéressant du développement que nous avons effectué.

Peut-être ne l'avez-vous pas remarqué, mais l'équation :

$$E^2 = c^2 p^2 + m_0^2 c^4 \quad (43.12)$$

peut prendre deux valeurs pour une impulsion donnée:

$$E = \pm \sqrt{c^2 p^2 + m_0^2 c^4} \quad (43.13)$$

l'une positive et l'autre négative. La valeur de l'énergie pourrait donc prendre toutes les valeurs de $] - \infty; +\infty[$

Jusqu'ici, nous avons implicitement admis en mécanique classique que les solutions négatives n'étaient pas physiques et devaient donc simplement être écartées. Cela ne peut se faire en théorie des champs quantifiés sans conduire à des incohérences graves. Plutôt que d'ignorer ces solutions d'énergie négative, il convient de leur trouver une interprétation physique.

Nous observons d'abord, que toutes les énergies négatives sont autorisées par la relation précédente (aussi bien que pour l'énergie positive). Nous disons que les états d'énergie négative sont tous occupés mais non observables; les électrons sont dits "**électrons virtuels**".

Imaginons un paquet d'onde constitué par une superposition d'ondes planes sur un intervalle étroit en impulsion. Ce paquet se déplace dans l'espace. Dans le cas unidimensionnel, il se propage à la vitesse:

$$v = \frac{dE}{dp} = \frac{c^2 p}{E} \quad (43.14)$$

Démonstration:

En nous basant toujours sur l'hypothèse que le champ de potentiel est nul, nous avons donc:

$$E = \frac{1}{2}mv^2 \Rightarrow \frac{dE}{dv} = mv \Rightarrow \frac{dE}{mdv} = v \Rightarrow \frac{dE}{dp} = v \quad (43.15)$$

et:

$$\begin{aligned} E^2 &= c^2 p^2 + m_0^2 c^4 \Leftrightarrow d(E^2) = E \cdot dE + dE \cdot E = 2E \cdot dE \\ \Rightarrow \frac{d(E^2)}{dp} &= \frac{2EdE}{dp} = \frac{d(c^2 p^2 + m_0^2 c^4)}{dp} = 2c^2 p \\ \Rightarrow \frac{2EdE}{dp} &= 2c^2 p \Leftrightarrow \frac{EdE}{dp} = c^2 p \Rightarrow \frac{dE}{dp} = \frac{c^2 p}{E} \end{aligned} \quad (43.16)$$

Donc pour résumer:

$$v = \frac{dE}{dp} = \frac{c^2 p}{E} \quad (43.17)$$

□C.Q.F.D.

Considérons d'abord une particule d'énergie positive E^+ . Sa position en fonction du temps est donnée par:

$$E^+ > 0 : x(t) = x_0 + \frac{c^2 p}{E^+} t \quad (43.18)$$

Une particule d'énergie négative E^- se déplace selon:

$$E^- < 0 : x(t) = x_0 + \frac{c^2 p}{E^-} t = x_0 + \frac{c^2 p}{|E^-|} (-t) \quad (43.19)$$

En d'autres termes, et ce sera notre première conclusion, nous pouvons dire qu'une particule d'énergie négative $E^- < 0$ est équivalente à une particule d'énergie positive $|E^-| > 0$ se déplaçant à l'envers dans le temps et ceci est ce que nous nommons une "antiparticule".

Il nous reste maintenant à voir quelle est l'interprétation à donner à une particule se déplaçant à l'envers dans le temps:

Pour simplifier, nous considérons une particule non relativiste de charge électrique $(-q)$ plongée dans un champ électrique $\vec{E}$ et magnétique $\vec{B}$ statiques. Elle satisfait à l'équation du mouvement:

$$m \frac{d^2 \vec{x}}{dt^2} = (-q) \vec{E} + (-q) \frac{d\vec{x}}{dt} \times \vec{B} \quad (43.20)$$

Nous avons étudié dans le chapitre d'Électrodynamique que les champs $\vec{E}$ et $\vec{B}$ pouvaient être construits à partir du quadripotiel $\vec{A}$. Donc nous pouvons réécrire l'équation précédente à partir des deux relations déterminées en électromagnétisme:

$$\vec{B} = \vec{\nabla} \times \vec{A} \text{ et } \vec{E} = -\vec{\nabla} \phi - \frac{\partial \vec{A}}{\partial t} \quad (43.21)$$

Cependant, il est toujours possible d'imposer la jauge suivante (nous laissons le soin au lecteur de faire la vérification en utilisant exactement la même méthodologie que celle utilisée dans le chapitre d'Électrodynamique):

$$\vec{\nabla} \phi = \vec{0} \quad (43.22)$$

L'équation du mouvement devient:

$$m \frac{d^2 \vec{x}}{dt^2} = -(-q) \frac{\partial \vec{A}}{\partial t} + (-q) \frac{d\vec{x}}{dt} \times (\vec{\nabla} \times \vec{A}) \quad (43.23)$$

ou encore:

$$m \frac{d^2 \vec{x}}{d(-t)^2} = -q \frac{\partial \vec{A}}{\partial (-t)} + q \frac{d\vec{x}}{d(-t)} \times (\vec{\nabla} \times \vec{A}) \quad (43.24)$$

Comparant les deux dernières équations, nous arrivons à notre seconde conclusion: une particule de charge q se déplaçant à l'envers dans le temps obéit aux mêmes équations du mouvement qu'une particule de charge opposée $-q$ se déplaçant vers l'avant dans le temps. L'interprétation physique de la deuxième particule est évidente.

La physique quantique relativiste implique donc l'existence d'antiparticules, qui sont effectivement observées.

Tout cela pour en arriver où exactement?

- Premièrement, la découverte théorique de l'antimatière permet d'avoir une possible explication de l'existence de l'Univers qui violait précédemment le principe de conservation de l'énergie. La théorie que nous venons de voir, prédit donc que l'Univers devrait contenir autant de matière que d'antimatière. Les scientifiques sont également à la recherche de la présence de cette antimatière.

- Deuxièmement, si nous considérons dans le vide un photon d'énergie $h\nu > 2m_e c^2$, il est capable de porter un électron virtuel vers un état d'énergie positive, où il devient réel. Il apparaît alors une lacune, ou un "trou" dans la région des énergies négatives. D'après le principe de la conservation de la charge, on voit apparaître un électron positif, ou positron, particule antimatériellement symétrique de l'électron.

Ainsi, le photon se matérialise sous la forme d'une paire e^+, e^- , avec:

$$h\nu = 2m_e c^2 + E_{cin} \quad (43.25)$$

Remarque: Certains résultats expérimentaux semblent montrer que les antiparticules ne sont pas les parfaits miroirs des particules que nous connaissons. Effectivement, la symétrie droite/gauche et temporelle ne semble pas être respectée (il y a brisure de symétrie). Nous n'avons encore rien rédigé à ce sujet sur le présent site mais nous le ferons dès que nous le pourrons.

2. ÉQUATION DE KLEIN-GORDON GÉNÉRALISÉE

L'équation de Klein-Gordon libre que nous avons initialement présentée plus haut ne prend pas en compte l'influence du champ magnétique sur l'observation du dédoublement des raies du spectre des atomes (constat expérimental). C'est pour cette raison que Klein et Gordon intégrèrent dans leur équation le champ magnétique. Cependant, ils le firent sans prendre en compte le spin de l'électron. C'est seulement après leur travail que Pauli développa son équation (dite "équation de Pauli") qui amena ensuite à l'équation de Dirac (voir plus loin).

Pour déterminer l'expression de l'équation de Klein-Gordon d'une particule chargée dans un champ magnétique et un potentiel électrostatique, utilisons la puissance du formalisme Lagrangien:

L'équation classique du mouvement admise (cf. chapitre de Mécanique Analytique), comme valable aussi en relativité, est donnée nous le savons par (équation d'Euler-Lagrange):

$$d_t \left(\partial_{\dot{q}_i} L \right) - \partial_{q_i} L \quad (43.26)$$

Dans le chapitre de Relativité Restreinte, nous avons vu que le lagrangien d'une particule libre a pour expression:

$$L_0 = -\frac{m_0 c^2}{\gamma} \quad \text{avec} \quad \gamma = (1 - \beta^2)^{-\frac{1}{2}} \quad \beta = \frac{v}{c} \quad (43.27)$$

et dans le chapitre d'Électrodynamique que le lagrangien total était:

$$L = -\frac{m_0 c^2}{\gamma} - qU + q\vec{v} \circ \vec{A} \quad (43.28)$$

Pour des besoins ultérieurs, commençons par calculer:

$$\partial_{\dot{x}} L = \partial_{\dot{x}} \left(-\frac{m_0 c^2}{\gamma} - qU + q\vec{v} \circ \vec{A} \right)_x \quad (43.29)$$

Calculons le premier terme:

$$\partial_{\dot{x}} \frac{1}{\gamma_x} = \partial_{\dot{x}} \left(1 - \frac{v_x^2}{c^2} \right)^{-\frac{1}{2}} = \frac{1}{2} \left(1 - \frac{v_x^2}{c^2} \right)^{-\frac{1}{2}} x - 2 \frac{v_x}{c^2} = -\gamma \frac{v_x}{c^2} \quad (43.30)$$

Comme le potentiel ne dépend pas de la vitesse, le terme :

$$\partial_{\dot{x}} (-qU) \quad (43.31)$$

est nul.

Le potentiel vecteur ne dépend pas de la vitesse de la particule dès lors:

$$\partial_{\dot{x}} qv_x A_x = q A_x \quad (43.32)$$

Il vient dans ce cas:

$$\partial_{\dot{x}} L = -m_0 c^2 \left(-\gamma \frac{v_x}{c^2} \right) + q A_x = \gamma m_0 v_x + q A_x = p_x + q A_x \quad (43.33)$$

L'hamiltonien classique s'écrit (cf. chapitre de Mécanique Analytique):

$$H = \sum_i p_i \dot{x}_i - L = \vec{v} \circ \partial_{\vec{v}} L - L \quad (43.34)$$

Nous avons donc démontré précédemment que:

$$\partial_{\dot{x}} L = p_x + q A_x \quad (43.35)$$

Nous pouvons donc écrire avec cette relation l'hamiltonien sous la forme:

$$\begin{aligned} H &= \vec{v} \circ (\vec{p} + q \vec{A}) + \frac{m_0 c^2}{\gamma} + q U - q \vec{v} \circ \vec{A} = \vec{v} \circ \vec{p} + q \vec{v} \circ \vec{A} + \frac{m_0 c^2}{\gamma} + q U - q \vec{v} \circ \vec{A} \\ &= \vec{v} \circ \vec{p} + \frac{m_0 c^2}{\gamma} + q U \end{aligned} \quad (43.36)$$

Le produit scalaire $\vec{v} \circ \vec{p}$ a pour expression (puisque $\vec{p}, \vec{v}$ sont colinéaires):

$$\vec{v} \circ \vec{p} = v \gamma m_0 v = \gamma m_0 v^2 \quad (43.37)$$

L'hamiltonien s'écrit alors:

$$H = \gamma m_0 v^2 + \frac{m_0 c^2}{\gamma} + q U \quad (43.38)$$

En travaillant sur les deux premiers termes:

$$\gamma m_0 v^2 + \frac{m_0 c^2}{\gamma} = \gamma m_0 c^2 \left(\frac{v^2}{c^2} + \frac{1}{\gamma^2} \right) \quad (43.39)$$

Or:

$$\gamma = \left(1 - \frac{v^2}{c^2} \right)^{-\frac{1}{2}} \Rightarrow \frac{1}{\gamma^2} = 1 - \frac{v^2}{c^2} \quad (43.40)$$

Dès lors:

$$\gamma m_0 v^2 + \frac{m_0 c^2}{\gamma} = \gamma m_0 c^2 \left(\frac{v^2}{c^2} + 1 - \frac{v^2}{c^2} \right) = \gamma m_0 c^2 \quad (43.41)$$

Finalement, nous obtenons (pour un système conservatif):

$$H = \gamma m_0 c^2 + q U = E_T \quad (43.42)$$

Toujours dans le cas d'une particule se déplaçant dans un champ électromagnétique, la relation entre l'énergie et l'impulsion (qui est différente de la quantité de mouvement par la présence d'un terme comprenant le potentiel vecteur) se calcule comme suit:

Comme (nous venons de le démontrer):

$$E = \gamma m_0 c^2 = H - qU \quad (43.43)$$

et comme nous l'avons démontré dans le chapitre de Physique Quantique Ondulatoire:

$$p^2 = (m_0 \gamma v)^2 = m_0^2 (\gamma v)^2 = m_0^2 \left(\frac{\vec{p} - q\vec{A}}{m_0} \right)^2 = (\vec{p} - q\vec{A})^2 \quad (43.44)$$

où il faut donc voir le dernier terme comme étant un abus d'écriture pour le carré de la norme.

Alors en substituant dans:

$$E^2 = p^2 c^2 + m_0^2 c^4 \quad (43.45)$$

nous obtenons (nous changeons de notation pour l'hamiltonien):

$$(E - qU)^2 - (\vec{p} - q\vec{A})^2 c^2 = m_0^2 c^4 \quad (43.46)$$

Si nous réécrivons cette relation en faisant usage des opérateurs correspondants (cf. chapitre de Physique Quantique Ondulatoire) de l'énergie et de la quantité de mouvement (quantification canonique):

$$E = i\hbar\partial_t \text{ et } \vec{p} = -i\hbar\vec{\nabla} \quad (43.47)$$

alors finalement nous pouvons écrire en analogie avec l'équation de Klein-Gordon libre (en l'absence de champ) "**l'équation de Klein-Gordon généralisée**":

$$\left[(i\hbar\partial_t - qU)^2 - (-i\hbar\vec{\nabla} - q\vec{A})^2 c^2 \right] \Psi = m_0^2 c^4 \Psi \quad (43.48)$$

Cette équation est celle de Klein-Gordon qui s'applique à une particule de charge q sans spin se déplaçant dans un champ électromagnétique.

Si $U = 0, \vec{A} = \vec{0}$ alors la relation précédente s'écrit:

$$\begin{aligned}
& \left[(i\hbar\partial_t)^2 - (-i\hbar\vec{\nabla})^2 c^2 \right] \Psi = m_0^2 c^4 \Psi \\
& \left(-\hbar^2 \partial_t^2 + \hbar^2 \Delta c^2 \right) \Psi = m_0^2 c^4 \Psi \\
& -\hbar^2 \left(\frac{1}{c^2} \partial_t^2 + \Delta \right) \Psi = m_0^2 c^2 \Psi \quad (43.49) \\
& \hbar^2 \left(\frac{1}{c^2} \partial_t^2 + \Delta \right) \Psi + m_0^2 c^2 \Psi = 0 \\
& (\hbar^2 \square + m_0^2 c^2) \Psi = 0
\end{aligned}$$

Nous retrouvons donc l'équation de Klein-Gordon d'une particule libre mais sans spin !

Il serait intéressant de regarder maintenant l'expression de l'équation de continuité (qui exprime rappelon-le: la conservation de l'énergie) avec la prise en compte du champ magnétique (parce qu'au fait elle posera toujours problème... et même un très gros). Pour cela, considérons le cas d'une particule libre se déplaçant avec une quantité de mouvement $\vec{p}$ et ayant une énergie E . Nous avons vu que nous pouvions lui associer une onde plane de la forme:

$$\psi = A e^{-i(Et - \vec{p} \cdot \vec{r})} \quad (43.50)$$

Soit l'équation Klein-Gordon libre et son expression en conjugué complexe (nous travaillons avec les unités naturelles ($c = \hbar = 1$))

$$\left. \begin{aligned}
-\partial_t^2 \Psi + \Delta \Psi - m_0^2 \Psi &= 0 & (1) \\
-\partial_t^2 \bar{\Psi} + \Delta \bar{\Psi} - m_0^2 \bar{\Psi} &= 0 & (2)
\end{aligned} \right\} \quad (43.51)$$

Nous multiplions (1) par $-i\bar{\Psi}$ et (b) par $-i\Psi$

$$\left. \begin{aligned}
-i\bar{\Psi} [-\partial_t^2 \Psi] - i\bar{\Psi} \Delta \Psi + i\bar{\Psi} m_0^2 \Psi &= 0 & (1) \\
-i\Psi [-\partial_t^2 \bar{\Psi}] - i\Psi \Delta \bar{\Psi} + i\Psi m_0^2 \bar{\Psi} &= 0 & (2)
\end{aligned} \right\} \quad (43.52)$$

Soit:

$$\left. \begin{aligned}
i\bar{\Psi} \partial_t^2 \Psi - i\bar{\Psi} \Delta \Psi + i\bar{\Psi} m_0^2 \Psi &= 0 & (1) \\
i\Psi \partial_t^2 \bar{\Psi} - i\Psi \Delta \bar{\Psi} + i\Psi m_0^2 \bar{\Psi} &= 0 & (2)
\end{aligned} \right\} \quad (43.53)$$

Par différence (1)-(2):

$$i \left[\bar{\Psi} \partial_t^2 \Psi - \Psi \partial_t^2 \bar{\Psi} \right] - i (\bar{\Psi} \Delta \Psi - \Psi \Delta \bar{\Psi}) = 0 \quad (43.54)$$

Le calcul des dérivées par rapport à t des fonctions suivantes:

$$\left. \begin{aligned}
\partial_t (\bar{\Psi} \partial_t \Psi) &= \partial_t \bar{\Psi} \partial_t \Psi + \bar{\Psi} \partial_t^2 \Psi & (1) \\
\partial_t (\Psi \partial_t \bar{\Psi}) &= \partial_t \Psi \partial_t \bar{\Psi} + \Psi \partial_t^2 \bar{\Psi} & (2)
\end{aligned} \right\} \quad (43.55)$$

Par différence (1)-(2)

$$\partial_t (\overline{\Psi} \partial_t \Psi - \Psi \partial_t \overline{\Psi}) = \boxed{\overline{\Psi} \partial_t^2 \Psi - \Psi \partial_t^2 \overline{\Psi}} \quad (43.56)$$

Ce qui nous donne finalement:

$$\partial_t \left[i (\overline{\Psi} \partial_t \Psi - \Psi \partial_t \overline{\Psi}) \right] - i \boxed{\overline{\Psi} \Delta \Psi - \Psi \Delta \overline{\Psi}} = 0 \quad (43.57)$$

Soit f un champ scalaire et $\vec{g}$ et un champ vectoriel. L'analyse vectorielle donne:

$$\vec{\nabla} \circ (f \cdot \vec{g}) = (\vec{\nabla} f) \circ \vec{g} + f (\vec{\nabla} \circ \vec{g}) \Rightarrow f (\vec{\nabla} \circ \vec{g}) = \vec{\nabla} \circ (f \cdot \vec{g}) - (\vec{\nabla} f) \circ \vec{g} \quad (43.58)$$

Posons:

$$f = \Psi, g = \vec{\nabla} \overline{\Psi} \quad (43.59)$$

Dès lors:

$$f (\vec{\nabla} \circ \vec{g}) = \vec{\nabla} \circ (f \cdot \vec{g}) - (\vec{\nabla} f) \circ \vec{g} \Rightarrow \Psi (\vec{\nabla} \circ \vec{\nabla} \overline{\Psi}) = \vec{\nabla} \circ (\Psi \vec{\nabla} \overline{\Psi}) - (\vec{\nabla} \Psi) \circ \vec{\nabla} \overline{\Psi} \quad (1) \quad (43.60)$$

Posons maintenant:

$$f = \overline{\Psi}, g = \vec{\nabla} \Psi \quad (43.61)$$

Dès lors:

$$f (\vec{\nabla} \circ \vec{g}) = \vec{\nabla} \circ (f \cdot \vec{g}) - (\vec{\nabla} f) \circ \vec{g} \Rightarrow \overline{\Psi} (\vec{\nabla} \circ \vec{\nabla} \Psi) = \vec{\nabla} \circ (\overline{\Psi} \vec{\nabla} \Psi) - (\vec{\nabla} \overline{\Psi}) \circ \vec{\nabla} \Psi \quad (2) \quad (43.62)$$

Soustrayons (1)-(2):

$$\Psi (\vec{\nabla} \circ \vec{\nabla} \overline{\Psi}) - \overline{\Psi} (\vec{\nabla} \circ \vec{\nabla} \Psi) = \vec{\nabla} \circ (\Psi \vec{\nabla} \overline{\Psi}) - \vec{\nabla} \circ (\overline{\Psi} \vec{\nabla} \Psi) \quad (43.63)$$

Comme $\vec{\nabla} \circ \vec{\nabla} = \vec{\nabla}^2 = \Delta$:

$$\Psi \Delta \overline{\Psi} - \overline{\Psi} \Delta \Psi = \vec{\nabla} \circ (\Psi \vec{\nabla} \overline{\Psi} - \overline{\Psi} \vec{\nabla} \Psi) \quad (43.64)$$

En changeant les signes:

$$\overline{\Psi} \Delta \Psi - \Psi \Delta \overline{\Psi} = \vec{\nabla} \circ (\overline{\Psi} \vec{\nabla} \Psi - \Psi \vec{\nabla} \overline{\Psi}) \quad (43.65)$$

Cette dernière relation et:

$$\partial_t \left[i (\overline{\Psi} \partial_t \Psi - \Psi \partial_t \overline{\Psi}) \right] - i (\overline{\Psi} \Delta \Psi - \Psi \Delta \overline{\Psi}) = 0 \quad (43.66)$$

donnent:

$$\partial_t \left[i (\bar{\Psi} \partial_t \Psi - \Psi \partial_t \bar{\Psi}) \right] - \vec{\nabla} \circ \left[i (\bar{\Psi} \vec{\nabla} \Psi - \Psi \vec{\nabla} \bar{\Psi}) \right] = 0 \quad (43.67)$$

À nouveau, rapprochons cette relation avec l'équation de continuité:

$$\partial_t \rho + \vec{\nabla} \circ \vec{j} = 0 \quad (43.68)$$

Rappelons que lors de notre première étude de l'équation de Klein-Gordon nous avons vu qu'en mécanique quantique son équivalent est donné par la même équation mais avec les significations suivantes: ρ est la densité de probabilité, $\vec{j}$ est la densité du flux de particules.

Nous avons donc:

$$\begin{aligned} \rho &= i (\bar{\Psi} \partial_t \Psi - \Psi \partial_t \bar{\Psi}) \\ \vec{j} &= -i (\bar{\Psi} \vec{\nabla} \Psi - \Psi \vec{\nabla} \bar{\Psi}) \end{aligned} \quad (43.69)$$

Si la fonction d'onde associée Ψ et sa conjuguée complexe $\bar{\Psi}$:

$$\Psi = A e^{-i(Et - \vec{p} \cdot \vec{r})}, \bar{\Psi} = A e^{i(Et - \vec{p} \cdot \vec{r})} \quad (43.70)$$

Les dérivées par rapport au temps de ces fonctions

$$\partial_t \Psi = -i E A e^{-i(Et - \vec{p} \cdot \vec{r})}, \partial_t \bar{\Psi} = i E A e^{i(Et - \vec{p} \cdot \vec{r})} \quad (43.71)$$

Les gradients se calculent comme suit:

$$\begin{aligned} \vec{\nabla} \Psi &= \partial_{\vec{r}} A e^{-i(Et - \vec{p} \cdot \vec{r})} = i \vec{p} A e^{-i(Et - \vec{p} \cdot \vec{r})} \\ \vec{\nabla} \bar{\Psi} &= -i \vec{p} A e^{i(Et - \vec{p} \cdot \vec{r})} \end{aligned} \quad (43.72)$$

En reprenant l'expression de la densité de probabilité et compte tenu de différentielles précédentes, il vient:

$$\begin{aligned} \rho &= i (\bar{\Psi} \partial_t \Psi - \Psi \partial_t \bar{\Psi}) \\ &= i \left(\cancel{A e^{i(Et - \vec{p} \cdot \vec{r})}} \cdot -i E \cancel{A e^{-i(Et - \vec{p} \cdot \vec{r})}} \right) - \left(\cancel{A e^{-i(Et - \vec{p} \cdot \vec{r})}} \cdot i E \cancel{A e^{i(Et - \vec{p} \cdot \vec{r})}} \right) \\ &= i (-i E A^2 - i E A^2) = i (-2 i E A^2) \end{aligned} \quad (43.73)$$

La densité de probabilité a donc pour expression:

$$\boxed{\rho = 2 E A^2} \quad (43.74)$$

En reprenant l'expression de la densité de courant et compte tenu des différentielles, il vient:

$$\begin{aligned}
\vec{J} &= -i(\bar{\Psi} \vec{\nabla} \Psi - \Psi \vec{\nabla} \bar{\Psi}) \\
&= -i \left(\left[A e^{i(Et - \vec{p} \cdot \vec{r})} \cdot i \vec{p} A e^{-i(Et - \vec{p} \cdot \vec{r})} \right] - \left[A e^{-i(Et - \vec{p} \cdot \vec{r})} \cdot -i \vec{p} A e^{i(Et - \vec{p} \cdot \vec{r})} \right] \right) \quad (43.75) \\
&= -i \left[i \vec{p} A^2 - (-i \vec{p} A^2) \right] = -i (2i \vec{p} A^2)
\end{aligned}$$

La densité de courant a pour expression:

$$\boxed{\vec{J} = 2 \vec{p} A^2} \quad (43.76)$$

En se plaçant dans la situation des connaissances de l'époque, l'équation de Klein-Gordon présente plusieurs pathologies et inconvénients.

- La densité de probabilité $\rho = 2 E A^2$ peut devenir négative (puisque comme nous l'avons vu, l'énergie peut l'être aussi), ce qui est inexplicable. Une telle situation n'existe pas avec l'équation de Schrödinger.
- L'équation de Klein-Gordon a l'inconvénient d'être du second ordre en t (l'équation de Schrödinger est elle du premier ordre). L'évolution temporelle nécessite donc la connaissance non seulement de $\Psi(t_0)$ mais également de sa dérivée $(\partial_t \Psi)_{t=t_0}$.
- Si nous appliquions cette équation à l'atome d'hydrogène, nous ne retrouverions pas les mêmes niveaux d'énergie en structure fine.

Tout ceci a conduit à l'époque qui précède les travaux de Dirac, à un rejet de cette équation qui, de plus, ne tenait pas compte du spin.

3. ÉQUATION DE DIRAC LIBRE CLASSIQUE

Jusqu'à présent, toute particule a été considérée comme ponctuelle et sans aucune structure ou degré de liberté interne. Dans cette optique, toute l'information sur l'état du système à l'instant t est alors réputée entièrement contenue dans la connaissance de la fonction d'onde $\Psi(x, y, z, t)$.

Une telle description est insuffisante, comme nous allons le voir. Cette insuffisance provient des preuves expérimentales démontrant qu'une particule telle que l'électron possède un moment magnétique propre, indépendamment de tout mouvement de rotation dans l'espace autour d'un centre. L'existence de ce moment magnétique entraîne à son tour l'existence d'un moment cinétique propre, ou intrinsèque, qui a été baptisé "spin" car on croyait au début que ce degré de liberté était lié à une rotation de la particule sur elle-même. Ce degré de liberté est "interne" - bien que l'électron continue à être considéré comme une particule ponctuelle ; c'est, au même titre que la charge ou la masse, un attribut intrinsèque, donné une fois pour toutes. Il s'avère impossible de donner du spin une image classique! Se représenter l'électron comme une petite bille de rayon non nul qui tourne sur elle-même conduit à des absurdités (par exemple, on trouve qu'un point situé à la périphérie de l'électron a une vitesse très supérieure à c). Il reste cependant que le spin d'une particule massive est son moment cinétique dans le référentiel où elle est au repos. L'hypothèse du spin de l'électron a été formulée par Uhlenbeck et Goudsmit en 1925 pour rendre compte des atomes complexes comme nous l'avons vu en physique quantique corpusculaire.

Le spin d'une particule est toujours demi-entier ou entier, c'est un fait d'expérience. Le caractère entier ou demi-entier du spin définit deux grandes familles de particules, les bosons (spin entier) et les fermions (spin demi-entier), obéissant à des statistiques très différentes telles que celles que nous avons présentées dans le chapitre de Mécanique Statistique (d'où l'existence d'une relation appelée "théorème spin-statistique").

Revenons au cas de l'électron. Les deux valeurs possibles révélées par une mesure de S (le μ_z que nous avons dans le chapitre de Physique Quantique Corpusculaire) sont donc les valeurs propres $\pm \hbar/2$ (cf. chapitre de Physique Quantique Ondulatoire) associées aux deux valeurs possibles d'un nombre quantique $m_s = \pm 1/2$ lui-même associé donc à l'état libre ($\vec{L} = \vec{0}$) au moment cinétique:

$$b \leq \sqrt{J(J+1)}\hbar = \sqrt{(L+S)((L+S)+1)}\hbar = \sqrt{S(S+1)}\hbar \quad (43.77)$$

Donc:

$$b \leq \sqrt{S(S+1)}\hbar = \sqrt{\pm \frac{1}{2}(\pm \frac{1}{2} + 1)}\hbar = \sqrt{\frac{3}{4}}\hbar = \sqrt{3}\hbar/2 \quad (43.78)$$

Une description complète de l'état de l'électron contient donc nécessairement une fonction d'onde donnant comme d'habitude la densité de probabilité de présence, mais prenant également en compte le degré de liberté du spin, d'où la notation $\Psi(x, y, z, m_s, t)$. Si les coordonnées d'espace prennent des valeurs réelles continues, en revanche la variable de spin est donc essentiellement discrète.

En maintenant l'interprétation usuelle, la quantité $|\Psi(x, y, z, m_s, t)|^2 d^3r$ est la probabilité de présence autour du point choisi avec la valeur $m_s \hbar$ pour le spin. La condition de normalisation des probabilités introduit comme toujours une sommation, qui porte non seulement sur les degrés orbitaux (sommation continue, c'est-à-dire intégration) mais également sur les degrés de spin (sommation discrète):

$$\int \left| \Psi \left(\vec{r}, +\frac{1}{2} \right) \right|^2 d^3r + \int \left| \Psi \left(\vec{r}, -\frac{1}{2} \right) \right|^2 d^3r = 1 \quad (43.79)$$

exprimant notamment le fait que nous épuisons toutes les possibilités du spin en sommant sur les deux valeurs possibles. En tout état de cause, l'électron n'a plus une mais deux fonctions d'onde, une pour chaque valeur de m_s .

La notation précédente n'est pas forcément la meilleure pour les particules libres de spin supérieur à 1/2 comme nous l'avons vu lors de notre étude du moment cinétique. S'agissant d'une variable prenant des valeurs discrètes, il est tout aussi légitime de mettre m_s en indice de Ψ et de poser $\Psi_{m_s}(\vec{r}, t)$. Enfin, il est commode d'utiliser une notation matricielle, rangeant en colonne les différentes fonctions correspondant aux valeurs possibles de la variable discrète m_s . Ainsi, pour l'électron, nous admettrons désormais que toute l'information au sens de la physique quantique ondulatoire est contenue dans un vecteur-colonne à deux lignes appelé "**spineur**" (cf. chapitre de Calcul Spinoriel) et noté:

$$\begin{bmatrix} \Psi_{+\frac{1}{2}}(\vec{r}, t) \\ \Psi_{-\frac{1}{2}}(\vec{r}, t) \end{bmatrix} \text{ ou } \begin{bmatrix} \Psi_+(\vec{r}, t) \\ \Psi_-(\vec{r}, t) \end{bmatrix} \quad (43.80)$$

Revenons maintenant sur l'équation de Klein-Gordon libre (plus générale que l'équation de Schrödinger bien évidemment mais moins que celle comportant le champ magnétique):

$$\hbar^2 \frac{\partial^2 \Psi}{\partial t^2} = \hbar^2 c^2 \Delta \Psi - m_0^2 c^4 \Psi \quad (43.81)$$

Cette équation est comme nous le savons malheureusement incomplète car elle ne contient aucune information sur le spin de l'électron.

Nous pouvons cependant, pour tenter de trouver une solution à ce problème, faire un parallèle avec le champ électromagnétique. Celui-ci comporte aussi un spin, résidant dans la polarisation du champ (cf. chapitre d'Électrodynamique). Cette polarisation est étroitement liée à la nature vectorielle du champ électromagnétique et transparaît dans les équations de Maxwell, qui sont du premier ordre en dérivées. Cependant, en combinant les équations de Maxwell, nous avons vu dans le chapitre d'Électrodynamique que nous pouvions obtenir les équations d'onde:

$$\nabla^2 \vec{E} - \frac{1}{c^2} \frac{\partial^2 \vec{E}}{\partial t^2} = \Delta \vec{E} - \frac{1}{c^2} \frac{\partial^2 \vec{E}}{\partial t^2} = \vec{0} \text{ et } \nabla^2 \vec{B} - \frac{1}{c^2} \frac{\partial^2 \vec{B}}{\partial t^2} = \Delta \vec{B} - \frac{1}{c^2} \frac{\partial^2 \vec{B}}{\partial t^2} = \vec{0} \quad (43.82)$$

qui sont (coïncidence très pertinente!) un cas particulier de l'équation de Klein-Gordon quand $m_0 = 0$:

$$\Delta \Psi - \frac{1}{c^2} \frac{\partial^2 \Psi}{\partial t^2} = 0 \quad (43.83)$$

Les équations d'onde recèlent cependant moins d'informations que les équations de Maxwell originales: elles ne contiennent explicitement aucune relation entre les différentes composantes des champs $\vec{E}$ et $\vec{B}$, comme par exemple le fait que, dans une onde électromagnétique de vecteur d'onde donné, les champs $\vec{E}$ et $\vec{B}$ sont mutuellement perpendiculaires et tous les deux perpendiculaires au vecteur d'onde. Pour établir ces contraintes, il faut retourner aux équations de Maxwell et donc à des équations avec des dérivées du premier ordre.

Il en est de même pour les fermions (les électrons en font partie). L'équation de Klein-Gordon, quoiqu'elle ne soit pas fautive, est incomplète. Il faut tenter ici d'établir une équation du premier ordre en dérivées qui décrive bien le spin 1/2 des électrons des fermions. Cette dernière condition signifie que cette équation doit donc faire intervenir les deux composantes d'un spineur (en analogie avec celui que nous nous sommes déterminé plus haut):

$$\chi = \begin{bmatrix} \chi_1 \\ \chi_2 \end{bmatrix} \quad (43.84)$$

Nous écrirons alors cette équation que nous cherchons comme:

$$D \chi = 0 \quad (43.85)$$

où D est une matrice 2×2 faisant intervenir des dérivées du premier ordre (un opérateur différentiel du premier ordre).

Pour donner un exemple avant d'aller plus loin, regardons comment l'équation de Klein-Gordon peut s'exprimer sous une telle forme.

Nous avons donc (équation de Klein-Gordon libre):

$$\hbar^2 \frac{\partial^2 \Psi}{\partial t^2} = \hbar^2 c^2 \Delta \Psi - m_0^2 c^4 \Psi \quad (43.86)$$

ou (équation de Klein-Gordon généralisée):

$$\left[(i\hbar \partial_t - qU)^2 - (-i\hbar \vec{\nabla} - q\vec{A})^2 c^2 \right] \Psi = m_0^2 c^4 \Psi \quad (43.87)$$

Ce qui s'écrit aussi pour l'équation de Klein-Gordon libre:

$$\left(1 + \frac{i\hbar}{m_0 c^2} \frac{\partial}{\partial t} \right) \left(1 - \frac{i\hbar}{m_0 c^2} \frac{\partial}{\partial t} \right) \Psi = \frac{\hbar^2}{m_0^2 c^2} \Delta \Psi \quad (43.88)$$

ou pour l'équation de Klein-Gordon généralisée:

$$\begin{aligned} & \left[-\hbar^2 \partial_t^2 - 2i\hbar \partial_t qU + q^2 U^2 + \hbar^2 \vec{\nabla}^2 c^2 - 2i\hbar \vec{\nabla} q \vec{A} c^2 - q^2 \vec{A}^2 c^2 \right] \Psi = m_0^2 c^4 \Psi \\ & \left[1 + \frac{\hbar^2 \partial_t^2}{m_0^2 c^4} \right] \Psi = \frac{(-q^2 \vec{A}^2 c^2 - 2i\hbar \vec{\nabla} q \vec{A} c^2 + \hbar^2 \Delta c^2 + q^2 U^2 - 2i\hbar \partial_t qU)}{m_0^2 c^4} \Psi \\ & \left(1 + \frac{i\hbar}{m_0 c^2} \frac{\partial}{\partial t} \right) \left(1 - \frac{i\hbar}{m_0 c^2} \frac{\partial}{\partial t} \right) \Psi = \frac{(-q^2 \vec{A}^2 c^2 - 2i\hbar \vec{\nabla} q \vec{A} c^2 + \hbar^2 \Delta c^2 + q^2 U^2 - 2i\hbar \partial_t qU)}{m_0^2 c^4} \Psi \end{aligned} \quad (43.89)$$

Restreignons-nous maintenant au cas de l'équation de Klein-Gordon libre (le raisonnement étant similaire mais... plus long pour la version généralisée).

La dernière expression de l'équation de Klein-Gordon libre suggère d'introduire les deux combinaisons trouvées après (paraît-il) de nombreux essais par nos prédécesseurs:

$$\begin{aligned} \Psi_a &= \Psi + \frac{i\hbar}{m_0 c^2} \frac{\partial \Psi}{\partial t} = \left(1 + \frac{i\hbar}{m_0 c^2} \frac{\partial}{\partial t} \right) \Psi \\ \Psi_b &= \Psi - \frac{i\hbar}{m_0 c^2} \frac{\partial \Psi}{\partial t} = \left(1 - \frac{i\hbar}{m_0 c^2} \frac{\partial}{\partial t} \right) \Psi \end{aligned} \quad (43.90)$$

d'où résulte:

$$\begin{aligned} \Psi &= \frac{1}{2} (\Psi_a + \Psi_b) \\ \frac{\partial \Psi}{\partial t} &= \frac{m_0 c^2}{2i\hbar} (\Psi_a - \Psi_b) \end{aligned} \quad (43.91)$$

Dès lors:

$$\left(1 + \frac{i\hbar}{m_0 c^2} \frac{\partial}{\partial t} \right) \left(1 - \frac{i\hbar}{m_0 c^2} \frac{\partial}{\partial t} \right) \Psi = \frac{\hbar^2}{m_0^2 c^2} \Delta \Psi \quad (43.92)$$

peut s'écrire de deux façons après substitutions respectives:

$$\begin{aligned} \left(1 - \frac{i\hbar}{m_0 c^2} \frac{\partial}{\partial t}\right) \Psi_a &= \frac{\hbar^2}{2m_0^2 c^2} \Delta (\Psi_a + \Psi_b) \\ \left(1 + \frac{i\hbar}{m_0 c^2} \frac{\partial}{\partial t}\right) \Psi_b &= \frac{\hbar^2}{2m_0^2 c^2} \Delta (\Psi_a + \Psi_b) \end{aligned} \quad (43.93)$$

Soit, sous forme matricielle:

$$i\hbar \frac{\partial}{\partial t} \begin{bmatrix} \Psi_a \\ \Psi_b \end{bmatrix} = \begin{bmatrix} m_0 c^2 - \frac{\hbar^2}{2m_0} \Delta & -\frac{\hbar^2}{2m_0} \Delta \\ \frac{\hbar^2}{2m_0} \Delta & -m_0 c^2 + \frac{\hbar^2}{2m_0} \Delta \end{bmatrix} \begin{bmatrix} \Psi_a \\ \Psi_b \end{bmatrix} \quad (43.94)$$

ou encore:

$$\left(i\hbar \frac{\partial}{\partial t} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} m_0 c^2 - \frac{\hbar^2}{2m_0} \Delta & -\frac{\hbar^2}{2m_0} \Delta \\ \frac{\hbar^2}{2m_0} \Delta & -m_0 c^2 + \frac{\hbar^2}{2m_0} \Delta \end{bmatrix} \right) \begin{bmatrix} \Psi_a \\ \Psi_b \end{bmatrix} = \vec{0} \quad (43.95)$$

Ce que nous pouvons écrire:

$$\begin{bmatrix} i\hbar \frac{\partial}{\partial t} - m_0 c^2 + \frac{\hbar^2}{2m_0} \Delta & \frac{\hbar^2}{2m_0} \Delta \\ \frac{\hbar^2}{2m_0} \Delta & -i\hbar \frac{\partial}{\partial t} - m_0 c^2 + \frac{\hbar^2}{2m_0} \Delta \end{bmatrix} \begin{bmatrix} \Psi_a \\ \Psi_b \end{bmatrix} = \vec{0} \quad (43.96)$$

Donc par rapport à notre idée initiale d'avoir une relation sous la forme:

$$D\chi = 0 \quad (43.97)$$

nous pouvons faire la similitude avec l'équation antérieure:

$$\chi \equiv \begin{bmatrix} \Psi_a \\ \Psi_b \end{bmatrix} \text{ et } D = \begin{bmatrix} i\hbar \frac{\partial}{\partial t} - m_0 c^2 + \frac{\hbar^2}{2m_0} \Delta & \frac{\hbar^2}{2m_0} \Delta \\ \frac{\hbar^2}{2m_0} \Delta & -i\hbar \frac{\partial}{\partial t} - m_0 c^2 + \frac{\hbar^2}{2m_0} \Delta \end{bmatrix} \quad (43.98)$$

où D est bien une matrice 2×2 .

Mais nous, nous recherchons toujours (pour faire le parallèle avec les équations de Maxwell) un système d'équations avec des différentielles du premier ordre. Dans l'objectif de chercher une forme plus générale incluant sous forme naturelle le spin, nous allons poser en analogie avec le résultat ci-dessus:

$$D = A + \vec{B} \circ \vec{\sigma} \quad (43.99)$$

où A est un vecteur (mais représenté par tradition comme un scalaire), $\vec{B}$ est un vecteur (et noté tel quel) et $\vec{\sigma}$ un vecteur composé de matrices symétriques 2×2 (en lisant la suite vous verrez que poser cela permet de trouver ce que nous cherchons...).

Rappelons que la multiplication entre $\vec{B}$ et $\vec{\sigma}$ constitue un produit scalaire tel que celui défini dans notre étude du chapitre de Calcul Spinoriel.

Remarque: Il faut être très prudent dans les développements qui vont suivre, car les notations traditionnelles dans le domaine rendent très difficiles les distinctions entre produits, produits scalaires, et produits de composantes de vecteurs formant un vecteur.

Posons (au fait nos prédécesseurs ont aussi fait de nombreux essais avant de poser cela...):

$$D = \frac{\partial}{\partial t} + \lambda \vec{\sigma} \circ \vec{\nabla} \quad (43.100)$$

Ce que nous cherchions initialement c'est-à-dire $D\chi = 0$ devient alors:

$$\frac{\partial}{\partial t} \chi + \lambda \vec{\sigma} \circ \vec{\nabla} \chi = \vec{0} \quad (43.101)$$

Ainsi, $A = \partial / \partial t$, $\vec{B} = \lambda \vec{\nabla}$ et $\vec{\sigma}$ reste (imaginons...) inconnu. Il nous faut également déterminer λ .

Toujours par analogie avec l'exemple fait plus haut, tentons de retrouver l'équation d'onde pour déterminer la constante λ :

$$\left(\frac{\partial}{\partial t} - \lambda \vec{\sigma} \circ \vec{\nabla} \right) \left(\frac{\partial}{\partial t} + \lambda \vec{\sigma} \circ \vec{\nabla} \right) \chi = \vec{0} \Rightarrow \left(\frac{\partial^2}{\partial t^2} - \lambda^2 (\vec{\sigma} \circ \vec{\nabla})(\vec{\sigma} \circ \vec{\nabla}) \right) \chi = \vec{0} \quad (43.102)$$

Pour que nous retrouvions l'équation d'onde, il faut donc que:

$$1. (\vec{\sigma} \circ \vec{\nabla})(\vec{\sigma} \circ \vec{\nabla}) = \vec{\nabla}^2$$

Effectivement (attention il faut que le lecteur ait étudié idéalement le chapitre de Calcul Spinoriel pour savoir en détails comment nous arrivons à ce développement!):

$$\begin{aligned} (\vec{\sigma} \circ \vec{\nabla})(\vec{\sigma} \circ \vec{\nabla}) &= \begin{bmatrix} \partial_x & \partial_x - i\partial_y \\ \partial_x + i\partial_y & -\partial_x \end{bmatrix} \begin{bmatrix} \partial_x & \partial_x - i\partial_y \\ \partial_x + i\partial_y & -\partial_x \end{bmatrix} \\ &= \begin{bmatrix} \partial_x^2 + (\partial_x + i\partial_y)(\partial_x - i\partial_y) & \partial_x(\partial_x - i\partial_y) - \partial_x(\partial_x - i\partial_y) \\ \partial_x(\partial_x + i\partial_y) - \partial_x(\partial_x + i\partial_y) & (\partial_x - i\partial_y)(\partial_x + i\partial_y) + \partial_x^2 \end{bmatrix} \\ &= \begin{bmatrix} \partial_x^2 + \partial_y^2 + \partial_z^2 & 0 \\ 0 & \partial_x^2 + \partial_y^2 + \partial_z^2 \end{bmatrix} = \mathbb{1} \vec{\nabla}^2 \equiv \vec{\nabla}^2 \end{aligned} \quad (43.103)$$

$$2. \lambda = c:$$

$$\left(\frac{\partial^2}{\partial t^2} - \lambda^2 \vec{\nabla}^2 \right) \chi = \vec{0} \quad (43.104)$$

Il y a donc deux possibilités qui peuvent s'appliquer à des champs différents que nous noterons $\chi^{(\pm)}$. Nous avons donc une sorte de double spineur tel que:

$$\left(\frac{\partial}{\partial t} \pm c \vec{\sigma} \circ \vec{\nabla} \right) \chi^{(\pm)} = \vec{0} \quad (43.105)$$

Ces équations sont appelées "**équations de Weyl**".

Il nous faut maintenant généraliser les équations de Weyl au cas d'un fermion de spin demi-entier avec masse. Cette nouvelle équation doit respecter les contraintes suivantes:

- C1. Elle doit se réduire aux équations de Weyl quand la masse tend vers zéro.
- C2. Elle doit mener à l'équation de Klein-Gordon libre.
- C3. Elle doit décrire des particules possédant un spin.

La solution consiste alors à coupler les deux équations de Weyl par un terme proportionnel à la masse:

$$\begin{aligned} i\hbar \left(\frac{\partial}{\partial t} + c \vec{\sigma} \circ \vec{\nabla} \right) \chi^{(+)} &= m_0 c^2 \chi^{(-)} \\ i\hbar \left(\frac{\partial}{\partial t} - c \vec{\sigma} \circ \vec{\nabla} \right) \chi^{(-)} &= m_0 c^2 \chi^{(+)} \end{aligned} \quad (43.106)$$

Pour vérifier que les facteurs ont été correctement choisis, nous appliquons $\hbar \left(\frac{\partial}{\partial t} + c \vec{\sigma} \circ \vec{\nabla} \right)$ sur la première équation et nous y substituons la deuxième. Nous trouvons:

$$i\hbar \left(\frac{\partial}{\partial t} - c \vec{\sigma} \circ \vec{\nabla} \right) \hbar \left(\frac{\partial}{\partial t} + c \vec{\sigma} \circ \vec{\nabla} \right) \chi^{(+)} = m_0 c^2 \hbar \left(\frac{\partial}{\partial t} - c \vec{\sigma} \circ \vec{\nabla} \right) \chi^{(-)} = -im_0^2 c^4 \chi^{(+)} \quad (43.107)$$

ou encore:

$$\hbar^2 \left(\frac{\partial^2}{\partial t^2} - c^2 \vec{\nabla}^2 \right) \chi^{(+)} = -m_0^2 c^4 \chi^{(+)} \quad (43.108)$$

à comparer avec:

$$\hbar^2 \left(\frac{\partial^2}{\partial t^2} - c^2 \vec{\nabla}^2 \right) \Psi = -m_0^2 c^4 \Psi \quad (43.109)$$

Ce qui est bel et bien l'équation de Klein-Gordon libre (nous démontrons la même correspondance pour la composante $\chi^{(-)}$) et renforce donc la validité des hypothèses et développements faits jusqu'à maintenant.

Il est usuel de rassembler les deux spineurs dans un seul spineur (cela devient alors un "**bispineur**") de quatre composantes (un spineur à quatre composantes dont deux sont en fait associées aux particules et deux aux antiparticules comme nous le verrons):

$$\Psi = \begin{bmatrix} \chi^{(+)} \\ \chi^{(-)} \end{bmatrix} = \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \Psi_4 \end{bmatrix} \quad (43.110)$$

et de définir les deux matrices 4×4 suivantes (sous une forme dite "**forme chirale**"):

$$\gamma^0 = \begin{bmatrix} 0 & \mathbb{1} \\ \mathbb{1} & 0 \end{bmatrix} \quad (43.111)$$

où $\mathbb{1}$ est la matrice unité 2×2 traditionnelle définie par:

$$\mathbb{1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (43.112)$$

et:

$$\gamma^i = \begin{bmatrix} 0 & -\sigma_i \\ \sigma_i & 0 \end{bmatrix} = \alpha \cdot \beta = \begin{bmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (43.113)$$

où les σ_i sont les "**matrices de Pauli**" données par (cf. chapitre de Calcul Spinoriel):

$$\sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \sigma_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (43.114)$$

qui doivent satisfaire rappelons-le (démontré plus haut):

$$(\vec{\sigma} \circ \vec{\nabla})(\vec{\sigma} \circ \vec{\nabla}) = \vec{\nabla}^2 \quad (43.115)$$

qui est pour rappel une matrice 2×2 . Les matrices de Pauli sont donc de bonnes candidates pour résoudre notre problème!

Remarques:

R1. Comme nous l'avons vu dans le chapitre de Calcul Spinoriel (section d'Algèbre), σ_0 n'est pas vraiment une matrice de Pauli en soi. Cependant, dans certains ouvrages, elle est indiquée comme en étant une (c'est aussi notre choix ici).

R2. Comme nous l'avons également vu dans le chapitre de Calcul Spinoriel, rappelons que les matrices de Pauli représentent implicitement des rotations spatiales infinitésimales d'un spineur.

R3. Attention dans les matrices de Pauli, l'écriture du 0 signifie qu'il s'agit des matrices 2×2 dont toutes les composantes sont nulles.

Ceci nous permet, enfin, de combiner les équations:

$$\begin{aligned} i\hbar \left(\frac{\partial}{\partial t} + c \vec{\sigma} \cdot \vec{\nabla} \right) \chi^{(+)} &= m_0 c^2 \chi^{(-)} \\ i\hbar \left(\frac{\partial}{\partial t} - c \vec{\sigma} \cdot \vec{\nabla} \right) \chi^{(-)} &= m_0 c^2 \chi^{(+)} \end{aligned} \quad (43.116)$$

en une seule (ne pas oublier l'association des opérateurs $E \rightarrow i\hbar \partial / \partial t, p \rightarrow -i\hbar \vec{\nabla}$):

$$\left(\gamma_0 i \frac{\partial}{\partial t} + \gamma_i \frac{\partial}{\partial x_i} - \frac{m_0 c}{\hbar} \right) \Psi = \vec{0} \quad (43.117)$$

en utilisant la notation d'usage en calcul tensoriel et en choisissant les unités naturelles $c = 1, \hbar = 1$ nous avons:

$$\boxed{(i\gamma^\mu \partial_\mu - m_0) \Psi = \vec{0}} \quad (43.118)$$

ce qui constitue la forme habituelle de "**l'équation de Dirac**" (implicitement un système de quatre équations différentielles couplées) ou "**équation relativiste de l'électron**" avec la "**dérivée covariante**":

$$\partial_\mu = \left(\frac{\partial}{\partial t}, \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) \quad (43.119)$$

et où le vecteur nul est donc (pour rappel...) un vecteur à quatre composantes toutes nulles!

Remarque: En physique des particules élémentaires, la relation antérieure est appelée "**équation relativiste covariante des fermions**" car elle décrit les particules de spin 1/2.

Les matrices γ^μ sont appelées "**matrices de Dirac**". Sous forme encore plus condensée (en utilisant le "**slash de Feynman**") l'équation de Dirac s'écrit parfois:

$$\boxed{(i \not{\partial} - m_0) \Psi = \vec{0}} \quad (43.120)$$

Nous avons ainsi, comme en analogie avec les équations de Maxwell, des équations différentielles du premier ordre qui ont comme propriété:

P1. De permettre de retomber sur l'équation de Klein-Gordon, in extenso sur l'équation d'onde (comme pour les équations de Maxwell)

P2. De prendre en compte (décrire) explicitement le caractère spinoriel des fonctions d'onde comme nous allons le voir en nous penchant de plus près sur les matrices de Pauli.

Remarque: Comme l'équation de Dirac s'applique aux particules de spin 1/2 elle s'applique aussi aux neutrinos dont la masse au repos est nulle (donc la résolution de l'équation de Dirac se simplifie largement).

Dans le but maintenant d'interpréter le contenu physique de l'équation de Dirac, nous allons utiliser une représentation différente des matrices de Pauli. Nous avons vu que la représentation par les matrices 4×4 :

$$\gamma^0 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \gamma^i = \begin{bmatrix} 0 & -\sigma_i \\ \sigma_i & 0 \end{bmatrix} \quad (43.121)$$

était dite "**représentation Chirale**" alors que nous allons utiliser maintenant la "**représentation de Dirac**" définie par:

$$\gamma^0 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \gamma^i = \begin{bmatrix} 0 & \sigma_i \\ -\sigma_i & 0 \end{bmatrix} \quad (43.122)$$

Nous vérifions facilement (algèbre linéaire élémentaire) que cette représentation s'obtient par la transformation:

$$\gamma^\mu \rightarrow U \gamma^\mu U^\dagger \text{ où } U = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad (43.123)$$

Rappelons que $U^\dagger$ est la matrice adjointe (la conjuguée de la matrice transposée) de U . Or, lorsque tous les éléments sont des réels comme c'est le cas ci-dessus et que la matrice est carrée alors (cf. chapitre d'Algèbre Linéaire) nous savons que $U = U^\dagger$.

Démonstration:

$$\begin{aligned} \gamma^0 &= U \gamma^0 U^\dagger = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1+1 & -1+1 \\ 1-1 & -1+-1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \end{aligned} \quad (43.124)$$

et:

$$\begin{aligned}
\gamma^j &= U \gamma^j U^\dagger = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 0 & -\sigma_i \\ \sigma_i & 0 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 0 & -\sigma_i \\ \sigma_i & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \\
&= \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} -\sigma_i & \sigma_i \\ \sigma_i & \sigma_i \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 0 & 2\sigma_i \\ -2\sigma_i & 0 \end{bmatrix} = \begin{bmatrix} 0 & \sigma_i \\ -\sigma_i & 0 \end{bmatrix}
\end{aligned} \quad (43.125)$$

Cherchons maintenant les solutions particulières à l'équation de Dirac sous la forme:

$$\Psi = \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} \quad (43.126)$$

En substituant dans l'équation de Dirac et après simplification par $e^{i(\vec{p} \circ \vec{r} - Et)}$ nous trouvons facilement:

$$\left(i\gamma^0(-iE) + i \sum_j \gamma^j (ip_j) - m_0 \right) \Psi_{\vec{p}} = \vec{0} \quad (43.127)$$

Effectivement en unités naturelles:

$$\begin{aligned}
&(i\gamma^\mu \partial_\mu - m_0) \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} = i\gamma^\mu \partial_\mu \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} - m_0 \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} \\
&= i\gamma^0 \partial_0 \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} \\
&+ i\gamma^1 \partial_1 \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} + i\gamma^2 \partial_2 \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} + i\gamma^3 \partial_3 \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} - m_0 \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} \\
&= i\gamma^0 \frac{\partial}{\partial t} \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} \\
&+ i\gamma^1 \frac{\partial}{\partial x} \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} + i\gamma^2 \frac{\partial}{\partial y} \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} + i\gamma^3 \frac{\partial}{\partial z} \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} - m_0 \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} \\
&= (-iE)i\gamma^0 \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} + i\gamma^0 e^{i(\vec{p} \circ \vec{r} - Et)} \frac{\partial}{\partial t} \Psi_{\vec{p}} \\
&+ i\gamma^1 \frac{\partial}{\partial x} \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} + i\gamma^2 \frac{\partial}{\partial y} \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} + i\gamma^3 \frac{\partial}{\partial z} \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} - m_0 \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} \\
&= (-iE)i\gamma^0 \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} \\
&+ i\gamma^1 ip_x \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} + i\gamma^2 ip_y \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} + i\gamma^3 ip_z \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} - m_0 \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} \\
&+ i\gamma^1 e^{i(\vec{p} \circ \vec{r} - Et)} \frac{\partial}{\partial x} \Psi_{\vec{p}} + i\gamma^2 e^{i(\vec{p} \circ \vec{r} - Et)} \frac{\partial}{\partial y} \Psi_{\vec{p}} + i\gamma^3 e^{i(\vec{p} \circ \vec{r} - Et)} \frac{\partial}{\partial z} \Psi_{\vec{p}} \\
&= (-iE)i\gamma^0 \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} \\
&+ i\gamma^1 ip_x \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} + i\gamma^2 ip_y \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} + i\gamma^3 ip_z \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)} - m_0 \Psi_{\vec{p}} e^{i(\vec{p} \circ \vec{r} - Et)}
\end{aligned} \quad (43.128)$$

Avec la représentation de Dirac, nous obtenons après développement (calcul trivial):

$$\begin{pmatrix} E - m_0 & -\vec{p} \circ \vec{\sigma} \\ \vec{p} \circ \vec{\sigma} & -E - m_0 \end{pmatrix} \Psi_{\vec{p}} = \vec{0} \quad (43.129)$$

ou plus explicitement:

$$\begin{pmatrix} (E - m_0) \mathbb{1} & -\vec{p} \circ \vec{\sigma} \\ \vec{p} \circ \vec{\sigma} & (-E - m_0) \mathbb{1} \end{pmatrix} \Psi_{\vec{p}} = \vec{0} \quad (43.130)$$

Effectivement avec la convention de sommation d'Einstein:

$$\begin{aligned} & \left(i\gamma^0(-iE) + i\sum_j \gamma^j(ip_j) - m_0 \right) \Psi_{\vec{p}} = (\gamma^0 E - \gamma^j p_j - m_0) \Psi_{\vec{p}} \\ & = \left(\begin{bmatrix} \mathbb{1} & 0 \\ 0 & -\mathbb{1} \end{bmatrix} E - \begin{bmatrix} 0 & \sigma_1 \\ -\sigma_1 & 0 \end{bmatrix} p_1 - \begin{bmatrix} 0 & \sigma_2 \\ -\sigma_2 & 0 \end{bmatrix} p_2 - \begin{bmatrix} 0 & \sigma_3 \\ -\sigma_3 & 0 \end{bmatrix} p_3 - m_0 \right) \Psi_{\vec{p}} \\ & = \left(\begin{bmatrix} E & 0 \\ 0 & -E \end{bmatrix} - \begin{bmatrix} 0 & \sigma_1 p_1 \\ -\sigma_1 p_1 & 0 \end{bmatrix} - \begin{bmatrix} 0 & \sigma_2 p_2 \\ -\sigma_2 p_2 & 0 \end{bmatrix} - \begin{bmatrix} 0 & \sigma_3 p_3 \\ -\sigma_3 p_3 & 0 \end{bmatrix} - m_0 \right) \Psi_{\vec{p}} \\ & = \left(\begin{bmatrix} E & 0 \\ 0 & -E \end{bmatrix} - \begin{bmatrix} 0 & \sigma_1 p_1 + \sigma_2 p_2 + \sigma_3 p_3 \\ -\sigma_1 p_1 - \sigma_2 p_2 - \sigma_3 p_3 & 0 \end{bmatrix} - m_0 \right) \Psi_{\vec{p}} \\ & = \left(\begin{bmatrix} E & 0 \\ 0 & -E \end{bmatrix} - \begin{bmatrix} 0 & \vec{p} \circ \vec{\sigma} \\ -\vec{p} \circ \vec{\sigma} & 0 \end{bmatrix} - m_0 \right) \Psi_{\vec{p}} = \left(\begin{bmatrix} E & 0 \\ 0 & -E \end{bmatrix} + \begin{bmatrix} 0 & -\vec{p} \circ \vec{\sigma} \\ \vec{p} \circ \vec{\sigma} & 0 \end{bmatrix} - m_0 \right) \Psi_{\vec{p}} \\ & = \left(\begin{bmatrix} E & -\vec{p} \circ \vec{\sigma} \\ \vec{p} \circ \vec{\sigma} & -E \end{bmatrix} - m_0 \right) \Psi_{\vec{p}} = \left(\begin{bmatrix} E & -\vec{p} \circ \vec{\sigma} \\ \vec{p} \circ \vec{\sigma} & -E \end{bmatrix} \Psi_{\vec{p}} - m_0 \Psi_{\vec{p}} \right) \\ & = \begin{pmatrix} E\phi_a - \vec{p} \circ \vec{\sigma} \phi_b - m_0 \phi_a \\ \vec{p} \circ \vec{\sigma} \phi_a - E\phi_b - m_0 \phi_b \end{pmatrix} = \begin{pmatrix} \phi_a(E - m_0) - \vec{p} \circ \vec{\sigma} \phi_b \\ \vec{p} \circ \vec{\sigma} \phi_a + \phi_b(-E - m_0) \end{pmatrix} = \begin{pmatrix} E - m_0 & -\vec{p} \circ \vec{\sigma} \\ \vec{p} \circ \vec{\sigma} & -E - m_0 \end{pmatrix} \begin{pmatrix} \phi_a \\ \phi_b \end{pmatrix} \\ & = \begin{pmatrix} E - m_0 & -\vec{p} \circ \vec{\sigma} \\ \vec{p} \circ \vec{\sigma} & -E - m_0 \end{pmatrix} \Psi_{\vec{p}} = \vec{0} \end{aligned} \quad (43.131)$$

Pour que cette équation matricielle ait des solutions non nulles, il faut comme d'habitude que le déterminant de la matrice soit nul (cf. chapitre d'Algèbre Linéaire). Nous vérifions facilement que:

$$\begin{aligned} \det \begin{pmatrix} E - m_0 & -\vec{p} \circ \vec{\sigma} \\ \vec{p} \circ \vec{\sigma} & -E - m_0 \end{pmatrix} &= (E - m_0)(-E - m_0) - (\vec{p} \circ \vec{\sigma})(-\vec{p} \circ \vec{\sigma}) \\ &= -(E - m_0)(E + m_0) + \|\vec{p}\|^2 = -(E^2 - m_0^2) + \|\vec{p}\|^2 = -E^2 + m_0^2 + \|\vec{p}\|^2 = 0 \end{aligned} \quad (43.132)$$

Ce qui implique (ne pas oublier que nous sommes en unités naturelles!):

$$\boxed{E = \pm \sqrt{\|\vec{p}\|^2 + m_0^2}} \quad (43.133)$$

Il s'agit donc d'un système d'équation follement amusant à résoudre (ne pas oublier que chaque terme de la relation ci-dessus en réalité une matrice 2×2).

Avec la représentation Chirale, nous aurions obtenu (toujours en adoptant la notation traditionnelle du site pour le produit scalaire):

$$\begin{aligned}
& \left(i\gamma^0(-iE) + i\sum_j \gamma^j(ip_j) - m_0 \right) \Psi_{\vec{p}} = \left(\gamma^0 E - \gamma^j p_j - m_0 \right) \Psi_{\vec{p}} \\
& = \left(\begin{bmatrix} 0 & \mathbb{1} \\ \mathbb{1} & 0 \end{bmatrix} E - \begin{bmatrix} 0 & -\sigma_1 \\ \sigma_1 & 0 \end{bmatrix} p_1 - \begin{bmatrix} 0 & -\sigma_2 \\ \sigma_2 & 0 \end{bmatrix} p_2 - \begin{bmatrix} 0 & -\sigma_3 \\ \sigma_3 & 0 \end{bmatrix} p_3 - m_0 \right) \Psi_{\vec{p}} \\
& = \left(\begin{bmatrix} 0 & E \\ E & 0 \end{bmatrix} - \begin{bmatrix} 0 & -\sigma_1 p_1 \\ \sigma_1 p_1 & 0 \end{bmatrix} - \begin{bmatrix} 0 & -\sigma_2 p_2 \\ \sigma_2 p_2 & 0 \end{bmatrix} - \begin{bmatrix} 0 & -\sigma_3 p_3 \\ \sigma_3 p_3 & 0 \end{bmatrix} - m_0 \right) \Psi_{\vec{p}} \\
& = \left(\begin{bmatrix} 0 & E \\ E & 0 \end{bmatrix} - \begin{bmatrix} 0 & -\sigma_1 p_1 - \sigma_2 p_2 - \sigma_3 p_3 \\ \sigma_1 p_1 + \sigma_2 p_2 + \sigma_3 p_3 & 0 \end{bmatrix} - m_0 \right) \Psi_{\vec{p}} \tag{43.134} \\
& = \left(\begin{bmatrix} 0 & E \\ E & 0 \end{bmatrix} - \begin{bmatrix} 0 & -\vec{p} \circ \vec{\sigma} \\ \vec{p} \circ \vec{\sigma} & 0 \end{bmatrix} - m_0 \right) \Psi_{\vec{p}} = \left(\begin{bmatrix} 0 & E \\ E & 0 \end{bmatrix} + \begin{bmatrix} 0 & \vec{p} \circ \vec{\sigma} \\ -\vec{p} \circ \vec{\sigma} & 0 \end{bmatrix} - m_0 \right) \Psi_{\vec{p}} \\
& = \left(\begin{bmatrix} 0 & \vec{p} \circ \vec{\sigma} + E \\ -\vec{p} \circ \vec{\sigma} + E & 0 \end{bmatrix} - m_0 \right) \Psi_{\vec{p}} = \left(\begin{bmatrix} 0 & \vec{p} \circ \vec{\sigma} + E \\ -\vec{p} \circ \vec{\sigma} + E & 0 \end{bmatrix} \Psi_{\vec{p}} - m_0 \Psi_{\vec{p}} \right) \\
& = \begin{pmatrix} 0\phi_a + (\vec{p} \circ \vec{\sigma} + E)\phi_b - m_0\phi_a \\ (-\vec{p} \circ \vec{\sigma} + E)\phi_a + 0\phi_b - m_0\phi_b \end{pmatrix} = \begin{pmatrix} -m_0\phi_a + (\vec{p} \circ \vec{\sigma} + E)\phi_b \\ (-\vec{p} \circ \vec{\sigma} + E)\phi_a - m_0\phi_b \end{pmatrix} = \begin{pmatrix} -m_0 & \vec{p} \circ \vec{\sigma} + E \\ -\vec{p} \circ \vec{\sigma} + E & -m_0 \end{pmatrix} \begin{pmatrix} \phi_a \\ \phi_b \end{pmatrix} \\
& = \begin{pmatrix} -m_0 & \vec{p} \circ \vec{\sigma} + E \\ -\vec{p} \circ \vec{\sigma} + E & -m_0 \end{pmatrix} \Psi_{\vec{p}} = \vec{0}
\end{aligned}$$

et nous ne serions pas tombés sur une condition aussi esthétique et physique pour qu'il y ait des solutions!

La masse étant toujours positive, l'équation de Dirac comporte donc quatre solutions linéairement indépendantes, dont deux avec une énergie positive :

$$E = E_{\vec{p}} = \sqrt{\|\vec{p}\|^2 + m_0^2} \tag{43.135}$$

et deux avec une énergie négative:

$$E = -E_{\vec{p}} \tag{43.136}$$

Il s'agit donc bien des antiparticules que nous avons déterminées lors de notre étude de l'équation de Klein-Gordon libre mais avec le spin en plus d'où le doublage des solutions supplémentaires (deux orientations du spin possibles par particule et par antiparticule). Avec la représentation Chirale, nous ne serions pas retombés sur ce résultat. D'où la nécessité de l'utilisation de la représentation de Dirac des matrices de Pauli.

Nous savons donc qu'il existe des solutions à l'équation de Dirac. Déterminons maintenant celles-ci. Posons:

$$\Psi_{\vec{p}} = \begin{pmatrix} \phi_a \\ \phi_b \end{pmatrix} \tag{43.137}$$

où ϕ_a, ϕ_b sont les deux doubles composantes du spineur. Nous écrivons ainsi le système d'équations (nous utilisons, comme le faisait Dirac, le simple symbole de multiplication en lieu et place du symbole du produit scalaire):

$$\begin{aligned}(E - m_0) \phi_a - \vec{p} \circ \vec{\sigma} \phi_b &= \vec{0} \\ \vec{p} \circ \vec{\sigma} \phi_a - (E + m_0) \phi_b &= \vec{0}\end{aligned}\quad (43.138)$$

ce qui nous donne (ne pas oublier que le terme au dénominateur est en réalité une matrice 2×2 ...):

$$\boxed{\phi_b = \frac{\vec{p} \circ \vec{\sigma}}{(E + m_0)} \phi_a} \text{ et } \boxed{\phi_a = \frac{\vec{p} \circ \vec{\sigma}}{(E - m_0)} \phi_b} \quad (43.139)$$

Ainsi, nous avons:

$$\Psi_{\vec{p}} = \begin{pmatrix} \phi_a \\ \phi_b \end{pmatrix} = \begin{pmatrix} \phi_a \\ \frac{\vec{p} \circ \vec{\sigma}}{(E + m_0)} \phi_a \end{pmatrix} = \begin{pmatrix} \frac{\vec{p} \circ \vec{\sigma}}{(E - m_0)} \phi_b \\ \phi_b \end{pmatrix} \quad (43.140)$$

Nous savons qu'il existe des solutions et la physique quantique nous impose que ces solutions soient linéairement indépendantes. Ainsi, choisissons les solutions pour ϕ_a, ϕ_b comme étant proportionnelles à:

$$\phi_a, \phi_b = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ ou } \phi_a, \phi_b = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad (43.141)$$

et comme (cf. chapitre de Calcul Spinoriel):

$$\vec{p} \circ \vec{\sigma} = \begin{bmatrix} p_z & p_x - ip_y \\ p_x + ip_y & -p_z \end{bmatrix} \quad (43.142)$$

nous avons alors les possibilités suivantes:

$$\begin{aligned}(1) \quad \phi_a &= \begin{pmatrix} 1 \\ 0 \end{pmatrix} \Rightarrow \phi_b = \frac{\vec{p} \circ \vec{\sigma}}{E + m_0} \phi_a = \frac{\vec{p} \circ \vec{\sigma}}{E + m_0} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{E + m_0} \begin{pmatrix} p_z \\ p_+ \end{pmatrix} \\ (2) \quad \phi_a &= \begin{pmatrix} 0 \\ 1 \end{pmatrix} \Rightarrow \phi_b = \frac{\vec{p} \circ \vec{\sigma}}{E + m_0} \phi_a = \frac{\vec{p} \circ \vec{\sigma}}{E + m_0} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{E + m_0} \begin{pmatrix} p_- \\ -p_z \end{pmatrix} \\ (3) \quad \phi_b &= \begin{pmatrix} 1 \\ 0 \end{pmatrix} \Rightarrow \phi_a = \frac{\vec{p} \circ \vec{\sigma}}{E - m_0} \phi_b = \frac{\vec{p} \circ \vec{\sigma}}{E - m_0} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{E - m_0} \begin{pmatrix} p_z \\ p_+ \end{pmatrix} \\ (4) \quad \phi_b &= \begin{pmatrix} 0 \\ 1 \end{pmatrix} \Rightarrow \phi_a = \frac{\vec{p} \circ \vec{\sigma}}{E - m_0} \phi_b = \frac{\vec{p} \circ \vec{\sigma}}{E - m_0} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{E - m_0} \begin{pmatrix} p_- \\ -p_z \end{pmatrix}\end{aligned} \quad (43.143)$$

La question est maintenant... devons-nous utiliser $E = +\sqrt{\|\vec{p}\|^2 + m_0^2}$ ou $E = -\sqrt{\|\vec{p}\|^2 + m_0^2}$? Eh bien, pour (1) et (2) nous devons utiliser $E = +\sqrt{}$ sinon $1/(E + m_0)$ devient une singularité pour $\vec{p} = \vec{0}$. Pour (3) et (4) nous devons utiliser $E = -\sqrt{}$ sinon $1/(E - m_0)$ devient une singularité pour $\vec{p} = \vec{0}$.

Remarque: Le terme $E = +\sqrt{}$ est souvent appelé "**solution particule**" dans la littérature et le terme $E = -\sqrt{}$ "**solution antiparticule**".

En reprenant

$$\Psi_{\vec{p}} = \begin{pmatrix} \phi_a \\ \phi_b \end{pmatrix} \quad (43.144)$$

et en notant les spineurs (nous changeons de notation):

$$\chi^{(+)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \chi^{(-)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad (43.145)$$

Nous avons finalement en utilisant (1) et (2) et en notant $N(\)$ la partie de solution que nous devrions normaliser les solutions suivantes possibles et qui sont indépendantes:

$$\begin{aligned} \Psi_{\vec{p}}^{(1)} = N \begin{pmatrix} \phi_a \\ \phi_b \end{pmatrix} &= N \begin{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ \frac{p_z}{E + m_0} \\ \frac{p_+}{E + m_0} \end{pmatrix} = N \begin{pmatrix} 1 \\ \frac{\vec{p} \circ \vec{\sigma}}{E + m_0} \end{pmatrix} \chi^{(+)} \\ \Psi_{\vec{p}}^{(2)} = N \begin{pmatrix} \phi_a \\ \phi_b \end{pmatrix} &= N \begin{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ \frac{p_-}{E + m_0} \\ \frac{-p_z}{E + m_0} \end{pmatrix} = N \begin{pmatrix} 1 \\ \frac{\vec{p} \circ \vec{\sigma}}{E + m_0} \end{pmatrix} \chi^{(-)} \end{aligned} \quad (43.146)$$

avec $E = +\sqrt{\quad}$ ainsi que:

$$\begin{aligned} \Psi_{\vec{p}}^{(3)} = N \begin{pmatrix} \phi_a \\ \phi_b \end{pmatrix} &= N \begin{pmatrix} \begin{pmatrix} \frac{p_z}{E - m_0} \\ \frac{p_+}{E - m_0} \end{pmatrix} \\ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \end{pmatrix} = N \begin{pmatrix} \frac{\vec{p} \circ \vec{\sigma}}{E - m_0} \\ 1 \end{pmatrix} \chi^{(+)} \\ \Psi_{\vec{p}}^{(4)} = N \begin{pmatrix} \phi_a \\ \phi_b \end{pmatrix} &= N \begin{pmatrix} \begin{pmatrix} \frac{p_-}{E - m_0} \\ \frac{-p_z}{E - m_0} \end{pmatrix} \\ \begin{pmatrix} 0 \\ 1 \end{pmatrix} \end{pmatrix} = N \begin{pmatrix} \frac{\vec{p} \circ \vec{\sigma}}{E - m_0} \\ 1 \end{pmatrix} \chi^{(-)} \end{aligned} \quad (43.147)$$

avec $E = -\sqrt{\quad}$.

Ce qui peut s'abréger:

$$\Psi_{\underline{E}_+}^{(\pm)}(\vec{p}) = N \begin{pmatrix} 1 \\ \frac{\vec{p} \circ \vec{\sigma}}{E + m_0} \end{pmatrix} \chi^{(\pm)}, \Psi_{\underline{E}_-}^{(\pm)}(\vec{p}) = N \begin{pmatrix} \frac{\vec{p} \circ \vec{\sigma}}{E - m_0} \\ 1 \end{pmatrix} \chi^{(\pm)} \quad (43.148)$$

Indiquons que le terme $\vec{p} \circ \vec{\sigma}$ qui est donc la projection de la quantité de mouvement (impulsion) sur une entité mathématique directement liée au spin est appelé "**hélicité**".

4. ÉQUATION DE DIRAC LIBRE LINÉARISÉE

Nous avons vu tout au début de notre étude la physique quantique ondulatoire que l'équation de Schrödinger classique d'évolution était:

$$i\hbar \partial_t \Psi = \left(E_{pot} - \frac{\hbar^2}{2m} \Delta \right) \Psi \quad (43.149)$$

soit une équation différentielle d'un premier ordre par rapport au temps et du second par rapport aux coordonnées spatiales.

Nous avons ensuite déterminé l'équation d'évolution relativiste de Schrödinger (équation de Klein-Gordon libre) donnée par:

$$\hbar^2 \partial_t^2 \Psi = (\hbar^2 c^2 \Delta - m_0^2 c^4) \Psi \quad (43.150)$$

Nous remarquons qu'en passant à une forme relativiste, nous avons maintenant une équation différentielle du second ordre dans le temps et dans l'espace.

Ensuite en passant par l'équation de Klein-Gordon généralisée qui contenait également une équation différentielle du second ordre en temps et en espace:

$$\left[(i\hbar \partial_t - qU)^2 - (-i\hbar \vec{\nabla} - q\vec{A})^2 c^2 \right] \Psi = m_0^2 c^4 \Psi \quad (43.151)$$

et dans l'équation de Dirac libre, nous obtenons de même une équation différentielle matricielle de premier ordre en temps et de deuxième ordre en espace:

$$i\hbar \frac{\partial}{\partial t} \begin{bmatrix} \Psi_a \\ \Psi_b \end{bmatrix} = \begin{bmatrix} m_0 c^2 - \frac{\hbar^2}{2m_0} \Delta & -\frac{\hbar^2}{2m_0} \Delta \\ \frac{\hbar^2}{2m_0} \Delta & -m_0 c^2 + \frac{\hbar^2}{2m_0} \Delta \end{bmatrix} \begin{bmatrix} \Psi_a \\ \Psi_b \end{bmatrix} \quad (43.152)$$

Ces changements d'ordre des différentielles d'un modèle relativiste ou non imposent bien sûr dans le cas d'un premier ordre de connaître les conditions initiales en temps et en espace de l'équation d'onde, ce qui est faisable. Cependant, lorsqu'un second ordre apparaît, il faut alors en plus connaître les conditions initiales des dérivées des fonctions d'onde (cf. chapitre de Calcul Différentiel Et Intégral). De plus, même si mathématiquement la rigueur nous a amené naturellement aux différents ordres obtenus, il est étrange en passant d'un modèle relativiste que nous changions d'ordre. Pourquoi ? : pour la simple raison qu'en approximant les équations relativistes, nous n'arrivons pas avec le facteur de la constante de Planck à faire des approximations (développement en série de v/c) qui nous ramèneraient à du premier ordre. Les équations relativistes et non relativistes sont donc a priori incompatibles dans les limites non relativistes !

La méthode de Dirac pour résoudre ce problème aura été la suivante:

Les ordres de l'équation différentielle de Klein-Gordon venant à la base de la relation (voir les débuts de nos développements de l'équation de Klein-Gordon libre) de l'énergie totale en l'absence de tout champ:

$$\begin{aligned} E^2 &= m_0^2 c^4 + c^2 \|\vec{p}\|^2 = m_0^2 c^4 + c^2 p_x^2 + c^2 p_y^2 + c^2 p_z^2 = m_0^2 c^4 + c^2 p_1^2 + c^2 p_2^2 + c^2 p_3^2 \\ &= m_0^2 c^4 + c^2 (p_1^2 + p_2^2 + p_3^2) = H^2 \end{aligned} \quad (43.153)$$

Dirac aurait eu donc l'idée géniale de linéariser cet hamiltonien en posant:

$$H_D = \beta m_0 c^2 + c \alpha_1 p_1 + c \alpha_2 p_2 + c \alpha_3 p_3 = \beta m_0 c^2 + c \vec{\alpha} \circ \vec{p} \quad (43.154)$$

dont nous devons déterminer les paramètres $\beta, \alpha(\alpha_1, \alpha_2, \alpha_3)$ qui pourront être des scalaires, des vecteurs ou des matrices (attendons un peu... la réponse viendra). Il s'ensuit que l'hamiltonien est lui aussi soit un scalaire, soit un vecteur, soit une matrice.

Ainsi, l'équation d'onde d'évolution relativiste la plus simple que nous pourrions construire sera:

$$i\hbar \partial_t \Psi = H \Psi \Rightarrow i\hbar \partial_t \Psi = H_D \Psi \Rightarrow i\hbar \partial_t \Psi = (\beta m_0 c^2 + c \vec{\alpha} \circ \vec{p}) \Psi \quad (43.155)$$

Sous une forme beaucoup plus commune dans la littérature:

$$(i\hbar \partial_t - \beta m_0 c^2 - c \vec{\alpha} \circ \vec{p}) \Psi = 0 \quad (43.156)$$

Comme ici:

$$\vec{p} = -i\hbar \vec{\nabla} \quad (43.157)$$

nous retrouvons alors la relation antérieure aussi sous la forme:

$$(i\hbar \partial_t + i\hbar c \vec{\alpha} \circ \vec{\nabla} - \beta m_0 c^2) \Psi = 0 \quad (43.158)$$

Si la quantité de mouvement venait à être nulle, nous retrouverions ainsi l'énergie au repos pour l'hamiltonien:

$$i\hbar \partial_t \Psi = \beta m_0 c^2 \Psi \quad (43.159)$$

où comme nous allons le voir plus loin $\beta = \pm 1$.

La validité de cette linéarisation devra être vérifiée en retrouvant les résultats obtenus lors de notre étude précédente de l'équation de Dirac.

Élevons maintenant l'opérateur au carré soit:

$$H_D^2 = (\beta m_0 c^2 + c \alpha_1 p_1 + c \alpha_2 p_2 + c \alpha_3 p_3)^2 \quad (43.160)$$

et posons:

$$A = (\beta m_0 c^2 + c \alpha_1 p_1) \text{ et } B = (c \alpha_2 p_2 + c \alpha_3 p_3) \quad (43.161)$$

À ce stade, il est important de remarquer que nous travaillons peut-être avec des opérateurs (des matrices typiquement) qui pourraient ne pas commuter car les α, β sont inconnus. Dès lors, l'élévation au carré sera effectuée comme suit:

$$H_D^2 = (A + B)^2 = A^2 + B^2 + AB + BA \quad (43.162)$$

soit simplement développé mais sans simplifier la somme $AB+BA$ en un $2AB$ ou en un $2BA$ puisque nous ne sommes pas sûr pour l'instant s'il y aura commutativité ou pas.

Nous développons ainsi l'hamiltonien de Dirac

$$\begin{aligned} H_D^2 = & (\beta^2 m_0^2 c^4 + c^2 \alpha_1^2 p_1^2 + \beta m_0 c^2 \cdot c \alpha_1 p_1 + c \alpha_1 p_1 \beta m_0 c^2) + \\ & (c^2 \alpha_2^2 p_2^2 + c^2 \alpha_3^2 p_3^2 + c \alpha_2 p_2 \cdot c \alpha_3 p_3 + c \alpha_3 p_3 \cdot c \alpha_2 p_2) + \\ & (\beta m_0 c^2 + c \alpha_1 p_1)(c \alpha_2 p_2 + c \alpha_3 p_3) + (c \alpha_2 p_2 + c \alpha_3 p_3)(\beta m_0 c^2 + c \alpha_1 p_1) \end{aligned} \quad (43.163)$$

En effectuant les produits des termes entre parenthèses et en respectant l'ordre des opérateurs, il vient:

$$\begin{aligned} H_D^2 = & \underbrace{\beta^2 m_0^2 c^4}_1 + \underbrace{c^2 \alpha_1^2 p_1^2}_2 + \underbrace{m_0 c^3 \beta \alpha_1 p_1}_{11} + \underbrace{m_0 c^3 \alpha_1 p_1 \beta}_{12} + \underbrace{c^2 \alpha_2^2 p_2^2}_3 + \underbrace{c^2 \alpha_3^2 p_3^2}_4 + \\ & \underbrace{c^2 \alpha_2 p_2 \alpha_3 p_3}_9 + \underbrace{c^2 \alpha_3 p_3 \alpha_2 p_2}_{10} + \underbrace{m_0 c^3 \beta \alpha_2 p_2}_{13} + \underbrace{m_0 c^3 \beta \alpha_3 p_3}_{15} + \underbrace{c^2 \alpha_1 p_1 \alpha_2 p_2}_5 + \\ & \underbrace{c^2 \alpha_1 p_1 \alpha_3 p_3}_7 + \underbrace{m_0 c^3 \alpha_2 p_2 \beta}_{14} + \underbrace{c^2 \alpha_2 p_2 \alpha_1 p_1}_6 + \underbrace{m_0 c^3 \alpha_3 p_3 \beta}_{16} + \underbrace{c^2 \alpha_3 p_3 \alpha_1 p_1}_8 \end{aligned} \quad (43.164)$$

En groupant certains termes:

$$\begin{aligned}
H_D^2 = & \underbrace{\beta^2 m_0^2 c^4}_1 + c^2 \left(\underbrace{\alpha_1^2 p_1^2}_2 + \underbrace{\alpha_2^2 p_2^2}_3 + \underbrace{\alpha_3^2 p_3^2}_4 \right) + c^2 \left(\underbrace{\alpha_1 \alpha_2}_5 + \underbrace{\alpha_2 \alpha_1}_6 \right) p_1 p_2 + \\
& c^2 \left(\underbrace{\alpha_1 \alpha_3}_7 + \underbrace{\alpha_3 \alpha_1}_8 \right) p_1 p_3 + c^2 \left(\underbrace{\alpha_2 \alpha_3}_9 + \underbrace{\alpha_3 \alpha_2}_{10} \right) p_2 p_3 + m_0 c^3 \left(\underbrace{\beta \alpha_1}_11 + \underbrace{\alpha_1 \beta}_{12} \right) p_1 + \\
& m_0 c^3 \left(\underbrace{\beta \alpha_2}_{13} + \underbrace{\alpha_2 \beta}_{14} \right) p_2 + m_0 c^3 \left(\underbrace{\beta \alpha_3}_{15} + \underbrace{\alpha_3 \beta}_{16} \right) p_3
\end{aligned}
\tag{43.165}$$

Pour être conforme à nos hypothèses de linéarisation, nous devons avoir:

$$\begin{aligned}
1 & \rightarrow \beta^2 m_0^2 c^4 \equiv m_0^2 c^4 \Rightarrow \beta^2 = 1 \\
2,3,4 & \rightarrow c^2 (\alpha_1^2 p_1^2 + \alpha_2^2 p_2^2 + \alpha_3^2 p_3^2) \equiv c^2 (p_1^2 + p_2^2 + p_3^2) \Rightarrow \alpha_1^2 = \alpha_2^2 = \alpha_3^2 = 1 \\
5,6 & \rightarrow c^2 (\alpha_1 \alpha_2 + \alpha_2 \alpha_1) p_1 p_2 = 0 \Rightarrow \alpha_1 \alpha_2 + \alpha_2 \alpha_1 = 0 \\
7,8 & \rightarrow c^2 (\alpha_1 \alpha_3 + \alpha_3 \alpha_1) p_1 p_3 = 0 \Rightarrow \alpha_1 \alpha_3 + \alpha_3 \alpha_1 = 0 \\
9,10 & \rightarrow c^2 (\alpha_2 \alpha_3 + \alpha_3 \alpha_2) p_2 p_3 = 0 \Rightarrow \alpha_2 \alpha_3 + \alpha_3 \alpha_2 = 0 \\
11,12 & \rightarrow m_0 c^3 (\beta \alpha_1 + \alpha_1 \beta) p_1 = 0 \Rightarrow \beta \alpha_1 + \alpha_1 \beta = 0 \\
13,14 & \rightarrow m_0 c^3 (\beta \alpha_2 + \alpha_2 \beta) p_2 = 0 \Rightarrow \beta \alpha_2 + \alpha_2 \beta = 0 \\
15,16 & \rightarrow m_0 c^3 (\beta \alpha_3 + \alpha_3 \beta) p_3 = 0 \Rightarrow \beta \alpha_3 + \alpha_3 \beta = 0
\end{aligned}
\tag{43.166}$$

et nous allons de suite voir que pour satisfaire à ces conditions les α, β devront être des matrices 4×4 .

Écrit sous forme de commutateurs, nous avons donc les trois conditions suivantes à satisfaire:

$$\begin{aligned}
\beta^2 &= \alpha_1^2 = \alpha_2^2 = \alpha_3^2 = 1 \\
[\alpha_i, \alpha_j]_+ &= [\alpha_i, \alpha_k]_+ = [\alpha_j, \alpha_k]_+ = 0 \\
[\alpha_i, \beta]_+ &= [\alpha_j, \beta]_+ = [\alpha_k, \beta]_+ = 0
\end{aligned}
\tag{43.167}$$

Nous observons ce qui suit:

- Le carré de chaque opérateur β et α_j est égal à 1 (ou à la matrice unitaire s'il s'agit de matrice...).
- $[\alpha_i, \alpha_j]_+ = 0, \forall i, j = \{1, 2, 3\}, i \neq j$ est un anticommutateur.
- $[\alpha_i, \beta]_+ = 0$ est un anticommutateur.

ces trois relations peuvent se résumer comme suit:

$$\beta^2 = 1, [\alpha_i, \alpha_j]_+ = 2\delta_{ij}, [\alpha_i, \beta]_+ = 0
\tag{43.168}$$

À ce stade, nous devons rechercher quels sont les objets mathématiques répondant aux trois conditions ci-dessus. Nous pourrions montrer qu'une matrice carrée de dimension 2 ou 3 ne répond pas aux trois conditions et un scalaire encore moins!

Dirac a alors adopté par analogie aux développements antérieurs, des matrices carrées de dimension 4 incluant des matrices de Pauli (comme par hasard...) et a admis pour β une matrice unité (ce choix fait par Dirac est particulier, il y a d'autres choix possibles).

Donc ce que nous notions "1" avant est au fait une matrice unitaire carrée de dimension 4! Et donc les fonctions d'onde doivent avoir 4 composantes et des êtres mathématiques qui ont cela sont les spineurs!

Les matrices considérées par Dirac (qui sont un choix particulier!) sont donc pour α_i et β :

$$\alpha_i = \begin{bmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{bmatrix} \quad \text{et} \quad \beta = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (43.169)$$

et par exemple dans certains domaines de la physique quantique on utilise le choix de Weyl:

$$\alpha_i = \begin{bmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{bmatrix}, \beta = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (43.170)$$

Dans le choix considéré par Dirac ou de Weyl, nous avons les matrices de Pauli et la matrice unitaire suivantes:

$$\sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \sigma_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \quad \sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad \mathbb{1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (43.171)$$

Ce qui conduit aux matrices α_i, β :

$$\begin{aligned} \alpha_1 &= \begin{bmatrix} 0 & \sigma_1 \\ \sigma_1 & 0 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \end{bmatrix} & \alpha_2 &= \begin{bmatrix} 0 & \sigma_2 \\ \sigma_2 & 0 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \end{bmatrix} \\ \alpha_3 &= \begin{bmatrix} 0 & \sigma_3 \\ \sigma_3 & 0 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \\ \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \end{bmatrix} & \beta &= \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & -\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \end{aligned} \quad (43.172)$$

On peut vérifier que les conditions de linéarisation sont vérifiées par les matrices précédentes:

- Première condition:

$$\beta^2 = 1 \rightarrow \left\{ \begin{array}{l} \beta_+ \cdot \beta_+ = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 1 \\ \beta_- \cdot \beta_- = - \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \left\{ - \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \right\} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 1 \end{array} \right\} \quad (43.173)$$

De même pour les α_j :

$$\begin{aligned} \alpha_1^2 &= \begin{bmatrix} 0 & \sigma_1 \\ \sigma_1 & 0 \end{bmatrix}^2 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = 1 \\ \alpha_2^2 &= \begin{bmatrix} 0 & \sigma_2 \\ \sigma_2 & 0 \end{bmatrix}^2 = \begin{bmatrix} 0 & 0 & 0 & -i \\ 0 & 0 & i & 0 \\ 0 & -i & 0 & 0 \\ i & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & -i \\ 0 & 0 & i & 0 \\ 0 & -i & 0 & 0 \\ i & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = 1 \\ \alpha_3^2 &= \begin{bmatrix} 0 & \sigma_3 \\ \sigma_3 & 0 \end{bmatrix}^2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = 1 \end{aligned} \quad (43.174)$$

La première condition est donc bien remplie!

- Deuxième condition (attention aux notations qui dérapent un peu par tradition entre matrices et scalaires!):

$$\begin{aligned} [\alpha_i, \alpha_j]_+ &= \begin{bmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{bmatrix} \begin{bmatrix} 0 & \sigma_j \\ \sigma_j & 0 \end{bmatrix} + \begin{bmatrix} 0 & \sigma_j \\ \sigma_j & 0 \end{bmatrix} \begin{bmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{bmatrix} = \\ &= \begin{bmatrix} \sigma_i \sigma_j & 0 \\ 0 & \sigma_i \sigma_j \end{bmatrix} + \begin{bmatrix} \sigma_j \sigma_i & 0 \\ 0 & \sigma_j \sigma_i \end{bmatrix} = \begin{bmatrix} \sigma_i \sigma_j + \sigma_j \sigma_i & 0 \\ 0 & \sigma_i \sigma_j + \sigma_j \sigma_i \end{bmatrix} \\ &= (\sigma_i \sigma_j + \sigma_j \sigma_i) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \sigma_i \sigma_j + \sigma_j \sigma_i \\ \text{Si } i=j &\rightarrow 2(\sigma_i \sigma_i) = 2(\sigma_i)^2 \end{aligned} \quad (43.175)$$

$$\text{Si } i=1 \rightarrow 2(\sigma_1)^2 = 2 \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}^2 = 2(-1)^2 = 2$$

$$\text{Si } i=2 \rightarrow 2(\sigma_2)^2 = 2 \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}^2 = 2(-1)^2 = 2$$

$$\text{Si } i=3 \rightarrow 2(\sigma_3)^2 = 2 \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}^2 = 2(-1)^2 = 2$$

et:

Si $i = 1$ et $j = 2$

$$\sigma_1 \sigma_2 + \sigma_2 \sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} + \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix} + \begin{bmatrix} -i & 0 \\ 0 & i \end{bmatrix} = 0 \quad (43.176)$$

Si $i = 1$ et $j = 3$

$$\sigma_1 \sigma_3 + \sigma_3 \sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = 0$$

Donc:

$$\sigma_i \sigma_j + \sigma_j \sigma_i = 2 \delta_{ij} \quad (43.177)$$

la deuxième condition est bien remplie.

- Troisième condition:

$$[\alpha_i, \beta]_+ = \begin{bmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{bmatrix} = \begin{bmatrix} 0 & -\sigma_i \\ \sigma_i & 0 \end{bmatrix} + \begin{bmatrix} 0 & \sigma_i \\ -\sigma_i & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad (43.178)$$

La troisième condition est bien remplie.

En se référant à l'équation de début écrite avec le formalisme de Dirac

$$i\hbar \partial_t \Psi = H_D \Psi \Rightarrow i\hbar \partial_t \Psi = \beta m_0 c^2 \Psi + c \alpha_1 p_1 \Psi + c \alpha_2 p_2 \Psi + c \alpha_3 p_3 \Psi \quad (43.179)$$

Avec:

$$\beta = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}, \alpha_1 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \alpha_2 = \begin{bmatrix} 0 & 0 & 0 & -i \\ 0 & 0 & i & 0 \\ 0 & -i & 0 & 0 \\ i & 0 & 0 & 0 \end{bmatrix}, \alpha_3 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} \quad (43.180)$$

Ce qui donne finalement:

$$i\hbar \partial_t \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \Psi_4 \end{bmatrix} = m_0 c^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \Psi_4 \end{bmatrix} + c \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} p_1 \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \Psi_4 \end{bmatrix} + c \begin{bmatrix} 0 & 0 & 0 & -i \\ 0 & 0 & i & 0 \\ 0 & -i & 0 & 0 \\ i & 0 & 0 & 0 \end{bmatrix} p_2 \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \Psi_4 \end{bmatrix} + c \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} p_3 \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \Psi_4 \end{bmatrix} \quad (43.181)$$

Nous nous retrouvons devant une fonction d'état possédant 4 composantes dans laquelle:

$$\begin{bmatrix} \Psi_1 \\ \Psi_2 \end{bmatrix} \text{ et } \begin{bmatrix} \Psi_3 \\ \Psi_4 \end{bmatrix} \quad (43.182)$$

sont des spineurs et l'ensemble:

$$\begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \Psi_4 \end{bmatrix} \quad (43.183)$$

est donc un "bispineur de Dirac" et nous notons:

$$\Psi = \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \Psi_4 \end{bmatrix} \quad (43.184)$$

la "fonction d'état de Dirac". Le lecteur remarquera que nous retrouvons les mêmes concepts que lors de notre étude de l'équation de Dirac libre non linéarisée.

En développant, il vient:

$$\begin{aligned} i\hbar\partial_t\Psi_1 &= m_0c^2\Psi_1 + cp_1\Psi_4 - icp_2\Psi_4 + cp_3\Psi_3 = \\ &= m_0c^2\Psi_1 + c(p_1 - ip_2)\Psi_4 + cp_3\Psi_3 \\ i\hbar\partial_t\Psi_2 &= m_0c^2\Psi_2 + cp_1\Psi_3 + icp_2\Psi_3 - cp_3\Psi_4 = \\ &= m_0c^2\Psi_2 + c(p_1 + ip_2)\Psi_3 - cp_3\Psi_4 \\ i\hbar\partial_t\Psi_3 &= -m_0c^2\Psi_3 + cp_1\Psi_2 - icp_2\Psi_2 + cp_3\Psi_4 = \\ &= -m_0c^2\Psi_3 + c(p_1 - ip_2)\Psi_2 + cp_3\Psi_1 \\ i\hbar\partial_t\Psi_4 &= -m_0c^2\Psi_4 + cp_1\Psi_1 + icp_2\Psi_1 - cp_3\Psi_2 = \\ &= -m_0c^2\Psi_4 + c(p_1 + ip_2)\Psi_1 - cp_3\Psi_2 \end{aligned} \quad (1)$$

(43.185)

Pour un électron libre, nous savons que la solution est:

$$\Psi(\vec{r}, t) = \Psi_0 \exp\left(-i \frac{Et - \vec{p} \circ \vec{r}}{\hbar}\right) \quad (43.186)$$

Avec le bispineur de Dirac, nous avons:

$$\Psi(\vec{r}, t) = \Psi_\alpha(\vec{r}) \exp\left(-i \frac{Et}{\hbar}\right) \quad (43.187)$$

avec:

$$\Psi_{\alpha}(\vec{r}) = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} \exp\left(i \frac{\vec{p} \circ \vec{r}}{\hbar}\right) \quad (43.188)$$

où b_1 à b_4 sont les composantes du bispineur de Dirac.

Nous noterons:

$$\left. \begin{aligned} \Psi_1 &= b_1 \begin{bmatrix} \\ \\ \\ \end{bmatrix} \\ \Psi_2 &= b_2 \begin{bmatrix} \\ \\ \\ \end{bmatrix} \\ \Psi_3 &= b_3 \begin{bmatrix} \\ \\ \\ \end{bmatrix} \\ \Psi_4 &= b_4 \begin{bmatrix} \\ \\ \\ \end{bmatrix} \end{aligned} \right\} \text{ avec } \begin{bmatrix} \\ \\ \\ \end{bmatrix} = \exp\left(i \frac{\vec{p} \circ \vec{r}}{\hbar}\right) \exp\left(-i \frac{Et}{\hbar}\right) \quad (2)$$

(43.189)

En calculant leurs dérivées par rapport à t :

$$\begin{aligned} i \hbar \partial_t \Psi_1 &= i \hbar \partial_t \left[b_1 \exp\left(i \frac{\vec{p} \circ \vec{r}}{\hbar}\right) \right] \exp\left(-i \frac{Et}{\hbar}\right) = i \hbar \left(-i \frac{E}{\hbar}\right) b_1 \begin{bmatrix} \\ \\ \\ \end{bmatrix} = E b_1 \begin{bmatrix} \\ \\ \\ \end{bmatrix} \\ i \hbar \partial_t \Psi_2 &= E b_2 \begin{bmatrix} \\ \\ \\ \end{bmatrix} \\ i \hbar \partial_t \Psi_3 &= E b_3 \begin{bmatrix} \\ \\ \\ \end{bmatrix} \\ i \hbar \partial_t \Psi_4 &= E b_4 \begin{bmatrix} \\ \\ \\ \end{bmatrix} \end{aligned} \quad (3)$$

(43.190)

Avec (2) et (3) dans (1), il vient

$$\begin{aligned} E b_1 \begin{bmatrix} \\ \\ \\ \end{bmatrix} &= m_0 c^2 b_1 \begin{bmatrix} \\ \\ \\ \end{bmatrix} + c(p_1 - i p_2) b_4 \begin{bmatrix} \\ \\ \\ \end{bmatrix} + c p_3 b_3 \begin{bmatrix} \\ \\ \\ \end{bmatrix} \\ E b_2 \begin{bmatrix} \\ \\ \\ \end{bmatrix} &= m_0 c^2 b_2 \begin{bmatrix} \\ \\ \\ \end{bmatrix} + c(p_1 + i p_2) b_3 \begin{bmatrix} \\ \\ \\ \end{bmatrix} - c p_3 b_4 \begin{bmatrix} \\ \\ \\ \end{bmatrix} \\ E b_3 \begin{bmatrix} \\ \\ \\ \end{bmatrix} &= -m_0 c^2 b_3 \begin{bmatrix} \\ \\ \\ \end{bmatrix} + c(p_1 - i p_2) b_2 \begin{bmatrix} \\ \\ \\ \end{bmatrix} + c p_3 b_1 \begin{bmatrix} \\ \\ \\ \end{bmatrix} \\ E b_4 \begin{bmatrix} \\ \\ \\ \end{bmatrix} &= -m_0 c^2 b_4 \begin{bmatrix} \\ \\ \\ \end{bmatrix} + c(p_1 + i p_2) b_1 \begin{bmatrix} \\ \\ \\ \end{bmatrix} - c p_3 b_2 \begin{bmatrix} \\ \\ \\ \end{bmatrix} \end{aligned} \quad (43.191)$$

soit un système d'équations dont les inconnues sont b_1, b_2, b_3, b_4 :

$$\begin{cases} (E - m_0 c^2) b_1 + 0 b_2 - c p_3 b_3 - c(p_1 - i p_2) b_4 = 0 & (a) \\ 0 b_1 + (E - m_0 c^2) b_2 - c(p_1 + i p_2) b_3 + c p_3 b_4 = 0 & (b) \\ -c p_3 b_1 - c(p_1 - i p_2) b_2 + (E + m_0 c^2) b_3 + 0 b_4 = 0 & (c) \\ -c(p_1 + i p_2) b_1 + c p_3 b_2 + 0 b_3 + (E + m_0 c^2) b_4 = 0 & (d) \end{cases} \quad (4)$$

(43.192)

Nous aurons des solutions non toutes nulles si et seulement si le déterminant des coefficients est nul (pour en connaître les raisons, se référer au chapitre d'Algèbre Linéaire) et donc une infinité de solutions (pour les composantes du spineur de Dirac) possibles. Soit:

$$\begin{vmatrix} E - m_0 c^2 & 0 & -c p_3 & -c(p_1 - i p_2) \\ 0 & E - m_0 c^2 & -c(p_1 + i p_2) & c p_3 \\ -c p_3 & -c(p_1 - i p_2) & E + m_0 c^2 & 0 \\ -c(p_1 + i p_2) & c p_3 & 0 & E + m_0 c^2 \end{vmatrix} = 0 \quad (43.193)$$

En simplifiant par c:

$$\begin{vmatrix} \frac{E}{c} - m_0 c & 0 & -p_3 & -(p_1 - i p_2) \\ 0 & \frac{E}{c} - m_0 c & -(p_1 + i p_2) & p_3 \\ -p_3 & -(p_1 - i p_2) & \frac{E}{c} + m_0 c & 0 \\ -(p_1 + i p_2) & p_3 & 0 & \frac{E}{c} + m_0 c \end{vmatrix} = 0 \quad (43.194)$$

La division dans le déterminant précédent permet le calcul des déterminants partiels (cf. chapitre d'Algèbre Linéaire):

$$\begin{vmatrix} \left(\frac{E}{c} - m_0 c\right)^2 & \begin{bmatrix} -p_3^2 - (p_1 - i p_2)(p_1 + i p_2) \\ -[p_3^2 + (p_1 - i p_2)(p_1 + i p_2)] \end{bmatrix} \\ \begin{bmatrix} -p_3^2 - (p_1 - i p_2)(p_1 + i p_2) \\ -[p_3^2 + (p_1 - i p_2)(p_1 + i p_2)] \end{bmatrix} & \left(\frac{E}{c} + m_0 c\right)^2 \end{vmatrix} = 0 \quad (43.195)$$

En résolvant le déterminant précédent, il vient:

$$\begin{aligned} & \left[\left(\frac{E}{c} - m_0 c \right) \left(\frac{E}{c} + m_0 c \right) \right]^2 - \left[p_3^2 + (p_1 - i p_2)(p_1 + i p_2) \right]^2 = 0 \\ & \left(\frac{E^2}{c^2} - m_0^2 c^2 \right)^2 - \left(p_3^2 + [p_1^2 - (i p_2)^2] \right)^2 = 0 \\ & \left(\frac{E^2}{c^2} - m_0^2 c^2 \right)^2 - \underbrace{(p_3^2 + p_1^2 + p_2^2)}_{(p^2)^2} = 0 \end{aligned} \quad (43.196)$$

D'où la relation suivante:

$$\frac{E^2}{c^2} - m_0^2 c^2 = p^2 \Leftrightarrow E^2 - m_0^2 c^4 = p^2 c^2 \quad (43.197)$$

Les valeurs de l'énergie données par l'équation de Dirac sont donc:

$$E = \pm \sqrt{m_0^2 c^4 + p^2 c^2} \quad (43.198)$$

Soit:

$$\begin{aligned} E_+ &= +\sqrt{m_0^2 c^4 + p^2 c^2} \\ E_- &= -\sqrt{m_0^2 c^4 + p^2 c^2} \end{aligned} \quad (43.199)$$

Si nous adoptons pour E_+ , deux valeurs constantes pour b_1 et b_2 , nous disposons de deux relations pour calculer b_3 et b_4 soit:

- Avec (4c):

$$-c p_3 b_1 - c(p_1 - i p_2) b_2 + (E_+ + m_0 c^2) b_3 = 0 \quad (43.200)$$

Soit:

$$b_3 = \frac{c[p_3 b_1 + (p_1 - i p_2) b_2]}{E_+ + m_0 c^2} \quad (43.201)$$

- Avec (4d):

$$-c(p_1 + i p_2) b_1 + c p_3 b_2 + (E_+ + m_0 c^2) b_4 = 0 \quad (43.202)$$

Soit:

$$b_4 = \frac{c[(p_1 + i p_2) b_1 - p_3 b_2]}{E_+ + m_0 c^2} \quad (43.203)$$

N.B: En adoptant $b_1 = 1, b_2 = 0$, il vient:

$$b_3 = \frac{c p_3}{E_+ + m_0 c^2} = \frac{c p_z}{E_+ + m_0 c^2}, b_4 = \frac{c(p_1 + i p_2)}{E_+ + m_0 c^2} = \frac{c p_+}{E_+ + m_0 c^2} \quad (43.204)$$

En prenant les unités naturelles:

$$b_3 = \frac{p_z}{E_+ + m_0}, b_4 = \frac{p_+}{E_+ + m_0} \quad (43.205)$$

En adoptant $b_1 = 0, b_2 = 1$, il vient:

$$b_3 = \frac{c(p_1 - i p_2)}{E_+ + m_0 c^2} = \frac{c p_-}{E_+ + m_0 c^2}, b_4 = \frac{-c p_3}{E_+ + m_0 c^2} = \frac{-c p_z}{E_+ + m_0 c^2} \quad (43.206)$$

En prenant les unités naturelles:

$$b_3 = \frac{p_-}{E_+ + m_0}, b_4 = \frac{-p_z}{E_+ + m_0} \quad (43.207)$$

Si nous adoptons pour E_- , deux valeurs constantes pour b_3, b_4 nous disposons de deux relations pour calculer b_1, b_2 soit:

- Avec (4a):

$$(E_- - m_0 c^2) b_1 - c p_3 b_3 - c(p_1 - i p_2) b_4 = 0 \quad (43.208)$$

Soit:

$$b_1 = \frac{c[p_3 b_3 + (p_1 - i p_2) b_4]}{E_- - m_0 c^2} \quad (43.209)$$

- Avec (4b):

$$(E_- - m_0 c^2) b_2 - c(p_1 + i p_2) b_3 + c p_3 b_4 = 0 \quad (43.210)$$

Soit:

$$b_2 = \frac{c[-p_3 b_4 + (p_1 + i p_2) b_3]}{E_- - m_0 c^2} \quad (43.211)$$

Notons, qu'en adoptant $b_3 = 1, b_4 = 0$, il vient:

$$b_1 = \frac{c p_3}{E_- - m_0 c^2} = \frac{c p_z}{E_- - m_0 c^2}, b_2 = \frac{c(p_1 + i p_2)}{E_- - m_0 c^2} = \frac{c p_+}{E_- - m_0 c^2} \quad (43.212)$$

Avec les unités naturelles:

$$b_1 = \frac{p_3}{E_- - m_0} = \frac{p_z}{E_- - m_0}, b_2 = \frac{p_1 + i p_2}{E_- - m_0} = \frac{p_+}{E_- - m_0} \quad (43.213)$$

En adoptant $b_3 = 0, b_4 = 1$, il vient:

$$b_1 = \frac{c(p_1 - i p_2)}{E_- - m_0 c^2} = \frac{c p_-}{E_- - m_0 c^2}, b_2 = \frac{-c p_3}{E_- - m_0 c^2} = \frac{-c p_z}{E_- - m_0 c^2} \quad (43.214)$$

Soit avec les unités naturelles:

$$b_1 = \frac{p_-}{E_- - m_0}, b_2 = \frac{-p_z}{E_- - m_0} \quad (43.215)$$

Bien que la méthode soit différente, nous retrouvons donc les coefficients des spineurs que nous avons obtenus dans notre étude de l'équation de Dirac libre classique. Cela nous rassure donc dans les hypothèses posées au début de cette linéarisation et valide ces résultats. De plus, les relations précédentes indiquent aussi une dégénérescence d'ordre deux de l'énergie pour chaque valeur de l'impulsion. En l'absence de champ extérieur, l'électron libre n'est donc pas influencé par l'orientation de son spin. Nous retrouvons donc les mêmes résultats que ce soit pour l'équation de Dirac libre classique ou linéarisée.

Cependant, l'explication donnée par Dirac pour expliquer les énergies positives et négatives est que son équation s'applique non seulement à l'état d'une particule à énergie positive (en l'occurrence l'électron) mais également à l'état d'une particule à énergie négative (son antiparticule soit le positron). La valeur absolue de ces deux énergies étant strictement égale.

La présence du signe négatif affectant l'énergie a posé problème à l'époque pour son interprétation (dans le cadre où nous omettons la variable du temps puisque nous avons vu lors de l'étude de l'équation de Klein-Gordon libre qu'une particule à énergie négative peut être vue comme une particule qui remonte le temps).

Si nous raisonnons dans le cas où le terme $p^2 c^2$ est faible comparé à $m_0 c^2$, nous nous posons la question: comment et quelles sont les conséquences d'une transition entre un état d'énergie $m_0 c^2$ à celui de l'état d'énergie $-m_0 c^2$ avec un saut ("gap") de $2m_0 c^2$ (nous retrouverons cette valeur lors de notre étude de la matérialisation dans le chapitre de Physique Nucléaire).

Dirac a recours à l'image d'une mer d'énergie négative (puisque rappelons-le, le nombre de solutions à notre système matriciel est infini, d'où l'analogie avec une mer plus qu'un contexte discret) dans laquelle tous les états d'énergie négatives sont occupés par les électrons et les états d'énergie positives seraient vides. Si un électron est soumis à une transition (via, par exemple un photon d'énergie supérieure à $2m_0 c^2$), il quitte cette mer en laissant derrière lui une lacune (le fameux "trou" de charge positive auquel les électroniciens font parfois référence...). Cette lacune devient une charge positive, d'énergie E_+ . L'apparition de cette lacune est assimilée à l'apparition d'une particule ayant une charge positive. Bien évidemment, nous pouvons nous imaginer le cas inverse, ce n'est qu'une question de conventions.

5. ÉQUATION DE DIRAC GÉNÉRALISÉE

Dans le cas de l'électron libre, nous avons donc maintes fois vu et démontré que l'hamiltonien a comme expression

$$E = \left(p^2 c^2 + m_0^2 c^4 \right)^{\frac{1}{2}} \quad (43.216)$$

Dans le cas d'un électron se déplaçant dans un champ électromagnétique, nous avons aussi démontré lors de notre étude de l'équation de Klein-Gordon au début de ce chapitre:

$$(E - qU)^2 = (\vec{p} - q\vec{A})^2 c^2 + m_0^2 c^4 \quad (43.217)$$

Soit:

$$E = \left[(\vec{p} - q\vec{A})^2 c^2 + m_0^2 c^4 \right]^{\frac{1}{2}} + qU \quad (43.218)$$

Bref fini pour le rappel!

Si maintenant, nous reprenons l'hamiltonien de Dirac pour l'électron libre démontré plus haut (qui est pour rappel une matrice 2×2):

$$H_D = c \vec{\alpha} \circ \vec{p} + \beta m_0 c^2 \quad (43.219)$$

et du fait qu'il faille rajouter à l'hamiltonien le terme de l'énergie potentielle électrostatique via la potentiel vecteur, nous avons:

$$H_D = c \vec{\alpha} \circ (\vec{p} - q\vec{A}) + qU + \beta m_0 c^2 \quad (43.220)$$

où le potentiel U devra donc aussi être exprimé sous la forme d'une matrice 2×2 diagonale.

Nous obtenons alors l'hamiltonien de Dirac généralisé sous la forme traditionnelle suivante:

$$(H_D - qU) = c \vec{\alpha} \circ (\vec{p} - q\vec{A}) + \beta m_0 c^2 \quad (43.221)$$

Nous avons donc sous une autre forme connue:

$$i\hbar \frac{\partial \Psi}{\partial t} = [c \vec{\alpha} \circ (\vec{p} - q\vec{A}) + qU + \beta m_0 c^2] \Psi \quad (43.222)$$

6. ÉQUATION DE PAULI

Considérons maintenant une représentation à deux composantes du spineur:

$$\Psi(\vec{r}, t) = \begin{bmatrix} \phi_a(\vec{r}, t) \\ \phi_b(\vec{r}, t) \end{bmatrix} \quad (43.223)$$

et rappelons que:

$$\alpha = \begin{bmatrix} 0 & \vec{\sigma} \\ \vec{\sigma} & 0 \end{bmatrix} \quad \text{et} \quad \beta = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (43.224)$$

Il vient alors:

$$i\hbar \frac{\partial \Psi}{\partial t} = c \begin{bmatrix} 0 & \vec{\sigma} \\ \vec{\sigma} & 0 \end{bmatrix} \circ (\vec{p} - q\vec{A}) \Psi + qU \Psi + \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} m_0 c^2 \Psi \quad (43.225)$$

Soit:

$$i\hbar \frac{\partial}{\partial t} \begin{bmatrix} \phi_a \\ \phi_b \end{bmatrix} = c \begin{bmatrix} 0 & \vec{\sigma} \\ \vec{\sigma} & 0 \end{bmatrix} \circ (\vec{p} - q\vec{A}) \begin{bmatrix} \phi_a \\ \phi_b \end{bmatrix} + qU \begin{bmatrix} \phi_a \\ \phi_b \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} m_0 c^2 \begin{bmatrix} \phi_a \\ \phi_b \end{bmatrix} \quad (43.226)$$

Ce qui après simplification donne:

$$i\hbar \frac{\partial}{\partial t} \begin{bmatrix} \phi_a \\ \phi_b \end{bmatrix} = \begin{pmatrix} c\vec{\sigma} \circ (\vec{p} - q\vec{A}) \phi_b \\ c\vec{\sigma} \circ (\vec{p} - q\vec{A}) \phi_a \end{pmatrix} + qU \begin{bmatrix} \phi_a \\ \phi_b \end{bmatrix} + m_0 c^2 \begin{bmatrix} \phi_a \\ -\phi_b \end{bmatrix} \quad (43.227)$$

Avant de continuer, ouvrons une parenthèse importante sinon quoi nous n'arriverons pas à trouver une solution à ces deux équations.

Rappelons qu'un des spineurs solutions de l'équation de Dirac libre était donné par (nous l'avons démontré plus haut):

$$\phi_b = \frac{\vec{p} \circ \vec{\sigma}}{E + m_0} \phi_a \quad (43.228)$$

Soit en unités S.I.:

$$\phi_b = \frac{c \vec{p} \circ \vec{\sigma}}{E + m_0 c^2} \phi_a \quad (43.229)$$

Afin de simplifier le calcul des équations antérieures nous abaisserons la situation à un cas non relativiste, c'est-à-dire lorsque l'énergie de masse est beaucoup plus grande que l'énergie cinétique. Donc la solution précédente devient (on oublie la deuxième qui poserait problème...):

$$\phi_b = \frac{\vec{p} \circ \vec{\sigma}}{2m_0 c} \phi_a \quad (43.230)$$

L'idée est alors de trouver une solution telle à:

$$i\hbar \frac{\partial}{\partial t} \begin{bmatrix} \phi_a \\ \phi_b \end{bmatrix} = \begin{pmatrix} c\vec{\sigma} \circ (\vec{p} - q\vec{A}) \phi_b \\ c\vec{\sigma} \circ (\vec{p} - q\vec{A}) \phi_a \end{pmatrix} + qU \begin{bmatrix} \phi_a \\ \phi_b \end{bmatrix} + m_0 c^2 \begin{bmatrix} \phi_a \\ -\phi_b \end{bmatrix} \quad (43.231)$$

qui lorsque nous faisons une approximation non relativiste et que nous annulons le champ magnétique (in extenso le potentiel vecteur), nous retombons sur:

$$\phi_b = \frac{\vec{p} \circ \vec{\sigma}}{2m_0 c} \phi_a \quad (43.232)$$

L'idée est simple mais il fallait y penser!

Après maints tâtonnements (eh oui la physique quantique ne s'est pas faite en un jour...) nous trouvons qu'une solution particulière satisfaisant à notre idée précédente est:

$$\begin{bmatrix} \phi_a \\ \phi_b \end{bmatrix} = \begin{bmatrix} \phi_{a0} \\ \phi_{b0} \end{bmatrix} e^{-imc^2 t / \hbar} \quad (43.233)$$

Effectivement:

$$\begin{aligned}
i\hbar \frac{\partial}{\partial t} \begin{bmatrix} \phi_{a0} \\ \phi_{b0} \end{bmatrix} e^{-imc^2 t/\hbar} &= \begin{pmatrix} c\vec{\sigma} \circ (\vec{p} - q\vec{A}) \phi_{a0} e^{-imc^2 t/\hbar} \\ c\vec{\sigma} \circ (\vec{p} - q\vec{A}) \phi_{a0} e^{-imc^2 t/\hbar} \end{pmatrix} + \left(qU \begin{bmatrix} \phi_{a0} \\ \phi_{b0} \end{bmatrix} + m_0 c^2 \begin{bmatrix} \phi_{a0} \\ -\phi_{b0} \end{bmatrix} \right) e^{-imc^2 t/\hbar} \\
i\hbar \begin{bmatrix} \frac{\partial}{\partial t} \phi_{a0} \\ \frac{\partial}{\partial t} \phi_{b0} \end{bmatrix} e^{-imc^2 t/\hbar} + i\hbar \begin{bmatrix} \phi_{a0} \\ \phi_{b0} \end{bmatrix} \frac{\partial}{\partial t} e^{-imc^2 t/\hbar} &= \begin{pmatrix} c\vec{\sigma} (\vec{p} - q\vec{A}) \phi_{a0} \\ c\vec{\sigma} (\vec{p} - q\vec{A}) \phi_{a0} \end{pmatrix} e^{-imc^2 t/\hbar} + \left(qU \begin{bmatrix} \phi_{a0} \\ \phi_{b0} \end{bmatrix} + m_0 c^2 \begin{bmatrix} \phi_{a0} \\ -\phi_{b0} \end{bmatrix} \right) e^{-imc^2 t/\hbar} \\
i\hbar \begin{bmatrix} \frac{\partial}{\partial t} \phi_{a0} \\ \frac{\partial}{\partial t} \phi_{b0} \end{bmatrix} e^{-imc^2 t/\hbar} + mc^2 \begin{bmatrix} \phi_{a0} \\ \phi_{b0} \end{bmatrix} e^{-imc^2 t/\hbar} &= \begin{pmatrix} c\vec{\sigma} (\vec{p} - q\vec{A}) \phi_{a0} \\ c\vec{\sigma} (\vec{p} - q\vec{A}) \phi_{a0} \end{pmatrix} e^{-imc^2 t/\hbar} + \left(qU \begin{bmatrix} \phi_{a0} \\ \phi_{b0} \end{bmatrix} + m_0 c^2 \begin{bmatrix} \phi_{a0} \\ -\phi_{b0} \end{bmatrix} \right) e^{-imc^2 t/\hbar} \\
i\hbar \frac{\partial}{\partial t} \begin{bmatrix} \phi_{a0} \\ \phi_{b0} \end{bmatrix} &= \begin{pmatrix} c\vec{\sigma} \circ (\vec{p} - q\vec{A}) \phi_{a0} \\ c\vec{\sigma} \circ (\vec{p} - q\vec{A}) \phi_{a0} \end{pmatrix} + qU \begin{bmatrix} \phi_{a0} \\ \phi_{b0} \end{bmatrix} - 2m_0 c^2 \begin{bmatrix} 0 \\ \phi_{b0} \end{bmatrix}
\end{aligned}
\tag{43.234}$$

Nous avons finalement deux équations:

$$\begin{aligned}
i\hbar \frac{\partial}{\partial t} \phi_{a0} &= c\vec{\sigma} \circ (\vec{p} - q\vec{A}) \phi_{b0} + qU \phi_{a0} \\
i\hbar \frac{\partial}{\partial t} \phi_{b0} &= c\vec{\sigma} \circ (\vec{p} - q\vec{A}) \phi_{a0} + qU \phi_{b0} - 2m_0 c^2 \phi_{b0}
\end{aligned}
\tag{43.235}$$

Maintenant, considérons uniquement la deuxième équation:

$$i\hbar \frac{\partial}{\partial t} \phi_{b0} = c\vec{\sigma} \circ (\vec{p} - q\vec{A}) \phi_{a0} + qU \phi_{b0} - 2m_0 c^2 \phi_{b0} \tag{43.236}$$

En supposant (gratuitement! après quoi il faudra comparer aux résultats expérimentaux) que le terme $|\partial_t \phi_b|$ est beaucoup plus petit que $|2m_0 c^2 \phi_{b0}|$ nous pouvons écrire:

$$0 \cong c\vec{\sigma} \circ (\vec{p} - q\vec{A}) \phi_{a0} + qU \phi_{b0} - 2m_0 c^2 \phi_{b0} \tag{43.237}$$

En faisant la même hypothèse avec $|qU \phi_{b0}|$ nous avons:

$$0 \cong c\vec{\sigma} \circ (\vec{p} - q\vec{A}) \phi_{a0} - 2m_0 c^2 \phi_{b0} \tag{43.238}$$

Nous avons alors (ne pas oublier que le dénominateur est une matrice diagonale en réalité...):

$$\phi_{b0} \cong \frac{\vec{\sigma} \circ (\vec{p} - q\vec{A})}{2m_0 c} \phi_{a0} \tag{43.239}$$

Or, nous voyons bien que si le champ magnétique (in extenso le potentiel vecteur) s'annule, nous retombons sur bien notre idée de départ! Le pari est donc bon!

À cause de toutes ces approximations vers le bas, la composante ϕ_b est souvent prise comme étant la "petite" composant de la fonction d'onde Ψ , relativement à la grosse composante ϕ_a .

La première équation:

$$i\hbar \frac{\partial}{\partial t} \phi_{a0} = c \vec{\sigma} \circ (\vec{p} - q\vec{A}) \phi_{b0} + qU \phi_{a0} \quad (43.240)$$

peut maintenant être simplifiée facilement en prenant la solution précédente tel que:

$$i\hbar \frac{\partial}{\partial t} \phi_{a0} \cong c \vec{\sigma} \circ (\vec{p} - q\vec{A}) \left(\frac{\vec{\sigma} \circ (\vec{p} - q\vec{A})}{2m_0 c} \right) \phi_{a0} + qU \phi_{a0} \quad (43.241)$$

Soit:

$$i\hbar \frac{\partial}{\partial t} \phi_{a0} \cong \frac{c \vec{\sigma} \circ (\vec{p} - q\vec{A}) \vec{\sigma} \circ (\vec{p} - q\vec{A})}{2m_0 c} \phi_{a0} + qU \phi_{a0} \quad (43.242)$$

En utilisant l'identité remarquable démontrée dans le chapitre de Calcul Spinoriel:

$$(\vec{\sigma} \circ \vec{a})(\vec{\sigma} \circ \vec{b}) = (\vec{a} \circ \vec{b}) \mathbb{1} + i \vec{\sigma} \circ (\vec{a} \times \vec{b}) \quad (43.243)$$

Nous avons:

$$\begin{aligned} \vec{\sigma} \circ (\vec{p} - q\vec{A}) \vec{\sigma} \circ (\vec{p} - q\vec{A}) &= (\vec{p} - q\vec{A})^2 + i \vec{\sigma} \circ [(\vec{p} - q\vec{A}) \times (\vec{p} - q\vec{A})] \\ &= (\vec{p} - q\vec{A})^2 + i \vec{\sigma} \circ [(-i\hbar \vec{\nabla} - q\vec{A}) \times (-i\hbar \vec{\nabla} - q\vec{A})] \end{aligned} \quad (43.244)$$

Détaillons le produit vectoriel en se rappelant qu'il agira comme opérateur sur ϕ_{a0} :

$$\begin{aligned} &\vec{\sigma} \circ (-i\hbar \vec{\nabla} - q\vec{A}) \times (-i\hbar \vec{\nabla} - q\vec{A}) \phi_{a0} \\ &= \underbrace{\vec{\sigma} \circ (-i\hbar \vec{\nabla} \times -i\hbar \vec{\nabla}) \phi_{a0}}_{=0} + \vec{\sigma} \circ (-i\hbar \vec{\nabla} \times -q\vec{A}) \phi_{a0} + \vec{\sigma} \circ (-q\vec{A} \times -i\hbar \vec{\nabla}) \phi_{a0} + \underbrace{\vec{\sigma} \circ (-q\vec{A} \times -q\vec{A}) \phi_{a0}}_{=0} \\ &= i\hbar q [\vec{\sigma} \circ (\vec{\nabla} \times \vec{A}) \phi_{a0} + \vec{\sigma} \circ (\vec{A} \times \vec{\nabla}) \phi_{a0}] \end{aligned} \quad (43.245)$$

Or, nous avons (cf. chapitre de Calcul Spinoriel):

$$\begin{aligned} &\vec{\sigma} \circ (\vec{\nabla} \times \vec{A}) + \vec{\sigma} \circ (\vec{A} \times \vec{\nabla}) = \\ &= \begin{bmatrix} \partial_x A_y - \partial_y A_x & (\partial_y A_z - \partial_z A_y) - i(\partial_z A_x - \partial_x A_z) \\ (\partial_y A_z - \partial_z A_y) + i(\partial_z A_x - \partial_x A_z) & -(\partial_x A_y - \partial_y A_x) \end{bmatrix} \\ &+ \begin{bmatrix} A_x \partial_y - A_y \partial_x & (A_y \partial_z - A_z \partial_y) - i(A_z \partial_x - A_x \partial_z) \\ (A_y \partial_z - A_z \partial_y) + i(A_z \partial_x - A_x \partial_z) & -(\partial_x A_y - \partial_y A_x) \end{bmatrix} \end{aligned} \quad (43.246)$$

Intéressons-nous juste à la composante dans le coin supérieur gauche (sinon les calculs sont trop longs) de cette somme de matrices. Il ne faut pas oublier que cette composante de la matrice agira sur la première composante en tant qu'opérateur sur ϕ_{a0} (notée de même...):

$$\begin{aligned} & \left(\partial_x A_y \phi_{a0} - \partial_y A_x \phi_{a0} \right) + \left(A_x \partial_y \phi_{a0} - A_y \partial_x \phi_{a0} \right) \\ &= \left(\partial_x (A_y \phi_{a0}) - \partial_y (A_x \phi_{a0}) \right) + \left(A_x \partial_y \phi_{a0} - A_y \partial_x \phi_{a0} \right) \quad (43.247) \\ &= \left(\partial_x (A_y \phi_{a0}) - A_y \partial_x \phi_{a0} \right) + \left(A_x \partial_y \phi_{a0} - \partial_y (A_x \phi_{a0}) \right) \end{aligned}$$

Or:

$$\begin{aligned} \partial_x (A_y \phi_{a0}) &= A_y \partial_x \phi_{a0} + \phi_{a0} \partial_x A_y \Rightarrow \partial_x (A_y \phi_{a0}) - A_y \partial_x \phi_{a0} = \phi_{a0} \partial_x A_y \\ \partial_y (A_x \phi_{a0}) &= A_x \partial_y \phi_{a0} + \phi_{a0} \partial_y A_x \Rightarrow A_x \partial_y \phi_{a0} - \partial_y (A_x \phi_{a0}) = -\phi_{a0} \partial_y A_x \end{aligned} \quad (43.248)$$

Donc:

$$\left(\partial_x (A_y \phi_{a0}) - A_y \partial_x \phi_{a0} \right) + \left(A_x \partial_y \phi_{a0} - \partial_y (A_x \phi_{a0}) \right) = \phi_{a0} \partial_x A_y - \phi_{a0} \partial_y A_x \quad (43.249)$$

Or, nous reconnaissons ici la troisième composante d'un produit vectoriel n'agissant pas comme opérateur. Finalement, il vient:

$$\vec{\sigma} \circ (-i\hbar \vec{\nabla} - q\vec{A}) \times (-i\hbar \vec{\nabla} - q\vec{A}) \phi_{a0} = \vec{\sigma} \circ (-i\hbar \vec{\nabla} \times -q\vec{A}) \phi_{a0} = i\hbar q \vec{\sigma} \circ (\vec{\nabla} \times \vec{A}) \phi_{a0} \quad (43.250)$$

Soit:

$$\begin{aligned} & (\vec{p} - q\vec{A})^2 + i \vec{\sigma} \circ [(-i\hbar \vec{\nabla} - q\vec{A}) \times (-i\hbar \vec{\nabla} - q\vec{A})] = (\vec{p} - q\vec{A})^2 - \hbar q \vec{\sigma} \circ (\vec{\nabla} \times \vec{A}) \\ &= (\vec{p} - q\vec{A})^2 - \hbar q \vec{\sigma} \circ \vec{B} \end{aligned} \quad (43.251)$$

Ainsi, la relation de la composante principale:

$$i\hbar \frac{\partial}{\partial t} \phi_{a0} \cong \frac{c \vec{\sigma} \circ (\vec{p} - q\vec{A}) \vec{\sigma} \circ (\vec{p} - q\vec{A})}{2m_0 c} \phi_{a0} + qU \phi_{a0} \quad (43.252)$$

Devient:

$$i\hbar \frac{\partial}{\partial t} \phi_{a0} \cong \frac{(\vec{p} - q\vec{A})^2 - \hbar q \vec{\sigma} \circ \vec{B}}{2m_0} \phi_{a0} + qU \phi_{a0} \quad (43.253)$$

Le lecteur remarquera que les notations ne sont pas des plus joyeuses (entre vecteurs, matrices et constantes il faut bien suivre pour savoir qu'est-ce qui est quoi là-dedans...).

Après réarrangement:

$$i\hbar \frac{\partial}{\partial t} \phi_{a0} \equiv \left[\frac{(\vec{p} - q\vec{A})^2}{2m_0} - \frac{\hbar q}{2m_0} \vec{\sigma} \circ \vec{B} + qU \right] \phi_{a0} \quad (43.254)$$

ce qui constitue "**l'équation de Pauli**" et décrit donc de manière relativiste les deux composantes ϕ_{a0} de liberté du spin de l'électron (il s'agit donc en réalité d'un système deux équations).

L'expression:

$$\frac{\hbar q}{2m_0} \vec{\sigma} \circ \vec{B} \quad (43.255)$$

est appelée "**terme de Stern-Gerlach**" et représente l'énergie d'interaction du champ magnétique avec le moment intrinsèque de l'électron (c'est une matrice 2×2 pour rappel).

L'équation de Pauli, et donc celle de Dirac (puisque cette dernière est plus générale), donnent le facteur gyromagnétique correct de $g = 2$ pour un électron libre. Pour vérifier ceci, prenons comme il a été fait expérimentalement, un champ magnétique constant:

Nous vérifions facilement que le choix d'un potentiel vecteur correspondant à un champ magnétique constant est alors:

$$\vec{A} = \frac{1}{2} \vec{B} \times \vec{r} \quad (43.256)$$

Ce choix va avoir pour effet de faire disparaître le potentiel vecteur au profit du champ magnétique dans l'équation de Pauli, ce qui fera apparaître l'interaction entre le moment angulaire orbital et le champ magnétique comme nous allons le voir:

Effectivement, nous avons:

$$\vec{B} = \vec{\nabla} \times \vec{A} = \vec{\nabla} \times \left(\frac{1}{2} \vec{B} \times \vec{r} \right) \quad (43.257)$$

Démonstration:

$$\begin{aligned} \vec{\nabla} \times \left(\frac{1}{2} \vec{B} \times \vec{r} \right) &= \frac{1}{2} \vec{\nabla} \times \begin{pmatrix} B_y z - B_z y \\ B_z x - B_x z \\ B_x y - B_y x \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \partial_y (B_x y - B_y x) - \partial_x (B_z x - B_x z) \\ \partial_x (B_y z - B_z y) - \partial_z (B_x y - B_y x) \\ \partial_x (B_z x - B_x z) - \partial_y (B_y z - B_z y) \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} B_x + B_x \\ B_y + B_y \\ B_z + B_z \end{pmatrix} = \begin{pmatrix} B_x \\ B_y \\ B_z \end{pmatrix} = \vec{B} \end{aligned} \quad (43.258)$$

Nous avons alors dans l'équation de Pauli:

$$\begin{aligned}
 (\vec{p} - q\vec{A})^2 &= \left(\vec{p} - \frac{q}{2} \vec{B} \times \vec{x} \right)^2 = \|\vec{p}\|^2 - q(\vec{B} \times \vec{x}) \circ \vec{p} + \underbrace{\frac{q^2}{4} \|\vec{B} \times \vec{x}\|^2}_{\cong 0} = \\
 &= \|\vec{p}\|^2 - q(\vec{B} \times \vec{x}) \circ \vec{p}
 \end{aligned} \quad (43.259)$$

Or, rappelons que nous avons vu dans le chapitre de Calcul Vectoriel que:

$$(\vec{x} \times \vec{y}) \circ \vec{z} = \vec{x} \circ (\vec{y} \times \vec{z}) \quad (43.260)$$

Cela nous donne donc:

$$\|\vec{p}\|^2 - q(\vec{B} \times \vec{x}) \circ \vec{p} = \|\vec{p}\|^2 - q\vec{B} \circ (\vec{x} \times \vec{p}) = \|\vec{p}\|^2 - q\vec{B} \circ \vec{L} \quad (43.261)$$

où:

$$\vec{L} = \vec{x} \times \vec{p} \quad (43.262)$$

noté aussi $\vec{b}$ (cf. chapitre de Mécanique Classique/Physique Quantique Corpusculaire) est donc un opérateur représentant le moment cinétique.

Nous avons donc:

$$i\hbar \frac{\partial}{\partial t} \phi_{a0} \cong \left[\frac{\|\vec{p}\|^2}{2m_0} - \frac{q\vec{B} \circ \vec{L}}{2m_0} - \frac{\hbar q}{2m_0} \vec{\sigma} \circ \vec{B} + qU \right] \phi_{a0} \quad (43.263)$$

En faisant comme l'avait Dirac en définissant l'opérateur spin comme étant la matrice (oh! on retrouve quelque chose de connu et vu dans le chapitre de Physique Quantique Ondulatoire!! c'est magnifique non?):

$$\boxed{S = \frac{1}{2} \hbar \vec{\sigma}} \quad (43.264)$$

Cette relation nous sera très utile dans le chapitre d'Informatique Quantique. Indiquons la notation suivante courante (logique) dans la littérature pour le moment magnétique intrinsèque:

$$\vec{M} = \frac{\hbar q}{2m_0} \vec{\sigma} \quad (43.265)$$

L'équation de Pauli s'écrit alors:

$$i\hbar \frac{\partial}{\partial t} \phi_{a0} \cong \left[\frac{\|\vec{p}\|^2}{2m_0} - \frac{q\vec{L} \circ \vec{B}}{2m_0} - \frac{q}{m_0} S \circ \vec{B} + qU \right] \phi_{a0} \quad (43.266)$$

ou encore:

$$i\hbar \frac{\partial}{\partial t} \phi_{a0} \cong \left[\frac{\|\vec{p}\|^2}{2m_0} - \frac{q}{2m_0} (\vec{L} + 2\vec{S}) \circ \vec{B} + qU \right] \phi_{a0} \quad (43.267)$$

ou encore en plus condensé en faisant bien attention à bien différencier ce qui est un opérateur, d'un vecteur et ce qui est un produit d'un produit scalaire et ce qui est une fonction d'un spineur... (que du bonheur...) et en posant:

$$g = 2 \quad (43.268)$$

comme étant le "**facteur de Landé**" ou "**facteur gyromagnétique**", nous avons:

$$i\hbar \frac{\partial}{\partial t} \phi_{a0} \cong \left[\frac{\|\vec{p}\|^2}{2m_0} - (\vec{\mu}_L + g\vec{\mu}_S) \circ \vec{B} + qU \right] \phi_{a0} \quad (43.269)$$

avec $\vec{\mu}_L$ étant donc le "**moment magnétique orbital**" (en comparant avec l'expression du moment magnétique démontré dans le chapitre de Magnétostatique vous verrez qu'il a effectivement la même forme et aussi les mêmes unités), $\vec{\mu}_S$ le "**moment magnétique de spin**".

On retrouve parfois l'expression entre parenthèses dans la relation ci-dessus notée de la manière suivante:

$$\vec{\mu} = \vec{\mu}_L + g\vec{\mu}_S \quad (43.270)$$

Avec tout cela le terme de Stern-Gerlach (moment magnétique) devient donc explicitement:

$$\vec{\mathcal{M}}_S = g \frac{q}{2m_0} \frac{1}{2} \hbar \vec{\sigma} = g \frac{q}{2m_0} \vec{S} \quad (43.271)$$

En se rappelant que:

$$\mu_B = \frac{q\hbar}{2m_0} \quad (43.272)$$

est le magnéton de Bohr que nous avons introduit rigoureusement dans le chapitre de Physique Quantique Corpusculaire, nous avons alors avec cette notation:

$$\vec{\mathcal{M}}_S = g \frac{q}{2m_0} \frac{1}{2} \hbar \vec{\sigma} = \frac{g}{2} \mu_B \vec{\sigma} = g \frac{q}{2m_0} \vec{S} \quad (43.273)$$

Chose intéressante... le neutron a (observation expérimentale) un moment magnétique de spin non nul et pourtant sa charge est nulle... donc il doit être constitué de particules chargées (si l'on reste dans le cadre d'un modèle d'explication de la nature par des particules...). Il est intéressant de savoir que dans la pratique nous avons pour l'électron, le proton et le neutron:

$$\begin{aligned}
 \mu_{\text{électron}} &\cong \mu_B \\
 \mu_{\text{proton}} &\cong 2.79 \mu_B \quad (43.274) \\
 \mu_{\text{neutron}} &\cong -1.91 \mu_B
 \end{aligned}$$

et les valeurs suivantes ont été mesurées pour les particules de spin 1/2 telles que l'électron, le proton et le neutron (attention le signe peut changer suivant la manière dont est notée l'équation de Dirac!):

$$\begin{aligned}
 g_e &\cong 2 \\
 g_p &\cong -5.585 \quad (43.275) \\
 g_n &= 3.826
 \end{aligned}$$

Donc la théorie de Dirac dans le cadre non relativiste prédit en bonne approximation que les particules élémentaire de spin 1/2 ont un facteur gyromagnétique de 2, et cette prédiction conforme à l'expérience pour l'électron est le plus grand triomphe de l'équation de Dirac. Les déviations de la valeur théorique (déviations conséquentes!) pour le proton et le neutron sont parfaitement expliquées dans le cadre de l'électrodynamique quantique. Ces déviations montrent que la structure du proton et celle du neutron sont plus complexes (composition de sous-particules) qu'une particule ponctuelle de spin 1/2 alors que dans le cas de l'électron, il semblerait qu'il n'y ait pas de sous-structure.

Remarque: Le facteur gyromagnétique est pris parfois comme étant négatif mais ce n'est qu'une question de convention.

C'est par ailleurs le terme d'interaction entre le champ magnétique et l'addition du moment cinétique orbital et intrinsèque de spin:

$$E_p = -\frac{q}{2m_0} (\vec{L} + 2\vec{S}) \circ \vec{B} \quad (43.276)$$

de l'hamiltonien de Pauli qui donne les valeurs mesurées par l'effet Zeeman! Raison pour laquelle cette expression est appelée parfois "**énergie de Zeeman**" et notée parfois sous la forme d'opérateur d'Hamiltonien:

$$\hat{H}_z = -\frac{q}{2m_0} (\vec{L} + 2\vec{S}) \circ \vec{B} \quad (43.277)$$

Nous savons que cette dernière relation peut aussi s'écrire pour toute particule seule (cf. chapitre de Physique Quantique Ondulatoire) plongé dans un champ magnétique colinéaire à l'axe Z et nul dans les autres directions:

$$\begin{aligned}
 E_p &= -(\vec{\mu}_l + g\vec{\mu}_s) \circ \vec{B} = -\frac{q}{2m_0} (\vec{L} + 2\vec{S}) \circ \vec{B} = -\frac{q}{2m_0} (L_z + 2S_z) B_z \\
 &= -\frac{q}{2m_0} (m_l \hbar + 2s \hbar) B = -\frac{q}{2m_0} \hbar (m_l + 2s) B = -\gamma \hbar (m_l + 2s) B = -\mu_B (m_l + 2s) B \quad (43.278)
 \end{aligned}$$

où γ est le "**rapport gyromagnétique**" et μ_B le "**magnéton de Bohr**" (et nous avons vu lors de la démonstration de cette relation que le magnéton de Bohr n'est qu'un cas particulier avec le facteur g de Landé égal à celui de l'électron).

Intéressons-nous maintenant seulement à la plus petite variation d'énergie entre deux états du moment cinétique orbital (puisque le plus petit pas de variation est de 1 de par la quantification de l). Il sera alors toujours à spin égal de la forme:

$$\Delta E_{p,l} = -\gamma \hbar B \quad (43.279)$$

Cette variation d'énergie sera restituée sous forme d'ondes électromagnétiques correspondant à:

$$h\nu = \hbar \frac{\omega}{2\pi} = -\gamma \hbar B = -\gamma \frac{h}{2\pi} B \quad (43.280)$$

d'où:

$$\omega = -\gamma B \quad (43.281)$$

qui est la "**relation de Larmor**" (à ne pas confondre avec le "rayon de Larmor" vu dans le chapitre de Magnétostatique). Mais dans la pratique nous utilisons surtout la relation:

$$\nu = -\gamma \frac{B}{2\pi} \quad (43.282)$$

qui donne ce que nous appelons la "**fréquence de résonance**" du moment cinétique orbital.

De même, en procédant pour le moment cinétique intrinsèque, nous obtenons à moment cinétique orbital constant (lors du transition de spin de $+1/2$ à $-1/2$ ou inversement):

$$\Delta E_{p,s} = -\Delta(\gamma \hbar 2s) B = -\gamma \hbar B \Delta(2s) = -2\gamma \hbar B \Delta s = -2\gamma \hbar B \quad (43.283)$$

et en reposant $g = 2$ il vient:

$$\Delta E_{p,s} = -g \gamma \hbar B \quad (43.284)$$

d'où:

$$\omega = -g \gamma B \quad (43.285)$$

et évidemment:

$$\boxed{\nu = -g \gamma \frac{B}{2\pi}} \quad (43.286)$$

Cette dernière étude de variation d'énergie due à l'application d'un champ magnétique et les fréquences d'émission d'énergie qui s'en suivent est à la base de la résonance magnétique nucléaire (RMN) qui ne marche donc que pour les particules possédant le moment magnétique de spin μ_s (par construction de l'hamiltonien de Pauli!).

La RMN consiste à modifier le moment magnétique nucléaire, autrement dit à faire passer le noyau d'un niveau d'énergie à un autre, en appliquant des champs magnétiques à l'échantillon qu'on veut étudier. Lorsque l'énergie des photons qui constituent ces champs magnétiques correspond à l'énergie de transition d'un niveau d'énergie à l'autre, ces photons peuvent être absorbés par le noyau: nous disons alors qu'il y a "**résonance nucléaire**".

Nous pouvons caractériser l'énergie de transition du moment magnétique de spin nucléaire en donnant la fréquence de l'onde électromagnétique qui permet la résonance. Pour les champs usuels (de l'ordre du tesla), la résonance du proton a lieu dans le domaine des ondes radio (100 [MHz] environ): 42 [MHz] dans un champ de 1.0 [T] et 63 [MHz] dans un champ de 1.5 [T].

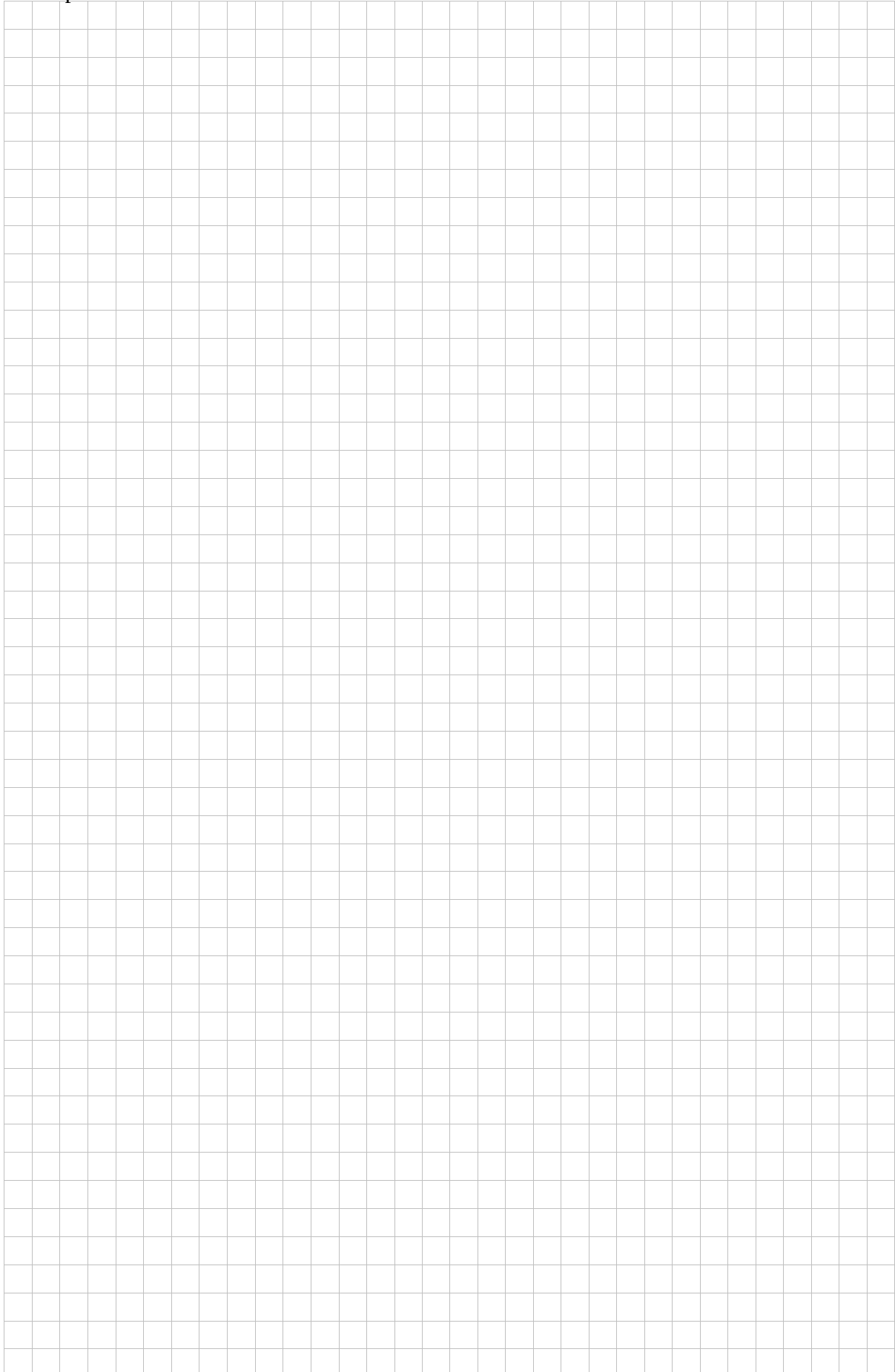
Signalons au passage que dans certains ouvrages, puisque nous avons:

$$\Delta E_{p,s} = -g\gamma\hbar B \quad (43.287)$$

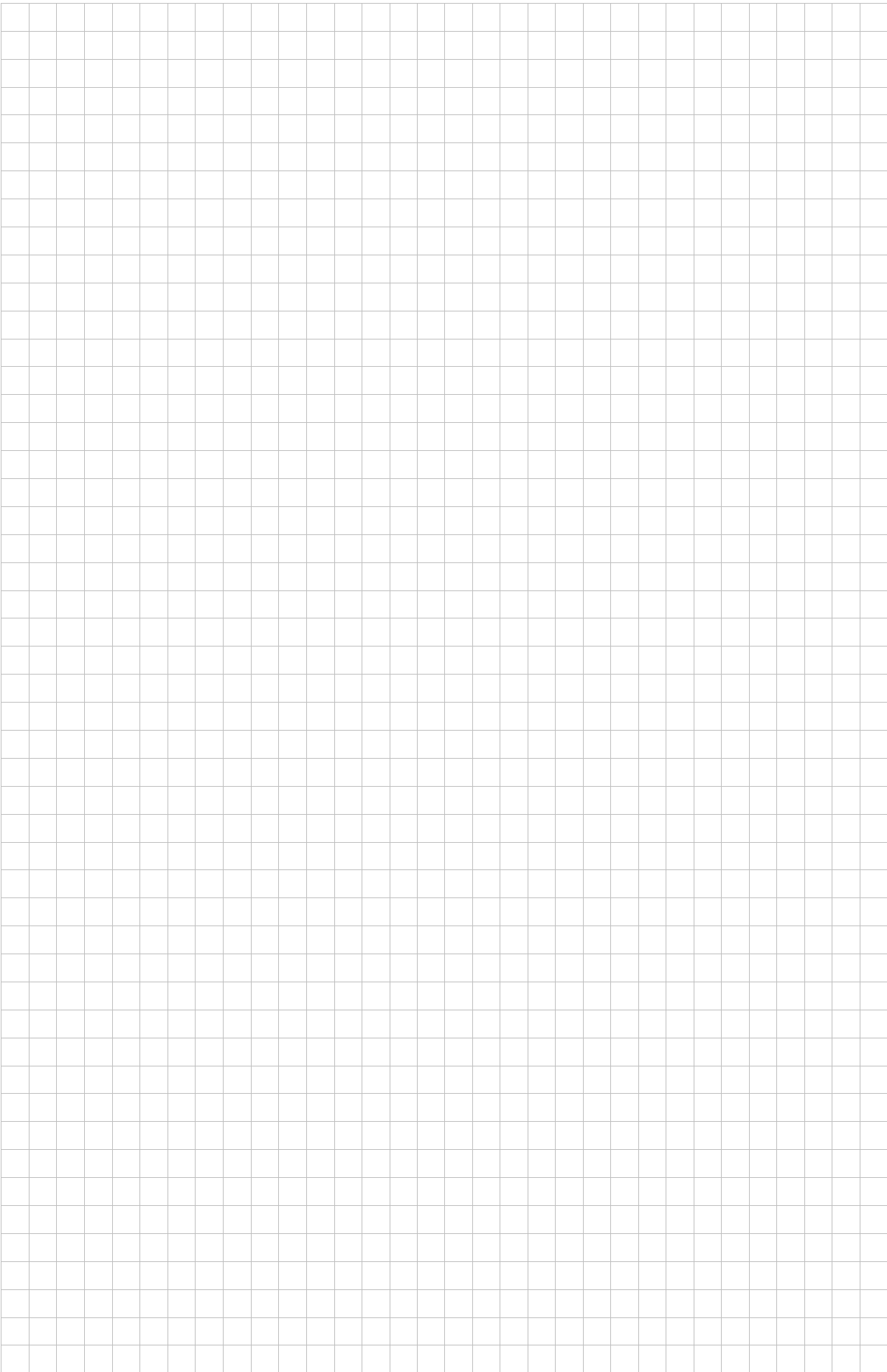
alors pour une particule avec que deux états de spin possible l'énergie de chacun des niveau est alors évidemment la moitié de la différence ci-dessus, ce qui est souvent noté:

$$E_{\pm,s} = \hbar\omega_{\pm} = -\frac{g\gamma\hbar B}{2} = -\frac{g\mu_B B}{2} \quad (43.288)$$

Notes personnelles:









44. PHYSIQUE NUCLÉAIRE

1. L'ARME NUCLÉAIRE

Sans souhaiter faire d'amalgame, nous considérons qu'il est indispensable, à l'époque où l'arme nucléaire sert de moyen de dissuasion et donc d'élément de stabilité planétaire (mais également de menace de destruction...), à la culture générale de l'ingénieur de connaître certaines propriétés élémentaires de la bombe atomique à fission. Nous allons donc exceptionnellement dans ce petit sous-chapitre sans mathématiques (les développements mathématiques de l'arme nucléaire et des centrales nucléaires seront présentés lors de notre étude de la neutronique plus loin) parler un petit peu de cette arme de destruction massive qui fascine souvent les jeunes étudiants.

Certes, nous étudierons plus tard théoriquement comment provoquer une réaction en chaîne divergente dans un volume donné. Cependant, il ne faudra évidemment pas s'attendre à acquérir les connaissances nécessaires à la fabrication d'une telle arme puisque celle-ci ne fait pas appel uniquement à des connaissances de la physique, mais également à de l'électronique, mécanique, chimie, mathématiques, etc.

S'agissant d'une explosion, l'usage s'est immédiatement introduit de comparer l'énergie d'une arme nucléaire à celle d'un explosif courant: le Trinitrotoluène (T.N.T). Ce T.N.T fournit 4'200'000 Joules par Kilo, mais les énergies des armes nucléaires sont telles qu'il est plus parlant de les évaluer en milliers de tonnes - ou kilotonnes de T.N.T (ultérieurement les armes thermonucléaires représentèrent un nouveau bond dans les énergies explosives: l'unité pratique de comparaison est le million de tonnes - Mégatonne de T.N.T).

La fission de 56 grammes d'Uranium 235 ou de Plutonium 239 donne l'équivalent de 1 kilotonne avec la fission de 145^{21} atomes (ce n'est même pas une valeur entière du nombre d'Avogadro!!).

La première explosion nucléaire expérimentale, à Alamogordo le 16 juillet 1945 - fut évaluée à 20 kt, avec une bonne précision car il avait été possible de mettre en place de multiples dispositifs de mesure.

Celles du 6 août, sur Hiroshima (à Uranium 235) puis du 9 août sur Nagasaki (au Plutonium 239) furent d'abord estimées aussi à 20 kt. Ultérieurement, et par étude fine sur les effets de souffle, leurs énergies furent respectivement ramenées à environ 17 et respectivement 15 kt. Cela n'en représentait pas moins l'équivalent d'un chargement en T.N.T. d'un convoi de l'ordre de 6'000 camions de l'US Army.

Voici un schéma à la fois intéressant et persuasif des effets d'une bombe atomique (pour information à partir d'une vitesse de 220 [km/h] un être humain moyen est soulevé du sol):

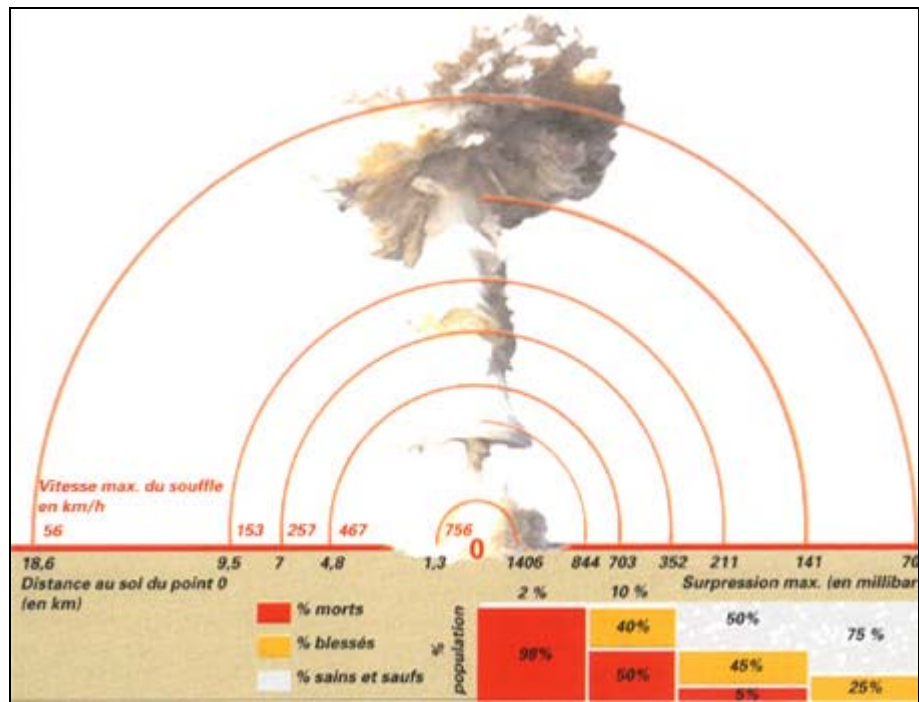


Figure: 44.1 - Effets d'une arme à fission de 1 Mt en fonction de la distance (source: Pour la Science)

Donc en d'autres termes voici en résumé et en approximations les effets d'une arme à fission de 1 Mt explosant à 2'450 mètres d'altitude (tout en sachant qu'aujourd'hui les Américains et les Russes ont des armes nucléaires à fusion dont la puissance de feu dépasserait les 50 mégatonnes...):

Jusqu'à 1.3 [km], tout est rasé, même les bâtiments en béton armé. Jusqu'à 4.8 [km], la plupart des usines et des bâtiments commerciaux sont détruits; les habitations faites de briques et de bois sont anéanties, et leurs débris éparpillés. Jusqu'à 7 [km], les ensembles commerciaux de structure légère et les résidences privées sont démolis. Les constructions plus lourdes sont sérieusement endommagées. Jusqu'à 9.5 [km], les murs des bâtiments légers sont renversés, les résidences privées gravement détériorées. Le souffle (ou surpression) est encore assez puissant pour tuer les personnes qui se trouvent à l'extérieur (explosion des poumons). Jusqu'à 18.6 [km], différents édifices sont endommagés, les rues sont jonchées de débris de vitres et de tuiles. 10 à 20 minutes après la déflagration, les débris aspirés dans la dépression de la tige du champignon atomique, retombent au sol. Suivent 1 à 2 heures après, les débris se situant dans le champignon (sa tête).

La plupart des effets représentés sur le schéma ne sont pas proportionnels à l'énergie. Il n'y a donc pas lieu de faire une simple multiplication pour une arme ayant une puissance multiple de celle utilisée pour le graphique ci-dessus!

Remarque: Pour un petit calcul sympathique sur les bombes nucléaires utilisant l'analyse dimensionnelle le lecteur pourra se référer au chapitre des Principes de la mécanique où nous donnons l'expression de l'énergie d'une bombe en fonction du rayon de sa boule de feu.

Avant de nous attaquer à la partie mathématique, nous tenons à souligner et à rappeler que nous allons nous limiter uniquement aux développements théoriques effectués entre 1890 et environ 1935 (au-delà la complexité des théories nécessite trop de pages pour un site Internet généraliste).

2. RADIOACTIVITÉ

Lorsque nous analysons expérimentalement la radioactivité, nous nous apercevons d'abord que le noyau n'émet pas ses constituants. Ensuite, nous découvrons de nouvelles forces, qui luttent et dominent à tour de rôle. Enfin, de nouvelles particules de matière, et même d'antimatière apparaissent. Le décryptage de ces énigmes a fourni une image cohérente du monde infiniment petit dont la radioactivité a révélé l'existence, un monde où les lois physiques échappent à une intuition issue de la pratique quotidienne de notre monde macroscopique.

D'emblée, la radioactivité a surpris. Dès 1900, il était connu que les rayonnements émis par l'Uranium et ses descendants avaient trois composantes, baptisées: "**alpha**" (α), "**bêta**" (β) et "**gamma**" (γ) séparables par l'action d'un champ magnétique comme représenté symboliquement dans l'image ci-dessous:

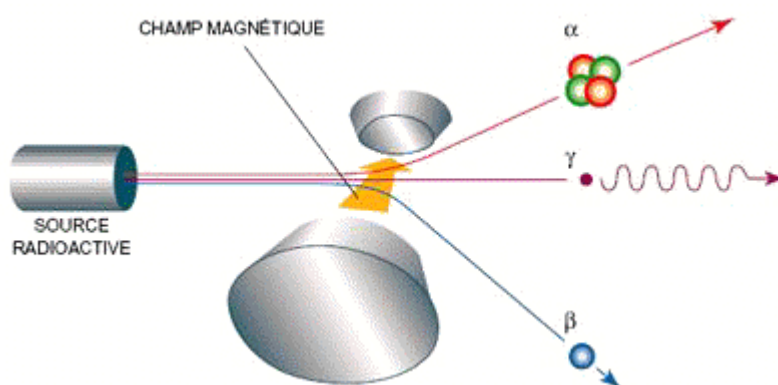


Figure: 44.2 - Séparation des rayonnements à l'aide d'un champ magnétique (source: Pour la Science)

Plus tard, il fut montré que la radioactivité alpha était l'émission de noyaux d'hélium (2 protons et 2 neutrons), la radioactivité bêta l'émission d'électrons. De ces observations, il était logique de déduire que le noyau était constitué de ces trois types de particules (protons, neutrons et électrons), ce qui n'est pas le cas: les constituants du noyau n'ayant été découverts par J. Chadwick qu'en 1932.

Alors, pourquoi les noyaux radioactifs n'émettent-ils pas des protons ou des neutrons? Comment les noyaux éjectent-ils autre chose que leurs constituants? Ces questions doivent être précédées d'une autre, sans doute plus fondamentale pourquoi certains noyaux sont-ils radioactifs? La réponse est la même pour tous les phénomènes physiques spontanés. La pomme tombant de l'arbre, par exemple: c'est parce que le système peut rejoindre un état plus stable en perdant de l'énergie potentielle, l'excédent d'énergie s'échappant sous forme d'énergie cinétique, c'est-à-dire sous la forme de mouvement.

Cette raison explique aussi pourquoi les isotopes n'émettent pas de protons ou neutrons seuls car souvent au niveau de la structure quantique du noyau, il est plus favorable au niveau énergétique d'émettre un petit noyau ou de changer un neutron en proton (l'étude quantique du noyau dépasse le cadre mathématique des sujets traités sur ce site web).

Définitions:

D1. Tout élément chimique (cf. section de chimie) est caractérisé par son nombre de protons Z appelé "**nombre atomique**".

D2. Le "**nombre de masse**" A est par définition le nombre de protons Z sommé du nombre de neutrons N de l'élément chimique donné. Ainsi, ce dernier se trouve entièrement caractérisé si nous connaissons son nom ou son nombre atomique Z et son nombre de neutrons N ou son nombre de masse. Nous notons usuellement n'importe quel élément sous la forme:

$${}^A_Z X_N \quad (44.1)$$

Les éléments chimiques d'une même espèce (même Z) peuvent avoir différents nombres de neutrons N , c'est-à-dire différents nombres de masse A , nous parlons alors "**d'isotopes**" ou de "**nucléides**" (aujourd'hui cela paraît évident comme définition mais il a fallu de nombreuses années de recherche pour arriver à ce concept et on le doit particulièrement à Niels Bohr). Les éléments chimiques qui ont le même A sont appelées des "**isobares**".

Évidemment, l'énergie nucléaire (du noyau) associée à un même élément chimique diffère selon le nombre de masse et il existe nous le verrons un nombre A pour lequel l'énergie est minimale. Les isotopes pour lesquels l'énergie n'est pas minimum pourront, pour certains d'entre eux et de façon spontanée, libérer l'excès d'énergie en se désintégrant.

D3. La propriété qu'ont certains atomes de modifier spontanément la structure de leurs noyaux pour atteindre un niveau d'énergie inférieur, plus fondamental, est appelée "**radioactivité**". Nous parlons alors de "**radio-isotopes**" pour les atomes concernés.

Les propriétés chimiques d'un atome dépendent (cf. [section de chimie](#)) du nombre et de la disposition des électrons dans son nuage. Ainsi tous les isotopes d'un même élément chimique ont les mêmes propriétés chimiques (c'est cette caractéristique chimique qui est à la base de la médecine nucléaire). Ce sont en quelque sorte des atomes "frères". Cependant, la légère différence de masse de leur noyau fait que leurs propriétés physiques se différencient quelque peu.

D4. Enfin, les "**isotones**" sont les isotopes de différents éléments chimiques (différents Z) ayant le même nombre de neutrons N .

Plutôt que de parler "d'éléments chimiques" lorsque l'on fait en réalité implicitement référence au noyau, on préfère utiliser le terme de "**nuclide**".

La petitesse des atomes pose un problème évident de mesure de masse. C'est pourquoi il a été préféré par les physiciens et les chimistes de mettre en place un système de masse atomique qui est un système de nombres proportionnels à la masse réelle des atomes.

Comme il y a une infinité de systèmes de nombres, un a été choisi judicieusement comme référence et c'est le chiffre 12 pour l'isotope 12 du Carbone:

$${}^{12}\text{C} = 12.00000 \text{ uma} \quad (44.2)$$

où "uma" est l'abréviation de "**unités de masse atomique**".

Ceci a pour conséquence intéressante de conférer au proton et au neutron des masses atomiques très voisines de l'unité.

Nous pouvons donc relier le système S.I. (cf. [chapitre Principes](#)) avec le système des unités de masse atomique (uma).

D5. "**L'unité de masse atomique**" est par définition la masse du 1/12 de l'atome de Carbone ^{12}C , nous avons (la masse des électrons est négligée car très faible par rapport à celle des nucléons):

$$1 \text{ uma} = \frac{1}{12 \frac{\text{nucléons}}{\text{at.}}} \cdot \frac{12 \frac{\text{g}}{\text{mole}}}{N_A} = \frac{1}{6.0221367 \cdot 10^{23}} = 1.6605402 \cdot 10^{-24} [\text{g}] \quad (44.3)$$

pour la valeur admise le lecteur devra se reporter à la littérature des normes internationales ad hoc (car celle-ci varie en fonction des nouvelles versions des normes).

Donc la masse du proton en uma vaut:

$$m_{uma,p} = \frac{1.6726231 \cdot 10^{-24} [\text{g}]}{1.6605402 \cdot 10^{-24} [\text{g}]} = 1.007276468 \text{ uma} \quad (44.4)$$

Attention, cependant la masse molaire d'un isotope différent que le ^{12}C ne peut pas être calculée par addition des masses des nucléons (protons et neutrons) qui compose son noyau, car il faut tenir compte du défaut de masse (notion que nous verrons plus loin).

Les masses peuvent être aussi exprimées en unités d'énergie puisqu'il y a équivalence masse-énergie comme nous l'avons vu en relativité restreinte d'après la relation $E = mc^2$ (cf. chapitre de Relativité Restreinte). L'unité d'énergie en physique nucléaire souvent utilisée est "**l'électronvolt**".

D6. Un "**électronvolt**" noté [eV] est l'énergie acquise par une charge élémentaire soumise à une différence de potentiel de 1 [V].

Ainsi, d'après la relation entre l'énergie et le potentiel électrostatique $E = q \cdot U$ (cf. chapitre d'Électrostatique), nous avons:

$$1 [\text{eV}] = 1.60217733 \cdot 10^{-19} [\text{J}] \quad (44.5)$$

Nous en tirons puisque la vitesse de la lumière dans le vide vaut $c = 299792458 [\text{ms}^{-1}]$:

$$\begin{aligned} 1 [\text{g}] &= 8.987551 \cdot 10^{13} [\text{J}] = 5.60959 \cdot 10^{32} [\text{eV}] \\ \Rightarrow 1 \text{ uma} &= 931.49432 [\text{MeV}] \end{aligned} \quad (44.6)$$

Encore une fois, pour la valeur admise le lecteur devra se reporter à la littérature des normes internationales ad hoc (car celle-ci varie en fonction des nouvelles versions des normes).

2.1. DÉSINTÉGRATION

Certains noyaux possèdent donc la propriété de modifier spontanément leur structure interne pour atteindre un niveau d'énergie plus fondamental. Cette transformation s'accompagne de l'émission de particules et/ou de rayonnements électromagnétiques. Le noyau résiduel peut être lui aussi radioactif et subir d'autres transformations par la suite ou être stable.

La désintégration radioactive d'un isotope est un phénomène aléatoire et nous ne pouvons jamais dire à quel moment un noyau va se désintégrer (probabilité sans effet de mémoire).

Remarque: Pour la démonstration de cette affirmation, le lecteur peut se reporter au chapitre de Techniques De Gestion dans la partie traitant de la théorie des files d'attente et en particulier la modélisation des arrivées. Effectivement, le développement est en tout point identique mais seulement l'objet d'étude change (ce ne sont alors plus des appels téléphoniques mais des désintégrations). Ainsi, on y démontre que sous certaines hypothèses le phénomène suit une loi de Poisson et nous y démontrons que celle-ci n'a pas de mémoire.

Nous ne pouvons donner que la probabilité de désintégration par unité de temps. Cette probabilité est donnée par la "**constante radioactive**" et a pour unité l'inverse du temps tel que $[\lambda] = s^{-1}$. Cette constante peut être calculée comme nous l'avons déjà vu lors de l'étude de l'effet tunnel dans le chapitre de Physique Quantique Ondulatoire.

La constante radioactive varie pour tous les isotopes connus:

$$1.57 \cdot 10^{-18} [s^{-1}] \leq \lambda \leq 3 \cdot 10^6 [s^{-1}] \quad (44.7)$$

Soit $N(t)$ le stock d'atomes d'un isotope radioactif au temps t . Le nombre d'atomes se désintégrant durant le temps infinitésimal dt est donc égal à:

$$\lambda \cdot N(t) dt \quad (44.8)$$

conduisant à une diminution du stock égale à:

$$- \lambda N(t) dt \quad (44.9)$$

L'équation différentielle (cf. chapitre de Calcul Intégral Et Différentiel) s'écrit donc:

$$dN = - \lambda N dt \quad (44.10)$$

ou:

$$\frac{dN}{dt} + \lambda N = 0 \quad (44.11)$$

qui a pour solution très simple (cf. chapitre de Calcul Différentiel Et Intégral):

$$N(t) = N_0 e^{-\lambda(t-t_0)} \quad (44.12)$$

avec $N_0 = N(t_0)$ le stock de noyaux au temps t_0 .

Remarque: $N(t)$ ne représente pas le nombre d'atomes restant au temps t mais le nombre le plus probable d'atomes radioactifs restant au temps t !!

Dans la pratique, la mesure de la constante radioactive se fait à l'aide de la décroissance de l'activité (voir plus loin) de l'isotope intéressé.

2.1.1. DEMI-VIE D'ISOTOPE

Définition: La "**période radioactive**" ou de "**demi-vie**" $T_{1/2}$ d'un isotope est le temps moyen qu'il faut attendre pour que 50% du stock de noyaux radioactifs d'un isotope donné soit désintégré:

$$\frac{N(t = T_{1/2})}{N_0} = \frac{1}{2} = e^{-\lambda T_{1/2}} \Rightarrow T_{1/2} = \frac{\ln(2)}{\lambda} \quad (44.13)$$

Nous avons ainsi une relation très importante entre la période de demi-vie et la constante radioactive!

Si le radio-isotope a le choix de se désintégrer selon deux voies de désintégrations distinctes caractérisées par deux périodes radioactives distinctes $T_{1/2}^1$ et $T_{1/2}^2$, la demi-vie de ce nucléide est définie par la moyenne:

$$T_{1/2} = \frac{1}{\frac{1}{T_{1/2}^1} + \frac{1}{T_{1/2}^2}} \quad (44.14)$$

et nous calculons le nombre de nucléides restant par la relation:

$$N(t) = N_0 e^{-(\lambda_1 + \lambda_2)t} \quad (44.15)$$

2.2. ACTIVITÉ RADIOACTIVE

Définition: L'activité A d'une source radioactive est le nombre de désintégrations par unité de temps.

Remarque: Son unité de mesure est le "**Becquerel**" noté $[A] = [Bq]$. 1 Becquerel correspondant donc à une désintégration par seconde.

L'ancienne unité de mesure de la radioactivité était le "**Curie**" $[Ci]$. Le Curie avait été défini dans un premier temps comme l'activité d'environ un gramme de radium, élément naturel que nous retrouvons dans les sols avec l'Uranium. Cette unité est beaucoup plus grande que la précédente car par définition 1 $[Ci]$ correspond à 37 milliards de désintégrations par seconde:

$$1 [Ci] = 3.7 \cdot 10^{10} [Bq] \quad (44.16)$$

L'activité s'obtient par la dérivation temporelle du stock d'atomes d'un échantillon donné:

$$A(t) = -\frac{dN(t)}{dt} = \lambda N_0 e^{-\lambda(t-t_0)} = \lambda N(t) \quad (44.17)$$

La relation dite "**équation d'activité**":

$$A = \lambda N \quad (44.18)$$

montre ainsi que l'activité d'un nombre donné d'atomes N d'un isotope radioactif est proportionnelle à ce nombre et inversement proportionnelle à la demi-vie de l'isotope (de par la relation vue plus haut entre la constante radioactive et la période de demi-vie).

 Exemple:

Un gramme de ^{225}Ra contient (il faut régulièrement se référer aux normes d'usage pour les valeurs des constantes utilisées):

$$\frac{1[\text{g}] \cdot 6.0221267 \cdot 10^{23} [\text{at / mole}]}{226.0254 [\text{g / mole}]} = 2.664 \cdot 10^{21} \text{ atomes} \quad (44.19)$$

donc l'activité de ce gramme vaut connaissant $T_{1/2} = 1'600$ années :

$$A = \lambda N = \frac{\ln(2)}{T_{1/2}} \cdot N = \frac{\ln(2)}{1600 \cdot 365.256 \cdot 24 \cdot 60 \cdot 60} \cdot 2.664 \cdot 10^{21} = 3.66 \cdot 10^{10} [\text{Bq}] \quad (44.20)$$

Par le même raisonnement, nous montrons que l'activité au cours du temps suit la même loi exponentielle que la diminution du nombre de nucléides:

$$A(t) = A_0 e^{-\lambda(t-t_0)} \quad (44.21)$$

avec:

$$A_0 = A(t_0) = \lambda N_0 \quad (44.22)$$

Expérimentalement pour déterminer la période de demi-vie d'un isotope radioactif, nous procédons de la manière suivante:

1. Nous choisissons un échantillon pur d'un isotope dont nous souhaitons déterminer la période de demi-vie.
2. Au temps t_0 nous mesurons à l'aide d'un détecteur pendant un intervalle de temps Δt fixé le nombre de désintégrations. Nous avons alors le nombre de désintégrations pendant un intervalle de temps en début d'expérience (l'unité de la mesure est alors les désintégrations et non pas le nombre de désintégrations par seconde).
3. Ensuite, pendant chaque Δt consécutif (l'intervalle de temps est fixé) nous mesurons le nombre de désintégrations pendant cet intervalle de temps. Cela nous donne donc une série de mesures du nombre de désintégrations observées pour $\Delta t_1, \Delta t_2, \Delta t_3, \dots$
4. À l'ensemble des mesures de désintégrations effectuées, nous enlevons le bruit de fond du laboratoire

Puisque:

$$A(t) \Delta t = A(t_0) \Delta t e^{-\lambda(t-t_0)} \quad (44.23)$$

En prenant le logarithme népérien nous avons:

$$\ln(A(t) \Delta t) = \ln(A(t_0) \Delta t) - \lambda(t - t_0) = -\lambda(t - t_0) + \ln(A(t_0) \Delta t) \quad (44.24)$$

Soit:

$$\ln(A(t)\Delta t) = -\lambda \cdot t + \ln(A(t_0)\Delta t) \quad (44.25)$$

Il s'agit donc de l'équation d'une droite de pente $-\lambda$ et d'ordonnée à l'origine $\ln(A(t_0)\Delta t)$. Ainsi, la constante radioactive est immédiatement mesurée et l'on en déduit rapidement la période de demi-vie à l'aide de la relation démontrée plus haut:

$$T_{1/2} = \frac{\ln(2)}{\lambda} \quad (44.26)$$

2.2.1. DATATION AU CARBONE 14

Certains éléments radioactifs naturels constituent de véritables chronomètres pour remonter dans le temps. Des méthodes de datation ont été mises au point, fondées sur la décroissance progressive de la radioactivité contenue dans les objets ou vestiges étudiés. On peut ainsi remonter jusqu'à des dizaines de milliers d'années dans le passé avec le carbone 14, voire bien d'avantage avec d'autres méthodes telles que la thermoluminescence ou la méthode uranium-thorium. La datation au carbone 14 permet d'aborder l'étude de l'histoire de l'homme et de son environnement pendant la période de 5'000 à 50'000 ans à partir du temps présent.

Le carbone naturel est composé de deux isotopes stables: le ^{12}C (98.892%) et ^{13}C (1.108%). Il n'existe donc pas de ^{14}C dans le carbone naturel. Ce dernier est produit en haute atmosphère par l'action de neutrons cosmiques sur le ^{13}C . Nous parlons alors de "**capture neutronique**" (voir plus loin) ou "activation $^{13}\text{C}(n, \gamma)^{14}\text{C}$ ". Ainsi, continûment du ^{14}C est produit en haute atmosphère et se désintègre naturellement avec une période de 5'700 ans. Nous nous imaginons aisément que la concentration en ^{14}C s'équilibre lorsque le taux de production est égal au taux de disparition suite au processus de désintégration radioactif (sinon quoi il n'y aurait plus que du ^{14}C partout).

Il se forme environ 2.5 atomes de ^{14}C par seconde et par cm^2 de surface Terrestre (ce chiffre est cependant dépendant de beaucoup de facteurs mais en amplitude négligeable sur le très long terme. Vous pouvez trouver des ouvrages entiers traitant du sujet), la contribution positive au nombre d'atome de ^{14}C vaut environ:

$$\dot{N}_+ = 2.5 \cdot 4\pi R^2 = 1.28 \cdot 10^{19} \text{ [atomes/s]} \quad (44.27)$$

R étant le rayon de la Terre.

Ou encore en débit de masse cela représente:

$$\dot{m}_+ = 1.28 \cdot 10^{19} \cdot \frac{14 \text{ [g/mole]}}{6.0221367 \cdot 10^{23} \text{ [atomes/mole]}} = 2.98 \cdot 10^{-4} \text{ [g/s]} \quad (44.28)$$

Le taux de disparition est égal au taux de production radioactif, c'est-à-dire:

$$-\dot{N}_+ = \dot{N}_- = -1.28 \cdot 10^{19} \text{ [atomes/s]} \quad (44.29)$$

car:

$$\frac{dN}{dt} \cong 0 \quad (44.30)$$

Comme le taux de disparition vaut:

$$A = \dot{N} = -\lambda N = \frac{-\ln(2)}{5'700 \cdot 365.256 \cdot 24 \cdot 60 \cdot 60} \cdot N = -1.28 \cdot 10^{19} \text{ [atomes / s]} \quad (44.31)$$

Nous en déduisons qu'il y a $N = 3.32 \cdot 10^{30}$ atomes de ^{14}C . Le lecteur pourra vérifier en divisant par la masse de l'atome de ^{14}C que cela fait une énorme masse présente dans l'atmosphère.

Ce radio-isotope se retrouve sous la forme chimique CO_2 et pénètre par photosynthèse et métabolisme dans le règne végétal et animal. À la mort de la plante ou de l'animal, la teneur en ^{14}C reste figée et commence à décroître par désintégration radioactive au cours des âges.

$$^{14}\text{C} + C_{\text{naturel}} = N = c^{\text{e}} \quad (44.32)$$

La datation n'est donc qu'une simple comparaison entre la concentration en ^{14}C de la matière vivante et de la matière morte. De fait, on détermine les activités spécifiques

$$\left(A_{^{14}\text{C}} \frac{\text{Bq}}{\text{kg}} \right)_{\text{matière morte}} = \left(A_{^{14}\text{C}} \frac{\text{Bq}}{\text{kg}} \right)_{\text{matière vivante}} e^{-\lambda(t-t_0)} \quad (44.33)$$

Les archéologues peuvent ainsi aisément dater ce qu'ils veulent avec une relativement bonne approximation.

2.3. FILIATION RADIOACTIVE

Définition: Une filiation radioactive (dite aussi "série de décroissance radioactive" ou encore "décroissance multiple") est par définition la stabilisation d'un noyau appelé "**noyau mère**" en une succession de désintégrations. Chaque étape est caractérisée par un état intermédiaire correspondant à un radionucléide appelé "**nucléide fille**" de l'élément mère. Nous avons:

$$X_1^* \rightarrow X_2^* \rightarrow X_3^* \rightarrow \dots \rightarrow X_n \quad (44.34)$$

où le petit symbole * désigne un isotope radioactif donné, X_n l'isotope stable de la filiation radioactive de l'élément mère X_1^* (les éléments entre deux étant tous des nucléides instables).

Exemple:

Considérons le problème à 2 corps (nous ne nous intéresserons pas aux cas supérieurs excepté sur demande). Supposons qu'à l'origine des temps, le premier descendant n'existe qu'en quantité négligeable:

Conditions Initiales (C.I.) à $t = 0$:

$$\begin{aligned} N_1(t=0) &= N_0 \\ N_2(t=0) &= 0 \end{aligned} \quad (44.35)$$

La variation de l'élément mère (1) n'est donnée que par une contribution négative, la désintégration de 1.

Nous avons:

$$dN_1 = -\lambda_1 N_1 dt \quad (44.36)$$

avec pour solution tenant compte des conditions initiales:

$$N_1(t) = N_{1_0} e^{-\lambda_1 t} \quad (44.37)$$

La variation de l'élément descendant (2), c'est-à-dire la fille de 1, est donnée par une contribution positive (les atomes de (1) désintégrés) et une négative, la désintégration de 2. On a:

$$\begin{aligned} dN_2 &= +\lambda_1 N_1 dt - \lambda_2 N_2 dt \\ \Rightarrow \frac{dN_2}{dt} + \lambda_2 N_2 &= +\lambda_1 N_1 = \lambda_1 N_{1_0} e^{-\lambda_1 t} \end{aligned} \quad (44.38)$$

il faut donc résoudre cette équation différentielle (cf. chapitre de Calcul Différentiel et Intégral).

Nous avons comme solution homogène (équation caractéristique ECAR):

$$\frac{dN_2}{dt} + \lambda_2 N_2 = 0 \quad (44.39)$$

Nous tirons la solution de l'équation homogène comme:

$$N_2^h(t) = N_{2_0}^h e^{-\lambda_2 t} \quad (44.40)$$

avec la lettre h pour signifier qu'il s'agit de la solution "homogène".

Déterminons maintenant la solution particulière de:

$$\frac{dN_2}{dt} + \lambda_2 N_2 = \lambda_1 N_{1_0} e^{-\lambda_1 t} \quad (44.41)$$

La démarche consiste à poser que:

$$N_2^p(t) = N_{2_0}^p e^{-\lambda_p t} \quad (44.42)$$

avec la lettre p pour "particulière". En substituant nous trouvons:

$$\begin{aligned} -\lambda_p \left(N_{2_0}^p e^{-\lambda_p t} \right) + \lambda_2 \left(N_{2_0}^p e^{-\lambda_p t} \right) &= (\lambda_2 - \lambda_p) N_{2_0}^p e^{-\lambda_p t} = \lambda_1 \left(N_{1_0} e^{-\lambda_1 t} \right) \\ \Rightarrow \lambda_p &= \lambda_1 \end{aligned} \quad (44.43)$$

Car si nous avions $\lambda_p = \lambda_2$ nous aurions une égalité nulle ce qui est absurde et nous avons dès lors:

$$-\lambda_1 N_{2_0}^p + \lambda_2 N_{2_0}^p = \lambda_1 N_{1_0} \quad (44.44)$$

d'où nous tirons que:

$$N_{2_0}^p = \frac{\lambda_1 N_{1_0}}{\lambda_2 - \lambda_1} \quad (44.45)$$

Finalement la solution générale est la somme de la solution homogène et de la particulière, ainsi:

$$N_2(t) = N_{2_0}^h e^{-\lambda_2 t} + \frac{\lambda_1 N_{1_0}}{\lambda_2 - \lambda_1} e^{-\lambda_1 t} \quad (44.46)$$

Appliquons les conditions initiales:

$$t = 0, N_2(t = 0) = 0 \Rightarrow N_{2_0}^h = \frac{-\lambda_1 N_{1_0}}{\lambda_2 - \lambda_1} \quad (44.47)$$

Finalement nous avons:

$$N_2(t) = \frac{\lambda_1 N_{1_0}}{\lambda_2 - \lambda_1} (e^{-\lambda_1 t} - e^{-\lambda_2 t}) \quad (44.48)$$

Nous laisserons le soin au lecteur de tracer les graphiques de:

$$N_1(t) = N_{1_0} e^{-\lambda_1 t} \text{ et } N_2(t) = \frac{\lambda_1 N_{1_0}}{\lambda_2 - \lambda_1} (e^{-\lambda_1 t} - e^{-\lambda_2 t}) \quad (44.49)$$

pour voir l'allure que cela a s'il en ressent le besoin.

$N_2(t)$ étant nul pour $t = 0$ et pour $t = \infty$, obligatoirement il passe, comme l'activité $A_2(t)$, par un maximum. Soit t^* le temps où le maximum est observé, nous avons:

$$\begin{aligned} \left. \frac{dA_2}{dt} \right|_{t=t^*} &= \left. \frac{d(\lambda_2 N_2)}{dt} \right|_{t=t^*} = 0 = \lambda_2 \frac{\lambda_1 N_{1_0}}{\lambda_2 - \lambda_1} (-\lambda_1 e^{-\lambda_1 t^*} + \lambda_2 e^{-\lambda_2 t^*}) \\ \Rightarrow (-\lambda_1 e^{-\lambda_1 t^*} + \lambda_2 e^{-\lambda_2 t^*}) &= 0 \end{aligned} \quad (44.50)$$

d'où:

$$t^* = \frac{\ln\left(\frac{\lambda_1}{\lambda_2}\right)}{\lambda_1 - \lambda_2} \quad (44.51)$$

La connaissance de t^* est importante en particulier en médecine nucléaire où nous désirons administrer le produit 1 à des fins radiodiagnostiques et minimiser les effets néfastes du/des produit(s) fille(s) de 1. Nous choisissons alors des produits tels que le temps t^* soit supérieur au temps d'élimination biologique (voies d'élimination naturelle de l'organisme) de sa fille. Nous y reviendrons d'ici quelques paragraphes après avoir étudié les trois scénarios classiques de la filiation radioactive:

2.3.1. ÉQUILIBRE SÉCULAIRE

Ce type de relation entre activités mère et fille intervient quand la demi-vie du noyau mère est beaucoup plus grande que celle du noyau fille. En d'autres termes, la décroissance du noyau mère est négligeable et l'activité du descendant tend vers celle du parent.

Nous avons alors dans ce cas particulier:

$$\lambda_1 \ll \lambda_2 \quad (44.52)$$

Donc nous avons pour l'activité en utilisant la relation précédemment démontrée:

$$\begin{aligned} \lambda_2 N_2(t) = A_2(t) &= \frac{\lambda_2 \lambda_1 N_{1_0}}{\lambda_2 - \lambda_1} \left(e^{-\lambda_1 t} - e^{-\lambda_2 t} \right) = \frac{\lambda_2 A_{1_0}}{\lambda_2 - \lambda_1} \left(e^{-\lambda_1 t} - e^{-\lambda_2 t} \right) \\ &\underset{\lambda_1 \ll \lambda_2}{=} A_{1_0} \left(e^{-\lambda_1 t} - e^{-\lambda_2 t} \right) \end{aligned} \quad (44.53)$$

Donc:

$$A_2(t) = A_{1_0} \left(e^{-\lambda_1 t} - e^{-\lambda_2 t} \right) \quad (44.54)$$

Nous voyons aussi que dans le cas où $\lambda_1 \ll \lambda_2$ et $t \rightarrow +\infty$, nous avons:

$$A_2(t) = A_{1_0} \left(e^{-\lambda_1 t} - e^{-\lambda_2 t} \right) \cong A_{1_0} e^{-\lambda_1 t} = A_1(t) \quad (44.55)$$

en d'autres termes, les activités mère et fille, deviennent équivalentes après un certain temps suffisamment grand. Par exemple, après un temps d'une demi-vie de l'isotope fille, nous avons déjà l'activité fille qui est égale à 50% de celle de l'activité mère:

$$A_2(T_{1/2}) \cong A_{1_0} e^{-\lambda_1 T_{1/2}} = \frac{1}{2} A_{1_0} \quad (44.56)$$

Si nous avons le cas où l'approximation suivante est acceptable:

$$\lambda_1 \cong 0 \quad (44.57)$$

nous aurons alors:

$$A_2(t) = A_{1_0} \left(1 - e^{-\lambda_2 t} \right)$$

2.3.2. ÉQUILIBRE TRANSITOIRE

Ce type de relation entre activités mère et fille intervient quand la demi-vie du noyau mère est plus grande que celle du noyau fille (mais pas beaucoup beaucoup plus grande contrairement au cas de l'équilibre séculaire). En d'autres termes, la décroissance du noyau mère et l'activité des descendants sont égales à un facteur constant près (en d'autres termes, leurs courbes de décroissance radioactive sont parallèles après un temps suffisamment long).

Nous avons alors dans ce cas particulier:

$$\lambda_1 < \lambda_2 \quad (44.58)$$

Donc nous avons pour l'activité en utilisant la relation précédemment démontrée:

$$\lambda_2 N_2(t) = A_2(t) = \frac{\lambda_2 \lambda_1 N_{1_0}}{\lambda_2 - \lambda_1} \left(e^{-\lambda_1 t} - e^{-\lambda_2 t} \right) = \frac{\lambda_2 A_{1_0}}{\lambda_2 - \lambda_1} \left(e^{-\lambda_1 t} - e^{-\lambda_2 t} \right) \quad (44.59)$$

Après un temps suffisamment long, il vient:

$$A_2(t) \cong \frac{\lambda_2 A_{1_0}}{\lambda_2 - \lambda_1} e^{-\lambda_1 t} = \frac{\lambda_2}{\lambda_2 - \lambda_1} A_1(t) \quad (44.60)$$

où nous voyons que le facteur:

$$\frac{\lambda_2}{\lambda_2 - \lambda_1} \quad (44.61)$$

est supérieure à l'unité. Donc après un temps suffisamment long, non seulement l'activité de l'isotope fille est parallèle à celle de la mère mais en plus elle lui est supérieure.

2.3.3. NON-ÉQUILIBRE

Ici le temps de demi-vie de l'élément-fils est supérieur à celui de l'élément mère. En d'autres termes nous avons l'hypothèse:

$$\lambda_1 > \lambda_2 \quad (44.62)$$

Ceci implique pour rappel que la probabilité de désintégration de l'élément mère (1) est supérieure à celle de l'élément fils (2) ce qui est logique ici.

L'activité de l'isotope fille croît alors dans l'échantillon suivant la relation démontrée juste plus haut:

$$A_2(t) = \frac{\lambda_2 A_{1_0}}{\lambda_2 - \lambda_1} \left(e^{-\lambda_1 t} - e^{-\lambda_2 t} \right) \quad (44.63)$$

Finalement, après un temps suffisamment long, seule l'activité de l'élément fils restera, puisque l'activité de l'élément mère disparaît à un taux plus élevé selon:

$$A_2(t) \cong \frac{\lambda_2 A_{1_0}}{\lambda_2 - \lambda_1} \left(-e^{-\lambda_2 t} \right) = \frac{\lambda_2}{\lambda_1 - \lambda_2} A_{1_0} e^{-\lambda_2 t}$$

Après un temps $t_{\max}$, l'activité de l'élément fils atteindra une valeur maximale pour:

$$\begin{aligned}\frac{dA_2(t)}{dt} = 0 &\cong \frac{d}{dt} \left(\frac{\lambda_2 A_0}{\lambda_2 - \lambda_1} \left(e^{-\lambda_1 t} - e^{-\lambda_2 t} \right) \right) = \frac{\lambda_2 A_0}{\lambda_2 - \lambda_1} \frac{d}{dt} \left(e^{-\lambda_1 t} - e^{-\lambda_2 t} \right) \\ &= \frac{\lambda_2 A_0}{\lambda_2 - \lambda_1} \left(\lambda_2 e^{-\lambda_2 t} - \lambda_1 e^{-\lambda_1 t} \right)\end{aligned}\quad (44.64)$$

Soit:

$$\frac{\lambda_2 A_0}{\lambda_2 - \lambda_1} \left(\lambda_2 e^{-\lambda_2 t} - \lambda_1 e^{-\lambda_1 t} \right) = 0 \quad (44.65)$$

Ce qui se simplifie en:

$$\lambda_1 e^{-\lambda_1 t} = \lambda_2 e^{-\lambda_2 t} \quad (44.66)$$

Il vient alors immédiatement le résultat déjà démontré lors de l'exemple présenté plus haut:

$$t_{\max} = \frac{1}{\lambda_2 - \lambda_1} \ln \left(\frac{\lambda_2}{\lambda_1} \right) \quad (44.67)$$

Enfin, considérons le cas extrême de la situation de non-équilibre qui consiste à considérer le cas où:

$$\lambda_2 = 0 \quad (44.68)$$

En d'autres termes que l'élément fils n'est pas radioactif. Nous tombons alors sur le cas classique:

$$A_2(t) = \frac{\lambda_2 A_0}{\lambda_2 - \lambda_1} \left(e^{-\lambda_1 t} - e^{-\lambda_2 t} \right) = 0 \quad (44.69)$$

2.4. PHÉNOMÈNES RADIOACTIFS

Lorsque nous "pesons" un noyau, nous constatons expérimentalement un fait très important!: sa masse est inférieure à la somme des masses de ses constituants. Cette différence est appelée le "**déficit de masse**" et est relativement bien déterminée avec des modèles théoriques simplificateurs.

Le déficit de masse est alors donné par définition:

$$\Delta m = Z \cdot m_p + (A - Z) \cdot m_n - m_{\text{noyau}} \quad (44.70)$$

avec m_{noyau} étant la masse du noyau dans son état fondamental, m_p la masse du proton et m_n la masse du neutron.

La masse d'un ensemble de nucléons liés est inférieure à la somme des masses des nucléons isolés (suffisamment éloignés en tout cas pour ne pas interagir). Nous tirons de la relativité restreinte (voir chapitre du même nom) que:

$$\Delta m = \frac{1}{c^2} E_{\text{liaison}}^{\text{nucléons}} \quad (44.71)$$

où $E_{\text{liaison}}^{\text{nucléons}}$ est l'énergie de liaisons des nucléons composants le noyau (>0).

Δm est donc positif pour tous les éléments (émission d'énergie et donc de masse vers le système extérieur). Si tel n'était pas le cas, les nucléons n'auraient aucune raison de se mettre ensemble afin de former naturellement des noyaux stables (ou plus stables...).

Soit $\bar{B} = f(A, Z)$ l'énergie moyenne par nucléon d'un atome donné. Nous avons:

$$\bar{B} = \frac{[Z \cdot m_p + N \cdot m_n - m_{\text{noyau}}(A, Z)] \cdot c^2}{A} \quad (44.72)$$

qui est donc par convention une valeur positive!

Remarquons que la masse du noyau est reliée à la masse de l'atome par:

$$m_{\text{atome}}(Z, A) = m_{\text{noyau}}(Z, A) + Z \cdot m_e \quad (44.73)$$

De même, la masse du noyau ajoutée à la masse de ses électrons isolés est supérieure à celle du noyau entouré de son cortège électronique. Notons que l'énergie de liaison électronique peut être souvent négligée à celle d'origine nucléaire et c'est une règle que nous adopterons tout au long de ce chapitre.

Cette énergie dégagée lors de la fusion, c'est-à-dire lors de la constitution de l'atome à partir de ses constituants, s'appelle aussi "**énergie de liaison**" (appellation qui pose souvent des problèmes d'interprétations aux jeunes étudiants) car c'est elle qu'il faut fournir si nous voulons, en sens inverse, séparer les éléments. Il ne faut jamais oublier que derrière le terme "énergie de liaison", il y a donc la variation d'énergie entre les éléments isolés et les éléments combinés d'un élément atomique.

L'expression générale pratique de l'énergie moyenne par nucléon d'un atome donné exprimée en unités de masse atomique est alors:

$$\bar{B} = \frac{[Z \cdot 1.00727647 + N \cdot 1.008664904 - m_{\text{noyau}}(A, Z)]}{A} \cdot 931.49432 [MeV] \quad (44.74)$$

Les principes de production d'énergie nucléaire de la fission ou de la fusion résultent de la forme de l'énergie moyenne par nucléon en fonction de A.

Nous avons en réalité la courbe suivante reliant l'énergie moyenne par nucléon (c'est-à-dire la variation d'énergie moyenne entre le nucléon seul et accompagné...) et le nombre de nucléons appelée "**courbe d'Aston**":

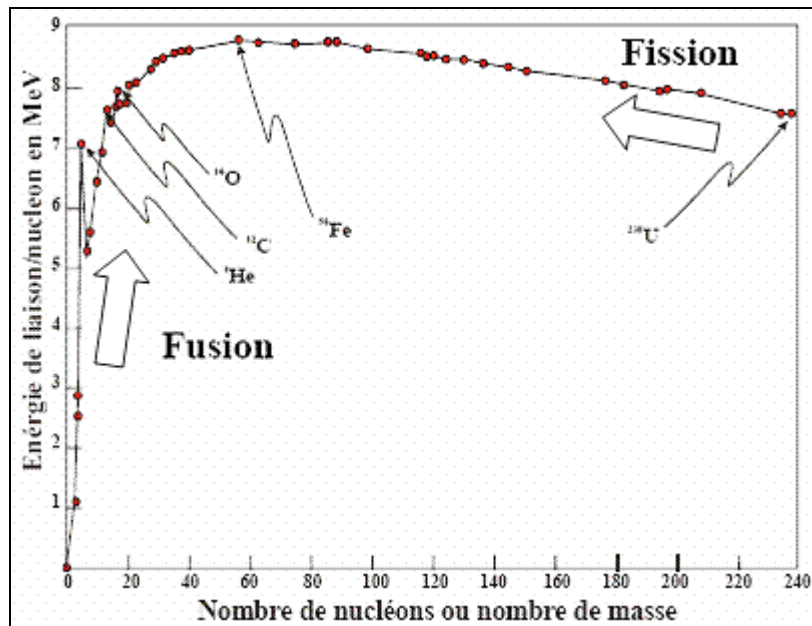


Figure: 44.3 - Courbe d'Aston

où nous voyons qu'à partir du Fer (élément qui est donc le plus "collé" et le plus stable en termes énergétiques car ayant la plus forte énergie de liaison moyenne) l'énergie moyenne diminue à nouveau. Cette diminution étant due au fait qu'à partir d'environ 70 nucléons il semblerait que la force électrostatique à l'intérieur du noyau commence à prendre le dessus sur une autre force qui règne dans les noyaux à très petite échelle (cette force sera nommée plus tard la "force forte" ou "interaction forte").

Au fait, ce qui est vraiment très important de remarquer dans le graphique ci-dessus c'est qu'il y a un point de flexion et que c'est celui-ci qui permet d'obtenir de l'énergie aussi bien avec la fusion, qu'avec la fission nucléaire! Nous voyons également que la variation est beaucoup plus grande sur la gauche que sur la droite, d'où le fait que la fusion libère des énergies beaucoup plus considérables.

Des phénomènes de radioactivité, nous en distinguons 8 dont certains sont qualifiés de "secondaires" car n'étant que les effets secondaires possibles des 6 premiers. Certains de ces phénomènes sont provoqués par l'homme, d'autres sont naturels et les autres sont purement probabilistes.

Voici un diagramme représentant en haut la "**vallée de stabilité**" des atomes et isotopes et en bas la même vallée mais mettant en évidence le type de désintégration:

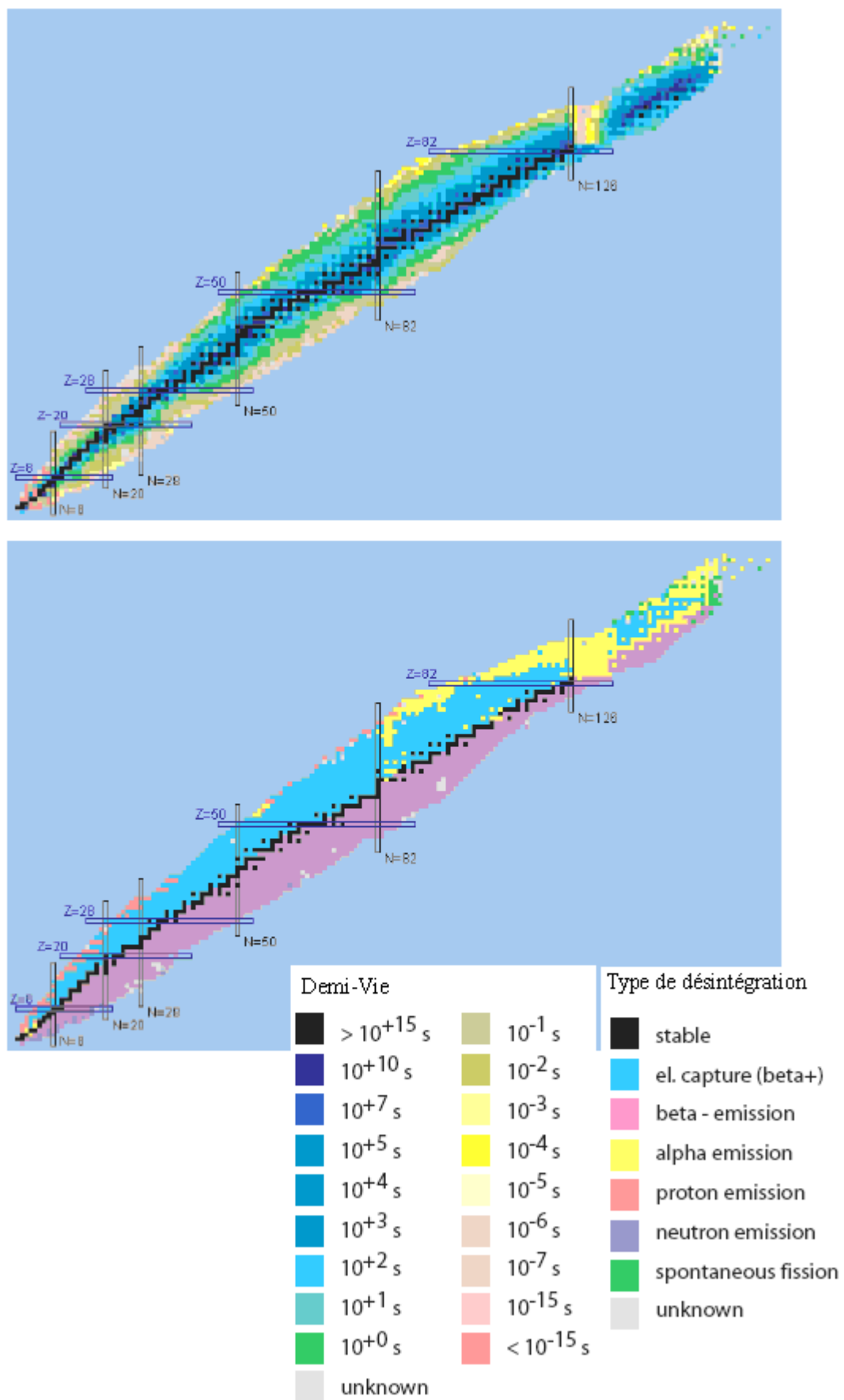


Figure: 44.4 - Vallée de stabilité

Voyons donc les types de désintégrations ou modifications de la structure de l'atome/noyau qui sont possibles dans les détails:

2.4.1. FUSION NUCLÉAIRE (1)

Si nous assemblons deux noyaux légers ${}^A_Z X$ et ${}^A_Y Y$ ($A_X \leq A_Y$) pour former un atome "lourd" ${}^A_Z Z$, alors conformément à la partie gauche de la courbe d'Aston vue plus haut, nous augmentons le défaut de masse puisque l'énergie moyenne par nucléon augmente. En effet:

- l'énergie de X vaut:

$$A_X \cdot \bar{B}(A_X, Z_X) = A_X \bar{B}_X \quad (44.75)$$

- l'énergie de Y vaut:

$$A_Y \cdot \bar{B}(A_Y, Z_Y) = A_Y \bar{B}_Y \quad (44.76)$$

- l'énergie de Z vaut:

$$A_Z \cdot \bar{B}(A_Z, Z_Z) = A_Z \bar{B}_Z \quad (44.77)$$

Comme:

$$\bar{B}_X \leq \bar{B}_Y < \bar{B}_Z \quad (44.78)$$

alors:

$$\boxed{\Delta E = A_Z \bar{B}_Z - (A_X \bar{B}_X + A_Y \bar{B}_Y)} \quad (44.79)$$

est strictement positive.

La fusion nucléaire est quasi exclusivement provoquée par l'homme (sur Terre en tout cas... car les étoiles la font toutes seules avec l'aide de la gravité). La probabilité d'observer une fusion nucléaire naturelle dans des conditions normales de température et de pression est tellement faible qu'il est inutile d'en parler du moins à ce jour...

2.4.2. FISSION NUCLÉAIRE (2)

De même, si nous cassons avec des moyens adéquats (souvent avec des neutrons car pour s'approcher du noyau et vaincre sa répulsion électrostatique c'est le moyen adéquat... c'est celui qu'utilisent les centrales nucléaires et les bombes nucléaires) un atome ${}^A_Z X$ lourd en deux atomes légers ${}^A_Y Y$ et ${}^A_Z Z$ ($A_Y \leq A_Z$) nous augmentons aussi le défaut de masse et l'énergie gagnée vaut:

$$\boxed{\Delta E = (A_Y \bar{B}_Y + A_Z \bar{B}_Z) - A_X \bar{B}_X} \quad (44.80)$$

Que ce soit dans le cas de la fission ou de la fusion, l'énergie dégagée se répartit alors en énergie cinétique des produits de fission, des neutrons émis et enfin des divers rayonnements.

Remarque: Un atome est dit "**fissile**" quand il faut des neutrons rapides pour produire la fission et "**fissile**" quand il suffit d'avoir des neutrons lents pour la fission (ce qui est plus rare).

L'énergie nucléaire est de loin une forme d'énergie beaucoup plus concentrée, puisque 1 kilogramme d'uranium naturel fournit une quantité de chaleur de 100'000 [kWh] dans une centrale électrique courante, alors que 1 kilogramme de charbon fournit en brûlant 8 [kWh]. C'est pourquoi on ne manipule que d'assez faibles masses de combustible nucléaire pour la production d'électricité: une centrale électronucléaire d'une puissance de 1'000 [MW] électrique consomme par an 27 tonnes d'uranium enrichi, le quart de son chargement, alors qu'une centrale thermique de même puissance consomme par an 1'500'000 tonnes de pétrole. Pour comparaison dans le soleil, 1 kilogramme d'hydrogène produit, par réactions nucléaires le transformant en hélium, 180 millions de kWh! Mais attention, industriellement nous ne savons extraire qu'une faible part de l'énergie nucléaire emmagasinée dans la matière. Sur les 27 tonnes d'uranium enrichi consommé en une année par une centrale, seule une petite quantité de noyau a été réellement consommée (d'où la nécessité économique de retraiter l'uranium après utilisation).

Nous nous rendons vite compte que le pouvoir calorifique de la fission est gigantesque par rapport à celui des énergies fossiles. Une estimation donne un rapport d'énergie dégagée par atome de 50'000 millions !!! Par contre le rapport de risque en termes d'exploitation et de gestion des déchets et d'impact sur l'environnement est de l'ordre du même facteur mais dans le sens inverse... C'est pour cette raison qu'il faut que l'humanité trouve soit une autre manière de produire de l'électricité, soit change ses habitudes de consommation.

Nous trouvons pour information en Suisse, rien que 5 centrales nucléaires (au début du 21ème siècle) pour une population de ~6 millions d'habitants (figure ci-dessous):

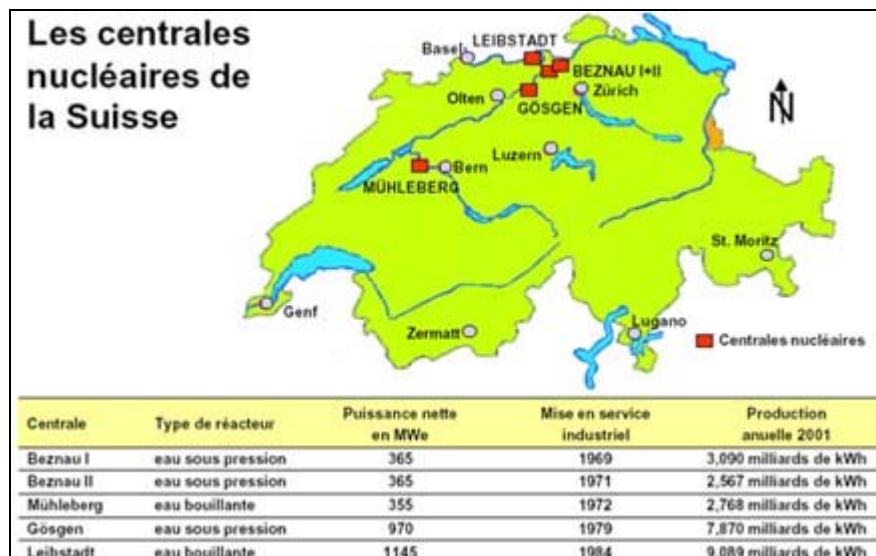
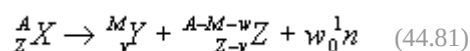


Figure: 44.5 - Centrales nucléaires en Suisse à la fin du 20ème siècle

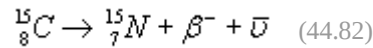
Cette petite introduction étant faite, revenons à la partie purement mathématique:

Dans le cas de la fission spontanée (ou naturelle) nous avons émission de deux produits de fission et de w neutrons. Ce que nous notons:



 **Exemple:**

L'isotope du Carbone 15 par fission spontanée donne un isotope de l'Azote avec émission d'un électron et d'un antineutrino:

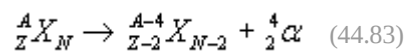


2.4.3. DÉSINTÉGRATION ALPHA (3)

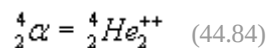
Définition: Lorsqu'un noyau lourd contient trop de protons et de neutrons (comme l'Uranium 238 par exemple), il va vider son trop-plein de nucléons en émettant une particule alpha (noyau d'hélium composé de 2 protons et deux neutrons) et le système final qui sera un nouveau noyau aura une masse plus faible et éventuellement stable. Ce mode de désintégration est la "**radioactivité alpha**".

La probabilité de désintégration est gouvernée par le mécanisme de barrière de pénétration (effet Tunnel) comme nous allons le démontrer un peu plus loin après une petite introduction.

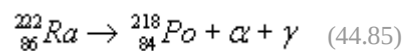
La décroissance radioactive selon la radioactivité alpha, peut être schématisée comme (avec quelques petites variations selon les cas):



où:



 **Exemple:**



L'énergie dégagée lors de la transmutation se calcule au moyen du défaut de masse:

$$Q_\alpha = (M_X - M_Y - M_{\text{He}}) \cdot c^2 \quad (44.86)$$

avec M_X étant la masse du noyau initial, M_Y la masse du noyau final et M_{He} la masse du noyau d'Hélium.

En négligeant l'énergie de liaison des électrons, nous avons:

$$m_X = M_X + Zm_e \text{ et } m_Y = M_Y + (Z-2)m_e \text{ et } m_{\text{He}} = m_{\text{He}} + 2m_e \quad (44.87)$$

Finalement:

$$Q_\alpha \cong (m_X - m_Y - m_{\text{He}}) \cdot c^2 \quad (44.88)$$

Cette expression montre que l'énergie des particules α est bien définie pour des noyaux initiaux et finaux donnés. De fait, nous observons en réalité un spectre énergétique discret. Nous en concluons que ces émissions mènent le noyau à des niveaux d'énergies intermédiaires correspondants à des états excités du noyau final. Nous pouvons expliquer ces observations par une structure nucléaire en couches. La désexcitation du final se faisant par émission de photons γ .

La conservation de l'énergie impose que l'énergie de la désintégration α se répartit entre l'énergie cinétique des deux produits résiduels.

$$Q_\alpha = E_{\text{cin}}^{\text{noyau}} + E_{\text{cin}}^\alpha = \frac{1}{2} m_{\text{noyau}} v_{\text{noyau}}^2 + \frac{1}{2} m_\alpha v_\alpha^2 \quad (44.89)$$

La conservation de la quantité de mouvement nous donne:

$$p_{\text{noyau}} = m_{\text{noyau}} \cdot v_{\text{noyau}} = p_\alpha = m_\alpha v_\alpha \quad (44.90)$$

et donc:

$$v_{\text{noyau}} = \frac{m_\alpha}{m_{\text{noyau}}} v_\alpha \quad (44.91)$$

que nous remplaçons dans l'équation de conservation de l'énergie:

$$\begin{aligned} Q_\alpha &= \frac{1}{2} m_{\text{noyau}} \left(\frac{m_\alpha}{m_{\text{noyau}}} v_\alpha \right)^2 + \frac{1}{2} m_\alpha v_\alpha^2 = \frac{m_\alpha}{m_{\text{noyau}}} \left(\frac{1}{2} m_\alpha v_\alpha^2 \right) + \frac{1}{2} m_\alpha v_\alpha^2 \\ &= \frac{1}{2} m_\alpha v_\alpha^2 \left(1 + \frac{m_\alpha}{m_{\text{noyau}}} \right) = E_{\text{cin}}^\alpha \left(1 + \frac{m_\alpha}{m_{\text{noyau}}} \right) \end{aligned} \quad (44.92)$$

et on en tire que l'énergie de la particule α vaut:

$$E_{\text{cin}}^\alpha = \frac{Q_\alpha}{\left(1 + \frac{m_\alpha}{m_{\text{noyau}}} \right)} = \frac{Q_\alpha}{\left(\frac{m_{\text{noyau}} + m_\alpha}{m_{\text{noyau}}} \right)} \cong \frac{A}{A+4} Q_\alpha \quad (44.93)$$

vu que les masses du noyau et de la particule α sont environ proportionnelles à leur nombre de masse, soit A et 4 respectivement.

Voyons les détails du mécanisme de la désintégration α avec une approche scolaire, simplifiée à l'extrême et donc approximative (mais suffisante quand même). Pour cette approche, nous allons utiliser les développements sur l'effet tunnel que nous avons effectué dans le chapitre de Physique Quantique Ondulatoire.

Pour des noyaux ayant un nombre de nucléons devenant trop important, la répulsion coulombienne entre protons prend des valeurs significatives par rapport à l'interaction forte qui assure la cohésion du noyau. On assiste alors au phénomène de saturation, qui donne lieu à la désintégration α qui est un cas particulier d'une fission spontanée.

Gamow a proposé une explication théorique à ce phénomène en 1928. Il suppose que la particule α préexiste dans le noyau et cogne sur les parois. Elle a alors une probabilité non nulle de franchir la barrière de potentiel du noyau par effet tunnel.

Si par la pensée nous débranchons les interactions coulombiennes, une telle particule α est liée au reste du noyau par un potentiel nucléaire de courte portée r_1 et de profondeur correspondant à une énergie potentielle que nous allons déterminer.

Schématiquement dans le cas de l'Uranium 238 la situation est considérée comme la suivante:

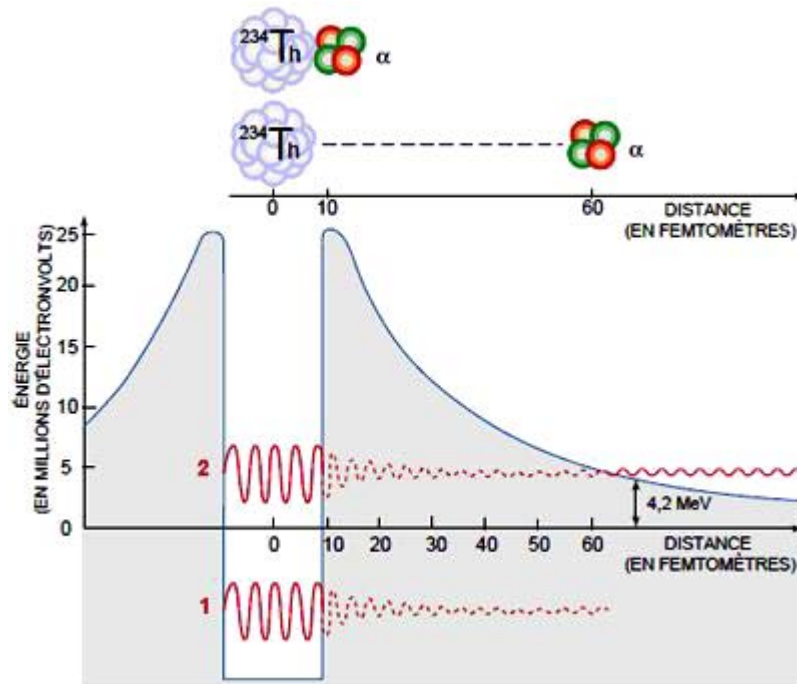


Figure: 44.6 - Représentation de l'idée de Gamow (source: Pour la Science)

En physique classique on représenterait l'émission α comme la fuite du noyau à partir du noyau. Cette représentation n'est pas valable, car elle implique que la particule α , subissant la répulsion électrostatique du noyau résiduel de Thorium 234 ne s'en éloignerait que moyennant une énergie d'environ 25 [MeV]. Or on ne retrouve la faible valeur observée expérimentalement (de seulement 4.2 [MeV]) qu'en faisant appel à la physique quantique.

Bon passons à la partie mathématique:

Branchons la répulsion coulombienne entre la particule α de charge $+2e$ (deux protons et deux neutrons) et le reste du noyau, alors de charge $+(Z-2)e$ à l'extérieur du puits de potentiel nucléaire.

Nous obtenons alors l'expression de l'énergie potentielle (cf. chapitre d'Électrostatique):

$$E_{p\alpha} = \frac{1}{4\pi\epsilon_0} \frac{(Z-2)e \cdot 2e}{r} = \frac{1}{4\pi\epsilon_0} \frac{2(Z-2)e^2}{r} \quad (44.94)$$

où r est la distance entre le centre du noyau et la particule α . L'énergie potentielle diminue donc avec la distance puisque la force est répulsive.

Maintenant, ayons une approche qualitative du phénomène. Nous allons maintenant noter la probabilité T de passage comme étant proportionnelle, selon nos résultats dans le chapitre de Physique Quantique Ondulatoire, à:

$$T \propto e^{-2kL} = \exp\left(-2L \frac{\sqrt{2m(E_{pot} - E_{tot})}}{\hbar}\right) \quad (44.95)$$

en sachant qu'il s'agit suite à nos approximations d'une borne inférieure indicative.

Si nous modélisons la barrière de potentiel du noyau par un profil non rectangulaire tel que présenté ci-dessous:

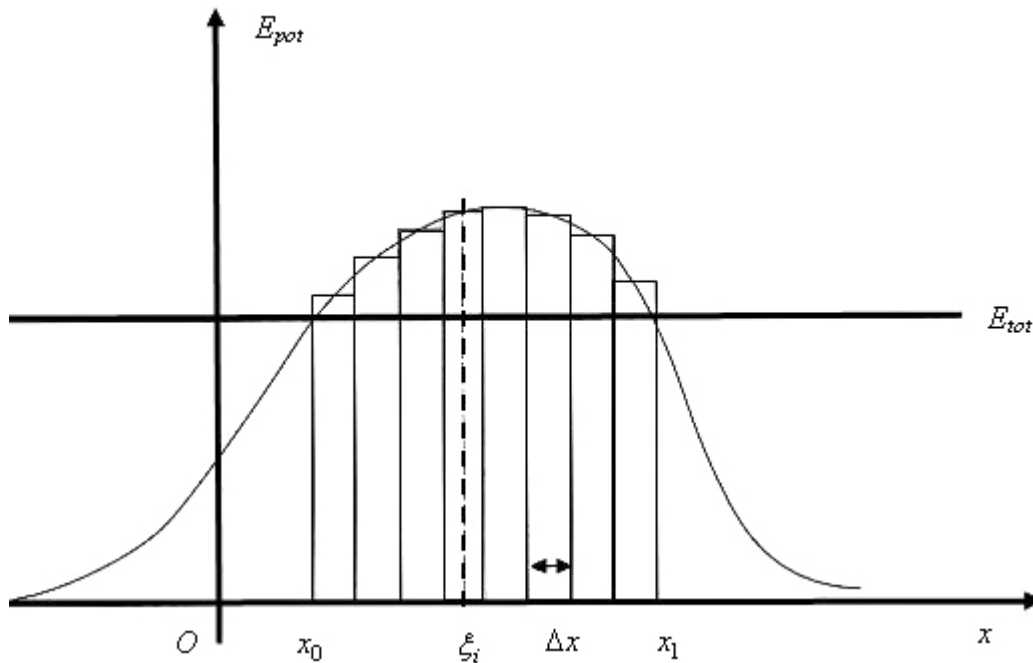


Figure: 44.7 - Profil de barrière non rectangulaire du noyau

où nous remplaçons le profil réel de la courbe par une série de barrières d'épaisseur Δx et où le potentiel est égal à $E_{pot}(\xi_i)$ au point ξ_i .

La probabilité de passer une barrière sera donc proportionnelle à:

$$T_i \propto \exp\left(-\frac{2}{\hbar} \Delta x \sqrt{2m(E_{pot}(\xi_i) - E_{tot})}\right) \quad (44.96)$$

et nous savons (cf. chapitre de Probabilités) que la probabilité de passer une des barrières est un événement indépendant (mutuellement exclusifs). Nous pouvons donc multiplier les probabilités tel que:

$$T \propto \exp\left(\sum_i -\frac{2}{\hbar} \Delta x \sqrt{2m(E_{pot}(\xi_i) - E_{tot})}\right) \quad (44.97)$$

et en passant à la limite il vient:

$$T \propto \exp\left(-\frac{2}{\hbar} \int_{x_0}^{x_1} \sqrt{2m(E_{pot}(x) - E_{tot})} dx\right) \quad (44.98)$$

et si x est assimilé à un rayon d'une configuration à symétrie sphérique:

$$T \propto \exp\left(-\frac{2}{\hbar} \int \sqrt{2m(E_{pot}(r) - E_{tot})} dr\right) \quad (44.99)$$

Dans le cas d'un noyau α , la barrière de potentiel va de r_1 où elle commence jusqu'à r_2 valeur où la barrière est considérée comme négligeable.

Or, l'énergie potentielle du noyau α en tout point distant r de l'extérieur du bord du noyau de l'atome radioactif sera égal, comme nous l'avons vu un peu plus haut à:

$$E_{p\alpha}(r) = \frac{1}{4\pi\epsilon_0} \frac{2(Z-2)e^2}{r} \quad (44.100)$$

Nous avons donc pour $r_1 \leq r \leq r_2$:

$$E_{p\alpha}(r) - E_{tot} = \frac{1}{4\pi\epsilon_0} \frac{2(Z-2)e^2}{r} - E_{tot} \quad (44.101)$$

Pour déterminer E_{tot} du noyau α émis, il faut savoir que son énergie totale est supposée conservée dans ce modèle simplifié. Elle est donc la même avant son passage dans la barrière de potentiel nucléaire lorsque $r_1 \leq r$, pendant, et après r_2 .

De plus, dans ce modèle, l'énergie cinétique aussi est supposée constante lorsque $r \leq r_2$. Autrement dit, puisque le noyau α préexiste dans le noyau de l'atome radioactif, il a déjà la vitesse finale qu'il aura lors du point de franchissement de la barrière du potentiel nucléaire...

Donc sous toutes ces hypothèses très simplificatrices... si nous savons déterminer l'énergie totale du noyau α en r_2 (par exemple), à la sortie de la barrière, nous avons son énergie totale lors de l'ensemble du phénomène de traversée de la barrière.

Réciproquement, son énergie totale nécessaire pour sortir en r_2 de la barrière de potentiel par effet tunnel en partant du noyau (et partir ensuite loin à l'infini et gagner en énergie cinétique et perdre toute son énergie potentielle coulombienne) est la même par hypothèse que l'énergie totale obtenue en calculant le travail de la force qui d'une distance infinie du noyau de l'atome radioactif ramènerait le noyau α à la vitesse précitée au point de sortie minimal r_2 (rayon minimal de sortie pris comme constant car très éloigné en ordres de grandeur par rapport au noyau de l'atome radioactif).

Ce qui correspond alors à la différence d'énergie potentielle entre un point à l'infini et r_2 . Et comme l'énergie potentielle est nulle à l'infini pour un système répulsif, il ne reste plus que le terme:

$$E_{tot} = \frac{1}{4\pi\epsilon_0} \frac{2(Z-2)e^2}{r_2} \quad (44.102)$$

Et finalement:

$$E_{p\alpha}(r) - E_{tot} = \frac{1}{4\pi\epsilon_0} \frac{2(Z-2)e^2}{r} - \frac{1}{4\pi\epsilon_0} \frac{2(Z-2)e^2}{r_2} = \frac{2(Z-2)e^2}{4\pi\epsilon_0} \left(\frac{1}{r} - \frac{1}{r_2} \right) \quad (44.103)$$

valable toujours que pour $r_1 \leq r \leq r_2$ (c'est comme si pendant la traversée de la barrière, le noyau α restituait de l'énergie cinétique au vide au fur et à mesure de son approche du point r_2 , ceci dit, en physique quantique on ne peut pas utiliser l'interprétation de la mécanique classique).

Or, très souvent dans les laboratoires, r_2 est exprimé comme une constante suffisamment loin du noyau de l'atome radioactif. Il est alors relativement naturel (même si c'est du bricolage) de prendre r comme variable d'intégration tel que:

$$T \propto \exp \left(-\frac{2}{\hbar} \int_{r_1}^{r_2} \sqrt{\frac{2m2(Z-2)e^2}{4\pi\epsilon_0} \left(\frac{1}{r} - \frac{1}{r_2} \right)} dr \right) = \exp \left(-2 \frac{1}{\hbar} \sqrt{\frac{m4(Z-2)e^2}{4\pi\epsilon_0}} \int_{r_1}^{r_2} \sqrt{\left(\frac{1}{r} - \frac{1}{r_2} \right)} dr \right) \quad (44.104)$$

et il est de tradition de prendre ensuite:

$$\gamma = \frac{1}{\hbar} \sqrt{\frac{m4(Z-2)e^2 r_2}{4\pi\epsilon_0}} \quad (44.105)$$

ce qui nous amène à:

$$T \propto \exp \left(-\frac{2\gamma}{\sqrt{r_2}} \int_{r_1}^{r_2} \sqrt{\left(\frac{1}{r} - \frac{1}{r_2} \right)} dr \right) \quad (44.106)$$

Faisons maintenant le changement de variables (la dérivation du $\cos^2 u$ est détaillée dans le chapitre de Calcul Différentiel Et Intégral):

$$\cos^2 u = \frac{r}{r_2} \Rightarrow \frac{dr}{du} = r_2 \frac{d \cos^2 u}{du} = -r_2 2 \cos(u) \sin(u) \quad (44.107)$$

d'où:

$$dr = -r_2 2 \cos(u) \sin(u) du \quad (44.108)$$

et en notant:

$$\cos^2 u_0 = \frac{r_1}{r_2} \quad (44.109)$$

L'intégrale:

$$\int_{r_1}^{r_2} \sqrt{\left(\frac{1}{r} - \frac{1}{r_2} \right)} dr \quad (44.110)$$

devient:

$$\begin{aligned}
\int_{r_1}^{r_2} \sqrt{\left(\frac{1}{r} - \frac{1}{r_2}\right)} dr &= \int_{r_1}^{r_2} \sqrt{\frac{1}{r} \left(1 - \frac{r}{r_2}\right)} dr = \int_{r_1}^{r_2} \sqrt{\frac{1}{r} (1 - \cos^2 u)} dr = \int_{r_1}^{r_2} \sqrt{\frac{1}{r} \sin^2 u} dr \\
&= \int_{r_1}^{r_2} \frac{1}{\sqrt{r}} \sin u dr = \int_{r_1}^{r_2} \frac{1}{\sqrt{r_2 \cos^2 u}} \sin u dr = \frac{1}{\sqrt{r_2}} \int_{r_1}^{r_2} \frac{\sin u}{\cos u} dr = \frac{1}{\sqrt{r_2}} \int_{r_1}^{r_2} \frac{\sin u}{\cos u} (-r_2 2 \cos(u) \sin(u) du) \quad (44.111) \\
&= -\frac{2r_2}{\sqrt{r_2}} \int_{r_1}^{r_2} \sin^2(u) du = -2\sqrt{r_2} \int_{r_1}^{r_2} \sin^2(u) du
\end{aligned}$$

Concernant les bornes nous avons pour rappel:

$$\cos^2 u = \frac{r}{r_2}, \quad \cos^2 u_0 = \frac{r_1}{r_2} \quad (44.112)$$

Donc si r vaut r_1 nous écrivons la borne comme étant u_0 et si r vaut r_2 alors:

$$\cos^2 u = \frac{r_2}{r_2} = 1 \Rightarrow u = 0 \quad (44.113)$$

Il vient alors:

$$\int_{r_1}^{r_2} \sqrt{\left(\frac{1}{r} - \frac{1}{r_2}\right)} dr = -2\sqrt{r_2} \int_{u_0}^0 \sin^2(u) du = 2\sqrt{r_2} \int_0^{u_0} \sin^2(u) du \quad (44.114)$$

Nous avons vu dans le chapitre de Calcul Différentiel Et Intégral:

$$I_n = \int \sin^n x dx = \frac{1}{n} (-\cos x \sin^{n-1} x + (n-1)I_{n-2}) \quad (44.115)$$

Donc:

$$\begin{aligned}
I_2 &= \int \sin^2 x dx = \frac{1}{2} (-\cos x \sin x + I_0) \\
I_0 &= \int \sin^0 x dx = \int dx = x
\end{aligned} \quad (44.116)$$

Alors:

$$I_2 = \int \sin^2 x dx = \frac{1}{2} (-\cos x \sin x + I_0) = \frac{1}{2} (-\cos x \sin x + x) \quad (44.117)$$

Ce qui fait:

$$\begin{aligned}
2\sqrt{r_2} \int_0^{u_0} \sin^2(u) du &= 2\sqrt{r_2} \frac{1}{2} (-\cos x \sin x + x) \Big|_0^{u_0} = \sqrt{r_2} (-\cos x \sin x + x) \Big|_0^{u_0} \\
&= \sqrt{r_2} [(-\cos u_0 \sin u_0 + u_0) - (-\cos 0 \sin 0 + 0)] = \sqrt{r_2} [(-\cos u_0 \sin u_0 + u_0)]
\end{aligned} \quad (44.118)$$

Or, nous avons aussi (cf. chapitre de Trigonométrie):

$$\sin(2u_0) = 2 \sin u_0 \cos u_0 \quad (44.119)$$

Donc:

$$2\sqrt{r_2} \int_0^{u_0} \sin^2(u) du = \sqrt{r_2} \left(-\frac{\sin(2u_0)}{2} + u_0 \right) = \sqrt{r_2} \frac{(2u_0 - \sin(2u_0))}{2} \quad (44.120)$$

Rappelons à nouveau que:

$$\cos^2 u_0 = \frac{r_1}{r_2} \Rightarrow u_0 = \arccos\left(\sqrt{\frac{r_1}{r_2}}\right) \quad (44.121)$$

Or, $r_2 \gg r_1$ donc $r_1/r_2 \cong 0$.

Si nous développons en série de Maclaurin (cf. chapitre de Suites et Séries) jusqu'au premier ordre:

$$\begin{aligned} u_0 &\cong \arccos(0) + \frac{1}{1!} \frac{d \arccos\left(\sqrt{\frac{r_1}{r_2}}\right)}{dr_1} r_1 = \frac{\pi}{2} - \frac{1}{\sqrt{\frac{r_1}{r_2}} r_2 \sqrt{1 - \frac{r_1}{r_2}}} r_1 \cong \frac{\pi}{2} - \frac{1}{\sqrt{\frac{r_1}{r_2}} r_2} r_1 \\ &= \frac{\pi}{2} - \sqrt{\frac{r_2}{r_1}} \frac{r_1}{r_2} = \frac{\pi}{2} - \sqrt{\frac{r_1}{r_2}} \end{aligned} \quad (44.122)$$

Alors:

$$u_0 \cong \frac{\pi}{2} - \sqrt{\frac{r_1}{r_2}} \quad (44.123)$$

Nous avons alors:

$$\begin{aligned}
\sqrt{r_2} \frac{(2u_0 - \sin(2u_0))}{2} &\cong \sqrt{r_2} \frac{\left(2\left(\frac{\pi}{2} - \sqrt{\frac{r_1}{r_2}}\right) - \sin\left(2\left(\frac{\pi}{2} - \sqrt{\frac{r_1}{r_2}}\right)\right)\right)}{2} \\
&= \sqrt{r_2} \frac{\left(\pi - 2\sqrt{\frac{r_1}{r_2}} - \sin\left(\pi - 2\sqrt{\frac{r_1}{r_2}}\right)\right)}{2} = \sqrt{r_2} \left(\frac{\pi}{2} - \sqrt{\frac{r_1}{r_2}} - \frac{1}{2}\sin\left(\pi - 2\sqrt{\frac{r_1}{r_2}}\right)\right) \quad (44.124) \\
&= \sqrt{r_2} \left[\frac{\pi}{2} - \sqrt{\frac{r_1}{r_2}} - \frac{1}{2}\left(\sin(\pi)\cos\left(-2\sqrt{\frac{r_1}{r_2}}\right) + \cos(\pi)\sin\left(-2\sqrt{\frac{r_1}{r_2}}\right)\right)\right] \\
&= \sqrt{r_2} \left[\frac{\pi}{2} - \sqrt{\frac{r_1}{r_2}} - \frac{1}{2}\left(\cos(\pi)\sin\left(-2\sqrt{\frac{r_1}{r_2}}\right)\right)\right] = \sqrt{r_2} \left[\frac{\pi}{2} - \sqrt{\frac{r_1}{r_2}} + \frac{1}{2}\sin\left(-2\sqrt{\frac{r_1}{r_2}}\right)\right]
\end{aligned}$$

Si on prend le développement de Maclaurin au premier ordre:

$$\sin\left(-2\sqrt{\frac{r_1}{r_2}}\right) \cong -2\sqrt{\frac{r_1}{r_2}} \quad (44.125)$$

Donc:

$$\begin{aligned}
\sqrt{r_2} \frac{(2u_0 - \sin(2u_0))}{2} &\cong \sqrt{r_2} \left[\frac{\pi}{2} - \sqrt{\frac{r_1}{r_2}} + \frac{1}{2}\sin\left(-2\sqrt{\frac{r_1}{r_2}}\right)\right] = \sqrt{r_2} \left[\frac{\pi}{2} - \sqrt{\frac{r_1}{r_2}} + \frac{1}{2}\left(-2\sqrt{\frac{r_1}{r_2}}\right)\right] \quad (44.126) \\
&= \sqrt{r_2} \left[\frac{\pi}{2} - \sqrt{\frac{r_1}{r_2}} - \sqrt{\frac{r_1}{r_2}}\right] = \sqrt{r_2} \left[\frac{\pi}{2} - 2\sqrt{\frac{r_1}{r_2}}\right]
\end{aligned}$$

Donc tout cela pour écrire finalement:

$$T \propto \exp\left(-\frac{2\gamma}{\sqrt{r_2}} \int_{r_1}^{r_2} \sqrt{\left(\frac{1}{r} - \frac{1}{r_2}\right)} dr\right) = \exp\left(-\frac{2\gamma}{\sqrt{r_2}} \sqrt{r_2} \left[\frac{\pi}{2} - 2\sqrt{\frac{r_1}{r_2}}\right]\right) = \exp\left(-\gamma \left[\pi - 4\sqrt{\frac{r_1}{r_2}}\right]\right) \quad (44.127)$$

Soit explicitement:

$$T \propto \exp\left(-\frac{1}{\hbar} \sqrt{\frac{m^4(Z-2)e^2 r_2}{4\pi\epsilon_0}} \cdot \left(\pi - 4\sqrt{\frac{r_1}{r_2}}\right)\right) \quad (44.128)$$

Relation dans laquelle nous pouvons remettre le coefficient de l'exponentielle que nous avons déterminé dans le chapitre de Physique Quantique Ondulatoire. C'est le facteur exponentiel dans la relation ci-dessus qui explique la grande variation des périodes radioactives des différents nuclides, alors que l'énergie des particules varie elle relativement peu.

Typiquement pour le noyau d'Uranium ^{238}U , nous prenons les valeurs dans les tables des constantes physiques et universelles qui sont dans la relation précédente pour obtenir une certaine valeur de T (je m'abstiendrai de montrer le calcul car les tables ne sont pas toutes d'accord entre elles...).

Ensuite, dans l'approximation semi-classique, le noyau α a, dans le puits, une vitesse de l'ordre de:

$$v = \sqrt{\frac{2E_\alpha}{m}} \quad (44.129)$$

et il effectue des allers-retours dans un noyau dont le rayon est de l'ordre de r_1 .

Ces allers-retours correspondent donc à un certain nombre d'oscillations par seconde. Effectivement, si nous notons τ la durée moyenne entre deux chocs successifs, nous avons alors:

$$\tau = \frac{2r_1}{v} = \frac{2r_1}{\sqrt{\frac{2E_\alpha}{m}}} \quad (44.130)$$

Donc la fréquence vaut:

$$f = \frac{1}{\tau} \quad (44.131)$$

À chaque fois elle a une probabilité T de franchir la barrière de potentiel. Cette probabilité par unité de temps est ainsi déterminée par:

$$\lambda \cong f \cdot T \quad (44.132)$$

et donne la constante de désintégration de l'isotope par émission α avec une relativement grosse erreur si on fait le calcul avec les valeurs numériques. Sinon, l'ordre de grandeur est par contre exact ce qui pas mal du tout! Le modèle (scolaire) présenté donne donc des résultats satisfaisants.

Ce qui est impressionnant dans ce résultat c'est que puisque T est très très sensible à $E_{tot} = E_\alpha$, les ordres de grandeur de λ varient énormément pour de petites variations de l'énergie. Et le modèle reste aussi satisfaisant sur environ 30 ordres de grandeur!!!

2.4.4. DÉSINTÉGRATION BÊTA- (4)

Définition: Lorsqu'un noyau est instable à cause d'un trop plein de neutrons (comme le Carbone 14 par exemple) il n'émettra pas de neutrons. En revanche il aura la faculté de changer un de ses neutrons en un proton. Lors de cette transformation, pour conserver la charge électrique totale du système, un électron sera créé. Cette transformation est la "**radioactivité bêta-**" (- car l'électron a une charge négative dans cette désintégration).

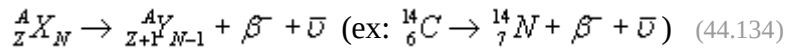
La désintégration dite β^- est donc une caractéristique des noyaux ayant un excès de neutrons. Les isotopes concernés se rendent plus stables en transformant un neutron en un proton avec émission d'un électron β^- et d'une particule appelée "**antineutrino**" dont nous justifierons la présence plus loin.

Nous avons alors pour le neutron concerné:

$${}^1_0n^{+1/2} \rightarrow p^{+1/2}_+ + \beta^{+1/2}_- + \bar{\nu}_0^{-1/2} \quad (44.133)$$

Nous avons mis en exposant droit le spin de la particule concernée et en indice droit le signe de charge de la particule. Ainsi, nous observons que le spin est une quantité conservée, ainsi que la charge.

Nous avons pour l'isotope concerné:



L'énergie dégagée lors de la transmutation se calcule au moyen du défaut de masse:

$$Q_{\beta^-} = (M_X - M_Y - m_e) \cdot c^2 \quad (44.135)$$

en négligeant l'énergie de liaison des électrons, nous avons:

$$m_X = M_X + Zm_e \text{ et } m_Y = M_Y + (Z+1)m_e \quad (44.136)$$

Attention! le Z dans l'égalité de m_Y est le même que celui que nous trouvons dans l'expression de m_X d'où le Z + 1.

Nous avons alors:

$$Q_{\beta^-} = (m_X - m_Y) \cdot c^2 \quad (44.137)$$

Chaque désintégration β^- pure est caractérisée par une énergie fixe de décroissance Q. Du fait que l'énergie cinétique du noyau est négligeable de par sa masse comparativement à celle de l'électron et de l'antineutrino réunis, l'énergie dégagée Q est partagée entre les énergies cinétiques de β^- et de $\bar{\nu}$. La masse de l'antineutrino étant de très loin inférieure à celle de l'électron, l'énergie cinétique de l'antineutrino peut aussi être négligée. L'énergie de β^- n'est cependant pas fixe et peut prendre n'importe quelle valeur entre 0 et Q. Nous observons donc un spectre d'énergie contrairement aux autres types de radioactivité (car l'électron peut avoir une énergie cinétique variable).

La forme des distributions observées permet de donner une valeur d'énergie moyenne aux β^- qui se situe autour de Q/3:

$$\bar{E} = \frac{Q_{\beta^-}}{3} = \frac{E_{\max}}{3} \quad (44.138)$$

L'existence de l'antineutrino a été postulée en 1933 (3 ans après le neutrino que nous verrons plus loin) par Wolfgang Pauli afin de satisfaire la conservation de spin. L'introduction d'une particule aussi étrange fut très controversée et mal acceptée (charge nulle, spin non nul, masse négligeable) et elle continue à poser quelques problèmes dans la physique contemporaine du 21ème siècle.

Indépendamment du neutrino d'électron (noté ν_e habituellement) accompagnant les particules β^- et β^+ (ce dernier ayant plusieurs noms "positon", "positron", "électron positif") il existe un neutrino de méson (muon) μ ou neutron tau (tauon) τ notés: ν_μ et ν_τ pour ne pas les confondre. Par la suite, n'étant pas confronté aux neutrinos de méson ou tau, nous noterons simplement le neutron électronique ν à la place de ν_e .

Remarque: Au début de sa découverte, la désintégration β^- était vue comme une transmutation du noyau..., dans les petites classes, encore aujourd'hui, on la voit comme la transformation d'un neutron en proton. Dans les théories contemporaines, elle est vue comme d'un quark d en quark u et elle a amené les physiciens à développer la théorie de l'interaction faible pour en expliquer l'origine.

2.4.5. DÉSINTÉGRATION β^+ (5)

Définition: Lorsqu'un noyau est instable à cause d'un trop plein de protons il n'émettra pas de protons. En revanche, il aura la faculté de changer un de ses protons en neutron, soit par capture d'un électron, phénomène appelé "**radioactivité par capture électronique**" (voir plus bas), soit par émission d'un électron positif (positon) ce qui correspond à la "**radioactivité bêta+**".

Cette transformation a une probabilité ridiculement faible puisque l'inverse de l'émission d'un électron et d'un antineutrino serait la capture simultanée de ces deux particules... et une telle rencontre serait un miracle. Pour surmonter cette difficulté, le noyau utilise un subterfuge quantique: l'émission d'une particule équivaut à la capture de son antiparticule. Ce joker offre alors les possibilités susmentionnées au noyau excédentaire en protons.

Lors de la désintégration β^+ un proton est dissocié en un neutron, un électron positif ("positon") noté β^+ et un neutrino (dont nous justifierons la présence un peu plus bas) et un neutrino.

Effectivement, pour effectuer l'inverse de la désintégration β^- , la solution consiste pour le noyau à utiliser la conservation de l'énergie et du spin en émettant un positon et en capturant dans l'énergie quantique du vide un antineutrino et d'émettre en échange un neutrino.

Nous écrivons cela:

$$p \rightarrow {}^1_0n + \beta^+ + \nu \quad (44.139)$$

ou:

$${}^A_ZX_N \rightarrow {}^A_{Z-1}Y_{N+1} + \beta^+ + \nu \quad (44.140)$$

L'énergie dégagée lors de la transmutation se calcule au moyen du défaut de masse:

$$Q_{\beta^+} = (M_X - M_Y - m_e)c^2 \quad (44.141)$$

en négligeant l'énergie de liaison des électrons, nous avons:

$$m_X = M_X + Zm_e \text{ et } m_Y = M_Y + (Z - 1)m_e \quad (44.142)$$

Attention! le Z dans l'égalité de m_Y est le même que celui que nous trouvons dans l'expression de m_X d'où le $Z - 1$

Nous avons ainsi:

$$Q_{\beta^+} = (m_X - m_Y - 2m_e)c^2 \quad (44.143)$$

La désintégration β^+ ne peut donc avoir lieu que si $Q_{\beta^+} \geq 0$, c'est-à-dire si:

$$(m_X - m_Y) \cdot c^2 \geq 2m_e c^2 = 1.022 \text{ [MeV]} \quad (44.144)$$

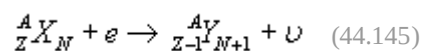
L'énergie massique de l'électron $E = m_e c^2$ est importante car c'est l'énergie d'un des deux photons résultant d'une annihilation d'un β^+ avec un électron.

Comme pour la désintégration β^- , l'énergie du β^+ n'est pas fixe et peut prendre n'importe quelle valeur entre 0 et Q . Nous observons donc un spectre d'énergie.

2.4.6. CAPTURE ÉLECTRONIQUE (6)

Définition: Lorsqu'un noyau est instable à cause d'un trop plein de protons par rapport aux neutrons, nous savons donc qu'une solution favorable du point de vue de son énergie est de transformer un de ses protons en neutron, c'est-à-dire de réaliser l'inverse de la radioactivité β^- . Nous avons vu tout à l'heure qu'une possibilité était pour le noyau via la désintégration β^+ d'attraper un antineutrino du vide et d'émettre un positon (perte de sa charge électrique) et un neutrino. Mais il peut aussi capturer un électron du cortège électronique (neutralisation de sa charge électrique) en lieu et place d'émettre un positon.

Ce sera le plus souvent un électron de la couche K . Ce qui se note:



L'énergie dégagée lors de la transmutation se calcule au moyen du défaut de masse:

$$Q_{EC} = (M_X + m_e - M_Y) c^2 \quad (44.146)$$

en supposant que l'énergie de liaison de l'électron K et celle de recul du noyau sont négligeables.

C'est donc le neutrino d'électron qui emporte toute l'énergie, d'où la nécessité qu'avait eue Wolfgang Pauli d'introduire cette nouvelle particule (ce qui lui avait fait horreur...!). Comme l'électron capturé occupait un niveau d'énergie précis dans l'atome, les neutrinos issus de la désintégration d'un isotope par capture électronique ont une énergie déterminée et présentent donc un spectre de raies.

En négligeant l'énergie de liaison des électrons, nous avons:

$$m_X = M_X + Zm_e \text{ et } m_Y = M_Y + (Z-1)m_e \quad (44.147)$$

donc:

$$Q_{EC} = (m_X - m_Y) c^2 \quad (44.148)$$

La désintégration par capture électronique n'est en concurrence avec la désintégration β^+ que si:

$$(m_X - m_Y) c^2 \geq 2m_e c^2 = 1.022 \text{ [MeV]} \quad (44.149)$$

Dans le cas où

$$(m_x - m_y)c^2 < 2m_e c^2 = 1.022 [MeV] \quad (44.150)$$

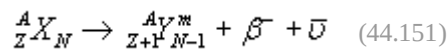
seule la désintégration par capture électronique est possible.

Cependant, le trou laissé par l'électron absorbé nécessite un réarrangement du cortège atomique et l'émission d'un rayonnement.

2.4.7. ÉMISSION GAMMA (7)

Définition: Pour le noyau, l'émission d'un rayonnement électromagnétique γ est une possibilité de gagner en stabilité. Cette émission suit généralement un phénomène de désintégration β^- , β^+ , α ou de capture électronique. On peut donc s'imaginer que lors de tels types de désintégration, la topologie des nucléons dans le noyau n'est pas idéale et que le réarrangement de ces derniers s'accompagnera d'une diminution d'énergie; cette dernière émise sous forme d'un ou de plusieurs photons γ .

Nous avons donc un schéma:



puis:

$${}^A_{Z+1} Y_{N-1}^m = {}^A_{Z+1} Y_{N-1} + \gamma \text{ (désintégration } \beta^- \text{)} \quad (44.152)$$

où le m signifie "métastable" ou "**isomère**" (on utilise ce dernier terme lorsque l'émission du rayonnement a lieu longtemps après la désintégration).

Remarque: "Isomère" veut dire que le noyau est excité. Il se désexcitera avec une période $T_{1/2}$. Généralement $T_{1/2}$ est extrêmement petit et le(s) photon(s) est (sont) émis immédiatement après l'électron dans le cas de notre exemple d'une désintégration β^- . Nous parlons alors d'état métastable ou isomère. Notons que ces radio-isotopes isomères sont particulièrement intéressants en imagerie médicale.

L'énergie du photon γ vaut:

$$Q_\gamma = (M_{Y^m} - M_Y)c^2 \quad (44.153)$$

Il est évident que dans cet exemple, nous avons considéré le cas le plus simple; soit la désexcitation de noyau Y^m en une seule étape avec émission d'un seul photon γ qui emporte toute l'énergie. De fait, selon le radio-isotope, cette désexcitation peut s'effectuer avec l'émission de plusieurs photons γ en cascade.

2.4.8. CONVERSION INTERNE (8)

La conversion interne I.C. est un processus lié aussi à l'émission d'un photon γ . En effet, il se peut que l'énergie soit transmise directement à un électron du cortège électronique, généralement de la couche K, qui se trouve éjecté de l'atome. Cet électron est appelé "**électron de conversion**". La place laissée dans le cortège électronique est par la suite comblée par un électron des couches supérieures et ainsi de suite. On a donc, comme dans le cas d'un processus de désintégration de capture électronique, un réarrangement du cortège électronique caractérisé par l'émission de rayons-X caractéristiques de l'élément Y.

L'énergie transmise vaut:

$$E_{e^-} = E_\gamma - E_b \quad (44.154)$$

avec E_{e^-} étant l'énergie cinétique de l'électron émis, E_γ l'énergie du photon percutant l'électron, E_b , l'énergie de liaison de l'électron considéré (K, L, M,...)

L'énergie du photon γ est transmise directement à un électron qui est éjecté; le processus est suivi du réarrangement des électrons (s'ensuivra une émission de rayons X). L'électron éjecté est appelé "**électron Auger**".

Si nous représentons sur un graphique tous les isotopes avec en ordonnée leur nombre atomique Z et en abscisse leur nombre de neutrons nous pouvons observer que les éléments stables existant dans la nature se trouvent tous dans la région nommée "vallée de stabilité". Les autres étant radioactifs. Nous pouvons remarquer que la ligne $Z = A$ est située presque partout en-dessous de la zone de stabilité.

Ces résultats ont été obtenus expérimentalement, car il est encore mal aisé aujourd'hui même avec les ordinateurs les plus puissants et ce en connaissant la théorie quantique, de simuler le comportement de noyaux ayant des nombres atomiques élevés.

L'émission d'un électron du cortège électronique appelé "**électron Auger**" est donc un processus similaire au processus de conversion interne (IC), mais le rayonnement électromagnétique ne provient pas d'une désexcitation du noyau (ce n'est pas un photon γ) mais d'un rayon-X produit lors du réarrangement du nuage électronique. Dans un processus radioactif, ce réarrangement électronique peut provenir soit d'une capture électronique (EC) soit d'une conversion interne (IC).

L'électron Auger éjecté provient principalement d'une orbitale externe et son énergie est l'énergie caractéristique du rayon-X moins son énergie de liaison. L'énergie des électrons Auger est donc faible (quelques [keV]) par rapport à une particule β^- et ces électrons sont donc souvent réabsorbés à l'intérieur de la source. Le processus d'émission d'un électron Auger est favorisé pour des éléments à faible numéro atomique à cause de leurs faibles énergies de liaison électronique.

Lors d'un réarrangement du nuage électronique tel que le passage d'un électron de la couche L à la couche K, l'énergie du rayon-X émis vaudra $E_K - E_L$. Cette différence d'énergie étant supérieure à l'énergie de liaison d'un autre électron se trouvant sur la couche L, ce dernier sera alors émis avec l'énergie cinétique:

$$E_{Auger} = h\nu_{L \rightarrow K} - E_L = (E_K - E_L) - E_L = E_K - 2E_L > 0 = c^2 \quad (44.155)$$

À leur tour, les 2 vacances laissées sur la couche L sont comblées par des électrons des couches supérieures. Fluorescence et électron Auger sont en compétition. Il se peut même que plusieurs électrons Auger soient émis lors de la désexcitation de l'atome. On parle alors de "**cascade Auger**" laissant l'atome considéré fortement ionisé, ce qui peut conduire à l'explosion coulombienne de la molécule dont il fait partie.

Pour conclure sur l'ensemble de ces phénomènes radioactifs indiquons l'ordre de grandeur des périodes radioactives de quelques éléments naturels et artificiels:

Nucléide	Décroissance	Période
Thorium 232	α	$\sim 10^{10}$ années
Uranium 238	α	$\sim 10^9$ années
Uranium 235	α	$\sim 10^8$ années
Uranium 233	α	$\sim 10^5$ années
Plutonium 239	α	$\sim 10^4$ années
Plutonium 238	α	~ 88 années
Radium 226	α	$\sim 10^3$ années
Curium 242	α	160 jours
Potassium 40	β^-	$\sim 10^9$ années
Carbone 14	β^-	$\sim 10^3$ années
Tritium	β^-	~ 12 années
Cobalt 60	β^-	~ 5.3 années
Iode 131	β^-	~ 8 jours
Azote 16	β^-	7.1 secondes
Technétium 97	CE	$\sim 10^6$ années
Cobalt 58	β^+	~ 20 minutes
Fluor 18	β^+	~ 110 minutes

Tableau: 28.1 - Ordres des grandeurs des périodes radioactives de quelques nucléides

Avec un exemple de la famille radioactive de l'uranium 238:

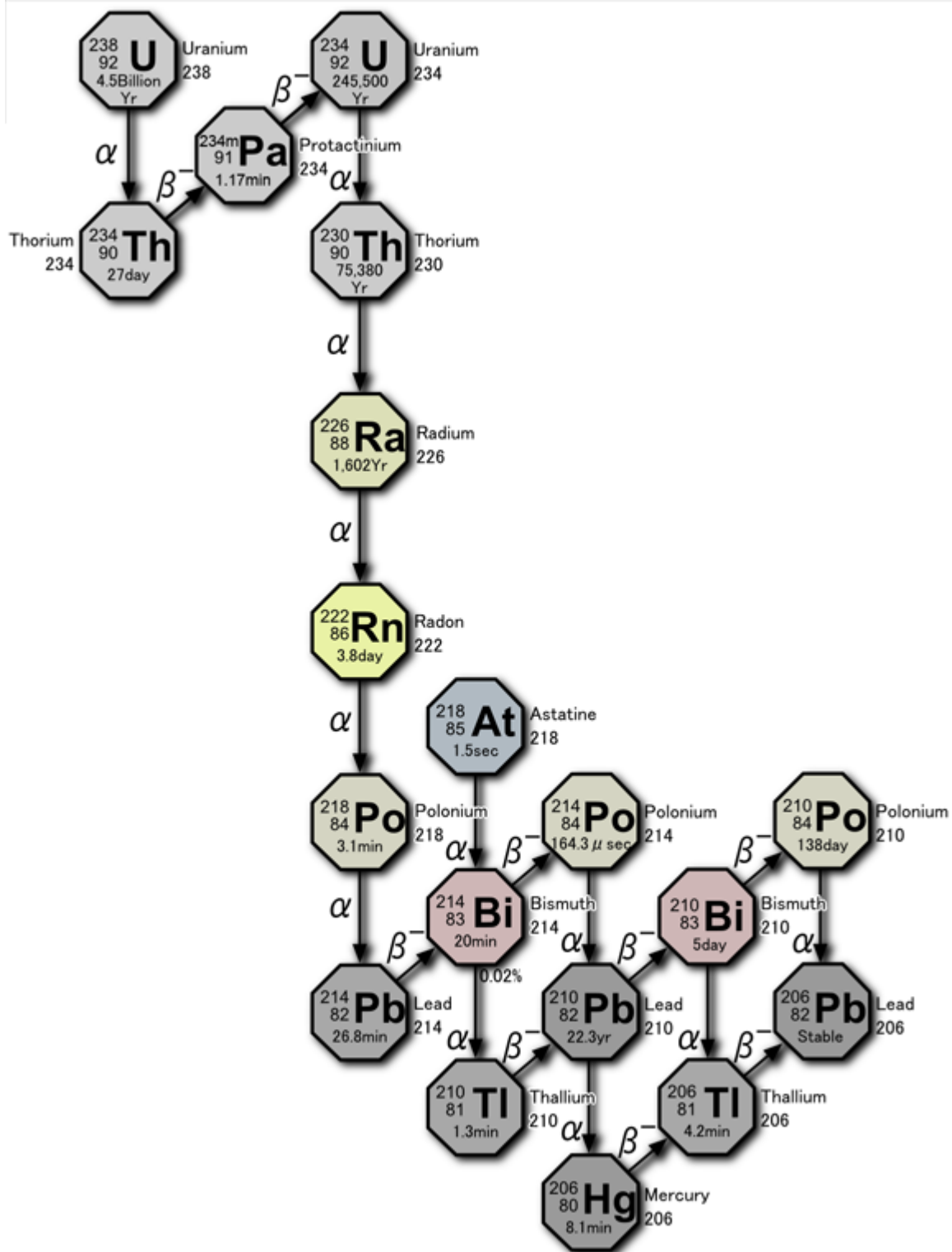


Figure: 44.8 - Famille radioactive de l'uranium 238 (source: Wikipédia)

3. RADIOPROTECTION

En physique nucléaire, il est très important de connaître la façon dont les divers rayonnements alpha, gamma, rayons-X ou neutroniques interagissent avec la matière (en gros les rayonnements non chargés ou chargés). Cela permet de connaître la façon dont leur énergie cinétique se répartit ou se dissipe dans la matière qu'ils rencontrent sur leur chemin et de s'en protéger de façon adaptée

3.1. FORMULE DE BETHE-BLOCH

Une particule chargée lourde ayant une énergie de un ou plusieurs MeV perd son énergie principalement par collisions avec les électrons des cortèges atomiques, électrons qui lui apparaissent comme quasi-libres. Le processus par lequel des électrons sont ainsi éjectés lors du passage d'une particule ionisante est appelé "**ionisation primaire**". Un électron pourra s'échapper s'il reçoit une énergie supérieure à son énergie de liaison.

Le transfert maximum d'énergie E_{max} qui peut se produire dans une collision non relativiste et élastique (où l'énergie du système est conservée car il n'y a par définition pas de dissipation de chaleur) est calculé simplement en utilisant le principe de conservation de la quantité de mouvement et d'énergie:

Soient M, v_M et m_e, v_e les masses et vitesses respectives de la particule incidente et de l'électron. Nous supposons que l'électron est immobile sur son orbite et que sa vitesse initiale est nulle $v_e = 0$. Après le choc, nous supposons que la particule incidente aura transféré toute son énergie cinétique à l'électron et se trouvera à son tour au repos telle que $v'_M = 0$.

Posons d'abord l'équation de conservation de la quantité de mouvement:

$$Mv_M + m_e v_e = Mv'_M + m_e v'_e \quad (44.156)$$

La conservation de l'énergie nous permet aussi d'écrire:

$$\frac{1}{2} Mv_M^2 + \frac{1}{2} m_e v_e^2 = \frac{1}{2} Mv'^2_M + \frac{1}{2} m_e v'^2_e \quad (44.157)$$

D'où après regroupement et simplification:

$$\begin{aligned} M(v_M - v'_M) &= m_e (v'_e - v_e) \\ M(v_M^2 - v'^2_M) &= m_e (v'^2_e - v_e^2) \end{aligned} \quad (44.158)$$

soit autrement écrit:

$$\begin{aligned} M(v_M - v'_M) &= m_e (v'_e - v_e) \\ M(v_M + v'_M)(v_M - v'_M) &= m_e (v'_e - v_e)(v'_e + v_e) \end{aligned} \quad (44.159)$$

Ensuite, après division de la deuxième équation par la première nous avons:

$$v_M + v'_M = v'_e + v_e \quad (44.160)$$

Nous avons alors le système:

$$\left. \begin{aligned} Mv_M + m_e v_e &= Mv'_M + m_e v'_e \\ v'_M &= v'_e + v_e - v_M \end{aligned} \right\} \Rightarrow m_e v'_e = Mv_M + m_e v_e - M(v'_e + v_e - v_M) \quad (44.161)$$

on déduit l'expression des vitesses après le choc:

$$v'_e = \frac{v_e(m_e - M) + 2Mv_M}{M + m_e} \quad (44.162)$$

relativement à nos hypothèses initiales, nous avons ($v_e = 0, v'_M = 0$) donc:

$$v'_e = \frac{2Mv_M}{M + m_e} \quad (44.163)$$

Manipulons un petit peu cette relation:

$$\begin{aligned} v'_e &= \frac{2Mv_M}{M + m_e} \Rightarrow \frac{1}{2} m_e v'_e v'_e = \frac{1}{2} m_e v'_e \frac{2Mv_M}{M + m_e} \Rightarrow E'_c = \frac{1}{2} M v_M v_M \frac{2m_e v'_e}{v_M (M + m_e)} \\ \Rightarrow E'_c &= E_c^M \frac{2m_e}{(M + m_e)} \frac{v'_e}{v_M} = E_c^M \frac{2m_e}{(M + m_e)} \left(\frac{2Mv_M}{M + m_e} \right) \frac{1}{v_M} = \frac{4Mm_e}{(M + m_e)^2} E_c^M \end{aligned} \quad (44.164)$$

Pour une particule lourde, avec $M \gg m_e$, nous pouvons écrire:

$$\max E'_c = \frac{4m_e}{M} \cdot E_c^M \quad (44.165)$$

Une ionisation ne pourra se produire que si $\max E'_c$ est au moins égale au seuil d'ionisation de l'électron que l'on notera I_0 et dont on a vu le calcul lors de l'étude du modèle de Bohr (cf. chapitre de Physique Quantique Corpusculaire).

L'énergie de la particule incidente devra donc au minimum être égale à:

$$E_{\min}^M = \frac{M \cdot I_0}{4m_e} \quad (44.166)$$

Donc, lors de son passage à travers la matière, le corps chargé de charge $Z \cdot e$ et de vitesse v_M cède son énergie en de nombreuses collisions avec les électrons des atomes rencontrés. L'interaction est coulombienne et à chaque fois, une diffusion se produit. L'énergie de recul de l'électron, supposé libre, peut se calculer de manière précise. Pour faire une estimation de la perte d'énergie, nous ferons ici l'approximation que la quantité de mouvement transférée Δp est égale au produit de la force d'interaction à la distance r multipliée par le temps nécessaire au projectile pour parcourir le trajet $2r$. Nous avons la force F de Coulomb donnée par:

$$F = \frac{1}{4\pi\epsilon_0} \cdot \frac{Ze^2}{r^2} \quad (44.167)$$

et la quantité de mouvement:

$$\Delta\varphi = F \cdot t = \left(\frac{1}{4\pi\epsilon_0} \frac{Ze^2}{r^2} \right) \cdot \frac{2r}{v_M} = \frac{1}{4\pi\epsilon_0} \frac{2Ze^2}{r \cdot v_M} \quad (44.168)$$

L'énergie cinétique transférée à un électron de masse m_e sera:

$$\Delta E_c^e = \frac{1}{2} m_e v_M^2 = \frac{1}{2} \frac{(\Delta\varphi)^2}{m_e} = \frac{1}{(4\pi\epsilon_0)^2} \cdot \frac{2Z^2 e^4}{r^2 \cdot v_M^2 \cdot m_e} \quad (44.169)$$

La perte d'énergie totale sera obtenue en intégrant sur tous les électrons rencontrés. À la distance comprise entre r et $r + dr$ de la trajectoire et sur le parcours dx , se trouvent:

$$dN = N \cdot Z' \cdot 2\pi r \cdot dr \cdot dx \quad (44.170)$$

électrons, où N est le nombre d'atomes de nombre atomique Z' par unité de volume. La perte d'énergie par unité de distance est donc:

$$\begin{aligned} -\frac{dE}{dx} &= \int_{r_{\min}}^{r_{\max}} \Delta E_c^e \cdot \frac{dN}{dx} = \int_{r_{\min}}^{r_{\max}} \frac{1}{(4\pi\epsilon_0)^2} \cdot \frac{2Z^2 e^4}{r^2 \cdot v_M^2 \cdot m_e} 2\pi Z' N r dr = \frac{4\pi N Z'}{(4\pi\epsilon_0)^2} \cdot \frac{Z^2 e^4}{m_e v_M^2} \int_{r_{\min}}^{r_{\max}} \frac{1}{r} dr \\ &= \frac{4\pi N Z'}{(4\pi\epsilon_0)^2} \cdot \frac{Z^2 e^4}{m_e v_M^2} \ln \left(\frac{r_{\max}}{r_{\min}} \right) \end{aligned} \quad (44.171)$$

La valeur de $r_{\min}$ est évaluée en remarquant que ce paramètre d'impact correspond au transfert d'énergie maximum. En utilisant les équations que nous avons démontrées précédemment:

$$r_{\max} \Delta E_c^e = \frac{1}{(4\pi\epsilon_0)^2} \cdot \frac{2Z^2 e^4}{r_{\min}^2 v_M^2 m_e} = \frac{4m_e}{M} \cdot E_c^M \quad (44.172)$$

Avec $E_c^M = \frac{1}{2} M v_M^2$, on peut obtenir le paramètre $r_{\min}$ par:

$$r_{\min}^2 = \frac{1}{(4\pi\epsilon_0)^2} \cdot \frac{2Z^2 e^4}{v_M^2 m_e} \cdot \frac{1}{2m_e v_M^2} = \frac{1}{(4\pi\epsilon_0)^2} \cdot \frac{Z^2 e^4}{v_M^4 m_e^2} \quad (44.173)$$

et nous obtenons:

$$r_{\min} = \frac{1}{4\pi\epsilon_0} \cdot \frac{Ze^2}{v_M^2 m_e} \quad (44.174)$$

Lorsque r devient très grand, le transfert d'énergie est plus petit que l'énergie moyenne d'ionisation notée $\bar{I}$ des électrons et le processus n'est plus efficace. Nous devons donc avoir la relation suivante:

$$r_{\min} \Delta E_c^e = \frac{1}{(4\pi\epsilon_0)^2} \cdot \frac{2Z^2 e^4}{r_{\max}^2 v_M^2 m_e} > \bar{I} \quad (44.175)$$

Nous en tirons une valeur pour $r_{\max}$:

$$r_{\max} < \frac{1}{4\pi\epsilon_0} \cdot \frac{Ze^2}{v_M} \sqrt{\frac{2}{m_e \cdot I}} \quad (44.176)$$

En remplaçant les valeurs de $r_{\min}$ et $r_{\max}$ des équations précédentes dans l'équation:

$$-\frac{dE}{dx} = \frac{4\pi NZ'}{(4\pi\epsilon_0)^2} \cdot \frac{Z^2 e^4}{m_e v_M^2} \ln \left(\frac{r_{\max}}{r_{\min}} \right) \quad (44.177)$$

nous obtenons:

$$-\frac{dE}{dx} = \frac{4\pi NZ'}{(4\pi\epsilon_0)^2} \cdot \frac{Z^2 e^4}{m_e v_M^2} \ln \left(\sqrt{\frac{2m_e v^2}{I}} \right) \quad (44.178)$$

Un traitement quantique plus rigoureux montrerait qu'il faudrait supprimer la racine de l'argument du logarithme en prenant en compte les effets relativistes ainsi que les propriétés intrinsèques de l'électron (constante de structure fine). Nous obtiendrions alors la formule de Bethe-Bloch:

$$-\frac{dE}{dx} = \frac{4\pi NZ'}{(4\pi\epsilon_0)^2} \cdot \frac{Z^2 e^4}{m_e v_M^2} \left[\ln \left(\frac{2m_e v_M^2}{I} \right) - \ln(1 - \beta^2) - \beta^2 - \frac{c_k}{Z} \right] \quad (44.179)$$

où $\beta = v/c$. c_k est quant à lui un terme de correction qui dépend de l'énergie et de Z lorsque nous tenons compte de la structure complète des noyaux (modèle en couche) de la matière.

Nous voyons finalement que la perte d'énergie linéique est proportionnelle au numéro atomique du rayonnement incident et de la matière pénétrée. Donc, des protections composées de matériaux à numéro atomique élevé (masse volumique élevée) auront un fort pouvoir de ralentissement et seront avantageux en radioprotection.

3.2. EFFET COMPTON

L'effet Compton s'observe lorsqu'un photon est diffusé inélastiquement par une particule chargée. En fait le photon est absorbé et puis réémis par la particule lui cédant ainsi au passage une partie de son énergie. C'est ce transfert d'énergie qui justifie le caractère inélastique de la diffusion.

Ainsi si la particule chargée est un électron, cet effet peut avoir lieu indifféremment sur un électron de n'importe quelle couche électronique voire sur un électron libre. L'énergie du photon et celle de l'électron dépendent de la direction d'émission de ces particules. Étant donné que cet effet dépend du nombre d'électrons disponibles par atome cible, la probabilité de diffusion Compton augmente linéairement avec le nombre atomique Z de l'absorbant. Mais comme cet effet est en concurrence avec la production d'une paire électron - positron que nous verrons plus loin, l'effet Compton est surtout important aux énergies et aux numéros atomiques moyens.

Nous avons démontré dans le chapitre de Relativité Restreinte, la relation d'Einstein:

$$E = mc^2 = E_c + m_0 c^2 \Rightarrow E_c = mc^2 - m_0 c^2 \quad (44.180)$$

et rappelons que nous avons ainsi pour la quantité de mouvement d'un photon:

$$E = mc^2 \Rightarrow \frac{E}{c} = mc = p \quad (44.181)$$

et nous y avons aussi démontré que, partant de l'énergie totale, la quantité de mouvement est donnée par:

$$\begin{aligned} E &= c\sqrt{m_0^2 c^2 + p^2} \Rightarrow p^2 = \frac{E^2}{c^2} - m_0^2 c^2 \Rightarrow p = \frac{1}{c}\sqrt{E^2 - m_0^2 c^4} = \frac{1}{c}\sqrt{(E_c + m_0 c^2)^2 - m_0^2 c^4} \\ &= \frac{1}{c}\sqrt{E_c^2 + 2E_c m_0 c^2 + m_0^2 c^4 - m_0^2 c^4} = \frac{1}{c}\sqrt{E_c^2 + 2E_c m_0 c^2} = \frac{1}{c}\sqrt{E_c (E_c + 2m_0 c^2)} \end{aligned} \quad (44.182)$$

d'où la relation, dont nous allons faire usage plus loin:

$$p^2 c^2 = E_c (E_c + 2m_0 c^2) \quad (44.183)$$

Avant l'interaction, photon-électron, nous avons (nous considérons grossièrement l'électron comme étant au repos) $h\nu_1 + m_0 c^2$ et après la collision $h\nu_2 + \sqrt{m_0^2 c^4 + p^2 c^2}$. La conservation de l'énergie nous amène donc à écrire:

$$h\nu_1 + m_0 c^2 = h\nu_2 + \sqrt{m_0^2 c^4 + p^2 c^2} \quad (44.184)$$

En ne considérant que les énergies cinétiques, nous avons en négligeant celle de l'électron avant le choc:

$$h\nu_1 = E_e + h\nu_2 = (mc^2 - m_0 c^2) + h\nu_2 \quad (44.185)$$

Soit la figure ci-dessous:

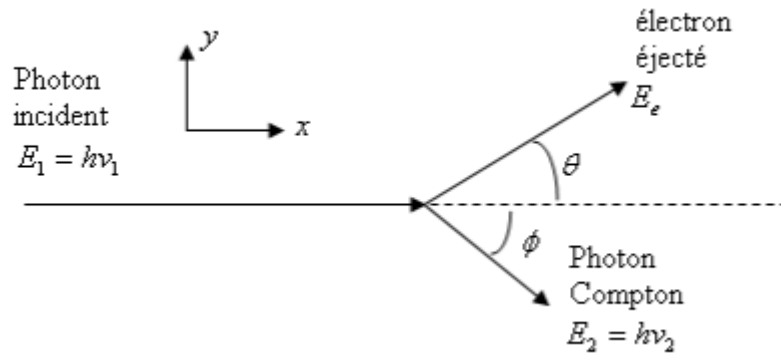


Figure: 44.9 - Illustration de l'effet Compton

La conservation de la quantité de mouvement nous donne:

Selon l'axe x:

$$\frac{E_1}{c} = \frac{E_2}{c} \cos \phi + p_e \cos \theta \Leftrightarrow p_e \cos \theta = \frac{E_1}{c} - \frac{E_2}{c} \cos \phi \quad (44.186)$$

Selon l'axe y:

$$0 = \frac{E_2}{c} \sin \phi - p_e \sin \theta \Leftrightarrow p_e \sin \theta = \frac{E_2}{c} \sin \phi \quad (44.187)$$

La somme de ces deux relations élevées au carré nous donne la quantité de mouvement totale:

$$\begin{aligned}
 p_e^2 \cos^2 \theta + p_e^2 \sin^2 \theta &= \left(\frac{E_1}{c} - \frac{E_2}{c} \cos \phi \right)^2 + \left(\frac{E_2}{c} \sin \phi \right)^2 \\
 \Rightarrow p_e^2 &= \frac{E_1^2}{c^2} - 2 \frac{E_1 E_2}{c^2} \cos \phi + \frac{E_2^2}{c^2} \cos^2 \phi + \frac{E_2^2}{c^2} \sin^2 \phi = \frac{E_1^2}{c^2} - 2 \frac{E_1 E_2}{c^2} \cos \phi + \frac{E_2^2}{c^2} \quad (44.188) \\
 \Rightarrow p_e^2 c^2 &= E_1^2 + E_2^2 - 2 E_1 E_2 \cos \phi
 \end{aligned}$$

Puis en substituant $p_e^2 c^2$:

$$E_e (E_e + 2m_0 c^2) = E_1^2 + E_2^2 - 2 E_1 E_2 \cos \phi \quad (44.189)$$

et comme $E_e = E_1 - E_2$:

$$E_2 = \frac{E_1}{1 + \frac{E_1}{m_0 c^2} (1 - \cos \phi)} \quad (44.190)$$

Lorsque l'énergie du photon est assez élevée, $E_1 \gg m_0 c^2$, celle du photon diffusé tend vers une limite donnée par (voir la règle de l'Hospital dans le chapitre de Calcul Différentiel Et Intégral):

$$\lim_{h\nu_1 \gg m_0 c^2} h\nu_2 = \lim_{h\nu_1 \gg m_0 c^2} E_2 = \frac{m_0 c^2}{(1 - \cos \phi)} \quad (44.191)$$

L'énergie acquise par l'électron Compton vaut finalement:

$$E_e = E_1 - E_2 = \frac{\frac{E_1^2}{m_0 c^2} (1 - \cos \phi)}{1 + \frac{E_1}{m_0 c^2} (1 - \cos \phi)} \quad (44.192)$$

Il est intéressant de remarquer que nous ne pouvons avoir $E_e = E_1$. Effectivement cela supposerait que:

$$E_2 = \frac{E_1}{1 + \frac{E_1}{m_0 c^2} (1 - \cos \phi)} = 0 \quad (44.193)$$

et nous voyons bien que quel que soit ϕ , nous avons toujours $E_2 \neq 0$.

La fréquence du photon diffusé est inférieure à celle du photon incident car son énergie E_2 est toujours plus faible et donc sa longueur d'onde λ_2 plus grande. Donc:

$$\Delta \lambda = \lambda_2 - \lambda_1 = c \left(\frac{1}{\nu_2} - \frac{1}{\nu_1} \right) = hc \left(\frac{1}{E_2} - \frac{1}{E_1} \right) = \frac{hc}{E_1} \left(\frac{E_1}{E_2} - 1 \right) \quad (44.194)$$

et puisque:

$$E_2 = \frac{E_1}{1 + \frac{E_1}{m_0 c^2} (1 - \cos \phi)} \quad (44.195)$$

Nous avons:

$$\Delta\lambda = \frac{hc}{E_1} \left(1 + \frac{E_1}{m_0 c^2} (1 - \cos \phi) - 1 \right) = \frac{hc}{m_0 c^2} (1 - \cos \phi) \quad (44.196)$$

ce qui s'écrit aussi en utilisant la définition de la constante de Planck et les relations trigonométriques habituelles:

$$\Delta\lambda = \frac{4\pi\hbar}{mc} \sin^2 \left(\frac{\phi}{2} \right) \quad (44.197)$$

Nous appelons le facteur $\frac{hc}{m_0 c^2}$ la "**longueur d'onde de Compton**" et elle vaut:

$$\lambda_C = 2.426 \cdot 10^{-12} [m] \quad (44.198)$$

Remarquons que si l'angle ϕ est nul, alors la variation de longueur d'onde est nulle et que si l'angle ϕ correspond à 180° alors la variation de longueur d'onde du photon incident est le double de la longueur d'onde de Compton!

3.3. EFFET PHOTOÉLECTRIQUE

L'effet photoélectrique est l'éjection d'électrons (dits alors "photoélectrons") de la surface de divers métaux exposée à une énergie de rayonnement. Ce rayonnement peut provenir du réarrangement du noyau de l'atome aussi bien que d'un rayonnement externe.

Par ailleurs, c'est par des mesures quantitatives de l'effet photoélectrique qu'Einstein proposa d'éprouver la validité de la théorie quantique de la lumière (transport d'énergie par paquets: quanta) et donc l'explication théorique lui valut le prix Nobel.

Exposons d'abord l'expérience mise en oeuvre: l'émission d'électrons par un métal ne contredit pas la théorie électromagnétique de la lumière. Si nous considérons un faisceau uniforme, son énergie est uniformément répartie sur tout le front d'onde. Plus la lumière est intense, plus grandes sont les amplitudes des champs électrique et magnétique en chaque point du front d'onde et plus l'énergie transmise par l'onde en une seconde est grande. Ces champs exercent des forces sur les électrons dans le métal et peuvent même en arracher de sa surface.

Voici l'expérience mise en place:

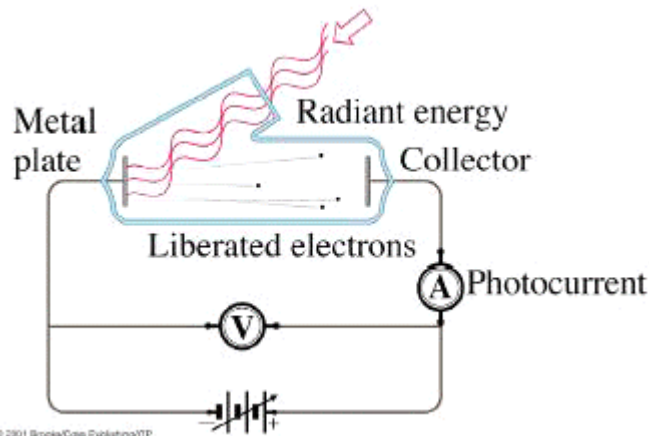


Figure: 44.10 - Expérience pour la mesure de l'effet photoélectrique

Si l'anode collectrice est à un potentiel positif relativement à la cathode émettrice, les photoélectrons parcourent le tube et constituent le courant mesuré par l'ampèremètre. Nous observons alors une proportionnalité entre l'intensité du faisceau incident et le courant.

Cependant, au moins trois problèmes persistent entre le modèle théorique et l'observation expérimentale:

1. La notion ondulatoire de la lumière ne convient pas pour expliquer le temps nécessaire à l'absorption de l'énergie d'extraction.

Effectivement, supposons une lampe de 100 [W] avec un rendement lumineux 15% placée à 0.5 [m] d'une plaque revêtue de potassium K d'énergie d'extraction E_i minimale de 2.25 [eV] en admettant un diamètre de 1[Å] pour l'atome de Potassium.

Nous avons alors:

$$I = \frac{P}{S_{\text{rayonnement}}} = \frac{100[W] \cdot 0.15}{4\pi R^2} = 4.77 \left[\frac{W}{m^2} \right] \quad (44.199)$$

La puissance lumineuse absorbée par l'atome par sa demi-surface qui fait face au rayonnement est alors:

$$P_K = I \cdot \frac{S_K}{2} = 4.77 \frac{4\pi R_K^2}{2} = 4.77 \cdot 2\pi (0.5 \cdot 10^{-10})^2 \cong 7.49 \cdot 10^{-20} [W] \quad (44.200)$$

La durée nécessaire pour l'absorption est alors:

$$t = \frac{E_i}{P_K} = \frac{2.25 \cdot 1.602 \cdot 10^{-19}}{7.49 \cdot 10^{-20}} \cong 4.8[s] \quad (44.201)$$

ce qui est en contradiction avec l'expérience où l'on observe que le phénomène est quasi instantané (le temps à la lumière pour se propager jusqu'au métal).

2. Si nous inversons les bornes, les électrons émis par le métal sont repoussés par l'électrode négative, mais si la tension inverse est faible les plus rapides pourront quand même l'atteindre et il se produira un courant. À un potentiel négatif, spécifique pour chaque métal, appelé potentiel d'arrêt U_a , tous les électrons émis sont repoussés et le courant est nul. L'énergie cinétique maximale de ces photoélectrons est alors:

$$E_{e,max} = e^-U_a \quad (44.202)$$

Or, nous trouvons expérimentalement que ce potentiel d'arrêt est indépendant de l'intensité du rayonnement. Dans la théorie ondulatoire, l'augmentation de l'intensité devrait augmenter le nombre d'électrons extraits (quel que soit leur niveau énergétique) et leur énergie cinétique maximale. Une plus grande intensité suppose une plus grande amplitude du champ électrique: $I \propto E^2$. Ainsi, un champ électrique plus grand devrait éjecter les électrons à plus grande vitesse toutes couches confondues au fur à mesure que l'intensité augmente.

3. Lorsque nous varions la fréquence ν de la lumière incidente et que nous mesurons U_a , nous observons que l'effet photoélectrique n'a pas lieu si $\nu < \nu_0$ (ν_0 est appelé le seuil de fréquence) et ceci quelle que soit l'intensité de la lumière. Ce qui est plutôt gênant... parce que dans la théorie ondulatoire, nous devons toujours pouvoir éjecter des électrons quelle que soit la fréquence, il suffit d'augmenter l'intensité.

Chaque problème peut être résolu en adoptant le point de vue suivant:

1. Dans l'aspect ondulatoire, la source est vue comme se propageant comme un front d'onde sphérique dont la densité superficielle d'énergie décroît comme $1/r^2$. Alors que pour expliquer l'observation expérimentale, il faut voir l'expérience d'un point de vue corpusculaire où le front est un front de corpuscules dont la densité superficielle de photons décroît en $1/r^2$ mais où l'énergie de chaque photon reste $h\nu$ (selon la relation de Planck-Einstein).
2. Si nous pensons en termes de photons, quand nous augmentons l'intensité, nous augmentons le nombre de photons, mais l'énergie par photon $E = h\nu$, reste inchangée. Ainsi, $E_{e,max}$ que peut avoir chaque photon ne change pas. D'où le fait que le potentiel d'arrêt est indépendant de l'intensité du champ.
3. Si nous pensons en termes de photons à nouveau, les électrons dans la cible sont retenus par les forces d'attraction, l'extraction d'un électron de la surface requiert une énergie minimale E_i qui dépend de chaque matériau (E_i est aussi appelé "**travail d'extraction**" qui est de l'ordre de quelques électronvolts). Si l'énergie du photon incident $E_i = h\nu$ est supérieure à E_i , un électron peut être arraché, par contre si elle est inférieure, aucun électron ne peut être arraché. L'apport d'énergie $h\nu$ est égal à l'énergie cinétique de sortie de l'électron plus l'énergie requise pour l'extraire du métal, soit:

$$E_e + E_i = E_i = h\nu \quad (44.203)$$

Ainsi, si l'on augmente la fréquence de la lumière, l'énergie cinétique maximale des électrons augmente linéairement. R.A. Millikan fit entre 1913-1914 des expériences rigoureuses dont les résultats corroborèrent parfaitement la théorie d'Einstein. Ce dernier reçut le prix Nobel en 1921 pour ses apports à la physique théorique, et surtout sa découverte de la loi de l'effet photoélectrique.

La lumière se propage d'un endroit à un autre comme si elle était une onde. Mais la lumière interagit avec la matière dans des processus d'absorption et d'émission comme si elle était un courant de particules. C'est ce que nous appelons la "**dualité onde-corpuscule**". Ainsi, celle-ci se trouvant dans les particules massives comme le suggère l'hypothèse de De Broglie que nous avons vue en physique quantique ondulatoire, se vérifie finalement également pour la lumière.

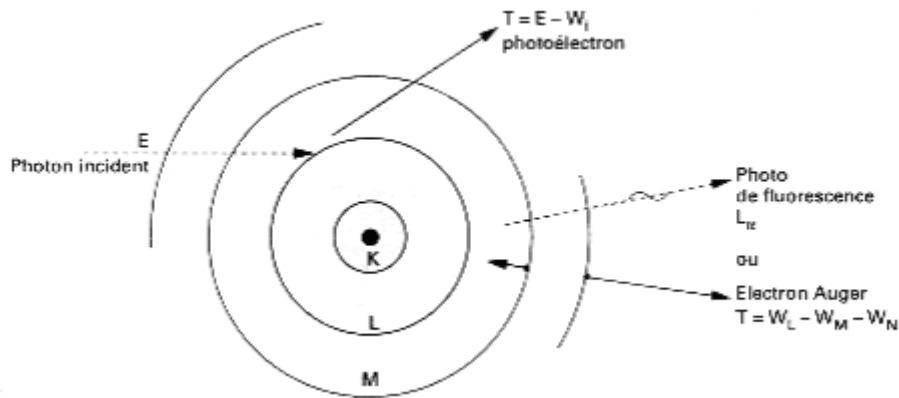


Figure: 44.11 - Principe de l'effet photoélectrique sur le modèle de Bohr de l'atome

Un photon d'énergie incidente E_i qui interagit avec un électron d'un atome cible peut éjecter cet électron de son orbite en lui communiquant une énergie cinétique E_c :

$$E_c = E_i - E_i \quad (44.204)$$

où E_i est l'énergie de liaison de l'électron éjecté de son orbite (cette relation est indiquée sous la forme $T = E - W_i$ dans la figure ci-dessus).

Si l'énergie du photon incident est inférieure à l'énergie de liaison de l'électron K (cf. chapitre de Physique Quantique Corpusculaire), l'effet photoélectrique se fait avec un électron de la couche L, etc.

Dans le cas où le rayonnement est absorbé, l'atome est dit "excité", car son état d'énergie n'est pas l'état minimal. Il s'ensuit donc une "relaxation" (ou "désexcitation"): un électron d'une couche supérieure vient combler la case quantique laissée vacante par l'électron éjecté.

Si l'énergie de transition est modérée (c'est-à-dire si le rayonnement incident avait une énergie modérée), la relaxation provoque l'émission d'un photon de faible énergie (visible ou ultra-violet), c'est le phénomène de fluorescence. Si l'énergie de transition est élevée, on peut avoir deux cas:

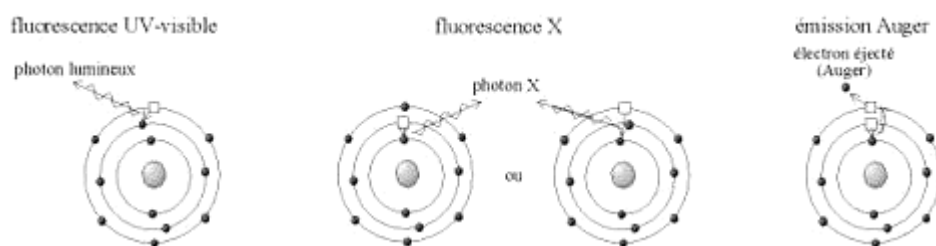


Figure: 44.12 - Types de transitions d'énergie selon le modèle de Bohr

1. Il y a émission d'un photon fluorescent, qui du fait de son énergie, est un photon X, nous parlons alors de "**fluorescence X**"

2. Ce photon X peut être recapturé par l'atome lui-même et provoquer l'éjection d'un électron périphérique, c'est "**l'émission Auger**" dont nous avons déjà parlé plus haut.

Pour résumer, nous avons vu jusqu'ici:

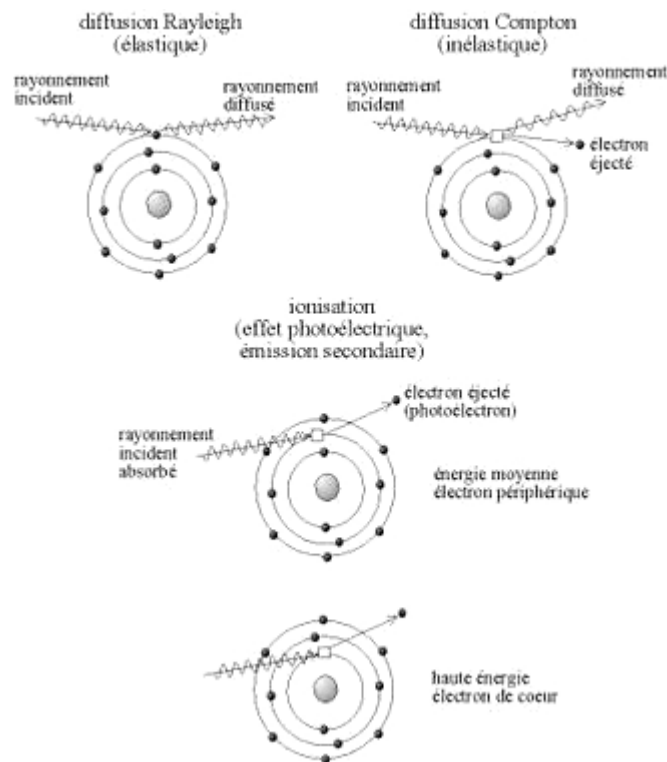


Figure: 44.13 - Diffusions et ionisations étudiées jusqu'ici

Indiquons qu'à l'époque de la découverte de l'effet, les premiers théoriciens avaient tenté de modéliser le phénomène par un processus mécanique fonction de l'intensité lumineuse (résonance mécanique). Or comme nous venons de le voir, ce n'est pas l'intensité de la lumière qui compte, mais sa fréquence.

3.4. DIFFUSION DE RUTHERFORD

Considérons la diffusion qu'une particule chargée subit quand elle est soumise à une force électrostatique répulsive inversement proportionnelle au carré de la distance entre la particule mobile et un point fixe ou centre de force. Ce problème est particulièrement intéressant en raison de son application à la physique atomique et nucléaire. Par exemple, quand un proton, accéléré par une machine telle qu'un cyclotron, passe près d'un noyau de la matière de la cible, il est dévié sous l'action d'une force de ce type, provenant de la répulsion électrostatique du noyau (c'est la raison pour laquelle nous parlons aussi de diffusion coulombienne).

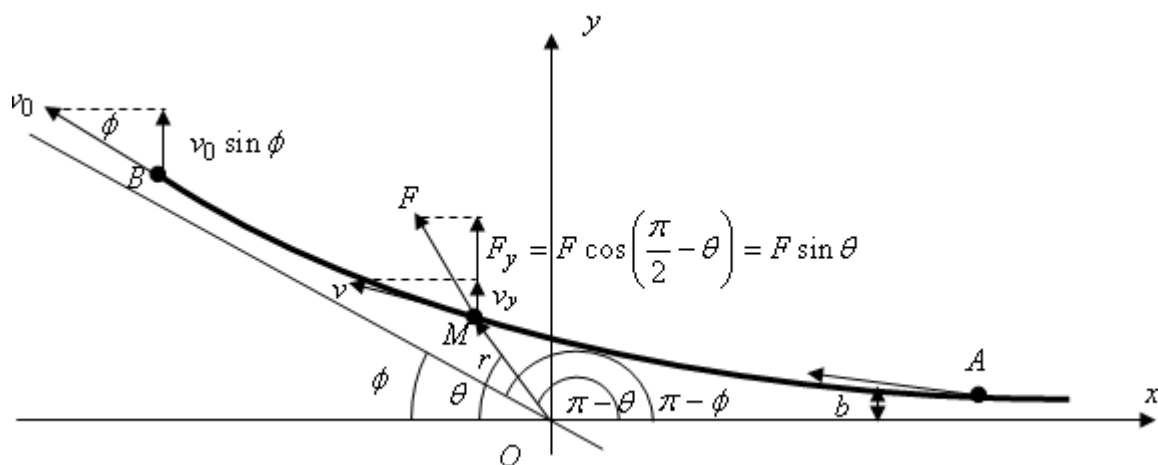


Figure: 44.14 - Diffusion de Rutherford

Soit O un centre de force et A une particule lancée contre O d'une grande distance avec la vitesse v_0 (voir figure ci-dessus). Nous choisisons l'axe des X passant par O et parallèle à v_0 . La distance b , appelée "**paramètre d'impact**", est la distance entre l'axe X des abscisses et le point A . En supposant que la force entre A et O est répulsive et centrale, la particule suivra la trajectoire AMB . La forme de la courbe dépend de la manière dont la force varie avec la distance. Si la force est inversement proportionnelle au carré de la distance, c'est-à-dire si:

$$F = k / r^2 \quad (44.205)$$

la trajectoire est une hyperbole. Avec bien évidemment (cf. chapitre d'Électrostatique):

$$k = \frac{1}{4\pi\epsilon_0} Z \cdot q \cdot Q \quad (44.206)$$

Quand la particule est en A son moment cinétique est mv_0b . Dans une position quelconque telle que M , son moment cinétique, est (cf. chapitre de Mécanique Classique) aussi donné par $mr^2 d\theta / dt$. Comme le moment cinétique doit rester constant puisque la force est centrale:

$$mr^2 \frac{d\theta}{dt} = mv_0b \quad (44.207)$$

L'équation du mouvement dans la direction OY est obtenue en combinant l'équation par:

$$m \frac{dv_y}{dt} = F_y = F \sin \theta = \frac{k}{r^2} \sin \theta \quad (44.208)$$

En éliminant r^2 à l'aide de l'avant-dernière équation nous pouvons écrire:

$$m \frac{dv_y}{dt} = \frac{k}{v_0b} \sin \theta \frac{d\theta}{dt} \quad (44.209)$$

Pour trouver la déviation de la particule, nous devons intégrer cette équation depuis l'une des extrémités de la trajectoire jusqu'à l'autre. En A la valeur de v_y est nulle car le mouvement initial est parallèle à l'axe des X et nous avons aussi $\theta = 0$. En B nous avons $v_y = v_0 \sin \phi$ et $\theta = \phi$ ou $\pi - \theta = \pi - \phi$. Remarquons qu'en B la vitesse est de nouveau v_0 car, par symétrie, la vitesse perdue quand la particule s'approche de O doit être regagnée quand elle s'en éloigne. Alors:

$$\int_0^{v_0 \sin \phi} dv_y = \frac{k}{mv_0b} \int_0^{\pi-\phi} \sin \theta d\theta \quad (44.210)$$

Ce qui donne:

$$v_0 \sin \phi = -\frac{k}{mv_0b} (\cos(\pi - \phi) - 1) \quad (44.211)$$

soit en utilisant les relations trigonométriques d'usage:

$$v_0 \sin \phi = \frac{k}{mv_0 b} (1 + \cos \phi) \quad (44.212)$$

et après réarrangement:

$$\frac{(1 + \cos \phi)}{\sin \phi} = \frac{mv_0^2 b}{k} \quad (44.213)$$

Rappelons (cf. chapitre de Trigonométrie) que:

$$\tan(\alpha) = \frac{\sin(2\alpha)}{1 + \cos(2\alpha)} = \frac{1}{\cot(\alpha)} \quad (44.214)$$

Ce qui nous donne alors:

$$\cot\left(\frac{1}{2}\phi\right) = \frac{mv_0^2 b}{k} \quad (44.215)$$

Soit de manière plus détaillée:

$$\cot\left(\frac{\phi}{2}\right) = \frac{4\pi\epsilon_0 mv_0^2 b}{ZQq} \quad (44.216)$$

Cette relation donne l'angle de déviation ϕ en fonction du paramètre d'impact b .

Ce qui nous donne aussi:

$$b = \frac{1}{4\pi\epsilon_0} \frac{ZQq}{mv_0^2} \cot\left(\frac{\phi}{2}\right) \quad (44.217)$$

Bien évidemment, dans les cas scolaires, on pose souvent $Q=q$ ce qui simplifie un peu la lourdeur de la relation mais on perd en généralisation.

Cette équation est appliquée à l'analyse de la déviation d'une particule chargée par les noyaux.

Remarquons que ce résultat n'est valable que pour une force inversement proportionnelle au carré de la distance. Si la force dépend de la distance selon une autre loi, l'angle de déviation satisfait à une autre équation. Les expériences de déviation sont donc très utiles quand nous voulons déterminer la loi de force entrant en jeu dans les interactions entre particules.

Trajectoires de particules alpha

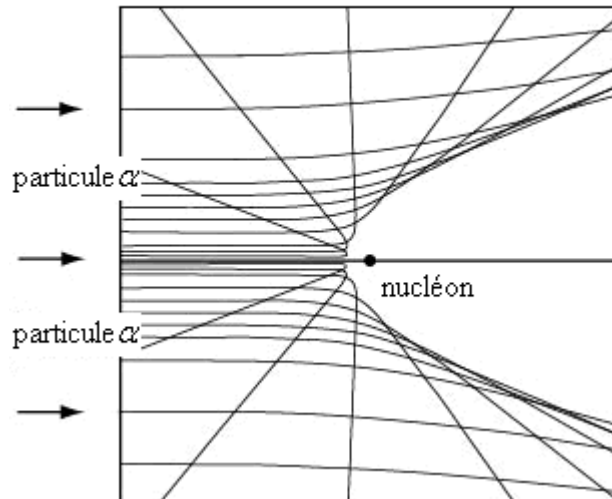


Figure: 44.15 - Représentation de la diffusion de Rutherford de noyaux alpha

Dans les laboratoires de physique nucléaire, on fait des expériences de diffusion en accélérant des électrons, des protons ou d'autres particules au moyen d'un cyclotron, d'un accélérateur de Van de Graaf ou de quelque autre dispositif semblable, et en observant la distribution angulaire des particules déviées.

Il est clair qu'une particule incidente dans une surface définie par un rayon compris entre b et $b + db$ sera respectivement comprise dans l'angle solide de diffusion:

$$d\sigma = 2\pi b db \quad (44.218)$$

avec (cf. chapitre de Trigonométrie):

$$d\Omega = 2\pi \sin \phi d\phi \quad (44.219)$$

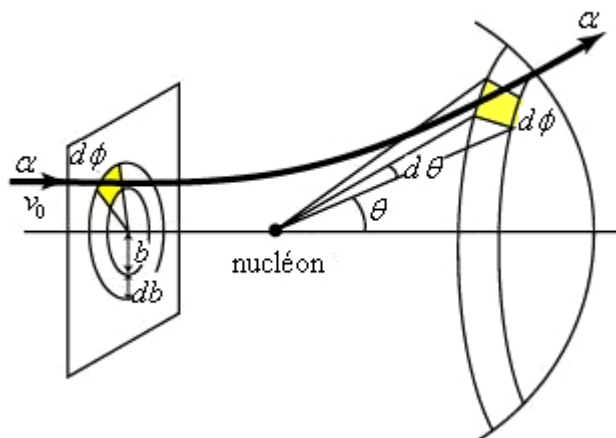


Figure: 44.16 - Représentation de l'angle solide de diffusion

La "**section efficace**" étant définie par:

$$\frac{d\sigma}{d\Omega} = \frac{2\pi b db}{2\pi \sin \phi d\phi} \quad (44.220)$$

Partant de (relation démontrée plus haut):

$$b = \frac{1}{4\pi\epsilon_0} \frac{qQ}{mv_0^2} \cot\left(\frac{\phi}{2}\right) \quad (44.221)$$

et en utilisant la dérivée usuelle suivante démontrée dans le chapitre de Calcul Différentiel Et Intégral:

$$\cot'(x) = -\frac{1}{\sin^2(x)} \quad (44.222)$$

nous avons:

$$\frac{db}{d\phi} = \frac{1}{4\pi\epsilon_0} \frac{qQ}{mv_0^2} \frac{d}{d\phi} \cot\left(\frac{\phi}{2}\right) = -\frac{1}{4\pi\epsilon_0} \frac{qQ}{mv_0^2} \frac{1}{2\sin^2\left(\frac{\phi}{2}\right)} \quad (44.223)$$

soit:

$$db = -\frac{1}{4\pi\epsilon_0} \frac{qQ}{mv_0^2} \frac{1}{2\sin^2\left(\frac{\phi}{2}\right)} d\phi \quad (44.224)$$

Dès lors:

$$bdb = -\left(\frac{1}{4\pi\epsilon_0}\right)^2 \left(\frac{qQ}{mv_0^2}\right)^2 \frac{\cot\left(\frac{\phi}{2}\right)}{2\sin^2\left(\frac{\phi}{2}\right)} d\phi = -\left(\frac{1}{4\pi\epsilon_0}\right)^2 \left(\frac{qQ}{mv_0^2}\right)^2 \frac{\cos\left(\frac{\phi}{2}\right)}{2\sin^3\left(\frac{\phi}{2}\right)} d\phi \quad (44.225)$$

Soit:

$$\begin{aligned} \frac{bdb}{\sin\phi d\phi} &= -\left(\frac{1}{4\pi\epsilon_0}\right)^2 \left(\frac{qQ}{mv_0^2}\right)^2 \frac{1}{2\sin^3\left(\frac{\phi}{2}\right)} \frac{\cos\left(\frac{\phi}{2}\right)}{\sin(\phi)} \\ &= -\left(\frac{1}{4\pi\epsilon_0}\right)^2 \left(\frac{qQ}{mv_0^2}\right)^2 \frac{1}{2\sin^3\left(\frac{\phi}{2}\right)} \frac{\cos\left(\frac{\phi}{2}\right)}{2\sin\left(\frac{\phi}{2}\right)\cos\left(\frac{\phi}{2}\right)} \quad (44.226) \\ &= -\left(\frac{1}{4\pi\epsilon_0}\right)^2 \left(\frac{qQ}{mv_0^2}\right)^2 \frac{1}{4\sin^4\left(\frac{\phi}{2}\right)} \end{aligned}$$

Nous avons alors, en se souvenant que le membre de gauche n'est rien d'autre que la section efficace:

$$\begin{aligned} \frac{d\sigma}{d\Omega} &= -\frac{(qQ)^2}{(4\pi\epsilon_0)^2 m^2 v_0^4} \frac{1}{4 \sin^4\left(\frac{\phi}{2}\right)} = -\frac{(qQ)^2}{(4\pi\epsilon_0)^2 4 \frac{1}{4} m^2 v_0^4} \frac{1}{4 \sin^4\left(\frac{\phi}{2}\right)} \\ &= -\frac{(qQ)^2}{16(4\pi\epsilon_0)^2 E_c^2 \sin^4\left(\frac{\phi}{2}\right)} \end{aligned} \quad (44.227)$$

et il est d'usage de prendre la valeur absolue pour définir la "**section (différentielle) efficace de Rutherford (ou de Coulomb)**":

$$\boxed{\frac{d\sigma}{d\Omega} = \frac{(qQ)^2}{16(4\pi\epsilon_0)^2 E_c^2 \sin^4\left(\frac{\phi}{2}\right)}} \quad (44.228)$$

Nous remarquons plusieurs petites choses intéressantes:

1. Pour un angle d'incidence nul, la section efficace diverge (à cause du sinus au dénominateur)
2. La section efficace décroît selon le carré de l'énergie cinétique de la particule incidente
3. L'expression est valable quelles que soient les charges mises en jeu (positives ou négatives)

À l'aide de la diffusion de Rutherford/Coulomb, Rutherford a pu déterminer une approximation de la taille du noyau de l'atome (bombardement d'une feuille d'or à l'aide de noyaux alpha) comme nous l'avons fait remarquer au début du chapitre de Physique Quantique Corpusculaire. Le raisonnement appliqué est le suivant pour déterminer une borne inférieure du rayon du noyau:

L'énergie totale d'un système en rotation est l'énergie cinétique de translation sommée à l'énergie cinétique de rotation, sommée à l'énergie potentielle. Ce qui nous donne:

$$E = \frac{1}{2} m \dot{r}^2 + \frac{1}{2} m \omega^2 r^2 + U(r) = \frac{1}{2} m \dot{r}^2 + \frac{1}{2} m r^2 \dot{\theta}^2 + U(r) \quad (44.229)$$

en notant L le moment cinétique donné par $L = m r^2 \dot{\theta}$ nous avons:

$$E = \frac{1}{2} m \dot{r}^2 + \frac{L^2}{2 m r^2} + U(r) \quad (44.230)$$

d'où:

$$\dot{r} = \frac{dr}{dt} = \sqrt{\frac{2}{m} \left(E - \frac{L^2}{2 m r^2} - U(r) \right)} = \frac{dr}{d\theta} \frac{d\theta}{dt} \quad (44.231)$$

Il en résulte donc:

$$d\theta = \frac{Ldr}{r^2 \sqrt{2m \left(E - \frac{L^2}{2mr^2} - U(r) \right)}} \quad (44.232)$$

D'où il découle que l'angle associé à deux distances radiales r_1, r_2 est donné par:

$$\Delta\theta = \int_{r_1}^{r_2} \frac{Ldr}{r^2 \sqrt{2m \left(E - \frac{L^2}{2mr^2} - U(r) \right)}} \quad (44.233)$$

La figure ci-dessous montre un processus de collision par un potentiel central $U(r)$. La particule incidente possède une vitesse initiale:

$$v = \sqrt{\frac{2E_c}{m}} \quad (44.234)$$

en $t = -\infty$ avec $z(t = -\infty) = -\infty$ et $y(t = +\infty) = b$ par symétrie à nouveau.

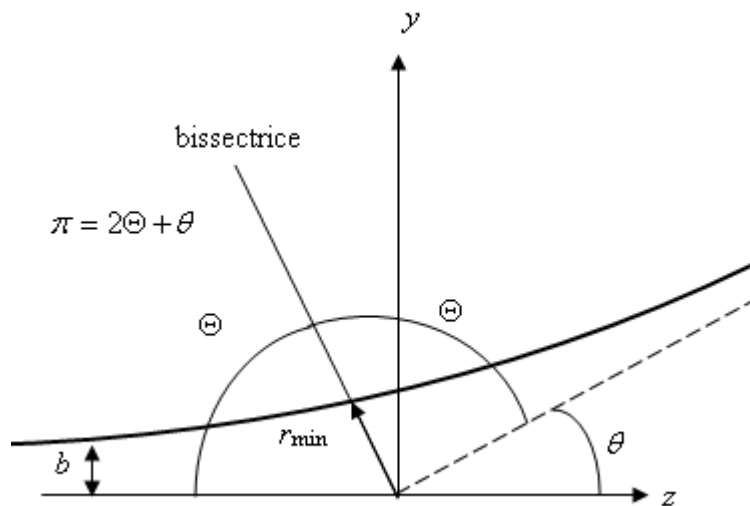


Figure: 44.17 - Approche schématique pour la détermination du rayon de la cible

L'angle Θ est l'angle de déflexion lorsque la particule incidente approche le diffuseur à la distance minimum $r_{\min}$.

Revenons-en à nos équations où le moment cinétique est lié au paramètre d'impact par la relation $L = mvb$ ou encore:

$$L = b\sqrt{2mE_c} \quad (44.235)$$

Nous pouvons donc écrire après simplifications:

$$\begin{aligned}
\Delta\theta &= \int_{r_1}^{r_2} \frac{Ldr}{r^2 \sqrt{2m \left(E - \frac{L^2}{2mr^2} - U(r) \right)}} \Rightarrow \Theta = \int_{r_{\min}}^{+\infty} \frac{b\sqrt{2mE_c}dr}{r^2 \sqrt{2m \left(E - \frac{L^2}{2mr^2} - U(r) \right)}} \\
&= \int_{r_{\min}}^{+\infty} \frac{bdr}{r^2 \frac{1}{\sqrt{2mE_c}} \sqrt{2m \left(E - \frac{L^2}{2mr^2} - U(r) \right)}} = \int_{r_{\min}}^{+\infty} \frac{bdr}{r^2 \sqrt{\frac{2m}{2mE_c} \left(E - \frac{b^2 E_c}{2mr^2} - U(r) \right)}} \\
&= \int_{r_{\min}}^{+\infty} \frac{bdr}{r^2 \sqrt{\frac{E}{E_c} - \frac{b^2}{r^2} - \frac{U(r)}{E_c}}} = \int_{r_{\min}}^{+\infty} \frac{bdr}{r^2 \sqrt{1 - \frac{b^2}{r^2} - \frac{U(r)}{E_c}}} = \int_{r_{\min}}^{+\infty} \frac{bdr}{\sqrt{r^4 - r^2 b^2 - \frac{r^4 U(r)}{E_c}}} \\
&= \int_{r_{\min}}^{+\infty} \frac{bdr}{\sqrt{r^2 \left(r^2 - r \frac{A}{E_c} - b^2 \right)}}
\end{aligned} \tag{44.236}$$

où nous avons posé $E \cong E_c$ (l'énergie de rotation et du potentiel considérés comme négligeables par rapport à l'énergie cinétique) et:

$$U(r) = \frac{1}{4\pi\epsilon_0} \frac{Z_1 Z_2 e^2}{r} = \frac{A}{r} \tag{44.237}$$

La distance minimale d'approche est donc déterminée par le plus grand zéro du dénominateur:

$$r^2 - \frac{A}{E_c} r - b^2 = 0 \tag{44.238}$$

c'est-à-dire (trivial):

$$r_{\min} = \frac{\frac{A}{E_c} + \sqrt{\left(\frac{A}{E_c}\right)^2 + 4b^2}}{2} = \frac{A}{2E_c} + \sqrt{\left(\frac{A}{2E_c}\right)^2 + b^2} \tag{44.239}$$

Nous avons donc:

$$r_{\min} = \frac{A}{2E_c} \left[1 + \sqrt{1 + \left(\frac{E_c b}{A}\right)^2} \right], A = \frac{Z_1 Z_2 e^2}{4\pi\epsilon_0}, E_c = \frac{1}{2} m v_0^2 \tag{44.240}$$

Comme nous le voyons dans cette dernière relation, la particule incidente subira une collision frontale lorsque $b = 0$. Dès lors, la valeur de l'approche maximale est:

$$r_{\min} = \frac{A}{E_c} = \frac{Z_1 Z_2 e^2}{4\pi\epsilon_0 E_c} \tag{44.241}$$

L'expérience de Rutherford permet d'estimer la taille du noyau atomique. En effet, les particules qui ont rebondi sur le noyau avec un angle de diffusion de 180° (nous parlons alors de "rétrodiffusion"), sont celles qui se sont approchées le plus près de ce dernier. Puisque nous avons:

$$\frac{1}{2}mv_0^2 = \frac{1}{4\pi\epsilon_0} \frac{Z_1 Z_2 e^2}{r_{\min}} \quad (44.242)$$

avec une énergie cinétique initiale de 7.7 [MeV], Rutherford trouva pour le rayon de l'atome d'or ($Z=79$) avec des particules alpha ($Z=2$) une valeur de:

$$a_0 = 3 \cdot 10^{-14} [m] \quad (44.243)$$

Ainsi, le noyau n'est pas ponctuel mais de l'ordre de la dizaine de femtomètres.

3.5. RAYONS-X ET GAMMA

La différence fondamentale de ce type de rayonnement, par rapport aux α , β^- , β^+ , est qu'il n'est pas porteur de charge électrique et n'a donc pas d'interaction coulombienne avec le cortège électronique du milieu traversé. Par conséquent, le photon suit un chemin rectiligne sans perte d'énergie jusqu'à ce qu'il rencontre sur sa trace une particule (électron, noyau) où il va subir une interaction modifiant profondément son état.

Le rayonnement gamma est une radiation électromagnétique de haute énergie produite par un phénomène nucléaire, alors que les rayons-X sont des radiations électromagnétiques de haute énergie produites lors de phénomènes atomiques ou moléculaires. Le photon est la particule élémentaire qui est associée à ces ondes électromagnétiques. Les photons gamma et X sont donc de même nature mais d'origines différentes, ils ont donc des propriétés identiques qui dépendent de leur énergie.

Rappelons que:

$$E = h\nu = \frac{hc}{\lambda}, p = \frac{h\nu}{c} = \frac{h}{\lambda} \quad (44.244)$$

En traversant la matière un photon peut interagir avec:

- Un des électrons de l'atome rencontré
- Le noyau de l'atome
- Le champ électrique des particules atomiques chargées
- Le champ mésique des nucléons (interaction forte)

Le résultat de l'interaction peut être schématisé comme suit:

- le photon est dévié en conservant son énergie, il y a alors "**diffusion totale**" de l'énergie et le processus est dit "cohérent" (élastique)
- le photon est dévié et son énergie diminuée, il y a alors "**diffusion partielle**" de l'énergie, l'autre partie est absorbée par la matière, le processus est dit alors "incohérent" (inélastique)

- le photon disparaît, il y a "**absorption (totale)**" de son énergie par la matière.

Nous pouvons démontrer que les caractéristiques macroscopiques de ces interactions dans le cadre d'un faisceau fin et collimaté conduisent à une loi exponentielle d'atténuation du rayonnement photonique dans la matière. Cela signifiant que pour les photons il n'y a pas de parcours fini (!) comme pour les particules chargées; on ne pourra jamais assurer qu'à une distance donnée tous les photons d'un faisceau aient subi une interaction.

Le nombre de particules interagissant avec la matière dépend évidemment de l'intensité I et du type de matière traversée (caractérisée par le "**coefficient d'atténuation linéique**" μ) et de son épaisseur x .

Nous avons:

$$dI = -\mu I dx \quad (44.245)$$

le signe "-" étant là pour mettre en évidence une diminution. Nous résolvons facilement cette équation différentielle (c'est simplement la loi de Beer-Lambert que nous avons déjà étudiée dans le chapitre d'Optique Géométrique):

$$I = I_0 e^{-\mu x} \quad (44.246)$$

avec I_0 l'intensité initiale ou "**débit de fluence**" et μ le coefficient d'atténuation linéique [cm^{-1}] qui tient compte de tous les effets d'atténuation possibles.

Remarque: Souvent dans les tables, nous trouvons le coefficient d'atténuation massique μ/ρ exprimé en [cm^2/g]. Nous avons alors:

$$I = I_0 e^{-\frac{\mu}{\rho} \rho x} \quad (44.247)$$

Dans le cas d'un absorbant contenant plusieurs éléments chimiques homogènement distribués, le coefficient d'atténuation vaut:

$$\mu = \sum_i w_i \mu_i \quad \text{ou} \quad \left(\frac{\mu}{\rho}\right) = \sum_i w_i \left(\frac{\mu_i}{\rho_i}\right) \quad (44.248)$$

où μ est le coefficient d'absorption de l'absorbant, μ_i le coefficient d'absorption de l'élément i , ρ la masse volumique de l'absorbant, ρ_i la masse volumique de l'élément i , w_i étant la fraction massique de l'élément i dans l'absorbant.

Faisons maintenant une approche microscopique:

Soit un faisceau de I [$particules/cm^2 \cdot s$] (où s est l'unité des secondes puisqu'il s'agit d'un flux de particules) frappant perpendiculairement la surface d'un matériau d'épaisseur dx et de densité atomique N [$atomes/cm^3$]. Si nous considérons les particules frappant la surface A , ces dernières peuvent théoriquement rencontrer $N \cdot A \cdot dx$ atomes cibles dans cette couche. Le nombre de particules interagissant sera proportionnel à l'intensité fois ce nombre et nous aurons:

$$dI \cdot A = -\sigma \cdot I \cdot N \cdot A \cdot dx \quad (44.249)$$

où σ est la constante de proportionnalité, appelée "**section efficace microscopique**". Ces unités sont souvent exprimées en "**barn**" ($1 [\text{barn}] = 10^{-24} [\text{cm}^2]$).

Remarques:

R1. La densité atomique N est égale à $\rho N_{Av} \cdot M^{-1}$ où ρ est la densité en $[\text{g}/\text{cm}^3]$ de la cible, N_{Av} le nombre d'Avogadro ($6.022 \cdot 10^{23} [\text{atomes}/\text{mole}]$) et M est la masse molaire de la cible exprimée en $[\text{g}/\text{mole}]$.

R2. Si nous admettons que les centres de diffusion sont les électrons et non pas les atomes cibles, alors il faut remplacer N par $N_e = N \cdot Z$.

D'où nous obtenons:

$$I = I_0 \exp(-NZ \sigma x) = I_0 \exp(-N_e \sigma x) = I_0 \exp\left(-\frac{ZN_{Av} \rho \sigma}{M} x\right) \quad (44.250)$$

En identifiant l'aspect macro et microscopique, nous voyons que μ joue le même rôle que $ZN_{Av} \rho \sigma / M$ et que nous trouvons que la section efficace peut s'écrire comme:

$$\sigma = \frac{\mu}{\rho} \frac{M}{ZN_{Av}} \quad (44.251)$$

et dans l'hypothèse où l'électron constitue une "**sphère d'action**" présentant une surface frontale πr_e^2 , r_e étant le rayon de la sphère d'action alors:

$$\sigma = \pi r_e^2 = \frac{\mu}{\rho} \frac{M}{ZN_{Av}} \quad (44.252)$$

et nous avons:

$$r_e = \sqrt{\frac{\mu}{\rho} \frac{M}{ZN_{Av} \pi}} \quad (44.253)$$

Par définition, nous appelons Coude de Demi-Atténuation (CDA) l'épaisseur du matériau qui divise le débit de fluence I d'un facteur deux. Ainsi:

$$\frac{I_0}{2} = I_0 e^{-\mu x} \Rightarrow \frac{1}{2} = e^{-\mu x} \Rightarrow \ln\left(\frac{1}{2}\right) = -\mu x \Rightarrow x = -\frac{\ln\left(\frac{1}{2}\right)}{\mu} = -\frac{\ln(2)}{\mu} = CDA \quad (44.254)$$

En radioprotection, nous utilisons parfois la notion de couche d'atténuation au dixième TVL (Tenth Value Layer) donnée par:

$$TVL = \frac{\ln(10)}{\mu} \quad (44.255)$$

Nous faisons usage parfois aussi de la "**longueur de relaxation**", qui représente l'épaisseur à partir de laquelle l'intensité d'un faisceau monoénergétique est diminuée d'un facteur e , et qui est donc donnée par:

$$\lambda = \frac{1}{\mu} \quad (44.256)$$

Cette valeur est beaucoup plus utile que les autres, car c'est aussi la distance moyenne à laquelle a lieu la première collision du photon.

Remarque: L'irradiation gamma est anecdotiquement utilisée dans le cadre de la conservation du patrimoine des objets organiques. Effectivement, lors de la découverte des archéologues d'oeuvres ou vestiges anciens, ces derniers sont attaqués par des micro-organismes qui vont détruire ces objets avec le temps. Le rayonnement gamma va permettre, sans détruire les objets, de tuer par irradiation gamma tous ces micro-organismes. L'exemple le plus connu étant l'irradiation de la momie de Toutankhamon pendant 10 heures dans les laboratoires du CEA.

Les causes microscopiques connues de l'atténuation d'un faisceau de photons (neutre au point de vue coulombien) qui méritent notre attention dans le domaine d'énergie des photons gamma ou rayons X sont au nombre de sept:

- Diffusion cohérente de Thomson
- Diffusion cohérente de Rayleigh
- Diffusion cohérente de Delbrück
- Diffusion cohérente de Compton (déjà vue partiellement plus haut)
- Absorption photoélectrique (déjà partiellement vue plus haut)
- Réaction photonucléaire
- Création de paires d'électrons-positrons (déjà partiellement vue plus haut)

Bien que nous puissions à ce jour parler de ces effets, il nous est impossible dans l'état actuel du site de présenter le formalisme mathématique permettant de déterminer la section efficace de chacune de ces diffusions.

3.5.1. CRÉATION PAIRES ÉLECTRON-POSITRON

Au cours de la création de paires, le photon absorbé dans le champ électrique du noyau peut générer une paire électron-positron. Pour que l'interaction puisse avoir lieu, il faut que l'énergie du photon soit supérieure à $2m_0c^2$ (soit environ 1.02 [MeV]), soit l'énergie au repos de la paire électron-positron.

Cet effet est important pour les hautes énergies et les numéros atomiques élevés. Le positron créé est freiné dans la matière tout comme un électron et, en fin de parcours, il s'annihile avec un électron pour donner lieu à deux photons de 0.511 [MeV] (photons d'annihilation) émis presque à 180° (toute la quantité de mouvement est transformée en énergie d'où la valeur de l'angle, ainsi la quantité de mouvement finale est nulle).

La création d'une paire coûte évidemment au moins l'énergie de masse de l'électron et du positron, soit $2m_0c^2$. Le solde d'énergie se répartit ensuite dans l'énergie cinétique des deux particules:

$$h\nu = 2m_0c^2 + E_c^{\beta^+} + E_c^{\beta^-} \quad (44.257)$$

La nécessité de satisfaire simultanément aux conditions de conservation de l'énergie masse et de la quantité de mouvement d'autre part impose à l'effet de matérialisation d'avoir lieu au voisinage d'une particule matérielle qui participe au phénomène. En effet, dans le vide, les deux conditions sont contradictoires ! La quantité de mouvement de chaque électron vaut:

$$p_{\beta^+\beta^-} = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}} v = \frac{E_{\text{tot}}}{c^2} v \quad (44.258)$$

où E_{tot} est l'énergie totale de chacun des électrons, c'est-à-dire:

$$E_{\text{tot}} = m_0c^2 + E_c^{\beta^+\beta^-} \quad (44.259)$$

Le photon d'origine a:

$$h\nu = 2E_{\text{tot}} \text{ et } p_\gamma = \frac{h\nu}{c} \quad (44.260)$$

que nous introduisons dans l'équation de conservation de l'énergie et avec l'aide la relation donnant $p_{\beta^+\beta^-}$ nous avons:

$$p_\gamma c = 2E_{\text{tot}} = 2p_{\beta^+\beta^-} \left(\frac{c^2}{v} \right) \quad (44.261)$$

ce qui montre bien que par le terme $c^2/v > 1$ que le noyau doit emporter une partie de la quantité de mouvement puisque:

$$p_{\beta^+\beta^-} < p_{\beta^+\beta^-} \left(\frac{c^2}{v} \right) \quad (44.262)$$

4. MODÈLE NUCLÉAIRE "GOUTTE LIQUIDE"

En ce début de 21ème siècle il n'existe pas de théorie générale qui sous-tende l'ensemble des propriétés expérimentalement découvertes relatives aux noyaux. Le noyau a de multiples facettes qui à ce jour ne sont pas réconciliables dans une théorie unique (la chromodynamique quantique n'arrivant pas à modéliser le noyau).

Les informations les plus précises sur les rayons des noyaux nucléaires et plus généralement sur la densité de charge des noyaux proviennent des mesures de diffusion d'électrons. Ces derniers considérés à ce jour comme particules élémentaires, ne subissent pas la force nucléaire forte garante de la cohésion du noyau et peuvent être considérés dans les traitements théoriques comme ponctuels.

Voici en gros l'état de nos connaissances sur les noyaux en ce début du 21ème siècle:

1. Le potentiel d'interaction nucléaire nucléon-nucléon est a priori attractif à l'échelle du noyau mais parfois aussi répulsif si la distance devient trop petite entre ses constituants.
2. L'interaction nucléaire est de faible portée et présente un phénomène de saturation comme si chaque nucléon n'était lié directement qu'à ses seuls voisins directs (après s'être lié à quelques nucléons sa possibilité de liaison est épuisée), contrairement à la force électrostatique.
3. La force nucléaire est a priori indépendante de la charge. Elle agit aussi bien entre neutrons, qu'entre neutrons et protons et protons-protons.

Dans le modèle naïf que nous allons étudier ici (qui explique assez bien l'énergie de liaison des nucléons), le noyau sera assimilable à une goutte liquide nucléaire en première approximation sphérique et incompressible, de densité volumique constante donc:

$$\frac{A}{\frac{4}{3}\pi R^3} = c^{te} \quad (44.263)$$

où A représente toujours dans le cadre ce chapitre le nombre de masse.

Ce qui entraîne:

$$R = \left(\frac{1}{\frac{4}{3}\pi c^{te}} \right)^{1/3} A^{1/3} := R_0 A^{1/3} \quad (44.264)$$

avec:

$$R_0 \cong 1.17 \cdot 10^{-15} [m] \quad (44.265)$$

Cependant la modélisation échoue déjà ici car les isotopes du mercure présentent des rayons plus grands que ceux prédits dans le cadre du modèle et ce avec de fortes variations (ces variations étant inconciliables avec l'évolution régulière d'une goutte en fonction de son nombre de constituants). Il y a même des cas où en enlevant des neutrons, le rayon du noyau augmente de façon très importante.

Passons maintenant en revue les différentes énergies en présence. Nous construirons chemin faisant la formule semi-empirique de von Weizsäcker.

4.1. ÉNERGIE DE LIAISON EN VOLUME

L'interaction nucléaire forte confère une "**énergie de liaison en volume**" de la forme:

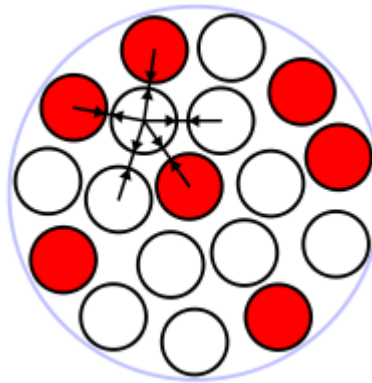
$$E_V = \alpha_V A \quad (44.266)$$

où est α_V une constante déterminée (dans un premier temps...) expérimentalement comme valant:

$$\alpha_V \cong 14 \text{ [MeV]} \quad (44.267)$$

et nous démontrerons comment la déterminer théoriquement un peu plus bas lors de notre détermination de l'énergie de Pauli.

La représentation schématique de l'idée "énergie de liaisons en volume" (avec voisins directement connexes) donne:



où les couleurs (rouge, blanc) ne sont là que pour illustrer l'ensemble des énergies de liaison que nous allons aussi voir par la suite. Il s'agit cependant bien évidemment indistinctement de nucléons (protons et neutrons).

4.2. ÉNERGIE DE LIAISON SUPERFICIELLE

Les nucléons proches de la surface externe du noyau sont moins liés via l'interaction forte que les nucléons situés en profondeur puisqu'ils ont moins de voisins directs. Il faut donc se départir de l'idée que chaque constituant possède la même énergie de liaison en volume et soustraire de la totalité de celle-ci une "**énergie de liaison superficielle**" proportionnelle à la surface du noyau qui est:

$$S = 4\pi R^2 = 4\pi \left(R_0 A^{1/3} \right)^2 = 4\pi R_0^2 A^{2/3} \quad (44.268)$$

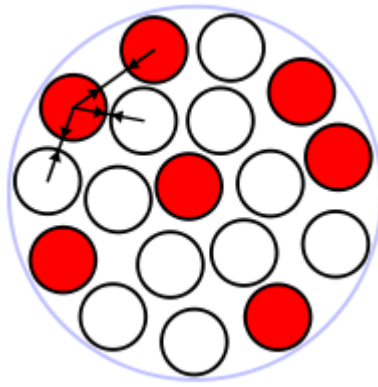
Nous avons alors pour l'énergie superficielle:

$$E_S = -\alpha_S A^{2/3} \quad (44.269)$$

où est α_S une constante déterminée expérimentalement comme valant:

$$\alpha_S \cong 17 \text{ [MeV]} \quad (44.270)$$

La représentation schématique de l'idée "énergie de liaison en surface" (avec voisins directement connexes) donne:



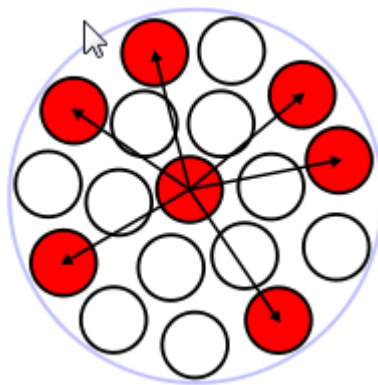
C'est ce terme qui fait penser à une goutte liquide. En effet, dans une goutte liquide, les forces sont également supposées à courte portée (forces de Van der Waals), et donc saturent ce qui provoque immédiatement une tension superficielle.

Nous avons donc jusqu'à maintenant l'énergie potentielle totale de liaison du noyau qui est donnée par une partie de la "**formule semi-empirique de von Weizsäcker**" (ou "**formule de Bethe-Weizsäcker**"):

$$E_n = E_V + E_S = a_V A - a_S A^{2/3} \quad (44.271)$$

4.3. ÉNERGIE DE RÉPULSION ÉLECTROSTATIQUE

Il nous faut aussi prendre en compte l'habituelle "**énergie de liaison électrostatique**" qui résulte de la force de répulsion électrostatique entre les protons:



Comme elle est répulsive, elle diminue l'énergie de liaison (donc cela sera un terme négatif). Pour obtenir l'énergie potentielle électrique, rappelons que nous avons déjà démontré dans le chapitre de Mécanique Classique que pour l'énergie gravitationnelle nous avons:

$$E_g = -\frac{3}{5} \frac{GM^2}{R} \quad (44.272)$$

Il vient alors immédiatement pour le cas électrostatique (coulombien):

$$E_{p,C} = -\frac{3}{5} \frac{1}{4\pi\epsilon_0} \frac{(Ze)^2}{R} = -\frac{3e^2}{20\pi\epsilon_0} \frac{Z^2}{R} = -\frac{3e^2}{20\pi\epsilon_0 R_0} \frac{Z^2}{A^{1/3}} = -a_C \frac{Z^2}{A^{1/3}} \quad (44.273)$$

avec la constante calculable:

$$a_C \cong 0.72 \text{ [MeV]} \quad (44.274)$$

Nous avons donc jusqu'à maintenant l'énergie potentielle totale de liaison du noyau qui est donnée par une partie de la formule semi-empirique de von Weizsäcker:

$$E_n = E_V + E_S + E_{p,C} = a_V A - a_S A^{2/3} - a_C \frac{Z^2}{A^{1/3}} \quad (44.275)$$

4.4. ÉNERGIE D'ASYMÉTRIE (ÉNERGIE DE PAULI)

Ce terme énergétique s'inspire du modèle du noyau basé sur le gaz de Fermi où l'on considère le noyau comme un ensemble de A nucléons (quasi)libres enfermés dans une boîte rectangulaire ayant les dimensions du noyau et respectant des règles de quantification (ce que l'expérience semble mettre en évidence). La physique quantique n'est donc pas propre au monde atomique mais aussi à la structure du noyau (on pouvait s'en douter...).

Nous avons alors démontré dans le chapitre d'Électrocinétique (lors de notre étude de la théorie des bandes) que sous certaines conditions (fortes!) bien précises, le nombre d'états maximum dans un volume sphérique était donné pour les fermions par:

$$n(k) = \frac{k_F^3}{3\pi^2} V \quad (44.276)$$

où pour rappel k_F est le nombre d'onde de Fermi qu'il est plus d'usage dans le domaine nucléaire d'écrire d'une autre façon en utilisant la relation de de Broglie (cf. chapitre de Physique Quantique Ondulatoire):

$$p = \hbar k \quad (44.277)$$

Soit:

$$n(p) = \frac{p_F^3}{3\pi^2 \hbar^3} V \quad (44.278)$$

Nous pouvons alors avoir un nombre de neutrons (N) et un nombre de protons (Z) respectivement égaux au maximum à:

$$N = \frac{(p_F^n)^3}{3\pi^2 \hbar^3} V \quad Z = \frac{(p_F^p)^3}{3\pi^2 \hbar^3} V \quad (44.279)$$

Connaissant l'expression du volume d'un noyau, sous l'hypothèse de modélisation par une goutte liquide sphérique et incompressible, nous avons explicitement:

$$n = n(p) = \frac{p_F^3}{3\pi^2 \hbar^3} V = \frac{p_F^3}{3\pi^2 \hbar^3} \frac{4}{3} \pi R^3 = \frac{4p_F^3}{9\pi \hbar^3} (R_0 A^{1/3})^3 = \frac{4p_F^3 A}{9\pi \hbar^3} R_0^3 \quad (44.280)$$

Nous avons donc in extenso:

$$p_F = \left(\frac{9\pi \hbar^3}{4A} \frac{n}{R_0^3} \right)^{1/3} = \left(\frac{9\pi}{4A} n \right)^{1/3} \frac{\hbar}{R_0} \quad (44.281)$$

En assumant que:

$$n = Z = N = A / 2 \quad (44.282)$$

Nous avons alors pour la quantité de mouvement (du nucléon considéré comme (quasi)libre dans le noyau... par construction des hypothèses de la théorie des bandes):

$$p = p_F^n = p_F^p = \left(\frac{9\pi A}{4A} \frac{A}{2} \right)^{1/3} \frac{\hbar}{R_0} = \left(\frac{9\pi}{8} \right)^{1/3} \frac{\hbar}{R_0} \quad (44.283)$$

Connaissant expérimentalement R_0 , nous en extrayons une valeur numérique. De celle-ci, nous pouvons en tirer, en prenant la formulation classique de l'énergie (donc non relativiste), que le niveau d'état d'énergie maximum (niveau de fermi du noyau) est alors:

$$E_F = \frac{1}{2} m v^2 = \frac{1}{2} \frac{p_F^2}{m} \cong 33 \text{ [MeV]} \quad (44.284)$$

Ceci étant fait, calculons maintenant l'énergie cinétique moyenne par nucléon. Nous avons alors:

$$\begin{aligned} \langle E_c \rangle &= \frac{\sum E_c}{N} = \frac{\int_0^{E_F} E_c dn}{\frac{p_F^3}{3\pi^2 \hbar^3} V} = \frac{\int_0^{E_F} E_c \frac{dn}{dp} dp}{\frac{p_F^3}{3\pi^2 \hbar^3} V} = \frac{\int_0^{p_F} E_c \left(\frac{3p^2}{3\pi^2 \hbar^3} V \right) dp}{\frac{p_F^3}{3\pi^2 \hbar^3} V} = \frac{3 \int_0^{p_F} E_c p^2 dp}{p_F^3} \\ &= \frac{3 \int_0^{p_F} \frac{p^2}{2m} p^2 dp}{p_F^3} = \frac{3}{2m} \frac{\int_0^{p_F} p^4 dp}{p_F^3} = \frac{3}{2m} \frac{1}{5} \frac{p_F^5}{p_F^3} = \frac{3}{5} \frac{p_F^2}{2m} \cong 20 \text{ [MeV]} \end{aligned} \quad (44.285)$$

L'énergie cinétique totale du noyau est alors (en approximant la masse du neutron comme étant égale à celle du proton):

$$E_c = N \cdot \langle E_c^n \rangle + Z \cdot \langle E_c^p \rangle = N \frac{3}{5} \frac{(p_F^n)^2}{2m_n} + Z \frac{3}{5} \frac{(p_F^p)^2}{2m_p} \quad (44.286)$$

Or, comme nous avons démontré plus haut que:

$$N = \frac{(p_F^n)^3}{3\pi^2 \hbar^3} V \quad Z = \frac{(p_F^p)^3}{3\pi^2 \hbar^3} V \quad (44.287)$$

Il vient:

$$\begin{aligned} E_c &= N \frac{3}{5} \frac{\left(\frac{N 3\pi^2 \hbar^3}{V} \right)^{2/3}}{2m_n} + Z \frac{3}{5} \frac{\left(\frac{Z 3\pi^2 \hbar^3}{V} \right)^{2/3}}{2m_p} = \frac{3\hbar^2}{10} \left(\frac{3\pi^2}{V} \right)^{2/3} \left[\frac{N^{5/3}}{m_n} + \frac{Z^{5/3}}{m_p} \right] \\ &= \frac{3\hbar^2}{10} \left(\frac{3\pi^2}{\frac{4}{3}\pi (R_0 A^{1/3})^3} \right)^{2/3} \left[\frac{N^{5/3}}{m_n} + \frac{Z^{5/3}}{m_p} \right] = \frac{3\hbar^2}{10} \left(\frac{3\pi^2}{\frac{4}{3}\pi R_0^3 A} \right)^{2/3} \left[\frac{N^{5/3}}{m_n} + \frac{Z^{5/3}}{m_p} \right] \\ &= \frac{3\hbar^2}{10 R_0^2} \left(\frac{9\pi}{4A} \right)^{2/3} \left[\frac{N^{5/3}}{m_n} + \frac{Z^{5/3}}{m_p} \right] \end{aligned} \quad (44.288)$$

Et si nous posons:

$$m_n \cong m_p \cong M \quad (44.289)$$

Nous avons alors:

$$E_c = \frac{3\hbar^2}{10 M R_0^2} \left(\frac{9\pi}{4} \right)^{2/3} \left[\frac{N^{5/3} + Z^{5/3}}{A^{2/3}} \right] \quad (44.290)$$

Nous savons que:

$$E_c = E_T - E_p \quad (44.291)$$

Nous allons chercher à obtenir une relation similaire en faisant une approximation astucieuse. Pour cela, en se rappelant que $A = Z + N$, nous allons considérer le rapport:

$$I = \frac{N - Z}{A} = \frac{A - 2Z}{A} = 1 - 2 \frac{Z}{A} \quad (44.292)$$

qui sera supposé petit... Nous avons alors:

$$N^{5/3} + Z^{5/3} = (A - Z)^{5/3} + Z^{5/3} = A^{5/3} \left[\left(1 - \frac{Z}{A}\right)^{5/3} + \left(\frac{Z}{A}\right)^{5/3} \right] \quad (44.293)$$

Nous avons aussi:

$$\frac{Z}{A} = \frac{1 - I}{2} \quad (44.294)$$

Il vient alors:

$$\begin{aligned} N^{5/3} + Z^{5/3} &= A^{5/3} \left[\left(1 - \frac{Z}{A}\right)^{5/3} + \left(\frac{Z}{A}\right)^{5/3} \right] = A^{5/3} \left[\left(1 - \frac{1 - I}{2}\right)^{5/3} + \left(\frac{1 - I}{2}\right)^{5/3} \right] \\ &= A^{5/3} \left[\left(\frac{1 + I}{2}\right)^{5/3} + \left(\frac{1 - I}{2}\right)^{5/3} \right] = \left(\frac{A}{2}\right)^{5/3} \left[(1 + I)^{5/3} + (1 - I)^{5/3} \right] \end{aligned} \quad (44.295)$$

Le développement en série de Taylor en I au deuxième ordre donne:

$$\begin{aligned} (1 + I)^{5/3} &\cong 1 + \frac{5}{3}I + \frac{5}{9}I^2 + \dots \\ (1 - I)^{5/3} &\cong 1 - \frac{5}{3}I + \frac{5}{9}I^2 - \dots \end{aligned} \quad (44.296)$$

et donc:

$$N^{5/3} + Z^{5/3} \cong \left(\frac{A}{2}\right)^{5/3} 2 \left[1 + \frac{5}{9}I^2 \right] = \left(\frac{A}{2}\right)^{5/3} 2 \left[1 + \frac{5}{9} \frac{(N - Z)^2}{A^2} \right] \quad (44.297)$$

Nous avons alors:

$$\begin{aligned}
E_c &= \frac{3\hbar^2}{10MR_0^2} \left(\frac{9\pi}{4}\right)^{2/3} \frac{1}{A^{2/3}} \left(\frac{A}{2}\right)^{5/3} 2 \left[1 + \frac{5(N-Z)^2}{9A^2}\right] \\
&= \frac{3\hbar^2}{5MR_0^2} \left(\frac{9\pi}{4}\right)^{2/3} \frac{1}{A^{5/3}} \left(\frac{A}{2}\right)^{5/3} \left[A + \frac{5(N-Z)^2}{9A}\right] \\
&= \frac{3\hbar^2}{5MR_0^2} \left(\frac{9\pi}{4}\right)^{2/3} \left(\frac{1}{2}\right)^{5/3} \left[A + \frac{5(N-Z)^2}{9A}\right] \\
&= \frac{3\hbar^2}{5MR_0^2} (9\pi)^{2/3} \left(\frac{1}{2^2}\right)^{2/3} \frac{1}{2^{5/3}} \left[A + \frac{5(N-Z)^2}{9A}\right] \\
&= \frac{3\hbar^2}{5MR_0^2} (9\pi)^{2/3} \frac{1}{2^{4/3}} \frac{1}{2^{5/3}} \left[A + \frac{5(N-Z)^2}{9A}\right] \\
&= \frac{3\hbar^2}{5MR_0^2} \frac{(9\pi)^{2/3}}{2^3} \left[A + \frac{5(N-Z)^2}{9A}\right] = a_V A + a_A \frac{(N-Z)^2}{A} = E_T - E_P
\end{aligned} \tag{44.298}$$

Ainsi nous avons fait d'une pierre deux coups: d'une part nous avons identifié la valeur du coefficient du terme d'énergie de liaison volumique a_V et d'autre part nous en déduisons le terme d'énergie d'asymétrie:

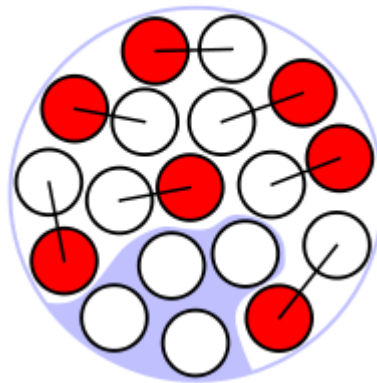
$$E_P = -a_A \frac{(N-Z)^2}{A} = E_A \tag{44.299}$$

Nous avons alors un terme d'énergie potentielle qui apparaît sous la forme:

$$E_A = -a_A \frac{(N-Z)^2}{A} \tag{44.300}$$

avec:

$$a_A \cong 21 [MeV]$$



Puisque ce terme est nul lorsque le nombre de neutrons est égal au nombre de protons, nous comprenons alors un peu mieux l'origine de la vallée de stabilité.

Nous avons donc jusqu'à maintenant l'énergie potentielle totale de liaison du noyau qui est donnée par une partie de la formule semi-empirique de von Weizsäcker:

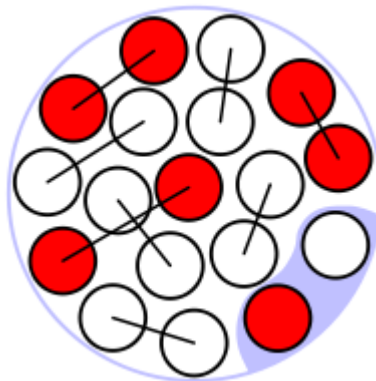
$$E_n = E_V + E_S + E_{p,C} + E_A = a_V A - a_S A^{2/3} - a_C \frac{Z^2}{A^{1/3}} - a_A \frac{(N-Z)^2}{A} \quad (44.301)$$

4.5. ÉNERGIE DE PAIRE (ÉNERGIE D'APPARIEMENT)

Une étude systématique des noyaux montre qu'ils sont plus stables quand ils sont constitués d'un nombre pair de neutrons ou de protons. Empiriquement, nous écrivons ce fait en soustrayant l'énergie de paire suivante:

$$E_p = -\frac{\delta}{\sqrt{A}} \quad (44.302)$$

où δ vaut -11.2 [MeV] si Z et N sont pairs, 0 si A est impair et 11.2 [MeV] si Z et N sont impairs.



Nous avons alors au final l'énergie potentielle totale de liaison du noyau qui est donnée par la formule semi-empirique de von Weizsäcker:

$$E_n = E_V + E_S + E_{p,C} + E_A + E_p = a_V A - a_S A^{2/3} - a_C \frac{Z^2}{A^{1/3}} - a_A \frac{(N-Z)^2}{A} - \frac{\delta}{\sqrt{A}} \quad (44.303)$$

Que l'on retrouve aussi sous la forme suivante:

$$E_n = \underbrace{a_V A}_{\text{volume}} - \underbrace{a_S A^{2/3}}_{\text{surface}} - \underbrace{a_C \frac{Z^2}{A^{1/3}}}_{\text{coulomb}} - \underbrace{a_A \frac{(A-2Z)^2}{A}}_{\text{Asymétrie}} - \underbrace{a_P A^{-1/2}}_{\text{Paire}} \quad (44.304)$$

L'énergie de masse du noyau peut alors s'écrire:

$$M_A c^2 = m_n(A-Z)c^2 + Zm_p c^2 - \left(a_V A - a_S A^{2/3} - a_C \frac{Z^2}{A^{1/3}} - a_A \frac{(A-2Z)^2}{A} - \frac{\delta}{\sqrt{A}} \right) \quad (44.305)$$

Remarque: Les valeurs numériques des constantes ne sont à ce jour pas encore normalisées et nous pouvons trouver dans la littérature plusieurs jeux de valeurs différentes.

Signalons que le modèle de la goutte liquide échoue aussi à expliquer que les éléments de fissions ne soient pas de taille symétrique (le modèle de goutte liquide privilégiant une fission en deux noyaux de même taille).

Nous avons pour le modèle théorique ci-dessus la représentation graphique correspondante:

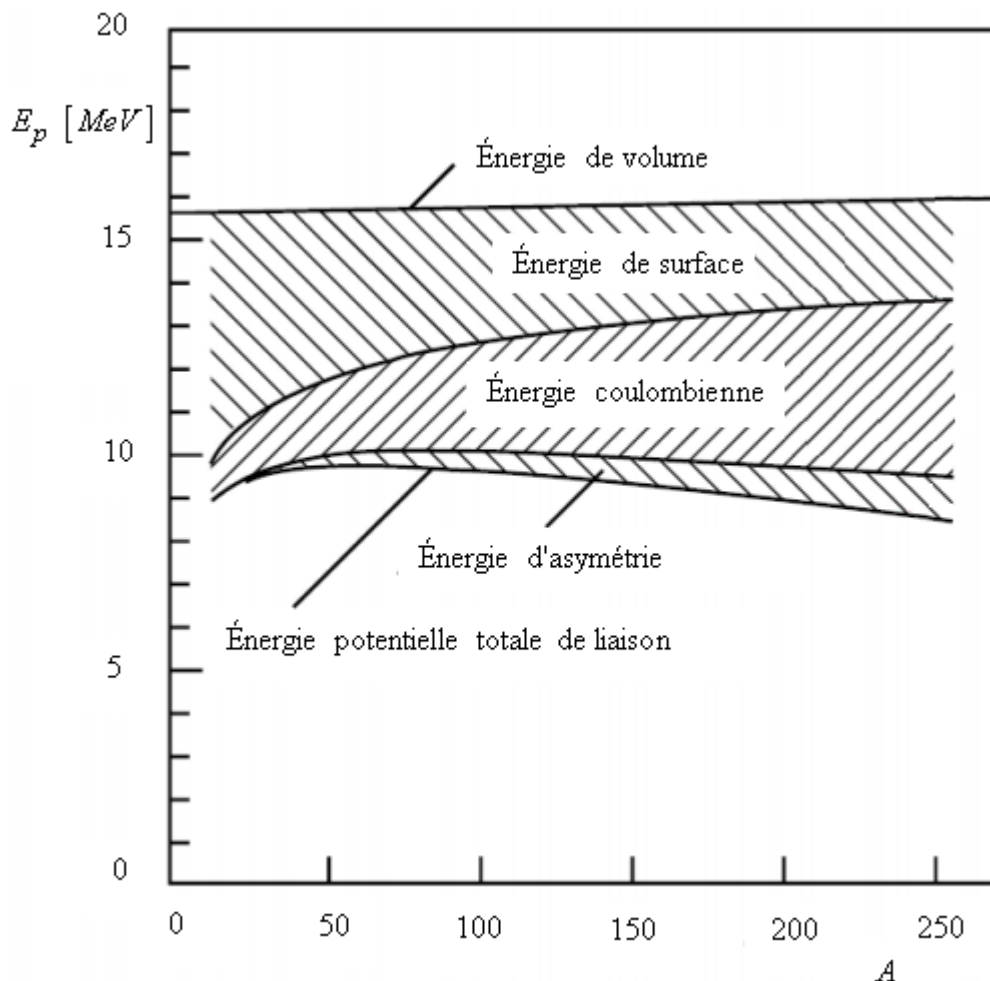


Figure: 44.18 - Représentation schématisée de la formule semi-empirique de von Weizsäcker

Nous voyons ci-dessus que l'énergie de liaison moyenne peut être considérée très approximativement comme constante. Ceci peut être interprété aussi comme une force exercée par un nombre limité de partenaires. Nous parlons alors de "**force saturante**". Par force saturante nous entendons que pour une force donnée, il existe une limite au nombre de nucléons mis côte à côte à partir de laquelle l'ajout d'un nucléon ne fait plus qu'apporter une énergie de liaison supplémentaire constante. Ce sont donc les proches voisins qui amènent la force à son niveau, l'arrivée de nouveaux voisins ne faisant plus par la suite que soutenir la valeur moyenne atteinte. C'est la raison pour laquelle nous disons que l'énergie de volume ne se calcule qu'avec les voisins directs et que le modèle est aussi assimilable à une goutte liquide (les molécules d'une goutte d'eau n'étant sensibles qu'aux molécules directement voisines).

Cette approche permet aussi d'expliquer l'augmentation de l'énergie de liaison par nucléon pour les faibles masses. Effectivement, considérons une force dite de type "**F2**". Nous avons alors la construction des noyaux grâce à ce type de géométrie:

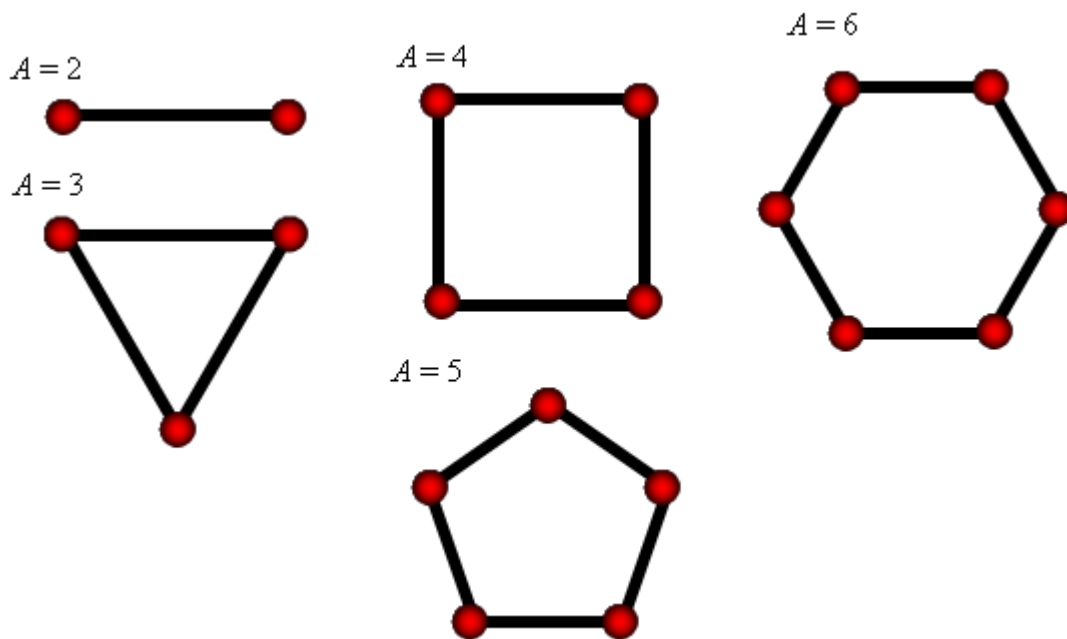
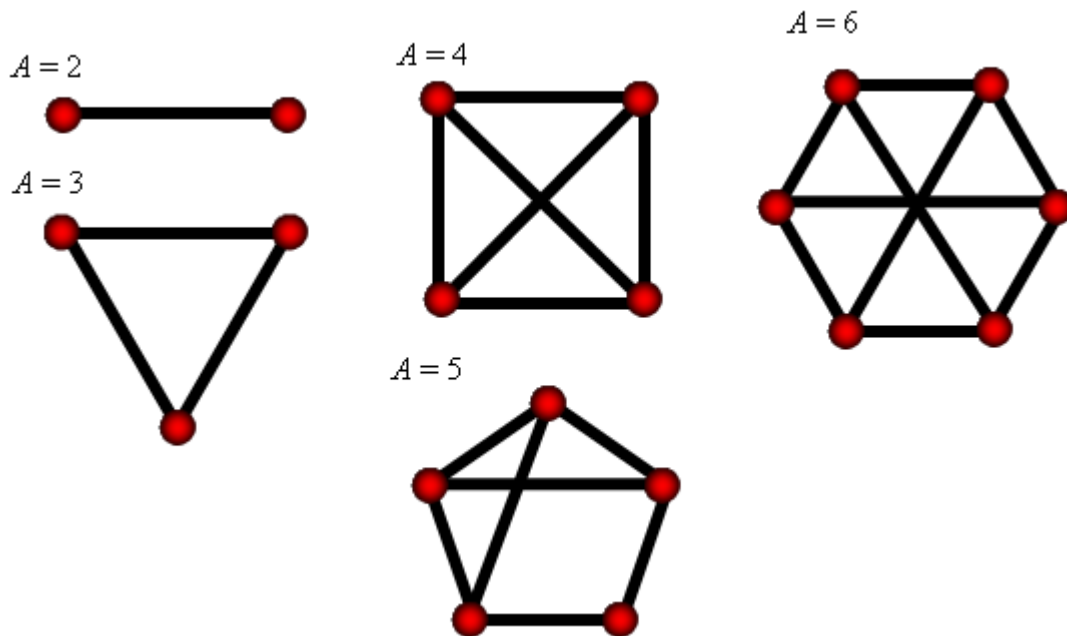


Figure: 44.19 - Représentation schématique de la force saturante de type F2

Chaque nucléon a deux liaisons (excepté le cas $A = 2$ ce qui explique que l'énergie de liaison augmente pour les petits A) et épuise les possibilités de la force de type F2. Cette limitation n'impose aucune restriction sur la taille des objets à construire et nous pouvons imaginer des édifices aussi gros que nous voulons, stabilisés par ce type de force.

Si nous calculons l'énergie totale de liaison, nous aurons $1E_P$ pour $A = 2$, où E_P est l'énergie potentielle de liaison apportée par une liaison. De même, nous aurons $3E_P$ pour $A = 3$, $4E_P$ pour $A = 4$, $5E_P$ pour $A = 5$, etc. L'énergie de liaison par nucléon sera alors $E_P/2$ pour $A = 2$ puis E_P pour toutes les autres masses, pour une force constante F2. Donc cette force sature au-delà de $A = 2$.

Si nous prenons une force de type F3, nous aurons:

Figure: 44.20 - Représentation schématique de la force saturante de type F_3

Si nous calculons l'énergie totale de liaison, nous aurons $1E_P$ pour $A = 2$, où E_P reste l'énergie potentielle de liaison apportée par une liaison. De même, nous aurons $3E_P$ pour $A = 3$, $6E_P$ pour $A = 4$, $7E_P$ pour $A = 5$, etc.

Remarque: Notons que pour $A = 5$ seules 2 liaisons peuvent partir du dernier sommet du pentagone sans quoi un autre sommet de celui-ci compterait 4 liaisons et non 3 (force F_3).

La valeur asymptotique de l'énergie de liaison par nucléon sera alors de $1.5E_P$ pour toutes les masses supérieures à $A = 3$, pour une force constante F_3 . Donc cette force sature au-delà de $A = 3$.

Nous pouvons continuer avec des noyaux incluant de plus en plus de nucléons.

Dans le cas particulier d'une force qui pourrait interagir avec tous autres nucléons environnants, il y aura comme nous l'avons démontré dans le chapitre de Théorie Des Graphes pour des graphes complets:

$$\frac{A(A-1)}{2} \quad (44.306)$$

liaisons au total et donc un comportement en $(A-1)/2$ dans l'énergie de liaison moyenne (puisque'il s'agit simplement de diviser par A le nombre de liaisons). Donc dans le cas classique, l'énergie de liaison ne ferait qu'augmenter ce qui n'est pas compatible avec l'expérience.

Exemple:

Voyons quelques applications de ce modèle en commençant par examiner les prédictions sur la fission de l'uranium ${}_{92}^{236}\text{U}$ en deux sous-produits égaux en masse et en charge:

$${}^A_ZX \rightarrow 2 {}^{A/2}_{Z/2}X \quad (44.307)$$

et ce en utilisant le modèle de la goutte liquide en négligeant les termes d'énergie d'asymétrie et d'appariement.

Pour cela, rappelons que:

$$E_p(A, Z) = a_V A - a_S A^{2/3} - a_C \frac{Z^2}{A^{1/3}} - a_A \frac{(A - 2Z)^2}{A} - \frac{\delta}{\sqrt{A}} \quad (44.308)$$

et nous prendrons comme valeurs des constantes:

$$\begin{aligned} a_V &\cong 15.85 [MeV] \\ a_S &\cong 18.34 [MeV] \\ a_C &\cong 0.71 [MeV] \\ a_A &\cong 23.21 [MeV] \\ \delta &\cong 12 [MeV] \end{aligned} \quad (44.309)$$

Évaluons la différence d'énergie entre d'une part les 2 noyaux issus de la fission et d'autre part le noyau de départ.

Nous avons alors:

$$\begin{aligned} \Delta E &= 2M\left({}^{A/2}_{Z/2}X_f\right)c^2 - M\left({}^A_ZX_m\right)c^2 \\ &\cong 2\left(a_V A_f - a_S A_f^{2/3} - a_C \frac{Z_f^2}{A_f^{1/3}}\right) - \left(a_V A_m - a_S A_m^{2/3} - a_C \frac{Z_m^2}{A_m^{1/3}}\right) \end{aligned} \quad (44.310)$$

comme:

$$2a_V A_f - a_V A_m = a_V (2A_f - A_m) = a_V (2A_f - 2A_f) = 0 \quad (44.311)$$

Le développement se simplifie en:

$$\begin{aligned}
\Delta E &= 2 \left(-a_S A_f^{2/3} - a_C \frac{Z_f^2}{A_f^{1/3}} \right) - \left(-a_S A_m^{2/3} - a_C \frac{Z_m^2}{A_m^{1/3}} \right) \\
&= a_S \left(A_m^{2/3} - 2 A_f^{2/3} \right) + a_C \left(\frac{Z_m^2}{A_m^{1/3}} - 2 \frac{Z_f^2}{A_f^{1/3}} \right) = \\
&= a_S \left(A_m^{2/3} - 2 \left(\frac{A_m}{2} \right)^{2/3} \right) + a_C \left(\frac{Z_m^2}{A_m^{1/3}} - 2 \frac{\left(\frac{Z_m}{2} \right)^2}{\left(\frac{A_m}{2} \right)^{1/3}} \right) \\
&= a_S A_m^{2/3} \left(1 - 2 \cdot 2^{-2/3} \right) + a_C \frac{Z_m^2}{A_m^{1/3}} \left(1 - 2 \frac{\left(\frac{1}{2} \right)^2}{\left(\frac{1}{2} \right)^{1/3}} \right) \\
&= a_S A_m^{2/3} \left(1 - 2^{1/3} \right) + a_C \frac{Z_m^2}{A_m^{1/3}} \left(1 - 2 \cdot \frac{2^{-2}}{2^{-1/3}} \right) = a_S A_m^{2/3} \left(1 - 2^{1/3} \right) + a_C \frac{Z_m^2}{A_m^{1/3}} \left(1 - 2^{-2/3} \right)
\end{aligned} \tag{44.312}$$

S'il y a fission, nous aurons $\Delta E \geq 0$, donc:

$$a_S A_m^{2/3} \left(1 - 2^{1/3} \right) \geq -a_C \frac{Z_m^2}{A_m^{1/3}} \left(1 - 2^{-2/3} \right) \tag{44.313}$$

Et donc:

$$\frac{a_S \left(1 - 2^{1/3} \right)}{a_C \left(1 - 2^{-2/3} \right)} \geq - \frac{Z_m^2}{A_m} \tag{44.314}$$

ou réarrangé:

$$\frac{Z_m^2}{A_m} \geq - \frac{a_S \left(1 - 2^{1/3} \right)}{a_C \left(1 - 2^{-2/3} \right)} \tag{44.315}$$

Soit en mettant les constantes, cela donne:

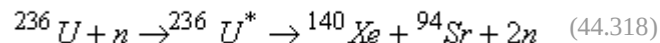
$$\frac{Z_m^2}{A_m} \geq 18 \tag{44.316}$$

Il s'agit donc de l'inégalité devant être satisfaite pour que comme énoncé dans l'exemple, la fission génère des produits identiques en masse et en charge.

Or, pour ${}^{236}_{92}\text{U}$ nous avons:

$$\frac{Z^2}{A_m} = \frac{92^2}{236} = 36 > 18 \quad (44.317)$$

Donc d'après le résultat ci-dessus ${}^{236}_{92}\text{U}$ peut fissionner de manière symétrique mais ce n'est pas le cas dans l'expérience, car dans la réalité, nous avons:



Il y a cependant une autre approche théorique possible qui donne un résultat plus en accord avec l'expérience. Effectivement, nous pouvons imaginer qu'un noyau peut subir une fission si la force dérivée de l'énergie de surface est exactement compensée par la force de Coulomb.

Au final, nous comparerons le rapport Z^2/A_m obtenu avec le résultat précédent.

Rappelons d'abord que nous avons vu dans le chapitre d'Électrostatique (entre autres...) que la force électrostatique dérive du potentiel électrostatique:

$$F = \left| -\frac{dU}{dR} \right| = \left| -\frac{1}{e} \frac{dE_P}{dR} \right| \quad (44.319)$$

où le rayon R a dans notre cas la valeur du rayon nucléaire avec pour rappel:

$$R = R_0 A^{1/3} \quad (44.320)$$

Nous avons alors:

$$\begin{aligned} U_S(R) &= -\frac{1}{e} a_S A^{2/3} = -\frac{1}{e} a_S \left(\frac{R}{R_0} \right)^2 = -\frac{1}{e} \frac{a_S}{R_0^2} R^2 \\ U_C(R) &= -\frac{1}{e} a_C Z^2 A^{-1/3} = -\frac{1}{e} a_C Z^2 \left(\frac{R}{R_0} \right)^{-1} = -\frac{1}{e} a_C Z^2 R_0 R^{-1} \end{aligned} \quad (44.321)$$

Donc:

$$\begin{aligned}
 F_S(R) &= \left| -\frac{d}{dR} \left(-\frac{a_S}{eR_0^2} R^2 \right) \right| = \left| \frac{2a_S}{eR_0^2} R \right| = \frac{2a_S}{|e|R_0^2} R_0 A^{1/3} = \frac{2a_S}{|e|R_0} A^{1/3} \\
 F_C(R) &= \left| -\frac{d}{dR} \left(\frac{a_C Z^2 R_0}{e} R^{-1} \right) \right| = \left| \frac{a_C Z^2 R_0}{e} R^{-2} \right| = \frac{a_C Z^2 R_0}{|e|} R_0^{-2} A^{-2/3} = \frac{a_C Z^2}{|e|R_0} A^{-2/3}
 \end{aligned} \tag{44.322}$$

Si la répulsion coulombienne l'emporte, la fission l'emporte et nous avons alors:

$$F_C(R) > F_S(R) \tag{44.323}$$

Soit:

$$\frac{a_C Z^2}{|e|R_0} A^{-2/3} > \frac{2a_S}{|e|R_0} A^{1/3} \tag{44.324}$$

Après simplification il reste:

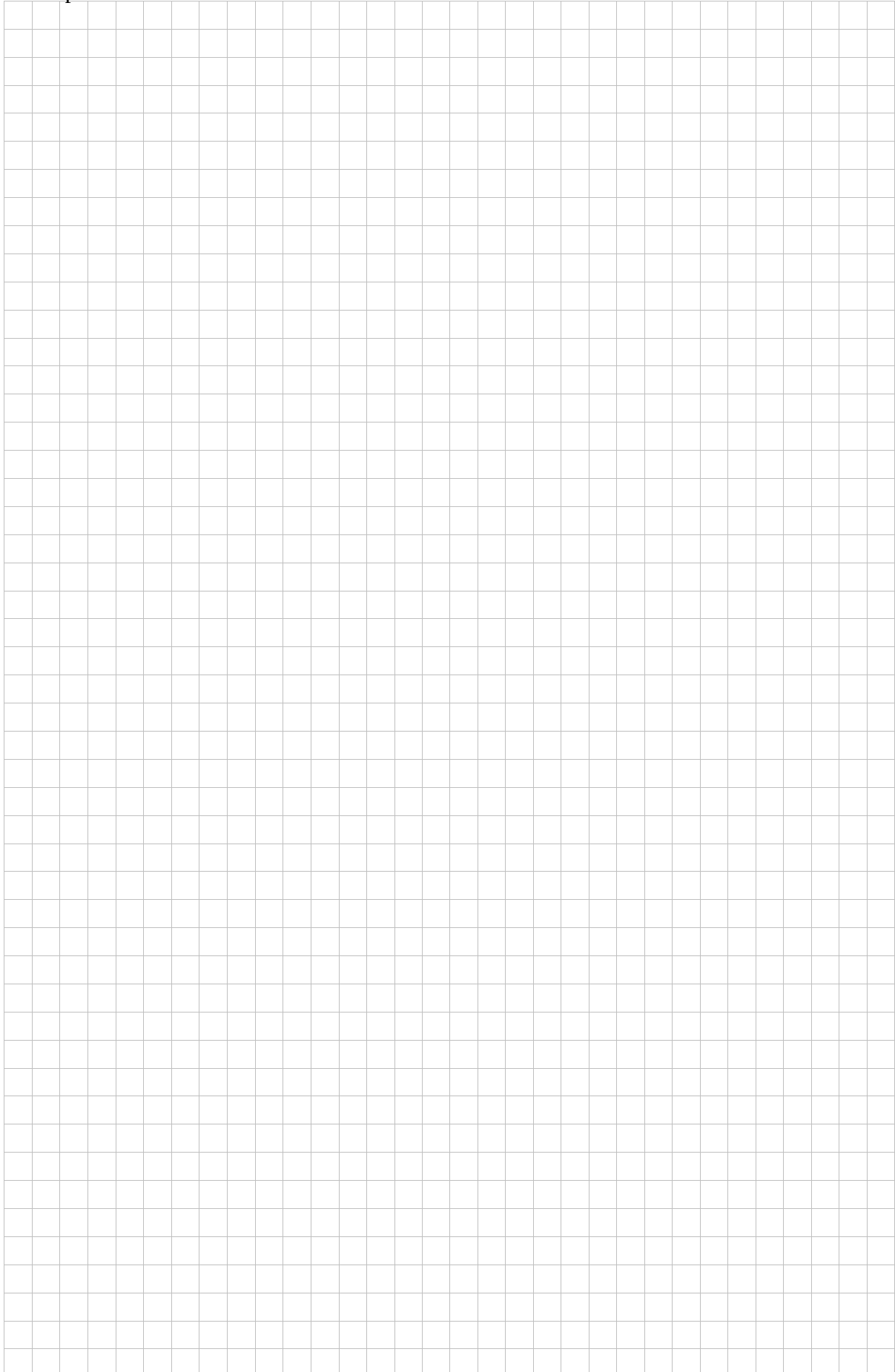
$$a_C Z^2 A^{-2/3} > 2a_S A^{1/3} \tag{44.325}$$

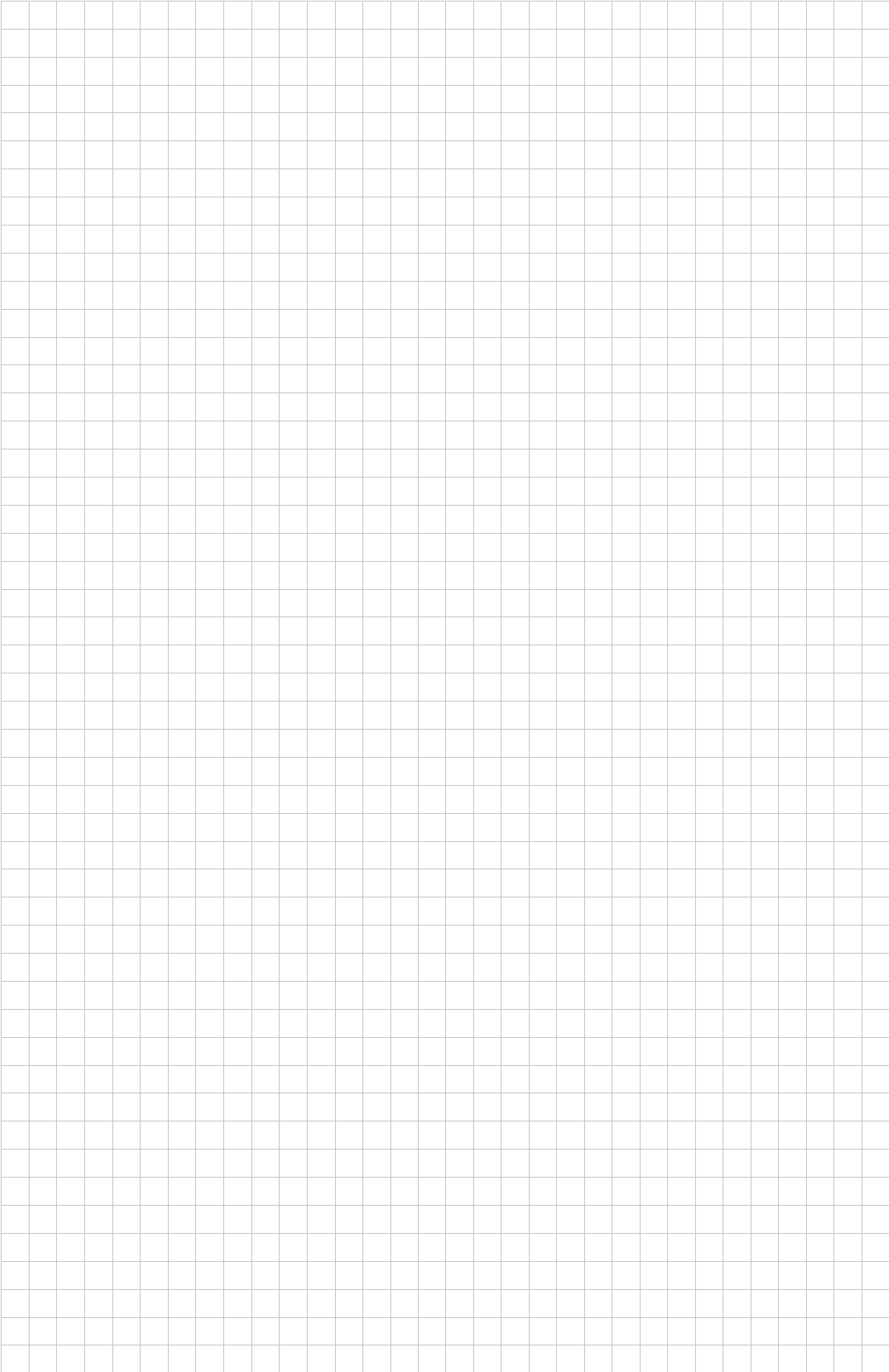
Soit:

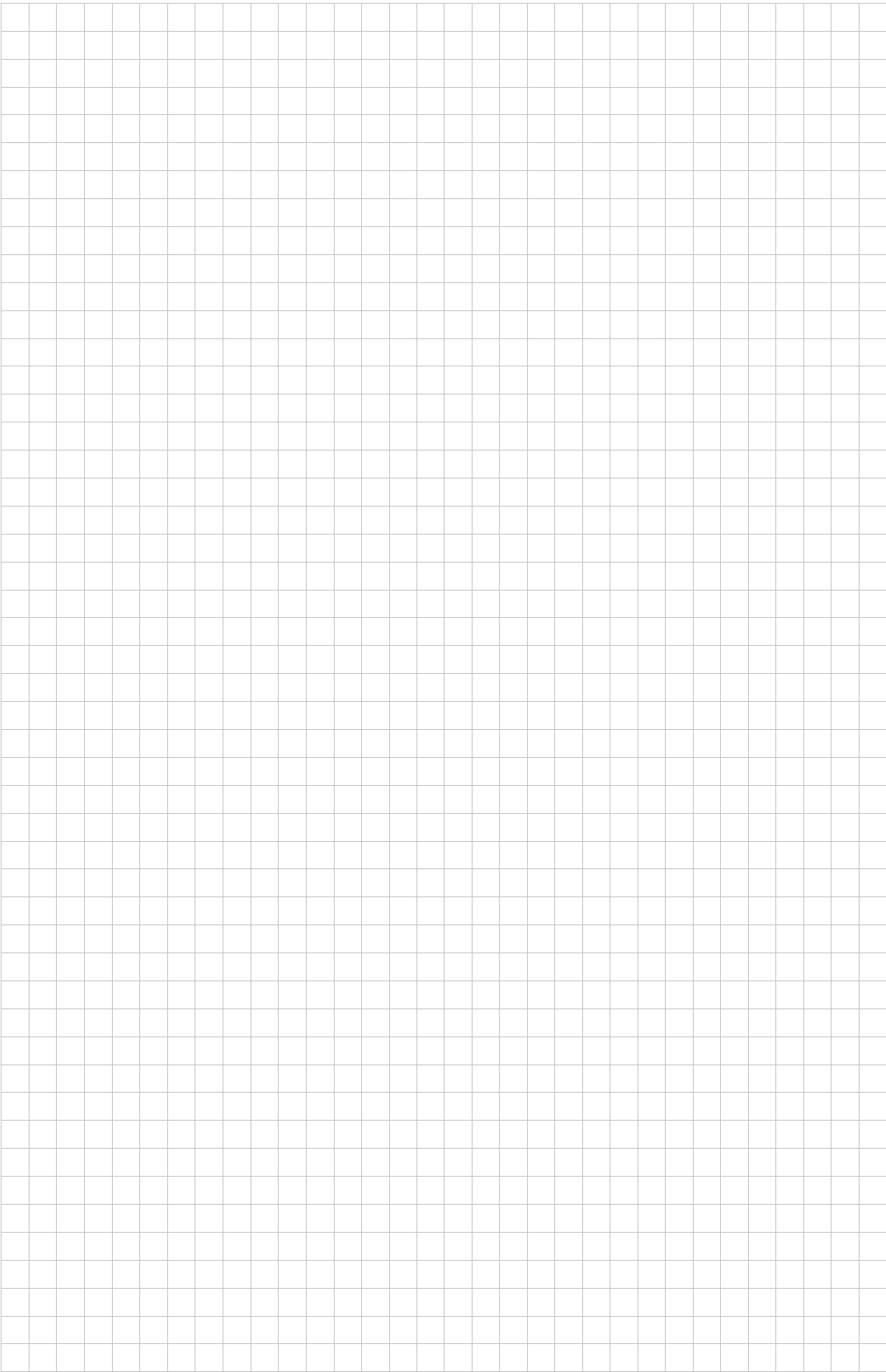
$$\frac{Z^2}{A} > \frac{2a_S}{a_C} \cong \frac{2 \cdot 18.34}{0.71} \cong 52 \tag{44.326}$$

Comme ce rapport pour ${}^{236}_{92}\text{U}$ est voisin de 36, donc inférieur à 52, cette approche expliquerait naïvement pourquoi il ne peut y avoir de fission de manière symétrique (ce que l'expérience confirme).

Notes personnelles:









45. PHYSIQUE QUANTIQUE DES CHAMPS

Avant la formulation de la physique quantique, les particules et les champs étaient considérés comme des entités distinctes mais liées; les particules possèdent certaines caractéristiques intrinsèques (comme la masse et la charge électrique) et produisent les champs (gravitationnels et électromagnétiques). Chaque champ de force émane des particules et remplit l'espace autour d'elles. Les champs emmagasinent et peuvent transporter de l'énergie; ils sont, en ce sens, des milieux continus réels qui lient les particules et communiquent les interactions entre elles. On considérerait que les particules étaient composées de matière et les champs étaient composés d'énergie. La notion de champ de force était l'alternative du 19ème siècle à l'ancienne action à distance assez mystérieuse. Des particules qui ne réagissent à aucun champ de force ne sont pas observables et physiquement n'ont aucun sens. De même, des champs de force qui n'agissent sur aucune particule sont également sans signification. Les notions de particules et de champs n'ont donc un sens que lorsqu'elles sont reliées.

La notion de champ a commencé à être modifiée fondamentalement avec l'introduction par Albert Einstein du concept de photon. Selon cette nouvelle conception, le champ électromagnétique n'a pas son énergie distribuée d'une façon continue dans l'espace. Le photon est le "**quantum du champ électromagnétique**". Il transporte l'énergie et la quantité de mouvement du champ. L'interaction électromagnétique de deux particules chargées et le transfert de l'énergie et de la quantité de mouvement d'une particule à l'autre doivent donc avoir lieu par l'échange des quanta d'énergie électromagnétique, les photons. La théorie de telles interactions (entre particules chargées), appelée "**électrodynamique quantique**" (Q.E.D.), a été la première application réussite de ces idées (elle permet de démontrer la structure fine du modèle de Sommerfeld, d'expliquer le spin de l'électron, etc.) et c'est à elle que nous allons nous intéresser ici.

Dans ce chapitre, nous n'avons pas souhaité faire un cours complet de physique quantique des champs car pour rappel, l'ensemble du site n'a que pour objectif de donner les bases de ce qu'un ingénieur doit connaître au début du 21ème siècle et accessoirement à l'auteur de se faire plaisir en étudiant des sujets qu'il n'avait pas pu voir pendant son cursus scolaire. À ce titre, le lecteur intéressé à approfondir plus cette matière pourra se reporter au meilleur ouvrage que nous ayons eu entre les mains à ce jour (développements détaillés, simples et pédagogiques avec de nombreux cas pratiques) sur ce sujet qui est celui *Quantum Mechanics with Basic Field Theory* de Bipin R. Desai (référence [96] dans la bibliographie).

Remarque: La théorie quantique des champs est l'application de la mécanique quantique aux champs. Elle fournit un cadre largement utilisé en physique des particules et en physique de la matière condensée. Les bases de la théorie quantique des champs auxquelles nous allons limiter notre étude furent développées entre 1935 et 1955, principalement par Paul Dirac, Wolfgang Pauli, Sin-Itiro Tomonaga, Julian Schwinger, Richard Feynman, et Freeman Dyson.

Avant de nous lancer dans les calculs (voir plus loin), montrons que l'approche proposée précédemment, peut être considérée à l'aide d'un formalisme fort simple (pédagogique).

Le lecteur doit cependant se rappeler que les approches simples nécessitent parfois des constructions mentales erronées (par trop simplificatrices) par rapport à la réalité, mais qui satisfont à l'objectif visé: avoir un modèle compréhensible et plus ou moins intuitif.

Considérons à ce titre la figure ci-dessous (représentation de la collision élastique de deux électrons):

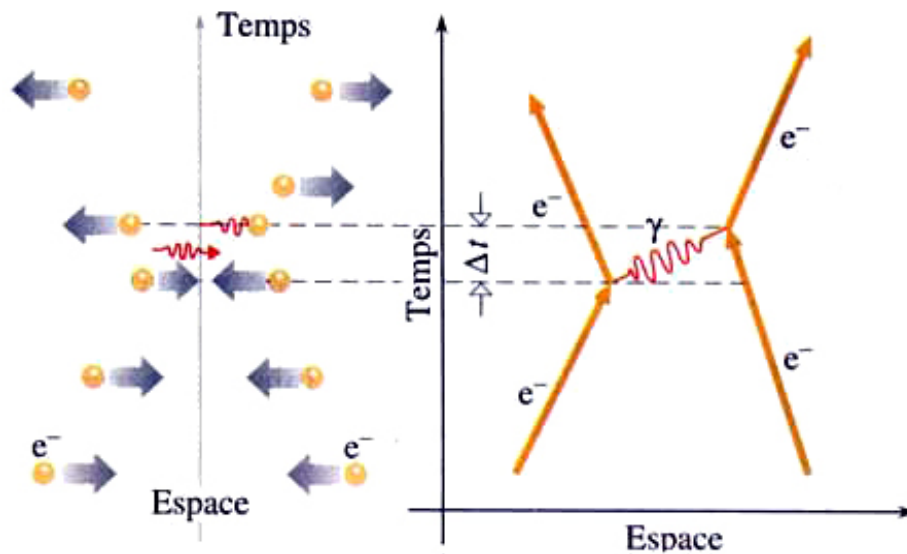


Figure: 45.1 - Exemple de pseudo-diagramme de Feynman

Cette figure est appelée, et à tort (!), dans de nombreux ouvrages pédagogiques "**diagramme de Feynman**" (en réalité il s'agit seulement d'un diagramme qui y ressemble un peu car les vrais diagrammes de Feynman sont destinés à calculer des produits de Wick dans une série perturbative).

Supposons que les deux électrons qui y sont représentés, se déplacent initialement à la même vitesse. Ils s'approchent d'abord puis s'éloignent l'un de l'autre le long d'une droite dans l'espace qui est projetée sur l'axe des temps, dans le sens des temps croissants. L'électron à gauche émet un photon (la ligne ondulée), et pendant un certain temps Δt , il y a deux électrons et un photon. L'électron à droite absorbe ensuite le photon et l'interaction est momentanément terminée; d'autres photons feront par la suite l'aller et retour entre les électrons. La force moyenne est proportionnelle au taux de transfert de la quantité de mouvement due à l'échange des photons. La probabilité de l'émission ou de l'absorption de photons par une particule est liée à sa charge. La force doit donc être proportionnelle au produit des charges en interaction (en accord avec la loi de Coulomb). Pensez à la force de répulsion entre deux astronautes flottant dans l'espace et échangeant une balle dans un sens puis dans l'autre (c'est une approche pédagogique du problème mais qui ne s'applique pas par exemple à l'attraction entre deux particules de charges opposées!). Cependant, le phénomène inverse d'attraction ne peut être visualisé de cette manière mais uniquement sous forme mathématique formelle.

La collision présentée dans la figure ci-dessus est élastique; l'énergie de chacun des électrons est inchangée dans la collision. Malgré cela, pendant un temps Δt , le système contient une quantité d'énergie supplémentaire $h\nu$ correspondant au photon. Pendant ce temps Δt , la conservation de l'énergie est apparemment violée! Peut-on tolérer cette situation? La réponse, donnée par la physique moderne, est oui; mais elle ne peut jamais être observée. Autrement dit, il y a toujours une certaine incertitude ΔE sur la valeur mesurée de l'énergie d'un système. Le principe d'incertitude de Heisenberg impliquant (voir démonstration dans le chapitre de Physique Quantique Ondulatoire) que:

$$\Delta t \cdot \Delta E \cong \hbar \quad (45.1)$$

Une violation de la loi de conservation de l'énergie jusqu'à une quantité ΔE sera cachée par l'incertitude sur l'énergie à condition que le temps disponible pour faire l'observation Δt soit suffisamment grand tel que

$$\Delta t \cong \frac{\hbar}{\Delta E} \quad (45.2)$$

évidemment une valeur inférieure à Δt satisfait également la condition. Nous pouvons donc écrire:

$$\Delta t \geq \frac{\hbar}{\Delta E} \quad (45.3)$$

L'incertitude sur l'énergie dépasse l'énergie d'un photon d'énergie $h\nu$ si le photon existe pendant un temps plus court que:

$$\Delta t = \frac{\hbar}{h\nu} = \frac{1}{4\pi f} \quad (45.4)$$

Ce photon est alors observable sur une distance maximale de:

$$R = c \cdot \Delta t = \frac{c}{4\pi f} \quad (45.5)$$

et comme la fréquence peut être arbitrairement petite, la portée de la force transmise par le photon sans masse est illimitée. Il peut paraître dans cette relation que la portée est limitée pour un photon libre. Mais ce serait oublier (cf. chapitre de Physique Quantique Ondulatoire) qu'un photon libre n'existe pas, car il aurait une fréquence totalement indéterminée. Donc la distance d'interaction le serait aussi.

Ces quanta d'échanges, qui sont inobservables, sont appelés des "**photons virtuels**". Comme les photons ne sont pas chargés, nous disons aussi que l'interaction s'effectue par "**courant neutre**".

Une approche beaucoup plus satisfaisante est celle qui consiste à utiliser la masse comme terme d'énergie:

$$\Delta t \leq \frac{\hbar}{m_0 c^2} \quad (45.6)$$

À l'aide de cette relation, il est possible de connaître le temps pendant lequel une particule virtuelle peut parcourir une distance qui correspondrait à:

$$R = c \cdot \Delta t = \frac{\hbar}{m_0 c} \quad (45.7)$$

Nous verrons plus loin comment déterminer approximativement la masse des particules virtuelles qui interviennent dans les forces nucléaires ce qui nous permettra d'estimer la durée des interactions comme étant de l'ordre de 10^{-24} [s].

Vers la fin des années 1920, il était devenu clair que l'on pouvait considérer chacune des particules connues (proton, électron, etc.) comme le quantum d'un champ spécifique. Dans cette vision, il y a un champ d'électron, un champ de proton, et ainsi de suite comme nous le démontrerons plus loin (l'Univers serait donc un ensemble de champs unifiés). Un objet quelconque est en réalité un ensemble de manifestations observables des quanta des champs.

Par ailleurs, nous avons vu que l'écriture des équations d'onde pour des particules relativistes (équation de Dirac et équation de Klein-Gordon vue en physique quantique relativiste) amène des problèmes insolubles classiquement, notamment des énergies négatives. En fait, cette approche n'est pas justifiée, car d'après l'équation d'Einstein masse et énergie sont équivalentes et si l'on rajoute à cela le principe d'incertitude d'Heisenberg énergie-temps nous constatons qu'un nombre infini de particules peuvent être créées ou annihilées, d'où la nécessité d'un modèle ne prenant plus en compte les propriétés d'une seule particule, mais d'un ensemble de particules, aussi bien réelles que virtuelles.

Remarque: Quand Fermi formula sa théorie des interactions faibles en 1932, il la fonda sur les mêmes principes que l'électrodynamique quantique (c'est une des raisons pour laquelle la QED est appelée "bijou de la physique" - le modèle standard est calqué sur cette théorie par ailleurs). Deux ans plus tard, le physicien japonais H. Yukawa proposa que l'interaction faible était due à l'échange d'un boson virtuel massif.

1. POTENTIEL DE YUKAWA

Le meilleur moyen pour argumenter l'exemple des quantums reste la "démonstration" de la loi de Coulomb (et de Newton) à partir des résultats que nous avons obtenus en physique quantique ondulatoire (nous devons ces développements à Yukawa).

Une version simplifiée de cette démonstration consiste d'abord à se rappeler de l'équation de Klein-Gordon libre (cf. chapitre Physique Quantique Ondulatoire):

$$\hbar^2 \frac{\partial^2 \Psi}{\partial t^2} = \hbar^2 c^2 \Delta \Psi - m_0^2 c^4 \Psi \quad (45.8)$$

cette équation décrit la dynamique d'amplitude de présence d'une particule sans spin dans le temps dans un potentiel donné.

Considérons une composante de Ψ statique (indépendante du temps) à symétrie sphérique:

$$\Psi = U(r) \quad (45.9)$$

L'équation de Klein-Gordon se réduit alors à:

$$\hbar^2 c^2 \Delta U(r) - m_0^2 c^4 U(r) = 0 \quad (45.10)$$

Si nous divisons des deux côtés de l'égalité par $\hbar^2 c^2$:

$$\Delta U(r) - \hbar^{-2} m_0^2 c^2 U(r) = 0 \quad (45.11)$$

Rappelons (cf. chapitre de Calcul Vectoriel) la notation du laplacien d'un champ scalaire:

$$\vec{\nabla}^2 = \Delta \quad (45.12)$$

ainsi que son expression en coordonnées sphériques où $r = 0$ est identifié à l'origine du champ (cf. chapitre de Calcul Vectoriel):

$$\nabla^2 f = \frac{\partial^2 f}{\partial r^2} + 2 \frac{1}{r} \frac{\partial f}{\partial r} + \frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2} + \frac{\cot \theta}{r^2} \frac{\partial f}{\partial \theta} + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 f}{\partial \phi^2} \quad (45.13)$$

Comme le champ $U(r)$ est à symétrie sphérique (dépendant de r uniquement) le laplacien se réduit à:

$$\nabla^2 = \frac{\partial^2}{\partial r^2} + 2 \frac{1}{r} \frac{\partial}{\partial r} = \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) \quad (45.14)$$

Donc l'équation du champ $U(r)$ s'écrit:

$$\frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} U(r) \right) - \hbar^{-2} m_0^2 c^2 U(r) = 0 \quad (45.15)$$

Cette équation différentielle a pour solution (on devine assez facilement que l'exponentielle est une solution possible):

$$U(r) = \frac{C}{r} e^{-r(\hbar^{-1} m_0 c)} \quad (45.16)$$

où C est une constante d'intégration.

Dans le cadre de l'utilisation des unités naturelles (ce qui est le plus fréquent à ce niveau dans la littérature scientifique) ce potentiel s'écrit:

$$U(r) = \frac{C}{r} e^{-rm_0} \quad (45.17)$$

et se nomme "**potentiel de Yukawa**".

Le lecteur remarquera que mise à part la distance r , l'autre variable dans l'exponentielle est la masse (les autres termes étant des constantes universelles). Conséquence: le potentiel de Yukawa est aussi bien un "**champ scalaire**" dans le cas où la masse est nulle (voir l'exemple ci-après) qu'un "**champ massique**" dans le cas où la masse est non nulle !

Cela nous conduit à l'hypothèse suivante: si c'est le champ électrique qui maintient les particules chargées entre elles dans l'atome (voir le traitement du champ non-massique ci-dessous), c'est le champ massique qui maintient les particules non chargées entre elles dans l'atome.

Autrement dit, si des particules interagissent par l'intermédiaire d'un champ massique de masse m_0 (au lieu d'interagir avec des photons de masse nulle), leur force mutuelle va décroître exponentiellement (ce qui est très rapide).

L'approche de Yukawa permet une nouvelle approche interprétation des phénomènes nucléaires, mais elle reste toutefois trop naïve pour expliquer adéquatement les interactions fortes dans leur ensemble.

1.1. CHAMPS MASSIQUES

Le physicien H. Yukawa proposa donc en 1935 que la force nucléaire devait sa très courte portée au fait qu'elle était transmise par des particules massives (plus la masse du quanta échangé est grande, plus

la portée de l'interaction est réduite), décrites par le champ massif ci-haut.

Remarque: Dans le cadre historique de l'époque ces particules hypothétiques étaient les "**mésos**". Mais nous verrons que cette hypothèse ne tiendra pas la route très longtemps.

Voyons cela de plus près. Notons le potentiel de Yukawa sous la forme suivante:

$$U(r) = \frac{C}{r} e^{-r/R} \quad (45.18)$$

avec:

$$R = \frac{\hbar}{m_0 c} \quad (45.19)$$

Cette notation n'est pas innocente, car comme nous le verrons en détails plus loin, lorsque $R \rightarrow +\infty$ (cas de l'interaction électromagnétique et gravitationnelle) alors $m_0 \rightarrow 0$ et nous retrouvons alors la loi fondamentale de l'électrodynamique ou de la gravitation où la particule d'interaction est respectivement le photon (masse nulle) pour la première et le graviton pour la deuxième.

Ainsi, en supposant que le rayon de l'interaction nucléaire forte (cohésion des nucléons entre eux) est $R \cong 1 \cdot 10^{-15} [m]$ et celui de l'interaction nucléaire faible (qui serait à l'origine de la désintégration bêta comme nous l'avons précisé dans le chapitre de Physique Nucléaire) $R \cong 1 \cdot 10^{-18} [m]$, nous obtenons alors les énergies de liaisons des interactions ainsi que leur masse approximative immédiatement:

- Pour "**l'interaction nucléaire forte**":

$$\begin{aligned} E &= m_0 c^2 = \frac{\hbar c}{R} \cong 3.15 \cdot 10^{-12} [J] \cong 196 [MeV] \\ m_0 &= \frac{\hbar}{cR} \cong 3.5 \cdot 10^{-28} [kg] = 386 m_e \cong \frac{1}{5} m_p \end{aligned} \quad (45.20)$$

soit environ 386 fois la masse de l'électron et 1/5 de la masse du proton.

Deux ans après cette prédiction de Yukawa, les physiciens découvrirent une particule correspondant à cette masse: le méson μ^- . Il s'avérera plus tard que ce n'était pas la bonne particule mais une particule de même type que l'électron, soit un lepton et donc un fermion (ce ne peut donc être une particule messagère). De plus, les expériences de diffusions et de collisions avec des protons, deutérons, etc. à des énergies de plus en plus hautes ont montré qu'il y avait une modification de l'intensité/forme de l'interaction forte incompatible avec l'hypothèse d'un seul méson. En outre, les résonnances hadroniques montraient qu'il existait des états excités des mésons ce qui est difficile à imaginer pour des particules considérées comme fondamentales en analogie avec le photon!!

Les particules détectées dans les laboratoires et qui semblaient être les meilleures candidates à l'époque (car il y en avait plusieurs...) de l'interaction nucléaire forte étaient les "**pions**" (ou "**mésos pi**") qui se présentent sous trois formes:

$$\pi^+, \pi^-, \pi^0 \quad (45.21)$$

et qui sont 270 fois plus massifs que l'électron. Donc cette différence de masse indique bien que le modèle de Yukawa n'est pas tout à fait exact.

Avant la découverte des quarks (dont sont constitués les mésons), les mésons pi (pions) étaient donc considérés comme les vecteurs de l'interaction forte (aujourd'hui nous savons qu'en fait ces vecteurs sont les gluons).

- Pour "l'interaction nucléaire faible":

$$\begin{aligned} E = m_0 c^2 &= \frac{\hbar c}{R} = 3.15 \cdot 10^{-8} [J] \cong 196 [GeV] \\ m_0 &= \frac{\hbar}{cR} \cong 3.5 \cdot 10^{-25} [kg] \end{aligned} \quad (45.22)$$

Il s'agit donc d'une masse colossale, une centaine de fois la masse du proton! Les vecteurs d'interactions ont des candidats qui ont été mis en évidence en 1983 dans les accélérateurs du CERN. Ces particules messagères de l'interaction nucléaire faible se nomment les "bosons intermédiaires" W^+, W^-, Z^0 .

Ces observations amenèrent l'hypothèse que la théorie de Yukawa n'était pas une théorie assez fondamentale puisqu'elle représente bien certaines de ses propriétés...

1.2. CHAMPS NON-MASSIQUES

Imaginons maintenant un champ scalaire à symétrie sphérique statique, dont le photon (particule sans spin) est l'hypothétique quantum d'échange.

Comme la masse du photon est nulle, l'expression de $U(r)$ se réduit à:

$$U(r) = \frac{C}{r} \quad (45.23)$$

Si nous interprétons $U(r)$ comme le potentiel électrostatique source d'une quantité $Q = n \cdot q$ de charges élémentaires q alors la constante C dans notre système métrique vaut:

$$C = \frac{Q}{4\pi\epsilon} \quad (45.24)$$

Tel que:

$$U(r) = \frac{1}{4\pi\epsilon} \frac{Q}{r} \quad (45.25)$$

Comme nous avons:

$$U(r) = - \int_{\infty}^r E dr \quad (45.26)$$

Il en découle:

$$\frac{dU(r)}{dr} = -E = -\frac{1}{4\pi\epsilon} \frac{Q}{r^2} \quad (45.27)$$

Ce qui nous donne:

$$E = \frac{1}{4\pi\epsilon} \frac{Q}{r^2} \quad (45.28)$$

Conclusion: Si une particule se trouve dans un champ de potentiel à symétrie sphérique $U(r)$ dont le photon est supposé être initialement le quantum d'interaction alors nous avons affaire à un champ électrostatique dont l'expression est identique à la loi de Coulomb (ceci valide donc encore une fois de façon magistrale la théorie de la physique quantique ondulatoire).

Remarque: Le photon est donc bien le quantum d'interaction du champ électrique à symétrie sphérique (lorsque les charges ont une vitesse relativiste le champ électrique n'est pas à symétrie sphérique et les équations deviennent un peu plus compliquées - voir le chapitre de Relativité Restreinte) et nous ne devrions plus parler de charge électrique mais de "transparence" aux photons. Effectivement, le neutron étant neutre globalement celui-ci ne devrait pas interagir avec le champ électrique, mais comme il est composé de particules chargées (les quarks) les expériences mettent en évidence une affluence en présence du champ électromagnétique (dont le photon est le quantum d'interaction).

Ceci dit, en appliquant le même raisonnement, nous pouvons de même retrouver le potentiel gravitationnel de Newton:

$$a = G \frac{M}{r^2} \quad (45.29)$$

Ce qui impliquerait que le quantum d'interaction du champ gravitationnel est aussi sans masse (du moins dans le cas des petites masses étant donné que nous savons que le potentiel de Newton n'est qu'une approximation de la relativité générale dans le cas des petites masses). Étant donné que le champ gravitationnel ne semble pas interagir avec la présence d'un champ magnétique ou électrostatique, cela nous amène à émettre l'hypothèse que le quantum d'interaction n'est pas le photon et à supposer qu'une autre particule, que nous appellerons "**graviton**", en est le messenger.

2. ÉQUATION D'EULER-LAGRANGE DES CHAMPS

La façon dont la théorie des champs fut introduite à partir des particules élémentaires par Dirac est connue pour des raisons historiques sous l'appellation de "**deuxième quantification**".

Il est peut-être utile de mettre en évidence une possible source de confusion: les champs ne sont pas liés à la dualité onde-corpuscule. Ce que nous entendons par "champ" est un concept qui permet la création ou l'annihilation de particules en tout point de l'espace comme nous le verrons dans les développements mathématiques.

Rappelons que nous avons défini dans le chapitre de Physique Quantique Ondulatoire lors de l'étude de l'équation d'évolution de Schrödinger l'opérateur d'Heisenberg, nécessaire à la condition de normalisation de De Broglie:

$$X(t) = e^{\frac{it}{\hbar}H} X e^{-\frac{it}{\hbar}H} \quad (45.30)$$

En dérivant cet opérateur par rapport au temps, nous avons trivialement:

$$\dot{X}(t) = -\frac{i}{\hbar}[X(t), H] \quad (45.31)$$

où rappelons-le, le commutateur de deux opérateurs est donné (comme nous l'avons déjà vu lors de notre étude des opérateurs adjoints et hermitiques en physique quantique ondulatoire) par définition par:

$$[\alpha, \beta] = \alpha\beta - \beta\alpha \quad (45.32)$$

C'est l'hamiltonien H qui fait irruption en premier dans la relation précédente. Mais nous pouvons tout aussi bien lui substituer un hamiltonien dépendant du temps $H(t)$ tel que:

$$\dot{X}(t) = -\frac{i}{\hbar}[X(t), H(t)] \quad (45.33)$$

Maintenant, nous pouvons substituer à $\dot{X}(t)$ des observables connues telles que:

$$\begin{cases} \dot{q}(t) = -\frac{i}{\hbar}[q(t), H(t)] \\ \dot{p}(t) = -\frac{i}{\hbar}[p(t), H(t)] \end{cases} \quad (45.34)$$

dites "**équations du mouvement de Heisenberg**". Ce qui est intéressant dans les deux relations obtenues précédemment, c'est la façon avec laquelle se réalise la jonction entre la physique quantique et la mécanique classique. Effectivement, nous avons démontré dans le chapitre de Mécanique Analytique que les relations ci-dessous sont et seront toujours valables quel que soit le domaine étudié:

$$\begin{aligned} \dot{q}(t) &= \frac{\partial H(t)}{\partial p(t)} \\ \dot{p}(t) &= -\frac{\partial H(t)}{\partial q(t)} \end{aligned} \quad (45.35)$$

ainsi que:

$$p = \frac{\partial L}{\partial \dot{q}} \quad (45.36)$$

et en supposant la généralisation à plusieurs degrés de liberté comme étant intuitive:

$$H = \sum p_n \dot{q}_n - L \quad (45.37)$$

La généralisation à plusieurs degrés de liberté est immédiate et nous donne l'ensemble des relations (nous allégeons les écritures en omettant l'écriture de la dépendance à la variable temporelle):

$$p_n = \frac{\partial L}{\partial \dot{q}_n} \text{ et } H = \sum p_n \dot{q}_n - L$$

$$\begin{cases} \dot{q}_n = \frac{\partial H}{\partial p_n} \\ \dot{p}_n = -\frac{\partial H}{\partial q_n} \end{cases} \text{ et } \begin{cases} \dot{q}_n = -\frac{i}{\hbar} [q_n, H] \\ \dot{p}_n = -\frac{i}{\hbar} [p_n, H] \end{cases} \quad (45.38)$$

Nous avons encore besoin de deux autres relations importantes que nous allons de suite déterminer. D'abord, d'après les définitions des commutateurs, il est inutile de démontrer que (trivial):

$$[q, q] = [p, p] = 0 \quad (45.39)$$

Par contre, il est un peu plus subtil de démontrer la valeur de $[q, p]$ (nous plaisantons...). Rappelons que nous avons démontré lors de notre étude des opérateurs linéaires fonctionnels que (nous nous restreignons au cas de la coordonnée x ici):

$$p_x = \frac{\hbar}{i} \frac{\partial}{\partial x} \quad (45.40)$$

et que q représente une coordonnée généralisée (x par exemple...). Nous avons donc (résultat déjà démontré dans le chapitre de Physique Quantique Ondulatoire...):

$$[q, p] = \left[x, \frac{\hbar}{i} \frac{\partial}{\partial x} \right] = x \frac{\hbar}{i} \frac{\partial}{\partial x} - \frac{\hbar}{i} \frac{\partial}{\partial x} x = 0 - \frac{\hbar}{i} = i\hbar \quad (45.41)$$

Les deux dernières relations peuvent être généralisées à toutes les composantes voulues telles que:

$$\boxed{\begin{aligned} [q_n, q_m] &= [p_n, p_m] = 0 \\ [q_n, p_m] &= i\hbar \delta_{nm} \end{aligned}} \quad (45.42)$$

avec rappelons-le (cf. chapitre de Calcul Tensoriel):

$$\delta_{ij} = \delta^{ij} = \delta_j^i = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases} \quad (45.43)$$

qui est le symbole de Kronecker.

Pour en arriver enfin à la théorie quantique des champs, il nous faut encore généraliser à une infinité continue de degrés de liberté. En effet, même le plus simple des champs est caractérisé, à un instant t , par une infinité continue de quantités:

$$\Phi(x, y, z, t) = \Phi^A \quad (45.44)$$

pour tout $\vec{x}$. Nous pourrions donc imaginer représenter la fonction Φ par ses valeurs $\Phi(\vec{x}, t)$ en un ensemble discret de points $\vec{x}_n$ que nous rendrons en fin de compte infiniment dense (prenez garde au fait que nous utilisons la notion de densité !). Nous pouvons aussi travailler, pour commencer, non pas dans tout l'espace, mais dans un volume fini que nous finirons par rendre très grand. En procédant ainsi,

nous pouvons trouver comment généraliser le formalisme canonique et le processus de quantification. Au niveau formel, nonobstant de subtiles questions de convergences (voir les parties mathématiques du site), la généralisation aux systèmes continus consiste principalement à remplacer les sommes sur des indices n par des intégrales sur des arguments $\vec{x}$, et les deltas de Kronecker par des deltas de Dirac (sur l'espace-temps):

$$\delta^4((x_1, y_1, z_1, t) - (x_2, y_2, z_2, t)) = \delta^4((A) - (B)) = \delta^4(\vec{x} - \vec{y}) = \delta_{AB}^4 \quad (45.45)$$

En considérant alors le principe variationnel comme nous l'avons étudié en mécanique analytique:

$$S(q, \dot{q}) = \int_{t_A}^{t_B} L(q, \dot{q}) dt \quad (45.46)$$

et le principe de moindre action nous imposant:

$$\delta S = \delta \int_{t_A}^{t_B} L dt = \int_{t_A}^{t_B} \delta L dt = 0 \quad (45.47)$$

où le lagrangien sera maintenant une fonction du champ $\Phi(\vec{x}, t)$ et de dérivée par rapport au champ $\phi(\vec{x}, t)$ (puisque'il n'y a pas de notion de quantité de mouvement pour un champ !).

Si nous divisons la relation précédente par $\delta q(t) = \delta q \neq 0$ nous obtenons:

$$\frac{\delta S(q, \dot{q})}{\delta q} = 0 \quad (45.48)$$

ce qui nous donne le droit d'écrire:

$$\delta S(q, \dot{q}) = \int \frac{\delta S(q, \dot{q})}{\delta q} \delta q dt = \int 0 \delta q dt = 0 \quad (45.49)$$

et en imposant une analogie avec un concept de champ:

$$\delta S(q, \dot{q}) = \int \frac{\delta S(q, \dot{q})}{\delta q} \delta q dt = \int \frac{\delta S(\Phi^A, \dot{\Phi}^A)}{\delta \Phi^A} \delta \Phi^A d^4x \quad (45.50)$$

où $x \equiv (\vec{x}, t)$ et $d^4x = dx_1 dx_2 dx_3 dt$.

Finalement, comme tous les termes suivants sont nuls, ils sont égaux (nous faisons intervenir l'équation d'Euler-Lagrange démontrée dans le chapitre de Mécanique Analytique):

$$\delta S(q, \dot{q}) = \frac{\delta S(q, \dot{q})}{\delta q(t)} = \frac{\partial L}{\partial q_i(t)} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i(t)} = 0 \quad (45.51)$$

en analogie avec le champ Φ^A nous obtenons (donc c'est une démarche intuitive à la sauce physicienne...):

$$\frac{\delta S(\Phi^A, \dot{\Phi}^A)}{\delta \Phi^A} = \frac{\partial L}{\partial \Phi^A} - \frac{d}{d(A)} \frac{\partial L}{\partial \dot{\Phi}^A} = 0 \quad (45.52)$$

Il s'agit donc maintenant d'un lagrangien avec des dérivées partielles fonctionnelles.

Cette écriture étant cependant peu commode, on prend pour habitude d'écrire les différentielles partielles (en utilisant les unités naturelles de la physique) aux composantes $\partial/\partial t, \partial/\partial y, \partial/\partial z, \partial/\partial x$ sous la forme ∂_μ , ce qui nous donne finalement:

$$\frac{\partial L}{\partial \Phi^A} - \partial_\mu \left(\frac{\partial L}{\partial (\partial_\mu \Phi^A)} \right) = 0 \quad (45.53)$$

Évidemment, si nous n'utilisons pas les unités naturelles, nous devrions adopter l'écriture:

$$\partial_\mu = \left(\frac{\partial}{c \partial t}, \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) \quad (45.54)$$

Le fait que la dérivation partielle ∂_μ porte maintenant sur toutes les composantes et non pas uniquement sur t est dû au passage de coordonnées généralisées q fonction uniquement de t à une fonction du champ Φ^A dépendante de x, y, z et t . La raison profonde réside dans le fait que le temps et les coordonnées spatiales jouent le même rôle, celui de décrire le continuum espace-temps sur lequel évolue le système physique.

Ceci nous amène aussi à écrire le principe de moindre action sous la forme suivante:

$$\begin{aligned} \delta S(q, \dot{q}) &= \int \delta L dt = 0 \\ \Rightarrow \delta S(\Phi^A, \dot{\Phi}^A) &= \int \frac{\delta L}{\delta \Phi^A} \delta \Phi^A dx^4 = \int \frac{\partial L}{\partial \Phi^A} \delta \Phi^A dx^4 = 0 \end{aligned} \quad (45.55)$$

Avec l'action des champs notée plus traditionnellement:

$$S = \int L(\Phi^A, \dot{\Phi}^A, \partial_\mu \Phi^A) dx^3 dt = \int L(\Phi^A, \dot{\Phi}^A, \partial_\mu \Phi^A) dx^4 \quad (45.56)$$

ou encore pour différencier lagrangien et densité lagrangienne (nous "stylisons" parfois le L):

$$S = \int \mathcal{L}(\Phi^A, \dot{\Phi}^A, \partial_\mu \Phi^A) dx^4 \quad (45.57)$$

à comparer à l'action de la particule:

$$S(q, \dot{q}) = \int_{t_a}^{t_b} L(q, \dot{q}) dt \quad (45.58)$$

En analogie avec $p_n = \frac{\partial L}{\partial \dot{q}_n}$ nous écrirons:

$$\Pi_A = \frac{\partial L}{\partial(\partial_\mu \Phi^A)} = \frac{\partial L}{\partial \dot{\Phi}^A} \quad (45.59)$$

et en analogie avec $H = \sum p_n \dot{q}_n - L$ nous écrivons:

$$H = \sum \Pi_A \dot{\Phi}^A - L \quad (45.60)$$

mais un champ est un milieu continu. La somme sigma n'est donc plus adaptée et il faut passer à une intégration sur tout l'espace-temps telle que:

$$H = \int \Pi_A \dot{\Phi}^A d^4x - L \quad (45.61)$$

En analogie avec les équations du mouvement de Heisenberg (cette démarche est souvent appelée "**principe de correspondance**"), nous écrivons:

$$\begin{cases} \dot{q}_n = \frac{\partial H}{\partial p_n} \\ \dot{p}_n = -\frac{\partial H}{\partial q_n} \end{cases} \Leftrightarrow \begin{cases} \dot{\Phi}^A = \frac{\partial H}{\partial \Pi_A} \\ \dot{\Pi}_A = -\frac{\partial H}{\partial \Phi^A} \end{cases} \quad (45.62)$$

Passons maintenant à la théorie quantique en postulant des champs d'opérateurs de Heisenberg correspondants. Rappelons que nous avons obtenu plus haut que:

$$\dot{q}_n = -\frac{i}{\hbar}[q_n, H], \dot{p}_n = -\frac{i}{\hbar}[p_n, H] \text{ et } [q_n, q_m] = [p_n, p_m] = 0, [q_n, p_m] = i\hbar\delta_{nm} \quad (45.63)$$

ce qui nous donne:

$$\dot{\Phi}^A = -\frac{i}{\hbar}[\Phi^A, H], \dot{\Pi}^A = -\frac{i}{\hbar}[\Pi^A, H] \text{ et } [\Phi^A, \Phi^B] = [\Pi^A, \Pi^B] = 0, [\Phi^A, \Pi^B] = i\hbar\delta_{AB}^3 \quad (45.64)$$

Si nous résumons un peu le tout et que nous affichons la comparaison avec la physique quantique ondulatoire, nous avons finalement:

1. En physique quantique ondulatoire (c'est joli à regarder non?):

$$\boxed{\begin{aligned} p_n &= \frac{\partial L}{\partial \dot{q}_n} \text{ et } H = \sum p_n \dot{q}_n - L \\ \begin{cases} \dot{q}_n = \frac{\partial H}{\partial p_n} \\ \dot{p}_n = -\frac{\partial H}{\partial q_n} \end{cases} &\text{ et } \begin{cases} \dot{q}_n = -\frac{i}{\hbar}[q_n, H] \\ \dot{p}_n = -\frac{i}{\hbar}[p_n, H] \end{cases} \\ \begin{cases} [q_n, q_m] = [p_n, p_m] = 0 \\ [q_n, p_m] = i\hbar\delta_{nm} \end{cases} \end{aligned}} \quad (45.65)$$

2. Et l'équivalent par principe de correspondance en physique quantique des champs (alors là... ça devient de l'art!):

$$\begin{aligned}
 &\Pi_A = \frac{\partial L}{\partial \dot{\Phi}^A} \text{ et } H = \int \Pi_A \dot{\Phi}^A d^4x - L \\
 &\left\{ \begin{aligned} \Phi^A &= \frac{\partial H}{\partial \Pi_A} \\ \dot{\Pi}_A &= -\frac{\partial H}{\partial \Phi^A} \end{aligned} \right. \text{ et } \left\{ \begin{aligned} \dot{\Phi}^A &= -\frac{i}{\hbar} [\Phi^A, H] \\ \dot{\Pi}^A &= -\frac{i}{\hbar} [\Pi^A, H] \end{aligned} \right. \quad (45.66) \\
 &\left\{ \begin{aligned} [\Phi^A, \Phi^B] &= [\Pi^A, \Pi^B] = 0 \\ [\Phi^A, \Pi^B] &= i\hbar \delta_{AB}^3 \end{aligned} \right.
 \end{aligned}$$

Et le tour est joué! Nous venons de passer les paramètres de la physique quantique où les corps ponctuels sont décrits par des fonctions d'onde, à une physique quantique où les corps ponctuels deviennent des champs continus.

Il ne reste plus qu'à appliquer ce schéma général à des exemples concrets:

Nous allons commencer par un premier exemple en tenant compte de l'aspect relativiste. Ainsi, la densité lagrangienne la plus simple non triviale que nous puissions construire est de la forme (vous allez de suite voir à quoi elle va mener, ce qui confirmera sa validité théorique - par ailleurs, le développement qui va suivre aurait très bien pu être présenté dans l'autre sens):

$$L = -\frac{1}{2} \partial_\mu \Phi^A \partial^\mu \Phi^A - \frac{1}{2} m_0^2 (\Phi^A)^2 \quad (45.67)$$

ou plus explicitement:

$$L = \underbrace{-\frac{1}{2} \partial_0 \Phi^A \partial^0 \Phi^A}_{\text{terme énergie cinétique de } L} - \underbrace{\left(\frac{1}{2} \underbrace{\partial_k \Phi^A \partial^k \Phi^A}_{\substack{\text{somme d'Einstein} \\ \text{uniquement sur les} \\ \text{composantes spatiales}}} + \frac{1}{2} m_0^2 (\Phi^A)^2 \right)}_{\text{terme énergie potentielle de } L} = T - V \quad (45.68)$$

que les physiciens appellent "**champ scalaire pour une particule libre et sans spin**" ou "**lagrangien de Klein-Gordon**" pour une particule sans spin où nous utilisons les notations condensées habituelles:

$$\partial_\mu \Phi^A = \frac{\partial \Phi^A}{\partial x_\mu} \quad \partial_\mu \Phi^A \partial^\mu \Phi^A = \sum_\mu \partial_\mu \Phi^A \partial^\mu \Phi^A \quad (45.69)$$

et les unités naturelles:

$$\hbar = c = 1 \quad (45.70)$$

calculons l'équation d'Euler-Lagrange y relative (normalement c'est trivial) en n'oubliant pas qu'il s'agit de dérivées fonctionnelles ce qui simplifie grandement le calcul (n'hésitez pas à nous demander les détails si besoin est):

$$\frac{\partial L}{\partial \Phi^A} - \partial_\mu \left(\frac{\partial L}{\partial (\partial_\mu \Phi^A)} \right) = -m_0^2 \Phi^A + \partial_\mu \partial^\mu \Phi^A \quad (45.71)$$

d'où l'équation du mouvement en unités naturelles et en osant une écriture dangereuse pour la double dérivée partielle (la sommation d'Einstein est alors sous-entendue...):

$$(\partial_\mu^2 - m_0^2) \Phi^A = 0 \quad (45.72)$$

où comme nous allons de suite le voir, pour retomber sur les résultats obtenus dans le chapitre de Physique Quantique Relativiste, nous sommes obligés d'introduire l'opérateur différentiel contravariant et covariant (cf. chapitre de Calcul Tensoriel) avec la signature -,+,+,+:

$$\begin{aligned} \partial^\mu &= \left(\frac{\partial^2}{c^2 \partial t^2}, \frac{\partial^2}{\partial x^2}, \frac{\partial^2}{\partial y^2}, \frac{\partial^2}{\partial z^2} \right) = \left(\frac{\partial^2}{c^2 \partial t^2}, \vec{\nabla}^2 \right) \\ \partial_\mu &= \eta_{\mu\nu} \partial^\nu = \left(-\frac{\partial^2}{c^2 \partial t^2}, \frac{\partial^2}{\partial x^2}, \frac{\partial^2}{\partial y^2}, \frac{\partial^2}{\partial z^2} \right) = \left(-\frac{\partial^2}{c^2 \partial t^2}, \vec{\nabla}^2 \right) \end{aligned} \quad (45.73)$$

Rappelons maintenant à titre de comparaison que dans le chapitre de Physique Quantique Ondulatoire, nous avons obtenu l'équation de Klein-Gordon libre suivante:

$$(\hbar^2 \square - m_0^2 c^2) \Psi = 0 \quad (45.74)$$

aussi avec la signature -,+,+,+ donc avec l'opérateur d'Alembertien:

$$\square = -\frac{\partial^2}{c^2 \partial t^2} + \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \quad (45.75)$$

Nous avons alors une correspondance parfaite entre l'équation de Klein-Gordon libre et l'équation des champs (notée ci-dessous en unités naturelles comme il est d'usage):

$$(\partial_\mu^2 - m_0^2) \Psi = 0 \text{ et } (\partial_\mu^2 - m_0^2) \Phi^A = 0 \quad (45.76)$$

et c'est ici que l'on peut éventuellement ressentir un frisson dans le dos et rester admiratif face à la puissance du formalisme mathématique ouvrant de nouvelles perspectives sur la manière de voir les rouages de l'Univers... Ainsi, dans la théorie des champs, l'équation de Klein-Gordon peut être réinterprétée comme une équation de champ!

La forme de l'équation de Klein-Gordon libre impliquant les champs est parfois appelée "**équation des champs de Klein-Gordon**".

Et encore... mieux...vous allez voir, nous allons le faire un peu à l'aveugle et... alors là ! Considérons maintenant le lagrangien suivant (que nous supposons obtenu par bricolages successifs... mais à nouveau nous aurions pu faire le développement dans l'autre sens) se voulant exprimer "**l'interaction d'un champ électromagnétique avec une densité courant**":

$$L = \frac{1}{4} F^{\mu\nu} F_{\mu\nu} - j^\nu A_\nu \quad (45.77)$$

où nous y reconnaissons les tenseurs du champ électromagnétique démontrés et déterminés dans le chapitre d'Électrodynamique et pour lesquels, rappelons-le:

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \quad (45.78)$$

Dans ce lagrangien, traitons le potentiel vecteur comme étant champ tel que:

$$\frac{\partial L}{\partial \Phi^A} - \partial_\mu \left(\frac{\partial L}{\partial (\partial_\mu \Phi^A)} \right) = \frac{\partial L}{\partial A_\nu} - \partial_\mu \left(\frac{\partial L}{\partial (\partial_\mu A_\nu)} \right) = 0 \quad (45.79)$$

Dès lors en décomposant les développements, nous obtenons très facilement:

$$\frac{\partial (j^\nu A_\nu)}{\partial (\partial_\mu A_\nu)} = 0 \text{ et } \frac{\partial (j^\nu A_\nu)}{\partial A_\nu} = j^\nu \text{ et } \frac{\partial (F^{\alpha\beta} F_{\alpha\beta})}{\partial A_\nu} = 0 \quad (45.80)$$

Dans un premier temps, le lecteur vérifiera en faisant un peu de calcul tensoriel relativement élémentaire que:

$$\frac{\partial F_{\alpha\beta}}{\partial (\partial_\mu A_\nu)} = \frac{\partial (\partial_\alpha A_\beta - \partial_\beta A_\alpha)}{\partial (\partial_\mu A_\nu)} = \delta_\alpha^\mu \delta_\beta^\nu - \delta_\beta^\mu \delta_\alpha^\nu \quad (45.81)$$

Puis:

$$\begin{aligned} \frac{\partial (F^{\alpha\beta} F_{\alpha\beta})}{\partial (\partial_\mu A_\nu)} &= F^{\alpha\beta} \frac{\partial F_{\alpha\beta}}{\partial (\partial_\mu A_\nu)} + \frac{\partial F^{\alpha\beta}}{\partial (\partial_\mu A_\nu)} F_{\alpha\beta} = 2 \frac{\partial F^{\alpha\beta}}{\partial (\partial_\mu A_\nu)} F_{\alpha\beta} \\ &= 2 \left[\delta_\alpha^\mu \delta_\beta^\nu - \delta_\beta^\mu \delta_\alpha^\nu \right] F^{\alpha\beta} = 2 \left[F^{\mu\nu} - F^{\nu\mu} \right] = 4 F^{\mu\nu} \end{aligned} \quad (45.82)$$

Dès lors, l'équation du champ s'écrit:

$$\frac{\partial L}{\partial A_\nu} - \partial_\mu \left(\frac{\partial L}{\partial (\partial_\mu A_\nu)} \right) = j^\nu - \partial_\mu \left(\frac{1}{4} 4 F^{\mu\nu} \right) = 0 \quad (45.83)$$

d'où:

$$\partial_\mu F^{\mu\nu} = j^\nu \quad (45.84)$$

Faut avouer que le résultat assez c'est beau même si la démarche n'est pas des plus rigoureuses!!! Nous retrouvons donc l'équation de Maxwell avec sources avec le même lagrangien du champ (cf. chapitre d'Électrodynamique). Ainsi, ce lagrangien sans masse est assimilé au lagrangien du champ vectoriel de spin 1 assimilé aux bosons (cf. chapitre de Mécanique Statistique).

Rappelons maintenant que nous avons obtenu dans le chapitre d'Électrodynamique l'action suivante

pour une particule chargée dans un champ électromagnétique (avant un long développement qui nous avait amené au tenseur du champ électromagnétique):

$$S = \int_{\alpha}^{\beta} -m_0 c ds - q A^i dx_i \quad (45.85)$$

et en se rappelant que (cf. chapitre d'Électrodynamique):

$$\frac{ds}{dt} = c \sqrt{1 - v^2 / c^2} \quad (45.86)$$

il vient:

$$S = \int_{\alpha}^{\beta} \left(-m_0 c^2 \sqrt{1 - v^2 / c^2} - q A^i dv_i \right) dt \quad (45.87)$$

La densité lagrangienne correspondante est donc:

$$L = -\rho_0 c^2 \sqrt{1 - v^2 / c^2} - j^v A_v \quad (45.88)$$

Nous avons donc finalement:

1. Le lagrangien (densité lagrangienne) d'une particule chargée dans un champ électromagnétique (que nous venons d'obtenir):

$$L = -\rho_0 c^2 \sqrt{1 - v^2 / c^2} - j^v A_v \quad (45.89)$$

2. Le lagrangien (densité lagrangienne) de tout à l'heure (qui nous a permis de retomber sur les équations de Maxwell avec sources):

$$L = \frac{1}{4} F^{\mu\nu} F_{\mu\nu} - j^v A_v \quad (45.90)$$

Dès lors, il est naturel d'écrire le "**lagrangien (densité lagrangienne) total du champ électromagnétique**":

$$L = -\rho_0 c^2 \sqrt{1 - v^2 / c^2} - j^v A_v + \frac{1}{4\mu_0} F^{\mu\nu} F_{\mu\nu} \quad (45.91)$$

Continuons maintenant notre bonhomme de chemin avec l'équation de Dirac libre! Rappelons que nous avons obtenu dans le chapitre de Physique Quantique Relativiste l'équation de Dirac libre sous la forme (fondamentalement rappelons qu'il s'agit d'une équation relativiste):

$$(i \gamma^\mu \partial_\mu - m_0) \Psi = (i \not{\partial} - m_0) \Psi = 0 \quad (45.92)$$

Maintenant rappelons (cf. chapitre d'Algèbre Linéaire) que $(AB)^\dagger = B^\dagger A^\dagger$. Dès lors, il vient:

$$\left((i \not{\partial} - m_0) \Psi \right)^\dagger = 0^\dagger \Rightarrow \Psi^\dagger (i \not{\partial} - m_0)^\dagger = \Psi^\dagger (i^\dagger \not{\partial}^\dagger - m_0) = \Psi^\dagger (-i \gamma^{\mu\dagger} \partial_\mu^\dagger - m_0) = 0 \quad (45.93)$$

Or, $\partial_\mu^\dagger = \partial_\mu$ et il est super facile de vérifier (ne pas oublier que nous utilisons la forme représentative de Dirac des matrices de Pauli !!!) $\gamma^{\mu\dagger} = \gamma^0 \gamma^\mu \gamma^0$ ce qui nous amène à écrire:

$$\Psi^\dagger (-i \gamma^0 \gamma^\mu \gamma^0 \partial_\mu - m_0) = 0 \quad (45.94)$$

Il est alors commode d'introduire "l'adjoint de Dirac":

$$\bar{\Psi} = \Psi^\dagger \gamma^0 \quad (45.95)$$

Remarque: Rappelons que Ψ est une matrice-colonne et $\Psi^\dagger$ une matrice-ligne. Il vient donc que $\bar{\Psi}$ est aussi une matrice-ligne!

Utilisant le fait que dans la représentation de Dirac $\gamma^0 \gamma^0 = \mathbb{1}$ nous pouvons écrire:

$$\begin{aligned} \Psi^\dagger (-i \gamma^0 \gamma^\mu \gamma^0 \partial_\mu - m_0 \gamma^0 \gamma^0) &= 0 \\ \Rightarrow \bar{\Psi} (-i \gamma^\mu \gamma^0 \partial_\mu - m_0 \gamma^0) &= 0 \quad (45.96) \\ \Rightarrow \bar{\Psi} (i \gamma^\mu \gamma^0 \partial_\mu + m_0 \gamma^0) &= 0 \end{aligned}$$

en simplifiant les γ^0 il vient l'équation de Dirac libre adjointe:

$$\bar{\Psi} (i \gamma^\mu \partial_\mu + m_0) = \bar{\Psi} (i \not{\partial} + m_0) = 0 \quad (45.97)$$

Ce que nous notons traditionnellement:

$$\bar{\Psi} (i \tilde{\not{\partial}} + m_0) = 0 \quad (45.98)$$

La notation $\tilde{\not{\partial}}$ signifiant que l'opérateur $\not{\partial}$ opère sur $\bar{\Psi}$ vers la gauche tel que:

$$\bar{\Psi} \tilde{\not{\partial}} = (\partial_\mu \bar{\Psi}) \gamma^\mu \quad (45.99)$$

Remarque: Certains auteurs écrivent $(i \not{\partial} + m_0) \bar{\Psi} = 0$ mais ceci est faux car $\bar{\Psi}$ est une matrice ligne comme nous l'avons fait remarquer plus haut!!!

Finalement nous avons pour les équations de Dirac libres:

$$\begin{aligned} (i \not{\partial} - m_0) \Psi &= 0 \Leftrightarrow i \gamma^\mu \partial_\mu \Psi - m_0 \Psi = 0 \\ \bar{\Psi} (i \tilde{\not{\partial}} + m_0) &= 0 \Leftrightarrow i (\partial_\mu \bar{\Psi}) \gamma^\mu + m_0 \bar{\Psi} = 0 \end{aligned} \quad (45.100)$$

Supposons maintenant que le "lagrangien du champ spinoriel de Dirac libre" soit de la forme (parce que finalement c'est le lagrangien qui nous intéresse):

$$L = \bar{\Psi} (i \gamma^\mu \partial_\mu - m_0) \Psi \quad (45.101)$$

où nous avons posé $\bar{\Psi} = \Psi^\dagger \gamma^0$. Il s'agit donc du lagrangien du champ spinoriel pour les particules de spin 1/2 qui sont donc des fermions libres.

En considérant les quantités $\bar{\Psi}, \Psi$ comme indépendantes (c'est ce qu'elles sont de toute façon puisque orthogonales) et choisissant le champ spinoriel comme $\Psi^\dagger$, nous avons l'équation d'Euler-Lagrange:

$$\frac{\partial \mathcal{L}}{\partial \bar{\Psi}} - \partial_\mu \left(\frac{\partial \mathcal{L}}{\partial (\partial_\mu \bar{\Psi})} \right) = 0 \quad (45.102)$$

Le deuxième terme est nul puisque le lagrangien de Dirac ne contient pas de termes en $\partial_\mu \Psi^\dagger$. De fait il reste:

$$\frac{\partial \mathcal{L}}{\partial \bar{\Psi}} = \frac{\partial (\bar{\Psi} i \gamma^\mu \partial_\mu \Psi - \bar{\Psi} \Psi m_0)}{\partial \bar{\Psi}} = i \gamma^\mu \partial_\mu \Psi - \Psi m_0 = (i \not{\partial} - m_0) \Psi = 0 \quad (45.103)$$

Nous retombons donc bien sur l'équation de Dirac libre (le même développement pouvant être fait pour l'équation de Dirac libre adjointe)! Ainsi, dans ce cadre, la seule manière d'expliquer les propriétés quantiques de la matière comportant des particules avec spin 1/2 est de faire intervenir des champs $\bar{\Psi}, \Psi$ représentant des particules chargées électriquement, les électrons et positrons comme nous le savons. Nous appelons alors ces entités des "**champs (spinoriels) de Dirac**".

3. THÉORIES DE JAUGE

Nous allons voir maintenant une approche simple d'un outil qui a révolutionné l'approche de la physique moderne des particules au milieu du 20ème siècle et qui a valu plusieurs prix Nobel à ceux qui y ont contribué.

Nous conseillons très fortement avant de lire ce qui va suivre que le lecteur aille jeter aussi un coup d'oeil préalable sur le sous-chapitre de théorie des Jauges du chapitre d'Électrodynamique, car c'est un premier exemple d'une invariance de jauge faisant apparaître un champ (le potentiel vecteur) indispensable pour expliquer certains phénomènes à l'échelle quantique comme l'explique clairement l'équation de Pauli (cf. chapitre de Physique Quantique Relativiste).

Depuis le début des années 80, les magazines de vulgarisation parlent beaucoup en physique quantique des théories de jauge. Les interactions électromagnétiques et les interactions faibles sont décrites conjointement par une théorie de jauge élaborée par Glashow, Weinberg et Salam. Les interactions fortes semblent aussi correctement décrites par une théorie de jauge. C'est dans le cadre de ces théories de jauge que les physiciens théoriciens tentent d'unifier les diverses interactions fondamentales de la nature. Il convient donc, même sur ce site qui traite de manière élémentaire de physique quantique, de parler de théorie de jauge dans le cadre de ce domaine.

Pour ce faire, nous considérerons déjà comme connu le contexte qui mena à la découverte de l'invariance de jauge dans le cadre de l'électrodynamique (voir chapitre du même nom pour les détails) et ferons un rapprochement avec certains développements vus dans le chapitre de Relativité Générale et le rôle qu'a joué Weyl dans la mise en évidence des principes fondamentaux d'une théorie de jauge.

Rappelons que la relativité restreinte et générale reposent sur le postulat qu'il n'existe dans l'univers

aucun référentiel absolu. Nous avons vu dans le chapitre de Relativité Restreinte en long et en large que les relations qui permettent de passer les lois de la physique d'un repère à l'autre ne dépendent que de la vitesse relative entre les référentiels. Ainsi, la relativité restreinte est une théorie à symétrie globale. Nous avons également vu en long et en large dans le chapitre de Relativité Générale que la connexion affine est le lien entre les référentiels de la théorie locale (approximation des champs faibles) qu'est la relativité générale.

En 1919 eut lieu la première observation expérimentale de la déviation de la lumière d'une étoile par le champ gravitationnel du Soleil. Cette confirmation spectaculaire de la théorie de la relativité générale inspira Hermann Weyl, qui proposa la même année une conception révolutionnaire de l'invariance de jauge: Si les effets d'un champ gravitationnel peuvent être décrits par une connexion exprimant l'orientation relative entre des référentiels locaux de l'espace-temps, d'autres forces de la nature telles que l'électromagnétisme peuvent-elles être associées aussi à des connexions similaires?

Nous considérons deux types de symétrie de jauge: l'une dite "jauge globale" et l'autre dite "jauge locale". Elles se distinguent par le paramètre caractérisant le changement de phase de la fonction d'onde (nous verrons cela en détails un peu plus loin).

3.1. INVARIANCE DE JAUGE GLOBALE

Nous allons donc étudier l'invariance de jauge à partir de l'équation de Schrödinger et montrer que même si les résultats peuvent paraître déroutants (dans le cadre d'applications complexes) ils n'en restent pas moins mathématiquement corrects.

Remarque: L'invariance de jauge globale est rigoureusement dénommée "symétrie globale".

Considérons donc l'équation de Schrödinger:

$$H\Psi = i\hbar \frac{\partial}{\partial t} \Psi \quad (45.104)$$

avec comme nous l'avons montré:

$$H = -\frac{\hbar^2}{2m} \nabla^2 + V \quad (45.105)$$

avec $\Psi = \Psi(\vec{r}, t)$. Soit dans le cas d'une particule libre:

$$H = -\frac{\hbar^2}{2m} \nabla^2 = \hbar^2 \frac{(-i\nabla)^2}{2m} \quad (45.106)$$

Cet opérateur est manifestement invariant dans la transformation qui fait passer de Ψ à Ψ' avec:

$$\Psi' = e^{ig\alpha} \Psi \Rightarrow \Psi = e^{-ig\alpha} \Psi' \quad (45.107)$$

où g est une constante de couplage (pour assurer l'homogénéité des unités et l'amplitude) étant considérée comme un nombre réel et α un paramètre réel indépendant des coordonnées (dans un premier temps...) d'espace et de temps.

$$H\Psi = i\hbar \frac{\partial}{\partial t} \Psi \quad (45.108)$$

devient:

$$He^{-ig\alpha}\Psi' = i\hbar \frac{\partial}{\partial t} e^{-ig\alpha}\Psi' \quad (45.109)$$

et comme α ne dépend ni de x, y, z, t alors:

$$e^{-ig\alpha} H\Psi' = e^{-ig\alpha} i\hbar \frac{\partial}{\partial t} \Psi' \quad (45.110)$$

Soit après simplification:

$$H\Psi' = i\hbar \frac{\partial}{\partial t} \Psi' \quad (45.111)$$

La forme de l'équation est restée la même lorsque nous avons fait le changement de Ψ en Ψ' .

Ainsi, la description d'un système libre n'est pas affectée par le changement de phase globale. En langage de la théorie des groupes (cf. chapitre d'Algèbre Ensembliste), nous parlons d'invariance sous le groupe $U(1)$ des phases.

En d'autres termes pour parler comme les physiciens...:

$$e^{ig\alpha}\Psi \quad (45.112)$$

définit une transformation de jauge par la rotation α (le paramètre au sens des groupes de Lie).

L'ensemble des rotations forme un groupe nommé $U(1)$ que l'usage appelle le groupe de jauge (isomorphe de $SO(2)$).

L'ensemble des $e^{ig\alpha}$ forme une représentation monodimensionnelle du groupe $U(1)$ que nous appelons la représentation g . Il y a bien entendu une infinité de représentations g (autant qu'il y a de valeurs de g !).

Comme le paramètre α ne dépend pas de la position et du temps, nous disons que le système est invariant par transformation de jauge globale (partout en même temps) ou simplement un invariant de $U(1)$ dans le temps et l'espace.

3.2. INVARIANCE DE JAUGE LOCALE

Mais mais... soit l'invariance de jauge globale montre que nous avons une équation qui reste valable dans le cadre d'un changement de phase fixe. Mais maintenant dans un laboratoire cette équation de Schrödinger doit être valable même si la phase dépend de la position et du temps. Cette contrainte s'appelle une "**invariance locale**".

Nous considérons cette fois que α est une fonction $\alpha(\vec{r}, t)$ et l'idée bien évidemment est de vérifier si l'équation de Schrödinger reste invariante dans la transformation:

$$\Psi \rightarrow \Psi' = e^{ig\alpha} \Psi \quad (45.113)$$

Il est dès lors évident que l'équation de Schrödinger:

$$H\Psi = i\hbar \frac{\partial}{\partial t} \Psi \quad (45.114)$$

n'est plus invariante. Effectivement, nous voyons rapidement que rien que l'opérateur $\vec{\nabla}^2$ dans l'hamiltonien va poser problème en faisant apparaître des termes gênants qui ne s'annuleront pas:

$$\begin{aligned} \vec{\nabla}\Psi &= e^{-ig\alpha} \left(\vec{\nabla}\Psi' \right) - ig e^{-ig\alpha} \left(\vec{\nabla}\alpha \right) \Psi' \\ \vec{\nabla}^2\Psi &= \vec{\nabla} \left(e^{-ig\alpha} \left(\vec{\nabla}\Psi' \right) \right) - ig \vec{\nabla} \left(e^{-ig\alpha} \left(\vec{\nabla}\alpha \right) \Psi' \right) \\ &= \left[e^{-ig\alpha} \left(\vec{\nabla}^2\Psi' \right) - ig e^{-ig\alpha} \left(\vec{\nabla}\alpha \right) \left(\vec{\nabla}\Psi' \right) \right] - ig \left[\left(-ig \right) e^{-ig\alpha} \left(\vec{\nabla}\alpha \right)^2 \Psi' + e^{-ig\alpha} \vec{\nabla}^2\alpha \Psi' + e^{-ig\alpha} \vec{\nabla}\alpha \vec{\nabla}\Psi' \right] \\ &\dots \end{aligned} \quad (45.115)$$

Pour contourner ce problème, nous introduisons le champ de force associé au potentiel vecteur et au potentiel électrique et nous verrons qu'il garantit l'invariance locale (donc il est impossible d'envisager un changement de phase invariant sans la présence d'un champ de force de ce type). L'invariance locale impose que la particule ne soit plus libre (il n'existe donc pas de particules chargées libres!).

Pour cela, reprenons l'hamiltonien de l'équation de Pauli (cf. chapitre de Physique Quantique Relativiste):

$$\frac{(\vec{p} - q\vec{A})^2}{2m_0} - \frac{\hbar q}{2m_0} \sigma \cdot \vec{B} + qU \quad (45.116)$$

et négligeons l'interaction entre le spin et le champ magnétique tel que l'hamiltonien devienne:

$$\frac{(\vec{p} - q\vec{A})^2}{2m_0} + qU \quad (45.117)$$

Soit:

$$\frac{(-i\hbar\vec{\nabla} - q\vec{A})^2}{2m_0} + qU \quad (45.118)$$

Nous obtenons donc l'équation de Schrödinger suivante:

$$\left[\frac{(-i\hbar\vec{\nabla} - q\vec{A})^2}{2m_0} + qU \right] \Psi = i\hbar \frac{\partial}{\partial t} \Psi \quad (45.119)$$

Ce qui par rapport à l'équation de Schrödinger libre:

$$\frac{(-i\hbar\vec{\nabla})^2}{2m}\Psi = i\hbar\frac{\partial}{\partial t}\Psi \quad (45.120)$$

fait intervenir les correspondances suivantes:

$$\begin{aligned}\vec{\nabla} &\rightarrow \vec{D} \equiv \vec{\nabla} - i\frac{q}{\hbar}\vec{A} = \\ \frac{\partial}{\partial t} &\rightarrow D^0 \equiv \frac{\partial}{\partial t} + i\frac{q}{\hbar}U\end{aligned} \quad (45.121)$$

Considérons la transformation de jauge (cf. chapitre d'Électrodynamique) en notant dorénavant le potentiel électrique par la lettre V :

$$\begin{aligned}\vec{A} &\rightarrow \vec{A}' = \vec{A} + \vec{\nabla}f \\ V &\rightarrow V' = V - \frac{\partial f}{\partial t}\end{aligned} \quad (45.122)$$

où $f = f(\vec{r}, t)$.

D'abord, nous voyons alors immédiatement que les opérateurs sont invariants. Effectivement:

$$\begin{aligned}\left(\vec{\nabla} - i\frac{q}{\hbar}\vec{A}'\right)e^{ig\alpha} &= \left(\vec{\nabla} - i\frac{q}{\hbar}(\vec{A} + \vec{\nabla}f)\right)e^{ig\alpha} = \left(\vec{\nabla} - i\frac{q}{\hbar}\vec{A} - i\frac{q}{\hbar}\vec{\nabla}f\right)e^{ig\alpha} \\ &= ig e^{ig\alpha}(\vec{\nabla}\alpha) + e^{ig\alpha}\vec{\nabla} - i\frac{q}{\hbar}\vec{A}e^{ig\alpha} - i\frac{q}{\hbar}\vec{\nabla}fe^{ig\alpha} \\ &= e^{ig\alpha}\left(ig\vec{\nabla}\alpha + \vec{\nabla} - i\frac{q}{\hbar}\vec{A} - i\frac{q}{\hbar}\vec{\nabla}f\right)\end{aligned} \quad (45.123)$$

Or, si g est posé comme étant $q/\hbar$ et f comme étant α alors nous avons:

$$\left(\vec{\nabla} - i\frac{q}{\hbar}\vec{A}'\right)e^{i\frac{q}{\hbar}\alpha} = e^{i\frac{q}{\hbar}\alpha}\left(i\frac{q}{\hbar}\cancel{\vec{\nabla}f} + \vec{\nabla} - i\frac{q}{\hbar}\vec{A} - i\frac{q}{\hbar}\cancel{\vec{\nabla}f}\right) = e^{i\frac{q}{\hbar}\alpha}\left(\vec{\nabla} - i\frac{q}{\hbar}\vec{A}\right) \quad (45.124)$$

Soit tout simplement:

$$\left(\vec{\nabla} - i\frac{q}{\hbar}\vec{A}'\right)e^{i\frac{q}{\hbar}\alpha} = e^{i\frac{q}{\hbar}\alpha}\left(\vec{\nabla} - i\frac{q}{\hbar}\vec{A}\right) \quad (45.125)$$

De même en sachant maintenant que f est α :

$$\begin{aligned}
\left(\frac{\partial}{\partial t} + i\frac{q}{\hbar}V\right)e^{i\frac{q}{\hbar}\alpha} &= \left(\frac{\partial}{\partial t} + i\frac{q}{\hbar}\left(V - \frac{\partial\alpha}{\partial t}\right)\right)e^{i\frac{q}{\hbar}\alpha} \\
&= i\frac{q}{\hbar}e^{i\frac{q}{\hbar}\alpha}\frac{\partial\alpha}{\partial t} + e^{i\frac{q}{\hbar}\alpha}\frac{\partial}{\partial t} + i\frac{q}{\hbar}Ve^{i\frac{q}{\hbar}\alpha} - i\frac{q}{\hbar}\frac{\partial\alpha}{\partial t}e^{i\frac{q}{\hbar}\alpha} = \quad (45.126) \\
&= e^{i\frac{q}{\hbar}\alpha}\left(\frac{\partial}{\partial t} + i\frac{q}{\hbar}V\right)
\end{aligned}$$

Nous avons donc:

$$\begin{aligned}
\left(\vec{\nabla} - i\frac{q}{\hbar}\vec{A}\right)e^{i\frac{q}{\hbar}\alpha} &= e^{i\frac{q}{\hbar}\alpha}\left(\vec{\nabla} - i\frac{q}{\hbar}\vec{A}\right) \\
\left(\frac{\partial}{\partial t} + i\frac{q}{\hbar}V\right)e^{i\frac{q}{\hbar}\alpha} &= e^{i\frac{q}{\hbar}\alpha}\left(\frac{\partial}{\partial t} + i\frac{q}{\hbar}V\right)
\end{aligned} \quad (45.127)$$

Soit:

$$\begin{aligned}
\vec{D}'e^{i\frac{q}{\hbar}\alpha} &= e^{i\frac{q}{\hbar}\alpha}\vec{D} \\
(D^0)'e^{i\frac{q}{\hbar}\alpha} &= e^{i\frac{q}{\hbar}\alpha}(D^0)
\end{aligned} \quad (45.128)$$

La relation:

$$\frac{(-i\hbar\vec{\nabla})^2}{2m}\Psi = i\hbar\frac{\partial}{\partial t}\Psi \quad (45.129)$$

devient alors avec les nouvelles correspondances:

$$\frac{(-i\hbar\vec{D})^2}{2m}\Psi = i\hbar D^0\Psi \quad (45.130)$$

et avec les développements antérieurs, nous avons donc:

$$\begin{aligned}
e^{ig\alpha}\frac{(-i\hbar\vec{D})^2}{2m}\Psi &= \frac{(-i\hbar\vec{D}')^2}{2m}e^{ig\alpha}\Psi \\
e^{ig\alpha}i\hbar D^0\Psi &= i\hbar D^0'e^{ig\alpha}\Psi
\end{aligned} \quad (45.131)$$

Soit:

$$\frac{(-i\hbar\vec{D}')^2}{2m}e^{ig\alpha}\Psi = i\hbar D^0'e^{ig\alpha}\Psi \quad (45.132)$$

Ce qui donne après simplification:

$$\frac{(-i\hbar\vec{D})^2}{2m}\Psi = i\hbar D^0\Psi \quad (45.133)$$

Ainsi, en demandant l'invariance de jauge, nous avons fait apparaître une interaction... et nous savons bien qu'elle est cette interaction!

L'équation de Schrödinger d'une particule se déplaçant dans un champ électromagnétique est donc invariante sous la transformation locale de phase. La phase d'une fonction d'onde est bel et bien une nouvelle variable locale au sens de Weyl et le potentiel électromagnétique peut être interprété, suivant Weyl, comme une connexion reliant les phases en différents points.

Nous en concluons que le champ électromagnétique est une conséquence de l'invariance de jauge locale fondée sur le groupe $U(1)$, groupe des matrices unitaires à une dimension (cf. chapitre d'Algèbre Ensembliste). L'intérêt qui s'en dégage est de construire des théories de jauge sur des groupes plus compliqués (non-abéliens): ces théories sont appelées "théories de Yang-Mills".

Maintenant allons un tout petit peu plus loin mais, sans trop approfondir... Nous avons montré plus haut que le lagrangien de l'équation de Dirac libre était:

$$L = \bar{\Psi} \left(i\gamma^\mu \partial_\mu - m_0 \right) \Psi \quad (45.134)$$

Or, cet opérateur ne faisant pas apparaître le champ électromagnétique, on se doute très fortement qu'il ne porte pas en lui une invariance pour une jauge locale...

Or, l'équivalent de l'opérateur divergence $\vec{\nabla}$ dans l'équation de Schrödinger libre est la dérivée covariante ∂_μ . Donc au même titre que nous avons associé pour garantir l'invariance locale de jauge de l'équation de Schrödinger libre:

$$\begin{aligned} \vec{\nabla} &\rightarrow \vec{D} \equiv \vec{\nabla} - i\frac{q}{\hbar}\vec{A} \\ \frac{\partial}{\partial t} &\rightarrow D^0 \equiv \frac{\partial}{\partial t} + i\frac{q}{\hbar}U \end{aligned} \quad (45.135)$$

il est tentant de combiner le tout en un nouvel opérateur:

$$D_\mu \equiv \partial_\mu - i\frac{q}{\hbar}A_\mu \quad (45.136)$$

avec:

$$D_0 \equiv \partial_0 + i\frac{q}{\hbar}A_0 = \frac{\partial}{c\partial t} + i\frac{q}{\hbar}(-U)$$

Le lagrangien de l'équation de Dirac libre s'écrirait alors:

$$L = \bar{\Psi} \left(i\gamma^\mu D_\mu - m_0 \right) \Psi \quad (45.137)$$

Soit:

$$\begin{aligned}
 L &= \bar{\Psi} \left(i \gamma^\mu \left(\partial_\mu - i \frac{q}{\hbar} A_\mu \right) - m_0 \right) \Psi = \bar{\Psi} \left(i \gamma^\mu \partial_\mu - m_0 \right) \Psi - \bar{\Psi} i \gamma^\mu \frac{q}{\hbar} A_\mu \Psi \\
 &= \bar{\Psi} \left(i \gamma^\mu \partial_\mu - m_0 \right) \Psi + \bar{\Psi} \gamma^\mu \Psi \frac{q}{\hbar} A_\mu = \bar{\Psi} \left(i \gamma^\mu \partial_\mu - m_0 \right) \Psi - j^\mu A_\mu
 \end{aligned} \tag{45.138}$$

avec:

$$j^\mu = -\frac{q}{\hbar} \bar{\Psi} \gamma^\mu \Psi \tag{45.139}$$

Il ne reste plus qu'à rajouter le terme du champ pour obtenir le lagrangien total de l'équation de Dirac (cela aurait été relativement dur de le trouver d'une autre manière...):

$$L = \bar{\Psi} \left(i \gamma^\mu \partial_\mu - m_0 \right) \Psi - j^\mu A_\mu + \frac{1}{4} \mu_0 F_{\mu\nu} F^{\mu\nu} \tag{45.140}$$

qui correspond aux équations de Dirac-Maxwell et qui est le "**lagrangien de l'électrodynamique quantique des champs**" où à gauche nous avons le terme des fermions et à droite la partie d'interaction des bosons de masse nulle (photons).

Donc le fait d'avoir rajouté au lagrangien libre une condition d'invariance par des transformations locales, nous a conduits à une théorie avec interaction que nous pouvons écrire avec plus de rigueur et sous forme développée:

$$\mathcal{L}_{QED} = \bar{\Psi} \left(i \gamma^\mu \partial_\mu - m_0 \right) \Psi + \frac{q}{\hbar} \bar{\Psi} \gamma^\mu \Psi A_\mu + \frac{1}{4} \mu_0 F_{\mu\nu} F^{\mu\nu} \tag{45.141}$$

ou encore en unités naturelles et avec la charge de l'électron:

$$\mathcal{L}_{QED} = \bar{\Psi} \left(i \gamma^\mu \partial_\mu - m_0 \right) \Psi + q \bar{\Psi} \gamma^\mu \Psi A_\mu + \frac{1}{4} F_{\mu\nu} F^{\mu\nu} \tag{45.142}$$

L'électrodynamique quantique a fait défaut cependant dans les années 1940 pour décrire bon nombre de particules mises en évidence par les accélérateurs. Certes, d'une certaine manière, elle a été étendue pour décrire de nouvelles particules. Mais beaucoup d'entre elles semblaient jouir de propriétés dont l'électrodynamique quantique ne pouvait rendre compte.

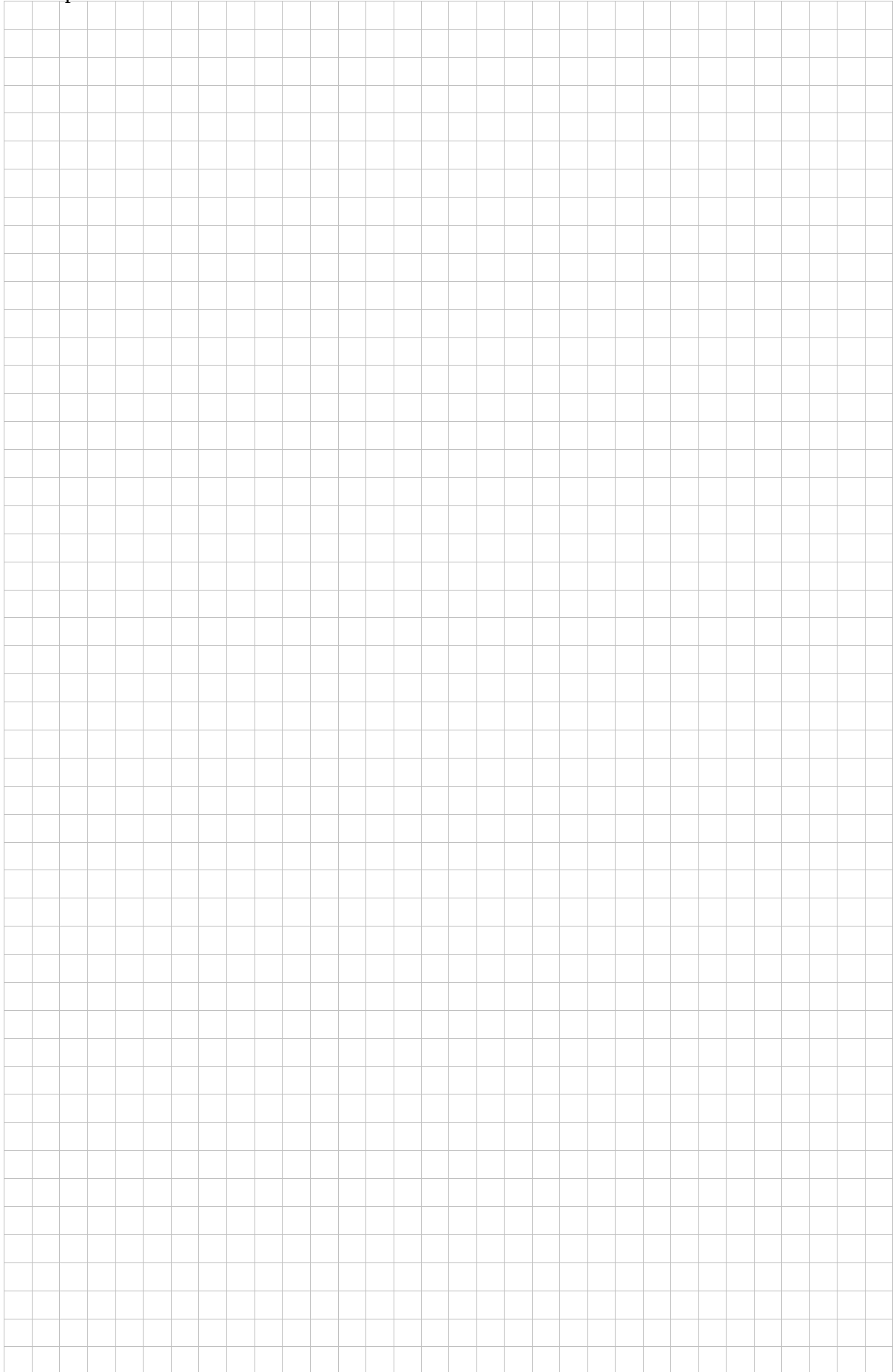
En fait, la raison est simple... c'est une théorie dans laquelle aucune solution exacte n'est connue, une situation qui perdure jusqu'à nos jours (2008). La seule méthode de calcul disponible est appelée développement perturbatif. L'idée est essentiellement la même que celle du développement limité que l'on pratique dans le domaine du calcul différentiel. En l'occurrence, si nous ne savons pas calculer la valeur d'une fonction, nous la décomposons en une séquence de polynômes et l'approximation s'affine au fur et à mesure que nous prenons en compte des termes de degrés de plus en plus élevés. Un tel développement en série commence par un terme d'ordre zéro, qui est juste la valeur de la fonction inconnue en un certain point où l'on sait calculer cette fonction.

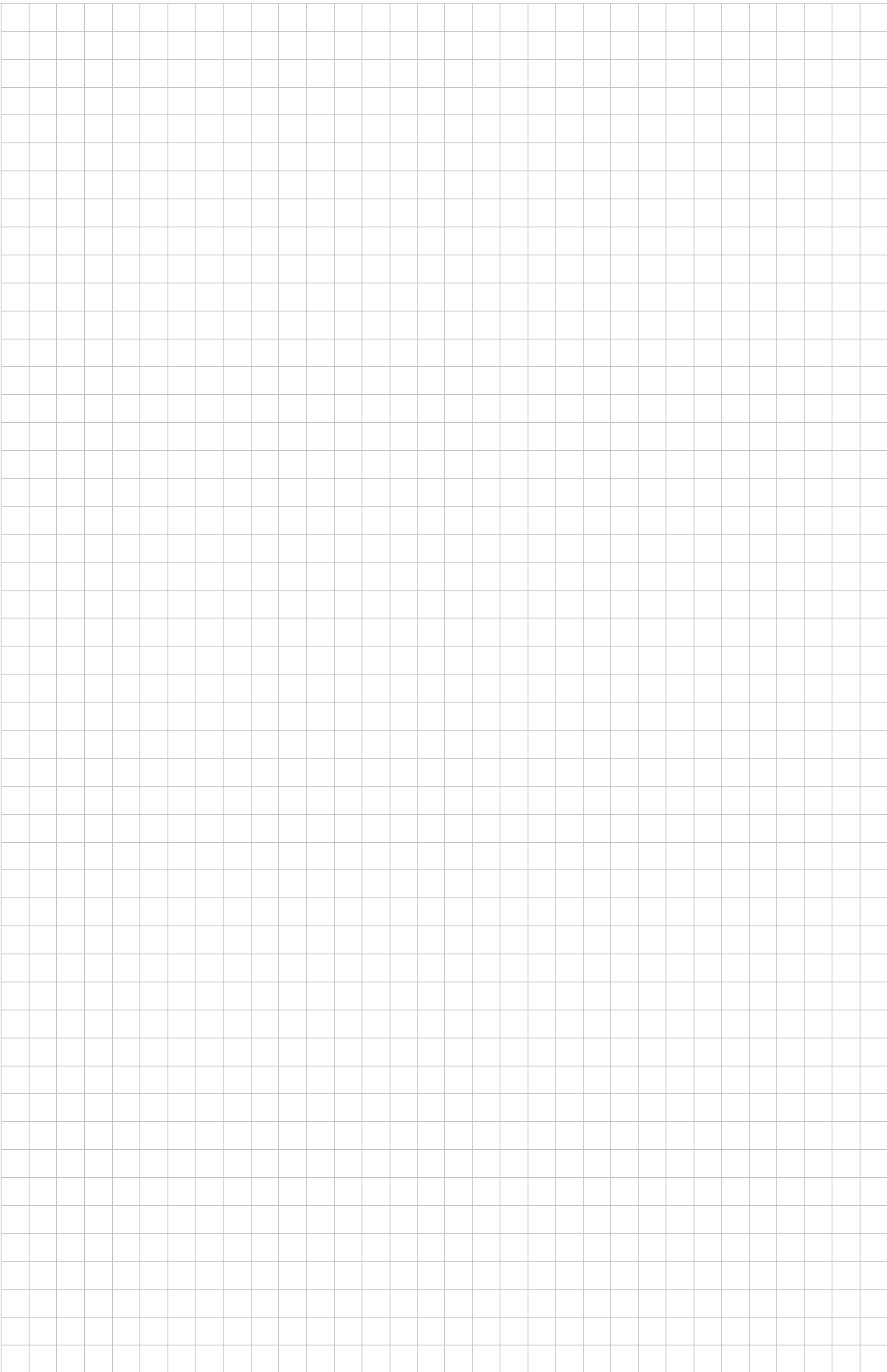
Dans le cas du développement perturbatif de l'électrodynamique quantique, le terme d'ordre zéro représente la propagation pure, sans interaction (l'intensité de l'interaction entre l'électron et le champ magnétique est mise à zéro). Dans cette approximation, l'électrodynamique quantique est une théorie des particules libres et elle est exactement calculable. Nous avons des électrons, des positons et des

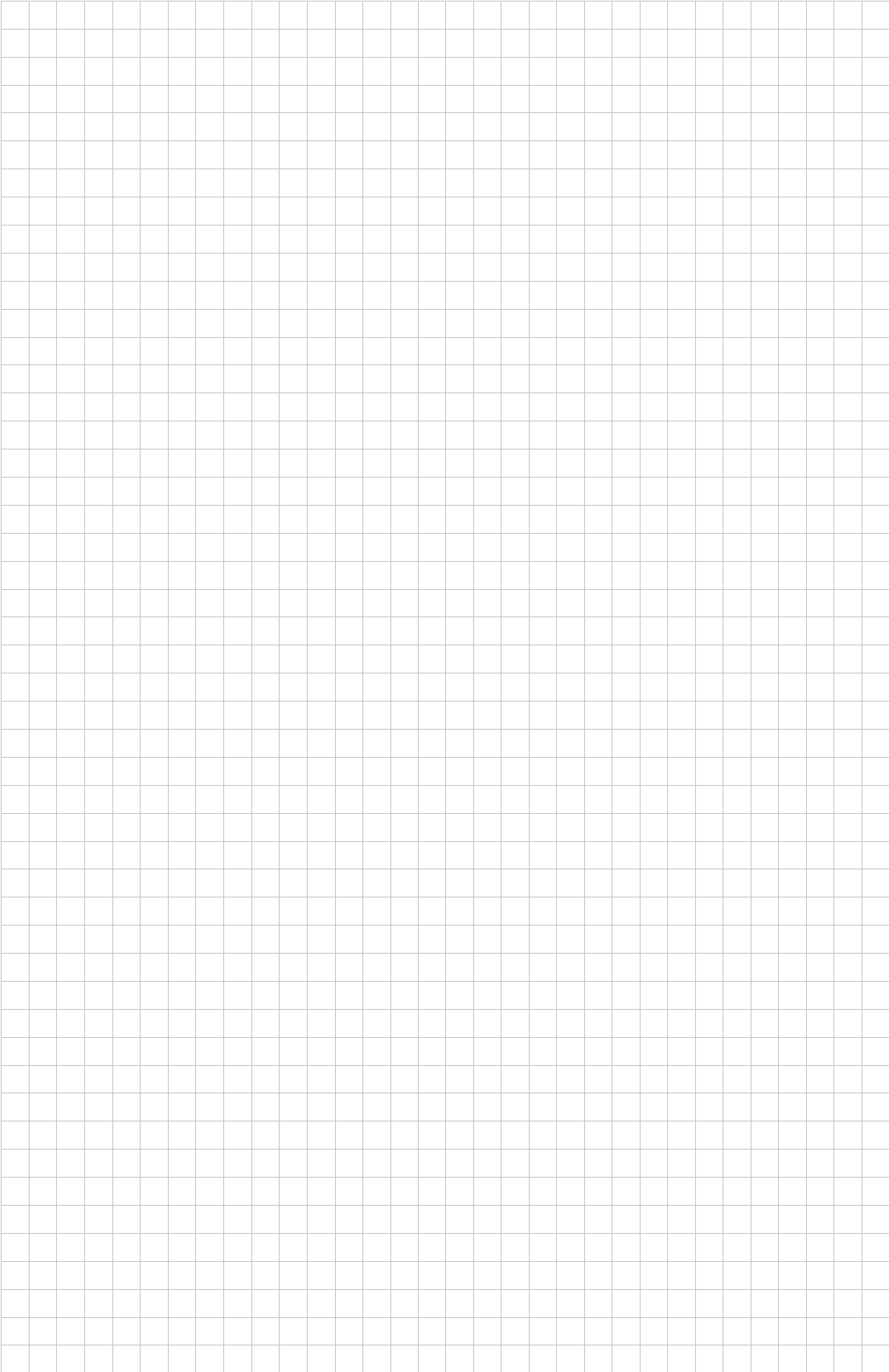
photons mais ils se croisent sans s'influencer. Le terme suivant dans le développement en série, celui du premier ordre, est aussi exactement calculable. Dans cette approximation, la théorie semble refléter assez fidèlement le monde réel. Des phénomènes physiques très intéressants apparaissent dans cette approximation de premier ordre de la théorie réelle de l'interaction photon-électron et la théorie s'accorde bien avec les résultats expérimentaux.

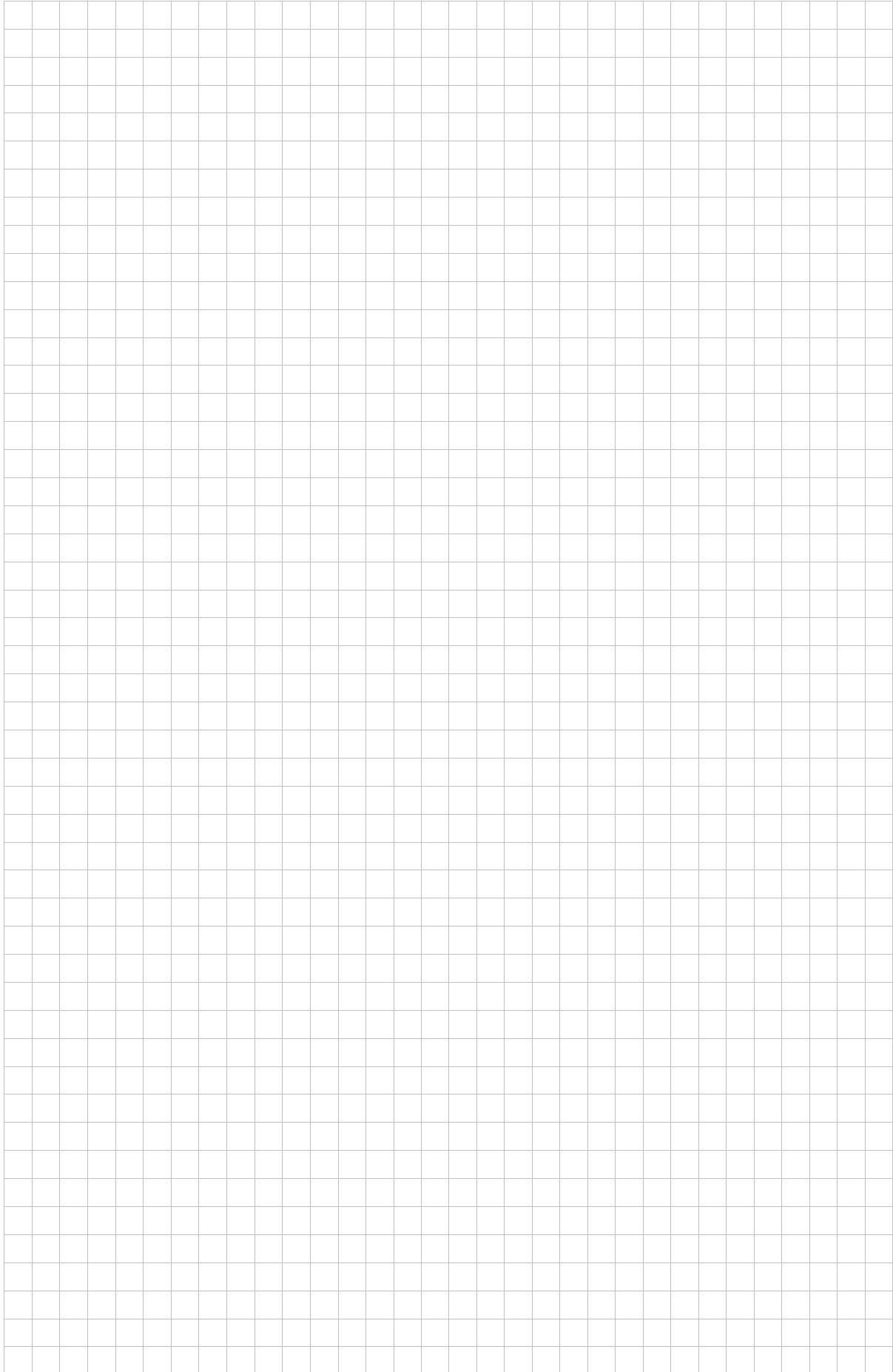
Malheureusement on eut tôt fait de découvrir que le calcul des termes de second ordre et des termes plus élevés semblait dénué de sens jusqu'à donner des valeurs infinies... aujourd'hui il n'existe encore que des méthodes de résolution approximatives et non totalement satisfaisantes dès lors on a été obligé de chercher une autre technique d'approximation se basant sur une renormalisation des équations... et les résultats sont extraordinairement bons (à la 11ème décimale près!) mais au fond cela sent un peu le bricolage sur mesure quand même...

Notes personnelles:









46. PHYSIQUE DES PARTICULES ÉLÉMENTAIRES

Nous avons déjà mentionné dans le chapitre de Physique Nucléaire que nous constatons donc expérimentalement que les noyaux radioactifs n'émettent pas de neutrons ou de protons. Mais nous pouvons nous interroger: comment font-ils pour synthétiser une particule alpha, ou transformer un neutron en proton, ou vice et versa? Pour répondre à ces questions, examinons les forces en présence.

Avant la découverte de la radioactivité, les physiciens avaient identifié deux forces fondamentales: la force de gravitation et la force électromagnétique. La découverte de la radioactivité et les études concernant le noyau atomique ont conduit les physiciens à introduire non pas une mais deux nouvelles forces fondamentales!

Avant même de connaître la composition exacte des noyaux, pour expliquer l'existence de ces systèmes minuscules et portants parfois de fortes charges positives, les physiciens avaient pressenti la nécessité d'une force de cohésion puissante capable de dominer la répulsion électrostatique s'exerçant entre ces charges (rappelons que nous avons vu en mécanique classique que la force gravitationnelle entre deux corps de masses équivalentes à celles de particules est totalement négligeable). Comme le noyau est petit, cette "**force nucléaire**" devait s'exercer à très courte distance. Quand J. Chadwick découvrit le neutron, il fut démontré expérimentalement que force attractive s'exerçait aussi bien entre deux neutrons, deux protons et entre un neutron et un proton. Dès 1935, H. Yukawa en élaborait une théorie dont les grandes lignes sont encore acceptées, mais qui doivent être améliorées suites aux défauts qui ont été mis en évidence (cf. chapitre de Physique Quantique Des Champs).

Cependant, comme nous le savons déjà, cette force nucléaire n'expliquait pas la transformation d'un neutron en proton, qui a lieu dans la radioactivité bêta-. Il fallut introduire une quatrième force fondamentale, d'intensité plus faible, baptisée pour cette raison "**interaction faible**", la force nucléaire devenant ipso facto "**l'interaction forte**".

Ainsi, en principe, la radioactivité met en jeu les quatre forces fondamentales de la Nature: la gravitation et la force électromagnétique, puisque les particules alpha et bêta possèdent une masse et une charge, et les deux interactions nucléaires, forte et faible (en fait, la gravitation, d'intensité bien moindre que les trois autres aux échelles subatomiques est souvent négligée).

Nous avons partiellement abordé dans le chapitre de Physique Quantique Des Champs les interactions fondamentales et leurs vecteurs d'interactions. Avant de nous lancer dans des calculs ardu, il est souhaitable d'abord d'acquérir un certain vocabulaire d'usage courant chez les physiciens théoriciens.

Le concept le plus simple à aborder dans le domaine de la physique des particules élémentaires est la comparaison des quatre forces élémentaires via leur constante de couplage respective (c'est un truc que les physiciens aiment bien...).

Remarque: Hubert Reeves et ses collègues astrophysiciens ont démontré qu'à l'époque de la genèse de l'Univers, la moindre déviance des constantes de couplage des valeurs nominales actuelles aurait provoqué l'instabilité des nucléons et aurait condamné l'évolution cosmique.

1. CONSTANTES DE COUPLAGES

Nous allons essayer ici de classer les quatre forces selon leur intensité via l'utilisation de "constantes de couplage".

Pour cela, il faut calculer les quatre interactions pour deux mêmes particules, par exemple deux protons, à des distances identiques, donc de type nucléaire, et les comparer à une grandeur commune de même dimension de sorte que leur rapport fournisse un nombre sans dimension.

Cette grandeur commune sera choisie comme étant le produit:

$$\frac{\hbar}{2\pi} c = \hbar c = 3.16 \cdot 10^{-26} [N \cdot m^2] \quad (46.1)$$

Nous trouvons ainsi:

1. Pour la force de gravitation (cf. chapitre d'Astronomie) où:

$$F \cdot r^2 = G \cdot m \cdot M [N \cdot m^2] \quad (46.2)$$

avec la masse du proton telle que $m = M = 1.67 \cdot 10^{-27} [kg]$, la constante de couplage de la force de gravitation vaut alors par définition:

$$\alpha_G = \frac{G \cdot m \cdot M}{\hbar c} = \frac{G \cdot m_p^2}{\hbar c} \cong 6 \cdot 10^{-39} \quad (46.3)$$

2. Pour la force électrique (cf. chapitre d'Électrostatique) où:

$$F \cdot r^2 = \frac{1}{4\pi\epsilon_0} \cdot q \cdot Q [N \cdot m^2] \quad (46.4)$$

avec les charges des protons telles que $q = Q = 1.602 \cdot 10^{-19} [C]$, la constante de couplage de la force électrique vaut alors par définition:

$$\alpha = \frac{\frac{1}{4\pi\epsilon_0} \cdot q \cdot Q}{\hbar c} \cong \frac{1}{137} \quad (46.5)$$

Remarque: Nous retrouvons ici la "constante de structure fine" que nous avons vue déjà dans le chapitre de Physique Quantique Corpusculaire. On comprend ainsi mieux le choix de départ pour la comparaison relative des interactions.

3. Pour la force nucléaire forte ("strong" en anglais), où F représente la "**charge nucléaire forte**", la constante de couplage forte vaut (attention la valeur dépend du modèle théorique choisi!):

$$\alpha_s = \frac{F \cdot F}{\hbar c} = \frac{g_s^2}{\hbar c} \cong 1 \quad (46.6)$$

d'où son nom.

4. Pour la force nucléaire faible ("weak" en anglais) responsable de la désintégration des particules, f représente la "**charge nucléaire faible**", et sa constante de couplage faible vaut (attention la valeur dépend du modèle théorique choisi!):

$$\alpha_w = \frac{f \cdot f}{\hbar c} = \frac{g_w^2}{\hbar c} \cong 10^{-6} \quad (46.7)$$

Ainsi tout cela peut se résumer dans le tableau phénoménologique suivant:

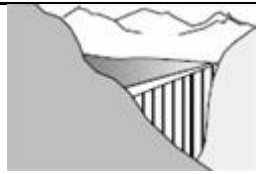
Les 4 interactions fondamentales	Description phénoménologique	Phénomènes macroscopiques	Intensité
Gravitationnelle	$F = G \frac{m_1 m_2}{r^2}$ Newton/Einstein	 Centrale hydraulique	$\alpha_G \cong 6 \cdot 10^{-39}$
Électromagnétique	$F = \frac{1}{4\pi\epsilon} \frac{q_1 q_2}{r^2} \frac{\vec{r}}{\ \vec{r}\ } = q \cdot \vec{E}$ $\vec{F} = q \cdot \vec{E} + \vec{v} \times \vec{B}$ Maxwell	 Groupes turbines-alternateurs	$\alpha \cong 1/137$
Force Faible	$U(r) = \frac{C}{r} e^{-r/R}$ $R = \frac{\hbar}{m_0 c}$ Yukawa	 Pile thermoélectrique avec isotope radioactif	$\alpha_w = g_w^2 / \hbar c$ $\cong 10^{-6}$
Forte	$U(r) = \frac{C}{r} e^{-r/R}$ $R = \frac{\hbar}{m_0 c}$ Yukawa	 Centrale nucléaire	$\alpha_s = g_s^2 / \hbar c$ $\cong 1$

Tableau: 46.1 - Résumé des 4 interactions fondamentales avec lois et constantes associées

ou encore avec le diagramme suivant (plus intéressant) où nous retrouvons, en tenant compte des résultats que nous avons tirés lors de notre étude des champs massiques et non massiques dans le cadre du modèle de Yukawa (cf. chapitre de Physique Quantique Des Champs):

1. En ordonnée à l'origine l'intensité des forces telle que calculée précédemment en fonction de la distance selon le modèle de Yukawa des champs massiques (interactions faible et forte) et non massiques (interactions électr. et gravitationnelle)
2. Les schémas représentatifs (diagrammes de Feynman) des interactions conformément aux résultats obtenus et particules déjà mentionnées dans le chapitre de Physique Quantique Des Champs.

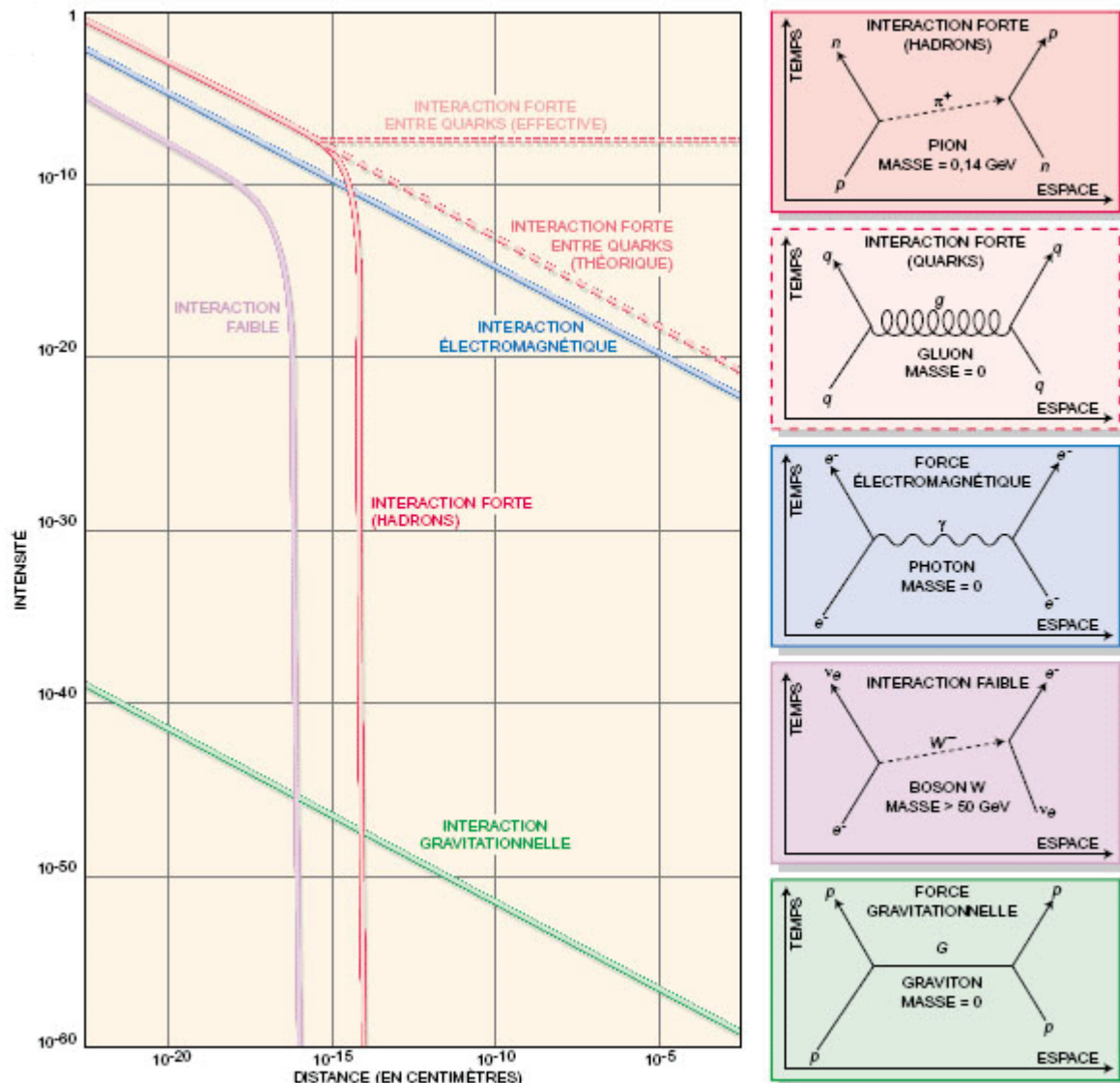


Figure: 46.1 - Diagrammes de Feynman et distance d'interaction des forces fondamentales

Il convient de préciser pour la culture générale que ces quatre forces sont décrites respectivement par quatre théories:

1. La relativité générale (englobe la mécanique classique) pour la gravitation
2. L'électrodynamique quantique (englobe l'électrodynamique) pour la force électromagnétique
3. La théorie électrofaible (qui englobe l'électrodynamique quantique) pour l'interaction faible
4. La chromodynamique quantique pour l'interaction forte

Les trois dernières étant regroupées dans le "**modèle standard**".

2. RÉSONANCE MAGNÉTIQUE DE SPIN

Nous avons hésité à mettre le traitement de ce sujet dans ce chapitre mais après réflexion, ce n'est pas de la physique nucléaire car les calculs ne s'appliquent pas qu'au noyau des atomes et cela n'est pas vraiment de la physique quantique relativiste pure car ne s'applique pas qu'à des particules élémentaires du type électron (comme le supposaient nos développements dans le chapitre de Physique Quantique Relativiste).

Lors de notre démonstration de l'équation de Pauli dans le chapitre de Physique Quantique Relativiste à la sauce physicienne.... nous avons vu que pour une particule de spin $\frac{1}{2}$ (qui pourrait être un noyau nucléaire de spin $\frac{1}{2}$), nous avons:

$$i\hbar \frac{\partial}{\partial t} \phi_{a0} \cong \left[\frac{\|\vec{p}\|^2}{2m_0} - \frac{q\vec{B} \circ \vec{L}}{2m_0} - \frac{\hbar q}{2m_0} \vec{\sigma} \circ \vec{B} + qU \right] \phi_{a0} \quad (46.8)$$

où il y a donc un terme propre au spin dans l'hamiltonien à savoir:

$$H_S = -\frac{\hbar q}{2m} \vec{\sigma} \circ \vec{B} = -\mu \vec{\sigma} \circ \vec{B} = -\frac{2}{2} \mu \vec{\sigma} \circ \vec{B} = -\frac{g\mu}{2} \vec{\sigma} \circ \vec{B} \quad (46.9)$$

en posant:

$$g = 2 \quad (46.10)$$

comme étant pour rappel "facteur de Landé" ou "facteur gyromagnétique" (cf. chapitre de Physique Quantique Relativiste).

Maintenant plongeons la particule appelée "sonde" dans un champ magnétique du type oscillant stable sur le plan horizontal et constant sur le plan vertical du type:

$$\vec{B} = \begin{pmatrix} B_A \cos(\omega t) \\ -B_A \sin(\omega t) \\ B_0 \end{pmatrix} \quad (46.11)$$

où dans la pratique le champ B_0 va de 0.1 à 17 Teslas.

Il vient alors:

$$\begin{aligned}
H_S &= -\frac{g\mu}{2} \vec{\sigma} \circ \vec{B} = -\frac{g\mu}{2} \begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \end{pmatrix} \circ \begin{pmatrix} B_A \cos(\omega t) \\ -B_A \sin(\omega t) \\ B_0 \end{pmatrix} \\
&= -\frac{g\mu}{2} (\sigma_1 B_A \cos(\omega t) - \sigma_2 B_A \sin(\omega t) + \sigma_3 B_0) \\
&= -\frac{g\mu}{2} \left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} B_A \cos(\omega t) - \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} B_A \sin(\omega t) + \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} B_0 \right) \\
&= -\frac{g\mu}{2} \left(\begin{bmatrix} 0 & B_A \cos(\omega t) \\ B_A \cos(\omega t) & 0 \end{bmatrix} - \begin{bmatrix} 0 & -i B_A \sin(\omega t) \\ i B_A \sin(\omega t) & 0 \end{bmatrix} + \begin{bmatrix} B_0 & 0 \\ 0 & -B_0 \end{bmatrix} \right) \\
&= -\frac{g\mu}{2} \begin{bmatrix} B_0 & B_A \cos(\omega t) + i B_A \sin(\omega t) \\ B_A \cos(\omega t) - i B_A \sin(\omega t) & -B_0 \end{bmatrix} \\
&= -\frac{g\mu}{2} \begin{bmatrix} B_0 & B_A e^{i\omega t} \\ B_A e^{-i\omega t} & -B_0 \end{bmatrix}
\end{aligned} \tag{46.12}$$

Nous avons alors en nous concentrant uniquement sur cet hamiltonien:

$$i\hbar \frac{\partial}{\partial t} \phi_{a0} = -\frac{g\mu}{2} \begin{bmatrix} B_0 & B_A e^{i\omega t} \\ B_A e^{-i\omega t} & -B_0 \end{bmatrix} \phi_{a0} \tag{46.13}$$

Et comme ϕ_{a0} est la double composante d'un spineur, notons-la explicitement:

$$i\hbar \frac{\partial}{\partial t} \begin{pmatrix} \alpha_+(t) \\ \alpha_-(t) \end{pmatrix} = -\frac{g\mu}{2} \begin{bmatrix} B_0 & B_A e^{i\omega t} \\ B_A e^{-i\omega t} & -B_0 \end{bmatrix} \begin{pmatrix} \alpha_+(t) \\ \alpha_-(t) \end{pmatrix} \tag{46.14}$$

Soit après réarrangement:

$$\frac{\partial}{\partial t} \begin{pmatrix} \alpha_+(t) \\ \alpha_-(t) \end{pmatrix} = \frac{i}{2} \begin{bmatrix} \frac{g\mu}{\hbar} B_0 & \frac{g\mu}{\hbar} B_A e^{i\omega t} \\ \frac{g\mu}{\hbar} B_A e^{-i\omega t} & -\frac{g\mu}{\hbar} B_0 \end{bmatrix} \begin{pmatrix} \alpha_+(t) \\ \alpha_-(t) \end{pmatrix} \tag{46.15}$$

Et rappelons que nous avons obtenu démontré tout à la fin du chapitre de Physique Quantique Relativiste que:

$$\omega = -g\gamma B = -\frac{g\mu_B B}{\hbar} \tag{46.16}$$

Il nous vient alors naturellement de poser que:

$$\begin{aligned}\omega_0 &= \frac{g\mu}{\hbar} B_0 \\ \omega_A &= \frac{g\mu}{\hbar} B_A\end{aligned}\quad (46.17)$$

Cela nous donne:

$$2 \begin{pmatrix} \dot{\alpha}_+(t) \\ \dot{\alpha}_-(t) \end{pmatrix} = i \begin{bmatrix} \omega_0 & \omega_A e^{i\omega t} \\ \omega_A e^{-i\omega t} & -\omega_0 \end{bmatrix} \begin{pmatrix} \alpha_+(t) \\ \alpha_-(t) \end{pmatrix} \quad (46.18)$$

Soit un système de deux équations différentielles:

$$\begin{aligned}2\dot{\alpha}_+(t) &= i \left(\omega_0 \alpha_+(t) + \omega_A e^{i\omega t} \alpha_-(t) \right) \\ 2\dot{\alpha}_-(t) &= i \left(\omega_A e^{-i\omega t} \alpha_+(t) - \omega_0 \alpha_-(t) \right)\end{aligned}\quad (46.19)$$

Si nous assumons que le spineur est orienté vers le haut initialement (conditions initiales) alors:

$$\begin{pmatrix} \alpha_+(0) \\ \alpha_-(0) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (46.20)$$

Maintenant pour résoudre ce système d'équations différentielles nous allons utiliser le travail laborieux de tâtonnement déjà effectué par nos illustres prédécesseurs en posant que la réponse est probablement du type:

$$\begin{aligned}\alpha_+(t) &= A_+ e^{i\omega_+ t} = |+\rangle \\ \alpha_-(t) &= A_- e^{i\omega_- t} = |-\rangle\end{aligned}\quad (46.21)$$

soit: des fonctions d'onde!

Alors en injectant ces solutions dans le système d'équations différentielles il vient:

$$\begin{aligned}2 \frac{d}{dt} A_+ e^{i\omega_+ t} &= i \left(\omega_0 A_+ e^{i\omega_+ t} + \omega_A e^{i\omega t} A_- e^{i\omega_- t} \right) \\ 2 \frac{d}{dt} A_- e^{i\omega_- t} &= i \left(\omega_A e^{-i\omega t} A_+ e^{i\omega_+ t} - \omega_0 A_- e^{i\omega_- t} \right)\end{aligned}\quad (46.22)$$

Soit:

$$\begin{aligned}2i\omega_+ A_+ e^{i\omega_+ t} &= i \left(\omega_0 A_+ e^{i\omega_+ t} + \omega_A A_- e^{i\omega t + i\omega_- t} \right) \\ 2i\omega_- A_- e^{i\omega_- t} &= i \left(\omega_A e^{-i\omega t + i\omega_+ t} A_+ - \omega_0 A_- e^{i\omega_- t} \right)\end{aligned}\quad (46.23)$$

et après simplification par l'exponentielle et l'imaginaire pur:

$$\begin{aligned} 2\omega_+ A_+ &= \omega_0 A_+ + \omega_A A_- e^{i\omega t + i\omega_+ t - i\omega_- t} \\ 2\omega_- A_- &= \omega_A A_+ e^{-i\omega t + i\omega_+ t - i\omega_- t} - \omega_0 A_- \end{aligned} \quad (46.24)$$

Soit:

$$\begin{aligned} 2\omega_+ A_+ &= \omega_0 A_+ + \omega_A A_- e^{i(\omega + \omega_+ - \omega_-)t} \\ 2\omega_- A_- &= \omega_A A_+ e^{-i(\omega - \omega_+ + \omega_-)t} - \omega_0 A_- \end{aligned} \quad (46.25)$$

Nous pouvons nous débarrasser de la dépendance en temps ci-dessus en posant:

$$\omega_+ = \omega_- + \omega \quad (46.26)$$

Notre système se réduit alors à:

$$\begin{aligned} 2\omega_+ A_+ &= \omega_0 A_+ + \omega_A A_- \\ 2\omega_- A_- &= \omega_A A_+ - \omega_0 A_- \end{aligned} \quad (46.27)$$

Soit en réarrangeant:

$$\begin{aligned} A_+ (2\omega_+ - \omega_0) - \omega_A A_- &= 0 \\ -\omega_A A_+ + A_- (2\omega_- + \omega_0) &= 0 \end{aligned} \quad (46.28)$$

ce qui peut s'écrire sous forme matricielle:

$$\begin{bmatrix} 2\omega_+ - \omega_0 & -\omega_A \\ -\omega_A & 2\omega_- + \omega_0 \end{bmatrix} \begin{bmatrix} A_+ \\ A_- \end{bmatrix} = 0 \quad (46.29)$$

Pour avoir une solution consistante (solutions non toutes nulles), le déterminant de la matrice doit être nul pour que celle-ci soit inversible (cf. chapitre d'Algèbre Linéaire). Or le déterminant de cette matrice est:

$$\begin{aligned} \det \begin{bmatrix} 2\omega_+ - \omega_0 & -\omega_A \\ -\omega_A & 2\omega_- + \omega_0 \end{bmatrix} &= (2\omega_+ - \omega_0)(2\omega_- + \omega_0) - (-\omega_A)(-\omega_A) \\ &= (2\omega_+ - \omega_0)(2\omega_- + \omega_0) - \omega_A^2 = 4\omega_+ \omega_- + 2\omega_+ \omega_0 - 2\omega_- \omega_0 - \omega_0^2 - \omega_A^2 \\ &= 4(\omega_- + \omega)\omega_- + 2(\omega_- + \omega)\omega_0 - 2\omega_- \omega_0 - \omega_0^2 - \omega_A^2 \\ &= 4\omega_-^2 + 4\omega\omega_- + 2\omega_- \omega_0 + 2\omega\omega_0 - 2\omega_- \omega_0 - \omega_0^2 - \omega_A^2 \\ &= 4\omega_-^2 + 4\omega\omega_- + 2\omega\omega_0 - \omega_0^2 - \omega_A^2 \end{aligned} \quad (46.30)$$

Or les racines de ce polynôme du deuxième degré en ω_- sont (cf. chapitre de Calcul Algébrique):

$$\begin{aligned}
\omega_{-,1,2} &= \frac{-4\omega \pm \sqrt{(4\omega)^2 - 4 \cdot 4 \cdot (2\omega\omega_0 - \omega_0^2 - \omega_A^2)}}{2 \cdot 4} \\
&= \frac{-\omega \pm \sqrt{\omega^2 - (2\omega\omega_0 - \omega_0^2 - \omega_A^2)}}{2} = \frac{-\omega \pm \sqrt{\omega^2 - (2\omega\omega_0 - \omega_0^2) + \omega_A^2}}{2} \\
&= \frac{-\omega \pm \sqrt{\omega^2 - 2\omega\omega_0 + \omega_0^2 + \omega_A^2}}{2} = \frac{-\omega \pm \sqrt{(\omega - \omega_0)^2 + \omega_A^2}}{2} = -\frac{\omega}{2} \pm \frac{\sqrt{(\omega - \omega_0)^2 + \omega_A^2}}{2} \\
&= -\frac{\omega}{2} \pm \Delta\omega
\end{aligned} \tag{46.31}$$

Nous en déduisons donc aussi:

$$\omega_+ = \omega_- + \omega \Rightarrow \omega_{+,1,2} = \omega_{-,1,2} + \omega = -\frac{\omega}{2} \pm \Delta\omega + \omega = +\frac{\omega}{2} \pm \Delta\omega \tag{46.32}$$

Nous avons donc pour résumer:

$$\begin{aligned}
\omega_{-,1,2} &= -\frac{\omega}{2} \pm \Delta\omega \\
\omega_{+,1,2} &= +\frac{\omega}{2} \pm \Delta\omega
\end{aligned} \tag{46.33}$$

Résultats à injecter dans:

$$\begin{aligned}
\alpha_+(t) &= A_+ e^{i\omega_+ t} \\
\alpha_-(t) &= A_- e^{i\omega_- t}
\end{aligned} \tag{46.34}$$

Donc comme chaque pulsation à deux solutions, la solution générale sera la somme des solutions particulières (cf. chapitre de Calcul Différentiel Et Intégral). Concentrons-nous sur la deuxième relation. Nous avons alors:

$$\alpha_-(t) = A_- e^{-i\frac{\omega}{2}t} e^{+i\Delta\omega t} + B_- e^{-i\frac{\omega}{2}t} e^{-i\Delta\omega t} \tag{46.35}$$

Or, rappelons la condition initiale que nous nous étions imposés pour la deuxième composante du spineur (liée au fait que le spineur est orienté vers le haut):

$$\alpha_-(0) = 0 \tag{46.36}$$

Nous avons alors:

$$A_- = -B_- := A \tag{46.37}$$

Donc:

$$\alpha_{-}(t) = A e^{-i\frac{\omega}{2}t} e^{+i\Delta\omega t} - A e^{-i\frac{\omega}{2}t} e^{-i\Delta\omega t} \quad (46.38)$$

avec:

$$\dot{\alpha}_{-}(t) = iA \left(-\frac{\omega}{2} + \Delta\omega \right) e^{-i\frac{\omega}{2}t} e^{+i\Delta\omega t} - iA \left(-\frac{\omega}{2} - \Delta\omega \right) e^{-i\frac{\omega}{2}t} e^{-i\Delta\omega t} \quad (46.39)$$

et:

$$\dot{\alpha}_{-}(0) = iA \left(-\frac{\omega}{2} + \Delta\omega \right) - iA \left(-\frac{\omega}{2} - \Delta\omega \right) = 2iA\Delta\omega \quad (46.40)$$

Maintenant, pour déterminer A , nous allons de façon astucieuse utiliser la différentielle de départ obtenue plus haut:

$$2\dot{\alpha}_{-}(t) = i \left(\omega_A e^{-i\omega t} \alpha_{+}(t) - \omega_0 \alpha_{-}(t) \right) \quad (46.41)$$

sans oublier les conditions initiales choisies, nous avons:

$$2\dot{\alpha}_{-}(0) = i \left(\omega_A e^{-i\omega_0} \alpha_{+}(0) - \omega_0 \alpha_{-}(0) \right) = i \left(\omega_A \cdot 1 - \omega_0 \cdot 0 \right) = i\omega_A \quad (46.42)$$

Soit:

$$\dot{\alpha}_{-}(0) = \frac{i\omega_A}{2} \quad (46.43)$$

Nous devons donc avoir l'égalité entre les deux expressions:

$$\begin{aligned} \dot{\alpha}_{-}(0) &= \frac{i\omega_A}{2} \\ \dot{\alpha}_{-}(0) &= 2iA\Delta\omega \end{aligned} \quad (46.44)$$

et cela ne peut être satisfait que si:

$$A = \frac{\omega_A}{4\Delta\omega} \quad (46.45)$$

Nous avons alors:

$$\begin{aligned}\alpha_{-}(t) &= \frac{\omega_A}{4\Delta\omega} e^{-i\frac{\omega}{2}t} \left(e^{+i\Delta\omega t} - e^{-i\Delta\omega t} \right) = \frac{\omega_A}{4\Delta\omega} e^{-i\frac{\omega}{2}t} 2i \sin(i\Delta\omega t) \\ &= \frac{\omega_A}{2\Delta\omega} i \sin(i\Delta\omega t) e^{-i\frac{\omega}{2}t}\end{aligned}\quad (46.46)$$

Soit explicitement:

$$\alpha_{-}(t) = \frac{\omega_A}{\sqrt{(\omega - \omega_0)^2 + \omega_A^2}} i \sin(i\Delta\omega t) e^{-i\frac{\omega}{2}t} \quad (46.47)$$

Puisque la deuxième composante du spineur représente l'angle et que ce que nous avons ci-dessus représente la position angulaire dans le temps, le terme:

$$\boxed{\frac{\omega_A}{\sqrt{(\omega - \omega_0)^2 + \omega_A^2}}} \quad (46.48)$$

peut être vu comme la valeur maximale de l'amplitude de l'angle (rotation de Larmor du spin de l'électron autour du champ magnétique). Cette amplitude maximale a elle-même un maximum si le dénominateur est le plus petit possible et donc:

$$\omega = \omega_0 = \frac{g\mu}{\hbar} B_0 \quad (46.49)$$

Nous disons alors qu'il y a "**résonance de spin**".

Donc le champ magnétique peut faire basculer l'état d'énergie de chaque spin si son amplitude d'oscillation est maximale. Comme nous l'avons démontré juste plus haut, pour une particule isolée, la variation d'énergie d'un état à l'autre est:

$$E_{\pm} = \pm \frac{1}{2} \hbar \omega \quad (46.50)$$

et donc lors d'un basculement de spin génère l'émission d'un rayonnement que l'on appelle "**Free Induction Decay**". Le signal recueilli dépend de plusieurs paramètres qui caractérisent la particule ou le noyau nucléaire. Bref dans la pratique c'est 99% d'ingénierie et 1% de théorie!

Lorsque nous appliquons ces résultats théoriques à un électron non apparié, c'est-à-dire pour la "**résonance magnétique électronique**", nous avons les données expérimentales suivantes pour un champ constant $B_0 \cong 0.3 [T]$:

$$f \cong 10 [GHz], \lambda = 3 [cm], E_{\gamma} = \Delta E \cong 4 \cdot 10^{-4} [eV]$$

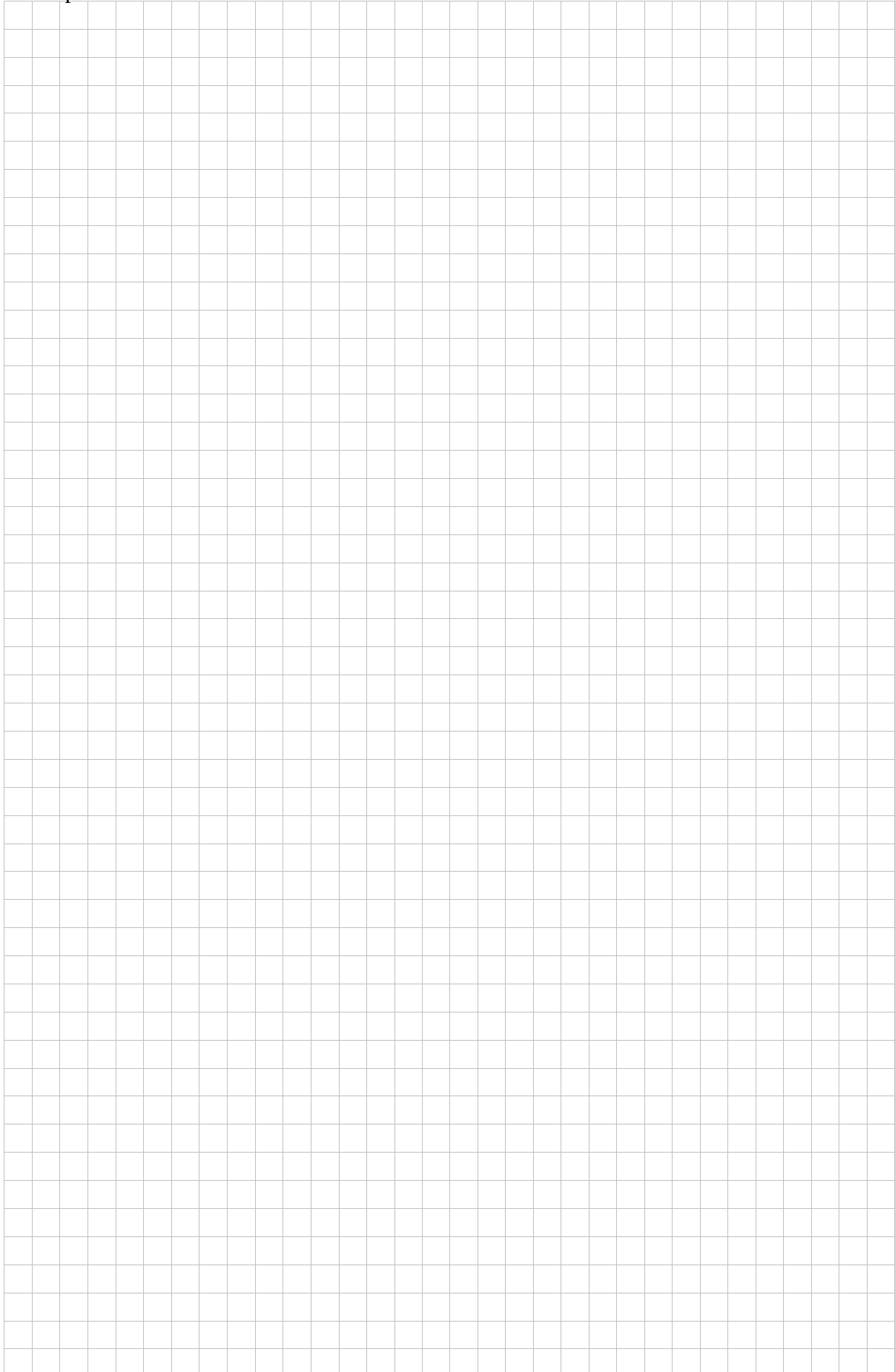
et donc la technologie de résonance magnétique électronique est basée sur la détection de micro-ondes.

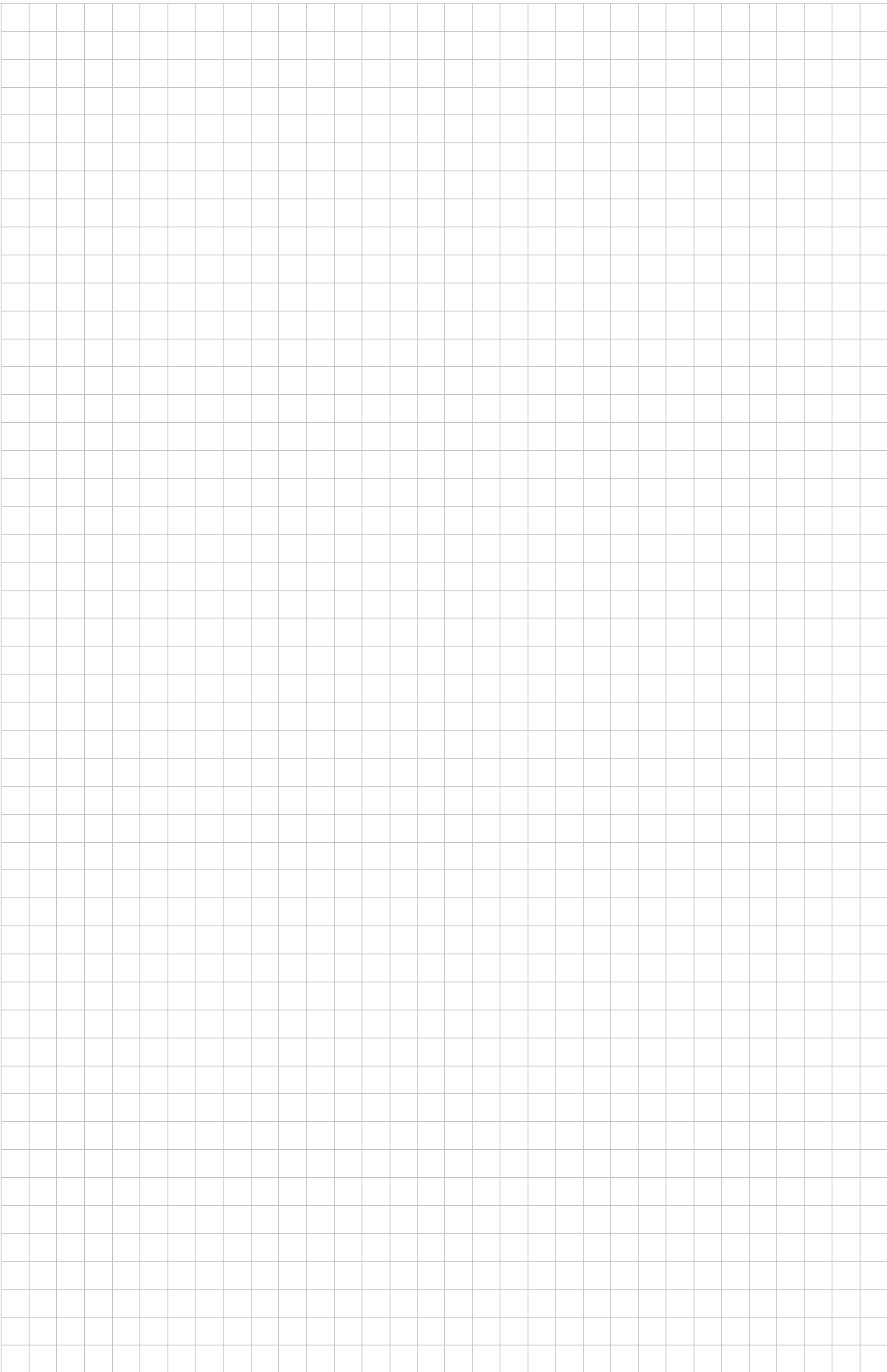
Lorsque nous appliquons ces résultats théoriques à un noyau nucléaire d'un seul proton, c'est-à-dire pour la "résonance magnétique nucléaire", nous avons les données expérimentales suivantes pour un champ constant $B_0 \cong 1 [T]$:

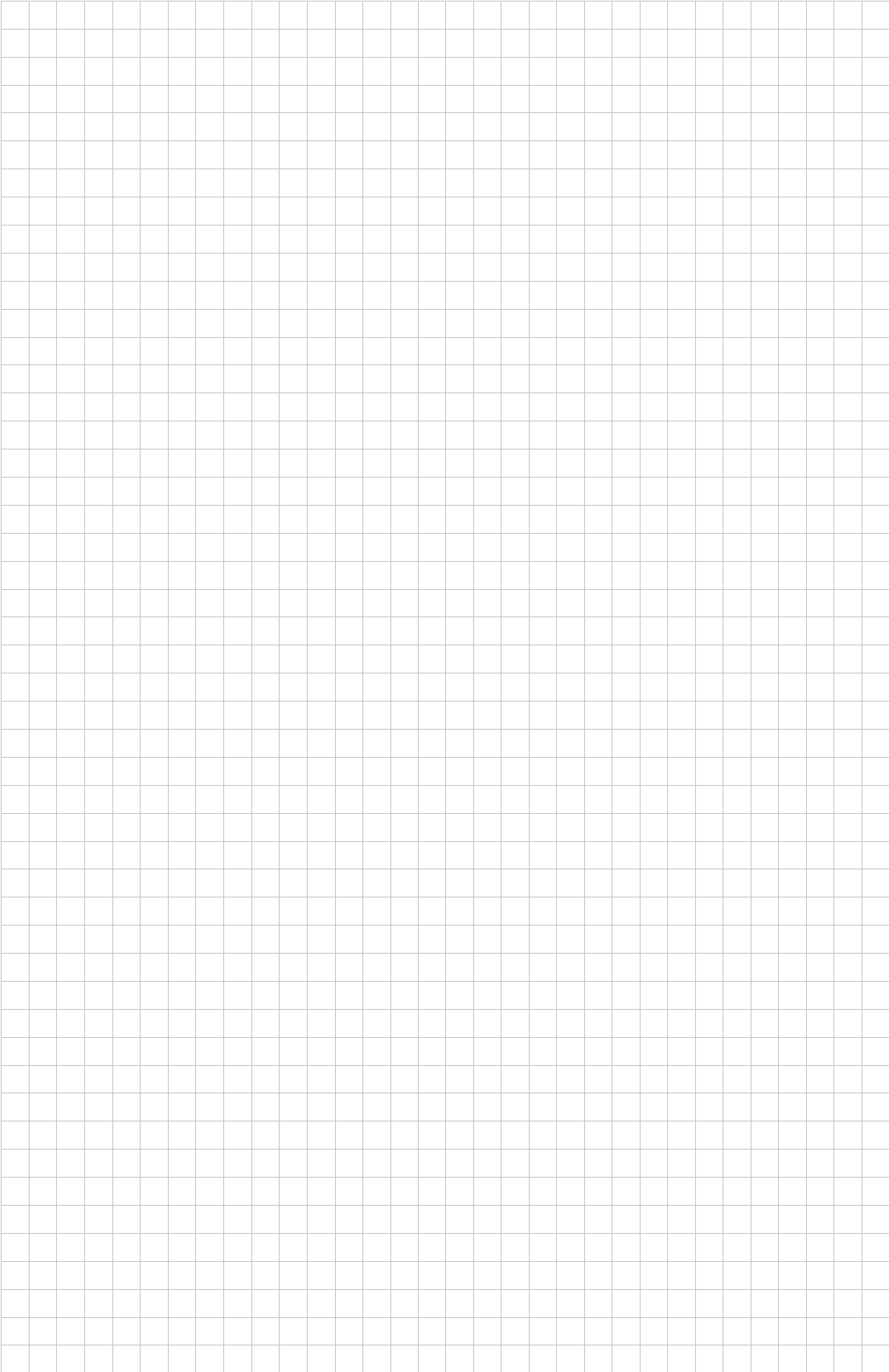
$$f \cong 40 [MHz], \lambda = 7 [m], E_\gamma = \Delta E \cong 2 \cdot 10^{-7} [eV]$$

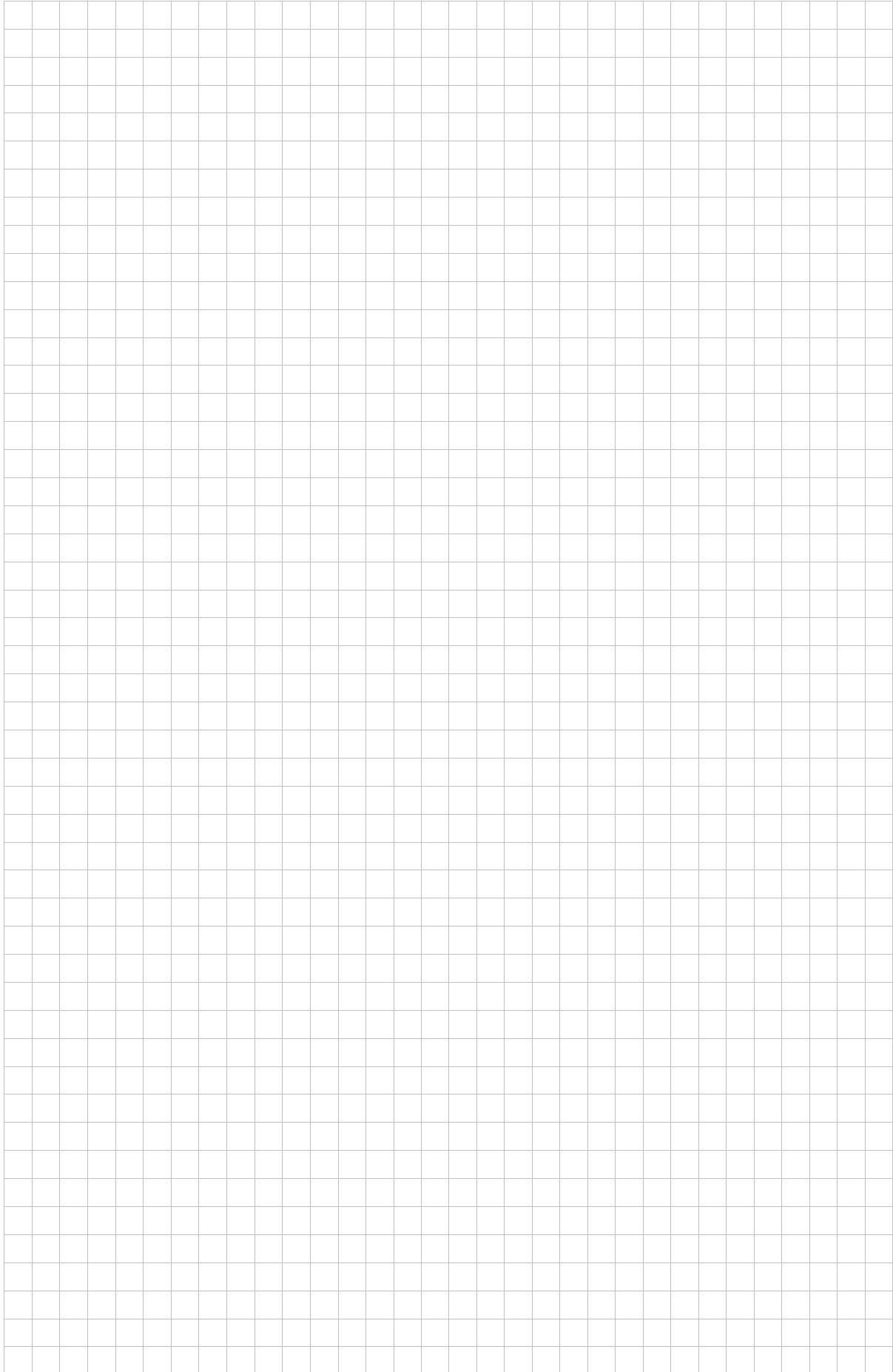
et donc la technologie de résonance magnétique nucléaire est basée sur la détection des ondes radio.

Notes personnelles:











Cosmologie

La mécanique céleste est la conséquence de la loi d'attraction universelle de Newton et du principe fondamental de la mécanique (cf. chapitre de Mécanique Classique). Elle a pour principal objectif la description du mouvement d'objets astronomiques tels que les étoiles et planètes à l'aide des théories physiques et mathématiques.

Nous allons dans ce chapitre aborder le sujet, comme toujours sur ce site, de la manière la plus élémentaire possible (à ce jour les sujets traités sur cette page ne dépassent pas techniquement le niveau de ce qui se faisait au milieu du 18ème siècle dans le domaine de l'astronomie).

D'abord, nous nous échaufferons avec une loi sympathique sur le vivant dans l'Univers... (l'équation de Drake). Une fois cet exercice de style accompli, nous commencerons à "énumérer" les lois de Kepler (en faisant souvent référence au chapitre de Mécanique Classique) pour ensuite étudier en détails les propriétés des orbitales Képlériennes à l'aide de la mécanique et ensuite à l'aide de la relativité restreinte, ce qui nous amènera à constater une précession théorique des orbitales concernées. Ensuite, nous nous amuserons à modéliser approximativement la variation de la durée de la journée (et de la nuit) sur la Terre en fonction du mois et de la latitude. Enfin, pour terminer en beauté, nous nous lancerons dans le calcul détaillé des cinq points de Lagrange!

ÉQUATION DE DRAKE

Cette équation a été inventée (...) par F. Drake dans les années 1960 dans l'intention d'estimer le nombre de civilisations extraterrestres dans notre galaxie avec lesquelles nous pourrions entrer en contact. Le principal objet de cette équation pour les scientifiques est de déterminer ses facteurs, afin de connaître le nombre probable et (très) estimé de civilisations extraterrestres.

Cette équation empirique (qui reste un amusement... et dont le principe peut être appliqué à pas mal de domaines différents de la physique et de la vie...) s'écrit:

$$N = N^* \cdot f_p \cdot n_e \cdot f_i \cdot f_c \cdot f_L \quad (47.1)$$

Les termes de cette formule (car c'en est une!) se définissent ainsi:

- N^* représente le nombre d'étoiles dans une seule et unique galaxie
- f_p est la fraction d'étoiles qui auraient une planète en orbite (comprise entre 0 et 1)
- n_e est le nombre de planètes par étoile qui remplissent les conditions au développement de la vie
- f_i est la fraction de planètes dont la vie s'est effectivement développée (comprise entre 0 et 1)
- f_c est la fraction de celles où une vie intelligente s'est développée (comprise entre 0 et 1)
- f_L est la fraction de f_i qui a mis en oeuvre des moyens de communication radio (comprise entre 0 et 1)
- f_L est la fraction de temps pendant laquelle les civilisations f_i vivront (comprise entre 0 et 1)

Dans la pratique, il faut remarquer que l'équation consiste à essayer de déterminer une quantité inconnue à partir d'autres quantités qui sont tout aussi inconnues qu'elles... Mais c'est une équation sympa à sortir et à évaluer entre amis pour passer le temps...

Il n'existe donc pas de garantie que l'on soit davantage fixé après cette estimation qu'avant (argument nommé parfois dans la littérature "garbage in, garbage out"....).

La valeur résultante peut motiver que les développements qui vont suivre ne sont pas applicables qu'à un seul système solaire dans l'Univers... peut-être... (cela ferait beaucoup de vide gâché sinon...).

LOIS DE KEPLER

En astronomie, les lois de Kepler décrivent les propriétés principales du mouvement des planètes autour d'un astre principal, sans les expliquer (à l'époque!). Elles ont été découvertes par Johannes Kepler à partir des observations et mesures (en quantité phénoménale) de la position des planètes faites par Tycho Brahe, mesures qui étaient très précises pour l'époque.

Les deux premières lois de Kepler furent publiées en 1609 et la troisième en 1618. Les orbites elliptiques, telles qu'énoncées dans ses deux premières lois, permettent d'expliquer la complexité du mouvement apparent des planètes.

Peu après, Isaac Newton découvrit en 1687 la loi de l'attraction gravitationnelle (ou gravitation), déduisant de celle-ci, par le calcul, les 3 lois de Kepler.

Nous allons maintenant nous efforcer de présenter ces lois de la manière la plus pertinente possible:

PREMIÈRE LOI

La "**première loi de Kepler**", appelée parfois aussi "**loi de conicité**" ou encore "**loi des orbites**" s'énonce ainsi: Les orbites des planètes sont des coniques (ellipses) dont le Soleil occupe l'un des foyers.

Au fait, il convient de préciser que ce n'est pas vraiment une "loi" dans le sens propre du terme puisque plus loin vous en trouverez la démonstration telle que:

$$r = \frac{p}{1 + e \cos \theta} \quad (47.2)$$

Remarque: Le lecteur qui aura lu au préalable le chapitre de Géométrie Analytique ne sera pas étranger à cette relation...

DEUXIÈME LOI

La "**deuxième loi de Kepler**", appelée parfois aussi "**loi des aires**", nous dit que le segment qui joint une planète au Soleil balaie des aires égales en des temps égaux (vitesse aréolaire constante) tel que:

$$\frac{\Delta S}{\Delta t} = c^{te} \quad (47.3)$$

C'est une relation qui découle de la conservation du moment cinétique comme nous l'avons déjà démontré dans le chapitre de Mécanique Classique où nous avons obtenu:

$$\frac{d\vec{S}}{dt} = \frac{\vec{b}}{2m} = \frac{1}{2m}(\vec{r} \times m \cdot \vec{v}) = c^{te} \quad (47.4)$$

Donc à nouveau, son statut de "loi" est discutable dans le langage de la physique moderne!

Par ailleurs, rappelons que nous avons aussi obtenu comme résultat que le mouvement soit et reste plan sans aucune action extérieure!

Nous constatons par ailleurs que cette loi nous donne que la vitesse de la planète est variable. Elle est plus grande au périhélie qu'à l'aphélie:

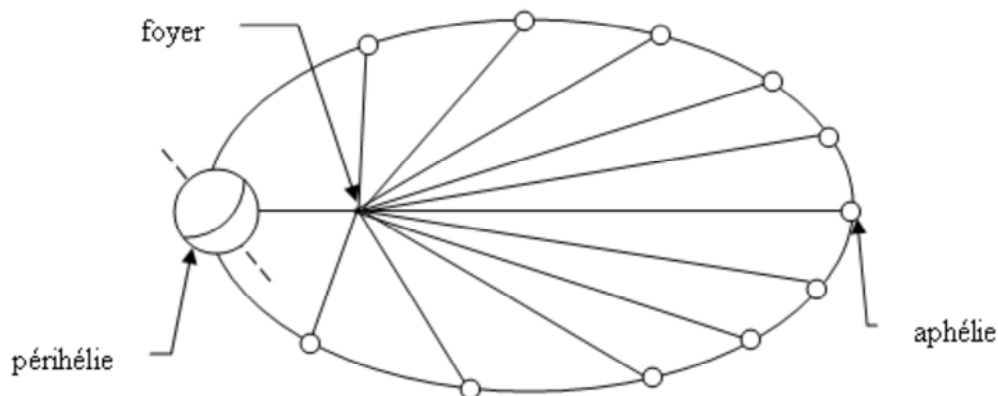


Figure: 47.1 - Représentation des surfaces balayées

Ceci se vérifie pour la Terre par exemple. En effet, cette dernière est plus proche du Soleil en hiver (pour l'hémisphère Nord) et elle a alors une vitesse sur trajectoire un peu plus élevée qu'en été; le temps de parcours est donc plus faible (l'hiver compte moins de jours que les autres saisons).

TROISIÈME LOI

La "**troisième loi de Kepler**", appelée parfois aussi "**loi des périodes**", s'énonce ainsi: Les carrés des périodes de révolution T sont proportionnels aux cubes des demi-grands axes D des orbites:

$$T^2 = c^{te} \cdot D^3 \quad (47.5)$$

À nouveau, nous verrons plus loin que le statut de "loi" n'est plus justifiable à notre époque puisqu'il est possible de démontrer que cette relation, dont l'expression sera détaillée, est réellement:

$$\boxed{\frac{R^3}{T^2} = \frac{Gm}{4\pi^2}} \quad (47.6)$$

Bien évidemment, Kepler n'a pas d'emblée publié ses trois lois dans cette provocante simplicité. Leur ordre actuel n'est d'ailleurs pas celui de leur énonciation... Elles sont en réalité à dénicher au milieu d'un foisonnement de spéculations physiques et de réflexions sur l'harmonie du monde.

LOI DE LA GRAVITATION DE NEWTON

Pour vérifier l'exactitude de son hypothèse, Newton (relativement longtemps après) retrouva les lois de Kepler à partir de la loi de la gravitation, donnant ainsi l'explication du mouvement général des planètes.

Newton considéra pour déterminer la loi de gravitation une planète théorique, gravitant autour du Soleil sur une orbite circulaire à vitesse constante v . Pendant une orbite complète, la planète parcourt une distance égale à la circonférence du cercle de rayon R , soit $2\pi R$, en un temps (sa période) égal à cette distance divisée par sa vitesse, soit:

$$T = \frac{2\pi R}{v} \quad (47.7)$$

Newton s'appuie ensuite sur la troisième loi de Kepler avec toujours l'hypothèse d'une orbite circulaire.

Nous avons donc:

$$T^2 = c^{te} \cdot R^3 \quad (47.8)$$

mais puisque:

$$T = \frac{2\pi R}{v} \text{ alors } T^2 = 4\pi^2 \frac{R^2}{v^2} \quad (47.9)$$

En comparant:

$$T^2 = c^{te} \cdot R^3 \text{ et } T^2 = 4\pi^2 \frac{R^2}{v^2} \quad (47.10)$$

et en posant maintenant que $4\pi^2$ divisé par la constante est une nouvelle constante (que nous noterons de la même manière que la première bien qu'elle ne lui soit pas égale) on obtient:

$$R^3 = \frac{4\pi^2}{c^{te}} \frac{R^2}{v^2} = c^{te} \frac{R^2}{v^2} \text{ d'où } R^2 = c^{te} \frac{R}{v^2} \quad (47.11)$$

Ensuite, si nous renversons les termes, cette expression devient (tout en notant que l'inverse de la constante d'origine est, elle aussi, une constante):

$$\frac{1}{R^2} = c^{te} \frac{v^2}{R} \quad (47.12)$$

Par un autre calcul, nous avons déjà établi dans le chapitre de Mécanique Classique l'expression de la force centrifuge:

$$F_{Cent} = m \frac{v^2}{R} \quad (47.13)$$

en rapprochant cette expression de l'expression précédente:

$$\frac{1}{c^2} \frac{1}{R^2} = \frac{v^2}{R} \quad (47.14)$$

nous obtenons:

$$F_{Cent} = m \frac{v^2}{R} = m \cdot \frac{1}{c^2} \frac{1}{R^2} = \frac{m \cdot c^2}{R^2} = F_G \quad (47.15)$$

Il existerait donc une force opposée à la force centrifuge qui maintient la cohésion orbitale et qui s'écrit:

$$F_G = \frac{m \cdot c^2}{R^2} \quad (47.16)$$

reste à déterminer la valeur de la constante!

Il est trivial que la masse centrale M du système orbital doit intervenir d'une façon ou d'une autre dans cette constante. Si la masse du corps secondaire intervient de façon proportionnelle dans la force centrifuge, l'envie est grande de faire de même avec la masse du corps central. Donc:

$$F_G = \frac{m \cdot M \cdot c^2}{R^2} \quad (47.17)$$

maintenant a priori il n'y aurait plus de paramètres à prendre en compte. La constante restante est là pour satisfaire à l'analyse dimensionnelle de telle façon que l'on ait des "Newtons" (nom donné à l'unité de force) des deux côtés de l'égalité. Les scientifiques ont déterminé avec grande précision cette "**constante gravitationnelle**" notée G qui a priori semble universelle et qui a comme valeur:

$$G = 6.6732 \cdot 10^{-11} [N \cdot m^2 \cdot Kg^{-2}] \quad (47.18)$$

Ce qui nous amène à écrire la "**loi de la gravitation de Newton**":

$$F_G = \frac{G \cdot m \cdot M}{R^2} \quad (47.19)$$

Évidemment il ne s'agit nullement d'une vraie démonstration car nous nous sommes basés sur les observations expérimentales de Kepler. Par contre, à partir de la relativité générale il est possible de la démontrer (sous certaines hypothèses...)!

Remarque: En égalisant force centrifuge et force gravitationnelle, il est assez facile d'obtenir une approximation de la vitesse de rotation des planètes sur leur orbite. Le lecteur qui fera le calcul verra que le chiffre tourne pour les planètes du système solaire autour d'une vitesse de l'ordre de 100'000 [km/h].

À partir de cette dernière relation, revenons brièvement sur notre troisième loi de Kepler et détaillons un peu pour montrer qu'elle est valable pour tout type d'orbite conique et afin de déterminer l'expression de sa constante.

Exprimée dans le repère de Frenet (cf. chapitre de Géométrie Différentielle), et décomposée en son accélération normale (centripète) et tangentielle, l'accélération par rapport à un référentiel

géocentrique (dans le cas d'un référentiel situé au centre de masse du système l'expression change un peu) s'écrit:

$$a_{tg} = 0, a_n = \frac{v^2}{R} \quad (47.20)$$

Des relations obtenues lors des développements précédents:

$$T^2 = c^{te} \cdot R^3 \text{ et } T^2 = 4\pi^2 \frac{R^3}{v^2} \quad (47.21)$$

la constante de la troisième loi de Kepler prend comme valeur (c'est une formulation utilisée parfois dans la pratique mais une étape non rigoureusement nécessaire dans le présent développement):

$$c^{te} = 4\pi^2 \frac{R}{v^2} \frac{1}{R^2} = 4\pi^2 \frac{1}{a_n} \frac{1}{R^2} \quad (47.22)$$

Or, puisque nous avons aussi:

$$a_n = g = \frac{v^2}{R} = \frac{Gm}{R^2} \Leftrightarrow v^2 = \frac{Gm}{R} \quad (47.23)$$

alors:

$$c^{te} = 4\pi^2 \frac{R}{\frac{Gm}{R}} \frac{1}{R^2} = 4\pi^2 \frac{R^2}{Gm} \frac{1}{R^2} = \frac{4\pi^2}{Gm} \quad (47.24)$$

d'où:

$$T^2 = \frac{4\pi^2}{Gm} R^3 \quad (47.25)$$

Finalement, la troisième loi de Kepler se retrouve alors fréquemment dans la littérature sous la forme suivante:

$$\boxed{\frac{R^3}{T^2} = \frac{Gm}{4\pi^2}} \quad (47.26)$$

Cet interlude effectué, revenons sur notre loi de la gravitation de Newton:

$$F_G = \frac{G \cdot m \cdot M}{R^2} \quad (47.27)$$

À partir de cette loi de la gravitation, nous pouvons retrouver toutes les lois de Kepler. D'ailleurs, nous l'avons déjà fait pour la deuxième et troisième loi de Kepler puisque ce sont ces dernières que nous avons utilisées pour obtenir cette relation (c'est cependant un peu le serpent qui se mange la queue...).

Sous forme vectorielle nous avons ainsi:

$$\vec{F} = m \cdot \frac{G \cdot M}{R^2} \cdot \frac{\vec{R}}{\|\vec{R}\|} \quad (47.28)$$

Identiquement au champ électrique (cf. chapitre d'Électrostatique), nous pouvons développer:

$$\vec{F} = m \cdot \sum_i \left(\frac{G \cdot m_i}{\|\vec{r} - \vec{r}_i\|^2} \cdot \frac{(\vec{r} - \vec{r}_i)}{\|\vec{r} - \vec{r}_i\|} \right) = m \cdot \vec{a} \quad (47.29)$$

Comme le champ électrique dérive d'un potentiel électrique, identiquement, le champ gravitationnel dérive lui aussi d'un potentiel gravitationnel. En effectuant le même développement qu'en électromagnétisme pour la première équation de Maxwell (cf. chapitre d'Électrodynamique), nous démontrons que:

$$\vec{a} = \vec{\nabla} \cdot \varphi \quad (47.30)$$

où φ est le "**potentiel gravitationnel**" qui varie en raison inverse de la distance relative des corps (ceci confirmant ce que nous avons démontré lors de notre étude du théorème de Noether dans le chapitre traitant des Principes) et vaut donc:

$$\varphi = \frac{GM}{r}$$

Remarque: Nous retrouverons souvent ce potentiel dans le chapitre de Relativité Générale. Il convient donc de s'en souvenir si possible.

Écriture qui implique bien évidemment la relation suivante:

$$\vec{F} = -m\vec{\nabla} \cdot \varphi \quad (47.31)$$

Remarque: Évidemment en l'absence de champ, nous avons $\varphi = c^{\text{te}}$ et donc $\vec{a}$ sera nul.

Comme en électromagnétisme à nouveau, nous démontrons comme nous l'avons fait pour la première équation de Maxwell que:

$$\Phi_a = \oint_{\Sigma} \vec{a} \cdot d\vec{S} = \oint_V (\vec{\nabla} \cdot \vec{a}) dV = 4\pi Gm \Leftrightarrow \vec{\nabla} \cdot \vec{a} = 4\pi G\rho \quad (47.32)$$

Si nous exprimons cette équation en fonction d'un potentiel gravitationnel φ (noté aussi souvent par la lettre U comme en Électrostatique...), nous obtenons:

$$\vec{\nabla} \cdot (\vec{\nabla} \cdot \varphi) = 4\pi G\rho \quad (47.33)$$

ce que l'on note de façon plus esthétique avec le laplacien scalaire (cf. chapitre de Calcul Vectoriel):

$$\Delta \cdot \varphi = 4\pi G\rho \quad (47.34)$$

qui n'est autre que "**l'équation de Newton-Poisson**" que nous retrouverons aussi lors de notre étude de la relativité générale (elle y a une place importante pour des raisons de validation de la théorie d'Einstein)!

Cette équation signifie que la théorie Newtonienne de la gravitation se résume à dire que le champ gravitationnel est décrit par un seul potentiel φ engendré par la densité volumique de masse et déterminant l'accélération d'une particule d'épreuve plongée dans le champ extérieur φ .

Amusons-nous maintenant un peu avec l'équation de la gravitation de Newton pour obtenir quelques résultats intéressants et curieux:

Soit un objet situé à l'extérieur de la Terre et r la distance de cet objet au centre de la Terre, nous avons:

$$m \cdot g(r) = G \cdot \frac{M \cdot m}{r^2} \quad (47.35)$$

il vient:

$$g(r) = G \cdot \frac{M}{r^2} \quad (47.36)$$

Si l'objet est situé à la surface de la Terre de rayon R , nous avons ($r = R$):

$$g_0 = G \cdot \frac{M}{R^2} \quad (47.37)$$

Des deux dernières relations il vient donc:

$$g(r) = g_0 \left(\frac{R}{r} \right)^2 \quad (47.38)$$

En surface nous avons donc bien (on s'y attendait...):

$$g_0 = G \cdot \frac{M}{R^2} \quad (47.39)$$

Maintenant, si l'objet est situé à l'intérieur de la Terre en notant la distance par rapport au centre par la lettre r et la masse centrale par M' , nous avons:

$$m \cdot g(r) = G \cdot \frac{M' \cdot m}{r^2} \quad (47.40)$$

Introduisons la masse volumique ρ que nous supposons égale partout.

$$\begin{aligned} M &= \frac{4}{3} \pi R^3 \cdot \rho \\ M' &= \frac{4}{3} \pi r^3 \cdot \rho \end{aligned} \quad (47.41)$$

En combinant ces quatre dernières relations, nous obtenons:

$$g(r) = g_0 \frac{r}{R} \quad (47.42)$$

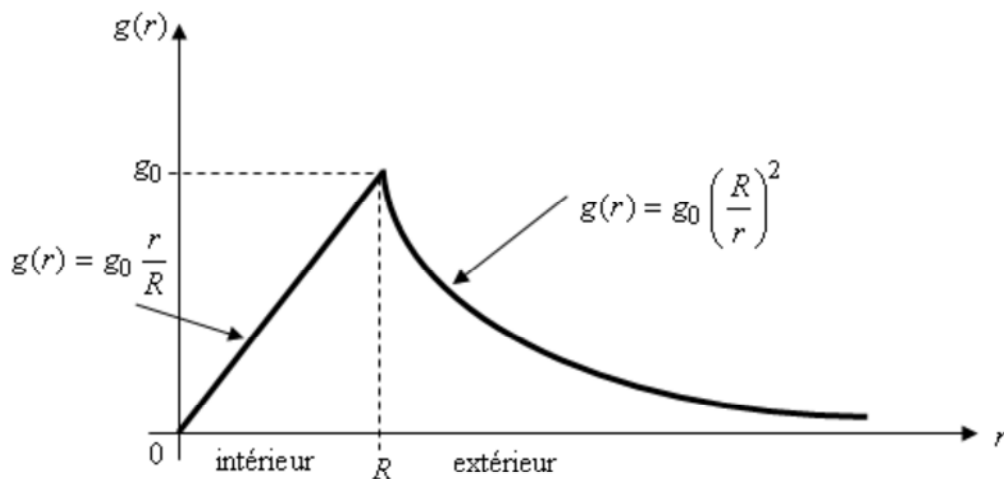


Figure: 47.2 - Profil intérieur/extérieur de l'accélération gravitationnelle

Pour de nombreuses personnes ce résultat est assez contre intuitif (faites un petit sondage dans votre entourage, vous verrez).

SPHÉRISATION DES CORPS CÉLESTES

À l'aide de la loi de Newton, nous pouvons répondre à pas mal de questions pertinentes de manière approximative et nous donnant des résultats tout à fait probants.

Un premier exemple est de se demander à quelle échelle il y a une transition du domaine des formes (les astéroïdes, lunes de Mars, comètes, etc.) au domaine des sphères (planètes et grandes lunes)? Pourquoi les satellites de Mars, Phobos et Deimos, ont une forme patatoïde tandis que notre lune est à peu près sphérique. Nous allons voir que ceci est dû à la masse qui est plus importante dans le cas de notre lune. Effectivement, à partir d'une certaine masse, les formes géométriques quelconques ne sont plus possibles.

Pour aborder cette étude, nous allons d'abord estimer la hauteur maximale d'une montagne sur une planète. Le Mont Everest a une altitude de 8.8 [km] tandis que le Mont Olympus sur Mars est de 27 [km]. Pourquoi de telles montagnes ne peuvent-elles exister sur Terre?

Pour prendre une approche simpliste, nous allons supposer qu'une montagne doit être en équilibre hydrostatique. Nous connaissons expérimentalement la pression limite type dans un réseau cristallin de roches au-delà de laquelle les roches commencent à "couler": $P_{\text{lim}} \cong 3 \cdot 10^8$ [Pa].

Nous connaissons de par notre étude la mécanique des milieux continus (cf. chapitre de Mécanique Des Milieux Continus) que la pression à la base d'une montagne de hauteur h sera donnée dans l'approximation hydrostatique par:

$$P_o = g \rho h \quad (47.43)$$

Pour que la montagne soit stable, il faut donc que:

$$P_o < P_{\text{lim}} \text{ et donc } h < h_0 \quad (47.44)$$

Ainsi:

$$h_0 = \frac{P_{\text{lim}}}{\rho g} = \frac{P_{\text{lim}} r^2}{\rho m G} \quad (47.45)$$

En supposant une densité moyenne de $\rho = 3'000 \left[\text{kg} \cdot \text{m}^{-3} \right]$ (croûte continentale de la Terre) nous obtenons:

- Terre: $h_0 \cong 10 \left[\text{km} \right]$

- Mars: $h_0 \cong 27 \left[\text{km} \right]$

Ce qui est remarquable comme résultat approximatif...

Pour estimer la taille minimale r_m d'un astre, à partir de laquelle la forme sphérique devient prédominante par rapport aux déformations de la surface (c'est-à-dire où la gravitation a pris le dessus sur les forces interatomiques), nous allons exiger que la taille r_m soit supérieure à la hauteur maximale d'une montagne h_0 . Nous supposons aussi que la densité ρ reste constante à travers l'astre. En reprenant la relation:

$$h_0 = \frac{P_{\text{lim}} r^2}{\rho m G} \quad (47.46)$$

nous avons:

$$h_0 = \frac{P_{\text{lim}} r^2}{\rho G} \left(\rho \frac{4}{3} \pi r^3 \right)^{-1} \quad (47.47)$$

d'où:

$$h_0 r = \frac{3 P_{\text{lim}}}{G 4 \pi \rho^2} \quad (47.48)$$

La limite r_m peut ensuite être estimée en fixant $r = r_m = h_0$ ainsi:

$$\boxed{r_m^2 = \frac{3 P_{\text{lim}}}{G 4 \pi \rho^2}} \quad (47.49)$$

bien évidemment pour $r \gg r_m$ nous serons encore plus proche de la forme sphérique.

APLATISSEMENT DES CORPS CÉLESTES

À cause de la symétrie du potentiel gravitationnel une étoile ou une planète devrait avoir une forme parfaitement sphérique à partir d'une certaine taille comme nous venons de le voir. Or, il n'en n'est pas ainsi.

À cause de la rotation propre de l'astre, un terme centrifuge vient modifier le potentiel. Ce terme dépend de la latitude ce qui explique la forme ellipsoïdale.

Rappelons que:

$$a = G \frac{M}{R^2} \quad (47.50)$$

où R est le rayon équatorial de l'astre à laquelle vient s'ajouter l'accélération centrifuge à une latitude donnée de rayon r :

$$a_c = \omega^2 r \quad (47.51)$$

Ainsi, l'accélération totale:

$$a_{tot} = G \frac{M}{R^2} + \omega^2 r \quad (47.52)$$

explique simplement que la Terre est aplatie aux pôles (ou selon le point de vue: étirée à l'équateur...) et que plus une planète tourne vite, plus elle sera aplatie aux pôles.

Sur Terre, le rayon équatorial est de 6379 [km] tandis que le rayon polaire est de 6357 [km]. La différence est de 22 [km]. "L'aplatissement" d'une planète peut être exprimé comme:

$$A = \frac{a - c}{a} \quad (47.53)$$

soit la différence entre rayon équatorial et le rayon polaire divisée par le rayon équatorial.

Bien qu'un ellipsoïde de révolution soit la meilleure description pour la forme d'une planète:



Figure: 47.3 - Image de la Terre avec son atmosphère

il y a des imperfections entre le modèle et la réalité pour certains corps du système solaire (en particulier les planètes telluriques, les satellites, et les petits corps rocheux). Le géopotential d'une planète réelle peut avoir une forme nettement plus compliquée à cause des influences des inhomogénéités visibles à la surface comme l'atteste cette image satellite de la Terre omettant les parties liquides (les déformations ont été un peu exagérées sur l'image ci-dessous):

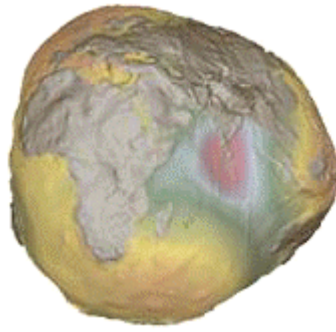


Figure: 47.4 - Image de la Terre sans son atmosphère

Les géodésistes tiennent compte de ces inhomogénéités. Ils mesurent et décrivent la forme des planètes qu'ils appellent "**géoides**".

STABILITÉ DES ATMOSPHÈRES

En comparant les vitesses de libération et les vitesses de divers gaz, nous pouvons expliquer la stabilité de certaines atmosphères et l'inexistence d'autres. Nous avons démontré dans le chapitre de Mécanique Classique que la vitesse de libération d'un astre sphérique était donnée par la relation suivante (sur laquelle nous reviendrons aussi dans le chapitre de Relativité Générale):

$$v = \sqrt{\frac{2GM}{R}} \quad (47.54)$$

Pour la Terre, une application numérique donne $v_L = 11.2 \cdot 10^3 \text{ [m/s]}$ et pour la Lune $v_L = 2.37 \cdot 10^3 \text{ [m/s]}$.

Rappelons que nous avons démontré dans le chapitre de Mécanique Des Milieux Continus lors de notre détermination de la température cinétique la relation suivante:

$$v = \sqrt{\frac{3kT}{m}} \quad (47.55)$$

En utilisant la masse molaire (cf. chapitre de Chimie Thermique):

$$v = \sqrt{\frac{3RT}{M_m}} \quad (47.56)$$

Une application numérique donne pour l'azote $v_{Az} = 517 \text{ [m/s]}$ et pour l'hydrogène $v_H = 1'934 \text{ [m/s]}$ avec une température arbitraire de 300 [K].

Donc l'azote est nettement piégé dans l'atmosphère terrestre. L'hydrogène, gaz léger, donc rapide l'est moins. Les deux gaz sont encore moins retenus par la Lune.

Remarque: En fait, la vitesse quadratique moyenne n'est pas la vitesse unique des molécules. Il y a une distribution des vitesses. Nous avons effectivement vu la distribution de Maxwell-Boltzmann d'un gaz à l'équilibre dans le chapitre de Mécanique Statistique.

LIMITE DE ROCHE

La limite de Roche est la distance théorique en dessous de laquelle un satellite commencerait à se disloquer sous l'action des forces de marées causées par le corps céleste autour duquel il orbite, ces forces dépassant la cohésion interne du satellite.

Nous pouvons simplifier le problème en considérant le satellite liquide et en le décomposant en deux petites masses m de rayon r et de masse volumique ρ_{sat} .

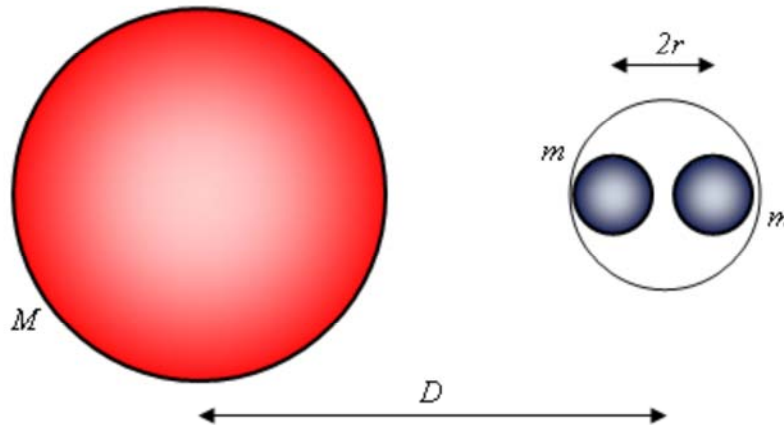


Figure: 47.5 - Configuration pour l'étude de la limite de Roche

La planète est une sphère de rayon R , de masse M , de masse volumique ρ_{pla} , située à une distance D de l'axe du satellite.

La planète exerce sur le satellite une attraction gravitationnelle:

$$F_a = G \frac{mM}{D^2} \quad (47.57)$$

La différence de force entre les 2 masses est:

$$\Delta F_a = G \frac{mM}{(D-r)^2} - G \frac{mM}{(D+r)^2} \quad (47.58)$$

Nous pouvons considérer $r \ll D$, ce qui donne:

$$\begin{aligned} \frac{1}{(D-r)^2} - \frac{1}{(D+r)^2} &\cong \frac{1}{D^2} \left(\frac{1}{1-2\frac{r}{D}} - \frac{1}{1+2\frac{r}{D}} \right) = \frac{1}{D^2} \left(\frac{4\frac{r}{D}}{\left(1-2\frac{r}{D}\right)\left(1+2\frac{r}{D}\right)} \right) \\ &\cong 4 \frac{r}{D^3} \end{aligned} \quad (47.59)$$

Donc la différence de force est

$$\Delta F_a = G \frac{4mM}{D^3} r \quad (47.60)$$

La force de cohésion du satellite résulte dans l'attraction gravitationnelle entre les 2 masses:

$$F_c = G \frac{m^2}{(2r)^2} \quad (47.61)$$

Le satellite est détruit si la différence de force entre les 2 masses est supérieure à la force de cohésion

$$G \frac{m^2}{4r^2} \leq G \frac{4mM}{D^3} r \quad (47.62)$$

Or nous avons les relations:

$$M = \frac{4}{3} \pi R^3 \rho_{pla} \text{ et } m = \frac{4}{3} \pi r^3 \rho_{sat} \quad (47.63)$$

donc nous obtenons:

$$\frac{\rho_{sat}}{\rho_{pla}} \leq 16 \frac{R^3}{D^3} \quad (47.64)$$

et nous en déduisons la "**limite de Roche**":

$$D_{Roche} = R \sqrt[3]{16 \frac{\rho_{pla}}{\rho_{sat}}} \quad (47.65)$$

Comme, dans ce calcul, nous avons considéré un satellite constitué de deux masses ponctuelles, et que de plus nous avons supposé que la cohésion du satellite était assurée exclusivement par les interactions gravitationnelles, cette valeur n'est qu'un ordre de grandeur (un minimum donc!).

Il faut savoir que les forces de marées de Jupiter ou de Saturne sur leurs satellites respectifs déforment de plusieurs centaines de mètres leur structure interne. Cela a pour effet de dégager de la chaleur par friction et déformation et provoque une dynamique de surface (éruptions) qui peut aider à l'émergence d'une forme de la vie élémentaire même dans les conditions extrêmes de température qui y règnent.

TRAJECTOIRES D'ORBITALES KÉPLÉRIENNES

L'observation (outil principal du physicien pour rappel) semble montrer qu'à première vue, les trajectoires suivies par les corps célestes en orbite autour d'astres sont bien du type conique (ouf!). Sachant cela, nous pouvons, afin de faciliter les calculs, anticiper la complexification des calculs et exprimer directement la dynamique d'un point matériel en des coordonnées polaires.

Comme nous l'avons vu dans le chapitre de Calcul Vectoriel, la vitesse en coordonnées polaires s'exprime par la relation (nous avons changé la lettre grecque de notation de l'angle pour nous adapter à la tradition):

$$\vec{v} = \frac{dr}{dt} \vec{u}_r + r \frac{d\theta}{dt} \vec{u}_\theta \quad (47.66)$$

où pour rappel le premier terme est la composante radiale de la vitesse et le second la composante tangentielle de la vitesse (angulaire).

Pour l'accélération (la démonstration se trouvant toujours dans le chapitre de Calcul Vectoriel):

$$\vec{a} = \left(\frac{d^2 r}{dt^2} - r \left(\frac{d\theta}{dt} \right)^2 \right) \vec{u}_r + \left(2 \frac{dr}{dt} \frac{d\theta}{dt} + r \frac{d^2 \theta}{dt^2} \right) \vec{u}_\theta \quad (47.67)$$

Maintenant que nous avons les outils nécessaires, attaquons-nous au cas des orbites Képlériennes dans le cas d'un champ Newtonien.

Il existe à notre connaissance deux manières principales de faire les développements mathématiques nécessaires mais qui n'amènent pas (à notre connaissance) au même niveau de détail quant au résultat. La première approche qui permet d'obtenir un résultat plus fin mais où la démarche est un peu du bricolage par moments... est basée sur l'utilisation de la vitesse radiale et d'une relation importante en astronomie, appelée "première formule de Binet". La seconde approche est la plus simple et la plus élégante, elle utilise l'accélération radiale pour l'approche du problème et une relation spéciale appelée "deuxième formule de Binet".

PREMIÈRE FORMULE DE BINET

Pour commencer par cette première approche du problème, rappelons que nous avons déjà démontré plus haut que:

$$\vec{F} = G \frac{Mm}{r^2} \vec{u} = -\overrightarrow{\text{grad}} \left(G \frac{Mm}{r} \right) = -m \cdot \overrightarrow{\text{grad}} (U) \quad (47.68)$$

Cependant, il est peu probable que le corps principal soit une sphère parfaite et homogène... Les astrophysiciens ont donc l'habitude de noter le potentiel Newtonien U sous la forme:

$$U = -\frac{\mu}{r} \cdot (f) \quad (47.69)$$

où $\mu = GM$ est appelée "**constante de gravitation de l'astre**" et où f est une fonction représentant les hétérogénéités de l'astre.

S'il est un endroit de l'Univers où les lois de la mécanique sont parfaitement vérifiables, c'est bien l'espace, parce que le frottement ou les causes de dissipation y sont extrêmement faibles. Dans le champ d'une seule force dérivant d'un potentiel, le mouvement vérifie la conservation de l'énergie mécanique.

Nous aboutissons ainsi à l'équation dite de l'énergie, dans laquelle E désigne "**l'énergie spécifique**" par unité de masse (kilogramme) envoyé.

$$E_T = E_c + U = \frac{1}{2} m v^2 - m \frac{\mu}{r} = m E \quad (47.70)$$

donc:

$$E = \frac{1}{2} v^2 - \frac{\mu}{r} \quad (47.71)$$

La force de gravitation Newtonienne est centrale, donc de couple nul au centre O du corps principal. Il en résulte la conservation du moment cinétique en norme et en direction, soit:

$$\vec{b} = \vec{r} \times m\vec{v} = m \cdot \vec{h} \Rightarrow \vec{h} = \vec{r} \times \vec{v} = K \cdot \vec{W} \quad (47.72)$$

Le vecteur $\vec{W}$ est l'unitaire de $\vec{b}$ ou de $\vec{h}$ appelé "**moment cinétique réduit**". K est la constante des aires (cf. chapitre de Mécanique Classique) telle que:

$$K = \|\vec{h}\| = \|\vec{r} \times \vec{v}\| \quad (47.73)$$

Nous rappelons que la norme de la vitesse exprimée en coordonnées polaires plane est donnée par la relation (n'oubliez pas que les deux vecteurs de la base polaire sont orthogonaux et que l'on peut donc appliquer le théorème de Pythagore pour calculer la norme comme il l'a été démontré dans le chapitre de Calcul Vectoriel du site):

$$v^2 = \left(\frac{dr}{dt} \right)^2 + r^2 \left(\frac{d\theta}{dt} \right)^2 \quad (47.74)$$

Ce qui nous permet d'écrire pour la constante des aires K :

$$K = \|\vec{h}\| = \|\vec{r} \times \vec{v}\| = \left\| \vec{r} \times \left(\frac{dr}{dt} \vec{u}_r + r \frac{d\theta}{dt} \vec{u}_\theta \right) \right\| = r^2 \frac{d\theta}{dt} \quad (47.75)$$

Plaçons-nous dans le plan orbital, en coordonnées polaires. Nous possédons deux intégrales premières dépendant des deux constantes essentielles E et K .

Soit la relation déjà démontrée:

$$\vec{v} = \frac{dr}{dt} \vec{u}_r + r \frac{d\theta}{dt} \vec{u}_\theta \quad (47.76)$$

et sa norme au carré:

$$v^2 = \left(\frac{dr}{dt} \right)^2 + r^2 \left(\frac{d\theta}{dt} \right)^2 \quad (47.77)$$

Or dans le cas d'une force centrale:

$$r^2 \frac{d\theta}{dt} = K \Leftrightarrow \frac{d\theta}{dt} = \frac{K}{r^2} \quad (47.78)$$

Mettons cela dans l'expression de v^2 antérieure, nous avons alors:

$$\begin{aligned}
 v^2 &= \left(\frac{dr}{dt}\right)^2 + r^2 \left(\frac{d\theta}{dt}\right)^2 = \left(\frac{dr}{dt} \frac{d\theta}{d\theta}\right)^2 + r^2 \left(\frac{K}{r^2}\right)^2 = \left(\frac{d\theta}{dt} \frac{dr}{d\theta}\right)^2 + \frac{K^2}{r^2} \\
 &= \left(\frac{K}{r^2} \frac{dr}{d\theta}\right)^2 + \frac{K^2}{r^2} = \left(-K \frac{d}{d\theta} \left(\frac{1}{r}\right)\right)^2 + \frac{K^2}{r^2} = K^2 \left(\left(\frac{d}{d\theta} \left(\frac{1}{r}\right)\right)^2 + \frac{1}{r^2}\right) \quad (47.79)
 \end{aligned}$$

Cette dernière relation est appelée "**première formule de Binet**".

En égalant avec l'expression de v^2 résultant de la conservation de l'énergie vue un peu plus haut, nous avons:

$$K^2 \left(\left(\frac{d}{d\theta} \left(\frac{1}{r} \right) \right)^2 + \frac{1}{r^2} \right) = \frac{2\mu}{r} + 2E \quad (47.80)$$

Ce qui nous donne une équation différentielle assez compliquée:

$$\left(\frac{d}{d\theta} \left(\frac{1}{r} \right) \right)^2 = - \left(\frac{1}{r} - \frac{\mu}{K^2} \right)^2 + \frac{\mu^2}{K^4} + \frac{2E}{K^2} \quad (47.81)$$

Et là nous nous demandons comment nous pouvons faire pour nous en sortir? Après quelques heures de réflexion... nous nous rendons compte qu'il faut faire une substitution. Après une autre heure de chaos neuronal cela finit par aboutir. Nous décidons de poser (nous en avons tout à fait le droit), sachant que r est une fonction u de θ :

$$\frac{1}{r} - \frac{\mu}{K^2} = \sqrt{\frac{\mu^2}{K^4} + \frac{2E}{K^2}} u \quad (47.82)$$

Dérivons allègrement par rapport à θ :

$$\frac{d \left(\frac{1}{r} \right)}{d\theta} = \sqrt{\frac{\mu^2}{K^4} + \frac{2E}{K^2}} \frac{du}{d\theta} \quad (47.83)$$

Substituons dans l'équation différentielle:

$$\left(\frac{\mu^2}{K^4} + \frac{2E}{K^2} \right) \left(\frac{du}{d\theta} \right)^2 = - \left(\frac{\mu^2}{K^4} + \frac{2E}{K^2} \right) u^2 + \left(\frac{\mu^2}{K^4} + \frac{2E}{K^2} \right) \quad (47.84)$$

Après simplification, nous obtenons:

$$\frac{du}{d\theta} = \pm \sqrt{1 - u^2} \quad (47.85)$$

Nous séparons les variables pour intégrer:

$$\int^{\pm} \frac{du}{\sqrt{1-u^2}} = \int d\theta \quad (47.86)$$

Nous avons deux solutions suivant le signe que nous choisissons. Cependant, à la fin de la résolution, nous remarquons que le seul choix physiquement intéressant est le signe négatif. Nous avons démontré dans le chapitre de Calcul Différentiel Et Intégral lors de notre étude des dérivées usuelles que:

$$\begin{aligned} + \frac{d \arccos(x)}{dx} &= - \frac{1}{\sqrt{1-x^2}} \\ - \frac{d \arcsin(x)}{dx} &= - \frac{1}{\sqrt{1-x^2}} \end{aligned} \quad (47.87)$$

Nous allons opter pour la primitive en cosinus et dès lors, nous avons:

$$\arccos(u) = \theta - \theta_0 \Rightarrow u = \cos(\theta - \theta_0) \quad (47.88)$$

Nous laissons, par approximation, de côté la constante d'intégration qui impliquerait des très faibles oscillations sur la trajectoire de l'orbite (si vous faites une étude ou un TP sur le sujet, communiquez-moi les graphiques que vous obtenez avec ou sans la constante, cela m'intéresserait).

Ce qui nous permet d'obtenir:

$$\frac{1}{r} = \frac{\mu}{K^2} + \sqrt{\frac{\mu^2}{K^4} + \frac{2E}{K^2}} \cos(\theta - \theta_0) \quad (47.89)$$

Or, nous voyons que notre choix du signe pour l'intégration se justifie pleinement puisque maintenant, si nous faisons un petit rappel sur les coniques (cf. chapitre de Géométrie Analytique), nous voyons que nous avons:

$$r = \frac{p}{1 + e \cos \theta} \quad (47.90)$$

où e est l'excentricité (rapport du petit axe $e = c/a < 1$) et p le paramètre focal ($p = b^2/a$) d'une ellipse. Ce qui correspond bien aux trajectoires que suivent les astres en orbite.

Nous retrouvons donc bien la première "loi" de Kepler...

Dans notre cas, nous avons après simplification:

$$p = \frac{K^2}{\mu} \quad \text{et} \quad e = \sqrt{1 + \frac{2EK^2}{\mu^2}} \quad (47.91)$$

où (pour rappel) K est la constante des aires:

$$K = \|\vec{h}\| = \|\vec{r} \times \vec{v}\| \quad (47.92)$$

et μ la constante de gravitation de l'astre:

$$\mu = GM \quad (47.93)$$

et enfin E l'énergie spécifique:

$$E = \frac{1}{2}v^2 - \frac{\mu}{r} \quad (47.94)$$

Le lecteur vérifiera comme nous l'avons vu dans le chapitre de Géométrie Analytique lors de notre étude des coniques que si:

- $E \geq 0, e = 1$ nous avons une orbite ouverte sous forme de parabole
- $e > 1$ nous avons une orbite ouverte sous forme d'hyperbole
- $E \leq 0, 0 \leq e < 1$ nous avons une orbite fermée sous forme d'une ellipse ou de cercle.

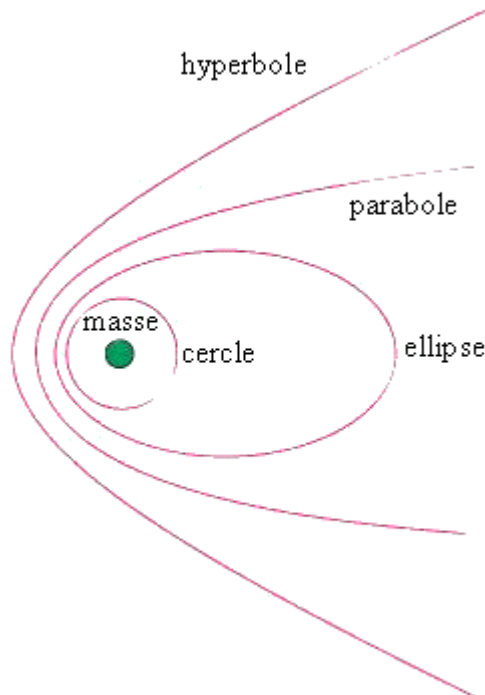


Figure: 47.6 - Rappels sur les coniques

Enfin, si nous injectons:

$$r = \frac{p}{1 + e \cos \theta} \quad (47.95)$$

dans la première formule de Binet:

$$v^2 = K^2 \left(\frac{d}{d\theta} \left(\frac{1}{r} \right)^2 + \frac{1}{r^2} \right) \quad (47.96)$$

nous obtenons alors la vitesse en n'importe quel point de l'ellipse en fonction du principal paramètre variable qui est donc l'angle.

DEUXIÈME FORMULE DE BINET

Voyons maintenant l'approche basée sur l'accélération radiale qui tout en étant plus élégante, nous permet d'obtenir un résultat moins fin concernant les paramètres de l'ellipse.

Nous partons donc l'expression de l'accélération en coordonnées polaires (cf. chapitre de Calcul Vectoriel):

$$\vec{a} = \left(\frac{d^2 r}{dt^2} - r \left(\frac{d\theta}{dt} \right)^2 \right) \vec{u}_r + \left(2 \frac{dr}{dt} \frac{d\theta}{dt} + r \frac{d^2 \theta}{dt^2} \right) \vec{u}_\theta \quad (47.97)$$

Nous pouvons simplifier l'écriture du deuxième terme:

$$2 \frac{dr}{dt} \frac{d\theta}{dt} + r \frac{d^2 \theta}{dt^2} = \frac{1}{r} \frac{d}{dt} \left(r^2 \frac{d\theta}{dt} \right) \quad (47.98)$$

Or, nous venons de voir plus haut que:

$$r^2 \frac{d\theta}{dt} = K = c^{te} \quad (47.99)$$

et donc:

$$\frac{1}{r} \frac{d}{dt} \left(r^2 \frac{d\theta}{dt} \right) = 0 \quad (47.100)$$

L'accélération se réduit alors à:

$$\vec{a} = \left(\frac{d^2 r}{dt^2} - r \left(\frac{d\theta}{dt} \right)^2 \right) \vec{u}_r \quad (47.101)$$

Nous pouvons éliminer le temps en écrivant:

$$\frac{dr}{dt} = \frac{dr}{d\theta} \frac{d\theta}{dt} \quad (47.102)$$

et:

$$\frac{d^2 r}{dt^2} = \frac{d}{d\theta} \left(\frac{dr}{d\theta} \frac{d\theta}{dt} \right) \frac{d\theta}{dt} \quad (47.103)$$

Alors nous obtenons:

$$\begin{aligned}
\vec{a} &= \left(\frac{d^2 r}{dt^2} - r \left(\frac{d\theta}{dt} \right)^2 \right) \vec{u}_r = \left(\frac{d}{d\theta} \left(\frac{dr}{d\theta} \frac{d\theta}{dt} \right) \frac{d\theta}{dt} - r \left(\frac{d\theta}{dt} \right)^2 \right) \vec{u}_r \\
&= \left(\frac{d}{d\theta} \left(\frac{dr}{d\theta} \frac{K}{r^2} \right) \frac{d\theta}{dt} - r \left(\frac{K}{r^2} \right)^2 \right) \vec{u}_r = \left(\frac{d}{d\theta} \left(\frac{dr}{d\theta} \frac{K}{r^2} \right) \frac{K}{r^2} - r \left(\frac{K}{r^2} \right)^2 \right) \vec{u}_r \\
&= \frac{K^2}{r^2} \left(\frac{d}{d\theta} \left(\frac{dr}{d\theta} \frac{1}{r^2} \right) - \frac{1}{r} \right) \vec{u}_r = \frac{K^2}{r^2} \left(\frac{d}{d\theta} \left(-\frac{d}{d\theta} \left(\frac{1}{r} \right) \right) - \frac{1}{r} \right) \vec{u}_r \\
&= -\frac{K^2}{r^2} \left(\frac{d^2}{d\theta^2} \left(\frac{1}{r} \right) + \frac{1}{r} \right) \vec{u}_r
\end{aligned} \tag{47.104}$$

Et donc il vient pour la norme la "**deuxième formule de Binet**":

$$a = -\frac{K^2}{r^2} \left(\frac{d^2}{d\theta^2} \left(\frac{1}{r} \right) + \frac{1}{r} \right) \tag{47.105}$$

Or, selon la deuxième loi de Newton et de sa loi de Gravitation, nous avons:

$$F = ma = \frac{GMm}{r^2} \Rightarrow a = \frac{GM}{r^2} = \frac{\mu}{r^2} \tag{47.106}$$

Nous avons alors avec la deuxième formule de Binet:

$$\frac{\mu}{r^2} = -\frac{K^2}{r^2} \left(\frac{d^2}{d\theta^2} \left(\frac{1}{r} \right) + \frac{1}{r} \right) \tag{47.107}$$

Soit après simplification et en choisissant le signe de l'accélération à notre convenance pour se débarrasser du signe "-", nous avons:

$$\mu = K^2 \left(\frac{d^2}{d\theta^2} \left(\frac{1}{r} \right) + \frac{1}{r} \right) \tag{47.108}$$

En isolant les constantes, il vient:

$$\frac{d^2}{d\theta^2} \left(\frac{1}{r} \right) + \frac{1}{r} = \frac{\mu}{K^2} \tag{47.109}$$

Après un petit changement de variables nous reconnaissons le cas particulier d'une équation différentielle du second ordre que nous avons déjà rencontrée plusieurs fois jusqu'à maintenant dans les différents chapitres du site et que nous rencontrerons encore:

$$\frac{d^2 u}{d\theta^2} + u = \frac{\mu}{K^2} \quad (47.110)$$

Comme il est d'usage, nous montrons toutefois les détails de sa résolution. L'équation sans second membre est (cf. chapitre de Calcul Différentiel Et Intégral):

$$\frac{d^2 u}{d\theta^2} + u = 0 \quad (47.111)$$

Nous avons alors le discriminant qui est négatif puisque:

$$\frac{-a \pm \sqrt{a^2 - 4b}}{2} = \frac{\sqrt{-4}}{2} = \sqrt{-1} \quad (47.112)$$

Nous avons alors vu dans le chapitre de Calcul Différentiel Et Intégral, que dans cette situation la solution de l'équation homogène était de la forme:

$$y_h = C' e^{\alpha x} \cos(\beta t - \phi) \quad (47.113)$$

Soit dans la situation qui nous concerne, nous avons:

$$u_h = A e^{0 \cdot \theta} \cos(1 \cdot \theta - \theta_0) = A \cos(\theta - \theta_0) \quad (47.114)$$

Nous injectons la solution homogène dans l'équation différentielle avec second membre:

$$\frac{d^2 A \cos(\theta - \theta_0)}{d\theta^2} + A \cos(\theta - \theta_0) = \frac{\mu}{K^2} \quad (47.115)$$

Et voyons immédiatement que pour que l'égalité soit satisfaite, la solution générale est:

$$u = \frac{1}{r} = A \cos(\theta - \theta_0) + \frac{\mu}{K^2} \quad (47.116)$$

Soit après réarrangement:

$$r = \frac{\frac{K^2}{\mu}}{1 + \frac{K^2 A}{\mu} \cos(\theta - \theta_0)} = \frac{p}{1 + e \cos(\theta - \theta_0)} \quad (47.117)$$

Et en choisissant l'angle initial comme étant nul, nous retrouvons donc bien:

$$\boxed{r = \frac{p}{1 + e \cos \theta}} \quad (47.118)$$

à la différence avec la première méthode de résolution que la valeur de la constante A reste inconnue.

PÉRIODE ORBITALE KÉPLÉRIENNE

La loi des aires permet, comme nous le savons déjà, de calculer la période orbitale Képlérienne T . En effet, l'aire S de l'ellipse valant $S = \pi \cdot a \cdot b$ (cf. chapitre sur les Formes Géométriques) et ayant déjà déterminé lors de la définition du moment cinétique la relation (cf. chapitre de Mécanique Classique):

$$\frac{dS}{dt} = \frac{K}{2} \quad (47.119)$$

Il vient naturellement:

$$\frac{\pi ab}{T} = \frac{K}{2} \quad (47.120)$$

Par ailleurs, l'étude des coniques (cf. chapitre de Géométrie Analytique) nous a montré que:

$$p = \frac{b^2}{a} = \frac{T^2 K^2}{4\pi^2 a^3} \quad (47.121)$$

et nous avons défini plus haut:

$$p = \frac{K^2}{\mu} \quad (47.122)$$

Nous avons donc la relation:

$$T = \sqrt{\frac{a^3 4\pi^2}{\mu}} \quad (47.123)$$

et nous retrouvons du même coup la troisième loi de Kepler...:

$$\frac{a^3}{T^2} = \frac{GM}{4\pi^2} \quad (47.124)$$

ce qui valide nos calculs précédents.

DÉFLEXION CLASSIQUE DE LA LUMIÈRE

Les calculs effectués précédemment peuvent s'appliquer à un cas intéressant: la déviation de la lumière par un astre selon une interprétation Newtonienne (bien évidemment!).

Attention!!! Newton ne savait pas à l'époque que le photon était sans masse. Les développements qui suivent sont donc une approche erronée à notre époque et qu'il convient de prendre avec des pincettes mais qui reste enseignée aujourd'hui car elle permet à des étudiants n'ayant pas encore étudié la relativité générale ou qui ne l'étudieront jamais (dans le chapitre de Relativité Générale, le lecteur trouvera la démonstration détaillée contemporaine de la déflexion de la lumière qui est d'un tout autre niveau) d'avoir une première approche... C'est comme dans tout en physique! Tant qu'on

n'a pas atteint le niveau de la licence universitaire, on apprend plein de choses "fausses" car simplifiées à l'extrême. Ensuite au Master et Doctorat, on apprend des théories un peu plus réalistes et valides.

Bon ceci étant rappelé (suite à une remarque d'un lecteur), nous avons donc montré plus haut que:

$$r = \frac{p}{1 + e \cos(\theta)} \quad (47.125)$$

Dans le cas d'un photon, nous aurions tendance à poser que $r \rightarrow \infty$ (donc une trajectoire de type hyperbolique) et donc pour cela il faut que dans la relation précédente nous ayons (ce qui est équivalent à dire que e est strictement supérieure à l'unité comme l'impose la trajectoire hyperbolique):

$$\cos(\theta) = -\frac{1}{e} \quad (47.126)$$

en posant $\varphi = 2\theta - \pi$ les relations trigonométriques élémentaires (cf. chapitre de Trigonométrie) nous donnent:

$$\sin\left(\frac{\varphi}{2}\right) = \sin\left(\theta - \frac{\pi}{2}\right) = -\cos(\theta) = \frac{1}{e} \quad (47.127)$$

et donc en utilisant encore les relations trigonométriques:

$$e^2 = \frac{1}{\sin^2\left(\frac{\varphi}{2}\right)} = 1 + \cot^2\left(\frac{\varphi}{2}\right) \quad (47.128)$$

soit:

$$\cot^2\left(\frac{\varphi}{2}\right) = e^2 - 1 \quad (47.129)$$

et nous savons que:

$$e^2 = 1 + \frac{2EK^2}{\mu^2} \quad (47.130)$$

donc:

$$e^2 - 1 = \frac{2EK^2}{\mu^2} \quad (47.131)$$

en négligeant l'énergie potentielle du photon puisque $r \rightarrow \infty$, nous avons (attention!!! rappelons que selon ce que nous avons vu dans le chapitre de Relativité Restreinte, le photon n'a pas de masse rigoureusement mais Newton n'en savait rien à l'époque!):

$$E = \frac{1}{2}v^2 \quad (47.132)$$

Donc:

$$\cot^2\left(\frac{\varphi}{2}\right) = \frac{v^2 K^2}{G^2 M^2} \quad (47.133)$$

donc:

$$\tan^2\left(\frac{\varphi}{2}\right) = \frac{G^2 M^2}{v^2 K^2} \quad (47.134)$$

après simplification:

$$\tan\left(\frac{\varphi}{2}\right) = \frac{GM}{vK} \quad (47.135)$$

et comme φ est supposé petit, nous avons à l'aide du développement de Taylor (cf. chapitre sur les Suites Et Séries) de la fonction tangente:

$$\tan\left(\frac{\varphi}{2}\right) \cong \frac{\varphi}{2} \quad (47.136)$$

il vient donc finalement:

$$\varphi \cong \frac{2GM}{vK} \quad (47.137)$$

Or, nous avons par définition:

$$K = r^2 \frac{d\theta}{dt} = r^2 \dot{\theta} \quad (47.138)$$

et nous savons que $v = \omega r = \dot{\theta} r$ (cf. chapitre de Mécanique Classique). Ainsi il vient:

$$\boxed{\varphi \cong \frac{2GM}{v^2 r}} \quad (47.139)$$

Si la particule est un photon passant au ras de la surface du Soleil alors:

$$\varphi \cong \frac{2GM}{c^2 R_{\odot}} \quad (47.140)$$

une application numérique donne:

$$\varphi \cong 4.237 \cdot 10^{-6} [rad] \underset{\times \frac{360}{2\pi}}{=} 2.427 \cdot 10^{-4} \circ \underset{\times 60 \times 60}{=} 0.87'' \quad (47.141)$$

La théorie Newtonienne prévoit donc une déviation de 0.87 secondes d'arc pour un rayon lumineux frôlant la surface du Soleil. Ce qui est deux fois moins que ce qui peut être observé expérimentalement et que ce que donne la relativité générale (cf. chapitre de Relativité Générale)!

PRÉCESSION DU PÉRIHÉLIE

Avant d'étudier la précession des orbites, nous souhaiterions rappeler que le champ gravitationnel est un champ conservatif et central. Ceci implique donc que le moment cinétique (cf. chapitre de Mécanique Classique) est constant et que la trajectoire a lieu dans un plan dont le vecteur normal à la surface conserve toujours la même direction (le vecteur moment cinétique est constant en norme et en direction pour rappel!).

Nous nous attaquerons à l'analyse de la précession du périhélie en prenant en compte les résultats de la théorie de la relativité restreinte (cela permettant d'être plus fin dans les résultats obtenus et de pouvoir appliquer ces mêmes résultats aux électrons en orbite autour du noyau de l'atome).

Définitions:

D1. Le "**périhélie**" est le point de l'orbite d'un corps céleste (planète, comète, etc.) qui est le plus rapproché de l'étoile autour de laquelle il tourne.

D2. "L'**aphélie**" est le point de l'orbite d'un objet (planète, comète, etc.) où il est le plus éloigné de l'étoile autour de laquelle il tourne.

D3. "L'**équinoxe**" est l'instant où l'étoile centrale traverse le plan de l'équateur de l'objet qui est en orbite autour de lui.

Remarque: Lorsque le Soleil passe de l'hémisphère Sud à l'hémisphère Nord de la Terre (en d'autres termes que le Soleil se trouve au Zénith à l'équateur à midi), c'est l'équinoxe de printemps (20 ou 21 mars), dans le sens inverse, c'est l'équinoxe d'automne (22 ou 23 septembre). À ces dates, il y a égalité du jour et de la nuit sur toute la Terre.

Évidemment, le résultat que nous obtiendrons ne sera pas complet, puisque comme nous le savons, il a fallu attendre le développement de la relativité générale pour donner avec exactitude la précession du périhélie de Mercure (nous y reviendrons).

Pour calculer cet effet de précession, nous allons rechercher l'équivalent des formules de Binet vues plus haut sous forme relativiste (nous verrons la forme classique dans le chapitre de Relativité Générale). Nous procédons comme suit:

Le lagrangien relativiste du système (cf. chapitre de Relativité Restreinte):

$$L = T + V = \mu c^2 (\gamma - 1) + V \quad (47.142)$$

Remarque: Nous soustrayons l'énergie au repos, car seul nous intéresse ici l'étude de l'énergie cinétique et potentielle. L'énergie potentielle est sommée dans le lagrangien ci-dessus (ce qui n'est pas conforme à l'usage) mais nous inverserons le signe plus loin lors des développements.

Avec:

$$\gamma = \left(1 - \frac{v^2}{c^2}\right)^{-\frac{1}{2}} = \left(1 - \frac{\dot{r}^2 + r^2 \dot{\theta}^2}{c^2}\right)^{-\frac{1}{2}} \quad (47.143)$$

et la masse réduite:

$$\mu = \frac{mM}{m + M} \quad (47.144)$$

Remarque: Pour déterminer l'expression de la vitesse en coordonnées polaires, nous avons utilisé le résultat de nos calculs du chapitre de Calcul Vectoriel.

Le moment cinétique:

$$\vec{b} = \mu \vec{r} \times \vec{v} \quad (47.145)$$

sous forme relativiste et appliqué à notre étude s'écrit:

$$\vec{b} = (\mu \gamma) \vec{r} \times (\vec{\omega} \times \vec{r}) = (\mu \gamma) \vec{r} \times (\vec{\omega} \times \vec{r}) \quad (47.146)$$

En prenant la norme, nous avons sans oublier que dans notre étude $\vec{\omega} \perp \vec{r}$ et donc $(\vec{\omega} \perp \vec{r}) \perp \vec{r}$:

$$b = (\mu \gamma) \omega r^2 \quad (47.147)$$

et rappelons que nous avons adopté l'écriture $\omega = \dot{\theta}$. Ce qui nous donne finalement:

$$b = (\mu \gamma) r^2 \dot{\theta} \quad (47.148)$$

Pour établir l'équivalent relativiste des formules de Binet:

- nous déduisons de l'expression du moment cinétique:

$$\dot{\theta} = d_t \theta = \frac{b}{\mu \gamma r^2} \quad (47.149)$$

- nous recherchons une relation du type $\dot{r} = \dot{r}(\theta)$ (puisque la trajectoire est une conique):

$$\dot{r} = d_t r = \frac{b}{\mu \gamma r^2} d_\theta r \quad (47.150)$$

Effectivement car rappelons qu'en coordonnées polaires la vitesse est donnée par l'expression suivante:

$$\vec{v} = \dot{\vec{r}} = \dot{r} \vec{e}_r + r \dot{\theta} \vec{e}_\theta \quad (47.151)$$

C'est-à-dire que $\dot{r} = f(r, \theta)$. Cette dernière expression permet d'écrire que:

$$\dot{r} = \frac{dr}{dt} = \frac{dr}{d\theta} \frac{d\theta}{dt} = d_{\theta} r \cdot \dot{\theta} \quad (47.152)$$

- nous cherchons ensuite une relation $\ddot{r} = \ddot{r}(\theta)$:

$$\ddot{r} = d_t(d_t r) = d_{\theta}(d_t r) d_t \theta = \left[d_{\theta} \left(\frac{b}{\mu \gamma r^2} d_{\theta} r \right) \right] \frac{b}{\mu \gamma r^2} \quad (47.153)$$

Soit:

$$\ddot{r} = \frac{b}{\mu \gamma r^2} \left[d_{\theta} \left(\frac{b}{\mu \gamma r^2} d_{\theta} r \right) \right] \quad (47.154)$$

À partir des équations obtenues précédemment, nous avons successivement:

$$\mu c^2 (\gamma - 1) + V = L \Leftrightarrow \gamma^2 = \left(1 + \frac{L - V}{\mu c^2} \right)^2 \quad (47.155)$$

Rappelons que nous avons défini en relativité restreinte:

$$\beta = \frac{v^2}{c^2} = \frac{\dot{r}^2 + r^2 \dot{\theta}^2}{c^2} \quad (47.156)$$

Avec les équations précédentes, cela nous donne:

$$\begin{aligned} \beta &= \frac{v^2}{c^2} = \frac{\dot{r}^2 + r^2 \dot{\theta}^2}{c^2} = \frac{(d_{\theta} r \dot{\theta})^2 + r^2 \left(\frac{b}{\mu \gamma r^2} \right)^2}{c^2} = \frac{\left(d_{\theta} r \frac{b}{\mu \gamma r^2} \right)^2 + r^2 \left(\frac{b}{\mu \gamma r^2} \right)^2}{c^2} \\ &= \frac{b^2}{\mu^2 \gamma^2 c^2} \left[\left(\frac{1}{r^2} d_{\theta} r \right)^2 + r^2 \left(\frac{1}{r^2} \right)^2 \right] = \frac{b^2}{\mu^2 \gamma^2 c^2} \left[\left(\frac{1}{r^2} d_{\theta} r \right)^2 + \frac{1}{r^2} \right] \end{aligned} \quad (47.157)$$

D'autre part:

$$\gamma^2 = \frac{1}{1 - \beta} = 1 + \frac{\beta}{1 - \beta} = 1 + \gamma^2 \beta \quad (47.158)$$

En introduisant l'avant-dernière relation dans cette dernière:

$$\gamma^2 = 1 + \cancel{\gamma^2} \frac{b^2}{\mu^2 \cancel{\gamma^2} c^2} \left[\left(\frac{1}{r^2} d_{\theta} r \right)^2 + \frac{1}{r^2} \right] \quad (47.159)$$

En posant $u = \frac{1}{r}$ et comme:

$$\frac{1}{r^2} d_{\theta} r = d_r \left(-\frac{1}{r} \right) d_{\theta} r = -\frac{d \left(-\frac{1}{r} \right)}{dr} \frac{dr}{d\theta} = -d_{\theta} \left(\frac{1}{r} \right) = -d_{\theta} u \quad (47.160)$$

L'avant-dernière relation devient avec cette dernière expression:

$$\gamma^2 = 1 + \frac{b^2}{\mu^2 c^2} \left[\left(-d_{\theta} u \right)^2 + u^2 \right] \quad (47.161)$$

En égalant cette dernière relation avec celle du lagrangien:

$$1 + \frac{b^2}{\mu^2 c^2} \left[d_{\theta}^2 u + u^2 \right] = \left(1 + \frac{L - V}{\mu c^2} \right)^2 \quad (47.162)$$

En dérivant cette dernière relation par rapport à θ :

$$\frac{b^2}{\mu^2 c^2} \left[2d_{\theta} u (d_{\theta}^2 u) + 2u d_{\theta} u \right] = 2 \left(1 + \frac{L - V}{\mu c^2} \right) \left(-\frac{1}{\mu c^2} d_{\theta} V \right) \quad (47.163)$$

Effectivement, le lagrangien étant constant au cours du temps (le système est conservatif !), nous avons donc:

$$d_{\theta} L = 0 \quad (47.164)$$

et également:

$$\frac{d}{d\theta} \left(\frac{du(\theta)}{d\theta} \frac{du(\theta)}{d\theta} \right) = \frac{d^2 u(\theta)}{d\theta^2} \frac{du(\theta)}{d\theta} + \frac{du(\theta)}{d\theta} \frac{d^2 u(\theta)}{d\theta^2} = 2 \frac{d}{d\theta} u(\theta) \frac{d^2 u(\theta)}{d\theta^2} \quad (47.165)$$

Or, si nous continuons:

$$-\frac{dV}{d\theta} = -\frac{dV}{dr} \frac{dr}{d\theta} \quad (47.166)$$

En se référant à:

$$\frac{1}{r^2} d_{\theta} r = -d_{\theta} u \rightarrow d_{\theta} r = -r^2 d_{\theta} u = -\frac{1}{u^2} d_{\theta} u \quad (47.167)$$

Nous obtenons donc:

$$-\frac{dV}{d\theta} = \left(-\frac{dV}{dr} \right) \cdot \left(-\frac{1}{u^2} d_{\theta} u \right) = \frac{1}{u^2} \frac{dV}{dr} d_{\theta} u \quad (47.168)$$

Ce qui donne finalement après quelques simplifications:

$$\frac{b^2}{\mu^2 c^2} \left[d_{\theta}^2 u + u \right] = \left(1 + \frac{L-V}{\mu c^2} \right) \frac{1}{\mu c^2} \frac{1}{u^2} \frac{dV}{dr} \quad (47.169)$$

En multipliant cette dernière par $\frac{\mu^2 c^2}{b^2}$:

$$\begin{aligned} d_{\theta}^2 u + u &= \frac{\mu^2 c^2}{b^2} \left(1 + \frac{L-V}{\mu c^2} \right) \frac{1}{\mu c^2} \frac{1}{u^2} \frac{dV}{dr} \\ d_{\theta}^2 u + u &= \frac{\mu}{b^2 u^2} \left(1 + \frac{L-V}{\mu c^2} \right) \frac{dV}{dr} \end{aligned} \quad (47.170)$$

Dans un potentiel gravitationnel:

$$V = -G \frac{mM}{r} \rightarrow \frac{dV}{dr} = - \left(-G \frac{mM}{r^2} \right) = G \frac{mM}{r^2} = GmMu^2 \quad (47.171)$$

L'équation de Binet en relativité restreinte est alors:

$$d_{\theta}^2 u + u = \frac{GmM\mu}{b^2} \left(1 + \frac{L + GmMu}{\mu c^2} \right) \quad (47.172)$$

Pour rechercher une solution à cette équation différentielle, nous allons grouper la variable u dans le membre de gauche:

$$\begin{aligned} d_{\theta}^2 u + u &= \frac{GmM\mu}{b^2} + \frac{GmML}{b^2 c^2} + \frac{G^2 m^2 M^2}{b^2 c^2} u \\ \Rightarrow d_{\theta}^2 u + u \left(1 - \frac{G^2 m^2 M^2}{b^2 c^2} \right) &= \frac{GmM\mu}{b^2} \left(1 + \frac{L}{\mu c^2} \right) \end{aligned} \quad (47.173)$$

Nous posons:

$$\Omega^2 = \left(1 - \frac{G^2 m^2 M^2}{b^2 c^2} \right) \quad \text{et} \quad B = \frac{GmM\mu}{b^2} \left(1 + \frac{L}{\mu c^2} \right) \quad (47.174)$$

L'équation différentielle s'écrit alors:

$$d_{\theta}^2 u + (\Omega^2 u - B) = 0 \quad (47.175)$$

Nous posons:

$$(\Omega^2 u - B) = W \quad (47.176)$$

En prenant la dérivée seconde:

$$\Omega^2 d_\theta^2 u = d_\theta^2 W \Rightarrow d_\theta^2 u = \frac{1}{\Omega^2} d_\theta^2 W \quad (47.177)$$

Nous trouvons alors une simple équation différentielle dont la solution est bien connue:

$$d_\theta^2 W + \Omega^2 W = 0 \quad (47.178)$$

Les solutions sont du type:

$$\begin{aligned} W &= C_1 \cos(\Omega \theta) + C_2 \sin(\Omega \theta) = \Omega^2 u - B \\ \Rightarrow u &= \frac{C_1}{\Omega^2} \cos(\Omega \theta) + \frac{C_2}{\Omega^2} \sin(\Omega \theta) + \frac{B}{\Omega^2} \end{aligned} \quad (47.179)$$

Ce qui s'écrit encore puisque Ω^2 est une constante:

$$u = k_1 \cos(\Omega \theta) + k_2 \sin(\Omega \theta) + \frac{B}{\Omega^2} \quad (47.180)$$

avec $k_1, k_2 = c^{te}$.

Pour déterminer les constantes k_1, k_2 nous nous plaçons d'abord dans la situation pour laquelle $\theta = 0$, où r est minimal et donc par définition u maximal.

Nous dérivons par rapport à θ :

$$(d_\theta u)_{\theta=0} = \left[\underbrace{-k_1 \Omega \sin(\Omega \theta)}_{=0} + \underbrace{k_2 \Omega \cos(\Omega \theta)}_{k_2 \neq 0} \right]_{\theta=0} = 0 \quad (47.181)$$

Donc $k_2 = 0$ ce qui fait que la relation:

$$u = k_1 \cos(\Omega \theta) + k_2 \sin(\Omega \theta) + \frac{B}{\Omega^2} \quad (47.182)$$

devient:

$$\frac{1}{r} = k_1 \cos(\Omega \theta) + \frac{B}{\Omega^2} = \frac{B}{\Omega^2} \left(1 + \frac{\Omega^2 k_1}{B} \cos(\Omega \theta) \right) \quad (47.183)$$

É crite autrement (en essayant de revenir sur une notation similaire à celle de l'étude des coniques) cela donne:

$$r = \frac{1}{\frac{B}{\Omega^2} \left(1 + \frac{k_1 \Omega^2}{B} \cos(\Omega \theta) \right)} = \frac{\frac{\Omega^2}{B}}{1 + \frac{k_1 \Omega^2}{B} \cos(\Omega \theta)} = \frac{p}{1 + e \cos(\Omega \theta)} \quad (47.184)$$

Et l'intérêt d'écrire cela ainsi est de remarquer que nous retombons en fin de compte sur l'équation d'une ellipse avec p étant le paramètre focal de la conique, paramètre focal donné par (cf. chapitre de Géométrie Analytique):

$$\boxed{p = a(1 - e^2)} \quad (47.185)$$

où a est le demi-grand axe de l'ellipse.

Maintenant posons:

$$D = \frac{B}{\Omega^2} \text{ et } \varepsilon = \frac{\Omega^2 k_1}{B} \quad (47.186)$$

Au premier passage par le périhélie $\theta = 0$ où:

$$\cos(\Omega \theta_0) = \cos(0) = 1 \quad (47.187)$$

nous avons donc:

$$\frac{1}{r} = D(1 + \varepsilon) \quad (47.188)$$

Au deuxième passage par le périhélie $\theta = 2\pi$, nous avons:

$$\cos(\Omega \theta_1) = \cos(2\pi) = 1 \quad (47.189)$$

nous avons donc également:

$$\frac{1}{r} = D(1 + \varepsilon) \quad (47.190)$$

La trajectoire est toujours une ellipse mais l'angle $\Omega \theta_0$ qui était nul au départ est devenu $\Omega \theta_1 = 2\pi$.

Soit si nous avons:

$$\theta_1 = \theta_0 + \Delta\theta \quad (47.191)$$

Alors:

$$\Omega(\theta_0 + \Delta\theta) - \Omega \theta_0 = 2\pi \quad (47.192)$$

Ce qui nous donne:

$$\Delta\theta = \frac{2\pi}{\Omega} = \frac{2\pi}{\sqrt{1 - \frac{G^2 m^2 M^2}{b^2 c^2}}} \quad (47.193)$$

Étant donné que $G^2 \ll c^2$, un développement en série de Taylor (cf. chapitre sur les Suites Et

Séries):

$$(1+x)^{-\frac{1}{2}} = 1 - \frac{1}{2}x + \frac{3}{8}x^2 - \dots \quad (47.194)$$

En se limitant à l'ordre 2:

$$\Delta\theta = 2\pi \left(1 + \frac{G^2 m^2 M^2}{2b^2 c^2} \right) = 2\pi + \frac{\pi G^2 m^2 M^2}{b^2 c^2} \quad (47.195)$$

Donc en conclusion, il y a un avancement du périhélie s'effectuant dans le sens de rotation du satellite. Pour un référentiel situé dans le plan de rotation du satellite, la trajectoire est toujours une ellipse.

Cette avance est de:

$$\Delta\alpha = \frac{\pi G^2 m^2 M^2}{b^2 c^2} \quad (47.196)$$

par période. Soit en explicitant le moment cinétique donné pour rappel par:

$$b = r \cdot m \cdot v \quad (47.197)$$

Il vient alors après simplification:

$$\Delta\alpha = \frac{\pi G^2 M^2}{c^2 v^2 r^2} \quad (47.198)$$

Nous allons maintenant nous permettre une approximation assez grossière (mélange de relativiste et non relativiste). Soit à considérer la dernière relation, nous avons obtenu lors de nos développements des trajectoires d'orbitales Képlériennes la relation:

$$p = \frac{K^2}{\mu} = \frac{\|\vec{r} \times \vec{v}\|^2}{\mu} = \frac{b^2}{\mu m^2} = \frac{b^2}{GMm^2} \quad (47.199)$$

Dès lors en injectant ceci dans la relation de $\Delta\alpha$ nous avons:

$$\Delta\alpha = \frac{\pi G^2 m^2 M^2}{b^2 c^2} = \frac{\pi G^2 m^2 M^2}{p GMm^2 c^2} = \frac{\pi GM}{pc^2} \quad (47.200)$$

Malheureusement, les valeurs numériques pour Mercure ne donnent qu'une précession de 7" d'angle par siècle et non pas les 43" d'angle par siècle attendus (...) il manque un facteur 6 que seulement la relativité générale (cf. [chapitre de Relativité Générale](#)) permet de trouver. Il est néanmoins intéressant de constater que la relativité, même restreinte, donne déjà une orbite qui précède là où Newton voit une ellipse stable et que cette approximation fonctionne pour toutes les planètes exceptées Mercure (planète la plus proche du Soleil et subissant de plein fouet la courbure de l'espace-temps).

Remarque: En appliquant exactement le même raisonnement pour la physique quantique corpusculaire (potentiel électrique) mais avec les constantes ad hoc vues dans le chapitre d'Électrostatique, nous trouvons:

$$\Delta\alpha = \frac{\pi k^2 Z^2 e^4}{b^2 c^2} \quad (47.201)$$

avec $\vec{b} = \mu \vec{r} \times \vec{v}$ étant le moment cinétique et dans le cas de l'atome, nous prendrons (cf. chapitres Physique Quantique Corpusculaire):

$$b = \mu \gamma r^2 \dot{\theta} = n_{\theta} \hbar \quad (47.202)$$

avec la masse réduite valant:

$$\mu = \frac{mM}{m + M} \quad (47.203)$$

Si les positions du périhélie (et donc de l'aphélie) du barycentre Terre-Lune étaient constantes dans le temps, la durée des différentes saisons serait, elle aussi constante. Mais l'orbite du barycentre Terre-Lune tourne elle aussi dans son plan dans le sens direct à raison d'environ 12" par an (soit une révolution en environ 108'000 ans).

La précession des équinoxes s'effectue dans le sens contraire (sens rétrograde) à raison d'environ 50" par an (soit une révolution en environ 26'000 ans). La combinaison de ces deux mouvements permet de calculer la période du passage du périhélie de la Terre par la direction de l'équinoxe de printemps, cette période d'environ 21'000 ans est appelée précession climatique.

En effet, tous les 10'500 ans (demi-période de la précession climatique) l'aphélie passe de l'été à l'hiver. Or même si la distance Terre-Soleil n'est de loin pas le facteur prédominant dans la nature des saisons, la combinaison du passage de la Terre à l'aphélie en hiver donne des hivers un peu plus rudes. La distance Terre-Soleil dépend également de la variation de l'excentricité de l'orbite terrestre (due aux planètes extérieures et intérieures). Ainsi, les périodes glaciaires sont corrélées avec les minima de l'excentricité de l'orbite terrestre.

Les travaux de l'institut de mécanique céleste (France), depuis les années 1970, auraient permis de confirmer définitivement les prédictions théoriques comme quoi l'excentricité de l'orbite terrestre subit de larges variations formées de nombreux termes périodiques dont les plus importants ont des périodes voisines de 100'000 ans, et pour l'un d'eux, une période de 400'000 ans. Ces résultats confirment les variations climatiques de la Terre au cours de l'ère quaternaire. Les paléoclimatologies montrent en effet la corrélation entre les variations des éléments de l'orbite terrestre et les grandes glaciations du quaternaire.

Remarque: Dans le cas de l'atome d'hydrogène (voir le chapitre de Physique Quantique Corpusculaire traitant du modèle relativiste de Sommerfeld) avec:

$$n = 2 ; n_{\theta} = 1 ; z = 1$$

et la constante de structure fine égale approximativement à $\sim 1/137$, nous obtenons pour la précession du périhélie de l'orbite donnée:

$$\Delta\alpha = \pi \left(\frac{1}{137} \right)^2 = 1.67382 \cdot 10^{-4} \text{ [rad} \cdot \text{s}^{-1}] \quad (47.204)$$

selon un point de vue corpusculaire de la matière! (ce qui nous le savons n'est plus à l'ordre du jour).

DURÉE DE L'ARC DIURNE

Nous allons nous intéresser à la durée du jour, plus exactement à la portion de journée où nous sommes éclairés par le Soleil, par rapport à la nuit où nous nous trouvons dans l'ombre.

Remarque: Merci à Xavier Hubaut pour ces très sympathiques développements.

Dans la réalité, la Terre tourne autour du Soleil et décrit une orbite presque circulaire en même temps qu'elle tourne sur elle-même autour de son axe qui est incliné d'environ $23^{\circ}27'$ sur le plan de son orbite (l'écliptique).

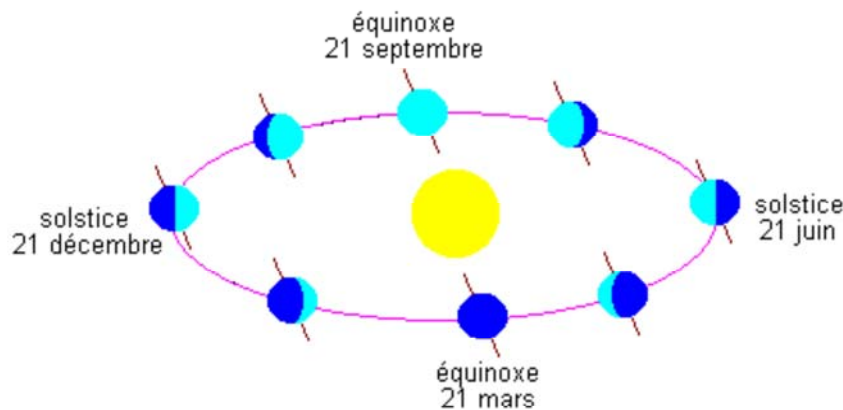


Figure: 47.7 - Représentation de la rotation de la Terre sur son orbite avec les phases majeures

Remarque: Il est évident qu'étant donnée la complexité du problème, nous le simplifierons en considérant une orbite circulaire, sans variations (précession, nutation) de l'axe de rotation de la Terre. Nous supposerons que le Soleil se réduit à un point (pas d'aurore, ni de crépuscule, etc.).

Rappelons que la précession est le changement graduel d'orientation de l'axe de rotation d'un objet quand un couple (de force) lui est appliqué alors que la nutation est un balancement périodique de l'axe de rotation de la Terre autour de sa position moyenne en plus de la précession.

Représentons la Terre avec son axe de rotation vertical; en conséquence l'équateur sera situé dans un

Supposons que ce jour-là, la Terre soit dans une position telle que les rayons du soleil forment un angle α avec le plan de l'équateur (ou que réciproquement l'axe de la Terre forme un angle avec le plan de l'équateur). Remarquons que cet angle α sera toujours compris selon les mesures actuelles entre $-23^{\circ}27'$ et $+23^{\circ}27'$.

Nous chercherons donc la durée du jour à un endroit situé à une latitude λ ? Pour fixer les idées, plaçons-nous dans les environs de Bruxelles à 50° de latitude Nord.

Enfin, γ désignera l'angle $\widehat{MOS}$ (où O est donc le centre du disque généré par le parallèle de Bruxelles) qui sous-tend la partie éclairée par le Soleil et r désignera le rayon $SO = MO$.

Pour simplifier le problème, supposons que pendant 24 heures la Terre tourne sur elle-même sans modifier la position de son axe de rotation par rapport au Soleil.

L'angle γ peut se calculer en remarquant que OP vaut, en valeur absolue:

$$PO = SO \cdot \cos\left(\pi - \frac{\gamma}{2}\right) = MO \cdot \cos\left(\pi - \frac{\gamma}{2}\right) = r \cdot \cos\left(\pi - \frac{\gamma}{2}\right) \quad (47.205)$$

où r représente le rayon du parallèle de Bruxelles.

En utilisant les propriétés des fonctions trigonométriques (cf. chapitre de Trigonométrie), nous avons:

$$PO = -r \cdot \cos\left(\frac{\gamma}{2}\right) \quad (47.206)$$

Or, il nous faut encore injecter le paramètre α . Connaissant la latitude λ de Bruxelles, nous avons:

$$r = OJ = R \cdot \cos(\lambda) \quad (47.207)$$

où R est le rayon de la Terre.

Nous avons aussi:

$$CO = R \cdot \sin(\lambda) \quad (47.208)$$

et dans le triangle COP :

$$PO = CO \cdot \tan(\alpha) = R \cdot \sin(\lambda) \cdot \tan(\alpha) \quad (47.209)$$

Enfin, en comparant les valeurs obtenues pour PO , nous obtenons:

$$\left. \begin{array}{l} PO = R \cdot \sin(\lambda) \cdot \tan(\alpha) \\ PO = -r \cdot \cos\left(\frac{\gamma}{2}\right) \end{array} \right\} \Rightarrow R \cdot \sin(\lambda) \cdot \tan(\alpha) = -r \cdot \cos\left(\frac{\gamma}{2}\right) \quad (47.210)$$

et comme:

$$r = OJ = R \cdot \cos(\lambda) \quad (47.211)$$

Nous obtenons finalement:

$$R \cdot \sin(\lambda) \cdot \tan(\alpha) = -R \cdot \cos(\lambda) \cdot \cos\left(\frac{\gamma}{2}\right) \quad (47.212)$$

et donc:

$$\gamma = -2 \arccos\left(\tan(\lambda) \cdot \tan(\alpha)\right) \quad (47.213)$$

Aux équinoxes (c'est-à-dire quand l'équateur est confondu avec le plan de l'écliptique), nous avons $\alpha = 0$ et donc:

$$\gamma = -2 \arccos(\tan(\lambda) \cdot \tan(\alpha)) = -2 \arccos(\tan(\lambda) \cdot 0) = -2 \arccos(0) = -\pi \quad (47.214)$$

Or, comme nous l'avons spécifié au début, il faut prendre la valeur absolue donc:

$$\gamma = \pi \quad (47.215)$$

En d'autres termes, quelle que soit la latitude que nous prenons, l'angle formé par la zone de nuit est égal à l'angle formé par la zone de jour (les deux étant égaux à π).

Prenons maintenant le solstice d'été, lorsque $\alpha = 23^\circ 27'$ en considérant toujours la latitude de Bruxelles $\lambda = 50^\circ$, nous avons:

$$\gamma = -2 \arccos(\tan(50^\circ) \cdot \tan(23^\circ 27')) = -2 \arccos(0.51695) \cong -118^\circ \quad (47.216)$$

ce qui, traduit en nombre d'heures:

$$-118^\circ \cdot \frac{24}{360} \cong -7.9 \quad (47.217)$$

Donc la journée de 24 heures perd 7.9 heures. Ce qui équivaut donc à une journée d'environ 16 heures.

En résumé pour calculer la durée du jour, il suffit de connaître deux choses: la latitude du lieu et l'angle selon lequel le soleil tombe sur le plan de l'équateur à la date choisie. La valeur de cet angle est bien connue aux équinoxes (il vaut 0°) et aux solstices (il vaut $+23^\circ 27'$ et $-23^\circ 27'$).

Mais aux autres dates?

La réponse est fort simple. Imaginons-nous, assis sur le Soleil regardant tout au long de l'année en direction du centre de la Terre.

Au cours de sa rotation autour du Soleil, l'axe de rotation de la Terre conserve son inclinaison sur l'écliptique. Vu du Soleil, cet axe tournera autour d'une normale au plan de l'écliptique et décrira donc un cône dont le demi-angle au sommet vaut $23^\circ 27'$ (voir figure plus bas).

L'angle d'attaque α des rayons solaires sur le plan de l'équateur variera donc en fonction de la date δ (nous associons à la date, l'angle δ parcouru par la Terre sur son orbite, à partir de sa position à l'équinoxe de printemps)

Par conséquent, l'angle α variera en fonction de la date δ de manière sinusoïdale.

Pour ceux qui ne seraient pas convaincus par ce raisonnement semi-intuitif, voici une autre approche:

Pour la lisibilité du schéma, nous avons fortement exagéré l'angle formé par l'axe de rotation de la Terre avec l'écliptique.

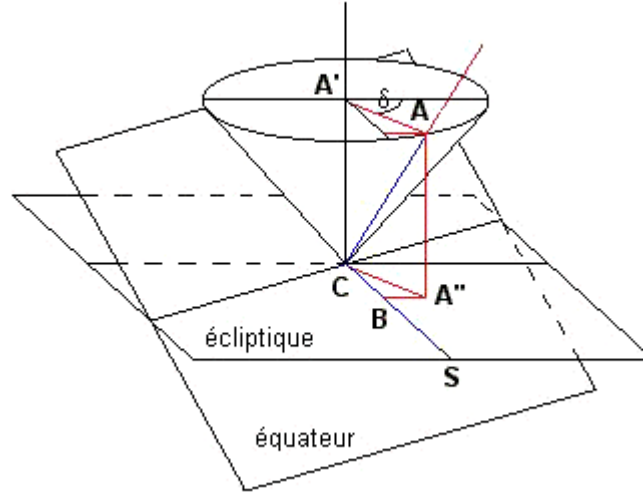


Figure: 47.9 - Cône généré par l'axe de rotation de la Terre (précession)

Soit C le centre de la Terre, A l'extrémité d'un vecteur unité $\overrightarrow{CA}$ dirigé suivant l'axe de rotation de la Terre (soit perpendiculaire au plan de l'équateur) et $\overrightarrow{CS}$ un autre vecteur unité dirigé vers le Soleil. Soit maintenant α l'angle du rayon CS avec le plan de l'équateur et β l'angle entre les vecteurs unitaires $\overrightarrow{CA}$ et $\overrightarrow{CS}$. Nous avons alors:

$$\overrightarrow{CS} \circ \overrightarrow{CA} = \underbrace{\|\overrightarrow{CS}\|}_{=1} \underbrace{\|\overrightarrow{CA}\|}_{=1} \cos(\beta) = \underbrace{\|\overrightarrow{CS}\|}_{=1} \underbrace{\|\overrightarrow{CA}\|}_{=1} \cos\left(\frac{\pi}{2} - \alpha\right) = \sin(\alpha) \quad (47.218)$$

Effectivement, le vecteur $\overrightarrow{CA}$ étant perpendiculaire au plan de l'équateur, il forme un angle droit avec celui-ci. Dès lors puisque l'angle β est l'angle entre ce vecteur et l'écliptique α en est le complémentaire.

Nous avons donc:

$$\overrightarrow{CS} \circ \overrightarrow{CA} = \sin(\alpha) \quad (47.219)$$

Décomposons maintenant $\overrightarrow{CA}$ en la somme de $\overrightarrow{CA}'$ dirigé perpendiculairement au plan de l'écliptique et de $\overrightarrow{CA}''$ situé dans le plan de l'écliptique:

$$\begin{aligned} \overrightarrow{CS} \circ \overrightarrow{CA} &= \underbrace{\overrightarrow{CS} \circ \overrightarrow{CA}'}_{\overrightarrow{CS} \perp \overrightarrow{CA}' = 0} + \overrightarrow{CS} \circ \overrightarrow{CA}'' = \overrightarrow{CS} \circ \overrightarrow{CA}'' = \underbrace{\|\overrightarrow{CS}\|}_{=1} \underbrace{\|\overrightarrow{CA}''\|}_{=1} \cdot \cos(\widehat{BCA}'') \\ &= \underbrace{\|\overrightarrow{CA}''\|}_{=1} \cdot \cos(\widehat{BCA}'') = \underbrace{\|\overrightarrow{CA}\|}_{=1} \cdot \cos(\beta) \cos(\widehat{BCA}'') = \cos\left(\frac{\pi}{2} - \alpha\right) \cdot \cos(\widehat{BCA}'') \quad (47.220) \\ &= \sin(\alpha) \cos(\widehat{BCA}'') \end{aligned}$$

Ainsi:

$$\overrightarrow{CS} \circ \overrightarrow{CA} = \sin(\alpha) \cos(\widehat{BCA}'') = \sin(23^\circ 27') \cos(\widehat{BCA}'') \quad (47.221)$$

Mais:

$$\cos(\widehat{BCA}) = \cos\left(\frac{\pi}{2} - \delta\right) = \sin(\delta) \quad (47.222)$$

Donc finalement:

$$\overrightarrow{CS} \circ \overrightarrow{CA} = \sin(23^{\circ}27') \sin(\delta) \quad (47.223)$$

et comme nous avons démontré que:

$$\overrightarrow{CS} \circ \overrightarrow{CA} = \sin(\alpha) \quad (47.224)$$

Nous obtenons finalement:

$$\sin(\alpha) = \sin(23^{\circ}27') \sin(\delta) = 0.3979 \cdot \sin(\delta) \quad (47.225)$$

À présent le problème est résolu et la durée du jour sera fonction de deux variables: la date δ et la latitude λ .

Il nous suffit donc maintenant de reprendre la relation:

$$\gamma = -2 \arccos(\tan(\lambda) \cdot \tan(\alpha)) \quad (47.226)$$

et d'y injecter le nouveau résultat:

$$\gamma = \left| -2 \arccos(\tan(\lambda) \cdot \tan(\arcsin(\sin(23^{\circ}27') \sin(\delta)))) \right| \quad (47.227)$$

Avec les outils informatiques à notre disposition, nous pouvons aisément calculer la valeur de γ . Nous avons par exemple ci-dessous les variations de la durée du jour sur une année à des latitudes allant de 0 à 90° réparties de 10 en 10°

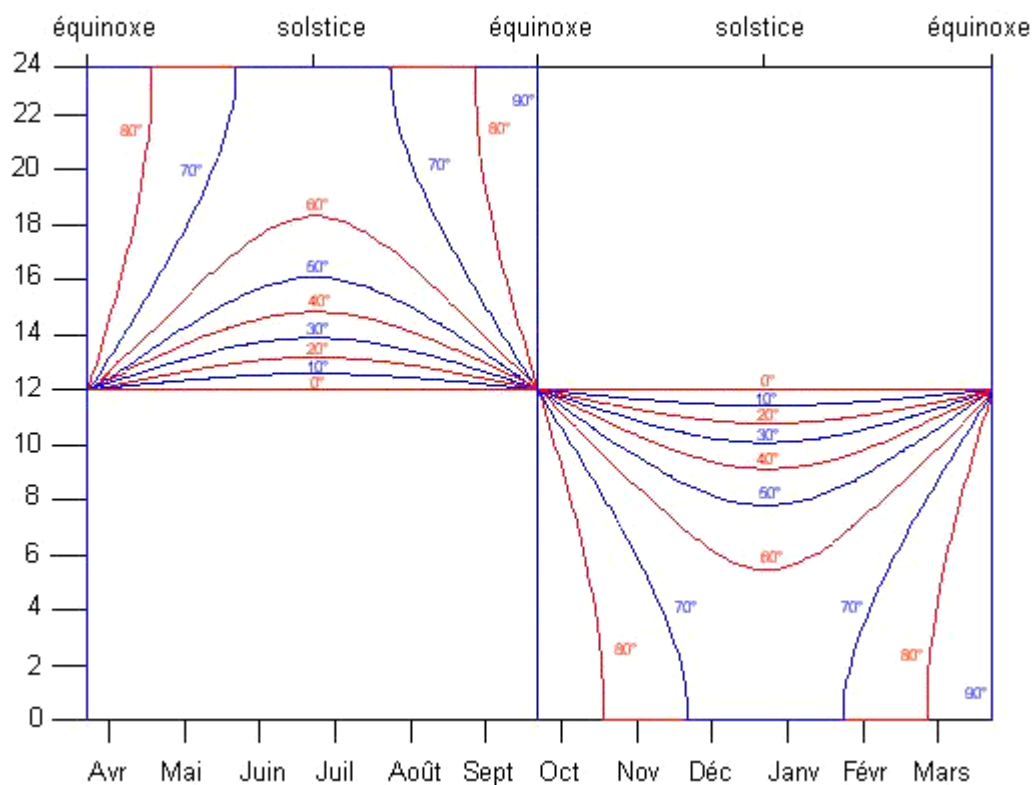


Figure: 47.10 - Angle de la partie éclairée de la Terre en fonction de la période de l'année

À partir de la latitude du cercle polaire, nous observons, en été, des périodes avec soleil ininterrompu (soleil de minuit) et, en hiver, des journées entières de nuit.

Pour Bruxelles (latitude=50°) nous voyons sur le graphique que la durée du jour varie approximativement entre les valeurs de 16h (solstice d'été) et 8h (solstice d'hiver).

MOUVEMENTS DES PLANÈTES

Nous allons brièvement nous intéresser aux mouvements des planètes en situations idéales et simplifiées. Nous ne considérerons que des mouvements dans le même plan (coplanaires) parfaitement circulaires et constants.

Définition: Les planètes qui sont plus proches du Soleil que la Terre (dont le rayon est inférieur à une unité astronomique), c'est-à-dire les planètes Mercure et Vénus, sont les "**planètes inférieures**". Les autres planètes (Mars et au-delà) sont appelées les "**planètes supérieures**".

PÉRIODE SYNODIQUE ET SIDÉRALE

La "**période synodique**" d'une planète est le temps mis par cette planète pour revenir à la même configuration Terre-planète-Soleil (si nous considérons ce cas particulier), c'est-à-dire à la même place dans le ciel par rapport au Soleil, vu de la Terre. Cette durée diffère de la période de révolution sidérale de la planète, car la Terre elle-même se déplace autour du Soleil. En conséquence, il s'agit de la période de révolution apparente, la durée entre deux conjonctions planète-Soleil, telle qu'observée depuis la Terre.

Le terme désigne de façon générale le temps séparant deux aspects identiques de l'objet (opposition, conjonction, quadrature, etc.) et dépend donc des trois corps impliqués.

Pour étudier mathématiquement le problème en question, considérons le schéma suivant avec deux

planètes décrivant une orbite parfaitement circulaire à vitesse angulaire constante et dans le même plan et dans le même sens et où nous avons $\omega_1 > \omega_2$ (donc la planète intérieure va plus vite que la planète extérieure):

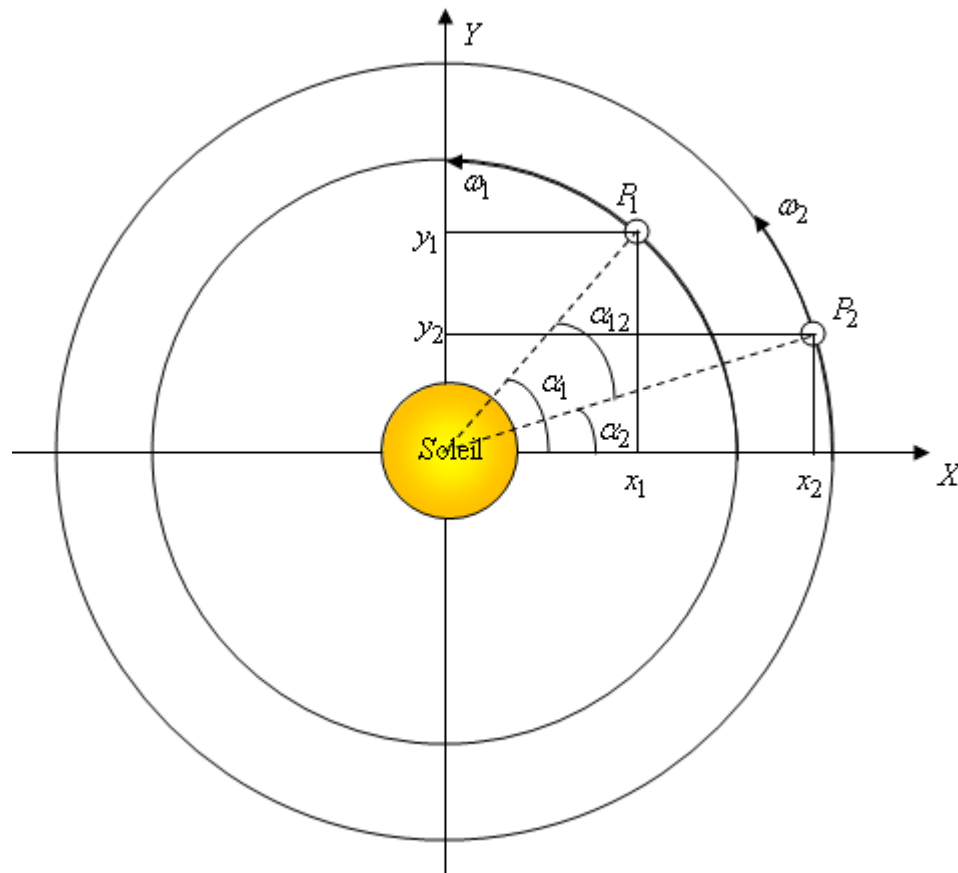


Figure: 47.11 - Schéma de base pour la détermination de la période synodique

où P_1 et P_2 sont deux planètes quelconques dont nous noterons les périodes sidérales respectives par T_1, T_2 et dont nous déduisons les vitesses angulaires:

$$\omega_1 = \frac{2\pi}{T_1} \quad \omega_2 = \frac{2\pi}{T_2} \quad (47.228)$$

Si nous prenons comme temps zéro, le temps où les deux planètes sont toutes deux alignées sur l'axe X et du même côté de l'axe (donc en conjonction inférieure), alors l'angle formé entre ce même axe et chacune des planètes est:

$$\alpha_1 = \omega_1 t \quad \alpha_2 = \omega_2 t \quad (47.229)$$

Nous avons respectivement:

$$\alpha_{12} = \alpha_1 - \alpha_2 = (\omega_1 - \omega_2)t = \omega_{12}t \quad (47.230)$$

Nous cherchons donc tous les instants t où la relation suivante est satisfaite pour un α_{12} fixé:

$$\alpha_{12} = (\omega_1 - \omega_2)t \quad (47.231)$$

Il vient alors:

$$T = \frac{\alpha_{12}}{\omega_1 - \omega_2} = \frac{\alpha_{12}}{\frac{2\pi}{T_1} - \frac{2\pi}{T_2}} = \frac{T_1 T_2 \alpha_{12}}{2\pi(T_2 - T_1)} \quad (47.232)$$

Si nous recherchons à partir du temps zéro la première (prochaine) conjonction (supérieure), cela revient à poser que $\alpha_{12} = \pi$ et donc que:

$$T = \frac{T_1 T_2}{2(T_2 - T_1)} \quad (47.233)$$

Si nous recherchons à partir du temps zéro la première (prochaine) conjonction (inférieure), cela revient à poser que $\alpha_{12} = 2\pi$ et donc que:

$$T = \frac{T_1 T_2}{T_2 - T_1} \quad (47.234)$$

Dans le cas où $\omega_2 > \omega_1$ (typiquement la Terre et une des planètes extérieures), le même raisonnement nous amène à:

$$T = \frac{T_1 T_2}{T_1 - T_2} \quad (47.235)$$

Voici quelques périodes synodiques et sidérales des planètes du système solaire:

Planète	Période synodique [j.]	Période sidérale [j.]
Mercure	115.878	87.969
Vénus	583.921	224.709
Mars	779.964	686.960
Jupiter	398.861	4'335.355
Saturne	378.094	10'757.737
Uranus	369.654	30'708.160
Neptune	367.486	60'224.904

À la vue de ce tableau, nous pouvons faire plusieurs constats empiriques:

1. Pour les planètes intérieures: Plus nous nous approchons du Soleil, plus la période synodique est courte, en effet dans la formule établie, plus T_1 est petit plus T diminue. Ainsi, s'il existait une planète tournant très près du Soleil, les deux périodes sidérale et synodique seraient pratiquement égales.
2. Quand nous nous approchons de la Terre, la période augmente. S'il existait une planète voisine de la Terre, nous aurions alors T_1 voisin de T_2 et la période synodique deviendrait très grande.
3. Pour les planètes extérieures: La période synodique diminue quand la planète est plus loin de la Terre et tend vers la période sidérale terrestre de 365 jours. Nous voyons bien pour Neptune, si on découvrait une planète située encore plus loin sa période synodique s'approcherait encore plus de 365

jours.

RÉTROGRADATION DES PLANÈTES

La "**rétrogradation**" ou "**mouvement rétrograde**" d'une planète est un mouvement apparent de cette planète qui donne l'impression de s'arrêter dans sa course dans le "mouvement direct" pour repartir en marche arrière. Ce phénomène est la résultante de la différence entre la vitesse de révolution de la planète et celle de la Terre autour du Soleil.

L'exemple ci-dessous illustre à peu près ce qu'un observateur Terrestre (point jaune) peut observer en surveillant mois après mois, le mouvement apparent de Mars (point cyan):

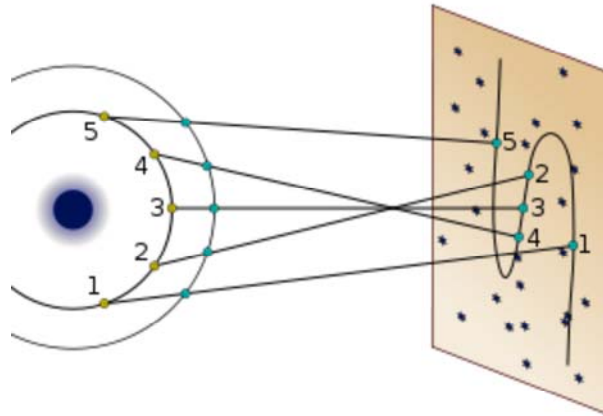


Figure: 47.12 - Principe de la rétrogradation (source: Wikipédia)

Pour étudier mathématiquement ce phénomène, nous allons considérer la situation suivante:

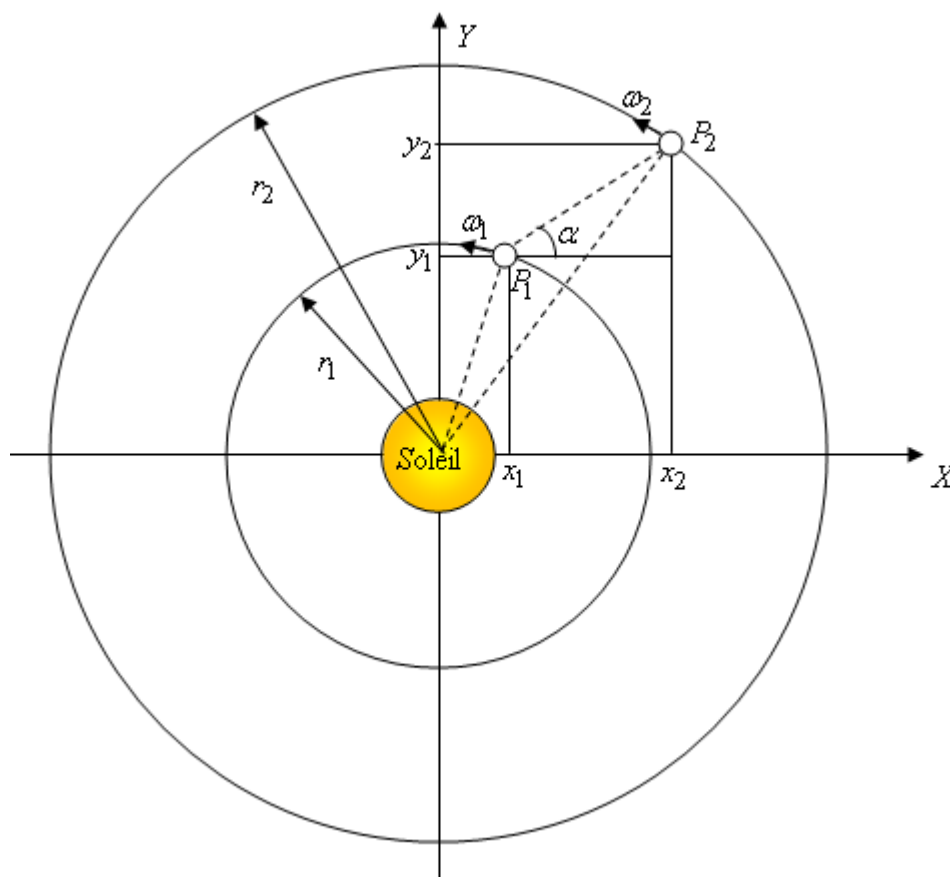


Figure: 47.13 - Schéma de base pour l'étude de la rétrogradation

avec deux planètes décrivant une orbite parfaitement circulaire à vitesse angulaire constante et dans le même plan et dans le même sens et où nous avons $\omega_1 > \omega_2$. Il est clair que la planète intérieure va dès lors rattraper la planète extérieure et elle va sembler avoir un mouvement rétrograde comme le montre la figure ci-dessous:

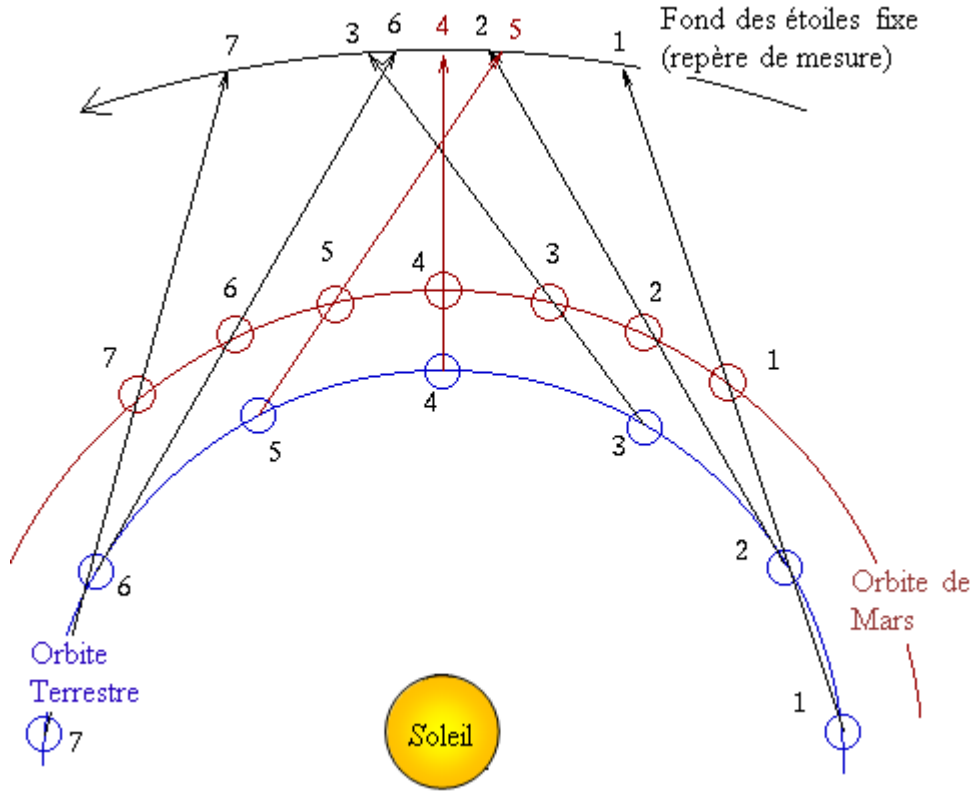


Figure: 47.14 - Schéma explicatif pour le choix du temps zéro

Comme le lecteur pourra le vérifier sur la figure ci-dessus nous observons que le mouvement rétrograde par rapport aux étoiles fixes commence lorsque l'angle entre les deux planètes est nul et qu'il finit lorsque l'angle entre les deux planètes passe par un maximum.

Dès lors, dans le schéma antérieur, nous avons:

$$\tan(\alpha) = \frac{y_2 - y_1}{x_2 - x_1} = \frac{r_2 \sin(\omega_2 t) - r_1 \sin(\omega_1 t)}{r_2 \cos(\omega_2 t) - r_1 \cos(\omega_1 t)} \quad (47.236)$$

Donc pour connaître le temps entre le moment où l'angle est nul entre les deux planètes, passe par un maximum et diminue à nouveau, il suffit de déterminer quand a lieu la variation de signe de la fonction précédente. Pour cela, il suffit de chercher quand la dérivée s'annule:

$$\frac{d}{dt} \left(\frac{r_2 \sin(\omega_2 t) - r_1 \sin(\omega_1 t)}{r_2 \cos(\omega_2 t) - r_1 \cos(\omega_1 t)} \right) = 0 \quad (47.237)$$

En appliquant les règles de dérivation vues dans le chapitre de Calcul Différentiel Et Intégral, nous avons:

$$\begin{aligned}
& \frac{d}{dt} \left(\left[r_2 \sin(\omega_2 t) - r_1 \sin(\omega_1 t) \right] \frac{1}{r_2 \cos(\omega_2 t) - r_1 \cos(\omega_1 t)} \right) \\
&= r_2 \omega_2 \cos(\omega_2 t) - r_1 \omega_1 \cos(\omega_1 t) \frac{1}{r_2 \cos(\omega_2 t) - r_1 \cos(\omega_1 t)} \\
&\quad - \frac{r_2 \sin(\omega_2 t) - r_1 \sin(\omega_1 t)}{(r_2 \cos(\omega_2 t) - r_1 \cos(\omega_1 t))^2} (-r_2 \omega_2 \sin(\omega_2 t) + r_1 \omega_1 \sin(\omega_1 t)) \\
&= \frac{(r_2 \omega_2 \cos(\omega_2 t) - r_1 \omega_1 \cos(\omega_1 t))(r_2 \cos(\omega_2 t) - r_1 \cos(\omega_1 t))}{(r_2 \cos(\omega_2 t) - r_1 \cos(\omega_1 t))^2} \\
&\quad - \frac{(r_2 \sin(\omega_2 t) - r_1 \sin(\omega_1 t))(-r_2 \omega_2 \sin(\omega_2 t) + r_1 \omega_1 \sin(\omega_1 t))}{(r_2 \cos(\omega_2 t) - r_1 \cos(\omega_1 t))^2} \\
&= 0
\end{aligned} \tag{47.238}$$

D'où après simplification:

$$\begin{aligned}
& (r_2 \omega_2 \cos(\omega_2 t) - r_1 \omega_1 \cos(\omega_1 t))(r_2 \cos(\omega_2 t) - r_1 \cos(\omega_1 t)) \\
& - (r_2 \sin(\omega_2 t) - r_1 \sin(\omega_1 t))(-r_2 \omega_2 \sin(\omega_2 t) + r_1 \omega_1 \sin(\omega_1 t)) = 0
\end{aligned} \tag{47.239}$$

Nous développons un peu cela:

$$\begin{aligned}
& r_2^2 \omega_2 \cos^2(\omega_2 t) - r_2 r_1 \omega_2 \cos(\omega_2 t) \cos(\omega_1 t) - r_1 r_2 \omega_1 \cos(\omega_1 t) \cos(\omega_2 t) + r_1^2 \omega_1 \cos^2(\omega_1 t) \\
& + r_2^2 \omega_2 \sin^2(\omega_2 t) - r_2 r_1 \omega_1 \sin(\omega_2 t) \sin(\omega_1 t) - r_1 r_2 \omega_2 \sin(\omega_1 t) \sin(\omega_2 t) + r_1^2 \omega_1 \sin^2(\omega_1 t) \\
& = 0
\end{aligned} \tag{47.240}$$

et nous simplifions un premier coup:

$$\begin{aligned}
& r_2^2 \omega_2 - r_2 r_1 \omega_2 \cos(\omega_2 t) \cos(\omega_1 t) - r_1 r_2 \omega_1 \cos(\omega_1 t) \cos(\omega_2 t) + r_1^2 \omega_1 \\
& - r_2 r_1 \omega_1 \sin(\omega_2 t) \sin(\omega_1 t) - r_1 r_2 \omega_2 \sin(\omega_1 t) \sin(\omega_2 t) \\
& = 0
\end{aligned} \tag{47.241}$$

et un deuxième:

$$r_2^2 \omega_2 + r_1^2 \omega_1 - r_1 r_2 \left[\omega_2 \cos(\omega_2 t) \cos(\omega_1 t) + \omega_1 \cos(\omega_1 t) \cos(\omega_2 t) + \omega_1 \sin(\omega_2 t) \sin(\omega_1 t) + \omega_2 \sin(\omega_1 t) \sin(\omega_2 t) \right] = 0 \tag{47.242}$$

et enfin une troisième:

$$r_2^2 \omega_2 + r_1^2 \omega_1 - r_1 r_2 (\omega_1 + \omega_2) [\cos(\omega_1 t) \cos(\omega_2 t) + \sin(\omega_1 t) \sin(\omega_2 t)] = 0 \tag{47.243}$$

et après réarrangement:

$$r_2^2 \omega_2 + r_1^2 \omega_1 = r_1 r_2 (\omega_1 + \omega_2) [\cos(\omega_1 t) \cos(\omega_2 t) + \sin(\omega_1 t) \sin(\omega_2 t)] \tag{47.244}$$

Nous simplifions en utilisant une des relations trigonométriques remarquables démontrées dans le chapitre de Trigonométrie:

$$r_2^2 \omega_2 + r_1^2 \omega_1 = r_1 r_2 (\omega_1 + \omega_2) \cos((\omega_1 - \omega_2)t) \quad (47.245)$$

Les valeurs de t qui satisfont cette relation nous donnent le changement de signe recherché.

Si t_0 est la première valeur de t qui vérifie l'équation, nous avons:

$$\cos((\omega_1 - \omega_2)t_0) = \frac{r_2^2 \omega_2 + r_1^2 \omega_1}{r_1 r_2 (\omega_1 + \omega_2)} \quad (47.246)$$

La valeur suivante de t sera telle que:

$$(\omega_1 - \omega_2)t_1 = 2\pi \pm (\omega_1 - \omega_2)t_0 \quad (47.247)$$

et donc:

$$t_1 = \frac{2\pi}{(\omega_1 - \omega_2)} \pm t_0 \quad (47.248)$$

Si nous faisons intervenir les périodes de rotation, nous avons:

$$t_1 = \frac{T_1 T_2}{T_2 - T_1} \pm t_0 \quad (47.249)$$

Pour en revenir à:

$$\cos((\omega_1 - \omega_2)t_0) = \frac{r_2^2 \omega_2 + r_1^2 \omega_1}{r_1 r_2 (\omega_1 + \omega_2)} \quad (47.250)$$

il peut être plus commode de l'écrire sous la forme traditionnelle suivante:

$$\begin{aligned} \cos\left(2\pi \frac{T_2 - T_1}{T_2 T_1} t_0\right) &= \frac{r_2^2 \frac{2\pi}{T_2} + r_1^2 \frac{2\pi}{T_1}}{r_1 r_2 \left(\frac{2\pi}{T_1} + \frac{2\pi}{T_2}\right)} = \frac{r_2^2 \frac{1}{T_2} + r_1^2 \frac{1}{T_1}}{r_1 r_2 \left(\frac{1}{T_1} + \frac{1}{T_2}\right)} = \frac{r_2^2 \frac{T_1}{T_2 T_1} + r_1^2 \frac{T_2}{T_1 T_2}}{r_1 r_2 \left(\frac{T_1 + T_2}{T_1 T_2}\right)} \\ &= \frac{\frac{r_2^2 T_1 + r_1^2 T_2}{T_1 T_2}}{r_1 r_2 \left(\frac{T_1 + T_2}{T_1 T_2}\right)} = \frac{r_2^2 T_1 + r_1^2 T_2}{r_1 r_2 (T_1 + T_2)} \end{aligned} \quad (47.251)$$

Jusqu'à maintenant, nous avons fait uniquement de la géométrie. Aucune loi de la gravitation n'est intervenue dans les calculs. Comme les rayons sont inconnus ou mal connus (historiquement parlant du moins), nous allons utiliser la troisième loi de Kepler (loi des périodes):

$$T^2 = c^{te} \cdot D^3 \quad (47.252)$$

où pour rappel, D est le demi-grand axe de l'orbite, et si celle-ci est circulaire, cela devient un simple rayon. Nous avons donc:

$$r = \frac{T^{2/3}}{\sqrt{c^{te}}} \quad (47.253)$$

Dès lors:

$$\cos \left(2\pi \frac{T_2 - T_1}{T_2 T_1} t_0 \right) = \frac{r_2^2 T_1 + r_1^2 T_2}{r_1 r_2 (T_1 + T_2)} = \frac{\frac{T_2^{4/3}}{c^{te}} T_1 + \frac{T_1^{4/3}}{c^{te}} T_2}{\frac{T_1^{2/3}}{\sqrt{c^{te}}} \frac{T_2^{2/3}}{\sqrt{c^{te}}} (T_1 + T_2)} = \frac{T_2^{4/3} T_1 + T_1^{4/3} T_2}{T_1^{2/3} T_2^{2/3} (T_1 + T_2)} \quad (47.254)$$

et donc:

$$t_0 = \frac{T_2 T_1}{2\pi (T_2 - T_1)} \arccos \left(\frac{T_2^{4/3} T_1 + T_1^{4/3} T_2}{T_1^{2/3} T_2^{2/3} (T_1 + T_2)} \right) \quad (47.255)$$

Une application numérique donne pour Mercure avec $T_1 \cong 87.95$ [j.] et pour la Terre $T_2 \cong 365.25$ [j.], nous obtenons:

$$t_0 \cong 11.44 \text{ [j.]} \quad (47.256)$$

Valeur que nous avons représentée sur le schéma ci-dessous:

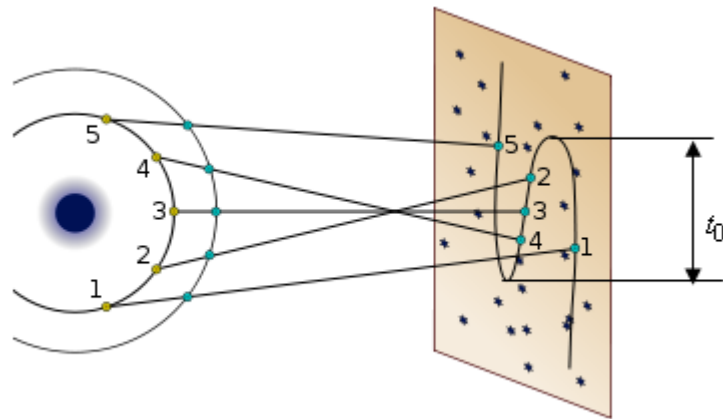


Figure: 47.15 - Représentation du temps calculé

et donc:

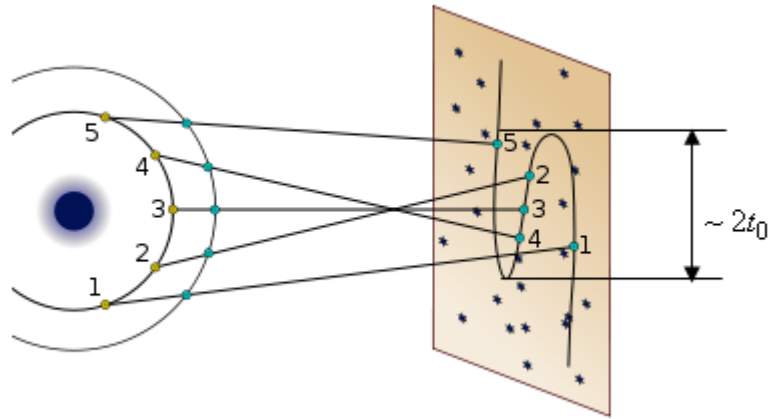


Figure: 47.16 - In extenso le temps calculé multiplié par deux

et nous avons donc:

$$t_1 = \frac{T_1 T_2}{T_2 - T_1} \pm t_0 \cong 115.84 \pm 11.44 \quad (47.257)$$

puis un nouveau cycle:

$$t_2 = \frac{T_1 T_2}{T_2 - T_1} \pm t_1 \quad (47.258)$$

etc. Ce qui donne sous forme schématique:

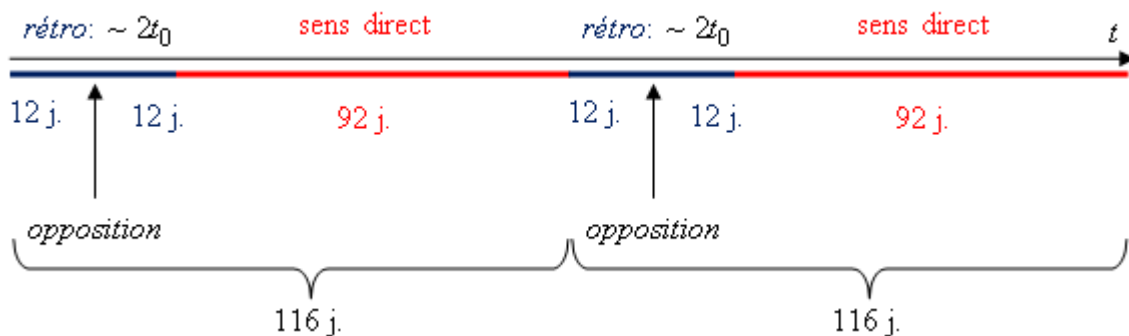


Figure: 47.17 - Schéma de principe du cycle

POINTS DE LAGRANGE

Un "**point de Lagrange**" (noté L), ou "**point de libration**", est une position de l'espace où les champs de gravité de deux corps en orbite l'un autour de l'autre, et de masses substantielles, se combinent de manière à fournir un point d'équilibre à un troisième corps de masse négligeable, tel que les positions relatives des trois corps soient fixes.

Nous allons dans les développements qui vont suivre nous attarder à démontrer au mieux que de tels points sont au nombre de 5 notés respectivement L_1 à L_5 .

Il peut être utile de faire une présentation de ces différents points et de leurs propriétés avant de passer à la partie calculatoire. Cela aidant peut-être à la compréhension du sujet.

Nous allons immédiatement considérer le schéma suivant:

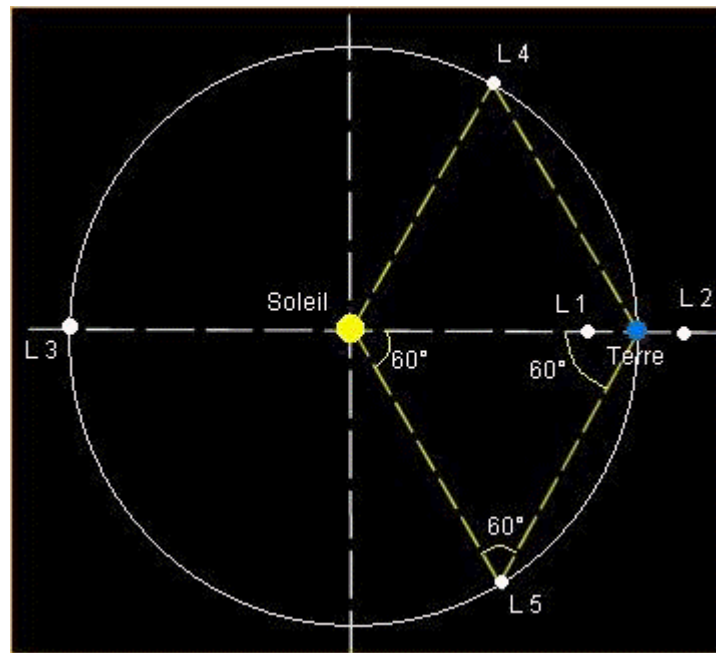


Figure: 47.18 - Représentation des cinq points de Lagrange du système Terre-Soleil

Il existe cinq points de Lagrange:

- *L1*: Sur la ligne définie par les deux masses, entre celles-ci.

Exemple:

Nous considérons un objet orbitant autour du Soleil, plus près de celui-ci que la Terre mais sur une même ligne. Cet objet subit une gravité solaire supérieure à celle de la Terre, et tourne donc plus rapidement autour du Soleil que ne le fait la Terre. Mais la gravité terrestre contrecarre en partie celle du Soleil, ce qui le ralentit. Plus on rapproche l'objet de la Terre, plus cet effet est important. À un certain point, le point *L1*, la vitesse angulaire de l'objet devient exactement égale à celle de la Terre.

- *L2*: Sur la ligne définie par les deux masses, au-delà de la plus petite.

Exemple:

Le principe est similaire au cas précédent, de l'autre côté de la Terre. L'objet devrait tourner moins vite que la Terre parce que la gravité solaire y est moindre, mais le champ gravitationnel supplémentaire dû à la Terre tend à l'accélérer. À un certain point, le point *L2*, l'objet tourne exactement à la même vitesse angulaire que la Terre autour du Soleil.

- *L3*: Sur la ligne définie par les deux masses, au-delà de la plus grande.

Exemple:

De manière identique au point *L2*, il existe un point situé un peu plus loin que l'opposé de la Terre par rapport au Soleil, où un objet de masse négligeable serait en équilibre.

- *L4* et *L5*: Sur les sommets des deux triangles équilatéraux dont la base est formée par les deux masses.

 Exemple:

Il s'agit d'un subtil équilibre entre la force centripète exercée par les deux masses principales et la force centrifuge des masses considérées aux points intéressés. L_4 est en avance sur la plus petite des masses, dans son orbite autour de la grande, et L_5 est en retard. Ces deux points sont parfois appelés "points de Lagrange triangulaires" ou "points Troyens".

Fait remarquable, ces deux derniers points ne dépendent en rien des masses relatives des deux autres corps comme nous le verrons.

Pour les trois premiers points de Lagrange, la stabilité n'apparaît que dans le plan perpendiculaire à la ligne occupée par les deux masses. Par exemple, pour le point L_1 , si nous déplaçons un objet perpendiculairement à la ligne entre les deux masses, les deux forces gravitationnelles vont jouer pour le ramener vers la position initiale. L'équilibre est stable. En revanche, si nous le déplaçons vers une des deux masses, alors le champ de celle-ci va l'emporter sur l'autre et l'objet tendra à se rapprocher encore plus. L'équilibre est instable. Pour les points L_4 et L_5 , la stabilité est obtenue grâce aux forces de Coriolis qui agissent sur les objets s'éloignant du point.

Étant données les questions de stabilité évoquées plus haut, nous ne trouvons pas d'objet naturel autour des points L_1 , L_2 et L_3 dans le système solaire. Cependant, ils représentent tout de même un intérêt pour les réalisations scientifiques, car ils permettent des économies de combustible pour le contrôle d'orbite et d'attitude. Ceci n'est pas valable pour le point L_3 , du fait de son éloignement de la Terre dont la seule application était que les auteurs de science-fiction et de bande dessinée aimaient y placer une Anti-Terre d'autant plus utopique que la masse de cette planète-jumelle y était bien trop élevée par rapport à la théorie énoncée plus haut. En revanche, des missions spatiales utilisent L_1 et L_2 : c'est le cas de la sonde SoHO (Solar and Heliospheric Observatory) une station d'observation du Soleil située au point L_1 .

L_4 et L_5 étant stables, nous y trouvons de nombreux corps naturels. Dans le système Jupiter-Soleil, plusieurs centaines d'astéroïdes, appelés astéroïdes Troyens, s'y agglutinent (près de 1800 en avril 2005). Nous en comptons quelques-uns dans les systèmes Neptune-Soleil et Mars-Soleil. Curieusement, il semblerait que le système Saturne-Soleil ne soit pas en mesure d'en accumuler, à cause des perturbations joviennes. Nous trouvons également des objets à ces points dans le système Saturne-satellites de Saturne: Saturne-Téthys avec Télésto et Calypso aux points L_4 et L_5 , et Saturne-Dioné avec Hélène au point L_4 et Pollux au point L_5 . Dans le système Terre-Soleil, il n'y a pas d'objet connu de grande taille aux points Troyens, mais on y a découvert une légère surabondance de poussière en 1950. De légers nuages de poussière sont également présents pour le système Terre-Lune; cela a fait renoncer à y placer un télescope spatial comme le projet en avait été envisagé. Le satellite SoHO occupe depuis 1995 le point L_1 à 1.5 million de kilomètres de la Terre. En 2007 le point L_2 sera occupé par le satellite Planck chargé d'étudier le fond diffus cosmologique à 2.7 [°K].

À strictement parler, ces 5 points existent uniquement pour deux corps en rotation circulaire l'un autour de l'autre. Dès que l'orbite des deux corps est elliptique, ces points ne sont plus des points d'équilibre. En pratique, si l'orbite est faiblement elliptique, comme c'est le cas pour les planètes réelles, on peut trouver des orbites oscillantes stables ne s'écartant pas beaucoup des régions correspondant aux points de Lagrange.

Nous allons donc considérer dans l'espace un système isolé de deux corps A et B , de masse M_A et M_B , en interaction gravitationnelle. Ces deux corps sont en orbite l'un autour de l'autre, à la manière d'un système de deux étoiles (système binaire) ou d'un système planète-satellite (Saturne-Titan par exemple). Nous cherchons à déterminer s'il existe des positions d'équilibre par rapport au système des deux corps en rotation pour un troisième corps (de masse suffisamment faible pour ne pas perturber le mouvement du système des deux corps principaux).

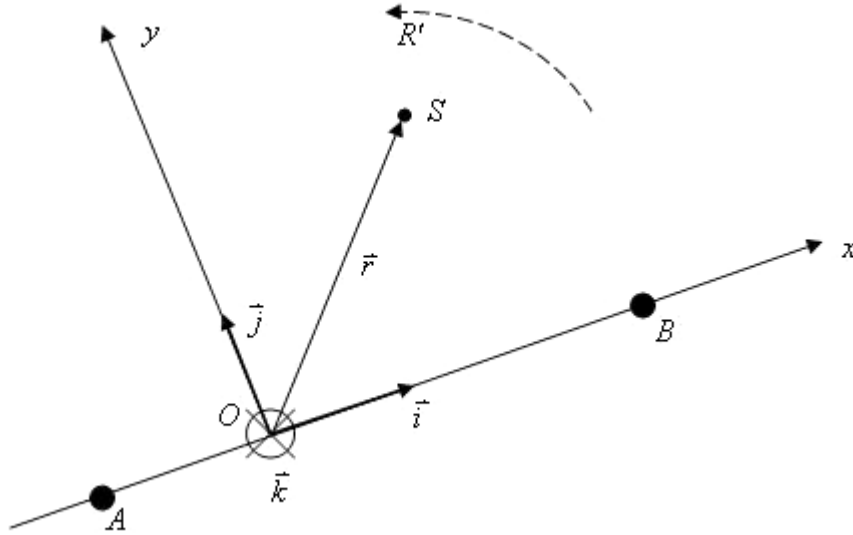


Figure: 47.19 - Configuration choisie pour l'étude mathématique

Soit O le barycentre (cf. chapitre de Mécanique Classique) de ces deux astres. Considérons un repère Galiléen (en mouvement rectiligne et uniforme donc!) d'origine O . Par rapport à ce repère, nous supposons que l'axe AB tourne à une vitesse angulaire constante ω d'axe fixe $\vec{k}$ (perpendiculaire à la page dans la figure ci-dessus et dirigé en direction du lecteur) et que les distances $r_A = \overline{AO}$ et $r_B = \overline{BO}$ restent également constantes.

Nous savons par notre étude de la mécanique classique que dans un mouvement circulaire la force centrifuge est donnée par:

$$F = ma = m \frac{v^2}{R} = m \frac{(\omega R)^2}{R} = m\omega^2 R \quad (47.259)$$

Nous avons donc (équilibre entre forces centrifuge et centripète) pour assurer l'équilibre:

$$M_A \omega^2 r_A = \frac{GM_A M_B}{(r_A + r_B)^2} \text{ et } M_B \omega^2 r_B = \frac{GM_A M_B}{(r_A + r_B)^2} \quad (47.260)$$

En simplifiant et en sommant ces deux relations:

$$\begin{aligned} \omega^2 r_A + \omega^2 r_B &= \frac{GM_A}{(r_A + r_B)^2} + \frac{GM_B}{(r_A + r_B)^2} \\ \Rightarrow \omega^2 &= \frac{G(M_A + M_B)}{(r_A + r_B)^3} = \frac{G(M_A + M_B)}{R^3} \end{aligned} \quad (47.261)$$

avec dans la suite $AB = r_A + r_B = R$.

Considérons un repère tournant R' lié à nos astres comme représenté sur la figure ci-dessus: $\vec{i}$ sera un vecteur unitaire colinéaire à AB , $\vec{j}$ un vecteur unitaire perpendiculaire à $\vec{i}$ et dans le plan de

rotation des planètes et finalement $\vec{k} = \vec{i} \times \vec{j}$ colinéaire à $\vec{\omega}$.

Nous considérons dans ce repère tournant (avec les astres) un troisième astre S de masse négligeable m devant M_A et M_B , soumis à l'attraction gravitationnelle de A et B .

Maintenant notons $\vec{a}_{R'}$ l'accélération de S par rapport à R' , $\vec{v}_{R'}$ sa vitesse et $\vec{e}_r$ le vecteur unitaire colinéaire à $\overrightarrow{OS'}$ où S' est le projeté de S dans le plan Oxy , et $r = OS'$ (dans la figure ci-dessus, nous avons supposé S dans le plan Oxy , et donc S et S' sont confondus).

S est donc soumis à deux forces, l'une $\vec{F}_A$ dirigée vers A et l'autre $\vec{F}_B$ dirigée vers B , forces d'intensités respectives:

$$\vec{F}_A = -\frac{GM_A m}{AS^3} \overrightarrow{AS} \quad \text{et} \quad \vec{F}_B = -\frac{GM_B m}{BS^3} \overrightarrow{BS} \quad (47.262)$$

Dans un repère Galiléen, ces deux forces imposent à S une accélération donnée par la loi de composition des accélérations dans un référentiel circulaire (cf. chapitre de Mécanique Classique):

$$\frac{1}{m}(\vec{F}_A + \vec{F}_B) = \vec{a}_a = \vec{a}_e + \vec{a}_r + 2\vec{\omega} \times \vec{v}_r + \frac{d\vec{\omega}}{dt} \times \vec{r} + \vec{\omega} \times (\vec{\omega} \times \vec{r}) \quad (47.263)$$

Or, dans notre configuration la pulsation est constante et l'accélération d'entraînement est nulle puisque nous avons posé R' comme référentiel principal. Il vient donc:

$$\frac{1}{m}(\vec{F}_A + \vec{F}_B) = \vec{a}_r + 2\vec{\omega} \times \vec{v}_r + \vec{\omega} \times (\vec{\omega} \times \vec{r}) = \vec{a}_{R'} + 2\vec{\omega} \times \vec{v}_r + \vec{\omega} \times (\vec{\omega} \times \vec{r}) \quad (47.264)$$

Nous avons également:

$$\vec{\omega} = \begin{pmatrix} 0 \\ 0 \\ \omega_z \end{pmatrix}, \vec{r} = \begin{pmatrix} r_x \\ r_y \\ 0 \end{pmatrix} \quad (47.265)$$

où selon schéma toutes les composantes sont positives. Le calcul du produit vectoriel donne (cf. chapitre de Calcul Vectoriel):

$$\vec{\omega} \times (\vec{\omega} \times \vec{r}) = \begin{pmatrix} -\omega^2 r_x \\ -\omega^2 r_y \\ 0 \end{pmatrix} = -\omega^2 \begin{pmatrix} r_x \\ r_y \\ 0 \end{pmatrix} = -\omega^2 \|\vec{r}\| \frac{\vec{r}}{\|\vec{r}\|} = -\omega^2 r \vec{e}_r \quad (47.266)$$

Donc finalement:

$$\frac{1}{m}(\vec{F}_A + \vec{F}_B) = \vec{a}_{R'} + 2\vec{\omega} \times \vec{v}_r + \vec{\omega} \times (\vec{\omega} \times \vec{r}) = \vec{a}_{R'} + 2\vec{\omega} \times \vec{v}_r - \omega^2 r \vec{e}_r \quad (47.267)$$

Écrivons plutôt cette relation sous la forme:

$$\vec{a}_R = \frac{1}{m} (\vec{F}_A + \vec{F}_B) + \omega^2 r \vec{e}_r - 2\vec{\omega} \times \vec{v}_r \quad (47.268)$$

Nous obtenons alors, en projetant sur les trois axes x, y et z , les dérivées prises par rapport au temps t le système suivant:

$$\begin{aligned} x'' &= -\frac{GM_A}{AS^3} (x+r_A) - \frac{GM_B}{BS^3} (x-r_B) + \omega^2 x - 2\omega y' \\ y'' &= -\frac{GM_A}{AS^3} y - \frac{GM_B}{BS^3} y + \omega^2 y - 2\omega x' \\ z'' &= -\frac{GM_A}{AS^3} z - \frac{GM_B}{BS^3} z \end{aligned} \quad (47.269)$$

avec:

$$AS = \sqrt{(x+r_A)^2 + y^2 + z^2} \quad \text{et} \quad BS = \sqrt{(x-r_B)^2 + y^2 + z^2} \quad (47.270)$$

pour que les coordonnées (x, y, z) du point S soient celles d'un point d'équilibre, il est bien évidemment que dans le référentiel tournant avec les astres A et B que:

$$x'' = y'' = z'' = 0 \quad \text{et} \quad x' = y' = z' = 0 \quad (47.271)$$

Nous obtenons alors le système suivant:

$$\begin{aligned} -\frac{GM_A}{AS^3} (x+r_A) - \frac{GM_B}{BS^3} (x-r_B) + \omega^2 x &= 0 \\ -\frac{GM_A}{AS^3} y - \frac{GM_B}{BS^3} y + \omega^2 y &= 0 \\ -\frac{GM_A}{AS^3} z - \frac{GM_B}{BS^3} z &= 0 \end{aligned} \quad (47.272)$$

Il vient par ailleurs immédiatement que la troisième équation a pour seule solution $z = 0$ et donc finalement le système se réduit à:

$$\begin{aligned} -\frac{GM_A}{AS^3} (x+r_A) - \frac{GM_B}{BS^3} (x-r_B) + \omega^2 x &= 0 \\ -\frac{GM_A}{AS^3} y - \frac{GM_B}{BS^3} y + \omega^2 y &= 0 \\ z &= 0 \end{aligned} \quad (47.273)$$

La troisième équation signifie simplement que les positions d'équilibre sont dans le plan Oxy (on pouvait s'en douter un peu...). Les deux autres, nous le verrons, nous amènent à considérer cinq solutions qui sont simplement nos cinq points de Lagrange $L1, \dots, L5$.

Si nous traçons avec un logiciel ad hoc l'accélération (respectivement la force) avec les isoclines mises en évidence (courbes sur lesquelles l'accélération a même norme) nous obtenons:

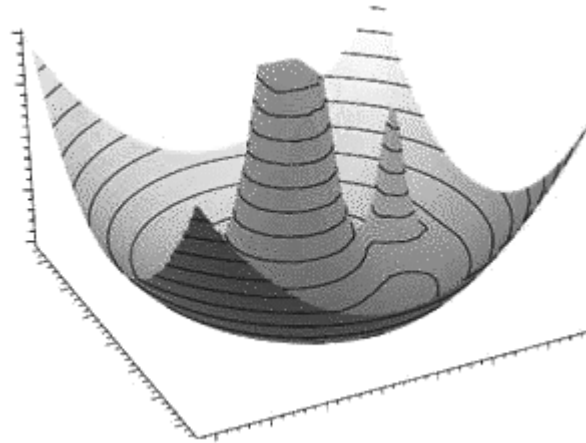


Figure: 47.20 - Isoclines du système à deux corps

et en demandant au logiciel de ne tracer que les isoclines projetées sur un plan:

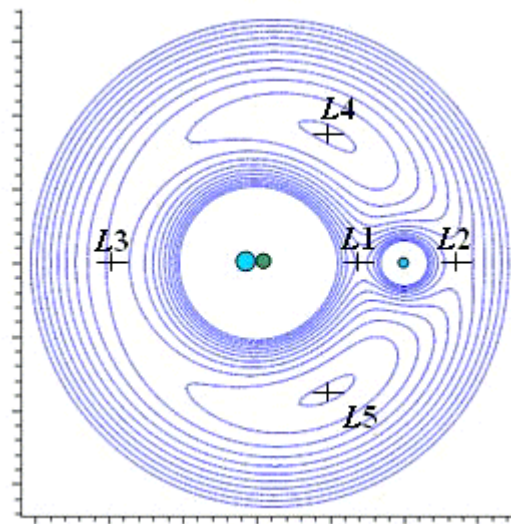


Figure: 47.21 - Projection des isoclines sur le plan

où nous avons mis en évidence les cinq points de Lagrange et où les astres sont représentés par des points bleus et le barycentre du système par un point vert.

Le lecteur remarquera qu'il est difficile de deviner intuitivement cette configuration du potentiel. Dans le référentiel tournant avec le barycentre des deux corps massifs, le potentiel résultant de la combinaison des potentiels gravitationnels et rotationnels présente 3 extrema $L1$, $L2$ et $L3$ sur la droite contenant les 2 corps. L'un de ces maxima se situe entre les 2 corps, ce que l'on attend intuitivement. Les deux autres maxima se trouvent sur la droite reliant les 2 objets, mais de part et d'autre ...ce qui est plus surprenant. Ils proviennent au fait de la contribution au potentiel du référentiel tournant ce qui peut être difficile à modéliser intuitivement.

POSITIONS D'ÉQUILIBRE DU PREMIER TYPE

Ce que nous entendons par les positions d'équilibre du premier type sont simplement les solutions situées sur la droite AB telles que $y = 0$ ce qui revient à étudier seulement:

$$-\frac{GM_A}{As^3}(x+r_A) - \frac{GM_B}{Bs^3}(x-r_B) + \omega^2 x = 0 \quad (47.274)$$

avec dès lors:

$$AS = |x + r_A| \text{ et } BS = |x - r_B| \quad (47.275)$$

À cette situation, nous allons considérer trois sous-cas possibles correspondants respectivement à $L1$, $L2$ et $L3$ comme nous allons de suite le voir.

POINT L1 DE LAGRANGE

Dans ce premier sous-cas, nous considérons:

$$-r_A < x < r_B \quad (47.276)$$

Ce qui revient aussi à avoir:

$$\begin{aligned} (x + r_A) &> 0 \\ (x - r_B) &< 0 \end{aligned} \quad (47.277)$$

Ce qui nous permet d'écrire:

$$-\frac{GM_A}{AS^3}(x + r_A) - \frac{GM_B}{BS^3}(x - r_B) + \omega^2 x = 0 \quad (47.278)$$

sous la forme simplifiée suivante:

$$-\frac{GM_A}{AS^2} + \frac{GM_B}{BS^2} + \omega^2 x = 0 \quad (47.279)$$

Maintenant pour dire quelque chose sur les solutions possibles de cette équation dérivons le membre de gauche. Nous obtenons alors:

$$\frac{2GM_A}{(x + r_A)^3} + \frac{2GM_B}{(r_B - x)^3} + \omega^2 \quad (47.280)$$

Ce terme est strictement croissant de $-\infty$ à $+\infty$ lorsque x décrit $] -r_A, r_B[$. Il y a donc une solution unique et un point d'équilibre noté $L1$ (premier point de Lagrange) entre A et B .

Si nous considérons typiquement le cas Soleil-Terre où $M_A > M_B$ et donc $r_A < r_B$ alors en $x = 0$ nous avons:

$$-\frac{GM_A}{(0 + r_A)^3}(0 + r_A) - \frac{GM_B}{(0 - r_B)^3}(0 - r_B) + \omega^2 0 = -\frac{GM_A}{r_A^2} - \frac{GM_B}{r_B^2} \quad (47.281)$$

ce qui est immédiatement négatif. La position d'équilibre sera donc obtenue pour une valeur positive de x que nous allons devoir déterminer.

Cette valeur peut être obtenue en considérant un cas limite: lorsque M_B tend vers 0 (correspondant à un astre massif A autour duquel tourne un astre B de masse beaucoup plus petit), A tend alors vers O ,

r_A vers 0 et donc:

$$\omega^2 = \frac{GM_A}{R^3} \quad (47.282)$$

avec $R = AB$. Dès lors, dans ce cas limite:

$$-\frac{GM_A}{AS^2} + \frac{GM_B}{BS^2} + \omega^2 x = 0 \quad (47.283)$$

devient en approximation:

$$-\frac{GM_A}{x^2} + \omega^2 x = 0 \quad (47.284)$$

et donc:

$$-\frac{GM_A}{x^2} + \left(\frac{GM_A}{R^3}\right)x = 0 \quad (47.285)$$

Donc la seule valeur de x satisfaisant cette relation sera $x = R$.

En d'autres termes, le point d'équilibre cherché $L1$ ici entre A et B se rapproche de B soit de l'astre le moins massif (ce qui correspond bien à la première figure que nous avons utilisé pour montrer l'emplacement des cinq points de Lagrange).

De par ce constat nous pouvons effectuer les calculs suivants:

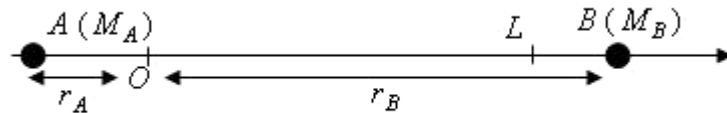


Figure: 47.22 - Configuration pour déterminer mathématiquement la position du point $L1$

Nous avons d'après la définition du barycentre:

$$\vec{r}_G = \frac{M_A \vec{r}_A + M_B \vec{r}_B}{M_A + M_B} \quad (47.286)$$

Comme notre étude se fait par rapport au barycentre nous avons $\vec{r}_G = \vec{0}$ et donc:

$$M_A \vec{r}_A + M_B \vec{r}_B = \vec{0} \quad (47.287)$$

De la relation précédente en prenant la norme, nous avons bien évidemment:

$$r_A = \frac{M_B}{M_A} r_B \quad (47.288)$$

La distance entre les deux astres A et B demeurant constante et égalant $r = r_A + r_B$ nous écrivons:

$$AB = r = r_A + r_B = \frac{M_B}{M_A} r_B + r_B = r_B \left(1 + \frac{M_B}{M_A} \right) = OB \left(1 + \frac{M_B}{M_A} \right) \quad (47.289)$$

Nous en déduisons trivialement deux relations (la deuxième étant obtenue par exactement le même raisonnement que la première):

$$OB = \frac{AB}{1 + \frac{M_B}{M_A}} \quad \text{et} \quad OA = \frac{AB}{1 + \frac{M_A}{M_B}} \quad (47.290)$$

Mais puisque $M_A \gg M_B$ nous pouvons grossièrement écrire la première relation sous la forme approximative suivante (série de Taylor):

$$OB \cong AB \left(1 - \frac{M_B}{M_A} \right) \quad (47.291)$$

et puisque:

$$r_A = \frac{M_B}{M_A} r_B \Leftrightarrow OA = \frac{M_B}{M_A} OB \quad (47.292)$$

nous avons aussi:

$$OA = \frac{M_B}{M_A} OB = \frac{M_B}{M_A} \frac{AB}{1 + \frac{M_B}{M_A}} \quad (47.293)$$

Donc avec $M_A \gg M_B$:

$$OA \cong \frac{M_B}{M_A} AB \quad (47.294)$$

Selon le cas limite étudié précédemment, nous pouvons donc supposer L voisin de B tel qu'abusivement il soit possible d'écrire:

$$OL \cong OB(1 - \varepsilon) \quad (47.295)$$

avec $\varepsilon \ll 1$.

Soit en utilisant:

$$OB \cong AB \left(1 - \frac{M_B}{M_A} \right) \quad (47.296)$$

Nous avons alors:

$$OL \cong AB(1 - \varepsilon) \left(1 - \frac{M_B}{M_A} \right) = AB \left(1 - \varepsilon - \frac{M_B}{M_A} \right) \quad (47.297)$$

en négligeant les infiniment petits d'ordre 2.

D'où:

$$\begin{aligned} LB &= OB - OL = AB \left(1 - \frac{M_B}{M_A} \right) - AB \left(1 - \varepsilon - \frac{M_B}{M_A} \right) = \varepsilon AB \\ LA &= OA + OL = \frac{M_B}{M_A} AB + AB \left(1 - \varepsilon - \frac{M_B}{M_A} \right) = AB(1 - \varepsilon) \end{aligned} \quad (47.298)$$

Maintenant dans la configuration mentionnée l'équilibre est donné par:

$$-\frac{GM_A}{LA^2} + \frac{GM_B}{LB^2} + \omega^2 OL = 0 \quad (47.299)$$

Donc:

$$-\frac{GM_A}{AB^2(1 - \varepsilon)^2} + \frac{GM_B}{\varepsilon^2 AB^2} + \omega^2 AB \left(1 - \varepsilon - \frac{M_B}{M_A} \right) = 0 \quad (47.300)$$

Maintenant la troisième loi de Kepler (cf. chapitre de Mécanique Classique) nous donne:

$$\frac{T^2}{AB^3} = \frac{4\pi^2}{GM_A} \Rightarrow \frac{GM_A}{AB^3} = \frac{4\pi^2}{T^2} = \left(\frac{2\pi}{T} \right)^2 = \left(\frac{2\pi}{\frac{2\pi}{\omega}} \right)^2 = \omega^2 \quad (47.301)$$

Soit:

$$-\frac{GM_A}{AB^2(1 - \varepsilon)^2} + \frac{GM_B}{\varepsilon^2 AB^2} + G \frac{M_A}{AB^3} AB \left(1 - \varepsilon - \frac{M_B}{M_A} \right) = 0 \quad (47.302)$$

Après simplification:

$$\begin{aligned} -\frac{M_A}{(1 - \varepsilon)^2} + \frac{M_B}{\varepsilon^2} + M_A \left(1 - \varepsilon - \frac{M_B}{M_A} \right) &= 0 \\ -\frac{M_A}{M_B} \frac{1}{(1 - \varepsilon)^2} + \frac{1}{\varepsilon^2} + \frac{M_A}{M_B} \left(1 - \varepsilon - \frac{M_B}{M_A} \right) &= 0 \end{aligned} \quad (47.303)$$

Soit:

$$\frac{1}{\varepsilon^2} - \frac{M_A}{M_B} (1 + 2\varepsilon) + \frac{M_A}{M_B} \left(1 - \varepsilon - \frac{M_B}{M_A} \right) \cong 0 \quad (47.304)$$

Donc:

$$\frac{1}{\varepsilon^2} - 3 \frac{M_A}{M_B} \varepsilon - 1 \cong 0 \quad (47.305)$$

Puisque $1/\varepsilon^2$ est très supérieur à 1 et en admettant que $3M_A/M_B$ le soit aussi nous avons:

$$\frac{1}{\varepsilon^2} \cong 3 \frac{M_A}{M_B} \varepsilon \quad (47.306)$$

Soit finalement:

$$\varepsilon \cong \sqrt[3]{\frac{M_B}{3M_A}} \quad (47.307)$$

et donc:

$$LA = AB(1 - \varepsilon) = AB \left(1 - \sqrt[3]{\frac{M_B}{3M_A}} \right) = R \left(1 - \sqrt[3]{\frac{M_B}{3M_A}} \right) \quad (47.308)$$

Si nous prenons le A Soleil et B la Terre, alors:

$$\frac{M_B}{M_A} \cong 3 \cdot 10^{-6}, R = 150 \cdot 10^6 [km] \quad (47.309)$$

Nous trouvons que la distance LB vaut à peu près:

$$LB = R \sqrt[3]{\frac{M_B}{3M_A}} = 1'500'000 [km] \quad (47.310)$$

qui est le point L1 auquel a été placé le satellite SoHO.

Un cas particulier du point L1 à considérer est lorsque $M_A = M_B = M$, alors $r_A = r_B = r$, O est alors le milieu de AB. Nous avons alors:

$$\omega^2 = \frac{GM}{R^3} = \frac{GM}{(2r)^3} = \frac{GM}{8r^3} \quad (47.311)$$

Dès lors:

$$-\frac{GM_A}{AS^2} + \frac{GM_B}{BS^2} + \omega^2 x = 0 \quad (47.312)$$

devient:

$$\begin{aligned}
& -\frac{GM}{(r+x)^2} + \frac{GM}{(x-r)^2} + \omega^2 x = 0 \\
\Rightarrow & -\frac{1}{(r+x)^2} + \frac{1}{(x-r)^2} = -\frac{x}{8r^3}
\end{aligned} \tag{47.313}$$

Parmi les quatre racines évidentes de cette équation la seule solution acceptable est $x = 0$ pour satisfaire $-r_A < x < r_B$. En d'autres termes, le point d'équilibre situé entre deux astres de même masse n'est autre que le barycentre de ces deux astres.

POINT L2 DE LAGRANGE

Dans ce deuxième sous-cas, nous considérons:

$$x > r_B \tag{47.314}$$

Nous cherchons donc les points d'équilibre au-delà de B .

Dès lors nous avons:

$$-\frac{GM_A}{AS^3}(x+r_A) - \frac{GM_B}{BS^3}(x-r_B) + \omega^2 x = 0 \tag{47.315}$$

qui devient simplement:

$$-\frac{GM_A}{AS^2} - \frac{GM_B}{BS^2} + \omega^2 x = 0 \tag{47.316}$$

Le membre de gauche est une fonction strictement croissante de x de $-\infty$ à $+\infty$ lorsque x décrit $]r_B, +\infty[$. Il y a donc une solution unique, et un point d'équilibre au-delà de B . Ce point est noté $L2$.

Cette valeur peut être obtenue en considérant un cas limite: lorsque M_B tend vers 0 (correspondant à un astre massif A autour duquel tourne un astre B de masse beaucoup plus petite), A tend alors vers O , r_A vers 0 et donc:

$$\omega^2 = \frac{GM_A}{R^3} \tag{47.317}$$

avec $R = AB$. Dès lors, dans ce cas limite:

$$-\frac{GM_A}{AS^2} - \frac{GM_B}{BS^2} + \omega^2 x = 0 \tag{47.318}$$

devient en approximation:

$$-\frac{GM_A}{x^2} + \omega^2 x = 0 \tag{47.319}$$

et donc:

$$-\frac{GM_A}{x^2} + \left(\frac{GM_A}{R^3}\right)x = 0 \quad (47.320)$$

Donc la seule valeur de x satisfaisant cette relation sera $x = R$. Le point $L2$ finit donc par se confondre avec B .

Connaissant ce cas limite, faisons une étude plus détaillée. Considérons le schéma suivant relativement à notre situation limite précédente:

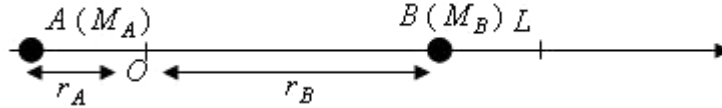


Figure: 47.23 - Configuration pour déterminer mathématiquement la position du point $L2$

et considérons $M_A \gg M_B$ sans oublier que dans ce scénario $x > r_B$

Nous avons alors quasiment les mêmes développements que pour $L1$ à la différence que:

$$OL \cong OB(1 - \varepsilon) \quad (47.321)$$

Devient:

$$OL \cong OB(1 + \varepsilon) \quad (47.322)$$

et que plutôt que d'avoir:

$$-\frac{GM_A}{LA^2} + \frac{GM_B}{LB^2} + \omega^2 OL = 0 \quad (47.323)$$

Nous avons:

$$-\frac{GM_A}{LA^2} - \frac{GM_B}{LB^2} + \omega^2 OL = 0 \quad (47.324)$$

et donc:

$$LA = AB(1 + \varepsilon) \quad (47.325)$$

Toujours avec:

$$\varepsilon \cong \sqrt[3]{\frac{M_B}{3M_A}} \quad (47.326)$$

et donc:

$$LA = AB(1 + \varepsilon) = AB \left(1 + \sqrt[3]{\frac{M_B}{3M_A}} \right) = R \left(1 + \sqrt[3]{\frac{M_B}{3M_A}} \right) \quad (47.327)$$

ce qui correspond au point de Lagrange L_2 .

Un cas particulier à nouveau de L_2 est lorsque $M_A = M_B = M$, alors $r_A = r_B = r$, O est alors le milieu de AB . Nous avons alors:

$$\omega^2 = \frac{GM}{R^3} = \frac{GM}{(2r)^3} = \frac{GM}{8r^3} \quad (47.328)$$

Dès lors:

$$-\frac{GM_A}{AS^2} - \frac{GM_B}{BS^2} + \omega^2 x = 0 \quad (47.329)$$

devient:

$$\begin{aligned} -\frac{GM}{(r+x)^2} - \frac{GM}{(x-r)^2} + \omega^2 x &= 0 \\ \Rightarrow \frac{1}{(r+x)^2} + \frac{1}{(x-r)^2} &= \frac{x}{8r^3} \end{aligned} \quad (47.330)$$

Il n'est plus possible d'extraire les racines ici. Il faut passer par une approximation numérique. Dans Maple 4.00b, il suffit de mettre:

```
>solve(-1/(r+x)^2-1/(x-r)^2=x/(8*r^3),x);allvalues("");
```

et la seule solution admissible dans $\mathbb{R}$ est $x \cong 2.8r$ les autres étant dans $\mathbb{C}$.

POINT L_3 DE LAGRANGE

Dans ce troisième sous-cas, nous considérons:

$$x < -r_A \quad (47.331)$$

Nous cherchons donc les points d'équilibre au-delà de A .

Dès lors nous avons:

$$-\frac{GM_A}{AS^3}(x+r_A) - \frac{GM_B}{BS^3}(x-r_B) + \omega^2 x = 0 \quad (47.332)$$

qui devient simplement:

$$-\frac{GM_A}{AS^2} - \frac{GM_B}{BS^2} + \omega^2 x = 0 \quad (47.333)$$

Le membre de gauche est une fonction strictement croissante de x de $-\infty$ à $+\infty$ lorsque x décrit $] -r_A, -\infty[$. Il y a donc une solution unique, et un point d'équilibre au-delà de A . Ce point est noté $L3$.

Cette valeur peut être obtenue en considérant un cas limite: lorsque M_B tend vers 0 (correspondant à un astre massif A autour duquel tourne un astre B de masse beaucoup plus petit), A tend alors vers O , r_A vers 0 et donc:

$$\omega^2 = \frac{GM_A}{R^3} \quad (47.334)$$

avec $R = AB$. Dès lors, dans ce cas limite:

$$-\frac{GM_A}{AS^2} - \frac{GM_B}{BS^2} + \omega^2 x = 0 \quad (47.335)$$

devient en approximation:

$$-\frac{GM_A}{x^2} + \omega^2 x = 0 \quad (47.336)$$

et donc:

$$-\frac{GM_A}{x^2} + \left(\frac{GM_A}{R^3} \right) x = 0 \quad (47.337)$$

Donc la seule valeur de x satisfaisant cette relation sera $x = R$. Le point $L3$ finit donc par se confondre avec la position diamétralement opposée à B .

Connaissant ce cas limite, faisons une étude plus détaillée. Considérons le schéma suivant relativement à notre situation limite précédente:

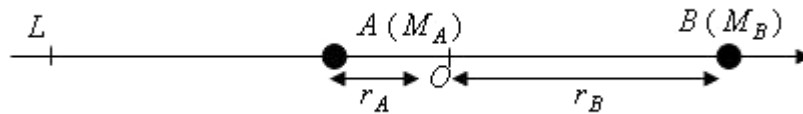


Figure: 47.24 - Configuration pour déterminer mathématiquement la position du point $L3$

et considérons toujours $M_A \gg M_B$ sans oublier que dans ce scénario $x < -r_A$

Nous allons considérer d'abord l'approximation suivante:

$$AL \cong AB(1 + \varepsilon) = R(1 + \varepsilon) \quad (47.338)$$

et celle-ci aussi (puisque OA tend vers zéro lorsque l'astre A devient très massif):

$$OA \cong \frac{M_B}{M_A} R \quad (47.339)$$

Dès lors:

$$OL = OA + AL \cong R \left(1 + \varepsilon + \frac{M_B}{M_A} \right) \quad (47.340)$$

Nous avons aussi (...):

$$\omega^2 = \frac{GM_A}{R^3} \cong \frac{GM_A}{R^3} \left(1 + \frac{M_B}{M_A} \right) \quad (47.341)$$

où à la limite où l'astre A est vraiment massif, nous retombons sur le premier terme...

Avec ces deux dernières relations, nous avons:

$$\omega^2 OL \cong R \left(1 + \varepsilon + \frac{M_B}{M_A} \right) \frac{GM_A}{R^3} \left(1 + \frac{M_B}{M_A} \right) \cong \frac{GM_A}{R^2} \left(1 + \varepsilon + 2 \frac{M_B}{M_A} \right) \quad (47.342)$$

si nous négligeons les termes du deuxième ordre.

Nous avons par ailleurs aussi:

$$BL = AB + AL = R + R(1 + \varepsilon) = 2R \left(1 + \frac{\varepsilon}{2} \right) \quad (47.343)$$

Rappelons la condition d'équilibre:

$$-\frac{GM_A}{AS^2} - \frac{GM_B}{BS^2} + \omega^2 x = -\frac{GM_A}{AL^2} - \frac{GM_B}{BL^2} + \omega^2 OL = 0 \quad (47.344)$$

Et mettons tout ce que nous avons obtenu avant là-dedans:

$$-\frac{GM_A}{\left(R(1 + \varepsilon) \right)^2} - \frac{GM_B}{\left(2R \left(1 + \frac{\varepsilon}{2} \right) \right)^2} + \frac{GM_A}{R^2} \left(1 + \varepsilon + 2 \frac{M_B}{M_A} \right) = 0 \quad (47.345)$$

Ce qui devient après simplifications:

$$\frac{M_A}{M_B} \frac{1}{(1 + \varepsilon)^2} + \frac{1}{4 \left(1 + \frac{\varepsilon}{2} \right)^2} = \frac{M_A}{M_B} \left(1 + \varepsilon + 2 \frac{M_B}{M_A} \right) \quad (47.346)$$

après une petite approximation:

$$4 \frac{M_A}{M_B} (1 - 2\varepsilon) + (1 - \varepsilon) = 4 \frac{M_A}{M_B} \left(1 + \varepsilon + 2 \frac{M_B}{M_A} \right) \quad (47.347)$$

après simplification:

$$-12\varepsilon \frac{M_A}{M_B} = 7 \quad (47.348)$$

D'où:

$$\varepsilon = -\frac{7}{12} \frac{M_B}{M_A} \quad (47.349)$$

et finalement:

$$OL \cong R \left(1 + \varepsilon + \frac{M_B}{M_A} \right) \cong R \left(1 - \frac{7}{12} \frac{M_B}{M_A} + \frac{M_B}{M_A} \right) = R \left(1 + \frac{5}{12} \frac{M_B}{M_A} \right) \quad (47.350)$$

Remarque: Chez certains auteurs de science-fiction, ce point L3 à l'opposé de la Terre par rapport au Soleil nous cacherait une hypothétique planète qui nous serait perpétuellement cachée par le Soleil.

POSITIONS D'ÉQUILIBRE DU DEUXIÈME TYPE

Les positions d'équilibre du deuxième type sont donc celles pour lesquelles $y \neq 0$. En d'autres termes les points situés hors de la droite AB, mais dans le plan Oxy.

Ainsi, notre système d'équations reste:

$$\begin{aligned} -\frac{GM_A}{AS^3}(x+r_A) - \frac{GM_B}{BS^3}(x-r_A) + \omega^2 x &= 0 \\ -\frac{GM_A}{AS^3}y - \frac{GM_B}{BS^3}y + \omega^2 y &= 0 \\ z &= 0 \end{aligned} \quad (47.351)$$

POINTS L4, L5 DE LAGRANGE

Pour déterminer les autres points d'équilibre restants, nous pouvons diviser la deuxième équation du système par y tel que le système devienne:

$$\begin{aligned} -\frac{GM_A}{AS^3}(x+r_A) - \frac{GM_B}{BS^3}(x-r_B) + \omega^2 x &= 0 \\ -\frac{GM_A}{AS^3} - \frac{GM_B}{BS^3} + \omega^2 &= 0 \end{aligned} \quad (47.352)$$

Retranchons de la première équation la deuxième multipliée par x. Nous obtenons alors pour la première:

$$-\frac{GM_A}{AS^3}r_A + \frac{GM_B}{BS^3}r_B = 0 \quad (47.353)$$

Soit:

$$\frac{M_A}{AS^3} r_A = \frac{M_B}{BS^3} r_B \quad (47.354)$$

Mais comme $r_A M_A = r_B M_B$ ceci se simplifie encore en:

$$\frac{1}{AS^3} = \frac{1}{BS^3} \quad (47.355)$$

Reprenons maintenant, en toute généralité, notre schéma du début en rajoutant quelques éléments:

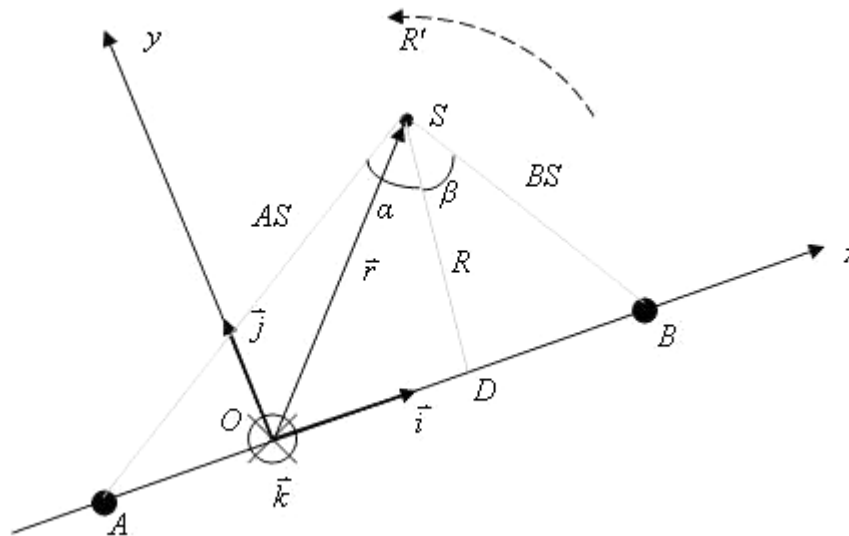


Figure: 47.25 - Configuration pour déterminer mathématiquement la position des points L4 et L5

où AB est la distance entre A et B et D est le centre de masse du système donné par:

$$AD = AB \frac{M_B}{M_A + M_B} \text{ ou } DB = AB \frac{M_A}{M_A + M_B} \quad (47.356)$$

qui sont donc les rayons de giration des corps A et B .

Il est facile de vérifier que la somme des deux distances précédentes est égale à AB et leur proportion M_B / M_A . Une autre forme de DB (qui nous sera utile) s'obtient en divisant numérateur et dénominateur par M_A :

$$DB = \frac{AB}{1 + \frac{M_B}{M_A}} \quad (47.357)$$

Nous savons selon nos calculs précédents que $AS = BS$ mais cela est insuffisant. Nous voulons encore connaître les angles des sommets A , B , S et c'est ce à quoi nous allons nous intéresser maintenant.

Dans ce cadre, si un satellite en S est en équilibre, il restera toujours à la même distance de A ou de B . Le centre de rotation des 3 points est le point D , la masse A elle-même tourne autour de lui. Si le

satellite, en S, reste stabilisé, les trois corps ont la même période orbitale T . Si S est immobile dans ce cadre en rotation il ne sera pas soumis à la force de Coriolis mais uniquement à la force centrifuge, aussi bien de celle de A que de B.

Notons v_B la vitesse de rotation de B et v_S la vitesse de rotation de S. Nous avons alors:

$$2\pi R = v_S \cdot T \text{ et } \frac{2\pi \frac{AB}{1 + \frac{M_B}{M_A}}}{1} = v_B \cdot T \quad (47.358)$$

Nous en tirons que:

$$\frac{2\pi}{T} = \frac{v_S}{R} \text{ et } \frac{2\pi}{T} = \frac{v_B}{AB} \cdot \left(1 + \frac{M_B}{M_A}\right) \quad (47.359)$$

Nous pouvons donc évaluer ces deux expressions:

$$\frac{v_S}{R} = \frac{v_B}{AB} \cdot \left(1 + \frac{M_B}{M_A}\right) \quad (47.360)$$

Cela exprime simplement le fait bien connu que si deux objets tournent conjointement, le plus éloigné de l'axe est le plus rapide. Les vitesses sont proportionnelles aux distances de l'axe.

La force centrifuge sur B est en équilibre avec la force gravitationnelle de A et cela s'exprime par:

$$M_B \frac{\frac{v_B^2}{AB}}{1 + \frac{M_B}{M_A}} = \frac{M_B}{AB} v_B^2 \left(1 + \frac{M_B}{M_A}\right) = G \frac{M_A M_B}{AB^2} \quad (47.361)$$

Soit en simplifiant:

$$G \frac{M_A}{AB} = v_B^2 \left(1 + \frac{M_B}{M_A}\right) \quad (47.362)$$

De même, la force centrifuge qui s'applique sur S est:

$$M_S \frac{v_S^2}{R} \quad (47.363)$$

Elle est équilibrée par les forces d'attraction $\vec{F}_A, \vec{F}_B$ des corps A et B. Néanmoins, seules les composantes de ces forces situées sur la ligne R s'opposent efficacement à cette force centrifuge. D'où:

$$M_S \frac{v_S^2}{R} = F_A \cos \alpha + F_B \cos \beta \quad (47.364)$$

et comme:

$$F_A = \frac{GM_A M_S}{AS^2} \text{ et } F_B = \frac{GM_B M_S}{BS^2} \quad (47.365)$$

Nous avons alors:

$$\frac{v_S^2}{R} = \frac{GM_A}{AS^2} \cos \alpha + \frac{GM_B}{BS^2} \cos \beta \quad (47.366)$$

En outre, les forces s'appliquant à S et perpendiculaires à R doivent s'annuler. Si non, le corps S suivrait la masse la plus importante et ne resterait pas en position et ne serait donc plus en équilibre. Il faut donc que:

$$F_A \sin \alpha = F_B \sin \beta \quad (47.367)$$

Soit, après substitution et simplification:

$$\frac{M_A}{AS^2} \sin \alpha = \frac{M_B}{BS^2} \sin \beta \quad (47.368)$$

De toutes les équations obtenues jusqu'à maintenant les seules qui nous dérangent sont celles contenant à la fois des vitesses et des angles α, β . Il faut donc que nous arrivions à éliminer ce qui convient pour n'avoir que les deux derniers paramètres (soit les angles).

Pour cela, nous portons au carré:

$$\frac{v_S^2}{R^2} = \frac{v_B^2}{AB^2} \cdot \left(1 + \frac{M_B}{M_A}\right)^2 \quad (47.369)$$

Nous multiplions des deux côtés par AB^2 et divisons par $1 + \frac{M_B}{M_A}$:

$$\frac{v_S^2 AB^2}{R^2 \left(1 + \frac{M_B}{M_A}\right)} = v_B^2 \left(1 + \frac{M_B}{M_A}\right) \quad (47.370)$$

qui est à rapprocher de:

$$G \frac{M_A}{AB} = v_B^2 \left(1 + \frac{M_B}{M_A}\right) \quad (47.371)$$

Donc en égalant:

$$\frac{v_S^2 AB^2}{R^2 \left(1 + \frac{M_B}{M_A}\right)} = G \frac{M_A}{AB} \quad (47.372)$$

Nous avons donc éliminé la vitesse de B . Maintenant, multiplions les deux côtés par $1 + \frac{M_B}{M_A}$ et divisons par AB^2 et multiplions par R :

$$\frac{v_S^2}{R} = \frac{GM_A}{AB^3} R \left(1 + \frac{M_B}{M_A} \right) \quad (47.373)$$

à rapprocher de:

$$\frac{v_S^2}{R} = \frac{GM_A}{AS^2} \cos \alpha + \frac{GM_B}{BS^2} \cos \beta \quad (47.374)$$

Donc:

$$\frac{GM_A}{AB^2} \frac{R}{AB} \left(1 + \frac{M_B}{M_A} \right) = \frac{GM_A}{AS^2} \cos \alpha + \frac{GM_B}{BS^2} \cos \beta \quad (47.375)$$

En divisant le tout par GM_A nous trouvons:

$$\frac{1}{AB^2} \frac{R}{AB} \left(1 + \frac{M_B}{M_A} \right) = \frac{1}{AS^2} \cos \alpha + \frac{M_B}{M_A} \frac{1}{BS^2} \cos \beta \quad (47.376)$$

Et comme nous avons démontré au début $AS = BS$ que nous noterons R' , nous avons alors:

$$\frac{R'^2}{AB^2} \frac{R}{AB} \left(1 + \frac{M_B}{M_A} \right) = \cos \alpha + \frac{M_B}{M_A} \cos \beta \quad (47.377)$$

et rappelons que nous avons:

$$\frac{M_A}{AS^2} \sin \alpha = \frac{M_B}{BS^2} \sin \beta \quad (47.378)$$

Soit:

$$\frac{M_B}{M_A} = \frac{\sin \alpha}{\sin \beta} \quad (47.379)$$

Ce qui nous permet d'écrire:

$$\frac{R'^2}{AB^2} \frac{R}{AB} \left(1 + \frac{M_B}{M_A} \right) = \cos \alpha + \frac{\sin \alpha}{\sin \beta} \cos \beta \quad (47.380)$$

En multipliant par $\sin \beta$:

$$\frac{R^2}{AB^2} \sin \beta \frac{R}{AB} \left(1 + \frac{M_B}{M_A} \right) = \sin \beta \cos \alpha + \sin \alpha \cos \beta = \sin(\alpha + \beta) \quad (47.381)$$

Soit:

$$\frac{R^2}{AB^2} \sin \beta \frac{1}{AB} \left(1 + \frac{M_B}{M_A} \right) = \frac{\sin(\alpha + \beta)}{R} \quad (47.382)$$

Nous pouvons maintenant remarquer une chose (faut le voir...). Si $R' = AB$ (soit que le triangle ABS est équilatéral) la relation précédente se simplifie en:

$$\sin \beta \frac{1}{AB} \left(1 + \frac{M_B}{M_A} \right) = \frac{\sin(\hat{S})}{R} \quad (47.383)$$

Or, si le triangle est bien équilatéral, nous avons alors $\sin(\hat{S}) = \sin(\hat{A}) = \sin(\hat{B})$. Dès lors:

$$\frac{\sin \beta}{AB \cdot \frac{M_A}{M_A + M_B}} = \frac{\sin(\hat{S})}{R} \quad (47.384)$$

Soit ce qui peut s'écrire finalement:

$$\frac{\sin \beta}{DB} = \frac{\sin(\hat{B})}{R} \quad (47.385)$$

Ce qui n'est autre que le théorème des sinus pour le triangle SDB (cf. chapitre de Trigonométrie) et est donc certain. En reprenant en arrière, nous pouvons maintenant prouver que toutes les équations précédentes sont satisfaites si et seulement si ABS est équilatéral. Si nous n'avions pas posé ABS comme équilatéral, nous aurions obtenu une relation différente du théorème des sinus, sans vérification possible, et l'ensemble des équations exigées pour l'équilibre au point S n'auraient pu être satisfaites.

Conclusion de la chose... le système donne comme solution:

$$AS = BS = R \quad (47.386)$$

ABS (ou ABL peu importe l'écriture), forme alors un triangle équilatéral. Les deux points d'équilibre sont notés $L4$ et $L5$. $L4$ est situé en avance par rapport à l'astre de masse la plus petite, et $L5$ en retard.

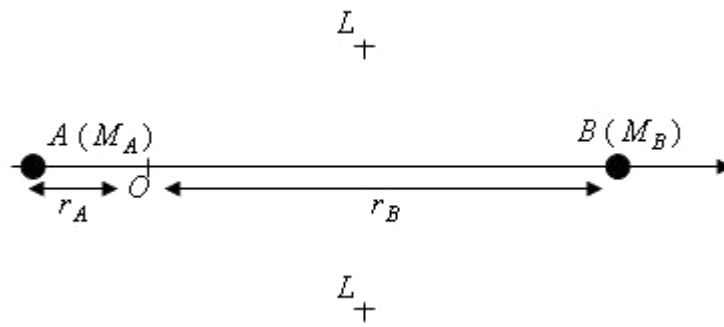
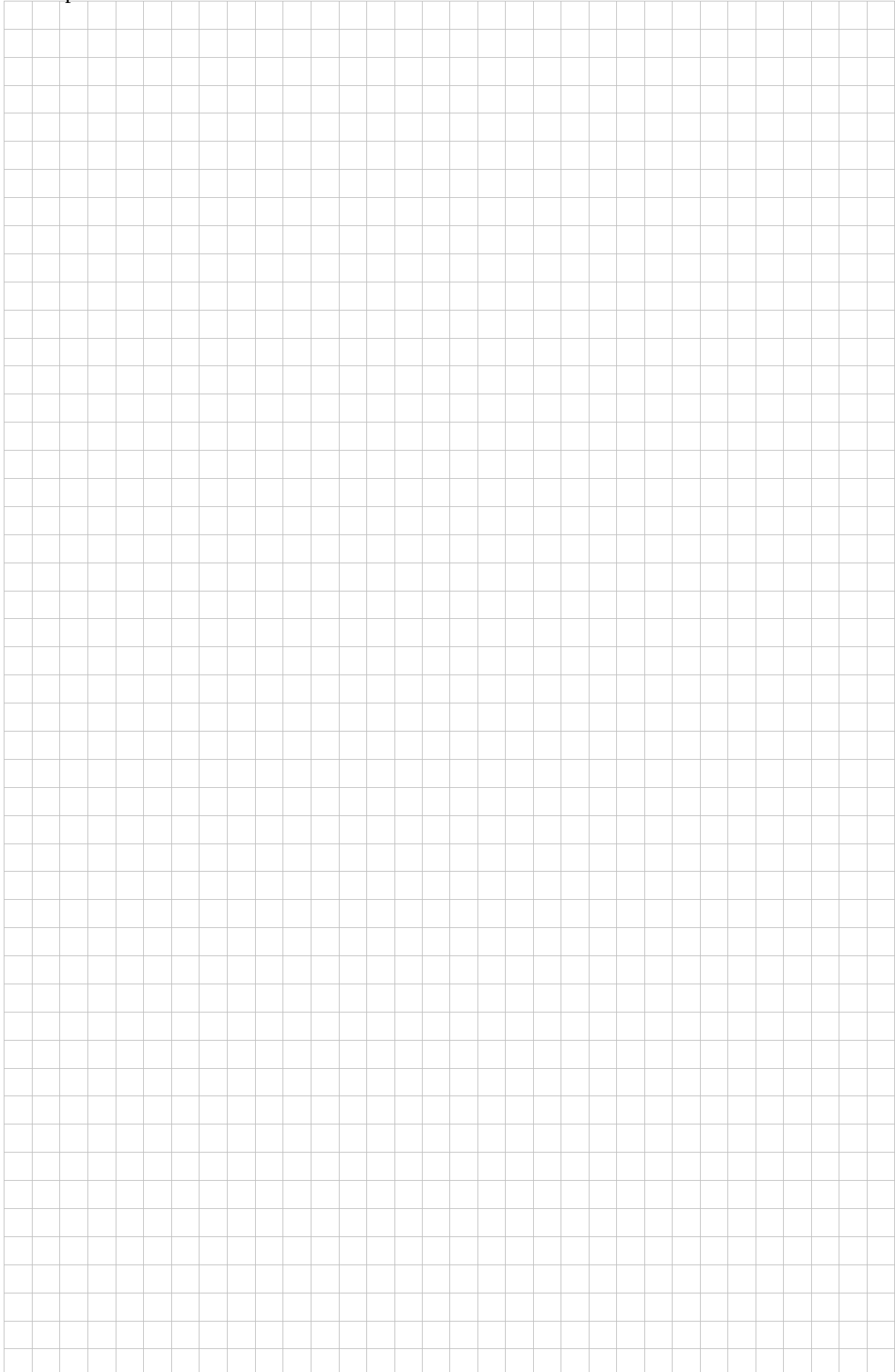
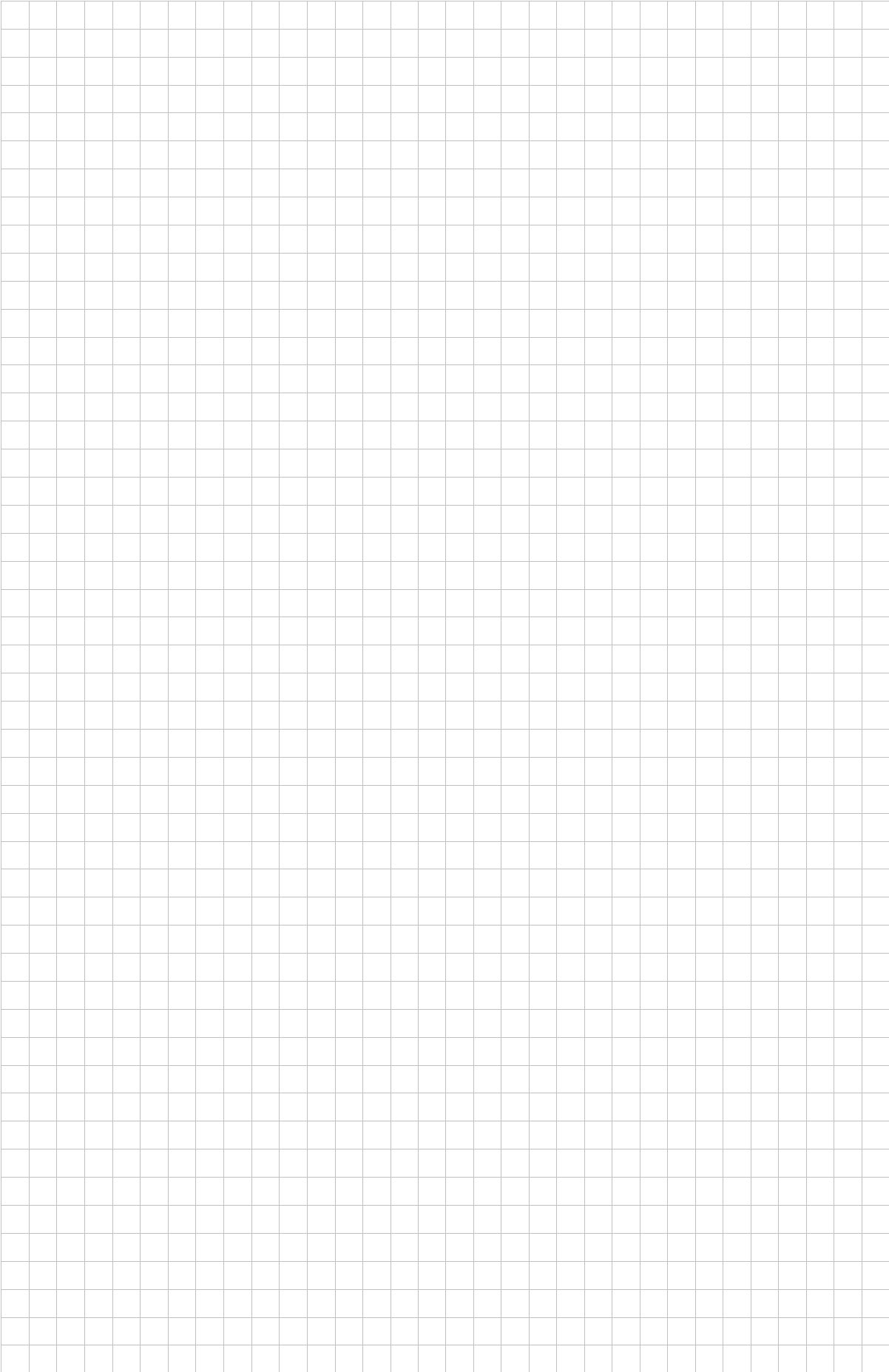


Figure: 47.26 - Triangle équilatéral formé par le système

En 2000, 385 astéroïdes en L_4 et 188 astéroïdes en L_5 ont été comptabilisés sur l'orbite de Jupiter, mais situés précisément selon un triangle équilatéral avec le Soleil et Jupiter de part et d'autre de Jupiter: ce sont les planètes Troyennes. Il a également été observé deux objets au point L_5 de Mars découverts en 1990 et 1998.

Notes personnelles:









48. ASTROPHYSIQUE

L'astrophysique est une branche interdisciplinaire de l'astronomie qui concerne principalement la physique et l'étude des propriétés des objets de l'Univers (étoiles, planètes, galaxies, milieu interstellaire par exemple), comme leur luminosité, leur densité, leur température et leur composition chimique. Les premières démarches scientifiques dans ce domaine dateraient du tout début du 19ème siècle.

Remarque: Actuellement, les astronomes ont une formation en astrophysique et leurs observations sont généralement étudiées dans un contexte astrophysique, de sorte qu'il y a moins de distinction entre ces deux disciplines qu'auparavant.

1. ÉTOILES

Avant d'aborder le formalisme mathématique relatif à la dynamique des étoiles, nous avons souhaité suite à une demande des lecteurs, écrire une introduction vulgarisée afin de compléter la culture générale relative à ce domaine.

Les étoiles sont donc des corps célestes gazeux dont la masse va de 0.05 masse solaire à 100 masses solaires. La luminosité d'une étoile (sa puissance) va de 10^{-6} à 10^6 fois celle du Soleil. Grossièrement, lorsque la masse double, la luminosité décuple. La plupart des étoiles visibles à l'oeil nu dans notre ciel sont des géantes bleues de 10^4 à 10^5 fois plus lumineuses que le Soleil ; elles ne représentent cependant que 10% des étoiles qui peuplent notre galaxie, les 90% restantes étant moins lumineuses que le Soleil.

Les astronomes (de Harvard entre 1918 à 1928) ont mis en place une méthode de classification des étoiles basée sur la position dans leur spectre, des raies spectrales d'absorption (spectroscopie). Autrefois classées de A à Q, l'évolution de la spectrométrie a permis leur regroupement et leur réorganisation. Les classes sont aujourd'hui définies par les lettres OBAFGKM, et chacune est divisée en 10 sous-classes, notées de 0 à 9. La classification spectrale (tirée d'un spectre continu qui se résume seulement à certaines raies du spectre après le passage de la lumière dans un milieu donné) peut être croisée avec les classes de luminosité de sorte que nous puissions en inférer la température à la surface de l'étoile (nous démontrerons comment obtenir mathématiquement cette information):

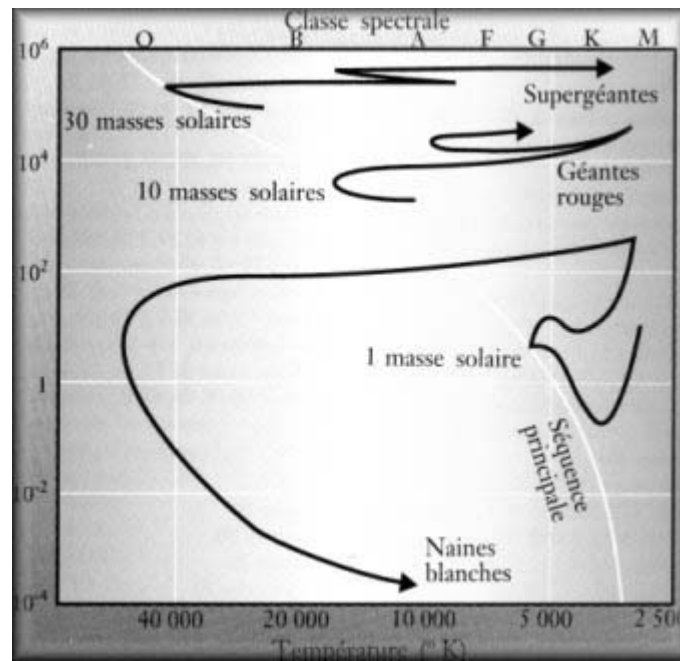


Figure: 48.1 - Méthode de classification de "Miss Cannon" des étoiles

Les étoiles O ont été découvertes à la fin du 19ème siècle. Elles sont très chaudes et leurs spectres les rapprochent des nébuleuses. Les B sont des étoiles à hélium, les A à hydrogène. L'élément prépondérant des F est le calcium. Les G sont du même type que le Soleil et les K en diffèrent assez peu. Les M sont caractérisées par l'oxyde de titane et les S par l'oxyde de zirconium, tandis que R et N contiennent des hydrocarbures et du cyanogène.

La grande courbe au centre indique l'évolution d'une étoile de même masse que le Soleil. Après un passage sur la séquence principale, elle devient une géante rouge, éventuellement une nébuleuse planétaire (éjection du combustible de l'étoile à de grandes distances), avant de terminer sa vie sous la forme d'une naine blanche. Par comparaison nous avons indiqué l'évolution d'étoiles 10 ou 30 fois plus massives que le Soleil: elles quittent la séquence principale pour devenir des supergéantes puis elles finissent en supernovae qui ne peuvent être représentées sur ce diagramme!

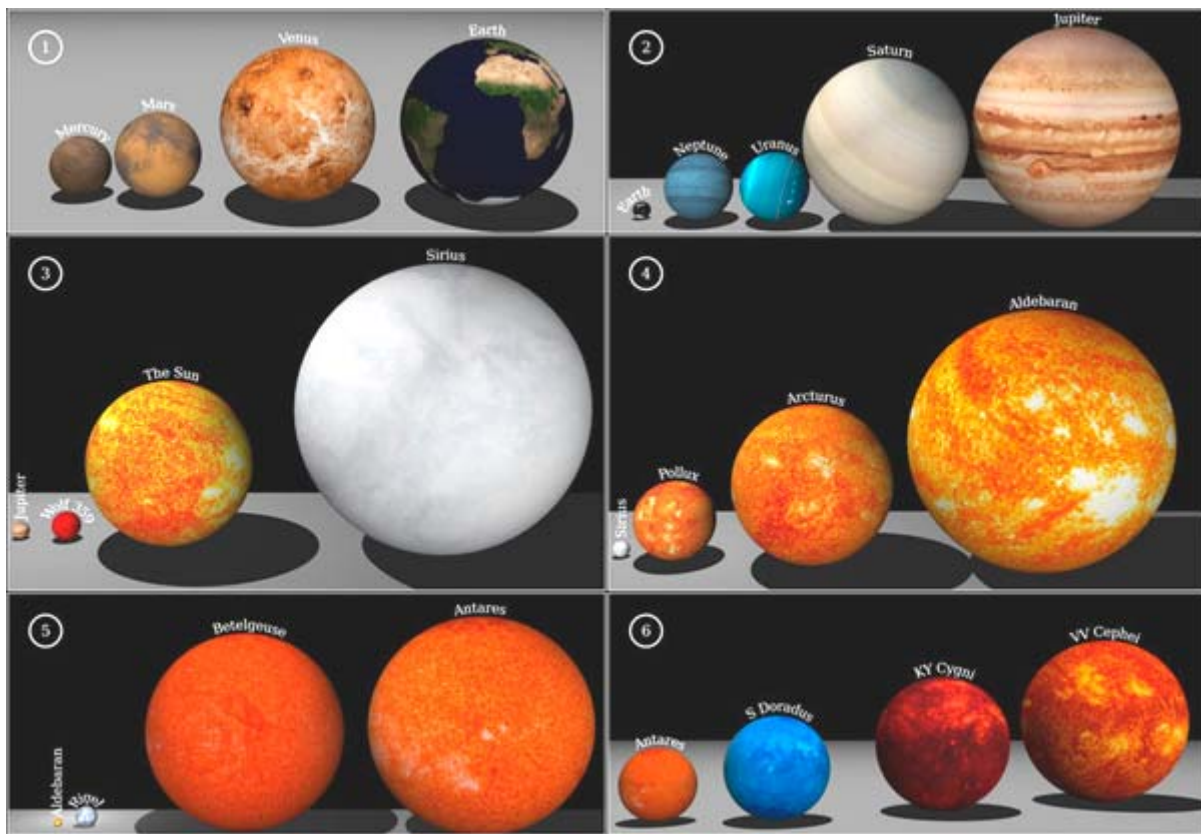


Figure: 48.2 - Proportion de quelques familles d'astres existant dans l'Univers

Une étoile est dans un premier temps en équilibre hydrostatique. Les forces gravitationnelles dues à sa masse sont compensées par les forces de pression interne dues à la température élevée entretenue par des réactions thermonucléaires à basse densité et à la pression de dégénérescence des électrons à densité élevée. Une étoile passe 90% de sa vie à fusionner de l'hydrogène en hélium qui s'accumule en son centre. Durant cette phase, elle évolue dans ce que nous appelons "**la séquence principale**" du diagramme de Hertzsprung-Russel représenté ci-dessous. Ce diagramme met en relation la température de surface (abscisse logarithmique présentée en ordre opposé) et la luminosité (ordonnée logarithmique) pour des populations d'étoiles. La séquence principale apparaît comme une diagonale. La température de surface et la luminosité étant directement fonction de la masse:

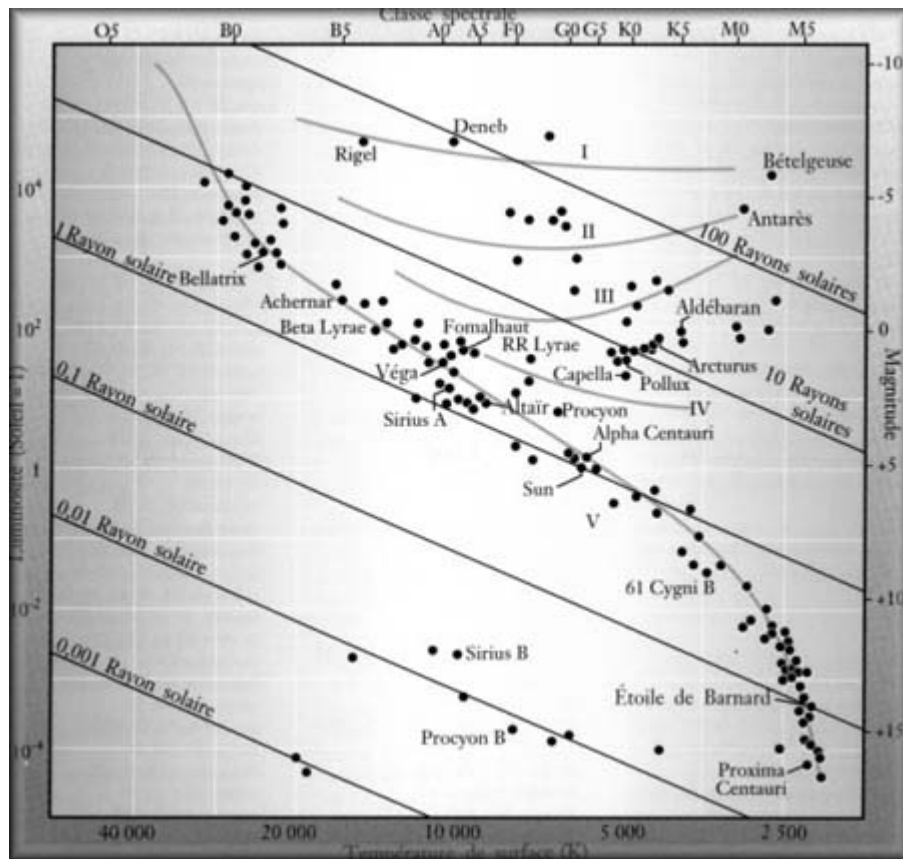


Figure: 48.3 - Diagramme de classification de Hertzsprung-Russell

Chacune des étoiles du ciel trouve sa place sur le diagramme introduit par Hertzsprung et Russell (diagramme H-R ci-dessus) dont les diverses régions permettent d'en repérer le stade d'évolution. Il est alors possible d'y tracer une courbe représentative de l'évolution d'une étoile donnée à partir de la connaissance de son état au moment de l'observation.

Ainsi, les étoiles massives évoluent plus vite que les étoiles de faible masse, mais ce résultat est déduit d'autres considérations que celles permettant de construire le diagramme. Le diagramme sert notamment à évaluer l'âge moyen d'un amas d'étoiles à partir de celui de ses composants. De même, il permet de caractériser les étoiles variables et leurs composantes telles les géantes rouges qui deviennent instables et pulsantes en vieillissant. Cette famille d'objets instables définit une bande d'instabilité sur le diagramme. Ce diagramme traduit la classification spectrale des étoiles ou leur température de surface en fonction de leur magnitude absolue ou de leur luminosité.

Ce diagramme, sur lequel toutes les étoiles trouvent leur place dès que nous connaissons leurs caractéristiques, fut développé indépendamment en Europe par Ejnar Hertzsprung et aux États-Unis par Henry Norris Russell. L'axe horizontal indique la classification spectrale en partant, à gauche, des étoiles les plus chaudes, les bleues, pour atteindre les moins chaudes, les rouges, à droite. Les étoiles se positionnent en groupes spécifiques sur le diagramme: celles qui évoluent sur leur séquence principale se situent sur une courbe incurvée qui commence en haut, à gauche, et se termine en bas, à droite. C'est sur cette courbe que se regroupent les étoiles stables qui brûlent leur hydrogène et, parmi elles, le Soleil qui se positionne au centre du diagramme. Les géantes et les supergéantes apparaissent dans la partie supérieure droite, tandis que les naines blanches se regroupent dans la partie inférieure gauche. Au fur et à mesure qu'elle évolue, chaque étoile décrit une courbe particulière: elle commence par suivre la trajectoire de Hayashi jusqu'à ce qu'elle atteigne sa séquence principale sur laquelle elle évolue tant que son noyau brûle de l'hydrogène. Lorsque commence la combustion de l'hélium, elle remonte vers le haut où se concentrent les géantes rouges et y reste jusqu'à ce que la fusion nucléaire s'arrête: elle

s'effondre alors sur elle-même pour rejoindre les naines blanches ou dans le cas d'une certaine valeur de masses solaires, les étoiles à neutrons, Trous Noirs ou encore, si sa masse est très élevée, explose en supernovae.



Figure: 48.4 - Région du ciel en lumière visible avant et après la supernova de 1987

Lorsque la masse d'hélium d'une étoile devient suffisante, l'augmentation de pression induit une augmentation de la température amorçant ainsi la fusion de l'hélium ("**flash de l'hélium**") en carbone, oxygène et néon créant un second front de combustion à l'intérieur du premier. Pour une étoile de masse solaire, les réactions s'arrêtent à ce stade. L'étoile grossit et se refroidit en surface. Elle devient une géante rouge 10^4 fois plus lumineuse qu'auparavant. Elle passe par des phases d'instabilité et finit par expulser progressivement ses couches externes en formant une "**nébuleuse planétaire**". Son noyau, dont la densité est de plusieurs tonnes par centimètre cube, se refroidit lentement: c'est la naine blanche (nous aborderons ce processus sous forme mathématique plus loin). L'équilibre y est maintenu par la pression de dégénérescence des électrons.

Pour une étoile plus massive, la température interne devient assez importante pour que le carbone et l'oxygène puissent fusionner en silicium. À son tour, s'il est en masse suffisante, le silicium fusionnera en fer. Les fronts de combustion se développent dans un schéma dit en pelures d'oignon. Le fer est le nucléotide le plus stable: il se trouve au fond de la vallée de stabilité (cf. chapitre de Physique Nucléaire). Il ne peut ni fusionner, ni fissionner. Lorsque la densité atteint une valeur critique (cela correspond à une masse totale de l'étoile de plus de 8 masses solaires), la pression de dégénérescence des électrons n'arrive plus à maintenir l'équilibre contre la gravitation. En un dixième de seconde, le noyau de fer s'effondre. Les autres couches du cœur se précipitent vers le noyau effondré sous forme d'une onde dont le maximum de vitesse correspond au rayon sonique.

La densité du noyau devient alors énorme. Il se produit des réactions β^- inverses où les protons capturent les électrons en formant des neutrons et libérant un flot de neutrinos. Lorsque le noyau de l'étoile atteint la densité nucléaire de $10^{18} \text{ [kg} \cdot \text{m}^{-3}]$, la compaction s'arrête brutalement (rayon d'environ 10km !). Les couches externes du noyau rebondissent par un choc superélastique et entrent en expansion. Lorsque cette onde de choc réfléchie rejoint le rayon sonique, la température monte tellement haut que la chiffrer n'a plus de sens. La matière subit une photodésintégration complète (tous les nucléotides sont désagrégés en gaz de nucléons). Finalement par un mécanisme pas clairement établi, toutes les couches externes de l'étoile sont éjectées dans l'espace: c'est une "**supernovae de type II**".

Le noyau effondré, presque entièrement constitué de neutrons, sera en rotation rapide si l'étoile initiale avait un moment cinétique non nul (conservation du moment cinétique oblige). Le champ magnétique est également conservé et dépasse de loin tout ce qui sera probablement jamais réalisable en laboratoire. Cela provoque un rayonnement synchrotron qui donne l'illusion que l'étoile clignote. Cela provoque un rayonnement synchrotron qui donne l'illusion que l'étoile clignote, c'est pourquoi ces jeunes "étoiles à neutrons" sont dénommées "pulsars".

Pour les étoiles très massives (plus de 50 masses solaires), la masse totale du coeur qui s'effondre pourrait dépasser 3 masses solaires. Dans ce cas, la gravité devient telle que sa masse s'effondre au-delà des dernières forces répulsives et se compacte en une singularité. La courbure de l'espace devient telle qu'aucune matière, rayonnement ou information ne peut plus s'échapper au-delà d'un volume appelé horizon ou sphère de Schwarzschild. C'est un "Trou noir". Tout ce qui y tombe perd son identité. Un trou noir ne présente plus que trois propriétés: sa masse, son moment cinétique et sa charge électrique. Nous disons qu'un trou noir n'a pas de chevelure. De plus, une telle singularité devrait toujours être cachée par un horizon, être habillée.

1.1. GENÈSE

Nous allons voir maintenant comment des astres nouveaux peuvent naître à partir d'immenses nuages de gaz qui s'étendent entre les étoiles dans les galaxies. Ce milieu interstellaire est une source potentielle d'étoiles nouvelles, qui une fois leur vie terminée (sous forme de géante rouge ou de supernovae), peuvent réinjecter une partie de leur matériau dans l'espace intersidéral.

Au fait, personne ne sait vraiment les détails de la façon dont un nuage interstellaire aboutit à une étoile car il s'agit d'un problème fort difficile, essentiellement à cause de l'apparition de toute une hiérarchie de structures, sous-structures, etc. dans le nuage à mesure qu'il s'effondre sur lui-même. Des mouvements turbulents apparaissent, qui ne peuvent être décrits de manière simple par les équations hydrodynamiques (cf. chapitre de Mécanique Des Milieux Continus). D'autres complications apparaissent lorsque nous voulons tenir compte du champ magnétique sur le gaz en contraction, ou d'explosions de supernovae dans le nuage...

Au moins, pouvons-nous donner les conditions nécessaires pour qu'une étoile puisse se former au sein d'un nuage interstellaire. Pour cela, plusieurs barrières doivent en fait être franchies. Une première barrière est thermique. Une deuxième barrière est rotationnelle: une proto-étoile qui se contracte tourne de plus en plus vite et peut littéralement exploser si sa vitesse de rotation devient trop importante (conservation du moment cinétique). Examinons ces deux effets.

1.1.1. EFFONDREMENT D'UN NUAGE INTERSTELLAIRE

Deux forces opposées sont présentes dans un nuage de masse M et de rayon R : une force d'autogravitation, qui tend à contracter le nuage, et une force de pression thermique, qui tend à le faire exploser.

Nous pouvons quantifier ces deux tendances opposées en termes d'énergie: le nuage possède une énergie potentielle de gravitation (négative) et une énergie cinétique (positive) due à l'agitation thermique de ses molécules.

Nous savons (cf. chapitre de Mécanique Classique) que l'énergie potentielle de gravitation de deux particules de masses m et m' séparées de r s'écrit $-\frac{Gmm'}{r}$. Donc l'énergie potentielle d'un nuage sphérique (...) de masse M et de rayon R est de l'ordre de:

$$E_p \cong -\frac{GM^2}{R} \quad (48.1)$$

Dans un gaz en équilibre thermodynamique, une particule a une énergie cinétique (cf. chapitre de Mécanique Des Milieux Continus) de $kT/2$ par degré de liberté (translation, rotation, etc.). Donc, si μ est la masse moyenne d'une molécule du nuage, l'énergie cinétique totale de cette dernière aura pour expression:

$$E_c = \frac{3}{2} kT \frac{M}{\mu} \cong kT \frac{M}{\mu} \quad (48.2)$$

Le nuage s'effondre alors si son énergie mécanique totale est négative, soit (selon l'approximation précédente):

$$kT \frac{M}{\mu} \leq \frac{GM^2}{R} \quad (48.3)$$

L'équation ci-dessus permet de définir la "**masse de Jeans**" (dans l'hypothèse d'une distribution sphérique et homogène). C'est la masse minimum (limite), à une température T et une masse volumique ρ données, pour que le nuage commence son effondrement jusqu'à ce qu'un autre processus physique intervienne éventuellement pour stopper la contraction du gaz.

En éliminant le rayon par:

$$M = 4\pi\rho R^3 / 3 \quad (48.4)$$

dans l'équation précédente, nous avons alors:

$$M \geq \frac{kT}{G\mu} R = \frac{kT}{G\mu} \left(\frac{3M}{4\pi\rho} \right)^{1/3} \Rightarrow M^{2/3} \geq \frac{kT}{G\mu} \left(\frac{3}{4\pi\rho} \right)^{1/3} \Rightarrow M \geq \left(\frac{kT}{G\mu} \right)^{3/2} \left(\frac{3}{4\pi} \right)^{1/2} \frac{1}{\sqrt{\rho}} \quad (48.5)$$

ce que les astrophysiciens notent à la suite de toutes les approximations faites...:

$$M \geq M_{Jeans} \cong C \left(\frac{kT}{G\mu} \right)^{3/2} \frac{1}{\sqrt{\rho}} \quad (48.6)$$

où C est une constante sans unités. En prenant un nuage composé d'hydrogène uniquement avec n atomes par mètre cube (c'est donc une densité!), nous aurons $\mu \cong m_p$ et $\rho \cong n \cdot m_p$ où m_p est la masse du proton. Nous pouvons alors exprimer la masse de Jeans en masses solaires de la manière suivante:

$$\begin{aligned} M_{Jeans} &\cong C \left(\frac{kT}{Gm_p} \right)^{3/2} \frac{1}{\sqrt{nm_p}} = C \left(\frac{k}{G} \right)^{3/2} \frac{T^{3/2}}{m_p^{3/2}} \frac{1}{n^{1/2} m_p^{1/2}} = \left[C \left(\frac{k}{G} \right)^{3/2} \frac{1}{m_p^2} \right] \frac{T^{3/2}}{n^{1/2}} \\ &= C' \cdot \frac{T^{3/2}}{n^{1/2}} M_{\odot} \end{aligned} \quad (48.7)$$

où nous avons la certitude que $C' \gg 1$ et où:

$$M_{\odot} = 1.9891 \cdot 10^{30} \text{ [Kg]} \quad (48.8)$$

est la masse solaire (unité de masse conventionnellement utilisée pour les étoiles ou les autres objets massifs). Cette masse est donc comme le laisse deviner son nom, égale à la masse de notre Soleil. Son symbole et sa valeur sont :

La masse solaire vaut environ 330 000 fois la masse de la Terre.

Nous voyons que la masse de Jeans varie comme $n^{-1/2}$. Ceci a une conséquence importante: à mesure que le nuage se contracte, n augmente, et donc M_{Jeans} diminue. Autrement dit, le nuage peut se fragmenter en sous-nuages une fois la masse de Jeans pour ces sous-nuages atteinte. Ces derniers vont à leur tour se scinder en sous-nuages, etc. Nous avons donc toute une hiérarchie d'effondrements, depuis les grandes masses vers les petites masses.

La chose importante à noter aussi est que la masse de Jeans d'un nuage est beaucoup plus grande que les masses stellaires individuelles (il suffit de voir les constantes contenues dans la relation précédente pour se rendre compte que les facteurs sont relativement conséquents!). Donc, les étoiles naissent en général par ensemble de plusieurs étoiles: nous ne pouvons pas former en principe un Soleil isolé dans une galaxie, à partir d'un tout petit nuage. Une fois formées, les étoiles se diluent dans la galaxie par les effets de rotations et de marées galactiques. Ainsi, le Soleil a perdu de vue ses soeurs depuis bien longtemps probablement...

1.1.2. RAYON DE JEANS

Nous pouvons également exprimer la condition d'effondrement en termes de "**rayon de Jeans**", toujours pour une température T et une masse volumique ρ données. Il suffit en fait d'éliminer M dans la relation:

$$kT \frac{M}{\mu} \leq \frac{GM^2}{R} \quad (48.9)$$

Ainsi, nous avons:

$$R \leq \mu \frac{GM}{kT} \Rightarrow R \leq \mu \frac{G4\pi\rho R^3}{kT3} \Rightarrow R \geq \sqrt{\frac{3kT}{\mu G4\pi\rho}} \Rightarrow R \geq \sqrt{\frac{3}{4\pi}} \sqrt{\frac{kT}{\mu G}} \frac{1}{\sqrt{\rho}} \quad (48.10)$$

Soit:

$$R \geq R_{\text{Jeans}} \equiv C'' \sqrt{\frac{kT}{\mu G}} \frac{1}{\sqrt{\rho}} \quad (48.11)$$

Ainsi, le rayon de Jeans est le rayon minimal pour qu'une sphère de masse donnée soit stable. Au-delà, le nuage stellaire va s'effondrer sur lui-même selon les mêmes conditions que la masse de Jeans.

Au vu des valeurs des paramètres de la relation précédente, nous voyons alors que les nuages de formation stellaire sont en fait immenses en ordre de grandeur. Ces véritables pépinières sont ensuite dispersées dans la galaxie par effet de marée galactique, comme nous le soulignons plus haut.

1.1.3. TEMPS DE CHUTE LIBRE

Nous avons vu pour l'instant que la masse d'un nuage doit être grande par rapport à celle du Soleil pour que l'effondrement se produise. Nous allons maintenant estimer le temps que va prendre le nuage pour s'effondrer sur lui-même.

Au début de l'effondrement, rien n'arrête la chute du nuage, la pression interne est encore très faible et l'énergie lumineuse provenant de l'échauffement progressif du nuage (lié à la contraction de ce dernier) est immédiatement évacuée car le nuage est encore transparent.

Une parcelle de nuage à la périphérie, in extenso à la distance R du centre du nuage, subit une accélération $a \cong GM / R^2$ de la part de ce dernier. Elle commence donc à tomber vers le centre avec la loi $x = at^2 / 2$ (cf. chapitre de Mécanique Classique). La parcelle aura atteint le centre quand $x \cong R$. Nous obtenons donc:

$$t_{chute} \cong \sqrt{\frac{2R}{a}} = \sqrt{\frac{2R^3}{GM}} \quad (48.12)$$

Nous pouvons exprimer ce temps uniquement en terme de masse volumique, puisque $M = 4\pi R^3 \rho / 3$:

$$t_{chute} \cong \sqrt{\frac{3}{2\pi G \rho}} \quad (48.13)$$

À noter que le temps de chute ne dépend pas de la taille de l'objet ni de sa masse, mais uniquement de sa masse volumique.

Une application numérique pour un nuage d'hydrogène donne alors:

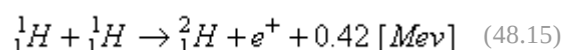
$$t_{chute} \cong \sqrt{\frac{3}{2\pi G m_p}} \frac{1}{\sqrt{n}} \cong 6.5 \cdot 10^{10} \frac{1}{\sqrt{n}} \text{ [années]} \quad (48.14)$$

Nous remarquons que ces temps restent petits par rapport à l'âge de l'Univers (13-14 milliards d'années). Ainsi, la genèse stellaire est un phénomène relativement rapide: plusieurs générations d'étoiles ont pu voir le jour depuis la formation des galaxies.

1.1.4. DURÉE DE VIE NUCLÉAIRE

L'âge des étoiles est principalement un problème de calcul du carburant nucléaire. La résolution de ce problème a été apportée par la relativité, et en particulier par l'équivalence masse-énergie (cf. chapitre de Relativité Restreinte).

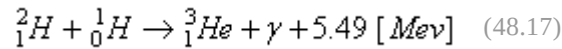
Même si la description détaillée des réactions nucléaires au coeur du Soleil n'a été fait qu'au milieu des années 1930 par Hans Bethe, les astrophysiciens ont soupçonné peu après les travaux d'Einstein que cette équivalence pouvait expliquer l'éclat du Soleil sur des milliards d'années, par exemple via la fusion de l'hydrogène (proton, p) en hélium (deux protons, deux neutrons) via une succession d'étapes (l'énergie indiquée est l'énergie cinétique des différents éléments):



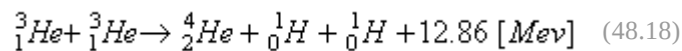
Le positron s'annihile immédiatement avec l'un des électrons d'un atome d'hydrogène environnant et leur masse-énergie est évacuée sous forme de deux photons gamma:

$$e^+ + e^- \rightarrow 2\gamma + 1.02 \text{ [MeV]} \quad (48.16)$$

Après ceci, le deutérium produit lors de la première étape peut fusionner avec un nouveau noyau d'hydrogène pour produire un isotope de l'hélium:



Finalement, deux isotopes de l'hélium ${}^3\text{He}$ peuvent fusionner et produire l'isotope normal de l'hélium ${}^4_2\text{He}$ ainsi que deux noyaux d'hydrogène qui peuvent commencer à nouveau la réaction de trois façons différentes appelées PP1, PP2 et PP3:



Et encore ces réactions ne se produisent pas toutes selon les mêmes probabilités et les mêmes températures...

La mesure de la masse du proton donne $m_p = 1.673 \cdot 10^{-27} \text{ [kg]}$, alors que l'hélium a une masse de $m_{\text{He}} = 6.645 \cdot 10^{-27} \text{ [kg]}$, soit une perte en masse atomique de (nous négligeons la masse des positrons qui est 10'000 fois plus petite ainsi que celle du neutrino):

$$4m_p - m_{\text{He}} = 0.047 \cdot 10^{-27} \text{ [kg]} \quad (48.19)$$

Donc une perte relative de masse par fusion (c'est la part de la réaction qui s'échappe du Soleil sous forme d'énergie cinétique):

$$\frac{\Delta m}{m} = \frac{0.047 \cdot 10^{-27}}{4 \cdot 1.673 \cdot 10^{-27}} \cong 0.0072 = 0.72\% \quad (48.20)$$

Nous allons démontrer plus bas que le Soleil émet une puissance de:

$$P \cong 3.861 \cdot 10^{26} \text{ [W]} \quad (48.21)$$

Donc sa consommation en masse par seconde est de:

$$\frac{\Delta m}{t} = \frac{E}{t \cdot c^2} = \frac{P}{c^2} \cong 4.4 \cdot 10^9 \text{ [kg} \cdot \text{s}^{-1}] \quad (48.22)$$

C'est à dire que sa masse diminue de 4.4 millions de tonnes par seconde...

Or nous savons que ce nombre correspond seulement à 0.72% de la masse mise en réaction dans la fusion. La masse totale mise en réaction est alors (règle de trois):

$$\frac{4.4 \cdot 10^9}{0.0072} \cong 6.27 \cdot 10^{11} \text{ [kg} \cdot \text{s}^{-1}] \quad (48.23)$$

Ainsi, à chaque seconde 627 millions de tonnes d'hydrogène (ionisé) 1 fusionnent en hélium 4 avec une perte de masse de 4.4 millions de tonnes qui est transformée en énergie.

En estimant que seul le centre du Soleil remplit les conditions thermiques pour la fusion, ceci nous amène à déterminer son temps de vie nucléaire:

$$t_{\text{nucl}} = \frac{0.1 \cdot M_{\odot}}{6.27 \cdot 10^{11}} \cong 3.17 \cdot 10^{17} [\text{s}] \quad (48.24)$$

En transformant cela en années nous avons:

$$t_{\text{nucl}} \cong 1.01 \cdot 10^{10} [\text{années}] \quad (48.25)$$

1.2. TEMPÉRATURE INTERNE

Les étoiles sont supposées être des amas sphériques d'hydrogène gazeux où les interactions entre molécules sont régies par l'attraction gravitationnelle.

Une étoile n'a pas de paroi qui la délimite, c'est-à-dire qu'il n'y a pas de forces extérieures donc:

$$PV = 0 \quad (48.26)$$

En utilisant le théorème du Viriel vu dans le chapitre de Mécanique Des Milieux Continus:

$$E_c = -\frac{1}{2} E_p \quad (48.27)$$

Nous avons pour une masse sphérique gazeuse de rayon R de masse M composée de N corps:

$$E_c = \frac{3}{2} kNT \text{ et } E_p = -\frac{3}{5} \frac{GM^2}{R} \quad (48.28)$$

Remarque: Pour le calcul de l'énergie potentielle, nous renvoyons le lecteur au chapitre de Mécanique Classique du site.

Donc:

$$\frac{3}{2} kNT = \frac{1}{2} \frac{3}{5} \frac{GM^2}{R} \quad (48.29)$$

où rappelons-le, k est la constante de Boltzmann.

Ce qui nous donne:

$$T = \frac{1}{5} \frac{GM^2}{NkR} \quad (48.30)$$

Avec pour une étoile donnée N étant le rapport de la masse totale de l'étoile sur la masse moyenne d'une molécule.

Pour le Soleil, il vient que $T = 10^7 [^\circ K]$.

C'est la température centrale du Soleil. Les mesures optiques mesurées depuis la Terre ne donnent que la température en surface (chromosphère), soit 6'000 [°K]. La température interne calculée est donc environ 1'600 fois plus élevée qu'à la surface. Des méthodes indépendantes basées sur les réactions nucléaires au centre du Soleil (mesure du flux de neutrinos solaires) donnent le même ordre de grandeur, mais les valeurs précises diffèrent d'un facteur 2 à 3.

1.3. TEMPÉRATURE EXTERNE

Nous avons démontré dans le chapitre de Thermodynamique que la loi de Stefan-Boltzmann, permet de calculer la température d'un corps chauffé à partir de son émittance ou de son énergie interne en termes de densité tel que:

$$\rho(T) = \sigma_{s.B.} T^4 = \frac{4}{c} M(T) \quad (48.31)$$

avec:

$$\sigma = \frac{\sigma_{s.B.} \cdot c}{4} \quad (48.32)$$

étant la constante de Stefan-Boltzmann.

Prenons un exemple intéressant qui nous concerne directement:

L'émittance moyenne dite aussi "**émittance moyenne bolométrique**" reçue par la Terre hors atmosphère, appelée aussi "**constante solaire**" (qui n'est au fait pas constante... sur une échelle de plusieurs milliards d'années), est directement mesurable en orbite et vaut 1'373 [$W \cdot m^{-2}$].

Connaissant la distance moyenne au Soleil comme étant d'environ $1.496 \cdot 10^{11} [m] = 1 [UA]$ (Unité Astronomique), nous pouvons calculer la surface de la sphère S à $R = 1 [UA]$ et donc la puissance solaire P . Ainsi:

$$S = 4\pi R^2 \cong 2.812 \cdot 10^{23} [m^2] \quad (48.33)$$

et:

$$P = S \cdot M(T) \cong 3.861 \cdot 10^{26} [W] \quad (48.34)$$

Supposant connu le rayon du Soleil comme valant $r \cong 6.9599 \cdot 10^8 [m]$, nous pouvons calculer sa surface S puis l'émittance radiative solaire $M(T)$. Ainsi:

$$S = 4\pi R^2 \cong 6.087 \cdot 10^{18} [m^2] \quad (48.35)$$

et:

$$M(T) = \frac{P}{S} \cong 6.343 \cdot 10^7 [W \cdot m^{-2}] \quad (48.36)$$

Remarque: La surface rayonnante d'une étoile est appelée "**photosphère**".

À l'aide de la loi de Stefan-Boltzmann, nous pouvons maintenant calculer la température thermodynamique de la photosphère:

$$T = \sqrt[4]{\frac{M(T)}{\sigma}} \cong 5780 \text{ [K]} \quad (48.37)$$

La loi de Planck (cf. chapitre de Thermodynamique) appliquée à cette température nous permettrait de calculer la distribution spectrale du rayonnement solaire et nous voyons alors que le maximum de l'intensité est dans le domaine visible (notre visibilité...) du spectre qui va de 400 [nm] à 700 [nm].

1.4. LUMINOSITÉ

La "**luminosité bolométrique intrinsèque**" d'une étoile correspond à sa puissance totale rayonnée dans tout le spectre électromagnétique dans la direction de l'observateur exprimée de façon relative à la puissance totale rayonnée par le Soleil. En supposant toutes les étoiles sphériques et isotropes, nous pouvons l'exprimer en unités solaires:

$$L_{bol} = \frac{P_{\text{étoile}}}{P_{\text{soleil}}} = \frac{P_{\text{étoile}}}{P_{\odot}} \quad (48.38)$$

La puissance rayonnée P se calcule elle, en multipliant bien évidemment l'émittance radiative (loi de Stefan-Boltzmann) par la surface de l'étoile:

$$P = M(T) \cdot S = \sigma \cdot T^4 \cdot 4\pi r^2 \quad (48.39)$$

La luminosité bolométrique intrinsèque d'une étoile est donc proportionnelle au carré de son rayon et à la quatrième puissance de sa température de surface. En prenant le Soleil comme référence, les constantes se simplifient. Nous pouvons alors écrire:

$$L_{bol} = \frac{P_{\text{étoile}}}{P_{\odot}} = \left(\frac{T}{T_{\odot}}\right)^4 \left(\frac{r}{r_{\odot}}\right)^2 = c^{\#} \cdot T^4 \cdot r^2 \quad (48.40)$$

avec $r_{\odot} = 6.9599 \cdot 10^8 \text{ [m]}$ et $T_{\odot} = 5780 \text{ [K]}$ d'où $c^{\#} = 1.85 \cdot 10^{-33} \text{ [K}^{-4} \text{ m}^{-2}]$

En astrophysique, nous utilisons également une échelle logarithmique pour exprimer la luminosité bolométrique d'une étoile: la magnitude absolue M . Cette unité a une origine empirique qui sera expliquée plus bas.

1.4.1. ÉCLAT

"**L'éclat**" d'une étoile est sa "**luminosité apparente**". L'éclat (luminosité apparente) d'une étoile correspond à la densité de rayonnement reçu par l'observateur, c'est-à-dire au flux et vaut le rapport entre la puissance de l'étoile et la surface de la sphère dont le rayon est égal à la distance d qui sépare l'observateur de l'étoile:

$$e = \frac{P}{S} = \frac{P}{4\pi d^2} \quad (48.41)$$

L'éclat diminue ainsi avec le carré de la distance. Il est important de remarquer que cette grandeur n'a aucune relation directe avec les propriétés intrinsèques physiques de l'étoile concernée (contrairement à la luminosité bolométrique!).

En astrophysique, nous utilisons également une autre échelle où la luminosité apparente est donnée par une autre grandeur d'origine empirique: la magnitude apparente, qui sera expliquée de suite ci-dessous.

1.4.2. MAGNITUDE APPARENTE

Ptolémée en 137 après J.-C. avait défini une échelle de six grandeurs pour exprimer l'éclat des étoiles, la première pour les plus brillantes et la sixième pour les étoiles tout juste visibles à l'oeil nu (6 grandeurs et donc 5 écarts).

Au cours du 19ème siècle, avec l'arrivée de nouvelles techniques d'observations photométriques (photographiques puis photoélectriques), l'échelle de grandeur a été remplacée par celle de "**magnitude apparente**" qui a été définie de telle sorte à ce que cette nouvelle échelle soit proche de l'ancienne.

La définition est la suivante:

- L'échelle est logarithmique en base 10 (par commodité des grandeurs manipulées)
- Il y a 5 écarts de magnitude correspondant à un rapport de luminosité apparente de 1 pour 100 (1:100)
- L'échelle est inverse (une magnitude élevée correspond à un faible éclat/luminosité apparente).

À l'aide de ces définitions, nous pouvons construire une règle liant de façon relative les éclats de deux étoiles à leur magnitude apparente m .

Pour une étoile 2, cent fois plus brillante ou éclatante qu'une étoile 1, l'étoile 1 est 5 unités de magnitude au-dessus de l'étoile 2 (n'oublions pas que l'échelle est inverse). Donc:

$$\frac{e_2}{e_1} = 10^2 \quad (48.42)$$

correspond à:

$$\Delta m = m_1 - m_2 = 5 \quad (48.43)$$

Nous pouvons alors poser les relations:

$$\log\left(\frac{e_2}{e_1}\right) \rightarrow 2 \text{ et } m_1 - m_2 \rightarrow 5 \quad (48.44)$$

Par application de la règle de trois, nous construisons:

$$\frac{\log\left(\frac{e_2}{e_1}\right)}{2} = \frac{m_1 - m_2}{5} \quad (48.45)$$

En simplifiant, nous trouvons la "**loi de Pogson**" qui exprime la relation entre magnitudes visuelles apparentes et éclats de deux étoiles:

$$m_1 - m_2 = 2.5 \cdot \log\left(\frac{e_2}{e_1}\right) \quad (48.46)$$

Ainsi définie, l'échelle de magnitudes visuelles n'est que relative. La référence photométrique est similaire à l'éclat de Véga (α Lyr).

Pour se faire une idée des magnitudes visuelles voici quelques exemples: Soleil -26.5, Pleine Lune -15, Vénus au maximum -4.8, Sirius la plus brillante des étoiles -1.5 (type spectral A1 et distante de 8.6 années-lumière), limite de la perception à l'oeil nu 6, limite de perception à travers un télescope amateur de 15 cm à ce jour (2003) 13, limite de perception du télescope spatial Hubble 30.

Il faut préciser que la magnitude apparente visuelle ne correspond pas exactement à la magnitude apparente réelle, car l'oeil n'a pas la même sensibilité pour toutes les longueurs d'onde. Les étoiles bleues ou rouges nous paraissent moins lumineuses à l'oeil qu'elles ne le sont en réalité car une partie du rayonnement se trouve dans les ultraviolets, respectivement dans l'infrarouge.

Il convient donc de préciser s'il s'agit d'une magnitude apparente visuelle ou bolométrique. En général, les astrophysiciens utilisent les grandeurs bolométriques dans leurs communiqués.

1.4.3. MAGNITUDE ABSOLUE

La magnitude absolue M (ne pas confondre avec la notation de l'émittance..) d'une étoile est une grandeur logarithmique aussi, qui exprime cette fois la luminosité L bolométrique. C'est la grandeur présentée en ordonnée du diagramme de Hertzsprung-Russel. L'échelle de cette grandeur est basée sur la magnitude visuelle.

La magnitude apparente et la magnitude absolue sont liées par la distance qui nous sépare de l'étoile. À luminosité apparente intrinsèque constante, la luminosité apparente décroît donc évidemment avec le carré de la distance comme nous l'avons déjà vu. Afin d'établir une relation, nous avons dû choisir une distance de référence par une nouvelle définition.

Définition: La "**magnitude absolue**" d'une étoile est égale à sa magnitude apparente si elle est distante de 10 parsecs (32.6 années-lumière).

Soit une étoile placée à une distance quelconque d . Son éclat e_2 est fonction de la distance et de son éclat e_1 si elle était située à 10 [pc] selon:

$$\begin{aligned} e_1 &\rightarrow 10^2 \\ e_2 &\rightarrow d^2 \end{aligned} \quad (48.47)$$

Par application de la règle de trois, nous construisons:

$$e_1 = e_2 \frac{1}{\left(\frac{d}{10}\right)^2} \quad (48.48)$$

En reprenant la loi de Pogson et en assimilant m_1 à la magnitude apparente m de l'étoile à la distance d quelconque, m_2 à la magnitude apparente de l'étoile à 10 [pc] (par définition de sa magnitude absolue M) ainsi que e_1 son éclat à 10 [pc] et e_2 son éclat à la distance quelconque, nous trouvons:

$$m_1 - m_2 = m - M = 2.51 \log \left(\frac{e_2}{e_1} \right) = 2.51 \log \left(\frac{e_2}{e_1} \left(\frac{d}{10} \right)^2 \right) = 2.51 \log \left(\frac{d^2}{100} \right) = 2.5(2 \log(d) - 2) \quad (48.49)$$

qui peut bien sûr aussi s'écrire:

$$m - M = 51 \log \left(\frac{d}{10} \right) = 5(\log(d) - 1) \quad (48.50)$$

En partant de cette définition, la magnitude absolue du Soleil est de 4.7. Sa magnitude apparente vue depuis la Terre est de -26.5. Elle est de 4.7 à 10 [pc] donc faiblement visible à l'oeil nu.

Cette dernière relation de comparaison de la magnitude absolue avec la magnitude apparente (qui est la magnitude observée effectivement sur Terre) permet une estimation de la distance d de l'objet en astrophysique.

Remarque: Pour avoir la magnitude absolue, il faut des modèles stellaires, et connaître la température de l'étoile comme nous allons de suite le voir. Dans la pratique, la seule quantité aisément accessible est évidemment la magnitude observée, qui est en fait la combinaison de la magnitude apparente et de l'absorption interstellaire.

La loi de Pogson exprime de même la relation entre magnitudes absolues M et luminosités bolométriques L de deux étoiles:

$$M_1 - M_2 = 2.51 \log \left(\frac{L_2}{L_1} \right) \quad (48.51)$$

Ainsi, Dénéb étant 300'000 fois plus lumineux que le Soleil, la magnitude absolue est de -9.

En reprenant la loi de Pogson, la magnitude absolue peut s'écrire relativement à la luminosité bolométrique absolue du Soleil:

$$M_{bol,\odot} - M = 2.51 \log \left(\frac{L_{bol}}{L_{bol,\odot}} \right) \quad (48.52)$$

Avec $M_{bol,\odot} = 4.7$ et $L_{bol,\odot} = 1$, la magnitude absolue bolométrique se calcule ainsi à partir de la luminosité bolométrique:

$$M = 4.7 - 2.51 \log(L_{bol}) \quad (48.53)$$

En reprenant l'expression de la luminosité bolométrique:

$$L_{bol} = \left(\frac{T}{T_{\odot}} \right)^4 \left(\frac{r}{r_{\odot}} \right)^2 \quad (48.54)$$

la magnitude (bolométrique) absolue d'une étoile étant directement fonction de sa température et de son rayon:

$$M = 4.7 - 2.5 \log \left[\left(\frac{T}{T_{\odot}} \right)^4 \left(\frac{r}{r_{\odot}} \right)^2 \right] = 4.7 - 10 \log \left(\frac{T}{T_{\odot}} \right) - 5 \log \left(\frac{r}{r_{\odot}} \right) \quad (48.55)$$

C'est le résultat que nous voulions montrer depuis le début: la magnitude absolue est directement liée à la luminosité bolométrique de l'étoile, raison pour laquelle c'est celle qui intéresse le plus les astrophysiciens.

Remarque: La distance d'étoiles proches a pu être déterminée grâce au satellite Hipparcos. Par mesure de la parallaxe (mesures de la position de l'étoile à six mois d'intervalle et par application des règles trigonométriques élémentaires). Mais, au-delà de quelques dizaines de parsecs, la mesure de la distance d'étoiles par parallaxe devient très imprécise. En étudiant le spectre de l'étoile, nous pouvons déterminer sa classe spectrale, sa température de surface et la placer dans le diagramme de Hertzsprung-Russell. Il est donc possible d'estimer sa magnitude absolue et de calculer approximativement sa distance.

Cet artifice de mesure est fondamental pour la cosmologie. C'est ainsi que l'on détermine la distance des galaxies proches en mesurant la période de certaines étoiles variables (nous y consacrons un petit chapitre ci-dessous).

La distance des galaxies lointaines se calcule en mesurant la magnitude apparente de supernovae qui s'y produisent fortuitement. En effet, les magnitudes absolues des supernovae du type Ia (nous les reconnaissons par l'absence de raies d'hydrogène et par la décroissance de leur luminosité) sont bien calibrées car l'énergie dégagée par ces explosions stellaires est relativement constante.

1.4.4. ÉTOILES VARIABLES

Les étoiles de la séquence principale du diagramme de Hertzsprung-Russell sont des objets très stables. La force de gravitation, qui tend à contracter l'astre, est exactement compensée par les forces de pression interne, qui tendent à le dilater. C'est au moment où l'étoile devient une géante rouge que parfois l'équilibre est rompu. Commence alors une phase d'instabilité qui se traduit par de fortes variations de la luminosité de l'étoile.

La rupture de l'équilibre est provoquée par un phénomène complexe qui met en jeu des variations de transparence des couches d'hélium près de la surface de l'étoile. À partir de là, l'astre se met à connaître une succession de dilatations et de contractions contrôlées par les forces qui assuraient auparavant l'équilibre. Lorsque la force de pression l'emporte, le volume de l'astre augmente. Mais la gravité freine le mouvement et finit par provoquer la contraction. Le volume de l'étoile passe alors sous sa valeur moyenne, jusqu'à ce que la pression interne s'oppose à la contraction et réussit à provoquer une nouvelle dilatation.

Ce ne sont pas les changements de taille qui provoquent les variations de luminosité, mais ceux de la température. Effectivement, comme nous l'avons vu précédemment, la luminosité d'une étoile varie avec la quatrième puissance de la température, alors qu'elle ne varie qu'avec le carré du rayon. Lorsque le volume de l'étoile est cependant plus faible qu'en moyenne, sa température est légèrement plus forte et la luminosité maximale. Dans le cas contraire, la température est légèrement plus basse qu'en moyenne et la luminosité minimale. L'éclat de l'étoile change donc de façon périodique, d'où le nom d'étoile variable.

Il existe dans le diagramme de Hertzsprung-Russel une bande d'instabilité qui traverse ce diagramme presque verticalement dans laquelle se produisent justement les phénomènes thermiques en question.

Les deux principaux types de variables pulsantes sont les Céphéides et les étoiles RR Lyrae. Ces astres jouent un rôle central en astrophysique. Les Céphéides sont des étoiles de quelques masses solaires. Elles sont dans la phase de combustion de l'hélium après avoir atteint le stade de géante rouge. Les étoiles de masse solaire arrivées à ce stade deviennent des RR-Lyrae. Leur luminosité varie avec une période comprise entre un jour et plusieurs semaines. La propriété remarquable des Céphéides est l'existence d'une relation entre leur luminosité moyenne et la période de leurs oscillations. Par exemple, leur luminosité moyenne est de 1'000 fois celle du Soleil pour une période de quelques jours et de 10'000 fois cette valeur pour une période de plusieurs semaines. C'est cette relation qui fait des Céphéides l'un des outils de base de l'astrophysique.

Si nous connaissons cette relation pour une étoile variable, il est relativement aisé, par la détermination de sa période d'en tirer la magnitude absolue M . En mesurant alors sa magnitude apparente m nous pouvons ensuite calculer sa distance d en parsec à l'aide de la relation (démontrée précédemment):

$$M = m + 5 - 5 \log(d) \quad (48.56)$$

La figure ci-dessous représente la courbe période-luminosité des Céphéides.

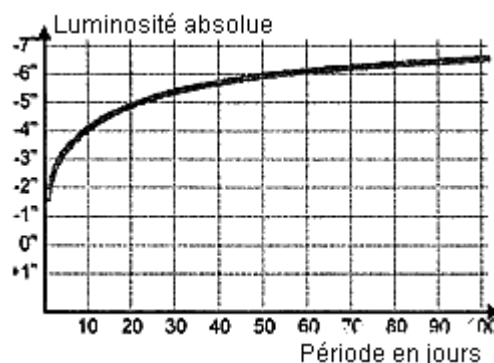


Figure: 48.5 - Courbe période-luminosité des Céphéides

L'étalonnage de cette courbe ne peut se faire que par des mesures de parallaxe sur des Céphéides proches. Il n'en existe malheureusement pas d'assez rapprochées pour qu'il soit possible d'utiliser la parallaxe annuelle. Il faut avoir recours à la parallaxe secondaire qui est basée sur le mouvement du Soleil dans la galaxie.

Exemple:

Nous repérons une Céphéide grâce à son type de classe spectrale. Sa période est de 50 jours et sa magnitude apparente $m = 3$. La figure précédente donne, pour cette étoile, une magnitude absolue $M = -6$.

En appliquant ensuite la formule donnée précédemment, nous trouvons:

$$d = 10^{(m+5-M)/5} = 10^{2.8} = 630 [pc] \quad (48.57)$$

Cette Céphéide est donc éloignée de 630 [pc].

Grâce aux propriétés des Céphéides, nous disposons d'un instrument de mesure qui porte jusqu'à quelques dizaines de millions d'années-lumière. Il est donc applicable au-delà de notre Voie lactée jusqu'aux galaxies proches comme les membres du groupe local. Au-delà, il devient difficile de détecter des Céphéides aux caractéristiques connues.

Les étoiles RR Lyrae sont quant à elles des étoiles peu massives et vieilles. Leur période d'oscillation est inférieure à un jour. Contrairement aux Céphéides, elles ont toutes la même luminosité moyenne (magnitude absolue de 0.5), environ 100 fois celle du Soleil.

Il existe encore une certaine quantité d'étoiles variables différentes (variables à éclipses, des variables explosives, variables binaires, etc.) pour lesquelles nous pouvons trouver une source abondante d'information sur l'Internet.

Il existe d'autres méthodes plus connues de mesure des distances que celle des Céphéides ou de l'effet Doppler (voir plus loin pour les développements concernant l'effet Doppler):

1.5. PARALLAXE TRIGONOMÉTRIQUE

La méthode de parallaxe trigonométrique est très simple (mais délicate à mettre en oeuvre à la surface de notre planète pour les étoiles très distantes). Tout astronome amateur constate la fuite de l'étoile qu'il observe dans son oculaire. Ce mouvement se nomme "**mouvement diurne**". Il est dû à la rotation de la Terre sur elle-même. L'étoile est également animée d'un mouvement elliptique beaucoup moins facilement détectable: le "**mouvement parallactique**".

Il est dû, comme le suggère le schéma ci-contre, à la rotation de la Terre autour du Soleil. Nous mesurons donc l'angle Π :

$$\frac{a}{d} = \tan(\Pi) \quad (48.58)$$

si l'angle est faible (ce qui est très fréquemment le cas étant donnée la distance des étoiles), nous pouvons prendre le premier terme du développement de Taylor (cf. chapitre sur les Suites Et Séries) de la fonction tangente:

$$\tan(x) = x + \frac{x^3}{3} + \frac{2x^5}{15} + \dots \cong x \quad (48.59)$$

Ce qui nous permet d'écrire:

$$d \cong \frac{a}{\Pi} \quad (48.60)$$

où d est la distance du Soleil à l'étoile et a celle de la Terre au Soleil comme représenté ci-dessous:

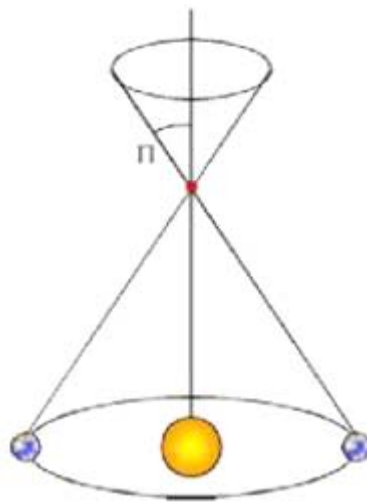


Figure: 48.6 - Principe de la parallaxe trigonométrique

1.6. L'EFFET DOPPLER-FIZEAU RELATIVISTE

L'effet Doppler-Fizeau est le décalage entre la fréquence de l'onde émise et de l'onde reçue lorsque l'émetteur et le récepteur sont en mouvement l'un par rapport à l'autre. C'est une technique très utilisée en astrophysique pour calculer la distance d'un astre en supposant sa longueur d'onde d'émission connue (ou estimée) et en mesurant sa longueur d'onde reçue ou encore pour mesurer la vitesse de rotation (vitesse radiale) des étoiles en observant très précisément et successivement leurs bords opposés et en mesurant le décalage du spectre obtenu.

En ce début de 21ème siècle la précision et la finesse des mesures de spectres a atteint un niveau tel qu'il permet même d'observer des variations minimales de la distance des étoiles et ainsi de spéculer sur d'éventuels satellites planétaires (ceci ne pouvant fonctionner que si le plan de l'orbite passe par la Terre).

L'effet Doppler des ondes électromagnétiques doit être discuté indépendamment de l'effet Doppler acoustique (appelé également "effet Doppler-Fizeau Galiléen") démontré dans le chapitre de Musique Mathématique. Premièrement parce que les ondes électromagnétiques ne consistent pas en un mouvement de matière et que par conséquent la vitesse de la source par rapport au milieu n'entre pas dans la discussion, ensuite parce que leur vitesse de propagation est c (la vitesse de la lumière) et reste la même pour tous les observateurs indépendamment de leurs mouvements relatifs. L'effet Doppler pour les ondes électromagnétiques se calcule donc nécessairement au moyen du principe de relativité et est symétrique par rapport aux mouvement relatifs de la source et de l'observateur (contrairement au cas acoustique).

Pour un observateur dans un repère d'inertie, une onde électromagnétique plane et harmonique peut être décrite par une fonction de la forme:

$$\sin(k \cdot x - \omega \cdot t) \quad (48.61)$$

multipliée par un facteur d'amplitude approprié. Pour un observateur attaché à un autre repère d'inertie, les coordonnées x et t doivent être remplacées par x' et t' , obtenues par la transformation de Lorentz (cf. chapitre de Relativité Restreinte), et celui-ci écrira par conséquent pour sa description la fonction:

$$\sin(k' \cdot x' - \omega' \cdot t') \quad (48.62)$$

où k' et ω' ne sont pas nécessairement les mêmes que pour l'autre observateur (justement c'est ce que nous cherchons à savoir). Par ailleurs, le principe de relativité nous a permis de démontrer dans le chapitre de Relativité Restreinte que:

$$\frac{\partial}{\partial t^2} - c^2 \frac{\partial^2}{\partial x^2} = \frac{\partial}{\partial t'^2} - c^2 \frac{\partial^2}{\partial x'^2} \quad (48.63)$$

Ce qui impose que l'expression:

$$k \cdot x - \omega \cdot t \quad (48.64)$$

reste invariante quand nous passons d'un observateur d'inertie à un autre. Nous aurons alors:

$$k \cdot x - \omega \cdot t = k' \cdot x' - \omega' \cdot t' \quad (48.65)$$

En utilisant les relations de transformation de Lorentz (cf. chapitre de Relativité Restreinte), nous avons:

$$k \cdot \frac{x' + vt'}{\sqrt{1 - v^2/c^2}} - \omega \cdot \frac{t' + v \cdot x'/c^2}{\sqrt{1 - v^2/c^2}} = k' \cdot x' - \omega' \cdot t' \quad (48.66)$$

Par identification, il vient immédiatement:

$$k' = \frac{k - \omega \cdot v/c^2}{\sqrt{1 - v^2/c^2}}, \omega' = \frac{\omega - k \cdot v}{\sqrt{1 - v^2/c^2}} \quad (48.67)$$

si nous tenons compte que:

$$\omega = c \cdot k \quad (48.68)$$

dans le cas des ondes électromagnétiques, nous pouvons écrire chacune de ces relations sous la forme:

$$k' = k \cdot \frac{1 - v/c}{\sqrt{1 - v^2/c^2}}, \omega' = \omega \cdot \frac{1 - v/c}{\sqrt{1 - v^2/c^2}} \quad (48.69)$$

Le rapport:

$$\frac{1 - v/c}{\sqrt{1 - v^2/c^2}} \quad (48.70)$$

donne le "**décalage spectral**" noté Z pour un mouvement de l'observateur par rapport à la source suivant la direction de propagation.

Par ailleurs, la dernière relation avec les pulsations est plus souvent donnée dans la littérature sous la forme suivante:

$$2\pi f = 2\pi f' \frac{1 - v/c}{\sqrt{1 - v^2/c^2}} = 2\pi f' \frac{1 - v/c}{\sqrt{(1 - v/c)(1 + v/c)}} = 2\pi f' \frac{\sqrt{1 - v/c}}{\sqrt{1 + v/c}} \quad (48.71)$$

Ce qui se note plus couramment sous la forme suivante:

$$f = f' \frac{\sqrt{1-\beta}}{\sqrt{1+\beta}} \quad (48.72)$$

Il faut bien se rappeler que le décalage de pulsation (et donc de fréquence) qui a lieu ici est dû à un mouvement relatif de l'observateur par rapport à la source et non à autre chose (ou réciproquement de la source par rapport à l'observateur). Effectivement, lors de notre étude la relativité générale (cf. [chapitre de Relativité Générale](#)), nous verrons qu'il y a également superposition d'un décalage à cause du champ gravitationnel environnant l'émetteur qui sera étudié comme étant causé par la courbure de l'espace-temps.

Enfin, pour les sceptiques qui veulent vérifier d'une autre manière que le phénomène Doppler est bien symétrique contrairement à l'effet Doppler acoustique démontré dans le chapitre de Musique Mathématique, voici une autre approche:

Considérons d'abord que c'est la source qui s'éloigne. Si on la calculait par la relation classique démontrée dans le chapitre de Musique Mathématique, la fréquence du signal à la réception serait:

$$f_{\text{éloignement}} = f \frac{1}{1 + \frac{v_{\text{source}}}{v_{\text{phase}}}} = f \frac{1}{1 + \frac{v}{c}} = \frac{1}{\Delta t} \frac{1}{1 + \beta} \quad (48.73)$$

et il faut prendre en compte la dilatation du temps pour f avec (cf. [chapitre de Relativité Restreinte](#)):

$$\Delta t = \frac{\Delta t'}{\sqrt{1 - (v/c)^2}}$$

car l'intervalle de temps de l'observateur fixe est plus long que celui de la source (le temps passe plus vite pour l'observateur fixe).

Il vient alors:

$$\begin{aligned} f_{\text{éloignement}} &= \frac{\sqrt{1-\beta^2}}{\Delta t'} \frac{1}{1+\beta} = f' \frac{\sqrt{1-\beta^2}}{1+\beta} = f' \frac{\sqrt{(1+\beta)(1-\beta)}}{1+\beta} = f' \sqrt{\frac{(1+\beta)(1-\beta)}{(1+\beta)^2}} \\ &= f' \sqrt{\frac{1-\beta}{1+\beta}} \end{aligned} \quad (48.74)$$

et dans le cas où c'est l'observateur qui s'éloigne de la source nous avons démontré dans le chapitre de Musique Mathématique que:

$$f_{\text{éloignement}} = f' \left(1 - \frac{v_{\text{obs}}}{v_{\text{phase}}} \right) = f' \left(1 - \frac{v}{c} \right) = \frac{1}{\Delta t'} (1 - \beta) \quad (48.75)$$

et selon les mêmes considérations, nous avons (simplement cette fois-ci c'est l'observateur qui se

déplace):

$$\begin{aligned}
 f_{\text{éloignement}} &= \frac{1}{\Delta t \sqrt{1-\beta^2}} (1-\beta) = \frac{1}{\Delta t} \frac{(1-\beta)}{\Delta t \sqrt{1-\beta^2}} = f \frac{(1-\beta)}{\sqrt{1-\beta^2}} = f \frac{(1-\beta)}{\sqrt{(1+\beta)(1-\beta)}} \\
 &= f \frac{\sqrt{(1-\beta)^2}}{\sqrt{(1+\beta)(1-\beta)}} = f \frac{\sqrt{(1-\beta)^2}}{\sqrt{(1+\beta)(1-\beta)}} = f \sqrt{\frac{(1-\beta)^2}{(1+\beta)(1-\beta)}} = f \sqrt{\frac{1-\beta}{1+\beta}}
 \end{aligned} \quad (48.76)$$

Les deux relations sont bien symétriques dans le cas relativiste (donc dans le cadre de l'électrodynamique)!

Remarque: En ce qui concerne la transformation de l'amplitude du champ électrique et du champ magnétique, il faut utiliser le tenseur de Maxwell démontré dans le chapitre de Relativité Restreinte.

Un très bon exemple de l'application de l'effet Doppler consiste à étudier les limites données par la mesure de la vitesse apparente. Voyons de quoi il s'agit:

1.6.1. VITESSE APPARENTE

En mesurant la vitesse apparente de déplacement d'objets très rapides dans le ciel (jets de plasma, etc.), les astrophysiciens ont obtenu des vitesses apparentes de déplacement supérieures à la vitesse de la lumière dans le vide!

Au fait, il s'agit d'une illusion qui peut se produire si la vitesse de l'objet est très proche de celle de la lumière qu'il émet, donc assez proche de c .

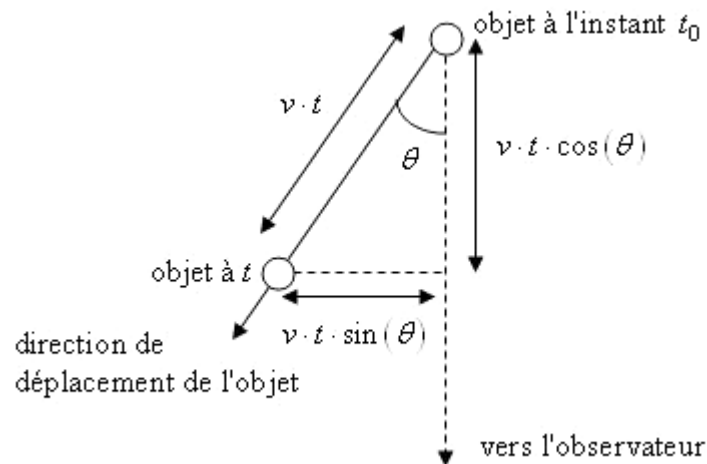


Figure: 48.7 - Principe de mise en situation de la vitesse apparente

L'objet émet de la lumière à l'instant t_0 , celle-ci ne nous atteint pas instantanément mais doit parcourir une distance d pour arriver à nous. Nous recevons après le temps:

$$t_1 = \frac{d}{c} \quad (48.77)$$

L'objet lui, se déplace à la vitesse v suivant un angle noté θ avec la direction d'observation, donc à

l'instant t , l'objet s'est déplacé d'une distance $v \cdot t$. La lumière émise par l'objet à l'instant t doit parcourir la distance (application de Pythagore):

$$d' = \sqrt{(d - v \cdot t \cdot \cos \theta)^2 + (v \cdot t \cdot \sin \theta)^2} \quad (48.78)$$

pour nous arriver (l'objet s'est avancé de $v \cdot t \cdot \cos \theta$ dans la direction d'observation mais s'est éloigné de l'axe d'observation de la distance $v \cdot t \cdot \sin \theta$), nous recevons donc la lumière qui a été émise par l'objet à l'instant t après un temps t_2 :

$$t_2 = t + \frac{d'}{c} = t + \frac{\sqrt{(d - v \cdot t \cdot \cos \theta)^2 + (v \cdot t \cdot \sin \theta)^2}}{c} \quad (48.79)$$

Entre les deux positions de l'objet, il s'est écoulé la durée t mais, vu de l'observateur, l'intervalle de temps entre la réception des images de ces deux positions est:

$$t_2 - t_1 = t + \frac{\sqrt{(d - v \cdot t \cdot \cos \theta)^2 + (v \cdot t \cdot \sin \theta)^2}}{c} - \frac{d}{c} \quad (48.80)$$

différent de t .

Pour un intervalle de temps t petit, nous avons, en développement limité de Taylor:

$$\begin{aligned} t_2 - t_1 &= t + \frac{\sqrt{d^2 - 2 \cdot d \cdot v \cdot t \cdot \cos \theta + v^2 \cdot t^2 \cdot \cos^2 \theta + v^2 \cdot t^2 \cdot \sin^2 \theta}}{c} - \frac{d}{c} \\ &= t + \sqrt{\frac{d^2}{c^2} - \frac{2 \cdot d \cdot v \cdot t \cdot \cos \theta}{c^2} + \frac{v^2 t^2}{c^2}} - \frac{d}{c} \cong t + \sqrt{\frac{d^2}{c^2} - \frac{2 \cdot d \cdot v \cdot t \cdot \cos \theta}{c^2}} - \frac{d}{c} \\ &= t + \frac{d}{c} \sqrt{1 + \left(-\frac{2 \cdot v \cdot t \cdot \cos \theta}{d} \right)} - \frac{d}{c} \cong t + \frac{d}{c} \left(1 - \frac{1}{2} \frac{2 \cdot v \cdot t \cdot \cos \theta}{d} \right) - \frac{d}{c} \\ &= t - \frac{v \cdot t \cdot \cos \theta}{c} = t \left(1 - \frac{v}{c} \cos \theta \right) \end{aligned} \quad (48.81)$$

Pendant cet intervalle de temps, toujours vu de l'observateur, l'objet semble s'être déplacé sur le plan du ciel de $v \cdot t \cdot \sin \theta$.

Ainsi, la vitesse apparente de l'objet est:

$$v_{app} = \frac{v \cdot t \cdot \sin \theta}{t_2 - t_1} = \frac{v \sin \theta}{1 - \frac{v}{c} \cos \theta} \quad (48.82)$$

Si nous posons l'angle θ comme étant très proche d'un angle droit, nous avons alors le deuxième terme du dénominateur qui est très petit ce qui nous permet avec un développement de Taylor d'écrire une relation que l'on retrouve assez souvent dans les manuels scolaires des petites classes:

$$v_{app} \cong v \cdot \left(1 + \frac{v}{c} \right) \quad (48.83)$$

Cherchons le maximum de cette fonction pour comprendre comment une telle observation est possible en dérivant par rapport à θ et en cherchant pour quelle valeur la dérivée s'annule:

$$\frac{dv_{app}}{d\theta} = \frac{v \cos \theta}{1 - \frac{v}{c} \cos \theta} - \frac{v}{c} \frac{v \sin^2 \theta}{\left(1 - \frac{v}{c} \cos \theta\right)^2} = 0 \quad (48.84)$$

et cela s'annule après simplification du dénominateur pour:

$$v \cos \theta \left(1 - \frac{v}{c} \cos \theta\right) - \frac{v}{c} v \sin^2 \theta = v \cos \theta - \frac{v^2}{c} (\cos^2 \theta + \sin^2 \theta) = v \cos \theta - \frac{v^2}{c} = 0 \quad (48.85)$$

d'où:

$$\cos \theta = \frac{v}{c} \quad (48.86)$$

La vitesse apparente est alors:

$$v_{app} = \frac{v \sin \theta}{1 - \frac{v^2}{c^2}} = \frac{v \sqrt{1 - \cos^2 \theta}}{1 - \frac{v^2}{c^2}} = \frac{v \sqrt{1 - \frac{v^2}{c^2}}}{1 - \frac{v^2}{c^2}} = \frac{v}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (48.87)$$

et elle est égale ou supérieure à c si déjà:

$$c = \frac{v}{\sqrt{1 - \frac{v^2}{c^2}}} \Rightarrow c^2 \left(1 - \frac{v^2}{c^2}\right) = v^2 \Rightarrow c^2 - v^2 = v^2 \Rightarrow 2v^2 = c^2 \quad (48.88)$$

donc:

$$v \geq \frac{1}{\sqrt{2}} c \quad (48.89)$$

Nous voyons ainsi qu'il est possible d'observer des mouvements apparents plus rapides que la lumière, alors même que l'objet est très rapide, certes, mais plus lent que c . Comme il ne s'agit que d'une illusion, il n'y a pas de contradiction avec la théorie de la relativité.

En connaissant la vitesse de déplacement d'un astre obtenue à l'aide de l'effet Doppler et la vitesse apparente à l'aide des observations, il est alors facile pour les astrophysiciens de déterminer l'angle θ en faisant un peu d'algèbre élémentaire à partir de la relation ci-dessous:

$$v_{app} = \frac{v \cdot t \cdot \sin \theta}{t_2 - t_1} = \frac{v \sin \theta}{1 - \frac{v}{c} \cos \theta} \quad (48.90)$$

1.7. LIMITE DE CHANDRASEKHAR

Nous avons déjà déterminé dans le chapitre de Mécanique Classique le rayon de Schwarzschild (sous sa forme classique) qui exprime le rayon critique d'un corps pour que la vitesse de libération à sa surface soit égale à la vitesse de la lumière. Nous avons obtenu la relation ci-dessous qui exprimait typiquement le rayon que devrait avoir un astre donné pour avoir une vitesse de libération égale à celle de la lumière:

$$R = \frac{2Gm}{c^2} \quad (48.91)$$

Dans ce cas particulier l'astre est ce que nous avons appelé un "**Trou Noir**". Cependant, avant le trou noir, une étoile passe comme nous en avons parlé, par plusieurs étapes intermédiaires par lesquelles elle peut d'ailleurs se stabiliser. Ainsi, vous avez dû souvent lire dans la littérature que pour qu'une naine blanche s'effondre en étoile à neutrons, sa masse devait être supérieure à 1.4 masses solaires mais sans démonstration mathématique. Eh bien c'est ce que nous allons démontrer maintenant!

Nous allons introduire le sujet par l'étude de l'influence du principe d'incertitude sur la taille d'un système atomique (il en limite la dimension minimale). Cet exemple est fort puissant car il montre que le principe d'incertitude ne régit pas seulement le processus de la mesure mais aussi le comportement global des systèmes quantiques.

Le premier exemple que nous pouvons donner est celui de l'atome d'hydrogène, non que nous attendions un résultat nouveau de cette méthode d'analyse, mais plutôt parce que nous pouvons exposer l'usage du principe d'incertitude et insister sur sa signification.

Nous admettons que le proton, dont la masse l'emporte de beaucoup sur celle de l'électron, peut être considéré comme fixe. L'énergie de l'électron s'écrit:

$$E_{tot} = \frac{1}{2} m_e v^2 - \frac{q_e^2}{4\pi\epsilon_0 r} = \frac{p^2}{2m_e} - \frac{q_e^2}{4\pi\epsilon_0 r} \quad (48.92)$$

En physique classique, un système dont l'énergie est donnée par la relation précédente ne possède pas de minimum: si nous faisons tendre r vers zéro en conservant la forme circulaire de l'orbite, il est facile de voir que E_{tot} tend vers $-\infty$. En revanche, en physique quantique, cette limite n'a pas de sens: le principe d'incertitude s'y oppose.

Dans ce cas, la recherche du minimum E_1 de E_{tot} prend un sens, car une contrainte apparaît qui maintient ce minimum à une valeur finie. Elle se détermine en physique quantique (voir le modèle de Bohr de l'atome dans le chapitre de Physique Quantique Corpusculaire) et impose:

$$p \cdot r = n \cdot \hbar = \hbar \quad \text{où } n \in \mathbb{N} \quad (48.93)$$

Cependant, cette relation mise à part, si le rayon r de l'atome devient trop faible sous des contraintes extérieures (attention! nous nous affranchissons des orbites quantifiées du modèle de Bohr de l'atome qui impose une contrainte à p) la quantité de mouvement p de l'électron ne peut être inférieure à l'incertitude Δp qu'impose le principe d'incertitude de Heisenberg, dès lors que Δx est de l'ordre du rayon r de l'atome. La forme même de la relation précédente limite la portée de la méthode: nous ne pouvons espérer déterminer mieux qu'un ordre de grandeur du minimum de E_{tot} .

Afin d'évaluer le minimum E_1 de l'énergie totale, que nous interprétons comme l'état fondamental de l'atome d'hydrogène, nous calculons le minimum de E_{tot} en éliminant p de l'expression:

$$E_{tot} = \frac{p^2}{2m_e} - \frac{q_e^2}{4\pi\epsilon_0 r} \text{ par } p \cdot r = n \cdot \hbar = \hbar \quad (48.94)$$

Nous obtenons:

$$E_{tot} \equiv E(r) = \frac{\hbar^2}{2m_e r^2} - \frac{q_e^2}{4\pi\epsilon_0 r} \quad (48.95)$$

Le rayon r_1 de l'atome dans l'état fondamental est la valeur de r qui donne à $E(r)$ sa valeur minimale:

$$\left. \frac{dE(r)}{dr} \right|_{r=r_1} = -\frac{\hbar^2}{m_e r_1^3} + \frac{q_e^2}{4\pi\epsilon_0 r_1^2} = 0 \quad (48.96)$$

si bien que:

$$r_1 = \frac{\hbar^2 4\pi\epsilon_0}{m_e q_e^2} = \frac{\hbar^2 \epsilon}{\pi m_e q_e^2} \quad (48.97)$$

qui est l'expression bien connue du rayon de Bohr vue en physique quantique corpusculaire lors de l'étude du modèle de Bohr de l'atome. L'énergie E_1 de l'état fondamental est donc maintenant facilement calculable.

Le but de cet exemple est de montrer qu'avec le principe d'incertitude de Heisenberg nous pouvons par un raisonnement très simple retrouver l'état fondamental d'un système. C'est exactement de cette façon que nous allons procéder pour déterminer les conditions qui font qu'un astre se retrouve dans son état fondamental.

Attaquons-nous maintenant à l'étude d'une étoile. Schématiquement celle-ci se compose d'un mélange de deux gaz: celui qui est formé de noyaux d'une part, le gaz électronique de l'autre.

Au cours de la vie de l'étoile, de nombreux processus de fusion ont eu lieu. Ils ont accru à chaque fois la taille et la masse des noyaux; Fe (le fer) qui est abondant à la fin de la vie d'une étoile, contient en moyenne 56 nucléons (voir la partie physique atomique du site).

Ces noyaux sont de nature chimique ou isotopique variée. Comme ils sont peu nombreux en comparaison des électrons, leur pression est celle d'un gaz classique chargé, neutralisé par la présence des électrons: elle peut être ignorée, et ce d'autant plus que la température est nulle.

La charge électronique seule ne permettrait pas aux électrons de résister à l'effondrement d'une étoile puisque la matière stellaire est neutre. À très basse température, quand le carburant est épuisé, la seule pression que le gaz électronique puisse opposer à la pression hydrostatique due à la pesanteur est d'origine quantique.

En première approximation, les électrons exercent donc l'un sur l'autre une répulsion apparente qui n'est pas d'origine Coulombienne (principe d'exclusion de Pauli). En première approximation, ils

obéissent à une relation analogue à celle de l'électron atomique et qui s'écrit dans le cas minimal (ou maximal de pression):

$$p \cdot d \cong \hbar \quad (48.98)$$

où d est la distance moyenne qui sépare deux électrons voisins.

À température $T = 0 [K]$, l'équilibre est atteint quand l'énergie (la matière de l'astre) totale du système est minimale.

Que se passe-t-il si nous essayons d'évaluer la variation du rayon r_{NB} de la Naine Blanche en fonction de sa masse m_{NB} ?

L'énergie potentielle gravifique d'une étoile est donnée en bonne approximation par (cf. chapitre de Mécanique Classique):

$$E_{pot} \cong -G \frac{m_{NB}^2}{r_{NB}} \quad (48.99)$$

m_{NB} étant approximativement donnée par:

$$m_{NB} = m_p N \quad (48.100)$$

où m_p est la masse du proton et N le nombre de nucléons que contient l'étoile: la contribution des électrons à la masse de l'astre est négligeable et il n'y a pas lieu de distinguer entre la masse du neutron et celle du proton, presque identiques.

La seconde contribution à l'énergie est essentiellement celle du gaz électronique dégénéré (la dégénérescence correspond à l'existence de plusieurs états ayant la même énergie), d'origine cinétique. Nous pourrions être tentés d'écrire simplement (en supposant que le nombre d'électrons est égal au nombre de nucléons puisque nous sommes pour rappel dans l'hypothèse simplificatrice d'un gaz d'hydrogène):

$$E_{cin} = N \frac{1}{2} m_e v^2 = \frac{N p^2}{2 m_e} \quad (48.101)$$

Cette manière de faire conduit à une impasse. Si nous exigeons que la somme $E_{pot} + E_{cin}$ atteigne une valeur minimale, nous aboutissons à une valeur du rayon de l'étoile tellement faible que, par application de la relation:

$$p \cdot d \cong \hbar \quad (48.102)$$

la vitesse moyenne des électrons v dépasserait celle de la lumière!

Pour éviter cette contradiction, nous devons recourir à la mécanique relativiste qui nous a montré que, dans ce cas (cf. chapitre de Relativité Restreinte), nous pouvons exprimer l'énergie totale comme:

$$E = c \sqrt{m_{0e}^2 c^2 + p^2} \quad (48.103)$$

si la valeur numérique de l'énergie cinétique l'emporte considérablement sur l'énergie de repos, nous avons alors:

$$E = Nc\sqrt{m_{0e}^2 c^2 + p^2} = N \cdot (E_{cin} + m_{0e} c^2) \quad (48.104)$$

et donc:

$$E \cong E_{cin} = N \cdot m_{0e} c^2 = N \cdot p \cdot c \quad (48.105)$$

La distance moyenne d entre électrons s'évalue en supposant que l'étoile est homogène, approximation suffisante dès lors que nous cherchons l'ordre de grandeur d'une moyenne. Nous simplifions encore la géométrie en admettant que chaque électron est entouré d'un domaine sphérique de rayon d dans lequel il n'y a pas d'autre électron de même spin et où nous ne pouvons compter qu'un électron de spin opposé. Dès lors:

$$d^3 \cong \frac{r_{NB}^3}{N} \quad (48.106)$$

Il reste à évaluer le minimum de la somme:

$$E_{tot} = E_{pot} + E_{cin} \cong -G \frac{m_p^2 N^2}{r_{NB}} + Npc \quad (48.107)$$

compte tenu de la condition:

$$m_{NB} = m_p N \quad (48.108)$$

Il vient encore:

$$p \cong \frac{\hbar}{d} \cong \frac{\hbar N^{1/3}}{r_{NB}} \quad (48.109)$$

puis:

$$E_{tot} = E_{NB}(r_{NB}) \cong -G \frac{m_p^2 N^2}{r_{NB}} + Nc \frac{\hbar}{r_{NB}} N^{1/3} = \frac{N^{4/3}}{r_{NB}} (c\hbar - Gm_p^2 N^{2/3}) \quad (48.110)$$

que nous écrivons finalement:

$$E_{tot} = E_{NB}(r_{NB}) \cong \frac{N^{4/3}}{r_{NB}} (c\hbar - Gm_p^{4/3} m_{NB}^{2/3}) \quad (48.111)$$

Face à ce résultat, nous sommes confrontés à une situation inattendue:

Si le facteur:

$$K = c\hbar - Gm_p^{4/3} m_{NB}^{2/3} \quad (48.112)$$

est positif, alors l'énergie totale de la naine blanche l'est aussi, ce qui signifie que le système n'est pas lié: l'étoile est totalement instable (elle n'a pas atteint son seuil d'énergie minimal). Elle ne peut réduire son énergie qu'en augmentant sans limites son rayon r .

Nous voyons que le facteur K est négatif si:

$$m_{NB} > \left(\frac{c\hbar}{G} \right)^{3/2} \frac{1}{m_p^2} \quad (48.113)$$

Si la Naine Blanche dépasse cette masse alors nous ne pouvons plus traiter le problème avec les équations précédentes. Elle satisfait alors aux équations régissant un astre composé de neutrons uniquement (étoile à neutrons) et ceci constitue alors un autre problème que nous n'aborderons pas ici pour l'instant.

La masse (approximative) de la fameuse "**limite de Chandrasekhar**" est donc donnée par:

$$m_{NB} = m_C = \left(\frac{c\hbar}{G} \right)^{3/2} \frac{1}{m_p^2} = 1.4m_{\odot} \quad (48.114)$$

Elle constitue la masse au-delà de laquelle une naine blanche s'effondre en étoile à neutrons.

Conventionnellement, les astrophysiciens associent cette valeur limite à un facteur multiplicateur de la masse du Soleil $m_{\odot}$. Nous avons effectivement (numériquement):

$$m_{NB} = 1.4m_{\odot} \quad (48.115)$$

1.8. LIMITE DE RUPTURE DE ROTATION

Faisons l'hypothèse simplificatrice que la vitesse limite de rotation possible d'un astre (planète ou étoile) est celle qui équilibre la force centrifuge et force gravitationnelle à la surface de l'astre telle que nous soyons amenés à écrire (cf. chapitre de Mécanique Classique):

$$m \frac{v^2}{R} = \frac{GMm}{R^2} \quad (48.116)$$

Poser cette relation suppose évidemment qu'il n'y ait aucune liaison autre que la gravité qui intervient dans la cohésion interne de l'astre. Donc les valeurs de temps de rotation que nous allons obtenir représentent une borne supérieure (dans le sens que la valeur réelle est probablement plus petite).

Il vient alors de la relation précédente:

$$v = \sqrt{\frac{GM}{R}} \quad (48.117)$$

Pour obtenir le temps de rotation auquel cela correspond il suffit de diviser le périmètre à l'équateur par cette vitesse:

$$t_r = \frac{2\pi R}{v} = \frac{2\pi R}{\sqrt{\frac{GM}{R}}} = \frac{2\pi R}{\sqrt{\frac{6.674 \cdot 10^{-11} M}{R}}} \quad (48.118)$$

Ainsi, pour la Terre, nous avons comme période de rotation limite avant rupture:

$$t_{r, Terre} = \frac{2\pi \cdot 6'378'137}{\sqrt{\frac{6.674 \cdot 10^{-11} \cdot 5.9736 \cdot 10^{24}}{6'378'137}}} \cong 84 \text{ [min.]} \quad (48.119)$$

Pour le Soleil:

$$t_{r, Soleil} = \frac{2\pi \cdot 1'392'000'000}{\sqrt{\frac{6.674 \cdot 10^{-11} \cdot 1.9891 \cdot 10^{30}}{1'392'000'000}}} \cong 7.8 \text{ [hrs.]} \quad (48.120)$$

Maintenant considérons le cas du pulsar NP0532 qui a une rotation de 33 millisecondes. Nous souhaiterions en déterminer le rayon. Nous avons alors en utilisant les relations précédentes:

$$R = \left(\frac{t_r \sqrt{GM}}{2\pi} \right)^{2/3} \quad (48.121)$$

En utilisant la relation théorique de la masse limite de Chandrasekhar (puisque'un pulsar est une étoile à neutrons tournant rapidement sur elle-même):

$$m_{NB} = m_C = \left(\frac{c \hbar}{G} \right)^{3/2} \frac{1}{m_p^2} = 1.4 m_\odot \quad (48.122)$$

Nous avons alors pour le rayon de plus petit pulsar possible selon ces hypothèses:

$$R = \left(\frac{t_r \sqrt{G \cdot 1.4 m_\odot}}{2\pi} \right)^{2/3} = \left(\frac{33 \cdot 10^{-3} \sqrt{6.674 \cdot 10^{-11} \cdot 1.4 \cdot 1.9891 \cdot 10^{30}}}{2\pi} \right)^{2/3} \cong 172 \text{ [km]} \quad (48.123)$$

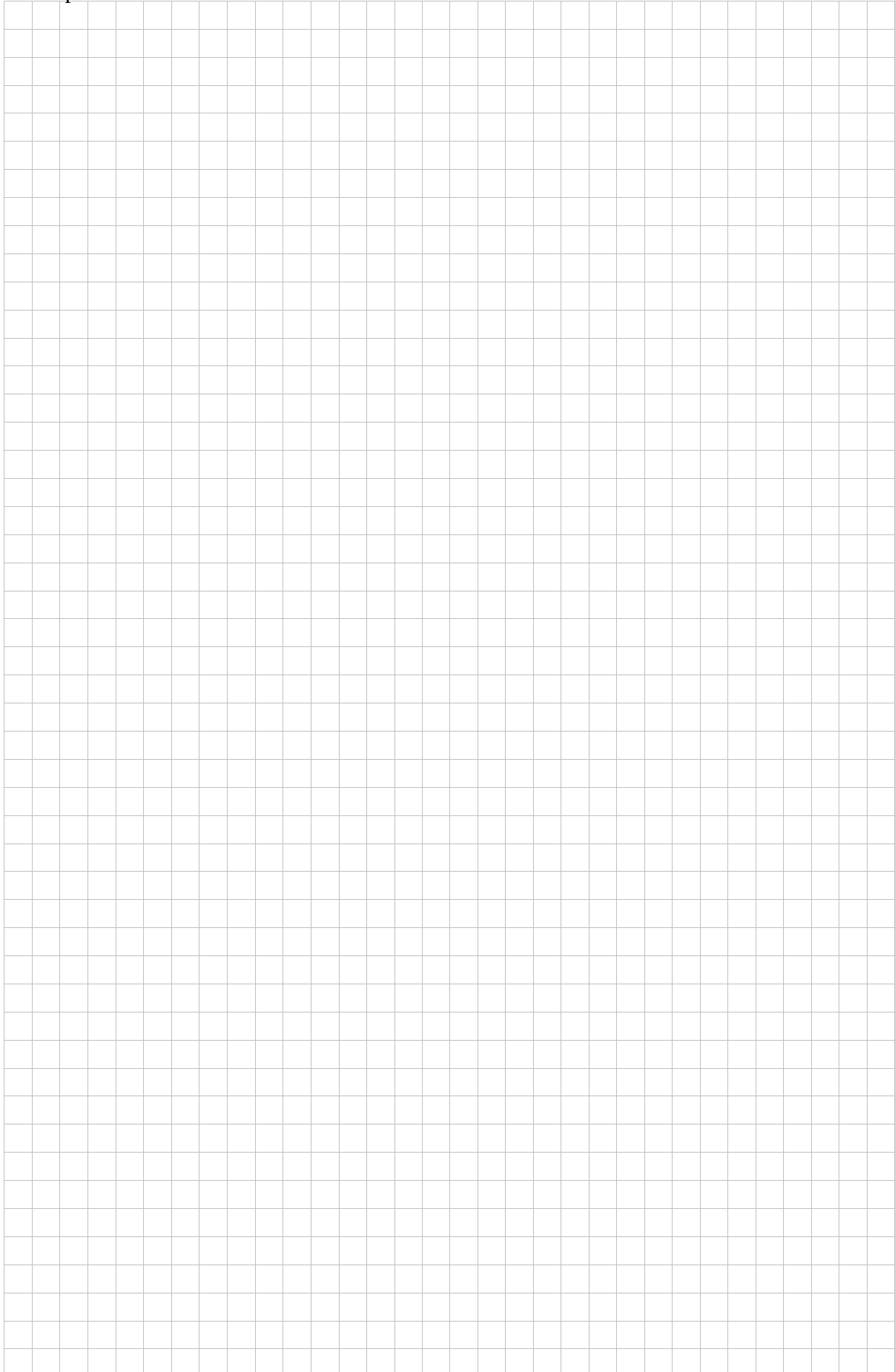
Avec le pulsar milliseconde PSR J1748-2446ad ayant une période de 1.39 millisecondes nous tombons alors sur:

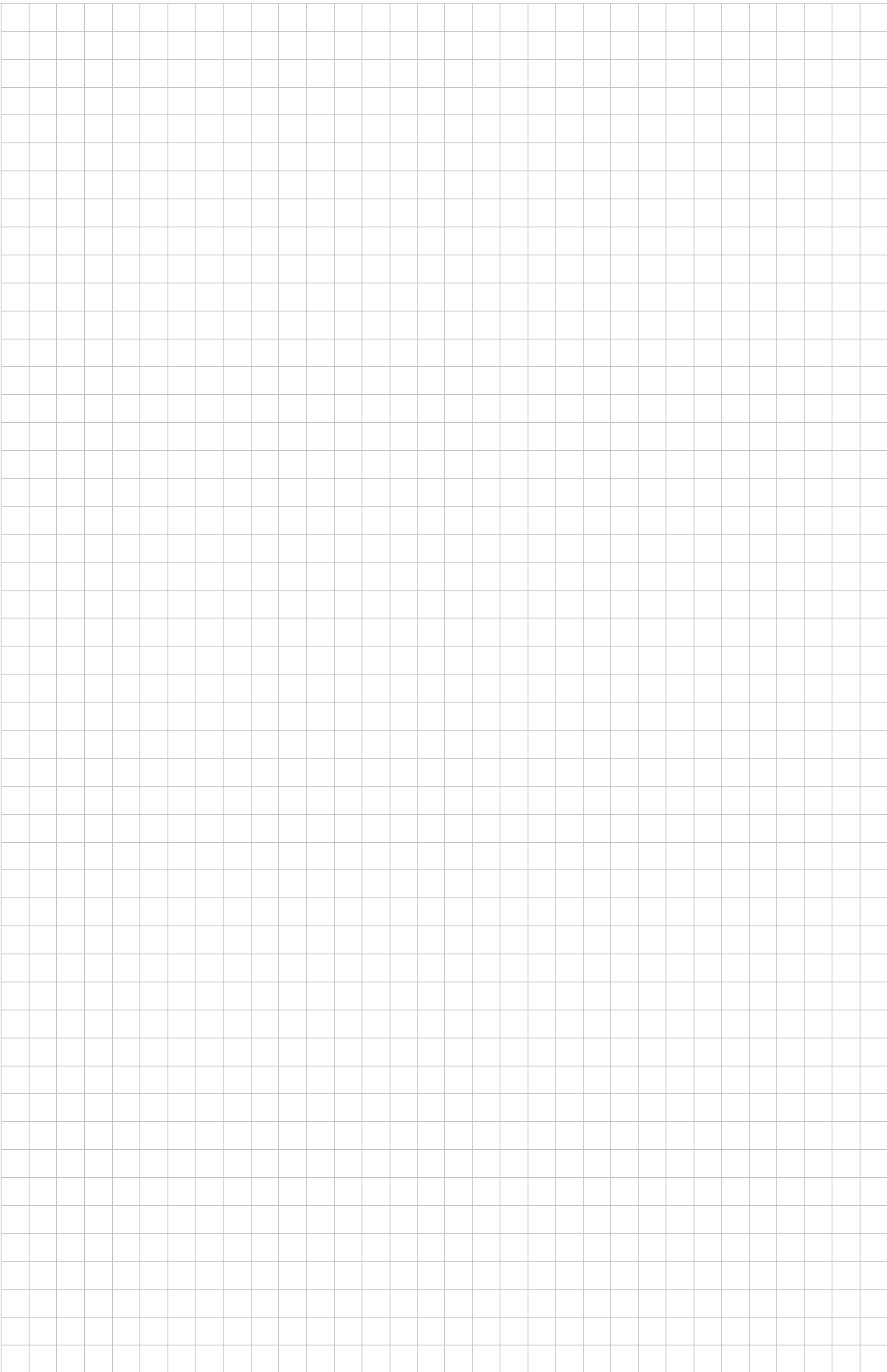
$$R = \left(\frac{t_r \sqrt{G \cdot 1.4 m_\odot}}{2\pi} \right)^{2/3} \cong 21 \text{ [km]} \quad (48.124)$$

ce qui est remarquable (même s'il s'agit d'une approximation) de penser qu'une telle masse peut être contenue dans un si petit rayon. À noter que pour ce dernier cela correspond à une densité de:

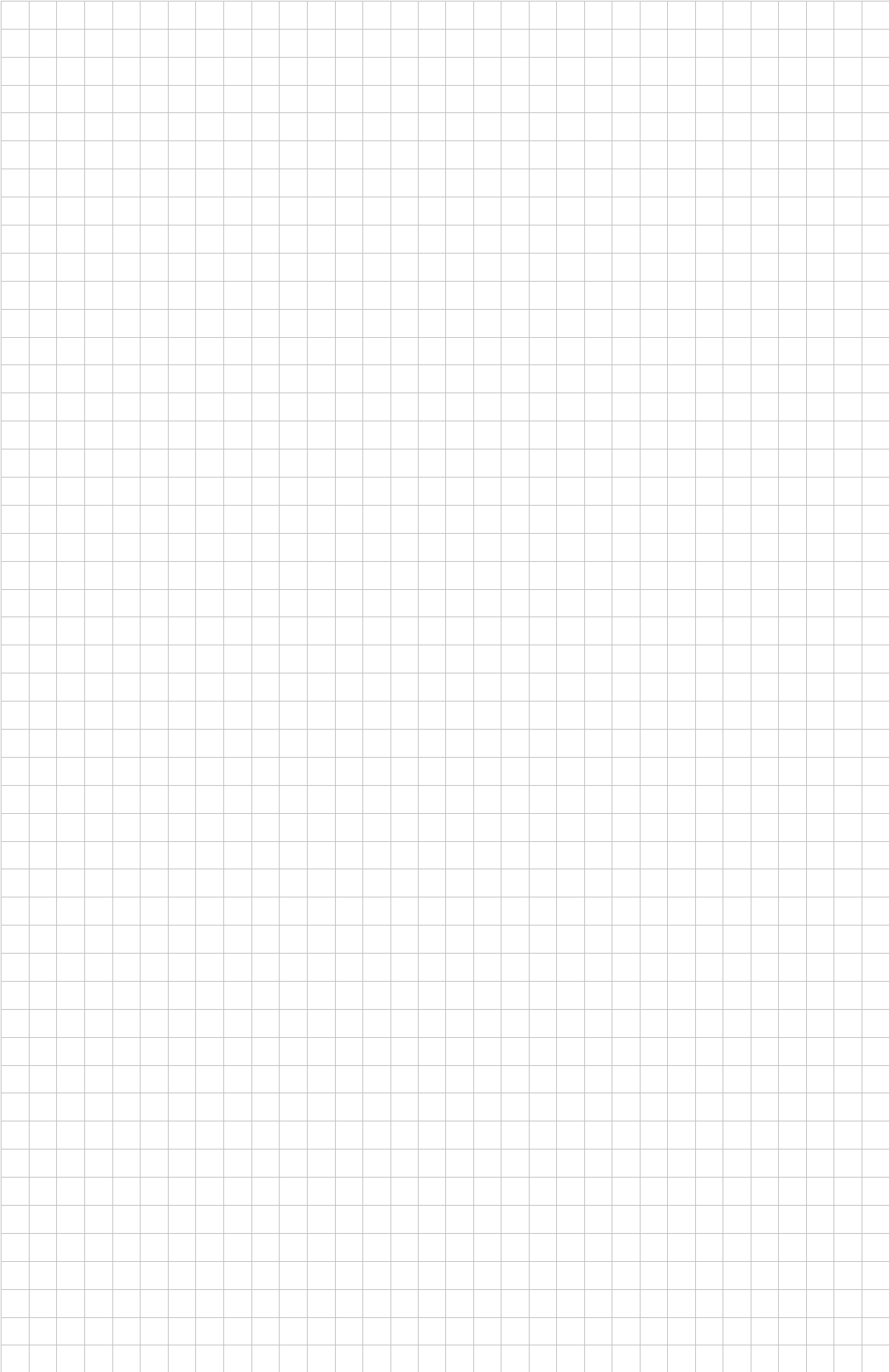
$$\rho = \frac{1.4 m_\odot}{\frac{4}{3} \pi R^3} \cong 15 \cdot 10^{14} \left[\frac{\text{kg}}{\text{mm}^3} \right] \quad (48.125)$$

Notes personnelles:









49. RELATIVITÉ RESTREINTE

Nous avons toujours considéré jusqu'à maintenant lors de tous nos développements que les interactions (relations de cause à effet) entre les corps se faisaient instantanément, ainsi que l'observation d'un phénomène avait lieu instantanément après que celui-ci avait eu lieu. Or, deux physiciens (Michelson et Morley) au cours d'une expérience découvrirent quelque chose qui allait changer radicalement toute la physique classique: la vitesse (célérité) de la lumière était invariante (constante) quel que soit le mouvement que l'on avait par rapport à elle !

Cette observation est d'autant plus importante que nous savons que c'est la lumière qui nous permet de percevoir et de ressentir les choses. Il convient également de prendre en considération que les champs électrostatique et magnétique sont, comme nous l'avons vu dans le chapitre de Physique Quantique Des Champs, véhiculés par le vecteur d'interaction qu'est le photon qui se déplace à la vitesse finie de la lumière c . Cette constatation nous permet aussi de supposer que le champ gravitationnel finalement a aussi un vecteur d'interaction (qui serait le "graviton" dont l'existence semble prouvée indirectement) qui se propage à la vitesse de la lumière. Il convient dès lors de prendre en compte cette non-instantanéité et les conséquences que cela entraîne dans les phénomènes observés pour départager finalement ce qui est réellement de ce qui semble être.

Avant de nous attaquer aux calculs, il nous faut définir un petit peu ce qui va être étudié dans ce chapitre (qui ne s'applique de loin pas qu'à la cosmologie mais bon... il me semblait préférable de le mettre dans cette section plutôt que dans celle de Mécanique).

Définition: La "**relativité restreinte**" est une théorie confinée aux référentiels inertiels isolés (Galiléens), c'est-à-dire à l'étude de référentiels animés d'un mouvement rectiligne uniforme (inertiels). La raison en sera donnée lors de l'énoncé du principe de relativité restreinte (voir plus bas).

 Remarques:

R1. Restreindre l'étude à des référentiels inertiels n'empêche bien évidemment pas qu'à l'intérieur de ceux-ci les corps peuvent être animés d'une vitesse uniforme ou non!

R2. La relativité générale a pour rôle de prendre en compte des référentiels non inertiels et dans n'importe quel système de coordonnées en faisant usage de la puissance du calcul tensoriel pour être applicable dans n'importe quel type d'espace (autre que plat donc !).

La relativité restreinte se base principalement sur trois concepts très importants:

1. Le postulat d'invariance (de la vitesse de la lumière).
2. Le principe cosmologique (voir plus bas).
3. Le principe de relativité restreinte (voir plus bas).

Il convient aussi de prévenir le lecteur que nous allons utiliser ici beaucoup de concepts vus dans les chapitres d'algèbre linéaire, calcul tensoriel, trigonométrie hyperbolique, calcul différentiel et intégral, mécanique analytique, mécanique classique, électrostatique, magnétostatique et électrodynamique. Il est fortement conseillé d'avoir parcouru ces différents sujets au risque de décrocher dans la lecture de ce qui va suivre.

1. PRINCIPES ET POSTULATS

Les lois physiques expriment des relations entre des grandeurs physiques fondamentales. Si les lois physiques sont invariantes par changement de référentiel Galiléen comme nous l'avons vu dans le chapitre de Mécanique Classique, il n'en est pas forcément de même des grandeurs physiques. Ces dernières peuvent se transformer d'un référentiel Galiléen à un autre selon une loi de transformation simple comme nous l'avons vu au chapitre de Mécanique Classique. Il en est de même en relativité restreinte, mais nous devons maintenant prendre en compte ce que nous avons négligé lors de notre étude des transformations de Galilée: le temps qui s'écoule n'est pas le même pour deux observateurs si la vitesse de la lumière est finie, par contre l'intervalle de temps est lui est conservé! (trivial)

1.1. POSTULAT D'INVARIANCE

Des mesures de laboratoire (expérience de Michelson-Morley comme nous en avons déjà fait mention) ont, depuis fort longtemps, montré que la vitesse c mesurée par un référentiel inertiel (en ligne droite et à vitesse constante) est bien constante quelle que soit sa vitesse d'entraînement. Nous devons alors postuler la propriété suivante:

Postulat d'invariance: la vitesse de la lumière (vecteur de transport de l'information) ne peut ni s'ajouter, ni se soustraire, à la vitesse d'entraînement du référentiel dans lequel nous la mesurons (plus clairement cela signifie que quelle que soit la vitesse à laquelle vous vous déplacerez vous mesurerez toujours la vitesse de lumière comme valant c numériquement constante et finie!).

Corollaire: le principe de relativité Galiléen (cf. chapitre de Mécanique Classique) selon ce postulat est complètement pris en défaut et il nous faut alors développer une nouvelle théorie qui prend en compte cette propriété de la lumière.

Remarque: Il est important de noter que nous considérons que la lumière est dans le cadre actuel de la relativité restreinte, le messenger de l'information d'un corps sur un autre!!!

1.2. PRINCIPE COSMOLOGIQUE

Nous supposons que notre position dans l'Univers est typique, non seulement dans l'espace comme l'affirme le modèle standard de l'Univers (cf. chapitre d'Astrophysique), mais aussi dans le temps. Ainsi, un astronome situé dans une galaxie éloignée doit observer les mêmes propriétés générales de l'Univers que nous, qu'il ait vécu un milliard d'années plus tôt, ou qu'il l'observe dans un milliard d'années.

En fait, il est relativement naturel d'aller plus loin et d'énoncer que: l'Univers présente le même aspect en chacun de ses points, c'est-à-dire qu'il est homogène. Cette homogénéité s'énonce donc sous la forme du "**principe cosmologique**".

Ce principe ne repose pas sur les observations, si fragmentaires par rapport à la démesure du cosmos qu'elles ne sauraient permettre d'établir sa validité. Il constitue bien un présupposé à toute étude physique de l'Univers. Sa raison d'être tient à son caractère, indispensable à toute cosmologie scientifique, et peut-être à une certaine réaction par rapport à l'ancienne vision géocentrique ou héliocentrique: il est supposé désormais qu'aucun lieu n'est privilégié dans le cosmos !

1.3. PRINCIPE DE RELATIVITÉ RESTREINTE

Rappelons (cf. chapitre de Mécanique Classique) que les transformations Galiléennes nous disent qu'aucun référentiel ne peut être considéré comme un référentiel absolu puisque les relations entre les grandeurs physiques sont identiques dans tous les référentiels Galiléens ("principe de relativité Galiléen"). Le mouvement Galiléen est donc relatif.

Au 20ème siècle les physiciens constatèrent qu'une importante catégorie de phénomènes physiques violait le principe de relativité Galiléen: les phénomènes électromagnétiques.

En appliquant les transformations Galiléennes aux équations de Maxwell, nous obtenons un jeu d'équations différent selon que l'observateur se trouve dans un référentiel fixe ou un référentiel mobile.

Effectivement, nous avons montré dans le chapitre d'Électrodynamique que l'équation de propagation du champ électrique ou magnétique s'écrivait sous la forme unidimensionnelle de l'équation de d'Alembert:

$$\left(\frac{\partial^2}{\partial x^2} - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) \psi = 0 \quad (49.1)$$

où ψ représente l'un quelconque des deux champs. Nous parlons alors très rarement de "l'équation de Hertz".

Nous avons aussi vu dans le chapitre de Mécanique Classique qu'un facteur important de la validité d'une théorie était l'invariance de l'expression de ses lois sous une transformation Galiléenne (transformée de Galilée) en posant:

$$x' = x - u_x t, y' = y, z' = z, t' = t \quad (49.2)$$

Nous avons également montré dans le chapitre de Calcul Différentiel Et Intégral que la différentielle totale d'une fonction s'écrivait (exemple à deux variables):

$$df(x', y') = \frac{\partial f}{\partial x'} dx' + \frac{\partial f}{\partial y'} dy' \quad (49.3)$$

Soit:

$$df(x', t') = \frac{\partial f}{\partial x'} dx' + \frac{\partial f}{\partial t'} dt' \quad (49.4)$$

Ce qui nous amène simplement à écrire:

$$\frac{\partial f}{\partial x} = \frac{1}{\partial x} \frac{\partial f}{\partial x'} dx' + \frac{1}{\partial x} \frac{\partial f}{\partial t'} dt' \quad (49.5)$$

Après élimination de f et en utilisant le théorème de Schwarz (cf. chapitre de Calcul Différentiel Et Intégral):

$$\begin{aligned}\frac{\partial}{\partial x} &= \frac{1}{\partial x} \frac{\partial}{\partial x'} dx' + \frac{1}{\partial x} \frac{\partial}{\partial t'} dt' = \frac{\partial}{\partial x'} \frac{\partial x'}{\partial x} + \frac{\partial}{\partial t'} \frac{\partial t'}{\partial x} = \frac{\partial}{\partial x'} \frac{\partial(x - u_x t)}{\partial x} + \frac{\partial}{\partial t'} \frac{\partial t}{\partial x} \\ &= \frac{\partial}{\partial x'} \frac{\partial(x - u_x t)}{\partial x} = \frac{\partial}{\partial x'}\end{aligned}\quad (49.6)$$

Si nous écrivons de même avec le temps:

$$\begin{aligned}\frac{\partial}{\partial t} &= \frac{1}{\partial t} \frac{\partial}{\partial x'} dx' + \frac{1}{\partial t} \frac{\partial}{\partial t'} dt' = \frac{\partial}{\partial x'} \frac{\partial x'}{\partial t} + \frac{\partial}{\partial t'} \frac{\partial t'}{\partial t} = \frac{\partial}{\partial x'} \frac{\partial(x - u_x t)}{\partial t} + \frac{\partial}{\partial t'} \frac{\partial t}{\partial t} \\ &= \frac{\partial}{\partial x'} \frac{\partial(x - u_x t)}{\partial t} + \frac{\partial}{\partial t'} = -u_x \frac{\partial}{\partial x'} + \frac{\partial}{\partial t'}\end{aligned}\quad (49.7)$$

En fin de compte la transformation Galiléenne de l'équation d'onde censée avoir une forme invariante devient:

$$\left(\frac{\partial^2}{\partial x'^2} - \frac{1}{c^2} \left(-u_x \frac{\partial}{\partial x'} + \frac{\partial}{\partial t'} \right)^2 \right) \psi = \left(\frac{\partial^2}{\partial x'^2} - \frac{u_x^2}{c^2} \frac{\partial^2}{\partial x'^2} + 2 \frac{u_x}{c^2} \frac{\partial^2}{\partial x' \partial t'} - \frac{1}{c^2} \frac{\partial^2}{\partial t'^2} \right) \psi = 0 \quad (49.8)$$

La forme de l'équation d'onde a donc été complètement altérée par la transformation. Au fait, nous savons que cela est dans un sens normal. Effectivement, après tout le champ magnétique, créé par des charges en mouvement disparaît quand nous utilisons un référentiel en mouvement avec les charges (ou inversement). Cependant, les champs électrique et magnétique ne se transforment pas correctement sous les transformations Galiléennes.

Pour fixer la situation, suite à ces deux exemples, nous avons trois hypothèses:

H1. Les équations de Maxwell sont fausses. Les équations correctes restent à être découvertes et devront être invariantes sous une transformation Galiléenne.

H2. L'invariance Galiléenne est valide pour la mécanique mais pas pour l'électromagnétisme (c'est la solution historique avant Einstein, un "éther" détermine l'existence d'une sorte de référentiel absolu où les équations de Maxwell ne changent pas).

H3. L'invariance Galiléenne est fausse. Il y a une invariance plus générale, qu'il reste à découvrir, qui préserve la forme des équations de Maxwell. La mécanique classique doit être reformulée telle qu'elle soit invariante sous cette nouvelle transformation.

Remarque: Il s'avère que les deux premières hypothèses sont exclues par les faits expérimentaux. De plus, les équations de Maxwell intégrant la vitesse de la lumière elle sont implicitement relativistes.

Albert Einstein n'admettait pas la violation du principe de relativité Galiléenne par l'électromagnétisme. De son point de vue, il fallait au contraire le généraliser à toutes les lois physiques.

Il postula donc que les lois physiques devaient être identiques dans tous les référentiels Galiléens, ce qui implique, implicitement, que du point de vue des lois physiques, il n'est pas possible de distinguer un référentiel Galiléen d'un autre. Ce résultat est plus fréquemment formulé sous la forme qu'aucun référentiel n'est privilégié. Ce principe fut baptisé "**principe de relativité**". En effet, cette relativité est restreinte aux cas des référentiels Galiléens (dits aussi "référentiels inertiels") exclusivement.

En d'autres termes, les lois physiques doivent rester inchangées après un changement de référentiel. Il nous faut donc déterminer de nouvelles transformations adéquates qui se substitueront aux transformations Galiléennes.

Dans le cas des référentiels non Galiléens les référentiels ne sont plus indiscernables. Effectivement, imaginons une personne se trouvant dans un train se déplaçant à une certaine vitesse constante et une autre personne sur la terre ferme. Chacun pourra alors dire que c'est l'autre qui est en mouvement (relatif) et ce indistinctement. Par contre, si le train se met à accélérer, bien que les deux individus puissent dire que c'est l'autre qui accélère, seul celui qui est dans le train ressentira l'effet de cette accélération... ainsi les référentiels ne sont plus indistinguables.

Einstein abolit ainsi aussi l'idée qu'il existe un point de référence absolu qui ne bouge pas et par rapport auquel on peut définir un temps absolu, une longueur absolue ou une masse absolue. On peut cependant définir un point de référence privilégié pour tout objet dans l'univers. Celui-ci est le référentiel se déplaçant à la même vitesse et dans la même direction que l'objet en question. Le temps mesuré dans ce référentiel privilégié est minimal et est appelé le "**temps propre**". Similairement, la dimension de l'objet y est maximale, c'est sa "**dimension propre**" ou "**distance propre**", et sa masse y est minimale, c'est sa "**masse au repos**" (nous ferons les développements mathématiques correspondants plus loin).

2. TRANSFORMATIONS DE LORENTZ

Pour que soit possible l'invariance de c (postulat d'invariance), nous devons admettre que le temps ne s'écoule pas de la même manière pour l'observateur immobile O que pour l'observateur O' dans un référentiel en translation uniforme en x (soit un référentiel inertiel) à vitesse relative (le terme "relative" est important!) v (attention ! la vitesse relative entre les référentiels est souvent notée u dans la littérature).

Remarque: Un cas particulier de disposition des référentiels dans laquelle les axes d'espaces sont parallèles amène à ce que nous appelons les: "**transformations de Lorentz pures**" ou encore "**transformations de Lorentz spéciales**" et le déplacement relatif selon un axe particulier est souvent appelé un: "**boost**".

Pour étudier le comportement des lois physiques, nous devons alors nous munir de deux horloges qui donnent t et t' (le référentiel qui contient son horloge/instrument de mesure est appelé "**référentiel propre**") .

Mettons en place l'expérience imaginaire suivante:

Lorsque les observateurs O et O' sont superposés, nous posons $t=0$ et $t'=0$ et nous émettons un flash lumineux dans la direction d'un point A repéré par $\vec{r}$ et $\vec{r}'$:

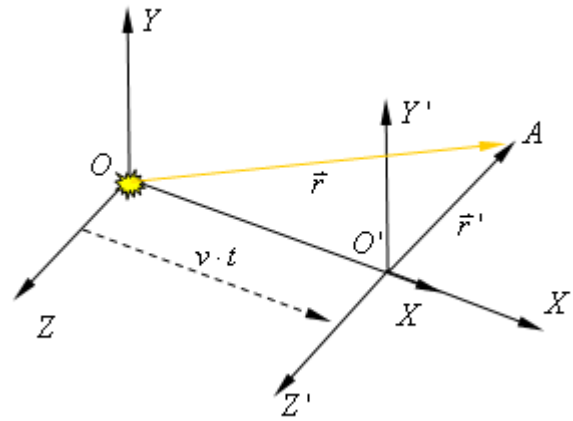


Figure: 49.1 - Configuration pour l'étude des effets relativistes

Il est évident que lorsque le flash arrivera en A, l'observateur O mesurera un temps t et O' un temps t' .

L'observateur O conclut dès lors:

$$r = c \cdot t = \sqrt{x^2 + y^2 + z^2} \Rightarrow x^2 + y^2 + z^2 = c^2 \cdot t^2 \quad (49.9)$$

L'observateur O' lui, conclut:

$$r' = c \cdot t' = \sqrt{x'^2 + y'^2 + z'^2} \Rightarrow x'^2 + y'^2 + z'^2 = c^2 \cdot t'^2 \quad (49.10)$$

Étant donné que le déplacement de O' ne se fait que selon l'axe OX , nous avons pour les deux observateurs:

$$y' = y, z' = z \quad (49.11)$$

De plus, si la trajectoire du rayon lumineux se confond dans Ox , nous avons:

$$y = y' = 0, z = z' = 0 \quad (49.12)$$

Ce qui nous donne dès lors $x^2 = c^2 t^2$ et $x'^2 = c^2 t'^2$ d'où:

$$x^2 - c^2 t^2 = 0 \quad \text{et} \quad x'^2 - c^2 t'^2 = 0 \quad (49.13)$$

Ces deux relations sont donc égales (nulles) en tout x, x', t, t' entre les deux observateurs. Ce sont les premiers "**invariants relativistes**" (valeurs égales quel que soit le référentiel) que nous retrouvons sous une forme plus généralisée lorsque qu'appliquée à tout l'espace:

$$\boxed{x^2 + y^2 + z^2 - c^2 t^2 = x'^2 + y'^2 + z'^2 - c^2 t'^2} \quad (49.14)$$

Il convient maintenant de se rappeler, que dans le modèle classique (relativité Galiléenne), nous aurions écrit que la position du point A pour l'observateur O à partir des informations données par O' serait $x = x' + v \cdot t'$ et réciproquement (cf. chapitre de Mécanique Classique) tel que:

$$x' = x - v \cdot t' = x - v \cdot t \quad (49.15)$$

Dans le modèle relativiste, nous devons par contre admettre que le temps t qui est en relation avec x n'est pas le même que t' qui est en relation avec x' , principe de relativité oblige (sinon quoi il serait donc impossible d'expliquer l'invariance de la vitesse de la lumière) !

Nous sommes alors amenés à poser la relation précédente sous la forme suivante:

$$x' = \lambda \cdot (x - v \cdot t) \quad (49.16)$$

où λ serait une valeur numérique à déterminer. Car pour expliquer la constance de la vitesse de la lumière c'est que l'espace doit s'ajuster en permanence en fonction de notre vitesse v . Ce qui est révolutionnaire comme nous l'avons déjà mentionné!

Remarque: Un lecteur nous a demandé pourquoi nous ne pourrions pas écrire la dernière relation sous la forme simplifiée suivante (en utilisant la relation $x = ct$ obtenue plus haut) si le point A se trouvait sur l'axe X :

$$x' = \lambda (x - vt) = \lambda \left(x - v \frac{x}{c} \right) = \lambda x \left(1 - \frac{v}{c} \right) \quad (49.17)$$

La seule raison tient au fait que plus tard nous allons introduire une écriture vectorielle (matricielle) de ce résultat faisant apparaître le concept de quadrivecteur et que c'est sous la première forme d'écriture (celle faisant explicitement référence au temps) que nous pouvons clairement faire apparaître le concept d'espace-temps.

De plus, si $t \neq t'$, nous devons aussi pouvoir exprimer t' comme fonction de t et de x sous une forme similaire:

$$t' = a \cdot (t - b \cdot x) \quad (49.18)$$

Résumons la forme du problème:

$$x' = \lambda \cdot (x - v \cdot t) \quad (49.19)$$

à déterminer λ . Et ensuite:

$$t' = a \cdot (t - b \cdot x) \quad (49.20)$$

à déterminer: a , b .

Nous cherchons alors à déterminer la relation permettant de connaître les valeurs des coefficients λ , a et b qui satisfont simultanément:

$$x^2 + y^2 + z^2 = c^2 \cdot t^2 \quad \text{et} \quad x'^2 + y'^2 + z'^2 = c^2 \cdot t'^2 \quad (49.21)$$

Compte tenu de ce qui précède et en se rappelant que $y' = y$ et $z' = z$, la dernière relation devient :

$$\lambda^2(x - v \cdot t)^2 + y^2 + z^2 = c^2 \cdot a^2 \cdot (t - b \cdot x)^2 \quad (49.22)$$

Distribuons:

$$(\lambda^2 - a^2 \cdot b^2 \cdot c^2) \cdot x^2 - 2 \cdot (\lambda^2 \cdot v - b \cdot a^2 \cdot c^2) \cdot x \cdot t + y^2 + z^2 = \left(a^2 - \frac{\lambda^2 \cdot v^2}{c^2} \right) \cdot c^2 \cdot t^2 \quad (49.23)$$

Pour satisfaire la relation:

$$x^2 + y^2 + z^2 = c^2 \cdot t^2 \quad (49.24)$$

Il faut que:

$$\lambda^2 - b^2 \cdot a^2 \cdot c^2 = 1 \quad (1)$$

$$\lambda^2 \cdot v - b \cdot a^2 \cdot c^2 = 0 \quad (2)$$

$$a^2 - \frac{\lambda^2 \cdot v^2}{c^2} = 1 \quad (3)$$

(49.25)

Il est facile de résoudre (2):

$$\lambda^2 = \frac{b \cdot a^2 \cdot c^2}{v} \quad (49.26)$$

Nous introduisons alors ce résultat dans (1) et (3) et nous arrivons à:

$$\lambda^2 - a^2 \cdot b^2 \cdot c^2 = \frac{b \cdot a^2 \cdot c^2}{v} (1 - b \cdot v) = 1 \quad (1')$$

$$a^2 - \frac{\lambda^2 \cdot v^2}{c^2} = a^2 (1 - b \cdot v) = 1 \quad (2')$$

(49.27)

Si nous divisons (1') par (2'), nous obtenons:

$$\frac{c^2 \cdot b}{v} = 1 \Rightarrow b = \frac{v}{c^2} \quad (49.28)$$

et en introduisant ce dernier résultat dans la relation :

$$a^2 (1 - b \cdot v) = 1 \quad (49.29)$$

nous obtenons le résultat remarquable suivant:

$$\alpha = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (49.30)$$

que nous notons souvent:

$$\alpha = \frac{1}{\sqrt{1 - \beta^2}} \quad (49.31)$$

et que nous appelons "**facteur de Michelson-Morley**" avec:

$$\beta = v/c \quad (49.32)$$

En introduisant également:

$$b = \frac{v}{c^2} \quad (49.33)$$

dans:

$$\lambda^2 = \frac{b \cdot \alpha^2 \cdot c^2}{v} \quad (49.34)$$

nous obtenons:

$$\lambda = \alpha = \frac{1}{\sqrt{1 - (v/c)^2}} \quad (49.35)$$

Posons maintenant (afin d'être conforme aux notations d'usage):

$$\gamma = \frac{1}{\sqrt{1 - \beta^2}} \quad (49.36)$$

avec donc le paramètre sans dimensions et toujours inférieur ou égal à l'unité:

$$\beta = v/c \quad (49.37)$$

2.1. QUADRI Vecteur DÉPLACEMENT

Nous en tirons les relations de "**transformation de Lorentz**" pour passer des valeurs mesurées par O' à celles mesurées par O et inversement:

$$\begin{aligned} x' &= \gamma(x - vt) & x &= \gamma(x' + vt') \\ y' &= y & y &= y' \\ z' &= z & z &= z' \end{aligned} \quad (49.38)$$

$$t' = \gamma \left(t - \frac{v}{c^2} x \right) \quad t = \gamma \left(t' + \frac{v}{c^2} x' \right)$$

qui ont par ailleurs comme propriété d'être covariantes (se traduisent par des relations ayant même structure lors d'un changement de référentiel Galiléen). Nous remarquons à travers ces relations que le concept de "temps" y est donc quelque chose d'individuel relatif au déplacement que nous avons par rapport aux autres (il s'agit du "**temps propre**"). Raison pour laquelle il n'est pas possible de définir un "temps commun" entre deux personnes en mouvement relatif qui ne connaissent pas leur vitesse respective l'une par rapport à l'autre (et encore nous ne prenons pas ici en compte la gravité qui déforme l'espace-temps... nous verrons cela dans le chapitre de Relativité Générale).

Remarque: Si v est beaucoup plus petit que c , nous retrouvons la transformation de Galilée.

Nous pouvons aussi écrire les dernières relations sous la forme (le lecteur remarquera que cette fois les unités de tous les termes à gauche de l'égalité sont identiques - il s'agit à chaque fois d'une distance!):

$$\begin{aligned}x &= \gamma(x' + \beta \cdot c \cdot t') \\y &= y' \\z &= z' \\t \cdot c &= \gamma(t' \cdot c + \beta \cdot x')\end{aligned} \quad (49.39)$$

Bien évidemment la différence est que la quatrième dimension constituant la partie temporelle de "**l'espace-temps**" semble contrairement aux coordonnées spatiales d'usage avoir une direction particulière privilégiée: la "**flèche du temps**" (on ne peut pas revenir à un instant temporel donné dans la réalité alors que cela est possible quand on parcourt une distance purement spatiale). Cette direction du temps est imposée par le deuxième principe de la thermodynamique comme quoi l'entropie ne peut faire qu'augmenter (cf. chapitre de Thermodynamique). Si cela n'était pas le cas alors tous les temps existeraient déjà et on pourrait parcourir le temps comme une distance et le futur serait déjà écrit et on pourrait revenir dans le passé. Cependant la thermodynamique ne donne pas une direction particulière au temps... donc si notre temps a la direction qu'il a.... c'est parce que notre Univers était organisé à sa création (donc qu'il avait une faible entropie).

Nous pouvons alors mettre les transformations de Lorentz des coordonnées et du temps sous la forme matricielle (cf. chapitre d'Algèbre Linéaire) traditionnelle suivante qui définit la "**matrice de Lorentz**" ou de "**matrice de Lorentz-Poincaré**":

$$\begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} = \begin{bmatrix} \gamma & \gamma\beta & 0 & 0 \\ \gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} \quad (49.40)$$

et réciproquement:

$$\begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} = \begin{bmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} \quad (49.41)$$

ce qui donne:

$$\begin{aligned}
 x' &= \gamma(x - vt) & x &= \gamma(x' + vt') \\
 y' &= y & y &= y' \\
 z' &= z & z &= z' \\
 t' &= \gamma\left(t - \frac{v}{c^2}x\right) & t &= \gamma\left(t' + \frac{v}{c^2}x'\right)
 \end{aligned} \quad (49.42)$$

sous forme indicielle cela est plus fréquemment noté:

$$\begin{pmatrix} x'^0 \\ x'^1 \\ x'^2 \\ x'^3 \end{pmatrix} = \begin{bmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{pmatrix} \quad (49.43)$$

ce qui sous forme tensorielle s'écrit:

$$x'^{\mu} = L^{\mu}_{\nu} x^{\nu} \quad \text{ou} \quad x^{\nu} = L^{\nu}_{\mu} x'^{\mu} \quad (49.44)$$

Il s'agit de la forme traditionnelle chez les physiciens de l'expression de changement de référentiel localement inertiel par une transformation de Lorentz.

Remarque: Nous retrouvons le tenseur (la matrice) de transformation de Lorentz dans certains ouvrages sous la forme condensée $L(\beta)$ voire parfois L^{μ}_{ν} ou encore Λ^{μ}_{ν} .

Le vecteur:

$$x^{\nu} = (ct, x, y, z) = (ct, \vec{x}) \quad (49.45)$$

est appelé le "**quadrivecteur d'espace-temps**" ou encore "**quadrivecteur déplacement**".

Remarquons que puisque:

$$x^2 + y^2 + z^2 - c^2 t^2 = x'^2 + y'^2 + z'^2 - c^2 t'^2 \quad (49.46)$$

la transformation par la matrice L^{μ}_{ν} conserve donc la norme. En termes géométriques il s'agit donc d'une "isométrie".

2.1.1. INVARIANCE DE L'ÉQUATION D'ONDE

Maintenant que nous avons déterminé les transformations de Lorentz, nous pouvons contrôler si l'équation d'onde est invariante relativement à ces dernières (rappelons que nous avons démontré plus haut qu'elle n'était pas invariante sous une transformation Galiléenne).

Partant de la transformation de Lorentz écrite en clair:

$$\begin{aligned}x' &= \gamma(x - vt) \\t' &= \gamma\left(t - \frac{v}{c^2}x\right)\end{aligned}\quad (49.47)$$

nous calculons les dérivées partielles par rapport à x et t (l'expression après la deuxième égalité ayant été démontrée plus haut dans ce chapitre):

$$\begin{aligned}\frac{\partial}{\partial x} &= \frac{\partial}{\partial x'} \frac{\partial x'}{\partial x} + \frac{\partial}{\partial t'} \frac{\partial t'}{\partial x} = \gamma \frac{\partial}{\partial x'} - \gamma \frac{v}{c^2} \frac{\partial}{\partial t'} \\ \frac{\partial}{\partial t} &= \frac{\partial}{\partial x'} \frac{\partial x'}{\partial t} + \frac{\partial}{\partial t'} \frac{\partial t'}{\partial t} = -\gamma v \frac{\partial}{\partial x'} + \gamma \frac{\partial}{\partial t'}\end{aligned}\quad (49.48)$$

Ces relations peuvent aussi s'écrire:

$$\begin{aligned}c \frac{\partial}{\partial x} &= \gamma \left(c \frac{\partial}{\partial x'} - \beta \frac{\partial}{\partial t'} \right) \\ \frac{\partial}{\partial t} &= -\gamma \left(v \frac{\partial}{\partial x'} - \frac{\partial}{\partial t'} \right)\end{aligned}\quad (49.49)$$

Au carré:

$$\begin{aligned}c^2 \frac{\partial^2}{\partial x^2} &= \gamma^2 \left(c^2 \frac{\partial^2}{\partial x'^2} - 2\beta c \frac{\partial^2}{\partial x' \partial t'} + \beta^2 \frac{\partial^2}{\partial t'^2} \right) \\ \frac{\partial^2}{\partial t^2} &= \gamma^2 \left(v^2 \frac{\partial^2}{\partial x'^2} - 2v \frac{\partial^2}{\partial x' \partial t'} + \frac{\partial^2}{\partial t'^2} \right)\end{aligned}\quad (49.50)$$

Dans les équations de Maxwell, ou plutôt dans l'équation de propagation du champ électrique ou magnétique dans le vide, nous avons montré (cf. chapitre d'Électrodynamique) que l'opérateur suivant apparaissait:

$$\frac{\partial^2}{\partial t^2} - c^2 \frac{\partial^2}{\partial x^2} \quad (49.51)$$

En y substituant les expressions différentielles précédentes:

$$\begin{aligned}\frac{\partial^2}{\partial t^2} - c^2 \frac{\partial^2}{\partial x^2} &= \gamma^2 \left(v^2 \frac{\partial^2}{\partial x'^2} - 2v \frac{\partial^2}{\partial x' \partial t'} + \frac{\partial^2}{\partial t'^2} - c^2 \frac{\partial^2}{\partial x'^2} + 2\beta c \frac{\partial^2}{\partial x' \partial t'} - \beta^2 \frac{\partial^2}{\partial t'^2} \right) \\ &= \gamma^2 \left[-c^2 \left(1 - v^2/c^2 \right) \frac{\partial^2}{\partial x'^2} - 2v \frac{\partial^2}{\partial x' \partial t'} + 2\beta c \frac{\partial^2}{\partial x' \partial t'} + \left(1 - \beta^2 \right) \frac{\partial^2}{\partial t'^2} \right] \\ &= \frac{\partial^2}{\partial t'^2} - c^2 \frac{\partial^2}{\partial x'^2} + 2\gamma^2 \left[-v \frac{\partial^2}{\partial x' \partial t'} + \frac{v}{c} c \frac{\partial^2}{\partial x' \partial t'} \right] = \frac{\partial^2}{\partial t'^2} - c^2 \frac{\partial^2}{\partial x'^2}\end{aligned}\quad (49.52)$$

Nous avons donc bien:

$$\frac{\partial^2}{\partial t^2} - c^2 \frac{\partial^2}{\partial x^2} = \frac{\partial^2}{\partial t'^2} - c^2 \frac{\partial^2}{\partial x'^2} \quad (49.53)$$

qui montre qu'une transformation de Lorentz laisse invariant cet opérateur (Jackpot!). Nous avons donc obtenu ce que nous cherchions (l'équation d'onde mais dans l'autre référentiel)!

Le lecteur remarquera aussi que cela ne fonctionne que si et seulement si la vitesse de propagation des ondes se fait à la vitesse de la lumière.

2.1.2. INTERPRÉTATION HYPERGÉOMETRIQUE

Revenons maintenant à nos transformations de Lorentz. Rappelons que nous nous sommes restreints au cas particulier où les axes d'espaces étaient parallèles (ce qui nous avait amené à définir le terme "transformations de Lorentz pures"). Cette configuration spéciale a une propriété géométrique intéressante dont parfois de nombreux ouvrages font usage.

Voyons de quoi il s'agit:

Nous avons vu dans le cadre des transformations de Lorentz des longueurs que nous avions une transformation spéciale (boost) selon un axe, à savoir l'axe OX , requérant dans ce cas pour les autres composantes :

$$y' = y, z' = z \quad (49.54)$$

Cela nous permet tout à fait de réduire la matrice de transformation L_{γ}^{μ} (matrice 4×4 que nous avons obtenue plus haut) à une matrice 2×2 de composantes A, B, C et D telle que:

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{pmatrix} ct \\ x \end{pmatrix} \quad (49.55)$$

Nous remarquons que les composantes A, B, C, D respectent par construction les expressions suivantes:

$$\begin{aligned} A^2 - C^2 &= \gamma^2 - (-\gamma\beta)^2 = \gamma^2 - \gamma^2 \beta^2 = \gamma^2 (1 - \beta^2) = \frac{1 - \beta^2}{1 - \beta^2} = 1 \\ D^2 - B^2 &= \gamma^2 - (-\gamma\beta)^2 = \gamma^2 - \gamma^2 \beta^2 = \gamma^2 (1 - \beta^2) = \frac{1 - \beta^2}{1 - \beta^2} = 1 \\ AB - CD &= \gamma(-\gamma\beta) - \gamma(-\gamma\beta) = -\gamma^2 \beta + \gamma^2 \beta = 0 \end{aligned} \quad (49.56)$$

La première relation peut être mise en relation avec la relation remarquable de la trigonométrie hyperbolique (cf. chapitre de Trigonométrie):

$$\cosh^2(\theta) - \sinh^2(\theta) = 1 \quad (49.57)$$

Et donc:

$$A = \cosh(\alpha_1) \text{ et } C = -\sinh(\alpha_1) \quad (49.58)$$

la deuxième qu'il existe α_2 tel que:

$$D = \cosh(\alpha_2) \text{ et } B = -\sinh(\alpha_2) \quad (49.59)$$

Remarque: Le choix du signe "-" pour C et B est utile car comme nous avons toujours $\beta \geq 0$ (de même pour γ qui est strictement positif) cela nous imposera à la fin des calculs d'avoir $\alpha \geq 0$. Dès lors, comme $-\gamma\beta \leq 0$ et $\alpha \geq 0$ la seule manière pour que C (ainsi que B) puisse être négatif c'est de mettre un "-".

La troisième donne alors la relation d'addition remarquable:

$$AB - CD = -\cosh(\alpha_1) \sinh(\alpha_2) + \sinh(\alpha_1) \cosh(\alpha_2) = \sinh(\alpha_1 - \alpha_2) = 0 \quad (49.60)$$

et donc la différence $\alpha_1 - \alpha_2$ que nous noterons plus simplement α est nulle. Ce qui valide les relations:

$$A = D \Leftrightarrow \cosh(\alpha_1) = \cosh(\alpha_2), B = C \Leftrightarrow -\sinh(\alpha_2) = -\sinh(\alpha_1) \quad (49.61)$$

La matrice se présente donc comme suit:

$$L_v^\mu = \begin{bmatrix} \cosh(\alpha_1) & -\sinh(\alpha_1) \\ -\sinh(\alpha_1) & \cosh(\alpha_1) \end{bmatrix} = \begin{bmatrix} \cosh(\alpha_2) & -\sinh(\alpha_2) \\ -\sinh(\alpha_2) & \cosh(\alpha_2) \end{bmatrix} \quad (49.62)$$

Finalement, les transformations de Lorentz spéciales de vitesse v suivant l'axe OX peuvent aussi s'écrire:

$$\begin{aligned} ct' &= \gamma(ct - \beta x) = \cosh(\alpha)ct - \sinh(\alpha)x = \cosh(\alpha) \left(ct - \frac{\sinh(\alpha)}{\cosh(\alpha)} x \right) \\ &= \cosh(\alpha) (ct - \tanh(\alpha)x) \\ x' &= \gamma(x - \beta ct) = \cosh(\alpha)x - \sinh(\alpha)ct = \cosh(\alpha) \left(x - \frac{\sinh(\alpha)}{\cosh(\alpha)} ct \right) \\ &= \cosh(\alpha) (x - \tanh(\alpha)ct) \end{aligned} \quad (49.63)$$

ce qui nous amène à écrire:

$$\beta = \frac{v}{c} = \tanh(\alpha) = \frac{\sinh(\alpha)}{\cosh(\alpha)} \text{ et } \gamma = \frac{1}{\sqrt{1 - \beta^2}} = \cosh(\alpha) \quad (49.64)$$

La quantité α (sans dimensions) est appelée "**rapidité**" par ceux qui l'utilisent en physique des hautes énergies. L'intérêt de travailler avec des angles est de rendre la combinaison de 2 boosts plus aisée.

Nous nous arrêtons ici en ce qui concerne l'étude géométrique de la relativité restreinte trouvant que cela a de moins en moins d'intérêt de procéder ainsi (bien que ce soit fort sympathique).

2.2. QUADRI-VECTEUR VITESSE

Nous pouvons de même déterminer les transformations de Lorentz des vitesses. Considérons une particule en mouvement dans le référentiel inertiel O' telle qu'au temps t' , ses coordonnées soient x', y', z' .

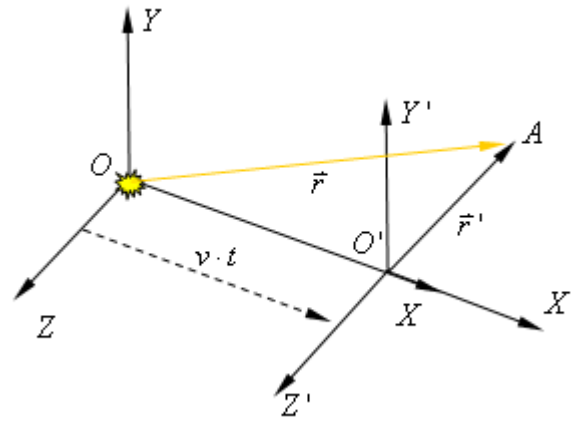


Figure: 49.2 - Configuration pour l'étude du quadrivecteur vitesse

Dès lors, les composantes de la vitesse v' sont:

$$v'_{x'} = \frac{dx'}{dt'}, v'_{y'} = \frac{dy'}{dt'}, v'_{z'} = \frac{dz'}{dt'} \quad (49.65)$$

Quelles sont alors les composantes de sa vitesse dans O (rappelons que O' s'éloigne à vitesse v) ?

À nouveau, nous écrivons:

$$v_x = \frac{dx}{dt}, v_y = \frac{dy}{dt}, v_z = \frac{dz}{dt} \quad (49.66)$$

Nous pouvons différencier les équations de transformation des composantes que nous avons obtenues avant et ainsi pouvons écrire:

$$\begin{aligned} dx &= \gamma(dx' + (\beta/c) dt') \\ dy &= dy' \\ dz &= dz' \\ dt &= \gamma(dt' + (\beta/c) dx') \end{aligned} \quad (49.67)$$

Dès lors, nous avons:

$$v_z = \frac{dz}{dt} = \frac{dz'}{\gamma(dt' + (\beta/c) dx')} = \frac{dz'}{\gamma(1 + (\beta/c) v'_{x'}) dt'} = \frac{v'_{z'}}{\gamma(1 + (\beta/c) v'_{x'})} \quad (49.68)$$

et de même:

$$v_y = \frac{v'_{y'}}{\gamma(1 + (\beta/c) v'_{x'})} \quad (49.69)$$

et:

$$v_x = \frac{dx}{dt} = \frac{\gamma(dx' + (\beta/c) dt')}{\gamma(dt' + (\beta/c) dx')} = \frac{(v'_{x'} + (\beta/c)) dt'}{(1 + (\beta/c) v'_{x'}) dt'} = \frac{v'_{x'} + (\beta/c)}{1 + (\beta/c) v'_{x'}} \quad (49.70)$$

Et comme la vitesse constante du référentiel O' est donnée par $\beta = v/c$, nous avons alors:

$$v_x = \frac{v_{x'} + v}{1 + (\beta/c) v_{x'}}, v_y = \frac{v_{y'}}{\gamma(1 + (v/c^2) v_{x'})}, v_z = \frac{v_{z'}}{\gamma(1 + (v/c^2) v_{x'})} \quad (49.71)$$

et inversement:

$$v_{x'} = \frac{v_x - v}{1 - (\beta/c) v_x}, v_{y'} = \frac{v_y}{\gamma(1 - (v/c^2) v_x)}, v_{z'} = \frac{v_z}{\gamma(1 - (v/c^2) v_x)} \quad (49.72)$$

Dans la limite de la mécanique classique, où la vitesse de la lumière était supposée instantanée et donc $c \rightarrow \infty$, nous retrouvons:

$$v_x = v_{x'} + v, v_y = v_{y'}, v_z = v_{z'} \quad (49.73)$$

qui sont les transformations de Galilée telles que nous les avons vues dans le chapitre de Mécanique Classique.

Comme nous pouvons le voir, les transformations des vitesses ne suivent pas trop la forme de la matrice de Lorentz que nous avons déterminée plus haut pour les coordonnées. Les physiciens, n'aimant pas ce qui est inhomogène, ont cherché à avoir les mêmes transformations pour les deux.

Ainsi, reprenons les relations de transformation des vitesses et réécrivons les telles que ci-dessous:

$$\frac{v_x}{c} = \frac{\frac{v_{x'}}{c} + \beta}{1 + \beta \frac{v_{x'}}{c}}, \frac{v_y}{c} = \frac{\frac{v_{y'}}{c}}{\gamma \left(1 + \beta \frac{v_{x'}}{c}\right)}, \frac{v_z}{c} = \frac{\frac{v_{z'}}{c}}{\gamma \left(1 + \beta \frac{v_{x'}}{c}\right)} \quad (49.74)$$

Ces relations peuvent s'écrire différemment si nous calculons:

$$1 - \frac{\|\vec{v}\|^2}{c^2} = 1 - \frac{v_x^2 + v_y^2 + v_z^2}{c^2} = 1 - \left(\frac{\frac{v_{x'}}{c} + \beta}{1 + \beta \frac{v_{x'}}{c}} \right)^2 - \left(1 - \frac{v^2}{c^2} \right) \left(\left(\frac{\frac{v_{y'}}{c}}{1 + \beta \frac{v_{x'}}{c}} \right)^2 + \left(\frac{\frac{v_{z'}}{c}}{1 + \beta \frac{v_{x'}}{c}} \right)^2 \right) \quad (49.75)$$

Soit en simplifiant un peu:

$$\begin{aligned}
\left(1 - \frac{\|\vec{v}\|^2}{c^2}\right) \left(1 + \beta \frac{v_{x'}}{c}\right)^2 &= \left(1 + \beta \frac{v_{x'}}{c}\right)^2 - \left(\frac{v_{x'}}{c} + \beta\right)^2 - \left(1 - \frac{v^2}{c^2}\right) \left(\left(\frac{v_{y'}}{c}\right)^2 + \left(\frac{v_{z'}}{c}\right)^2\right) \\
&= 1 + 2\beta \frac{v_{x'}}{c} + \beta^2 \left(\frac{v_{x'}}{c}\right)^2 - \left(\frac{v_{x'}}{c}\right)^2 - 2\beta \frac{v_{x'}}{c} - \beta^2 - \left(1 - \frac{v^2}{c^2}\right) \left(\left(\frac{v_{y'}}{c}\right)^2 + \left(\frac{v_{z'}}{c}\right)^2\right) \\
&= 1 + \beta^2 \left(\frac{v_{x'}}{c}\right)^2 - \left(\frac{v_{x'}}{c}\right)^2 - \beta^2 - \left(1 - \frac{v^2}{c^2}\right) \left(\left(\frac{v_{y'}}{c}\right)^2 + \left(\frac{v_{z'}}{c}\right)^2\right) \\
&= (1 - \beta^2) \left(1 - \left(\frac{v_{x'}}{c}\right)^2\right) - \left(1 - \frac{v^2}{c^2}\right) \left(\left(\frac{v_{y'}}{c}\right)^2 + \left(\frac{v_{z'}}{c}\right)^2\right) \\
&= \left(1 - \frac{v^2}{c^2}\right) \left(1 - \left(\frac{v_{x'}}{c}\right)^2 - \left(\frac{v_{y'}}{c}\right)^2 - \left(\frac{v_{z'}}{c}\right)^2\right) = \left(1 - \frac{v^2}{c^2}\right) \left(1 - \frac{\|\vec{v}'\|^2}{c^2}\right)
\end{aligned} \tag{49.76}$$

Posons:

$$\Gamma(\vec{v}) = \frac{1}{\sqrt{1 - \frac{\|\vec{v}\|^2}{c^2}}} = \frac{1}{\sqrt{1 - \left(\frac{x}{ct}\right)^2 - \left(\frac{y}{ct}\right)^2 - \left(\frac{z}{ct}\right)^2}} = \Gamma_{\vec{v}} \tag{49.77}$$

et:

$$\Gamma(\vec{v}') = \frac{1}{\sqrt{1 - \frac{\|\vec{v}'\|^2}{c^2}}} = \frac{1}{\sqrt{1 - \left(\frac{x'}{ct'}\right)^2 - \left(\frac{y'}{ct'}\right)^2 - \left(\frac{z'}{ct'}\right)^2}} = \Gamma_{\vec{v}'} \tag{49.78}$$

et:

$$\gamma = \gamma(v) = \Gamma(\vec{v}) = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{49.79}$$

Avec cette notation il nous sera aisé de déterminer la composante temporelle, en effet la relation:

$$\left(1 - \frac{\|\vec{v}\|^2}{c^2}\right) \left(1 + \beta \frac{v_{x'}}{c}\right)^2 = \left(1 - \frac{v^2}{c^2}\right) \left(1 - \frac{\|\vec{v}'\|^2}{c^2}\right) \tag{49.80}$$

s'écrit:

$$\begin{aligned}
\frac{1}{\sqrt{1 - \frac{\|\vec{v}\|^2}{c^2}}} \frac{1}{1 + \beta \frac{v'_x}{c}} &= \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \frac{1}{\sqrt{1 - \frac{\|\vec{v}'\|^2}{c^2}}} \\
\Gamma_{\vec{v}} \frac{1}{1 + \beta \frac{v'_x}{c}} &= \Gamma_{\vec{v}'} \gamma(v) \\
\Gamma_{\vec{v}} &= \left(1 + \beta \frac{v'_x}{c} \right) \Gamma_{\vec{v}'} \gamma(v) \\
\Gamma_{\vec{v}} &= \gamma(v) \left(\Gamma_{\vec{v}'} + \beta \frac{v'_x}{c} \Gamma_{\vec{v}'} \right) \\
\Gamma_{\vec{v}} c &= \gamma \left(\Gamma_{\vec{v}'} c + \beta v'_x \Gamma_{\vec{v}'} \right)
\end{aligned} \tag{49.81}$$

En procédant de même pour chacune des composantes spatiales, nous aurons au final:

$$\begin{aligned}
\Gamma_{\vec{v}} c &= \gamma \left(\Gamma_{\vec{v}'} c + \beta v'_x \Gamma_{\vec{v}'} \right) \\
\Gamma_{\vec{v}} v_x &= \gamma \left(\Gamma_{\vec{v}'} v'_x + \beta \Gamma_{\vec{v}'} c \right) \\
\Gamma_{\vec{v}} v_y &= \Gamma_{\vec{v}'} v'_y \\
\Gamma_{\vec{v}} v_z &= \Gamma_{\vec{v}'} v'_z
\end{aligned} \tag{49.82}$$

et nous avons atteint ici notre objectif d'homogénéisation qui nous permet d'écrire si nous posons $\Gamma_{\vec{v}} = \gamma(\vec{v})$, $\Gamma_{\vec{v}'} = \gamma(\vec{v}')$:

$$\begin{pmatrix} \gamma(\vec{v})c \\ \gamma(\vec{v})v_x \\ \gamma(\vec{v})v_y \\ \gamma(\vec{v})v_z \end{pmatrix} = \gamma(\vec{v}) \begin{pmatrix} c \\ v_x \\ v_y \\ v_z \end{pmatrix} = \begin{bmatrix} \gamma & \gamma\beta & 0 & 0 \\ \gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} \gamma(\vec{v}')c \\ \gamma(\vec{v}')v_x \\ \gamma(\vec{v}')v_y \\ \gamma(\vec{v}')v_z \end{pmatrix} \tag{49.83}$$

ce qui sous forme tensorielle s'écrit:

$$\boxed{u'^{\mu} = L^{\mu}_{\nu} u^{\nu}} \text{ ou } \boxed{u^{\nu} = L^{\nu}_{\mu} u'^{\mu}} \tag{49.84}$$

Le vecteur:

$$\boxed{u^{\nu} = \gamma(\vec{v})(c, v_x, v_y, v_z) = \gamma(\vec{v})(c, \vec{v})} \tag{49.85}$$

est quant à lui appelé le "**quadrivecteur vitesse**".

2.3. QUADRIVECTEUR COURANT

Nous avons défini naturellement lors de notre introduction du tenseur du champ électromagnétique (cf. chapitre d'Électrodynamique) le quadrivecteur courant:

$$j^{\nu} = (\rho c, \vec{j}) \tag{49.86}$$

que nous pouvons écrire:

$$j^{\mathbf{v}} = \rho(c, \vec{v}) \quad (49.87)$$

Dès lors, en considérant ρ_0 comme la densité de charge dans le référentiel propre se déplaçant à la vitesse v par rapport au référentiel O' et du fait de la contraction des longueurs dans la direction de la vitesse, le volume occupé par une charge donnée sera multiplié par le facteur $\gamma(\vec{v})$ de sorte que:

$$j^{\mathbf{v}} = \rho_0 \gamma(\vec{v}) (c, \vec{v}) \quad (49.88)$$

qui n'est autre que le "**quadrivecteur courant**" où nous retrouvons le quadrivecteur vitesse déterminé précédemment.

2.4. QUADRIVECTEUR ACCÉLÉRATION

Ayant obtenu précédemment un quadrivecteur vitesse transformable à l'aide de la matrice de Lorentz cherchons aussi l'équivalent pour l'accélération.

Le quadrivecteur accélération s'exprime naturellement comme la dérivée par rapport au temps propre de la quadri-vitesse u tel que:

$$a^{\mathbf{v}} = \frac{du^{\mathbf{v}}}{dt'} = \frac{dt}{dt'} \frac{du^{\mathbf{v}}}{dt} \quad (49.89)$$

Rappelons que le temps propre d'une particule est le temps mesuré dans le repère de cette particule, c'est-à-dire dans le repère où elle est immobile. Le temps propre dans la littérature spécialisée est souvent noté τ .

Remarque: Attention!! Si le lecteur a compris les développements jusqu'à maintenant, l'accélération que nous cherchons à calculer est celle d'un objet accéléré dans un des référentiels en mouvement relatif uniforme par rapport à un autre (ce ne sont donc pas les référentiels qui sont en mouvement accéléré ici!!).

Il faudra d'abord que le lecteur admette (nous le démontrons cependant un peu plus loin) que:

$$\frac{dt}{dt'} = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = \gamma \quad (49.90)$$

Dès lors, nous avons:

$$a^{\mathbf{v}} = \frac{dt}{dt'} \frac{du^{\mathbf{v}}}{dt} = \gamma \frac{d}{dt} (c\gamma, \gamma\vec{v}) \quad (49.91)$$

Si nous introduisons l'accélération ordinaire $\vec{a} = d\vec{v}/dt$ nous voyons que:

$$\frac{d\gamma}{dt} = \frac{1}{2c^2} \gamma^3 \frac{d}{dt} (\|\vec{v}\|^2) = \frac{1}{2c^2} \gamma^3 \frac{d}{dt} (\vec{v} \circ \vec{v}) = \frac{\gamma^3}{2c^2} (\vec{a} \circ \vec{v} + \vec{v} \circ \vec{a}) = \frac{\gamma^3}{c^2} \vec{v} \circ \vec{a} \quad (49.92)$$

alors:

$$a^v = \gamma \frac{d}{dt}(c\gamma, \gamma\vec{v}) = \gamma \left(\frac{\gamma^3}{c} \vec{v} \circ \vec{a}, \gamma\vec{a} + \frac{\gamma^3}{c^2} (\vec{v} \circ \vec{a}) \vec{v} \right) \quad (49.93)$$

En utilisant la relation (cf. chapitre de Calcul Vectoriel):

$$(\vec{v} \circ \vec{a}) \vec{v} = \vec{v} \times (\vec{v} \times \vec{a}) + (\vec{v} \circ \vec{v}) \vec{a} \quad (49.94)$$

nous trouvons que le quadrivecteur accélération peut être écrit:

$$\begin{aligned} a^v &= \gamma^4 \left(\vec{\beta} \circ \vec{a}, \frac{\vec{a}}{\gamma^2} + \frac{1}{c^2} (\vec{v} \circ \vec{a}) \vec{v} \right) = \gamma^4 \left(\vec{\beta} \circ \vec{a}, \frac{\vec{a}}{\gamma^2} + \frac{1}{c^2} (\vec{v} \times (\vec{v} \times \vec{a}) + (\vec{v} \circ \vec{v}) \vec{a}) \right) \\ &= \gamma^4 \left(\vec{\beta} \circ \vec{a}, \frac{\vec{a}}{\gamma^2} + (\vec{\beta} \times (\vec{\beta} \times \vec{a}) + (\vec{\beta} \circ \vec{\beta}) \vec{a}) \right) = \gamma^4 \left(\vec{\beta} \circ \vec{a}, \left(\frac{1}{\gamma^2} + \vec{\beta} \circ \vec{\beta} \right) \vec{a} + \vec{\beta} \times (\vec{\beta} \times \vec{a}) \right) \quad (49.95) \\ &= \gamma^4 \left(\vec{\beta} \circ \vec{a}, ((1 - \vec{\beta} \circ \vec{\beta}) + \vec{\beta} \circ \vec{\beta}) \vec{a} + \vec{\beta} \times (\vec{\beta} \times \vec{a}) \right) = \gamma^4 \left(\vec{\beta} \circ \vec{a}, \vec{a} + \vec{\beta} \times (\vec{\beta} \times \vec{a}) \right) \end{aligned}$$

Le vecteur:

$$a^v = \gamma^4 \left(\vec{\beta} \circ \vec{a}, \vec{a} + \vec{\beta} \times (\vec{\beta} \times \vec{a}) \right) \quad (49.96)$$

est appelé "**quadrivecteur accélération**" et se transforme donc aussi à l'aide de la matrice de Lorentz.

Nous voyons que si $v \ll c$ et $\vec{a} = \vec{a}_0$ cette dernière relation se simplifie en:

$$a^v = \gamma^4 \left(\vec{\beta} \circ \vec{a}, \vec{a} + \vec{\beta} \times (\vec{\beta} \times \vec{a}) \right) = 1 \left(\vec{0} \circ \vec{a}, \vec{a} + \vec{0} \times (\vec{0} \times \vec{a}) \right) = (0, \vec{a}) \quad (49.97)$$

Nous retrouvons donc l'accélération classique.

En utilisant la métrique de Minkowski (voir sa définition plus loin), notée η_{uv} , calculons la norme du quadrivecteur accélération:

$$\begin{aligned} a^v a_v &= a^v (\eta_{uv} a^u) = a^0 a_0 - a_i a^i = \gamma^8 \left[(\vec{\beta} \circ \vec{a})^2 - (\vec{a} + \vec{\beta} \times (\vec{\beta} \times \vec{a}))^2 \right] \\ &= \gamma^8 \left[(\vec{\beta} \circ \vec{a})^2 - \left(\vec{a} + \underbrace{(\vec{\beta} \circ \vec{a}) \vec{\beta} - (\vec{\beta} \circ \vec{\beta}) \vec{a}}_{=\vec{c}} \right)^2 \right] = \gamma^8 \left[(\vec{\beta} \circ \vec{a})^2 - (\vec{a} + \vec{c})^2 \right] \quad (49.98) \\ &= \gamma^8 \left[(\vec{\beta} \circ \vec{a})^2 - (\vec{a}^2 + 2\vec{a}\vec{c} + \vec{c}^2) \right] \end{aligned}$$

Remarque: Il faut bien comprendre que quand nous écrivons $(\vec{a} + \vec{\beta} \times (\vec{\beta} \times \vec{a}))^2$ il s'agit dans ce cas implicitement de la somme des carrés des composantes du calcul entre parenthèses.

Et comme:

$$\begin{aligned}\vec{c}^2 &= (\vec{\beta} \circ \vec{a})^2 \vec{\beta}^2 - 2 (\vec{\beta} \circ \vec{\beta})(\vec{\beta} \circ \vec{a}) \vec{\beta} \vec{a} + (\vec{\beta} \circ \vec{\beta})^2 \vec{a}^2 \\ \vec{a} \vec{c} &= (\vec{\beta} \circ \vec{a}) \vec{\beta} \vec{a} - (\vec{\beta} \circ \vec{\beta}) \vec{a}^2\end{aligned}\quad (49.99)$$

nous rassemblons cela:

$$a^v a_v = \gamma^8 \left[(\vec{\beta} \circ \vec{a})^2 - \left(\vec{a}^2 + 2 (\vec{\beta} \circ \vec{a}) \vec{\beta} \vec{a} - 2 (\vec{\beta} \circ \vec{\beta}) \vec{a}^2 + (\vec{\beta} \circ \vec{a})^2 \vec{\beta}^2 - 2 (\vec{\beta} \circ \vec{\beta})(\vec{\beta} \circ \vec{a}) \vec{\beta} \vec{a} + (\vec{\beta} \circ \vec{\beta})^2 \vec{a}^2 \right) \right] \quad (49.100)$$

Maintenant, nous développons la somme $a_i a^i$ de la grosse parenthèse qui devient dès lors:

$$\begin{aligned}a^v a_v &= \gamma^8 \left[(\beta_i a^i)^2 - \left(a_i a^i + 2 (\beta_i a^i)(\beta_i a^i) - 2 (\beta_i \beta^i) a_i a^i + (\beta_i a^i)^2 (\beta_i \beta^i) - 2 (\beta_i \beta^i)(\beta_i a^i)(\beta_i a^i) + (\beta_i \beta^i)^2 a_i a^i \right) \right] \\ &= \gamma^8 \left[(\beta_i a^i)^2 - \left(a_i a^i + 2 (\beta_i a^i)^2 - 2 (\beta_i \beta^i) a_i a^i + (\beta_i a^i)^2 (\beta_i \beta^i) - 2 (\beta_i \beta^i)(\beta_i a^i)^2 + (\beta_i \beta^i)^2 a_i a^i \right) \right]\end{aligned}\quad (49.101)$$

Nous simplifions:

$$\begin{aligned}a^v a_v &= \gamma^8 \left[(\beta_i a^i)^2 (1 - 2 - \beta_i \beta^i + 2 \beta_i \beta^i) + a_i a^i (-1 + 2 \beta_i \beta^i - (\beta_i \beta^i)^2) \right] \\ &= \gamma^8 \left[(\vec{\beta} \circ \vec{a})^2 (1 - 2 - \vec{\beta}^2 + 2 \vec{\beta}^2) + \vec{a}^2 (-1 + 2 \vec{\beta}^2 - \vec{\beta}^4) \right]\end{aligned}\quad (49.102)$$

d'où:

$$\begin{aligned}a^v a_v &= \gamma^8 \left[(\vec{\beta} \circ \vec{a})^2 (1 - 2 - \vec{\beta}^2 + 2 \vec{\beta}^2) + \vec{a}^2 (-1 + 2 \vec{\beta}^2 - \vec{\beta}^4) \right] \\ &= \gamma^8 \left[(\vec{\beta} \circ \vec{a})^2 (-1 + \vec{\beta}^2) + \vec{a}^2 (-1 + 2 \vec{\beta}^2 - \vec{\beta}^4) \right] \\ &= -\gamma^8 (1 - \vec{\beta}^2) \left[(\vec{\beta} \circ \vec{a})^2 + \vec{a}^2 (1 - \vec{\beta}^2) \right] = -\gamma^8 \gamma^{-2} \left[(\vec{\beta} \circ \vec{a})^2 + \vec{a}^2 (1 - \vec{\beta}^2) \right] \\ &= -\gamma^6 \left[(\vec{a} \circ \vec{\beta})^2 + \vec{a}^2 (1 - \vec{\beta}^2) \right] = -\gamma^6 \left\{ \vec{a}^2 - \left[\vec{a}^2 \vec{\beta}^2 - (\vec{a} \circ \vec{\beta})^2 \right] \right\}\end{aligned}\quad (49.103)$$

Or, nous avons la relation:

$$\|\vec{a} \times \vec{\beta}\|^2 = \vec{a}^2 \vec{\beta}^2 - (\vec{a} \circ \vec{\beta})^2 = (a_i a^i)(\beta_i \beta^i) - (a_i \beta^i)^2 \quad (49.104)$$

et la propriété du produit vectoriel:

$$\vec{a} \times \vec{\beta} = -\vec{\beta} \times \vec{a} \quad (49.105)$$

Ce qui nous donne finalement:

$$a^v a_v = -\gamma^6 \left\{ \vec{a}^2 - (\vec{a} \times \vec{\beta})^2 \right\} = -\gamma^6 \left\{ \vec{a} \circ \vec{a} - (\vec{a} \times \vec{\beta}) \circ (\vec{a} \times \vec{\beta}) \right\} = -\gamma^6 \left\{ \vec{a}^2 - (\vec{\beta} \times \vec{a})^2 \right\} \quad (49.106)$$

Imaginons maintenant un objet avec un mouvement relatif uniformément accéléré $\vec{a}_0^2$ (accélération constante) dans notre propre référentiel. Si nous supposons notre référentiel fixe, nous avons $\vec{v} = \vec{0} \rightarrow \vec{\beta} = 0$. Dès lors:

$$a^v a_v = -\gamma^6 \vec{a}^2 = -\vec{a}_0^2 \quad (49.107)$$

In extenso, si le mouvement accéléré ne se fait que le long d'une seule composante:

$$\gamma^3 \frac{du}{dt} = a_0 \quad (49.108)$$

Or, nous avons aussi:

$$\frac{d}{dt}(\gamma u) = \gamma \frac{du}{dt} + u \frac{d\gamma}{dt} = \gamma^3 \frac{du}{dt} \left(\gamma^{-2} + \frac{u^2}{c^2} \right) = \gamma^3 \frac{du}{dt} \left(1 - \frac{u^2}{c^2} + \frac{u^2}{c^2} \right) = \gamma^3 \frac{du}{dt} \quad (49.109)$$

Donc finalement, nous pouvons écrire:

$$\gamma^3 \frac{du}{dt} = \frac{d}{dt}(\gamma u) = \frac{d}{dt} \frac{u}{\sqrt{1 - \frac{u^2}{c^2}}} = a_0 \quad (49.110)$$

Ce qui après intégration donne:

$$\boxed{u = \frac{dx}{dt} = \frac{a_0 t}{\sqrt{1 + \frac{a_0^2 t^2}{c^2}}}} \quad (49.111)$$

Nous voyons que la vitesse u n'atteint jamais c alors que la force est toujours la même!

Nous avons donc:

$$x = \frac{c^2}{a_0} \int_0^{a_0 t/c} \frac{x}{\sqrt{1+x^2}} dx = \frac{c^2}{a_0} \sqrt{1 + \frac{a_0^2 t^2}{c^2}} - \frac{c^2}{a_0} \quad (49.112)$$

ce qui nous donne:

$$\left(x + \frac{c^2}{a_0} \right)^2 - c^2 t^2 = \frac{c^4}{a_0^2} \quad (49.113)$$

Après réarrangement, nous écrivons cela:

$$x = \frac{c^2}{a_0} \left(\sqrt{1 + \frac{a_0^2 t^2}{c^2}} - 1 \right) \quad (49.114)$$

Nous sommes bien loin de la relation du mouvement uniformément accéléré que nous avons en mécanique classique. Cependant, pour t proche de zéro, nous retrouvons la relation de la mécanique classique en prenant le développement de Taylor au deuxième ordre de la racine (cf. chapitre Suites Et Séries):

$$x = \frac{c^2}{a_0} \left(\sqrt{1 + \frac{a_0^2 t^2}{c^2}} - 1 \right) \cong \frac{c^2}{a_0} \left(1 + \frac{a_0^2 t^2}{2c^2} - 1 \right) = \frac{1}{2} a_0 t^2 \quad (49.115)$$

Cependant, ceci ne nous donne pas les relations de transformations de composantes de l'accélération sous une forme simple. Voyons donc comment les obtenir.

Rappelons d'abord que nous avons obtenu pour la vitesse:

$$v_x = \frac{v'_{x'} + v}{1 + \frac{v}{c^2} v'_{x'}}, v_y = \frac{v'_{y'}}{\gamma \left(1 + \frac{v}{c^2} v'_{x'} \right)}, v_z = \frac{v'_{z'}}{\gamma \left(1 + \frac{v}{c^2} v'_{x'} \right)} \quad (49.116)$$

Il vient en les différenciant:

$$\begin{aligned} dv_x &= d \left((v'_{x'} + v) \frac{1}{1 + \frac{v}{c^2} v'_{x'}} \right) = \frac{dv'_{x'}}{1 + \frac{v}{c^2} v'_{x'}} + \frac{\frac{dv}{dt}}{1 + \frac{v}{c^2} v'_{x'}} + (v'_{x'} + v) d \left(1 + \frac{v}{c^2} v'_{x'} \right)^{-1} \\ &= \frac{dv'_{x'}}{1 + \frac{v}{c^2} v'_{x'}} - \frac{(v'_{x'} + v) \frac{v}{c^2} dv'_{x'}}{\left(1 + \frac{v}{c^2} v'_{x'} \right)^2} = \frac{dv'_{x'} \left(1 + \frac{v}{c^2} v'_{x'} \right) - (v'_{x'} + v) \frac{v}{c^2} dv'_{x'}}{\left(1 + \frac{v}{c^2} v'_{x'} \right)^2} \\ &= \frac{dv'_{x'} + \frac{v}{c^2} v'_{x'} dv'_{x'} - v'_{x'} \frac{v}{c^2} dv'_{x'} - v \frac{v}{c^2} dv'_{x'}}{\left(1 + \frac{v}{c^2} v'_{x'} \right)^2} = \frac{dv'_{x'} - v \frac{v}{c^2} dv'_{x'}}{\left(1 + \frac{v}{c^2} v'_{x'} \right)^2} \\ &= \frac{dv'_{x'} \left(1 - \frac{v^2}{c^2} \right)}{\left(1 + \frac{v}{c^2} v'_{x'} \right)^2} = \frac{dv'_{x'}}{\gamma^2 \left(1 + \frac{v}{c^2} v'_{x'} \right)^2} \end{aligned} \quad (49.117)$$

et donc:

$$\begin{aligned} \frac{dv_x}{dt} &= \frac{1}{\gamma^2 \left(1 + \frac{v}{c^2} v'_{x'} \right)^2} \frac{dv'_{x'}}{dt} = \frac{1}{\gamma^2 \left(1 + \frac{v}{c^2} v'_{x'} \right)^2} \frac{dt'}{dt} \frac{dv'_{x'}}{dt'} \\ &= \frac{1}{\gamma^2 \left(1 + \frac{v}{c^2} v'_{x'} \right)^2} \frac{dt'}{dt} a'_{x'} \end{aligned} \quad (49.118)$$

Rappelons maintenant que nous avons démontré que:

$$c \cdot t = \gamma(t' \cdot c + \beta \cdot x') \quad (49.119)$$

en différenciant il vient:

$$\begin{aligned} c \cdot dt &= \gamma(dt' \cdot c + \beta \cdot dx') \Rightarrow c \cdot \frac{dt}{dt'} = \frac{\gamma}{dt'}(dt' \cdot c + \beta \cdot dx') \Rightarrow \frac{dt}{dt'} = \gamma \left(1 + \frac{\beta \cdot v'_x}{c} \right) \\ \Rightarrow \frac{dt'}{dt} &= \frac{1}{\gamma \left(1 + \frac{\beta \cdot v'_x}{c} \right)} = \frac{1}{\gamma \left(1 + \frac{v \cdot v'_x}{c^2} \right)} \end{aligned} \quad (49.120)$$

d'où finalement:

$$a_x = \frac{dv_x}{dt} = \frac{1}{\gamma^2 \left(1 + \frac{v}{c^2} v'_x \right)^2} \frac{dt'}{dt} a'_{x'} = \frac{1}{\gamma^3 \left(1 + \frac{v}{c^2} v'_x \right)^3} a'_{x'} \quad (49.121)$$

et pour les composantes y, z:

$$v_{y/z} = \frac{v'_{y/z'}}{\gamma \left(1 + \frac{v}{c^2} v'_x \right)} \Rightarrow dv_{y/z} = \frac{dv'_{y/z'}}{\gamma \left(1 + \frac{v}{c^2} v'_x \right)} - \frac{v'_{y/z'}}{\gamma \left(1 + \frac{v}{c^2} v'_x \right)^2} \frac{v}{c^2} dv'_x \quad (49.122)$$

et donc:

$$\begin{aligned} \frac{dv_{y/z}}{dt} &= \left(\frac{1}{\gamma \left(1 + \frac{v}{c^2} v'_x \right)} \frac{dv'_{y/z'}}{dt} - \frac{v'_{y/z'}}{\gamma \left(1 + \frac{v}{c^2} v'_x \right)^2} \frac{v}{c^2} \frac{dv'_x}{dt} \right) \frac{dt'}{dt} \\ &= \frac{a'_{y/z'}}{\gamma^2 \left(1 + \frac{v}{c^2} v'_x \right)^2} - \frac{v \cdot v'_{y/z'}}{c^2} \frac{a'_{x'}}{\gamma^2 \left(1 + \frac{v}{c^2} v'_x \right)^3} \end{aligned} \quad (49.123)$$

Donc finalement:

$$\boxed{\begin{aligned} a_x &= \frac{1}{\gamma^3 \left(1 + \frac{v}{c^2} v'_x \right)^3} a'_{x'} \\ a_y &= \frac{a'_{y'}}{\gamma^2 \left(1 + \frac{v}{c^2} v'_x \right)^2} - \frac{v \cdot v'_y}{c^2} \frac{a'_{x'}}{\gamma^2 \left(1 + \frac{v}{c^2} v'_x \right)^3} \\ a_z &= \frac{a'_{z'}}{\gamma^2 \left(1 + \frac{v}{c^2} v'_x \right)^2} - \frac{v \cdot v'_z}{c^2} \frac{a'_{x'}}{\gamma^2 \left(1 + \frac{v}{c^2} v'_x \right)^3} \end{aligned}} \quad (49.124)$$

Rappelons que ces relations s'appliquent lorsque les mouvements des référentiels sont de translation uniforme!

2.5. ADDITION RELATIVISTE DES VITESSES

Comme la vitesse de la lumière est une vitesse supposée indépassable, nous en venons maintenant à nous demander quelle sera alors finalement la vitesse d'un objet lancé à une vitesse proche de celle de la lumière (par exemple...) à partir d'un référentiel se déplaçant lui aussi à une vitesse proche de la lumière (pourquoi pas non plus...).

Il nous faut alors trouver une relation qui donne la vitesse réelle V à partir de la vitesse de lancement v_2 et de la vitesse du référentiel v_1 .

Nous savons que pour l'objet lancé:

$$x' = v_2 \cdot t' \quad (49.125)$$

Comme celui qui est intéressé ne connaît pas la vitesse réelle V , il se doit d'utiliser les transformations de Lorentz. Ainsi, compte tenu de l'expression de t' que nous avons vue plus haut il vient:

$$x' = v_2 \cdot \frac{t - \frac{v_1 \cdot x}{c^2}}{\sqrt{1 - (v_1/c)^2}} \quad (49.126)$$

et compte tenu de l'expression de x' , avons également:

$$\frac{x - v_1 \cdot t}{\sqrt{1 - (v_1/c)^2}} = v_2 \cdot \frac{t - \frac{v_1 \cdot x}{c^2}}{\sqrt{1 - (v_1/c)^2}} \quad (49.127)$$

d'où:

$$x - v_1 t = v_2 \left(t - \frac{v_1 x}{c^2} \right) \Rightarrow x - v_1 t = v_2 t - \frac{v_2 v_1 x}{c^2} \Rightarrow x \left(1 + \frac{v_2 v_1}{c^2} \right) = t (v_2 + v_1) \quad (49.128)$$

Nous savons que $v = x/t$ d'où finalement la "**loi de compositions des vitesses relativistes**":

$$\boxed{V = \frac{x}{t} = \frac{v_1 + v_2}{1 + \frac{v_2 v_1}{c^2}}} \quad (49.129)$$

qui est donc la vitesse d'un corps en mouvement dans le référentiel en mouvement par rapport au référentiel au repos (ou autrement dit: vu par le référentiel au repos).

Et réciproquement vu de l'autre référentiel en mouvement, nous avons en faisant les mêmes développements (avec inversion des signes et des vitesses bien sûr):

$$\boxed{V' = \frac{x'}{t'} = \frac{v_2 - v_1}{1 - \frac{v_1 v_2}{c^2}}} \quad (49.130)$$

qui est donc la vitesse d'un corps en mouvement dans le référentiel au repos par rapport au référentiel en mouvement (ou autrement dit: vu par le référentiel en mouvement).

2.6. VARIATION RELATIVISTE DES LONGUEURS

Considérons maintenant que la longueur d'un objet est donnée par la distance entre ses deux extrémités A et B . Considérons cet objet AB immobile dans le référentiel O' en translation uniforme et orienté selon l'axe $O'X'$.

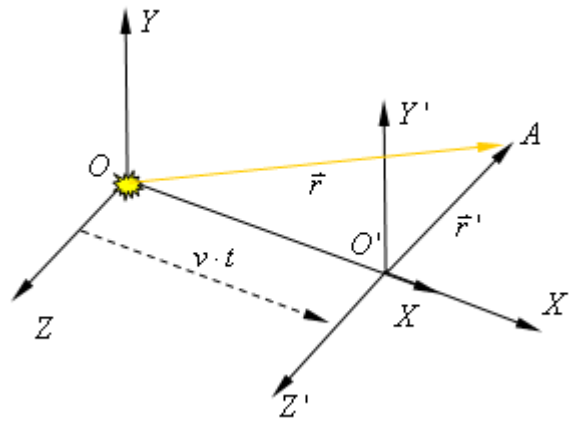


Figure: 49.3 - Configuration pour l'étude de la variation relativiste des longueurs

Sa longueur est donc la distance entre ses deux extrémités:

$$L' = x'_B - x'_A \quad (49.131)$$

Pour l'observateur O , l'objet est en mouvement. Les positions de A et B devraient donc être mesurées simultanément:

$$L = x_B - x_A \quad (49.132)$$

Il vient donc en utilisant la relation démontrée au début de ce chapitre:

$$x' = \gamma(x - vt) = \frac{x - vt}{\sqrt{1 - (v/c)^2}} \quad (49.133)$$

la différence suivante:

$$L' = x'_B - x'_A = \frac{x_B - vt}{\sqrt{1 - (v/c)^2}} - \frac{x_A - vt}{\sqrt{1 - (v/c)^2}} = \frac{x_B - x_A}{\sqrt{1 - (v/c)^2}} = \frac{L}{\sqrt{1 - (v/c)^2}} \quad (49.134)$$

d'où le résultat remarquable:

$$L = L' \sqrt{1 - (v/c)^2} \quad (49.135)$$

que nous retrouvons aussi fréquemment dans la littérature sous la forme suivante:

$$\Delta L = \Delta L' \sqrt{1 - \frac{v^2}{c^2}} \quad (49.136)$$

Ainsi, la longueur d'une règle observée dans un référentiel mobile par rapport au référentiel propre de la règle est inférieure à sa longueur propre (que l'on peut assimiler en toute généralité à une "**distance propre**"). En d'autres termes, la longueur d'un objet en déplacement mesuré par le référentiel immobile sera plus courte que sa grandeur réelle. Ce phénomène porte le nom de "contraction des longueurs".

2.7. VARIATION RELATIVISTE DU TEMPS

Un événement est un phénomène qui se produit en un endroit donné et à un instant donné. L'origine du temps étant difficile à préciser, nous préférons souvent définir la notion d'intervalle de temps comme le temps qui s'écoule entre deux événements comme il est fréquemment d'usage.

Considérons maintenant deux événements A et B consécutifs qui se produisent au même endroit x' (!) dans le référentiel en translation uniforme:

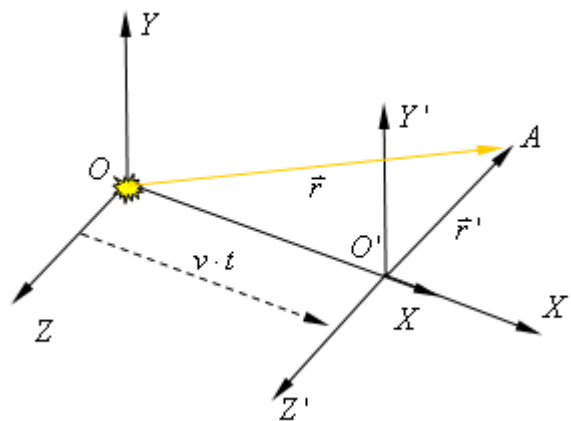


Figure: 49.4 - Configuration pour l'étude de la variation relativiste du temps

Pour l'observateur O' , l'intervalle de temps est simplement:

$$\Delta t' = t'_B - t'_A \quad (49.137)$$

Pour mesurer cet intervalle, l'observateur O dans le référentiel fixe, doit aussi imposer que x' est commun aux deux événements. Alors en utilisant la relation démontrée au début de ce chapitre:

$$t' = \gamma \left(t - \frac{v}{c^2} x \right) = \frac{t - \frac{v}{c^2} x}{\sqrt{1 - (v/c)^2}} \quad (49.138)$$

nous obtenons:

$$\Delta t = t_B - t_A = \frac{t'_B + \frac{vx'_1}{c^2}}{\sqrt{1 - (v/c)^2}} - \frac{t'_A + \frac{vx'_1}{c^2}}{\sqrt{1 - (v/c)^2}} = \frac{t'_B - t'_A}{\sqrt{1 - (v/c)^2}} = \frac{\Delta t'}{\sqrt{1 - (v/c)^2}} \quad (49.139)$$

d'où le résultat remarquable ci-dessous:

$$\Delta t = \frac{\Delta t'}{\sqrt{1 - (v/c)^2}} \quad (49.140)$$

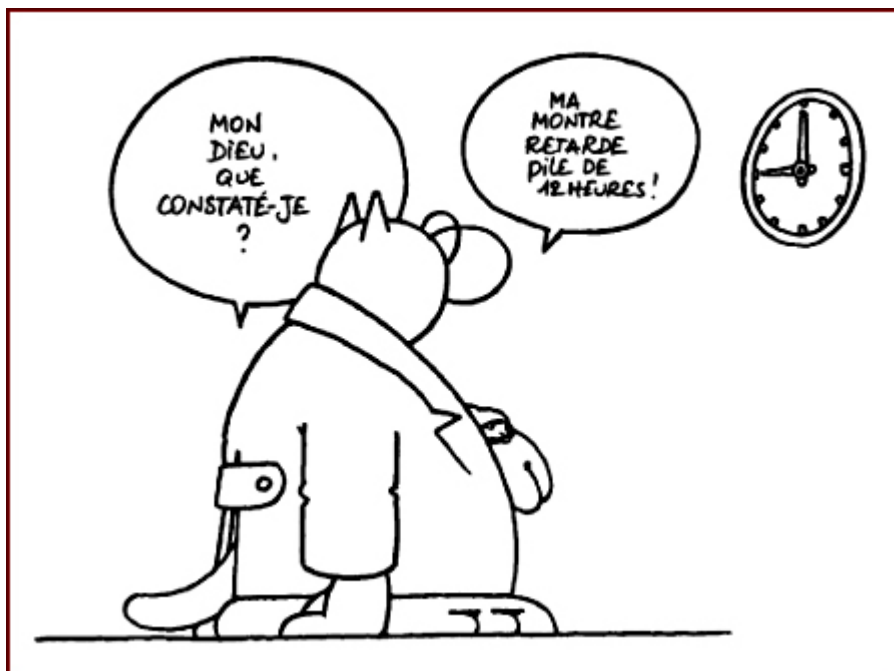
ce qui se note sous forme condensée traditionnelle:

$$t = \gamma t' \quad (49.141)$$

Nous en déduisons aussi en prenant un élément infinitésimal:

$$\frac{dt}{dt'} = \frac{1}{\sqrt{1 - v^2/c^2}} \quad (49.142)$$

Donc l'observateur O (immobile) mesure un intervalle de temps d'autant plus grand que le référentiel dans lequel se déroule le phénomène se déplace rapidement. Le temps dans le référentiel fixe (donc le "**temps propre**" du référentiel fixe!) semble comme dilaté par rapport à celui en vigueur dans le référentiel mobile (dont par rapport au "temps propre du référentiel mobile!").



Voyons un exemple d'application sympathique et (cependant) simplifié à l'extrême:

En 1971, une vérification expérimentale directe de la dilatation du temps fut effectuée. Deux avions à bord desquels avaient été placée une horloge atomique au césium pendant leurs vols commerciaux réguliers (l'un vers l'Est, l'autre vers l'Ouest) comparèrent leur horloge à une troisième horloge atomique restée au sol. Cette expérimentation devenue célèbre par le temps est appelée "**expérience de Hafele-Keating**".

L'avion volant vers l'Est perd 59 [ns] alors que l'avion volant vers l'Ouest gagna 279 [ns] (la Terre tourne sur elle-même en un jour, d'Ouest en Est). Il fut donc mesuré une différence totale de:

$$\Delta t = t_{Est} - t_{Ouest} = (-59) - (+279) \cong -332 \text{ [ns]} \quad (49.143)$$

entre les deux horloges cette différence est nettement supérieure à celle qu'implique la relativité restreinte.

Analysons l'expérience en considérant que tous les référentiels sont inertiels (ce qui élimine donc la relativité générale).

Remarque: En toute rigueur l'effet de la relativité générale (ralentissement des horloges en fonction de l'altitude conformément à l'effet Einstein vu dans le chapitre de Relativité Générale) n'est absolument pas négligeable puisqu'il est d'une amplitude équivalente à celle de la relativité restreinte.

Considérons pour l'étude trois repères inertiels, un situé au pôle Nord, un sur Terre (ailleurs qu'au pôle nord dans l'idée!) et un dans un avion. Les intervalles de temps t_{Nord} , t_{Terre} et t_{Avion} respectivement (que nous noterons de manière abrégée t_N , t_T , t_A pour la suite), sont reliés entre eux par la relation démontrée précédemment (donc le pôle nord est pris comme le référentiel au repos dans cette expérience et donc la référence du temps propre!):

$$t_N = \gamma_T t_T = \gamma_A t_A \quad (49.144)$$

où nous avons donc:

$$\gamma_T = \frac{1}{\sqrt{1 - (v_T/c)^2}} \quad \gamma_A = \frac{1}{\sqrt{1 - (v_A/c)^2}} \quad (49.145)$$

Les repères sur Terre et dans l'avion ont donc des vitesses relatives v_T et v_A par rapport au pôle nord. Les temps en avion et sur Terre sont donc reliés par:

$$t_A = \frac{\gamma_T}{\gamma_A} t_T = \frac{\sqrt{1 - (v_T/c)^2}}{\sqrt{1 - (v_A/c)^2}} t_T \quad (49.146)$$

Nous allons réécrire cette relation:

$$t_A = \frac{\gamma_T}{\gamma_A} t_T = \sqrt{\frac{1 - (v_A/c)^2 - (v_T/c)^2 + (v_A/c)^2}{1 - (v_A/c)^2}} t_T = \sqrt{1 + \frac{(v_A/c)^2 - (v_T/c)^2}{1 - (v_A/c)^2}} t_T \quad (49.147)$$

Nous allons accepter l'approximation suivante:

$$t_A = \frac{\gamma_T}{\gamma_A} t_T \cong \sqrt{1 + \left[(v_A/c)^2 - (v_T/c)^2 \right]} t_T \quad (49.148)$$

où nous avons supposé au dénominateur que:

$$\left(v_A/c\right)^2 \ll 1 \quad (49.149)$$

Pour les racines dont la valeur est de toute façon proche de 1 (puisque c est beaucoup plus grand que les vitesses relatives considérées), nous pouvons faire un développement de Taylor au deuxième ordre lorsque x tend vers zéro:

$$\sqrt{1+x} \cong (1+x) + \frac{1}{1!} \frac{1}{2\sqrt{1+a}} (x-a) \Big|_{a=0} = 1 + \frac{1}{2}x \quad (49.150)$$

Nous pouvons alors écrire:

$$t_A = \frac{\gamma_T}{\gamma_A} t_T \cong \left(1 + \frac{\left(v_A/c\right)^2 - \left(v_T/c\right)^2}{2} \right) t_T \quad (49.151)$$

Grâce à ces approximations successives, nous pouvons facilement écrire la différence entre les deux horloges qui est alors de:

$$\Delta t_{AT} = t_A - t_T \cong t_T \frac{\left(v_A/c\right)^2 - \left(v_T/c\right)^2}{2} \quad (49.152)$$

Selon les hypothèses initiales, la vitesse de croisière des deux avions par rapport au sol est constante et sera notée v . La vitesse de chaque avion (non relativiste selon les approximations précédentes!) est alors:

$$v_A = v_T - v \quad (49.153)$$

suivant que l'avion va vers l'Est et:

$$v_A = v_T + v \quad (49.154)$$

suivant que l'avion va respectivement vers l'Ouest. Alors:

$$\begin{aligned} \Delta t = t_{Est} - t_{Ouest} &= \left(t_T \frac{\left(\frac{v_T - v}{c}\right)^2 - \left(\frac{v_T}{c}\right)^2}{2} \right) - \left(t_T \frac{\left(\frac{v_T + v}{c}\right)^2 - \left(\frac{v_T}{c}\right)^2}{2} \right) \\ &= t_T \left[\frac{\left(\frac{v_T - v}{c}\right)^2}{2} - \frac{\left(\frac{v_T + v}{c}\right)^2}{2} \right] \end{aligned} \quad (49.155)$$

Nous allons considérer que (c'est assez grossier...):

$$\frac{v_T^2}{c^2} \ll \frac{2vv_T}{c^2} \quad (49.156)$$

$$\frac{v^2}{c^2} \ll \frac{2vv_T}{c^2}$$

Donc il reste:

$$\Delta t = t_{Est} - t_{Ouest} \cong t_T \left[\frac{-\frac{2vv_T}{c^2}}{2} - \frac{\frac{2vv_T}{c^2}}{2} \right] = t_T \left(-\frac{vv_T}{c^2} - \frac{vv_T}{c^2} \right) = -t_T \frac{2vv_T}{c^2} \quad (49.157)$$

Nous voyons bien évidemment qu'avec les approximations effectuées, nous perdons l'asymétrie de la dilatation du temps entre l'Est et l'Ouest. Celui que cela dérange peut appliquer alors directement les valeurs numériques à la relation antérieure.

Le résultat précédant des approximations successives permet déjà de voir de manière formelle et rapide que le signe sera en accord avec le résultat expérimental.

Pour une application pratique, nous prendrons la vitesse constante des avions commerciaux de l'époque qui valait:

$$v \cong 220 \left[m \cdot s^{-1} \right] \quad (49.158)$$

et le voyage total des avions dura 41 heures selon la mesure au sol soit:

$$t_T = 147'600 [s] \quad (49.159)$$

et un point de la surface Terrestre va à la vitesse:

$$v_T = \frac{2\pi R}{t_{jour}} \quad (49.160)$$

où le rayon de la Terre étant de 6'371 kilomètres (cela suppose que les avions sont sur le rayon de l'équateur!). Nous avons donc en application numérique:

$$\Delta t = -t_T \frac{2vv_T}{c^2} = -147'600 \cdot \frac{2 \cdot 220 \cdot \frac{2\pi \cdot 6'371'000}{24 \cdot 60 \cdot 60}}{299'792'458^2} = -147'600 \frac{2 \cdot 220 \cdot 463}{299'792'458^2} \cong -334 [ns] \quad (49.161)$$

ce qui mène à un résultat très proche de la mesure qui fut effectuée.

Et en utilisant directement la version non approximée:

$$\Delta t = t_{Est} - t_{Ouest} = \left(t_T \frac{\left(\frac{v_T - v}{c} \right)^2 - \left(\frac{v_T}{c} \right)^2}{2} \right) - \left(t_T \frac{\left(\frac{v_T + v}{c} \right)^2 - \left(\frac{v_T}{c} \right)^2}{2} \right) \quad (49.162)$$

$$= -92 - (+173) = -265 [ns]$$

où nous avons pris cette fois-ci la vitesse de rotation de la terre à la latitude conforme à l'expérience faite en 1971:

$$v_T = \frac{2\pi R \cos(38^\circ)}{t_{jour}} \cong 365 [m \cdot s^{-1}] \quad (49.163)$$

Donc nous voyons que le résultat n'est dès lors plus très conforme à l'expérience! Effectivement, il faut maintenant prendre en compte dans ce cas non approximé l'accélération du temps dû à la gravité. Nous allons devoir utiliser la relation de l'effet Einstein démontrée dans le chapitre de Relativité Générale:

$$\frac{\Delta t}{t_0} = \frac{gh}{c^2} \quad (49.164)$$

qui exprime donc que le temps au sol s'écoule moins rapidement que le temps à l'altitude h .

D'après le compte-rendu de l'expérience, les avions ont volé à 10'000 [m] d'altitude. Ce qui donne (l'accélération g n'est pas la même au sol qu'à l'altitude pour rappel!) une accélération du temps:

$$\Delta t = \frac{gh}{c^2} t_0 \cong \frac{9.5 \cdot 10^4}{299792458^2} \cdot 147'600 = +156 [ns] \quad (49.165)$$

Or, nous voyons que les deux avions étant tous deux à la même hauteur, nous avons toujours:

$$\Delta t = t_{Est} - t_{Ouest} = (-92 - 156) - (+173 - 156) = -265 [ns] \quad (49.166)$$

Donc, soit il y a d'autres effets, de l'ordre de la Relativité Générale, qui devraient être pris en compte pour expliquer les 67 [ns] de différence par rapport à l'expérience, soit il s'agit d'un problème de précision des appareils de l'époque.

Au fait, nous verrons une étude détaillée de cette expérience dans le chapitre de Relativité Générale et nous verrons que les valeurs théoriques sont alors en très très bon accord avec les résultats expérimentaux.

2.7.1. PARADOXE DES Jumeaux

Nous pouvons déjà considérer le fameux paradoxe des jumeaux dans le cas de la relativité restreinte pour montrer que cette théorie ne s'applique pas aux systèmes non inertiels. Il s'agit d'un approche grossière (sachant que la manière rigoureuse sera abordée dans le chapitre de Relativité Générale).

Considérons une fusée décollant au temps t nul de la Terre et accélérant à 20 fois l'accélération de la gravité terrestre g jusqu'à une vitesse de croisière 90% de celle de la lumière c . Supposons que la fusée continue à cette vitesse pendant une année terrestre et décélère avec la même intensité pour reprendre son voyage vers la Terre et accélère à nouveau pour lors de son approche vers le Terre décélérer encore une fois pour avoir au final une vitesse nulle.

$$\Delta v = a \cdot \Delta t = 20g(\Delta t_{acc}) = 0.9c \Rightarrow \Delta t_{acc} = \sqrt{\frac{0.9c}{20g}} = \sqrt{\frac{0.9c}{20 \cdot 10}} \cong 1.35 \cdot 10^6 \cong 15.6 \text{ jours} \quad (49.167)$$

Ainsi, le temps propre passé pour un terrien resté sur Terre est au total de:

$$\Delta t'_T \cong 15.6 \cdot 4 + 365 \cdot 2 \cong 792.4 \text{ jours} \quad (49.168)$$

Pour le voyageur dans la fusée, le temps propre pendant les phases d'accélération sera donné grossièrement par:

$$dt'_{a,f} = \sqrt{1 - \frac{v^2}{c^2}} dt = \sqrt{1 - \frac{1}{c^2} (v)^2} dt = \sqrt{1 - \frac{1}{c^2} (gt)^2} dt = \sqrt{1 - \frac{g^2}{c^2} t^2} dt \quad (49.169)$$

Soit en intégrant (en utilisant la primitive usuelle démontrée dans le chapitre de Calcul Différentiel Et Intégral) pour une des phases d'accélération de la fusée cela donne:

$$\begin{aligned} \Delta t'_{a,f} &= \int_0^{\Delta t_{acc}} \sqrt{1 - \frac{g^2}{c^2} t^2} dt = \frac{g}{c} \int_0^{\Delta t_{acc}} \sqrt{\left(\frac{c}{g}\right)^2 - t^2} dt \\ &= \left[\frac{\left(\frac{c}{g}\right)^2}{2} \arcsin\left(\frac{t}{\left(\frac{c}{g}\right)}\right) + \frac{t}{2} \sqrt{\left(\frac{c}{g}\right)^2 - t^2} \right]_0^{1.3510^6} \\ &\cong 1.13 \cdot 10^6 \cong 13.1 \text{ jours} \end{aligned} \quad (49.170)$$

Et le temps propre pour la partie en vitesse de croisière constante:

$$\Delta t'_{croisière,f} = \int_0^{365j} \sqrt{1 - \frac{v^2}{c^2}} dt \cong \int_0^{365j} \sqrt{1 - 0.81} dt = \sqrt{0.19} \int_0^{365j} dt \cong 159.1 \text{ jours} \quad (49.171)$$

Et donc le temps propre total dans la fusée est alors:

$$\Delta t'_f = 13.1 \cdot 4 + 159.1 \cdot 2 = 370.6 \text{ jours} \quad (49.172)$$

Donc comparé à la personne restée sur Terre, celle qui était dans la fusée à vieilli environ deux fois moins!!! Il s'agit d'un paradoxe (plutôt un "sophisme" en réalité) car on ne peut appliquer la relativité restreinte à des référentiels non inertiels. Il n'empêche que même avec la Relativité Générale, il y a une différence de temps!

2.8. VARIATION RELATIVISTE DE LA MASSE

Bon d'abord attention le titre est abusif par tradition! Nous verrons un peu plus loin pourquoi.

En attendant, imaginons une collision frontale entre deux objets identiques (1), (2) ayant dans le référentiel R_0 des vitesses égales mais opposées. Nous supposons que cette collision est élastique, c'est-à-dire que l'énergie cinétique et la quantité de mouvement sont conservées.

Avant le choc, les composantes des vitesses des objets (1) et (2) sont:

$$1 = \begin{pmatrix} u_1 \\ v_1 \end{pmatrix} \quad 2 = \begin{pmatrix} u_2 \\ v_2 \end{pmatrix} = \begin{pmatrix} -u_1 \\ -v_1 \end{pmatrix} \quad (49.173)$$

comme indiqué ci-dessous:

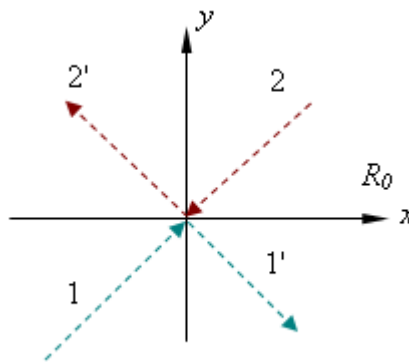


Figure: 49.5 - Configuration pour l'étude de la variation relativiste de la masse

Après le choc, nous avons:

$$1' = \begin{pmatrix} u'_1 \\ v'_1 \end{pmatrix} = \begin{pmatrix} u_1 \\ -v_1 \end{pmatrix} \quad 2' = \begin{pmatrix} u'_2 \\ v'_2 \end{pmatrix} = \begin{pmatrix} u_2 \\ -v_2 \end{pmatrix} = \begin{pmatrix} -u_1 \\ v_1 \end{pmatrix} \quad (49.174)$$

Nous allons maintenant faire la transformation de Lorentz suivante :

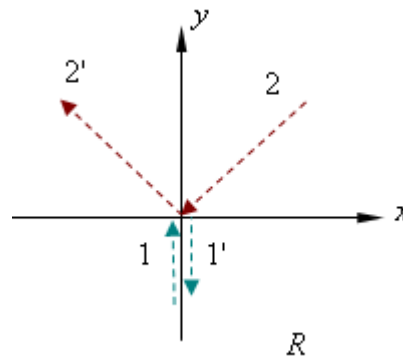
- Nous nous donnons un autre référentiel R et supposons que les référentiels R_0 et R sont en translation uniforme de vitesse u_1 selon l'axe OX dans le sens positif (c'est-à-dire dans la même direction et à la même vitesse horizontale que la particule 1).
- Pour notre particule 1 sa trajectoire est devenue telle qu'elle ne présente pas de vitesse selon l'axe OX .

Allons-y! Plaçons-nous dans un référentiel R qui se déplace par rapport à R_0 avec la vitesse u_1 suivant OX , les composantes des vitesses dans ce référentiel sont avant choc:

$$1 = \begin{pmatrix} U_1 \\ V_1 \end{pmatrix} \quad 2 = \begin{pmatrix} U_2 \\ V_2 \end{pmatrix} \quad (49.175)$$

et après le choc:

$$1' = \begin{pmatrix} U'_1 \\ V'_1 \end{pmatrix} \quad 2' = \begin{pmatrix} U'_2 \\ V'_2 \end{pmatrix}$$



(49.176)

Nous avons donc trivialement dans le référentiel R :

$$U'_1 = U_1 = 0 \quad V'_1 = -V_1 \quad U'_2 = U_2 \quad V'_2 = -V_2 \quad (49.177)$$

mais en appliquant la loi de composition des vitesses démontrée plus haut:

$$V'_1 = \frac{x'}{t'} = \frac{v_2 - v_1}{1 - \frac{v_1 v_2}{c^2}} \quad (49.178)$$

pour les composantes de l'axe horizontal nous avons toujours dans le référentiel R :

$$U_1 = 0$$

$$U_2 = \frac{u_2 - u_1}{1 - \frac{u_2 u_1}{c^2}} = \frac{(-u_1) - u_1}{1 - \frac{(-u_1)u_1}{c^2}} = -\frac{2u_1}{1 + \frac{u_1^2}{c^2}} \quad (49.179)$$

et pour le mouvement vertical, nous avons vu plus haut que:

$$v'_{y'} = \frac{v_y}{\gamma \left(1 - \left(v/c^2\right)v_x\right)} \quad (49.180)$$

Ainsi, il vient:

$$V'_1 = \frac{v_1}{\gamma \left(1 - \left(u_1/c^2\right)u_1\right)} = \frac{v_1}{\gamma \left(1 - u_1^2/c^2\right)} = \frac{\sqrt{1 - u_1^2/c^2} v_1}{\left(1 - u_1^2/c^2\right)} = \frac{v_1}{\sqrt{1 - u_1^2/c^2}} = v_1 \gamma$$

$$V'_2 = \frac{v_2}{\gamma \left(1 - \left(u_1/c^2\right)u_2\right)} = \frac{v_2}{\gamma \left(1 - u_1 u_2/c^2\right)} = \frac{-v_1}{\gamma} \frac{1}{1 + \frac{u_1^2}{c^2}} \quad (49.181)$$

En passant de R_0 à R , la composante suivant y de la quantité de mouvement totale doit rester nulle (comme c'était le cas dans R_0 initialement). Or:

$$V_1 + V_2 = \frac{v_1}{\gamma} \frac{1}{1 - \frac{u_1^2}{c^2}} - \frac{v_1}{\gamma} \frac{1}{1 + \frac{u_1^2}{c^2}} \neq 0 \quad (49.182)$$

Pour sortir de cette impasse, il faut admettre que les masses respectives m_1, m_2 des objets (1) et (2) ne peuvent être identiques dans R . Alors cela nous amène à imposer:

$$|m_1 V_1| = \frac{m_1 v_1}{\gamma} \frac{1}{1 - \frac{u_1^2}{c^2}} = |m_2 V_2| = \frac{m_2 v_1}{\gamma} \frac{1}{1 + \frac{u_1^2}{c^2}} \quad (49.183)$$

ce qui entraîne:

$$\frac{m_2}{m_1} = \frac{1 + \frac{u_1^2}{c^2}}{1 - \frac{u_1^2}{c^2}} \quad (49.184)$$

Dans R , le carré de la norme des vitesses des deux objets donne:

$$\begin{aligned} U_1^2 + V_1^2 &= V_1^2 = v_1^2 \gamma^2 \\ U_2^2 + V_2^2 &= \left(\frac{-2u_1}{1 + \frac{u_1^2}{c^2}} \right)^2 + \left(\frac{-v_1}{\gamma} \frac{1}{1 + \frac{u_1^2}{c^2}} \right)^2 = \frac{4u_1^2 + v_1^2 \left(1 - \frac{u_1^2}{c^2} \right)}{\left(1 + \frac{u_1^2}{c^2} \right)^2} = \frac{4u_1^2 + V_1^2 \left(1 - \frac{u_1^2}{c^2} \right)}{\left(1 + \frac{u_1^2}{c^2} \right)^2} \end{aligned} \quad (49.185)$$

La dernière relation peut s'écrire:

$$\frac{U_2^2 + V_2^2}{c^2} \left(1 + \frac{u_1^2}{c^2} \right)^2 = 4 \frac{u_1^2}{c^2} + \frac{V_1^2}{c^2} \left(1 - \frac{u_1^2}{c^2} \right)^2 = \left(1 + \frac{u_1^2}{c^2} \right)^2 - \left(1 - \frac{u_1^2}{c^2} \right)^2 + \frac{V_1^2}{c^2} \left(1 - \frac{u_1^2}{c^2} \right)^2 \quad (49.186)$$

de sorte qu'après réarrangement et factorisation:

$$\left(\frac{U_2^2 + V_2^2}{c^2} - 1 \right) \left(1 + \frac{u_1^2}{c^2} \right)^2 = \left(1 - \frac{u_1^2}{c^2} \right)^2 \left(\frac{V_1^2}{c^2} - 1 \right) \quad (49.187)$$

Soit:

$$\frac{\left(1 + \frac{u_1^2}{c^2} \right)^2}{\left(1 - \frac{u_1^2}{c^2} \right)^2} = \frac{1 - \frac{V_1^2}{c^2}}{1 - \frac{U_2^2 + V_2^2}{c^2}} = \frac{\Gamma_2^2}{\Gamma_1^2} \quad (49.188)$$

où nous avons posé:

$$\Gamma_1^2 = \frac{1}{1 - \frac{V_1^2}{c^2}} \quad \Gamma_2^2 = \frac{1}{1 - \frac{U_2^2 + V_2^2}{c^2}} \quad (49.189)$$

Nous trouvons ainsi:

$$\frac{m_2}{m_1} = \frac{1 + \frac{u_1^2}{c^2}}{1 - \frac{u_1^2}{c^2}} = \frac{\Gamma_2^2}{\Gamma_1^2} \Leftrightarrow \frac{m_2}{\Gamma_2^2} = \frac{m_1}{\Gamma_1^2} \quad (49.190)$$

Nous poserons maintenant:

$$\frac{m_2}{\Gamma_2^2} = \frac{m_1}{\Gamma_1^2} = m(\Gamma = 1) = m_0 \quad (49.191)$$

où m_0 est évidemment la masse au repos de l'un ou l'autre des objets identiques (1) et (2).

Le raisonnement que nous venons de faire sur un exemple simple montre que l'inertie (et non la masse!) d'un objet semble dépendre de sa vitesse v dans un référentiel donné. Au fait, pour être plus exact, c'est le facteur de Michelson-Morley qui varie et non pas la masse en elle-même car celle-ci est un invariant relativiste!

D'une façon générale, m_0 étant la "**masse au repos**":

$$m = \gamma m_0 = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (49.192)$$

Puisque la masse est donc fonction de v (en apparence), certains physiciens notent la masse au repos comme une fonction, c'est-à-dire: $m(0)$. Mais il est plutôt d'usage de la noter m_0 afin de ne pas avoir trop de parenthèses dans les développements...

Ainsi, le facteur de Michelson-Morley tend vers l'infini lorsque la vitesse tend vers la vitesse c de la lumière dans le vide. C'est une raison supplémentaire pour affirmer que c est la limite supérieure assignée à la vitesse de tout objet matériel, ce qui est conforme à la fois à l'expérience et aux conséquences déjà formulées de la transformation de Lorentz.

2.8.1. ÉQUIVALENCE MASSE-ÉNERGIE

Sous l'action d'une force F , la vitesse d'une masse m augmente ou diminue sur chaque portion de la trajectoire. Le travail de la composante E_c peut s'interpréter alors en énergie cinétique E_c :

Dans la théorie relativiste, la masse varie avec la vitesse, donc:

$$E_c = \int F_t \cdot dx = \int_0^v \frac{dp}{dt} \cdot dx = \int_0^v dp \cdot \frac{dx}{dt} = \int_0^v d(mv) \cdot v \quad (49.193)$$

L'intégration par parties (cf. chapitre de Calcul Différentiel Et Intégral):

$$\int a \cdot db = a \cdot b - \int b \cdot da \quad (49.194)$$

nous donne:

$$\begin{aligned} E_c &= mv^2 - \int_0^v mvdv = \frac{m_0 v^2}{\sqrt{1-(v/c)^2}} - \int_0^v \frac{m_0 v}{\sqrt{1-(v/c)^2}} dv \\ &= \frac{m_0 v^2}{\sqrt{1-(v/c)^2}} - \frac{m_0 c^2}{2} \int_0^{(v/c)^2} \frac{1}{\sqrt{1-x}} dx = \frac{m_0 v^2}{\sqrt{1-(v/c)^2}} - \frac{m_0 c^2}{2} [-2\sqrt{1-x}]_0^{(v/c)^2} \\ &= \frac{m_0 v^2}{\sqrt{1-(v/c)^2}} + m_0 c^2 [\sqrt{1-x}]_0^{(v/c)^2} = \frac{m_0 v^2}{\sqrt{1-(v/c)^2}} + m_0 c^2 (\sqrt{1-(v/c)^2} - 1) \\ &= \frac{m_0 v^2}{\sqrt{1-(v/c)^2}} + m_0 c^2 \sqrt{1-(v/c)^2} - m_0 c^2 = \frac{m_0 v^2 + m_0 c^2 (1-(v/c)^2)}{\sqrt{1-(v/c)^2}} - m_0 c^2 \\ &= \frac{m_0 v^2 + m_0 c^2 - m_0 v^2}{\sqrt{1-(v/c)^2}} - m_0 c^2 = \frac{m_0 c^2}{\sqrt{1-(v/c)^2}} - m_0 c^2 = mc^2 - m_0 c^2 = \Delta mc^2 \end{aligned} \quad (49.195)$$

Le gain d'énergie cinétique d'une particule peut donc être considéré comme gain de sa masse. Puisque m_0 est la masse au repos, la quantité $m_0 c^2$ est appelée "**énergie au repos**" de la particule.

Nous avons donc:

$$E_c = mc^2 - m_0 c^2 \Leftrightarrow mc^2 = E_c + m_0 c^2 \quad (49.196)$$

où E_c représente donc l'énergie de mouvement (l'énergie cinétique).

La somme:

$$E_c + m_0 c^2 \quad (49.197)$$

représente donc l'énergie totale E de la particule en absence du champ de potentiel. Ce qui nous amène à écrire:

$$\boxed{mc^2 = E_c + m_0 c^2 = E \Rightarrow E = mc^2} \quad (49.198)$$

Finalement nous aurions aussi pu obtenir le même résultat autrement:



2.8.2. LAGRANGIEN RELATIVISTE

Les développements suivants vont nous permettre dans l'étude de l'électrodynamique (si ce chapitre n'a pas encore été lu), de déterminer l'expression du tenseur du champ électromagnétique ainsi qu'en physique quantique relativiste de déterminer l'équation de Klein-Gordon avec champ magnétique. Il faut donc bien lire ce qui va suivre.

En relativité, nous voulons donc que les équations du mouvement aient la même forme dans tous les référentiels inertiels. Pour cela, il faut que l'action S (cf. chapitre de Mécanique Analytique) soit donc invariante par rapport aux transformations de Lorentz. Guidés par ce principe, essayons d'obtenir l'action d'une particule libre. Supposons que l'action soit dans le référentiel O' :

$$S_0 = \int L_0 dt' = -mc^2 \int dt' \quad (49.199)$$

Remarques:

R1. Le choix du signe moins deviendra évident lors de notre étude de l'électrodynamique.

R2. La notation L_0 au lieu de L pour le lagrangien permet simplement de mettre en évidence qu'il s'agit d'un cas d'étude où le système est libre. Cette distinction de notation sera utile lors de notre étude de la relativité générale et de la détermination du tenseur du champ électromagnétique.

R3. Nous ne sommes pas censés savoir à quel type de masse nous avons affaire (masse au repos ou inertielle) d'où le fait que dans l'ignorance, nous travaillerons avec la masse inertielle m quitte à corriger cette hypothèse plus loin.

Et rappelons que:

$$dt = \frac{dt'}{\sqrt{1-v^2/c^2}} \Leftrightarrow dt' = \sqrt{1-v^2/c^2} dt \quad (49.200)$$

Dans le référentiel O , nous avons alors "l'action invariante de Lorentz":

$$S = -mc^2 \int \sqrt{1-v^2/c^2} dt \quad (49.201)$$

Donc selon notre hypothèse initiale, nous avons pour le lagrangien relativiste (en l'absence du champ de potentiel donc... puisque le système est "libre"):

$$L_0 = -mc^2 \sqrt{1-v^2/c^2} \quad (49.202)$$

Dans l'approximation non-relativiste $v \ll c$, nous avons selon le développement de Maclaurin:

$$L_0 = -mc^2 \sqrt{1-v^2/c^2} \cong -mc^2 \left(1 - \frac{1}{2} \frac{v^2}{c^2} \right) = -mc^2 + \frac{1}{2} mv^2 \quad (49.203)$$

Nous retrouvons donc le lagrangien habituel d'un système libre en mouvement mais plus une constante ($-mc^2$) qui n'affecte cependant pas les équations du mouvement que nous obtenons en mécanique classique mais qui nous sera absolument nécessaire en électrodynamique.

Rappelons maintenant que le moment généralisé (cf. chapitre de Mécanique Analytique) est défini par:

$$p_i = \frac{\partial L}{\partial \dot{q}_i} \quad (49.204)$$

Nous allons voir maintenant que cette définition n'est pas fortuite. Effectivement:

$$\begin{aligned} p = \frac{\partial L_0}{\partial v} &= p = \frac{\partial}{\partial v} \left(-mc^2 \sqrt{1-v^2/c^2} \right) = -mc^2 \frac{\partial}{\partial v} \left(\sqrt{1-v^2/c^2} \right) = mc^2 \frac{v}{c^2 \sqrt{1-v^2/c^2}} \\ &= \frac{mv}{\sqrt{1-v^2/c^2}} \end{aligned} \quad (49.205)$$

L'hamiltonien (cf. chapitre de Mécanique Analytique) vaut:

$$H(q_i, p_i) = \sum_{i=1}^n \dot{q}_i p_i - L(q_i, \dot{q}_i) \quad (49.206)$$

ce qui donne:

$$H_0 = v \cdot p - L_0 = \frac{mv^2}{\sqrt{1-v^2/c^2}} + mc^2 \sqrt{1-v^2/c^2} = \frac{mc^2}{\sqrt{1-v^2/c^2}} \quad (49.207)$$

L'hamiltonien est dans ce cas égal à l'énergie totale de la particule. Son expression nous conduit à changer quelque peu notre hypothèse initiale et finalement à écrire m_0 au lieu de m dans l'expression de l'action S .

Ainsi, nous avons finalement:

$$L_0 = -m_0 c^2 \sqrt{1 - v^2 / c^2} \quad (49.208)$$

et:

$$H_0 = \frac{m_0 c^2}{\sqrt{1 - v^2 / c^2}} = m c^2 \quad (49.209)$$

Dans l'approximation non relativiste $v \ll c$, H_0 devient avec un développement de Maclaurin (cf. chapitre sur les Suites Et Séries):

$$H_0 \cong m_0 c^2 + \frac{1}{2} m_0 v^2 \quad (49.210)$$

Nous reconnaissons l'énergie cinétique usuelle, plus une constante: l'énergie au repos. Ce qui correspond bien aux calculs que nous avons faits avant où nous avons obtenu:

$$m c^2 = E_c + m_0 c^2 = E \Rightarrow E = m c^2 \quad (49.211)$$

2.9. QUANTITÉ DE MOUVEMENT RELATIVISTE

L'énergie totale E et la quantité de mouvement $p = m \cdot v$ d'une particule peuvent donc prendre n'importe quelle valeur positive (si la vitesse tend vers la valeur limite c , la masse s'adapte pour que le produit $p = m \cdot v$ ne soit pas borné).

Dans l'expression de E , nous pouvons remplacer la vitesse v^2 par une fonction de p^2 :

$$p^2 = m^2 \cdot v^2 = \frac{E^2}{c^4} \cdot v^2 \Rightarrow v^2 = \frac{p^2 \cdot c^4}{E^2} \quad (49.212)$$

introduit dans:

$$E = m \cdot c^2 = \frac{m_0 \cdot c^2}{\sqrt{1 - v^2 / c^2}} \quad (49.213)$$

nous avons:

$$\begin{aligned} E &= \frac{m_0 \cdot c^2}{\sqrt{1 - (p^2 c^4) / (E^2 c^2)}} = \frac{m_0 \cdot c^2}{\sqrt{E^2 - p^2 c^2}} \cdot E \\ \Rightarrow \sqrt{E^2 - p^2 c^2} &= m_0 \cdot c^2 \Rightarrow E^2 - p^2 c^2 = m_0^2 \cdot c^4 \end{aligned} \quad (49.214)$$

d'où (nous reviendrons sur cette relation de la plus haute importance lors de notre démonstration de la relation d'Einstein):

$$E = c \sqrt{m_0^2 \cdot c^2 + p^2} \quad (49.215)$$

Nous n'avons pas gardé la partie négative de l'égalité précédente car elle n'a aucun sens en physique classique. Cependant, lorsque nous étudierons la physique quantique relativiste, il s'avérera indispensable de la préserver sinon quoi nous arriverons à des absurdités.

Cependant, nous pouvons bien évidemment écrire cette dernière relation aussi sous la forme:

$$E^2 = m_0^2 c^4 + p^2 c^2 \quad (49.216)$$

ou encore (beurk!):

$$E^2 = (E(m_0))^2 + (E(c))^2 \quad (49.217)$$

En d'autres termes, l'énergie totale d'une particule en mouvement est égale à son énergie de masse additionnée par son énergie cinétique (rien de fondamentalement nouveau).

Cette relation présente deux cas limites où nous pouvons réduire la formule:

C1. Pour une particule au repos ($p=0$), nous pouvons réduire l'expression à $E = m_0 c^2$ (en omettant l'énergie négative...pour l'instant).

C2. Nous pouvons appliquer l'équation à une particule sans masse de manière à éliminer le premier terme, ce qui nous donne alors $E = pc$. Un photon, par exemple, a une masse nulle au repos mais il n'est jamais au repos... Par définition, c'est un quantum d'énergie, son énergie cinétique n'est donc jamais nulle et il a donc une masse correspondant à son énergie cinétique. Ainsi, une particule de masse nulle au repos se déplace à la vitesse de la lumière, quel que soit le référentiel choisi! À l'inverse, une particule ayant une masse au repos non-nulle ne pourra jamais atteindre la vitesse de la lumière dans aucun référentiel.

Remarques:

R1. Comme nous le démontrerons plus loin (voir la "relation d'Einstein"), à partir de la définition de la loi de Planck, nous pourrions écrire $E = pc = h\nu$.

R2. La masse du photon peut difficilement être non nulle! Effectivement, la théorie quantique serait alors fautive dans le cas contraire. Or, elle n'a jamais été mise à défaut à ce jour (cf. chapitre de Physique Quantique Ondulatoire). On aurait également un léger changement sur la loi des forces électrostatiques et gravitationnelles selon le potentiel de Yukawa (cf. chapitre de Physique Quantique Des Champs) et cela se remarquerait.

Cherchons maintenant les relations entre p et p' ainsi qu'entre E et E' , pour qu'il soit possible à O' d'écrire:

$$E' = c \sqrt{m_0^2 \cdot c^2 + p'^2} \quad (49.218)$$

Nous commençons alors à nous débarrasser de la racine carrée:

$$E = c \sqrt{m_0^2 \cdot c^2 + p^2} \Rightarrow p^2 - \frac{E^2}{c^2} = -m_0^2 \cdot c^2 \quad (49.219)$$

Si O écrit:

$$p_x^2 + p_y^2 + p_z^2 - \frac{E^2}{c^2} = -m_0^2 \cdot c^2 \quad (49.220)$$

O' doit pouvoir écrire:

$$p_{x'}^2 + p_{y'}^2 + p_{z'}^2 - \frac{E'^2}{c^2} = -m_0^2 \cdot c^2 \quad (49.221)$$

Nous avons donc:

$$p_x^2 + p_y^2 + p_z^2 - \frac{E^2}{c^2} = p_{x'}^2 + p_{y'}^2 + p_{z'}^2 - \frac{E'^2}{c^2} \quad (49.222)$$

Si nous comparons:

$$x \equiv p_x, \quad y \equiv p_y, \quad z \equiv p_z \quad \text{et} \quad c \cdot t \equiv \frac{E}{c} \quad (49.223)$$

nous obtenons des expressions semblables à celles utilisées pour les transformations de Lorentz des composantes spatiales et temporelles. Nous pouvons alors écrire, par similitude, que les transformations pour la quantité de mouvement et l'énergie sont dès lors données par:

$$\begin{aligned} p_{x'} &= \frac{p_x - v \cdot \frac{E}{c^2}}{\sqrt{1 - v^2/c^2}} \Leftrightarrow p_x = \frac{p_{x'} + v \cdot \frac{E'}{c^2}}{\sqrt{1 - v^2/c^2}} \\ p_{y'} &= p_y \Leftrightarrow p_y = p_{y'} \\ p_{z'} &= p_z \Leftrightarrow p_z = p_{z'} \\ E' &= \frac{E - v \cdot p_x}{\sqrt{1 - v^2/c^2}} \Leftrightarrow E = \frac{E' + v \cdot p_{x'}}{\sqrt{1 - v^2/c^2}} \end{aligned} \quad (49.224)$$

À nouveau, si nous prenons:

$$\gamma = \frac{1}{\sqrt{1 - \beta^2}} \quad (49.225)$$

toujours avec $\beta = v/c$.

Nous avons dès lors en exprimant toutes les relations précédentes de transformation dans les mêmes unités en se souvenant que $E \equiv p \cdot c$:

$$\begin{aligned} p_x \cdot c &= \gamma(p_{x'} \cdot c + \beta \cdot E') \\ p_y \cdot c &= p_{y'} \cdot c \\ p_z \cdot c &= p_{z'} \cdot c \\ E &= \gamma(E' + \beta \cdot c \cdot p_{x'}) \end{aligned} \quad (49.226)$$

Nous pouvons alors définir une matrice telle que:

$$\begin{pmatrix} E \\ p_x c \\ p_y c \\ p_z c \end{pmatrix} = \begin{bmatrix} \gamma & \gamma\beta & 0 & 0 \\ \gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} E' \\ p'_x c \\ p'_y c \\ p'_z c \end{pmatrix} \quad (49.227)$$

où nous retrouvons la "matrice de Lorentz" ou "tenseur symétrique de Lorentz" $L^\mu_{\mu'}$.

Le vecteur:

$$p^\nu = \left(\frac{E}{c}, p_x, p_y, p_z \right) = \left(\frac{E}{c}, \vec{p} \right) = m\gamma(\vec{v})(c, \vec{v}) \quad (49.228)$$

est quant à lui, appelé le "**quadrivecteur d'énergie-impulsion**" ou plus simplement "**quadrivecteur impulsion**". Son utilité est que sa valeur est conservée, lors d'une réaction nucléaire. Si nous additionnons ces vecteurs sur toutes les particules (sans oublier les photons) avant et après la réaction, nous trouvons les mêmes sommes pour les 4 composantes.

Remarques:

R1. La transformation inverse étant effectuée bien évidemment avec la matrice inverse que nous avons déjà exposée plus haut.

R2. Nous utilisons en optique relativiste le quadrivecteur $(\omega/c, \vec{k})$, où ω est la pulsation de l'onde et $\vec{k}$ le vecteur d'onde (cf. chapitre de Mécanique Ondulatoire ou Optique Ondulatoire). Ce quadrivecteur est l'équivalent pour une onde électromagnétique du quadrivecteur $(E/c, \vec{p})$ pour une particule, multiplié par la constante de Planck $\hbar = h/2\pi$. En effet, la dualité onde-corpuscule (cf. chapitre de Physique Quantique Ondulatoire) attribue à une onde une énergie:

$$E = h\nu = \frac{h}{2\pi} \omega = \hbar \omega \quad (49.229)$$

et une quantité de mouvement dont la norme est:

$$p = \frac{E}{c} = \hbar \frac{\omega}{c} = \hbar k \quad (49.230)$$

Revenons sur le relation suivante qui est central dans certains domaines de la physique quantique:

$$p_x^2 + p_y^2 + p_z^2 - \frac{E^2}{c^2} = -m_0^2 \cdot c^2 \quad (49.231)$$

Soit:

$$\frac{E^2}{c^2} - (p_x^2 + p_y^2 + p_z^2) = m_0^2 c^2 \quad (49.232)$$

Ce qui peut s'écrire sous forme vectorielle (forme très courante):

$$\frac{E^2}{c^2} - \|\vec{p}\|^2 = m_0^2 c^2 \quad (49.233)$$

Cette dernière relation nous sera très utile dans le chapitre de Physique Quantique Relativiste pour calculer l'énergie de photons virtuels d'échange.

Pour les photons, puisque la masse est nulle, nous avons:

$$\frac{E^2}{c^2} - \|\vec{p}\|^2 = 0 \quad (49.234)$$

2.9.1. RELATION D'EINSTEIN

Suivant le principe de relativité, nous souhaitons que la relation entre la quantité de mouvement et l'énergie d'une onde électromagnétique s'écrive de la même manière pour deux observateurs d'inertie en translation l'un par rapport à l'autre:

Si O écrit:

$$E = c \sqrt{m_0^2 \cdot c^2 + p^2} \quad (49.235)$$

alors O' doit pouvoir écrire:

$$E' = c \sqrt{m_0^2 \cdot c^2 + p'^2} \quad (49.236)$$

Reprenons la première relation ci-dessus et mettons-la au carré sans oublier que le photon a une masse nulle $m_0 = 0$. Alors:

$$E^2 = c^4 m_0^2 + p^2 c^2 \quad (49.237)$$

et comme $m_0 = 0$:

$$E = pc \quad (49.238)$$

Étant donnée connue la relation de Planck (définie en thermodynamique):

$$E = h\nu$$

nous sommes amenés à écrire la fameuse "**relation d'Einstein**" que nous retrouverons très souvent en physique quantique ainsi qu'en thermodynamique:

$$p = \frac{h\nu}{c} \quad (49.239)$$

2.9.2. FORCE RELATIVISTE

Suivant le principe de la relativité, nous voulons que la relation entre la force et la quantité de mouvement s'écrive de la même manière par deux observateurs d'inertie en translation l'un par rapport à l'autre:

Ainsi, si O écrit:

$$\vec{F} = \frac{d\vec{p}}{dt} \quad (49.240)$$

O' doit pouvoir écrire:

$$\vec{F}' = \frac{d\vec{p}'}{dt'} \quad (49.241)$$

La relation entre $\vec{F}$ et $\vec{F}'$ est assez compliquée dans le cas général. Nous nous limiterons ici au cas particulier où un corps est momentanément immobile dans O' et où donc l'observateur O' ne tiendra compte que de la force $\vec{F}'$ qu'il applique. Il l'appellera par ailleurs "**force propre**", car il n'a pas à se préoccuper d'autres forces (comme une force centrifuge, par exemple).

Il faut substituer p' et t' par p et t dans:

$$F'_x = \frac{dp'_x}{dt'}, F'_y = \frac{dp'_y}{dt'}, F'_z = \frac{dp'_z}{dt'} \quad (49.242)$$

Puisque:

$$\frac{dt}{dt'} = \frac{1}{\sqrt{1 - v^2/c^2}} \quad (49.243)$$

nous aurons:

$$\begin{aligned} F'_x &= \frac{dp'_x}{dt'} = \frac{dt}{dt'} \frac{dp'_x}{dt} = \frac{dt}{dt'} \frac{d}{dt} \left(\frac{p_x - v \frac{E}{c^2}}{\sqrt{1 - v^2/c^2}} \right) = \frac{dt}{dt'} \frac{d}{dt} \left(\frac{1}{\sqrt{1 - v^2/c^2}} \left(p_x - v \frac{E}{c^2} \right) \right) \\ &= \frac{dt}{dt'} \left(\frac{1}{\sqrt{1 - v^2/c^2}} \right) \frac{d}{dt} \left(p_x - v \frac{E}{c^2} \right) = \left(\frac{1}{\sqrt{1 - v^2/c^2}} \right) \left(\frac{1}{\sqrt{1 - v^2/c^2}} \right) \frac{d}{dt} \left(p_x - v \frac{E}{c^2} \right) \quad (49.244) \\ &= \frac{1}{1 - v^2/c^2} \frac{d}{dt} \left(p_x - v \frac{E}{c^2} \right) = \frac{1}{1 - v^2/c^2} \left(\frac{dp_x}{dt} - \frac{v}{c^2} \frac{dE}{dt} \right) \end{aligned}$$

Nous avons par ailleurs vu que:

$$E = E_c + m_0 c^2 \Rightarrow \frac{dE}{dt} = \frac{dE_c}{dt} + 0 = \frac{F_x dx}{dt} = F_x \cdot v \quad (49.245)$$

Il reste donc:

$$F'_x = \frac{1}{1 - v^2/c^2} \cdot \left(F_x - \frac{v}{c^2} F_y v \right) = F_x \quad (49.246)$$

La composante de la force est donc invariante dans la direction du déplacement.

Pour les directions y, z perpendiculaires au déplacement:

$$F'_y = \frac{dp'_y}{dt'} = \frac{dt}{dt'} \cdot \frac{dp_y}{dt} = \frac{F_y}{\sqrt{1 - v^2/c^2}} \quad \text{et} \quad F'_z = \frac{dp'_z}{dt'} = \frac{dt}{dt'} \cdot \frac{dp_z}{dt} = \frac{F_z}{\sqrt{1 - v^2/c^2}} \quad (49.247)$$

En résumé:

$$\boxed{\begin{aligned} F'_x &= F_x \Leftrightarrow F_x = F'_x \\ F'_y &= \frac{F_y}{\sqrt{1 - v^2/c^2}} \Leftrightarrow F_y = F'_y \sqrt{1 - v^2/c^2} \\ F'_z &= \frac{F_z}{\sqrt{1 - v^2/c^2}} \Leftrightarrow F_z = F'_z \sqrt{1 - v^2/c^2} \end{aligned}} \quad (49.248)$$

Cependant, pour passer d'un référentiel à un autre, il vaut mieux utiliser le "**quadrivecteur force**" défini comme la dérivée du quadrivecteur impulsion par rapport au temps propre:

$$\vec{F}^\nu = \frac{dp^\nu}{dt'} = \frac{dt}{dt'} \frac{dp^\nu}{dt} = \gamma \frac{dp^\nu}{dt} = \gamma \frac{d}{dt} \left(\frac{E}{c}, \vec{p} \right) = \gamma \left(\frac{\vec{F} \circ \vec{v}}{c}, \vec{F} \right) \quad (49.249)$$

Effectivement, rappelons que:

$$\frac{dE}{dt} = \vec{F} \circ \vec{v} \quad \vec{F} = \frac{d\vec{p}}{dt} \quad (49.250)$$

2.10. ÉLECTRODYNAMIQUE RELATIVISTE

Avec un spectromètre de masse, nous établissons que le rapport m/q de la masse m d'une particule par sa charge électrique q varie de la même manière que la masse m lorsque la vitesse v de la particule varie:

$$\frac{m}{q} = \frac{m_0/q_0}{\sqrt{1 - v^2/c^2}} = \frac{1}{q_0} \frac{m_0}{\sqrt{1 - v^2/c^2}} = \frac{m}{q_0} \quad (49.251)$$

Ainsi, il vient que:

$$q = q_0 \quad (49.252)$$

La charge d'une particule est donc indépendante de sa vitesse comme nous l'avons démontré dans la section d'électromagnétisme (cf. chapitre d'Électrodynamique) lors de la détermination de l'équation de conservation de la charge.

Considérons maintenant deux charges q et Q immobiles dans le référentiel O' en translation à vitesse v par rapport à O :

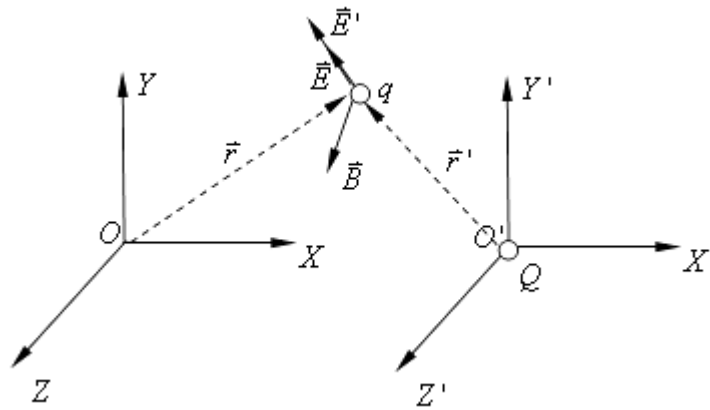


Figure: 49.6 - Configuration pour l'étude des transformations des champs électrique et magnétique

Nous allons nous restreindre au cas où la vitesse $\vec{v}$ est parallèle à l'axe OX :

$$\vec{v} = (v_x, 0, 0) \quad (49.253)$$

et nous notons le vecteur à l'horizontale pour économiser un peu de papier...

La charge Q est placée en O' et elle est donc immobile pour O' . L'observateur O' conclut qu'une force électrostatique:

$$\vec{F}' = q \cdot \vec{E}' \quad (49.254)$$

agit donc sur la charge témoin q placée en $\vec{r}'$.

$$F'_x = q \cdot E'_x, F'_y = q \cdot E'_y, F'_z = q \cdot E'_z \quad (49.255)$$

L'observateur O voit également un champ électrostatique $\vec{E}$ en $\vec{r}$, mais il voit aussi que Q est en mouvement selon l'axe OX . Il en déduit donc l'existence d'un champ magnétique $\vec{B}$ en $\vec{r}$ orienté dans le plan YZ :

$$\vec{B} = (0, B_y, B_z) \quad (49.256)$$

Il mesure donc la force (cf. chapitre de Magnétostatique) de Lorentz (supposée connue):

$$\vec{F} = q(\vec{E} + \vec{v} \times \vec{B}) \quad (49.257)$$

Mais:

$$\vec{v} \times \vec{B} = \begin{pmatrix} v_x \\ 0 \\ 0 \end{pmatrix} \times \begin{pmatrix} 0 \\ B_y \\ B_z \end{pmatrix} = \begin{pmatrix} 0 - 0 \\ 0 - v_x B_z \\ v_x B_y \end{pmatrix} = \begin{pmatrix} 0 \\ -v B_z \\ v B_y \end{pmatrix} \quad (49.258)$$

Donc:

$$\begin{aligned}
 F_x &= q \cdot E_x \\
 F_y &= q \cdot (E_y - vB_z) \\
 F_z &= q \cdot (E_z + vB_y)
 \end{aligned} \quad (49.259)$$

Nous avons vu maintenant:

$$\begin{aligned}
 F'_x &= F_x = q \cdot E_x \\
 F'_y &= \frac{F_y}{\sqrt{1-v^2/c^2}} = \frac{q \cdot (E_y - vB_z)}{\sqrt{1-v^2/c^2}} \\
 F'_z &= \frac{F_z}{\sqrt{1-v^2/c^2}} = \frac{q \cdot (E_z + vB_y)}{\sqrt{1-v^2/c^2}}
 \end{aligned} \quad (49.260)$$

La comparaison des expressions ci-dessus donne les transformations relativistes du champ électrique:

$$\begin{aligned}
 E'_x &= E_x \Leftrightarrow E_x = E'_x \\
 E'_y &= \frac{E_y - vB_z}{\sqrt{1-v^2/c^2}} \Leftrightarrow E_y = \frac{E'_y + vB'_z}{\sqrt{1-v^2/c^2}} \\
 E'_z &= \frac{E_z + vB_y}{\sqrt{1-v^2/c^2}} \Leftrightarrow E_z = \frac{E'_z - vB'_y}{\sqrt{1-v^2/c^2}}
 \end{aligned} \quad (49.261)$$

Comme pour la transformation de Lorentz des composantes spatiales et temporelles, nous avons obtenu les transformations inverses en échangeant les champs et en considérant que O' voit O reculer (nous remplaçons donc v par $-v$).

Pour obtenir les transformations relativistes du champ magnétique, nous procédons comme ci-dessous:

$$E'_y = \frac{E_y - vB_z}{\sqrt{1-v^2/c^2}} = \frac{\frac{E'_y + vB'_z}{\sqrt{1-v^2/c^2}} - vB_z}{\sqrt{1-v^2/c^2}} \quad (49.262)$$

Après quelques petites manipulations d'algèbre très élémentaire, nous obtenons:

$$B_z = \frac{B'_z + v/c^2 E'_y}{\sqrt{1-v^2/c^2}} \quad (49.263)$$

Nous faisons identiquement:

$$E_y = \frac{E'_y + vB'_z}{\sqrt{1-v^2/c^2}} = \frac{\frac{B'_z + v/c^2 E'_y}{\sqrt{1-v^2/c^2}} + vB_z}{\sqrt{1-v^2/c^2}} \quad (49.264)$$

Après encore une fois quelques petites manipulations d'algèbre très élémentaire, nous obtenons:

$$B'_z = \frac{B_z - v/c^2 E_y}{\sqrt{1 - v^2/c^2}} \quad (49.265)$$

et ainsi de suite. Nous obtenons finalement:

$$\begin{aligned} B'_x &= B_x \Leftrightarrow B_x = B'_x \\ B'_y &= \frac{B_y + v/c^2 E_z}{\sqrt{1 - v^2/c^2}} \Leftrightarrow B_y = \frac{B'_y - v/c^2 E'_z}{\sqrt{1 - v^2/c^2}} \\ B'_z &= \frac{B_z - v/c^2 E_y}{\sqrt{1 - v^2/c^2}} \Leftrightarrow B_z = \frac{B'_z + v/c^2 E'_y}{\sqrt{1 - v^2/c^2}} \end{aligned} \quad (49.266)$$

Étudions maintenant le comportement du champ électromagnétique d'une charge en mouvement:

Soient deux référentiels parallèles O et O' , en translation à vitesse constante v selon l'axe XX' :

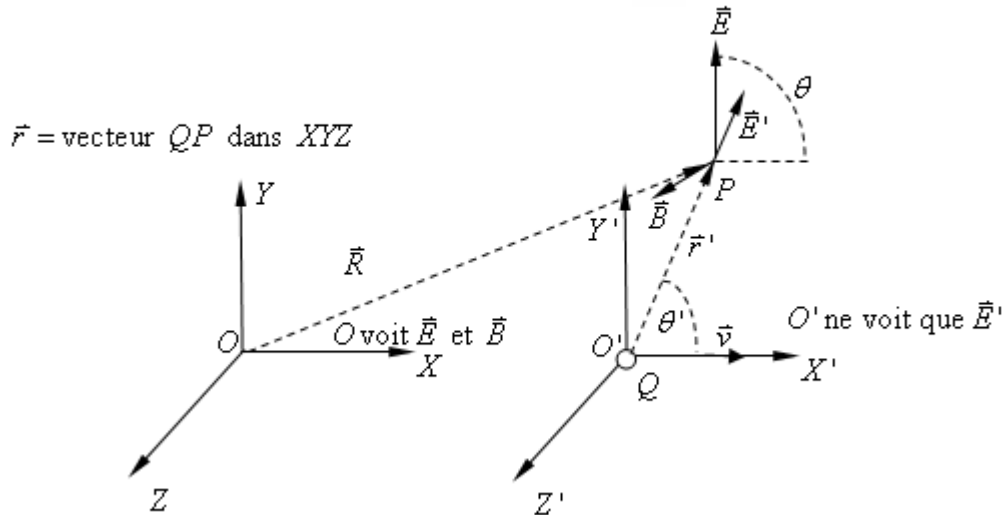


Figure: 49.7 - Configuration pour l'étude de transformations électrodynamiques

où une charge immobile Q est placée en O' .

Il est clair alors que l'observateur O mesure $\vec{B}' = 0$ partout et qu'au point P du plan $X'Y'$, en $\vec{r}' = (x', y', 0)$ il mesure le champ électrostatique (cf. chapitre Électrostatique):

$$\vec{E}' = (E'_x, E'_y, 0) = \frac{Q}{4\pi\epsilon_0 r'^3} \cdot \vec{r}' = \frac{Q}{4\pi\epsilon_0 r'^3} \cdot (x', y', 0) \quad (49.267)$$

Si l'observateur O est informé des valeurs de $\vec{E}'$ et de $\vec{B}' = 0$, il peut les introduire dans les transformées relativistes donnant le champ électrique $\vec{E}$ qu'il observe:

$$E_x = E'_x$$

$$E_y = \frac{E'_y + vB'_z}{\sqrt{1 - v^2/c^2}} = \frac{E'_y}{\sqrt{1 - v^2/c^2}}$$

$$E_z = \frac{E'_z - vB'_y}{\sqrt{1 - v^2/c^2}} = \frac{E'_z}{\sqrt{1 - v^2/c^2}}$$

(49.268)

Pour écrire une expression du champ $\vec{E}$ au point P , l'observateur O doit déterminer, à un instant t de son temps local, les composantes du vecteur $\vec{r} = (x, y, 0)$ qui sépare le point P de la charge Q (en sommant les vecteurs positions de ces deux derniers points matériels).

Les coordonnées du point P et de la charge Q qu'il voit dans le plan XYZ sont données par les transformations de Lorentz habituelles:

$$\vec{R}_P = \left(\frac{x' + vt}{\sqrt{1 - v^2/c^2}}, y', 0 \right) \text{ et } \vec{R}_Q = \left(\frac{vt}{\sqrt{1 - v^2/c^2}}, 0, 0 \right) \quad (49.269)$$

Il en déduit donc facilement, par sommation les distances x, y .

Une autre méthode, plus simple, est qu'étant donné que la composante x est une longueur, elle subit donc les transformations de Lorentz et:

$$\vec{r} = (x, y, z) = \left(x' \sqrt{1 - v^2/c^2}, y', 0 \right) \quad (49.270)$$

Car rappelons-le:

$$x' = \frac{x}{\sqrt{1 - v^2/c^2}} \text{ et } y' = y \quad (49.271)$$

La transformée relativiste du champ électrique donne alors:

$$E_x = E'_x = \frac{Q}{4\pi\epsilon_0 r'^3} x' = \frac{Q}{4\pi\epsilon_0 r'^3} \frac{x}{\sqrt{1 - v^2/c^2}}$$

$$E_y = \frac{E'_y}{\sqrt{1 - v^2/c^2}} = \frac{1}{\sqrt{1 - v^2/c^2}} \frac{Q}{4\pi\epsilon_0 r'^3} y' = \frac{Q}{4\pi\epsilon_0 r'^3} \frac{y}{\sqrt{1 - v^2/c^2}}$$

(49.272)

et:

$$E_z = E'_z = 0 \quad (49.273)$$

Écrit sous forme vectorielle:

$$\vec{E} = \frac{Q \cdot \vec{r}}{4\pi\epsilon_0 \sqrt{1 - v^2/c^2} r'^3} \quad (49.274)$$

Il nous faut encore déterminer comment exprimer r' en fonction de r :

$$r'^2 = x'^2 + y'^2 = \frac{x^2}{1 - v^2/c^2} + y^2 = \frac{x^2 + (1 - v^2/c^2)y^2}{1 - v^2/c^2} = \frac{r^2 - (v^2/c^2)y^2}{1 - v^2/c^2} \quad (49.275)$$

car (théorème de Pythagore):

$$x^2 + y^2 = r^2 \quad (49.276)$$

L'écriture se simplifie si nous utilisons l'angle formé par le vecteur champ électrique et l'axe OX. Nous notons alors θ' dans O' et θ dans O les angles donnés par:

$$\tan \theta' = y'/x' \text{ et } \tan \theta = y/x \quad (49.277)$$

avec $\theta \geq \theta'$ à cause de la dilatation des longueurs selon l'axe OX.

Nous éliminons y avec:

$$y^2 = r^2 \sin^2 \theta \Rightarrow r' = \frac{r[1 - (v^2/c^2) \sin^2 \theta]^{1/2}}{\sqrt{1 - v^2/c^2}} \quad (49.278)$$

Ainsi, le champ électrique $\vec{E}$ que voit O est donné par:

$$\vec{E} = \frac{Q \cdot \vec{r}}{4\pi\epsilon_0 \sqrt{1 - v^2/c^2}} \cdot \frac{1}{r'^3} = \frac{(1 - v^2/c^2)}{[1 - (v^2/c^2) \sin^2 \theta]^{3/2}} \frac{Q \cdot \vec{r}}{4\pi\epsilon_0 r^3} \quad (49.279)$$

Le facteur contenant $\sin \theta$ montre que le champ électrique $\vec{E}$ d'une charge en mouvement n'a plus la symétrie sphérique. Il dépend de la direction du vecteur $\vec{r}$.

À distance égales, le champ électrique est plus intense dans la direction verticale à celle du déplacement ($\theta = \pi/2$) que dans la direction du déplacement de la charge ($\theta = 0$).

Si $v=0$, nous retrouvons par ailleurs l'expression classique et connue:

$$\vec{E} = \frac{Q \cdot \vec{r}}{4\pi\epsilon_0 r^3} \quad (49.280)$$

Remarque: Rappelons que nous avons effectué (et continuons dans ce sens) ici une étude d'une charge en mouvement rectiligne uniforme, c'est-à-dire à vitesse constante.

Pour trouver maintenant l'expression du champ magnétique $\vec{B}$, nous introduisons:

$$\begin{aligned} \vec{B} &= (B'_x, B'_y, B'_z) = (0, 0, 0) \text{ et } E'_x = E_x \\ E'_y &= E_y \sqrt{1 - v^2/c^2} \text{ et } E'_z = E_z \sqrt{1 - v^2/c^2} \end{aligned} \quad (49.281)$$

dans:

$$B_x = B'_x, B_y = \frac{B'_y - vE'_z/c^2}{\sqrt{1-v^2/c^2}}, B_z = \frac{B'_z + vE'_y/c^2}{\sqrt{1-v^2/c^2}} \quad (49.282)$$

Nous obtenons dès lors:

$$B_x = 0, B_y = -\frac{vE_z}{c^2}, B_z = \frac{vE_y}{c^2} \quad (49.283)$$

qui sont les composantes de:

$$\vec{B} = \vec{v} \times \frac{\vec{E}}{c^2} \quad (49.284)$$

Pour connaître $\vec{B}$ en fonction de $\vec{r}$, nous substituons l'expression obtenue pour $\vec{E}$:

$$\vec{B} = \frac{Q}{c^2 \epsilon_0 4\pi r^3} \frac{(1-v^2/c^2)}{[1-(v^2/c^2)\sin^2\theta]^{3/2}} \cdot \vec{v} \times \vec{r} \quad (49.285)$$

Dans le cas où la vitesse est faible, le terme relativiste tend vers 1 et le champ $\vec{B}$ d'une charge Q se déplaçant à la vitesse v devient:

$$\vec{B} = \frac{Q}{c^2 \epsilon_0 4\pi r^3} \cdot \vec{v} \times \vec{r} = \frac{\mu_0 Q}{4\pi r^3} \cdot \vec{v} \times \vec{r} \quad (49.286)$$

car comme nous l'avons dans le chapitre d'Électrodynamique: $\mu_0 = 1/(c^2 \epsilon_0)$

Remarques:

R1. En chaque endroit, les lignes du champ $\vec{B}$ sont contenues dans un plan perpendiculaire à la direction de déplacement de la charge Q (produit vectoriel oblige).

R2. Si la charge en mouvement est vue comme un dQ attaché au point O' , nous pouvons interpréter son déplacement à vitesse v comme un courant I en un point du référentiel O où se trouve O' . Ainsi:

$$\begin{aligned} dQ \cdot \vec{v} &= dQ \cdot \frac{d\vec{x}}{dt} = \frac{dQ}{dt} \cdot d\vec{x} = I \cdot d\vec{x} \\ \Rightarrow d\vec{B}(\vec{r}) &= \frac{\mu_0 dQ}{4\pi r^3} \cdot \vec{v} \times \vec{r} = \frac{\mu_0 \cdot I}{4\pi r^3} \cdot d\vec{x} \times \vec{r} \end{aligned} \quad (49.287)$$

Cette dernière relation est connue sous le nom de la "**loi de Biot et Savart**" et nous la retrouverons au début de la section traitant de l'électromagnétisme. Cet état de fait valide encore le modèle relativiste.

Il est intéressant de se rappeler qu'une particule chargée en mouvement sera vue dans le référentiel de la particule comme n'émettant aucun champ électromagnétique (il y aura juste un champ électrostatique). Ce qui n'est pas le cas pour un référentiel au repos. Il y a donc ici une sorte de contradiction contre-intuitive flagrante.

Mais cela pose alors un autre problème, dans un référentiel en mouvement accéléré, une particule chargée émet normalement un rayonnement d'accélération, ce rayonnement en mécanique quantique doit s'accompagner forcément de l'émission d'un quanta, qui lui existe ou n'existe pas (un moyen terme n'existe pas). L'existence même des photons serait donc purement relative. Et pourtant c'est le cas! Certaines particules n'ont qu'une existence relative. La réponse compliquée, c'est donc de savoir ce que sont devenus les photons.

Mais là, nous touchons à la limite de ce que nous maîtrisons parfaitement dans la physique de la fin du 20ème siècle, car nous parlons de référentiels accélérés (ce qui implique d'être en relativité générale et non restreinte) et de théorie quantique des champs. Le cadre rigoureux pour traiter ça (qui engloberait une gravitation quantique) n'existe pas encore. Mais un premier pas a été franchi avec le développement de la théorie quantique des champs en espace courbe.

2.10.1. TRANSFORMATION DU TENSEUR DE CHAMP

Nous avons vu et démontré dans le chapitre d'Électrodynamique que l'ensemble du champ électromagnétique se résumait au tenseur du même nom. Il serait alors bon de regarder comment se transforme ce tenseur et s'il le fait correctement relativement aux résultats obtenus plus haut.

Considérons la transformation (où le tenseur du champ électromagnétique est en unités naturelles!!!):

$$F'^{\mu\nu} = L^\mu_\alpha L^\nu_\beta F^{\alpha\beta} \quad (49.288)$$

avec le tenseur du champ électromagnétique en composantes contravariantes dans la métrique de Minkowski + - - - :

$$F^{\mu\nu} = \begin{bmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & -B_z & B_y \\ E_y & B_z & 0 & -B_x \\ E_z & -B_y & B_x & 0 \end{bmatrix} \quad (49.289)$$

et aussi par construction:

$$F'^{\mu\nu} = \begin{bmatrix} 0 & -E'_x & -E'_y & -E'_z \\ E'_x & 0 & -B'_z & B'_y \\ E'_y & B'_z & 0 & -B'_x \\ E'_z & -B'_y & B'_x & 0 \end{bmatrix} \quad (49.290)$$

Prenons, par exemple, la vitesse parallèle à l'axe OX, alors nous avons démontré plus haut que:

$$L = \begin{bmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (49.291)$$

Soit, donc:

$$L_0^0 = \gamma, L_1^0 = -\beta\gamma, L_0^1 = -\beta\gamma, L_1^1 = \gamma \quad (49.292)$$

où comme nous pouvons le voir, il est souvent d'usage dans le domaine de la relativité restreinte et de l'électrodynamique de numéroté les composantes des matrices/tenseurs à partir de 0 (au lieu de 1 pour tous les autres chapitres du site).

Nous calculons les transformées (se rappeler que le tenseur du champ électromagnétique est antisymétrique!):

$$\begin{aligned} F'^{01} &= L_\alpha^0 L_\beta^1 F^{\alpha\beta} = L_0^0 L_0^1 F^{00} + L_1^0 L_0^1 F^{10} + L_0^0 L_1^1 F^{01} + L_1^0 L_1^1 F^{11} = L_1^0 L_0^1 F^{10} + L_0^0 L_1^1 F^{01} \\ &= \beta^2 \gamma^2 F^{10} + \gamma^2 F^{01} = F^{01} \gamma^2 (1 - \beta^2) = F^{01} \end{aligned} \quad (49.293)$$

Nous en déduisons donc, pour le champ électrique (ce qui correspond parfaitement à ce que nous avons obtenu plus haut):

$$E'_x = E_x \quad (49.294)$$

Nous faisons un second calcul pour la composante perpendiculaire:

$$F'^{02} = L_\alpha^0 L_\beta^2 F^{\alpha\beta} = L_\alpha^0 F^{\alpha 2} = L_0^0 F^{02} + L_1^0 F^{12} \quad (49.295)$$

d'où:

$$-E'_y = \gamma(-E_y + \beta B_z) \Rightarrow E'_y = \gamma(E_y - \beta B_z) \quad (49.296)$$

ce qui correspond à nouveau parfaitement à ce que nous avons obtenu plus haut (en unités naturelles, ne pas oublier que nous avons alors $\beta = v$) !

La vérification se fait de même pour le champ magnétique:

$$F'^{23} = L_\alpha^2 L_\beta^3 F^{\alpha\beta} = F^{23} \Rightarrow B'_x = B_x \quad (49.297)$$

et:

$$F'^{12} = L_\alpha^1 L_\beta^2 F^{\alpha\beta} = L_\alpha^1 F^{\alpha 2} = L_1^1 F^{12} + L_0^1 F^{02} = \gamma F^{12} - \gamma\beta F^{02} \quad (49.298)$$

ce qui donne (en unités naturelles, ne pas oublier que nous avons alors $\beta = v$) ! :

$$-B'_z = \gamma(-B_z + \beta E_y) \Rightarrow B'_z = \gamma(B_z - \beta E_y) \quad (49.299)$$

etc.

3. ESPACE-TEMPS DE MINKOWSKI

Nous avons démontré plus haut que:

$$t = \frac{t'}{\sqrt{1 - v^2/c^2}} \quad (49.300)$$

Écrivons cela sous la forme:

$$\frac{t'^2}{t^2} = 1 - \frac{v^2}{c^2} \quad (49.301)$$

Multiplions les deux membres par $(ct)^2$:

$$\frac{t'^2 \cdot c^2 \cdot t^2}{t^2} = c^2 \cdot t^2 - \frac{v^2 \cdot c^2 \cdot t^2}{c^2} \quad (49.302)$$

ce qui nous donne:

$$t'^2 \cdot c^2 = c^2 \cdot t^2 - v^2 \cdot t^2 \quad (49.303)$$

Si $v = c$, l'équation s'annule:

$$t'^2 \cdot c^2 = 0 \quad (49.304)$$

Ce résultat traduit, que les dimensions d'espace et de temps sont comme arrêtées dans le référentiel relativiste, car la vitesse relative de l'objet est égale à celle de la lumière!

Imaginons maintenant qu'un faisceau lumineux soit émis à l'instant $t = 0$ et se propage depuis l'origine du référentiel. Nous savons que dans l'espace-temps (application du théorème de Pythagore dans l'espace euclidien à trois dimensions) la distance parcourue par le photon lumineux est:

$$c = \frac{\sqrt{x^2 + y^2 + z^2}}{t} \quad (49.305)$$

En changeant t de membre et en portant le tout au carré pour supprimer la racine, nous obtenons:

$$ct = \sqrt{x^2 + y^2 + z^2} \Leftrightarrow ct^2 = x^2 + y^2 + z^2$$

$$ct^2 - (x^2 + y^2 + z^2) = 0 \quad (49.306)$$

Remarque: Nous pouvons assimiler cette équation à la représentation d'un front d'onde sphérique d'une onde lumineuse se propageant à la vitesse de la lumière (voir l'équation d'une sphère centrée à l'origine dans le chapitre de Géométrie Analytique).

Considérons maintenant deux événements de coordonnées (x_1, y_1, z_1, t_1) et (x_2, y_2, z_2, t_2) et pour éviter la confusion changeons de lettre $s := x$. Nous pouvons dès lors écrire l'intervalle spatio-temporel tel quel:

$$(ds)^2 = c^2 \cdot (t_2 - t_1)^2 - \left((x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2 \right) \quad (49.307)$$

En passant à la limite, nous obtenons la forme quadratique:

$$ds^2 = c^2 dt^2 - (dx^2 + dy^2 + dz^2) = c^2 dt^2 - dL^2 \quad (49.308)$$

qui a la même forme et même valeur quel que soit le référentiel considéré. L'intervalle infinitésimal d'espace-temps ds^2 entre deux événements infiniment voisins est donc un invariant relativiste que nous appelons souvent "**abscisse curviligne d'espace-temps**"; c'est l'intervalle d'espace-temps ou, comme le disait simplement Einstein, le "carré de la distance".... Le fait que cette grandeur puisse être positive, négative (!) ou nulle est lié au caractère absolu de la vitesse de la lumière (nous y reviendrons juste après).

Nous pouvons aussi maintenant nous intéresser au caractère relativiste de cette métrique. Si elle est invariante, c'est qu'elle doit aussi l'être par les transformations de Lorentz. Nous disons alors que "**la métrique est invariante par transformation de Lorentz**". Une telle transformation peut être trouvée en s'inspirant de celle utilisée pour le tenseur du champ électromagnétique (voir plus haut). Le lecteur vérifiera sans peine en s'inspirant de l'exemple détaillé du champ électromagnétique que pour le tenseur métrique, nous avons la relation:

$$\eta_{\mu\nu}' = L_{\alpha}^{\mu} L_{\beta}^{\nu} \eta_{\alpha\beta} \quad (49.309)$$

L'abscisse curviligne peut s'exprimer aussi par la norme du quadrivecteur déplacement que nous avons défini plus haut comme étant $(c \cdot t, x, y, z)$. Effectivement, la norme (cf. chapitre de Calcul Tensoriel) s'écrit en descendant les indices à l'aide de la "**métrique de Minkowski**" $\eta_{\mu\nu}$ ou "**métrique pseudo-riemannienne**":

$$ds^2 = dx^{\nu} (\eta_{\mu\nu} dx^{\mu}) = dx^{\nu} dx_{\nu} = c^2 dt^2 - dx^2 - dy^2 - dz^2 \quad (49.310)$$

avec par définition (nous reviendrons là-dessus dans les détails au début de notre étude de la relativité générale) la "**matrice de Minkowski**":

$$\eta_{\mu\nu} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \quad (49.311)$$

où comme d'habitude sur ce site nous faisons l'abus de notation (déjà mentionné dans le chapitre de Calcul Tensoriel) de ne pas mettre $\eta_{\mu\nu}$ entre crochets (puisque un tenseur et sa forme matricielle sont normalement deux choses distinctes en toute rigueur).

Si nous mettons les deux relations suivantes en correspondance:

$$c^2 \cdot t^2 - (x^2 + y^2 + z^2) = 0 \text{ et } ds^2 = c^2 \cdot dt^2 - (dx^2 + dy^2 + dz^2) \quad (49.312)$$

nous avons alors $ds^2 = 0$ lorsque que les deux événements sont reliés à la vitesse de la lumière.

De plus, si nous posons $dx^2 + dy^2 + dz^2 = dr^2$ nous pouvons alors écrire:

$$ds^2 + dr^2 = c^2 dt^2 \quad (49.313)$$

Ceci n'est rien d'autre que l'équation d'un cône (cf. chapitre de Géométrie Analytique) d'axe d'ordonnée $c^2 dt^2 \dots$ le fameux "**cône d'Univers**" (auquel nous consacrons une étude plus loin). Tout événement est donc par extension dans ce cône et l'évolution de tout système peut donc y être décrit (par sa position spatiale et temporelle), par ce que nous appelons sa "**ligne d'Univers**". La ligne d'Univers d'une particule est donc la séquence d'événements qu'elle déroule durant sa vie.

3.1. QUADRI Vecteurs

Nous venons de définir ce qu'était la métrique de Minkowski, nous pouvons maintenant définir correctement le concept de quadri vecteur que nous avons déjà abordé sans toutefois toujours savoir ce que l'on faisait.

Définition: Dans un espace à quatre dimensions de type Minkowski, quatre grandeurs (K_x, K_y, K_z, K_t) (peu importe l'ordre des termes pour cette définition ou que les indices soient des chiffres ou des lettres correspondant aux quatre composantes spatio-temporelles) forment un quadri vecteur covariant si elles se transforment suivant la transformation de Lorentz:

$$\begin{aligned} K'_x &= \gamma(K_x - \beta K_t) & K_x &= \gamma(K'_x + \beta K'_t) \\ K'_y &= K_y & K_y &= K'_y \\ K'_z &= K_z & K_z &= K'_z \\ K'_t &= \gamma(K_t - \beta K_x) & K_t &= \gamma(K'_t + \beta K'_x) \end{aligned} \quad (49.314)$$

La pseudo-norme d'un quadri vecteur dans un espace de Minkowski à métrique $\eta_{\mu\nu}$ est alors:

$$K^\mu K_\mu = K^\mu (\eta_{\mu\nu} K^\nu) = K_t^2 - K_x^2 - K_y^2 - K_z^2 \quad (49.315)$$

où nous voyons que le quadri vecteur contravariant multiplié par la métrique redonne le quadri vecteur covariant.

La quantité suivante étant invariante par changement de référentiel Galiléen comme nous l'avons vu presque tout au début de ce chapitre:

$$K_\mu K^\mu = K'_\mu K'^\mu \quad (49.316)$$

Cette propriété d'invariance par changement de référentiel Galiléen des quadri vecteurs est leur propriété principale. Ainsi, deux observateurs en mouvement relatif uniforme l'un par rapport à l'autre doivent pour comparer les résultats d'une même mesure utiliser la norme des quadri vecteurs. De même, les lois qu'ils cherchent à déterminer pour être les plus générales possible doivent utiliser ces quantités invariantes!

Nous pouvons par ailleurs aussi écrire la norme sous la forme:

$$\begin{aligned} K_\mu K^\mu &= K_0 K^0 + K_1 K^1 + K_2 K^2 + K_3 K^3 \\ &= (K_0)^2 - (K_1)^2 - (K_2)^2 - (K_3)^2 \\ &= (K^0)^2 - (K^1)^2 - (K^2)^2 - (K^3)^2 \\ &= (K^0)^2 - \|\vec{K}\|^2 \end{aligned} \quad (49.317)$$

et les quadrivecteurs sous la forme:

$$K^\mu = (K^0, \vec{K}) \quad K_\mu = (K_0, -\vec{K}) \quad (49.318)$$

3.2. CÔNE D'UNIVERS

La topologie du cône de lumière trouve son origine dans les relations d'antériorité et postériorité des événements relativistes, ce qui permet de faire la distinction entre un événement dans le passé d'un autre ou dans le futur de celui-ci.

Les cônes de lumière ont pour objectif principal dans les ouvrages de vulgarisation de la physique théorique de schématiser l'histoire d'impulsions lumineuses émises en un point de l'espace où peuvent régner certaines conditions. Les points sont représentés dans l'espace par une série d'instantanés à divers instants t_1, t_2, t_3 , etc. (voir figure ci-dessous), le front d'onde sphérique de la lumière grossissant dans l'espace. Dans l'espace-temps, le même événement (en bas sur la figure) est représenté par un "**cône de lumière**", dont le sommet est le point d'émission.

Sur une feuille de papier, nous devons supprimer l'une des dimensions spatiales. Les axes spatiaux sont dessinés dans le plan horizontal et l'axe temporel dirigé vers le haut. Les sections du cône aux instants t_1, t_2, t_3 correspondent aux instantanés de la représentation spatiale: les fronts d'ondes à deux dimensions sont des cercles dont le rayon est celui du front d'onde sphérique à l'instant considéré. Le cône de lumière montre en un seul diagramme l'histoire continue du front d'onde d'un signal lumineux.

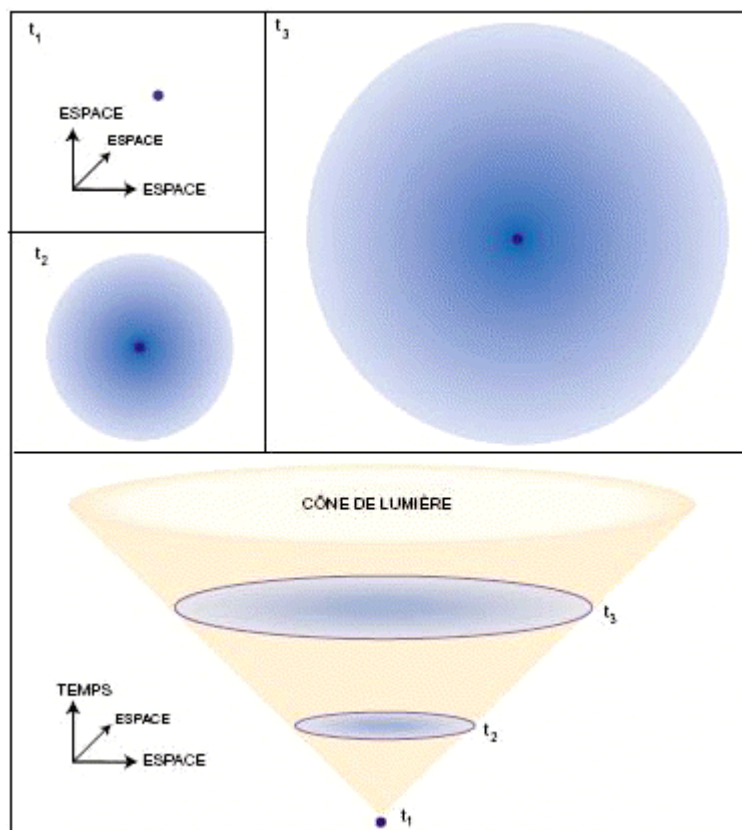


Figure: 49.8 - Principe du cône de lumière

Plus rigoureusement, les "instantanés" dont il a été fait mention plus haut sont appelés des "**événements ponctuels**" et ceux-ci apparaissent instantanés (approximation reposée sur l'optique géométrique) à tout observateur capable de les voir. Une collision entre deux particules ponctuelles fournit un exemple d'événement ponctuel. Il est tout à fait possible qu'un événement instantané non ponctuel apparaisse instantané à un certain observateur mais, à cause de la vitesse de propagation finie de la lumière, non instantané à un autre observateur.

Définitions:

D1. Nous disons par définition que deux événements ponctuels occupent le même point d'espace-temps s'ils apparaissent simultanés à tout observateur capable de les voir.

D2. L'ensemble M de tous les points de l'espace-temps est appelé "**l'espace-temps**".

D3. La frontière définie par le cône d'Univers est appelée "**horizon cosmologique**".

Rappelons que si aucune force n'agit sur une particule ponctuelle, nous la qualifions "d'inertielle" ou de "libre". Nous disons également qu'elle est en "**mouvement inertiel**".

Étant donné le point p , $N(p)$ est une structure géométrique absolue, indépendante de l'observateur. Sa composante future sera notée $N^+(p)$; sa composante passée $N^-(p)$ et elle sera représentée par le cône suivant:

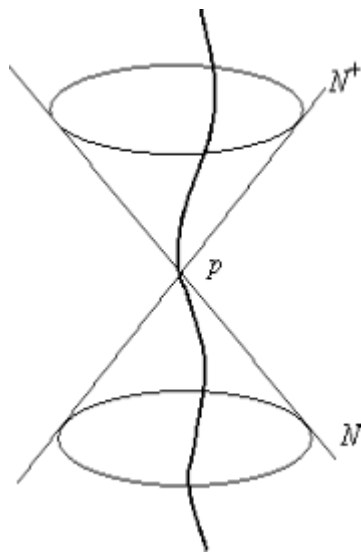


Figure: 49.9 - Cône de lumière passé et futur

Effectivement, rappelons que l'équation de Minkowski est invariante puisque:

$$cdt^2 - (dx^2 + dy^2 + dz^2) = cdt^2 - (dx^2 + dy^2 + dz^2) \quad (49.319)$$

Rapporté à trois paramètres (nous enlevons une dimension spatiale) nous avons, si les événements ponctuels sont reliés à la vitesse de la lumière (voir plus haut):

$$cdt^2 - (dx^2 + dy^2) = 0 \quad (49.320)$$

Ce que nous pouvons aussi écrire sous la forme:

$$cdt^2 = dx^2 + dy^2 \quad (49.321)$$

à comparer avec l'équation d'un cône (cf. chapitre de Géométrie Analytique):

$$(z \tan \theta)^2 = x^2 + y^2 \quad (49.322)$$

lorsque nous posons $c=1$ (ce qui est fréquent en physique théorique comme nous en avons déjà fait mention de nombreuses fois).

Donc l'équation de Minkowski peut donc bien être représentée par un cône.

Remarque: Si nous gardions les trois paramètres spatiaux et l'intervalle de temps constant, le lecteur remarquera certainement que nous tomberions non plus sur l'équation d'un cône mais sur celle d'une sphère. Il s'agit de la "**sphère céleste**" où à un instant donné, à sa surface, se créent de multiples cônes de lumière.

La ligne d'Univers de tout observateur qui occupe instantanément p et dont la ligne d'Univers passe donc par p lui-même, est contenue à l'intérieur de $N(p)$ défini par un point unique sur sa sphère céleste (celle qui est donc décrite par le vecteur d'information - qu'est le photon - dans toutes les directions de l'espace). Cela veut dire qu'il peut y avoir, in extenso, autant de rayons nuls (foyers des cônes) passant par p que de points sur une sphère.

L'exemple suivant paraîtra plus évident:

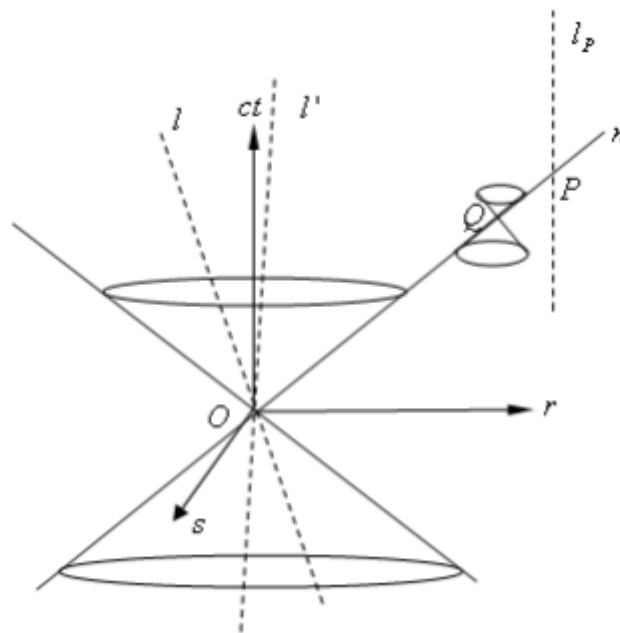


Figure: 49.10 - Principe de ligne d'Univers avec cône associé

Comme illustré sur la figure ci-dessus, un événement lumineux au point O de l'espace-temps produit un faisceau de photons, tous dans le cône nul du futur O , $N^+(O)$ (ces photons ont été émis par des atomes dans des états de mouvement variés, dont les lignes d'univers l et l' passent par O , mais sont entièrement contenues à l'intérieur de $N^+(O)$). La ligne d'univers n peut seulement être décrite par une particule se mouvant à la vitesse de la lumière car elle définit la frontière du cône (nous disons alors que la ligne d'Univers est "du genre lumière").

Remarque: La représentation des lignes d'Univers dans la partie inférieure (cône renversé) vient du fait qu'un événement peut également avoir un passé... donc le schéma généralise l'exemple particulier.

Soit l_P la ligne d'univers d'un personnage P immobile (d'où la verticalité de sa ligne d'Univers sur la figure ci-dessus) et n celle d'un rayon lumineux ayant pour origine O . Tous deux résident dans l'espace à quatre dimensions et ils se coupent selon un point unique P . Les points O et P se situent sur un rayon nul (d'un cône du futur), n , de $N^+(O)$. En P , le personnage P voit un flash soudain dans la direction définie par n , pour lui la direction de l'événement lumineux (décrite uniquement par sa vitesse donc, ainsi une ligne d'univers d'une particule inertielle peut être décrite uniquement par le temps et sa vitesse).

Un atome dont la ligne d'Univers coupe n au point Q , absorbe un photon de l'événement lumineux O et réémet peu de temps après un faisceau de photons. Ceux-ci forment alors à leur tour des rayons nuls dans $N^+(Q)$, mais seuls ceux de direction n atteindront le personnage P et seront vus par lui au point P .

Si P se trouve à l'intérieur de $N(O)$, le cône nul de O , nous dirons que sa ligne d'Univers est de "genre temps". Dans ce cas, O et P sont situés sur la ligne d'Univers d'un observateur ou d'une particule massive. Il existe bien évidemment deux types de déplacements de genre temps:

1. Si P est dans le futur de O (selon un observateur dont la ligne d'univers passe par O et P), nous dirons que P "pointe vers le futur".
2. Dans le cas contraire, nous dirons bien entendu qu'il "pointe vers le passé".

Si P se situe sur $N(O)$ - soit à la surface du cône - nous dirons alors qu'il est "nul" ou de "genre lumière" et si P n'est ni nul ni de genre temps, alors P se situe à l'extérieur de $N(O)$. Nous disons alors qu'il est de "genre espace":

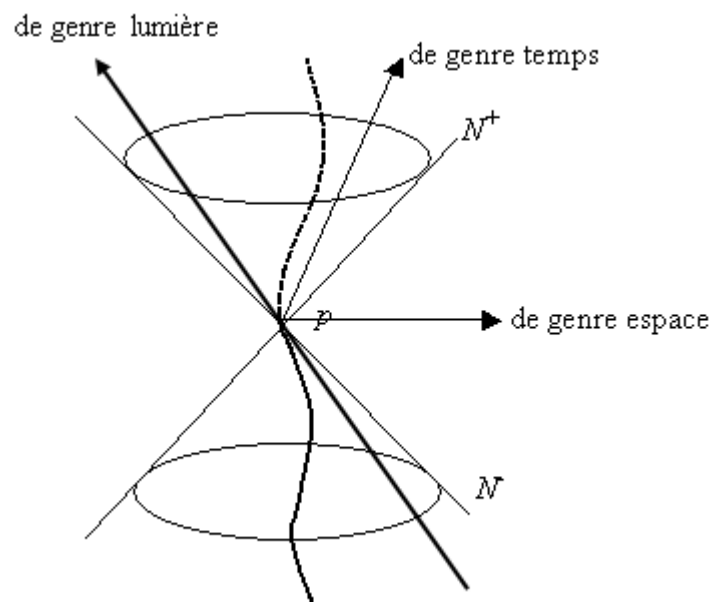


Figure: 49.11 - Types de lignes d'Univers

Cela se traduit mathématiquement par en se rappelant (voir plus haut) que:

$$ds^2 = dx^\nu (\eta_{\mu\nu} dx^\mu) = dx^\nu dx_\nu = c^2 dt^2 - dx^2 - dy^2 - dz^2 \quad (49.323)$$

D1. $ds^2 = 0$: la ligne d'Univers est donc de "**type lumière**" et c'est elle qui décrit la surface du cône par définition (selon ce que nous avons démontré précédemment et quel que soit le choix de la métrique) soit telle que:

$$c^2 dt^2 = dx^2 + dy^2 + dz^2 \quad (49.324)$$

ce qui est le cas d'un photon de lumière (d'où le nom...).

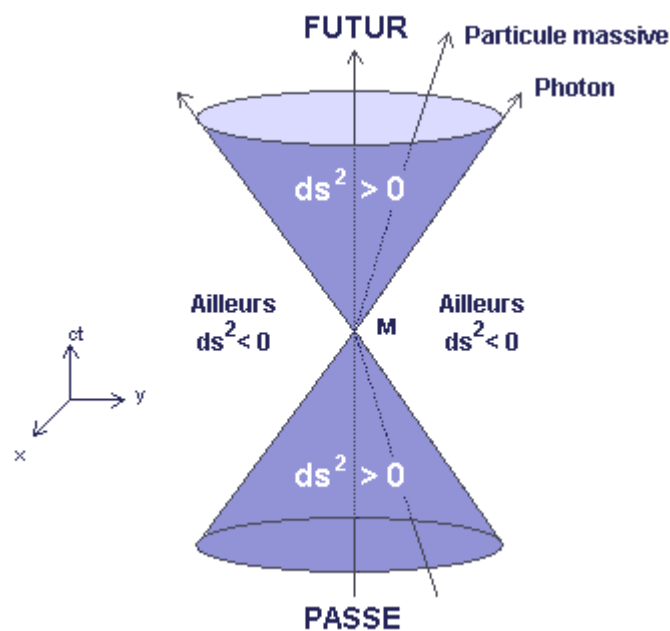
D2. $ds^2 < 0$: nous disons alors que la ligne d'Univers est de "**type espace**" soit telle que:

$$c^2 dt^2 < dx^2 + dy^2 + dz^2 \quad (49.325)$$

Deux événements qui ont lieu simultanément mais à des lieux différents sont donc de type espace.

D3. $ds^2 > 0$: nous disons alors que l'intervalle ou la ligne d'Univers sont de "**type temps**" soit telle que:

$$c^2 dt^2 > dx^2 + dy^2 + dz^2 \quad (49.326)$$



D4. Une "**ligne causale**" est une ligne de genre lumière ou temps qui est toujours orientée vers le futur.

Revenons à nos équations après ce petit interlude... les équations conduisent donc à faire plusieurs observations. Ainsi, dans l'Univers euclidien à quatre dimensions de Minkowski, les trajectoires des objets dans l'espace-temps sont toujours des droites. Effectivement, l'exemple trivial consiste à considérer que l'objet reste au repos, seul le temps continue alors de s'écouler. Nous avons dès lors:

$$t'^2 \cdot c^2 = c^2 \cdot t^2 - v^2 \cdot t^2 \quad (49.327)$$

en posant $v = 0$, cela nous donne:

$$\begin{aligned} c^2 \cdot t'^2 &= c^2 \cdot t^2 \\ c^2 \cdot dt'^2 &= c^2 \cdot dt^2 \end{aligned} \quad (49.328)$$

donc:

$$ds^2 = c^2 \cdot dt^2 \quad (49.329)$$

et aussi:

$$ds = c \cdot dt' \quad (49.330)$$

La primitive étant (constante d'intégration nulle):

$$s = c \cdot t' \quad (49.331)$$

qui est bien une droite et représente donc la ligne d'Univers de l'objet considéré dans le cône d'Univers. Nous pouvons aussi observer aussi que dans ce cas, l'évolution du phénomène est purement temporelle quand l'intervalle est positif (ce qui appuie ce que nous avons dit tout à l'heure).

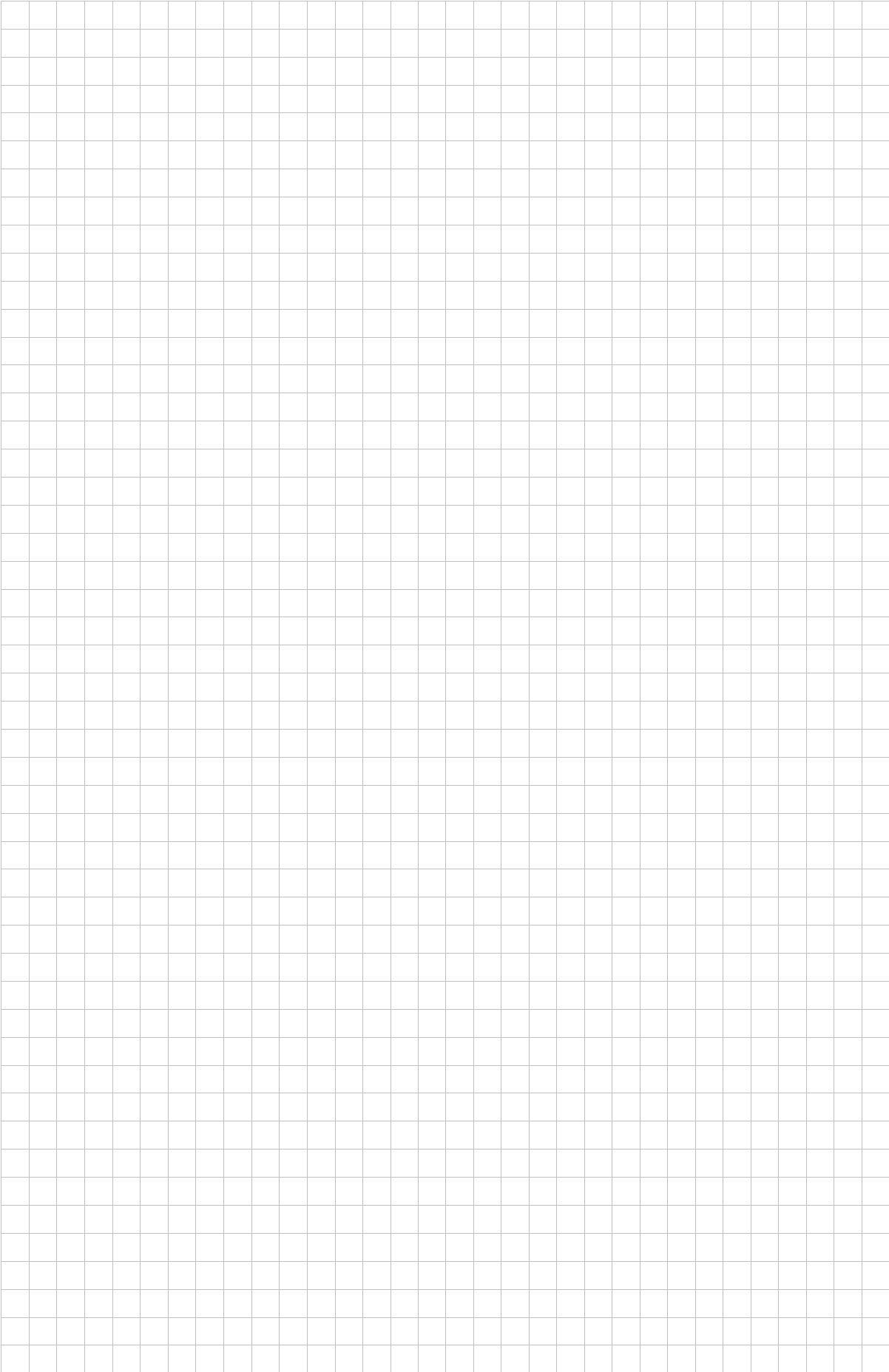
Remarques:

R1. Si la vitesse de la lumière est infinie, nous retrouvons le cas particulier de l'univers newtonien, où un phénomène peut instantanément se produire. Le temps y est absolu et il n'existe pas d'horizon cosmologique car le cône a une ouverture maximale (angle droit).

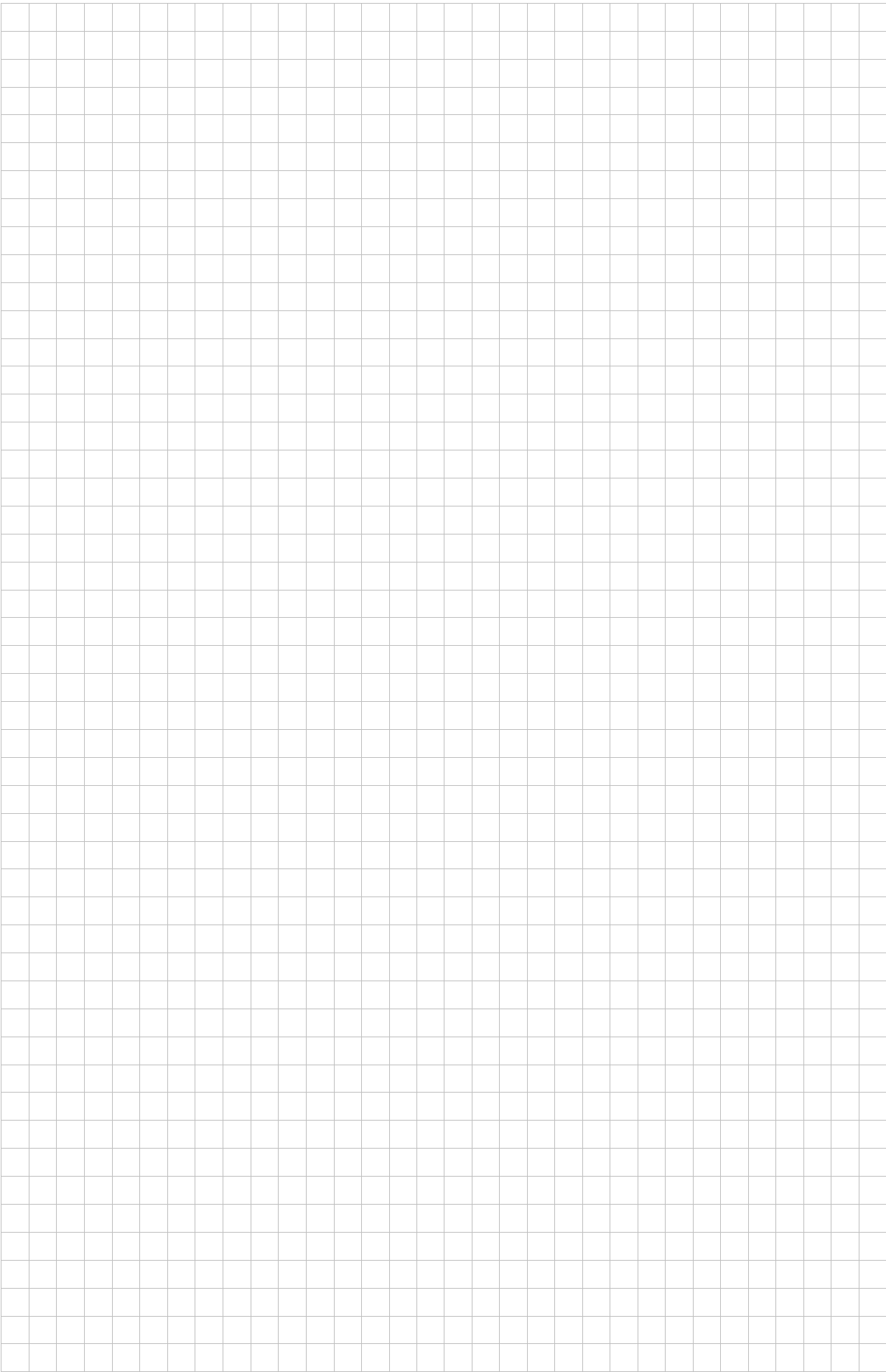
R2. Si nous posons que la vitesse de lumière est égale à l'unité, comme nous l'avons fait, l'axe des ordonnées du cône est dit "**axe purement temporel**".

R3. Il faut comprendre par soi-même que l'Univers a son propre cône d'Univers (cône... si l'espace est de type Minkowskien bien sûr...).

Enfin, indiquons que la théorie de la relativité restreinte, au même titre que celle de la relativité générale, n'impose pas un nombre de dimensions spatiales données pour rester consistante: ce qui est dommage pour les physiciens théoriciens qui souhaiteraient une théorie qui s'impose à elle-même un nombre fini de dimensions pour rester consistante (ce que par contre la théorie des cordes fait avec 25 dimensions... et celle des supercordes avec 11).







50. RELATIVITÉ GÉNÉRALE

Comme nous l'avons vu dans le chapitre précédent, la relativité restreinte est une réussite remarquable d'un point de vue théorique aussi bien que d'un point de vue pratique en formant un continuum d'espace-temps où les grandeurs d'espace et de temps se voient donner la même dimension physique (celle d'une distance métrique pour rappel!). Cependant, celle-ci s'applique aux repères euclidiens seulement et aux référentiels inertiels/Galiléens (à vitesse constante pour rappel...). Il convient donc de généraliser l'ensemble de la mécanique d'abord en exprimant ses principes et ses résultats fondamentaux sous une forme généralisée indépendante du type de systèmes de coordonnées choisis (in extenso: du type d'espace) en faisant usage du calcul tensoriel et ensuite de prendre en compte les systèmes non inertiels.

Il convient aussi de prendre en compte que le fait que la relativité restreinte ne s'applique qu'aux référentiels Galiléens est restrictif car toute masse crée un champ gravitationnel dont la portée est infinie. Pour pouvoir trouver un vrai référentiel galiléen, il est donc nécessaire de se situer infiniment loin de toute masse. La mécanique relativiste bâtie à partir de la relativité restreinte ne constitue donc qu'une approximation des lois de la nature, dans le cas où les champs gravitationnels ou les accélérations sont suffisamment faibles. Cette limitation d'application n'est plus adaptée à l'astrophysique relativiste donc l'activité s'est intensifiée à la fin du 20ème siècle.

C'est ici qu'encore une fois intervient Albert Einstein et nombre de ses confrères à travers le temps!

1. POSTULATS ET PRINCIPES

Einstein croyait en une physique ne devant privilégier aucun référentiel puisque telle était à ses yeux la réalité de l'Univers (nous en avons déjà fait mention). Mais comment se soustraire alors au phénomène d'accélération. L'idée géniale fut d'énoncer le "**postulat d'équivalence**" ci-dessous (qui encore aujourd'hui en ce début du 21ème siècle n'est toujours pas pris en défaut par les expériences récentes) en plus du postulat d'invariance et du principe cosmologique que nous avons énoncés dans le chapitre de Relativité Restreinte et de l'hypothèse selon laquelle le mouvement d'une particule qui ne subit aucune autre interaction que la gravitation est une ligne géodésique (voir plus loin la démonstration).

1.1. POSTULAT D'ÉQUIVALENCE

Dans un premier temps, Albert Einstein va améliorer le postulat d'équivalence dont les versions les plus anciennes sont dues à Galilée et à Newton:

Postulat: L'accélération (uniforme!) d'une masse (hors champ gravitationnel) due à l'application d'une force mécanique et l'accélération de cette même masse soumise à un champ gravitationnel (uniforme!) sont supposées parfaitement équivalentes. Ainsi, les résultats des analyses mathématiques dans un cas, peuvent s'appliquer dans l'autre (déjà là c'est fort mais cohérent... l'idée est très très bonne encore fallait-il l'avoir...!)

Autrement dit, le champ de gravité possède une propriété fondamentale qui le distingue de tous les autres champs connus dans la nature: le mouvement de chute libre des corps est universel, indépendant de la masse et de la composition des corps.

Corollaire: La masse au repos d'un corps doit alors être la même qu'elle soit mesurée dans un référentiel

dans un champ gravitationnel ou hors champ gravitationnel (nous parlons alors de masse inertielle et de masse pesante comme nous l'avons vu au tout début de notre étude de la mécanique classique).

Remarque: Il faut bien prendre garde et vérifier que le corollaire du postulat d'équivalence soit vrai sinon toute la relativité générale s'écroulerait (en ce début de 21ème siècle des expériences sont toujours en cours pour essayer de mettre en défaut cette équivalence)!

In extenso, tout champ de gravitation statique et uniforme est équivalent à un référentiel accéléré dans le vide. Nous pouvons physiquement considérer tout champ de gravitation comme statique et uniforme dans une région assez petite de l'espace et pendant un laps de temps assez court pour éviter les effets de marées. Nous sommes donc amenés à poser le "**principe d'équivalence faible**" (PEF): Pour tout événement de l'espace-temps dans un champ de gravitation arbitraire, nous pouvons choisir un référentiel dit "**référentiel localement inertielle**" tel que dans un voisinage de l'événement en question le mouvement libre de tous les corps (qui sont donc aussi dans le champ de gravité) soit rectiligne et uniforme tel qu'on puisse y appliquer les transformations de Lorentz (cf. chapitre de Relativité Restreinte).

Si nous mettons expérimentalement PEF en défaut, alors nous mettons en défaut le principe d'équivalence lui-même... ce qui n'a jamais pu être réalisé en laboratoire à ce jour!

Remarque: Le concept de localité est très important car il n'existe pas naturellement de champ de gravité uniforme. Par exemple, sur Terre, deux corps ponctuels distants d'une certaine longueur lâchés d'une certaine hauteur tomberont au sol avec une distance plus courte que la distance qui les séparait au moment où ils ont été lâchés. C'est ce que nous appelons en physique les "**effets de marées**": le champ gravitationnel n'est jamais uniforme (dans la nature en tout cas...).

Donc le principe d'équivalence (qui inclut le principe d'équivalence faible) affirme finalement que la force de Newton:

$$\vec{F} = m_i \vec{a} \quad (50.1)$$

et celle de la gravitation selon la forme de la loi de Newton-Poisson (cf. chapitre d'Astronomie):

$$\vec{F}_g = m_g \vec{g} = -m_g \vec{\nabla} \cdot \varphi \quad (50.2)$$

sont équivalentes telles que la masse inerte égale la masse pesante et l'accélération égale la pesanteur et qu'il n'est pas possible de distinguer les deux:

$$m_i = m_g \quad (50.3)$$

$$\vec{a} = \vec{g} \quad (50.4)$$

En quoi ce postulat permet-il de résoudre toutes les difficultés alors ? C'est simple ! L'idée est la suivante:

Lorsque nous allons considérer un corps en accélération, nous allons d'abord toujours assimiler celle-ci à l'accélération due à la chute dans un champ gravitationnel (de par l'application du principe

d'équivalence). Ensuite, nous allons supposer, et devons le vérifier (démonstration plus bas) en retrouvant la loi de Newton, que l'accélération due à ce champ gravitationnel n'est pas due au champ lui-même mais à la géométrie de l'espace déformé par la présence de la masse (in extenso l'énergie) qui crée le champ gravitationnel. Ainsi, l'objet n'est plus en "chute libre" mais sera vu comme glissant sur la trame spatiale déformée pour acquérir ainsi son accélération.

Au fait, l'enjeu est double:

1. Si le calcul tensoriel permet d'exprimer les lois de la mécanique classique et relativiste restreinte dans n'importe quel système de coordonnées, il est alors possible de voir comment le système de coordonnées (la métrique) agit sur l'expression des lois de l'Univers (Albert Einstein ne le savait pas tant qu'il n'avait pas terminé ses calculs mais le pressentait) !
2. Si l'expression tensorielle naturelle des lois de la mécanique fait apparaître le glissement (in extenso l'accélération) sur la trame spatiale suivant la métrique (locale) considérée, alors le pari est gagné et alors l'accélération peut être vue comme un effet dont la cause est purement géométrique.

Ainsi, l'extension de la relativité restreinte ne se fait plus en prenant en compte les systèmes non inertiels mais la géométrie du système!! Nous pouvons (et arrivons!) ainsi (à) contourner le problème initial et le pire... c'est que cela marche!!!!

Exemple:

Supposons que deux fusées, que nous nommerons *A* et *B*, se trouvent dans une région de l'espace éloignée de toute masse. Leurs moteurs sont arrêtés ce qui se traduit physiquement par un mouvement rectiligne uniforme. Dans chaque fusée, des physiciens réalisent des expériences de mécanique avec des objets dont ils connaissent la masse inerte. Soudain, le moteur de la fusée *A* démarre et lui communique une accélération dont l'effet ressenti à l'intérieur du vaisseau spatial est une force d'inertie qui plaque les objets vers le plancher. Pour les physiciens de la fusée *A* les lois de la mécanique sont alors les mêmes que celles que l'on observe dans un champ gravitationnel. Ils sont donc logiquement amenés à interpréter la force d'inertie comme la manifestation d'un champ gravitationnel. À l'aide d'une balance, ils peuvent alors peser leurs objets et leur attribuer une masse gravitationnelle.

Supposons que les physiciens de la fusée *B* puissent observer ce qui se passe dans la fusée *A*. Ils savent que ce que leurs collègues interprètent comme le poids des objets n'est en fait qu'une force d'inertie. La force d'inertie étant proportionnelle à l'accélération et à la masse inerte. Si la masse gravitationnelle était différente de la masse inerte les physiciens de la fusée *A* pourraient distinguer les effets des forces d'inertie de ceux d'un champ de gravitation car les masses mesurées seraient distinctes. Or, nous savons que la masse inerte et la masse gravitationnelle sont équivalentes (principe d'équivalence Galiléen). Il s'ensuit que les physiciens de la fusée *A* n'ont aucun moyen de faire la différence entre des forces d'inertie résultant d'un mouvement accéléré de leur vaisseau spatial et les forces d'attraction gravitationnelles.

Il faut toutefois tempérer les conclusions de cette expérience: les vrais champs de gravitation se distinguent d'un référentiel accéléré dans la mesure où l'accélération gravitationnelle varie avec la distance qui sépare les corps alors que dans un référentiel accéléré, l'accélération est identique en tout point de l'espace. Cependant, localement, un champ gravitationnel et un référentiel accéléré ne peuvent être différenciés.

Nous sommes donc amenés à énoncer le "**principe d'équivalence d'Einstein**" (PEE) tel que l'a fait

Einstein: localement, toutes les lois de la physique sont les mêmes dans un champ gravitationnel et dans un référentiel uniformément accéléré.

Ceci a une conséquence: Si la masse (qui est équivalente à de l'énergie comme nous l'avons vu en relativité restreinte) d'un objet n'est pas différenciable que nous soyons dans un champ gravitationnel ou dans un référentiel uniformément accéléré c'est que tous les types d'énergie (énergie de cohésion nucléaire, énergie électrostatique, énergie gravifique propre de l'objet, etc.) de cet objet ne sont pas différenciables. Donc les lois de la relativité restreinte sont valables quel que soit le référentiel considéré!

Si les lois ne sont pas les mêmes, alors PEE est mis en défaut, donc in extenso PEF aussi et plus généralement le principe d'équivalence dans sa généralité mais ceci n'est encore jamais arrivé expérimentalement.

Remarque: De par le PEF, il est intéressant de constater que le champ gravitationnel agit aussi sur l'énergie potentielle gravitationnelle des autres corps. Nous disons alors que le champ gravitationnel est un "**champ couplé**".

Étant donné qu'en relativité générale, le champ gravitationnel est censé être décrit par la métrique $g_{\mu\nu}$ (dont est munie la variété différentiable à 4 dimensions que constitue l'espace-temps), nous pouvons voir un référentiel localement inertiel comme un système de coordonnées de l'espace-temps dans lequel la métrique $g_{\mu\nu}$ devient plate (pseudo-Riemannienne):

$$g_{\mu\nu} = \eta_{\mu\nu} \quad (50.5)$$

Un tel système de coordonnées existe toujours, ce qui traduit l'existence, pour tout champ gravitationnel, de référentiels localement inertiels!

Lorsque la métrique n'est cependant pas plate, les coordonnées sont appelées "**coordonnées normales de Riemann**" et la métrique décrit alors un espace Riemannien (espace courbe) et dépend elle-même de manière non triviale des coordonnées du système.

1.2. PRINCIPE DE MACH

Si le principe d'équivalence met en évidence l'égalité des masses inerte et gravitationnelle, il ne nous éclaire pas sur la nature de ces deux masses. Finalement, que sont les masses inerte et gravitationnelle?

La nature profonde de la masse inerte devrait nous renseigner sur celle de l'inertie elle-même. L'inertie se manifeste sous une forme passive - le principe d'inertie - et une forme active - la seconde loi de Newton. D'une manière générale, elle exprime un comportement universel des corps à résister au changement du mouvement. Or nous savons que le mouvement inertiel est relatif, c'est-à-dire qu'il n'existe aucun référentiel absolu. En est-il de même du mouvement accéléré? Considérons, pour illustrer cette interrogation, une fusée dans laquelle se trouve un physicien et réalisons deux expériences:

- Première expérience. La fusée accélère: le physicien est soumis à une force d'inertie orientée dans la direction opposée à celle de l'accélération.
- Deuxième expérience. Maintenant supposons que l'on imprime à l'ensemble de l'Univers - à

l'exception de la fusée qui se déplace selon un mouvement inertiel - une accélération exactement opposée à celle de la fusée lors de l'expérience précédente.

Si le mouvement accéléré est relatif alors, pour un observateur, il n'est pas possible de distinguer les deux expériences. Notamment, le physicien situé à l'intérieur de la fusée doit observer l'apparition d'une force d'inertie absolument identique à celle qu'il a notée lors de la première expérience. La masse inerte trouverait alors son origine dans les interactions de la masse gravitationnelle des corps avec l'ensemble des masses gravitationnelles de l'Univers! Selon Ernst Mach, un physicien et philosophe du 19ème siècle, le mouvement quel qu'il soit, inertiel ou accéléré, serait relatif.

Cette théorie fut baptisée par Einstein "**principe de Mach**". Jusqu'à ce jour, le principe de Mach n'a pas été confirmé, mais pas davantage infirmé. Il est vrai que sa vérification expérimentale dépasse de beaucoup les capacités humaines!

Tout se passe comme si en déplaçant toutes les masses de l'Univers, celles-ci entraînaient avec elles les objets se trouvant dans la fusée, dont le physicien qui ressent alors une force qui le tire dans le même sens que l'accélération appliquée aux étoiles.

2. MÉTRIQUES

Einstein supposa donc que la gravitation n'était que la manifestation de déformations de l'espace-temps. Pour tenter d'illustrer de façon simpliste mais très imagée l'idée d'Einstein, considérons une roue dentée roulant à vitesse constante (disons une dent à la seconde) sur une crémaillère. Imaginons que nous ayons le pouvoir de modifier simultanément le pas de la crémaillère et celui de la roue quand et où nous le désirons. Faisons alors en sorte que le pas de la crémaillère augmente légèrement d'une dent à l'autre. Pour des observateurs fixes la roue est alors animée d'un mouvement uniformément accéléré car, en effet, à chaque tour celle-ci parcourt une distance toujours plus grande. En revanche, si l'on choisit la crémaillère comme référentiel et le pas de celle-ci comme étalon de mesure, le mouvement de la roue est alors uniforme (une dent par seconde). L'accélération de la roue est la conséquence de l'augmentation du pas de la crémaillère.

Poursuivons l'analogie: le pas de la crémaillère joue le rôle d'étalon de mesure local dans notre espace à une dimension que constitue la crémaillère. En géométrie, il porte le nom de "**métrique**". La métrique est ce qui permet de déterminer la distance entre deux points, elle représente en quelque sorte l'étalon infinitésimal d'un espace. En géométrie euclidienne la métrique est une constante, ce qui nous permet de créer des étalons de mesure universels. Bernhard Riemann, inventa une géométrie où la métrique peut varier d'un point à un autre de l'espace, ce qui lui permit de décrire des espaces courbes comme la surface d'une sphère par exemple (cf. chapitre de Géométries Non-Euclidiennes).

Lors de notre étude du calcul tensoriel, des géométries non-euclidiennes et de la géométrie différentielle (chapitres dont la lecture est plus que recommandée!!!), nous avons vu que la mesure de la distance ds entre deux points positionnés dans un espace à deux ou trois dimensions peut s'effectuer au moyen d'un grand nombre de système de coordonnées par "**l'équation métrique**" (cf. chapitre de Calcul Tensoriel):

$$ds^2 = dx_m dx^n = g_{mn} dx^m dx^n \quad (50.6)$$

En relativité générale, l'idée est de rendre le modèle théorique indépendant du fond et donc le construire sous une forme covariante (ce que certains assimilent à un postulat sous le nom de "**principe de covariance**"). Un excellent candidat pour ce type de démarche est donc d'utiliser le formalisme

tensoriel. Raison pour laquelle l'équation de métrique en constitue aussi un des piliers.

Exemples:

E1. Les coordonnées rectangulaires (dans $\mathbb{R}^3$):

$$\boxed{ds^2 = dx^2 + dy^2 + dz^2} \quad (50.7)$$

Si la distance au carré satisfait à cette relation alors nous sommes dans un espace plat (cf. chapitre de Géométries Non-Euclidiennes).

E2. Les coordonnées polaires (dans $\mathbb{R}^2$):

$$\begin{aligned} x &= r \cos \phi \Rightarrow dx = dr \cos \phi - r \sin \phi d\phi \\ y &= r \sin \phi \Rightarrow dy = dr \sin \phi + r \cos \phi d\phi \end{aligned} \quad (50.8)$$

d'où:

$$\begin{aligned} ds^2 &= dx^2 + dy^2 = (dr \cos \phi - r \sin \phi d\phi)^2 + (dr \sin \phi + r \cos \phi d\phi)^2 \\ &= dr^2 \cos^2 \phi - \cancel{2rdr \sin \phi \cos \phi d\phi} + r^2 \sin^2 \phi d\phi^2 + dr^2 \sin^2 \phi + \cancel{2rdr \sin \phi \cos \phi d\phi} + r^2 \cos^2 \phi d\phi^2 \\ &= dr^2 (\cos^2 \phi + \sin^2 \phi) + r^2 (\cos^2 \phi + \sin^2 \phi) d\phi^2 \end{aligned} \quad (50.9)$$

d'où:

$$\boxed{ds^2 = dr^2 + r^2 d\phi^2} \quad (50.10)$$

Si la distance au carré satisfait à cette relation alors nous sommes dans un espace plat (cf. chapitre de Géométries Non-Euclidiennes).

E3. Les coordonnées cylindriques (dans $\mathbb{R}^3$) pour lesquelles nous avons:

$$\begin{aligned} x &= r \cos \phi \Rightarrow dx = dr \cos \phi - r \sin \phi d\phi \\ y &= r \sin \phi \Rightarrow dy = dr \sin \phi + r \cos \phi d\phi \\ z &= z \Rightarrow dz = dz \end{aligned} \quad (50.11)$$

à remplacer dans $ds^2 = dx^2 + dy^2 + dz^2$ nous obtenons de façon quasiment identique à précédemment:

$$ds^2 = dr^2 + r^2 d\phi^2 + dz^2 \quad (50.12)$$

Si la distance au carré satisfait à cette relation alors nous sommes dans un espace courbe (de type cylindrique) mais qui localement peut être plat (cf. chapitre de Géométries Non-Euclidiennes). En réalité, pour avoir la métrique de la surface du cylindre et non pas simplement du plan exprimé en coordonnées cylindriques, il faudra prendre la métrique suivante:

$$\boxed{ds^2 = dr^2 + r^2 d\phi^2} \quad (50.13)$$

dont l'origine a été démontrée dans le chapitre de Géométrie Différentielle.

E4. Les coordonnées sphériques (dans $\mathbb{R}^3$) pour lesquelles nous avons:

$$\begin{aligned}x &= r \sin \theta \cos \phi \Rightarrow dx = dr \sin \theta \cos \phi + r \cos \theta d\theta \cos \phi - r \sin \theta \sin \phi d\phi \\y &= r \sin \theta \sin \phi \Rightarrow dy = dr \sin \theta \sin \phi + r \cos \theta d\theta \sin \phi + r \sin \theta \cos \phi d\phi \\z &= r \cos \theta \Rightarrow dz = dr \cos \theta - r \sin \theta d\theta\end{aligned} \quad (50.14)$$

à remplacer dans $ds^2 = dx^2 + dy^2 + dz^2$ nous obtenons:

$$\begin{aligned}ds^2 &= \\& (dr \sin \theta \cos \phi + r \cos \theta d\theta \cos \phi - r \sin \theta \sin \phi d\phi)^2 \\& + (dr \sin \theta \sin \phi + r \cos \theta d\theta \sin \phi + r \sin \theta \cos \phi d\phi)^2 \\& + (dr \cos \theta - r \sin \theta d\theta)^2\end{aligned} \quad (50.15)$$

Petit rappel préalable:

$$\begin{aligned}(a+b-c)^2 &= a^2 + b^2 + c^2 + 2(ab - ac - bc) \\(a+b+c)^2 &= a^2 + b^2 + c^2 + 2(ab + ac + bc) \\(a-b)^2 &= a^2 - 2ab + b^2\end{aligned} \quad (50.16)$$

Donc:

$$\begin{aligned}ds^2 &= \\& dr^2 \sin^2 \theta \cos^2 \phi + r^2 \cos^2 \theta d\theta^2 \cos^2 \phi + r^2 \sin^2 \theta \sin^2 \phi d\phi^2 \\& + 2(rdrd\theta \cos \theta \sin \theta \cos^2 \phi - rdrd\phi \sin^2 \theta \cos \phi \sin \phi - r^2 d\theta d\phi \cos \theta \sin \theta \cos \phi \sin \phi) \\& + dr^2 \sin^2 \theta \sin^2 \phi + r^2 \cos^2 \theta d\theta^2 \sin^2 \phi + r^2 \sin^2 \theta \cos^2 \phi d\phi^2 \\& + 2(rdrd\theta \sin \theta \cos \theta \sin^2 \phi + rdrd\phi \sin^2 \theta \sin \phi \cos \phi + r^2 d\theta d\phi \cos \theta \sin \theta \sin \phi \cos \phi) \\& + dr^2 \cos^2 \theta - 2rdrd\theta \cos \theta \sin \theta + r^2 \sin^2 \theta d\theta^2\end{aligned} \quad (50.17)$$

Après une première série de mise en commun et de simplifications élémentaires des termes identiques, nous obtenons:

$$ds^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (50.18)$$

Si la distance au carré satisfait à cette relation alors nous sommes dans un espace courbe (de type sphérique) mais qui localement peut être plat (cf. chapitre de Géométries Non-Euclidiennes). En réalité, pour avoir la métrique de la surface de la sphère et non pas simplement du plan exprimé en coordonnées sphériques, il faudra prendre la métrique suivante:

$$ds^2 = r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (50.19)$$

dont l'origine a été démontrée dans le chapitre de Géométrie Différentielle. Nous avons par ailleurs vérifié dans le chapitre de Calcul Tensoriel, que la courbure de Ricci de la métrique sphérique antécédente était nulle. Par contre, nous avons tout de suite après vérifié que si nous prenions la métrique précédente de la surface de la sphère, la courbure de Ricci était non nulle (et c'est encore heureux!).

Jusque-là, vous vous demandez peut-être où nous voulons en venir. Au fait, nous cherchons à définir à partir de ces relations, un être mathématique qui en concordance avec l'hypothèse d'Einstein, exprime les propriétés géométriques d'espaces donnés.

Comment allons-nous faire?: Nous allons d'abord changer d'écriture tout simplement. Au lieu d'utiliser les symboles $x, y, z, \theta, \varphi, r$ nous allons écrire $x^1, x^2, x^3, \dots$. Attention! Les chiffres en suffixes ne sont pas des puissances. Ce sont des valeurs muettes qui sont là uniquement pour symboliser la x -ème coordonnée d'un repère donné.

Écrivons maintenant à nouveau nos équations métriques avec cette nouvelle notation en considérant qu'il ne s'agit que d'exemples particuliers qui n'ont pas nécessairement un sens physique pertinent (nous l'avons par ailleurs mentionné précédemment!):

- Coordonnées rectangulaires:

$$ds^2 = (dx^1)^2 + (dx^2)^2 + (dx^3)^2 \quad (50.20)$$

- Coordonnées polaires:

$$ds^2 = (dx^1)^2 + (x^1)^2 (dx^2)^2 \quad (50.21)$$

- Coordonnées cylindriques:

$$ds^2 = (dx^1)^2 + (x^1)^2 (dx^2)^2 + (dx^3)^2 \quad (50.22)$$

- Coordonnées sphériques:

$$ds^2 = (dx^1)^2 + (x^1)^2 (dx^2)^2 + (x^1)^2 \sin^2(x^2) (dx^3)^2 \quad (50.23)$$

Maintenant rappelons encore une fois que le "**tenseur métrique**" (nommé ainsi car il étalonne l'espace-temps) noté:

$$g_{mn} = \begin{bmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ g_{31} & g_{32} & g_{33} \end{bmatrix} \quad (50.24)$$

intervient dans l'équation métrique de la manière suivante:

$$ds^2 = dx_m dx^n = g_{mn} dx^m dx^n \quad (50.25)$$

et remarquez que les composantes de la matrice sont sans dimensions aussi.

Cet être mathématique qui est un tenseur contient donc les paramètres de la courbure (nous disons parfois aussi de la "contrainte" ou de la "tension") dans lequel un espace se trouve. Mais alors que contient le tenseur métrique d'espace-temps pour un espace euclidien plat?

Selon la convention d'écriture de sommation d'Einstein (cf. chapitre de Calcul Tensoriel) par exemple, pour $m = n = 2$ nous avons:

$$ds^2 = g_{mn} dx^m dx^n = g_{11} dx^1 dx^1 + g_{12} dx^1 dx^2 + g_{21} dx^2 dx^1 + g_{22} dx^2 dx^2 \quad (50.26)$$

Donc si nous revenons à notre tenseur pour l'espace euclidien plat, nous savons déjà (cf. chapitre de Calcul Tensoriel) que m et n vont de 1 à 3 et que nous avons dans notre tenseur $g_{mn} = 0$ pour $m \neq n$ et $g_{mn} = 1$ pour $m = n$ (tenseur symétrique). Donc:

$$\begin{aligned} ds^2 &= g_{mn} dx^m dx^n = g_{11} dx^1 dx^1 + g_{22} dx^2 dx^2 + g_{33} dx^3 dx^3 \\ &= 1 \cdot (dx^1)^2 + 1 \cdot (dx^2)^2 + 1 \cdot (dx^3)^2 = (dx^1)^2 + (dx^2)^2 + (dx^3)^2 \end{aligned} \quad (50.27)$$

Ainsi:

$$g_{mn} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (50.28)$$

où comme d'habitude sur ce site nous faisons l'abus de notation (déjà mentionné dans le chapitre de Calcul Tensoriel) de ne pas mettre g_{mn} entre crochets (puisque un tenseur et sa forme matricielle sont normalement deux choses distinctes en toute rigueur).

Ce résultat est remarquable, car le tenseur métrique va donc nous permettre de définir les propriétés d'un espace à partir d'un simple être mathématique facilement manipulable formellement.

En coordonnées polaires le tenseur g_{mn} s'écrit:

$$g_{mn}(r, \phi) = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & (x^1)^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & r^2 \end{bmatrix} \quad (50.29)$$

Vérification:

$$ds^2 = g_{mn} dx^m dx^n = g_{11} dx^1 dx^1 + g_{22} dx^2 dx^2 = (dx^1)^2 + (x^1)^2 (dx^2)^2 \quad (50.30)$$

En coordonnées cylindriques le tenseur g_{mn} s'écrit:

$$g_{mn}(r, \phi, z) = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{33} & x_{32} & x_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & (x^1)^2 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (50.31)$$

La vérification ne se fait même plus tellement le résultat est évident.

En coordonnées sphériques le tenseur g_{mn} est un peu plus complexe et s'écrit:

$$g_{mn}(r, \theta, \varphi) = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & (x^1)^2 & 0 \\ 0 & 0 & (x^1 \sin(x^2))^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2 \theta \end{bmatrix} \quad (50.32)$$

La vérification ne se fait même plus tellement le résultat est évident.

En relativité restreinte, nous avons vu que les notions d'espace et de temps étaient implicitement liées. Ainsi, pour étudier la physique (cela intéresse peu le mathématicien), nous avons besoin d'ajouter à notre tenseur métrique la composante du temps pour obtenir ce que nous appelons le "**tenseur métrique d'espace-temps**".

Pour déterminer l'écriture de ce tenseur, nous allons nous placer dans un premier temps dans un espace de Minkowski où nous avons rappelés-le (cf. chapitre de Relativité Restreinte):

$$ds^2 = c^2 \cdot dt^2 - (dx^2 + dy^2 + dz^2) \quad (50.33)$$

qui est donc l'intervalle infinitésimal d'espace-temps entre deux événements infiniment voisins (ou considérés comme tels à une certaine échelle...).

Ainsi, en posant:

$$x^0 = ct, x^1 = x, x^2 = y, x^3 = z \quad (50.34)$$

Nous avons:

$$g_{\mu\nu} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \quad (50.35)$$

avec la "**signature**":

$$\text{sign}(g_{\mu\nu}) = (1, -1, -1, -1) = +--- \quad (50.36)$$

Remarque: Pour tous les tenseurs métriques que nous avons déterminés avant, si nous les exprimons dans l'espace-temps (donc en rajoutant le temps), les composantes spatiales ont toutes un signe négatif!

Nous verrons par la suite d'autres métriques beaucoup moins intuitives une fois que nous aurons démontré bien plus loin l'équation d'Einstein des champs.

2.1. CRITÈRE DE SCHILD

Nous allons maintenant voir que pour étudier la gravitation, la géométrie courbe est nécessaire après quoi (il nous faudra démontrer l'équation des géodésiques avant!) nous montrerons qu'elle est également suffisante. Nous verrons que la gravitation telle qu'elle est formulée en mécanique

newtonienne est entièrement descriptible à partir d'une formulation de courbure de l'espace-temps.

Imaginons d'abord une tour d'une très grande hauteur h construite à la surface de la Terre. Un homme A se trouve au pied de la tour, et envoie un signal de pulsation ω_A à son collègue B situé en haut de la tour. Il se trouve, et nous allons de suite le démontrer, que la pulsation ω_B de l'onde reçue par B diffère de ω_A selon:

$$\frac{\omega_A}{\omega_B} = 1 + \frac{gh}{c^2} \quad (50.37)$$

D'où:

$$\Delta\omega = \omega_A - \omega_B > 0 \quad (50.38)$$

Ce décalage des pulsations (respectivement des fréquences) dans un champ gravitationnel est ce que nous appelons "**l'effet Einstein**", ou encore "**redshift gravitationnel**".

Nous allons démontrer cette relation à l'aide d'arguments classiques et connus maintenant.

Un corps matériel envoyé du sol vers le ciel doit lutter contre la force de gravitation qui l'attire vers le bas. Il perdra donc une certaine quantité d'énergie, équivalant à l'énergie potentielle gravitationnelle gagnée durant le trajet. L'énergie E_A du corps au niveau du sol est donc son énergie de masse à laquelle s'ajoute l'énergie potentielle à la hauteur de la tour:

$$E_A = mc^2 + mgh \quad (50.39)$$

L'énergie de ce corps une fois arrivé en haut de la tour est simplement son énergie de masse:

$$E_B = mc^2 \quad (50.40)$$

car il a dû dépenser la quantité d'énergie mgh durant le trajet. Le rapport des énergies est alors:

$$\frac{E_A}{E_B} = \frac{mc^2 + mgh}{mc^2} = 1 + \frac{gh}{c^2} \quad (50.41)$$

Ce rapport étant indépendant de la masse, on peut prendre la limite $m \rightarrow 0$ afin d'avoir la relation pour le photon. Nous obtenons alors:

$$\frac{E_A}{E_B} = \frac{\hbar\omega_A}{\hbar\omega_B} = \frac{\omega_A}{\omega_B} = \frac{f_A}{f_B} = 1 + \frac{gh}{c^2} \quad (50.42)$$

ce qui implique:

$$\omega_A > \omega_B \quad (50.43)$$

Nous allons maintenant étudier ce phénomène dans le cadre de l'espace-temps de Minkowski. Nous verrons apparaître une contradiction, ce qui motivera le passage vers un espace-temps courbe: c'est l'argument en faveur d'une géométrie courbe qui a été utilisé par Schild.

Considérons à nouveau le schéma d'expérience de l'homme A qui envoie une onde vers son ami B . Soit Δt_A le temps mis par A pour émettre exactement 1 cycle de l'onde (cf. chapitre de Mécanique Ondulatoire):

$$\Delta t_A = \frac{\lambda_A}{c} = \frac{1}{\omega_A} \quad (50.44)$$

et Δt_B le temps mis par B pour recevoir ce cycle:

$$\Delta t_B = \frac{\lambda_B}{c} = \frac{1}{\omega_B} \quad (50.45)$$

À cause de l'effet Einstein, nous savons que $\omega_A > \omega_B$ et donc $\Delta t_A < \Delta t_B$ en temps propre! Soit en fait que le temps passe plus lentement pour quelqu'un au sol (A) que pour une autre personne en haut d'une montagne (B)!

Mais comme nous sommes en géométrie plate et que le champ gravitationnel est supposé statique, nous en déduisons que les trajectoires d'espace-temps décrites par les signaux doivent être parallèles. Ceci mène à la conclusion que l'intervalle de temps propre serait $\Delta t_A = \Delta t_B$ (selon la relativité restreinte).

Si nous optons pour un espace courbe, nous pouvons préserver la relation $\Delta t_A < \Delta t_B$, c'est-à-dire le fait que le temps avance plus lentement pour A que pour B . Ceci se traduit simplement par le fait qu'en géométrie courbe, le temps propre (!) d'un observateur dépend de la métrique.

Les mêmes développements peuvent être faits en assimilant l'expérience précédente à un train qui se déplace avec une accélération constante g . L'observateur A se trouve dans le compartiment arrière (équivalant au sol de la Terre dans l'expérience précédente) et envoie une onde à son collègue B situé à l'avant du train (à une distance h).

L'observateur B reçoit l'onde après un temps $\Delta t = h/c$. Durant ce laps de temps, le train a accéléré, et sa vitesse a augmenté d'une valeur $\Delta v = g\Delta t = gh/c$. Par conséquent, l'onde perçue par B sera altérée par l'effet Doppler conventionnel (cf. chapitre de Mécanique Ondulatoire):

$$\frac{\Delta \lambda}{\lambda_A} = \frac{\Delta v \Delta t}{c \Delta t} = \frac{\Delta v}{c} = \frac{gh}{c^2} \quad (50.46)$$

Nous retrouvons le résultat initial de l'effet Einstein en écrivant simplement:

$$\frac{\Delta \lambda}{\lambda_A} = \frac{\lambda_B - \lambda_A}{\lambda_A} = \frac{\lambda_B}{\lambda_A} - 1 = \frac{\omega_B}{\omega_A} - 1 \quad (50.47)$$

ce qui donne glorieusement:

$$\left. \begin{aligned} \frac{\Delta \lambda}{\lambda_A} &= \frac{gh}{c^2} \\ \frac{\Delta \lambda}{\lambda_A} &= \frac{\omega_B}{\omega_A} - 1 \end{aligned} \right\} : \frac{\omega_B}{\omega_A} = 1 + \frac{gh}{c^2} \quad (50.48)$$

Nous retrouvons plus souvent cette relation sous la forme ci-dessous dans la littérature en utilisant les relations entre pulsation et fréquence et la force de gravitation de Newton pour expliciter g et en posant h comme valant 1:

$$\frac{f_B - f_A}{f_A} = \frac{GM}{c^2} \quad (50.49)$$

Nous retrouvons également cette dernière relation sous la forme condensée suivante:

$$\frac{\Delta f}{f_0} = \frac{\Delta \lambda}{\lambda_0} = \frac{a_g}{c^2} \quad (50.50)$$

Le même résultat peut être obtenu en utilisant la métrique de Schwarzschild (voir plus loin), d'où le nom de cet effet qui peut aussi être obtenu à partir des outils de la relativité générale d'Einstein. Nous démontrerons simplement plus tard à l'aide de cette métrique que le temps s'écoule effectivement moins vite dans un champ gravitationnel (hypothèse que nous avons faite quelques paragraphes plus haut).

Nous voyons dans tous les cas que $f_B > f_A$ puisque le terme de droite est positif et non nul. Cela signifie simplement que l'onde électromagnétique en analogie au spectre des couleurs se décale vers le rouge. Ainsi, l'effet Einstein est bien un redshift gravitationnel.

La différence de fréquence est très faible et par conséquent difficilement mesurable même avec les meilleurs spectroscopes. La moindre perturbation peut totalement masquer l'effet Einstein. Il faudra véritablement attendre 1960 pour que l'expérience de Pound et Rebka permette de mesurer un décalage de fréquences avec une précision de 1% ne laissant dès lors plus aucun doute quant à la réalité du phénomène.

3. ÉQUATIONS DU MOUVEMENT

Nous allons démontrer ici que l'équation du mouvement d'une particule libre est constante le long de sa ligne d'Univers en nous limitant d'abord au cas d'un espace plat (de type espace de Minkowski). Après quoi, nous généraliserons ce résultat à tout type d'espace en utilisant un développement simple, pour montrer de manière assez évidente que l'équation du mouvement est indépendante de la masse et suit la courbure de l'espace!!! Enfin, nous présenterons une deuxième démonstration dans tout type d'espace en utilisant le principe variationnel.

Commençons donc par démontrer l'équation du mouvement d'une particule libre dans un espace plat.

Lors de notre étude de la relativité restreinte, nous avons démontré le lagrangien relativiste d'une particule libre donné par (attention! la notation m est celle de la masse au repos m_0 de la particule conformément à ce que nous avons montré dans le chapitre de Relativité Restreinte!!!):

$$L = -mc^2 \sqrt{1 - v^2/c^2} \quad (50.51)$$

et pour cela nous étions partis de l'action (hypothétique):

$$S_0 = -mc^2 \int dt' \quad (50.52)$$

et nous étions arrivés à écrire:

$$S_0 = -mc^2 \int \sqrt{1 - \frac{v^2}{c^2}} dt \quad (50.53)$$

Maintenant, montrons quelque chose d'intéressant. Rappelons que pour l'espace-temps de Minkowski, nous avons obtenu:

$$ds^2 = c^2 \cdot dt^2 - (dx^2 + dy^2 + dz^2) \quad (50.54)$$

en nous restreignant à une seule dimension spatiale, nous obtenons comme relation:

$$ds^2 = c^2 dt^2 - dx^2 \Leftrightarrow ds^2 = c^2 dt^2 \left(1 - \frac{dx^2}{c^2 dt^2} \right) = c^2 dt^2 \left(1 - \frac{v^2}{c^2} \right) \quad (50.55)$$

$$ds = c dt \sqrt{1 - \frac{v^2}{c^2}}$$

et alors... eh bien voilà au fait, si nous posons:

$$S = -mc \int ds \quad (50.56)$$

nous avons finalement:

$$S = -mc \int ds = -mc \int c dt \sqrt{1 - \frac{v^2}{c^2}} = -mc^2 \int \sqrt{1 - \frac{v^2}{c^2}} dt \quad (50.57)$$

nous retrouvons donc la même action à partir d'une forme plus générale (pure) de l'action qui est:

$$S = -mc \int ds \quad (50.58)$$

résultat que nous avons aussi démontré dans le chapitre d'Électrodynamique!! Nous pouvons même faire mieux en termes d'élégance...! Si nous observons bien les développements des lignes précédentes, nous observons qu'au fait la relation:

$$ds = c dt \sqrt{1 - \frac{v^2}{c^2}} \quad (50.59)$$

est le cas particulier à une dimension de la relation:

$$ds = dt \sqrt{\frac{g_{\mu\nu} \dot{x}^\mu \dot{x}^\nu}{c^2}} = \frac{1}{c} \sqrt{g_{\mu\nu} \dot{x}^\mu \dot{x}^\nu} dt \quad (50.60)$$

avec comme défini plus haut:

$$x^0 = ct, x^1 = x, x^2 = y, x^3 = z \quad (50.61)$$

et donc:

$$\dot{x}^0 = c, \dot{x}^1 = v_x, \dot{x}^2 = v_y, \dot{x}^3 = v_z \quad (50.62)$$

Effectivement, si nous prenons le cas à une dimension dans un espace plat de Minkowski:

$$ds = dt \sqrt{g_{\mu\nu} \dot{x}^\mu \dot{x}^\nu} = dt \sqrt{\eta_{00} \dot{x}^0 \dot{x}^0 + \eta_{11} \dot{x}^1 \dot{x}^1} = dt \sqrt{c^2 - 1 \cdot v_x^2} = c dt \sqrt{1 - \frac{v_x^2}{c^2}} \quad (50.63)$$

Ainsi, nous avons le facteur de Fitzgerald-Lorentz qui est donné en toute généralité par:

$$\gamma = \frac{1}{c} \sqrt{g_{\mu\nu} \dot{x}^\mu \dot{x}^\nu} \quad (50.64)$$

comme généralisation de la Relativité Restreinte!

Ceci étant fait, revenons à nos moutons... Dans un espace sans champ de potentiel, nous avons démontré dans le chapitre de Mécanique Analytique que le lagrangien se réduit à la simple expression de l'énergie cinétique tel que:

$$L = \frac{1}{2} m v^2 = \frac{1}{2} m \left(\frac{dx}{dt} \right)^2 \quad (50.65)$$

Si nous souhaitons généraliser cette relation pour qu'elle soit valable dans n'importe quel type d'espace (courbe ou plat), il nous faut introduire les coordonnées curvilignes telles que nous les avons étudiées en calcul tensoriel (cf. chapitre de Calcul Tensoriel).

Dans un premier temps, cela donne:

$$L = \frac{1}{2} m \left(\frac{ds}{dt} \right)^2 \quad (50.66)$$

où rappelons-le ds est l'abscisse curviligne de la trajectoire.

Et nous avons démontré en calcul tensoriel que:

$$ds = \left[g_{ij} \frac{\partial u^i}{\partial \alpha} \frac{\partial u^j}{\partial \alpha} \right]^{1/2} d\alpha \quad (50.67)$$

Cette dernière relation s'écrit dans le contexte de la mécanique relativiste de manière plus standard:

$$ds = \left(g_{\alpha\beta} \frac{\partial x^\alpha}{\partial t} \frac{\partial x^\beta}{\partial t} \right)^{1/2} dt \quad (50.68)$$

t étant un paramètre qui correspond en mécanique au temps propre de la particule et qui dans la littérature spécialisée est souvent noté τ .

Avant de nous intéresser aux espaces courbes décrits par la métrique g_{ij} (ce que nous ferons lors de

notre démonstration du lagrangien libre généralisé), restreignons-nous à l'espace euclidien avec la métrique (ce sera un bon exercice pour bien comprendre) donnée par la matrice de Minkowski (cf. chapitre de Relativité Restreinte):

$$\xi_{\alpha\beta} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \quad (50.69)$$

que nous noterons $\eta_{\alpha\beta}$ pour la différencier des autres (car plus souvent utilisée). Nous avons finalement dans l'espace euclidien:

$$S = -mc \int ds = -mc \int \left(\eta_{\alpha\beta} \frac{\partial x^\alpha}{\partial t} \frac{\partial x^\beta}{\partial t} \right)^{1/2} dt \quad (50.70)$$

Maintenant, appliquons le principe variationnel:

$$\delta S = -mc \int \delta(ds) \quad (50.71)$$

La variation de ds peut être trouvée plus simplement à partir de la variation de ds^2 :

$$(ds)^2 = ds^2 = \eta_{\alpha\beta} \frac{\partial x^\alpha}{\partial t} \frac{\partial x^\beta}{\partial t} dt^2 = \eta_{\alpha\beta} dx^\alpha dx^\beta \quad (50.72)$$

nous trouvons:

$$\delta(ds)^2 = 2ds\delta(ds) = \eta_{\alpha\beta} \delta(dx^\alpha dx^\beta) = \eta_{\alpha\beta} \delta dx^\alpha dx^\beta + \eta_{\alpha\beta} dx^\alpha \delta dx^\beta = 2\eta_{\alpha\beta} \delta(dx^\alpha) dx^\beta \quad (50.73)$$

Le facteur "2" provient du fait que par symétrie de l'espace euclidien, les variations de dx^α et dx^β sont égales.

Remarque: Comme nous le verrons après, cette relation de $\delta(ds)^2$ ne sera plus identique lorsque nous traiterons des espaces courbes.

En simplifiant un peu, nous obtenons:

$$2ds\delta(ds) = 2\eta_{\alpha\beta} \delta(dx^\alpha) dx^\beta \Leftrightarrow \delta(ds) = \eta_{\alpha\beta} \delta(dx^\alpha) \frac{dx^\beta}{ds} \quad (50.74)$$

Ce qui est équivalent à écrire:

$$\delta(ds) = \eta_{\alpha\beta} d(\delta x^\alpha) \frac{dx^\beta}{ds} = \eta_{\alpha\beta} \frac{d(\delta x^\alpha)}{dt} \frac{dx^\beta}{ds} dt \quad (50.75)$$

Nous pouvons maintenant revenir à l'action:

$$\delta S = -mc \int \delta(ds) = -mc \int \eta_{\alpha\beta} \frac{d(\delta x^\alpha)}{dt} \frac{dx^\beta}{ds} dt \quad (50.76)$$

Nous réécrivons l'intégrale précédente (ce sera plus simple à traiter):

$$\delta S = -mc \int \eta_{\alpha\beta} \frac{d(\delta x^\alpha)}{dt} \frac{dx^\beta}{ds} dt = -mc \int_{t_1}^{t_2} dt \frac{d}{dt} \left(\eta_{\alpha\beta} \delta x^\alpha \frac{dx^\beta}{ds} \right) + \int_{t_1}^{t_2} dt \delta x^\alpha \left(mc \eta_{\alpha\beta} \frac{d}{dt} \frac{dx^\beta}{ds} \right) \quad (50.77)$$

Effectivement, vérifions que cette forme est bien équivalente:

$$\begin{aligned} \delta S &= -mc \int_{t_1}^{t_2} dt \frac{d}{dt} \left(\eta_{\alpha\beta} \delta x^\alpha \frac{dx^\beta}{ds} \right) + \int_{t_1}^{t_2} dt \delta x^\alpha \left(mc \eta_{\alpha\beta} \frac{d}{dt} \frac{dx^\beta}{ds} \right) \\ &= -mc \int_{t_1}^{t_2} dt \left(\frac{d\eta_{\alpha\beta}}{dt} \delta x^\alpha \frac{dx^\beta}{ds} \right) - mc \int_{t_1}^{t_2} dt \left(\eta_{\alpha\beta} \frac{d(\delta x^\alpha)}{dt} \frac{dx^\beta}{ds} \right) - mc \int_{t_1}^{t_2} dt \left(\eta_{\alpha\beta} \delta x^\alpha \frac{d}{dt} \frac{dx^\beta}{ds} \right) \\ &\quad + \int_{t_1}^{t_2} dt \delta x^\alpha \left(mc \eta_{\alpha\beta} \frac{d}{dt} \frac{dx^\beta}{ds} \right) = -mc \int_{t_1}^{t_2} dt \eta_{\alpha\beta} \frac{d(\delta x^\alpha)}{dt} \frac{dx^\beta}{ds} \end{aligned} \quad (50.78)$$

Donc revenons à notre intégrale:

$$\delta S = -mc \int_{t_1}^{t_2} dt \frac{d}{dt} \left(\eta_{\alpha\beta} \delta x^\alpha \frac{dx^\beta}{ds} \right) + \int_{t_1}^{t_2} dt \delta x^\alpha \left(mc \eta_{\alpha\beta} \frac{d}{dt} \frac{dx^\beta}{ds} \right) \quad (50.79)$$

Nous avons donc deux intégrales qu'il va être un peu plus simple d'analyser. La première intégrale:

$$-mc \int_{t_1}^{t_2} dt \left(\eta_{\alpha\beta} \delta x^\alpha \frac{dx^\beta}{ds} \right) = -mc \left[\eta_{\alpha\beta} \delta x^\alpha \frac{dx^\beta}{ds} \right]_{t_1}^{t_2} \quad (50.80)$$

donne simplement une expression évaluée aux extrémités temporelles (t_1, t_2) . Dès lors, comme les valeurs de x^α sont parfaitement connues aux extrémités temporelles, le variationnel δx^α est nul aux deux bornes et cette intégrale est nulle.

Il ne nous reste alors plus que l'intégrale:

$$\delta S = \int_{t_1}^{t_2} dt \delta x^\alpha \left(mc \eta_{\alpha\beta} \frac{d}{dt} \frac{dx^\beta}{ds} \right) = \int_{t_1}^{t_2} dt \delta x^\alpha \eta_{\alpha\beta} \frac{d}{dt} \left(mc \frac{dx^\beta}{ds} \right) \quad (50.81)$$

Donc pour que le principe variationnel $\delta S = 0$ (cf. chapitre de Mécanique Analytique) soit respecté, il faut que nous ayons:

$$\frac{d}{dt} \left(mc \frac{dx^\beta}{ds} \right) = 0 \quad (50.82)$$

Or, nous pouvons réécrire une partie de cette expression. Effectivement, nous avons:

$$mc \frac{dx^\beta}{ds} = mc \left(\frac{d(ct)}{ds}, \frac{dx}{ds}, \frac{dy}{ds}, \frac{dz}{ds} \right) \quad (50.83)$$

Rappelons par ailleurs que nous avons démontré plus haut que:

$$ds = c dt \sqrt{1 - \beta^2} \Leftrightarrow \frac{dt}{ds} = \frac{1}{c \sqrt{1 - \beta^2}} = \frac{\gamma}{c} \quad (50.84)$$

et que nous avons:

$$\frac{dx}{ds} = \frac{dx}{dt} \frac{dt}{ds} = v_x \frac{dt}{ds} = \frac{v_x \gamma}{c} \quad (50.85)$$

Donc:

$$mc \left(\frac{d(ct)}{ds}, \frac{dx}{ds}, \frac{dy}{ds}, \frac{dz}{ds} \right) = m\gamma (c, v_x, v_y, v_z) \quad (50.86)$$

Maintenant, rappelons que lors de notre étude de la relativité restreinte, nous avons démontré le cheminement qui nous amenait à définir le quadrivecteur d'énergie impulsion:

$$p^\nu = \left(\frac{E}{c}, p_x, p_y, p_z \right) = \left(\frac{E}{c}, \vec{p} \right) = m\gamma (c, \vec{v}) \quad (50.87)$$

Donc finalement, ce qui annule le variationnel de l'intégrale d'action peut s'écrire:

$$\frac{dp^\nu}{dt} = 0 \quad (50.88)$$

Nous retrouvons donc l'équation de conservation de la quantité de mouvement (conservation de l'impulsion) que nous appelons dans le cadre de la relativité générale "**équation du mouvement**". Cette forme de l'équation du mouvement semble dépendante de la masse mais en fouillant un peu, nous verrons qu'il n'en est rien.

En multipliant cette relation par $\eta_{\mu\nu}$ nous pouvons aussi écrire:

$$\frac{dp_\mu}{dt} = 0 \quad (50.89)$$

et de même pour un autre observateur:

$$\frac{dp_\mu}{dt \gamma} = 0 \Leftrightarrow \frac{dp_\mu}{dt'} = 0 \quad (50.90)$$

En d'autres termes, l'impulsion de la particule reste constante sur toute sa ligne d'Univers.

Mais nous pouvons aussi écrire:

$$\frac{dp^v}{dt} = \frac{dp^v}{\left(\frac{ds}{c} \gamma\right)} = 0 \quad (50.91)$$

donc:

$$\boxed{\frac{dp^v}{ds} = 0} \quad (50.92)$$

Une forme plus importante encore de l'équation du mouvement peut être obtenue. Effectivement:

$$p^v = mc \frac{dx^v}{ds} \Rightarrow \frac{dp^v}{ds} = \frac{d}{ds} \left(mc \frac{dx^v}{ds} \right) = mc \frac{d}{ds} \left(\frac{dx^v}{ds} \right) = 0 \quad (50.93)$$

alors:

$$\boxed{\frac{d^2 x^v}{ds^2} = 0} \quad (50.94)$$

cette relation est donc la forme "sans masse" de l'équation du mouvement dans un espace euclidien ou autrement dit, dans un espace-temps de type Minkowski. Autrement dit, il existe donc un système de coordonnées en chute libre dans lequel le mouvement de la particule est celui d'un déplacement uniforme dans l'espace-temps.

Il sera très intéressant de la comparer avec l'équation du mouvement dans un espace courbe que nous verrons plus loin (appelée "équation des géodésiques").

Remarque: Il est équivalent d'écrire les relations des équations du mouvement par rapport à l'abscisse curviligne propre ds ou au temps propre dt (noté traditionnellement $d\tau$)

Nous pouvons maintenant montrer que l'équation du mouvement, au même titre que l'équation des géodésiques que nous verrons de suite après, est invariante par transformation de Lorentz:

$$L_v^\mu \frac{d^2 x^v}{ds^2} = \frac{d^2 (L_v^\mu x^v)}{ds^2} = \frac{d^2 x'^v}{ds^2} = 0 \quad (50.95)$$

Maintenant, voyons une forme plus générale de l'équation du mouvement pour tout type d'espace. L'objectif ici est de mettre en évidence, et ce en quelques lignes de calculs, que le mouvement suivi par une particule libre est indépendant de sa masse (vous pouvez déjà anticiper sur l'interprétation de la trajectoire d'un photon dans un espace courbe...!).

Nous avons démontré en calcul tensoriel (et précédemment) que:

$$ds = \left[g_{ij} \frac{\partial u^i}{\partial \alpha} \frac{\partial u^j}{\partial \alpha} \right]^{1/2} d\alpha \quad (50.96)$$

ce qui donne pour le lagrangien généralisé d'une particule libre avec $d\alpha = dt, u^i = x^\alpha, u^j = x^\beta$ (nous

retrouvons bien l'expression générale de l'énergie cinétique):

$$L = \frac{1}{2} m g_{\alpha\beta} \frac{\partial x^\alpha}{\partial t} \frac{\partial x^\beta}{\partial t} \quad (50.97)$$

où t est le temps propre de la particule, c'est un invariant !

Remarque: Cette relation est appelée "**lagrangien géodésique**" par certains auteurs.

Rappel: Le temps propre est une sorte d'horloge imaginaire qui voyage sur la particule et quels que soient les observateurs qui regardent l'horloge, ils seront mathématiquement d'accord sur la valeur de l'intervalle de temps entre deux "TIC" de l'horloge.

Ce qui nous permet d'écrire (attention il faut bien se rappeler des différentes relations que nous avons déterminées lors de notre étude du formalisme lagrangien dans le chapitre traitant de la Mécanique Analytique):

$$p_\alpha = \frac{\partial L}{\partial \dot{x}^\alpha} = m g_{\alpha\beta} \frac{dx^\beta}{dt} \quad (50.98)$$

Remarque: L'élimination du facteur 1/2 du Lagrangien provient de la symétrie du tenseur métrique. Si ce dernier n'est pas symétrique, nous pouvons toujours le caractériser par un tenseur qui l'est.

Effectivement, soit $\vec{x}$ un vecteur de coordonnées $x_1, \dots, x_n$ et soit:

$$q = A_{ij} x_i x_j \quad (50.99)$$

Les A_{ij} ne sont pas symétriques a priori, mais nous pouvons écrire:

$$q = \frac{1}{2} (A_{ij} + A_{ji}) x_i x_j \quad (50.100)$$

Nous posons ensuite:

$$B_{ij} = \frac{1}{2} (A_{ij} + A_{ji}) \quad (50.101)$$

Donc:

$$q = B_{ij} x_i x_j \quad (50.102)$$

et les B_{ij} sont symétriques.

La forme quadratique q peut donc toujours s'écrire avec une matrice symétrique, il y a même bijection. La conclusion étant qu'un tenseur métrique doit être symétrique si l'on veut le caractériser par la forme quadratique qu'il définit.

L'interlude mathématique étant terminé, continuons notre développement physique. En conséquence de la dernière relation, l'expression de l'hamiltonien devient bien évidemment:

$$H = p_\alpha \dot{x}^\alpha - L = L \quad (50.103)$$

puisque nous considérons être dans un espace sans champ de potentiel. Le carré de la vitesse étant dès lors constant sur toute la trajectoire, nous avons:

$$\frac{dH}{dt} = \frac{dL}{dt} = 0 \quad (50.104)$$

Établissons maintenant les équations du mouvement de tout corps. Nous avons:

$$\frac{\partial L}{\partial x^\nu} = \frac{1}{2} m \frac{\partial g_{\alpha\beta}}{\partial x^\nu} \dot{x}^\alpha \dot{x}^\beta \quad \text{et} \quad \frac{\partial L}{\partial \dot{x}^\nu} = m g_{\nu\beta} \dot{x}^\beta \quad (50.105)$$

et comme:

$$\frac{\partial L}{\partial x^\nu} = \frac{d}{dt} \frac{\partial L}{\partial \dot{x}^\nu} = \frac{d}{dt} \frac{\partial}{\partial \frac{d}{dt}(x^\nu)} \quad (50.106)$$

alors:

$$\frac{1}{2} m \frac{\partial g_{\alpha\beta}}{\partial x^\nu} \dot{x}^\alpha \dot{x}^\beta = \frac{d}{dt} (m g_{\nu\beta} \dot{x}^\beta) \quad (50.107)$$

d'où:

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{x}^\nu} \right) = m \frac{d}{dt} (g_{\nu\beta} \dot{x}^\beta) = m g_{\nu\beta} \ddot{x}^\beta + m \frac{d g_{\nu\beta}}{dt} \frac{\partial x^\alpha}{\partial x^\alpha} \dot{x}^\beta = m g_{\nu\beta} \ddot{x}^\beta + m \frac{\partial g_{\nu\beta}}{\partial x^\alpha} \dot{x}^\alpha \dot{x}^\beta \quad (50.108)$$

en mettant en commun:

$$m g_{\nu\beta} \ddot{x}^\beta + m \frac{\partial g_{\nu\beta}}{\partial x^\alpha} \dot{x}^\alpha \dot{x}^\beta = \frac{1}{2} m \frac{\partial g_{\alpha\beta}}{\partial x^\nu} \dot{x}^\alpha \dot{x}^\beta \quad (50.109)$$

que nous pouvons écrire identiquement pour les $\ddot{x}^\alpha$ en procédant de façon identique à ci-dessus.

La relation précédente donne donc la trajectoire d'un corps en mouvement, dans un espace sans champ de potentiel, en fonction de ses coordonnées curvilignes et de la métrique de l'espace considéré.

Ce qui est particulièrement intéressant dans ce résultat, c'est que la masse m (à nouveau) s'élimine identiquement dans cette équation du mouvement:

$$g_{\nu\beta} \ddot{x}^\beta + \left(\frac{\partial g_{\nu\beta}}{\partial x^\alpha} - \frac{1}{2} \frac{\partial g_{\alpha\beta}}{\partial x^\nu} \right) \dot{x}^\alpha \dot{x}^\beta = 0 \quad (50.110)$$

Remarquez, que nous aurions pu utiliser aussi un autre paramètre invariant que le temps propre tel que l'abscisse curviligne ds . Dès lors l'équation précédente s'écrirait:

$$\boxed{g_{\nu\beta} \frac{\partial^2 x^\beta}{\partial s^2} + \left(\frac{\partial g_{\nu\beta}}{\partial x^\alpha} - \frac{1}{2} \frac{\partial g_{\alpha\beta}}{\partial x^\nu} \right) \frac{\partial x^\alpha}{\partial s} \frac{\partial x^\beta}{\partial s} = 0} \quad (50.111)$$

Nous pouvons encore simplifier cette relation mais nous garderons cette simplification pour la deuxième démonstration de l'équation du mouvement dans un espace quelconque (en faisant usage du principe variationnel cette fois) juste après.

Il est très (très) intéressant d'observer que si nous restreignons la métrique à celle d'un espace euclidien:

$$\eta_{\nu\beta} \frac{\partial^2 x^\beta}{\partial s^2} + \left(\frac{\partial \eta_{\nu\beta}}{\partial x^\alpha} - \frac{1}{2} \frac{\partial \eta_{\alpha\beta}}{\partial x^\nu} \right) \frac{\partial x^\alpha}{\partial s} \frac{\partial x^\beta}{\partial s} = 0 \quad (50.112)$$

avec:

$$\frac{\partial \eta_{\nu\beta}}{\partial x^\alpha} = \frac{\partial \eta_{\alpha\beta}}{\partial x^\nu} = 0 \quad (50.113)$$

Nous obtenons alors la simplification:

$$\eta_{\nu\beta} \frac{\partial^2 x^\beta}{\partial s^2} = \frac{\partial^2 x_\nu}{\partial s^2} = 0 \quad (50.114)$$

Nous retrouvons donc la première équation du mouvement obtenue pour un espace plat! Le résultat est remarquable !

Conclusion: Aux mêmes conditions initiales de position et de vitesse curvilignes dans un espace (plat ou courbe) sans champ de potentiel (c'est ce que nous pourrions penser du moins selon nos hypothèses initiales...), correspond la même trajectoire quelle que soit la masse m de la particule (même pour les photons - la lumière - dont la masse est nulle!!).

Nous pouvons maintenant étudier le principe de moindre action dans le but de rechercher le plus court chemin (aussi bien au niveau spatial que temporel) entre deux points dans un espace de géométrie donnée avant de s'attaquer au cas beaucoup plus complexe du lagrangien qui prend en compte le tenseur des champs...

3.1. ÉQUATION DES GÉODÉSIQUES

Intéressons-nous maintenant à obtenir le même résultat mais en faisant usage cette fois-ci du principe variationnel. Nous retomberons sur la même équation que précédemment pour tout type d'espace à la différence que cette fois-ci, nous prendrons la peine de la simplifier pour arriver à "**l'équation des géodésiques**".

En partant de (voir développements précédents):

$$ds = \left(g_{ij} dx^i dx^j \right)^{1/2} \quad (50.115)$$

avec une paramétrisation telle que x^i et x^j sont fonction d'un paramètre temporel ou spatial.

Pour une surface donnée sous forme paramétrique, nous cherchons donc à minimiser la longueur d'un arc ds en appliquant donc le principe variationnel (non dépendant du temps car les photons ne peuvent avoir un chemin plus rapide au sens temporel du terme entre deux points mais uniquement un chemin plus court - au sens métrique du terme):

$$\delta S = \int_a^b \delta(ds) = \int_a^b \delta(ds) = 0 \quad (50.116)$$

en unités naturelles. Or:

$$\delta(ds^2) = 2ds\delta ds = \delta(g_{ij}dx^i dx^j) \quad (50.117)$$

En développant, et comme les indices ont le même domaine de variation:

$$\begin{aligned} \delta(g_{ij}dx^i dx^j) &= \left(\frac{\partial g_{ij}}{\partial x^k} \delta dx^k \right) dx^i dx^j + g_{ij} dx^i \delta dx^j + g_{ij} dx^i \delta dx^j \\ &= \left(\frac{\partial g_{ij}}{\partial x^k} \delta dx^k \right) dx^i dx^j + 2g_{ij} dx^i \delta dx^j \end{aligned} \quad (50.118)$$

d'où (nous avons déjà multiplié l'expression après la seconde égalité par ds/ds par anticipation de l'intégrale qui va suivre):

$$\delta(ds) = \frac{\delta(ds^2)}{2ds} = \frac{1}{2} \frac{dx^i}{ds} \frac{dx^j}{ds} \left(\frac{\partial g_{ij}}{\partial x^k} \delta dx^k \right) ds + g_{ij} \frac{dx^i}{ds} \frac{\delta dx^j}{ds} ds \quad (50.119)$$

Ensuite, il nous faut donc introduire ce développement sous l'intégrale:

$$\int_a^b \delta(ds) = \int_a^b \frac{1}{2} \frac{dx^i}{ds} \frac{dx^j}{ds} \left(\frac{\partial g_{ij}}{\partial x^k} \delta dx^k \right) ds + \int_a^b g_{ij} \frac{dx^i}{ds} \frac{\delta dx^j}{ds} ds = 0 \quad (50.120)$$

En travaillant sur la seconde intégrale (après l'égalité), nous posons:

$$u = g_{ij} \frac{dx^i}{ds} \quad \text{et} \quad v' = \frac{\delta dx^j}{ds} ds \quad (50.121)$$

Donc par l'intégration par partie (cf. chapitre de Calcul Différentiel Et Intégral):

$$\int_a^b g_{ij} \frac{dx^i}{ds} \frac{\delta dx^j}{ds} ds = \int_a^b uv' = uv \Big|_a^b - \int_a^b u'v \quad (50.122)$$

devient:

$$\int_a^b g_{ij} \frac{dx^i}{ds} \frac{\delta dx^j}{ds} ds = g_{ij} \frac{dx^i}{ds} \delta dx^j ds \Big|_a^b - \int_a^b \frac{d}{ds} \left(g_{ij} \frac{dx^i}{ds} \right) \delta dx^j ds \quad (50.123)$$

Soit finalement:

$$\int_a^b \delta(ds) = \int_a^b \frac{1}{2} \frac{dx^i}{ds} \frac{dx^j}{ds} \left(\frac{\partial g_{ij}}{\partial x^k} \delta dx^k \right) ds + g_{ij} \frac{dx^i}{ds} \delta dx^j ds \Big|_a^b - \int_a^b \frac{d}{ds} \left(g_{ij} \frac{dx^i}{ds} \right) \delta dx^j ds = 0 \quad (50.124)$$

Le terme non intégré ci-dessous est négligeable à cause de la présence du facteur δdx^j :

$$g_{ij} \frac{dx^i}{ds} \delta dx^j ds \Big|_a^b \cong 0 \quad (50.125)$$

Donc nous avons:

$$\begin{aligned} \int_a^b \delta(ds) &= \int_a^b \frac{1}{2} \frac{dx^i}{ds} \frac{dx^j}{ds} \left(\frac{\partial g_{ij}}{\partial x^k} \delta dx^k \right) ds - \int_a^b \frac{d}{ds} \left(g_{ij} \frac{dx^i}{ds} \right) \delta dx^j ds \\ &= \int_a^b \left[\frac{1}{2} \frac{dx^i}{ds} \frac{dx^j}{ds} \left(\frac{\partial g_{ij}}{\partial x^k} \delta dx^k \right) - \frac{d}{ds} \left(g_{ij} \frac{dx^i}{ds} \right) \delta dx^j \right] ds = 0 \end{aligned} \quad (50.126)$$

Nous effectuons un changement d'indice:

$$\int_a^b \delta ds = \int_a^b \left[\frac{1}{2} \frac{dx^i}{ds} \frac{dx^j}{ds} \left(\frac{\partial g_{ij}}{\partial x^k} \delta dx^k \right) - \frac{d}{ds} \left(g_{ik} \frac{dx^i}{ds} \right) \delta dx^k \right] ds = 0 \quad (50.127)$$

ce qui nous permet de factoriser δdx^k :

$$\int_a^b \delta ds = \int_a^b \left[\frac{1}{2} \frac{dx^i}{ds} \frac{dx^j}{ds} \frac{\partial g_{ij}}{\partial x^k} - \frac{d}{ds} \left(g_{ik} \frac{dx^i}{ds} \right) \right] \delta dx^k ds = 0 \quad (50.128)$$

Comme δdx^k et ds sont différents de zéro, c'est l'intégrande qui doit être nulle:

$$\frac{1}{2} \frac{dx^i}{ds} \frac{dx^j}{ds} \frac{\partial g_{ij}}{\partial x^k} - \frac{d}{ds} \left(g_{ik} \frac{dx^i}{ds} \right) = 0 \quad (50.129)$$

En développant le second terme:

$$\frac{1}{2} \frac{dx^i}{ds} \frac{dx^j}{ds} \frac{\partial g_{ij}}{\partial x^k} - \frac{dg_{ik}}{ds} \frac{dx^i}{ds} - g_{ik} \frac{d}{ds} \frac{dx^i}{ds} = 0 \quad (50.130)$$

Qui s'écrit encore:

$$\frac{1}{2} \frac{dx^i}{ds} \frac{dx^j}{ds} \frac{\partial g_{ij}}{\partial x^k} - \frac{dx^i}{ds} \frac{\partial g_{ik}}{\partial x^j} \frac{dx^j}{ds} - g_{ik} \frac{d}{ds} \frac{dx^i}{ds} = 0 \quad (50.131)$$

et qui se simplifie en:

$$g_{ik} \frac{d^2 x^i}{ds^2} + \left(\frac{\partial g_{ik}}{\partial x^j} - \frac{1}{2} \frac{\partial g_{ij}}{\partial x^k} \right) \frac{dx^i}{ds} \frac{dx^j}{ds} = 0 \quad (50.132)$$

Nous obtenons (à nouveau!!!) le système d'équations qui définit les "géodésiques", c'est-à-dire les droites de $\mathcal{E}_\kappa$. Ces dernières constituent donc les extrémales de l'intégrale qui mesure la longueur d'un arc de courbe joignant deux points donnés dans $\mathcal{E}_\kappa$.

Cette dernière équation, est celle qui nous intéresse dans le cas du lagrangien libre. Effectivement, si nous prenons le cas extrême de la lumière (ou des photons si vous préférez), cette dernière ne va pas chercher le chemin le plus rapide (le plus vite) au niveau temporel. Ce serait totalement en contradiction avec le postulat d'invariance de voir la lumière accélérer en fonction du chemin!!! Dans ce contexte, cela signifie que sur la trame spatio-temporelle, la seule chose qui a un sens est le plus court chemin spatial et non le plus court chemin temporel! C'est la raison pour laquelle cette dernière équation est appelée "équation des géodésiques" ou encore "équation d'Euler-Lagrange généralisée".

Cependant, nous pouvons écrire cette dernière équation de façon plus condensée en introduisant les symboles de Christoffel si la métrique est un tenseur symétrique tel que $g_{\alpha\beta} = g_{\beta\alpha}$. Effectivement:

$$g_{ik} \frac{d^2 x^i}{ds^2} + \frac{1}{2} \left(2 \frac{\partial g_{ik}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^k} \right) \frac{dx^i}{ds} \frac{dx^j}{ds} = g_{ik} \frac{d^2 x^i}{ds^2} + \frac{1}{2} \left(\frac{\partial g_{ik}}{\partial x^j} + \frac{\partial g_{jk}}{\partial x^i} - \frac{\partial g_{ij}}{\partial x^k} \right) \frac{dx^i}{ds} \frac{dx^j}{ds} = 0 \quad (50.133)$$

et comme le symbole de Christoffel de première espèce (cf. chapitre de Calcul Tensoriel) est défini par:

$$\Gamma_{ijk} = \frac{1}{2} \left(\partial_j g_{ik} + \partial_i g_{jk} - \partial_k g_{ij} \right) \quad (50.134)$$

Remarque: Il est important de se rappeler que ce symbole contient toute l'information sur la métrique de l'espace-temps. Nous verrons un exemple plus bas comme quoi dans un référentiel localement inertiel ce symbole de Christoffel est nul.

Alors l'équation d'Euler-Lagrange s'écrit:

$$g_{ik} \frac{d^2 x^i}{ds^2} + \Gamma_{ijk} \frac{dx^i}{ds} \frac{dx^j}{ds} = 0 \quad (50.135)$$

La multiplication contractée (cf. chapitre de Calcul Tensoriel) de la relation précédente dans la base canonique par g^{kl} nous donne:

$$\begin{aligned} g^{kl} g_{ik} \frac{d^2 x^i}{ds^2} + g^{kl} \Gamma_{ijk} \frac{dx^i}{ds} \frac{dx^j}{ds} &= \delta_{il} \frac{d^2 x^i}{ds^2} + \Gamma^l_{ij} \frac{dx^i}{ds} \frac{dx^j}{ds} \\ \frac{d^2 x^l}{ds^2} + \Gamma^l_{ij} \frac{dx^i}{ds} \frac{dx^j}{ds} &= 0 \end{aligned} \quad (50.136)$$

dans la littérature un changement d'indice est souvent effectué afin d'avoir au final (c'est toujours la même expression étant donné que les indices ont le même domaine de variation!):

$$\frac{d^2 x^\mu}{ds^2} + \Gamma_{\alpha\beta}^\mu \frac{dx^\alpha}{ds} \frac{dx^\beta}{ds} = 0 \quad (50.137)$$

avec $\Gamma_{\alpha\beta}^\mu$ étant donc le symbole de Christoffel de deuxième espèce (cf. chapitre de Calcul Tensoriel) donné par:

$$\Gamma_{\alpha\beta}^\mu = g^{\mu\lambda} \Gamma_{\alpha\lambda\beta} = \frac{1}{2} g^{\mu\lambda} (\partial_\alpha g_{\lambda\beta} + \partial_\beta g_{\alpha\lambda} - \partial_\lambda g_{\alpha\beta}) \quad (50.138)$$

et est appelé dans le cadre de la relativité générale la "**connexion affine**" ou encore "**coefficients de connexion**" et qui permet de trouver le système de coordonnées (via la résolution d'un système d'équations différentielles) en chute libre dans lequel l'équation de la particule est celle d'un déplacement uniforme dans l'espace-temps en fonction d'un système de référence (les deux systèmes étant donc reliés par la connexion affine!).

Cette relation, de la plus haute importance, nous permet de déterminer comment un corps en mouvement va naturellement se déplacer dans un espace courbe et ce peut-être... indépendamment de sa masse !!! Elle nous donne donc la métrique dans laquelle nous devons poser un référentiel pour qu'il soit inertiel par rapport au corps considéré.

L'équation des géodésiques antéprécédente est aussi l'équation différentielle du second ordre que doit donc satisfaire la représentation paramétrique d'une ligne sur une surface où s est la longueur le long de la ligne afin que sa longueur totale soit extrémale.

Selon le principe d'équivalence, nous sommes donc en droit d'interpréter cette relation comme l'équation du mouvement dans un champ de gravitation quelconque, et donc d'interpréter le deuxième terme supplémentaire de l'équation comme l'opposé d'un terme de force gravitationnelle par unité de masse, c'est-à-dire comme l'opposé d'un champ gravitationnel.

Remarque: Nous pouvons également écrire l'équation des géodésiques et utilisant le temps propre:

$$\frac{d^2 x^\mu}{\left(\frac{dt \cdot c}{\gamma}\right)^2} + \Gamma_{\alpha\beta}^\mu \left(\frac{dx^\alpha}{\frac{dt \cdot c}{\gamma}}\right) \left(\frac{dx^\beta}{\frac{dt \cdot c}{\gamma}}\right) = \frac{d^2 x^\mu}{dt^2} + \Gamma_{\alpha\beta}^\mu \frac{dx^\alpha}{dt} \frac{dx^\beta}{dt} = 0 \quad (50.139)$$

ou encore en utilisant la quadrivitesse:

$$\frac{du^\mu}{dt} + \Gamma_{\alpha\beta}^\mu du^\alpha du^\beta = 0 \quad (50.140)$$

Encore une fois, si nous nous restreignons à un espace-temps plat, nous voyons trivialement que nous retombons sur la première équation du mouvement que nous avons obtenue:

$$\frac{d^2 x^\mu}{ds^2} = \frac{d^2 x^\mu}{\left(\frac{dt \cdot c}{\gamma}\right)^2} = \frac{d^2 x^\mu}{dt^2} = 0 \quad (50.141)$$

car les composantes de la métrique de Minkowski étant constantes les coefficients de Christoffel sont tous nuls.

Les solutions de cette dernière équation sont des lignes droites ordinaires données par:

$$x^\mu(s) = a^\mu s - b^\mu \quad (50.142)$$

Bien évidemment, dans un espace-temps courbe général, les géodésiques ne pourront pas être globalement représentées par des lignes droites. Cependant avec une approximation au deuxième ordre en développement de Taylor (cf. [chapitre sur les Suites Et Séries](#)) nous arrivons à nous ramener à des droites (ce qui revient à ramener l'espace courbe à un espace plat).

L'important dans tout cela, c'est que l'équation des géodésiques permet de constater que la courbure de l'espace détermine les trajectoires des corps qui s'y meuvent quelle que soit leur masse, qu'ils soient en mouvement uniforme ou non (observez la dérivée seconde dans l'équation des géodésiques!). Il ne nous reste plus alors qu'à effectuer la fin du travail et de mettre en relation la courbure de l'espace-temps avec l'énergie qui s'y trouve !

3.2. LIMITE NEWTONIENNE

Nous avons montré plus haut (argument de Shild) que pour étudier la gravitation (en particulier l'effet Einstein), la géométrie courbe est nécessaire. Nous avons promis de montrer aussi qu'elle était suffisante. Il est temps maintenant de le faire !

Définition: La "**limite Newtonienne**" est une situation physique où les trois conditions ci-dessous sont satisfaites:

C1. Les particules se déplacent lentement par rapport à la vitesse de la lumière. Ce qui s'exprime comme le fait que les variations des composantes spatiales de leur quadrivecteur sont très inférieures à celles de la composante temporelle (t étant le temps propre):

$$\frac{dx^i}{dt} \ll \frac{dx^0}{dt} \quad (50.143)$$

C2. Le champ de gravitation est statique. En d'autres termes, toute dérivée temporelle de la métrique est nulle.

C3. Le champ gravitationnel est faible, c'est-à-dire qu'il peut être vu comme une faible perturbation d'un espace plat:

$$g_{\mu\nu} \cong \eta_{\mu\nu} + h_{\mu\nu} \text{ avec } |h_{\mu\nu}| \ll 1 \quad (50.144)$$

et où $\eta_{\mu\nu}$ est constant (seul $h_{\mu\nu}$ dépend des coordonnées).

Considérons l'équation des géodésiques obtenue précédemment:

$$\frac{d^2 x^\mu}{dt^2} + \Gamma_{\alpha\beta}^\mu \frac{dx^\alpha}{dt} \frac{dx^\beta}{dt} = 0 \quad (50.145)$$

La première condition (C1) nous amène à la simplifier sous la forme:

$$\frac{d^2 x^\mu}{dt^2} + \Gamma_{00}^\mu \left(\frac{dx^0}{dt} \right)^2 = 0 \quad (50.146)$$

Les deux autres conditions (C2 et C3 dont l'application a été mise en évidence dans le développement ci-dessous) nous offrent plusieurs simplifications dans l'expression du symbole de Christoffel de deuxième espèce:

$$\begin{aligned} \Gamma_{00}^\mu &= \frac{1}{2} g^{\mu\lambda} (\partial_0 g_{\lambda 0} + \partial_0 g_{0\lambda} - \partial_\lambda g_{00}) \\ &= -\frac{1}{2} g^{\mu\lambda} \partial_\lambda g_{00} \quad (C2) \\ &\cong -\frac{1}{2} (\eta^{\mu\lambda} + h^{\mu\lambda}) \left(\frac{\partial_\lambda \eta_{00}}{\substack{=0(C3)}} + \partial_\lambda h_{00} \right) = -\frac{1}{2} (\eta^{\mu\lambda} + h^{\mu\lambda}) \partial_\lambda h_{00} \cong -\frac{1}{2} \eta^{\mu\lambda} \partial_\lambda h_{00} \end{aligned} \quad (50.147)$$

L'équation des géodésiques devient alors:

$$\frac{d^2 x^\mu}{dt^2} - \frac{1}{2} \eta^{\mu\lambda} \partial_\lambda h_{00} \left(\frac{dx^0}{dt} \right)^2 = 0 \quad (50.148)$$

et vaut alors pour la composante temporelle ($\mu = 0$):

$$\frac{d^2 x^0}{dt^2} - \frac{1}{2} \eta^{0\lambda} \partial_\lambda h_{00} \left(\frac{dx^0}{dt} \right)^2 = 0 \quad (50.149)$$

Or (rappel de la métrique de Minkowski) $\eta^{0\lambda} = 0$ pour $\lambda > 0$ et pour $\lambda = 0$ nous avons (métrique statique) $\partial_0 h_{00} = 0$. Donc obligatoirement, nous devons conclure que dx^0/dt est une constante (et ce quel que soit le choix de la signature de la métrique de Minkowski).

Quant aux composantes spatiales, nous savons que $\eta^{\mu\nu}$ lorsque réduit à sa partie spatiale est une simple matrice identité 3×3 , ce qui donne pour chaque composante spatiale dans le cas où nous choisissons (par tradition uniquement!) la signature - + + de la métrique de Minkowski:

$$\frac{d^2 x^i}{dt^2} - \frac{1}{2} \partial_i h_{00} \left(\frac{dx^0}{dt} \right)^2 = 0 \quad (50.150)$$

Bien évidemment, le lecteur peut s'amuser à faire le développement qui va suivre avec la signature inverse (+ - -) et il verra que cela change seulement le signe du potentiel dans le résultat final du développement):

Notons maintenant le temps propre $d\tau$ comme il est d'usage. Nous avons alors:

$$\frac{d^2 x^i}{d\tau^2} = \frac{1}{2} \partial_i h_{00} \left(\frac{dx^0}{d\tau} \right)^2 \quad (50.151)$$

En divisant par $(dx^0/d\tau)^2$ et en rétablissant $x^0 = ct$, nous obtenons:

$$\begin{aligned}\frac{d^2 x^i}{d\tau^2} \left(\frac{d\tau}{dx^0} \right)^2 &= \frac{1}{2} \partial_i h_{00} \\ \frac{d^2 x^i}{(dx^0)^2} &= \frac{1}{2} \partial_i h_{00} \\ \frac{d^2 x^i}{(dct)^2} &= \frac{1}{2} \partial_i h_{00} \\ \frac{d^2 x^i}{dt^2} &= \frac{c^2}{2} \partial_i h_{00}\end{aligned}\quad (50.152)$$

À partir d'ici nous posons (car nos illustres prédécesseurs ont tâtonné avant nous):

$$h_{00} = -\frac{2}{c^2} \varphi = -\frac{2GM}{rc^2} \quad (50.153)$$

tel que (relation qui nous sera très utile lors de l'étude de la métrique de Schwarzschild plus loin):

$$g_{00} = \eta_{00} + h_{00} = 1 - \frac{2}{c^2} \varphi = 1 - \frac{2GM}{rc^2} = \left(1 - \frac{2GM}{rc^2} \right) \quad (50.154)$$

où φ est le potentiel gravitationnel, nous retrouvons l'expression de l'accélération gravitationnelle (équation de Newton-Poisson) de la mécanique Newtonienne (cf. chapitre de Mécanique Classique):

$$\vec{a} = \frac{d^2 x^i}{dt^2} = -\frac{c^2}{2} \partial_i \left(\frac{2\varphi}{c^2} \right) = -\partial_i \varphi = -\vec{\nabla} \cdot \varphi \quad (50.155)$$

avec $i = 1, 2, 3$.

Ce développement, simple mais néanmoins remarquable par son interprétation, prouve que la géométrie courbe est suffisante pour décrire la gravitation (et donc la théorie de Newton)!! Ce vérification est nommée par certains le "**principe de correspondance**".

4. TENSEUR D'ÉNERGIE-IMPULSION

Le tenseur d'énergie-impulsion (T.E.I.) est un outil mathématique utilisé (notamment) en relativité générale afin de représenter la répartition de masse et d'énergie dans l'espace-temps.

Prenons pour exemple le T.E.I. qui considère en relativité générale la matière comme pouvant être approximée par un fluide parfait. Dans le chapitre de Mécanique Des Milieux Continus nous avons démontré:

$$N_i = \sigma_{ij} n_j \quad (50.156)$$

où N_i a les unités d'une force et n_j celles d'une surface. Ainsi:

$$F_i = \sigma_{ij} n_j \quad (50.157)$$

sous forme variationnelle cela donne:

$$\sigma_{ij} = \left. \frac{\Delta F_i}{\Delta n_j} \right|_{\Delta a_j \rightarrow 0} \quad (50.158)$$

Calculons maintenant:

$$\sigma_{ij} = \frac{\Delta F_i}{\Delta n_j} = \frac{\Delta(ma_i)}{dn_j} = \frac{\Delta(\rho V a_i)}{\Delta n_j} \quad (50.159)$$

Remarque: Nous ne travaillons exprès pas avec des éléments différentiels afin de ne pas être coincé plus tard. C'est un peu du bricolage à la physicienne, mais bon cela marche (confirmé par l'expérience).

En supposant que seuls le volume et le temps font que la force varie (ce qui suppose une densité constante quand même et que le système est inertiel) nous avons alors:

$$\begin{aligned} \sigma_{ij} &= \frac{\Delta(\rho V a_i)}{\Delta n_j} = \frac{1}{\Delta n_j} \Delta\left(\frac{V}{t}\right) \rho v_i = \frac{1}{\Delta n_j} \frac{\Delta V}{\Delta t} \rho v_i = \frac{\Delta x_i \Delta x_j \Delta x_k}{\Delta n_j} \rho \frac{v_i}{\Delta t} \\ &= \frac{\Delta x_i \Delta x_j \Delta x_k}{\Delta x_i \Delta x_k} \rho \frac{v_i}{\Delta t} = \frac{\Delta x_j}{\Delta t} \rho v_i = v_j \rho v_i = \rho v_j v_i \end{aligned} \quad (50.160)$$

Ce qui donne simplement (cf. chapitre de Calcul Tensoriel) le produit tensoriel des vitesses:

$$\vec{\sigma} = \rho \vec{v} \otimes \vec{v} \quad (50.161)$$

Si nous généralisons cette relation aux quadrivecteurs-vitesse de la relativité restreinte, nous avons alors par définition le "tenseur d'énergie-impulsion":

$$\vec{T} = \rho \vec{u} \otimes \vec{u} \quad (50.162)$$

ou sous forme indicielle:

$$T_{ij} = \rho u_i u_j \quad (50.163)$$

soit sous forme contravariante:

$$T^{ij} = \rho u^i u^j \quad (50.164)$$

Cette relation est la justification pour laquelle la relativité générale est aussi indiquée comme étant une théorie des milieux continus par certains spécialistes.

Maintenant démontrons que la dérivée:

$$\partial_j T^{0j} = \partial_j (\rho u^0 u^j) = 0 \quad (50.165)$$

Remarque: Ce qui comme nous l'avons déjà signalé dans le chapitre de Calcul Tensoriel s'écrit $T^{0j}_{,j} = 0$ dans les vieux livres.

D'abord, rappelons que (cf. chapitre de Relativité Restreinte):

$$u^\nu = (u^0, u^1, u^2, u^3) = \gamma(\vec{v})(c, v^1, v^2, v^3) = \gamma(\vec{v})(c, v^x, v^y, v^z) \quad (50.166)$$

et admettons que nous sommes dans les faibles vitesses telles que $\gamma = 1$. Dès lors dans une métrique de Minkowski (+, -, -, -):

$$\begin{aligned} \partial_j T^{0j} &= \partial_j (\rho u^0 u^j) = \partial_j (\rho c u^j) \\ &= \partial_0 (\rho c u^0) + \partial_1 (\rho c u^1) + \partial_2 (\rho c u^2) + \partial_3 (\rho c u^3) \\ &= \frac{\partial(\rho c c)}{\partial c t} - \frac{\partial(\rho c v^x)}{\partial x} - \frac{\partial(\rho c v^y)}{\partial y} - \frac{\partial(\rho c v^z)}{\partial z} \\ &= c \left(\frac{\partial \rho}{\partial t} - \frac{\partial(\rho v^x)}{\partial x} - \frac{\partial(\rho v^y)}{\partial y} - \frac{\partial(\rho v^z)}{\partial z} \right) \\ &= c \left(\frac{\partial \rho}{\partial t} - \vec{\nabla} \cdot (\rho \cdot \vec{v}) \right) \end{aligned} \quad (50.167)$$

Or, nous reconnaissons dans les parenthèses l'équation de continuité (conservation de la masse) que nous avons démontré dans le chapitre de Thermodynamique et qui nous le savons est nulle! Ainsi:

$$\partial_j T^{0j} = c \left(\frac{\partial \rho}{\partial t} - \vec{\nabla} \cdot (\rho \cdot \vec{v}) \right) = c \cdot 0 = 0 \quad (50.168)$$

Regardons par ailleurs ce que contient la composante T^{00} du T.E.I.:

$$T^{00} = \rho u^0 u^0 = \rho \gamma c \gamma c = c^2 \gamma^2 \rho \quad (50.169)$$

En termes d'unités, il s'agit d'une densité d'énergie (nous voyons directement que cette grandeur ne peut être que positive).

Regardons maintenant les autres composantes avec $i = 0$ et $j = 1, 2, 3$:

$$T^{0j} = \rho u^0 u^j = \rho (\gamma c) (\gamma v^j) = c \rho \gamma^2 v^j = c p^j \quad (50.170)$$

où p^j a les unités d'une densité de quantité de mouvement.

Regardons maintenant les composantes du tenseur lorsque $i, j = 1 \dots 3$ (nous omettons donc la première ligne et la première colonne):

$$T^{ij} = \rho(\gamma u^i)(\gamma u^j) = \rho \gamma^2 u^i u^j = \sigma_{ij} \quad (50.171)$$

Nous retrouvons donc les composantes du tenseur des contraintes d'un fluide parfait.

Donc finalement, le T.E.I. peut s'écrire sous la forme d'une matrice 4×4 réelle symétrique:

$$T^{ij} = \begin{bmatrix} T^{00} & T^{01} & T^{02} & T^{03} \\ T^{10} & T^{11} & T^{12} & T^{13} \\ T^{20} & T^{21} & T^{22} & T^{23} \\ T^{30} & T^{31} & T^{32} & T^{33} \end{bmatrix} = \begin{bmatrix} \rho \gamma^2 c^2 & \rho \gamma^2 c v^x & \rho \gamma^2 c v^y & \rho \gamma^2 c v^z \\ \rho \gamma^2 c v^x & \rho \gamma^2 v^x v^x & \rho \gamma^2 v^x v^y & \rho \gamma^2 v^x v^z \\ \rho \gamma^2 c v^y & \rho \gamma^2 v^y v^x & \rho \gamma^2 v^y v^y & \rho \gamma^2 v^y v^z \\ \rho \gamma^2 c v^z & \rho \gamma^2 v^z v^x & \rho \gamma^2 v^z v^y & \rho \gamma^2 v^z v^z \end{bmatrix} \quad (50.172)$$

Dans le cas où les vitesses sont faibles, c'est-à-dire $\gamma \rightarrow 1$, nous avons:

$$T^{ij} = \begin{bmatrix} \rho c^2 & \rho c v^x & \rho c v^y & \rho c v^z \\ \rho c v^x & \rho v^x v^x & \rho v^x v^y & \rho v^x v^z \\ \rho c v^y & \rho v^y v^x & \rho v^y v^y & \rho v^y v^z \\ \rho c v^z & \rho v^z v^x & \rho v^z v^y & \rho v^z v^z \end{bmatrix} = \begin{bmatrix} \frac{mc^2}{V} & \rho c v^x & \rho c v^y & \rho c v^z \\ \rho c v^x & \sigma_{xx} & \tau^{xy} & \tau^{xz} \\ \rho c v^y & \tau^{yx} & \sigma_{yy} & \tau^{yz} \\ \rho c v^z & \tau^{zx} & \tau^{zy} & \sigma_{zz} \end{bmatrix} \quad (50.173)$$

$$= \begin{bmatrix} E_{tot}/V & \rho c v^x & \rho c v^y & \rho c v^z \\ \rho c v^x & \sigma_{xx} & \tau^{xy} & \tau^{xz} \\ \rho c v^y & \tau^{yx} & \sigma_{yy} & \tau^{yz} \\ \rho c v^z & \tau^{zx} & \tau^{zy} & \sigma_{zz} \end{bmatrix}$$

Nous retrouvons donc dans ce tenseur les interprétations suivantes des grandeurs physiques (bien que rigoureusement toutes les composantes aient des unités qui peuvent être vues comme densité d'énergie soit comme une pression).

- T^{00} est la densité volumique d'énergie (elle est positive)
- T^{10}, T^{20}, T^{30} sont les densités de moments
- T^{01}, T^{02}, T^{03} sont les flux d'énergie

Remarque: La sous-matrice des composantes spatiales:

$$T^{ij} = \begin{bmatrix} T^{11} & T^{12} & T^{13} \\ T^{21} & T^{22} & T^{23} \\ T^{31} & T^{32} & T^{33} \end{bmatrix} \quad (50.174)$$

est la matrice dite "**matrice des flux de moments**" (appellation tout à fait discutable...). En mécanique des milieux continus (cf. chapitre de Mécanique Des Milieux Continus), nous avons

démontré que sa diagonale correspond à la pression, et les autres composantes aux efforts tangentiels dus à la viscosité dynamique.

Bref, pour résumer sous forme covariante:

$$\begin{array}{c}
 \begin{array}{cc}
 \text{densité} & \text{flux d'énergie} \\
 \text{d'énergie} &
 \end{array} \\
 \begin{array}{c}
 \begin{array}{cccc}
 T_{00} & T_{01} & T_{02} & T_{03} \\
 T_{10} & T_{11} & T_{12} & T_{13} \\
 T_{20} & T_{21} & T_{22} & T_{23} \\
 T_{30} & T_{31} & T_{32} & T_{33}
 \end{array} \\
 \begin{array}{cc}
 \text{densité de} & \text{flux de} \\
 \text{moment} & \text{moment}
 \end{array}
 \end{array}
 \end{array}
 \quad \begin{array}{l}
 \text{viscosité} \\
 \text{pression}
 \end{array}
 \quad (50.175)$$

Montrons que la dérivée covariante (cf. chapitre de Calcul Tensoriel) du tenseur d'énergie-impulsion est nulle telle que:

$$\nabla_{\mu} T^{\mu\nu} = 0 \quad (50.176)$$

Donc:

$$\nabla_{\mu} T^{\mu\nu} = \nabla_{\mu} (\rho u^{\mu} u^{\nu}) = u^{\nu} \nabla_{\mu} (\rho u^{\mu}) + \rho u^{\mu} \nabla_{\mu} (u^{\nu}) \quad (50.177)$$

Commençons par développer le premier terme:

$$u^{\nu} \nabla_{\mu} (\rho u^{\mu}) = u^{\nu} \left(\frac{\partial(\rho \gamma c)}{\partial ct} - \frac{\partial(\rho \gamma u^x)}{\partial x} - \frac{\partial(\rho \gamma u^y)}{\partial y} - \frac{\partial(\rho \gamma u^z)}{\partial z} \right) \quad (50.178)$$

Or, nous avons:

$$\frac{\partial \gamma}{\partial x} = \frac{\partial \gamma}{\partial y} = \frac{\partial \gamma}{\partial z} = 0 \quad (50.179)$$

d'où:

$$u^{\nu} \nabla_{\mu} (\rho u^{\mu}) = u^{\nu} \left(\frac{\partial(\rho \gamma)}{\partial t} - \gamma \vec{\nabla} \circ (\rho \vec{u}) \right) = u^{\nu} \left(\rho \frac{\partial \gamma}{\partial t} + \gamma \left[\frac{\partial(\rho)}{\partial t} - \vec{\nabla} \circ (\rho \vec{u}) \right] \right) \quad (50.180)$$

Nous retrouvons entre les crochets l'équation de continuité qui est nulle. Par contre, le premier terme entre parenthèses est non nul comme nous l'avons vu lors de notre étude du quadrivecteur accélération dans le chapitre de Relativité Restreinte:

$$u^\nu \nabla_\mu (\rho u^\mu) = u^\nu \rho \frac{\partial \gamma}{\partial t} = u^\nu \rho \left(\frac{\gamma^3}{c^2} \vec{v} \circ \vec{a} \right) \quad (50.181)$$

Mais selon le principe d'équivalence faible (PEF), nous pouvons toujours nous placer dans un référentiel tel que localement l'accélération soit nulle tel que (pour rappel, on ne met pas de flèches de vecteur pour les quadrivecteurs):

$$a = 0 \quad (50.182)$$

et il vient alors:

$$u^\nu \nabla_\mu (\rho u^\mu) = 0 \quad (50.183)$$

Donc nous avons maintenant:

$$\nabla_\mu T^{\mu\nu} = u^\nu \nabla_\mu (\rho u^\mu) + \rho u^\mu \nabla_\mu (u^\nu) = 0 + \rho u^\mu \nabla_\mu (u^\nu) = \rho u^\mu \nabla_\mu (u^\nu) \quad (50.184)$$

Regardons ce que donne ce dernier terme mais en rappelant d'abord que dans le chapitre de Relativité Restreinte nous avons démontré que la quadri-accélération s'exprimait selon:

$$a^\nu = \left(\frac{\gamma^4}{c} \vec{v} \circ \vec{a}, \gamma^2 \vec{a} + \frac{\gamma^4}{c^2} (\vec{v} \circ \vec{a}) \vec{v} \right) \quad (50.185)$$

Soit (nous ne prenons que les deux premières composantes comme exemples):

$$\begin{aligned} a^0 &= \frac{\gamma^4}{c} (v^x a^x + v^y a^y + v^z a^z) \\ a^1 &= \gamma^2 a^x + \frac{\gamma^4}{c^2} (v^x a^x + v^y a^y + v^z a^z) v^x \\ &\dots \end{aligned} \quad (50.186)$$

Nous allons maintenant au fait montrer que:

$$u^\mu \nabla_\mu (u^\nu) = a^\nu \quad (50.187)$$

Commençons par montrer que $u^\mu \nabla_\mu (u^0) = a^0$:

$$u^\mu \nabla_\mu (u^0) = \gamma c \frac{\partial(\gamma c)}{\partial(ct)} + \gamma v^x \frac{\partial(\gamma c)}{\partial x} + \gamma v^y \frac{\partial(\gamma c)}{\partial y} + \gamma v^z \frac{\partial(\gamma c)}{\partial z} \quad (50.188)$$

Or:

$$\frac{\partial \gamma}{\partial x} = \frac{\partial \gamma}{\partial y} = \frac{\partial \gamma}{\partial z} = 0 \quad \text{et} \quad \frac{\partial c}{\partial x} = \frac{\partial c}{\partial y} = \frac{\partial c}{\partial z} = 0 \quad (50.189)$$

d'où:

$$u^\mu \nabla_\mu (u^0) = \gamma c \frac{\partial(\gamma c)}{\partial(ct)} = \gamma c \frac{\partial \gamma}{\partial t} = \gamma c \left(\frac{\gamma^3}{c^2} \vec{v} \circ \vec{a} \right) = \frac{\gamma^4}{c} \vec{v} \circ \vec{a} = a^0 \quad (50.190)$$

Maintenant montrons que $u^\mu \nabla_\mu (u^1) = a^1$ (les autres composantes se vérifiant alors automatiquement):

$$\begin{aligned} u^\mu \nabla_\mu (u^1) &= \gamma c \frac{\partial(\gamma v^x)}{\partial(ct)} + \underbrace{\gamma v^x \frac{\partial(\gamma v^x)}{\partial x} + \gamma v^y \frac{\partial(\gamma v^x)}{\partial y} + \gamma v^z \frac{\partial(\gamma v^x)}{\partial z}}_{=0} \\ &= \gamma \frac{\partial(\gamma v^x)}{\partial t} = \gamma \left(v^x \frac{\partial \gamma}{\partial t} + \gamma \frac{\partial v^x}{\partial t} \right) = \gamma \left(v^x \left(\frac{\gamma^3}{c^2} \vec{v} \circ \vec{a} \right) + \gamma a^x \right) \\ &= \gamma^2 a^x + \left(\frac{\gamma^4}{c^2} \vec{v} \circ \vec{a} \right) v^x = a^1 \end{aligned} \quad (50.191)$$

et donc nous avons bien:

$$u^\mu \nabla_\mu (u^\nu) = a^\nu \quad (50.192)$$

mais selon le PEF $a^\nu = 0$ alors:

$$u^\mu \nabla_\mu (u^\nu) = 0 \quad (50.193)$$

et nous avons donc bien finalement:

$$\boxed{\nabla_\mu T^{\mu\nu} = 0} \quad (50.194)$$

qui est l'expression de la conservation de l'énergie en relativité générale! En abaissant les indices, il vient:

$$\nabla^\rho T_{\rho\sigma} = g^{\rho\mu} \nabla_\rho (g_{\sigma\mu} g_{\alpha\nu} T^{\alpha\nu}) = \nabla_\mu T^{\mu\nu} = 0 \quad (50.195)$$

5. ÉQUATION D'EINSTEIN DES CHAMPS

Il est temps maintenant de nous attaquer au plus beau, à l'une des équations les plus fameuses de notre époque et qui fait briller les yeux de beaucoup de jeunes étudiants: l'équation d'Einstein des champs. Celle qui explique pourquoi la matière (l'énergie) courbe l'espace. Il existe plusieurs manières de l'obtenir. Les deux plus courantes consistent soit:

1. À avoir une approche "à l'ingénieur": C'est-à-dire que nous procédons par comparaison avec un résultat limite connu qui est la loi de gravitation de Newton (c'est celle que nous avons choisie)
2. À avoir une approche "matheuse" (très élégante mais un peu tombée du ciel): C'est-à-dire que nous utilisons le formalisme lagrangien et cherchons par tâtonnements une densité lagrangienne qui nous permet de retomber sur quelque chose de connu.

Bon ceci ayant été dit, rappelons avant de commencer quelques résultats que nous avons obtenus

jusqu'ici. Premièrement, nous avons réussi à démontrer avec brio que toute particule (supposée libre mais cela est laissé à l'interprétation... dans un espace courbe...) suit l'équation du mouvement des géodésiques:

$$\frac{d^2 x^\mu}{ds^2} + \Gamma_{\alpha\beta}^\mu \frac{dx^\alpha}{ds} \frac{dx^\beta}{ds} = 0 \quad (50.196)$$

Dans le chapitre de Calcul Tensoriel, nous avons démontré (non sans peine) que ce que nous appelons le "**tenseur d'Einstein**" (qui est une constante dans un espace Riemannien donné) est donné par:

$$G^{\mu\nu} = R^{\mu\nu} - \frac{1}{2} g^{\mu\nu} R \quad (50.197)$$

où $R^{\mu\nu}$ est le tenseur de Ricci (cf. chapitre de Calcul Tensoriel).

Puisque la dérivée covariante du tenseur d'Einstein est nulle et que nous avons démontré que la dérivée covariante de T.E.I. l'est aussi, il est tentant de poser:

$$\nabla^\rho G_{\rho\sigma} = \kappa \nabla^\rho T_{\rho\sigma} = 0 \quad (50.198)$$

où κ est une constante de normalisation et devant satisfaire la relation pour qu'elle soit homogène au niveau des unités. Ainsi, il vient:

$$G_{\rho\sigma} = \kappa T_{\rho\sigma} \quad (50.199)$$

Pour trouver l'expression de la constante, nous allons nous placer en limite Newtonienne et exiger que la relation précédente reproduise l'équation de Poisson pour le potentiel gravitationnel Φ (cf. chapitre d'Astronomie):

$$\nabla^2 \Phi = 4\pi G \rho \quad (50.200)$$

Remarque: Cette relation montre que le potentiel de gravitation est relié à la matière de façon linéaire par l'intermédiaire de ses dérivées secondes. Einstein pensa donc que le premier membre des équations du champ en relativité générale, membre supposé décrire la géométrie de l'espace-temps, devait donc inclure d'une manière ou d'une autre les dérivées secondes, non pas du potentiel de gravitation, mais des potentiels de la métrique. En fait, Einstein essaya de généraliser le membre de droite de l'équation de Poisson: la grandeur recherchée devait inclure non seulement la densité de matière mais aussi l'impulsion (dès que le corps est en mouvement, son énergie augmente et donc sa masse). Pour évaluer l'effet gravitationnel d'un corps il fallait donc combiner sa masse au repos avec son impulsion. Il s'agissait finalement du T.E.I. de rang 2 qui est la généralisation du quadrivecteur impulsion de la relativité restreinte.

Nous avons montré plus haut que dans la limite Newtonienne (approximation du champ faible):

$$h_{00} = -\frac{2\Phi}{c^2} \quad (50.201)$$

et dans notre définition du T.E.I., pour une distribution de matière au repos (ou dans un référentiel

comobile c'est selon....) seule la composante suivante est non nulle:

$$T_{00} = \rho c^2 \quad (50.202)$$

Il vient dès lors que l'équation de Poisson peut s'écrire:

$$\Delta \cdot h_{00} = -\frac{8\pi G}{c^4} T_{00} \quad (50.203)$$

Maintenant revenons sur la relation:

$$G_{\rho\sigma} = \kappa T_{\rho\sigma} \quad (50.204)$$

En contractant les deux membres de la relation précédente, il vient:

$$\begin{aligned} g^{\rho\sigma} G_{\rho\sigma} &= \kappa g^{\rho\sigma} T_{\rho\sigma} \\ g^{\rho\sigma} \left(R_{\rho\sigma} - \frac{1}{2} R g_{\rho\sigma} \right) &= \kappa g^{\rho\sigma} T_{\rho\sigma} \end{aligned} \quad (50.205)$$

Or, le scalaire de Ricci (cf. chapitre de Calcul Tensoriel) est donné par $R = g^{\rho\sigma} R_{\rho\sigma}$. Il vient donc:

$$\left(R - \frac{1}{2} R g^{\rho\sigma} g_{\rho\sigma} \right) = \kappa g^{\rho\sigma} T_{\rho\sigma} \quad (50.206)$$

Or dans la métrique de Minkowski (avec la signature -,+,+,+) il est immédiat que:

$$g^{\rho\sigma} g_{\rho\sigma} = 4 \quad (50.207)$$

Donc:

$$\begin{aligned} R - 2R &= \kappa g^{\rho\sigma} T_{\rho\sigma} \\ R &= -\kappa g^{\rho\sigma} T_{\rho\sigma} \end{aligned} \quad (50.208)$$

En utilisant cette dernière relation, l'équation:

$$R_{\rho\sigma} - \frac{1}{2} R g_{\rho\sigma} = \kappa T_{\rho\sigma} \quad (50.209)$$

qui peut s'écrire aussi:

$$R_{\rho\sigma} = \kappa T_{\rho\sigma} + \frac{1}{2} R g_{\rho\sigma} \quad (50.210)$$

peut finalement se mettre sous la forme:

$$R_{\rho\sigma} = \kappa T_{\rho\sigma} + \frac{1}{2} \left(-\kappa g^{\rho\sigma} T_{\rho\sigma} \right) g_{\rho\sigma} = \kappa \left(T_{\rho\sigma} - \frac{1}{2} g^{\rho\sigma} T_{\rho\sigma} g_{\rho\sigma} \right) \quad (50.211)$$

Intéressons-nous à la composante $\rho = \sigma = 0$ telle que la relation précédente s'écrive:

$$\begin{aligned}
 R_{00} &= \kappa \left(T_{00} - \frac{1}{2} (g^{00} T_{00}) g_{00} \right) = \kappa \left(T_{00} - \frac{1}{2} ((-1) \cdot T_{00}) (-1) \right) \\
 &= \kappa \left(T_{00} - \frac{1}{2} T_{00} \right) = \kappa \frac{1}{2} T_{00}
 \end{aligned} \quad (50.212)$$

Explicitons cette dernière relation en utilisant la définition du tenseur de Ricci (cf. chapitre de Calcul Tensoriel):

$$R_{\alpha\beta} = R^{\mu}_{\alpha\mu\beta} \quad (50.213)$$

Il vient alors:

$$R_{00} = R^{\mu}_{0\mu0} \quad (50.214)$$

Or, le tenseur de Riemann-Christoffel sous forme développée dans ce cas particulier est donné par (cf. chapitre de Calcul Tensoriel):

$$R^{\mu}_{0,\mu0} = R^{\mu}_{0\mu0} = \partial_{\mu} \Gamma^{\mu}_{00} - \partial_0 \Gamma^{\mu}_{0\mu} + \Gamma^{\nu}_{00} \Gamma^{\mu}_{\mu\nu} - \Gamma^{\nu}_{0\mu} \Gamma^{\mu}_{0\nu} \quad (50.215)$$

Remarque: En absence de champ gravitationnel et en coordonnées cartésiennes, il est logique que tous les symboles de Christoffel soient nuls. En effet, les symboles de Christoffel ne traduisent rien de plus que les forces d'inertie. Mais quand nous avons un champ de gravitation, les trajectoires suivies ne sont plus des droites, même dans le cas Newtonien alors les Christoffel sont non nuls...

À l'approximation du champ faible lentement variable dans le temps, les symboles de Christoffel sont d'ordre O et leurs produits sont d'ordre O^2 et les dérivées temporelles sont négligeables devant les dérivées spatiales. Il reste donc seulement les termes d'ordre O tel que:

$$R_{00} = R^{\mu}_{0\mu0} \cong \partial_{\mu} \Gamma^{\mu}_{00} \quad (50.216)$$

Or, nous avons vu dans le chapitre de Calcul Tensoriel que:

$$\Gamma^i_{ij} = \frac{1}{2} g^{il} (\partial_k g_{lj} + \partial_j g_{kl} - \partial_l g_{kj}) \quad (50.217)$$

Dès lors:

$$R_{00} \cong \partial_{\mu} \Gamma^{\mu}_{00} = \partial_{\mu} \left(\frac{1}{2} g^{\mu l} (\partial_0 g_{l0} + \partial_0 g_{0l} - \partial_l g_{00}) \right) \quad (50.218)$$

Or dans l'approximation du champ faible la variation de la métrique par rapport au temps étant négligeable par rapport à la variation spatiale (l'approximation est un peu tirée par les cheveux il faut dire...):

$$\begin{aligned}
R_{00} &\equiv \partial_\mu \Gamma_{00}^\mu = \partial_\mu \left(\frac{1}{2} g^{\mu\lambda} (-\partial_\lambda g_{00}) \right) = -\frac{1}{2} \partial_\mu \left[(\eta^{\mu\lambda} + h^{\mu\lambda}) \partial_\lambda (\eta_{00} + h_{00}) \right] \\
&\equiv -\frac{1}{2} \partial_\mu (\eta^{\mu\lambda} \partial_\lambda h_{00}) = -\frac{1}{2} \eta^{\mu\lambda} \partial_\mu \partial_\lambda h_{00} - \frac{1}{2} (\partial_\lambda h_{00}) (\partial_\mu \eta^{\mu\lambda}) \equiv -\frac{1}{2} \eta^{\mu\lambda} \partial_\mu \partial_\lambda h_{00} \quad (50.219) \\
&= -\frac{1}{2} \nabla^2 h_{00}
\end{aligned}$$

Par conséquent, la relation:

$$R_{00} = \kappa \frac{1}{2} T_{00} \quad (50.220)$$

devient:

$$\nabla^2 h_{00} = -\kappa T_{00} \quad (50.221)$$

et nous constatons immédiatement qu'il s'agit de l'équation de Poisson si et seulement si:

$$\kappa = \frac{8\pi G}{c^4} \quad (50.222)$$

constante qui est parfois appelée "**constante d'Einstein**". Il s'ensuit immédiatement que le scalaire de Ricci est positif et donc que nous sommes localement dans un espace à courbure de type sphérique.

L'équation d'Einstein des champs est donc sous forme définitive:

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu} \quad (50.223)$$

ou de manière plus conventionnelle:

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu} \quad (50.224)$$

La partie de gauche représente la courbure de l'espace-temps telle qu'elle est déterminée par la métrique et l'expression de droite représente une modélisation du contenu masse/énergie de l'espace-temps. Cette équation peut alors être interprétée comme un ensemble d'équations décrivant comment la courbure de l'espace-temps est liée au contenu masse/énergie de l'Univers. Ces équations, ainsi que l'équation de la géodésique, forment le coeur de la formulation mathématique de la relativité générale.

L'équation d'Einstein est donc une équation dynamique qui décrit comment la matière et l'énergie modifie la géométrie de l'espace-temps. Cette courbure de la géométrie autour d'une source de matière est alors interprétée comme le champ gravitationnel de cette source. Le mouvement des objets dans ce champ étant décrit très précisément par l'équation de sa géodésique.

Par ailleurs, nous venons aussi de voir que l'équation d'Einstein se réduit aux lois de la gravité de Newton en utilisant l'approximation des champs faibles et des mouvements lents.

Puisque le tenseur d'énergie-impulsion comporte 16 composantes dont au fait 10 sont réellement

uniques (indépendantes) puisque le tenseur est symétrique, nous pouvons voir l'équation d'Einstein des champs comme dix équations différentielles du second ordre couplées sur tenseur de champ métrique $g_{\mu\nu}$.

Ces équations différentielles sont en général cauchemardesques à résoudre, les scalaires et tenseurs de Ricci sont des contractions du tenseur de Riemann, qui incluent les dérivées et les produits des symboles de Christoffel, qui eux-mêmes sont construits sur le tenseur métrique inverse et sur les dérivées de celui-ci. Pour corser le tout, il est possible de construire des tenseurs d'énergie-impulsion qui peuvent invoquer la métrique aussi. Il est donc très difficile de résoudre les équations d'Einstein des champs dans le cas général et nous devons donc souvent nous appuyer sur des hypothèses simplificatrices.

6. SOLUTION DE SCHWARZSCHILD

La "**métrique de Schwarzschild**" (1916) est une solution de l'équation d'Einstein dans le cas d'un champ gravitationnel isotrope. Elle fournit les trois preuves principales de la Relativité Générale: le décalage des horloges, la déviation de la lumière par le Soleil et l'avance du périhélie de Mercure. Ces trois preuves sont très importantes car l'équation d'Einstein n'était pas démontrée expérimentalement à l'époque.

Pour introduire cette métrique imaginons une source (par exemple le Soleil) qui produit un champ de gravitation à l'aide de sa masse M . Nous cherchons, pour comparer par rapport à l'expérience, les solutions de l'équation d'Einstein (en d'autres termes: la métrique) en dehors de la source (du Soleil donc...) de masse M .

Remarque: Il existe plusieurs techniques mathématiques pour introduire la métrique de Schwarzschild. Le lecteur intéressé pourra chercher par exemple dans la littérature ou sur Internet celle utilisant une transformation de jauge ("jauge d'Einstein" avec la "jauge harmonique") pour la contrainte de perturbation locale. Cette deuxième méthode est très élégante mais plutôt "matheuse" et nous préférons la méthode "à l'ingénieur"...

En d'autres termes, cela revient à avoir dans la région de l'espace qui nous intéresse (en considérant qu'il n'y a que l'astre en question et rien d'autre autour, n'y même l'énergie/masse propre au champ gravitationnel) la propriété suivante:

$$T_{\mu\nu} = 0 \quad (50.225)$$

Donc l'équation d'Einstein des champs démontrée juste plus haut:

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu} \quad (50.226)$$

devient alors:

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = 0 \quad (50.227)$$

Mais nous avons montré plus haut que cette dernière relation peut aussi s'écrire à l'aide de la définition

du scalaire de Ricci $R = g^{\mu\nu} R_{\mu\nu}$:

$$R_{\mu\nu} - \frac{1}{2} R_{\mu\nu} g^{\mu\nu} g_{\mu\nu} = R_{\mu\nu} \left(1 - \frac{1}{2} g^{\mu\nu} g_{\mu\nu} \right) = 0 \quad (50.228)$$

et comme la paratnèse n'est pas nulle puisque nous avons démontré plus haut que $g^{\alpha\sigma} g_{\alpha\sigma} = 4$, il reste:

$$R_{\mu\nu} = 0 \quad (50.229)$$

et donc in extenso le scalaire de Ricci est nul aussi.

Nous devons donc trouver la métrique qui satisfait cette relation (en d'autres termes, une métrique qui loin de la source correspond à un espace plat puisque le tenseur de Ricci est nul). Comme il y en a plusieurs intéressons-nous à un cas particulièrement élégant avec comme les aiment les physiciens... plein de symétries.

L'idée est donc de trouver une métrique si possible indépendante du temps (donc le champ gravitationnel aussi) et... à symétrie sphérique (l'astre étant lui-même de cette forme), prenant en compte la masse de l'astre central (c'est l'objectif majeur!) et telle qu'assez loin de la source (...) ou lorsque la masse est nulle nous retrouvons la métrique classique connue vue plus haut:

$$ds^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (50.230)$$

Mais ceci n'est pas totalement exact! Effectivement, nous travaillons dans l'espace-temps. Or, nous avons vu que l'équation de la métrique curviligne est donné dans un espace-temps plat par:

$$ds^2 = c^2 dt^2 - (dx^2 + dy^2 + dz^2) \quad (50.231)$$

en passant en coordonnées sphériques nous avons alors:

$$ds^2 = c^2 dt^2 - (dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2) \quad (50.232)$$

Et c'est sur cette équation de la métrique que nous devons retomber lorsque nous sommes éloignés de la source ou que la masse de celle-ci est extrêmement faible (la métrique de Schwarzschild doit donc être asymptotiquement plate, c'est-à-dire correspondre alors à l'espace plat de Minkowski).

Donc mettons-nous à la tâche. D'abord, nous partons de ce que nous savons (vaut mieux!). C'est-à-dire que:

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu \quad (50.233)$$

et en coordonnées sphériques avec le temps nous avons pour composantes r, θ, ϕ, t . En toute rigueur, nous notons:

$$(x^\alpha) = (ct, r, \theta, \phi) \quad (50.234)$$

les "coordonnées de Schwarzschild".

Sur un total de 16 termes qu'implique la relation antécédente, nous en retenons finalement 10 à savoir les 4 termes de la diagonale et 6 autres termes d'interaction de sorte d'obtenir:

$$ds^2 = A dt^2 - \left(B dr^2 + C d\theta^2 + D d\phi^2 + F dr d\theta + G dr d\phi + H dr dt + I d\theta d\phi + J d\theta dt + K d\phi dt \right) \quad (50.235)$$

où $A, B, C, \dots$ sont des coefficients à déterminer.

Avant de s'attaquer à ce travail, nous savons que selon une de nos contraintes de départ, lorsque la masse est faible ou que nous sommes éloignés de la source, nous devons donc retomber sur:

$$ds^2 = c^2 dt^2 - \left(dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \right) \quad (50.236)$$

dès lors intuitivement nous pouvons déjà écrire:

$$ds^2 = A c^2 dt^2 - \left(B dr^2 + C r^2 d\theta^2 + D r^2 \sin^2 \theta d\phi^2 + F dr d\theta + G dr d\phi + H dr dt + I d\theta d\phi + J d\theta dt + K d\phi dt \right) \quad (50.237)$$

ce qui admettons-le... est un net progrès...!

Si comme nous nous le sommes imposés au début, l'équation de la métrique est indépendante du temps, nous pouvons par symétrie du temps (hypothèse...) faire le changement de variable suivant:

$$dt = -dt \quad (50.238)$$

sans que cela ne change quoi que ce soit dans notre ds^2 . Or, nous nous rendons tout de suite compte que cela ne sera pas le cas. Immédiatement, pour que cela soit satisfait il faut:

$$H = J = K = 0 \quad (50.239)$$

ce qui nous amène (c'est déjà mieux!) à:

$$ds^2 = A c^2 dt^2 - \left(B dr^2 + C r^2 d\theta^2 + D r^2 \sin^2 \theta d\phi^2 + F dr d\theta + G dr d\phi + I d\theta d\phi \right) \quad (50.240)$$

Maintenant si le système est bien sphérique, l'équation de la métrique doit être invariante par la transformation $d\phi = -d\phi$ (le contraire se saurait depuis longtemps si ce n'était pas le cas expérimentalement) et/ou également pour la transformation $d\theta = -d\theta$.

Donc pour que cela soit juste, nous voyons immédiatement que dans la relation précédente, nous devons imposer:

$$F = G = I = 0 \quad (50.241)$$

Donc finalement nous n'avons plus que:

$$ds^2 = A c^2 dt^2 - \left(B dr^2 + C r^2 d\theta^2 + D r^2 \sin^2 \theta d\phi^2 \right) \quad (50.242)$$

où A, B, C, D seront bien évidemment indépendants du temps (le contraire contredirait notre contrainte

initiale) mais peuvent par symétrie de la sphère être dépendants de r tel que:

$$ds^2 = A(r)c^2 dt^2 - \left(B(r) dr^2 + C(r)r^2 d\theta^2 + D(r)r^2 \sin^2 \theta d\phi^2 \right) \quad (50.243)$$

Maintenant, imaginons-nous sur la sphère (rigoureusement c'est une hypersphère mais cela aide quand même...) à une distance r fixe du centre de la source du champ à un instant donné t fixé. Nous n'avons alors plus que:

$$ds^2 = C(r)r^2 d\theta^2 + D(r)r^2 \sin^2 \theta d\phi^2 \quad (50.244)$$

puisque dt est nul (temps fixé) et dr aussi (distance r fixée).

Nous avons par ailleurs enlevé le signe - car nous avons anticipé le fait qu'il va s'éliminer à la troisième égalité qui va suivre et nous le remettrons ensuite.

Maintenant, imaginons-nous proche du pôle nord de la sphère ($\theta = 0$) nous n'avons alors plus qu'en première approximation:

$$ds^2 = C(r)r^2 d\theta^2 \quad (50.245)$$

et à l'équateur ($\theta = \pi/2$):

$$ds^2 = D(r)r^2 d\phi^2 \quad (50.246)$$

Par symétrie du champ, un déplacement angulaire infinitésimal en chacune de ces deux zones particulières doit pourtant être égal. Dès lors, nous ne pouvons que poser:

$$C(r) = D(r) \quad (50.247)$$

Dès lors, l'équation de la métrique se réduit à:

$$ds^2 = A(r)c^2 dt^2 - B(r) dr^2 - C(r)r^2 \left(d\theta^2 + \sin^2 \theta d\phi^2 \right) \quad (50.248)$$

Montrons maintenant que nous pouvons choisir un système de coordonnées pour lequel $C(r) = 1$.

Introduisons pour cela une distance définie par:

$$\bar{r} = \sqrt{C(r)}r \quad (50.249)$$

d'où:

$$C(r)r^2 = \bar{r}^2 \quad (50.250)$$

Il vient dès lors:

$$\frac{d(\bar{r}^2)}{dr} = C'(r)r^2 + 2C(r)r \quad (50.251)$$

d'où:

$$2\bar{r}d\bar{r} = C'(r)r^2 dr + 2C(r)rdr = 2C(r)rdr \left[1 + \frac{r}{2C(r)} C'(r) \right] \quad (50.252)$$

Ce qui se simplifie encore en:

$$\bar{r}d\bar{r} = C(r)rdr \left[1 + \frac{r}{2C(r)} C'(r) \right] \quad (50.253)$$

Mettons le tout au carré et divisons à gauche et à droite par $C(r)r^2 = \bar{r}^2$:

$$\begin{aligned} (d\bar{r})^2 &= C(r)(dr)^2 \left[1 + \frac{r}{2C(r)} C'(r) \right]^2 \\ \Rightarrow dr^2 &= \frac{1}{C(r)} \left[1 + \frac{r}{2C(r)} C'(r) \right]^{-2} d\bar{r}^2 \end{aligned} \quad (50.254)$$

d'où:

$$Bdr^2 = \frac{B}{C(r)} \left[1 + \frac{r}{2C(r)} C'(r) \right]^{-2} d\bar{r}^2 = B(\bar{r})d\bar{r}^2 \quad (50.255)$$

Dès lors, l'équation de la métrique s'écrit:

$$ds^2 = A(\bar{r})c^2 dt^2 - B(\bar{r})d\bar{r}^2 - \bar{r}^2 (d\theta^2 + \sin^2 \theta d\phi^2) \quad (50.256)$$

C'est donc comme si $C(r) = 1$:

$$ds^2 = A(r)c^2 dt^2 - B(r)dr^2 - r^2 (d\theta^2 + \sin^2 \theta d\phi^2) \quad (50.257)$$

Donc:

$$g_{\mu\nu} = \text{Diag} \left(A(r), -B(r), -r^2, -r^2 \sin^2 \theta \right) \quad (50.258)$$

Soit:

$$g_{\mu\nu} = \begin{bmatrix} A(r) & 0 & 0 & 0 \\ 0 & -B(r) & 0 & 0 \\ 0 & 0 & -r^2 & 0 \\ 0 & 0 & 0 & -r^2 \sin^2 \theta \end{bmatrix} \quad (50.259)$$

et le tenseur métrique contravariant correspondant (dont nous allons avoir besoin plus loin):

$$g^{\mu\nu} = \begin{bmatrix} 1/A(r) & 0 & 0 & 0 \\ 0 & -1/B(r) & 0 & 0 \\ 0 & 0 & -1/r^2 & 0 \\ 0 & 0 & 0 & -1/(r^2 \sin^2 \theta) \end{bmatrix} \quad (50.260)$$

tel que (cf. chapitre de Calcul Tensoriel):

$$g^{\mu\nu} g_{\mu\nu} = \delta_\nu^\mu \quad (50.261)$$

Maintenant, pour déterminer les coefficients restants (soit A et B) nous allons nous aider de la relation que doit satisfaire la métrique:

$$R_{\mu\nu} = R_{\mu\alpha\nu}^\alpha = R_{\mu,\alpha\nu}^\alpha = 0 \quad (50.262)$$

Soit sous forme développée (cf. chapitre de Calcul Tensoriel):

$$\begin{aligned} R_{\mu,\alpha\nu}^\alpha &= \partial_\alpha \Gamma_{\mu\nu}^\alpha - \partial_\nu \Gamma_{\mu\alpha}^\alpha + \Gamma_{\mu\nu}^\beta \Gamma_{\alpha\beta}^\alpha - \Gamma_{\mu\alpha}^\beta \Gamma_{\nu\beta}^\alpha \\ \Gamma_{\mu\nu,\alpha}^\alpha - \Gamma_{\mu\alpha,\nu}^\alpha + \Gamma_{\mu\nu}^\beta \Gamma_{\alpha\beta}^\alpha - \Gamma_{\mu\alpha}^\beta \Gamma_{\nu\beta}^\alpha &= 0 \end{aligned} \quad (50.263)$$

avec bien évidemment (cf. chapitre de Calcul Tensoriel):

$$\Gamma_{ij}^i = \frac{1}{2} g^{il} (\partial_k g_{lj} + \partial_j g_{kl} - \partial_l g_{kj}) = \frac{1}{2} g^{il} (g_{lj,k} + g_{kl,j} - g_{kj,l}) \quad (50.264)$$

C'est dire que l'on a du travail sur la planche... Bon d'abord puisque la métrique est simple les seules dérivées non nulles sont:

$$\begin{aligned} g_{tt,r} &= A' & g_{rr,r} &= -B' & g_{\theta\theta,r} &= -2r \\ g_{\varphi\varphi,r} &= -2r \sin^2 \theta & g_{\varphi\varphi,\theta} &= -2r^2 \sin \theta \cos \theta \end{aligned} \quad (50.265)$$

Nous en déduisons simplement les 9 éléments de la connexion (nous pouvons détailler sur demande...) non nuls:

$$\begin{aligned} \Gamma_{tr}^t &= \Gamma_{rt}^t = -A'/2A & \Gamma_{tt}^r &= -A'/2B & \Gamma_{rr}^r &= -B'/2B \\ \Gamma_{\theta r}^\theta &= \Gamma_{r\theta}^\theta = -1/r & \Gamma_{\theta\theta}^r &= r/B & \Gamma_{\varphi r}^\varphi &= \Gamma_{r\varphi}^\varphi = -1/r \\ \Gamma_{\varphi\varphi}^r &= r \sin^2 \theta / B & \Gamma_{\varphi\theta}^\varphi &= \Gamma_{\theta\varphi}^\varphi = -\cot \theta & \Gamma_{\varphi\varphi}^\theta &= \sin \theta \cos \theta \end{aligned} \quad (50.266)$$

Maintenant que nous avons ces termes de la connexion, il nous faut calculer leur dérivée pour pouvoir exprimer les deux premiers termes de:

$$R_{\mu\nu} = R_{\mu,\alpha\nu}^\alpha = \Gamma_{\mu\nu,\alpha}^\alpha - \Gamma_{\mu\alpha,\nu}^\alpha + \Gamma_{\mu\nu}^\beta \Gamma_{\alpha\beta}^\alpha - \Gamma_{\mu\alpha}^\beta \Gamma_{\nu\beta}^\alpha = 0 \quad (50.267)$$

Il y a alors 10 termes non nuls qui sont:

$$\begin{aligned}
\Gamma_{t r, r}^t &= \Gamma_{r t, r}^t = -\frac{A''}{2A} + \frac{A'^2}{2A^2} & \Gamma_{t t, r}^r &= -\frac{A''}{2B} + \frac{A' B'}{2B^2} \\
\Gamma_{r r, r}^r &= -\frac{B''}{2B} + \frac{B'^2}{2B^2} & \Gamma_{\theta r, r}^\theta &= \Gamma_{r \theta, r}^\theta = \frac{1}{r^2} \\
\Gamma_{\theta \theta, r}^r &= \frac{1}{B} - \frac{r B'}{B^2} & \Gamma_{\varphi r, r}^\varphi &= \Gamma_{r \varphi, r}^\varphi = \frac{1}{r^2} \\
\Gamma_{\varphi \varphi, r}^r &= \frac{\sin^2 \theta}{B} - \frac{r \sin^2 \theta B'}{B^2} & \Gamma_{\varphi \theta, \theta}^\varphi &= \Gamma_{\theta \varphi, \theta}^\varphi = \frac{1}{\sin^2 \theta} \\
\Gamma_{\varphi \varphi, \theta}^\theta &= \frac{r 2 \sin \theta \cos \theta}{B} & \Gamma_{\varphi \varphi, \theta}^\theta &= \cos^2 \theta - \sin^2 \theta
\end{aligned} \quad (50.268)$$

Nous avons finalement pour chaque composante du tenseur de Ricci:

$$\begin{aligned}
R_{tt} &= \left(\frac{A''}{2B} + \frac{A' B'}{2B^2} \right) - (0) + \left(\frac{A'^2}{2AB} \right) - \left(\frac{A'^2}{4AB} + \frac{A' B'}{4B^2} + \frac{A'}{rB} \right) = -\frac{A''}{2B} + \frac{A'^2}{4AB} + \frac{A' B'}{4B^2} - \frac{A'}{rB} \\
R_{rr} &= \left(-\frac{B''}{2B} + \frac{B'^2}{2B^2} \right) - \left(-\frac{A''}{2A} + \frac{A'^2}{2A^2} - \frac{B''}{2B} \right) + \left(\frac{A'^2}{4A^2} + \frac{B'^2}{4B^2} + \frac{2}{r^2} \right) - \left(\frac{A' B'}{4AB} + \frac{B'^2}{4B^2} + \frac{B'}{rB} \right) \\
&= \frac{A''}{2A} - \frac{A'^2}{4A^2} - \frac{A' B'}{4AB} - \frac{B'}{rB} \\
R_{\theta\theta} &= \left(\frac{1}{B} - \frac{r B'}{B^2} \right) - \left(\frac{1}{\sin^2 \theta} \right) + \left(-\frac{2}{B} + \cot^2 \theta \right) - \left(-\frac{A' r}{2AB} - \frac{B' r}{2B^2} - \frac{2}{B} \right) = \frac{A' r}{2AB} - \frac{r B'}{2B^2} + \frac{1}{B} - 1 \\
R_{\varphi\varphi} &= \left(\sin^2 \theta \left[\frac{1}{B} - \frac{r B'}{B^2} - 2 \right] + 1 \right) - (0) + \left(-\sin^2 \theta \left[\frac{2}{B} - 2 \right] - 2 \right) \\
&- \left(-\sin^2 \theta \left[\frac{A' r}{2AB} + \frac{B' r}{2B^2} + \frac{2}{B} - 1 \right] - 1 \right) = \sin^2 \theta \left[\frac{A' r}{2AB} - \frac{B' r}{2B^2} + \frac{1}{B} - 1 \right] \\
R_{r\theta} &= (0) - (0) + (\cot \theta / r) - (\cot \theta / r) = 0
\end{aligned} \quad (50.269)$$

Les seuls éléments directement non nuls sont:

$$\begin{aligned}
R_{tt} &= -\frac{A''}{2B} + \frac{A'^2}{4AB} + \frac{A' B'}{4B^2} - \frac{A'}{rB} = 0 \\
R_{rr} &= \frac{A''}{2A} - \frac{A'^2}{4A^2} - \frac{A' B'}{4AB} - \frac{B'}{rB} = 0 \\
R_{\theta\theta} &= \frac{A' r}{2AB} - \frac{r B'}{2B^2} + \frac{1}{B} - 1 = 0 \\
R_{\varphi\varphi} &= \sin^2 \theta \left[\frac{A' r}{2AB} - \frac{B' r}{2B^2} + \frac{1}{B} - 1 \right] = \sin^2 \theta R_{\theta\theta} = 0
\end{aligned} \quad (50.270)$$

Sous une forme plus conventionnelle (conforme à la littérature) nous pouvons simplifier un peu et par ailleurs ne garder que les trois premières équations:

$$\begin{aligned} R_{tt} &= -2rA\partial_r^2(A)B + r\partial_r(A)^2B + rA\partial_r(A)\partial_r(B) - 4A\partial_r(A)B = 0 \\ R_{rr} &= 2rA\partial_r^2(A)B - r\partial_r(A)^2B - rA\partial_r(A)\partial_r(B) - 4A^2\partial_r(B) = 0 \\ R_{\theta\theta} &= r\partial_r(A)B - r\partial_r(B)A + 2AB - 2AB^2 = 0 \end{aligned} \quad (50.271)$$

Si nous additionnons les deux premières équations, il nous reste:

$$\partial_r(A)B + A\partial_r(B) = 0 \quad (50.272)$$

ce qui équivaut à:

$$\partial_r(A \cdot B) = \partial_r(A)B + A\partial_r(B) = 0 \quad (50.273)$$

et cela nous donne aussi:

$$\partial_r(A)B = -A\partial_r(B) \quad (50.274)$$

Nous avons donc:

$$R_{\theta\theta} = r\partial_r(A)B - r\partial_r(B)A + 2AB - 2AB^2 = 0 \quad (50.275)$$

qui devient:

$$\begin{aligned} r(\partial_r(A)B - \partial_r(B)A) + 2AB - 2AB^2 &= 0 \\ \Rightarrow r(-\partial_r(B)A - \partial_r(B)A) + 2AB - 2AB^2 &= 0 \\ \Rightarrow -r\partial_r(B) + B(1-B) &= 0 \\ \Rightarrow r\partial_r B(r) = B(r)(1-B(r)) \end{aligned} \quad (50.276)$$

où nous avons divient par $2A$ lors du passage de la deuxième à la troisième ligne.

Le lecteur pourra vérifier qu'une solution de l'équation différentielle est:

$$B(r) = \left(1 + \frac{1}{S \cdot r}\right)^{-1} \quad (50.277)$$

où S est une constante réelle non nulle. En conséquence, la métrique pour une solution statique, symétriquement sphérique et dans le vide (...), s'écrit:

$$ds^2 = A(r)c^2 dt^2 - \left(1 + \frac{1}{S \cdot r}\right)^{-1} dr^2 - r^2 (d\theta^2 + \sin^2 \theta d\phi^2) \quad (50.278)$$

Il nous reste à déterminer un coefficient. Mais comme:

$$\partial_r (A \cdot B) = 0 \quad (50.279)$$

il vient:

$$A \cdot B = K = c^{te} \quad (50.280)$$

Donc:

$$A = \frac{K}{B} = K \left(1 + \frac{1}{S \cdot r} \right) \quad (50.281)$$

Donc finalement:

$$ds^2 = K \left(1 + \frac{1}{S \cdot r} \right) c^2 dt^2 - \left(1 + \frac{1}{S \cdot r} \right)^{-1} dr^2 - r^2 (d\theta^2 + \sin^2 \theta d\phi^2) \quad (50.282)$$

Notons que l'espace-temps représenté par cette métrique est asymptotiquement plat, ou, en d'autres termes lorsque $r \rightarrow \infty$, la métrique s'approche de celle de Minkowski, et la variété de l'espace-temps ressemble à celle de l'espace de Minkowski.

Pour calculer les constantes K et S , nous utilisons l'approximation du champ faible. En d'autres termes, nous nous plaçons loin du centre, là où le champ de gravitation est faible. Dans ce cas, la composante g_{tt} de la métrique peut être calculée.

Effectivement, nous avons étudié plus haut la limite newtonienne et avons obtenu la relation suivante:

$$g_{00} = \eta_{00} + h_{00} = \left(1 - \frac{2GM}{rc^2} \right) \quad (50.283)$$

avec (cf. chapitre d'Astronomie) $\varphi = GM / rc^2$. Donc in extenso nous pouvons poser sans trop de crainte:

$$g_{00} = K \left(1 + \frac{1}{S \cdot r} \right) = \left(1 - \frac{2GM}{rc^2} \right) \quad (50.284)$$

soit:

$$K = 1 \text{ et } \frac{1}{S} = -\frac{2GM}{c^2} \quad (50.285)$$

Finalement nous avons pour la "**métrique de Schwarzschild**":

$$ds^2 = \left(1 - \frac{2GM}{c^2 r} \right) c^2 dt^2 - \frac{1}{\left(1 - \frac{2GM}{c^2 r} \right)} dr^2 - r^2 (d\theta^2 + \sin^2 \theta d\phi^2) \quad (50.286)$$

soit en unités naturelles:

$$ds^2 = \left(1 - \frac{2GM}{r}\right) dt^2 - \frac{1}{\left(1 - \frac{2GM}{r}\right)} dr^2 - r^2 (d\theta^2 + \sin^2 \theta d\phi^2) \quad (50.287)$$

Attention!!! Certains ouvrages de référence ont la métrique de Schwarzschild avec des signes différents car ils prennent la métrique $- , + , + , +$ au lieu de la métrique $+ , - , - , -$.

Une singularité toute (physiquement) apparente apparaît lorsque:

$$\left(1 - \frac{2GM}{c^2 r}\right) = 0 \quad (50.288)$$

ou en d'autres termes, lorsque la coordonnée du rayon r vaut:

$$r = \frac{2GM}{c^2} \quad (50.289)$$

Ce rayon, que nous avons déjà déterminé lors de notre étude la mécanique classique, est appelé "**rayon de Schwarzschild**".

Le rayon de Schwarzschild est défini comme le rayon critique prévu par la géométrie de Schwarzschild, en deçà duquel rien ne peut s'échapper: si une étoile ou tout autre objet atteint un rayon égal ou inférieur à son rayon de Schwarzschild (qui dépend de sa masse, cf. ci-dessous), alors elle devient un Trou Noir, et tout objet s'approchant à une distance de celui-ci inférieure au rayon de Schwarzschild ne pourra s'en échapper. Le terme est utilisé en physique et en astronomie pour donner un ordre de grandeur de la taille caractéristique à laquelle des effets de relativité générale deviennent nécessaires pour la description d'objets d'une masse donnée. Les seuls objets qui ne sont pas des trous noirs et dont la taille est du même ordre que leur rayon de Schwarzschild sont les étoiles à neutrons (ou pulsars), ainsi, curieusement, que l'univers observable en son entier.

Remarques:

R1. La singularité dans la métrique lorsque l'on atteint le rayon de Schwarzschild est apparente car il ne s'agit que d'un effet du système de coordonnées utilisées.

R2. Un théorème remarquable affirme que la métrique de Schwarzschild est l'unique solution aux équations d'Einstein dans le vide possédant la symétrie sphérique. Comme la métrique de Schwarzschild est également statique, ceci montre qu'en fait dans le vide toute solution sphérique est automatiquement statique. Une des conséquences intéressantes de ce théorème est que n'importe quelle étoile pulsante qui reste à symétrie sphérique ne peut pas générer d'ondes gravitationnelles (puisque la région de l'espace-temps extérieure à l'étoile doit rester statique).

Maintenant que nous avons la métrique de Schwarzschild revenons sur le critère de Schild que nous avons vu lors de notre étude classique de l'effet Einstein.

Si nous réécrivons la métrique de Schwarzschild pour un corps immobile, nous avons la métrique qui se simplifie en:

$$ds^2 = \left(1 - \frac{2GM}{c^2 r}\right) c^2 dt^2 \quad (50.290)$$

En faisant intervenir le potentiel gravitationnel (cf. chapitre d'Astronomie):

$$\varphi = \frac{GM}{r} \quad (50.291)$$

la métrique s'écrit:

$$ds^2 = \left(1 - \frac{2\varphi}{c^2}\right) c^2 dt^2 \quad (50.292)$$

d'où en introduisant le temps propre:

$$c^2 d\tau^2 = \left(1 - \frac{2\varphi}{c^2}\right) c^2 dt^2 \quad (50.293)$$

d'où:

$$d\tau = \left(1 - \frac{2\varphi}{c^2}\right)^{1/2} dt \quad (50.294)$$

soit:

$$dt = \left(1 - \frac{2\varphi}{c^2}\right)^{-1/2} d\tau \quad (50.295)$$

Le développement au deuxième ordre en série de Maclaurin (cf. chapitre de Suite Et Séries) de la racine négative donne:

$$dt \cong \left(1 + \frac{\varphi}{c^2}\right) d\tau \quad (50.296)$$

Ainsi, nous avons:

$$dt \cong \left(1 + \frac{\varphi}{c^2}\right) d\tau \geq d\tau \quad (50.297)$$

Donc cela démontre que la courbure (la gravitation) engendre une dilatation du temps d'autant plus importante (dans le sens qu'il s'écoule plus vite) que le champ de gravité est intense (la masse M est grande) ou que nous sommes près du corps sous l'influence du champ (rayon r petit).

Or, pour la Terre, le terme:

$$\varphi = \frac{GM}{r}$$

est relativement faible. Mais pour un Trou Noir ou une étoile à Neutrons, ce n'est plus vraiment le cas

et la dilatation devient importante et les effets accessibles à la mesure.

7. VÉRIFICATIONS EXPÉRIMENTALES

Nous allons maintenant passer en revue les quatre vérifications expérimentales classiques du 20ème siècle de la théorie de la relativité générale qui sont:

1. La précession du périhélie qui au niveau des résultats numériques nous posait problème avec les outils de la mécanique classique (cf. chapitre d'Astronomie).
2. La déflexion des ondes électromagnétiques (lumière) passant proches d'un corps stellaire massif qui au niveau des résultats numériques nous posait aussi problème avec les outils de la mécanique classique (cf. chapitre d'Astronomie).
3. La démonstration du critère de Schild (déjà faite dans les paragraphes précédents) comme seul moyen d'expliquer rigoureusement le redshift gravitationnel et l'hypothèse de ralentissement du temps dans un champ gravitationnel.
4. Le retard des signaux électromagnétiques se propageant près du corps massif. Retard désigné sous le nom "d'effet Shapiro" dont les applications numériques sont utilisées pour le fonctionnement du G.P.S et que nous verrons plus loin.

7.1. PRÉCESSION DU PÉRIHÉLIE DE MERCURE

Traisons donc maintenant un des plus fameux exemples de la relativité générale: la précession du périhélie de Mercure. Nous avons déjà traité ce cas dans le chapitre d'Astronomie, mais nous avons mentionné que le résultat théorique numérique ne correspondait pas à l'expérience. Nous allons voir en l'équivalent d'une dizaine de pages A4 de développements détaillés comment la relativité générale permet de réconcilier théorie et expérience.

Pour étudier ce cas, nous allons utiliser le formalisme lagrangien vu dans le chapitre de Mécanique Analytique.

D'abord, rappelons que nous avons obtenu pour la métrique de Schwarzschild:

$$ds^2 = \left(1 - \frac{2GM}{c^2 r}\right) c^2 dt^2 - \frac{1}{\left(1 - \frac{2GM}{c^2 r}\right)} dr^2 - r^2 (d\theta^2 + \sin^2 \theta d\phi^2) \quad (50.298)$$

Ce que nous noterons en divisant par ds^2 :

$$1 = \left(1 - \frac{2GM}{c^2 r}\right) \dot{t}^2 - \frac{1}{\left(1 - \frac{2GM}{c^2 r}\right)} \dot{r}^2 - r^2 (\dot{\theta}^2 + \sin^2 \theta \dot{\phi}^2) \quad (50.299)$$

et pour abrégé les notations, nous posons $l = GM/c^2$ tel que:

$$1 = \left(1 - \frac{2l}{r}\right) c^2 \dot{t}^2 - \frac{1}{\left(1 - \frac{2l}{r}\right)} \dot{r}^2 - r^2 \left(\dot{\theta}^2 + \sin^2 \theta \dot{\phi}^2\right) \quad (50.300)$$

Maintenant rappelons que (cf. chapitre de Mécanique Analytique) en unités naturelles:

$$\delta S = \int \delta(ds) = \int \delta(1 \cdot ds) = 0 \quad (50.301)$$

Donc (c'est très grossier mais cela fonctionne... c'est aussi ça parfois la physique...):

$$\delta S = \int \delta \left(\left(1 - \frac{2l}{r}\right) c^2 \dot{t}^2 - \frac{1}{\left(1 - \frac{2l}{r}\right)} \dot{r}^2 - r^2 \left(\dot{\theta}^2 + \sin^2 \theta \dot{\phi}^2\right) \right) ds = 0 \quad (50.302)$$

Enfin cela signifie que le lagrangien est:

$$L = \left(1 - \frac{2l}{r}\right) c^2 \dot{t}^2 - \frac{1}{\left(1 - \frac{2l}{r}\right)} \dot{r}^2 - r^2 \left(\dot{\theta}^2 + \sin^2 \theta \dot{\phi}^2\right) \quad (50.303)$$

Les équations de Lagrange nous donnent pour la coordonnée θ :

$$\frac{d}{ds} \frac{\partial L}{\partial \dot{\theta}} = \frac{\partial L}{\partial \theta} \quad (50.304)$$

avec donc:

$$\frac{\partial L}{\partial \dot{\theta}} = -2r^2 \dot{\theta} \quad (50.305)$$

d'où:

$$\frac{d}{ds} \frac{\partial L}{\partial \dot{\theta}} = -2r^2 \ddot{\theta} - 4r \dot{\theta} \dot{r} \quad (50.306)$$

et:

$$\frac{\partial L}{\partial \theta} = -2r^2 \sin \theta \cos \theta \dot{\phi}^2 \quad (50.307)$$

d'où finalement pour la coordonnée θ :

$$\frac{d}{ds} (r^2 \dot{\theta}) = r^2 \sin \theta \cos \theta \dot{\phi}^2 \quad (50.308)$$

Faisons de même pour ϕ . Nous avons d'abord:

$$\frac{\partial L}{\partial \dot{\phi}} = -2r^2 \sin^2 \theta \dot{\phi} \quad (50.309)$$

et:

$$\frac{\partial L}{\partial \phi} = 0 \quad (50.310)$$

et il vient immédiatement de par l'application de l'équation d'Euler-Lagrange:

$$\frac{d}{ds} (r^2 \sin^2 \theta \dot{\phi}) = 0 \quad (50.311)$$

Faisons de même pour t :

$$\frac{\partial L}{\partial \dot{t}} = 2 \left(1 - \frac{2l}{r} \right) c^2 \dot{t} \quad (50.312)$$

et il vient ici aussi immédiatement:

$$\frac{d}{ds} \left[\left(1 - \frac{2l}{r} \right) \dot{t} \right] = 0 \quad (50.313)$$

Dès lors:

$$\left(1 - \frac{2l}{r} \right) \dot{t} = c^{te} = A \quad (50.314)$$

Maintenant nous allons supposer que le mouvement de Mercure est dans le plan équatorial tel que $\theta = \pi/2$. Dès lors, la relation obtenue plus haut:

$$\frac{d}{ds} (r^2 \sin^2 \theta \dot{\phi}) = 0 \quad (50.315)$$

se simplifie en:

$$\frac{d}{ds} (r^2 \dot{\phi}) = 0 \quad (50.316)$$

d'où:

$$r^2 \dot{\phi} = c^{te} = B \quad (50.317)$$

Nous avons aussi dès lors l'expression de la ligne d'Univers qui pour rappel est:

$$1 = \left(1 - \frac{2l}{r} \right) c^2 \dot{t}^2 - \frac{1}{\left(1 - \frac{2l}{r} \right)} \dot{r}^2 - r^2 \left(\dot{\theta}^2 + \sin^2 \theta \dot{\phi}^2 \right) \quad (50.318)$$

qui puisque $\theta = \pi/2$ (qui est donc une constante) se simplifie en:

$$1 = \left(1 - \frac{2l}{r}\right) c^2 \dot{t}^2 - \frac{1}{\left(1 - \frac{2l}{r}\right)} \dot{r}^2 - r^2 \dot{\phi}^2 \quad (50.319)$$

Faisons maintenant le remplacement suivant:

$$r^2 \dot{\phi} = B \quad (50.320)$$

qui est donc une constante comme nous l'avons démontré juste plus haut ainsi que le remplacement suivant (qui est aussi une constante comme nous l'avons démontré juste plus haut):

$$\dot{t} = \frac{A}{\left(1 - \frac{2l}{r}\right)^2} \quad (50.321)$$

dans l'élément de ligne d'Univers et nous obtenons::

$$1 = \left(1 - \frac{2l}{r}\right)^{-1} c^2 A^2 - \left(1 - \frac{2l}{r}\right)^{-1} \dot{r}^2 - \frac{B^2}{r^2} \quad (50.322)$$

Considérons aussi r comme fonction ϕ alors:

$$r' = \frac{dr}{d\phi} = \frac{\frac{dr}{ds}}{\frac{d\phi}{ds}} = \frac{\dot{r}}{\dot{\phi}} \quad (50.323)$$

d'où:

$$\dot{r} = r' \dot{\phi} = r' \frac{B}{r^2} \quad (50.324)$$

Ainsi, nous pouvons réécrire la ligne d'Univers sous la forme:

$$\left(1 - \frac{2l}{r}\right) = c^2 A^2 - r'^2 \frac{B^2}{r^4} - \left(1 - \frac{2l}{r}\right) \frac{B^2}{r^2} \quad (50.325)$$

Faisons un changement de variable en posant:

$$r = \frac{1}{u} \Rightarrow \frac{dr}{du} = -\frac{1}{u^2} \Rightarrow dr = -\frac{du}{u^2} \quad (50.326)$$

d'où:

$$\frac{dr}{d\phi} = -\frac{\frac{du}{d\phi}}{u^2} \Rightarrow r' = -\frac{u'}{u^2} \quad (50.327)$$

Ce qui donne pour notre ligne d'Univers:

$$(1 - 2lu) = c^2 A^2 - u'^2 B^2 - (1 - 2lu)u^2 B^2 \quad (50.328)$$

ou:

$$\begin{aligned} u'^2 &= \frac{c^2 A^2}{B^2} - \frac{(1 - 2lu)u^2 B^2}{B^2} - \frac{(1 - 2lu)}{B^2} \\ &= \frac{c^2 A^2}{B^2} - u^2 + 2lu^3 - \frac{1}{B^2} + \frac{2lu}{B^2} \\ &= \frac{c^2 A^2 - 1}{B^2} - u^2 + 2lu^3 + \frac{2lu}{B^2} \end{aligned} \quad (50.329)$$

en différenciant:

$$2u'u'' = -2uu' + 6lu^2 u' + \frac{2lu'}{B^2} \quad (50.330)$$

ou écrit autrement:

$$2u'u'' + 2uu' - 6lu^2 u' - \frac{2lu'}{B^2} = 0 \quad (50.331)$$

ce qui se simplifie et se factorise en:

$$u' \left(u'' - \frac{l}{B^2} + u - 3lu^2 \right) = 0 \quad (50.332)$$

La première solution possible est bien évidemment:

$$u' = 0 \quad (50.333)$$

d'où comme $r=1/u$:

$$r = c^{te} \quad (50.334)$$

Le mouvement circulaire est donc aussi une solution du problème de Kepler en relativité générale dans un champ de Schwarzschild.

L'autre solution sera:

$$u'' - \frac{l}{B^2} + u - 3lu^2 = 0 \quad (50.335)$$

Soit écrit autrement:

$$u'' + u = \frac{l}{B^2} + 3lu^2 \quad (50.336)$$

elle correspond à l'orbite du problème de Kepler.

Faisons la comparaison en considérant en mécanique de Newton le mouvement d'une particule de masse m dans un potentiel U . Le lagrangien (cf. chapitre de Mécanique Analytique) est alors:

$$L = T - V = \frac{1}{2}mv^2 - mf(r) \quad (50.337)$$

En coordonnées polaires nous avons déjà vu dans différents chapitres (de Calcul Vectoriel et d'Astronomie) que la vitesse s'écrit alors:

$$L = \frac{1}{2}m \left[\left(\frac{dr}{dt} \right)^2 + r^2 \left(\frac{d\phi}{dt} \right)^2 \right] - mf(r) \quad (50.338)$$

En utilisant l'équation d'Euler-Lagrange nous avons l'équation du mouvement:

$$\frac{\partial L}{\partial r} - \frac{d}{dt} \frac{\partial L}{\partial \dot{r}} = 0 \quad (50.339)$$

ce qui donne:

$$\frac{\partial L}{\partial r} = rm \left(\frac{d\phi}{dt} \right)^2 - mf'(r) \text{ et } \frac{d}{dt} \frac{\partial L}{\partial \dot{r}} = m \frac{d^2 r}{dt^2} \quad (50.340)$$

d'où:

$$\frac{d^2 r}{dt^2} = r \left(\frac{d\phi}{dt} \right)^2 - f'(r) \quad (50.341)$$

et comme nous l'avons vu dans le chapitre d'Astronomie:

$$r^2 \frac{d\phi}{dt} = K \quad (50.342)$$

est la constante des aires. Introduisons:

$$u(\phi) = \frac{1}{r(\phi)} \quad (50.343)$$

d'où:

$$u'(\phi) = -\frac{1}{r^2} r'(\phi) \quad (50.344)$$

et donc:

$$\frac{dr}{dt} = \frac{dr}{d\phi} \frac{d\phi}{dt} = r'(\phi) \frac{K}{r^2} = -Ku'(\phi) \quad (50.345)$$

Ainsi:

$$\begin{aligned} \frac{d}{dt} \frac{dr}{dt} &= \frac{d}{d\phi} \frac{d\phi}{dt} [-Ku'(\phi)] = \frac{d}{d\phi} [-Ku'(\phi)] \frac{d\phi}{dt} = -Ku''(\phi) \frac{K}{r^2} \\ &= -K^2 u''(\phi) u^2 \end{aligned} \quad (50.346)$$

L'équation:

$$\frac{d^2 r}{dt^2} = r \left(\frac{d\phi}{dt} \right)^2 - f'(r) \quad (50.347)$$

devient alors:

$$-K^2 u''(\phi) u^2 = r \left(\frac{d\phi}{dt} \right)^2 - f'(r) \quad (50.348)$$

Or:

$$\frac{d\phi}{dt} = \frac{K}{r^2} \quad (50.349)$$

d'où:

$$-K^2 u''(\phi) u^2 = \frac{K^2}{r^3} - f'(r) \quad (50.350)$$

soit:

$$-K^2 u''(\phi) u^2 = K^2 u^3 - f'(r) \quad (50.351)$$

ou:

$$u''(\phi) + u(\phi) = \frac{1}{K^2} \frac{f'(r)}{u^2} \quad (50.352)$$

Il s'agit donc de la "**formule de Binet non relativiste**" qui donne donc la relation entre $u=1/r$ et ϕ pour une force centrale (cf. chapitre d'Astronomie). Dans le cas d'un potentiel newtonien:

$$f(r) = -G \frac{M}{r} \Rightarrow f'(r) = G \frac{M}{r^2} = GMu^2 \quad (50.353)$$

d'où:

$$u''(\phi) + u(\phi) = \frac{GM}{K^2} \quad (50.354)$$

avec pour rappel:

$$K = r^2 \frac{d\phi}{dt} \quad (50.355)$$

Or, rappelons la forme de celle que nous avons obtenue avec la relativité générale:

$$u'' + u = \frac{l}{B^2} + 3lu^2 \quad (50.356)$$

Ainsi, nous voyons que le terme analogue en relativité est:

$$\frac{l}{B^2} = \frac{\frac{GM}{c^2}}{B^2} \quad (50.357)$$

et que la relativité générale ajoute le terme $3lu^2$. Or, comme en relativité générale:

$$B = r^2 \dot{\phi} \quad (50.358)$$

Alors:

$$\frac{l}{B^2} = \frac{\frac{GM}{c^2}}{r^4 \left(\frac{d\phi}{ds} \right)^2} = \frac{GM}{c^2 r^4 \left(\frac{d\phi}{dt} \right)^2 \left(\frac{dt}{ds} \right)^2} \quad (50.359)$$

Or, dans le cas de l'approximation des champs faibles:

$$ds^2 = c^2 dt^2 - (dx_1^2 + dx_2^2 + dx_3^2) \quad (50.360)$$

d'où:

$$\frac{ds^2}{dt^2} = c^2 [1 - \beta^2] \underset{v \ll c}{\cong} c^2 \quad (50.361)$$

donc finalement:

$$\frac{l}{B^2} = \frac{GM}{c^2 r^4 \left(\frac{d\phi}{dt} \right)^2 \left(\frac{dt}{ds} \right)^2} \cong \frac{GM}{r^4 \left(\frac{d\phi}{dt} \right)^2} = \frac{GM}{K^2} \quad (50.362)$$

Ceci dit, il est vraiment intéressant de remarquer que l'équation pour la relativité générale:

$$u'' + u = \frac{l}{B^2} + 3lu^2 \quad (50.363)$$

peut être interprétée comme l'équation de Binet pour la mécanique classique:

$$u'' + u = \frac{1}{K^2} \frac{f'(r)}{u^2} \quad (50.364)$$

avec le potentiel:

$$f(r) = -\frac{GM}{r} - \frac{\gamma}{r^3} \quad (50.365)$$

avec $\gamma = lK^2$.

Revenons maintenant à notre équation:

$$u'' + u = \frac{l}{B^2} + 3lu^2 \quad (50.366)$$

Nous aimerions savoir si le deuxième terme à droite de l'égalité est négligeable ou non par rapport au premier terme de droite de l'égalité et ce afin de pouvoir appliquer la théorie des perturbations.

Nous allons d'abord poser à l'aide de l'approximation des champs faibles faite plus haut:

$$C = \frac{l}{B^2} \cong \frac{GM}{K^2} \quad (50.367)$$

Maintenant calculons le rapport:

$$\begin{aligned} \frac{3lu^2}{\frac{l}{B^2}} &= 3B^2u^2 = 3\left(r^2\dot{\phi}\right)^2 \frac{1}{r^2} = 3r^2\dot{\phi}^2 = 3\left(r^2\left(\frac{d\phi}{ds}\right)^2\right) \\ &= 3\left(r^2\left(\frac{d\phi}{dt}\right)^2\left(\frac{dt}{ds}\right)^2\right) \cong 3\left(r\frac{d\phi}{dt}\right)^2 \frac{1}{c^2} \end{aligned} \quad (50.368)$$

Rappelons qu'en coordonnées polaires:

$$v^2 = \left(\frac{dr}{dt}\right)^2 + r^2\left(\frac{d\phi}{dt}\right)^2 \quad (50.369)$$

en approximation, nous pouvons grossièrement poser que:

$$v^2 \cong r^2\left(\frac{d\phi}{dt}\right)^2 \quad (50.370)$$

Dès lors pour Mercure....:

$$\frac{3lu^2}{\frac{l}{B^2}} \cong 3\frac{v^2}{c^2} \cong 7.7 \cdot 10^{-8} \quad (50.371)$$

Donc nous voyons de suite que nous pourrions appliquer les théories variationnelles sur le terme $3lu^2$.

Ainsi, posons:

$$\varepsilon = 3lC = 3l \frac{l}{B^2} \cong 3 \frac{GM}{c^2} \frac{GM}{K^2} = 3 \frac{G^2 M^2}{c^2 K^2} \quad (50.372)$$

L'équation:

$$u'' + u = \frac{l}{B^2} + 3lu^2 \quad (50.373)$$

prend alors la forme:

$$u'' + u = C + \frac{\varepsilon}{C} u^2 \quad (50.374)$$

Pour résoudre cette équation différentielle, nous allons utiliser l'approche de la théorie des perturbations (cf. chapitre de Calcul Différentiel Et Intégral). Nous allons donc nous intéresser à une solution de la forme d'un développement de Taylor en deuxième ordre seulement en ε :

$$u = u_0 + \varepsilon u_1 + O(\varepsilon^2) \quad (50.375)$$

où u_0, u_1 sont bien évidemment dépendants de ϕ et devront être déterminés! Pour cela, nous savons qu'il faut remplacer l'expression précédente dans l'équation différentielle telle que:

$$\left(u_0 + \varepsilon u_1 + O(\varepsilon^2)\right)'' + \left(u_0 + \varepsilon u_1 + O(\varepsilon^2)\right) = C + \frac{\varepsilon}{C} \left(u_0 + \varepsilon u_1 + O(\varepsilon^2)\right)^2 \quad (50.376)$$

Ce qui se simplifie en:

$$u_0'' + \varepsilon u_1'' + u_0 + \varepsilon u_1 = C + \frac{\varepsilon u_0^2}{C} + O(\varepsilon^2) \quad (50.377)$$

où rappelons que:

$$u_0'' + u_0 = C \quad (50.378)$$

est l'équation classique obtenue plus haut:

$$u'' + u = \frac{GM}{K^2} \quad (50.379)$$

considérons la solution du type:

$$u_0 = C + D \cos \phi \quad (50.380)$$

où D est une constante arbitraire. Or, comme nous l'avons vu dans le chapitre d'Astronomie dans le cas de la précession du périhélie:

$$\frac{1}{r} = \frac{B}{\Omega^2} + k_1 \cos(\Omega\theta) \quad (50.381)$$

est au fait une ellipse. Ce qui signifie que toute solution de la forme:

$$u_0 = C + D \cos(\phi) \quad (50.382)$$

est aussi une ellipse!

Pour l'équation en ε :

$$\varepsilon u_1'' + \varepsilon u_1 = \frac{\varepsilon u_0^2}{C} \quad (50.383)$$

qui se simplifie en:

$$u_1'' + u_1 = \frac{(C + D \cos \phi)^2}{C} = C + 2D \cos \phi + \frac{D^2}{C} \cos^2 \phi \quad (50.384)$$

Puisque (cf. chapitre de Trigonométrie):

$$\cos^2 \phi = \frac{\cos(2\phi) + 1}{2} \quad (50.385)$$

Il vient:

$$\begin{aligned} u_1'' + u_1 &= C + 2D \cos \phi + \frac{D^2}{C} \frac{\cos(2\phi) + 1}{2} \\ &= \left(C + \frac{D^2}{2C} \right) + 2D \cos \phi + \frac{D^2}{2C} \cos(2\phi) \end{aligned} \quad (50.386)$$

Pour déterminer u_1 , décomposons-le en trois termes:

$$u_1 = X_1 + X_2 + X_3 \quad (50.387)$$

Ce qui nous donne immédiatement (en injectant les trois termes respectivement dans la dérivée seconde et le terme seul):

$$\begin{aligned} X_1'' + X_1 &= C + \frac{D^2}{2C} \Rightarrow X_1 = C + \frac{D^2}{2C} \\ X_2'' + X_2 &= 2D \cos \phi \Rightarrow X_2 = D \phi \sin \phi \\ X_3'' + X_3 &= \frac{D^2}{2C} \cos(2\phi) \Rightarrow X_3 = -\frac{D^2}{6C} \cos(2\phi) \end{aligned} \quad (50.388)$$

Donc finalement:

$$u_1 = C + \frac{D^2}{2C} + D\phi \sin \phi - \frac{D^2}{6C} \cos(2\phi) \quad (50.389)$$

La solution cherchée est finalement:

$$\begin{aligned} u &= u_0 + \varepsilon u_1 + O(\varepsilon^2) = \left(C + D \cos \phi \right) + \varepsilon \left(C + \frac{D^2}{2C} + D\phi \sin \phi - \frac{D^2}{6C} \cos(2\phi) \right) \\ &= \left(C + \varepsilon C + \frac{\varepsilon D^2}{2C} \right) + \left(D \cos \phi - \varepsilon \frac{D^2}{6C} \cos(2\phi) \right) + \varepsilon D \phi \sin \phi \end{aligned} \quad (50.390)$$

C'est donc avec:

$$u(\phi) = \left(C + \varepsilon C + \frac{\varepsilon D^2}{2C} \right) + \left(D \cos \phi - \varepsilon \frac{D^2}{6C} \cos(2\phi) \right) + \varepsilon D \phi \sin \phi \quad (50.391)$$

qu'il faut calculer le déplacement du périhélie (on y arrive...).

Nous voyons relativement vite en observant la relation précédente que le seul terme dont l'amplitude n'est pas constante est $\varepsilon D \phi \sin \phi$.

Rappelons alors que (cf. chapitre de Trigonométrie):

$$\cos(\phi - \varepsilon\phi) = \cos \phi \cos(\varepsilon\phi) + \sin \phi \sin(\varepsilon\phi) \quad (50.392)$$

Ce qui peut grossièrement s'écrire aussi en première approximation en utilisant les développements de Maclaurin au premier ordre (cf. chapitre de Suites Et Séries):

$$\cos(\phi - \varepsilon\phi) \cong \cos \phi + \phi \sin \phi \varepsilon \quad (50.393)$$

d'où:

$$\begin{aligned} u(\phi) &= \left(C + \varepsilon C + \frac{\varepsilon D^2}{2C} \right) + \left(D \cos \phi - \varepsilon \frac{D^2}{6C} \cos(2\phi) \right) + \varepsilon D \phi \sin \phi \\ &= \left(C + \varepsilon C + \frac{\varepsilon D^2}{2C} \right) + \left(-\varepsilon \frac{D^2}{6C} \cos(2\phi) \right) + D \cos(\phi - \varepsilon\phi) \\ &= C + D \cos(\phi - \varepsilon\phi) + \varepsilon \left(C + \frac{D^2}{2C} - \frac{D^2}{6C} \cos(2\phi) \right) \end{aligned} \quad (50.394)$$

Nous savons que l'orbite d'ordre zéro est:

$$u_0 = C + D \cos(\phi) \quad (50.395)$$

L'effet du dernier terme:

$$\varepsilon \left(C + \frac{D^2}{2C} - \frac{D^2}{6C} \cos(2\phi) \right) \quad (50.396)$$

est donc d'introduire une petite variation périodique dans la distance radiale. Ce terme n'affecte pas le déplacement du périhélie. C'est le terme $\varepsilon\phi$ dans:

$$\cos(\phi - \varepsilon\phi) = \cos(\phi(1 - \varepsilon)) \quad (50.397)$$

qui introduit une non-périodicité qui peut être non négligeable dans le cas où ϕ est grand.

Le périhélie (point le plus proche du Soleil) se présente donc quand r est minimum soit $u = 1/r$ maximum. Or, u est maximum quand le terme qui nous intéresse est maximum, c'est-à-dire:

$$\phi(1 - \varepsilon) = 2\pi n \quad n \in \mathbb{Z} \quad (50.398)$$

Nous avons approximativement:

$$\phi = \frac{2\pi n}{(1 - \varepsilon)} \cong 2\pi n(1 + \varepsilon) \quad (50.399)$$

Pour deux périhélies successifs, nous avons un intervalle:

$$\Delta\phi = 2\pi(1 + \varepsilon) = 2\pi + 2\pi\varepsilon = 2\pi + \delta\phi \quad (50.400)$$

au lieu de 2π . Ainsi, le déplacement pour une révolution est:

$$\delta\phi = 2\pi\varepsilon = 2\pi \left(3 \frac{G^2 M^2}{c^2 K^2} \right) = \frac{6\pi G^2 M^2}{c^2 K^2} = \frac{6\pi G^2 M^2}{c^2 \left(r^2 \frac{d\phi}{dt} \right)^2} = \frac{6\pi G^2 M^2}{c^2 r^2 v^2} \quad (50.401)$$

où K est donc la constante des aires et M la masse de l'astre central et puisque:

$$v = \omega \cdot r = \frac{d\phi}{dt} \cdot r \quad (50.402)$$

Bref, nous avons au final:

$$\Delta\phi = \frac{6\pi G^2 M^2}{c^2 v^2 r^2} \quad (50.403)$$

Relation à comparer avec celle que nous avons obtenue dans le chapitre d'Astronomie avec un traitement newtonien classique:

$$\Delta\alpha = \frac{\pi G^2 M^2}{c^2 v^2 r^2} \quad (50.404)$$

Nous retrouvons donc à la perfection le facteur 6 qui manquait dans les traitements classiques!

Pour Mercure une application numérique donne:

$$\delta\phi = \frac{6\pi G^2 M^2}{c^2 r^2 v^2} \cong 42.89'' \quad (50.405)$$

et l'expérience donne $\delta\phi \cong 42.6'' \pm 1.0''$. De l'aveu même d'Einstein, en obtenant ce résultat il eut des palpitations et l'impression de frôler la crise cardiaque et satisfait de son effort herculéen qui l'avait totalement épuisé il prit une longue période de repos.

Pour terminer sur ce sujet, signalons une deuxième écriture fréquente dans la littérature concernant le résultat obtenu. Effectivement, nous avons démontré dans le chapitre d'Astronomie que le paramètre focal était donné par:

$$p = \frac{\|\vec{v} \times \vec{r}\|^2}{GM} = \frac{v^2 \cdot r^2}{GM} \quad (50.406)$$

Il reste donc:

$$\Delta\phi = \frac{6\pi GM}{c^2 p} \quad (50.407)$$

et nous avons démontré aussi dans le chapitre de Géométrie Analytique que:

$$p = a(1 - e^2) \quad (50.408)$$

Il vient donc au final la forme la plus classique:

$$\Delta\phi = \frac{6\pi GM}{c^2 a(1 - e^2)} \quad (50.409)$$

7.2. DÉFLÉXION DE LA LUMIÈRE

Nous avons donc montré que:

$$u'' + u = \frac{l}{B^2} + 3lu^2 \quad (50.410)$$

en remplaçant les facteurs par leurs valeurs respectives, nous avons:

$$u'' + u = \frac{GM}{c^2 B^2} + 3\frac{GM}{c^2} u^2 \quad (50.411)$$

Mais nous avons vu plus haut que:

$$B^2 = \frac{l K^2}{GM} \quad (50.412)$$

et comme K est la constante des aires donnée par la conservation du moment cinétique lui-même constant (cf. chapitre de Mécanique Classique):

$$b = mr^2 \dot{\phi} = mr^2 \frac{d\phi}{dt} = mK \Rightarrow K = \frac{b}{m} \quad (50.413)$$

Nous avons alors pour un photon $m \rightarrow 0 \Rightarrow K \rightarrow +\infty$.

Finalement l'équation du mouvement d'un photon se résume à:

$$u'' + u = 3 \frac{GM}{c^2} u^2 \quad (50.414)$$

Posons maintenant pour simplifier les notations:

$$r_g = \frac{2GM}{c^2} \quad (50.415)$$

alors:

$$u'' + u = \frac{3}{2} r_g u^2 \quad (50.416)$$

Le terme à droite de l'égalité est petit (vu les constantes qui y interviennent...) si bien qu'une forme approchée de l'équation différentielle est:

$$u'' + u = 0 \quad (50.417)$$

Dont une solution particulière, qui nous le savons d'avance, est intéressante:

$$u = u_0 \cos \varphi \quad (50.418)$$

Nous portons cette solution approximée dans l'équation différentielle initiale et nous obtenons:

$$u'' + u = \frac{3}{2} r_g u_0^2 \cos^2 \varphi \quad (50.419)$$

Soit:

$$u'' + u = \frac{3}{2} r_g u_0^2 \left(\frac{1}{2} + \frac{1}{2} \cos(2\varphi) \right) = \frac{3}{4} r_g u_0^2 + \frac{3}{4} r_g u_0^2 \cos(2\varphi) \quad (50.420)$$

Soit:

$$\frac{\partial^2 u}{\partial \varphi^2} + u = \frac{3}{4} r_g u_0^2 + \frac{3}{4} r_g u_0^2 \cos(2\varphi) \quad (50.421)$$

La suite va être très subtile (comment deviner quelque chose comme cela...?). D'abord nous allons créer une nouvelle équation différentielle:

$$\frac{\partial^2 v}{\partial \varphi^2} + v = \frac{3}{4} r_g u_0^2 + \frac{3}{4} r_g u_0^2 \sin(2\varphi) \quad (50.422)$$

L'astuce consiste à multiplier cette équation par i et la sommer à l'équation différentielle d'origine:

$$\frac{\partial^2}{\partial \phi^2}(u + iv) + (u + iv) = \frac{3}{4} r_{\varepsilon} u_0^2 (1 + i) + \frac{3}{4} r_{\varepsilon} u_0^2 e^{i2\varphi} \quad (50.423)$$

Ce que nous noterons:

$$\frac{\partial^2 \bar{u}}{\partial \phi^2} + \bar{u} = \frac{3}{4} r_{\varepsilon} u_0^2 \sqrt{2} e^{i\frac{\pi}{4}} + \frac{3}{4} r_{\varepsilon} u_0^2 e^{i2\varphi} \quad (50.424)$$

L'astuce est de chercher une solution particulière de la relation précédente sous la forme:

$$\bar{u} = A + B e^{i2\varphi} \quad (50.425)$$

Nous avons alors:

$$\begin{aligned} \bar{u}' &= 2iB e^{i2\varphi} \\ \bar{u}'' &= -4B e^{i2\varphi} \end{aligned} \quad (50.426)$$

Ceci injecté dans notre nouvelle équation différentielle donne:

$$-4B e^{i2\varphi} + A + B e^{i2\varphi} = \frac{3}{4} r_{\varepsilon} u_0^2 \sqrt{2} e^{i\frac{\pi}{4}} + \frac{3}{4} r_{\varepsilon} u_0^2 e^{i2\varphi} \quad (50.427)$$

Nous en déduisons immédiatement:

$$\begin{aligned} A &= \frac{3}{4} r_{\varepsilon} u_0^2 \sqrt{2} e^{i\frac{\pi}{4}} \\ B &= -\frac{1}{4} r_{\varepsilon} u_0^2 \end{aligned} \quad (50.428)$$

Une solution particulière de l'équation différentielle d'origine est donc:

$$\begin{aligned} u &= \Re(\bar{u}) = \Re\left(\frac{3}{4} r_{\varepsilon} u_0^2 \sqrt{2} e^{i\frac{\pi}{4}} - \frac{1}{4} r_{\varepsilon} u_0^2 e^{i2\varphi}\right) = \frac{3}{4} r_{\varepsilon} u_0^2 - \frac{1}{4} r_{\varepsilon} u_0^2 \cos(2\varphi) \\ &= \frac{1}{2} r_{\varepsilon} u_0^2 \left(\frac{3}{2} - \frac{1}{2} \cos(2\varphi)\right) \end{aligned} \quad (50.429)$$

Soit en utilisant les relations trigonométriques remarquables:

$$\cos(2\varphi) = 1 - 2\sin^2(\varphi) \quad (50.430)$$

Il vient:

$$u = \frac{1}{2} r_{\varepsilon} u_0^2 \left(\frac{3}{2} - \frac{1}{2} + \sin^2 \varphi\right) = \frac{1}{2} r_{\varepsilon} u_0^2 (1 + \sin^2 \varphi) \quad (50.431)$$

La solution générale est finalement:

$$u = u_0 \cos \varphi + \frac{1}{2} r_g u_0^2 (1 + \sin^2 \varphi) \quad (50.432)$$

Si nous admettons que la lumière est très faiblement déviée par le Soleil, le rayon de courbure ($1/r$) de sa trajectoire sera très faible.

Ainsi:

$$u = \frac{1}{r} \rightarrow 0 \quad (50.433)$$

tel que:

$$u_0 \cos \varphi + \frac{1}{2} r_g u_0^2 (1 + \sin^2 \varphi) \cong 0 \quad (50.434)$$

Le premier terme est prédominant par rapport au deuxième à cause du facteur r_g qui est très petit sur le deuxième. Pour la suite, nous procédons comme dans le chapitre d'Astronomie (juste les notations changent) pour l'étude de l'angle de déflexion (si vous n'y revenez pas vous ne pourrez comprendre la justification de ce qui va être fait!). Nous posons sans perdre en généralité que:

$$\theta = 2\varphi - \pi \quad (50.435)$$

Soit:

$$\varphi = \frac{\pi}{2} + \frac{\theta}{2} \quad (50.436)$$

et comme:

$$\begin{aligned} \cos \varphi &= \cos \left(\frac{\pi}{2} + \frac{\theta}{2} \right) = -\sin \left(\frac{\theta}{2} \right) \\ \sin^2 \varphi &= \sin^2 \left(\frac{\pi}{2} + \frac{\theta}{2} \right) = \cos^2 \left(\frac{\theta}{2} \right) \end{aligned} \quad (50.437)$$

il vient:

$$-u_0 \sin \left(\frac{\theta}{2} \right) + \frac{1}{2} r_g u_0^2 \left(1 + \cos^2 \left(\frac{\theta}{2} \right) \right) \cong 0 \quad (50.438)$$

En utilisant les relations trigonométriques à nouveau:

$$\cos^2 \left(\frac{\theta}{2} \right) = \frac{1 + \cos \left(2 \frac{\theta}{2} \right)}{2} \quad (50.439)$$

Il vient:

$$-\sin\left(\frac{\theta}{2}\right) + \frac{1}{2}r_g u_0 \left(\frac{3}{2} + \frac{1}{2}\cos(\theta)\right) \cong 0 \quad (50.440)$$

δ étant supposé très petit nous faisons un développement de Maclaurin (cf. chapitre de Suites Et Séries) au premier ordre des fonctions trigonométriques:

$$-\frac{\theta}{2} + \frac{1}{2}r_g u_0 \left(\frac{3}{2} + \frac{1}{2}\right) \cong 0 \quad (50.441)$$

Ce qui donne:

$$-\frac{\theta}{2} + r_g u_0 \cong 0 \Rightarrow \theta = 2r_g u_0 = \frac{4GM}{c^2} \frac{1}{r_0} \quad (50.442)$$

Donc après une série d'approximations... et d'hypothèses limites acceptables..., nous arrivons à:

$$\theta = \frac{4GM}{c^2 R} \quad (50.443)$$

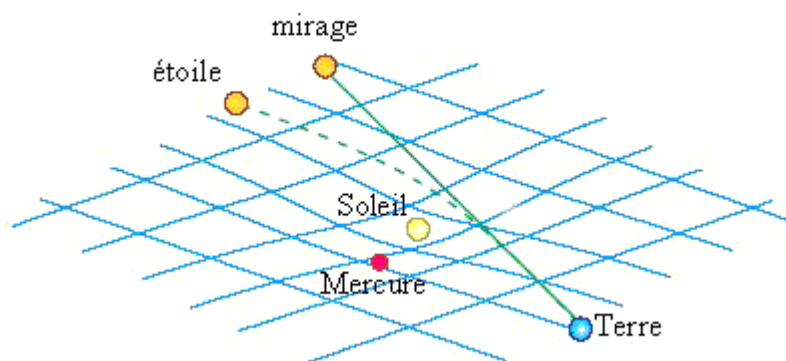
au lieu du résultat que nous avons obtenu selon l'approche newtonienne dans le chapitre d'Astronomie:

$$\varphi \cong \frac{2GM}{v^2 r} \quad (50.444)$$

Nous trouvons donc le facteur 2 qui faisait défaut au traitement classique du problème, relativement aux mesures expérimentales, que nous avons vues dans le chapitre d'Astronomie.

$$\delta_{relativiste} = 2 \cdot \delta_{newtonien} = 1.74'' \quad (50.445)$$

Ce qui est souvent représenté de façon imagée dans les médias par le dessin suivant:



Cette déviation a pu être mise en évidence en mesurant la position des étoiles au voisinage du disque solaire lors de l'éclipse de 1919 par Arthur Eddington et son équipe. Après l'avance du périhélie de Mercure, il s'agissait du second test passé avec succès par la Relativité Générale. C'est cet événement qui a rendu Albert Einstein célèbre auprès du grand public. Aujourd'hui, la déviation des rayons lumineux a pu être mesurée avec beaucoup plus de précision en considérant les signaux radio émis par des sources extragalactiques (quasars, AGN, etc.): la prédiction de la Relativité Générale a été confirmée au millième près.

La déviation des rayons lumineux est aujourd'hui très importante en cosmologie observationnelle, puisqu'elle est à l'origine du phénomène de mirage gravitationnel, encore appelée "**lentille gravitationnelle**".

Il est intéressant de remarquer que toute la théorie des mirages gravitationnels est basée sur la relation:

$$\theta = \frac{4GM}{c^2 R} \quad (50.446)$$

du moins pour un détecteur ponctuel. C'est le seul ingrédient de Relativité Générale utilisé dans le calcul des images.

7.3. EFFET SHAPIRO

En 1964, Shapiro démontra qu'un rayon lumineux n'était pas seulement dévié en passant près d'une masse, mais également que la durée de son trajet était allongée par rapport à une géométrie euclidienne. Il calcula que le retard devait atteindre environ 200 microsecondes, donc parfaitement mesurable, pour une ligne de visée rasant le Soleil. Il suggéra alors de mesurer systématiquement la durée mise par un signal radar pour effectuer le trajet aller-retour entre la Terre et une planète passant derrière le Soleil (pour que l'effet soit maximal). Cela fut d'abord accompli avec des échos radar sur Mars, Vénus ou Mercure, avec une précision de l'ordre de 20%. Le résultat est très net: la durée nécessaire à un signal radar pour faire l'aller-retour Terre-Planète augmente brutalement juste avant que la planète passe derrière le Soleil et diminue tout aussi brutalement quand celle-ci réapparaît.

Remarque: Nous parlons parfois de "ralentissement de la lumière" près du Soleil pour décrire l'effet Shapiro mais c'est une expression maladroite et erronée. Comme cela a déjà été mentionné, la vitesse de la lumière est constante en relativité générale aussi bien qu'en relativité restreinte. Dans le cas de l'effet Shapiro (et dans d'autres cas similaires), ce qui change c'est l'écoulement du temps là où passe la lumière, par rapport à ce qu'il est là où se situe l'observateur.

Bien qu'il s'agisse d'un effet faible, on a pu le vérifier très précisément depuis l'arrivée des sondes Viking sur Mars en 1976, à l'aide de signaux envoyés depuis la Terre vers Mars et réfléchis sur cette dernière par les sondes (voir le principe de l'expérience sur la figure suivante). En outre, il existe même désormais un objet de plus en plus courant pour le fonctionnement duquel l'effet Shapiro doit être pris en compte: le "G.P.S." (Global Positioning System). En effet, malgré la faiblesse du champ de gravitation terrestre, une précision géographique de quelques mètres nécessite de tels détails dans les calculs. Toutefois, un satellite a été lancé récemment dont le but est de vérifier, dans le champ de gravitation terrestre, un effet encore plus faible prédit par la relativité générale et qui n'intervient même pas dans le GPS: l'entraînement de l'espace-temps, aussi nommé "**effet Lense-Thirring**".

Signalons pour le GPS que deux phénomènes d'erreur sont connus dans le cadre de la relativité:

1. Les satellites tournant autour de la Terre à une vitesse approximative de 20'000 kilomètres par heure retardent alors de 7 milliardièmes de seconde par jour (relativité restreinte).
2. À l'altitude de 20'200 kilomètres, celle de l'orbite des satellites, le champ gravitationnel plus faible

fait avancer les horloges satellitaires de 45 millièmes de seconde par jour.

La somme des deux corrections donne une dérive de 38 millièmes de seconde par jour, un chiffre ahurissant pour un système GPS dont la précision se doit d'être de 50 milliardièmes de seconde par jour.

Faisons le calcul pour un rayon frôlant la surface du Soleil. Pour cela, nous reprenons notre métrique de Schwarzschild:

$$ds^2 = \left(1 - \frac{r_g}{r}\right) c^2 dt^2 - \frac{1}{\left(1 - \frac{r_g}{r}\right)} dr^2 - r^2 (d\theta^2 + \sin^2 \theta d\phi^2) \quad (50.447)$$

avec:

$$r_g = \frac{2GM}{c^2} \quad (50.448)$$

Pour un photon, nous savons que $ds = 0$ et donc l'équation de la métrique de Schwarzschild s'écrit alors:

$$c^2 dt^2 = \frac{1}{\left(1 - \frac{r_g}{r}\right)^2} dr^2 + \frac{r^2 (d\theta^2 + \sin^2 \theta d\phi^2)}{\left(1 - \frac{r_g}{r}\right)} \quad (50.449)$$

La trajectoire du photon ayant lieu dans le plan équatorial du Soleil, nous posons:

$$\theta = \pi/2 = \frac{\pi}{2} \quad (50.450)$$

ce qui simplifie encore l'équation de la métrique en:

$$c^2 dt^2 = \frac{1}{\left(1 - \frac{r_g}{r}\right)^2} dr^2 + \frac{r^2 d\phi^2}{\left(1 - \frac{r_g}{r}\right)} = \frac{1}{\left(1 - \frac{r_g}{r}\right)^2} \left(dr^2 + \left(1 - \frac{r_g}{r}\right) r^2 d\phi^2 \right) \quad (50.451)$$

Pour simplifier encore plus nous faisons l'hypothèse que la trajectoire (en coordonnées polaires) du photon rasant le Soleil est rectiligne telle que (pour une des composantes polaires du plan):

$$r \cos \phi = r_\odot \quad (50.452)$$

où $r_\odot$ est le rayon du Soleil. Nous allons utiliser cette hypothèse pour simplifier l'équation de la métrique. Pour cela nous réarrangeons:

$$r = \frac{r_\odot}{\cos \phi} \quad (50.453)$$

Nous dérivons (cf. chapitre de Calcul Différentiel Et Intégral):

$$\frac{dr}{d\phi} = \frac{r_{\odot} \sin \phi}{\cos^2 \phi} = \frac{r_{\odot} \sin \phi}{\frac{r_{\odot}^2}{r^2}} = \frac{r^2 \sin \phi}{r_{\odot}} \quad (50.454)$$

Si nous mettons le tout au carré:

$$\left(\frac{dr}{d\phi}\right)^2 = \frac{r^4 \sin^2 \phi}{r_{\odot}^2} = \frac{r^4 (1 - \cos^2 \phi)}{r_{\odot}^2} = \frac{r^4}{r_{\odot}^2} - \frac{r^4 \cos^2 \phi}{r_{\odot}^2} = \frac{r^4}{r_{\odot}^2} - r^2 = \frac{r^2 (r^2 - r_{\odot}^2)}{r_{\odot}^2} \quad (50.455)$$

d'où:

$$\frac{r_{\odot}^2}{(r^2 - r_{\odot}^2)} dr^2 = r^2 d\phi^2 \quad (50.456)$$

Nous pouvons maintenant récrire l'équation de la métrique:

$$\begin{aligned} c^2 dt^2 &= \frac{1}{\left(1 - \frac{r_g}{r}\right)^2} \left(dr^2 + \left(1 - \frac{r_g}{r}\right) r^2 d\phi^2 \right) = \frac{dr^2}{\left(1 - \frac{r_g}{r}\right)^2} \left(1 + \left(1 - \frac{r_g}{r}\right) \frac{r_{\odot}^2}{(r^2 - r_{\odot}^2)} \right) \\ &= \frac{dr^2}{\left(1 - \frac{r_g}{r}\right)^2} \left(\frac{(r^2 - r_{\odot}^2) + r_{\odot}^2}{(r^2 - r_{\odot}^2)} - \frac{r_g}{r} \frac{r_{\odot}^2}{(r^2 - r_{\odot}^2)} \right) = \frac{dr^2}{\left(1 - \frac{r_g}{r}\right)^2} \left(\frac{r^2}{(r^2 - r_{\odot}^2)} - \frac{r_g}{r} \frac{r_{\odot}^2}{(r^2 - r_{\odot}^2)} \right) \\ &= \frac{dr^2}{\left(1 - \frac{r_g}{r}\right)^2} \frac{r^2}{(r^2 - r_{\odot}^2)} \left(1 - \frac{r_g r_{\odot}^2}{r^3} \right) \end{aligned} \quad (50.457)$$

En prenant la racine:

$$cdt = \frac{dr}{\left(1 - \frac{r_g}{r}\right)} \frac{r}{(r^2 - r_{\odot}^2)^{1/2}} \left(1 - \frac{r_g r_{\odot}^2}{r^3} \right)^{1/2} \quad (50.458)$$

Étant donné que $r \geq r_{\odot}$ et que $r_g \ll 0$ alors:

$$\frac{r_g}{r} \ll 1 \quad \frac{r_g r_{\odot}^2}{r^3} \ll 1 \quad (50.459)$$

Dès lors nous avons en utilisant les développements de Maclaurin (cf. chapitre de Suites Et Séries) au premier ordre:

$$cdt \cong dr \left(1 + \frac{r_g}{r} \right) \frac{r}{(r^2 - r_\odot^2)^{1/2}} \left(1 - \frac{1}{2} \frac{r_g r_\odot^2}{r^3} \right) \quad (50.460)$$

Nous avons alors:

$$\begin{aligned} dt &\cong \left(\frac{1}{c} dr + \frac{r_g}{cr} dr \right) \left(\frac{r}{(r^2 - r_\odot^2)^{1/2}} - \frac{1}{2} \frac{r_g r_\odot^2}{r^3} \frac{r}{(r^2 - r_\odot^2)^{1/2}} \right) \\ &= \frac{1}{c} dr \frac{r}{(r^2 - r_\odot^2)^{1/2}} + \frac{r_g}{cr} dr \frac{r}{(r^2 - r_\odot^2)^{1/2}} - \frac{1}{c} dr \frac{1}{2} \frac{r_g r_\odot^2}{r^3} \frac{r}{(r^2 - r_\odot^2)^{1/2}} \\ &\quad - \frac{r_g}{cr} dr \frac{1}{2} \frac{r_g r_\odot^2}{r^3} \frac{r}{(r^2 - r_\odot^2)^{1/2}} \end{aligned} \quad (50.461)$$

Nous avons finalement une fois condensé:

$$dt \cong \frac{r}{c(r^2 - r_\odot^2)^{1/2}} dr + \frac{r_g}{c(r^2 - r_\odot^2)^{1/2}} dr - \frac{r_g r_\odot^2}{2cr^2(r^2 - r_\odot^2)^{1/2}} dr - \frac{r_g^2 r_\odot^2}{2cr^3(r^2 - r_\odot^2)^{1/2}} dr \quad (50.462)$$

Ce qu'il est de tradition de noter (nous sortons le $1/c$ des différents termes):

$$dt \cong \frac{1}{c} e(r) dr + \frac{1}{c} f(r) dr - \frac{1}{c} g(r) dr - \frac{1}{c} h(r) dr \quad (50.463)$$

S'il n'y a pas de masse alors l'espace-temps est plat et $r_g = 0$. Dès lors:

$$dt \cong \frac{r}{c(r^2 - r_\odot^2)^{1/2}} dr \quad (50.464)$$

Nous pouvons ainsi distinguer le temps classique du temps supplémentaire engendré par l'espace courbe. Le "retard" sera donc donné par:

$$\int dt = \Delta t \cong \frac{1}{c} \int e(r) dr + \frac{1}{c} \int f(r) dr - \frac{1}{c} \int g(r) dr - \frac{1}{c} \int h(r) dr \quad (50.465)$$

Ensuite, pour intégrer les quatre fonctions de r il faut se placer dans un référentiel placé si possible au centre de l'astre principal (le Soleil typiquement) puisque la métrique de Schwarzschild est basée sur cette hypothèse pour rappel... Ainsi, pour connaître le retard d'un rayon lumineux partant du Soleil jusqu'à la Terre, nous choisirons logiquement comme rayon de départ celui du Soleil lui-même et comme rayon d'arrivée, la distance Soleil-Terre (donc cela correspondra une fois les primitives calculées aux bornes d'intégration).

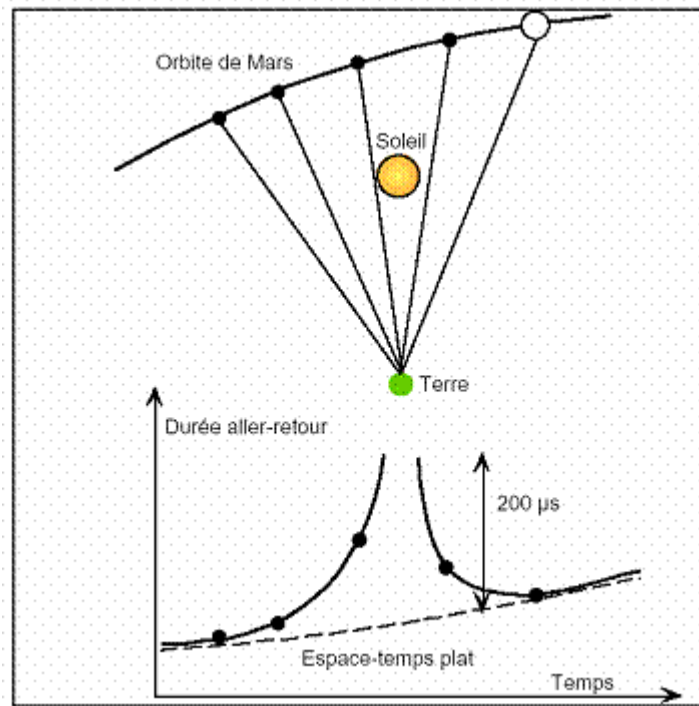


Figure: 50.1 - Temps d'aller-retour d'un signal en fonction de la position de Mars

Bon cela dit c'est bien joli de connaître les notations d'usage, mais c'est encore mieux de faire une application numérique! Nous allons donc d'abord déterminer la primitive de chacun des termes ci-dessous:

$$\int \frac{r}{\sqrt{r^2 - r_{\odot}^2}} dr, \int \frac{r_g}{\sqrt{r^2 - r_{\odot}^2}} dr, \int \frac{r_g r_{\odot}^2}{2r^2 \sqrt{r^2 - r_{\odot}^2}} dr, \int \frac{r_g^2 r_{\odot}^2}{2r^3 \sqrt{r^2 - r_{\odot}^2}} dr \quad (50.466)$$

Les deux premières primitives sont simples car il s'agit de primitives usuelles démontrées en détail dans le chapitre de Calcul Différentiel Et Intégral:

$$\begin{aligned} \int \frac{r}{\sqrt{r^2 - r_{\odot}^2}} dr &= \sqrt{r^2 - r_{\odot}^2} \\ \int \frac{r_g}{\sqrt{r^2 - r_{\odot}^2}} dr &= r_g \operatorname{arccosh} \left(\frac{r}{r_{\odot}} \right) = r_g \ln \left(\frac{r}{r_{\odot}} + \sqrt{\frac{r^2}{r_{\odot}^2} - 1} \right) \end{aligned} \quad (50.467)$$

où pour la dernière primitive nous avons préservé la constante d'intégration (contrairement à ce qui a été fait dans le chapitre de Calcul Différentiel Et Intégral car $r_{\odot} \neq 1$).

Maintenant il nous reste les deux dernières intégrales. Commençons dans l'ordre par:

$$\int \frac{r_g r_{\odot}^2}{2r^2 \sqrt{r^2 - r_{\odot}^2}} dr = \frac{r_g r_{\odot}^2}{2} \int \frac{1}{r^2 \sqrt{r^2 - r_{\odot}^2}} dr \quad (50.468)$$

En posant:

$$r = r_{\odot} \cosh(t), t = \operatorname{arccosh}\left(\frac{r}{r_{\odot}}\right), dr = r_{\odot} \sinh(t) dt \quad (50.469)$$

et en utilisant les résultats démontrés dans le chapitre de Calcul Différentiel Et Intégral, nous avons alors:

$$\begin{aligned} \frac{r_{\mathcal{E}} r_{\odot}^2}{2} \int \frac{1}{r^2 \sqrt{r^2 - r_{\odot}^2}} dr &= \frac{r_{\mathcal{E}} r_{\odot}^2}{2} \int \frac{r_{\odot} \sinh(t)}{r_{\odot}^2 \cosh^2(t) \sqrt{r_{\odot}^2 \cosh^2(t) - r_{\odot}^2}} dt \\ &= \frac{r_{\mathcal{E}}}{2} \int \frac{r_{\odot} \sinh(t)}{\cosh^2(t) \sqrt{r_{\odot}^2 \sinh^2(t)}} dt = \frac{r_{\mathcal{E}}}{2} \int \frac{1}{\cosh^2(t)} dt = \frac{r_{\mathcal{E}}}{2} \tanh(t) = \frac{r_{\mathcal{E}}}{2} \frac{\sinh(t)}{\cosh(t)} \\ &= \frac{r_{\mathcal{E}}}{2} \frac{\sinh\left(\operatorname{arccosh}\left(\frac{r}{r_{\odot}}\right)\right)}{\cosh\left(\operatorname{arccosh}\left(\frac{r}{r_{\odot}}\right)\right)} = \frac{r_{\mathcal{E}}}{2} \frac{\sinh\left(\operatorname{arccosh}\left(\frac{r}{r_{\odot}}\right)\right)}{\frac{r}{r_{\odot}}} \end{aligned} \quad (50.470)$$

Puisque nous avons (cf. chapitre de Trigonométrie):

$$\sinh(x) = \sqrt{\cosh^2(x) - 1} \quad (50.471)$$

Alors:

$$\begin{aligned} \frac{r_{\mathcal{E}} r_{\odot}^2}{2} \int \frac{1}{r^2 \sqrt{r^2 - r_{\odot}^2}} dr &= \frac{r_{\mathcal{E}}}{2} \frac{\sinh\left(\operatorname{arccosh}\left(\frac{r}{r_{\odot}}\right)\right)}{\frac{r}{r_{\odot}}} = \frac{r_{\mathcal{E}}}{2} \frac{\sqrt{\cosh^2\left(\operatorname{arccosh}\left(\frac{r}{r_{\odot}}\right)\right) - 1}}{\frac{r}{r_{\odot}}} \\ &= \frac{r_{\mathcal{E}}}{2} \frac{\sqrt{\left(\frac{r}{r_{\odot}}\right)^2 - 1}}{\frac{r}{r_{\odot}}} = \frac{r_{\mathcal{E}}}{2} \frac{\frac{\sqrt{r^2 - r_{\odot}^2}}{r_{\odot}}}{\frac{r}{r_{\odot}}} = r_{\mathcal{E}} \frac{\sqrt{r^2 - r_{\odot}^2}}{2r} \end{aligned} \quad (50.472)$$

Enfin, il reste la dernière primitive:

$$\int \frac{r_{\mathcal{E}}^2 r_{\odot}^2}{2r^3 \sqrt{r^2 - r_{\odot}^2}} dr = \frac{r_{\mathcal{E}}^2 r_{\odot}^2}{2} \int \frac{1}{r^3 \sqrt{r^2 - r_{\odot}^2}} dr = \frac{r_{\mathcal{E}}^2 r_{\odot}^2}{2} \int \frac{1}{r^4 \sqrt{1 - \left(\frac{r_{\odot}}{r}\right)^2}} dr \quad (50.473)$$

Nous posons pour la suite:

$$r = \frac{r_{\odot}}{x}, dr = -\frac{r_{\odot}}{x^2} dx \quad (50.474)$$

Il vient alors:

$$\begin{aligned} \frac{r_g^2 r_\odot^2}{2} \int \frac{1}{r^4 \sqrt{1 - \left(\frac{r_\odot}{r}\right)^2}} dr &= \frac{r_g^2}{2r_\odot} \int \frac{-x^2}{\sqrt{1-x^2}} dx = \frac{r_g^2}{2r_\odot} \int \frac{1-x^2-1}{\sqrt{1-x^2}} dx \\ &= \frac{r_g^2}{2r_\odot} \left[\int \sqrt{1-x^2} dx - \int \frac{1}{\sqrt{1-x^2}} dx \right] \end{aligned} \quad (50.475)$$

Dans le chapitre de Calcul Différentiel Et Intégral nous avons démontré que:

$$\int \sqrt{1-x^2} dx = \frac{1}{2} \arcsin x + \frac{x}{2} \sqrt{1-x^2} \quad (50.476)$$

et que:

$$\int \frac{1}{\sqrt{1-x^2}} dx = \arcsin x \quad (50.477)$$

Donc:

$$\begin{aligned} \frac{r_g^2}{2r_\odot} \left[\int \sqrt{1-x^2} dx - \int \frac{1}{\sqrt{1-x^2}} dx \right] &= \frac{r_g^2}{2r_\odot} \left[\frac{1}{2} \arcsin x + \frac{x}{2} \sqrt{1-x^2} - \arcsin x \right] \\ &= \frac{r_g^2}{2r_\odot} \left[\frac{x}{2} \sqrt{1-x^2} - \frac{1}{2} \arcsin x \right] \end{aligned} \quad (50.478)$$

Pour revenir à l'intégrale du début on se rappelle que $r = \frac{r_\odot}{x}$. Donc:

$$\int \frac{r_g^2 r_\odot^2}{2r^3 \sqrt{r^2 - r_\odot^2}} dr = \frac{r_g^2}{2r_\odot} \left[\frac{r_\odot}{2r^2} \sqrt{r^2 - r_\odot^2} - \frac{1}{2} \arcsin \left(\frac{r_\odot}{r} \right) \right] \quad (50.479)$$

Nous avons donc au final en prenant tous les primitives calculés plus haut en en choisissant une borne de départ et d'arrivée pour le calcul:

$$\Delta t = \frac{1}{c} \left[\sqrt{r^2 - r_\odot^2} + r_g \ln \left(\frac{r}{r_\odot} + \sqrt{\frac{r^2}{r_\odot^2} - 1} \right) - r_g \frac{\sqrt{r^2 - r_\odot^2}}{2r} - \frac{r_g^2}{2r_\odot} \left[\frac{r_\odot}{2r^2} \sqrt{r^2 - r_\odot^2} - \frac{1}{2} \arcsin \left(\frac{r_\odot}{r} \right) \right] \right] \Bigg|_{R_0}^{R_1} \quad (50.480)$$

Nous voyons dans le cas limite Newtonien où $r_g = 0$, cette relation se réduit comme peau de chagrin à:

$$\Delta t \cong \frac{1}{c} \sqrt{r^2 - r_\odot^2} \Bigg|_{R_0}^{R_1} = \frac{1}{c} \left(\sqrt{R_0^2 - r_\odot^2} - \sqrt{R_1^2 - r_\odot^2} \right) \quad (50.481)$$

Donc pour un aller-retour (entre planète et satellite par exemple), il vient alors dans ce cas simplifié:

$$\Delta t \cong \frac{2}{c} \left(\sqrt{R_0^2 - r_\odot^2} - \sqrt{R_1^2 - r_\odot^2} \right) \quad (50.482)$$

7.4. TROUS NOIRS

En restant toujours à notre métrique de Schwarzschild....Une trajectoire radiale de type lumière implique:

$$ds^2 = 0 \quad (50.483)$$

donc:

$$\left(1 - \frac{2GM}{c^2 r}\right) c^2 dt^2 - \frac{1}{\left(1 - \frac{2GM}{c^2 r}\right)} dr^2 - r^2 (d\theta^2 + \sin^2 \theta d\phi^2) = 0 \quad (50.484)$$

et dans une trajectoire radiale directe (par définition) nous avons aussi:

$$d\theta = 0 \text{ et } d\phi = 0 \quad (50.485)$$

donc:

$$\left(1 - \frac{2GM}{c^2 r}\right) c^2 dt^2 - \frac{1}{\left(1 - \frac{2GM}{c^2 r}\right)} dr^2 = 0 \quad (50.486)$$

Dès lors:

$$\left(1 - \frac{2GM}{c^2 r}\right) c^2 dt^2 = \frac{1}{\left(1 - \frac{2GM}{c^2 r}\right)} dr^2 \quad (50.487)$$

Il vient alors:

$$\frac{dt^2}{dr^2} = \frac{1}{\left(1 - \frac{2GM}{c^2 r}\right)^2 c^2} \quad (50.488)$$

D'où:

$$\frac{dt}{dr} = \pm \frac{1}{c} \left(1 - \frac{2GM}{c^2 r}\right)^{-1} \quad (50.489)$$

Posons en unités naturelles $c = 1$. Il vient alors:

$$\frac{dt}{dr} = \pm \frac{1}{\left(1 - \frac{2GM}{r}\right)} \quad (50.490)$$

Lorsque $r \rightarrow 2GM$ le membre de droite de l'égalité tend vers $\pm\infty$, donc l'évolution du temps t (observateur extérieur) en fonction de r tend vers l'infini par rapport au temps propre de la lumière.

La sphère donnée par le rayon:

$$r = r_H = \frac{2GM}{c^2} \quad (50.491)$$

définit "l'horizon du Trou Noir de Schwarzschild".

Vers cette frontière limite, la lumière semble mettre un temps infini par rapport à un observateur extérieur à se déplacer lorsqu'elle approche un Trou Noir. Elle ne parvient donc jamais vraiment à l'atteindre par rapport à l'observateur, d'où le fait que les Trous Noirs peuvent être entourés en fonction de leur environnement d'un halo lumineux aux abords du rayon de Schwarzschild. De plus, puisque le temps semble arrêté, la fréquence de la lumière environnant le Trou Noir tend vers zéro et donc vers l'infra-rouge.

Signalons encore un point très important. Avant Einstein, la géométrie était considérée comme partie intégrante des lois. Einstein a montré que la géométrie de l'espace évolue dans le temps selon d'autres lois, encore plus profondes. Il est important de bien comprendre ce point. La géométrie de l'espace ne fait pas partie des lois de la nature. Par conséquent, rien que nous puissions trouver dans ces lois ne dit ce qu'est la géométrie de l'espace. Ainsi, avant de commencer à résoudre les équations de la théorie générale de la relativité d'Einstein, nous n'avons strictement aucune idée de ce qu'est la géométrie. Nous la découvrons seulement une fois les équations résolues.

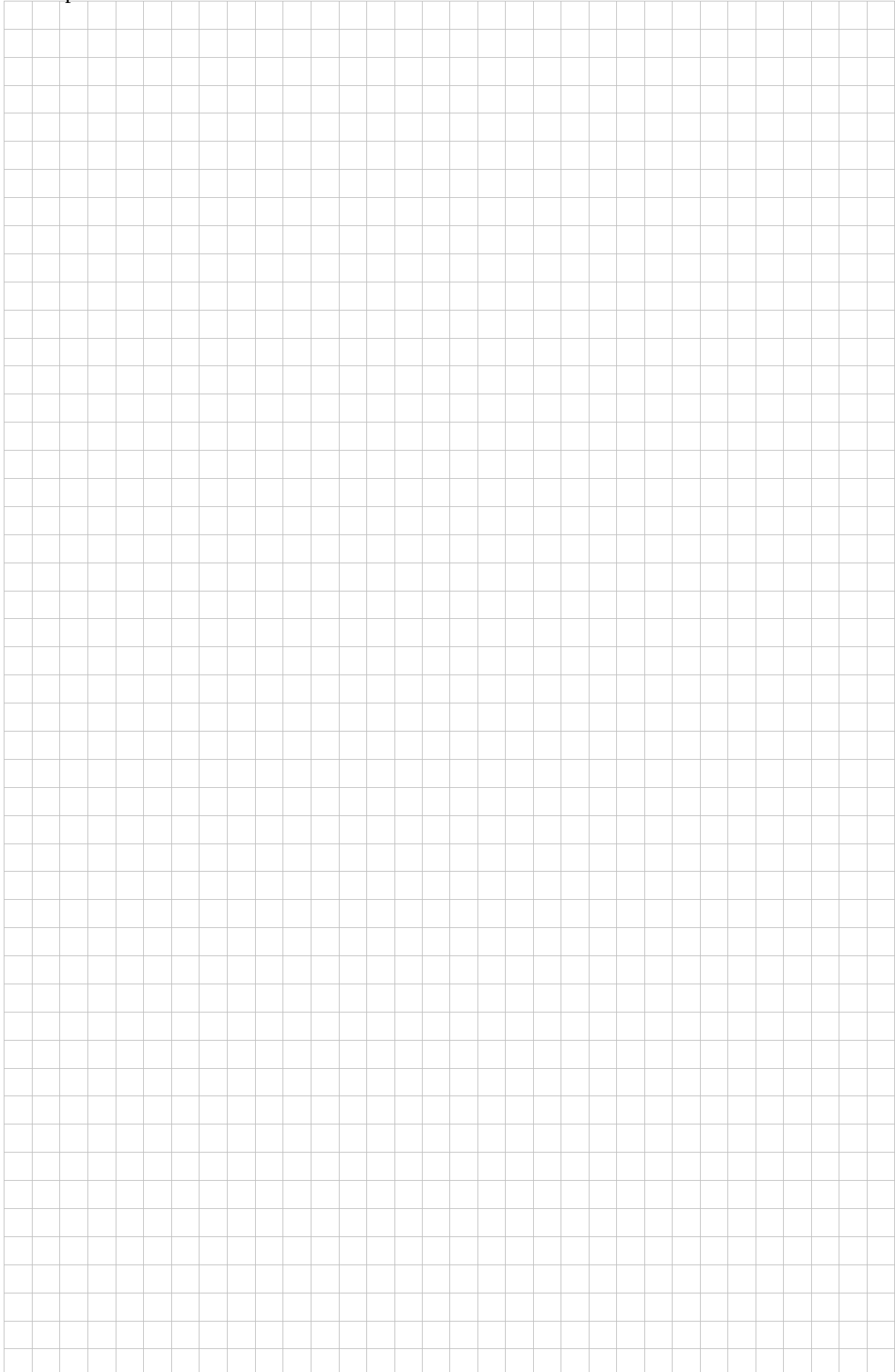
Cela signifie que les lois de la nature doivent s'exprimer sous une forme qui ne présuppose pas que l'espace ait une géométrie fixe. C'est le coeur de la leçon einsteinienne. Cette forme se traduit en un principe appelé "indépendance par rapport au fond". Ce principe énonce donc que les lois de la nature peuvent être décrites dans leur totalité sans présupposer la géométrie de l'espace.

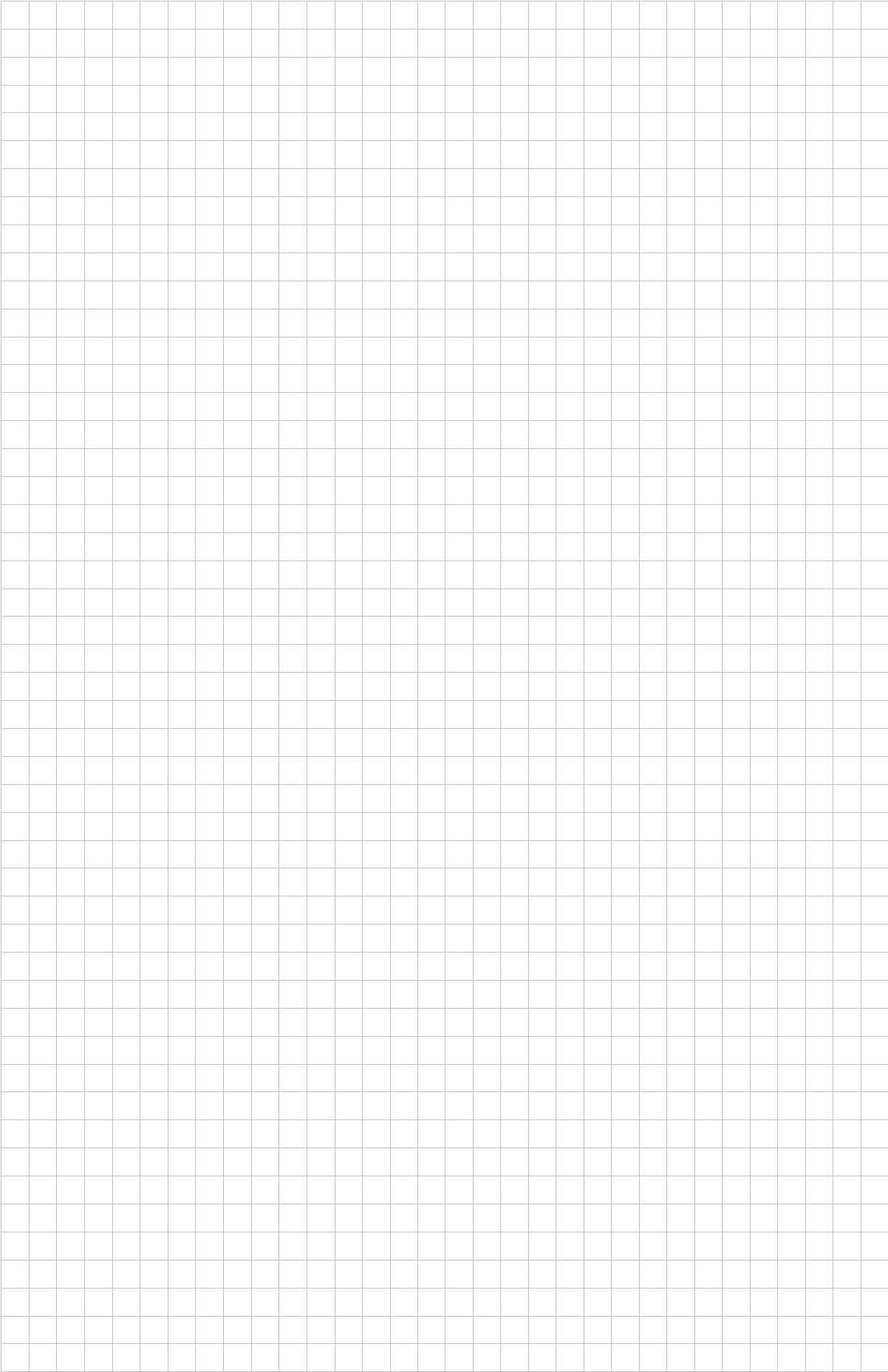
In extenso, le choix des quatre dimensions fait partie du fond. Serait-il possible qu'une autre théorie plus profonde ne nécessite pas de présupposer le nombre de dimensions?

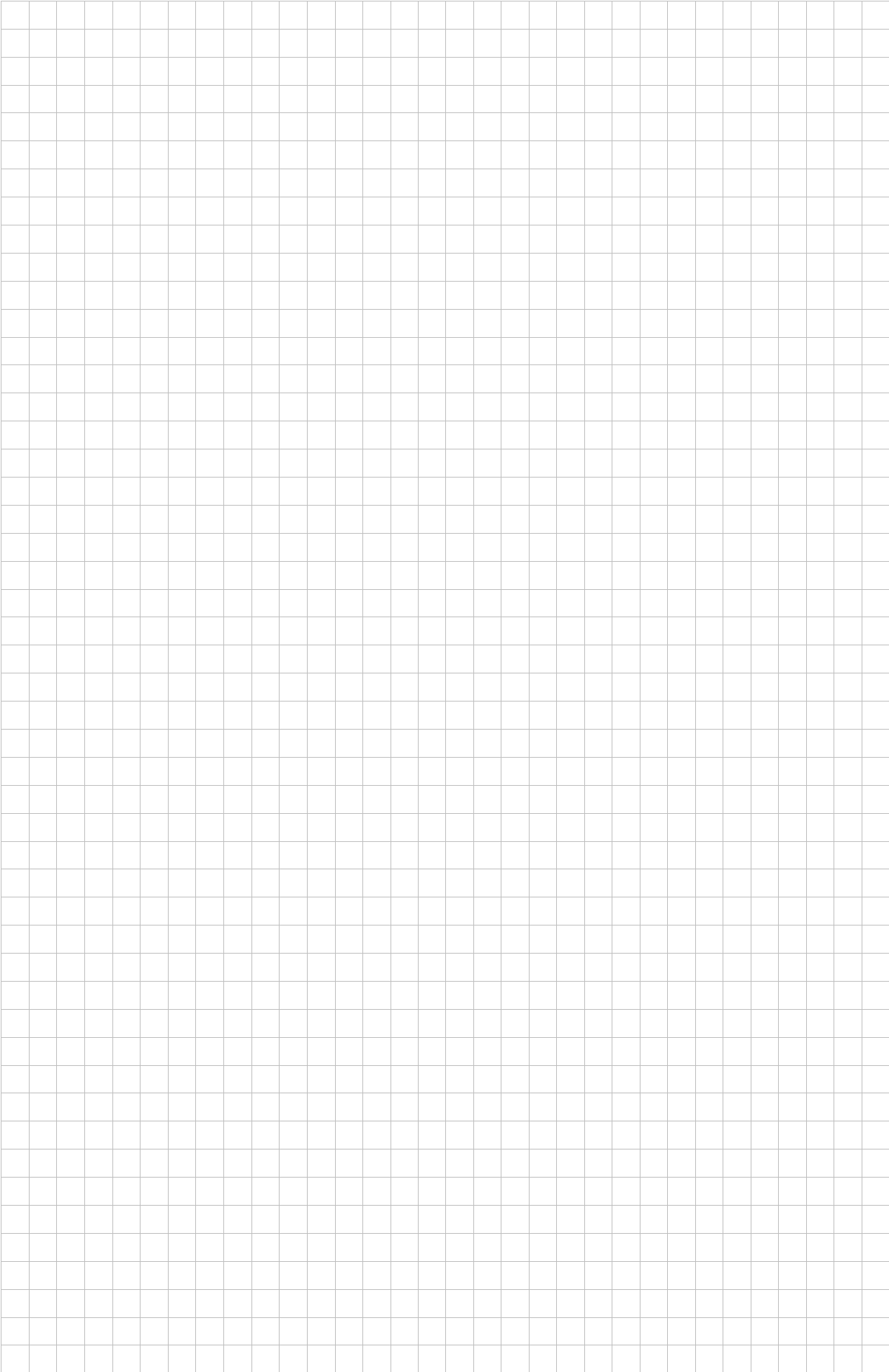
En résumé, l'idée de l'indépendance par rapport au fond, dans sa formulation la plus générale est une façon sage de faire de la physique: faite de meilleures théories, dans lesquelles les choses qui, avant, étaient postulées, seront expliquées en permettant à de telles choses d'évoluer dans le temps en fonction de lois nouvelles.

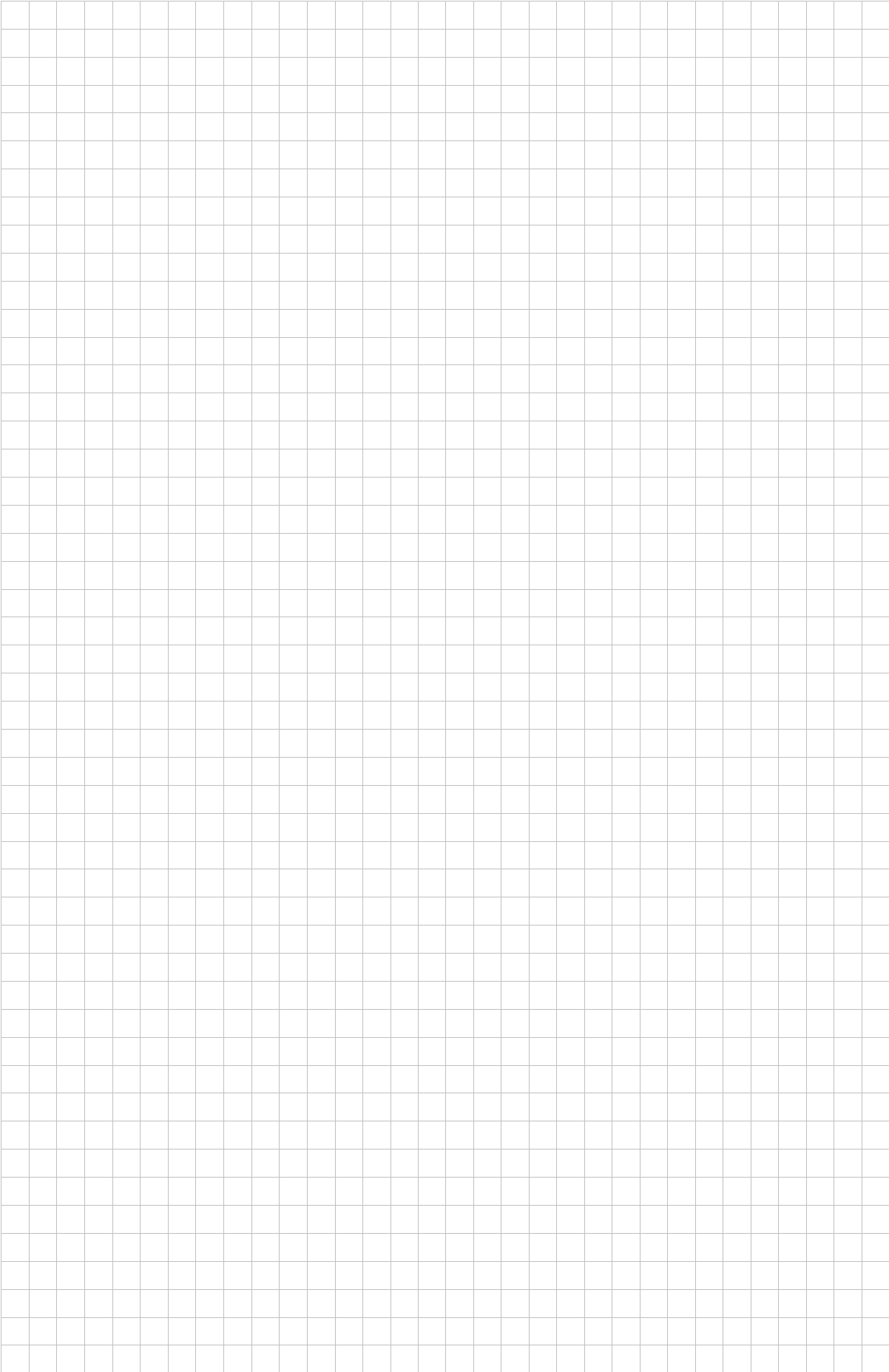
C'est là aussi une difficulté de la théorie quantique. Elle est dépendante de fond contrairement à la relativité générale.

Notes personnelles:









51. COSMOLOGIE

La cosmologie s'occupe de comprendre la naissance et l'évolution de l'Univers par la méthode scientifique. C'est uniquement par ce jeu entre théories physiques, modélisations et observations que nous aborderons cette question ici. Nous éviterons soigneusement toute digression métaphysique. Les problèmes spécifiques de la cosmologie tiennent dans sa définition même: la statistique qui est une des grandes méthodes scientifiques est apparemment pauvre: nous n'avons qu'un univers à notre disposition. En outre, nous n'observons que le passé de l'Univers. Peut-on parler de prédictions dans ces conditions? Les théories sont cependant fiables dans la mesure où elles prédisent des comportements que des observations peuvent tester.

La cosmologie utilise principalement l'arsenal des mathématiques, de la physique théorique, de la physique des particules, de la physique nucléaire, de la physique des détecteurs et de l'astrophysique. Elle est donc interdisciplinaire. La cosmologie traite des échelles supérieures à la taille d'une galaxie jusqu'aux échelles définies par elle-même comme les horizons. Encore que la limite soit volontairement floue, la cosmologie ne traite pas des détails internes de la naissance et de l'évolution d'objets astrophysiques (comme les galaxies, les amas globulaires, ou des amas de galaxies) qui relèvent plus de la "**cosmogonie**".

1. MODÈLE COSMOLOGIQUE NEWTONIEN

Un modèle cosmologique est une représentation mathématique de l'Univers qui cherche à expliquer les raisons de son aspect actuel, et à décrire son évolution au cours du temps (appelé "**temps cosmologique**") mais pas de sa création!

Le modèle newtonien s'applique dans le cadre des hypothèses de la mécanique de Newton (action instantanée). Les résultats que nous allons étudier ici ont été découverts avant le développement de la Relativité Générale mais publiés après! Mais ce modèle présente l'avantage de la simplicité tout en étant capable de mettre en évidence et de discuter de la dynamique de l'Univers et de se préparer à l'étude des modèles d'Univers faisant usage des résultats de la Relativité Générale. Son inconvénient, outre le fait qu'il ne correspond pas tout à fait aux résultats expérimentaux, est de n'être plus valable dans des conditions extrêmes donc de ne pas être extrapolable à l'instant du Big Bang.

Avant de commencer, nous devons définir le "**principe cosmologique**" formé des deux assertions suivantes (en gros, il assure que nous ne sommes pas des observateurs privilégiés, et que ce que nous observons est bien représentatif de l'ensemble de l'Univers):

- L'espace (Univers) est homogène, c'est-à-dire qu'il présente les mêmes propriétés dans toutes ses régions. Ceci doit s'entendre à très grande échelle, au-delà du millier de Mpc (Mégaparsecs). Il est clair qu'à petite échelle existent des inhomogénéités, nous par exemple.
- L'espace (Univers) est isotrope, c'est à dire qu'il n'existe pas de direction particulière de l'espace, comme une direction d'aplatissement, ou un mouvement d'ensemble à l'échelle universelle par exemple.

Remarque: Cette hypothèse de l'isotropie de l'Univers et qui marche relativement bien dans les modèles théoriques (voir ci-après) impose une constatation intéressante si nous admettons un commencement à l'Univers. Cette constatation implique que l'Univers a eu une phase dans son histoire où il n'a pas laissé à la matière le temps de s'agglutiner pour former à ses débuts de groupes de matières inhomogènes et anisotropes qui seraient visibles aujourd'hui à nos télescopes. De ceci, il découle qu'à un moment de son histoire, l'Univers a eu un taux d'expansion supérieur à celui que l'on pourrait faire correspondre à la vitesse de la lumière (c'est mal dit mais j'espère que c'est quand même acceptable).

Nous allons poser quelques autres hypothèses de travail:

H1. L'Univers est un fluide gazeux non visqueux dont les particules sont les galaxies. Sous l'hypothèse du principe cosmologique, le mouvement des galaxies, constituants de ce "fluide" est par construction, statistiquement au repos.

H2. L'Univers est thermodynamiquement un système fermé, sans travail et adiabatique (pas d'échange de chaleur avec l'extérieur).

H3. L'Univers en expansion homothétique (en expansion proportionnelle dans toutes ses dimensions) est pris comme ayant une géométrie sphérique avec un centre (eh oui c'est le modèle newtonien...).

H4. Sa masse volumique est uniquement fonction du temps et il y a conservation de la masse (et donc de l'énergie). Donc la quantité de matière y est constante!

H5. Nous acceptons la dynamique (approximation) newtonienne pour construire les modèles à suivre dans ce chapitre.

H6. L'origine du temps est assimilée à l'origine de création (horizon) de l'Univers et le référentiel d'étude est comobile aux particules (et se déplace donc avec les galaxies posées sur la trame de l'espace-temps) et appelé "**référentiel matériel**" (les galaxies sont donc immobiles dans ce référentiel!).

1.1. LOI DE HUBBLE

Sous l'hypothèse du principe cosmologique et des hypothèses précédentes, la distance d'un point origine O à un point M quelconque de l'Univers peut varier en fonction du temps (de manière indétectable à l'échelle humaine) sous la forme:

$$\overline{OM}(t) = \overline{OM}(t_0) \cdot F(t) \quad (51.1)$$

où $F(t)$ est le "**facteur d'échelle**" (noté $R(t)$ suivant le contexte...).

En écrivant cette relation, nous considérons que les points O et M sont sur un plan à courbure nulle. Effectivement, si nous imaginons deux points sur une surface courbe circulaire (par exemple la surface d'une sphère) voyons ce qui se passe:

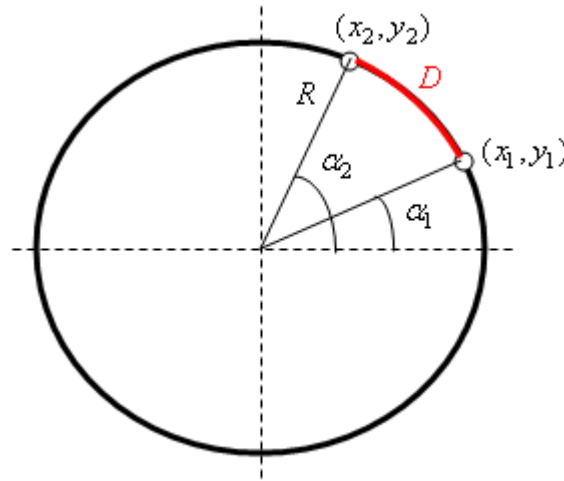


Figure: 51.1 - Illustration de la limite de validité du modèle

La distance entre deux points du cercle (in extenso de l'espace sphérique) est donnée par:

$$D = R \Delta \alpha = R(\alpha_2 - \alpha_1) = R \left(\arcsin \left(\frac{y_2}{R} \right) - \arcsin \left(\frac{y_1}{R} \right) \right) = R \left(\arccos \left(\frac{x_2}{R} \right) - \arccos \left(\frac{x_1}{R} \right) \right) \quad (51.2)$$

Nous voyons très bien dans cette relation que si le rayon (de l'Univers sphérique) change d'un facteur F , alors la variation de la distance entre les 2 points n'est pas linéairement proportionnelle à ce facteur!! Ce qui n'est pas le cas dans un plan à courbure nulle.

Conséquence: Notre modèle newtonien n'est valable que dans un Univers plat alors que la relativité générale ou une approche purement énergétique classique (voir plus loin) peut prendre en compte différents types de courbure !

Nous voyons tout de suite que la relation:

$$\overline{\vec{OM}}(t) = \overline{\vec{OM}}(t_o) \cdot F(t) \quad (51.3)$$

est indépendante de l'origine choisie, en effet, si nous l'appliquons à deux points A, B quelconques, nous avons:

$$\overline{\vec{OA}}(t) = \overline{\vec{OA}}(t_o) \cdot F(t) \quad \overline{\vec{OB}}(t) = \overline{\vec{OB}}(t_o) \cdot F(t) \quad (51.4)$$

Soit par différence:

$$\overline{\vec{AB}}(t) = \overline{\vec{AB}}(t_o) \cdot F(t) \quad (51.5)$$

Remarques:

R1. Au temps $t = t_0 = 0$ il est évident que la relation précédente s'écrit:

$$\overrightarrow{AB}(t_0) = \overrightarrow{AB}(t_0) \cdot F(t_0) \quad (51.6)$$

et nous impose $F(t_0) = F(0) = 1$. Cette remarque est importante et nous y reviendrons plusieurs fois pendant les développements qui vont suivre.

R2. La loi s'applique donc à un segment $\overrightarrow{AB}$ quelconque dans l'Univers. C'est pourquoi l'Univers ne comporte pas de centre géométrique et que nous pouvons nous donner une image suggestive pour se donner une idée de l'expansion de la trame de l'Univers: soit un ballon mi-gonflé sur la surface duquel nous traçons deux repères (par exemple: deux croix tracées à l'encre). En le gonflant davantage, nous constaterons que ces deux croix s'écartent l'une de l'autre et donc la distance qui les sépare, s'accroît. C'est ce que nous constatons avec les galaxies.

Dérivons par rapport au temps la relation:

$$\overrightarrow{OM}(t) = \overrightarrow{OM}(t_0) \cdot F(t) \quad (51.7)$$

Le premier membre donne alors la vitesse des particules (ou de tout autre objet) au point $\vec{r}(t) = \overrightarrow{OM}(t)$:

$$\vec{v}(\vec{r}, t) = \overrightarrow{OM}(t_0) \cdot d_t F(t) = \vec{r}(t_0) \cdot d_t F(t) \quad (51.8)$$

Soit en éliminant $\overrightarrow{OM}(t_0)$:

$$\vec{v}(\vec{r}, t) = \frac{\overrightarrow{OM}(t)}{F(t)} \cdot d_t F(t) = \frac{\vec{r}(t)}{F(t)} \cdot d_t F(t) = \frac{1}{F(t)} \cdot d_t F(t) \vec{r}(t) \quad (51.9)$$

Nous posons pour simplifier l'écriture:

$$H(t) = \frac{1}{F(t)} \cdot d_t F(t) \quad (51.10)$$

Nous avons donc:

$$\boxed{\vec{v}(\vec{r}, t) = H(t) \vec{r}(t)} \quad (51.11)$$

Cette relation est connue sous le nom de "**loi de Hubble**" (et qui d'après des recherches historiques devrait sa paternité plutôt à Georges Lemaître...).

Avant d'aller plus loin, il convient de s'arrêter sur cette équation pour l'instant présent t_0 :

$$\vec{v}(\vec{r}, t_0) = H(t_0) \vec{r}(t_0) = H_0 \vec{r}_0 \quad (51.12)$$

Cette équation dit que les objets de l'Univers s'éloignent avec une vitesse proportionnelle à leur

éloignement dans tous les points de l'Univers sans référentiel privilégié (aucune galaxie ne semble être fixe alors qu'elles le sont dans le référentiel matériel!).

Remarque: Cette relation permet d'avoir des vitesses supérieures à celles de la lumière. Mais cela n'est pas une violation de la relativité relativement à la constance de la vitesse de la lumière! Effectivement, il ne faut pas oublier que la loi de Hubble prend en compte l'expansion de la "trame" de l'espace-temps sur laquelle se meut la lumière. Dès lors si la trame s'étend selon un facteur d'expansion F supérieur à l'unité, cela donne l'impression que la lumière va plus vite que c et c'est ce qui donne des redshift parfois de 4 ou 5!

La constante $H(t_0)$ étant bien sûr identifiable à la "**constante de Hubble**" telle qu'elle est mesurée actuellement au début des années 2000 et valant environ: $75 [km \cdot s^{-1} / Mpc]$.

En unités S.I., puisqu'un mégaparsec vaut $\sim 3.085 \cdot 10^{22} [m]$ alors nous avons:

$$H_0 = 2.431 \cdot 10^{-18} [s^{-1}] \quad (51.13)$$

Ainsi, une estimation actuelle de l'âge (horizon) de l'Univers pourrait être interprétée comme l'inverse de la constante de Hubble qui donne le "**temps de Hubble**":

$$t_H = 1 / H_0 \cong 13 \cdot 10^9 \text{ années} \quad (51.14)$$

soit environ 13 milliards d'années (nous verrons une meilleure approche plus loin).

Inversement, nous pouvons nous amuser à calculer la distance à partir de laquelle nous pouvons atteindre la vitesse de la lumière $299'792'775 = 3997 [Mpc]$ avec la relation:

$$\vec{v}(\vec{r}, t_0) = H(t_0) \vec{r}(t_0) = H_0 \vec{r}_0 \quad (51.15)$$

et une application numérique donne grosso modo 13 milliards d'années-lumière. Telle est la distance de "**l'horizon cosmologique**".

1.2. ÉQUATIONS DE FRIEDMANN

Considérons maintenant un anneau sphérique de matière de rayon r et de masse constante m en expansion à la vitesse v , et contenant une boule de matière de masse M (elle aussi en expansion à la vitesse v).

Rappelons que selon le principe cosmologique, M et m n'ont pas la même densité ρ à cause de la masse constante de l'Univers.

Nous pouvons appliquer à ce système la conservation de l'énergie mécanique car il est isolé (c'est d'ailleurs le seul vrai système isolé...). Nous obtenons alors l'équation:

$$\frac{1}{2} m v^2 - \frac{GMm}{r} = k_1 \quad (51.16)$$

où k_1 est une constante. En divisant par m chaque membre et en remplaçant M par son expression en

fonction de la densité, nous obtenons:

$$\frac{1}{2}v^2 - \frac{4\pi G\rho r^3}{3r} = \frac{k_1}{m} \quad (51.17)$$

Remarque: Si cela peut aider le lecteur à comprendre ce que nous avons fait avec le terme de l'énergie potentielle, il peut se référer au chapitre de Mécanique Classique lorsque nous avons développé les calculs de l'énergie potentielle d'une sphère de matière.

Or la loi de Hubble nous donne selon ce que nous avons vu plus haut:

$$v = H \cdot r \quad (51.18)$$

et:

$$r = r_0 \cdot F \quad (51.19)$$

Nous obtenons:

$$\frac{1}{2}H^2 r_0^2 F^2 - \frac{4\pi G\rho r_0^2 F^2}{3} = \frac{k_1}{m} \quad (51.20)$$

que nous simplifions en:

$$H^2 = \frac{8\pi G\rho}{3} + \frac{2k_1}{mr_0^2} F^{-2} \quad (51.21)$$

Or, k_1, m, r_0 sont des constantes. Nous introduisons une nouvelle constante k définie par (afin de simplifier les écritures):

$$k = \frac{-2k_1}{mr_0^2} \quad (51.22)$$

Nous obtenons donc l'équation:

$$H^2 = \frac{8\pi G}{3}\rho - kF^{-2} \quad (51.23)$$

qui est n'est autre que la "**première équation de Friedmann**" que nous retrouvons fréquemment dans la littérature sous la forme suivante (parmi tant d'autres...):

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G\rho}{3} - \frac{k}{R^2} \quad (51.24)$$

Il est possible d'obtenir la même équation à partir de l'équation d'Einstein des champs (cf. chapitre de Relativité Générale) et la métrique de Friedmann-Robertson-Walker (je n'ai rien écrit sur cette métrique dans le chapitre de Relativité Générale car elle est à ce jour trop compliquée à comprendre pour moi).

Signalons quand même une forme très courante de cette dernière relation. Lorsqu'en utilisant l'équivalence masse et énergie d'Einstein la densité ρ n'est plus une densité massique mais une densité d'énergie, il faudra la diviser par la vitesse de la lumière au carré pour avoir à nouveau une densité massique. Il est en est de même si au dénominateur de la constante k , la masse m est remplacée par l'énergie, il faudra alors multiplier k par la vitesse de la lumière au carré. Nous avons alors en notant le rayon a (comme cela est souvent d'usage dans les ouvrages spécialisés) et en redistribuant les termes, la forme suivante de la première équation de Friedmann:

$$\left(\frac{a'}{a}\right)^2 + \frac{kc^2}{a^2} = \frac{8\pi G\rho}{3c^2} \quad (51.25)$$

Remarque: Einstein rajouta à cette équation pour des raisons de convictions personnelles et quasi religieuses une constante cosmologique Λ qui lui permettait de rendre statique le facteur d'échelle de l'Univers. Nous (les auteurs du site) rejetons cette constante arbitraire, même si dans la physique contemporaine, elle est revenue à la mode (sa valeur a été cependant définie mathématiquement plutôt que religieusement) car elle permettrait d'expliquer la provenance de la matière sombre, les lois actuelles de notre Univers, la période inflationniste de notre Univers ainsi que sa géométrie. Ainsi, la première équation de Friedmann avec cette constante cosmologique, qui est un total artifice de travail, s'écrit alors:

$$H^2 = \frac{8\pi G}{3}\rho - kF^{-2} + \frac{\Lambda}{3} \quad (51.26)$$

avec:

$$\Lambda = \frac{8\pi G}{c^4}F \quad (51.27)$$

C'est Andreï Sakharov qui a défini la valeur de cette constante cosmologique qui s'apparenterait soi-disant à l'énergie quantique du vide (fonction des champs de Higgs).

Deux idées guident les chercheurs de ce début de 21ème siècle: en physique quantique les équations du champ associées aux particules élémentaires servent à définir la théorie du Big Bang. La célèbre équation d'équivalence d'Einstein nous dit que l'énergie crée un champ gravitationnel comme l'électron en mouvement provoque un champ électromagnétique. Il découle de ces deux observations qu'en mesurant le champ gravitationnel nous avons un moyen de déterminer l'énergie du vide. Le champ gravitationnel ne concerne plus la matière mais bien la densité d'énergie du vide. Or la constante cosmologique est directement proportionnelle à la constante de la gravitation, G . Sa mesure est un jeu très dangereux car de sa valeur dépendent plusieurs lois fondamentales de physique et des propriétés non négligeables quant à la dynamique de notre Univers. Le débat reste donc complètement ouvert et si nous (les auteurs du site) trouvons une démonstration valable et rigoureuse de cette constante, nous mettrons à disposition du lecteur les conséquences de cette constante sur les modèles que nous allons voir ci-après.

Utilisons maintenant le premier principe de la thermodynamique (cf. [chapitre de Thermodynamique](#)) pour un système par définition fermé dont la somme de l'énergie cinétique et potentielle est constante (et donc la somme des variations est nulle pour ces deux énergies). Nous avons alors la variation

d'énergie totale qui n'est donnée que par la variation d'énergie interne (cas le plus courant en thermodynamique pour les objets macroscopiques):

$$d(E_{tot}) = d(U) \quad (51.28)$$

et nous avons également vu dans le chapitre de Thermodynamique l'équation caractéristique du fluide à l'équilibre:

$$d(U) = Td(S) - Pd(V) \quad (51.29)$$

Si le système est adiabatique (aucun transfert chaleur entre le système et l'extérieur), alors nous avons selon ce qui a été vu dans le chapitre de Thermodynamique:

$$dS = \frac{\delta Q}{T} = \frac{0}{T} = 0 \quad (51.30)$$

Donc:

$$d(U) + Pd(V) = 0 \quad (51.31)$$

Puisque l'Univers est supposé sphérique dans notre modèle, nous avons:

$$V = \frac{4}{3} \pi R^3 \quad (51.32)$$

et dans le référentiel matériel où les galaxies (particules du fluide cosmique) sont immobiles:

$$U = E_{tot} = mc^2 = \rho V c^2 \quad (51.33)$$

Soit:

$$d(\rho V c^2) + Pd\left(\frac{4}{3} \pi R^3\right) = d\left(\rho \frac{4}{3} \pi R^3 c^2\right) + Pd\left(\frac{4}{3} \pi R^3\right) = 0 \quad (51.34)$$

ce qui se simplifie en:

$$d(\rho R^3 c^2) + Pd(R^3) = 0 \quad (51.35)$$

en prenant la dérivée par rapport au temps cosmique t :

$$\begin{aligned} c^2 \frac{d(\rho R^3)}{dt} + P \frac{d(R^3)}{dt} &= c^2 \dot{\rho} R^3 + c^2 \rho 3R^2 \dot{R} + 3PR^2 \dot{R} \\ &= c^2 \dot{\rho} R^3 + 3R^2 (c^2 \rho + P) \dot{R} = 0 \end{aligned} \quad (51.36)$$

d'où:

$$\dot{\rho} = -3 \left(\rho + \frac{P}{c^2} \right) \frac{\dot{R}}{R} \quad (51.37)$$

Reprenons maintenant la première équation de Friedmann, obtenue plus haut, sous la forme:

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G\rho}{3} - \frac{k}{R^2} \quad (51.38)$$

et mettons-la sous la forme suivante:

$$\dot{R}^2 = \frac{8\pi G\rho R^2}{3} - k \quad (51.39)$$

Si nous différencions:

$$\frac{d}{dt}(\dot{R}^2) = \frac{d}{dt}\left(\frac{8\pi G\rho R^2}{3} - k\right) \quad (51.40)$$

Nous obtenons alors:

$$2\dot{R}\ddot{R} = \frac{8\pi G}{3} \frac{d}{dt}(\rho R^2) = \frac{8\pi G}{3} \left(R^2 \frac{d\rho}{dt} + \rho \frac{dR^2}{dt} \right) = \frac{8\pi G}{3} (R^2 \dot{\rho} + 2\rho R\dot{R}) \quad (51.41)$$

Injectons:

$$\dot{\rho} = -3\left(\rho + \frac{P}{c^2}\right) \frac{\dot{R}}{R} \quad (51.42)$$

dans la relation:

$$2\dot{R}\ddot{R} = \frac{8\pi G}{3} (\dot{\rho} R^2 + 2\rho R\dot{R}) \quad (51.43)$$

Nous obtenons alors:

$$2\dot{R}\ddot{R} = \frac{8\pi G}{3} \left(-3\left(\rho + \frac{P}{c^2}\right) \frac{\dot{R}}{R} R^2 + 2\rho R\dot{R} \right) = \frac{8\pi G}{3} R\dot{R} \left(-3\left(\rho + \frac{P}{c^2}\right) + 2\rho \right) \quad (51.44)$$

Soit:

$$\frac{\ddot{R}}{R} = \frac{4\pi G}{3} \left(-3\rho - 3\frac{P}{c^2} + 2\rho \right) = -\frac{4\pi G}{3} \left(\rho + 3\frac{P}{c^2} \right) \quad (51.45)$$

La relation suivante:

$$\boxed{\frac{\ddot{R}}{R} = -\frac{4\pi G}{3} \left(\rho + 3\frac{P}{c^2} \right)} \quad (51.46)$$

est la "**deuxième équation de Friedmann**" qui serait aussi appelée parfois "**équation de Raychaudhuri**".

1.2.1. DENSITÉ CRITIQUE

Revenons à notre première équation de Friedmann sans constante cosmologique. Nous avons donc démontré plus haut que:

$$H = \frac{1}{F} d_t F \quad (51.47)$$

Nous obtenons alors en injectant cette dernière relation dans la première équation de Friedmann la relation suivante:

$$(d_t F)^2 = \frac{8\pi G}{3} \rho F^2 - k \quad (51.48)$$

qui se réarrange avec:

$$\rho = \rho_0 \cdot F^{-3} \quad (51.49)$$

en:

$$(d_t F)^2 = \frac{8\pi}{3} G \rho_0 F^{-1} - k \quad (51.50)$$

L'exposant du terme de gauche impose que le terme de droite soit positif ou nul tel que:

$$\frac{8\pi}{3} G \rho_0 F^{-1} \geq k \quad (51.51)$$

Rappelons que les conditions initiales nous imposent qu'au temps $t_0 = 0$ nous ayons:

$$F(t_0) = 1 \text{ et } d_t F(t)|_{t=t_0} = H(t_0) \quad (51.52)$$

Effectivement:

$$H(t_0) = \frac{1}{F(t_0)} d_t F|_0 = d_t F|_0 \quad (51.53)$$

Il vient alors:

$$k = \frac{8\pi}{3} G \rho_0 - H_0^2 \quad (51.54)$$

Ce terme devrait être accessible à l'observation, hélas H_0^2 est très mal connu et ρ_0 encore plus.

Autrement dit, compte tenu du signe "-" dans l'expression de k , nous ne connaissons aujourd'hui même pas le signe de cette constante.

Cependant, il peut être important de noter qu'il existe une valeur ρ_0 appelée "**densité critique**" qui annule k et donc aussi (voir plus haut):

$$k = \frac{-2k_1}{mr_0^2} \quad (51.55)$$

ce qui implique que l'énergie totale de l'Univers serait nulle (selon des considérations de la cosmologie quantique).

Cette valeur de ρ_0 est donc trivialement:

$$\rho_0 = \frac{3H_0^2}{8\pi G} = \rho_c \quad (51.56)$$

Pour $H_0 = 75 [kg \cdot m \cdot s^{-1} / Mpc]$ (valeur actuelle) nous trouvons:

$$\rho_c = 10.57 \cdot 10^{-27} [kg \cdot m^{-3}] \quad (51.57)$$

À titre de comparaison, un atome d'hydrogène pèse $1.7 \cdot 10^{-27} [Kg]$, la densité critique correspondrait donc à six atomes d'hydrogène par mètre cube.

Les physiciens ont défini une constante (variant dans temps) notée par la lettre grecque Ω et appelée "**paramètre de densité cosmologique**" et donnée par le rapport des densités massiques (ou des densités énergétiques puisque le rapport sera la même!):

$$\Omega = \frac{\rho}{\rho_0} \quad (51.58)$$

souvent les astrophysiciens décomposent le paramètre de densité cosmologique en trois termes:

$$\Omega = \Omega_{matière} + \Omega_{radiation} + \Omega_{vide} = \Omega_m + \Omega_r + \Omega_\Lambda \quad (51.59)$$

Il est intéressant de travailler avec cette constante car dans le cas où:

- $\Omega = 1$:

Nous avons:

$$\rho_0 = \rho_c = \frac{3H_0^2}{8\pi G} \quad (51.60)$$

ce qui en remplaçant dans l'équation de Friedmann donne: $k = 0$ (un Univers plat comme nous le verrons dans notre étude du modèle relativiste).

- $\Omega > 1$:

En effectuant le même raisonnement, et toujours en inégalités, nous avons alors: $k > 0$ (un Univers à courbure positive (fermé) comme nous le verrons dans notre étude du modèle relativiste).

- $\Omega < 1$:

En effectuant le même raisonnement, mais en inégalités, nous avons alors: $k < 0$ (un Univers à courbure négative (ouvert) comme nous le verrons dans notre étude du modèle relativiste).

Ces trois situations peuvent se résumer géométriquement à:

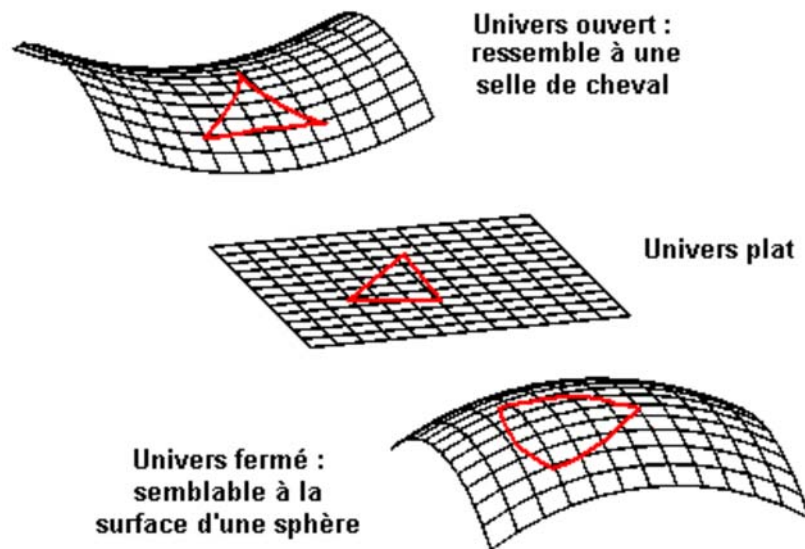


Figure: 51.2 - Illustration des différents types de courbure

Remarques:

R1. Toutes les mesures qui ont pu être faites jusqu'à présent n'ont pas permis de mettre en évidence une courbure de l'Univers. Les mesures du rayonnement fossile par le ballon BOOMERANG et le satellite COBE tendent cependant à accréditer l'hypothèse d'un univers plat relativement aux simulations numériques:

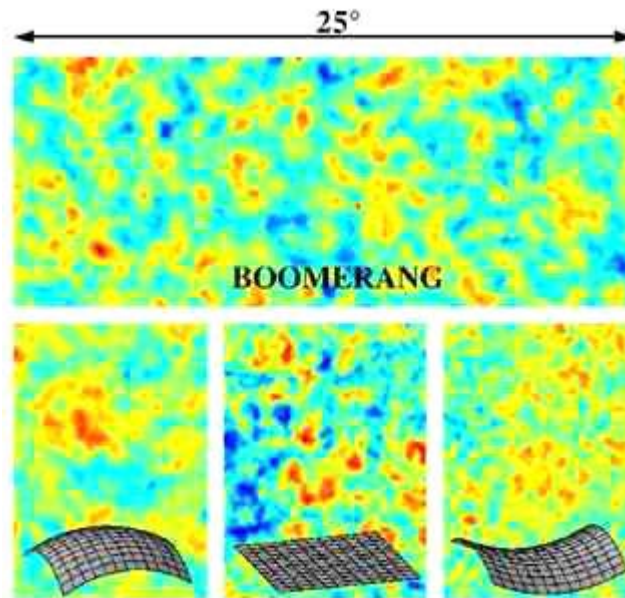


Figure: 51.3 - Illustration de ce que donnerait l'expérience en fonction du type de courbure

R2. La notion de topologie de l'Univers et son ouverture sont en fait normalement deux notions distinctes. Quand nous parlons d'Univers fermé ou ouvert nous ne parlons normalement pas de sa topologie mais de son destin. Ainsi, un Univers ouvert est en expansion indéfiniment et un Univers fermé se recontracte sur lui-même au bout d'un certain temps. Cela dit, dans les modèles que nous étudions dans ce chapitre (à constante cosmologique nulle), la courbure est directement liée à la densité, et donc à son ouverture.

Revenons à l'équation:

$$(d_t F)^2 = \frac{8\pi}{3} G \rho_0 F^{-1} - k \quad (51.61)$$

Nous pouvons écrire:

$$d_t F(t) = \sqrt{\frac{A}{F} - k} \quad (51.62)$$

En adoptant la notation:

$$A = \frac{8\pi}{3} G \rho_0 \quad (51.63)$$

Remarque: Les mesures actuelles donnent:

$$A \cong 0.2793923067 \cdot 10^{-35} \quad (51.64)$$

D'où:

$$dF(t) = \pm \sqrt{\frac{A}{F} - k} dt \quad (51.65)$$

Il convient maintenant pour nous de considérer trois situations:

$$k = 0 \quad k < 0 \quad k > 0 \quad (51.66)$$

qui correspondent donc respectivement en paramètre de densité cosmologique à:

$$\Omega = 1 \quad \Omega < 1 \quad \Omega > 1 \quad (51.67)$$

Remarque: Nous ne pouvons poser $\rho_0 = 0$ car dans nos hypothèses initiales se trouvait le principe de conservation de l'énergie.

2. MODÈLES COSMOLOGIQUES DE FRIEDMANN-LEMAÎTRE

Les modèles cosmologiques euclidiens de Friedmann-Lemaître consistent dans la limite newtonienne à étudier "**l'équation fondamentale des modèles de Friedmann**":

$$dF(t) = \pm \sqrt{\frac{A}{F} - k} dt \quad (51.68)$$

en considérant les trois situations:

$$k = 0 \quad k < 0 \quad k > 0 \quad (51.69)$$

Remarque: Il est possible dans le cadre de la relativité générale de trouver rigoureusement une solution aux équations d'Einstein des champs appelée "métrique de Robertson-Walker" qui dans le cas d'une approximation newtonienne nous redonne les équations de Friedmann obtenues dans le présent texte (souvent ce sont ces approximations qui sont utilisées dans la littérature car la solution exacte est hors de portée du cadre des cours universitaires traditionnels).

2.1. ESPACE PLAT (K=0)

Le modèle d'espace plat (euclidien) consiste à supposer que $k = 0$. Autrement dit, nous sommes dans un Univers dont la densité est dite "**densité critique**" ou également simplement "**plat**" (comme nous le verrons avec le modèle relativiste).

Nous avons alors l'équation:

$$dF(t) = \pm \sqrt{\frac{A}{F}} - k dt = \pm \sqrt{\frac{A}{F}} dt \quad (51.70)$$

En disposant les termes de manière adéquate:

$$\sqrt{F} dF(t) = \pm \sqrt{A} dt \quad (51.71)$$

et en intégrant, il vient:

$$\frac{2}{3} F^{3/2} = \pm \sqrt{A} t = \pm \sqrt{\frac{8\pi G \rho_0}{3}} t \quad (51.72)$$

Qui se simplifie en (nous élevons au carré d'où la suppression du double signe $\pm$):

$$F^3 = 6\pi G \rho_0 t^2 \quad (51.73)$$

Nous avons donc dans ce modèle la relation:

$$F = \sqrt[3]{6\pi G \rho_0} t^{2/3} \quad (51.74)$$

à laquelle il nous faut rajouter une constante pour avoir la condition correspondant à aujourd'hui:

$$F(t_0) = F(0) = 1 \quad (51.75)$$

qui reste satisfaite:

$$F(t) = \sqrt[3]{6\pi G \rho_0} t^{2/3} + 1 \quad (51.76)$$

Ce qui nous donne sur un tracé une fonction à l'allure suivante (ne pas se fier aux valeurs indiquées sur l'axe horizontal car elles sont arbitraires):

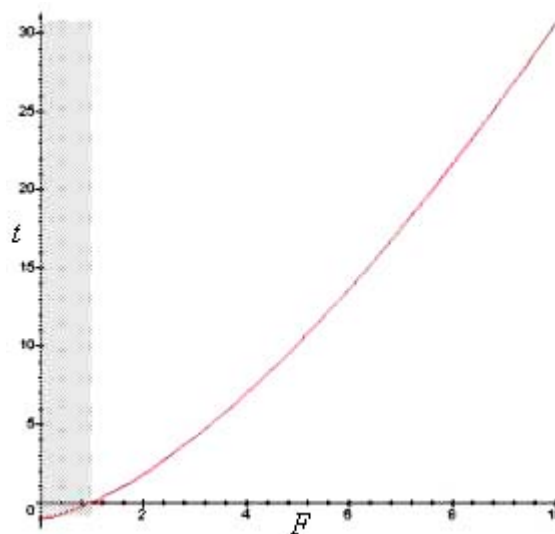


Figure: 51.4 - Évolution du facteur d'échelle pour un espace à courbure nulle

Nous avons mis la zone où $F(t) < 1$ en évidence pour bien rappeler que cette partie de la solution est à rejeter.

Nous avons donc un modèle d'Univers dont le facteur d'échelle croît de façon exponentielle et ce indéfiniment.

Remarque: Plus ρ_0 est grand, plus la croissance du facteur d'échelle est grande (sous-entendu que la pente est bien évidemment plus grande).

2.1.1. ESPACE PLAT DOMINÉ PAR LA MATIÈRE

Il existe également une autre approche beaucoup plus élégante et subtile à mon goût que la démonstration précédente (je ne l'ai découverte que de nombreuses années après avoir rédigé la version précédente). Elle a en plus l'avantage de mettre en évidence une hypothèse qui n'est pas apparue avec les développements précédents.

Nous partons de la première équation de Friedmann:

$$\left(\frac{R'}{R}\right)^2 = \frac{8\pi G\rho}{3} - \frac{k}{R^2} \quad (51.77)$$

en posant toujours k comme étant nul et ensuite nous utilisons l'astuce qui consiste à partir de la relation démontrée plus haut (utilisée pour démontrer la deuxième équation de Friedmann):

$$\dot{\rho} = -3\left(\rho + \frac{P}{c^2}\right)\frac{\dot{R}}{R} \quad (51.78)$$

d'imposer que la pression P du fluide (quel qu'il soit: gaz ou radiation!) soit nulle. Nous disons alors que l'Univers est un univers dominé par la matière et nous en déduisons:

$$\dot{\rho} + 3\rho\frac{R'}{R} = 0 \quad (51.79)$$

ce qui nous donne:

$$R^3\dot{\rho} + 3\rho R^2R' = 0 \Rightarrow \frac{d}{dt}(R^3\rho) = 0 \Rightarrow R^3\rho = R_0^3\rho_0 = c^{te} \quad (51.80)$$

Dans ces conditions, la première équation de Friedmann devient:

$$\frac{dR}{R} = \sqrt{\frac{8\pi G}{3} \left(\rho_0 \frac{R_0^3}{R^3}\right)} dt \quad (51.81)$$

en réarrangeant et en simplifiant, nous avons alors:

$$\int R^{1/2} dR = \int \sqrt{\frac{8\pi G\rho_0 R_0^3}{3}} dt \quad (51.82)$$

ce qui donne:

$$\frac{2}{3}R^{3/2} + c^{te} = \sqrt{\frac{8\pi G \rho_0 R_0^3}{3}} t \quad (51.83)$$

Au temps $t = 0$, nous avons le facteur d'échelle qui vaut $R = 0$ et donc la constante est nulle. Nous avons alors:

$$R(t) = \left(6\pi G \rho_0 R_0^3\right)^{1/3} t^{2/3} = \left(6\pi G \frac{3H_0^2}{8\pi G} R_0^3\right)^{1/3} t^{2/3} = \left(\frac{9H_0^2}{4} R_0^3\right)^{1/3} t^{2/3} \quad (51.84)$$

En posant qu'au temps $t = 0$, le facteur d'échelle était unitaire, cette dernière relation se simplifie:

$$R(t) = \left(\frac{9H_0^2}{4}\right)^{1/3} t^{2/3} \quad (51.85)$$

si nous posons que le facteur d'échelle est aujourd'hui pris comme référence unitaire, il vient alors:

$$t = \frac{2}{3H_0} \quad (51.86)$$

et en y remplaçant les valeurs numériques actuellement connues pour la constante de Hubble, il vient que l'Univers est actuellement âgé d'environ 8.6 milliards d'années (à comparer aux 13 milliards du temps de Hubble obtenu plus haut!).

2.1.2. ESPACE PLAT DOMINÉ PAR LA RADIATION

Nous avons démontré dans le chapitre de Thermodynamique lors de notre étude de la loi de Stefan-Boltzmann que la pression était liée à la densité d'énergie par la relation suivante:

$$P = \frac{1}{3} \rho_E \quad (51.87)$$

Pour un univers dominé par la radiation, la relation suivante démontrée plus haut:

$$\dot{\rho} = -3 \left(\rho + \frac{P}{c^2} \right) \frac{\dot{R}}{R} \quad (51.88)$$

exprimée avec une densité d'énergie et non une densité de masse devient:

$$\dot{\rho} c^2 = -3 \left(\rho c^2 + \frac{P c^2}{c^2} \right) \frac{\dot{R}}{R} \quad (51.89)$$

Soit:

$$\dot{\rho}_E = -3(\rho_E + P) \frac{\dot{R}}{R} \quad (51.90)$$

et en utilisant la relation liant pression et densité d'énergie, il vient:

$$\dot{\rho}_E = -3\left(\rho_E + \frac{\rho_E}{3}\right) \frac{\dot{R}}{R} \quad (51.91)$$

Après un petit réarrangement, nous avons:

$$\dot{\rho}_E + 4\rho_E \frac{\dot{R}}{R} = 0 \quad (51.92)$$

d'où nous tirons que:

$$\dot{\rho}_E + 4\rho_E \frac{\dot{R}}{R} = 0 \Rightarrow R^4 \dot{\rho}_E + 4\rho_E \dot{R} R^3 = 0 \Rightarrow \frac{d}{dt}(R^4 \rho_E) = 0 \Rightarrow R^4 \rho_E = R_0^4 \rho_{E,0} = c^{te} \quad (51.93)$$

Dans ces conditions, la première équation de Friedmann:

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G \rho}{3} - \frac{k}{R^2} \quad (51.94)$$

devient d'abord en passant en densité d'énergie et en posant k comme étant nul:

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G \rho_E}{3c^2} \quad (51.95)$$

et nous pouvons donc substituer la densité d'énergie par le résultant obtenu juste précédemment:

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G \rho_{E,0} R_0^4}{3c^2} \left(\frac{R_0}{R}\right)^4 \quad (51.96)$$

Nous avons alors:

$$\frac{dR}{dt} = \sqrt{\frac{8\pi G \rho_{E,0} R_0^4}{3c^2}} \frac{1}{R} \quad (51.97)$$

d'où:

$$R dR = \sqrt{\frac{8\pi G \rho_{E,0} R_0^4}{3c^2}} dt \quad (51.98)$$

La primitive donne donc:

$$2R^2 + c^2t = \sqrt{\frac{8\pi G \rho_{E,0} R_0^4}{3c^2}} t \quad (51.99)$$

Au moment du Big Bang en $t = 0$, nous avons le facteur d'échelle qui vaut $R = 0$ et donc la constante est nulle. Nous avons alors:

$$2R^2 = \sqrt{\frac{8\pi G \rho_{E,0} R_0^4}{3c^2}} t \quad (51.100)$$

En posant qu'au temps $t = 0$, le facteur d'échelle était unitaire, cette dernière relation se simplifie:

$$2R^2 = \sqrt{\frac{8\pi G \rho_{E,0}}{3c^2}} t \quad (51.101)$$

Dès lors, il vient:

$$R(t) = \left(\frac{2\pi G \rho_{E,0}}{3c^2} \right)^{1/4} t^{1/2} = \left(\frac{2\pi G}{3} \rho_0 \right)^{1/4} t^{1/2} = \left(\frac{2\pi G}{3} \frac{3H_0^2}{8\pi G} \right)^{1/4} t^{1/2} \quad (51.102)$$

Soit après simplification:

$$R(t) = \left(\frac{H_0}{2} \right)^{1/2} t^{1/2} \quad (51.103)$$

Ainsi, un univers plat dominé par la radiation a son facteur d'échelle qui croît légèrement plus lentement qu'un univers plat dominé par la matière.

Pour comparaison avec Maple 4.00b (en bleu: Univers plat dominé par la matière, en rouge: Univers plat dominé par la radiation):

```
>plot([t^(2/3),t^(1/2)],t=0..2*Pi,0..3,color=[blue,red]);
```

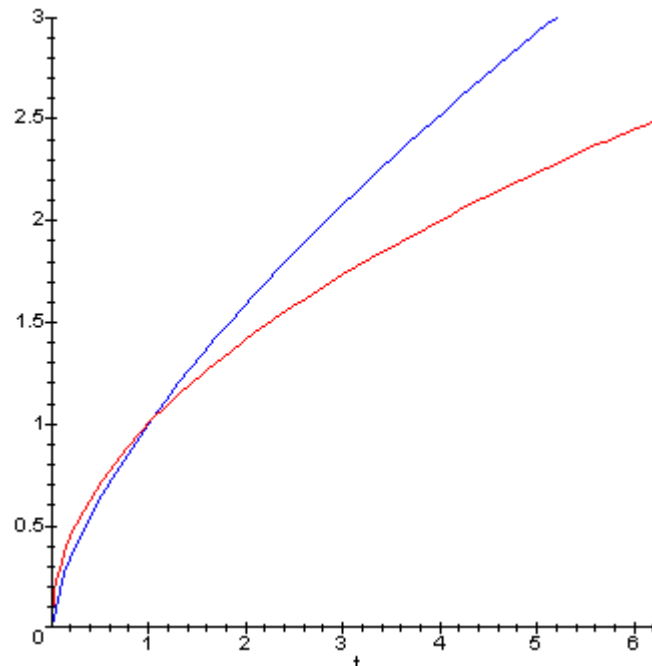


Figure: 51.5 - Évolution du facteur R pour un espace à courbure 0 dominé par la matière ou par la radiation

2.2. MODÈLE SPHÉRIQUE ($K>0$)

Dans ce modèle (appelé aussi parfois "modèle elliptique"), nous considérons $k > 0$. Donc l'équation à traiter reste:

$$dF(t) = \pm \sqrt{\frac{A}{F} - k} dt \quad (51.104)$$

Ce qui s'écrit aussi:

$$\Delta t = \pm \int_{t_0}^t \frac{1}{\sqrt{\frac{A}{F} - k}} dF(t) \geq 0 \quad (51.105)$$

Rappelons que nous avons supposé pour $t = t_0$ que $F(t_0) = 1$ si nous effectuons le changement de variable $U = 1/F(t)$, nous obtenons l'intégrale suivante:

$$\Delta t = \pm \int_1^{1/F} \frac{dU}{U^2 \sqrt{AU - k}} \quad (51.106)$$

Nous recherchons donc une primitive de:

$$\Delta t = \int \frac{dU}{U^2 \sqrt{AU - k}} \quad (51.107)$$

et nous discuterons du signe $\pm$ après avoir trouvé la primitive.

Nous effectuons encore un changement de variable en posant $x^2 = AU - k$ donc $2x dx = A dU$ ce qui nous donne la primitive suivante à calculer:

$$\frac{2}{A} \int \frac{1}{U^2} dx = \frac{2A^2}{A} \int \frac{1}{(x^2 + k)^2} dx = \frac{2A}{k^2} \int \frac{1}{\left(\left(\frac{x}{\sqrt{k}} \right)^2 + 1 \right)^2} dx \quad (51.108)$$

en refaisant un changement de variable:

$$y = \frac{x}{\sqrt{k}} \Rightarrow dx = \sqrt{k} dy \quad (51.109)$$

d'où à une constante multiplicative près:

$$\frac{2A}{k^{3/2}} \quad (51.110)$$

nous avons:

$$\Delta t = \int \frac{dy}{(1+y^2)^2} \quad (51.111)$$

Dans le chapitre de Calcul Différentiel Et Intégral nous avons vu que cette forme de primitive se résout par la relation (nous rajoutons la constante d'intégration à la fin car nous faisons de la physique et il faut satisfaire des conditions initiales auxquelles nous ne nous intéressons pas nécessairement en mathématiques):

$$2nI_{n+1} = \frac{y}{(1+y^2)^n} + (2n-1)I_n + c^{te} \quad (51.112)$$

avec:

$$I_n = \int \frac{dy}{(1+y^2)^n} \quad (51.113)$$

d'où:

$$I_2 = \frac{1}{2} \frac{y}{(1+y^2)} + \frac{1}{2} I_1 \quad (51.114)$$

Il nous faut encore calculer I_1 (cf. chapitre de Calcul Différentiel Et Intégral):

$$I_1 = \int \frac{1}{(1+y^2)} dy = \arctan(y) + c^{te} \quad (51.115)$$

Enfin:

$$I_2 = \frac{1}{2} \frac{y}{(1+y^2)} + \frac{1}{2} \arctan(y) + c^{te} \quad (51.116)$$

en remettant en place tous les changements de variables et en introduisant à nouveau la constante multiplicative, nous avons dans le cas où $k > 0$:

$$I_2 = \frac{\sqrt{AU - k}}{kU} + \frac{A}{k^{3/2}} \arctan\left(\frac{\sqrt{AU - k}}{\sqrt{k}}\right) + \frac{2A}{k^{3/2}} c^{te} \quad (51.117)$$

Entre les deux bornes d'intégration $(1/F, 1)$ nous avons donc (la constante d'intégration s'annule et nous reprenons le $\pm$ qui se trouvait initialement dans l'intégrale):

$$\Delta t = \pm \left(\frac{F\sqrt{A/F - k}}{k} + \frac{A}{k^{3/2}} \arctan\left(\frac{\sqrt{A/F - k}}{\sqrt{k}}\right) \right) - \left(\frac{\sqrt{A - k}}{k} + \frac{A}{k^{3/2}} \arctan\left(\frac{\sqrt{A - k}}{\sqrt{k}}\right) \right) \quad (51.118)$$

où rappelons que la théorie nous impose $\Delta t \geq 0$.

Si nous traçons cette fonction pour une valeur $k > 0$ fixe. Nous avons le tracé suivant dans Maple 4.00b (nous ne considérerons que le cas avec le signe "-" ci-dessous pour l'instant car le signe "+" nous donnerait un tracé dans les différentiels de temps négatifs: $\Delta t < 0$):

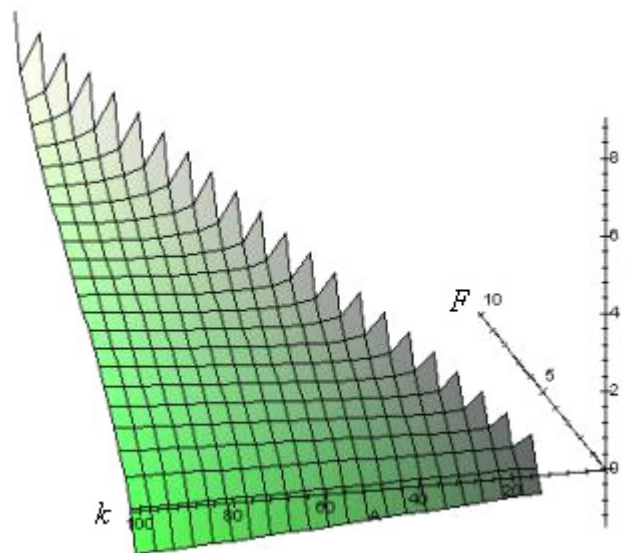


Figure: 51.6 - Évolution de facteur d'échelle pour un espace à courbure positive

Remarque: Le temps est ici toujours représenté sur l'axe vertical ainsi que pour tous les diagrammes suivants (il vous faut tourner un peu la tête si habituellement vous mettez le temps sur l'axe des abscisses...).

Nous voyons que plus la constante A est petite, plus l'Univers arrive rapidement à une valeur finale. De plus pour une valeur de k fixée, certaines valeurs de A sont interdites (c'est à cause de la condition d'intégration).

En fixant une valeur de A , nous obtenons la représentation bidimensionnelle suivante:

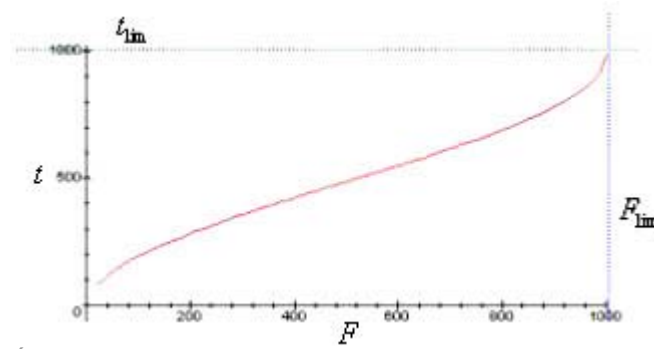


Figure: 51.7 - Évolution particulière du facteur d'échelle pour un espace à courbure positive

Si nous effectuons un zoom au niveau $t = t_0 = 0$, nous avons:

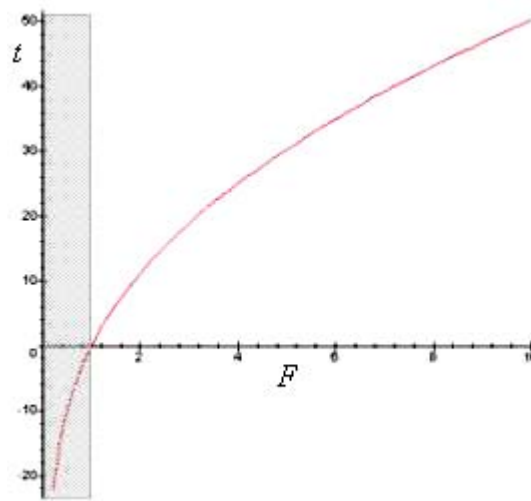


Figure: 51.8 - Évolution particulière du facteur d'échelle pour un espace à courbure positive (zoom)

Nous voyons que le critère $F(t_0) = F(0) = 1$ est parfaitement et naturellement respecté sans introduction d'une quelconque constante. Il suffit par ailleurs de remplacer F par 1 dans l'équation que nous avons obtenue pour voir que nous trouvons $\Delta t = 0$.

Remarque: Comme nous l'avons déjà précisé, toutes les valeurs de $F(t)$ inférieures à 1 sont à rejeter!

Analysons l'avant-dernier tracé en rappelant que:

$$(d_t F)^2 = \frac{8\pi}{3} G \rho_0 F^{-1} - k \quad (51.119)$$

Une condition limite (condition d'intégration) pour que le terme de droite de l'égalité soit positif est que:

$$F_{\text{lim}} = \frac{8\pi G \rho_0}{3k} \quad \text{ou} \quad F_{\text{lim}}^{-1} = \frac{3k}{8\pi G \rho_0} = \frac{k}{A} \quad (51.120)$$

Donc, si F^{-1} est plus petit que F_{lim}^{-1} , nous ne sommes plus dans un domaine valable (réel) du modèle.

Il faut donc que:

$$F^{-1} \geq F_{\text{lim}}^{-1} \text{ ou } F \leq F_{\text{lim}} \quad (51.121)$$

Cette limite a été représentée par une ligne verticale bleue sur l'avant-dernier tracé. Nous y avons également représenté par une ligne horizontale verte la limite temporelle t_{lim} correspondante à F_{lim} .

Au fait, au-delà de cette limite temporelle, ce que ne sait pas l'ordinateur qui a tracé notre fonction, c'est qu'il devrait basculer sur la fonction d'échelle avec le signe "+". Ainsi, lorsque nous exécutons le tracé des deux fonctions avec les bornes adéquates:

$$\begin{aligned} \Delta t &= -f(F) \Big|_{t_0}^{t_{\text{lim}}} \\ \Delta t &= +f(F) \Big|_{t_{\text{lim}}}^{t_f} \end{aligned} \quad (51.122)$$

nous obtenons alors (le temps est représenté sur l'axe vertical!):

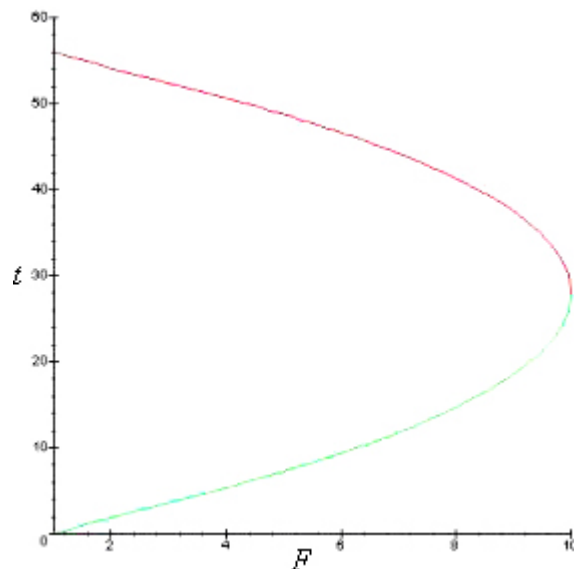


Figure: 51.9 - Évolution particulière du facteur d'échelle pour un espace à courbure positive

Nous voyons que alors que pour $t > t_{\text{lim}}$ l'Univers entre dans une phase de contraction que nous appelons communément "**Big Crunch**". Après cette phase de rétraction, il est possible soit que l'Univers disparaisse totalement, soit qu'il entre à nouveau dans une phase dynamique cyclique (mathématiquement les deux issues sont possibles).

2.2.1. ESPACE SPHÉRIQUE DOMINÉ PAR LA MATIÈRE

Au même titre que pour le modèle à espace plat, il existe également une autre approche beaucoup plus élégante et subtile à mon goût que la démonstration précédente (je ne l'ai découverte aussi que de nombreuses années après avoir écrit le texte précédent). Elle a l'avantage de mettre en évidence une hypothèse qui n'est pas apparue avec les développements précédents et permet de tracer plus simplement dans Maple 4.00b le comportement du facteur d'échelle de l'Univers. On retrouve ainsi exactement le fameux graphique représentant l'évolution du facteur d'échelle de l'Univers disponible dans la quasi-totalité des ouvrages de vulgarisation sur le sujet.

Nous partons toujours de la première équation de Friedmann:

$$\left(\frac{R'}{R}\right)^2 = \frac{8\pi G \rho}{3} - \frac{k}{R^2} \quad (51.123)$$

Il est d'usage pour ce modèle de poser $k = 1$ (quitte à prendre tout nombre positif au moins en prendre un qui est sympathique...) et nous avons montré que lorsque la matière domine, nous avons:

$$R^3 \rho = R_0^3 \rho_0 = c^{te} \quad (51.124)$$

Dès lors:

$$\frac{R'^2}{R^2} = \frac{8\pi G}{3} \rho_0 \left(\frac{R_0}{R}\right)^3 - \frac{1}{R^2} \quad (51.125)$$

et il vient immédiatement:

$$\frac{dR}{dt} = \sqrt{\frac{8\pi G \rho_0 R_0^3}{3R} - 1} \quad (51.126)$$

Soit:

$$\int dt = \int \frac{1}{\sqrt{\frac{8\pi G \rho_0 R_0^3}{3R} - 1}} dR \quad (51.127)$$

Si nous passons en temps comobile (déjà vulgarisé tout au début de ce chapitre) défini mathématiquement par:

$$d\eta = \frac{dt}{R} \quad (51.128)$$

Nous avons alors:

$$\int d\eta = \int \frac{1}{\sqrt{\frac{8\pi G \rho_0 R_0^3}{3} R - R^2}} dR \quad (51.129)$$

Notons cela sous la forme:

$$\int d\eta = \int \frac{1}{\sqrt{A \cdot R - R^2}} dR \quad (51.130)$$

où A est donc strictement positif. Faisons-y la substitution:

$$R = x + \frac{A}{2} \quad (51.131)$$

nous avons alors:

$$\begin{aligned} \int d\eta &= \int \frac{1}{\sqrt{A \cdot \left(x + \frac{A}{2}\right) - \left(x + \frac{A}{2}\right)^2}} dx = \int \frac{1}{\sqrt{Ax + \frac{A^2}{2} - \left(x^2 + Ax + \frac{A^2}{4}\right)}} dx \\ &= \int \frac{1}{\sqrt{\frac{A^2}{4} - x^2}} dx \end{aligned} \quad (51.132)$$

et nous avons démontré dans le chapitre de Calcul Différentiel Et Intégral que la primitive est:

$$\int \frac{1}{\sqrt{\frac{A^2}{4} - x^2}} dx = \arcsin\left(\frac{2x}{A}\right) + c^{te} \quad (51.133)$$

Donc:

$$\eta = \arcsin\left(\frac{2R - A}{A}\right) + c^{te} \quad (51.134)$$

Pour qu'au temps $\eta = 0$ nous ayons $R = 0$, il faut que la constante soit telle que:

$$\eta = \arcsin\left(\frac{2R - A}{A}\right) + \frac{\pi}{2} \quad (51.135)$$

Dès lors:

$$\frac{2R - A}{A} = \sin\left(\eta - \frac{\pi}{2}\right) = -\cos(\eta) \quad (51.136)$$

D'où:

$$R = \frac{A}{2}(1 - \cos(\eta)) \quad (51.137)$$

Maintenant, rappelons que:

$$d\eta = \frac{dt}{R} \Rightarrow dt = R d\eta \quad (51.138)$$

Dès lors:

$$\int dt = \int R d\eta = \int \frac{A}{2}(1 - \cos(\eta)) d\eta \quad (51.139)$$

Soit:

$$t = \frac{A}{2}(\eta - \sin(\eta)) + c^{te} \quad (51.140)$$

et comme nous devons avoir au temps $t = 0$ le temps comobile qui est aussi nul, la constante est donc nulle. Dès lors nous avons au final le système paramétrique suivant:

$$\begin{aligned} R &= \frac{A}{2}(1 - \cos(\eta)) \\ t &= \frac{A}{2}(\eta - \sin(\eta)) \end{aligned} \quad (51.141)$$

avec Maple 4.00b nous avons alors en comparant l'Univers plat dominé par la matière (en bleu), l'Univers plat dominé par la radiation (en rouge) et enfin l'Univers à courbure positive dominé par la matière (vert) et en mettant des coefficients artificiels pour mieux distinguer les tracés:

```
>plot([t^(2/3),t^(1/2),[0.5*(t-sin(t)),0.5*(1-cos(t)),t=0..2*Pi],
t=0...Pi,0..3,color=[blue,red,green]);
```

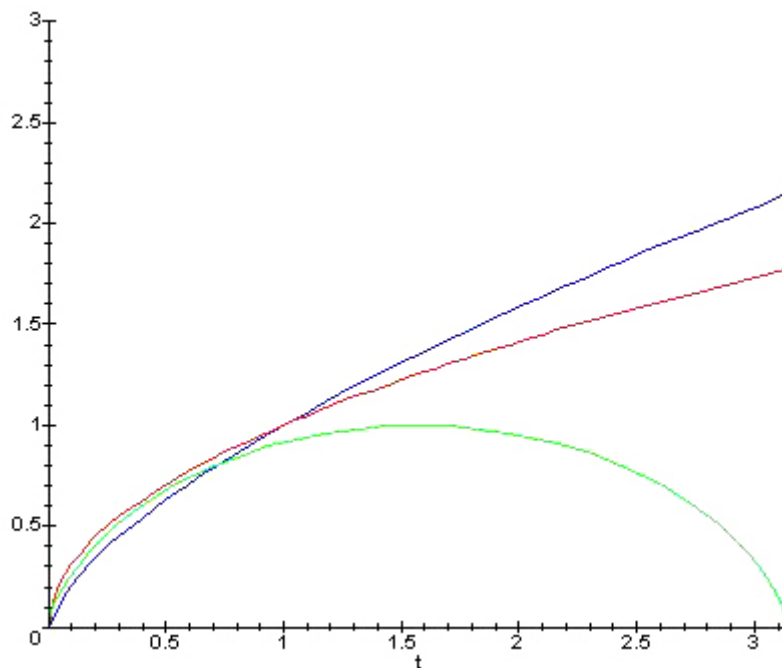


Figure: 51.10 - Évolution du facteur R pour les configurations d'espace obtenues étudiées jusqu'à présent

Nous comprenons alors mieux pourquoi le modèle d'Univers à courbure positive est aussi considéré comme un modèle d'Univers fermé.

2.2.2. ESPACE SPHÉRIQUE DOMINÉ PAR LA RADIATION

Considérons maintenant un univers dominé par la radiation. Nous avons démontré que dans cette situation nous avons:

$$P = \frac{1}{3} \rho_E \quad (51.142)$$

et:

$$R^4 \rho_E = R_0^4 \rho_{E,0} = c^{\text{te}} \quad (51.143)$$

Dans ce cas l'équation de Friedmann en termes de densité d'énergie s'écrit en posant $k = 1$:

$$\left(\frac{R'}{R}\right)^2 = \frac{8\pi G \rho_E}{3c^2} - \frac{1}{R^2} \quad (51.144)$$

Ce qui devient:

$$\frac{dR}{dt} = \sqrt{\frac{8\pi G \rho_E}{3c^2} R^2 - 1} \quad (51.145)$$

En injectant $R^4 \rho_E = R_0^4 \rho_{E,0}$, il vient:

$$\frac{dR}{dt} = \sqrt{\frac{8\pi G \rho_{E,0} R_0^4}{3c^2 R^2} - 1} \quad (51.146)$$

Soit:

$$\int dt = \int \frac{1}{\sqrt{\frac{8\pi G \rho_{E,0} R_0^4}{3c^2 R^2} - 1}} dR \quad (51.147)$$

Notons cela sous la forme:

$$\int dt = \int \frac{1}{\sqrt{\frac{A}{R^2} - 1}} dR \quad (51.148)$$

Si nous passons aussi en temps comobile:

$$d\eta = \frac{dt}{R} \quad (51.149)$$

Nous avons alors:

$$\int d\eta = \int \frac{1}{\sqrt{A - R^2}} dR \quad (51.150)$$

Dans le chapitre de Calcul Différentiel Et Intégral nous avons démontré comment déterminer exactement la même primitive (car il s'agit d'une primitive usuelle). Nous avons:

$$\eta = \arcsin\left(\frac{R}{\sqrt{A}}\right) + c^{te} \quad (51.151)$$

Pour qu'au temps $\eta = 0$ nous ayons $R = 0$, il faut que la constante soit telle que:

$$\eta = \arcsin\left(\frac{R}{\sqrt{A}}\right) \quad (51.152)$$

D'où:

$$R = \sqrt{A} \sin(\eta) \quad (51.153)$$

Maintenant, rappelons que:

$$d\eta = \frac{dt}{R} \Rightarrow dt = R d\eta \quad (51.154)$$

Dès lors:

$$\int dt = \int R d\eta = \int \sqrt{A} \sin(\eta) d\eta \quad (51.155)$$

Soit:

$$t = -\sqrt{A} \cos(\eta) + c^{te} \quad (51.156)$$

et comme nous devons avoir au temps $t = 0$ le temps comobile qui est aussi nul, la constante vaut donc $\sqrt{A}$. Dès lors nous avons au final le système paramétrique suivant:

$$\begin{aligned} R &= \sqrt{A} \sin(\eta) \\ t &= \sqrt{A} (1 - \cos(\eta)) \end{aligned} \quad (51.157)$$

avec Maple 4.00b nous avons alors en comparant l'Univers plat dominé par la matière (en bleu), l'Univers plat dominé par la radiation (en rouge), l'Univers à courbure positive dominé par la matière (vert), l'Univers à courbure positive dominé par la radiation (noir):

```
>plot([t^(2/3),t^(1/2),[0.5*(t-sin(t)),0.5*(1-cos(t)),t=0..2*Pi],
[0.5*(1-cos(t)),0.5*sin(t),t=0..2*Pi]],t=0...Pi,0..3,color=[blue,red,green,black]);
```

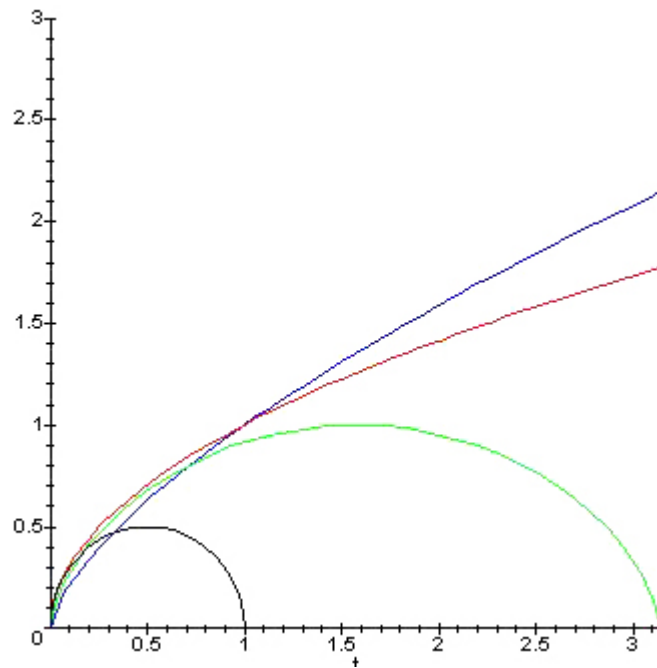


Figure: 51.11 - Évolution du facteur R pour un espace à courbure $0/+$ dominé par la matière ou par la radiation

Nous avons donc meilleur intérêt à être dans un Univers sphérique dominé par la matière (ou un mélange matière-radiation)...

2.3. MODÈLE HYPERBOLIQUE ($K < 0$)

Dans ce modèle, nous considérons $k < 0$. Donc l'équation à traiter peut s'écrire:

$$dF(t) = \pm \sqrt{\frac{A}{F} + |k|} dt \quad (51.158)$$

Ce qui s'écrit aussi:

$$\Delta t = \pm \int_{t_0}^t \frac{1}{\sqrt{\frac{A}{F} + |k|}} dF(t) \quad (51.159)$$

Rappelons que nous avons supposé pour $t = t_0$ que $F(t_0) = 1$. Si nous effectuons le changement de variable $U = 1/F(t)$, nous obtenons l'intégrale suivante:

$$\Delta t = \pm \int \frac{1/F}{U^2 \sqrt{AU + |k|}} dU \geq 0 \quad (51.160)$$

Nous recherchons donc une primitive de:

$$\Delta t = \int \frac{dU}{U^2 \sqrt{AU + |k|}} \quad (51.161)$$

et nous discuterons du signe $\pm$ après avoir trouvé la primitive.

Nous effectuons encore un changement de variable en posant $x^2 = AU + k$ donc $2xdx = AdU$ ce qui nous donne la primitive suivante à calculer:

$$\frac{2}{A} \int \frac{1}{U^2} dx = \frac{2A^2}{A} \int \frac{1}{(x^2 - k)^2} dx = \frac{2A}{k^2} \int \frac{1}{\left(\left(\frac{x}{\sqrt{k}}\right)^2 - 1\right)^2} dx \quad (51.162)$$

en refaisant un changement de variable:

$$y = \frac{x}{\sqrt{k}} \Rightarrow dx = \sqrt{k} dy \quad (51.163)$$

d'où à une constante multiplicative près:

$$\frac{2A}{k^{3/2}} \quad (51.164)$$

nous avons:

$$\int \frac{dy}{(y^2 - 1)^2} = \int \frac{dy}{(1 - y^2)^2} \quad (51.165)$$

Dans le chapitre de Calcul Différentiel Et Intégral nous avons vu que cette forme de primitive se résout par la relation (nous rajoutons la constante d'intégration à la fin car nous faisons de la physique et il faut satisfaire des conditions initiales auxquelles nous ne nous intéressons pas nécessairement en mathématiques):

$$2nI_{n+1} = \frac{y}{(1-y^2)^n} + (2n-1)I_n + c^{te} \quad (51.166)$$

avec:

$$I_n = \int \frac{dy}{(1-y^2)^n} \quad (51.167)$$

d'où:

$$I_2 = \frac{1}{2} \frac{y}{(1-y^2)} + \frac{1}{2} I_1 \quad (51.168)$$

Il nous faut encore calculer I_1 :

$$I_1 = \int \frac{1}{(1-y^2)} dy = \frac{1}{2} \int \left(\frac{1}{1+y} + \frac{1}{1-y} \right) dy = \frac{1}{2} \ln|1+y| - \ln|1-y| = \frac{1}{2} \ln \left| \frac{1+y}{1-y} \right| + c^{te} \quad (51.169)$$

Enfin:

$$I_2 = \frac{1}{2} \frac{y}{(1-y^2)} + \frac{1}{4} \ln \left| \frac{1+y}{1-y} \right| + c^{\pm} \quad (51.170)$$

en remettant en place tous les changements de variables et en introduisant à nouveau la constante multiplicative, nous avons dans le cas où $k > 0$:

$$I_2 = -\frac{\sqrt{AU + |k|}}{kU} + \frac{A}{2|k|^{3/2}} \ln \left| \frac{\sqrt{|k|} + \sqrt{AU + |k|}}{\sqrt{|k|} - \sqrt{AU + |k|}} \right| + \frac{2A}{|k|^{3/2}} c^{\pm} \quad (51.171)$$

Entre les deux bornes d'intégration $(1/F, 1)$ nous avons donc (la constante d'intégration s'annule):

$$\Delta t = \left(\frac{-F \sqrt{A/F + |k|}}{k} + \frac{A}{2|k|^{3/2}} \ln \left| \frac{\sqrt{|k|} + \sqrt{A/F + |k|}}{\sqrt{|k|} - \sqrt{A/F + |k|}} \right| \right) - \left(\frac{-\sqrt{A + |k|}}{k} + \frac{A}{2|k|^{3/2}} \ln \left| \frac{\sqrt{|k|} + \sqrt{A + |k|}}{\sqrt{|k|} - \sqrt{A + |k|}} \right| \right) \quad (51.172)$$

Nous devons évidemment avoir (nous reprenons le $\pm$ qui se trouvait initialement dans l'intégrale):

$$\Delta t = \pm \left(\frac{-F \sqrt{A/F + |k|} + \sqrt{A + |k|}}{k} + \frac{A}{2|k|^{3/2}} \ln \left| \frac{\frac{\sqrt{|k|} + \sqrt{A/F + |k|}}{\sqrt{|k|} - \sqrt{A/F + |k|}}}{\frac{\sqrt{|k|} + \sqrt{A + |k|}}{\sqrt{|k|} - \sqrt{A + |k|}}} \right| \right) \geq 0 \quad (51.173)$$

Si nous traçons cette fonction pour une valeur $k > 0$ fixe. Nous avons le tracé suivant dans Maple 4.00b (nous ne considérerons que le cas avec le signe "-" car celui avec le signe "+" n'a pas de sens physique même translaté):

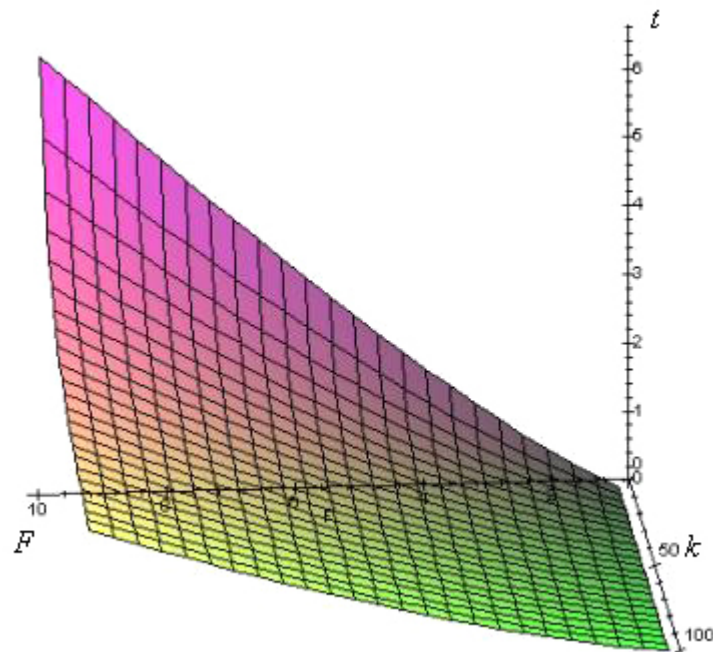


Figure: 51.12 - Évolution du facteur d'échelle pour un espace à courbure négative

Nous voyons que plus la constante A est petite, plus l'Univers croît indéfiniment rapidement. De plus pour une valeur de k fixée, certaines valeurs de A sont interdites (il s'agit au toujours fait de la condition d'intégration).

Nous voyons à nouveau que le critère $F(t_0) = F(0) = 1$ est naturellement parfaitement respecté. Toutes les valeurs de $F(t)$ inférieures à 1 sont à rejeter !

Nous avons donc dans ce modèle hyperbolique un Univers qui croît indéfiniment de façon exponentielle (comme le modèle plat de Friedmann-Lemaître) car étant donné que $k < 0$, il n'y a plus de condition limite d'intégration (contrairement au modèle elliptique précédent).

2.3.1. ESPACE HYPERBOLIQUE DOMINÉ PAR LA MATIÈRE

Au même titre que pour les modèles à espace plat et sphérique, il existe également une autre approche beaucoup plus élégante et subtile à mon goût que la démonstration précédente (je ne l'ai découverte aussi que de nombreuses années après avoir écrit la version précédente). Elle aussi l'avantage de mettre en évidence une hypothèse qui n'est pas apparue avec les développements précédents et permet de tracer plus simplement dans Maple 4.00b le comportement du facteur d'échelle de l'Univers. On retrouve ainsi exactement le fameux graphique représentant l'évolution du facteur d'échelle de l'Univers disponible dans la quasi-totalité des ouvrages de vulgarisation sur le sujet.

Nous partons toujours de la première équation de Friedmann:

$$\left(\frac{R'}{R}\right)^2 = \frac{8\pi G\rho}{3} - \frac{k}{R^2} \quad (51.174)$$

Il est d'usage pour ce modèle de poser $k = -1$ (quitte à prendre tout nombre positif au moins en prendre un qui est sympathique...) et nous avons montré que lorsque la radiation domine, nous avons:

$$R^3 \rho = R_0^3 \rho_0 = c^{te} \quad (51.175)$$

Dans ce cas l'équation de Friedmann en termes de densité d'énergie s'écrit en posant $k = -1$:

$$\left(\frac{R'}{R}\right)^2 = \frac{8\pi G\rho_E}{3c^2} - \frac{1}{R^2} \quad (51.176)$$

La première équation de Friedmann devient alors:

$$\left(\frac{R'}{R}\right) = \frac{8\pi G\rho}{3} + \frac{1}{R^2} = \frac{8\pi G\rho_0}{3} \left(\frac{R_0}{R}\right)^3 + \frac{1}{R^2} \quad (51.177)$$

Nous avons alors:

$$\frac{dR}{dt} = \sqrt{\frac{8\pi G\rho_0 R_0^3}{3R} + 1} \quad (51.178)$$

d'où:

$$\int dt = \int \frac{1}{\sqrt{\frac{8\pi G \rho_0 R_0^3}{3R} + 1}} dR \quad (51.179)$$

Il s'agit exactement de la même intégrale que celle de l'univers sphérique dominé par la matière à la différence que sous la racine, nous avons +1 au lieu de -1. Nous allons donc procéder de la même manière en utilisant le temps comobile:

$$d\eta = \frac{dt}{R} \quad (51.180)$$

Il vient alors:

$$\int d\eta = \int \frac{1}{\sqrt{\frac{8\pi G \rho_0 R_0^3}{3} R + R^2}} dR \quad (51.181)$$

Notons cela sous la forme:

$$\int d\eta = \int \frac{1}{\sqrt{A \cdot R + R^2}} dR \quad (51.182)$$

où A est donc strictement positif. Faisons-y la substitution:

$$R = x - \frac{A}{2} \quad (51.183)$$

nous avons alors:

$$\begin{aligned} \int d\eta &= \int \frac{1}{\sqrt{A \cdot \left(x - \frac{A}{2}\right) + \left(x - \frac{A}{2}\right)^2}} dx = \int \frac{1}{\sqrt{Ax - \frac{A^2}{2} + \left(x^2 - Ax + \frac{A^2}{4}\right)}} dx \\ &= \int \frac{1}{\sqrt{x^2 - \frac{A^2}{4}}} dx \end{aligned} \quad (51.184)$$

En utilisant la primitive usuelle démontrée dans le chapitre de Calcul Différentiel Et Intégral il vient:

$$\int \frac{1}{\sqrt{x^2 - \left(\frac{A}{2}\right)^2}} dx = \operatorname{arccosh}\left(\frac{2x}{A}\right) + c^{\text{te}} \quad (51.185)$$

Soit en refaisant le changement de variables:

$$\operatorname{arccosh}\left(\frac{4x}{A^2}\right) + c^{te} = \operatorname{arccosh}\left(\frac{2\left(R + \frac{A}{2}\right)}{A}\right) + c^{te} = \operatorname{arccosh}\left(\frac{2R + A}{A}\right) + c^{te} \quad (51.186)$$

Donc:

$$\eta = \operatorname{arccosh}\left(\frac{2R + A}{A}\right) + c^{te} \quad (51.187)$$

Pour qu'au temps $\eta = 0$ nous ayons $R = 0$, il faut que la constante soit telle que:

$$\frac{\cosh(-c^{te})A - A}{2} = 0 \quad (51.188)$$

Ce qui nous amène à ce que la constante soit nulle et donc:

$$\eta = \operatorname{arccosh}\left(\frac{2R + A}{A}\right) \quad (51.189)$$

Dès lors:

$$\frac{2R + A}{A} = \cosh(\eta) \quad (51.190)$$

D'où:

$$R = \frac{A}{2}(\cosh(\eta) - 1) \quad (51.191)$$

et comme:

$$d\eta = \frac{dt}{R} \quad (51.192)$$

Nous avons:

$$\int dt = \int R d\eta = \int \frac{A}{2}(\cosh(\eta) - 1) d\eta \quad (51.193)$$

Ce qui donne:

$$t = \frac{A}{2}(\sinh(\eta) - \eta) + c^{te} \quad (51.194)$$

Comme au temps $t = 0$, nous devons avoir $\eta = 0$, il vient que la constante doit être nulle. Donc pour finir, nous avons:

$$R = \frac{A}{2} (\cosh(\eta) - 1) \quad (51.195)$$

$$t = \frac{A}{2} (\sinh(\eta) - \eta)$$

avec Maple 4.00b nous avons alors en comparant l'Univers plat dominé par la matière (en bleu), l'Univers plat dominé par la radiation (en rouge), l'Univers à courbure positive dominé par la matière (vert), l'Univers à courbure positive dominé par la radiation (noir), l'Univers à courbure négative dominé par la matière (gris):

```
>plot([t^(2/3),t^(1/2),[0.5*(t-sin(t)),0.5*(1-cos(t)),t=0..2*Pi],
[0.5*(1-cos(t)),0.5*sin(t),t=0..2*Pi],[0.5*(sinh(t)-t),
0.5*(cosh(t)-1),t=0..2*Pi]],t=0...Pi,0..3,color=[blue,red,green,black,gray]);
```

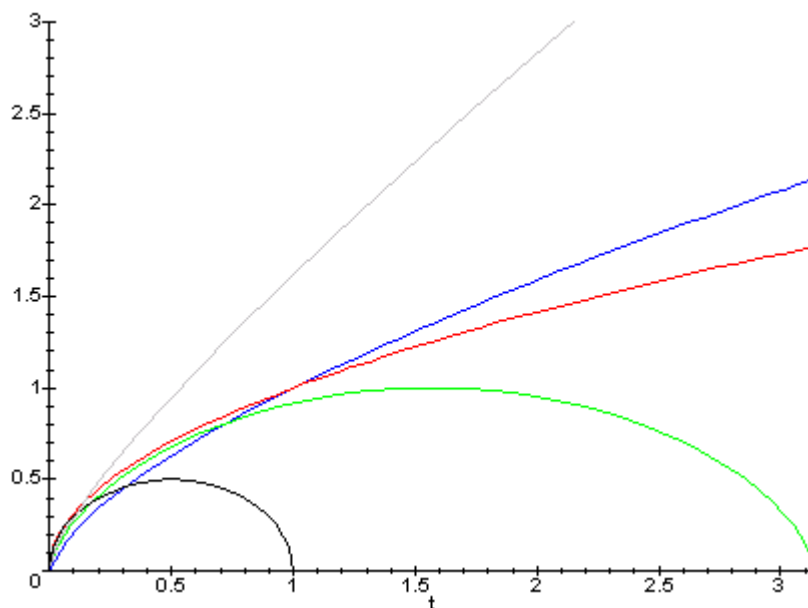


Figure: 51.13 - Évolution du facteur R pour les configurations d'espace obtenues étudiées jusqu'à présent

Nous pouvons donc observer que pour une courbure négative (du type hyperbolique), l'expansion croît nettement plus vite que pour un univers plat et ce sans fin.

2.3.2. ESPACE HYPERBOLIQUE DOMINÉ PAR LA RADIATION

Considérons maintenant un univers dominé par la radiation. Nous avons démontré que dans cette situation nous avons:

$$P = \frac{1}{3} \rho_E \quad (51.196)$$

et:

$$R^4 \rho_E = R_0^4 \rho_{E,0} = c^{t_0} \quad (51.197)$$

La première équation de Friedmann devient alors:

$$\left(\frac{R'}{R}\right)^2 = \frac{8\pi G \rho_E}{3c^2} + \frac{1}{R^2} = \frac{8\pi G \rho_{E,0}}{3c^2} \frac{R_0^4}{R^4} + \frac{1}{R^2} \quad (51.198)$$

Nous avons alors:

$$\frac{dR}{dt} = \sqrt{\frac{8\pi G \rho_{E,0} R_0^4}{3R^2 c^2} + 1} \quad (51.199)$$

d'où:

$$\int dt = \int \frac{1}{\sqrt{\frac{8\pi G \rho_{E,0} R_0^4}{3R^2 c^2} + 1}} dR \quad (51.200)$$

Il s'agit exactement de la même intégrale que l'univers sphérique dominé par la matière à la différence que sous la racine, nous avons +1 au lieu de -1. Nous allons donc procéder de la même manière en utilisant le temps comobile:

$$d\eta = \frac{dt}{R} \quad (51.201)$$

Il vient alors:

$$\int d\eta = \int \frac{1}{\sqrt{\frac{8\pi G \rho_{E,0} R_0^4}{3c^2} + R^2}} dR \quad (51.202)$$

Notons cela sous la forme:

$$\int d\eta = \int \frac{1}{\sqrt{A + R^2}} dR \quad (51.203)$$

où A est donc strictement positif. Dans le chapitre de Calcul Différentiel Et Intégral nous avons démontré comment déterminer exactement la même primitive (car il s'agit d'une primitive usuelle). Nous avons:

$$\eta = \operatorname{arcsinh}\left(\frac{R}{\sqrt{A}}\right) + c^{te} \quad (51.204)$$

Pour qu'au temps $\eta = 0$ nous ayons $R = 0$, il faut que la constante soit nulle. Donc:

$$\eta = \operatorname{arcsinh}\left(\frac{R}{\sqrt{A}}\right) \quad (51.205)$$

D'où:

$$R = \sqrt{A} \sinh(\eta) \quad (51.206)$$

et comme:

$$d\eta = \frac{dt}{R} \quad (51.207)$$

Nous avons:

$$\int dt = \int R d\eta = \int \sqrt{A} \sinh(\eta) d\eta \quad (51.208)$$

Ce qui donne:

$$t = \sqrt{A} \cosh(\eta) + c^{te} \quad (51.209)$$

Comme au temps $t = 0$, nous devons avoir $\eta = 0$, il vient que la constante doit être égale à $-\sqrt{A}$.
Donc pour finir, nous avons:

$$\begin{aligned} R &= \sqrt{A} \sinh(\eta) \\ t &= \sqrt{A} (\cosh(\eta) - 1) \end{aligned} \quad (51.210)$$

avec Maple 4.00b nous avons alors en comparant l'Univers plat dominé par la matière (en bleu), l'Univers plat dominé par la radiation (en rouge), l'Univers à courbure positive dominé par la matière (vert), l'Univers à courbure positive dominé par la radiation (noir), l'Univers à courbure négative dominé par la matière (gris), l'Univers à courbure négative dominé par la radiation (brun):

```
>plot([t^(2/3),t^(1/2),[0.5*(t-sin(t)),0.5*(1-cos(t)),t=0..2*Pi],
[0.5*(1-cos(t)),0.5*sin(t),t=0..2*Pi],[0.5*(sinh(t)-t),0.5*(cosh(t)-1),t=0..2*Pi],
[0.5*(cosh(t)-1),0.5*(sinh(t)),t=0..2*Pi]],t=0...Pi,0..3
,color=[blue,red,green,black,gray,brown]);
```

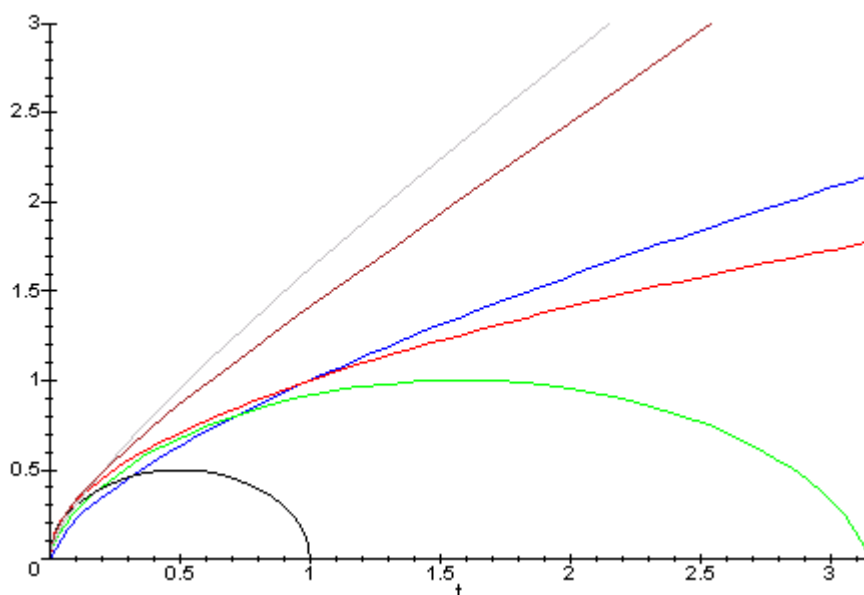


Figure: 51.14 - Évolution du facteur R pour un espace à courbure $0/\pm$ dominé par la matière ou par la radiation

Nous pouvons donc observer que pour une courbure négative (du type hyperbolique), l'expansion d'un Univers dominé par la radiation croît moins rapidement qu'un Univers dominé par la matière (c'est un peu intuitif...).

Soit pour résumer un peu mieux tout cela avec des légendes ce dernier graphique devient (faut quand même s'appliquer car l'Univers nous concerne tous...):

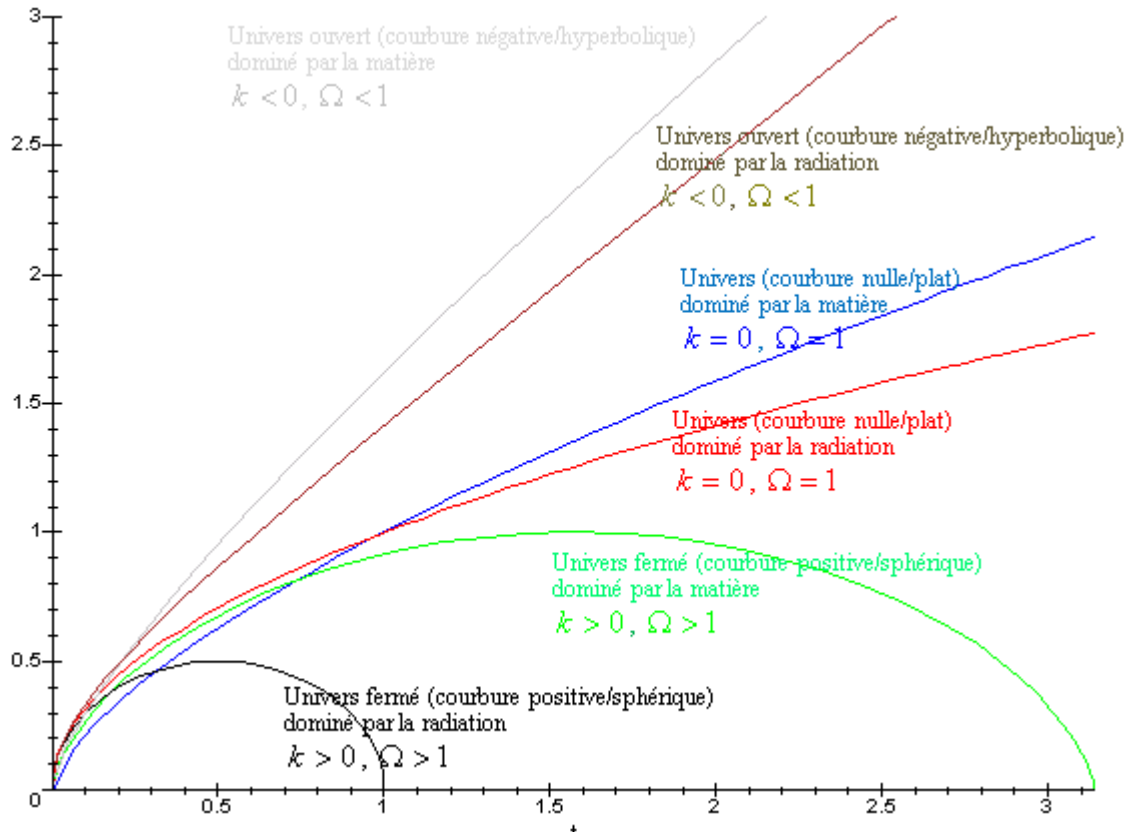


Figure: 51.15 - Résumé des modèles Newtonien d'Univers

2.4. UNIVERS OBSERVABLE

Nous avons déterminé plus haut une estimation actuelle de l'âge (horizon) de l'Univers comme pouvant être interprétée comme l'inverse de la constante de Hubble ce qui nous a donné:

$$t_H = 1/H_0 \cong 13 \cdot 10^9 \text{ années} \quad (51.211)$$

soit environ 13 milliards d'années.

Remarques:

R1. Nous noterons que les articles populaires et professionnels de recherche en cosmologie emploient souvent le terme "Univers" dans le sens de "**Univers observable**".

R2. Il faudrait au fait être plus rigoureux lorsque nous parlons d'âge de l'Univers. Au fait, nous devrions plutôt dire que l'horizon de l'Univers, pour un observateur comobile depuis les époques les plus reculées, est de 13 milliards d'années. En clair, c'est le temps que mesurerait quelqu'un qui serait resté observateur inertiel (en chute libre: ne subissant aucune autre force que la gravitation) tout au long de l'évolution de l'Univers et dans un référentiel tel qu'il aurait toujours perçu cet Univers comme homogène et isotrope.

À ce jour, nous ne savons pas si l'Univers est fini ou infini, bien que la majorité des théoriciens favorisent actuellement un Univers infini.

L'Univers observable se compose ainsi de tous les endroits qui pourraient nous avoir affectés depuis le Big Bang (attention! malgré son nom, cette théorie du Big Bang n'a rien à dire sur le début! Elle se contente de décrire l'évolution et l'expansion de l'Univers), en tenant compte que la vitesse de la lumière est certainement finie. L'horizon cosmique se trouve quant à lui à une distance de 14 à 15 milliards d'années-lumière selon les observations expérimentales de la fin du 20ème siècle.

La taille actuelle (la "**distance comobile**") de l'Univers observable est plus grande, puisque l'Univers a continué de s'étendre pendant le temps que la lumière met à nous parvenir, nous estimons qu'elle est d'environ ~40 milliards d'années-lumière.

Ce chiffre peut être obtenu en prenant un objet visible qui est à 13 milliards d'années de notre Terre. Celui-ci aura donc mis 13 milliards d'années pour s'éloigner de nous, sa lumière aura mis 13 milliards d'années pour arriver jusqu'à nous et pendant ce temps de parcours de la lumière, il se sera éloigné de 13 milliards d'années (puisque les objets à l'horizon cosmologique vont à la vitesse de la lumière). Soit un total de ~39 milliard d'années.

Celui-ci contiendrait d'après des estimations toutes heuristiques environ $7 \cdot 10^{22}$ étoiles, répandues dans environ 10^{10} galaxies, elles-mêmes organisées en amas et superamas de galaxies (le nombre de galaxies pourrait être encore plus grand, selon le "**champ profond de Hubble**" observé avec le télescope spatial Hubble.)

Cependant il est difficile de s'imaginer ce que cela représente. À ce titre, nous avons trouvé sur Internet une magnifique série d'illustrations (<http://atunivers.free.fr>) que nous vous proposons:

1. L'Univers jusqu'à 13 milliards d'années-lumière (l'Univers visible):

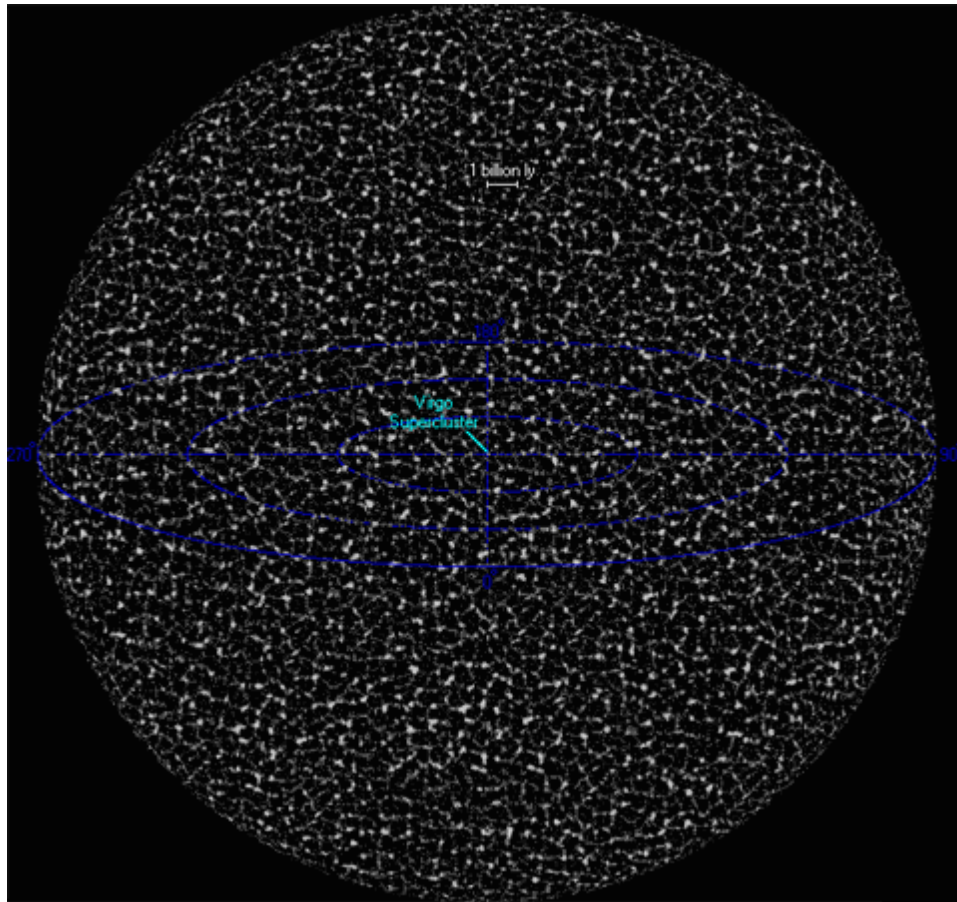


Figure: 51.16 - Illustration simplifiée de l'Univers observable (source: <http://atunivers.free.fr>)

Cette carte essaie de montrer l'ensemble de l'Univers visible. Les galaxies dans l'Univers ont tendance à se rassembler en vastes feuilles et "**superamas**" de galaxies, entourant de grands vides, ce qui confère à l'univers une apparence cellulaire. Parce que la lumière dans l'univers ne voyage qu'à une vitesse finie, nous voyons les objets sur le bord de l'Univers quand celui-ci était très jeune, il y a 13 milliards d'années.

Quelques chiffres (estimations):

- Nombre de superamas de l'univers visible = 10 millions
- Nombre de groupes de galaxies de l'univers visible = 25 milliards
- Nombre de grandes galaxies de l'univers visible = 350 milliards
- Nombre de galaxies naines de l'univers visible = 7 trillions
- Nombre d'étoiles de l'univers visible = 30 milliards de trillions

2. L'Univers jusqu'à 1 milliard d'années-lumière (les superamas voisins):

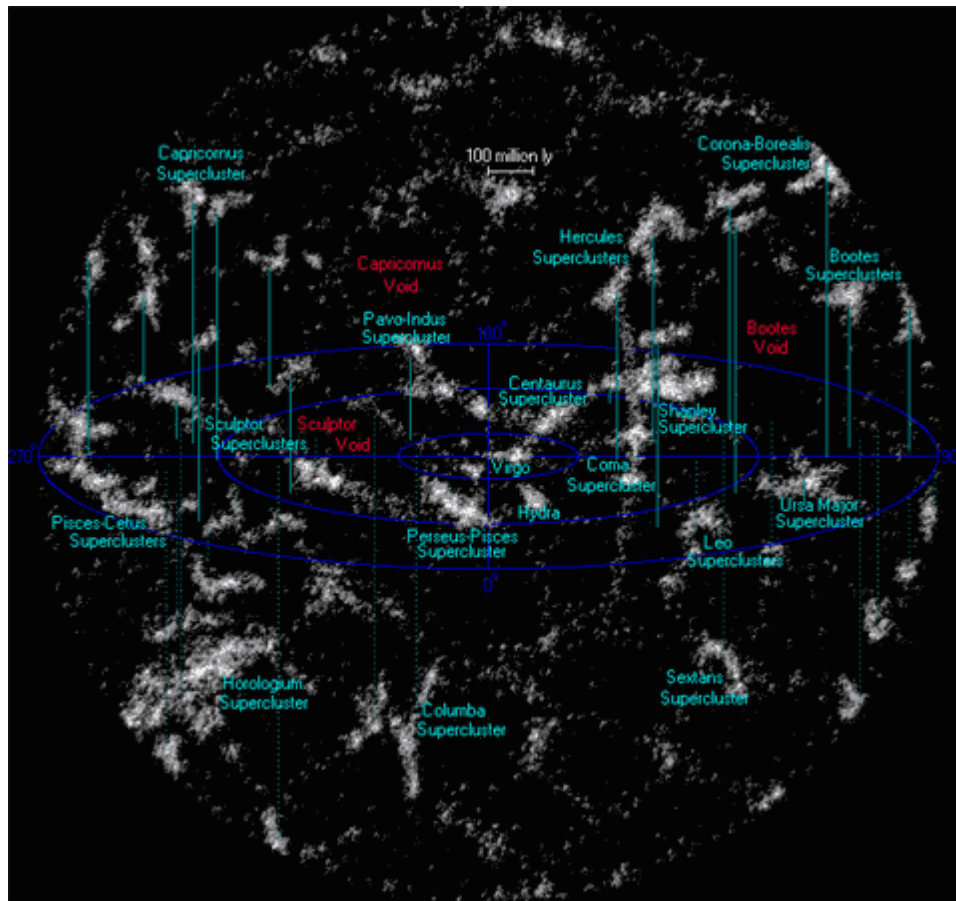


Figure: 51.17 - Illustration simplifiée de l'Univers à une échelle de 1 milliard d'a.-l. (source: <http://atunivers.free.fr>)

Quelques chiffres (estimations):

Les Galaxies et les amas de galaxies ne sont pas distribués régulièrement dans l'Univers. Au lieu de cela, ils sont rassemblés en de larges amas, feuillets et murs de galaxies séparés par de larges vides dans lesquels peu de galaxies semblent se trouver. La carte ci-dessus montre un certain nombre de ces superamas, y compris celui de la Vierge - un superamas plutôt petit dont notre galaxie fait partie. La carte entière représente à peu près 7% du diamètre de l'Univers visible. Les galaxies sont trop petites pour apparaître individuellement sur cette carte, chaque point y représente un groupe de galaxies.

Quelques chiffres (estimations):

- Nombre de superamas jusqu'à 1 milliard d'années-lumière = 100
- Nombre de groupes galactiques jusqu'à 1 milliard d'années-lumière = 240'000
- Nombre de grandes galaxies jusqu'à 1 milliard d'années-lumière = 3 millions
- Nombre de galaxies naines jusqu'à 1 milliard d'années-lumière = 60 millions
- Nombre d'étoiles jusqu'à 1 milliard d'années-lumière = 250 trillions

3. L'Univers jusqu'à 100 millions d'années-lumière (le superamas de la Vierge):

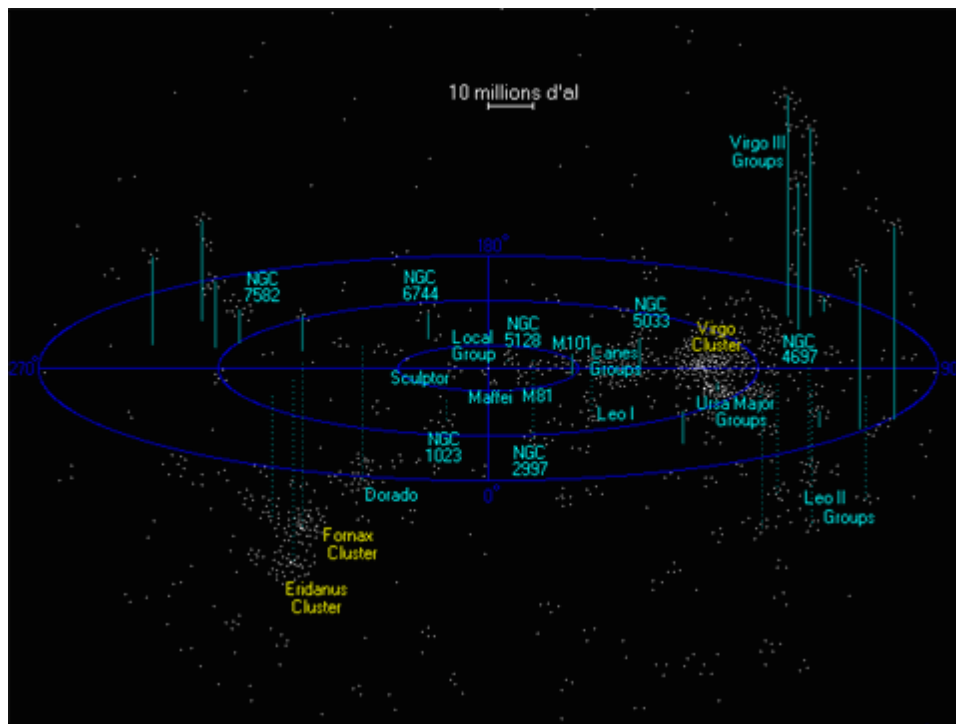


Figure: 51.18 - Illustration simplifiée de l'Univers à une échelle de 100 millions d'a.-l. (source: <http://atunivers.free.fr>)

Notre galaxie n'est qu'une parmi des milliers d'autres qui se trouvent à moins de 100 millions d'années-lumière. La carte ci-dessus montre comment les galaxies tendent à s'amasser par groupes, le plus important des amas proches étant l'amas de la Vierge (Virgo), une concentration de plusieurs centaines de galaxies qui domine les groupes de galaxies environnants. Collectivement, l'ensemble de ces groupes est connu sous le nom de Superamas de la Vierge. Le second amas le plus riche de ce volume est l'amas du Fourneau (Fornax), mais il est bien moins riche que celui de la Vierge. Seules les galaxies brillantes sont dessinées ici, notre galaxie est le point tout au centre.

Quelques chiffres (estimations):

- Nombre de groupes de galaxies jusqu'à 100 millions d'années-lumière = 200
- Nombre de grandes galaxies jusqu'à 100 millions d'années-lumière = 2'500
- Nombre de galaxies naines jusqu'à 100 millions d'années-lumière = 50'000
- Nombre d'étoiles jusqu'à 100 millions d'années-lumière = 200 trillions

4. L'Univers à moins de 5 millions d'années-lumière (le groupe local de galaxies):

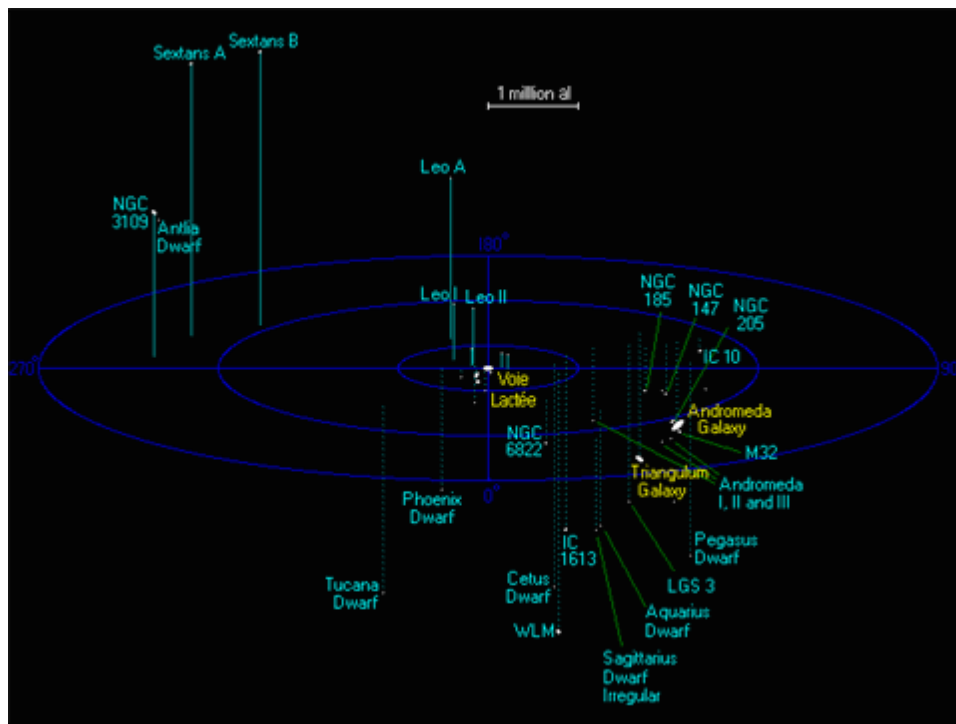


Figure: 51.19 - Illustration simplifiée de l'Univers à une échelle de 5 millions d'a.-l. (source: <http://atunivers.free.fr>)

La Voie Lactée est une des trois grandes galaxies du groupe appelé "**Groupe Local**" qui contient aussi plusieurs dizaines de galaxies naines. La plupart de ces galaxies sont portées sur cette carte, mais il faut noter que beaucoup de ces galaxies naines sont très peu brillantes, et qu'il y en a donc certainement d'autres à découvrir.

Quelques chiffres (estimations):

- Nombre de grandes galaxies à moins de 5 millions d'années-lumière = 3
- Nombre de galaxies naines à moins de 5 millions d'années-lumière = 42
- Nombre d'étoiles à moins de 5 millions d'années-lumière = 700 milliards

5. L'Univers jusqu'à 500'000 années-lumière (les galaxies satellites):

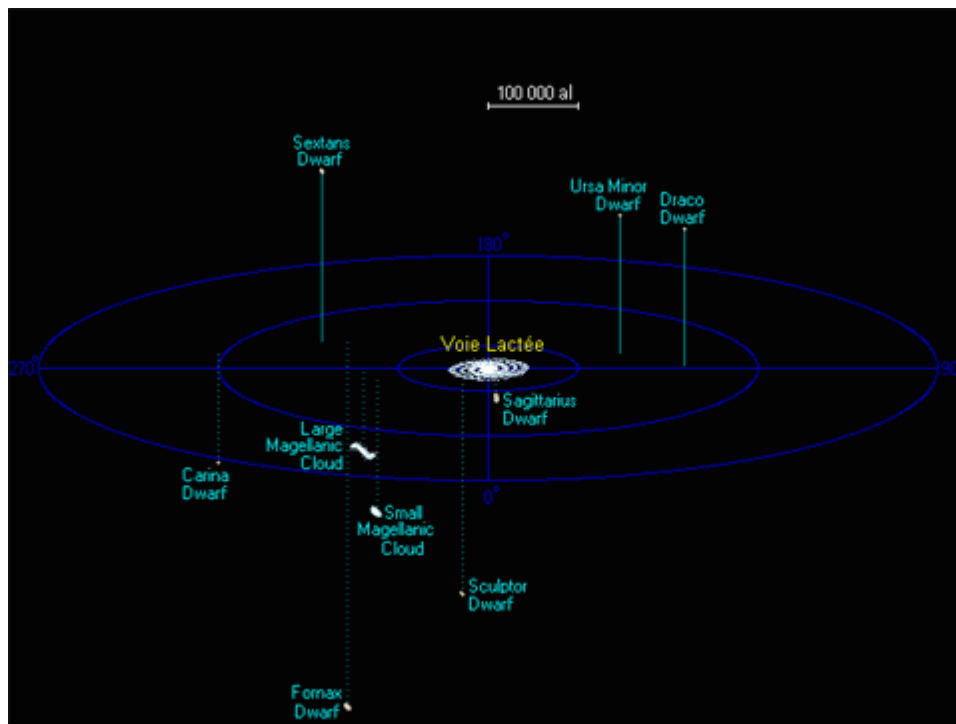


Figure: 51.20 - Illustration simplifiée de l'Univers à une échelle de 500'000 a.-l. (source: <http://atunivers.free.fr>)

La Voie Lactée est entourée par plusieurs galaxies naines, qui contiennent chacune quelques dizaines de millions d'étoiles, ce qui est insignifiant comparé à la population de la Voie Lactée elle-même. La carte ci-dessus montre l'ensemble des galaxies naines les plus proches, elles sont liées gravitationnellement à la Voie Lactée, et gravitent autour d'elle en quelques milliards d'années.

Quelques chiffres (estimations):

- Nombre de grandes galaxies jusqu'à 500'000 années-lumière = 1
- Nombre de galaxies naines jusqu'à 500'000 années-lumière = 12
- Nombre d'étoiles jusqu'à 500'000 années-lumière = 225 milliards

6. L'Univers jusqu'à 50'000 années-lumière (la Voie Lactée):

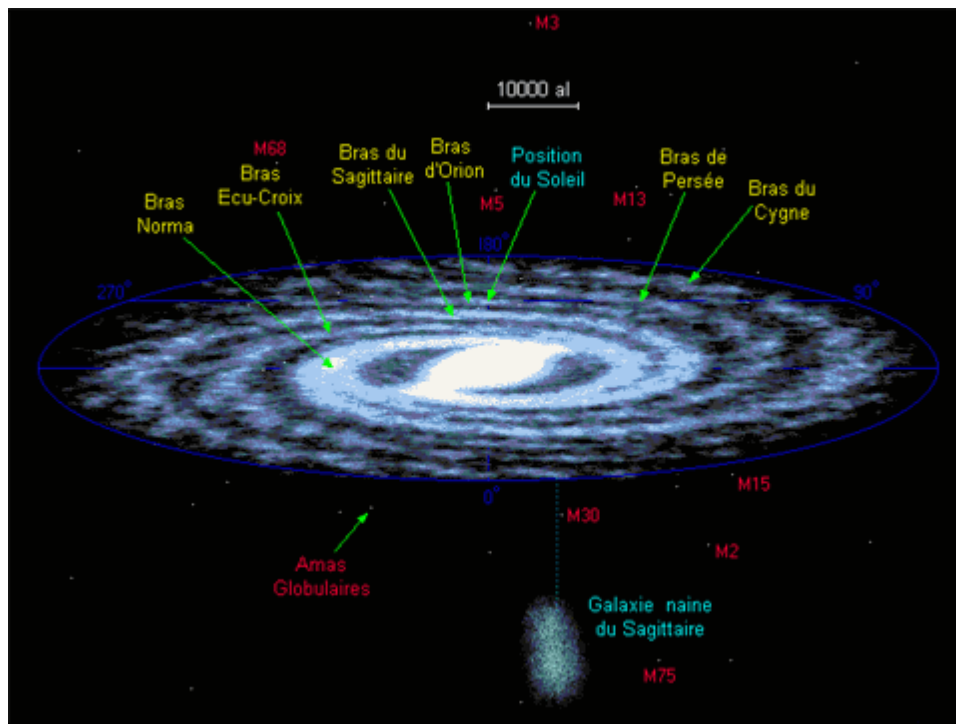


Figure: 51.21 - Illustration simplifiée de l'Univers à une échelle de 50'000 a.-l. (source: <http://atunivers.free.fr>)

Cette carte montre la Voie Lactée dans son ensemble - une galaxie spirale d'au moins deux cents milliards d'étoiles. Notre Soleil est profondément enfoui dans le Bras d'Orion à environ 26'000 années lumière du centre. Vers le centre de la Galaxie, les étoiles sont beaucoup plus proches les unes des autres qu'à la périphérie où nous vivons. Notez également la présence de petits amas globulaires bien en dehors du plan galactique, et la présence d'une galaxie naine voisine - dite du Sagittaire - qui est en train d'être lentement avalée par notre propre Galaxie.

Quelques chiffres (estimations):

- Nombre d'étoiles jusqu'à 50'000 années-lumière = 200 milliards

7. L'Univers jusqu'à 5'000 années-lumière (le Bras d'Orion):

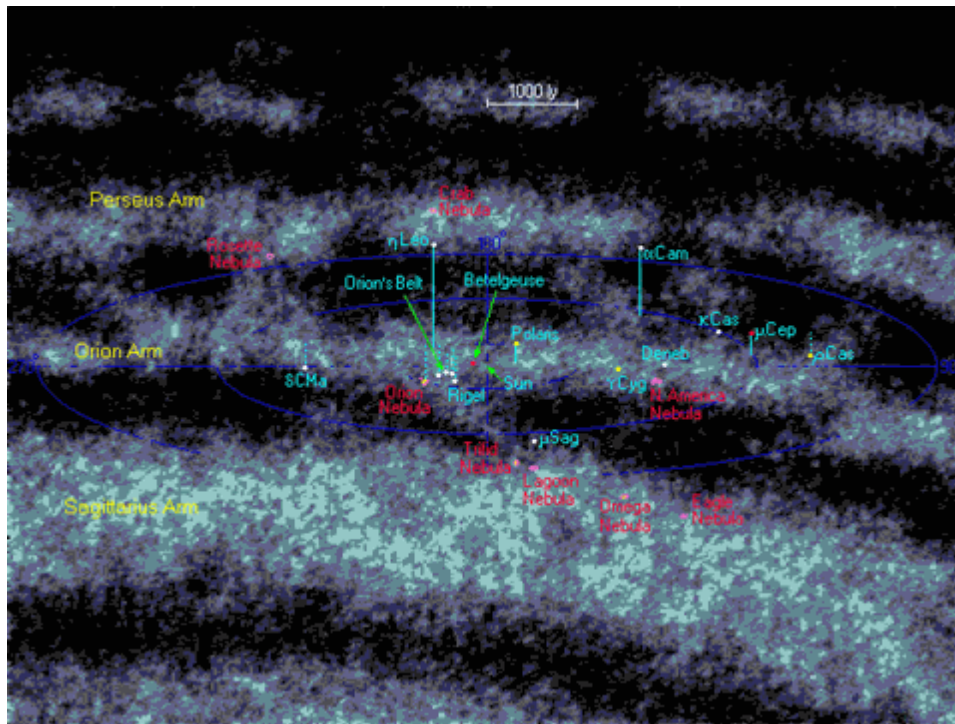


Figure: 51.22 - Illustration simplifiée de l'Univers à une échelle de 5'000 a.-l. (source: <http://atunivers.free.fr>)

Ceci est une carte de notre coin de la Voie Lactée. Le Soleil est situé dans le Bras d'Orion - un bras assez petit comparé au Bras du Sagittaire, qui se situe plus près du centre galactique. La carte montre plusieurs étoiles visibles à l'oeil nu, situées loin dans le bras d'Orion. Le groupe d'étoiles le plus marquant est composé des étoiles principales de la constellation d'Orion - de laquelle le bras spiral tire son nom. Toutes ces étoiles sont des géantes et supergéantes lumineuses, des milliers de fois plus lumineuses que le Soleil. L'étoile la plus brillante de la carte est Rho Cassiopeia - à 4'000 années-lumière de nous, c'est juste une étoile à peine visible à l'oeil nu, mais en réalité c'est une supergéante 100'000 fois plus lumineuse que le Soleil.

Quelques chiffres (estimations):

- Nombre d'étoiles jusqu'à 5'000 années-lumière = 600 millions

8. L'Univers jusqu'à 250 années-lumière (le voisinage du Soleil):

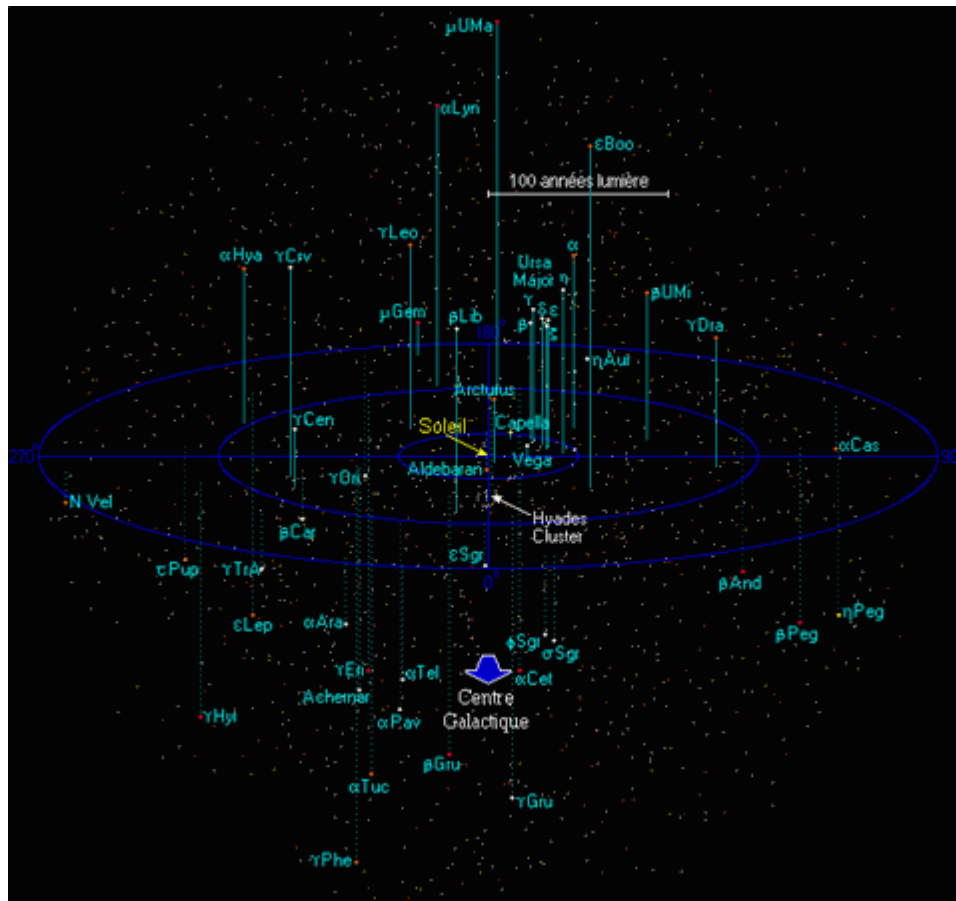


Figure: 51.23 - Illustration simplifiée de l'Univers à une échelle de 250 a.-l. (source: <http://atunivers.free.fr>)

Cette carte indique les 1'500 étoiles les plus lumineuses situées à moins de 250 années-lumière. Toutes ces étoiles sont bien plus lumineuses que le Soleil, et la plupart sont visibles à l'oeil nu. Environ un tiers des étoiles visibles à l'oeil nu sont situées à moins de 250 années-lumière, même si cette zone ne représente qu'une toute petite partie de notre galaxie.

Quelques chiffres (estimations):

- Nombre d'étoiles jusqu'à 250 années-lumière = 260'000

9. L'Univers jusqu'à 12.5 années-lumière (les étoiles les plus proches):

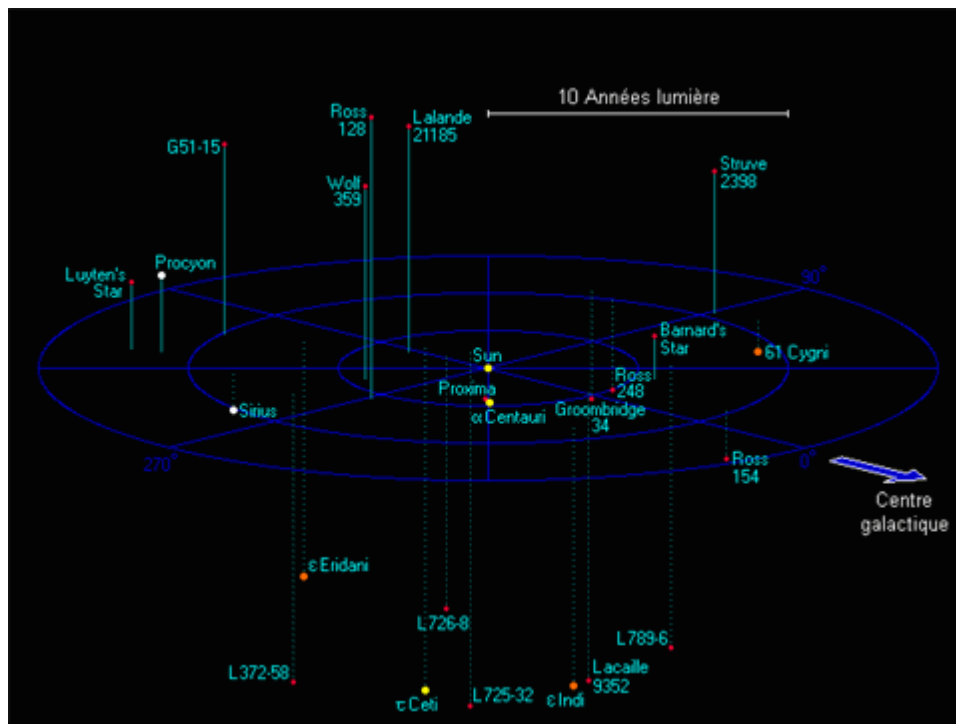


Figure: 51.24 - Illustration simplifiée de l'Univers à une échelle de 12.5 a.-l. (source: <http://atunivers.free.fr>)

Cette carte montre certaines étoiles jusqu'à une distance de 12.5 années-lumière de notre Soleil (il y en aurait 33 d'identifiées à ce jour). La plupart de ces étoiles sont des naines rouges - des étoiles avec une masse du dixième de celle du Soleil et une luminosité cent fois moins grande. Environ 80% des étoiles de l'Univers sont des naines rouges, et l'étoile la plus proche - Proxima du Centaure- en est un exemple typique.

Cette carte montre toutes les étoiles connues situées à moins de 20 années-lumière. On y trouve un total de 77 systèmes contenant 110 étoiles.

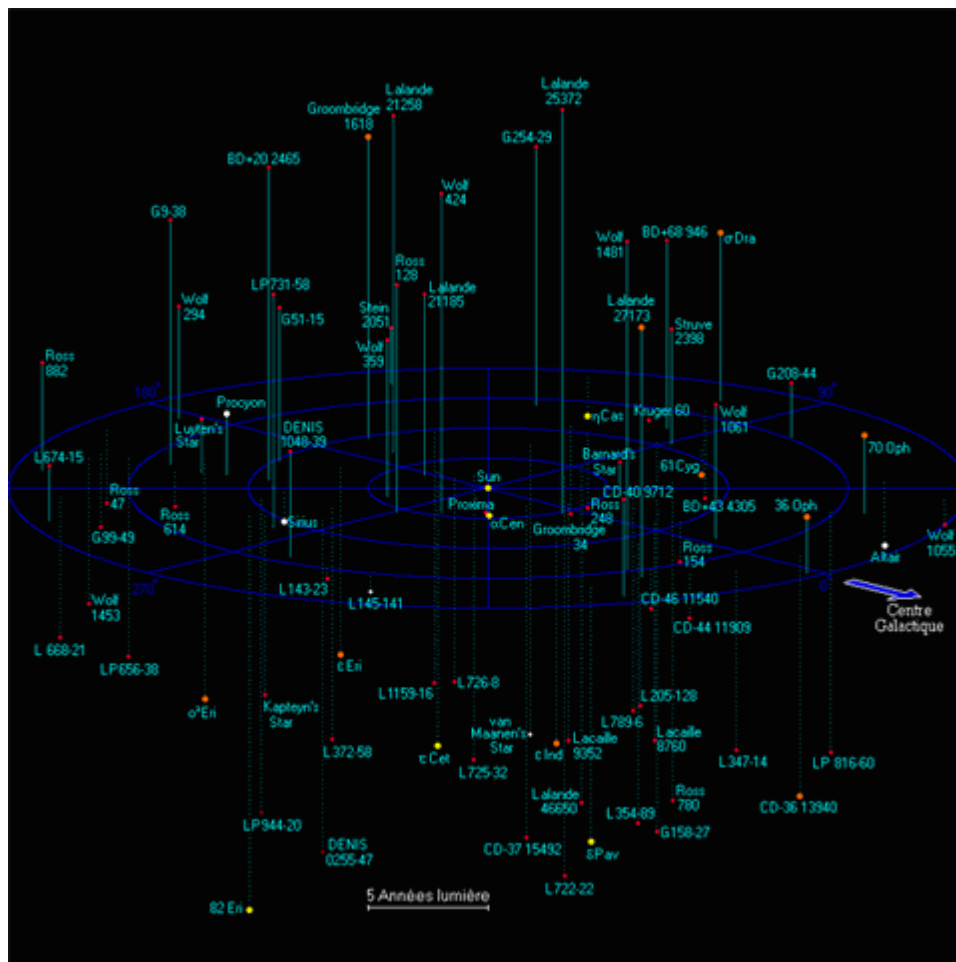


Figure: 51.25 - Illustration simplifiée de l'Univers à une échelle de moins de 20 a.-l. (source: <http://atunivers.free.fr>)

Les distances entre les étoiles sont énormes. La distance du Soleil à Proxima Centauri est de 4.22 années-lumière, soit quarante trillions de kilomètres. Marcher sur cette distance prendrait un milliard d'années. Même les sondes spatiales les plus rapides mettraient soixante mille ans pour faire le voyage. Il y a actuellement quatre sondes qui quittent le système solaire - Pioneer 10 et 11, et Voyager 1 et 2 mais nous perdrons vraisemblablement le contact avec elles d'ici une vingtaine d'années. Le schéma ci-dessous essaie de montrer ces distances en élargissant le champ depuis le système solaire intérieur jusqu'à Alpha du Centaure.

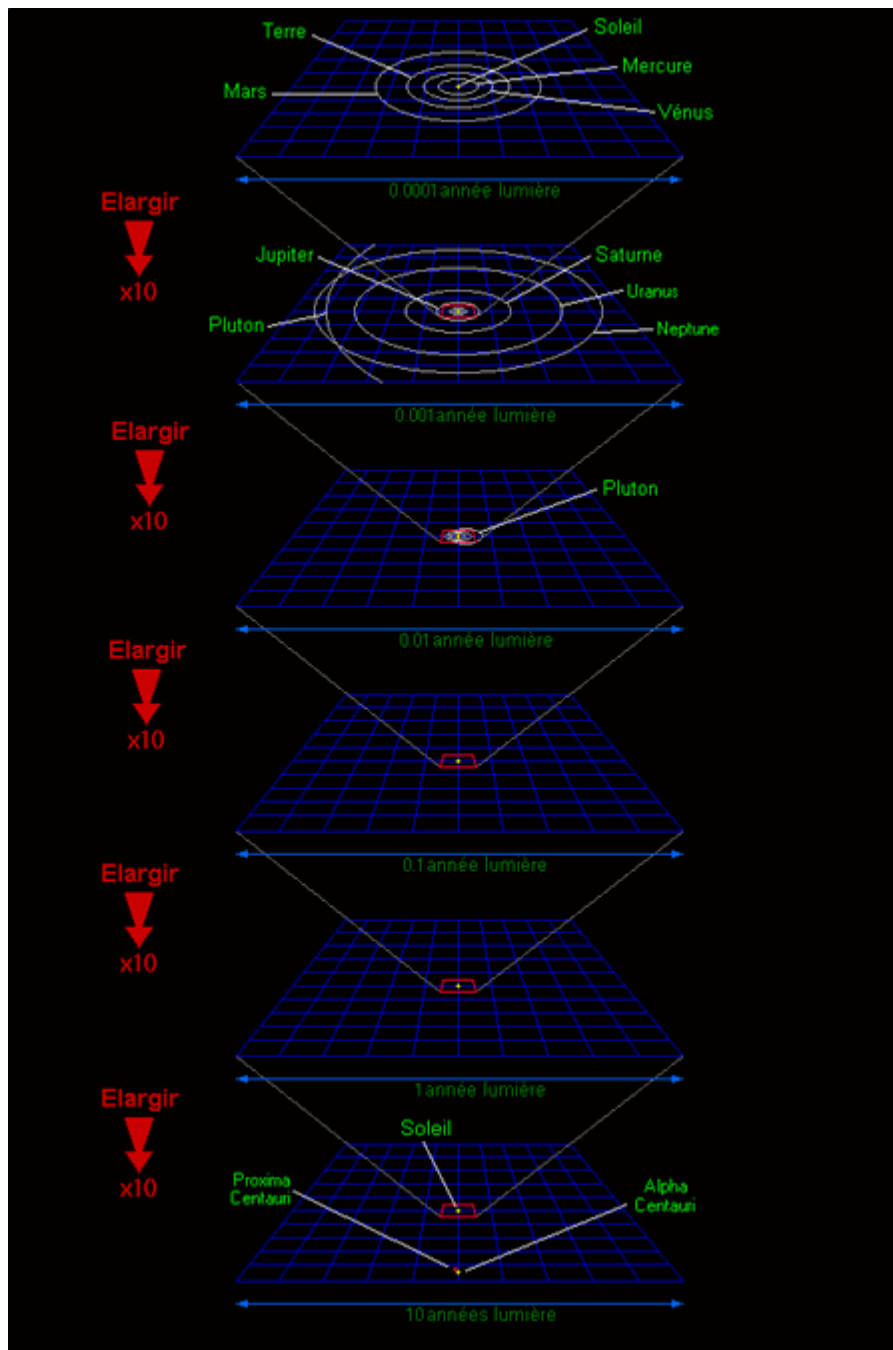


Figure: 51.26 - Zoom de bas en haut pour illustrer les échelles (source: <http://atunivers.free.fr>)

3. RAYONNEMENT FOSSILE

L'existence et les propriétés du rayonnement cosmique découvert par Penzias et Wilson s'expliquent essentiellement par les deux phénomènes physiques que nous allons maintenant décrire dans leurs grandes lignes.

L'expansion de l'Univers a pour conséquence son refroidissement graduel. À partir des valeurs fantastiquement élevées qui ont dû régner aussitôt après le Big Bang qui a engendré l'Univers, sa température a progressivement décliné. Lorsqu'elle atteint environ 3'000 [K] se produit le premier des deux phénomènes cruciaux qui nous intéressent ici: le rayonnement, qui jusque-là était en équilibre thermique avec les particules matérielles, cesse pratiquement d'interagir avec elles et en devient indépendant. Dans le "**modèle standard**" d'évolution de l'Univers, nous calculons que ce moment crucial se situe $0.68 \cdot 10^6$ ans après le Big Bang.

Nous pouvons comprendre qualitativement les raisons physiques de ce phénomène. Un peu avant, lorsque par exemple la température était de 100'000 [K], l'Univers contenait essentiellement des photons, des électrons et des noyaux atomiques nus (principalement des protons, et, dans une moindre proportion, des particules α , noyaux d'hélium 4). La température était trop élevée pour que les électrons et les noyaux puissent former des atomes, autrement que de manière transitoire et labile. L'interaction entre les photons et les particules chargées (surtout les électrons, les plus légères d'entre elles) est suffisamment intense, et la densité de ces dernières était alors suffisamment forte, pour que les photons soient sans arrêt diffusés, émis et absorbés. Malgré son expansion, l'Univers était alors à chaque instant en équilibre; sa température T était constamment bien définie, bien que décroissante au cours du temps, l'énergie des photons, c'est-à-dire la pulsation du rayonnement, était donc distribuée suivant la loi de Planck correspondant à cette température T .

La diminution de la température a ensuite permis la formation d'atomes à partir des électrons et des noyaux. Ce processus a entraîné une chute rapide de la section efficace moyenne d'interaction entre les photons et les particules matérielles (principalement à cause de la disparition des électrons libres), de sorte que l'Univers est devenu transparent aux photons. Une évaluation quantitative des caractéristiques du phénomène situe ce découplage au moment où la température est descendue à 3'000 [K].

Au moment du découplage, la densité volumique d'énergie du rayonnement est distribuée dans le spectre des pulsations selon la loi de Planck (cf. chapitre de Thermodynamique):

$$R(\omega, T) = \frac{2\hbar\omega^3}{\pi c^3} \frac{1}{\exp\left(\frac{\hbar\omega}{kT}\right) - 1} \quad [J \cdot s \cdot m^{-3}] \quad (51.212)$$

où nous admettrons que T est la température (3'000 [K] environ - température d'ionisation des atomes les plus simples) à ce moment-là. Cette distribution va ensuite évoluer sous l'influence de l'expansion de l'Univers.

Considérons les photons situés, à cet instant t dans le volume r^3 , et dont la pulsation est ω à $d\omega$ près. Leur nombre est donc égal à:

$$dN = \frac{r^3}{\hbar\omega} R(\omega, T) = r^3 \frac{2\omega^2}{\pi c^3} \frac{\omega^2 d\omega}{\exp\left(\frac{\hbar\omega}{kT}\right) - 1} \quad (51.213)$$

Comme il n'y a plus d'absorption ni d'émission de photons à cette température (c'est une hypothèse mais comme les mesures expérimentales semblent confirmer ce modèle à défaut de mieux...), ce nombre va rester constant. Mais à cause de l'expansion de l'Univers, ces photons en nombre constant vont occuper un volume plus grand, et acquérir une longueur d'onde plus grande (selon l'expansion de la structure même de l'espace due à la valeur positive de la constante de Hubble) c'est-à-dire une pulsation plus petite (l'équivalent de l'effet Doppler). Pour préciser, examinons la situation à un instant t' ultérieur. Toutes les longueurs de l'Univers ont été multipliées entre, entre t et t' , par le même facteur d'échelle F selon la loi de Hubble: l'arête r du volume cubique choisi est ainsi devenue:

$$r' = F \cdot r \quad (51.214)$$

et la longueur d'onde des photons considérés:

$$\lambda' = F \cdot \lambda \quad (51.215)$$

de sorte que leur pulsation vaut à l'instant t' :

$$\omega' = \frac{2\pi c}{\lambda'} = \frac{\omega}{F} \quad (51.216)$$

Donc, l'énergie contenue à cet instant dans le volume $V' = r'^3$ et dans la bande de pulsations $(\omega', \omega' + d\omega')$, que nous écrirons $r'^3 R'(\omega', T') d\omega'$ est donnée par:

$$r'^3 R'(\omega', T') d\omega' = \hbar \omega' (dN) = \hbar \omega' \left(\frac{2r'^3}{\pi c^3} \frac{\omega'^2 d\omega'}{\exp\left(\frac{\hbar \omega'}{kT}\right) - 1} \right) \quad (51.217)$$

La densité volumique d'énergie $R'(\omega', T') d\omega'$ à l'instant t' , pour la bande de pulsation $(\omega', \omega' + d\omega')$, s'écrit donc:

$$R'(\omega', T') d\omega' = \hbar \omega' \frac{(dN)}{r'^3} = \hbar \omega' \left(\frac{2r'^3}{r'^3 \pi c^3} \frac{\omega'^2 d\omega'}{\exp\left(\frac{\hbar \omega'}{kT}\right) - 1} \right) = \hbar \omega' \left(\frac{2}{F^3 \pi c^3} \frac{F^2 \omega'^2 \cdot F d\omega'}{\exp\left(\frac{\hbar F \omega'}{kT}\right) - 1} \right) \quad (51.218)$$

Il s'ensuit que la distribution spectrale de l'énergie est encore à l'instant t' celle du corps noir:

$$R'(\omega', T') = \hbar \omega' \frac{2\omega'^2}{\pi c^3} \frac{1}{\exp\left(\frac{\hbar F \omega'}{kT}\right) - 1} = \quad (51.219)$$

où la température correspondante T' est telle que:

$$T' = \frac{T}{F} \quad (51.220)$$

Ainsi, après son découplage d'avec la matière, le rayonnement cosmique évolue en conservant la distribution d'un corps noir dont la température décroît régulièrement, dans la même proportion que s'accroissent les distances au cours de l'expansion de l'Univers (depuis le moment du découplage, le facteur F d'échelle est très voisin de 1'000 puisque pour passer de 3'000 [K] aux 2.7 [K] actuels il y a un facteur 1000...). Cette valeur de 1000 nous permet à partir du modèle de Friedmann-Lemaître que nous avons démontré en partie ci-dessus de facilement calculer à quel moment de l'âge (horizon) de l'Univers ce découplage a eu lieu. C'est ainsi que nous trouvons une valeur d'à peu près $0.68 \cdot 10^6$ années.

C'est en se fondant sur ce raisonnement que divers auteurs furent amenés à prédire l'existence dans l'Univers actuel, d'un rayonnement fossile de quelques kelvins. La découverte de Penzias et Wilson, confirme parfaitement le plus solide argument en faveur du modèle (cosmologique) standard, qui reconstitue l'histoire de l'Univers à partir de la "grande explosion" initiale.

4. L'UNIVERS TROU NOIR

Une hypothèse assez récente dans l'histoire de la cosmologie (quelques décennies) et qui est au coeur de nombreuses recherches théoriques (Hawking, Penrose et autres) est la possibilité d'assimiler notre Univers à un Trou Noir (cf. chapitre de Relativité Générale).

L'origine de l'idée peut se faire à partir d'un calcul très simple:

Nous savons que le rayon de l'Univers (actuel) est donné selon nos calculs précédents par:

$$r_0 = c \cdot t_0 = cH_0^{-1} \cong 1.2 \cdot 10^{26} [m] \quad (51.221)$$

Or, nous avons démontré dans le chapitre de Relativité Générale (et de Mécanique Classique) que le rayon de Schwarzschild est donné par:

$$R = \frac{2GM}{c^2} \quad (51.222)$$

Ce que nous pouvons écrire pour l'Univers sous la forme suivante:

$$S_r = \frac{2GM}{c^2} = \frac{2G\rho_0 V}{c^2} = \frac{2G\rho_0 \left(\frac{4}{3}\pi r_0^3\right)}{c^2} \quad (51.223)$$

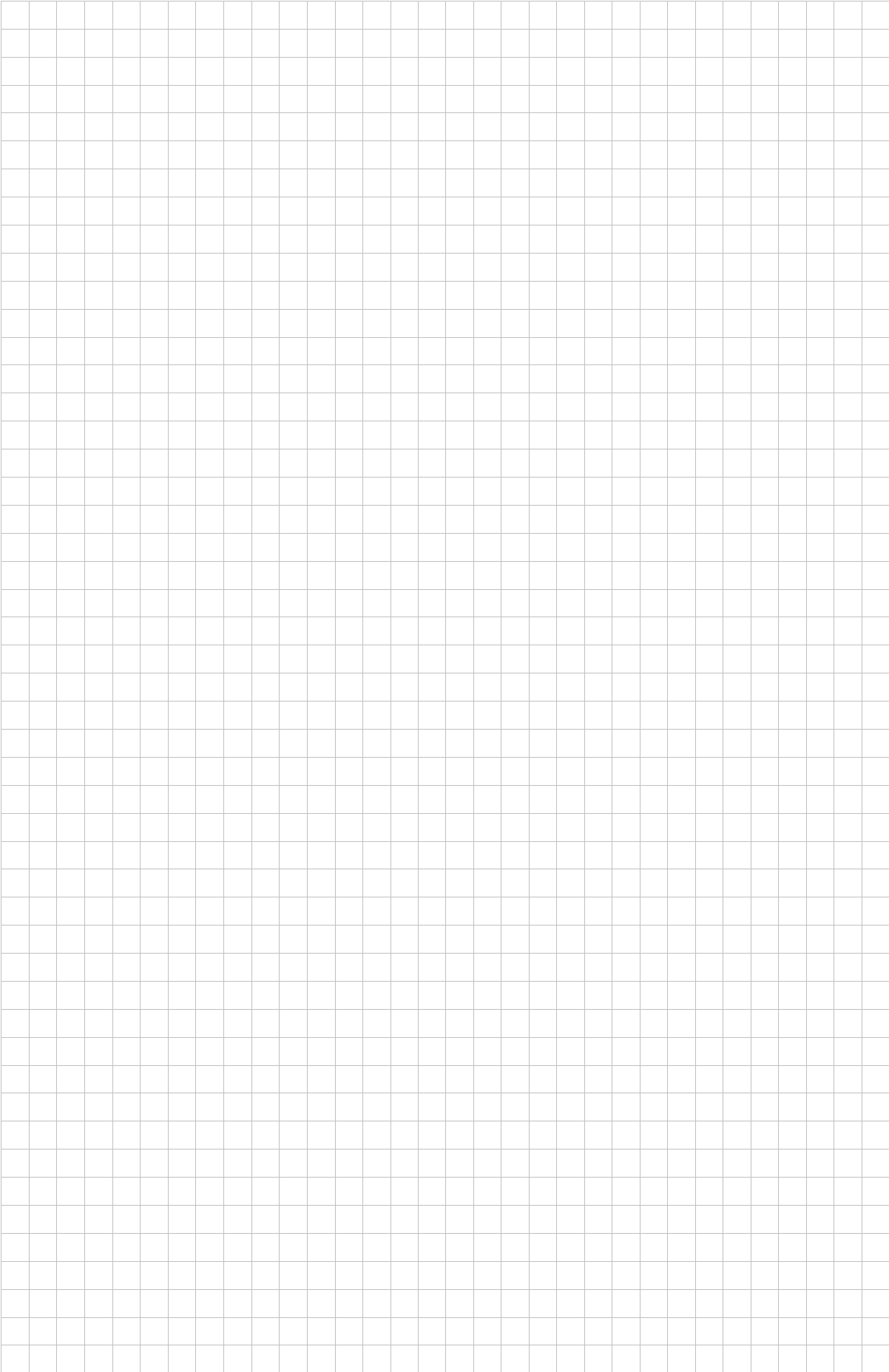
ce qui avec les valeurs de la densité critique et du rayon de l'horizon cosmologique calculé plus haut donne:

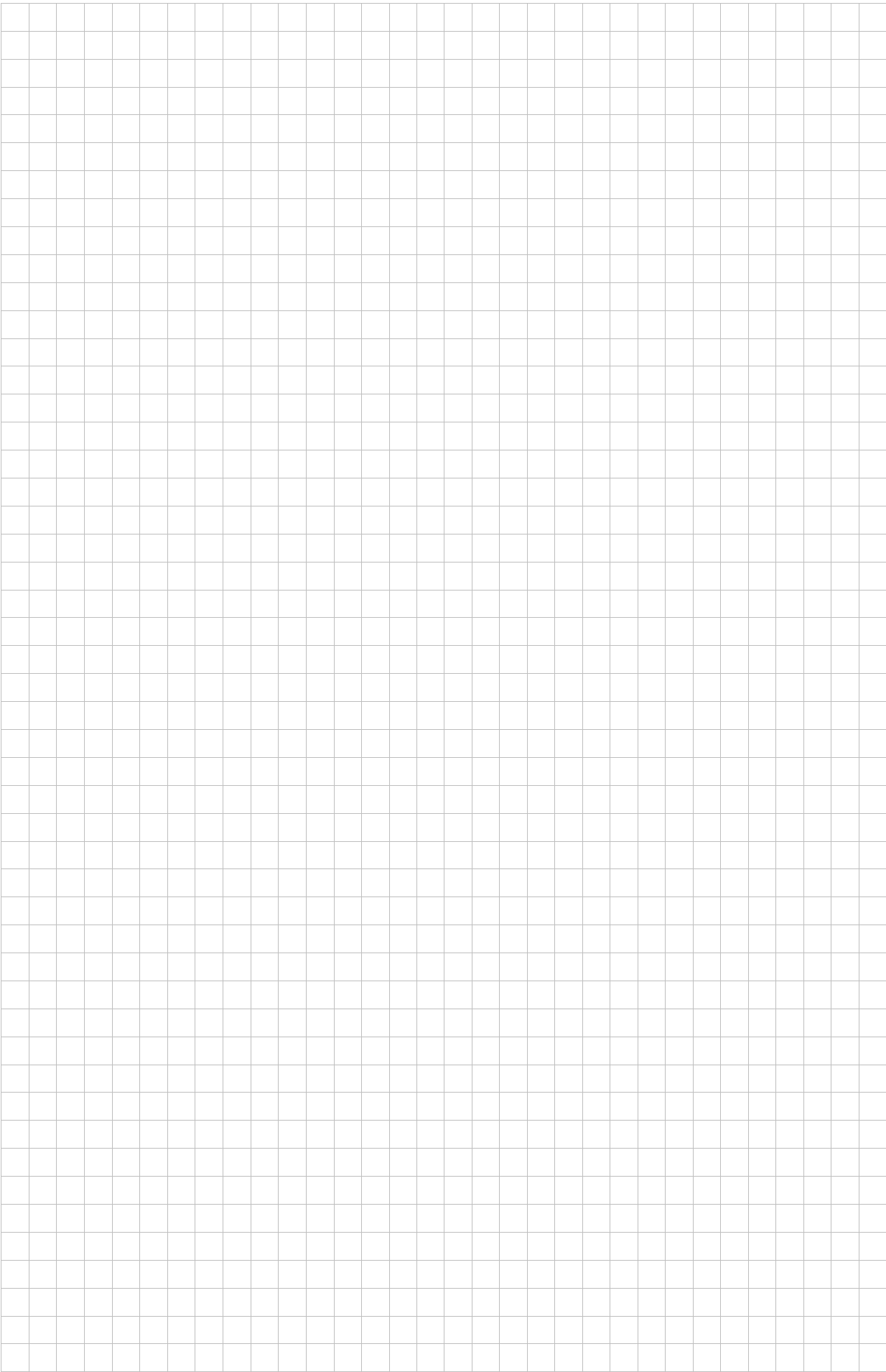
$$S_r = \frac{2G\rho_0 \left(\frac{4}{3}\pi r_0^3\right)}{c^2} \cong \frac{2G \cdot 3.61 \cdot 10^{52}}{c^2} \cong 0.53 \cdot 10^{26} \quad (51.224)$$

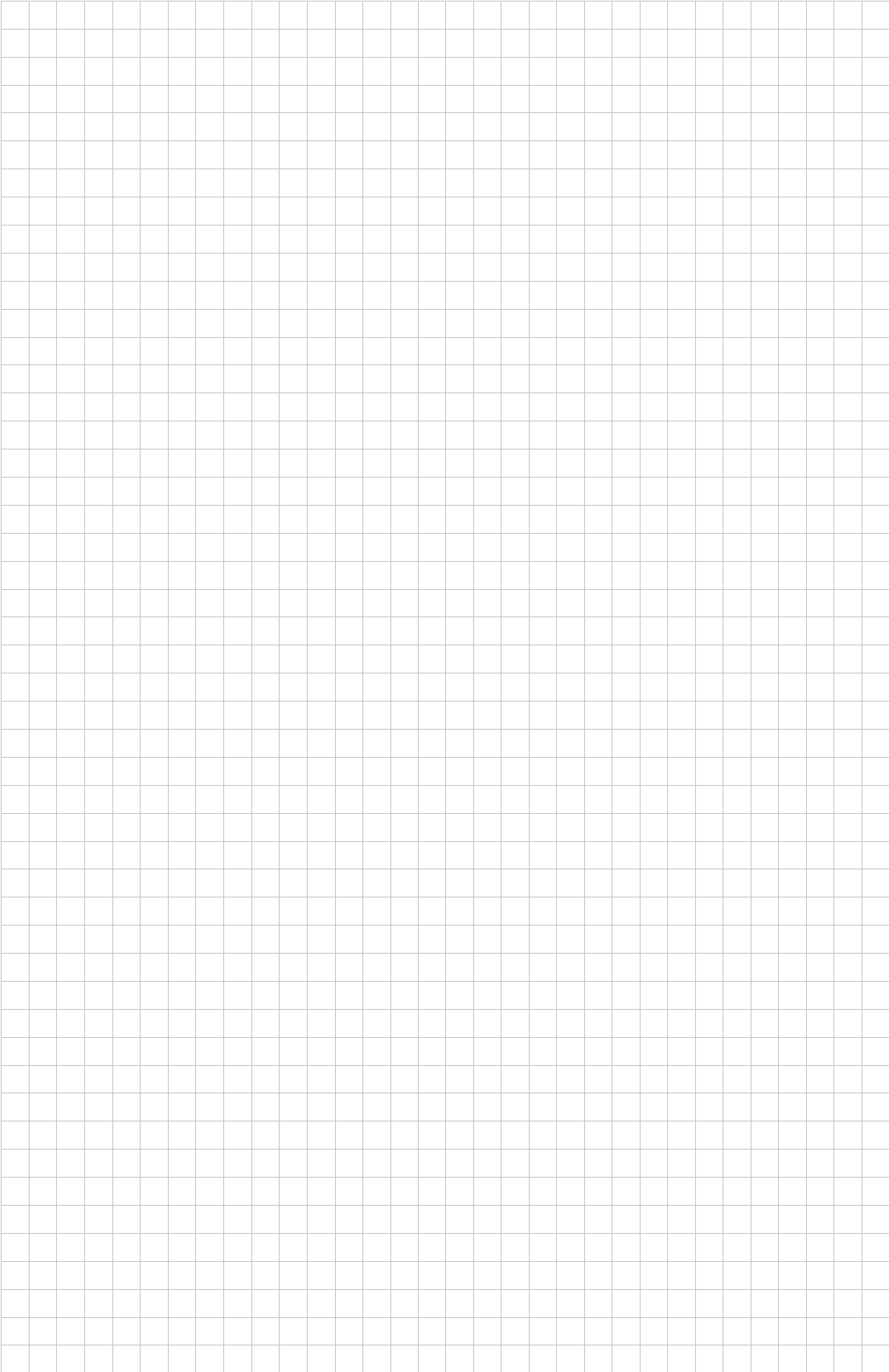
Donc grosso modo, connaissant toutes les incertitudes que nous avons accumulées en particulier celle sur la constante de Hubble nous voyons que le rayon de Schwarzschild n'est pas très loin du rayon de l'Univers actuel.

Aussi curieux que cela puisse sembler, cette question n'est pas si farfelue et est très sérieusement étudiée. Il est donc théoriquement possible que tout notre univers soit encapsulé dans un gigantesque Trou Noir (donc de très grande masse et très faible densité comme nous le voyons avec nos valeurs numériques) d'un autre univers inaccessible...

Ce qui est sûr est que si tel était le cas, l'expansion de l'Univers (observée actuellement), ne pourrait pas se poursuivre au-delà de l'horizon de ce super Trou Noir, car rien venant de l'intérieur ne peut franchir cet horizon. Or, des observations récentes semblent montrer que l'expansion de l'Univers est loin de ralentir et tend plutôt à s'accélérer avec le temps, ce qui est en contradiction avec un tel Trou Noir Univers...







52. THÉORIE DES CORDES

Il faut bien considérer dans le présent chapitre que la théorie des cordes (et in extenso des supercordes) est actuellement spéculative et n'a pas pu être vérifiée (confirmée) ni falsifiée par l'expérience comme le veut la démarche scientifique. Il convient donc de prendre avec prudence les développements qui vont suivre et d'être le plus critique possible !

Il s'agit par ailleurs d'une théorie (nous ne pouvons pas parler de modèle actuellement) d'unification des forces qui n'est pas nouvelle puisqu'elle a bientôt plus de trente ans et qui tente de combler les défauts du modèle standard des particules et aussi de réunir la relativité générale et physique quantique (ce qui n'est pas sans mal puisque cette dernière est dépendante du fond contrairement à la relativité générale). Elle est une des nombreuses théories qui existent en physique moderne et qui tentent en ce début de 21ème siècle cette unification (il en existe une dizaine d'autres plus ou moins connues).

Remarque: Si ce sujet est traité dans la section de cosmologie et non d'atomistique c'est uniquement pour une raison pédagogique. Effectivement, le formalisme de base de la théorie des cordes est beaucoup plus proche de la mécanique relativiste (relativité restreinte et générale) que de celle de la physique quantique ondulatoire ou de la physique quantique des champs. Il nous a semblé donc plus adapté, à ce jour (!), de proposer une continuité dans le formalisme mathématique et son interprétation plutôt qu'une continuité thématique avec une approche relativement différente de celle du formalisme habituel de la physique quantique.

L'avantage indéniable de la théorie des cordes, outre le fait que mathématiquement elle soit assez indigeste mais pas vraiment pire que la relativité générale, est qu'elle permet d'éviter dans un certain ordre... de nombreuses singularités dans les calculs à l'inverse d'autres théories contemporaines qui considèrent les objets comme des points (donc de volume et longueur nuls...).

Cette théorie, bien qu'étant esthétique et remarquable dans le sens qu'elle utilise pour ses fondements des bases de calculs qui ont plus de 200 ans, a pour défaut selon nous de s'imposer par analogies successives, comme nous le verrons, avec les théories relativistes et quantiques actuelles. Même si cela n'est pas dramatique en soit, la théorie peut sembler perdre un peu de son autonomie propre même si au fait il n'en est rien. Il ne faut donc pas être surpris en mal lors du parcours des développements qui vont suivre...

La principale particularité de la théorie des cordes est que son ambition ne s'arrête pas à cette réconciliation, mais qu'elle prétend réussir à unifier les quatre interactions élémentaires connues, on parle de théorie du tout, tout en reposant sur deux hypothèses:

H1. Les briques fondamentales de l'Univers ne seraient pas des particules ponctuelles, mais des sortes de cordelettes vibrantes possédant une tension à la manière d'un élastique. Ce que nous percevons comme des particules de caractéristiques (masse, etc.) distinctes ne seraient que des cordes vibrant différemment. Avec cette hypothèse, les théories des cordes admettent une échelle minimale et permettent d'éviter facilement l'apparition de certaines quantités infinies qui sont inévitables dans les théories quantiques des champs habituelles.

H2. L'Univers contiendrait plus de trois dimensions spatiales. Certaines d'entre elles, repliées sur elles-

mêmes, passant inaperçues à nos échelles (par une procédure appelée "réduction dimensionnelle").

Malgré de premiers résultats partiels très prometteurs ainsi qu'une richesse mathématique remarquable la théorie des cordes reste toutefois incomplète. D'une part, une multitude de solutions aux équations de la théorie des cordes existe, ce qui pose un problème de sélection de notre Univers et, d'autre part, même si beaucoup de modèles voisins ont pu être obtenus, aucun d'entre eux ne permet de rendre compte précisément du modèle standard de la physique des particules...

Ceci étant dit... commençons notre initiation:

1. ÉQUATION D'ONDE NON RELATIVISTE D'UNE CORDE TRANSVERSALE

L'objectif ici va être dans un premier temps de déterminer l'équation d'onde non relativiste d'une corde excitée transversalement à l'aide des calculs que nous avons effectués dans le chapitre de Mécanique Ondulatoire. Une fois ce travail effectué, nous passerons à l'étude des cordes relativistes et nous verrons que leur équation d'onde, au même titre que la version non relativiste, peut s'assimiler à l'équation de conservation du courant que nous avons démontrée en électrodynamique.

Nous commençons en rappelant la forme de l'action que nous avons obtenue dans le chapitre de Mécanique Ondulatoire pour une corde non relativiste:

$$S = \int_{t_i}^{t_f} dt \int_0^L dx \mathcal{L} \left(\frac{\partial y}{\partial t}, \frac{\partial y}{\partial x} \right) \quad (52.1)$$

avec donc:

$$L = \int_0^L \left[\frac{1}{2} \rho_t \left(\frac{\partial y}{\partial t} \right)^2 - \frac{1}{2} F \left(\frac{\partial y}{\partial x} \right)^2 \right] dx = \int_0^L \mathcal{L} dx \quad (52.2)$$

Maintenant, de manière identique à ce que nous avons fait dans le chapitre de Mécanique Analytique (ainsi que dans celui de Physique Quantique Des Champs), nous allons définir une notation par une analogie aux moments canoniques de la corde:

$$\mathcal{P}^t \equiv \frac{\partial \mathcal{L}}{\partial \dot{y}} \quad \mathcal{P}^x \equiv \frac{\partial \mathcal{L}}{\partial y'} \quad (52.3)$$

avec $y' = \partial y / \partial x$. Il s'agit simplement des dérivées de la densité lagrangienne en fonction respectivement du premier et second argument. De manière plus explicite, nous avons alors directement en faisant le calcul (cf. chapitre de Mécanique Ondulatoire):

$$\mathcal{P}^t \equiv \rho_t \frac{\partial y}{\partial t} \quad \mathcal{P}^x \equiv -F \frac{\partial y}{\partial x} \quad (52.4)$$

Ainsi, si nous réécrivons le variationnel d'action obtenu dans le chapitre de Mécanique Ondulatoire avec cette notation canonique, nous obtenons:

$$S = \int_{t_i}^{t_f} dt \int_0^L dx \left[\rho_t \frac{\partial y}{\partial t} \frac{\partial (\delta y)}{\partial t} - F \frac{\partial y}{\partial x} \frac{\partial (\delta y)}{\partial x} \right] = \int_{t_i}^{t_f} dt \int_0^L dx \left[\frac{\partial \mathcal{L}}{\partial \dot{y}} \delta \dot{y} + \frac{\partial \mathcal{L}}{\partial y'} \delta y' \right] \quad (52.5)$$

Faisant usage des mêmes méthodes que dans le chapitre de Mécanique Ondulatoire, notre variationnel s'exprime après simplification à nouveau sous la forme de trois termes:

$$\begin{aligned}\delta S &= \int_0^L \left[\rho_t \frac{\partial y}{\partial t} \delta y \right]_{t_i}^{t_f} dx + \int_{t_i}^{t_f} \left[-F \frac{\partial y}{\partial x} \delta y \right]_0^L dt - \int_{t_i}^{t_f} dt \int_0^L dx \left(\rho_t \frac{\partial^2 y}{\partial t^2} - F \frac{\partial^2 y}{\partial x^2} \right) \delta y = \\ &= \int_0^L \left[\mathcal{P}^* \delta y \right]_{t_i}^{t_f} dx + \int_{t_i}^{t_f} \left[\mathcal{P}^* \delta y \right]_0^L dt - \int_{t_i}^{t_f} dt \int_0^L dx \left(\frac{\partial \mathcal{P}^*}{\partial t} + \frac{\partial \mathcal{P}^*}{\partial x} \right) \delta y\end{aligned}\quad (52.6)$$

Les conditions pour trouver l'extremum (selon le principe de moindre action) restent les mêmes que celles vues dans le chapitre de Mécanique Ondulatoire. Ainsi, pour le troisième terme, nous avons bien l'équation d'onde d'une corde excitée de manière transversale donnée avec la forme canonique par:

$$\boxed{\frac{\partial \mathcal{P}^*}{\partial t} + \frac{\partial \mathcal{P}^*}{\partial x} = 0} \quad (52.7)$$

Remarque: Il convient bien évidemment de remarquer que cette forme d'écriture va considérablement nous faciliter la tâche!

Il faut bien observer (car c'est remarquable!) aussi que comme dans le chapitre de Mécanique Analytique, le moment canonique $\mathcal{P}^*$ tel que défini plus haut, coïncide parfaitement (le hasard fait bien les choses...) avec la densité de quantité de mouvement que nous avons obtenue dans le chapitre de Mécanique Ondulatoire. Effectivement:

$$p_y = \int_0^L (\rho_t dx) \frac{\partial y}{\partial t} = \int_0^L \rho_t \frac{\partial y}{\partial t} dx = \int_0^L \mathcal{P}^* dx \quad (52.8)$$

Ainsi, par analogie avec la mécanique analytique (où rappelons-le, la dérivée du lagrangien par rapport à la vitesse donne la quantité de mouvement), $\dot{y}$ joue bien le rôle de la vitesse et ainsi la dérivée de la densité lagrangienne par celui-ci donne la densité de quantité de mouvement $\mathcal{P}^*$!!!

Rappelons aussi un autre point qui a été vu dans le chapitre de Mécanique Ondulatoire, l'extremum de l'action ($\delta S = 0$) nous impose les conditions de Neumann, ce qui nous amène à écrire $\mathcal{P}^* = 0$.

De plus, il convient aussi de rappeler pour ce qui va suivre, que pour les conditions de Dirichlet nous avons aussi $\mathcal{P}^* = 0$.

Remarque: Dans le cadre de la théorie des cordes relativistes à plus de 3 dimensions, il est possible de généraliser le concept de conditions aux limites en considérant les contraintes dans l'espace comme des hypersurfaces nommées Dp -branes à p dimensions. Les conditions aux limites de Dirichlet usuelles correspondent alors à la situation où les bouts d'une corde sont contraints par une 0-brane. La condition de Neumann pour une corde libre dans p dimensions correspond à une corde contrainte sur une Dp -brane.

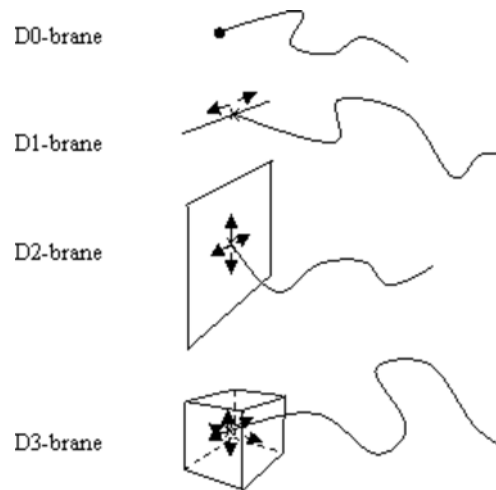


Figure: 52.1 - Illustration des Dp -branes

2. ÉQUATION D'ONDE RELATIVISTE D'UNE CORDE TRANSVERSALE

Nous allons maintenant déterminer l'action d'une corde relativiste. Nous pouvons, pour poser les bases de notre étude, nous rappeler qu'une particule ponctuelle trace une ligne dans l'espace-temps (chaque point de la ligne étant repéré par une coordonnée temporelle et trois spatiales). Dès lors, par extension, une corde qui est un élément bidimensionnel (si nous la considérons sans épaisseur) trace une surface dans l'espace-temps.

Ainsi, au même titre que la ligne que trace une particule dans l'espace-temps est appelée une "ligne d'Univers" (cf. chapitre de Relativité Restreinte), la surface tracée par une corde sera appelée par analogie une "**surface d'Univers**".

Une corde fermée dans l'espace-temps de Minkowski trace, par exemple, un tube, alors qu'une corde ouverte tracera une bande:

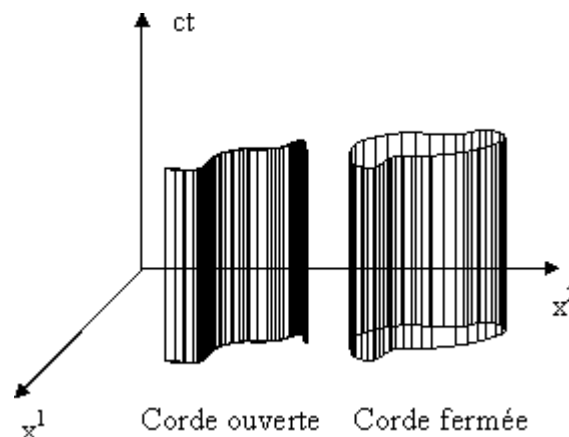


Figure: 52.2 - Surface d'Univers générée par une corde ouverte/respectivement fermée

Sur la figure ci-dessus, à deux dimensions spatiales et une temporelle, la corde est immobile dans notre espace courant. Elle ne se meut que dans l'espace-temps (car le temps s'écoule sur l'axe vertical) mais pas dans l'espace dans l'exemple ci-dessus (il faudrait une composante spatiale supplémentaire pour voir un tel mouvement).

Remarques:

R1. Attention! Rappelez-vous bien que le schéma ci-dessous est dans trois dimensions alors que l'espace-temps compte lui quatre dimensions.

R2. Rappelez-vous également que le vecteur temps de la base orthogonale est toujours perpendiculaire à toutes les autres composantes spatiales (cette remarque sera utile lors de notre démonstration de l'action de Nambu-Goto).

Lors de notre démonstration de l'équation du mouvement dans le chapitre de Relativité Générale, nous avons reparamétrisé la ligne d'Univers de la particule à l'aide d'un paramètre qui était le temps propre de la particule t . Effectivement, il suffit de se rappeler des équations paramétriques qui représentent des courbes. Par exemple avec Maple 4.00b:

```
> with(plots):
> spacecurve([cos(t),sin(t),t],t=0..4*Pi,axes=boxed);
```

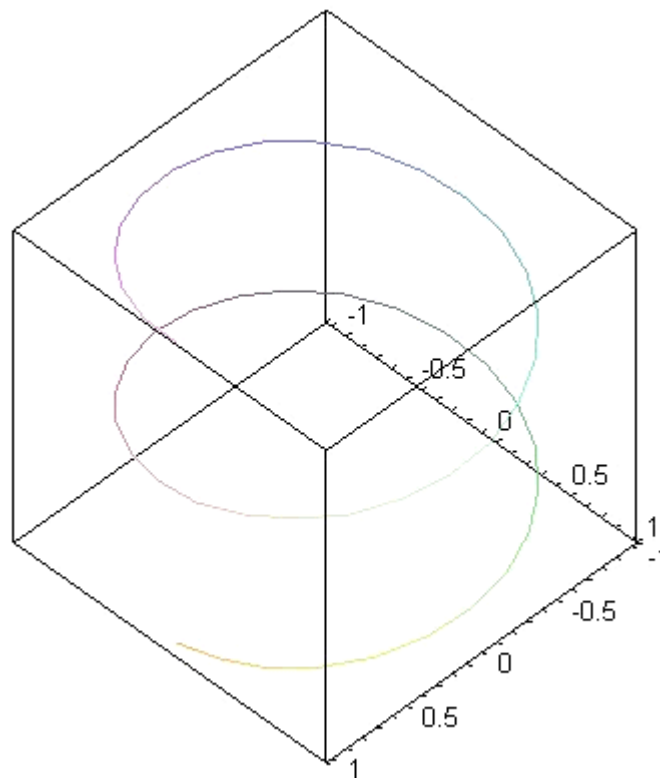


Figure: 52.3 - Rappel élémentaire illustré de la paramétrisation d'une courbe

et la même procédure est valable pour une ligne en quatre dimensions (espace + temps).

Nous étions ainsi arrivés à construire l'expression de l'action S de celle-ci avant d'y appliquer le principe

variationnel.

Nous allons faire de même pour une corde relativiste à la différence que nous allons reparamétriser les surfaces engendrées par les cordes cette fois-ci. Les contraintes que nous nous imposons sont que les paramètres choisis devront aussi (en faisant référence au cas de la particule) être des invariants relativistes.

Comme nous l'avons donc vu dans le chapitre de Relativité Générale, une ligne d'Univers peut être reparamétrisée naturellement en utilisant seulement un paramètre (abscisse curviligne). Une surface dans l'espace étant cependant un objet bidimensionnel, nous supposons qu'il requiert par extension deux paramètres ξ^1, ξ^2 (un de plus) pour être décrit complètement.

Effectivement, nous devinons, qu'un des deux paramètres sera le temps propre (pour faire évoluer la surface dans le temps), le second paramètre permettra de donner une "épaisseur" à ce qui ne serait qu'une ligne d'Univers s'il n'existait pas. Il suffirait dans un espace à trois dimensions que ce deuxième paramètre ait pour générer une surface les dimensions d'une longueur mais dans l'espace-temps à quatre dimensions il faut que second paramètre ait les unités d'une surface.

Étant donnée une surface paramétrée, nous pouvons dessiner sur celle-ci les isolignes des paramètres ξ^1, ξ^2 (les lignes où les deux paramètres sont constants sur toute la surface). Ces isolignes couvrent la surface comme une grille (voir figure un peu plus bas).

L'équation paramétrique d'un volume requiert dans l'espace trois paramètres comme nous l'avons vu dans le chapitre de Géométrie Analytique. Ainsi, si une surface paramétrée peut dans l'espace euclidien être représentée par un vecteur du type:

$$\vec{f} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \quad (52.9)$$

lors d'une reparamétrisation et en faisant usage de la notation tensorielle de l'espace-temps de Minkowski telle que vue dans le chapitre de Relativité Générale, nous aurons (en nous restreignant pour l'instant aux cas particuliers de deux dimensions spatiales et une temporelle):

$$\vec{x} = \begin{pmatrix} x^1(\xi^1, \xi^2) \\ x^2(\xi^1, \xi^2) \\ x^3(\xi^1, \xi^2) \end{pmatrix} \quad (52.10)$$

Ainsi, la surface est l'image des paramètres ξ^1, ξ^2 . Alternativement, nous pouvons voir les composantes ξ^1, ξ^2 comme les coordonnées de temps et d'espace de la surface, au moins localement!

Nous voulons maintenant calculer la surface d'un élément de n'importe quel type d'espace au même titre que nous l'avons fait pour l'abscisse curviligne de n'importe quelle ligne d'Univers dans le chapitre de Relativité Générale. Se pose alors la question de la forme de l'élément différentiel de surface ??? Faut-il prendre la multiplication du différentiel des deux paramètres choisis précédemment pour un carré, un rectangle, un cercle ou autre ?

Au fait, nous allons reporter notre choix sur un parallélogramme ! Ce choix peut sembler complètement

arbitraire pour l'instant mais comme nous allons le voir quelques lignes plus loin, ce choix coïncide pour des raisons mathématiques à ce que nous appelons la "métrique induite" de la surface elle-même (résultat assez remarquable!).

Ainsi, notons $d\vec{v}_1$ et $d\vec{v}_2$ les côtés du parallélogramme. Ils sont l'image par $\vec{x}$ des couples $(\xi^1, 0)$ et respectivement $(0, \xi^2)$:

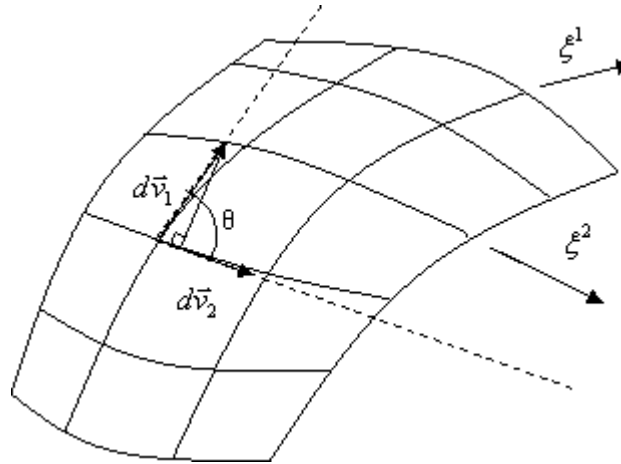


Figure: 52.4 - Configuration pour l'étude d'un élément de surface élémentaire

Ainsi, nous pouvons écrire:

$$\frac{d\vec{v}_1}{d\xi^1} = \frac{\partial \vec{x}}{\partial \xi^1}, \quad \frac{d\vec{v}_2}{d\xi^2} = \frac{\partial \vec{x}}{\partial \xi^2} \quad (52.11)$$

et donc:

$$d\vec{v}_1 = \frac{\partial \vec{x}}{\partial \xi^1} d\xi^1, \quad d\vec{v}_2 = \frac{\partial \vec{x}}{\partial \xi^2} d\xi^2 \quad (52.12)$$

Maintenant calculons la surface dA (nous ne prendrons pas la lettre S pour éviter la confusion avec l'action dans ce chapitre) du parallélogramme (cf. chapitre de Calcul Vectoriel):

$$dA = \|d\vec{v}_1\| \|d\vec{v}_2\| \sin \theta = \|d\vec{v}_1\| \|d\vec{v}_2\| \sqrt{1 - \cos^2 \theta} = \sqrt{\|d\vec{v}_1\|^2 \|d\vec{v}_2\|^2 - \|d\vec{v}_1\|^2 \|d\vec{v}_2\|^2 \cos^2 \theta} \quad (52.13)$$

en utilisant le produit scalaire, cela peut se réécrire:

$$dA = \sqrt{(d\vec{v}_1 \circ d\vec{v}_1)(d\vec{v}_2 \circ d\vec{v}_2) - (d\vec{v}_1 \circ d\vec{v}_2)^2} \quad (52.14)$$

en utilisant, les relations établies précédemment cela peut s'écrire:

$$dA = d\xi^1 d\xi^2 \sqrt{\left(\frac{\partial \vec{x}}{\partial \xi^1} \circ \frac{\partial \vec{x}}{\partial \xi^1} \right) \left(\frac{\partial \vec{x}}{\partial \xi^2} \circ \frac{\partial \vec{x}}{\partial \xi^2} \right) - \left(\frac{\partial \vec{x}}{\partial \xi^1} \circ \frac{\partial \vec{x}}{\partial \xi^2} \right)^2} \quad (52.15)$$

cette dernière relation est la forme générale d'un élément de surface d'une nappe paramétrée. La surface totale étant évidemment donnée par:

$$A = \int d\xi^1 d\xi^2 \sqrt{\left(\frac{\partial \vec{x}}{\partial \xi^1} \circ \frac{\partial \vec{x}}{\partial \xi^1} \right) \left(\frac{\partial \vec{x}}{\partial \xi^2} \circ \frac{\partial \vec{x}}{\partial \xi^2} \right) - \left(\frac{\partial \vec{x}}{\partial \xi^1} \circ \frac{\partial \vec{x}}{\partial \xi^2} \right)^2} \quad (52.16)$$

Au même titre que dans le cadre de l'étude principe de moindre action (cf. chapitre de Mécanique Analytique) nous avons cherché le chemin optimum pour une particule parcourant une ligne d'Univers, pour une corde, nous aurons à optimiser la surface A en minimisant la fonction $\vec{x}(\xi^1, \xi^2)$.

Cette dernière forme est cependant un peu lourde et ne fait rien ressortir de particulier ou de similaire à quelque forme déjà connue dans un autre domaine de la physique. Nous allons voir cependant qu'en creusant un peu il est cependant possible d'obtenir quelque chose de pas mal du tout.

Considérons maintenant un vecteur $d\vec{x}$ et sa longueur (norme) au carré donnée par son produit scalaire:

$$ds^2 = (ds)^2 = d\vec{x} \circ d\vec{x} \quad (52.17)$$

Attention à l'avenir de ne pas "voir" le s comme étant élevé au carré dans le ds (comme c'est le cas en relativité restreinte et générale) mais rappelez-vous bien qu'il s'agit du ds en entier qui est mis au carré (la notation peut amener à confusion...).

Le vecteur $d\vec{x}$ peut être exprimé sous forme de termes de dérivées partielles de $d\xi^1, d\xi^2$, tel que nous obtenions sa différentielle totale exacte (cf. chapitre de Calcul Différentiel Et Intégral):

$$d\vec{x} = \frac{\partial \vec{x}}{\partial \xi^1} d\xi^1 + \frac{\partial \vec{x}}{\partial \xi^2} d\xi^2 = \frac{\partial \vec{x}}{\partial \xi^i} d\xi^i \quad (52.18)$$

Ainsi, la longueur au carré de $d\vec{x}$ peut s'exprimer sous la forme tensorielle:

$$ds^2 = \left(\frac{\partial \vec{x}}{\partial \xi^i} d\xi^i \right) \circ \left(\frac{\partial \vec{x}}{\partial \xi^j} d\xi^j \right) = \frac{\partial \vec{x}}{\partial \xi^i} \circ \frac{\partial \vec{x}}{\partial \xi^j} d\xi^i d\xi^j \quad (52.19)$$

ce que nous noterons par convention à l'avenir:

$$ds^2 = g_{ij}(\xi) d\xi^i d\xi^j \quad (52.20)$$

La quantité $g_{ij}(\xi)$ est appelée la "**métrique induite de la surface paramétrée**" (car contient un produit scalaire ce qui en toute généralité fait appel à une métrique... d'où le terme "induite") et il s'agit donc d'une matrice de dimensions 2×2 . Il est évident que le choix de cette dénomination provient de la ressemblance avec la métrique habituelle telle que nous l'avons définie lors de notre étude du calcul tensoriel et de son utilisation en relativité restreinte et générale.

La matrice g_{ij} a donc par construction et définition la forme:

$$g_{ij} = \begin{bmatrix} \frac{\partial \vec{x}}{\partial \xi^1} \circ \frac{\partial \vec{x}}{\partial \xi^1} & \frac{\partial \vec{x}}{\partial \xi^1} \circ \frac{\partial \vec{x}}{\partial \xi^2} \\ \frac{\partial \vec{x}}{\partial \xi^2} \circ \frac{\partial \vec{x}}{\partial \xi^1} & \frac{\partial \vec{x}}{\partial \xi^2} \circ \frac{\partial \vec{x}}{\partial \xi^2} \end{bmatrix} \quad (52.21)$$

Revenons maintenant à notre expression de la surface engendrée par la corde:

$$A = \int d\xi^1 d\xi^2 \sqrt{\left(\frac{\partial \vec{x}}{\partial \xi^1} \circ \frac{\partial \vec{x}}{\partial \xi^1} \right) \left(\frac{\partial \vec{x}}{\partial \xi^2} \circ \frac{\partial \vec{x}}{\partial \xi^2} \right) - \left(\frac{\partial \vec{x}}{\partial \xi^1} \circ \frac{\partial \vec{x}}{\partial \xi^2} \right)^2} \quad (52.22)$$

et calculons rapidement le déterminant (cf. chapitre d'Algèbre Linéaire) de la matrice g_{ij} :

$$g = g_{ij} = g_{11}g_{22} - g_{21}g_{12} = \left(\frac{\partial \vec{x}}{\partial \xi^1} \circ \frac{\partial \vec{x}}{\partial \xi^1} \right) \left(\frac{\partial \vec{x}}{\partial \xi^2} \circ \frac{\partial \vec{x}}{\partial \xi^2} \right) - \left(\frac{\partial \vec{x}}{\partial \xi^1} \circ \frac{\partial \vec{x}}{\partial \xi^2} \right)^2 \quad (52.23)$$

et donc quoi ? Eh ben voilà:

$$A = \int d\xi^1 d\xi^2 \sqrt{g} \quad (52.24)$$

Ainsi, le choix du parallélogramme comme surface élémentaire s'explique mieux ici!

Maintenant, nous allons adopter les écritures traditionnelles de la théorie des cordes relativement à l'expression de la surface. Ainsi, au même titre que les coordonnées d'espace-temps sont décrites en relativité restreinte par le quadrivecteur temps-espace:

$$x^\mu = (x^0, x^1, \dots, x^d) \quad (52.25)$$

nous décrirons les surfaces d'Univers par (nous passons maintenant à l'écriture faisant usage des 4 dimensions de l'espace-temps):

$$x^\mu(\tau, \sigma) = X^\mu(\tau, \sigma) \quad (52.26)$$

Cette notation nous évitera à l'avenir d'avoir à confondre, si la théorie nous y amène, les coordonnées d'espace-temps x^μ traditionnelles avec la fonction image de la surface d'Univers $x^\mu(\tau, \sigma)$ et ce d'autant plus que les physiciens étant un peu flemmards abrègent parfois cette dernière x^μ ... d'où le choix de la majuscule.

Il est donc beaucoup plus convenable et sage de changer de notation...

À partir de maintenant, nous appellerons "**coordonnées de corde**" la surface d'Univers décrite par X^μ .

Ce petit changement de notation ne change évidemment pas l'interprétation de la fonction image. Étant donnée un couple (τ, σ) associant élément de temps propre dans l'ensemble et élément de surface des pré-images, ce point est projeté sur un élément de surface de l'espace-temps de la corde de coordonnées:

$$(X^0(\tau, \sigma), X^1(\tau, \sigma), \dots, X^d(\tau, \sigma)) \quad (52.27)$$

2.1. ACTION DE NAMBU-GOTO

Dans le cas d'une surface d'Univers les paramètres sont donc par convention τ et σ , où comme en relativité restreinte et générale le temps propre peut être compris dans l'intervalle:

$$\tau \in]-\infty, +\infty[\quad (52.28)$$

le deuxième paramètre par contre ne pouvant être que positif puisqu'il s'agit d'une surface:

$$\sigma \in]0, \sigma_1[\quad (52.29)$$

et les coordonnées de cette surface qui correspondent à l'espace des paramètres étant donc:

$$X^\mu(\tau, \sigma) \quad (52.30)$$

Où encore une fois pour rappel, le paramètre τ est considéré comme la variable décrivant l'écoulement du temps (il en faut bien une!), et σ la variable décrivant l'extension dans l'espace d'une corde (i.e. la condition $\sigma \in]0, \sigma_1[$ impliquant la longueur finie de cette corde).

Les paramètres (τ, σ) décrivent ainsi une surface de l'espace des pré-images:

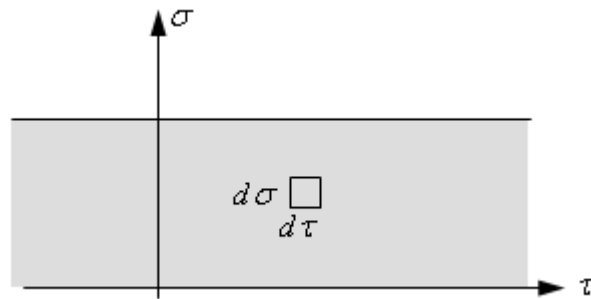


Figure: 52.5 - Paramétrisation d'une surface de l'espace-temps

Les extrémités de la corde ont une valeur σ constante. Cependant, comme le temps s'écoule et que les extrémités de la corde sur la surface d'Univers se meuvent il faut noter une condition essentielle de la surface d'Univers concernant les deux bouts d'une corde ouverte:

$$\frac{\partial X^0}{\partial \tau} \neq 0 \quad (52.31)$$

Remarque: Cette condition se fait sur la composante X^0 car elle correspond à la composante x^0 du quadrivecteur d'espace-temps qui n'est autre, en unités naturelles, que t (le temps propre). Dès lors, le temps s'écoule et n'est jamais constant d'où le fait d'imposer cette dérivée comme différente de zéro.

Et en utilisant les conventions habituelles en physique pour la notation des dérivées par rapport au temps ou composante spatiale, nous convenons d'adopter aussi maintenant les écritures suivantes:

$$\dot{X} = \frac{\partial X^\mu}{\partial \tau} \quad X' = \frac{\partial X^\mu}{\partial \sigma} \quad (52.32)$$

où puisque:

$$X^\mu(\tau, \sigma) = X^\mu(t, \sigma) = (ct, \vec{X}(t, \sigma)) \quad (52.33)$$

alors:

$$\begin{aligned} \frac{\partial X^\mu}{\partial \sigma} &= \left(\frac{\partial X^0}{\partial \sigma}, \frac{\partial \vec{X}}{\partial \sigma} \right) = \left(0, \frac{\partial \vec{X}}{\partial \sigma} \right) \\ \frac{\partial X^\mu}{\partial \tau} &= \left(\frac{\partial X^0}{\partial \tau}, \frac{\partial \vec{X}}{\partial \tau} \right) = \left(c, \frac{\partial \vec{X}}{\partial \tau} \right) \end{aligned} \quad (52.34)$$

La surface s'écrit donc:

$$\begin{aligned} A &= \int d\xi^1 d\xi^2 \sqrt{\left(\frac{\partial \vec{x}}{\partial \xi^1} \circ \frac{\partial \vec{x}}{\partial \xi^1} \right) \left(\frac{\partial \vec{x}}{\partial \xi^2} \circ \frac{\partial \vec{x}}{\partial \xi^2} \right) - \left(\frac{\partial \vec{x}}{\partial \xi^1} \circ \frac{\partial \vec{x}}{\partial \xi^2} \right)^2} \\ &= \int d\tau d\sigma \sqrt{\left(\frac{\partial X^\mu}{\partial \tau} \frac{\partial X_\mu}{\partial \tau} \right) \left(\frac{\partial X^\nu}{\partial \sigma} \frac{\partial X_\nu}{\partial \sigma} \right) - \left(\frac{\partial X^\mu}{\partial \tau} \frac{\partial X_\mu}{\partial \sigma} \right)^2} \\ &= \int d\tau d\sigma \sqrt{(\dot{X})^2 (X')^2 - (\dot{X} \circ X')^2} \end{aligned} \quad (52.35)$$

Cependant, il y a un problème ici ! Effectivement, regardons si le radicande (terme sous la racine) a une réalité physique tangible...

Pour cela, il faut d'abord considérer la partie gauche de la figure ci-dessous qui représente la surface (nappe) décrite par une corde ouverte:

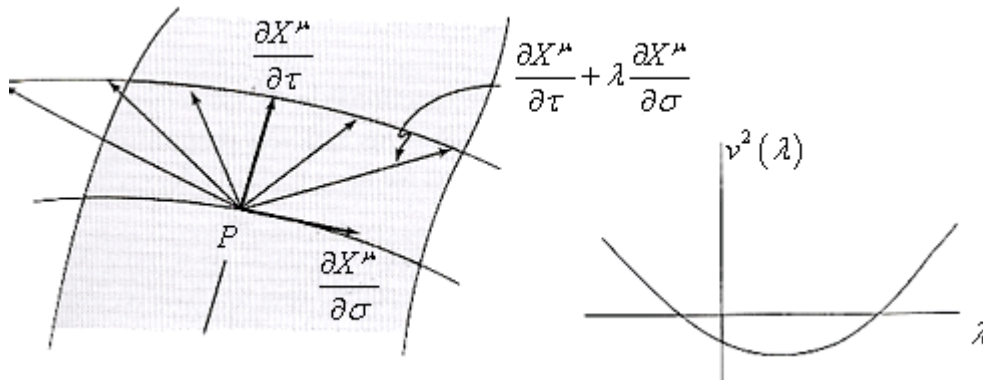


Figure: 52.6 - Configuration pour l'étude du radicande

En chaque point P de cette nappe (supposée dérivable en tout point) il existe une infinité de tangentes, toutes dans le même plan, que nous noterons pour l'exemple $\vec{v}$ et qui forment donc une surface tangente au point P .

Maintenant, comme l'espace dans lequel la nappe de la corde est plongée se déroule dans une base orthonormale spatiale et temporelle, les vecteurs tangents $\vec{v}$ peuvent alors à leur tour être décomposés dans une base orthogonale spatiale et temporelle locale bidimensionnelle au point P tel que les vecteurs de cette base soient deux vecteurs:

$$\frac{\partial X^\mu}{\partial \tau} \quad \frac{\partial X^\mu}{\partial \sigma} \quad (52.36)$$

tous les autres vecteurs tangents s'expriment comme combinaison linéaire de ceux-ci.

Cependant un problème subsiste dans notre décomposition (...): les unités des vecteurs de la base orthogonale locale au point P ont des unités qui diffèrent. Pour cela, rajoutons un facteur de dimensionnement à la composante spatiale (cela est arbitraire car la conclusion sera identique quelle que soit la composante sur laquelle vous mettez le facteur de dimensionnement):

$$\frac{\partial X^\mu}{\partial \tau} \quad \lambda \frac{\partial X^\mu}{\partial \sigma} \quad (52.37)$$

ce facteur de dimensionnement peut aussi être utilisé pour obtenir tous les vecteurs tangents tel que:

$$v^\mu = \frac{\partial X^\mu}{\partial \tau} + \lambda \frac{\partial X^\mu}{\partial \sigma} \quad (52.38)$$

Effectivement, si $\lambda \in [-\infty, +\infty]$, alors pour $\lambda = 0$ nous obtenons le vecteur $\partial X^\mu / \partial \tau$ et pour $\lambda = +\infty$ le vecteur $\partial X^\mu / \partial \sigma$. Et pour toutes les valeurs intermédiaires, nous obtenons tous les vecteurs tangents comme indiqué sur la partie gauche de la figure précédente.

Maintenons, rappelons que nous avons vu dans le chapitre de Relativité Restreinte qu'il existait selon l'abscisse curviligne:

$$ds^2 = dx^\nu (\eta_{\nu\mu} dx^\mu) = dx^\nu dx_\nu = c^2 dt^2 - dx^2 - dy^2 - dz^2 \quad (52.39)$$

des lignes d'Univers de type lumière ($ds^2 = 0$), espace ($ds^2 < 0$) ou temps ($ds^2 > 0$) si nous considérons les quadrivecteurs x^μ .

Il doit en être de même par analogie pour les vecteurs tangents à la surface et donc données par:

$$v^\mu = \frac{\partial X^\mu}{\partial \tau} + \lambda \frac{\partial X^\mu}{\partial \sigma} \quad (52.40)$$

Ainsi:

$$v^\mu v_\mu = \lambda^2 \left(\frac{\partial X^\mu}{\partial \sigma} \right)^2 + 2\lambda \left(\frac{\partial X^\mu}{\partial \tau} \circ \frac{\partial X^\mu}{\partial \sigma} \right) + \left(\frac{\partial X^\mu}{\partial \tau} \right)^2 \quad (52.41)$$

ce qui correspond à une équation du deuxième degré en λ , doit pour avoir des valeurs négatives (surface d'Univers de type temps) ou positives (surface d'Univers de type espace) avoir au moins deux racines (voir partie droite de la figure précédente). Cela nous ramène à la condition que le discriminant soit strictement positif (cf. chapitre de Calcul Algébrique):

$$\Delta = b^2 - 4ac = v^\mu v_\mu = 4 \left(\frac{\partial X^\mu}{\partial \tau} \circ \frac{\partial X^\mu}{\partial \sigma} \right)^2 - 4 \left(\frac{\partial X^\mu}{\partial \sigma} \right)^2 \left(\frac{\partial X^\mu}{\partial \tau} \right)^2 > 0 \quad (52.42)$$

Soit:

$$\left(\frac{\partial X^\mu}{\partial \tau} \circ \frac{\partial X^\mu}{\partial \sigma} \right)^2 - \left(\frac{\partial X^\mu}{\partial \sigma} \right)^2 \left(\frac{\partial X^\mu}{\partial \tau} \right)^2 > 0 \quad (52.43)$$

sous forme condensée cela nous ramène à écrire:

$$\left(\dot{X} \circ X' \right)^2 - \left(\dot{X} \right)^2 \left(X' \right)^2 > 0 \quad (52.44)$$

La surface doit donc alors s'écrire en fin de compte:

$$A = \int d\tau d\sigma \sqrt{\left(\dot{X} \circ X' \right)^2 - \left(\dot{X} \right)^2 \left(X' \right)^2} \quad (52.45)$$

si nous voulons que le radicande ait un sens physique.

Rappelons maintenant que l'action S d'une particule ponctuelle est proportionnelle à sa ligne d'Univers. Ainsi par analogie, l'action S d'une corde sera proportionnelle à la surface d'Univers:

$$S \propto \int dA \Leftrightarrow S \propto A \quad (52.46)$$

ce qui donne:

$$S \propto \int d\tau d\sigma \sqrt{\left(\dot{X} \circ X' \right)^2 - \left(\dot{X} \right)^2 \left(X' \right)^2} \quad (52.47)$$

Ce qui nous amène très fréquemment dans la littérature à trouver l'action d'une corde sous la forme suivante:

$$S = \int d\sigma d\tau \mathcal{L}(\sigma, \tau) \quad (52.48)$$

Relation à comparer avec le lagrangien d'une particule libre (cf. chapitre de Mécanique Analytique) et la densité lagrangienne d'un champ (cf. chapitre de Physique Quantique Des Champs):

$$S(q, \dot{q}) = \int_{t_i}^{t_f} L(q, \dot{q}) dt \quad \text{et} \quad S = \int \mathcal{L}(\Phi^A, \dot{\Phi}^A, \partial_\mu \Phi^A) dx^4 \quad (52.49)$$

La fonctionnelle S a pour unités celles d'une surface. Cela parce que les X^μ ont une unité de longueur et dans la racine chacun est à la puissance quatrième et que les unités τ, σ s'annulent entre l'intérieure de la racine et les différentielles en dehors.

Maintenant, par définition même de l'action, les unités que nous devons obtenir doivent correspondre à celle d'une énergie multipliée par le temps, des joules J ou en utilisant le système international, des $\text{kg} \cdot \text{m}^2 \cdot \text{s}^{-1}$. Pour l'instant, nous avons:

$$[\tau] = s \quad [\sigma] = m \quad [X^\mu] = m \quad [A] = m^2 \quad (52.50)$$

Pour obtenir pour l'action les unités que nous voulons, il nous faut alors multiplier l'expression de la

surface A par une quantité ayant pour unités des $\text{kg} \cdot \text{s}^{-1}$. Pour choisir ces quantités, nous allons nous inspirer de notre étude de la mécanique ondulatoire. Quand nous avons travaillé avec des cordes (non relativistes) nous avons vu que les propriétés à prendre en compte étaient la tension et la vitesse de l'onde de propagation de la corde. Nous allons donc faire l'essai de prendre le rapport tension/vitesse suivant:

$$\frac{\text{tension}}{\text{vitesse}} = \frac{N}{\text{m} \cdot \text{s}^{-1}} = \frac{\text{kg} \cdot \text{m} \cdot \text{s}^{-2}}{\text{m} \cdot \text{s}^{-1}} = \text{kg} \cdot \text{s}^{-1} = \frac{T}{v} = \frac{T_0}{c} \quad (52.51)$$

où apparaissent donc la "**tension de la corde au repos**" et la vitesse de la lumière.

Remarque: Cela est similaire à la physique du point où dans l'action nous retrouvons la masse au repos (équivalente de la tension au repos de la corde) et la vitesse de la lumière (cf. chapitre de Relativité Restreinte).

Ainsi, "**l'action de Nambu-Goto**" s'écrit maintenant:

$$S = -\frac{T_0}{c} \int d\tau d\sigma \sqrt{(\dot{X} \circ X')^2 - (\dot{X})^2 (X')^2} \quad (52.52)$$

Remarque: Nous démontrerons pourquoi nous avons posé un facteur "-" plus loin. Cependant, une petite analogie avec l'action d'une particule ponctuelle, pour laquelle nous avons aussi un signe "-" (cf. chapitre de Relativité Restreinte), peut facilement déjà se faire.

Définissons pour la suite:

$$\gamma_{\alpha\beta} = \frac{\partial X}{\partial \tau} \circ \frac{\partial X}{\partial \sigma} = \frac{\partial X^\mu}{\partial \tau} \frac{\partial X_\mu}{\partial \sigma} \quad (52.53)$$

ce que nous pouvons aussi écrire sous forme matricielle:

$$\gamma_{\alpha\beta} = \begin{bmatrix} (\dot{X})^2 & \dot{X} \circ X' \\ \dot{X} \circ X' & (X')^2 \end{bmatrix} \quad (52.54)$$

Et en utilisant le déterminant de cette matrice il vient:

$$\gamma = \det(\gamma_{\alpha\beta}) = (\dot{X})^2 (X')^2 - (\dot{X} \circ X')^2 \quad (52.55)$$

Donc nous pouvons alors réécrire l'action d'une corde relativiste sous la forme finale condensée suivante:

$$S = -\frac{T_0}{c} \int d\tau d\sigma \sqrt{-\gamma} \quad (52.56)$$

qui n'est autre que "**l'action de Nambu-Goto condensée**" d'une corde relativiste.

Nous allons maintenant obtenir l'équation du mouvement en faisant varier l'action. Nous allons pour cela nous inspirer exactement des méthodes vues lors de la détermination en début de ce chapitre de l'équation d'onde non-relativiste d'une corde.

Ainsi, nous réécrivons l'action de Nambu-Goto en définissant une densité lagrangienne $\mathcal{L}$ telle que:

$$S = -\frac{T_0}{c} \int d\tau d\sigma \sqrt{-\gamma} = \int_{t_i}^{t_f} d\tau \int_0^{\sigma_f} d\sigma \mathcal{L} \left(\frac{\partial X^\mu}{\partial t}, \frac{\partial X^\mu}{\partial \sigma} \right) = \int_{t_i}^{t_f} d\tau \int_0^{\sigma_f} d\sigma \mathcal{L}(\dot{X}^\mu, X^{\mu'}) \quad (52.57)$$

où $\mathcal{L}$ est donc définie par:

$$\mathcal{L}(\dot{X}^\mu, X^{\mu'}) = -\frac{T_0}{c} \sqrt{(\dot{X}^\mu X^{\mu'})^2 - (\dot{X}^\mu)^2 (X^{\mu'})^2} \quad (52.58)$$

Nous allons maintenant appliquer le principe variationnel sur l'action afin d'en tirer l'équation de mouvement d'une corde. Le développement et l'approximation sont parfaitement similaires à ceux vus dans le chapitre de Mécanique Ondulatoire pour la corde non relativiste. Rappelons que nous avons obtenu comme densité lagrangienne et comme expression de l'action:

$$\mathcal{L} = \frac{1}{2} \rho_t \left(\frac{\partial y}{\partial t} \right)^2 - \frac{1}{2} F \left(\frac{\partial y}{\partial x} \right)^2 \quad \text{et} \quad S = \int_{t_i}^{t_f} dt \int_0^L dx \mathcal{L} \left(\frac{\partial y}{\partial t}, \frac{\partial y}{\partial x} \right) = \int_{t_i}^{t_f} dt \int_0^L dx \mathcal{L}(\dot{y}, y') \quad (52.59)$$

et que l'application du principe variationnel nous avait donné:

$$\delta S = \int_{t_i}^{t_f} dt \int_0^L dx \left[\rho_t \frac{\partial y}{\partial t} \frac{\partial (\delta y)}{\partial t} - F \frac{\partial y}{\partial x} \frac{\partial (\delta y)}{\partial x} \right] \quad (52.60)$$

Or, ce que nous n'avions pas vu dans le chapitre de Mécanique Ondulatoire, c'est que cette dernière relation pouvait facilement s'écrire aussi à partir de la densité lagrangienne:

$$\delta S = \int_{t_i}^{t_f} dt \int_0^L dx \left[\frac{\partial \mathcal{L}}{\partial \left(\frac{\partial y}{\partial t} \right)} \frac{\partial (\delta y)}{\partial t} + \frac{\partial \mathcal{L}}{\partial \left(\frac{\partial y}{\partial x} \right)} \frac{\partial (\delta y)}{\partial x} \right] = \int_{t_i}^{t_f} dt \int_0^L dx \left[\frac{\partial \mathcal{L}}{\partial \dot{y}} \frac{\partial (\delta y)}{\partial t} + \frac{\partial \mathcal{L}}{\partial y'} \frac{\partial (\delta y)}{\partial x} \right] \quad (52.61)$$

Dès lors, pour la corde relativiste, nous avons une forme identique en appliquant des développements en tous points similaires (et ce même si la densité lagrangienne a une forme différente):

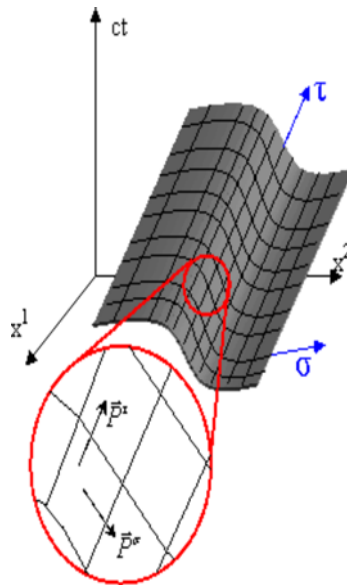
$$\delta S = \int_{t_i}^{t_f} d\tau \int_0^{\sigma_1} d\sigma \left[\frac{\partial \mathcal{L}}{\partial \dot{X}^\mu} \frac{\partial (\delta X^\mu)}{\partial \tau} + \frac{\partial \mathcal{L}}{\partial X^{\mu'}} \frac{\partial (\delta X^\mu)}{\partial \sigma} \right] \quad (52.62)$$

et comme nous l'avons fait au début de ce chapitre pour les cordes non relativistes, nous allons introduire les moments canoniques (densités d'impulsion/quantité de mouvement si vous préférez) de la corde en optant pour la notation:

$$\mathcal{P}_\mu^\tau \equiv \frac{\partial \mathcal{L}}{\partial \dot{X}^\mu} \quad \mathcal{P}_\mu^\sigma \equiv \frac{\partial \mathcal{L}}{\partial X^{\mu,1}}$$

où dans les détails, nous obtenons très facilement (c'est une simple dérivée mais si vous le souhaitez en nous contactant, nous pouvons vous le détailler) les moments longitudinaux et transverses:

$$\begin{aligned} \mathcal{P}_\mu^\tau &\equiv \frac{\partial \mathcal{L}}{\partial \dot{X}^\mu} = -\frac{T_0}{c} \frac{(\dot{X} \circ X') X'_\mu - (X')^2 \dot{X}_\mu}{\sqrt{(\dot{X} \circ X')^2 - (\dot{X})^2 (X')^2}} \\ \mathcal{P}_\mu^\sigma &\equiv \frac{\partial \mathcal{L}}{\partial X^{\mu,1}} = -\frac{T_0}{c} \frac{(\dot{X} \circ X') \dot{X}_\mu - (\dot{X})^2 X'_\mu}{\sqrt{(\dot{X} \circ X')^2 - (\dot{X})^2 (X')^2}} \end{aligned} \quad (52.63)$$



en faisant usage de cette notation, nous pouvons alors écrire:

$$\delta S = \int_{\tau_i}^{\tau_f} d\tau \int_0^{\sigma_1} d\sigma \left[\mathcal{P}_\mu^\tau \frac{\partial (\delta X^\mu)}{\partial \tau} + \mathcal{P}_\mu^\sigma \frac{\partial (\delta X^\mu)}{\partial \sigma} \right] \quad (52.64)$$

Faisant usage des mêmes méthodes que celles vues dans le chapitre de Mécanique Ondulatoire, notre variationnel s'exprime après simplification à nouveau sous la forme de trois termes:

$$\delta S = \int_{\tau_i}^{\tau_f} d\tau \int_0^{\sigma_1} d\sigma \left[\frac{\partial}{\partial \tau} (\delta X^\mu \mathcal{P}_\mu^\tau) + \frac{\partial}{\partial \sigma} (\delta X^\mu \mathcal{P}_\mu^\sigma) - \delta X^\mu \left(\frac{\partial (\mathcal{P}_\mu^\tau)}{\partial \tau} + \frac{\partial (\mathcal{P}_\mu^\sigma)}{\partial \sigma} \right) \right] \quad (52.65)$$

Les conditions pour trouver l'extremum (selon le principe de moindre action) restent les mêmes qu'en mécanique ondulatoire. Ainsi, pour le troisième terme, nous avons bien l'équation d'onde d'une corde excitée de manière transversale donnée avec la forme canonique par:

$$\frac{\partial (\mathcal{P}_\mu^\tau)}{\partial \tau} + \frac{\partial (\mathcal{P}_\mu^\sigma)}{\partial \sigma} = 0 \quad (52.66)$$

Il s'agit de l'équation du mouvement (ou onde) d'une corde ouverte ou même fermée (car finalement dans les développements précédents à aucun moment nous n'avions contraints les termes à être ouverts ou fermés).

Cette équation est horriblement difficile à résoudre mais le choix d'une paramétrisation adéquate peut néanmoins simplifier la tâche.

3. LAGRANGIEN D'UNE CORDE

Rappelons que nous avons:

$$\begin{aligned}\frac{\partial X^\mu}{\partial \sigma} &= \left(\frac{\partial X^0}{\partial \sigma}, \frac{\partial \vec{X}}{\partial \sigma} \right) = \left(0, \frac{\partial \vec{X}}{\partial \sigma} \right) \\ \frac{\partial X^\mu}{\partial \tau} &= \left(\frac{\partial X^0}{\partial \tau}, \frac{\partial \vec{X}}{\partial \tau} \right) = \left(c, \frac{\partial \vec{X}}{\partial \tau} \right)\end{aligned}\quad (52.67)$$

et qu'avec ce choix, nous avons donc:

$$\begin{aligned}S &= -\frac{T_0}{c} \int dt d\sigma \sqrt{(\dot{X}^\mu \circ X^\nu)^2 - (\dot{X}^\mu)^2 (X^\nu)^2} \\ &= -\frac{T_0}{c} \int dt d\sigma \sqrt{\left(\begin{pmatrix} c \\ \frac{\partial \vec{X}}{\partial \tau} \end{pmatrix} \circ \begin{pmatrix} 0 \\ -\frac{\partial \vec{X}}{\partial \sigma} \end{pmatrix} \right)^2 - \begin{pmatrix} c \\ \frac{\partial \vec{X}}{\partial \tau} \end{pmatrix} \circ \begin{pmatrix} c \\ -\frac{\partial \vec{X}}{\partial \tau} \end{pmatrix} \begin{pmatrix} 0 \\ \frac{\partial \vec{X}}{\partial \sigma} \end{pmatrix} \circ \begin{pmatrix} 0 \\ -\frac{\partial \vec{X}}{\partial \sigma} \end{pmatrix}} \\ &= -\frac{T_0}{c} \int dt d\sigma \sqrt{\left(\frac{\partial \vec{X}}{\partial \tau} \circ \begin{pmatrix} -\frac{\partial \vec{X}}{\partial \sigma} \end{pmatrix} \right)^2 - \left(c^2 - \frac{\partial \vec{X}}{\partial \tau} \circ \frac{\partial \vec{X}}{\partial \tau} \right) \left(-\frac{\partial \vec{X}}{\partial \sigma} \circ \frac{\partial \vec{X}}{\partial \sigma} \right)} \\ &= -\frac{T_0}{c} \int dt d\sigma \sqrt{(-1)^2 \left(\frac{\partial \vec{X}}{\partial \tau} \circ \frac{\partial \vec{X}}{\partial \sigma} \right)^2 - \left(c^2 - \frac{\partial \vec{X}}{\partial \tau} \circ \frac{\partial \vec{X}}{\partial \tau} \right) \left(-\frac{\partial \vec{X}}{\partial \sigma} \circ \frac{\partial \vec{X}}{\partial \sigma} \right)} \\ &= -\frac{T_0}{c} \int dt d\sigma \sqrt{\left(\frac{\partial \vec{X}}{\partial \tau} \circ \frac{\partial \vec{X}}{\partial \sigma} \right)^2 + \left(c^2 - \left(\frac{\partial \vec{X}}{\partial \tau} \right)^2 \right) \left(\frac{\partial \vec{X}}{\partial \sigma} \right)^2} \\ &= -\frac{T_0}{c} \int dt d\sigma \sqrt{\left(\frac{\partial \vec{X}}{\partial \tau} \circ \frac{\partial \vec{X}}{\partial \sigma} \right)^2 + \left[c^2 - \left(\frac{\partial \vec{X}}{\partial \tau} \right)^2 \right] \left(\frac{\partial \vec{X}}{\partial \sigma} \right)^2} \\ &= -\frac{T_0}{c} \int dt d\sigma \sqrt{\left(\frac{\partial \vec{X}}{\partial \tau} \circ \frac{\partial \vec{X}}{\partial \sigma} \right)^2 + \left[c^2 \left(\frac{\partial \vec{X}}{\partial \sigma} \right)^2 - \left(\frac{\partial \vec{X}}{\partial \tau} \right)^2 \left(\frac{\partial \vec{X}}{\partial \sigma} \right)^2 \right]} \\ &= -\frac{T_0}{c} \int dt d\sigma \sqrt{\left(\frac{\partial \vec{X}}{\partial \tau} \circ \frac{\partial \vec{X}}{\partial \sigma} \right)^2 + c^2 \left(\frac{\partial \vec{X}}{\partial \sigma} \right)^2 - \left(\frac{\partial \vec{X}}{\partial \tau} \right)^2 \left(\frac{\partial \vec{X}}{\partial \sigma} \right)^2}\end{aligned}\quad (52.68)$$

Maintenant, utilisons ce que nous avons vu dans le chapitre de Géométrie Différentielle avec le trièdre de Frenet:

$$\vec{T} = \frac{d\vec{X}}{ds} \quad (52.69)$$

où $\vec{T}$ est donc la tangente à la surface d'Univers à un instant t au voisinage d'un point donné. Nous avons par ailleurs montré dans ce même chapitre que par définition:

$$\|\vec{T}\| = \sqrt{\frac{d\vec{X}}{ds} \circ \frac{d\vec{X}}{ds}} = 1 \Rightarrow \frac{d\vec{X}}{ds} \circ \frac{d\vec{X}}{ds} = 1 \quad (52.70)$$

Or, nous pouvons écrire:

$$\frac{d\vec{X}}{ds} \circ \frac{d\vec{X}}{ds} = \left(\frac{d\vec{X}}{ds} \right)^2 = \frac{\partial \vec{X}}{\partial \sigma} \circ \frac{\partial \vec{X}}{\partial \sigma} = \frac{\partial \vec{X}}{\partial \sigma} \circ \frac{\partial \vec{X}}{\partial \sigma} \left(\frac{d\sigma}{ds} \right)^2 = \frac{\partial \vec{X}}{\partial \sigma} \frac{\partial \vec{X}}{\partial \sigma} \left(\frac{d\sigma}{ds} \right)^2 = 1 \quad (52.71)$$

où il ne faut pas oublier que $\partial \vec{X} / \partial \sigma$ est pris à un temps t fixé. Comme les lignes de la surface d'Univers de constante t décrivent la corde, alors $\partial \vec{X} / \partial \sigma$ est tangent à la corde.

Et comme:

$$\frac{d\vec{X}}{ds} = \frac{\partial \vec{X}}{\partial \sigma} \frac{d\sigma}{ds} \quad (52.72)$$

Alors $d\vec{X} / ds$ est colinéaire à $\partial \vec{X} / \partial \sigma$ et donc aussi tangent à la corde (information que nous n'avions pas quelques lignes plus haut!). Ces petites constatations étant faites, revenons à:

$$\begin{aligned} S &= -\frac{T_0}{c} \int dt d\sigma \sqrt{\left(\frac{\partial \vec{X}}{\partial t} \circ \frac{\partial \vec{X}}{\partial \sigma} \right)^2 + c^2 \left(\frac{\partial \vec{X}}{\partial \sigma} \right)^2 - \left(\frac{\partial \vec{X}}{\partial t} \right)^2 \left(\frac{\partial \vec{X}}{\partial \sigma} \right)^2} \\ &= -\frac{T_0}{c} \int dt d\sigma \sqrt{\left(\frac{ds}{d\sigma} \right)^2 \left[\left(\frac{\partial \vec{X}}{\partial t} \circ \frac{\partial \vec{X}}{\partial s} \right)^2 + c^2 \left(\frac{\partial \vec{X}}{\partial s} \right)^2 - \left(\frac{\partial \vec{X}}{\partial t} \right)^2 \left(\frac{\partial \vec{X}}{\partial s} \right)^2 \right]} \\ &= -\frac{T_0}{c} \int dt d\sigma \sqrt{\left(\frac{ds}{d\sigma} \right)^2 \left[\left(\frac{\partial \vec{X}}{\partial t} \circ \frac{\partial \vec{X}}{\partial s} \right)^2 + c^2 - \left(\frac{\partial \vec{X}}{\partial t} \right)^2 \right]} \\ &= -\frac{T_0}{c} \int dt ds \sqrt{c^2 - \left[\left(\frac{\partial \vec{X}}{\partial t} \right)^2 - \left(\frac{\partial \vec{X}}{\partial t} \circ \frac{\partial \vec{X}}{\partial s} \right)^2 \right]} \end{aligned} \quad (52.73)$$

cela devient déjà un peu plus intéressant!

Considérons maintenant le schéma suivant:

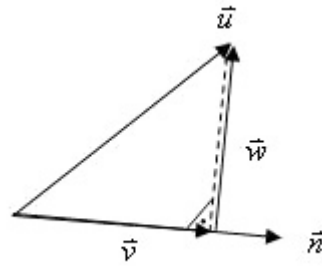


Figure: 52.7 - Rappel illustré du produit scalaire

où $\vec{u}$ est un vecteur quelconque et $\vec{n}$ un vecteur unitaire (sans dimensions) et $\vec{v}$, la projection orthogonale de $\vec{u}$ sur $\vec{n}$. Nous avons alors (cf. chapitre de Calcul Vectoriel):

$$\vec{u} \circ \vec{n} = \|\vec{u}\| \|\vec{n}\| \cos \theta = (\|\vec{u}\| \cos \theta) \|\vec{n}\| = \|\vec{v}\| \cdot \|\vec{n}\| \quad (52.74)$$

Or, si nous recherchons le vecteur $\vec{v}$ il faudra multiplier le tout par $\vec{n}$:

$$\vec{v} = (\vec{u} \circ \vec{n}) \vec{n} = \|\vec{v}\| \cdot \|\vec{n}\| \cdot \vec{n} \quad (52.75)$$

enfin, si nous recherchons l'expression du vecteur $\vec{w}$ il vient immédiatement:

$$\vec{v} + \vec{w} = \vec{u} \Rightarrow \vec{w} = \vec{u} - \vec{v} = \vec{u} - (\vec{u} \circ \vec{n}) \vec{n} \quad (52.76)$$

Dès lors, par analogie, nous pouvons écrire:

$$\vec{w} = \vec{u} - (\vec{u} \circ \vec{n}) \vec{n} = \frac{\partial \vec{X}}{\partial t} - \left(\frac{\partial \vec{X}}{\partial t} \circ \frac{d\vec{X}}{ds} \right) \frac{d\vec{X}}{ds} \quad (52.77)$$

où $\vec{w}$ est donc perpendiculaire à $d\vec{X}/ds$ et a comme unité celle d'une vitesse. Par construction, $\vec{w}$ est donc la vitesse transversale à la corde à un instant t donné puisque $d\vec{X}/ds$ est tangent à celle-ci. Nous noterons alors:

$$\vec{v}_\perp = \frac{\partial \vec{X}}{\partial t} - \left(\frac{\partial \vec{X}}{\partial t} \circ \frac{d\vec{X}}{ds} \right) \frac{d\vec{X}}{ds} \quad (52.78)$$

Mettons maintenant, pour des besoins ultérieurs, la norme au carré de cette dernière relation (attention on fait le traitement des composantes des vecteurs directement en généralisant à la notation vectorielle!):

$$\begin{aligned} v_\perp^2 &= \left(\frac{\partial \vec{X}}{\partial t} - \left(\frac{\partial \vec{X}}{\partial t} \circ \frac{d\vec{X}}{ds} \right) \frac{d\vec{X}}{ds} \right)^2 = \left(\frac{\partial \vec{X}}{\partial t} \right)^2 - 2 \left(\frac{\partial \vec{X}}{\partial t} \circ \frac{d\vec{X}}{ds} \right) \left(\frac{d\vec{X}}{ds} \circ \frac{\partial \vec{X}}{\partial t} \right) + \left(\frac{\partial \vec{X}}{\partial t} \circ \frac{d\vec{X}}{ds} \right)^2 \frac{d\vec{X}}{ds} \circ \frac{d\vec{X}}{ds} \\ &= \left(\frac{\partial \vec{X}}{\partial t} \right)^2 - 2 \left(\frac{\partial \vec{X}}{\partial t} \circ \frac{d\vec{X}}{ds} \right) \left(\frac{d\vec{X}}{ds} \circ \frac{\partial \vec{X}}{\partial t} \right) + \left(\frac{\partial \vec{X}}{\partial t} \circ \frac{d\vec{X}}{ds} \right)^2 = \left(\frac{\partial \vec{X}}{\partial t} \right)^2 - 2 \left(\frac{\partial \vec{X}}{\partial t} \circ \frac{d\vec{X}}{ds} \right)^2 + \left(\frac{\partial \vec{X}}{\partial t} \circ \frac{d\vec{X}}{ds} \right)^2 \\ &= \left(\frac{\partial \vec{X}}{\partial t} \right)^2 - \left(\frac{\partial \vec{X}}{\partial t} \circ \frac{d\vec{X}}{ds} \right)^2 \end{aligned} \quad (52.79)$$

et si nous revenons maintenant à:

$$\begin{aligned}
 S &= -\frac{T_0}{c} \int dt ds \sqrt{c^2 - \left[\left(\frac{\partial \vec{X}}{\partial t} \right)^2 - \left(\frac{\partial \vec{X}}{\partial t} \circ \frac{\partial \vec{X}}{\partial s} \right)^2 \right]} = -\frac{T_0}{c} \int dt ds \sqrt{c^2 - v_1^2} \\
 &= -T_0 \int dt ds \sqrt{1 - \frac{v_1^2}{c^2}}
 \end{aligned} \tag{52.80}$$

Le lagrangien associé est alors directement (ne pas confondre avec la densité lagrangienne!):

$$L = -T_0 \int \sqrt{1 - \frac{v_1^2}{c^2}} ds \tag{52.81}$$

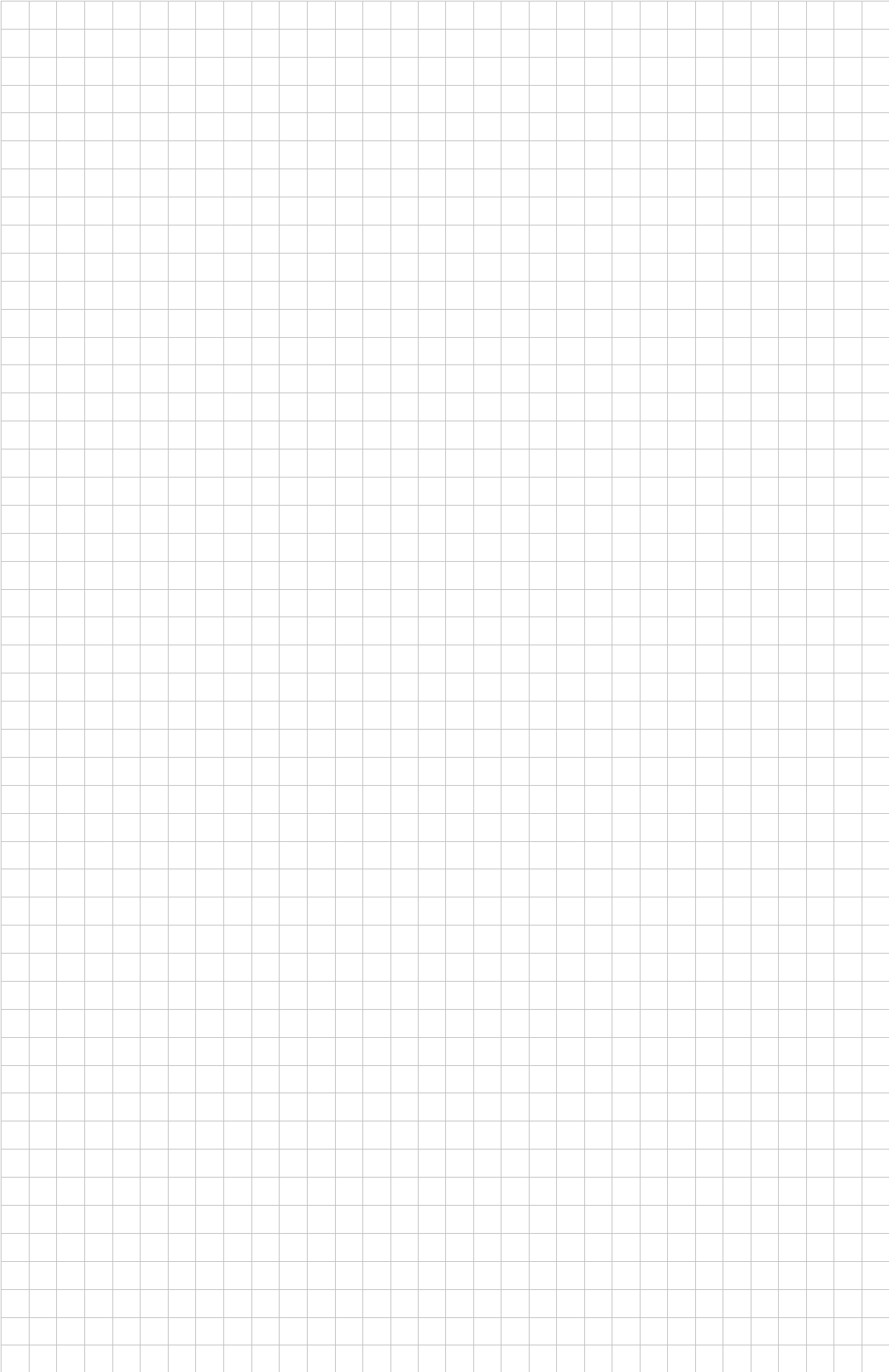
puisque:

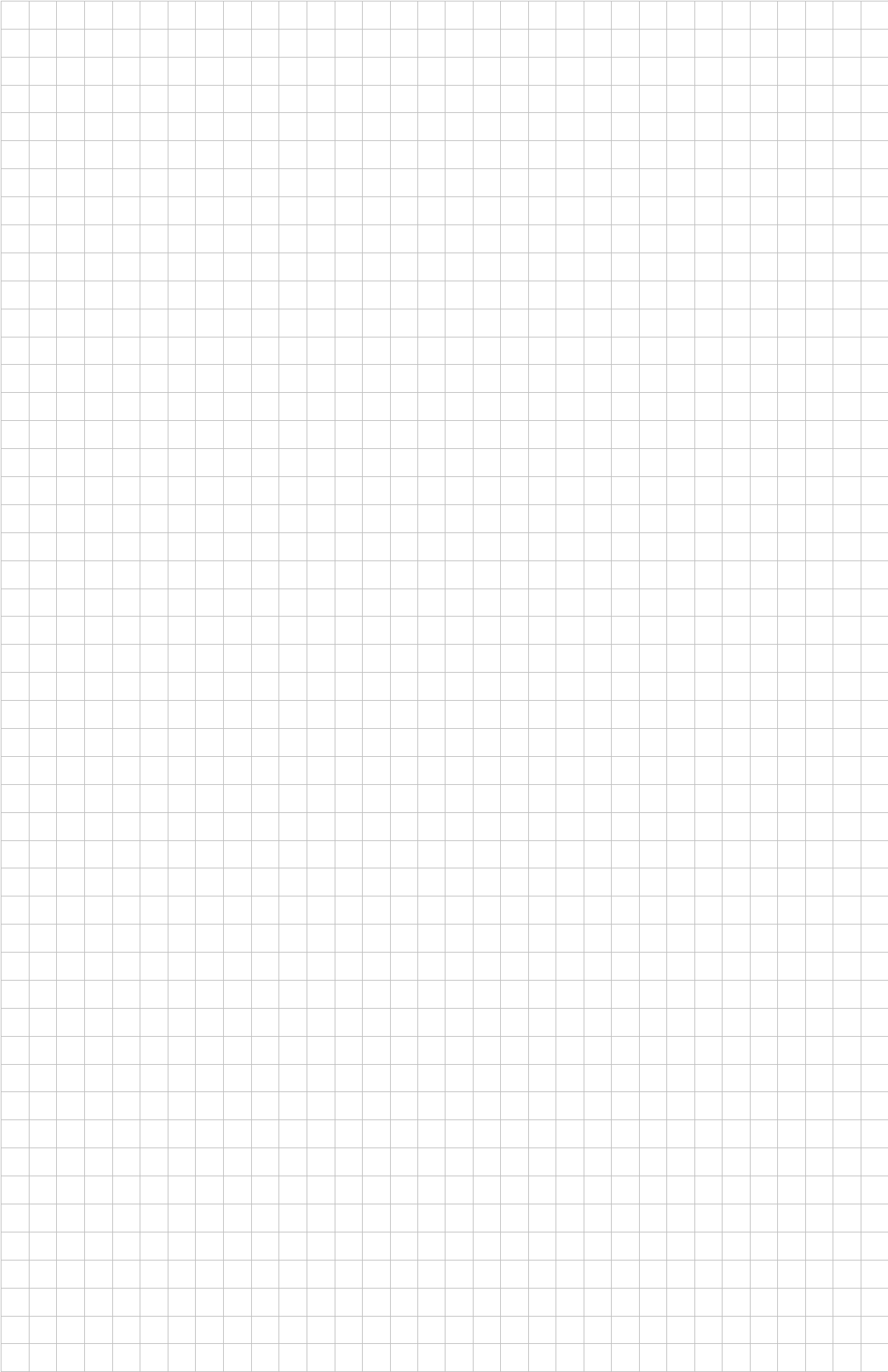
$$S = \int_{t_i}^{t_f} L dt = \int_{t_i}^{t_f} dt \int_0^{\sigma_1} d\sigma \mathcal{L} \tag{52.82}$$

Le lagrangien de la relation antéprécédente est considéré par les spécialistes de la théorie des cordes comme la généralisation naturelle du lagrangien de la particule libre obtenu dans le chapitre de Relativité Restreinte:

$$L = -mc^2 \sqrt{1 - \frac{v^2}{c^2}} \tag{52.83}$$









Chimie

53. CHIMIE QUANTIQUE

Avant que le lecteur aille plus loin dans la lecture de cette section du site, nous souhaitons rappeler que le site traite de mathématique appliquée et de physique théorique. Ainsi, nous traiterons dans cette section uniquement de chimie théorique (chimie quantique théorique, thermochimie théorique, chimie cinétique théorique, ...).

Ce choix fait suite à l'évolution du visage de la chimie ces vingt dernières décennies: de science en grande partie descriptive, elle tend à devenir déductive. C'est-à-dire qu'à côté de l'expérience, la place du calcul grandit constamment et ce particulièrement depuis le développement de l'informatique moderne qui aide grandement les chimistes à la modélisation numérique.

La chimie théorique, appelée également "**chimie physique**" - application des méthodes de la physique à la chimie - est trop souvent encore considérée comme une discipline en soi. Au fait, sous ce terme toute la chimie moderne est incluse. Ainsi, l'abord d'un problème quelconque de pointe en chimie exige son concours (ainsi que souvent celui de la chance...) et le chimiste doit en posséder une connaissance approfondie. Au niveau de l'enseignement de la chimie comme branche secondaire, ce rôle de la chimie-physique se manifeste déjà: il en résulte un relèvement du niveau, une augmentation du degré d'abstraction, et un risque de rebuter l'étudiant moyen. Enfin, il ne s'agit pas d'alourdir les connaissances en y incorporant plus d'éléments nouveaux, mais bien de convertir le mode d'approche de cette discipline en substituant le plus souvent à l'énoncé de connaissances encyclopédiques un exposé raisonné des questions fondamentales.

Une bonne compréhension de la chimie nécessite selon nous obligatoirement un crochet obligatoire par la physique quantique (cf. [section d'Atomistique](#)) afin d'avoir au moins une approche de ce qu'est un atome et de ses différentes orbites électroniques avant de parler de liaisons, des différentes méthodes de remplissage des orbites électroniques, d'oxydo-réduction, de remplissage des couches, et autres...

Dans ce sens, nous allons commencer par l'étude du cas particulier de l'atome d'hydrogène qui revêt une importance capitale pour toute la suite (étude des atomes polyélectroniques). Il convient donc au lecteur de parcourir les prochaines lignes avec toute l'attention possible et d'en saisir au mieux les subtilités!

1. Puits de potentiel rectangulaire tridimensionnel infini

Nous avons étudié dans le chapitre de Physique Quantique Corpusculaire l'atome d'hydrogène de Bohr-Sommerfeld en utilisant les résultats démontrés dans le chapitre de Relativité Restreinte. De ce modèle, en est ressorti une quantification simpliste (mais pas trop fausse comme nous le verrons plus loin) de certaines propriétés de la matière.

Dans le chapitre de Physique Quantique Ondulatoire, nous avons étudié le puits de potentiel rectiligne infini et l'oscillateur harmonique sans donner beaucoup plus d'exemples. Maintenant nous allons nous orienter vers la résolution de problèmes plus proches de ceux utiles en chimie avec l'objectif d'étudier l'atome hydrogénoïde.

Nous allons considérer maintenant une particule se mouvant librement dans la boîte tridimensionnelle ci-dessous.

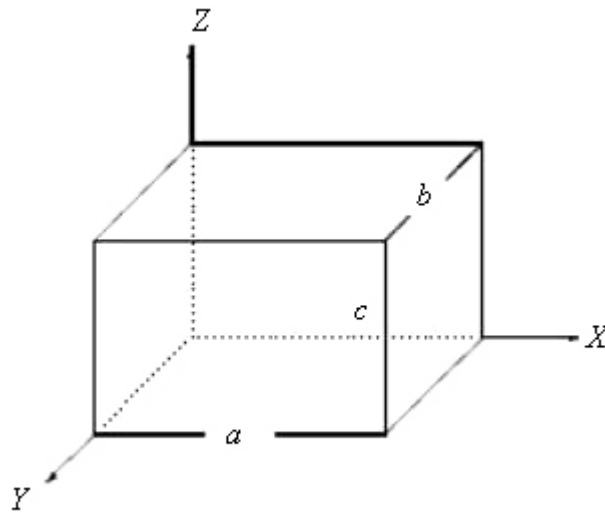


Figure: 53.1 - Boîte tridimensionnelle imaginaire dans laquelle la particule se meut

L'énergie potentielle de ce système est donnée par:

$$\begin{aligned} E_{pot}(x,y,z) &= 0 & 0 < x < a, 0 < y < b, 0 < z < c \\ E_{pot}(x,y,z) &= \infty & \text{ailleurs} \end{aligned} \quad (53.1)$$

Comme dans le cas unidimensionnel (voir section de physique atomique), les murs de potentiel infini empêchent la particule de quitter la boîte, et la fonction d'onde n'est non nulle que pour $\vec{r}$ se trouvant à l'intérieur de la boîte. Elle s'annule nécessairement dès que l'un des murs est atteint. L'équation de Schrödinger que nous devons résoudre est donc (cf. chapitre de Physique Quantique Ondulatoire):

$$\Delta \Psi + \frac{2m}{\hbar^2} (E_{tot} - E_{pot}) \Psi = 0 \quad (53.2)$$

et les conditions aux bornes se lisent:

$$\begin{aligned} \Psi(x=0, y, z) &= \Psi(x=a, y, z) = 0 \\ \Psi(x, y=0, z) &= \Psi(x, y=b, z) = 0 \\ \Psi(x, y, z=0) &= \Psi(x, y, z=c) = 0 \end{aligned} \quad (53.3)$$

Notons que l'hamiltonien peut s'écrire comme la somme de l'hamiltonien selon chaque axe (nous parlons des opérateurs bien évidemment!). Nous avons donc:

$$H = H_x + H_y + H_z \quad (53.4)$$

où:

$$H_x = \left(\frac{-\hbar^2}{2m} \right) \frac{\partial^2}{\partial x^2}, H_y = \left(\frac{-\hbar^2}{2m} \right) \frac{\partial^2}{\partial y^2}, H_z = \left(\frac{-\hbar^2}{2m} \right) \frac{\partial^2}{\partial z^2} \quad (53.5)$$

relations dont nous avons démontré la provenance dans la section de physique quantique ondulatoire de ce site.

Une telle forme est dite "**forme séparable**": l'hamiltonien est la somme d'opérateurs individuels H_i chacun ne dépendant que d'une seule variable ou degré de liberté q_i . Cette forme traduit le caractère indépendant des mouvements décrits par les variables q_i .

Rappelons-nous que la probabilité conjointe de deux événements indépendants est le produit des probabilités individuelles des deux événements, pris séparément (cf. chapitre de Probabilités). Nous nous attendons donc à ce que la densité de probabilité de présence dans l'espace (cf. chapitre de Physique Quantique Ondulatoire) de configuration multidimensionnel soit, dans le cas où l'hamiltonien est de forme séparable, un simple produit de densités de probabilités individuelles. En fait, la forme séparable de l'hamiltonien permet une séparation de variables sur la fonction d'onde elle-même.

Écrivons donc les solutions de l'équation de Schrödinger sous la forme:

$$\Psi(x, y, z) = \xi(x) \vartheta(y) \zeta(z) \quad (53.6)$$

(lire "zêta de x, thêta y, xi z") d'un produit de trois facteurs chacun ne dépendant que d'une coordonnée.

Substituant cette écriture dans l'équation de Schrödinger, on obtient sans développements (algèbre élémentaire):

$$\vartheta(y) \zeta(z) [H_x \xi(x)] + \xi(x) \zeta(z) [H_y \vartheta(y)] + \xi(x) \vartheta(y) [H_z \zeta(z)] = (E_{\text{tot}} - E_{\text{pot}}) \xi(x) \vartheta(y) \zeta(z) \quad (53.7)$$

ou encore, en divisant les deux membres de ceci par $\xi(x) \vartheta(y) \zeta(z)$:

$$\frac{1}{\xi(x)} [H_x \xi(x)] + \frac{1}{\vartheta(y)} [H_y \vartheta(y)] + \frac{1}{\zeta(z)} [H_z \zeta(z)] = (E_{\text{tot}} - E_{\text{pot}}) \quad (53.8)$$

ce qui est une forme beaucoup plus esthétique et facile à mémoriser.

Cette équation demande que la somme des trois termes dans le membre de gauche soit égale à une constante dans le cadre d'un système conservatif (c'est ce qui intéresse souvent les chimistes)! Chacun de ces trois termes ne dépendant que d'une et une seule variable, pour que leur somme soit égale à une constante, il faut que chaque terme soit lui-même constant. En effet, en prenant la dérivée des deux membres de la relation précédente par rapport à x , par exemple, nous avons:

$$\frac{d}{dx} \left\{ \frac{1}{\xi(x)} [H_x \xi(x)] \right\} = 0 \quad (53.9)$$

ce qui signifie que $\frac{1}{\xi(x)} [H_x \xi(x)]$ doit bien être une constante que nous noterons E_x (car ce terme exprime une énergie). Nous avons alors (tiens donc...):

$$H_x \xi(x) = E_x \xi(x) \quad (53.10)$$

De même, nous obtenons:

$$H_y \vartheta(y) = E_y \vartheta(y) \text{ et } H_z \zeta(z) = E_z \zeta(z) \quad (53.11)$$

Notons que chacune des équations séparées que nous venons d'obtenir, pour le mouvement de la particule dans les trois directions spatiales, est une équation de Schrödinger dans une boîte unidimensionnelle. Ainsi, les trois relations obtenues précédemment décrivent chacune indépendamment le mouvement dans la direction respective x, y, z , limité aux intervalles respectifs:

$$[0, a], [0, b], [0, c] \quad (53.12)$$

et doit être respectivement résolue avec conditions aux bornes:

$$\begin{aligned} \xi(x=0) &= 0 = \xi(x=a) \\ \mathfrak{g}(y=0) &= 0 = \mathfrak{g}(y=b) \\ \zeta(z=0) &= 0 = \zeta(z=c) \end{aligned} \quad (53.13)$$

Les résultats obtenus dans le chapitre de Physique Quantique Ondulatoire lors de la résolution de l'équation de Schrödinger dans le cas du puits rectiligne nous donnent directement:

$$\begin{aligned} \xi_{\lambda}(x) &= \sqrt{\frac{2}{a}} \sin\left(\frac{\lambda\pi x}{a}\right), E_{x,\lambda} = \frac{\lambda^2 \hbar^2}{8ma^2} \text{ avec } \lambda \in \mathbb{N}^* \\ \mathfrak{g}_{\mu}(y) &= \sqrt{\frac{2}{b}} \sin\left(\frac{\mu\pi y}{b}\right), E_{y,\mu} = \frac{\mu^2 \hbar^2}{8mb^2} \text{ avec } \mu \in \mathbb{N}^* \\ \zeta_{\nu}(z) &= \sqrt{\frac{2}{c}} \sin\left(\frac{\nu\pi z}{c}\right), E_{z,\nu} = \frac{\nu^2 \hbar^2}{8mc^2} \text{ avec } \nu \in \mathbb{N}^* \end{aligned} \quad (53.14)$$

En résumé les états stationnaires de la particule dans la boîte tridimensionnelle sont spécifiés par trois nombres quantiques entiers strictement positifs λ, μ, ν . La fonction d'onde est finalement:

$$\Psi_{\lambda\mu\nu} = \sqrt{\frac{8}{abc}} \sin\left(\frac{\lambda\pi x}{a}\right) \sin\left(\frac{\mu\pi y}{b}\right) \sin\left(\frac{\nu\pi z}{c}\right) \quad (53.15)$$

et ses énergies respectives (valeurs propres):

$$E_{\lambda\mu\nu} = \frac{\hbar^2}{8m} \left(\frac{\lambda^2}{a^2} + \frac{\mu^2}{b^2} + \frac{\nu^2}{c^2} \right) \quad (53.16)$$

La technique de séparation de variables détaillée ci-haut, n'est applicable que parce que l'hamiltonien est de forme séparable. Il vient automatiquement dès lors que la densité de probabilité tridimensionnelle $|\Psi_{\lambda\mu\nu}(x, y, z)|^2$ est le produit des densités de probabilités $|\xi_{\lambda}(x)|^2, |\mathfrak{g}_{\mu}(y)|^2, |\zeta_{\nu}(z)|^2$, comme nous l'avions anticipé. Nous notons aussi que l'énergie de mouvement dans l'espace tridimensionnel est la somme des énergies de mouvements dans les trois directions spatiales: l'indépendance de ces trois directions ou degrés de liberté, implique donc l'additivité de leur énergie.

2. VIBRATIONS MOLÉCULAIRES

Nous avons étudié dans le chapitre de Physique Quantique Ondulatoire l'oscillateur harmonique. C'est maintenant en chimie que l'on va utiliser toute la puissance des résultats obtenus lors de l'étude de ce système.

L'oscillateur harmonique est un modèle des vibrations moléculaires, et est représenté par un potentiel parabolique de type:

$$E_p = \frac{1}{2} c^x x^2 \quad (53.17)$$

pour une molécule diatomique. Mais nous avons vu dans le chapitre de Physique Nucléaire que $c^x = m\omega_0^2$ ce qui fait que nous avons finalement pour une molécule diatomique:

$$E_p = \frac{m\omega_0^2}{2} x^2 \quad (53.18)$$

Pour une molécule polyatomique, nous aurons in extenso (par l'additivité de l'énergie):

$$E_{pot}(q_i) = \frac{1}{2} \sum_i \omega_i^2 q_i^2 \quad (53.19)$$

Les quantités ω_0 et ω_i sont les fréquences (ou plutôt, plus correctement: les pulsations) vibrationnelles d'une molécule, diatomique dans le premier cas, et polyatomique dans le second cas. Dans la première équation, la variable x représente l'élongation de la liaison entre les deux atomes A et B (comme avec un ressort) dans une molécule diatomique, c'est-à-dire $x = (R - R_{eq})$, où R est la longueur instantanée de cette liaison, et R_{eq} est sa valeur d'équilibre.

Dans le cas d'une molécule polyatomique, le potentiel décrivant les vibrations moléculaires ne prend la forme séparable en termes de sommation ci-dessus que si l'on envisage des variables spéciales q_i dénotant des mouvements collectifs des noyaux, et qui sont appelées "**modes normaux de vibrations**".

Nous avons aussi vu dans le chapitre de Physique Quantique Ondulatoire que l'hamiltonien d'une molécule diatomique (problème de l'oscillateur harmonique) peut s'écrire sous la forme:

$$H = -\frac{\hbar^2}{2m} \frac{d^2}{dx^2} + \frac{m\omega^2}{2} x^2 \quad (53.20)$$

Pour une molécule polyatomique cette relation devient logiquement:

$$H = \sum_i \left(-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial q_i^2} + \frac{\omega_i^2 q_i^2}{2} \right) \quad (53.21)$$

L'hamiltonien ci-dessus est clairement de forme séparable: c'est une somme d'hamiltoniens unidimensionnels, chacun ne dépendant que d'un seul mode q_i comme variable, et décrivant ce mode comme étant un ressort unique, ou oscillateur harmonique de masse unitaire ($m = 1$) et de pulsation d'oscillation ω_i . Par conséquent, une séparation des variables q_i est possible, réduisant l'équation de Schrödinger indépendante du temps en un certain nombre d'équations du même type que celle d'un oscillateur harmonique unidimensionnel. Il suffit donc de connaître l'expression de la fonction d'onde pour un oscillateur harmonique unidimensionnel, ce que nous avons déjà fait dans le chapitre de Physique Quantique Ondulatoire où nous avons obtenu:

$$E_n = \hbar \omega \left(n + \frac{1}{2} \right) \quad (53.22)$$

avec $n \in \mathbb{N}$ et:

$$\Psi_n = (A\sqrt{\pi n!} 2^n)^{-1/2} e^{-Q^2/2} H_n(Q) \quad (53.23)$$

La figure ci-dessous montre le graphique des premières fonctions d'onde de la relation précédente ainsi que celui de leurs densités de probabilité de présence respectives. On note les mêmes structures modales que celles des fonctions propres d'une particule dans une boîte unidimensionnelle.

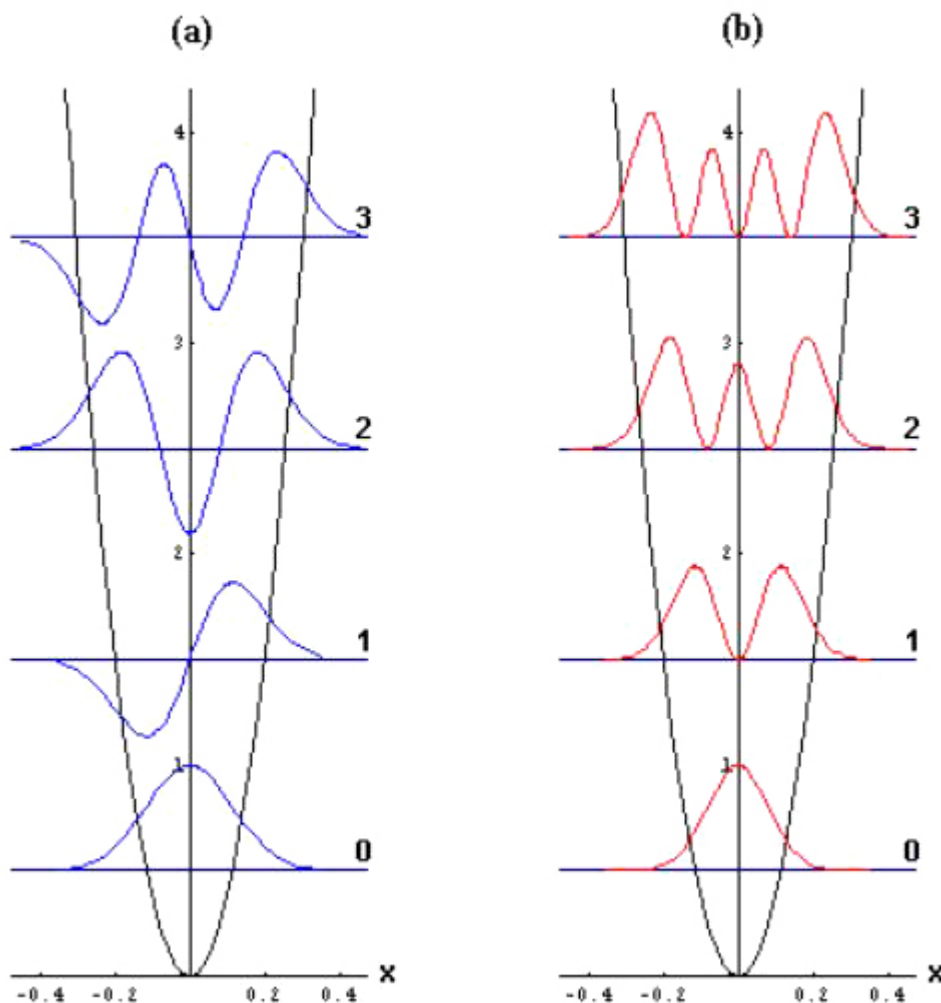


Figure: 53.2 - Fonctions d'onde et de densité de probabilité de l'oscillateur harmonique unidimensionnel

Ci-dessus les premiers niveaux d'énergie de l'oscillateur unidimensionnel avec (a) leur fonction propre associée, (b) la distribution de probabilité de présence associée.

Dans la limite des très grandes valeurs de n , la distribution de probabilité se rapproche de plus en plus de celle prédite par la mécanique classique, l'oscillateur réside pour la majeure partie du temps au voisinage des points de rebroussement définis par l'intersection du potentiel E_{pot} avec le niveau n . Cette tendance est illustrée ci-dessous:

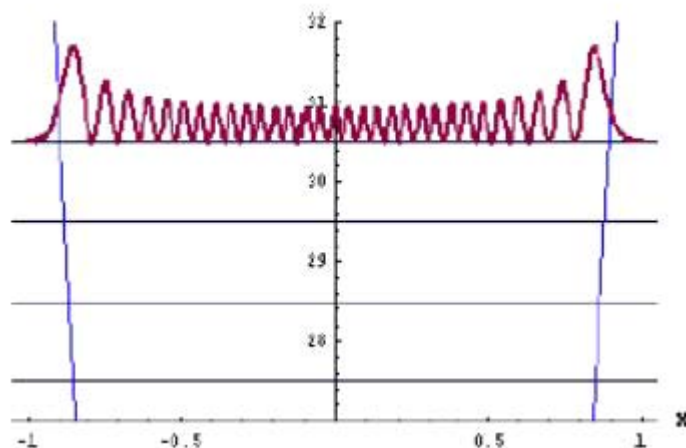


Figure: 53.3 - Fonctions de densité de probabilité de l'oscillateur harmonique unidimensionnel pour de grands n

Pour une molécule polyatomique l'expression de la quantification d'énergie devient:

$$E_n = \sum_i \hbar \omega_i \left(n_i + \frac{1}{2} \right) \quad (53.24)$$

avec $n \in \mathbb{N}, \forall i = 1, 2, 3, \dots, n$.

et les fonctions/états propres deviennent:

$$\Psi_n(q_i) = (A\sqrt{\pi n!} 2^n)^{-1/2} \prod_i H_n(Q_i) e^{-Q_i^2/2} \quad (53.25)$$

avec:

$$Q_i = \sqrt{\frac{m_i}{\hbar}} q_i \quad (53.26)$$

Ces deux dernières relations sont très importantes parce qu'elles permettent parmi tant d'autres de:

- Prédire le spectre de la molécule (spectroscopie)
- D'étudier les bandes d'énergie (d'où viennent les notions de bande de conduction et de valence)
- De localiser les liaisons entre atomes et donc les propriétés chimiques

3. ATOME HYDROGÉNOÏDE

Nous considérons ici la quantification d'un système générique comportant deux corps (particules) en interaction mutuelle et se mouvant dans l'espace tridimensionnel. On démontrera dans un premier temps que, si la séparation des variables dynamiques décrivant individuellement chacun des deux corps est impossible, par contre, le mouvement d'ensemble du système (celui du centre de masse) et le mouvement interne, dit encore mouvement relatif, sont séparables. En outre, si le potentiel est centrosymétrique, le mouvement interne peut encore se décomposer en un mouvement de rotation et un mouvement radial. La quantification du mouvement rotationnel est intimement liée à celle du moment cinétique.

Nous nous intéresserons ici à la mécanique d'un système atomique ne comportant qu'un seul électron. C'est un système à deux particules: un noyau, de masse M et de charge $+Ze$, et un électron de masse m_e et de charge $-e$.

Le système atomique est décrit par l'hamiltonien suivant:

$$\begin{aligned}
 H &= E_c^{\text{Noyau}} + E_c^{\text{électron}} + E_{\text{pot}}^{\text{système}} = \frac{p_N^2}{2M} + \frac{p_e^2}{2m_e} + E_{\text{pot}}(\vec{r}_e - \vec{R}_N) \\
 &= -\frac{\hbar^2}{2M} \nabla_N^2 - \frac{\hbar^2}{2m_e} \nabla_e^2 + E_{\text{pot}}(\vec{r}_e - \vec{R}_N)
 \end{aligned} \quad (53.27)$$

Rappelez-vous que dans le chapitre de Physique Quantique Ondulatoire, nous avons démontré lors de l'étude des opérateurs fonctionnels que:

$$\vec{p} = \frac{1}{i} \hbar \vec{\nabla} = \frac{1}{i} \frac{h}{2\pi} \vec{\nabla} \quad (53.28)$$

et rappelez-vous aussi pour la suite que $\vec{r}_e$ et $\vec{R}_N$ sont respectivement les vecteurs de position de l'électron et du noyau dans la relation antérieure.

L'énergie potentielle étant donnée par (cf. chapitre d'Électrostatique):

$$E_{\text{pot}}(\vec{r}_e - \vec{R}_N) = -\frac{Ze^2}{4\pi\epsilon_0 \|\vec{r}_e - \vec{R}_N\|} \quad (53.29)$$

Les mouvements des deux particules sont corrélés car les deux charges interagissent à travers leur champ électrique mutuel. On ne peut donc pas effectuer une séparation de variables entre $\vec{r}_e$ et $\vec{R}_N$. Par contre, une séparation de variables est possible entre la coordonnée du centre de masse (voir la définition du centre de masse dans le chapitre de Mécanique Classique):

$$\vec{R}_{\text{CM}} = \frac{m_e \vec{r}_e + M \vec{R}_N}{m_e + M} \quad (53.30)$$

et la coordonnée relative de l'électron par rapport au noyau:

$$\vec{r}_{\text{rel}} = \vec{r}_e - \vec{R}_N \quad (53.31)$$

Nous obtenons dès lors:

$$\vec{p}_{CM} = \frac{\hbar}{i} \vec{\nabla}_{CM} = \frac{\hbar}{i} (\vec{\nabla}_e + \vec{\nabla}_N) = \vec{p}_e + \vec{p}_N \quad (53.32)$$

et:

$$\vec{p}_{rel} = \frac{M\vec{p}_e - m_e\vec{p}_N}{m_e + M} \quad (53.33)$$

L'hamiltonien dans le référentiel du centre de masse s'écrit donc:

$$H = \frac{p_{CM}^2}{2M_{tot}} + \frac{p_{rel}^2}{2\mu} + E_{pot}(\vec{r}_{rel}) \quad (53.34)$$

où $M_{tot} = m_e + M$ est la masse totale du système, et:

$$\mu = \frac{m_e M}{M_{tot}} \quad (53.35)$$

est sa masse réduite.

Nous voyons clairement que l'hamiltonien H est cette fois mis sous une forme séparable et qu'on peut l'écrire sous la forme suivante:

$$H = H_{CM} + H_{rel} \quad (53.36)$$

avec:

$$H_{CM} = \frac{p_{CM}^2}{2M_{tot}} \text{ et } H_{rel} = \frac{p_{rel}^2}{2\mu} + E_{pot}(\vec{r}_{rel}) \quad (53.37)$$

En termes des coordonnées $\vec{R}_{CM}$ et $\vec{r}_{rel}$, la fonction décrivant un état stationnaire du système à deux corps est donc un produit de fonctions d'onde individuelles (rappelons que la probabilité conjointe de deux événements est le produit de leur probabilité), l'une pour le mouvement du centre de masse, l'autre pour le mouvement relatif:

$$\Psi_{tot}(\vec{R}_{CM}, \vec{r}_{rel}) = \Psi_{CM}(\vec{R}_{CM}) \Psi_{rel}(\vec{r}_{rel}) \quad (53.38)$$

et l'énergie de cet état est la somme des énergies de mouvement respectives:

$$E_{tot} = E_{CM} + E_{rel} \quad (53.39)$$

avec:

$$H_{CM} \Psi_{CM}(\vec{R}_{CM}) = E_{CM} \Psi_{CM}(\vec{R}_{CM}) \text{ et } H_{rel} \Psi_{rel}(\vec{r}_{rel}) = E_{rel} \Psi_{rel}(\vec{r}_{rel}) \quad (53.40)$$

Remarque: Cette approche qui consiste à séparer la fonction d'onde en la composition d'une fonction d'onde du centre de masse et du mouvement relatif est utilisée également dans le cadre de l'étude des atomes poly-électroniques mais à une différence près: comme le noyau est alors beaucoup plus massif que le cortège des électrons (en approximation...), le centre de masse est assimilé au noyau de l'atome et le mouvement relatif à l'ensemble du cortège électronique. Cette approche approximative est très connue sous la dénomination "**d'approximation de Born-Oppenheimer**".

L'hamiltonien apparaissant dans la première de ces deux relations a été défini plus haut comme valant:

$$H_{CM} = \frac{P_{CM}^2}{2M_{wt}} \quad (53.41)$$

Ce mouvement est celui d'une particule de masse M_{wt} dans une boîte tridimensionnelle de volume infini. Les fonctions propres et valeurs propres pour ce mouvement ayant déjà été obtenues dans notre étude précédente, on se limitera à l'étude de l'équation séparée pour le mouvement relatif, ou mouvement interne. Comme aucune confusion ne sera dès lors possible entre les différents hamiltoniens, nous laisserons tomber, pour simplifier les notations, la mention *rel* en indice inférieur.

Avec H_{rel} donné par la relation que nous avons démontrée précédemment:

$$H_{rel} = \frac{p_{rel}^2}{2\mu} + E_{pot}(\vec{r}_{rel}) \quad (53.42)$$

et la relation (aussi démontrée précédemment):

$$H_{rel} \Psi_{rel}(\vec{r}_{rel}) = E_{rel} \Psi_{rel}(\vec{r}_{rel}) \quad (53.43)$$

nous obtenons alors l'équation de Schrödinger pour le mouvement relatif:

$$\left[\frac{p^2}{2\mu} + E_{pot}(r) \right] \Psi(\vec{r}) = E \Psi(\vec{r}) \quad (53.44)$$

ou écrit autrement:

$$\left[-\frac{\hbar^2}{2\mu} \nabla^2 + E_{pot}(r) \right] \Psi(\vec{r}) = E \Psi(\vec{r}) \quad (53.45)$$

Remarquons que dans le cas où l'énergie potentielle E_{pot} est de source centrosymétrique, c'est-à-dire qu'elle ne dépend que de la longueur du vecteur position $\vec{r}$, et non de son orientation, l'équation précédente, telle qu'écrite, en coordonnées cartésiennes, n'est pas séparable: en effet, en coordonnées cartésiennes, la longueur de $\vec{r}$ est donnée par:

$$\|\vec{r}\| = \sqrt{x^2 + y^2 + z^2} \quad (53.46)$$

et l'énergie potentielle n'est pas séparable en trois composantes chacune ne dépendant que d'une seule des trois variables x, y, z . L'hamiltonien n'est donc toujours pas de forme séparable et nous n'avons donc pas atteint notre objectif. Cependant, l'équation précédente est séparable dès lors que nous effectuons un changement de coordonnées vers les coordonnées sphériques. En effet, dans ce système de coordonnées, le potentiel ne dépend que d'une des trois variables sphériques, le rayon r . Il est indépendant des deux angles θ et φ .

Si nous nous référons au résultat obtenu lors de l'étude des expressions du laplacien dans différents systèmes de coordonnées, dans le chapitre de Calcul Vectoriel, nous avons obtenu pour le laplacien d'un champ scalaire en coordonnées sphériques, l'expression suivante:

$$\nabla^2 u = \frac{\partial^2 u}{\partial r^2} + 2 \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} + \frac{\cot \theta}{r^2} \frac{\partial u}{\partial \theta} + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 u}{\partial \varphi^2} \quad (53.47)$$

L'hamiltonien:

$$\left[-\frac{\hbar^2}{2\mu} \nabla^2 + E_{\text{pot}}(\vec{r}) \right] \quad (53.48)$$

devient dès lors (simple distribution et nouvelle manière de noter):

$$H = T_r + \frac{L^2}{2\mu r^2} + E_{\text{pot}}(\vec{r}) \quad (53.49)$$

où:

$$T_r = -\frac{\hbar^2}{2\mu} \left(\frac{\partial^2}{\partial r^2} + 2 \frac{1}{r} \frac{\partial}{\partial r} \right) = -\frac{\hbar^2}{2\mu} \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) \quad (53.50)$$

est l'opérateur énergie cinétique pour le mouvement radial de l'électron par rapport au noyau, et L^2 est l'opérateur "associé" au carré du vecteur moment cinétique:

$$L^2 = -\hbar^2 \left[\frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2}{\partial \varphi^2} \right] \quad (53.51)$$

Le terme:

$$\frac{L^2}{2\mu r^2} \quad (53.52)$$

est donc une énergie associée au moment cinétique.

Pour bien saisir la nature de cet opérateur L^2 un crochet par la notion de rotateur rigide s'impose.

3.1. ROTATEUR RIGIDE

Si nous considérons maintenant le cas d'un système appelé "**rotateur rigide**" où nous négligeons ("restreignons" serait un terme plus adapté...) les degrés de liberté d'oscillation (c'est à ce système que l'on a affaire dans le cas des molécules diatomiques ou polyatomiques linéaires), les seules coordonnées mises en jeu étant les angles ϑ et φ qui fixent l'orientation du rotateur.

Ainsi, dans ce cas r est fixé et nous avons:

$$T_r = 0, \quad E_{pot}(r) = \begin{cases} 0 & r = R_e \\ \infty & r \neq R_e \end{cases} \quad (53.53)$$

et au vu des contraintes sur le potentiel, il est normalement assez facile de comprendre pourquoi le rotateur est dit "rigide". Dans le cas précité, l'hamiltonien se réduit alors à:

$$H = \frac{\vec{L}^2}{2J} \quad (53.54)$$

où:

$$J = \mu r^2 \quad (53.55)$$

est le moment d'inertie (cf. chapitre de Mécanique Classique) de la masse réduite du système.

Pour la suite, nous associons l'opérateur $\vec{L}^2$ à un moment cinétique, pour la simple raison qu'il en a les unités... Effectivement, rappelons que nous avons démontré dans le chapitre de Physique Quantique Ondulatoire que lorsque le spin est nul (donc dans le cadre de notre étude de l'atome hydrogénoïde ici présent, le spin ne sera pas pris en compte dans un premier temps) et que nous avons affaire à une seule particule alors le moment cinétique (que nous noterons donc L dans ce chapitre au lieu de b) est donné par:

$$L \leq \sqrt{l(l+1)}\hbar \Rightarrow L^2 \leq [l(l+1)]\hbar^2 \quad (53.56)$$

avec $l \in \mathbb{N}$ et:

$$\vec{L} = \hbar \vec{l} \quad (53.57)$$

où les composantes du vecteur $\vec{l}$ sont aussi des entiers naturels. En faisant cette similitude, nous pouvons alors écrire l'équation de Schrödinger sous la forme:

$$\left[T_r + \frac{l(l+1)\hbar^2}{2\mu r^2} + E_{pot} \right] \Psi = E \Psi \quad (53.58)$$

Rappelons aussi au passage que nous avons obtenu dans le chapitre de Physique Quantique Ondulatoire que:

$$L_x = \frac{\hbar}{i} (y \partial_z - z \partial_y) \quad (53.59)$$

par le produit vectoriel.

Passons maintenant des coordonnées rectangulaires x, y, z aux coordonnées sphériques r, θ, ϕ . Rappelons pour cela (cf. chapitre de Calcul Vectoriel) que:

$$\begin{aligned} r &= \sqrt{x^2 + y^2 + z^2}, 0 \leq r < \infty \\ x &= r \sin \theta \cos \phi & \theta &= \arccos \frac{z}{\sqrt{x^2 + y^2 + z^2}}, 0 \leq \theta \leq \pi \\ y &= r \sin \theta \sin \phi \text{ et} & \phi &= \arccos \frac{x}{\sqrt{x^2 + y^2}}, 0 \leq \phi < 2\pi \\ z &= r \cos \theta \end{aligned} \quad (53.60)$$

exprimons les différentielles totales:

$$\begin{aligned} dx &= \sin \theta \cos \phi dr + r \cos \theta \cos \phi d\theta - r \sin \theta \sin \phi d\phi \\ dy &= \sin \theta \sin \phi dr + r \cos \theta \sin \phi d\theta + r \sin \theta \cos \phi d\phi \\ dz &= \cos \theta dr - r \sin \theta d\theta \end{aligned} \quad (53.61)$$

Ces relations peuvent être écrites comme une transformation orthogonale des différentielles totales $dr, r d\theta, r \sin \theta d\phi$ par:

$$\begin{pmatrix} dx \\ dy \\ dz \end{pmatrix} = \begin{bmatrix} \sin \theta \cos \phi & \cos \theta \cos \phi & -\sin \phi \\ \sin \theta \sin \phi & \cos \theta \sin \phi & \cos \phi \\ \cos \theta & -\sin \theta & 0 \end{bmatrix} \begin{pmatrix} dr \\ r d\theta \\ r \sin \theta d\phi \end{pmatrix} \quad (53.62)$$

ou encore par la transformation inverse (au besoin... il suffit de vérifier que les deux matrices de transformation multipliées entre elles donnent la matrice unité):

$$\begin{pmatrix} dr \\ r d\theta \\ r \sin \theta d\phi \end{pmatrix} = \begin{bmatrix} \sin \theta \cos \phi & \sin \theta \sin \phi & \cos \theta \\ \cos \theta \cos \phi & \cos \theta \sin \phi & -\sin \theta \\ -\sin \phi & \cos \phi & 0 \end{bmatrix} \begin{pmatrix} dx \\ dy \\ dz \end{pmatrix} \quad (53.63)$$

Il en résulte par exemple:

$$\begin{aligned} dr &= \sin \theta \cos \phi dx + \sin \theta \sin \phi dy + \cos \theta dz \Rightarrow \frac{\partial r}{\partial x} = \sin \theta \cos \phi \\ r d\theta &= \cos \theta \cos \phi dx + \cos \theta \sin \phi dy - \sin \theta dz \Rightarrow \frac{\partial \theta}{\partial x} = \frac{\cos \theta \cos \phi}{r} \\ r \sin \theta d\phi &= -\sin \phi dx + \cos \phi dy + 0 dz \Rightarrow \frac{\partial \phi}{\partial x} = \frac{-\sin \phi}{r \sin \theta} \end{aligned} \quad (53.64)$$

et finalement (la méthode pour les deuxième et troisième lignes est la même que pour la première!):

$$\begin{aligned}
\frac{\partial}{\partial x} &= \frac{\partial r}{\partial x} \frac{\partial}{\partial r} + \frac{\partial \theta}{\partial x} \frac{\partial}{\partial \theta} + \frac{\partial \phi}{\partial x} \frac{\partial}{\partial \phi} = \sin \theta \cos \phi \frac{\partial}{\partial r} + \frac{\cos \theta \cos \phi}{r} \frac{\partial}{\partial \theta} - \frac{\sin \phi}{r \sin \theta} \frac{\partial}{\partial \phi} \\
\frac{\partial}{\partial y} &= \sin \theta \sin \phi \frac{\partial}{\partial r} + \frac{\cos \theta \sin \phi}{r} \frac{\partial}{\partial \theta} + \frac{\cos \phi}{r \sin \theta} \frac{\partial}{\partial \phi} \\
\frac{\partial}{\partial z} &= \cos \theta \frac{\partial}{\partial r} - \frac{\sin \theta}{r} \frac{\partial}{\partial \theta}
\end{aligned} \tag{53.65}$$

Ainsi, en tenant compte de ces relations, nous obtenons par exemple, dans le cas de l'opérateur:

$$L_x = \frac{\hbar}{i} (y \partial_z - z \partial_y) \tag{53.66}$$

les développements:

$$\begin{aligned}
y \frac{\partial}{\partial z} - z \frac{\partial}{\partial y} &= y \left(\cos \theta \frac{\partial}{\partial r} - \frac{\sin \theta}{r} \frac{\partial}{\partial \theta} \right) - z \left(\sin \theta \sin \phi \frac{\partial}{\partial r} + \frac{\cos \theta \sin \phi}{r} \frac{\partial}{\partial \theta} + \frac{\cos \phi}{r \sin \theta} \frac{\partial}{\partial \phi} \right) \\
&= r \sin \theta \sin \phi \left(\cos \theta \frac{\partial}{\partial r} - \frac{\sin \theta}{r} \frac{\partial}{\partial \theta} \right) - r \cos \theta \left(\sin \theta \sin \phi \frac{\partial}{\partial r} + \frac{\cos \theta \sin \phi}{r} \frac{\partial}{\partial \theta} + \frac{\cos \phi}{r \sin \theta} \frac{\partial}{\partial \phi} \right) \\
&= r \sin \theta \sin \phi \left(-\frac{\sin \theta}{r} \frac{\partial}{\partial \theta} \right) - r \cos \theta \left(\frac{\cos \theta \sin \phi}{r} \frac{\partial}{\partial \theta} + \frac{\cos \phi}{r \sin \theta} \frac{\partial}{\partial \phi} \right) \\
&= \sin \theta \sin \phi \left(-\sin \theta \frac{\partial}{\partial \theta} \right) - \cos \theta \left(\cos \theta \sin \phi \frac{\partial}{\partial \theta} + \frac{\cos \phi}{\sin \theta} \frac{\partial}{\partial \phi} \right) \\
&= (-\sin \theta \sin \theta \sin \phi - \cos \theta \cos \theta \sin \phi) \frac{\partial}{\partial \theta} - \cos \theta \left(\frac{\cos \phi}{\sin \theta} \frac{\partial}{\partial \phi} \right) \\
&= -\sin \phi \frac{\partial}{\partial \theta} - \cot \theta \cos \phi \frac{\partial}{\partial \phi}
\end{aligned} \tag{53.67}$$

ce qui donne le résultat suivant:

$$L_x = i\hbar \left(\sin \phi \frac{\partial}{\partial \theta} + \cot \theta \cos \phi \frac{\partial}{\partial \phi} \right) \tag{53.68}$$

En procédant de même avec:

$$L_y = \frac{\hbar}{i} (z \partial_x - x \partial_z) \tag{53.69}$$

en faisant les développements:

$$\begin{aligned}
z \frac{\partial}{\partial x} - x \frac{\partial}{\partial z} &= z \left(\sin \theta \cos \phi \frac{\partial}{\partial r} + \frac{\cos \theta \cos \phi}{r} \frac{\partial}{\partial \theta} - \frac{\sin \phi}{r \sin \theta} \frac{\partial}{\partial \phi} \right) - x \left(\cos \theta \frac{\partial}{\partial r} - \frac{\sin \theta}{r} \frac{\partial}{\partial \theta} \right) \\
&= r \cos \theta \left(\sin \theta \cos \phi \frac{\partial}{\partial r} + \frac{\cos \theta \cos \phi}{r} \frac{\partial}{\partial \theta} - \frac{\sin \phi}{r \sin \theta} \frac{\partial}{\partial \phi} \right) - r \sin \theta \cos \phi \left(\cos \theta \frac{\partial}{\partial r} - \frac{\sin \theta}{r} \frac{\partial}{\partial \theta} \right) \\
&= r \cos \theta \left(\frac{\cos \theta \cos \phi}{r} \frac{\partial}{\partial \theta} - \frac{\sin \phi}{r \sin \theta} \frac{\partial}{\partial \phi} \right) - r \sin \theta \cos \phi \left(-\frac{\sin \theta}{r} \frac{\partial}{\partial \theta} \right) \\
&= \cos \theta \left(\cos \theta \cos \phi \frac{\partial}{\partial \theta} - \frac{\sin \phi}{\sin \theta} \frac{\partial}{\partial \phi} \right) - \sin \theta \cos \phi \left(-\sin \theta \frac{\partial}{\partial \theta} \right) \\
&= (\cos \theta \cos \theta \cos \phi + \sin \theta \cos \phi \sin \theta) \frac{\partial}{\partial \theta} - \cos \theta \frac{\sin \phi}{\sin \theta} \frac{\partial}{\partial \phi} \\
&= \cos \phi (\cos^2 \theta + \sin^2 \theta) \frac{\partial}{\partial \theta} - \sin \phi \cot \theta \frac{\partial}{\partial \phi} \\
&= \cos \phi \frac{\partial}{\partial \theta} - \sin \phi \cot \theta \frac{\partial}{\partial \phi}
\end{aligned}$$

(53.70)

nous avons le résultat suivant:

$$L_y = -i\hbar \left(\cos \phi \frac{\partial}{\partial \theta} - \sin \phi \cot \theta \frac{\partial}{\partial \phi} \right) \quad (53.71)$$

Et pour finir avec:

$$L_z = \frac{\hbar}{i} (x \partial_y - y \partial_x) \quad (53.72)$$

en faisant les développements:

$$\begin{aligned}
x \frac{\partial}{\partial y} - y \frac{\partial}{\partial x} &= x \left(\sin \theta \sin \phi \frac{\partial}{\partial r} + \frac{\cos \theta \sin \phi}{r} \frac{\partial}{\partial \theta} + \frac{\cos \phi}{r \sin \theta} \frac{\partial}{\partial \phi} \right) \\
&- y \left(\sin \theta \cos \phi \frac{\partial}{\partial r} + \frac{\cos \theta \cos \phi}{r} \frac{\partial}{\partial \theta} - \frac{\sin \phi}{r \sin \theta} \frac{\partial}{\partial \phi} \right) \\
&= r \sin \theta \cos \phi \left(\sin \theta \sin \phi \frac{\partial}{\partial r} + \frac{\cos \theta \sin \phi}{r} \frac{\partial}{\partial \theta} + \frac{\cos \phi}{r \sin \theta} \frac{\partial}{\partial \phi} \right) \\
&- r \sin \theta \sin \phi \left(\sin \theta \cos \phi \frac{\partial}{\partial r} + \frac{\cos \theta \cos \phi}{r} \frac{\partial}{\partial \theta} - \frac{\sin \phi}{r \sin \theta} \frac{\partial}{\partial \phi} \right) \\
&= r \sin \theta \cos \phi \left(\frac{\cos \theta \sin \phi}{r} \frac{\partial}{\partial \theta} + \frac{\cos \phi}{r \sin \theta} \frac{\partial}{\partial \phi} \right) - r \sin \theta \sin \phi \left(\frac{\cos \theta \cos \phi}{r} \frac{\partial}{\partial \theta} - \frac{\sin \phi}{r \sin \theta} \frac{\partial}{\partial \phi} \right) \\
&= r \sin \theta \cos \phi \left(\frac{\cos \phi}{r \sin \theta} \frac{\partial}{\partial \phi} \right) - r \sin \theta \sin \phi \left(- \frac{\sin \phi}{r \sin \theta} \frac{\partial}{\partial \phi} \right) \\
&= \cos^2 \phi \frac{\partial}{\partial \phi} + \sin^2 \phi \frac{\partial}{\partial \phi} = \frac{\partial}{\partial \phi}
\end{aligned}$$

(53.73)

nous avons le résultat suivant:

$$L_z = -i\hbar \frac{\partial}{\partial \phi} \quad (53.74)$$

Finalement, nous avons donc peu de libertés de mouvement pour notre rotateur rigide (car il est très rigide...) et nous pouvons écrire pour l'équation de Schrödinger:

$$H_{\text{rot}} = \frac{L^2}{2J} \Rightarrow H_{\text{rot}} \Psi = E \Psi \quad (53.75)$$

où H_{rot} est rappelons-le, vu comme un opérateur linéaire fonctionnel, et l'énergie totale E comme sa valeur propre correspondante.

Dès lors, nous pouvons écrire que l'opérateur moment cinétique (nous changeons la notation afin de ne pas confondre par la suite opérateur et valeur propre conformément aux remarques que nous avons faites lors des énoncés des postulats de la physique quantique ondulatoire):

$$\hat{L}_z = -i\hbar \partial_\phi \quad (53.76)$$

Ainsi, les fonctions propres $\Phi(\phi)$ de $\hat{L}_z$ sont solutions de l'équation aux valeurs et fonctions propres:

$$\boxed{\hat{L}_z \Phi(\phi) = L_z \Phi(\phi)} \quad (53.77)$$

c'est-à-dire de l'équation différentielle:

$$\boxed{-i\hbar \partial_\phi \Phi(\phi) = L_z \Phi(\phi)} \quad (53.78)$$

où L_z est bien évidemment la valeur propre de $\hat{L}_z$. Une solution simple de cette équation différentielle serait:

$$\Phi(\phi) = c \cdot e^{\frac{i}{\hbar} L_z \phi} \quad (53.79)$$

avec comme condition d'uniformité selon les propriétés des formes complexes (cf. chapitre sur les Nombres):

$$\Phi(\phi + 2\pi) = \Phi(\phi) \quad (53.80)$$

Cette condition mathématique, impose la quantification évidente et remarquable suivante:

$$L_z = \hbar m_l \text{ avec } m_l = 0, \pm 1, \pm 2, \dots \quad (53.81)$$

où (rappel) m_l est le nombre quantique magnétique.

Sachant que (cf. chapitre de Physique Quantique Corpusculaire):

$$-l \leq m_l \leq +l \quad (53.82)$$

Nous pouvons écrire:

$$L_z = \hbar l \quad (53.83)$$

Dès lors, nous retrouvons le(s) résultat(s) que nous avons obtenus dans les chapitres de Physique Quantique Corpusculaire et de Physique Quantique Ondulatoire:

$$b_z = \hbar l \Rightarrow \vec{b} = \hbar \vec{l} \quad (53.84)$$

Ce qui est assez satisfaisant, même remarquable et jouissif (pour ne pas le dire...).

Ainsi, la mesure d'une composante du moment cinétique fournit toujours un multiple entier relatif de $\hbar$ qui apparaît comme l'unité naturelle du moment cinétique.

Les fonctions propres communes (!!!) aux opérateurs $\hat{L}^2$ et $\hat{L}_z$ sont dans un cadre plus général nécessairement de la forme (méthode de séparation des variables):

$$R(r)\Phi(\phi)\Theta(\theta) \quad (53.85)$$

Comme le rotateur est rigide, nous avons $R(r) = c^*$. Ce facteur s'éliminera de lui-même dans l'équation aux valeurs propres et fonctions propres que nous déterminerons plus loin. Donc nous pouvons ne pas le prendre en compte. Finalement, nous pouvons écrire au vu des développements qui ont précédé:

$$\Phi(\phi)\Theta(\theta) = e^{im_l\phi} \Theta(\theta) \quad (53.86)$$

Ce qui nous amène à l'équation aux valeurs et fonctions propres:

$$\hat{L}^2 \Phi(\phi)\Theta(\theta) = L^2 e^{im_l\phi} \Theta(\theta) \quad (53.87)$$

C'est-à-dire:

$$\boxed{-\hbar^2 \left(\frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2}{\partial \phi^2} \right) \Phi(\phi) \Theta(\theta) = L^2 e^{im_l \phi} \Theta(\theta)} \quad (53.88)$$

d'où:

$$\begin{aligned} & -\hbar^2 \left(\frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial (e^{im_l \phi} \Theta(\theta))}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2 (e^{im_l \phi} \Theta(\theta))}{\partial \phi^2} \right) = L^2 e^{im_l \phi} \Theta(\theta) \\ & -\hbar^2 \left(\frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \Theta(\theta) \underbrace{\frac{\partial (e^{im_l \phi})}{\partial \theta}}_{=0} + \sin \theta e^{im_l \phi} \frac{\partial \Theta(\theta)}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} e^{im_l \phi} \underbrace{\frac{\partial^2 (\Theta(\theta))}{\partial \phi^2}}_{=0} \right. \\ & \quad \left. + \frac{1}{\sin^2 \theta} \Theta(\theta) \frac{\partial^2 (e^{im_l \phi})}{\partial \phi^2} \right) = L^2 e^{im_l \phi} \Theta(\theta) \\ & -\hbar^2 \left(\frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \sin \theta e^{im_l \phi} \frac{\partial \Theta(\theta)}{\partial \theta} + \frac{1}{\sin^2 \theta} \Theta(\theta) \frac{\partial^2 (e^{im_l \phi})}{\partial \phi^2} \right) = L^2 e^{im_l \phi} \Theta(\theta) \\ & -\hbar^2 \left(\cancel{e^{im_l \phi}} \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \sin \theta \frac{\partial \Theta(\theta)}{\partial \theta} + \frac{1}{\sin^2 \theta} \Theta(\theta) (im_l)^2 \cancel{e^{im_l \phi}} \right) = L^2 \cancel{e^{im_l \phi}} \Theta(\theta) \\ & -\hbar^2 \left(\frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \sin \theta \frac{\partial \Theta(\theta)}{\partial \theta} - \frac{\Theta(\theta) m_l^2}{\sin^2 \theta} \right) = L^2 \Theta(\theta) \end{aligned} \quad (53.89)$$

En posant:

$$\xi = \cos \theta \Rightarrow \frac{\partial \xi}{\partial \theta} = -\sin(\theta) \Rightarrow \partial \theta = -\frac{\partial \xi}{\sin(\theta)} \quad (53.90)$$

et donc:

$$\cos^2 \theta + \sin^2 \theta = 1 \Rightarrow \xi^2 + \sin^2 \theta = 1 \Rightarrow \sin^2 \theta = 1 - \xi^2 \quad (53.91)$$

nous obtenons une équation différentielle du type "Fuchs" donnée par:

$$\begin{aligned}
& -\hbar^2 \left(\frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \sin \theta \frac{\partial \Theta(\theta)}{\partial \theta} - \frac{\Theta(\theta) m_l^2}{\sin^2 \theta} \right) = L^2 \Theta(\theta) \Rightarrow \\
& -\hbar^2 \left(\frac{1}{\cancel{\sin \theta}} \frac{\partial}{\partial \xi} \sin \theta \frac{\partial \Theta(\xi)}{\partial \xi} - \frac{\Theta(\xi) m_l^2}{\cancel{\sin(\theta)}} \right) - L^2 \Theta(\xi) = 0 \\
& -\hbar^2 \left(\frac{\partial}{\partial \xi} \sin^2 \theta \frac{\partial \Theta(\xi)}{\partial \xi} - \frac{\Theta(\xi) m_l^2}{\sin^2 \theta} \right) - L^2 \Theta(\xi) = 0 \\
& \left(\frac{\partial}{\partial \xi} (1 - \xi^2) \frac{\partial \Theta(\xi)}{\partial \xi} - \frac{\Theta(\xi) m_l^2}{1 - \xi^2} \right) + \frac{L^2}{\hbar^2} \Theta(\xi) = 0 \quad (53.92) \\
& \left(\frac{\partial}{\partial \xi} (1 - \xi^2) \frac{\partial \Theta(\xi)}{\partial \xi} \right) + \frac{L^2}{\hbar^2} \Theta(\xi) - \frac{\Theta(\xi) m_l^2}{1 - \xi^2} = 0 \\
& \left(\frac{\partial}{\partial \xi} (1 - \xi^2) \frac{\partial \Theta(\xi)}{\partial \xi} \right) + \left(\frac{L^2}{\hbar^2} - \frac{m_l^2}{1 - \xi^2} \right) \Theta(\xi) = 0 \\
& \left(\frac{\partial (1 - \xi^2)}{\partial \xi} \frac{\partial \Theta(\xi)}{\partial \xi} + (1 - \xi^2) \frac{\partial}{\partial \xi} \frac{\partial \Theta(\xi)}{\partial \xi} \right) + \left(\frac{L^2}{\hbar^2} - \frac{m_l^2}{1 - \xi^2} \right) \Theta(\xi) = 0 \\
& (-2\xi \Theta'(\xi) + (1 - \xi^2) \Theta''(\xi)) + \left(\frac{L^2}{\hbar^2} - \frac{m_l^2}{1 - \xi^2} \right) \Theta(\xi) = 0
\end{aligned}$$

D'où finalement:

$$\Theta''(\xi) - \frac{2\xi \Theta'(\xi)}{(1 - \xi^2)} + \left(\frac{L^2}{\hbar^2} - \frac{m_l^2}{1 - \xi^2} \right) \frac{\Theta(\xi)}{(1 - \xi^2)} = 0 \quad (53.93)$$

Dont les coefficients présentent des pôles (singularités) en $\xi = \pm 1$. Or, rappelons que nous avons:

$$L^2 = \hbar^2 l(l+1) \Rightarrow \frac{L^2}{\hbar^2} = l(l+1) \quad (53.94)$$

Ce qui fait que nous retrouvons souvent l'équation différentielle précédente sous la forme suivante dans la littérature après factorisation élémentaire de certains termes:

$$\frac{d}{d\xi} \left[(1 - \xi^2) \frac{d\Theta(\xi)}{d\xi} \right] + \left[l(l+1) - \frac{m_l^2}{1 - \xi^2} \right] \Theta(\xi) = 0 \quad (53.95)$$

Une solution non triviale étant, connaissant les équations différentielles de type Fuchs, ce qu'il est d'usage d'appeler les "**polynômes associés de Legendre**" (bien que ce ne soit pas un polynôme rigoureusement parlant....) car contenant en partie les polynômes de Legendre (cf. chapitre de Calcul Algébrique):

$$P_{m_l, l} = \Theta(\xi) = \frac{(-1)^l}{2^l l!} (1 - \xi^2)^{\frac{m_l}{2}} \frac{d^{m_l+l}}{d\xi^{m_l+l}} (1 - \xi^2)^l \quad (53.96)$$

ce que vous pouvez vérifier en injectant cette solution dans l'équation différentielle antécédente.

Suite à la demande d'un lecteur voici un exemple de vérification avant de poursuivre:

Le cas $m_l = l = 0$ avec se vérifie immédiatement. Prenons le cas où $m_l = l = 1$:

$$\begin{aligned}
 P_{m_l, l} &= P_{1,1} = \frac{(-1)^l}{2^l l!} (1 - \xi^2)^{\frac{m_l}{2}} \frac{d^{m_l+l}}{d\xi^{m_l+l}} (1 - \xi^2)^l \\
 &= \frac{(-1)^1}{2^1 1!} (1 - \xi^2)^{\frac{1}{2}} \frac{d^2}{d\xi^2} (1 - \xi^2)^1 = -\frac{1}{2} (1 - \xi^2)^{\frac{1}{2}} \frac{d^2}{d\xi^2} (1 - \xi^2) \quad (53.97) \\
 &= \frac{1}{2} (1 - \xi^2)^{\frac{1}{2}} \frac{d^2}{d\xi^2} \xi^2 = \sqrt{1 - \xi^2}
 \end{aligned}$$

L'équation différentielle devient elle pour $m_l = l = 1$:

$$\frac{d}{d\xi} \left[(1 - \xi^2) \frac{d\Theta(\xi)}{d\xi} \right] + \left[1(1+1) - \frac{1^2}{1 - \xi^2} \right] \Theta(\xi) = 0 \quad (53.98)$$

Soit:

$$\frac{d}{d\xi} \left[(1 - \xi^2) \frac{d\Theta(\xi)}{d\xi} \right] + \left[\frac{1 - 2\xi^2}{1 - \xi^2} \right] \Theta(\xi) = 0 \quad (53.99)$$

Et nous injectons le polynôme associé de Lagrange:

$$\frac{d}{d\xi} \left[(1 - \xi^2) \frac{d\sqrt{1 - \xi^2}}{d\xi} \right] + \left[\frac{1 - 2\xi^2}{1 - \xi^2} \right] \sqrt{1 - \xi^2} = 0 \quad (53.100)$$

Soit:

$$\frac{d}{d\xi} \left[-(1 - \xi^2) \frac{\xi}{\sqrt{1 - \xi^2}} \right] + \left[\frac{1 - 2\xi^2}{1 - \xi^2} \right] \sqrt{1 - \xi^2} = 0 \quad (53.101)$$

Ce qui donne après une petite simplification:

$$\frac{d}{d\xi} \left[-\sqrt{1 - \xi^2} \xi \right] + \left[\frac{1 - 2\xi^2}{1 - \xi^2} \right] \sqrt{1 - \xi^2} = 0 \quad (53.102)$$

Et en dérivant:

$$\frac{\xi^2}{\sqrt{1-\xi^2}} - \sqrt{1-\xi^2} + \left[\frac{1-2\xi^2}{1-\xi^2} \right] \sqrt{1-\xi^2} = 0 \quad (53.103)$$

Concentrons-nous sur la partie gauche pour voir ce qu'elle vaut en mettant tout au même dénominateur:

$$\frac{\xi^2 (1-\xi^2) - \sqrt{1-\xi^2} \sqrt{1-\xi^2} (1-\xi^2) + (1-2\xi^2) \sqrt{1-\xi^2} \sqrt{1-\xi^2}}{\sqrt{1-\xi^2} (1-\xi^2)} \quad (53.104)$$

En simplifiant le numérateur, celui-ci devrait donc être nul. Voyons cela en le simplifiant une première fois:

$$\xi^2 (1-\xi^2) - (1-\xi^2)^2 + (1-2\xi^2)(1-\xi^2) \quad (53.105)$$

et en distribuant:

$$\xi^2 - \xi^4 - 1 + 2\xi^2 - \xi^4 + 1 - \xi^2 - 2\xi^2 + 2\xi^4 \quad (53.106)$$

Ce qui est bien nul!!

Donc finalement, nous avons des fonctions propres communes (car rappelons que les polynômes de Legendre sont orthogonaux entre eux) qui seront:

$$Y_{m,l}(\theta, \phi) = N_{m,l} P_{m,l}(\xi) e^{im_l \phi} \quad (53.107)$$

Pour normaliser la partie dépendante de ϕ , nous savons que l'exponentielle doit être multipliée par $(2\pi)^{-1/2}$ (voir chapitre de Statistiques, la partie concernant la fonction de distribution de Gauss-Laplace) ce qui nous permet déjà d'écrire:

$$Y_{m,l}(\theta, \phi) = \frac{N_{m,l}}{\sqrt{2\pi}} P_{m,l}(\xi) e^{im_l \phi} \quad (53.108)$$

Remarque: Il n'est pas besoin de faire de calculs compliqués pour calculer le facteur de normalisation de l'exponentielle, car dans le cadre d'une intégration sur tout l'espace, les trois facteurs de $Y_{m,l}(\theta, \phi)$ sont indépendants les uns des autres. Ainsi l'intégrale sera le produit des intégrales (cf. chapitre de Calcul Différentiel Et Intégral).

Finalement, nous devons trouver $N_{m,l}$ tel que:

$$\int |Y_{m,l}(\theta, \phi)|^2 d\Omega = \int Y_{m,l}(\theta, \phi) \bar{Y}_{m,l}(\theta, \phi) d\Omega = 1 \quad (53.109)$$

et nous verrons (ce que nous allons démontrer juste plus bas) que:

$$N_{m,l} = \sqrt{\frac{(2l+1) \cdot (l-m)!}{2(l+m)!}} \quad (53.110)$$

En résumé, nous écrivons (nous devrions plutôt conjuguer au futur...):

$$Y_{m,l}(\theta, \phi) = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{(2l+1) \cdot (l-m)!}{2(l+m)!}} \frac{1}{2^l l!} (1-\xi^2)^{m/2} \frac{d^{m+l}}{d\xi^{m+l}} (1-\xi^2)^l e^{im\phi} \quad (53.111)$$

où nous avons omis le facteur $(-1)^n$ puisque de toute façon dans le module de la fonction ce terme se multiplie à lui-même et donne alors $(-1)^{2n} = 1$.

Vérification la relation encadrée précédente (attention c'est un peu long et il est conseillé de le relire plusieurs fois):

Nous considérons les fonctions définies par:

$$Y_{m,l}(\theta, \phi) = \frac{N_{m,l}}{\sqrt{2\pi}} P_{m,l}(\cos \theta) \cdot e^{im\phi} \quad (53.112)$$

où:

$$P_{m,l}(x) = \frac{1}{2^l \cdot l!} (1-x^2)^{m/2} \cdot \frac{d^{m+l}}{dx^{m+l}} (1-x^2)^l \quad (53.113)$$

avec:

$$l \in \mathbb{N}, -l \leq m \leq l \quad \text{et} \quad \theta \in [0, \pi], \phi \in [0, 2\pi[\quad (53.114)$$

Le but va être donc de montrer que ces fonctions sont d'abord orthogonales et ensuite de trouver les constantes $N_{m,l}$ telles que $\|Y_{m,l}\| = 1$. Bref il va falloir retrousser les manches... du cerveau...

D'abord, démontrons pour des besoins ultérieurs que:

$$\frac{d^{m_l}}{dx^{m_l}} \left((1-x^2)^{m_l} \cdot \frac{d^{m_l+l}}{dx^{m_l+l}} (1-x^2)^l \right) = (-1)^{m_l} \frac{(l+m_l)!}{(l-m_l)!} \cdot \frac{d^l}{dx^l} (1-x^2)^l \quad (53.115)$$

Démonstration:

Si et seulement si $l = m_l = 0$ l'égalité est évidente. Supposons $l \geq 1$ (donc le cas général en dehors du cas particulier évident précédent) et soit P un polynôme réel de degré $\leq l-1$.

Posons:

$$Z(x) := (1-x^2)^{m_l} \cdot \frac{d^{m_l+l}}{dx^{m_l+l}} (1-x^2)^l \quad (53.116)$$

Montrons que (produit scalaire fonctionnel):

$$\left\langle \frac{d^{m_i}}{dx^{m_i}} Z \middle| P \right\rangle = 0 \quad (53.117)$$

dans $\mathcal{C}(x \in [-1, 1], \mathbb{R})$.

Effectivement, rappelons que nous avons fait le changement de variable:

$$x = \cos(\theta), -\pi \leq \theta \leq +\pi \Rightarrow -1 \leq x \leq +1 \quad (53.118)$$

En intégrant par parties, nous obtenons:

$$\left\langle \frac{d^{m_i}}{dx^{m_i}} Z \middle| P \right\rangle = \int_{-1}^1 \frac{d^{m_i}}{dx^{m_i}} Z(x) \cdot P(x) dx = \frac{d^{m_i-1}}{dx^{m_i-1}} Z(x) \cdot P(x) \Big|_{-1}^1 - \int_{-1}^1 \frac{d^{m_i-1}}{dx^{m_i-1}} Z(x) \cdot \frac{d}{dx} P(x) dx$$

remarquons que pour tout $0 \leq j \leq m_i - 1$, $\frac{d^j}{dx^j} Z(x)$ est nul en $x = \pm 1$. Par suite (par extension), la relation précédente se simplifie en:

$$\left\langle \frac{d^{m_i}}{dx^{m_i}} Z \middle| P \right\rangle = - \int_{-1}^1 \frac{d^{m_i-1}}{dx^{m_i-1}} Z(x) \cdot \frac{d}{dx} P(x) dx \quad (53.119)$$

Après m_i intégrations par parties, nous obtenons:

$$\left\langle \frac{d^{m_i}}{dx^{m_i}} Z \middle| P \right\rangle = (-1)^{m_i} \int_{-1}^1 Z(x) \cdot \frac{d^{m_i}}{dx^{m_i}} P(x) dx \quad (53.120)$$

Si $\deg(P) < m_i$ alors l'expression précédente montre trivialement que:

$$\left\langle \frac{d^{m_i}}{dx^{m_i}} Z \middle| P \right\rangle = 0 \quad (53.121)$$

Si $\deg(P) \geq m_i$ alors en posant:

$$h(x) := (1 - x^2)^{m_i} \cdot \frac{d^{m_i}}{dx^{m_i}} P(x) \quad (53.122)$$

Nous obtenons:

$$\begin{aligned} \left\langle \frac{d^{m_i}}{dx^{m_i}} Z \middle| P \right\rangle &= (-1)^{m_i} \int_{-1}^1 (1 - x^2)^{m_i} \cdot \frac{d^{m_i+l}}{dx^{m_i+l}} (1 - x^2)^l \cdot \frac{d^{m_i}}{dx^{m_i}} P(x) dx \\ &= (-1)^{m_i} \int_{-1}^1 \frac{d^{m_i+l}}{dx^{m_i+l}} (1 - x^2)^l \cdot h(x) dx \end{aligned} \quad (53.123)$$

remarquons encore une fois que $\frac{d^j}{dx^j} h(x)$ s'annule en $x = \pm 1$ pour tout $j \leq m_l - 1$. En intégrant m_l fois par parties l'expression précédente, nous trouvons:

$$\left\langle \frac{d^{m_l}}{dx^{m_l}} Z \middle| P \right\rangle = \int_{-1}^1 \frac{d^l}{dx^l} (1-x^2)^l \cdot \frac{d^{m_l}}{dx^{m_l}} h(x) dx \quad (53.124)$$

or h est un polynôme de degré $m_l + \deg(P)$.

Effectivement, le premier facteur est de degré $2m$ et la dérivée m_l -ème de $P(x)$ est de degré $P - m_l$, dès lors:

$$2m_l + P - m_l = m_l + P = m_l + \deg(P) \quad (53.125)$$

donc $\frac{d^{m_l}}{dx^{m_l}} h(x)$ est un polynôme de degré $\deg(P) \leq l - 1$ et sachant que $\frac{d^l}{dx^l} (1-x^2)^l$ est à une constante près égal au l -ème polynôme de Legendre (cf. chapitre de Calcul Algébrique) nous avons alors:

$$\left\langle \frac{d^{m_l}}{dx^{m_l}} Z \middle| P \right\rangle = \int_{-1}^1 \frac{d^l}{dx^l} (1-x^2)^l \cdot \frac{d^{m_l}}{dx^{m_l}} h(x) dx = 0 \quad (53.126)$$

Nous venons de montrer que $\frac{d^{m_l}}{dx^{m_l}} Z$ est orthogonal à tout polynôme de degré $\leq l - 1$.

□C.Q.F.D.

$\frac{d^{m_l}}{dx^{m_l}} Z$ est un polynôme de degré l (il suffit de le vérifier pour quelques valeurs) donc cherchons s'il existe une constante $C \in \mathbb{R}$ telle que:

$$\frac{d^{m_l}}{dx^{m_l}} Z = C \cdot \frac{d^l}{dx^l} (1-x^2)^l \quad (53.127)$$

avec rappelons-le:

$$Z(x) := (1-x^2)^{m_l} \cdot \frac{d^{m_l+l}}{dx^{m_l+l}} (1-x^2)^l \quad (53.128)$$

Nous pouvons déterminer la constante C en comparant les coefficients dominants des polynômes:

$$\frac{d^{m_l}}{dx^{m_l}} Z \text{ et } \frac{d^l}{dx^l} (1-x^2)^l \quad (53.129)$$

Le coefficient dominant de $\frac{d^{m_l}}{dx^{m_l}} Z$ est:

$$(-1)^{l+m_l} \frac{(2l)!}{(l-m_l)!} \frac{(l+m_l)!}{l!} \quad (53.130)$$

et le coefficient dominant de $\frac{d^l}{dx^l}(1-x^2)^l$ est:

$$(-1)^l \frac{(2l)!}{l!} \quad (53.131)$$

Ainsi:

$$(-1)^{l+m_l} \frac{(2l)!}{(l-m_l)!} \frac{(l+m_l)!}{l!} = C \cdot (-1)^l \frac{(2l)!}{l!} \quad (53.132)$$

c'est-à-dire:

$$C = (-1)^{m_l} \frac{(l+m_l)!}{(l-m_l)!} \quad (53.133)$$

Nous aurions donc pour $l \geq m_l \geq 0$ (on intègre par parties autant de fois qu'il le faut à gauche et à droite - nécessairement - pour obtenir ce résultat):

$$Z(x) = (1-x^2)^{m_l} \cdot \frac{d^{m_l+l}}{dx^{m_l+l}}(1-x^2)^l = (-1)^{m_l} \frac{(l+m_l)!}{(l-m_l)!} \cdot \frac{d^{l-m_l}}{dx^{l-m_l}}(1-x^2)^l \quad (53.134)$$

Maintenant, établissons une relation remarquable qui existerait entre $P_{-m_l, l}$ et $P_{m_l, l}$ (et qui nous sera utile par la suite). Supposons pour cela $0 \leq m_l \leq l$ et rappelons qu'à la base:

$$P_{m_l, l}(x) = \frac{1}{2^l \cdot l!} (1-x^2)^{m_l/2} \cdot \frac{d^{m_l+l}}{dx^{m_l+l}}(1-x^2)^l \quad (53.135)$$

Donc cela nous amène à écrire (rien de particulier):

$$P_{-m_l, l}(x) = \frac{1}{2^l \cdot l!} (1-x^2)^{-m_l/2} \cdot \left[\frac{d^{l-m_l}}{dx^{l-m_l}}(1-x^2)^l \right] \quad (53.136)$$

remplacé par le résultat obtenu plus haut

Par les résultats précédents ($(-1)^{m_l} = (-1)^{-m_l}$):

$$\begin{aligned} (1-x^2)^{m_l} \cdot \frac{d^{m_l+l}}{dx^{m_l+l}}(1-x^2)^l &= (-1)^{m_l} \frac{(l+m_l)!}{(l-m_l)!} \cdot \frac{d^{l-m_l}}{dx^{l-m_l}}(1-x^2)^l \\ \Rightarrow \frac{d^{l-m_l}}{dx^{l-m_l}}(1-x^2)^l &= (-1)^{m_l} \frac{(l-m_l)!}{(l+m_l)!} (1-x^2)^{m_l} \cdot \frac{d^{m_l+l}}{dx^{m_l+l}}(1-x^2)^l \end{aligned} \quad (53.137)$$

cela nous amène à écrire:

$$\begin{aligned}
 P_{-m_l, l}(x) &= \frac{(-1)^{m_l}}{2^l \cdot l!} \cdot \frac{(l - m_l)!}{(l + m_l)!} (1 - x^2)^{-m_l/2} \cdot (1 - x^2)^{m_l} \cdot \frac{d^{m_l+l}}{dx^{m_l+l}} (1 - x^2)^l \\
 &= \frac{(-1)^{m_l}}{2^l \cdot l!} \cdot \frac{(l - m_l)!}{(l + m_l)!} (1 - x^2)^{m_l/2} \cdot \frac{d^{m_l+l}}{dx^{m_l+l}} (1 - x^2)^l = (-1)^{m_l} \cdot \frac{(l - m_l)!}{(l + m_l)!} P_{m_l, l}
 \end{aligned} \quad (53.138)$$

Ainsi, nous obtenons:

$$P_{-m_l, l}(x) = (-1)^{m_l} \cdot \frac{(l - m_l)!}{(l + m_l)!} P_{m_l, l} \quad (53.139)$$

Nous allons à présent (enfin !) montrer que les fonctions $Y_{m_l, l}$ sont orthogonales.

D'abord, démontrons que:

$$\int_{-1}^1 P_l(x) \cdot x^l dx = \frac{(l!)^2}{(2l+1)!} 2^{l+1} \quad (53.140)$$

où P_l est le l -ème polynôme de Legendre (d'où le nom d'origine de "polynôme associé de Legendre" ...).

Démonstration:

D'abord, nous avons démontré que les polynômes de Legendre satisfont à la relation de récurrence suivante (cf. chapitre de Calcul Algébrique):

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x) \quad (53.141)$$

pour $n \geq 1$.

En multipliant l'égalité ci-dessus par x^{n-1} et en intégrant, nous obtenons:

$$(n+1) \int_{-1}^1 P_{n+1}(x) x^{n-1} dx = (2n+1) \int_{-1}^1 P_n(x) x^n dx - n \int_{-1}^1 P_{n-1}(x) x^{n-1} dx \quad (53.142)$$

Or:

$$\int_{-1}^1 P_{n+1}(x) x^{n-1} dx = 0 \quad (53.143)$$

Rappelons que les P_{n+1} polynômes forment une base orthogonale dont les polynômes qui l'engendrent sont de degré croissant de 0 à n , donc un polynôme d'ordre inférieur - exprimé dans un sous-espace vectoriel - sera toujours perpendiculaire aux vecteurs (polynômes) générant les dimensions supérieures. Ainsi, si nous prenons l'exemple de $\mathbb{R}^3$ engendré par la base $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$, alors un vecteur $\vec{v}$ exprimé par la combinaison linéaire de $\vec{e}_1, \vec{e}_2$, sera toujours perpendiculaire à $\vec{e}_3$ et donc un produit scalaire avec celui-ci.

Et donc par suite:

$$\int_{-1}^1 P_n(x) x^n dx = \frac{n}{(2n+1)} \int_{-1}^1 P_{n-1}(x) x^{n-1} dx \quad (53.144)$$

Posons:

$$S_n = \int_{-1}^1 P_n(x) x^n dx \quad (53.145)$$

L'expression précédente devient (pour rappel $P_0(x) = 1$):

$$S_n = \frac{n}{(2n+1)} S_{n-1} \text{ avec } S_0 = S_n = \int_{-1}^1 P_0(x) x^0 dx = \int_{-1}^1 1 dx = 2 \quad (53.146)$$

Ainsi par récurrence:

$$S_n = \frac{n}{(2n+1)} S_{n-1} = 2 \cdot \frac{1}{3} \cdot \frac{2}{5} \cdot \frac{3}{7} \cdot \dots \cdot \frac{n}{(2n+1)} = \frac{2}{1} \cdot \frac{1}{3} \cdot \frac{2}{5} \cdot \frac{3}{7} \cdot \dots \cdot \frac{n}{(2n+1)} \quad (53.147)$$

De plus comme:

$$1 \cdot 2 \cdot 3 \cdot 4 \cdot \dots \cdot n = n! \quad (53.148)$$

Nous avons alors pour la relation antérieure le dénominateur qui peut bien évidemment se récrire:

$$1 \cdot 3 \cdot 5 \cdot 7 \cdot \dots \cdot (2n+1) = \frac{(2n+1)!}{2 \cdot 4 \cdot 6 \cdot \dots \cdot (2n)} \quad (53.149)$$

Nous avons aussi:

$$2 \cdot 4 \cdot 6 \cdot \dots \cdot (2n) = \prod_{j=1}^n 2j = 2^n n! \quad (53.150)$$

Donc au final nous pouvons simplifier le dénominateur de la façon suivante:

$$1 \cdot 3 \cdot 5 \cdot \dots \cdot (2n+1) = \frac{(2n+1)!}{2 \cdot 4 \cdot 6 \cdot \dots \cdot (2n)} = \frac{(2n+1)!}{2^n \cdot n!} \quad (53.151)$$

et:

$$S_n = 2 \cdot \frac{1}{3} \cdot \frac{2}{5} \cdot \frac{3}{7} \cdot \dots \cdot \frac{n}{(2n+1)} = 2^{n+1} \frac{(n!)^2}{(2n+1)!} \quad (53.152)$$

Nous avons donc bien démontré (c'est juste au cas où... vous ne suivriez plus...) que:

$$\int_{-1}^1 P_l(x) \cdot x^l dx = \frac{(l!)^2}{(2l+1)!} 2^{l+1} \quad (53.153)$$

Attaquons-nous enfin à ce qui nous intéresse. C'est-à-dire démontrer que:

$$\langle Y_{m,l} | Y_{j,k} \rangle = \begin{cases} \frac{2N_{m,l}^2}{2l+1} \cdot \frac{(l+m_l)!}{(l-m_l)!} & \text{si } m_l = j \text{ et } l = k \\ 0 & \text{sinon} \end{cases} \quad (53.154)$$

Démonstration:

Si $m_l \neq j$:

$$\langle Y_{m,l} | Y_{j,k} \rangle = \frac{N_{m,l} \cdot N_{j,k}}{2\pi} \cdot \int_0^{2\pi} \int_0^\pi P_{m,l}(\cos \theta) \cdot P_{j,k}(\cos \theta) \cdot e^{i\phi(m_l-j)} d\Omega = 0 \quad (53.155)$$

où:

$$d\Omega = \sin \theta d\theta d\phi \quad (53.156)$$

Remarque: Rappelons que le jacobien en coordonnées sphériques est $r^2 \sin \theta$ (cf. chapitre de Calcul Différentiel Et Intégral) et comme la fonction intégrée ci-dessus n'est pas dépendante de r , nous avons sorti le terme $r^2 dr$ de cette intégrale (par contre nous le retrouverons ce même terme dans la fonction $R(r)$ présente dans l'équation de Schrödinger).

Et avec:

$$\int_0^{2\pi} e^{i\phi(m_l-j)} d\phi = 0 \quad (53.157)$$

Si $l > k$ et $m_l = j$ alors d'abord le produit scalaire se simplifie:

$$\langle Y_{m,l} | Y_{m,k} \rangle = N_{m,l} \cdot N_{m,k} \cdot \int_0^\pi P_{m,l}(\cos \theta) \cdot P_{m,k}(\cos \theta) \sin \theta d\theta \quad (53.158)$$

En faisant le changement de variable $x = \cos \theta$ nous obtenons:

$$\langle Y_{m,l} | Y_{m,k} \rangle = N_{m,l} \cdot N_{m,k} \cdot \int_{-1}^1 P_{m,l}(x) \cdot P_{m,k}(x) dx \quad (53.159)$$

Supposons $m_l \geq 0$:

$$P_{m,l}(x) = \frac{1}{2^l \cdot l!} (1-x^2)^{m_l/2} \cdot \frac{d^{m_l+l}}{dx^{m_l+l}} (1-x^2)^l = (-1)^l (1-x^2)^{m_l/2} \frac{d^{m_l}}{dx^{m_l}} P_l(x) \quad (53.160)$$

où $P_l(x)$ est le l -ème polynôme de Legendre. Ainsi l'expression du produit scalaire devient:

$$\langle Y_{m_i, l} | Y_{m_i, k} \rangle = (-1)^{l+k} N_{m_i, l} \cdot N_{m_i, k} \cdot \int_{-1}^1 (1-x^2)^{m_i} \frac{d^{m_i}}{dx^{m_i}} P_l(x) \cdot \frac{d^{m_i}}{dx^{m_i}} P_k(x) dx \quad (53.161)$$

Si nous posons:

$$h(x) := (1-x^2)^{m_i} \cdot \frac{d^{m_i}}{dx^{m_i}} P_k(x) \quad (53.162)$$

alors la relation devient:

$$\langle Y_{m_i, l} | Y_{m_i, k} \rangle = (-1)^{l+k} N_{m_i, l} \cdot N_{m_i, k} \cdot \int_{-1}^1 \frac{d^{m_i}}{dx^{m_i}} P_l(x) \cdot h(x) dx. \quad (53.163)$$

En intégrant m fois par parties l'expression ci-dessus nous obtenons:

$$\langle Y_{m_i, l} | Y_{m_i, k} \rangle = (-1)^{l+k+m_i} N_{m_i, l} \cdot N_{m_i, k} \cdot \int_{-1}^1 P_l(x) \cdot \frac{d^{m_i}}{dx^{m_i}} h(x) dx \quad (53.164)$$

Or $\frac{d^{m_i}}{dx^{m_i}} h(x)$ est un polynôme de degré k . Sachant que $l > k$, cette dernière intégrale est nulle pour les mêmes raisons que celles évoquées précédemment. Donc:

$$\langle Y_{m_i, l} | Y_{m_i, k} \rangle = 0 \quad (53.165)$$

Si $m_i < 0$ alors nous avons démontré que:

$$P_{-m_i, l}(x) = (-1)^{m_i} \cdot \frac{(l-m_i)!}{(l+m_i)!} P_{m_i, l} \quad (53.166)$$

et donc:

$$\langle Y_{m_i, l} | Y_{m_i, k} \rangle = C \cdot \int_{-1}^1 P_{-m_i, l}(x) \cdot P_{-m_i, k}(x) dx = 0 \quad (53.167)$$

car $-m_i \geq 0$.

Il ne reste qu'à traiter le cas $m_i = j, l = k$. Supposons à nouveau $m_i \geq 0$. Alors comme avant nous avons:

$$P_{m_i, l}(x) = (-1)^l (1-x^2)^{m_i/2} \frac{d^{m_i}}{dx^{m_i}} P_l(x) \quad (53.168)$$

et:

$$\langle Y_{m_i, l} | Y_{m_i, l} \rangle = N_{m_i, l}^2 \cdot \int_{-1}^1 P_{m_i, l}(x)^2 dx = N_{m_i, l}^2 \cdot \int_{-1}^1 (1-x^2)^{m_i} \frac{d^{m_i}}{dx^{m_i}} P_l(x) \cdot \frac{d^{m_i}}{dx^{m_i}} P_l(x) dx \quad (53.169)$$

Posons:

$$h(x) := (1 - x^2)^{m_l} \frac{d^{m_l}}{dx^{m_l}} P_l \quad (53.170)$$

La relation devient alors:

$$\langle Y_{m_l} | Y_{m_l} \rangle = N_{m_l}^2 \cdot \int_{-1}^1 \frac{d^{m_l}}{dx^{m_l}} P_l(x) \cdot h(x) dx \quad (53.171)$$

En intégrant m fois par parties, nous trouvons:

$$\langle Y_{m_l} | Y_{m_l} \rangle = N_{m_l}^2 \cdot (-1)^{m_l} \int_{-1}^1 P_l(x) \cdot \frac{d^{m_l}}{dx^{m_l}} h(x) dx \quad (53.172)$$

$\frac{d^{m_l}}{dx^{m_l}} h(x)$ est un polynôme de degré l dont le coefficient dominant vaut:

$$(-1)^{m_l} \frac{(2l)!}{(l!)^2 \cdot 2^l} \cdot \frac{(l + m_l)!}{(l - m_l)!} \quad (53.173)$$

P_l étant orthogonal à tout polynôme de degré strictement inférieur à l , l'expression peut s'écrire:

$$\langle Y_{m_l} | Y_{m_l} \rangle = N_{m_l}^2 \cdot \frac{(2l)!}{(l!)^2 \cdot 2^l} \cdot \frac{(l + m_l)!}{(l - m_l)!} \int_{-1}^1 P_l(x) \cdot x^l dx \quad (53.174)$$

Or, nous avons démontré que:

$$\int_{-1}^1 P_l(x) \cdot x^l dx = \frac{(l!)^2}{(2l + 1)!} 2^{l+1} \quad (53.175)$$

donc:

$$\langle Y_{m_l} | Y_{m_l} \rangle = N_{m_l}^2 \cdot \frac{(2l)!}{(l!)^2 \cdot 2^l} \cdot \frac{(l + m_l)!}{(l - m_l)!} \frac{(l!)^2}{(2l + 1)!} 2^{l+1} = \frac{2N_{m_l}^2}{2l + 1} \cdot \frac{(l + m_l)!}{(l - m_l)!} \quad (53.176)$$

Si $m_l \leq 0$ nous savons que nous obtenons le résultat.

□C.Q.F.D.

Finalement ce dernier résultat nous donne aussi la condition de normalisation:

$$\langle Y_{m_l} | Y_{m_l} \rangle = \frac{2N_{m_l}^2}{2l + 1} \cdot \frac{(l + m_l)!}{(l - m_l)!} = 1 \Rightarrow N_{m_l} = \sqrt{\frac{(2l + 1) \cdot (l - m_l)!}{2(l + m_l)!}} \quad (53.177)$$

Et donc finalement:

$$Y_{m,l}(\theta, \phi) = \frac{N_{m,l}}{\sqrt{2\pi}} P_{m,l}(\cos \theta) \cdot e^{im\phi} \quad (53.178)$$

est bien une famille orthonormale. Soit explicitement:

$$Y_{m,l}(\theta, \phi) = \sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}} \frac{1}{2^l \cdot l!} (1 - \cos^2 \theta)^{m/2} \frac{d^{m+l}}{d \cos \theta^{m+l}} (1 - \cos^2 \theta)^l e^{im\phi} \quad (53.179)$$

Enfin, après cette interlude fortement mathématique (mais instructif quant à la méthodologie d'approche), nous voyons (ce qui est logique) qu'à chaque valeur de l correspondent donc $2l+1$ fonctions propres $Y_{m,l}(\theta, \phi)$. Nous disons encore que la valeur $\hbar^2 l(l+1)$ est $2l+1$ fois dégénérée puisque:

$$\begin{aligned} \hat{L}^2 Y_{m,l}(\theta, \phi) &= L^2 Y_{m,l}(\theta, \phi) \\ &\Updownarrow \\ \hat{L}^2 Y_{m,l}(\theta, \phi) &= \hbar^2 l(l+1) Y_{m,l}(\theta, \phi) \end{aligned} \quad (53.180)$$

Voici quelques valeurs de la fonction $Y_{m,l}(\theta, \phi)$ qui génère ce que nous appelons communément des "harmoniques sphériques":

$$\begin{aligned} Y_{0,0} &= \frac{1}{\sqrt{4\pi}} \\ Y_{0,1} &= \sqrt{\frac{3}{4\pi}} \cos \theta \\ Y_{1,1} &= \sqrt{\frac{3}{8\pi}} \sin \theta e^{i\phi} \\ Y_{-1,1} &= \sqrt{\frac{3}{8\pi}} \sin \theta e^{-i\phi} \\ Y_{0,2} &= \sqrt{\frac{5}{16\pi}} (3 \cos^2 \theta - 1) \\ Y_{1,2} &= \sqrt{\frac{15}{8\pi}} \sin \theta \cos \theta e^{i\phi} \\ Y_{-1,2} &= \sqrt{\frac{15}{8\pi}} \sin \theta \cos \theta e^{-i\phi} \\ Y_{2,2} &= \sqrt{\frac{15}{32\pi}} \sin^2 \theta e^{2i\phi} \\ Y_{-2,2} &= \sqrt{\frac{15}{32\pi}} \sin^2 \theta e^{-2i\phi} \end{aligned} \quad (53.181)$$

Voyons quelques tracés de ces magnifiques harmoniques sphériques que l'on peut obtenir avec Maple 4.00b en utilisant la commande suivante (il s'agit de la 6ème fonction d'harmonique sphérique ci-dessus):

```
>plot3d(Re(sqrt(15/(8*Pi))*(sin(theta)*cos(theta)*exp(I*phi)))^2,phi=0..2*Pi,theta=0..Pi,
coords=spherical,scaling=constrained);
```

- $Y_{0,0}$ (correspondant à $n = 1$!) donne une sphère (valeur constante quels que soient θ, ϕ) dont la densité de probabilité peut être représentée par la "carte photographique" ou "carte de densité" (la densité dans un état donné y est représenté par la densité de points clairs sur un fond foncé):

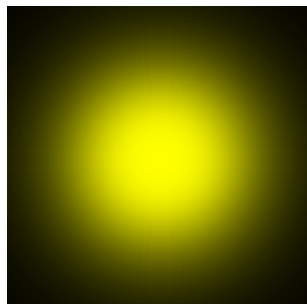


Figure: 53.4 - Carte de densité de 1s

Ce qui représente les orbitales 1s possibles.

- $Y_{0,1}, Y_{1,1}, Y_{-1,1}$ donnent (pour $n = 2$ au moins!):

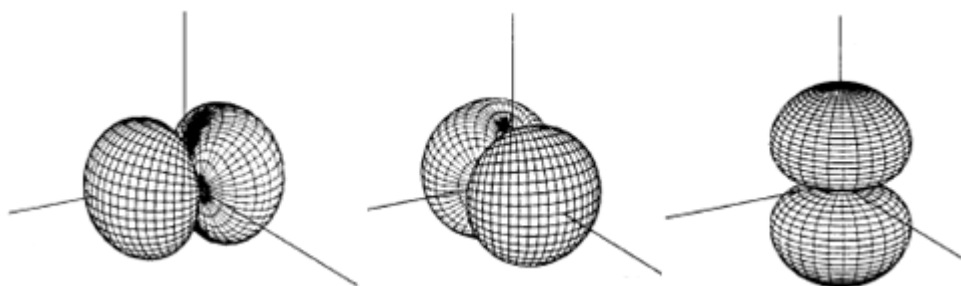


Figure: 53.5 - Orbitales 2p

Ce qui représente les orbitales 2p possibles, dont la densité de probabilité peut être représentée par sa carte de densité et d'isodensité:

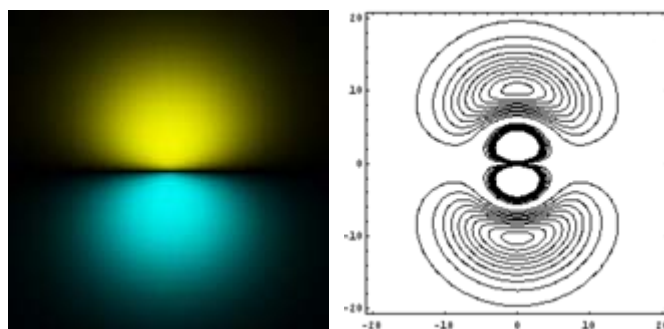


Figure: 53.6 - Carte de densité et d'isodensité de 2p

- $Y_{-2,2}, Y_{-1,2}, Y_{1,2}, Y_{2,2}, Y_{0,2}$ (pour $n = 3$ au moins!):

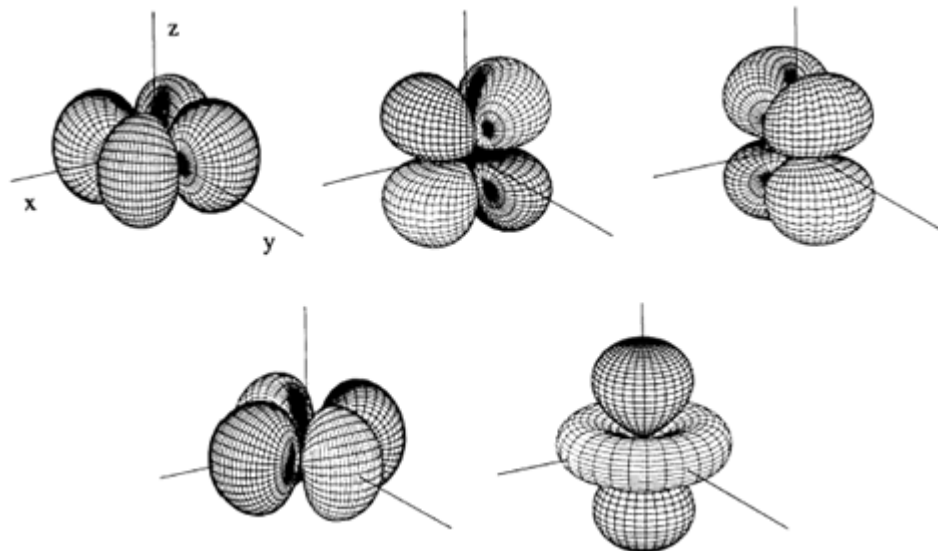


Figure: 53.7 - Orbitales 3d

Ce qui représente 5 orbitales centrosymétriques 3d possibles, dont la densité de probabilité peut être représentée par les (les deux dernières cartes représentent $Y_{0,2}$) cartes de densité:

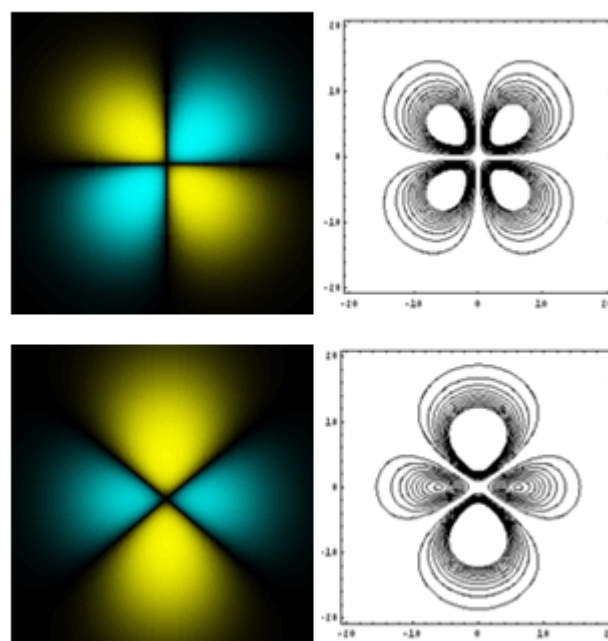


Figure: 53.8 - Carte de densité et d'isodensité de 3d

- $Y_{-3,3}, Y_{-2,3}, Y_{-1,3}, Y_{+1,3}, Y_{+2,3}, Y_{+3,3}, Y_{0,3}$ (pour $n = 4$ au moins!):

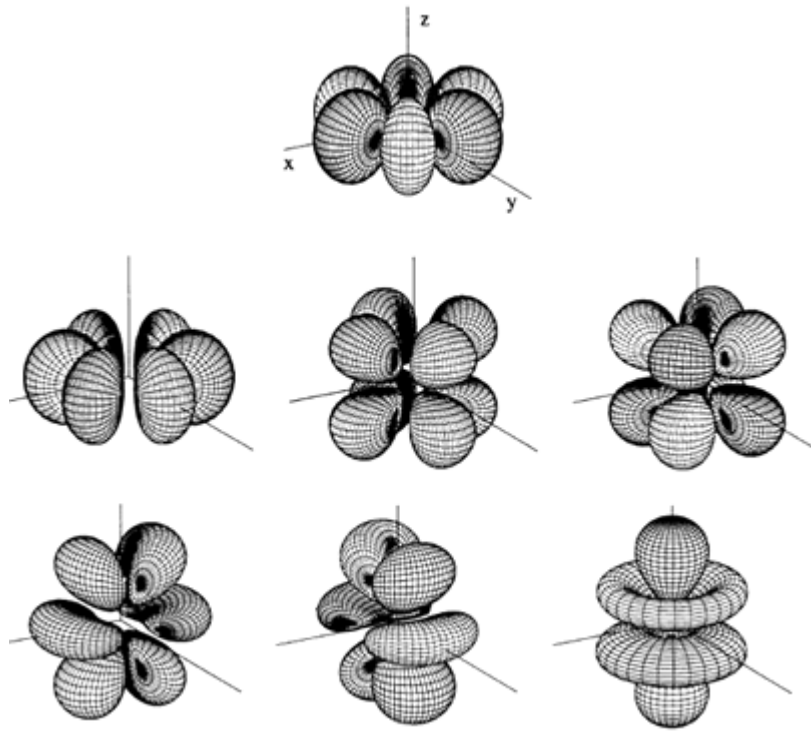


Figure: 53.9 - Orbitales 4f

Ce qui représente 7 orbitales anti-centrosymétriques 4f possibles, dont la densité de probabilité peut être représentée par (dans l'ordre: $Y_{0,3}, Y_{\pm 1,3}, Y_{\pm 2,3}, Y_{\pm 3,3}$):

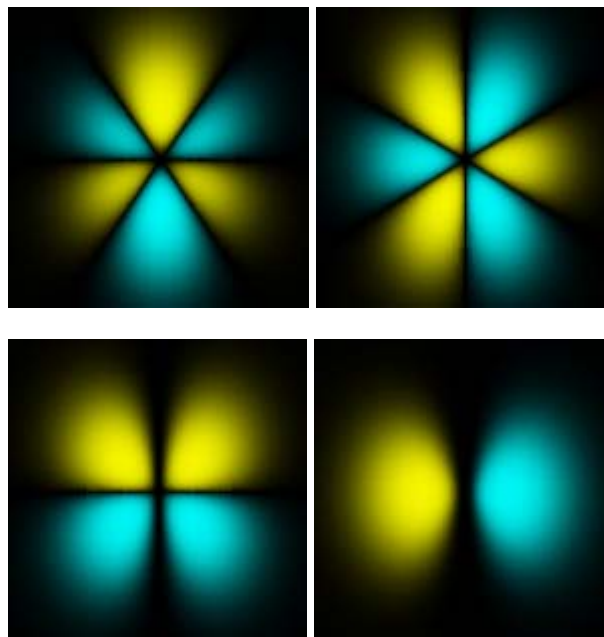


Figure: 53.10 - Carte de densité et d'isodensité de 4f

Les résultats précédents nous amènent donc à écrire:

$$\Psi(r, \theta, \phi) = R(r)\Phi(\phi)\Theta(\theta) = R(r)Y_{m,l}(\theta, \phi) \quad (53.182)$$

Remplaçant ceci dans l'équation de Schrödinger:

$$\left[T_r + \frac{l(l+1)\hbar^2}{2\mu r^2} + E_{pot} \right] \Psi = E \Psi \quad (53.183)$$

Nous obtenons ($T_r = 0$ dans le rotateur rigide mais $\neq 0$ dans le cas de l'atome hydrogénoïde):

$$\left[T_r + \frac{l(l+1)\hbar^2}{2\mu r^2} + E_{pot} \right] R(r) Y_{m,l}(\theta, \phi) = E R(r) Y_{m,l}(\theta, \phi) \quad (53.184)$$

Comme il n'y a dans cette relation aucun opérateur qui agit sur $Y_{m,l}(\theta, \phi)$, nous pouvons la simplifier de façon à obtenir:

$$\left[T_r + \frac{l(l+1)\hbar^2}{2\mu r^2} + E_{pot} \right] R(r) = E R(r) \quad (1) \quad (53.185)$$

où nous voyons dans ce cas général de l'atome isolé que les niveaux d'énergie ne dépendent plus de m_l (en raison de la symétrie sphérique du potentiel). Nous disons alors que les niveaux correspondants aux mêmes valeurs de n et de l sont tous confondus quelles que soient les valeurs de m_l .

Dans le cas où E_{pot} dérive du potentiel de Coulomb en $1/r$, cette équation radiale ne donne lieu à une solution $R(r)$ normalisable (différente de zéro aussi donc...) que pour des valeurs de l'énergie répondant à la loi de quantification suivante (tiens donc... quelle coïncidence, nous retrouvons l'expression démontrée dans les vieux modèles de la physique quantique corpusculaire!):

$$E_n = -hcR_H \frac{Z^2}{n^2} \quad (53.186)$$

où R_H est la constante de Rydberg telle que nous l'avons déterminée dans le chapitre de Physique Quantique Corpusculaire. Ainsi, dans ce cas les niveaux d'énergie correspondants aux mêmes valeurs de n sont tous confondus quel que soit l .

Pour une valeur donnée du nombre quantique principal n (rappelons que nous avons vu dans le chapitre de Physique Quantique Corpusculaire que $l \leq n-1$), il est possible de vérifier qu'il existe plusieurs solutions pour la fonction $R(r)$ selon la valeur du nombre quantique azimutal l . D'où l'identification des solutions par la paire (n, l) . Nous les notons $R_{n,l}(r)$. Ce sont des fonctions réelles de la variable r données par (il suffit de vérifier... car si elles marchent elles satisfont alors à l'équation de Schrödinger, nous ferons un exemple un peu plus loin):

$$\begin{aligned}
R_{10}(r) &= \left(\frac{Z}{a_0}\right)^{3/2} 2e^{-(Zr/a_0)} \\
R_{20}(r) &= \left(\frac{Z}{a_0}\right)^{3/2} \frac{1}{2\sqrt{2}} \left(2 - \frac{Zr}{a_0}\right) e^{-(Zr/2a_0)} \\
R_{21}(r) &= \left(\frac{Z}{a_0}\right)^{3/2} \frac{1}{2\sqrt{6}} \frac{Zr}{a_0} e^{-(Zr/2a_0)} \\
R_{30}(r) &= \left(\frac{Z}{a_0}\right)^{3/2} \frac{1}{9\sqrt{3}} \left(6 - \frac{4Zr}{a_0} + \frac{4Z^2 r^2}{9a_0^2}\right) e^{-(Zr/3a_0)} \\
R_{31}(r) &= \left(\frac{Z}{a_0}\right)^{3/2} \frac{1}{9\sqrt{6}} \left(\frac{8Zr}{3a_0} + \frac{4Z^2 r^2}{9a_0^2}\right) e^{-(Zr/3a_0)} \\
R_{32}(r) &= \left(\frac{Z}{a_0}\right)^{3/2} \frac{1}{9\sqrt{30}} \frac{4Z^2 r^2}{9a_0^2} e^{-(Zr/3a_0)}
\end{aligned} \tag{53.187}$$

où (attention certains ouvrages le donnent en unités naturelles!):

$$a_0 = 4\pi\epsilon_0 \frac{\hbar^2}{\mu e^2} \tag{53.188}$$

est l'équivalent du rayon de Bohr (pour la masse réduite) que nous avons déterminé dans le chapitre de Physique Quantique Corpusculaire avec comme différence que nous avons ici une masse réduite au lieu d'une masse simple.

Voyons cependant si notre équation de Schrödinger est bien vérifiée (prenons $n = 1, l = 0$):

$$\begin{aligned}
& \left[T_r + \frac{l(l+1)\hbar^2}{2\mu r^2} + E_{pot} \right] R(r) = E_1 R(r) \\
& \left[-\frac{\hbar^2}{2\mu r^2} \partial_r (r^2 \partial_r) - \frac{Ze^2}{4\pi\epsilon_0 r} \right] \left(\frac{Z}{a_0} \right)^{3/2} 2e^{-(Zr/a_0)} = E_1 \left(\frac{Z}{a_0} \right)^{3/2} 2e^{-(Zr/a_0)} \\
& \left[-\frac{\hbar^2}{2\mu r^2} \partial_r (r^2 \partial_r) - \frac{Ze^2}{4\pi\epsilon_0 r} \right] e^{-(Zr/a_0)} = E_1 e^{-(Zr/a_0)} \\
& -\frac{\hbar^2}{2\mu r^2} \partial_r (r^2 \partial_r e^{-(Zr/a_0)}) - \frac{Ze^2}{4\pi\epsilon_0 r} e^{-(Zr/a_0)} = E_1 e^{-(Zr/a_0)} \\
& -\frac{\hbar^2}{2\mu r^2} \partial_r \left(-\frac{Zr^2}{a_0} e^{-(Zr/a_0)} \right) - \frac{Ze^2}{4\pi\epsilon_0 r} e^{-(Zr/a_0)} = E_1 e^{-(Zr/a_0)} \\
& -\frac{\hbar^2}{2\mu r^2} \left(\frac{-2Zre^{-(Zr/a_0)}}{a_0} + \frac{Z^2 r^2 e^{-(Zr/a_0)}}{a_0^2} \right) - \frac{Ze^2}{4\pi\epsilon_0 r} e^{-(Zr/a_0)} = E_1 e^{-(Zr/a_0)} \quad (53.189) \\
& -\frac{\hbar^2}{2\mu r^2} \left(\frac{-2Zr}{4\pi\epsilon_0 \frac{\hbar^2}{\mu e^2}} + \frac{Z^2 r^2}{\left(4\pi\epsilon_0 \frac{\hbar^2}{\mu e^2} \right)^2} \right) - \frac{Ze^2}{4\pi\epsilon_0 r} = E_1 \\
& \frac{Ze^2}{4\pi\epsilon_0 r} - \frac{\hbar^2 Z^2}{2\mu \left(4\pi\epsilon_0 \frac{\hbar^2}{\mu e^2} \right)^2} - \frac{Ze^2}{4\pi\epsilon_0 r} = E_1 \\
& -\frac{1}{(4\pi\epsilon_0)^2} \frac{Z^2 \mu e^4}{2\hbar^2} = -\frac{1}{16\pi^2 \epsilon_0^2} \frac{Z^2 \mu e^4}{2\hbar^2} = -Z^2 \frac{\mu e^4}{8\epsilon_0^2 \hbar^2} = -hcR_H Z^2 = E_1
\end{aligned}$$

Ce qui correspond bien au résultat attendu.

Ce qui sous forme graphique nous donne pour la partie radiale $R_{n,l}(r)$:

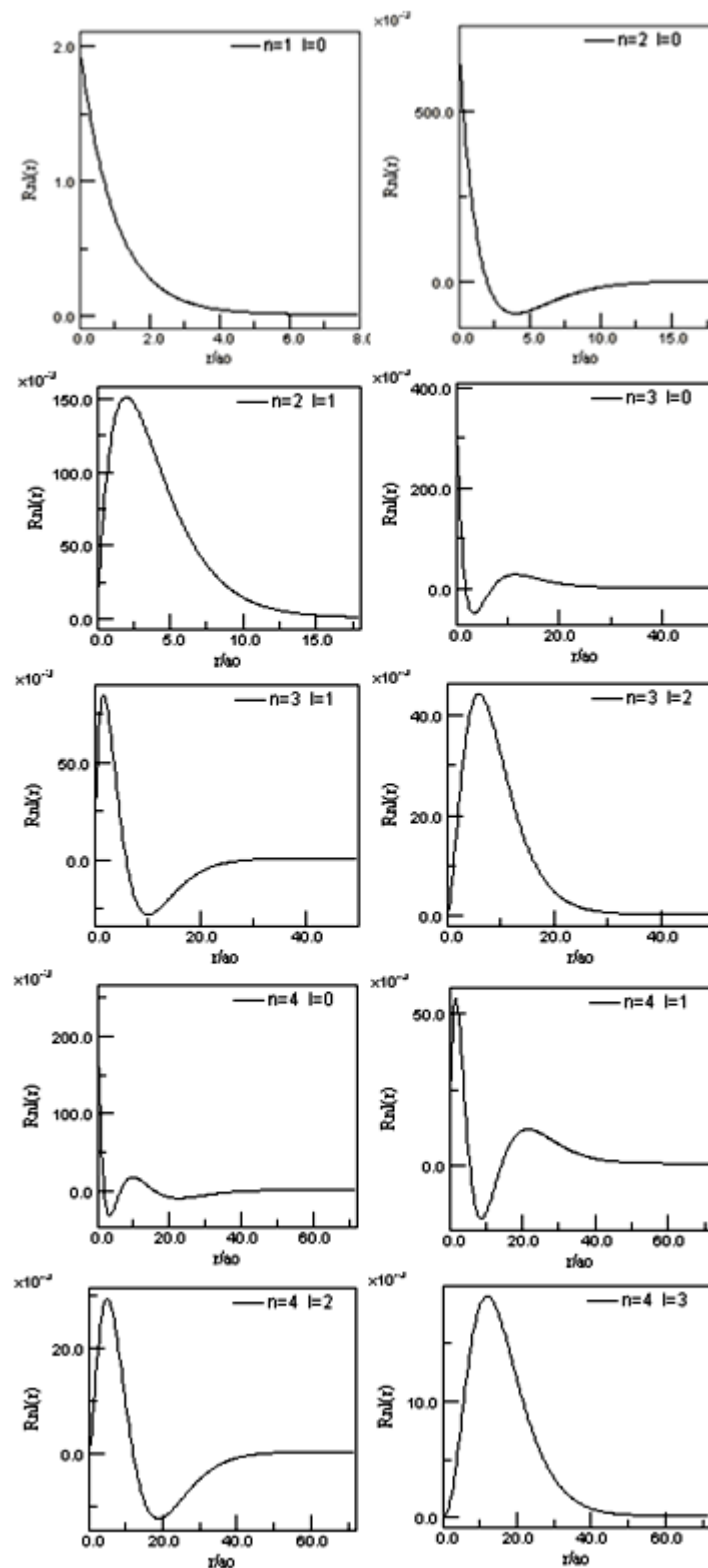


Figure: 53.11 - Tracé de quelques fonctions radiales

Étudions un peu plus en détail la fonction radiale dans le cas de l'atome d'hydrogène!:

Dans le cas de l'orbitale atomique 1s (cas particulier mais nous pourrions faire les mêmes calculs que ceux qui suivent avec toutes les autres orbitales!) nous avons donc pour l'atome d'hydrogène:

$$R_{10}(r) = \left(\frac{Z}{a_0}\right)^{3/2} 2e^{-(Zr/a_0)} = \left(\frac{1}{a_0}\right)^{3/2} 2e^{-(r/a_0)} = \frac{2}{a_0^{3/2}} e^{-(r/a_0)} \quad (53.190)$$

C'est donc bien une fonction exponentielle décroissante comme le montraient les graphiques ci-dessus. Avant de continuer rappelons (cf. chapitre de Physique Quantique Ondulatoire) que:

$$\int_{\infty} |\Psi|^2 dV = \int_{\infty} |\Psi|^2 dx dy dz = 1 \quad (53.191)$$

Or, en coordonnées sphériques (voir le début de ce chapitre):

$$dx dy dz = r^2 dr \sin \theta d\theta d\phi \quad (53.192)$$

Il vient alors comme nous l'avons vu plus haut:

$$\begin{aligned} \int_{\infty} |\Psi|^2 dx dy dz &= \int_{\infty} |\Psi(r, \theta, \phi)|^2 r^2 dr \sin \theta d\theta d\phi = \int_{\infty} |R(r)Y(\theta, \phi)|^2 r^2 dr \sin \theta d\theta d\phi \\ &= \int_{\infty} |R(r)|^2 |Y(\theta, \phi)|^2 r^2 dr \sin \theta d\theta d\phi = \int_{\infty} |R(r)|^2 r^2 dr \int_{\infty} |Y(\theta, \phi)|^2 \sin \theta d\theta d\phi = 1 \end{aligned} \quad (53.193)$$

Il s'ensuit que:

$$\begin{aligned} \int_{\infty} |R(r)|^2 r^2 dr &= 1 \\ \int_{\infty} |Y(\theta, \phi)|^2 \sin \theta d\theta d\phi &= 1 \end{aligned} \quad (53.194)$$

Avec ce résultat nous allons pouvoir calculer la probabilité radiale de trouver l'électron sur chaque orbitale atomique! Ainsi, il vient immédiatement avec le résultat précédent:

$$\frac{dP_r}{dr} = |R(r)|^2 r^2 \quad (53.195)$$

Et donc dans le cas de notre orbitale atomique 1s:

$$\frac{dP_r}{dr} = \left(\frac{2}{a_0^{3/2}} e^{-(r/a_0)}\right)^2 r^2 = \frac{4}{a_0^3} e^{-(2r/a_0)} r^2 \quad (53.196)$$

Il est maintenant super intéressant de calculer le point r où la probabilité de trouver l'électron est maximale sur l'orbitale 1s!

Pour cela, nous remarquons que $\frac{dP_r}{dr}$ passe par un maximum lorsque nous avons trivialement:

$$\frac{d^2 P_r}{dr^2} = 0 \quad (53.197)$$

soit:

$$\frac{d^2 P_r}{dr^2} = \frac{d}{dr} \left(\frac{4}{a_0^3} e^{-(2r/a_0)} r^2 \right) = \frac{4}{a_0^3} e^{-(2r/a_0)} \left(-\frac{2}{a_0} r^2 + 2r \right) = 0 \quad (53.198)$$

Soit:

$$r = a_0 \quad (53.199)$$

Ce qui est remarquable, car nous retrouvons le résultat du modèle de Bohr.

Pour résumer un peu tout cela, les états stationnaires de l'atome d'hydrogène sont spécifiés par trois nombres quantiques $n \in \mathbb{N}^*$, $l \leq n-1$, $|m_l| \leq l$ et la fonction d'onde de Schrödinger étant donnée au final par:

$$\Psi_{n,l,m_l}(r, \theta, \phi) = R_{n,l}(r) Y_{m_l,l}(\theta, \phi) \quad (53.200)$$

Nous avons alors la nomenclature traditionnelle suivante dans le cas de l'atome d'hydrogène:

n	l	m_l	fonction	nomenclature
1	0	0	$\Psi_{1,0,0}$	1s
2	0	0	$\Psi_{2,0,0}$	2s
	1	1	$\Psi_{2,1,1}$	2p ₁
		0	$\Psi_{2,1,0}$	2p ₀
		-1	$\Psi_{2,1,-1}$	2p ₋₁
3	0	0	$\Psi_{3,0,0}$	3s
	1	1	$\Psi_{3,1,1}$	3p ₁
		0	$\Psi_{3,1,0}$	3p ₀
		-1	$\Psi_{3,1,-1}$	3p ₋₁
	2	2	$\Psi_{3,2,2}$	3d ₂
		1	$\Psi_{3,2,1}$	3d ₁
		0	$\Psi_{3,2,0}$	3d ₀
		-1	$\Psi_{3,2,-1}$	3d ₋₁
		-2	$\Psi_{3,2,-2}$	3d ₋₂

Tableau: 53.1 - Nomenclature des couches et sous-couches de l'atome d'hydrogène

Nous pouvons inclure le spin de l'électron dans la description de la structure électronique de l'atome. Si nous traitons le spin comme un degré de liberté additionnel alors, l'absence de terme d'interaction entre les degrés de liberté classiques (positions dans l'espace réel) et le spin, interaction appelée "couplage spin-orbite", dans l'hamiltonien précédent, implique que nous pouvons écrire la fonction d'onde totale, spin inclus, sous la forme d'un produit:

$$\Psi_{n,l,m_l,m_s}(r,\theta,\phi) = \Psi_{n,l,m_l}(r,\theta,\phi) \varpi_{m_s} = R_{n,l}(r) Y_{m_l,l}(\theta,\phi) \varpi_{m_s} \quad (53.201)$$

où nous avons ajouté le nombre quantique de spin $m_s = \pm 1/2$ (cf. chapitre de Physique Quantique Corpusculaire).

La même remarque que nous avons faite dans le chapitre de Physique Quantique Corpusculaire s'applique dès lors: les niveaux restent $2s^2$ fois dégénérés.

3.2. PROFIL DE POTENTIEL

Revenons sur un point important qui est souvent utilisé dans les ouvrages de physique mais quasiment jamais démontré: le profil de potentiel quantique de l'atome hydrogénoïde. Beaucoup d'ouvrages parlent parfois de "**modèle harmonique de liaison atomique**" mais il semblerait que cela soit plutôt a priori un abus de langage.

Nous avons donc vu beaucoup plus haut dans le présent chapitre que:

$$\left[T_r + \frac{l(l+1)\hbar^2}{2\mu r^2} + E_{pot} \right] \Psi = E \Psi \quad (53.202)$$

Au vu de l'interprétation des trois termes de l'hamiltonien, il est d'usage de dire que les deux termes:

$$\frac{l(l+1)\hbar^2}{2\mu r^2} + E_{pot} \quad (53.203)$$

constituent "**l'énergie potentielle effective**" soit explicitement:

$$E_{pot,eff} = \underbrace{\frac{l(l+1)\hbar^2}{2\mu r^2}}_{\geq 0} + \underbrace{\left(-\frac{Zq_e^2}{4\pi\epsilon_0 \|\vec{r}_e - \vec{R}_N\|} \right)}_{< 0} \quad (53.204)$$

Donc le premier terme est (logiquement) répulsif alors que le deuxième est attractif. Un plot sous Maple 4.00b de l'énergie potentielle effective donne avec des valeurs expérimentales réelles concernant le rayon avec les valeurs réelles des constantes:

```
>plot([-2.31E-28/r+6.11E-39*0*(0+1)/r^2,-2.31E-28/r+6.11E-39*1*(1+1)/r^2,-2.31E-28/r+6.11E-39*2*(2+1)/r^2,-2.31E-28/r+6.11E-39*3*(3+1)/r^2,-2.31E-28/r+6.11E-39*10*(10+1)/r^2],r=5E-11..10E-10,y=-0.5E-17..0.5E-17,thickness=2);
```

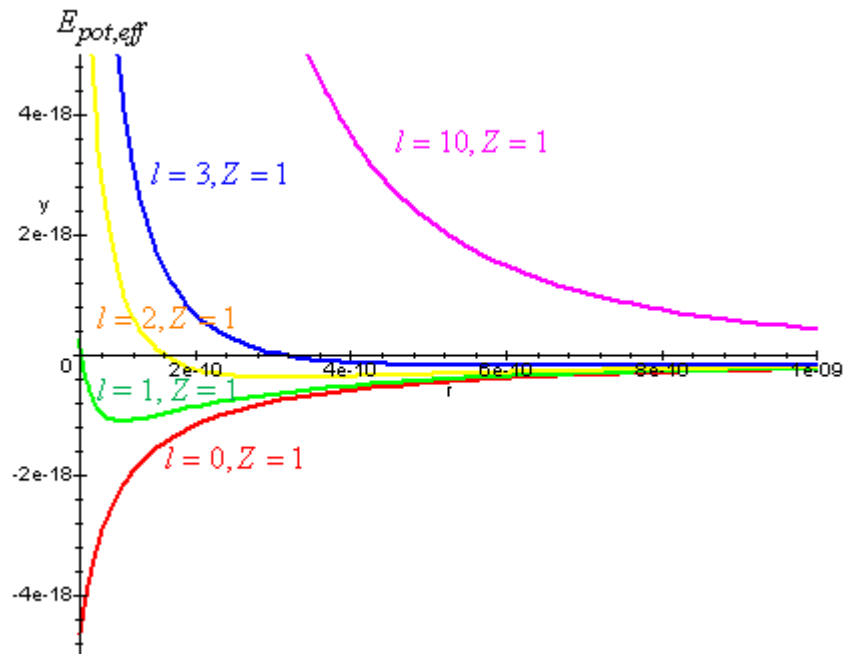
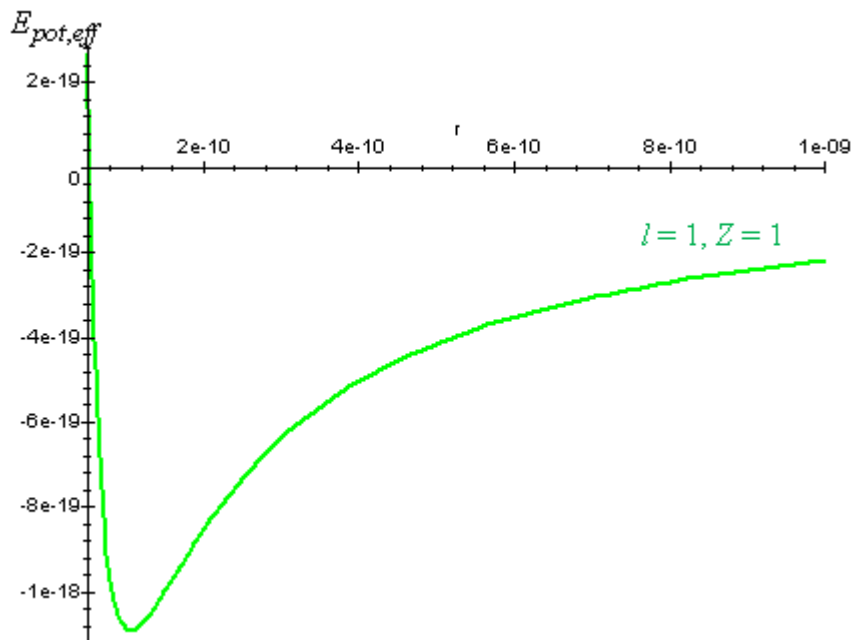



Figure: 53.12 - Plot de l'énergie potentielle effective

où les légendes ont été rajoutées après coup avec un logiciel de traitement de texte. Le lecteur y remarquera plus particulièrement le cas où l vaut 1 qui correspond au cas de figure indiqué par la majorité des ouvrages de physique de niveau licence. Soit en zoom:

```
>plot(-2.31E-28/r+6.11E-39*1*(1+1)/r^2,r=5E-11..10E-10,thickness=2, color=green);
```

Figure: 53.13 - Plot de l'énergie potentielle effective pour l valant 1

Le premier graphique nous indique aussi assez clairement que pour l valant 0 l'électron a une énergie potentielle négative qui le tient fermement dans l'orbite du proton. Par contre déjà à l valant 1 nous devinons que le point de stabilité de l'électron se trouve là où la dérivée s'annule. Au-delà de l valant 1, dans le cas d'un noyau avec un seul proton, l'électron n'est plus naturellement lié puisque son énergie potentielle tend à être positive. Le lecteur pourra d'ailleurs s'amuser avec Maple en faisant varier Z et l . Il verra que l'énergie potentielle effective est très sensible à ces deux paramètres. Pour l'exemple, le graphique ci-dessous montre l'énergie potentielle effective avec $l=4$ et $Z=1$ (donc atome instable) et ensuite avec $l=4$ et $Z=6$ (ce qui correspond plutôt à un état excité):

```
> plot([-2.31E-28*1/r+6.11E-39*4*(4+1)/r^2,-2.31E-28*6/r+6.11E-39*4*(4+1)/r^2]
,r=5E-11..10E-10,thickness=2);
```

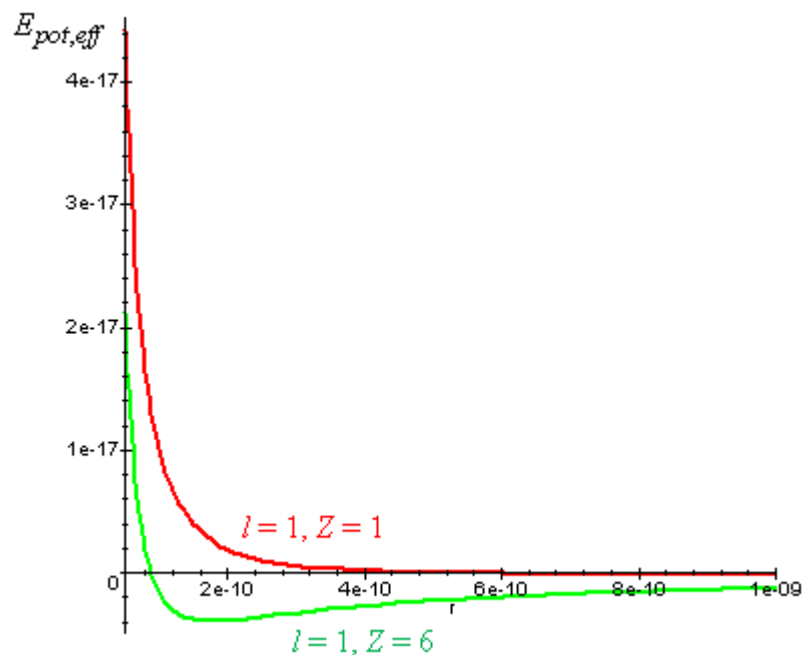


Figure: 53.14 - Plot de l'énergie potentielle effective pour l valant 1

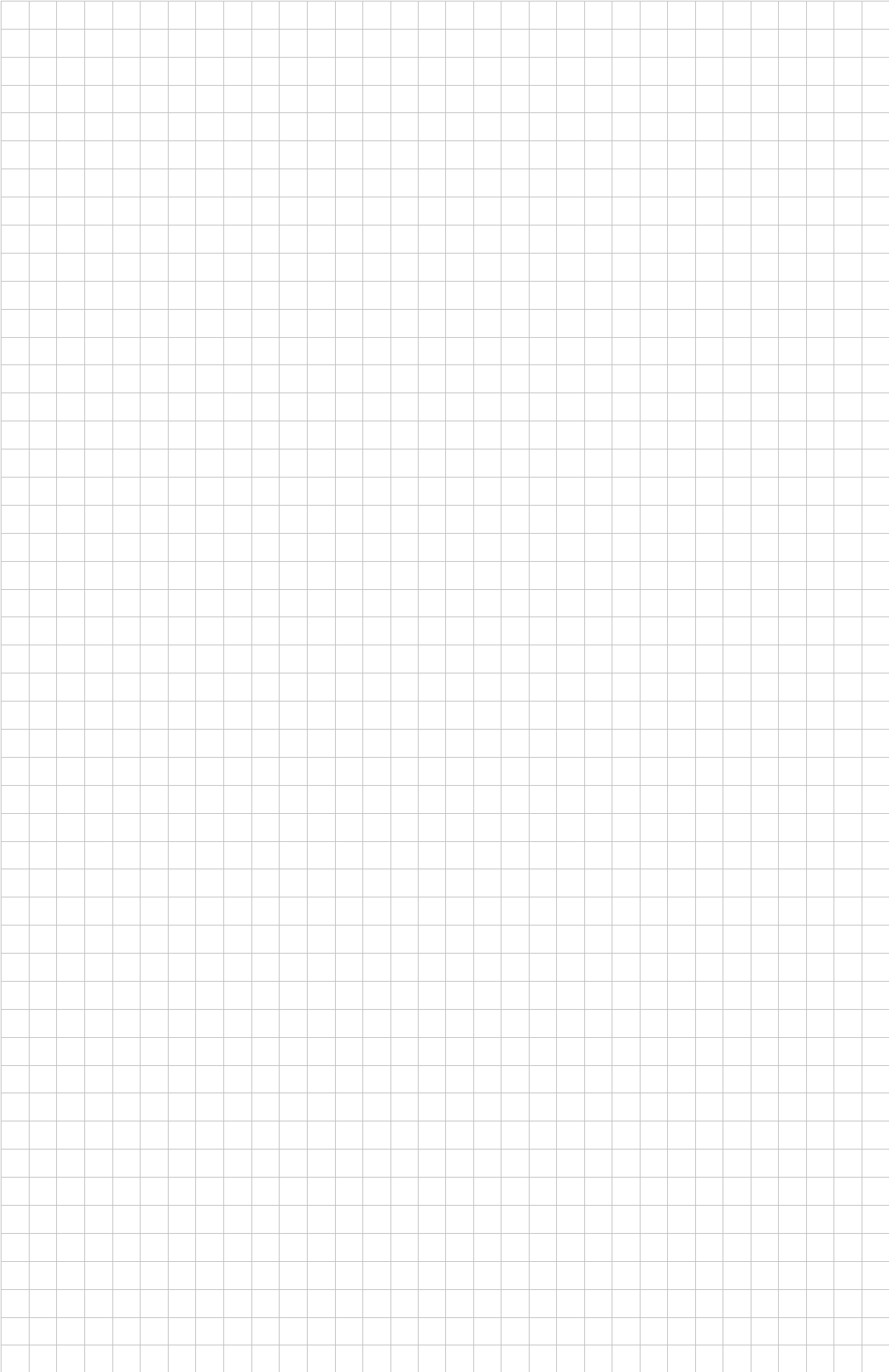
Il est d'usage dans la pratique considérer que:

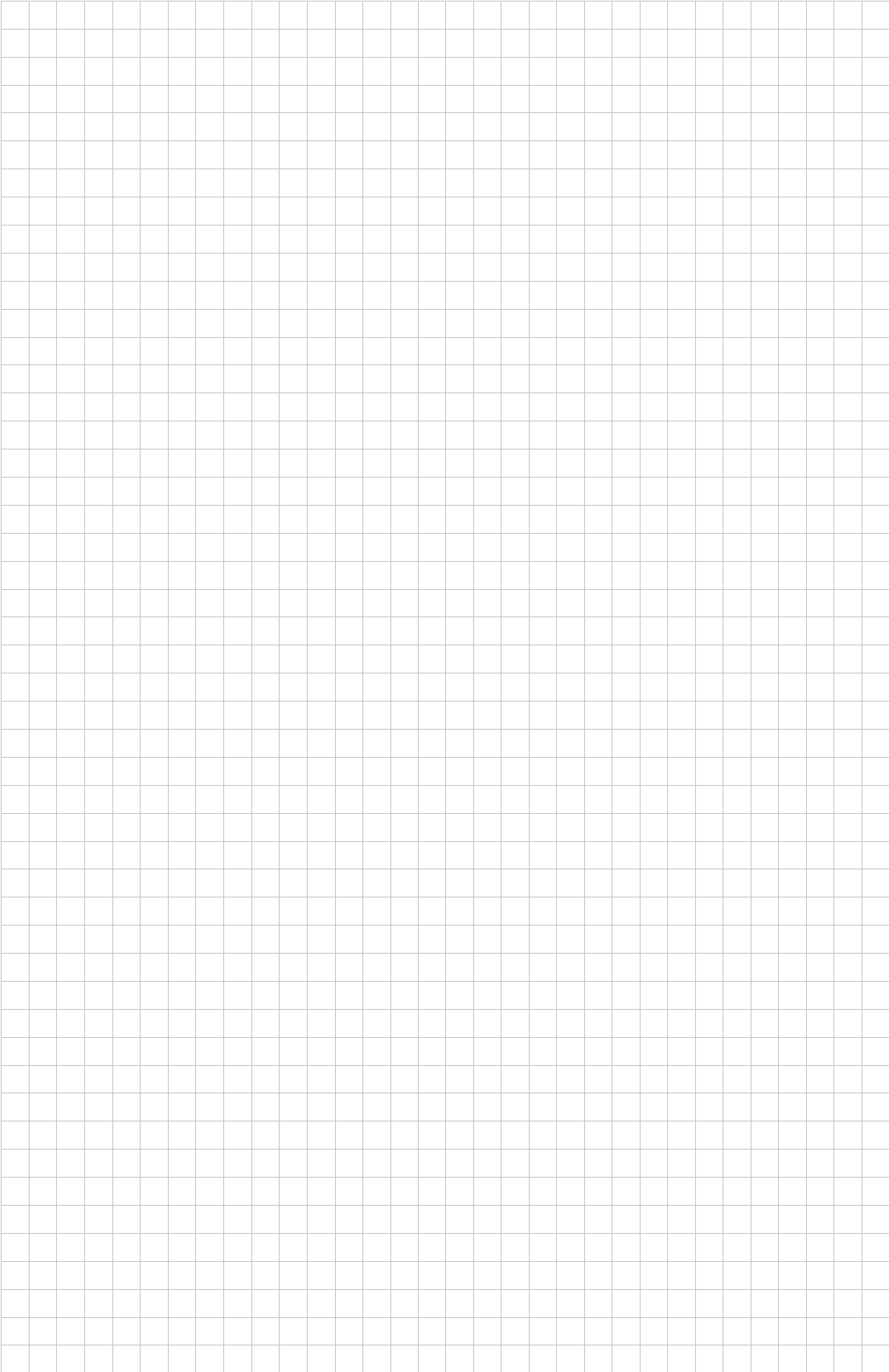
$$E_{pot,eff} = \frac{l(l+1)\hbar^2}{2\mu r^2} - \frac{Zq_e^2}{4\pi\epsilon_0 \|\vec{r}_e - \vec{R}_N\|} \quad (53.205)$$

est à un facteur près (facteur de charge électrique) un "**potentiel électrique effectif**" ou "**potentiel électrique écranté**". Effectivement par la définition du potentiel électrique (cf. chapitre d'Électrostatique), il y a seulement un rapport de facteur de charge entre l'énergie potentielle électrique et le potentiel électrique. Nous avons donc:

$$U_{eff} = \frac{1}{Zq} E_{pot,eff} \Rightarrow U_{eff} \propto E_{pot,eff} \quad (53.206)$$







54. CHIMIE MOLÉCULAIRE

La chimie moléculaire est le domaine central qui interconnecte par l'étude des molécules de nombreuses techniques de pointe prometteuses du début 21ème siècle qui sont pour ne citer que les plus connues: la biologie moléculaire, les matériaux moléculaires, l'électronique moléculaire, les polymères...

Sachant qu'il a été découvert expérimentalement qu'une même molécule peut avoir plusieurs fonctions très différentes, son étude théorique permet de bien les utiliser (avec parfois un meilleur rendement en termes de R&D) dans ses domaines d'application. Le lecteur aura donc compris que, comme à l'habitude sur ce site, nous allons nous concentrer ici uniquement sur l'aspect théorique (mathématique) de la chimie moléculaire même si nous allons nous limiter uniquement aux développements théoriques effectués entre 1910 et environ 1935 (au-delà la complexité des théories nécessite trop de pages pour un site Internet généraliste).

Nous ne sommes à notre époque qu'aux balbutiements de la découverte de ce que la nature a fait avec beaucoup de temps et de hasard: c'est-à-dire des molécules complexes fonctionnant comme des nanomachines capables localement (site actif) de filtrer, d'oxyder, de faire de catalyses... et bien d'autres manipulations (y a qu'à observer votre corps!).

Une molécule est souvent traitée dans les classes d'école avec l'équation de Schrödinger (donc pas de cas relativistes et pas de prise en compte des spins) sous la forme habituelle (cf. [chapitre de Physique Quantique Ondulatoire](#)):

$$H\Psi = E\Psi \quad (54.1)$$

soit sous une forme également stationnaire (indépendante du temps) où pour rappel Ψ est une fonction propre et E une valeur propre de l'application H .

En réalité, les fonctions d'ondes sont impossibles à calculer formellement avec les outils mathématiques contemporains et la seule chose que nous savons faire sont des calculs numériques (méthode des perturbations). C'est la raison pour laquelle certains centres de chimie se transforment avec le temps en des centres de calculs où le caractère prédictif (et peu coûteux) de la chimie quantique s'affirme de plus en plus.

Il reste cependant bien évidemment indispensable, comme toujours, de comprendre comment les modèles théoriques sont construits et leurs hypothèses sous-jacentes.

Mais nous pouvons quand même avec les calculs prédire la forme de molécules de taille raisonnable, l'énergie de leurs liaisons internes, leur capacité énergétique sous l'effort d'une déformation, la forme des orbitales moléculaires (O.M.), l'énergie de transitions d'états (lorsque des parties de la molécule se déplacent dans celle-ci), leur réactivité vis-à-vis d'un milieu réactionnel...

Nous distinguons couramment deux cas d'études de la chimie moléculaire:

- Mécanique quantique: toutes les interactions entre particules sont prises en compte sous l'hypothèse de quelques simplifications acceptables.

- Mécanique moléculaire: Pour de grosses molécules, nous ne nous occupons plus du problème électronique, mais des interactions de certains paramètres sur lesquels nous voulons nous concentrer.

Par exemple, l'hémoglobine (protéine transportant l'oxygène dans les muscles) est une énorme structure moléculaire dont on ne va étudier que le site actif avec les outils de la mécanique quantique. Le comportement global de la molécule elle-même sera traité avec les outils de la mécanique moléculaire.

Il s'ensuit qu'excepté pour les atomes hydrogénoïdes, nous ne pouvons pas décrire analytiquement une molécule d'un point de vue purement quantique! Toutes les méthodes quantiques actuelles reposent sur une ou plusieurs approximations. Les fonctions d'onde ne sont donc qu'approchées et le niveau de calcul est adapté en fonction de ce que nous voulons montrer et de la précision que nous recherchons (en cherchant à minimiser le temps de calcul pour des problèmes de coûts...). La bonne compréhension des approximations permet de formuler des modèles simples ne nécessitant qu'un minimum de calculs (souvent triviaux).

Nous nous proposons ici de montrer deux modèles courants (et les plus simples):

1. APPROXIMATION ORBITALAIRE

Une molécule est bien évidemment un problème extrêmement complexe: N noyaux, n électrons et le tout est en mouvement!

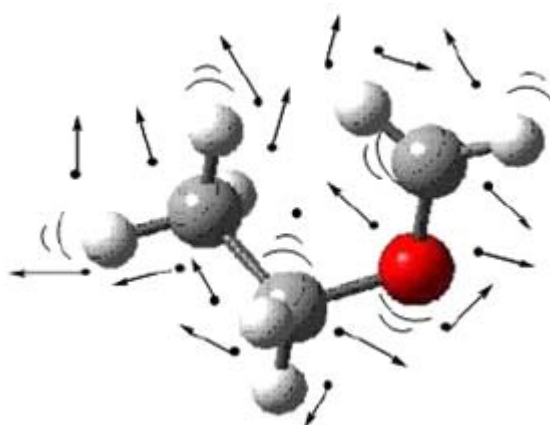


Figure: 54.1 - Exemple de molécule où un peu tout est en mouvement

L'hamiltonien (cf. chapitre de Physique Quantique Ondulatoire):

$$H = E_{cin} + E_{pot} = -\frac{\hbar^2}{2m} \cdot \vec{\nabla}^2 + V(q_1, q_2, \dots, t) = \sum_{k=1}^n \left(-\frac{\hbar^2}{2m_k} \cdot \frac{\partial^2}{\partial q_k^2} \right) + V(q_1, q_2, \dots, t) \quad (54.2)$$

est alors de la forme cauchemardesque mais intuitive (le G en indice de l'hamiltonien signifie "Général") suivante:

$$H_G = \underbrace{\left(-\sum_{k=1}^N \frac{\hbar^2}{2M_k} \vec{\nabla}_k^2 - \sum_{i=1}^n \frac{\hbar^2}{2m_e} \vec{\nabla}_i^2 \right)}_{E_c} + \underbrace{\left(-\sum_{k=1}^N \sum_{i=1}^n \frac{Ze^2}{4\pi\epsilon_0 r_{ik}} + \sum_{\substack{i=1 \\ j>i}}^{n-1} \frac{e^2}{4\pi\epsilon_0 r_{ij}} + \sum_{\substack{k=1 \\ i>k}}^{N-1} \frac{Z_k Z_i e^2}{4\pi\epsilon_0 r_{ki}} \right)}_{E_{pot}} \quad (54.3)$$

où:

1. $-\sum_{k=1}^N \frac{\hbar^2}{2M_k} \nabla_k^2$ est l'énergie cinétique des k noyaux de masse M_k de la molécule.

2. $-\sum_{i=1}^n \frac{\hbar^2}{2m_e} \nabla_i^2$ est l'énergie cinétique des n électrons de masse m_e

3. $-\sum_{k=1}^N \sum_{i=1}^n \frac{Ze^2}{4\pi\epsilon_0 r_{ik}}$ est l'énergie potentielle due à l'attraction électron(-)/noyaux(+)

4. $\sum_{i=1}^{n-1} \sum_{j>i} \frac{e^2}{4\pi\epsilon_0 r_{ij}}$ est l'énergie potentielle de répulsion électron(-)/électron(-)

5. $\sum_{k=1}^{N-1} \sum_{i>k} \frac{Z_k Z_i e^2}{4\pi\epsilon_0 r_{ki}}$ est l'énergie potentielle de répulsion noyau(+)/noyau(+)

Souvent nous retrouvons ces termes sous la forme suivante de l'équation de Schrödinger dans la littérature:

$$\left[(T_e + T_n) + (V_{ee} + V_{en} + V_{nn}) \right] \Psi_n \Psi_{el} = E \Psi_n \Psi_{el} \quad (54.4)$$

Une première approximation que nous pourrions tenter c'est de découpler le mouvement des noyaux de celui des électrons. Effectivement, comme le noyau est beaucoup plus massif (environ 2000 fois) que le cortège des électrons, le centre de masse est assimilé au noyau de l'atome et le mouvement relatif à l'ensemble du cortège électronique. Cette approche approximative est très connue sous la dénomination "**d'approximation de Born-Oppenheimer**":

$$H_{el} = -\sum_{i=1}^n \frac{\hbar^2}{2m_e} \nabla_i^2 - \sum_{k=1}^N \sum_{i=1}^n \frac{Ze^2}{4\pi\epsilon_0 r_{ik}} + \sum_{i=1}^{n-1} \sum_{j>i} \frac{e^2}{4\pi\epsilon_0 r_{ij}} = -\sum_{i=1}^n \frac{\hbar^2}{2m_e} \nabla_i^2 + \frac{e^2}{4\pi\epsilon_0} \left(-\sum_{k=1}^N \sum_{i=1}^n \frac{Z}{r_{ik}} + \sum_{i=1}^{n-1} \sum_{j>i} \frac{1}{r_{ij}} \right) \quad (54.5)$$

qui nous permet alors d'étudier les orbitales moléculaires. Mais malheureusement cette approximation n'est pas suffisante à cause du terme de répulsion interélectronique (la double somme) qui empêche d'utiliser la technique de séparation des variables comme nous l'avons fait dans le chapitre de Chimie Quantique avec l'atome d'hydrogène.

Par ailleurs, cette dernière équation s'écrit également sous la forme de la première ligne du couple d'équation ci-dessous (équation de Schrödinger des électrons et des noyaux):

$$\begin{aligned} \underbrace{(T_e + V_{ee} + V_{en})}_{H_{el}} \Psi_{el} &= E \Psi_{el} \\ \underbrace{(T_n + V_{nn})}_{H_{noy}} \Psi_n \Psi_{el} &= E \Psi_n \Psi_{el} \end{aligned} \quad (54.6)$$

Ce système d'équations constitue ce que certains appellent "l'approximation adiabatique" (???).

L'idée qui vient alors à l'esprit va être d'utiliser la propriété suivante:

Soit deux opérateurs A et B , $f(u)$ et $g(v)$ leurs fonctions propres respectives associées aux valeurs propres a et b . Alors $f(u)g(v)$ est fonction propre de l'opérateur somme $A+B$ avec la valeur propre associée $a+b$.

Ce qui s'écrit:

$$\left. \begin{array}{l} Af(u) = af(u) \\ Bg(v) = bg(v) \end{array} \right\} \Rightarrow (A+B)f(u)g(v) = (a+b)f(u)g(v) \quad (54.7)$$

Démonstration:

$$\begin{aligned} (A+B)f(u)g(v) &= [Af(u)]g(v) + [Bf(u)]g(v) \\ &= [af(u)]g(v) + [bf(u)]g(v) = (a+b)f(u)g(v) \end{aligned} \quad (54.8)$$

C.Q.F.D.

Et c'est ce que nous allons faire pour décomposer l'hamiltonien n -électronique H_{el} en somme d'hamiltoniens monoélectroniques indépendants en sachant de ce qui précède que si nous trouvons la fonction propre pour chacun (ce qui est relativement plus facile) il suffira de les multiplier pour avoir la fonction propre globale.

Ainsi, nous écrivons:

$$H_{el} = H_1 + H_2 + \dots + H_i + \dots + H_n \quad (54.9)$$

et donc nous devons trouver pour chaque i :

$$H_i \Psi_i = E_i \Psi_i \quad (54.10)$$

Pour ensuite avoir:

$$H_{el} \Psi_{el} = \sum_{i=1}^n H_i \prod_{i=1}^n \Psi_i = \sum_{i=1}^n E_i \prod_{i=1}^n \Psi_i = E_{el} \Psi_{el} \quad (54.11)$$

avec donc:

$$\Psi_{el} = \Psi_1 \cdot \Psi_2 \cdot \dots \cdot \Psi_i \cdot \dots \cdot \Psi_n \quad (54.12)$$

Cette approche par hamiltoniens monoélectroniques va nous amener à remplacer:

$$H_{el} = -\sum_{i=1}^n \frac{\hbar^2}{2m_e} \nabla_i^2 - \sum_{k=1}^N \sum_{i=1}^n \frac{Ze^2}{4\pi\epsilon_0 r_{ik}} + \sum_{i=1}^{n-1} \sum_{j>i} \frac{e^2}{4\pi\epsilon_0 r_{ij}} \quad (54.13)$$

par la somme des hamiltoniens pour un électron appelés "hamiltoniens effectifs":

$$H_i = -\frac{\hbar^2}{2m_e} \nabla_i^2 - \frac{Ze^2}{4\pi\epsilon_0 r_i} + \frac{e^2}{4\pi\epsilon_0 r_i} = -\frac{\hbar^2}{2m_e} \nabla_i^2 - \frac{e^2(Z-1)}{4\pi\epsilon_0 r_i} = -\frac{\hbar^2}{2m_e} \nabla_i^2 - \frac{e^2 Z^*}{4\pi\epsilon_0 r_i} \quad (54.14)$$

Cette méthode d'approximation est parfois appelée en chimie théorique "**approximation des électrons indépendants**" ou encore "**approximation orbitale**". Elle consiste donc à globaliser les interactions individuelles électron-électron et à écrire que chaque électron évolue dans un potentiel moyen résultant de la présence de l'ensemble des autres électrons.

La "**méthode de Slater**" consiste par définition à écrire cette dernière relation sous la forme:

$$H_i \cong -\frac{\hbar^2}{2m_e} \nabla_i^2 - \frac{e^2(Z-\sigma)}{4\pi\epsilon_0 r_i} \quad (54.15)$$

où σ est appelée "**constante d'écran**".

Cette méthode de Slater revient en gros à remplacer les termes purement électroniques par une constante. Elle peut être considérée comme une méthode paramétrique car les constantes ont été déterminées de façon purement expérimentale.

Le principe de calcul empirique de la constante d'écran est relativement simple. Dans un atome poly-électronique, les électrons de coeur sont sur des orbites beaucoup plus contractées alors que les électrons de valence qui vont être responsables des propriétés chimiques de l'atome en question sont sur des orbites beaucoup plus relâchées.

L'attraction exercée par le noyau sur ces électrons est beaucoup plus faible que celle exercée sur les électrons de coeur et ces électrons ne perçoivent qu'une partie de la charge atomique.

Slater a alors proposé que cette charge effective, qui est le plus souvent notée Z^* pouvait être calculée en tenant compte de la constante d'écran. Cette constante représente donc l'effet moyen des autres électrons sur l'électron considéré de l'hamiltonien effectif i .

$$Z^* = Z - \sigma \quad (54.16)$$

Pour un électron périphérique, nous devons donc considérer que sa constante d'écran est due à tous les électrons placés sur des orbites égales ou inférieures. La tradition (ou plutôt le "truc") veut que le calcul se fasse en regroupant les orbitales atomiques en plusieurs groupes 1s/2s, 2p/3s, 3p/3d/4s, 4p/4d/4f/5s, 5p/etc...

Ensuite le calcul est simple car il se fait sur la base d'un tableau de valeurs prédéfinies et il suffit d'additionner les contributions d'écran de tous les électrons selon le tableau ci-dessous:

	$n' < n-1$	$n' = n-1$	$n' = n$	$n' > n$
1s			0.30	0
ns, np	1	0.85	0.35	0
nd, nf	1	1	0.35	0

Tableau: 7.1 - Contributions d'écran des électrons

Ce tableau mérite quelques explications bien évidemment!:

L'indice indique le nombre du groupe qui contribue à la constante d'écran alors que n est le nombre du groupe de l'électron que nous considérons.

Exemple:

Dans le cas du Carbone de configuration $1s^2 2s^2 2p^2$, la charge nucléaire est $Z = 6$. Un électron $1s$ n'est écrané que par l'autre électron $1s$, la charge effective qu'il voit est donc de:

$$Z^* = Z - \sigma = 6 - 0.30 = 5.70 \quad (54.17)$$

Un électron $2s$ ou $2p$ est écrané par les deux électrons $1s$ et par les 3 autres électrons $2s$ et $2p$. La charge effective par laquelle il est attiré est alors:

$$Z^* = Z - \sigma = 6 - (2 \cdot 0.85 + 3 \cdot 0.35) = 3.25 \quad (54.18)$$

Nous voyons donc bien que la charge effective ressentie décroît assez rapidement!

2. MÉTHODE C.L.O.A.

Cette méthode, plutôt qualitative, consiste à considérer que la fonction d'onde moléculaire est une "**Combinaison Linéaire des Orbitales Atomiques**" (C.L.O.A.) contrairement à la méthode précédente où nous multiplions les hamiltoniens effectifs.

Cette méthode est importante car elle est à la base d'une grande partie du vocabulaire courant des chimistes dès que la chimie effectuée est assez pointue!

Basons-nous sur l'exemple de la molécule de dihydrogène H_2 . L'idée est alors la suivante:

Si nous avons la fonction de l'orbitale atomique $1s_A$ de H_A et respectivement la fonction $1s_B$ de H_B , alors nous posons que l'orbitale moléculaire dicentrique (liée à deux atomes) correspondante est donnée par:

$$\Psi_{AB} = a\Psi_A + b\Psi_B \quad (54.19)$$

qui définit un système quantique à deux états propres.

Mais comme nous le savons bien, en réalité, seul le carré de la fonction d'onde possède un sens physique (probabilité de présence). Ainsi, si nous supposons que la fonction d'onde n'est pas à valeur dans $\mathbb{C}$, nous avons pour l'unique électron qui nous intéresse ($1s$):

$$\Psi_{AB}^2 = (a\Psi_A + b\Psi_B)^2 = a^2\Psi_A^2 + 2ab\Psi_A\Psi_B + b^2\Psi_B^2 \quad (54.20)$$

où nous admettons que:

- $a^2\Psi_A^2$ représente la probabilité de présence d'être près de A
- $b^2\Psi_B^2$ représente la probabilité de présence d'être près de B
- $2ab\Psi_A\Psi_B$ représente la probabilité de présence de l'électron qui joue la liaison A-B

Dans le cas particulier de la molécule diatomique symétrique que nous avons choisie comme exemple, les atomes A et B jouent le même rôle et il n'y a donc pas de raison que l'électron soit plus près de A que de B ou inversement.

Ainsi, la probabilité de trouver l'électron près de A est donc égale à la probabilité de le trouver près de B .

$$\alpha^2 \Psi_A^2 = b^2 \Psi_B^2 \quad (54.21)$$

De plus dans ce cas les orbitales Ψ_A et Ψ_B sont totalement identiques (orbitales 1s toutes deux du même atome) et il n'y a donc pas lieu de les distinguer. Nous avons donc:

$$\alpha^2 = b^2 \Rightarrow b = \pm \alpha \quad (54.22)$$

Nous avons donc deux solutions pour Ψ_{AB} qui sont (ces deux solutions peuvent se trouver sous des notations très variées dans la littérature):

$$\Psi_{AB}^1 = \alpha \Psi_A + \alpha \Psi_B = \alpha (\Psi_A + \Psi_B) \quad (54.23)$$

si $b = +\alpha$ et:

$$\Psi_{AB}^2 = \alpha \Psi_A - \alpha \Psi_B = \alpha (\Psi_A - \Psi_B) \quad (54.24)$$

si $b = -\alpha$.

Remarque: Attention! Nous ne pouvons pas poser pour les deux dernières relations que $\Psi_A = \Psi_B$. Cette dernière égalité n'a lieu en tout point que si la distance qui sépare les deux noyaux est nulle (ce qui est peu probable) ou, s'ils sont distants d'une certaine valeur D située au milieu de celle-ci.

Ces deux expressions seront simultanément solutions de l'équation de Schrödinger. Nous obtenons donc deux orbitales moléculaires à partir des deux orbitales atomiques dans le cas de la molécule diatomique symétrique.

La fonction:

$$\Psi_A + \Psi_B \quad (54.25)$$

est appelée "**fonction liante**" car elle correspond à un renforcement de la probabilité de présence de l'électron entre les atomes A et B ce qui correspond à la création de la liaison!

Inversement, la fonction:

$$\Psi_A - \Psi_B \quad (54.26)$$

est appelée "**fonction antiliante**" car elle correspond à une diminution de la probabilité de présence de l'électron entre les atomes A et B ce qui correspond à la destruction de la liaison!

En définitive, en se recouvrant, les deux orbitales atomiques de même énergie donnent naissance à deux orbitales moléculaires d'énergies différentes, l'une liante stabilisée et l'autre antiliante déstabilisée.

Nous avons évidemment au vu de ce qui précède que, dans des cas plus complexes, le niveau d'énergie de l'orbitale moléculaire liante est plus petit que celle de l'antiliante (nous allons le démontrer rigoureusement en détails plus bas).

Ainsi, il faut respectivement plus d'énergie pour ioniser l'électron de l'orbitale liante σ qu'il n'en faut pour ioniser l'électron de l'orbitale antiliante σ^* . Donc il est communément admis de dire que l'énergie de la liaison liante est plus forte que l'antiliante (mais nous en ferons la démonstration juste un peu plus loin).

Indiquons aussi qu'en chimie, une liaison chimique dans laquelle chacun des atomes liés met en commun un électron d'une de ses couches externes afin de former un doublet d'électrons liant les deux atomes est communément appelée "**liaison covalente**".

Les chimistes disent alors que la liaison covalente implique le partage équitable d'une seule paire d'électrons, appelé "**doublet liant**" (mais dont au fait qu'un seul des électrons l'est vraiment). Chaque atome fournissant un électron, la paire d'électrons est alors délocalisée entre les deux atomes comme nous l'avons démontré.

Ce sont les raisons qui font que nous disons couramment que la liaison σ est une liaison chimique covalente créée entre deux atomes par recouvrement axial d'orbitales.

Creusons maintenant cette approche! Les orbitales moléculaires doivent être normées. Ce qui signifie que:

$$\int_{-\infty}^{+\infty} |\Psi|^2 dx dy dz = 1 \quad (54.27)$$

Ce qui donne, puisque les orbitales atomiques sont normées pour Ψ_1 et sont des fonctions réelles:

$$\int_{-\infty}^{+\infty} |\Psi_{AB}^1|^2 dx dy dz = \int_{-\infty}^{+\infty} \left| a(\Psi_A + \Psi_B) \right|^2 dx dy dz = \int_{-\infty}^{+\infty} (a^2 \Psi_A^2 + 2a^2 \Psi_A \Psi_B + a^2 \Psi_B^2) dx dy dz = 1 \quad (54.28)$$

Puisque a (nombre réel dans notre cas particulier) est imposé comme étant une constante, il vient immédiatement:

$$\begin{aligned} a^2 \left(\int_{-\infty}^{+\infty} \Psi_A^2 dx dy dz + \int_{-\infty}^{+\infty} \Psi_B^2 dx dy dz + 2 \int_{-\infty}^{+\infty} \Psi_A \Psi_B dx dy dz \right) &= a^2 \left(2 + 2 \int_{-\infty}^{+\infty} \Psi_A \Psi_B dx dy dz \right) \\ &= a^2 (2 + 2S_{12}) = 1 \end{aligned} \quad (54.29)$$

D'où pour Ψ_{AB}^1 :

$$a = \frac{1}{\sqrt{2(1 + S_{12})}} \quad (54.30)$$

De manière identique, nous avons pour Ψ_{AB}^2 :

$$\alpha = \frac{1}{\sqrt{2(1-S_{12})}} \quad (54.31)$$

Si nous avons $S_{12} \ll 1$, il vient la forme suivante que nous trouvons dans la littérature:

$$\begin{aligned} \Psi_{AB}^1 &\cong \frac{1}{\sqrt{2}} (\Psi_A + \Psi_B) \\ \Psi_{AB}^2 &\cong \frac{1}{\sqrt{2}} (\Psi_A - \Psi_B) \end{aligned} \quad (54.32)$$

Faisons un petit exemple en utilisant comme orbitale, l'orbitale atomique la plus basse (1s) de l'atome d'hydrogène dans le cas d'une liaison dihydrogène H_2 pour laquelle nous avons démontré à la fin du chapitre de Chimie Quantique que:

$$\Psi_{1,0,0} = R_{1,0} Y_{0,0} = \frac{2}{a_0^{3/2}} e^{-\frac{r}{a_0}} \frac{1}{\sqrt{4\pi}} = \frac{1}{\sqrt{\pi}} \frac{1}{a_0^{3/2}} e^{-\frac{r}{a_0}} \quad (54.33)$$

Il vient alors que:

$$\Psi_A = \frac{1}{\sqrt{\pi}} \frac{1}{a_0^{3/2}} e^{-\frac{r_1}{a_0}} = \lambda e^{-\frac{r_1}{a_0}} \text{ et } \Psi_B = \frac{1}{\sqrt{\pi}} \frac{1}{a_0^{3/2}} e^{-\frac{r_2}{a_0}} = \lambda e^{-\frac{r_2}{a_0}} \quad (54.34)$$

avec pour rappel:

$$a_0 = 4\pi\epsilon_0 \frac{\hbar^2}{\mu e^2} \quad (54.35)$$

Il vient alors pour l'orbitale moléculaire liante de niveau s:

$$\sigma_s = \frac{\lambda}{\sqrt{2}} \left(e^{-\frac{r_1}{a_0}} + e^{-\frac{r_2}{a_0}} \right) \quad (54.36)$$

et pour l'antiliante de niveau s:

$$\sigma_s^* = \frac{\lambda}{\sqrt{2}} \left(e^{-\frac{r_1}{a_0}} - e^{-\frac{r_2}{a_0}} \right) \quad (54.37)$$

Nous voyons alors immédiatement que σ_s^* s'annule au milieu des deux protons car à cet endroit $r_1 = r_2$. L'orbitale moléculaire antiliante présente donc un plan nodal et les électrons sont essentiellement localisés sur les protons.

Par contre, pour l'orbitale moléculaire σ_s , la densité ne s'annule pas. Nous comprenons alors aisément qu'un électron de σ_s assure la stabilité de la molécule et qu'il est donc responsable de la liaison chimique.

Nous en déduisons donc que la stabilisation électronique due à l'interaction de deux orbitales identiques est proportionnelle à leur recouvrement. Plus le recouvrement est grand, plus la stabilisation est importante.

Il y a une approche encore plus technique utilisant la notation de Dirac et qui a l'avantage de permettre la détermination des valeurs propres de l'énergie.

D'abord nous écrivons l'expression générale de l'équation indépendante du temps de Schrödinger avec la notation Ket-Bra pour une orbitale moléculaire, superposition de deux orbitales atomiques:

$$H|\Psi_{AB}\rangle = E|\Psi_{AB}\rangle \quad (54.38)$$

Soit sous forme explicite:

$$H|a\Psi_A + b\Psi_B\rangle = E|a\Psi_A + b\Psi_B\rangle \quad (54.39)$$

Si nous multiplions par le bra $\langle\Psi_A|$ à gauche et en tenant compte que a, b ainsi que les valeurs propres de l'énergie sont des constantes, nous obtenons l'équation suivante:

$$a\langle\Psi_A|H|\Psi_A\rangle + b\langle\Psi_A|H|\Psi_B\rangle = aE\langle\Psi_A|\Psi_A\rangle + bE\langle\Psi_A|\Psi_B\rangle \quad (54.40)$$

De même, nous obtenons avec le bra $\langle\Psi_B|$:

$$a\langle\Psi_B|H|\Psi_A\rangle + b\langle\Psi_B|H|\Psi_B\rangle = aE\langle\Psi_B|\Psi_A\rangle + bE\langle\Psi_B|\Psi_B\rangle \quad (54.41)$$

Simplifions encore les écritures:

$$\begin{aligned} aH_{AA} + bH_{AB} &= aES_{AA} + bES_{AB} \\ aH_{BA} + bH_{BB} &= aES_{BA} + bES_{BB} \end{aligned} \quad (54.42)$$

Par symétrie du problème dans le cas du dihydrogène, nous posons:

$$\langle\Psi_A|H|\Psi_B\rangle = H_{AB} = \langle\Psi_B|H|\Psi_A\rangle = H_{BA} \quad (54.43)$$

qui sont appelées "**intégrales de résonance**" car il s'agit d'un terme relatif à la combinaison (résonance) des deux orbitales atomiques propres aux deux atomes de base de la structure moléculaire.

Nous avons également:

$$\begin{aligned} \langle\Psi_A|H|\Psi_A\rangle &= H_{AA} \\ \langle\Psi_B|H|\Psi_B\rangle &= H_{BB} \end{aligned} \quad (54.44)$$

qui sont appelées "**intégrales coulombiennes**" car elles correspondent selon le cinquième postulat de la

physique quantique ondulatoire (voir chapitre du même nom) à la valeur moyenne de l'énergie totale de l'électron.

Nous avons évidemment aussi:

$$\langle \Psi_A | \Psi_A \rangle = S_{AA} = \langle \Psi_B | \Psi_B \rangle = S_{BB} = 1 \quad (54.45)$$

qui sont appelées "**intégrales de recouvrement**" car les deux orbitales atomiques de même type de chacun des atomes se recouvrent.

Et enfin, nous avons toujours par symétrie de notre cas particulier:

$$\langle \Psi_B | \Psi_A \rangle = S_{BA} = \langle \Psi_A | \Psi_B \rangle = S_{AB} \quad (54.46)$$

Nous pouvons alors écrire, puisque les intégrales de recouvrement sont unitaires:

$$\begin{aligned} \alpha(H_{AA} - E) + b(H_{AB} - ES_{AB}) &= 0 \\ \alpha(H_{BA} - ES_{BA}) + b(H_{BB} - E) &= 0 \end{aligned} \quad (54.47)$$

Ces deux équations sont appelées "**équations séculaires**". La solution triviale, mais a priori non physique car elle signifierait que l'électron aurait une densité de probabilité nulle en tout point de l'espace correspond à $\alpha = b = 0$.

Il existe une solution non triviale, si et seulement si, le déterminant (cf. chapitre d'Algèbre Linéaire), appelé dans la chimie moléculaire "**déterminant séculaire**", suivant est nul:

$$\begin{vmatrix} H_{AA} - E & H_{AB} - ES_{AB} \\ H_{BA} - ES_{BA} & H_{BB} - E \end{vmatrix} = (H_{AA} - E)(H_{BB} - E) - (H_{AB} - ES_{AB})(H_{BA} - ES_{BA}) = 0 \quad (54.48)$$

Comme nous avons par symétrie de notre cas particulier:

$$H_{AA} = H_{BB}, H_{AB} = H_{BA}, S_{AB} = S_{BA} \quad (54.49)$$

Il vient alors:

$$(H_{AA} - E)^2 - (H_{AB} - ES_{AB})^2 = 0 \quad (54.50)$$

D'où:

$$(H_{AA} - E) = \pm(H_{AB} - ES_{AB}) \quad (54.51)$$

Ce qui nous donne deux solutions (+):

$$-E_1 + E_1 S_{AB} = H_{AB} - H_{AA} \Rightarrow E_1(1 - S_{AB}) = H_{AA} - H_{AB} \Rightarrow E_1 = \frac{H_{AA} - H_{AB}}{1 - S_{AB}} \quad (54.52)$$

et (-):

$$-E_2 - E_2 S_{AB} = -H_{AB} - H_{AA} \Rightarrow -E_2(1 + S_{AB}) = -(H_{AA} + H_{AB}) \Rightarrow E_2 = \frac{H_{AA} + H_{AB}}{1 + S_{AB}} \quad (54.53)$$

Nous avons donc:

$$E_1 = \frac{H_{AA} - H_{AB}}{1 - S_{AB}} \text{ et } E_2 = \frac{H_{AA} + H_{AB}}{1 + S_{AB}} \quad (54.54)$$

Mais pour pouvoir faire le calcul des niveaux d'énergie de manière détaillée, encore faut-il avoir la forme de l'hamiltonien... et celui utilisant les deux électrons de la molécule de dihydrogène est déjà bien tordu... Pour simplifier l'étude, nous nous réduisons au cas du cation (ion positif) H_2^+ constitué de deux protons et un électron:



Figure: 54.2 - Étude simplifiée du cation hydrogène H_2^+

Nous avons alors en nous basant sur la relation obtenue au début de ce chapitre:

$$H_{el} = -\sum_{i=1}^n \frac{\hbar^2}{2m_e} \nabla_i^2 - \sum_{k=1}^N \sum_{i=1}^n \frac{Ze^2}{4\pi\epsilon_0 r_{ik}} + \sum_{i=1}^{n-1} \sum_{j>i} \frac{e^2}{4\pi\epsilon_0 r_{ij}} \quad (54.55)$$

la relation suivante:

$$H = -\frac{\hbar^2}{2m_e} \nabla^2 + \frac{e^2}{4\pi\epsilon_0} \left(-\frac{1}{r_1} - \frac{1}{r_2} + \frac{1}{r} \right) \quad (54.56)$$

où les deux premiers termes dans la parenthèse sont pour rappel associés à l'énergie potentielle de l'électron et le dernier à l'énergie potentielle de répulsion des protons (le premier terme à droite de l'égalité étant l'énergie cinétique de l'électron).

Essayons maintenant d'ordonner l'énergie de ces deux orbitales moléculaires. Pour cela, écrivons:

$$E_1 = \frac{H_{AA} - H_{AB}}{1 - S_{AB}} = \frac{\alpha - \beta}{1 - S} \text{ et } E_2 = \frac{H_{AA} + H_{AB}}{1 + S_{AB}} = \frac{\alpha + \beta}{1 + S} \quad (54.57)$$

Rappelons que pour qu'un système soit stable, les énergies E_n doivent être négatives, ce qui correspond à des états stables (il faut un apport de l'énergie pour les défaire) et nous impose à cause de la forme de E_2 :

$$\alpha < 0, \beta < 0 \quad (54.58)$$

Sachant ceci, il vient:

$$E_2 < E_1 < 0 \quad (54.59)$$

Dès lors, nous voyons que les notations ne sont pas conformes à l'usage de la physique quantique car normalement l'indice 1 est réservé au niveau de plus basse énergie. Nous écrirons donc à l'avenir:

$$E_1 = \frac{\alpha + \beta}{1 + S} \text{ et } E_2 = \frac{\alpha - \beta}{1 - S} \quad (54.60)$$

avec les fonctions propres associées Ψ_1 et Ψ_2 et donc:

$$E_1 < E_2 < 0 \quad (54.61)$$

Nous pouvons également remarquer une chose importante! C'est que si nous considérons les atomes isolément, les termes d'interactions s'annulent et nous avons:

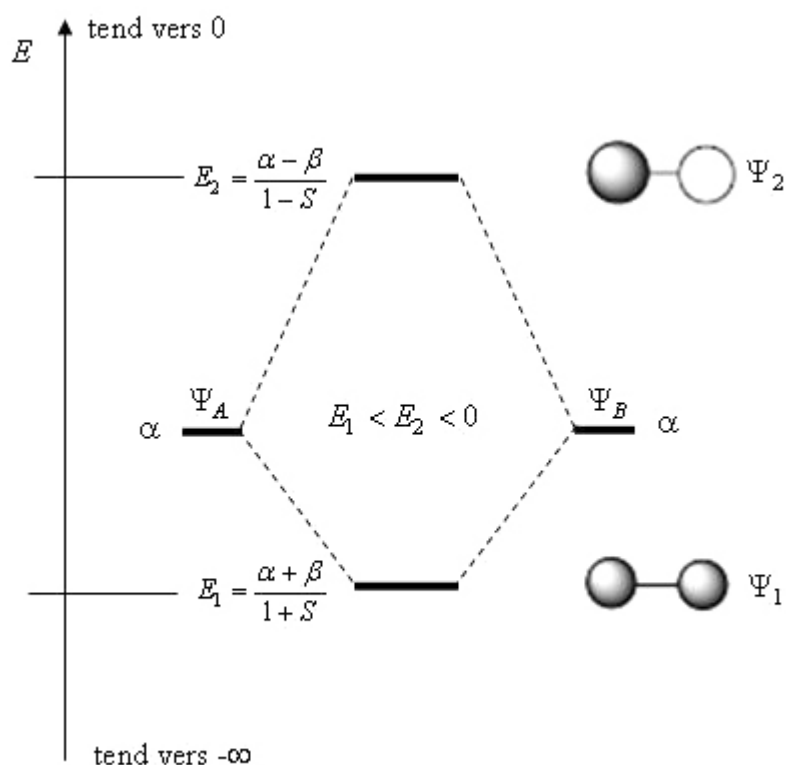
$$E_{1,2} \cong \alpha \quad (54.62)$$

Donc nous avons la différence qualitative entre un atome isolé et un système diatomique simple:

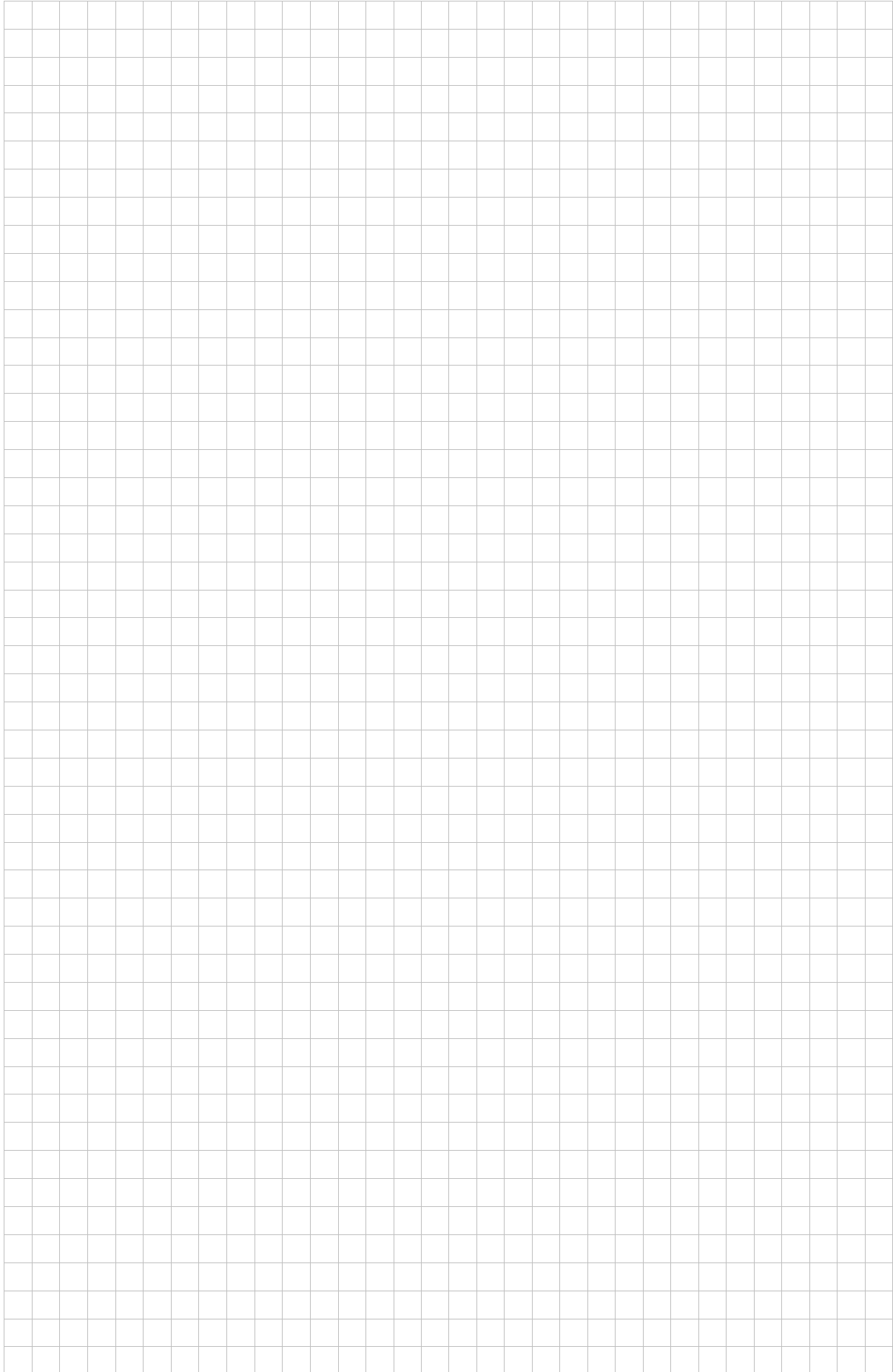
$$\frac{\alpha + \beta}{1 + S} < \alpha < 0 \quad (54.63)$$

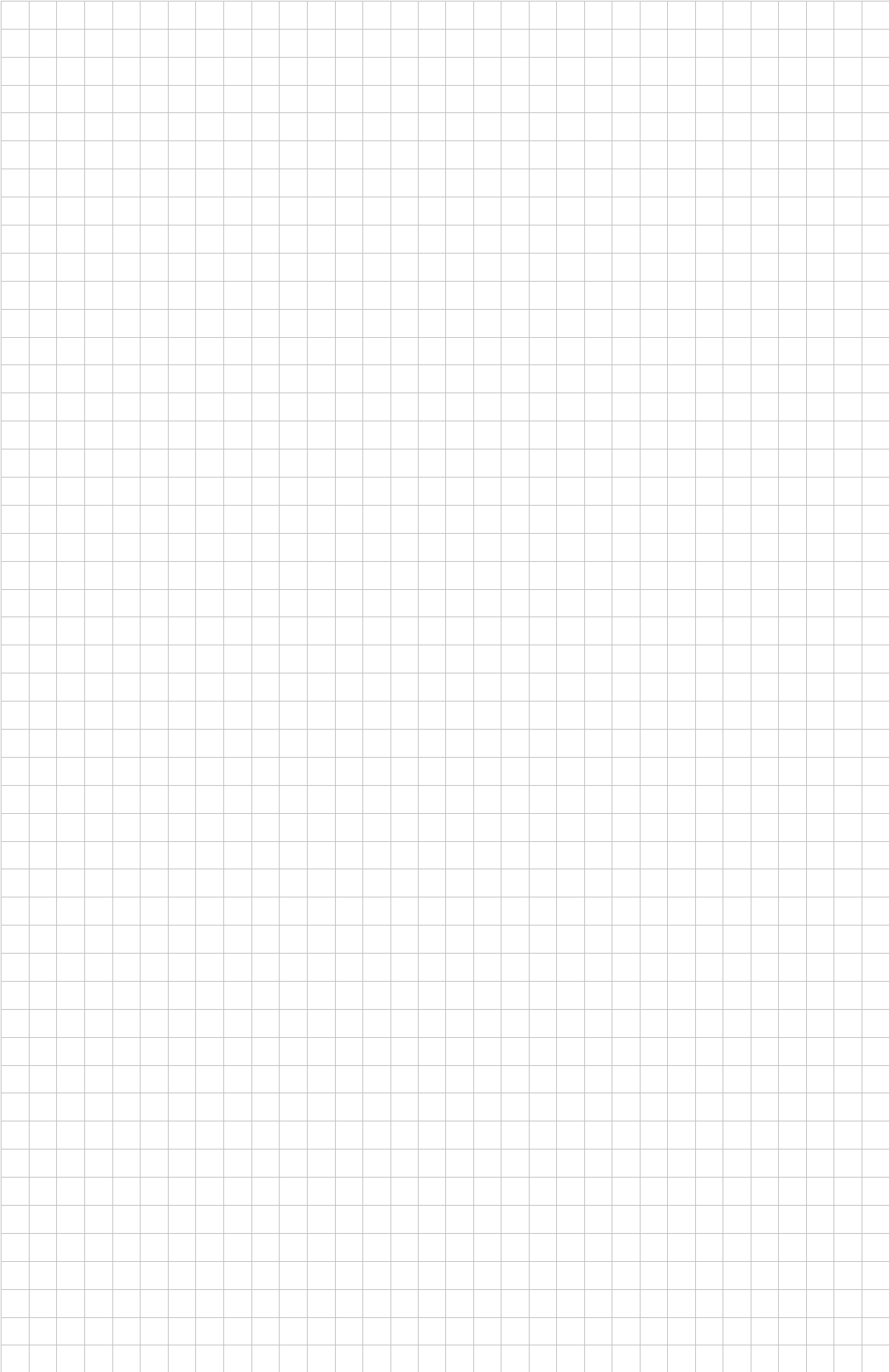
Ce qui signifie que l'énergie du plus bas niveau d'une molécule diatomique est inférieure à l'énergie d'un atome isolé qui est proche de α . Cette observation confirme que le système est stabilisé en énergie par rapport aux deux atomes isolés, ce qui paraît cohérent avec le constat expérimental de l'existence de telles molécules.

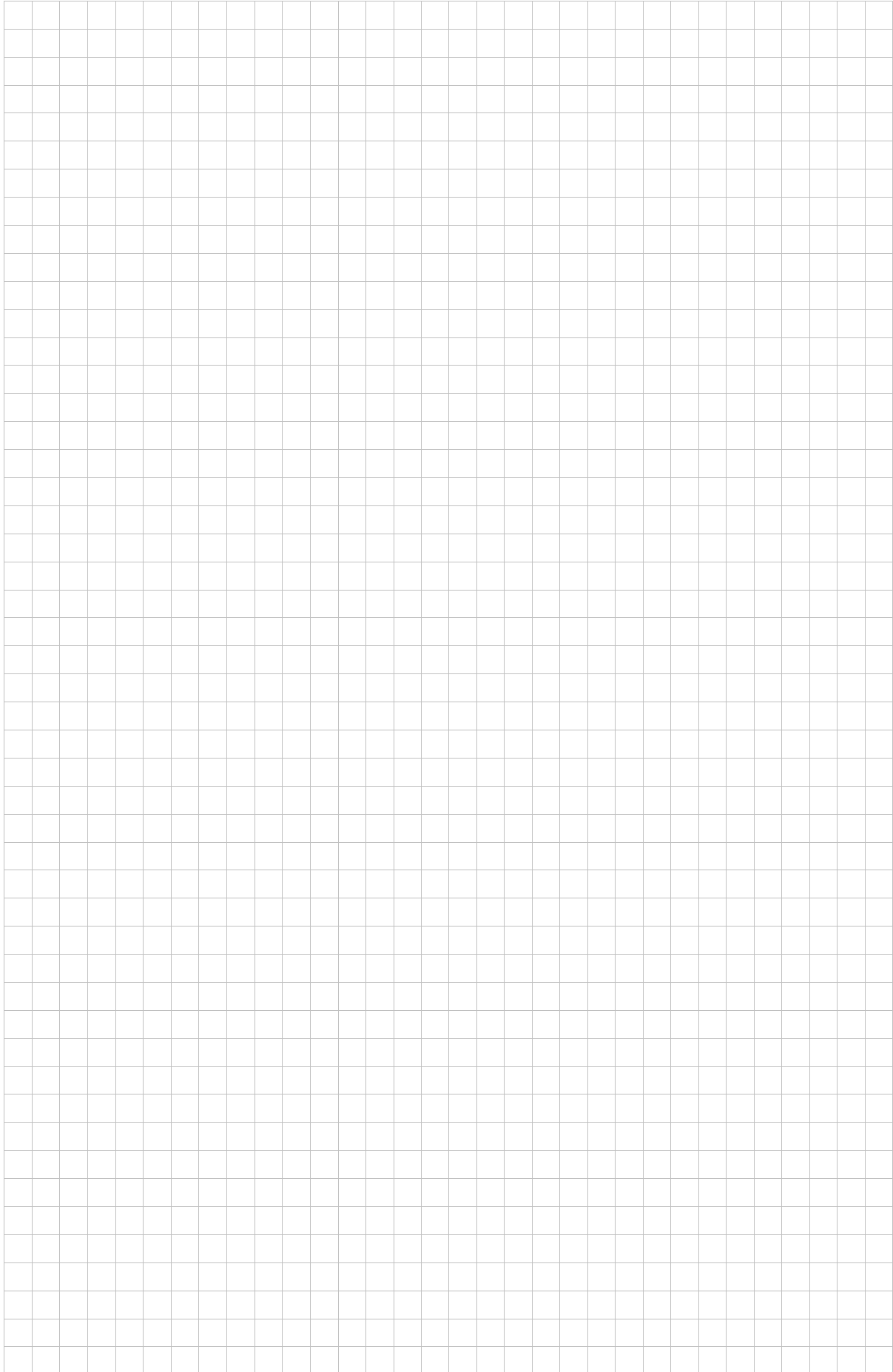
L'usage veut que les chimistes représentent les différences d'énergie sous la forme suivante pour notre cas particulier:

Figure: 54.3 - Niveaux d'énergie du cation hydrogène H_2^+

Nous concluons donc que lorsque deux atomes (contribuant chacun avec un électron) s'unissent, leurs orbitales atomiques vont se combiner pour générer deux orbitales moléculaires, l'une Ψ_1 d'énergie plus basse et une seconde Ψ_2 d'énergie plus élevée que celle des atomes isolés. Ainsi, la rupture qui fera partir l'électron avec un des atomes sera exothermique par rapport aux atomes isolés.







55. CHIMIE ANALYTIQUE

La chimie est une science très complexe à n corps que la mathématique ne peut expliquer sans l'apport de simulations numériques sur ordinateur ou d'approximations quant à l'utilisation de la théorie quantique (voir section d'atomistique). D'ici que ces outils soient suffisamment puissants et accessibles à tout le monde, la chimie reste une science principalement expérimentale basée sur l'observation de différentes propriétés de la matière dont voici quelques définitions fort importantes (que nous retrouvons d'ailleurs dans d'autres domaines que la chimie).

Définitions:

D1. Une "**propriété subjective**" est une propriété basée sur une impression personnelle/individuelle, par exemple, la beauté, la sympathie, la couleur, l'utilité, etc.

D2. Une "**propriété objective**" est une propriété ressentie (qui ne peut être contredite), par exemple, sa masse, son volume, sa forme, etc.

D3. Une "**propriété qualitative**" est une propriété descriptive donnée à l'aide de mots. Par exemple: forme ovale, magnétique, conducteur, etc.

D4. Une "**propriété quantitative**" est une propriété qui se chiffre. Par exemple: sa masse, son volume, sa densité, etc.

D5. Une "**propriété caractéristique**" est une propriété exclusive qui permet d'identifier une substance pure. Elle ne change pas même si l'on transforme physiquement la matière, par exemple, sa masse volumique, son point d'ébullition, son point de fusion, etc.

D6. Un corps est dit un "**corps pur**" lorsque tout échantillon de ce corps présente des valeurs de constantes bien définies et identiques à celles de l'ensemble (densité, point de fusion, point d'ébullition, indice de réfraction, etc.).

Remarque: Nous connaissons environ 2'000'000 de substances pures différentes en ce début du 21ème siècle (c'est dire...qu'il y a du travail derrière).

D7. Nous nommons "**corps composés**", les corps qui, soumis à des procédés chimiques, restituent leurs composants sous forme de corps purs.

D8. Si nous effectuons la séparation des mélanges et la décomposition des combinaisons, nous obtenons finalement des corps qui sont indécomposables par les méthodes chimiques classiques; nous les nommons "**éléments**" ou "**corps simples**".

Ainsi, la plus petite partie d'une combinaison chimique présentant encore toutes les propriétés de celle-ci est la "**molécule**" de cette combinaison. La plus petite partie d'un élément ou corps simple est "**l'atome**" de cet élément.

Quelques rappels d'ordre général au préalable:

R1. Un mélange est dit "**hétérogène**" en chimie si ces constituants sont immédiatement discernables à l'oeil nu ou au microscope

R2. Un mélange est dit "**homogène**" en chimie si ces constituants ne sont pas discernables à l'oeil nu ou au microscope.

1. MÉLANGES SIMPLES

Avant de partir dans des équations plus ou moins compliquées, le cas le plus simple d'application des mathématiques à la chimie par lequel nous pouvons commencer est la gestion des mélanges pour les opérations d'analyse et de contrôle de réactions chimiques simples avec deux mélanges.

Considérons deux exemples typiques et particuliers:

1. Soit une solution (jaune) de 10 millilitres d'une solution qui contient une concentration d'acide à 30%. Combien de millilitres d'acide pur (bleu) devons-nous rajouter pour augmenter la concentration (verte) à 50%?

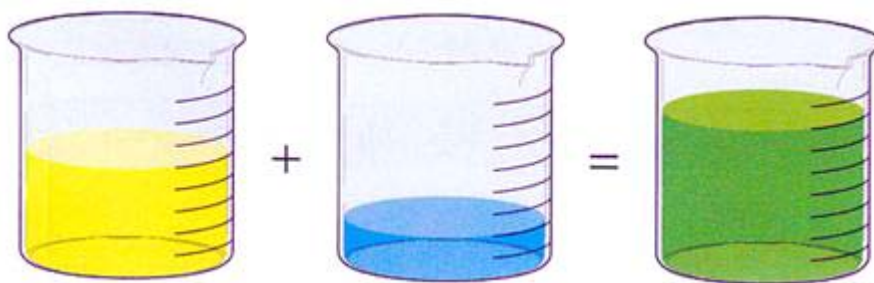


Figure: 55.1 - La joie des mélanges...

Puisque l'inconnue est la quantité d'acide pur à rajouter, nous la noterons x . Nous avons alors:

$$10 \cdot 30\% + x = 50\% \cdot (10 + x) \quad (55.1)$$

Ce qui donne:

$$3 + x = 5 + 0.5x \quad (55.2)$$

Il vient alors trivialement:

$$x = 4 \quad (55.3)$$

Donc 4 millilitres d'acide devraient être ajoutés à la solution d'origine.

2. Un jerricane contient 8 litres d'essence et d'huile pour faire fonctionner un agrégat. Si 40% du mélange initial est de l'essence, combien devrions-nous enlever du mélange (reste en rose) pour le remplacer par de l'essence pure (vert clair) pour que le mélange final (vert kaki) contienne 60% d'essence?



Figure: 55.2 - La joie des mélanges chez les bricoleurs et militaires...

Nous noterons x l'inconnue qui est le nombre de litres du mélange initial à retirer et à remplacés par l'essence pure qui étant de quantité égale est aussi x . Nous avons alors:

$$(8 \cdot 40\% - x \cdot 40\%) + x = 8 \cdot 60\% \quad (55.4)$$

Ce qui donne:

$$0.4(8 - x) + x = 8 \cdot 0.6 \quad (55.5)$$

Il vient alors trivialement:

$$x = 2.\bar{6} \quad (55.6)$$

Donc 2.6 litres devraient être enlevés du mélange d'origine et être remplacés par 2.6 d'essence pure.

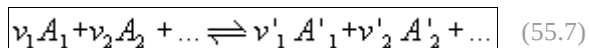
Bref voilà pour les mélanges. Nous pouvons aller beaucoup plus loin, et faire beaucoup plus compliqué avec plus d'inconnues mais nous nous arrêterons là pour l'instant.

2. RÉACTIONS

Puisque l'étude principale en chimie consiste à observer les résultats de mélanges de corps purs et/ou composés, il convient d'abord de nous attarder sur les règles de bases qui régissent ces mélanges dans des conditions normales de pression et de température (C.N.T.P.).

Il convient au préalable de préciser que nous n'allons pas étudier dans ce chapitre ce qui crée les liaisons entre les éléments, car ceci est le rôle de la chimie quantique et moléculaire (voir chapitres précédents). De plus, nous insistons sur le fait que chaque élément théorique sera illustré d'un exemple pratique auquel il peut être utile de se reporter parfois pour mieux comprendre.

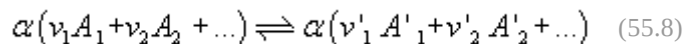
Considérons maintenant un système chimique fermé (sans transfert de masse donc!). Nous traduisons la modification de la composition (s'il y a lieu et si elle existe) du système chimique par une équation de réaction de la forme (le système ne va pas toujours dans les deux sens!):



appelée "**équation de bilan**" où les coefficients $\nu_i \in \mathbb{N}^*$ sont appelés "**coefficients stoechiométriques**" dans le sens où ils indiquent les "proportions d'or", rigoureusement appelées "**proportions stoechiométriques**", nécessaires telles qu'à des conditions normales la réaction puisse avoir lieu et où les A_i sont les produits réagissants (purs ou composés) et les A'_i produits formés.

Attention! Dans l'écriture de l'équation ci-dessus, nous imposons que tous A_i sans exception réagissent à la réaction chimique et donc que tous les ν_i sont dépendants.

Si les proportions d'or sont respectées (tels que les coefficients soient bien stoechiométriques!) et existent lors de l'écriture de l'équation de réaction, alors pour tout $\alpha \in \mathbb{N}$ nous avons:



cette proposition n'est démontrable que si les coefficients stoechiométriques d'un côté ou l'autre de la réaction varient de manière proportionnelle. L'expérience montre que dans des conditions normales de température et de pression (C.N.T.P) cela est bien le cas.

Dès lors, la stoechiométrie de la réaction impose que s'il disparaît dans le système x_1 moles de A_1 , x_2 moles de A_2 avec respectivement une variation de matière des produits $dn_1, dn_2, \dots$, il apparaîtra en conséquence x'_1 moles de A'_1 , x'_2 moles de A'_2 , ... avec respectivement une variation de matière des produits $dn'_1, dn'_2, \dots$ en respectant les proportionnalités des coefficients stoechiométriques tel que nous puissions écrire "**l'équation du bilan de matière**":

$$\frac{-dn_1}{\nu_1} = \frac{-dn_2}{\nu_2} = \dots = \frac{dn'_1}{\nu'_1} = \frac{dn'_2}{\nu'_2} = \dots = d\xi \quad (55.9)$$

où $d\xi$ est appelé "**avancement élémentaire de la réaction**" (fréquemment on prendra les valeurs absolues des rapports pour ne pas avoir à réfléchir sur le signe des variations).

La division des variations dn_1, dn_2 et dn'_1, dn'_2 par leurs coefficients stoechiométriques se justifie uniquement pour des raisons de normalisation ayant pour objectif de rapporter $d\xi$ à une valeur comprise entre 0 et 1 (soit entre 0% et 100%...).

Ces dernières égalités indiquent simplement que si l'un des produits réactif disparaît en une quantité donnée, les autres produits réactifs voient leur quantité diminuée en rapport à leur coefficient stoechiométrique de manière à conserver la proportion d'or de la réaction.

L'écriture du bilan énergétique peut être allégée par l'introduction des coefficients stoechiométriques algébriques ν_i tels que: $\nu_i > 0$ pour un produit formé, $\nu_i < 0$ pour un produit réagissant.

Finalement nous pouvons écrire:

$$d\xi = \frac{dn_i}{\nu_i} \quad (55.10)$$

que nous retrouvons également souvent dans la littérature avec la valeur absolue au numérateur!

Dès lors, avec cette convention algébrique, l'équation de réaction comme elle existe, permet d'écrire:

$$\sum_i \nu_i A_i = 0 \quad (55.11)$$

ce qui signifie que la somme algébrique du nombre total de composés purs des réactifs et produits

formés est toujours nulle.

Il est clair qu'à l'instant initial de la réaction nous choisissons pour l'avancement la valeur $\xi = 0$ (sa valeur maximale étant égale à l'unité), instant auquel les quantités de matière sont $n_{i,0}$.

L'intégration de l'expression différentielle du bilan de matière donne bien évidemment:

$$\int_{t_0}^t \nu_i d\xi = \int_{t_0}^t dn_i \Rightarrow \nu_i \left[\xi(t) - \underbrace{\xi(t_0)}_{=0} \right] = n_i(t) - \underbrace{n_i(t_0)}_{n_{i,0}} \quad (55.12)$$

$$\Rightarrow \nu_i \xi = n_i(t) - n_{i,0} \Rightarrow \boxed{n_i = n_{i,0} + \nu_i \xi}$$

relation que nous retrouvons dans les tableaux d'avancements (voir plus bas) en se souvenant bien que $\nu_i > 0$ pour un produit formé, $\nu_i < 0$ pour un produit réagissant.

Se pose alors la question: Quelle est la valeur maximale $\xi_{\max}$ de l'avancement d'une réaction ? Eh bien la réponse à cette question est ma foi fort simple. La valeur maximale d'avancement d'une réaction ayant les proportions stoechiométriques (nous respectons la tradition ainsi plutôt qu'en parlant de proportions d'or...) est telle qu'elle a lieu lorsque les réactifs auront tous disparu et dès lors elle est nécessairement donnée par:

$$\xi_{\max} = \left| \frac{n_i}{\nu_i} \right| \quad (55.13)$$

pour ce que nous appelons le "**réactif limitant**", c'est-à-dire le produit réagissant qui disparaît (a toujours la plus petite valeur de molarité) en premier et qui arrête la réaction attendue! S'il n'y a pas de réactif limitant, c'est qu'à la fin de la réaction tous les réactifs ont été transformés: on dit que les réactifs étaient en proportion stœchiométriques.

Il peut être utile de définir le "**taux d'avancement**" A_i donné par la grandeur intensive:

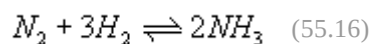
$$\tau_i = \frac{\text{quantité de } A_i \text{ ayant réagi}}{\text{quantité de } A_i \text{ initiale}} \quad (55.14)$$

ce qui de manière formelle donne:

$$\tau_i = \frac{|\nu_i| \xi}{n_{i,0}} \quad (55.15)$$

Exemple:

Considérons pour illustrer tous ces concepts la réaction (diazote avec hydrogène donnant de l'ammoniac):



où les lettres latines représentent des corps purs (atomes) dont le nom ne nous importe pas (notation proposée par Jöns Jacob Berzelius en 1813). Les indices représentent tout simplement la combinaison

des atomes pour obtenir une molécule. Nous avons dans cette réaction:

$$\nu_1 = 1, \nu_2 = 3, \nu'_1 = 2 \text{ et } A_1 \equiv N_2, A_2 \equiv H_2, A'_1 \equiv NH_3 \quad (55.17)$$

Le lecteur remarque que nous avons bien selon notre convention pour le bilan de masse:

$$\sum_i \nu_i A_i = 1 \cdot (2N) + 3 \cdot (2H) - 2(N) - 2(3H) = 0 \quad (55.18)$$

Si nous considérons qu'il y a une mole de chaque corps composé, cela nous donne pour les proportions stoechiométriques (à un facteur $x \in \mathbb{R}^*$ près pour toutes les valeurs):

$$n_1 = \nu_1 \cdot 1 = 1 \cdot 1 = 1[\text{mol}], n_2 = \nu_2 \cdot 1 = 3 \cdot 1 = 3[\text{mol}], n_3 = \nu'_1 \cdot 2 = 1 \cdot 2 = 2[\text{mol}] \quad (55.19)$$

Si à un moment $t \neq t_0$ donné, nous obtenons par mesure:

$$n_3 = 0.2[\text{mol}] \quad (55.20)$$

Quel est l'avancement de la réaction ?

$$d\xi = \frac{dn_i}{\nu_i} \Rightarrow \xi = \frac{\Delta n(NH_3)}{2} = \frac{0.2}{2} = 0.1 \quad (55.21)$$

soit autrement dit, nous en sommes à 10% (logique!).

Le taux de conversion de NH_3 y relatif est donc:

$$\tau_i = \frac{|\nu_i| \xi}{n_{i,0}} = \frac{|-3| 0.1}{3} = 0.1 \quad (55.22)$$

Et quelle est la valeur maximale d'avancement $\xi_{\max}$ du réactif limitant ?

Donc dans le cadre de l'exemple ci-dessus où nous avons $n_{1,0} = 1[\text{mol}]$ pour le N_2 alors:

$$\xi_{\max} = \left| \frac{n_i}{\nu_i} \right| = \left| \frac{1}{-1} \right| = 1[\text{mol}] \quad (55.23)$$

Les chimistes utilisent également souvent ce qu'ils appellent un "tableau d'avancement".

Voyons de quoi il s'agit avec notre exemple:

	Équation	N_2	$+ 3H_2$	$= 2NH_3$
État initial	$n_{i,0}$	1	3	0
État intermédiaire	$n_i = n_{i,0} + \nu_i \xi$	$1 - 1 \cdot \xi$	$3 - 3 \cdot \xi$	$0 + 2 \cdot \xi$
État final	$\xi_{\max}$	$1 - 1 \cdot \xi_{\max}$	$3 - 3 \cdot \xi_{\max}$	$0 + 2 \cdot \xi_{\max}$

Tableau: 55.1 - Avancement d'une réaction chimique

Recherchons $\xi_{\max}$ à partir de ce tableau. Le réactif limitant est donc soit N_2 , soit $3H_2$.

Donc pour N_2 :

$$1 - 1 \cdot \xi_{\max} = 0 \Rightarrow \xi_{\max} = \frac{n_i}{\nu_i} = 1 \quad (55.24)$$

et pour $3H_2$:

$$3 - 3 \cdot \xi_{\max} = 0 \Rightarrow \xi_{\max} = \frac{n_i}{\nu_i} = 1 \quad (55.25)$$

Chaque réactif possédant le même avancement $\xi_{\max}$, celui-ci est donc aussi le $\xi_{\max}$ minimal. En conséquence, selon la définition de l'actif limitant, vu que les proportions sont stoechiométriques dans l'exemple donné aucun des réactifs n'est limitant.

Notes personnelles:







56. CHIMIE THERMIQUE (THERMOCHIMIE)

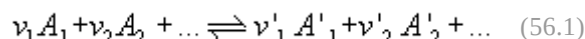
La thermochimie est la branche qui s'intéresse, historiquement, aux phénomènes thermiques et aux bilans qui accompagnent les réactions chimiques. Elle puise principalement ses fondements dans la thermodynamique.

Nous ne saurions donc que recommander au lecteur d'avoir lu ou parcouru le chapitre traitant de la thermodynamique dans la section de mécanique car de nombreuses notions qui y ont été vues seront supposées comme connues dans ce chapitre.

Par ailleurs, il est fortement recommandé de lire le présent chapitre en parallèle de celui de Chimie Analytique (cela peut être embêtant mais il faut faire avec).

1. TRANSFORMATIONS CHIMIQUES

Soit le système fermé siège de la réaction chimique (cf. chapitre de Chimie Analytique):



Nous allons pour simplifier... considérer que la réaction chimique est totale et que les produits réagissants sont utilisés en quantités stoechiométriques (état 1: Σ_1) pour donner les produits formés, également en quantités stoechiométriques (état 2: Σ_2).

Si la transformation s'effectue à volume (quasi-)constant, le travail sur l'atmosphère ambiante est nul car (cf. chapitre de Thermodynamique):

$$\Delta W = - \int_1^2 P dV = 0 \quad (56.2)$$

L'application du premier principe se réduit et permet alors d'écrire:

$$U_2 - U_1 = Q_V \quad (56.3)$$

où Q_V est dans le cadre de la chimie thermique appelée "**chaleur de réaction à volume constant**", échangée bien sûr entre le système et le milieu extérieur (nous n'écrivons pas le delta pour indiquer qu'il s'agit d'une variation par tradition...).

Rappelons que:

- Si $Q_V > 0$, la réaction est dite "**endothermique**" (le système reçoit de la chaleur du milieu extérieur)
- Si $Q_V < 0$ la réaction est dite "**exothermique**" (le système cède de la chaleur au milieu extérieur)
- Si $Q_V = 0$ la réaction est dite "**athermique**" (pas d'échange de chaleur avec le milieu extérieur)

Remarque: Rappelons aussi qu'un système fermé n'est pas un système isolé! Pour un rappel des différentes définitions, nous renvoyons le lecteur au chapitre de Thermodynamique.

Si la réaction se fait à pression constante (cas le plus courant dans la pratique), soit isobare, nous avons alors:

$$\Delta W = - \int_1^3 P dV = -P(V_3 - V_1) \quad (56.4)$$

Remarque: Le choix des indices d'intégration sont différents à précédemment pour différencier le fait qu'une réaction à pression ou à volume constant ne sont pas nécessairement identiques.

L'application du premier principe de la thermodynamique, entre les deux états, permet d'écrire:

$$U_3 - U_1 = Q_P - P(V_3 - V_1) \quad (56.5)$$

où Q_P est la quantité de chaleur, appelée "**chaleur de réaction à pression constante**", échangée entre le système et le milieu extérieur (Q_P est donc une variation... même si la notation traditionnelle malheureuse des thermodynamiciens ne le met pas en évidence...).

En utilisant la définition de l'enthalpie, nous pouvons écrire cette dernière relation sous la forme:

$$Q_P = U_3 - U_1 + P(V_3 - V_1) = (U_3 + PV_3) - (U_1 + PV_1) = H_3 - H_1 \quad (56.6)$$

Si nous travaillons avec les volumes molaires, ceux des phases condensées (solides et liquides donc) étant négligeables devant le volume molaire gazeux, seuls les constituants gazeux auront une enthalpie très différente de leur énergie interne (voir l'exemple dans le chapitre de Thermodynamique). Nous aurons dès lors dans le cadre de l'approximation des gaz parfaits (cf. chapitre de Thermodynamique):

$$H_i = U_i + n_i RT \quad (56.7)$$

Dans le cadre des gaz parfaits, la relation antérieure peut donc s'écrire:

$$\begin{aligned} Q_P = H_3 - H_1 &= U_3 - U_1 + (n_3 - n_1) RT = U_3 - U_1 + U_2 + (n_3 - n_1) RT - U_2 \\ &= Q_V + (n_3 - n_1) RT + U_3 - U_2 \end{aligned} \quad (56.8)$$

Or, comme (cf. chapitre de Mécanique Des Milieux Continus) U_2 et U_3 sont tous deux les mêmes états finaux d'une même réaction complète et que pour un gaz monoatomique:

$$U_i = \frac{3}{2} N_i kT \quad (56.9)$$

donc l'énergie interne U_2 et U_3 ne dépend que du nombre de constituants mais... ils sont égaux puisque même état final d'une même réaction!

Ainsi, nous avons:

$$U_2 = U_3 \Rightarrow U_3 - U_2 = 0 \quad (56.10) .$$

En posant $\Delta n = n_2 - n_1$ (différence entre le nombre de moles de gaz des produits formés et ceux des produits réagissant), nous pouvons écrire pour une réaction chimique:

$$Q_p = Q_v + \Delta n RT \quad (56.11)$$

ce qui permet de différencier l'énergie mise en jeu entre réaction isobare et isochore et de chercher le meilleur choix à faire en fonction des objectifs industriels.

Il est intéressant de remarquer que si le delta du nombre de moles est nul. Les variations de chaleur isobare ou isochore sont égales et qu'il n'y a donc pas de raison de préférer a priori l'une ou l'autre des transformations.

Évidemment dans la pratique le problème consiste à connaître les valeurs des différentes variables de la relation. Ces valeurs se trouvant dans d'immenses bases de données auxquelles les chimistes ont accès... cette relation n'est que très très rarement utilisée dans la pratique et de toute façon elle est basée sur des hypothèses simplificatrices trop contraignantes pour être d'un vrai intérêt pratique.

2. GRANDEURS MOLAIRES

Définitions:

D1. Par convention, la "**mole**" est la quantité de matière d'un système contenant autant d'espèces chimiques qu'il y a d'atomes de carbone dans 12 [g] de carbone 12 (cf. chapitre de Physique Nucléaire).

Le nombre d'atomes de carbone contenus dans 12 [g] de carbone est égal au nombre d'Avogadro donné approximativement par:

$$N_{Av.} \cong 6.0221367 \cdot 10^{23} [mol^{-1}] \quad (56.12)$$

Cela signifie in extenso et par construction qu'une mole d'eau, de fer, d'électrons, contient respectivement toujours un nombre d'atomes égal au nombre d'Avogadro.

Remarquons qu'avec un système mélangé c'est un non-sens mathématique de faire la somme des masses molaires des constituants pour avoir la masse molaire totale. La masse molaire est donc une grandeur intensive!

D2. La "**masse molaire atomique**" est la masse d'une mole d'atomes des éléments chimiques concernés.

Remarque: Nous trouvons ces masses molaires atomiques dans les classifications périodiques. Mais il faut surtout savoir que celles qui sont indiquées prennent en compte les isotopes naturels (ce qui est bien normal puisque chimiquement ils sont indiscernables à part pour le chimiste nucléaire ou le physicien nucléaire). Donc la valeur indiquée dans les tables se calcule comme la somme des proportions respectives des masses molaires des différents isotopes (la validité de cette manière de calculer est toute relative...).

D3. La "**masse molaire moléculaire**" est égale à la somme des masses molaires atomiques des éléments chimiques qui la constitue.

Il vient dès lors immédiatement le constat suivant: la masse m d'un échantillon constitué d'une quantité de n moles d'espèces chimiques tout identiques de masse molaire M_m est donnée par la relation:

$$m = n \cdot M_m \quad (56.13)$$

De manière un peu plus formelle et sous un aspect thermodynamique, voici aussi comment nous pouvons définir la masse molaire:

Soit X une grandeur extensive relative à un système monophasique (voir le chapitre de Thermodynamique pour le vocabulaire dont il est fait usage) et soit un élément de volume dV de ce système, autour d'un point courant M et contenant la quantité de matière dn . Nous lui associons la grandeur extensive dX proportionnelle à dn tel que:

$$dX = X_m dn \Leftrightarrow X_m = \frac{dX}{dn} \quad (56.14)$$

ce qui fait que X_m est une grandeur intensive (quotient de deux grandeurs extensives selon ce qui a été vu dans le chapitre de Thermodynamique) que nous appellerons par définition "**grandeur molaire associée**" à X .

Nous en déduisons que:

$$X = \int X_m dn \quad (56.15)$$

l'intégrale portant sur tout le système monophasique.

Dans le cas d'une phase uniforme, X_m étant constant en tout point, nous pouvons simplement écrire:

$$X = nX_m \quad (56.16)$$

Remarque: En gros l'idée est de dire que la masse d'un système chimique monophasique est proportionnelle à la masse molaire de celui-ci à un facteur entier près représentant le nombre de ses constituants (où le nombre de moles pour être plus exact).

D4. Dès lors que le système est hétérogène, nous utilisons le concept de "**titre molaire**", défini par:

$$x_i = \frac{n_i}{n} \quad (56.17)$$

x_i étant donc le titre molaire d'une espèce A_i dont la quantité de matière (le nombre de moles par exemple) est n_i avec $n = \sum_i n_i$ étant la quantité de matière totale de la phase envisagée.

Il en résulte que, pour toutes les espèces de la phase envisagée, $\sum_i x_i = 1$, ce qui veut dire que, s'il y a n espèces chimiques, il suffit de connaître $n-1$ titres molaires pour toutes les connaître.

Si la phase considérée est gazeuse et en considérant un gaz parfait suivant la loi de Boyle-Mariotte (approximation de l'équation de Van der Waals démontrée dans le chapitre de Mécanique Statistique) nous avons:

$$P_i V = n_i RT \text{ et } PV = n RT \quad (56.18)$$

nous avons dès lors aussi la possibilité dans le cas de phases gazeuses d'exprimer le titre molaire comme:

$$x_i = \frac{n_i}{n} \cdot \frac{nRT}{nRT} = \frac{P_i V}{PV} = \frac{P_i}{P} \quad (56.19)$$

Remarque: Nous pourrions faire de même avec le volume.

D5. Nous définissons le "**titre massique associée à l'espèce**" A_i par le rapport:

$$w_i = \frac{m_i}{m} \quad (56.20)$$

avec $m = \sum_i m_i$ étant la masse totale de la phase étudiée. Nous avons bien sûr également $\sum_i w_i = 1$.

D6. Nous définissons la "**concentration molaire volumique**" ou "**molarité**" par le rapport (attention à ne pas confondre la notation avec la chaleur massique):

$$c_i = \frac{n_i}{V} \quad (56.21)$$

Remarque: Il existe d'autres grandeurs de composition beaucoup moins utilisées que x_i ou c_i . Nous pouvons citer la "**concentration massique volumique**" m_i/V , la "**molalité**" (quotient de la quantité de matière de l'espèce A_i par la masse du solvant", etc.

En chimie thermique, nous considérons très souvent les gaz comme parfaits et nous faisons jouer un rôle privilégié à une pression particulière appelée "**pression standard**" P° et dont la valeur vaut (pression atmosphérique):

$$P^\circ = 1[\text{bar}] = 10^5[\text{Pa}] \cong 1[\text{atmosphère}] \quad (56.22)$$

D6. Nous disons qu'un gaz (parfait) est dans l'état standard si sa pression est égale à la pression standard P° de 1 [bar].

D7. Nous appelons "**grandeur molaire standard**" $X_{m,i}^\circ$ d'un constituant la valeur de la grandeur molaire de ce même constituant pris à l'état standard, c'est-à-dire sous la pression P° .

Remarques:

R1. Toute grandeur molaire standard est bien évidemment intensive: la pression étant fixée par l'état standard, celle-ci ne dépend plus que de la température.

R2. Toute grandeur standard est notée grâce à l'exposant "°": V_m^* est donc le volume molaire standard. Par contre, la grandeur molaire standard n'est pas toujours spécifiée avec le petit m en indice, il faut donc parfois être prudent avec ce que l'on manipule.

Dans le cas du gaz parfait, le volume molaire se calcule grâce à l'équation d'état du gaz parfait. Il en résulte:

$$V_m^* = \frac{RT}{P^*} \quad (56.23)$$

Nous constatons bien évidemment que le volume molaire standard du gaz parfait dépend de la température.

Si nous faisons ce calcul aux "**conditions normales de température et de pression**" (abrégé C.N.T.P.), c'est-à-dire à une température de 273.15 [K] (soit 0 [°C]) et à une pression de 1 [atm] (soit 101.325 [kPa]), nous trouvons alors un volume de 22.4 [L/mol] qui est une valeur très connue par les chimistes.

Remarques:

R1. Dans une large gamme de températures et de pressions, le volume molaire des gaz réels n'est généralement pas très différent de celui d'un gaz parfait.

R2. Dans le cas d'un état condensé, nous ne disposons pas en général d'une équation d'état mais nous pouvons néanmoins mesurer le volume molaire.

Nous pouvons définir ainsi par extension d'autres grandeurs standard découlant de celles que nous avons définies dans le chapitre de Thermodynamique:

- "**L'énergie interne molaire standard**" (grandeur intensive puisque exprimée par unité molaire) U_m^*
- "**L'enthalpie molaire standard**" (grandeur intensive puisque exprimée par unité molaire) H_m^* avec:

$$H_m^* = U_m^* + P^* V_m^* = U_m^* + RT \quad (56.24)$$

Il est important que le lecteur observe que l'enthalpie ne dépend que de la température (et de l'énergie interne).

Remarque: Pour les états condensés le volume standard est très faible en unités S.I. ce qui fait que $H_m^* \cong U_m^*$. Il est cependant très difficile de parler de pression pour les états condensés donc cette approximation est à prendre avec des pincettes.

Si nous considérons maintenant une fonction extensive X (comme le volume par exemple!) définie sur un système chimique gazeux en évolution. Nous pouvons a priori, exprimer X en fonction des deux variables intensives T, P (car une fonction extensive est soit produit/rapport de deux grandeurs intensives, soit une somme de grandeurs extensives) et des différentes quantités de matières n_i, n'_i de A_i, A'_i tel que:

$$X = X(T, P, \dots n_i \dots) \quad (56.25)$$

Si tous les produits (réactifs et formés) sont dans leur état standard la fonction extensive, notée dès lors X^* , se met sous la forme:

$$X^* = X^*(T, \dots n_i \dots) \quad (56.26)$$

où la pression n'intervient plus puisque fixée à sa valeur standard. Le gaz n'est alors décrit que par sa température et de la quantité de ses constituants !

Si nous considérons cependant une évolution infiniment petite du système à température et pression constantes (car nous supposons les proportions des composants du gaz comme variant toujours en proportion) les différentes quantités de matières varient alors selon la différentielle totale exacte (cf. chapitre de Calcul Différentiel Et Intégral):

$$dX_T^* = \sum_i \frac{\partial X^*}{\partial n_i} dn_i \quad (56.27)$$

où n'interviennent bien évidemment par la contrainte précitée (température et pression constantes) que les différentielles par rapport aux quantités de matières qui varient.

Nous pouvons alors être amenés à définir artificiellement (rien ne nous en empêche, ce n'est pas faux!) la grandeur molaire standard intensive qui ne dépend que de la température:

$$X_i^* = \frac{\partial X^*}{\partial n_i} \quad (56.28)$$

Dès lors:

$$dX_T^* = \sum_i X_i^* dn_i \quad (56.29)$$

mais nous avons aussi défini en dans le chapitre de Chimie Analytique la relation:

$$dn = \nu d\xi \quad (56.30)$$

qui exprimait, rappelons-le, la variation de la quantité de matière d'un des composés de la réaction par rapport à sa proportion stoechiométrique (constante) et l'avancement de la réaction. Il vient dès lors:

$$dX_r^* = \left(\sum_i X_i^* \nu_i \right) d\xi \quad (56.31)$$

ainsi que:

$$\begin{aligned} dX_r^* &= \sum_i \frac{\partial X^*}{\partial n_i} dn_i = \sum_i \frac{\partial X^*}{\nu_i \partial \xi} \nu_i d\xi = \left(\sum_i \frac{\partial X^*}{\partial \xi} \right) d\xi = \left(\sum_i X_i^* \nu_i \right) d\xi \\ \Rightarrow \frac{dX_r^*}{d\xi} &= \sum_i X_i^* \nu_i = \sum_i \frac{\partial X^*}{\partial \xi} \end{aligned} \quad (56.32)$$

Par définition, nous appelons cette dernière somme algébrique "**grandeur standard de réaction associée à la fonction extensive X** " et la notons (notation mal choisie par les chimistes...):

$$\Delta_r X^* = \sum_i X_i^* \nu_i \quad (56.33)$$

qui est une grandeur intensive qui ne dépend que de la température et qui représente une variation relative (d'où le r en indice!). Cette relation peut aussi s'écrire:

$$\Delta_r X^* = \sum_i X_i^* \nu_i = \sum_i \frac{\partial X^*}{\partial \xi} = \frac{dX_r^*}{d\xi} \quad (56.34)$$

De manière générale, les chimistes appellent "**opérateur de Lewis**", noté Δ_r , la dérivée d'une grandeur X (standardisée ou non), par rapport à l'état d'avancement de la réaction ξ à température et pression constantes.

Remarque: Le symbole figure avec la lettre r en indice pour bien montrer qu'il s'agit d'une grandeur de réaction. En d'autres termes, il s'agit de variation de la grandeur molaire standard considérée au cours de la réaction concernée pour un avancement de réaction d'une mole à une pression de 1 bar pour un gaz parfait.

Il ne faut par ailleurs pas oublier que les coefficients stoechiométriques des réactifs sont positifs et ceux des produits négatifs (cf. chapitre de Chimie Analytique).

Il existe deux grandeurs de réaction jouant des rôles importants en chimie:

1. L'énergie interne molaire de réaction, appelée plus souvent "**énergie interne de réaction standard**", d'un système chimique:

$$\Delta_r U^* = \sum_i U_i^* \nu_i = \frac{dU_r^*}{d\xi} \quad (56.35)$$

2. L'enthalpie molaire de réaction, appelée plus souvent "**enthalpie de réaction standard**", d'un système chimique:

$$\Delta_r H^* = \sum_i H_i^* \nu_i = \frac{dH_r^*}{d\xi} \quad (56.36)$$

2.1. ENTHALPIE STANDARD DE RÉACTION

Dès lors nous pouvons considérer les deux cas suivants suivant en connaissant la relation (cf. chapitre de Thermodynamique):

$$H_i^* = U_i^* + V_i^* P^* : \quad (56.37)$$

1. Si les A_i sont dans un état condensé, puisque la pression interne ne s'applique pas nous avons:

$$H_i^* \cong U_i^* \quad (56.38)$$

qui reste quand même à prendre avec des pincettes suivant les scénarios!

2. Si les A_i sont à l'état gazeux (supposés parfaits):

$$H_i^* = U_i^* + RT \quad (56.39)$$

Il en résulte que seuls les gaz vont intervenir dans cette relation:

$$\Delta_r H^* = \sum_i H_i^* \nu_i = \sum_i \nu_i (U_i^* + RT) = \Delta_r U^* + \sum_i \nu_i RT \quad (56.40)$$

que nous écrivons conventionnellement:

$$\Delta_r H^* = \Delta_r U^* + \Delta_r \nu_g RT \quad (56.41)$$

Il en résulte que dans le cas particulier où:

$$\Delta_r \nu_g = 0 \quad (56.42)$$

(qui est au fait une notation... malheureuse... pour la somme algébrique des proportions stoechiométriques qui serait donc nulle) pour une température donnée alors nous avons:

$$\Delta_r H^* = \Delta_r U^* \quad (56.43)$$

où il faut se rappeler que les coefficients stoechiométriques des produits sont comptés comme positifs, alors que ceux des réactifs sont négatifs (cf. chapitre de Chimie Analytique).

Ainsi, la variation de la fonction enthalpie correspond à la variation de la quantité de chaleur mise en jeu dans une transformation isobare à une température T donnée. Raison pour laquelle celle-ci est parfois notée $\Delta_r H_{T,P}$.

Une réaction chimique qui a une enthalpie de réaction (qui est pour rappel la variation instantanée de l'enthalpie pendant une réaction) négative est dite exothermique, puisqu'elle libère de la chaleur dans l'environnement (pression constante oblige par définition de cette enthalpie de réaction!), alors qu'une réaction chimique dont l'enthalpie de réaction est positive est dite endothermique puisqu'elle nécessite alors un apport de chaleur pour se produire.

Ainsi, selon les développements précédents, si nous notons avec un indice p les produits et avec un indice i les réactifs, nous trouvons souvent l'enthalpie standard de réaction sous la forme suivante si les coefficients stoechiométriques sont tous comptés comme positifs:

$$\Delta_r H_{T,P} = \sum_p \nu_p H_p^\circ - \sum_r \nu_r H_r^\circ \quad (56.44)$$

sous cette forme, nous voyons alors bien que l'enthalpie standard de réaction correspond donc à la différence des enthalpies molaires partielles entre les produits et les réactifs. Il s'agit ni plus ni moins de la "**loi de Hess**" définie au 19ème siècle par le chimiste suisse Henri Hess.

Puisque dans un système à l'équilibre, l'énergie initiale est toujours supérieure ou égale à l'énergie finale (les systèmes tendent toujours à aller vers un état plus stable ayant une énergie minimale), alors l'enthalpie standard de réaction $\Delta_r H^\circ$ ne peut être que négative ou nulle.

Si une réaction chimique à pression constante et à une température donnée ne donne qu'un seul et unique composé chimique (produit) alors l'enthalpie standard de réaction est appelée "**enthalpie standard de formation**" et notée $\Delta_f H^\circ$.

Au fait, l'intérêt de la relation antécédente est que le chimiste peut simplement, sans avoir à connaître les quantités de matières mises en jeu, déterminer juste en connaissant les coefficients stoechiométriques d'une réaction chimique gazeuse ou condensée (s'il accepte qu'il s'agira alors d'une approximation pour ce dernier cas) isobare si la variation instantanée de l'énergie interne molaire lors de l'avancement d'une réaction à une température donnée est égale à la variation instantanée de l'enthalpie molaire.

Deux cas pratiques se présentent alors:

1. La différence entre la variation instantanée de l'énergie interne molaire et l'enthalpie molaire est nulle: Dès lors, la réaction chimique (à la température donnée) n'occupe pas instantanément un volume plus grand et ne perd donc aucune énergie à repousser ("inutilement") la pression du gaz environnant la réaction étudiée (cela peut être vu comme une économie d'argent en termes énergétiques dans une industrie chimique). Dans ce cas, l'enthalpie standard de réaction est égale simplement à la chaleur de réaction à pression constante Q_P .
2. La différence entre la variation instantanée de l'énergie interne molaire et l'enthalpie molaire est positive: Dès lors, la réaction chimique (à la température donnée) occupe instantanément un volume plus grand et perd donc une part d'énergie à repousser ("inutilement") la pression du gaz environnant la réaction étudiée (cela peut être vu comme une perte d'argent en termes de rendement énergétique dans une industrie chimique).

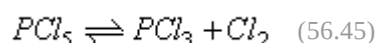
Remarque: Évidemment, il est tout à fait possible d'imaginer une entreprise qui tire profit de la variation du volume d'une réaction (le cas 2) qui repousse le gaz environnant avec un système de piston pour produire ensuite de l'énergie mécanique... donc il est serait possible dans certaines situations de perdre beaucoup moins d'argent (in extenso d'énergie...).

Une petite difficulté se présente cependant...l'enthalpie standard d'un corps pur simple (corps formé que d'un seul type d'atomes) ne peut être calculée de façon absolue car elle dépend de l'énergie interne qui est très difficilement calculable (il faut utiliser les outils de la physique quantique qui soulèvent des problèmes insurmontables aujourd'hui). Cela suppose de définir une échelle arbitraire d'enthalpies molaires en définissant un zéro arbitraire d'enthalpie et adopté à l'échelle internationale (ce qui n'est malheureusement pas le cas à ma connaissance!).

Ainsi, pour pouvoir dresser des tables d'enthalpies molaires standards, il a été choisi de définir l'échelle d'enthalpie de la manière suivante: l'enthalpie molaire standard d'un corps pur simple stable dans l'état standard est nulle à 298 [K]. Il s'ensuit que l'enthalpie de formation standard d'un corps pur simple est donc toujours nulle.

Exemple:

Soit la réaction:



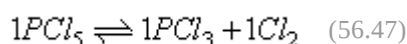
C'est-à-dire la dissociation du pentachlorure de phosphore en chlore et trichlorure de phosphore. Les tables nous donnent à la température de $T = 1'000$ [K] la valeur suivante de l'enthalpie standard molaire de cette réaction:

$$\Delta_r H_m^\circ = 156 \text{ [kJ} \cdot \text{mol}^{-1}] \quad (56.46)$$

où nous n'avons pas indiqué le petit m en indice

La variation de la valeur de l'enthalpie molaire de réaction étant positive, il s'ensuit que la réaction est endothermique (puisque nécessite un apport de chaleur d'où la température de dissociation de 1'000 [K]) et donc le produit est plus instable que le réactif initial.

Nous avons la somme algébrique suivante des coefficients stoechiométriques de la réaction:



Soit (qui est sans dimensions puisque l'enthalpie est en valeur molaire!):

$$\Delta_r \nu_g = -1 + 1 + 1 = 1 \quad (56.48)$$

donc la réaction fait augmenter la pression en créant une mole supplémentaire par mole de réactif.

Puisque:

$$\Delta_r H_m^\circ = \Delta_r U_m^\circ + \Delta_r \nu RT = \Delta_r U^\circ + 1 \cdot RT \quad (56.49)$$

il vient alors:

$$\Delta_r U^\circ = 148 \text{ [kJ} \cdot \text{mol}^{-1}] \quad (56.50)$$

C'est donc la part d'énergie interne absorbée par le système sur les $156 \text{ [kJ} \cdot \text{mol}^{-1}]$. Le reste (différence) étant juste passé dans la poussée de l'atmosphère environnante du réacteur chimique.

2.1.1. RELATION ENTHALPIQUE DE KIRCHHOFF

La relation enthalpique de Kirchhoff permet d'exprimer les variations de l'enthalpie de réaction (molaire ou non) en fonction de la température à partir de la connaissance de la capacité calorifique à pression constante des réactifs gazeux.

Nous avons donc défini dans les développements précédents la relation:

$$\Delta_r H^* = \sum_i H_i^* \nu_i = \frac{dH_r^*}{d\xi} \quad (56.51)$$

qui est donc l'enthalpie de réaction standard à une température donnée dans un système à la pression standard.

Nous avons également mentionné que Δ_r était pour rappel une notation un peu malheureuse pour l'opérateur différentiel (de Lewis) d'avancement de la réaction $d/d\xi$.

Si nous nous intéressons à l'influence de la température T sur $\Delta_r H^*$ nous n'avons qu'à écrire la différentielle exacte:

$$\frac{d(\Delta_r H^*)}{dT} = \frac{d}{dT} \left(\sum_i H_i^* \nu_i \right) \quad (56.52)$$

puisque la variation algébrique de l'enthalpie standard par définition ne dépend que de la température.

Les coefficients stoechiométriques ν_i ne sont pas dépendants de la température tant que cette dernière n'arrive pas à un niveau tel qu'elle change l'essence même de la transformation étudiée.

Nous avons alors dans le cadre de cette approximation:

$$\frac{d(\Delta_r H^*)}{dT} = \frac{d}{dT} \left(\sum_i H_i^* \nu_i \right) = \sum_i \nu_i \frac{dH_i^*}{dT} \quad (56.53)$$

Or, nous avons défini dans le chapitre de Thermodynamique la capacité calorifique à pression constante qui s'écrit:

$$C_p = \left(\frac{\partial Q}{\partial T} \right)_p = \frac{\partial H}{\partial T} \quad (56.54)$$

Donc si les conditions sont standards (l'enthalpie de dépend donc plus que de la température), il vient la différentielle exacte:

$$C_{p,i}^* = \frac{dH_i^*}{dT} \quad (56.55)$$

Nous avons alors:

$$\frac{d(\Delta_r H^\circ)}{dT} = \sum_i \nu_i \frac{dH_i^\circ}{dT} = \sum_i \nu_i C_{P,i}^\circ = \Delta_r C_P^\circ \quad (56.56)$$

Nous pouvons bien évidemment intégrer cette dernière relation pour avoir sa forme commune d'utilisation dans la pratique:

$$\int_{T_0}^T d(\Delta_r H^\circ) = \int_{T_0}^T \Delta_r C_P^\circ dT \quad (56.57)$$

Nous avons alors:

$$\Delta_r H^\circ(T) = \Delta_r H^\circ(T_0) + \int_{T_0}^T \Delta_r C_P^\circ dT$$

où T_0 est une température particulière pour laquelle $\Delta_r H^\circ(T_0)$ est connu.

Dans un domaine de température très proche de T_0 les chimistes approximent parfois la variation comme étant linéaire. Ce qui revient à poser:

$$\Delta_r C_P^\circ = \alpha \quad (56.58)$$

Il vient alors immédiatement de la relation antérieure:

$$\Delta_r H^\circ(T) = \Delta_r H^\circ(T_0) + \alpha \int_{T_0}^T dT = \Delta_r H^\circ(T_0) + \alpha(T - T_0) \quad (56.59)$$

Remarque: Assez souvent, la variation de l'enthalpie de réaction avec la température est négligeable!

Exemple:

Pour la réaction (graphite + oxygène donnant du dioxyde de carbone) nous aimerions connaître $\Delta_r H^\circ$ à 1'000 [K]:



Pour cela, il est donné dans les tables pour cette réaction à 298 [K]:

$$\Delta_r H_m^\circ(298) = -393.5 [kJ \cdot mol^{-1}] \quad (56.61)$$

et:

$$c_{P,C}^\circ = 8.5 [J \cdot K^{-1} \cdot mol^{-1}] \quad c_{P,O_2}^\circ = 29.4 [J \cdot K^{-1} \cdot mol^{-1}] \quad c_{P,CO_2}^\circ = 37.1 [J \cdot K^{-1} \cdot mol^{-1}] \quad (56.62)$$

Remarque: Lorsque l'enthalpie de réaction est donnée à la température de référence (à ce jour...) de 298 [K] les chimistes parlent alors comme nous en avons déjà fait mention plus haut "d'enthalpie standard de formation".

La valeur de l'enthalpie molaire de réaction étant négative, il s'ensuit que la réaction est exothermique (c'est une tendance de la nature à basse température de privilégier les réactions exothermiques pour stabiliser énergétiquement le système).

Nous avons alors immédiatement:

$$\begin{aligned}
 \Delta_r H_m^\circ(T) &= \Delta_r H_m^\circ(T_0) + \int_{T_0}^T \Delta_r c_p^\circ dT = \Delta_r H_m^\circ(T_0) + \int_{T_0}^T (c_{p,C}^\circ - c_{p,O_2}^\circ - c_{p,CO_2}^\circ) dT \\
 &\cong \Delta_r H_m^\circ(T_0) + (c_{p,C}^\circ - c_{p,O_2}^\circ - c_{p,CO_2}^\circ) \int_{T_0}^T dT \\
 &= \Delta_r H_m^\circ(T_0) + (c_{p,C}^\circ - c_{p,O_2}^\circ - c_{p,CO_2}^\circ)(T - T_0) \\
 &= -393'500 + (37.1 - 8.5 - 29.4) \cdot (1'000 - 298) \\
 &= -394'060 [J \cdot mol^{-1}]
 \end{aligned} \tag{56.63}$$

Donc la variation est de $-560 [J \cdot mol^{-1}]$ soit une variation d'environ +0.1%. Il s'ensuit que plus la température augmente, plus la réaction est exothermique. Au fait, le choix de cette température particulière de 1'000 [K] n'est pas innocent car à partir de cette température l'expérience montre que la réaction produit également du monoxyde de carbone.

Nous pouvons aussi conclure que certaines réactions exothermiques et ayant une enthalpie de réaction qui diminue très vite avec la température peuvent s'emballer!

Enfin, indiquons que dans la pratique nous utilisons souvent la notion de "**pouvoir calorifique**" ou "**énergie spécifique de combustion**", qui est au fait simplement... l'enthalpie de réaction par unité de masse de combustible ou l'énergie obtenue par la combustion d'un kilogramme de combustible.

Ainsi, pour l'essence, nous avons selon les tables (sous réserve que ce chiffre soit juste):

$$\Delta_r H^\circ(298) = -47.3 [MJ \cdot kg^{-1}] \tag{56.64}$$

Et nous pouvons nous amuser à calculer la quantité d'essence nécessaire pour accélérer une voiture de 1'000 [kg] à 100 [km/h] avec un rendement à 35%. Nous avons donc:

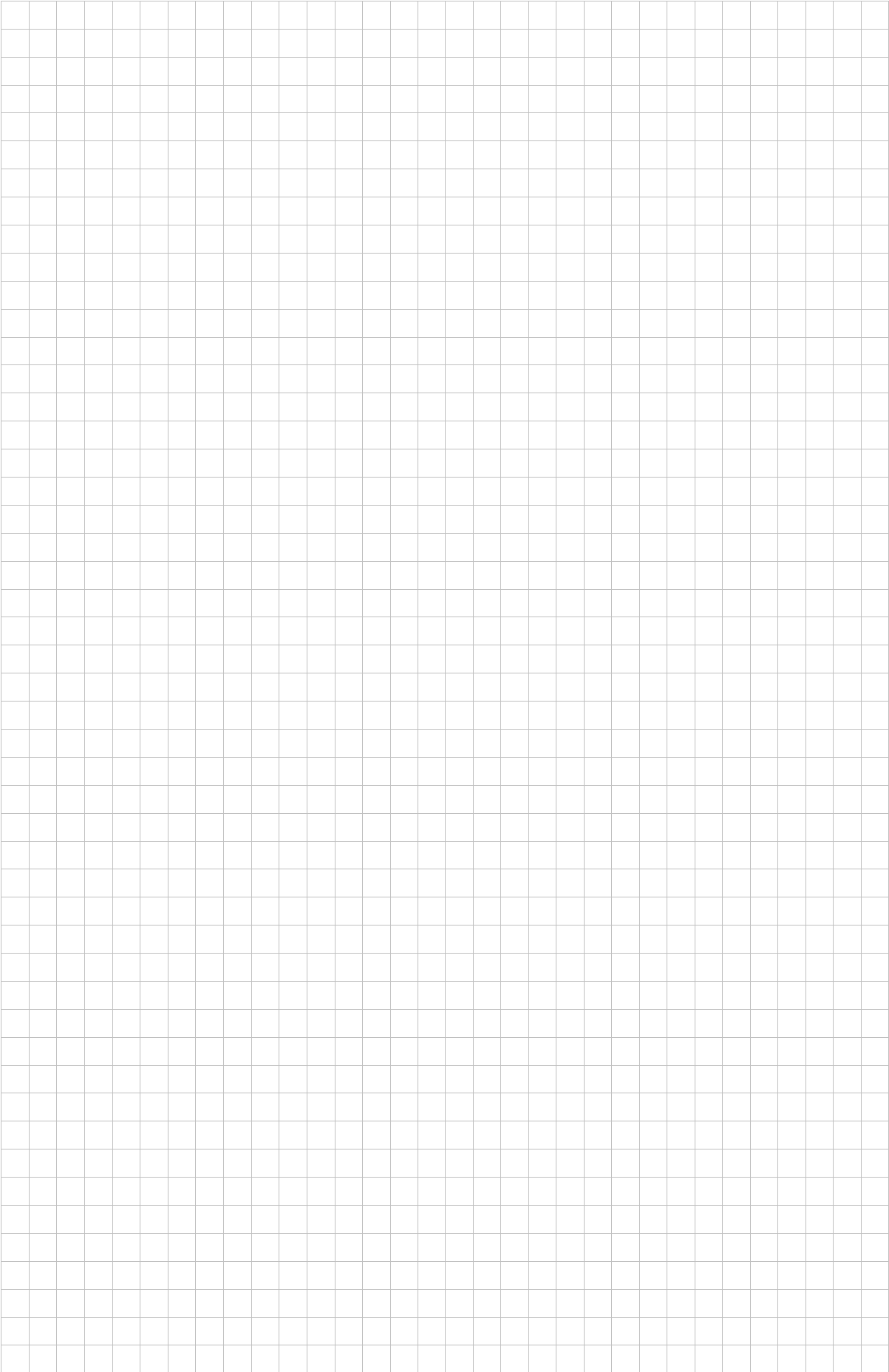
$$M = \frac{\frac{1}{2}mv^2}{\Delta_r H^\circ(298) \cdot \eta} \cong \frac{392 \cdot 10^3}{47.3 \cdot 10^6 \cdot 35\%} = 23.67 [g] \tag{56.65}$$

et pour obtenir la quantité d'essence en litres les tables nous donnent pour l'essence une densité d'environ 700 [kg · m³] ce qui donne au final:

$$L \cong \frac{0.00828}{700} \cdot 1000 = 0.033 [L] \quad (56.66)$$

Notes personnelles:









Informatique théorique

57. MÉTHODES (ANALYSES) NUMÉRIQUES

L'analyse numérique est une discipline des mathématiques. Elle s'intéresse tant aux fondements théoriques qu'à la mise en pratique des méthodes permettant de résoudre, par des calculs purement numériques, des problèmes d'analyse mathématique.

Définition: "L'analyse numérique" est l'étude des algorithmes permettant de résoudre les problèmes de mathématiques continues (distinguées des mathématiques discrètes).

Cela signifie qu'elle s'occupe principalement de répondre numériquement à des questions à variable réelle ou complexe comme l'algèbre linéaire numérique sur les champs réels ou complexes, la recherche de solutions numériques d'équations différentielles et d'autres problèmes liés survenant dans les sciences physiques et l'ingénierie.

Certains problèmes de mathématique continue peuvent être résolus de façon exacte par un algorithme. Ces algorithmes sont appelés alors "méthodes directes". Des exemples sont l'élimination de Gauss-Jordan pour la résolution d'un système d'équations linéaires et l'algorithme du simplexe en programmation linéaire (voir plus loin).

Cependant, pour certains problèmes aucune méthode directe n'est connue (et il est même démontré que pour une classe de problèmes dits "NP complets", il n'existe aucun algorithme fini de calcul direct en temps polynomial). Dans de tels cas, il est parfois possible d'utiliser une méthode itérative pour tenter de déterminer une approximation de la solution. Une telle méthode démarre à partir d'une valeur devinée ou estimée grossièrement et trouve des approximations successives qui devraient converger vers la solution sous certaines conditions. Même quand une méthode directe existe cependant, une méthode itérative peut être préférable car elle est souvent plus efficace et même souvent plus stable (notamment elle permet le plus souvent de corriger des erreurs mineures dans les calculs intermédiaires).

L'utilisation de l'analyse numérique est grandement facilitée par les ordinateurs. L'accroissement de la disponibilité et de la puissance des ordinateurs depuis la seconde moitié du 20ème siècle a permis l'application de l'analyse numérique dans de nombreux domaines scientifiques, techniques et économiques, avec souvent des effets révolutionnaires.

Lors de simulations numériques de systèmes physiques, les conditions initiales sont très importantes dans la résolution d'équations différentielles (voir les différents chapitres du site où apparaissent des effets chaotiques). Le fait que nous ne puissions les connaître avec exactitude implique que les résultats des calculs ne peuvent jamais être parfaitement exacts (nous le savons très bien pour la météo qui en est l'exemple connu le plus flagrant). Cet effet est une conséquence des résultats de la physique fondamentale (basée sur des mathématiques pures) qui démontre que l'on ne peut connaître parfaitement un système en y effectuant des mesures puisqu'elles perturbent directement ce dernier (principe d'incertitude de Heisenberg) et ces perturbations font l'objet de la théorie du Chaos (classique ou quantique).



Avec les nouveaux outils informatiques à disposition en ce début de 21ème siècle, il est devenu pratique et passionnant de connaître les méthodes numériques afin de s'amuser avec certains logiciels (OpenGL, 3D Studio Max, Maple, Matlab, Mathematica, Comsol, etc.) à simuler des systèmes physiques sous forme graphique 2D ou 3D.

Remarques:

R1. Beaucoup de méthodes numériques utilisées en informatique se basent sur des raisonnements qui ont déjà été étudiés dans d'autres sections de ce site, méthodes sur lesquelles nous ne reviendrons pas.

R2. Ce chapitre étant à la limite entre l'ingénierie et la mathématique appliquée, nous avons décidé de donner parfois des exemples d'applications des outils développés.

Définition: Un "**algorithme**" est une suite finie de règles à appliquer dans un ordre déterminé à un nombre fini de données pour arriver, en un nombre fini d'étapes (dont la quantité, ou réciproquement le temps d'exécution est définie par le terme "**coût**"), à un certain résultat, et cela indépendamment du type de données.

Les algorithmes sont intégrés dans des calculateurs par l'intermédiaire de "programmes".

Définition: Un "**programme**" est la réalisation (l'implémentation) d'un algorithme au moyen d'un langage donné (sur une architecture donnée). Il s'agit de la mise en oeuvre du principe.

Axiomes de la programmation (anecdotes):

A1. Plus nous écrivons de code, plus nous produirons d'erreurs.

A2. Il n'existe pas de programmes sans de possibles erreurs (dus au programme lui-même, à l'électronique sous-jacente ou le plus souvent à l'utilisateur même).

Basiquement voici la démarche minimale à suivre lors du développement d'un produit informatique (au niveau du code):

M1. Auditer les besoins présents et anticiper les besoins futurs.

M2. Définir les objectifs.

M3. Calculer la faisabilité, le risque d'erreur, la durée nécessaire au développement.

M4. Créer les algorithmes en langage formel (cela comprend la gestion des erreurs!).

M5. Optimiser la complexité et contrôler les algorithmes.

M6. Choisir une stratégie de développement (modulable ou autre).

M7. Traduire les algorithmes dans la technologie choisie (ce choix est un sujet assez sensible...).

Remarque: Il faudrait dans l'étape (7) veiller à respecter les normes de nommage et de représentation du code ainsi que des conventions (traditions) de fréquence d'insertion des commentaires.

M8. Tests (maintenance préventive)

Remarque: Le débogage (la gestion des erreurs) d'un programme et les tests de fonctionnement doivent prendre autant, voire plus, de temps que le développement du programme lui-même.

Lors du développement d'un programme scientifique, il peut être intéressant, voire même rigoureux de s'attarder sur la notion de "complexité" de ce dernier. Sans aller trop loin, voyons un peu de quoi il s'agit:

1. COMPLEXITÉ

La complexité d'un algorithme est la mesure du nombre d'opérations fondamentales qu'il effectue dans le pire des cas sur un jeu de données. La complexité est donc exprimée comme une fonction de la taille du jeu de données.

Les hypothèses permettant un calcul de cette complexité sont:

H1. $+$ $\equiv$ $*$ $\equiv$ $-$ $\equiv$ $\div$ (les quatre opérations fondamentales ont le même coût)

H2. Un accès mémoire $\ll$ une opération arithmétique

H3. Un contrôle de comparaison $\ll$ une opération arithmétique

H4. Un seul processeur

Nous notons D_n l'ensemble des données de taille n et $T(n)$ le coût (en temps) de l'algorithme sur la donnée ou le jeu de données de taille n .

- La "complexité au meilleur" est donnée par la fonction:

$$T_{\min}(n) = \min_{d \in D_n} C(d) \quad (57.1)$$

C'est le plus petit temps qu'aura à exécuter l'algorithme sur un jeu de données (des lignes de code) de taille fixée, ici à n , dont le coût (la durée) d'exécution est $C(d)$. C'est une borne inférieure de la complexité de l'algorithme sur un jeu de données de taille n .

- La "**complexité au pire**" (la plus intéressante pour le praticien):

$$T_{\max}(n) = \max_{d \in D_n} C(d) \quad (57.2)$$

C'est le plus grand temps qu'aura à exécuter l'algorithme sur un jeu de données de taille fixée, ici à n . Il s'agit donc d'un maximum, et l'algorithme finira toujours avant d'avoir effectué $T_{\max}(n)$ opérations. Cependant, cette complexité peut ne pas refléter le comportement usuel de l'algorithme, le pire cas ne pouvant se produire que très rarement, mais il n'est pas rare que le cas moyen soit quasi aussi mauvais que le pire des cas.

Si l'expression de la complexité au pire d'un algorithme contient plusieurs termes (additions ou soustractions), on ne garde que le terme qui croît le plus vite. Ainsi, un algorithme ayant une complexité du type:

$$An^2 + Bn + n! \quad (57.3)$$

se résumera à avoir une complexité d'ordre $O(n!)$.

- La "**complexité moyenne**":

$$T_{\text{moy}}(n) = \frac{\sum_{d \in D_n} C(d)}{D_n} \quad (57.4)$$

Il s'agit de la moyenne des complexités de l'algorithme sur des jeux de données de taille n (en toute rigueur, il faut bien évidemment tenir compte de la probabilité d'apparition de chacun des jeux de données). Cette moyenne reflète le comportement général de l'algorithme si les cas extrêmes sont rares ou si la complexité varie peu en fonction des données. Cependant, la complexité en pratique sur un jeu de données particulier peut être nettement plus importante que la complexité moyenne; dans ce cas la complexité moyenne ne donnera pas une bonne indication du comportement de l'algorithme.

En pratique, nous ne nous intéressons qu'à la complexité au pire, et à la complexité moyenne.

Définition: Un algorithme est dit "**algorithme optimal**" si sa complexité est de complexité minimale parmi les algorithmes de sa classe.

Comme nous l'avons fait sous-entendre précédemment, nous nous intéressons quasi exclusivement à la complexité en temps des algorithmes. Il est parfois intéressant de s'intéresser à d'autres de leurs caractéristiques, comme la complexité en espace (taille de l'espace mémoire utilisé), la largeur de bande passante requise, etc.

Pour que le résultat de l'analyse d'un algorithme soit pertinent, il faut avoir un modèle de la machine sur laquelle l'algorithme sera implémenté (sous forme de programme). Nous prenons habituellement comme référence, la "machine à accès aléatoire (RAM)" et à processeur unique, où les instructions sont exécutées l'une après l'autre, sans opérations simultanées et sans processus stochastiques (contrairement aux possibles machines quantiques futures).

Les algorithmes usuels peuvent être classés en un certain nombre de grandes classes de complexité dont l'ordre O varie d'une certaine manière:

- Les algorithmes sublinéaires dont la complexité est en général de l'ordre $O(\log(n))$
- Les algorithmes linéaires en complexité $O(n)$ et ceux en complexité en $O(n \log(n))$ sont considérés comme rapides.
- Les algorithmes polynomiaux en $O(n^k)$ pour $k > 3$ sont considérés comme lents, sans parler des algorithmes exponentiels (dont la complexité est supérieure à tout en n) que l'on s'accorde à dire impraticables dès que la taille des données est supérieure à quelques dizaines d'unités.

Remarque: Une bonne complexité est du type $O(n^k)$ pour $k \leq 3$. Une mauvaise complexité est du type $O(e^n)$, $O(n!)$ ou $O(n^n)$.

Exemples:

E1. Évaluation d'un polynôme:

$$P(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n \quad (57.5)$$

L'évaluation directe de la valeur de $P(x)$ conduit à une complexité

$$C_{P(x)} = O\left(\underbrace{\binom{n}{}}_{\text{additions}} + \underbrace{\left(1 + 2 + 3 + \dots + (n+1) = \frac{n(n+1)}{2}\right)}_{\text{multiplications}}\right) = O(n^2) \quad (57.6)$$

Nous devons à Horner un algorithme plus performant qui utilise une factorisation du polynôme sous la forme:

$$P(x) = a_0 + x(a_1 + x(a_2 + \dots x(a_{n-1} + xa_n))) \quad (57.7)$$

Nous pouvons facilement montrer que cette factorisation maintient le même nombre d'additions (n) mais réduit le nombre de multiplications à (n) . La complexité qui en découle est donc $O(n)$. Le gain est incontestablement important. De plus, cette factorisation évite tout calcul de puissance.

E2. Un autre exemple connu de complexité algorithmique est la recherche d'une information dans une colonne triée. Un algorithme simple appelé "recherche dichotomique" consiste à prendre la cellule située à mi-colonne et de voir si nous y trouvons la valeur recherchée. Sinon, la recherche doit continuer sur le même mode opératoire dans la partie supérieure ou inférieure du tableau (cela dépend de l'ordre lexicographique).

L'algorithme est récursif et permet à chaque étape, de diviser par deux, la taille de l'espace de recherche. Si cette taille, initialement, est de n cellules dans la colonne, elle est de $n/2$ à l'étape 1, $n/2^2$ à l'étape 2, et plus généralement de $n/2^k$ à l'étape k .

Au pire, la recherche se termine quand il n'y a plus qu'une seule cellule de la colonne à explorer, autrement dit quand k est tel que $n < 2^k$. Nous en déduisons le nombre maximal d'étapes: c'est le plus petit k tel que $n < 2^k$, soit $\log_2(n) < k$, soit:

$$T_{\max}(n) = k = \log_2(n) \quad (57.8)$$

à comparer avec une recherche séquentielle (utile lorsque le tri est trop coûteux en ressources). Par exemple, dans une colonne de 25'000 données dont la complexité linéaire est $O(n)$ soit 25'000 alors qu'avec la méthode dichotomique, nous avons une complexité sublinéaire $\log_2(25000) \cong 15$. Le gain est donc considérable (à condition que les données soient triées)!

E3. Soient A et B deux matrices carrées de dimension n . Les principales opérations sur ces matrices ont des ordres de complexité suivants (nous laisserons le soin au lecteur de vérifier car c'est vraiment trivial):

- Lecture (itérations): $O(n^2)$

- Calcul de la trace: $tr(A) = \sum_{i=1}^n a_{ii} \rightarrow O(n)$

- Addition: $A + B = C$ telle que $c_{ij} = a_{ij} + b_{ij} \rightarrow O(n^2)$

- Multiplication: $A \cdot B = C$ telle que $c_{ij} = \sum_{k=1}^n a_{ik} \cdot b_{kj} \rightarrow O(n^3)$

- Déterminant (par la méthode directe de Cramer). Nous renvoyons le lecteur au chapitre d'Algèbre Linéaire pour le détail des méthodes de calcul du déterminant d'une matrice. Nous pouvons alors montrer que la complexité d'un déterminant d'ordre n est n multiplications, $n-1$ additions plus n fois la complexité d'un déterminant d'ordre $n-1$. Par cumul, nous arrivons à:

$$\det(A) \rightarrow O(n \cdot n!) \quad (57.9)$$

En faisant l'hypothèse que l'ordinateur utilisé effectue une opération élémentaire en 10^{-9} secondes (ce qui est déjà une bonne machine), nous obtenons alors les temps de calculs suivants pour plusieurs valeurs de n :

n	5	10	15	20	50
t	$6 \cdot 10^{-7}[s]$	$0.04[s]$	$5.5[h]$	$1543[ans]$	$4.8 \cdot 10^{46}[ans]$

Tableau: 57.1 - Temps de calcul d'un déterminant

d'où la nécessité de faire un calcul de complexité avant de mettre l'algorithme en route (à moins que vous ne travailliez exclusivement pour les générations futures, à condition qu'il y en ait encore...).

Finalement, pour résumer un peu, nous distinguons quelques types de complexités classiques: $O(1)$ indépendant de la taille de la donnée, $O(\log(n))$ complexité logarithmique, $O(n)$ complexité linéaire, $O(n \log(n))$ complexité quasi-linéaire, $O(n^2)$ complexité quadratique, $O(n^3)$ complexité cubique, $O(k^n)$ complexité exponentielle, $O(n!)$ complexité factorielle.

1.1. NP-COMPLÉTUDE

Nous allons introduire pour la culture générale le concept de NP-complétude, c'est-à-dire que nous allons tenter de définir sans trop de formalisme (comme à l'habitude) la classe des problèmes dit "**NP-complets**". Ces problèmes sont ceux pour lesquels personne à l'heure actuelle ne connaît d'algorithme efficace (i.e. seulement des algorithmes exponentiels). Ce sont des problèmes vraiment difficiles par opposition à ceux pour lesquels nous connaissons des algorithmes de complexité polynomiale.

Définitions:

D1. Les problèmes de "**classe P**" sont de bons problèmes dans le sens où le calcul de leur solution est faisable dans un temps raisonnable avec un algorithme à complexité polynomiale. Plus formellement, ce sont les problèmes pour lesquels nous pouvons construire une machine déterministe (voir la remarque après les définitions) dont le temps d'exécution est de complexité polynomiale (le sigle "P" signifiant "Polynomial Time").

D2. Les problèmes de "**classe E**" sont des problèmes dans le sens où le calcul de leur solution est faisable dans un temps exponentiel par nature du type k^n . Plus formellement, ce sont les problèmes pour lesquels nous pouvons construire une machine déterministe dont le temps d'exécution est de complexité exponentielle (le sigle "E" signifiant "Exponential Time").

D3. Les problèmes de la "**classe NP**" sont ceux pour lesquels nous pouvons construire une machine de Turing non déterministe (voir la remarque après les définitions) dont le temps d'exécution est de complexité polynomiale (le sigle "NP" provient de "Nondeterministic Polynomial time" et non de "Non Polynomial").

Remarque: Contrairement aux machines déterministes qui exécutent une séquence d'instructions bien déterminée, les machines non déterministes ont la remarquable capacité de toujours choisir la meilleure séquence d'instructions qui mène à la bonne réponse lorsque celle-ci existe. Il va sans dire qu'une telle machine ne peut pas exister autrement que dans l'esprit d'un théoricien (à moins qu'avec les ordinateurs quantiques...)! Néanmoins, comme nous le verrons par la suite, ce concept abstrait n'est pas sans intérêt et constitue en fait la base de toute la théorie de la NP-complétude.

Il importe de remarquer à ce stade de la discussion que la classe P est incluse dans la classe NP, nous écrivons $P \subseteq NP$, car si nous pouvons construire une machine déterministe pour résoudre efficacement (en un temps au pire polynomial) un problème, nous pouvons certainement (du moins dans notre esprit) en construire une non déterministe qui résout aussi efficacement le même problème. Par ailleurs, il ne faut pas croire que ce concept de divination optimale qu'est la machine non déterministe permet de tout résoudre puisqu'il existe en informatique théorique d'autres types de problèmes n'appartenant pas à la classe NP qui sont les problèmes indécidables.

Pour savoir si un problème donné appartient ou non à la classe NP, il s'agit simplement de l'exprimer sous une forme dont la réponse est soit OUI, soit NON. Le problème appartient alors à la classe NP si par définition, nous arrivons à démontrer à l'aide d'un algorithme de complexité polynomiale que n'importe quelle instance OUI du problème est bel et bien correcte. Nous n'avons pas à nous préoccuper des instances NON du problème puisque la machine non déterministe, par définition, prend toujours la bonne décision (lorsque celle-ci existe).

Par exemple, le problème consistant à trouver un cycle hamiltonien (cycle qui passe une et une seule fois par tous les sommets du graphe - voir chapitre de Théorie Des Graphes) dans un graphe appartient à NP puisque, étant donné un cycle, il est trivial de vérifier en temps linéaire qu'il contient bien une et une seule fois chaque sommet.

Un autre exemple de problème difficile des mathématiques est la factorisation d'un entier en produit de facteurs premiers. Nous ne savons pas à ce jour s'il existe un algorithme polynomial qui puisse réussir cette opération. Autrement dit, nous ne savons pas si ce problème est dans P. En revanche, étant donnés des nombres premiers $p_1, \dots, p_k$ il est trivial de vérifier que $n = p_1 \dots p_k$: ce problème est donc dans NP. Il semblerait (nous n'avons pas vérifié ce résultat et ni la possibilité de le faire) que la complexité du meilleur algorithme de factorisation en nombres premiers soit du type:

$$O\left(\exp\left[\left(\frac{64}{9}\log(n)\right)^{1/3} \left(\log(\log(n))\right)^{2/3}\right]\right) \quad (57.10)$$

il resterait donc du travail à faire (si un internaute pouvait nous fournir les détails qui ont amené à ce résultat, nous sommes preneurs).

Remarque: Si un problème est dans NP, alors il existera un algorithme en temps exponentiel pour le résoudre mais le contraire n'est pas toujours vrai (il faut donc être prudent).

Parmi l'ensemble des problèmes NP, il en existe un sous-ensemble qui contient les problèmes les plus difficiles: nous les appelons les problèmes "**NP-complet**" (N.P.C.). Ainsi, un problème NP-complet possède la propriété que tout problème dans NP peut être transformé en celui-ci en un temps polynomial.

La raison essentielle pour laquelle les problèmes NPC sont les plus difficiles parmi les problèmes de NP est que ces premiers peuvent toujours servir comme des sous-routines pour solutionner ces derniers. Cette réduction à une ou des sous-routines est assez facilement faisable puisque réalisable, si elle existe, en un temps polynomial. Un problème NPC est donc complet en ce sens qu'il contient l'essentiel de la complexité des problèmes appartenant à NP, et qu'une solution polynomiale à ce problème implique une solution polynomiale à tous les problèmes NP.

Autrement formulé, les problèmes NPC ont une complexité exponentielle et ils ont tous la même classe de complexité (modulo les polynômes).

Finalement, ce qu'il importe de bien comprendre et de retenir de toute cette théorie, son idée maîtresse, est que si nous trouvons un jour un algorithme de complexité polynomiale pour un seul de ces problèmes vraiment difficiles que sont les problèmes NPC, alors d'un seul coup NP devient égal à P et tous les problèmes difficiles deviennent faciles!

Pour résumer, être dans P, c'est trouver une solution en un temps polynomial, tandis qu'être dans NP, c'est prouver en un temps polynomial qu'une proposition de réponse est une solution du problème. Ainsi, tout problème qui est dans P se trouve dans NP. Un champ de recherche majeur des mathématiques actuelles est l'investigation de la réciproque: et donc finalement a-t-on $P=NP$? Autrement dit, peut-on trouver en un temps polynomial ce que l'on peut prouver en temps polynomial?

Remarque: Ce problème est si important en informatique qu'il fait partie (arbitrairement) des 7 problèmes du millénaire, dont la résolution est primée 1 million de dollars par le Clay Mathematic Institute.

Passons maintenant à l'étude de quelques applications types des méthodes numériques dont il est très souvent fait usage dans l'industrie. Nous irons du plus simple au plus compliqué et sans oublier que beaucoup de méthodes ne se trouvant pas dans ce chapitre peuvent parfois être trouvées dans d'autres sections du site!

2. PARTIE ENTIÈRE

Le plus grand entier inférieur ou égal à un nombre réel x est exprimé par $[x]$, qui se lit "**partie entière de x** ".

Ainsi, le nombre M est entier si et seulement si $[M]=M$. De même, le naturel A est divisible dans l'ensemble des naturels par le naturel B si et seulement si:

$$\left[\frac{A}{B} \right] = \frac{A}{B} \quad (57.11)$$

Nous notons aussi $\{x\}$ pour désigner la partie fractionnaire de x . Nous avons ainsi:

$$\{x\} = x - [x] \quad (57.12)$$

Soit $x, y \in \mathbb{R}$. Alors nous avons les propriétés suivantes:

$$P1. [x] \leq x < [x] + 1, 0 \leq x - [x] < 1, x = [x] + \theta \text{ où } 0 \leq \theta < 1$$

$$P2. [x] = \sum_{n \leq x} 1, \text{ lorsque } x \geq 0$$

$$P3. [x + m] = [x] + m, \text{ si } m \in \mathbb{Z}$$

$$P4. [x] + [y] \leq [x + y] \leq [x] + [y] + 1$$

$$P5. [-x] = -[x] \text{ si } x \in \mathbb{Z}, [-x] = -[x] - 1 \text{ si } x \notin \mathbb{Z}$$

$$P6. \left[\frac{[x]}{m} \right] = \left[\frac{x}{m} \right] \text{ si } m \in \mathbb{N}$$

P7. Si $a \in \mathbb{N}$, alors $[x/a]$ représente le nombre d'entiers positifs inférieurs ou égaux à x qui sont divisibles par a .

Démonstrations:

La première partie de P1 est simplement la définition de $[x]$ sous forme algébrique. Les deux autres parties sont des réarrangements de la première partie. Dans ce cas, nous pouvons écrire

$$x = [x] + \theta \quad (57.13)$$

où $0 \leq \theta < 1$.

Pour P2, la somme est vide (ne contient aucun terme autrement dit) pour $x < 1$ et, dans ce cas, on adopte la convention selon laquelle la somme vaut 0. Alors, pour $x \geq 1$, la somme compte le nombre d'entiers positifs n qui sont plus petits ou égaux à x . Ce nombre est évidemment $[x]$.

La démonstration de P3 sera supposée évidente.

Pour prouver P4, nous écrivons:

$$x = n + \theta, y = m + \phi \quad (57.14)$$

où n et m sont des entiers et où $0 \leq \theta < 1$ et $0 \leq \phi < 1$. Alors:

$$\begin{aligned} [x] + [y] &= n + m = [n + m] \leq [n + \theta + m + \phi] = [x + y] \\ &= n + m + [\theta + \phi] \leq n + m + 1 = [x] + [y] + 1 \end{aligned} \quad (57.15)$$

En écrivant $x = n + \theta$, où $0 \leq \theta < 1$, nous avons

$$-x = -n - 1 + 1 - \theta \quad (57.16)$$

où $0 < 1 - \theta \leq 1$.

Il s'ensuit que:

$$[x] + [-x] = n + [-n - 1 + 1 - \theta] = n - n - 1 + [1 - \theta] = 0 \quad (57.17)$$

si $\theta = 0$ et:

$$[x] + [-x] = n + [-n - 1 + 1 - \theta] = n - n - 1 + [1 - \theta] = -1$$

si $0 < \theta < 1$. Et nous obtenons du même coup la démonstration P5.

Pour démontrer P6, nous écrivons:

$$x = [x] + \theta \quad (57.18)$$

où $0 \leq \theta < 1$, et:

$$n = qm + r \quad (57.19)$$

où $0 \leq r < m - 1$. Nous obtenons ainsi:

$$\left[\frac{x}{m} \right] = \left[\frac{qm + r + \theta}{m} \right] = q + \left[\frac{r + \theta}{m} \right] = q \quad (57.20)$$

puisque $0 \leq r + \theta \leq m$. Par ailleurs:

$$\left[\frac{[x]}{m} \right] = \left[\frac{n}{m} \right] = \left[q + \frac{r}{m} \right] = q \quad (57.21)$$

et nous avons ainsi le résultat.

Pour la dernière partie, nous observons que, si $a, 2a, \dots, ma$ sont tous les entiers positifs $\leq x$ qui sont divisibles par a , il suffit de prouver que $[x/a] = m$. Puisque $(m+1)a > x$, alors:

$$ma \leq x < (m+1)a \quad (57.22)$$

c'est-à-dire:

$$m \leq \frac{x}{a} < m+1 \quad (57.23)$$

soit le résultat attendu.

□C.Q.F.D.

Remarque: La méthode d'arrondi de valeurs réelles est donnée dans le chapitre d'Économie.

3. ALGORITHME D'HÉRON

Soit à calculer la racine carrée:

$$x = \sqrt{A} \quad (57.24)$$

Il existe un algorithme dit "**algorithme d'Héron**" ou encore "**algorithme d'Héron d'Alexandrie**" qui permet de calculer la valeur de cette racine carrée.

Démonstration:

Voici une pseudo-démonstration car historiquement l'algorithme a été construit sur des considérations purement intuitives (car 100 ans avant J.C. l'algèbre n'existait pas...). Dans les classes le résultat est donné en tant que définition et on en étudie la convergence.

Donc pour la démonstration, nous procéderons ainsi:

$$\begin{aligned} x^2 &= A \\ \Rightarrow 2x^2 &= x^2 + A \\ \Rightarrow 2x &= (x^2 + A) / x \\ \Rightarrow x &= (x^2 + A) / (2x) = (x + A/x) / 2 \end{aligned} \quad (57.25)$$

Et l'astuce consiste à poser que:

$$x_{i+1} = \left(\frac{x_i}{2} + \frac{A}{2x_i} \right) \quad (57.26)$$

où il faut prendre comme valeur initiale $x_1 = A$ (ce qui évidemment est en totale contradiction avec les développements mais l'on voit alors mieux pourquoi c'est une pseudo-démonstration...).

□C.Q.F.D.

Exemple:

Soit à calculer:

$$\sqrt{10} = 3.162277660168... \quad (57.27)$$

Nous prendrons donc $A = 10$ et cela nous donne le tableau d'itérations suivant:

Itération	$x_i/2$	$A/2x_i$	x_{i+1}	Écart
1	5	0.5	5.50	~ 2.3
2	2.750	0.90909	3.659 090 909	~ 0.49
3	1.82954	1.3664	3.196 005 083	~ 0.033
4	1.59800	1.5644	3.162 455 624	~ 0.0002
5	1.58122	1.5810	3.162 277 665	$\sim 0.5 \cdot 10^{-8}$

Tableau: 57.2 - Itérations pour l'algorithme d'Héron

Dans le cas de la racine cubique, la démonstration est semblable et nous obtenons:

$$x_{i+1} = \left(\frac{2x_i}{3} + \frac{A}{3x_i^2} \right) \quad (57.28)$$

Signalons encore que le lecteur pourra trouver dans le chapitre de Théorie des Nombres la méthode utilisée pendant l'antiquité (du moins une analogie) et utilisant les fractions continues.

4. ALGORITHME D'ARCHIMÈDE

Le calcul de la constante universelle "pi" notée π est très certainement le plus grand intérêt de l'algorithmique puisque l'on retrouve cette constante un peu partout en physique et mathématique (nous pouvons vous conseiller un très bon ouvrage sur le sujet).

Nous rappelons que nous n'en avons pas donné la valeur ni en géométrie, ni dans les autres sections de ce site jusqu'à maintenant. Nous allons donc nous atteler à cette tâche.

Nous définissons en géométrie le nombre π dit "**pi**", quel que soit le système métrique utilisé (ce qui fait son universalité), comme le rapport de la moitié du périmètre d'un cercle par son rayon tel que:

$$\pi = \frac{P/2}{R} \quad (57.29)$$

Nous devons le premier algorithme du calcul de cette constante à Archimède (287-212 av. J.-C.) dont voici la démonstration.

Démonstration:

Soit un n -polygone inscrit dans un cercle:

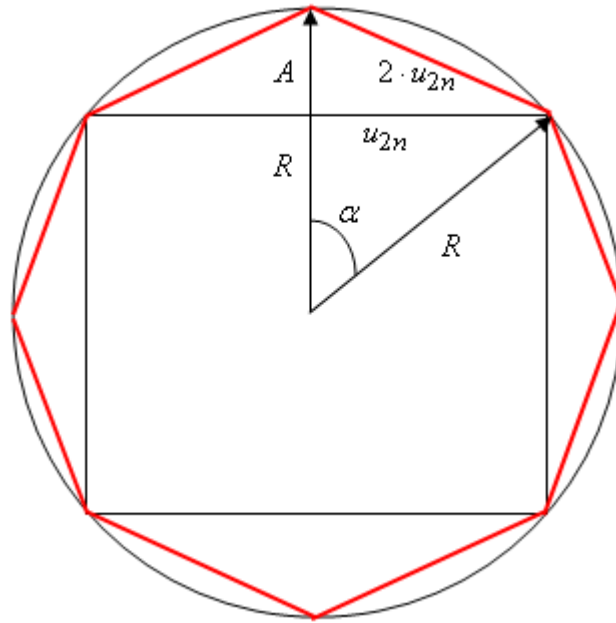


Figure: 57.1 - Principe illustré de l'algorithme d'Archimède

Le principe de l'algorithme d'Archimède est le suivant: Soit P_n le périmètre d'un polygone régulier de n côtés inscrit dans un cercle de rayon $1/2$. Archimède arrive à montrer par induction que:

$$P_n = 2 \cdot n \cdot u_n \underset{n \rightarrow \infty}{=} 2 \cdot R \cdot \pi = 2 \cdot \frac{1}{2} \cdot \pi = \pi \quad (57.30)$$

Nous avons pour le périmètre d'un n -polygone:

$$u_n = R \sin \alpha = \frac{1}{2} \sin \left(\frac{2\pi}{n} \right) = \frac{1}{2} \sin \left(\frac{\pi}{n} \right) \text{ et } (2u_{2n})^2 = u_n^2 + A^2 \quad (57.31)$$

Avec:

$$A = R - R \cos \alpha = \frac{1}{2} - \frac{1}{2} \cos \left(\frac{\pi}{n} \right) \quad (57.32)$$

Donc:

$$\begin{aligned} 4 \cdot u_{2n}^2 &= \left[\frac{1}{2} \sin \left(\frac{\pi}{n} \right) \right]^2 + \left[\frac{1}{2} - \frac{1}{2} \cos \left(\frac{\pi}{n} \right) \right]^2 = \frac{1}{4} \sin^2 \left(\frac{\pi}{n} \right) + \frac{1}{4} - \frac{1}{2} \cos \left(\frac{\pi}{n} \right) + \frac{1}{4} \cos^2 \left(\frac{\pi}{n} \right) \\ &= \frac{1}{4} \left[\sin^2 \left(\frac{\pi}{n} \right) + \cos^2 \left(\frac{\pi}{n} \right) \right] + \frac{1}{4} - \frac{1}{2} \cos \left(\frac{\pi}{n} \right) = \frac{2}{4} - \frac{1}{2} \cos \left(\frac{\pi}{n} \right) = \frac{1}{2} - \frac{1}{2} \sqrt{1 - \sin^2 \left(\frac{\pi}{n} \right)} \\ &= \frac{1}{2} \left(1 - \sqrt{1 - \sin^2 \left(\frac{\pi}{n} \right)} \right) = \frac{1}{2} \left(1 - \sqrt{1 - 4u_n^2} \right) \\ \Rightarrow u_{2n} &= \frac{1}{2} \sqrt{\frac{1}{2} \left(1 - \sqrt{1 - 4u_n^2} \right)} \end{aligned} \quad (57.33)$$

Il suffit d'un ordinateur ensuite et de plusieurs itérations pour évaluer avec une bonne précision la valeur de π . Évidemment, on utilise l'algorithme d'Héron pour calculer la racine...

□C.Q.F.D.

Remarque: Il existe un très grand nombre d'algorithmes pour calculer π . Celui présenté ci-dessus, sans être le plus esthétique, serait historiquement le premier.

5. CALCUL DU NOMBRE D'EULER

Hors la constante π , il existe d'autres constantes mathématiques importantes qu'il faut pouvoir générer à l'ordinateur (de nos jours les valeurs constantes sont stockées telles quelles et ne sont plus recalculées systématiquement). Parmi celles-ci, se trouve le "nombre d'Euler" noté e (cf. chapitre d'Analyse Fonctionnelle). Voyons comment calculer ce nombre:

Soit la série de Taylor (cf. chapitre sur les Suites Et Séries), pour une fonction indéfiniment dérivable f donnée par:

$$f(u) = \sum_{n=0}^{\infty} \frac{d^n f(0)}{du^n} \frac{u^n}{n!} \quad (57.34)$$

Comme (cf. chapitre de Calcul Différentiel Et Intégral):

$$\frac{d^n (e^u)}{du^n} = e^u \quad (57.35)$$

nous avons:

$$f(u) = e^u = \sum_{n=0}^{\infty} \frac{d^n f}{du^n}(0) \frac{u^n}{n!} = \sum_{n=0}^{\infty} \frac{d^n (e^u)}{du^n}(0) \frac{u^n}{n!} = \sum_{n=0}^{\infty} 1 \frac{u^n}{n!} = \sum_{n=0}^{\infty} \frac{u^n}{n!} \quad (57.36)$$

Donc en résumé:

$$e^u = \sum_{n=0}^{\infty} \frac{u^n}{n!} \quad (57.37)$$

Cette relation donne un algorithme pour calculer l'exponentielle e^u à un ordre n donné de précision.

Remarque: Pour diminuer la complexité de cet algorithme, la factorielle peut être calculée avec la formule exposée ci-après.

6. CALCUL DE LA FACTORIELLE (FORMULE DE STIRLING)

Évidemment, la factorielle pourrait être calculée avec une simple itération. Cependant, ce genre de méthode génère un algorithme à complexité exponentielle ce qui n'est pas le mieux. Il existe alors une autre méthode:

Soit, la définition de la factorielle:

$$n! = n(n-1)(n-2)\dots 1 \quad (57.38)$$

Et d'après les propriétés des logarithmes:

$$\ln(n!) = \sum_{m=1}^n \ln(m) \quad (57.39)$$

Si n est très grand (mais alors très...) alors:

$$\ln(n!) = \int_1^n \ln(m) dm = (m \ln(m) - m) \Big|_1^n \quad (57.40)$$

Lorsque $n \gg 1$, la limite inférieure est négligeable et alors (approximation qui nous est très utile dans le chapitre de Mécanique Statistique):

$$\ln(n!) \cong n \ln(n) - n \Rightarrow n! = e^{n(\ln(n)-1)} \quad (57.41)$$

Après une petite simplification élémentaire, nous obtenons:

$$n! = e^{n \ln(n)} e^{-n} = \left(\frac{n}{e} \right)^n \quad (57.42)$$

Cette dernière relation est utile si l'on suppose bien évidemment que la constante d'Euler est une valeur stockée dans la machine...

7. SYSTÈMES D'ÉQUATIONS LINÉAIRES

Il existe de nombreuses méthodes de résolution de systèmes d'équations linéaires. La plupart d'entre elles ont été mises au point pour traiter des systèmes particuliers. Nous en étudierons une, appelée la "**méthode de réduction de Gauss**" ou "**pivot de Gauss**", qui est bien adaptée à la résolution des petits systèmes d'équations (jusqu'à 50 inconnues).

Remarques:

R1. La validité de certaines des opérations que nous allons effectuer ici pour résoudre les systèmes linéaires se démontre dans le chapitre traitant de l'Algèbre Linéaire. Au fait, pour être bref, le tout fait appel à des espaces vectoriels dont les vecteurs-colonnes sont linéairement indépendants.

R2. Les systèmes admettent une solution si et seulement si (rappel) le rang de la matrice augmentée est inférieur ou égal au nombre d'équations (cf. chapitre d'Algèbre Linéaire).

7.1. UNE ÉQUATION À UNE INCONNUE

Considérons le cas le plus simple: celui d'une équation à une inconnue:

$$ax = b \quad (57.43)$$

a et b sont les coefficients de l'équation et x en est l'inconnue. Résoudre cette équation, c'est déterminer x en fonction de a et b . Si a est différent de 0 alors:

$$x = b/a \quad (57.44)$$

est la solution de l'équation. Si a est nul et si b est différent de 0 alors l'équation n'admet pas de solution. Si a et b sont nuls, alors l'équation possède une infinité de solutions.

7.2. DEUX ÉQUATIONS À DEUX INCONNUES

Un système (linéaire) de deux équations à deux inconnues s'écrit:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1 \\ a_{21}x_1 + a_{22}x_2 &= b_2 \end{aligned} \quad (57.45)$$

$a_{11}, a_{12}, b_1, a_{21}, a_{22}, b_2$ sont les coefficients du système d'équations, x_1 et x_2 en sont les inconnues.

Remarque: Les notations usitées ci-dessus n'ont rien à voir avec le calcul tensoriel.

Pour résoudre le système, nous procédons comme suit:

À l'aide de manipulations algébriques (addition ou soustraction des différentes égalités entre elles - manipulations autorisées par l'indépendance linéaire des vecteurs-lignes) nous transformons le système en un autre donné par:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1 \\ 0 \cdot x_1 + a'_{22}x_2 &= b'_2 \end{aligned} \quad (57.46)$$

Ensuite, nous résolvons l'équation $a'_{22}x_2 = b'_2$, ce qui donne:

$$x_2 = \frac{b'_2}{a'_{22}} \quad (57.47)$$

Nous en déduisons:

$$x_1 = \frac{(b_1 - a_{12}x_2)}{a_{11}} \quad (57.48)$$

La transformation entre les deux systèmes:

$$\left. \begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1 \\ a_{21}x_1 + a_{22}x_2 &= b_2 \end{aligned} \right\} \Rightarrow \left\{ \begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1 \\ 0 \cdot x_1 + a'_{22}x_2 &= b'_2 \end{aligned} \right. \quad (57.49)$$

s'effectue simplement en multipliant chaque coefficient de la première égalité par a_{21}/a_{11} et en soustrayant les résultats obtenus des coefficients correspondants de la seconde égalité. Dans ce cas, l'élément a_{11} est appelé "**pivot**".

7.3. TROIS ÉQUATIONS À TROIS INCONNUES

Examinons encore le cas des systèmes de trois équations à trois inconnues:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned} \quad (57.50)$$

Nous pouvons par la suite des opérations élémentaires (cf. chapitre d'Algèbre Linéaire) réduire le système linéaire précédent en le système suivant:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ 0 \cdot x_1 + a''_{22}x_2 + a''_{23}x_3 &= b''_2 \\ 0 \cdot x_1 + 0 \cdot x_2 + a''_{33}x_3 &= b''_3 \end{aligned} \quad (57.51)$$

et ensuite résoudre l'équation:

$$x_3 = \frac{b''_3}{a''_{33}} \quad (57.52)$$

puis la deuxième:

$$x_2 = \frac{(b''_2 - a''_{23}x_3)}{a''_{22}} \quad (57.53)$$

et enfin:

$$x_1 = \frac{(b_1 - a_{13}x_3 - a_{12}x_2)}{a_{11}} \quad (57.54)$$

Revenons à la transformation des systèmes. Elle s'effectue en deux étapes:

1. Dans la première, nous choisissons a_{11} comme pivot et nous éliminons les coefficients a_{21} et a_{31} de la manière suivante: il faut multiplier chaque coefficient de la première égalité par a_{21}/a_{11} et soustraire les résultats obtenus de la deuxième égalité, ainsi a_{21} devient nul. De même, en multipliant les coefficients de la première équation par a_{31}/a_{11} et en soustrayant les résultats obtenus de la troisième égalité, a_{31} disparaît. Le système d'équation s'écrivant alors:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ 0 \cdot x_1 + a'_{22}x_2 + a'_{23}x_3 &= b'_2 \\ 0 \cdot x_1 + a'_{32}x_2 + a'_{33}x_3 &= b'_3 \end{aligned} \quad (57.55)$$

2. La deuxième étape consiste à traiter le système de deux équations à deux inconnues formé des deuxième et troisième égalités du système précédent, et ce, en choisissant a'_{22} comme pivot. Cette méthode de résolution peut paraître compliquée, mais elle a l'avantage de pouvoir être généralisée et être appliquée à la résolution de systèmes de n équations à n inconnues.

7.4. N ÉQUATIONS À N INCONNUES

Pour simplifier l'écriture, les coefficients seront toujours notés a_{ij} et non pas a'_{ij}, a''_{ij} , etc. lors de chaque étape du calcul.

Soit le système linéaire (nous pourrions très bien le représenter sous la forme d'une matrice augmentée afin d'alléger les écritures):

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= b_2 \\ \dots & \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n &= b_n \end{aligned} \quad (57.56)$$

Il faut choisir a_{11} comme pivot pour éliminer $a_{21}, a_{31}, \dots, a_{n1}$. Ensuite, l'élimination de $a_{32}, a_{42}, \dots, a_{n2}$ s'effectue en prenant a_{22} comme pivot. Le dernier pivot à considérer est bien évidemment $a_{n-1, n-1}$, il permet d'éliminer a_{nn} . Le système prend alors la forme:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1 \\ 0 \cdot x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= b_2 \\ \dots & \\ 0 \cdot x_1 + 0 \cdot x_2 + \dots + a_{nn}x_n &= b_n \end{aligned} \quad (57.57)$$

Et de résoudre d'abord la dernière équation, puis l'avant-dernière et ainsi de suite jusqu'à la première.

Cette méthode, appelée "**méthode de résolution de Gauss**" ou encore "**méthode du pivot**" doit cependant être affinée pour éviter les pivots de valeur 0. L'astuce consiste à permuter l'ordre dans lequel sont écrites les équations pour choisir comme pivot le coefficient dont la valeur absolue est la plus grande. Ainsi, dans la première colonne, le meilleur pivot est l'élément a_{j1} tel que:

$$|a_{j1}| > |a_{i1}| \quad \forall i \quad (57.58)$$

Il est amené à a_{11} par échange des première et j-ème lignes. L'élimination du reste de la première colonne peut alors être effectuée. Ensuite, on recommence avec les $n-1$ équations qui restent.

8. POLYNÔMES

L'ensemble des polynômes à coefficients réels a été étudié dans le chapitre d'Analyse Fonctionnelle en détails. Nous allons ici traiter de l'aspect numérique de quelques problèmes liés aux polynômes.

Mises à part l'addition et la soustraction de polynômes que nous supposons comme triviales (optimisation de la complexité mise à part comme le schéma de Horner), nous allons voir comment multiplier et diviser deux polynômes.

Voyons d'abord comment multiplier deux polynômes:

Soient:

$$\begin{aligned} f(x) &= a_0 + a_1x + a_2x^2 + \dots + a_nx^n \\ g(x) &= b_0 + b_1x + b_2x^2 + \dots + b_mx^m \end{aligned} \quad (57.59)$$

alors:

$$\begin{aligned} f(x) \cdot g(x) &= a_0b_0 + (a_0b_1 + a_1b_0)x + (a_0b_2 + a_1b_1 + a_2b_0)x^2 + \dots \\ f(x) \cdot g(x) &= c_0 + c_1x + c_2x^2 + \dots \end{aligned} \quad (57.60)$$

avec:

$$c_k = \sum_{i=0}^k a_i b_{k-i} \quad (k = 0, 1, 2, \dots, n+m) \quad (57.61)$$

C'était simple...

Un tout petit peu plus difficile maintenant: la division euclidienne de polynômes (cf. chapitre de Calcul Algébrique).

Reprenons:

$$\begin{aligned} f(x) &= a_0 + a_1x + a_2x^2 + \dots + a_nx^n \\ g(x) &= b_0 + b_1x + b_2x^2 + \dots + b_mx^m \end{aligned} \quad (57.62)$$

mais en imposant cette fois-ci $n \geq m$.

La division s'écrit donc nous le savons:

$$f(x) = g(x)q(x) + r(x) \quad (57.63)$$

avec:

$$\deg(r(x)) < \deg(g(x)) \quad (57.64)$$

ou sinon $r(x) = 0$.

Il est normalement connu d'avance (car démontré dans le chapitre de Calcul Algébrique) que nous avons:

$$\deg(q(x)) = n - m \quad (57.65)$$

et:

$$\deg(r(x)) < m \quad (57.66)$$

Nous avons donc par définition $q(x)$ qui est le quotient de la division et $r(x)$ le reste de la division euclidienne de $f(x)$ par $g(x)$.

Rien ne nous interdit donc de poser de la manière la plus générale qui soit:

$$\begin{aligned} q(x) &= q_0 + q_1x + \dots + q_{n-m}x^{n-m} \\ r(x) &= r_0 + r_1x + \dots + r_{m-1}x^{m-1} \end{aligned} \quad (57.67)$$

Pour démontrer l'expression des différents q_i , nous avons préféré pour des raisons pédagogiques de passer par un exemple particulier (voir ci-dessous) dont nous généraliserons le résultat.

Exemple:

Soient:

$$\begin{aligned} f(x) &= a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 \\ g(x) &= b_0 + b_1x + b_2x^2 \end{aligned} \quad (57.68)$$

donc de ce que nous avons dit précédemment, il vient (point de départ):

$$\begin{aligned} q(x) &= q_0 + q_1x + q_2x^2 \\ r(x) &= r_0 + r_1x \end{aligned} \quad (57.69)$$

En utilisant le fait que (pour rappel):

$$f(x) = g(x)q(x) + r(x) \quad (57.70)$$

nous avons donc immédiatement:

$$\begin{aligned} q_2 &= \frac{a_4}{b_2} \\ r_1(x) &= (a_3 - q_2b_1)x^3 + (a_2 - q_2b_0)x^2 + a_1x + a_0 = c_3x^3 + c_2x^2 + c_1x^1 + c_0 \end{aligned} \quad (57.71)$$

Ensuite (toujours en procédant de façon identique):

$$\begin{aligned} q_1 &= \frac{c_3}{b_2} \\ r_2(x) &= (c_2 - q_1b_1)x^2 + (c_1 - q_1b_0)x + c_0 = c'_2x^2 + c'_1x + c'_0 \end{aligned} \quad (57.72)$$

et enfin:

$$\begin{aligned} q_0 &= \frac{c'_2}{b_2} \\ r_3(x) &= r(x) = (c'_1 - q_0b_1)x + (c'_0 - q_0b_0) = c''_1x + c''_0 \end{aligned} \quad (57.73)$$

Donc de manière générale:

$$q_{n-m} = \frac{a_n}{b_m} \quad (57.74)$$

comme:

$$f(x) = g(x)q(x) + r(x) \Leftrightarrow r(x) = f(x) - q(x)g(x) \quad (57.75)$$

le premier reste est donc:

$$\begin{aligned}
 r_1(x) &= f(x) - q_{n-m}g(x) \\
 &= \underbrace{(a_n - q_{n-m}b_m)}_{=0}x^n + (a_{n-1} - q_{n-m}b_{m-1})x^{n-1} + (a_{n-2} - q_{n-m}b_{m-2})x^{n-2} + \dots \\
 &\quad + (a_{n-m} - q_{n-m}b_0)x^{n-m} + a_{n-m-1}x^{n-m-1} + \dots + a_0 = c_{n-1}x^{n-1} + c_{n-2}x^{n-2} + \dots + c_0
 \end{aligned} \tag{57.76}$$

Ensuite:

$$q_{n-m-1} = \frac{c_{n-1}}{b_m} \tag{57.77}$$

Le deuxième reste est alors:

$$\begin{aligned}
 r_2(x) &= r_1(x) - q_{n-m-1}g(x) \\
 &= \underbrace{(c_{n-1} - q_{n-m-1}b_m)}_{=0}x^{n-1} + (c_{n-2} - q_{n-m-1}b_{m-1})x^{n-2} + (c_{n-3} - q_{n-m-1}b_{m-2})x^{n-3} + \dots \\
 &\quad + (c_{n-m} - q_{n-m-1}b_0)x^{n-m} + a_{n-m-1}x^{n-m-1} + \dots + a_0 = c'_{n-2}x^{n-2} + c'_{n-3}x^{n-3} + \dots + c'_0
 \end{aligned} \tag{57.78}$$

etc.

Nous continuons jusqu'à ce que $\deg(r_k(x)) < m$.

9. RÉGRESSIONS ET INTERPOLATIONS

Les régressions (ou "interpolations") sont des outils très utiles aux statisticiens, ingénieurs, informaticiens souhaitant établir une loi de corrélation entre deux (ou plus) variables dans un contexte d'études et d'analyse ou d'extrapolation.

Il existe un grand nombre de méthodes d'interpolation: de la simple résolution d'équations du premier degré (lorsque uniquement deux points d'une mesure sont connus) aux équations permettant d'obtenir à partir d'un grand nombre de points des informations essentielles à l'établissement d'une loi (ou fonction) de régression linéaire, polynomiale ou encore logistique.

Listons les techniques les plus utilisées dans les entreprises et administrations (dont les modèles mathématiques ne sont pas tous présentés ici):

- Modèle de régression linéaire à une ou plusieurs variables explicatives basé sur la méthode des moindres carrés avec variables explicatives binaires ou continues avec réponse réelle. Présenté en détails dans le présent chapitre.

- Modèle de régression linéaire gaussien (approche statistique de la régression linéaire basée sur la méthode des moindres carrés) avec variables explicatives binaires ou continues avec réponse réelle. Présenté en détails dans le présent chapitre.

- Modèles de régression non-linéaires avec variables explicatives binaires ou continues avec réponse réelle. Présenté en détails dans le présent chapitre dans le cas où ils peuvent être ramenés à des cas linéaires ou non mais alors sans interactions des variables explicatives. Sinon basé sur des techniques de type quasi-Newton ou de Gauss-Newton présentées dans le présent chapitre.

- Modèle de régression polynomial par la méthode des B-spline et du polynôme de collocation. Présenté en détails dans le présent chapitre.
- Modèle de régression logistique (régression binomiale) avec variables explicatives binaires, nominales (catégorielles) ou ordinales ou continues avec réponse bornée entre 0 et 1. Présenté sommairement et naïvement dans le présent chapitre.
- Modèle de régression de comptage de Poisson (Poisson MLE, PMLE, GLM) ou binomial négatif (binomial MLE et QGPMLE) avec variables explicatives binaires, nominales (catégorielles) ou ordinales ou continues avec réponse entière positive.
- Modèle de régression linéaire orthogonal qui est utilisée comme complément au test- T pour données appariées pour voir la stabilité des instruments de mesure dans les laboratoires. Il s'agit d'un cas où la variable explicative et expliquée sont entachées d'une incertitude (je présenterai la démonstration quand le temps me le permettra).

9.1. RÉGRESSION LINÉAIRE À UNE VARIABLE EXPLICATIVE

Nous présentons ici plusieurs algorithmes (méthodes) utiles et connus dans les sciences expérimentales (nous en avons déjà parlé lors de notre étude des statistiques). L'objectif est de chercher à exprimer la relation linéaire entre deux variables x et y indépendantes par un "**modèle linéaire**" (ML) le plus simple possible (sinon quoi il faudrait des centaines de pages pour présenter le sujet!).

- x est la variable indépendante ou "**explicative**" appelée aussi "**covariable**" ou "**variable prédictive**" (et en économie "**variable exogène**"...). Les valeurs de x sont fixées par l'expérimentateur et sont supposées connues sans erreur

- y est la variable dépendante ou "**expliquée**" (exemple: réponse de l'analyseur) appelée aussi en économiste "**variable endogène**". Les valeurs de y sont entachées d'une erreur de mesure. L'un des buts de la régression sera précisément d'estimer cette erreur.

Nous cherchons une relation de la forme:

$$y = f(x) = ax + b \quad (57.79)$$

C'est l'équation d'une droite (fonction affine), d'où le terme de "**régression linéaire**" avec a qui est appelé dans ce cadre d'étude: "**coefficient de régression**".

Dans la vie réelle, les relations linéaires constituent vraiment une exception car la majorité des relations sont non linéaires dans la réalité et même non continues dans certaines situations... De plus, ce n'est pas parce qu'elles sont linéaires dans l'intervalle de mesures effectuées qu'elles le sont à plus petite échelle ou à plus grande échelle!

Cependant, dans la pratique nous nous arrangeons pour linéariser les fonctions soit par des transformations algébriques élémentaires comme celles qu'utilisent les tableurs (par exemple Microsoft Excel) pour la linéarisation d'une fonction logarithmique en faisant un simple changement de variables:

$$y = a \cdot \underbrace{\ln(x)}_{:=x} + b = ax + b \quad (57.80)$$

ou encore pour les fonctions puissance et exponentielles en faisant aussi une petite manipulation algébrique avec les propriétés des logarithmes démontrées dans le chapitre d'Analyse Fonctionnelle (à condition que a soit strictement positif):

$$\begin{aligned} y = ax^b &\Rightarrow \ln(y) = \ln(ax^b) = \ln(a) + \ln(x^b) = b \ln(x) + \ln(a) \\ y = ae^{bx} &\Rightarrow \ln(y) = \ln(ae^{bx}) = \ln(e^{bx}) + \ln(a) = bx + \ln(a) \end{aligned} \quad (57.81)$$

soit en faisant des approximations en série de Taylor (cf. chapitre de Suites Et Séries).

Remarques: Si nous cherchons à déterminer la valeur de y pour un x non mesuré et se situant au-delà de l'étendue de mesure d'origine, nous parlons alors "d'intervalle de prédiction pour x ".

9.1.1. DROITE DE RÉGRESSION

Il existe aussi une autre manière commune de faire une régression linéaire du type:

$$y = f(x) = ax + b \quad (57.82)$$

qui consiste à se baser sur les propriétés de la covariance et de l'espérance (cf. chapitre de Statistiques) et très utilisée entre autres en finance (mais aussi dans n'importe quel domaine où l'on fait un peu de statistiques).

Soient x, y deux variables dont l'une dépend de l'autre (souvent c'est y qui dépend de x) nous avons selon la propriété de linéarité de la covariance qui est, rappelons-le:

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y) \quad (57.83)$$

la relation suivante:

$$\begin{aligned} \text{cov}(X, aX + b) &= E(aX^2 + bX) - E(X)E(aX + b) \\ &= E(aX^2) + E(bX) - E(X)(E(aX) + E(b)) \\ &= aE(X^2) + bE(X) - E(X)(aE(X) + b) = aE(X^2) + bE(X) - aE(X)^2 - bE(X) \\ &= a(E(X^2) - E(X)^2) = a \text{cov}(X, X) = a \text{Var}(X) \end{aligned} \quad (57.84)$$

Il vient donc pour la pente (nous réutiliserons cette relation lors de l'étude du rendement d'un portefeuille selon le modèle de Sharpe dans le chapitre d'Économie):

$$a = \frac{\text{cov}(X, Y)}{\text{Var}(X)} \quad (57.85)$$

Soit sous la forme plus explicite que nous utiliserons plus loin (cf. chapitre de Statistiques):

$$a = \frac{\text{cov}(X,Y)}{V(X)} = \frac{E((x_k - \bar{x})(y_k - \bar{y}))}{\frac{1}{n} \sum_k (x_k - \bar{x})^2} = \frac{\frac{1}{n} \sum_k (x_k - \bar{x})(y_k - \bar{y})}{\frac{1}{n} \sum_k (x_k - \bar{x})^2} = \frac{\sum_k (x_k - \bar{x})(y_k - \bar{y})}{\sum_k (x_k - \bar{x})^2} \quad (57.86)$$

Pour déterminer l'ordonnée à l'origine nous utilisons les propriétés de l'espérance démontrées dans le chapitre de Statistiques:

$$E(Y) = E(aX + b) = aE(X) + E(b) = aE(x) + b \quad (57.87)$$

ce qui donne b sous la forme:

$$b = E(Y) - aE(X) \quad (57.88)$$

9.1.2. MÉTHODE DES MOINDRES CARRÉS

Du fait de l'erreur sur y , les points expérimentaux, de coordonnées (x_k, y_k) , ne se situent pas exactement sur la droite théorique. Il faut donc trouver l'équation de la droite expérimentale qui passe le plus près possible de ces points.

La "**méthode des moindres carrés**" (DMC) consistera dans le cadre d'étude particulier qui nous intéresse à chercher les valeurs des paramètres a, b qui rendent minimale la somme des carrés des écarts e_i résiduels (SS_r : Sum of Squared Residuals) entre les valeurs observées y_k et les valeurs calculées théoriques de y' . Nous parlons alors de "**méthode des moindres carrés des écarts d'ordonnées**":

$$SS_r = \sum_{k=1}^n (y_k - y'_k)^2 \quad (57.89)$$

où n est le nombre de points et:

$$y'_k = ax_k + b \quad (57.90)$$

d'où autrement écrit:

$$SS_r = \sum_{k=1}^n (y_k - ax_k - b)^2 = \Phi(a, b) \quad (57.91)$$

Cette relation fait apparaître la somme des carrés des écarts comme une fonction des paramètres a, b . Lorsque cette fonction est minimale (extrémale), les dérivées par rapport à ses paramètres s'annulent:

$$\begin{cases} \frac{\partial \Phi}{\partial a}(a, b) = -2 \sum_{k=1}^n x_k (y_k - ax_k - b) = 0 \\ \frac{\partial \Phi}{\partial b}(a, b) = -2 \sum_{k=1}^n (y_k - ax_k - b) = 0 \end{cases} \quad (57.92)$$

Remarque: Cette méthode de recherche de minimum (optimisation) est nommée "**méthode des multiplicateurs de Lagrange**" dans le monde de l'économétrie. Dans notre exemple SS_y est la grandeur scalaire qui fait office de multiplicateur de Lagrange.

Soit après simplification:

$$\begin{cases} b \sum_{k=1}^n x_k + a \sum_{k=1}^n x_k^2 = \sum_{k=1}^n x_k y_k \\ b n + a \sum_{k=1}^n x_k = \sum_{k=1}^n y_k \end{cases} \quad (57.93)$$

Le système ci-dessus est appelé "**système des équations normales**". C'est un système linéaire de deux équations à deux inconnues. Notons pour simplifier:

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k, \bar{y} = \frac{1}{n} \sum_{k=1}^n y_k \quad (57.94)$$

Le système devient:

$$\begin{cases} b n \bar{x} + a \sum_{k=1}^n x_k^2 = \sum_{k=1}^n x_k y_k \\ b + a \bar{x} = \bar{y} \end{cases} \quad (57.95)$$

De la deuxième équation, nous tirons:

$$b = \bar{y} - a \bar{x} \quad (57.96)$$

En remplaçant dans la première, nous obtenons:

$$\begin{aligned} (\bar{y} - a \bar{x}) n \bar{x} + a \sum_{k=1}^n x_k^2 &= \sum_{k=1}^n x_k y_k \Leftrightarrow n \bar{x} \bar{y} - n a \bar{x}^2 + a \sum_{k=1}^n x_k^2 = \sum_{k=1}^n x_k y_k \Leftrightarrow \\ a &= \frac{\sum_{k=1}^n x_k y_k - n \bar{x} \bar{y}}{\sum_{k=1}^n x_k^2 - n \bar{x}^2} = \frac{n \sum_{k=1}^n x_k y_k - \sum_{k=1}^n x_k \sum_{k=1}^n y_k}{n \sum_{k=1}^n x_k^2 - \left(\sum_{k=1}^n x_k \right)^2} \end{aligned} \quad (57.97)$$

De là nous avons:

$$\begin{aligned}
 b &= \frac{1}{n} \sum_{k=1}^n y_k - \frac{\left(n \sum_{k=1}^n x_k y_k - \sum_{k=1}^n x_k \sum_{k=1}^n y_k \right) \frac{1}{n} \sum_{k=1}^n x_k}{n \sum_{k=1}^n x_k^2 - \left(\sum_{k=1}^n x_k \right)^2} \\
 &= \frac{\left(n \sum_{k=1}^n x_k^2 - \left(\sum_{k=1}^n x_k \right)^2 \right) \frac{1}{n} \sum_{k=1}^n y_k - \left(n \sum_{k=1}^n x_k y_k - \sum_{k=1}^n x_k \sum_{k=1}^n y_k \right) \frac{1}{n} \sum_{k=1}^n x_k}{n \sum_{k=1}^n x_k^2 - \left(\sum_{k=1}^n x_k \right)^2} \quad (57.98) \\
 &= \frac{\sum_{k=1}^n x_k^2 \sum_{k=1}^n y_k - \sum_{k=1}^n x_k \sum_{k=1}^n x_k y_k}{n \sum_{k=1}^n x_k^2 - \left(\sum_{k=1}^n x_k \right)^2}
 \end{aligned}$$

Ainsi, les expressions de la pente et de l'ordonnée à l'origine de l'équation recherchée sont:

$$a = \frac{n \sum_{k=1}^n x_k y_k - \sum_{k=1}^n x_k \sum_{k=1}^n y_k}{n \sum_{k=1}^n x_k^2 - \left(\sum_{k=1}^n x_k \right)^2}; b = \frac{\sum_{k=1}^n x_k^2 \sum_{k=1}^n y_k - \sum_{k=1}^n x_k \sum_{k=1}^n x_k y_k}{n \sum_{k=1}^n x_k^2 - \left(\sum_{k=1}^n x_k \right)^2} \quad (57.99)$$

Les deux dernières relations sont utilisées par la majorité des tableurs comme par exemple dans la version française de Microsoft Excel 11.8346 lors de l'utilisation de la fonction REGRESSION(). Le terme b (l'ordonnée à l'origine) peut être obtenu directement avec la fonction ORDONNEE.ORIGINE() et a avec la fonction PENTE() et l'ensemble avec la fonction DROITEREG().

Voici si jamais un petit listing intéressant de quelques cas très pratiques avec ce tableur:

- Pour une droite:

a: =PENTE(y,x)

b: =ORDONNEE.ORIGINE(y,x)

- Pour une fonction logarithmique (nous retrouvons le changement de variable donné au début):

a: =INDEX(REGRESSION(y,LN(x)),1)

b: =INDEX(REGRESSION(y,LN(x)),1,2)

- Pour une fonction puissance (nous retrouvons le changement de variable donné au début):

a: =EXP(INDEX(REGRESSION(LN(y),LN(x),,),1,2))

b: =INDEX(REGRESSION(LN(y),LN(x),,),1)

- Pour une fonction exponentielle (nous retrouvons aussi le changement de variable donné au début):

a: =EXP(INDEX(REGRESSION(LN(y),x),1,2))

b: =INDEX(REGRESSION(LN(y),x),1)

Remarque: Il faut bien avoir en tête que la droite des moindres carrés, qui permet de résumer au mieux le nuage de points des observations, passe nécessairement par le centre de gravité du nuage, c'est-à-dire par un point moyen qui ne correspond que rarement à une observation (moyenne des abscisses et moyenne des ordonnées).

9.1.3. ANALYSE DE LA VARIANCE DE LA RÉGRESSION BIVARIÉE

Nous avons donc maintenant pour la droite des moindres carrés:

$$y = ax + b \quad (57.100)$$

soit sous forme discrète:

$$y_k = ax_k + b \quad (57.101)$$

ainsi que par construction de la méthode la relation suivante:

$$\bar{y} = a\bar{x} + b \quad (57.102)$$

Maintenant, nous faisons l'hypothèse que chaque valeur mesurée est entachée d'un résidu tel que:

$$y_k = ax_k + b + e_k \quad (57.103)$$

Soit en soustrayant les deux dernières relations:

$$(y_k - \bar{y}) = a(x_k - \bar{x}) + e_k \Rightarrow e_k = (y_k - \bar{y}) - a(x_k - \bar{x}) \quad (57.104)$$

Maintenant, passons par un résultat intermédiaire. Rappelons que nous avons obtenu plus haut:

$$\begin{aligned} \sum_k x_k (y_k - ax_k - b) &= 0 \\ \sum_k (y_k - ax_k - b) &= 0 \end{aligned} \quad (57.105)$$

En remplaçant b par sa valeur:

$$b = \bar{y} - a\bar{x} \quad (57.106)$$

nous avons alors:

$$\begin{aligned} \sum_k x_k (y_k - ax_k - \bar{y} + a\bar{x}) &= \sum_k x_k ((y_k - \bar{y}) - a(x_k - \bar{x})) = 0 \\ \sum_k (y_k - ax_k - \bar{y} + a\bar{x}) &= \sum_k ((y_k - \bar{y}) - a(x_k - \bar{x})) = 0 \end{aligned} \quad (57.107)$$

Multipliant la deuxième relation ci-dessus par $\bar{x}$ et retranchant de la première, nous obtenons:

$$\sum_k x_k ((y_k - \bar{y}) - a(x_k - \bar{x})) - \bar{x} \sum_k ((y_k - \bar{y}) - a(x_k - \bar{x})) = 0 \quad (57.108)$$

Soit après réarrangement:

$$\sum_k (x_k - \bar{x}) ((y_k - \bar{y}) - a(x_k - \bar{x})) = 0 \quad (57.109)$$

Revenons maintenant à:

$$(y_k - \bar{y}) = a(x_k - \bar{x}) + e_k \quad (57.110)$$

Si nous mettons le tout au carré et en sommant pour toutes les observations, nous avons:

$$\sum_k (y_k - \bar{y})^2 = a^2 \sum_k (x_k - \bar{x})^2 + 2a \sum_k (x_k - \bar{x}) e_k + \sum_k e_k^2 \quad (57.111)$$

soit:

$$\sum_k (y_k - \bar{y})^2 = a^2 \sum_k (x_k - \bar{x})^2 + 2a \sum_k (x_k - \bar{x}) ((y_k - \bar{y}) - a(x_k - \bar{x})) + \sum_k e_k^2 \quad (57.112)$$

Or, nous venons de montrer avant que le double produit était nul. Donc:

$$\sum_k (y_k - \bar{y})^2 = a^2 \sum_k (x_k - \bar{x})^2 + \sum_k e_k^2 \quad (57.113)$$

Cette dernière relation est appelée "**équation d'analyse de la variance**". En fait, il s'agit de sommes de carrés. Il faudrait diviser par n pour obtenir des variances biaisées.

Cette dernière relation s'écrit aussi souvent:

$$SCT = SCE + SCR \quad (57.114)$$

où SCT est la "**somme des carrés totale**", SCE la "**somme des carrés expliquée**" et SCR la "**somme des carrés résiduelle**".

Cette dernière relation se trouve également souvent sous la forme suivante dans la littérature:

$$\sum_k (y_k - \bar{y})^2 = \sum_k (ax_k + b - a\bar{x} - b)^2 + \sum_k e_k^2 = \sum_k (ax_k + b - \bar{y})^2 + \sum_k e_k^2 \quad (57.115)$$

Notons maintenant les y_k sans erreurs d'une manière différente et appelons cela le "**modèle linéaire a priori**":

$$\hat{y}_k = ax_k + b \quad (57.116)$$

Il est effectivement important dans la pratique de différencier le modèle a priori qui ne prend pas en compte les erreurs du modèle réel entaché d'erreurs!

Nous avons alors:

$$\sum_k (y_k - \bar{y})^2 = \sum_k (\hat{y}_k - \bar{y})^2 + \sum_k e_k^2 \quad (57.117)$$

où la première somme après l'égalité est très souvent appelée "**manque d'ajustement**" ("**lack of fit**" en anglais). Cette dernière relation peut se représenter graphiquement sous la forme suivante:

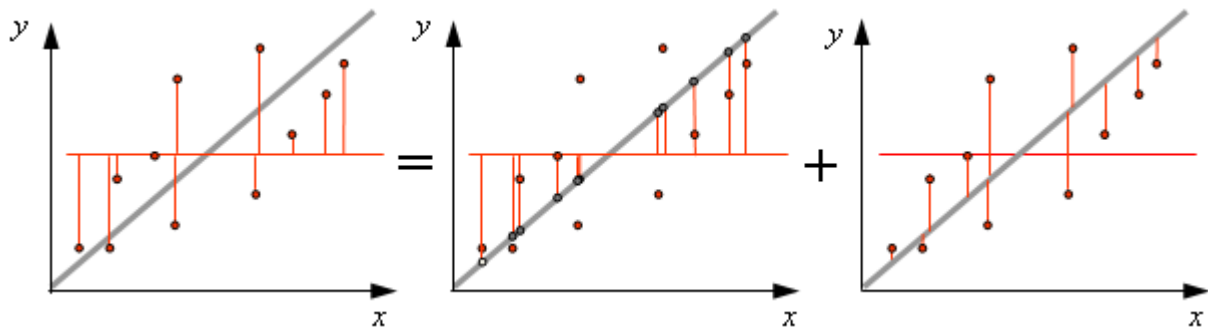


Figure: 57.2 - Représentation graphique de SCT, SCE, SCR

La dernière relation est souvent notée dans la littérature sous la forme plus pédagogique suivante:

$$\sum_k (y_k - \bar{y})^2 = \sum_k (f_k - \bar{y})^2 + \sum_k e_k^2 = \sum_k (f_k - \bar{y})^2 + \sum_k (y_k - f_k)^2 \quad (57.118)$$

qui est simplement une autre manière d'écrire la décomposition de la variance (implicite):

$$SCT = SCE + SCR \quad (57.119)$$

et il vient alors immédiatement la relation utilisée parfois dans la pratique pour calculer les résidus (connaissant les valeurs calculées et les valeurs mesurées):

$$\sum_k e_k^2 = \sum_k (y_k - \bar{y})^2 - \sum_k (\hat{y}_k - \bar{y})^2 \quad (57.120)$$

Il est important de se rappeler (ou de constater) que les relations ci-dessus entre SCT , SCE et SCR ne sont valables que dans le cas d'un modèle linéaire!

Rappelons maintenant que dans le chapitre de Statistiques, nous avons démontré que le coefficient de corrélation s'exprimait (était défini) par:

$$R_{X,Y} = \frac{\text{cov}_{X,Y}}{\sqrt{V(X)V(Y)}} \Rightarrow R_{X,Y}^2 = \frac{(\text{cov}_{X,Y})^2}{V(X)V(Y)} \quad (57.121)$$

Ou sinon puisque nous avons démontré plus haut que (se rappeler que la variance indiquée est la variance biaisée...):

$$a = \frac{\text{cov}(X,Y)}{V(X)} \quad (57.122)$$

nous pouvons aussi écrire le coefficient de corrélation sous la forme:

$$R_{X,Y}^2 = \frac{a^2 V(X)}{V(Y)} \quad (57.123)$$

Donc nous en déduisons):

$$\begin{aligned}
 R_{X,Y}^2 &= \frac{(\text{cov}_{X,Y})^2}{V(X)V(Y)} = \frac{(E((x_k - \bar{x})(y_k - \bar{y})))^2}{E((x_k - \bar{x})^2)E((y_k - \bar{y})^2)} = \frac{\left(\frac{1}{n} \sum_k (x_k - \bar{x})(y_k - \bar{y})\right)^2}{\frac{1}{n} \sum_k (x_k - \bar{x})^2 \frac{1}{n} \sum_k (y_k - \bar{y})^2} \\
 &= \frac{\left(\sum_k (x_k - \bar{x})(y_k - \bar{y})\right)^2}{\sum_k (x_k - \bar{x})^2 \sum_k (y_k - \bar{y})^2} = \frac{a^2 \sum_k (x_k - \bar{x})^2}{\sum_k (y_k - \bar{y})^2}
 \end{aligned} \tag{57.124}$$

Montrons que ceci est égal à (notation souvent utilisée dans la littérature spécialisée):

$$R_{X,Y}^2 = \frac{SCE}{SCT} \tag{57.125}$$

Remarque: Cette formulation du coefficient de corrélation est extrêmement utile car, contrairement à la formulation statistique, cette dernière se généralise immédiatement à la régression linéaire multiple que nous verrons un peu plus loin.

Démonstration:

Nous partons donc de:

$$R_{X,Y}^2 = \frac{SCE}{SCT} = \frac{a^2 \sum_k (x_k - \bar{x})^2}{\sum_k (y_k - \bar{y})^2} \tag{57.126}$$

et puisque nous avons montré que:

$$\sum_k (x_k - \bar{x})(y_k - \bar{y}) - a \sum_k (x_k - \bar{x})^2 = 0 \Rightarrow a = \frac{\sum_k (x_k - \bar{x})(y_k - \bar{y})}{\sum_k (x_k - \bar{x})^2} \tag{57.127}$$

Donc:

$$R_{X,Y}^2 = \frac{\left(\frac{\sum_k (x_k - \bar{x})(y_k - \bar{y})}{\sum_k (x_k - \bar{x})^2}\right)^2 \sum_k (x_k - \bar{x})^2}{\sum_k (y_k - \bar{y})^2} = \frac{\left(\sum_k (x_k - \bar{x})(y_k - \bar{y})\right)^2}{\sum_k (x_k - \bar{x})^2 \sum_k (y_k - \bar{y})^2} \tag{57.128}$$

□C.Q.F.D.

et dans le cadre des modèles de régression voici quelques cas typiques de la valeur de ce coefficient avec un corrélation linéaire pour les deux premières lignes et non linéaire pour la troisième ligne:

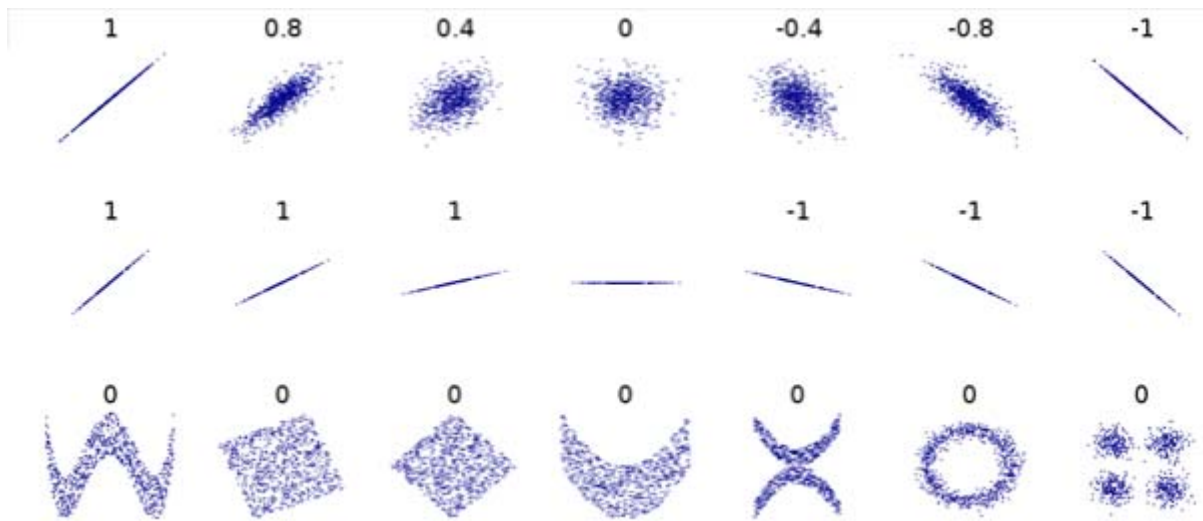


Figure: 57.3 - Quelques valeurs du coefficient de corrélation linéaire (source: Wikipédia)

Enfin, indiquons que nous retrouvons aussi très souvent le coefficient de corrélation linéaire sous la forme suivante dans les logiciels et la littérature:

$$R^2 = R_{X,Y}^2 = \frac{SCE}{SCT} = \frac{SCT - SCR}{SCT} = 1 - \frac{SCR}{SCT} \quad (57.129)$$

Cette dernière forme met mieux en évidence que si la somme des carrés des résidus SCR est nulle, le modèle est parfaitement modélisable par une relation linéaire dans l'intervalle d'étude considéré.

Enfin, remarquons que l'ordonnée à l'origine n'intervient pas dans la valeur du coefficient de corrélation puisque (propriété de bilinéarité de la covariance démontrée dans le chapitre de Statistiques):

$$\text{cov}(X, Y) = \text{cov}(X, aX + b) = a \text{cov}(X, X) + b \text{cov}(X, 1) = a\sigma_X^2 + b \cdot 0 \quad (57.130)$$

9.2. MODÈLE LINÉAIRE GAUSSIEN

Nous admettrons que, pour un individu i prélevé au hasard dans la population, x_k est connu sans erreur, et que y_k est une réalisation d'une variable aléatoire que nous noterons dorénavant Y_k et la droite théorique des moindres carrés s'écrira maintenant:

$$Y_k = \alpha x_k + \beta + \varepsilon_k \quad (57.131)$$

où ε_k est par hypothèse un résidu identiquement distribué et indépendant (pas d'autocorrélation) pour chaque point i selon une loi Normale centrée (de moyenne nulle et d'écart-type égal pour tout k) tel que:

$$\boxed{\varepsilon_k = N(0, \sigma^2)} \text{ et } \boxed{\text{cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j} \quad (57.132)$$

donc autrement dit:

$$E(\varepsilon_k) = 0, \sigma(\varepsilon_k) = \sigma^2 \quad (57.133)$$

où nous avons le résidu qui est donc donné par la différence entre l'ordonnée théorique et l'ordonnée mesurée (expérimentale):

$$\varepsilon_k = Y_k - y_k \quad (57.134)$$

et puisque par hypothèse $\varepsilon_k = N(0, \sigma^2)$, il vient immédiatement que:

$$Y_k = N(\alpha x_k + \beta, \sigma^2) \quad (57.135)$$

Raison pour laquelle le modèle s'appelle "modèle linéaire gaussien" Explicitement, cela donne donc:

$$Y_k = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\alpha x_k + \beta - \mu_k}{\sigma}\right)^2} \quad (57.136)$$

Ce qui équivaut graphiquement à avoir:

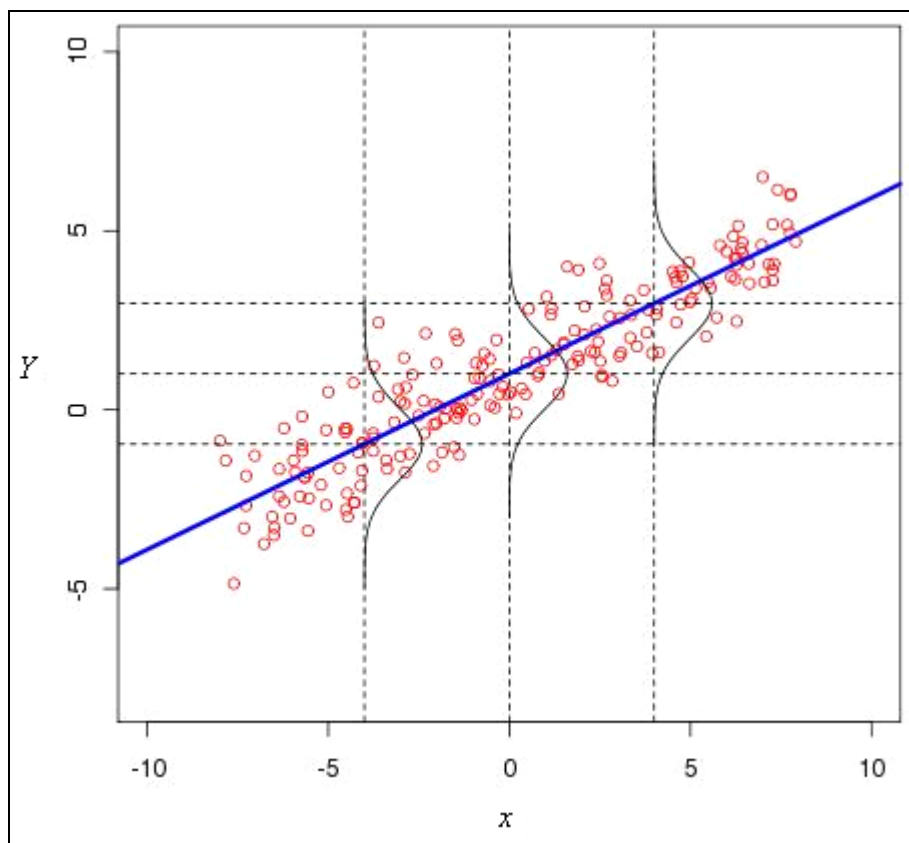


Figure: 57.4 - Représentation graphique du principe de distribution Normale

Presque tous les logiciels d'analyse statistique proposent de faire un tracé (graphique) des résidus en fonction des valeurs de x . Ainsi, deux graphiques du type suivant amèneraient à rejeter le modèle linéaire gaussien:

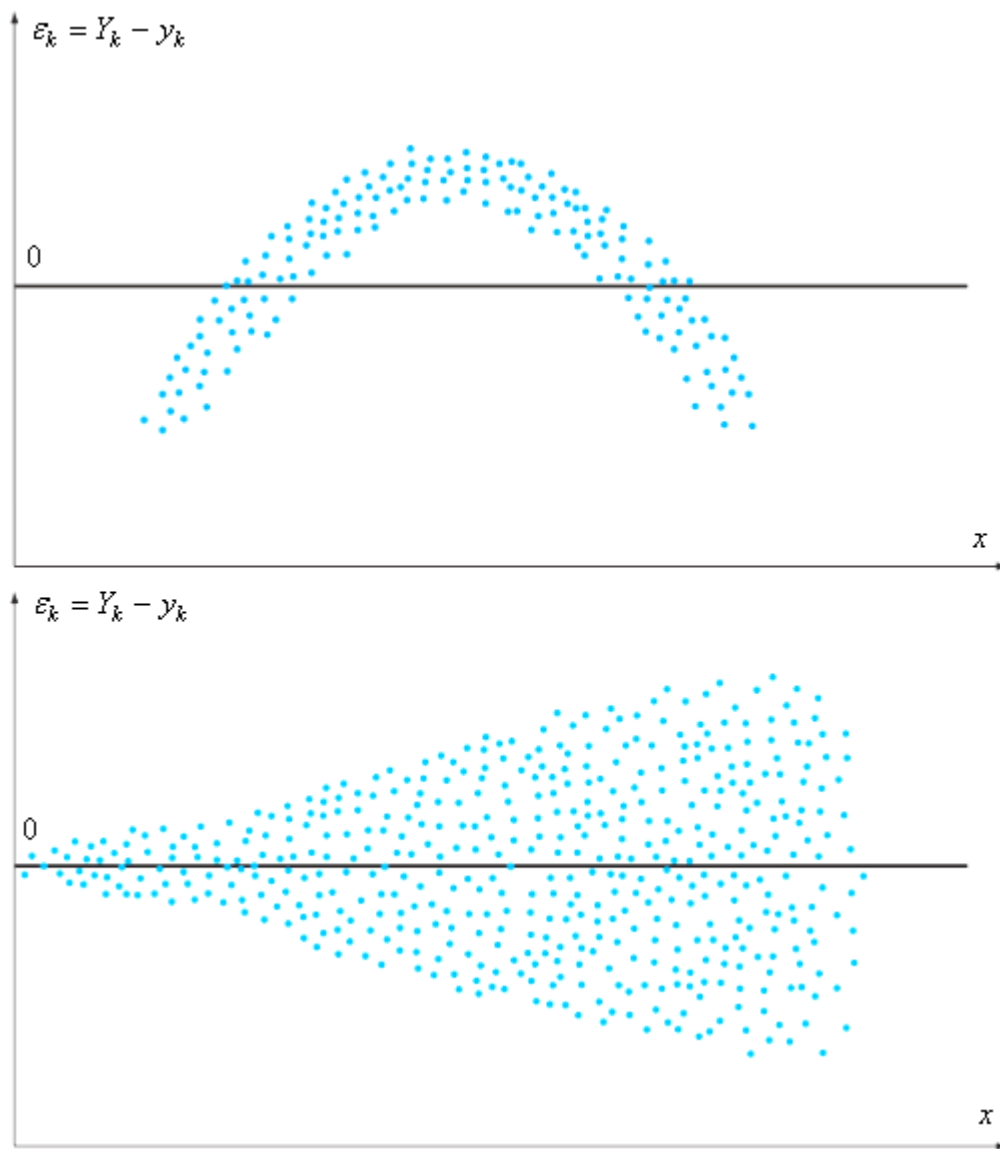


Figure: 57.5 - Exemple de "plot" de résidus qui n'annoncent rien de bon...

Les hypothèses précédentes concernant les moments des résidus sont appelées "**hypothèses de Gauss-Markov**" et l'hypothèse particulière d'égalité des variances s'appelle comme nous l'avons vu dans le chapitre de Statistiques "l'homoscédasticité" (tandis que le fait que les variances ne soient pas égales s'appelle pour rappel "l'hétéroscédasticité").

Remarque: La majorité des logiciels (dont Microsoft Excel 11.8346) proposent un graphique qui montre les résidus en fonction des valeurs des ordonnées x . Bien évidemment, il vaut mieux que les points représentant les résidus ne soient pas trop divergents... sinon quoi l'hypothèse d'homoscédasticité est violée.

Nous avons de par la propriété de l'espérance (cf. chapitre de Statistiques):

$$\bar{Y} = E(Y_k) = E(\alpha x_k + \beta + \varepsilon_k) = E(\alpha x_k) + E(\beta) + E(\varepsilon_k) = \alpha \bar{x} + \beta \quad (57.137)$$

Alors sous les hypothèses ci-dessus, nous allons montrer que a et b sont les estimateurs sans biais (cf. chapitre de Statistiques) de α et β et qu'il est possible d'estimer l'écart-type à partir de SCR ce qui est un résultat important appelé "**théorème de Gauss-Markov**".

Avant de voir la démonstration faisons un rappel et donnons quelques définitions des variables que nous avons déjà manipulées et des nouvelles que nous allons manipuler (si le vocabulaire semble technique au lecteur alors il devra au préalable lire ou relire le chapitre de Probabilités et de Statistiques):

Variable	Description
a	Pente (coefficient directeur) du modèle linéaire de la méthode des moindres carrés. Il s'agit d'une valeur ponctuelle (déterministe).
b	Ordonnée à l'origine du modèle linéaire de la méthode des moindres carrés. Il s'agit d'une valeur ponctuelle (déterministe).
A	Variable aléatoire de la pente (coefficient directeur) selon l'approche statistique du modèle gaussien et dont a est une réalisation. L'espérance de A étant un estimateur non biaisé de α .
B	Variable aléatoire de l'ordonnée à l'origine selon l'approche statistique du modèle linéaire gaussien et dont b est une réalisation. L'espérance de B étant un estimateur non biaisé de β .
α	Espérance non biaisée de la variable A représentant la pente (coefficient directeur) dans le cadre de l'approche statistique du modèle linéaire gaussien.
β	Espérance non biaisée de la variable B représentant l'ordonnée à l'origine dans le cadre de l'approche statistique du modèle linéaire gaussien.

Tableau: 57.3 - Rappel des notations pour l'étude de la régression linéaire

Démonstration:

Conformément au modèle adopté, a est à considérer maintenant comme une réalisation de la variable aléatoire donnée par (démontrée plus haut comme étant le rapport de la covariance et de la variance):

$$A = \frac{\sum_k (x_k - \bar{x})(Y_k - \bar{Y})}{\sum_k (x_k - \bar{x})^2} \quad (57.138)$$

et b comme une réalisation de la variable aléatoire donnée par:

$$B = \bar{Y} - A\bar{x} \quad (57.139)$$

Donc nous différencions les valeurs aléatoires et non aléatoires des coefficients par le passage de la notation minuscule à majuscule.

Tenant compte de ce que la variable expliquée théorique considérée comme la réalisation d'une variable aléatoire est donnée par:

$$Y_k = \alpha x_k + \beta + \varepsilon_k \quad (57.140)$$

nous pouvons mettre A sous la forme:

$$\begin{aligned}
A &= \frac{\sum_k (x_k - \bar{x})(Y_k - \bar{Y})}{\sum_k (x_k - \bar{x})^2} = \frac{\sum_k (x_k - \bar{x})(\alpha x_k + \beta + \varepsilon_k - \bar{Y})}{\sum_k (x_k - \bar{x})^2} \\
&= \frac{\sum_k (x_k - \bar{x})(\alpha x_k + \beta + \varepsilon_k - E(Y))}{\sum_k (x_k - \bar{x})^2} = \frac{\sum_k (x_k - \bar{x})(\alpha x_k + \beta + \varepsilon_k - \alpha \bar{x} - \beta)}{\sum_k (x_k - \bar{x})^2} \\
&= \frac{\sum_k (x_k - \bar{x})(\alpha x_k + \varepsilon_k - \alpha \bar{x})}{\sum_k (x_k - \bar{x})^2} = \frac{\alpha \sum_k (x_k - \bar{x})(x_k - \bar{x}) + \sum_k (x_k - \bar{x}) \varepsilon_k}{\sum_k (x_k - \bar{x})^2} \\
&= \frac{\alpha \sum_k (x_k - \bar{x})^2}{\sum_k (x_k - \bar{x})^2} + \frac{\sum_k (x_k - \bar{x}) \varepsilon_k}{\sum_k (x_k - \bar{x})^2} = \alpha + \frac{\sum_k (x_k - \bar{x}) \varepsilon_k}{\sum_k (x_k - \bar{x})^2}
\end{aligned} \tag{57.141}$$

et B :

$$B = \bar{Y} - A\bar{x} = E(Y) - A\bar{x} = \alpha\bar{x} + \beta - A\bar{x} = \beta - \bar{x}(A - \alpha) \tag{57.142}$$

Nous en déduisons les espérances pour A :

$$\begin{aligned}
E(A) &= E\left(\alpha + \frac{\sum_k (x_k - \bar{x}) \varepsilon_k}{\sum_k (x_k - \bar{x})^2}\right) = E(\alpha) + E\left(\frac{\sum_k (x_k - \bar{x}) \varepsilon_k}{\sum_k (x_k - \bar{x})^2}\right) \\
&= \alpha + \frac{\sum_k (x_k - \bar{x}) E(\varepsilon_k)}{\sum_k (x_k - \bar{x})^2} = \alpha
\end{aligned} \tag{57.143}$$

ce qui montre que notre hypothèse initiale pour l'expression de A n'est pas trop mauvaise... puisque l'espérance de A est visiblement un estimateur non biaisé de α .

Respectivement, pour B il vient:

$$\begin{aligned}
E(B) &= E(\beta - \bar{x}(A - \alpha)) = E(\beta) - E(\bar{x}(A - \alpha)) = \beta - \bar{x}E(A - \alpha) \\
&= \beta - \bar{x}E(A) + \bar{x}E(\alpha) = \beta - \bar{x}\alpha + \bar{x}\alpha = \beta
\end{aligned} \tag{57.144}$$

avec la même conclusion.

□C.Q.F.D.

Donc l'espérance de A et B sont bien des estimateurs sans biais (donc de variance minimale pour rappel) de α, β . Comme ce sont des estimateurs, dans la littérature spécialisée, α, β sont souvent notés $\hat{\alpha}, \hat{\beta}$ et dès lors il vient la notation courante alternative:

$$\hat{Y}_k = \hat{a}x_k + \hat{b} \quad (57.145)$$

Nous devons enfin calculer aussi les variances de A et B en utilisant ces propriétés (cf. chapitre de Statistiques) et les hypothèses sur les résidus, nous avons:

$$\begin{aligned} V(A) &= V\left(\alpha + \frac{\sum_k (x_k - \bar{x}) \varepsilon_k}{\sum_k (x_k - \bar{x})^2}\right) = V\left(\frac{\sum_k (x_k - \bar{x}) \varepsilon_k}{\sum_k (x_k - \bar{x})^2}\right) = \frac{V\left(\sum_k (x_k - \bar{x}) \varepsilon_k\right)}{\left[\sum_k (x_k - \bar{x})^2\right]^2} \\ &= \frac{\sum_k V((x_k - \bar{x}) \varepsilon_k)}{\left[\sum_k (x_k - \bar{x})^2\right]^2} = \frac{\sum_k (x_k - \bar{x})^2 V(\varepsilon_k)}{\left[\sum_k (x_k - \bar{x})^2\right]^2} \end{aligned} \quad (57.146)$$

Comme par hypothèse nous avons tous les $V(\varepsilon_k)$ qui sont égaux, et qu'il n'y a pas d'autocorrélation, nous pouvons alors écrire:

$$V(A) = \frac{\sum_k (x_k - \bar{x})^2 V(\varepsilon)}{\left[\sum_k (x_k - \bar{x})^2\right]^2} = \frac{V(\varepsilon) \sum_k (x_k - \bar{x})^2}{\left[\sum_k (x_k - \bar{x})^2\right]^2} = \frac{V(\varepsilon)}{\sum_k (x_k - \bar{x})^2} \quad (57.147)$$

Soit si n est assez grand nous écrivons:

$$\sigma^2(A) = V(A) = \frac{\sigma^2}{\sum_k (x_k - \bar{x})^2} = \frac{\sigma^2}{n \cdot V(x)} \quad (57.148)$$

Avant de déterminer la variance de B , rappelons que par hypothèse:

$$Y_k = N(\alpha x_k + \beta, \sigma^2) \quad (57.149)$$

donc par propriété de linéarité de la loi Normale, les variables aléatoires A et B suivent donc aussi une loi Normale.

Du rappel de cette hypothèse, il découle immédiatement (cf. chapitre de Statistiques):

$$V(Y_k) = \sigma^2 \Rightarrow V(\bar{Y}) = \frac{\sigma^2}{n} \quad (57.150)$$

Dès lors:

$$\begin{aligned}
 V(B) &= V(\bar{Y} - A\bar{x}) = V(\bar{Y}) + \bar{x}^2 V(A) - 2\bar{x} \operatorname{cov}(\bar{Y}, A) = V(\bar{Y}) + \bar{x}^2 V(A) \\
 &= \frac{\sigma^2}{n} + \bar{x}^2 \sigma^2(A) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{\sum_k (x_k - \bar{x})^2} = \frac{\sigma^2 \left(\sum_k (x_k - \bar{x})^2 + n\bar{x}^2 \right)}{n \sum_k (x_k - \bar{x})^2} \quad (57.151)
 \end{aligned}$$

En rappelant la relation de Huyghens (cf. chapitre de Statistiques):

$$\sum_k (x_k - \bar{x})^2 = \sum_k x_k^2 - n\bar{x}^2 \quad (57.152)$$

Nous avons finalement:

$$\sigma^2(B) = V(B) = \frac{\sigma^2 \sum_k x_k^2}{n \sum_k (x_k - \bar{x})^2} = \frac{\sigma^2 \sum_k x_k^2}{n^2 \cdot V(x)} \quad (57.153)$$

où évidemment la notation de la variance au dénominateur est très abusive (puisque x n'est pas une variable aléatoire) mais très pratique pour condenser les écritures.

Le problème réside maintenant dans la détermination de σ^2 . Évidemment pour ce faire nous allons être obligés de passer par un estimateur statistique.

Nous savons que nous pouvons écrire selon ce qui a été vu dans le chapitre de Statistiques en ce qui concerne les estimateurs:

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum (\varepsilon_i - \bar{\varepsilon})^2 = \frac{1}{n-2} \sum \varepsilon_i^2 = \frac{1}{n-2} \sum (Y_i - y_i)^2 \quad (57.154)$$

puisque la loi Normale est centrée pour les résidus donc $\bar{\varepsilon} = 0 \dots$ et que le résidu est une variable aléatoire implicitement dépendante de la somme de deux variables aléatoires que sont A et B d'où la minoration de deux fois l'erreur standard.

Indiquons aussi que dans la pratique nous notons fréquemment ce dernier résultat en mélangeant les notations de l'aspect aléatoire et déterministe (donc en notant tout avec des minuscules):

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum (y'_k - y_k)^2 = \frac{1}{n-2} \sum e_k^2 = \frac{SCR}{n-2} = SEE \quad (57.155)$$

où SEE signifie en anglais "**Standard Error of Estimate**" (Erreur Standard de l'Estimation). Il s'agit en français de "**l'erreur standard de la régression**".

Nous avons donc les estimateurs non biaisés des variances de A et de B :

$$\hat{V}(A) = \frac{\hat{\sigma}^2}{n \cdot V(x)} = \frac{\frac{1}{n-2} \sum e_k^2}{n \cdot V(x)} \quad \hat{V}(B) = \frac{\hat{\sigma}^2 \sum_k x_k^2}{n^2 \cdot V(x)} = \frac{\frac{1}{n-2} \sum e_k^2 \sum_k x_k^2}{n^2 \cdot V(x)} \quad (57.156)$$

relations qui ne sont donc valables que pour une régression linéaire à une variable explicative (et sous les hypothèses de construction du modèle linéaire Gaussien). Sachant par construction de l'hypothèse de départ que A et B suivent une loi Normale d'espérance respective α, β et dont la variance est donnée juste ci-dessus, nous connaissons donc entièrement la distribution qui les caractérise.

Ce qui est sympathique connaissant ces variances, c'est que nous pouvons aussi estimer la variance de la variable expliquée de notre régression facilement (en utilisant les propriétés de la variance vues dans le chapitre de Statistiques).

Il serait intéressant aussi de faire de l'inférence statistique sur l'espérance des paramètres A et B (donc la pente et l'ordonnée à l'origine) étant donné leur espérance empirique connue. Dès lors, nous avons démontré dans le chapitre de Statistiques l'intervalle de confiance suivant:

$$\bar{X} - \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \quad (57.157)$$

Il s'ensuit en faisant un parallèle digne de l'ingénieur... que comme A est un estimateur non biaisé de la moyenne de la pente a et que:

$$\hat{\sigma}_A = \hat{V}(A) = \sqrt{\frac{\frac{1}{n-2} \sum e_k^2}{nV(x)}} \quad (57.158)$$

est en fait l'erreur standard de la moyenne A , alors par analogies:

$$\begin{aligned} \mu &\rightarrow a \\ \bar{X} &\rightarrow A \\ \frac{S}{\sqrt{n}} &\rightarrow \sqrt{\frac{\frac{1}{n-2} \sum e_k^2}{nV(x)}} \\ T_{\alpha/2}(n-1) &\rightarrow T_{\alpha/2}(n-2) \end{aligned} \quad (57.159)$$

et alors (c'est un raisonnement à prendre avec des pincettes et il vaut mieux utiliser les développements qui vont suivre par la suite):

$$A - T_{\alpha/2}(n-2) \sqrt{\frac{\frac{1}{n-2} \sum e_k^2}{nV(x)}} \leq a \leq A + T_{\alpha/2}(n-2) \sqrt{\frac{\frac{1}{n-2} \sum e_k^2}{nV(x)}} \quad (57.160)$$

ce qui donne donc l'intervalle de confiance de la pente d'un modèle linéaire Gaussien à une variable explicative (c'est ce que donne Microsoft Excel 11.8346 pour chaque coefficient). La démarche est la même pour l'ordonnée à l'origine.

Dans le cas d'une régression linéaire ayant plusieurs variables explicatives, assimilées alors au concept de "**degrés de liberté**" (d.d.l.) ou en anglais de "**degrees of freedom**", l'idée est la même mais les calculs sont plus longs (pas le courage de les faire...).

Enfin, rappelons que nous avons obtenu pour le coefficient de corrélation empirique:

$$R_{X,Y}^2 = \frac{a^2 V(X)}{V(Y)} \quad (57.161)$$

Nous avons alors in extenso le fameux intervalle de confiance du coefficient de corrélation:

$$\left(A - T_{\alpha/2}(n-2) \sqrt{\frac{1}{n-2} \frac{\sum e_k^2}{nV(x)}} \right)^2 \frac{V(X)}{V(Y)} \leq R_{X,Y}^2 \leq \left(A + T_{\alpha/2}(n-2) \sqrt{\frac{1}{n-2} \frac{\sum e_k^2}{nV(x)}} \right)^2 \frac{V(X)}{V(Y)} \quad (57.162)$$

Il faut savoir qu'étant donné que calculer l'intervalle de confiance pour la pente ou pour le coefficient de corrélation est une démarche équivalente, nombreux sont les logiciels (Tanagra, Minitab, Microsoft Excel, etc.) qui donnent la valeur de la distribution de Student et la valeur critique de celle-ci uniquement pour la pente et supposent que le lecteur sait qu'il en est de même pour le coefficient de corrélation.

9.2.1. TEST DU COEFFICIENT DE CORRÉLATION DE PEARSON

Le calcul obtenu plus haut pour l'intervalle de confiance du coefficient de corrélation est un peu pénible dans la pratique. C'est pour cette raison que de nombreux praticiens et logiciels de statistiques implémentent une alternative très simple communiquée uniquement sous la forme minimale qu'est la p -value.

Pour voir cette approche, rappelons que nous avons vu dans le chapitre de Statistiques que (cette fois-ci nous allons adopter la notation correcte...):

$$\rho = \frac{\text{cov}(X,Y)}{\sqrt{V(X)V(Y)}} \quad (57.163)$$

Et que nous avons vu juste plus haut que:

$$\rho = \frac{\text{cov}(X,Y)}{\sqrt{V(X)}} \frac{1}{\sqrt{V(X)V(Y)}} = a \frac{1}{\sqrt{V(X)V(Y)}} = \frac{a\sigma_X}{\sigma_Y} \quad (57.164)$$

De la même manière, nous avons l'estimateur du coefficient de corrélation de Pearson qui est (en adaptant ici aussi les notations d'usage possibles que nous pouvons trouver dans la littérature):

$$R_{X,Y} = \sqrt{\frac{SCE}{SCT}} = \frac{A \sqrt{\sum_k (x_k - \bar{x})^2}}{\sqrt{\sum_k (y_k - \bar{y})^2}} = \frac{A \sqrt{\frac{1}{n-1} \sum_k (x_k - \bar{x})^2}}{\sqrt{\frac{1}{n-1} \sum_k (y_k - \bar{y})^2}} = \frac{A\hat{\sigma}_X}{\hat{\sigma}_Y} \quad (57.165)$$

et donc:

$$A = \frac{R_{X,Y}\hat{\sigma}_Y}{\hat{\sigma}_X} \quad (57.166)$$

Le test d'hypothèse que nous voulons faire est alors:

$$H_0 : \rho = 0 \quad (57.167)$$

et donc équivalent à:

$$H_0 : \alpha = 0 \quad (57.168)$$

l'hypothèse nulle étant évidemment que le coefficient de corrélation de Pearson est statistiquement significativement différent de zéro. Il s'agit donc d'un test bilatéral!

Pour trouver une forme simple du test, rappelons que nous avons obtenu:

$$(n-2) \hat{\sigma}^2 = \sum_k (y'_k - y_k)^2 = \sum \varepsilon_k^2 = SCR \quad (57.169)$$

ainsi que:

$$R_{X,Y}^2 = \frac{SCE}{SCT} = \frac{SCT - SCR}{SCT} = 1 - \frac{SCR}{SCT} \quad (57.170)$$

ce qui nous amène en mixant les deux à avoir:

$$SCR = SCT(1 - R_{X,Y}^2) = \sum_k (y_k - \bar{y})^2 (1 - R_{X,Y}^2) = (n-1) \hat{\sigma}_Y^2 (1 - R_{X,Y}^2) \quad (57.171)$$

d'où:

$$\hat{\sigma}^2 = \frac{(n-1) \hat{\sigma}_Y^2 (1 - R_{X,Y}^2)}{n-2} \quad (57.172)$$

Or, nous avons aussi montré que si n est assez grand:

$$A - T_{\alpha/2}(n-2) \sqrt{\frac{1}{n-2} \frac{\sum \varepsilon_k^2}{nV(x)}} \leq a \leq A + T_{\alpha/2}(n-2) \sqrt{\frac{1}{n-2} \frac{\sum \varepsilon_k^2}{nV(x)}} \quad (57.173)$$

Mais si n est assez petit, nous écrirons donc:

$$A - T_{\alpha/2}(n-2) \sqrt{\frac{1}{n-2} \frac{\sum \varepsilon_k^2}{(n-1) \hat{\sigma}_X^2}} \leq a \leq A + T_{\alpha/2}(n-2) \sqrt{\frac{1}{n-2} \frac{\sum \varepsilon_k^2}{(n-1) \hat{\sigma}_X^2}} \quad (57.174)$$

Donc:

$$T_{\alpha/2}(n-2) \geq \frac{A-a}{\sqrt{\frac{1}{n-2} \sum e_k^2}} = \frac{A-a}{\sqrt{\frac{1}{n-2} SCR}} \quad (57.175)$$

$$\sqrt{(n-1) \hat{\sigma}_X^2}$$

et avec l'hypothèse nulle $\alpha = 0$, il vient:

$$T_{\alpha/2}(n-2) \geq \frac{A}{\sqrt{\frac{1}{n-2} \sum e_k^2}} = \frac{R_{X,Y} \frac{\hat{\sigma}_Y}{\hat{\sigma}_X}}{\sqrt{\frac{SCR}{n-2}}} = \frac{R_{X,Y} \hat{\sigma}_Y}{\hat{\sigma}_X \sqrt{\frac{\hat{\sigma}^2}{(n-1) \hat{\sigma}_X^2}}} = \frac{R_{X,Y} \hat{\sigma}_Y}{\hat{\sigma} \sqrt{\frac{1}{n-1}}} \quad (57.176)$$

$$= \frac{R_{X,Y} \hat{\sigma}_Y}{\sqrt{\frac{1}{n-1}} \sqrt{\frac{(n-1) \hat{\sigma}_Y^2 (1-R_{X,Y}^2)}{n-2}}} = \frac{R_{X,Y}}{\sqrt{\frac{1-R_{X,Y}^2}{n-2}}}$$

Il faut faire attention à l'utilisation de ce test suivant si le coefficient de corrélation de Pearson est négatif ou positif et ne pas oublier qu'il est bilatéral.

Exemples:

E1. Nous avons calculé pour une série de données un coefficient de corrélation de Pearson positif R valant 0.298 et la variable explicative comporte 7 valeurs. Nous avons donc avec la version anglaise de Microsoft Excel 11.8346 la p -value qui est (nous retrouvons exactement la même valeur qu'un logiciel comme Minitab 15.1.1):

$$=2*(1-T.DIST(0.298/SQRT((1-0.298^2)/(7-2));7-2;1))=2*(1-0.741869)=0.51626$$

Dans le cas présent nous ne rejetons donc pas l'hypothèse nulle comme quoi le coefficient de corrélation de Pearson est nul à un niveau de confiance de 5%.

E2. Nous avons calculé pour une série de données un coefficient de corrélation positif de Pearson R valant -0.084 et la variable explicative comporte 19 valeurs. Nous avons donc avec la version anglaise de Microsoft Excel 11.8346 la p -value qui est (nous retrouvons exactement la même valeur qu'un logiciel comme Minitab 15.1.1):

$$=2* T.DIST((-0.084)/SQRT((1-(-0.084)^2)/(19-2));19-2;1)=2*(1-0.366)=0.74186$$

Dans le cas présent nous ne rejetons donc pas l'hypothèse nulle comme quoi le coefficient de corrélation de Pearson est nul à un niveau de confiance de 5%.

Ce petit piège fait que finalement on prend la valeur absolue du coefficient de corrélation de Pearson et on fait le calcul toujours d'une seule manière

9.2.2. INTERVALLE DE CONFIANCE DES VALEURS PRÉDICTIVES

Nous souhaiterions pour chaque valeur mesurée de la variable expliquée, connaître l'intervalle de confiance. En d'autres termes, nous souhaiterions connaître la variance de l'estimateur statistique de Y (nous n'écrivons plus les indices pour gagner du temps):

$$V(Y) = V(Ax + B) = V(Ax) + V(B) + 2\text{cov}(Ax, B) \quad (57.177)$$

Malheureusement, nous allons nous casser les dents, car la covariance est difficile à calculer (A et B ne sont pas indépendants comme le montrent les expressions que nous avons obtenues plus haut).

Par contre, en étant bon observateur, nous voyons que si nous utilisons le résultat vu plus haut:

$$B = \bar{Y} - A\bar{x} \quad (57.178)$$

alors:

$$V(Y) = V(Ax + B) = V(Ax + (\bar{Y} - A\bar{x})) = V(\bar{Y} + A(x - \bar{x})) \quad (57.179)$$

Le problème étant contourné, nous avons maintenant en utilisant les propriétés de la variance:

$$\begin{aligned} V(Y) &= V(\bar{Y} + A(x - \bar{x})) = V(\bar{Y}) + V(A(x - \bar{x})) + \underbrace{2\text{cov}(\bar{Y}, A(x - \bar{x}))}_{=0} \\ &= V(\bar{Y}) + V(A(x - \bar{x})) = V(\bar{Y}) + (x - \bar{x})^2 V(A) = \frac{\sigma^2}{n} + (x - \bar{x})^2 \frac{\sigma^2}{\sum_k (x_k - \bar{x})^2} \end{aligned} \quad (57.180)$$

d'où:

$$V(Y) = \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_k (x_k - \bar{x})^2} \right) \quad (57.181)$$

Donc nous avons:

$$Y = N \left(Ax + B, \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_k (x_k - \bar{x})^2} \right) \right) \quad (57.182)$$

Maintenant, rappelons (cf. chapitre de Statistiques) que:

$$Z = \frac{X - \mu}{\sigma} \quad (57.183)$$

et comme Y est distribué selon une loi Normale dont l'estimateur non biaisé de la moyenne ainsi que l'écart-type sont donnés par la relation antérieure il vient immédiatement:

$$Z = \frac{Y - (Ax + B)}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_k (x_k - \bar{x})^2} \right)}} = \frac{Y - (Ax + B)}{\sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}}} \quad (57.184)$$

Dans la pratique il faut aussi vérifier que ce rapport suit effectivement une loi Normale pour pouvoir faire les intervalles de confiance et tests statistiques qui s'en suivent.

Rappelons maintenant que nous avons démontré dans le chapitre de Statistiques que:

$$T(k) = \frac{Z}{\sqrt{U/k}} \quad (57.185)$$

suit donc une loi de Student de degré de liberté k et U une loi du Khi-deux de degré de liberté k .

Maintenant revenons à l'expression du Z précédemment obtenue et rappelons que:

$$\sigma^2 = \frac{1}{n-2} \sum \varepsilon_i^2 \quad (57.186)$$

Donc:

$$\sigma = \sqrt{\left(\sum \varepsilon_i^2 \right) / (n-2)} \quad (57.187)$$

Or comme:

$$\varepsilon_i = N(0, \sigma^2) \quad (57.188)$$

Alors:

$$\frac{\varepsilon_i}{\sigma} = N(0, 1) \quad (57.189)$$

et donc:

$$\sum \frac{\varepsilon_i^2}{\sigma^2} = \sum \left(\frac{\varepsilon_i}{\sigma} \right)^2 \quad (57.190)$$

correspond à une somme de carrés de lois Normales centrées réduites. Et donc conformément à ce que nous avons démontré dans le chapitre de Statistiques il vient que:

$$\sum \left(\frac{\varepsilon_i}{\sigma} \right)^2 = \chi^2 (n-2) = \chi_{n-2}^2 \quad (57.191)$$

Donc:

$$\begin{aligned} T(n-2) = T_{n-2} &= \frac{Z}{\sqrt{U/k}} = \frac{Z}{\sqrt{\sum \left(\frac{\varepsilon_i}{\sigma} \right)^2 / (n-2)}} = \frac{Z}{\sqrt{\frac{1}{\sigma^2} \sum \varepsilon_i^2 / (n-2)}} \\ &= \frac{\sigma Z}{\sqrt{\sum \varepsilon_i^2 / (n-2)}} = \frac{\sigma Z}{\hat{\sigma}} \end{aligned} \quad (57.192)$$

Soit:

$$T(n-2) = \frac{\sigma Z}{\hat{\sigma}} = \frac{\sigma \left(\frac{Y - (Ax + B)}{\sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}}} \right)}{\hat{\sigma}} = \frac{Y - (Ax + B)}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}}} \quad (57.193)$$

C'est une raison pour laquelle beaucoup de statisticiens notent directement et sans détours:

$$Z = \frac{Y - (Ax + B)}{\sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}}} \Rightarrow T(n-2) = \frac{Y - (Ax + B)}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}}} \quad (57.194)$$

Ce qui n'est pas forcément évident au premier coup d'oeil. Raison pour laquelle, suite à la demande d'un lecteur, nous avons détaillé un peu de manière exagérée, le mécanisme qui se cache derrière cette implication.

Bon ceci étant, fait, nous avons donc:

$$T(n-2) = \frac{Y - (Ax + B)}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}}} \quad (57.195)$$

et il en découle en bilatéral, un intervalle de confiance à un niveau α donné et pour un x fixé par:

$$Y = Ax + B \pm T_{\alpha/2}(n-2) \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}} \quad (57.196)$$

Remarque: Le même type de développement peut être fait pour la pente et l'ordonnée à l'origine. Raison pour laquelle des logiciels comme Microsoft Excel, IBM SPSS, Minitab, Statistica, etc. donnent la valeur de la distribution de Student ainsi que l'intervalle de confiance à un niveau de α choisi. Mais pour que cela ait un sens, il faut que toutes les hypothèses du modèle construit soient satisfaites.

Raison pour laquelle de nombreux logiciels de statistiques donnent le graphique suivant lors de régressions linéaires (nous y voyons bien l'intervalle de confiance):

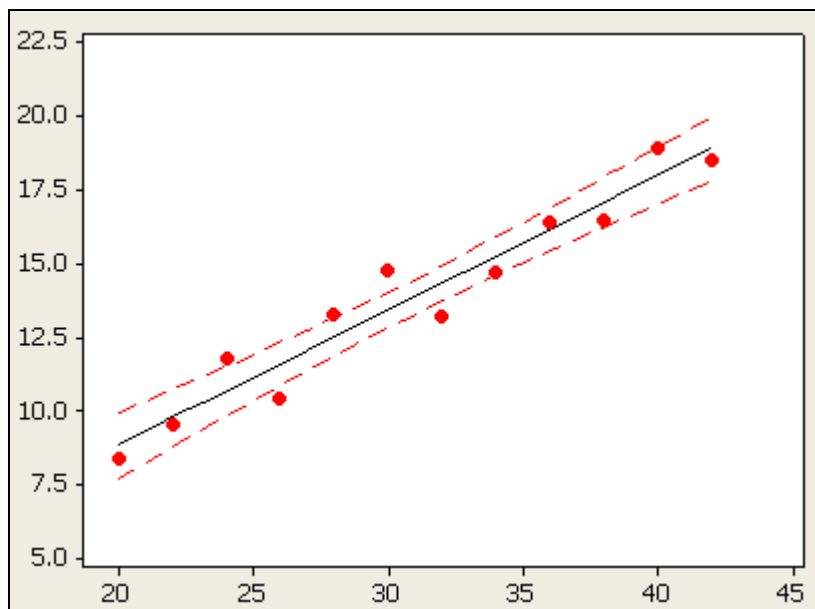


Figure: 57.6 - Capture du logiciel Minitab 15 pour l'intervalle de confiance

Le lecteur aura remarqué que:

1. Il est fort pénible sans logiciel d'obtenir le tracé de l'intervalle de confiance pour la droite des moindres carrés puisqu'il faudrait le calculer pour chaque point...
2. L'intervalle de confiance est courbe ce qui est parfois considéré comme relevant du bon sens, du moins dans la version temporelle de la régression : plus l'échéance de la prévision est lointaine, moins elle est sûre.

La valeur vraie de Y est donc donnée par:

$$y = a + bx + \varepsilon \quad (57.197)$$

avec la variance:

$$V(y) = \varepsilon = \sigma^2 \quad (57.198)$$

qui est indépendante de l'estimateur Y . Dès lors, la différence entre Y et y (donc entre estimateur et réel) a pour variance:

$$V(Y - y) = V(Y) + V(y) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_k (x_k - \bar{x})^2} \right) \quad (57.199)$$

Dès lors il est d'usage de considérer que "**l'intervalle de prédiction**" (à ne pas confondre avec l'intervalle de confiance de l'estimateur) est pris comme étant:

$$Y = Ax + B \pm T_{\alpha/2}(n-2) \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}} \quad (57.200)$$

Ce qui nous donne avec un logiciel comme Minitab 15:

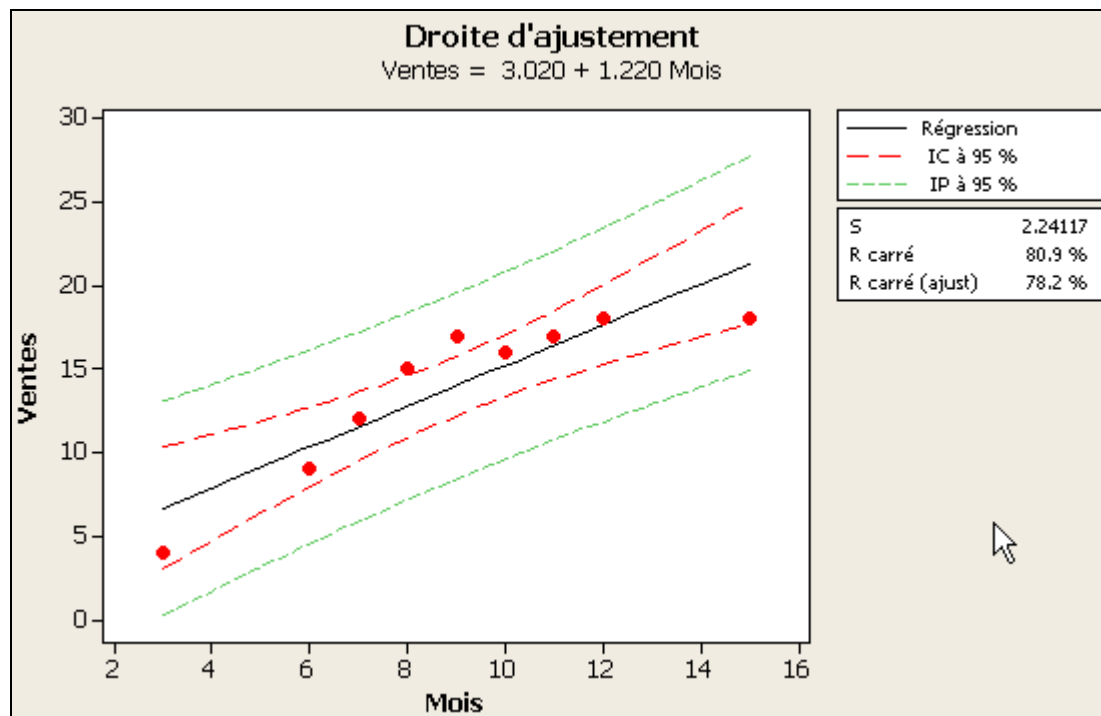


Figure: 57.7 - Capture du logiciel Minitab 15 pour l'intervalle de confiance et de prédiction

où nous distinguons bien l'intervalle de confiance (en rouge) de l'intervalle de prévision (en vert).

9.3. RÉGRESSION LINÉAIRE À UNE VARIABLE EXPLICATIVE FORCÉE PAR L'ORIGINE

Un cas très fréquent et demandé dans les laboratoires (et globalement dans d'autres départements) des entreprises, est de forcer la régression linéaire à passer par l'origine.

Nous allons voir ici que l'approche est seulement une variante simplifiée de la méthode des moindres carrés.

Nous utilisons comme précédemment:

$$SS_y = \sum_{k=1}^n (y_k - y'_k)^2 \quad (57.201)$$

où n est le nombre de points. Mais cette fois-ci, nous écrivons:

$$y'_k = ax_k \quad (57.202)$$

d'où autrement écrit:

$$SS_y = \sum_{k=1}^n (y_k - ax_k)^2 = \Phi(a) \quad (57.203)$$

Cette relation fait apparaître la somme des carrés des écarts comme une fonction du paramètre a . Lorsque cette fonction est minimale (extrémale), les dérivées par rapport à ses paramètres s'annulent:

$$\frac{\partial \Phi}{\partial a}(a) = -2 \sum_{k=1}^n x_k (y_k - ax_k) = 0 \quad (57.204)$$

Soit après simplification:

$$a \sum_{k=1}^n x_k^2 = \sum_{k=1}^n x_k y_k \quad (57.205)$$

Enfin:

$$a = \frac{\sum_{k=1}^n x_k y_k}{\sum_{k=1}^n x_k^2} \quad (57.206)$$

Vous pouvez vérifier aussi avec un tableur quelconque (Microsoft Excel par exemple) que les calculs correspondent bien.

9.4. RÉGRESSION LINÉAIRE MULTIPLE

Bien évidemment, dans certaines situations la régression linéaire est trop simpliste ou simplement pas adaptée. Le cas le plus typique qui nous concerne dans ce qui va suivre étant les situations où nous avons plusieurs variables explicatives.

Le principe de la régression linéaire multiple est relativement simple. Nous voulons déterminer la variable expliquée y à partir de $p-1$ variables explicatives indépendantes (donc en absence de "colinéarité") reliées par une loi linéaire de la forme générale:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1} + \varepsilon \quad (57.207)$$

Dans un échantillon de n individus, nous mesurons $y_i, x_{i,1}, \dots, x_{i,p-1}$ pour $i = 1 \dots n$:

Observations	y_i	$x_{i,1}$	...	$x_{i,p-1}$
1	y_1	$x_{1,1}$	...	$x_{1,p-1}$
2	y_2	$x_{2,1}$	...	$x_{2,p-1}$
n	y_n	$x_{n,1}$	...	$x_{n,p-1}$

Au fait, pour estimer les paramètres $\beta_0, \dots, \beta_{p-1}$ (valeurs estimées que nous noterons $\hat{\beta}_0, \dots, \hat{\beta}_{p-1}$ afin de respecter cette fois-ci la tradition) l'approche est très simple car elle consiste juste en une généralisation de la méthode des moindres carrés que nous avons vue plus haut pour la régression linéaire simple.

Donc en fin de compte nous réécrivons la relation du Sum of Squared Residuals vue plus haut en la modifiant un tout petit peu puisque maintenant nous avons du multilinéaire:

$$SS_y = \sum_{i=1}^k (y_i - y'_i)^2 \quad (57.208)$$

avec le modèle théorique estimé:

$$y'_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \dots + \hat{\beta}_{p-1} x_{i,p-1} \quad (57.209)$$

Donc il nous faut minimiser:

$$SS_y = \sum_{i=1}^k \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \dots + \hat{\beta}_{p-1} x_{i,p-1} \right) \right)^2 \quad (57.210)$$

Les parenthèses ci-dessus peuvent être réécrites sous la forme:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p-1} \end{bmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_0 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (57.211)$$

$$\vec{y} = X \vec{\hat{\beta}} + \vec{\varepsilon}$$

Nous avons le vecteur des résidus qui vaut donc:

$$\vec{\varepsilon} = \vec{y} - \vec{y}' = \vec{y} - X\vec{\beta} \quad (57.212)$$

Comme nous le savons, la méthode des moindres carrés consiste à trouver le vecteur $\vec{\beta}$ qui minimise:

$$\|\vec{\varepsilon}\|^2 = \vec{\varepsilon} \circ \vec{\varepsilon} = \vec{\varepsilon}^T \circ \vec{\varepsilon} \quad (57.213)$$

Soit explicitement:

$$\|\vec{\varepsilon}\|^2 = \vec{\varepsilon}^T \circ \vec{\varepsilon} = (\vec{y} - X\vec{\beta})^T (\vec{y} - X\vec{\beta}) = \vec{y}^T \vec{y} - \vec{y}^T (X\vec{\beta}) - (X\vec{\beta})^T \vec{y} + (X\vec{\beta})^T (X\vec{\beta}) \quad (57.214)$$

à remarquer que nous avons:

$$(X\vec{\beta})^T (X\vec{\beta}) = (X\vec{\beta})(X\vec{\beta})^T \quad (57.215)$$

et:

$$\vec{y}^T (X\vec{\beta}) = (X\vec{\beta})^T \vec{y} \quad (57.216)$$

puisque chacun des éléments de la multiplication est un simple vecteur.

Nous avons alors (attention! ne pas oublier que certaines multiplications dans la relation qui va suivre sont des produits scalaires!!!) la fonction quadratique multivariée de coefficients de vecteurs:

$$\|\vec{\varepsilon}\|^2 = \vec{\varepsilon} \circ \vec{\varepsilon} = (\vec{y} - X\vec{\beta}) \circ (\vec{y} - X\vec{\beta}) = \vec{y} \circ \vec{y} - 2(X\vec{\beta}) \circ \vec{y} + (X\vec{\beta}) \circ (X\vec{\beta}) = F \quad (57.217)$$

Dérivons cette dernière "fonction objet F " à l'ordre du vecteur $\vec{\beta}$ (il s'agit d'une dérivée intérieure composante par composante). Ce que nous écrivons:

$$\begin{aligned}
\frac{\partial F}{\partial \vec{\beta}} &= \frac{\partial}{\partial \vec{\beta}} \left(\vec{y} \circ \vec{y} - 2 \left(X \vec{\beta} \right) \circ \vec{y} + \left(X \vec{\beta} \right) \circ \left(X \vec{\beta} \right) \right) \\
&= \frac{\partial}{\partial \vec{\beta}} \left(-2 \left(X \vec{\beta} \right) \circ \vec{y} + \left(X \vec{\beta} \right) \circ \left(X \vec{\beta} \right) \right) \\
&= -2 \left(X \vec{u} \right) \circ \vec{y} + \frac{\partial}{\partial \vec{\beta}} \left(\left(X \vec{\beta} \right) \circ \left(X \vec{\beta} \right) \right) \\
&= -2 \left(X \vec{u} \right) \circ \vec{y} + \frac{\partial \left(X \vec{\beta} \right)}{\partial \vec{\beta}} \circ \left(X \vec{\beta} \right) + \left(X \vec{\beta} \right) \circ \frac{\partial \left(X \vec{\beta} \right)}{\partial \vec{\beta}} \\
&= -2 \left(X \vec{u} \right) \circ \vec{y} + \left(X \vec{u} \right) \circ \left(X \vec{\beta} \right) + \left(X \vec{\beta} \right) \circ X \vec{u} \\
&= -2 \left(X \vec{u} \right) \circ \vec{y} + 2 \left(X \vec{u} \right) \circ \left(X \vec{\beta} \right)
\end{aligned} \tag{57.218}$$

Maintenant, passons d'une écriture vectorielle à une écriture matricielle pure:

$$\frac{\partial F}{\partial \vec{\beta}} = -2 \left(X \vec{u} \right)^T \vec{y} + 2 \left(X \vec{u} \right)^T \left(X \vec{\beta} \right) \tag{57.219}$$

Nous cherchons donc $\vec{\beta}$ qui annule cette dérivée. Donc nous devons résoudre l'équation suivante:

$$-2 \left(X \vec{u} \right)^T \vec{y} + 2 \left(X \vec{u} \right)^T \left(X \vec{\beta} \right) = 0 \tag{57.220}$$

Soit:

$$\left(X \vec{u} \right)^T \left(X \vec{\beta} \right) = \left(X \vec{u} \right)^T \vec{y} \tag{57.221}$$

Rappelons avant d'aller plus loin que:

$$\left(X \vec{u} \right)^T = \vec{u}^T X^T \tag{57.222}$$

Donc:

$$\left(\vec{u}^T X^T \right) \left(X \vec{\beta} \right) = \left(\vec{u}^T X^T \right) \vec{y} \tag{57.223}$$

Puisque l'algèbre linéaire est associative, écrivons sans les parenthèses:

$$\vec{u}^T X^T X \vec{\beta} = \vec{u}^T X^T \vec{y} \tag{57.224}$$

Nous ne pouvons évidemment pas simplifier à droite et gauche par $\vec{u}^T X^T$ car comme il ne s'agit pas d'une matrice carrée, ce terme est obligatoirement non-inversible. La seule chose que nous puissions faire c'est identifier les éléments tels que:

$$\vec{u}^T \left(X^T X \vec{\beta} \right) = \vec{u}^T \left(X^T \vec{y} \right) \quad (57.225)$$

impose forcément que:

$$X^T X \vec{\beta} = X^T \vec{y} \quad (57.226)$$

Nous trouvons alors que si la matrice carrée $X^T X$ est inversible alors:

$$\boxed{\vec{\beta} = \left(X^T X \right)^{-1} X^T \vec{y}} \quad (57.227)$$

Pour montrer que cela semble juste, retrouvons les résultats de la régression linéaire simple:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 \quad (57.228)$$

donc en ne supposant que deux observations, nous avons alors:

$$\begin{aligned} \vec{\beta} &= \left(X^T X \right)^{-1} X^T \vec{y} \\ &= \left(\begin{bmatrix} 1 & x_{1,1} \\ 1 & x_{2,1} \end{bmatrix}^T \begin{bmatrix} 1 & x_{1,1} \\ 1 & x_{2,1} \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & x_{1,1} \\ 1 & x_{2,1} \end{bmatrix}^T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\ &= \left(\begin{bmatrix} 1 & 1 \\ x_{1,1} & x_{2,1} \end{bmatrix} \begin{bmatrix} 1 & x_{1,1} \\ 1 & x_{2,1} \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 \\ x_{1,1} & x_{2,1} \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad (57.229) \\ &= \left(\begin{bmatrix} 1 \cdot 1 + 1 \cdot 1 & x_{1,1} + x_{2,1} \\ x_{1,1} + x_{2,1} & x_{1,1}^2 + x_{2,1}^2 \end{bmatrix} \right)^{-1} \begin{pmatrix} y_1 + y_2 \\ x_{1,1} y_1 + x_{2,1} y_2 \end{pmatrix} \\ &= \left(\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \right)^{-1} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \end{aligned}$$

En utilisant la relation démontrée dans le chapitre d'Algèbre Linéaire pour calculer en toute généralité l'inverse d'une matrice $A(a_{ij})$ en $A^{-1}(b_{ij})$:

$$b_{ij} = \frac{1}{\det(A)} \cdot (-1)^{i+j} \cdot \det(A_{ji}) \quad (57.230)$$

Nous avons dans le cas présent:

$$\det(X^T X) = a_{11}a_{22} - a_{12}a_{21} = n \sum x_i^2 - \sum x_i \sum x_i = n \sum x_i^2 - \left(\sum x_i \right)^2 \quad (57.231)$$

Nous avons donc:

$$\begin{aligned} (X^T X)^{-1} &= \frac{1}{n \sum x_i^2 - \left(\sum x_i \right)^2} (-1)^{i+j} \det(X^T X_{ji}) \\ &= \frac{1}{n \sum x_i^2 - \left(\sum x_i \right)^2} \begin{bmatrix} \det(X^T X_{11}) & -\det(X^T X_{21}) \\ -\det(X^T X_{12}) & \det(X^T X_{22}) \end{bmatrix} \quad (57.232) \end{aligned}$$

et vu que nous avons une matrice carrée de dimension 2 seulement, le calcul des quatre déterminants se réduit au fait à sélectionner les composantes de $X^T X$ (cf. chapitre Algèbre Linéaire)

$$(X^T X)^{-1} = \frac{1}{n \sum x_i^2 - \left(\sum x_i \right)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \quad (57.233)$$

Donc:

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T \vec{y} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \\ &= \frac{1}{n \sum x_i^2 - \left(\sum x_i \right)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \\ &= \frac{1}{n \sum x_i^2 - \left(\sum x_i \right)^2} \begin{pmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ -\sum x_i \sum y_i + n \sum x_i y_i \end{pmatrix} \quad (57.234) \\ &= \begin{pmatrix} \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - \left(\sum x_i \right)^2} \\ \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i \right)^2} \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} b \\ a \end{pmatrix} \end{aligned}$$

Donc à un changement de notation près pour les indices et les mesures expérimentales, nous retrouvons bien le résultat que nous avons obtenu lors de notre étude de la régression linéaire simple qui était (pour rappel...):

$$a = \frac{n \sum_{k=1}^n x_k y_k - \sum_{k=1}^n x_k \sum_{k=1}^n y_k}{n \sum_{k=1}^n x_k^2 - \left(\sum_{k=1}^n x_k \right)^2}; b = \frac{\sum_{k=1}^n x_k^2 \sum_{k=1}^n y_k - \sum_{k=1}^n x_k \sum_{k=1}^n x_k y_k}{n \sum_{k=1}^n x_k^2 - \left(\sum_{k=1}^n x_k \right)^2} \quad (57.235)$$

Maintenant, il nous faut un indicateur de qualité en ce qui concerne notre régression multilinéaire.

Rappelons que dans le cadre de notre étude de la régression linéaire à une variable explicative, nous avons démontré que le coefficient de corrélation linéaire pouvait être écrit sous la forme:

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_k (\hat{y}_k - \bar{y})^2}{\sum_k (y_k - \bar{y})^2} \quad (57.236)$$

et au fait celui-ci est applicable aussi directement à la régression linéaire multiple puisque qu'il ne présuppose pas le nombre de variables explicatives!!

Remarque: Le lecteur intéressé par la mise en pratique de ces résultats pourra, au même titre que pour la régression simple, se référer au serveur d'exercices - section Méthodes Numériques - où il y a des exemples pratiques avec Microsoft Excel.

Évidemment avec la régression linéaire multiple, nous pouvons maintenant, moyennant une astuce, faire de la régression linéaire de polynômes (nous verrons plus loin comment appliquer cependant directement la méthode des moindres carrés sur un polynôme). Effectivement, considérons un polynôme de la forme:

$$y = ax^3 + bx^2 + cx + d \quad (57.237)$$

Ce que nous pouvons considérer sous la forme:

$$y = az + bw + cx + d \quad (57.238)$$

Donc dans des tableurs du type Microsoft Excel, il suffit d'utiliser l'utilitaire d'analyse de la régression avec en variable d'entrée la colonne x , une deuxième colonne que l'on aura pris soin de créer avec le carré de x et qui sera donc considérée comme la variable explicative w et enfin la troisième colonne que l'on aura aussi pris le soin de créer comme étant le cube de x et qui sera donc considérée comme la variable explicative z .

Nous pouvons aussi obtenir directement les coefficients de polynômes avec les fonctions déjà susmentionnées (mais vous n'aurez pas tous les résultats pertinents de l'Utilitaire d'Analyse). Par exemple pour un polynôme du deuxième degré:

```
a: =INDEX(REGRESSION(y,x^{1,2}),1)
b: =INDEX(REGRESSION(y,x^{1,2}),1,2)
c: =INDEX(REGRESSION(y,x^{1,2}),1,3)
```

et pour un polynôme du troisième degré:

a: =INDEX(REGRESSION(y,x^{1,2,3}),1)
 b: =INDEX(REGRESSION(y,x^{1,2,3}),1,2)
 c: =INDEX(REGRESSION(y,x^{1,2,3}),1,3)
 d: =INDEX(REGRESSION(y,x^{1,2,3}),1,4)

C'est cette astuce qui permet de comprendre pourquoi et comment le coefficient de corrélation linéaire s'applique aussi aux polynômes dans la majorité des tableurs. Cependant nous pouvons avoir une approche plus directe qui ne nécessite pas de transformation mais qui dès lors est un peu plus longue.

9.5. RÉGRESSION POLYNOMIALE

Nous allons voir maintenant comment déterminer par exemple le meilleur polynôme du deuxième degré qui passe par un nombre quelconque de points mais sans transformer la fonction contrairement à ce que nous venons de faire juste avant! Comme nous aimons bien la physique sur ce site, nous allons reprendre un cas classique de la cinématique afin de joindre l'utile à l'agréable...

Considérons donc que nous recherchons un polynôme du deuxième degré de la forme:

$$x = x_0 + v_0 t + \frac{1}{2} a t^2 \quad (57.239)$$

sachant que la méthode est très facilement applicable à des polynômes de degré supérieur.

Relation qu'il est d'usage d'écrire dans le domaine de la régression polynomiale sous la forme suivante:

$$d_i = c_1 + c_2 t_i + c_3 t_i^2 \quad (57.240)$$

où i représente le nombre de points à notre disposition.

Pour la suite, nous allons nous baser encore une fois sur la méthode des moindres carrés. En d'autres termes, nous cherchons les coefficients c_1, c_2, c_3 qui minimisent l'erreur:

$$SS_r = \sum_{i=1}^n (d_i - d_i')^2 = \sum_{i=1}^n (d_i' - d_i)^2 = \sum_{i=1}^n \left((c_1 + c_2 t_i + c_3 t_i^2) - d_i \right)^2 = \Phi(c_1, c_2, c_3) \quad (57.241)$$

et nous réattaquons avec des dérivées partielles pour chaque coefficient:

$$\frac{\partial \Phi}{\partial c_1} = \sum_{i=1}^n 2 \left(c_1 + c_2 t_i + c_3 t_i^2 - d_i \right) = 2 \left(n c_1 + c_2 \sum_{i=1}^n t_i + c_3 \sum_{i=1}^n t_i^2 - \sum_{i=1}^n d_i \right) = 0 \quad (57.242)$$

Soit après simplification et un petit réarrangement:

$$n c_1 + c_2 \sum_{i=1}^n t_i + c_3 \sum_{i=1}^n t_i^2 = \sum_{i=1}^n d_i \quad (57.243)$$

De même:

$$\frac{\partial \Phi}{\partial c_2} = \sum_{i=1}^n 2t_i (c_1 + c_2 t_i + c_3 t_i^2 - d_i) = 2 \left(c_1 \sum_{i=1}^n t_i + c_2 \sum_{i=1}^n t_i^2 + c_3 \sum_{i=1}^n t_i^3 - \sum_{i=1}^n d_i t_i \right) = 0 \quad (57.244)$$

Soit après simplification et un petit réarrangement:

$$c_1 \sum_{i=1}^n t_i + c_2 \sum_{i=1}^n t_i^2 + c_3 \sum_{i=1}^n t_i^3 = \sum_{i=1}^n d_i t_i \quad (57.245)$$

Et enfin:

$$\frac{\partial \Phi}{\partial c_3} = \sum_{i=1}^n 2t_i^2 (c_1 + c_2 t_i + c_3 t_i^2 - d_i) = 2 \left(c_1 \sum_{i=1}^n t_i^2 + c_2 \sum_{i=1}^n t_i^3 + c_3 \sum_{i=1}^n t_i^4 - \sum_{i=1}^n d_i t_i^2 \right) = 0 \quad (57.246)$$

Soit après simplification et un petit réarrangement:

$$c_1 \sum_{i=1}^n t_i^2 + c_2 \sum_{i=1}^n t_i^3 + c_3 \sum_{i=1}^n t_i^4 = \sum_{i=1}^n d_i t_i^2 \quad (57.247)$$

Donc en utilisant la notation de l'algèbre linéaire, nous avons finalement le système suivant à résoudre:

$$\begin{bmatrix} n & \sum_{i=1}^n t_i & \sum_{i=1}^n t_i^2 \\ \sum_{i=1}^n t_i & \sum_{i=1}^n t_i^2 & \sum_{i=1}^n t_i^3 \\ \sum_{i=1}^n t_i^2 & \sum_{i=1}^n t_i^3 & \sum_{i=1}^n t_i^4 \end{bmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n d_i \\ \sum_{i=1}^n d_i t_i \\ \sum_{i=1}^n d_i t_i^2 \end{pmatrix} \quad (57.248)$$

et donc il suffit de résoudre ce simple système linéaire soit à la main en utilisant les relations démontrées dans le chapitre d'Algèbre Linéaire, soit avec un simple tableur (du type Microsoft Excel).

9.6. RÉGRESSION LOGISTIQUE (LOGIT)

Bien souvent, les données statistiques disponibles sont relatives à des caractères qualitatifs. Or, comme nous allons le voir, les méthodes d'inférence traditionnelles ne permettent pas de modéliser et d'étudier ce type de caractères. Des méthodes spécifiques doivent être utilisées tenant compte par exemple de l'absence de continuité des variables traitées ou de l'absence d'ordre naturel entre les modalités que peut prendre le caractère qualitatif. Ce sont ces méthodes spécifiques les plus usuelles qui feront l'objet du texte qui suit.

Comme nous l'avons vu plus haut, la régression linéaire simple a donc pour but de modéliser la relation entre une variable dépendante quantitative (non bornée) et une variable explicative quantitative.

Lorsque la "**variable de classe**" Y à expliquer est binaire (oui-non, présence-absence, 0-1, etc.), l'idée est de nous approcher dans un premier temps de celle-ci par une fonction de probabilité $P(Y = 1)$, qui nous donne à l'opposé la probabilité d'appartenir à la classe $Y = 0$, que nous nommerons "**régression logistique**" ou "**régression logit**" ou encore "**régression binomiale**" (très souvent utilisée dans le cadre des réseaux de neurones formels que nous verrons plus loin). Ensuite, dans une deuxième étape, nous définissons pour un cas binaire une valeur "**cutoff**". Par exemple, si nous prenons un cutoff de 0.5 alors les cas pour lesquels $P(Y = 1) > 0.5$ appartiendront à la classe 1 (et inversement dans le cas contraire).

Remarques:

R1. Au fait, la régression logistique n'est qu'une simple loi de distribution de probabilités dans le cas qui nous intéresse (nous verrons une autre régression logistique dans le chapitre d'Économie lors de notre étude des séries temporelles et encore une autre dans le chapitre de Dynamique des Populations).

R2. Il n'est évidemment pas possible d'appliquer systématiquement la régression logistique à n'importe quel type d'échantillon de données! Parfois il faut chercher ailleurs...

R3. Lorsque le nombre de modalités est égal à 2, nous parlons de "**variable dichotomique**" (oui-non) ou d'un "**modèle dichotomique**"; s'il est supérieur à 2, nous parlons de "variables polytomiques" (régression logistique polytomique). Le modèle logit est donc un "**modèle dichotomique**".

Considérons par exemple la variable dichotomique: fin des études. Celle-ci prend deux modalités: "en cours", "a fini". L'âge est une variable explicative de cette variable et nous cherchons à modéliser la probabilité d'avoir terminé ses études en fonction de l'âge.

Exemple:

Pour construire le graphique suivant, nous avons calculé et représenté en ordonnées, pour des jeunes d'âges x différents, le pourcentage de ceux qui ont arrêté leurs études.

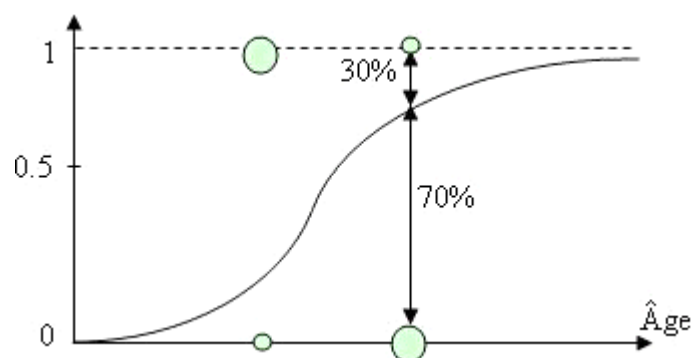


Figure: 57.8 - Partie de la population aux études en fonction de l'âge

Mais comment obtenons-nous pareil graphique avec une variable dichotomique??? Au fait c'est relativement simple... Imaginez un échantillon de 100 individus. Pour ces 100 individus supposez pour un âge donné que 70% "a fini" et 30% sont "en cours". Eh bien la courbe représente simplement la proportion des deux classes pour l'âge donné. Il est même parfois indiqué la grosseur des classes avec des cercles sur toute la longueur des asymptotes horizontales pour bien signifier qu'il s'agit d'une variable dichotomique.

Les points sont distribués selon une courbe en S (une sigmoïde): il y a deux asymptotes horizontales car la proportion est comprise entre 0 et 1. Nous voyons immédiatement qu'un modèle linéaire serait manifestement inadapté (d'autant plus que la variable expliquée d'un modèle linéaire balaie tout l'ensemble des réels et ne se borne pas à l'intervalle $[0, 1]$).

Cette courbe évoquera pour certains, à juste titre, une courbe cumulative représentant une fonction de répartition (d'une loi Normale par exemple, mais d'autres lois continues ont la même allure). Ainsi, pour ajuster une courbe à cette représentation, nous pourrions nous orienter vers la fonction de répartition d'une loi Normale, et au lieu d'estimer les paramètres a et b de la régression linéaire, nous pourrions estimer les paramètres μ, σ de la loi Normale (qui est très similaire à la loi logistique comme nous le démontrerons plus loin). Nous parlons alors d'un "modèle Probit".

La loi qui va nous intéresser cependant est la loi logistique. Contrairement à la loi Normale, nous savons calculer l'expression de sa fonction de répartition (probabilité cumulée) qui est du type (c'est son premier avantage!):

$$P(X \leq x) = \frac{e^{ax+b}}{1 + e^{ax+b}} \quad (57.249)$$

pour une variable de prédiction (explicative) x où P est donc une probabilité bien évidemment comprise entre 0 et 1. Nous verrons un peu plus loin la raison historique de ce choix.

Nous voyons immédiatement que cette dernière relation étant la primitive de la fonction de distribution (voir la démonstration un peu plus bas) il s'agit donc bien d'une fonction de répartition puisque:

$$\int_{-\infty}^{+\infty} f(x)dx = F(x) \Big|_{-\infty}^{+\infty} = \frac{e^{ax+b}}{1 + e^{ax+b}} \Big|_{-\infty}^{+\infty} = \frac{e^{ax+b}}{1 + e^{ax+b}} \Big|_{-\infty}^{+\infty} = 1 - 0 = 1 \quad (57.250)$$

S'il y a plusieurs variables prédictives nous avons alors:

$$P = \frac{e^{a_0 + a_1 x_1 + a_2 x_2 + \dots + a_n x_n}}{1 + e^{a_0 + a_1 x_1 + a_2 x_2 + \dots + a_n x_n}} \quad (57.251)$$

Lorsque nous optons pour la fonction de répartition de la loi logistique, nous obtenons le modèle de régression logistique ou "modèle Logit" pour le choix de la "fonction de lien" et c'est là son deuxième avantage (le plus important!): nous pouvons faire des statistiques sur une variable binaire comme si nous faisons une simple régression linéaire!

Ainsi, nous estimerons la probabilité cumulée d'avoir fini ses études pour un individu d'âge x par (il existe plusieurs manières d'écrire cette loi suivant les habitudes et le contexte) la fonction de répartition logistique suivante:

$$P(X \leq x) = \frac{e^{ax+b}}{1+e^{ax+b}} = \frac{1}{\frac{1+e^{ax+b}}{e^{ax+b}}} = \frac{1}{1+e^{-(ax+b)}} \quad (57.252)$$

il en découle donc la fonction de distribution:

$$f(x) = \frac{dP(X)}{dx} = \frac{d}{dx} \left(1 + e^{-(ax+b)} \right)^{-1} = \frac{ae^{-(ax+b)}}{\left(1 + e^{-(ax+b)} \right)^2} \quad (57.253)$$

Évidemment, en fonction de la valeur de la probabilité nous associons à l'âge x le fait de ne pas avoir fini ses études (état associé à la valeur binaire: 0) ou de les avoir finies (état associé à la valeur binaire: 1).

Indiquons que si a est posé comme unitaire, et b comme nul, nous avons alors la "**loi logistique standard**" donnée par:

$$P(X \leq x) = \frac{1}{1+e^{-x}} \quad (57.254)$$

Nous pouvons calculer aussi l'espérance de la fonction de distribution en appliquant ce qui a déjà été vu au chapitre de Statistiques mais une partie de cette intégrale ne peut être résolue que numériquement par contre... si nous posons:

$$t = ax + b \quad (57.255)$$

comme étant la variable aléatoire alors nous pouvons calculer formellement l'espérance de la loi logistique (le lecteur aura peut-être remarqué que c'est comme si nous posions a comme valant 1 et b comme valant 0). Effectivement, partant de:

$$P(X \leq x) = \frac{1}{1+e^{-(ax+b)}} = \frac{1}{1+e^{-t}} \Rightarrow f(x) = \frac{e^{-t}}{\left(1 + e^{-t} \right)^2} \quad (57.256)$$

il vient alors:

$$E(X) = \int_{-\infty}^{+\infty} f(t) t dt = \int_{-\infty}^{+\infty} \frac{e^{-t}}{\left(1 + e^{-t} \right)^2} t dt \quad (57.257)$$

qui après une intégration numérique donne 0. Nous obtenons alors aussi le résultat suivant (si quelqu'un a le résultat analytique détaillé nous sommes preneurs!):

$$V(X) = E(X^2) - E(X)^2 = E(X^2) = \int_{-\infty}^{+\infty} f(x) x^2 dx = \int_{-\infty}^{+\infty} \frac{e^{-t}}{\left(1 + e^{-t} \right)^2} t^2 dt = \frac{\pi^2}{3} \quad (57.258)$$

Ainsi, nous voyons que si nous posons:

$$P(X \leq x) = \frac{1}{1 + e^{-\frac{\pi}{\sqrt{3}}(ax+b)}} \quad (57.259)$$

Nous retombons sur une fonction de répartition ayant parfaitement les mêmes paramètres de position et de dispersion qu'une loi Normale centrée réduite (moyenne nulle et variance unitaire).

La fonction de répartition:

$$P(X \leq x) = \frac{e^{ax+b}}{1 + e^{ax+b}} \quad (57.260)$$

peut par ailleurs être transformée de la façon suivante:

$$\begin{aligned} P \cdot (1 + e^{ax+b}) &= e^{ax+b} \Rightarrow P + P e^{ax+b} = e^{ax+b} \Rightarrow P = e^{ax+b} (1 - P) \\ \Rightarrow e^{ax+b} &= \frac{P}{1 - P} \end{aligned} \quad (57.261)$$

d'où:

$$ax + b = \ln\left(\frac{P}{1 - P}\right) = \text{logit}(P) \quad (57.262)$$

Au fait c'est ici que réside l'astuce d'origine historique de la régression logistique. Nous transformons une variable P prenant ces valeurs dans 0 à 1 à l'aide du logarithme du rapport $P/(1-P)$ en une variable prenant ses valeurs sur l'ensemble des réels et dès lors il est possible d'y associer une régression linéaire. Certes c'est empirique mais l'idée est bonne!

Ce que certains écrivent aussi...:

$$\text{logit}(P(Y = 1 / X = x)) = \log\left(\frac{P(Y = 1 / X = x)}{1 - P(Y = 1 / X = x)}\right) \quad (57.263)$$

Le résultat de cette dernière transformation est appelé "**logit**". Il est égal au logarithme de "**l'odds**" (sur lequel nous reviendrons très vite plus en détails):

$$\text{odds} = \frac{P}{1 - P} \quad (57.264)$$

Donc lorsque les coefficients a et b ont été déterminés, l'expression précédente permet de déterminer P connaissant x facilement (il s'agit de résoudre une équation linéaire) et inversement! Par ailleurs, puisque x est une variable dichotomique les coefficients sont très facilement interprétables.

Remarque: L'odds est également appelé "**cote**" par analogie à la cote des chevaux au Tiercé. Par exemple, si un cheval a 3 chances sur 4 d'être gagnant (donc in extenso 1 chance sur 4 d'être non-gagnant) sa cote est de 3 contre un 1, soit un odds égal à 3.

Revenons un peu sur l'odds car il est possible d'introduire la notion de fonction logistique en faisant la démarche inverse de celle présentée ci-dessus (soit de commencer par la définition de l'odds pour aller jusqu'au logit) et ceci peut parfois même s'avérer plus pédagogique.

Supposons que nous partions de la taille (hauteur) d'une personne pour prédire si cette personne est un homme ou une femme. Nous pouvons donc parler de probabilité d'être un homme ou une femme. Imaginons que la probabilité d'être un homme pour une hauteur donnée est 90%. Alors l'odds d'être un homme est:

$$\boxed{odds = \frac{P}{1-P}} \quad (57.265)$$

Dans notre exemple, l'odds sera de 0.90/0.10 soit 9. Maintenant, la probabilité d'être une femme sera donc de 0.10/0.90 soit 0.11. Cette asymétrie des valeurs est peu parlante parce que l'odds d'être un homme devrait être l'opposé de l'odds d'être une femme idéalement. Nous résolvons justement cette asymétrie à l'aide du logarithme naturel. Ainsi, nous avons:

$$\ln\left(\frac{P}{1-P}\right) = \ln(9) = 2.197 \text{ et } \ln\left(\frac{(1-P)}{1-(1-P)}\right) = \ln(0.11) = -2.197 \quad (57.266)$$

Ainsi, le logit (logarithme de l'odds) est exactement l'opposé de celui d'être une femme de par la propriété du logarithme:

$$\ln\left(\frac{x}{y}\right) = \ln(x) - \ln(y) = -(\ln(y) - \ln(x)) = -\ln\left(\frac{y}{x}\right) \quad (57.267)$$

Exemple:

Imaginons qu'une banque souhaite faire un scoring de ses débiteurs. Comme elle a plusieurs succursales elle (la banque) construit les tables de données (fictives...) suivantes pour certaines d'entre elles (toutes les succursales ne sont donc pas représentées):

- 1^{ère} succursale:

Montant crédit	Payé	Non Payé
27'200	1	9
27'700	7	3
28'300	13	0
28'400	7	3
29'900	10	1

Tableau: 57.4 - Scoring débiteurs par montant de crédit succursale 1

- 2^{ème} succursale:

Montant crédit	Payé	Non Payé
27'200	0	8
27'700	4	2
28'300	6	3
28'400	5	3
29'900	8	0

Tableau: 57.5 - Scoring débiteurs par montant de crédit succursale 2

- 3^{ème} succursale:

Montant crédit	Payé	Non Payé
27'200	1	8
27'700	6	2
28'300	7	1
28'400	7	2
29'900	9	0

Tableau: 57.6 - Scoring débiteurs par montant de crédit succursale 3

Nous pouvons voir que la proportion totale des bons débiteurs dans les trois succursales est de $91/136 \cong 0.67$.

Quand le crédit est inférieur à 27'500, la proportion de bons débiteurs est de $2/27 \cong 0.07$. Quand le montant des crédits est inférieur à 28'000 la proportion de bons débiteurs est de $19/51 \cong 0.37$.

Quand le montant des crédits est inférieur à 28'500, la proportion de bons débiteurs est de $64/108 \cong 0.59$, et pour des montants inférieurs à 30'000 la proportion est de $91/136 \cong 0.67$.

Nous poserons pour cette régression logistique que $Y = 1$ est un bon risque de crédit et $Y = 0$ est un mauvais risque. Ensuite, nous créons le tableau suivant qui est un récapitulatif des données de toutes les succursales:

Montant crédit	Proportion P
27'200	$=2/27=0.0741$
27'700	$=17/24=0.7083$
28'300	$=26/30=0.8667$
28'400	$=19/27=0.7037$
29'900	$=27/28=0.9643$

Tableau: 57.7 - Proportion des bons débiteurs

Ce qui donne graphiquement en Kilo-francs:

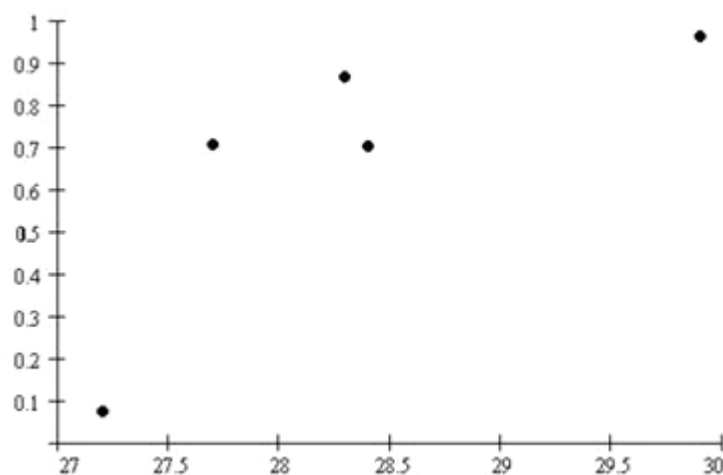


Figure: 57.9 - Pourcentage cumulé des bons débiteurs en fonction du crédit

Une fois ceci fait, nous utilisons la transformation en logit:

$$ax + b = \ln\left(\frac{P}{1-P}\right) = \text{logit}(P) \quad (57.268)$$

Ce qui donne:

Montant crédit	Proportion P	Logit
27'200	0.0741	-2.5257
27'700	0.7083	0.8873
28'300	0.8667	1.8718
28'400	0.7037	0.8650
29'900	0.9643	3.2958

Tableau: 57.8 - Proportion des bons débiteurs et Logit

Une régression linéaire par la méthode des moindres carrés donne:

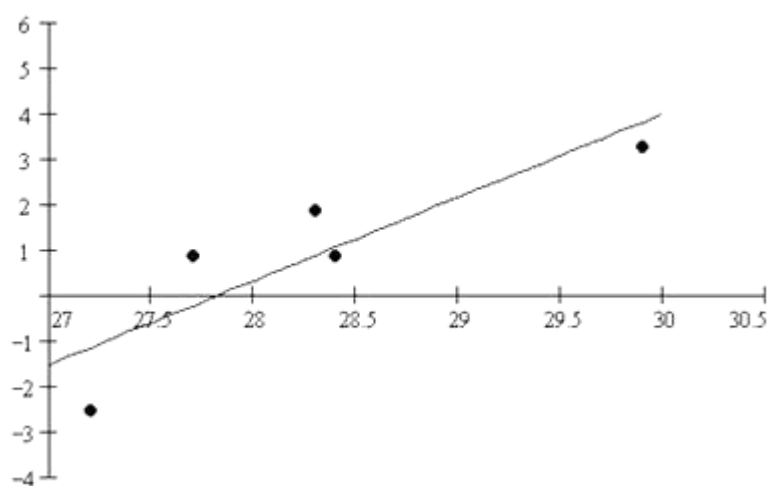


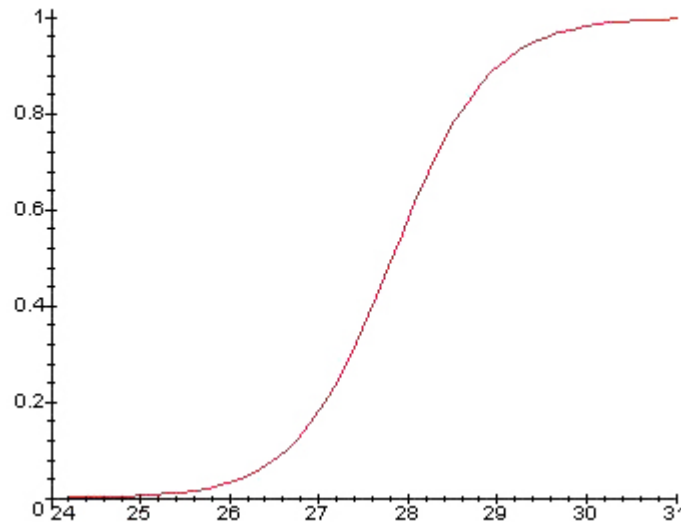
Figure: 57.10 - Logit des bons débiteurs en fonction du montant du crédit

avec pour équation:

$$-51.1116 + 0.0018371x = \ln\left(\frac{P}{1-P}\right) = \text{logit}(P) \quad (57.269)$$

La fonction logistique avec sa représentation vient alors immédiatement (les unités de x sont en milliers de francs!):

$$P(X \leq x) = \frac{1}{1 + e^{-(51.1116 + 0.0018371x)}} \quad (57.270)$$



Ainsi, il est possible de dire dans cet exemple, quelle est la proportion P de bons ou mauvais payeurs en fonction d'une valeur de crédit X plus petite ou égale à une certaine valeur donnée. Puisque 0 est un mauvais risque de crédit, nous voyons que plus les crédits sont élevés, moins le risque est gros (dans ce cas fictif...). Par ailleurs, avec un logiciel comme Minitab, la différence entre les calculs (grossiers) effectués ici à la main et ceux effectués avec l'outil de régression logistique binaire du logiciel sont de l'ordre de 10% (car évidemment... Minitab utilise le concept des estimateurs de maximum de vraisemblance vus dans le chapitre de Statistiques pour déterminer les coefficients et la constante).

Remarque: Les logiciels de statistiques n'utilisent pas les méthode des moindres carrés pour déterminer les coefficients mais la méthode de la log-vraisemblance (cf. chapitre de Statistiques).

9.7. COEFFICIENT DE CORRÉLATION (DÉTERMINATION) GÉNÉRALISÉ

Nous avons vu plus haut ainsi que dans le chapitre de Statistiques plusieurs formulations du coefficient de corrélation linéaire, qui mis au carré était alors nommé coefficient de détermination de Pearson.

Nous avons également vu dans le chapitre de Statistiques le coefficient de corrélation non-paramétrique de Spearman mais celui-ci aussi est uniquement valable dans le cadre d'une relation linéaire (affine).

Afin de généraliser cela, une définition empirique plus générale a été donnée (je n'en ai jamais vu la démonstration ni le nom de la personne qui est à l'origine de cette définition...):

$$R_{x,y}^2 = 1 - \frac{SCR}{SCT} = 1 - \frac{\sum_k (y_k - f_k)^2}{\sum_k (y_k - \bar{y})^2} \quad (57.271)$$

Ainsi, pour n'importe quelle dépendance, nous cherchons les paramètres des équations souhaitées avec les techniques de recherche opérationnelle (voir plus bas dans ce chapitre) qui minimisent le coefficient de corrélation.

Ainsi dans le cas particulier de la régression linéaire puisque nous avons:

$$SCT = SCE + SCR \quad (57.272)$$

Nous retrouvons alors:

$$R_{x,y}^2 = 1 - \frac{SCR}{SCT} = 1 - \frac{SCT - SCE}{SCT} = \frac{SCE}{SCT} \quad (57.273)$$

9.8. INTERPOLATION POLYNOMIALE

Il existe de nombreuses techniques d'interpolation de polynômes plus ou moins complexes et élaborées. Nous nous proposons ici d'en présenter quelques-unes dans un ordre croissant de difficulté et de puissance d'application.

9.8.1. COURBES DE BÉZIER (B-SPLINE)

L'ingénieur russe Pierre Bézier (Peugeot), aux débuts de la Conception Assistée par Ordinateur (C.A.O), dans les années 60, a donné un moyen de définir des courbes et des surfaces à partir de points. Ceci permet la manipulation directe, géométrique, des courbes sans avoir à donner d'équation à la machine!!

Le thème des Courbes de Bézier est une notion à multiples facettes, vraiment très riche, au croisement de nombreux domaines mathématiques très divers: Analyse, Cinématique, Géométrie Différentielle, Géométrie Affine, Géométrie Projective, Géométrie Fractale, Probabilités, Finance (courbe de Taux), ...

Les Courbes de Bézier sont par ailleurs devenues incontournables dans leurs applications concrètes dans l'industrie, l'infographie, ...

Voici l'approche mathématique de cette technique:

D'abord, nous savons que l'équation d'une droite que nous noterons dans le domaine M (par respect des traditions) joignant deux points $A(x_1, y_1), B(x_2, y_2)$ est:

$$M(t) = t(B - A) + A = (1-t)A + tB \quad (57.274)$$

Ce qui est juste puisque lorsque $t = 0$ nous sommes en A et lorsque $t = 1$ nous sommes en B . Donc $t \in [0, 1]$ et le point M parcourt tout le segment $[AB]$. Par construction, si A et B étaient des masses physiques égales à l'unité, $M(t)$ représenterait le barycentre (centre de gravité) du système pour un t donné.

Par définition, le segment $[AB]$ est une "courbe de Bézier de degré 1" avec les points de contrôle A et B et les Polynômes $1-t$ et t sont les "polynômes de Bernstein de degré 1".

Construisons maintenant une courbe paramétrique en rajoutant une 2ème étape à ce qui précède:

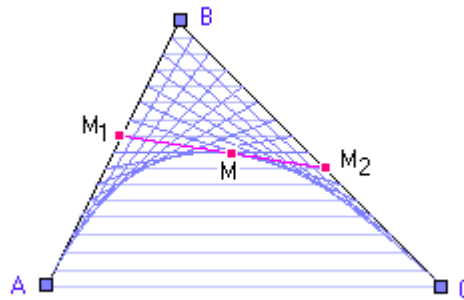


Figure: 57.11 - Courbe de Bézier avec un point supplémentaire

1ère étape:

- Soit $M_1(t)$ le barycentre de $(A, 1-t)$ et (B, t) et où $M_1(t)$ décrit $[AB]$.
- Soit $M_2(t)$ le barycentre de $(B, 1-t)$ et (C, t) et où $M_2(t)$ décrit $[BC]$.

2ème étape:

- Soit $M(t)$ le barycentre de $(M_1(t), 1-t)$, $(M_2(t), t)$.

Par construction, $M(t)$ se situe donc à la même proportion du segment $[M_1(t), M_2(t)]$ que $M_1(t)$ par rapport au segment $[AB]$ ou $M_2(t)$ par rapport au segment $[BC]$.

La courbe obtenue est alors l'enveloppe des segments $[M_1(t), M_2(t)]$: en tout point M , la tangente à la courbe est donc le segment $[M_1(t), M_2(t)]$.

$M(t)$ décrit alors une Courbe de Bézier de degré 2, qui, par construction commence en A et se finit en C , et a pour tangentes $[AB]$ en A et $[BC]$ en C .

C'est en fait un arc de parabole (que nous pourrions noter très logiquement $[ABC]$):

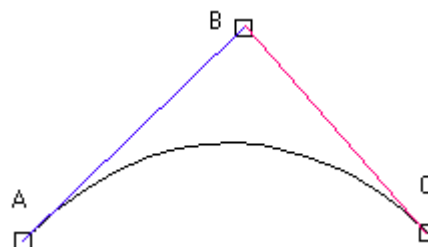


Figure: 57.12 - Arc de parabole $[ABC]$

Par le même schéma, nous pouvons définir une courbe de Bézier de n points P_i soit $i = 1 \dots n$. C'est ce que nous appelons "l'algorithme de Casteljau". Ainsi, soit:

$$\forall r \in [1, n-1] \quad (57.275)$$

Nous avons:

$$b_i^r(t) = (1-t)b_i^{r-1}(t) + t \cdot b_{i+1}^{r-1}(t) \quad (57.276)$$

La récurrence se terminant pour:

$$b_i^{(0)}(t) = P_i \quad (57.277)$$

Ainsi, pour $n = 2$ nous avons:

$$M(t) = b_1^{(1)}(t) = (1-t)b_1^0(t) + t \cdot b_2^0(t) \quad (57.278)$$

Soit:

$$M(t) = b_1^{(1)}(t) = (1-t)b_1^0(t) + t \cdot b_2^0(t) = (1-t)P_1 + t \cdot P_2 \quad (57.279)$$

Ainsi, nous avons forcément avec deux points l'équation d'une droite.

Considérons maintenant $M(t)$ la courbe de Bézier d'ordre 3, nous avons donc les points définis par:

$$M(t) = b_1^{(r)}(t) = (1-t)b_i^{r-1}(t) + t \cdot b_{i+1}^{r-1}(t) \quad (57.280)$$

Nous avons par la relation de récurrence:

$$\begin{aligned} b_1^{(0)}(t) &= P_1 \\ b_2^{(0)}(t) &= P_2 \\ b_3^{(0)}(t) &= P_3 \\ b_1^{(1)}(t) &= (1-t)b_1^{(0)}(t) + tb_2^{(0)}(t) \\ b_2^{(1)}(t) &= (1-t)b_2^{(0)}(t) + tb_3^{(0)}(t) \\ b_3^{(1)}(t) &= (1-t)b_3^{(0)}(t) + tb_4^{(0)}(t) \\ b_1^{(2)}(t) &= (1-t)b_1^{(1)}(t) + tb_2^{(1)}(t) \\ b_2^{(2)}(t) &= (1-t)b_2^{(1)}(t) + tb_3^{(1)}(t) \\ b_3^{(2)}(t) &= (1-t)b_3^{(1)}(t) + tb_4^{(1)}(t) \end{aligned} \quad (57.281)$$

où nous avons éliminé les termes contenant des points non déterminés.

Nous avons donc:

$$\begin{aligned}
b_1^{(1)}(t) &= (1-t)P_1 + tP_2 \\
b_2^{(1)}(t) &= (1-t)P_2 + tP_3 \\
b_1^{(2)}(t) &= (1-t)b_1^{(1)}(t) + tb_2^{(1)}(t) \\
b_2^{(2)}(t) &= (1-t)b_2^{(1)}(t) + tb_3^{(1)}(t)
\end{aligned} \quad (57.282)$$

Il vient alors:

$$M(t) = (1-t) \left((1-t)P_1 + tP_2 \right) + t \left((1-t)P_2 + tP_3 \right) \quad (57.283)$$

Et donc:

$$M(t) = (1-t)^2 P_1 + 2t(1-t)P_2 + t^2 P_3 \quad (57.284)$$

soit sous forme vectorielle (plus conforme à la notation mathématique d'usage):

$$\overrightarrow{OP} = \vec{M}(t) = \begin{pmatrix} x \\ y \end{pmatrix} = (1-t)^2 \overrightarrow{OP_1} + 2t(1-t) \overrightarrow{OP_2} + t^2 \overrightarrow{OP_3} \quad (57.285)$$

et sous forme matricielle:

$$M(t) = \begin{bmatrix} t^2 & t & 1 \end{bmatrix} \begin{bmatrix} 1 & -2 & 1 \\ -2 & 2 & 0 \\ 1 & 0 & 0 \end{bmatrix} \cdot \begin{pmatrix} P_1 \\ P_2 \\ P_3 \end{pmatrix} \quad (57.286)$$

Par le même raisonnement, nous avons pour une courbe de Bézier d'ordre 4:

$$M(t) = (1-t)^3 P_1 + 3t(1-t)^2 P_2 + 3t^2(1-t)P_3 + t^3 P_4 \quad (57.287)$$

soit sous forme vectorielle:

$$\overrightarrow{OP} = \vec{M}(t) = \begin{pmatrix} x \\ y \end{pmatrix} = (1-t)^3 \overrightarrow{OP_1} + 3t(1-t)^2 \overrightarrow{OP_2} + 3t^2(1-t) \overrightarrow{OP_3} + t^3 \overrightarrow{OP_4} \quad (57.288)$$

Ce qui correspond de manière générique à:

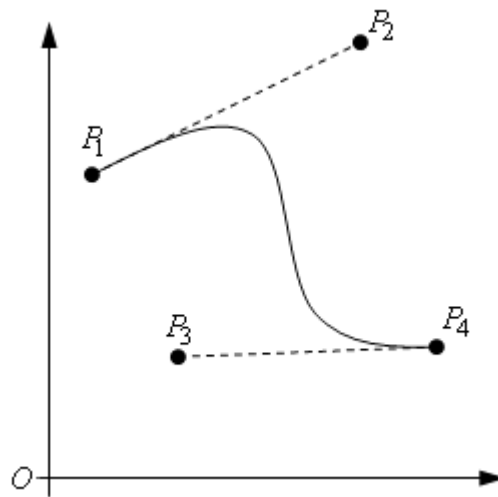


Figure: 57.13 - Courbe de Bézier d'ordre 4

Et là nous pouvons creuser un peu les coefficients en comparant les coefficients de Bézier des courbes d'ordre 2, 3 et 4.

D'abord, reprenons la courbe de Bézier précédente:

$$M(t) = B_1 P_1 + B_2 P_2 + B_3 P_3 + B_4 P_4 = \sum_{i=1}^4 B_i P_i \quad (57.289)$$

Nous remarquons d'abord aisément la proportionnalité suivante:

$$B_i \propto t^{i-1} (1-t)^{n-i} \quad (57.290)$$

et si nous regardons de plus près les coefficients, nous remarquons que nous avons aussi:

$$\begin{array}{ll} \text{ordre 2:} & 1 \\ \text{ordre 3:} & 1 \ 2 \ 1 \\ \text{ordre 4:} & 1 \ 3 \ 3 \ 1 \end{array} \quad (57.291)$$

Il ne s'agit ni plus ni moins que du triangle de Pascal!! Et donc les constantes sont simplement les coefficients binomiaux (cf. chapitre de Calcul Algébrique) donnés pour l'ordre n dans notre cas par:

$$C_{i-1}^{n-1} = \frac{(n-1)!}{(i-1)!((n-1)-(i-1))!} = \frac{(n-1)!}{(i-1)!(n-i)!} \quad (57.292)$$

Ainsi, les polynômes de Bernstein sont donnés par:

$$B_{i,n}(t) = \frac{(n-1)!}{(i-1)!(n-i)!} t^{i-1} (1-t)^{n-i} = C_{i-1}^{n-1} t^{i-1} (1-t)^{n-i} \quad (57.293)$$

et finalement les courbes de Bernstein d'ordre n par:

$$M(t) = \sum_{i=1}^n C_{i-1}^{n-1} t^{i-1} (1-t)^{n-i} P_i \quad (57.294)$$

Au fait, si nous avons noté plus haut la somme sous la forme suivante:

$$M(t) = B_0 P_0 + B_1 P_1 + B_2 P_2 + B_3 P_3 = \sum_{i=0}^3 B_i P_i \quad (57.295)$$

Nous aurions alors les polynômes de Bernstein qui seraient donnés (ce qui est plus respectueux des traditions...) par:

$$B_{i,n}(t) = C_i^n t^i (1-t)^{n-i} \quad (57.296)$$

C'est une relation très pratique car elle permet de calculer facilement et rapidement le polynôme correspondant à une courbe de Bézier d'ordre n .

Nous avons alors:

$$\begin{aligned} x(t) &= \sum_{i=0}^n C_i^n t^i (1-t)^{n-i} x_i \\ y(t) &= \sum_{i=0}^n C_i^n t^i (1-t)^{n-i} y_i \end{aligned} \quad (57.297)$$

Remarque: Une courbe de Bézier est totalement modifiée dès que l'on déplace un point de contrôle. Nous disons alors que la méthode de Bézier est une "méthode globale".

Un exemple très connu des courbes de Bézier d'ordre 3 est l'outil Plume des produits phares Adobe Photoshop ou Adobe Illustrator. Effectivement, ces deux outils créent une succession de courbes de Bézier d'ordre 3 dont le point P_2 est défini après coup à la souris en utilisant des poignées appelées "torseurs" dans le langage de spécialistes Adobe... Voici un exemple pris d'un de ces logiciels fait avec un tracé à la plume de 5 points (soit 4 splines):

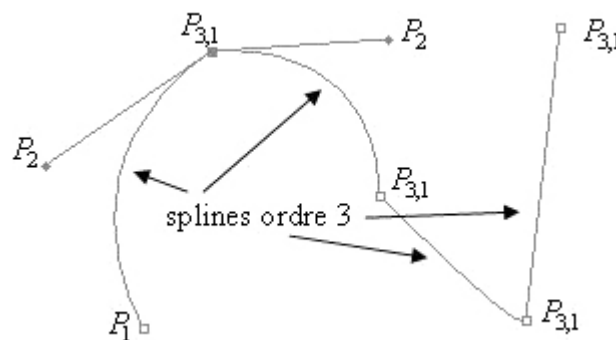


Figure: 57.14 - Exemples de splines avec un logiciel de dessin

Tant que l'utilisateur ne bouge pas les poignées de points alors tous les points sont alignés sur la droite. Nous avons alors l'impression d'avoir une spline d'ordre 2.

 **Exemple:**

Un cercle, dessiné par un logiciel graphique est en pratique composé de 4 arcs de Bézier. Pour observer cette particularité, il suffit de dessiner un cercle avec Illustrator par exemple, puis de le sélectionner pour voir apparaître les points de contrôle des arcs de Bézier qui le forment.

Nous allons nous intéresser à la meilleure façon de choisir les points de contrôle de ces arcs de sorte qu'ils ressemblent à des quarts de cercle, puis nous observerons la différence entre le dessin produit et un vrai cercle:

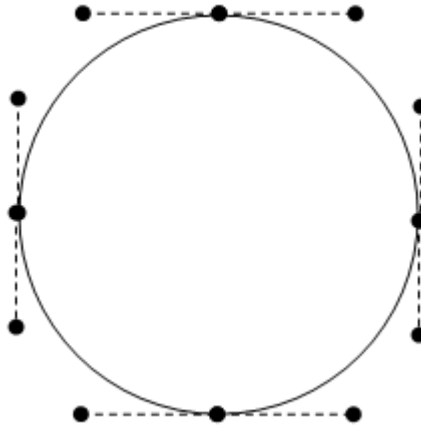
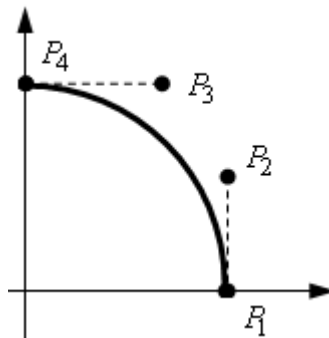


Figure: 57.15 - Exemple de construction d'un cercle avec des courbes de Bézier

Prenons le quart de cercle de rayon 1 centré à l'origine:



Il est approché par un arc de Bézier dont les points de contrôle sont P_1, P_2, P_3, P_4 . Les extrémités de l'arc de Bézier étant P_1 et P_4 , il est naturel de choisir $P_1(1,0)$ et $P_4(0,1)$.

L'intuition nous amène à choisir $P_2(1,k)$ et $P_3(k,1)$ et il reste à trouver une valeur positive de k de sorte que l'arc de Bézier ressemble à un arc de cercle.

Nous obtenons ainsi l'équation paramétrique de l'arc de Bézier

$$\begin{pmatrix} x \\ y \end{pmatrix} = (1-t)^3 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 3t(1-t)^2 \begin{pmatrix} 1 \\ k \end{pmatrix} + 3t^2(1-t) \begin{pmatrix} k \\ 1 \end{pmatrix} + t^3 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad (57.298)$$

Soit:

$$\begin{aligned} x(t) &= (1-t)^3 + 3t(1-t)^2 + 3t^2(1-t)k \\ y(t) &= 3t(1-t)^2k + 3t^2(1-t) + t^3 \end{aligned} \quad (57.299)$$

Nous pouvons par exemple chercher la valeur de k pour laquelle l'arc passe par le point:

$$P = \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2} \right) \quad (57.300)$$

en $t = 0.5$. Il devient alors très simple à partir de l'équation paramétrique de déterminer k . Il s'agit simplement pour x (ou pour y) d'une simple équation à une inconnue.

9.8.2. MÉTHODE D'EULER

Il s'agit là de la méthode numérique la plus simple (elle est triviale et dans l'idée assez proche de la méthode de Newton même si leur objectif n'est pas le même). En fait elle ne fournit qu'une approximation (au sens très large du terme) d'une fonction $f(x)$ dont la dérivée est connue.

Ici les points choisis sont équidistants, c'est-à-dire $x_i = x_0 + ih$ (h étant le "pas"). Nous notons $f'(x_i)$ la valeur exacte et y_i , la valeur approchée.

Il y a plusieurs méthodes pour procéder (comme souvent):

1. Graphiquement:

Nous nous déplaçons d'un pas h en x en suivant le vecteur de pente $f'(x,y)$

Par construction, nous savons (cf. chapitre de Calcul Différentiel Et Intégral) que (nous adoptons une notation un peu particulière dans ce contexte):

$$\frac{y_{i+1} - y_i}{h} = f'(x_i, y_i) \quad (57.301)$$

qui correspond donc bien à la pente (non instantanée bien sûr!) en (x_i, y_i) de la "courbe intégrale" passant par ce point. D'où:

$$y_{i+1} = y_i + hf'(x_i, y_i) \quad (57.302)$$

2. Analytiquement:

Nous remplaçons dans la dernière relation y_{i+1} par $y(x_i + h)$. Nous obtenons alors:

$$y(x_i + h) = y_i + hf'(x_i, y_i) \quad (57.303)$$

appelée "équation aux différences pour la méthode d'Euler".

L'application en est triviale et ne nécessite pas d'exemples particuliers.

9.8.3. POLYNÔME DE COLLOCATION

Soit $y = f(x)$ une fonction connue sous forme explicite ou sous forme tabulée, et supposons qu'un certain nombre de valeurs:

$$f_i = f(x_i) \quad i = 0, 1, \dots, n \quad (57.304)$$

en sont données. Les points (x_i, f_i) sont appelés les "points d'appui".

"Interpoler" f signifie estimer les valeurs de f pour des abscisses situées entre x_0 et x_n , c'est-à-dire dans l'intervalle d'interpolation, par une fonction approximante $y = p(x)$, qui vérifie les "conditions de collocations" (rien à voir avec votre colocataire!):

$$p(x_i) = f_i \quad i = 0, 1, \dots, n$$

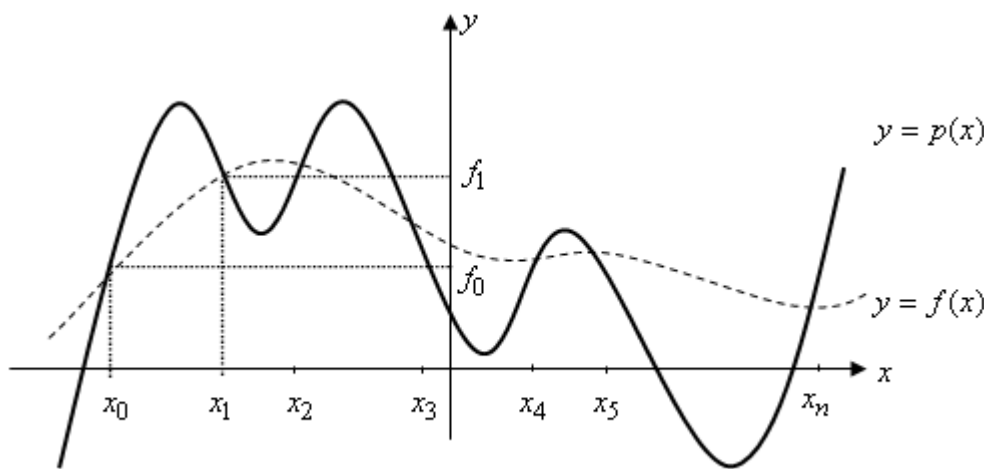


Figure: 57.16 - Illustration du concept d'interpolation

La fonction p s'appelle "fonction de collocation" sur les x_i . Lorsque p est un polynôme, nous parlons de "polynôme de collocation" ou de "polynôme d'interpolation".

"Extrapoler" f signifie approcher $f(x)$ par $p(x)$ pour des abscisses situées "hors" de l'intervalle d'interpolation.

Remarque: Il va sans dire que l'interpolation est un outil très important pour tous les chercheurs, statisticiens et autres.

Quand nous connaissons un polynôme de degré n en $n+1$ points, nous pouvons donc connaître par une méthode simple (mais pas très rapide - mais il existe plusieurs méthodes) complètement ce polynôme.

Pour déterminer le polynôme, nous allons utiliser les résultats exposés précédemment lors de notre étude des systèmes d'équations linéaires. Le désavantage de la méthode présentée ici est qu'il faut deviner à quel type de polynôme nous avons affaire et savoir quels sont les bons points qu'il faut choisir...

Un exemple particulier devrait suffire à la compréhension de cette méthode, la généralisation en étant assez simple (voir plus loin).

Soit un polynôme du second degré:

$$P_2(x) = a_0 + a_1x + a_2x^2 \quad (57.305)$$

et nous avons connaissance des points suivants (dont vous remarquerez l'ingéniosité des points choisis par les auteurs de ces lignes...):

$$P(0) = 4, P(1) = 8, P(-1) = 2 \quad (57.306)$$

Nous en déduisons donc le système d'équations:

$$\begin{aligned} a_0 + a_1 \cdot 0 + a_2 \cdot 0^2 &= 4 \\ a_0 + a_1 \cdot 1 + a_2 \cdot 1^2 &= 2 \\ a_0 + a_1 \cdot (-1) + a_2 \cdot (-1)^2 &= 8 \end{aligned} \quad (57.307)$$

Système qui une fois résolu dans les règles de l'art (cf. chapitre d'Algèbre Linéaire) nous donne:

$$a_0 = 4, a_1 = -3, a_2 = 1 \quad (57.308)$$

Voyons le cas général:

Théorème: Soient (x_i, f_i) des points d'appui, avec $x_i \neq x_j$ si $i \neq j$. Alors il existe un polynôme $p_n(x)$ de degré inférieur ou égal à n , et un seul, tel que $p_n(x_i) = f_i$ pour $i = 0, 1, \dots, n$.

Démonstration:

Posons:

$$p_n(x) = a_0 + a_1x + \dots + a_nx^n \quad (57.309)$$

Les conditions de collocation:

$$p(x_i) = f_i \quad i = 0, 1, \dots, n \quad (57.310)$$

s'écrivent donc:

$$\begin{aligned} a_0 + a_1x_0 + a_2x_0^2 + \dots + a_nx_0^n &= f_0 \\ a_0 + a_1x_1 + a_2x_1^2 + \dots + a_nx_1^n &= f_1 \\ \vdots & \\ a_0 + a_1x_n + a_2x_n^2 + \dots + a_nx_n^n &= f_n \end{aligned} \quad (57.311)$$

Il s'agit d'un système de $n+1$ équations à $n+1$ inconnues.

Son déterminant s'écrit:

$$D_n = \det \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix} = \prod_{n \geq i > j \geq 0} (x_i - x_j) \quad (57.312)$$

relation que nous appelons "**déterminant de Vandermonde**". Nous savons que si le système a une solution, le déterminant du système doit être non nul (cf. chapitre d'Algèbre linéaire).

Montrons par l'exemple (en reprenant un polynôme du même degré que celui que nous avons utilisé plus haut) que le déterminant se calcule selon la relation suivante précédente (le lecteur généralisera par récurrence):

Donc dans le cas $n = 2$, nous considérons le déterminant:

$$\det \begin{bmatrix} 1 & x_0^1 & x_0^2 \\ 1 & x_1^1 & x_1^2 \\ 1 & x_2^1 & x_2^2 \end{bmatrix} \quad (57.313)$$

qui correspond donc au système (pour rappel):

$$\begin{aligned} a_0 \cdot 1 + a_1 x_0^1 + a_2 x_0^2 &= f_0 \\ a_0 \cdot 1 + a_1 x_1^1 + a_2 x_1^2 &= f_1 \\ a_0 \cdot 1 + a_1 x_2^1 + a_2 x_2^2 &= f_2 \end{aligned} \quad (57.314)$$

Calculons ce déterminant suivant la colonne 1 (en faisant usage des cofacteurs comme démontré dans le chapitre d'Algèbre linéaire):

$$\begin{aligned} \det \begin{bmatrix} 1 & x_0^1 & x_0^2 \\ 1 & x_1^1 & x_1^2 \\ 1 & x_2^1 & x_2^2 \end{bmatrix} &= 1 \cdot \det \begin{bmatrix} x_1^1 & x_1^2 \\ x_2^1 & x_2^2 \end{bmatrix} - 1 \cdot \det \begin{bmatrix} x_0^1 & x_0^2 \\ x_2^1 & x_2^2 \end{bmatrix} + 1 \cdot \det \begin{bmatrix} x_0^1 & x_0^2 \\ x_1^1 & x_1^2 \end{bmatrix} \\ &= x_1^1 x_2^2 - x_2^1 x_1^2 - x_0^1 x_2^2 + x_2^1 x_0^2 + x_0^1 x_1^2 - x_1^1 x_0^2 \end{aligned} \quad (57.315)$$

Ce dernier polynôme peut s'écrire:

$$D_n = (x_1 - x_0)(x_2 - x_0)(x_2 - x_1) \quad (57.316)$$

Ce qui s'écrit:

$$D_n = \prod_{n \geq i > j \geq 0}^{n-2} (x_i - x_j) \quad (57.317)$$

Comme les x_i sont dans l'énoncé de notre problème tous différents tels que $x_i \neq x_j$ alors le système a une solution unique! Ce qui prouve qu'il existe toujours alors un polynôme d'interpolation.

□C.Q.F.D.

Il convient cependant de noter qu'il ne s'agit pas d'une méthode de régression polynomiale. Car avec une méthode de régression polynomiale, nous pourrions choisir un degré supérieur au nombre de points que nous avons!

9.8.4. MÉTHODE DE LAGRANGE

La méthode d'interpolation polynomiale de Lagrange (très utilisée dans la pratique car l'algorithme est très simple et donc efficace) considère que nous avons initialement $n + 1$ points tels que:

$$\begin{aligned} y_0 &= f(x_0) \\ y_1 &= f(x_1) \\ &\dots \\ y_i &= f(x_i) \\ &\dots \\ y_n &= f(x_n) \end{aligned} \quad (57.318)$$

Et que nous cherchons donc un polynôme de collocation qui passe par tous les points. L'idée de l'interpolation polynomiale de Lagrange est alors simple et très astucieuse (comme toujours il fallait y penser...). Observez le graphique ci-dessous où nous avons 5 points par lesquels nous cherchons à faire passer un polynôme de collocation:

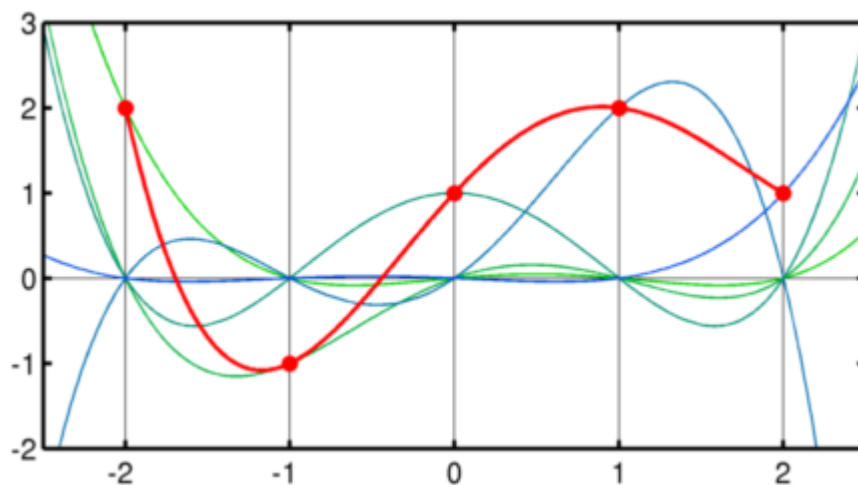


Figure: 57.17 - Illustration du concept d'interpolation de Lagrange (source: Wikipédia)

Pour trouver le polynôme de collocation en rouge (supposons que nous ne le connaissons préalablement pas), l'idée est que pour chaque x_i nous associons un polynôme de degré n et non nul en $f(x_i)$ mais nul en tous les autres points mesurés $x_j \neq x_i$ et que tous les autres polynômes associés aux autres points x_i y soient nul. Pour cela, comme nous le voyons dans le graphique ci-dessus, il faut donc que pour chaque point i nous ayons un polynôme associé qui ait n racines tel que:

$$\begin{aligned}
f_0(x) &= A_0 (x - x_1)(x - x_2)(x - x_3)(x - x_4) \\
f_1(x) &= A_1 (x - x_0)(x - x_2)(x - x_3)(x - x_4) \\
f_2(x) &= A_2 (x - x_0)(x - x_1)(x - x_3)(x - x_4) \\
f_3(x) &= A_3 (x - x_0)(x - x_1)(x - x_2)(x - x_4) \\
f_4(x) &= A_4 (x - x_0)(x - x_1)(x - x_2)(x - x_3)
\end{aligned} \quad (57.319)$$

Donc nous voyons bien que nous avons 5 polynômes, chacun avec 4 racines (donc de degré 4) et qui sont respectivement nuls sur tous les points $x_j \neq x_i$ (vous pouvez le voir sur le graphique). Les coefficients A_i sont des constantes à déterminer.

Maintenant, rien nous nous empêche de sommer puisque pour chaque x_i nous aurons toujours la bonne valeur $f(x_i)$ (puisque les autres polynômes y sont nuls). Soit en généralisant l'écriture, la somme devient:

$$\begin{aligned}
P(x) &= A_0 (x - x_1)(x - x_2)(x - x_3) \dots (x - x_n) + A_1 (x - x_0)(x - x_2)(x - x_3) \dots (x - x_n) \\
&+ A_i (x - x_0) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_n) + A_n (x - x_0)(x - x_1)(x - x_2) \dots (x - x_{n-1})
\end{aligned} \quad (57.320)$$

En injectant respectivement $x_0, x_1, \dots, x_i, \dots, x_n$ dans la relation précédente, il vient:

$$\begin{aligned}
P(x_0) &= f(x_0) = A_0 (x_0 - x_1)(x_0 - x_2)(x_0 - x_3) \dots (x_0 - x_n) \\
P(x_1) &= f(x_1) = A_1 (x_1 - x_0)(x_1 - x_2)(x_1 - x_3) \dots (x_1 - x_n) \\
P(x_2) &= f(x_2) = A_2 (x_2 - x_0)(x_2 - x_1)(x_2 - x_3) \dots (x_2 - x_n) \\
&\dots \\
P(x_n) &= f(x_n) = A_n (x_n - x_0)(x_n - x_1)(x_n - x_2) \dots (x_n - x_{n-1})
\end{aligned} \quad (57.321)$$

Nous en déduisons alors immédiatement:

$$\begin{aligned}
A_0 &= \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3) \dots (x_0 - x_n)} \\
A_1 &= \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3) \dots (x_1 - x_n)} \\
A_2 &= \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3) \dots (x_2 - x_n)} \\
&\dots \\
A_n &= \frac{f(x_n)}{(x_n - x_0)(x_n - x_1)(x_n - x_2) \dots (x_n - x_{n-1})}
\end{aligned} \quad (57.322)$$

En substituant les valeurs des constantes dans l'expression initiale de la somme, nous obtenons:

$$\begin{aligned}
 P(x) = & \frac{(x-x_1)(x-x_2)(x-x_3)\dots(x-x_n)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)\dots(x_0-x_n)} f(x_0) \\
 & + \frac{(x-x_0)(x-x_2)(x-x_3)\dots(x-x_n)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)\dots(x_1-x_n)} f(x_1) + \dots \\
 & + \frac{(x-x_0)(x-x_1)(x-x_2)\dots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)(x_n-x_2)\dots(x_n-x_{n-1})} f(x_n)
 \end{aligned} \quad (57.323)$$

Ce qui peut s'écrire sous forme condensée:

$$P(x) = \sum_{j=0}^n \left(\prod_{i=0, i \neq j}^n \frac{x-x_i}{x_j-x_i} \right) f(x_j) \quad (57.324)$$

Où le terme:

$$L(x_j) = \prod_{i=0, i \neq j}^n \frac{x-x_i}{x_j-x_i} \quad (57.325)$$

est appelé "**coefficients d'interpolations de polynômes de Lagrange**".

10. RECHERCHE DES RACINES

Bien des équations rencontrées en pratique ou en théorie ne peuvent pas être résolues exactement par des méthodes formelles ou analytiques. En conséquence, seule une solution numérique approchée peut être obtenue en un nombre fini d'opérations.

Évariste Galois a démontré, en particulier, que l'équation $P_n(x) = 0$ ne possède pas de solution algébrique (sauf accident...) si $P_n(x)$ est un polynôme de degré supérieur à 4.

Il existe un grand nombre d'algorithmes permettant de calculer les racines de l'équation $f(x) = 0$ avec une précision théorique arbitraire. Nous n'en verrons que les principaux. Attention, la mise en oeuvre de tels algorithmes nécessite toujours une connaissance approximative de la valeur cherchée et celle du comportement de la fonction près de la racine. Un tableau des valeurs de la fonction et sa représentation graphique permettent souvent d'acquérir ces connaissances préliminaires.

Si l'équation à résoudre est mise sous la forme $g(x) = h(x)$, nous traçons les courbes représentant g et h . Les racines de l'équation $g(x) = h(x)$ étant données par les abscisses des points d'intersection des deux courbes.

Remarque: Avant de résoudre numériquement l'équation $f(x) = 0$, il faut vérifier que la fonction f choisie satisfasse certaines contraintes. Il faut par exemple, que la fonction f soit strictement monotone au voisinage de la racine $\bar{x}$, lorsque la méthode de Newton est appliquée. Il est souvent utile, voire indispensable, de déterminer un intervalle $[a, b]$ tel que:

- f est continue sur $[a, b]$ ou $\mathbb{C}^1$
- $f(a) \cdot f(b) < 0$
- $\exists \bar{x}$ unique, $\bar{x} \in [a, b], f(\bar{x}) = 0$

10.1. MÉTHODE DES PARTIES PROPORTIONNELLES

La mise en oeuvre, sur calculatrice, de cette méthode est particulièrement simple. Les conditions à vérifier étant seulement:

- f est continue
- f est monotone dans un voisinage de la racine $\bar{x}$

Dans un petit intervalle, nous pouvons remplacer une courbe par un segment de droite. Il y a plusieurs situations possibles mais en voici une particulière généralisable facilement à n'importe quoi:

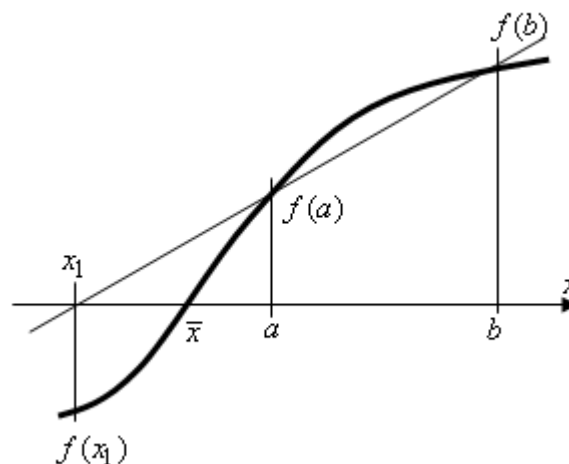


Figure: 57.18 - Approximation d'une courbe par un segment de droite

Sur cette figure, nous tirons à l'aide des théorèmes de Thalès (cf. chapitre de Géométrie Euclidienne):

$$\frac{x_1 - a}{f(a)} = \frac{x_1 - b}{f(b)} \quad (57.326)$$

d'où:

$$x_1 = \frac{a \cdot f(b) - b \cdot f(a)}{f(b) - f(a)} = a - (b - a) \frac{f(a)}{f(b) - f(a)} \quad (57.327)$$

Si $|f(a)| \ll |f(b)|$, nous pouvons négliger $f(a)$ au dénominateur et il vient:

$$x_1 \cong a - (b - a) \frac{f(a)}{f(b)} = a - (b - a)k \quad (57.328)$$

L'algorithme consiste donc à réaliser les étapes suivantes:

1. Choisir a et b , calculer $f(a)$ et $f(b)$
2. Déterminer $x_1 = a - (b - a)k$. Si $|f(x_1)|$ est assez petit, nous arrêtons le calcul et affichons x_1 et $f(x_1)$
3. Sinon nous procédons comme suit:
 - nous remplaçons b par a et $f(b)$ par $f(a)$
 - nous remplaçons a par x_1 et $f(a)$ par $f(x_1)$
 - nous retournons au point (2)

10.2. MÉTHODE DE LA BISSECTION

La condition préalable à satisfaire pour cette méthode est de trouver un intervalle $[a, b]$ tel que:

1. $f(x)$ est continue sur $[a, b]$
2. $f(a) \cdot f(b) < 0$

Il faut encore fixer $\varepsilon > 0$ qui est défini comme la borne supérieure de l'erreur admissible.

La méthode consiste à appliquer successivement les 4 étapes suivantes:

1. Calcul de $x = (a + b) / 2$
2. Évaluation de $f(x)$
3. Si $|f(x)| < \varepsilon$ alors le travail est terminé, il faut afficher x et $f(x)$
4. Sinon on procède comme suit:
 - on remplace a par x si $f(x) \cdot f(a) > 0$
 - on remplace b par x si $f(x) \cdot f(b) > 0$ ou $f(x) \cdot f(a) < 0$
 - on retourne en (1)

L'étape (3) impose la condition $|f(x)| < \varepsilon$ pour l'arrêt des calculs. Il est parfois préférable de choisir un autre critère de fin de calcul. Celui-ci impose à la solution calculée d'être confinée dans un intervalle de longueur ε contenant $\bar{x}$. Ce test s'énonce comme suit:

3'. Si $|b - a| < \varepsilon$, le travail est terminé et $x = (a + b)/2$ est affiché. Il est bien sûr évident que $|x - \bar{x}| < \varepsilon/2$

10.3. MÉTHODE DE LA SÉCANTE (REGULA FALSI)

La méthode de la sécante (ou "regula falsi" pour: régulièrement fausse) en méthodes numériques, est toujours un algorithme de recherche d'un zéro. Pour voir cela, considérons le schéma suivant:

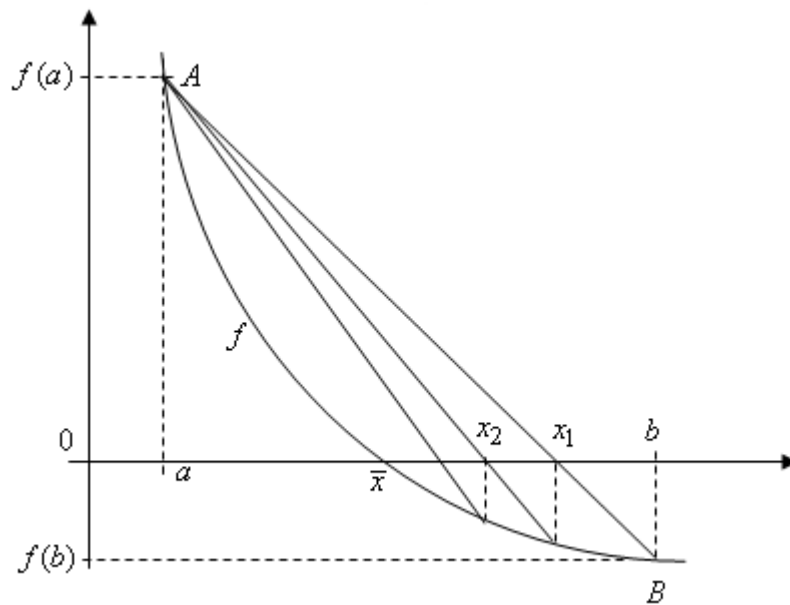


Figure: 57.19 - Illustration de la méthode de la sécante

Les conditions préalables sont les suivantes:

Il faut déterminer un intervalle $[a, b]$ tel que:

1. $f(x)$ est continue sur $[a, b]$
2. $f(a) \cdot f(b) < 0$

Si P_n est le point de coordonnées $(x_n, 0)$, alors les points A, B, P_n sont alignés sur la sécante. La proportion suivante (Thalès) est donc vraie:

$$\frac{a - b}{a - x_n} = \frac{f(a) - f(b)}{f(a) - 0} \quad (57.329)$$

nous en déduisons:

$$x_n = \frac{af(b) - bf(a)}{f(b) - f(a)} \quad (57.330)$$

La méthode consiste à appliquer successivement les étapes suivantes:

1. Calcul de $x_n = \frac{af(b) - bf(a)}{f(b) - f(a)}$

2. Évaluation de $f(x_n)$

3. Si $|f(x_n)| < \varepsilon$, le travail est terminé. Il faut afficher $x_n < \varepsilon$

4. Sinon nous procédons comme suit:

- nous remplaçons a par x_n si $f(x_n) \cdot f(a) > 0$

- nous remplaçons b par x_n si $f(x_n) \cdot f(b) > 0$ ou $f(x_n) \cdot f(a) < 0$

- nous retournons en (1)

La condition (3) peut être remplacée par la condition:

3'. Si $|x_{n+1} - x_n| < \varepsilon$, alors le travail est terminé et nous affichons x_{n+1}

Remarque: Si l'intervalle $[a,b]$ contient plusieurs racines, cette méthode converge vers l'une d'entre elles. Toutes les autres sont malheureusement perdues.

10.4. MÉTHODE DE NEWTON

Quelques années après sa découverte de théorie de la gravitation, Newton développa une technique extraordinaire qui permet de calculer les solutions d'une équation quelconque avec une rapidité (convergence phénoménale). Cette convergence surnaturellement rapide a été utilisée pour démontrer certains des résultats théoriques les plus marquants du 20ème siècle: le théorème de stabilité de Kolmogorov, le théorème de plongement isométrique de Nash... À elle seule, cette technique transcende la distinction entre mathématique pure et mathématique appliquée.

Pour étudier la méthode de Newton (appelée aussi "**méthode de Newton-Raphson**" ou encore "**schéma d'approximation de Newton**") dans le plan (donc à une variable explicative), considérons la figure suivante:

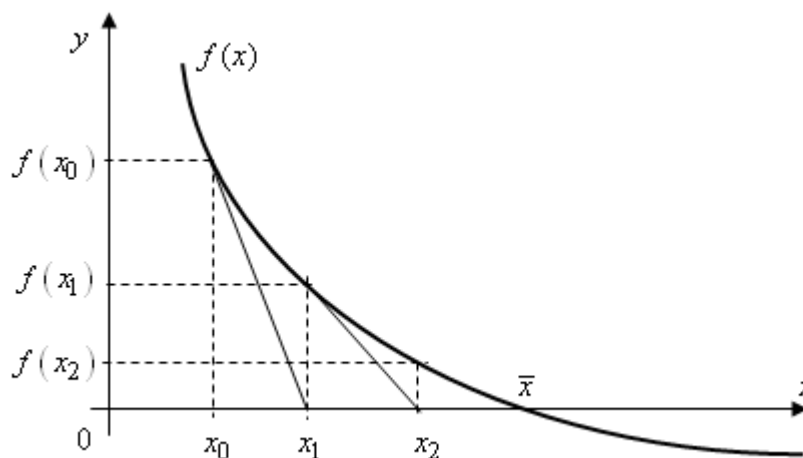


Figure: 57.20 - Illustration de la méthode de Newton dans le plan

Si x_0 est une approximation de la racine $\bar{x}$, nous remarquons que x_1 en est une meilleure. x_1 est l'intersection de la tangente à la courbe en $(x_0, f(x_0))$ et de l'axe des abscisses. x_2 est encore une meilleure approximation de $\bar{x}$, x_2 est obtenu de la même manière que x_1 mais à partir de $(x_1, f(x_1))$.

La méthode de Newton consiste en la formalisation de cette constatation géométrique.

Pour utiliser cette technique, rappelons que si nous prenons une fonction f qui est dérivable en x_0 , alors nous pouvons la réécrire sous la forme (cf. chapitre de Calcul Différentiel Et Intégral):

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \mathcal{G}(x - x_0) \quad (57.331)$$

où $f'(x_0)$ est la dérivée de f en x_0 et $\mathcal{G}(x - x_0)$ est une fonction qui tend vers 0 comme $(x - x_0)^n$ pour $n \geq 2$ lorsque x tend vers x_0 (c'est un terme correctif qui sous-tend la suite des termes du développement de Taylor).

En appliquant ce résultat à la résolution de $f(x) = 0$, nous obtenons:

$$0 = f(\bar{x}) = f(x_0) + f'(x_0)(\bar{x} - x_0) + \mathcal{G}(\bar{x} - x_0) \quad (57.332)$$

La fonction $\mathcal{G}(\bar{x} - x_0)$ empêche la résolution de cette équation par rapport à $\bar{x}$. En négligeant le terme $\mathcal{G}(\bar{x} - x_0)$, l'équation se réécrit:

$$0 = f(\bar{x}) = f(x_0) + f'(x_0)(\bar{x} - x_0) \quad (57.333)$$

et se résout aisément par rapport à $\bar{x}$. Pour voir cela, commençons par $\bar{x} = x_1$:

$$\boxed{x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}} \quad (57.334)$$

Mais x_1 ne satisfait pas, en général, l'égalité $f(x_1) = 0$. Mais comme nous l'avons déjà souligné, $|f(x_1)|$ est plus petit que $|f(x_0)|$ si la fonction f satisfait à certaines conditions.

La méthode de Newton consiste à remplacer l'équation:

$$0 = f(\bar{x}) = f(x_0) + f'(x_0)(\bar{x} - x_0) \quad (57.335)$$

par:

$$0 = f(x_n) + f'(x_n)(x_{n+1} - x_n) \quad (57.336)$$

et à résoudre itérativement cette équation.

Les conditions suivantes sont suffisantes pour assurer la convergence de la méthode:

Dans un intervalle $[a, b]$ contenant x_0 et $\bar{x}$ il faut que:

1. La fonction soit deux fois dérivable

2. La dérivée f' ne s'annule pas (monotonie)

3. La deuxième dérivée soit continue et ne s'annule pas (pas de point d'inflexion)

Remarque: Il suffit souvent de vérifier les conditions (1) et (2) pour que le processus soit convergent.

La condition (2) est évidente, en effet si $f'(x) = 0$ alors l'itération peut conduire à une erreur de calcul (singularité).

La condition (3) est moins évidente, mais le graphique suivant illustre un cas de non-convergence. Dans ce cas, le processus a une boucle calculant alternativement x_i et x_j .

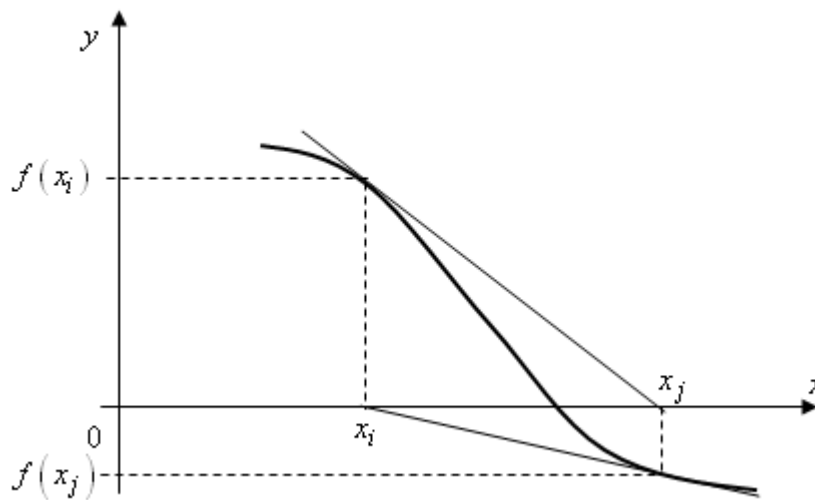


Figure: 57.21 - Exemple de cas de non-convergence avec la méthode de Newton

Si la fonction f est donnée analytiquement, sa dérivée peut être déterminée analytiquement. Mais dans bien des cas, il est utile, voire indispensable de remplacer $f'(x_n)$ par le quotient différentiel:

$$f'(x_n) \cong \frac{f(x_n + h) - f(x_n)}{h} \quad (57.337)$$

où h doit être choisi suffisamment petit pour que la différence:

$$\left| f'(x_n) - \frac{f(x_n + h) - f(x_n)}{h} \right| \quad (57.338)$$

soit elle aussi suffisamment petite.

L'itération s'écrit alors:

$$x_{n+1} = x_n - h \frac{f(x_n)}{f(x_n + h) - f(x_n)} \quad (57.339)$$

Si la méthode de résolution est convergente, l'écart entre x_{n+1} et $\bar{x}$ diminue à chaque itération. Ceci est assuré, par exemple, si l'intervalle $[a,b]$ contenant x_{n+1} , voit sa longueur diminuer à chaque étape. La méthode de Newton est intéressante car la convergence est quadratique:

$$|x_{n+1} - \bar{x}| < \alpha |x_n - \bar{x}|^2 \quad (57.340)$$

alors que la convergence des autres méthodes est linéaire:

$$|x_{n+1} - \bar{x}| < \alpha |x_n - \bar{x}| \quad (57.341)$$

Considérons, par exemple, la méthode de la bisection vue précédemment. À chaque itération la longueur de l'intervalle $[a,b]$ diminue de moitié. Ceci nous assure que l'écart $|x_{n+1} - \bar{x}|$ est réduit de moitié à chaque étape du calcul:

$$|x_{n+1} - \bar{x}| \leq 0.5 |x_n - \bar{x}| \quad (57.342)$$

Pour démontrer la convergence quadratique de la méthode de Newton, il faut utiliser les développements limités de f et f' au voisinage de $\bar{x}$:

$$\begin{aligned} f(x_n) &\cong \underbrace{f(\bar{x})}_{=0} + \frac{1}{1!}(x_n - \bar{x})f'(\bar{x}) + \frac{1}{2!}(x_n - \bar{x})^2 f''(\bar{x}) \\ f'(x_n) &= f'(\bar{x}) + \frac{1}{1!}(x_n - \bar{x})f''(\bar{x}) \end{aligned} \quad (57.343)$$

Mais:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad (57.344)$$

donc:

$$x_{n+1} \cong x_n - \frac{\frac{1}{1!}(x_n - \bar{x})f'(\bar{x}) + \frac{1}{2!}(x_n - \bar{x})^2 f''(\bar{x})}{f'(\bar{x}) + \frac{1}{1!}(x_n - \bar{x})f''(\bar{x})} \quad (57.345)$$

En soustrayant $\bar{x}$ à gauche et à droite de l'égalité et en mettant les deux termes du second membre au même dénominateur, il vient:

$$x_{n+1} - \bar{x} \cong (x_n - \bar{x})^2 \frac{0.5 f''(\bar{x})}{f'(\bar{x}) + (x_n - \bar{x}) f''(\bar{x})} \quad (57.346)$$

et dès que $x_n - \bar{x}$ est assez petit, le dénominateur peut être simplifié.

$$x_{n+1} - \bar{x} \cong (x_n - \bar{x})^2 0.5 \frac{f''(\bar{x})}{f'(\bar{x})} \quad (57.347)$$

ce qui montre bien que la convergence est quadratique.

Signalons enfin que nous étudierons la méthode de Newton à plusieurs variables dans le cadre de l'étude de la recherche (optimisation) non linéaire. C'est le choix pédagogique qui nous a semblé le plus judicieux.

Voici une application avec Maple 4.00b de cette méthode:

```
> with(plots): with(plottools):
Une fonction quelconque dont on cherche la racine
> f:=x->exp(x)*x^2-36;
> D(f)(x);
> Définition du point de départ
> x[0]:=3;
Définition du nombre d'étapes
> n:=7;
Équation de la tangente
> g:=x->f(x[i-1])+D(f)(x[i-1])*(x-x[i-1]);
> for i from 1 by 1 to n do;
> x[i]:=evalf(solve(g(x)=0,x));
> od;
> lines:={}:
> for i from 1 by 1 to n do;
> lines:=lines union {line([x[i-1],0],[x[i-1],f(x[i-1])],color=green), line([x[i-1],f(x[i-1])],[x[i],0],
,color=green)};
> od;
On peut jouer avec le x=.... pour mieux voir les itérations sur le graphique
> display({plot(f(x),x=2..3.01)} union lines);
```

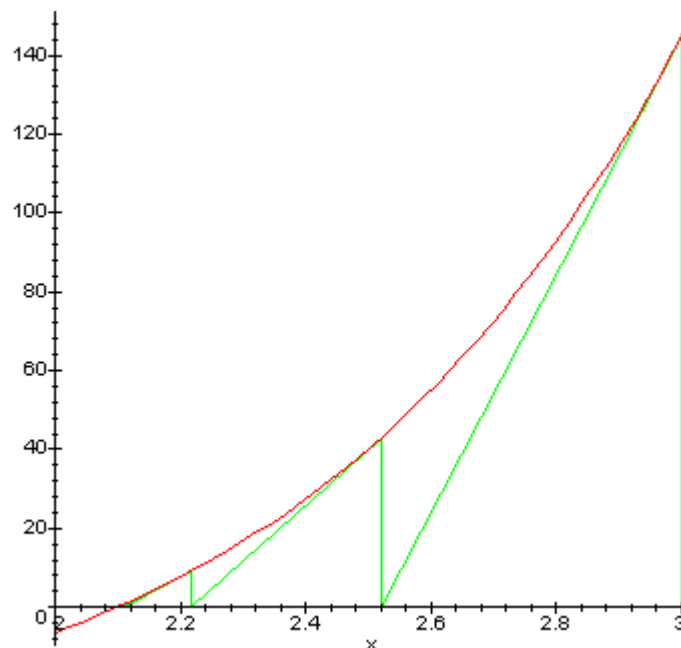


Figure: 57.22 - Application avec Maple 4.00b de la méthode de Newton

11. DÉRIVÉES NUMÉRIQUES

De nombreuses techniques de modélisation ou de résolution numériques que nous verrons plus loin utilisent les dérivées comme la recherche d'optimum (voir plus loin), les méthodes des éléments finis (voir plus loin). Par exemple, pour ne citer que le cas le plus connu, le solveur de MS Office Excel des versions 2007 et antérieures propose quelques unes des dérivées numériques que nous allons étudier ici et réutiliser plus loin:

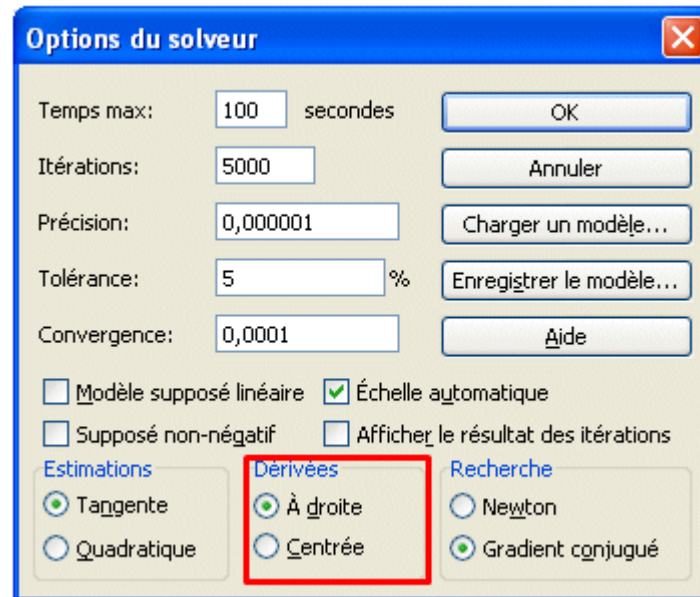


Figure: 57.23 - Capture d'écran du solveur de MS Excel 2003

Afin de permettre donc un traitement sur calculateur, les différentes dérivées présentes dans de nombreux algorithmes doivent être approchées numériquement. Pour ce faire, nous utilisons le principe des différences finies centrées qui s'appuie sur les développements en série de Taylor suivants (cf. chapitre Suites Et Séries):

$$f(x) = \sum_{k=0}^n \frac{1}{k!} \frac{d^k f(x_0)}{dx^k} (x - x_0)^k = \sum_{k=0}^n \frac{d^{(k)} f(x_0)}{k!} (x - x_0)^k \quad (57.348)$$

Nous avons alors sur la base de ce principe le développement au deuxième ordre:

$$\begin{aligned} f\left(x_0 + \frac{d_x}{2}\right) &= f(x_0) + \frac{d_x}{2} f'(x_0) + \frac{1}{2} \left(\frac{d_x}{2}\right)^2 f''(x_0) + \dots \\ f\left(x_0 - \frac{d_x}{2}\right) &= f(x_0) - \frac{d_x}{2} f'(x_0) + \frac{1}{2} \left(\frac{d_x}{2}\right)^2 f''(x_0) - \dots \end{aligned} \quad (57.349)$$

Il alors en négligeant les termes d'ordre supérieur et en soustrayant et simplifiant les deux séries ci-dessus:

$$f'(x_0) = \frac{f\left(x_0 + \frac{d_x}{2}\right) - f\left(x_0 - \frac{d_x}{2}\right)}{d_x} + \varepsilon(d_x^2) \quad (57.350)$$

Relation que nous appelons "**dérivée centrée avec estimation tangente**" (car nous négligeons tous les termes non linéaires). Nous retrouvons également souvent cette dernière relation sous la forme suivante:

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0 - h)}{2h} + O(h^2) \quad (57.351)$$

Maintenant, voyons ce que nous appelons les "**dérivées à droite**" ou aussi "**dérivées avant**" ("**forward derivate**" en anglais) qui consistent simplement dans l'application de l'algorithme intuitif suivant:

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \quad (57.352)$$

et accessoirement nous pouvons aussi définir les "**dérivées à gauche**", appelées également "**dérivées arrières**" ("**backward derivate**" en anglais):

$$f'(x_i) = \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \quad (57.353)$$

Donc nous voyons que les dérivées centrales nécessitent plus de calculs mais sont aussi plus précises.

Nous pouvons également développer des relations plus élaborées en prenant des développements de Taylor à des ordres supérieurs et donc c'est sans fin...

12. INTÉGRALES NUMÉRIQUES

Considérons la figure suivante:

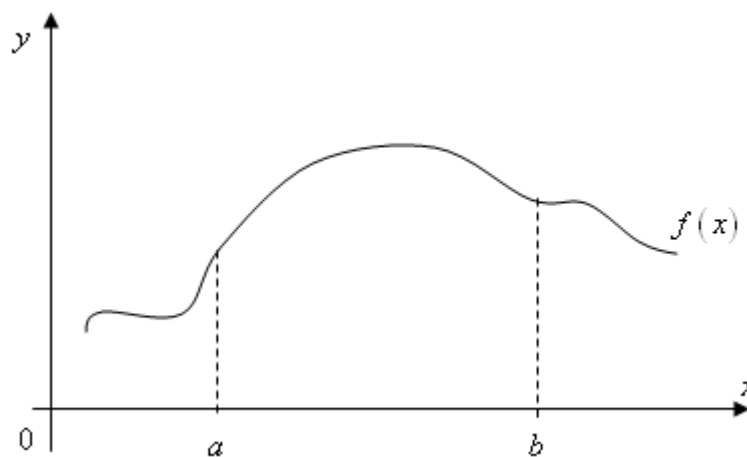


Figure: 57.24 - Illustration d'un intervalle sous une courbe

Nous désirons calculer l'aire comprise entre l'axe x , la courbe de f et les droites d'équations $x = a$ et $x = b$. Nous supposons dans ce cas que la fonction f est à valeurs positives:

$$f(x) \geq 0 \quad \forall x \in [a, b] \quad (57.354)$$

Ce problème, dans sa généralité, est difficile voire impossible à résoudre analytiquement. Voici donc quelques méthodes numériques permettant le calcul approché de cette aire (ces méthodes sont utilisées parfois dans les entreprises par les employés qui n'ont que des tableurs du type Microsoft Excel ou OpenOffice Calc pour calculer des intégrales).

12.1. MÉTHODE DES RECTANGLES

Nous subdivisons l'intervalle $[a, b]$ en n sous-intervalles dont les bornes sont x_i . Les longueurs de ces sous-intervalles sont $h_i = x_{i+1} - x_i$. Nous construisons les rectangles dont les côtés sont h_i et $f(x_i)$.

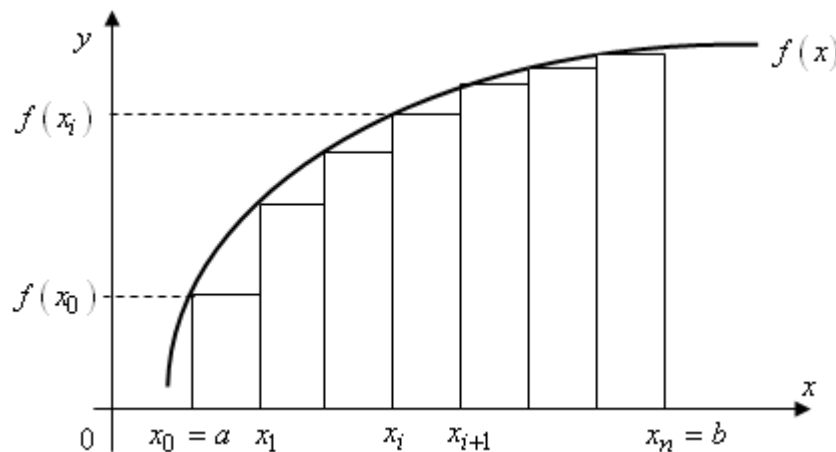


Figure: 57.25 - Approche de l'aire sous une courbe par des rectangles inférieurs gauche

L'aire de ces rectangles vaut:

$$A_G = \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_i) = \sum_{i=0}^{n-1} h_i f(x_i) \quad (57.355)$$

Si les h_i sont suffisamment petits, A_G est une bonne approximation de l'aire cherchée approchée par la gauche.

Nous pouvons recommencer cet exercice en choisissant h_i et $f(x_{i+1})$ comme côtés des rectangles (donc l'approche par la droite). Nous obtenons alors:

$$A_D = \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_{i+1}) = \sum_{i=0}^{n-1} h_i f(x_{i+1}) \quad (57.356)$$

La figure correspondante est la suivante:

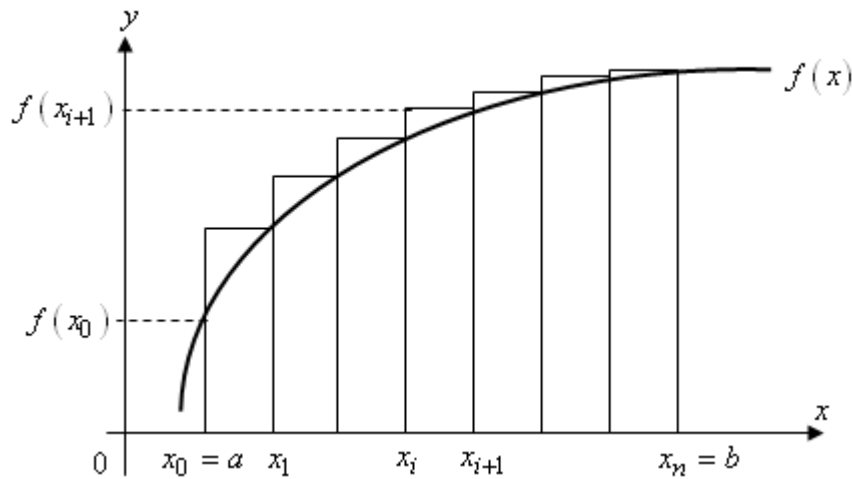


Figure: 57.26 - Approche de l'aire sous une courbe par des rectangles supérieurs droite

Encore une fois, l'aire de ces rectangles approche l'aire cherchée. Afin de simplifier la programmation, il est utile de choisir des intervalles de longueur identique:

$$|x_{i+1} - x_i| = h \quad \forall i \quad (57.357)$$

Si nous avons n rectangles, h vaut alors $(b - a) / n$. Les aires A_D et A_G deviennent:

$$A_D = h \sum_{i=0}^{n-1} f(x_{i+1}) \quad A_G = h \sum_{i=0}^{n-1} f(x_i) \quad (57.358)$$

12.2. MÉTHODE DES TRAPÈZES

Afin d'augmenter la précision des calculs, il est possible de calculer:

$$A_T = \frac{(A_D + A_G)}{2} \quad (57.359)$$

Dans le cas où tous les intervalles sont de longueur égale, A_T vaut:

$$\begin{aligned} A_T &= \frac{1}{2} h \left(\sum_{i=0}^{n-1} f(x_{i+1}) + \sum_{i=0}^{n-1} f(x_i) \right) = \frac{1}{2} h \left(\sum_{i=1}^n f(x_i) + \sum_{i=0}^{n-1} f(x_i) \right) \\ &= \frac{1}{2} h \left(f(x_0) + 2 \sum_{i=1}^{n-1} f(x_i) + f(x_n) \right) \end{aligned} \quad (57.360)$$

Il existe une foule d'autres méthodes permettant la résolution de ce problème (dont la méthode de Monte-Carlo que nous verrons plus loin).

Dans le cas où la fonction f n'est pas à valeurs positives, nous ne parlons plus d'aire mais de "**somme de Riemann**". Les sommes à calculer sont alors:

$$S_D = h \sum_{i=0}^{n-1} f(x_{i+1}) \quad S_G = h \sum_{i=0}^{n-1} f(x_i) \quad (57.361)$$

et:

$$S_T = \frac{1}{2} h \left(f(x_0) + 2 \sum_{i=1}^{n-1} f(x_i) + f(x_n) \right) \quad (57.362)$$

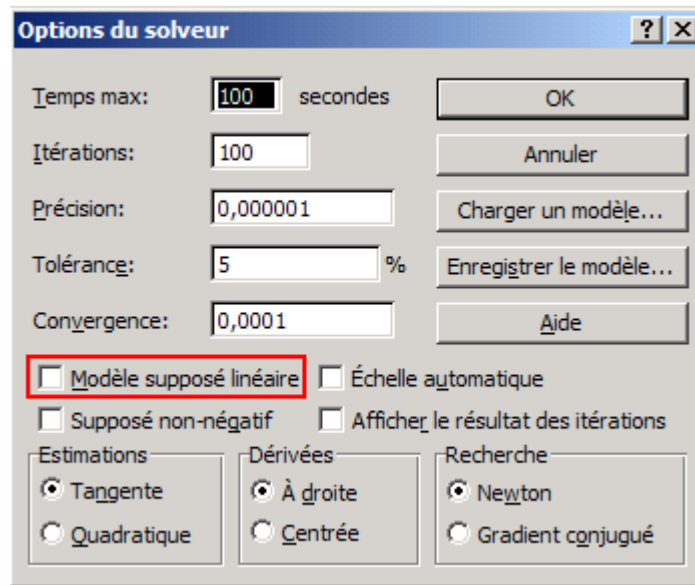
Tous les calculs doivent être conduits de la même manière, mais les résultats peuvent être positifs, négatifs ou nuls.

13. PROGRAMMATION (OPTIMISATION) LINÉAIRE

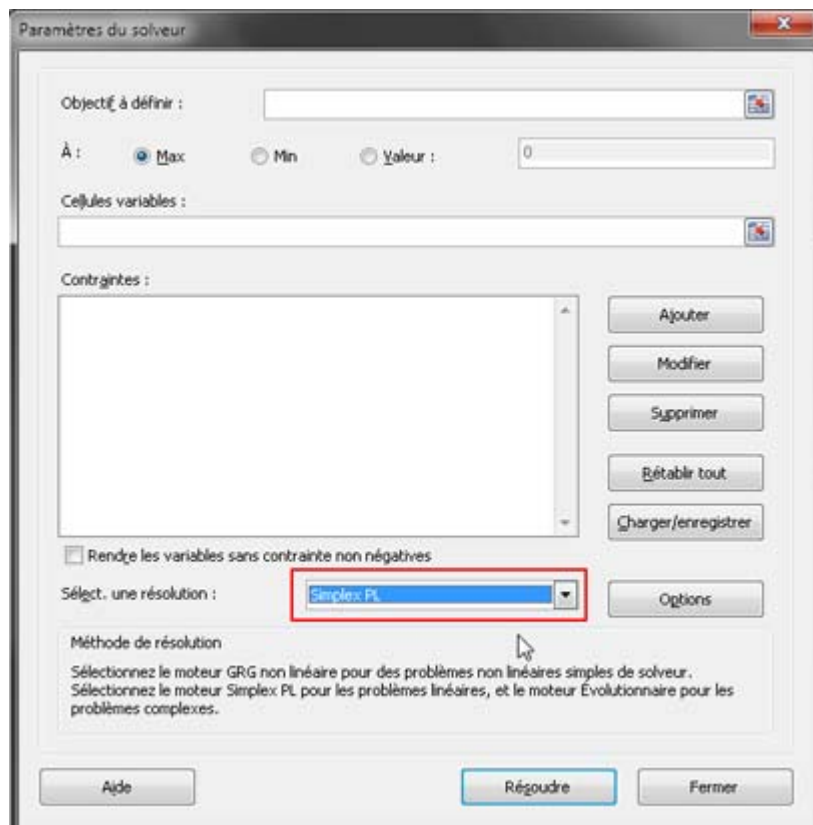
L'objectif de la programmation linéaire (P.L.) est de trouver la valeur optimale d'une fonction linéaire soumise à un système d'équations d'inégalités constitué de contraintes elles aussi linéaires. La fonction à optimiser est baptisée "**fonction économique**" (utilisée en économie dans le cadre d'optimisations) et on la résout en utilisant une méthode dite "**méthode du simplexe**" (voir plus loin) dont la représentation graphique consiste en un "**polygone des contraintes**".

Remarques:

R1. La programmation linéaire est beaucoup utilisée (pour ne citer que les cas les plus connus) dans la logistique (problème à flot maximal dit aussi "**problème de transport**"), la finance d'entreprise ou encore aussi en théorie de la décision lorsque nous devons résoudre un jeu à stratégie mixte (voir le chapitre de Théorie de la décision et des jeux pour un exemple pratique). C'est pour cette raison que Microsoft Excel 12.0 et antérieur intègre un outil appelé le "solveur" dans lequel il existe une option appelée "modèle supposé linéaire" qui alors impose l'utilisation du modèle du simplexe que nous allons voir ci-après:



ou encore depuis la version 2010 du même logiciel (dont l'interface a complètement changé):



R2. Dans le cadre de la résolution de problèmes où interviennent des produits de deux variables, nous parlons alors logiquement de "**programmation quadratique**" ou plus simplement de "**programmation non linéaire**". C'est typiquement le cas en économétrie dans la modélisation des portefeuilles (cf. chapitre d'Économie) ou dans le prévisionnel. Nous étudierons par ailleurs plus loin une version simplifiée et particulière des modèles correspondants qui sont: la méthode de Newton, de quasi-Newton, des gradients conjugués et la GRG non linéaire.

R3. Les programmations quadratique et linéaire sont réunies dans le cadre général de ce que nous appelons la "**recherche opérationnelle**".

La recherche opérationnelle a pour domaine l'étude de l'optimisation de processus quels qu'ils soient. Il existe de nombreux algorithmes s'inspirant des problèmes du type de ceux exposés lors de notre étude de la programmation linéaire. Nous nous attarderons en particulier sur l'algorithme le plus utilisé qui est "**l'algorithme du simplexe**".

Lorsque l'on peut modéliser un problème sous forme d'une fonction économique à maximiser dans le respect de certaines contraintes, alors on est typiquement dans le cadre de la programmation linéaire.

Soit une fonction économique Z telle que:

$$Z = c_1 x_1 + c_2 x_2 + \dots + c_n x_n = \sum_{i=1}^n c_i x_i \quad (57.363)$$

où les x_i sont des variables qui influent sur la valeur de Z , et c_i les poids respectifs de ces variables modélisant l'importance relative de chacune de ces variables sur la valeur de la fonction économique.

Les contraintes relatives aux variables s'expriment par le système linéaire suivant:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &\leq B_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &\leq B_2 \\ &\dots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n &\leq B_m \end{aligned} \quad (57.364)$$

Sous forme générale et matricielle ce genre de problème s'écrit:

$$\begin{aligned} \min \sum_{j=1}^n c_j x_j &= Z \\ \sum_{j=1}^n a_{ij} x_j &= b_i, \quad i = 1, \dots, m \\ x_j &\geq 0, \quad j = 1, \dots, n \end{aligned} \quad (57.365)$$

Exemple:

Une usine fabrique 2 types de pièces $P1$ et $P2$ usinées dans deux ateliers $A1$ et $A2$. Les temps d'usage sont pour $P1$ de 3 heures dans l'atelier $A1$ et de 6 heures dans l'atelier $A2$ et pour $P2$ de 4 heures dans l'atelier $A1$ et de 3 heures dans l'atelier $A2$.

Le temps de disponibilité hebdomadaire des ressources humaines (ouvriers) de l'atelier A1 est de 160 heures et celui de l'atelier A2 de 180 heures.

La marge bénéficiaire est de 1'200.- pour une pièce P1 et 1'000.- pour une pièce P2.

Quelle production de chaque type doit-on fabriquer pour maximiser la marge hebdomadaire?

Le problème peut se formaliser de la façon suivante (formulation canonique):

$$\begin{aligned} A1: 3x_1 + 4x_2 &\leq 160 \\ A2: 6x_1 + 3x_2 &\leq 180 \\ x_1, x_2 &\geq 0 \end{aligned} \quad (57.366)$$

La fonction économique à maximiser étant:

$$Z = 1'200x_1 + 1'000x_2 \quad (57.367)$$

Résolution graphique du problème (ou méthode du "polygone des contraintes"): les contraintes économiques et de signe sont représentées graphiquement par des demi-plans. Les solutions, si elles existent, appartiennent donc à cet ensemble appelé "**région des solutions admissibles**":

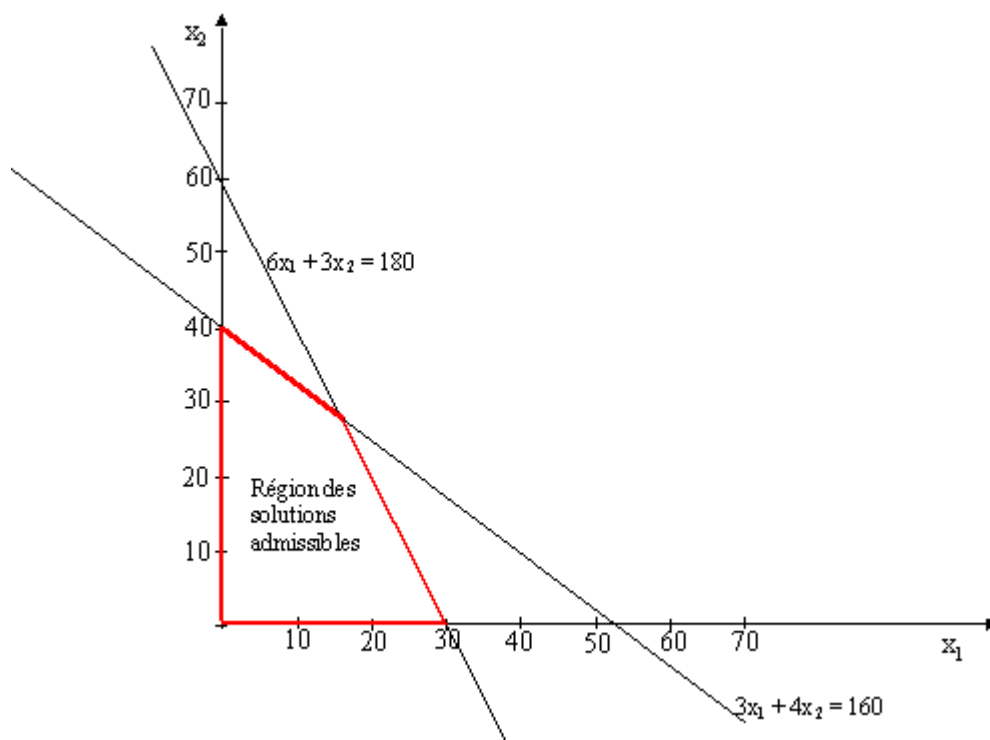


Figure: 57.27 - Illustration d'un problème de recherche opérationnelle simple avec région des solutions admissibles

Remarque: Dans le cas général, pour ceux qui aiment le vocabulaire des mathématiciens..., la donnée d'une contrainte linéaire correspond géométriquement à la donnée d'un demi-espace d'un espace à n dimensions (n étant le nombre de variables). Dans les cas élémentaires, l'ensemble des points de l'espace qui vérifient toutes les contraintes est un convexe limité par des portions d'hyperplan (voir le cas 2 variables, facile à illustrer). Si la fonction de coût est linéaire, l'extremum est à un sommet (facile à voir). L'algorithme du simplexe de base (voir plus loin) part d'un sommet et va au sommet d'à côté qui maximise localement le coût. Et recommence tant que c'est possible.

Pour trouver les coordonnées des sommets, on peut utiliser le graphique si les points sont faciles à déterminer.

Il s'agit donc de chercher à l'intérieur de ce domaine (connexe), le couple (x_1, x_2) maximisant la fonction économique.

Or, l'équation Z est représentée par une droite de pente constante (-1.2) dont tous les points (x_1, x_2) fournissent la même valeur Z pour la fonction économique.

En particulier, la droite $1'200x_1 + 1'000x_2$ passe par l'origine et celle-ci fournit une valeur nulle à la fonction économique. Pour augmenter la valeur de Z et donc la fonction économique, il suffit d'éloigner de l'origine (dans le quart de plan $x_1 \geq 0, x_2 \geq 0$) la droite de pente -1.2. Évidemment on remarque alors très vite que la méthode du simplexe ne fonctionnera plus si le polygone des contraintes ne contient pas l'origine!

Pour respecter les contraintes, cette droite sera déplacée, jusqu'à l'extrême limite où elle n'aura plus qu'un point d'intersection (éventuellement un segment) avec la région des solutions admissibles.

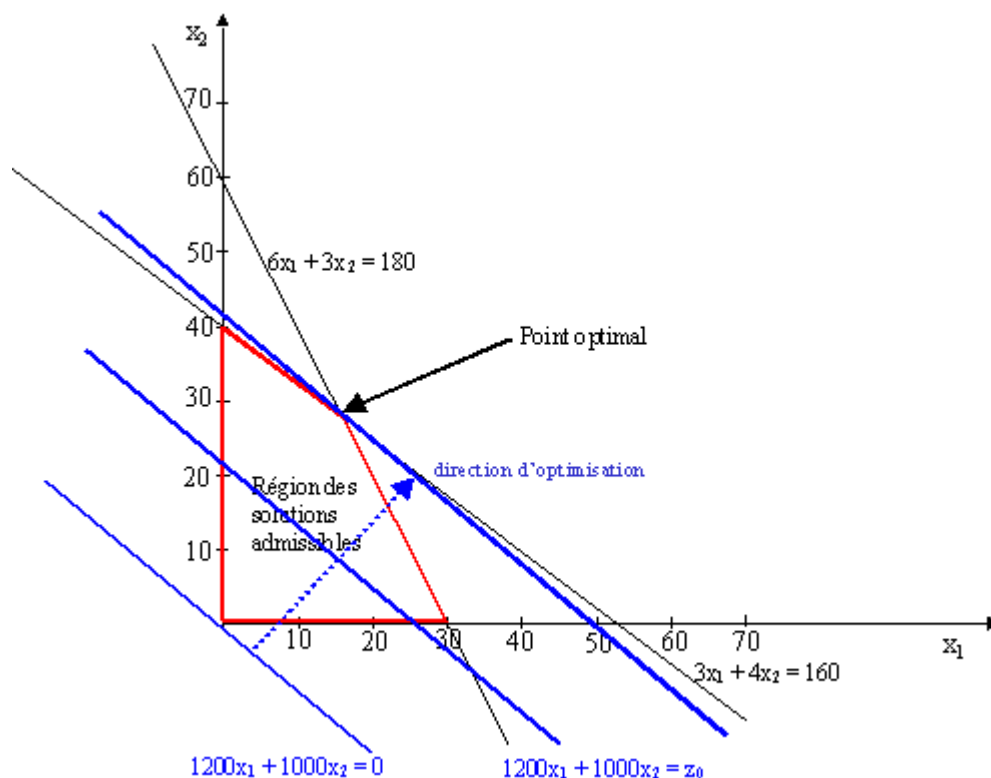


Figure: 57.28 - Recherche de solutions graphiquement avec la droite de la fonction économique

La solution optimale se trouve donc nécessairement sur le pourtour de la région des solutions admissibles et les parallèles formées par la translation de la fonction économique s'appellent les "**droites isoquantes**" ou "**droites d'isocoûts**"...

Voyons maintenant comment résoudre ce problème de manière analytique avant de passer à la partie théorique.

Nous avons donc le "**système canonique**":

$$\begin{aligned} A1: 3x_1 + 4x_2 &\leq 160 \\ A2: 6x_1 + 3x_2 &\leq 180 \\ x_1, x_2 &\geq 0 \end{aligned} \quad (57.368)$$

avec:

$$Z = 1'200x_1 + 1'000x_2 \quad (57.369)$$

Nous introduisons d'abord les "**variables d'écart**" afin de transformer les 2 inégalités en des égalités. Le système d'équations prend alors une "**forme standard**":

$$\begin{aligned} 3x_1 + 4x_2 + a \cdot x_3 &= 160 \\ 6x_1 + 3x_2 + b \cdot x_4 &= 180 \\ 1'200x_1 + 1'000x_2 &= Z \end{aligned} \quad (57.370)$$

Donc pour $x_1, x_2 \geq 0$ fixés, les variables d'écarts dont les coefficients sont toujours unitaires, mesurent la distance à parcourir jusqu'aux sommets.

Il va sans dire que la technique des variables d'écart peut être utilisée pour les systèmes linéaires (ou non linéaires). Dès lors, un système d'optimisation contrainte avec inégalités, peut toujours être ramené à un système d'optimisation avec égalités.

Remarque: Il y a autant de variables d'écart que d'inéquations!

Pour la suite, nous avons remarqué, suite à une relecture de ce chapitre, que la technique des tableaux souvent présentée dans les livres et sur les sites Internet n'apportait finalement absolument rien à la compréhension profonde du mécanisme de résolution (même si pour programmer informatiquement la méthode cela est plus pratique). Comme le but de ce site est de démontrer toujours avec le maximum de détails le principe de fonctionnement des choses il va donc de soi que nous allons opter pour une approche purement algébrique. Voyons donc cela en revenant au système avec les variables d'écart et la fonction économique mais un peu réarrangée:

$$\begin{aligned} x_3 &= 160 - (3x_1 + 4x_2) \\ x_4 &= 180 - (6x_1 + 3x_2) \\ Z &= 1'200x_1 + 1'000x_2 \end{aligned} \quad (57.371)$$

La contrainte A1 devient alors:

$$x_3 \geq 0 \quad (57.372)$$

et la contrainte A2 devient respectivement:

$$x_4 \geq 0 \quad (57.373)$$

Dès lors, le problème se résume à maximiser Z avec les contraintes:

$$x_1, x_2, x_3, x_4 \geq 0 \quad (57.374)$$

Partons donc d'une solution réalisable évidente qui au vu des contraintes est trivialement:

$$x_1 = x_2 = 0 \quad (57.375)$$

Dès lors, avec le système:

$$\begin{aligned} x_3 &= 160 - (3x_1 + 4x_2) \\ x_4 &= 180 - (6x_1 + 3x_2) \\ Z &= 1'200x_1 + 1'000x_2 \end{aligned} \quad (57.376)$$

nous trouvons immédiatement:

$$\begin{aligned} x_3 &= 160 \\ x_4 &= 180 \\ Z &= 0 \end{aligned} \quad (57.377)$$

Les paramètres dans l'état actuel peuvent se résumer à:

$$x_1 = 0, x_2 = 0, x_3 = 160, x_4 = 180 \quad (57.378)$$

Pour progresser, le but sera de faire croître Z et pour cela nous allons ne faire croître qu'une variable, en choisissant celle qui a le plus grand coefficient (poids) dans:

$$Z = 1'200x_1 + 1'000x_2 \quad (57.379)$$

c'est-à-dire x_1 (nous pensons que c'est ainsi que Z croîtra le plus vite). Nous parlons alors de x_1 comme "**direction du pivot**". Nous gardons donc $x_2 = 0$ et nous faisons croître x_1 avec le système qui se réduit alors à:

$$\begin{aligned} x_3 &= 160 - 3x_1 \\ x_4 &= 180 - 6x_1 \\ Z &= 1'200x_1 \end{aligned} \quad (57.380)$$

Avec donc $x_2 = 0$ et pour commencer $x_1 = 1$, nous avons:

$$\begin{aligned}
 x_3 &= 160 - 3 = 157 \\
 x_4 &= 180 - 6 = 154 \quad (57.381) \\
 Z &= 1'200
 \end{aligned}$$

et nous voyons que les contraintes $x_1, x_2, x_3, x_4 \geq 0$ sont toujours respectées, il en va de même si x_1 vaut 2, 3, 4, 5, ... et ce jusqu'à 31, car dès lors:

$$\begin{aligned}
 x_3 &= 160 - 3 \cdot 31 = 67 \\
 x_4 &= 180 - 6 \cdot 31 = -6 \quad (57.382) \\
 Z &= 1'200 \cdot 31 = 37'200
 \end{aligned}$$

et l'une des variables d'écart étant devenue négative, les contraintes $x_1, x_2, x_3, x_4 \geq 0$ ne sont plus toutes satisfaites et donc cette solution n'est pas réalisable.

La question dans le cas général consiste à se demander jusqu'à combien (la valeur la plus contraignante, in extenso la plus petite) nous pouvons faire croître x_1 tout en maintenant la condition $x_1, x_2, x_3, x_4 \geq 0$ lorsque $x_2 = 0$? Et la réponse est assez simple:

$$\begin{aligned}
 x_3 &= 160 - 3x_1 \Rightarrow x_1 = \frac{160}{3} \\
 x_4 &= 180 - 6x_1 \Rightarrow x_1 = \frac{180}{6} = 30 \quad (57.383)
 \end{aligned}$$

et donc il s'agit de:

$$x_1 = 30 \quad (57.384)$$

Nous parlons alors parfois du "**pas du pivot**". Nous avons alors la solution actuelle:

$$x_1 = 30, x_2 = 0, x_3 = 70, x_4 = 0 \quad (57.385)$$

Ce qui donne:

$$Z = 1'200 \cdot 30 = 36'000 \quad (57.386)$$

Graphiquement parlant, voilà à quoi correspond ce que nous venons de faire:

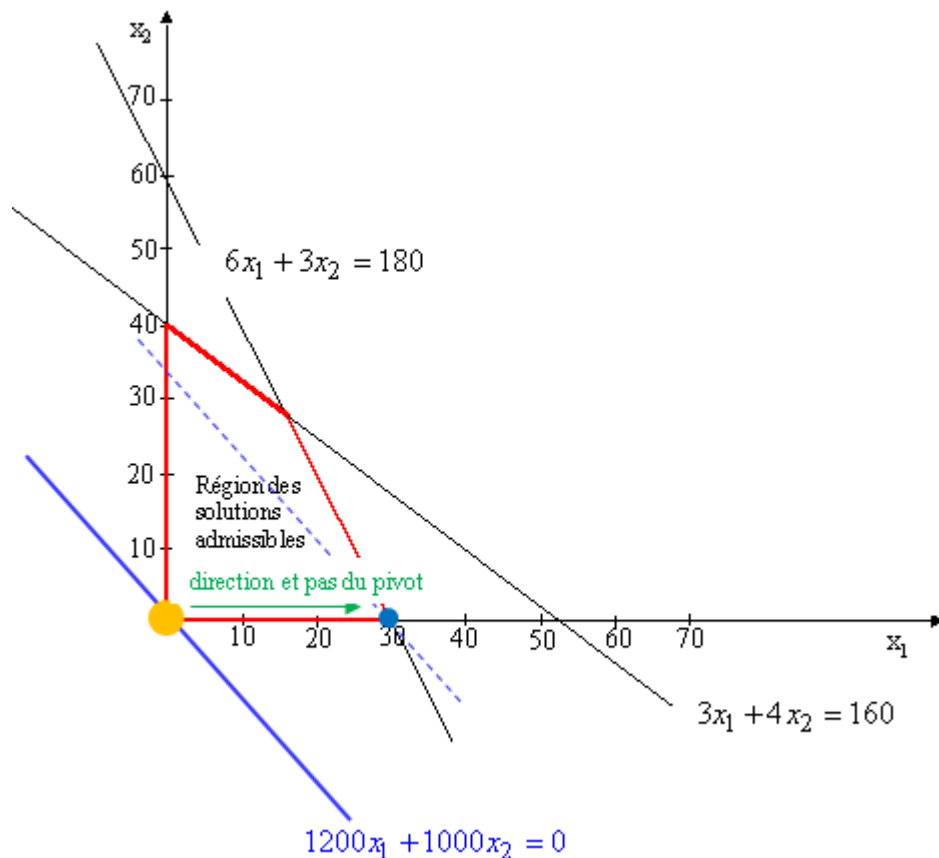


Figure: 57.29 - Direction du pivot avec arrivée au point (30, 0)

Pour continuer à faire croître aussi simplement Z (en ne faisant croître qu'une variable), il nous faut un nouveau système d'équations similaire au système initial:

$$\begin{aligned} x_3 &= 160 - (3x_1 + 4x_2) \\ x_4 &= 180 - (6x_1 + 3x_2) \quad (57.387) \\ Z &= 1'200x_1 + 1'000x_2 \end{aligned}$$

où nous avons exprimé les variables qui prennent une valeur non nulle en fonction des autres qui prennent une valeur nulle, c'est-à-dire x_3, x_4 en fonction de x_1, x_2 puisque nous avons pour rappel:

$$x_1 = 0, x_2 = 0, x_3 = 160, x_4 = 180 \quad (57.388)$$

Pour la suite, il nous faut exprimer x_1, x_3 ainsi que Z en fonction de x_2, x_4 puisque nous venons d'obtenir pour rappel:

$$x_1 = 30, x_2 = 0, x_3 = 70, x_4 = 0 \quad (57.389)$$

Avant d'obtenir le nouveau système fonction de x_1, x_3 , faisons quelques manipulations algébriques:

$$\left. \begin{aligned} x_3 &= 160 - (3x_1 + 4x_2) \\ x_4 &= 180 - (6x_1 + 3x_2) \end{aligned} \right\} \Rightarrow 2x_3 - x_4 = (320 - (6x_1 + 8x_2)) - (180 - (6x_1 + 3x_2)) \quad (57.390)$$

ce qui nous donne après simplification:

$$2x_3 - x_4 = 140 - 5x_2 \Rightarrow x_3 = 70 - 2.5x_2 + 0.5x_4 \quad (57.391)$$

et dès lors, il vient:

$$\left. \begin{array}{l} x_3 = 160 - (3x_1 + 4x_2) \\ x_3 = 70 - 2.5x_2 + 0.5x_4 \end{array} \right\} \Rightarrow x_3 - x_3 = (160 - (3x_1 + 4x_2)) - (70 - 2.5x_2 + 0.5x_4) \quad (57.392)$$

ce qui nous donne après simplification:

$$x_1 = 30 - 0.5x_2 - \frac{1}{6}x_4 \quad (57.393)$$

et nous avons de même:

$$Z = 1'200x_1 + 1'000x_2 = 1'200\left(30 - 0.5x_2 - \frac{1}{6}x_4\right) + 1'000x_2 = 36'000 + 400x_2 - 200x_4 \quad (57.394)$$

Donc au final, le système est:

$$\begin{aligned} x_1 &= 30 - 0.5x_2 - \frac{1}{6}x_4 \\ x_3 &= 70 - 2.5x_2 + 0.5x_4 \\ Z &= 36'000 + 400x_2 - 200x_4 \end{aligned} \quad (57.395)$$

à partir duquel nous itérons le processus (nous ne faisons croître qu'une variable dans Z en gardant les autres à 0). Quand nous ne pouvons plus faire augmenter Z car tous les coefficients sont négatifs, et bien c'est que nous sommes au maximum (merci la convexité). Voyons cela...

Dans Z le plus gros coefficient est maintenant x_2 et donc cela nous amène à poser $x_4 = 0$. La valeur la plus contraignante de x_2 qui permet de respecter toujours les contraintes $x_1, x_2, x_3, x_4 \geq 0$ est dès lors:

$$x_2 = 28 \quad (57.396)$$

Et pour cette valeur, nous avons:

$$x_1 = 30 - 0.5x_2 = 30 - 0.5 \cdot 28 = 16 \quad (57.397)$$

La fonction économique initiale prend alors la valeur:

$$1'200 \cdot x_1 + 1'000 \cdot x_2 = Z \rightarrow 1'200 \cdot 16 + 1'000 \cdot 28 = 47'200 \quad (57.398)$$

ce qui correspond donc graphiquement à:

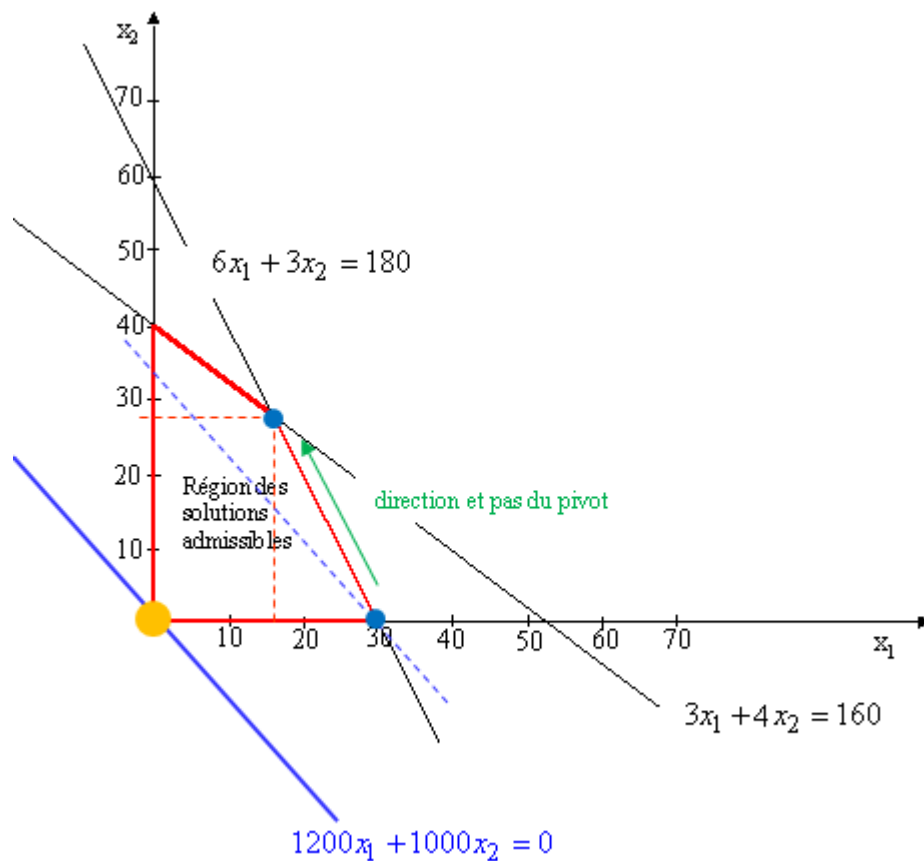


Figure: 57.30 - Direction du pivot avec arrivée au point (16, 28) pour la deuxième itération

Donc nous voyons bien ci-dessus que nous sommes arrivés à la valeur optimale visible sur le graphique donné au début de l'exercice. Mais comment savoir que nous sommes arrivés au point final si nous n'avons pas de graphique ou que nous travaillons dans des dimensions supérieures?

Au fait, le processus est terminé soit quand tous les coefficients de la fonction économique sont négatifs ou que la valeur la plus contraignante qui respecte les contraintes est nulle!!! Voyons si c'est bien le cas! Nous avons donc dans le cas présent:

$$x_1 = 16, x_2 = 28, x_3 = 0, x_4 = 0 \quad (57.399)$$

Et nous allons donc réécrire le système:

$$\begin{aligned} x_1 &= 30 - 0.5x_2 - \frac{1}{6}x_4 \\ x_3 &= 70 - 2.5x_2 + 0.5x_4 \\ Z &= 36'000 + 400x_2 - 200x_4 \end{aligned} \quad (57.400)$$

avec cette fois-ci x_1, x_2 ainsi que Z en fonction de x_3, x_4 . Nous avons alors d'abord:

$$x_1 = 30 - 0.5 \left(\frac{70 - x_3 + 0.5x_4}{2.5} \right) - \frac{1}{6}x_4 = 16 + \frac{1}{5}x_3 - \frac{4}{15}x_4 \quad (57.401)$$

et nous avons:

$$\begin{aligned}
 x_1 = 30 - 0.5x_2 - \frac{1}{6}x_4 &\Rightarrow x_2 = 60 - 2x_1 - \frac{1}{3}x_4 = 60 - 2\left(16 + \frac{1}{5}x_3 - \frac{4}{15}x_4\right) - \frac{1}{3}x_4 \\
 &= 28 - \frac{2}{5}x_3 + \frac{1}{5}x_4
 \end{aligned}
 \tag{57.402}$$

Et donc pour Z il vient (les coefficients y sont tous négatifs donc nous devinons la suite...):

$$Z = 36000 + 400\left(28 - \frac{2}{5}x_3 + \frac{1}{5}x_4\right) - 200x_4 = 47200 - 160x_3 - 120x_4 \tag{57.403}$$

Nous avons donc le nouveau système:

$$\begin{aligned}
 x_1 &= 16 + \frac{1}{5}x_3 - \frac{4}{15}x_4 \\
 x_2 &= 28 - \frac{2}{5}x_3 + \frac{1}{5}x_4 \\
 Z &= 47200 - 160x_3 - 120x_4
 \end{aligned}
 \tag{57.404}$$

Comme tous les coefficients de Z sont négatifs, nous sommes bloqués car nous partirions dans la mauvaise direction. Il faut donc nous arrêter ici et nous adoptons au final la solution:

$$\begin{aligned}
 x_1 &= 16 \\
 x_2 &= 28
 \end{aligned}
 \tag{57.405}$$

La méthode des tableaux que l'on trouve souvent dans la littérature consiste à ne noter que les coefficients des variables du système dans un tableau, mais les transformations que l'on fait sont exactement celles que l'on vient de faire algébriquement (à part qu'elle masque le sens de la méthode).

13.1. ALGORITHME DU SIMPLEXE

Pour mettre en oeuvre cet algorithme, nous devons poser le problème sous une forme "standard" et introduire la notion de "programme de base" qui est l'expression algébrique correspondant à la notion de "point extrême du polyèdre des programmes admissibles" étudiée lors de la programmation linéaire (notée ci-après P.L.). En effet, nous verrons que la solution d'un problème du type P.L. si elle existe, peut toujours être obtenue en un programme de base. La méthode du simplexe va donc consister à trouver un premier programme de base puis à construire une suite de programmes de base améliorant constamment la fonction économique et donc conduisant à l'optimum.

Un problème de P.L. est donc mis sous sa "forme standard" s'il implique la recherche du minimum de la fonction objectif sous des contraintes ayant la forme d'équations linéaires et de conditions de non négativité des variables, c'est-à-dire s'il se pose sous la forme que nous avons vue lors de notre étude de la programmation linéaire:

$$\begin{aligned}
 \min \sum_{j=1}^n c_j x_j &= Z \\
 \sum_{j=1}^n a_{ij} x_j &= b_i, \quad i = 1, \dots, m \\
 x_j &\geq 0, \quad j = 1, \dots, n
 \end{aligned} \quad (57.406)$$

C'est-à-dire aussi, en utilisant des notations matricielles:

$$\begin{aligned}
 \min \quad & c^T x = z \\
 & Ax = b \\
 & x \geq 0
 \end{aligned} \quad (57.407)$$

où les matrices $c : n \times 1$, $A : m \times n$, $b : m \times 1$ correspondent, respectivement, aux coefficients des niveaux d'activité dans la fonction objectif, aux coefficients techniques des activités et aux seconds membres des contraintes.

Nous allons voir maintenant comment un problème général de P.L. peut toujours être ramené à une forme standard. La notion de "variable d'écart" est essentielle pour effectuer cette "réduction".

Chercher le maximum d'une fonction $f(x)$ revient à chercher le minimum de la fonction de signe opposé $-f(x)$. D'autre part, une contrainte qui se présente comme une inéquation:

$$\sum_{j=1}^n a_{ij} x_j \leq b_i \quad (57.408)$$

peut être remplacée par l'équation:

$$\sum_{j=1}^n a_{ij} x_j + e_i = b_i \quad (57.409)$$

impliquant une variable supplémentaire, e_i , appelée donc "variable d'écart", et soumise à la contrainte de non-négativité, $e_i \geq 0$.

Bien évidemment, dans un cas contraire tel où le système est du type:

$$\sum_{j=1}^n a_{ij} x_j \geq b_i \quad (57.410)$$

Nous écrirons:

$$\sum_{j=1}^n a_{ij} x_j - e_i = b_i \quad (57.411)$$

impliquant donc également une variable supplémentaire et soumise à la contrainte de non-négativité, $e_i \geq 0$.

Ce travail de mise en forme standard nous permet donc de retrouver un système d'équations linéaires à résoudre (nous avons vu précédemment sur le site comment résoudre ce genre de système avec l'algorithme du pivot).

La matrice A qui représente les composantes du système d'équations peut s'exprimer dans différentes variantes en fonction de la base vectorielle choisie (voir le chapitre d'Analyse vectorielle dans la section d'algèbre). Nous allons introduire la notion de "forme canonique utilisable" associée au choix d'une base et nous montrerons que cette reformulation du système de contraintes va nous permettre de progresser vers l'optimum.

La matrice A peut, après introduction des variables d'écart se décomposer en deux sous-matrices $[D | B]$, une contenant les variables initiales D et l'autre comportant les variables d'écart B tel que:

$$A = [D | B] \quad (57.412)$$

Remarque: Les variables d'écart sont des variables et non des constantes!! Il convient dans un système où les variables sont au nombre de n et les équations au nombre de m tel qu'une équation du système s'écrirait:

$$a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n \leq b_i \quad (57.413)$$

d'ajouter une variable d'écart tel que:

$$a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n + x_{n+1} \leq b_i \quad (57.414)$$

où $x_{n+1} = e_i$ et sur chaque ligne m , la variable d'écart ajoutée devant être différente de celles déjà insérées dans le système. C'est la raison pour laquelle nous pouvons décomposer la matrice en deux sous-matrices.

Les colonnes de la matrice B sont bien évidemment, par définition de la méthode, des colonnes unités, linéairement indépendantes. Ces colonnes forment une base de l'espace vectoriel des colonnes à m éléments (ou dimensions) - le nombre de lignes du système. Nous appelons B la "**matrice de la base**".

Les variables associées aux composantes colonnes de la matrice B seront dès maintenant appelées les "**variables de bases**". Dans ce cas, les variables de base sont donc essentiellement les variables d'écart $x_{n+1}, x_{n+2}, \dots, x_{n+m}$. Les variables associées aux colonnes de la matrice D seront appelées les "variables hors-base"; il s'agit des variables $x_1, x_2, \dots, x_n$.

Remarque: Rappelons que dans l'expression de la fonction économique, seules les variables hors-base apparaissent.

En résumé, toute P.L. une fois mise sous forme standard est telle que:

- il existe une sous-matrice carrée de la matrice A des coefficients techniques, qui est appelée matrice de base et qui est égale à la matrice carrée unité I de dimension $m \times m$ (effectivement il y a autant de variables d'écart que de lignes dans le système d'équations original - au nombre de m - et autant de colonnes puisque chaque variable d'écart à un indice différent).

- les variables de base associées n'apparaissent pas dans l'expression de la fonction économique.
- le second membre des contraintes est constitué d'éléments non négatifs.

Nous disons que le problème est mis sous "forme canonique utilisable associée à la base B , correspondant aux variables de base $x_{n+1}, x_{n+2}, \dots, x_{n+m}$ ".

Remarque: Nous pouvons intervertir les matrices (et donc changer de base canonique) B et D (ce qui revient à dire que la matrice des variables de base devient la matrice des variables hors-base et inversement).

Il est maintenant commode d'introduire les notations suivantes:

$$x_B = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} \quad x_D = \begin{bmatrix} x_{n+1} \\ x_{n+2} \\ \dots \\ x_{n+m} \end{bmatrix} \quad (57.415)$$

qui sont respectivement le vecteur des variables de base et le vecteur des variables hors-base.

Ainsi, le système d'équations décrivant les contraintes peut s'écrire indifféremment:

$$Ax = b \quad (57.416)$$

ou bien aussi:

$$Bx_B + Dx_D = b \quad (57.417)$$

Si la matrice B est une matrice de base, elle est non singulière et admet donc une matrice inverse B^{-1} . En multipliant cette équation, à gauche et à droite, par B^{-1} nous obtenons:

$$x_B + B^{-1}Dx_D = B^{-1}b \quad (57.418)$$

Le système d'équations aura alors été mis sous une forme résolue en x_B .

Pour obtenir une forme canonique utilisable associée à la base B , correspondant aux variables de base, il ne reste plus qu'à éliminer les variables de base de l'expression de la fonction économique.

Écrivons cette fonction en séparant les vecteurs x_B et x_D , nous obtenons:

$$z = c_B^T x_B + c_D^T x_D \quad (57.419)$$

Nous pouvons alors facilement exprimer x_B en fonction de x_D . En utilisant le système d'équations mis sous forme résolue en x_B , nous avons dans un premier temps:

$$x_B = B^{-1}b - B^{-1}Dx_D \quad (57.420)$$

que nous substituons dans la fonction économique, pour obtenir:

$$z = c_B^T (B^{-1}b - B^{-1}Dx_D) + c_D^T x_D \quad (57.421)$$

Nous regroupons les termes en x_D et nous avons:

$$z = c_B^T B^{-1}b + (c_D^T - c_B^T B^{-1}D)x_D \quad (57.422)$$

Nous avons alors toujours un système d'équations mais ne comportant plus d'inégalités mais au contraire des égalités !!! Reste plus qu'à démontrer que la solution de ce système dit "**programme de base**" par la méthode du pivot est optimale.

Définition: Nous appelons "**coût réduit**" de l'activité hors base j , le coefficient correspondant $c_j - z_j$ de la ligne $c_D^T - c_B^T B^{-1}D$.

Soit un problème de programmation linéaire sous forme standard:

$$\begin{aligned} \min c^T x &= z \\ Ax &= b \quad A: m \times n, n > m \quad (57.423) \\ x &\geq 0 \end{aligned}$$

La matrice A à m lignes (autant qu'il y a de contraintes) et n colonnes, avec $n > m$. Si nous sélectionnons m variables de base et si nous annulons les $n - m$ variables hors base, la matrice A :

$$A = [D \mid B] \quad (57.424)$$

et le système se réduit à:

$$Bx_B = b \quad (57.425)$$

La matrice B est de dimension $m \times m$. Si elle définit une base, elle admet une matrice inverse B^{-1} . Une solution du système est donc:

$$x_D = 0, x_B = B^{-1}b \quad (57.426)$$

Si l'expression $B^{-1}b$ est non négative (≥ 0), nous avons une solution admissible qui vérifie les contraintes et que nous appellerons un "**programme de base**":

$$x = (x_B, x_D) = (B^{-1}b, 0) \geq 0 \quad (57.427)$$

Le problème de programmation linéaire, s'écrit aussi sous la forme suivante, que nous appelons "**forme canonique utilisable associée au programme de base**":

$$\begin{aligned} x_B &= B^{-1}b - B^{-1}Dx_D \\ z &= c_B^T B^{-1}b + (c_D^T - c_B^T B^{-1}D)x_D \quad (57.428) \\ x_B &\geq 0, x_D \geq 0 \end{aligned}$$

À partir des développements effectués précédemment nous pouvons énoncer le résultat suivant:

Proposition 1: Si dans la forme canonique utilisable associée à un programme de base, tous les coûts réduits sont ≥ 0 alors le programme de base est optimal.

Proposition 2: La solution d'un problème de P.L., si elle existe, peut toujours être trouvée en un programme de base.

La démonstration précise de ce résultat est assez délicate. Nous pouvons cependant en avoir une intuition en considérant, une fois de plus, la notion de coût réduit.

En effet, pour un programme de base donné, considérons la forme canonique utilisable associée à la base. Sur cette forme nous pouvons vérifier que, ou bien le programme de base est optimal (tous les coûts réduits sont ≥ 0), ou bien que le programme de base peut être amélioré et remplacé par un nouveau programme de base donnant à z une valeur plus petite (un coût réduit est négatif et la variable hors-base associée peut être augmentée jusqu'à ce qu'une ancienne variable de base s'annule). Comme il y a un nombre fini de programmes de base (au plus égal au nombre C_m^x), la solution de P.L. se trouve nécessairement en un programme de base.

14. PROGRAMMATION (OPTIMISATION) NON LINÉAIRE

Un programme d'optimisation non linéaire (N.L.O.) est une généralisation de la programmation linéaire (algorithme du simplexe) mais à des fonctions non linéaires et pouvant comporter aussi des contraintes et des fonctions économiques non linéaires.

Le but de ce qui va suivre est donc de comprendre dans les grandes lignes mais avec un niveau de rigueur acceptable les outils d'optimisation que proposent de nombreux tableurs comme les versions antérieures à MS Excel 2007:

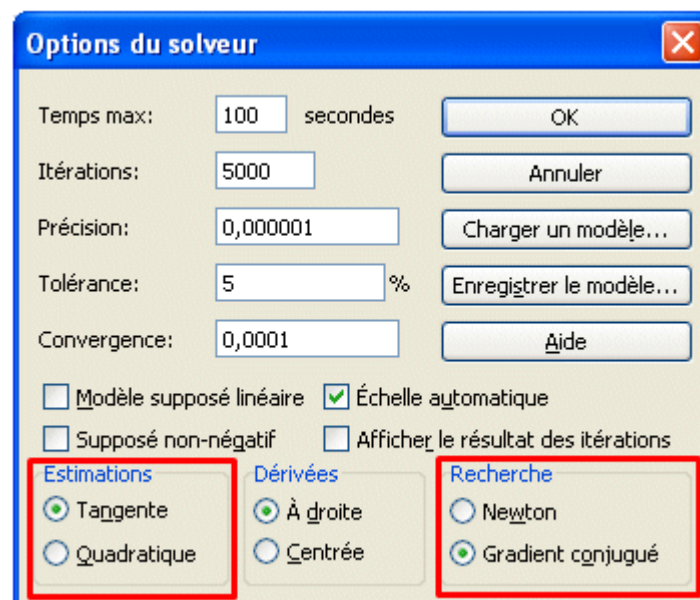


Figure: 57.31 - Exemple d'optimisation non linéaire du solveur de MS Excel

Nous allons en particulier voir maintenant en quoi consiste la recherche Newton (sous-entendu: Gauss-Newton) avec les estimations *Tangente* et *Quadratique*. Après quoi nous étudierons la méthode de recherche des *Gradients conjugués* aussi avec les méthodes tangente et respectivement quadratique.

Nous le verrons plus loin, mais nous le devinons déjà, que le choix *Tangente* utilise une approximation linéaire de la tangente à la fonction à optimiser au point considéré alors que l'option *Quadratique* fera une estimation d'une fonction du deuxième degré au point considéré (typiquement une parabole). Si au point considéré, la fonction se modélise bien par une quadrique, alors l'option Quadratique peut faire économiser du temps en choisissant un meilleur point initial qui demandera moins de pas supplémentaires à chaque recherche. Si vous n'avez pas d'idée du comportement a priori de la fonction, la méthode *Tangente* est alors plus lente mais plus sûre.

Un exemple connu dans la littérature pour introduire la recherche d'optimum de fonctions non linéaires, avant de passer à la partie intégrant les contraintes du système, est la fonction "baleine à bosse" qui consiste à trouver le minimum de:

$$x_1^2 \left(4 - 2.1x_1^2 + \frac{1}{3}x_1^4 \right) + x_1 x_2 + x_2^2 \left(-4 + 4x_2^2 \right) \quad (57.429)$$

avec les contraintes:

$$\begin{aligned} -2 &\leq x_1 \leq 2 \\ -1 &\leq x_2 \leq 1 \end{aligned} \quad (57.430)$$

ce que nous pouvons en effet vérifier visuellement.

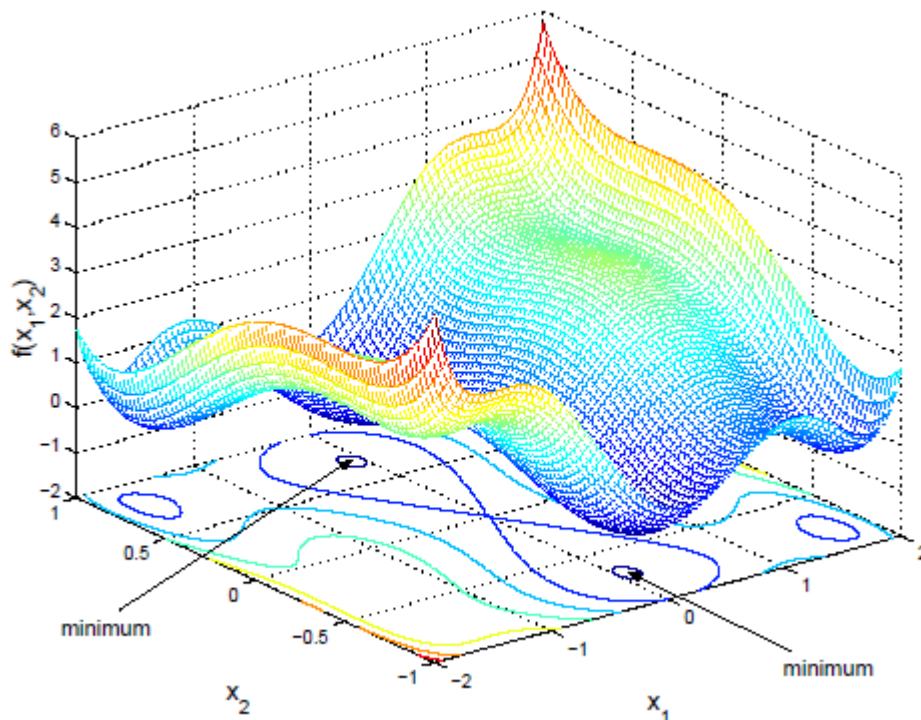


Figure: 57.32 - Tracé de la fonction baleine à bosse avec minimum déjà visibles

Comme nous pouvons le voir, cette fonction est également un excellent exemple de minimaux locaux multiples.

14.1. MÉTHODE DE NEWTON-RAPHSON (NEWTON QUADRATIQUE)

La "**méthode de Newton-Raphson**" est une technique permettant de chercher l'extremum d'une fonction ou également, comme nous le verrons plus loin lorsque nous comparerons dans un cas particulier d'application la méthode de Gauss-Newton à celle de Newton-Raphson, pour la régression non linéaire.

La méthode de Newton-Raphson, qui dans les versions antérieures à MS Excel 2007 s'activait dans le solveur en sélectionnant l'option *Newton* et *Quadratique*, utilise les approximations de Taylor au deuxième ordre (donc avec les dérivées du deuxième ordre) pour avoir une fonction quadratique (parabole) qui converge si le point d'origine de la recherche est proche de l'optimum. Cette approximation est réitérée à chaque itération.

Pour commencer, rappelons que nous avons démontré dans le chapitre de Suites Et Séries qu'un développement de Taylor pour une fonction à deux variables pouvait s'écrire en approximation quadratique:

$$f(x_0 + h, y_0 + k) = f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)h + \frac{\partial f}{\partial y}(x_0, y_0)k + \frac{1}{2} \begin{pmatrix} h & k \end{pmatrix} H \begin{pmatrix} h \\ k \end{pmatrix} \quad (57.431)$$

où pour rappel h et k sont des variables et x_0, y_0 sont fixés et où nous avons la matrice hessienne:

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} \quad (57.432)$$

que les américains spécialistes du domaine ont pour habitude (malheureuse à mon avis...) de noter:

$$H = \left[\frac{\partial^2 f}{\partial x_i \partial x_j} \right]_{ij} = \left[\tilde{\nabla}^2 f \right]_{ij} = \tilde{\nabla}^2 f \quad (57.433)$$

la dernière expression étant la plus courante peut être très trompeuse avec la notation du laplacien.

Dans le domaine des méthodes numériques il est d'usage d'écrire la série de Taylor ci-dessus avec quelques changements de notation en posant d'abord:

$$\begin{aligned} x_0 + h &= x \Rightarrow h = x - x_0 \\ y_0 + k &= y \Rightarrow k = y - y_0 \\ \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} + \begin{pmatrix} h \\ k \end{pmatrix} &= \begin{pmatrix} x_0 + h \\ y_0 + k \end{pmatrix} = \vec{x}_0 + \begin{pmatrix} h \\ k \end{pmatrix} = \vec{x} \Rightarrow \begin{pmatrix} h \\ k \end{pmatrix} = \vec{x} - \vec{x}_0 \\ \tilde{\nabla} &= \begin{pmatrix} \partial_x \\ \partial_y \end{pmatrix} \end{aligned} \quad (57.434)$$

Ce qui nous donne une forme plus condensée et technique de la série de Taylor autour de $\vec{x}$:

$$f(\vec{x}) \cong f(\vec{x}_0) + \vec{\nabla}(f(\vec{x}_0))(\vec{x} - \vec{x}_0) + \frac{1}{2}(\vec{x} - \vec{x}_0)^T H(\vec{x}_0)(\vec{x} - \vec{x}_0) \quad (57.435)$$

En changeant un peu la notation:

$$f(\vec{x}) \cong f(\vec{x}_i) + \vec{\nabla}(f(\vec{x}_i))(\vec{x} - \vec{x}_i) + \frac{1}{2}(\vec{x} - \vec{x}_i)^T H(\vec{x}_i)(\vec{x} - \vec{x}_i) \quad (57.436)$$

Nous retrouvons donc l'expression d'usage d'une fonction de $\mathbb{R}^2 \rightarrow \mathbb{R}$ évaluée en série de Taylor centrée en $\vec{x}$.

Mais si nous cherchons un extrema local (appelé aussi parfois "point critique"), il faudra dans un premier temps que la dérivée de l'ensemble de la série de Taylor soit nulle. C'est-à-dire:

$$\vec{\nabla}(f(\vec{x})) = \begin{pmatrix} \partial_x f(\vec{x}) \\ \partial_y f(\vec{x}) \end{pmatrix} = \vec{\nabla} \left(f(\vec{x}_i) + \nabla f(\vec{x}_i)(\vec{x} - \vec{x}_i) + \frac{1}{2}(\vec{x} - \vec{x}_i)^T H(\vec{x}_i)(\vec{x} - \vec{x}_i) \right) = \vec{0} \quad (57.437)$$

et que le déterminant de la matrice hessienne soit positif (cf. chapitre Suites Et Séries). Et pour savoir si nous sommes sur un maximum local ou minimum local, il nous faudra regarder le signe de $\partial_x^2 f(\vec{x})$.

Récrivons la relation ci-dessus explicitement comme nous l'avions démontrée dans le chapitre de Suites Et Séries pour des raisons pédagogiques:

$$\begin{aligned} & f(x_0 + h, y_0 + k) \\ &= f(x_0, y_0) + \frac{1}{1!} \frac{\partial f}{\partial x}(x_0, y_0)h + \frac{1}{1!} \frac{\partial f}{\partial y}(x_0, y_0)k \\ &+ \frac{1}{2!} \left[\frac{\partial^2 f}{\partial x^2}(x_0, y_0)h^2 + 2 \frac{\partial^2 f}{\partial x \partial y}(x_0, y_0)hk + \frac{\partial^2 f}{\partial y^2}(x_0, y_0)k^2 \right] \end{aligned} \quad (57.438)$$

Et rappelons que tous les termes x_0, y_0 sont des constantes car il s'agit soit de la fonction évaluée sur le point particulier x_0, y_0 , soit de la dérivée partielle évaluée en ce même point, soit de la dérivée partielle seconde toujours évaluée en ce même point, etc.

Donc le gradient donnera finalement:

$$\begin{aligned}
\vec{\nabla}(f(x_0+h, y_0+h)) &= \begin{pmatrix} \partial_h f(x_0+h, y_0+h) \\ \partial_k f(x_0+h, y_0+h) \end{pmatrix} \\
&= \begin{pmatrix} \underbrace{\partial_h f(x_0, y_0)}_{=0} + \underbrace{\partial_h(\partial_x f(x_0, y_0)h)}_{\partial_x f(x_0, y_0)h} + \underbrace{\partial_h(\partial_y f(x_0, y_0)k)}_{=0} + \frac{1}{2} \underbrace{\partial_h(\partial_x^2 f(x_0, y_0)h^2)}_{\partial_x^2 f(x_0, y_0)h} \\ + \frac{1}{2} \underbrace{\partial_h(\partial_{xy}^2 f(x_0, y_0)hk)}_{\frac{1}{2}\partial_{xy}^2 f(x_0, y_0)k} + \frac{1}{2} \underbrace{\partial_h(\partial_y^2 f(x_0, y_0)k^2)}_{=0} \\ \underbrace{\partial_k f(x_0, y_0)}_{=0} + \underbrace{\partial_k(\partial_x f(x_0, y_0)h)}_{=0} + \underbrace{\partial_k(\partial_y f(x_0, y_0)k)}_{=\partial_y f(x_0, y_0)} + \frac{1}{2} \underbrace{\partial_k(\partial_x^2 f(x_0, y_0)h^2)}_{=0} \\ + \frac{1}{2} \underbrace{\partial_k(\partial_{xy}^2 f(x_0, y_0)hk)}_{\frac{1}{2}\partial_{xy}^2 f(x_0, y_0)h} + \frac{1}{2} \underbrace{\partial_k(\partial_y^2 f(x_0, y_0)k^2)}_{=\partial_y^2 f(x_0, y_0)k} \end{pmatrix} \quad (57.439) \\
&= \begin{pmatrix} \partial_x f(x_0, y_0) + \partial_x^2 f(x_0, y_0)h + \frac{1}{2}\partial_{xy}^2 f(x_0, y_0)k \\ \partial_y f(x_0, y_0) + \partial_y^2 f(x_0, y_0)k + \frac{1}{2}\partial_{xy}^2 f(x_0, y_0)h \end{pmatrix} = \vec{\nabla}(f(x_0, y_0)) + H \begin{pmatrix} h \\ k \end{pmatrix}
\end{aligned}$$

et en revenant aux notations d'usage du domaine des méthodes numériques, nous avons alors:

$$\vec{\nabla}(f(\vec{x})) = \vec{\nabla}(f(\vec{x}_i)) + H(\vec{x}_i)(\vec{x} - \vec{x}_i) \quad (57.440)$$

Et donc le gradient devant être nul, nous avons:

$$\vec{\nabla}(f(\vec{x}_i)) + H(\vec{x}_i)(\vec{x} - \vec{x}_i) = 0 \quad (57.441)$$

et après un premier réarrangement:

$$H^{-1}(\vec{x}_i) \vec{\nabla}(f(\vec{x}_i)) + \vec{x} - \vec{x}_i = 0 \quad (57.442)$$

et un deuxième réarrangement:

$$\boxed{\vec{x} = \vec{x}_i - H^{-1}(\vec{x}_i) \vec{\nabla}(f(\vec{x}_i))} \quad (57.443)$$

ce qui se note souvent:

$$\vec{x}_{i+1} = \vec{x}_i - H^{-1}(\vec{x}_i) \vec{\nabla}(f(\vec{x}_i)) \quad (57.444)$$

et par les américains...:

$$\bar{x}_{i+1} = \bar{x}_i - \left[\bar{\nabla}^2 f(\bar{x}_i) \right]^{-1} \bar{\nabla} f(\bar{x}_i) \quad (57.445)$$

Enfin avant de passer à un exemple concret il est important que le lecteur se souvienne de la relation vu juste plus haut:

$$\bar{\nabla} f(\bar{x}_i) + H(\bar{x}_i)(\bar{x} - \bar{x}_i) = 0 \Rightarrow -\bar{\nabla} f(\bar{x}_i) = H(\bar{x}_i)(\bar{x} - \bar{x}_i) \quad (57.446)$$

Le fait que la gradient négatif apparaisse fait que la technique de Newton-Raphson (Newton Quadratique) appartient à la famille des techniques dites de "**descente rapide**".

Cette dernière égalité étant souvent noté chez les américains (... sans commentaires...) par:

$$-\nabla f(\bar{x}_i) = \nabla^2 f(\bar{x}_i) p \quad (57.447)$$

Évidemment un tableur comme MS Excel ne pouvant pas déterminer les dérivées il va les calculer en utilisant la méthode numérique des dérivées à droite ou centrées comme nous l'avons présenté un peu plus haut.

Exemple:

Nous cherchons un extrema local de la fonction "baleine à bosse" représentée plus haut:

$$f(x, y) = x^2 \left(4 - 2.1x^2 + \frac{1}{3}x^4 \right) + xy + y^2 (-4 + 4y^2) \quad (57.448)$$

avec le point de départ (arbitraire):

$$\bar{x}_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad (57.449)$$

Pour effectuer la recherche, nous calculons d'abord le gradient:

$$\bar{\nabla} f(\bar{x}_i) = \begin{pmatrix} \partial_x f(\bar{x}_i) \\ \partial_y f(\bar{x}_i) \end{pmatrix} = \begin{pmatrix} 8x_i - 8.4x_i^3 + 2x_i^5 + y_i \\ x_i - 8y_i + 16y_i^3 \end{pmatrix} \quad (57.450)$$

et la matrice hessienne:

$$H(\bar{x}_i) = \begin{bmatrix} \partial_x^2 f(x_i, y_i) & \partial_{xy}^2 f(x_i, y_i) \\ \partial_{xy}^2 f(x_i, y_i) & \partial_y^2 f(x_i, y_i) \end{bmatrix} = \begin{bmatrix} 8 - 25.2x_i^2 + 10x_i^4 & 0 \\ 0 & -8 + 48y_i^2 \end{bmatrix} \quad (57.451)$$

Nous avons alors:

$$\begin{aligned}
\vec{\nabla}(f(\vec{x}_0)) &= \begin{pmatrix} \partial_x f(x_0) \\ \partial_y f(y_0) \end{pmatrix} = \begin{pmatrix} \partial_x f(1) \\ \partial_y f(1) \end{pmatrix} = \begin{pmatrix} 9x_0 - 8.4x_0^3 + 2x_0^5 + y_0 \\ x_0 - 8y_0 + 16y_0^3 \end{pmatrix} = \begin{pmatrix} 2.60 \\ 9.00 \end{pmatrix} \\
H(\vec{x}_0) &= \begin{bmatrix} \partial_{xx}^2 f(x_0, y_0) & \partial_{xy}^2 f(x_0, y_0) \\ \partial_{xy}^2 f(x_0, y_0) & \partial_{yy}^2 f(x_0, y_0) \end{bmatrix} = \begin{bmatrix} 8 - 25.2x_0^2 + 10x_0^4 & 0 \\ 0 & -8 + 48y_0^2 \end{bmatrix} \quad (57.452) \\
&= \begin{bmatrix} -7.20 & 0 \\ 0 & 40.00 \end{bmatrix} \\
H^{-1}(\vec{x}_0) &= \begin{bmatrix} -7.20 & 0 \\ 0 & 40.00 \end{bmatrix}^{-1} = \begin{bmatrix} -0.138 & 0 \\ 0 & 0.025 \end{bmatrix}
\end{aligned}$$

et donc:

$$\vec{x}_1 = \vec{x}_0 - H^{-1}(\vec{x}_0) \vec{\nabla}(f(\vec{x}_0)) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{bmatrix} -0.138 & 0 \\ 0 & 0.025 \end{bmatrix} \begin{pmatrix} 2.60 \\ 9.00 \end{pmatrix} = \begin{pmatrix} 1.361 \\ 0.775 \end{pmatrix} \quad (57.453)$$

et nous recommençons (bon on va se passer de tous les détails maintenant car faut pas exagérer non plus...):

$$\begin{aligned}
\vec{\nabla}(f(\vec{x}_1)) &= \begin{pmatrix} \partial_x f(1.361) \\ \partial_y f(0.775) \end{pmatrix} = \begin{pmatrix} -0.17 \\ 2.61 \end{pmatrix} \\
H(\vec{x}_1) &= \begin{bmatrix} 8 - 25.2x_0^2 + 10x_0^4 & 0 \\ 0 & -8 + 48y_0^2 \end{bmatrix} = \begin{bmatrix} -4.36 & 0 \\ 0 & 20.83 \end{bmatrix} \quad (57.454) \\
H^{-1}(\vec{x}_1) &= \begin{bmatrix} -4.36 & 0 \\ 0 & 20.83 \end{bmatrix}^{-1} = \begin{bmatrix} -0.229 & 0 \\ 0 & 0.048 \end{bmatrix}
\end{aligned}$$

et donc:

$$\vec{x}_2 = \vec{x}_1 - H^{-1}(\vec{x}_1) \vec{\nabla}(f(\vec{x}_1)) = \begin{pmatrix} 1.361 \\ 0.775 \end{pmatrix} - \begin{bmatrix} -0.229 & 0 \\ 0 & 0.048 \end{bmatrix} \begin{pmatrix} -0.17 \\ 2.61 \end{pmatrix} = \begin{pmatrix} 1.321 \\ -0.650 \end{pmatrix} \quad (57.455)$$

et nous recommençons (avec encore moins de détails):

$$\begin{aligned}
\vec{\nabla}(f(\vec{x}_2)) &= \begin{pmatrix} -0.01 \\ 0.51 \end{pmatrix} \\
H(\vec{x}_2) &= \begin{bmatrix} -5.52 & 0 \\ 0 & 12.26 \end{bmatrix} \quad (57.456) \\
H^{-1}(\vec{x}_2) &= \begin{bmatrix} -0.181 & 0 \\ 0 & 0.081 \end{bmatrix}
\end{aligned}$$

et donc:

$$\bar{x}_3 = \bar{x}_2 - H^{-1}(\bar{x}_2) \bar{\nabla}(f(\bar{x}_2)) = \begin{pmatrix} 1.303 \\ 0.608 \end{pmatrix} \quad (57.457)$$

et encore une fois (avec encore moins de détails):

$$\begin{aligned} \bar{\nabla}(f(\bar{x}_3)) &= \begin{pmatrix} -0.04 \\ 0.03 \end{pmatrix} \\ H^{-1}(\bar{x}_3) &= \begin{bmatrix} -0.167 & 0 \\ 0 & 0.102 \end{bmatrix} \end{aligned} \quad (57.458)$$

et donc:

$$\bar{x}_4 = \bar{x}_3 - H^{-1}(\bar{x}_3) \bar{\nabla}(f(\bar{x}_3)) = \begin{pmatrix} 1.296 \\ 0.604 \end{pmatrix} \quad (57.459)$$

et encore une fois (avec encore moins de détails):

$$\bar{x}_5 = \bar{x}_4 - H^{-1}(\bar{x}_4) \bar{\nabla}(f(\bar{x}_4)) = \begin{pmatrix} 1.296 \\ 0.605 \end{pmatrix} \quad (57.460)$$

et les valeurs ne bougeront plus. Mais si nous regardons le graphique d'origine où nous avons mis en évidence le point de convergence par un point rouge:

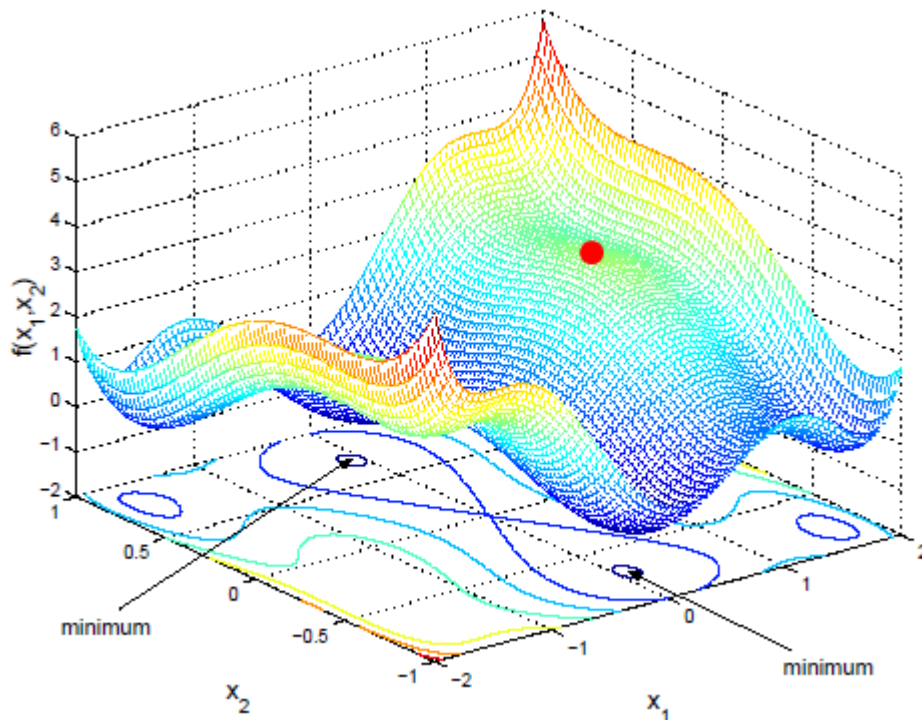


Figure: 57.33 - Mise en évidence du point de convergence dans la fonction baleine à bosse

nous constatons que ce système ne cherche pas un extremum global mais local comme nous l'avions déjà spécifié. En réalité, comme le lecteur pourra le tester lui-même, la convergence est très sensible au point de départ initial.

14.2. MÉTHODE DE GAUSS-NEWTON (NEWTON TANGENTE)

La méthode de Gauss-Newton est une approximation puissante sans les dérivées du deuxième ordre de la méthode de Newton-Raphson qui dans les versions antérieures à MS Excel 2007 s'activait dans le solveur en sélectionnant l'option Newton et Tangente,

Pour aborder ce sujet, partons tout de suite accompagnés d'un exemple concret. Supposons que nous avons obtenu les données suivantes:

t_i	y_i
1	3.2939
2	4.2699
4	7.1749
5	9.3008
8	20.259

Tableau: 57.9 - Données mesurées

et nous supposons que les données suivent le modèle théorique suivant (nous aurions pu faire n'importe quel autre choix!):

$$y_i \cong x_1 e^{x_2 t_i} \quad (57.461)$$

Nous cherchons donc x_1, x_2 qui minimisent la somme des carrés des écarts entre les valeurs expérimentales et théoriques tels que:

$$\min_{x_1, x_2} f(x_1, x_2) = \sum_{i=1}^5 \left(x_1 e^{x_2 t_i} - y_i \right)^2 \quad (57.462)$$

avec donc:

$$f_i(x_1, x_2) = f_i(\vec{x}) = x_1 e^{x_2 t_i} - y_i \quad (57.463)$$

Notons pour la suite comme le veut la tradition dans le domaine:

$$F(\vec{x}) = \begin{pmatrix} f_1 \\ f_2 \\ \dots \\ f_n \end{pmatrix} \quad (57.464)$$

Nous avons alors la notation courante:

$$\min_{x_1, x_2} f(x_1, x_2) = \sum_i f_i^2 \Rightarrow \min_{\vec{x}} f(\vec{x}) = F(\vec{x})^T F(\vec{x}) \quad (57.465)$$

Maintenant, imaginons que nous ayons trouvé une valeur du couple $(x_1, x_2) = (\vec{x})$ qui donne ce minimum et notons le $(\vec{x}_*)$ et en n'oubliant pas évidemment que cela ne sera qu'une solution locale! Considérons un cas particulier que nous appelons "**solution compatible**" et définie par le fait que le couple qui minimise la somme des carrées des erreurs est aussi tel que pour tout i nous ayons:

$$f_i(\vec{x}_*) = 0 \quad (57.466)$$

Dès lors, il en découle immédiatement que:

$$F(\vec{x}_*) = 0 \quad (57.467)$$

Avant d'aller plus loin, remarquons que par exemple que pour une composante j (ce qui correspond dans notre cas à chaque variable de la fonction théorique supposée a priori):

$$\begin{aligned} \vec{\nabla} \left(F(\vec{x})^T F(\vec{x}) \right)_j &= \frac{\partial}{\partial x_j} \sum_i f_i^2 = \sum_i \frac{\partial f_i^2}{\partial x_j} = \sum_i 2f_i \frac{\partial f_i}{\partial x_j} = 2 \sum_i f_i \frac{\partial f_i}{\partial x_j} \\ &= 2 \sum_i \frac{\partial f_i}{\partial x_j} f_i = 2 \left(\vec{\nabla} \left(F(\vec{x})^T \right) F(\vec{x}) \right)_j \end{aligned} \quad (57.468)$$

où la dernière égalité condensée est souvent loi d'être triviale d'autant plus qu'elle fait usage du gradient d'un champ de vecteur (cf. chapitre de Calcul Vectoriel) que l'on retrouve rarement dans la pratique. Le lecteur déstabilisé pourra se reporter si besoin directement à l'exemple plus bas afin d'éclairer sa lanterne.

Nous en déduisons donc que:

$$\vec{\nabla} \left(\sum_i f_i^2 \right) = \vec{\nabla} (f(\vec{x})) = 2 \vec{\nabla} \left(F(\vec{x})^T \right) F(\vec{x}) \quad (57.469)$$

et la solution compatible nous amène donc à bien évidemment à:

$$\vec{\nabla} (f(\vec{x}_*)) = 2 \vec{\nabla} \left(F(\vec{x}_*)^T \right) F(\vec{x}_*) = 0 \quad (57.470)$$

De la même manière, nous avons:

$$\begin{aligned}
 \tilde{\nabla}^2 \left(F(\bar{x})^T F(\bar{x}) \right)_j &= \tilde{\nabla} \left(2 \sum_i f_i \frac{\partial f_i}{\partial x_j} \right) = 2 \tilde{\nabla} \left(\sum_i f_i \frac{\partial f_i}{\partial x_j} \right) \\
 &= 2 \left(\sum_{i,k} \frac{\partial f_i}{\partial x_k} \frac{\partial f_i}{\partial x_j} + f_i \frac{\partial f_i}{\partial x_k \partial x_j} \right) = 2 \left(\sum_{i,k} \frac{\partial f_i}{\partial x_k} \frac{\partial f_i}{\partial x_j} + \sum_{i,k} f_i \frac{\partial f_i}{\partial x_k \partial x_j} \right) \quad (57.471) \\
 &= 2 \left(\tilde{\nabla} \left(F(\bar{x})^T \right) \left(\tilde{\nabla} F(\bar{x}) \right) + \sum_i f_i \tilde{\nabla}^2 f_i(\bar{x}) \right)_j
 \end{aligned}$$

Nous avons donc au final les deux relations suivantes:

$$\begin{aligned}
 \tilde{\nabla} (f(\bar{x})) &= 2 \tilde{\nabla} \left(F(\bar{x})^T \right) F(\bar{x}) \\
 \tilde{\nabla}^2 \left(F(\bar{x})^T F(\bar{x}) \right) &= 2 \left(\tilde{\nabla} \left(F(\bar{x})^T \right) \tilde{\nabla} (F(\bar{x})) + \sum_i f_i \tilde{\nabla}^2 f_i(\bar{x}) \right) \quad (57.472)
 \end{aligned}$$

Étant donné que pour la solution compatible nous avons:

$$f_i(\bar{x}_*) = 0 \quad (57.473)$$

Il s'ensuit que dans ce cas que la deuxième relation devient:

$$\begin{aligned}
 \tilde{\nabla}^2 \left(F(\bar{x}_*)^T F(\bar{x}_*) \right) &\cong 2 \left(\tilde{\nabla} \left(F(\bar{x}_*)^T \right) \tilde{\nabla} (F(\bar{x}_*)) + \sum_i \underbrace{f_i(\bar{x}_*)}_{=0} \tilde{\nabla}^2 f_i(\bar{x}_*) \right) \quad (57.474) \\
 &= 2 \tilde{\nabla} \left(F(\bar{x}_*)^T \right) \tilde{\nabla} (F(\bar{x}_*)) = 2 \begin{bmatrix} 1 & 1 \end{bmatrix} H \begin{bmatrix} 1 \\ 1 \end{bmatrix}
 \end{aligned}$$

où H est la matrice Hessienne (cf. chapitre de Suites Et Séries) ce que les américains notent simplement:

$$\tilde{\nabla} \left(F(\bar{x}_*)^T \right) \tilde{\nabla} (F(\bar{x}_*)) \cong H(\bar{x}_*) \quad (57.475)$$

Donc nous pouvons approximer dans le cas de la solution compatible, la Hessienne qui contient des dérivées d'ordre deux par des dérivées premières.

Donc nous avons finalement dans ce cas particulier les deux relations qui sont à la base de méthode de Gauss-Newton:

$$\begin{aligned}
 \tilde{\nabla} (f(\bar{x}_*)) &= 2 \tilde{\nabla} \left(F(\bar{x}_*)^T \right) F(\bar{x}_*) \\
 \tilde{\nabla} \left(F(\bar{x}_*)^T \right) \tilde{\nabla} (F(\bar{x}_*)) &\cong H(\bar{x}_*)
 \end{aligned} \quad (57.476)$$

Maintenant, rappelons la relation de base de la méthode de Newton-Raphson obtenu plus haut:

$$-\vec{\nabla}(f(\vec{x}_i)) = H(\vec{x}_i)(\vec{x} - \vec{x}_i) \quad (57.477)$$

et pour information, toute technique mathématique (car elles sont nombreuses!), permettant de simplifier la matrice Hessien à droite de l'égalité fait alors partie de la famille des "méthodes de quasi-Newton".

Eh bien la méthode de Gauss-Newton qui nous intéresse ici et qui est donc une des techniques de la famille des méthodes de quasi-Newton consiste simplement dans un premier temps à se débarrasser des dérivées secondes de la Hessienne de la méthode de Newton-Raphson à droite de l'égalité à l'aide de la relation établie précédemment tel que (attention à se rappeler de l'abus d'écriture!):

$$-\vec{\nabla}(f(\vec{x}_i)) \cong \vec{\nabla}(F(\vec{x}_{i*})^T) \vec{\nabla}(F(\vec{x}_{i*}))(\vec{x} - \vec{x}_{i*}) \quad (57.478)$$

et dans un deuxième temps récrire le gradient à gauche de l'égalité à l'aide aussi de la relation précédemment établie. Ce qui nous donne:

$$-2\vec{\nabla}(F(\vec{x}_*)^T) F(\vec{x}_*) \cong \vec{\nabla}(F(\vec{x}_{i*})^T) \vec{\nabla}(F(\vec{x}_{i*}))(\vec{x} - \vec{x}_{i*}) \quad (57.479)$$

Le facteur 2 n'étant pas très esthétique, la quasi-totalité des ouvrages de référence optimisent le problème avec la relation de départ suivante:

$$\min_{x_1, x_2} f(x_1, x_2) = \frac{1}{2} \sum_{i=1}^5 \left(x_1 e^{x_2 t_i} - y_i \right)^2 \quad (57.480)$$

Donc en multipliant simplement par un facteur 1/2 (ce qui ne change rien au résultat). Nous avons alors:

$$-\vec{\nabla}(F(\vec{x}_*)^T) F(\vec{x}_*) \cong \vec{\nabla}(F(\vec{x}_{i*})^T) \vec{\nabla}(F(\vec{x}_{i*}))(\vec{x} - \vec{x}_{i*}) \quad (57.481)$$

Rappelons encore une fois qu'un tableur comme MS Excel ne pouvant pas déterminer les dérivées il va les calculer en utilisant la méthode numérique des dérivées à droite ou centrées comme nous l'avons présenté un peu plus haut.

Revenons maintenant à notre exemple du début! Nous avons donc:

$$t = \begin{pmatrix} 1 \\ 2 \\ 4 \\ 5 \\ 8 \end{pmatrix}, y = \begin{pmatrix} 3.2939 \\ 4.2699 \\ 7.1749 \\ 9.3008 \\ 20.259 \end{pmatrix} \quad (57.482)$$

Nous commençons avec un couple qui nous semble proche de la solution cherchée:

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_{1,0^*} \\ x_{2,0^*} \end{pmatrix} = \begin{pmatrix} 2.5 \\ 0.25 \end{pmatrix} \quad (57.483)$$

Nous avons alors:

$$F(\vec{x}) = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{pmatrix} = \begin{pmatrix} x_1 e^{x_2 t_1} - y_1 \\ x_1 e^{x_2 t_2} - y_2 \\ x_1 e^{x_2 t_3} - y_3 \\ x_1 e^{x_2 t_4} - y_4 \\ x_1 e^{x_2 t_5} - y_5 \end{pmatrix} = \begin{pmatrix} 2.5 e^{0.25 \cdot 1} - 3.2939 \\ 2.5 e^{0.25 \cdot 2} - 4.2699 \\ 2.5 e^{0.25 \cdot 4} - 7.1749 \\ 2.5 e^{0.25 \cdot 5} - 9.3008 \\ 2.5 e^{0.25 \cdot 8} - 20.259 \end{pmatrix} = \begin{pmatrix} -0.0838 \\ -0.1481 \\ -0.3792 \\ -0.5749 \\ -1.7864 \end{pmatrix} \quad (57.484)$$

et donc nous avons:

$$\min_{x_1, x_2} f(x_1, x_2) = \frac{1}{2} \sum_{i=1}^5 \left(x_1 e^{x_2 t_i} - y_i \right)^2 = 1.847 \cong 2 \quad (57.485)$$

Pour la suite, il vient alors:

$$\bar{\nabla}(f(\vec{x})) = \frac{1}{2} \begin{pmatrix} \sum_{i=1}^5 2(x_1 e^{x_2 t_i} - y_i) e^{x_2 t_i} \\ \sum_{i=1}^5 2(x_1 e^{x_2 t_i} - y_i) x_1 t_i e^{x_2 t_i} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^5 (x_1 e^{x_2 t_i} - y_i) e^{x_2 t_i} \\ \sum_{i=1}^5 (x_1 e^{x_2 t_i} - y_i) x_1 t_i e^{x_2 t_i} \end{pmatrix} \quad (57.486)$$

Ce que nous pouvons donc récrire sous la forme:

$$\begin{aligned} \bar{\nabla}(f(\vec{x})) &= \bar{\nabla}(F(\vec{x})^T) F(\vec{x}) \\ &= \begin{bmatrix} e^{x_2 t_1} & e^{x_2 t_2} & e^{x_2 t_3} & e^{x_2 t_4} & e^{x_2 t_5} \\ x_1 t_1 e^{x_2 t_1} & x_1 t_2 e^{x_2 t_2} & x_1 t_3 e^{x_2 t_3} & x_1 t_4 e^{x_2 t_4} & x_1 t_5 e^{x_2 t_5} \end{bmatrix} \begin{pmatrix} x_1 e^{x_2 t_1} - y_1 \\ x_1 e^{x_2 t_2} - y_2 \\ x_1 e^{x_2 t_3} - y_3 \\ x_1 e^{x_2 t_4} - y_4 \\ x_1 e^{x_2 t_5} - y_5 \end{pmatrix} \\ &= \begin{bmatrix} 1.2840 & 1.6487 & 2.7183 & 3.4903 & 7.3891 \\ 3.2101 & 8.2435 & 27.1828 & 43.6293 & 147.7811 \end{bmatrix} \begin{pmatrix} -0.0838 \\ -0.1481 \\ -0.3792 \\ -0.5749 \\ -1.7864 \end{pmatrix} = \begin{bmatrix} -16.5888 \\ -300.8722 \end{bmatrix} \end{aligned} \quad (57.487)$$

Nous avons aussi in extenso:

$$\begin{aligned}
 & \vec{\nabla} \left(F(\vec{x})^T \right) \vec{\nabla} \left(F(\vec{x}) \right) \\
 &= \begin{bmatrix} e^{x_2 t_1} & e^{x_2 t_2} & e^{x_2 t_3} & e^{x_2 t_4} & e^{x_2 t_5} \\ x_1 t_1 e^{x_2 t_1} & x_1 t_2 e^{x_2 t_2} & x_1 t_3 e^{x_2 t_3} & x_1 t_4 e^{x_2 t_4} & x_1 t_5 e^{x_2 t_5} \end{bmatrix} \begin{bmatrix} e^{x_2 t_1} & x_1 t_1 e^{x_2 t_1} \\ e^{x_2 t_2} & x_1 t_2 e^{x_2 t_2} \\ e^{x_2 t_3} & x_1 t_3 e^{x_2 t_3} \\ e^{x_2 t_4} & x_1 t_4 e^{x_2 t_4} \\ e^{x_2 t_5} & x_1 t_5 e^{x_2 t_5} \end{bmatrix} \quad (57.488) \\
 &= \begin{bmatrix} 78.5367 & 1'335.8479 \\ 1'335.8479 & 24'559.9419 \end{bmatrix}
 \end{aligned}$$

Ensuite, nous appliquons la relation démontrée plus haut:

$$-\vec{\nabla} \left(F(\vec{x}_*)^T \right) F(\vec{x}_*) \cong \vec{\nabla} \left(F(\vec{x}_i^*)^T \right) \vec{\nabla} \left(F(\vec{x}_i^*) \right) (\vec{x} - \vec{x}_i^*) \quad (57.489)$$

Soit:

$$-\begin{bmatrix} -16.5888 \\ -300.8722 \end{bmatrix} = \begin{bmatrix} 78.5367 & 1'335.8479 \\ 1'335.8479 & 24'559.9419 \end{bmatrix} \begin{pmatrix} x_1 - x_{1,0*} \\ x_2 - x_{2,0*} \end{pmatrix} \quad (57.490)$$

et après une petite simplification mineure:

$$\begin{bmatrix} 16.5888 \\ 300.8722 \end{bmatrix} = \begin{bmatrix} 78.5367 & 1'335.8479 \\ 1'335.8479 & 24'559.9419 \end{bmatrix} \begin{pmatrix} x_1 - x_{1,0*} \\ x_2 - x_{2,0*} \end{pmatrix} \quad (57.491)$$

Soit:

$$\begin{pmatrix} x_1 - x_{1,0*} \\ x_2 - x_{2,0*} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} x_{1,0*} \\ x_{2,0*} \end{pmatrix} = \begin{bmatrix} 78.5367 & 1'335.8479 \\ 1'335.8479 & 24'559.9419 \end{bmatrix}^{-1} \begin{bmatrix} 16.5888 \\ 300.8722 \end{bmatrix} = \begin{pmatrix} 0.0381 \\ 0.0102 \end{pmatrix} \quad (57.492)$$

et donc le prochain couple pour l'itération sera:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_{1,1*} \\ x_{2,1*} \end{pmatrix} = \begin{pmatrix} 0.0381 \\ 0.0102 \end{pmatrix} + \begin{pmatrix} x_{1,0*} \\ x_{2,0*} \end{pmatrix} = \begin{pmatrix} 0.0381 \\ 0.0102 \end{pmatrix} + \begin{pmatrix} 2.5 \\ 0.25 \end{pmatrix} = \begin{pmatrix} 2.5381 \\ 0.2602 \end{pmatrix} \quad (57.493)$$

Ce qui correspond bien évidemment aux valeurs pour la première itération:

$$\begin{pmatrix} x_{1,1*} \\ x_{2,1*} \end{pmatrix} = \begin{pmatrix} 2.5381 \\ 0.2602 \end{pmatrix} \quad (57.494)$$

Nous n'allons pas refaire aussi explicitement les autres itérations. Donc voilà ce que cela donne au final:

i	$\min_{x_1, x_2} f(x_1, x_2)$
0	$2 \cdot 10^0$
1	$4 \cdot 10^{-3}$
2	$2 \cdot 10^{-8}$
3	$3 \cdot 10^{-9}$
4	$3 \cdot 10^{-9}$

Tableau: 57.10 - Itérations Gauss-Newton

avec donc pour solution locale à la 4ème itération:

$$\begin{pmatrix} x_{1,4*} \\ x_{2,4*} \end{pmatrix} = \begin{pmatrix} 2.5411 \\ 0.2595 \end{pmatrix} \quad (57.495)$$

Faisons pour clore une comparaison avec la méthode de Newton-Raphson pour la première itération en utilisant le même couple de départ. Rappelons encore une fois que pour cette méthode, les itérations sont basées sur la relation:

$$-\vec{\nabla}(f(\vec{x}_i)) = H(\vec{x}_i)(\vec{x} - \vec{x}_i) \quad (57.496)$$

et nous écrivons la fonction de la façon suivant pour la méthode de Newton-Raphson:

$$f(x_1, x_2) = f(\vec{x}) = \frac{1}{2} \sum_{i=1}^5 \left(x_1 e^{x_2 t_i} - y_i \right)^2 \quad (57.497)$$

où nous avons donc pour la première itération (nous retrouvons le même vecteur que pour la méthode de Gauss-Newton):

$$\vec{\nabla}(f(\vec{x}_0)) = \vec{\nabla}(F(\vec{x}_0)^T) F(\vec{x}_0) = \begin{bmatrix} -16.5888 \\ -300.8722 \end{bmatrix} \quad (57.498)$$

et:

$$\begin{aligned}
H(\vec{x}_i) &= \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} f(x_i, y_i) & \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} f(x_i, y_i) \\ \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} f(x_i, y_i) & \frac{\partial^2}{\partial x_2^2} f(x_i, y_i) \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} \frac{1}{2} \sum_{i=1}^5 (x_1 e^{x_2 t_i} - y_i)^2 & \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} \frac{1}{2} \sum_{i=1}^5 (x_1 e^{x_2 t_i} - y_i)^2 \\ \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} \frac{1}{2} \sum_{i=1}^5 (x_1 e^{x_2 t_i} - y_i)^2 & \frac{\partial^2}{\partial x_2^2} \frac{1}{2} \sum_{i=1}^5 (x_1 e^{x_2 t_i} - y_i)^2 \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial}{\partial x_1} \sum_{i=1}^5 (x_1 e^{x_2 t_i} - y_i) e^{x_2 t_i} & \frac{\partial}{\partial x_1} \sum_{i=1}^5 (x_1 e^{x_2 t_i} - y_i) x_1 t_i e^{x_2 t_i} \\ \frac{\partial}{\partial x_1} \sum_{i=1}^5 (x_1 e^{x_2 t_i} - y_i) x_1 t_i e^{x_2 t_i} & \frac{\partial}{\partial x_2} \sum_{i=1}^5 (x_1 e^{x_2 t_i} - y_i) x_1 t_i e^{x_2 t_i} \end{bmatrix} \quad (57.499) \\
&= \begin{bmatrix} \sum_{i=1}^5 e^{2x_2 t_i} & \sum_{i=1}^5 2x_1 t_i e^{x_2 t_i} - y_i t_i e^{x_2 t_i} \\ \sum_{i=1}^5 2x_1 t_i e^{x_2 t_i} - y_i t_i e^{x_2 t_i} & \sum_{i=1}^5 2x_1^2 t_i e^{2x_2 t_i} - x_1 y_i t_i^2 e^{x_2 t_i} \end{bmatrix} \\
&= \begin{bmatrix} 78.5367 & 1'215.4991 \\ 1'215.4991 & 22'278.6570 \end{bmatrix}
\end{aligned}$$

Nous avons donc:

$$-\vec{\nabla}(f(\vec{x}_i)) = H(\vec{x}_i)(\vec{x} - \vec{x}_i) \quad (57.500)$$

qui devient:

$$\begin{bmatrix} 16.5888 \\ 300.8722 \end{bmatrix} = \begin{bmatrix} 78.5367 & 1'215.4991 \\ 1'215.4991 & 22'278.6570 \end{bmatrix} \begin{bmatrix} x_1 - x_{1,0*} \\ x_2 - x_{2,0*} \end{bmatrix} \quad (57.501)$$

Soit:

$$\begin{pmatrix} x_1 - x_{1,0*} \\ x_2 - x_{2,0*} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} x_{1,0*} \\ x_{2,0*} \end{pmatrix} = \begin{bmatrix} 78.5367 & 1'215.4991 \\ 1'215.4991 & 22'278.6570 \end{bmatrix}^{-1} \begin{bmatrix} 16.5888 \\ 300.8722 \end{bmatrix} = \begin{pmatrix} 0.0142 \\ 0.0127 \end{pmatrix} \quad (57.502)$$

et donc le prochain couple pour l'itération sera:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_{1,1} \\ x_{2,1} \end{pmatrix} = \begin{pmatrix} 0.0381 \\ 0.0102 \end{pmatrix} + \begin{pmatrix} x_{1,0*} \\ x_{2,0*} \end{pmatrix} = \begin{pmatrix} 0.0142 \\ 0.0127 \end{pmatrix} + \begin{pmatrix} 2.5 \\ 0.25 \end{pmatrix} = \begin{pmatrix} 2.5142 \\ 0.2627 \end{pmatrix} \quad (57.503)$$

Ce qui correspond bien évidemment aux valeurs pour la première itération:

$$\begin{pmatrix} x_{1,1*} \\ x_{2,1*} \end{pmatrix} = \begin{pmatrix} 2.5381 \\ 0.2602 \end{pmatrix} \quad (57.504)$$

Nous n'allons pas refaire aussi explicitement les autres itérations. Donc voilà ce que cela donne au final au niveau de la fonction à minimiser:

i	$\min_{x_1, x_2} f(x_1, x_2)$
0	$2 \cdot 10^0$
1	$1 \cdot 10^{-1}$
2	$2 \cdot 10^{-4}$
3	$5 \cdot 10^{-9}$
4	$6 \cdot 10^{-9}$
5	$3 \cdot 10^{-9}$

Tableau: 57.11 - Itérations Newton-Raphson

donc la méthode de Newton-Raphson converge dans ce cas particulier moins vite que celle de Gauss-Newton.

15. MÉTHODES DE MONTE-CARLO

La méthode de Monte-Carlo est un moyen très efficace de contourner les problèmes mathématiques et physiques les plus complexes. Elle trouve ses applications dans des domaines variés dont voici quelques exemples:

- Problèmes de neutronique liés à la bombe atomique (ou tout autre problème de la même famille)
- Calculs d'intégrales ou de paramètres divers de variables aléatoires (finance, risque)
- Résolution d'équations elliptiques ou paraboliques
- Résolution de systèmes linéaires
- Résolution de problèmes d'optimisation (recherche opérationnelle, gestion de projets)
- Création de tests statistiques (Anderson-Darling, Kolmogorov, Levene, Brown-Forsythe, etc.)

Il existe donc deux types de problèmes qui peuvent être traités par la méthode de Monte-Carlo: les problèmes probabilistes, qui ont un comportement aléatoire, et les problèmes déterministes, qui n'en ont pas.

Pour ce qui est du cas probabiliste, il consiste à observer le comportement d'une série de nombres aléatoires qui simule le fonctionnement du problème réel et à en tirer les solutions.

Pour le cas déterministe, le système étudié est complètement défini et on peut en principe prévoir son évolution, mais certains paramètres du problème peuvent être traités comme s'il s'agissait de variables aléatoires. Le problème déterministe devient alors probabiliste et résoluble de façon numérique. On parle alors d'estimation de "Monte-Carlo" ou d'une approche de "MC élaborée".

La dénomination de méthode de "Monte-Carlo" date des alentours de 1944. Des chercheurs isolés ont cependant utilisé bien avant des méthodes statistiques du même genre: par exemple, Hall pour la détermination expérimentale de la vitesse de la lumière (1873), ou Kelvin dans une discussion de l'équation de Boltzmann (1901), mais la véritable utilisation des méthodes de Monte-Carlo commença avec les recherches sur la bombe atomique.

Au cours de l'immédiate après-guerre, Von Neumann, Fermi et Ulam avertirent le public scientifique des possibilités d'application de la méthode de Monte-Carlo (par exemple, pour l'approximation des valeurs propres de l'équation de Schrödinger). L'étude systématique en fut faite par Harris et Hermann Khan en 1948. Après une éclipse due à une utilisation trop intensive pendant les années 1950, la méthode de Monte-Carlo est revenue en faveur pour de nombreux problèmes: en sciences physiques, en sciences économiques, pour des prévisions électorales, etc., bref, partout où il est fructueux d'employer des procédés de simulation.

15.1. GÉNÉRATION DES VARIABLES ALÉATOIRES

Le mieux pour comprendre la méthode de Monte-Carlo c'est de faire des exemples. Mais pour cela, il faut d'abord d'avoir un très bon générateur de nombres aléatoires (ce qui est très difficile). C'est un domaine très délicat et sensible donc pour lequel des normes internationales ont été édictées (ISO 28640:2010).

Prenons comme exemple le générateur de Maple 4.00b:

```
>rand();
```

427419669081

```
>restart;rand();
```

427419669081

Nous voyons donc que la fonction par défaut de générateur de nombres aléatoires de Maple 4.00b est à utiliser avec la plus grande prudence puisqu'une réinitialisation du système suffit à retrouver des valeurs aléatoires... égales. Il s'agit donc d'un "**générateur pseudo-aléatoire**" permettant de faire des simulations appelées parfois "**pseudo Monte-Carlo**".

Cependant il existe des libraires spécialisées dans Maple 4.00b telles que:

```
>restart;readlib(randomize):randomize():rand();
```

536427565027

```
>restart;readlib(randomize):randomize():rand();
```

818686572281

Épreuve a priori réussie (au fait, il nous faudrait faire un beaucoup plus grand nombre d'essais afin de bien vérifier que le générateur ne suit pas une loi de distribution connue... ce qui n'est malheureusement jamais le cas).

Les fonctions ALEA() et ALEA.ENTRE.BORNES() de la version française de Microsoft Excel 14.0.6123 sont aussi des générateurs pseudo-aléatoires dont voici un échantillon de 100 simulations (évidemment dans Microsoft Excel le graphique ci-dessous changera à chaque fois que vous activerez la touche F9 du clavier):

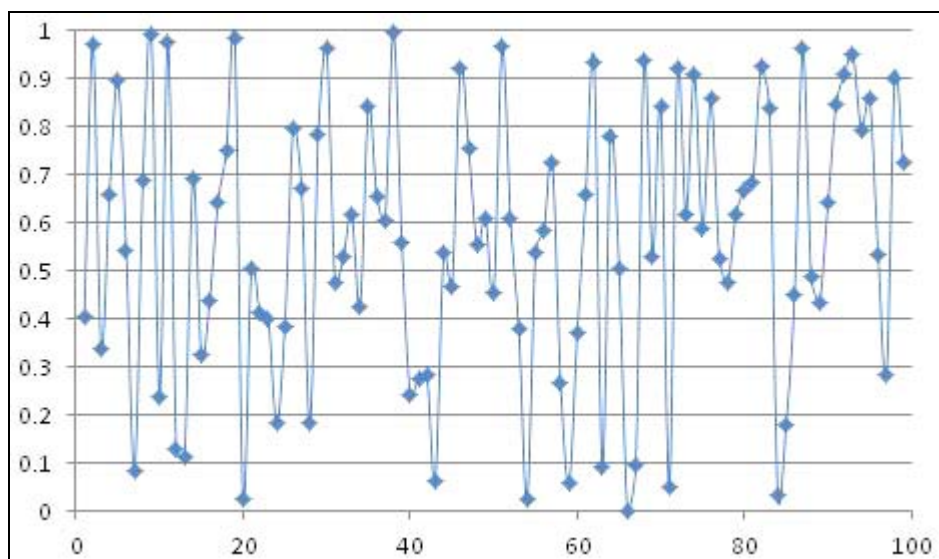


Figure: 57.34 - Illustration d'une séquence de nombres pseudo-aléatoires avec Microsoft Excel 14.0.6123

Il peut malheureusement arriver avec les nombres pseudo-aléatoires que les nombres générés se présentent en grappes, c'est-à-dire en séries de nombres rapprochés les uns des autres, ce qui nuit à l'efficacité de la simulation de Monte-Carlo.

Une technique empirique consiste à faire appel à des séquences de nombres générés sur la base d'algorithmes qui balaient à coup sûr l'intervalle $[0,1]$. Nous parlons alors de "**nombres quasi-aléatoires**" permettant de faire des simulations appelées parfois "**quasi Monte-Carlo**". Avec la version française de Microsoft Excel 11.8346, il est possible de créer une fonction qui remplacera les générateurs pseudo-aléatoires que sont les fonctions ALEA() ou ALEA.ENTRE.BORNES().

Voici donc un exemple de fonction en V.B.A. (Visual Basic for Application) qui génère des nombres quasi-aléatoires appelés "**séquence de Fauré**":

```

Function SequenceFaure(n) As Double
Dim f As Double, sb As Double
Dim i As Integer, n1 As Integer, n2 As Integer
n1 = n
sb = 1 / 2
Do While n1 > 0
n2 = Int(n1 / 2)
i = n1 - n2 * 2
f = f + sb * i
sb = sb / 2
n1 = n2
Loop
SequenceFaure = f
End Function

```

Ce qui donnera la séquence suivante pour un échantillon de 100 simulations:

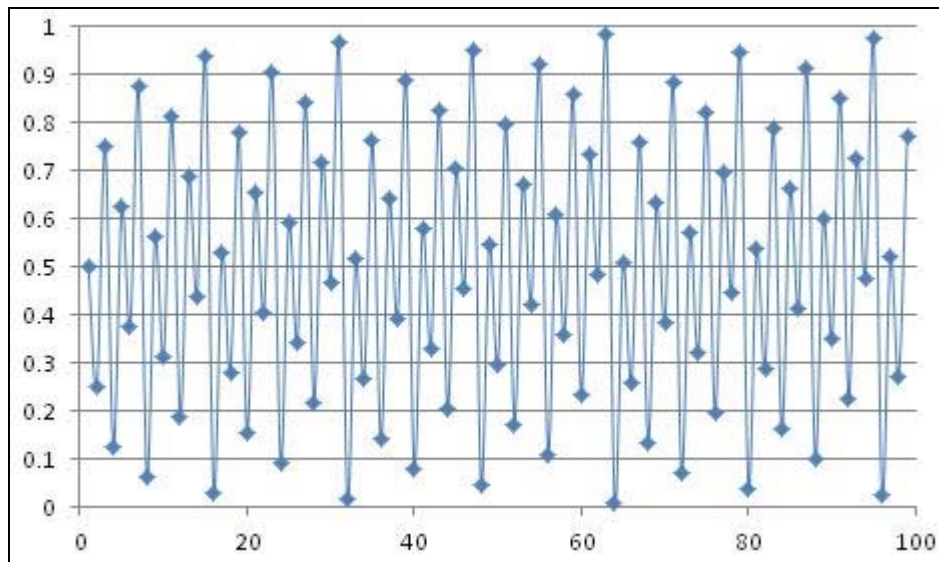


Figure: 57.35 - Illustration d'une séquence de nombres quasi-aléatoires avec Microsoft Excel 11.8346

où nous voyons bien que la séquence couvre bien la surface comprise entre 0 et 1 (nous disons alors qu'elle couvre plus rapidement la surface d'intégration). Cette technique est parfois appréciée car elle a pour avantage de conserver les valeurs de la simulation à chaque fois que l'on relance la simulation (donc dans Microsoft Excel 11.8346 le graphique ci-dessous ne changera pas quand vous activerez la touche F9 du clavier).

Par contre les générateurs de séquence ont une grande faiblesse: ils ne sont applicables (à ma connaissance du moins) que pour des problèmes de simulations avec une seule et unique variable aléatoire (typiquement du pricing d'options selon Black & Scholes). Effectivement si nous avons plusieurs variables aléatoires (et c'est le cas le plus courant!), alors les variables sont artificiellement corrélées (coefficient de corrélation égal à 1) car elles parcourent toutes la surface comprise entre 0 et 1 de la même manière. Donc une bonne simulation avec plusieurs variables est une simulation dont les variables traitées ont un coefficient de corrélation qui tend vers zéro.

De plus, les générateurs de séquence nécessitent des algorithmes qui sont très gourmands lorsqu'il y a de nombreuses variables par rapport à un générateur pseudo-aléatoire, raison pour laquelle dans la majorité des situations, on préférera cette bonne vieille méthode.

Remarque: Se référer à la norme internationale ISO 28640:2010 pour les ingénieurs ayant besoin d'implémenter des générateurs de nombres aléatoires dans leurs logiciels.

Une fois le générateur créé et testé, nous pouvons voir quelques résultats de la méthode de Monte-Carlo. Ainsi, dans le calcul des intégrales, celle-ci s'avère très utile et très rapide en termes de vitesse de convergence.

15.2. CALCUL D'UNE INTÉGRALE

Soit à calculer l'intégrale d'une fonction f définie et positive sur l'intervalle $[a,b]$:

$$I = \int_a^b f(x) dx \quad (57.505)$$

Soit:

$$m = \max(f(a), f(b)) \quad (57.506)$$

la valeur maximale de la fonction entre les bornes $[a,b]$.

Nous considérons le rectangle englobant la fonction sur $[a,b]$ défini par les points $(a, 0), (b, 0), (b, m), (a, m)$:

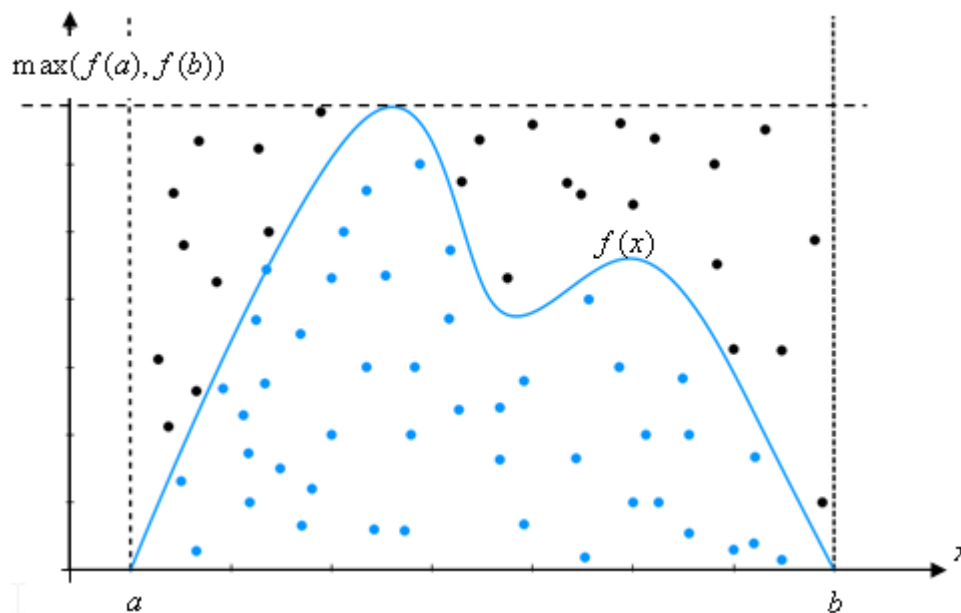


Figure: 57.36 - Principe de base du calcul de l'intégrale avec Monte-Carlo

Nous tirons un grand nombre N de points au hasard dans ce rectangle. Pour chaque point, nous testons s'il est au-dessous de la courbe. Soit F la proportion de points situés au-dessous, nous avons:

$$I = (b - a) \cdot m \cdot \frac{F}{N} \quad (57.507)$$

L'algorithme Maple 4.00b est donné par:

```

intmonte:=proc(f,a,b,N)
local i,al,bl,m,F,aleaabs,aleaord,estaudessous;
m:=round(max(a,b)*10^4);
al:=round(a*10^4);
bl:=round(b*10^4);
aleaabs:=rand(al..bl);
aleaord:=rand(0..m);
F:=0;
for i from 1 to N do
    estaudessous:=(f(aleaabs()/10^4)-aleaord()/10^4)>=0;
    if estaudessous then
        F:=F+1;
    fi
od;
RETURN((b-a)*max(a,b)*F/N)
end:

```

Remarque: Pour appeler cette procédure, il suffit d'écrire `>intmonte(f,a,b,N)` mais en remplaçant le premier argument passé en paramètre par l'expression d'une fonction et les autres arguments par des valeurs numériques bien évidemment.

15.3. CALCUL DE PI

Pour le calcul de π le principe est le même et consiste donc à utiliser la proportion du nombre de points dans un quartier de cercle (cela permet de simplifier l'algorithme en se restreignant aux coordonnées strictement positives) inscrit dans un carré relativement au nombre de points totaux (pour tester si un point est à l'extérieur du cercle, nous utilisons bien évidemment le théorème de Pythagore) tel que:

$$\pi = 4 \cdot \frac{F}{N} \quad (57.508)$$

L'algorithme Maple 4.00b est donné par:

```

estalinterieur:=proc(x,y) x^2+y^2<1 end;
calculepi:=proc(N)
local i,F,abs,ord,alea,erreur,result;
alea:=rand(-10^4..10^4);
F:=0;
for i from 1 to N do
    abs:=alea()/10^4;ord:=alea()/10^4;
    if estalinterieur(abs,ord) then
        F:=F+1;
    fi
od;
RETURN(4*F/N)
end:

```

15.4. MODÉLISATION

L'application la plus courante de la méthode par Monte-Carlo est certainement l'étude de variables aléatoires. Par ailleurs, cette méthode fait partie intégrante de la norme ISO 31010 de gestion du risque sous le nom "analyse de Monte-Carlo" tellement elle est courante et utile. De nombreuses entreprises font de la modélisation de Monte-Carlo avec un tableur comme Microsoft Excel (même les multinationales!) et dans une moindre mesure avec des logiciels comme @Risk, CrystalBall, TreeAge ou encore Isograph.

Les avantages de cette méthode dans la modélisation de variables aléatoires sont les suivants:

- On peut intégrer n'importe quelle distribution dans une variable d'entrée, y compris des empiriques!
- Les modèles sont très simples à développer et peuvent être étendus à mesure des besoins
- Toutes les influences ou relations se produisant dans la réalité peuvent être représentées
- L'analyse de la sensibilité (cf. [chapitre Techniques De Gestion](#)) peut être appliquée
- Les modèles sont aisément compréhensibles et fournissent une mesure de l'exactitude d'un résultat
- De nombreux logiciels sont disponibles et peu onéreux

Considérons un cas simple mais concret (très utilisé dans les entreprises) d'un petit projet de deux tâches notées *A* et *B* qui se succèdent sans marge libre. La durée de chacune des tâches a été estimée conformément à la recommandation du Project Management Institute avec une loi bêta (cf. [chapitre de Statistiques](#)) comme l'apprennent tous les responsables de projets lors de leur cursus de formation (cf. [chapitre Techniques De Gestion](#)).

Pour cet exemple, la tâche *A* a une durée optimiste de 5 jours et pessimiste de 8 jours. La tâche *B* une durée optimiste de 1 jour et pessimiste de 4 jours. Nous souhaiterions dans Microsoft Excel à l'aide d'une simulation de pseudo Monte-Carlo (basée donc obligatoirement sur une variable pseudo-aléatoire) introduire les trois informations traditionnelles minimales suivantes:

- Un tableau avec les 3 colonnes (durée de *A*, de *B* et somme des deux) de 10'000 simulations
- La fonction de distribution de la somme des deux variables aléatoires sous forme graphique
- La convergence du 95ème centile des 100 premières simulations (utile pour le sujet d'après).

Nous construisons alors le tableau suivant sur 10'000 lignes (la capture d'écran ne prend que les premières...):

	A	B	C
1	Tâche A	Tâche B	Somme
2	7.66	3.69	10.58
3	7.72	3.53	11.25
4	6.82	3.16	9.98
5	6.75	3.04	9.79
6	7.77	3.36	11.13
7	7.64	2.68	10.31
8	7.69	2.95	10.64
9	7.72	3.07	10.78
10	7.31	3.55	10.86
11	6.76	3.62	10.38
12	6.50	3.04	9.54
13	6.53	3.68	10.21
14	7.26	2.77	10.02
15	6.99	3.54	10.54
16	6.19	2.99	9.18
17	7.36	3.54	10.90
18	7.77	3.81	11.58
19	7.62	3.61	11.23
20	7.67	3.27	10.94
21	7.59	3.38	10.97
22	7.33	3.25	10.58

Figure: 57.37 - Mise en place d'une petite simulation de Monte-Carlo avec Microsoft Excel 11.8346

où toutes les cellules de la colonne A contiennent la fonction suivante (version français de Microsoft Excel 14.0.6123):

=BETA.INVERSE.N(ALEA.ENTRE.BORNES(1;9999)/10000;3+RACINE(2);3-RACINE(2);5;8)

et toutes les cellules de la colonne B contiennent la fonction suivante (version français de Microsoft Excel 14.0.6123):

=BETA.INVERSE.N(ALEA.ENTRE.BORNES(1;9999)/10000;3+RACINE(2);3-RACINE(2);1;4)

et enfin la colonne C contient la fonction suivante:

=A2+B2

évidemment les valeurs dans Microsoft Excel 14.0.6123 changeront à chaque fois que vous activerez la touche F9 du clavier.

Cela nous donne alors pour l'histogramme (dont je ne vais pas détailler la construction car il s'agit d'un sujet élémentaire en bureautique):

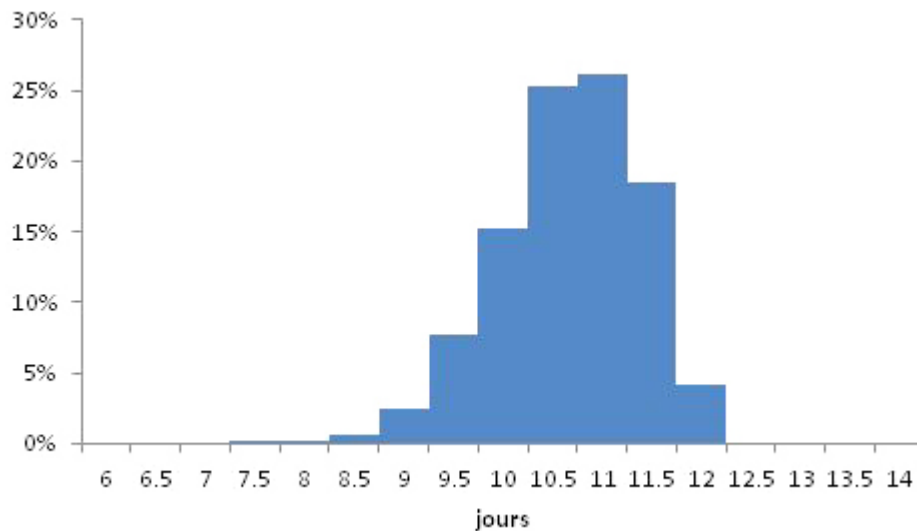


Figure: 57.38 - Distribution obtenue de la somme des variables aléatoires

et la convergence du 95ème centile sur les 100 premières simulations (car le problème étant simple, le système converge suffisamment rapidement pour ne pas avoir besoin d'en prendre plus de 100 pour l'exemple):

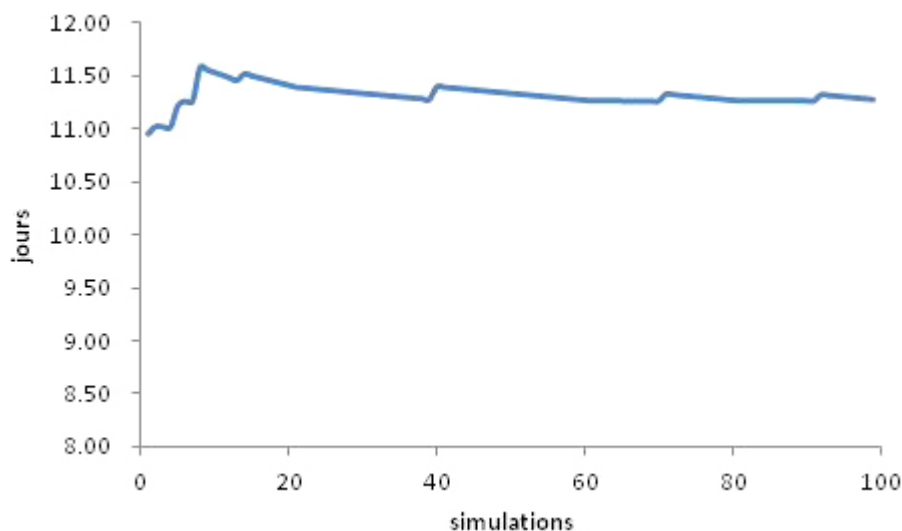


Figure: 57.39 - Illustration de la convergence du 95ème centile

Évidemment dans Microsoft Excel 14.0.6123 les graphiques ci-dessus changeront à chaque fois que vous activez la touche F9 du clavier.

Remarque: Dans le cas de simulations de variables aléatoires, on peut dans les cas simples impliquant uniquement des sommes ou soustractions de variables aléatoires, comme c'est le cas pour l'exemple ci-dessus, déterminer l'espérance et l'écart-type du résultat analytiquement en utilisant la propriété de linéarité de l'espérance et de la variance (car normalement pour la variance de deux variables aléatoires indépendantes, la covariance est nulle). En analysant la différence entre la valeur analytique et celle obtenue par la simulation numérique, on peut corriger le décalage de certains autres indicateurs statistiques par simple ajout ou soustraction du différentiel. On parle alors de la technique des "**variables de contrôle**".

Il existe d'autres techniques de réduction de la variance (ou in extenso: de l'écart-type) que la méthode de quasi Monte-Carlo permettant de réduire la variance des estimateurs de Monte-Carlo dans certaines conditions particulières:

1. Une de ces techniques est l'usage des "**variables antithétiques**" qui consiste très simplement (la programmation de cette technique est du niveau du lycée) à décorréler les simulations pour rendre la covariance entre les variables négatives et ainsi de réduire la variance (puisque comme nous l'avons vu dans le chapitre de Statistiques, la variance de la somme de deux variables aléatoires fait apparaître un terme de covariance). Malheureusement, cette technique ne fonctionne de manière satisfaisante qu'avec des distributions symétriques ce qui fait qu'à ma connaissance elle n'est pas implémentée dans les logiciels de simulation disponibles sur le marché.
2. Il existe aussi la technique "**d'échantillonnage stratifié**" qui consiste à découper l'espace des pré-images de la variable aléatoire en intervalles réguliers (la programmation de cette technique est aussi du niveau du lycée). Cette technique fonctionne très bien lorsque le nombre de simulations doit être faible mais seulement dans le cas d'une unique variable. Raison pour laquelle elle n'est pas implémentée à ma connaissance dans les logiciels de simulation disponibles sur le marché.
3. Il existe une généralisation de l'échantillonnage stratifié (la programmation de cette technique est aussi du niveau du lycée) pour les simulations comportant plusieurs variables et qui se nomme "**Latin Hypercube**" (abrégié "LHS" pour Latin Hypercube Stratification). Cette technique assure donc que chaque n -uplet de variables aléatoires (correspondant à un espace à n -dimensions) utilise une pré-image unique à chaque itération, d'où le nom de la technique (Latin: fait référence aux carrés magiques où chaque valeur apparaît de manière unique, Hypercube: car il s'agit d'une généralisation à n dimensions d'un carré magique). Certains logiciels de simulation disponibles sur le marché implémentent cette technique (@Risk, CrystalBall).

Pour résumer, que ce soit la technique des générateurs de séquence de Fauré, des variables antithétiques, des variables de contrôle, de l'échantillonnage stratifié ou de Latin Hypercube même si ces techniques sont toutes faciles à programmer, la méthode utilisant les variables pseudo aléatoires est privilégiée car est la plus adaptée à la majorité des situations courantes de l'économie mondiale.

16. BOOTSTRAPPING

En Statistiques, les techniques de bootstrap sont des méthodes d'inférence statistique requérant des calculs informatiques relativement intensifs. L'objectif est de connaître certaines indications sur une statistique : son estimation bien sûr, mais aussi la dispersion (variance, écart-type), des intervalles de confiance voire un test d'hypothèse. Cette méthode est basée sur des simulations, comme les méthodes de Monte-Carlo, à la différence près que le bootstrap ne nécessite pas d'information supplémentaire que celle disponible dans l'échantillon. En général, il est basé sur de nouveaux échantillons obtenus par tirage avec remise à partir de l'échantillon initial (on parle de rééchantillonnage).

Nous distinguons en général, deux types de bootstrap:

- Les bootstraps qui ne font aucune hypothèse sur la loi de distribution des données analysées. Nous parlons alors de "**bootstrap non-paramétrique**". C'est le cas le plus courant et nous ferons un exemple uniquement pour celui-ci.

- Les bootstraps qui remplacent chacune des données mesurées par celles correspondantes à l'expression analytique de la loi distribution de probabilité supposée. Nous parlons alors de "**bootstrap paramétrique**". Une fois le remplacement de chacune de valeurs d'origine effectué, la démarche est exactement celle du bootstrap non-paramétrique.

Nous allons illustrer le principe du bootstrap (dit aussi "bootstrapping") sur l'exemple de l'intervalle de confiance de l'espérance μ d'une variable aléatoire. Pour cet exemple, l'intervalle de confiance de l'espérance d'une variable aléatoire est parfaitement déterminé à partir de la moyenne et de la variance calculées sur l'échantillon (cf. chapitre de Statistiques).

Nous considérons un échantillon de la variable aléatoire composé de $n = 10$ estimations:

$$X = \{16, 12, 14, 6, 43, 7, 0, 54, 25, 13\} \quad (57.509)$$

La moyenne arithmétique de l'échantillon est:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{10} \sum_{i=1}^{10} x_i = 19 \quad (57.510)$$

et son écart-type (estimateur de maximum de vraisemblance non biaisé):

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2} = \sqrt{\frac{1}{9} \sum_{i=1}^{10} (x_i - 19)^2} = 17.09 \quad (57.511)$$

Comme nous sommes dans la situation d'une moyenne empirique connue et d'une variance empirique connue, pour faire le calcul d'un intervalle de confiance, nous avons alors démontré dans le chapitre de Statistiques qu'il fallait utiliser:

$$\bar{X} - \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \quad (57.512)$$

où S est une autre notation traditionnelle dans certains domaines de la statistique pour la notation de l'écart-type empirique (cf. chapitre de Statistiques). Nous avons alors pour l'intervalle de confiance à 95% de l'espérance:

$$19 - \frac{17.09}{\sqrt{10}} T_{2.5\%}(9) \leq \mu \leq 19 + \frac{17.09}{\sqrt{10}} T_{2.5\%}(9) \quad (57.513)$$

Soit:

$$19 - \frac{17.09}{\sqrt{10}} 2.26 \leq \mu \leq 19 + \frac{17.09}{\sqrt{10}} 2.26 \quad (57.514)$$

Ce qui donne:

$$4.49 \leq \mu \leq 33.51 \quad (57.515)$$

L'intervalle de confiance peut être également calculé par bootstrap. Il est alors obtenu par l'algorithme suivant:

À partir de l'échantillon initial, nous simulons de nouveaux échantillons, appelés "**répliques**", de taille n , par tirages aléatoires avec remise. Par exemple avec la série précédente, nous pourrions obtenir la réplique suivante:

$$X^* = \{54, 0, 16, 7, 43, 54, 0, 25, 25, 6\} \quad (57.516)$$

dans laquelle certaines valeurs de l'échantillon initial ne figurent pas, et où d'autres apparaissent plusieurs fois. Plusieurs échantillons sont ainsi simulés. Nous pouvons ainsi former un nombre de répliques (arrangements) égal à (cf. chapitre de Probabilités):

$$\bar{A}_n^n = n^n \quad (57.517)$$

Pour chaque échantillon simulé, une moyenne est calculée (plusieurs milliers de moyennes!). L'intervalle de confiance à 95% est défini sur cet ensemble de moyennes typiquement à l'aide du calcul des centiles (via les fonctions d'un tableur ou d'un langage de programmation).

Évidemment pour chaque ensemble de plusieurs milliers de valeurs, les centiles ne seront pas les mêmes donc il est même possible de créer un intervalle de confiance pour les centiles eux-mêmes!

Il est très facile (au même titre que la méthode de Monte-Carlo) de créer des répliques avec des tableurs (de type Microsoft Excel) sans faire de la programmation informatique! En plus la technique du bootstrap est très puissante car elle ne fait appel à aucune hypothèse sur la distribution statistique sous-jacente. Le domaine le plus courant et simple d'application du bootstrapping est la gestion de projets où lors de réunions avec une dizaine de ressources chacune estime la durée d'une tâche ou d'une phase.

Le bootstrap peut donc être appliqué à tout estimateur autre que la moyenne, tel que la médiane, le coefficient de corrélation entre deux variables aléatoires ou la valeur propre principale d'une matrice de variance-covariance (pour l'analyse en composantes principales) et c'est là sa grande force!!! Effectivement, pour ces estimateurs, il n'existe pas de relation mathématique qui définisse l'erreur-standard ou l'intervalle de confiance. Les seules méthodes applicables sont des "**méthodes de rééchantillonnage**" (resampling) comme en fait partie le bootstrapping.

Exemple:

Avec par exemple le tableau Microsoft Excel 14.0.6123 et en s'interdisant de faire de la programmation VBA, nous construisons un petit tableau avec l'échantillon:

	A
4	X
5	16
6	12
7	14
8	6
9	43
10	7
11	0
12	54
13	25
14	13

Figure: 57.40 - Échantillon de base

Nous souhaiterions pouvoir déterminer un intervalle de confiance de la médiane (nous faisons exprès de prendre un indicateur statistique pour lequel il n'existe pas d'intervalle de confiance analytique). Pour cela, nous calculons la médiane de plusieurs milliers de répliques où chaque réplique correspond à une ligne:

	F
4	Médiane Réplique
5	12.50
6	9.50
7	13.00
8	12.50
9	12.00
10	20.50
11	12.00
12	25.00
13	20.50
14	15.00
15	15.00
16	14.50
17	13.50
18	12.00
19	15.00
20	7.00
21	15.00
22	10.50
23	13.00
24	7.00
25	14.50

Figure: 57.41 - Médianes de répliques

avec la longue formule pour la version française de Microsoft Excel 14.0.6123 suivante qu'il faut mettre dans la cellule F5 et tirer ensuite jusqu'à la fin de la feuille:

```
=MEDIANE(INDEX($A$5:$A$14;ALEA.ENTRE.BORNES(1;10);1);
INDEX($A$5:$A$14;ALEA.ENTRE.BORNES(1;10);1);
INDEX($A$5:$A$14;ALEA.ENTRE.BORNES(1;10);1);
INDEX($A$5:$A$14;ALEA.ENTRE.BORNES(1;10);1);
INDEX($A$5:$A$14;ALEA.ENTRE.BORNES(1;10);1);
INDEX($A$5:$A$14;ALEA.ENTRE.BORNES(1;10);1);
INDEX($A$5:$A$14;ALEA.ENTRE.BORNES(1;10);1);
INDEX($A$5:$A$14;ALEA.ENTRE.BORNES(1;10);1);
INDEX($A$5:$A$14;ALEA.ENTRE.BORNES(1;10);1);
INDEX($A$5:$A$14;ALEA.ENTRE.BORNES(1;10);1))
```

Nous pourrions donc en avoir 10 milliards pas plus... (comme Microsoft Excel 14.0.6123 est limité à un million de lignes, cela coupe net toute discussion...).

Il suffit ensuite dans une cellule de votre choix de mettre:

```
=CENTILE(F5:F2003;0.025)
```

et dans une autre:

```
=CENTILE(F5:F2003;0.975)
```

ce qui avec 2'000 réplifications donnera respectivement 7 et 29.5.

Avec des connaissances élémentaires du tableur, il est possible de montrer graphiquement la convergence de la moyenne de la médiane en fonction du nombre de réplifications (ci-dessous avec les 100 premières réplifications):

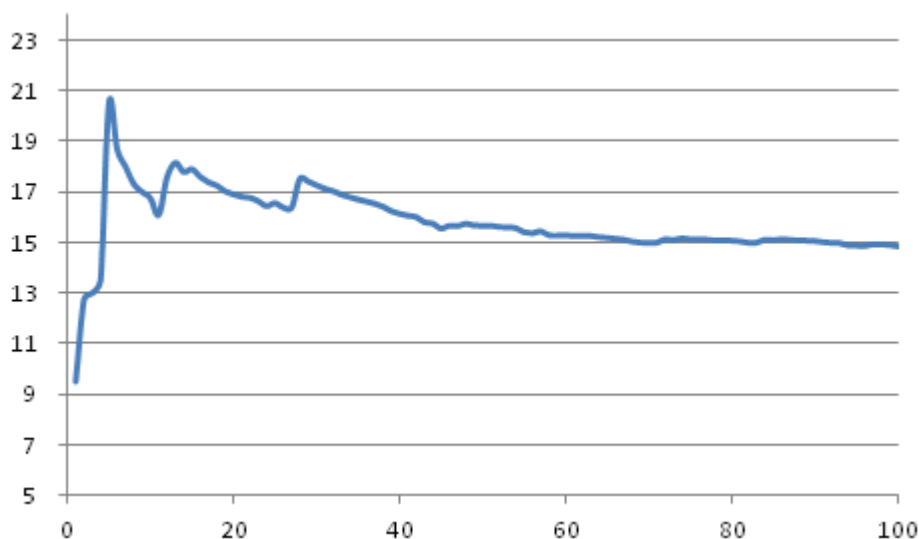


Figure: 57.42 - Convergence de la médiane en fonction du nombre de réplifications

Évidemment, ce graphique aura un aspect différent à chaque fois que vous relancerez la simulation dans Microsoft Excel 14.0.6123 en appuyant sur la touche F9 du clavier

17. DICHOTOMIE

La dichotomie consiste pour un objet de taille N à exécuter un algorithme de façon à réduire la recherche à un objet de taille $N/2$. On répète l'algorithme de réduction sur ce dernier objet. Ce type d'algorithme est souvent implémenté de manière récursive. Lorsque cette technique est utilisable, elle conduit à un algorithme très efficace et très lisible.

Un exemple simple est la recherche de la racine d'une fonction continue (nous avons déjà étudié différentes méthodes plus haut pour résoudre ce genre de problèmes). C'est-à-dire le point pour lequel la fonction f s'annule.

Supposons qu'une fonction soit croissante et continue sur un intervalle $[a,b]$ et telle que la racine cherchée soit entre a et b . Nous avons donc par le fait que la fonction soit croissante dans l'intervalle:

$$f(a) < f(b) \quad (57.518)$$

et le fait que la racine se trouve dans l'intervalle:

$$f(a) \cdot f(b) < 0 \quad (57.519)$$

Nous calculons:

$$M = f\left(\frac{a+b}{2}\right) \quad (57.520)$$

Si $M > 0$ alors la racine est dans l'intervalle $[f^{-1}(M), b]$ sinon elle est dans l'intervalle $[a, f^{-1}(M)]$.

Nous avons donc ramené le problème à une taille inférieure. Nous arrêterons l'algorithme quand la précision sera suffisante.

L'algorithme Maple 4.00b est donné par:

```
zero:=proc(f,a,b,pre)
local M;
M:=f((a+b)/2);
if abs(M)<pre then
  RETURN((a+b)/2)
elif M>0 then
  zero(f,a,(a+b)/2,pre)
else zero(f,(a+b)/2,b,pre)
fi
end;
```

et ce ne sont que quelques exemples auxquels la méthode est applicable!!

18. ANALYSE EN COMPOSANTES PRINCIPALES (A.C.P.)

L'analyse en composantes principales (A.C.P.) est une méthode mathématique d'analyse graphique de données qui consiste à rechercher les directions de l'espace qui représentent le mieux les corrélations entre n variables aléatoires (relations linéaires entre elles).

Simplement dit, une A.C.P. permet par exemple de trouver des similitudes de comportement d'achat entre les classes des données observées.

Même si l'A.C.P. est majoritairement utilisée pour visualiser des données, il ne faut pas oublier que c'est aussi un moyen:

- De décorrélérer ces données. Dans la nouvelle base, constituée des nouveaux axes, les points ont une corrélation nulle (nous le démontrerons).
- De classer ces données en amas (clusters) corrélés (dans l'industrie c'est surtout cette possibilité qui est intéressante!).

Remarque: Il existe plusieurs versions de l'A.C.P. connues sous le nom de "**transformée de Karhunen-Loève**" ou de "**transformée de Hotelling**" et qui peuvent aussi bien être appliquées sans programmation V.B.A. dans Microsoft Excel que dans des logiciels spécialisés (où le temps de calcul sera par contre plus bref... et les résultats plus précis aussi...).

Lorsque nous ne considérons que deux effets, il est usuel de caractériser leur effet conjoint via le coefficient de corrélation. Lorsque l'on se place en dimension deux, les points disponibles (l'échantillon de points tirés suivant la loi conjointe) peuvent être représentés dans un plan. Le résultat d'une A.C.P. dans ce plan consiste en la détermination des deux axes qui expliquent le mieux la dispersion des points disponibles.

Lorsqu'il y a plus de deux effets, par exemple trois effets, il y a trois coefficients de corrélations à prendre en compte. La question qui a donné naissance à l'A.C.P. est: comment avoir une intuition rapide des effets conjoints?

En dimension plus grande que deux, une A.C.P. va toujours déterminer les axes qui expliquent le mieux la dispersion du nuage des points disponibles.

L'objectif de l'A.C.P. est de décrire graphiquement un tableau de données d'individus avec leurs variables quantitatives de grande taille:

individus/variables	$var_1 \cdots var_j \cdots var_p$
ind_1	x_{ij}
$\vdots$	
ind_i	
$\vdots$	
ind_n	

Tableau: 57.12 - Représentation-type d'un tableau A.C.P.

Afin de ne pas alourdir l'exposé de cette méthode et de permettre au lecteur de refaire complètement les calculs, nous travaillerons sur un exemple.

Considérons pour l'exemple une étude d'un botaniste qui a mesuré les dimensions de 15 fleurs d'iris. Les trois variables ($p = 3$) mesurées sont:

- x_1 : longueur du sépale
- x_2 : largeur du sépale
- x_3 : longueur du pétale

Les données sont les suivantes:

Fleur n°	x_1	x_2	x_3
1	5.1	3.5	1.4
2	4.9	3.0	1.4
3	4.7	3.2	1.3
4	4.6	3.1	1.5
5	5.0	3.6	1.4
6	7.0	3.2	4.7
7	6.4	3.2	4.5
8	6.9	3.1	4.9
9	5.5	2.3	4.0
10	6.5	2.8	4.6
11	6.3	3.3	6.0
12	5.8	2.7	5.1
13	7.1	3.0	5.9
14	6.3	2.9	5.6
15	6.5	3.0	5.8

Tableau: 57.13 - Exemple pratique de données tabulaires A.C.P.

Pour nous un tel tableau de données sera tout simplement une matrice réelle à n lignes (les individus) et à p colonnes (les variables):

$$X = (x_{ij})_{\substack{i=1 \dots n \\ j=1 \dots p}} \quad (57.521)$$

Par la suite l'indice i correspondra à l'indice ligne et donc aux individus. Nous identifierons donc l'individu i avec le point ligne $x_i = (x_{i1}, \dots, x_{ip})$ qui sera considéré comme un point dans un espace affine (cf. chapitre de Calcul Vectoriel) de dimension p . L'indice j correspondra à l'indice colonne donc aux variables. Nous identifierons la variable j avec le vecteur-colonne:

$$\vec{x}_{\cdot j} = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix} \quad (57.522)$$

c'est donc un vecteur dans l'espace vectoriel de dimension n dans $\mathbb{R}^n$.

Nous nous placerons dans la suite suivant deux points de vue: soit nous prendrons le tableau de données comme n points dans un espace affine de dimension p , soit nous prendrons ce tableau comme p points d'un espace vectoriel de dimension n . Nous verrons qu'il y a des dualités entre ces deux points de vue.

L'outil mathématique que nous allons utiliser ici est l'algèbre linéaire (cf. chapitre d'Algèbre Linéaire), avec les notions de produit scalaire, de norme euclidienne et de distance euclidienne.

Afin de simplifier la présentation, nous allons dans un premier temps considérer que chaque individu, comme chaque variable, a la même importance, le même poids. Nous ne considérerons aussi que le cas de la distance euclidienne.

Nous allons commencer en centrant les données, c'est-à-dire mettre l'origine du système d'axes au centre de gravité du nuage de points. Ceci ne modifie pas l'aspect du nuage, mais permet d'avoir les coordonnées du point M égales aux coordonnées du vecteur $\overrightarrow{GM}$ et donc de se placer dans l'espace vectoriel pour pouvoir y faire les calculs! Comme nous supposons dans toute la suite que les poids des individus sont identiques, nous prendrons donc $m_i = 1/n$ avec $i = 1 \dots n$.

Nous considérons le repère orthonormé $(O, \vec{e}_1, \vec{e}_2, \dots, \vec{e}_p)$ dans la base canonique $(\vec{e}_1, \vec{e}_2, \dots, \vec{e}_p)$ de $\mathbb{R}^p$. Soit donc G le centre de gravité du nuage de points. Comme chaque variable ou chaque individu est supposé avoir le même poids, G a alors pour coordonnées dans le repère $(O, \vec{e}_1, \vec{e}_2, \dots, \vec{e}_p)$:

$$\overrightarrow{OG} = \begin{pmatrix} \bar{x}_{.1} \\ \vdots \\ \bar{x}_{.j} \\ \vdots \\ \bar{x}_{.p} \end{pmatrix} \quad (57.523)$$

avec:

$$\bar{x}_{.j} = \frac{1}{n} \sum_{i=1}^n x_{ij} \quad (57.524)$$

Nous avons alors pour l'instant sous forme graphique:

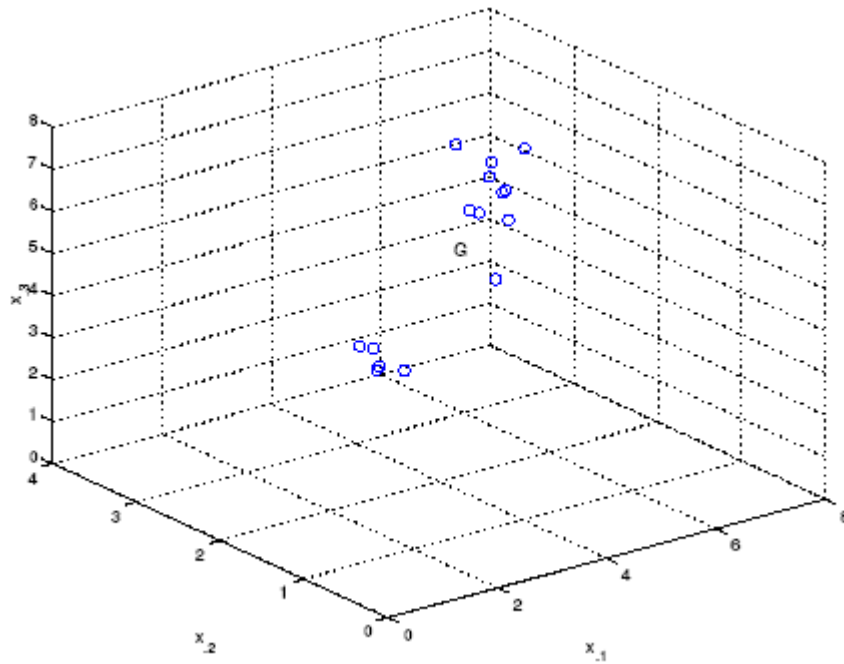


Figure: 57.43 - Points de mesures et centre de gravité

Nous appelons "**matrice centrée**" la matrice:

$$X_c = \begin{bmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1j} - \bar{x}_j & \cdots & x_{1p} - \bar{x}_p \\ \vdots & & \vdots & & \vdots \\ x_{i1} - \bar{x}_1 & \cdots & x_{ij} - \bar{x}_j & \cdots & x_{ip} - \bar{x}_p \\ \vdots & & \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{nj} - \bar{x}_j & \cdots & x_{np} - \bar{x}_p \end{bmatrix} \quad (57.525)$$

Remarque: La matrice des données centrées contient les coordonnées centrées (que nous noterons $x_{c_{ij}}$) des individus dans le repère $(G, \vec{e}_1, \vec{e}_2, \dots, \vec{e}_p)$. Nous nous placerons dans la suite toujours dans ce repère pour le nuage de points des individus et nous prendrons $O = G$.

Pour notre exemple, nous avons:

$$G = \bar{x} = (5.91, 3.06, 3.87) \quad (57.526)$$

et pour la matrice centrée:

$$X_c = \begin{bmatrix} -0.8067 & 0.4400 & -2.4733 \\ -1.0067 & -0.0600 & -2.4733 \\ -1.2067 & 0.1400 & -2.5733 \\ -1.3067 & 0.0400 & -2.3733 \\ -0.9067 & 0.1400 & -2.4733 \\ 1.0933 & 0.5400 & 0.8267 \\ 0.4933 & 0.1400 & 0.6267 \\ 0.9933 & 0.0400 & 1.0267 \\ -0.4067 & -0.7600 & 0.1267 \\ 0.5933 & -0.2600 & 0.7267 \\ 0.3933 & 0.2400 & 2.1267 \\ -0.1067 & -0.3600 & 1.2267 \\ 1.1933 & -0.0600 & 2.0267 \\ 0.3933 & -0.1600 & 1.7267 \\ 0.5933 & -0.0600 & 1.9267 \end{bmatrix} \quad (57.527)$$

et sous forme graphique:

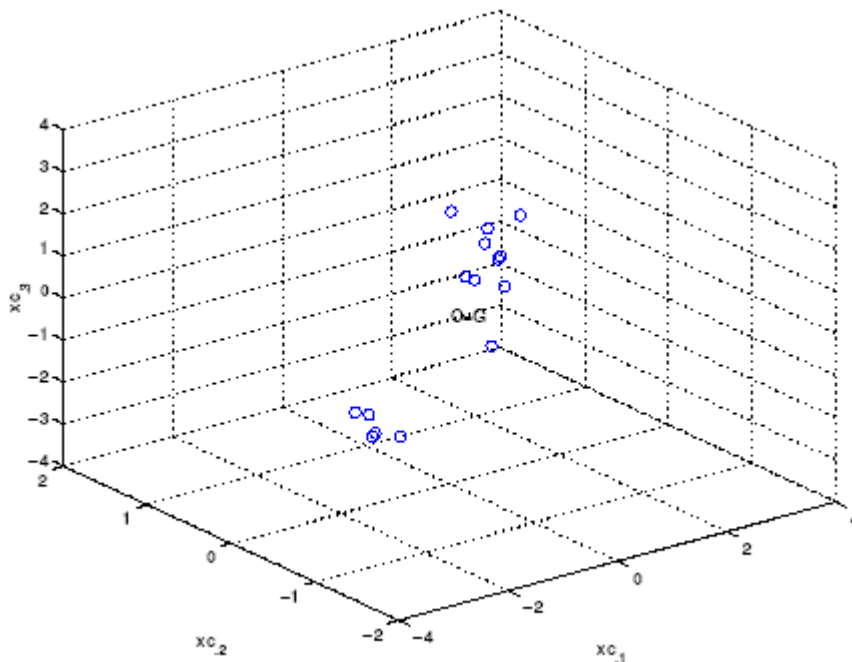


Figure: 57.44 - Points de mesures centrés

Pour donner une importance identique à chaque variable afin que le type d'unités des mesures n'influence pas l'analyse, nous travaillerons avec les données centrées réduites (cf. chapitre de Statistiques). Pour cela, nous noterons d'abord:

$$\sigma_j^2 = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 = \frac{1}{n} \sum_{i=1}^n c_{ij}^2 = \frac{1}{n} \left\| \bar{c}_j \right\|^2 \quad (57.528)$$

La variance d'échantillon de la variable centrée est donc égale à un facteur $1/n$ près à la norme de cette même variable mais centrée. La matrice des données centrées réduites (sans dimensions) est alors:

$$Y = \begin{bmatrix} (x_{11} - \bar{x}_1)/\sigma_1 & \cdots & (x_{1j} - \bar{x}_j)/\sigma_j & \cdots & (x_{1p} - \bar{x}_p)/\sigma_p \\ \vdots & & \vdots & & \vdots \\ (x_{i1} - \bar{x}_1)/\sigma_1 & \cdots & (x_{ij} - \bar{x}_j)/\sigma_j & \cdots & (x_{ip} - \bar{x}_p)/\sigma_p \\ \vdots & & \vdots & & \vdots \\ (x_{n1} - \bar{x}_1)/\sigma_1 & \cdots & (x_{nj} - \bar{x}_j)/\sigma_j & \cdots & (x_{np} - \bar{x}_p)/\sigma_p \end{bmatrix} \quad (57.529)$$

Si nous notons $D_{1/\sigma}$ la matrice diagonale suivante:

$$D_{1/\sigma} = \begin{bmatrix} 1/\sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_p \end{bmatrix} \quad (57.530)$$

Nous avons alors:

$$Y = X_e D_{1/\sigma} \quad (57.531)$$

Remarque: Chaque composante de la matrice Y est donc de moyenne nulle et de variance unitaire (ce qui revient à dire que la norme de la variable centrée réduite est unitaire comme nous allons de suite le démontrer).

Nous définissons la "**matrice des données centrées normées**" par:

$$Z = \frac{1}{\sqrt{n}} Y \quad (57.532)$$

Soit encore (il s'agit simplement de l'erreur quadratique moyenne que nous avons introduite dans le chapitre de Statistiques):

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{n}\sigma_j} \quad (57.533)$$

La terminologie vient bien évidemment du fait que la somme des carrés des composantes de chaque colonne de la matrice Z est de norme unitaire. En effet:

$$\|\bar{z}_j\|^2 = \sum_{i=1}^n z_{ij}^2 = \sum_{i=1}^n \left(\frac{x_{ij} - \bar{x}_j}{\sqrt{n}\sigma_j} \right)^2 = \frac{1}{\sigma_j^2} \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 = 1 \quad (57.534)$$

Ce qui donne:

$$Z = \begin{bmatrix} -0.2383 & 0.3572 & -0.3375 \\ -0.2974 & -0.0487 & -0.3375 \\ -0.3565 & 0.1136 & -0.3512 \\ -0.3861 & 0.0324 & -0.3239 \\ -0.2679 & 0.4384 & -0.3375 \\ 0.3230 & 0.1136 & 0.1128 \\ 0.1457 & 0.1136 & 0.0855 \\ 0.2935 & 0.0324 & 0.1401 \\ -0.1201 & -0.6170 & 0.0172 \\ 0.1753 & -0.2110 & 0.0991 \\ 0.1162 & 0.1948 & 0.2902 \\ -0.0315 & -0.2922 & 0.1674 \\ 0.3526 & -0.0487 & 0.2765 \\ 0.1162 & -0.1298 & 0.2356 \\ 0.1753 & -0.0487 & 0.2629 \end{bmatrix} \quad (57.535)$$

Nous avons graphiquement:

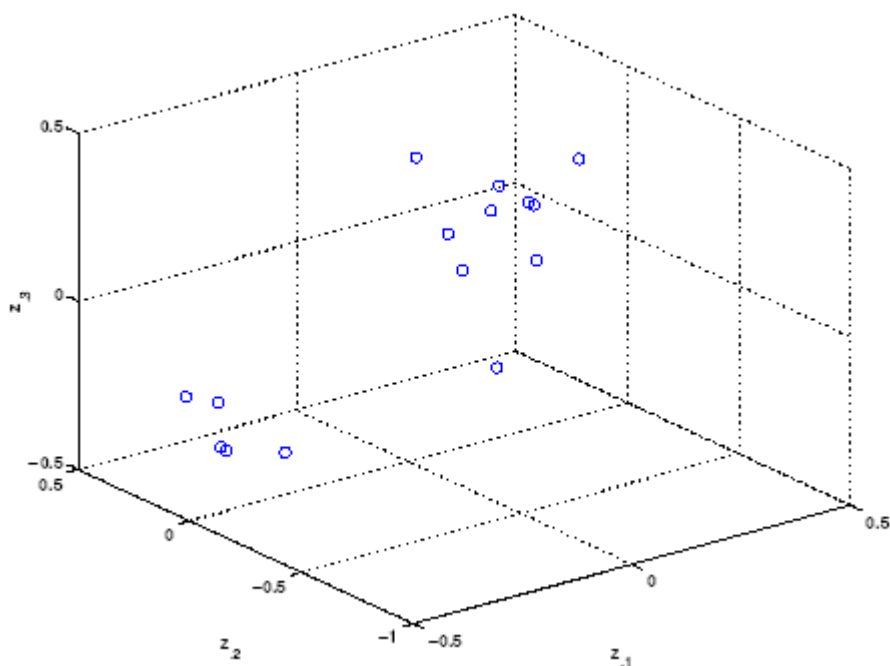


Figure: 57.45 - Points de mesures centrés et réduits

Représenter le nuage de points des données centrées réduites ou centrées normées ne modifie rien à la forme de celui-ci. En effet, la différence entre les deux n'est qu'un changement d'échelle.

L'information intéressante pour les individus est la distance entre les points! En effet plus cette distance sera grande entre deux individus z_i et $z_{i'}$, plus les deux individus seront différents et mieux on pourra les caractériser. Mais il faut d'abord choisir une distance. Nous prendrons la distance euclidienne (cf. chapitre de Topologie):

$$d^2(z_i, z_{i'}) = \|\overline{z_i z_{i'}}\|^2 = \sum_{j=1}^p (z_{ij} - z_{i'j})^2 \quad (57.536)$$

Les figures suivantes montrent les projections orthogonales dans l'espace de ce nuage de points respectivement dans les plans $(O, \vec{e}_1, \vec{e}_2)$, $(O, \vec{e}_2, \vec{e}_3)$ et enfin dans $(O, \vec{u}_1, \vec{u}_3)$ qui est la meilleure projection, appelée "**plan factoriel**" (ou parfois "**diagramme des scores**"), dans le sens où elle respecte le mieux les distances entre les individus (in extenso, elle déforme moins le nuage de points dans l'espace). L'objectif de l'A.C.P. est de déterminer ce meilleur plan et nous démontrerons comment.

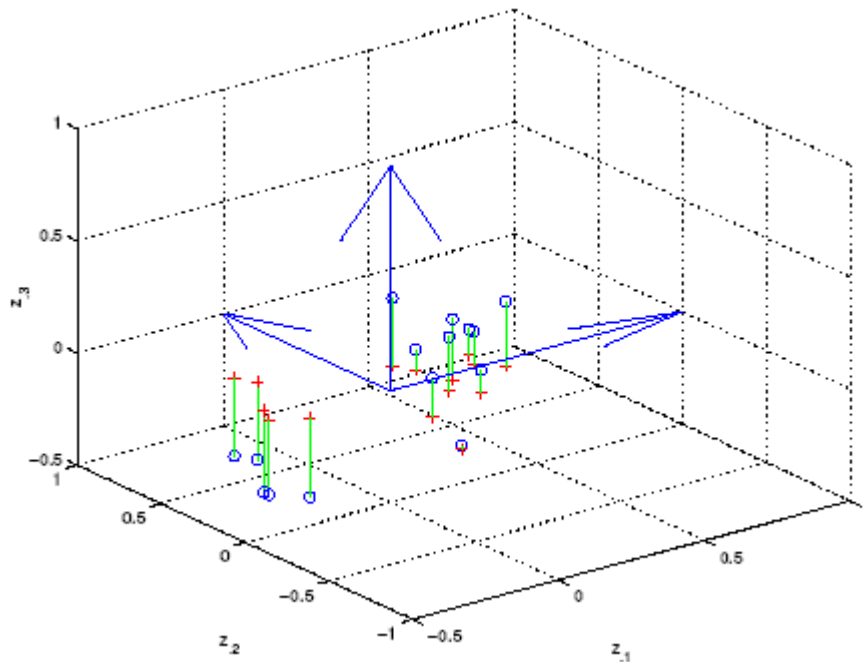


Figure: 57.46 - Projection des points sur le plan horizontal du repère centré

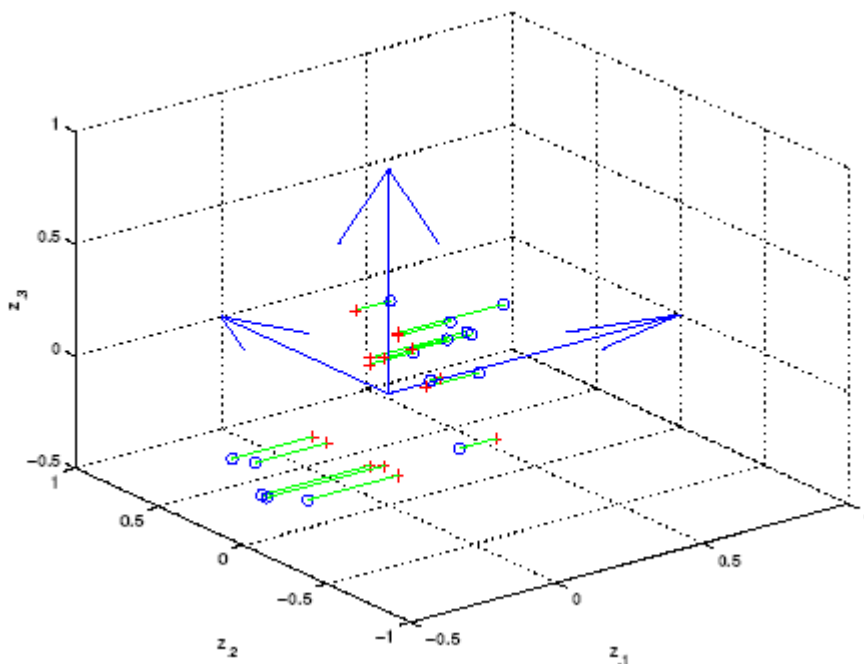


Figure: 57.47 - Projection des points sur le plan vertical du repère centré

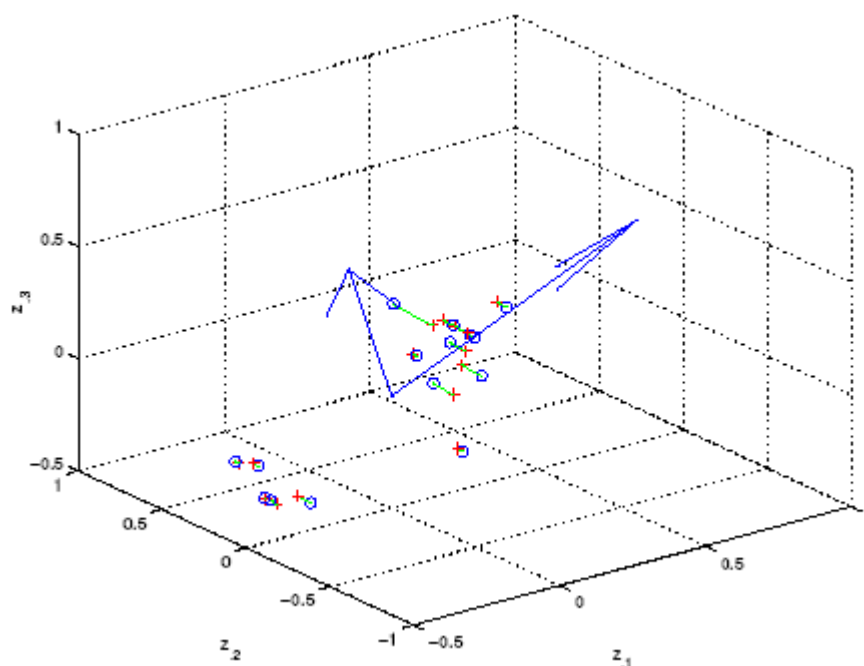


Figure: 57.48 - Projection des points sur le plan factoriel

Et la vue plane de chacune des projections:

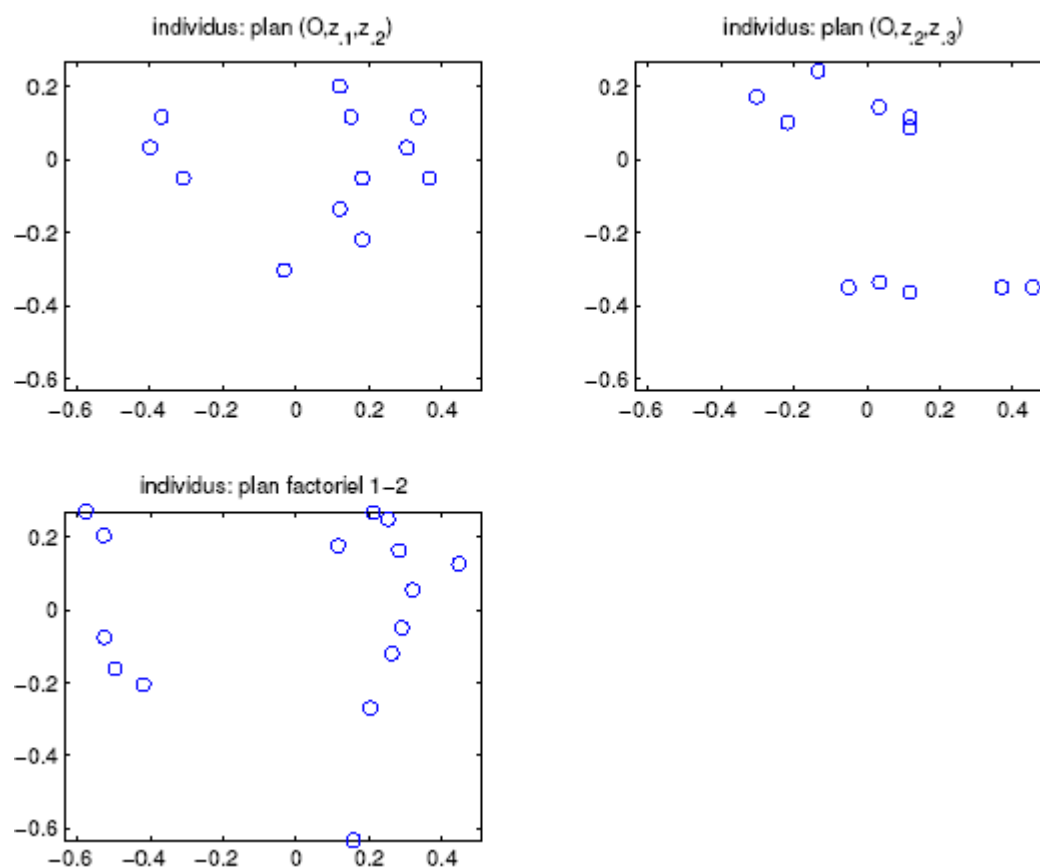


Figure: 57.49 - Vue plane de chacune des projections

Avant de déterminer le plan factoriel, nous allons maintenant chercher à détecter les liens possibles entre les variables.

Nous rappelons (cf. chapitre de Statistiques) que la covariance entre deux variables x_j et $x_{j'}$ est donnée par:

$$\begin{aligned} \text{cov}(x_j, x_{j'}) &= E[(x_j - \bar{x}_j)(x_{j'} - \bar{x}_{j'})] = \frac{1}{n} \sum_{i=1}^n (x_{ji} - \bar{x}_j)(x_{ji'} - \bar{x}_{j'}) \\ &= \frac{1}{n} \sum_{i=1}^n c_{ji} c_{ji'} \end{aligned} \quad (57.537)$$

et que le coefficient de corrélation linéaire (cf. chapitre de Statistiques) est:

$$R(x_j, x_{j'}) = \frac{\text{cov}(x_j, x_{j'})}{\sigma_j \sigma_{j'}} \quad (57.538)$$

Nous noterons par la suite:

$$c_{j,j'} = \left(\text{cov}(x_j, x_{j'}) \right) \quad \text{et} \quad r_{j,j'} = \left(R(x_j, x_{j'}) \right) \quad (57.539)$$

les matrices des covariances et de corrélations carrées (toutes deux étant pour rappel des matrices carrées et symétriques) avec $j = 1 \dots p, j' = 1 \dots p$.

Nous voyons relativement facilement que la matrice des covariances est au coefficient $1/n$ près, la matrice des produits scalaires canoniques des vecteurs de la matrice des données centrées X_c (en d'autres termes, chaque composante de la matrice des covariances est égale au produit scalaire des variables centrées). Nous en déduisons la relation suivante:

$$c_{j,j'} = \frac{1}{n} X_c^T X_c \quad (57.540)$$

La matrice des covariances-variances (puisque comme nous l'avons vu dans le chapitre de Statistiques, la diagonale contient les variances) est un outil connu d'interprétation sur ce site. Par contre, ce qui est nouveau et va nous être très utile pour déterminer le plan factoriel est la matrice des corrélations linéaires qui peut aussi être écrite sous la forme suivante:

$$R = r_{j,j'} = \frac{1}{n} Y^T Y = Z^T Z \quad (57.541)$$

Ce qui donne pour notre exemple où nous avons trois variables (très facile à calculer avec un tableur comme Microsoft Excel), la matrice carrée suivante (que les données soient centrées ou non les composantes de la matrice sont identiques):

$$R = \begin{bmatrix} 1 & -0.1609 & 0.8854 \\ -0.1609 & 1 & -0.3818 \\ 0.8854 & -0.3818 & 1 \end{bmatrix} \quad (57.542)$$

Pour continuer, toujours dans le but de déterminer le plan factoriel, définissons le concept d'inertie d'un nuage de points.

Définition: Nous appelons "**inertie d'un nuage de points**" la quantité:

$$I = \frac{1}{n} \sum_{i=1}^n d^2(G, M_i) \quad (57.543)$$

où G est le centre de gravité du nuage de points et M_i un point de $\mathbb{R}^p$ de coordonnées x_i^t .

Remarque: Le carré de la distance est pris par anticipation des développements qui vont suivre.

Ensuite, démontrons que nous avons la relation suivante:

$$\sum_{i=1}^n \sum_{i'=1}^n d^2(M_i, M_{i'}) = 2nI \quad (57.544)$$

Démonstration:

$$\begin{aligned} \sum_{i=1}^n \sum_{i'=1}^n d^2(M_i, M_{i'}) &= \sum_{i,i'} \|\overrightarrow{M_i M_{i'}}\|^2 = \sum_{i,i'} \|\overrightarrow{M_i G} + \overrightarrow{G M_{i'}}\|^2 \\ &= \sum_{i,i'} \left(\|\overrightarrow{M_i G}\|^2 + \|\overrightarrow{G M_{i'}}\|^2 + 2\overrightarrow{M_i G} \circ \overrightarrow{G M_{i'}} \right) \\ &= \sum_{i'} \left(\sum_i \|\overrightarrow{M_i G}\|^2 \right) + \sum_i \left(\sum_{i'} \|\overrightarrow{G M_{i'}}\|^2 \right) + 2 \sum_i \left(\overrightarrow{M_i G} \circ \sum_{i'} \overrightarrow{G M_{i'}} \right) \quad (57.545) \\ &= \sum_{i'} \left(\sum_i d^2(M_i, G) \right) + \sum_i \left(\sum_{i'} d^2(G, M_{i'}) \right) + 2 \sum_i (\overrightarrow{M_i G} \circ \vec{0}) \\ &= \sum_{i'} nI + \sum_i nI + 0 \\ &= 2nI \end{aligned}$$

□C.Q.F.D

Nous allons dans toute la suite travailler avec les données centrées normées, in extenso avec la matrice Z . Les points M_i auront donc ici comme coordonnées z_i^t .

Le problème est maintenant de trouver le meilleur espace affine de dimension p dans le sens où il respecte au mieux les distances entre les points. Pour cela, nous allons rechercher la meilleure droite vectorielle $\Delta_{\vec{u}}$ qui est parfaitement déterminée par le vecteur $\vec{u}$. Appelons H_i la projection orthogonale de M_i sur la droite $\Delta_{\vec{u}}$. Alors notre problème est de trouver la droite (in extenso le vecteur $\vec{u}$) qui fasse que la somme des carrés des distances entre les points H_i soit maximale. Nous écrirons le problème sous la forme d'un problème de programmation quadratique:

$$\begin{aligned} \max_{\vec{u} \in \mathbb{R}^p} \sum_{i,i'} d^2(H_i, H_{i'}) \\ \|\vec{u}\| = 1 \end{aligned} \quad (57.546)$$

Or ici, nous avons:

$$\sum_{i,i'} d^2(H_i, H_{i'}) = 2nI \quad (57.547)$$

En effet, le centre de gravité du nuage de points projetés est aussi l'origine. Par suite, notre problème peut s'écrire:

$$\begin{aligned} \max I \\ \vec{u} \in \mathbb{R}^p \\ \|\vec{u}\| = 1 \end{aligned} \quad (57.548)$$

Lui-même équivalant donc à:

$$\begin{aligned} \max \sum_i d^2(O, H_i) \\ \vec{u} \in \mathbb{R}^p \\ \|\vec{u}\| = 1 \end{aligned} \quad (57.549)$$

Résolvons donc ce problème:

Tout d'abord, puisque H_i est la projection orthogonale du point M_i sur $\Delta_{\vec{u}}$ nous avons $\overrightarrow{OH_i} = \alpha_i \vec{u}$ pour tout i avec $\alpha_i = \overrightarrow{OM_i} \circ \vec{u}$. Par suite les coordonnées des points H_i sur la droite $\Delta_{\vec{u}}$ sont:

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = Z \cdot \vec{u} \quad (57.550)$$

Par suite, nous avons:

$$\sum_i d^2(O, H_i) = \sum_i \alpha_i^2 = \|Z \cdot \vec{u}\|^2 = Z \cdot \vec{u} \circ Z \cdot \vec{u} \quad (57.551)$$

Ici nous cherchons le vecteur unitaire $\vec{u}$. La matrice Z nous est parfaitement connue. Or, nous avons:

$$(Z \cdot \vec{u}) \circ (Z \cdot \vec{u}) = (Z^T Z \vec{u}) \circ \vec{u} = (R \cdot \vec{u}) \circ \vec{u} \quad (57.552)$$

La matrice de corrélation R est symétrique donc, selon le théorème spectral vu dans le chapitre d'Algèbre Linéaire, elle est diagonalisable dans une base orthonormée de vecteurs propres (théorème d'Eckart-Young). Ainsi, nous avons démontré dans le théorème spectral que:

$$\Lambda = S^{-1}RS \quad (57.553)$$

est diagonale (libre à nous d'en choisir le contenu) si R est symétrique et S orthogonale (qui est donc dans notre exemple une matrice carrée 3×3 !). Alors nous en déduisons la relation suivante qu'il est d'usage d'appeler la "**décomposition spectrale**":

$$R = S\Lambda S^{-1} \quad (57.554)$$

et comme S avait été démontrée comme étant orthogonale (et qu'il existe une famille de vecteurs propres pour cela!), nous avons (cf. chapitre d'Algèbre Linéaire):

$$S^{-1} = S^T \quad (57.555)$$

Donc:

$$R = S\Lambda S^T \quad (57.556)$$

où nous choisissons pour Λ la matrice diagonale des valeurs propres rangées en ordre décroissant:

$$\lambda_1 \geq \lambda_2 \dots \geq \lambda_p.$$

Nous avons donc:

$$(R \cdot \vec{u}) \circ \vec{u} = (S\Lambda S^T \vec{u}) \circ \vec{u} = (\Lambda S^T \vec{u}) \circ S^T \vec{u} = (\Lambda \vec{w}) \circ \vec{w} = \sum_{j=1}^p \lambda_j w_j^2 \quad (57.557)$$

Dans la littérature, cette somme est souvent notée de la manière suivante (très souvent mentionnée dans les logiciels statistiques):

$$\sum_{j=1}^p \lambda_j w_j^2 = \begin{pmatrix} \sqrt{\lambda_1} w_1 & \sqrt{\lambda_2} w_2 & \dots & \sqrt{\lambda_p} w_p \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1} w_1 \\ \sqrt{\lambda_2} w_2 \\ \vdots \\ \sqrt{\lambda_p} w_p \end{pmatrix} \quad (57.558)$$

Mais S étant orthogonale, nous avons par conséquent:

$$\left\{ \vec{w} = S^T \vec{u} \mid \vec{u} \in \mathbb{R}^p \quad \|\vec{u}\| = 1 \right\} \Rightarrow \left\{ \vec{w} \in \mathbb{R}^p \quad \|\vec{w}\| = 1 \right\} \quad (57.559)$$

et ceci provient du fait que la matrice orthogonale est comme nous l'avions démontré dans le chapitre d'Algèbre Linéaire une isométrie (elle conserve donc la norme!).

Comme les valeurs propres sont dans l'ordre décroissant, nous écrivons:

$$\sum_{j=1}^p \lambda_j w_j^2 = \lambda_1 \left(w_1^2 + \sum_{j=2}^p \frac{\lambda_j}{\lambda_1} w_j^2 \right) \quad (57.560)$$

Or le terme entre parenthèses est strictement inférieur ou égal à 1 de par l'implication précédente. Donc:

$$\lambda_1 \left(w_1^2 + \sum_{j=2}^p \frac{\lambda_j}{\lambda_1} w_j^2 \right) \leq \lambda_1 \quad (57.561)$$

Soit:

$$\sum_{j=1}^p \lambda_j w_j^2 \leq \lambda_1 \quad (57.562)$$

Or rappelons que notre objectif est de maximiser cette inégalité. En d'autres termes de chercher w_1 tel que l'égalité soit respectée. Nous voyons assez vite qu'il en sera ainsi si $w_1 = 1$ et que les autres termes soient nuls. Ainsi, une solution triviale de notre problème de maximisation est donc:

$$\vec{w} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (57.563)$$

soit puisque:

$$\vec{w} = S^T \vec{u} = S^{-1} \vec{u} \Rightarrow \vec{u} = S \cdot \vec{w} \quad (57.564)$$

qui est alors le premier vecteur propre de R (puisque R se diagonalise dans cette base) associé à la plus grande valeur propre λ_1 . D'où le fait que cette solution soit souvent notée sous la forme:

$$\vec{u}_1 = S \cdot \vec{w}_1 \quad (57.565)$$

toujours avec $\Lambda = S^{-1} R S$ (il est donc relativement aisé de déterminer S avec des logiciels lorsque R et Λ sont connus).

Une fois que l'on a trouvé la première droite vectorielle, nous cherchons une deuxième droite dans le sous-espace vectoriel orthogonal à la droite vectorielle qui maximise l'inertie du nuage de points projetés. Nous démontrons, et devinons, que la solution est donnée par la droite vectorielle dirigée par le vecteur propre associé à la deuxième valeur propre de la matrice de corrélation et ainsi de suite...

Ainsi, nous obtenons une nouvelle base $(\vec{u}_1, \dots, \vec{u}_p)$ dont un des plans constitue le plan factoriel.

Cependant, il nous faut connaître les composantes de Z dans cette base. Comme cette base a été construite sous la condition que R y est diagonalisable via la matrice S alors cette dernière matrice est l'application linéaire qui va nous permettre d'exprimer Z dans la base $(\vec{u}_1, \dots, \vec{u}_p)$ via la relation:

$$\psi = Z \cdot S \quad (57.566)$$

Ainsi, dans notre exemple les trois valeurs propres sont de la matrice de corrélation R sont (cf. chapitre d'Algèbre Linéaire):

$$\lambda_1 = 2.0303 \quad \lambda_2 = 0.8846 \quad \lambda_3 = 0.0853 \quad (57.567)$$

et donc:

$$\Lambda = \begin{bmatrix} 2.0303 & 0 & 0 \\ 0 & 0.8846 & 0 \\ 0 & 0 & 0.0853 \end{bmatrix} \quad (57.568)$$

Remarque: Certains logiciels indiquent les poids en % respectifs et cumulés pour chacune des valeurs propres. Ainsi, nous avons dans le cas présent les poids respectifs suivants en % du total:

$$P_{\lambda 1} = \frac{2.0303}{\sum_i \lambda_i} = 66.67\% \quad P_{\lambda 2} = \frac{0.8846}{\sum_i \lambda_i} = 29.48\% \quad P_{\lambda 3} = \frac{0.0853}{\sum_i \lambda_i} = 2.84\% \quad (57.569)$$

En ayant les trois valeurs propres, pour déterminer les trois vecteurs propres $(\vec{u}_1, \vec{u}_2, \vec{u}_3)$ qui forment la base principale, il nous faut donc résoudre le système de trois équations à trois inconnues (cf. chapitre Algèbre Linéaire) suivant pour chaque valeur propre:

$$\underbrace{\begin{bmatrix} 1 - \lambda_i & -0.1609 & 0.8854 \\ -0.1609 & 1 - \lambda_i & -0.3818 \\ 0.8854 & -0.3818 & 1 - \lambda_i \end{bmatrix}}_R \underbrace{\begin{bmatrix} u_{i,1} \\ u_{i,2} \\ u_{i,3} \end{bmatrix}}_{\vec{u}_i} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad (57.570)$$

Ce qui donne donc (nous nous passerons de ce calcul élémentaire qui peut être fait à la main ou avec un simple tableur):

$$S = (\vec{u}_1, \vec{u}_2, \vec{u}_3) = \begin{bmatrix} 0.6410 & 0.3803 & -0.6666 \\ -0.3527 & 0.9174 & 0.1841 \\ 0.6816 & 0.1171 & 0.7222 \end{bmatrix} \quad (57.571)$$

qui vérifie donc:

$$R = S \Lambda S^T \quad (57.572)$$

ou autrement écrit (suite à la remarque d'un lecteur qui a voulu vérifier les calculs et qui s'est fait piéger):

$$R = S \left(\Lambda S^T \right) \quad (57.573)$$

Nous avons alors comme coordonnées des points M_i dans la base $(\vec{u}_1, \vec{u}_2, \vec{u}_3)$ en utilisant:

$$\psi = Z \cdot S \quad (57.574)$$

la matrice suivante:

$$\psi = \begin{bmatrix} -0.5268 & 0.2045 & -0.0198 \\ -0.4178 & -0.2043 & -0.0564 \\ -0.5259 & -0.0750 & 0.0052 \\ -0.4966 & -0.1604 & 0.0305 \\ -0.5760 & 0.2699 & 0.0161 \\ 0.2525 & 0.2488 & -0.1169 \\ 0.1156 & 0.1757 & -0.0150 \\ 0.2818 & 0.1634 & -0.0916 \\ 0.1578 & -0.6311 & -0.0218 \\ 0.2634 & -0.1194 & -0.0871 \\ 0.2108 & 0.2660 & 0.1739 \\ 0.2039 & -0.2697 & 0.0912 \\ 0.4469 & 0.1261 & -0.0459 \\ 0.2908 & -0.0490 & 0.0712 \\ 0.3196 & 0.0546 & 0.0663 \end{bmatrix} \quad (57.575)$$

Les coordonnées des projections du nuage de points dans le meilleur plan défini par les vecteurs $(\vec{u}_1, \vec{u}_2)$ sont donc les deux premières colonnes de la matrice précédente (correspondant donc à la longueur du sépale et la largeur du sépale).

Effectivement, nous voyons immédiatement que ce sont ces deux colonnes qui maximiseront la somme des normes dans le plan donné:

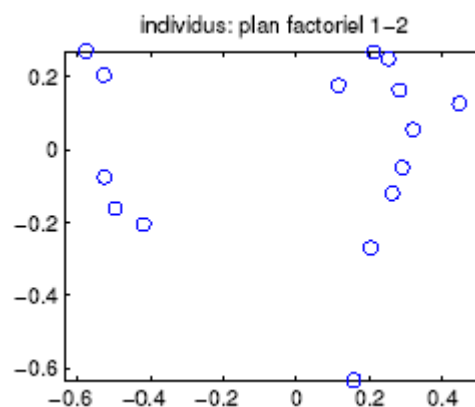


Figure: 57.50 - Plan factoriel déjà montré plus haut...

Un logiciel comme Minitab 15.1 (référence dans l'industrie de la gestion de la qualité) donne les informations suivantes pour les valeurs propres (info pas très utile sous forme graphique... à mon avis):

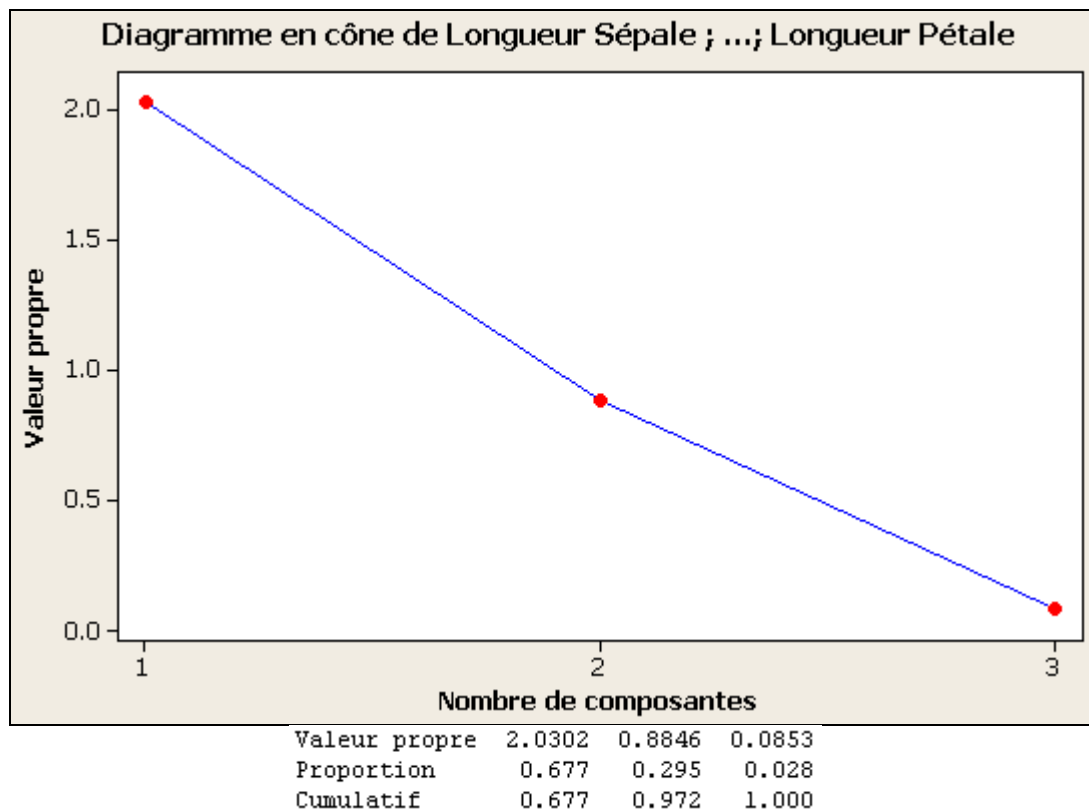


Figure: 57.51 - Valeurs propres pour l'ACP données par Minitab

et le plan factoriel suivant (resterait à savoir comment les valeurs sont calculées car elles ne sont pas identiques à celles que nous avons obtenues ici... mais la forme graphique est bien juste et c'est le principal!):

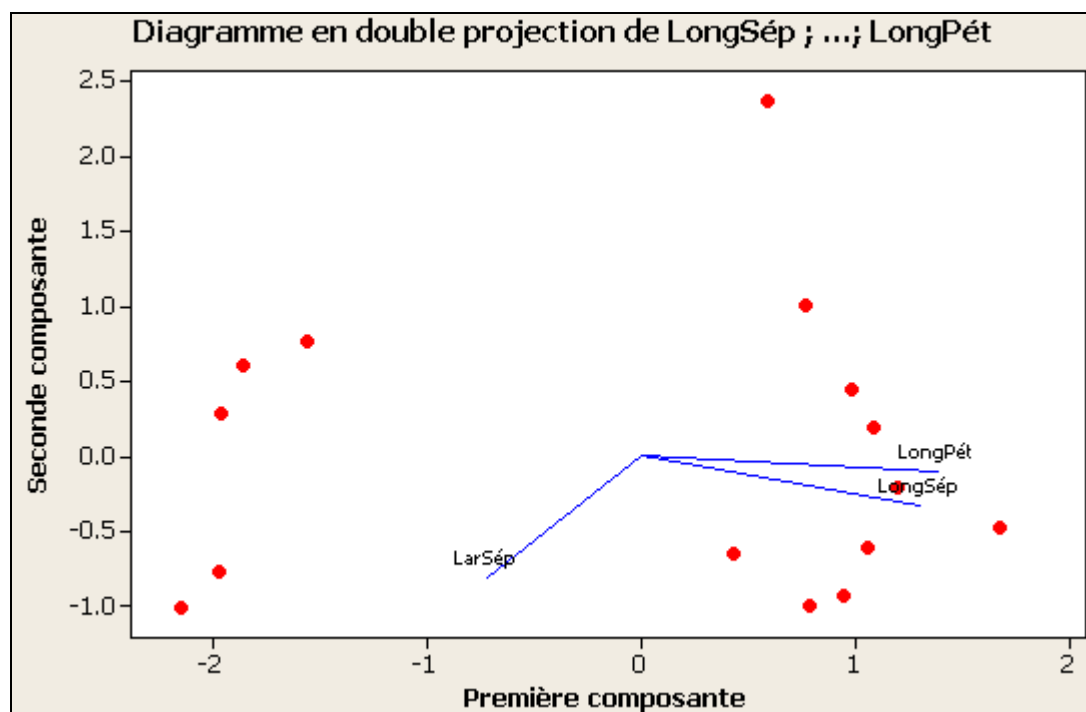


Figure: 57.52 - Plan factoriel donné par Minitab

Pour clore ce sujet, signalons que de nombreux logiciels utilisent le fait que les vecteurs $z_{.j}$ sont unitaires pour faire le produit scalaire qui correspond alors dans ce cas particulier simplement au cosinus entre les différents vecteurs tel que:

$$z_j \circ z_i = \|z_j\| \|z_i\| \cos(\widehat{z_j z_i}) = \cos(\widehat{z_j z_i}) \quad (57.576)$$

et comme nous avons démontré plus haut que:

$$R = r_{j,j'} = \frac{1}{n} Y^T Y = Z^T Z \quad (57.577)$$

Il vient alors:

$$R = r_{j,j'} = \cos(\widehat{z_j z_{j'}}) \quad (57.578)$$

et comme dans notre exemple, nous avons 3 vecteurs z_j , il y a donc 3 produits scalaires possibles si nous omettons les produits scalaires des vecteurs avec eux-mêmes. Donc la matrice:

$$R = \begin{bmatrix} 1 & -0.1609 & 0.8854 \\ -0.1609 & 1 & -0.3818 \\ 0.8854 & -0.3818 & 1 \end{bmatrix} \quad (57.579)$$

contient aussi les angles entre les vecteurs z_j .

Enfin, signalons que l'A.C.P. étant sensible aux données aberrantes, il vaut mieux parfois transformer les valeurs du tableau d'origine en leurs rangs respectifs (cf. chapitre de Statistiques) et appliquer exactement le même algorithme. Nous parlons alors d'une "**A.C.P. de rangs**".

19. ANALYSE FACTORIELLE DES CORRESPONDANCES (A.F.C.)

L'analyse factorielle des correspondances, en abrégé A.F.C., est une méthode statistique d'analyse des données (très utilisée en biostatistique et dans l'analyse de sondages). La technique de l'A.F.C. est essentiellement utilisée pour de grands tableaux de données toutes comparables entre elles (si possible exprimées toutes dans la même unité, comme une monnaie, une dimension, une fréquence ou toute autre grandeur mesurable). Elle peut en particulier permettre d'étudier des tableaux de contingence (ou tableau croisé de co-occurrence) et de décrire la liaison entre deux variables qualitatives. Elle sert à déterminer et à hiérarchiser toutes les dépendances entre les lignes et les colonnes du tableau.

Le cas de plus de deux variables qualitatives est l'analyse de correspondance multiples (A.C.M.)

Attaquons la théorie directement avec un exemple. Pour cela nous considérons le tableau (avec deux variables qualitatives) suivant des superficies des types de peuplements d'arbres en Picardie en 1984 en hectares:

	Feuillus	Résineux	Mixtes	Total par dép.
L'Aisne (A)	106'500	3'380	1'470	111'350
L'Oise (O)	101'700	10'000	0	111'700
La Somme (S)	45'200	4'350	50	49'600
Total	253'400	17'730	1'520	272'650

Tableau: 57.14 - Tableau de contingences (tableau croisé) de l'A.F.C.

Les spécialistes du domaine appellent parfois les totaux des lignes et des colonnes respectivement les "**marges de ligne**" et "**marges de colonne**". Lorsque l'ensemble du tableau est mis sous forme de pourcentage, par rapport au total des totaux, on parle de "**représentation en fréquences conjointes**":

	Feuillus	Résineux	Mixtes	Total par dép.
L'Aisne (A)	39.06%	1.24%	0.54%	40.84%
L'Oise (O)	37.30%	3.67%	0%	40.97%
La Somme (S)	16.58%	1.60%	0.02%	18.19%
Total	92.93%	6.50%	0.56%	100%

Tableau: 57.15 - Tableau des fréquences conjointes de l'A.F.C.

Nous souhaitons analyser s'il existe les degrés de ressemblance et de différence entre les variables. Remarquons, que nous ne cherchons pas à comparer l'égalité des moyennes ou des variances donc les outils statistiques vus dans le chapitre du même nom ne sont pas adaptés à ce genre d'analyse.

Si nous choisissons la distance euclidienne:

$$d^2(x_i, x_{i'}) = \|\overline{x_i x_{i'}}\|^2 = \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \quad (57.580)$$

sur les données brutes pour mesurer ces différences entre départements, nous obtenons les écarts suivants:

$$\begin{aligned} d^2(A, O) &= \sum_{j=1}^3 (x_{1j} - x_{2j})^2 = (106'500 - 101'700)^2 + (3'380 - 10'000)^2 + (1'470 - 0)^2 \\ &= 69'025'300 \Rightarrow d(A, O) = \sqrt{69'025'300} = 8'308.1 [ha] \end{aligned} \quad (57.581)$$

et ainsi de suite pour les autres régions. Nous obtenons alors:

$$\begin{aligned} d(A, O) &= 8'308.1 [ha] \\ d(A, S) &= 61'324.1 [ha] \\ d(O, S) &= 56'781.8 [ha] \end{aligned} \quad (57.582)$$

Nous voyons en regardant le tableau et avant tout calcul que les départements de l'Aisne et l'Oise se ressemblent alors que le département de la Somme se diffère nettement. Les distances obtenues mettent en évidence cette observation.

Mais dans le tableau ci-dessus les profils de l'Oise et de la Somme, avec une forêt mixte très faible, sont pourtant très proches en proportion.

Dans ce contexte, nous voyons que la distance euclidienne transcrit les différences de masse entre les départements. En d'autres termes, l'Aisne et l'Oise se ressemblent car leurs superficies sont proches. Pour éliminer l'artefact lié aux ordres de grandeur, il nous faut transformer les données en pourcentage (pourcentages des régions). Nous obtenons alors:

	Feuillus	Résineux	Mixtes	%Région
Aisne	95.6	3.0	1.3	40.8
Oise	91.0	9.0	0.0	41.0
Somme	91.1	8.8	0.1	18.2

Tableau: 57.16 - Transformation du tableau de contingences en pourcents

où les spécialistes du domaine appellent parfois la colonne en pourcents des régions "**profil marginal des lignes**" ou "**masse**" (et respectivement quand ils indiquent la ligne des pourcents des arbres).

Si nous choisissons la distance euclidienne sur les proportions (données relatives), nous obtenons:

$$d^2(p_i, p_{i'}) = \left\| \overrightarrow{p_i p_{i'}} \right\|^2 = \sum_{j=1}^p \left(\frac{x_{ij}}{x_{i.}} - \frac{x_{i'j}}{x_{i'.}} \right)^2 = \sum_{j=1}^p \left(\frac{x_{ij}}{\sum_j x_{ij}} - \frac{x_{i'j}}{\sum_j x_{i'j}} \right)^2 \quad (57.583)$$

soit:

$$\begin{aligned} d(A, O) &= 7.6\% \\ d(A, S) &= 7.4\% \\ d(O, S) &= 0.2\% \end{aligned} \quad (57.584)$$

Cette fois, l'Oise et la Somme apparaissent bien comme se ressemblant le plus avec leurs forêts. Nous voyons que travailler avec les données relatives semble donc plus pertinent dans ce cas!

Maintenant, nous allons emprunter une idée aux économistes qui, lorsqu'ils ont des tableaux du même genre que le précédent, calculent ce qu'ils appellent "**l'index**" ou "**élasticité**" (souvent appelé "**indice de spécificité**" en statistiques et qui est donné par le rapports entre la fréquence conjointe et la fréquence marginale:

$$I = \frac{\frac{x_{ij}}{x_{i.}}}{\frac{x_{.j}}{x_{..}}} = \frac{\frac{x_{ij}}{x_{.j}}}{\frac{x_{i.}}{x_{..}}} \quad (57.585)$$

Voici un exemple obtenu avec les tableaux croisés dynamiques de Microsoft Excel 11.8346 qui inclut la fonction Index. D'abord le tableau de départ:

Sum	Region			
Item	Alberta	Ontario	Quebec	Total
Binder	1'279.36	5762.63	2'535.66	9'577.65
Desk	825.00	875.00		1700.00
Pen	151.24	539.73	1'354.25	2'045.22
Pen Set		2'421.39	1748.48	4'169.87
Pencil	231.12	1'540.32	363.70	2'135.14
Total	2'486.72	11'139.07	6'002.09	19'627.88

Figure: 57.53 - Tableau croisé dynamique Microsoft Excel 11.8346 de départ

et en activant la fonction Index:

Index	Region		
Item	Alberta	Ontario	Quebec
Binder	1.05	1.06	0.87
Desk	3.83	0.91	-
Pen	0.58	0.47	2.17
Pen Set	-	1.02	1.37
Pencil	0.85	1.27	0.56

Figure: 57.54 - Tableau croisé dynamique Microsoft Excel 11.8346 avec la fonction Index

Pour voir d'où viennent ces valeurs, regardons par exemple l'article *Desk* dans la région *Alberta*. Il a un rendement (fréquence conjointe) de:

$$\frac{x_{21}}{x_{2.}} = \frac{825}{1700} \cong 48.52\% \quad (57.586)$$

par rapport à toutes les régions ce qui est au-dessus de la valeur de 33.33% qu'aurait comme rendement cet article toutes régions confondues s'il n'y avait pas de préférences de région!

La région *Alberta* a elle un rendement (fréquence marginale) de:

$$\frac{x_{.1}}{x_{..}} = \frac{2486.72}{19627.88} \cong 12.66\% \quad (57.587)$$

par rapport à toutes les régions ce qui est en-dessous des 33.33% de rendement qu'elle aurait s'il n'y avait pas de préférences de région. Ainsi, ce tableau d'index permet de savoir si les différences sont qualitativement significatives!!

Le rapport donne donc:

$$I = \frac{\frac{x_{21}}{x_{2.}}}{\frac{x_{.1}}{x_{..}}} \cong 3.83 \quad (57.588)$$

ce qui montre un fort décalage entre la valeur obtenue et la valeur que nous aurions si les proportions étaient respectées (surreprésentation de 283%).

C'est donc une sorte de calcul de conformité: si le rapport valait 1, c'est que le rendement régional des ventes de cet article particulier serait conforme par rapport à toutes les ventes de cette région relativement à un marché national. Il n'y aurait alors pas d'anomalies. Voyons cela par exemple pour nos arbres où nous avons les effectifs observés:

	Feuillus	Résineux	Mixtes	Total par dép.
L'Aisne (A)	106'500	3'380	1'470	111'350
L'Oise (O)	101'700	10'000	0	111'700
La Somme (S)	45'200	4'350	50	49'600
Total	253'400	17'730	1'520	272'650

Tableau: 57.17 - Tableau de contingences (tableau croisé) de l'A.F.C.

et pour lequel nous obtenons le tableau croisé dynamique des index effectifs observés suivant dans Microsoft Excel 11.8346 :

Index	Arbre		
Région	Feuillus	Résineux	Mixtes
La Somme (S)	0.98	1.35	0.18
L'Aisne (A)	1.03	0.47	2.37
L'Oise (O)	0.98	1.38	-

Figure: 57.55 - Tableau croisé dynamique Microsoft Excel 11.8346 avec la fonction Index

et nous voyons encore clairement à l'aide de ce tableau que ce sont l'Oise et la Somme qui se ressemblent le plus!

Avant de continuer, nous pourrions nous poser la question extrêmement importante suivante: Quels seraient les effectifs théoriques qui auraient été obtenus si les proportions des arbres dans les régions étaient rigoureusement équivalentes à la proportion d'ensemble (soit de telle manière à ce que les index soient tous unitaires)?

Eh bien simplement en faisant le calcul suivant (il s'agit simplement d'une règle de trois calculée dans chaque cellule) dont il faut bien - si possible - comprendre le sens sans l'appliquer bêtement:

	Feuillus	Résineux	Mixtes
Aisne	$= (253'400 / 272'650) * 111'350$ =103'488	$= (17'730 / 272'650) * 111'350$ =7'241	$= (1'520 / 272'650) * 111'350$ =621
Oise	$= (253'400 / 272'650) * 111'700$ =103'813	$= (17'730 / 272'650) * 111'700$ =7'264	$= (1'520 / 272'650) * 111'700$ =623
Somme	$= (253'400 / 272'650) * 49'600$ =46'098	$= (17'730 / 272'650) * 49'600$ =3'225	$= (1'520 / 272'650) * 49'600$ =276
Total	253'400	17'730	1'520

Tableau: 57.18 - Respect des proportions de l'A.F.C.

Et nous obtenons avec ces nouvelles valeurs le tableau des index des effectifs théoriques suivant dans Microsoft Excel 11.8346:

Index	Arbre		
Région	Feuillus	Résineux	Mixtes
La Somme (S)	1.00	1.00	1.00
L'Aisne (A)	1.00	1.00	1.00
L'Oise (O)	1.00	1.00	1.00

Figure: 57.56 - Tableau croisé dynamique Microsoft Excel 11.8346 de l'index des effectifs théoriques

ce qui montre que les proportions sont maintenant respectées! Parenthèse fermée (mais sur laquelle nous reviendrons un peu plus loin)!

Eh bien quand nous voulons faire de l'analyse factorielle des correspondances, notre relation:

$$d^2(x_i, x_{i'}) = \sum_{j=1}^p \left(\frac{x_{ij}}{x_{i.}} - \frac{x_{i'j}}{x_{i'.}} \right)^2 \quad (57.589)$$

devient alors:

$$d^2(x_i, x_{i'}) = \sum_{j=1}^p \frac{\left(\frac{x_{ij}}{x_{i.}} - \frac{x_{i'j}}{x_{i'.}} \right)^2}{\frac{x_{.j}}{x_{..}}} \quad (57.590)$$

soit:

$$\begin{aligned} d(A, O) &= 0.30 \\ d(A, S) &= 0.39 \\ d(O, S) &= 0.02 \end{aligned} \quad (57.591)$$

Cette fois encore, l'Oise et la Somme apparaissent bien comme se ressemblant le plus.

La distance ci-dessus se nomme la "**métrique du Khi-deux**" car elle ressemble (mais c'est tout!) à la distance utilisée dans le test d'ajustement du même nom (cf. chapitre de Statistiques) mais ici, elle permet seulement de mettre en place une hiérarchie dans le cadre d'un tableau de contingences et d'observer les variables similaires de manière plus aisée!!

Remarque: Il existe une autre manière de calculer une AFC en se basant sur une distance euclidienne mais en ayant pris soin au préalable de transformer les données du tableau de contingences de manière particulière et ce pour que le calcul soit identique qu'en lorsqu'on utilise la métrique du Khi-deux.

20. TEST D'INDÉPENDANCE DU KHI-2

Pendant l'introduction de la méthode précédente permettant de comparer des effectifs (valeurs) et détecter lesquels étaient les plus proches, nous avons donné le tableau des effectifs observés:

	Feuillus	Résineux	Mixtes	Total par dép.
L'Aisne (A)	106'500	3'380	1'470	111'350
L'Oise (O)	101'700	10'000	0	111'700
La Somme (S)	45'200	4'350	50	49'600
Total	253'400	17'730	1'520	272'650

Tableau: 57.19 - Tableau de contingences de l'A.F.C.

et nous avons montré comment trouver le tableau des effectifs théoriques (arrondis à l'entier le plus proche) dans les cas où les proportions auraient dû éventuellement être respectées:

	Feuillus	Résineux	Mixtes	Total par dép.
L'Aisne (A)	103'488	7'241	621	111'350
L'Oise (O)	103'813	7'264	623	111'700
La Somme (S)	46'098	3'225	276	49'600
Total	253'400	17'730	1'520	272'650

Tableau: 57.20 - Tableau de contingences avec effectifs théoriques

La construction du dernier tableau ci-dessus présuppose par exemple que les trois régions sont dans des conditions identiques pour tout ce qui concerne la croissance et la multiplication des arbres et que le nombre d'arbres est en relation de cause à effet directe!!!! avec les régions et qu'il n'y a pas d'autres causes intermédiaires... ce qui est une hypothèse forte!

Mais sous cette hypothèse, supposons que nous souhaiterions savoir si les différences observées entre le nombre d'arbres et les régions sont statistiquement significatives ou purement aléatoires à cause de l'échantillon expérimental? En d'autres termes, nous voulons savoir si le nombre d'arbres dépend réellement des régions dans lesquelles ils poussent ou si ces valeurs ne sont que dues au hasard de l'échantillon? Raison pour laquelle ce test s'appelle le "**test d'indépendance du Khi-deux**".

Remarque: Le test d'indépendance du Khi-deux est recommandé en analyse sensorielle par la norme ISO 8588-1987 sous le nom de "test A-Non A".

Pour répondre à cette question il faut d'abord une référence. Et cette référence est justement l'hypothèse de lien causal direct (proportions respectées) que nous avons donnée juste précédemment.

Si nous considérons que chaque case du tableau des effectifs observés correspond à l'issue d'une variable aléatoire de loi inconnue et que chaque case du tableau théorique (du moins la classe d'effectifs) est considérée comme issue d'une variable aléatoire suivant une loi binomiale (et asymptotique d'une loi Normale) alors nous pouvons utiliser le test d'ajustement du Khi-deux:

$$D = \sum_{i=1}^N E_i^2 = \sum_{i=1}^N \frac{(k_i - np_i)^2}{np_i} \quad (57.592)$$

(cf. chapitre de Statistiques) pour avoir une bonne idée (mais qui reste quand même très approximative au vu des hypothèses!) si les différences entre les valeurs des effectifs observés sont dues au hasard ou sont réelles. Or, si D est petit, la probabilité que ce soit dû au hasard est grande mais si D est grand alors nous avons une différence réelle (donc nous utilisons le test d'ajustement du Khi-deux mais dans le sens inverse!).

Reste à déterminer le nombre de degrés de liberté de la loi χ^2 que suit cette somme dans ce type de configuration!

Dans le cas particulier (mais facilement généralisable par récurrence) d'une table à deux entrées avec deux variables catégorisées X avec l niveaux et Y avec c niveaux nous aurons respectivement l lignes et c colonnes.

Ainsi, la table aura bien évidemment $l \cdot c$ cellules. Dans la table des effectifs théoriques (dont chaque cellule est considérée comme une variable aléatoire) chaque cellule sera entièrement déterminée par la somme des autres de telle sorte que le nombre de degrés de liberté sera alors en toute logique comme nous l'avons vu dans le chapitre de Statistiques:

$$N - 1 = l \cdot c - 1 \quad (57.593)$$

Ainsi, en prenant notre exemple des forêts, c'est le total des totaux de 272'650 qui nous permet d'écrire cette dernière relation et ainsi de déterminer la valeur d'une cellule éventuellement vide, toutes les autres étant données!

Un test du Khi-deux sur ce type de table teste l'hypothèse d'indépendance contre l'hypothèse alternative de dépendance. Sous l'hypothèse d'indépendance nous estimons qu'il y a besoin de seulement:

$$(l - 1) + (c - 1) \quad (57.594)$$

valeurs sur les N pour pouvoir en déterminer la totalité (en supposant implicitement connues les sommes par ligne et par colonne).

Ainsi, si vous avez une table de 2 lignes par 2 colonnes, il vous suffit si vous connaissez les totaux des lignes et des colonnes, d'avoir 2 valeurs (soit $(2-1)+(2-1)$) pour déterminer les 2 manquantes. Le raisonnement s'applique aussi pour une table de 3 lignes par 3 colonnes où il vous suffit d'avoir au moins 4 valeurs (soit $(3-1)+(3-1)$) pour déterminer les 5 manquantes.

Les degrés de liberté pour le Khi-deux sont alors:

$$ddl = (l \cdot c - 1) - [(l - 1) + (c - 1)] = l \cdot c - 1 - l + 1 - c + 1 = l \cdot c - l - c + 1 = (l - 1)(c - 1) \quad (57.595)$$

C'est cette relation qui nous dit (trivialement!) que si dans un tableau de 2 lignes par 2 colonnes comprenant donc 4 cellules (totaux des lignes et colonnes étant aussi connus!) qu'étant donnée une seule des valeurs (ddl valant 1), nous pouvons déterminer les 3 autres valeurs manquantes.

Voici donc une définition possible du nombre de degrés de liberté: C'est le nombre maximum de valeurs du modèle telles qu'aucune d'entre elles n'est calculable à partir des autres.

De même, pour un tableau de 3 lignes par 3 colonnes comprenant 9 cellules comme c'est le cas de notre exemple dans ce chapitre avec les forêts, la connaissance de 4 cellules seules permet grâce aux totaux en ligne et colonnes de déterminer les 5 autres qui seraient éventuellement non connues.

D'où la relation dans le cadre de l'application du Khi-deux de la relation finale:

$$\chi^2 ((l - 1)(c - 1)) = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (57.596)$$

en faisant usage des notations utilisées dans l'industrie. Le terme:

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (57.597)$$

est souvent appelé "**carré du résidu standardisé**". Et le rapport:

$$\frac{\frac{(O_{ij} - E_{ij})^2}{E_{ij}}}{\sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}} \quad (57.598)$$

est souvent appelé "**contribution au Khi-deux d'indépendance**".

Remarque: Pour utiliser ce test à bon escient, il faudrait donc vérifier d'abord que les différences (numérateur) suivent une loi Normale ou que l'ensemble des termes de la somme forment une loi du Khi-deux ou approximativement (asymptotiquement) une loi Normale centrée réduite.

Dans le cadre de notre exemple, nous avons:

$$D = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^3 \sum_{j=1}^3 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \cong 5'602.35 \quad (57.599)$$

et la p -value de cette valeur avec la loi du Khi-deux à quatre degrés de liberté:

$$\chi^2((l-1)(c-1)) = \chi^2(4) \quad (57.600)$$

est tellement proche de zéro (non statistiquement significatif) que nous n'avons aucune chance de nous tromper en affirmant que les différences observées dans le tableau sont statistiquement significatives entre les 3 forêts et donc qu'il y a très probablement indépendance.

Nous obtenons un résultat similaire entre l'Oise et la Somme alors qu'avec l'AFC nous avons vu que ces deux forêts se ressemblaient beaucoup.

Remarque: Dans la pratique il est souvent d'usage de prendre le p -value à 5% pour considérer la probabilité attachée aux écarts observés comme statistiquement significative ou non significative.

20.1. V DE CRAMER

Nous avons vu plus haut que le test d'indépendance du Khi-deux peut être utilisé pour mesurer le degré d'association de deux variables catégorielles dans une table de contingences de l lignes et c colonnes:

$$D = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (57.601)$$

et que cette distance suit une loi du Khi-deux à $(l-1)(c-1)$ degrés de liberté. Nous allons démontrer de façon intuitive que la valeur maximum de la distance D est:

$$D_{\max} = N((\min(l, c) - 1)) \quad (57.602)$$

et que cette valeur maximale n'est atteinte que si et seulement si chaque ligne ou chaque colonne contient une et une seule valeur non nulle. Sous cette dernière condition, nous pouvons toujours réarranger le tableau de contingences de façon à avoir tous les termes non nuls sur la diagonale du tableau.

Évidemment, si le tableau n'est pas carré comme ci-dessous:

Ligne/Colonne	c1	c2	...	c4	Total
l1	a_1	0	0	0	a_1
l2	0	a_2	0	0	a_2
l3	0	0	a_3	0	a_3
l4	0	0	0	a_4	a_4
l5	0	0	0	0	0
Total	a_1	a_2	a_3	a_4	N

Tableau: 57.21 - Tableau de contingences rectangulaire

le cas qui maximise D impose donc des termes en diagonale sur la dimension la plus petite en ligne ou colonne notée traditionnellement q . Dans le cas présent, nous avons donc:

$$q = \min(l, c) = c = 4 \quad (57.603)$$

Dans ce cas extrême et bien évidemment théorique, les lignes ou colonnes qui n'ont que des zéros peuvent être mises de côté et dès lors tout tableau peut se résumer à:

Ligne/Colonne	c1	c2	...	cq	Total
l1	a_1	0	...	0	a_1
l2	0	a_2	...	0	a_2
l3	0	0	...	0	a_3
...	...	...	...	...	...
lq	0	0	...	a_q	a_q
Total	a_1	a_2	...	a_q	N

Tableau: 57.22 - Tableau de contingences rectangulaire transformé en carré

Évidemment, faire abstraction des lignes ou colonnes qui n'ont que des valeurs nulles suppose que dans la distance D du Khi-deux nous posions que:

$$\frac{(0-0)^2}{0} = 0 \quad (57.604)$$

ce qui est tout à fait discutable... Pour la suite, nous allons avoir besoin des relations suivantes:

$$N = \sum_{i=1}^q a_i \quad (57.605)$$

et:

$$N^2 = \left(\sum_{i=1}^q a_i \right)^2 = \sum_{i=1}^q a_i \sum_{j=1}^q a_j = \sum_{i=1}^q \sum_{j=1}^q a_i a_j = \sum_{i=1}^q a_i^2 + \sum_{i \neq j}^q a_i a_j \quad (57.606)$$

Dès lors, il vient:

$$\begin{aligned}
D_{\max} &= \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_i \frac{\left(a_i - \frac{a_i^2}{N}\right)^2}{\frac{a_i^2}{N}} + \sum_{i \neq j} \frac{\left(0 - \frac{a_i a_j}{N}\right)^2}{\frac{a_i a_j}{N}} = \sum_i \frac{\left(\frac{Na_i - a_i^2}{N}\right)^2}{\frac{a_i^2}{N}} + \sum_{i \neq j} \frac{a_i a_j}{N} \\
&= \sum_i \frac{\left(a_i (N - a_i)\right)^2}{a_i^2 N} + \frac{1}{N} \left(\sum_{i \neq j} a_i a_j + \sum_i a_i^2 \right) - \frac{1}{N} \sum_i a_i^2 \\
&= \sum_i \frac{(N - a_i)^2}{N} + \frac{1}{N} N^2 - \frac{1}{N} \sum_i a_i^2 = \frac{1}{N} \sum_i (N - a_i)^2 - \frac{1}{N} \sum_i a_i^2 + \frac{1}{N} N^2 \\
&= \frac{1}{N} \left[\sum_i (N - a_i)^2 - \sum_i a_i^2 \right] + N = \frac{1}{N} \sum_i \left((N - a_i)^2 - a_i^2 \right) + N = \sum_i \frac{\left((N - a_i)^2 - a_i^2 \right)}{N} + N \\
&= \sum_i \frac{(N^2 - 2Na_i + a_i^2 - a_i^2)}{N} + N = \sum_i \frac{(N^2 - 2Na_i)}{N} + N = \sum_i (N - 2a_i) + N \\
&= qN - 2 \sum_i a_i + N = qN - 2N + N = N(q - 1) = N(\min(l, c) - 1)
\end{aligned}$$

(57.607)

Nous définissons alors la valeur suivante:

$$V = \sqrt{V^2} = \sqrt{\frac{\chi^2 (l-1)(c-1)}{N(q-1)}} \quad (57.608)$$

comme étant le "**coefficient V de Cramer**" (la majorité des logiciels donnent cependant la valeur de V au carré). Ce dernier est tel qu'il ne dépasse jamais 1 et permet une interprétation plus intuitive du degré d'association dans un tableau de contingences.

Dans le cas où le tableau est de dimension 2 en ligne et 2 en colonne, la relation précédente se réduit alors immédiatement à:

$$V = \sqrt{\frac{\chi^2}{N}} \quad (57.609)$$

Relation qu'il est de tradition de noter dans ce cas particulier sous la forme suivante:

$$\phi = \sqrt{\frac{\chi^2}{N}} \quad (57.610)$$

et de nommer "**phi de Cramer**".

 Exemple:

Considérons le tableau suivant (même si les conditions ne sont pas satisfaites pour un test d'indépendance du Khi-deux):

Projets	Chefs de projet certifiés	Chefs de projet non certifiés	Total
Délais respectés	8	1	9
Délais non respectés	4	5	9
Total	12	6	18

Tableau: 57.23 - Un exemple de tableau de contingences pour le calcul du V de Cramer

avec les effectifs théoriques:

Projets	Chefs de projet certifiés	Chefs de projet non certifiés	Total
Délais respectés	(12/18)*9=6	(6/18)*9=3	9
Délais non respectés	(12/18)*9=6	(6/18)*9=3	9
Total	12	6	18

Tableau: 57.24 - Tableau de contingences de l'exemple avec effectifs théoriques

Nous avons alors:

$$\begin{aligned}
 D &= \chi^2 (l-1)(c-1) = \chi^2 (1) = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\
 &= \frac{(8-6)^2}{6} + \frac{(4-6)^2}{6} + \frac{(1-3)^2}{3} + \frac{(5-3)^2}{3} \\
 &= 0.\bar{6} + 0.\bar{6} + 1.\bar{3} + 1.\bar{3} = 4
 \end{aligned} \quad (57.611)$$

Et à un niveau de 95%, nous obtenons avec la version française de Microsoft Excel 14.0.6123:

$$\chi_{95\%}^2(1) = \text{LOI.KHIDEUX.INVERSE}(95\%; 1) \cong 3.841 \quad (57.612)$$

ou avec la p -value:

$$= 1 - \text{LOI.KHIDEUX.N}(4; 1; 1) = 0.045 \quad (57.613)$$

Il vient alors immédiatement:

$$V = \sqrt{V^2} = \sqrt{\frac{\chi^2 (l-1)(c-1)}{N(q-1)}} = \sqrt{\frac{4}{18(2-1)}} = \sqrt{0.\bar{2}} \cong 0.4714 \quad (57.614)$$

Nous sommes un peu limite ici... étant donné que la p -value est toute proche du seuil de 0.05. Cependant, prendre une décision dans le cas présent alors que de toute façon les effectifs sont si faibles reviendrait à conclure n'importe quoi et ce d'autant plus que l'outil est construit sur une cumulation d'approximations.

21. TEST EXACT DE FISHER

Lorsque les effectifs dans la table de contingences sont trop petits ou que les valeurs sont vraiment trop irrégulières, l'utilisation du Khi-deux (test de Pearson) n'est plus possible car les conditions d'application ne sont plus valables. Nous allons voir que le test exact de Fisher peut être formalisé analytiquement dans des tableaux de contingences de 2 lignes par 2 colonnes (la majorité des logiciels de statistiques ne gèrent que ce scénario particulier pour le test exact de Fisher) sinon quoi il faut recourir à des simulations de Monte-Carlo.

Le principe du "**test exact de Fisher**" (utilisable aussi bien en unilatéral qu'en bilatéral même si ce dernier est largement plus courant dans la pratique), basé donc sur la fréquence de croisement, est de déterminer si la configuration observée dans le tableau de contingence est une situation extrême par rapport aux situations possibles. Comme nous allons le démontrer, ce test a pour propriété particulière que n'importe quelle case du tableau peut servir de référence pour le test car les distributions sous-jacentes (lois marginales) de probabilité sont équivalentes.

Pour étudier ce test, comme souvent dans ce chapitre, passons directement par un exemple. Considérons le tableau de contingence suivant (qui maintenant nous est connu...):

Projets	Chefs de projet certifiés	Chefs de projet non certifiés	Total
Délais respectés	8	1	9
Délais non respectés	4	5	9
Total	12	6	18

Tableau: 57.25 - Tableau de contingences de départ pour l'étude du test exact de Fisher

qui est donc en réalité pas adapté pour un test d'indépendance du Khi-deux puisque le contenu des cellules est inférieur à 10 unités et le nombre de degrés de liberté serait lui égal à l'unité.

Ce même tableau en pourcentages donnera (même si c'est inutile pour le test que nous étudions il arrive souvent que les logiciels de statistiques communiquent ces valeurs):

Projets	Chefs de projet certifiés	Chefs de projet non certifiés	Total
Délais respectés	88.88%	11.12%	50%
Délais non respectés	44.44%	55.56%	50%
Total	66.66%	33.34%	100%

Tableau: 57.26 - Même tableau avec les pourcentages marginaux

Les effectifs théoriques sont donnés par (même si c'est aussi inutile pour le test que nous étudions il arrive souvent que les logiciels de statistiques communiquent ces valeurs):

Projets	Chefs de projet certifiés	Chefs de projet non certifiés	Total
Délais respectés	$(12/18)*9=6$	$(6/18)*9=3$	9
Délais non respectés	$(12/18)*9=6$	$(6/18)*9=3$	9
Total	12	6	18

Tableau: 57.27 - Toujours le même tableau mais avec les effectifs théoriques (règle de trois)

La question que nous allons commencer par nous poser est la suivante: connaissant les totaux pour chaque ligne et chaque colonne, quelle est la probabilité d'avoir les valeurs présentes actuellement dans chacune des cases!

Cette question peut être formalisée si nous changeons le tableau sous la forme générique suivante:

Projets	Chefs de projet certifiés	Chefs de projet non certifiés	Total
Délais respectés	$a=k$	b	$a+b=m$
Délais non respectés	c	d	$c+d$
Total	$a+c=p$	$b+d$	$a+b+c+d=n$

Tableau: 57.28 - Représentation algébrique du contenu du tableau

Explicitement et relativement à notre exemple en adoptant la notation d'usage de la loi hypergéométrique, la question est de savoir quelle était la probabilité d'avoir 8 ($a=k$) projets parmi les 18 (n) dont les délais ont été respectés par des chefs de projets certifiés sachant qu'il y a 9 projets (m) au total dont les délais ont été respectés et 12 projets (p) au total gérés par des chefs de projets certifiés.

Nous avons alors vu dans le chapitre de Statistiques que dans ce cas il s'agit d'un tirage exhaustif, il faut donc utiliser la loi hypergéométrique donnée par:

$$P(n, p, m, k) = H(k) = \frac{C_k^m C_{p-k}^{n-m}}{C_p^n} = \frac{\frac{m!}{k!(m-k)!} \frac{(n-m)!}{(p-k)!((n-m)-(p-k))!}}{\frac{n!}{p!(n-p)!}} \quad (57.615)$$

Soit avec la version française de Microsoft Excel 11.8346:

$$\begin{aligned} &= \text{LOI.HYPERGEOMETRIQUE}(k;p;m;n) \\ &= \text{LOI.HYPERGEOMETRIQUE}(8;12;9;18)=0.06108597 \end{aligned}$$

où pour rappel (cf. chapitre de Statistiques), k est le nombre de succès dans l'échantillon, p est la taille de l'échantillon, m le nombre de succès dans la population et n la taille de la population.

Au fait les probabilités sont toutes égales quelle que soit la cellule choisie du tableau de contingences!! Cela peut se vérifier numériquement pour les sceptiques avec à nouveau un tableur comme la version française de Microsoft Excel 14.0.6123 en créant la structure suivante:

	A	B	C
1	Tableau de contingence:		
2	=ALEA.ENTRE.BORNES(0;15)	=ALEA.ENTRE.BORNES(0;15)	=SOMME(A2:B2)
3	=ALEA.ENTRE.BORNES(0;15)	=ALEA.ENTRE.BORNES(0;15)	=SOMME(A3:B3)
4	=SOMME(A2:A3)	=SOMME(B2:B3)	=SOMME(A4:B4)
5			
6	Contrôle:		
7	=LOI.HYPERGEOMETRIQUE(A2;A4;C2;C4)	=LOI.HYPERGEOMETRIQUE(B2;B4;C2;C4)	
8	=LOI.HYPERGEOMETRIQUE(A3;A4;C3;C4)	=LOI.HYPERGEOMETRIQUE(B3;B4;C3;C4)	

et donc à chaque fois que le lecteur appuiera sur la touche F9 de son clavier il pourra constater que toutes les probabilités sont toujours égales.

Cela peut se vérifier aussi formellement en choisissant une cellule du tableau et en écrivant:

$$P(n, p, m, k) = \frac{\frac{m!}{k!(m-k)!} \frac{(n-m)!}{(p-k)!((n-m)-(p-k))!}}{\frac{n!}{p!(n-p)!}} \quad (57.616)$$

et pour une autre cellule de la même colonne, nous aurons:

$$\begin{aligned} P(n, p, n-m, p-k) &= \frac{\frac{(n-m)!}{(p-k)!((n-m)-(p-k))!} \frac{(n-(n-m))!}{(p-(p-k))!((n-(n-m))-(p-(p-k)))!}}{\frac{n!}{p!(n-p)!}} \\ &= \frac{\frac{(n-m)!}{(p-k)!((n-m)-(p-k))!} \frac{m!}{k!(m-k)!}}{\frac{n!}{p!(n-p)!}} \end{aligned} \quad (57.617)$$

et donc:

$$P(n, p, m, k) = P(n, p, n-m, p-k) \quad (57.618)$$

et ainsi de suite...

Bref, ceci étant dit, nous avons donc dans la case supérieure gauche la valeur 8 alors que l'effectif théorique est de 6. La première chose à laquelle nous pouvons répondre est de savoir si cette valeur de 8 est anormalement grande ou pas par rapport à l'effectif théorique. Pour cela, nous calculons par exemple en unilatéral la probabilité cumulée d'être inférieur ou égal à 8. Nous avons alors avec la version française de Microsoft Excel 14.0.6123 et ultérieur (le dernier paramètre à 1 de la fonction permettant d'indiquer au logiciel que nous voulons la probabilité cumulée):

$$=LOI.HYPERGEOMETRIQUE.N(8;12;9;18;1)=0.995475113$$

Il apparaît donc avec un seuil de 5% en unilatéral que cette valeur est anormalement grande. Nous sommes donc dans une situation extrême.

Par contre, même si les probabilités sont égales pour toutes les cases, la probabilité cumulée elle ne l'est pas! Ainsi, nous avons par exemple pour la valeur de la case inférieure gauche (afin de vérifier si elle est anormalement petite par rapport à l'effectif théorique de 6):

$$=LOI.HYPERGEOMETRIQUE.N(4;12;9;18;1)=0.06561086$$

Donc c'est une valeur qui n'est pas anormalement petite. Cependant, nous souhaiterions avoir un test permettant de conclure si l'ensemble du tableau est dans une configuration extrême ou pas. Or, en faisant le calcul case par case, nous n'allons pas arriver à grand-chose...

L'idée consiste alors à construire tous les tableaux dont les fréquences marginales sont de 9;9 et 12;6 et de calculer la probabilité d'une case donnée (l'avantage de cette technique est que la conclusion sera la même quelle que soit la case prise pour référence du calcul):

Projets	Chefs de projet certifiés	Chefs de projet non certifiés	Total
Délais respectés	9	0	9
Délais non respectés	3	6	9
Total	12	6	18

$$=LOI.HYPERGEOMETRIQUE(9;12;9;18)=0.004524887$$

Projets	Chefs de projet certifiés	Chefs de projet non certifiés	Total
Délais respectés	8	1	9
Délais non respectés	4	5	9
Total	12	6	18

$$=LOI.HYPERGEOMETRIQUE(8;12;9;18)=0.06108597$$

Projets	Chefs de projet certifiés	Chefs de projet non certifiés	Total
Délais respectés	7	2	9
Délais non respectés	5	4	9
Total	12	6	18

$$=LOI.HYPERGEOMETRIQUE(7;12;9;18)=0.244343891$$

Projets	Chefs de projet certifiés	Chefs de projet non certifiés	Total
Délais respectés	6	3	9
Délais non respectés	6	3	9
Total	12	6	18

$$=LOI.HYPERGEOMETRIQUE(6;12;9;18)=0.380090498$$

Projets	Chefs de projet certifiés	Chefs de projet non certifiés	Total
Délais respectés	5	4	9
Délais non respectés	7	2	9
Total	12	6	18

$$=LOI.HYPERGEOMETRIQUE(5;12;9;18)=0.244343891$$

Projets	Chefs de projet certifiés	Chefs de projet non certifiés	Total
Délais respectés	4	5	9
Délais non respectés	8	1	9
Total	12	6	18

$$=LOI.HYPERGEOMETRIQUE(4;12;9;18)=0.061085973$$

Projets	Chefs de projet certifiés	Chefs de projet non certifiés	Total
Délais respectés	3	6	9
Délais non respectés	9	0	9
Total	12	6	18

$$=LOI.HYPERGEOMETRIQUE(3;12;9;18)=0.004524887$$

Soit pour résumer avec les valeurs de k (correspondantes à la cellule supérieure gauche):

k	Probabilité
9	0.00452489
8	0.06108597
7	0.24434389
6	0.3800905
5	0.24434389
4	0.06108597
3	0.00452489
Somme:	1

Tableau: 57.29 - Probabilités de la loi Hypergéométrique correspondantes à la combinaison

Comme dans la colonne du tableau original avec laquelle nous venons de travailler la plus petite valeur est 4 et la plus grande 8, nous allons prendre les probabilités de queues pour savoir quelle est la p -value d'être au-dessus ou égal à 8 et en-dessous ou égal 4 (donc il s'agit d'un test en bilatéral). Nous avons alors:

k	Probabilité
9	0.00452489
8	0.06108597
4	0.06108597
3	0.00452489
Somme:	0.131221719

Tableau: 57.30 - Sélection des valeurs d'intérêt

Donc la p -value est de 13.12%. Nous ne pouvons dès lors pas dire que notre tableau d'origine est dans une configuration extrême si par exemple nous choisissons un seuil de 5%. De nombreux logiciels ne donnent que la p -value.

Remarque: Le choix des bornes est discutable avec ce test car si nous choisissons par exemple de nous concentrer sur la probabilité cumulée d'être dans l'intervalle borné fermé de 4 à 8 (donc inclus!), nous aurions un résultat de 99.09% et donc il faudrait considérer que nous sommes dans une configuration extrême à un seuil de 5%. Donc le choix des bornes avec une loi discrète est toujours délicat dans un test contrairement à un test basé sur une loi continue qui ne souffre absolument pas de ce problème. La majorité des logiciels de statistiques que nous connaissons prennent un intervalle borné ouvert pour l'intervalle (ce qui correspond alors premier calcul que nous avons effectué avec la p -value de 13.12%)

Enfin, pour clore, le lecteur pourra vérifier qu'il tombera sur le même résultat quelle que soit la case qu'il choisit comme référence.

22. KAPPA D'AGRÈMENT DE COHEN

Si nos jugements reflètent notre pensée, ils sont plus rarement en accord avec ceux d'autrui.

Cette variabilité interindividuelle bénéfique pour l'Homme, est cependant pénalisante dans de nombreuses disciplines scientifiques, où il est souvent nécessaire d'évaluer et d'améliorer l'accord entre des informations de même nature appliquées au même objet dans une exigence de contrôle de la qualité ou d'assurance qualité ou encore d'analyse sensorielle.

Le test non paramétrique kappa de Cohen permet par exemple de chiffrer l'accord binaire (dichotomique) entre deux ou plusieurs observateurs ou techniciens lorsque les jugements sont qualitatifs.

Prenons le cas dans le domaine médical où deux ou plusieurs praticiens examinant le même patient proposent des diagnostics différents ou des décisions thérapeutiques différentes. En l'absence d'une référence, cette multiplication des avis n'apporte pas la sécurité attendue d'un parfait accord diagnostique ou thérapeutique pour le médecin traitant et le patient. Il est donc important que l'accord dans une équipe de travail ou entre plusieurs équipes soit le meilleur possible pour garantir la qualité et la continuité des soins.

Une solution consiste ici à réaliser une séance de "mise en concordance" entre les médecins pour estimer leur taux d'accord par le coefficient kappa et d'étudier leurs désaccords pour y remédier.

Plus généralement, le test statistique Kappa est utilisé dans les études de reproductibilité qui nécessitent d'estimer l'agrément entre deux ou plusieurs cotations lorsqu'on étudie une variable discontinue.

Le "**Kappa de Cohen**" est un indicateur empirique permettant de comparer le degré de concordance entre deux observateurs (donneurs d'opinion). C'est un système basé sur l'analyse d'une table de contingences de données appariées (puisque les deux observateurs analysent les mêmes objets).

Pour illustrer le principe, considérons le cas très important où deux responsables qualité ont analysé 11 pièces pour les rejeter ou les accepter. Ils ont obtenu:

		Bob		
		Rejeté	Accepté	Total
Alice	Rejeté	3	2	5
	Accepté	1	5	6
	Total	4	7	11

Tableau: 57.31 - Tableau de contingences dichotomique d'exemple

Les fréquences théoriques étant obtenues par une règle de trois (même calcul de règle de trois que pour le test d'indépendance du Khi-deux et le test exact de Fisher vus plus haut):

		Bob		
		Rejeté	Accepté	Total
Alice	Rejeté	$(4/11)*5=1.82$	$(7/11)*5=3.18$	5
	Accepté	$(4/11)*6=2.18$	$(7/11)*6=3.82$	6
	Total	4	7	11

Tableau: 57.32 - Fréquences théoriques

Le Kappa de Cohen est défini par le rapport:

$$\kappa_c = \frac{\text{Observés}_{\text{Accords}} - \text{Attendus}_{\text{Accords}}}{\text{Essais} - \text{Attendus}_{\text{Accords}}} = \frac{(3+5) - (1.82 + 3.82)}{11 - (1.82 + 3.82)} = 0.441 \quad (57.619)$$

Cette valeur de 0.441 indique un accord modéré entre les deux responsables.

Si plutôt que d'avoir des fréquences, nous travaillons en pourcents (proportions) du total, le Kappa s'écrit alors:

$$\kappa_c = \frac{\% \text{Observés}_{\text{Accords}} - \% \text{Attendus}_{\text{Accords}}}{1 - \% \text{Attendus}_{\text{Accords}}} = \frac{p_{oA} - p_{aA}}{1 - p_{aA}} \quad (57.620)$$

Ce qui donnera avec notre exemple:

$$\kappa_c = \frac{\frac{(3+5)}{11} - \frac{1}{11^2}(4 \cdot 5 + 7 \cdot 6)}{1 - \frac{1}{11^2}(4 \cdot 5 + 7 \cdot 6)} = 0.441 \quad (57.621)$$

Ce qu'il est aussi d'usage d'écrire de manière plus condensée sous la forme:

$$\kappa_c = \frac{\sum_i p_{ii} - \sum_i p_{i+} p_{+i}}{1 - \sum_i p_{i+} p_{+i}} = \frac{(p_{00} + p_{11}) - (p_{0+} \cdot p_{0+} + p_{1+} p_{1+})}{1 - (p_{0+} \cdot p_{0+} + p_{1+} p_{1+})} \quad (57.622)$$

Avec:

$$-1 \leq \kappa_c \leq 1 \quad (57.623)$$

où +1 correspond à un accord parfait et -1 à un désaccord parfait. Évidemment, pour qu'il y ait accord parfait, il faut que les cellules (Rejeté, Rejeté) et (Accepté, Accepté) soient égales et que les autres soient nulles.

L'interprétation suivante est d'usage pour la partie positive (la négative ayant aucun intérêt):

Intervalle	Interprétation
entre 0.8 et 1	Très bon accord
entre 0.6 et 0.8	Bon accord
entre 0.4 et 0.6	Accord modéré
entre 0.2 et 0.4	Accord faible
entre 0.0 et 0.2	Accord nul

Tableau: 57.33 - Interprétations d'usage

Il faut cependant être très critique (comme toujours!) en utilisant ce type d'outil. Comprendre sa construction permet aussi d'en identifier les faiblesses et hypothèses qui sont tout à fait discutables.

23. TEST DE MCNEMAR

Le test de McNemar pourrait très bien se calculer en même temps que le Kappa de Cohen (le premier étant un test d'hypothèse statistiques et le deuxième un uniquement un estimateur ponctuel empirique de concordance). L'idée est que sous l'hypothèse nulle (appelée dans ce cas particulier "**hypothèse de symétrie**"), l'une des diagonales du tableau devrait avoir des valeurs égales. En d'autres termes sous la forme des proportions et en ne nous concentrant que sur une des deux diagonales:

$$H_0 : p_{11} = p_{22} \quad (57.624)$$

ou de fréquences:

$$H_0 : n_{11} = n_{22} \quad (57.625)$$

Sachant que:

$$n_{11} + n_{22} = n \quad (57.626)$$

et sous la condition que n est suffisamment grand, nous pouvons écrire en nous basant sur une loi Binomiale dont le comportement est asymptotiquement Normal:

$$Z = \frac{\frac{n_{11}}{n} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{n}}} \quad (57.627)$$

Nous pouvons nous rendre compte que cela équivaut à écrire:

$$Z = \frac{\frac{n_{11}}{n} - \frac{1}{2} \frac{n_{11} + n_{22}}{n}}{\sqrt{\frac{\left(\frac{n_{11} + n_{22}}{4n}\right)}}} = \frac{n_{11} - \left(\frac{n_{11} + n_{22}}{2}\right)}{\frac{\sqrt{n_{11} + n_{22}}}{2}} \quad (57.628)$$

Dans la littérature spécialisée, nous retrouvons souvent cette dernière relation sous la forme:

$$Z = \frac{b - \left(\frac{b+c}{2}\right)}{\frac{\sqrt{b+c}}{2}} = \frac{b-c}{\sqrt{b+c}} \quad (57.629)$$

Pour en revenir à notre relation initiale, certains en prennent le carré et approximent alors le carré de Z comme une loi du Khi-deux à un degré de liberté (mais bon dans la réalité approximer une loi Normale centrée réduite par une loi du Khi-deux à un degré de liberté c'est un peu n'importe quoi...):

$$Z^2 = \frac{\left(\frac{n_{11}}{n} - 0.5\right)^2}{\frac{0.5(1-0.5)}{n}} = \frac{(b-c)^2}{b+c} \cong \chi^2(1) \quad (57.630)$$

qui est souvent la relation définie dans les livres (sans démonstration...) comme étant le "test de McNemar".

Le test est normalement effectué en bilatéral. L'avantage du test de McNemar est la facilité avec laquelle nous pouvons construire un intervalle de confiance de la différence de la diagonale. Effectivement, partons de l'estimateur de la différence:

$$\hat{d} = p_{11} - p_{22} = \frac{n_{11}}{n} - \frac{n_{22}}{n} \quad (57.631)$$

Nous avons alors en utilisant l'expression de la variance de la loi Binomiale qui est pour rappel (en adaptant juste un peu la notation à ce que nous avons démontré tout au début du chapitre de Statistiques):

$$V(k) = np(1-p) \quad (57.632)$$

Dès lors en utilisant aussi la covariance démontrée lors de notre étude de la loi multinomiale dans le chapitre de Statistiques, il vient:

$$\begin{aligned} V(\hat{d}) &= V\left(\frac{n_{11}}{n} - \frac{n_{22}}{n}\right) = \frac{1}{n^2} V(n_{11} - n_{22}) = \frac{1}{n^2} [V(n_{11}) + V(n_{22}) - 2 \text{cov}(n_{11}, n_{22})] \\ &= \frac{1}{n^2} [np_{11}(1-p_{11}) + np_{22}(1-p_{22}) - 2(-np_{11}p_{22})] \\ &= \frac{1}{n} [p_{11}(1-p_{11}) + p_{22}(1-p_{22}) + 2p_{11}p_{22}] \\ &= \frac{1}{n} \left[\frac{n_{11}}{n} \left(1 - \frac{n_{11}}{n}\right) + \frac{n_{22}}{n} \left(1 - \frac{n_{22}}{n}\right) + 2 \frac{n_{11}}{n} \frac{n_{22}}{n} \right] \end{aligned} \quad (57.633)$$

et donc nous pouvons faire un intervalle de confiance approximatif si les conditions habituelles sont respectées sous la forme:

$$\hat{d} - Z_{\alpha/2} \sqrt{V(\hat{d})} \leq d \leq \hat{d} + Z_{\alpha/2} \sqrt{V(\hat{d})} \quad (57.634)$$

 **Exemple:**

Lors d'un audit social, un sondage est mené sur 200 salariées à propose de l'organisation du travail. Après des réaménagements, la même question est posée. Peut-on considérer qu'il y a eu de réels changements?

	Oui avant	Non avant
Oui après	55	25
Non après	38	82

Tableau: 57.34 - Sondage avant/après (apparié)

Nous avons alors en prenant arbitrairement la diagonale (25, 38) pour l'analyse:

$$Z^2 = \frac{\left(\frac{n_{11}}{n} - 0.5\right)^2}{\frac{0.5(1-0.5)}{n}} = \frac{(b-c)^2}{b+c} = 2.683 \quad (57.635)$$

La majorité des logiciels de statistique ne vous donneront pas 2.683 car ils appliquent une correction empirique à cause de l'approximation grossière qu'est la loi de Khi-deux. Vous aurez alors sur l'écran des logiciels la valeur de 2.285. pour Z^2 .

La p -value sera elle donnée (sans la correction) avec un logiciel comme la version française de Microsoft Excel 14.0.6123 par :

$$=1-LOI.KHIDEUX.N(2.683;1;1)=10.14\%$$

et avec la correction elle donnerait environ 13%. Donc dans les deux cas la p -value étant supérieure au seuil de 5%, on rejette l'hypothèse nulle comme quoi la différence des deux valeurs est grande.

Nous ne calculerons pas dans l'exemple ici présent l'intervalle de confiance de l'écart d car les logiciels de statistiques ont presque tous des méthodes différentes de calcul pour cette valeur.

Enfin, pour clore le sujet concernant le test de McNemar, signalons un indicateur empirique souvent utilisé est qui s'appelle le "**coefficient de Yule**" défini par:

$$Q = \frac{P(X \text{ et } Y \text{ concordants}) - P(X \text{ et } Y \text{ discordants})}{P(X \text{ et } Y \text{ concordants}) + P(X \text{ et } Y \text{ discordants})} \quad (57.636)$$

Ainsi, si Q vaut 1, il est certain que les valeurs concordent et donc que Y est positivement lié à X . Si Q vaut -1, il est respectivement certain que les valeurs sont discordantes et que Y est négativement lié à X . Lorsque les valeurs sont indépendantes, Q sera nul.

Encore une fois, insistons sur le fait qu'il faut cependant être très critique (comme toujours!) en utilisant ce type d'outil (coefficient de Yule ou test de McNemar). Comprendre sa construction permet aussi d'en identifier les faiblesses et hypothèses qui sont tout à fait discutables.

24. TEST DE COCHRAN-MANTEL-HAENZEL

Le "test logarithmique des rangs" appelé aussi "test de (Cochran)-Mantel-Haenzel à temps stratifié" ou encore "test de Mantel-Cox", ou encore "test du khi-2 de Cochran-Mantel-Heanzel" ou plus simplement encore "test CMH"... a pour objectif (principal dans la pratique) de tester l'hypothèse nulle comme quoi deux courbes de survies (groupe Contrôle versus groupe Test), comme celles visibles ci-dessous (où la population survivant a été normalisée sur l'axe des ordonnées), sont significativement différentes ou non sous les hypothèses que:

H1. Chaque strate est indépendante de la précédente

H2. À chaque strate on s'attend à ce que la proportion attendue de survivants/décès soit identique toute chose égale par ailleurs (voir plus loin si ce n'est pas clair).

H3. Chaque strate est distribuée selon une loi hypergéométrique.

H4. La loi hypergéométrique est approximable par une loi Normale (ce que nous allons le rappeler, ne peut être fait que sous certaines conditions).

Autrement dit, l'hypothèse nulle du test est que le traitement (médical, mécanique ou autre) n'a pas d'influence entre le groupe de Contrôle et le groupe de Test.

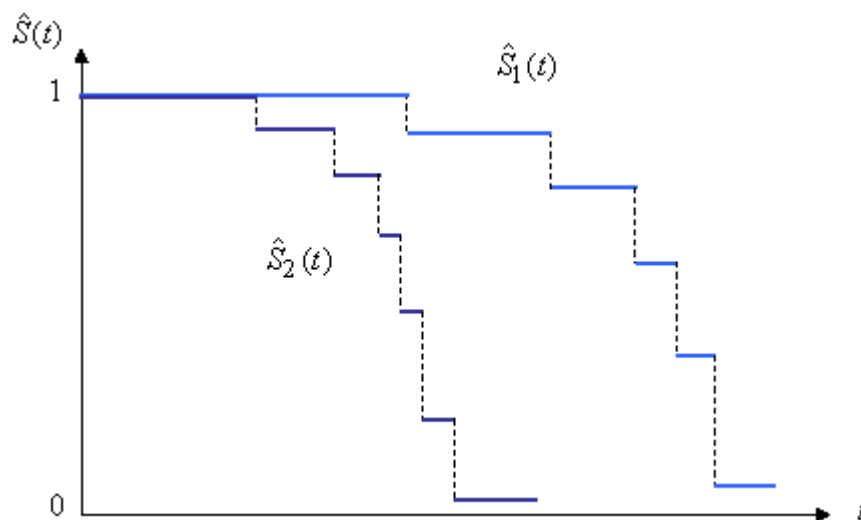


Figure: 57.57 - Courbes de survie type

Autrement dit, l'hypothèse nulle du test est que le traitement (médical, mécanique ou autre) n'a pas d'influence entre le groupe de Contrôle et le groupe de Test.

Pour introduire ce test, nous créons une table de contingence 2×2 pour chaque intervalle (strate) de temps $[t_i, t_{i+1}]$ où $i = 1, 2, \dots, T$ (chaque intervalle de temps peut aussi être assimilé à un hôpital différent pour une même étude clinique) qui aura la structure suivante:

	Morts observés	Survivants observés	Total
Groupe 1 (contrôle)	a_i	b_i	L_{1i}
Groupe 2 (test)	c_i	d_i	L_{2i}
Total	C_{1i}	C_{2i}	n_i

Tableau: 57.35 - Tableau de contingence type pour le test CMH

Nous parlons alors de table générale stratifiée $2 \times 2 \times n$ ou de "tableau de contingence à trois dimensions"...

Remarque: Rappelons que si nous avons qu'une seule et unique table et que nous voulons seulement comparer la proportion de survivants ou morts pour les deux groupes, un test des différences des proportions sera appliqué (cf. chapitre de Statistiques). Nous pourrions aussi faire un test du Khi-deux si nous avons toujours qu'une seule et unique table (voir plus haut) et que les conditions ad hoc sont réunies. Ou encore un test exact de Fisher si les conditions du test du Khi-deux ne sont pas réunies. C'est la raison pour laquelle dans les logiciels de statistiques (comme Minitab par exemple), ces trois tests sont disponibles les uns à côté des autres.

Ainsi, sous l'hypothèse que toutes les choses sont égales par ailleurs (H2), et c'est là l'essence même du test!, le nombre d'individus attendus pour chaque cellule de la période i sera comme pour le test exact de Fisher ou le kappa d'agrément de Kohen égal à:

$$\begin{aligned}
 E_{a,i} &= L_{1i} \frac{C_{1i}}{n_i} = \frac{(a_i + b_i)(a_i + c_i)}{n_i} \\
 E_{c,i} &= L_{2i} \frac{C_{1i}}{n_i} = \frac{(c_i + d_i)(a_i + c_i)}{n_i} \\
 E_{b,i} &= L_{1i} \frac{C_{2i}}{n_i} = \frac{(a_i + b_i)(b_i + d_i)}{n_i} \\
 E_{d,i} &= L_{2i} \frac{C_{2i}}{n_i} = \frac{(c_i + d_i)(b_i + d_i)}{n_i}
 \end{aligned}
 \quad (57.637)$$

Donc il faut bien comprendre que par exemple, $E_{a,i}$ représente alors le nombre d'individus morts attendus du groupe Contrôle si celui-ci se comportait comme l'ensemble des individus Contrôle+Test. Donc si les deux groupes (courbes de survie) se comportent identiquement, la valeur attendue devrait être alors égale à la valeur observée.

Pour bien comprendre, illustrons le concept avec un exemple particulier:

	Morts observés	Survivants observés	Total
Contrôle	200	800	1000
Test	280	1'120	1'400
Total	480	1'920	2'400

Tableau: 57.36 - Cas particulier de valeurs observées

Ici, les valeurs attendues donnent:

	Morts observés	Survivants observés
Contrôle	200	800
Test	280	1'120

Tableau: 57.37 - Valeurs attendues correspondantes au tableau précédent

et nous remarquons que dans ce cas particulier, les valeurs attendues sont égales aux observées (simplement pour la raison que les rapports 800/1'000 et 1'120/1'400 représentent dans le tableau des observations le même pourcentage (la même proportion) de 80%). Évidemment, si la proportion des survivants observés pour les deux groupes sont égaux toutes choses égales par ailleurs, il en va de même pour les morts observés. Donc ce qu'il faut bien comprendre, quand nous faisons le test de MCH, c'est que nous sommes libre de choisir la colonne à analyser (puisque finalement cela revient au même!).

Nous remarquons que chacune des relations:

$$E_{a,i} = L_{1i} \frac{C_{1i}}{n_i}, E_{c,i} = L_{2i} \frac{C_{1i}}{n_i}, E_{b,i} = L_{1i} \frac{C_{2i}}{n_i}, E_{d,i} = L_{2i} \frac{C_{2i}}{n_i} \quad (57.638)$$

représente en réalité l'espérance d'une loi binomiale ou hypergéométrique (cf. chapitre de Statistiques) puisque les deux lois ont la même expression pour l'espérance. Cependant, comme la taille des individus dans les cellules pourrait significative par rapport au total des lignes ou des colonnes, les tirages ne peuvent pas être indépendants. Nous devons alors nous rabattre sur la loi hypergéométrique (H3).

Nous voyons aussi que par symétrie des relations ci-dessus, que la variable d'intérêt soit l'attribut de colonne ou de ligne ne change absolument rien en réalité au résultat du calcul puisque (par exemple):

$$E_{a,i} = L_{1i} \frac{C_{1i}}{n_i} = C_{1i} \frac{L_{1i}}{n_i} \quad (57.639)$$

La variance pour chaque cellule sera alors celle de la loi hypergéométrique que nous avons déjà démontré dans le chapitre de Statistiques comme étant donnée par:

$$V(X) = kpq \frac{n-k}{n-1} = k \left(\frac{r}{n} \right) \left(\frac{s}{n} \right) \frac{n-k}{n-1} = \frac{krs(n-k)}{n^2(n-1)} = \frac{kls}{n^2(n-1)} \quad (57.640)$$

avec pour rappel $l = n - k$. Si nous adaptons l'écriture du chapitre de Statistiques au cas de notre tableau ci-dessus, cela donne pour toutes les cellules (de par l'écriture de la variance de la loi hypergéométrique la variance a la même expression pour chaque cellule!):

$$V_i = L_{1i} \frac{L_{2i} C_{1i} C_{2i}}{n_i^2 (n_i - 1)} = L_{2i} \frac{L_{1i} C_{1i} C_{2i}}{n_i^2 (n_i - 1)} \quad (57.641)$$

avec respectivement par exemple (pour ceux qui veulent faire l'analogie avec notre traitement de la loi hypergéométrique dans le chapitre de Statistiques):

$$\begin{aligned}
 n &= n_i \\
 r &= C_{2i} \\
 s &= C_{1i} \\
 k &= L_{1i} \\
 l &= n - k = n_i - L_{1i} = L_{2i}
 \end{aligned}
 \quad (57.642)$$

Là aussi, nous voyons que peu importe que la variable d'intérêt soit en ligne ou en colonne, la valeur calculée de la variance reste la même!!!

Remarque: Une question qui revient souvent de ceux qui étudient ce test est de savoir pourquoi nous ne pouvons pas simplement faire la somme de toutes les strates dans une seule et unique table (car ce serait bien plus simple évidemment...)? Eh bien nous ne pouvons mettre réunir les tables dans une seule car elles ne suivent pas forcément la même loi (non identiquement distribuées dans le sens que la loi hypergéométrique n'a pas les mêmes paramètres d'une table à l'autre) et malheureusement la loi hypergéométrique n'est pas stable par l'addition.

Et puis....??? En quoi cela va nous aider à comparer si les deux courbes de survie sont identiques à travers le temps (ou à travers différents hôpitaux s'il s'agit d'un test clinique)?!

Eh bien prenons comme exemple de départ le tableau des valeurs observées suivantes:

	Morts observés	Survivants observés	Total
Contrôle	200	900	1'100
Test	280	1'120	1'400
Total	480	2'020	2'500

Tableau: 57.38 - Exemple particulier de valeurs observées

Nous avons alors pour les valeurs attendues:

	Morts observés	Survivants observés	Total
Contrôle	211.2	888.8	1'100
Test	268.8	1'131.2	1'400
Total	480	2'020	2'500

Tableau: 57.39 - Valeurs attendues correspondantes au tableau précédent

Soit la différence:

	Morts observés	Survivants observés
Contrôle	-11.2	11.2
Test	11.2	-11.2

Tableau: 57.40 - Différences entre valeurs attendues et observées

Nous comprenons alors déjà mieux pourquoi cela n'a aucune influence de choisir une cellule en particulier pour faire le test. Nous prendrons simplement celle qui nous arrangera le plus (en fonction de ce qui va suivre...).

Et prenant donc au hasard (puisque le choix n'influence - du moins pour l'instant.... - pas le résultat du test comme nous l'avons montré dans les calculs précédents) la colonne des Survivants et particulièrement les observés du groupe de Test. Sa valeur attendue est alors:

Survivants observés
$E_{d,1} = \frac{L_{21}C_{21}}{n_1} = \frac{1'400 \cdot 2'020}{2'500} = 1'131.2$

Tableau: 57.41 - Valeurs attendue de la cellule Tests-Survivants

La différence entre observés et attendus donne:

$$1'120 - 1'131.2 = -11.2 \quad (57.643)$$

Cette différence serait évidemment nulle si l'observé était égal aux attendus! Avec pour variance:

$$V_1 = \frac{L_{1i}L_{2i}C_{1i}C_{2i}}{n_i^2(n_i-1)} = \frac{480 \cdot 2'020 \cdot 1'100 \cdot 1'400}{2'500^2(2'500-1)} \cong 95.602 \quad (57.644)$$

Maintenant, sous les conditions

$$L_{2i} > 9 \left(\frac{n_i - C_{2i}}{C_{2i}} \right) \quad L_{2i} > 9 \left(\frac{C_{2i}}{n_i - C_{2i}} \right) \quad L_{2i} < \frac{n_i}{10} \quad (57.645)$$

vues dans le chapitre de Statistiques, la loi hypergéométrique peut être approximée par une loi Normale (H4).

Donc dans notre cas, cette approximation n'est pas acceptable (la troisième condition est disqualifiante) et donc le test ne peut être effectué (normalement on s'arrange pour prendre la colonne et la ligne qui permettent cette approximation puisque le choix n'influe pas sur la valeur du test mais sur l'autorisation de faire l'approximation mentionnée!). Mais si elle l'avait été, nous aurions pu approximer la loi hypergéométrique par une loi Normale (cf. chapitre de Statistiques):

$$N(\mu_i, \sigma_i) = \left(L_{2i} \frac{C_{2i}}{n_i}, \sqrt{\frac{L_{2i}L_{1i}C_{1i}C_{2i}}{n_i^2(n_i-1)}} \right) \quad (57.646)$$

Ce qui peut se ramener à une loi Normale centrée réduite si nous prenons dans notre cellule du tableau des observés:

$$\frac{d_i - L_{2i} \frac{C_{2i}}{n_i}}{\sqrt{\frac{L_{2i}L_{1i}C_{1i}C_{2i}}{n_i^2(n_i-1)}}} \cong N(0,1) \quad (57.647)$$

Et dès lors il suffit pour une strate i de savoir si nous sommes en dehors de l'intervalle de confiance que nous nous sommes fixés. Mais... nous avons plusieurs strates! L'idée est alors soit de sommer sur les T strates par indépendance (H1):

$$\sum_{i=1}^T \frac{d_i - L_{2i} \frac{C_{2i}}{n_i}}{\sqrt{\frac{L_{2i} L_{1i} C_{1i} C_{2i}}{n_i^2 (n_i - 1)}}} \cong \sum_{i=1}^T N(0, 1) = N(0, \sqrt{T}) \quad (57.648)$$

ce qui est peu pratique (était du moins à l'époque où pas tout le monde avait un ordinateur), raison pour laquelle on préfère faire la somme suivante:

$$\frac{\sum_{i=1}^T d_i - L_{2i} \frac{C_{2i}}{n_i}}{\sqrt{\sum_{i=1}^T \frac{L_{2i} L_{1i} C_{1i} C_{2i}}{n_i^2 (n_i - 1)}}} \cong N(0, 1) \quad (57.649)$$

Soit sous forme condensée (encore une fois! peu importe le choix de la cellule!):

$$\frac{\sum (d_i - E_{d,i})}{\sqrt{\sum V_i}} \cong N(0, 1) \quad (57.650)$$

Et si les différences entre attendu et observé n'est pas trop grande à travers toutes les strates, la valeur de cette expression se trouvera dans un certain intervalle de confiance de la loi Normes. En dehors l'hypothèse d'égalité des deux courbes de survie sera rejetée.

Cependant, la majorité des logiciels de statistique prennent le carré de cette dernière relation.

Il vient alors que la carré suit une loi du khi-2 à un degré de liberté (nous l'avons démontré dans le chapitre de Statistiques) tel que:

$$\frac{(\sum (d_i - E_{d,i}))^2}{\sum V_i} \cong \chi^2(1) \quad (57.651)$$

Pour des raisons pratiques, on ajoute un terme 0.5 à la somme, ceci assure une meilleure approximation à la loi Normale. Donc:

$$\frac{(\sum (d_i - E_{d,i}) + 0.5)^2}{\sum V_i} \cong \chi^2(1) \quad (57.652)$$

Exemple:

Considérons les valeurs observées suivantes de deux hôpitaux A et B pour un test clinique:

A	Morts observés	Survivants observés	Total
Contrôle	288	4	292
Test	400	50	450
Total	688	54	742

<i>B</i>	Morts observés	Survivants observés	Total
Contrôle	300	10	310
Test	450	40	490
Total	750	50	800

Tableau: 57.42 - Test clinique stratifié sur deux niveaux (valeurs observées)

Nous souhaiterions donc non pas savoir si les deux hôpitaux sont différents ou pas (là n'est pas le propos), mais de savoir si les différences entre le groupe de Contrôle et de Test sur l'ensemble des hôpitaux est significativement différent ou non.

Bon nous devinons déjà le résultat intuitivement au vu des valeurs... mais faisons quand même les calculs.

Nous voyons déjà que pour l'hôpital A, la colonne des Survivants satisfait les trois conditions pour l'approximation par une loi Normale (ce qui n'est pas le cas pour la colonne des Morts):

$$54 > 9 \left(\frac{742 - 450}{742} \right) \quad 54 > 9 \left(\frac{742}{742 - 450} \right) \quad 54 < \frac{742}{10} \quad (57.653)$$

De même pour l'hôpital B (et c'est heureux car une fois une colonne choisie pour une strate, il faut que la condition d'approximation soit applicable à la même colonne de toutes les autres strates):

$$50 > 9 \left(\frac{800 - 490}{800} \right) \quad 50 > 9 \left(\frac{800}{800 - 490} \right) \quad 50 < \frac{80}{10} \quad (57.654)$$

Les valeurs attendues respectives donnent pour la colonne d'intérêt:

<i>A</i>	Survivants observés	Total
Contrôle	21.25	292
Test	32.75	450
Total	54	742

<i>B</i>	Survivants observés	Total
Contrôle	19.38	310
Test	30.62	490
Total	50	800

Tableau: 57.43 - Valeurs attendues de la colonne d'intérêt du test clinique sur deux niveaux

Ce qui donne pour les différences:

<i>A</i>	Survivants observés
Contrôle	-17.25
Test	17.25

<i>B</i>	Survivants observés
Contrôle	-9.38
Test	9.38

Tableau: 57.44 - Différences valeurs observées-attendues de la colonne d'intérêt

Nous voyons alors rapidement pourquoi une fois la colonne choisie, celui de la ligne n'a plus d'importance (soit une somme des valeurs négatives, soit on somme des valeurs positives et comme on prend le carré de la somme, cela ne change rien au final!). Choisissons arbitrairement la deuxième ligne. Nous avons alors:

$$\frac{\left(\sum_{i=1}^2 (d_i - E_{d,i})\right)^2}{\sum_{i=1}^2 V_i} = \frac{\left(\sum_{i=1}^2 \left(d_i - \frac{L_{2i} C_{2i}}{n_i}\right)\right)^2}{\sum_{i=1}^2 \frac{L_{2i} L_{1i} C_{1i} C_{2i}}{n_i^2 (n_i - 1)}} = \frac{\left(\sum_{i=1}^2 \left(d_i - \frac{L_{2i} C_{2i}}{n_i}\right)\right)^2}{\sum_{i=1}^2 \frac{L_{2i} L_{1i} C_{1i} C_{2i}}{n_i^2 (n_i - 1)}} \quad (57.655)$$

$$= \frac{(17.25 + 9.38)^2}{\frac{292 \cdot 450 \cdot 688 \cdot 54}{742^2 (742 - 1)} + \frac{310 \cdot 490 \cdot 750 \cdot 50}{258^2 (258 - 1)}} \cong \frac{736}{11.966 + 11.139} \cong 31.845$$

Et nous avons avec un tableur standard comme Microsoft Excel 14.0.6123 en unilatéral gauche à un seuil de risque de 5%:

$$\chi_{95\%}^2(1) = \text{LOI.KHIDEUX.INVERSE}(95\%;1) \cong 3.841 \quad (57.656)$$

Donc sur le cumul des deux hôpitaux (strates) le groupe de Contrôle est significativement différent du groupe de Test. La p-value du test est typiquement donnée avec Microsoft Excel 14.0.6123:

$$1 - \text{LOI.KHIDEUX.N}(31.845;1;1) = 0.000002\% \quad (57.657)$$

25. MÉTHODE DES DIFFÉRENCES FINIES

Dans le domaine des méthodes numériques, nous pouvons être amenés à rechercher la solution d'une équation aux dérivées partielles. Parmi les méthodes de résolution couramment pratiquées, la méthode des différences finies ou M.D.F. est la plus facile d'accès, puisqu'elle repose sur deux notions: la discrétisation des opérateurs de dérivation/différenciation (assez intuitive) d'une part, et la convergence du schéma numérique ainsi obtenu d'autre part.

Prenons deux exemples fameux (car très scolaires) qui ne sont que des cas particuliers et simplistes d'application de la M.D.F.

25.1. M.D.F À UNE DIMENSION SPATIALE

Rappelons que nous avons démontré dans le chapitre de Thermodynamique l'équation de la chaleur suivante (nous présentons ici cette équation réduite à une dimension spatiale):

$$\frac{\partial T(x,t)}{\partial t} - \frac{\kappa}{\rho c_p} \frac{\partial^2 T(x,t)}{\partial x^2} = 0 \quad (57.658)$$

et remarquons que cette équation n'est pas très générale... (elle n'est pas relativiste et ne prend pas en compte la chaleur dégagée sous forme de rayonnement par le matériau considéré ni plein d'autres facteurs...).

Nous pouvons considérer (cf. chapitre de Calcul Différentiel Et Intégral) que:

$$\frac{\partial T(x,t)}{\partial x} \cong \frac{T(x+h,t) - T(x,t)}{h} \quad (57.659)$$

et:

$$\begin{aligned} \frac{\partial^2 T(x,t)}{\partial^2 x} &\cong \frac{1}{h} \left(\frac{T(x+h,t) - T(x,t)}{h} - \frac{T(x,t) - T(x-h,t)}{h} \right) \\ &= \frac{1}{h^2} (T(x+h,t) - 2T(x,t) + T(x-h,t)) \end{aligned} \quad (57.660)$$

De même:

$$\frac{\partial T(x,t)}{\partial t} \cong \frac{T(x,t+k) - T(x,t)}{k} \quad (57.661)$$

L'équation de la chaleur devient alors:

$$\frac{T(x,t+k) - T(x,t)}{k} - \frac{\kappa}{\rho c_p} \frac{1}{h^2} (T(x+h,t) - 2T(x,t) + T(x-h,t)) = 0 \quad (57.662)$$

Après réarrangement, nous avons:

$$\begin{aligned} T(x,t+k) &= k \frac{\kappa}{\rho c_p} \frac{1}{h^2} (T(x+h,t) - 2T(x,t) + T(x-h,t)) + T(x,t) \\ &= \left(1 - \frac{2k\kappa}{\rho c_p h^2} \right) T(x,t) + k \frac{\kappa}{\rho c_p} \frac{1}{h^2} (T(x+h,t) + T(x-h,t)) \end{aligned} \quad (57.663)$$

Si nous regardons cette relation de plus près, nous observons qu'il s'agit d'une simple récursivité. Il suffit de connaître la distribution $T(x,0)$ pour déterminer ensuite toutes les autres valeurs puisque:

$$T(x,0+k) = \left(1 - \frac{2k\kappa}{\rho c_p h^2} \right) T(x,0) + k \frac{\kappa}{\rho c_p} \frac{1}{h^2} (T(x+h,0) + T(x-h,0)) \quad (57.664)$$

et:

$$T(x,2k) = \left(1 - \frac{2k\kappa}{\rho c_p h^2} \right) T(x,k) + k \frac{\kappa}{\rho c_p} \frac{1}{h^2} (T(x+h,k) + T(x-h,k)) \quad (57.665)$$

etc. Il est possible de mettre en oeuvre une telle simulation rien qu'avec un petit tableau et un peu de temps... h est appelé alors le "**pas de maillage**" du modèle.

Pour le lecteur souhaitant s'entraîner... une barre de Fer longitudinale de 1 kilogramme a une capacité calorifique massique de $450 [J \cdot kg^{-1} \cdot K^{-1}]$, une densité de $7.88 [kg \cdot m^{-3}]$ et sa conductivité thermique est de $82 [J \cdot s^{-1} \cdot m^{-1} \cdot K^{-1}]$.

25.2. M.D.F SPATIO-TEMPORELLE

La M.D.F. est donc une méthode numérique très importante dans la pratique car elle permet aussi de résoudre par exemple directement les équations de Maxwell dans le domaine temps et dans l'espace. Elle se classe alors dans la famille des méthodes 3D (discrétisation tridimensionnelle de l'espace) et temporelles et trouve ses principales applications dans les domaines de la conception (antennes et circuits), de la compatibilité électromagnétique, de la diffraction, de la propagation et de la dosimétrie électromagnétique (interaction ondes- matière vivante).

Nous allons présenter ici les bases du concept car dans la pratique, la programmation de la M.D.F. est un métier à lui seul (comme tout le reste du site bien évidemment mais il est parfois utile de le rappeler).

Dans un problème d'électrodynamique traité par M.D.F., la première opération nécessaire consiste à délimiter le volume V de l'espace et l'intervalle de temps $I=[0,T]$ pour lesquels on souhaite effectuer cette résolution (il est illusoire d'espérer résoudre les équations de Maxwell dans l'espace infini et pour une durée illimitée!). Le volume de calcul contient l'objet (antenne, circuit, ...) que l'on souhaite caractériser, en réponse à une excitation donnée. Dans un deuxième temps, il convient de procéder à un échantillonnage de l'espace (maillage de V) et du temps (discrétisation de I) afin de permettre une implémentation numérique de la résolution. Le problème devient dès lors celui de la détermination du champ en tout point du maillage et pour tout instant discret de l'intervalle d'observation. Les échantillonnages spatial et temporel seront précisés dans la suite et découleront naturellement de la physique des équations à résoudre. Ils conditionnent bien évidemment à la fois la précision des résultats du calcul et les ressources informatiques requises pour le mener à bien.

La structuration du maillage M.D.F. et le cheminement de la résolution résultent directement des équations à résoudre.

Dans un milieu linéaire, homogène, isotrope, non dispersif et non magnétique (...), celles-ci s'écrivent explicitement en se basant sur la troisième équation de Maxwell (cf. chapitre d'Électrodynamique):

$$-\frac{\partial \vec{B}}{\partial t} = \vec{\nabla} \times \vec{E} \quad (57.666)$$

Soit explicitement et avec le signe négatif passé de l'autre côté de l'égalité:

$$\begin{aligned} \frac{\partial B_x}{\partial t} &= -(\partial_y E_z - \partial_z E_y) \\ \frac{\partial B_y}{\partial t} &= -(\partial_z E_x - \partial_x E_z) \\ \frac{\partial B_z}{\partial t} &= -(\partial_x E_y - \partial_y E_x) \end{aligned} \quad (57.667)$$

Et nous utiliserons aussi la quatrième équation de Maxwell sans sources:

$$\vec{\nabla} \times \vec{B} = \mu\epsilon \frac{\partial \vec{E}}{\partial t} \quad (57.668)$$

Soit explicitement et réarrangé:

$$\begin{aligned}
\frac{\partial E_x}{\partial t} &= \frac{1}{\mu\epsilon} (\partial_y B_z - \partial_z B_y) \\
\frac{\partial E_y}{\partial t} &= -(\partial_z B_x - \partial_x B_z) \\
\frac{\partial E_z}{\partial t} &= -(\partial_x B_y - \partial_y B_x)
\end{aligned} \quad (57.669)$$

Soit pour résumer:

$$\begin{cases}
\frac{\partial}{\partial t} B_x(\vec{r}, t) = -(\partial_y E_z(\vec{r}, t) - \partial_z E_y(\vec{r}, t)) \\
\frac{\partial}{\partial t} B_y(\vec{r}, t) = -(\partial_z E_x(\vec{r}, t) - \partial_x E_z(\vec{r}, t)) \\
\frac{\partial}{\partial t} B_z(\vec{r}, t) = -(\partial_x E_y(\vec{r}, t) - \partial_y E_x(\vec{r}, t)) \\
\frac{d}{dt} E_x(\vec{r}, t) = \frac{1}{\mu\epsilon} (\partial_y B_z(\vec{r}, t) - \partial_z B_y(\vec{r}, t)) \\
\frac{d}{dt} E_y(\vec{r}, t) = -(\partial_z B_x(\vec{r}, t) - \partial_x B_z(\vec{r}, t)) \\
\frac{d}{dt} E_z(\vec{r}, t) = -(\partial_x B_y(\vec{r}, t) - \partial_y B_x(\vec{r}, t))
\end{cases} \quad (57.670)$$

Dans la suite, nous nous intéresserons uniquement à la première équation, les autres conduisant à des développements similaires.

Afin de permettre un traitement sur calculateur, les différentes dérivées présentes dans l'équation doivent être approchées numériquement. Pour ce faire, nous utilisons le principe des différences finies centrées qui s'appuie sur les développements en série de Taylor suivants (cf. chapitre Suites Et Séries):

$$f(x) = \sum_{k=0}^{\infty} \frac{1}{k!} \frac{d^k f(x_0)}{dx^k} (x - x_0)^k = \sum_{k=0}^{\infty} \frac{d^{(k)} f(x_0)}{k!} (x - x_0)^k \quad (57.671)$$

Nous avons alors sur la base de ce principe:

$$\begin{aligned}
f\left(x_0 + \frac{d_x}{2}\right) &= f(x_0) + \frac{d_x}{2} f'(x_0) + \frac{1}{2} \left(\frac{d_x}{2}\right)^2 f''(x_0) + \dots \\
f\left(x_0 - \frac{d_x}{2}\right) &= f(x_0) - \frac{d_x}{2} f'(x_0) + \frac{1}{2} \left(\frac{d_x}{2}\right)^2 f''(x_0) - \dots
\end{aligned} \quad (57.672)$$

Si nous négligeons les termes du deuxième ordre, il vient en soustrayant les deux séries:

$$f'(x_0) = \frac{f\left(x_0 + \frac{d_x}{2}\right) - f\left(x_0 - \frac{d_x}{2}\right)}{d_x} + \mathcal{E}(d_x^2) \quad (57.673)$$

où $\varepsilon(d_x^2)$ est une erreur d'ordre 2, négligée par la suite (nous noterons que c'est le centrage qui, en permettant une compensation des dérivées secondes, minimise l'erreur dans l'approximation).

En appliquant ce principe aux dérivées temporelles et spatiales de:

$$\begin{aligned}\frac{\partial}{\partial t} B_x(\vec{r}, t) &= -(\partial_y E_z(\vec{r}, t) - \partial_z E_y(\vec{r}, t)) \\ \frac{\partial}{\partial t} B_y(\vec{r}, t) &= -(\partial_z E_x(\vec{r}, t) - \partial_x E_z(\vec{r}, t)) \\ \frac{\partial}{\partial t} B_z(\vec{r}, t) &= -(\partial_x E_y(\vec{r}, t) - \partial_y E_x(\vec{r}, t))\end{aligned} \quad (57.674)$$

nous obtenons:

$$\begin{aligned}& \frac{B_x\left(x, y, z, t + \frac{d_t}{2}\right) - B_x\left(x, y, z, t - \frac{d_t}{2}\right)}{d_t} \\ &= - \left(\frac{E_z\left(x, y + \frac{d_y}{2}, z, t\right) - E_z\left(x, y - \frac{d_y}{2}, z, t\right)}{d_y} - \frac{E_y\left(x, y, z + \frac{d_z}{2}, t\right) - E_y\left(x, y, z - \frac{d_z}{2}, t\right)}{d_z} \right)\end{aligned} \quad (57.675)$$

ou encore après réarrangement:

$$\begin{aligned}& B_x\left(x, y, z, t + \frac{d_t}{2}\right) = B_x\left(x, y, z, t - \frac{d_t}{2}\right) \\ & + d_t \left(\frac{E_y\left(x, y, z + \frac{d_z}{2}, t\right) - E_y\left(x, y, z - \frac{d_z}{2}, t\right)}{d_z} - \frac{E_z\left(x, y + \frac{d_y}{2}, z, t\right) - E_z\left(x, y - \frac{d_y}{2}, z, t\right)}{d_y} \right)\end{aligned} \quad (57.676)$$

Cette relation montre que si nous connaissons les composantes E_y, E_z du champ électrique à l'instant t et la composante B_x du champ magnétique à l'instant antérieur $t - d_t/2$, il est possible de déterminer B_x à l'instant $t + d_t/2$. Bien évidemment la démarche est parfaitement identique pour toutes les composantes et montre le même décalage temporel. Ce résultat suggère donc d'utiliser une résolution numérique itérative, dans laquelle les champs électrique et magnétique sont évalués alternativement, respectivement aux instants discrets nd_t et $(n+1/2)d_t$, d_t étant le pas du temps. Il est d'usage dans la littérature spécialisée de noter $B_x^{n+1/2}$ la composante du champ magnétique à l'instant $(n+1/2)d_t$.

Le même constat s'applique pour la distribution spatiale des points d'observation du champ. Ainsi, l'évaluation de B_x au point (x, y, z) s'appuie sur la connaissance de E_y aux points $(x, y, z + d_z/2)$ et $(x, y, z - d_z/2)$ et de E_z aux points $(x, y + d_y/2, z)$ et $(x, y - d_y/2, z)$.

Nous pouvons donc résumer cela sous la forme géométrique suivante:

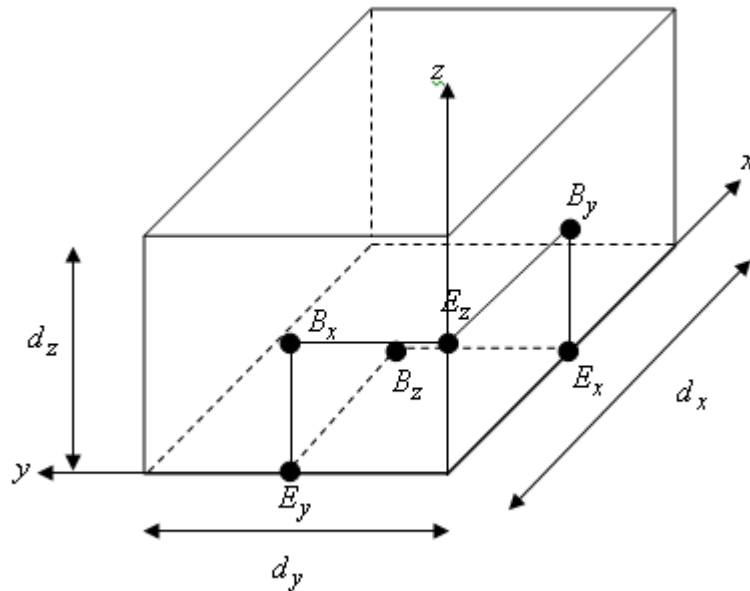


Figure: 57.58 - Cellule de Yee

Les composantes du champ électrique sont évaluées aux centres des arêtes du maillage et les composantes du champ magnétique aux centres des faces de façon à garantir l'alternance imposée par les équations (on appelle "**cellule de Yee**" la maille élémentaire dotée de cette répartition de points).

Et donc globalement, nous avons un maillage de l'espace qui peut être représenté sous la forme suivante:

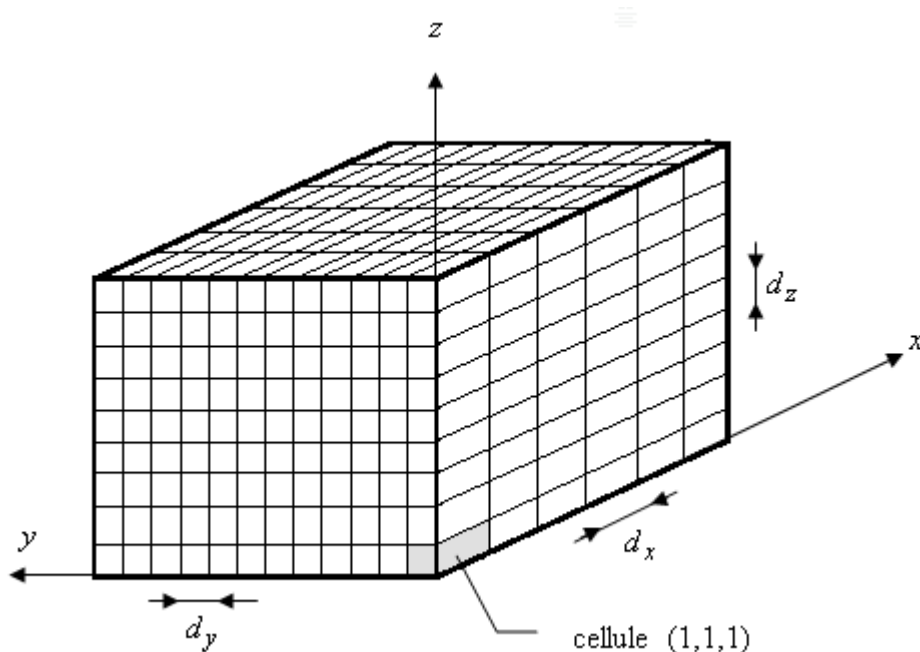


Figure: 57.59 - Exemple de maillage par parallélépipède

Évidemment, nous pouvons exécuter les routines de calcul avec une valeur de la permittivité et de la perméabilité du vide non forcément égales dans toutes les cellules. Ce qui permet en plus de modéliser la propagation d'ondes électromagnétiques dans des milieux hétérogènes et non isotropes.

Enfin, il est important de noter que le pas d'échantillonnage spatial et temporel doit pouvoir être paramétrable par l'utilisateur qui exécute ce type de simulations. Ceci pour des raisons de temps de calculs et aussi de précision. Effectivement, nous n'effectuons pas les mêmes modélisations pour un système multiphysique dans le vide à basse fréquence que dans la matière non isotrope à haute fréquence et ce, que ce soit sur un ordinateur de bureau ou un supercalculateur.

26. CLUSTERING

En analyse de données statistiques, le clustering (data clustering pour les anglophones) décrit des méthodes empiriques de classification de données (méthode de regroupement hiérarchique ou méthode de partitionnement de données).

Il s'agit de techniques permettant typiquement la segmentation de l'ensemble des clients d'une entreprise en fonction de leur démographie ou de leurs habitudes d'achat, de grouper des documents pour des présentations, d'identifier de nouvelles espèces animales ou végétales, de regrouper de l'information ou des individus par par intérêts.

Nous allons voir ici quelques techniques triviales que nous compléterons avec le temps...

26.1. ARBRE DE RÉGRESSION ET DE CLASSIFICATION

Les arbres de régression et de classification (en anglais CART pour Classification and Regression Tree) sont un ensemble d'algorithmes heuristiques très utilisés en marketing ou dans les sciences sociales pour discriminer (catégoriser) une très grande population. Bien évidemment, ces algorithmes ne feront jamais mieux qu'un être humain... mais demandez aussi à un employé de créer des groupes dans une population de 5 millions de clients sur la base d'une dizaine de variables explicatives. Vous allez pouvoir attendre la réponse longtemps...

Bien que ces techniques automatisées de classification soient très utiles dans les situations susmentionnées, elles ont cependant un problème majeur qui fait que nous ne nous étendrons pas trop sur le sujet:

1. Ces techniques sont très sensibles à la population analysée et donnent des résultats très différents.
2. Les différentes techniques de classifications existantes donnent des résultats souvent totalement différents pour une même population.

Il vaut donc mieux être prudent quant aux conclusions que nous pouvons en tirer et comparer les résultats de plusieurs méthodes et en fonction du retour d'expérience choisir celle qui subjectivement semble la mieux adaptée.

Considérons un ensemble de variables catégorielles $x_1, \dots, x_p$. Le partitionnement récursif a pour but de diviser l'espace des p variables en rectangles qui ne se chevauchent pas.

Par exemple, soit la variable x_i et une valeur s_i de cette variable, on trouve que le partitionnement où $x_i < s_i$ et $s_i < x_i$ sépare bien les données en deux ensembles disjoints. Ensuite, une des deux parties est à son tour divisée par une valeur de x_i ou par la valeur d'une autre variable. Nous aboutissons alors à trois rectangles et ainsi de suite...

L'idée est de créer n rectangles de telle sorte que l'ensemble de données contenues dans un rectangle soit homogène (c'est-à-dire ne contienne qu'une seule famille de points).

Pour traiter de ce sujet, considérons un cas pratique:

Un concessionnaire souhaiterait pour sa ville trouver un moyen de classer les familles qui sont à même d'acheter une voiture (propriétaires) et celles qui ne sont pas prêtes à en acheter (non-propriétaires). Un échantillon de 12 propriétaires (valeur "1" dans la figure ci-dessous) et 12 non propriétaires (valeur "2" dans la figure ci-dessous) est choisi. Les deux variables indépendantes sont x_1 (revenus en kilo-francs) et x_2 (surface de leur habitat):

	A	B	C	D
1	Observation	Revenus	Surface	Propriétaire
2	1	60	18.4	1
3	2	85.5	16.8	1
4	3	64.8	21.6	1
5	4	61.5	20.8	1
6	5	87	23.6	1
7	6	110.1	19.2	1
8	7	108	17.6	1
9	8	82.8	22.4	1
10	9	69	20	1
11	10	93	20.8	1
12	11	51	22	1
13	12	81	20	1
14	13	75	19.6	2
15	14	52.8	20.8	2
16	15	64.8	17.2	2
17	16	43.2	20.4	2
18	17	84	17.6	2
19	18	49.2	17.6	2
20	19	59.4	16	2
21	20	66	18.4	2
22	21	47.4	16.4	2
23	22	33	18.8	2
24	23	51	14	2
25	24	63	14.8	2

Figure: 57.60 - Données à classifier

Nous voyons que nous avons autant de propriétaires que de non-propriétaires (fréquence d'apparition égale dans toute la population). Dès lors la probabilité d'appartenance à une classe est de 50%.

Soit graphiquement:

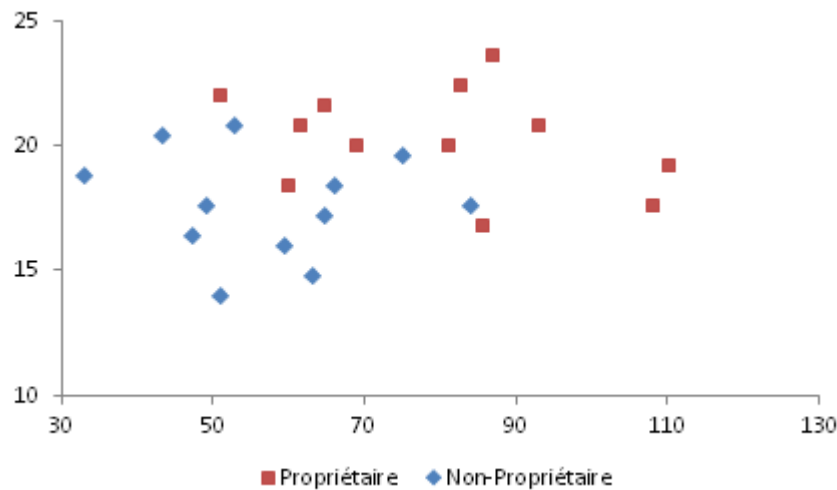


Figure: 57.61 - Représentation des points (revenus, surface)

Si nous appliquons l'algorithme CART sur ces données, nous voyons que nous devons choisir x_2 (Surface) comme premier choix de division avec la valeur de division 19 (nous allons justifier pourquoi!). L'espace (x_1, x_2) est maintenant divisé en deux rectangles (il était facile de deviner cette étape de discrimination même sans faire appel aux mathématiques):

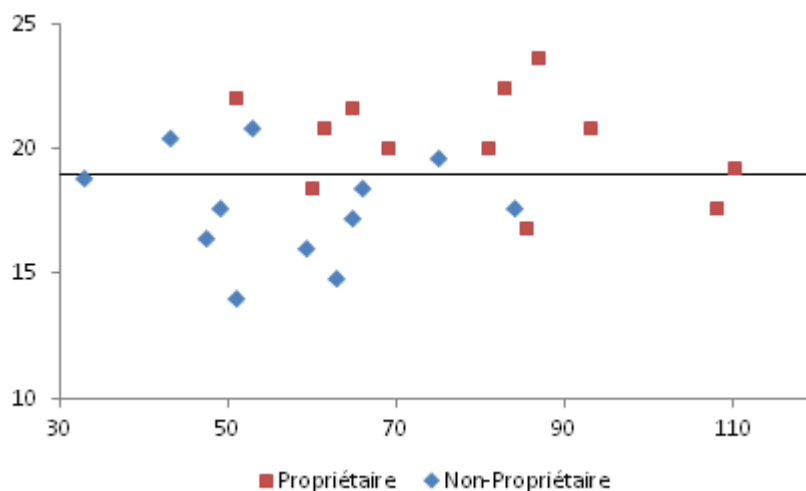


Figure: 57.62 - Première itération de la classification

Notez comment la division en deux rectangles a créé deux zones (split) qui sont plus homogènes que le graphique initial. Le rectangle supérieur contient des points qui sont davantage des propriétaires tandis que le rectangle inférieur contient davantage de non-propriétaires.

Pour déterminer cette division, l'algorithme CART a examiné chaque variable et toutes les valeurs possibles de division pour chaque variable de façon à trouver la meilleure division.

Ainsi, les points de division possible pour x_1 sont (remarquez qu'il s'agit à chaque fois de la moyenne de deux valeurs connexes du tableau):

$$x_1 : \{38.1, 45.3, 50.1, \dots, 109.5\} \quad (57.677)$$

et ceux pour x_2 sont:

$$x_2 : \{14.4, 15.4, 16.2, \dots, 23\} \quad (57.678)$$

Ces points sont ordonnés par l'algorithme d'après la façon dont ils réduisent "l'impureté" (hétérogénéité de composition) dans le rectangle que génère le "split".

Il existe un grand nombre de façons empiriques de mesurer l'impureté. La plus commune à ce jour est l'utilisation d'un indicateur inspiré du coefficient de Gini (cf. chapitre de Techniques De Gestion). Ainsi, si nous dénotons les classes par $k = 1, 2, 3, \dots, C$ où C est le nombre total de classes à prédire, "l'indice d'impureté de Gini" pour le rectangle A est défini par:

$$I(A) = 1 - \sum_{k=1}^C p_k^2 \quad (57.679)$$

où p_k est la fraction d'observations dans le rectangle A qui appartiennent à la

classe k . Dans le cas qui nous concerne, nous avons toujours deux classes: Propriétaires/Non-Propriétaires.

Ensuite, l'indice de Gini global est défini comme la moyenne pondérée des indices de Gini.

Ainsi dans le cadre de notre exemple, nous avons deux classes, donc $C = 2$. Avant la séparation, nous avons:

$$I(A) = 1 - \sum_{k=1}^C p_k^2 = 1 - \left(\frac{12}{24}\right)^2 - \left(\frac{12}{24}\right)^2 = 1 - 2(0.5)^2 = 0.5 \quad (57.680)$$

La séparation trouvée en 19 (voir le deuxième graphique) donne par exemple pour le rectangle supérieur de la première subdivision:

$$\begin{aligned} I(A_{\text{sup}}) &= 1 - \sum_{k=1}^2 p_k^2 = 1 - \left[\left(\frac{\text{points bleus}}{n} \right)^2 + \left(\frac{\text{points rouges}}{n} \right)^2 \right] \\ &= 1 - \left[\left(\frac{3}{12} \right)^2 + \left(\frac{9}{12} \right)^2 \right] = 0.375 \end{aligned} \quad (57.681)$$

Pour la partie inférieure:

$$\begin{aligned} I(A_{\text{inf}}) &= 1 - \sum_{k=1}^2 p_k^2 = 1 - \left[\left(\frac{\text{points bleus}}{n} \right)^2 + \left(\frac{\text{points rouges}}{n} \right)^2 \right] \\ &= 1 - \left[\left(\frac{9}{12} \right)^2 + \left(\frac{3}{12} \right)^2 \right] = 0.375 \end{aligned} \quad (57.682)$$

Le hasard faisant, l'impureté est la même pour les deux rectangles (du haut et du bas).

L'indice de Gini global est alors donné par:

$$I(A) = \frac{12}{24} 0.375 + \frac{12}{24} 0.375 = 0.375 \quad (57.683)$$

Remarquons avant de continuer que si la subdivision est parfaite (n'avoir qu'une seule famille de points dans un des rectangles), nous avons alors:

$$\begin{aligned} I(A_{\text{inf,sup}}) &= 1 - \sum_{k=1}^2 p_k^2 \\ &= 1 - \left[\left(\frac{\text{points bleus}}{n} \right)^2 + \left(\frac{\text{points rouges}}{n} \right)^2 \right] = 1 - \left[\left(\frac{12}{12} \right)^2 + \left(\frac{0}{12} \right)^2 \right] \quad (57.684) \\ &= 1 - \left[\left(\frac{\text{points bleus}}{12} \right)^2 + \left(\frac{\text{points rouges}}{12} \right)^2 \right] = 1 - \left[\left(\frac{0}{12} \right)^2 + \left(\frac{12}{12} \right)^2 \right] \\ &= 0 \end{aligned}$$

donc l'impureté est nulle... Et si tous les points apparaissent en proportions égales dans chacun des rectangles (pire des cas en quelque sorte), la valeur est alors:

$$\begin{aligned} I(A_{\text{inf,sup}}) &= 1 - \sum_{k=1}^2 p_k^2 \\ &= 1 - \left[\left(\frac{\text{points bleus}}{n_1} \right)^2 + \left(\frac{\text{points rouges}}{n_2} \right)^2 \right] = 1 - \left[\left(\frac{x}{n_1} \right)^2 + \left(\frac{y}{n_2} \right)^2 \right] \quad (57.685) \\ &\stackrel{x=y}{n_1=n_2=n} = 1 - \left[\left(\frac{x}{n} \right)^2 + \left(\frac{x}{n} \right)^2 \right] = 1 - 2 \left(\frac{x}{n} \right)^2 \stackrel{x=\frac{1}{2}n}{=} 1 - 2 \left(\frac{1}{2} \right)^2 = 1 - 2 \frac{1}{4} = 1 - \frac{1}{2} = 0.5 \end{aligned}$$

Si nous généralisons à C classes (C dimensions spatiales), il vient immédiatement:

$$I(A_{\text{inf,sup}}) = 1 - \sum_{k=1}^C p_k^2 = 1 - \frac{1}{C} = \frac{C-1}{C} \quad (57.686)$$

ce qui est l'impureté maximale.

Donc l'impureté est toujours définie par une valeur dans l'intervalle:

$$I(A) \in \left[0, \frac{C-1}{C} \right] \quad (57.687)$$

Maintenant, pour continuer avec notre exemple, même sans faire appel à un algorithme informatique et sans même calculer l'impureté, il est relativement aisé de deviner où va se trouver la prochaine discrimination: elle sera en $x_1 = 84.75$ (revenus). Ce qui donnera:

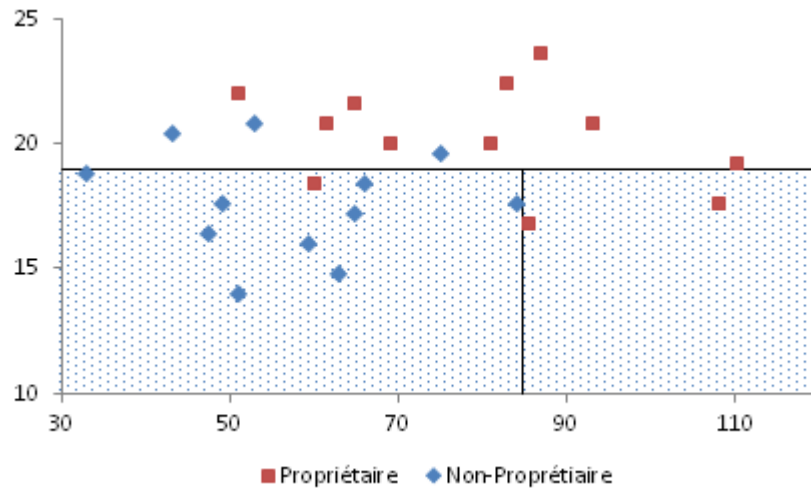


Figure: 57.63 - Deuxième itération de la classification

Ce qui était aussi facile à deviner même sans faire appel aux calculs (essayez avec votre entourage, vous verrez que très souvent ils trouvent les deux premières discriminations).

L'impureté devra donc être calculée dans la nouvelle zone discriminée mise en évidence par un hachurage. Nous y avons donc:

$$\begin{aligned}
 I(A_{\text{gauche}}) &= 1 - \sum_{k=1}^2 p_k^2 = 1 - \left[\left(\frac{\text{points bleus}}{n} \right)^2 + \left(\frac{\text{points rouges}}{n} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{9}{10} \right)^2 + \left(\frac{1}{10} \right)^2 \right] = 0.18 \\
 I(A_{\text{droite}}) &= 1 - \sum_{k=1}^2 p_k^2 = 1 - \left[\left(\frac{\text{points bleus}}{n} \right)^2 + \left(\frac{\text{points rouges}}{n} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0}{2} \right)^2 + \left(\frac{2}{2} \right)^2 \right] = 0
 \end{aligned} \tag{57.688}$$

L'indice de Gini global est alors:

$$I(A) = \frac{10}{12} \cdot 0.18 + \frac{2}{12} \cdot 0 = 0.15 \tag{57.689}$$

Nous continuons, mais sachez que la suite est moins facile à deviner. La majorité des individus interrogés se trompent sans faire appel à la définition mathématique de l'impureté et proposent intuitivement à tort l'une ou l'autre des discriminations suivantes:

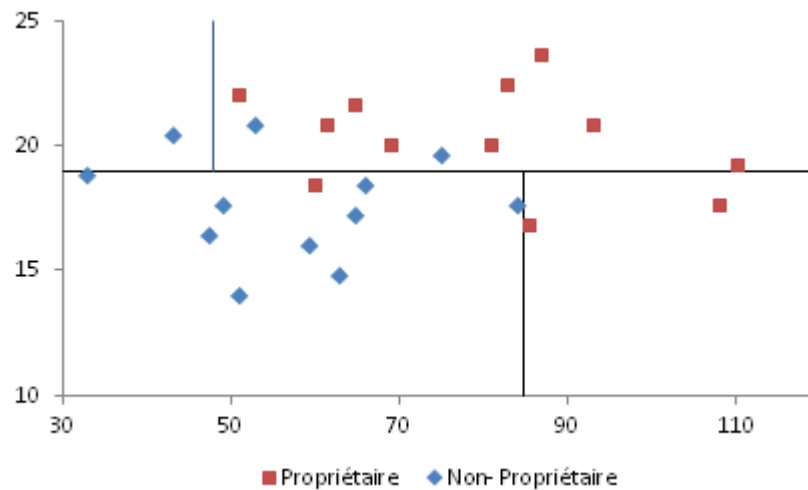
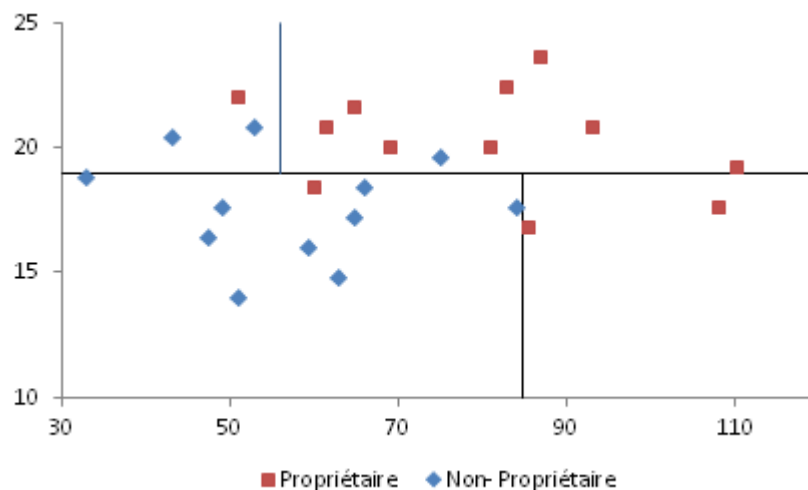


Figure: 57.64 - Troisième itération de la classification

avec (vous pouvez faire le calcul) une impureté totale de 0.2727. Alors qu'en réalité, la discrimination optimale est:



avec une impureté totale de 0.2592. Effectivement:

$$I(A_{gauche}) = 1 - \sum_{k=1}^2 p_k^2 = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 \right] = 0.4$$

$$I(A_{droite}) = 1 - \sum_{k=1}^2 p_k^2 = 1 - \left[\left(\frac{1}{9}\right)^2 + \left(\frac{8}{9}\right)^2 \right] \cong 0.197$$

(57.690)

L'indice de Gini global est alors:

$$I(A) = \frac{3}{12} \cdot 0.4 + \frac{9}{12} \cdot 0.197 \cong 0.2592$$

(57.691)

À l'étape suivante, nous avons:

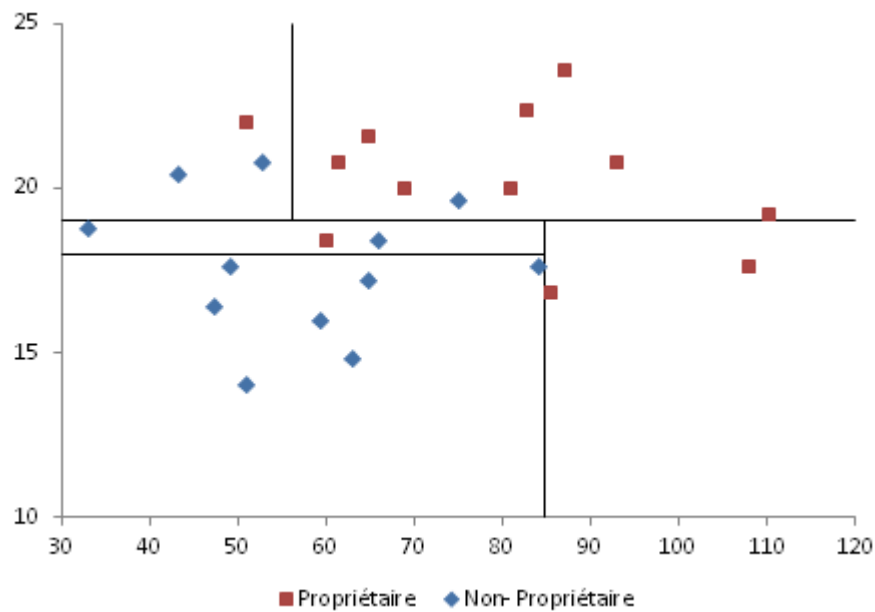


Figure: 57.65 - Quatrième itération de la classification

À l'étape suivante:

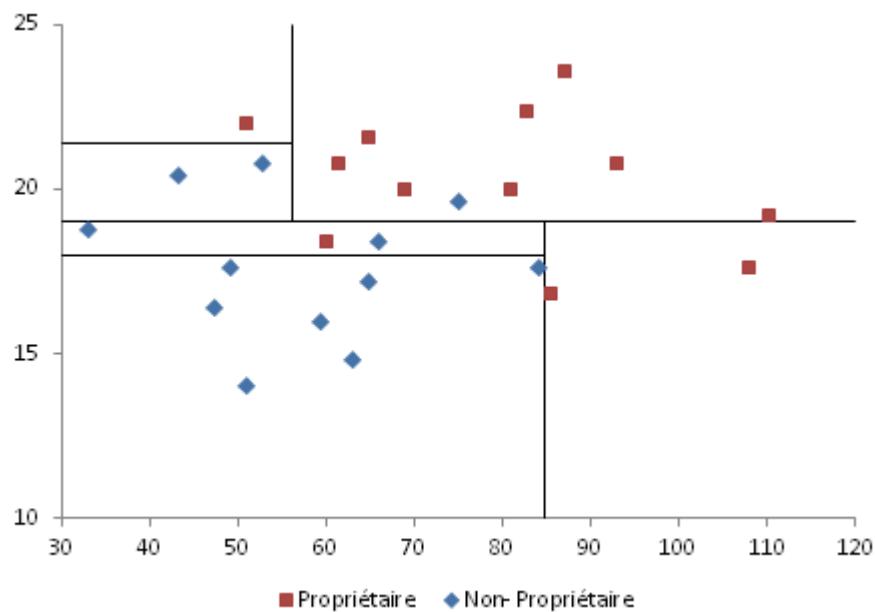


Figure: 57.66 - Cinquième itération de la classification

etc. jusqu'à obtenir au final:

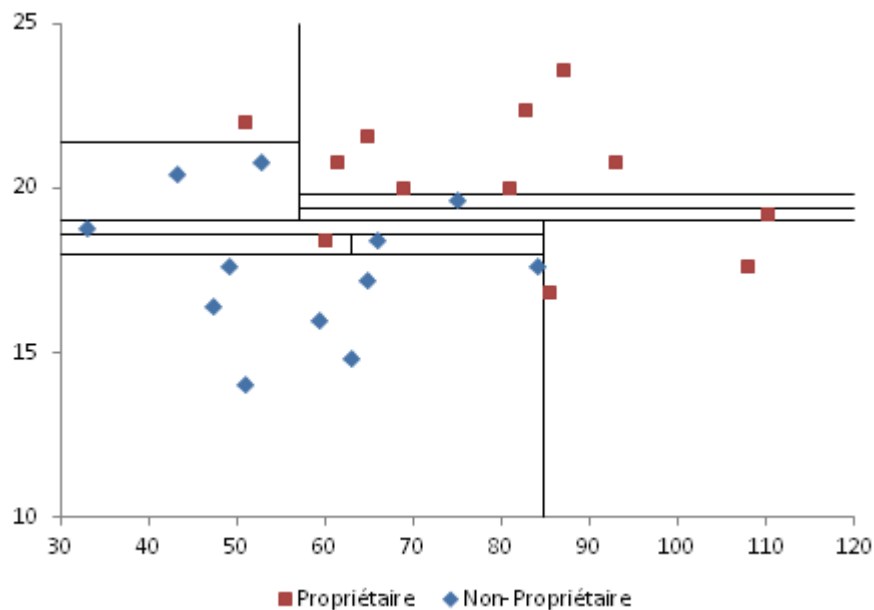


Figure: 57.67 - Résultat final de la classification

où chaque rectangle est pur (ne contient des données que d'une des deux classes).

La raison pour laquelle la méthode est appelée algorithme d'arbre de classification est que chaque division peut être figurée comme la division d'un nœud en deux nœuds successeurs. La première division est montrée comme un branchement du nœud racine de l'arbre. Voici par exemple (faute de place sur la page) les six premières itérations de l'algorithme.

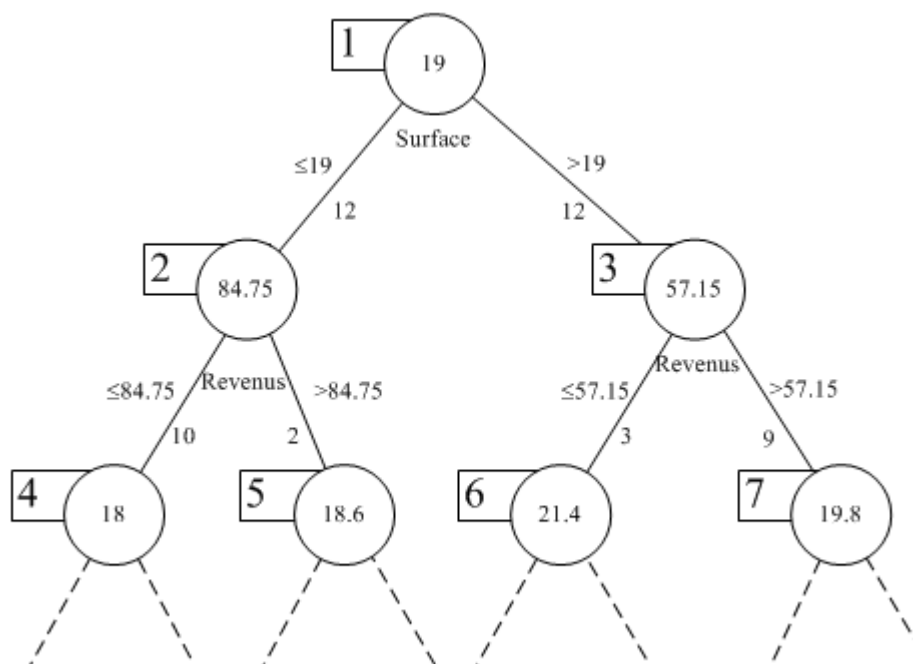


Figure: 57.68 - Représentation du résultat sous forme d'arbre/organigramme

et comme mentionné plus haut, nous nous arrêtons ici puisque l'ensemble des techniques empiriques forment un métier à part entier tellement elles sont nombreuses (et presque aucune ne donne le même résultat). Indiquons enfin qu'il existe également des réseaux de neurones (voir ci-après) spécialisés dans la classification.

26.2. K-MEANS

L'algorithme des K-moyennes ou "**K-means**" en anglais, appelé aussi parfois "**classification avec méthode des nuées dynamiques**", est en statistiques et en apprentissage automatique (plus précisément en apprentissage non supervisé), un algorithme de partitionnement de données, c'est-à-dire une méthode dont le but est de diviser des observations en K partitions (clusters) dans lesquelles chaque observation appartient à la partition avec la moyenne la plus proche.

Les étapes de base de l'algorithme sont les suivantes:

E1. Nous choisissons un partitionnement en K groupes

E2. Nous générons K moyennes (centres) au hasard

E3. Les données sont affectées au groupe dont le centre leur est le plus proche

E4. Nous calculons la moyenne de chaque groupe en utilisant les données affectées

E5. Nous retournons à l'étape E3.

L'algorithme K -means ne converge cependant pas forcément vers une solution optimale. Rappelons qu'un calcul d'optimisation globale est incompatible avec les volumes de données utilisés, quelle que soit la puissance des ordinateurs. Ainsi, les K -means vont utiliser des algorithmes itératifs permettant d'arriver à un optimum local. Ils vont tâtonner pour minimiser des matrices de covariances intra-classes. C'est aussi un algorithme qui va tenter de trouver les meilleurs K points initiaux.

Certains logiciels vous donnent la possibilité de paramétrer les valeurs initiales et cela va influencer la qualité finale de la typologie, tout en sachant qu'il n'existe pas UN bon choix initial. Cela varie en fonction de la configuration des données et même du hasard...

Il existe trois solutions fréquentes:

- 1ère solution: Le logiciel les détermine les K points initiaux aléatoirement. Il peut procéder à un certain nombre d'essais et il choisira le plus concluant.

- 2ème solution: L'avis d'expert, suppose que quelqu'un ait une assez bonne connaissance de la population étudiée pour rattacher à chaque classe un type idéal. Ce dernier peut être ou non un individu réel.

- 3ème solution: Le logiciel répartit les K points initiaux non aléatoirement mais selon certains algorithmes empiriques.

Plutôt que de présenter l'algorithme avec des équations mathématiques, il nous a semblé plus pédagogique de montrer comment implémenter cette technique dans un tableur comme Microsoft Excel 14.0.6123 car l'expérience nous montre que cela est beaucoup plus parlant et efficace.

Pour cela, nous allons considérer d'abord la structure suivante dont l'idée est de déterminer 3 centres (donc il s'agit d'un 3-means):

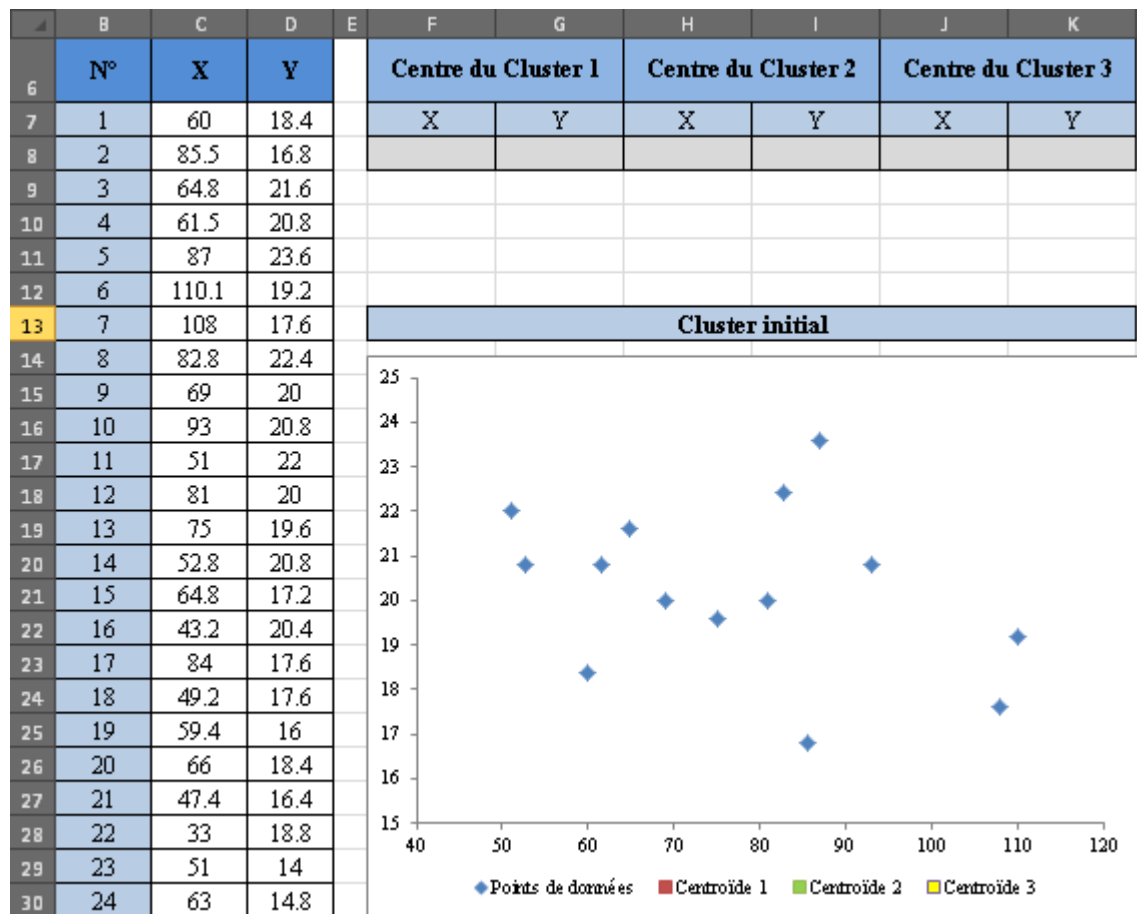


Figure: 57.69 - Structure de départ de base avec Microsoft Excel 14.0.6123

où nous avons à gauche des données d'une population basées sur caractéristiques X et Y avec un petit tableau qui affichera les coordonnées des trois centroïdes. Nous créons sur la même feuille le tableau suivant:

	M	N	O	P	Q	R
6	N°	Distance au centroïde 1	Distance au centroïde 2	Distance au centroïde 3	Classe	Distance
7	1	62.76	62.76	62.76	Cluster1	62.76
8	2	87.13	87.13	87.13	Cluster1	87.13
9	3	68.31	68.31	68.31	Cluster1	68.31
10	4	64.92	64.92	64.92	Cluster1	64.92
11	5	90.14	90.14	90.14	Cluster1	90.14
12	6	111.76	111.76	111.76	Cluster1	111.76
13	7	109.42	109.42	109.42	Cluster1	109.42
14	8	85.78	85.78	85.78	Cluster1	85.78
15	9	71.84	71.84	71.84	Cluster1	71.84
16	10	95.30	95.30	95.30	Cluster1	95.30
17	11	55.54	55.54	55.54	Cluster1	55.54
18	12	83.43	83.43	83.43	Cluster1	83.43
19	13	77.52	77.52	77.52	Cluster1	77.52
20	14	56.75	56.75	56.75	Cluster1	56.75
21	15	67.04	67.04	67.04	Cluster1	67.04
22	16	47.77	47.77	47.77	Cluster1	47.77
23	17	85.82	85.82	85.82	Cluster1	85.82
24	18	52.25	52.25	52.25	Cluster1	52.25
25	19	61.52	61.52	61.52	Cluster1	61.52
26	20	68.52	68.52	68.52	Cluster1	68.52
27	21	50.16	50.16	50.16	Cluster1	50.16
28	22	37.98	37.98	37.98	Cluster1	37.98
29	23	52.89	52.89	52.89	Cluster1	52.89
30	24	64.72	64.72	64.72	Cluster1	64.72
31		Somme des distances				1'120.61

Figure: 57.70 - Tableau principal dans Microsoft Excel 14.0.6123 pour l'algorithme

avec les formules triviales pour les trois colonnes N, O, P où l'on retrouve la norme de la distance euclidienne:

	N	O	P
6	Distance au centroïde 1	Distance au centroïde 2	Distance au centroïde 3
7	=RACINE((C7-\$F\$8)^2+(D7-\$G\$8)^2)	=RACINE((C7-\$H\$8)^2+(D7-\$I\$8)^2)	=RACINE((C7-\$J\$8)^2+(D7-\$K\$8)^2)
8	=RACINE((C8-\$F\$8)^2+(D8-\$G\$8)^2)	=RACINE((C8-\$H\$8)^2+(D8-\$I\$8)^2)	=RACINE((C8-\$J\$8)^2+(D8-\$K\$8)^2)
9	=RACINE((C9-\$F\$8)^2+(D9-\$G\$8)^2)	=RACINE((C9-\$H\$8)^2+(D9-\$I\$8)^2)	=RACINE((C9-\$J\$8)^2+(D9-\$K\$8)^2)
10	=RACINE((C10-\$F\$8)^2+(D10-\$G\$8)^2)	=RACINE((C10-\$H\$8)^2+(D10-\$I\$8)^2)	=RACINE((C10-\$J\$8)^2+(D10-\$K\$8)^2)
11	=RACINE((C11-\$F\$8)^2+(D11-\$G\$8)^2)	=RACINE((C11-\$H\$8)^2+(D11-\$I\$8)^2)	=RACINE((C11-\$J\$8)^2+(D11-\$K\$8)^2)
12	=RACINE((C12-\$F\$8)^2+(D12-\$G\$8)^2)	=RACINE((C12-\$H\$8)^2+(D12-\$I\$8)^2)	=RACINE((C12-\$J\$8)^2+(D12-\$K\$8)^2)
13	=RACINE((C13-\$F\$8)^2+(D13-\$G\$8)^2)	=RACINE((C13-\$H\$8)^2+(D13-\$I\$8)^2)	=RACINE((C13-\$J\$8)^2+(D13-\$K\$8)^2)
14	=RACINE((C14-\$F\$8)^2+(D14-\$G\$8)^2)	=RACINE((C14-\$H\$8)^2+(D14-\$I\$8)^2)	=RACINE((C14-\$J\$8)^2+(D14-\$K\$8)^2)
15	=RACINE((C15-\$F\$8)^2+(D15-\$G\$8)^2)	=RACINE((C15-\$H\$8)^2+(D15-\$I\$8)^2)	=RACINE((C15-\$J\$8)^2+(D15-\$K\$8)^2)
16	=RACINE((C16-\$F\$8)^2+(D16-\$G\$8)^2)	=RACINE((C16-\$H\$8)^2+(D16-\$I\$8)^2)	=RACINE((C16-\$J\$8)^2+(D16-\$K\$8)^2)
17	=RACINE((C17-\$F\$8)^2+(D17-\$G\$8)^2)	=RACINE((C17-\$H\$8)^2+(D17-\$I\$8)^2)	=RACINE((C17-\$J\$8)^2+(D17-\$K\$8)^2)
18	=RACINE((C18-\$F\$8)^2+(D18-\$G\$8)^2)	=RACINE((C18-\$H\$8)^2+(D18-\$I\$8)^2)	=RACINE((C18-\$J\$8)^2+(D18-\$K\$8)^2)
19	=RACINE((C19-\$F\$8)^2+(D19-\$G\$8)^2)	=RACINE((C19-\$H\$8)^2+(D19-\$I\$8)^2)	=RACINE((C19-\$J\$8)^2+(D19-\$K\$8)^2)
20	=RACINE((C20-\$F\$8)^2+(D20-\$G\$8)^2)	=RACINE((C20-\$H\$8)^2+(D20-\$I\$8)^2)	=RACINE((C20-\$J\$8)^2+(D20-\$K\$8)^2)
21	=RACINE((C21-\$F\$8)^2+(D21-\$G\$8)^2)	=RACINE((C21-\$H\$8)^2+(D21-\$I\$8)^2)	=RACINE((C21-\$J\$8)^2+(D21-\$K\$8)^2)
22	=RACINE((C22-\$F\$8)^2+(D22-\$G\$8)^2)	=RACINE((C22-\$H\$8)^2+(D22-\$I\$8)^2)	=RACINE((C22-\$J\$8)^2+(D22-\$K\$8)^2)
23	=RACINE((C23-\$F\$8)^2+(D23-\$G\$8)^2)	=RACINE((C23-\$H\$8)^2+(D23-\$I\$8)^2)	=RACINE((C23-\$J\$8)^2+(D23-\$K\$8)^2)
24	=RACINE((C24-\$F\$8)^2+(D24-\$G\$8)^2)	=RACINE((C24-\$H\$8)^2+(D24-\$I\$8)^2)	=RACINE((C24-\$J\$8)^2+(D24-\$K\$8)^2)
25	=RACINE((C25-\$F\$8)^2+(D25-\$G\$8)^2)	=RACINE((C25-\$H\$8)^2+(D25-\$I\$8)^2)	=RACINE((C25-\$J\$8)^2+(D25-\$K\$8)^2)
26	=RACINE((C26-\$F\$8)^2+(D26-\$G\$8)^2)	=RACINE((C26-\$H\$8)^2+(D26-\$I\$8)^2)	=RACINE((C26-\$J\$8)^2+(D26-\$K\$8)^2)
27	=RACINE((C27-\$F\$8)^2+(D27-\$G\$8)^2)	=RACINE((C27-\$H\$8)^2+(D27-\$I\$8)^2)	=RACINE((C27-\$J\$8)^2+(D27-\$K\$8)^2)
28	=RACINE((C28-\$F\$8)^2+(D28-\$G\$8)^2)	=RACINE((C28-\$H\$8)^2+(D28-\$I\$8)^2)	=RACINE((C28-\$J\$8)^2+(D28-\$K\$8)^2)
29	=RACINE((C29-\$F\$8)^2+(D29-\$G\$8)^2)	=RACINE((C29-\$H\$8)^2+(D29-\$I\$8)^2)	=RACINE((C29-\$J\$8)^2+(D29-\$K\$8)^2)
30	=RACINE((C30-\$F\$8)^2+(D30-\$G\$8)^2)	=RACINE((C30-\$H\$8)^2+(D30-\$I\$8)^2)	=RACINE((C30-\$J\$8)^2+(D30-\$K\$8)^2)
31	Somme des d		

Figure: 57.71 - Formules correspondantes à la figure précédente

et les formules suivantes pour les deux colonnes restantes:

	Q	R
6	Classe	Distance
7	=SI(MIN(N7:P7)=N7;"Cluster1";SI(MIN(N7:P7)=O7;"Cluster2";"Cluster3"))	=MIN(N7:P7)
8	=SI(MIN(N8:P8)=N8;"Cluster1";SI(MIN(N8:P8)=O8;"Cluster2";"Cluster3"))	=MIN(N8:P8)
9	=SI(MIN(N9:P9)=N9;"Cluster1";SI(MIN(N9:P9)=O9;"Cluster2";"Cluster3"))	=MIN(N9:P9)
10	=SI(MIN(N10:P10)=N10;"Cluster1";SI(MIN(N10:P10)=O10;"Cluster2";"Cluster3"))	=MIN(N10:P10)
11	=SI(MIN(N11:P11)=N11;"Cluster1";SI(MIN(N11:P11)=O11;"Cluster2";"Cluster3"))	=MIN(N11:P11)
12	=SI(MIN(N12:P12)=N12;"Cluster1";SI(MIN(N12:P12)=O12;"Cluster2";"Cluster3"))	=MIN(N12:P12)
13	=SI(MIN(N13:P13)=N13;"Cluster1";SI(MIN(N13:P13)=O13;"Cluster2";"Cluster3"))	=MIN(N13:P13)
14	=SI(MIN(N14:P14)=N14;"Cluster1";SI(MIN(N14:P14)=O14;"Cluster2";"Cluster3"))	=MIN(N14:P14)
15	=SI(MIN(N15:P15)=N15;"Cluster1";SI(MIN(N15:P15)=O15;"Cluster2";"Cluster3"))	=MIN(N15:P15)
16	=SI(MIN(N16:P16)=N16;"Cluster1";SI(MIN(N16:P16)=O16;"Cluster2";"Cluster3"))	=MIN(N16:P16)
17	=SI(MIN(N17:P17)=N17;"Cluster1";SI(MIN(N17:P17)=O17;"Cluster2";"Cluster3"))	=MIN(N17:P17)
18	=SI(MIN(N18:P18)=N18;"Cluster1";SI(MIN(N18:P18)=O18;"Cluster2";"Cluster3"))	=MIN(N18:P18)
19	=SI(MIN(N19:P19)=N19;"Cluster1";SI(MIN(N19:P19)=O19;"Cluster2";"Cluster3"))	=MIN(N19:P19)
20	=SI(MIN(N20:P20)=N20;"Cluster1";SI(MIN(N20:P20)=O20;"Cluster2";"Cluster3"))	=MIN(N20:P20)
21	=SI(MIN(N21:P21)=N21;"Cluster1";SI(MIN(N21:P21)=O21;"Cluster2";"Cluster3"))	=MIN(N21:P21)
22	=SI(MIN(N22:P22)=N22;"Cluster1";SI(MIN(N22:P22)=O22;"Cluster2";"Cluster3"))	=MIN(N22:P22)
23	=SI(MIN(N23:P23)=N23;"Cluster1";SI(MIN(N23:P23)=O23;"Cluster2";"Cluster3"))	=MIN(N23:P23)
24	=SI(MIN(N24:P24)=N24;"Cluster1";SI(MIN(N24:P24)=O24;"Cluster2";"Cluster3"))	=MIN(N24:P24)
25	=SI(MIN(N25:P25)=N25;"Cluster1";SI(MIN(N25:P25)=O25;"Cluster2";"Cluster3"))	=MIN(N25:P25)
26	=SI(MIN(N26:P26)=N26;"Cluster1";SI(MIN(N26:P26)=O26;"Cluster2";"Cluster3"))	=MIN(N26:P26)
27	=SI(MIN(N27:P27)=N27;"Cluster1";SI(MIN(N27:P27)=O27;"Cluster2";"Cluster3"))	=MIN(N27:P27)
28	=SI(MIN(N28:P28)=N28;"Cluster1";SI(MIN(N28:P28)=O28;"Cluster2";"Cluster3"))	=MIN(N28:P28)
29	=SI(MIN(N29:P29)=N29;"Cluster1";SI(MIN(N29:P29)=O29;"Cluster2";"Cluster3"))	=MIN(N29:P29)
30	=SI(MIN(N30:P30)=N30;"Cluster1";SI(MIN(N30:P30)=O30;"Cluster2";"Cluster3"))	=MIN(N30:P30)
31	s distances	=SOMME(R7:R20)

Figure: 57.72 - Formules pour l'assignation au cluster et la minimisation des distances

Ensuite, nous lançons le solveur de Microsoft Excel 14.0.6123 avec les paramètres suivants en faisant bien attention à prendre l'algorithme évolutionnaire ce qui fait que nous assumons qu'à chaque fois que nous pourrions avoir un résultat différent:

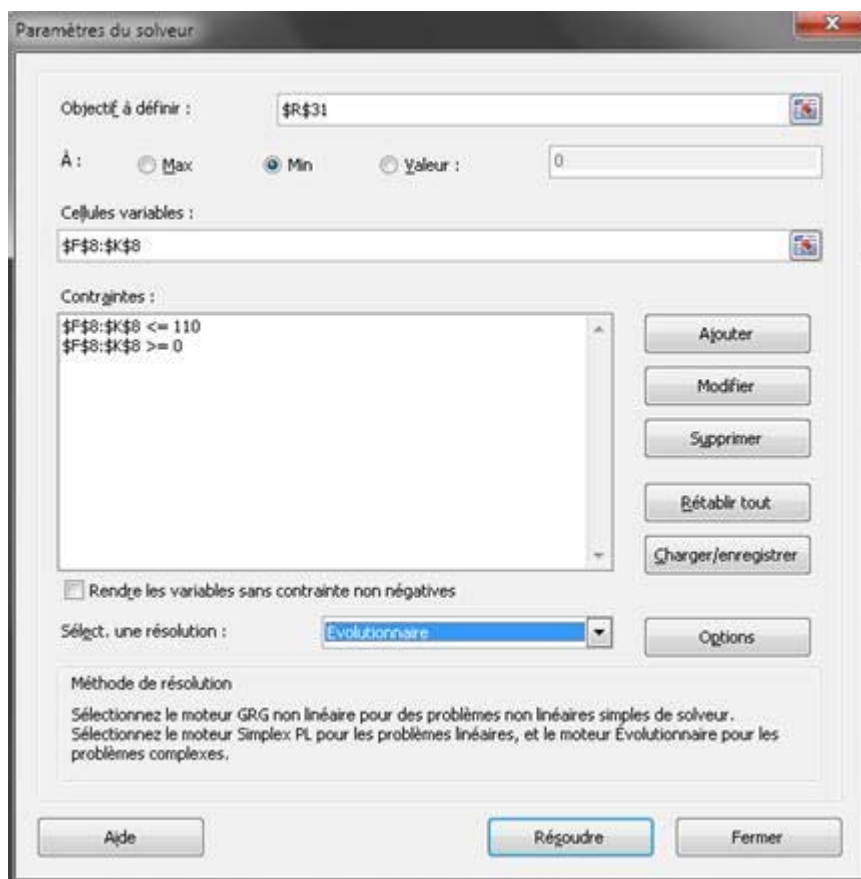
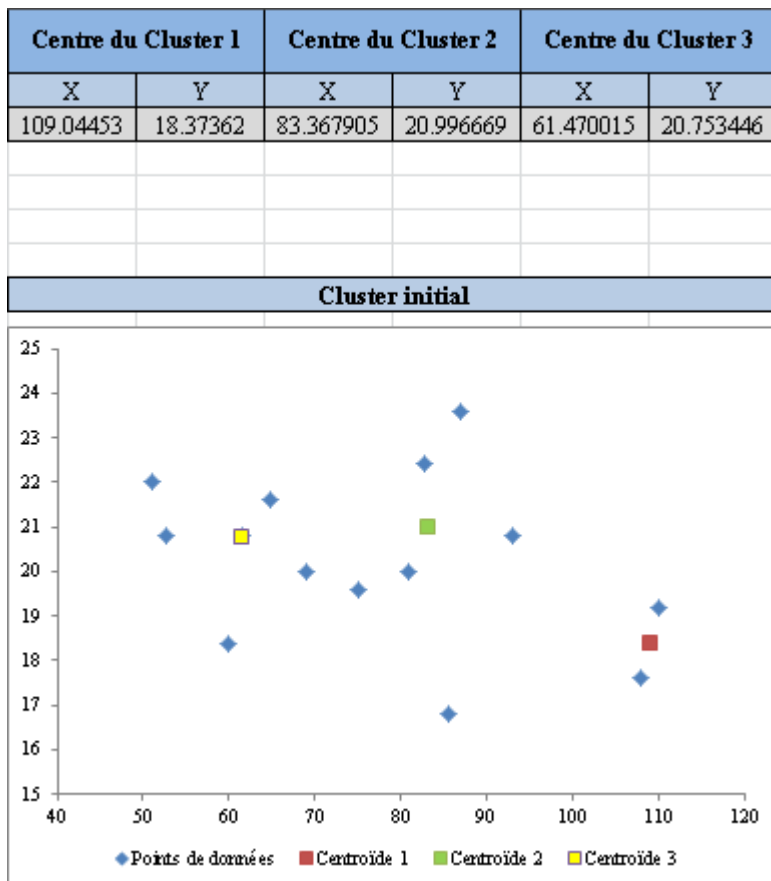


Figure: 57.73 - Paramétrages du solveur de Microsoft Excel 14.0.6123

et nous avons en lançant le solveur:



avec le tableau:

N°	Distance au centroïde 1	Distance au centroïde 2	Distance au centroïde 2	CLASS	Minimum Distance
1	49.04	23.51	2.77	Cluster3	2.77
2	23.60	4.71	24.35	Cluster2	4.71
3	44.36	18.58	3.44	Cluster3	3.44
4	47.61	21.87	0.06	Cluster3	0.06
5	22.66	4.47	25.69	Cluster2	4.47
6	1.34	26.79	48.65	Cluster1	1.34
7	1.30	24.87	46.64	Cluster1	1.30
8	26.55	1.51	21.39	Cluster2	1.51
9	40.08	14.40	7.57	Cluster3	7.57
10	16.23	9.63	31.53	Cluster2	9.63
11	58.16	32.38	10.54	Cluster3	10.54
12	28.09	2.57	19.54	Cluster2	2.57
13	34.07	8.48	13.58	Cluster2	8.48
14	56.30	30.57	8.67	Cluster3	8.67
15	44.26	18.95	4.87	Cluster3	4.87
16	65.88	40.17	18.27	Cluster3	18.27
17	25.06	3.45	22.75	Cluster2	3.45
18	59.85	34.34	12.67	Cluster3	12.67
19	49.70	24.48	5.18	Cluster3	5.18
20	43.04	17.56	5.10	Cluster3	5.10
21	61.68	36.26	14.73	Cluster3	14.73
22	76.05	50.42	28.54	Cluster3	28.54
23	58.21	33.12	12.46	Cluster3	12.46
24	46.18	21.29	6.15	Cluster3	6.15
				Sum of Minimum distances	67.06

Figure: 57.75 - Tableau principal correspondant au résultat obtenu après exécution du solveur

Des logiciels spécialisés de statistiques comme Minitab 15.1.1 ou de Data Mining comme Tanagra 1.4.44 donnent en comparaison les valeurs des 3 centres et qui sont une fois reportées dans Microsoft Excel 14.0.6123:

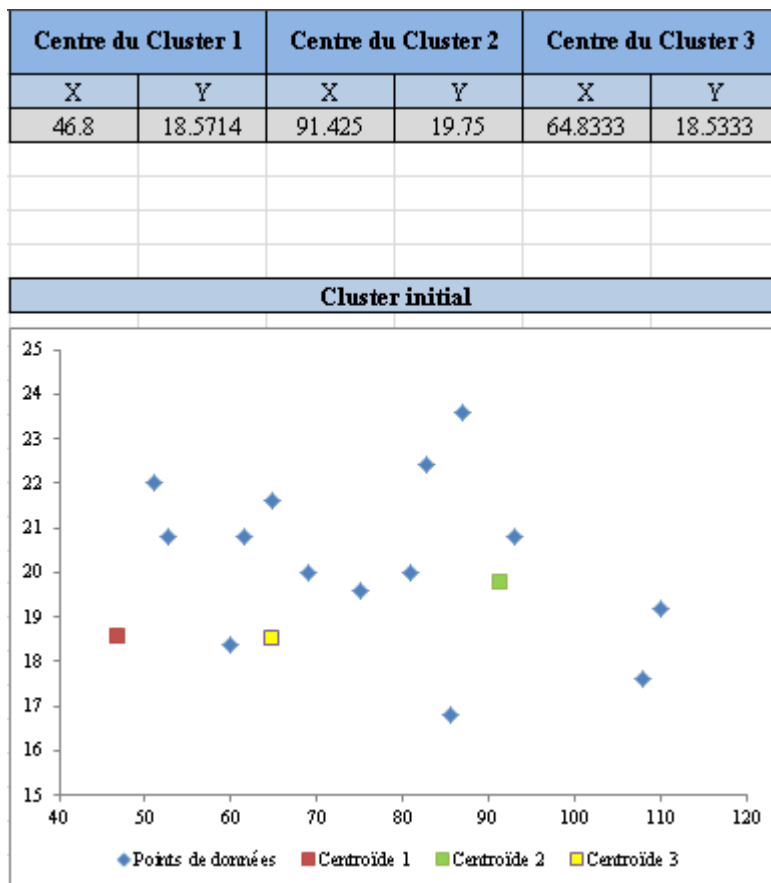


Figure: 57.76 - Report du résultat de Minitab 15.1.1/Tanagra 1.4.44 dans Microsoft Excel 14.0.6123

avec le tableau:

N°	Distance au centroïde 1	Distance au centroïde 2	Distance au centroïde 2	CLASS	Minimum Distance
1	13.20	31.45	4.84	Cluster3	4.84
2	38.74	6.62	20.74	Cluster2	6.62
3	18.25	26.69	3.07	Cluster3	3.07
4	14.87	29.94	4.03	Cluster3	4.03
5	40.51	5.87	22.74	Cluster2	5.87
6	63.30	18.68	45.27	Cluster2	18.68
7	61.21	16.71	43.18	Cluster2	16.71
8	36.20	9.02	18.38	Cluster2	9.02
9	22.25	22.43	4.42	Cluster3	4.42
10	46.25	1.89	28.26	Cluster2	1.89
11	5.42	40.49	14.26	Cluster1	5.42
12	34.23	10.43	16.23	Cluster2	10.43
13	28.22	16.43	10.22	Cluster3	10.22
14	6.40	38.64	12.24	Cluster1	6.40
15	18.05	26.75	1.33	Cluster3	1.33
16	4.04	48.23	21.71	Cluster1	4.04
17	37.21	7.73	19.19	Cluster2	7.73
18	2.59	42.28	15.66	Cluster1	2.59
19	12.86	32.24	5.99	Cluster3	5.99
20	19.20	25.46	1.17	Cluster3	1.17
21	2.25	44.15	17.56	Cluster1	2.25
22	13.80	58.43	31.83	Cluster1	13.80
23	6.21	40.83	14.56	Cluster1	6.21
24	16.63	28.85	4.16	Cluster3	4.16
				Sum of Minimum distances	107.62

Figure: 57.77 - Tableau principal correspondant aux centroïdes obtenus avec Minitab 15.1.1 et Tanagra 1.4.44

La différence s'explique assez simplement! Un logiciel comme Microsoft Excel 14.0.6123 minimise la distance des points aux centres mais est incapable en même temps de maximiser la distance entre les centres. Par contre les logiciels de statistiques ont les algorithmes qu'il faut pour cela.

Minitab 15.1.1 ne donne pas de graphique mais Tanagra 1.4.44 lui donne:

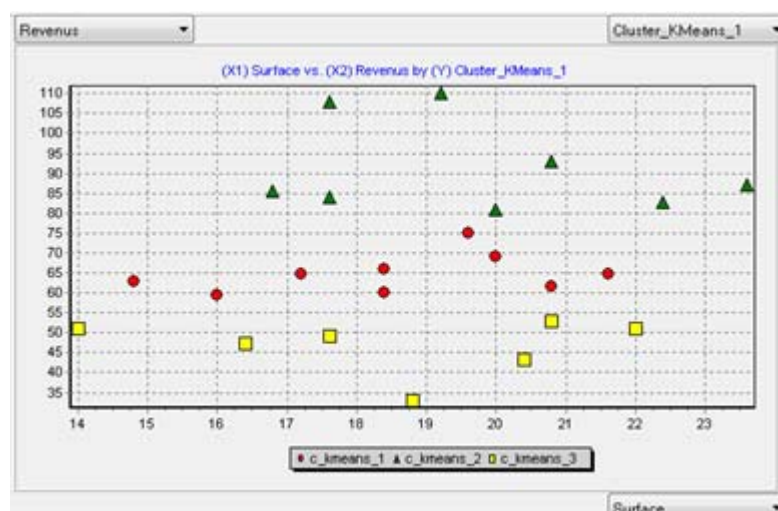


Figure: 57.78 - Graphique obtenu avec Tanagra 1.4.44

Bien évidemment il est possible de trouver des logiciels faisant encore mieux!

26.3. DENDROGRAMME

Pour introduire cette autre technique de clustering, considérons directement l'exemple basé sur le tableau suivant:

	A	B	C	D
1	Observation	Revenus	Surface	Propriétaire
2	1	60	18.4	1
3	2	85.5	16.8	1
4	3	64.8	21.6	1
5	4	61.5	20.8	1
6	5	87	23.6	1
7	6	110.1	19.2	1
8	7	108	17.6	1
9	8	82.8	22.4	1
10	9	69	20	1
11	10	93	20.8	1
12	11	51	22	1
13	12	81	20	1
14	13	75	19.6	2
15	14	52.8	20.8	2
16	15	64.8	17.2	2
17	16	43.2	20.4	2
18	17	84	17.6	2
19	18	49.2	17.6	2
20	19	59.4	16	2
21	20	66	18.4	2
22	21	47.4	16.4	2
23	22	33	18.8	2
24	23	51	14	2
25	24	63	14.8	2

Figure: 57.79 - Données de base pour l'étude des dendrogrammes

Nous souhaitons avoir une organisation hiérarchique de ressemblance des individus en fonction de leurs revenus et de leur surface d'habitation. Une technique consiste à définir une distance pour cette mesure de ressemblance. Par exemple, la distance euclidienne (si chaque individu à deux coordonnées):

$$d_{ij}(x, y) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

est un choix particulier qui permettra d'associer deux individus dont la distance est minimale. Nous parlons alors dans le domaine du clustering de "lien simple".

Ainsi, nous pouvons facilement avec un tableur comme Microsoft Excel 14.0.6123 créer une "**matrice des distances**" qui est une matrice symétrique à diagonale nulle et qui relativement au tableau donné ci-dessus donnera:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
2				60	85.5	64.8	61.5	87	110	108	82.8	69	93	51	81	75	52.8	64.8	43.2	84	49.2	59.4	66	47.4	33	51	63
3				18.4	16.8	21.6	20.8	23.6	19.2	17.6	22.4	20	20.8	22	20	19.6	20.8	17.2	20.4	17.6	17.6	16	18.4	16.4	18.8	14	14.8
4	1	60	18.4	0.00	25.55	5.77	2.83	27.50	50.11	48.01	23.15	9.14	33.09	9.69	21.06	15.05	7.59	4.95	16.92	24.01	10.83	2.47	6.00	12.76	27.00	10.02	4.69
5	2	85.5	16.8	25.55	0.00	21.25	24.33	6.96	24.72	22.51	6.22	16.81	8.50	34.89	5.52	10.87	32.94	20.70	42.45	1.70	36.31	26.11	19.57	38.10	52.54	34.61	22.59
6	3	64.8	21.6	5.77	21.25	0.00	3.40	22.29	45.36	43.38	18.02	4.49	28.21	13.81	16.28	10.39	12.03	4.40	21.63	19.61	16.10	7.78	3.42	18.16	31.92	15.75	7.03
7	4	61.5	20.8	2.83	24.33	3.40	0.00	25.65	48.63	46.61	21.36	7.54	31.50	10.57	19.52	13.55	8.70	4.88	18.30	22.73	12.71	5.24	5.10	14.77	28.57	12.51	6.18
8	5	87	23.6	27.50	6.96	22.29	25.65	0.00	23.52	21.84	4.37	18.36	6.62	36.04	7.00	12.65	34.31	23.10	43.92	6.71	38.27	28.63	21.63	40.25	54.21	37.26	25.56
9	6	110	19.2	50.11	24.72	45.36	48.63	23.52	0.00	2.64	27.49	41.11	17.17	59.17	29.11	35.10	57.32	45.34	66.91	26.15	60.92	50.80	44.11	62.76	77.10	59.33	47.31
10	7	108	17.6	48.01	22.51	43.38	46.61	21.84	2.64	0.00	25.65	39.07	15.34	57.17	27.11	33.06	55.29	43.20	64.86	24.00	58.80	48.63	42.01	60.61	75.01	57.11	45.09
11	8	82.8	22.4	23.15	6.22	18.02	21.36	4.37	27.49	25.65	0.00	14.01	10.32	31.80	3.00	8.29	30.04	18.74	29.65	4.95	33.94	24.26	17.27	35.90	49.93	32.89	21.21
12	9	69	20	9.14	16.81	4.49	7.54	18.36	41.11	39.07	14.01	0.00	24.01	18.11	12.00	6.01	16.22	5.05	25.80	15.19	19.94	10.40	3.40	21.90	36.02	18.97	7.94
13	10	93	20.8	33.09	8.50	28.21	31.50	6.62	17.17	15.34	10.32	24.01	0.00	42.02	12.03	18.04	40.20	28.43	49.80	9.55	43.92	33.94	27.11	45.81	60.03	42.55	30.59
14	11	51	22	9.69	34.89	13.81	10.57	36.04	59.17	57.17	31.80	18.11	42.02	0.00	30.07	24.12	2.16	14.61	7.96	33.29	4.75	10.32	15.43	6.66	18.28	8.00	13.99
15	12	81	20	21.06	5.52	16.28	19.52	7.00	29.11	27.11	3.00	12.00	12.03	30.07	0.00	6.01	28.21	16.44	37.80	3.84	31.89	21.97	15.09	33.79	48.01	30.59	18.74
16	13	75	19.6	15.05	10.87	10.39	13.55	12.65	35.10	33.06	8.29	6.01	18.04	24.12	6.01	0.00	22.23	10.48	31.81	9.22	25.88	16.01	9.08	27.78	42.01	24.64	12.92
17	14	52.8	20.8	7.59	32.94	12.03	8.70	34.31	57.32	55.29	30.04	16.22	40.20	2.16	28.21	22.23	0.00	12.53	9.61	31.36	4.82	8.16	13.42	6.97	19.90	7.03	11.83
18	15	64.8	17.2	4.95	20.70	4.40	4.88	23.10	45.34	43.20	18.74	5.05	28.43	14.61	16.44	10.48	12.53	0.00	21.84	19.20	15.61	5.53	1.70	17.42	31.84	14.17	3.00
19	16	43.2	20.4	16.92	42.45	21.63	18.30	43.92	66.91	64.86	39.65	25.80	49.80	7.96	37.80	31.81	9.61	21.84	0.00	40.90	6.62	16.79	22.89	5.80	10.32	10.09	20.58
20	17	84	17.6	24.01	1.70	19.61	22.73	6.71	26.15	24.00	4.95	15.19	9.55	33.29	3.84	9.22	31.36	19.20	40.90	0.00	34.80	24.65	18.02	36.62	51.01	33.20	21.19
21	18	49.2	17.6	10.83	36.31	16.10	12.71	38.27	60.92	58.80	33.94	19.94	43.92	4.75	31.89	25.88	4.82	15.61	6.62	34.80	0.00	10.32	16.82	2.16	16.24	4.02	14.08
22	19	59.4	16	2.47	26.11	7.78	5.24	28.63	50.80	48.63	24.26	10.40	33.94	10.32	21.97	16.01	8.16	5.53	16.79	24.65	10.32	0.00	7.02	12.01	26.55	8.63	3.79
23	20	66	18.4	6.00	19.57	3.42	5.10	21.63	44.11	42.01	17.27	3.40	27.11	15.43	15.09	9.08	13.42	1.70	22.89	18.02	16.82	7.02	0.00	18.71	33.00	15.63	4.69
24	21	47.4	16.4	12.76	38.10	18.16	14.77	40.25	62.76	60.61	35.90	21.90	45.81	6.66	33.79	27.78	6.97	17.42	5.80	36.62	2.16	12.01	18.71	0.00	14.60	4.33	15.68
25	22	33	18.8	27.00	52.54	31.92	28.57	54.21	77.10	75.01	49.93	36.02	60.03	18.28	48.01	42.01	19.90	31.84	30.32	51.01	16.24	26.55	33.00	14.60	0.00	18.63	30.27
26	23	51	14	10.02	34.61	15.75	12.51	37.26	59.33	57.11	32.89	18.97	42.55	8.00	30.59	24.64	7.03	14.17	10.09	33.20	4.02	8.63	15.63	4.33	18.63	0.00	12.03
27	24	63	14.8	4.69	22.59	7.03	6.18	25.56	47.31	45.09	21.21	7.94	30.59	13.99	18.74	12.92	11.83	3.00	20.58	21.19	14.08	3.79	4.69	15.68	30.27	12.03	0.00

Figure: 57.80 - Tableau des distances euclidiennes avec Microsoft Excel 14.0.6123

où nous avons mis dans la cellule D4 le calcul de la distance euclidienne:

$$=RACINE((\$B4-D$2)^2+(\$C4-D$3)^2)$$

que nous avons ensuite tiré sur tout le reste du tableau jusqu'à la cellule AA27.

Ensuite, nous utilisons la méthode ascendante (agglomération) où nous combinons les groupes jusqu'à ce qu'il n'y ait plus qu'un seul groupe (contenant toutes les données).

Ce qui donnera sous forme tabulaire avec un logiciel comme Minitab 15.1.1.0 (les valeurs n'y sont par arrondies au centièmes contrairement au petit tableau donné précédemment):

	Nombre de groupes	Niveau de similarité	Niveau de distance	Groupes liés	Nouveau groupe	Nombre d'obs. dans le nouveau groupe
Etape						
1	23	97.7989	1.6971	15 20	15	2
2	22	97.7951	1.7000	2 17	2	2
3	21	97.1942	2.1633	11 14	11	2
4	20	97.1942	2.1633	18 21	18	2
5	19	96.7914	2.4739	1 19	1	2
6	18	96.5758	2.6401	6 7	6	2
7	17	96.3292	2.8302	1 4	1	3
8	16	96.1090	3.0000	15 24	15	3
9	15	96.1090	3.0000	8 12	8	2
10	14	95.5959	3.3956	1 3	1	4
11	13	95.5902	3.4000	9 15	9	4
12	12	95.5674	3.4176	1 9	1	8
13	11	95.0171	3.8419	2 8	2	4
14	10	94.7797	4.0249	18 23	18	3
15	9	94.3346	4.3681	2 5	2	5
16	8	93.8341	4.7539	11 18	11	5
17	7	92.4774	5.8000	11 16	11	6
18	6	92.2007	6.0133	2 13	2	6
19	5	92.2007	6.0133	1 2	1	14
20	4	91.4123	6.6212	1 10	1	15
21	3	90.1565	7.5895	1 11	1	21

où le niveau de similarité du groupe lié i, j est défini empiriquement par:

$$S_{ij} = 1 - \frac{d_{ij}(x, y)}{\max(d_{ij}(x, y))}$$

Ainsi, pour la première ligne nous avons par exemple:

$$S_{15,20} = 1 - \frac{d_{15,20}(x, y)}{\max(d_{ij}(x, y))} = 1 - \frac{1.6971}{77.1010} = 97.7989\%$$

La liste précédente est plus agréable à analyser si, comme il est d'usage, nous la représentons sous forme de "dendrogramme" (diagramme d'arrangement) suivant:

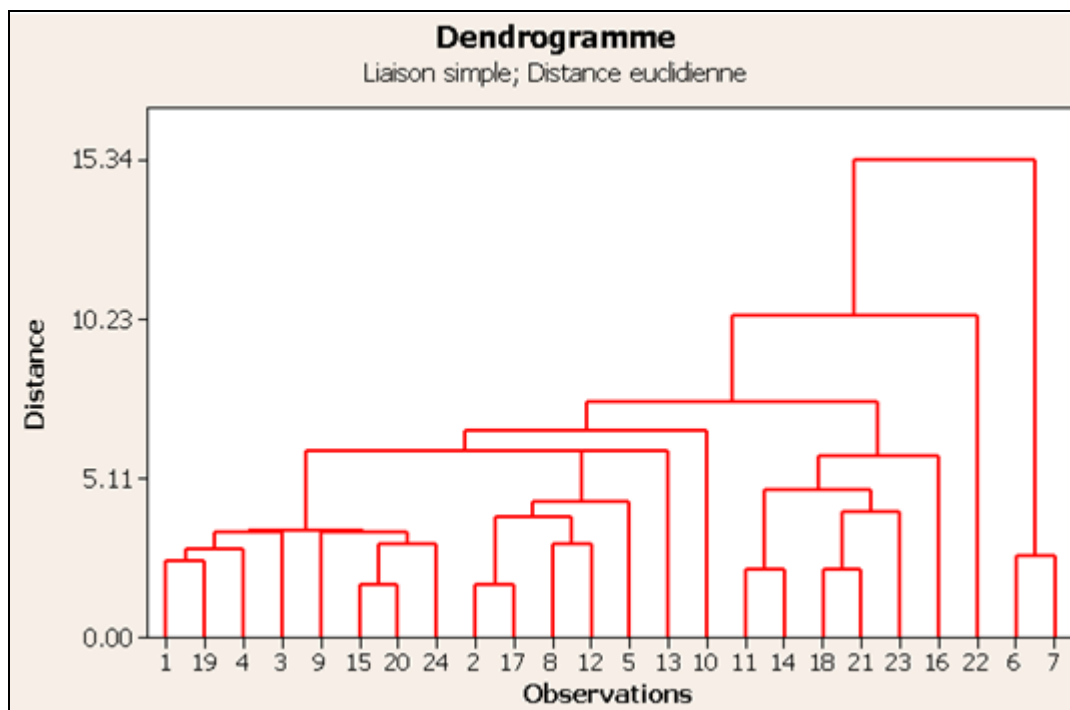


Figure: 57.81 - Graphique obtenu avec Minitab 15.1.1.0

27. RÉSEAUX DE NEURONES FORMELS

Les réseaux de neurones, fabriqués à partir de structures cellulaires artificielles, constituent une approche permettant d'aborder sous des angles nouveaux les problèmes de perception, de mémoire, d'apprentissage et de raisonnement non-linéaires (en d'autres termes... d'intelligence artificielle ou en abrégé "I.A.") au même titre que les algorithmes génétiques. Ils s'avèrent aussi des alternatives très prometteuses pour contourner certaines des limitations des méthodes numériques classiques. Grâce à leur traitement en parallèle de l'information et à leurs mécanismes inspirés des cellules nerveuses (neurones), ils infèrent des propriétés émergentes permettant de solutionner des problèmes jadis qualifiés de complexes.

Cependant, le problème majeur des neurones artificiels (à ma connaissance...), c'est qu'ils ne s'auto-organisent pas ni ne s'auto-structurent de manière intelligente. Il faut donc à ce jour, procéder de manière heuristique pour trouver la meilleure structure de réseau de neurones adaptée à un problème et c'est là leur grande faiblesse actuelle (on utilise soit la force brute via une base de données contenant des millions de modèles, soit les algorithmes génétiques que nous verrons un peu plus loin).

Nous aborderons ici les principales architectures des réseaux de neurones. Il ne s'agit pas de les étudier toutes, car elles sont trop nombreuses, mais plutôt d'en comprendre les mécanismes internes fondamentaux et de savoir comment et quand les utiliser. Nous aborderons également certaines notions relatives aux ensembles flous et à la logique (cf. chapitre de Logique Floue) dans la mesure où ces dernières sont incorporées dans certaines architectures de réseaux de neurones que nous étudierons.

Le cerveau humain contient environ 100 milliards de neurones. Ces neurones nous permettent entre autres, de lire un texte tout en maintenant une respiration régulière permettant d'oxygéner notre sang, en actionnant notre cœur qui assure une circulation efficace de ce sang pour nourrir nos cellules, etc. Ils nous permettent même de comprendre certaines idées (...)

Chacun de ces neurones est par ailleurs fort complexe. Essentiellement, il s'agit de tissu vivant et de chimie. Les neuro-physiciens commencent à peine à comprendre quelques-uns de leurs mécanismes internes. Nous pensons en général que leurs différentes fonctions neuronales, y compris celle de la mémoire sont stockées au niveau des connexions (synapses) entre les neurones. C'est ce genre de théorie qui a inspiré la plupart des architectures de réseaux de neurones artificiels (dits "formels"). L'apprentissage consiste alors soit à établir de nouvelles connexions, soit à en modifier des existantes (nous nous concentrerons en particulier sur cette dernière possibilité).

Ceci nous amène à poser une question fondamentale: en se basant sur nos connaissances actuelles, peut-on construire des modèles approximatifs de neurones et les entraîner pour, éventuellement, réaliser des tâches utiles ? Eh bien, la réponse courte est: oui !, même si les réseaux que nous allons développer ne possèdent qu'une infime fraction de la puissance du cerveau humain, et c'est l'objectif ici de montrer comment nous pouvons y arriver.

Les réseaux de neurones servent aujourd'hui à toutes sortes d'application dans divers domaines. Par exemple, il existe des réseaux de neurones développés pour les autopilotes d'avions, ou encore pour les systèmes de guidage pour automobiles, pour les systèmes de lecture automatique de chèques bancaires et d'adresses postales, pour le traitement du signal pour différentes applications militaires, pour la synthèse de la parole, pour bâtir des systèmes de vision par ordinateur, pour faire des prévisions sur les marchés monétaires, pour évaluer le risque financier ou en assurance, pour différents processus manufacturiers, pour le diagnostic médical, pour l'exploration pétrolière ou gazière, en robotique, en télécommunication, pour la classification et bien d'autres. Bref, les réseaux de neurones ont aujourd'hui un impact considérable et, il y a fort à parier, que leur importance ira grandissante dans le futur.

27.1. MODÈLE DU NEURONE

Le modèle mathématique d'un neurone artificiel, ou "**perceptron**", est illustré à la figure ci-dessous. Un neurone est essentiellement constitué d'un intégrateur qui effectue la somme pondérée de ses entrées (comme l'espérance statistique!). Le résultat n de cette somme est ensuite transformé par une fonction de transfert f qui produit la sortie a du neurone.

Les R entrées du neurone correspondent au vecteur noté traditionnellement en ligne (mais dans la pratique on représente sa transposée):

$$\vec{p} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_R \end{pmatrix} \quad (57.692)$$

alors que:

$$\vec{w} = \begin{pmatrix} w_{1,1} \\ w_{1,2} \\ \vdots \\ w_{1,R} \end{pmatrix} \quad (57.693)$$

représente le vecteur des poids du neurone (nous les distinguons pour préparer le terrain à des neurones un peu plus complexes).

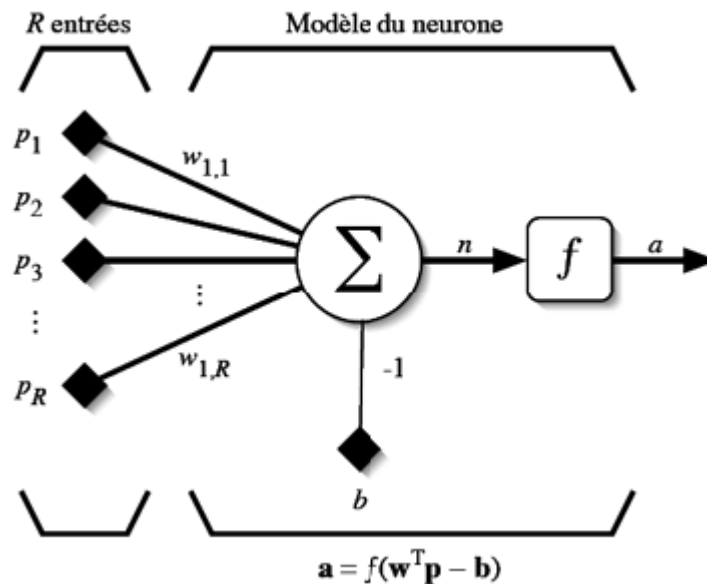


Figure: 57.82 - Exemple de neurone formel à une couche avec vecteur d'entrée et scalaire de sortie

La sortie n de l'intégrateur est définie (car il s'agit d'une technique de l'ingénieur) par l'équation suivante:

$$n = \sum_{j=1}^R w_{1,j} p_j - b = w_{1,1} p_1 + w_{1,2} p_2 + \dots + w_{1,R} p_R - b \quad (57.694)$$

que nous pouvons aussi écrire sous forme matricielle (on pourrait aussi l'écrire sous forme tensorielle mais bon...):

$$n = \vec{w}^T \vec{p} - b \quad (57.695)$$

Cette sortie correspond à une somme pondérée des poids et des entrées moins que ce nous nommons "**le biais b du neurone**" (facteur correctif décidé par tâtonnement et souvent nul dans la pratique). La somme pondérée s'appelle le "**niveau d'activation du neurone**". Le biais b s'appelle aussi le "**seuil d'activation du neurone**". Lorsque le niveau d'activation atteint ou dépasse le seuil b , alors n , l'argument de f , devient nul ou bien évidemment positif. Sinon, il est négatif.

Nous pouvons faire un parallèle entre ce modèle mathématique et certaines informations que nous connaissons (ou que nous croyons connaître) à propos du neurone biologique. Ce dernier possède trois principales composantes: les dendrites, le corps cellulaire et l'axone:

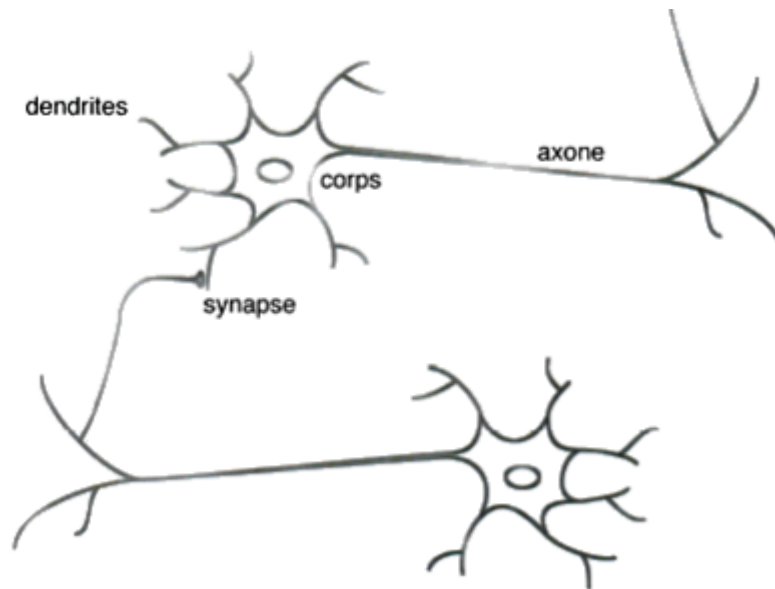


Figure: 57.83 - Représentation simplifiée du vocabulaire des neurones humains

Les dendrites forment un maillage de récepteurs nerveux qui permettent d'acheminer vers le corps du neurone des signaux électriques en provenance d'autres neurones. Celui-ci agit comme une espèce d'intégrateur en accumulant des charges électriques. Lorsque le neurone devient suffisamment excité (lorsque la charge accumulée dépasse un certain seuil), par un processus électrochimique, il engendre un potentiel électrique qui se propage à travers son axone pour éventuellement venir exciter d'autres neurones. Le point de contact entre l'axone d'un neurone et la dendrite d'un autre neurone s'appelle le "synapse". Il semble que c'est l'arrangement spatial des neurones et leur axone, ainsi que la qualité des connexions synaptiques individuelles qui déterminent la fonction précise d'un réseau de neurones biologique. C'est en se basant sur ces connaissances que le modèle mathématique décrit ci-dessus a été défini.

Un poids d'un neurone artificiel représente donc en quelque sorte l'efficacité d'une connexion synaptique. Un poids négatif inhibe en quelque sorte une entrée, alors qu'un poids positif vient l'accentuer. Il importe de retenir que ceci est une grossière approximation d'une véritable synapse qui résulte en fait d'un processus chimique très complexe et dépendant de nombreux facteurs extérieurs encore mal connus. Il faut bien comprendre que notre neurone artificiel est un modèle pragmatique qui, comme nous le verrons plus tard, nous permettra d'accomplir des tâches intéressantes. La vraisemblance biologique de ce modèle nous importe peu. Ce qui compte est le résultat que ce modèle nous permettra d'atteindre.

Un autre facteur limitatif dans le modèle que nous nous sommes donnés concerne son caractère discret. En effet, pour pouvoir simuler un réseau de neurones, nous allons rendre le temps discret dans nos équations. Autrement dit, nous allons supposer que tous les neurones sont synchrones, c'est-à-dire qu'à chaque temps t , ils vont simultanément calculer leur somme pondérée et produire une sortie $a(t) = f(n(t))$. Dans les réseaux biologiques, tous les neurones sont en fait asynchrones.

Revenons donc à notre modèle tel que formulé par l'équation précédente et ajoutons la fonction d'activation f pour obtenir la sortie du neurone:

$$a = f(n) = f(\vec{w}^T \vec{p} - b) \quad (57.696)$$

Il est temps maintenant de remplacer (parce que la notation est un peu lourde à la longue) $\vec{w}^T$ par une matrice d'une seule ligne que nous noterons $W = \vec{w}^T$. Nous obtenons alors une forme générale que nous adopterons tout au long de notre étude:

$$a = f(n) = f(W\vec{p} - b) \quad (57.697)$$

Cette équation nous amène à introduire un nouveau schéma plus formel de notre RNF (ou perceptron):

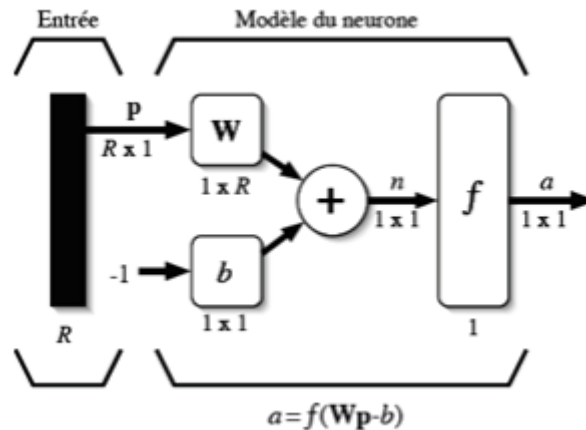


Figure: 57.84 - Exemple plus abouti du neurone formel

Nous y représentons les R entrées comme un rectangle noir (le nombre d'entrées est indiqué sous le rectangle). De ce rectangle sort le vecteur $\vec{p}$ dont la dimension matricielle est $R \times 1$. Ce vecteur est multiplié par une matrice W qui contient les poids (synaptiques) du neurone. Dans le cas d'un neurone simple, cette matrice possède la dimension $1 \times R$. Le résultat de la multiplication correspond au niveau d'activation qui est ensuite comparé au seuil b (un scalaire) par soustraction. Finalement, la sortie du neurone est calculée par la fonction f . La sortie d'un neurone simple est alors toujours un scalaire.

Pour trouver les composantes de la matrice W (poids d'entrée du neurone), ainsi que le biais b nous utilisons des techniques de recherche opérationnelle (méthode du simplexe, méthode des gradients conjugués, algorithmes évolutionnaires, etc.) sur un échantillon des données de l'entreprise de taille n afin "d'entraîner le modèle du neurone". Il s'agira à l'aide de ce modèle de minimiser l'erreur donnée par:

$$\min_{w,b} \sum_{k=1}^n \left(a(x^k) - y^k \right)^2 \quad (57.698)$$

et ensuite de tester le résultat obtenu sur un échantillon test avant de faire de la simulation pour des données non encore existantes (nous ferons un exemple détaillé avec Microsoft Excel juste après avoir présenté les fonctions de transfert).

27.2. FONCTIONS DE TRANSFERT

Jusqu'à présent, nous n'avons pas spécifié la nature de la fonction d'activation $a = f(n)$ de notre modèle. Il se trouve que plusieurs possibilités existent et celles-ci sont empiriques et à adapter en fonction des situations. Les plus courantes et les plus citées dans la littérature sont énumérées dans la figure ci-dessous:

Nom de la fonction	Relation d'entrée/sortie	Icône
seuil	$a = 0$ si $n < 0$ $a = 1$ si $n \geq 0$	
seuil symétrique	$a = -1$ si $n < 0$ $a = 1$ si $n \geq 0$	
linéaire	$a = n$	
linéaire saturée	$a = 0$ si $n < 0$ $a = n$ si $0 \leq n \leq 1$ $a = 1$ si $n > 1$	
linéaire saturée symétrique	$a = -1$ si $n < -1$ $a = n$ si $-1 \leq n \leq 1$ $a = 1$ si $n > 1$	
linéaire positive	$a = 0$ si $n < 0$ $a = n$ si $n \geq 0$	
sigmoïde	$a = \frac{1}{1+\exp^{-n}}$	
tangente hyperbolique	$a = \frac{e^n - e^{-n}}{e^n + e^{-n}}$	
compétitive	$a = 1$ si n maximum $a = 0$ autrement	

Tableau: 57.45 - Types de fonctions de transfert pour les R.N.F.

Les trois les plus utilisées dans le domaine de l'ingénierie sont les fonctions "**seuil**" (a) (en anglais "hard limit"), "**linéaire**" (b) et "**sigmoïde**" (c) comme représentées ci-dessous:

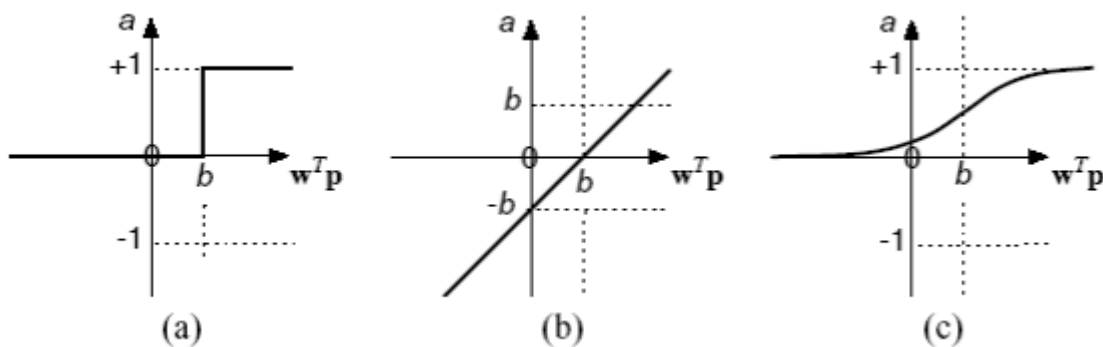


Figure: 57.85 - Fonctions de transfert les plus utilisées à ce jour dans le domaine

Comme son nom l'indique, la fonction seuil applique un seuil sur son entrée. Plus précisément, une entrée négative ne passe pas le seuil, la fonction retourne la valeur 0 (faux), alors qu'une entrée positive ou nulle dépasse le seuil, et la fonction retourne 1 (vrai). Il est évident que ce genre de fonction permet de prendre des décisions binaires (cette fonction peut aussi être assimilée à la fonction de Heaviside pour ceux qui connaissent...).

La fonction linéaire est quant à elle très simple, elle affecte directement son entrée à sa sortie selon la relation $a = f(n) = n$. Il est évident que la sortie du neurone correspond alors à son niveau d'activation dont le passage à zéro (l'ordonnée à l'origine) se produit lorsque $W \cdot \vec{p} = b$.

La fonction de transfert sigmoïde est quant à elle définie par la relation mathématique:

$$\alpha = \frac{1}{1 + e^{-x}} \quad (57.699)$$

elle ressemble soit à la fonction seuil, soit à la fonction linéaire, selon que nous sommes loin ou près de b respectivement. La fonction seuil est très non linéaire car il y a une discontinuité lorsque $W \cdot \vec{p} = b$. De son côté, la fonction linéaire est tout à fait linéaire. Elle ne comporte aucun changement de pente. La sigmoïde est un compromis intéressant entre les deux précédentes. Notons finalement que la fonction "**tangente hyperbolique**" est une version symétrique de la sigmoïde.

27.3. ARCHITECTURE DE RÉSEAU

Par définition, un réseau de neurones est un maillage de plusieurs neurones, généralement organisés en couches. Pour construire une couche de S neurones, il s'agit simplement de les assembler comme dans la figure ci-dessous:

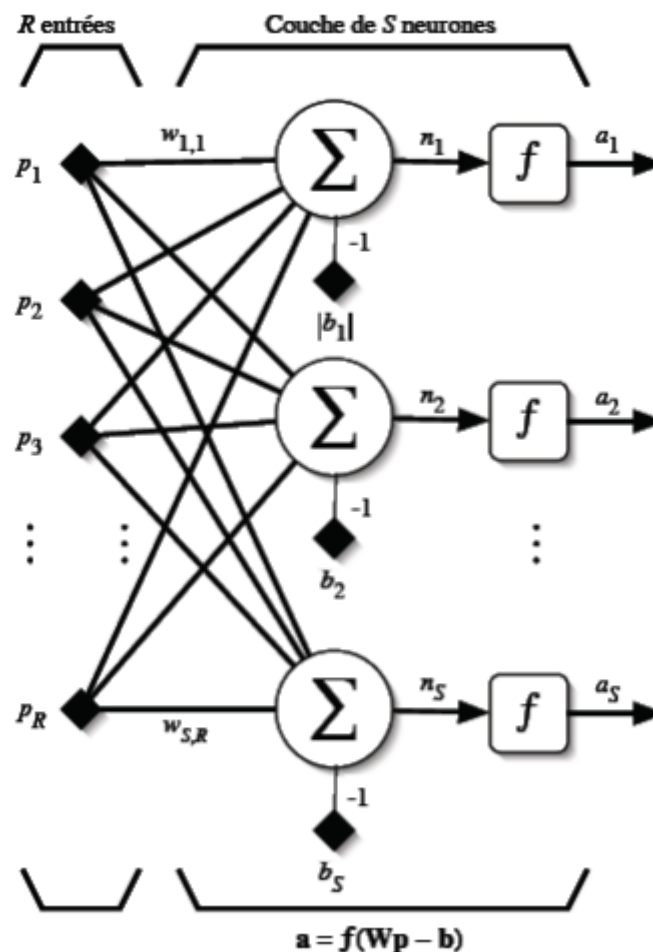


Figure: 57.86 - Illustration d'un réseau de neurones formel

Les S neurones d'une même couche sont tous branchés aux R entrées dans la figure ci-dessous. Nous disons alors que la couche est "**totale-ment connectée**". Mais c'est un cas particulier et non pas une généralité. Souvent, les entrées d'un neurone sont différentes de celles d'un autre neurone, etc.

Un poids $w_{i,j}$ est associé à chacune des connexions. Nous noterons toujours le premier indice par i et le deuxième par j (jamais l'inverse). Le premier indice (rangée) désigne toujours le numéro du neurone sur la couche, alors que le deuxième indice (colonne) spécifie le numéro de l'entrée. Ainsi, $w_{i,j}$ désigne le poids de la connexion qui relie le neurone i à son entrée j . L'ensemble des poids d'une couche forme donc une matrice W de dimension $S \times R$:

$$W = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,R} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,R} \\ \vdots & \vdots & \ddots & \vdots \\ w_{S,1} & w_{S,2} & \cdots & w_{S,R} \end{bmatrix} \quad (57.700)$$

Il faut bien sûr prendre en compte que dimensionnellement nous n'avons pas nécessairement $S = R$ dans le cas général (les nombres de neurones et d'entrées sont indépendants). Si nous considérons que les S neurones forment un vecteur de neurones, alors nous pouvons créer les vecteurs:

$$\vec{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_S \end{pmatrix}, \vec{n} = \begin{pmatrix} n_1 \\ n_2 \\ \vdots \\ n_S \end{pmatrix}, \vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_S \end{pmatrix} \quad (57.701)$$

Ceci nous amène à la représentation simplifiée illustrée ci-dessous:

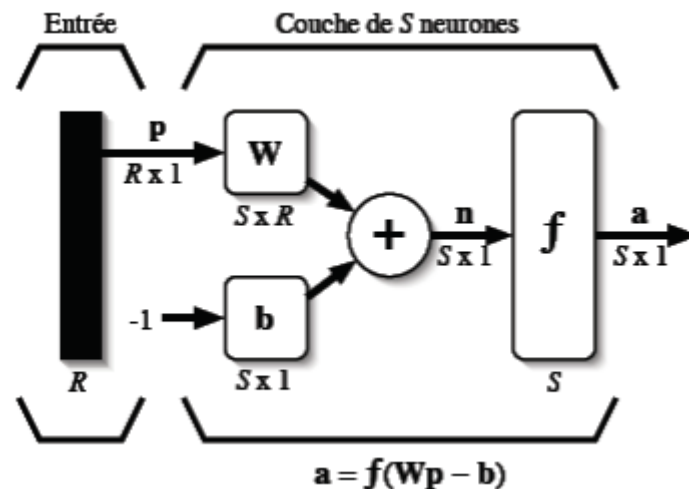


Figure: 57.87 - Principe fonctionnel d'un réseau de neurones formel

Finalement, pour construire un réseau de neurones (ou PMC pour "**Perceptron Multi-Couches**"), il ne suffit plus que de combiner des couches comme ci-dessous:

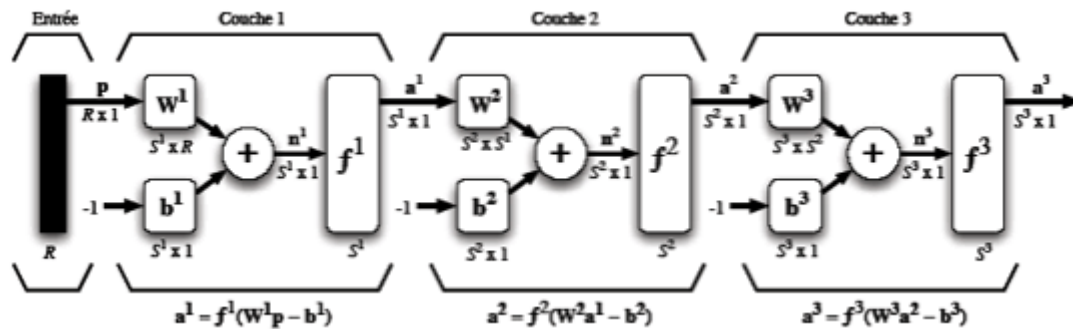


Figure: 57.88 - Principe de construction d'un PMC

Cet exemple comporte R entrées et trois couches de neurones comptant respectivement S^1, S^2, S^3 neurones. Dans le cas général, de nouveau ces nombres ne sont pas nécessairement égaux. Chaque couche possède aussi sa propre matrice de poids W^k , où k désigne l'indice de couche. Dans le contexte des vecteurs et des matrices relatives à une couche, nous emploierons toujours un exposant pour désigner cet indice. Ainsi, les vecteurs $\vec{p}, \vec{x}^k, \vec{a}^k$ sont aussi associés à la couche k .

Il importe de remarquer dans cet exemple que les couches qui suivent la première ont comme entrée la sortie de la couche précédente. Ainsi, nous pouvons enfilez autant de couches que nous voulons, du moins en théorie. Nous pouvons fixer un nombre quelconque de neurones sur chaque couche. En pratique, nous verrons plus tard qu'il n'est cependant pas souhaitable d'utiliser trop de neurones. Notons aussi que rien ne nous empêche de changer de fonction de transfert d'une couche à l'autre. Ainsi, dans le cas général nous n'avons pas nécessairement $f^1 = f^2 = f^3$.

Définition: La dernière couche est nommée "**couche de sortie**". Les couches qui précèdent la couche de sortie sont nommées "**couches cachées**".

Remarque: Les réseaux multicouches sont beaucoup plus puissants que les réseaux simples à une seule couche bien évidemment. En utilisant deux couches, à condition d'employer une fonction d'activation sigmoïde sur la couche cachée, nous pouvons "entraîner" un réseau à produire une approximation de la plupart des fonctions, avec une précision arbitraire. Sauf dans de rares cas, les réseaux de neurones formels exploitent deux ou trois couches.

Définition: "**Entraîner**" un réseau de neurones signifie modifier la valeur de ses poids et de ses biais pour qu'il réalise la fonction d'entrée sortie (I/O). Nous étudierons en détails différents algorithmes et méthodes d'approche heuristiques pour y parvenir dans différents contextes.

Exemple:

Une entreprise a mesuré pendant 14 semaines ses ventes réelles (Colonne: *Valeur à prédire*) en fonction des ventes prévisionnelles de cinq de ses succursales (*Variable1, Variable2*, etc.) et les a reproduites dans Microsoft Excel 14.0.6123:

	A	B	C	D	E	F	G	H
20			Valeur à prédire	Variable1	Variable2	Variable3	Variable4	Variable5
21	Semaine		y^k	x^1	x^2	x^3	x^4	x^5
22	03.01.1997		32.50	9.25	42.50	52.38	34.63	57.88
23	10.01.1997		33.50	8.88	49.00	54.50	35.75	61.13
24	17.01.1997		33.00	9.00	48.63	55.00	34.38	60.13
25	24.01.1997		33.63	8.63	45.63	54.25	35.25	62.50
26	31.01.1997		32.13	8.38	46.63	54.50	34.88	59.00
27	07.02.1997		32.25	8.25	45.50	52.50	34.13	56.75
28	14.02.1997		32.75	7.75	44.75	53.63	34.50	58.75
29	21.02.1997		33.13	7.88	43.38	53.75	35.13	58.50
30	28.02.1997		32.88	8.25	42.38	52.75	34.00	57.88
31	07.03.1997		32.25	8.13	42.63	53.38	31.88	56.63
32	14.03.1997		32.13	8.50	42.50	53.88	30.50	58.00
33	21.03.1997		31.75	8.25	40.88	54.50	30.25	57.00
34	27.03.1997		30.88	8.50	40.13	54.25	30.25	56.25
35	31.03.1997		31.38	8.25	40.25	52.38	30.00	55.38

Figure: 57.89 - Données d'entraînement pour le réseau de neurones

Signalons qu'il n'y a absolument aucune formule dans le tableau ci-dessus! Le retour d'expérience (et surtout l'emplacement...) nous dit qu'il vaudrait mieux faire une architecture de réseau à deux neurones, un premier avec les succursales {1,2,3} basé sur une sigmoïde et un deuxième avec les succursales {4,5} lui aussi basé sur une sigmoïde. De plus, l'ensemble devra avoir un seul et unique biais et les deux neurones devraient avoir un poids particulier en comparaison de l'un et l'autre.

Nous préparons alors le tableau suivant:

	B	C	D
6			Valeurs
7	Vecteur	w_{11}	0.000
8		w_{12}	0.000
9		w_{13}	0.000
10		w_{21}	0.000
11		w_{22}	0.000
12	Biais	θ_0	0.000
13	Pondérations	θ_1	0.000
14		θ_2	0.000

Figure: 57.90 - Poids, biais et pondérations à déterminer par recherche opérationnelle

Une fois le tableau des poids, biais et pondérations construit, nous écrivons notre réseau à deux neurones avec la fonction de sigmoïde par exemple juste à côté des données d'entraînement (ce qui facilitera la comparaison):

	C	D	E	F	G	H	I	J	K
19									
20	Valeur à prédire	Variable1	Variable2	Variable3	Variable4	Variable5	Réseau de neurone		
21	y^k	x^1	x^2	x^3	x^4	x^5	RNF1	RNF2	Sortie
22	32.50	9.25	42.50	52.38	34.63	57.88	0.500	0.500	0.000
23	33.50	8.88	49.00	54.50	35.75	61.13	0.500	0.500	0.000
24	33.00	9.00	48.63	55.00	34.38	60.13	0.500	0.500	0.000
25	33.63	8.63	45.63	54.25	35.25	62.50	0.500	0.500	0.000
26	32.13	8.38	46.63	54.50	34.88	59.00	0.500	0.500	0.000
27	32.25	8.25	45.50	52.50	34.13	56.75	0.500	0.500	0.000
28	32.75	7.75	44.75	53.63	34.50	58.75	0.500	0.500	0.000
29	33.13	7.88	43.38	53.75	35.13	58.50	0.500	0.500	0.000
30	32.88	8.25	42.38	52.75	34.00	57.88	0.500	0.500	0.000
31	32.25	8.13	42.63	53.38	31.88	56.63	0.500	0.500	0.000
32	32.13	8.50	42.50	53.88	30.50	58.00	0.500	0.500	0.000
33	31.75	8.25	40.88	54.50	30.25	57.00	0.500	0.500	0.000
34	30.88	8.50	40.13	54.25	30.25	56.25	0.500	0.500	0.000
35	31.38	8.25	40.25	52.38	30.00	55.38	0.500	0.500	0.000

Figure: 57.91 - Poids, biais et pondérations à déterminer par recherche opérationnelle

Soit avec les formules explicites pour les trois dernières colonnes qui nous intéressent:

	I	J	K
19			
20	Réseau de neurones		
21	RNF1	RNF2	Sortie
22	=1/(1+EXP(-(D22*D7+E22*D8+F22*D9)))	=1/(1+EXP(-(G22*D10+H22*D11)))	=D12*D13*I22+D14*J22
23	=1/(1+EXP(-(D23*D7+E23*D8+F23*D9)))	=1/(1+EXP(-(G23*D10+H23*D11)))	=D12*D13*I23+D14*J23
24	=1/(1+EXP(-(D24*D7+E24*D8+F24*D9)))	=1/(1+EXP(-(G24*D10+H24*D11)))	=D12*D13*I24+D14*J24
25	=1/(1+EXP(-(D25*D7+E25*D8+F25*D9)))	=1/(1+EXP(-(G25*D10+H25*D11)))	=D12*D13*I25+D14*J25
26	=1/(1+EXP(-(D26*D7+E26*D8+F26*D9)))	=1/(1+EXP(-(G26*D10+H26*D11)))	=D12*D13*I26+D14*J26
27	=1/(1+EXP(-(D27*D7+E27*D8+F27*D9)))	=1/(1+EXP(-(G27*D10+H27*D11)))	=D12*D13*I27+D14*J27
28	=1/(1+EXP(-(D28*D7+E28*D8+F28*D9)))	=1/(1+EXP(-(G28*D10+H28*D11)))	=D12*D13*I28+D14*J28
29	=1/(1+EXP(-(D29*D7+E29*D8+F29*D9)))	=1/(1+EXP(-(G29*D10+H29*D11)))	=D12*D13*I29+D14*J29
30	=1/(1+EXP(-(D30*D7+E30*D8+F30*D9)))	=1/(1+EXP(-(G30*D10+H30*D11)))	=D12*D13*I30+D14*J30
31	=1/(1+EXP(-(D31*D7+E31*D8+F31*D9)))	=1/(1+EXP(-(G31*D10+H31*D11)))	=D12*D13*I31+D14*J31
32	=1/(1+EXP(-(D32*D7+E32*D8+F32*D9)))	=1/(1+EXP(-(G32*D10+H32*D11)))	=D12*D13*I32+D14*J32
33	=1/(1+EXP(-(D33*D7+E33*D8+F33*D9)))	=1/(1+EXP(-(G33*D10+H33*D11)))	=D12*D13*I33+D14*J33
34	=1/(1+EXP(-(D34*D7+E34*D8+F34*D9)))	=1/(1+EXP(-(G34*D10+H34*D11)))	=D12*D13*I34+D14*J34
35	=1/(1+EXP(-(D35*D7+E35*D8+F35*D9)))	=1/(1+EXP(-(G35*D10+H35*D11)))	=D12*D13*I35+D14*J35

Figure: 57.92 - Le réseau de neurones avec les deux neurones, la fonction sigmoïde et la pondération à la sortie

Pour pouvoir appliquer une technique de recherche opérationnelle, il nous faut minimiser ou maximiser quelque chose. Dès lors, nous allons chercher à minimiser la somme des erreurs quadratiques en créant la colonne suivante:

	I	J	K	L
20	Réseau de neurones			Erreur
21	RNF1	RNF2	Sortie	
22	0.500	0.500	0.000	1'056.2500
23	0.500	0.500	0.000	1'122.2500
24	0.500	0.500	0.000	1'089.0000
25	0.500	0.500	0.000	1'130.6406
26	0.500	0.500	0.000	1'032.0156
27	0.500	0.500	0.000	1'040.0625
28	0.500	0.500	0.000	1'072.5625
29	0.500	0.500	0.000	1'097.2656
30	0.500	0.500	0.000	1'080.7656
31	0.500	0.500	0.000	1'040.0625
32	0.500	0.500	0.000	1'032.0156
33	0.500	0.500	0.000	1'008.0625
34	0.500	0.500	0.000	953.2656
35	0.500	0.500	0.000	984.3906
36				14738.609

Figure: 57.93 - Ajout de la colonne avec l'erreur quadratique à minimiser

Soit avec les formules explicites (nous voyons bien que cela correspond au carré de la différence entre les mesures et le modèle):

	L
20	Erreur
21	
22	=(C22-K22)^2
23	=(C23-K23)^2
24	=(C24-K24)^2
25	=(C25-K25)^2
26	=(C26-K26)^2
27	=(C27-K27)^2
28	=(C28-K28)^2
29	=(C29-K29)^2
30	=(C30-K30)^2
31	=(C31-K31)^2
32	=(C32-K32)^2
33	=(C33-K33)^2
34	=(C34-K34)^2
35	=(C35-K35)^2
36	=SOMME(L22:L35)

Figure: 57.94 - Implémentation de la colonne avec l'erreur quadratique à minimiser

Il va s'agir maintenant avec le solveur de Microsoft Excel 14.0.6123 de minimiser le contenu de la cellule L36:

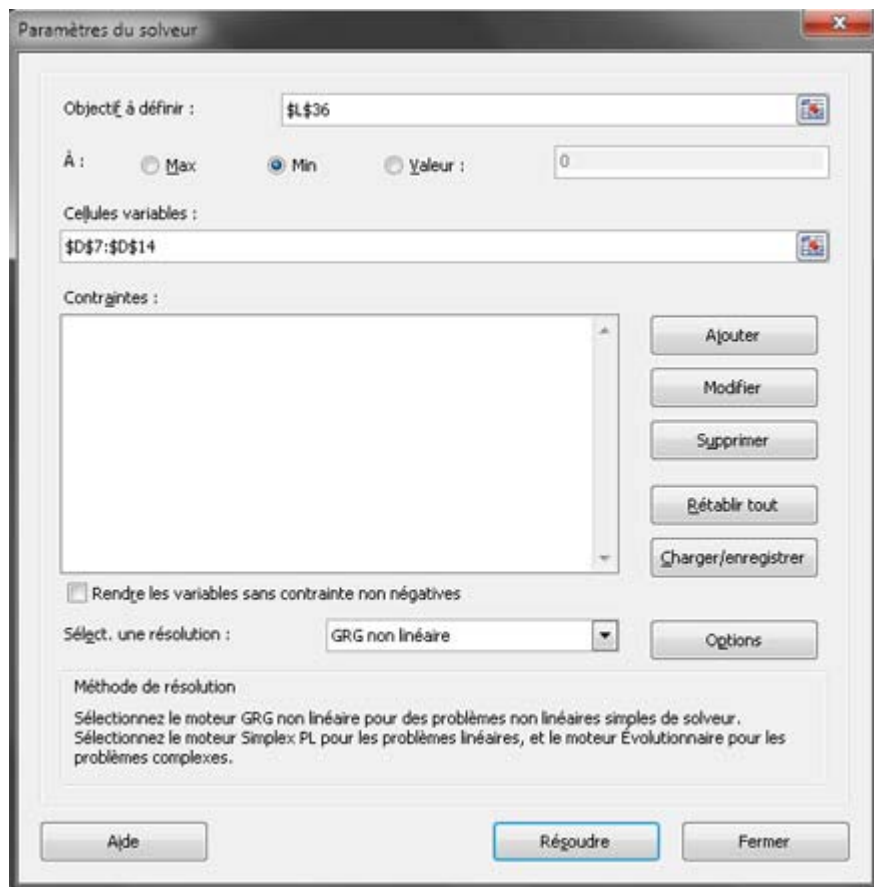


Figure: 57.95 - Ouverture du solveur et choix du GRG non linéaire

et il ne faut pas être trop regardant sur la précision des contraintes et il faut jouer un peu avec ce paramètre pour obtenir un résultat satisfaisant:

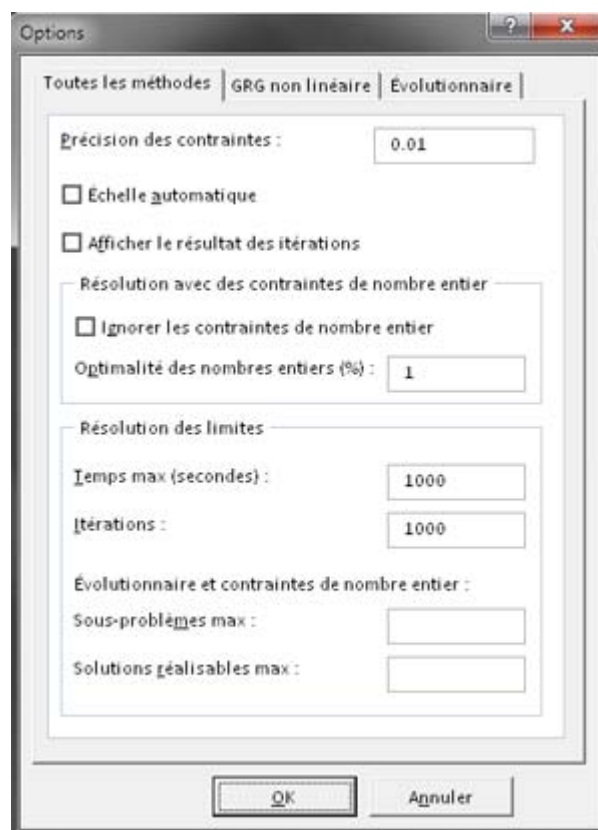


Figure: 57.96 - Options de précision des contraintes

Pour obtenir un résultat satisfaisant, il faudra mettre dans le cas présent une précision de 0.001. Ce qui donnera lors de l'exécution de la recherche par le solveur une erreur quadratique totale de 0.8479 (cellule L36) et pour les paramètres du réseau de neurones:

	A	B	C	D
5				
6				Valeurs
7		Vecteur	w_{11}	-2.629
8			w_{12}	1.293
9			w_{13}	-0.488
10			w_{21}	0.015
11			w_{22}	0.028
12		Biais	θ_0	-55.833
13		Pondérations	θ_1	20.630
14			θ_2	75.775

Figure: 57.97 - Paramètres du réseau de neurones

Des logiciels spécialisés feront mieux avec une erreur quadratique totale de 0.8405 (cellule L36) et pour les paramètres du réseau de neurones:

	A	B	C	D
5				
6				Valeurs
7		Vecteur	w_{11}	-2.695
8			w_{12}	1.309
9			w_{13}	-0.486
10			w_{21}	0.002
11			w_{22}	0.004
12		Biais	θ_0	-101.559
13		Pondérations	θ_1	25.337
14			θ_2	187.317

Figure: 57.98 - Paramètres du réseau de neurones

Nous pouvons comparer graphiquement les mesures utilisées pour entraîner le réseau de neurones et le résultat du modèle de réseau de neurones lui-même. Nous avons alors:

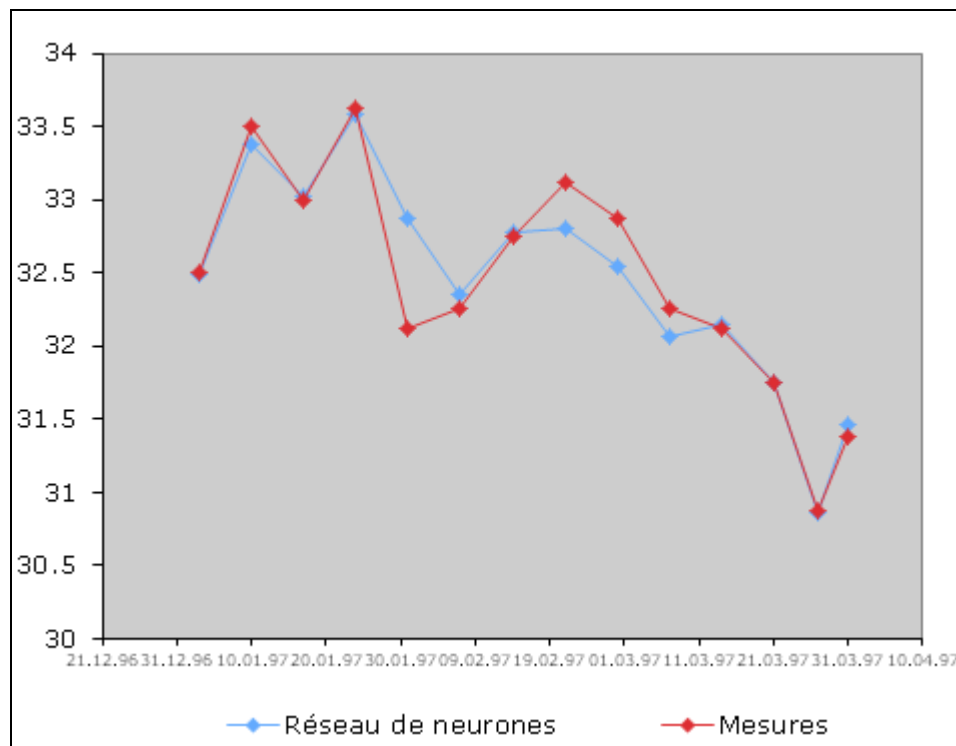


Figure: 57.99 - Comparaison réel/modèle

Ce qui est pas mal du tout pour un modèle non linéaire! Mais une fois le modèle entraîné, il faut voir s'il s'applique sur d'autres données:

	A	E	C	D	E	F	G	H	I	J	K
37		Mesures							Réseau de neurones		
38	04.04.1997	30.88	8.25	39.13	51.00	30.13	54.00	0.98	0.57	31.00	
39	11.04.1997	32.25	8.00	37.63	50.25	28.88	53.00	0.96	0.57	30.11	
40	18.04.1997	34.25	8.13	38.88	52.25	30.38	56.25	0.97	0.58	31.15	
41	25.04.1997	34.25	8.00	39.13	51.63	29.25	54.88	0.99	0.57	31.19	
42	02.05.1997	34.75	8.63	41.38	53.38	29.75	57.00	0.99	0.58	31.74	
43	09.05.1997	36.63	9.25	42.25	53.88	31.25	57.88	0.98	0.58	31.87	

Figure: 57.100 - Données de test pour voir si le modèle est bien entraîné

Toujours avec les mêmes formules:

	I	J	K
37		Réseau de neurones	
38	=1/(1+EXP(-(D38*D7+E38*D8+F38*D9)))	=1/(1+EXP(-(G38*D10+H38*D11)))	=D12*D13*I38+D14*J38
39	=1/(1+EXP(-(D39*D7+E39*D8+F39*D9)))	=1/(1+EXP(-(G39*D10+H39*D11)))	=D12*D13*I39+D14*J39
40	=1/(1+EXP(-(D40*D7+E40*D8+F40*D9)))	=1/(1+EXP(-(G40*D10+H40*D11)))	=D12*D13*I40+D14*J40
41	=1/(1+EXP(-(D41*D7+E41*D8+F41*D9)))	=1/(1+EXP(-(G41*D10+H41*D11)))	=D12*D13*I41+D14*J41
42	=1/(1+EXP(-(D42*D7+E42*D8+F42*D9)))	=1/(1+EXP(-(G42*D10+H42*D11)))	=D12*D13*I42+D14*J42
43	=1/(1+EXP(-(D43*D7+E43*D8+F43*D9)))	=1/(1+EXP(-(G43*D10+H43*D11)))	=D12*D13*I43+D14*J43

Figure: 57.101 - Données de test pour voir si le modèle est bien entraîné

Et si nous comparons aussi graphiquement les données réelles et celles modélisées, nous obtenons:

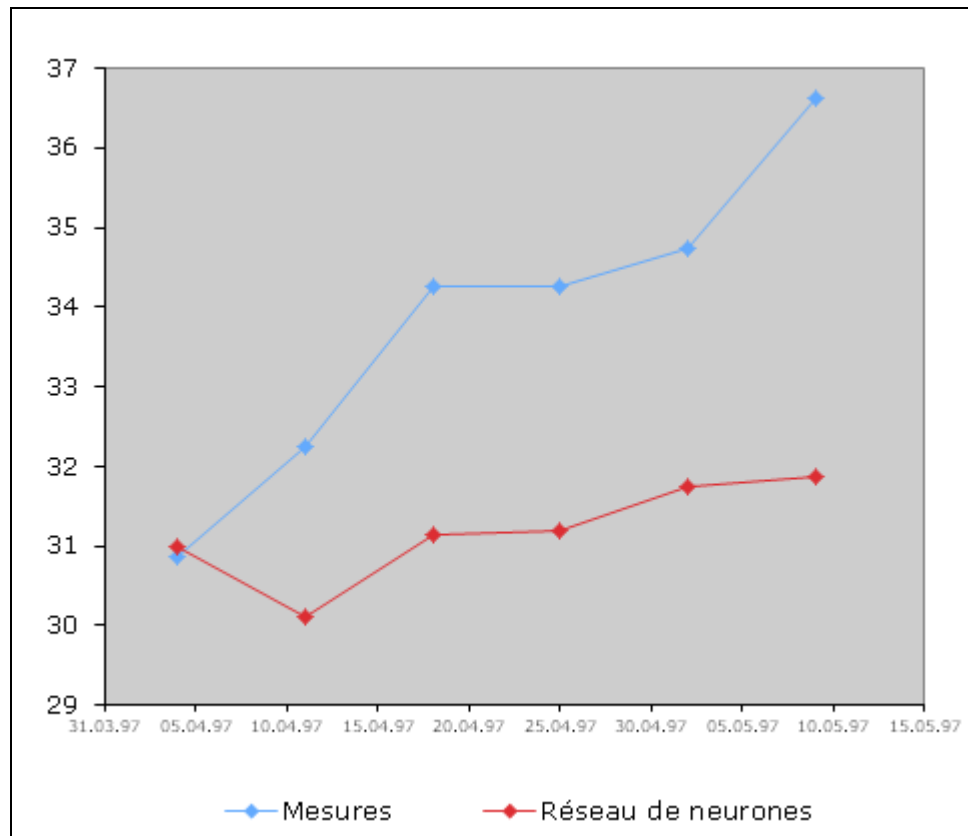


Figure: 57.102 - Comparaison réel/modèle

et là nous voyons que le modèle est nettement moins bon. Mais c'est ainsi! La science prédictive n'est pas une science exacte mais une méthode heuristique...

28. ALGORITHMES GÉNÉTIQUES

Les algorithmes génétiques (AGs) sont des algorithmes d'optimisation stochastiques itérés fondés sur les mécanismes de la sélection naturelle et de la génétique appartenant à la famille des "**algorithmes évolutionnaires**". Il s'agit d'une technique d'optimisation qui s'est beaucoup répandue depuis le début du 21ème siècle grâce la version 14.0.6123 de Microsoft Excel où le solveur intègre un algorithme évolutionnaire par défaut comme en atteste la capture d'écran ci-dessous:

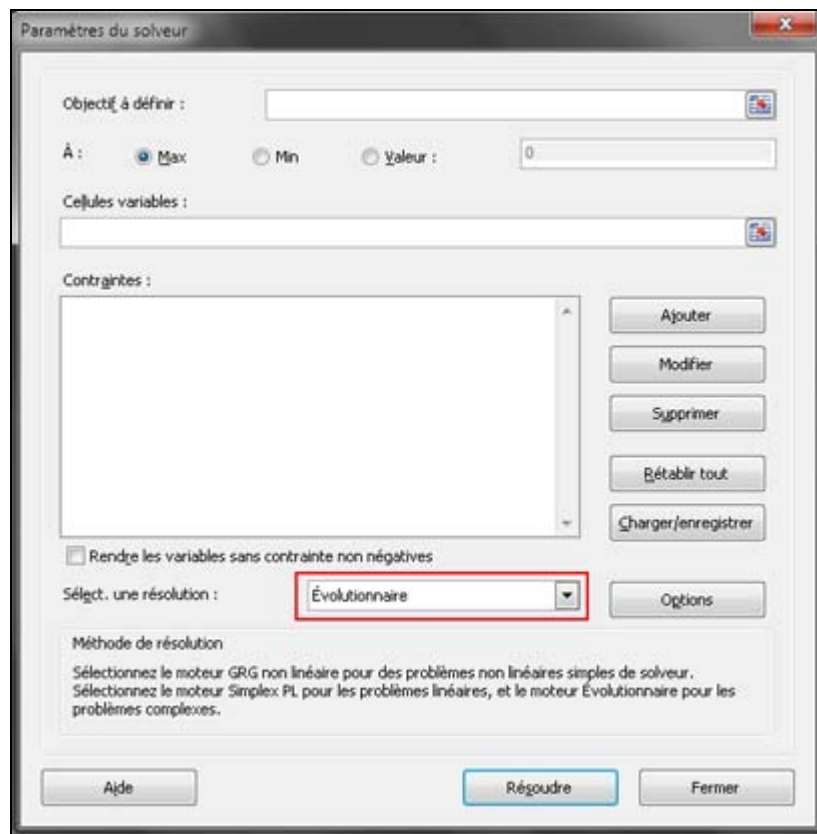


Figure: 57.103 - Capture d'écran du solveur de Microsoft Excel 14.0.6123

avec les options avancées correspondantes:



Figure: 57.104 - Options avancées de la résolution évolutionnaire du solveur de Microsoft Excel 14.0.6123

Le fonctionnement des algorithmes génétiques est relativement simple:

1. Nous partons avec une population de solutions potentielles (chromosomes) initiales arbitrairement choisies
2. Nous évaluons leur performance (fitness) relative
3. Sur la base de ces performances, nous créons une nouvelle population de solutions potentielles en utilisant des opérateurs évolutionnaires simples: la sélection, le croisement et la mutation.
4. Nous recommençons ce cycle jusqu'à ce que nous trouvions une solution satisfaisante.

Les AGs ont été initialement développés par John Holland (1975). C'est au livre de Goldberg (1989) que nous devons leur popularisation. Leurs champs d'application sont très vastes. Outre l'économie (minimisation du risque des portefeuilles), ils sont utilisés pour l'optimisation de fonctions, en finance, en théorie du contrôle optimal (recherche opérationnelle), ou encore en théorie des jeux répétés et différentiels (en l'occurrence dans les jeux évolutionnaires et le dilemme du prisonnier) et la recherche d'information (Google) ainsi que la recherche des plus courts chemins en théorie des graphes (routages Internet ou GPS). La raison de ce grand nombre d'applications est claire: simplicité et efficacité. Bien sûr, d'autres techniques d'exploration stochastiques existent, la méthode de Monte-Carlo peut être considérée comme un concept similaire.

Pour résumer, Lerman et Ngouenet (1995) distinguent quatre principales propriétés qui font la différence fondamentale entre ces algorithmes et les autres méthodes:

- P1. Les algorithmes génétiques utilisent un codage des paramètres, et non les paramètres eux-mêmes.
- P2. Les algorithmes génétiques travaillent sur une population de points, au lieu d'un point unique.
- P3. Les algorithmes génétiques n'utilisent que les valeurs de la fonction étudiée, pas sa dérivée, ou une autre connaissance auxiliaire.
- P4. Les algorithmes utilisent des règles de transition probabilistes, et non déterministes.

La simplicité de leurs mécanismes, la facilité de leur mise en application et leur efficacité même pour des problèmes complexes ont conduit à un nombre croissant de travaux ces dernières années.

Définitions:

- D1. Un "**algorithme génétique**" est défini par un individu/chromosome/séquence et une solution potentielle au problème donné.
- D2. Une "**population**" est un ensemble de chromosomes ou de points de l'espace de recherche
- D3. "**L'environnement**" est assimilé à l'espace de recherche
- D4. La fonction que nous cherchons à optimiser est appelée "**fonction de fitness**"

Avant d'aller plus loin, il faut définir de manière plus formelle les concepts précédents (mais sous l'hypothèse particulière ! de codage binaire).

Définition: Les organismes en compétition sont appelés "**individu**". Soit un alphabet $A = \{a_1, a_2, \dots, a_l\}$, nous supposons que chaque individu peut être représenté par un mot de longueur fixe l pris dans A^* . Le mot A associé à un individu de la population sera appelé "**chromosome**" ou "**séquence**" (le terme n'est pas tout à fait équivalent à son homonyme biologique, cependant c'est une pratique courante que d'utiliser ce terme ici aussi) et donné donc par A de longueur $l(A)$ avec $\forall i \in [1, l], a_i \in V = \{0, 1\}$ (cause: hypothèse du codage binaire).

Dans la mesure où il n'y aura pas de risque de confusion, nous identifierons les termes individu et chromosome.

Les individus forment une population P de taille N , notée:

$$P = \{A_i\}_{i=1..N} \quad (57.702)$$

Nous allons faire une autre assertion importante, c'est-à-dire, qu'il existe une fonction f d'une séquence à valeurs positives que nous noterons $f(A)$, dite "**fonction de fitness**" qui à tout A_i associe un réel tel que:

$$f(A_i) > f(A_j), i \neq j \quad (57.703)$$

si et seulement si A_i est mieux adapté au milieu que A_j .

Remarquons que le terme "adapté" n'est pas défini. Pour cela, il faudrait caractériser le milieu dans lequel évoluent les individus, ce que nous ne ferons pas. En fait, puisque nous supposons l'existence d'une telle fonction et que nous posons l'équivalence avec le degré d'adaptation, celui-ci est automatiquement défini par la donnée de f .

Nous nommerons "**génération**" une population à un instant t , ce qu'il faut mettre en relation avec la notion de durée de vie ou d'âge. Cependant, nous nous placerons ici dans le cas particulier où chaque individu a une durée de vie égale à 1, donc la génération $(t+1)$ est constituée d'individus différents de la génération t , nous les appellerons les "**descendants**". Réciproquement, les individus de la génération t seront les "**ancêtres**" des individus de la génération $(t+1)$. Nous désignerons la génération t par $P(t)$ soit la population à l'instant t .

Ainsi, un chromosome est vu comme une suite de bits en codage binaire appelée aussi "**chaîne binaire**". Dans le cas d'un codage non binaire, tel que le codage réel par exemple, alors la suite A ne contient qu'un point, nous avons $A = \{a\}$ avec $a \in \mathbb{R}$.

Remarque: La fitness (efficacité) est donc donnée par une fonction à valeurs positives réelles. Dans le cas d'un codage binaire, nous utiliserons souvent une fonction de décodage d qui permettra de passer d'une chaîne binaire à un chiffre à valeur réelle:

$$d : \{0, 1\}^l \rightarrow \mathbb{R} \quad (57.704)$$

La fonction de fitness est alors choisie telle qu'elle transforme cette valeur en valeur positive soit:

$$f : d\{0, 1\}^l \rightarrow \mathbb{R}_+^* \quad (57.705)$$

Le but d'un algorithme génétique est alors simplement de trouver la chaîne qui maximise cette fonction f . Bien évidemment, chaque problème particulier nécessitera ses propres fonctions d et f .

Les AGs sont alors basés sur les phases suivantes:

1. Initialisation: une population initiale de N chromosomes est tirée aléatoirement
2. Évaluation: chaque chromosome est décodé puis évalué
3. Sélection: création d'une nouvelle population de N chromosomes par l'utilisation d'une méthode de sélection appropriée.
4. Reproduction: possibilité de croisement et mutation au sein de la nouvelle population
5. Retour à la phase d'évaluation jusqu'à l'arrêt de l'algorithme

28.1. CODAGE ET POPULATION INITIALE

Il existerait trois principaux types de codage: (1) Binaire, (2) Gray, (3) Réel. Nous pouvons facilement passer d'un codage à l'autre. Certains auteurs n'hésitent pas, par ailleurs, à faire le parallèle avec la biologie en parlant de génotype (cf. chapitre de Dynamique Des Populations) en ce qui concerne la représentation binaire d'un individu, et de phénotype (cf. chapitre de Dynamique Des Populations) pour ce qui est de sa valeur réelle correspondante dans l'espace de recherche.

Rappelons que la transformation la plus simple (fonction de décodage d) d'une chaîne binaire A en nombre entier x s'opère par la règle suivante (cf. chapitre sur les Nombres):

$$x = d(A) = \sum_{i=0}^l a_i 2^i \quad (57.706)$$

où l est le nombre de chiffres de la chaîne -1.

Ainsi, le chromosome $A = \{1, 0, 1, 1\}$ vaut trivialement:

$$1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 = 8 + 2 + 1 = 11 \quad (57.707)$$

Évidemment, la fonction d sera adaptée (par tâtonnements!) selon le problème. Ainsi, si nous cherchons à maximiser une fonction $f: [0,1] \rightarrow [0,1]$ une méthode possible serait la suivante (la taille du chromosome dépendant bien évidemment de la précision voulue):

$$x = d(A) = \sum_{i=1}^{l+1} a_i 2^{-i} \quad (57.708)$$

ce qui peut s'assimiler à une série harmonique. Pour une précision au cinquième chiffre après la virgule, nous imposerons $l = 17 - 1$ puisque:

$$d\left(\left\{\underbrace{1, \dots, 1}_{17}\right\}\right) \cong 0.999992 \rightarrow 1 \quad (57.709)$$

Une autre façon de faire serait de choisir d telle que:

$$x = d(A) = \sum_{i=0}^l \frac{a_i 2^i}{2^{l+1} - 1} \quad (57.710)$$

Petite explication:

$$\begin{aligned} x^n - 1 &= (x - 1) \cdot (1 + x + x^2 + \dots + x^{n-1}) \\ \frac{x^n - 1}{(x - 1)} &= 1 + x + x^2 + x^3 + \dots + x^{n-1} = \sum_{i=0}^{n-1} x^i \quad (57.711) \\ 2^n - 1 &= \sum_{i=0}^{n-1} 2^i \Leftrightarrow 1 = \sum_{i=0}^{n-1} \frac{2^i}{2^n - 1} \end{aligned}$$

Posons $l = n - 1$:

$$1 = \sum_{i=0}^l \frac{2^i}{2^{l+1} - 1} \quad (57.712)$$

Ainsi, avec $l = 16$ nous avons $2^{17} - 1 = 131071$ et:

$$d\left(\left\{\underbrace{1, \dots, 1}_{17}\right\}\right) = \frac{131071}{131071} = 1 \quad (57.713)$$

Cette dernière règle peut se généraliser. Ainsi, admettons que nous cherchons à maximiser (normaliser serait un terme peut-être plus correct) f en fonction d'une variable réelle x . Soit $D = [x_{\min}, x_{\max}]$, avec $D \subset \mathbb{R}$, l'espace de recherche permis avec $x_{\min}$ et $x_{\max}$ les bornes inférieures et supérieures. Soit $prec$ la précision (chiffre après la virgule) avec laquelle nous cherchons x . Soit:

$$ld = x_{\max} - x_{\min} \quad (57.714)$$

la longueur de l'intervalle D . Nous devons alors diviser cet intervalle au pire en:

$$n = ld \cdot 10^{prec} \quad (57.715)$$

sous-intervalles égaux afin de respecter la précision. Par exemple, soit $D = [-1, 2]$ nous avons donc $ld = 3$ si nous voulions une précision $prec = 6$, alors il nous faut diviser cet intervalle en $n = 3'000'000$ sous-intervalles.

Notons s l'entier naturel tel que $2^s > n$, ce qui dans notre exemple implique $s = 22$ puisque:

$$2^{21} = 2'097'152 < 3'000'000 < 2^{22} = 4'194'304 \quad (57.716)$$

la transformation d'une chaîne binaire $A = \{a_1, \dots, a_l\}$ en un nombre réel x peut alors s'exécuter en trois étapes:

1. Conversion (base 2 en base 10):

$$x' = \sum_{i=0}^l \alpha_i 2^i \quad (57.717)$$

2. Normalisation:

$$x'' = x' \frac{x_{\max} - x_{\min}}{2^{l+1} - 1} = \frac{x' l d}{2^{l+1} - 1} \quad (57.718)$$

3. Maximisation:

$$x = x_{\min} + x'' \quad (57.719)$$

ou ce qui revient au même directement en une seule étape par:

$$x = x_{\min} + \sum_{i=0}^l \frac{\alpha_i 2^i l d}{2^{l+1} - 1} \quad (57.720)$$

Ainsi pour $f : [0,1] \rightarrow [0,1]$, et $\forall i, \alpha_i = 1$ nous retrouvons bien:

$$x = 0 + \sum_{i=0}^l \frac{2^i \cdot 1}{2^{l+1} - 1} = \sum_{i=0}^l \frac{2^i}{2^{l+1} - 1} \quad (57.721)$$

Pour ce qui est de la phase d'initialisation, la procédure est assez simple. Elle consiste en un tirage aléatoire de N individus dans l'espace des individus permis. En codage binaire, selon la taille l de la chaîne, nous effectuons pour un chromosome l tirages dans $\{0,1\}$ avec équiprobabilité.

28.2. LES OPÉRATEURS

Les opérateurs jouent un rôle prépondérant dans la possible réussite d'un AG. Nous en dénombrons trois principaux: l'opérateur de sélection, de croisement et de mutation. Si le principe de chacun de ces opérateurs est facilement compréhensible, il est toutefois difficile d'expliquer l'importance isolée de chacun de ces opérateurs dans la réussite de l'AG. Cela tient pour partie au fait que chacun de ces opérateurs agit selon divers critères qui lui sont propres (valeur sélective des individus, probabilité d'activation de l'opérateur, etc.).

28.2.1. OPÉRATEUR DE SÉLECTION

Cet opérateur est peut-être le plus important puisqu'il permet aux individus d'une population de survivre, de se reproduire ou de mourir. En règle générale, la probabilité de survie d'un individu sera directement reliée à son efficacité relative au sein de la population.

Il existe plusieurs méthodes pour la reproduction. La méthode la plus connue et utilisée est sans nul doute, la roue de loterie biaisée (roulette wheel) de Goldberg (1989). Selon cette méthode, chaque chromosome sera dupliqué dans une nouvelle population proportionnellement à sa valeur d'adaptation. Nous effectuons, en quelque sorte, autant de tirages avec remise qu'il y a d'éléments dans la population. Ainsi, dans le cas d'un codage binaire, la fitness d'un chromosome particulier étant $f(d(A))$, la probabilité avec laquelle il sera réintroduit dans la nouvelle population de taille N est:

$$\frac{f(d(A))}{\sum_{j=1}^N f(d(A))} \quad (57.722)$$

Les individus ayant une grande fitness ont donc plus de chance d'être sélectionnés. Nous parlons alors de "**sélection proportionnelle**".

L'inconvénient majeur de cette méthode repose sur le fait qu'un individu n'étant pas le meilleur pourrait tout de même dominer la sélection (imaginez la recherche du maxima d'une fonction dans $\mathbb{R}^2$, il peut y en avoir plusieurs - de maxima - et donc mauvaise sélection...), nous parlerons alors à juste titre de "**convergence prématurée**" et c'est l'un des problèmes les plus fréquents lors de l'utilisation des algorithmes génétiques. Elle peut aussi donc engendrer une perte de diversité par la domination d'un super individu. Un autre inconvénient est sa faible performance vers la fin quand l'ensemble des individus se ressemblent.

Une solution à ce problème ne tient pas dans l'utilisation d'une autre méthode de sélection mais dans l'utilisation d'une fonction de fitness modifiée. Ainsi, nous pouvons utiliser un changement d'échelle (scaling) afin de diminuer ou accroître de manière artificielle l'écart relatif entre les fitness des individus.

Brièvement, il existe d'autres méthodes, la plus connue étant celle du tournoi (tournament selection): nous tirons deux individus aléatoirement dans la population et nous reproduisons le meilleur des deux dans la nouvelle population. Nous refaisons cette procédure jusqu'à ce que la nouvelle population soit complète. Cette méthode donne de bons résultats. Toutefois, aussi importante que soit la phase de sélection, elle ne crée pas de nouveaux individus dans la population. Ceci est le rôle des opérateurs de croisement et de mutation.

28.2.2. OPÉRATEUR DE CROISEMENT

L'opérateur de croisement permet la création de nouveaux individus selon un processus fort simple. Il permet donc l'échange d'informations entre les chromosomes (individus). Tout d'abord, deux individus, qui forment alors un couple, sont tirés au sein de la nouvelle population issue de la reproduction. Puis un (potentiellement plusieurs) site de croisement est tiré aléatoirement (chiffre entre 1 et $l-1$). Enfin, selon une probabilité P_c que le croisement s'effectue, les segments finaux (dans le cas d'un seul site de croisement) des deux parents sont alors échangés autour de ce site:

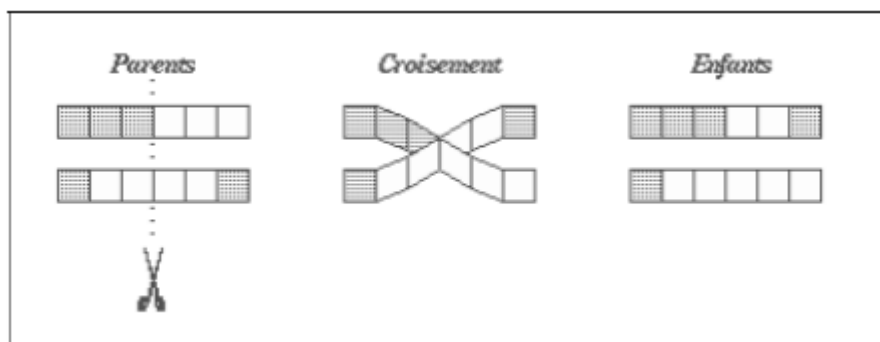


Figure: 57.105 - Exemple d'opérateur de croisement

Cet opérateur permet la création de deux nouveaux individus. Toutefois, un individu sélectionné lors de la reproduction ne subit pas nécessairement l'action d'un croisement. Ce dernier ne s'effectue qu'avec une certaine probabilité. Plus cette probabilité est élevée et plus la population subira de changement.

Quoi qu'il en soit, il se peut que l'action conjointe de la reproduction et du croisement soit insuffisante pour assurer la réussite de l'AG. Ainsi, dans le cas du codage binaire que nous avons choisi jusqu'ici, certaines informations (in extenso: caractères de l'alphabet) peuvent disparaître de la population. Ainsi aucun individu de la population initiale ne contient de 1 en dernière position de la chaîne, et que ce 1 fasse partie de la chaîne optimale à trouver, tous les croisements possibles ne permettront pas de faire apparaître ce 1 initialement inconnu. En codage réel, une telle situation peut arriver si utilisant un opérateur simple de croisement, il se trouvait qu'initialement toute la population soit comprise entre 0 et 40 et que la valeur optimale était de 50. Toutes les combinaisons convexes possibles de chiffres appartenant à l'intervalle $[0,40]$ ne permettront jamais d'aboutir à un chiffre de 50. C'est pour remédier entre autres à ce problème que l'opération de mutation est utilisée.

28.2.3. OPÉRATEUR DE MUTATION

Le rôle de cet opérateur est de modifier aléatoirement, avec une certaine probabilité, la valeur d'un composant de l'individu. Dans le cas du codage binaire, chaque bit $a_i \in \{0,1\}$ est remplacé selon une probabilité P_m par son inverse $a'_i = 1 - a_i$. C'est ce qui est illustré à la figure ci-dessous. Tout comme plusieurs lieux de croisement peuvent être possibles, nous pouvons très bien admettre qu'une même chaîne puisse subir plusieurs mutations.

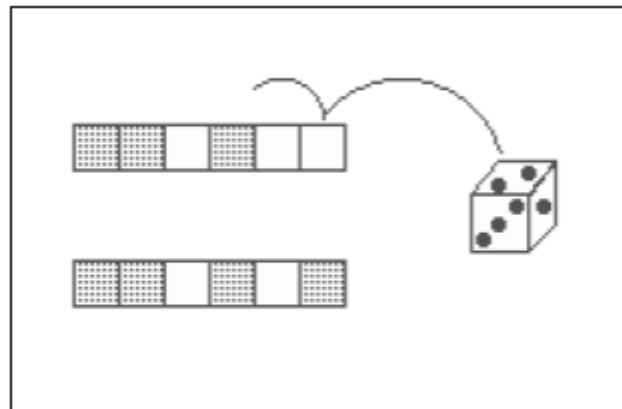


Figure: 57.106 - Principe de l'opérateur de mutation

La mutation est traditionnellement considérée comme un opérateur marginal bien qu'elle confère en quelque sorte aux algorithmes génétiques la propriété d'ergodicité (in extenso: tous les points de l'espace de recherche peuvent être atteints). Cet opérateur est d'une grande importance. Il a de fait un double rôle: celui d'effectuer une recherche locale et/ou de sortir d'une trappe (recherche éloignée).

Les opérateurs de l'algorithme génétique sont guidés par un certain nombre de paramètres fixés à l'avance. La valeur de ces paramètres influence la réussite ou non d'un algorithme génétique. Ces paramètres sont les suivants:

- La taille de la population N , et la longueur du codage de chaque individu l (dans le cas de codage binaire). Si N est trop grand, le temps de calcul de l'algorithme peut s'avérer très important, et si N est trop petit, il peut converger trop rapidement vers un mauvais chromosome.
- La probabilité de croisement P_c . Elle dépend de la forme de la fonction de fitness. Son choix est en général heuristique (tout comme pour P_m). Plus elle est élevée, plus la population subit évidemment de changements importants. Les valeurs généralement admises sont comprises entre 0.5 et 0.9.

- La probabilité de mutation P_m . Ce taux est généralement faible puisqu'un taux élevé risque de conduire à une solution sous-optimale.

Plutôt que de réduire P_m , une autre façon d'éviter que les meilleurs individus soient altérés est d'utiliser la reconduite explicite de l'élite dans une certaine proportion. Ainsi, bien souvent, les meilleurs 5%, par exemple, de la population sont directement reproduits à l'identique, l'opérateur de reproduction ne jouant alors que sur les 95% restant. Cela est appelé une "**stratégie élitiste**".

Partant du constat que les valeurs des paramètres des différents opérateurs sont elles-mêmes inconnues et ne peuvent être améliorées au fur et à mesure que de manière expérimentale, certains auteurs, tels que Novkovic et Sverko (1997), proposent d'utiliser une sorte de méta-AG: l'un pour trouver l'individu optimal et l'autre pour trouver la valeur optimale des paramètres. Ces deux algorithmes tourneraient alors simultanément ou séquentiellement. Toutefois, il est inévitable que le temps de calcul (la complexité algorithmique) s'alourdisse en conséquence.

Voyons maintenant un exemple d'AG (exemple de Goldberg - 1989). Il consiste à trouver le maximum de la fonction $f(x) = x$ sur l'intervalle $[0,31]$ où x est un entier. La première étape consiste à coder la fonction. Par exemple, nous utilisons un codage binaire de x , la séquence (chromosome) contenant au maximum 5 bits. Ainsi, nous avons $x = 2 \rightarrow \{0, 0, 0, 1, 0\}$, de même $x = 31 \rightarrow \{1, 1, 1, 1, 1\}$. Nous recherchons donc le maximum d'une fonction de fitness (nous choisirons $f(x)$ lui-même dans cet exemple simple) dans un espace de 32 valeurs possible de x .

1. Tirage et évaluation de la population initiale

Nous fixons la taille de la population à $N = 4$. Nous tirons donc de façon aléatoire 4 chromosomes sachant qu'un chromosome est composé de 5 bits, et chaque bit dispose d'une probabilité $\frac{1}{2}$ d'avoir une valeur 0 ou 1. Le maximum, (au hasard) 16 est atteint par la deuxième séquence. Voyons comment l'algorithme va tenter d'améliorer ce résultat.

D'abord, nous obtenons le tableau suivant:

N°	Chromosome	Valeur	Fitness $f(x)$	P_i %
1	00101	5	5	14.3
2	10000	16	16	45.7
3	00010	2	2	5.7
4	01100	12	12	34.3
Total			35	100

Tableau: 57.46 - Évolution (mutation) des chromosomes

Nous tournons donc quatre fois cette roue pour obtenir la séquence suivante:

Tirage	Chromosome
1	10000
2	01100
3	00101
4	10000

Tableau: 57.47 - Séquence tirage des chromosomes

Nous voyons bien ici le risque que nous aurions à perdre la séquence N° 2 dès le départ... c'est le problème de cette méthode. Elle peut converger moins vite que d'autres. Cependant, le lecteur remarquera que nous avons perdu la séquence N° 3.

Nous passons maintenant à la partie du croisement: les parents sont sélectionnés au hasard. Nous tirons aléatoirement un lieu de croisement ("site" ou "locus") dans la séquence. Le croisement s'opère alors à ce lieu avec une probabilité P_c . Le tableau ci-dessous donne les conséquences de cet opérateur en supposant que les chromosomes 1 et 3, puis 2 et 4 sont appariés et qu'à chaque fois le croisement s'opère (par exemple avec $P_c = 1$).

	$l=2$	$l=3$
Séquences d'origine	100 00	01 100
	001 01	10 000
Séquences croisées	10001	01000
	00100	10100

Tableau: 57.48 - Croisement des chromosomes

Nous passons maintenant à la partie mutation: dans cet exemple à codage binaire, la mutation est la modification aléatoire occasionnelle (de faible probabilité) de la valeur d'un bit (inversion d'un bit). Nous tirons ainsi pour chaque bit un chiffre aléatoire entre 0 et 1 et si ce chiffre est inférieur à P_m alors la mutation s'opère. Le tableau ci-dessous avec $P_m = 0.05$ met en évidence ce processus:

Anc. Chr.	Tirage aléa.	Nouveau bit	Nouveau Chr.
10001	15 25 36 04 12	1	10011
00100	26 89 13 48 59	-	00100
01000	32 45 87 22 65	-	01000
10100	47 01 85 62 35	1	11100

Tableau: 57.49 - Mutation des chromosomes

Maintenant que la nouvelle population est entièrement créée, nous pouvons à nouveau l'évaluer:

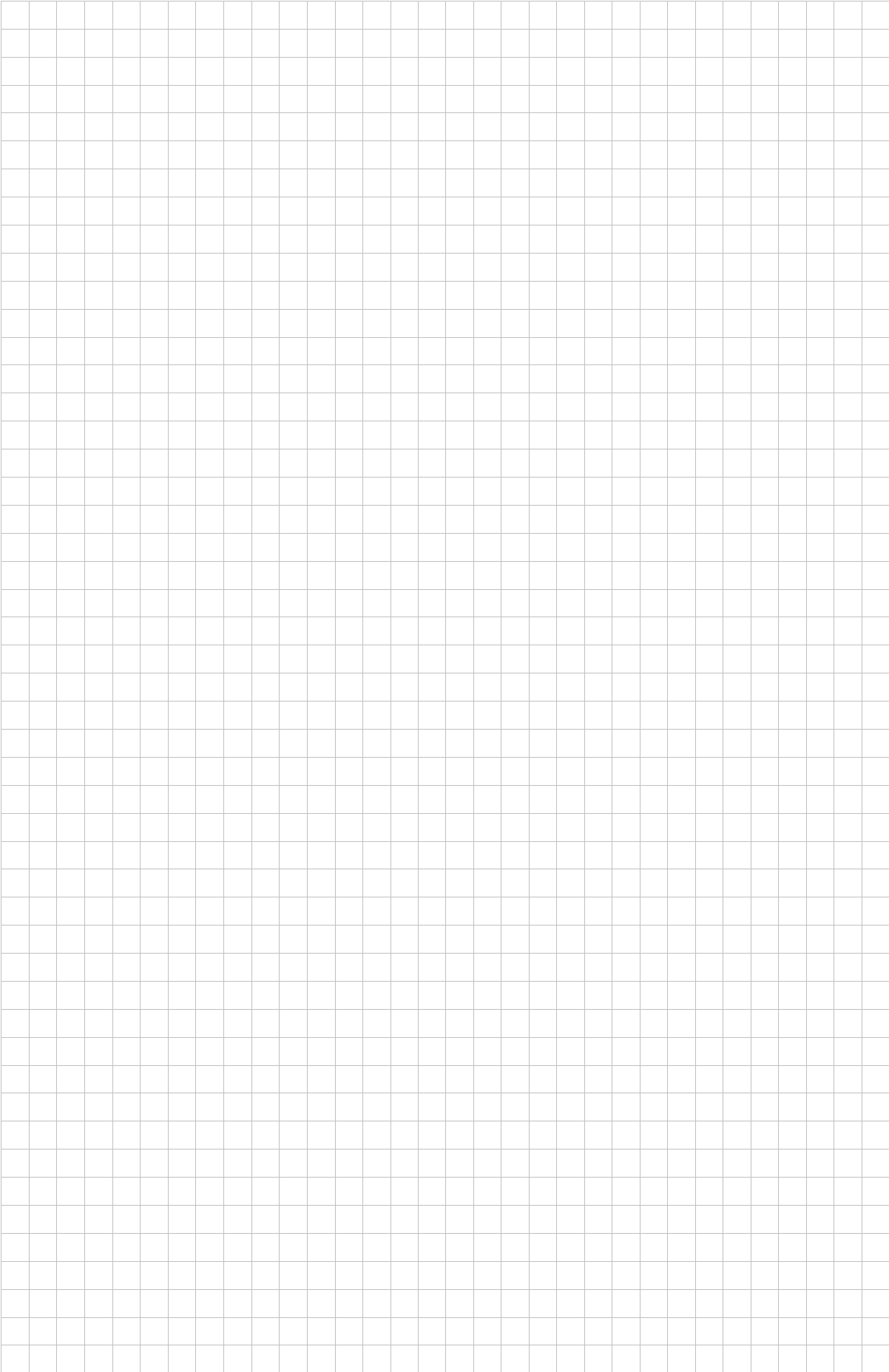
N°	Chromosome	Valeur	Fitness $f(x)$	P_i %
1	10011	19	19	32.2
2	00100	4	4	6.8
3	01000	8	8	13.5
4	11100	28	28	47.5
Total			59	100

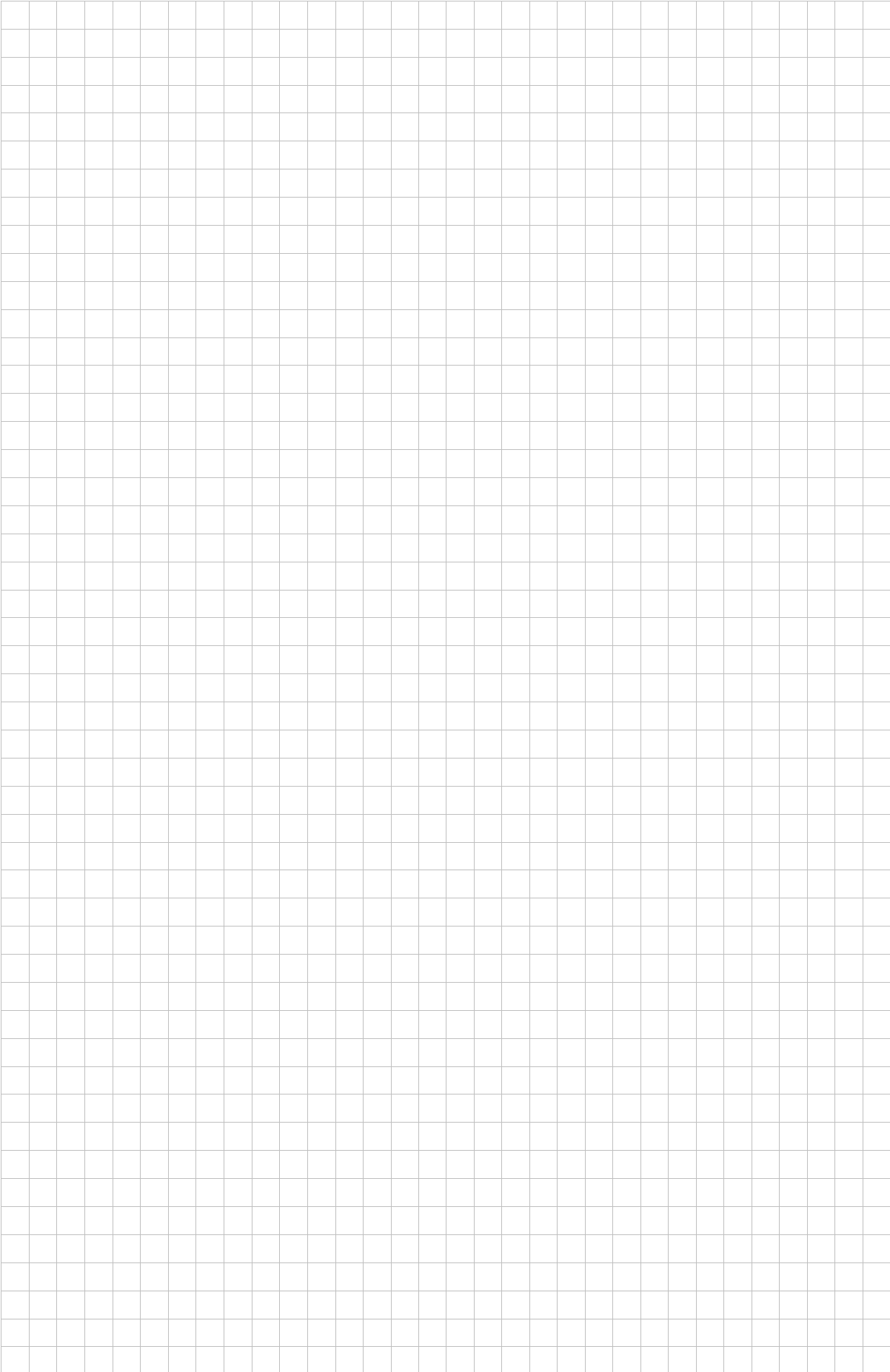
Tableau: 57.50 - Évaluation de la mutation des chromosomes

Le maximum est maintenant 28 (N° 4). Nous sommes donc passés de 16 à 28 après une seule génération. Bien sûr, nous devons recommencer la procédure à partir de l'étape de sélection jusqu'à ce que le maximum global, 31, soit obtenu, ou bien qu'un critère d'arrêt ait été satisfait.

Remarque: Il est possible de démontrer mathématiquement, ce qui est remarquable !!!, que les portions de chromosomes qui se retrouvent chez les meilleurs individus vont avoir tendance à se reproduire...







58. FRACTALES

Les fractales sont des figures invariantes par changement d'échelle (nous parlons aussi de "**structures autosimilaires**") et sont la représentation graphique de suites récurrentes contractantes (pour les fractales IFS que nous verrons plus loin) ou non divergentes (pour les fractales à temps d'échappement que nous verrons aussi plus loin).

L'idée de base - simple et géniale... à la fois - consiste souvent à prendre un point de départ, de construire son image via une fonction mathématique donnée, de prendre l'image de l'image et ainsi de suite. Le but étant d'étudier comment se répartissent les points successifs dans l'ensemble global, s'ils s'approchent d'une limite ou s'ils errent entre diverses valeurs que nous pouvons expliciter, s'il y a plus de points dans telle partie de l'ensemble que dans telle autre?

L'intérêt de ce type de questions concerne aussi bien l'étude de l'évolution de populations biologiques que celle de l'avenir du système solaire, la 3D en informatique (l'origine étant la génération de montagnes pour des paysages 3D), les variations des cours de la bourse ou la génération de nombres aléatoires dans des domaines particuliers ou encore le domaine du diagnostic médical.

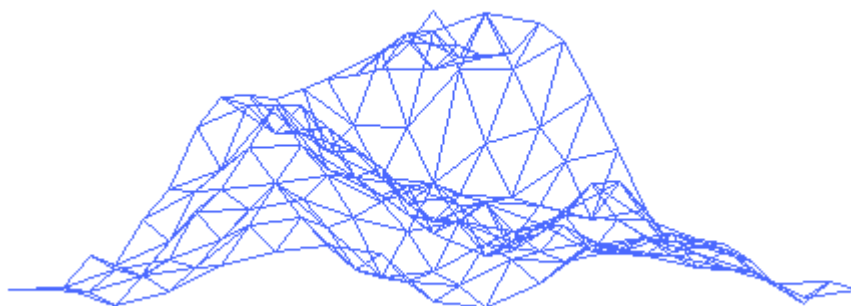


Figure: 58.1 - Génération de pseudo-reliefs à partir d'une fractale aléatoire (probabiliste)

Pour le commun des mortels, les fractales servent à faire joli. Mais elles ont des applications infiniment plus sérieuses: nous avons vu par exemple sur le présent site web que certaines de ces "séduisantes" images reproduisaient des phénomènes physiques (dynamique des populations pour la fractale de Feigenbaum, turbulences dans un fluide pour l'attracteur de Lorentz, dispositions des galaxies, L-Fractales, amas et superamas de galaxies,...). Les fractales ont également trouvé des applications en musique (avec des logiciels générant de la musique fractale) et dans le cinéma (3D). Enfin, dans le domaine de l'infographie, les fractales permettent de compresser très efficacement les images, avec une qualité constante quel que soit le zoom, elles permettent de créer des textures réalistes, et peuvent permettre de tramer une image avec de bons résultats. Les fractales sont aussi utilisées pour réduire la taille des antennes de réception et élargir leur spectre de fréquence utile (nos téléphones portables du début du 21ème siècle ont des récepteurs fractals de type "tapis de Sierpinski" - voir plus loin - à cause de tous les types de fréquences qu'ils doivent pouvoir gérer!). Dans le génie civil les fractales sont utilisées pour la construction de certains murs absorbants de son. Et bien d'autres choses encore...

Cette géométrie fractale se différencie de la géométrie euclidienne par sa définition d'une part: les figures de la géométrie euclidienne sont en général déterminées par des relations algébriques, alors que les courbes fractales sont définies de façon récursive comme nous l'avons déjà mentionné. Les fractales ont aussi des dimensions fractionnaires (nous avons déjà traité ce sujet dans le chapitre de Géométrie Euclidienne lors de la définition du concept de dimension). D'autre part, il ne faut pas non plus négliger leur aspect autosemblable: chaque partie d'une fractale peut être observée à n'importe quelle échelle: chaque partie est (sensiblement) une copie de l'ensemble.

Remarque: Les développements qui vont suivre auraient très bien pu être mis dans le chapitre de Suites Et Séries ou encore d'Analyse Fonctionnelle ou encore vus comme un cas particulier du chapitre de Topologie réduit à l'espace euclidien (raison pour laquelle vous y trouverez par ailleurs de nombreuses références). Notre choix se veut pédagogique au même titre que pour le chapitre de Cryptographie, dans le sens qu'il est beaucoup plus intéressant pour un étudiant d'une petite classe de voir une application des concepts abstraits de la topologie dans un cadre pratique (et par ailleurs esthétique) où ils sont absolument nécessaires à la bonne compréhension du sujet plutôt que dans un cadre où l'on peut très bien s'y soustraire sans avoir à trop en souffrir. Le lecteur retrouvera ici certains développements et théorèmes proposés ailleurs sur le site ceci dans le but de lui éviter d'avoir trop à "tourner les pages".

Les objets fractals naturels sont dits "**objets non déterministes**", car le processus dynamique qui permet leur création varie lui-même avec le temps de façon aléatoire (voir le chapitre de Dynamique Des Populations pour un excellent exemple). Nous pouvons néanmoins essayer de modéliser des systèmes dynamiques permettant d'aboutir à des objets fractals, sous une forme mathématique rigoureuse (c'est encore un bon exemple de la façon avec laquelle les mathématiciens arrivent à rendre un concept concret simple et intuitif en un modèle mathématique abstrait et un peu confus).

Dans le cadre de ce chapitre, nous envisagerons l'étude de deux familles de fractales qui seront dans l'ordre:

- Les fractales déterministes basées sur des fonctions itérées qui sont strictement autosimilaires. Elles sont générées, comme nous le verrons, par l'application récursive de fonctions contractantes sur des sous-ensembles d'un espace métrique. Le théorème du point fixe assurera (comme nous le verrons aussi!) l'existence et l'unicité d'un "**sous-ensemble fixe**" de l'espace métrique, vers lequel tout sous-ensemble converge.
- Les fractales à temps d'échappement (dites aussi fractales par récurrence) qui sont non strictement autosimilaires: Elles sont générées comme nous le verrons par des suites récurrentes non divergentes. Le théorème du point fixe servant de garantie pour la non-divergence de la fonction relativement aux points de départ choisis.

1. FRACTALES IFS

Commençons par étudier la première famille de fractales découverte par Michael Barnsley en 1987: les "**systèmes de fonctions itérées déterministes**" ou souvent appelées en anglais "deterministic iterated function systems" (IFS).

De toutes les figures fractales, seules celles construites au moyen de systèmes de fonctions itérées affichent habituellement la propriété d'autosimilarité, signifiant que leur complexité est invariante par changement d'échelle.

Commençons par "borner" la chose... :

Nous nous donnons un objet géométrique initial E_0 de l'espace E , une fonction f de E dans E telle que:

$$f(E_0) \subset E_0 \quad (58.1)$$

(ce qui impose que l'objet initial ne pourra pas sortir de son propre domaine de définition via l'itération à travers la fonction f) et nous créons le système dynamique discret défini par:

$$S_D \begin{cases} E_0 \\ E_{n+1} = f(E_n) \end{cases} \quad (58.2)$$

Sous certaines conditions que nous allons de suite voir, la suite d'objets géométriques (E_n) "tend" vers une limite, qui est souvent un objet fractal (nous en verrons par ailleurs quelques exemples).

Naturellement, il existe un cadre mathématique rigoureux dans lequel les conditions évoquées et le verbe "tendre" ont une définition précise. En particulier, les objets E_n sont tous des compacts de E , c'est-à-dire des sous-ensembles bornés (que nous pouvons inclure dans un segment si E est une droite, un disque si E est un plan ou une boule si E est l'espace à trois dimensions) et fermés (toute suite convergente de E_n a sa limite dans E). Nous nous plaçons alors dans l'espace métrique des compacts, muni de la distance de Hausdorff (voir plus loin la définition), dont nous allons montrer qu'il est complet lorsqu'il s'agit de compacts du plan et de l'espace, et nous vérifierons que f est un "opérateur de Hutchinson", c'est à dire une application contractante de l'espace des compacts dans lui-même pour cette distance. Il ne restera alors plus qu'à appliquer le théorème du point fixe.

Des systèmes dynamiques de ce type sont dits déterministes, et donc appelés IFS (deterministic iterated function systems). Précisons que la limite de l'IFS s'appelle "l'attracteur de l'IFS". Nous pouvons montrer que sous les conditions évoquées plus haut, cet attracteur ne dépend pas de la forme de l'objet géométrique initial (nous verrons des exemples pratiques plus bas).

Dans un premier temps, nous limiterons notre étude à $\mathbb{R}$ (le cas général étant donné dans le chapitre de Topologie du site) sachant de toute façon qu'une généralisation à l'espace euclidien de dimension deux ne nécessite pas un travail intellectuel trop grand et que l'ensemble des complexes y est isomorphe.

Définition: Pour nous permettre de définir les frontières de nos fonctions fractales considérons $X \subseteq \mathbb{R}$. Nous disons que ξ est le "supremum" de X et nous notons:

$$\xi = \sup(X) \quad (58.3)$$

si ξ est le plus petit des "majorants" de X (un majorant de X est un nombre a qui vérifie $\forall x \in X, x \leq a$).

De la même façon, nous disons que ξ est "l'infimum" de X et nous notons:

$$\xi = \inf(X) \quad (58.4)$$

si ξ est le plus grand des "**minorants**" de X (un minorant de X est un nombre a qui vérifie $\forall x \in X, a \leq x$). Il existe des sous-ensembles de $\mathbb{R}$ qui n'ont pas de supremum (respectivement d'infimum) par exemple $[0, +\infty[$ (respectivement $]-\infty, 0]$).

Remarque: Nous utilisons souvent la caractérisation suivante du sup:

$$\xi = \sup(X) \quad (58.5)$$

si et seulement si:

$$\forall \varepsilon > 0, \exists x \in X, \xi - \varepsilon \leq x \leq \xi \quad (58.6)$$

ce qui est évident car nous pouvons nous approcher aussi près que l'on veut de ξ par des éléments de X (penser à ε petit). Pour information, nous avons alors aussi dans la même idée:

$$\xi = \inf(X) \quad (58.7)$$

si et seulement si:

$$\forall \varepsilon > 0, \exists x \in X, \xi \leq x \leq \xi + \varepsilon \quad (58.8)$$

Nous considérerons comme intuitif que si $X \subseteq \mathbb{R}$ est majoré, c'est-à-dire s'il existe $a \in \mathbb{R}$ tel que $\forall x \in X, x \leq a$ (respectivement minoré), alors X possède un supremum (respectivement un infimum).

Nous verrons plus tard que c'est cette propriété qui permettra de montrer que $\mathbb{R}$ est un "**espace métrique complet**"!

Remarque: En passant, soulignons l'importance de prendre $\mathbb{R}$ comme espace métrique de définition pour que cette propriété soit satisfaite. Nous pouvons effectivement remarquer qu'elle n'est pas vérifiée dans l'ensemble $\mathbb{Q}$ des nombres rationnels avec l'exemple simple suivant:

$$X = \{x \in \mathbb{Q} \mid x^2 \leq 2\} \quad (58.9)$$

qui est majoré mais n'a pas de supremum dans $\mathbb{Q}$ car ce supremum se situe dans $\mathbb{R}$ puisque:

$$x \leq \sqrt{2} \quad (58.10)$$

donc:

$$\sup(X) = \sqrt{2} \quad (58.11)$$

C'est ce qui fait que $\mathbb{Q}$ n'est pas "complet".

Définition: Nous disons que $X \subseteq \mathbb{R}$ est "**borné**" si X est majoré et minoré.

De la définition suit immédiatement que X est borné si et seulement s'il existe $a, b \in \mathbb{R}$ avec $a < b$ tels que $X \subseteq [a, b]$.

Maintenant que le concept de borne est relativement bien défini, voyons comment une suite peut s'y comporter:

Définition: Nous disons qu'une suite $(a_n)_{n \in \mathbb{N}}$ de $\mathbb{R}$ est une "**suite croissante**" (respectivement "décroissante") si:

$$\forall n \in \mathbb{N}, a_n \leq a_{n+1} \quad (58.12)$$

respectivement:

$$\forall n \in \mathbb{N}, a_n \geq a_{n+1} \quad (58.13)$$

Nous disons que la suite $(a_n)_{n \in \mathbb{N}}$ est "**monotone**" si elle est croissante ou décroissante comme nous l'avons déjà vu dans le chapitre de Suites Et Séries.

Définition: Soit $T = \{n_0, n_1, n_2, \dots\}$ un sous-ensemble infini de $\mathbb{N}$ avec $n_0 < n_1 < \dots$. Nous disons que la suite $(a_{n_i})_{i \in \mathbb{N}}$ est une "**sous-suite**" de la suite $(a_n)_{n \in \mathbb{N}}$.

Montrons maintenant que toute suite $(a_n)_{n \in \mathbb{N}}$ de $\mathbb{R}$ admet une sous-suite monotone (c'est un peu l'idée de fractale!).

Démonstration:

Nous disons que a_m est un "**pic de la suite**" si:

$$\forall n \geq m, a_m \geq a_n \quad (58.14)$$

Considérons l'ensemble P des pics de la suite $(a_n)_{n \in \mathbb{N}}$.

- Si P est infini alors la sous-suite $(a_n)_{n \in P}$ est monotone car décroissante.

- Si P est fini ou vide alors soit:

$$m_1 = \max(P) + 1 \quad (58.15)$$

(si $P = \emptyset$ nous choisissons m_1 quelconque). a_{m_1} n'est donc par construction pas un pic, donc il existe $m_2 \geq m_1$ tel que $a_{m_1} \leq a_{m_2}$. À son tour a_{m_2} n'est pas un pic, donc il existe $m_3 \geq m_2$ tel que $a_{m_2} \leq a_{m_3}$ etc. Nous voyons que nous définissons ainsi une sous-suite croissante.

□C.Q.F.D.

Définition: Nous disons que la suite $(a_n)_{n \in \mathbb{N}}$ "**converge**" vers $a \in \mathbb{R}$ et nous notons $\lim_{n \rightarrow \infty} a_n = a$ si:

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, n \geq N \Rightarrow |a_n - a| < \varepsilon \quad (58.16)$$

Nous disons dans ce cas que a est la "**limite de la suite**" $(a_n)_{\mathbb{N}}$.

Dans l'exemple de la figure ci-dessous où la suite semble converger vers 1.13 nous observons que pour un ε positif non nul particulier donné, il existe un n particulier que nous noterons N (valant 17) à partir duquel la suite converge.

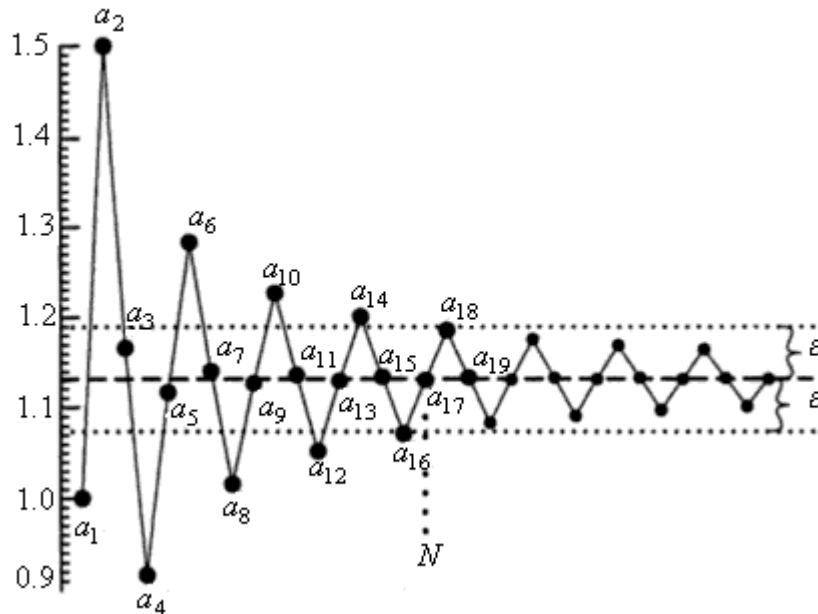


Figure: 58.2 - Illustration du principe de convergence d'une suite

S'il n'existe pas de a (respectivement de N) pour lequel la relation précédente est vraie, nous disons que la suite "**diverge**".

Démontrons maintenant que toute suite $(a_n)_{\mathbb{N}}$ croissante (resp. décroissante) et majorée (resp. minorée) converge.

En d'autres termes, nous cherchons à démontrer que toute suite $(a_n)_{\mathbb{N}}$ monotone et bornée converge (forcément... par construction).

Remarque: Si elle ne convergerait pas, nous pourrions difficilement savoir quel est son minorant et son majorant... d'où le fait que la nécessité du théorème devient triviale.

Démonstration:

Ce théorème est au fait assez intuitif. Considérons pour cela une suite croissante. Nous nous doutons que:

$$a = \sup\{a_0, a_1, a_2, \dots\} \quad (58.17)$$

est la limite de cette suite. Remarquons tout d'abord que $a = \sup\{a_0, a_1, a_2, \dots\}$ existe car $(a_n)_{\mathbb{N}}$ est... majorée (cf. premier théorème).

Soit $\varepsilon > 0$. Il existe un a_N tel que $a - \varepsilon \leq a_N \leq a$. Mais dans ce cas vu que la suite est croissante, nous avons $\forall n \geq N, a - \varepsilon \leq a_n \leq a$. C'est-à-dire $|a_n - a| \leq \varepsilon$. Dans le cas où la suite est décroissante en procédant de la même façon, nous montrons que $a = \inf \{a_0, a_1, a_2, \dots\}$ est la limite de cette suite.

□C.Q.F.D.

Et voici maintenant le résultat important à retenir suite à tout cela: Toute suite bornée de nombres réels possède une sous-suite convergente (c'est intuitif... mais encore une fois... formalisé cela devient parfois moins intuitif...).

C'est ce que les mathématiciens appellent le "**théorème de Bolzano-Weierstrass**" et il est extrêmement important dans de nombreux domaines des mathématiques:

Démonstration:

Soit $(a_n)_{\mathbb{N}}$ une telle suite. Par une proposition précédente nous savons qu'il existe une sous-suite monotone que nous noterons $(b_n)_{\mathbb{N}}$. $(b_n)_{\mathbb{N}}$ est donc une suite monotone et bornée et par le théorème précédent, $(b_n)_{\mathbb{N}}$ converge.

Donc, si nous n'arrivons pas à déterminer si la sous-suite converge ni sa limite exacte (ce qui dans la pratique est souvent très difficile), il nous suffit de savoir que la sous-suite est monotone et bornée pour nous assurer qu'elle converge (ce qui est beaucoup plus simple).

□C.Q.F.D.

Rappelons que nous avons vu dans le chapitre de Suites Et Séries qu'une suite de Cauchy, est une suite $(a_n)_{\mathbb{N}}$ qui vérifie (nous nous restreignons à un rappel particulier sur une distance euclidienne):

$$\forall \varepsilon > 0, \exists N \in \mathbb{N} \quad n, m \geq N \Rightarrow |a_n - a_m| \leq \varepsilon \quad (58.18)$$

La différence entre deux termes d'une suite de Cauchy peut être rendue arbitrairement petite pourvu que les indices de ces termes soient assez grands.

Nous avons aussi démontré (à nouveau dans le chapitre de Suites Et Séries) que dans le cas d'une distance dans le sens topologique général toute suite convergente est une suite de Cauchy (par contre la réciproque est pas toujours vraie à condition qu'on ne complète pas l'ensemble... sinon la réciproque est toujours vraie). Par exemple, une suite de nombres rationnels qui converge vers un réel n'est pas une suite de Cauchy, excepté si on complète l'espace des rationnels pour avoir l'espace des réels.

Refaisons la démonstration restreinte à la distance euclidienne (la méthode est exactement la même comme le lecteur pourra le remarquer):

Démonstration:

Soit $\varepsilon > 0$, nous devons montrer qu'il existe:

$$N \in \mathbb{N} \text{ tel que } n, m \geq N \Rightarrow |a_n - a_m| \leq \varepsilon \quad (58.19)$$

Mais $(a_n)_{\mathbb{N}}$ tend vers a donc il existe $N \in \mathbb{N}$ tel que $n \geq N \Rightarrow |a_n - a| \leq \varepsilon/2$. Pour $n, m \geq N$ nous avons donc:

$$|a_n - a_m| \leq |a_n - a| + |a - a_m| \leq \varepsilon/2 + \varepsilon/2 = \varepsilon \quad (58.20)$$

□C.Q.F.D.

Montrons maintenant que toute suite de Cauchy est bornée (nous n'en avons pas parlé jusque-là où que ce soit sur le site d'où la nécessité d'une démonstration). Puisqu'actuellement nous avons juste démontré que toute sous-suite convergente est une suite de Cauchy...

Démonstration:

Si $(a_n)_{\mathbb{N}}$ est une suite de Cauchy alors en particulier pour $\varepsilon = 1$ (choisi au hasard) nous savons qu'il existe $N \in \mathbb{N}$ tel que $n, m \geq N \Rightarrow |a_n - a_m| \leq 1$. Donc si nous fixons m , nous obtenons:

$$\forall n \geq N, |a_n| \leq 1 + |a_m| \quad (58.21)$$

□C.Q.F.D.

Voici à présent le théorème fondamental (c'est à ce niveau qu'il y a un impact énorme sur la compréhension de ce qu'est réellement une fractale!) découlant des quelques lignes précédentes.

Nous devons démontrer que toute suite de Cauchy de nombres réels est convergente (par construction...). Nous disons alors que l'espace métrique $\mathbb{R}$ muni de la distance euclidienne $|\cdot|$ (valeur absolue) est un "espace complet".

Remarques:

R1. La propriété de complétude est liée à la métrique (donc ce théorème aurait tout aussi bien sa place dans le chapitre de Topologie!): un même espace peut être complet pour une distance et incomplet pour une autre. Il est donc important de toujours préciser la distance que l'on prend quand on parle d'espace complet.

R2. Intuitivement, un espace est complet s'il n'a pas de trous. L'ensemble des rationnels n'est par exemple complet que si on lui ajoute les nombres réels.

Considérons d'abord $(a_n)_{\mathbb{N}}$ une suite de Cauchy. Nous avons vu que $(a_n)_{\mathbb{N}}$ est bornée donc par le théorème de Bolzano-Weierstrass, il existe une sous-suite $(a_{n_i})_{i \in \mathbb{N}}$ convergente. Notons a la limite de la sous-suite $(a_{n_i})_{i \in \mathbb{N}}$. Nous allons montrer maintenant que la suite $(a_n)_{\mathbb{N}}$ est convergente de limite a .

Démonstration:

Soit $\varepsilon > 0$, il existe $N \in \mathbb{N}$ tel que (application de la définition de la convergence pour une sous-suite):

$$i \geq N \Rightarrow |a_{n_i} - a| \leq \varepsilon/2 \quad (58.22)$$

Pour ce même ε il existe $M \geq 0$ tel que (application de la définition de la convergence pour une suite de Cauchy):

$$p, q \geq M \Rightarrow |a_p - a_q| \leq \varepsilon/2 \quad (58.23)$$

Soit $C = N + M$. Choisissons $i \geq C$. Nous avons donc, $|a_{n_i} - a| \leq \varepsilon/2$ et pour $n, i \geq C$, $|a_n - a_{n_i}| \leq \varepsilon/2$.
Donc par l'inégalité triangulaire (cf. chapitre de Calcul Vectoriel), pour tout $n \geq C$:

$$|a_n - a| \leq |a_n - a_{n_i}| + |a_{n_i} - a| \leq \varepsilon/2 + \varepsilon/2 = \varepsilon \quad (58.24)$$

Ce qui veut justement dire que $(a_n)_{n \in \mathbb{N}}$ converge vers a .

□C.Q.F.D.

Fondamentalement c'est un résultat intuitif mais à l'époque où les nombres réels n'étaient pas connus ou pas rigoureusement définis c'était une autre paire de manches. En réalité, il suffit de compléter tout ensemble par les nombres réels pour avoir un espace complet. Par ailleurs, certains mathématiciens définissent l'ensemble des réels en disant que c'est l'ensemble pour lequel toute suite de Cauchy converge.

Définition (intuitive): Un "**point adhérent**" est un point dont nous pouvons nous approcher autant que nous voulons à l'aide d'éléments d'un ensemble X donné (nous nous en approcherons par exemple avec une suite). Cependant, ce point adhérent peut aussi bien être à l'intérieur qu'à l'extérieur de X (tous les points à l'intérieur de X sont bien évidemment des points adhérents). Une bonne image est de voir une suite qui se rapproche de ce point adhérent et de définir des cercles autour de celui-ci qui deviennent de plus en plus petits contenant des éléments de la suite.

On peut imaginer comme exemple une suite définie par l'ensemble X des rationnels qui tend vers un irrationnel ou vers un nombre transcendant (ces deux points adhérents étant à l'extérieur de l'ensemble des rationnels). Donc dans ce cas, le point adhérent est extérieur à X (l'ensemble des rationnels). Par contre, tout point adhérent qui serait rationnel pour une suite de rationnels sera forcément... dans X .

Dès lors, vient la définition suivante:

Définition (formelle): Soit $X \subseteq \mathbb{R}^n$. Nous disons que $x \in \mathbb{R}^n$ est un "**point adhérent**" à X si pour toute boule $B(x, r)$ de rayon r centrée en x nous avons:

$$X \cap B(x, r) \neq \emptyset \quad (58.25)$$

L'ensemble des points adhérents à X est "**l'adhérence**" de X et est noté $\overline{X}$. Nous avons évidemment (il suffit de se le conceptualiser de manière abstraite pour toutes les boules possibles) $X \subseteq \overline{X}$. Sans oublier qu'il faut penser en termes d'ensemble de nombres pour X ...!

 **Exemple:**

Prenons l'intervalle $]0,1]$ avec la boule $B(0,1)$. L'intersection entre la boule et l'intervalle est non nulle, nous pouvons alors dire que 0 est adhérent! Mais maintenant prenons une suite $1/n$ par exemple, dans l'intervalle $]0,1]$. Cette suite tend vers zéro mais pourtant 0 n'appartient pas à l'intervalle. C'est donc bien un exemple qui montre que $X \subseteq \overline{X}$.

Nous pouvons faire dès lors la proposition suivante:

Montrons maintenant que $x \in \mathbb{R}^n$ est adhérent à X si et seulement si il existe une suite $(u_n)_{n \in \mathbb{N}}$ dans X qui converge vers x (attention, l'exemple précédent nous montre que x n'est pas nécessairement dans X).

Au fait, nous allons plutôt démontrer (si l'on peut appeler cela une démonstration...) que si nous choisissons un point adhérent x alors nous pouvons toujours trouver une suite $(u_n)_{n \in \mathbb{N}}$ dans X qui converge vers x .

Démonstration:

Si $x \in \mathbb{R}^n$ est adhérent à X alors considérons la suite des boules concentriques $B\left(x, \frac{1}{n}\right)$ avec $n \geq 1$ tel que:

$$X \cap B\left(x, \frac{1}{n}\right) \neq \emptyset \quad (58.26)$$

et alors il existe toujours des éléments u_n qui satisfont:

$$u_n \in \left(X \cap B\left(x, \frac{1}{n}\right)\right) \quad (58.27)$$

avec lesquels nous pouvons créer une suite par l'infinité des suites existantes.

□C.Q.F.D.

Définition: Nous disons que $X \subseteq \mathbb{R}^n$ est un "espace fermé" si $X = \overline{X}$.

Des propositions précédentes découle le fait que dans tout fermé F , une suite $(x_n)_{n \in \mathbb{N}}$ qui converge a sa limite dans F .

Nous considérerons comme trivial que si $(F_i)_{i \in I}$ est une famille de fermés indexée sur un ensemble I quelconque. Alors $\bigcap_{i \in I} F_i$ est fermé.

Définition: $X \subseteq \mathbb{R}^n$ est un "espace compact" si X est fermé et borné.

Le théorème suivant donne une caractérisation des compacts à partir des suites:

$X \subseteq \mathbb{R}^n$ est compact si et seulement si toute suite $(x_n)_{n \in \mathbb{N}}$ de X possède une sous-suite qui converge dans X .

Démonstration:

Montrons que X est fermé:

Si X est compact et $(x_n)_{n \in \mathbb{N}}$ est une suite de X alors par le théorème de Bolzano-Weierstrass, $(x_n)_{n \in \mathbb{N}}$ possède une sous-suite convergente de limite $x \in \mathbb{R}^n$. Mais puisque X est fermé, nous avons $x \in X$.

Réciproquement, supposons que toute suite $(x_n)_{n \in \mathbb{N}}$ de X possède une sous-suite qui converge dans X . Alors X est fermé car si $x \in \overline{X}$ il existe une suite $(x_n)_{n \in \mathbb{N}}$ de X qui tend vers x . Par hypothèse, $(x_n)_{n \in \mathbb{N}}$ possède une sous-suite qui converge vers $y \in X$. $(x_n)_{n \in \mathbb{N}}$ étant convergente toute les sous-suites convergent vers la même valeur, donc $x = y \in X$ (c'est pas beau ça ?!!). Ainsi $X = \overline{X}$ c'est-à-dire X est fermé.

Montrons que X est borné:

Supposons le contraire. Il existe donc une suite $(x_n)_{n \in \mathbb{N}}$ de X telle que $\|x_n\| \geq n$. Mais dans ce cas, aucune sous-suite de $(x_n)_{n \in \mathbb{N}}$ n'est convergente, ce qui est une contradiction. Donc X est borné. En conclusion, X est compact.

Une propriété des compacts est que si nous considérons $(A_n)_{n \in \mathbb{N}}$ une suite décroissante de compacts non vides, c'est-à-dire $A_{n+1} \subseteq A_n$, alors $\bigcap_{n \in \mathbb{N}} A_n$ est un compact non vide. Nous nous passerons de la démonstration qui est relativement triviale de par la définition du concept d'ensemble d'adhérence qui oblige qu'un compact soit par construction non vide... !

Exemple:

Nous obtenons l'ensemble C de Cantor de la manière suivante:

Nous commençons par considérer l'intervalle fermé borné $C_0 = [0,1]$ de $\mathbb{R}$ qui est donc un espace compact (ensemble borné et fermé). Nous partageons C_0 en trois parties égales et nous enlevons l'intervalle du milieu. Nous obtenons ainsi l'ensemble:

$$C_1 = [0, 1/3] \cup [2/3, 1] \quad (58.28)$$

qui peut être considéré aussi comme l'application d'une homothétie de facteur contractant $1/3$ sur l'intervalle fermé borné de départ dont on translate le centre d'homothétie.

Nous recommençons avec les deux intervalles $[0, 1/3], [2/3, 1]$ pour obtenir:

$$C_2 = [0, 1/9] \cup [2/9, 1/3] \cup [2/3, 7/9] \cup [8/9, 1] \quad (58.29)$$

réunion disjointe de quatre intervalles. Et ainsi de suite. Nous obtenons donc une suite décroissante C_n de compacts. Nous définissons:

$$C = \bigcap_{n \geq 0} C_n \quad (58.30)$$

Grâce à la proposition précédente, nous savons que C est non vide et qu'il est compact ce qui montre que les compacts ne sont pas tous "triviaux" comme des intervalles. L'ensemble de Cantor (car il avait joué avec en faisant le dessin comme ci-dessous en partant du bas) est un exemple de fractale (de compact):



Figure: 58.3 - Ensemble de Cantor avec Maple 4.00b

qu'il est possible d'obtenir avec le code Maple 4.00b suivant (si le copier/coller de la page web dans Maple 4.00b ne marche pas, réécrivez simplement le code):

```
>with(plots):
line := proc(a:: list, b:: list)
local plotoptionen, n;
if nargs > 2 then
plotoptionen := seq(args[n], n=3 .. nargs)
else
plotoptionen := NULL
fi;
plot([a, b], style=line, plotoptionen);
end:
cree_segment := (a,b,h) -> line([a,h],[b,h],color=black):
f1:=x->x/3: f2:=x->(x+2)/3:

f := s -> s union map(f1, s) union map(f2, s):

sequence_de_segments := proc(l,h)
local accu, i;
accu := NULL;
for i to nops(l) by 2 do
accu := accu,cree_segment(l[i], l[i+1], h) od;
accu
end:
```

```

Cantor := proc(n) local s, i;
option remember;
s := sequence_de_segments([0,1], 1);
for i from 1 to n do
s := sequence_de_segments(sort([op((f@@i)({0,1}))), (1-i/n)), s;
od;
display({s} union {seq(textplot([[0,(i+1/2)/n, '0'], [1, (i+1/2)/n, '1']]
), i=0 .. n)}, color=blue, axes=NONE, thickness=7)
end;

>Cantor(7);

```

Il est très intéressant de remarquer que l'on converge vers la fractale de Cantor (en termes de géométrie mais aussi de valeurs!) quel que soit le compact de départ choisi (l'intervalle fermé borné) et aussi... quel que soit le facteur contractant choisi!

Mandelbrot observa aussi ce type de structure autosimilaire lors de l'analyse de signaux transmis électriquement à l'époque par IBM sur des câbles de cuivre (IBM avait des problèmes de perte d'informations par transmission).

Regardons pour finir comment se comportent les compacts vis-à-vis des applications continues (nous en avons besoin pour montrer comment déterminer la distance d'un point à un ensemble ce qui nous sera indispensable après pour déterminer les propriétés de la distance de Hausdorff).

Nous rappelons (cf. chapitre d'Analyse Fonctionnelle) qu'une application $f : X \rightarrow \mathbb{R}^m$ où $X \subseteq \mathbb{R}^n$ est quelconque, est continue en un point $x \in X$ si:

$$\forall \varepsilon > 0, \exists \delta > 0, \forall y \in X, d(x, y) \leq \delta \Rightarrow d(f(x), f(y)) \leq \varepsilon \quad (58.31)$$

Ce qui traduit le fait que pour y assez proche de x , $f(y)$ est arbitrairement proche de $f(x)$. Nous disons aussi que f est continue sur X si elle est continue en tout point de X .

Proposition: Soit $f : X \rightarrow \mathbb{R}^m$ une application continue en $x \in X$ et $(x_n)_{\mathbb{N}}$ une suite de X avec:

$$\lim_{n \rightarrow \infty} x_n = x \in X \quad (58.32)$$

Alors la suite $f(x_n)$ converge et (cette proposition est très importante!):

$$\boxed{\lim_{n \rightarrow \infty} f(x_n) = f(x)} \quad (58.33)$$

En d'autres termes, si nous utilisons en tant qu'ensemble de départ les valeurs d'une suite convergente, alors la fonction qui prendra en entrée les valeurs de cette suite convergera elle aussi!

Démonstration:

Soit $\varepsilon > 0$. f est continue en x , donc il existe $\delta > 0$ tel que:

$$d(x_n, x) \leq \delta \Rightarrow d(f(x_n), f(x)) \leq \varepsilon \quad (58.34)$$

$(x_n)_{n \in \mathbb{N}}$ tend vers x donc il existe $N \in \mathbb{N}$ tel que:

$$n \geq N \Rightarrow d(x, x_n) \leq \delta \quad (58.35)$$

Par suite pour $n \geq N$, nous avons:

$$d(f(x_n), f(x)) \leq \varepsilon \quad (58.36)$$

□C.Q.F.D.

Si nous considérons maintenant $X \subseteq \mathbb{R}^n$ un compact et $f : X \rightarrow \mathbb{R}^m$ une application continue. $f(X)$ est compact. En particulier $\sup(f)$ et $\inf(f)$ seront atteints par définition et construction d'un compact (ensemble borné et fermé) qui est égal à son adhérence.

Autrement dit, une fonction à valeurs réelles continue sur un compact y atteint toujours son supremum ou son infimum.

Démonstration:

- Montrons que $f(X)$ est fermé: En effet, soit $f(x_n)$ une suite qui tend vers $y \in \overline{f(X)}$ (nous prenons l'adhérence au fait pour espérer montrer qu'elle est égale à l'ensemble lui-même) alors X étant compact, $(x_n)_{n \in \mathbb{N}}$ possède une sous-suite $(x_{n_i})_{i \in \mathbb{N}}$ convergente.

Posons:

$$\lim_{i \rightarrow \infty} x_{n_i} = x \in X \quad (58.37)$$

f est continue, donc:

$$\lim_{i \rightarrow \infty} f(x_{n_i}) = f(x) \quad (58.38)$$

Mais comme:

$$y = \lim_{n \rightarrow \infty} f(x_n) = \lim_{i \rightarrow \infty} f(x_{n_i}) \quad (58.39)$$

nous avons $y = f(x) \in f(X)$. Ceci prouve que:

$$f(X) = \overline{f(X)} \quad (58.40)$$

et donc que $f(X)$ est fermé.

- Montrons que $f(X)$ est borné: Supposons le contraire. Il existe donc une suite $f(x_n)$ telle que:

$$\|f(x_n)\| \geq n \quad (58.41)$$

pour tout n entier naturel (puisque justement il est supposé non borné). Soit $(x_{n_i})_{i \in \mathbb{N}}$ une sous-suite convergente de $(x_n)_{n \in \mathbb{N}}$ avec:

$$\lim_{i \rightarrow \infty} x_{n_i} = x \in X \quad (58.42)$$

Alors:

$$\lim_{i \rightarrow \infty} f(x_{n_i}) = f(x) \quad (58.43)$$

et par suite:

$$\lim_{i \rightarrow \infty} \|f(x_{n_i})\| = \|f(x)\| \quad (58.44)$$

mais ceci est en contradiction avec:

$$\lim_{i \rightarrow \infty} \|f(x_{n_i})\| = \lim_{i \rightarrow \infty} n_i = +\infty \quad (58.45)$$

Donc $f(X)$ est borné. Donc $f(X)$ étant fermé et borné il est compact.

□C.Q.F.D.

Appliquons maintenant cela (car c'est ce qui nous intéresse dans le cadre des espaces fractals) au calcul de la distance d'un point à un ensemble:

Soit $x \in \mathbb{R}^n$, l'application $f: \mathbb{R}^n \rightarrow \mathbb{R}$ définie par $f(y) = d(x, y)$ est continue.

Démonstration:

Pour tout $y, z \in \mathbb{R}^n$, l'inégalité triangulaire nous donne:

$$d(x, y) \leq d(x, z) + d(z, y) \Rightarrow d(x, y) - d(x, z) \leq d(z, y) \quad (58.46)$$

En changeant les rôles de y, z nous obtenons:

$$d(x, z) - d(x, y) \leq d(z, y) \quad (58.47)$$

et donc:

$$|d(x, z) - d(x, y)| \leq d(z, y) \quad (58.48)$$

Ainsi pour $\varepsilon > 0$ donné $d(z, y) \leq \varepsilon$ implique:

$$|d(x, z) - d(x, y)| \leq \varepsilon \quad (58.49)$$

c'est-à-dire:

$$|f(z) - f(y)| \leq \varepsilon \quad (58.50)$$

et f est donc continue en y .

□C.Q.F.D.

Définition: Pour $x \in \mathbb{R}^n$ et $A \subseteq \mathbb{R}^n$ nous définissons la distance de x à A comme étant la valeur:

$$d(x, A) = \inf \{d(x, a) \mid a \in A\} \quad (58.51)$$

Si $x \in A$ alors $d(x, A) = 0$ (trivial). La réciproque n'est pas vraie. En effet dans le cas $x = 0$ et $A =]0, 1]$ nous avons bien $d(x, A) = 0$ mais $x \notin A$. Nous avons donc la proposition (importante!):

$$d(x, A) = 0 \Leftrightarrow x \in \overline{A} \quad (58.52)$$

Démonstration:

$d(x, A) = 0$ entraîne l'existence d'une suite $(a_n)_{n \in \mathbb{N}}$ d'éléments de A telle que:

$$\lim_{n \rightarrow \infty} d(x, a_n) = 0 \quad (58.53)$$

ce qui veut dire:

$$\lim_{n \rightarrow \infty} a_n = x \quad (58.54)$$

donc $x \in \overline{A}$ (voir développements plus hauts).

Réciproquement, si $x \in \overline{A}$ alors pour tout $\varepsilon > 0$ il existe $a \in A$ tel que $d(x, a) \leq \varepsilon$. Mais $d(x, A) \leq d(x, a) \leq \varepsilon$. Ainsi pour tout $\varepsilon > 0$, $d(x, A) \leq \varepsilon$. C'est-à-dire:

$$d(x, A) = 0 \quad (58.55)$$

□C.Q.F.D.

En général la distance de x à A n'est pas atteinte. C'est-à-dire qu'il n'existe pas de $a \in A$ tel que $d(x, A) = d(x, a)$ il suffit pour cela de considérer l'exemple $x = -1$ et $A =]0, 1]$ nous avons $d(x, A) = 1$ mais pour tout $a \in A$, $d(x, a) > 1$. Si A est compact, la situation est bien évidemment différente selon la proposition (la plus importante pour la distance de Hausdorff) suivante:

Si $A \subseteq \mathbb{R}^n$ est compact, il existe $a \in A$ tel que $d(x, A) = d(x, a)$. Ainsi:

$$d(x, A) = \min \{d(x, y) \mid y \in A\} \quad (58.56)$$

Démonstration:

L'application $f: A \rightarrow \mathbb{R}$ définie par $f(a) = d(x, a)$ est continue comme déjà montré. Par conséquent $f(A)$ est compact (cf. une proposition précédente). Ainsi, f atteint ses bornes, c'est-à-dire, il existe $a \in A$ tel que $f(a) = \inf(f(A))$. Donc:

$$d(x, a) = f(a) = \inf(f(A)) = d(x, A) \quad (58.57)$$

Remarque: La proposition précédente ne dit pas que a est unique, d'ailleurs en général, il en existe plusieurs.

1.1. ESPACE MÉTRIQUE DES FRACTALES

Les fractales sont souvent perçues par les gens comme de jolis dessins sur une feuille, mais lorsque nous voulons regarder en détail la géométrie fractale, nous avons besoin d'un espace particulier où l'étudier, un peu comme le biologiste qui met des petits vers sur une plaquette pour les observer en détail au microscope. Nous allons faire de même pour nos fractales en les plaçant dans un endroit qu'ils apprécient.

Cet endroit a de fortes chances d'être un sous-espace de $\mathbb{R}^2$ ou $\mathbb{R}^3$, puisqu'en fin de compte il s'agira de produire des dessins, et pour illustrer nos propos nous nous placerons souvent dans le cas $X = \mathbb{R}^2$ (avec la métrique euclidienne) et sauf mention du contraire, nous considérerons toujours le cas où (X, d) est un espace métrique complet.

Rassemblons différents éléments afin de pouvoir construire cet espace:

Définition: Nous définissons $\mathcal{H}(X)$ comme l'espace dont les points sont les sous-ensembles compacts de X , autres que l'ensemble. Désormais nous appellerons "**fractale**" n'importe quel élément de $\mathcal{H}(X)$.

Exemple:

Il est immédiat que si $x, y \in \mathcal{H}(X)$, alors $x \cup y \in \mathcal{H}(X)$, mais $x \cap y$ n'est pas forcément dans $\mathcal{H}(X)$. Il suffit de voir la figure avec les deux ensembles compacts (fermés, bornés donc) ci-dessous de $\mathbb{R}^2$. Ce sont donc deux points de $\mathcal{H}(X)$. Leur réunion est encore un ensemble compact, et donc:

$$z = x \cup y \in \mathcal{H}(\mathbb{R}^2) \quad (58.58)$$

Par contre, si les ensembles sont disjoints (comme ici), $x \cap y = \emptyset$ et par conséquent n'est pas un point de $\mathcal{H}(\mathbb{R}^2)$ (voir la théorie précédente).

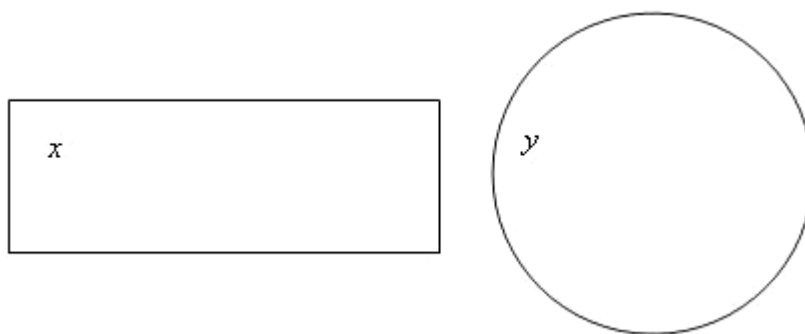


Figure: 58.4 - Source: IFS et L-Système, V. Rezzonico, C. Hebeisen

Un autre exemple consiste à prendre la fractale de Cantor...

Définition: Soit $x \in X$ et $B \in \mathcal{H}(X)$, nous définissons la distance d'un point x à l'ensemble B , et nous la notons $d(x, B)$ comme étant:

$$d(x, B) = \min\{d(x, y) \mid y \in B\} \quad (58.59)$$

Remarques:

R1. Cette définition est tout à fait générale et s'applique à n'importe quel sous-ensemble non vide de X , en remplaçant min par inf. Mais dans le cas particulier, nous sommes intéressés à prendre précisément $\mathcal{H}(X)$ comme sous-espace.

R2. Cette distance est bien définie (elle existe) du fait que B est non vide et compact.

R3. Il est trivial de voir que si cette distance est nulle, alors $x \in \bar{B}$.

Exemple:

Illustration dans le cas où $X = \mathbb{R}^2$:

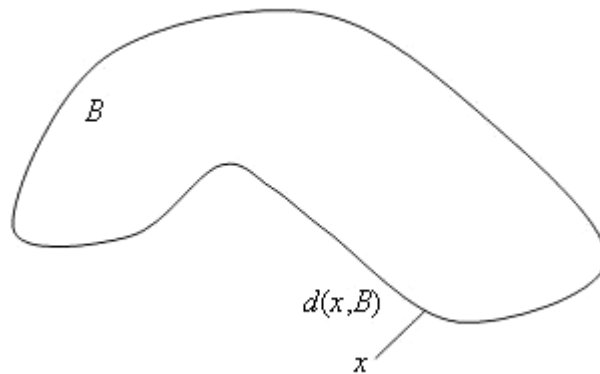


Figure: 58.5 - Source: IFS et L-Système, V. Rezzonico, C. Hebeisen

Définition: Soient $A, B \in \mathcal{H}(X)$. Nous définissons la distance de A à B et nous la notons $d(A, B)$ comme étant:

$$d(A, B) = \max\{d(x, B) \mid x \in A\} \quad (58.60)$$

Remarques:

R1. Comme avant, cette définition a un sens, et en particulier il existe deux points $\hat{x} \in A, \hat{y} \in B$ tels que $d(A, B) = d(\hat{x}, \hat{y})$.

R2. Nous constatons que cette distance ne fournit pas de métrique $\mathcal{H}(X)$: en effet, $d(A, B) \neq d(B, A)$ en général (prendre par exemple la fractale de Cantor où pour certains compacts nous avons $A \subset B$ avec $A \neq B$, nous aurons alors $d(A, B)=0$ mais $d(B, A) > 0$).

Définition: Soient $A, B \in \mathcal{H}(X)$. Nous définissons la "**distance de Hausdorff**" entre deux ensembles $A, B \in \mathcal{H}(X)$, et nous la notons $h(A, B)$, comme étant:

$$h(A, B) = \max \{d(A, B), d(B, A)\} \quad (58.61)$$

Cette fois-ci, de par cette dernière définition, nous avons bien une métrique sur $\mathcal{H}(X)$.

En effet, vérifions que les 5 propriétés d'une distance soient vérifiées (cf. chapitre de Topologie):

Soient $A, B, C \in \mathcal{H}(X)$. Clairement nous avons sans démonstration (symétrie, nullité sur la diagonale et séparation):

$$\begin{aligned} h(A, B) &= h(B, A) \\ h(A, A) &= d(A, A) = 0 \\ h(A, B) &= 0 \Leftrightarrow A = B \end{aligned} \quad (58.62)$$

De plus, comme A et B sont compacts, $h(A, B) = d(a, b)$ (cf. une des propositions précédentes) pour un certain $a \in A$ et un certain $b \in B$. Or, puisque $d(a, b) > 0$ par définition nous avons (propriété de positivité) finalement $h(A, B) > 0$ tel que $a \in A, a \notin B$:

$$h(A, B) = \max \{d(A, B), d(B, A)\} \geq d(A, B) = \max \{d(a, B) \mid a \in A\} \geq d(a, B) > 0 \quad (58.63)$$

puisque B est fermé.

Enfin, puisque $h(A, B) = d(a, b)$ (cf. extension d'une des propositions précédentes), l'inégalité triangulaire est alors forcément respectée et alors:

$$h(A, B) \leq h(A, C) + h(C, B) \quad (58.64)$$

Donc h est bien une métrique sur $\mathcal{H}(X)$, ce qui fait de $(\mathcal{H}(X), h)$ un espace métrique. C'est déjà un premier pas dans la direction souhaitée: nous avons désormais les moyens de comparer deux ensembles appartenant à $\mathcal{H}(X)$ par la distance de Hausdorff qui les sépare. Si les deux ne sont pas "trop différents", alors intuitivement cette distance devrait être assez petite.

Si nous choisissons une fonction $f: X \rightarrow X$ strictement contractante de constante λ . Alors, l'application:

$$T_\lambda: \mathcal{H}(X) \rightarrow \mathcal{H}(X) \quad (58.65)$$

définie par

$$T_\lambda(A) = f(A) \quad (58.66)$$

est par construction aussi strictement contractante de constante λ .

Soit $f_i: \mathbb{R}^n \rightarrow \mathbb{R}^n$, $1 \leq i \leq k$ des applications strictement contractantes de constantes de contraction, $0 \leq \lambda_i < 1$. Alors, il existe un unique compact $A \in \mathcal{H}(\mathbb{R}^n)$ tel que:

$$A = T_{\lambda_1, \dots, \lambda_k}(A) = f_1(A) \cup \dots \cup f_k(A) \quad (58.67)$$

(A est l'unique point fixe de $T_{\lambda_1, \dots, \lambda_k}$) et pour tout compact B , on a:

$$\lim_{m \rightarrow \infty} T_{\lambda_1, \dots, \lambda_k}^m(B) = A \quad (58.68)$$

où $T_{\lambda_1, \dots, \lambda_k}^m(B)$ est le m -ième itéré de B par $T_{\lambda_1, \dots, \lambda_k}$.

Ce résultat découle du théorème du point fixe (cf. chapitre Suites Et Séries) appliqué à l'espace $\mathcal{H}(\mathbb{R}^n)$ qui est complet.

Avec les mêmes notations, nous disons dit que $f_1, \dots, f_k$ est un codage IFS (Iterated Function Systems) du compact A . Ainsi les fonctions $f_1, \dots, f_k$ définissent le compact A . Ce qui est surprenant, comme nous allons le voir dans les quelques exemples qui suivent, c'est que les $f_1, \dots, f_k$ sont généralement assez simples (comme des homothéties du plan) tandis que le compact A est dans bien des cas relativement "compliqué".

Le cas $k = 1$ est sans intérêt, on aurait $A = \{x\}$ où x est le point fixe de f_1 . Avec $k = 2$ nous obtenons déjà des résultats non triviaux.

Remarque: Lorsque les fonctions itératives contractantes sont toutes des homothéties, nous parlons alors de "fractale de Sierpinski". Ainsi, la fractale de Cantor appartient à la famille des fractales de Sierpinski.

Une méthode fréquemment utilisée pour générer informatiquement des fractales IFS (comme ce sera le cas ci-dessous avec Maple 4.00b) est de considérer un point dans le plan (x_n, y_n) auquel nous pouvons sans autre forme de procès appliquer une transformation affine pour obtenir un nouveau point (x_{n+1}, y_{n+1}) tel que:

$$\begin{aligned} x_{n+1} &= ax_n + by_n + e \\ y_{n+1} &= cx_n + dy_n + f \end{aligned} \quad (58.69)$$

où a, b, c, d, e et f sont des constantes quelconques, et (x_0, y_0) est donné.

Nous pouvons dès lors considérer une application W qui décrit notre transformation, et sous forme matricielle nous pouvons écrire le système précédent comme suit:

$$W \begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix} \quad (58.70)$$

ou encore:

$$W(\vec{x}) = A\vec{x} + \vec{b} \quad (58.71)$$

De façon tout à faire générale, le vecteur $\vec{b}$ décrit simplement une translation, et la matrice A est la composition de rotations et d'un changement d'échelle (cf. chapitre de Géométrie Euclidienne). Les programmes informatiques (comme ce sera le cas dans les exemples plus bas), ne demandent donc souvent que les six paramètres a, b, c, d, e et f .

1.2. FRACTALE DE CANTOR

Revoyons la fractale de Cantor vue plus haut mais cette fois-ci avec le point de vue de l'application de deux fonctions itératives contractantes f_1, f_2 (donc correspondant à $k=2$).

Nous partons donc de l'ensemble borné fermé suivant:

$$[[0,1]] \quad (58.72)$$

soit:



Figure: 58.6 - Ensemble de départ de la fractale de Cantor

Nous partageons donc en trois parties égales et nous enlevons l'intervalle du milieu. Nous obtenons ainsi l'ensemble:

$$[[0,1/3],[2/3,1]] \quad (58.73)$$

Soit:



Figure: 58.7 - Première itération de la fractale de Cantor

Nous pouvons constater que $[0,1/3]$ peut être obtenu par l'homothétie de facteur $1/3$ centrée en $(0,0)$ suivante:

$$f_1 = h_{(x,y)}^1 = h_{(0,0)}^{1/3} \quad (58.74)$$

et que $[2/3,1]$ peut être obtenu par l'homothétie de facteur $1/3$ centrée en $(1,0)$ suivante:

$$f_2 = h_{(x,y)}^2 = h_{(1,0)}^{1/3} \quad (58.75)$$

et ainsi de suite, nous obtenons comme nous le savons déjà (voir code Maple 4.00b déjà donné plus haut):



Figure: 58.8 - Attracteur de Cantor après 6 itérations

Ce qui correspond en reprenant le formalisme vu plus haut à:

$$\begin{aligned}
 A_0 &= [0,1] \\
 A_1 &= f_1([0,1]) \cup f_2([0,1]) = \left[\left[\frac{0}{3}, \frac{1}{3} \right], \left[\frac{2}{3}, 1 \right] \right] \quad (58.76) \\
 A_2 &= f_1(A_1) \cup f_2(A_1) = \dots
 \end{aligned}$$

Mais voyons que cela marche avec n'importe quel compact de l'ensemble des réels comme un carré par exemple avec le code Maple 4.00b suivant (nous montrerons toujours tous les détails de Maple 4.00b, car rien ne dit que les lecteurs le possèdent ni que le logiciel existera toujours dans 50 ans...).

Attention!!! Si le copier/coller de la page web dans Maple 4.00b ne marche pas, réécrivez simplement le code.

```

>transforme_point := proc(t, p)
    [t[1]*p[1]+t[2]*p[2]+t[5], t[3]*p[1]+t[4]*p[2]+t[6]]
end:

>IFSS := proc(n, liste_de_transformations,col)
    local i, j, k, s, seq_square:
    seq_square := [[0,0],[1,0],[1,1],[0,1]];
    for j to n do
        s := NULL;
        for i to nops(liste_de_transformations) do
            s := s,
            seq(transform_square(liste_de_transformations[i],
            op(k, [seq_square])),
            k=1 .. nops([seq_square]))
        od;
        seq_square := s
    od;
    plots[polygonplot]([seq_square], axes=none, color=col, scaling=constrained)
end:

```

```
>cantor:=[[evalf(1/3),0,0,evalf(1/3),0,0],[evalf(1/3),0,0,evalf(1/3),evalf(2/3),0]]:
```

```
>IFSS(1, cantor,blue);
```



Figure: 58.9 - Première itération sur l'ensemble de Cantor avec des carrés

```
>IFSS(2, cantor,blue);
```



Figure: 58.10 - Deuxième itération sur l'ensemble de Cantor avec des carrés

```
>IFSS(3, cantor,blue);
```



Figure: 58.11 - Troisième itération sur l'ensemble de Cantor avec des carrés

etc.

Donc, quel que soit l'ensemble de départ, la suite de compacts obtenue par application successive de ces deux homothéties du plan converge toujours (au sens de la distance de Hausdorff) vers le même compact/attracteur (assimilé au point fixe du théorème du point fixe...) appelée fractale de Cantor (appartenant donc à la famille des fractales de Serpienski). La dernière figure ci-dessus est une bonne approximation de cet ensemble.

1.3. FRACTALE DU TRIANGLE (TAMIS) DE SIERPINSKI

Pour construire la fractale de Sierpinski (que l'on retrouve comme curiosité sur le coquillage Cymbiola innexa REEVE), basée sur trois fonctions itératives contractantes f_1, f_2, f_3 (donc correspondant à $k=3$), nous partons par exemple des trois points de $\mathbb{R}^2$ suivants:

$$[[0, 0], [1, 0], [0.5, 1]] \quad (58.77)$$

Ce qui donne avec Maple 4.00b:

```
>plots[polygonplot]([ [0, 0], [1, 0], [0.5, 1]], axes=none, color=black, scaling=constrained);
```

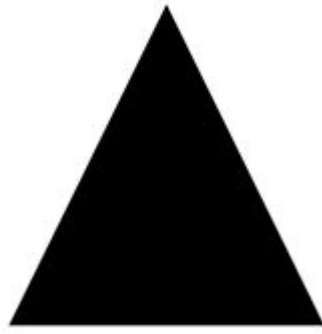


Figure: 58.12 - Ensemble de départ du triangle de Sierpinski

c'est un triangle, mais nous pourrions partir de n'importe quelle forme et nous arriverions toujours au même résultat que nous allons voir plus loin.

Nous appliquons sur chaque ensemble une fonction contractante de facteur 0.5, ce qui donne le triangle:

$$[[0,0],[0.5,0],[0.25,0.5]] \quad (58.78)$$

et nous noterons cette homothétie de facteur 0.5 et de centre (0,0) sur le triangle d'origine:

$$f_1 = h_{(x,y)}^A = h_{(0,0)}^{1/2} \quad (58.79)$$

Nous effectuons maintenant sur ce triangle une translation de 0.5 dans la direction de l'axe des X, ce qui donne le triangle:

$$[[0.5,0], [1,0], [0.75,0.5]] \quad (58.80)$$

ce qui correspond à une homothétie de facteur 0.5 et de centre (1,0) sur le triangle d'origine:

$$f_2 = h_{(x,y)}^A = h_{(1,0)}^{1/2} \quad (58.81)$$

Nous translatons maintenant $[[0,0],[0.5,0],[0.25,0.5]]$ de 0.25 selon l'axe de X et de 0.5 selon l'axe des Y pour avoir:

$$[[0.25,0.5], [0.75,0.5], [0.5, 1]] \quad (58.82)$$

ce qui correspond à une homothétie de facteur 0.5 et de centre (0.5,0.75) sur le triangle d'origine:

$$f_3 = h_{(x,y)}^A = h_{(0.5,0.75)}^{1/2} \quad (58.83)$$

Avec Maple 4.00b cela donne maintenant pour les trois triangles:

```
>plots[polygonplot]( [[[0,0],[0.5,0],[0.25,0.5]], [[0.5,0],[1,0],[0.75,0.5]], [[0.25,0.5],[0.75,0.5],
, [0.5,1]]], axes=none,color=black, scaling=constrained);
```



Figure: 58.13 - Première itération du triangle de Sierpinski

et ainsi de suite:



Figure: 58.14 - Deuxième itération du triangle de Sierpinski

et ainsi de suite:



Figure: 58.15 - Troisième itération du triangle de Sierpinski

et ainsi de suite:

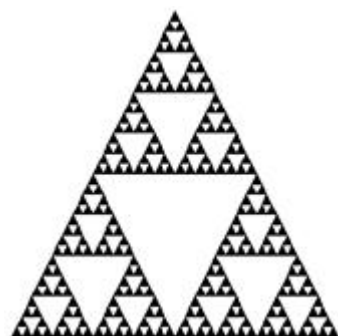


Figure: 58.16 - Quatrième itération du triangle de Sierpinski

et ainsi de suite:

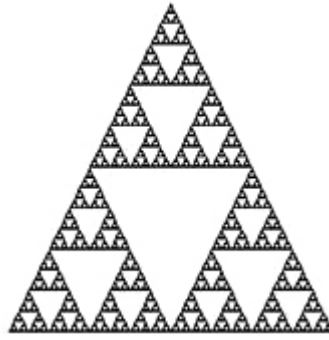


Figure: 58.17 - Cinquième itération du triangle de Sierpinski

etc.

Ce qui correspond en reprenant le formalisme vu plus haut:

$$\begin{aligned} \mathcal{A}_0 &= [[0,0], [1,0], [0.5,1]] \\ \mathcal{A}_1 &= f_1([0,0]) \cup f_2([1,0]) \cup f_3([0.5,1]) \quad (58.84) \\ &\dots \end{aligned}$$

Nous pouvons faire la même remarque que lorsque nous avons présenté la fractale de Cantor la toute première fois: quel que soit l'ensemble de départ, la suite de compacts obtenue par application successive de ces trois homothéties converge toujours (au sens de la distance de Hausdorff) vers le même compact/attracteur (assimilé au point fixe du théorème du point fixe...) appelé triangle Serpienski. La dernière figure ci-dessus est une bonne approximation de cet ensemble.

Voyons cela avec un code Maple 4.00b (si le copier/coller de la page web dans Maple 4.00b ne marche pas, réécrivez simplement le code).

```
> transforme_triangle := proc(t, triangle)
    local i;
    [seq(transforme_point(t, triangle[i]), i=1 .. 3)]
end;

> IFS := proc(n, liste_de_transformations, col)
    local i, j, k, s, sequence_de_triangles;
    options `Copyright by Alain Schaubert, 1996`;
    sequence_de_triangles := [[0, 0], [1, 0], [0.5, 1]];
    for j to n do
        s := NULL;
        for i to nops(liste_de_transformations) do
            s := s,
            seq(transforme_triangle(liste_de_transformations[i],
            op(k, [sequence_de_triangles])),
            k=1 .. nops([sequence_de_triangles]))
        od;
        sequence_de_triangles := s
    od;
    plots[polygonplot]([sequence_de_triangles], axes=none, color=col, scaling=constrained)
end;
```

```
> triangle_de_Sierpinski:=[[0.5,0,0,0.5,0,0],[0.5,0,0,0.5,0.5,0],[0.5,0,0,0.5,0.25,0.5]]:
>IFS(6, triangle_de_Sierpinski,blue);
```

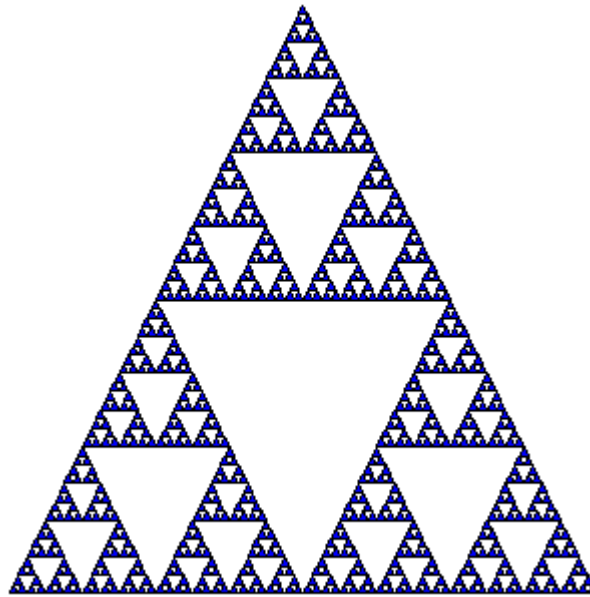


Figure: 58.18 - Attracteur du triangle de Sierpinski

Et cette fois-ci nous partons non plus d'un triangle, mais d'un carré (IFS Square) avec le code Maple 4.00b suivant:

```
> transform_square := proc(t, square)
  local i;
  [seq(transforme_point(t, square[i]), i=1 .. 4)]
end:

>IFSS := proc(n, liste_de_transformations,col)
  local i, j, k, s, seq_square:
  seq_square :=[[0,0],[1,0],[1,1],[0,1]];
  for j to n do
    s := NULL;
    for i to nops(liste_de_transformations) do
      s := s,
      seq(transform_square(liste_de_transformations[i],
      op(k, [seq_square])),
      k=1 .. nops([seq_square]))
    od;
    seq_square := s
  od;
  plots[polygonplot]([seq_square], axes=none, color=col, scaling=constrained)
end:

> triangle_de_Sierpinski:=[[0.5,0,0,0.5,0,0],[0.5,0,0,0.5,0.5,0],[0.5,0,0,0.5,0.25,0.5]]:
>IFSS(1, triangle_de_Sierpinski,green);
```

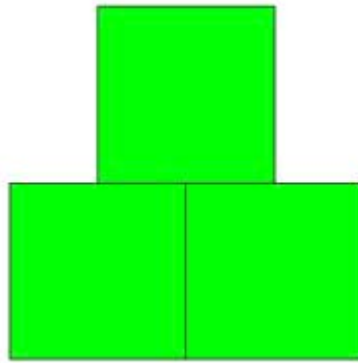



Figure: 58.19 - Première itération du triangle de Sierpinski avec des carrés

```
>IFSS(2, triangle_de_Sierpinski,green);
```

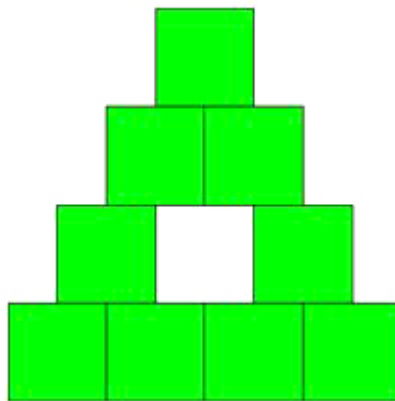


Figure: 58.20 - Deuxième itération du triangle de Sierpinski avec des carrés

```
> IFSS(3, triangle_de_Sierpinski,green);
```

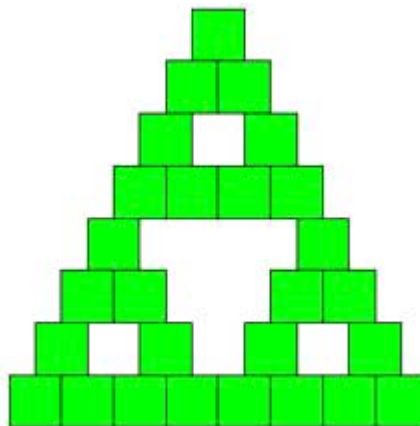


Figure: 58.21 - Troisième itération du triangle de Sierpinski avec des carrés

etc. jusqu'à:

```
> IFSS(6, triangle_de_Sierpinski,green);
```

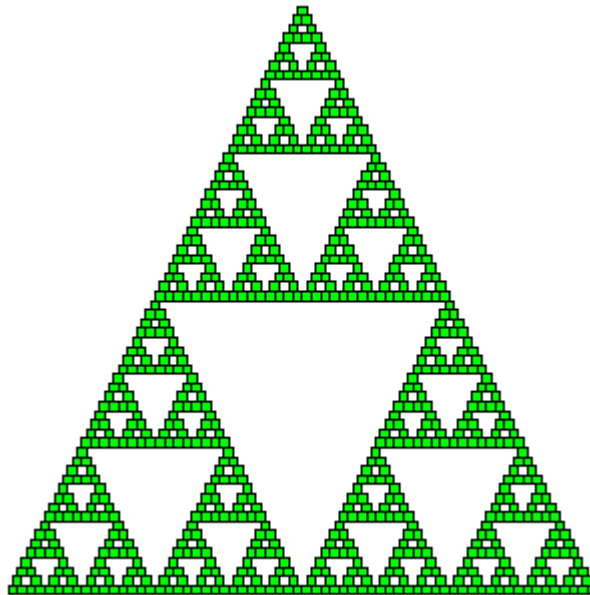


Figure: 58.22 - Attracteur du triangle de Sierpinski avec des carrés

Basiquement, le triangle de Sierpinski peut bien évidemment être aussi vu comme un triangle auquel on enlève le triangle du milieu et où pour chacun des triangles restants, on recommence la procédure!

1.4. FRACTALE DU TAPIS DE SIERPINSKI

Le tapis de Sierpinski est l'attracteur de 8 fonctions itératives contractantes d'homothéties de rapport $1/3$ centrées aux sommets et aux milieux des côtés d'un carré dans lequel peut se trouver n'importe quelle forme géométrique.

Cette fois-ci, dans $\mathbb{R}^2$ nous considérons les huit homothéties (h):

$$h_{(0,0)}^{1/3}, h_{(0,0.5)}^{1/3}, h_{(0,1)}^{1/3}, h_{(0.5,1)}^{1/3}, h_{(1,1)}^{1/3}, h_{(1,0.5)}^{1/3}, h_{(1,0)}^{1/3} \text{ et } h_{(0.5,0)}^{1/3} \quad (58.85)$$

et nous partons par exemple des quatre points de $\mathbb{R}^2$ suivants:

$$[[0, 0], [1, 0], [1, 1], [0, 1]] \quad (58.86)$$

ce qui correspond à un carré plein (mais nous pourrions choisir n'importe quoi d'autre!):



Figure: 58.23 - Ensemble de départ du tapis de Sierpinski

Après l'application des huit fonctions d'homothéties (nous laissons le soin au lecteur de faire manuellement les calculs comme nous les avons déjà détaillés pour le triangle), nous obtenons la forme suivante composée de huit carrés:

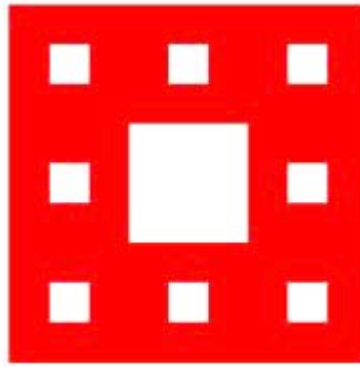


Figure: 58.24 - Deuxième itération du tapis de Sierpinski

et en réappliquant les huit homothéties encore une fois (heureusement qu'il y a l'ordinateur...):

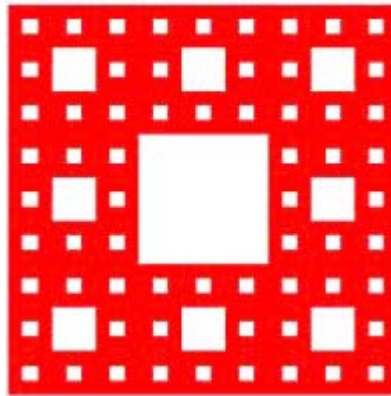


Figure: 58.25 - Troisième itération du tapis de Sierpinski

et encore une fois:

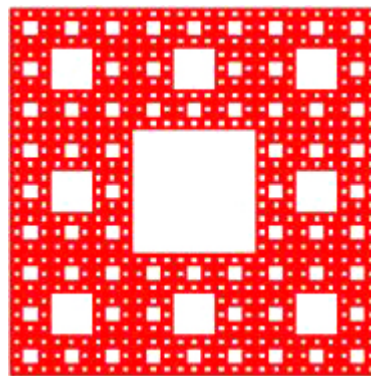


Figure: 58.26 - Quatrième itération du tapis de Sierpinski

etc.

Le point fixe obtenu (attracteur) s'appelle donc cette fois-ci "tapis de Sierpinski" et c'est la forme qu'à l'antenne réceptrice de la majorité de nos téléphones portables en ce début de 21ème siècle.

Les figures précédentes peuvent être obtenues successivement avec le code Maple 4.00b suivant (si le copier/coller de la page web dans Maple 4.00b ne marche pas, réécrivez simplement le code):

```

>transforme_point := proc(t, p)
  [t[1]*p[1]+t[2]*p[2]+t[5], t[3]*p[1]+t[4]*p[2]+t[6]]
end:

transform_square := proc(t, square)
  local i;
  [seq(transforme_point(t, square[i]), i=1 .. 4)]
end:

>IFSS := proc(n, liste_de_transformations,col)
  local i, j, k, s, seq_square:
  seq_square :=[[0,0],[1,0],[1,1],[0,1]];
  for j to n do
    s := NULL;
    for i to nops(liste_de_transformations) do
      s := s,
      seq(transform_square(liste_de_transformations[i],
      op(k, [seq_square])),
      k=1 .. nops([seq_square]))
    od;
    seq_square := s
  od;
  plots[polygonplot]([seq_square], axes=none, color=col, scaling=constrained)
end:

> dywan:= [[evalf(1/3),0,0,evalf(1/3),0,0],[evalf(1/3),0,0,evalf(1/3),evalf(1/3),0],
[evalf(1/3),0,0,evalf(1/3),evalf(2/3),0], [evalf(1/3),0,0,evalf(1/3),0,evalf(2/3)],
[evalf(1/3),0,0,evalf(1/3),evalf(1/3),evalf(2/3)], [evalf(1/3),0,0,evalf(1/3),evalf(2/3),evalf(2/3)],
[evalf(1/3),0,0,evalf(1/3),0,evalf(1/3)],[evalf(1/3),0,0,evalf(1/3),evalf(2/3),evalf(1/3)]];

> IFSS(0, dywan, blue);

```

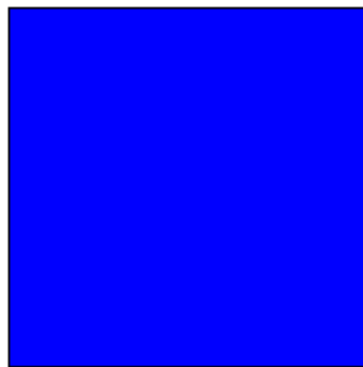


Figure: 58.27 - Ensemble de départ du tapis de Sierpinski

```

> IFSS(1, dywan, blue);

```

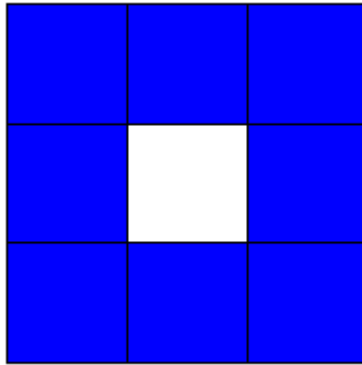


Figure: 58.28 - Première itération du tapis de Sierpinski

> IFSS(2, dywan, blue);

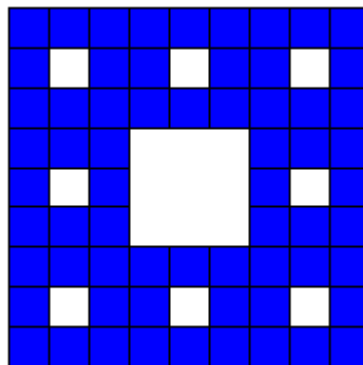


Figure: 58.29 - Deuxième itération du tapis de Sierpinski

> IFSS(3, dywan, blue);

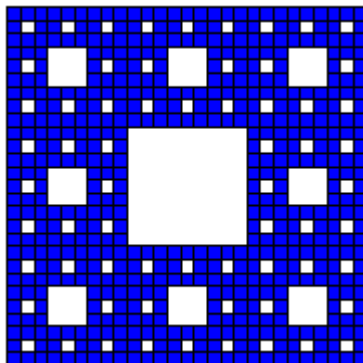


Figure: 58.30 - Troisième itération du tapis de Sierpinski

> IFSS(4, dywan, blue);

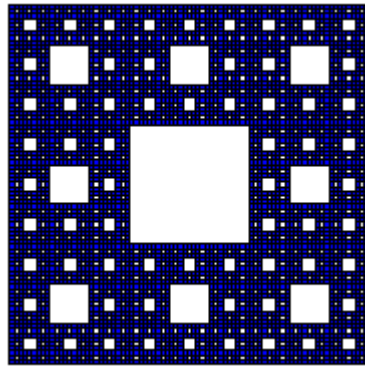


Figure: 58.31 - Quatrième itération du tapis de Sierpinski

1.5. FRACTALE SPIRALE

Nous avons vu deux fractales de la famille des fractales de Sierpinski basées donc uniquement sur des homothéties contractantes. Voyons maintenant une fractale qui combine rotation et homothétie contractantes.

Dans $\mathbb{R}^2$ nous considérons les deux applications d'homothéties (h) de rotations (R) suivantes:

$$h_{(0.25,0.5)}^{0.2} \circ R_{(0.25,0.5)}^{5\pi/6} \quad h_{(0.5,0.5)}^{0.95} \circ R_{(0.5,0.5)}^{\pi/6} \quad (58.87)$$

Avec un triangle et toujours avec Maple 4.00b, cela nous donne (si le copier/coller de la page web dans Maple 4.00b ne marche pas, réécrivez simplement le code):

```
>transforme_triangle := proc(t, triangle)
  local i;
  [seq(transforme_point(t, triangle[i]), i=1 .. 3)]
end;

>IFS := proc(n, liste_de_transformations,col)
  local i, j, k, s, sequence_de_triangles;
  options `Copyright by Alain Schaubert, 1996`;
  sequence_de_triangles := [[0, 0], [1, 0], [0.5, 1]];
  for j to n do
    s := NULL;
    for i to nops(liste_de_transformations) do
      s := s,
      seq(transforme_triangle(liste_de_transformations[i],
      op(k, [sequence_de_triangles])),
      k=1 .. nops([sequence_de_triangles]))
    od;
    sequence_de_triangles := s
  od;
  plots[polygonplot]([sequence_de_triangles], axes=none, color=col, scaling=constrained)
end;
```

```

>a:=evalf(5*Pi/6);b:=evalf(Pi/6);
> c1x:=0.25;c1y:=0.5;c2x:=0.5;c2y:=0.5;
> h1:=0.2;h2:=0.95;
> spirale:=[[h1*cos(a),-h1*sin(a),h1*sin(a),h1*cos(a),(1-h1*cos(a))*c1x+h1*sin(a)*c1y,
-h1*sin(a)*c1x+(1-h1*cos(a))*c1y],[h2*cos(b),-h2*sin(b),h2*sin(b),h2*cos(b),
(1-h2*cos(b))*c2x+h2*sin(b)*c2y,-h2*sin(b)*c2x+(1-h2*cos(b))*c2y]]:

>IFS(1,spirale,blue);

```

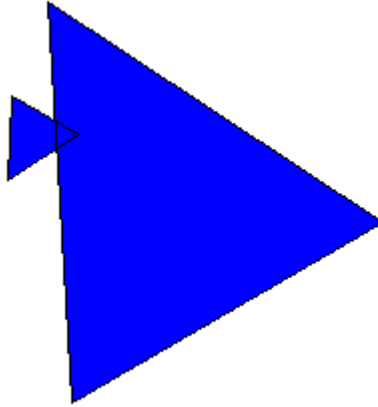


Figure: 58.32 - Première itération de la fractale spirale

et encore et comme la convergence est très longue, nous allons faire par pas de 5 itérations.

```

>IFS(6,spirale,blue);

```

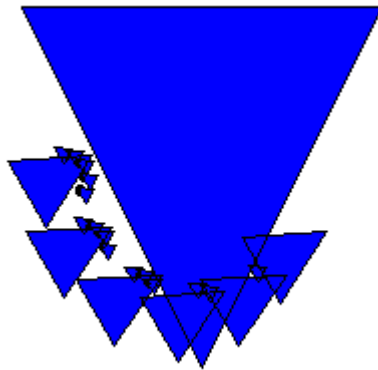


Figure: 58.33 - Deuxième itération de la fractale spirale

```

>IFS(11,spirale,blue);

```

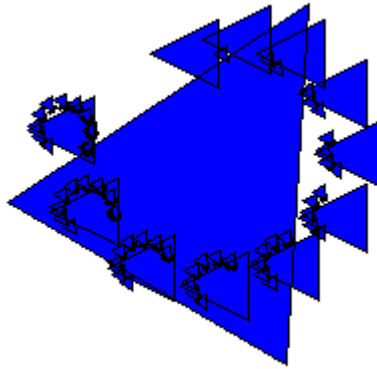


Figure: 58.34 - Troisième itération de la fractale spirale

>IFS(16,spirale,blue);

Nous obtenons après plusieurs centaines d'itérations le point fixe suivant:

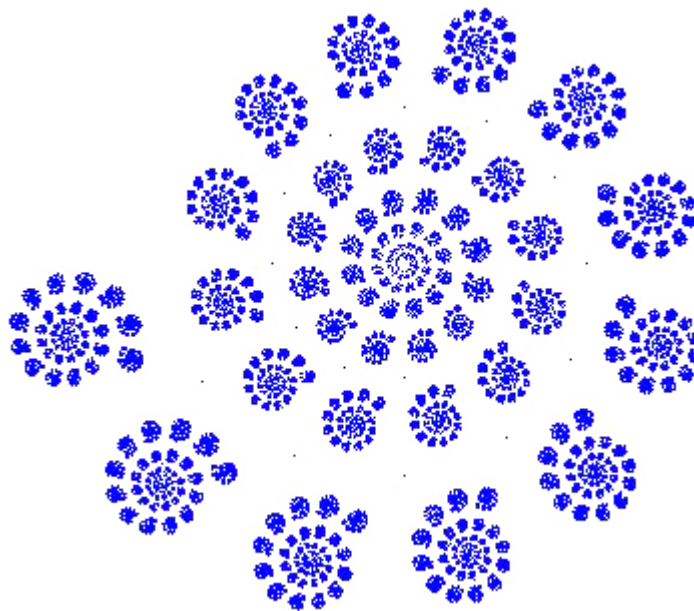


Figure: 58.35 - Attracteur de la fractale spirale

1.6. FRACTALE DE VON KOCH

Toujours, dans les fractales obtenues par homothéties (h) et rotations (R) contractantes mais auxquelles nous rajoutons en plus une translation (T), la courbe de Von Koch est une fractale assez connue, elle peut être obtenue par les applications suivantes:

$$h_{(0,0)}^{1/3} \quad h_{(1,0)}^{1/3} \quad T_{(1/3,0)} \circ h_{(0,0)}^{1/3} \circ R_{(0,0)}^{\pi/3} \quad T_{(-1/3,0)} \circ h_{(1,0)}^{1/3} \circ R_{(1,0)}^{-\pi/3} \quad (58.88)$$

Donnons directement le résultat avec Maple 4.00b (si le copier/coller de la page web dans Maple 4.00b ne marche pas, réécrivez simplement le code):


```

>koch := proc(p:: numeric)
  local m, n, k, l, s, h, x, y, pts, t, i;
  h := 3^(-p);
  pts := table([]): # [0, 0];
  pts[0]:=[0,0];
  x := 0; y := 0;
  for n from 0 to (4^p) do
    m := n;
    s := 0;
    for l from 0 to p-1 do
      t := irem(m, 4);
      m := iquo(m, 4);
      s := s+irem((t+1), 3) - 1
    od; # end of for l
    x := evalhf(x+cos(Pi*s/3)*h);
    y := evalhf(y+sin(Pi*s/3)*h);
    pts[n+1] := [x, y];
  od;
  [seq(pts[i], i=0 .. n-1)];
end:

> plot(koch(0), scaling=constrained, style=LINE, axes=NONE, color=blue,thickness=2);

```



Figure: 58.36 - Ensemble de départ de la fractale de Von Koch

```

> plot(koch(1), scaling=constrained, style=LINE, axes=NONE, color=blue,thickness=2);

```



Figure: 58.37 - Première itération de la fractale de Von Koch

```

> plot(koch(2), scaling=constrained, style=LINE, axes=NONE, color=blue,thickness=2);

```

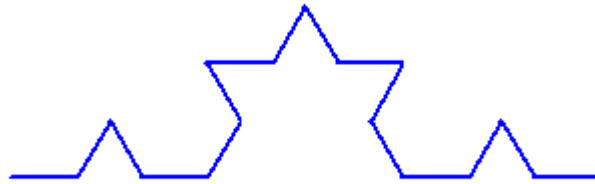


Figure: 58.38 - Deuxième itération de la fractale de Von Koch

```
> plot(koch(3), scaling=constrained, style=LINE, axes=NONE, color=blue,thickness=2);
```



Figure: 58.39 - Troisième itération de la fractale de Von Koch

etc. etc. Jusqu'à obtenir l'attracteur suivant:

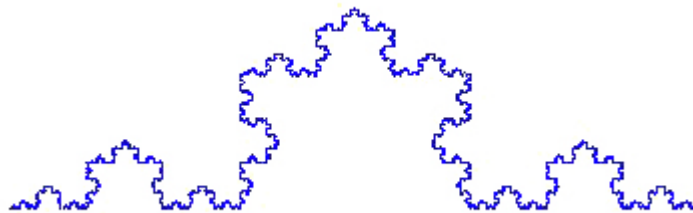


Figure: 58.40 -Attracteur de la fractale de Von Koch

Ce qui est aussi dérangent avec la fractale de Von Koch c'est que nous partons d'une ligne de longueur finie, pour arriver à la fin à une ligne de longueur infinie si nous réitérons la structure à l'infini. Pourtant visuellement elle est finie... c'est une courbe "pathologique" comme disent parfois les mathématiciens dans le domaine.

Effectivement, la première intuition conduit à penser que le périmètre de cette figure tend vers une valeur limite finie, puisqu'on ajoute des détails de plus en plus petits au fur et à mesure des itérations successives. En réalité, à la première itération la longueur L de chaque côté est remplacée par 4 segments de longueur $L/3$; À la deuxième, elle devient $16 L/9$... À chaque itération la longueur est donc multipliée par $4/3$, ce qui signifie que (contrairement à l'intuition première) la longueur d'une courbe de Koch tend vers l'infini pour un nombre d'itérations infini (série géométrique de raison $4/3$). Et, pourtant, cette courbe ne déborde à aucun moment des limites constituées à l'extérieur par le cercle circonscrit au triangle initial, et à l'intérieur par le cercle inscrit dans ce triangle! En d'autres termes une surface de dimension finie est limitée par une frontière de longueur infinie.

Cette courbe de Koch est légendaire, car elle a servi à Mandelbrot à écrire un article concernant le problème de mesure de la longueur des côtes des littoraux de mer (car plus l'unité de mesure de base prise était petite, plus le périmètre du littoral était grand). Il proposa de considérer les littoraux comme des fractales dont il est impossible de mesurer le périmètre mais bien "l'arborescence fractale", soit en d'autres termes: la dimension fractale.

Lorsque nous accolons trois courbes de Koch aux sommets d'un triangle équilatéral, nous obtenons une élégante figure à symétrie hexagonale dénommée flocon de Koch (ou Île de Koch):

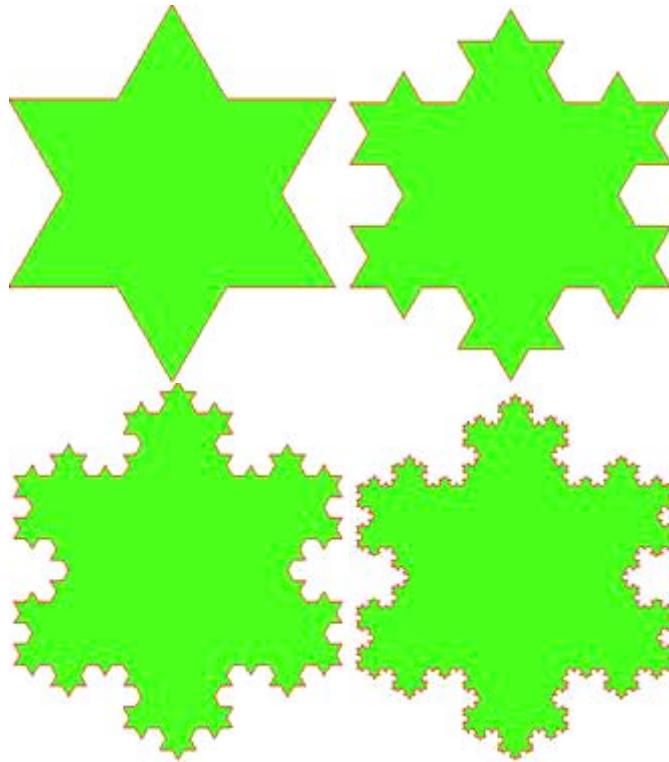


Figure: 58.41 - Exemples de l'île de Koch

1.7. FRACTALES NATURELLES

Outre l'aspect purement mathématique des fractales, on peut via des méthodes heuristiques trouver les applications contractantes pour des fractales similaires aux formes que nous pouvons retrouver dans la nature. Voyons quelques exemples toujours avec Maple 4.00b en prenant d'abord pour base commune de toutes les fractales qui vont suivre, les procédures suivantes (si le copier/coller de la page web dans Maple 4.00b ne marche pas, réécrivez simplement le code).

```
>transforme_point := proc(t, p)
  [t[1]*p[1]+t[2]*p[2]+t[5], t[3]*p[1]+t[4]*p[2]+t[6]]
end;
```

```
transforme_triangle := proc(t, triangle)
  local i;
  [seq(transforme_point(t, triangle[i]), i=1 .. 3)]
end;
```

```
transform_square := proc(t, square)
  local i;
  [seq(transforme_point(t, square[i]), i=1 .. 4)]
end;
```

```

>IFS := proc(n, liste_de_transformations,col)
  local i, j, k, s, sequence_de_triangles:
  options `Copyright by Alain Schaubert, 1996`;
  sequence_de_triangles := [[0, 0], [1, 0], [0.5, 1]];
  for j to n do
    s := NULL;
    for i to nops(liste_de_transformations) do
      s := s,
      seq(transforme_triangle(liste_de_transformations[i],
        op(k, [sequence_de_triangles])),
        k=1 .. nops([sequence_de_triangles]))
    od;
    sequence_de_triangles := s
  od;
  plots[polygonplot]([sequence_de_triangles], axes=none, color=col, scaling=constrained)
end:

```

```

> IFSS := proc(n, liste_de_transformations,col)
  local i, j, k, s, seq_square:
  seq_square := [[0,0],[1,0],[1,1],[0,1]];
  for j to n do
    s := NULL;
    for i to nops(liste_de_transformations) do
      s := s,
      seq(transform_square(liste_de_transformations[i],
        op(k, [seq_square])),
        k=1 .. nops([seq_square]))
    od;
    seq_square := s
  od;
  plots[polygonplot]([seq_square], axes=none, color=col, scaling=constrained)
end:

```

1.7.1. RAMEAU

On part de:

```

>rameau:=[[.387,.430,.430,-.387,.2560,.5220], [.441,-.091,-.009,-.322,.4219,.5059],
[-.468,.020,-.113,.015,.4,.4]]:

```

Et on obtient:

```

> IFSS(0,rameau,green);

```

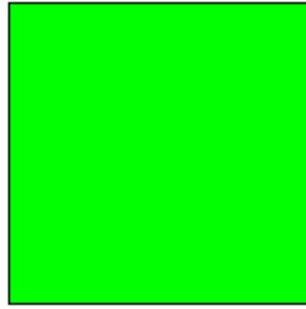


Figure: 58.42 - Ensemble de départ de la fractale du rameau

```
> IFSS(1,rameau,green);
```

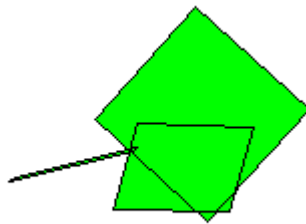


Figure: 58.43 - Première itération pour la fractale du rameau

```
> IFSS(2,rameau,green);
```

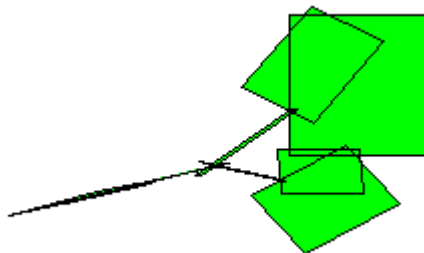


Figure: 58.44 - Deuxième itération pour la fractale du rameau

```
> IFSS(3,rameau,green);
```

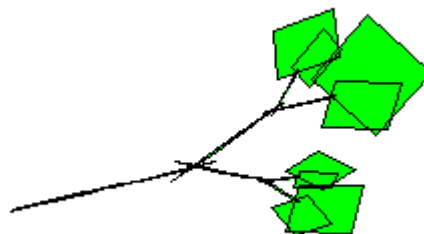


Figure: 58.45 - Troisième itération pour la fractale du rameau

```
> IFSS(4,rameau,green);
```

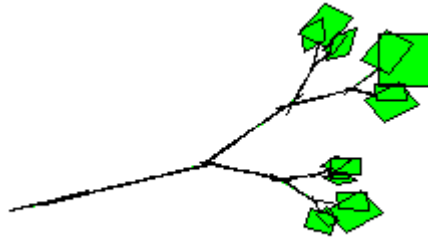


Figure: 58.46 - Quatrième itération pour la fractale du rameau

```
> IFSS(5,rameau,green);
```

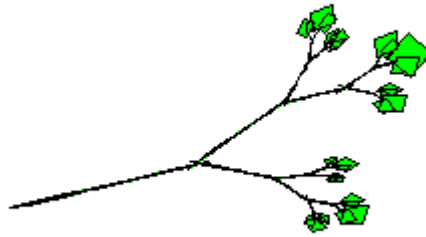


Figure: 58.47 - Cinquième itération pour la fractale du rameau

```
> IFSS(6,rameau,green);
```



Figure: 58.48 - Sixième itération pour la fractale du rameau

1.7.2. FLOCON DE NEIGE

On part de:

```
> cristal:=[[.255,0,0,.255,.3726,.6714],[.255,0,0,.255,.1146,.2232],  
[.255,0,0,.255,.6306,.2232],[.37,-.642,.642,.37,.6356,-.0061]]:
```

Et on obtient:

```
> IFSS(0, cristal,green);
```

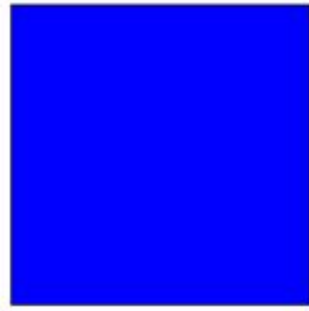


Figure: 58.49 - Ensemble de départ pour le flocon de neige

```
> IFSS(1, cristal,green);
```

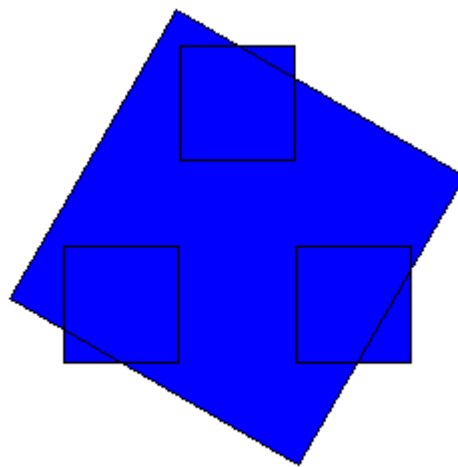


Figure: 58.50 - Première itération pour la fractale du flocon de neige

```
>IFSS(2, cristal,green);
```

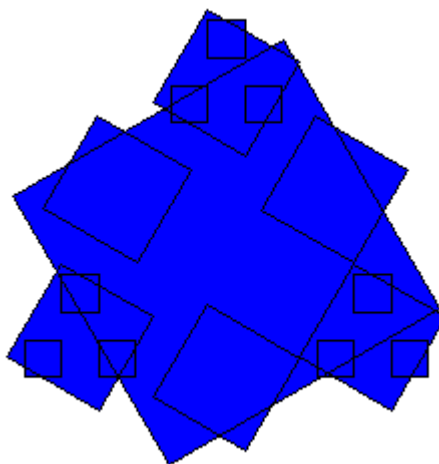


Figure: 58.51 - Deuxième itération pour la fractale du flocon de neige

```
>IFSS(3,cristal,green);
```

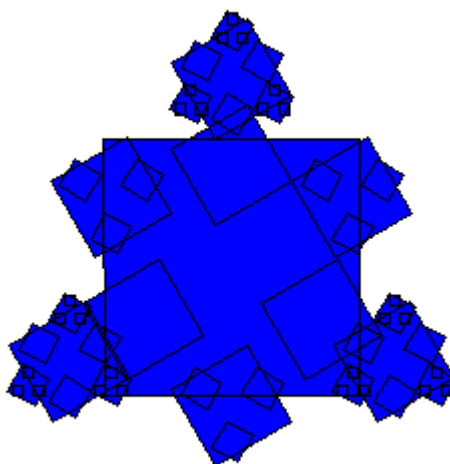


Figure: 58.52 - Troisième itération pour la fractale du flocon de neige

>IFSS(4,cristal,blue);

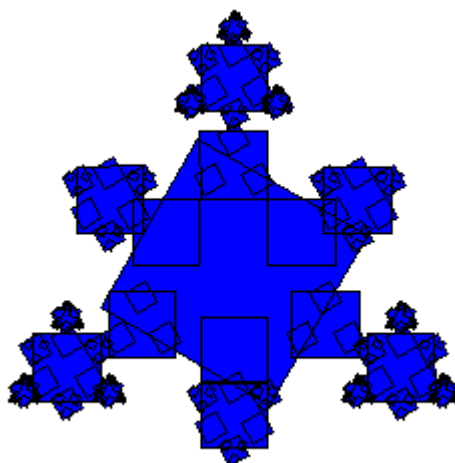


Figure: 58.53 - Quatrième itération pour la fractale du flocon de neige

>IFSS(5,cristal,blue);

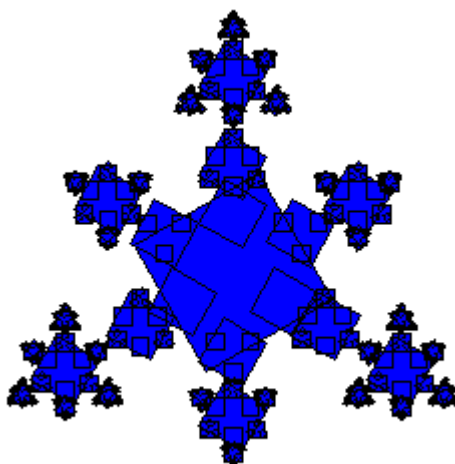


Figure: 58.54 - Cinquième itération pour la fractale du flocon de neige


```
>IFSS(6,cristal,blue);
```

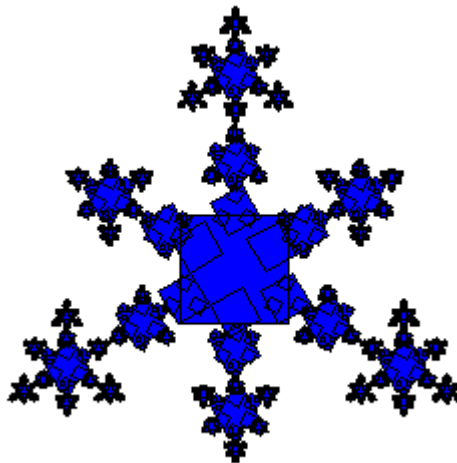


Figure: 58.55 - Sixième itération pour la fractale du flocon de neige

```
>IFSS(7,cristal,blue);
```

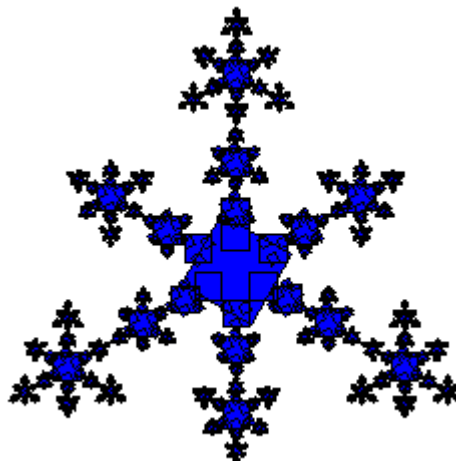


Figure: 58.56 - Septième itération pour la fractale du flocon de neige

1.7.3. ARBRE

On part de:

```
> tree := [[-0.04, 0, -0.23, -0.65, -0.08, 0.26], [0.61, 0, 0, 0.31, 0.07, 2.5],  
[0.65, 0.29, -0.3, 0.48, 0.54, 0.39], [0.64, -0.3, 0.16, 0.56, -0.56, 0.4]];
```

```
> IFS(0, tree, green);
```

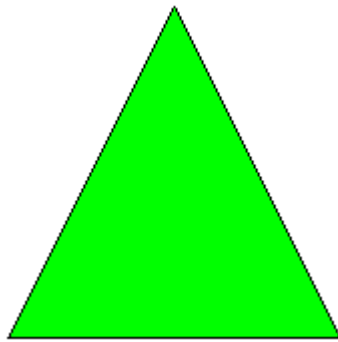


Figure: 58.57 - Ensemble de départ pour la fractale de l'arbre

```
> IFS(1, tree, green);
```

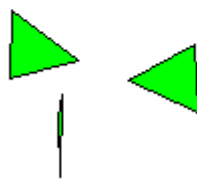


Figure: 58.58 - Première itération la fractale de l'arbre

```
> IFS(2, tree, green);
```

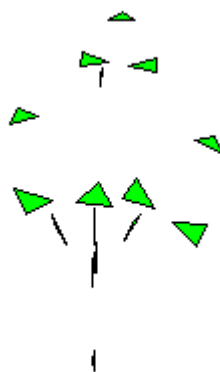


Figure: 58.59 - Deuxième itération pour la fractale de l'arbre

```
> IFS(3, tree, green);
```

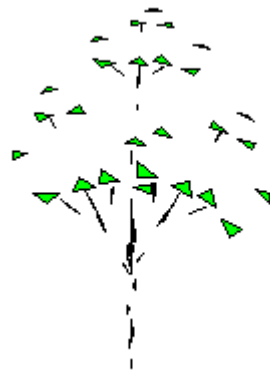


Figure: 58.60 - Troisième itération pour la fractale de l'arbre

> IFS(4, tree, green);

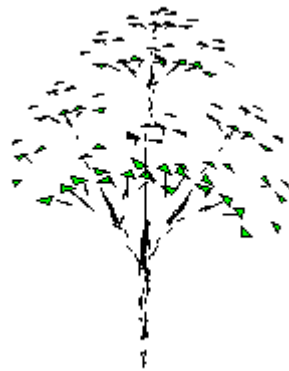


Figure: 58.61 - Quatrième itération pour la fractale de l'arbre

> IFS(5, tree, green);



Figure: 58.62 - Cinquième itération pour la fractale de l'arbre

1.7.4. FOUGÈRE

On part de:

```
> fougere:=[[0,0,0,0.16,0,0],[0.2,-0.26,0.23,0.22,0,1.6],[-0.15,0.28,0.26,0.24,0,0.44],  
,[0.85,0.04,-0.04,0.85,0,1.6]]:  
  
> IFS(0, fougere, blue);
```

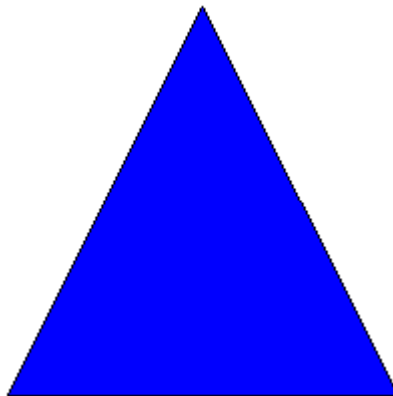


Figure: 58.63 - Ensemble de départ pour la fractale de la fougère

```
> IFS(1, fougere, blue);
```



Figure: 58.64 - Première itération pour la fractale de la fougère

```
> IFS(2, fougere, blue);
```



Figure: 58.65 - Deuxième itération pour la fractale de la fougère

```
> IFS(3, fougere, blue);
```



Figure: 58.66 - Troisième itération pour la fractale de la fougère

```
> IFS(4, fougere, blue);
```



Figure: 58.67 - Quatrième itération pour la fractale de la fougère

```
> IFS(5, fougere, blue);
```



Figure: 58.68 - Cinquième itération pour la fractale de la fougère

```
> IFS(6, fougere, blue);
```



Figure: 58.69 - Sixième itération pour la fractale de la fougère

> IFS(7, fougere, blue);



Figure: 58.70 - Septième itération pour la fractale de la fougère

etc. Jusqu'à obtenir:

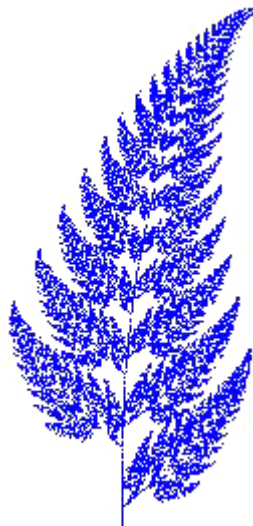


Figure: 58.71 - Attracteur de la fractale de la fougère

et nous allons nous arrêter ici car les exemples sont non dénombrables...

2. FRACTALES À TEMPS D'ÉCHAPPEMENT

Plusieurs méthodes ont donc été proposées pour construire des images fractales comme nous l'avons mentionné tout au début de ce chapitre. Nous allons donc maintenant nous intéresser aux méthodes dites "**méthodes d'échappement**".

Pour cela, on se place dans le plan complexe formé des points M de coordonnées (x, y) d'affixe:

$$z = x + iy \quad (58.89)$$

où i représente le nombre complexe tel que:

$$i^2 = -1 \quad (58.90)$$

On considère une suite complexe définie par:

$$u(0) = z_0 \quad (58.91)$$

et:

$$u(n+1) = f(u(n)) \quad (58.92)$$

f étant une fonction continue complexe. On suppose que f a un point fixe x_0 , c'est-à-dire qu'il existe x_0 tel que:

$$f(x_0) = x_0 \quad (58.93)$$

Il s'agit donc simplement de l'application du théorème du point fixe déjà mentionné plusieurs fois jusqu'ici. Sous certaines conditions sur f et sur z_0 , on constate que la suite des points ne diverge pas (ce qui veut dire qu'on ne s'intéresse pas qu'aux points qui convergent mais à tous ceux qui ne divergent pas!). Cette méthode est à la base de la construction des ensembles de Mandelbrot et de Julia.

Construire une image fractale à partir d'un ensemble de suites ainsi définies, revient à étudier pour chaque couple (x, y) du plan le comportement de la suite. On associe alors une couleur à chaque suite (c'est-à-dire à chaque couple (x, y)) représentant la "**rapidité**" de divergence de la suite.

Pour étudier la convergence d'une suite, on regarde ses n premiers éléments, si on détecte que les conditions de divergence sont vérifiées alors on peut dire que cette suite diverge, sinon, cette suite est potentiellement convergente. On remarque que plus n est grand, plus les résultats seront précis (mais plus le temps de calcul sera grand).

L'algorithme de base est le suivant:

Fractal=proc(x,y)

z:= valeur de z0;

j:=nombre max itérations

```

Tant que condition_de_divergence non vérifiée(z) et j non atteint faire
    z:=formule_iteration(z)
    changer de couleur;
Fin Tant que
Renvoyer couleur finale

```

Fin Fractal

2.1. ENSEMBLE DE MANDELBROT

On construit l'ensemble de Mandelbrot grâce à des itérations dans le plan complexe (nous parlons alors de "**dynamique holomorphe**"). La fonction est de la forme:

$$f(z) = z^2 + c \quad (58.94)$$

où c est un paramètre constant tel que $c \in \mathbb{C}$. Le premier terme de la suite est nul. On a donc la suite U définie par:

$$z_0 = 0 \text{ et } z_{n+1} = z_n^2 + c \quad (58.95)$$

Pourquoi commence-t-on avec $z_0 = 0$? Car zéro est le point critique de $z^2 + c$, c'est-à-dire le point qui satisfait à l'extremum:

$$\frac{d(z^2 + c)}{dz} = 0 \quad (58.96)$$

Pour chaque point d'affixe $x+iy$ du plan, on étudie la suite U pour $c = x + iy$. Si la suite diverge, on dit que le point testé n'appartient pas à l'ensemble M , si la suite converge, on dit que le point appartient à M .

Pour reproduire l'ensemble de Mandelbrot, on associe à c des valeurs du plan complexe. On considère généralement la portion du plan complexe ayant comme partie réelle, les valeurs entre -2.5 et 1.5 et comme partie imaginaire, les valeurs entre -1.5 et 1.5. Cette portion du plan complexe est subdivisée de façon à former une grille dont les éléments seront associés à des valeurs de C . Pour chaque valeur de C , on obtient une suite dont les modules peuvent converger ou diverger.

En pratique, on considère que la suite des modules converge si les 30 premiers modules sont inférieurs à 2. Lorsque la suite des modules converge, on colorie en noir le point de la grille. Après avoir considéré tous les points de la grille, on obtient un ensemble de points noircis: "**l'ensemble de Mandelbrot**" noté M . Ce qui constitue un résultat remarquablement curieux!

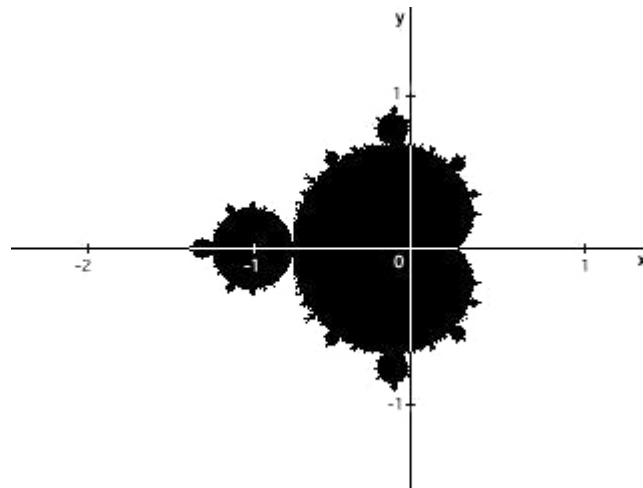


Figure: 58.72 - Ensemble de Mandelbrot

La liste des z_i générés par l'itération s'appelle "**l'orbite**" de z_0 .

On peut colorier les points à l'extérieur de l'ensemble de Mandelbrot en utilisant des couleurs qui dépendent du nombre de termes calculés avant d'obtenir un module supérieur ou égal à 2. Les points d'une même couleur peuvent être interprétés comme étant des points s'éloignant à la même vitesse de l'ensemble de Mandelbrot.

On peut aussi faire une incursion dans l'ensemble de Mandelbrot en utilisant Maple 4.00b (disponible habituellement au collège). Il suffit de copier le programme ci-dessous sur une feuille de travail du logiciel et d'indiquer à la place de -2 .. 1, -1.5 .. 1.5 de la dernière ligne, l'étendue des parties réelles et imaginaires de c que l'on désire visualiser:

```
>restart: with(plots):
>couleur:=proc(a,b)
local x,y,xi,yi,n;
x:=a;
y:=b;
for n from 0 to 30 while evalf(x^2+y^2) < 4 do;
  xi:=evalf(x^2-y^2+a);
  yi:=evalf(2*x*y+b);
  x:=xi;
  y:=yi;
od;
n
end;

>plot3d(0,-2..1,-1.5..1.5,orientation=[-90,0],style=patchnogrid,
scaling=constrained,axes=framed,numpoints=20000,color=couleur);
```

Vous obtiendrez dès lors le résultat ci-dessous:

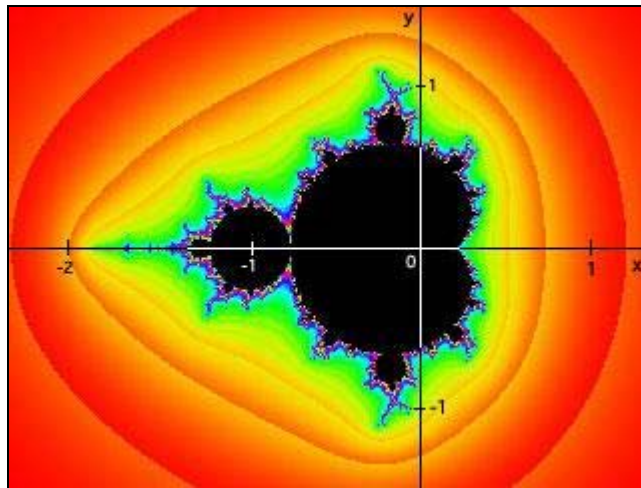
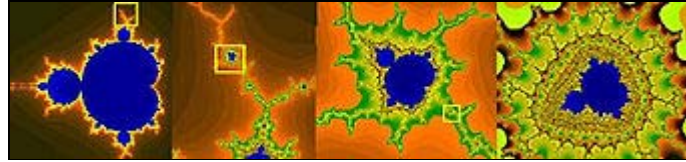


Figure: 58.73 - Ensemble (fractale) de Mandelbrot avec Maple 4.00b

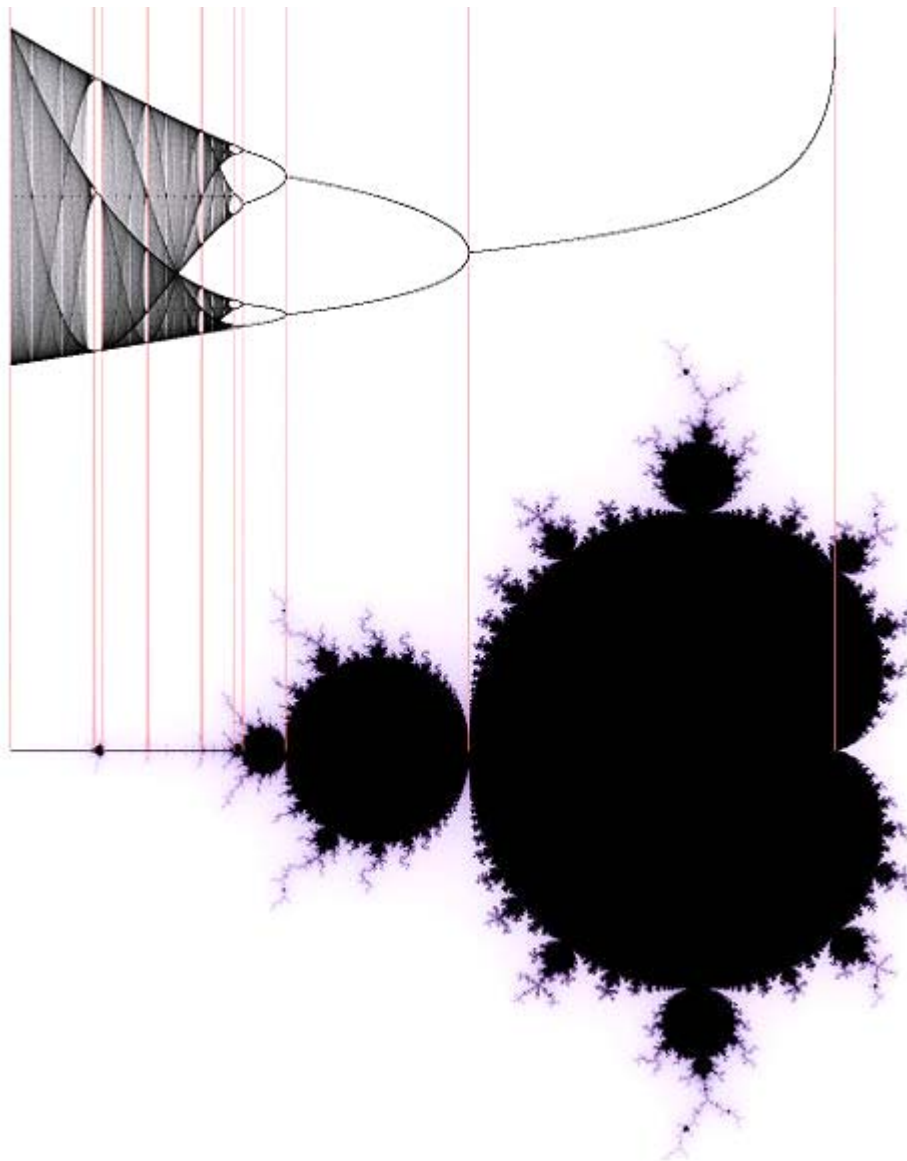
Pour information, le domaine de l'analyse complexe qui étudie des systèmes dynamiques s'intéressant principalement à l'étude d'itérations d'applications holomorphes (cf. chapitre d'Analyse Complexe) se nomme la "**dynamique holomorphe**".

Remarques:

R1. L'ensemble de Mandelbrot est auto-similaire dans le voisinage de points dits "points de Misiurewicz":



R2. Il paraît qu'on peut démontrer (je cherche la démonstration...) que la fractale de Mandelbrot peut être mise en correspondance avec le diagramme de bifurcation que nous avons étudié dans le chapitre de Dynamique des populations tel que:



2.2. ENSEMBLES DE JULIA

L'ensemble de Julia se construit presque de la même façon que l'ensemble de Mandelbrot (puisque l'ensemble de Julia en est en fait un sous-ensemble!). Dans l'ensemble de Mandelbrot, c balaye le plan. Pour l'ensemble de Julia, c est fixé pendant tout le calcul de l'image. À chaque c correspond donc un ensemble particulier que l'on notera $J(c)$ et qui est donc "**l'ensemble de Julia**". Ce qui varie, c'est z_0 , qui prend la valeur du point à tester. C'est donc z_0 qui balaie le plan.

Le point A de coordonnées (x, y) et d'affixe $x + iy$ appartient à $J(c)$ si et seulement si la suite définie par:

$$z_0 = x + iy \text{ et } z_{n+1} = z_n^2 + c \quad (58.97)$$

converge.

En fait, l'ensemble de Mandelbrot est l'ensemble des points c tels que l'ensemble de Julia de paramètre c soit connexe (donc l'ensemble de Mandelbrot généralise tous les ensembles de Julia!!!). Donc la figure de l'ensemble de Mandelbrot contient les figures de tous les ensembles de Julia, ce qui est remarquable (mais aussi logique...!):

Si à nouveau nous développons l'algorithme, nous obtenons à un facteur d'échelle près donné, la fractale représentée ci-dessous (obtenue grâce au petit programme Maple 4.00b modifié utilisé pour la fractale de Mandelbrot):

```
restart; with(plots):
julia:= proc(c,x, y)local z, m;
z:= evalf(x+y*I);
for m from 0 to 30 while abs(z) < 3 do
  z:= z^2 + c
od;
m
end;
```

```
J:= proc(d)
global phonyvar;
phonyvar:= d;
(x, y) -> julia(phonyvar, x, y)
end;
```

```
plot3d(0, -2 .. 2, -1.3 ..1.3, style=patchnogrid,orientation=[-90,0], grid=[270, 270],scaling=constrained,
color=J(-1.25));
```

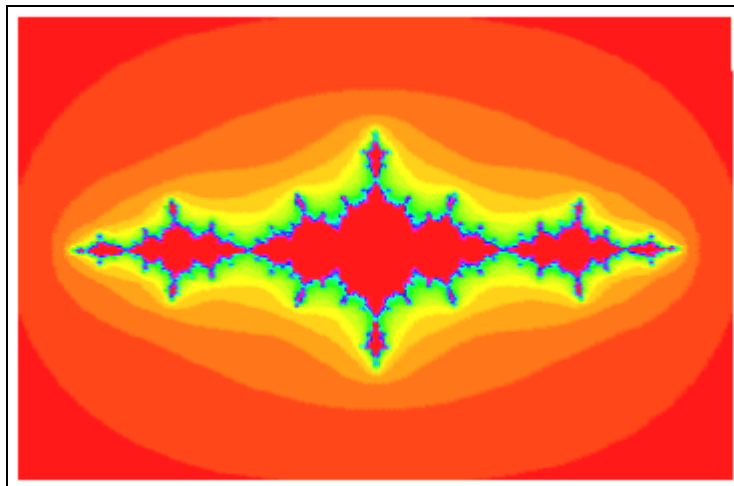


Figure: 58.74 - Ensemble de Julia (fractale) avec Maple 4.00b

et pour montrer que l'ensemble de Mandelbrot contient tous les ensembles de Julia:

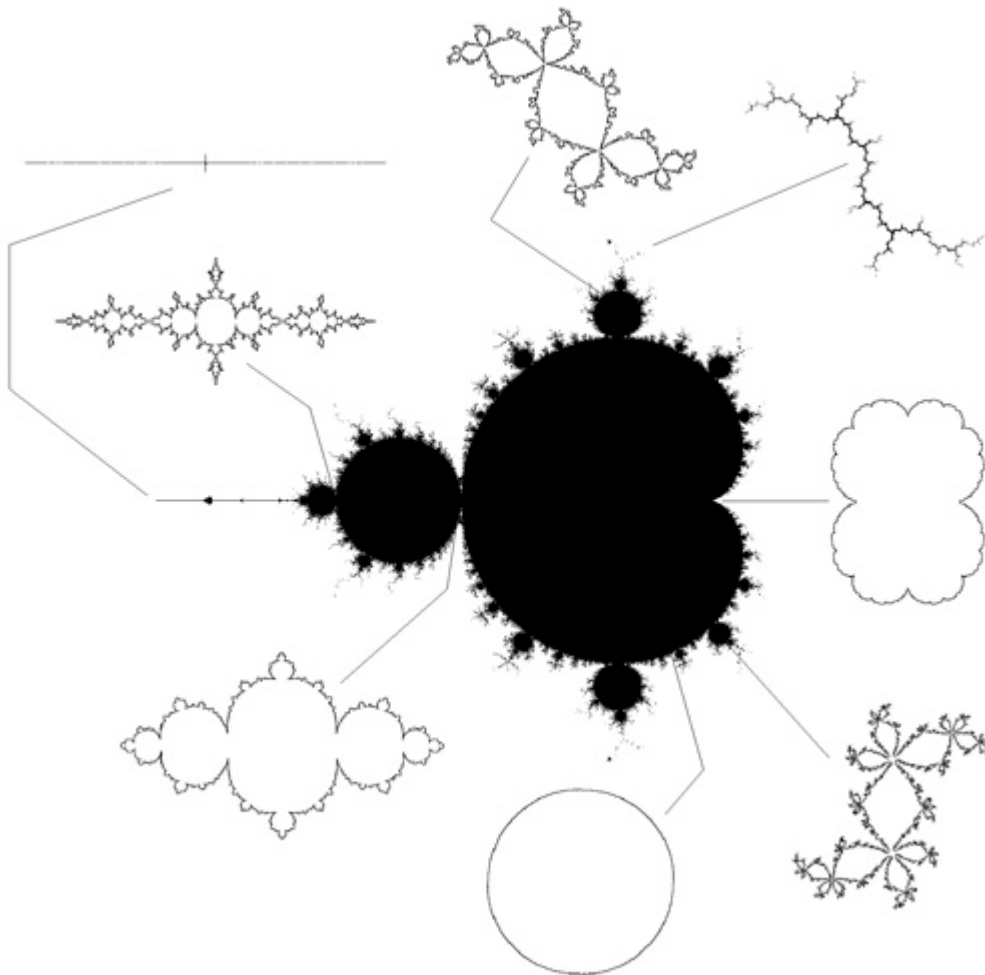
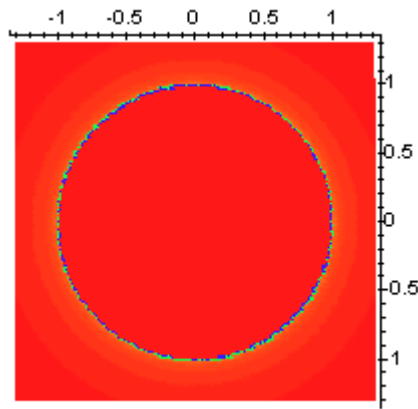
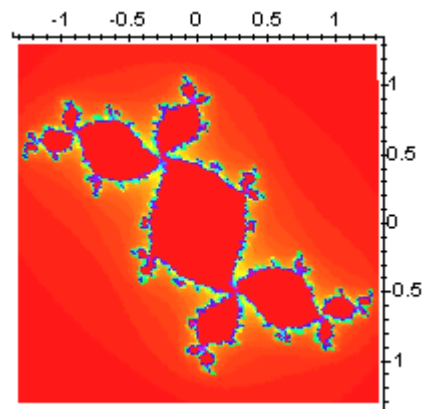


Figure: 58.75 - Illustration de la paternité... de l'ensemble de Mandelbrot

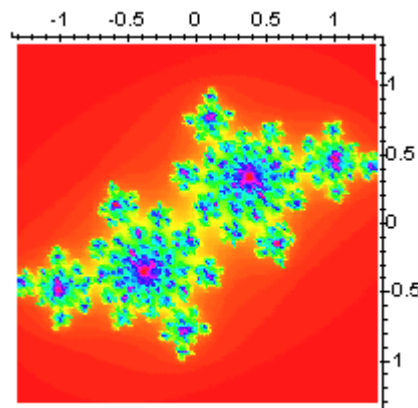
Nous devons donc pouvoir écrire un unique algorithme (voir plus bas) qui permette d'obtenir tous les ensembles de Julia en choisissant simplement bien le point de départ comme le montre les figures ci-dessous:



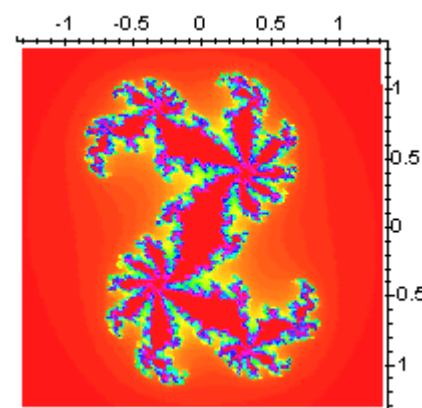
$$z_0 = 0 + 0i$$



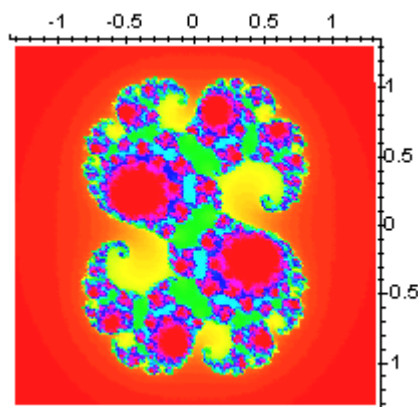
$$z_0 = 0 + 0i$$



$$z_0 = -0.414 - 0.612i$$



$$z_0 = 0.382 + 0.147i$$



$$z_0 = 0.284 - 0.0122i$$

et ainsi de suite...

Obtenu donc avec l'algorithme Maple 4.00b suivant:

```

couleur:=proc(a,b)
local x,y,xi,yi,n;
global reel,imaginaire;
x:=a;
y:=b;
for n from 0 to 100 while evalf(x^2+y^2)<4 do;
xi:=evalf(x^2-y^2+reel);
yi:=evalf(2*x*y+imaginaire);
x:=xi;
y:=yi;
od;
n;
end;

reel:=-0.181;
imaginaire:=-0.667;

plot3d(0,(-13/10)..(13/10),(-13/10)..(13/10),orientation=[-90,0],
style=patchnogrid,scaling=constrained,axes=framed,numpoints=20000,color=couleur);

```

2.3. ENSEMBLES DE NEWTON

Les ensembles de Newton sont ainsi appelés car ils découlent de la résolution du problème de la recherche des zéros d'une fonction par la méthode de Newton.

Soit une fonction f à valeur dans $\mathbb{C}$, et dérivable dans $\mathbb{C}$, on prend z_0 dans $\mathbb{C}$ tel que:

$$z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)} \quad (58.98)$$

Il y a alors deux manières de procéder:

1. Soit nous nous intéressons à $|z_0 - z_i|$ et alors nous faisons comme précédemment
2. Soit nous nous demandons vers quel zéro r_k la suite converge et nous nous intéressons à $|z_i - z_k|$

Si à nouveau nous développons l'algorithme, nous obtenons à un facteur d'échelle près donné, la fractale représentée ci-dessous obtenue à nouveau avec Maple 4.00b:

```

restart;
newton:= proc(x, y)
local z, m;
z:= evalf(x+y*I);
for m from 0 to 50 while abs(z^3-1) >= 0.001 do
z:= z - (z^3-1)/(3*z^2)
od;
m
end;

```

```
plot3d(0, -2 .. 2, -1.5 .. 1.5, orientation=[-90,0],grid=[250, 250], style=patchnogrid,  
scaling=constrained,color=newton);
```

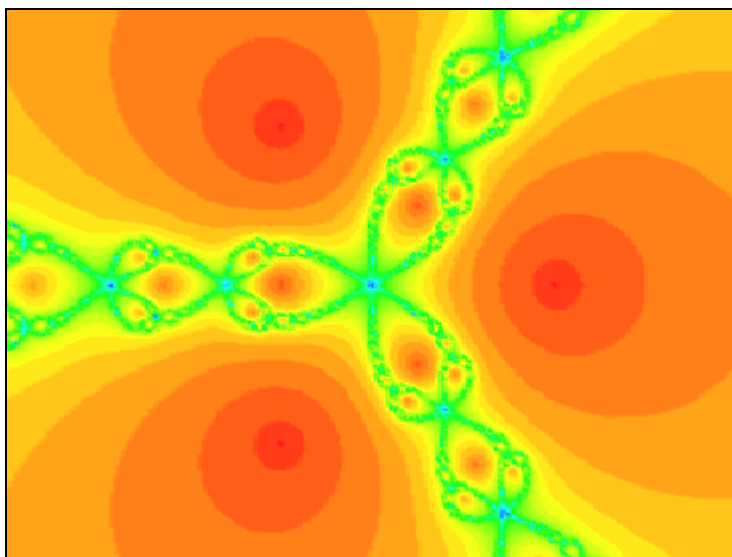
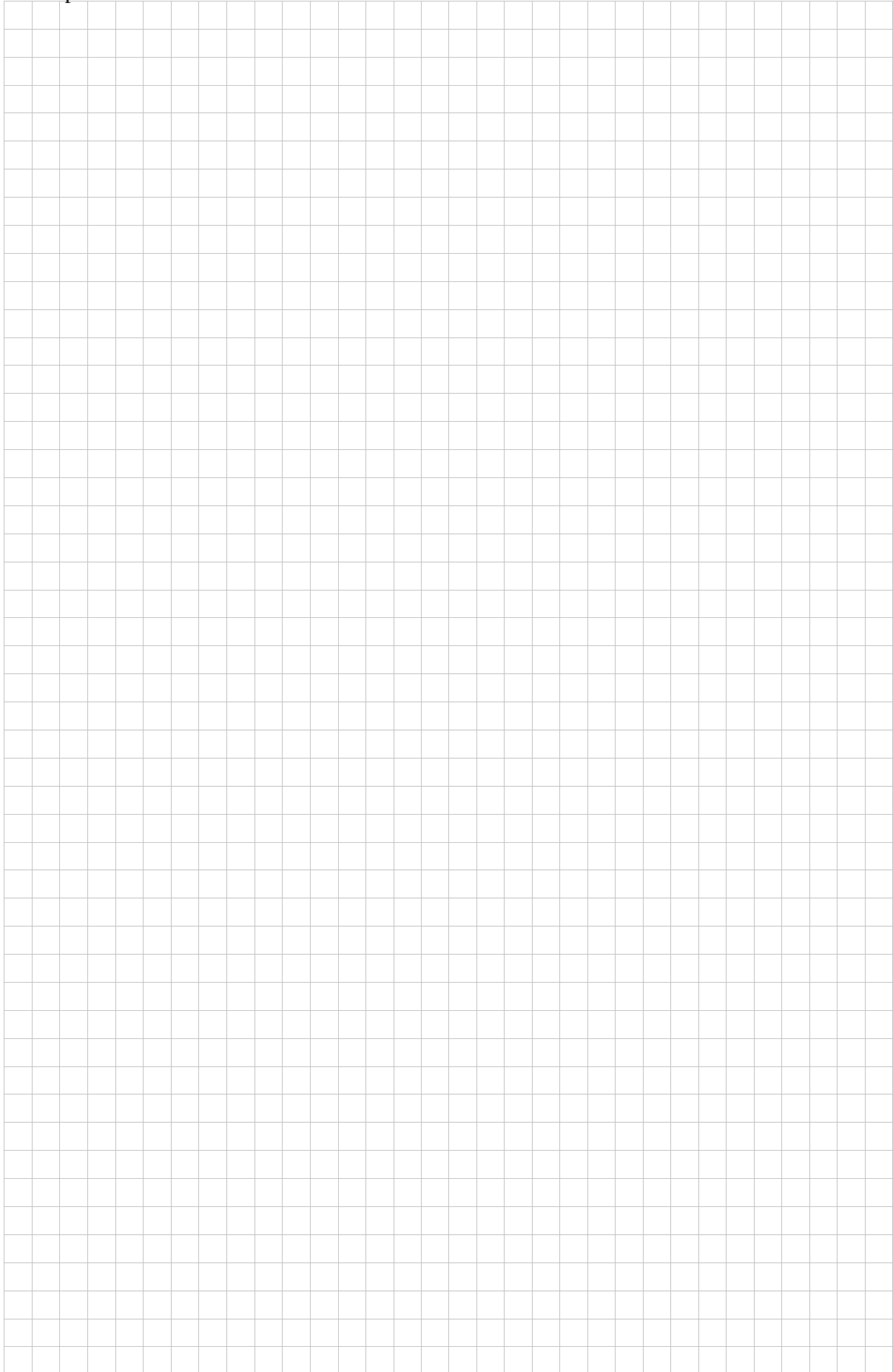
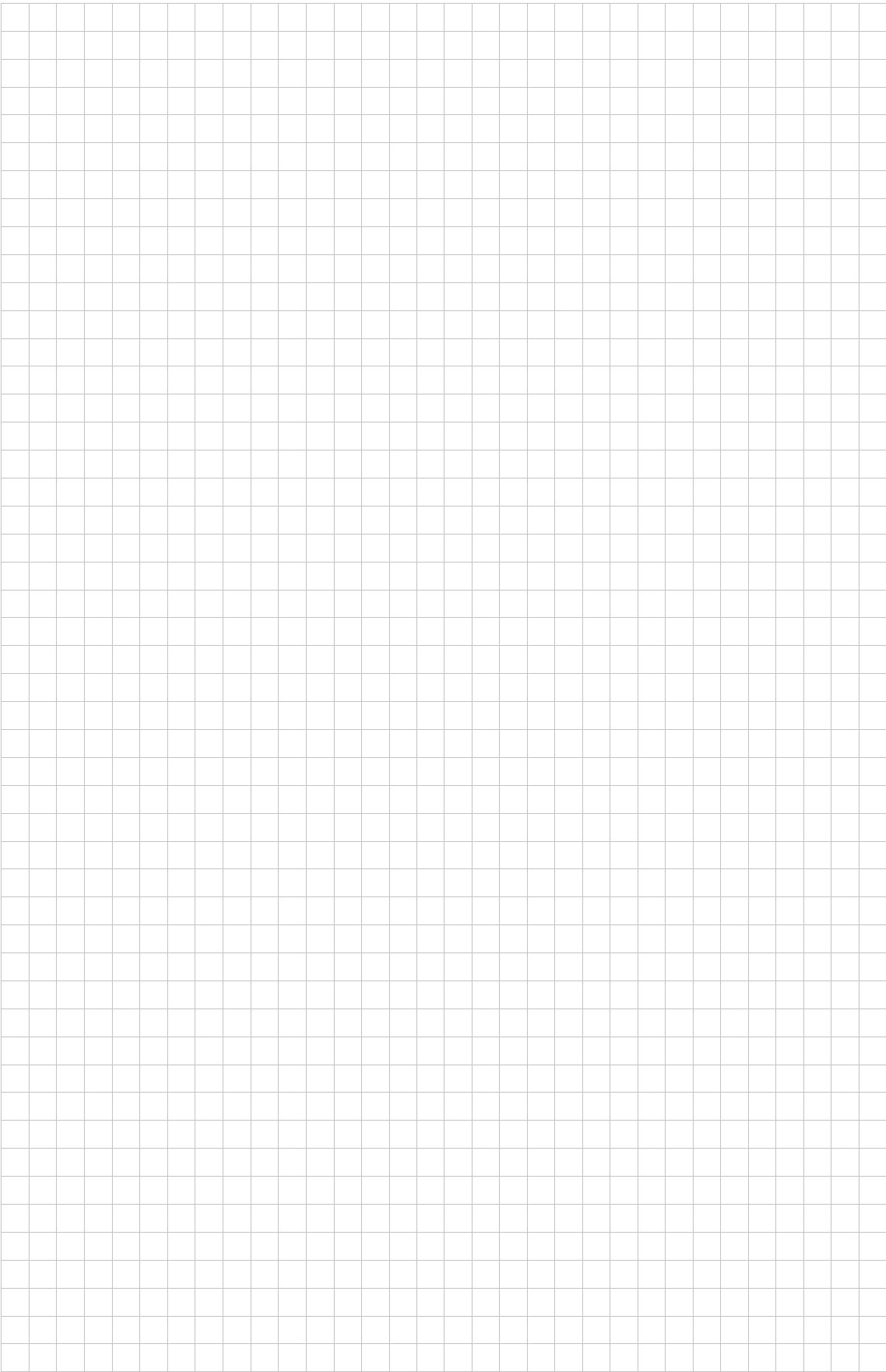


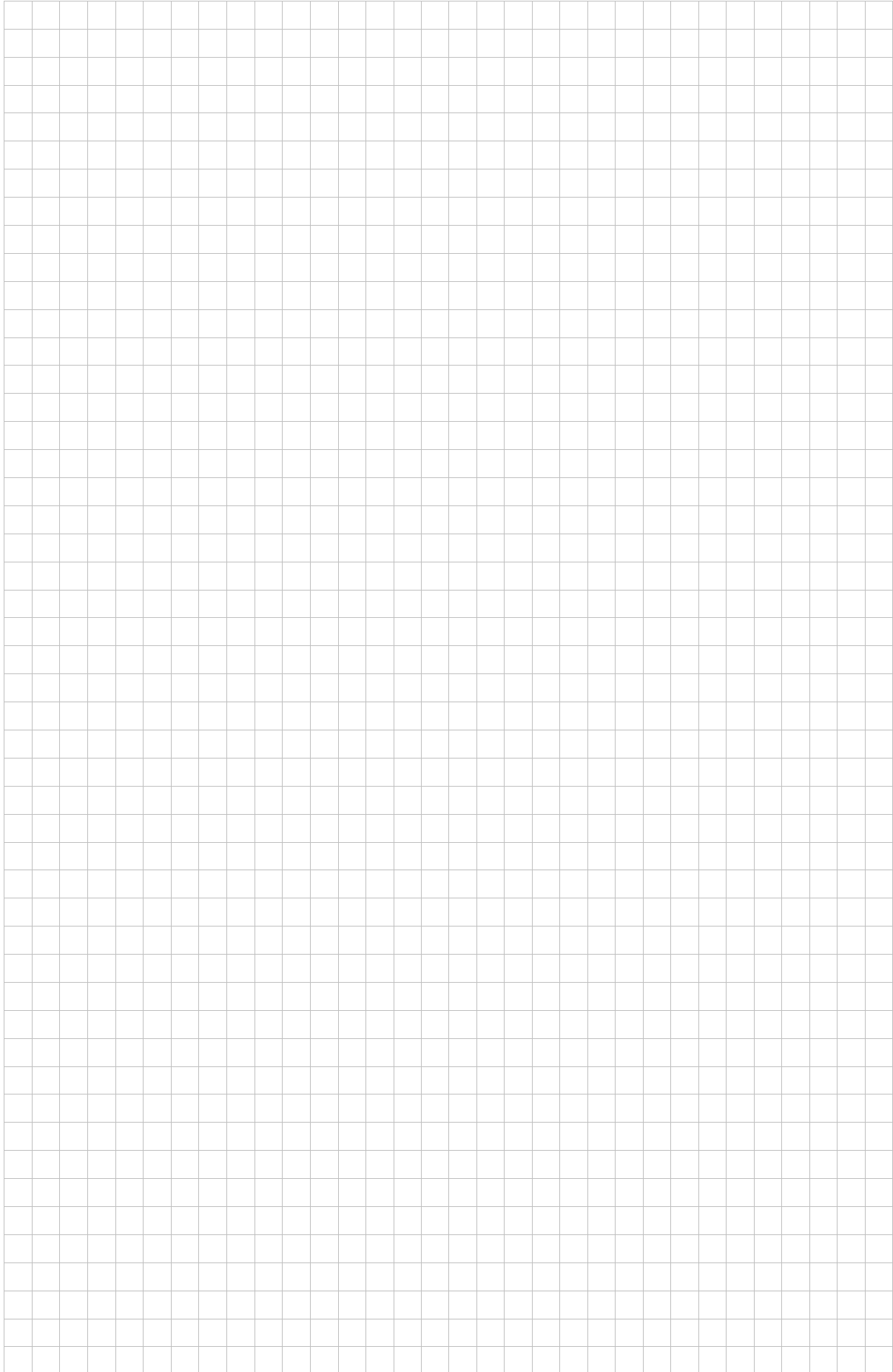
Figure: 58.76 - Ensemble de Newton (fractale) avec Maple 4.00b

Notes personnelles:









59. SYSTÈMES LOGIQUES FORMELS

Le lecteur connaissant bien l'objectif de ce site ne doit pas s'attendre à voir ici de quelconques schémas de boutons poussoirs, interrupteurs, chronogrammes ou encore de schémas électriques de norme MIL ou autres. Nous resterons donc dans un cadre purement formel des systèmes logiques et de leurs outils.

Définitions:

D1. Nous parlons de "**modèle logique asynchrone**" (couramment appelé "modèle logique séquentiel") lorsque les sorties d'un système dépendent de l'ordre chronologique dans lequel se succèdent les entrées.

D2. Nous parlons de "**modèle logique combinatoire**" lorsque les sorties d'un système dépendent uniquement de la combinaison des variables d'entrées.

Remarque: Nous différencions la "**logique stricte**" de la "**logique floue**" qui seront toutes deux définies dans les détails plus loin.

LOGIQUE STRICTE

Considérons dans un premier temps un ensemble que nous noterons B à deux éléments $\{0, 1\}$ (plus formellement notés $\{\perp, \top\}$).

Définitions:

D1. Une "**variable logique stricte**" ou "**variable booléenne**" est un élément de B qui ne possède que deux états 0 et 1 (à l'opposé d'une variable logique floue dont la valeur peut être comprise entre 0 et 1). Elle est représentée par des lettres latines majuscules ou minuscules (au choix).

D2. Une "**fonction logique F** " de plusieurs variables applique $B \times B \times B \dots$ dans B . Elle associe à un n -uplet de variables logiques $(b_0, b_1, \dots, b_{n-1})$ une valeur $F(b_0, b_1, \dots, b_{n-1})$.

D3. Il existe différentes manières d'exprimer une fonction logique (ou "fonction booléenne"). Une fonction de n variables est entièrement décrite par l'énoncé des valeurs de cette fonction pour l'ensemble (ou le sous-ensemble de définition) des combinaisons du n -uplet de variables:

$$F(0, \dots, 0, 0), F(0, \dots, 0, 1), F(0, \dots, 1, 0), \dots, F(1, \dots, 1, 1) \quad (59.1)$$

Cet énoncé prend généralement la forme d'un tableau à $n+1$ colonnes et au plus 2^n lignes, chaque ligne exposant une combinaison des variables et la valeur correspondante de la fonction. Le tableau suivant donne la forme générale d'une "**table de vérité**" de fonctions de trois variables totalement (fonction F) définies:

A	B	C	F(A,B,C)
0	0	0	F(0,0,0)
0	0	1	F(0,0,1)
0	1	0	F(0,1,0)
0	1	1	F(0,1,1)
1	0	0	F(1,0,0)
1	0	1	F(1,0,1)
1	1	0	F(1,1,0)
1	1	1	F(1,1,1)

Tableau: 59.1 - Table de vérité générique

Les éléments d'entrées des systèmes seront considérés comme des variables booléennes sur lesquelles nous pouvons construire une structure de base en anneau que, par ajout d'un axiome particulier, nous pouvons munir d'une "algèbre" (dans le sens calculatoire du terme et non ensembliste!) appelée couramment "**algèbre de Boole**".

ALGÈBRE DE BOOLE

L'algèbre de Boole (ou "anneau de Boole" à un axiome près...) est donc une structure qui est le plus souvent utilisée en électronique (ou micro-électronique). Ainsi, un processeur est composé de transistors permettant de réaliser des fonctions sur des signaux numériques. Ces transistors, assemblés entre eux forment des composants permettant de réaliser des fonctions très simples. À partir de ces composants il est possible de créer des circuits réalisant des opérations assez complexes. L'algèbre de Boole (du nom du mathématicien anglais George Boole 1815 - 1864) est un moyen d'arriver à créer plus ou moins facilement de tels circuits.

L'algèbre de Boole est donc une algèbre sur elle-même (avec une structure d'anneau comme nous allons le définir rigoureusement plus loin) se proposant de traduire des signaux dont la valeur est du type 0/1 (assimilé à: Vrai/Faux) en expressions mathématiques. Pour cela, nous définissons chaque signal élémentaire par des "variables logiques" et leur traitement par des "fonctions logiques". Des méthodes ("tables de vérité") permettent de définir les opérations que nous désirons réaliser, et à transcrire le résultat en une expression algébrique. Grâce à des règles que nous verrons plus loin, ces expressions peuvent être simplifiées. Cela va permettre de représenter grâce à des symboles simples un circuit logique capable d'effectuer des opérations arithmétiques élémentaires, c'est-à-dire un circuit qui schématise l'agencement des composants de base (au niveau logique) sans se préoccuper de la réalisation au moyen de transistors (niveau physique).

Remarque: Il serait préférable avant de commencer la lecture de ce chapitre, de parcourir la partie traitant de la logique au chapitre de la Théorie De La Démonstration et des structures algébriques dans le chapitre de Théorie Des Ensembles.

Il est nécessaire pour obtenir une définition rigoureuse d'une algèbre de Boole de se la donner en des termes d'algèbre abstraite.

Rappel: une "**algèbre de Boole**" est un ensemble $(\mathcal{B}, \vee, \wedge)$ contenant deux éléments particuliers,

$\top, \perp$, (formes abstraites du 0 et du 1) muni de deux lois de composition internes, $\wedge, \vee$ (ET et OU logiques) et qui vérifie les axiomes suivants pour former une structure d'anneau telle que $\forall a, b, c \in B$:

A1. $(a \vee b) \vee c = a \vee (b \vee c)$ et $(a \wedge b) \wedge c = a \wedge (b \wedge c)$ (associativité)

A2. $a \vee b = b \vee a$ et $a \wedge b = b \wedge a$ (commutativité)

A3. $a \wedge (a \vee b) = a$ et $a \vee (a \wedge b) = a$ (absorption)

A4. $(a \vee b) \wedge c = (a \wedge c) \vee (b \wedge c)$ et $(a \wedge b) \vee c = (a \vee c) \wedge (b \vee c)$ (distributivité)

A5. $a \vee a = a$ et $a \wedge a = a$ (idempotence)

A6. a possède un complément noté $\neg a$ ou $\bar{a}$ (NON) tel que: $a \vee \neg a = \top$ et $a \wedge \neg a = \perp$ ("complémentation" ou "inversion")

Remarque: Les quatre premiers axiomes établissent une structure d'anneau. Le cinquième axiome (idempotence) ajouté aux quatre premiers définit le concept "d'algèbre de Boole".

Rigoureusement, pour former une algèbre de Boole il faut un élément symétrique (cf. le chapitre de Théorie Des Ensembles) ce qui n'est pas le cas de l'opération $\vee$. C'est la raison pour laquelle les vrais opérateurs d'une algèbre de Boole sont normalement le $\wedge$ (ET) et la Δ (différence symétrique) donnée par l'opération logique:

$$x \Delta y = (x \wedge \bar{y}) \vee (y \wedge \bar{x}) \quad (59.2)$$

mais pour simplifier, dans les petites classes, il est fréquent que nous y fassions implicitement référence sans entrer dans les détails.

Il s'ensuit que l'ensemble binaire $B = \{0, 1\}$ constitue donc bien par rapport aux lois $\wedge, \vee$ un "groupe abélien". Dès lors, $(B, \vee)$ étant un groupe abélien, la loi $\wedge$ étant associative et distributive par rapport à $\vee$, $(B, \vee, \wedge)$ est un "anneau commutatif unitaire" (vu que B possède un élément neutre pour la loi $\vee$).

Remarques:

R1. Ainsi, les opérations $\wedge$ et $\vee$ admettent chacune un élément neutre tel que le chiffre 1 est l'élément neutre de $\wedge$ et le chiffre 0 l'élément neutre de $\vee$

R2. Les deux opérations que nous utilisons habituellement pour former une algèbre de Boole sont le "ou inclusif" noté rigoureusement $\vee$ mais plus fréquemment par le signe d'addition "+", et le "et inclusif" noté rigoureusement $\wedge$ mais plus fréquemment par le signe de multiplication ".".

Les axiomes précédents peuvent cependant se démontrer à partir des "axiomes de la définition":

$$A1. \top = \neg \perp$$

$$A2. \perp = \neg \top$$

$$A3. \neg \neg a = \bar{\bar{a}} = a \text{ (double complémentation)}$$

$$A4. \top \text{ est l'élément neutre de la loi } \wedge$$

$$A5. \perp \text{ est l'élément neutre de la loi } \vee$$

$$A6. \neg(a \vee b) = \neg a \wedge \neg b \text{ et } \neg(a \wedge b) = \neg a \vee \neg b \text{ (ces deux dernières formulations forment le théorème de De Morgan).}$$

Remarque: Le théorème de De Morgan se démontre à l'aide d'une simple table de vérité ou algébriquement comme nous le verrons juste un peu plus loin.

Il s'ensuit donc le tableau suivant:

$$\begin{array}{l}
 \left. \begin{array}{l} 0 \cdot 0 = 0 \\ 1 + 1 = 1 \end{array} \right\} \text{expression duale} \\
 \left. \begin{array}{l} 1 \cdot 1 = 1 \\ 0 + 0 = 0 \end{array} \right\} \text{expression duale} \\
 \left. \begin{array}{l} 1 \cdot 0 = 0 \\ 0 + 1 = 1 \end{array} \right\} \text{expression duale} \\
 \left. \begin{array}{l} \bar{0} = 1 \\ \bar{1} = 0 \end{array} \right\} \text{expression duale}
 \end{array} \quad (59.3)$$

Nous appelons ces expressions "**duales**" car en remplaçant dans une même équation logique, les 0 par 1, les $\cdot$ par des + est inversement, cette équation reste vérifiée.

Voyons maintenant ce que nous appelons le "**théorème des constantes**" qui consiste à prouver que:

$$\begin{array}{l}
 a \cdot 0 = 0 \\
 a \cdot 1 = a \\
 a + 0 = a \\
 a + 1 = 1
 \end{array} \quad (59.4)$$

La démonstration est triviale (au besoin faire une table de vérité) car elle provient de la propriété même du concept "d'anneau" (de Boole) et de son élément neutre (1) par rapport au $\wedge$ et à son élément neutre (0) par rapport au $\vee$.

Démontrons maintenant la relation suivante:

$$a + \bar{a}b = a + b \quad (59.5)$$

Démonstration:

La distributivité nous amène à écrire:

$$a + \bar{a}b = (a + \bar{a})(a + b) \quad (59.6)$$

et en appliquant la complémentation:

$$(a + \bar{a})(a + b) = 1(a + b) \quad (59.7)$$

en appliquant la commutativité:

$$1(a + b) = (a + b)1 \quad (59.8)$$

et enfin en appliquant le théorème des constantes:

$$(a + b)1 = a + b \quad (59.9)$$

□C.Q.F.D.

Cette démonstration va nous permettre de démontrer le fameux "théorème du consensus":

$$\begin{aligned} a \cdot x + b \cdot \bar{x} + a \cdot b &= a \cdot x + b \cdot \bar{x} \\ (a + x) \cdot (b + \bar{x}) \cdot (a + b) &= (a + x) \cdot (b + \bar{x}) \end{aligned} \quad (59.10)$$

Démonstration:

Pour vérifier le théorème du consensus relatif au produit logique:

$$a \cdot x + b \cdot \bar{x} + a \cdot b = a \cdot x + b \cdot \bar{x} \quad (59.11)$$

nous pouvons faire usage d'un diagramme de Venn où nous voyons bien que le terme $a \cdot b$ est contenu dans les deux autres:

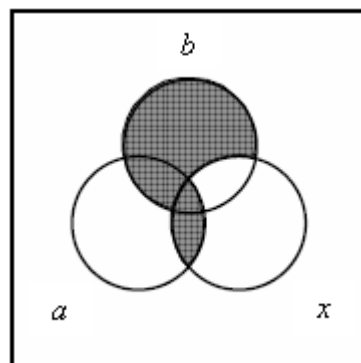


Figure: 59.1 - Diagramme de Venn du théorème du consensus

En procédant de même avec un diagramme de Venn, le lecteur verra sans aucun problème que:

$$(a + x) \cdot (b + \bar{x}) \cdot (a + b) = (a + x) \cdot (b + \bar{x}) \quad (59.12)$$

□C.Q.F.D.

Et enfin les très fameux "**théorèmes de Shannon**" (à ne pas confondre avec le théorème de Shannon en théorie du signal!):

$$\begin{aligned}a + b &= a \cdot b + a \cdot \bar{b} + \bar{a} \cdot b \\a \cdot b &= (a + b) \cdot (a + \bar{b}) \cdot (\bar{a} + b) \quad (59.13)\end{aligned}$$

Démonstrations:

$$\begin{aligned}a + b &= a \cdot b + a \cdot \bar{b} + \bar{a} \cdot b \\&= ab + a \cdot \bar{b} + \bar{a} \cdot b + ab \\&= a \cdot (b + \bar{b}) + b \cdot (\bar{a} + a) \\&= a \cdot 1 + b \cdot 1 \\&= a + b\end{aligned}$$

et:

$$\begin{aligned}a \cdot b &= (a + b) \cdot (a + \bar{b}) \cdot (\bar{a} + b) \\&= (aa + a\bar{b} + ab + b\bar{b}) \cdot (\bar{a} + b) \\&= (a + a \cdot (\bar{b} + b) + 0) \cdot (\bar{a} + b) \quad (59.14) \\&= (a + a \cdot 1 + 0) \cdot (\bar{a} + b) \\&= a \cdot (\bar{a} + b) = a \cdot \bar{a} + a \cdot b = a \cdot b\end{aligned}$$

□C.Q.F.D.

Revenons maintenant sur les théorèmes de De Morgan précédemment présentés comme des axiomes:

$$\overline{(a + b)} = \bar{a} \cdot \bar{b} \quad \overline{(a \cdot b)} = \bar{a} + \bar{b} \quad (59.15)$$

Ces deux relations expriment donc que l'inverse (ou l'opposé) d'un produit (respectivement de la somme) de deux variables est égal à la somme (respectivement au produit) des inverses de ces variables.

Démonstration:

Supposons $\overline{(a + b)} = \bar{a} \cdot \bar{b}$ juste. Alors en vertu des relations $a + \bar{a} = 1$ et $a \cdot \bar{a} = 0$ (axiome de complémentation) nous devons avoir:

$$\begin{aligned}(a + b) + \bar{a} \cdot \bar{b} &= 1 \\(a + b) \cdot \bar{a} \cdot \bar{b} &= 0 \quad (59.16)\end{aligned}$$

Donc il nous faut prouver que ces deux relations sont exactes:

$$\begin{aligned}
 (a+b) + \bar{a} \cdot \bar{b} &= (a+b+\bar{a})(a+b+\bar{b}) = (a+\bar{a}+b)(a+b+\bar{b}) \\
 &= (1+b)(a+1) = 1
 \end{aligned}
 \quad (59.17)$$

et:

$$(a+b) \cdot \bar{a} \cdot \bar{b} = a \cdot \bar{a} \cdot \bar{b} + \bar{a} \cdot b \cdot \bar{b} = 0 \quad (59.18)$$

Le deuxième théorème de De Morgan se démontre de la même façon.

□C.Q.F.D.

Remarque: Ces deux théorèmes peuvent s'étendre à un nombre quelconque de variables.

Corollaires:

$$\begin{aligned}
 a \cdot b &= \overline{\bar{a} + \bar{b}} \\
 a + b &= \overline{\bar{a} \cdot \bar{b}} \\
 \overline{a \cdot b} &= \bar{a} + \bar{b} \\
 \overline{a + b} &= \bar{a} \cdot \bar{b} \\
 \overline{a + b + c + \dots} &= \bar{a} \cdot \bar{b} \cdot \bar{c} \cdot \dots \\
 \overline{a \cdot b \cdot c \cdot \dots} &= \bar{a} + \bar{b} + \bar{c} + \dots
 \end{aligned}
 \quad (59.19)$$

Les expressions logiques, nous l'avons vu à l'aide des axiomes, propriétés et particulièrement des théorèmes précédents, doivent donc toujours pouvoir se mettre sous deux formes (en jouant avec les oppositions $\neg$ aussi donc):

F1. Sous la forme d'une somme de produits logiques, appelée "**forme normale disjonctive F.N.D.**", tel que (exemple):

$$F = \sum \prod (b_i) = \bar{a} \cdot \bar{c} \cdot d + \bar{a} \cdot c \cdot \bar{d} \quad (59.20)$$

Les termes constitutifs de ce polynôme sont les monômes $\bar{a} \cdot \bar{c} \cdot d, \bar{a} \cdot c \cdot \bar{d}$. Les variables ou complémentaires de ces monômes sont les "lettres" $\bar{a}, \bar{c}, d, c, \bar{d}$.

Remarque: Si chacun (tous) des produits contient toutes les variables d'entrée sous une forme directe ou complémentée, alors la forme est appelée "**première forme canonique**" ou "**forme canonique disjonctive**". Chacun des produits est alors appelé "**minterme**".

F2. Sous la forme d'un produit de sommes logiques, appelée "**forme normale conjonctive F.N.C.**", tel que (exemple):

$$F = \prod \sum (b_i) = (\bar{a} + \bar{c} + d) \cdot (\bar{a} + c + \bar{d}) \quad (59.21)$$

Remarque: Si chacune des sommes contient toutes les variables d'entrée sous une forme directe ou complémentée, alors la forme est appelée "**deuxième forme canonique**" ou "**forme canonique conjonctive**". Chacune des sommes est alors appelée "**maxterme**".

Ainsi, une forme normale disjonctive est soit un littéral (une lettre), soit une disjonction de formules écrites comme conjonctions de littéraux. Une forme normale conjonctive est soit un littéral, soit une conjonction de formules écrites comme disjonctions de littéraux.

Les méthodes de simplification que nous verrons par la suite viseront à minimiser le nombre de lettres des expressions de manière à réduire le nombre d'entrées de notre système logique.

Remarque: La simplification algébrique d'une expression consiste à la transformer de manière à réduire au maximum le nombre de ses lettres en lui appliquant les théorèmes vus précédemment.

Pour simplifier les expressions (ou les déterminer) une technique connue consiste donc à utiliser les "**tables de Karnaugh**" que nous verrons plus loin dans les détails.

FONCTIONS LOGIQUES

Donc quand nous parlons d'algèbre de Boole sauf mention contraire, nous faisons référence aux trois opérations booléennes élémentaires (ET, OU, NON) et quelques autres fonctions logiques qui en découlent dont voici les symboles tels que définis en théorie des circuits (norme MIL sauf erreur...):

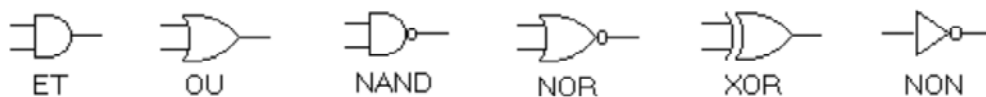


Figure: 59.2 - Portes logiques

et leurs "**tables de vérité**" respectives:

Table de vérité ET ($\wedge$)		
ET	0	1
0	0	0
1	0	1

Tableau: 59.2 - Table de vérité ET

Table de vérité OU ($\vee$)		
OU	0	1
0	0	1
1	1	1

Tableau: 59.3 - Table de vérité OU

Table de vérité <i>NON</i> ($\neg$)	
<i>NON</i>	-
0	1
1	0

Tableau: 59.4 - Table de vérité NON

Toutes les autres "**fonctions logiques**" connues (communes) peuvent être composées de ces deux opérateurs fondamentaux. Tels que par définition (données avec leur définition standard dans la première ligne et avec leurs différentes formes algébriques sous leur table de vérité respective)

$$\neg(a \wedge b) = \overline{(a \cdot b)} = \bar{a} + \bar{b}$$

NON-ET (NAND): NON (<i>a</i> ET <i>b</i>)		
<i>NAND</i>	0	1
0	1	1
1	1	0

Tableau: 59.5 - Table de vérité NON-ET

$$\neg(a \vee b) = \overline{(a + b)} = \bar{a} \cdot \bar{b}$$

NON-OU (NOR): NON (<i>a</i> OU <i>b</i>)		
<i>NOR</i>	0	1
0	1	0
1	0	0

Tableau: 59.6 - Table de vérité NON-OU

$$a \oplus b = (a \vee b) \wedge \neg(a \wedge b) = (a + b) \cdot \overline{(a \cdot b)}$$

$$a \oplus b = a \wedge \neg b + b \wedge \neg a = a \cdot \bar{b} + b \cdot \bar{a}$$

$$a \oplus b = \neg(a \wedge b + \neg a \wedge \neg b) = \overline{a \cdot b + \bar{a} \bar{b}}$$

OU EXCLUSIF (XOR): [<i>a</i> OU <i>b</i>] ET [NON (<i>a</i> ET <i>b</i>)]		
<i>XOR</i>	0	1
0	0	1
1	1	0

Tableau: 59.7 - Table de vérité OU EXCLUSIF

où *a* et *b* sont, vous l'aurez compris, des variables (ou "bit" de Binary Digit) pouvant prendre arbitrairement les valeurs binaires 0 ou 1.

Remarque: La fonction logique XOR est souvent notée dans la littérature par l'opérateur ($\oplus$) et nous considérerons comme évident que le XOR est également une loi de groupe et permet de construire ainsi un groupe commutatif abélien. Cette propriété du XOR est particulièrement utilisée en cryptographie.

TABLES DE KARNAUGH

À part la table de vérité qui simplifie en général la présentation d'un problème logique (mathématique, électronique, micro-électronique, fiabilité des systèmes), il existe d'autres formes tabulées en particulier la table de Karnaugh qui est dans de nombreux cas un outil de travail facile à manipuler.

Considérons pour exemple la fonction:

$$F = ab + \bar{a}b + \bar{a}\bar{b} \quad (59.22)$$

(sous forme F.N.D) et sa table de vérité respective:

b	a	F
0	0	1
0	1	0
1	0	1
1	1	1

Tableau: 59.8 - Exemple de base pour la table de Karnaugh

La table de Karnaugh est définie par une représentation comme celle ci-dessous:

		b	
	$a \backslash b$	0	1
a	0	1 $\bar{a}\bar{b}$	1 $\bar{a}b$
	1	0 $a\bar{b}$	1 ab

Figure: 59.3 - Représentation traditionnelle de la table de Karnaugh

La table de Karnaugh d'une fonction logique comporte donc autant de cases que de combinaisons possibles de variables qui la composent, soit quatre cases pour une fonction à deux variables, et 2^n cases pour une fonction à n variables. Chaque case, qui se trouve à l'intersection d'une ligne et d'une colonne de la table de Karnaugh porte l'état 0 ou 1 que prend la fonction pour le produit logique correspondant des variables (mintermes).

Dans l'exemple précédent nous pouvons voir cependant quelque chose d'intéressant, la fonction F , nous le voyons très bien, peut se simplifier de deux manières:

$$F = b + \bar{a}\bar{b} \quad (59.23)$$

ou encore:

$$F = ab + \bar{a} \quad (59.24)$$

Cette simplification possible ce fait toujours avec deux mintermes adjacents dans la table de Karnaugh tels que:

		b	
$a \backslash b$		0	1
0		1 $\bar{a}\bar{b}$	1 $\bar{a}b$
1		0 $a\bar{b}$	1 ab

Figure: 59.4 - Table de Karnaugh associée

Nous voyons que le premier regroupement/simplification (horizontal) se fait sur la ligne $\bar{a}$ et le second regroupement/simplification (vertical) se fait sur la colonne b tous deux résultats de la simplification algébrique de la fonction.

Donc nous pourrions émettre l'hypothèse que la table de Karnaugh a pour propriété:

P1. De nous donner la forme disjonctive normale d'une fonction.

P2. Que toutes cases adjacentes mises ayant pour valeur 1 peuvent se simplifier en la lettre respective de leur réunion.

C'est donc un outil extrêmement puissant pour simplifier et déterminer des fonctions logiques.

Voyons un exemple à trois variables:

		c			
$a \backslash bc$		00	01	11	10
0		1 $\bar{a}\bar{b}\bar{c}$	1 $\bar{a}\bar{b}c$	1 $\bar{a}bc$	1 $\bar{a}b\bar{c}$
1		0 $a\bar{b}\bar{c}$	0 $a\bar{b}c$	1 abc	1 $ab\bar{c}$

Figure: 59.5 - Table de Karnaugh à trois variables

La F.N.D est donc $F = \bar{a}\bar{b}\bar{c} + \bar{a}\bar{b}c + \bar{a}bc + \bar{a}b\bar{c} + a\bar{b}\bar{c} + a\bar{b}c + abc + ab\bar{c}$ mais elle peut se simplifier algébriquement sous la forme $F = b\bar{c} + bc + \bar{a}b$ mais nous voyons que ceci peut encore se simplifier en $F = b + \bar{a}b$ où nous voyons que les quatre cases adjacentes sont là où b vaut partout 1.

Une autre manière de simplifier:

		c			
$a \backslash bc$		00	01	11	10
0		1 $\bar{a}\bar{b}\bar{c}$	1 $\bar{a}\bar{b}c$	1 $\bar{a}bc$	1 $\bar{a}b\bar{c}$
1		0 $a\bar{b}\bar{c}$	0 $a\bar{b}c$	1 abc	1 $ab\bar{c}$

Figure: 59.6 - Table de Karnaugh à trois variables avec autre méthode de simplification

Donne $F = \bar{a} + ab$.

Une difficulté subsiste cependant avec cette technique: comment choisir la meilleure construction du

tableau (choix des lettres en colonnes ou en ligne) ?

Au fait, il existe une manière spécifique qui consiste à associer la règle de complémentation de l'algèbre de Boole avec ce que nous appelons le "code de Gray".

Définition: Dans le code de Gray, deux termes successifs ne diffèrent que par un seul bit. Les termes ne différant que par un seul bit sont appelés "adjacents".

En utilisant le code de Gray nous pouvons créer des tables de Karnaugh optimales. La raison en est simple, le code de Gray ne change qu'un bit à la fois à chaque incrémentation. En pratique ceci signifie que pour deux valeurs qui se suivent, un et deux par exemple, une des deux variables sera le contraire.

Exemple:

Soit 1=01 correspondant à $\bar{b}a$ et 2 = 11 correspondant à ba , la somme (forme disjonctive) nous donnerait donc $F = a\bar{b} + ab$ ce qui se réduit à l'aide de la règle de complémentation directement à $F = a$.

Tout cela pour dire que quand deux formules se retrouvent côte à côte dans le tableau de Karnaugh, nous conservons des éléments semblables seulement.

Les règles sont telles que nous pouvons réduire quand (voir l'exemple concret précédant):

R1. Deux 1 sont juxtaposés dans le tableau:

	BA	BA	BA	BA
C	00	01	11	10
0	1			
1	1		1	

Figure: 59.7 - Premier type de réduction possible

R2. Quand deux 1 sont aux extrémités du tableau:

	BA	BA	BA	BA
C	00	01	11	10
0	1			1
1	1		1	

Figure: 59.8 - Deuxième type de réduction possible

R3. Quand une rangée pleine fait disparaître les deux variables BA dans ce cas:

	BA	BA	BA	BA
C	00	01	11	10
0	1	1	1	1
1				

Figure: 59.9 - Troisième type de réduction possible

R4. Une colonne pleine fait disparaître deux variables DC dans ce cas:

	BA	BA	BA	BA
DC	00	01	11	10
00		1		1
01		1		
11		1		
10		1		

Figure: 59.10 - Quatrième type de réduction possible

R5. Quatre cases font disparaître deux variables A et C dans ce cas:

	BA	BA	BA	BA
DC	00	01	11	10
00	1	1		
01	1	1		
11				
10				

Figure: 59.11 - Cinquième type de réduction possible

R6. La même case peut servir à deux réductions:

	BA	BA	BA	BA
DC	00	01	11	10
00	1	1	1	
01	1	1	1	
11				
10				

Figure: 59.12 - Septième type de réduction possible

R7. La même case peut servir à deux réductions:

	BA	BA	BA	BA
DC	00	01	11	10
00	1	1	1	
01				
11				
10				

Figure: 59.13 - Huitième type de réduction possible

et sauf erreur... c'est tout mais c'est déjà pas mal.

OPÉRATIONS ARITHMÉTIQUES

À l'aide de tous les éléments démontrés et donnés précédemment, nous sommes maintenant capables de déterminer rigoureusement la fonction logique permettant l'addition et la soustraction booléenne. Rappelons aussi que ceci étant fait, nous pouvons construire la multiplication et la division à l'aide respectivement de l'addition et de la soustraction.

Cependant, nous ne pouvons avec les systèmes numériques formels construire des éléments permettant l'intégration et la différenciation. Pour cela, nous renvoyons le lecteur au chapitre d'Électrocinétique où il est montré comment utiliser des inductances et des condensateurs pour

effectuer de telles opérations avec des signaux.

Remarque: Nous travaillerons sur des nombres entiers mais le lecteur doit se rappeler que les nombres rationnels peuvent toujours être augmentés en puissance pour être représentés de manière entière (reste après à effectuer l'opération inverse au besoin).

La somme de deux bytes sera notée S , la retenue C_s (retenue sortante) et la retenue reportée C_e (retenue entrante).

La table de vérité sera construite avec pour astuce que les entrées du système (a, b, C_e) prennent toutes les valeurs possibles sur 3 bits (trois lettres) soit $2^3 = 8$ lignes que nous avons représentées dans la table suivante:

C_e	a	b
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Tableau: 59.9 - Identifications des retenues reportées pour la somme de bits

et maintenant l'idée consiste à rajouter la colonne constituée par la somme:

$$S = C_e + a + b \quad (59.25)$$

ligne par ligne (sans penser à la retenue sortante C_s , que nous allons voir un tout petit peu plus loin):

C_e	a	b	S
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1

Tableau: 59.10 - Retenues reportées pour la somme de bits

Maintenant, ligne par ligne, nous rajoutons la retenue sortante C_s (qui n'est autre que la valeur qui est envoyée à la retenue entrante de la ligne suivante) de la somme S :

C_e	a	b	S	C_s	mintermes
0	0	0	0	0	$\bar{C}_e \bar{a} \bar{b}$
0	0	1	1	0	$\bar{C}_e \bar{a} b$
0	1	0	1	0	$\bar{C}_e a \bar{b}$
0	1	1	0	1	$\bar{C}_e a b$
1	0	0	1	0	$C_e \bar{a} \bar{b}$
1	0	1	0	1	$C_e \bar{a} b$
1	1	0	0	1	$C_e a \bar{b}$
1	1	1	1	1	$C_e a b$

Tableau: 59.11 - Identification des mintermes de la somme

Il vient alors quatre mintermes (c'est-à-dire les termes pour lesquels S est non nul aux lignes 2,3,5,8) tel que la F.N.D s'écrit:

$$S = \bar{C}_e \bar{a} b + \bar{C}_e a \bar{b} + C_e \bar{a} \bar{b} + C_e a b \quad (59.26)$$

Une simplification possible est:

$$S = C_e (\bar{a} \bar{b} + ab) + \bar{C}_e (\bar{a} b + a \bar{b}) = C_e \cdot \bar{a} + \bar{C}_e a = C_e \oplus a = C_e \oplus a \oplus b \quad (59.27)$$

Il vient également pour la retenue sortante les mintermes suivants:

$$C_s = \bar{C}_e ab + C_e \bar{a} b + C_e a \bar{b} + C_e ab = ab + C_e \cdot (\bar{a} b + a \bar{b}) = ab + C_e \cdot a \oplus b \quad (59.28)$$

Donc finalement nous avons:

$$\begin{aligned} S &= C_e \oplus a \oplus b \\ C_s &= ab + C_e \cdot a \oplus b \end{aligned} \quad (59.29)$$

Remarque: La table de vérité de l'addition sans retenue entrante est appelée "**demi-additionneur**".

La soustraction (différence) de deux bytes sera notée D , l'emprunt e_s (emprunt sortant) et l'emprunt reporté e_e (emprunt entrant). La table de vérité sera construite avec dans un premier temps comme pour l'addition. C'est-à-dire que les entrées du système (a, b, C_e) prennent toutes les valeurs possibles

sur 3 bits (trois lettres) soit $2^3 = 8$ lignes. Ainsi:

e_e	a	b
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Tableau: 59.12 - Identifications des emprunts reportés pour la soustraction de bits

Mais nous allons rajouter une petite subtilité. Plutôt que de nous ennuyer à calculer $D = a - b - e_e$, nous allons calculer $D = a + (-b) + (-e_e)$ de manière à travailler avec la table de vérité ci-dessous:

$-e_e$	a	$-b$
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Tableau: 59.13 - Inversion des emprunts reportés pour la soustraction de bits

et maintenant l'idée consiste à rajouter la colonne de différence $D = a + (-b) + (-e_e)$ ligne par ligne (sans penser à l'emprunt e_e) qui sera strictement identique à la table de vérité de la somme:

e_e	a	b	D
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1

Tableau: 59.14 - Emprunts reportés pour la soustraction de bits

Maintenant, ligne par ligne, nous rajoutons l'emprunt sortant de la différence $D = a + (-b) + (-e_e)$ qui écrite ainsi, devient donc une somme S :

e_e	a	b	$S(D)$	e_s	mintermes
0	0	0	0	0	$\bar{e}_e \bar{a} \bar{b}$
0	0	1	1	1	$\bar{e}_e \bar{a} b$
0	1	0	1	0	$\bar{e}_e a \bar{b}$
0	1	1	0	0	$\bar{e}_e a b$
1	0	0	1	1	$e_e \bar{a} \bar{b}$
1	0	1	0	1	$e_e \bar{a} b$
1	1	0	0	0	$e_e a \bar{b}$
1	1	1	1	1	$e_e a b$

Tableau: 59.15 - Identification des mintermes de la soustraction

Il vient alors quatre mintermes (c'est-à-dire les termes pour lesquels D est non nul aux lignes 2,3,5,8) tel que la F.N.D s'écrit:

$$D = \bar{e}_e \bar{a} b + \bar{e}_e a \bar{b} + e_e \bar{a} \bar{b} + e_e a b \quad (59.30)$$

Une simplification triviale possible est:

$$D = \bar{e}_e (\bar{a} b + a \bar{b}) + e_e (\bar{a} \bar{b} + a b) = e_e \oplus a \oplus b \quad (59.31)$$

Il vient également pour l'emprunt sortant les mintermes suivants:

$$e_s = \bar{e}_e \bar{a} b + e_e \bar{a} \bar{b} + e_e \bar{a} b + e_e a b = \bar{a} b + e_e \cdot (\overline{a \oplus b}) \quad (59.32)$$

Donc finalement:

$$\begin{aligned} D &= e_e \oplus a \oplus b \\ e_s &= \bar{a} b + e_e \cdot (\overline{a \oplus b}) \end{aligned} \quad (59.33)$$

Remarque: La table de vérité de la soustraction sans emprunt entrant est appelée "**demi-soustracteur**".

LOGIQUE FLOUE

La plupart des problèmes rencontrés sont modélisables mathématiquement. Mais ces modèles nécessitent des hypothèses parfois trop restrictives, rendant délicate l'application au monde réel. Les problèmes de ce monde doivent tenir compte d'informations imprécises, incertaines. Prenons l'exemple d'une climatisation: si nous voulons obtenir une température fraîche, nous pouvons nous demander quelle gamme de températures conviendra (la demande est imprécise); de plus la fiabilité des capteurs entre en jeu (la mesure de la température ambiante est incertaine). Nous voyons apparaître la difficulté d'interprétation des variables linguistiques comme frais, chaud, ...ainsi que du traitement de ces données entachées d'incertitude.

Une approche fut développée à partir de 1965 par Loft. A. Zadeh, professeur à l'Université de Californie à Berkeley, basée sur la théorie des sous-ensembles flous ("fuzzy sets" en anglais), généralisant la théorie des ensembles classiques. Dans la nouvelle théorie de Zadeh, un élément peut plus ou moins appartenir à un certain ensemble. Les imprécisions et les incertitudes peuvent ainsi être modélisées, et les raisonnements acquièrent une flexibilité que ne permet pas la logique classique: la "logique floue" était née. De nombreuses applications se sont alors développées dans divers domaines, là où aucun modèle déterministe n'existe ou n'est pratiquement implémentable, ainsi que dans des situations pour lesquelles l'imprécision sur les données rend le contrôle par des méthodes classiques impossible.

Dans ce qui suit, nous développerons d'abord la théorie des sous-ensembles flous, puis nous préciserons le raisonnement en logique floue, nous examinerons les méthodes d'exploitation des résultats obtenus, et enfin nous verrons une application effective.

Avant de passer au côté formel de la chose (mathématiquement parlant) il peut être préférable (puisque'il s'agit quand même d'une technique de l'ingénieur) de présenter brièvement les concepts de la logique floue de manière imagée.

La logique floue est une technique utilisée dans des domaines aussi variés que l'automatisme (freins ABS), la robotique (reconnaissance de formes), la gestion de la circulation routière (feux rouges), le contrôle aérien, l'environnement (météorologie, climatologie, sismologie), la médecine (aide au diagnostic) et bien d'autres.

À l'inverse de la logique booléenne, la logique floue permet à une condition d'être en un autre état que vrai ou faux. Il y a des degrés dans la vérification d'une condition.

Considérons par exemple la vitesse d'un véhicule sur une route nationale. La vitesse normale est de 90 [km/h]. Une vitesse peut être considérée comme élevée au-dessus de 100 [km/h], et comme plus du tout élevée en dessous de 80 [km/h]. La logique booléenne envisagerait les choses de la manière suivante:

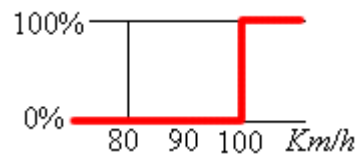


Figure: 59.14 - Application de la logique booléenne

La vitesse est considérée à 100% comme élevée à partir de 100 km/h, et à 0% en dessous.

La logique floue, à l'inverse, permet des degrés de vérification de la condition " La vitesse est-elle élevée? " selon:

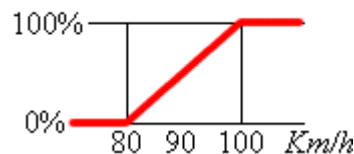


Figure: 59.15 - Aspect flou de la logique... floue

La vitesse est considérée comme pas du tout élevée en dessous de 80 km/h. On peut donc dire qu'en dessous de 80 km/h, la vitesse est élevée à 0%. La vitesse est considérée comme élevée au-dessus de 100 km/h. La vitesse est donc élevée à 100% au-dessus de 100 km/h. La vitesse est donc élevée à 50% à 90 km/h, et à 25% à 85 km/h.

De même, la fonction "La vitesse est-elle peu élevée?" sera évaluée de la manière suivante selon:

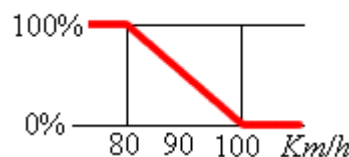


Figure: 59.16 - A l'inverse... le flou demeure...

La vitesse est considérée comme peu élevée en dessous de 80 km/h. Elle est donc peu élevée à 100%. La vitesse est considérée comme pas du tout peu élevée au-dessus de 100 km/h. Elle est donc peu élevée à 0%. La vitesse est donc peu élevée à 50% à 90 km/h, et à 75% à 85 km/h.

Nous pouvons également définir une fonction "La vitesse est-elle moyenne?" selon:

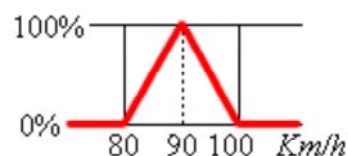


Figure: 59.17 - Application pertinente de la logique floue

La vitesse est moyenne à 90 km/h. À cette allure, la vitesse est moyenne à 100%. La vitesse n'est pas du tout moyenne en dessous de 80 km/h et au-dessus de 100 km/h. Hors de cet intervalle, la vitesse est moyenne à 0%. La vitesse est donc moyenne à 50% à 85 km/h et 95 km/h.

Il n'est pas obligatoire que la transition soit linéaire. Des transitions hyperboliques (comme une sigmoïde ou une tangente hyperbolique), exponentielles, gaussiennes (dans le cas d'un état moyen) ou de toutes autres nature, sont utilisables telles que les méthodes que nous avons vues lors de notre étude des réseaux de neurones dans le chapitre de Méthodes Numériques:

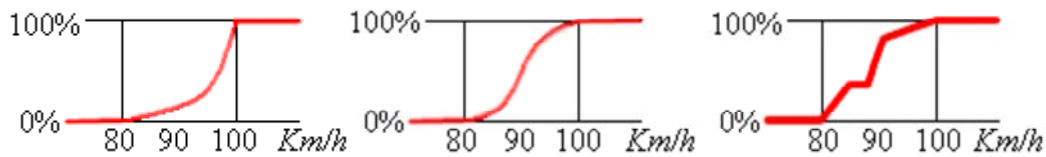


Figure: 59.18 - Exemples de transitions non linéaires

Une fois évaluée la valeur de l'entrée ("La vitesse est-elle élevée?"), une valeur peut être déterminée pour une fonction de sortie. Considérons la fonction " Si la fièvre est forte, alors administrer de l'aspirine ". Une telle fonction est appelée "commande floue". Elle est composée de deux parties:

1. Une entrée: "La fièvre est-elle forte?". Nous considérons qu'une fièvre n'est pas forte en dessous de 38°C, et qu'elle est forte au-dessus de 40°C.

2. Une sortie: "Administrer de l'aspirine"

Ces deux parties sont liées. Nous pouvons les représenter ensemble comme ci-dessous:

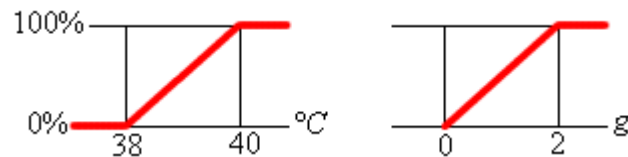


Figure: 59.19 - Entrée/Sortie floue

Il existe plusieurs techniques pour déterminer la valeur de la sortie (dans l'exemple: la quantité d'aspirine à administrer):

Un exemple consiste à prendre l'horizontale passant par le point d'ordonnée correspondant sur la courbe de départ à l'abscisse de la valeur de l'entrée et de regarder où cette horizontale coupe la courbe de sortie. L'abscisse de ce point d'intersection est une valeur de sortie possible comme représenté ci-dessous:

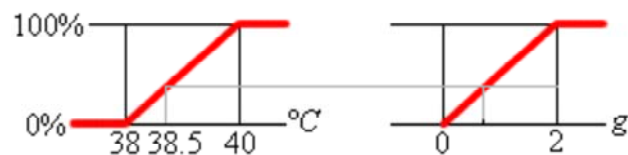


Figure: 59.20 - Transposition entrée/sortie

Un deuxième exemple consiste à prendre comme valeur de sortie possible le centre de gravité du trapèze en grisé délimité par l'horizontale et la courbe de sortie comme représenté sur la figure ci-dessous:

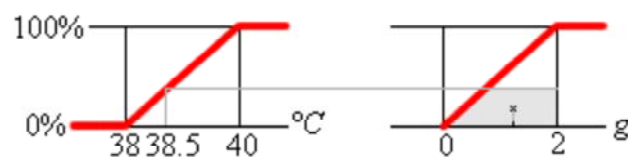


Figure: 59.21 - Autre transposition possible

De par ces deux exemples, nous voyons bien que nous sommes à la frontière science/ingénierie puisqu'il y a un choix technique ou/et statistique à faire dans la méthode à choisir.

ENSEMBLE FLOU

Définition: Soit X un ensemble. Un "sous-ensemble flou" A de X est défini par une fonction d'appartenance f_A sur X à valeurs dans l'intervalle $[0,1]$.

Remarque: La fonction d'appartenance peut être fixée arbitrairement. Un problème des applications pratiques est pour l'ingénieur de définir ces fonctions (nous faisons généralement appel à des données statistiques ou à l'avis d'un expert).

La notion de sous-ensemble flou englobe celle de sous-ensemble classique pour laquelle f_A est la fonction indicatrice:

Définition: Si A et B sont deux ensembles, tels que A est inclus dans B , nous appelons "fonction indicatrice" de A (relativement à B), la fonction 1_A définie dans $\{0,1\}$, et telle que:

$$\begin{aligned} 1_A(x) &= 1 \text{ si } x \text{ est dans } A \\ 1_A(x) &= 0 \text{ si } x \text{ n'est pas dans } A \end{aligned} \quad (59.34)$$

Les fonctions indicatrices sont souvent des intermédiaires techniques très pratiques!

 Exemple:

Une fonction caractéristique possible pour définir le sous-ensemble A flou "avoir une vingtaine d'années" sur l'ensemble X des réels:

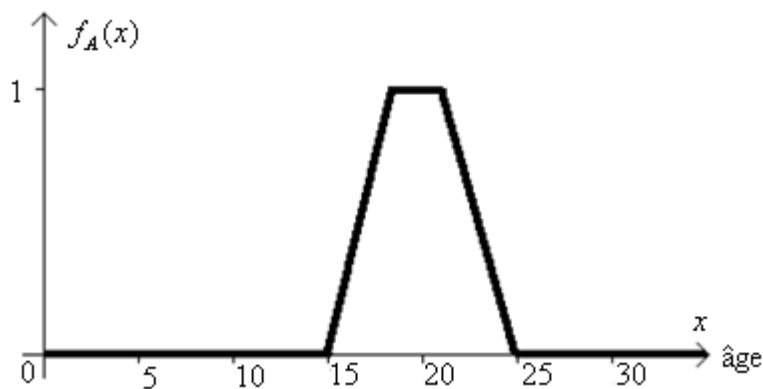


Figure: 59.22 - Fonction floue centrée

Les notions suivantes sont caractéristiques de A :

Définitions:

D1. Support de A : $\text{supp}(A) = \{x \in X, f_A(x) \neq 0\}$

D2. Hauteur de A : $h(A) = \sup_{x \in X} f_A(x)$

D3. A est dit normalisée si $h(A) = 1$

Remarque: Les sous-ensembles flous considérés seront tous supposés normalisés, in extenso de hauteur égale à 1.

D4. Noyau de A : $\text{noy}(A) = \{x \in X, f_A(x) = 1\}$

D5. Cardinalité de A : $|A| = \sum_{x \in X} f_A(x)$

Exemple:

Avec l'exemple de la figure précédente:

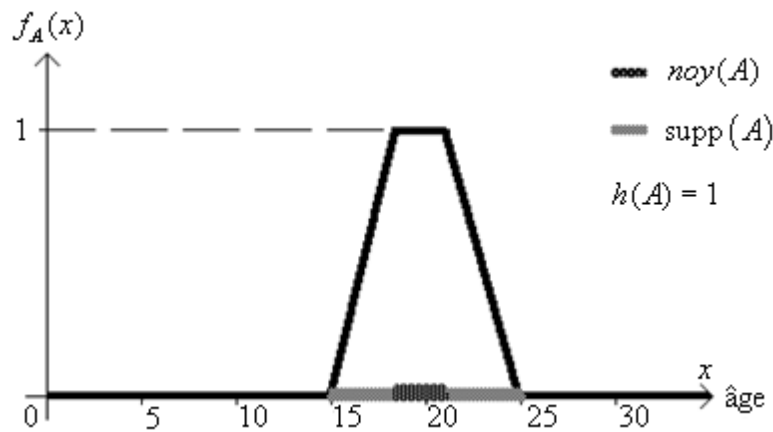


Figure: 59.23 - Décomposition de la fonction floue centrée

D6. Si A et B sont deux sous-ensembles flous de l'ensemble X , nous disons que:

1. A est "plus spécifique" que B si:

$$\text{noy}(A) \subsetneq \text{noy}(B) \text{ et } \text{supp}(A) \subseteq \text{supp}(B) \quad (59.35)$$

2. A est "plus précis" que B si:

$$\text{noy}(A) = \text{noy}(B), \text{ et } \text{supp}(A) \subsetneq \text{supp}(B) \quad (59.36)$$

D7. Il y a égalité entre deux sous-ensembles flous si et seulement si:

$$\forall x \in X, f_A(x) = f_B(x) \quad (59.37)$$

D8. Il y a inclusion si et seulement si:

$$\forall x \in X, f_A(x) \leq f_B(x) \quad (59.38)$$

D9. L'intersection $A \cap B$ est définie par:

$$f_{A \cap B} : x \rightarrow \min(f_A(x), f_B(x)) \quad (59.39)$$

D10. L'union $A \cup B$ est définie par:

$$f_{A \cup B} : x \rightarrow \max(f_A(x), f_B(x)) \quad (59.40)$$

Exemple:

Reprenons le cas déjà envisagé. Nous considérons les personnes ayant une "vingtaine d'années", et celles "ayant la majorité" (c'est-à-dire celles qui sont majeures) (en pointillés sur la figure: nous considérons celle-ci comme un sous ensemble non-flou!):

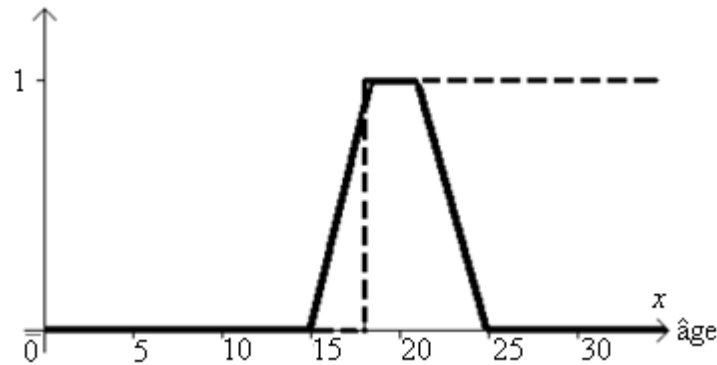


Figure: 59.24 - Décomposition de la fonction floue centrée

Selon les définitions de l'intersection ("ET logique" ou multiplication logique selon l'algèbre de Boole) et de l'union ("OU logique" ou addition logique selon l'algèbre de Boole), nous pouvons caractériser les sous-ensembles flous correspondant aux personnes "ayant une vingtaine d'années et la majorité" (figure de gauche ci-dessous) ainsi que celui des personnes "ayant une vingtaine d'années ou la majorité" (figure de droite ci-dessous):

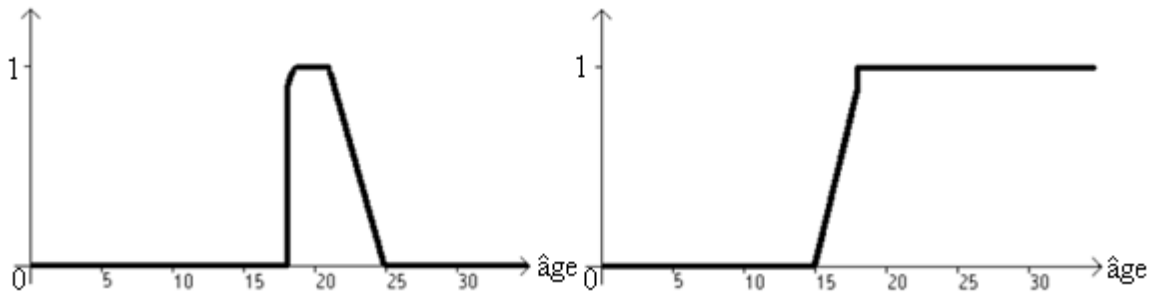
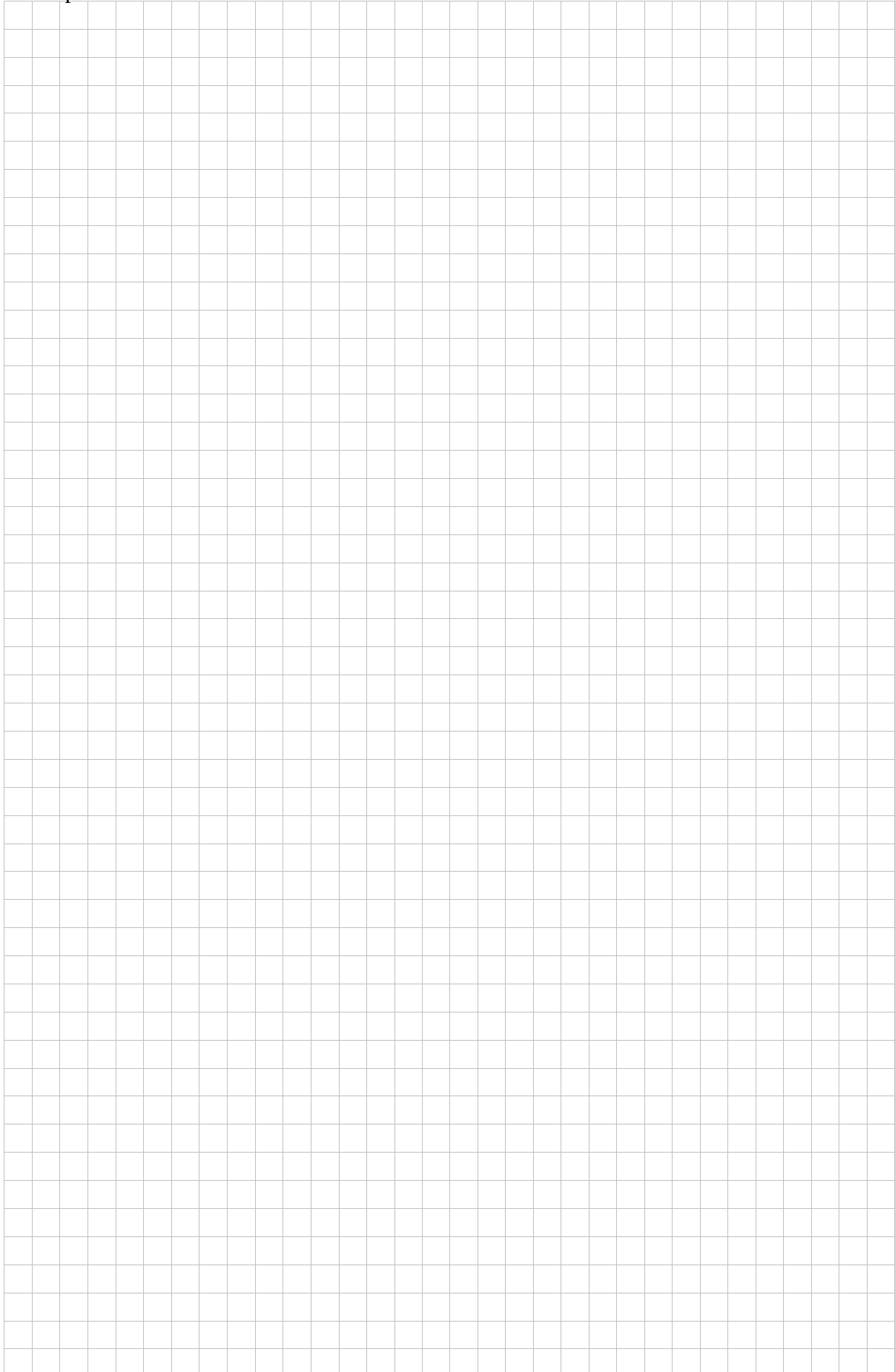


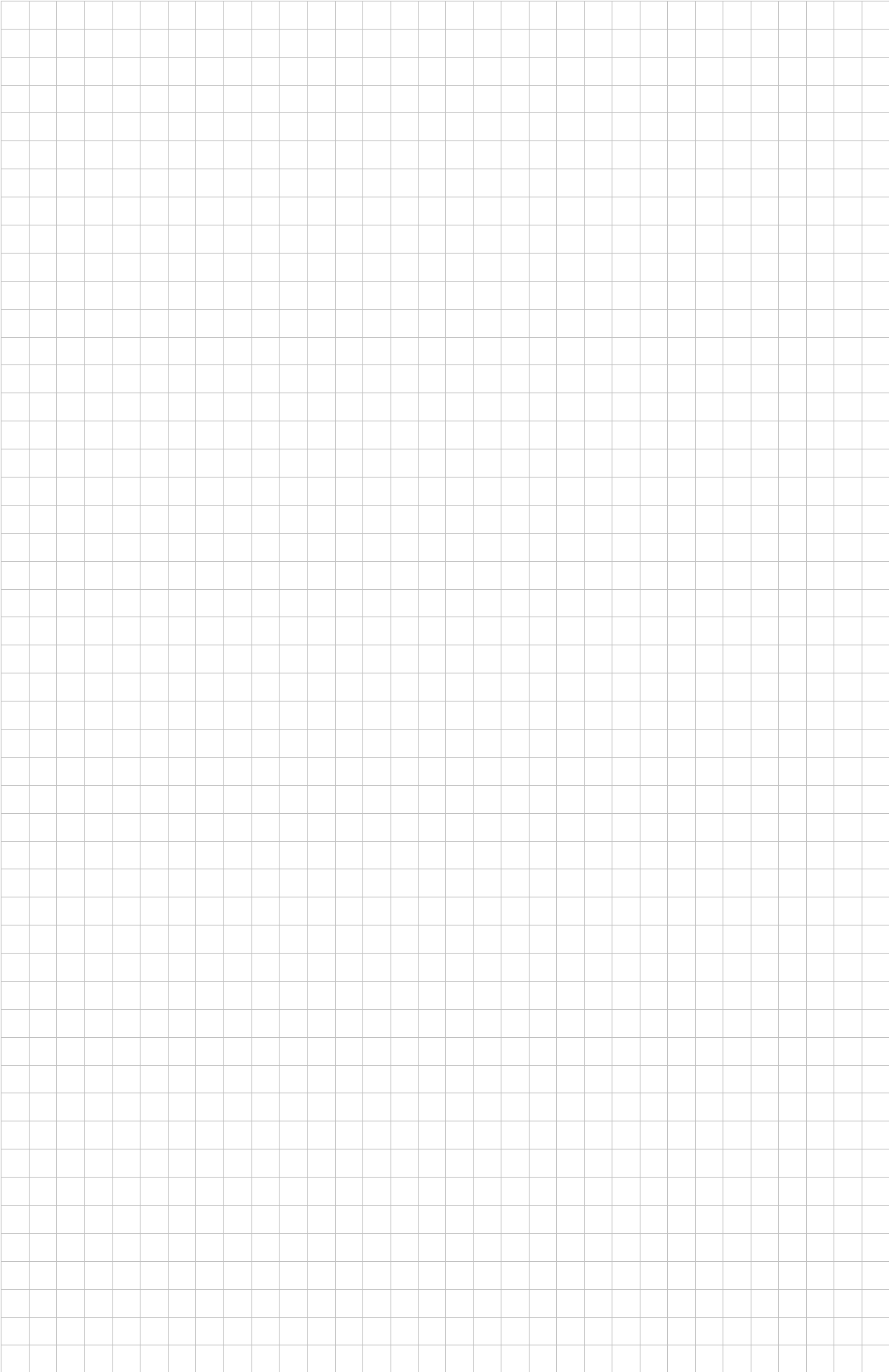
Figure: 59.25 - Résumé des résultats des différentes opérations logiques

Notes personnelles:









60. CODES CORRECTEURS

Si la première moitié du vingtième siècle a été celle de la révolution analogique par la radio et la télévision, la seconde moitié de ce siècle est celle de la révolution numérique et de l'utilisation systématique de l'algèbre dans la transmission de données.

Les codes correcteurs d'erreurs sont utilisés pour ajouter de la redondance aux données pour les rendre tolérantes à la transmission (jusqu'à un certain degré). Basiquement, l'idée est d'encoder d'une manière ou d'une autre une séquence d'information et d'ajouter de l'encodage aux données d'origine en guise de contrôle de l'intégrité des données. Ainsi, même si une certaine partie de l'information est corrompue, mais pas trop,... la redondance permettra d'identifier les parties erronées du message.

Ainsi, après l'apparition du CD audio dans les années 1980, il faut compter sur le développement de la diffusion par satellite et de nouveaux moyens de communication tels la télécopie, le réseau Internet ou le téléphone numérique qui utilisent tous les codes correcteurs d'erreurs (C.C.D). Même la photographie et la radio deviennent par ailleurs numériques.

Les techniques de restitution d'images ou de sons sont liées à la transmission et à la lecture correcte de nombreux messages numériques, encore appelés "**mots**". Un message est formé de mots, eux-mêmes constitués de symboles (un exemple particulier étant le "bit" qui est pour rappel, la contraction du terme "binary digit") pris dans un alphabet. Si l'alphabet est binaire alors chaque symbole sera donc un bit.

Prenons le message 00101 formé de 5 bits valant chacun 0 ou 1. Si nous transmettons le message tel quel, une erreur de transmission ou de lecture peut avoir lieu et rendre le message inintelligible. Décidons de répéter ce message trois fois et d'envoyer:

001010010100101 (60.1)

Si le message reçu comporte une erreur, cette erreur peut être corrigée. S'il comporte deux erreurs, le récepteur est capable de détecter qu'il y a eu erreur mais ne peut pas toujours récupérer le message originel. Enfin, s'il se produit plus de deux erreurs pendant la transmission, le récepteur peut ne pas les détecter.

Nous venons ainsi de voir un premier exemple de C.C.D., appelé "**code à répétition**". Ce code, qui corrige une erreur et en détecte deux, est utilisé dans certains lecteurs de CD Audio possédant trois têtes de lecteur. Le signal 0 ou 1 est lu indépendamment par chacune de ces trois têtes pour donner un mot de trois chiffres, et une erreur de lecture peut être corrigée.

Remarquons qu'il est naturel d'allonger un message pour le protéger. Prenons les mots d'un langage. Ils sont en général très éloignés les uns des autres, deux mots différant selon leurs longueurs et selon les lettres et les syllabes utilisées. Ainsi, nous confondrons difficilement les mots "bibliothèque" et "armoire" même si ces mots sont mal prononcés ou entendus, et nous reconstituerons naturellement le message dans une conversation quand bien même certains sons seraient supprimés ou déformés. Les militaires quant à eux épellent leurs numéros d'immatriculation en disant "alpha zoulou" pour "AZ"...

Un deuxième exemple de détection d'erreurs largement utilisé en informatique est l'adjonction d'un "**bit de parité**". Reprenons le message 00101 et ajoutons-lui un dernier bit obtenu en additionnant les cinq

bits du départ modulo 2. Le message devient 001010 et permet de détecter une erreur sans toutefois pouvoir la corriger. Pour cela, nous faisons la somme de tous les bits pour obtenir 0 s'il n'y a pas eu d'erreur, et 1 dans le cas contraire. Ce code appelé "**code de parité**" est utilisé un peu partout: dans les numéros de sécurité sociale où l'on rajoute la clé, dans ceux des comptes bancaires ou encore dans les code-barres des supermarchés où c'est le 13ème chiffre qui constitue la clé de contrôle (Voyager II est l'un des nombreux utilisateurs des codes de parité pour communiquer de manière fiable ainsi que le 8ème bit de l'ASCII qui est utilisé comme bit de parité).

Pendant de nombreuses années, les boîtiers de DRAM géraient des mots d'un bit ; il fallait donc placer sur la carte mémoire 8 boîtiers pour travailler avec des octets (8 bits). Pourtant, de nombreuses cartes comportaient non pas 8, mais 9 boîtiers ! Le neuvième boîtier était destiné à stocker un bit de parité lors de chaque mise en mémoire d'un octet. Lors de la lecture d'un octet, on vérifiait si, entre le moment de l'écriture et celui de la lecture, la parité n'avait pas été modifiée (suite à un parasite, par exemple).



Figure: 60.1 - Exemple de boîtier de DRAM

Enfin, voyons un troisième exemple utilisé sur certains serveurs informatiques qui utilisent des disques parallèles en RAID4 ou RAID6, ce dernier utilisant les codes de Hamming que nous verrons plus loin.

Supposons que nous ayons trois disques durs de données, et que le contenu du premier octet de chaque disque soit le suivant:

$D1: 11110000$
 $D2: 10101010$ (60.2)
 $D3: 00110011$

Alors il suffit de prendre chaque colonne et de compter le nombre p de 1 dans la colonne. La valeur du bit de parité est alors p modulo 2. Nous avons pour la première colonne p qui vaut 2. Donc le bit de parité vaut 0, etc. Nous avons alors sur le disque de contrôle DC :

$DC: 01101001$ (60.3)

Dès lors, dès qu'un des trois disques tombe en panne il devient aisé de le restaurer grâce au disque de contrôle en faisant la procédure inverse.

Ces trois exemples fondamentaux sont à la base de la théorie du codage et montrent que nous pouvons maîtriser l'apparition d'erreur en allongeant volontairement le message avant sa transmission ou sa lecture. Des techniques algébriques plus sophistiquées sont ensuite utilisées pour améliorer les performances du codage, le but étant de:

- savoir si des erreurs se sont produites (problème de la détection);
- retrouver le message correct à partir du message reçu (problème de la correction);

- corriger le plus d'erreurs possible tout en utilisant le moins de bits supplémentaires possibles (problème de la performance du codage).

Du point de vue mathématique, l'un des intérêts de la théorie des codes est de montrer que l'algèbre s'applique fondamentalement dans notre vie de tous les jours dès que nous écoutons de la musique ou que nous nous installons devant notre téléviseur, et que des notions aussi abstraites que celles d'espaces vectoriels ou de polynômes sur des corps finis nous permettent de lire des messages, d'écouter de la musique ou de regarder des films dans des conditions optimales.

Nous distinguons les deux classes suivantes de C.C.D.: les codes en blocs et les codes en treillis. La figure ci-dessous donne un simple aperçu de la grande famille de codage. Dans la première classe (à droite sur la figure), citons les codes les plus célèbres comme les codes BCH, Reed-Solomon et Goppa, Golay et Hamming. La deuxième classe (à gauche sur la figure) est moins riche en variété mais présente beaucoup plus de souplesse surtout dans le choix des paramètres et des algorithmes de décodage disponibles. Citons par exemple, les codes convolutifs binaires systématiques récurrents très utilisés dans les modulations codées (TCM) et les codes concaténés parallèles (Turbo Codes).

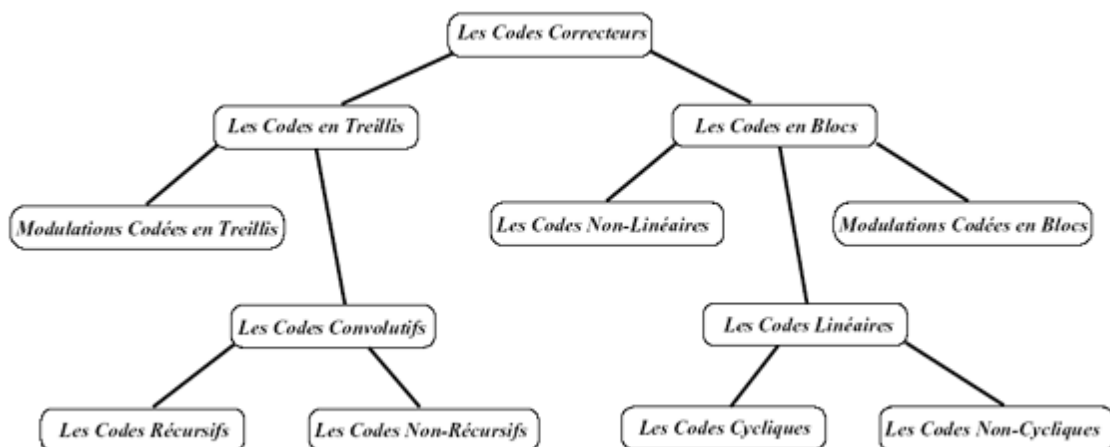


Figure: 60.2 - Organigramme des codes correcteurs

Remarque: Pour aborder les fondements de la théorie des codes correcteurs, nous conseillons au lecteur d'avoir parcouru au préalable les chapitres de Mécanique Statistique (où se trouve la théorie de l'information), des systèmes numériques formels, et de topologie.

1. CHECKSUM

Avant de commencer la partie de mathématiques pures, nous souhaiterions faire une petite introduction sur le "checksum" (somme de contrôle) qui est un outil très fréquemment utilisé dans les entreprises lors de l'échange de fichiers de plus de quelques GB entre deux ordinateurs ou encore lors du téléchargement sur Internet.

La somme de contrôle, parfois appelée aussi "empreinte", est un concept de base de la théorie des codes utilisé pour les codes correcteurs. Elle correspond à un cas particulier de contrôle par redondance. Elle est largement utilisée en informatique et en télécommunications numériques.

Les codes utilisant les sommes de contrôle permettent de valider un message. Si le nombre d'altérations durant la transmission est suffisamment petit, alors elles sont détectées. L'utilisation d'une unique somme de contrôle permet la détection mais non la correction des erreurs.

La technique de base consiste à prendre la somme de certaines longueurs de bits (octets, word ou autres...) et de calculer le modulo 255 (FF). Par exemple, si nous prenons deux mots et nous basons sur leur code ASCII hexadécimal (vous pouvez trouver les tables ASCII un peu partout sur Internet):

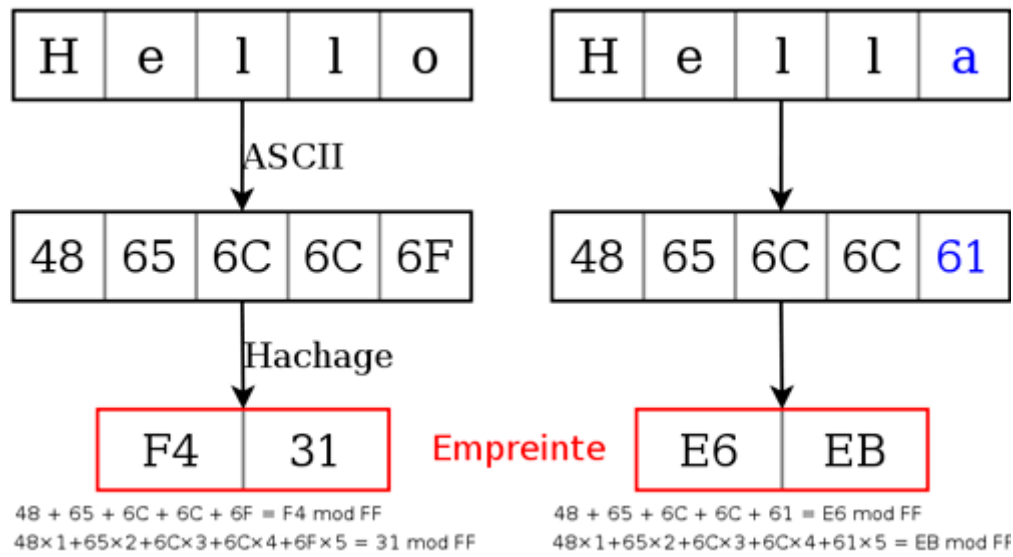


Figure: 60.3 - Principe du Checksum (source: Wikipédia)

Certains utilisent aussi le MD5 (Message Digest 5) pour avoir une empreinte d'un message (cf. chapitre de Cryptographie).

2. ENCODEURS

Soit Q un ensemble fini à q éléments (bits, alphabets). Soient k et n deux entiers naturels non nuls avec $k \leq n$. L'ensemble des messages sera une partie E de Q^k , et nous introduisons une application bijective (du moins c'est le but):

$$\begin{aligned} f: E &\rightarrow Q^n \\ a = (a_1, a_2, \dots, a_k) &\rightarrow c = (c_1, c_2, \dots, c_n) \end{aligned} \quad (60.4)$$

appelée "**application de codage**" ou "**encodeur**". Le message ou mot a est un élément de E . Il est modifié pour fournir le mot $f(a) = c \in Q^n$. C'est le mot c qui sera transmis et lu par un système quelconque pour donner un message reçu $x = (x_1, \dots, x_n)$ éventuellement entaché d'erreurs.

Notons $C = f(E)$ l'image de f . Comme f est injective, f réalise une bijection de E sur C et C peut être considéré comme l'ensemble de tous les messages possibles. C est appelé "**code de longueur n** ", et les éléments de C s'appellent les "**mots**" du code. Le cardinal du code est par définition celui de C . Nous le noterons $\#C$ ou M . Pour mesurer le degré de différence entre deux mots x et y de Q^n , nous utilisons la "**distance de Hamming**" d définie par:

$$\forall x, y \in Q^n \quad d(x, y) = \#\{i \in \mathbb{N} \mid x_i \neq y_i\} \quad (60.5)$$

Démonstration:

Sur un ensemble Q quelconque, nous définissons donc l'application $d: Q^n \times Q^n \rightarrow \mathbb{R}$ par:

$$d(x, y) = \#\{i \in \mathbb{N} \mid x_i \neq y_i\} \quad (60.6)$$

Si nous notons par δ_x la fonction caractéristique de x :

$$\delta_x(y) = \begin{cases} 1 & \text{si } y = x \\ 0 & \text{sinon} \end{cases} \quad (60.7)$$

alors:

$$d(x, y) = \sum_{i=1}^n (1 - \delta_{x_i}(y_i)) \quad (60.8)$$

Montrons que d est une distance (cf. chapitre de Topologie):

1. Il sera requis que:

$$\forall x, y \in Q^n, d(x, y) \geq 0 \quad (60.9)$$

est évident.

2. Il sera requis que:

$$\forall x, y \in Q^n, d(x, y) = d(y, x) \quad (60.10)$$

est évident aussi.

3. Il sera également requis que:

$$\forall x \in Q^n, d(x, x) = 0 \quad (60.11)$$

est évident aussi.

4. Nous avons aussi:

$$\forall x, y \in Q^n, d(x, y) = 0 \Rightarrow x = y \quad (60.12)$$

où $d(x, y) = 0$ signifie que $x_i = y_i$ pour $i = 1 \dots n$ donc que $x = y$.

5. Enfin:

$$\forall x, y, z \in Q^n, d(x, z) \leq d(x, y) + d(y, z) \quad (60.13)$$

En effet:

$$d(x, z) = \sum_{i=1}^n (1 - \delta_{x_i}(z_i)) \quad (60.14)$$

mais comme:

$$\left(1 - \delta_{x_i}(z_i)\right) \leq \left(1 - \delta_{x_i}(y_i)\right) + \left(1 - \delta_{y_i}(z_i)\right) \quad (60.15)$$

car $\left(1 - \delta_{x_i}(z_i)\right)$ vaut 1 si $x_i \neq z_i$ et 0 sinon, alors:

$$d(x, z) = \sum_{i=1}^n \left(1 - \delta_{x_i}(z_i)\right) \leq \sum_{i=1}^n \left(1 - \delta_{x_i}(y_i)\right) + \left(1 - \delta_{y_i}(z_i)\right) = d(x, y) + d(y, z) \quad (60.16)$$

□C.Q.F.D.

Par exemple, la distance de Hamming entre les mots "ramer" et "cases" ou entre "0100" et "1001" est de 3.

Remarque: Attention les vecteurs seront notés sans la flèche par respect de la tradition dans ce cadre d'étude.

La distance de Hamming (notée d_H dans certains ouvrages) est bien une métrique (cf. chapitre de Topologie) comme nous venons de le démontrer et nous appelons alors "espace de Hamming" sur Q l'ensemble Q^n muni de la métrique d_H .

Définition: Si Q est un groupe, le "poids de Hamming" $w_H(x)$ d'un mot $x \in Q^n$ est le nombre de ses coordonnées non nulles:

$$w_H(x) = d(x, 0) \quad (60.17)$$

où 0 est le mot (vecteur) de Q^n ayant toutes ses coordonnées égales à l'élément neutre de Q . Nous avons par ailleurs la propriété triviale suivante:

$$d(x, y) = d(x - y, 0) = w(x - y) \quad (60.18)$$

Remarque: Lorsque $Q = \{0, 1\}$ nous parlerons de "code binaire" (nous allons voir de suite plus loin une autre forme d'écriture pour cet ensemble binaire) de dimension n égale à 2.

La "distance minimale" du code C est la distance... minimale entre deux mots distincts de ce code. Nous notons ce nombre entier $d(C)$ ou simplement d et donc:

$$d = \min\{d(x, y) \mid x, y \in C \text{ et } x \neq y\} \quad (60.19)$$

ou encore en utilisant la propriété du poids de Hamming $d(x, y) = w(x - y) = w(z)$:

$$d = \min\{w(z) \mid z \in C^*\} \quad (60.20)$$

Exemple:

Considérons le code redondant noté (5, 4) pour 4 mots codés de longueur 5 suivant:

Mot original	Mot code	Poids	Identifiant code
00	00000	0	C1
01	01110	3	C2
10	10011	3	C3
11	11101	4	C4

Distance de Hamming de chacun des couples du code:

$$\begin{aligned}
 d(C1, C2) &= 3 & d(C2, C3) &= 4 \\
 d(C1, C3) &= 3 & d(C2, C4) &= 3 \\
 d(C1, C4) &= 4 & d(C3, C4) &= 3
 \end{aligned} \quad (60.21)$$

Le plus petit poids minimum non nul est donc 3 et la distance de Hamming la plus petite est 3.

Définitions:

D1. Un code C de longueur n , de cardinal M et de distance minimale d est appelé "**code (n, M, d)** ". Les nombres n, M, d sont les "**paramètres du code**". Ainsi, le code $(7, 4, 3)$ est un code de longueur sept, c'est-à-dire que le récepteur reçoit sept bits, de longueur quatre c'est-à-dire qu'une fois décodé, le message contient quatre symboles (lettres) et la distance minimale entre chaque mot de code est de trois.

D2. Nous appelons "**poids minimum**" d'un code C l'entier:

$$\min \{w_H(x) \mid x \in C, x \neq 0\} \quad (60.22)$$

d joue un rôle important car il se trouve en relation étroite avec le nombre d'erreurs susceptibles d'être corrigées. Supposons que le message codé soit $c = (c_1, \dots, c_n)$ et qu'il y ait eu moins de e erreurs de transmission ou de lecture. Le message obtenu $x = (x_1, \dots, x_n)$ vérifie $d(x, c) \leq e$. Nous pouvons retrouver c à partir de x si, et seulement si, il existe un seul mot de code situé à une distance de x inférieure ou égale à e (donc que la distance centre à centre entre deux boules soit égale à $2e$). Autrement dit, il faut et il suffit que les boules fermées de rayon e et centrées sur les éléments du code C soient disjointes. Un code corrigera e erreurs si cette condition est vérifiée.

Ainsi, un code C de distance minimale d corrige au plus:

$$e = \left\lfloor \frac{d-1}{2} \right\rfloor \quad (60.23)$$

erreurs (où $\lfloor \cdot \rfloor$ représente la partie entière d'un nombre réel).

Effectivement, si un message du code se trouve à $d/2$ nous ne pourrions savoir à quel message du code (centre de boule) il appartient puisque se trouvant (de manière imagée) à la tangente de deux boules.

C'est la raison pour laquelle nous prendrons $\frac{d-1}{2}$ qui représente alors la "**distance sûre**" pour corriger au mieux un message erroné du code. De plus, comme le nombre d'erreurs est un entier, il vient naturellement l'écriture:

$$e = \left\lfloor \frac{d-1}{2} \right\rfloor \quad (60.24)$$

Il est clair que le code permet de détecter au plus $d-1$ erreurs. Effectivement, sinon comment distinguer un code erroné d'un code juste (code = message codé) ? Mis à part le fait que chaque élément du code soit différent (application injective de l'ensemble des messages dans l'ensemble des messages codés) il faut en plus pouvoir différencier parmi ceux-ci, ceux qui sont des codes erronés de ceux qui ne le sont pas. D'où le $d-1$.

Exemple:

Considérons le code redondant noté $(5, 4)$ pour 4 mots codés de longueur 5 suivants dont la distance minimale était donc de 3:

Mot original	Mot code	Poids	Identifiant code
00	00000	0	C1
01	01110	3	C2
10	10011	3	C3
11	11101	4	C4

Ce code ne permet donc de détecter au plus que:

$$d-1 = 3-1 = 2 \quad (60.25)$$

erreurs et d'en corriger au plus:

$$e = \left\lfloor \frac{d-1}{2} \right\rfloor = \left\lfloor \frac{3-1}{2} \right\rfloor = \left\lfloor \frac{2}{2} \right\rfloor = 1 \quad (60.26)$$

2.1. CODES EN BLOCS - LINÉAIRES

Définition: Un "**code en bloc**" de taille M et de longueur n défini sur un alphabet de q symboles (1 et 0 pour le langage binaire par exemple) est un ensemble de M vecteurs appelés "**mots du code**". L'idée est que chaque mot d'information composé de k symboles est associé à un mot de code unique composé de n symboles. Les vecteurs sont donc de longueur $n \geq k$ et leurs composantes sont q -aires (donc "2-aires" dans le cas du langage binaire).

La linéarité des codes en blocs signifie que les n symboles du mot code sont obtenus par une combinaison linéaire des k symboles du mot d'information.

Exemple:

Partons des M vecteurs suivants basés sur $q = 3$ symboles binaires (donc $k = 2$). Dans ce cas, nous avons $M = q^k$:

Vecteur
000
100
010
110
001
101
011
111

Nous choisissons alors n comme valant 6 et nous définissons une application bijective telle que:

Vecteur Message	Vecteur Code
000	000 000
100	110 100
010	011 010
110	101 110
001	101 001
101	011 101
011	110 011
111	000 111

Comme le montre cet exemple du code en bloc particulier noté traditionnellement $(n, k) = (6, 3)$, le code ne présente aucune structure particulière. L'opération de décodage implique de faire la comparaison exhaustive du mot reçu en sortie du canal avec l'ensemble des modes de code avant de déterminer le mot de code le plus probable.

Ainsi, selon la définition ci-dessus, un code en blocs C est la résultante d'une application bijective qui associe à chaque vecteur formé de k symboles q -aires (k symboles d'information), un vecteur image de longueur n avec des composantes dans le même alphabet (n symboles codés). Le codage ajoute au débit initial $n-k$ symboles supplémentaires. La quantité:

$$R = \frac{k}{n} \quad (60.27)$$

est appelée le "**rendement de C** ", ou encore "**taux de codage**". L'opération de codage en blocs est "**sans mémoire**", in extenso les blocs sont codés de manière indépendante sans aucune corrélation entre deux blocs consécutifs.

Maintenant il convient de revenir un peu sur les algèbres de Boole (cf. chapitre de Systèmes Logiques). Aux cinq axiomes qui définissent une algèbre de Boole ajoutons en un sixième qui lui confère une structure de corps:

A6. L'algèbre de Boole (extension d'un anneau unitaire par un axiome) munie de la loi $*$ ($\wedge$) est un groupe.

Rappel (théorie des ensembles): un corps est un anneau non nul dans lequel tout élément non nul est inversible.

Si nous prenons l'algèbre de Boole formée des $q = 2$ éléments $\{0,1\}$ formant un ensemble binaire, nous avons effectivement 1 qui est inversible puisqu'il existe x tel que $1 * x = 1$ qui est 1 lui-même.

Ce corps est noté $\mathbb{F}_q = \mathbb{F}_2$. Dans les codes correcteurs, nous travaillons souvent dans $\mathbb{F}_2$ (unique corps à deux éléments) où pour rappel l'addition est définie par:

$$0+0=0, 1+1=0 \text{ (donc } 1=-1), 1+0=1 \quad (60.28)$$

La multiplication étant définie par:

$$1 \cdot 1 = 1, 1 \cdot 0 = 0, 0 \cdot 0 = 0 \quad (60.29)$$

Pour en revenir à notre théorie des codes: l'ensemble des messages $E = \mathbb{F}_q^k$ peut être muni d'une structure d'espace vectoriel de dimension k sur $\mathbb{F}_q$ (cf. chapitre de Théorie Des Ensembles).

Effectivement, il suffisait pour cela que $(E, +)$ soit un groupe abélien et $*$ une loi externe définie par $\mathbb{F}_q^k \times \mathbb{F}_q \rightarrow \mathbb{F}_q^k$. Si nous décidons de n'utiliser que des encodeurs qui sont (des applications) linéaires, le code $C = f(\mathbb{F}_q^k)$ devient un sous-espace vectoriel de $\mathbb{F}_q^n$ (car même si l'application est bijective, comme le corps des messages est fini, nous avons nécessairement un sous-espace vectoriel de l'espace vectoriel de tous les messages codés possibles).

Définition: Un "code linéaire" de dimension k et de "longueur" n est un sous-espace vectoriel de dimension k de $\mathbb{F}_q^n$ (c'est ainsi que cela se dit...). Si la distance minimale de C est d , nous disons que C est un "code $[n, k, d]$ " ou simplement "code $[n, k]$ ".

Remarque: Les codes linéaires sont donc un cas particulier des codes en blocs comme le montre le schéma hiérarchique au début de ce chapitre.

L'ajout de la contrainte de linéarité pourrait nuire à la qualité du code recherché, mais heureusement l'étude des performances montre que les codes linéaires sont très proches des meilleurs codes en blocs. Ainsi, la linéarité facilite l'étude des codes en blocs et permet l'utilisation d'outils algébriques très puissants, sans réduire la classe des blocs linéaires à une classe inefficace.

Notons G la matrice de l'application linéaire $f: \mathbb{F}_q^k \rightarrow \mathbb{F}_q^n$. G est du type $n \times k$ et tout mot c de C s'obtient à partir de tout mot x de E par $c = f(x) = Gx$ où $c = (c_1, \dots, c_n) \in \mathbb{F}_q^n$ et $x = (x_1, \dots, x_k) \in \mathbb{F}_q^k$ sont des vecteurs-lignes avec toujours $n \geq k$. Ainsi, $f(\mathbb{F}_q^k) = C$.

Remarque: Les bases de $\mathbb{F}_2^k, \mathbb{F}_2^n$ sont les bases canoniques courantes (celles dont nous avons souvent fait usage dans le chapitre de Calcul Vectoriel).

Définition: soit C un code linéaire $[n, k]$ et soit $\{g_1, g_2, \dots, g_k\}$ la base de C . Une "matrice génératrice" de C est donc une matrice $n \times k$ dont les colonnes sont formées par les vecteurs g_i de la base.

Soit $u = (u_1, \dots, u_k)$ le mot d'information, in extenso le vecteur contenant les k symboles d'information. Alors, nous pouvons écrire la relation matricielle liant le mot de code c et le mot d'information u :

$$c = Gu \quad (60.30)$$

Définition: Soit C un code en blocs $[n, k]$. Ce code est dit "**code systématique**", si l'ensemble des mots de code contient les k symboles d'informations non modifiés. Les $n-k$ symboles restants sont appelés "**symboles de parité**".

Remarque: Le "code de Hamming" est un tel code. Par ailleurs, les codes systématiques sont des cas particuliers des codes en blocs et nous reviendrons sur leur étude plus loin.

Définition: Soit H une matrice $(n-k) \times n$ à éléments dans $\mathbb{F}_q$, qui vérifie $Hc = 0$ pour tout mot c d'un code linéaire C (en d'autres termes: dont le noyau est C). Alors, H est dite "**matrice de contrôle**" du code C . Réciproquement, c appartient au Code si et seulement si $Hc = 0$. Sinon quoi il y a une erreur !

Remarque: Il est facile de trouver H car celle-ci est "orthogonale" à G puisque la définition ci-dessus implique $H(Gu) = (GH)u = 0$ donc $GH = 0$ (évidemment il ne faut pas prendre $H=0\dots$).

Voyons un exemple de tout cela avec le code de Hamming qui est un code en blocs systématique (attention !! il existerait plusieurs définitions d'un "code de Hamming!):

Cette méthode consiste à doubler l'information, en envoyant autant de bits de parité que de bits de données. Ainsi, à l'aide de matrices, il est possible de détecter et corriger les erreurs qui figurent dans les quartets. Une première matrice est :

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \quad (60.31)$$

Celle-ci est la matrice de codage G . Elle est de dimension $n \times k$, où n est le nombre de bits reçus par paquet, et k le nombre de bits par message contenant l'information (ici $n = 8$ et $k = 4$). Elle permet de générer automatiquement les bits de parité propres à un message. Par exemple pour envoyer le message 1101, il faut, pour respecter la règle de multiplication des matrices, considérer ce quartet comme un vecteur-colonne:

$$\begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} \quad (60.32)$$

Donc en multipliant, nous obtenons:

$$c = Gm = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \quad (60.33)$$

Nous enverrons donc l'octet 11010110, dont les quatre premiers bits forment le message et les quatre derniers les bits de parité, qui servent à vérifier la véracité/intégrité du message.

La matrice de contrôle correspondante H est:

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (60.34)$$

Ainsi lorsque le destinataire reçoit l'octet 11000110 au lieu de 11010110, le décodage donne comme "syndrome":

$$s = Hc = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \quad (60.35)$$

Le vecteur-colonne obtenu n'est donc pas nul. Il y a donc une erreur. Avec la matrice de contrôle, la théorie permet d'affirmer que comme le vecteur obtenu est le même que celui qui est en quatrième position dans la matrice de décodage, l'erreur est due au quatrième bit. Comme nous sommes en base 2, il suffit de changer le 0 en 1. Ce codage de l'information est coûteux, car il occupe deux fois plus de bande passante. Cependant c'est l'un des moyens les plus efficaces pour sécuriser l'information.

Pour montrer que le syndrome d'un code de Hamming correspond à une des colonnes de la matrice de

contrôle, notons e_i les vecteurs-colonnes de la base canonique sur $\mathbb{F}_2^n$, $e_i = (0, \dots, 0, 1, 0, \dots, 0)$ avec 1 à la i -ème place. Soit x un mot de code. Nous avons donc par définition de H , $Hx = 0$. Supposons que le mot reçu, que nous noterons $\tilde{x}$, soit entaché d'une seule erreur et que cette erreur soit sur le j -ème bit. Nous avons donc:

$$x + \tilde{x} = e_j \text{ et } He_j = H(x + \tilde{x}) = Hx + H\tilde{x} = H\tilde{x} \quad (60.36)$$

ainsi il vient que $H\tilde{x} = He_j$, mais He_j est le j -ème vecteur colonne de la matrice H .

Ceci nous montre bien que, lorsque nous recevons $\tilde{x}$ et que nous calculons $H\tilde{x}$ nous obtenons le vecteur-colonne de la matrice H situé exactement à l'emplacement de l'erreur (en l'occurrence j).

Remarque: Un syndrome nul ne signifie pas l'absence d'erreur(s). Il existe donc des configurations d'erreurs indétectables.

Notons maintenant:

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \text{ et } I_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (60.37)$$

alors nous remarquerons que G et H sont formées par les blocs I et A de la manière suivante:

$$G = (I_k \mid A) = (I_4 \mid A) \quad (60.38)$$

et:

$$H = (A \mid I_k) \quad (60.39)$$

Ainsi:

$$HG = (A \mid I_k) \begin{pmatrix} I_k \\ A \end{pmatrix} = AI_k + I_k A = A + A = A(1+1) = 0 \quad (60.40)$$

car $1+1=0$ dans $\mathbb{F}_2$.

De façon générale, si nous travaillons avec l'alphabet $\mathbb{F}_2$ et si $G = \begin{pmatrix} I_k \\ A \end{pmatrix}$ où A est une matrice $(n-k) \times k$ alors $H = (A \mid -I_k)$ est aussi une matrice de contrôle car de nouveau:

$$HG = (A \mid -I_k) \begin{pmatrix} I_k \\ A \end{pmatrix} = AI_k - I_k A = 0 \quad (60.41)$$

Remarque: Dans $\mathbb{F}_2$, $I_k = -I_k$ car $1 = -1$, c'est pour ça que nous avons écrit $H = \begin{pmatrix} A & I_k \end{pmatrix}$ dans l'exemple précédent.

2.2. CODES SYSTÉMATIQUES

Construire un code systématique consiste à adjoindre à chaque mot $x = (x_1, \dots, x_k)$ du message $n-k$ symboles $c_{k+1}, \dots, c_{k+n}$ dépendants linéairement des x_i pour obtenir le mot de code $c = f(x)$.

Les symboles sont appelés "**bits de contrôle**" et (nous en verrons un exemple juste plus bas):

$$c = (x_1, \dots, x_k, c_{k+1}, \dots, c_{k+n}) = (I_k \mid A)(x_1, \dots, x_k) \quad (60.42)$$

où, pour rappel, $(I_k \mid A)$ désigne la matrice $n \times k$ obtenue en écrivant l'une en dessous de l'autre, la matrice identité I_k de taille k et une matrice quelconque A .

Nous dirons qu'un code C est systématique s'il possède une matrice génératrice de la forme $G = (I_k \mid A)$

Exemple:

Nous nous proposons de construire un code linéaire systématique avec $n=k=3$. Nous notons a_1, a_2, a_3 les bits d'information. Les bits de contrôle a_4, a_5, a_6 seront définis par:

$$\begin{aligned} a_4 &= a_1 + a_2 \\ a_5 &= a_2 + a_3 \\ a_6 &= a_1 + a_3 \end{aligned} \quad (60.43)$$

La matrice génératrice G est telle que sa partie supérieure est la matrice identité de dimension 3 (nous avons la même chose pour le code de Hamming). La première ligne (110) de la matrice A correspond à l'expression du bit de contrôle a_4 :

$$a_4 = 1 \cdot a_1 + 1 \cdot a_2 + 0 \cdot a_3 \quad (60.44)$$

etc. pour chaque bit de contrôle.

La matrice génératrice G s'écrit alors:

$$G = \begin{bmatrix} I_k \\ A \end{bmatrix} = \begin{bmatrix} I_3 \\ A \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \quad (60.45)$$

En multipliant cette matrice par les $2^3 = 8$ vecteurs possibles (les mots constitués de trois bits d'information), nous obtenons les mots code:

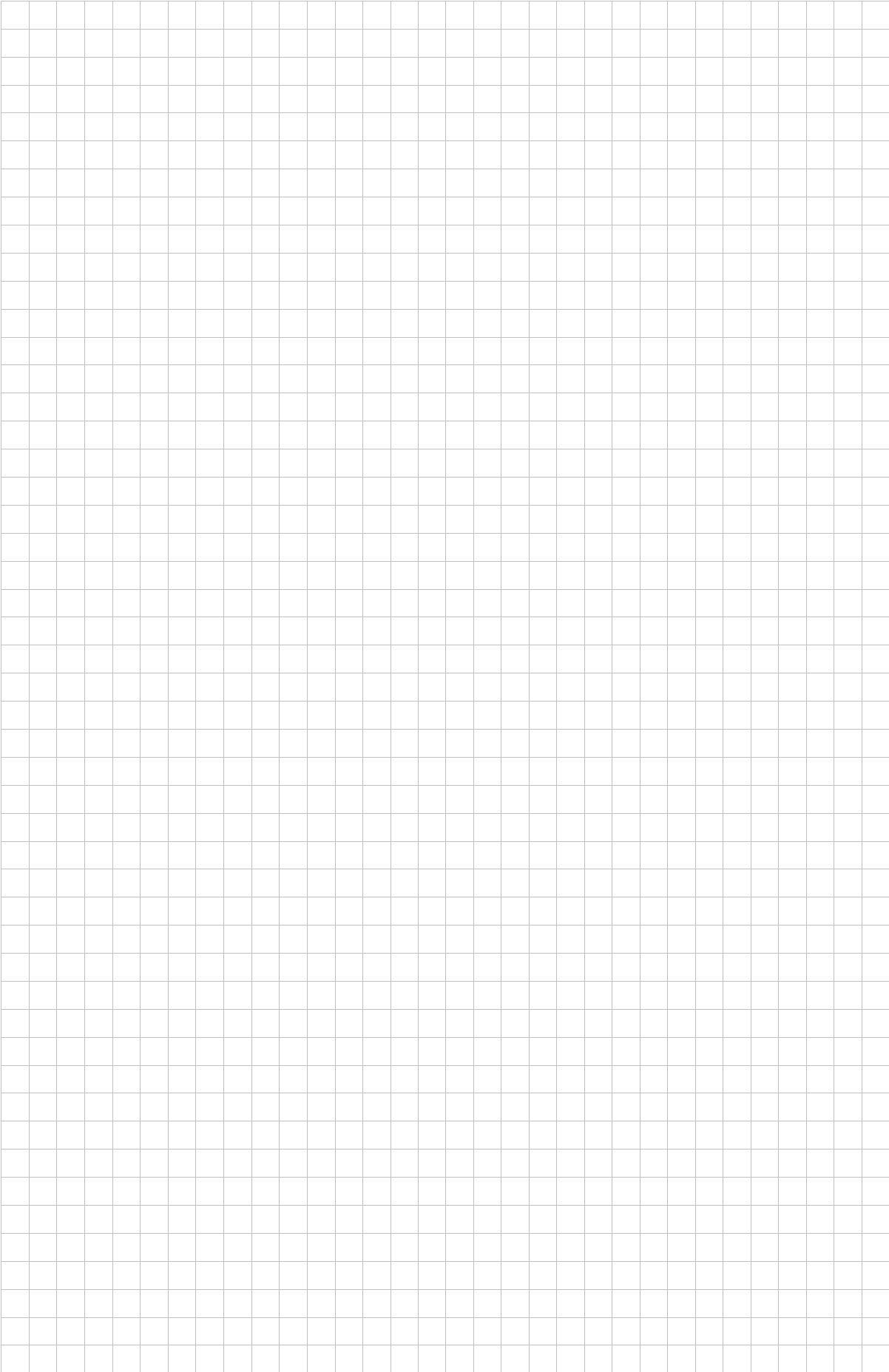
a1	a2	a3	a4	a5	a6
0	0	0	0	0	0
0	0	1	0	1	1
0	1	0	1	1	0
0	1	1	1	0	1
1	0	0	1	0	1
1	0	1	1	1	0
1	1	0	0	1	1
1	1	1	0	0	0

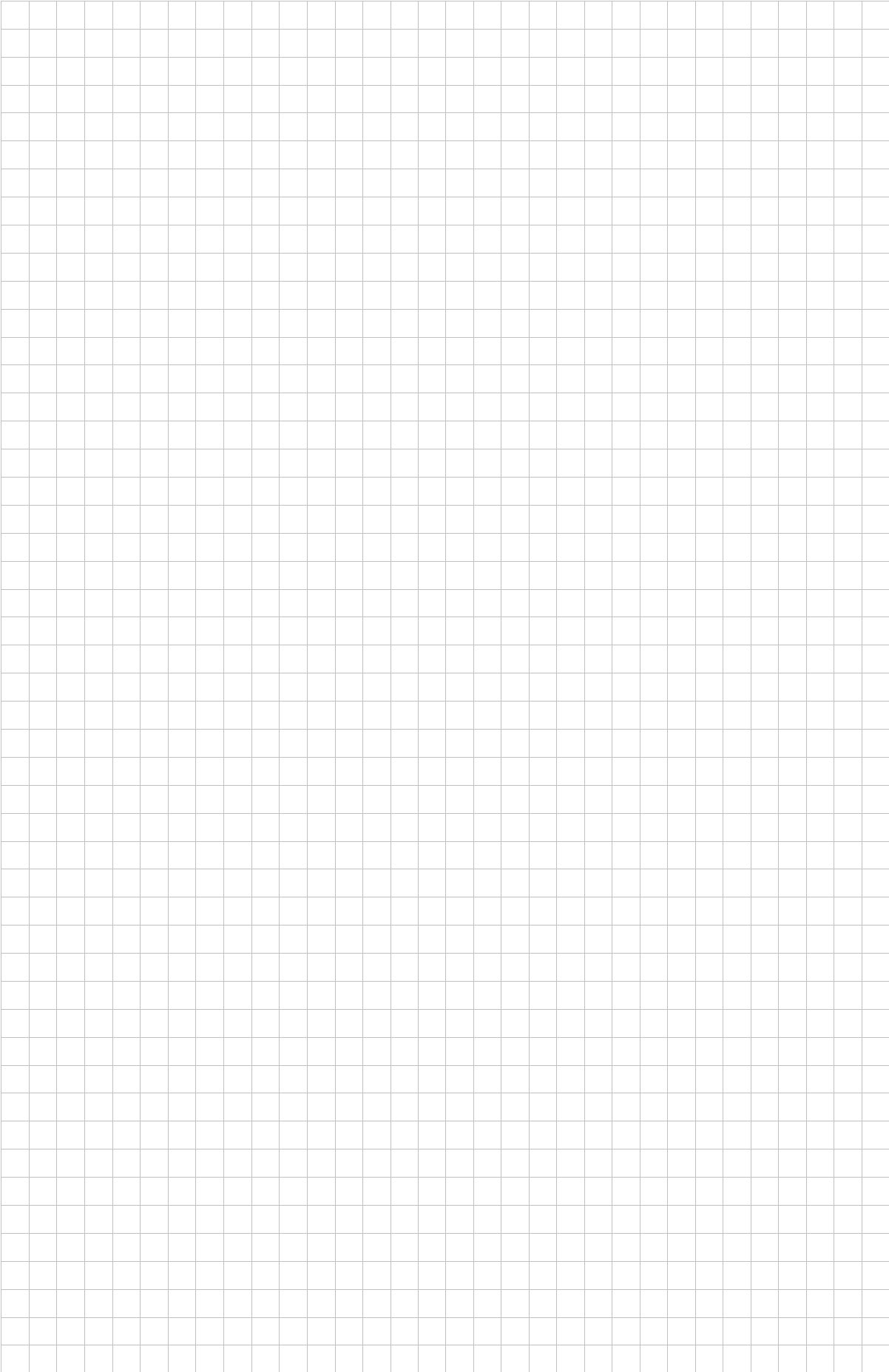
Tableau: 60.1 - Exemples de mots code

Nous constatons donc que le poids minimum des mots code est 3. Donc le code détecte $3-1=2$ erreurs et peut en corriger $\left\lfloor \frac{3-1}{2} \right\rfloor = 1$.

Notes personnelles:







61. CRYPTOGRAPHIE

La cryptographie est une des disciplines de la cryptologie s'attachant à protéger des messages (assurant confidentialité et/ou authenticité) que deux personnes souhaitent s'échanger à travers un canal peu sûr en s'aidant souvent de secrets ou clés.

L'histoire de la cryptographie est déjà longue et passionnante (puisque c'est en quelque sorte un "jeu"). Nous rapportons son utilisation en Égypte il y a 4'000 ans. Toutefois, pendant des siècles, les méthodes utilisées étaient restées souvent très primitives. D'autre part, sa mise en oeuvre était limitée aux besoins de l'armée et de la diplomatie. Ainsi, les méthodes de chiffrement et de cryptanalyse (le casse de code) ont connu un développement très important au cours de la seconde guerre mondiale et ont eu une profonde influence sur le cours de celle-ci.

À la fin du 20ème siècle (en particulier !), avec la prolifération des ordinateurs et des moyens électroniques de communication, il était devenu de plus en plus important d'utiliser des codes secrets pour la transmission des données entre les organismes à caractère militaire ou privé. Ainsi, les ingénieurs ont dû chercher à cette même époque des méthodes numériques solides et dont la mise en oeuvre et l'usage était à portée de presque tout un chacun (nation, entreprise et individu) tout en faisant en sorte que les attaques extérieures nécessitent des outils hors d'atteinte d'un individu ou groupe d'individus équipé d'outils informatiques standards ou performants (en puissance de calcul donc). Les ingénieurs et chercheurs se sont alors plongés dans la mathématique pour chercher les outils satisfaisant ce cahier des charges et pour les systèmes les plus connus, les théories mathématiques qui furent adoptées avaient plus de 200 ans (cryptographie quantique mise à part) d'ancienneté.

Les techniques de stéganographie (art de dissimuler un message dans un autre, ou dans une image) doivent toutefois être préservées car rien ne nous dit que la puissance de calcul de l'informatique sera toujours disponible en temps de guerre. Il convient de souligner aussi que la stéganographie déploie des trésors d'imagination. Signalons par exemple: les permutations de lettres, les formatages spéciaux et subtiles de caractères, l'utilisation de synonymes, les messages cachés dans la virgule d'un texte ou derrière un timbre-poste, dans des coups de jeux d'échecs (d'où le fait que ces jeux aient été interdits par les américains pendant quelques années après l'attaque de Pearl Harbor), dans des images/dessins, dans des partitions musicales, etc. Toutes ces techniques font que pendant la deuxième guerre mondiale, l'office de censure aux États-Unis occupait 10'000 employés à plein temps qui analysaient le courrier des citoyens, les petites annonces, les textes radios, etc.

Remarques:

R1. Pour aborder les fondements de la théorie de la cryptographie, nous conseillons au lecteur d'avoir parcouru au préalable les chapitres de Théorie des Nombres, Théorie des Ensembles, de Méthodes Numériques (surtout la partie traitant de la complexité algorithmique), des Systèmes Numériques Formels, de la Mécanique Statistique (où se trouve la théorie de l'information) et pour la partie de la cryptographie quantique: le chapitre d'Informatique Quantique du site.

R2. Il faut rester conscient à nouveau que la cryptographie étant plus une science de l'ingénieur qu'une science du physicien (mis à part en ce qui concerne la cryptographie quantique) il ne faut pas s'étonner alors de voir apparaître des algorithmes tombés un peu de nulle part et adoptés par l'industrie parce qu'ils marchent bien... par ailleurs il est certain que seulement quelques années après avoir écrit ce texte il soit déjà considéré comme obsolète (c'est tout l'art de l'ingénierie... l'obsolescence programmée)

1. SYSTÈMES CRYPTOGRAPHIQUES**Définitions:**

Un "système cryptographique" est composé de:

D1. Un ensemble fini P appelé "l'espace des textes clairs"

D2. Un ensemble fini C appelé "l'espace des textes chiffrés"

D3. Un ensemble fini K appelé "l'espace de clés"

Pour chaque clé $k \in K$, nous cherchons une fonction de chiffrement:

$$e_k : P \rightarrow C \quad (61.1)$$

et une fonction de déchiffrement (decryption):

$$d_k : C \rightarrow P \quad (61.2)$$

telles que (cf. chapitre de Théorie Des Ensembles):

$$\boxed{d_k \circ e_k = Id_P} \quad (61.3)$$

Autrement dit, ces deux fonctions doivent être injectives!

Pour arriver à ce résultat, deux types de techniques cryptographiques se distinguent, englobant toutes les méthodes de cryptage modernes connues (pour les détails voir plus loin):

1. Les premières concernent les systèmes de chiffrement "symétriques à clé secrète".

Remarque: Les clés publiques font souvent référence au protocole D.E.S. (voir plus loin): Data Encryption System.

2. Les secondes concernent les systèmes de chiffrement "asymétriques à clé publique".

Remarque: Ce type de clé fait souvent référence par exemple au protocole R.S.A., du nom des personnes à qui on en a attribué le développement: Rivest, Shamir et Adleman. Elles sont beaucoup utilisées de par la rapidité du temps de cryptage et de décryptage ainsi que de leur grande entropie (voir définition plus loin).

Par nature, ces deux types de clés sont très différents. Essayons d'en comprendre les raisons:

Un chiffrement symétrique désigne un système où la clé utilisée dans l'opération de chiffrement est aussi celle utilisée dans l'opération de déchiffrement. Dans ce cas, lors d'un échange sécurisé (supposé), les deux parties de la correspondance doivent partager un secret: la clé utilisée ou "**clé de session**".

Un chiffrement asymétrique désigne un système de chiffrement où la clé utilisée pour le chiffrement (clé privée de l'expéditeur) diffère de celle utilisée pour le déchiffrement (clé privée du destinataire). Le seul échange qu'il y a entre les membres du groupe est la clé publique, qui permet à chacun des membres d'adapter son chiffrement (ou cryptage) en fonction de la clé privée des autres membres (parmi les nombreux systèmes asymétriques qui ont été proposés, l'un des plus répandu en ce début de 21ème siècle est le R.S.A.).

Remarques:

R1. En 2001, MS Internet Explorer (navigateur internet de Microsoft) fonctionnait avec un système asynchrone de 1024 bits certifié par un système synchrone et Adobe Acrobat (PDF) en 2004 avec un système A.E.S. (Advanced Encryption System) de 128 bits pour les protections basses de même que pour l'iPhone 4S et 5.

R2. MS Windows et son système E.F.S. (Encryption File System) utilise une clé symétrique (pour chiffrer le fichier) appelée "**File Encryption Key**" et la cryptographie asymétrique pour coder la clé symétrique dans l'en-tête du fichier selon le schéma suivant (les clés étant mises à jour régulièrement via les certificats racines de Windows Update):

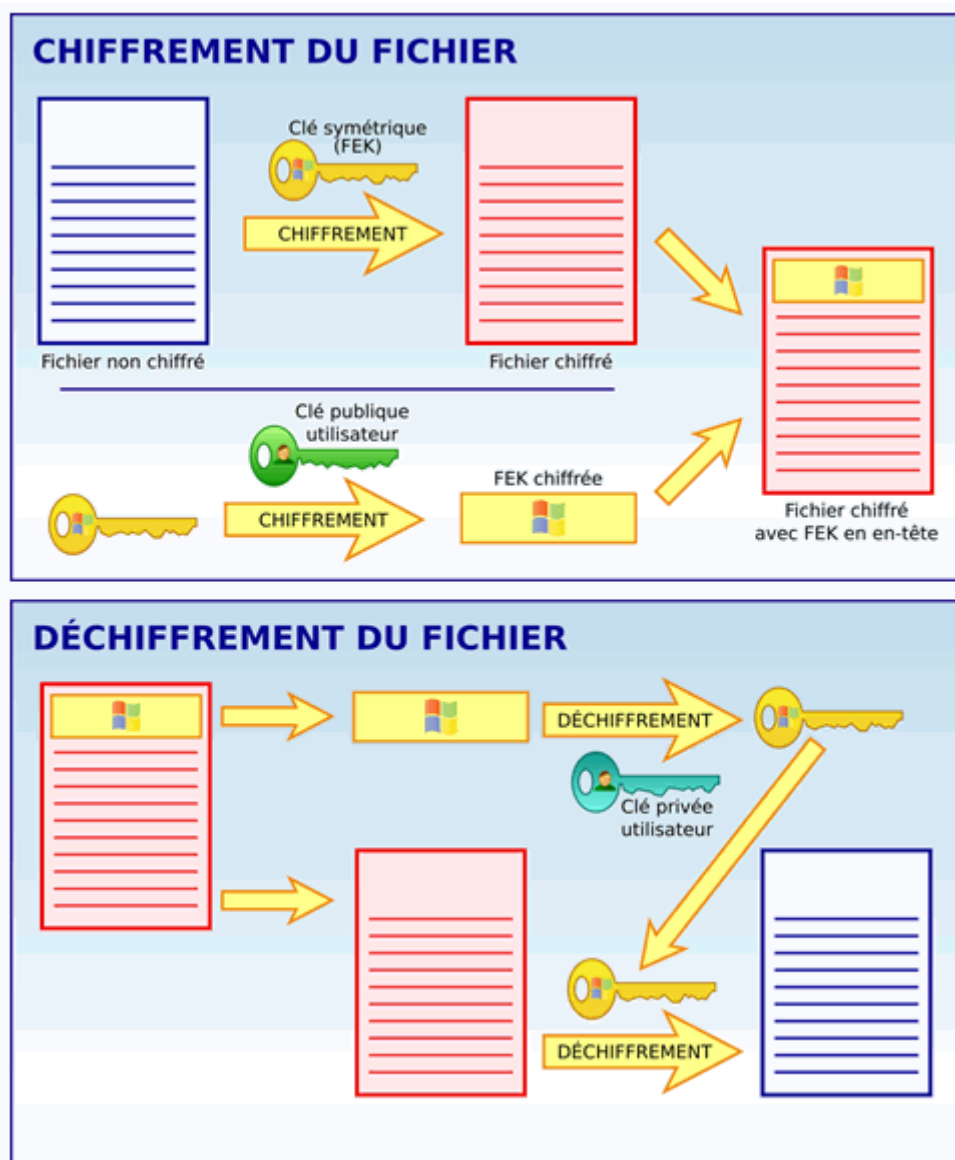


Figure: 61.1 - Principe de l'E.F.S. dans l'O.S. Microsoft Windows (source: Wikipédia)

Ces méthodes demeurent toujours déchiffrables, à condition que l'intercepteur possède "assez de temps et de papier" (exception à ce jour pour le cryptage quantique).

Voici un petit tableau résumé des clés cassées et de leur taille respective pour chacun des deux

systèmes classiques:

Clé secrète (système symétrique) Recherche exhaustive	
Nombre de bits	Année
40	Cassée en 1995
56	Cassée en 1998
64	Cassable
128	Cassable en ~2100
256	?
Clé publique RSA (système asymétrique) Factorisation	
Nombre de bits	Année
256	Cassée en 1985
512	Cassée en 1999
1024	Cassée 2010
2048	Cassable en ~2100
4096	?

Tableau: 61.1 - Systèmes de clés et cassages récents

1.1. PRINCIPE DE KERCKHOFFS

La première fonction de la cryptographie consiste donc à assurer la confidentialité d'un échange d'informations. Deux parties d'un échange confidentiel s'accordent d'abord sur une convention secrète pour rédiger leurs messages, et si elles l'ont soigneusement choisie, personne d'autre ne devrait pouvoir saisir leur échange.

Si le caractère secret de telles conventions est envisageable entre quelques personnes isolées pour une période limitée, il est inconcevable à grande échelle et pour une durée assez longue. C'est ce qu'avait compris Auguste Kerckhoffs lorsqu'il établit les principes de base de la cryptographie pratique dont un principe fondamental exige un système de chiffrement: "qui n'exige pas le secret, et qui puisse sans inconvénient tomber entre les mains de l'ennemi".

Un autre principe précise que: "la clé doit pouvoir être changée ou modifiée au gré des correspondants".

Le premier de ces deux principes, connu aujourd'hui sous le nom de "**principe de Kerckhoffs**", stipule donc que la sécurité d'un système de chiffrement n'est pas fondée sur le secret de la procédure qu'il suit, mais uniquement sur un paramètre utilisé lors de sa mise en oeuvre: la clé. Cette clé est le seul secret de la convention d'échange.

Ce principe a cependant été reformulé par Claude Shannon: "l'adversaire connaît le système". Cette formulation est connue sous le nom de la "**maxime de Shannon**". C'est le principe le plus souvent adopté par les cryptologues, par opposition à la sécurité par l'obscurité.

1.2. TRAPPES

Il existe parfois ce que nous nommons des "**trappes**" dans les clés publiques et secrètes. Ceci est dû au

fait que lors de la génération de la clé, qui doit se faire aléatoirement en respectant certaines contraintes théoriques prédéfinies, le générateur aléatoire peut avoir un défaut (parfois le défaut est volontaire de la part du fournisseur... espionnage oblige).

Dans les clés secrètes, les trappes se situent au niveau de "**l'entropie de la clé**" (cf. chapitre de Mécanique Statistique), directement liée à l'entropie du générateur aléatoire. Nous pouvons de manière simpliste définir l'entropie d'un générateur de clés par le nombre moyen optimal de questions binaires (c'est-à-dire donnant lieu à des réponses du type Oui/Non) qu'il faut poser à quelqu'un connaissant une clé produite par ce générateur, pour la déterminer. Plus l'entropie d'un générateur de clé est élevée, plus il faut de questions pour déterminer cette clé. À l'inverse, plus l'entropie est faible, moins il faut de questions, de sorte que la recherche d'une clé est facilitée.

L'introduction de trappes dans les clés de systèmes asymétriques est beaucoup plus difficile, puisque ce type de clé possède déjà une structure mathématique intrinsèque: leur construction n'est pas due au hasard, mais résulte de règles mathématiques. Le hasard est ici dans le choix des grands nombres premiers utilisés. Le fait que les systèmes asymétriques puissent être aisément calculés, mais difficiles à inverser font qu'ils sont parfois appelés "**fonctions trappes à sens unique**".

Remarque: Si le générateur aléatoire qui engendre ces nombres premiers est biaisé (cf. chapitre de Statistiques), ce biais facilitera la recherche des nombres premiers ayant servi à l'élaboration de la clé qu'un attaquant tente de casser.

2. SYSTÈME DE CHIFFREMENT A CLÉ SECRÈTE

Le "**chiffre à usage unique**" est un algorithme de chiffrement à clé secrète prouvé inconditionnellement sûr. Correctement utilisé (et c'est un point important), il fournit un chiffrement incassable en des temps raisonnables.

Les bases théoriques de ce système de cryptage sont les suivantes:

Soit un message M sous forme binaire à transmettre entre des personnes A (créateur et expéditeur du message M) et B (lecteur et destinataire). Nous engendrons une grande quantité de bits "réellement aléatoires" qui forment une clé secrète K de même taille que le message à transmettre (les programmes informatiques, déterministes par essence, ne peuvent engendrer des bits vraiment aléatoires).

Cette clé sera transmise à B par un canal supposé sûr... Un laps de temps donné après la transmission de cette clé, A va encoder son message en C en effectuant l'opération:

$$C = K * M \quad (61.4)$$

où $*$ est un opérateur qui doit satisfaire à une loi de groupe (cf. chapitre de Théorie Des Ensembles) sur un ensemble fini (qui contient un nombre fini d'éléments).

L'intérêt en informatique est d'utiliser la loi de groupe XOR (aussi nommée OU EXCLUSIF) notée $\oplus$ par la suite (cf. chapitre de Systèmes Logiques).

Finalement, l'expéditeur A transmet la version cryptée C de son message par une voie pas nécessairement sécurisée. B retrouve le message original M en utilisant l'opération inverse $(*)^{-1}$ de $*$ (l'opérateur XOR est son propre inverse comme le montre sa table de vérité dans le chapitre de

Systèmes numériques). Ainsi B va effectuer l'opération suivante:

$$M = C \oplus K \quad (61.5)$$

Sous réserve que la clé K ait bien été engendrée de façon totalement aléatoire et que chaque bit la composant n'ait été utilisé qu'une seule fois pour crypter le message, un intercepteur n'obtient aucune information sur le message clair M s'il intercepte C . En effet, dans ces conditions, on ne peut établir aucune corrélation entre M et C sans la connaissance de K .

Même avec de futurs ordinateurs quantiques ultra-puissants, le problème est insoluble, car rien ne relie les informations dont on dispose et le problème à résoudre. En conséquence, le "chiffre à usage unique" est un algorithme de chiffrement "inconditionnellement sûr". La preuve de sa sécurité ne fait pas appel à des conjectures mathématiques non démontrées et les tentatives de déchiffrement d'un intercepteur muni d'une puissance de calcul infinie sont vaines.

Cependant, chaque étape du chiffrement est source d'erreurs possibles. En effet, la clé K peut avoir été mal élaborée. La moindre déviation statistique sur K par rapport à du "vrai" aléatoire fournit des informations sur le message clair M à partir de sa version cryptée. C'est la raison pour laquelle les bits de K ne doivent servir qu'une seule fois.

Effectivement, supposons qu'une même clé ait servi à chiffrer les messages de langue française M_1 et M_2 et qu'une personne malveillante arrive à intercepter les deux messages correspondants cryptés C_1 et C_2 . À partir de C_1 et C_2 l'intercepteur peut facilement obtenir des informations sur M_1 et M_2 du fait des particularités des langues. En effet, puisque:

$$C_1 = M_1 \oplus K \text{ et } C_2 = M_2 \oplus K \quad (61.6)$$

alors l'intercepteur connaît un résultat simple qui fait intervenir M_1 et M_2 , sans la clé K :

$$C_1 \oplus C_2 = M_1 \oplus M_2 \quad (61.7)$$

car:

$$K \oplus K = 0 \quad (61.8)$$

(au besoin faire la table de vérité pour s'en convaincre). Or, si M_1 et M_2 sont dans la même langue, on saura en général, grâce aux redondances des langues (par exemple la lettre "e" apparaît très souvent dans la langue française), retrouver à partir de $M_1 \oplus M_2$, chacun des deux messages (le travail est quand même laborieux).

Imaginons que nous souhaitons envoyer un tout petit message codé en binaire par 1101 et que nous avons généré une clé aléatoire qui a donné 0101.

Nous avons alors:

$$C = M \oplus K = 1101 \oplus 0101 = 1000 \quad (61.9)$$

et donc:

$$M = C \oplus K = 1000 \oplus 0101 = 1101 \quad (61.10)$$

Évidemment dans ce genre de petites situations on peut deviner sans trop de difficultés M rien qu'en ayant C s'il n'y a comme ici qu'une seule étape de chiffrement. Raisons pour lesquelles il existe des schémas comme nous allons le voir maintenant.

Le problème principal de cette technique est donc la création d'une clé aussi aléatoire que possible. Pour pallier à cela, les mathématiciens font passer la clé par une série de fonctions imbriquées, dont le résultat, après un grand nombre d'itérations, devient "pseudo-aléatoire".

Construire une itération pseudo-aléatoire est une chose, construire une bijection pseudo-aléatoire en est cependant une autre. En effet, il faut pouvoir décrypter le message par la suite, c'est pourquoi l'on a absolument besoin d'un système bijectif (qui a tout élément d'arrivée - message crypté - fait correspondre un unique élément de départ - message décrypté - et inversement).

2.1. SCHÉMA DE FEISTEL

Même si les algorithmes de chiffrement de fin du 20ème siècle et du début du 21ème siècle se contentent d'une clé à nombre de bits fini, l'objectif demeure l'élaboration à partir d'un message M d'une suite aléatoire de chiffres, ou, à défaut, qui paraisse aléatoire, que la détention de la clé K permet de déchiffrer. Concrètement, cet objectif demande de construire ou d'identifier une fonction qui, d'une part, fasse correspondre à chaque chiffre de M un chiffre de C qui semble tiré au hasard (mais dont la valeur dépend en réalité de la clé) et, d'autre part, qui autorise le cheminement inverse, c'est-à-dire qu'à partir d'un chiffre de C , on puisse remonter de façon univoque au chiffre correspondant de M (donc une fonction bijective!). Nous désirons ainsi trouver une bijection pseudo-aléatoire.

Dans les années 1950, un mathématicien (Horst Feistel) a montré qu'une fonction pseudo-aléatoire se transforme, par une méthode relativement simple, en bijection. Aujourd'hui, la "**méthode de Feistel**" est la plus utilisée dans les chiffrements à clé secrète et est aussi à la base du D.E.S. (Data Encryption System). Comment fonctionne-t-elle ?

En voici le principe:

Le message initial à chiffrer a une taille de $2n$ bits. On sépare le message M en deux blocs, G et D , de longueurs égales (G regroupe les n premiers bits et D les suivants) et on construit la transformation φ qui associe à G et D les nombres S et T telle que:

$$S = D \text{ et } T = G \oplus f_1(D) \quad (61.11)$$

où le signe $\oplus$ représente toujours l'opération XOR bit à bit et f_1 une fonction quelconque, pas nécessairement bijective, de n bits vers n bits qui utilise la clé secrète K .

La transformation $\varphi(G, D) = (S, T)$ est bien bijective, car on remonte de façon univoque (unique) à partir de S et de T à G et D par les opérations:

$$D = S \text{ et } G = T \oplus f_1(S) \quad (61.12)$$

On ne doit évidemment pas s'arrêter-là puisque la partie droite du message, D , n'a pas été chiffrée, elle est simplement passée à gauche. Cependant, comme φ est bijective, on peut réitérer le processus. Un schéma de Feistel où l'on applique n fois la fonction φ est nommé "**schéma à n étapes**".

Exemple:

Nous allons chiffrer par la méthode de Feistel à deux étapes un message constitué de quatre bits (donc 16 possibilités de messages), ce qui vient à construire une bijection de quatre bits vers quatre bits à partir de deux fonctions f_1, f_2 de deux bits vers deux bits. Les fonctions f_1, f_2 possèdent en entrée à la fois le message à chiffrer et la clé secrète. Nous considérerons que pour une certaine clé entrée, ces fonctions sont les suivantes:

Entrée	f_1	Sortie	Entrée	f_2	Sortie
00	→	01	00	→	11
01	→	11	01	→	00
10	→	10	10	→	00
11	→	01	11	→	01

Tableau: 61.2 - Correspondances entrées/sorties de clés par fonctions

Notons que ni f_1 , ni f_2 ne sont des bijections ($f_1(00) = f_1(11) = 01, f_2(01) = f_2(10) = 00$). À titre d'exemple, chiffrons le message 1101. G désigne la moitié gauche du message à chiffrer, D la moitié droite:

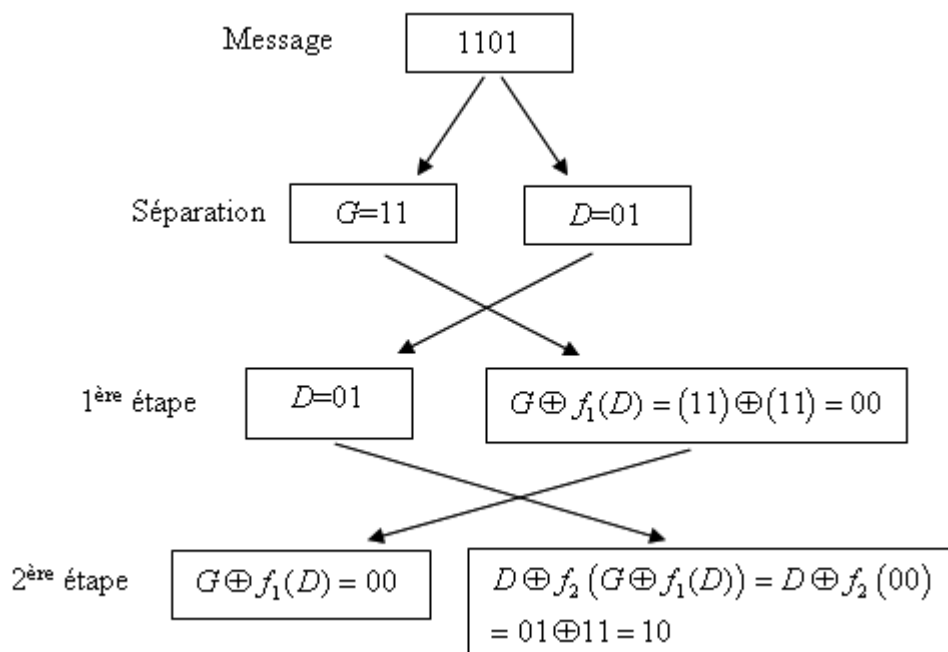


Figure: 61.2 - Chiffrement de 1101 par la méthode de Feistel

Le résultat est 0010. Nous calculerons l'image des 15 autres messages possibles et nous vérifierons qu'il y ait une correspondance univoque entre chaque message et son image par le schéma de Feistel: nous avons construit une bijection à partir de deux fonctions qui n'en sont pas.

Des résultats théoriques complexes garantissent la sécurité cryptographique des schémas de Feistel à partir de quatre étapes lorsque n est assez grand et lorsque les fonctions f_1, f_2, f_3, f_4 sont indiscernables de fonctions réellement aléatoires. En pratique, plutôt que d'utiliser quatre étapes et des fonctions f_i qui ont l'air aléatoires, on préfère en général utiliser plus d'étapes et des fonctions f_i plus

simples. Au bout de quelques étapes, la bijection obtenue devient souvent très difficile à distinguer des bijections aléatoires. Et pour des paramètres bien choisis, on ne sait plus du tout comment les distinguer de bijections réellement aléatoires !!!

La plupart des algorithmes à chiffrement à clé secrète utilisés actuellement dans le monde civil sont des schémas de Feistel. En particulier, l'algorithme D.E.S. (Data Encryption System) qui est un schéma de Feistel à 16 étapes comme représenté dans la figure ci-dessous et l'algorithme Triple DES (TDES) qui est un schéma de Feistel à 48 étapes et l'algorithme Blowfish (que nous n'aborderons pas ici).

Remarque: Il y a, par exemple, dans beaucoup de cartes bancaires, une clé DES (ou TDES depuis octobre 2001) qui apporte la preuve de la légitimité de la carte entre le centre de contrôle de la banque et le terminal du commerçant en plus de la partie publique d'une clé RSA pour s'assurer de la saisie du code utilisateur (contrôle fait par une puce interne à la carte dont la fabrication doit alors se faire dans des locaux très sécurisés).

Rigoureusement le schéma de Feistel est un peu autre car il fait intervenir des clés, ce que nous n'avons pas utilisé dans l'exemple cité précédemment. Voici au fait en un peu plus détaillé en quoi consiste ce schéma de Feistel (voir figures ci-dessous).

Principe du schéma: Un message à chiffrer est découpé en blocs de 64 bits, chacun d'eux étant séparé en deux sous-blocs de 32 bits, le bloc de gauche (G) et le bloc de droite (D). À chaque itération, l'ancien bloc droit devient le nouveau bloc gauche et le nouveau bloc droit résulte de la combinaison par l'opération XOR de l'ancien bloc droit, dont les bits sont mélangés par une fonction de confusion, et de l'ancien bloc gauche. On répète l'itération 16 fois.

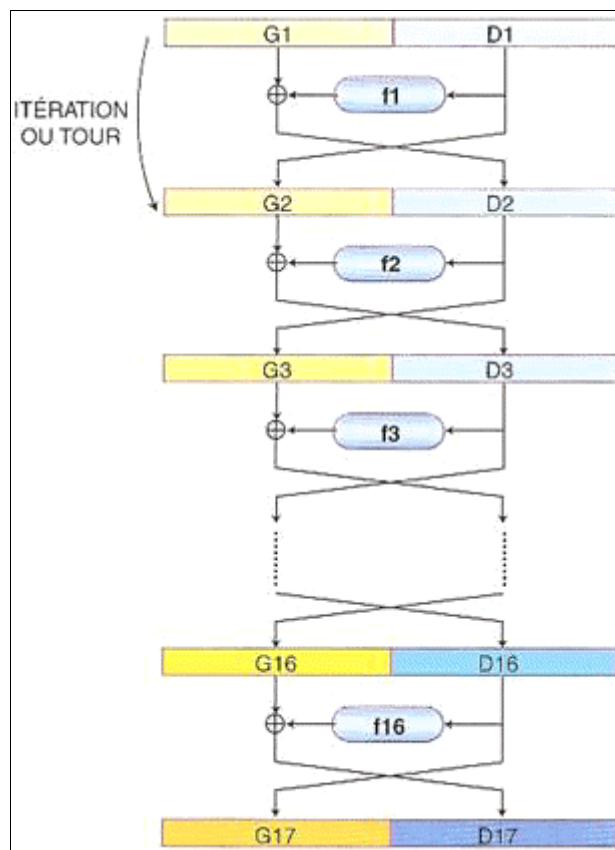


Figure: 61.3 - Schéma de Feistel un peu plus réaliste

La fonction de confusion (f), qui agit sur les blocs de 32 bits, mélange les bits selon les processus suivants (à droite sur la figure). D'abord, elle transforme le bloc de 32 bits en un bloc de 48 bits par duplication de certains bits (expansion). Ensuite, elle ajoute à ce bloc, une sous-clé de 48 bits (clé de tour) extraite de la clé secrète de 56 bits puis elle transforme chaque ensemble de 6 bits en 4 bits par des transformations locales (transformation S). On aboutit à un bloc de 32 bits que l'on mélange enfin suivant une permutation fixe.

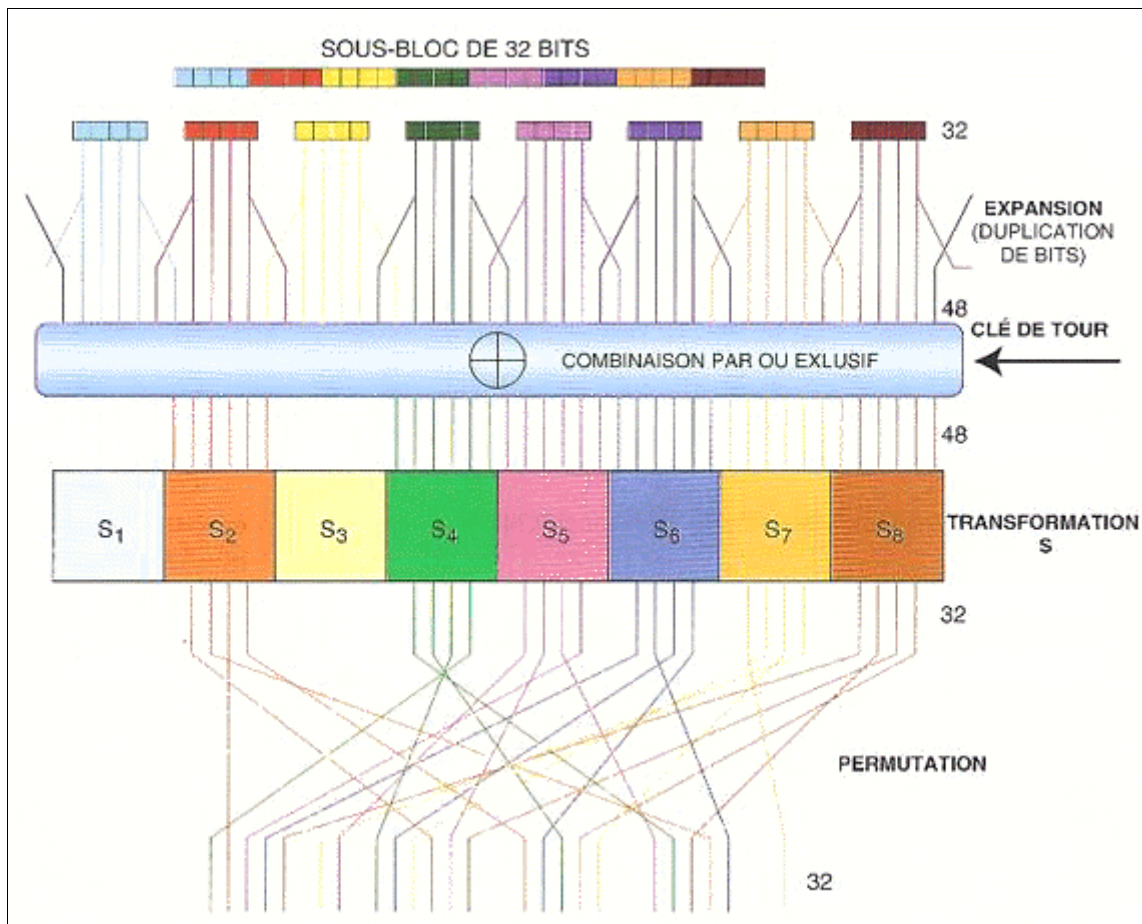


Figure: 61.4 - Schéma de la fonction de confusion

3. SYSTÈME DE CHIFFREMENT A CLÉ PUBLIQUE

En 1975, W. Diffie et M.E. Hellman révolutionnaient la science de la cryptographie en démontrant l'existence d'un protocole qui ne pouvait être déchiffré par un intercepteur à moins que ce dernier ne disposât de conséquentes ressources informatiques. Le plus fascinant dans leur méthode - dont le principe est encore en usage aujourd'hui - c'est que le code utilisé ne nécessite pas le camouflage de la méthode employée et qu'il peut servir à maintes reprises sans aucune modification (principe de Kerckhoffs). Ils ont à l'époque tout simplement créé le concept de cryptographie à clé publique, ou cryptographie asymétrique (dont nous avons déjà fait mention au tout début de ce chapitre), invention qui suscita l'émergence d'une communauté universitaire et industrielle dynamique.

Remarque: Contrairement à ce que l'on pourrait croire, la cryptographie à clé publique n'a pas relégué la cryptographie à clé secrète aux oubliettes, bien au contraire: ces deux types de cryptographie s'utilisent le plus souvent conjointement dans des cryptosystèmes hybrides où l'authentification des clés publiées est assurée par une "**autorité de certification**".

Avant d'exposer dans le détail le protocole de Diffie-Hellman, rappelons que le protocole d'échange des "clés secrètes" n'était fiable à l'époque (et ne l'est toujours pas aujourd'hui) puisqu'il voyage/transite entre les interlocuteurs, l'élément permettant de crypter et donc décrypter les messages. De plus, même si rien qu'une seule clé venait à voyager, toute personne ayant une puissance de calculs suffisante pourrait briser le code. D'où la nécessité qu'il y avait de changer (malheur de plus!) périodiquement les clés (cryptopériode). Deux solutions s'offrent alors:

S1. Ne pas échanger de clé (c'est possible mais c'est long comme nous allons le voir dans la figure ci-dessous)

S2. Échanger une clé secrète utilisant une fonction mathématique non inversible ou très difficilement inversible (c'est le protocole de Diffie-Hellman que nous verrons également dans une figure plus bas).

Voyons en quoi consiste la première solution et son désavantage flagrant:

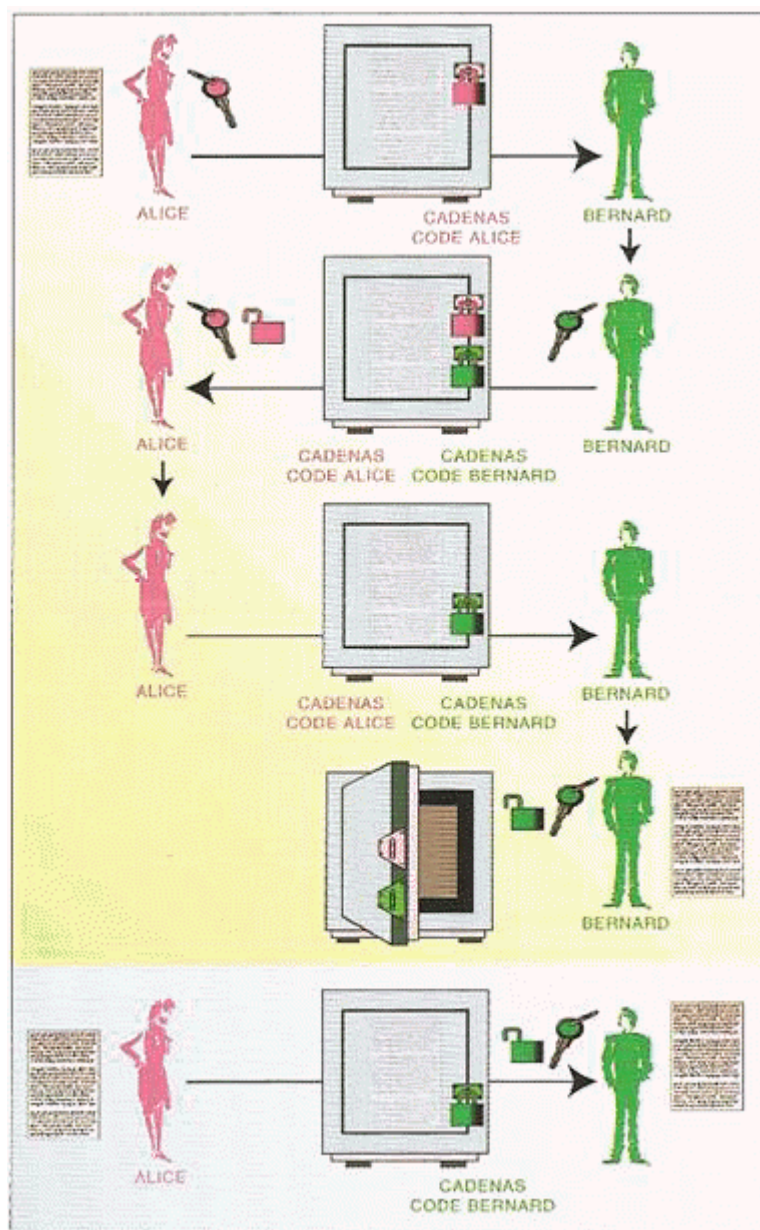


Figure: 61.5 - Principe du chiffrement à clé publique

Explication: Alice et Bernard veulent transmettre un message sur une ligne non sécurisée et sans

échanger de clé. Pour cela, Alice met sa lettre dans un coffre qu'elle ferme avec sa clé et l'envoie à Bernard. Ce dernier renvoie le coffre à Alice où il a ajouté son propre cadenas qu'il a fermé avec sa propre clé. Quand Alice reçoit le coffre, elle ôte son cadenas et renvoie à Bernard un coffre qui ne comprend plus que le cadenas de Bernard fermé avec la clé de Bernard. Celui-ci n'a donc plus qu'à ouvrir le coffre pour lire la lettre. Cette opération est sûre et ne nécessite pas d'échange de clé. En revanche, elle requiert plusieurs trajets (l'ensemble est représenté par les 4 premières transactions de la figure ci-dessus).

Le principe de la clé publique doit autoriser des échanges sécurisés, sans clé secrète, en un seul trajet. Bernard distribue largement des exemplaires de son cadenas public. Alice s'en procure un, mais n'importe qui pourrait faire de même. Alice place le message dans le coffre et le ferme avec le cadenas code de Bernard, puis elle lui envoie le coffre (représenté par la cinquième transaction de la figure ci-dessus). En recevant le coffre, Bernard peut ouvrir le coffre, puisque lui seul détient la clé qui ouvre ce cadenas. Le transfert est sûr en un seul voyage. En cryptographie, la clé publique équivaut au cadenas code, qui est disponible par exemple dans des annuaires, tandis que la clé qui ouvre ce cadenas est la clé privée, détenue uniquement par leur propriétaire et qui n'est jamais divulguée. Les clés privée et publique (le "**trousseau de clé**" comme on dit...) sont construites à partir d'une fonction mathématique supposée "à sens unique".

Voyons donc maintenant la deuxième solution faisant usage de clé publique selon le protocole de Diffie-Hellman:

3.1. PROTOCOLE DE DIFFIE-HELLMAN

Comme son nom l'indique, une fonction à sens unique donne un résultat facilement, mais l'opération inverse est très difficile. Trouver de telles fonctions dans le monde mathématique semblait fort ardu aux mathématiciens. Comment imaginer une fonction qui soit à sens unique pour tout le monde, excepté pour son créateur qui peut l'inverser grâce à la connaissance d'une information particulière. Ainsi, W. Diffie et M. Hellman ont été les premiers à proposer publiquement une fonction à sens unique pour résoudre le problème de la mise en accord sur un secret commun. L'idée de base consiste à calculer des valeurs du type:

$$\alpha^a \bmod(p) \quad (61.13)$$

où α et a sont imposés comme étant des entiers et p un nombre premier.

Les mathématiciens appelant ce genre d'opérations une "**exponentiation modulaire**" ou "**exponentielle discrète**" et il est d'usage de noter le corps fini des entiers modulo p (où p est un nombre premier) par $\mathbb{GF}(p)$ en l'honneur d'Évariste Galois.

Pour expliciter un tel calcul (pour rappel de ce qui a été vu dans le chapitre de Théorie Des Nombres...), nous élevons un nombre α à la puissance a , puis nous divisons le résultat par un grand nombre premier p et nous conservons finalement le reste de cette division (opération modulo p).

L'opération inverse est un problème redoutable: même si nous connaissons les valeurs numériques de α , de p et de $\alpha^a \bmod(p)$, il est extrêmement difficile en pratique de retrouver le bon nombre a . Les fonctions à sens unique telles que celle ci-dessus provenant de l'arithmétique modulaire se comportent de manière très irrégulière comme l'atteste le tableau avec l'exemple particulier ci-dessous:

a	$\alpha^a = 3^a$	$\alpha^a \bmod(p) = 3^a \bmod(7)$
0	1	1
1	3	3
2	9	2
3	27	6
4	81	4
5	243	5
6	729	1
7	2187	3
8	6561	2

Tableau: 61.3 - Exemple d'application de l'exponentiation modulaire

Donc même s'il est facile de calculer une exponentielle discrète, il est presque impossible de retrouver le nombre de départ a à partir du résultat et ce particulièrement quand cette fonction modulaire s'applique à des nombres premiers p très grands.

La sécurité de ce protocole est calculatoire. Elle se fonde sur l'hypothèse qu'avec une puissance de calcul et un temps limités, un adversaire ne peut inverser la fonction exponentielle modulaire (en faisant usage des propriétés des logarithmes avec les fonctions exponentielles comme nous l'avons vu dans le chapitre d'analyse fonctionnelle) et donc ne peut trouver le secret a à partir des éléments échangés. Cette difficulté calculatoire est due au fait que le temps de calcul nécessaire à l'inversion d'une fonction à sens unique n'a pas une complexité algorithmique (cf. chapitre de Méthodes Numériques) polynomiale mais exponentielle avec p .

Voyons un exemple schématique:

ALICE	Publique (Internet)	BERNARD
On choisit un nombre arbitraire commun: $p = 419$ et un nombre aléatoire commun inférieur à p : $\alpha = 7$. On suppose que ces deux valeurs sont secrètes.		
Alice choisit un nombre aléatoire secret: $a = 178$		Bernard choisit un nombre aléatoire secret: $b = 344$
Avec le nombre a Alice génère l'élément public: $k_a = \alpha^a \bmod(p) = 181$		Avec le nombre b Bernard génère l'élément public: $k_b = \alpha^b \bmod(p) = 351$
Envoi du résultat à Bernard: $k_a = 181$	$\longrightarrow$	$k_a = 181$
$k_b = 351$	$\longleftarrow$	Envoi du résultat à Alice: $k_b = 351$
Le secret partagé est alors: $K = (k_b)^a \bmod(p) = 493$		Le secret partagé est alors: $K = (k_a)^b \bmod(p) = 493$
$\longleftarrow$ Les échanges sont alors chiffrés avec la clé secrète K $\longrightarrow$		

Tableau: 61.4 - Exemple d'échange de clé entre Alice et Bernard

Alice et Bernard ont calculé le même secret commun: 493. On se sert de 493 pour chiffrer les données échangées (dans la pratique, on utilise des nombres beaucoup plus grands). L'espion n'est supposé pouvoir intervenir qu'après l'échange du choix commun de p et α .

Rappel: La clé K est obtenue par le fait que l'opération puissance est compatible avec la relation d'équivalence modulo p (cf. chapitre de Théorie Des Nombres) telle que:

$$x \bmod(y) = z \Leftrightarrow (x \bmod(y))^w = (x^w \bmod(y)) = z \quad (61.14)$$

Exemple:

$5 \bmod(3) = 2$, alors qu'avec $w = 2$ nous avons $5^2 \bmod(3^2) \neq 2^2$ alors que $5^2 \bmod(3) = 2^2$.

Ainsi, puisque $x < p$, le second modulo n'a plus de sens donc nous pouvons écrire:

$$(k_b)^a \bmod(p) = (\alpha^b \bmod(p))^a \bmod(p) = (\alpha^b \bmod(p))^a = \alpha^{b+a} \bmod(p) \quad (61.15)$$

de même:

$$(k_a)^b \bmod(p) = (\alpha^a \bmod(p))^b \bmod(p) = (\alpha^a \bmod(p))^b = \alpha^{a+b} \bmod(p) \quad (61.16)$$

et donc:

$$(k_b)^a \bmod(p) = (k_a)^b \bmod(p) = K \quad (61.17)$$

Malgré ces précautions, des experts ont établi au début du 21ème siècle un record en utilisant un nouvel algorithme, ils ont réussi à inverser la fonction exponentielle modulaire pour un nombre p de 120 chiffres (environ 400 bits), à l'aide d'un ordinateur intégrant quatre processeurs de 525 MHz. Ce record montre que la sécurité du protocole dépend grandement des progrès constants réalisés dans le domaine de la complexité algorithmique.

L'astucieux schéma de Diffie-Hellman reste un schéma de principe. Son principal inconvénient est qu'il ne permet pas d'assurer les services de sécurité classiques: authentification des deux intervenants, contrôle de l'intégrité de la clé et anti-rejeu (vérification qu'une information déjà transmise ne le soit pas à nouveau). Il s'ensuit qu'un attaquant peut, par exemple, usurper l'identité d'Alice en remplaçant l'élément public d'Alice par son propre élément public. Pour pallier à cet inconvénient, des versions sécurisées de ce protocole générique ont été publiées, par exemple un protocole nommé "**STS**" (Station To Station) qui utilise notamment la signature électronique pour assurer l'authentification des intervenants (voir plus loin). Cette stratégie est à la base de la connexion Internet sécurisée (IPSec).

Le protocole de Diffie-Hellman a ouvert la voie à toute une série d'algorithmes, celui du chiffrement à clé publique étant le premier. L'idée était de rompre la symétrie du chiffrement et du déchiffrement en utilisant les fonctions à sens unique.

3.2. SYSTÈME R.S.A.

Curieusement, le système de chiffrement R.S.A. le premier apparu dans la littérature est conceptuellement assez éloigné du protocole de Diffie-Hellman: il n'utilise pas l'exponentielle discrète, mais la factorisation des grands nombres. Ce système à clé publique a été inventé en 1977 par Ron Rivest, Adi Shamir et Leonard Adleman. Vite devenu un standard international, la technique R.S.A. a été commercialisée par plus de 400 entreprises et nous estimons que plus de 400 millions de logiciels l'utilisent. Elle est implémentée dans les navigateurs Web, comme Netscape Navigator, Microsoft Internet Explorer ou encore dans certaines cartes à puce bancaires, comme les cartes VISA.

Le système RSA est fondé sur la difficulté de factoriser des grands nombres et la fonction à sens unique utilisée est une fonction "puissance". Le protocole de chiffrement R.S.A. se décompose en trois phases:

1. Création des clés (publiques et privées)
2. Chiffrement à l'aide de la clé publique du destinataire
3. Déchiffrement à l'aide de la clé privée

Ce concept repose sur un théorème fameux appelé "théorème d'Euler" (rien à voir avec le théorème du même nom vu dans les chapitres de Théorie Des Graphes ou Formes Géométriques). Voyons de quoi il s'agit (attention c'est relativement long!).

3.2.1. THÉORÈME D'EULER

Avant de voir en quoi consiste le théorème d'Euler, il nous faut définir deux éléments qui y sont inclus. Outre le concept de congruence que nous avons déjà étudié dans le chapitre de Théorie des Nombres, il

reste une fonction spéciale appelée "**indicatrice d'Euler**" et définie en toute généralité par:

$$\phi(m) = \sum_{\substack{k=1 \\ (k,m)=1}}^m 1 \quad (61.18)$$

Autrement dit, la fonction ϕ du nombre entier m a pour résultat un nombre n strictement inférieur à m , donné par le nombre d'éléments compris entre 1 et m dont le p.g.c.d. (cf. chapitre de Théorie Des Ensembles) avec m est 1. Nous avons déjà donné un exemple pratique de l'utilité de cette fonction indicatrice dans le chapitre de Théorie Des Nombres dans le cadre des systèmes réduits de résidus et qui sont au centre de la démonstration du théorème d'Euler.

Ce qui peut encore se formuler sous la forme suivante: l'indicateur ϕ du nombre entier m est défini comme le nombre d'entiers positifs inférieurs ou égaux à m et premiers avec m .

Cette fonction a donc la propriété remarquable de compter le nombre d'entiers positifs plus petits que m et "relativement premiers" (c'est-à-dire qui ont le p.g.c.d. égal à 1) avec m .

Voici quelques valeurs de 0 à 19:

$\phi(m)$	0	1	2	3	4	5	6	7	8	9
0+		1	1	2	2	4	2	6	4	6
10+	4	10	4	12	6	8	8	16	6	18

Tableau: 61.5 - Valeurs de l'indicateur d'Euler

P1. Nous remarquons la propriété (triviale) de cette fonction lorsque nous notons un nombre premier (se rappeler que 1 n'est pas un nombre premier!) quelconque par la lettre p alors:

$$\phi(p) = p - 1 \quad (61.19)$$

Remarque: Cette fonction se trouve parfois dans la littérature sous la dénomination "**indicateur d'Euler**" au lieu de "**fonction phi d'Euler**".

P2. L'indicateur d'Euler peut s'écrire aussi sous la forme suivante si p et q sont premiers (il s'agit du cadenas du système R.S.A qui est plus compliqué que la simple multiplication de p et q):

$$\phi(n) = \phi(pq) = \phi(p)\phi(q) = (p-1)(q-1) \quad (61.20)$$

cette dernière relation peut se vérifier aisément (sans démonstration) en prenant quelques valeurs du tableau précédent.

Ceci fait, soit $(a, m) = 1$ (le p.g.c.d. de a et m), le "**théorème d'Euler**" dit que si m est un entier naturel et a est relativement premier avec m alors nous avons:

$$a^{\phi(m)} \equiv 1 \pmod{m} \quad (61.21)$$

dans laquelle nous voyons apparaître l'indicateur d'Euler définie juste plus haut. C'est une relation assez surprenante. Voyons si elle marche avec 7 et 2 qui sont premiers entre eux:

$$2^{\phi(7)} \equiv 1 \pmod{7} \Rightarrow 2^{\phi(7)} = 2^6 = 64 = 9 \cdot 7 + 1 \quad (61.22)$$

le reste étant donc bien 1 quand on fait 64 modulo 7.

Démonstration:

Rappelons d'abord (cf. chapitre de Théorie Des Nombres) qu'un système réduit de résidus modulo m est un ensemble d'entiers r_i qui satisfont les trois propriétés:

P1. $(r_i, m) = 1$

P2. r_i n'est pas congru r_j modulo m lorsque $i \neq j$

P3. Chaque entier x relativement premier avec m est congru à un certain r_i modulo m .

Par exemple, l'ensemble $\{1,5\}$ est un système réduit de résidus modulo 6 ou autre exemple; $\{1,2,3,4,5,6\}$ est un système réduit de résidus modulo 7. Nous vérifions également pour le premier exemple, que 1 n'est pas congru 5 modulo 6 (effectivement, 6 ne divise pas (5-1)) et que 5 qui est relativement premier à 6 est congru à lui-même.

Pour le deuxième, exemple, nous remarquons que le cardinal de l'ensemble de résidus correspond à la valeur de l'indicateur d'Euler pour le nombre 7.

Ainsi, soit $\{r_1, r_2, \dots, r_{\phi(m)}\}$ un système réduit de résidus modulo m . Nous avons besoin pour la démonstration du théorème d'Euler, de démontrer au préalable que $\{ar_1, ar_2, \dots, ar_{\phi(m)}\}$ est aussi un système réduit de résidus modulo m .

Remarque: Comme nous en avons déjà fait mention dans l'exemple précédent, vous pouvez observer que le cardinal de l'ensemble des résidus correspond, pour un modulo m premier donné, au résultat défini par la propriété P1 de la fonction ϕ d'Euler. Cette propriété n'est à ce jour qu'une "conjecture", c'est-à-dire une supposition fondée sur des probabilités (car non démontrée paraît-il!).

Pour cela, rappelons-nous que par la propriété d'un système réduit:

$$(r_i, m) = 1$$

et par hypothèse:

$$(a, m) = 1 \quad (61.23)$$

alors nous voulons démontrer que:

$$(ar_i, m) = 1 \quad (61.24)$$

est aussi satisfait.

Posons pour cela $d = 1$ (par tradition...). Nous avons alors puisque $(r_i, m) = d$ que $d \mid r_i$ et $d \mid m$ et

identiquement pour $(a, m) = d$ que $d \mid a$ et que $d \mid m$. Maintenant si d divise bien a ou r_i dans ce cas nous avons que $d \mid r_i(a)$ ou autrement $d \mid a(r_i)$. Donc $d \mid a \cdot r_i$ et $d \mid m$ ce qui nous permet d'écrire:

$$(ar_i, m) = (r_i a, m) = 1 \quad (61.25)$$

Revenons à notre théorème d'Euler... si vous suivez toujours... Nous venons de démontrer qu'il y a bijection entre les deux ensembles de résidus. C'est-à-dire que pour chaque résidu r_i du système réduit modulo m , nous aurons un résidu ar_i du système réduit modulo m selon la propriété fondamentale de la congruence qui rappelons-le dit que: nous pouvons multiplier les deux membres d'une congruence par un même nombre entier et il restera congru modulo m et modulo m multiplié par ce nombre entier.

Exemple:

Prenons:

$$35 \equiv 5 \pmod{6} \quad (61.26)$$

effectivement:

$$6 \mid (35 - 5 = 30) \quad (61.27)$$

car le reste de la division de 30 par 6 est bien nul. Si nous prenons par exemple:

$$2 \cdot 35 \equiv 2 \cdot 5 \pmod{2 \cdot 6} \quad (61.28)$$

alors nous avons également:

$$12 \mid (70 - 10 = 60) \quad (61.29)$$

et le reste est aussi nul...

Petit rappel sur la bijection (cf. chapitre de Théorie Des Ensembles): nous disons que nous avons une bijection, si à chaque élément d'un ensemble de départ correspond un et un seul élément dans l'ensemble d'arrivée (s'il y avait pour chaque homme sur Terre une femme - à proportions égales donc - il y aurait bijection entre l'ensemble des Hommes et des Femmes).

Bref, comme il y a bijection entre les deux ensembles de résidus, nous pouvons écrire:

$$\prod_{i=1}^{\phi(m)} ar_i \equiv \prod_{i=1}^{\phi(m)} r_i \pmod{m} \quad (61.30)$$

Exemple:

L'ensemble $\{1, 5\}$ est un système réduit de résidus modulo 6 comme nous l'avons déjà vu. Nous avons donc:

$$\begin{aligned} (r_1, m) &= (1, 6) = 1 \\ (r_2, m) &= (5, 6) = 1 \end{aligned} \quad (61.31)$$

Nous avons alors:

$$\prod_{i=1}^{\phi(m)} r_i \bmod(m) = \prod_{i=1}^{\phi(6)} r_i \bmod(6) = \prod_{i=1}^2 r_i \bmod(6) = 1 \cdot 5 = 5 \quad (61.32)$$

Si nous prenons un a tel que $(a, m) = 1$, par exemple $a = 7$ car effectivement $(7, 6) = 1$, alors:

$$7 \cdot 5 \equiv 5 \bmod(6) \quad (61.33)$$

car $6 \mid (35 - 5 = 30)$. Effectivement 6 divise bien 30 avec un reste nul.

Donc revenons à notre bijection qui peut s'écrire par les règles élémentaires de l'algèbre:

$$a^{\phi(m)} \prod_{i=1}^{\phi(m)} r_i \equiv \prod_{i=1}^{\phi(m)} r_i \bmod(m) \quad (61.34)$$

Puisque:

$$\left(m, \prod_{i=1}^{\phi(m)} r_i \right) = 1 \quad (61.35)$$

(vous pouvez vérifier mais cela découle de la définition même d'un ensemble de résidus!), nous sommes bien obligés de conclure que:

$$a^{\phi(m)} \equiv 1 \bmod(m) \quad (61.36)$$

et de toute façon, même si cela ne vous semble pas évident, vous n'avez qu'à multiplier chacun des membres de l'égalité de la congruence par:

$$\left(\prod_{i=1}^{\phi(m)} r_i \right)^{-1} \quad (61.37)$$

comme nous l'autorise une des propriétés intrinsèque de la congruence démontrée précédemment.

□C.Q.F.D.

Cet interlude théorie étant fait, considérons un nombre N dont nous souhaitons décider s'il est premier ou non.

Nous savons d'après le théorème d'Euler et de la propriété P1 de l'indicateur d'Euler, que si N est un nombre premier et si $a \in \mathbb{N}$, où $a < N$, alors:

$$\boxed{a^{N-1} \equiv 1 \bmod(N)} \quad (61.38)$$

qui est appelé le "**petit théorème de Fermat**".

Remarque: Cette relation découle des propriétés que nous avons exposées lors de notre démonstration du théorème d'Euler:

$$a^{\phi(m)} \equiv 1 \pmod{m} \quad (61.39)$$

et de la propriété P1 de la fonction $\phi(m)$ pour un nombre p premier:

$$\phi(p) = p - 1 \quad (61.40)$$

Le petit théorème de Fermat est cependant également valable pour quelques nombres N qui ne sont pas premiers. Mais les nombres qui vérifient ça sans être premiers sont rares, et du coup ça vaut la peine de déclencher un algorithme plus sophistiqué pour savoir si N est réellement premier ou non (disons que dans ce cas, N est un bon candidat à la primalité et est alors appelé "**nombre pseudo-premier**"). Pour tester si le nombre, le cas échéant N non-premier est "suffisamment premier", on essaie avec un algorithme de tester le petit théorème de Fermat pour un nombre maximal de $a \in \mathbb{N}$ avec $a < N$.

D'après la propriété de la congruence (voir plus haut), nous avons également:

$$a \cdot a^{N-1} \equiv a \cdot 1 \pmod{N} \Rightarrow a^N \equiv a \pmod{N} \quad (61.41)$$

Nous pouvons appliquer ce dernier théorème sur un nombre N à propos duquel nous aimerions savoir au mieux s'il est premier ou non.

Il existe une grande quantité d'autres méthodes non optimales pour déterminer si N est premier; dont les essais préliminaires de division par 2, 3, 5, 7, 11 et des nombres premiers petits jusqu'à $p \leq \sqrt{N}$ selon la méthode du crible d'Ératosthène qui est la plus connue dans les petites écoles.

Remarque: En fait, avec l'aide d'un ordinateur assez puissant, nous pouvons décider si un nombre naturel de l'ordre de 10^{300} (10 suivi 300 zéros) est premier ou non en l'espace de quelques minutes voir secondes. Ce qu'il est important de savoir, c'est que, étant donné un nombre naturel N , on peut décider en relativement peu de temps s'il est premier ou non, sans pour autant connaître ses facteurs premiers.

Cependant, selon le théorème fondamental de l'arithmétique nous avons que:

Tout nombre naturel $n > 1$ peut s'écrire comme un produit de nombres premiers, et cette représentation est unique, à part l'ordre dans lequel les facteurs premiers sont disposés.

Démonstration:

Si n est premier, alors la preuve est terminée. Supposons que n n'est pas premier et considérons l'ensemble:

$$D = \{d \mid d \mid n \text{ et } 1 < d < n\} \quad (61.42)$$

Alors, $D \subset \mathbb{N}$ et, puisque n est supposé composé, nous avons que $D \neq \emptyset$, D'après le principe du

bon ordre (tout ensemble non vide contient un plus petit élément), D possède un plus petit élément p_1 qui est premier, sans quoi le choix minimal de p_1 serait contredit. Nous pouvons donc écrire $n = p_1 n_1$. Si n_1 est premier, alors la preuve est terminée. Si n_1 est composé, alors nous répétons le même argument que précédemment et nous en déduisons l'existence d'un nombre premier p_2 et d'un entier $n_2 < n_1$ tels que $n = p_1 p_2 n_2$. En poursuivant ainsi, nous arrivons forcément à la conclusion que n_k sera premier.

□C.Q.F.D.

Donc finalement nous avons bien démontré qu'un nombre quelconque est décomposable en facteurs de nombres premiers à l'aide du principe du bon ordre. Il existe dans l'ensemble des nombres naturels, certains qui peuvent s'exprimer, ou qui s'expriment tout court, uniquement par 2 facteurs premiers notés traditionnellement p et q . Ce sont ces nombres que nous utilisons en cryptographie à clé publique selon le protocole R.S.A.

Remarque: Nous ne connaissons pas à ce jour de loi qui permette de calculer facilement et rapidement le n -ième facteur premier (p_n) d'un nombre. En fait, même avec les ordinateurs les plus puissants d'aujourd'hui, il faudrait plusieurs années pour arriver à trouver les deux facteurs premiers p et q d'un nombre $N = pq$ où p et q sont de l'ordre de 10^{100} chacun. Et il semble peu probable que l'on découvre dans un avenir proche un algorithme assez efficace pour améliorer de façon appréciable ce temps de calcul. Notons qu'il est possible de déterminer en moins de 5 minutes (en 2002) si un nombre de 200 chiffres est premier ou non. Cependant, pour factoriser un nombre de 200 chiffres en deux nombres premiers, il faudrait au moins 100 ans. Chose merveilleuse: les théories qui permettent ces exploits sont très profondes et ont été élaborées en partie il y a longtemps dans un cadre très différent.

Le fait qu'il soit beaucoup plus difficile de trouver les facteurs premiers d'un nombre N que de découvrir si N est premier ou composé, est précisément ce qui a permis d'élaborer cette méthode très ingénieuse de codage et décodage de messages selon le protocole R.S.A.

Exemple:

Considérons maintenant un groupe d'individus qui se transmettent régulièrement des messages par courrier électronique et pour lequel il est important que les messages ne soient connus que de l'expéditeur et du destinataire. Alors, le membre du groupe (ici Alice) qui souhaite recevoir des informations cryptées, se trouve deux nombres premiers p et q très grands de l'ordre de 10^{100} . Pour trouver de si grands nombres premiers, nous choisissons au hasard un nombre de 100 chiffres et nous vérifions par un des algorithmes connus s'il est premier ou non et nous répétons l'expérience jusqu'à ce que nous obtenions ainsi un nombre premier. Une fois ceci fait avec ces deux nombres premiers, nous calculons l'expression:

$$n = pq \quad (61.43)$$

appelée "**modulus**".

Ensuite, Alice qui souhaite recevoir les informations cryptées (qui est la seule en possession du nombre n pour l'instant) choisit un entier positif a tel (p.g.c.d.) que:

$$(\alpha, \phi(n)) = 1 \quad (61.44)$$

Donc a (souvent noté e dans la littérature) est un entier premier avec $\phi(n)$ appelé parfois le "générateur".

Et comme:

$$\phi(n) = \phi(pq) = \phi(p)\phi(q) = (p-1)(q-1) \quad (61.45)$$

par conséquent, par essais répétés, il est facile de trouver un tel nombre a . Le membre principal, trouve donc un n et un a pour son contact, qu'il lui envoie sans aucune protection. Car, même dans le cas où il y aurait d'éventuels intercepteurs à l'affût sur la ligne, il leur sera extrêmement difficile de retrouver les facteurs premiers de n dans un laps de temps relativement bref.

Supposons qu'une agence secrète souhaite recevoir un message d'un de ses agents.

L'agence envoie la clé publique, définie donc par le couple:

$$\{a, n\} \quad (61.46)$$

à l'agent.

L'agent reçoit la clé publique et souhaite envoyer le message "déclencher l'opération rouge". Pour ce faire, l'agent transforme d'abord le message en chiffres en utilisant la convention que chaque lettre est remplacée par sa position correspondante dans l'alphabet en commençant à compter à partir de 01 (le caractère "espace" sera chiffré "27").

Ainsi le message clair noté M par la suite devient:

$$M = \underbrace{04050312051403080518}_{\text{DECLENCHER}} \underbrace{27}_{\text{L}} \underbrace{151605180120091514}_{\text{OPERATION}} \underbrace{27}_{\text{L}} \underbrace{1815210705}_{\text{ROUGE}} \quad (61.47)$$

Point technique: il faut que M et n n'aient pas de diviseur commun autre que 1 (sinon quoi, un éventuel espion pourrait réduire le problème de deux très grands nombres difficilement manipulables à celui de plus petits nombres, plus facilement manipulables). Sinon quoi, on ajoute à la fin de M des chiffres sans valeur, comme 01 (par exemple), pour finalement avoir M et n sans diviseur commun autre que 1. On peut aussi briser M en morceaux M_i dont le nombre de chiffres n'excède pas 99 (rappelez-vous que nous avons fixé une limite inférieure d'une puissance de 100 pour p et q et qu'il suffirait donc qu'un des deux nombres premiers soit 1 et l'autre exactement un nombre avec un exposant 100 pour être à limite du nombre n comportant alors au pire 100 chiffres), auquel cas on aura toujours:

$$(M_i, p) = (M_i, q) = (M_i, n) = 1 \quad (61.48)$$

On découpe M en morceaux, chacun étant plus petit que n :

$$\begin{array}{cccccccccccccccc} 0405 & 0312 & 0514 & 0308 & 0518 & 27 & 12 & 27 & 1516 & 0518 & 0120 & 0915 & 1427 & 1815 & 2107 & 05 \\ M_1 & M_2 & M_3 & M_4 & M_5 & & & & M_6 & M_7 & M_8 & M_9 & M_{10} & M_{11} & M_{12} & \\ M = & \underbrace{04050312051403080518}_{\text{DECLENCHER}} & \underbrace{27}_{\text{L}} & \underbrace{151605180120091514}_{\text{OPERATION}} & \underbrace{27}_{\text{L}} & \underbrace{1815210705}_{\text{ROUGE}} \end{array} \quad (61.49)$$

et on travaille successivement avec chaque morceau $M_1, M_2, \dots, M_{12}$ du message.

On considère la puissance a de M_1 , c'est-à-dire M_1^a . On remplace M_1 par le nombre $\bar{M}_1$, qui est le reste de la division par n du nombre M_1^a . On procède de même pour tous les autres morceaux M_i tels que:

$$\bar{M}_i = M_i^a \bmod(n) \quad (61.50)$$

L'agent envoie alors le message codé à l'agence:

$$\bar{M} = \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|} \hline \bar{M}_1 & \bar{M}_2 & \bar{M}_3 & \bar{M}_4 & \bar{M}_5 & \bar{M}_6 & \bar{M}_7 & \bar{M}_8 & \bar{M}_9 & \bar{M}_{10} & \bar{M}_{11} & \bar{M}_{12} \\ \hline \end{array} \quad (61.51)$$

Un intercepteur (non-mathématicien) du message codé et de la clé publique, connaissant l'algorithme de chiffrement, aurait donc à résoudre le problème d'une équation à deux inconnues (équation obtenue simplement à partir de l'expression mathématique du chiffrement):

$$M_i = \sqrt[a]{n \cdot q + \bar{M}_i} \quad (61.52)$$

Problème évidemment indéterminé !

Pour voir comment l'agence décrypte le message, on a besoin d'un outil mathématique supplémentaire.

Rappelons que l'agence choisit a de telle sorte que $(a, \phi(n)) = 1$, ce qui implique, d'après le théorème de Bézout (cf. chapitre de Théorie Des Nombres), que si a et $\phi(n)$ sont premiers entre eux (que leur plus grand diviseur commun est 1) il existe des entiers x et y tels que (on peut supposer que $x > 0$, auquel cas $y < 0$):

$$ax + \phi(n)y = 1 \quad (61.53)$$

ou autrement écrit:

$$ax \equiv 1 \bmod(\phi(n)) \quad (61.54)$$

C'est ainsi que nous allons déterminer la valeur de x (il faut utiliser des algorithmes pour trouver la solution x à cette équation).

Ce qui signifie:

1. Que si a est premier avec $\phi(n)$ alors par les propriétés de la congruence il l'est également avec $p-1$ et $q-1$.
2. Que a est inversible modulo $\phi(n)$

Effectivement, car:

$$ax - 1 = \phi(n)y \quad (61.55)$$

et d'après la définition de la congruence ($m \mid (a - b)$) nous avons effectivement:

$$ax \equiv 1 \pmod{\phi(n)} \quad (61.56)$$

puisque $\phi(n)$ divise le membre de droite de $ax - 1 = \phi(n)y$ et donc de par l'égalité, le membre de gauche.

Seule l'agence, qui reçoit le message, peut facilement calculer le nombre $x \leq a$ utilisé ci-dessus. Car pour cela, il faut pouvoir calculer la valeur de la fonction ϕ indicatrice d'Euler et donc connaître p et q .

Si M_i est le message (valeur numérique) d'origine et $\bar{M}_i$ est le message (valeur numérique) codé reçu, alors nous avons la relation suivante:

$$M_i^a \equiv \bar{M}_i \pmod{n} \quad (61.57)$$

Ce qui est complètement logique puisque la différence $M_i^a - \bar{M}_i$, où rappelons-le, $\bar{M}_i$ est le reste de la division de M_i^a par n , ne peut donc qu'être divisible par n .

L'agence reçoit donc le message codé $\bar{M}$ et élève à la puissance x les nombres $\bar{M}_i$ et obtient ainsi le message initial.

En effet, elle va appliquer pour chaque $\bar{M}_i$ la propriété mathématique suivante de la congruence:

$$\bar{M}_i^x \equiv M_i^{ax} \equiv M_i^{1-\phi(n)y} \equiv M_i \pmod{n} \quad (61.58)$$

La clé privée (permettant de décrypter le message et qui peut être connue facilement uniquement par l'agence) est donc définie par le couple:

$$\text{🔑} \{x, n\} \quad (61.59)$$

Explications:

Nous avons déjà montré que:

$$M_i^a \equiv \bar{M}_i \pmod{n} \quad (61.60)$$

et selon la propriété de symétrie de la congruence (cf. chapitre de Théorie Des Nombres), nous pouvons écrire:

$$M_i^a \equiv \bar{M}_i \pmod{n} \Leftrightarrow \bar{M}_i \equiv M_i^a \pmod{n} \quad (61.61)$$

Effectivement:

$$\bar{M}_i^x \equiv M_i^{ax} \pmod{n} \quad (61.62)$$

selon la deuxième principale propriété de la congruence qui dit que l'on peut élever à une même puissance les deux membres d'une congruence. Ce qui nous donne aussi directement:

$$\bar{M}_i^{ax} \equiv M_i^{1-\phi(n)y} \pmod{n} \quad (61.63)$$

puisque nous avons démontré précédemment que:

$$ax + \phi(n)y = 1 \quad (61.64)$$

donc que:

$$ax = 1 - \phi(n)y \quad (61.65)$$

Reste à démontrer que:

$$M_i^{1-\phi(n)y} \equiv \bar{M}_i^x \pmod{n} \quad (61.66)$$

où l'on peut écrire $M_i^{1-\phi(n)y}$ sous la forme:

$$M_i \cdot (M_i^{\phi(n)})^{-y} \quad (61.67)$$

Or, rappelez-vous que nous avons démontré le théorème d'Euler:

$$M_i^{\phi(n)} \equiv 1 \pmod{n} \quad (61.68)$$

et qu'une des propriétés de la congruence nous donne le droit d'élever à une puissance quelconque les deux membres de la congruence tel que:

$$(M_i^{\phi(n)})^{-y} \equiv (1)^{-y} \pmod{n} \quad (61.69)$$

Mais comme 1 élevé à n'importe quelle puissance fait 1, nous avons:

$$(M_i^{\phi(n)})^{-y} \equiv 1 \pmod{n} \quad (61.70)$$

Cette dernière relation nous permet donc de vérifier que l'on peut s'autoriser à écrire:

$$M_i \cdot (M_i^{\phi(n)})^{-y} \equiv M_i \pmod{n} \quad (61.71)$$

Puisque les deux membres de gauche sont bien modulus n .

Donc si on reprend tout ça, l'agence reçoit un morceau $\bar{M}_i$ et l'élève par automatisme à la puissance x pour obtenir un nombre qui selon elle devrait être le M_i véritable. Pour en être sûr, elle applique la vérification imparable:

$$\bar{M}_i^x \equiv M_i \pmod{n} \quad (61.72)$$

Il est facile de voir que tout intercepteur ne peut décoder et en plus vérifier si le décodage est bien le bon, car pour cela il devrait connaître la valeur de x , laquelle à son tour dépend de $\phi(n)$, qu'il ne connaît pas non plus, parce qu'il ne connaît pas les facteurs premiers de n .

Il est d'usage de dire que le système RSA utilise les nombres p, q (secrets), n (publique), a (publique) et x (secret). Le tout se résumant au triplet n, a, x noté parfois dans la littérature n, e, d .

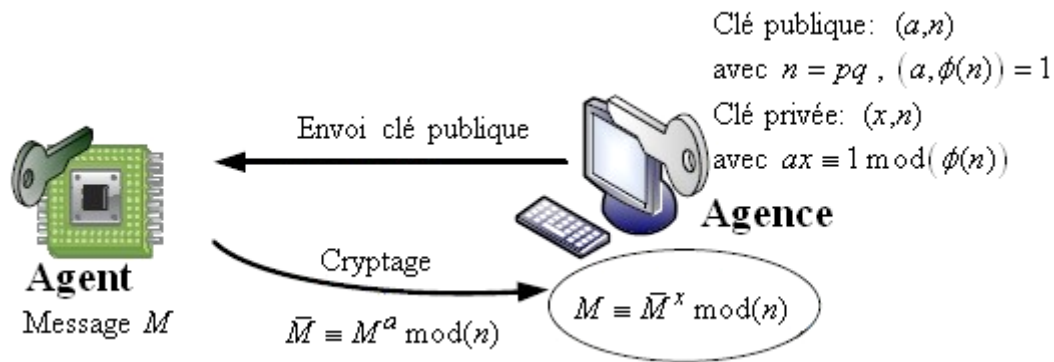


Figure: 61.6 - Principe du chiffrement à clé publique RSA

Et une petite application pratique avec Maple 4.00b:

```
> #initialisation du générateur aléatoire de Maple 4.00b
> randomize();
> #définition de la taille souhaitée du modulus (c'est un nombre pair)
> t:=30;
> #génération de deux entiers de t/2 bits
> x:=rand(2^(t/2-1)..2^(t/2))();
> y:=rand(2^(t/2-1)..2^(t/2))();
> #calcul des nombres premiers qui suivent
> p:=nextprime(x);
> q:=nextprime(y);
> #modulus public de la clé RSA
> n:=p*q;
> phi:=(p-1)*(q-1);
> #on choisit a empiriquement
> a:=65537;
> #on vérifie qu'il est premier avec phi
> igcd(a,phi);
> #on calcule l'inverse de a modulo phi
> x:=1/a mod phi;
> #on choisit un message comme étant 1234
> m:=1234;
> # on code
> c:=m&^a mod n;
> # on décode
> c&^x mod n;
```

Suite à la demande d'un lecteur voici un résumé littéral de ce que nous avons vu jusqu'à maintenant pour les premières étapes de l'algorithme ci-dessus:

1) Nous choisissons p et q premiers suffisamment grands:

$$p = 593, q = 613 \quad (61.73)$$

nous avons alors:

$$n = p \cdot q = 593 \cdot 613 = 363'509 \quad (61.74)$$

2) Nous calculons l'indicatrice d'Euler:

$$\phi(n) = (p-1) \cdot (q-1) = (593-1) \cdot (613-1) = 592 \cdot 612 = 362'304 \quad (61.75)$$

3) Nous choisissons le générateur a tel que:

$$(a, \phi(n)) = 1 \quad (61.76)$$

et pour cela nous prendrons $a = 5$. Le couple (n, a) est la clé publique (peut être distribuée à tout le monde pendant un temps déterminé).

4) Ensuite, nous calculons:

$$x = \frac{1 \bmod \phi(n)}{a} = \frac{362'304 + 1}{5} = 72'461 \quad (61.77)$$

Donc le couple (x, a) est la clé privée (à garder secrète).

Pour des raisons de sécurité, la cryptographie à clé publique est utilisée conjointement avec la cryptographie à clé secrète. Par exemple, au jour de l'écriture de ces lignes, le protocole SSL pour les pages Internet utilise le RSA pour échanger une clé secrète (système symétrique) et crypte ensuite les données grâce à un algorithme symétrique classique.

Signalons en terminant cette brève présentation du codage des messages, que le gouvernement américain surveille de très près les activités des mathématiciens qui travaillent sur la factorisation des grands nombres. En effet, si un de ceux-ci arrivait à trouver un algorithme permettant de factoriser en peu de temps un nombre de deux cents chiffres (supérieur à 524 bits non signé), cela mettrait en péril le caractère secret de plusieurs communications d'ordre militaire. Cette surveillance a d'ailleurs soulevé aux États-Unis un mouvement de protestation de la part des hommes de sciences, qui voient ainsi brimer leur liberté professionnelle (Notices of American Mathematical Society, janvier 1983).

Pour information technique, le logiciel PGP (Pretty Good Privacy) du MIT, utilise un système de chiffrement RSA.

4. FONCTIONS DE HACHAGE

Une fonction de hash (anglicisme) ou fonction de hachage est une fonction qui associe à un grand ensemble de données un ensemble beaucoup plus petit (de l'ordre de quelques centaines de bits) qui est caractéristique de l'ensemble de départ. Cette propriété fait qu'elle est très utilisée en informatique, en particulier pour accéder rapidement à des données grâce à des "**tables de hachage**". En effet, une fonction de hachage permet d'associer à une chaîne de caractères un entier particulier. Ainsi, si nous connaissons l'empreinte des chaînes de caractères stockées, nous pouvons rapidement vérifier si une chaîne se trouve ou non dans cette table (en $O(1)$ si la fonction de hachage est suffisamment bonne). Les fonctions de hachage sont aussi extrêmement utiles en cryptographie pour accélérer le cryptage.

Les 2 algorithmes de condensation les plus utilisés sont le "**SHA**" (Secure Hash Algorithm) qui calcule un résumé de 160 bits, et le MD5 (Message Digest 5 - Run Rivest 1992), qui calcule un résumé de 128

bits nommé "Message Digest".

4.1. FONCTION DE CONDENSATION MESSAGE DIGEST MD5

Cet algorithme est (était) surtout utilisé pour les signatures numériques (notion utilisée, lors de la validation de certificats d'authenticité comme nous le verrons plus loin).

Voici les différentes étapes de son fonctionnement:

Étape 1: Complétion

Le message est constitué de b bits. On complète le message par un 1, et suffisamment de 0 pour que le message étendu ait une longueur multiple de 512 bits. Après ce traitement initial, on manipule le texte d'entrée par blocs de 512 bits divisés en 16 sous-blocs $M[i]$ de 32 bits.

Étape 2: Initialisation

On définit les variables de chaînage de 32 bits A, B, C et D initialisées ainsi (les chiffres sont hexadécimaux):

$$A=01234567, B=89ABCDEF, C=FEDCBA98, D=76543210$$

On définit aussi 4 fonctions non linéaires F, G, H et I, qui prennent des arguments codés sur 32 bits, et renvoient une valeur sur 32 bits, les opérations se faisant bit à bit.

$$F(X,Y,Z) = (X \text{ AND } Y) \text{ OR } (\text{NOT}(X) \text{ AND } Z)$$

$$G(X,Y,Z) = (X \text{ AND } Z) \text{ OR } (Y \text{ AND } \text{NOT}(Z))$$

$$H(X,Y,Z) = X \text{ XOR } Y \text{ XOR } Z$$

$$I(X,Y,Z) = Y \text{ XOR } (X \text{ OR } \text{NOT}(Z))$$

Ce qu'il y a d'important avec ces 4 fonctions est que si les bits de leurs arguments X,Y et Z sont indépendants, les bits du résultat le sont aussi.

Étape 3: Calcul itératif

La boucle principale a 4 rondes qui utilisent chacune une fonction non linéaire différente (d'où le fait qu'il y en ait 4...). Chaque ronde consiste donc en 16 exécutions d'une opération (car 16 sous-blocs).

Chaque opération calcule une fonction non linéaire de trois des variables A, B, C et D, y ajoute un sous-bloc $M[i]$ du texte à chiffrer, une constante s prédéfinie (codée sur 32 bits) et effectue un décalage circulaire vers la gauche, d'un nombre variable n de bits. Voici l'exemple pour A:

$$- A = B + A + F(B,C,D) + M[i] + s \mid \text{décalé circulairement de } n \text{ vers la gauche}$$

$$- A = B + A + G(B,C,D) + M[i] + s \mid \text{décalé circulairement de } n \text{ vers la gauche}$$

$$- A = B + A + H(B,C,D) + M[i] + s \mid \text{décalé circulairement de } n \text{ vers la gauche}$$

$$- A = B + A + I(B,C,D) + M[i] + s \mid \text{décalé circulairement de } n \text{ vers la gauche}$$

Cette nouvelle valeur de A est ensuite sommée avec l'ancienne.

Étape 4: Ecriture du résumé

Le résumé sur 128 bits est obtenu en mettant bout à bout les 4 variables de chaînage A, B, C, D de 32 bits obtenues à la fin de l'itération.

La fonction MD5 n'est pas sûre et pas unique (deux entrées différentes peuvent donner la même signature: nous parlons alors de collision). Cependant, la fonction MD5 reste encore largement utilisée comme outil de vérification lors des téléchargements et l'utilisateur peut valider l'intégrité de la version téléchargée grâce à l'empreinte. Ceci peut se faire avec un programme comme *md5sum* pour MD5 et *sha1sum* pour SHA-1. (cf. chapitre de Codes Correcteurs).

Voici l'empreinte (appelée abusivement "signature") obtenue sur une phrase (que nous avons pris sans accents):

MD5("Wikipedia, l'encyclopedie libre et gratuite") = d6aa97d33d459ea3670056e737c99a3d

En modifiant un caractère, cette empreinte change radicalement:

MD5("Wikipedia, l'encyclopedie libre et gratuitE") = 5da8aa7126701c9840f99f8e9fa54976

Très concrètement, la vérification de l'empreinte ou somme de contrôle MD5 peut être réalisée de la façon suivante: lors du téléchargement d'un programme, nous notons la série de caractères indiquée sur la page de téléchargement. Quand ce téléchargement est terminé, nous lançons un des utilitaires susmentionné.

5. FONCTION DE CONDENSATION SECURE HASH ALGORITHM SHA-1

Le SHA-1 est (était) utilisé en concurrence du MD5 pour les signatures numériques (Digital Signature Algorithm) comme spécifié par le Digital Signature Standard (DSS). Pour un message de longueur inférieure à 2^{64} , le SHA-1 génère un condensé de 160 bits du message appelé "**hash**". À nouveau, à l'identique du MD5, une modification infime du message d'origine doit avoir une grosse répercussion sur le message condensé et il ne doit pas exister de Message Digest identique pour deux messages d'origine différente.

Comme pour le MD5, on travaille sur des messages dont la longueur est un multiple de 512 bits.

Étape 1: Complétion

Si le message n'a pas une longueur de 512 bits, on rajoute autant de 1 que nécessaire à la fin de ce dernier. Les derniers 64 bits du bloc de 512 bits sont utilisés pour définir la longueur d'origine du message. On transforme ensuite le bloc de 512 bits en sous-blocs $M[i]$ de 32 bits chacun exprimés en hexadécimal ($0 \leq i \leq 15$).

Étape 2: Initialisation

Comme pour le MD5, on définit cette fois 80 variables de chaînage de 32 bits $K[i]$ initialisées ainsi (les chiffres sont hexadécimaux):

$K[t] = 01234567 \mid 0 \leq t \leq 19$

$K[t] = 89ABCDEF \mid 20 \leq t \leq 39$

$K[t] = FEDCBA98 \mid 40 \leq t \leq 59$

$K[t] = 76543210 \mid 60 \leq t \leq 79$

On définit aussi 80 fonctions non linéaires $F[0], F[1], F[2], \dots, F[79]$ qui prennent des arguments codés sur 32 bits, et renvoient une valeur sur 32 bits, les opérations se faisant bit à bit.

$F[t](X,Y,Z) = (X \text{ AND } Y) \text{ OR } (\text{NOT}(X) \text{ AND } Z) \mid 0 \leq t \leq 19$

$F[t](X,Y,Z) = (X \text{ XOR } Y) \text{ XOR } D \mid 20 \leq t \leq 39$

$F[t](X,Y,Z) = (X \text{ AND } Y) \text{ OR } (X \text{ AND } Z) \text{ OR } (Y \text{ AND } Z) \mid 40 \leq t \leq 59$

$F[t](X,Y,Z) = X \text{ XOR } Y \text{ XOR } Z \mid 60 \leq t \leq 79$

Ce qu'il y a d'important avec ces 80 fonctions est que si les bits de leurs arguments X, Y et Z sont indépendants, les bits du résultat le sont aussi.

Étape 3: Calcul Itératif

L'itération utilise deux buffers, chacun consistant en l'utilisation de 5 variables de chaînage. Les variables de chaînage du premier buffer sont notées A, B, C, D, E . Le second paquet de 5 contient les variables de chaînage notées $H[0], H[1], H[2], H[3], H[4]$.

Par ailleurs, notons S^n le décalage circulaire de n bits vers la gauche

Voici l'algorithme SHA-1:

Pour $t = 16$ à 79 faire

$M[t] = S^1(M[t-16] \text{ XOR } M[t-15] \text{ XOR } M[t-14] \text{ XOR } M[t-13])$

Fin Pour

$A = H[0]; B = H[1]; C = H[2]; D = H[3]; E = H[4]$

Pour $t = 0$ à 79 faire

$TEMP = S^5(A) + F[t](B,C,D) + E + M[t] + K[t]$

$E = D; D = C; C = S^{30}(B); B = A; A = TEMP$

Fin Pour

$H[0] = H[0] + A; H[1] = H[1] + B; H[2] = H[2] + C; H[3] = H[3] + D; H[4] = H[4] + E$

Après l'exécution de cet algorithme, on obtient un message 160 bits (5×32) représentés par les 5 variables de chaînage $H[0], H[1], H[2], H[3], H[4]$.

6. CERTIFICATS D'AUTHENTIFICATION

Nous avons vu lors de la cryptographie à clé publique et à clé secrète, qu'il subsistait une faille dans le système de transmission des clés au début de la communication.

Ainsi dans les deux systèmes, la faille réside dans le fait que quelqu'un de malveillant puisse se substituer à l'interlocuteur réel et envoyer ainsi soit une fausse clé secrète, soit une fausse clé publique (en fonction des cas).

Ainsi, un certificat d'authenticité permet d'associer une clé à une entité (une personne, une machine, ...) afin d'en assurer la validité (l'association à la "vraie personne"). Le certificat est en quelque sorte la carte d'identité de la clé ou la "**signature numérique**", délivrée par un organisme appelé "**autorité de certification**".

La technologie faisant usage des signatures numériques fait partie d'un ensemble plus vaste connu sous l'acronyme "**PKI**" (Public Key Infrastructure). L'ensemble se déroule moyennant des certificats que vous pouvez obtenir auprès d'une Autorité de certification (voir exemple plus bas). Lorsque vous demandez votre certificat, votre ordinateur crée la paire de clefs composées d'une clé privée (la jaune sur le schéma) et une clé publique (la noire). Votre clé privée est secrète et c'est seulement vous qui y avez accès alors que la clé publique est librement disponible pour tout le monde. Votre clef publique sera attachée à votre certificat que vous obtiendrez de la part de l'autorité de certification à qui vous avez remis votre demande de certificat.

Le PKI (sur lequel est basée la connexion IPSec) vise essentiellement 4 points importants:

1. l'authentification (le destinataire de votre courriel doit pouvoir vérifier que c'est bien vous qui avez envoyé l'objet et pas un autre). Une personne peut intercepter votre mail, extraire votre mot de passe, etc.
2. l'intégrité (s'assurer que le contenu n'a pas été changé en chemin).
3. la confidentialité (s'assurer que le contenu n'est lisible que par le destinataire).
4. la non-répudiation (découlant des 3 premiers points)

L'autorité de certification est chargée de délivrer les certificats, de leur assigner une date de validité (1 jour), ainsi que de révoquer éventuellement des certificats avant cette date en cas de compromission de la clé.

Les certificats sont de petits fichiers divisés en deux parties:

- La partie contenant les informations
- La partie contenant la signature de l'autorité de certification (voir Internet Explorer)

La structure des certificats est normalisée par le standard X.509 de l'Union International des Télécommunication (UIT), qui définit les informations contenues dans le certificat:

- Le nom de l'autorité de certification (VeriSign par exemple)
- Le nom du propriétaire du certificat (la banque UBS par exemple)
- La date de validité du certificat (1 jour à partir de la date courante)
- L'algorithme de chiffrement utilisé (MD5RSA)
- La clé publique du propriétaire

Voici un très bon exemple:

Pour signer le message que vous expédiez (point (5) sur le schéma ci-dessous), il suffit en effet de lui appliquer une fonction de hachage (point (1)) qui produit un résumé (code haché) du message (les algorithmes de hachage les plus connus sont le MD5 (128 bits (Message Digest 5)) et le SHA-1 (160 bits (Secure Hash Algorithm 1)). Le résumé obtenu est propre à chaque message, à l'image d'une empreinte digitale. Cet algorithme assure que si un seul bit du texte original est modifié et si l'on

refaisait un nouveau hachage (empreinte), ce dernier serait radicalement différent du premier. Le code haché peut ensuite être chiffré à l'aide de votre clé privée et annexé à votre message (points (2) et (3)). C'est ce code qui constitue la signature numérique. Le destinataire du message (point (6)) peut ensuite vérifier que vous en êtes bien l'expéditeur en déchiffrant la signature numérique (point (7)), au moyen de votre clé publique (point (8)), que vous lui avez transmis automatiquement avec le mail (point (4)), pour obtenir le code haché (point (9)). Le destinataire applique ensuite la même fonction de hachage au message reçu (point (10) sur le schéma); si les deux codes (points 11 et 12 sur le schéma) sont identiques, vous êtes bien l'expéditeur du message (authentification) et le message n'a pas été altéré (intégrité).

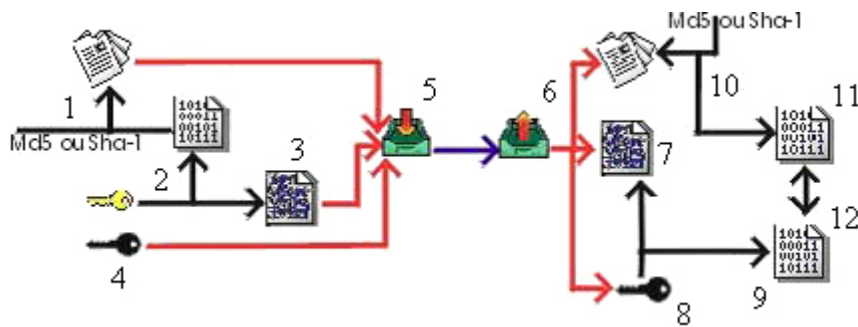


Figure: 61.7 - Principe des certificats d'authentification 1/2

Tout cela a l'air bien compliqué, mais en pratique, vous n'avez qu'à cliquer sur une icône à l'écran pour lancer tout le processus.

Sinon voyons un autre schéma peut-être un peu plus clair:

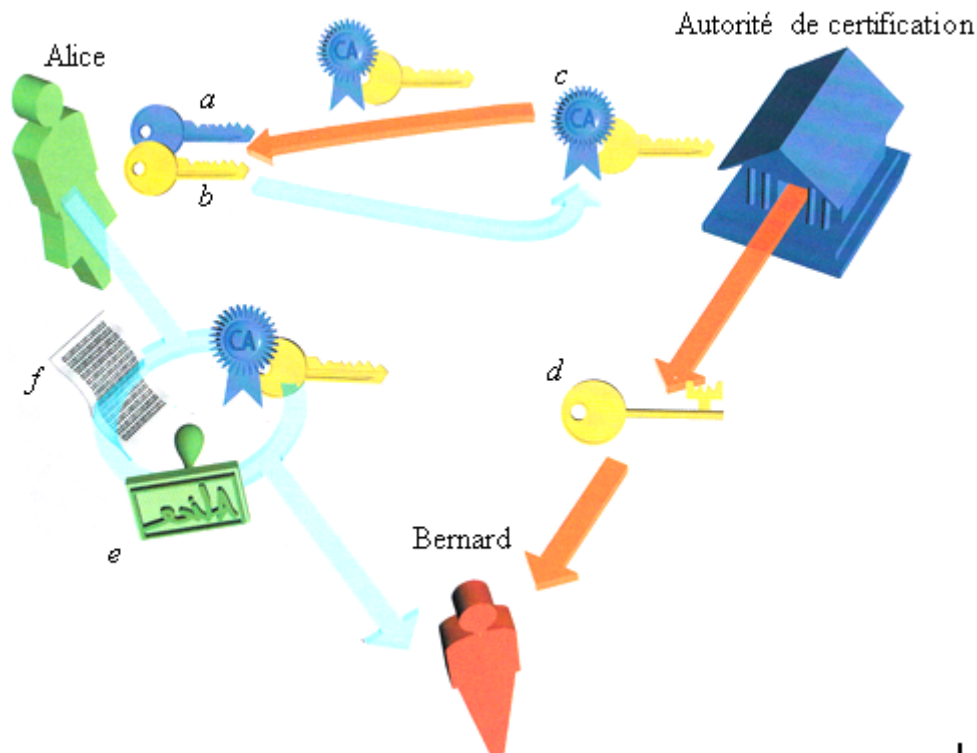


Figure: 61.8 - Principe des certificats d'authentification 2/2 (source: Dossier Pour La Science)

1. Alice utilise une clé secrète (a) ainsi qu'une clé publique (b) reçue d'une autorité de certification qui a transmis par courrier classique les clés privée et publique à Alice dans une carte à puce contenant un

certificat numérique (c). Sur ce certificat, se trouve aussi la signature de l'autorité de certification, laquelle peut être vérifiée par toute personne (ou logiciel) connaissant ou pouvant accéder à la clé publique de cet organisme.

2. Le clé publique (d) de l'autorité de certification est fournie à ceux qui en ont besoin, Bernard par exemple. Cette clé peut être incluse dans les programmes de navigation sur le réseau Internet et dans d'autres logiciels utilisés pour les communications informatiques sécurisées.

3. Alice signe numériquement le message qu'elle envoie à Bernard. Tout d'abord, elle crée un condensé du message en lui appliquant une fonction de hachage. Le condensé ainsi créé est ensuite chiffré à l'aide de la clé secrète d'Alice ce qui donne la signature numérique du message (e). Cette signature est envoyée à Bernard en même temps que le message crypté (f) et la clé publique.

4. Bernard utilise la clé publique de l'autorité de certification pour vérifier si la signature numérique officielle apposée sur le certificat est authentique et que la clé publique qui l'accompagne est celle d'Alice. Il utilise alors cette clé pour déchiffrer la signature numérique d'Alice et obtient le condensé du message. Enfin, Bernard applique la fonction de hachage au message envoyé par Alice et obtient ainsi un condensé du message. Si ce condensé est identique à celui qui est obtenu par le chiffrement numérique d'Alice, Bernard est certain que le message provient bien d'Alice et qu'il n'a pas été altéré par une tierce personne.

7. CRYPTOGRAPHIE QUANTIQUE

La "cryptographie quantique" est une expression médiatique, mais quelque peu trompeuse: en effet, il ne s'agit pas de chiffrer un message à l'aide de la physique quantique, mais d'utiliser celle-ci pour s'assurer que la transmission de la clé n'a pas été espionnée. Comme nous l'avons déjà expliqué en informatique quantique, la transmission d'un message, chiffré ou non, peut se faire en utilisant les deux états de polarisation linéaire orthogonaux d'un photon, par exemple $|x\rangle, |y\rangle$. Nous pouvons décider d'attribuer par convention la valeur 1 à la polarisation $|x\rangle$ et la valeur 0 à la polarisation $|y\rangle$: chaque photon transporte donc un bit d'information. Tout message chiffré ou non peut être alors écrit en langage binaire, comme une suite de 0 et 1, et le message 1001110 sera codé par Alice grâce à la séquence de photons $xyyxxxy$, qu'elle expédiera à Bob par exemple par une fibre optique. À l'aide d'une lame biréfringente, Bob sépare les photons de polarisation verticale et horizontale et deux détecteurs placés derrière la lame lui permettent de décider si le photon était polarisé horizontalement ou verticalement:

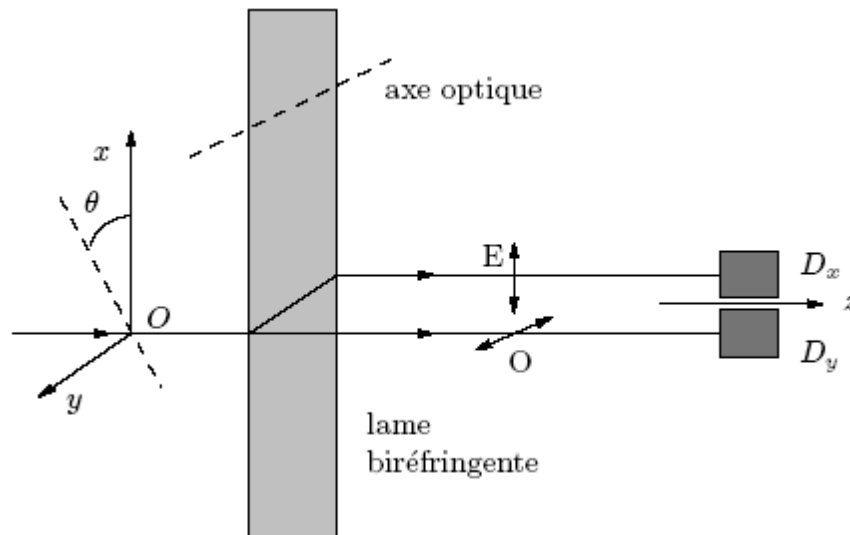


Figure: 61.9 - Expérience de pensée pour la cryptographie quantique

il peut donc reconstituer le message. S'il s'agissait d'un message ordinaire, il y aurait bien sûr des façons bien plus simples et efficaces de le transmettre! Remarquons simplement que si Ève s'installe sur la fibre, détecte les photons et renvoie à Bob des photons de polarisation identique à ceux expédiés par Alice, Bob ne peut pas savoir que la ligne a été espionnée. Il en serait de même pour tout dispositif fonctionnant de façon classique (c'est-à-dire sans utiliser le principe de superposition): si l'espion prend suffisamment de précautions, il est indétectable.

C'est ici que la physique quantique et le principe de superposition viennent au secours d'Alice et de Bob, en leur permettant de s'assurer que leur message n'a pas été intercepté. Ce message n'a pas besoin d'être long (le système de transmission par la polarisation est à ce jour très peu performant). Il s'agira en général de transmettre une clé permettant de chiffrer un message ultérieur, clé qui pourra être remplacée à la demande. Tout ceci satisfaisant le principe de Kerckhoffs.

Avant de passer à la partie très formelle, voyons le principe (vulgarisé) de fonctionnement de ce système:

Dans le transport de "**clé quantique**", l'information est donc transportée par les photons. Chaque photon peut être polarisé, c'est-à-dire que l'on impose une direction à son champ électrique (cf. chapitre d'Optique Ondulatoire).

La polarisation est mesurée par un angle qui varie de 0° à 180° . Dans le protocole que nous décrivons la polarisation peut prendre 4 valeurs: 0° , 45° , 90° , 135° . Pour les photons polarisés à 0° et à 90° , nous parlons de "**polarisation rectiligne**", pour ceux polarisés à 45° et à 135° , de "**polarisation diagonale**":



Figure: 61.10 - Rappels de quelques polarisations classiques

Il nous faut pouvoir détecter la polarisation des photons. Pour cela, nous utilisons un filtre polarisant suivi d'un détecteur de photons. Si un photon polarisé à 0° rencontre un filtre polarisant orienté à 0° , il

traverse ce filtre polarisant et est enregistré par le détecteur placé juste après. Si un photon polarisé à 90° rencontre le même filtre, il est immédiatement stoppé, et le détecteur n'enregistre rien. Maintenant, si le photon est polarisé diagonalement (45° ou 135°), une fois sur deux, il traverse le filtre (superposition de deux états polarisés de manière rectiligne), et une fois sur deux, il est stoppé. Si nous pouvons distinguer entre une polarisation à 0° et à 90° , il est impossible de distinguer en même temps entre une polarisation à 45° et à 135° ! De la même façon, on peut utiliser un filtre polarisant orienté à 45° : il laisse passer les photons polarisés à 45° , stoppe ceux polarisés à 135° , et se comporte aléatoirement avec ceux à 0° et 90° !

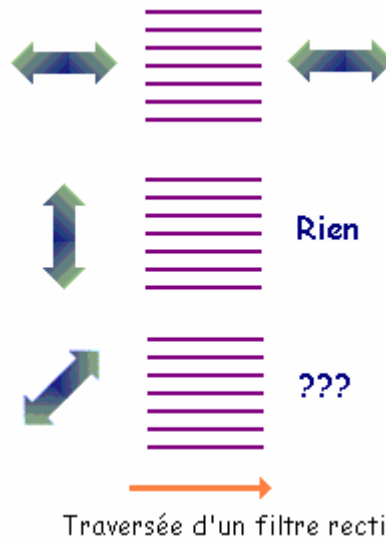


Figure: 61.11 - Passage à travers un filtre polarisant

Décrivons alors le protocole qu'Alice et Bob doivent respecter pour qu'Alice envoie à Bob une clé secrète constituée de 0 et de 1; ils disposent de 2 canaux d'échange: un "canal quantique", où ils peuvent s'échanger des photons polarisés, et un canal radio (non protégé) par lequel ils peuvent discuter. Ils conviennent avant que les photons polarisés à 0° ou 45° représentent 0, et ceux polarisés à 90° ou 135° représentent 1. Alice émet, sur le canal quantique, une suite de photons polarisés au hasard parmi 0° , 45° , 90° et 135° . À l'autre bout, Bob reçoit les photons et mesure aléatoirement ou leur polarisation rectiligne (filtre placé à 0°), ou leur polarisation diagonale (filtre placé à 45°). Si le photon traverse le filtre, Bob note 0, sinon il note 1.

Bien sûr, certaines mesures de Bob (en moyenne, une sur deux) n'ont pas d'intérêt (c'est là que toute l'astuce réside !!!): il a pu essayer de mesurer la polarisation rectiligne d'un photon polarisé à 45° , ce qui n'a pas de sens et donne un résultat aléatoire (par exemple, le photon a été bloqué par le filtre, Bob note donc 1 alors qu'Alice avait envoyé 0). Pour éliminer ces bits sans sens, il indique à Alice, par le canal radio, quel type de mesure (rectiligne ou diagonale) il a faite pour chaque photon. Par le même canal radio, Alice lui indique quelles sont les mesures correctes (photon polarisé à 0° ou 90° avec filtre rectiligne, photon à 45° ou 135° avec filtre diagonal), dans l'exemple ci-dessous la 1, la 3, la 4, et la 7. Les bits 1,3,4,7 sont désormais connus à la fois de Bob et d'Alice, et constituent leur clé secrète commune.

Dans la figure ci-dessous qui représente donc de façon schématique le concept, il ne faut pas oublier que la valeur peut être correcte mais non fiable. Ainsi, l'arbre de décision est le suivant (afin que les choses soient bien claires!):

- soit la valeur fournie par Bob est incorrecte et on élimine la colonne

- soit la valeur est correcte et le filtre est adéquat: cela devient une valeur de la clé
- soit la valeur est correcte mais le filtrage est aléatoire: pour toute sûreté on élimine la colonne.

Alice émet des photons								
Valeur en bit :	0	0	1	1	1	0	0	1
Bob reçoit les photons à travers un filtre								
Le photon passe?	OUI	NON	NON	NON	NON	OUI	OUI	OUI
Valeur en bit :	0	1	1	1	1	0	0	0
---Canal radio---								
Bob : ma mesure	diag	diag	rect	rect	rect	rect	rect	diag
Alice : fiable	oui	non	oui	oui	non	non	oui	non
Clé reconstituée	0	×	1	1	×	×	0	×

Figure: 61.12 - Principe général résumé

Il faut encore vérifier que ce protocole est sûr. Si Caroline écoute le canal quantique, elle peut faire la même chose que Bob: intercepter les photons en plaçant un filtre polarisant tantôt rectiligne, tantôt diagonal. Pour que Bob ne se doute de rien, elle doit réémettre un photon polarisé. Elle va essayer d'envoyer le même photon qu'Alice, mais comme elle a une chance sur deux d'avoir choisi le mauvais filtre, elle a une chance sur deux de se tromper. Quand Bob reçoit le photon, s'il est mal polarisé par Caroline, il a une chance sur deux d'avoir un résultat différent d'avec le photon original, et finalement, pour chaque photon intercepté par Caroline, il y a une chance sur 4 que Bob reçoive une information erronée.

Alice et Bob décident alors de "sacrifier" une partie de leur clé commune. Parmi tous les bits qu'ils ont en commun, ils en choisissent quelques-uns au hasard, et les compare publiquement par le canal radio: s'ils sont différents, ils ont une preuve qu'ils ont été écoutés, et ils oublient vite cette clé. En comparant suffisamment de bits, ils ont une garantie presque absolue de ne pas avoir été écoutés.

Alice émet des photons								
Valeur en bit :	0	0	1	1	1	0	0	1
Caroline intercepte...								
Elle lit :	1	0	1	0	0	0	0	1
et réémet :								
Bob reçoit les photons à travers un filtre								
Le photon passe?	OUI	NON	NON	NON	NON	OUI	OUI	OUI
Valeur en bit :	0	1	1	1	1	0	0	0
---Canal radio---								
Bob : ma mesure	diag	diag	rect	rect	rect	rect	rect	diag
Alice : fiable	oui	non	oui	oui	non	non	oui	non
Clé reconstituée	0	×	1	1	×	×	0	×

Figure: 61.13 - En cas d'interception

Puis... Bob: j'ai peur que nous ayons été espionnés, sacrifions le premier bit de notre clé - j'obtiens 1.
 Alice: je t'avais envoyé 0, nous avons été espionnés...

Remarquons que même non repérée, Caroline n'avait pas la bonne clé, puisque le troisième bit de la clé qu'elle obtient (par rapport à la clé reconstituée d'Alice et Bob) est 0 alors qu'Alice avait envoyé 1 !

Remarque: Le protocole décrit ci-dessus est appelé BB84, du nom de ses inventeurs Bennett et Brassard.

Passons maintenant à la partie formelle (il faut si possible avoir parcouru le début du chapitre d'informatique quantique au préalable).

Les états du système quantique sont les états de polarisation d'un photon: les mesures (de l'observable) auront aussi pour valeur ses états de polarisation. Les mesures possibles seront du type:

$$\begin{aligned}\lambda_0 &= 0 \\ \lambda_1 &= 1\end{aligned}\quad (61.78)$$

nous noterons les états correspondants $|\nu_0\rangle, |\nu_1\rangle$ (base orthonormée de l'espace des états (de polarisation): c'est la base H/V (Horizontale/Verticale).

Prenons plusieurs cas:

C1. Soit un photon dans l'état $|\psi\rangle = |\nu_0\rangle$ alors comme nous l'avons vu en informatique quantique, nous aurons:

$$\begin{aligned}P(T=1) &= P(\lambda = \lambda_0) = 1 \\ P(T=0) &= P(\lambda = \lambda_1) = 0\end{aligned}\quad (61.79)$$

C2. Soit un photon dans l'état:

$$|\psi\rangle = \frac{|\nu_0\rangle + |\nu_1\rangle}{\sqrt{2}} \quad (61.80)$$

Remarques:

R1. Cette (fameuse) valeur est choisie à des fins de normalisation telle que $\langle\psi|\psi\rangle = 1$!!! Beaucoup de gens se posent la question d'où vient la racine carrée en informatique quantique. La réponse est simplement pour la normalisation.

R2. Notons que ce photon n'est pas polarisé dans la direction " $\vec{\nu}_0 + \vec{\nu}_1$ " (c'est-à-dire dans la direction oblique) mais est dans une superposition quantique de ces deux polarisations.

Alors (nous appliquons comme nous l'avons vu dans le chapitre d'Informatique Quantique, le test $|\nu_0\rangle$ à l'état $|\psi\rangle$):

$$\begin{aligned}
 P(0) &= \|\langle v_0 | \psi \rangle\|^2 = \left\| \langle v_0 | \left(\frac{|v_0\rangle + |v_1\rangle}{\sqrt{2}} \right) \right\|^2 = \frac{1}{2} \|\langle v_0 | v_0 \rangle + \langle v_0 | v_1 \rangle\|^2 \\
 &= \frac{1}{2} \left\| \begin{pmatrix} 0 \\ 1 \end{pmatrix} \circ \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \circ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\|^2 = \frac{1}{2} \|1\|^2 = \frac{1}{2}
 \end{aligned}
 \tag{61.81}$$

et:

$$\begin{aligned}
 P(1) &= \|\langle v_1 | \psi \rangle\|^2 = \left\| \langle v_1 | \left(\frac{|v_0\rangle + |v_1\rangle}{\sqrt{2}} \right) \right\|^2 = \frac{1}{2} \|\langle v_1 | v_0 \rangle + \langle v_1 | v_1 \rangle\|^2 \\
 &= \frac{1}{2} \left\| \begin{pmatrix} 1 \\ 0 \end{pmatrix} \circ \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} \circ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\|^2 = \frac{1}{2} \|1\|^2 = \frac{1}{2}
 \end{aligned}
 \tag{61.82}$$

Remarque: Rappelons que sur ce site, nous notons en physique quantique le module d'un nombre complexe et la norme, indistinctement par le symbole $\| \cdot \|$ donc attention aux confusions!

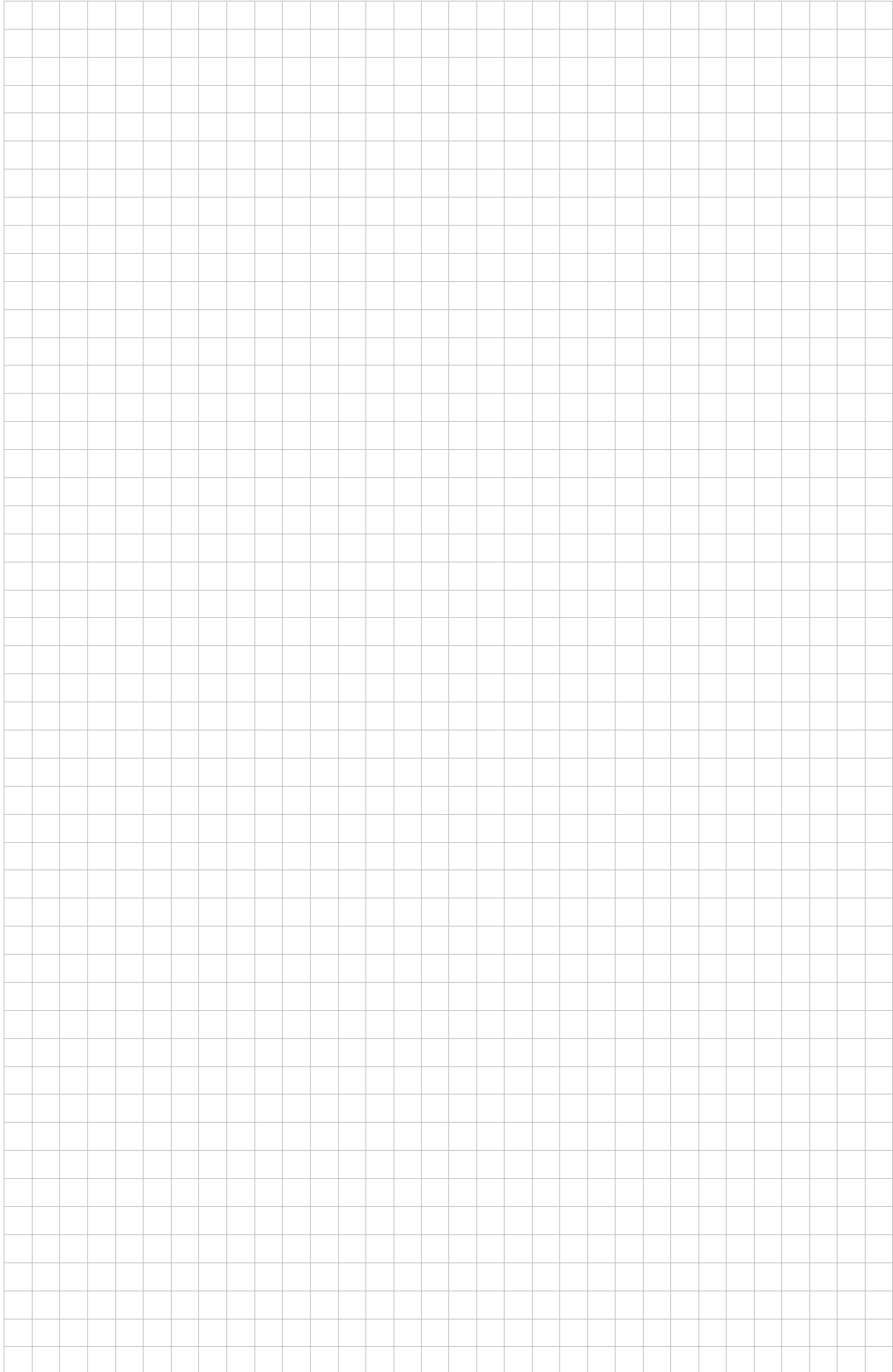
8. CRYPTOGRAPHIE ALTERNATIVE

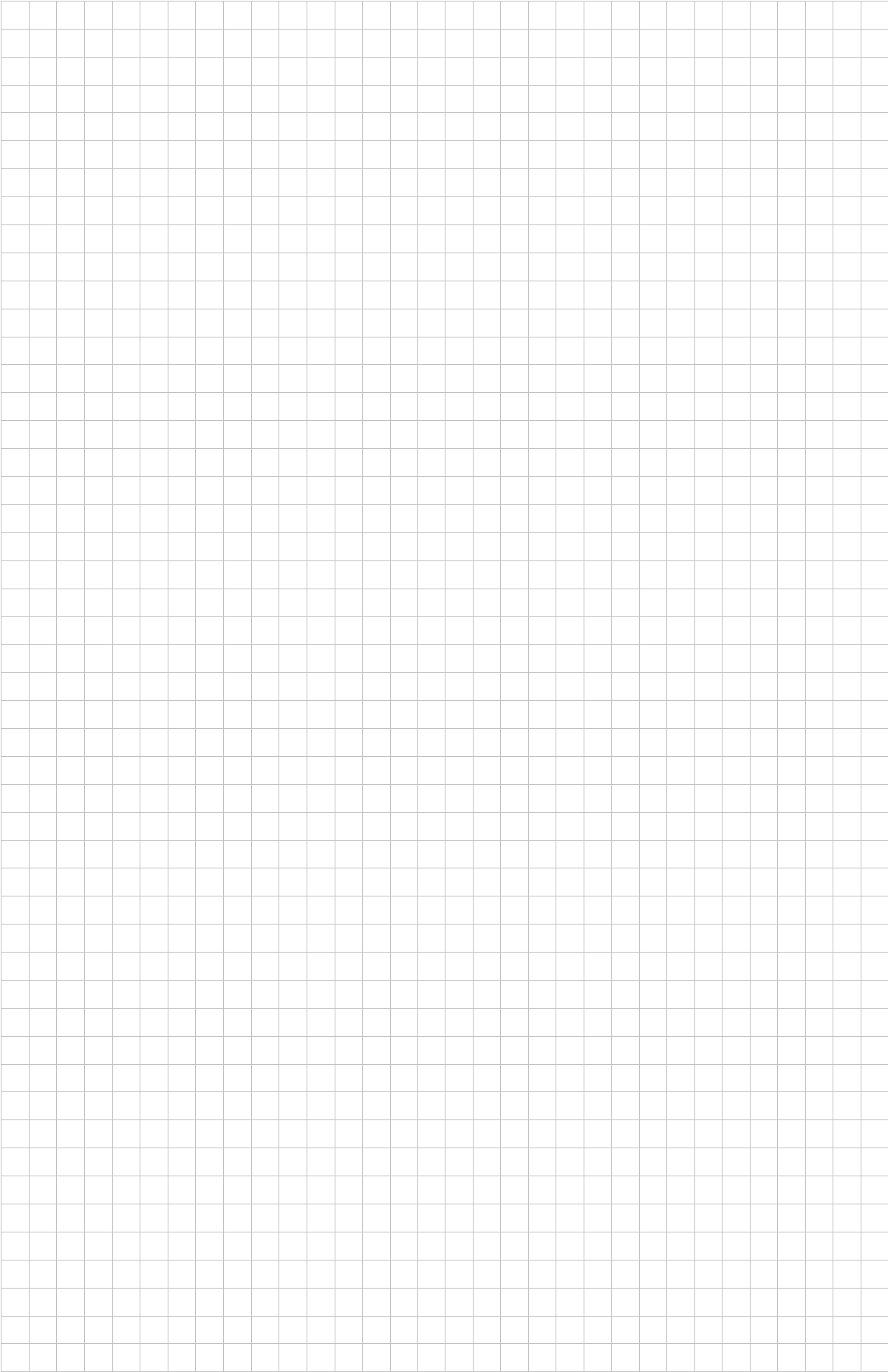
Les mathématiciens s'aventurent parfois hors des sentiers battus de la théorie des nombres: ils inventent des cryptosystèmes fondés sur des tresses ou des réseaux (théorie des noeuds et des graphes). Les physiciens ne sont pas en reste et proposent des méthodes de chiffrement qui utilisent la théorie du chaos ou la physique quantique. Cette dernière apporterait une solution définitive au délicat problème de l'échange de clés et mettrait en péril les cryptosystèmes fondés sur la factorisation.

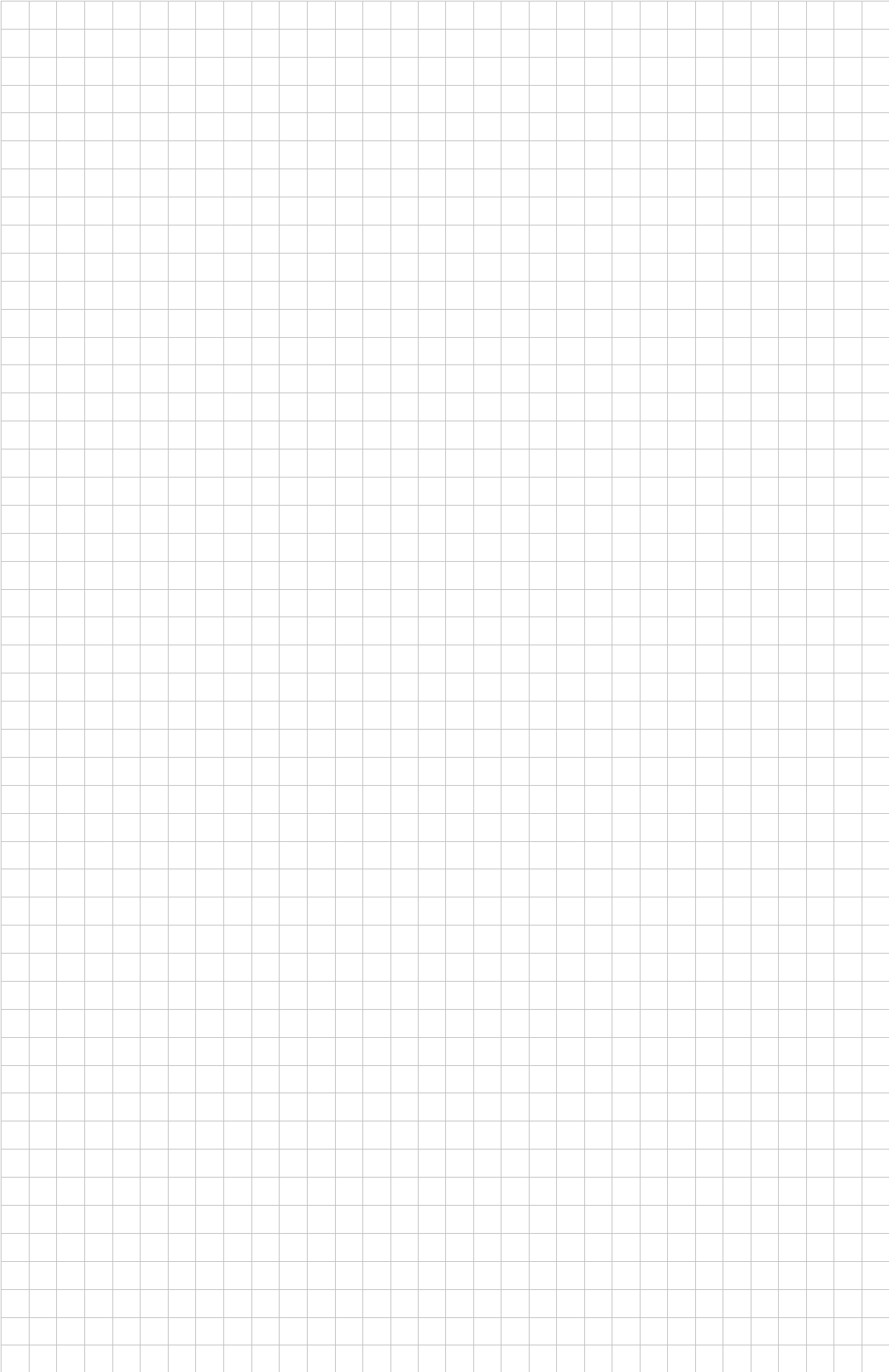
La plupart de ces méthodes sortent pour l'instant du contexte du contenu de ce site mais on peut citer cependant:

- l'algorithme LLL basé sur la structure en maille d'ensembles de nombres et se basant sur le théorème de Minkowski assurant que le contenu d'un disque de rayon donné en un point contient au moins un autre point du réseau
- la cryptographie ultravariante dans laquelle les données passent par des systèmes d'équations quadratiques superposées.
- l'hyperchaos optique, obtenu par le passage d'un LASER dans un anneau d'IKEDA dans lequel se présente un matériau non linéaire en longueur d'onde.
- la cryptographie quantique, basée sur le principe d'incertitude de Heisenberg et l'implication de l'annulation des transferts de données. Les scientifiques cherchent aujourd'hui des moyens de communication moins onéreux des clés quantiques en utilisant entre autres, les propriétés du condensat de Bose-Einstein qui permettrait de contrôler l'émission de photons ainsi que la transmission instantanée d'un message sans liaison physique...

L'avenir nous dira le reste!







62. AUTOMATES (THÉORIE DES LANGAGES)

Le propos de ce chapitre est d'étudier le fonctionnement théorique du concept ordinateur, ou machine. Nous nous situons ici au niveau des mathématiques et de la logique mathématique, indépendamment de toute référence à une machine (ou un logiciel) concrète existante. Nous nous pencherons sur la façon dont cette machine théorique va prendre connaissance de données numériques, de quelque nature qu'elles soient, pour en effectuer un traitement, en vue de résoudre un problème d'ordre général. Nous serons alors amenés à constater que, de ce point de vue, toute machine théorique est réductible, dans son principe de fonctionnement, à une machine idéale. Ainsi, nous pouvons dire que tous les ordinateurs, ou tous les programmes, sont équivalents entre eux, puisque le propre d'un ordinateur, dans sa définition théorique, est l'universalité, c'est-à-dire la capacité à traiter tous les problèmes traitables effectivement.

Remarque: Ce chapitre aurait normalement sa place en tout premier de la section d'informatique théorique mais il nous a semblé plus judicieux de se faire au préalable la main sur des exemples concrets de l'informatique théorique avant de passer au formalisme abstrait de leurs exécutions. C'est une des raisons pour lesquelles nous reviendrons ici brièvement sur les concepts d'algorithmes, de complexité, de systèmes logiques formels, de théorie de la démonstration et de l'information (voir les chapitres du même nom). Par ailleurs, pour ce chapitre, une expérience dans le développement de logiciels informatiques est un grand plus pour comprendre certaines notions (ou pour s'imaginer les applications pratiques).

Avant de commencer, il convient de faire un tour d'horizon très sommaire des questions mises en jeu par ces premiers mots. Mais d'abord citons quelques domaines où la théorie du langage et les automates sont utilisés: spécification des langages de programmation, compilation, recherche de motifs (dans un texte, dans une base de données, sur le web, dans les gènes, ...), compression de textes, preuves de programmes, électronique des ordinateurs, codage pour la transmission, cryptographie, décodage du génome, linguistique, sciences cognitives, etc.

1. MISE EN PERSPECTIVE

L'informatique moderne est née de la recherche entreprise au début du 20ème siècle par Bertrand Russel et Alfred North Whitehead pour constituer la mathématique en un système formel où toute proposition pourrait être démontrée par un calcul logique (cf. chapitre de Théorie De La Démonstration). David Hilbert et Kurt Gödel accomplirent des pas décisifs dans l'exploration de ce programme. En 1931, Gödel démontre que (rappel):

1. Il se peut que dans certains cas, nous puissions démontrer une chose et son contraire (inconsistance)
2. Dans tout système mathématique formel, il existe des vérités mathématiques qu'il est impossible de démontrer (incomplétude)

Le théorème de Gödel ruine ainsi le rêve de réunir la mathématique en un système déductif parfaitement cohérent, mais de l'effervescence intellectuelle autour du projet des *Principia* de Russel et Whitehead vont sortir les idées fondatrices de l'informatique. Cela amène en 1936 Alan Turing, à la suite de Gödel, à s'attaquer au problème de la décidabilité.

Définition: Un système est appelé "**système décidable**" s'il existe une procédure effective pour distinguer les propositions démontrables des autres. Pour définir plus rigoureusement la notion de procédure effective, Turing élabore le concept "**d'automate**", appelé par la suite "**machine de Turing**" (voir exemple plus bas), qui lui permet de préciser la notion d'exécution d'un "**algorithme**" (cf. chapitre de Méthodes Numériques).

Inventer des procédures effectives (des algorithmes) consiste à déterminer un enchaînement d'opérations élémentaires qui exécuteront les calculs nécessaires à la solution de problèmes pour lesquels existent des solutions calculables (il y a des problèmes sans solution et des solutions incalculables comme nous l'avons vu lors de notre étude de la complexité dans le chapitre de méthodes numériques). Turing démontre en outre que son modèle de calcul est universel, c'est-à-dire que toutes les machines de Turing sont équivalentes (nous le démontrerons plus loin). Il formule l'hypothèse selon laquelle tout algorithme est calculable par une machine de Turing. Ces idées fondent la théorie de la programmation des ordinateurs.

1.1. MACHINE DE VON NEUMANN

Il revient à von Neumann de concevoir en 1945 l'architecture générale des appareils concrets qui vont réaliser les calculs selon le modèle de Turing, architecture si efficiente et élégante que les ordinateurs d'aujourd'hui sont encore construits, pour l'essentiel, selon ces principes.

Remarque: Nous pouvons d'une certaine façon dire que cette décennie entre 1936 et 1945 a vu la naissance de l'informatique, qui est passée du stade de construction intellectuelle mathématique et logique à celui de l'application de ces idées à la réalisation de systèmes physiques concrets.

Voici le schéma de l'architecture de von Neumann:

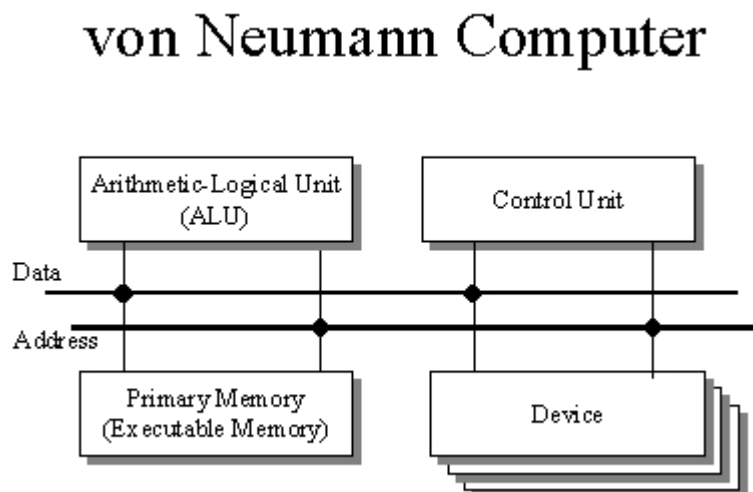


Figure: 62.1 - Principe de la machine de von Neumann

Les unités de contrôle (Control Unit), arithmétique (ALU) et de mémoire primaire (Primary Memory) constituent à elles trois l'unité centrale, ou le "processeur" de l'ordinateur. Le processeur est constitué de circuits électroniques qui peuvent exécuter des actions. L'ensemble des actions câblées dans le processeur constitue le jeu d'instructions du processeur et détermine le langage élémentaire de son utilisation, appelé "**langage machine**".

Le rôle de l'unité de contrôle consiste à permettre le déclenchement de l'action (l'instruction) voulue au

moment voulu. Cette instruction peut appartenir à l'unité arithmétique, à l'unité de mémoire ou à l'unité de contrôle elle-même. Une instruction peut en outre consulter le contenu de la mémoire (la "lire") ou modifier le contenu de la mémoire (y "écrire"). De façon générale, une action consiste soit à consulter ou à modifier l'état de la mémoire ou d'un des registres (qui sont des éléments de mémoire spéciaux incorporés à l'unité centrale), soit à déclencher une opération d'entrée-sortie (communication avec le monde extérieur et notamment l'utilisateur humain).

Exemple:

Comment indiquons-nous à l'unité de contrôle le moment voulu pour déclencher telle ou telle action?: C'est écrit dans le texte d'un programme. Où est le programme?: Dans la mémoire.

La mémoire est constituée d'éléments susceptibles de prendre des états. Un élément de base de la mémoire peut prendre deux états distincts et peut servir à représenter une donnée élémentaire, ou bit (cf. chapitre sur les Systèmes Logiques). Cette représentation d'une donnée par un élément de mémoire s'appelle un "**code**". Une mémoire avec beaucoup de bits permet le codage de données complexes, dans la limite de la taille de la mémoire.

Le chemin par lequel unité centrale, mémoire et organes d'entrée-sortie (Devices) communiquent s'appelle de façon générique un "**bus**" (c'est en quelque sorte l'autoroute où circulent des données d'un point à un autre à l'aide d'adresses). De façon un peu formelle, un bus est un graphe connexe complet (cf. chapitre de Théorie Des Graphes), ce qui veut dire en langage courant que tous les éléments connectés au bus peuvent communiquer entre eux.

Remarque: Le codage fait correspondre des groupes de bits à des symboles. Les symboles les plus simples sont les chiffres et les lettres. Pour représenter des données complexes on peut définir des méthodes, des règles pour regrouper des symboles puis associer un élément de données à un groupe de symboles construit selon les règles.

Définition: Nous appellerons "**langage**" un ensemble de symboles ou de groupes de symboles, construits selon certaines règles, et qui sont les mots du langage. La "**syntaxe du langage**" est l'ensemble des règles de construction des mots du langage.

La mémoire de l'ordinateur (c'est l'idée fondamentale de von Neumann) contient des informations de deux types: des programmes et des données. Les programmes et les données sont représentés avec les mêmes symboles, seule la sémantique permet d'interpréter leurs textes respectifs. D'ailleurs, le texte d'un programme peut parfois être envisagé comme des données pour un autre programme, par exemple un programme de traduction d'un langage dans un autre.

1.2. MACHINE DE TURING

Il importe de se convaincre (ce ne sera pas en un jour) que tous les programmes que nous pourrions écrire dans différents langages sont équivalents. La machine de Turing est un modèle d'automate dont on trouvera ci-dessous une description très terre à terre (avant de passer à une définition beaucoup plus formelle). L'architecture de von Neumann, conçue pour réaliser efficacement les traitements décrits par une machine de Turing, engendre les langages impératifs (voir définition en R1 ci-dessous). Tout programme, fonctionnel ou impératif, destiné à être exécuté, sera traduit dans un langage impératif, le langage machine de l'ordinateur utilisé. La cohérence de l'informatique, et l'équivalence sémantique des programmes écrits en divers langages, qui assurent la validité de cette opération, sont le fruit non pas du

hasard, mais d'une conception théorique originelle commune. Gödel, Church, von Neumann et Turing étaient tous à Princeton en 1936.

Remarques:

R1. Les premiers langages évolués qui apparurent sont des langages dits "**langages impératifs**", fondés sur la notion d'état de la mémoire (c'est l'assembleur au fait!). Ces langages, inspirés du modèle de John von Neumann, comportent comme le langage machine des instructions qui produisent des modifications de la mémoire (instruction d'affectation). La rédaction d'un programme en langage impératif consiste à écrire la suite des instructions qui vont causer les états successifs par lesquels devra passer la mémoire pour que, en partant d'un état initial permettant l'initialisation du programme, elle arrive dans un état final fournissant les résultats recherchés.

R2. Outre les langages impératifs, en informatique, nous distinguons les langages séquentiels, interprétés et compilés.

Un modèle formel pour une procédure effective (pour décrire un algorithme) doit posséder certaines propriétés. Premièrement, chaque procédure doit recevoir une définition finie. Deuxièmement, la procédure doit être composée d'étapes distinctes, dont chacune doit pouvoir être accomplie mécaniquement. Dans sa simplicité, la machine de Turing composée des éléments suivants répond à ce programme:

1. Une mémoire infinie représentée par un ruban divisé en cases. Chaque case du ruban peut recevoir un symbole de l'alphabet défini pour la machine ;
2. Une tête de lecture capable de parcourir le ruban dans les deux sens ;
3. Un ensemble fini d'états parmi lesquels on distingue un état initial et les autres états, dits "**états accepteurs**";
4. Une fonction de transition qui, pour chaque état de la machine et chaque symbole figurant sous la tête de lecture, précise: l'état suivant, le caractère qui sera écrit sur le ruban à la place de celui qui se trouvait sous la tête de lecture, le sens du prochain déplacement de la tête de lecture.

On peut doter sa Machine de Turing de l'alphabet fini de son choix. Son ruban peut être infini dans les deux sens ou dans un seul. Elle peut même avoir plusieurs rubans. On montre que ces diverses machines sont équivalentes.

Nous sommes alors amenés à la définition simpliste suivante:

Un automate fini est un modèle mathématique des systèmes ayant un nombre fini d'états et que des actions (externes ou internes) peuvent faire passer d'un état à un autre. Les actions externes sont représentées par les symboles d'un alphabet A ; les actions internes (invisibles, silencieuses, ou spontanées) sont représentées par un symbole n'appartenant pas à l'alphabet précité.

Un automate est représenté par un graphe (cf. chapitre de Théorie Des Graphes) dont les sommets sont des états et à chaque arc est associée la reconnaissance d'une ou plusieurs lettres.

Les automates finis permettent de modéliser et de contrôler des systèmes à nombre d'états finis et de

résoudre des problèmes courants: analyse lexicale, recherche de motifs dans un texte, analyse du génome, etc.

Exemples:

E1. Automate fini et déterministe qui reconnaît tous les entiers dont l'écriture est normalisée (langage régulier), c'est-à-dire ne commençant pas par 0 (les chiffres dans les cercles sont juste là pour décrire l'ordre dans lequel l'automate exécute l'opération):

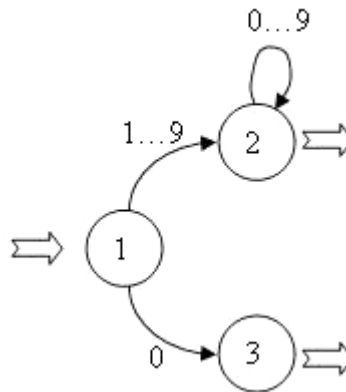


Figure: 62.2 - Premier exemple d'automate fini

Explication: L'automate reçoit dans l'entrée (1) un nombre entier, il regarde si ce nombre commence par un 0 ou est un nombre compris entre 1 et 9. Si le nombre commence par zéro, l'automate sort et s'arrête en (3). Sinon, l'automate va en (2) et analyse les chiffres du nombre les uns après les autres jusqu'à ce qu'il arrive à la fin après quoi il s'arrête et sort en (3).

E2. Automate fini et déterministe qui reconnaît une entrée numérique dans un tableur type langage régulier (par exemple: +12,3 ou 08 ou -15 ou 5E12 ou 14E-3):

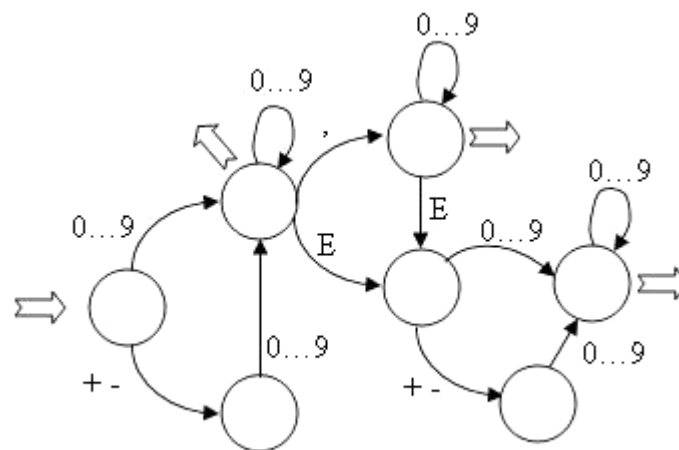


Figure: 62.3 - Deuxième exemple d'automate fini

En d'autres termes, il suffit de reconnaître un langage de la forme:

$$\{-, +, \varepsilon\} A^*, A^* E \{-, +, \varepsilon\} A^* \quad (62.1)$$

qui est bien régulier où ε est le mot vide, A est l'alphabet $\{0, 1, \dots, 9\}$ et A^* l'ensemble des mots (in extenso des nombres) qu'on peut écrire avec A .

E3. Automate fini et déterministe reconnaissant tous les multiples de 3, type langage régulier (en d'autres termes si un tel multiple est trouvé, l'automate effectue une sortie, sinon rien):

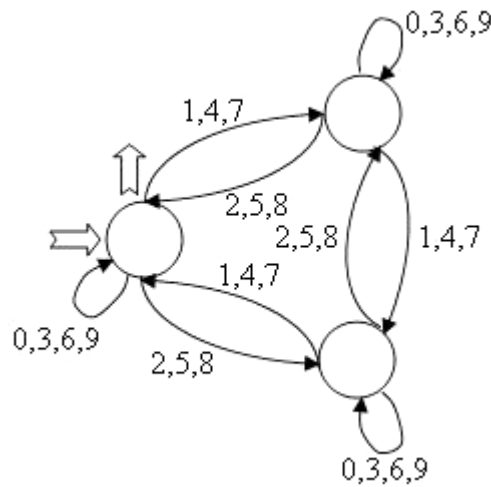


Figure: 62.4 - Troisième exemple d'automate fini

1.3. HIÉRARCHIE DE CHOMSKY

La hiérarchie de Chomsky est une classification des langages décrits par les grammaires formelles proposée en 1956, par le linguiste Noam Chomsky. Elle est aujourd'hui largement utilisée en informatique, en particulier pour la conception d'interpréteurs ou de compilateurs, ou encore pour l'analyse des langages naturels.

Il convient au préalable de définir certaines notions:

1.3.1. LANGAGE FORMEL

Définition (simpliste): Dans de nombreux contextes (scientifique, légal, etc.) l'on désigne par "**langage formel**" un mode d'expression plus formalisé et plus précis (les deux n'allant pas nécessairement de pair) que le langage de tous les jours (voir langage naturel).

En mathématiques, logique et informatique, un langage formel est formé:

1. D'un ensemble de mots obéissant à des règles logiques (grammaire formelle ou syntaxe) strictes.
2. Éventuellement d'une sémantique sous-jacente (la force des langages formels est de pouvoir faire abstraction d'une telle sémantique, ce qui rend les théories réutilisables dans plusieurs modèles)

Remarque: Ainsi, alors qu'un calcul particulier de paye ou de matrice inverse restera toujours un calcul de paye ou de matrice inverse, un théorème sur les groupes s'appliquera aussi bien sur l'ensemble des entiers que sur les transformations du cube de Rubik.

Le langage formel d'une discipline scientifique c'est donc effectivement un langage obéissant à une syntaxe formelle stricte et qui va servir à exposer des énoncés de manière précise, si possible concise et sans ambiguïté, et s'oppose en cela au langage naturel.

Le langage formel a pour avantage de rendre aisées la manipulation et la transformation de ces énoncés. En effet, on va disposer en général de règles de transformation précises (développement de

formules logiques, formes normales, contrapositions, commutativité, associativité, etc.) qui peuvent être appliquées sans même connaître la signification de l'énoncé à transformer ou la signification de la transformation. C'est donc un outil d'exploration puissant, et c'est le seul langage qui permette aux machines de "faire des mathématiques".

L'inconvénient est évident: ne pas connaître le sens de l'énoncé empêche de savoir quelles sont les transformations pertinentes et nuit à l'intuition du raisonnement. Ainsi, il est bon de savoir lire rapidement un énoncé en langage formel et de le traduire tout aussi rapidement en un ou plusieurs énoncés du langage naturel, plus parlants.

C'est là que se trouve la limite de ce que nous appelons les "logiciels d'aide à la preuve": naturellement, l'ordinateur n'a pas d'intuition. Toute l'habileté du concepteur d'un tel programme consiste à trouver des moyens pour que l'ordinateur comprenne.

Donner un sens pertinent à un langage de programmation en vue d'exécuter ses programmes est relativement aisé, du fait que ces langages formels ont été conçus pour signifier des suites d'actions élémentaires de la machine. Pour prouver un programme (démontrer que l'algorithme se termine en un nombre fini de fois) ou un théorème de mathématiques (ce qui revient au même), il n'y a, en revanche, aucune méthode infaillible, la correction d'un programme étant un problème de décision indécidable. Ainsi, le prouveur doit se contenter d'appliquer certaines heuristiques (technique consistant à apprendre petit à petit en tenant compte de ce qui a été fait au préalable) et souvent appeler à l'aide l'utilisateur humain. Cependant, grâce à ses heuristiques et à sa puissance de calcul, l'ordinateur explore des milliers de voies que l'utilisateur humain n'aurait pas pu tester en plusieurs années, accélérant ainsi le travail du mathématicien.

Définition (un peu plus formelle): En tant qu'objet d'étude, un "**langage formel**" est défini comme un ensemble de mots de longueur finie (i.e. chaînes de caractères) déduits d'un certain alphabet fini, c'est-à-dire une partie du monoïde libre (l'ensemble des mots formé sur un alphabet, muni de la loi interne de concaténation (qui est une loi de composition), est un monoïde que nous appelons monoïde libre, dont le mot vide est l'élément neutre) sur cet alphabet.

Remarque: Il faut tout de même que cet ensemble de mots ait un sens, soit pertinent, soit opérationnel, serve à quelque chose. Sinon toute collection de groupements finis de caractères sera un langage formel. En somme que ces mots puissent s'articuler entre eux pour former sens, ou du moins construire une pensée, une démarche, un mécanisme logique, une technique de calcul...

1.3.2. SYNTAXE

Définition: La "**syntaxe**" est la branche de la linguistique qui étudie la façon dont les "**morphèmes libres**" (les mots) se combinent pour former des "**syntagmes**" (nominaux ou verbaux) pouvant mener à des propositions, lesquelles peuvent se combiner à leur tour pour former des énoncés.

Exemple:

Le syntagme "une modeste maison de briques rouges" est englobé dans le syntagme supérieur, c'est-à-dire, la phrase complète. Mais ce même syntagme "une modeste maison de briques rouges" inclut parmi ses éléments, le syntagme inférieur "de briques rouges", complément du nom "maison".

Définitions:

D1. En grammaire scolaire, une "**proposition**" est un syntagme articulé autour d'un verbe. Cette notion est surtout utilisée dans l'apprentissage des langues.

D2. Un "**énoncé**", en linguistique est tout ce qui est prononcé par un locuteur entre deux pauses. Syntactiquement, l'énoncé peut donc s'étendre du simple mot à la phrase (voire au discours) en passant par le syntagme.

Le terme de syntaxe est aussi utilisé en informatique, où sa définition est similaire, modulo une terminologie différente. Ainsi la syntaxe est le respect, ou le non-respect, de la grammaire formelle d'un langage, c'est-à-dire des règles d'agencement des lexèmes (qui, en informatique, ne sont que des entités lexicales) en des termes plus complexes, souvent des programmes. Dans la théorie des langages formels, ce qui joue le rôle de lexème est en général appelé "**lettre**" ou "**symbole**", et les termes produits sont appelés "**mots**".

D'un point de vue purement grammatical, l'étude de la syntaxe concerne trois sortes d'unités:

U1. La phrase, qui est la limite supérieure de la syntaxe

U2. Le mot, qui en est le constituant de base, parfois appelé "élément terminal"

U3. Le syntagme (ou groupe), qui en est l'unité intermédiaire

Les relations syntaxiques entre ces différentes unités peuvent être de deux ordres:

O1. La coordination lorsque les éléments sont de même statut

O2. La subordination dans le cas contraire (lorsqu'il y a subordination, l'élément subordonné remplit une fonction syntaxique déterminée par rapport à l'unité de niveau supérieur)

L'étude de la syntaxe tiendra compte, notamment, de la nature (ou catégorie ou espèce) des mots, de leur forme (morphologie) et de leur fonction. C'est ainsi que l'on parlera plus généralement de rapports morphosyntaxiques.

1.3.3. GRAMMAIRE FORMELLE

Définition (simpliste): Une "**grammaire formelle**" est un formalisme permettant de définir une syntaxe et donc un langage formel, c'est-à-dire un ensemble de mots sur un alphabet donné.

La notion de grammaire formelle est particulièrement utilisée dans les domaines suivants:

- La compilation (analyse syntaxique)
- L'analyse et le traitement des langues naturelles
- Les modèles de calcul (automates, circuits, machines de Turing, etc.)

Pour définir une grammaire, nous avons besoin (voir l'exemple plus bas pour comprendre):

1. D'un alphabet de non-terminaux ;

2. D'un alphabet de terminaux ;
3. D'un symbole initial (l'axiome) pris parmi les non-terminaux ;
4. D'un ensemble de règles de production.

Exemples:

E1. Nous pouvons définir des expressions arithmétiques de la façon suivante (écritures que nous retrouvons fréquemment en théorie de la démonstration):

$$\begin{aligned} \text{exp} &\rightarrow \text{exp} + \text{exp} \mid \text{exp} * \text{exp} \mid (\text{exp}) \mid \text{num} \\ \text{num} &\rightarrow 0\text{num} \mid 1\text{num} \mid 2\text{num} \mid \dots \mid 9\text{num} \mid 1 \mid 2 \mid \dots \mid 9 \end{aligned} \quad (62.2)$$

Les non-terminaux sont ici implicitement *exp* et *num*, les terminaux sont + , * ,(,) et les chiffres. L'axiome est *exp*.

Une utilisation de cette grammaire (règle de production) peut-être la suivante:

$$\text{exp} \rightarrow \text{exp} * \text{exp} \rightarrow \text{exp} * \text{num} \rightarrow 3 * \text{num} \rightarrow 3 * 1\text{num} \rightarrow 3 * 18 \quad (62.3)$$

E2. la syntaxe de la logique propositionnelle classique ou calcul des propositions peut se définir de la façon suivante (cf. chapitre de Théorie De La Démonstration):

$$F = \text{Atom} \mid F \vee F \mid F \wedge F \mid F \rightarrow F \mid F \neg F \mid \exists x F \mid \forall x F \quad (62.4)$$

Les types de grammaires les plus couramment utilisées sont:

1. Les grammaires linéaires gauches qui produisent les mêmes langages que les expressions régulières (c'est ce qui va nous intéresser dans ce chapitre)
2. Les grammaires hors-contexte (exemple ci-dessus)
3. Les grammaires contextuelles (ce type de grammaire requiert un formalisme mathématique et ne peut être défini sans celui-ci)

Un langage est ainsi un ensemble de mots, qui sont simplement des séquences de symboles choisis dans un ensemble fini appelé "**alphabet**". Les langages de la hiérarchie de Chomsky sont formés de mots qui respectent une grammaire formelle particulière. Ce qui les distingue dans le cadre de la classification est la nature de la grammaire.

Remarque: Le plus souvent, les symboles que l'on considère sont formés de plusieurs caractères, de sorte qu'ils correspondent plutôt à ce que l'on appelle des mots dans la langue courante. Lorsqu'il y a ambiguïté, par exemple en analyse lexicale (vocabulaire) et syntaxique (partie de la grammaire qui traite de la fonction et de la disposition des mots et des propositions dans la phrase), on parle de caractères pour les symboles de l'alphabet utilisé pour coder les informations, et de lexèmes pour les symboles de l'alphabet abstrait, qui sont les unités de base du langage. De même, les mots du langage correspondent plutôt à des phrases ou à des textes.)

Définition (plus formelle): Une "**grammaire formelle**", ou simplement "**grammaire**", est formée d'un ensemble fini de symboles terminaux (qui sont les lettres ou les mots du langage), d'un ensemble fini de non-terminaux, d'un ensemble de productions dont les membres gauches et droits sont des mots formés de terminaux et de non-terminaux, et d'un axiome. Appliquer une production consiste à remplacer son membre de gauche par son membre de droite ; l'application successive d'un certain nombre de productions s'appelle une dérivation. Le langage défini par une grammaire est l'ensemble des mots formés uniquement de symboles terminaux qui peuvent être atteints par dérivation à partir de l'axiome.

La hiérarchie de Chomsky est formée des quatre niveaux suivants, du plus restrictif au plus large.

N1. Les "**langages de type 3**", ou "**langages réguliers**": ce sont les langages définis par une grammaire régulière ou une expression régulière, ou bien encore les langages reconnus par un automate fini.

N2. Les "**langages de type 2**", ou "**langages algébriques**": ce sont les langages définis par une grammaire formelle hors-contexte, ou bien encore les langages reconnaissables par un automate à pile non déterministe. Dans cette catégorie se trouvent les langages de programmation informatique.

N3. Les "**langages de type 1**", ou "**langages contextuels**": ce sont les langages définis par une grammaire contextuelle, ou encore les langages reconnaissables par une machine de Turing non-déterministe à ruban de longueur bornée par un multiple fixé de la longueur du mot d'entrée (ce type de langages requièrent un formalisme mathématique et ne peuvent être définis sans celui-ci).

N4. Les langages de type 0, ou "**langages récursivement énumérables**". Cet ensemble inclut tous les langages définis par une grammaire formelle. C'est aussi l'ensemble des langages acceptables par une machine de Turing (que l'on autorise à boucler sur un mot qui n'est pas du langage).

Remarques:

R1. Outre les 4 types de la hiérarchie de Chomsky, il existe des classes intermédiaires remarquables: entre 3 et 2: les langages non-contextuels déterministes, reconnaissables par automate à pile déterministe et les langages compris entre 1 et 0: les langages récursifs, c'est-à-dire reconnaissables par une machine de Turing (celle-ci doit refuser les mots qui ne sont pas du langage).

R2. Les six types ci-dessus sont strictement inclus les uns dans les autres.

Un analyseur pour un langage formel est un programme informatique qui décide si un mot donné en entrée appartient ou non au langage, et éventuellement en construit une dérivation.

On dispose de méthodes systématiques pour écrire des programmes d'analyse des langages de type 2 ou 3. Les interpréteurs ou compilateurs comprennent presque toujours une phase d'analyse lexicale, qui consiste à reconnaître des langages de type 3, suivie d'une phase d'analyse syntaxique qui est une analyse de langage de type 2.

Nous pouvons maintenant de manière vulgarisée (toujours dans l'objectif de préparer le terrain) définir ce qu'est un automate dans le cadre de la hiérarchie de Chomsky.

1.3.4. AUTOMATES ASSOCIÉS

Définition (simpliste): Dans le domaine de l'informatique, un "**automate**" est une machine à traiter de l'information par un modèle formel (une machine de Turing) sur un langage donné. Ainsi:

1. Sur un "**langage fini**" (langage contenant un nombre fini de mots), l'automate associé est une machine comparant un texte avec celui qui est enregistré dans une mémoire morte. La grammaire associée à un langage fini est une liste des mots du langage.
2. Sur un "**langage régulier**" (langage où la correction syntaxique se vérifie en ne mémorisant qu'un nombre fini d'informations), l'automate associé est "l'automate fini déterministe" (c'est-à-dire que, pour chaque mot entré, il n'existe qu'un parcours possible du graphe) ou "l'automate fini indéterministe". La grammaire associée à un langage régulier est une grammaire linéaire gauche.
3. Sur un "**langage algébrique**" (langage où la principale contrainte syntaxique est le parenthésage), l'automate associé est "**l'automate à piles indéterministe**". La grammaire associée est la grammaire algébrique.
4. Sur un "**langage borné**" (description nécessitant un formalisme mathématique), l'automate associé est "**l'automate linéairement borné**". La grammaire associée est la grammaire contextuelle.
5. Sur un "**langage décidable**" (un être intelligent arrive à trouver un procédé pour savoir si on est ou pas dans le langage), l'automate associé est une machine de Turing qui s'arrête sur toutes les données. Il n'y a aucune grammaire associée.
6. Sur un langage "**semi-décidable**" (un être intelligent arrive à trouver un procédé pour savoir si on est dans le langage, mais peut rester dans l'expectative si on est à l'extérieur d'un tel langage), l'automate associé est la machine de Turing. Les grammaires associées sont des grammaires "**semi-Turingienne**", "**de Vangarden**", ou "**grammaires affixes**".

2. TERMINOLOGIE

Les automates travaillent donc essentiellement sur des lettres, mots, phrases et langues. Afin de construire des méthodes d'analyses et de traitement rigoureuses et optimales sur le sujet il est intéressant de formaliser les objets traités. C'est ce que nous allons faire dans un premier temps en définissant ceux-ci et leurs propriétés mathématiques (qui sont très intuitives).

2.1. MOTS

Définitions:

D1. Un "**alphabet**" est un ensemble dont les éléments sont des "**lettres**". Les alphabets sont toujours supposés finis.

D2. Un "**mot**" est une suite finie de lettres que nous notons par juxtaposition:

$$w = a_1 a_2 \dots a_n \quad a_i \in A \quad (62.5)$$

D3. Le "**mot vide**", noté ε , est le seul mot composé d'aucune lettre.

D4. La "**longueur**" d'un mot w est le nombre de lettres qui le composent, et est notée $|w|$ (le mot vide est

le seul mot de longueur 0).

Le "**produit de concaténation**" de deux mots $x = a_1 a_2 \dots a_n$ et $y = b_1 b_2 \dots b_m$ est le mot xy obtenu par juxtaposition (concaténation):

$$xy = a_1 a_2 \dots a_n b_1 b_2 \dots b_m \quad (62.6)$$

Bien entendu (trivial), nous avons $|xy| = |x| + |y|$. Nous notons A^* l'ensemble des mots sur A .

Exemple:

Les gènes sont des mots sur l'alphabet ACGT, les protéines sont des mots sur un alphabet à 20 lettres. Les entiers naturels, écrits en base 10, sont des mots sur l'alphabet des dix chiffres décimaux...

Soit A un alphabet. Soit B une partie de A . Pour tout mot $w \in A^*$, la longueur en B de w est le nombre d'occurrences de lettres de B dans le mot w . Ce nombre sera noté $|w|_B$.

Remarque: En particulier, nous avons trivialement $|w| = |w|_A$.

Pour toute lettre $a \in A$, $|w|_a$ est le nombre d'occurrences de a dans w . Nous avons:

$$|w|_B = \sum_{a \in B} |w|_a \quad (62.7)$$

Soit $w = a_1 \dots a_n$, avec $a_1 \dots a_n \in A$. Le mot miroir de w est le mot noté $\tilde{w}$ ou $w^\sim$ défini par:

$$w^\sim = a_n \dots a_1 \quad (62.8)$$

Évidemment:

$$(uv)^\sim = \tilde{v}\tilde{u} \text{ et } (w^\sim)^\sim = w \quad (62.9)$$

Définitions:

D1. Un mot u est un "**préfixe**" ou "**facteur gauche**" d'un mot v s'il existe un mot x tel que $ux = v$. Le mot u est "préfixe strict" ou "propre" si, de plus, $u \neq v$. De manière symétrique, u est "**suffixe**" ou "**facteur droit**" de v si $xu = v$ pour un mot x . Si $u \neq v$, alors u est "**suffixe strict**" ou "**propre**". Le nombre de préfixes d'un mot v non vide est $1 + |v|$ (le mot vide étant toujours un préfixe, nous avons toujours n'importe quel mot non vide qui a comme préfixe au moins le mot vide).

D2. Un mot u est "**facteur**" d'un mot v s'il existe x, y non vides tels que $v = xuy$.

Exemple:

Le mot $(v) aabab$ sur $A = \{a, b\}$ a 12 facteurs (u) différents possibles:

$$\{\varepsilon, a, b, aa, ab, ba, aab, aba, bab, aaba, abab, aabab\} \quad (62.10)$$

Lemme de Levy: Soient x, y, z, t des mots tels que $xy = zt$. Alors il existe un mot w tel que:

$$xw = z \quad (62.11)$$

avec bien évidemment:

$$y = wt \quad (62.12)$$

ou:

$$x = zw \quad (62.13)$$

avec aussi in extenso:

$$wy = t \quad (62.14)$$

Il en résulte logiquement en particulier que si $|x| = |z|$, le mot w est vide, et donc $x = z$ et $y = t$. En d'autres termes, un monoïde libre (voir le rappel plus bas) est simplifiable à gauche et à droite.

Démonstration (bof!):

Posons:

$$x = a_1 \dots a_n, y = a_{n+1} \dots a_m \quad (62.15)$$

avec $a_i \in A$ de même:

$$z = b_1 \dots b_p, t = b_{p+1} \dots b_q \quad (62.16)$$

avec $b_i \in A$.

Comme:

$$xy = zt \quad (62.17)$$

nous avons:

$$m = q \quad (62.18)$$

(mais pas nécessairement $n = p$) et:

$$a_i = b_i \quad (62.19)$$

pour $i = 1, \dots, m$ de sorte que:

$$z = a_1 \dots a_p \text{ et } t = a_{p+1} \dots a_m \quad (62.20)$$

Si $|z| = p \leq n = |x|$, posons $w = x_{p+1} \dots x_n$. Alors:

$$x = zw \text{ et } wy = t \quad (62.21)$$

Si $|z| > |x|$, posons $w = x_{n+1} \dots x_p$. Alors:

$$xw = z \text{ et } y = wt \quad (62.22)$$

□C.Q.F.D.

Rappel: Dans le cadre des automates un "monoïde libre" est un ensemble A (l'alphabet), dont les éléments sont les lettres.

$$(A, \cdot) \text{ est un monoïde libre si } \Rightarrow \begin{cases} \cdot \text{ est associative (concaténation)} \\ \cdot \text{ il existe un élément neutre } (\varepsilon) \text{ dans } A \end{cases}$$

Remarque: Dans la section d'arithmétique, chapitre de théorie des ensembles, nous parlons simplement de "monoïde". Le monoïde $(A, \cdot) = (A, \text{concaténation})$.

2.2. LANGAGES

Les sous-ensembles de A^* sont appelés des "**langages formels**". Par exemple, pour $A = \{a, b\}$, l'ensemble $\{a^n b^n \mid n \geq 0\}$ est un langage.

Nous définissons sur les langages plusieurs opérations. Les opérations ensemblistes sont l'union, l'intersection, la complémentation et la différence qui s'en déduit. Si X et Y sont deux parties de A^* , alors:

$$\begin{aligned} X \cup Y &= \{z \in A^* \mid z \in X \text{ ou } z \in Y\} \\ X \cap Y &= \{z \in A^* \mid z \in X \text{ et } z \in Y\} \\ X^c &= A^* \setminus X = \{z \in A^* \mid z \notin X\} \\ X \setminus Y &= X \cap Y^c = \{z \in A^* \mid z \in X \text{ et } z \notin Y\} \end{aligned} \quad (62.23)$$

Le produit (de concaténation) de deux langages X et Y est le langage:

$$XY = \{xy \mid x \in X \text{ et } y \in Y\} \quad (62.24)$$

Remarque: Nous avons bien évidemment $X\{\varepsilon\} = \{\varepsilon\}X = X$

et le quotient gauche de Y par X :

$$X^{-1} \cdot Y = \{w \mid \exists u \in X \text{ et } v \in Y \mid v = uw\} \quad (62.25)$$

 Exemple:

$$\begin{aligned}
X &= \{0, 1, 2, 3, 4\} \\
Y &= \{5, 6, 7, 8, 9\} \\
Z &= \{0, 1\} \\
X \cup Y &= \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\} \\
XY &= \{05, 06, 07, 08, 09, 15, 16, 17, 18, 19, 25, 26, 27, 28, 29, 35, \dots\} \\
Z^{-1} \cdot (XY) &= \{5, 6, 7, 8, 9\} = Y \\
Z^{-1} \cdot X &= \{\varepsilon\} \\
Y^{-1} \cdot X &= \emptyset \\
(X \cup Y) \setminus \{\varepsilon\} &\text{ est l'ensemble de toutes les suites d'écritures } \\
Z^+ &= \{0, 1, 00, 01, 10, 11, \dots\}
\end{aligned} \tag{62.26}$$

Suite à la demande d'un lecteur détaillons la 6ème ligne en rappelant la définition:

$$Z^{-1} \cdot Y = \{w \mid \exists u \in X \text{ et } v \in Y \mid v = uw\} \tag{62.27}$$

Nous avons alors explicitement:

$$\{0, 1\}^{-1} \cdot \{05, 06, 07, 08, 09, 15, 16, 17, 18, 19, 25, 26, \dots\} \tag{62.28}$$

et dans le produit de concaténation des deux langages X et Y , les seuls mots où nous trouvons les éléments $\{0, 1\}$ du langage Z en tant que préfixe sont:

$$\{05, 06, 07, 08, 09, 15, 16, 17, 18, 19\} \tag{62.29}$$

et comme par définition $Z^{-1} \cdot (XY)$ est l'ensemble unique des termes w qui suivent les préfixes que constitue les termes de Z il ne reste alors plus que:

$$\{5, 6, 7, 8, 9\} \tag{62.30}$$

Nous avons les propriétés suivantes:

$$P1. X(Y \cup Z) = XY \cup XZ \text{ (trivial)}$$

P2. $X(Y \cap Z) \subset (XY \cap XZ)$ où l'inclusion est généralement stricte. Pour conceptualiser cette propriété, il ne faut surtout pas oublier que X est un ensemble de mots et que Y, Z n'ont pas nécessairement des mots de la même longueur !

Les puissances de X sont définies par $X^0 = \{\varepsilon\}$, $X^1 = X$ et $X^{n+1} = X^n X$, $n \geq 1$.

En particulier, si A est un alphabet, A^n est l'ensemble des mots de longueur n .

Définitions:

D1. "L'étoile" (ou "fermeture de Kleene") de X est l'ensemble:

$$X^* = \bigcup_{n \geq 0} X^n = \{x_1 \dots x_n \mid n \geq 0, x_1, \dots, x_n \in X\} \quad (62.31)$$

 Exemple:

Soit $X = \{a, ba\}$. Les mots de X^* , classés par longueur sont:

$$\begin{array}{ll} 0 & \varepsilon \\ 1 & a \\ 2 & aa, ba \\ 3 & aaa, aba, baa \\ 4 & aaaa, aaba, abaa, baaa, baba \\ \dots & \end{array} \quad (62.32)$$

D2. L'opération "plus" est définie de manière similaire:

$$X^+ = \bigcup_{n > 0} X^n = \{x_1 \dots x_n \mid n > 0, x_1, \dots, x_n \in X\} \quad (62.33)$$

2.3. ÉQUATIONS

D'abord voyons un petit quelque chose dont nous allons avoir besoin par la suite: soient u et v deux mots non vides. Les trois conditions suivantes sont équivalentes (sans démonstration car assez trivial de tête):

C1. $uv = vu$

C2. Il existe deux entiers $n, m \geq 1$ tels que $u^n = v^m$

C3. Il existe un mot w non vide et deux entiers $k, l \geq 1$ tels que $u = w^k, v = w^l$

Remarque: Rappelons à nouveau que nous n'avons pas forcément $|u| = |v|$ mais que nous pouvons très bien avoir $|u| > |v|$.

Passons maintenant aux choses intéressantes (certains points flous du chapitre de théorie de la démonstration peuvent s'éclaircir ici parfois...):

Définitions:

D1. Soient V et A deux alphabets disjoints (vous pouvez les imaginer comme l'ensemble des variables et respectivement celui des constantes par exemple...). Une "équation en mots" avec constantes sur A est un couple $e = (\alpha, \beta)$ de mots de $(V \cup A)^*$. Une telle équation est représentée par $\alpha = \beta$. Il faut donc voir les deux mots choisis comme le membre de gauche et respectivement de droite d'une équation

D2. Une équation $e = (\alpha, \beta)$ est dite "**équation non triviale**" si $\alpha \neq \beta$

Exemple:

Soient $A = \{a\}, V = \{x\}$ et définissons $\alpha = ax, \beta = xa$, dès lors nous avons $xa = ax \Leftrightarrow \alpha = \beta$

D3. Une équation e est dite "**équation sans constante**" si $\alpha, \beta \in V^*$.

D4. Une "**solution**" de l'équation e est un homomorphisme de monoïde (cf. chapitre de Théorie Des Ensembles) $h: (V \cup A)^* \rightarrow A^*$ invariant (car toute lettre sur A est envoyée sur A^* et par conséquent tout mot sur A est envoyé sur A) sur A tel que:

$$h(\alpha) = h(\beta) \quad (62.34)$$

Rappel: La définition de l'homomorphisme est telle que si $\alpha = xy$ alors $h(\alpha) = h(xy) = h(x)h(y)$.

Exemple:

Soient $A = \{a, b\}, V = \{x, y\}$. Maintenant considérons les mots suivants:

$$\alpha = xab \quad \beta = byx \quad (62.35)$$

définissons h tel qu'il envoie x sur b , y sur a , a sur a , b sur b . Dès lors nous avons bien:

$$h(\alpha) = h(xab) = bab \quad h(\beta) = h(byx) = bab \quad (62.36)$$

et nous aurons toujours pour tout couple:

$$(\alpha, \beta) \rightarrow h(\alpha) = h(\beta) \quad (62.37)$$

D4. Une solution h est dite "**solution cyclique**" s'il existe un mot w (appartenant à A) tel que $h(x) \in w^*$ (en considérant le mot lui-même comme un alphabet donc) pour toute variable x .

Remarque: Selon cette définition, les solutions de l'équation sans constantes $xy = yx$ en deux variables x, y sont cycliques telles que:

$$h(xy) = yx, h(yx) = xy \Leftrightarrow h(x) = y, h(y) = x \quad (62.38)$$

2.4. CODES

Définition: Nous appelons "**code**" toute partie C d'un monoïde libre A^* qui vérifie la condition suivante pour tout (mot) $x_1, \dots, x_n, y_1, \dots, y_m \in C$:

$$x_1 \dots x_n = y_1 \dots y_m \Rightarrow n = m, x_i = y_i, i = 1, \dots, n \quad (62.39)$$

En d'autres termes, C est un code si tout mot de C^* (mot composé de mots) se factorise, de manière unique, en un produit de mots de C . Lorsqu'un ensemble n'est pas un code, nous nous en apercevons en général assez facilement.

 Exemples:

E1. L'ensemble (langage) $\{a, ab, ba\}$ n'est pas un code puisque le mot aba s'écrit à la fois comme produit $a \cdot ba$ et comme produit $ab \cdot a$.

Remarque: Les codes les plus simples sont les "**codes uniformes**". Ce sont des ensembles dont tous les mots ont une même longueur (ce qui fait qu'étant donné que chaque mot est différent, la combinatoire des mots peu très difficilement ne pas différer).

E2. L'ensemble A^n des mots de longueur n est un code, si $n \geq 1$. Le code ASCII qui associe à certains caractères des mots binaires de longueur 7 (voir table ASCII) est un code uniforme.

2.4.1. CODES PRÉFIXES

Le morse code la lettre E , la plus fréquente, par un '.' et la lettre Y , plus rare, par '-.--': c'est un exemple de code de longueur variable, ce qui permet de représenter les lettres ou les mots les plus fréquents par des mots plus courts. Une propriété importante est l'unicité du décodage (application injective), problème qui ne se pose pas pour les codes de longueur constante. Il peut être résolu, mais de façon trop coûteuse, quand un symbole spécial sépare deux mots successifs du code (le blanc dans le cas du morse).

On peut ne pas recourir à cette technique si aucun mot du code n'est le préfixe d'un autre mot du code; un code présentant cette propriété est appelé un "**code préfixe**".

 Exemple:

Supposons que nous décidons d'une convention de code à taille variable, qui fait correspondre, entre autres, les valeurs suivantes:

0	"11"
2	"11010"
12	"00"
127	"0111100"
255	"0100"

Tableau: 62.1 - Correspondance de code arbitraire

Supposons alors que nous ayons à décoder la séquence: 1101000111100

Plusieurs interprétations (factorisation) sont alors possibles:

$$1101000111100 = 11 \ 0100 \ 0111100 = 0 \ 255 \ 127$$

$$1101000111100 = 11010 \ 00 \ 11 \ 11 \ 00 = 2 \ 12 \ 0 \ 0 \ 12$$

Et maintenant, nous sommes bien embarrassés! Avec plusieurs possibilités équivalentes entre lesquelles on ne sait pas trancher, on est incapable de retranscrire le code initial.

Le problème qui s'est posé ici est que certains codes sont le début d'autres codes. Ici, "11" est le code du nombre 0, mais c'est aussi le début de "11010", code du nombre 2. D'où l'ambiguïté.

Nous appelons alors "**code-préfixe**" un codage dans lequel aucun code n'est le début d'un autre. Ainsi, pour qu'il n'y ait aucune ambiguïté au moment du décodage, il nous faut absolument un code-préfixe.

Exemple:

L'ensemble $X = \{ab, ababa, baa\}$ est un code. Ici, la connaissance du début *abababa* ne permet pas encore de savoir si la décomposition commence par $ab \cdot ab \cdot ab \cdot a$ ou par $ababa \cdot ba$. Ce n'est qu'après la lecture de la lettre suivante (non encore indiquée dans l'exemple) que l'on sait si la décomposition commence par ab (si la lettre est b) ou par $ababa$ (si la lettre est a)

Définition: Un code est à "**délai de déchiffre fini**" s'il existe un entier d tel que, chaque fois qu'un message w commence par $p = x_1 x_2 \dots x_{d+1}$ avec $x_1, x_2, \dots, x_{d+1} \in X$ alors la factorisation complète de w commence par x_1 . C'est donc après un "délai" de d mots de code que nous pouvons affirmer que le premier mot trouvé est le bon.

Exemples:

Le code:

$$X = \{a, ba, bb\} \quad \text{Tableau: 62.2}$$

a un délai de 0.

Le code:

$$X = \{ab, ababa, baa\} \quad \text{Tableau: 62.3}$$

a un délai de 2.

3. ALGORITHMES LINGUISTIQUES

Mettons un peu en pratique ce qui a déjà été vu jusqu'ici afin de nous appuyer un peu sur du concret "utile" ! Nous verrons plus loin une fois les automates introduits, quelques algorithmes de recherche de sous-chaînes (tels qu'utilisés en génomique).

3.1. ALGORITHME DE HUFFMAN

Rappel: en informatique, nous décidons de coder un nombre entier compris entre 0 et 255 par une séquence de 8 chiffres binaires (binary digits, ou "bits" en anglais, valant 0 ou 1), encore appelée octet.

A priori, il suffit pour cela de se donner une table de correspondance du genre:

...	...
138	10001010
139	10001011
...	...

Tableau: 62.4 - Correspondance de code arbitraire

La correspondance peut être quelconque, dictée par notre imagination, du moment qu'à chaque entier entre 0 et 255 est affecté un code binaire et un seul. Une fois une correspondance fixée, il suffit de la

prendre comme convention.

En pratique, la convention qui a été choisie est celle de la représentation en base 2. Elle a le mérite d'être naturelle et logique. Dans la table ci-dessus, 10001010 est la représentation en base 2 de 138. C'est la manière de noter les chiffres que nous aurions adoptée si nous n'avions que deux doigts au lieu de dix.

L'octet défini selon cette convention est l'unité de base de stockage des données. Tout fichier informatique est une suite d'octets rangés selon un ordre bien défini. La taille du fichier est simplement le nombre d'octets qui le constituent. Le kilo-octet (Ko) correspond à 1024 (et non pas 1000) octets, le méga-octet (Mo) à 1024×1024 octets (convention absurde des informaticiens).

Il convient de noter que cette représentation en base 2, n'est qu'une convention. D'autres conventions sont possibles, qui conviendraient tout autant pour peu que tout le monde s'accorde à employer la même convention.

Le problème que nous nous posons est: y aurait-il une autre manière de coder les chiffres, peut-être moins logique mais plus judicieuse, de telle manière que la taille du même fichier réécrit selon la nouvelle convention, serait moins grande?

La convention de codage binaire est finalement très démocratique: que vous soyez un Zéro ou un DeuxCentCinquanteCinq, nous vous allouons 8 bits de toute manière pour pouvoir vous coder. Autrement dit, chaque entrée possible (un nombre entre 0 et 255) est codée sur 8 bits. Il s'agit d'un codage à taille fixe.

Du point de vue de notre problème (la compression de données), cela ne nous importerait pas si chacune des valeurs possibles (0 ... 255) était représentée aussi fréquemment que les autres. Mais en général c'est n'est pas le cas.

À titre d'exemple, voir l'analyse des octets du fichier Wordpad.exe (cf. dans votre dossier Accessoires...). Dans ce tableau, en abscisse comme en ordonnée sont portées les valeurs possibles d'un octet donné (donc 0 ... 255). Sur la diagonale, à l'abscisse et ordonnée correspondante la taille du cercle est proportionnelle au nombre d'octets ayant cette valeur dans le fichier:

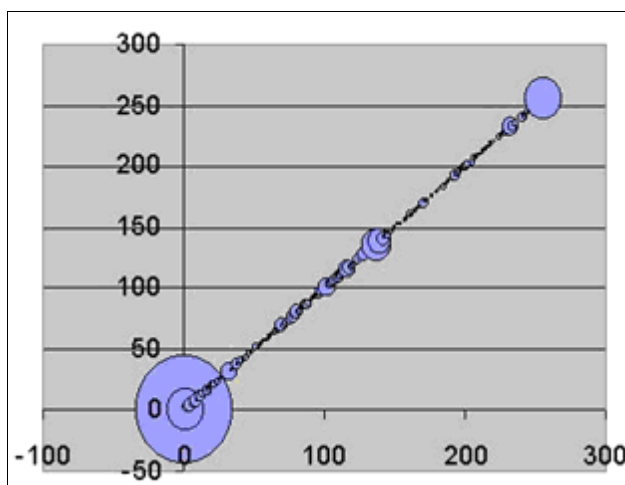


Figure: 62.5 - Analyse de la répartition des octets d'un fichier particulier

Nous voyons clairement que les valeurs 0, 128 et 255 sont nettement plus fréquentes que les autres!

À titre indicatif, voici quelques valeurs:

Valeur	Nombre	Fréquence
0	71891	34.4%
2	1119	0.53%
128	1968	0.94%
130	79	0.038%
255	10422	4.99%

Tableau: 62.5 - Correspondance valeurs/fréquences

Nous allons maintenant décider d'une convention de codage à taille variable, qui représente une valeur fréquente par un petit nombre de bits, et une valeur peu fréquente par un grand nombre de bits.

Par exemple, 0 sera maintenant représenté par la séquence "11" (anciennement "00000000"), 128 par "1011010" (anciennement "10000000"), 255 par "0100" (anciennement "11111111"), etc.

Étant donné que 0 représente à lui seul un tiers du fichier, nous avons gagné une place considérable en le codant sur deux bits au lieu de huit! Et idem pour les autres valeurs fréquentes...

L'algorithme de Huffman est une recette pour générer un code-préfixe à taille variable, à partir de la table des fréquences d'une suite de valeurs. Donc, si vous nous avez bien suivis dans la théorie jusqu'ici, c'est une solution à notre problème.

Supposons que notre fichier soit extrêmement simple, et constitué d'un mot unique:

anticonstitutionnellement

Il y a 25 caractères dans ce fichier; chaque caractère étant codé par un octet de 8 bits (codage ASCII) cela signifie donc 25 octets, ou encore 200 bits. Voyons ce que nous pouvons faire de cela.

Table des fréquences:

Clé	Fréquence
' a '	1
' c '	1
' s '	1
' u '	1
' m '	1
' o '	2
' l '	2
' i '	3
' e '	3
' n '	5
' t '	5

Tableau: 62.6 - Fréquence des lettres dans un mot

Tous les autres octets ont une fréquence nulle: ils ne sont pas représentés dans le fichier.

Première étape, nous créons maintenant un "noeud terminal" pour chaque entrée du tableau:

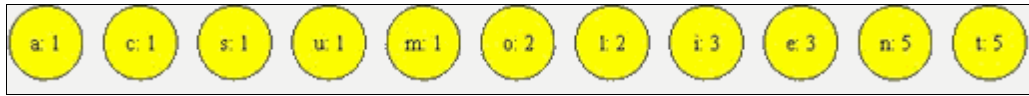


Figure: 62.6 - Noeud terminal associé à chaque entrée du tableau

Ce qui nous fait pour l'instant 11 arbres contenant un seul noeud chacun.

Nous commençons maintenant une itération: à chaque fois nous supprimons les deux arbres de gauche et les remplaçons par un "arbre somme". Le nouvel arbre est inséré dans la liste en respectant l'ordre croissant, et l'on recommence jusqu'à ce qu'il ne reste plus qu'un seul arbre:

Itération 1:

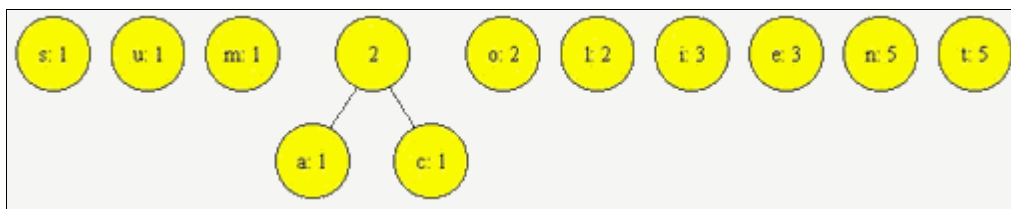


Figure: 62.7 - Première itération

Itération 2:

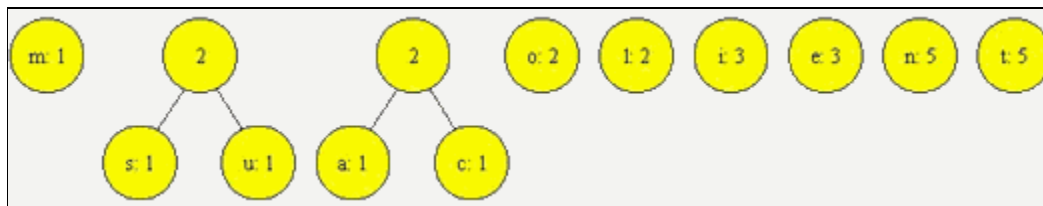


Figure: 62.8 - Deuxième itération

Itération 3:

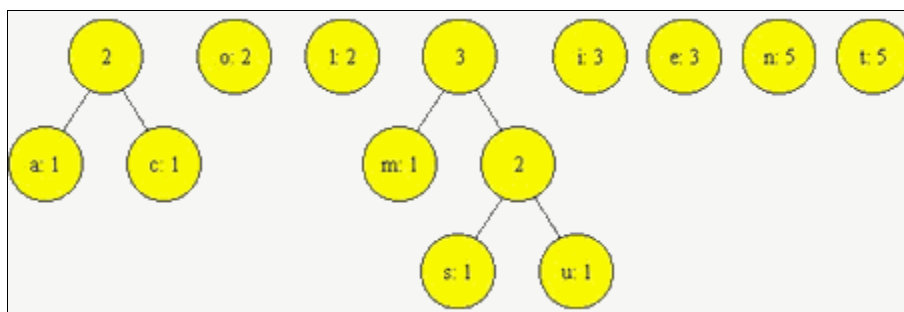


Figure: 62.9 - Troisième itération

Itération 4:

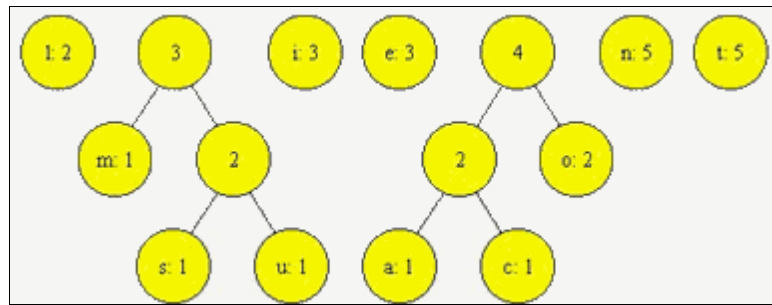


Figure: 62.10 - Quatrième itération

etc.

L'arbre final:

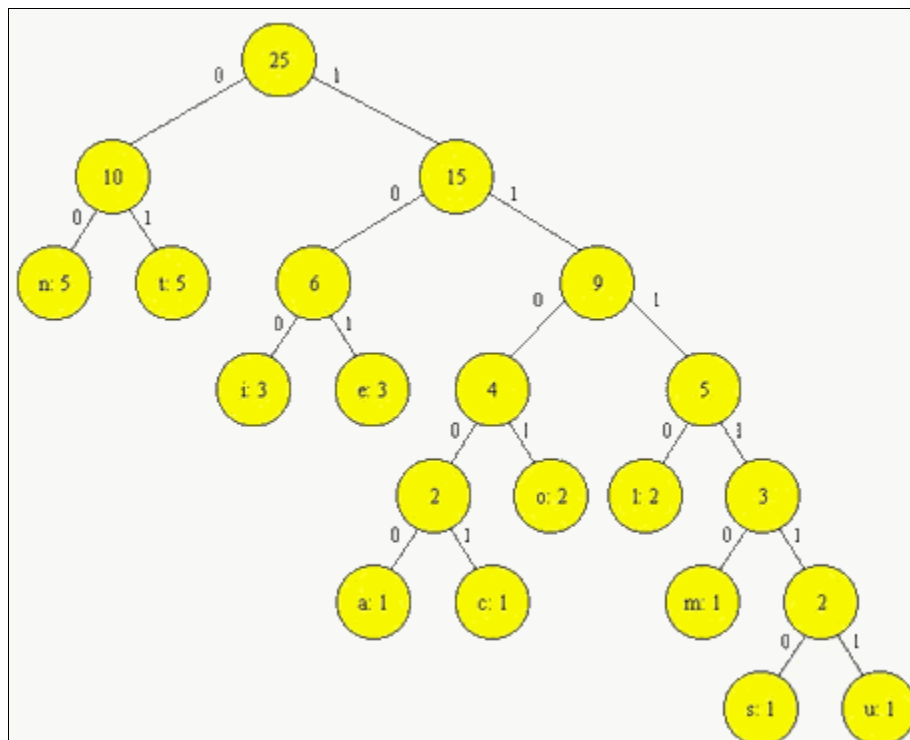


Figure: 62.11 - Dixième itération

Et voilà le travail!

Maintenant, le code associé à chaque Clé n'est autre que le chemin d'accès au noeud terminal correspondant, depuis la racine, en notant 0 chaque branche de gauche et 1 chaque branche de droite.

Finalement:

Clé	Code binaire
n	00
t	01
i	100
e	101
a	11000
c	11001

o	1101
l	1110
m	11110
s	111110
u	111111

Tableau: 62.7 - Correspondance lettre/code binaire

Et voici maintenant, transcrit avec notre nouveau code, le mot de départ:

11000000110011001110100111110011000111111011001101000010111101110101111101010001

ce qui nous fait 81 bits, au lieu de 200 au départ! Cela correspond à un taux de compression de 60%.

Le fait d'avoir généré un code en se servant d'un arbre binaire assure qu'aucun code ne peut être le préfixe d'un autre. Vous pouvez vérifier qu'à l'aide de la table de codage, il n'y a aucune ambiguïté possible pour décoder notre mot compressé!

En pratique, la table de codage étant spécifique à chaque fichier, il est indispensable de l'incorporer au fichier compressé, de manière à ce que le décryptage soit possible. Ce qui signifie que la taille du fichier compressé doit être augmentée d'autant. Dans le cas de notre fichier exemple, il faudrait incorporer au minimum 22 octets de plus pour insérer la table de codage, et le taux de compression n'est plus aussi bon.

Toutefois, pour des fichiers suffisamment larges (à partir de quelques kilo-octets) le surplus de taille généré par la table de codage devient négligeable par rapport à l'ensemble du fichier. Concrètement, l'algorithme de Huffman permet d'obtenir des taux de compression typiques compris entre 30% et 60%. Sans être aussi bien que ce que réalise Winzip (en moyenne 20% de mieux), c'est déjà pas mal!

3.2. ALGORITHME DE SARDINAS ET PATTERSON

Face à de longs codes, la difficulté est de vérifier si un code en est vraiment un... Pour cela, nous pouvons utiliser l'algorithme de Sardinas et Patterson (la démonstration de cet algorithme se fera lors de la prochaine mise à jour de ce chapitre).

Pour ce faire, nous construisons un graphe $G(X) = (P, U)$, où P est l'ensemble des préfixes non vides (selon la définition des "préfixes" donnée plus haut) de mots de X , et U l'ensemble des couples (u, v) tels que l'un ou l'autre des possibilités suivantes soient satisfaites:

P1. $uv \in X$ (en éliminant les couples doubles au besoin)

ou

P2. $v \notin X$ et il existe $x \in X$ tel que $ux = v$

Exemple:

Pour $X = \{a, bb, abbbba, babab\}$, l'ensemble P contient, en plus de X (qui sont préfixes de ε), les mots $\{b, ab, abb, abbb, b, ba, bab, baba\}$ (respectivement préfixes de $\{b, bba, ba, a, abab, bab, ab, b\}$)

D'abord nous voyons de suite que l'ensemble $X = \{a, bb, abbbba, babab\}$ n'est pas un code parce que:

$$a \cdot bb \cdot babab \cdot abbbba = abbbba \cdot babab \cdot bb \cdot a \quad (62.40)$$

Maintenant, les couples de U sont pour la première possibilité P1:

$$(b, b), (abbb, a), (abb, ba), (ba, bab), (bab, ab), (baba, b) \quad (62.41)$$

et pour la deuxième possibilité P2:

$$(a, abb), (ab, abbb), (b, ba), (bab, baba) \quad (62.42)$$

pour lesquels le x qui sert à former le v vaut respectivement bb, bb, a, a).

Le graphique de Sardinas et Patterson sera:

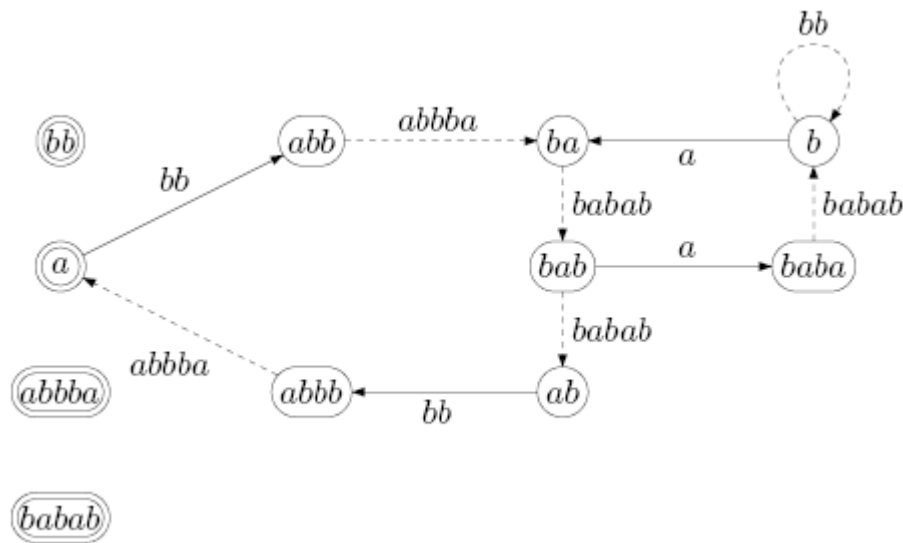


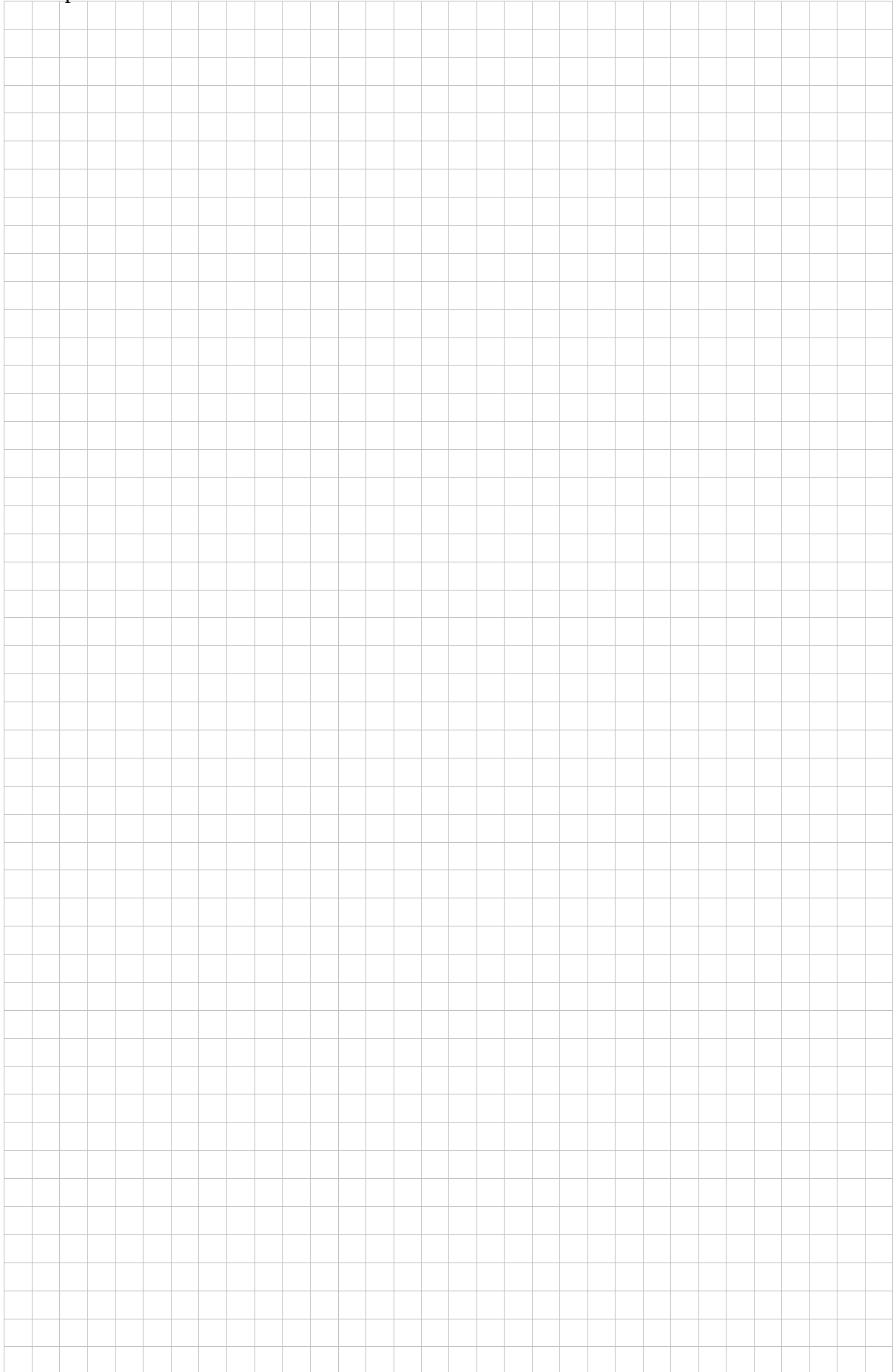
Figure: 62.12 - Graphe de Sardinas et Patterson correspondant

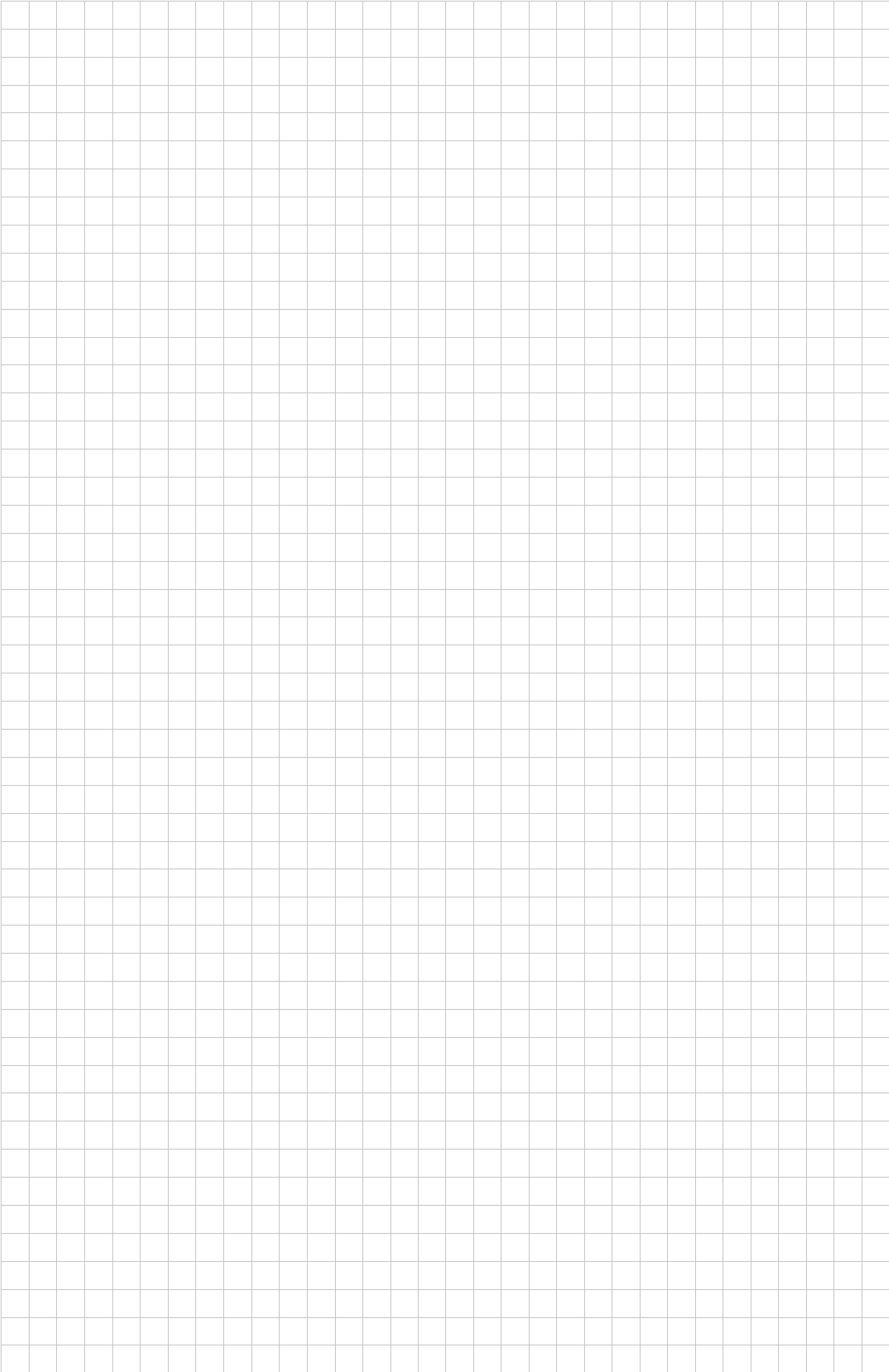
où les sommets correspondants aux mots de X sont doublement cerclés. L'étiquette de chaque arc est un mot de l'ensemble X . Les "arcs croisés" sont tracés en pointillés et les "arcs avant" en traits pleins. Si l'arc (u, v) est croisé, alors l'étiquette est uv , sinon c'est le mot x tel que $ux=v$.

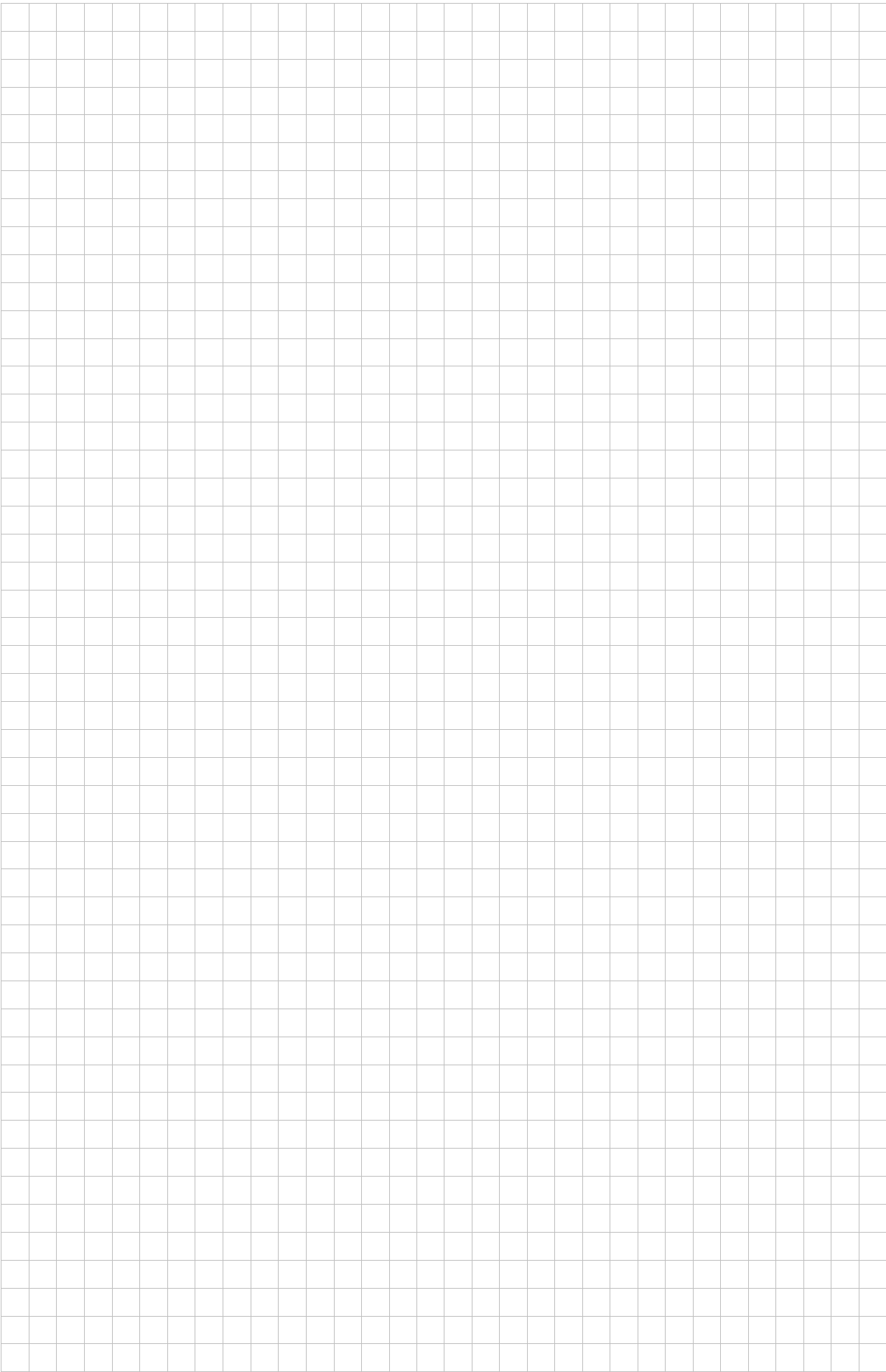
Dans notre exemple, il y a un chemin qui part de a et arrive en a . En vertu du théorème de Sardinas et Patterson (que nous démontrerons lors de la prochaine mise à jour de ce chapitre), l'ensemble X n'est pas un code (il suffit d'un seul et unique chemin entre deux sommets quelconques de X).

L'ensemble X est un code si et seulement si il n'y a pas de chemin non trivial dans $\mathcal{G}(X)$ d'un sommet de X à un sommet de X (en d'autres termes, un ensemble X est un code si et seulement si il n'y a que le seul et unique chemin trivial qui mène d'un sommet de X à un autre - ce sommet pouvant être le même comme dans l'exemple précédent).

Notes personnelles:









63. INFORMATIQUE QUANTIQUE

L'informatique quantique (nous devrions plutôt parler de "calcul quantique" car nous sommes actuellement très loin d'un système d'entrée/sortie) est un exemple royal de l'utilisation des spécificités des modèles théoriques de la physique quantique pour le traitement et la transmission de l'information.

Toutefois il faut aussi se rappeler que le comportement des transistors gravés sur la puce de votre ordinateur n'a pu être imaginé en 1947 par Bardeen, Brattain et Shockley qu'à partir de leurs connaissances en physique quantique. Donc la totalité de nos appareils électroniques fonctionnant déjà sur la base de semi-conducteurs fonctionnent à l'aide de développements obtenus grâce à la physique quantique.

La grande nouveauté, depuis le début des années 1980, est la possibilité pour les physiciens de manipuler et d'observer des objets quantiques élémentaires individuels: photons, atomes, ions, etc. C'est cette possibilité de manipuler et d'observer des objets quantiques élémentaires qui est à l'origine de l'information quantique, où ces objets quantiques élémentaires permettront de construire physiquement les qubits. Cela dit, aucun concept fondamentalement nouveau n'a été introduit depuis les années 1930, et les pères fondateurs de la physique quantique (Heisenberg, Schrödinger, Dirac, Planck, Einstein,...), s'ils ressuscitaient aujourd'hui, ne seraient pas surpris par l'informatique quantique, même s'ils seraient sûrement éblouis par les prouesses des expérimentateurs qui réalisent aujourd'hui des expériences qualifiées à l'époque de "gedanken experiment" (expérience imaginaire).

Il vaut aussi la peine de signaler que la miniaturisation croissante de l'électronique va trouver ses limites en raison des effets quantiques, qui vont devenir incontournables en dessous du nanomètre. Ainsi, nous estimons que la loi de Moore (hypothèse selon quoi la puissance de calcul des machines double à peu près tous les 18 mois) pourrait ne plus être valable d'ici 2015-2020.

Il est fort à parier que la mode de l'étude de la physique quantique et son application à l'informatique quantique (et l'électronique quantique et la télécommunication quantique) va exploser dans les décennies à venir (surtout vers la fin du 21ème siècle). Ainsi, les écoles d'ingénieurs intégreront presque dans tous les domaines la physique quantique dans les programmes scolaires. Ce que les physiciens étudient depuis bientôt déjà presque 100 ans dans leur cursus.

Avant de passer au côté formel, nous avons jugé cependant intéressant de faire un petit passage vulgarisé car nous avons remarqué que cela aide à comprendre les calculs qui seront faits par la suite.

Dans les années 70 et 80, les premiers ordinateurs quantiques naissent par retournement dans l'esprit de physiciens tels que Richard Feynman, Paul Bénioff, David Deutsch ou Charles Bennett. L'idée de Feynman était qu'au lieu de nous plaindre que la simulation des phénomènes quantiques demande des puissances énormes à nos ordinateurs actuels, que nous utilisions alors la puissance de calcul des phénomènes quantiques pour faire plus rapides que nos ordinateurs actuels.

Longtemps les physiciens ont douté que les calculateurs quantiques utilisables puissent exister, et même que nous puissions en faire quelque chose de viable s'ils existaient. Mais:

- En 1994, Peter Shor, un scientifique d'AT&T montre qu'il est possible de factoriser des grands nombres dans un temps raisonnable à l'aide d'un calculateur quantique. Cette découverte débloque brusquement des crédits.

- En 1996, Lov Grover, invente un algorithme basé sur les calculateurs quantiques permettant de trouver une entrée dans une base de données non triée.
- En 1998, IBM est le premier à présenter un ordinateur quantique de 2 qubits (pour "Quantum Bit").
- En 1999, l'équipe d'IBM utilise l'algorithme de Grover pour la recherche quantique rapide sur une base de données (quantum database search) sur un ordinateur de 3 qubits et battent leur record l'année suivante avec un ordinateur de 5 qubits.
- En 2001, IBM crée un ordinateur quantique de 7 qubits et factorise le nombre 15... grâce à l'algorithme de Shor (cf. chapitre de Méthodes Numériques). Les ordinateurs à 7 qubits sont bâtis autour de molécules de chloroforme et leur durée de vie utile ne dépasse pas quelques minutes.
- En 2007, la compagnie canadienne D-Wave lors d'une démonstration a présenté un ordinateur quantique à 16 qubits.

La mémoire d'un ordinateur classique est faite donc de bits (cf. chapitre de Systèmes Logiques). Chaque bit porte soit un 1 soit un 0 (mode bipolaire). La machine calcule en manipulant ces bits. Un ordinateur quantique travaille sur un jeu de qubits. Un qubit peut porter soit un 1, soit un 0, soit une superposition d'un 1 et d'un 0 (ou, plus exactement, il porte une distribution de phase). Le ordinateur quantique calcule en manipulant ces distributions comme nous le verrons dans les détails plus loin.

Interroger un qubit dont la phase n'est pas de 0° ou de 90° ne sert pas à grand-chose: nous obtiendrons la réponse 0 avec une probabilité donnée, et la réponse 1 avec une autre probabilité... et il est possible de construire des générateurs aléatoires bien moins onéreux! En revanche, si nous arrivons à créer un algorithme qui le conduit systématiquement à une phase 0° ou 90° , nous obtiendrons un résultat déterministe. Encore faut-il que celui-ci corresponde à une réponse cherchée.

Un ordinateur quantique pourrait être implémenté à partir de toute particule pouvant avoir deux états. Il peut être construit à partir de photons, ou à partir de n'importe quelle particule ou atome comportant un spin.

Comme nous le savons, un ordinateur classique ayant trois bits de mémoire peut stocker uniquement des nombres de trois chiffres composés de uns ou de zéros digitaux (cf. chapitre de Systèmes Logiques) pour un total de 8 états qu'il doit traiter à part. À un moment donné, il pourrait contenir les bits 101.

Un ordinateur quantique ayant trois qubits peut en fait stocker 16 valeurs, assemblées deux par deux pour former $8 (2^3)$ nombres complexes. Il pourrait, par exemple, contenir ceci (nous démontrerons bien sûr d'où vient cela un peu plus loin!) à un instant donné:

État	Amplitude	Probabilité
	$a+ib$	$(a^2 + b^2)$
000	0.37+i0.04	0.14
001	0.11+i0.18	0.04
010	0.09+i0.31	0.10
011	0.03+i0.30	0.18
100	0.35+i0.43	0.31
101	0.40+i0.01	0.16
110	0.09+i0.12	0.02
111	0.15+i0.16	0.05

Tableau: 63.1 - Exemple de valeurs de qubits

La première colonne montre tous les états possibles pour trois bits. Un ordinateur classique peut donc seulement porter un de ces états à la fois. Un ordinateur quantique, lui, peut être dans une superposition de ces 8 états à la fois. La deuxième colonne montre l'amplitude pour chacun des 8 états. Ces 8 nombres complexes sont un instantané du contenu d'un ordinateur quantique à un moment donné. Durant le calcul, ces trois nombres changeront et interagiront les uns avec les autres. En ce sens, un ordinateur quantique à trois qubits a bien plus de mémoire qu'un ordinateur classique à trois bits.

Cependant, il n'est pas possible de voir directement ces trois nombres. Quand l'algorithme est fini, une seule mesure est accomplie. La mesure retourne une simple chaîne (string) de 3 bits et efface les 8 nombres quantiques. De plus, le report de calculs intermédiaire doit utiliser l'intrication quantique ce qui est loin d'être simple.

La troisième colonne donne la probabilité pour chacune des chaînes possibles. Dans cet exemple, il y a 14% de chance que la chaîne retournée soit 000, 4% que ce soit 001 et ainsi de suite. Chaque nombre complexe est nommé "**amplitude**" et chaque probabilité une "**amplitude carrée**". La somme des huit probabilités est égale à un.

Typiquement, un algorithme d'un ordinateur quantique initialisera tous les nombres complexes à des valeurs égales, donc tous les états auront les mêmes probabilités. La liste des nombres complexes peut être imaginée comme un vecteur à 8 éléments. À chaque étape de l'algorithme, le vecteur est modifié par son produit avec une matrice. La matrice provient de la physique de la machine et sera toujours inversible, et s'assurera que la somme des probabilités continue à être égale à un.

Il faut d'abord réaliser un calcul conduisant à un état non superposé. En effet, si on interroge un qubit qui se trouve dans une superposition d'états, la réponse sera aléatoire et ne nous apprendra pas grand-chose. Il faut donc trouver un algorithme donnant une réponse unique pour tous les chemins de calcul possibles. C'est un problème semblable à celui des énigmes où l'on doit obtenir une réponse toujours vraie en la posant à une série d'intermédiaires dont on sait que certains mentent toujours et d'autres jamais. La question est donc de trouver un calcul parvenant à cet invariant, par exemple dans le cas du cassage d'un code la clé de chiffrement que l'on cherche à déterminer.

Un autre problème ensuite est de le mesurer: la lecture d'un seul bit d'un état quantique détruit la totalité de cet état. Il faudra donc refaire le calcul autant de fois que la réponse souhaitée comporte de bits, mais le temps correspondant sera juste proportionnel à ce nombre de bits et non exponentiellement

plus grand, ce qui est justement le but recherché.

Commençons par comprendre les concepts sous-jacents à la théorie quantique de l'informatique quantique avec une étude de la polarisation du photon.

1. POLARISATION DU PHOTON

Depuis Einstein, nous savons que la lumière est composée de photons, ou particules de lumière et celle-ci a un aspect dual onde-particule (cf. chapitre d'Optique Ondulatoire). Si nous réduisons l'intensité lumineuse d'un faisceau de photons, nous devrions pouvoir étudier la polarisation des photons individuels, que nous savons parfaitement détecter à l'aide de photomultiplicateurs. Supposons que l'expérience détecte N photons. Lorsque $N \rightarrow \infty$, nous devons retrouver les résultats de l'optique ondulatoire (voir chapitre du même nom).

Effectuons par exemple l'expérience suivante:

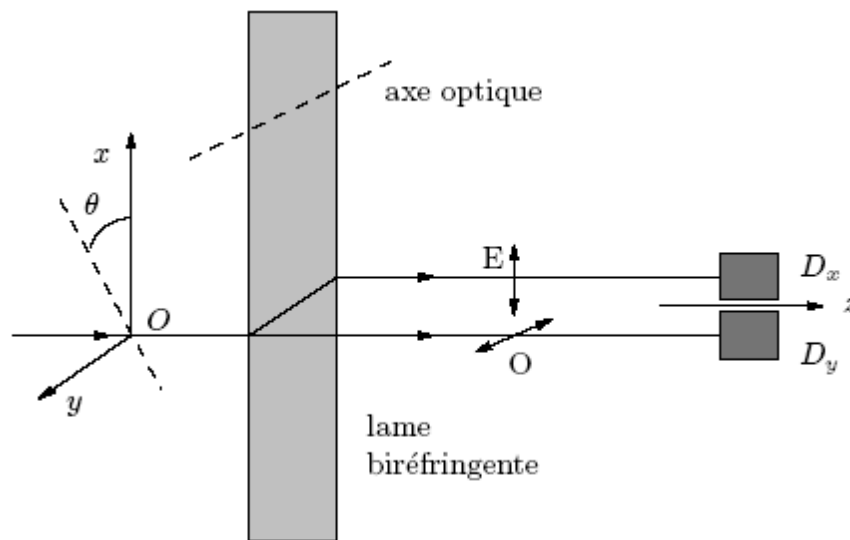


Figure: 63.1 - Expérience imaginaire pour la mesure de la polarisation

Une lame biréfringente sépare un faisceau lumineux dont la polarisation fait un angle θ avec Ox en un faisceau polarisé suivant Ox et un faisceau polarisé suivant Oy , les intensités étant respectivement $I \cos^2 \theta$ et $I \sin^2 \theta$ (selon la démonstration de la loi de Malus faite dans le chapitre d'Optique Ondulatoire).

Réduisons l'intensité de telle sorte que les photons arrivent un à un, et plaçons deux photodétecteurs D_x, D_y derrière la lame. L'expérience montre que D_x, D_y ne cliquent jamais simultanément (sauf cas de "dark count" où un compteur se déclenche spontanément): un photon arrive entier soit sur D_x , soit sur D_y , un photon ne se divise donc pas. D'autre part, l'expérience montre que la probabilité P_x (P_y) de détection d'un photon par D_x (D_y) est de $\cos^2 \theta$ ($\sin^2 \theta$). Ainsi, si l'expérience détecte N photons, nous aurons donc N_x (N_y) photons détectés par D_x (D_y):

$$N_x \cong N \cos^2 \theta \quad N_y \cong N \sin^2 \theta \quad (63.1)$$

où le $\cong$ tient compte des fluctuations statistiques. Comme l'intensité lumineuse est proportionnelle au

nombre de photons, nous retrouvons bien la loi de Malus à la limite $N \rightarrow \infty$.

Cependant, nous notons deux problèmes:

1. Pouvons-nous prévoir, pour un photon donné, s'il va déclencher D_x ou D_y ? La réponse de la théorie quantique est NON, énoncé qui a profondément choqué Einstein (Dieu ne joue pas aux dés!). Certains physiciens (dont Einstein) ont été tentés de supposer que la théorie quantique était incomplète, et qu'il y avait des "variables cachées" dont la connaissance permettrait de prévoir le sort individuel de chaque photon. Moyennant des hypothèses très raisonnables sur lesquelles nous reviendrons, nous savons aujourd'hui que de telles variables cachées sont exclues. Les probabilités de la théorie quantique sont, nous le savons (cf. chapitre de Physique Quantique Ondulatoire), intrinsèques! Elles ne sont pas liées à une connaissance imparfaite de la situation physique, comme c'est le cas par exemple dans le jeu de pile ou face.
2. Si nous recombinaisons les deux faisceaux de la première lame biréfringente, en utilisant une seconde lame symétrique à la première:

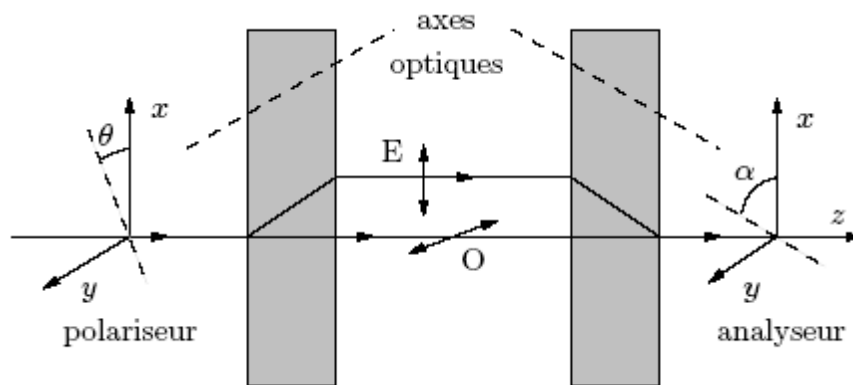


Figure: 63.2 - Expérience imaginaire recombinaison des deux faisceaux

et si nous cherchons la probabilité qu'un photon traverse l'analyseur, un photon peut choisir le trajet x avec une probabilité $\cos^2 \theta$, il a ensuite une probabilité $\cos^2 \alpha$ de traverser l'analyseur soit une probabilité totale $\cos^2 \theta \cos^2 \alpha$. S'il choisit le trajet y , il aura une probabilité $\sin^2 \theta \sin^2 \alpha$ de traverser l'analyseur. La probabilité totale s'obtient donc en additionnant les probabilités des deux trajets possibles:

$$P'_{\text{tot}} = \cos^2 \theta \cos^2 \alpha + \sin^2 \theta \sin^2 \alpha \quad (63.2)$$

Ce résultat est FAUX! En effet, l'optique classique nous apprend que l'intensité est $I \cos^2 (\theta - \alpha)$ (cf. chapitre d'Optique Ondulatoire) et le résultat correct, confirmé par l'expérience est:

$$P_{\text{tot}} = \cos^2 (\theta - \alpha) \quad (63.3)$$

Ce qui n'est pas du tout la même chose!

En fait, pour retrouver les résultats de l'optique ondulatoire, il faut se rappeler que la probabilité en physique quantique s'obtient par la norme au carré de l'amplitude de probabilité (cf. chapitre de Physique Quantique Ondulatoire). Ainsi:

$$\begin{aligned} \alpha(\theta \rightarrow x) &= \cos \theta & \alpha(x \rightarrow \alpha) &= \cos \alpha \\ \alpha(\theta \rightarrow y) &= \sin \theta & \alpha(y \rightarrow \alpha) &= \sin \alpha \end{aligned} \quad (63.4)$$

et nous devons additionner les amplitudes pour des trajets indiscernables et en utilisant les relations trigonométriques de base, nous obtenons:

$$\alpha_{wt} = \cos \theta \cos \alpha + \sin \theta \sin \alpha = \cos(\theta - \alpha) \quad (63.5)$$

ce qui redonne bien:

$$P_{wt} = \|\alpha_{wt}\|^2 = \cos^2(\theta - \alpha) \quad (63.6)$$

Supposons que nous ayons un moyen de savoir si le photon emprunte le trajet x ou le trajet y (impossible dans notre cas, mais des expériences analogiques répondant à la question "Quel trajet?" ont été réalisées avec des atomes). Nous pourrions alors diviser les photons en deux classes, ceux qui ont choisi le trajet x et ceux qui ont choisi le trajet y .

Pour les photons ayant choisi le trajet x , nous pourrions bloquer le trajet y par un cache sans rien changer, et inversement pour les photons ayant choisi le trajet y nous pourrions bloquer le trajet x . Bien évidemment, le résultat ne peut être alors que P'_{wt} . Si nous arrivons à discriminer entre les trajets, le résultat n'est plus le même, les trajets ne sont plus indiscernables. Dans les conditions expérimentales où il est impossible en principe de distinguer entre les trajets, nous pouvons dire au choix:

1. Soit que le photon emprunte les deux trajets à la fois (...)
2. Soit que cela n'a pas de sens de poser la question "Quel trajet?", puisque les conditions expérimentales ne permettent pas d'y répondre.

Il faut noter que si l'expérience permet de décider entre les deux trajets, le résultat est P'_{wt} , même si nous décidons de ne pas les observer. Il suffit que les conditions expérimentales permettent, en principe, de distinguer entre les deux trajets.

2. QUBIT

Nous pouvons utiliser la polarisation des photons pour transmettre de l'information, par exemple par une fibre optique. Nous décidons tout à fait arbitrairement, d'attribuer la valeur 1 du bit à un photon polarisé suivant Ox et la valeur 0 à un photon polarisé suivant Oy .

Pour étudier la théorie, il est devenu traditionnel de se représenter que les deux personnes qui échangent de l'information sont appelées conventionnellement Alice (A) et Bob (B). Alice envoie par exemple à Bob la suite de photons polarisés suivante:

$$yyxyxyyyx... \quad (63.7)$$

Bob analyse la polarisation de ces photons à l'aide d'une lame biréfringente et en déduit le message d'Alice:

$$001010001... \quad (63.8)$$

Ce n'est évidemment pas une façon très efficace d'échanger des messages, mais c'est à la base de la cryptographie quantique (cf. chapitre de Cryptographie). Cependant, la question intéressante est maintenant: quelle est la valeur du bit que nous pouvons attribuer par exemple à un photon polarisé à 45° ? Suivant les résultats précédents, un photon polarisé à 45° est une superposition linéaire d'un photon polarisé suivant O_x et d'un photon polarisé suivant O_y . Un qubit est donc une entité beaucoup plus riche qu'un bit ordinaire, qui ne peut prendre en logique stricte que les valeurs 0 et 1.

En un certain sens, un qubit peut prendre toutes les valeurs intermédiaires entre 0 et 1 et contiendrait donc une quantité infinie d'information ! Cependant, cet énoncé optimiste est immédiatement démenti lorsque nous nous rendons compte que la mesure du qubit ne peut donner que le résultat 0 ou 1, quelle que soit la base choisie. Malgré tout, nous pouvons nous poser la question de cette "information cachée" dans la superposition linéaire et nous verrons que nous pouvons l'exploiter sous certaines conditions.

Afin de rendre compte de la possibilité des superpositions linéaires, il est naturel d'introduire pour la description mathématique de la polarisation un espace vectoriel complexe (cause: phaseurs) à deux dimensions correspondant au plan de polarisation comme nous l'avons vu dans le chapitre d'Optique Ondulatoire. Nous noterons cet espace vectoriel $\mathcal{H}$ (nous reprenons la notation des espaces de Hilbert) et l'appellerons "**l'espace de Hilbert des états de polarisation**".

Nous pouvons très bien décomposer le vecteur correspondant aux polarisations linéaires O_x et O_y en deux vecteurs kets $|x\rangle$ et $|y\rangle$ tel que tout état de polarisation (qu'il soit linéaire, circulaire ou autre) pourra se décomposer suivant cette base:

$$|\Phi\rangle = \lambda|x\rangle + \mu|y\rangle \quad (63.9)$$

Ainsi, une polarisation linéaire sera décrite par des coefficients λ, μ réels, mais la description d'une polarisation circulaire ou elliptique exigera bien évidemment de faire appel à des coefficients complexes!

Les amplitudes de probabilité vont correspondre à un produit scalaire sur cet espace. Soit deux vecteurs correspondant donc à deux polarisations différentes:

$$\begin{aligned} |\Phi\rangle &= \lambda|x\rangle + \mu|y\rangle \\ |\Psi\rangle &= \nu|x\rangle + \sigma|y\rangle \end{aligned} \quad (63.10)$$

Le produit scalaire hermitique (cf. chapitre de Calcul Vectoriel) sera donc:

$$\langle\Psi|\Phi\rangle = \bar{\nu}\lambda + \bar{\sigma}\mu = \overline{\langle\Psi|\Phi\rangle} \quad (63.11)$$

Maintenant, un état de polarisation linéaire (cf. chapitre d'optique Ondulatoire) suivant θ sera donné logiquement par (si nous nous restreignons au cas linéaire donc!):

$$|\theta\rangle = \cos\theta|x\rangle + \sin\theta|y\rangle \quad (63.12)$$

où $|x\rangle, |y\rangle$ sont des vecteurs de norme unité. Ce qui est conforme à la représentation mentale:

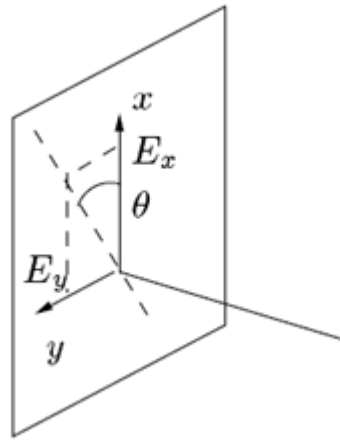


Figure: 63.3 - Rappel du principe de décomposition du champ

où l'amplitude du champ est normalisée à l'unité.

L'amplitude de probabilité pour qu'un photon polarisé suivant θ traverse un analyseur orienté suivant α pourra maintenant s'écrire:

$$\langle \alpha | \theta \rangle = \cos \alpha \cos \theta + \sin \alpha \sin \theta = \cos(\theta - \alpha) \quad (63.13)$$

et la probabilité de traverser l'analyseur sera toujours donnée par la norme au carré de cette amplitude comme nous l'avons démontré plus haut:

$$\|\langle \alpha | \theta \rangle\|^2 = \cos^2(\theta - \alpha) \quad (63.14)$$

De façon générale, nous définirons des amplitudes de probabilité, où $|\Phi\rangle, |\Psi\rangle$ sont des états de polarisation:

$$\alpha(\Phi \rightarrow \Psi) = \langle \Psi | \Phi \rangle \quad (63.15)$$

et la probabilité correspondante sera:

$$p(\Phi \rightarrow \Psi) = \|\alpha(\Phi \rightarrow \Psi)\|^2 = \|\langle \Psi | \Phi \rangle\|^2 \quad (63.16)$$

Nous sommes maintenant prêts à aborder la question cruciale de la mesure dans le cadre de cette expérience quantique. Reprenons l'ensemble polariseur/analyseur, en supposant que l'analyseur soit orienté suivant Ox . Si le polariseur est aussi orienté suivant Ox , un photon sortant du polariseur traverse l'analyseur avec une probabilité de 100%; si le polariseur est orienté suivant Oy , la probabilité est nulle. L'analyseur effectue un test (de la polarisation), et le résultat du test est 1 ou 0. Le test permet donc de connaître l'état de polarisation du photon.

Mais ceci n'est pas le cas général!

Supposons que le polariseur soit orienté suivant la direction générale θ ou la direction orthogonale $\theta_\perp$ (il y a une rotation de $\pi/2$ [rad]). Nous avons alors en utilisant les propriétés du cercle trigonométrique:

$$\begin{aligned} |\theta\rangle &= \cos \theta |x\rangle + \sin \theta |y\rangle \\ |\theta_\perp\rangle &= -\sin \theta |x\rangle + \cos \theta |y\rangle \end{aligned} \quad (63.17)$$

et donc si le polariseur prépare par exemple le photon dans l'état $|\theta\rangle$ et que l'analyseur est orienté suivant Ox , la probabilité de réussite du test sera toujours $\cos^2 \theta$ quel que soit le type de polarisation!! Rappelons que dans cet exemple, après le passage dans l'analyseur, l'état de polarisation du photon n'est plus $|\theta\rangle$, mais $|x\rangle$. La mesure modifie donc l'état de polarisation.

Remarque: Bien sûr, une autre manière de voir que les deux vecteurs plus hauts sont orthogonaux est d'effectuer un produit scalaire et de constater que celui-ci est nul.

Nous constatons une différence de principe entre la mesure en physique classique et la mesure en physique quantique. En physique classique, la quantité physique à mesurer préexiste à la mesure: si un radar mesure la vitesse de votre voiture à 180 Km/h sur l'autoroute, cette vitesse préexistait à sa mesure par le gendarme. Au contraire, dans la mesure de polarisation d'un photon $|\theta\rangle$ par un analyseur orienté suivant Ox , le fait que le test donne une polarisation suivant Ox ne permet pas de conclure que le photon testé avait au préalable sa polarisation suivant Ox .

Donc nous avons un dispositif préparant le système quantique dans l'état $|\Phi\rangle$ et un second capable de le "préparer" dans l'état $|\Psi\rangle$ que nous utiliserons comme analyseur. Après le test, le système quantique sera donc dans l'état $|\Psi\rangle$, ce qui veut dire du point de vue mathématique que nous réalisons une projection orthogonale sur $|\Psi\rangle$.

Soit $\mathcal{P}_\Psi$ ce projecteur, alors la projection orthogonale vectorielle (cf. chapitre de Calcul Vectoriel) est donnée par:

$$\mathcal{P}_\Psi |\Phi\rangle = |\Psi\rangle \langle \Psi | \Phi \rangle \quad (63.18)$$

ce qui consiste (pour rappel) en un simple produit scalaire (projection orthogonale scalaire) multiplié par le vecteur $|\Psi\rangle$. Ceci se voit aisément en plaçant judicieusement les parenthèses:

$$\mathcal{P}_\Psi |\Phi\rangle = (|\Psi\rangle \langle \Psi|) |\Phi\rangle \quad (63.19)$$

et donc:

$$\mathcal{P}_\Psi = |\Psi\rangle \langle \Psi| \quad (63.20)$$

La projection du vecteur d'état est appelée, comme nous l'avons déjà vu (cf. chapitre de Physique Quantique Ondulatoire), dans l'interprétation de Copenhague de la physique quantique "**réduction du vecteur d'état**", ou, pour des raisons historiques, "**réduction du paquet d'ondes**". Cette réduction du vecteur d'état est une fiction commode de l'interprétation de Copenhague, qui évite d'avoir à se poser des questions sur le processus de mesure...

Le lecteur habitué à l'algèbre linéaire (voir chapitre du même nom) remarquera trivialement que nous pouvons manipuler la convention de notation du projecteur comme une matrice (application linéaire)

telle que dans deux cas particuliers simples (ceux qui nous intéressent!):

$$\mathcal{P}_x = |x\rangle\langle x| = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \mathcal{P}_y = |y\rangle\langle y| = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \quad (63.21)$$

Un lecteur nous ayant demandé de préciser la démarche, voyons comme nous arrivons à cet aspect matriciel de la projection orthogonale avec un exemple particulier de deux vecteurs dans un espace réel à deux dimensions. Considérons:

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \Phi = \begin{pmatrix} \Phi_1 \\ \Phi_2 \end{pmatrix} \quad (63.22)$$

et donc (la démarche est la même pour y):

$$\begin{aligned} |x\rangle\langle x|\Phi\rangle &= \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \circ \begin{pmatrix} \Phi_1 \\ \Phi_2 \end{pmatrix} \right) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} (x_1 \Phi_1 + x_2 \Phi_2) = \begin{pmatrix} x_1 (x_1 \Phi_1 + x_2 \Phi_2) \\ x_2 (x_1 \Phi_1 + x_2 \Phi_2) \end{pmatrix} \\ &= \begin{pmatrix} x_1^2 \Phi_1 + x_1 x_2 \Phi_2 \\ x_2 x_1 \Phi_1 + x_2^2 \Phi_2 \end{pmatrix} = \begin{pmatrix} x_1^2 \Phi_1 + x_1 x_2 \Phi_2 \\ x_2 x_1 \Phi_1 + x_2^2 \Phi_2 \end{pmatrix} = \underbrace{\begin{bmatrix} x_1^2 & x_1 x_2 \\ x_1 x_2 & x_2^2 \end{bmatrix}}_{\mathcal{P}_x} \begin{pmatrix} \Phi_1 \\ \Phi_2 \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} \Phi_1 \\ \Phi_2 \end{pmatrix} \end{aligned} \quad (63.23)$$

Remarque: L'écriture matricielle du projecteur orthogonal est souvent présentée en tant qu'une définition d'un outil mathématique appelé "**produit extérieur**".

Donc finalement, pour en revenir à nos moutons, nous avons:

$$\mathcal{P}_x|\Phi\rangle = |x\rangle\langle x|\Phi\rangle = |x\rangle 1 \cdot \Phi_x = \begin{pmatrix} \Phi_x \\ 0 \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \Phi \quad (63.24)$$

et idem pour l'autre composante.

Comme:

$$|\theta\rangle = \cos \theta |x\rangle + \sin \theta |y\rangle \quad (63.25)$$

Nous avons alors:

$$\begin{aligned} \mathcal{P}_\theta|\Phi\rangle &= |\theta\rangle\langle\theta|\Phi\rangle = |\theta\rangle (\cos \theta \Phi_x + \sin \theta \Phi_y) = \begin{pmatrix} \cos^2 \theta \Phi_x + \cos \theta \sin \theta \Phi_y \\ \sin \theta \cos \theta \Phi_x + \sin^2 \theta \Phi_y \end{pmatrix} \\ &= \begin{bmatrix} \cos^2 \theta & \sin \theta \cos \theta \\ \sin \theta \cos \theta & \sin^2 \theta \end{bmatrix} \begin{pmatrix} \Phi_x \\ \Phi_y \end{pmatrix} \end{aligned} \quad (63.26)$$

Nous remarquerons que l'opérateur identité peut être écrit comme la somme des deux projecteurs $\mathcal{P}_x, \mathcal{P}_y$:

$$\mathcal{P}_x + \mathcal{P}_y = |x\rangle\langle x| + |y\rangle\langle y| = \mathbb{1} \quad (63.27)$$

relation dite "**relation de fermeture**", qui se généralise à une base orthonormée d'un espace de Hilbert $\mathcal{H}$ de dimension N :

$$\sum_{i=1}^N |i\rangle\langle i| = \mathbb{1} \quad \langle i|j\rangle = \delta_{ij} \quad (63.28)$$

Par ailleurs, les projecteurs $\mathcal{P}_x, \mathcal{P}_y$ commutent (vérification triviale):

$$[\mathcal{P}_x, \mathcal{P}_y] = \mathcal{P}_x \mathcal{P}_y - \mathcal{P}_y \mathcal{P}_x = 0 \quad (63.29)$$

Ainsi, les tests $|x\rangle, |y\rangle$ sont compatibles (quel que soit le sens de la mesure le résultat en est indépendant). En revanche, les projecteurs $\mathcal{P}_\theta, \mathcal{P}_{\theta\perp}$:

$$\mathcal{P}_\theta = |\theta\rangle\langle\theta| = \begin{bmatrix} \cos^2 \theta & \sin \theta \cos \theta \\ \sin \theta \cos \theta & \sin^2 \theta \end{bmatrix} \quad \mathcal{P}_{\theta\perp} = |\theta_\perp\rangle\langle\theta_\perp| = \begin{bmatrix} \sin^2 \theta & -\sin \theta \cos \theta \\ -\sin \theta \cos \theta & \cos^2 \theta \end{bmatrix} \quad (63.30)$$

qui satisfont (vérification triviale) à:

$$|\theta\rangle\langle\theta| + |\theta_\perp\rangle\langle\theta_\perp| = \mathbb{1} \quad \langle\theta|\theta_\perp\rangle = 0 \quad (63.31)$$

ainsi qu'à (vérification triviale):

$$\mathcal{P}_\theta |\theta\rangle = |\theta\rangle\langle\theta|\theta\rangle = |\theta\rangle \quad \mathcal{P}_{\theta\perp} |\theta_\perp\rangle = |\theta_\perp\rangle\langle\theta_\perp|\theta_\perp\rangle = |\theta_\perp\rangle \quad (63.32)$$

ne commutent pas avec $\mathcal{P}_x, \mathcal{P}_y$:

$$[\mathcal{P}_x, \mathcal{P}_\theta] = \begin{bmatrix} 0 & \sin \theta \cos \theta \\ -\sin \theta \cos \theta & 0 \end{bmatrix} \quad (63.33)$$

et donc les tests $|x\rangle, |\theta\rangle$ sont incompatibles.

Pour des développements ultérieurs, il sera utile de remarquer que la connaissance des probabilités de réussite d'un test T permet de définir une valeur moyenne (espérance):

$$\langle T \rangle = 1 \cdot P(T=1) + 0 \cdot P(T=0) = P(T=1) \quad (63.34)$$

En analogie avec le contexte, nous pouvons lire cela ainsi: l'espérance du test est égale à la valeur représentative du photon orienté selon Ox (correspondant à la valeur 1 arbitrairement) multipliée par la probabilité de passer l'analyseur orienté aussi selon Ox (donc test concluant à 100%) sommée à la valeur représentative du photon orienté selon Oy (correspondant à la valeur 0 arbitrairement) multiplié par la probabilité de passage l'analyseur toujours orienté selon Ox (donc 0% des photons Oy passeront le test Ox).

Par exemple, si le test T est représenté par la procédure $|\Psi\rangle$ et que nous l'appliquons à un état $|\Phi\rangle$ (contenant comme nous l'avons vu plus haut les valeurs représentatives du photon polarisé en linéaire

ou autre...) alors le test correspond à un produit scalaire:

$$P(\Psi) = \|\langle \Psi | \Phi \rangle\|^2 = \langle \Psi | \Phi \rangle \langle \Psi | \Phi \rangle = \langle \Phi | (|\Psi\rangle\langle\Psi|) | \Phi \rangle = \langle \Phi | \mathcal{P}_\Psi | \Phi \rangle \quad (63.35)$$

et comme nous l'avons vu dans le chapitre de Physique Quantique Ondulatoire, nous savons qu'on a fait:

$$\langle M \rangle_\Phi = \langle \Phi | \mathcal{P}_\Psi | \Phi \rangle \quad (63.36)$$

est la valeur moyenne d'un opérateur M dans l'état $|\Phi\rangle$. Ainsi, au test T auquel est associée une procédure $|\Psi\rangle$, nous pouvons associer le projecteur $\mathcal{P}_\Psi$ dont la valeur moyenne dans l'état $|\Phi\rangle$ donne la probabilité de réussite du test.

La généralisation de cette observation permet de construire les propriétés physiques d'un système quantique. Donnons un exemple en revenant au cas de la polarisation. Supposons que nous construisions (de manière tout à fait arbitraire) une propriété $\mathcal{M}$ d'un photon de la façon suivante:

$\mathcal{M}$ vaut +1 si le photon est polarisé suivant Ox

$\mathcal{M}$ vaut -1 si le photon est polarisé selon Oy .

Nous pouvons associer à la propriété physique $\mathcal{M}$ l'opérateur hermitique:

$$M = \mathcal{P}_x - \mathcal{P}_y \quad (63.37)$$

qui vérifie bien (trivial) la relation entre opérateur, valeur propre et vecteur:

$$\begin{aligned} M|x\rangle &= +1 \cdot |x\rangle = +|x\rangle \\ M|y\rangle &= -1 \cdot |y\rangle = -|y\rangle \end{aligned} \quad (63.38)$$

La valeur moyenne (espérance) de M étant alors (par définition):

$$\langle M \rangle = (+1) \cdot P(M=1) + (-1) \cdot P(M=-1) \quad (63.39)$$

Supposons le photon dans l'état de polarisation linéaire θ , alors la valeur moyenne $\langle M \rangle_\theta$ dans l'état $|\theta\rangle$ est (trivial):

$$\langle M \rangle_\theta = \langle \theta | \mathcal{P}_x | \theta \rangle - \langle \theta | \mathcal{P}_y | \theta \rangle = \cos^2 \theta - \sin^2 \theta = \cos(2\theta) \quad (63.40)$$

Avant de voir, comment nous pouvons construire un tel opérateur M avec un autre objet que le photon et ce avec les mêmes propriétés, introduisons un outil mathématique généralisant les conditions et la configuration d'une onde polarisée quelconque:

2.1. SPHÈRE DE BLOCH

La sphère de Bloch est comme nous allons le voir une représentation géométrique des états des qubits comme points de la surface d'une sphère.

Un certain nombre d'opérations élémentaires faites en informatique quantique peuvent sous le choix

d'un projecteur adéquat être réalisées avec cette sphère.

Nous allons montrer qu'un état d'un qubit arbitraire peut être écrit:

$$|\Psi\rangle = e^{i\gamma} \left(\cos \frac{\theta}{2} |0\rangle + e^{i\phi} \sin \frac{\theta}{2} |1\rangle \right) \quad (63.41)$$

où $\gamma \in \mathbb{R}$ et $0 \leq \theta \leq \pi$, $0 \leq \phi \leq 2\pi$ définissent un point sur la sphère tridimensionnelle de Bloch.

Les qubits représentés par des valeurs arbitraires de γ (invariance de jauge globale selon $U(1)$) sont tous représentés par le même point sur la sphère de Bloch parce que nous allons montrer que le facteur n'a pas d'effet observable et que nous pourrions alors écrire sans perdre en généralité:

$$|\Psi\rangle = \cos \frac{\theta}{2} |0\rangle + e^{i\phi} \sin \frac{\theta}{2} |1\rangle \quad (63.42)$$

ce qui est représenté comme nous le justifierons plus loin par la figure ci-dessous:

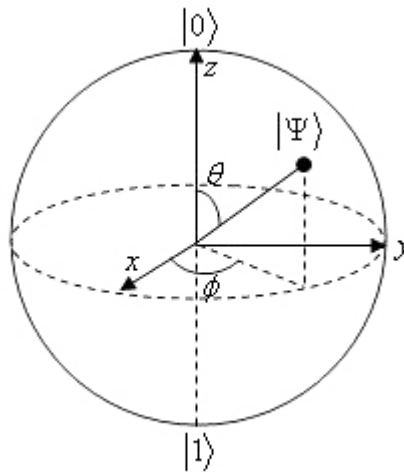


Figure: 63.4 - Sphère de Bloch

La sphère de Bloch est une généralisation de la représentation d'un nombre complexe z avec $|z|^2 = 1$ comme un point du cercle dans le plan (de Gauss) comme nous l'avons vu lors de notre étude des nombres complexes dans le chapitre traitant des Nombres.

Nous avons vu également dans ce même chapitre qu'un nombre complexe pouvait être représenté par une exponentielle complexe telle que:

$$z = x + iy = re^{i\theta} \quad (63.43)$$

et si le cercle était unitaire:

$$z = e^{i\theta} \quad (63.44)$$

Notons que la contrainte $|z|^2 = 1$ élimine un degré de liberté.

Nous allons maintenant noter la décomposition d'un état de polarisation sous la forme:

$$|\Psi\rangle = \alpha|x\rangle + \beta|y\rangle \quad (63.45)$$

et celle-ci sous une forme plus traditionnelle en informatique quantique (logique vu les bases...):

$$|\Psi\rangle = \alpha|0\rangle + \beta|1\rangle \quad (63.46)$$

où $\alpha, \beta \in \mathbb{C}$ (eh oui! on est plus dans le cas simple d'une onde polarisée linéairement maintenant...!) sans oublier la condition de normalisation:

$$\langle\Psi|\Psi\rangle = \Psi\overline{\Psi} = |\alpha|^2 + |\beta|^2 = 1 \quad (63.47)$$

Nous pouvons donc écrire le qubit sous la forme:

$$|\Psi\rangle = r_\alpha e^{i\phi_\alpha} |0\rangle + r_\beta e^{i\phi_\beta} |1\rangle \quad (63.48)$$

Remarque: Attention il est très important d'avoir lu la partie traitant de la polarisation de la lumière dans le chapitre d'Optique Ondulatoire pour comprendre que cela ne tombe pas du ciel! À la différence, que nous ne travaillons pas ici avec des phaseurs car la solution de l'équation d'évolution de Schrödinger comporte des exponentielles complexes comme nous l'avons vu dans le cadre de la résolution de celle-ci pour un mode propre d'une particule libre.

Ajouter un facteur de phase globale ne devrait avoir aucune influence sur les coefficients α, β car:

$$|e^{i\gamma}\alpha|^2 = (e^{i\gamma}\alpha)(\overline{e^{i\gamma}\alpha}) = (e^{i\gamma}\alpha)(e^{-i\gamma}\overline{\alpha}) = \alpha\overline{\alpha} = |\alpha|^2 \quad (63.49)$$

et similairement pour $|\beta|^2$. Ainsi, nous sommes libres de multiplier notre qubit polarisé et normalisé:

$$|\Psi\rangle = r_\alpha e^{i\phi_\alpha} |0\rangle + r_\beta e^{i\phi_\beta} |1\rangle \quad (63.50)$$

par la phase globale $e^{-i\phi_\alpha}$ ce qui donne:

$$|\Psi'\rangle = r_\alpha e^{-i\phi_\alpha} e^{i\phi_\alpha} |0\rangle + r_\beta e^{-i\phi_\alpha} e^{i\phi_\beta} |1\rangle = r_\alpha |0\rangle + r_\beta e^{i\phi_\beta - i\phi_\alpha} |1\rangle = r_\alpha |0\rangle + r_\beta e^{i\phi} |1\rangle \quad (63.51)$$

En plus, nous avons toujours la condition de normalisation $\langle\Psi'|\Psi'\rangle = 1$ à respecter (imposer).

En revenant aux coordonnées cartésiennes, nous avons:

$$|\Psi\rangle = r_\alpha |0\rangle + r_\beta e^{i\phi} |1\rangle = r_\alpha |0\rangle + (x + iy) |1\rangle \quad (63.52)$$

et la contrainte de normalisation donne alors:

$$|r_\alpha|^2 + |x + iy|^2 = r_\alpha^2 + (x + iy)(\overline{x + iy}) = r_\alpha^2 + x^2 + y^2 = 1 \quad (63.53)$$

Ce qui est l'équation d'une sphère unitaire dans l'espace $\mathbb{R}^3$ avec les coordonnées cartésiennes

(x, y, r_α) . D'où l'origine quantique de la sphère de Bloch!

Nous savons (cf. chapitre de Calcul Vectoriel) que les coordonnées cartésiennes sont reliées aux coordonnées sphériques par les relations:

$$x = r \sin \theta \cos \phi \quad y = r \sin \theta \sin \phi \quad z = r \cos \theta \quad (63.54)$$

donc en renommant $r_\alpha \rightarrow z$ et en se rappelant que $r = 1$, nous pouvons écrire:

$$|\Psi\rangle = z|0\rangle + (x + iy)|1\rangle = \cos \theta |0\rangle + \sin \theta (\cos \phi + i \sin \phi) |1\rangle = \cos \theta |0\rangle + \sin \theta e^{i\phi} |1\rangle \quad (63.55)$$

Nous n'avons maintenant plus que deux paramètres utiles à connaître pour définir notre point sur la sphère unité (et ce toujours à l'arbitraire de phase près).

Le lecteur remarquera que contrairement au qubit polarisé linéairement, le cas général ci-dessus rajoute un terme complexe et qu'inversement en enlevant ce terme supplémentaire, nous retombons sur la relation de l'onde polarisée linéairement vue en début de chapitre. Cependant, à noter (pour la culture générale), que si nous ajoutons le même terme complexe au premier terme, nous avons alors aussi la représentation suivante très courante de l'onde polarisée linéairement et dont l'écriture se nomme "**vecteur de Jones**":

$$|\Psi\rangle = \cos \theta e^{i\phi} |0\rangle + \sin \theta e^{i\phi} |1\rangle = \cos \theta e^{i\phi} \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \sin \theta e^{i\phi} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \sin \theta \\ \cos \theta \end{pmatrix} e^{i\phi} \quad (63.56)$$

Mais nous n'en avons pas encore terminé!

Revenons donc à (nous enlevons l'apostrophe pour l'état de polarisation):

$$|\Psi\rangle = \cos \theta |0\rangle + \sin \theta e^{i\phi} |1\rangle \quad (63.57)$$

et remarquons que:

$$\theta = 0 \Rightarrow |\Psi\rangle = |0\rangle \quad (63.58)$$

et:

$$\theta = \frac{\pi}{2} \Rightarrow |\Psi\rangle = e^{i\phi} |1\rangle \quad (63.59)$$

en n'oubliant pas que $\phi \in \mathbb{R}$ et que dans ce cas le facteur exponentiel devant $|1\rangle$ est un changement de phase global sans influence.

Tout cela suggère que $0 \leq \theta \leq \frac{\pi}{2}$ est suffisant pour décrire n'importe quel état de polarisation et donc tous les points de la sphère de Bloch.

Par ailleurs, nous pouvons voir que dans le système (r, θ, ϕ) le point de coordonnées $(1, \pi - \theta, \phi + \pi)$ est le point opposé à celui de coordonnées $(1, \theta, \phi)$:

$$\begin{aligned}
|\Psi\rangle &= \cos(\pi - \theta)|0\rangle + \sin(\pi - \theta)e^{i(\phi + \pi)}|1\rangle \\
&= -\cos(\theta)|0\rangle + \sin(\theta)e^{i\pi}e^{i\phi}|1\rangle \\
&= -\cos(\theta)|0\rangle - \sin(\theta)e^{i\phi}|1\rangle = -|\Psi\rangle = |\Psi'\rangle
\end{aligned} \tag{63.60}$$

Nous avons par ailleurs (c'est immédiat):

$$\langle\Psi|\Psi'\rangle = 0 \tag{63.61}$$

et donc des points opposés sur la sphère de Bloch correspondent à des qubits (états de polarisation) orthogonaux!

Ainsi, nous pouvons considérer seulement l'hémisphère supérieur de la sphère de Bloch puisque les points opposés ne diffèrent que d'un facteur de phase -1 et sont donc équivalents dans la représentation de la sphère de Bloch.

Ainsi, la relation:

$$|\Psi\rangle = \cos\frac{\theta}{2}|0\rangle + e^{i\phi}\sin\frac{\theta}{2}|1\rangle \tag{63.62}$$

est suffisante pour décrire toute la sphère de Bloch dans un espace complexe de dimension 2.

Par construction, chaque point donné par la relation précédente de dimension 2 contient une double représentation d'une rotation dans l'espace réel de dimension 3.

Nous avons vu par ailleurs dans le chapitre de calcul spinoriel qu'une rotation pouvait s'écrire sous la forme:

$$R\left(\vec{L}, \frac{\theta}{2}\right) = \cos\frac{\theta}{2}\sigma_0 - i\sin\frac{\theta}{2}\vec{L} \circ \vec{\sigma} \tag{63.63}$$

avec pour rappel les matrices de Pauli:

$$\sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \sigma_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{63.64}$$

Soit avec l'écriture habituelle (traditionnelle) du domaine de l'informatique quantique:

$$\begin{aligned}
R_x(\theta) &= \cos\left(\frac{\theta}{2}\right)I - i \sin\left(\frac{\theta}{2}\right)\sigma_1 = \begin{bmatrix} \cos\left(\frac{\theta}{2}\right) & -i \sin\left(\frac{\theta}{2}\right) \\ -i \sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) \end{bmatrix} \\
R_y(\theta) &= \cos\left(\frac{\theta}{2}\right)I - i \sin\left(\frac{\theta}{2}\right)\sigma_2 = \begin{bmatrix} \cos\left(\frac{\theta}{2}\right) & -\sin\left(\frac{\theta}{2}\right) \\ \sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) \end{bmatrix} \\
R_z(\theta) &= \cos\left(\frac{\theta}{2}\right)I - i \sin\left(\frac{\theta}{2}\right)\sigma_3 = \begin{bmatrix} \cos\left(\frac{\theta}{2}\right) - i \sin\left(\frac{\theta}{2}\right) & 0 \\ 0 & \cos\left(\frac{\theta}{2}\right) + i \sin\left(\frac{\theta}{2}\right) \end{bmatrix}
\end{aligned} \tag{63.65}$$

Soit après simplification de la dernière matrice:

$$\begin{aligned}
R_x(\theta) &= \begin{bmatrix} \cos\left(\frac{\theta}{2}\right) & -i \sin\left(\frac{\theta}{2}\right) \\ -i \sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) \end{bmatrix} \\
R_y(\theta) &= \begin{bmatrix} \cos\left(\frac{\theta}{2}\right) & -\sin\left(\frac{\theta}{2}\right) \\ \sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) \end{bmatrix} \\
R_z(\theta) &= \begin{bmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{bmatrix}
\end{aligned} \tag{63.66}$$

Maintenant considérons la rotation de notre vecteur d'état de polarisation (dû à un projecteur):

$$R_z(\alpha)|\Psi\rangle = \begin{bmatrix} e^{-i\alpha/2} & 0 \\ 0 & e^{i\alpha/2} \end{bmatrix} \begin{pmatrix} \cos\frac{\theta}{2} \\ e^{i\phi} \sin\frac{\theta}{2} \end{pmatrix} = \begin{pmatrix} e^{-i\alpha/2} \cos\frac{\theta}{2} \\ e^{i\alpha/2} e^{i\phi} \sin\frac{\theta}{2} \end{pmatrix} \tag{63.67}$$

Pour obtenir un coefficient de $|0\rangle$ réel (afin d'avoir une observable dans la projection d'un axe), nous multiplions par un facteur de phase $e^{i\alpha/2}$ donnant:

$$e^{i\alpha/2} \begin{pmatrix} e^{-i\alpha/2} \cos\frac{\theta}{2} \\ e^{i\alpha/2} e^{i\phi} \sin\frac{\theta}{2} \end{pmatrix} = \begin{pmatrix} \cos\frac{\theta}{2} \\ e^{i\alpha} e^{i\phi} \sin\frac{\theta}{2} \end{pmatrix} \tag{63.68}$$

donc pour obtenir une rotation autour de l'axe z il suffit de changer $\phi \rightarrow \phi + \alpha$.

Donc si nous revenons sur:

$$R(\vec{L}, \theta) = \cos \frac{\theta}{2} I - i \sin \frac{\theta}{2} \vec{L} \cdot \vec{\sigma} \quad (63.69)$$

il peut être montré de la même manière que dans un cadre général un opérateur qubit unitaire peut être écrit sous la forme observable:

$$U = e^{i\alpha} R(\vec{L}, \theta) \quad (63.70)$$

Il faut choisir ensuite les angles et l'axe de rotation pour définir complètement l'opérateur.

2.1.1. QUBIT DE POLARISATION

Nous allons revenir ici sur le cas de la polarisation du photon mais ce coup-ci, nous allons pouvoir généraliser grâce à la formulation de la sphère de Bloch à n'importe quel type de polarisation.

Imaginons pour cela un polariseur qui ne laisse passer que des photons polarisés verticalement suivis d'un photodétecteur, qui fait 'clic' si un photon est détecté et rien sinon. Ce dispositif nous permet de détecter les photons polarisés verticalement.

Traduisons ceci dans le langage de la mécanique quantique:

Les états du système sont donc les états de polarisation d'un photon. Les mesures de l'observable auront aussi pour valeur ses états de polarisation.

Les mesures possibles sont:

$$\begin{aligned} \lambda_0 &= 0 = \text{'clic'} \\ \lambda_1 &= 1 = \text{'pas clic'} \end{aligned} \quad (63.71)$$

Nous noterons les états correspondants $|x\rangle, |y\rangle$. Dans notre configuration, il est alors évident que le couple (λ_0, λ_1) représente les valeurs propres et $|x\rangle, |y\rangle$ les vecteurs propres d'un opérateur (que nous ne connaissons pas) et que nous noterons donc $\mathcal{P}$.

Comme nous le savons, $|x\rangle, |y\rangle$ est une base orthonormée de l'espace des états (de polarisation). C'est la base appelée "**base H/V**" pour Horizontale/Verticale et qui est notée normalement:

$$|0\rangle, |1\rangle \quad (63.72)$$

Prenons maintenant plusieurs cas:

1. Soit un photon dans l'état:

$$|\Psi\rangle = |0\rangle \quad (63.73)$$

Alors:

$$\begin{aligned} P(\lambda = \lambda_0) &= P(0) = \langle 0|0\rangle = 1 \\ P(\lambda = \lambda_1) &= P(1) = \langle 1|0\rangle = 0 \end{aligned} \quad (63.74)$$

2. Soit un photon dans l'état mi-vertical/horizontal, c'est-à-dire oblique (ce qui peut être assimilé à la superposition quantique de ses deux polarisations):

$$|\Psi\rangle = \frac{|0\rangle + |1\rangle}{\sqrt{2}} \quad (63.75)$$

où la racine est juste là pour assurer la condition de normalisation $\langle\Psi|\Psi\rangle = 1$

Alors puisqu'il y a superposition (soit la moitié de chaque dans l'onde totale):

$$\begin{aligned} P(\lambda = \lambda_0) = P(0) &= \langle 0|\Psi\rangle = \frac{1}{2} \\ P(\lambda = \lambda_1) = P(1) &= \langle 1|\Psi\rangle = \frac{1}{2} \end{aligned} \quad (63.76)$$

3. Prenons maintenant n'importe quelle polarisation (et c'est cela que nous n'avions pas avant!):

$$|\Psi\rangle = \cos\frac{\theta}{2}|0\rangle + e^{i\phi}\sin\frac{\theta}{2}|1\rangle \quad (63.77)$$

qui est bien normé comme nous le savons.

Alors:

$$\begin{aligned} P(\lambda = \lambda_0) = P(0) &= \langle 0|\Psi\rangle = \left(\cos\frac{\theta}{2}\right)^2 \\ P(\lambda = \lambda_1) = P(1) &= \langle 1|\Psi\rangle = \left(\sin\frac{\theta}{2}\right)^2 \end{aligned} \quad (63.78)$$

La somme des deux probabilités donnant bien 1!

Maintenant, imaginons que nous tournions le polariseur de $\pi/4$. Nous noterons la nouvelle base de ce polariseur $|0'\rangle, |1'\rangle$ déterminée par une rotation d'angle $\pi/4$ avec par construction:

$$\begin{aligned} |0'\rangle &= \frac{|0\rangle + |1\rangle}{\sqrt{2}} \\ |1'\rangle &= \frac{|0\rangle - |1\rangle}{\sqrt{2}} \end{aligned} \quad (63.79)$$

où la première base correspond donc à la polarisation diagonale et la seconde est appelée polarisation antidiagonale. Il s'agit donc de la "base **D/A**" (Diagonale/Antidiagonale). Sous la forme des vecteurs de Jones ces deux dernières relations s'écrivent:

$$\begin{aligned}
 |0'\rangle &= \frac{|0\rangle + |1\rangle}{\sqrt{2}} = \frac{\begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix}}{\sqrt{2}} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
 |1'\rangle &= \frac{|0\rangle - |1\rangle}{\sqrt{2}} = \frac{\begin{pmatrix} 0 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \end{pmatrix}}{\sqrt{2}} = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix}
 \end{aligned}
 \quad (63.80)$$

Si nous imaginons que nous avons deux polariseurs qui se suivent. Le premier ayant la base D/A et le deuxième la base H/V, le premier va préparer le photon polarisé de manière générale dans un état (polarisation) particulier qui sera par construction l'état oblique pour la base H/V. Ainsi notre deuxième polariseur n'aura que des situations du type:

$$\begin{aligned}
 P(\lambda = \lambda_0) &= P(0) = \langle 0 | \Psi \rangle = \frac{1}{2} \\
 P(\lambda = \lambda_1) &= P(1) = \langle 1 | \Psi \rangle = \frac{1}{2}
 \end{aligned}
 \quad (63.81)$$

Ainsi, cela montre que toute mesure perturbe bien évidemment l'état de polarisation du photon et perturbe donc l'état du système.

Ce résultat est utilisé en cryptographie quantique!

2.2. QUBIT DE SPIN 1/2

Nous allons voir ici comment construire un qubit basé sur une particule possédant un spin de $\frac{1}{2}$.

Lors de l'étude de la sphère de Bloch, nous avons examiné un qubit à un instant déterminé et nous avons vu que dans un espace de Hilbert H , ce qubit est décrit (par choix) par un vecteur unitaire:

$$|\Psi\rangle = \alpha|0\rangle + \beta|1\rangle \quad |\alpha|^2 + |\beta|^2 = 1 \quad (63.82)$$

décomposé dans une base orthonormée $\{|0\rangle, |1\rangle\}$.

Considérons l'état initial le plus général et minimal correspondant à une orientation arbitraire d'un spin:

$$|\Psi\rangle = \cos \frac{\theta}{2} |0\rangle + e^{i\phi} \sin \frac{\theta}{2} |1\rangle \quad (63.83)$$

qui correspond comme nous le savons à deux états opposés (et superposés) sur la sphère de Bloch.

Remarquons que nous avons bien une probabilité normalisée de la forme:

$$\langle \Psi | \Psi \rangle = \cos^2 \frac{\theta}{2} + \sin^2 \frac{\theta}{2} = 1 \quad (63.84)$$

Nous avons également vu que la projection selon z par l'opérateur $R_z(\alpha)$ de l'état $|\Psi\rangle$ est donnée à l'arbitraire de phase par:

$$e^{i\alpha/2} R_z(\alpha) |\Psi\rangle = \begin{pmatrix} \cos \frac{\theta}{2} \\ e^{i\alpha} e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} = \begin{pmatrix} \cos \frac{\theta}{2} \\ e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} \quad (63.85)$$

Maintenant rappelons notre exemple:

$$\begin{aligned} M|x\rangle &= +1 \cdot |x\rangle = +|x\rangle \\ M|y\rangle &= -1 \cdot |y\rangle = -|y\rangle \end{aligned} \quad (63.86)$$

où M était donc un opérateur hermitique. Or les matrices de Pauli sont des opérateurs hermitiques simples. De plus, comme nous l'avons démontré dans le chapitre de calcul spinoriel, l'opérateur hermitique σ_x (assimilé à M) a comme par hasard les mêmes valeurs propres et vecteurs propres correspondant aux deux relations. Mais s'écrit alors de manière traditionnelle comme nous l'avons vu dans le chapitre de Calcul Spinoriel:

$$\begin{aligned} \sigma_x |+\rangle &= +|+\rangle \\ \sigma_x |-\rangle &= -|-\rangle \end{aligned} \quad (63.87)$$

ou de manière encore plus condensée:

$$\sigma_x |\pm\rangle = \pm |\pm\rangle \quad (63.88)$$

De plus cet opérateur satisfait aussi la relation:

$$\sigma_x = M = \mathcal{P}_x - \mathcal{P}_y = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (63.89)$$

Et quelle est la propriété physique associée à σ_x ? Eh bien il s'agit du spin et nous y reviendrons donc un peu plus bas car cela signifie que nous pouvons utiliser le spin 1/2 comme qubit.

Maintenant introduisons l'évolution du système sur cette projection car celui-ci n'est pas statique (mais ceci dit cela ne changerait rien à ce cas particulier de le considérer comme statique).

Nous avons vu dans le chapitre de Physique Quantique Ondulatoire que cette opération consistait dans un cas simple (comme ici) à introduire un terme de phase dépendant du temps du type:

$$e^{-\frac{i\hbar}{\hbar} H} \quad (63.90)$$

Ainsi:

$$|\Psi(\alpha, t)\rangle = \begin{pmatrix} e^{-\frac{i\hbar}{\hbar} H} \cos \frac{\theta}{2} \\ e^{-\frac{i\hbar}{\hbar} H} e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} = \begin{pmatrix} e^{-\frac{i\hbar}{\hbar} H} \cos \frac{\theta}{2} \\ e^{-\frac{i\hbar}{\hbar} H} e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} = \begin{pmatrix} e^{-2\pi i t \nu} \cos \frac{\theta}{2} \\ e^{-2\pi i t \nu} e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} = \begin{pmatrix} e^{-i\alpha t} \cos \frac{\theta}{2} \\ e^{-i\alpha t} e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} \quad (63.91)$$

Ce qui est noté plus sobrement:

$$|\Psi(t)\rangle = e^{-i\alpha t} \begin{pmatrix} \cos \frac{\theta}{2} \\ e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} \quad (63.92)$$

Rappelons que l'état d'une particule de spin 1/2 est bidimensionnel et décrit par la matrice d'état (cf. chapitre de Physique Quantique Relativiste):

$$\mathcal{M} = S = \frac{1}{2} \hbar \sigma \quad (63.93)$$

Soit:

$$\begin{aligned} S_x &= \frac{\hbar}{2} \sigma_1 = \frac{\hbar}{2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \\ S_y &= \frac{\hbar}{2} \sigma_2 = \frac{\hbar}{2} \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \\ S_z &= \frac{\hbar}{2} \sigma_3 = \frac{\hbar}{2} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \end{aligned} \quad (63.94)$$

Si nous voulons calculer la moyenne (l'espérance) de l'observable (propriété physique) selon chaque axe nous avons alors en utilisant le cinquième postulat dans le cas de l'axe x:

$$\begin{aligned} \langle S_x \rangle &= \langle \Psi(t) | S_x | \Psi(t) \rangle = e^{+i\alpha t} \begin{pmatrix} \cos \frac{\theta}{2} & e^{-i\phi} \sin \frac{\theta}{2} \end{pmatrix} \frac{\hbar}{2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} e^{-i\alpha t} \begin{pmatrix} \cos \frac{\theta}{2} \\ e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} \\ &= \frac{\hbar}{2} \begin{pmatrix} \cos \frac{\theta}{2} & e^{-i\phi} \sin \frac{\theta}{2} \end{pmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{pmatrix} \cos \frac{\theta}{2} \\ e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} = \frac{\hbar}{2} \begin{pmatrix} \cos \frac{\theta}{2} & e^{-i\phi} \sin \frac{\theta}{2} \end{pmatrix} \begin{pmatrix} e^{i\phi} \sin \frac{\theta}{2} \\ \cos \frac{\theta}{2} \end{pmatrix} \quad (63.95) \\ &= \frac{\hbar}{2} \left(\cos \frac{\theta}{2} e^{i\phi} \sin \frac{\theta}{2} + e^{-i\phi} \sin \frac{\theta}{2} \cos \frac{\theta}{2} \right) = \frac{\hbar}{2} \sin \frac{\theta}{2} \cos \frac{\theta}{2} (e^{i\phi} + e^{-i\phi}) \\ &= \frac{\hbar}{2} \sin \frac{\theta}{2} \cos \frac{\theta}{2} (\cos \phi + i \sin \phi + \cos \phi - i \sin \phi) = \hbar \sin \frac{\theta}{2} \cos \frac{\theta}{2} \cos \phi \\ &= \frac{\hbar}{2} \sin \theta \cos \phi \end{aligned}$$

Selon y:

$$\begin{aligned}
\langle S_y \rangle &= \langle \Psi(t) | S_y | \Psi(t) \rangle = e^{+i\alpha x} \left(\cos \frac{\theta}{2} \quad e^{-i\phi} \sin \frac{\theta}{2} \right) \frac{\hbar}{2} \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} e^{-i\alpha x} \begin{pmatrix} \cos \frac{\theta}{2} \\ e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} \\
&= \frac{\hbar}{2} \left(\cos \frac{\theta}{2} \quad e^{-i\phi} \sin \frac{\theta}{2} \right) \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \begin{pmatrix} \cos \frac{\theta}{2} \\ e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} = \frac{\hbar}{2} \left(\cos \frac{\theta}{2} \quad e^{-i\phi} \sin \frac{\theta}{2} \right) \begin{pmatrix} -ie^{i\phi} \sin \frac{\theta}{2} \\ i \cos \frac{\theta}{2} \end{pmatrix} \quad (63.96) \\
&= \frac{\hbar}{2} \left(-i \cos \frac{\theta}{2} e^{i\phi} \sin \frac{\theta}{2} + i e^{-i\phi} \sin \frac{\theta}{2} \cos \frac{\theta}{2} \right) = \frac{\hbar}{2} \sin \frac{\theta}{2} \cos \frac{\theta}{2} (-ie^{i\phi} + ie^{-i\phi}) \\
&= \frac{\hbar}{2} \sin \frac{\theta}{2} \cos \frac{\theta}{2} (-i \cos(\phi) + \sin(\phi) + i \cos(\phi) + \sin(\phi)) = \hbar \sin \frac{\theta}{2} \cos \frac{\theta}{2} \sin(\phi) \\
&= \frac{\hbar}{2} \sin \theta \sin \phi
\end{aligned}$$

Et selon z :

$$\begin{aligned}
\langle S_z \rangle &= \langle \Psi(t) | S_z | \Psi(t) \rangle = e^{+i\alpha x} \left(\cos \frac{\theta}{2} \quad e^{-i\phi} \sin \frac{\theta}{2} \right) \frac{\hbar}{2} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} e^{-i\alpha x} \begin{pmatrix} \cos \frac{\theta}{2} \\ e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} \\
&= \frac{\hbar}{2} \left(\cos \frac{\theta}{2} \quad e^{-i\phi} \sin \frac{\theta}{2} \right) \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{pmatrix} \cos \frac{\theta}{2} \\ e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} = \frac{\hbar}{2} \left(\cos \frac{\theta}{2} \quad e^{-i\phi} \sin \frac{\theta}{2} \right) \begin{pmatrix} \cos \frac{\theta}{2} \\ -e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} \quad (63.97) \\
&= \frac{\hbar}{2} \left(\cos^2 \frac{\theta}{2} - \sin^2 \frac{\theta}{2} \right) = \frac{\hbar}{2} \cos \theta
\end{aligned}$$

Nous retrouvons donc bien pour la composante z (car c'est la seule qui nous intéresse ici) le résultat qui était imposé plus haut sous la forme :

$$\langle M \rangle_\theta = \langle \theta | \mathcal{P}_x | \theta \rangle - \langle \theta | \mathcal{P}_y | \theta \rangle = \cos^2 \theta - \sin^2 \theta = \cos(2\theta) \quad (63.98)$$

à une différence d'angle qui n'est qu'histoire de substitution et d'une amplitude qui permet de mettre en adéquation la particularité de la configuration du système. Nous avons dès lors des relations mathématiques similaires en tout point en ce qui concerne la manipulation des qubits de spins orientés ou des qubits de photons polarisés.

La question que nous pouvons nous poser sur le spin est comment le préparer dans l'état $|\Psi\rangle$? Au fait nous pouvons le faire avec un champ magnétique en copiant à l'identique l'expérience de Stern-Gerlach qui permet de séparer un faisceau de particules de spin $\frac{1}{2}$ en deux faisceaux distincts.

Puisque nous connaissons maintenant toutes les valeurs propres et vecteurs propres de l'opérateur de spin S , nous pouvons alors déterminer la forme générale de l'opérateur de spin sous une orientation quelconque :

$$\begin{aligned}
S_n &= S_x \sin \theta \cos \phi - S_y \sin \theta \sin \phi + S_z \cos \theta \\
&= \frac{\hbar}{2} \left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \sin \theta \cos \phi + \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \sin \theta \sin \phi + \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \cos \theta \right) \\
&= \frac{\hbar}{2} \begin{bmatrix} \cos \theta & \sin \theta \cos \phi - i \sin \theta \sin \phi \\ \sin \theta \cos \phi + i \sin \theta \sin \phi & -\cos \theta \end{bmatrix} \quad (63.99) \\
&= \frac{\hbar}{2} \begin{bmatrix} \cos \theta & \sin \theta (\cos \phi - i \sin \phi) \\ \sin \theta (\cos \phi + i \sin \phi) & -\cos \theta \end{bmatrix} = \frac{\hbar}{2} \begin{bmatrix} \cos \theta & e^{-i\phi} \sin \theta \\ e^{i\phi} \sin \theta & -\cos \theta \end{bmatrix}
\end{aligned}$$

Nous avons par ailleurs:

$$\begin{aligned}
S_n |\Psi\rangle &= S_n \left(\cos \frac{\theta}{2} |0\rangle + e^{i\phi} \sin \frac{\theta}{2} |1\rangle \right) = \frac{\hbar}{2} \begin{bmatrix} \cos \theta & e^{-i\phi} \sin \theta \\ e^{i\phi} \sin \theta & -\cos \theta \end{bmatrix} \begin{pmatrix} \cos \frac{\theta}{2} \\ e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} \\
&= \frac{\hbar}{2} \begin{pmatrix} \cos \theta \cos \frac{\theta}{2} + \sin \theta \sin \frac{\theta}{2} \\ e^{i\phi} \sin \theta \cos \frac{\theta}{2} - \cos \theta e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} = \frac{\hbar}{2} \begin{pmatrix} \cos \frac{\theta}{2} \\ e^{i\phi} \left(\sin \theta \cos \frac{\theta}{2} - \cos \theta \sin \frac{\theta}{2} \right) \end{pmatrix} \quad (63.100) \\
&= \frac{\hbar}{2} \begin{pmatrix} \cos \frac{\theta}{2} \\ e^{i\phi} \left(\sin \theta \cos \frac{\theta}{2} - \cos \theta \sin \frac{\theta}{2} \right) \end{pmatrix} = \frac{\hbar}{2} \begin{pmatrix} \cos \frac{\theta}{2} \\ e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} = \frac{\hbar}{2} |\Psi\rangle
\end{aligned}$$

Nous avons $|\Psi\rangle$ qui est donc un vecteur propre de S_n avec la valeur propre $\frac{\hbar}{2}$.

Maintenant, nous savons en utilisant le 5ème postulat que la probabilité de trouver la valeur propre $\hbar/2$ (de l'opérateur S_n), lors d'une mesure de la propriété S selon l'axe z effectuée au temps t sur le système quantique préparé dans l'état Ψ , est donnée par le carré du module de la projection de la fonction ou vecteur d'état Ψ sur le vecteur ou vecteur propre φ associé à la valeur propre $\hbar/2$ (et de son opérateur selon cet axe).

Or l'opérateur selon z est:

$$\sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (63.101)$$

Pour le même opérateur, nous avons vu dans le chapitre de Calcul Spinoriel qu'il avait comme vecteurs propres:

$$|+\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, |-\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad (63.102)$$

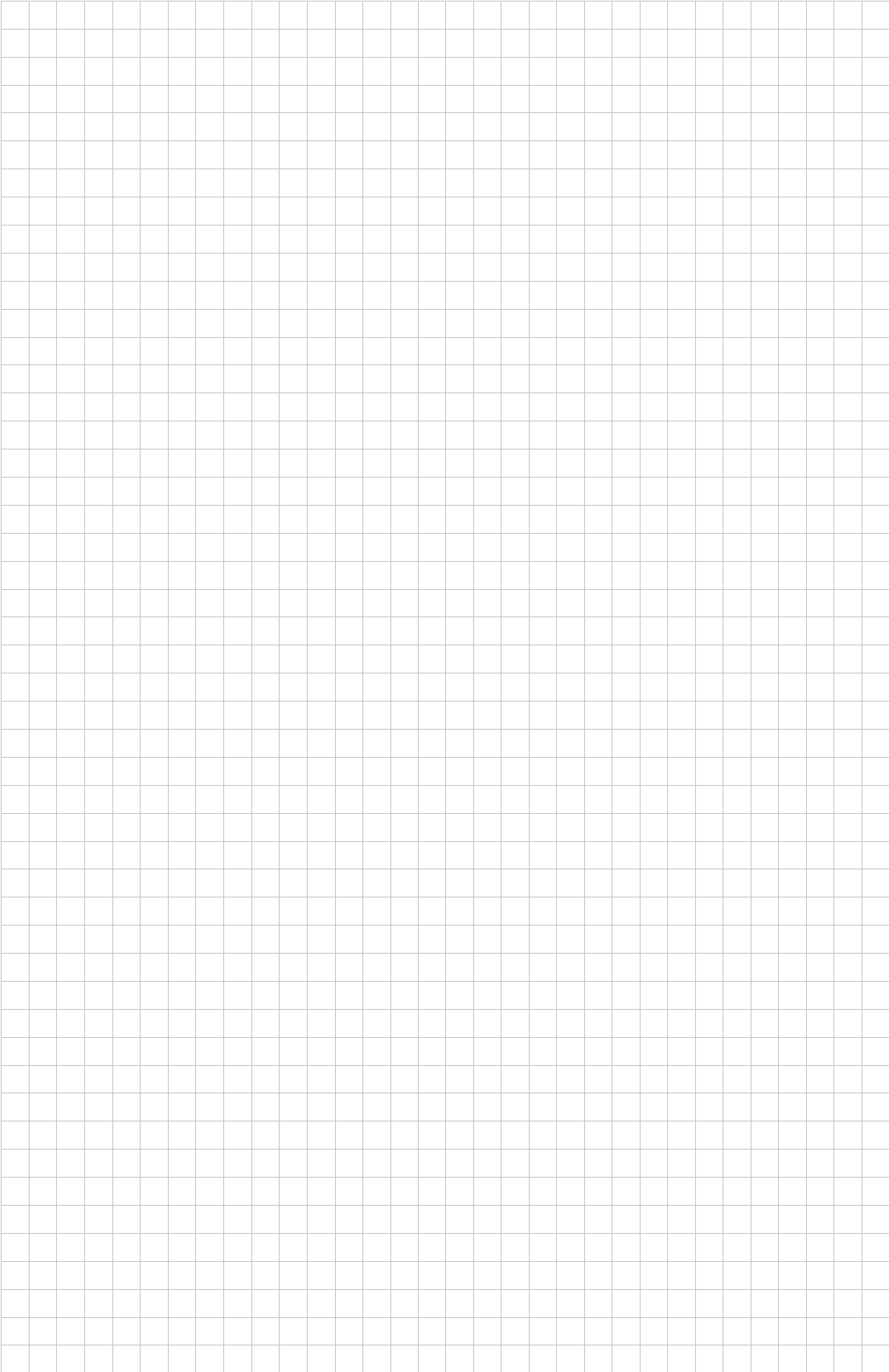
Prenons le premier vecteur propre orienté donc selon z sur la sphère de Bloch. Nous avons alors:

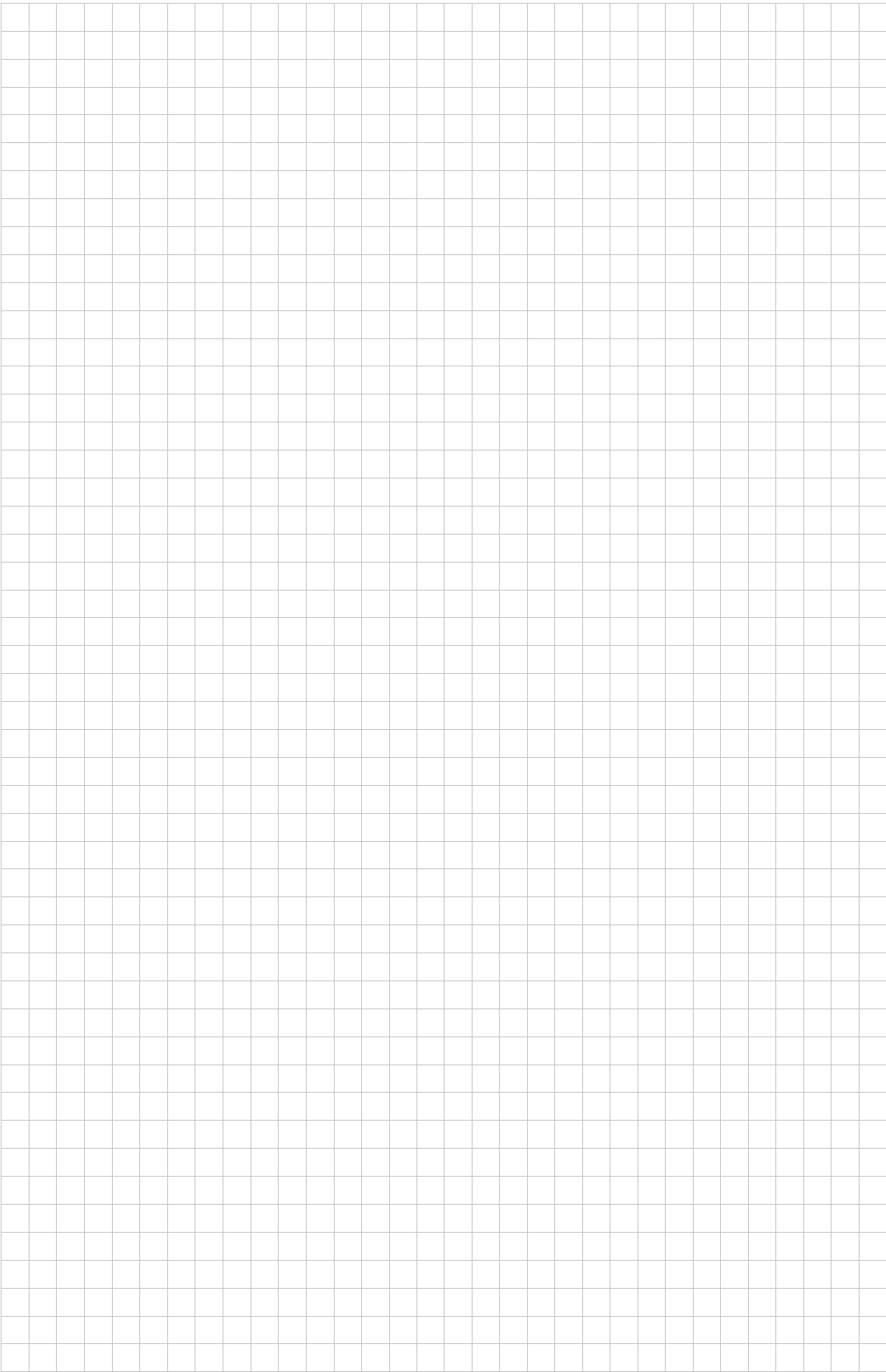
$$P = |\langle + | \Psi \rangle|^2 = \left| \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \cos \frac{\theta}{2} \\ e^{i\phi} \sin \frac{\theta}{2} \end{pmatrix} \right|^2 = \left| \cos \frac{\theta}{2} \right|^2 = \cos^2 \frac{\theta}{2} \quad (63.103)$$

Et la probabilité selon l'autre vecteur propre donnerait la même expression mais avec un sinus. La somme des deux probabilités nous donnerait alors bien 1!

Nous remarquons donc que les relations sont très similaires entre le photon et le spin dans notre cas d'étude. Ce qui est normal puisque les deux sont des systèmes à deux niveaux, d'où les résultats similaires.









XI

Mathématiques sociales

64. DYNAMIQUE DES POPULATIONS

Einstein, à la fin de sa vie, avait coutume d'évoquer les trois explosions qui allaient sous peu menacer l'humanité: l'explosion des bombes nucléaires, l'explosion de nos savoirs, l'explosion de l'effectif des hommes. Le rôle des scientifiques à ce niveau n'est pas seulement d'améliorer leur savoir, mais de le partager et surtout de diffuser la conscience qu'ils ont acquise des conséquences qu'ils savent discerner. Einstein avait à coeur de jouer ce rôle. Au-delà des péripéties immédiates, il savait voir les évolutions à long terme. Parmi ces évolutions, celle du nombre des hommes lui apparaissait à l'évidence comme celle des plus grands dangers (déjà pressentis par les initiateurs de la science démographique au début du 18ème siècle!). Aujourd'hui, en début de ces années 2000, nous pouvons constater leur perspicacité. L'humanité est réellement prise à la gorge par l'accroissement de son effectif ou de son maintien actuel sans la forcer à la baisse (car les politiciens favorisent la croissance infinie et la participation aux aides sociales qui effectivement imposent une augmentation sans fin de la population). Les problèmes majeurs étant de garantir le plein emploi (éviter le "chômage technologique" à un certain niveau) et le fait que l'accroissement de la population enrichit naturellement les propriétaires et appauvrit les salariés (parmi tant d'autres problèmes...).

En attendant que les politiciens de tous les grands pays occidentaux comprennent et montrent l'exemple en osant proposer la limitation de la population (et donc du nombre de naissances) au niveau mondial et que la mentalité de la population de la rue soit ouverte à ce type de proposition (c'est elle qui élit ses politiciens...), l'étude théorique du sujet peut permettre d'ouvrir quelques perspectives... en ce qui concerne le renouvellement et le cycle des ressources de notre unique planète... (car pour l'instant nous n'en avons qu'une...).

Pour introduire les mathématiques sociales, il nous faut d'abord déterminer les caractéristiques qui décrivent la dynamique du nombre d'individus avant de formaliser les propriétés uniques qui les caractérisent.

1. TABLES DE MORTALITÉ ET NATALITÉ (FONCTIONS BIOMÉTRIQUES)

Nul ne sait ni le jour ni l'heure de sa mort. Cette évidence individuelle n'est plus pertinente si nous nous intéressons non à telle personne, mais à une collectivité assez nombreuse. Alors jouent les compensations entre ceux qui succombent à des accidents prématurés et ceux qui échappent quasi miraculeusement aux pires dangers. Nous pouvons alors décrire la façon dont est payé globalement le tribut à la mort en considérant un grand nombre d'enfants nés la même année et en précisant, grâce aux données de l'état civil, comment leur effectif diminue peu à peu pour un jour s'annuler.

Un tel ensemble de conscrits est appelé par les démographes une "**cohorte**". Considérons donc la cohorte des Français mâles nés en 1900. Chaque année, nous pouvons, en regroupant les indications de l'état civil, calculer le nombre de ceux qui sont décédés dans la cohorte.

Représentons par d_a le quotient du nombre des décès entre les anniversaires a et $a+1$ par l'effectif initial de la cohorte.

Remarque: Nous allons dans ce qui suit définir les bases des probabilités nécessaires au calcul actuariel. En effet, la mathématique actuarielle réunit le calcul financier et le calcul des probabilités. Le paiement d'un capital n'est plus certain et dépend par exemple de la survie d'une personne.

La suite de ces nombres contient la totalité de l'information nécessaire pour étudier la mortalité de cette cohorte.

Nous pouvons en déduire la proportion d'entre eux survivant à l'âge a :

$$S_a = 1 - d_0 - d_1 - d_2 - \dots - d_{a-1} \quad (64.1)$$

Nous pouvons également caractériser l'intensité de la mortalité à chaque âge en divisant le nombre des décès entre les âges a et $a+1$ par le nombre de survivants. Ce nombre est alors appelé le "**quotient de mortalité**":

$$q_a = \frac{S_a - S_{a+1}}{S_a} = \frac{D_a}{S_a} \quad (64.2)$$

La liste âge par âge de ces trois paramètres d , S , et q est la "**table de mortalité**" de la cohorte étudiée. Le tableau ci-dessous en donne le résumé pour les âges multiples de 5.

Âge a	Survivants S_a	Décès de l'âge a à $a+5$ $S_a - S_{a+5}$	Probabilité des décès $q_a = \frac{S_a - S_{a+5}}{S_a}$
0	1	0.228	0.228
5	0.772	0.013	0.017
10	0.759	0.009	0.012
15	0.750	0.023	0.031
20	0.727	0.023	0.032
25	0.704	0.020	0.028
30	0.684	0.021	0.031
35	0.663	0.025	0.038
40	0.638	0.029	0.045
45	0.609	0.026	0.043
50	0.583	0.036	0.062
55	0.547	0.047	0.086
60	0.500	0.062	0.124
65	0.438	0.078	0.178
70	0.360	0.360	

Tableau: 64.1 - Table de mortalité

Remarque: Suivre une génération réelle d'individus tout au long de son existence ou sur une période de temps déterminée est appelé "**analyse longitudinale**" par contraste avec "**l'analyse transversale**", qui consiste à étudier les caractéristiques d'une population à un moment donné. Le tableau ci-dessus est donc un exemple d'analyse longitudinale.

Choisissons au hasard dans ce tableau le nom d'un enfant inconnu dans la liste des naissances de l'année 1900. La grande question lors de cette naissance était: combien d'années vivra-t-il? Aujourd'hui, nous sommes en mesure de répondre rétroactivement à cette question, tout au moins en évoquant des probabilités, car nous connaissons la table de mortalité de cette cohorte.

Si la seule information dont nous disposons aujourd'hui à propos de cet enfant est le fait qu'il est né en 1900, nous pouvons déclarer que la probabilité qu'il ait encore été vivant à l'âge de 5 ans est égale à 0.772, à l'âge de 50 ans de 0.583... Autrement dit que la probabilité pour qu'il soit mort avant 5 ans est égale à 0.228.

Si nous désignons aujourd'hui un individu inconnu repris sur la liste de ceux qui ont été incorporés dans la cohorte au cours de l'année 1920 à l'âge de 20 ans, nous pouvons de même calculer les probabilités des diverses durées de sa vie, mais nous avons une information supplémentaire: il était encore vivant à 20 ans, il a évité les risques de mort avant cet âge. La probabilité qu'il atteigne alors l'âge de 50 ans est devenue:

$${}_20P_{20} = \frac{S_{50}}{S_{20}} = \frac{0.583}{0.727} = 0.802 \quad (64.3)$$

Donc la probabilité pour une personne d'âge a d'être en vie à l'âge $a+n$ est égale à:

$${}_nP_a = \frac{S_{a+n}}{S_a} \quad (64.4)$$

La probabilité pour une personne d'âge a de décéder entre l'âge a et $a+n$ est logiquement donnée par:

$${}_nq_a = \frac{S_a - S_{a+n}}{S_a} = \frac{D_{a+n}}{S_a} \quad (64.5)$$

Ainsi à chaque âge, nous pouvons donner la loi de la variable "durée encore à vivre". Cette loi peut être résumée en indiquant son espérance (cf. chapitre de Statistiques). Un calcul immédiat permet d'en donner la valeur à chaque âge en fonction de la table de mortalité.

Ce résultat peut être passionnant pour un historien, mais il donne une réponse à une question posée il y a longtemps et depuis oubliée. Ce qui nous intéresse est le présent. Cet enfant qui vient de naître, quelle est son espérance de vie? Pour répondre, il faudrait connaître la table de mortalité de sa génération, or le don de prémonition n'existe pas. L'attitude probabiliste permet de contourner cette difficulté à condition de bien préciser les hypothèses sous-jacentes.

Ainsi, pour répondre à la question: quelle est l'espérance de vie des nouveau-nés de 1990, nous faisons alors l'hypothèse, tout à fait gratuite, qu'ils rencontreront à chaque âge, à l'avenir, les conditions qu'ont rencontrées en 1990 les individus de ces âges: en l'an 2000 ils subiront la même mortalité que celle subie en 1990 par ceux qui sont nés en 1980. Bien sûr, personne n'imagine que la réalité sera conforme

à cette hypothèse, mais le calcul qu'elle permet fournit une image synthétique des conditions actuelles de la lutte contre la mort.

C'est cette hypothèse qui fait que l'espérance de vie rajeunit parfois les vieux (leur espérance de vie ayant tendance à augmenter au fur et à mesure qu'ils deviennent plus âgés grâce aux progrès de la science...).

Prenons pour le calcul de l'espérance de vie, la table de mortalité pour les hommes en Suisse en 1983-1993:

Âge n	Survivants S_a	Espérance de vie $E(a)$	Âge n	Survivants S_a	Espérance de vie $E(a)$
0	1	73.68823	55	0.90224	22.81093
1	0.99246	73.24806	56	0.89581	21.97466
2	0.99183	72.29459	57	0.88875	21.14922
3	0.99148	71.32011	58	0.88099	20.33551
4	0.99117	70.34241	59	0.87247	19.53409
5	0.9909	69.36158	60	0.86312	18.7457
6	0.99066	68.37838	61	0.85288	17.97077
7	0.99044	67.39357	62	0.84169	17.20969
8	0.99022	66.40855	63	0.82948	16.46301
9	0.99001	65.42263	64	0.81618	15.73128
10	0.98971	64.44246	65	0.80174	15.01462
11	0.9896	63.44963	66	0.78609	14.31354
12	0.98938	62.46373	67	0.76918	13.62821
13	0.98915	61.47826	68	0.75096	12.95887
14	0.98889	60.49442	69	0.73138	12.30579
15	0.9886	59.51217	70	0.7104	11.66921
16	0.98823	58.53445	71	0.68798	11.04949
17	0.9877	57.56586	72	0.66409	10.44699
18	0.98692	56.61136	73	0.63873	9.861773
19	0.98581	55.6751	74	0.6119	9.294182
20	0.98439	54.75541	75	0.58362	8.744543
21	0.98285	53.84121	76	0.55393	8.21324
22	0.98131	52.9257	77	0.52291	7.700465
23	0.97975	52.00997	78	0.49066	7.206599
24	0.97815	51.09505	79	0.45733	6.731813
25	0.97653	50.17981	80	0.42309	6.276608
26	0.9749	49.26371	81	0.38819	5.840903
27	0.97328	48.34571	82	0.35288	5.425357
28	0.97166	47.42631	83	0.31748	5.030301
29	0.97007	46.50405	84	0.28234	4.656372
30	0.9685	45.57943	85	0.24785	4.304337
31	0.96694	44.65297	86	0.21446	3.974494

32	0.96541	43.72373	87	0.18263	3.667196
33	0.96388	42.79314	88	0.15281	3.382828
34	0.96236	41.86073	89	0.12546	3.120277
35	0.96082	40.92782	90	0.10091	2.879397
36	0.95927	39.99395	91	0.07942	2.658524
37	0.95768	39.06035	92	0.06111	2.455081
38	0.95604	38.12736	93	0.04593	2.266492
39	0.95434	37.19528	94	0.03373	2.086273
40	0.95257	36.26439	95	0.02414	1.915079
41	0.95071	35.33534	96	0.0168	1.751786
42	0.94874	34.40871	97	0.01134	1.595238
43	0.94665	33.48468	98	0.00739	1.447903
44	0.94441	32.5641	99	0.00464	1.306034
45	0.94201	31.64706	100	0.00279	1.172043
46	0.93942	30.73431	101	0.0016	1.04375
47	0.93662	29.82619	102	0.00087	0.91954
48	0.93356	28.92396	103	0.00045	0.777778
49	0.93023	28.0275	104	0.00021	0.666667
50	0.92659	27.1376	105	0.00009	0.555556
51	0.92259	26.25526	106	0.00004	0.25
52	0.91821	25.3805	107	0.00001	0
53	0.91338	24.51471	108	0	
54	0.90808	23.65779			

Tableau: 64.2 - Table de mortalité en Suisse en 1983-1993

Donc la courbe représentative, appelée "l'ordre des vivants" est:

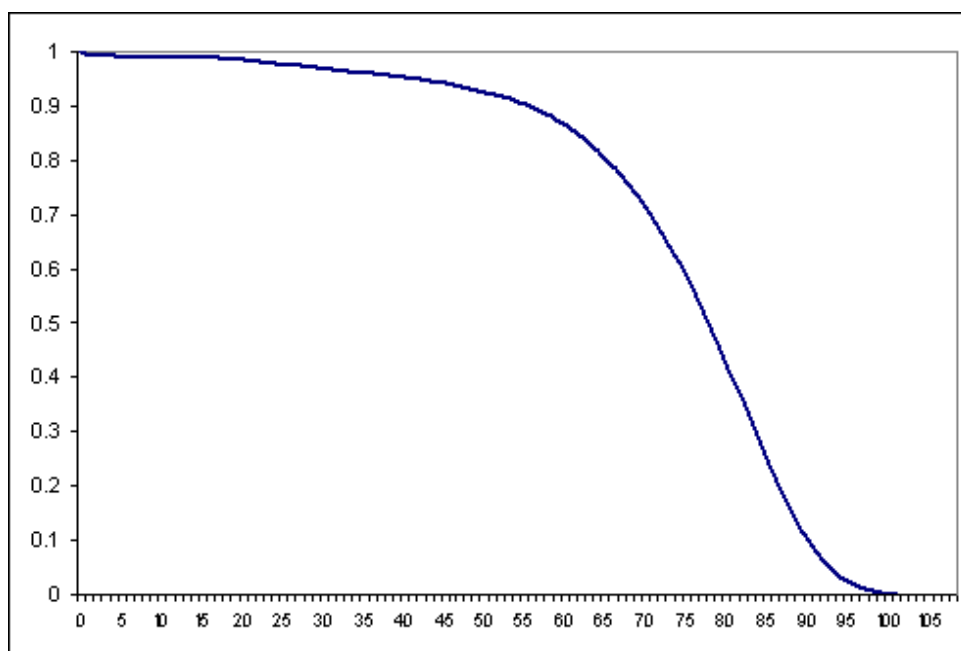


Figure: 64.1 - Probabilité cumulée de survie en fonction de l'âge

Voyons comment calculer l'espérance de vie. Pour cela, considérons un homme en vie lors de son

a -ème anniversaire. Le nombre d'années qui lui reste à vivre est une variable aléatoire dont nous pouvons calculer l'espérance mathématique (cf. chapitre de Probabilités). En négligeant les fractions d'années, cette espérance peut s'écrire:

$$\begin{aligned}
 E(a) &= 0 \cdot q_a + 1 \cdot {}_1q_a + 2 \cdot {}_2q_a + \dots \\
 &= \frac{1}{S_a} (0 \cdot D_a + 1 \cdot D_{a+1} + 2 \cdot D_{a+2} + \dots) \\
 &= \frac{1}{S_a} [0 \cdot (S_a - S_{a+1}) + 1 \cdot (S_{a+1} - S_{a+2}) + 2 \cdot (S_{a+2} - S_{a+3}) + \dots] \\
 &= \frac{1}{S_a} [S_{a+1} - S_{a+2} + 2S_{a+2} - 2S_{a+3} + \dots] \quad (64.6) \\
 &= \frac{1}{S_a} [S_{a+1} + S_{a+2} + S_{a+3} + \dots] \\
 &= \frac{1}{S_a} \sum_{n=1}^{\infty} S_{a+n}
 \end{aligned}$$

Si a est pris comme étant égal à zéro, les démographes parlent de EDVN (espérance de vie à la naissance).

Voici pour information l'augmentation de l'espérance de vie chez les hommes (source: Institut National Français d'Études Démographiques) depuis 1996 à 2006:

1996	74.1
1997	74.5
1998	74.8
1999	75.0
2000	75.3
2001	75.5
2002	75.7
2003	75.9
2004	76.7
2005	76.8
2006	77.2

Tableau: 64.3 - Augmentation de l'espérance de vie à la naissance

Nous pouvons donc observer à l'aide du tableau ci-dessus que l'espérance de vie augmente d'un peu plus d'un an tous les quatre ans depuis plus de 50 ans (et jusqu'à quand...)

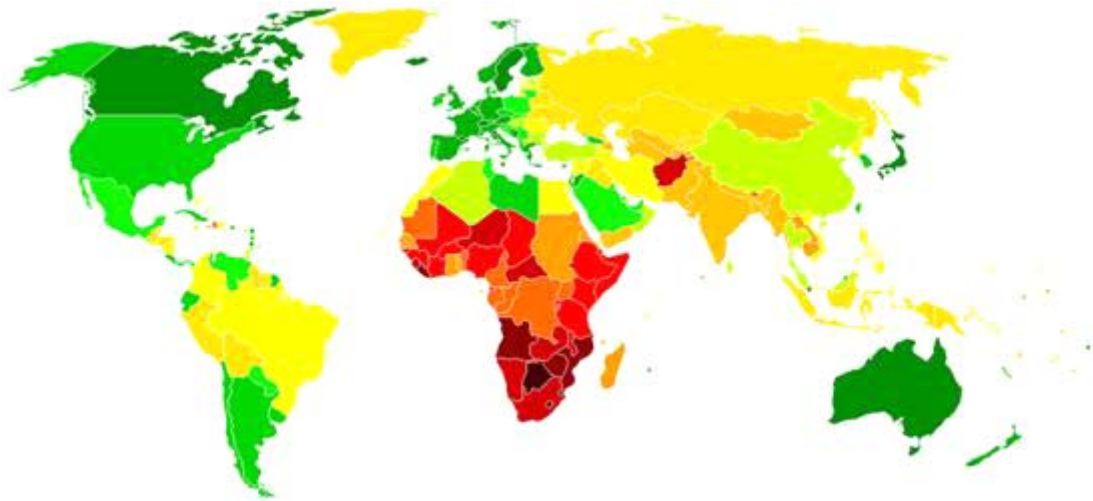
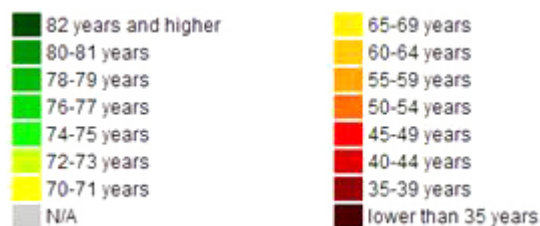


Figure: 64.2 - Espérance de vie au début du 21ème siècle (source: Wikipédia)



Enfin, signalons que les actuaires définissent l'espérance de vie conditionnelle comme étant la somme:

$$e(a, s) = a + E(a) \quad (64.7)$$

à l'âge a pour le sexe s . Ces calculs nous seront utiles pour le calcul des primes d'assurance retraite dans le chapitre Techniques De Gestion.

Passons à un tout autre sujet. Nous savons qu'à ce jour, pour faire un enfant, il faut être deux, certes les démographes le savent aussi. Mais cette double source de chacun des nouveau-nés leur pose de tels problèmes de description et d'analyse de la fécondité qu'ils préfèrent en général l'ignorer. Leur attitude est justifiée par le fait que seule la conception nécessite l'intervention de deux acteurs. Au moment de la naissance, la mère agit seule. Or la démographie ne s'intéresse pas aux conceptions, inaccessibles à l'observation, mais seulement aux naissances.

Nous avons vu comment nous pouvons suivre un ensemble d'hommes ou de femmes nés une année donnée et enregistrer les décès successifs, ce qui permet d'établir la table de mortalité de cette cohorte. De même, nous pouvons noter les effectifs des enfants auxquels ils donnent naissance année après année. Nous obtenons ainsi la "**table de fécondité**".

Il suffit de suivre une cohorte de femmes de 15 jusqu'à 50 ans pour avoir une description complète de son comportement procréateur. Les données fournies par l'état civil permettent de calculer chaque année le nombre des enfants auxquels ont donné naissance les femmes de cette cohorte, regroupées par âge ou par groupes d'âge. En divisant par le taux des femmes survivantes à cet âge, nous obtenons le "**taux de fécondité**" F_a .

Si nous additionnons l'ensemble de ces taux, nous obtenons le nombre d'enfants qu'auraient eu, en moyenne, les femmes de cette cohorte si leur mortalité avait été nulle. Tel n'a évidemment pas été le cas. Pour caractériser la façon dont elles ont assuré le renouvellement de leur génération, il faut

additionner les nombres réels moyens de naissances, produits du taux F_a par le taux de survie S_a .

L'ensemble de ces données est présenté dans une table de fécondité dont voici un exemple où $n(a, a+5)$ est le nombre d'effectifs de naissances chez les femmes survivantes à cette tranche d'âge et où F_a est le nombre de naissances de 1'000 femmes de cette tranche d'âge.

Femmes françaises nées vers 1830:

Âge n	Taux de survie S_a	Effectifs naissances $n(a, a+5)$	Taux de fécondité F_a
15	0.672	91	135
20	0.645	464	720
25	0.616	589	955
30	0.587	475	810
35	0.558	328	588
40	0.528	153	290
Total	1	2'100	3'498

Tableau: 64.4 - Table de fécondité en France en 1830

Par exemple, sur 1'000 petites filles nées en 1830, 645 on atteint l'âge de 20 ans et ont eu 464 enfants entre 20 et 25 ans. L'intensité de la fécondité est mesurée par le nombre de naissances qu'auraient eu 1000 femmes de cet âge:

$$\frac{464}{645} \cdot 1000 = 720 \quad (64.8)$$

Et la moyenne des naissances:

Âge n	Survivantes	Effectifs naissances $n(a, a+5)$
15	672	91
20	645	464
25	616	589
30	587	475
35	558	328
40	528	153
Total		2'100

Tableau: 64.5 - Moyenne des naissances en 1830

Nous pouvons observer qu'en 1830, de 20 à 30 ans les femmes donnaient naissance en moyenne à:

$$\frac{464 + 589}{616} = \frac{1053}{616} = 1.709 \quad (64.9)$$

ce chiffre étant assimilé par le grand public au "**taux de fécondité**" (donc attention à ne pas confondre avec F_a).

Pour faire le calcul sur l'ensemble de tous les âges, il suffit de prendre le rapport du nombre d'enfants

sur le nombre initial de femmes. Cela donne mondialement les chiffres indiqués dans la carte ci-dessous:

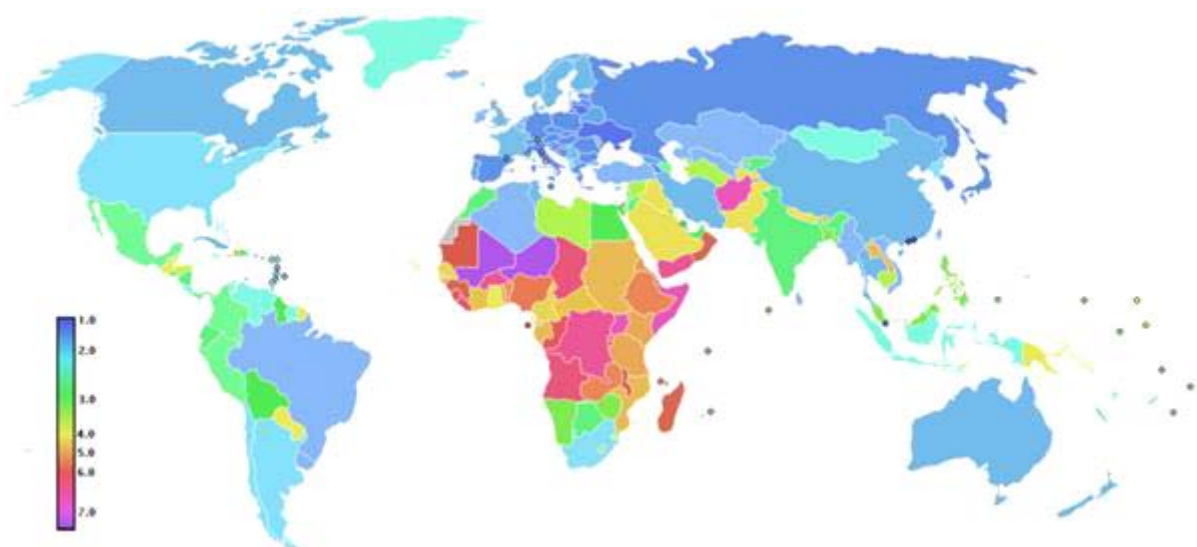


Figure: 64.3 - Taux de fécondité au début du 21ème siècle (source: Wikipédia)

1.1. RENOUVELLEMENT DE LA POPULATION

La question essentielle pour un ensemble humain en renouvellement permanent en raison des flux d'entrées et de sorties que sont les naissances et les décès est: notre effectif est-il en décroissance ou en croissance?

La table de natalité féminine permet de répondre, grâce au rapport du nombre de garçons à la naissance sur le nombre de filles. Ainsi, en ce début du 21ème siècle: il naît en moyenne 105 garçons pour 100 filles dans le monde. Ainsi, la proportion des filles est donc de:

$$\frac{100}{205} = 0.49 \quad (64.10)$$

Vers quarante ans (49 en France selon l'INED), la prépondérance s'inverse et le nombre de femmes l'emporte généralement sur le nombre d'hommes, malgré de notables disparités régionales.

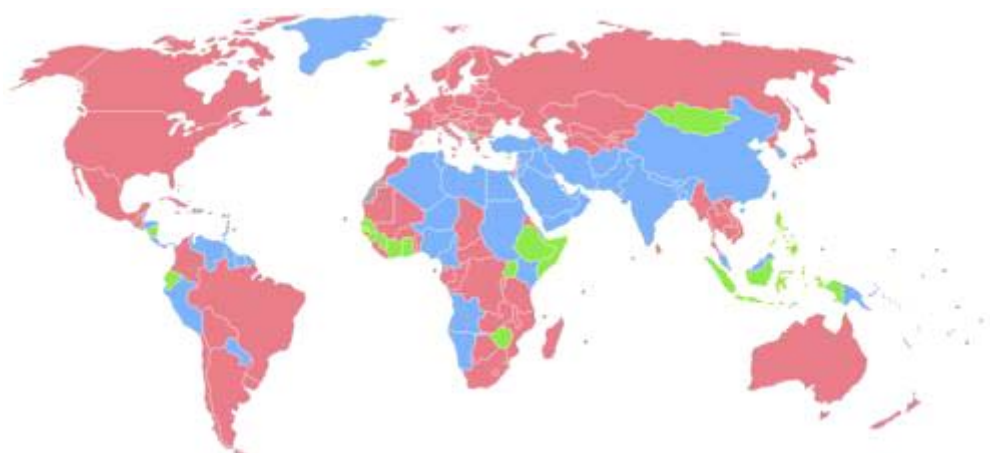
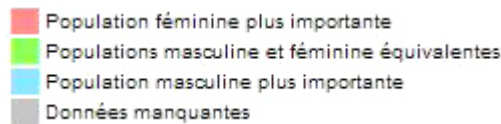


Figure: 64.4 - Répartition Hommes/Femmes au début du 21ème siècle (source: Wikipédia)



En supposant que la proportion était la même à l'époque, la table de fécondité de 1830 montre donc qu'en moyenne une femme de la cohorte de 1830 a engendré:

$$\frac{2100}{1000} \cdot 0.49 = 2.1 \cdot 0.49 = 1.03$$

filles. Soit une augmentation que nous noterons k (en analogie avec le modèle exponentiel que nous verrons plus loin) de 3%. À tort, certains politiciens avancent la valeur (scolaire) 2.1 comme étant le taux de fécondité... qu'il faut pour assurer le renouvellement des générations ce qui n'est donc pas tout à fait exact.

L'effectif féminin était donc en accroissement. Donc la population pouvait assurer son renouvellement (tant que le rapport est supérieur à 1).

Le nombre ainsi obtenu est le "**taux net de reproduction**". Ce taux est donc normalement constitué par le rapport entre le nombre de filles mises au monde par cent femmes, rapport corrigé par la mortalité prévue entre la naissance de ces filles et l'âge moyen à la reproduction, car une partie des filles n'atteindra pas l'âge de la reproduction, étant donnés les décès survenus parmi elles entre leur naissance et leur âge à la maternité. L'âge moyen à la reproduction est donné par l'âge moyen des mères à la naissance.

2. MODÈLES DE POPULATIONS

Il existe de nombreux modèles mathématiques permettant d'étudier la croissance d'une population. Le terme "**population**" est utilisé ici au sens le plus large - il peut s'agir d'une population d'humains, d'animaux, de plantes, de personnes infectées par un virus, etc.

Pour construire un modèle mathématique, il est nécessaire de faire des hypothèses. Ces hypothèses jouent deux rôles: préserver certaines caractéristiques essentielles de la réalité et simplifier suffisamment cette réalité afin qu'elle puisse être étudiée par la mathématique.

2.1. MODÈLE EXPONENTIEL

Dans ce modèle de dynamique des populations (un des plus simples), l'hypothèse sera la suivante: le taux de variation de la population est proportionnel, en tout temps t , à la population $P(t)$ présente au temps t .

Nous pouvons penser, a priori, que cette hypothèse est raisonnable pour une foule de situations. Par exemple, plus la population humaine est grande et plus le taux de variation de cette population, exprimé en nombre de personnes qui s'ajoutent par unité de temps, sera grand. De même, plus il y a de personnes infectées par un virus et plus, dans les semaines qui viennent, il y aura de nouveaux cas de personnes infectées.

Mathématiquement, cette hypothèse peut se traduire à l'aide de l'équation différentielle (cf. chapitre de Calcul Différentiel Et Intégral):

$$\frac{dP}{dt} = kP \quad (64.11)$$

Cette équation différentielle est un modèle mathématique représentant une situation où le taux de croissance de la population est proportionnel à la grandeur de la population en tout temps. Dans ce cas, k est une constante appelée "**taux d'accroissement**" et nous verrons plus loin comment nous pouvons la déterminer. Dans certaines situations, la valeur de k est négative indiquant le fait que la population diminue avec le temps au lieu de croître. Il est évident qu'une solution à cette équation différentielle (cf. chapitre de Calcul Différentiel Et Intégral) est:

$$P(t) = c^{te} e^{kt} \quad (64.12)$$

En premier lieu nous déterminerons la valeur de la constante k , à partir de données démographiques pour l'année 1965. À cette époque, il y avait 3 milliards de personnes sur la planète Terre. De plus, à cette époque, la population augmentait de 54 millions par année.

Ainsi, en 1965:

$$\frac{dP}{dt} = 54 \cdot 10^6 = kP = k \cdot 3 \cdot 10^9 \quad (64.13)$$

Ce qui nous donne:

$$k = 0.018 \quad (64.14)$$

Soit 1.8%. L'équation différentielle est alors:

$$P(t) = c^{te} e^{0.018t} \quad (64.15)$$

Pour déterminer la constante multiplicative, il suffit de poser $t = 0$ et de la choisir en conséquence (puisque'elle correspond à la condition initiale). Ainsi, en 1965 nous avons:

$$P(t) = c_{1965}^{te} e^{0.018t} = 3 \cdot 10^9 e^{0.018t} \quad (64.16)$$

Si ce modèle mathématique est conforme à la réalité, la solution trouvée nous permettra d'estimer la population d'humains sur la terre pour des temps ultérieurs à 1965. Voici le graphique de la fonction:

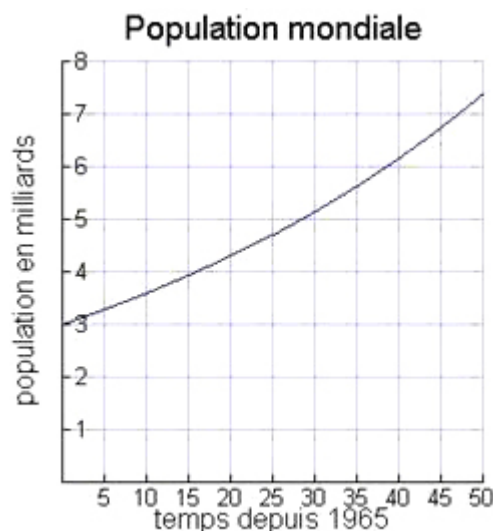


Figure: 64.5 - Évolution de la population mondiale depuis 1965

Si nous évaluons $P(37)$, cela nous fournira la prédiction pour la population en 2002. Nous trouvons 5.84 milliards ce qui est assez proche de la réalité.

Le lecteur pourra également facilement vérifier le tableau suivant selon les valeurs de k :

Taux d'accroissement par an k	Temps de doublement en années
0.5%	139
1%	69
1.5%	46
2%	35
2.5%	28
3%	23
3.5%	20
4%	17

Tableau: 64.6 - Taux d'accroissement de la population selon modèle exponentiel

Nous voyons alors immédiatement, que selon le modèle exponentiel, une croissance qui peut paraître lente, de l'ordre de 3% par an, est en fait véritablement explosive puisqu'elle entraîne un doublement tous les 23 ans. Soit une multiplication par plus de 17 en un siècle. Or, un siècle est vite passé!

Cette simple constatation montre à quel point dans ce modèle sont mensongers les arguments qui nous font chercher la solution d'un problème économique ou social dans la croissance, et même, comme il est souvent admis, dans une croissance durable. Il est clair qu'aucune croissance ne peut véritablement durer, elle n'est qu'un épisode transitoire, nécessairement suivie d'un palier ou même d'une décroissance. Résoudre une difficulté par la croissance ou le maintien de celle-ci c'est reporter le problème à plus tard, à la période où il faudra trouver les moyens à la fois d'arrêter cette croissance et de résoudre autrement cette difficulté restée un temps camouflée.

Le modèle de croissance que nous avons présenté ne peut par contre être valable sur de très longues périodes de temps. En effet, si nous calculions, en utilisant l'équation précédente, la population dans 7 siècles le résultat serait que sur chaque mètre carré de la terre, excluant l'eau, il y aurait, en moyenne, dix humains! De même, une population de personnes infectées par un virus ne peut pas vraiment être décrite par un tel modèle.

Ces derniers résultats nous enseignent que si nous voulons élaborer une modélisation de la croissance d'une population, qui soit davantage conforme à la réalité, il va falloir modifier nos hypothèses initiales. Ce que nous allons de suite voir avec les modèles logistiques déterministes.

Mais avant cela donnons un exemple sympathique où une explosion exponentielle a aussi lieu.

Chaque individu sur Terre a normalement, nous le savons, deux parents, quatre grands-parents et huit arrière-grands-parents et ainsi de suite à chaque génération, si nous remontons le temps, le nombre d'ancêtres double. Mais cela ne peut pas doubler ainsi indéfiniment et la croissance exponentielle doit s'arrêter à un moment ou à un autre! Si le nombre d'ancêtres double à chaque génération et que nous remontons seulement 2'000 ans en arrière, nous aurions à titre individuel à peu près 10^{23} ancêtres en ne considérant que 25 ans pour une génération... Deux mille ans en arrière, ce qui n'est pas si éloigné, et ces ancêtre représenteraient à peu près l'équivalent de la masse totale de la Terre si cette masse n'était faite que de corps humains! La réponse à ce paradoxe apparent est bien sûr qu'il y a énormément (vraiment énormément!) de répétitions dans chaque arbre généalogique (des parents ont plusieurs enfants... et les pères plusieurs femmes... et réciproquement) et que donc nous sommes tous sur une période assez courte cousins...

Nous avons aussi les données actuelles concernant l'évolution de la population mondiale passée:

Année	Population mondiale
-100000	0.5 million
-10000	1 à 10 millions
-6500	5 à 10 millions
-5000	5 à 20 millions
-200	150 à 231 millions
1	170 à 400 millions
200	190 à 256 millions
400	190 à 206 millions
500	190 à 206 millions
600	200 à 206 millions
700	207 à 210 millions
800	220 à 224 millions
900	226 à 240 millions
1000	254 à 345 millions
1100	301 à 320 millions
1200	360 à 450 millions
1250	400 à 416 millions
1300	360 à 432 millions
1400	350 à 374 millions
1500	425 à 540 millions
1600	545 à 579 millions
1650	470 à 545 millions
1700	600 à 679 millions
1750	629 à 691 millions

Année	Population mondiale
1910	1.750 milliard
1920	1.860 milliard
1930	2.07 milliards
1940	2.3 milliards
1950	2.519 milliards
1955	2.757 milliards
1960	3.023 milliards
1965	3.337 milliards
1970	3.696 milliards
1975	4.073 milliards
1980	4.442 milliards
1985	4.843 milliards
1990	5.279 milliards
1995	5.692 milliards
2000	6.085 milliards
2005	6.5 milliards

1800	0.813 à 1.125 milliard
1850	1.128 à 1.402 milliard
1900	1.550 à 1.762 milliard

Tableau: 64.7 - Évolution de la population mondiale (Wikipédia)

Nous y observons bien une croissance exponentielle pour l'instant... mais quand même pas avec des valeurs correspondant au petit calcul de notre exemple de l'arbre généalogique et ce même en cumulé!

2.2. MODÈLE LOGISTIQUE DÉTERMINISTE

Nous allons maintenant nous intéresser à un type de modèle autre qu'exponentiel (où la population explose) qui a l'avantage d'avoir un comportement asymptotique plutôt que divergent.

Ce type de comportement est intéressant car les ressources sont normalement limitées et il y a une compétition entre individus. Le modèle logistique, également appelé "**modèle de Verhulst**" permet de rendre compte de cela relativement bien.

Soit $N(t)$ la population au temps t . Posons:

$$\frac{dN(t)}{dt} = rN(t) \quad (64.17)$$

où r est le taux d'accroissement qui cette fois-ci ne sera pas constant et sera défini comme valant:

$$r = r_0 \left(1 - \frac{N(t)}{K} \right) \quad (64.18)$$

où K est la capacité maximale du milieu. Nous voyons que si K est infini que nous retombons immédiatement sur le modèle exponentiel et que si $N(t)$ égale K alors r est nul.

Finalement, nous avons:

$$\frac{dN(t)}{dt} = r_0 \left(1 - \frac{N(t)}{K} \right) N(t) \quad (64.19)$$

ce qui donne après réarrangement:

$$\frac{dN(t)}{dt} - r_0 N(t) + r_0 \frac{N^2(t)}{K} = 0 \quad (64.20)$$

Soit sous forme mathématique:

$$y' = r_0 \left(y - \frac{y^2}{K} \right) \quad (64.21)$$

avec comme condition initiale que $y(0) = 0$.

Il s'agit maintenant de résoudre cette équation différentielle (cf. chapitre de Calcul Différentiel Et Intégral).

Donc nous avons à résoudre l'équation différentielle:

$$y' = r_0 y \left(1 - \frac{y}{K} \right) \quad (64.22)$$

Nous posons $y(x) = 1/f(x)$. L'équation différentielle devient alors:

$$-\frac{f'}{f^2} = r_0 \frac{1}{f} \left(1 - \frac{1}{Kf} \right) \quad (64.23)$$

ce qui se simplifie en:

$$f' + r_0 f - \frac{r_0}{K} = 0 \quad (64.24)$$

Cette équation différentielle se résout comme à l'habitude. Nous écrivons d'abord l'équation homogène:

$$f' + r_0 f = 0 \quad (64.25)$$

Une solution particulière immédiate est:

$$f(x) = Ae^{-r_0 x} \quad (64.26)$$

où A est une constante. En injectant cette solution dans l'équation différentielle initiale cela donne:

$$-r_0 Ae^{-r_0 x} + r_0 Ae^{-r_0 x} - \frac{r_0}{K} = -\frac{r_0}{K} = 0 \quad (64.27)$$

Nous voyons donc immédiatement que pour que l'égalité soit satisfaite notre solution particulière devient la solution générale si nous l'écrivons sous la forme:

$$f(x) = Ae^{-r_0 x} + \frac{1}{K} \quad (64.28)$$

La solution de l'équation différentielle du début est donc:

$$y = \frac{1}{f} = \frac{1}{Ae^{-r_0 x} + \frac{1}{K}} = \frac{K}{KAe^{-r_0 x} + 1} \quad (64.29)$$

La condition initiale:

$$y(0) = y_0 \quad (64.30)$$

Implique:

$$y(0) = \frac{K}{KAe^{-r_0 \cdot 0} + 1} = \frac{K}{KAe^{-r_0 \cdot 0} + 1} = \frac{K}{KA + 1} = y_0 \quad (64.31)$$

Soit:

$$K = y_0 (KA + 1) \Rightarrow A = \frac{K - y_0}{Ky_0} \quad (64.32)$$

Ce qui nous donne:

$$\begin{aligned} y &= \frac{K}{KAe^{-r_0x} + 1} = \frac{K}{K \frac{K - y_0}{Ky_0} e^{-r_0x} + 1} = \frac{K}{\frac{K - y_0}{y_0} e^{-r_0x} + 1} = \frac{K}{1 + \left(\frac{K}{y_0} - 1 \right) e^{-r_0x}} \\ &= \frac{y_0}{y_0} \frac{K}{1 + \left(\frac{K}{y_0} - 1 \right) e^{-r_0x}} = \frac{y_0 K}{y_0 + (K - y_0) e^{-r_0x}} \end{aligned} \quad (64.33)$$

Soit sous forme traditionnelle dans le domaine de la dynamique de la population:

$$N(t) = \frac{N_0 K}{N_0 + (K - N_0) e^{-r_0 t}} \quad (64.34)$$

qui est donc la relation finale du modèle logistique déterministe. Nous voyons d'ailleurs aisément que:

- si t tend vers l'infini alors l'asymptote horizontale est K
- au temps t nul, la population initiale est bien N_0

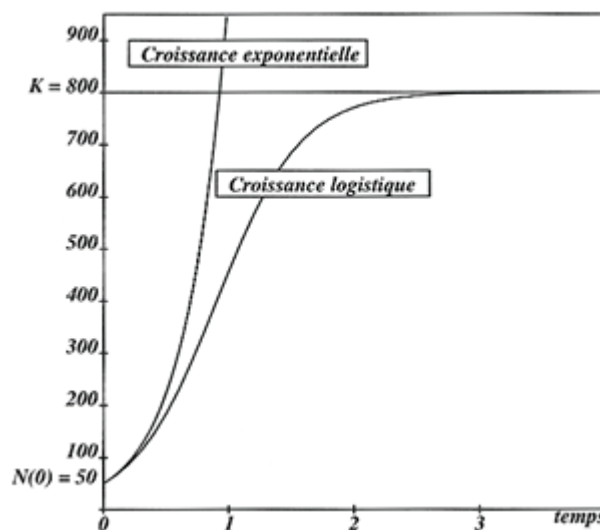


Figure: 64.6 - Différence entre croissance logistique et croissance exponentielle

Pour clore en revenant à la forme suivante de notre développement:

$$y = \frac{K}{1 + \left(\frac{K}{y_0} - 1 \right) e^{-r_0x}} \quad (64.35)$$

Si nous la récrivons:

$$f(x) = \frac{K}{1 + ae^{-rx}} \quad (64.36)$$

nous obtenons alors ce qui est communément appelé par sa forme une "fonction logistique".

2.3. MODÈLE LOGISTIQUE CHAOTIQUE

Nous allons supposer que chaque génération est proportionnelle à la précédente. D'une période à la suivante, l'évolution de la population peut alors se traduire par une suite du type:

$$x_{t+1} = kx_t \quad (64.37)$$

où x_t représente la population à la période t , x_{t+1} la population à la période $t + 1$ et k le taux de reproduction.

Mais on se rend très rapidement compte qu'un tel modèle est irréaliste: la terre aurait depuis longtemps été submergée par une marée humaine. En effet, le modèle exponentiel ignore des facteurs importants, comme celui des ressources qui ne sont pas illimitées.

Pour rétablir un modèle plus satisfaisant, il faut tenir compte du fait que les ressources disponibles sont limitées et que, pour tout territoire, il existe une population maximale au-delà de laquelle la population décroît, quelle que soit l'espèce. Pour trouver une fonction qui traduirait de façon plus réaliste l'évolution d'une population, énumérons, en les simplifiant un peu, les propriétés que doit vérifier cette évolution:

1. C'est un phénomène itératif: si x_{t+1} représente l'effectif de la population de la période $t + 1$, il dépend de celui de la période précédente x_t .
2. Une population ne peut croître indéfiniment sur un territoire délimité: il existe un maximum après lequel elle décroît. Il faut donc prévoir un facteur rétroactif limitant la hausse de population quand sa densité devient trop élevée. Pour simplifier les calculs, x ne représentera pas la population en nombre absolu, mais un pourcentage de ce maximum correspondant à un territoire donné; x ne peut donc fluctuer qu'entre 0 et 1.
3. k représente le taux de croissance effectif d'une période à la suivante.

L'effectif de la population de la période $t + 1$ sera la valeur x_{t+1} , exprimée en pourcentage de la population maximale que peut accueillir le territoire donné, et sera obtenu à partir de l'effectif de la population x_t de la période précédente t par l'équation:

$$x_{t+1} = kx_t(1 - x_t) \quad (64.38)$$

le facteur $(1 - x_t)$ représentant l'effet rétroactif. Cette suite est souvent appelée "suite logistique" et nous la retrouvons également dans de nombreux autres domaines de la physique non-linéaire.

En effet, quand la densité de la population est élevée, proche de la saturation, alors x_t est proche de 1 et, par conséquent, $1 - x_t$ sera proche de 0. Donc, ce facteur rétroactif aura tendance à minimiser la population x_{t+1} .

Nous allons voir que selon la valeur de son taux de croissance effectif k , une population animale, peut tendre vers un état d'équilibre, ou fluctuer entre deux ou quatre ou huit valeurs, ou varier de façon totalement aléatoire.

Ainsi, pour certaines espèces animales, il est normal que les populations varient régulièrement, tandis que pour d'autres, il est normal de tendre vers une situation d'équilibre. Cette variété, ce "chaos" dans certains cas, est liée aux propriétés mathématiques même de l'équation:

$$x_{t+1} = kx_t(1 - x_t) \quad (64.39)$$

La complexité devient la règle et non l'exception, et ce qui semble "chaotique" découle des propriétés bien précises d'une fonction bien précise.

Ce sont ces résultats que l'on va retrouver. Pour cela, on va étudier l'évolution de populations animales dont les taux de croissance effectifs k varient de 1 à 4. On étudiera ces évaluations par calcul et par l'étude du graphe de la fonction logistique.

Voyons quelques cas particuliers:

- Si la valeur initiale x_0 vaut 0.1 et $k = 1$ l'évolution de la population pour les quarante premières périodes est représentée par la courbe ci-dessous:

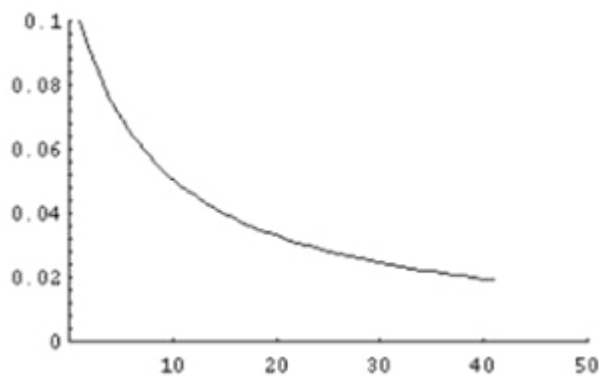
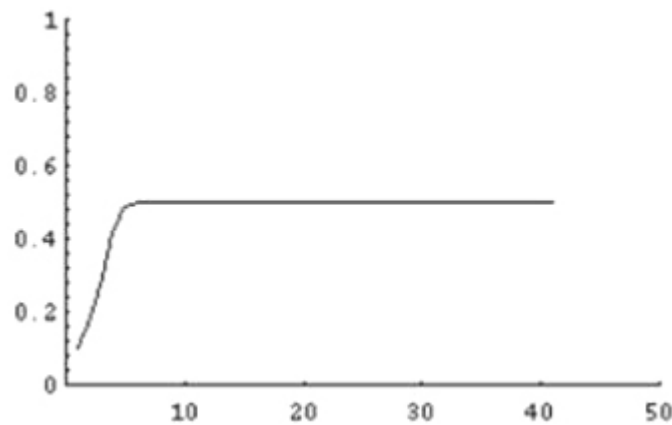


Figure: 64.7 - Plot de l'équation logistique avec $x_0=0.1$ et $k=1$

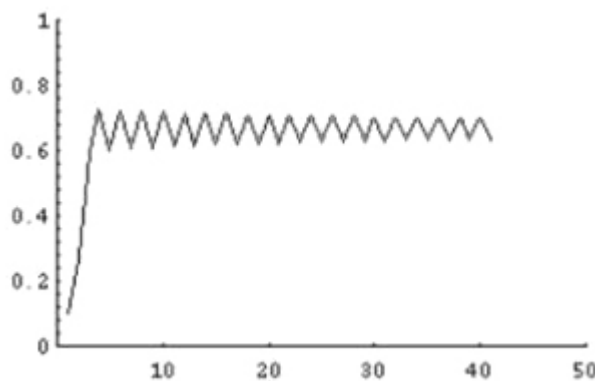
On constate que pour ce taux de croissance et cette condition initiale, la population décroît et va tendre vers zéro.

- Si la valeur initiale x_0 vaut 0.1 et $k = 2$ l'évolution de la population pour les quarante premières périodes est représentée par la courbe ci-dessous:

Figure: 64.8 - Plot de l'équation logistique avec $x_0=0.1$ et $k=2$

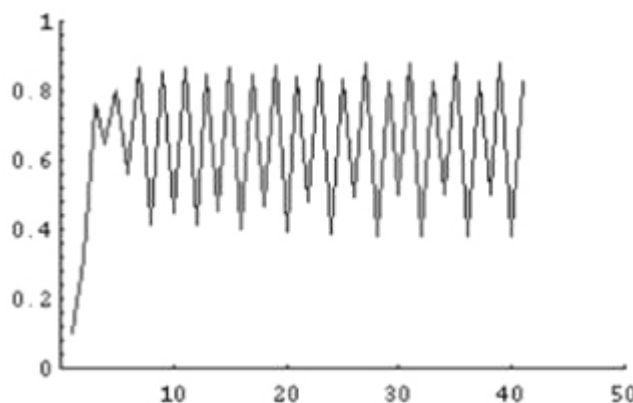
On constate que pour ce taux de croissance, la population va se stabiliser autour d'un nombre correspondant à la moitié de la population que le territoire pourrait accueillir.

- Si la valeur initiale x_0 vaut 0.1 et $k = 3$ l'évolution de la population pour les quarante premières périodes est représentée par la courbe ci-dessous:

Figure: 64.9 - Plot de l'équation logistique avec $x_0=0.1$ et $k=3$

Avec un taux de croissance effectif de 3, le nombre d'individus de cette population se met à osciller entre deux valeurs que l'on peut calculer à l'aide de l'ordinateur. On obtient alors les valeurs 0.64 et 0.68.

- Si la valeur initiale x_0 vaut 0.1 et $k = 3.5$ l'évolution de la population pour les quarante premières périodes est représentée par la courbe ci-dessous:

Figure: 64.10 - Plot de l'équation logistique avec $x_0=0.1$ et $k=3.5$

Avec un taux de croissance effectif de 3.5 la population oscille entre quatre valeurs: 0.39, puis 0.83, puis 0.49, et enfin 0.87. L'évolution d'une population qui a un tel taux de croissance effectif est nettement cyclique.

- Si la valeur initiale x_0 vaut 0.1 et $k = 4.5$ l'évolution de la population pour les quarante premières périodes est représentée par la courbe ci-dessous:

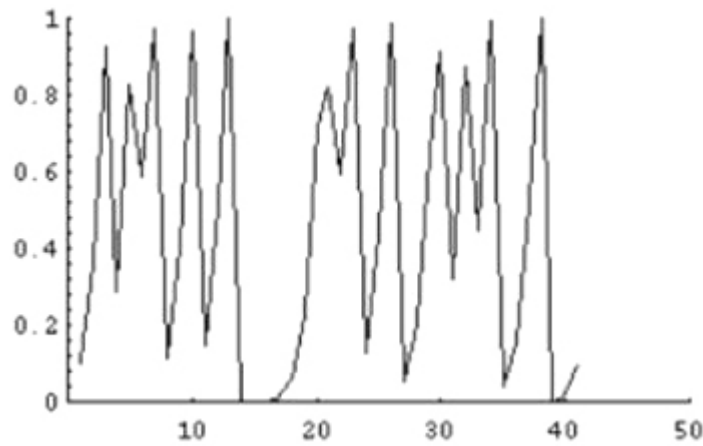


Figure: 64.11 - Plot de l'équation logistique avec $x_0=0.1$ et $k=4.5$

On constate que, lorsque le taux de croissance effectif k est égal à 4, le nombre d'individus des populations des périodes successives semble osciller irrégulièrement, de façon "chaotique" entre les deux extrêmes: la saturation quand x tend vers 1 et l'extinction quand x tend vers 0.

La question maintenant, c'est quel est le comportement de cette fonction pour d'autres valeurs initiales de population?

Pour le savoir, nous reprenons les calculs précédents avec une population initiale représentant 0.8 de 1 population maximale pour un territoire donné et nous comparons les résultats obtenus.

La croissance de la population est toujours donnée par l'équation:

$$x_{t+1} = kx_t(1 - x_t) \quad (64.40)$$

- Selon que la population initiale soit $x_0 = 0.1$ ou $x_0 = 0.8$, on obtient pour $k = 1$:

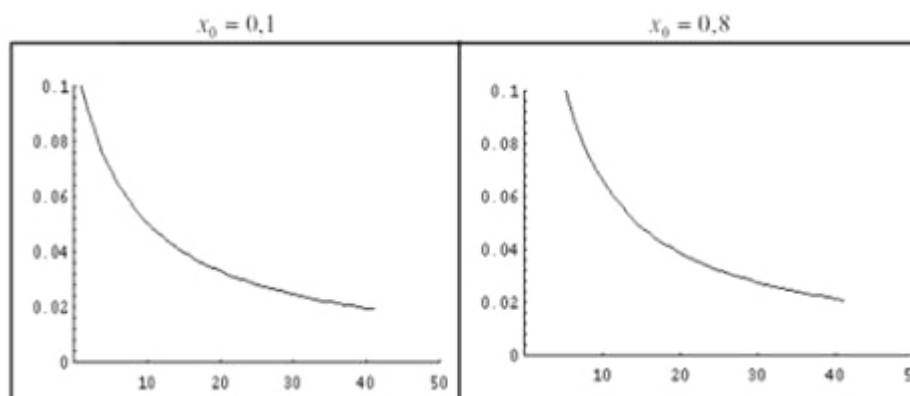


Figure: 64.12 - Plots comparés des équations logistiques

On constate que, quelle que soit la population initiale, elle tend vers zéro.

- Selon que la population initiale soit $x_0 = 0.1$ ou $x_0 = 0.8$, on obtient pour $k = 2$:

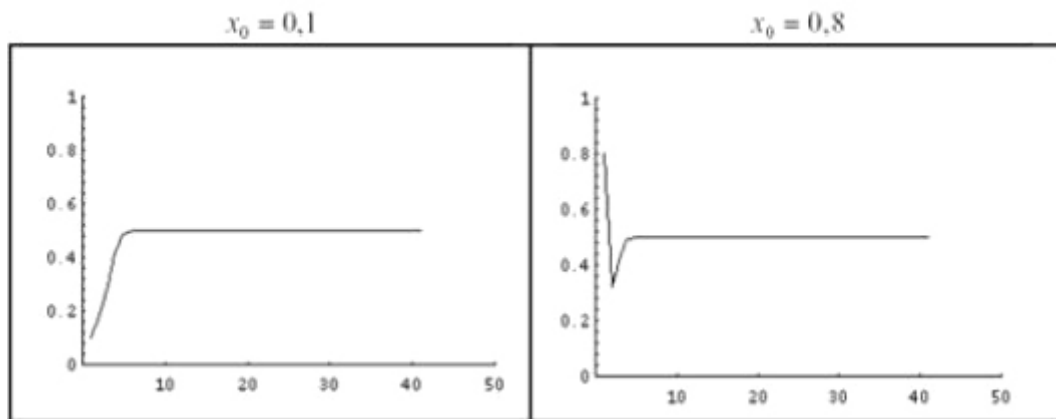


Figure: 64.13 - Plots comparés des équations logistiques

On constate que, quelle que soit la population initiale, elle tend vers une stabilisation.

- Selon que la population initiale soit $x_0 = 0.1$ ou $x_0 = 0.8$, on obtient pour $k = 3.8$:

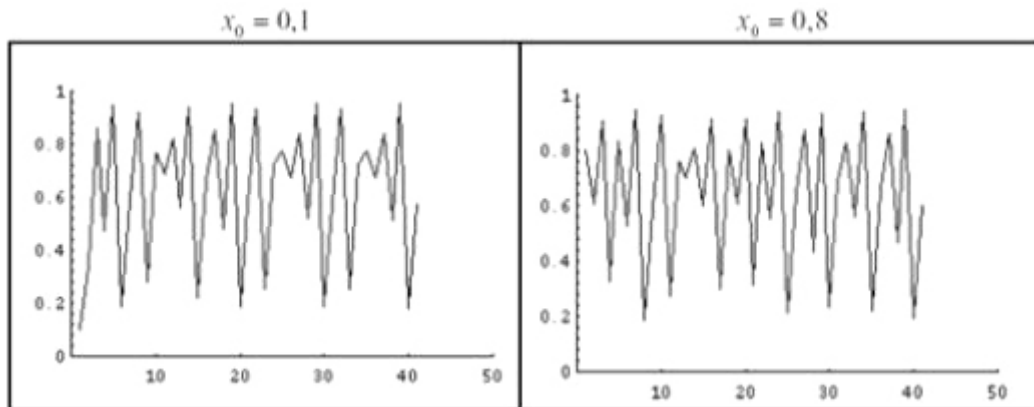


Figure: 64.14 - Plots comparés des équations logistiques

On constate ici que l'évolution des populations est très différente. On a pris des populations initiales éloignées. Si on avait pris des populations initiales très proches, quelle aurait été l'évolution des deux populations ?

- Selon que la population initiale soit $x_0 = 0.8$ ou $x_0 = 0.8001$, on obtient pour $k = 3.8$:

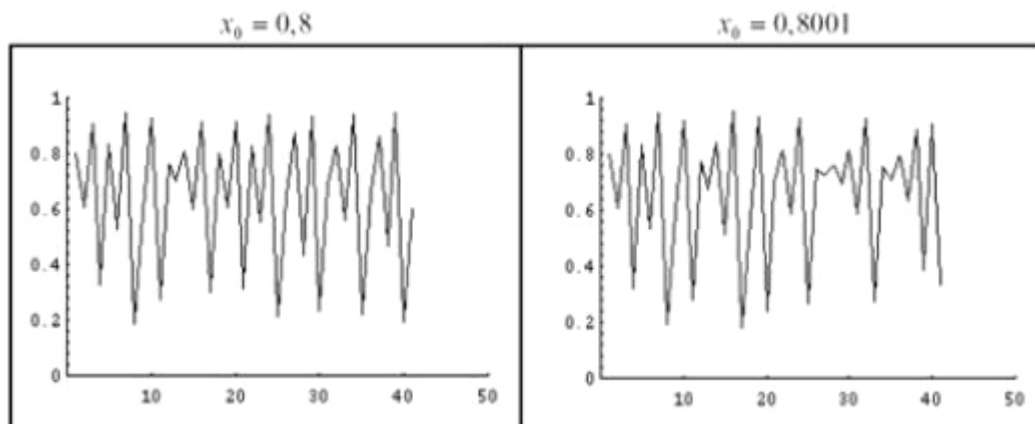


Figure: 64.15 - Plots de l'équation logistique

On constate ici que l'évolution des populations est très différente même si elles étaient très proches initialement et même si l'évolution de ces populations découle d'une fonction très simple. Si on avait pris comme population initiale $x_0 = 0.800001$ on aurait constaté également une évolution très différente de l'évolution des populations.

Conclusion: en dehors de la phase chaotique, la valeur initiale n'a aucune importance, mais dans cette phase chaotique, au contraire, la plus petite variation de valeur initiale change du tout au tout les valeurs suivantes. C'est ce qu'on appelle "**l'effet papillon**".

2.3.1. DIAGRAMME DE FEIGENBAUM

Pour comprendre l'évolution d'une population selon le modèle logistique, on a représenté l'évolution dans le temps, quarante périodes, d'une population correspondant à une valeur initiale précise et à une constante k déterminée. On a constaté, en prenant quelques valeurs particulières de k , que pour ces différentes valeurs, la population avait une évolution différente. On va donc étudier le comportement de la fonction logistique en prenant k comme variable.

En donnant à k des valeurs comprises entre 0 et 4 avec un pas de 0.02, nous allons, pour chacune de ces valeurs, calculer ce que deviendra la population pour chaque période comprise entre la 30^{ème} et la 130^{ème}.

Le graphique ci-dessous représente donc pour chaque valeur du taux de croissance effectif k en abscisse, cent valeurs successives de x_t en ordonnées pour t variant de 30 à 130.

Soit avec Maple 4.00b:

```
>with(plots): with(plottools);
>feigenbaum:=proc(début,fin,pas) local k,itéré,a,b,s;
s:={}; a:=début;
while a<=fin do itéré:=0.1;
  for k to 50 do itéré:=a*itéré*(1-itéré) od;
  for k to 100 do itéré:=a*itéré*(1-itéré);
    s:=s union {[a,evalf(itéré,4)]};
  od;
  a:=a+pas
od;
plot([op(s)],'a'=début..fin,style=POINT,symbol=POINT)
end:
>feigenbaum(1,4,0.01);
```

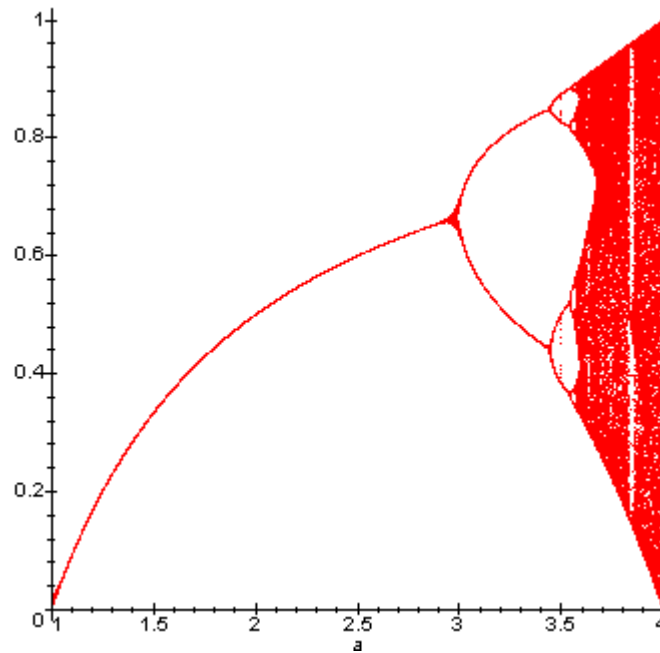


Figure: 64.16 - Diagramme de Feigenbaum

Ce graphique (où l'on peut observer des bifurcations), est appelé "**diagramme de Feigenbaum**", du nom du physicien Mitchell Feigenbaum qui l'a étudié en profondeur et a montré qu'on le retrouvait dans de nombreux phénomènes naturels.

On va maintenant examiner deux propriétés remarquables de ce diagramme: le doublement de période et sa dimension fractale.

Si on revient à notre exemple où la fonction logistique donne l'évolution d'une population animale selon son taux de croissance effectif k , le diagramme de Feigenbaum indique que, quand ce taux est inférieur à 3, le système tend vers un état final stable. Cela correspond en général à notre intuition influencée par notre désir inconscient d'ordre et de simplicité.

Mais à partir de 3, cela se complique: le nombre d'individus par génération se met à osciller entre 2 puis 4, puis 8 valeurs... pour enfin entrer dans une zone "chaotique" où toute valeur semble possible. Il y a déjà quelque chose de fascinant. Mais si on observe de plus près cet intervalle de doublement qui précède la zone chaotique, on va découvrir des résultats encore plus curieux.

On prend ainsi d'abord la séquence des trois premiers doublements avec 2, puis 4, puis 8 branches, pour un taux de croissance effectif portant sur un intervalle $k = [3, 3.564407]$ et on obtient le graphique suivant:

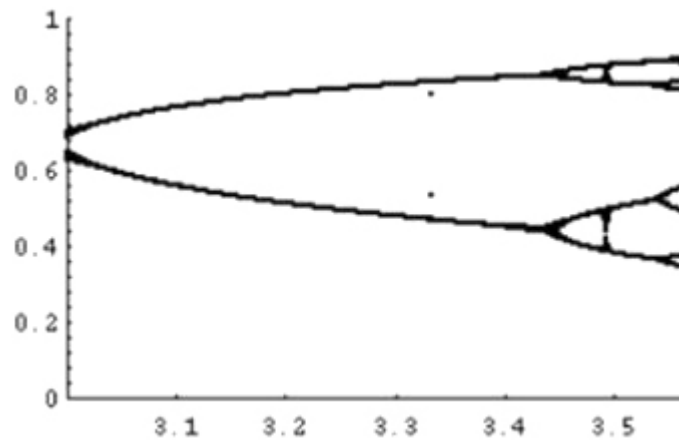


Figure: 64.17 - Bifurcations dans le diagramme de Feigenbaum

Il est remarquable d'observer également sur le graphique ci-dessus qu'il y a deux points qui sont totalement stables (ceux juste après 3.3)!

Si on prend maintenant l'intervalle $k = [3.5, 3.569891]$ on obtient le diagramme de Feigenbaum suivant:

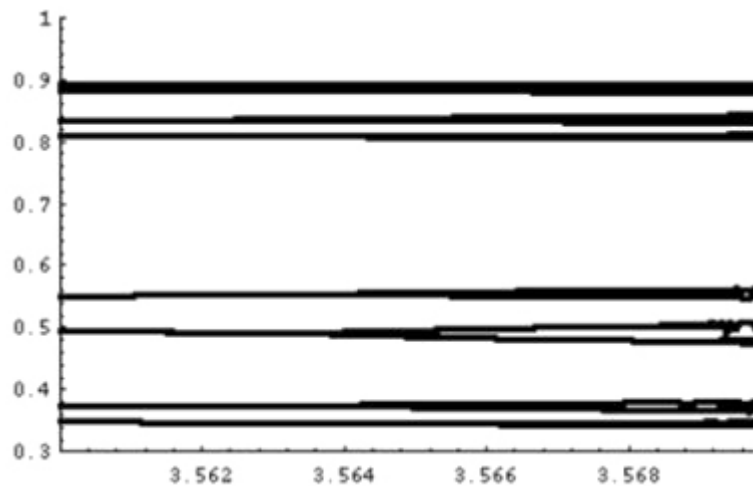


Figure: 64.18 - Multiplication des bifurcations

On constate sur le graphique ci-dessus que les bifurcations se multiplient à partir de points de plus en plus rapprochés et sur des intervalles de plus en plus courts. Le physicien Feigenbaum a démontré deux résultats curieux:

- les bifurcations vont se multiplier à l'infini (!) sur un intervalle qui ne dépassera pas le point de d'abscisse 3.5699456 appelé le "point de Feigenbaum" ou "porte d'entrée sur le chaos", car, après ce point, le système devient chaotique. Il se met à fluctuer entre des valeurs "imprévisibles" et devient extrêmement sensible aux conditions initiales.
- la longueur des intervalles propres aux différentes classes de bifurcations (2, 4, 8, 16,...) diminue dans un rapport constant, 4.6692... appelé, bien sûr, "**constante de Feigenbaum**".

Peitgen indique que ces bifurcations, points et constante de Feigenbaum, ne se retrouvent pas seulement dans le cas de la fonction logistique étudiée par May (voir plus loin), mais dans de nombreux phénomènes physiques comme l'hydrodynamique, l'électronique, les lasers ou l'acoustique.

Remarquons enfin un dernier phénomène bien curieux, celui des "fenêtres". Si on examine plus attentivement la zone chaotique comprise entre le point de Feigenbaum et $k = 4$, on constate qu'il

existe des zones étroites où la fonction logistique se remet à osciller entre un nombre fini de bifurcations avant de replonger dans le chaos.

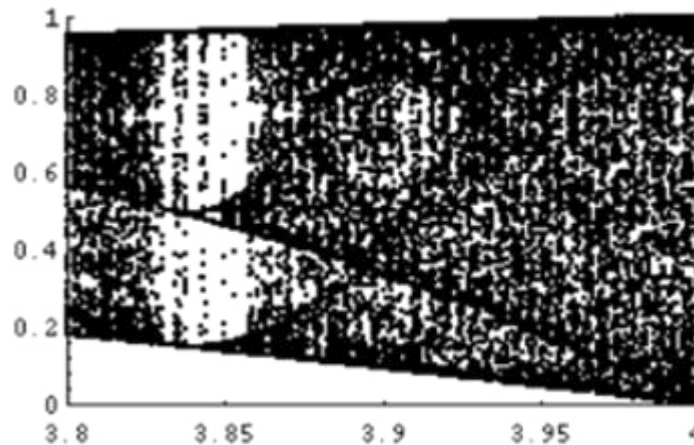


Figure: 64.19 - Visualisation des fluctuations de fond

On constate que, dans cette fenêtre, il y a des bifurcations qui ressembleraient au diagramme de Feigenbaum lui-même si on l'agrandissait.

On aurait pu reprendre les calculs entre $k = [3.848, 3.850]$ et on aurait alors constaté qu'on retrouve la même figure et que, si on l'agrandissait à l'infini, on retrouverait toujours (!) la même figure toujours et toujours:

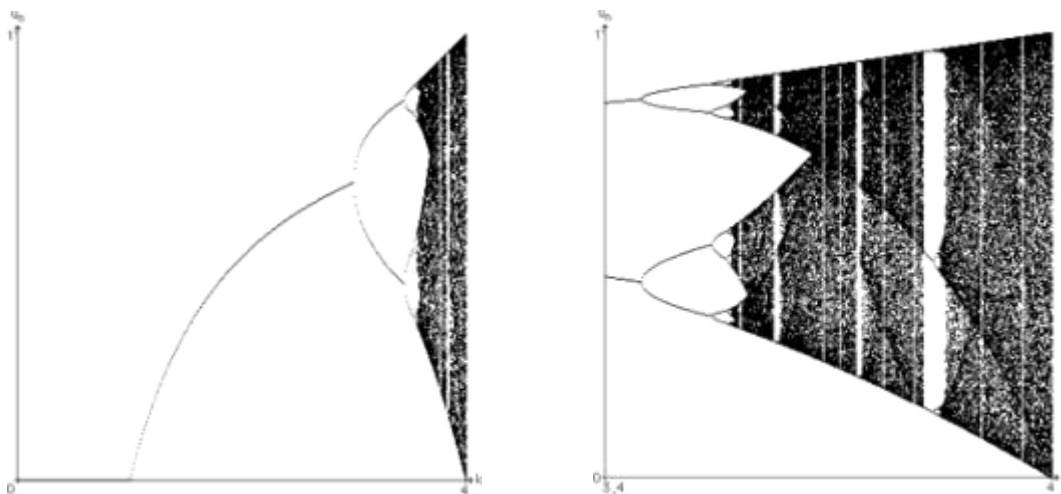


Figure: 64.20 - Autosimilarité du diagramme de Feigenbaum

On vient ainsi de découvrir une figure fractale: "**l'attracteur étrange de la fonction logistique**" appelé aussi "**diagramme de bifurcation de l'équation logistique**" ou encore simplement "**diagramme de Feigenbaum**".

Nous verrons (sans démonstration) dans le chapitre sur les Fractales que le diagramme de Feigenbaum ci-dessus est intimement lié à la fractale de Mandelbrot.

Grâce à la théorie mathématique du chaos appliquée à la dynamique des populations, l'écologie reçoit une stimulation décisive. Jusqu'au début des années 1960, le débat sur la dynamique des populations opposait les tenants d'une théorie déterministe, voyant des évolutions régulières de populations ne subissant qu'exceptionnellement de brutales variations, à ceux pensant ces évolutions comme purement aléatoires. Mais de nombreux faits restaient mal expliqués. En particulier les explosions cycliques de

certaines populations et leur périodicité étrange n'entraient dans aucune des deux explications. En montrant que des modèles déterministes peuvent donner naissance à un comportement aléatoire, Robert May, réconcilie ces points de vue à partir d'une théorie plus profonde. Ce qui apparaît à niveau d'appréhension comme une instabilité généralisée peut se concevoir à un autre comme un chaos stable. Une situation mathématique chaotique peut se révéler stable au point de vue écologique.

Révolution aux conséquences profondes dans la théorie de l'écologie, et aux implications pratiques non moins importantes, l'approche de Robert May est un trait de lumière dans la situation théorique passablement confuse qui règne encore dans la dynamique des populations.

Si, à cause de la montée en puissance des ordinateurs, la théorie du chaos a permis des avancées importantes dans différentes disciplines, elle soulève un débat majeur, celui du déterminisme. Dans quelle mesure la science permet-elle de prédire l'avenir ? Pour certains, les résultats obtenus jusqu'ici dans le cadre de la théorie du chaos prouvent l'importance des conditions initiales. Pour ces personnes, les équations déterministes n'ont qu'une portée limitée et l'avenir restera imprévisible. Pour d'autres au contraire, les résultats obtenus montrent que l'on peut trouver un ordre et des lois dans ce qui peut sembler chaotique. Ces lois sont tout simplement plus complexes.

C'est un débat ouvert et vif qui déborde largement le cadre des milieux scientifiques.

2.4. LOI DE MALTHUS

Nous avons vu précédemment quelques modèles déterministes et chaotiques utilisant des taux d'accroissement pour la simulation. Introduisons maintenant un autre type de modèle utilisant le numéraire des enfants, des femmes et de leur fécondité respective.

Pour ce modèle, nous énonçons les hypothèses suivantes:

H1. Au départ, N individus dont $N/2$ sont des femelles.

H2. Le taux d'accroissement r est supposé constant à cause d'un taux de fécondité f constant aussi.

H3. La population s'accroît.

Nous avons donc comme données: la population au temps t notée $N(t)$, la population féminine notée $N_F(t)$ et $N_E(t)$ le nombre d'enfants.

Avec les relations suivantes:

$$\begin{aligned} N_F(t) &= N(t) / 2 \\ N_E(t) &= f \cdot N_F(t) \end{aligned} \quad (64.41)$$

et donc:

$$N(t+1) = N(t) + N_E(t) = N(t) + f N_F(t) = N(t) + f \frac{N(t)}{2} = N(t) \left(1 + \frac{f}{2} \right) = r N(t) \quad (64.42)$$

donc le taux d'accroissement est:

$$r = 1 + \frac{f}{2} \quad (64.43)$$

Finalement nous obtenons la loi de Malthus:

$$N(t) = r^t N(0) \quad (64.44)$$

ce modèle est donc continu, prend en compte la fécondité mais diverge...

2.5. MODÈLE DE LESLIE

Le modèle de Leslie est un peu plus perfectionné que les autres modèles déterministes mais tout aussi empirique (il est possible comme dans tout domaine de la science de construire des modèles théoriques valides et toujours plus complets et complexes).

Outre le taux de fécondité et de mortalité, il permet de prendre en compte les tranches d'âges de la population et certaines de leurs propriétés relativement aux deux facteurs précités. L'inconvénient de ce modèle est cependant les trop nombreux paramètres à déterminer pour l'ensemble des classes d'âge...

Le système se base sur la découpe de la population en tranches d'âges tels que par exemple:

- $N_1(t)$: est le nombre d'individus de 1 an
- $N_2(t)$: est le nombre d'individus de 2 ans
- ...
- $N_{10}(t)$: est le nombre d'individus de plus 10 ans
- ...

Ensuite, l'idée est que l'évolution d'une classe d'âge dépend des autres classes d'âge. Par exemple, les naissances sont données par le taux de reproduction r sommé sur toutes les classes d'âge (bien évidemment pour certaines d'entre elles le taux est nul...) tel que:

$$N_1(t+1) = \sum_{i=2}^T r_i N_i(t) \quad (64.45)$$

Il est de tradition et raisonnable pour l'être humain d'admettre que normalement seules les classes d'âge pour lesquelles $i \geq 11$ et $T = 55$ entrent en ligne de compte en ce qui concerne la reproduction ce qui implique que normalement les r_i sont non nuls que pour cet intervalle.

Ensuite, le vieillissement et la mortalité m seront pris en compte par les relations (nous reprenons les indices sur toute la population car évidemment il n'y pas que les individus entre 11 et 55 qui décèdent...) :

$$N_i(t+1) = N_{i-1}(t) - m_{i-1} N_{i-1}(t) \quad \forall i \in \mathbb{N}, i = 2 \dots T \quad (64.46)$$

Il est relativement aisé de voir que ces équations peuvent être mises sous forme matricielle (cf. chapitre d'Algèbre Linéaire) de la manière suivante:

$$\begin{pmatrix} N_1(t+1) \\ N_2(t+1) \\ \vdots \\ N_i(t+1) \\ \vdots \\ N_T(t+1) \end{pmatrix} = \begin{bmatrix} 0 & r_2 & \cdots & r_i & \cdots & r_T \\ 1-m_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1-m_2 & 0 & 0 & 0 & 0 \\ \vdots & 0 & \ddots & 0 & 0 & \vdots \\ \vdots & 0 & 0 & 1-m_i & 0 & \vdots \\ 0 & 0 & 0 & 0 & 1-m_{T-1} & 0 \end{bmatrix} \begin{pmatrix} N_1(t) \\ N_2(t) \\ \vdots \\ N_i(t) \\ \vdots \\ N_T(t) \end{pmatrix} \quad (64.47)$$

où la matrice contenant les coefficients de mortalité et de reproduction est appelée "**matrice de Leslie**".

De manière plus compacte cela s'écrit:

$$N(t+1) = A \cdot N(t) \quad (64.48)$$

ou en partant de la population initiale:

$$N(t) = A \cdot N(0) \quad (64.49)$$

Il est possible de faire des analyses très intéressantes sur ce modèle relativement à l'âge de faire des enfants et aux conséquences y relatives. Ce modèle est relativement beaucoup utilisé en biologie marine.

3. PROPAGATION DES ÉPIDÉMIES

La "**théorie des épidémies**" fournit de nombreux systèmes intuitifs d'équations différentielles ou aux dérivées partielles en temps discret dans lesquelles interviennent des phénomènes de contamination, de diffusion, etc.

Nous allons ici prendre l'exemple de la diffusion de la rage dans une population de renards, et en présenter un modèle assez simple où nous considérons que les individus (les renards) peuvent se trouver dans trois états différents:

1. Il y a des renards sains S
2. Des renards infectés I
3. Et des renards morts R

Cette hypothèse de classification en trois catégories fait que ce modèle de propagation des épidémies est souvent appelé "**modèle SIR**".

Pour les plus optimistes, nous pouvons aussi utiliser le modèle qui suppose que certains individus sont immunisés et ne peuvent donc plus être infectés. Finalement, à trois états différents, nous pouvons lier trois équations d'évolutions propres telles que:

$$\begin{aligned}
\frac{\Delta S}{\Delta t} &= S' = -r \cdot I \cdot S \\
\frac{\Delta I}{\Delta t} &= I' = r \cdot I \cdot S - \alpha \cdot I \quad (64.50) \\
\frac{\Delta R}{\Delta t} &= R' = \alpha \cdot I
\end{aligned}$$

avec:

$$\begin{aligned}
S_{t+1} &= S + S' \cdot \Delta t \\
I_{t+1} &= I + I' \cdot \Delta t \quad (64.51) \\
R_{t+1} &= R + R' \cdot \Delta t
\end{aligned}$$

où les S , I , R sont donnés en % de la population totale.

Dans ce modèle, la première équation correspond au phénomène de contamination: lorsque cohabitent des individus sains et des individus infectés, un certain nombre d'individus sains sont infectés. Il est par ailleurs naturel de considérer ce terme comme proportionnel au produit $I \cdot S$. En effet, la quantité de microbes dans le milieu défini (donc le facteur r pour un individu sain donné d'être infecté en présence d'un individu infecté) est proportionnelle à I selon le facteur. Il nous faut ensuite multiplier cette probabilité par le nombre d'individus sains, c'est-à-dire par S . Le signe négatif est quant à lui présent pour signifier la diminution de la population saine.

Pour ce qui est des variations de I , le premier terme correspond aux individus contaminés (qui augmentent I). Le deuxième terme, en $-aI$, correspond aux individus qui meurent.

Le tracé ci-dessous, montre un calcul de l'évolution des populations S , I , R en fonction du temps (une seconde étant considérée comme une boucle de l'algorithme). Selon les valeurs des paramètres a , r les comportements sont complètement différents: si a est trop élevé par exemple, les individus infectés meurent presque tout de suite et n'ont pas le temps d'en infecter beaucoup d'autres (...). Voici un tracé de l'évolution des populations au cours du temps, effectué avec l'algorithme suivant avec $r = 4, a = 2$

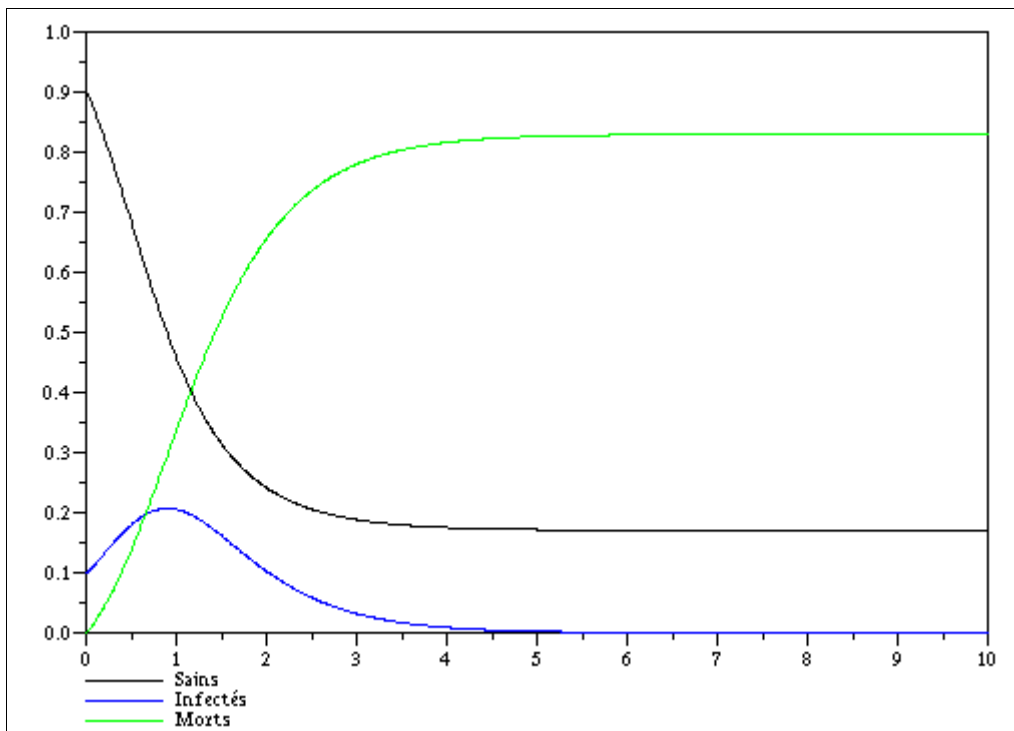


Figure: 64.21 - Évolution des renards sains, infectés et morts en fonction du temps

L'algorithme étant du type suivant:

Pour $i=1$ à (temps max)/(par du temps) faire

$$dS = -r \cdot S(i) \cdot I(i);$$

$$dI = r \cdot S(i) \cdot I(i) - a \cdot I(i);$$

$$dR = a \cdot I(i);$$

$$S(i+1) = S(i) + dt \cdot dS;$$

$$I(i+1) = I(i) + dt \cdot dI;$$

$$R(i+1) = R(i) + dt \cdot dR;$$

Fin Pour

Nous pouvons bien sûr arriver à de nombreux modèles suivant les hypothèses que nous faisons sur la maladie étudiée. Il sera souvent intéressant de définir plus de catégories de populations. Ainsi, pour des modèles de M.S.T. (Maladies Sexuellement Transmissibles), nous différencierons les populations suivant leur sexe (il faudrait aussi prendre l'âge en compte), ce qui multiplie le nombre de possibilités déjà par deux. Pour définir le système différentiel sous-jacent, il devient ainsi préférable de modéliser le processus par un schéma.

Ainsi, pour étudier une infection du type SIDA (Syndrome d'Immunodéficience Acquise), nous pouvons par exemple faire un modèle dans lequel nous considérons quatre groupes de populations: les individus sains X, les individus infectés Y, les porteurs sains Z, et enfin les individus malades A. Dans le schéma qui suit, les flèches entre les différents états représentent les transitions possibles pour un individu donné avec les probabilités correspondantes:

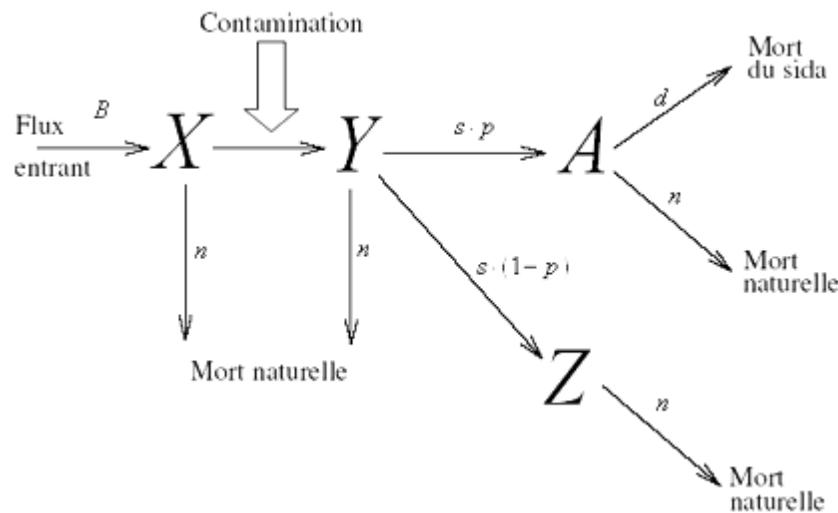


Figure: 64.22 - Chaîne de Markov d'une contamination mortelle

En effet, du point de vue de l'étude du SIDA (et uniquement du SIDA), un individu donné, quel que soit son état, peut mourir d'autre chose que le SIDA, c'est ce que nous rangeons dans "mort naturelle". Nous considérons qu'il y a un taux de mortalité fixe, noté n , et ce pour toutes les populations considérées (sains, infectés, malades ou porteurs sains). D'autre part, un individu sain X peut aussi être contaminé, auquel cas il devient un individu infecté Y . Un individu infecté peut lui évoluer vers deux stades possibles: malade A ou porteur sain Z .

Le nombre s représente la vitesse d'évolution de la maladie: plus ce taux est grand, plus les individus infectés évoluent vite vers l'un des stades suivants. Le nombre p , compris entre 0 et 1, représente la probabilité pour que cette évolution se fasse vers l'état de malade. Bien sûr, si p diminue, la quantité de porteurs sains, immunisés contre la maladie, augmente. La maladie est alors moins dangereuse globalement. Enfin, un malade peut mourir de sa maladie, ce avec un taux de mortalité d considéré comme fixe, alors qu'un porteur sain, immunisé, ne peut plus évoluer du point de vue de cette maladie. Toutes ces hypothèses conduisent à poser des équations du système de la forme:

$$\begin{aligned}
 X' &= B - nX - \lambda_c X \\
 Y' &= \lambda_c X - (n + s)Y \\
 A' &= s \cdot p \cdot Y - (d + n)A \\
 Z' &= s(1 - p)Y - nZ
 \end{aligned} \quad (64.52)$$

où:

$$\lambda_c = \frac{\beta Y}{X + Y + Z + A} \quad (64.53)$$

En effet, ce coefficient représente un taux de contamination, qu'il est naturel de relier au rapport $Y / (X + Y + Z + A)$, probabilité lors d'une rencontre pour que la personne rencontrée soit infectée, donc contaminante.

4. MODÈLE PROIES-PRÉDATEURS (DE LOTKA-VOLTERRA)

Ce modèle d'interaction proies-prédateurs a été proposé par Volterra après la première guerre mondiale. Il s'agissait d'expliquer la dynamique des populations de sardines et de requins en mer Adriatique (et hop! un peu d'océanographie sur le site...); expliquer notamment pourquoi les quantités de sardines pêchées après l'interruption due à la guerre n'étaient plus aussi importantes que précédemment (ce qui peut sembler contre-intuitif) et pourquoi à la reprise de la pêche la proportion observée de requins avait augmenté.

Supposons d'abord qu'il n'y ait pas d'interaction entre les 2 populations (sardines et requins)!

Reprenant ce que nous avons vu tout au début de ce chapitre et en notant $N(t)$ le nombre d'individus proies et $P(t)$ le nombre d'individus prédateurs, nous avons (soit implicitement une croissance exponentielle du nombre de proies en l'absence de pêche):

$$\frac{dN}{dt} = rN \quad (64.54)$$

et en l'absence de proies (dans le cas d'une pêche), le nombre de prédateurs diminuant exponentiellement, nous avons respectivement:

$$\frac{dP}{dt} = -mP \quad (64.55)$$

À ce point du discours, nous devons considérer deux espèces (proies et prédateurs) qui ne sont pas isolées, mais en interaction. Pour quantifier la contribution de l'interaction entre espèces, nous considérerons seulement la prédation, en assumant que sa valeur ou intensité (de l'interaction) est fonction de la probabilité de rencontre des proies-prédateurs qui sera elle supposée proportionnelle au produit $N \cdot P$ des pourcentages des deux populations.

Ces rencontres n'ont pas les mêmes effets sur les deux espèces. Premièrement, bien sûr, chaque proie mangée par un prédateur est un gain net pour la population de ce dernier et une perte nette pour le premier. Ainsi, si l'effet des interactions est accepté comme étant proportionnel à $N \cdot P$, les signes de l'influence d'interaction différeront pour les deux espèces de sorte que nos équations avec interaction cette fois deviennent:

$$\begin{cases} \frac{dN}{dt} = rN - cNP \\ \frac{dP}{dt} = -mP + bNP \end{cases} \quad (64.56)$$

Ce système d'équations est appelé le "**modèle de Lotka-Volterra**".

Avant d'aller plus loin, cherchons les valeurs pour lesquelles les dérivées s'annulent (ce qui nous donnera au fait le point d'équilibre du système):

$$\begin{aligned} rN' - cN'P' &= 0 \\ -mP' + bN'P' &= 0 \end{aligned} \quad (64.57)$$

où N' et P' sont les valeurs qui annulent les dérivées (à ne pas confondre cette fois-ci avec la notation

parfois d'usage pour condenser l'écriture d'une dérivée!!!).

D'où:

$$\begin{aligned}(r - cP')N' &= 0 \\ (bN' - m)P' &= 0\end{aligned}\quad (64.58)$$

Une solution triviale est la "**solution d'extinction**" (appelée aussi parfois "**solution critique**") donnée par:

$$N'_e = P'_e = 0 \quad (64.59)$$

Sinon, nous avons aussi comme possible solution:

$$\begin{aligned}r = cP' &\rightarrow P' = \frac{r}{c} \\ bN' = m &\rightarrow N' = \frac{m}{b}\end{aligned}\quad (64.60)$$

Maintenant, normalisons les équations de Lotka-Volterra en écrivant (ainsi elles sont sans dimensions):

$$\begin{aligned}x = \frac{N}{N'} = \frac{b}{m} N &\rightarrow N = \frac{m}{b} x \\ y = \frac{P}{P'} = \frac{c}{r} P &\rightarrow P = \frac{r}{c} y\end{aligned}\quad (64.61)$$

Avec cette normalisation, le modèle de Lotka-Volterra s'écrit:

$$\begin{aligned}\frac{m}{b} \frac{dx}{dt} &= \frac{mr}{b} x - \frac{cmr}{bc} xy \\ \frac{r}{c} \frac{dy}{dt} &= \frac{bmr}{bc} xy - \frac{mr}{c} y\end{aligned}\quad (64.62)$$

En réarrangeant les coefficients, le système s'écrit finalement à l'équilibre (hors de la solution d'extinction):

$$\begin{aligned}\frac{dx}{dt} &= rx(1 - y) \\ \frac{dy}{dt} &= my(x - 1)\end{aligned}\quad (64.63)$$

pour lequel les dérivées s'annulent aux points unitaires.

Le tracé discret de ce système d'équations (dans lequel nous reconnaissons les termes logistiques vus plus haut), donne avec $m = r = b = c = 1$ et les conditions initiales $(x_0, y_0) = (4, 1)$:

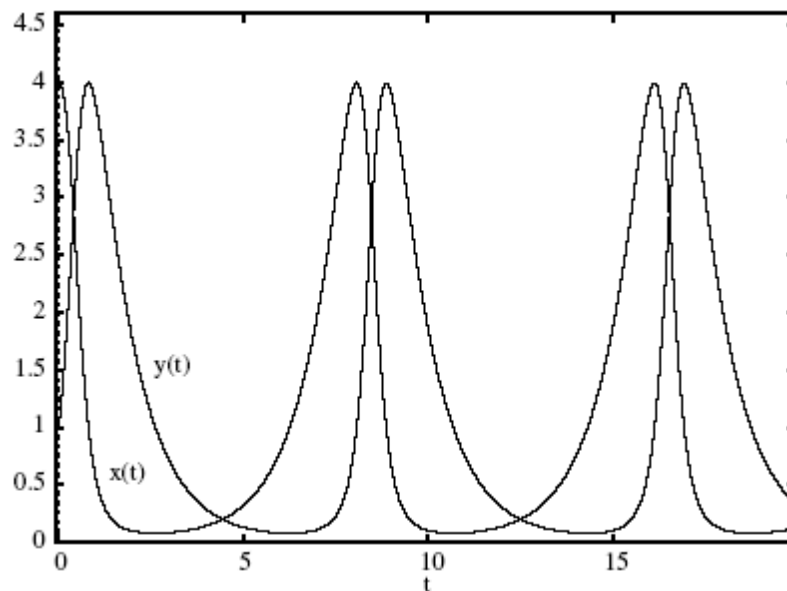


Figure: 64.23 - Fluctuation théorique des populations

En comparaison, voici un exemple pratique (réel) de mesures de proies-prédateurs (Lièvres ("Hares" en anglais)- Lynx) par la Hudson Bay-Company (quand on cherche des données qui concordent à un modèle, on finit souvent par trouver...):

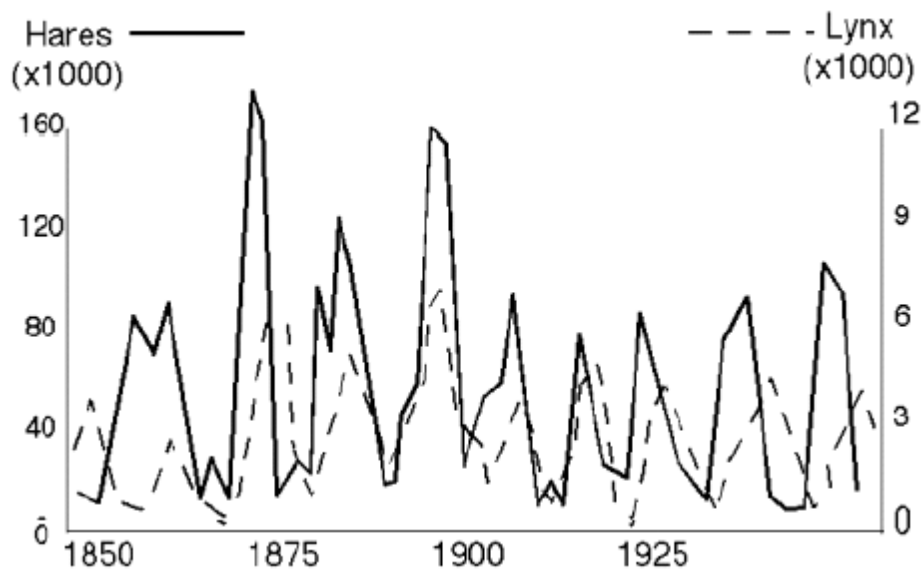


Figure: 64.24 - Fluctuation mesurée des populations

Les deux diagrammes précédents représentent donc les variables x, y en fonction du temps. Cependant, ce qui peut être intéressant pour un scientifique (ou un écologiste...) c'est la représentation de y en fonction de x et inversement. Ainsi, nous obtenons pour les mêmes conditions initiales m, r, b, c précédentes et pour des conditions (x_0, y_0) diverses (les prédateurs sont en ordonnée et les proies en abscisse):

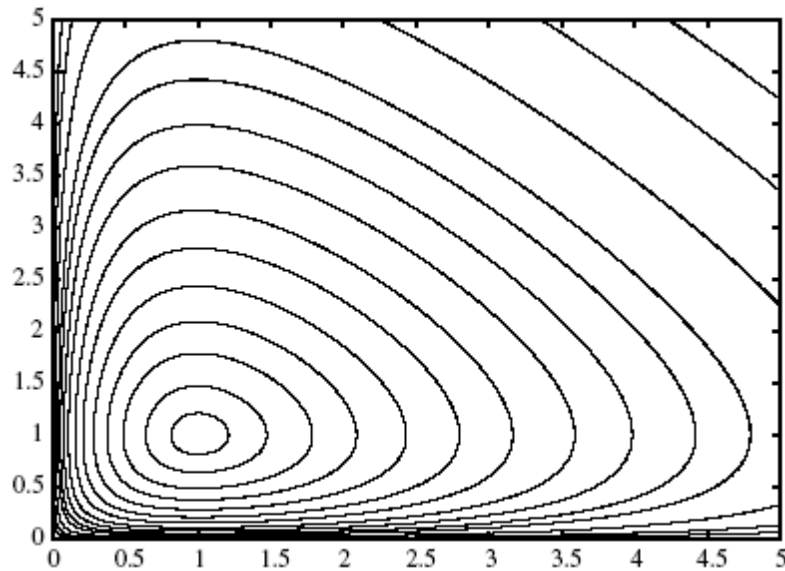


Figure: 64.25 - Espace des phases x-y

Nous voyons ainsi (dans cette représentation de l'espace des phases particulier x-y) que pour des conditions initiales fixes données, le système est périodique (oscillant) déphasé et a un point d'équilibre en:

$$(x, y) = (1, 1) \quad (64.64)$$

appelé "point fixe" dans le domaine de la physique non-linéaire. Ce qui correspond aux points où:

$$\frac{dx}{dt} = \frac{dy}{dt} = 0 \quad (64.65)$$

Finalement, nous avons deux couples de points d'équilibre (c'est trivial en regardant le système d'équation):

$$(x_1^e, y_1^e) = (0, 0) \quad \text{et} \quad (x_2^e, y_2^e) = (1, 1) \quad (64.66)$$

La question qui se pose est le sens de rotation (représentation) du plan des phases. Ainsi, en représentant les directions à l'aide d'un champ de vecteurs, nous obtenons la représentation (plus intéressante pour comprendre les observations de la pêche après la guerre...).

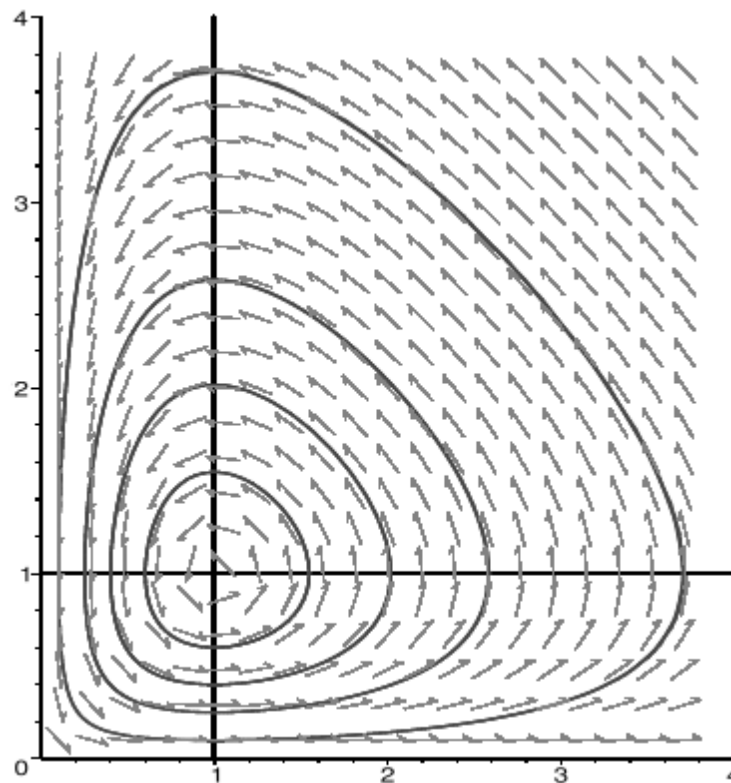


Figure: 64.26 - Mise en évidence de la direction des cycles dans l'espace des phases

Pour savoir dans quelle direction nous nous dirigeons dans l'espace des phases à un moment donné, il suffit donc de connaître la dérivée dy/dx (ou réciproquement dx/dy). Ainsi, nous avons:

$$\frac{dy}{dx} = \frac{dy}{dt} \frac{dt}{dx} = \frac{dy}{dt} \left(\frac{dx}{dt} \right)^{-1} = \frac{my(x-1)}{rx(1-y)} \quad (64.67)$$

Ceci dit, nous voyons bien sur le diagramme des phases sous sa forme de champ de vecteurs qu'il arrive un moment dans le cycle de ce modèle où le nombre de prédateurs est très élevé pour un faible nombre de proies. Donc le modèle mathématique (théorique) explique bien ce qui peut être a priori contre-intuitif pour l'être humain. Il devient dès lors évident de deviner en quels points du cycle se trouvait l'écosystème à la reprise de la pêche après la guerre.

Cependant, nous pouvons (devons) nous poser la question de ce qu'il se passe après une petite perturbation autour des points d'équilibres (ce qui est de la plus haute importance en écologie).

Nous avons donc le système de Lotka-Volterra à l'équilibre:

$$\begin{aligned} \frac{dx}{dt} &= rx(1-y) \\ \frac{dy}{dt} &= my(x-1) \end{aligned} \quad (64.68)$$

En y mettant une perturbation infiniment petite, celui-ci s'écrit:

$$\begin{aligned}\frac{dx}{dt} &= r(x+x^*)(1-(y+y^*)) \\ \frac{dy}{dt} &= m(y+y^*)((x+x^*)-1)\end{aligned}\quad (64.69)$$

En nous concentrant d'abord sur l'étude proche du point d'extinction (0,0) nous pouvons négliger les termes xy , nous obtenons:

$$\begin{aligned}\frac{dx}{dt} &\cong r\left[(1-y^*)x + (1-y-y^*)x^*\right] \\ \frac{dy}{dt} &\cong -m\left[(1-x^*)y + (1-x-x^*)y^*\right]\end{aligned}\quad (64.70)$$

Mais c'est encore trop compliqué. Toujours en nous considérant proche du point d'extinction (0,0), nous considérons un peu abusivement mais astucieusement (sinon nous ne pourrions toujours pas résoudre le problème analytiquement) les approximations suivantes:

$$\begin{aligned}\frac{dx}{dt} &\cong r\left[x - \underbrace{y^*x}_{=0} + x^* - \underbrace{yx^*}_{=0} - \underbrace{y^*x^*}_{=0}\right] \cong r(x+x^*) \underset{x \gg x^*}{\cong} rx \\ \frac{dy}{dt} &\cong -m\left[y - \underbrace{x^*y}_{=0} + y^* - \underbrace{xy^*}_{=0} - \underbrace{x^*y^*}_{=0}\right] \cong -m(y+y^*) \underset{y \gg y^*}{\cong} -my\end{aligned}\quad (64.71)$$

Donc:

$$\begin{aligned}\frac{dx}{dt} &\cong rx \\ \frac{dy}{dt} &\cong -my\end{aligned}\quad (64.72)$$

Ce qui nous montre pour le système à l'équilibre, proche du point d'extinction, que le nombre d'individus des prédateurs diminue exponentiellement alors que les proies augmentent elles exponentiellement. Ceci a un sens biologique: quand il y a peu d'individus prédateurs, les proies se multiplient alors qu'au fur et à mesure que le nombre de proies augmente, les prédateurs se multiplient et se concentrent de plus en plus sur leurs proies; respectivement quand il y a peu de proies, le nombre de prédateurs diminue et à mesure le nombre de proies augmente (ahhh la nature...).

Remarque: Dans la littérature nous retrouvons tantôt le signe "-" en haut ou en bas dans le système d'équations précédent. En réalité, la position du signe "-" n'a aucune importance car il s'agit juste du choix de départ dans la dynamique du système.

Proche du point d'équilibre (1,1), nous avons donc (hop! nous changeons la position du signe exprès afin de d'insister sur le fait que ce n'est qu'un choix de départ!!!):

$$\begin{aligned}\frac{dx}{dt} &\cong -ry \\ \frac{dy}{dt} &\cong mx\end{aligned}\quad (64.73)$$

Pour résoudre ce système, différencions l'équation des proies par rapport au temps telle que:

$$\frac{d^2x}{dt^2} = -r \frac{dy}{dt} \quad (64.74)$$

et en y injectant dy/dt (prédateurs):

$$\frac{d^2x}{dt^2} = -rmx \quad (64.75)$$

Nous obtenons donc une petite équation différentielle du deuxième ordre (cf. chapitre de Calcul Différentiel Et Intégral) dont la solution type est:

$$x = x_0 e^{-\lambda t} \quad (64.76)$$

En injectant cette solution dans l'équation différentielle, nous obtenons après simplification des exponentielles une simple équation du deuxième degré (cf. chapitre de Calcul Algébrique):

$$\lambda^2 x_0 + rm x_0 = 0 \quad (64.77)$$

Dont la solution est triviale:

$$\lambda_{1,2} = \pm i \sqrt{rm} \quad (64.78)$$

Ainsi, la solution générale de l'équation différentielle est la combinaison linéaire des deux solutions telle que:

$$x(t) = x_0 \left(e^{-i\sqrt{rm}t} + e^{i\sqrt{rm}t} \right) \quad (64.79)$$

Mais nous avons donc:

$$\frac{dx}{dt} = -ry \Rightarrow y(t) = -\frac{1}{r} \frac{dx}{dt} \quad (64.80)$$

Dès lors, connaissant $x(t)$ nous obtenons facilement:

$$y(t) = -x_0 \sqrt{\frac{m}{r}} \left(-ie^{-i\sqrt{rm}t} + ie^{i\sqrt{rm}t} \right) \quad (64.81)$$

Utilisons maintenant la relation d'Euler (voir le traitement des nombres complexes dans le chapitre des Nombres dans la section Arithmétique):

$$e^{ix} = \cos(x) + i \sin(x) \quad (64.82)$$

Ainsi, pour les proies, nous avons:

$$x(t) = x_0 \left[\cos(-\sqrt{rmt}) + i \sin(-\sqrt{rmt}) + \cos(\sqrt{rmt}) + i \sin(\sqrt{rmt}) \right] \quad (64.83)$$

et comme (cf. chapitre de Trigonométrie) :

$$\cos(x) = \cos(-x), -\sin(x) = \sin(-x) \quad (64.84)$$

nous avons alors:

$$x(t) = 2x_0 \cos(\sqrt{rmt}) \quad (64.85)$$

et de manière similaire, nous obtenons pour les prédateurs:

$$\begin{aligned} y(t) &= -x_0 \sqrt{\frac{m}{r}} i \left(- \left(\cos(-\sqrt{rmt}) + i \sin(-\sqrt{rmt}) \right) + \left(\cos(\sqrt{rmt}) + i \sin(\sqrt{rmt}) \right) \right) \\ &= -x_0 \sqrt{\frac{m}{r}} i \left(- \left(\cos(\sqrt{rmt}) - i \sin(\sqrt{rmt}) \right) + \cos(\sqrt{rmt}) + i \sin(\sqrt{rmt}) \right) \\ &= -x_0 \sqrt{\frac{m}{r}} i \left(-\cos(\sqrt{rmt}) + i \sin(\sqrt{rmt}) + \cos(\sqrt{rmt}) + i \sin(\sqrt{rmt}) \right) \\ &= -x_0 \sqrt{\frac{m}{r}} i \left(2i \sin(\sqrt{rmt}) \right) = 2x_0 \sqrt{\frac{m}{r}} \sin(\sqrt{rmt}) \end{aligned} \quad (64.86)$$

soit:

$$y(t) = 2x_0 \sqrt{\frac{m}{r}} \sin(\sqrt{rmt}) \quad (64.87)$$

Ainsi, autour du point d'équilibre (1,1), les perturbations suffisamment petites pour valider la linéarisation (annulation des termes quadratiques) oscillent comme des ellipses (ou cercles) dont les axes sont définis par les deux équations précédentes.

Nous pouvons obtenir les graphes ci-dessus avec Maple 4.00b (car il s'agit d'un joli exemple d'application du logiciel):

```
>restart: with(plots): with(DEtools):
>rate_eqn1:= diff(h(t),t)=(0.1)*h-(0.005)*h*(1/60)*u; rate_eqn2:=diff(u(t),t)=(0.00004)*h*u-(0.04)*u;
vars:= [h(t), u(t)];
> init1:=[h(0)=2000,u(0)=600]; init2:=[h(0)=2000,u(0)=1200]; init3:=[h(0)=2000, u(0)=3000];domain
:= 0 .. 320;
> L:= DEplot({rate_eqn1, rate_eqn2}, vars, domain,{init1 }, stepsize=0.5, scene=[t, u],
arrows=NONE):
H:= DEplot({rate_eqn1, rate_eqn2}, vars, domain,{init1 }, stepsize=0.5, scene=[t, h], arrows=NONE):
> display( {L,H} , title = `Lièvres et 60 * Lynx vs. temps` );
```

Lièvre and 60 * Lynx vs. temps

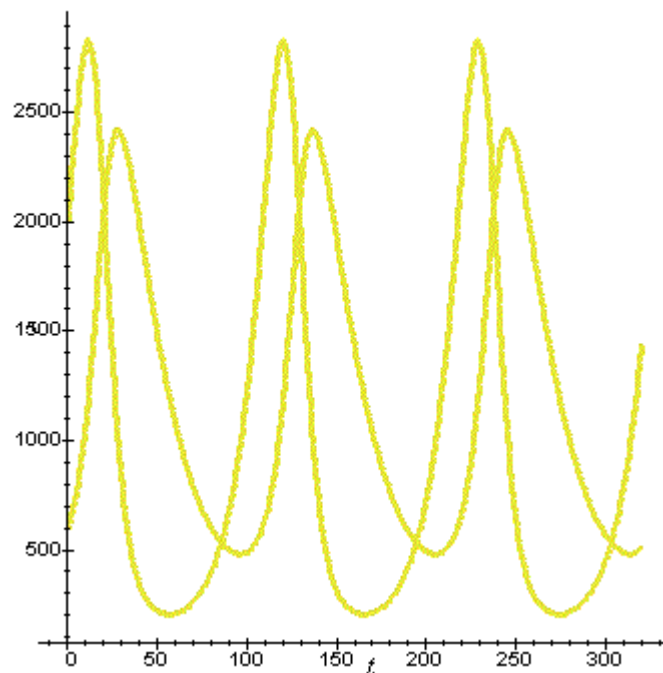


Figure: 64.27 - Plot sous Maple 4.00b du modèle théorique

```
>DEplot({rate_eqn1, rate_eqn2}, vars, t= 0 .. 160, {init1, init2, init3}, stepsize=0.5, scene=[h,u],
title=`Lièvres vs. 60 * Lynx pour t = 0 .. 160`, arrows=slim);
```

Lièvres vs. 60 * Lynx pour t = 0 .. 160

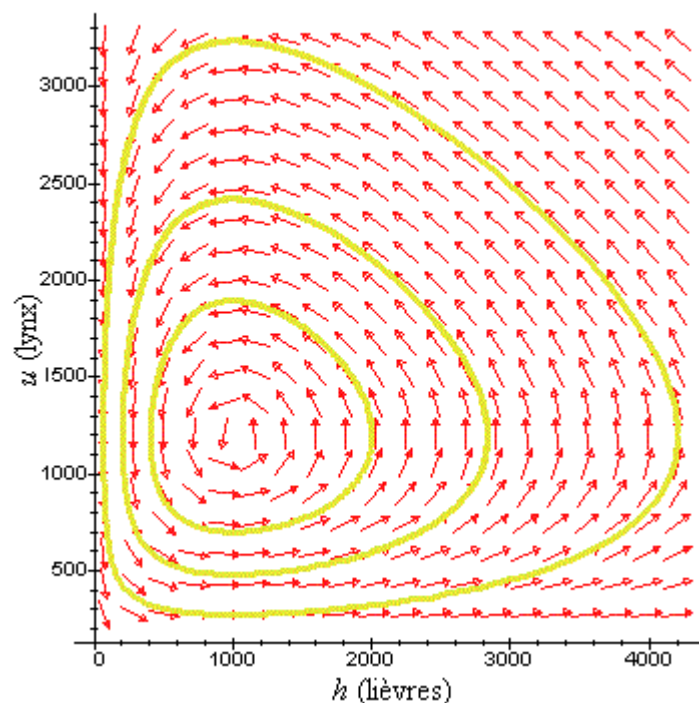


Figure: 64.28 - Plot sous Maple 4.00b de la direction des cycles dans l'espace des phases x-y

5. MODÈLE DE CAPTURE OPTIMALE DE SCHEAFER

Un principe utile pour la gestion des quotas de pêche, de la chasse, de la cueillette sauvage ou de l'exploitation de matières premières est celui du "rendement équilibré maximal" qui est directement lié à une bonne gestion de la dynamique de la population concernée par l'exploitation.

Définition: Le "**rendement équilibré maximal**", c'est la quantité maximale de poissons, cerfs, oiseaux ou champignons, etc. qui peuvent être tués/récoltés dans une population sans mettre en péril sa survie.

Déterminer ce rendement équilibré maximal permet d'atteindre l'objectif de pêcher, chasser, cueillir le plus possible, sans toutefois tomber dans la surconsommation qui entraînerait potentiellement la disparition de la population.

La distinction entre une consommation de ressources naturelles équilibrée et exagérée est souvent difficile à faire. Des modèles mathématiques et statistiques sont alors comme toujours utiles et nécessaires pour établir cette frontière.

Ces calculs interviennent dans les calculs des quotas même encore en ce début du 21ème siècle. Le modèle simpliste que nous allons présenter est utilisé par certains grands États (mais il va avec le temps disparaître pour faire place à un modèle plus élaboré avec des paramètres probabilistes).

Nous nous intéressons donc à la biomasse d'une population d'un bien consommable. Le modèle mis en avant par Schaefer en 1954 fait intervenir deux paramètres seulement:

- La biomasse maximale M qui peut vivre en permanence dans un système
- La taux de reproduction r

Nous partons sur l'hypothèse (à vérifier dans la pratique à cause du Modèle de Lotka-Volterra) qu'à l'année zéro (avant le début de la consommation de la ressource), la biomasse correspondante est:

$$X_0 = M \quad (64.88)$$

Le modèle empirique s'écrit:

$$X_{n+1} = X_n + rX_n \left(1 - \frac{X_n}{M}\right) - C \quad (64.89)$$

Pour bien comprendre ce modèle, supposons d'abord que la population n'est pas exploitée, c'est-à-dire que $C = 0$. À l'année n , la population s'accroît alors de:

$$r \left(1 - \frac{X_n}{M}\right) \quad (64.90)$$

Ainsi, en l'absence de pêche, si X_0 est plus petit que M , la population va croître tranquillement jusqu'à atteindre sa capacité limite M .

Dans le cas extrême, où la consommation C est égale à M , nous avons:

$$X_{0+1} = X_0 + rX_0 \left(1 - \frac{X_0}{M}\right) - C = M + rM \left(1 - \frac{M}{M}\right) - C = M + 0 - M = 0 \quad (64.91)$$

ce qui est logique... (il y a extinction totale).

Maintenant, réécrivons le modèle de Schaefer sous la forme suivante:

$$X_{n+1} - X_n = \Delta X = rX_n \left(1 - \frac{X_n}{M}\right) - C \quad (64.92)$$

Nous souhaiterions savoir pour quelle valeur de X_n , nous aurions la biomasse qui ne varie plus. C'est-à-dire:

$$\Delta X = rX_n \left(1 - \frac{X_n}{M}\right) - C = 0 \quad (64.93)$$

Dès lors:

$$rX_n - r\frac{X_n^2}{M} - C = \frac{r}{M}X_n^2 - rX_n + C = 0 \quad (64.94)$$

Il s'agit donc d'une simple équation du deuxième degré! Elle n'a de solution réelle que si le discriminant (cf. chapitre Calcul Algébrique) est positif ou nul. Soit, sachant que r est positif ou nul par construction:

$$r^2 - 4\frac{r}{M}C \geq 0 \Leftrightarrow r\left(r - \frac{4C}{M}\right) \geq 0 \Leftrightarrow r - \frac{4C}{M} \geq 0 \Leftrightarrow C \leq \frac{rM}{4} \quad (64.95)$$

Ainsi, la plus grande valeur possible de C est $rM/4$. Lorsque nous remplaçons C par cette valeur dans l'équation initiale:

$$\frac{r}{M}X_n^2 - rX_n + C = 0 \quad (64.96)$$

nous obtenons:

$$\frac{r}{M}X_n^2 - rX_n + \frac{rM}{4} = 0 \quad (64.97)$$

Soit:

$$\begin{aligned} \frac{r}{M} \left(X_n^2 - MX_n + \frac{M^2}{4} \right) &= 0 \\ X_n^2 - MX_n + \frac{M^2}{4} &= 0 \quad (64.98) \\ \left(X_n - \frac{M}{2} \right)^2 &= 0 \end{aligned}$$

Donc, la solution est:

$$X_n = \frac{M}{2} \quad (64.99)$$

Ainsi, il existe un temps n (infini) pour lequel la population tend asymptotiquement vers $M/2$ à la condition que $X_0 \geq \frac{M}{2}$.

Cela dit, dans la pratique, on choisira une fraction de la valeur suivante afin de se prémunir des variations naturelles négatives de la reproduction:

$$C \leq \frac{rM}{4} \quad (64.100)$$

Malgré tout, sa grande simplicité et son côté déterministe font que le modèle de Schaefer est encore utilisé en ce début de 21ème siècle dans la gestion des pêches.

6. MODÈLE DE HARDY-WEINBERG

Le modèle de Hardy-Weinberg (développé en 1908) décrit la diversité (plutôt que l'évolution) du vivant dans le sens de la génétique des populations. Cependant, ce modèle permet quand même de comprendre l'impact d'un agent de mutation sur une population d'individus et ainsi de résoudre les problèmes liés au refus de modèle de la sélection (évolution) naturelle de Darwin. Effectivement, les travaux de Darwin portaient sur des caractères qui variaient de façon continue. Or, il ne pouvait expliquer comment les individus se transmettent ces variations (c'est effectivement gênant). La génétique de Mendel, quant à elle imposait que seuls les caractères discontinus étaient héréditaires.

Il y avait donc conflit entre le modèle continu (de Darwin) et discontinu (de Mendel) et ce fut la génétique qui au vingtième siècle permis d'établir qu'en fait c'est le mélange des gènes qui crée la diversité et in extenso la sélection naturelle.

Remarque: Nous démontrerons en méthodes numériques (section d'informatique théorique) dans le cadre de l'étude des algorithmes génétiques que le chromosome ayant la valeur de fonction de "fitness" la plus grande sera celui qui aura le plus de chance de se reproduire. Ceci fait partie du théorème nommé "The Schema Theorem"...

Conclusion: L'évolution ne se fait pas à l'échelle de l'individu mais de la population (c'est une bonne leçon à retenir ça...)

Définitions:

D1. Un gène ou "**génotype**" est composé "**d'allèles**" (deux habituellement, donc nous considérerons les individus comme "diploïdes")

D2. Un "**pool génique**" est un ensemble de gènes présents dans une population

D3. Une "**micro-évolution**" est un changement de la fréquence des allèles d'un pool génétique d'une population.

D4. Un allèle est dit "**allèle fixe**" lorsque tous les membres d'une population portent deux allèles identiques.

D5. Une "**spéciation**" est l'apparition à long terme d'une micro-évolution faisant apparaître une nouvelle espèce.

D6. Une "**espèce**" est un ensemble de populations dont tous les membres ont le potentiel de se reproduire entre eux dans un environnement naturel (nous supposons dans ce modèle que les individus se reproduisent selon un mode sexué).

D7. La "**mutation**" est un agent évolutif qui modifie le pool génique d'une population en produisant de nouveaux gènes (nous considérons donc qu'il n'y a pas d'agents évolutifs à l'oeuvre dans la population dans ce modèle).

Remarque: La mutation est un événement rare et le plus souvent néfaste. Ses effets quantitatifs sont plus importants chez les organismes à temps de génération court (bactéries et virus), l'effet étant nettement moins marqué chez les organismes à temps de génération plus long (animaux et végétaux).

Considérons maintenant une population composée de 13 individus (1 gène, chacun composé de deux allèles pris parmi 3):

$$P = \left\{ \begin{array}{l} X1X1, X2X1, X3X2, X1X1, X1X1, X1X1, X2X3 \\ X1X3, X3X2, X1X1, X1X3, X3X3, X3X3 \end{array} \right\} \quad (64.101)$$

La fréquence de l'allèle X1 est dans ce cas particulier:

$$p_{X1} = 13/26 = 0.5 \quad (64.102)$$

La fréquence de l'allèle X2 est:

$$p_{X2} = 4/26 = 0.154 \quad (64.103)$$

Enfin, la fréquence de l'allèle X3 est:

$$p_{X3} = 9/26 = 0.346 \quad (64.104)$$

Et bien évidemment:

$$p_{X1} + p_{X2} + p_{X3} = 1 \quad (64.105)$$

Pour résumer, si nous considérons maintenant une population d'individus possédant des gènes composés de seulement deux types d'allèles, nous avons dès lors (cas plus concret par rapport à la biologie du vivant sur notre Terre):

$$BB, Bb, bb \quad (64.106)$$

Nous avons donc pour la fréquence relative du nombre d'allèles B:

$$p = \frac{2 \cdot (\text{nombre d'individus } BB) + 1 \cdot (\text{nombre d'individus } Bb)}{\text{nombre d'individus total}} \quad (64.107)$$

et pour la fréquence relative d'allèles b :

$$q = \frac{2 \cdot (\text{nombre d'individus } bb) + 1 \cdot (\text{nombre d'individus } Bb)}{\text{nombre d'individus total}} \quad (64.108)$$

Avec donc:

$$p + q = 1 \quad (64.109)$$

Nous pouvons faire quelques observations suite à la manière de construire ce modèle. Effectivement, la fréquence des allèles (équilibre génétique) dans une population demeure approximativement constante génération après génération si:

1. La population est de très grande taille (une perturbation infime ne change que très peu la fréquence des types de génotypes)
2. Pas d'émigration ou d'immigration (il n'y a pas de nouveaux types d'allèles dans la population)
3. Pas de mutations modifiant les allèles (cela contient le point 2)
4. Accouplements au hasard (non influencés par le type d'allèle étudié)
5. Pas de sélection naturelle (cela contient le point 4)

Il peut être pertinent (pour la culture générale) de mettre en évidence le fait que nous distinguons deux types d'accouplements non aléatoires:

1. L'endogamie: croisement entre individus du voisinage, donc ayant des liens de parenté ; chez les espèces se dispersant peu, particulièrement dans le règne végétal.
2. L'homogamie: choix des partenaires qui se ressemblent pour certains caractères ; répandu chez les animaux.

L'endogamie et l'homogamie ne sont pas une cause directe de microévolution puisqu'elles ne modifient pas le pool génique d'une population. Cependant, si certains des individus issus de ces croisements non aléatoires ont plus de chance de s'accoupler que d'autres (par sélection naturelle), il s'ensuivra une modification de la fréquence allélique dans la population descendante et par conséquent, une microévolution. En conséquence, les accouplements non aléatoires sont un facteur potentiel de microévolution.

Finalement, pour en revenir à notre modèle, la probabilité qu'un individu possède l'un des trois génotypes possibles (relativement à un génotype) dans la population (relativement à notre exemple) est:

- Probabilité qu'un individu soit BB :

$$BB = p \cdot p = p^2 \quad (64.110)$$

- Probabilité qu'un individu soit bb :

$$bb = q \cdot q = q^2 \quad (64.111)$$

- Probabilité qu'un individu soit Bb ou bB :

$$(bB, Bb) = pq + qp = 2pq \quad (64.112)$$

Finalement, nous obtenons "l'équation de Hardy-Weinberg":

$$p^2 + 2pq + q^2 = (p + q)^2 = 1 \quad (64.113)$$

Que nous pouvons représenter par le graphique:

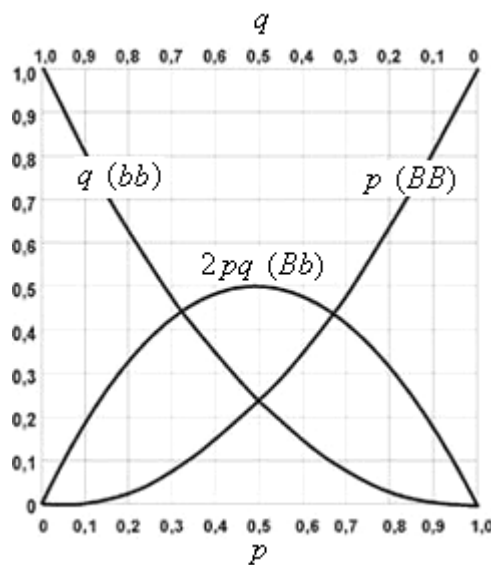


Figure: 64.29 - Tracé de l'équation de Hardy-Weinberg

Après cette étude, nous pouvons enfin (re)définir la "**loi de Darwin**" de manière à ce qu'elle se soustraie à la grande question (problématique) de la modification continue ou discrète:

La sélection naturelle est un agent évolutif qui modifie le pool génique d'une population en augmentant la fréquence des allèles produisant les phénotypes (génotypes définissant le comportement d'un individu) les mieux adaptés au milieu. Ainsi, les individus les mieux adaptés à leur environnement se reproduisent plus que les autres et contribuent davantage au patrimoine génétique de la descendance. La fréquence des bons gènes augmente graduellement dans la population, de génération en génération. Ainsi, la sélection naturelle dirige l'adaptation d'une population à son milieu en accumulant les génotypes qui favorisent la survie dans le milieu.

Remarques:

R1. Le modèle de Hardy-Weinberg n'existe pas dans la nature. Effectivement, il est impossible que toutes les conditions nécessaires soient présentes en même temps. Cependant, c'est un modèle théorique qui permet d'évaluer s'il y a de la microévolution dans une population. Ainsi, on mesure la fréquence des allèles dans la population mère et l'on applique l'équation. Si la structure génétique de la population fille s'écarte de celle prédite par la loi, alors on sait qu'au moins un agent de l'évolution est à l'oeuvre.

R2. Ce modèle théorique permet de prédire les coûts sociaux des maladies génétiques. Effectivement, si on connaît le pourcentage d'individus qui présentent une maladie récessive dans une population, on peut appliquer l'équation et évaluer le nombre de transmetteurs sains puis prévoir la probabilité d'apparition de la maladie dans les générations futures. Si celle-ci est grande (ce raisonnement n'est valable que dans un cadre libéral et capitaliste), on peut décider d'investir dans la recherche ou prévoir des budgets pour soigner les malades.

 Exemples:

E1. Soit un génotype de séquences possibles $\{BB, bb, Bb\}$ et considérons que $p_B = 0.6, p_b = 0.4$. Nous aurons alors:

$$p^2 = 0.36 \quad 2pq = 0.48 \quad q^2 = 0.16 \quad (64.114)$$

Ainsi, dans une population de 1'000'000 d'individus, nous devrions avoir:

$$\begin{aligned} 0.36 \cdot 1 \cdot 10^6 &= 360'000 BB \\ 0.48 \cdot 1 \cdot 10^6 &= 480'000 Bb \\ 0.16 \cdot 1 \cdot 10^6 &= 160'000 bb \end{aligned} \quad (64.115)$$

E2. Si au Québec (6 millions d'habitants), une personne sur cinq a les yeux bleus ($bb \rightarrow q^2 = 1/5$), combien de personnes sont de génotype BB et combien sont Bb ?

$$\begin{aligned} q &= \sqrt{1/5} = 0.44721 \\ p &= (1 - q) = 0.5528 \end{aligned} \quad (64.116)$$

Donc:

$$p^2 = 0.3056 \quad 2pq = 0.4944 \quad q^2 = 0.2 \quad (64.117)$$

Ainsi nous avons:

$$\begin{aligned} 0.3056 \cdot 6 \cdot 10^6 &= 1'833'600 BB \\ 0.4944 \cdot 6 \cdot 10^6 &= 2'966'400 Bb \\ 0.2 \cdot 6 \cdot 10^6 &= 1'200'000 bb = \frac{1}{5} \cdot 6 \cdot 10^6 \end{aligned} \quad (64.118)$$

E3. Nous introduisons dans un étang 1'000 grenouilles tachetées, homozygotes pour ce caractère (TT) et 250 grenouilles sans taches également homozygotes pour ce caractère (tt). Si nous laissons les grenouilles se reproduire pendant quelques années, en supposant que la population demeure stable, quel nombre de grenouilles TT , Tt , tt devrait-on alors observer ?

Nous avons donc une population de 1'250 individus dans laquelle nous avons 1'000 TT (2'000 allèles T) et 250 tt (500 allèles t). Donc:

$$\left. \begin{aligned} p(T) &= \frac{2'000}{2'500} = 0.8 \\ q(t) &= \frac{500}{2'500} = 0.2 \end{aligned} \right\} \Rightarrow \begin{cases} p^2 = 0.64 \\ 2pq = 0.32 \\ q^2 = 0.04 \end{cases} \quad (64.119)$$

Si la population est de 1'250 individus, nous aurons donc:

$$\begin{aligned} 0.64 \cdot 1'250 &= 800 \text{ individus } TT \\ 0.32 \cdot 1'250 &= 400 \text{ individus } Tt \\ 0.04 \cdot 1'250 &= 50 \text{ individus } tt \end{aligned} \quad (64.120)$$

7. TAUX DE CROISSANCE AVEC LA TEMPÉRATURE

Pour de nombreuses espèces (mammifères, poissons, micro-organismes), il est raisonnable de considérer qu'en première approximation, la relation du taux de croissance de la population avec la température de l'environnement est un polynôme du second degré:

$$\mu = aT^2 + bT + c \quad (64.121)$$

où les valeurs a , b , c vont dépendre de l'espèce considérée.

Nous savons en plus qu'il existe pour chaque individu un optimum de croissance à une température donnée T_{opt} , ainsi que des températures minimale T_{min} et maximale T_{max} en deçà et au-delà desquelles il n'y a plus de croissance.

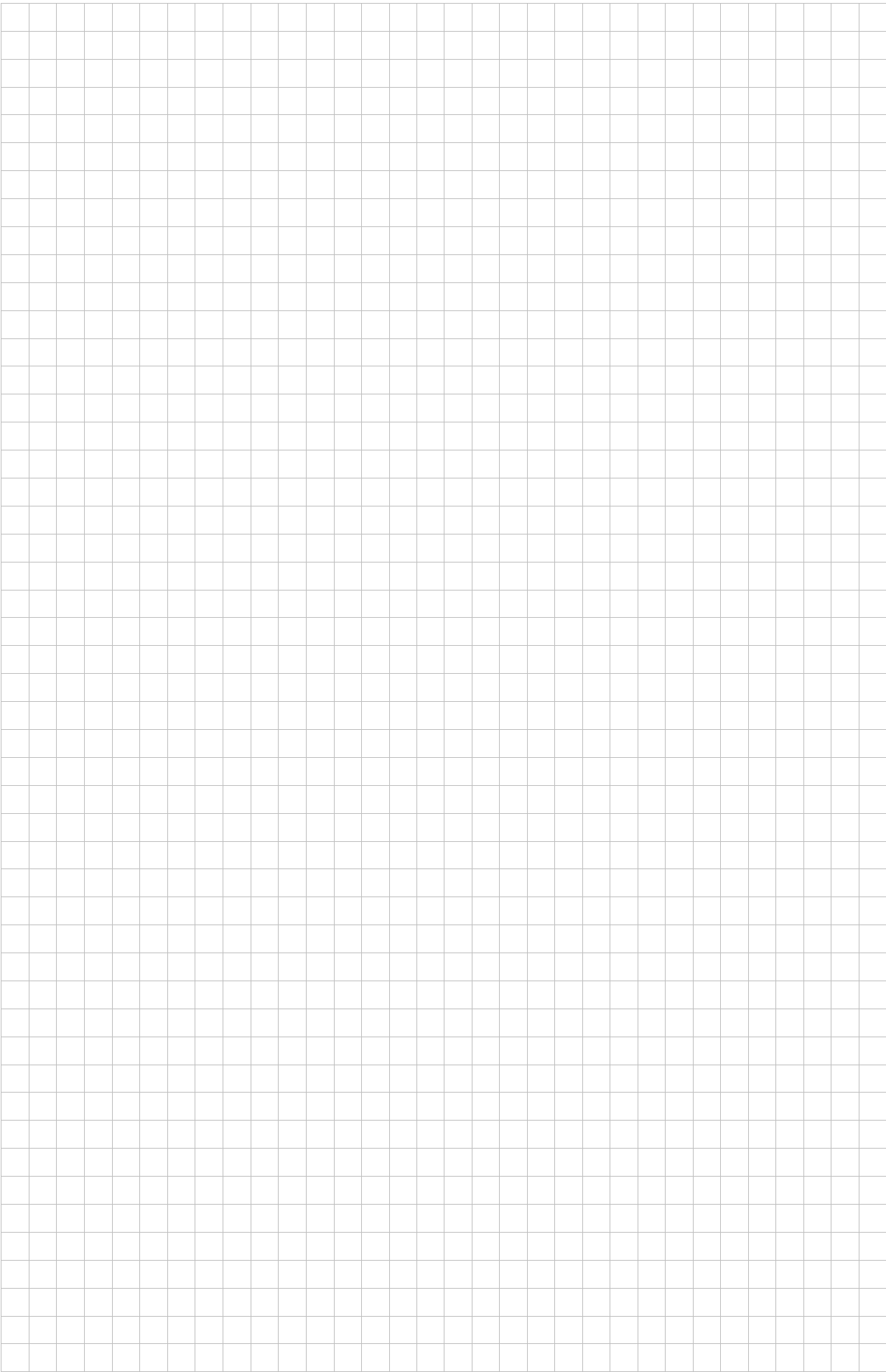
Ainsi, les paramètres a , b , c doivent vérifier les relations suivantes:

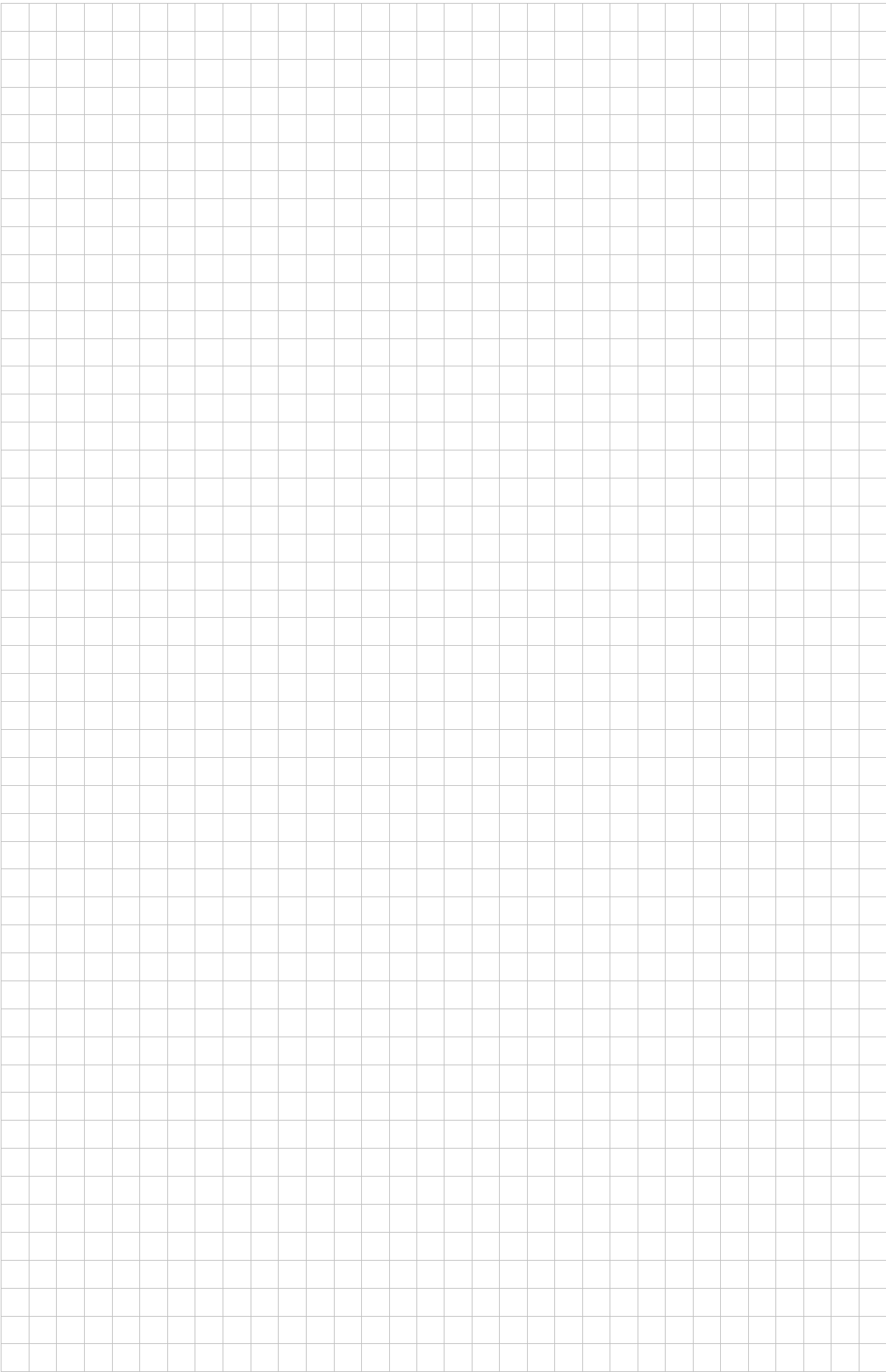
$$\begin{aligned} \frac{d\mu(T_{opt})}{dT} &= 0 \\ \mu(T_{min}) &= \mu(T_{max}) = 0 \end{aligned} \quad (64.122)$$

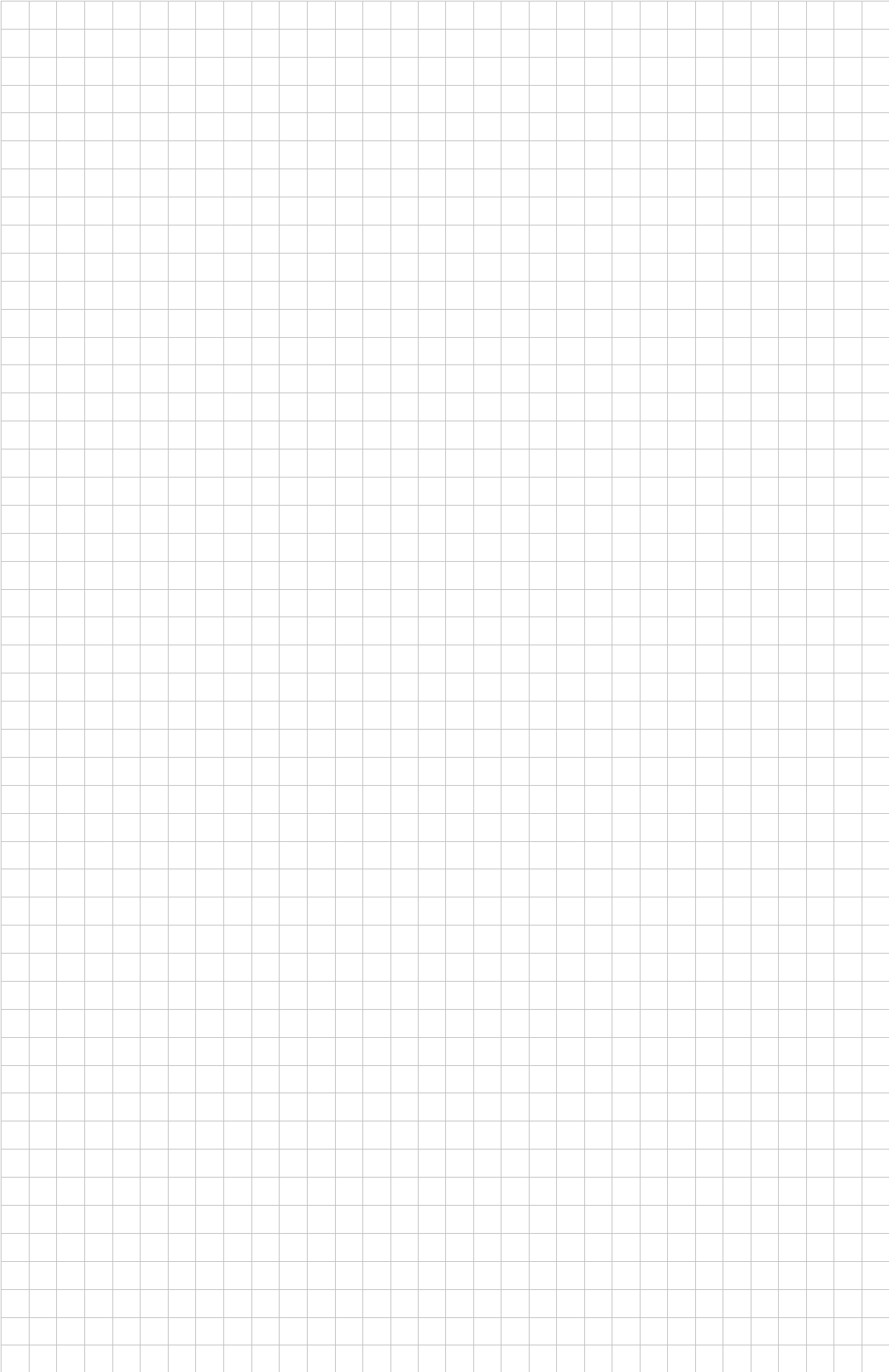
Ce qui implique pour les trois températures cardinales T_{opt} , T_{min} , T_{max} de respecter les relations démontrées dans le cadre de l'étude des polynômes du deuxième degré:

$$\begin{aligned} T_{opt} &= -\frac{b}{2a} \\ T_{min} &= \frac{-b - \sqrt{b^2 - 4ac}}{2a} = \frac{-b}{2a} - \frac{\sqrt{b^2 - 4ac}}{2a} = T_{opt} - \frac{\sqrt{b^2 - 4ac}}{2a} \\ T_{max} &= \frac{-b + \sqrt{b^2 - 4ac}}{2a} = \frac{-b}{2a} + \frac{\sqrt{b^2 - 4ac}}{2a} = T_{opt} + \frac{\sqrt{b^2 - 4ac}}{2a} \end{aligned} \quad (64.123)$$

Donc si nous connaissons pour un objet vivant T_{opt} et T_{min} il est facile d'en déduire théoriquement une approximation de T_{max} .







65. THÉORIE DES JEUX ET DE LA DÉCISION

La théorie de la décision et des jeux dépasse très largement le cadre étroit des jeux de société, même si ces derniers ont constitué son premier objet d'étude et lui ont donné son nom dans la plupart des ouvrages disponibles dans le commerce.

Par ailleurs, les deux théories sont très proches l'une de l'autre d'où le fait qu'elles soient très souvent non différenciées dans la littérature.

Définition:

D1. La "**théorie des jeux**" est l'étude des modèles de prise de décision en avenir incertain non probabilisable.

D2. La "**théorie de la décision**" (appelée aussi parfois "**analyse décisionnelle**") est l'étude des modèles de prise de décision en avenir incertain probabilisable (objectivement ou subjectivement).

Chacune des méthodes d'analyse de ces deux théories se fait principalement sous forme tabulaire (tableau) ou sous la forme d'un arbre vertical ou horizontal.

Voici un schéma assez connu par les coordinateurs de projets qui résume assez bien la situation globale:

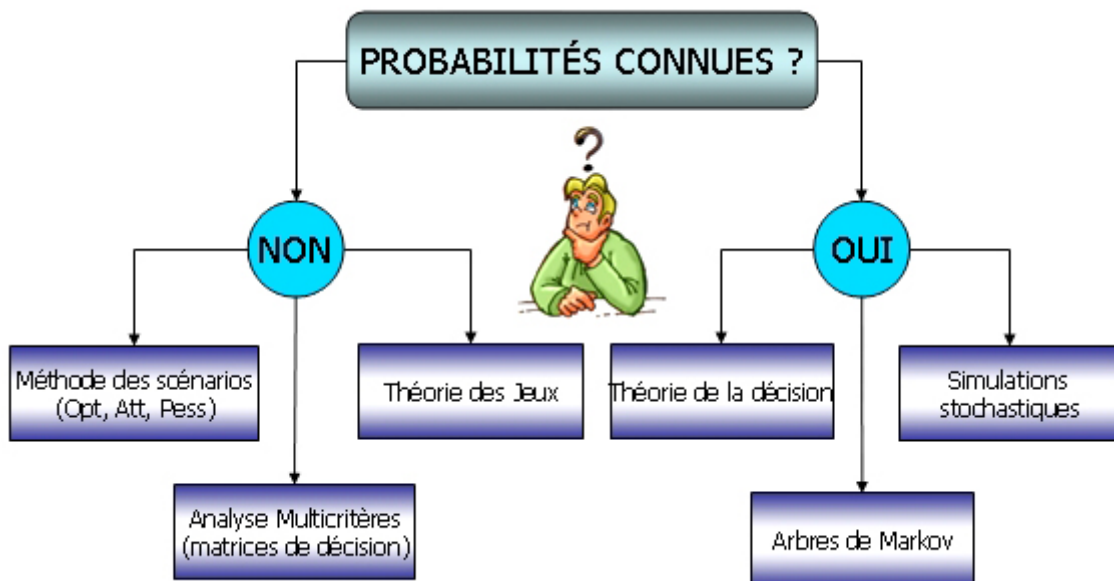


Figure: 65.1 - Classification élémentaire des techniques de décision

Ces outils ont pour objectif de tenter de formaliser comment statuer que telle configuration ou décision est meilleure qu'une autre? Nous chercherons pour cela à trouver l'optimum de certains paramètres qui permettent de quantifier la qualité stratégique d'une situation. Il faut également déterminer quelles conditions conduisent à une configuration qui est jugée optimale.

La théorie des jeux et de la décision est aujourd'hui assez répandue et utilisée dans les milieux universitaires, non seulement en économie (finance d'entreprise particulièrement), mais également par toute une classe d'autres sciences dans lesquelles l'étude des situations de conflits est pertinente:

sociologie, biologie, évolution, informatique (jeux vidéo), marketing...

Remarque: Dans le monde de l'industrie les techniques de décisions sont inconnues de la quasi-totalité des dirigeants dont les choix sont souvent plus qualitatifs, instinctifs que scientifiques...

Nous tenterons, comme toujours sur ce site, de minimiser au mieux le nombre de définitions et concepts afin de ne pas noyer la rigueur de l'analyse mathématique sous le chaos d'un vocabulaire inutile et non nécessaire à une telle analyse (et dans le cadre de la théorie de jeux c'est un peu comme dans la théorie des graphes vraiment le cauchemar!).

Définition: Un "**jeu**" est une situation où des joueurs sont conduits à faire des choix stratégiques parmi un certain nombre d'actions possibles, et ce dans le cadre défini à l'avance par les "**règles du jeu**", le résultat de ces choix constituant une "**issue du jeu**", à laquelle est associé un "**gain**" (ou paiement), positif ou négatif, pour chacun des participants.

Remarque: Un joueur peut être une personne, un groupe de personnes, une société, une région, un parti politique, un pays ou la Nature...

Postulats (nous les retrouverons en économétrie):

- P1. Le marché est régi par la compétition et la coopération...
- P2. Les comportements des agents économiques sont rationnels (...)
- P3. Il est possible de formaliser les comportements compétitifs
- P4. Tous les phénomènes compétitifs ont une dimension utilitaire

Nous différencions et définissons quatre types de situations (que nous formaliserons plus loin):

D1. Les "**jeux coopératifs ou non-coopératifs**": un jeu est dit coopératif lorsque les joueurs peuvent communiquer librement entre eux et passer des accords (par ex. sous forme d'un contrat). Ils forment alors une coalition et recherchent l'intérêt général suivi d'un partage des gains entre tous les joueurs. Dans un jeu non-coopératif, les joueurs (qui ne communiquent pas ou ne peuvent pas communiquer entre eux) agissent selon le principe de rationalité économique: chacun cherche à prendre les meilleures décisions pour lui-même (c'est à dire cherche à maximiser égoïstement ses gains individuels). Ce dernier type de jeu fait intervenir les probabilités.

D2. Les "**jeux à somme nulle ou non nulle**": un jeu est dit à "somme nulle" lorsque la somme des gains des joueurs est constante (ou par le choix subtil d'une fonction d'utilité peut l'être...) ou autrement dit: ce que l'un gagne est nécessairement perdu par un autre (échecs, poker...). Les jeux de société sont souvent des jeux à somme nulle mais les situations réelles sont souvent mieux décrites par les jeux non-coopératifs à somme non nulle car certaines issues sont profitables pour tous, ou dommageables pour tous (vie politique, situations d'affaires...).

Remarques:

R1. Certains théoriciens critiquent les jeux à somme nulle, au moins dans le giron des situations économiques, aux motifs qu'un échange économique est en principe mutuellement avantageux et que les jeux à somme nulle seraient totalement irréalistes.

R2. Les jeux à somme nulle sont parfois appelés "**jeux antagonistes**".

R3. Depuis l'invention de l'arme atomique, l'équilibre de la terreur repose sur la doctrine de dissuasion offensive. Contrés par les capacités réciproques à s'infliger des dégâts colossaux, les arsenaux nucléaires respectifs s'auto-annulent, dans un jeu à somme nulle, par un principe de destruction mutuelle assurée.

D3. Les "**jeux avec ou sans équilibre**": un jeu à somme non nulle coopératif ou non est dit avec "**équilibre de Nash**" s'il existe un couple de stratégies (dans le cas d'un jeu à deux joueurs) tel qu'aucun des joueurs n'a intérêt à changer unilatéralement de stratégie et ceci afin de s'assurer le maximum des minimum (le "maximin") des gains.

D4. Les "**jeux compétitifs ou non-compétitifs**": un jeu non-compétitif est à l'opposé d'un jeu compétitif tel que par définition, lorsque tout couple de stratégies (dans le cas d'un jeu à deux joueurs) est tel qu'il fait perdre ou gagner simultanément à tous les joueurs un gain donné (quand je perds quelque chose, tu perds quelque chose, quand je gagne quelque chose tu gagnes aussi quelque chose).

1. REPRÉSENTATIONS DE JEUX

Il existe différentes manières de formaliser la théorie des jeux et de la décision et ce d'autant plus suivant le type de situations dont il s'agit. Ainsi, nous distinguons:

1. Les "**formes extensives**" qui sont des formes synoptiques (arbre, branche, feuille) utiles à une compréhension simple des stratégies possibles et où l'issue d'un jeu est assimilée à une feuille dans laquelle nous retrouvons le vecteur des gains (ou "payements") respectifs des joueurs. Ce genre de représentation devient compliqué (longue à dessiner) lors de jeux répétitifs.

Lorsqu'une forme extensive fait appel aux probabilités, nous faisons alors référence à un "**arbre de décision**", car comme nous l'avons mentionné au début, qui dit intervention des probabilités dit théorie à part entière: la théorie de la décision.

2. Les "**formes normales**" qui permettent de réduire considérablement la taille et le temps de représentation graphique d'un jeu sous forme d'un tableau (matrice) de gains (ou "payements") mais qui sont inadaptées aux jeux répétitifs.

Deux sous-catégories principales peuvent en plus être distinguées (il en existe donc d'autres!):

- Les "**formes normales des jeux à somme nulle**" (jeux strictement compétitifs) où selon un choix adapté, il est possible de simplifier la représentation de la matrice (ou "**bimatrice**") en demi-matrice puisque les gains sont égaux et opposés pour les joueurs pour chaque stratégie donnée.

- Les "**formes normales des jeux à somme non nulle**" (jeux compétitifs).

Remarque: Chaque cellule du tableau/matrice contient donc un "vecteur" dont les composantes sont les gains respectifs des joueurs. Si le jeu est à somme nulle chaque cellule ne contient qu'une seule valeur puisque ce qui est gagné par un joueur est perdu par l'autre. Nous en verrons de nombreux exemples.

3. Les "**formes ensemblistes**" qui ont une approche ensembliste orientée probabiliste qui va nous permettre d'étudier la dernière forme ci-dessous.

4. Les "**formes graphiques**" qui sont sympathiques à regarder et que nous introduirons comme approche complémentaire car faisant appel à la recherche opérationnelle (cf. chapitre de Méthodes Numériques).

1.1. FORME EXTENSIVE D'UN JEU

Les règles d'un jeu stratégique et les gains contingents qui y sont associés peuvent donc être représentés sous une forme extensive plus couramment nommée par les spécialistes "**arbre de Kuhn**".

Exemple:

Nous considérons deux firmes d'ordinateurs qui ont à faire un choix de système d'exploitation. La compatibilité entre les systèmes serait socialement préférable, mais pour des raisons liées à l'histoire des deux firmes, chacune préférerait que ce soit l'autre qui fasse l'effort de s'adapter. Si les deux firmes choisissent CAM, MBI (J_1) gagne 600 M\$ et Poire (J_2) 200 M\$. Si elles choisissent MAC, c'est Poire qui gagne 600 M\$ et MBI 200 M\$. S'ils ne sont pas compatibles, ils gagnent chacun 100 M\$.

Remarque: Nous appelons ce type de jeu, un "**jeu de coordination**". Par exemple, le choix de standards de télévision ou de lecteur des Mac et PC correspondent à ce type de jeux. Chaque constructeur voudrait imposer son propre standard mais en cas de désaccord, les consommateurs pourraient refuser d'acheter le produit.

Les firmes jouent séquentiellement tel que le jeu puisse être représenté sous la forme d'un arbre de décision:

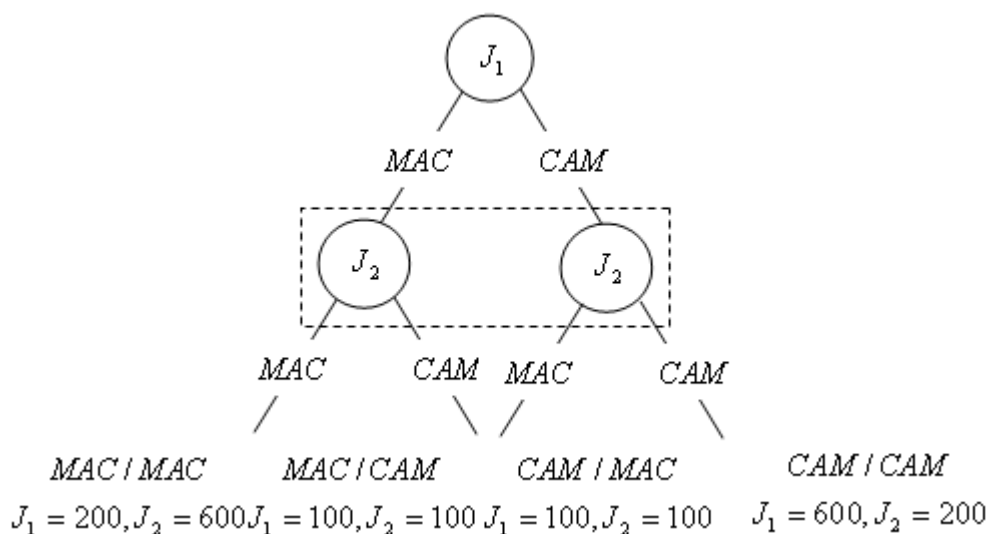


Figure: 65.2 - Jeu séquentiel sous forme d'un arbre de décision (ou arbre de Kuhn)

Remarques:

R1. La structure informationnelle mise en évidence fait référence à l'information dont dispose chaque joueur à chaque noeud de décision du jeu.

R2. MAC/CAM est un jeu à "**information parfaite**" dans le sens que les joueurs connaissent exactement l'éventail de leurs stratégies et de celles de leur adversaire et les conséquences précises de ces stratégies. Ainsi, chaque noeud de la forme extensive est visible par les joueurs (nous définirons le concept d'information parfaite de manière formelle un peu plus loin).

Une analyse plus simple de la meilleure stratégie à adopter dans le cadre d'un jeu consiste à passer directement à la forme normale comme nous le verrons un peu plus loin (mais cette forme normale n'est pas adaptée pour une forme extensive d'une décision).

1.2. FORME EXTENSIVE D'UNE DÉCISION

Comme nous l'avons mentionné au début de ce chapitre, les théories des jeux et de la décision se différencient par le fait que les données de départ de la première se trouvent dans un univers totalement déterministe alors que pour la seconde, elles sont totalement probabilistes. Ce dernier contexte est tellement important dans l'industrie qu'il existe comme nous allons le voir de suite des logiciels (@Risk, Isograph, Treeage) spécialisés sur le marché pour la gestion des formes extensives (ces dernières faisant partie intégrante des méthodes de gestion de risque de la norme ISO 31010).

Exemples:

E1. Le cas de décision le plus simple dans l'industrie est l'analyse par arbres d'événements qui fait appel (lorsque les probabilités sont fixes) uniquement aux axiomes élémentaires des probabilités (cf. [chapitre de Probabilités](#)). Si les probabilités ne sont pas fixes (ce qui est fréquent dans la réalité), il faudra faire appel à des logiciels intégrant les méthodes de Monte-Carlo (cf. [chapitre de Méthodes Numériques](#)).

L'analyse par arbre d'événements est une technique graphique permettant de représenter les séquences d'événements mutuellement exclusifs suivant un événement initiateur dépendant du fonctionnement/non fonctionnement (ou enclenchement/non enclenchement) de divers systèmes conçus pour limiter ses conséquences.

Voici un exemple simple d'arbre d'événement effectué et automatisé par mes soins au niveau des calculs avec le logiciel bureautique Microsoft Office Visio:

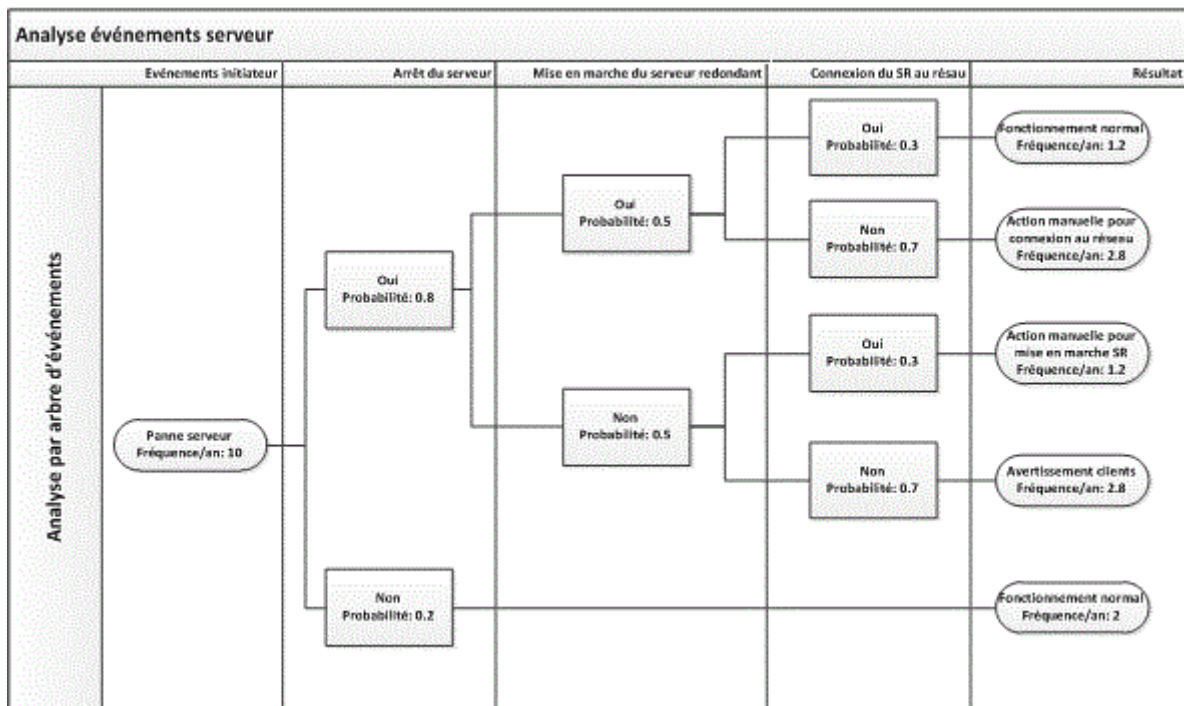


Figure: 65.3 - Arbre d'événements avec probabilités fixes sous MS Office Visio

La fréquence annuelle dans la colonne se trouvant à l'extrême droite du tableau (les feuilles) est simplement égale à l'estimation de la fréquence annuelle de l'événement initiateur multipliée par le produit des probabilités d'une branche telle que:

$$f_r = f_0 \prod_i P_i \quad (65.1)$$

et il faut bien évidemment faire attention à ce que la somme des probabilités dans chaque branche soit égale à 100%.

E2. Cet exemple va être plus compliqué et nettement plus long que l'analyse par arbre d'événements vue avant. Imaginons une société informatique *B* en concurrence potentielle, pour une migration informatique internationale chez *X* avec une autre société *A* (cette dernière pouvant être vue comme un ensemble de concurrents aussi!).

En simplifiant quelque peu, mais sans être toutefois hors de la réalité, considérons que deux choix sont ouverts à *B*: viser "cher" ou viser "bas".

Supposons que nous sachions également que dans le passé *B* a soumis une proposition pour chaque appel d'offres de ce type, alors que le groupe *A* ne l'a fait que dans 60% des cas (pas de fonction de distribution de probabilité dans notre scénario!).

Nous savons également que:

- Si *B* soumet cher et est le seul à soumettre une proposition, son bénéfice attendu est de 22 millions.
- Si *B* soumet un prix élevé mais se trouve en concurrence avec le groupe *A*, il obtiendra le contrat selon le niveau de prix demandé par le groupe *A*. Dans ce cas, il sait qu'il obtiendra en moyenne 1 million.

- Si enfin B soumet à un prix bas, il est sûr d'obtenir le contrat et de réaliser un bénéfice de 10 millions.

Donc dans le cadre où le choix du projet est déterminé uniquement par son prix (au détriment de la qualité comme souvent dans la réalité...) les questions qui se posent sont alors les suivantes:

Q1. Que doit faire B , si aucune information complémentaire ne peut être obtenue?

Ceci constitue une situation du type: "**décision sans informations**"

Q2. À supposer qu'un espion au sein du groupe A puisse informer B si le groupe A soumettra une offre ou non, combien vaudrait cette information pour B ?

Ceci constitue une situation du type: "**décision avec information parfaite**"

Q3. Une société de conseil spécialisée peut donner son avis, mais son expertise, chère, s'élève à 1 million par étude. Pour envisager le recours à ses services, nous savons que, dans le passé, sur les 30 fois où le groupe A avait en fait soumis une proposition, la société de conseil l'avait prévu 24 fois. Et, sur les 20 fois où il n'en avait pas soumis, elle l'avait prévu 17 fois. Faut-il lui commander une étude (bon de toute façon dans la réalité ce genre d'information est quasi impossible à obtenir...)?

Ceci constitue une situation du type: "**décision avec information imparfaite**"

S1. Pour répondre à la première question (Q1), nous représentons tout d'abord le problème à résoudre sous la forme graphique d'un arbre de décision (qui est pour l'instant assez simple à mettre aussi sous forme de tableau) avec le logiciel TreeAge par exemple:

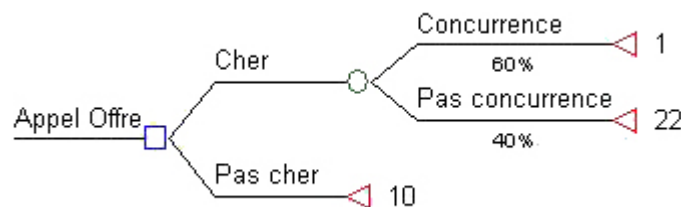


Figure: 65.4 - Arbre de décision sans informations avec TreeAge

Indiquons qu'il est possible de faire le même type d'arbres avec MS Office Visio mais la modélisation de Monte-Carlo n'y est pas incorporée et la création des formules prend beaucoup de temps (compter un facteur temps de 10 à 20 par rapport à TreeAge, Isograph ou @Risk).

Ensuite, en lançant le calcul de l'espérance à chaque branchement, appelée "**valeur monétaire espérée**" (VME), TreeAge nous donne simplement (ce logiciel a une option pour faire de la modélisation de Monte-Carlo mais l'exemple ici étant avec des probabilités fixes cette option est inutile à ce niveau):

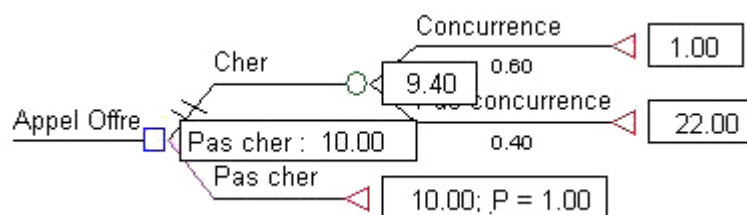


Figure: 65.5 - Espérances calculées sur l'arbre de décision

Ainsi, la réponse à la première question est que la stratégie donnant l'espérance de gain la plus grande

est la stratégie "Pas Cher" car il y a un gain espéré de 10 millions.

Avec la première décision (Cher) nous ne gagnerions en moyenne que:

$$22 \cdot 0.4 + 1 \cdot 0.6 = 9.4 \quad (65.2)$$

Remarque: Dans les arbres de décision construits avec TreeAge une règle de base est d'avoir à chaque branche probabiliste la somme des probabilités qui vaut 1!

S2. Pour répondre à la deuxième question (Q2) qui est de connaître la valeur financière de l'information donnée par l'espion, nous devons d'abord construire l'arbre (bon l'exemple est tellement simple qu'ici ce n'est pas vraiment nécessaire mais bon...) d'une situation dite de concurrence à "information parfaite" (car l'espion peut nous fournir une information tout à fait sûre).

L'arbre est facile à construire. Si l'espion nous dit que le groupe A va faire une offre, alors nous allons devoir proposer l'offre la moins chère. Dans le cas contraire, nous allons proposer l'offre la plus chère. Le scénario est donc le suivant:

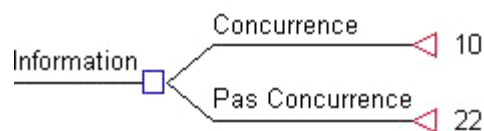


Figure: 65.6 - Arbre de décision avec information parfaite sous TreeAge

La probabilité qu'il y ait concurrence est de 60% et 40% qu'il n'y en ait pas. Donc "**l'espérance de la valeur monétaire de l'information parfaite**" (EVMIP) est dans une situation à information parfaite:

$$40\% \cdot 22 + 60\% \cdot 10 = 14.8 \quad (65.3)$$

Donc par rapport à la meilleure situation précédente nous avons un delta de 4.8 millions. C'est donc la valeur de l'information parfaite de l'espion.

S3. Concernant la troisième question (Q3) qui consiste à déterminer la valeur de l'information imparfaite fournie par une société de conseil. La seule certitude de bon sens que nous ayons est que cette information ne peut avoir une valeur supérieure à celle de l'information parfaite: elle aura donc une valeur comprise entre 0 et 4.8 millions.

Pour commencer, rappelons que selon l'énoncé, nous pensons que pour le mandat actuel, il y a 60% de probabilité qu'il y ait concurrence et la société de conseil dans le passé a eu 80% du temps raison (24 fois sur 30) lorsqu'elle avait dit qu'il y aurait concurrence (et donc 20% des autres fois tort...).

Respectivement, nous pensons pour le mandat actuel qu'il y a 40% de probabilité qu'il y ait non concurrence (1-60%) et la société de conseil dans le passé au eu 85% du temps raison (17 fois sur 20) lorsqu'elle avait dit qu'il n'y aurait pas de concurrence (et donc 15% des autres fois tort...).

Ce qui se résume sous forme de tableau:

		Proba.	Prévisions	
			Concurrence	Sans concurrence
Réalité	Concurrence	60%	80%	20%
	Sans concurrence	40%	15%	85%

Tableau: 65.1 - Décision avec information imparfaite (forme initiale)

Nous aimerions maintenant:

1. Calculer la probabilité qu'il y ait réellement concurrence ET que la société de conseil ait prévu de la concurrence.
2. Calculer la probabilité qu'il n'y ait réellement pas de concurrence ET que la société de conseil ait prévu de la concurrence.
3. Calculer la probabilité qu'il n'y ait réellement pas de concurrence ET que la société de conseil ait prévu pas de concurrence.
4. Calculer la probabilité qu'il y ait réellement concurrence ET que la société de conseil ait prévu pas de concurrence.

Pour calculer ces probabilités, nous allons utiliser la formule de Bayes démontrée dans le chapitre de Probabilités. Ainsi pour rappel, les probabilités a posteriori et a priori sont données par:

$$P(B/A) = \frac{P(A \cap B)}{P(A)} \text{ et } P(A/B) = \frac{P(A \cap B)}{P(B)} \quad (65.4)$$

d'où:

$$P(B \cap A) = P(B/A)P(A) = P(A/B)P(B) \quad (65.5)$$

Nous pouvons maintenant:

1. Calculer la probabilité qu'il y ait réellement concurrence ET que la société de conseil ait prévu de la concurrence. Nous avons alors en utilisant le tableau précédent:

$$P(A \cap B) = P(B/A)P(A) = 80\% \cdot 60\% = 48\% \quad (65.6)$$

où dans cette situation A est donc l'événement "il y a eu réellement concurrence" et B est l'événement "prévision de concurrence par la société de conseil". B/A est donc l'événement "prévision de concurrence par la société de conseil sachant qu'il y a eu réellement concurrence".

2. Calculer la probabilité qu'il n'y ait réellement pas concurrence ET que la société de conseil ait prévu de la concurrence. Nous avons alors en utilisant le tableau précédent:

$$P(A \cap B) = P(B/A)P(A) = 15\% \cdot 40\% = 6\% \quad (65.7)$$

où dans cette situation A est donc l'événement "il y a eu réellement pas de concurrence" et B est l'événement "prévision de concurrence par la société de conseil". B/A est donc l'événement "prévision de concurrence par la société de conseil sachant qu'il y a eu réellement pas de concurrence".

3. Calculer la probabilité qu'il n'y ait réellement pas concurrence ET que la société de conseil ait prévu pas concurrence. Nous avons alors en utilisant le tableau précédent:

$$P(A \cap B) = P(B / A)P(A) = 85\% \cdot 40\% = 34\% \quad (65.8)$$

où dans cette situation A est donc l'événement "il y a eu réellement pas de concurrence" et B est l'événement "prévision de pas de concurrence par la société de conseil". B/A est donc l'événement "prévision de pas concurrence par la société de conseil sachant qu'il y a eu réellement pas de concurrence".

4. Calculer la probabilité qu'il y ait réellement concurrence ET que la société de conseil ait prévu pas de concurrence. Nous avons alors en utilisant le tableau précédent:

$$P(A \cap B) = P(B / A)P(A) = 20\% \cdot 60\% = 12\% \quad (65.9)$$

où dans cette situation A est donc l'événement "il y a eu réellement concurrence" et B est l'événement "prévision de pas de concurrence par la société de conseil". B/A est donc l'événement "prévision de pas concurrence par la société de conseil sachant qu'il y a eu réellement concurrence".

Nous avons alors le tableau suivant qui résume de manière utilisable les scénarios possibles:

Réalité/Prévisions	Concurrence	Sans concurrence
Concurrence	48%	12%
Sans concurrence	6%	34%
Total	54%	46%

Tableau: 65.2 - Décision avec information imparfaite (forme finale)

À l'aide de ce tableau, nous vérifions bien que la somme des 4 premières cases donne 100% (c'est-à-dire l'ensemble des éventualités).

Nous voyons donc que 54% du temps la société de conseil prévoit de la concurrence (quelle que soit l'issue réelle) et 46% du temps aucune concurrence (quelle que soit l'issue réelle).

Nous avons alors l'arbre de décision suivant dans le logiciel TreeAge:

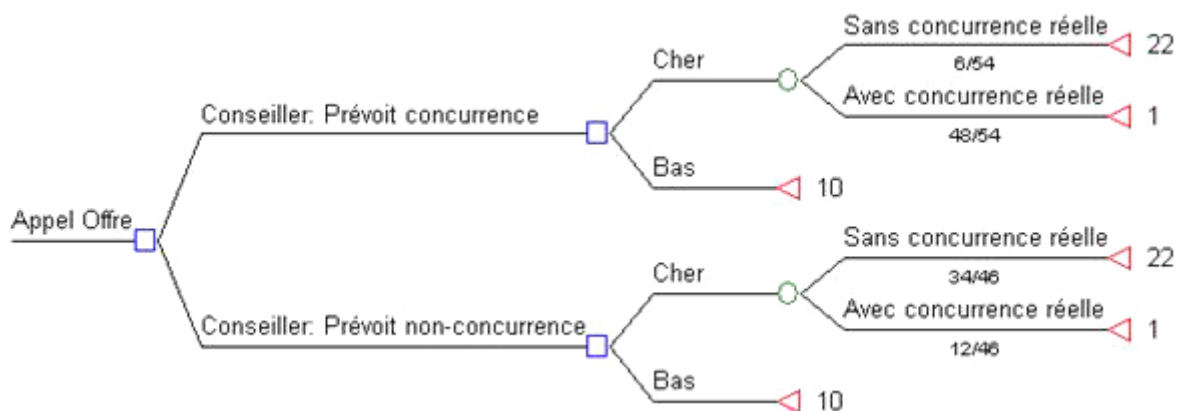


Figure: 65.7 - Arbre de décision avec information imparfaite avec TreeAge

Ce qui donne après calculs:

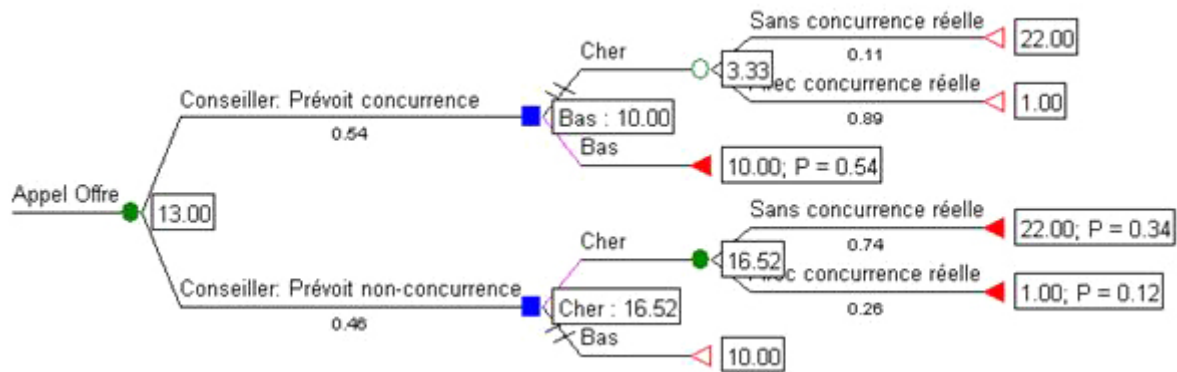


Figure: 65.8 - Espérances calculées sur l'arbre de décision

Nous nous retrouvons donc avec une espérance de gain de 13 millions moins le 1 million de paiement pour la société de conseil cela fait 12 millions.

Ainsi, la valeur de l'information imparfaite est de 2 millions par rapport aux 4.8 que rapporte l'information parfaite. Ce résultat est donc tout à fait logique.

Remarques:

R1. Ce type d'arbre est souvent utilisé en pharmaco-économie (analyse des coûts). Ainsi, la racine de l'arbre sera une infection X pour laquelle il existe plusieurs antibiotiques (branche A/B) et dont chacun à deux issues (traitement réussi/raté) avec deux résultats possibles (effets secondaires oui/non). Une probabilité est associée à chaque nœud et pour les nœuds terminaux des coûts de traitement. Ainsi, en calculant l'espérance, il est possible de déterminer le choix économique du meilleur antibiotique pour l'institut médical (évidemment ce n'est utile déontologiquement que lorsque le taux de succès des deux antibiotiques est proche et que le taux d'effets secondaires oui/non l'est aussi... sinon cette méthode ferait scandale!).

R2. Dans l'industrie, mes clients utilisent ces arbres avec des distributions de probabilités définies sur chaque nœud. Ils font ensuite une simulation de Monte-Carlo sur l'ensemble et font une analyse de la sensibilité (graphe Tornado) avec des suites décisionnelles comme celle de @Risk de Palisade.

1.3. FORME NORMALE D'UN JEU

Pour passer à la forme normale ou encore "**forme stratégique**", nous définissons une stratégie comme un plan d'action complet pour chaque joueur, qui spécifie un choix pour chaque nœud de l'arbre et donc pour chaque situation pouvant survenir au cours du jeu. La "**matrice des gains**" représente la situation stratégique des joueurs et les gains qu'ils recevront pour chaque stratégie.

Nous reprenons l'exemple précédant MAC/CAM et obtenons:

J_1 / J_2	CAM	MAC
CAM	600 , 200	100 , 100
MAC	100 , 100	200 , 600

Tableau: 65.3 - Matrice des gains d'un jeu à somme non nulle

Il s'agit donc d'une simple forme tabulaire du jeu.

Remarques:

R1. Nous voyons dans cette matrice que les intérêts des deux entreprises ne sont pas complètement opposés, elles progressent à chaque fois dans la même direction lorsque les stratégies sont opposées (si un perd, l'autre perd aussi et inversement). Ainsi, le jeu MAC/CAM est un jeu dont les gains ne progressent pas dans des directions (stratégies) opposées. Nous parlons alors de "**jeu non strictement compétitif**" (nous définirons ce concept de manière formelle un peu plus loin).

R2. Nous voyons également que quelle que soit la stratégie choisie par un des joueurs, chaque choix possible par l'autre joueur amènera toujours à des gains équivalents. Dès lors, nous disons alors que c'est un "**jeu sans tactique prudente**".

Définition: Une stratégie donnée est dite à "**tactique prudente**" (c'est le choix du numéro de la ligne pour le joueur ligne, ou du numéro de la colonne pour le joueur colonne) lorsque le gain d'un des joueurs est tel que lorsque par rapport à une stratégie choisie, l'ensemble des choix de son concurrent apporte un gain maximal à ce dernier. Le gain minimal assuré de J_1 est appelé le "**niveau de sécurité**" de J_1 .

Exemple:

J_1 / J_2	b_1	b_2	b_3	b_4
a_1	5, 5	6, 4	0, 10	4, 6
a_2	1, 9	7, 3	5, 5	6, 4
a_3	6, 4	7, 3	7, 3	8, 1
a_4	4, 6	8, 1	0, 10	2, 8
a_5	3, 7	5, 5	9, 0	0, 10

Tableau: 65.4 - Matrice des gains d'un jeu avec tactique prudente

Le joueur A peut penser que le joueur B est très perspicace, ou a beaucoup de chance, et est ainsi en mesure de choisir la meilleure réponse possible à toute tactique de A.

Ainsi:

- Si A choisit a_1 , B le devinant choisirait b_3 , et A aurait gagné 0 (tandis que B aurait gagné 10)
- Si A choisit a_2 , B le devinant choisirait b_1 , et A aurait gagné 1 (tandis que B aurait gagné 9)
- Si A choisit a_3 , B le devinant choisirait b_1 , et A aurait gagné 6 (tandis que B aurait gagné 4)
- Si A choisit a_4 , B le devinant choisirait b_3 , et A aurait gagné 0 (tandis que B aurait gagné 10)
- Si A choisit a_5 , B le devinant choisirait b_4 , et A aurait gagné 0 (tandis que B aurait gagné 10)

Le choix prudent de A est donc a_3 , qui lui assure de gagner au moins 6. Ce gain minimal assuré est le

niveau de sécurité. En faisant de même pour B s'il redoute l'extrême perspicacité de A le choix est b_1 . Cette tactique lui assure un gain de 4, qui est aussi son niveau de sécurité.

Si nous étudions le jeu MAC/CAM par sa matrice de gains, nous pouvons nous rendre compte qu'il y a deux issues remarquables où le gain des deux entreprises est maximum par rapport aux autres stratégies. Ces deux issues sont intéressantes à plus d'un titre:

Effectivement, les deux entreprises n'ont aucun regret quant à leur choix de stratégie. S'ils considèrent la stratégie de leur adversaire comme inéluctable, leur propre choix de stratégie est le meilleur possible. Nous disons que les deux issues sont des "**équilibres de Nash**" (nous définirons ce concept de manière formelle un peu plus loin). L'équilibre de Nash caractérise ainsi en quelque sorte la rationalité individuelle!

Remarque: Le jeu MAC/CAM comporte deux équilibres de Nash. Dès lors, nous ne sommes pas capables, sans aucune information complémentaire, de prédire quelle sera exactement la solution du jeu. Les deux résultats sont également vraisemblables.

C'est ainsi que la théorie des jeux fait apparaître la stratégie sociale la plus favorable aux deux joueurs: que les deux joueurs adoptent au moins le même système. Quant à savoir lequel... le jeu devra dès lors être coopératif.

Dans l'exemple précédent aussi, la conjonction des tactiques prudentes (a_3, b_1) constitue un équilibre de Nash (dans le sens où chacun des joueurs n'a pas intérêt à changer unilatéralement de stratégie s'il veut préserver le gain minimum). Cela tient à une particularité de ce jeu ! En d'autres termes:

1. Il existe de nombreux jeux qui n'ont pas d'équilibre
2. Il existe de nombreux jeux qui ont des équilibres qui ne correspondent pas à la conjonction des tactiques prudentes.

Remarque: Si dans un jeu, un couple d'issues est tel qu'il est impossible d'améliorer le score de l'un des deux joueurs sans diminuer le score de l'autre, nous disons que ces issues sont "**pareto-optimales**" ou "**pareto-efficientes**" (nous définirons ce concept de manière formelle un peu plus loin).

Exemple:

Dans ce jeu, deux joueurs J_1, J_2 s'affrontent à pierre, ciseaux, papier (PCP...). De façon générale, la pierre bat les ciseaux (en les émoussant), les ciseaux battent le papier (en le coupant), le papier bat la pierre (en l'enveloppant). Ainsi chaque coup bat un autre coup, fait match nul contre le deuxième (son homologue) et est battu par le troisième.

La forme extensive de ce jeu est trivialement:

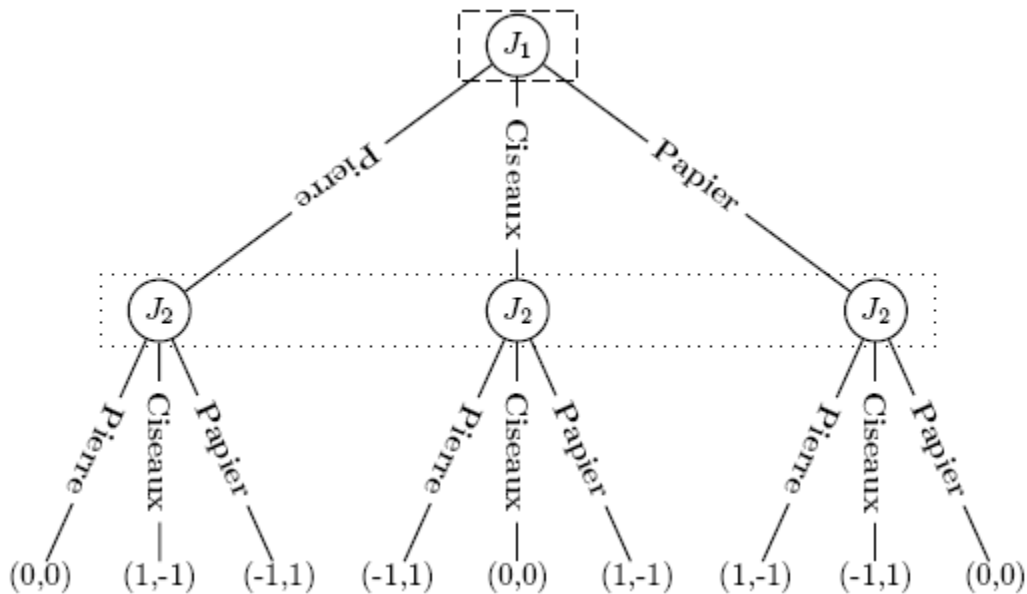


Figure: 65.9 - Forme extensive du jeu (arbre de Kuhn)

Pour faire apparaître la simultanéité du jeu sur la représentation, nous avons entouré les ensembles d'informations. J_2 sait que J_1 a choisi un élément, mais il ne sait pas lequel, donc il ne connaît pas le noeud exact où son propre choix va intervenir, et donc il est incapable de déterminer l'issue du jeu qui va être atteinte. Le jeu est donc à "information imparfaite".

Sous forme normale, nous avons donc:

J_1 / J_2	Pierre	Ciseaux	Papier
Pierre	0 , 0	1 , -1	-1 , 1
Ciseaux	-1 , 1	0 , 0	1 , -1
Papier	1 , -1	- 1 , 1	0 , 0

Tableau: 65.5 - Matrice des gains d'un jeu à somme nulle

Ce jeu est un jeu à somme nulle dans le sens où tout ce qui est gagné par l'un est perdu par l'autre. En d'autres termes, nous avons déjà vu que nous pouvons parler dès lors de jeux "strictement compétitifs".

Les jeux à somme nulle ont ceci de particulier en plus qu'il est toujours possible comme nous l'avons déjà mentionné de les représenter par leur demi-matrice (par rapport à un seul joueur donc) qui résume à elle seule tout le jeu puisque ce qui est gagné par ce joueur est perdu par l'autre et inversement:

J_1 / J_2	Pierre	Ciseaux	Papier
Pierre	0	1	-1
Ciseaux	-1	0	1
Papier	1	- 1	0

Tableau: 65.6 - Demi-matrice des gains du jeu

Au besoin, si les gains et pertes respectives de jeu n'ont pas le même "delta", il suffit de définir une fonction d'utilité adéquate pour l'autre joueur telle qu'il soit toujours possible pour n'importe quel jeu strictement compétitif où les gains ne sont pas opposés et égaux d'être mis sous la forme d'une demi-matrice. Nous démontrerons qu'il existe une telle fonction d'utilité.

Remarque: Sur la demi-matrice d'un jeu à somme nulle, il est très facile de reconnaître s'il existe un équilibre de Nash ou non. Par exemple:

2	0
1	3

Tableau: 65.7 - Demi-matrice sans équilibre de Nash

Dans ce jeu, la ligne 2 est la tactique prudente du joueur ligne et le joueur colonne choisira la colonne 1 comme tactique prudente dans laquelle ne se trouve pas la plus grande perte. Dès lors, le joueur ligne aura intérêt à se déplacer en première ligne donc les tactiques prudentes conjointes ne sont pas un équilibre et par ailleurs, il n'y a pas d'équilibre de Nash!!

Dans le duel tactique ainsi défini, l'espérance du joueur ligne est le maximum des minimums de lignes, c'est-à-dire le "maximin", tandis que l'espérance du joueur colonne est le minimum des maximums de colonnes, c'est-à-dire le "minimax".

Définitions:

D1. Le "**maximin**", appelé aussi parfois "**critère de Wald**", est un critère pessimiste. Il s'agit effectivement selon ce critère, de maximiser le résultat minimum. Pour le mettre en oeuvre, il convient dans un jeu à somme nulle et à information parfaite:

- Pour chaque décision (ou stratégie), de retenir le résultat le plus faible
- Parmi, les moins bons résultats, choisir le plus élevé des moins bons résultats des différentes stratégies.

Ce genre d'approche peut donc outre sous forme tabulaire être représenté sous la forme d'un arbre de décision.

D2. Le "**maximax**", selon la même logique que le critère précédent consiste à retenir le meilleur des résultats des différentes stratégies possibles, c'est donc un critère optimiste. Pour le mettre en oeuvre, il convient dans un jeu à somme nulle et à information parfaite:

- Pour chaque décision (ou stratégie), de retenir le résultat le plus attendu le plus élevé
- Parmi, les meilleurs résultats, choisir le plus élevé des meilleurs résultats des différentes stratégies.

D3. Le "**minimax**", est un critère appelé parfois "**critère de Von Neumann**" prudent qui consiste à retenir le plus petit des meilleures résultats des stratégies possibles. Pour le mettre en oeuvre, il convient dans un jeu à somme nulle et à information parfaite:

- Pour chaque décision (ou stratégie), de retenir le résultat le plus attendu le plus élevé

- Parmi, les meilleurs résultats, choisir le moins élevé des meilleurs résultats des différentes stratégies.

Ce genre d'approche peut donc outre sous forme tabulaire être représenté sous la forme d'un arbre de décision.

Si et seulement si le maximin est égal au minimax, leur valeur commune, qui est l'espérance commune aux deux adversaires, est appelée la "**valeur du jeu**" (nous le démontrons juste quelques lignes en dessous), et tout couple formé par une telle tactique prudente du joueur ligne et une tactique prudente du joueur colonne défini un équilibre (pour cette raison l'exemple précédent n'a pas d'équilibre).

Exemple:

2	4
1	3
0	4

Tableau: 65.8 - Demi-matrice avec équilibre de Nash

Dans ce jeu, la ligne 1 est la meilleure tactique prudente du joueur ligne et le joueur colonne choisira la colonne 1 comme tactique prudente dans laquelle se trouvent les plus petites pertes. Dès lors, la cellule supérieure gauche correspond aux tactiques prudentes conjointes et correspond comme nous le voyons à un équilibre de Nash.

Définition: Dans un jeu à somme nulle nous appelons "**col**", l'utilité (vue dans le sens du gain ou de la perte) qui est à la fois minimum dans sa ligne et maximum dans sa colonne (ce qui est le cas de l'exemple précédent où l'équilibre est un col).

Démontrons maintenant que dans tout jeu à somme nulle, si et seulement si les niveaux de sécurité des deux joueurs sont opposés (le minimax est égal au maximin), la conjonction des tactiques prudentes est toujours un équilibre.

Reprenons la définition d'un couple formé:

- d'une tactique prudente α_p pour le joueur A, lui assurant de gagner au moins v
- d'une tactique prudente b_p pour le joueur B, lui assurant de gagner au moins $-w$

Dans le cas d'un jeu à somme nulle, nous pouvons toujours redéfinir la fonction d'utilité d'un des joueurs de manière à obtenir $v = w$ comme nous l'avons vu afin de pouvoir écrire la demi-matrice. Dès lors, observons ce qu'il se passe (en se rappelant bien que dans un tel jeu, le gain équivaut à la perte donc par extension quand le gain est minimal pour l'un la perte est minimale pour l'autre):

Le couple (α_p, b_p) , comme tout couple qui contient α_p , assure A de gagner au moins v et in extenso assure B de gagner au moins $-v$ (puisque $v = w$).

A n'a donc aucun intérêt à s'écarter unilatéralement de α_p , puisque B s'est assuré de perdre au plus v dans la stratégie de A. De même, B n'a aucun intérêt à s'écarter unilatéralement de la tactique b_p , puisque A s'est assuré de gagner au moins v .

Par conséquent, dans le cas où les niveaux de sécurité des deux joueurs sont égaux et opposés, la

conjonction des tactiques prudentes est un équilibre.

Nous avons déjà vu précédemment un exemple dans lequel les niveaux n'étaient pas exactement opposés.

1.3.1. JEUX RÉPÉTITIFS

Supposons qu'un homme J_1 et une femme J_2 aillent au cinéma. Une fois sur place, ils doivent choisir entre aller voir un documentaire ou une comédie. L'un des deux préfère les documentaires et l'autre les comédies, mais tous deux préfèrent voir un film ensemble plutôt que séparément: c'est... la guerre des sexes (GDS...)

Les stratégies disponibles pour chacun des deux joueurs, en considérant qu'ils font leur choix simultanément (ce qui est peu vraisemblable dans un cas réel, la galanterie obligeant à désynchroniser le jeu au profit de la femme:-)), sont alors:

- Aller voir un documentaire, ce que nous noterons *Doc*
- Aller voir une comédie, ce que nous noterons *Com*

La matrice des gains sera alors:

J_1 / J_2	<i>Doc</i>	<i>Com</i>
<i>Doc</i>	2 , 3	1, 1
<i>Com</i>	1 , 1	3 , 2

Tableau: 65.9 - Matrice pour jeu répétitif

D'abord, nous pouvons remarquer que GDS n'est pas un jeu strictement compétitif (donc inutile d'essayer de le représenter sous la forme d'une demi-matrice) et qu'il s'agit d'un jeu de coordination. Deuxièmement, nous remarquons que les deux issues à gain maximum sont des équilibres de Nash (nous ne pouvons donc prédire l'issue du jeu).

Ce jeu a cependant ceci de particulier par rapport aux précédents ce que c'est un jeu à une seule étape. Supposons ainsi maintenant que le couple retourne au cinéma la semaine suivante, et qu'il doive à nouveau faire ce choix. Nous pouvons de nouveau représenter cette situation par un jeu, qui n'est en fait que la répétition de GDS, notons le GDS2.

Si nous considérons que lors de la deuxième étape chacun des deux joueurs sait ce que l'autre a choisi lors de la première étape, les stratégies disponibles sont maintenant des stratégies conditionnelles: elles peuvent tenir compte des coups joués par l'adversaire lors d'étapes précédentes.

La description de ces stratégies suit le schéma suivant: nous jouons X_1 au premier coup, puis si l'autre a choisi le documentaire lors de la première sortie, alors nous jouons X_{2d} , sinon X_{2c} , avec X_1, X_{2d}, X_{2c} prenant leur valeur dans l'ensemble $\{Doc, Com\}$. Nous noterons cette stratégie:

$$(X_1, X_{2d} | [X_1, Doc]; X_{2c} | [X_1, Com]) \quad (65.10)$$

Nous pouvons lire cette notation de la manière suivante: nous jouons X_1 , puis si nous nous retrouvons

en $[X_1, Doc]$ alors nous jouons X_{2d} , ou si nous nous retrouvons en $[X_1, Com]$ alors nous jouons X_{2c} .
 Dans le cas GDS2, nous avons donc 8 stratégies:

1. $(Doc, Doc | [Doc, Doc]; Doc | [Doc, Com])$: nous choisissons toujours le documentaire
2. $(Doc, Doc | [Doc, Doc]; Com | [Doc, Com])$: nous choisissons toujours le documentaire, sauf si la première fois nous nous sommes retrouvé(e)s seul(e)s
3. $(Doc, Com | [Doc, Doc]; Doc | [Doc, Com])$: nous choisissons toujours le documentaire, sauf si la première fois nous avons tous les deux choisi le documentaire
4. $(Doc, Com | [Doc, Doc]; Com | [Doc, Com])$: la première fois nous choisissons le documentaire et la seconde, la comédie.
5. $(Com, Doc | [Com, Doc]; Doc | [Com, Com])$: la première fois nous choisissons la comédie et la seconde, le documentaire
6. $(Com, Doc | [Com, Doc]; Com | [Com, Com])$: nous choisissons toujours la comédie, sauf si la première fois, nous nous sommes retrouvé(s) seul(e)s.
7. $(Com, Com | [Com, Doc]; Doc | [Com, Com])$: nous choisissons la comédie sauf si la première fois, nous avons tous les deux choisi la comédie.
8. $(Com, Com | [Com, Doc]; Com | [Com, Com])$: nous choisissons toujours la comédie.

Pour chaque issue de GDS2, les vecteurs d'utilité sont déterminés en effectuant la somme des vecteurs obtenus pour chacune des étapes considérées comme des issues de GDS. Nous dirons que GDS2 est un "super jeu" dont GDS est le "jeu constitutif".

Définition: Un "équilibre parfait en sous-jeux" correspond à une combinaison stratégique dont les actions choisies pour chaque sous-jeu sont des équilibres de Nash.

Remarque: Un "sous-jeu" est simplement un sous-arbre de l'arbre de jeu.

Voyons maintenant tous ces concepts de manière ensembliste (accrochez-vous un peu ;-)

1.4. FORME ENSEMBLISTE D'UN JEU

Nous avons donc vu jusqu'à maintenant qu'il existe un certain nombre d'éléments qui composent un jeu: les joueurs, les actions et stratégies des joueurs, les déroulements et les étapes du jeu, les résultats du jeu et les informations dont disposent les joueurs de chaque choix d'action.

Définitions:

D1. Les règles d'un jeu indiquent:

- La succession des étapes du jeu, et l'ordre dans lequel interviennent les joueurs
- Les actions qui sont autorisées à chaque étape
- Les informations dont dispose le joueur chaque fois qu'il doit prendre une décision

Nous avons vu qu'il y a deux formes de représentations possibles pour un jeu jusqu'à maintenant. L'une d'entre elles utilise un arbre (une forme extensive) et l'autre une table (forme normale). Sous une expression formelle cela donne:

D2. Un arbre de jeu $A = (D, I, p)$ est la donnée:

- D'un ensemble D de noeuds de décisions, ou situations de jeu
- D'un ensemble I d'issues de jeu, avec $D \cap I = \emptyset$ (donc un noeud n'est pas considéré comme une issue!)
- D'un élément d_0 de D et d'une fonction p de $D \cup I \setminus \{d_0\}$ dans D telle que:

$$p : D \cup I \setminus \{d_0\} \rightarrow D \quad (65.11)$$

appelée "**fonction prédécesseur**", qui pour chaque situation de jeu, ou issue, indique l'unique action (décision ou situation, d'où le fait que nous enlevons au moins un élément D de l'ensemble de départ) qui a permis d'arriver à cette situation, ou issue.

Pour déterminer l'issue d'un jeu, il suffit de connaître les stratégies utilisées par chacun des joueurs. Une stratégie est une combinaison d'actions autorisées par les règles du jeu jusqu'à la fin de celui-ci. Il existe plus précisément trois types de stratégies.

D3. Une "**stratégie pure**" s pour un joueur n est une application de l'ensemble D_n des noeuds de décision de ce joueur vers l'ensemble D de tous les noeuds de décision du jeu telle que:

$$\boxed{\forall d \in D_n, p(s(d)) = d} \quad (65.12)$$

Plus simplement dit, une stratégie pure est une stratégie ne faisant intervenir aucune forme de hasard, qui est donc complètement déterministe.

Remarque: La fonction stratégie pure n'est que la fonction réciproque de p telle que $p^{-1} = s$.

D4. Une "**stratégie mixte**" pour un joueur n est une distribution de probabilité $(P_1, P_2, \dots, P_m)$ avec $\sum_i P_i = 1$ sur l'ensemble de ses stratégies pures $\mathcal{S} = \{s_1, s_2, \dots, s_m\}$.

Exemple:

Les tirs aux buts (penaltys) sont une forme de jeu à stratégie mixte. Effectivement, le gardien de but doit anticiper le tir et ne peut l'analyser. Il doit donc choisir au hasard s'il restera au milieu, s'il ira à gauche ou à droite. Idem, pour l'attaquant (normalement le gardien doit se lancer au moment même où l'attaquant tire) qui ne sachant pas où se lancera le gardien tirera donc au hasard.

Remarques:

R1. Une stratégie pure peut être regardée ainsi comme une stratégie qui donne la probabilité 1 à s_i et 0 à toutes les autres.

R2. Dans notre définition de l'ensemble des stratégies, il y a un nombre fini de stratégies pour chaque agent mais en économie, les ensembles des stratégies sont souvent continus et contiennent une infinité de stratégies possibles (choix de quantité, de prix, etc.).

Naturellement, le résultat obtenu par le joueur ne peut pas être garanti de façon certaine, puisque le processus de choix de la décision fait intervenir des probabilités.

Une stratégie pure est donc une stratégie faisant le choix d'une parmi toutes les stratégies mixtes et qui utilise celle-ci durant toute la durée de jeu. Un joueur utilisant une stratégie mixte face à un joueur utilisant une stratégie pure utilisera (sera forcé) donc lui aussi une stratégie pure pour une rencontre, mais n'utilisera pas toujours la même stratégie pure lors de toutes leurs rencontres.

D5. Une "**stratégie de comportement**" pour un joueur n est un ensemble $C_n = \{..., c_i, ...\}$ où i est un élément de D_n (donc un numéro de noeud) et c_i une distribution de probabilité sur le sous-ensemble $p^{-1}(i)$ des successeurs du noeud de décision i .

D6. Une "**combinaison stratégique**" est un vecteur de stratégies dont chaque élément correspond à la stratégie utilisée par un joueur participant au jeu. La donnée d'une combinaison stratégique détermine donc de manière complète l'issue du jeu.

Les joueurs doivent avoir des préférences parmi les issues qui sont à leur portée. C'est avec la définition de ces préférences que nous pouvons caractériser la rationalité d'un joueur. La relation de préférence que nous noterons $\succsim$, est une relation binaire sur l'ensemble des issues d'un jeu. Nous noterons $x \succsim y$ et nous dirons que " x est au moins aussi bon que y ". Nous pouvons alors définir la préférence stricte $\succ$ telle que:

$$x \succ y \Leftrightarrow x \succsim y \text{ mais pas } y \succsim x \quad (65.13)$$

que nous lirons " x est préféré à y ", et la relation d'indifférence:

$$x \sim y \Leftrightarrow x \succsim y \text{ et } y \succsim x \quad (65.14)$$

Remarque: Nous réutiliserons ces concepts en économétrie lors de notre étude de la théorie de la préférence.

D7. Une relation de préférence $\succsim$ est dite "**relation rationnelle**", si elle est complète (réflexive) et transitive. Dans ce cas, comme nous l'avons vu dans le chapitre des Opérateurs (section Arithmétique), nous avons affaire à un préordre.

D8. Une "**fonction d'utilité**", ou encore "**fonction de paiement**" ("**payoff function**" en anglais) est une fonction de l'ensemble des issues d'un jeu à n joueurs vers $\mathbb{R}^n$ qui associe les utilités retirées par chaque

joueur.

Si U est une fonction d'utilité, nous noterons U_i la fonction de l'ensemble des issues d'un jeu vers $\mathbb{R}$ correspondant aux utilités du joueur i . Une telle fonction sera dite représentant de la relation de préférence $\succsim$ si pour toute issue $x, y \in I$, nous avons:

$$x \succsim y \Leftrightarrow U_i(x) \geq U_i(y) \quad (65.15)$$

La théorie de l'utilité dont fait usage la théorie des jeux axiomatise le fait que seule cette notion de préférence est importante. En bref, nous dirons que seul l'ordre de préférence de l'utilité des issues est important, la valeur des gains apportés par chaque issue étant sans importance.

Nous pouvons maintenant étendre la définition du jeu:

D9. Un "jeu sous forme développée" $J = (A, N, U, F)$ est la donnée:

- d'un arbre de jeu $A = (D, I, p)$
- d'un ensemble N de joueurs
- d'une fonction d'utilité U qui donne pour un joueur donné son gain
- d'un ensemble de partitions d'informations F , dont chaque élément est une partition de D et indique les états du jeu que le joueur est capable de distinguer

Remarque: Comme nous l'avons déjà précisé, un jeu sous forme développée est également dit "forme extensive", ou encore "arbre de Kuhn".

D10. Un jeu est à "information complète" quand chaque joueur connaît l'ensemble des composantes du jeu, et à "information incomplète" sinon. Il est à noter que de parler d'un jeu à information complète revient à dire que F ne contient qu'une seule partition et donc que les joueurs n'ont qu'une seule vue sur l'arbre de jeu.

D11. Un jeu est à "information parfaite" quand l'unique élément de F se réduit à une partition de D où chaque noeud de décision forme un sous-ensemble, c'est-à-dire que chaque élément de la partition est un noeud de l'arbre et réciproquement. Plus simplement, nous pouvons dire que dans ce cas les joueurs peuvent savoir à chaque instant quel noeud de l'arbre est atteint. Dans le cas contraire le jeu est dit à information imparfaite.

Remarque: Nous pouvons remarquer que tous les jeux simultanés, c'est-à-dire dans lesquels les joueurs font leur choix en même temps, sont des jeux à information imparfaite. En effet, au moment de son choix, le joueur ne sait pas sur quel noeud de décision il se trouve.

Maintenant nous pouvons en venir à définir ce qu'est la matrice des gains:

D12. Un "jeu sous forme normale" $J = (N, S, U)$ est la donnée:

- d'un ensemble N de joueurs
- d'un ensemble S de combinaisons stratégiques
- d'une fonction d'utilité U définie sur S

Ainsi, un jeu sous forme normale est également dit "**jeu sous forme stratégique**". Nous simplifions d'ailleurs la donnée du jeu à la donnée de la fonction d'utilité, sous la forme d'une matrice de gains (ou de paiements).

D13. Un jeu est "**concurrentiel pur**" ou "**strictement compétitif**" si:

$$\boxed{\forall i, j \in I \times I, \exists m, n \in N \times N \quad (U_m(j) - U_m(i)) \cdot (U_n(j) - U_n(i)) < 0} \quad (65.16)$$

Donc un jeu est strictement compétitif si pour un ensemble couple d'issues, les gains d'un au moins des joueurs diminuent. Si les deux joueurs ont pour un couple d'issues, leurs gains respectifs qui augmentent ou diminuent, alors nous avons:

$$\underbrace{(U_m(j) - U_m(i))}_{>0} \cdot \underbrace{(U_n(j) - U_n(i))}_{>0} > 0 \quad (65.17)$$

le jeu n'est dès lors plus strictement positif. Nous en avons par ailleurs donné des exemples au début de ce chapitre.

D14. Un jeu strictement compétitif est un "**jeu à somme nulle**" si:

$$\boxed{\forall i \in I, \sum_{n \in N} U_n(i) = 0} \quad (65.18)$$

Un jeu est à somme nulle quand les intérêts des joueurs sont diamétralement opposés. Dans un jeu à deux joueurs à somme nulle, par exemple, ce qui est gagné par l'un est perdu par l'autre. Ce terme trouve son origine dans les jeux de salon comme le poker où un joueur qui veut gagner de l'argent doit le faire aux dépens des autres. Les échecs sont un jeu à somme nulle.

D15. Un "**super jeu**" $\mathcal{J} = (J, T, \Omega)$ est la donnée:

- d'un jeu constitutif $J = (N, S, U)$
- du nombre de répétitions T
- du vecteur $\Omega = (\omega_1, \dots, \omega_n)$ de taux d'escompte d'utilité, ω_i étant le taux d'escompte du joueur i (souvent pris comme égal à l'unité)

Ainsi, comme nous en avons déjà fait mention lors de notre jeu répétitif GDS2, nous considérons qu'à une étape t le choix dicté par une combinaison stratégique s au joueur n est noté $s_{n,t}$ et que l'utilité, pour ce même joueur, obtenue à cette étape du jeu, c'est-à-dire l'utilité issue du jeu constitutif correspondant, est notée $U_n(s_{n,t})$, alors l'utilité associée à l'issue du super jeu est:

$$U_n = \sum_{t=0}^{T-1} \omega_n^t U_n(s_{n,t}) \quad (65.19)$$

il est clair que si $\forall n, t \ \omega_n^t = 1$ nous retrouvons une définition intuitive simple de la cumulation des gains.

1.5. FORME GRAPHIQUE D'UN JEU

Nous avons maintenant amassé suffisamment d'éléments pour avoir une approche probabiliste et opérationnelle de jeux à somme nulle relativement simples.

Comme il est toujours relativement difficile de ne pas être trop théorique pour que ce domaine reste compréhensible étudions les formes graphiques via un exemple.

Considérons deux sociétés que nous nommerons respectivement $S1$ et $S2$ qui sont spécialisées dans la vente à grande échelle d'un certain produit et qui forment un oligopole bilatéral en concurrence parfaite (cf. chapitre d'Économie). La société $S1$ décide d'investir un nouveau marché, constitué par un ensemble de régions d'importances comparables.

La pénétration dans différentes régions s'opère grâce à l'installation d'un présentoir dans des chaînes de magasins $C1$ ou $C2$ dans chacune des régions. Pour mieux motiver ses détaillants, la société $S1$ ne choisira qu'une seule chaîne de distribution ($C1$ ou $C2$) par région pour vendre ses produits.

La société $S2$ ayant pris connaissance du projet de la société $S1$ décide alors aussi d'investir le marché de manière similaire.

Le problème pour chaque société est de savoir, pour chaque région, s'il vaut mieux faire installer un présentoir dans la chaîne de magasins $C1$ ou $C2$ ou ne pas en faire installer du tout, c'est-à-dire nulle part (ce que nous noterons NP).

Suite à une étude de marché (il faut bien obtenir au moins quelques chiffres au départ pour faire des maths...) la société $S1$ apprend que ses gains par rapport au concurrent seraient ceux représentés dans le tableau ci-dessous:

$S1 / S2$	$C1$	$C2$	NP
$C1$	0	2	4
$C2$	6	-3	8
NP	-3	-5	0

Tableau: 65.10 - Préparation du jeu pour analyse graphique

La société $S2$ arrive au même résultat suite à une étude de marché (nous simplifions par cette hypothèse l'analyse du problème).

Remarques:

R1. Puisque tout ce que gagne un concurrent serait perdu par l'autre, le jeu est à somme nulle (d'où le fait qu'il n'y ait qu'une seule valeur dans chaque cellule)

R2. Nous supposons que les deux sociétés ne peuvent et ne veulent pas communiquer entre elles, en d'autres termes qu'il s'agit d'un jeu non coopératif.

Commençons par analyser quelles sont les stratégies qui n'ont aucun intérêt pour l'une ou l'autre des

sociétés.

Pour cela, regardons s'il y a une stratégie qui ne sera jamais choisie par S1 quelle que soit la stratégie de S2:

1. Si S2 choisit C1 alors S1 aura pour meilleur intérêt à choisir C2
2. Si S2 choisit C2 alors S1 aura pour meilleur intérêt à choisir C1
2. Si S2 choisit NP alors S1 aura pour meilleur intérêt à choisir C2

Nous voyons ici que quel que soit le choix de S2, la société S1 ne choisira jamais NP. Donc la stratégie NP pour S1 est totalement dominée et peut être éliminée.

De même, regardons s'il y a une stratégie qui ne sera jamais choisie par S2 quelle que soit la stratégie de S1.

1. Si S1 choisit C1 alors S2 aura pour meilleur intérêt à choisir C1 (car ainsi S2 ne perdra rien (0))
2. Si S1 choisit C2 alors S2 aura pour meilleur intérêt à choisir C2 (car ainsi S2 gagnera 3)
3. Si S1 choisit NP alors S2 aura pour meilleur intérêt à choisir C2 (car ainsi S2 gagnera 5)

Nous voyons ici que quel que soit le choix de S1, la société S2 ne choisira jamais NP. Donc la stratégie NP pour S2 est totalement dominée et peut être éliminée.

Le tableau se simplifie alors de la manière suivante:

S1 / S2	C1	C2
C1	0	2
C2	6	-3

Tableau: 65.11 - Simplification du jeu pour analyse graphique

Par ailleurs, ce jeu ne contient pas d'équilibre de Nash (donc aucune stratégie pure n'est avantageuse). Il est donc sans équilibres. Effectivement, si S1 choisit C1 alors S2 a intérêt à choisir aussi C1. Mais S1 a alors meilleur intérêt à jouer C2. Mais S2 a maintenant intérêt à choisir plutôt C2. Ce qui redonne à S1 l'envie de choisir C1...

Étudions maintenant l'aspect ensembliste, en d'autres termes l'aspect du jeu qui va donner la stratégie mixte à adopter par S1 avec la répartition du choix ad hoc pour que celle-ci ait un gain maximal.

Pour cela, appelons p et q les fréquences avec lesquelles les sociétés S1 et S2 choisissent la chaîne de magasin C1.

S1 / S2	C1	C2	
C1	0	2	p
C2	6	-3	$1-p$
	q	$1-q$	

Tableau: 65.12 - Mise sous forme paramétrique du jeu pour analyse graphique

Ces probabilités doivent être interprétées de la manière suivante:

1. Si p et q sont égaux à l'unité alors pour toutes les régions, ce sera la chaîne $C1$ qui s'occupera de la commercialisation
2. Si p et q sont par exemple $9/11$ et respectivement $5/11$ cela signifiera que la société $S1$ donnera le droit de vente à la chaîne de magasins $C1$ dans 9 régions sur 11 (les deux restantes étant pour $C2$) et respectivement la société $S2$ donnera le droit de vente à la chaîne de magasins $C1$ dans 5 régions sur 11 (les 6 restantes étant pour $C2$).

Donc commençons notre étude. Nous allons nous mettre dans une optique d'analyse dans laquelle la société $S1$ cherche sa stratégie mixte de manière à maximiser son gain (ou utilité) que nous noterons v et à connaître la stratégie mixte de la société $S2$ afin qu'elle minimise sa perte v (puisque c'est un jeu à somme nulle et tout ce que gagne l'un l'autre le perd).

Le système d'équation sera alors naturellement pour la société $S1$:

$$\begin{aligned} & \max(v) \\ & 6 \cdot (1-p) + 0 \cdot p \geq v \\ & 2 \cdot p - 3 \cdot (1-p) \geq v \\ & 0 \leq p \leq 1 \end{aligned} \quad (65.20)$$

et pour la société $S2$:

$$\begin{aligned} & \min(v) \\ & 2 \cdot (1-q) + 0 \cdot q \leq v \\ & 6 \cdot q - 3 \cdot (1-q) \leq v \\ & 0 \leq q \leq 1 \end{aligned} \quad (65.21)$$

Or, nous retrouvons ici une situation remarquable. Effectivement, il ne s'agit que de deux formes standards de programmation linéaire (cf. chapitre de Méthodes Numériques). Nous avons vu lors de notre étude de celle-ci que lorsqu'il n'y a qu'une seule inconnue par forme (ou système) alors il est possible de passer par une résolution graphique sans faire usage de l'algorithme du simplexe.

Après simplification cela donne:

$$\begin{array}{cc} S1 & S2 \\ \max(v) & \min(v) \\ v \leq -6 \cdot p + 6 & v \geq -2 \cdot q + 2 \\ v \leq 5 \cdot p - 3 & v \geq 9 \cdot q - 3 \\ 0 \leq p \leq 1 & 0 \leq q \leq 1 \end{array} \quad (65.22)$$

et la représentation graphique de v en fonction de p correspondante:

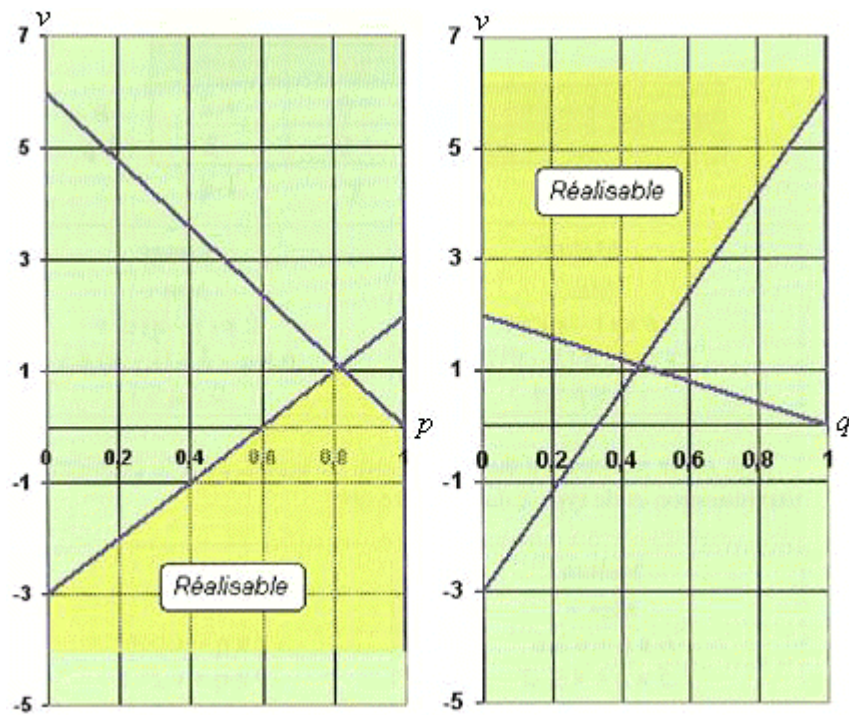


Figure: 65.10 - Correspondance des inéquations du problème d'optimisation

En résolvant avec l'algorithme du simplexe, nous avons comme valeurs optimales pour les deux systèmes respectifs (il est aussi possible de lire la valeur approximative sur les graphiques mais bon...):

$$\begin{array}{cc} S1 & S2 \\ v = 12/11 & v = 12/11 \\ p = 9/11 & q = 5/11 \end{array} \quad (65.23)$$

La société S1 peut par conséquent se garantir un gain moyen v (nous devrions parler "d'espérance" pour être rigoureux) de $12/11$. Effectivement:

$$v = E(U_i) \begin{cases} = 6 \cdot (1-p) + 0 \cdot p = 6 \cdot (1-9/11) = 12/11 \\ = 2 \cdot p - 3 \cdot (1-p) = 2 \cdot 9/11 - 3 \cdot (1-9/11) = 12/11 \end{cases} \quad (65.24)$$

et la probabilité p donnant au fait la distribution entre les chaînes de magasins C1 qui aura $9/11$ du marché de l'ensemble des régions et C2 le reste soit $2/11$ (la somme devant faire bien évidemment 1).

La société S2 peut par conséquent se garantir aussi un gain moyen v de $12/11$. Effectivement:

$$v = E(U_i) \begin{cases} = 2 \cdot (1-q) + 0 \cdot q = 2 \cdot (1-5/11) = 12/11 \\ = 6 \cdot q - 3 \cdot (1-q) = 6 \cdot 5/11 - 3 \cdot (1-5/11) = 12/11 \end{cases} \quad (65.25)$$

et la probabilité q donnant la distribution entre les chaînes de magasins C1 qui aura $5/11$ du marché de l'ensemble des régions et C2 le reste soit $6/11$.

2. JEUX COOPÉRATIFS ET NON-COOPÉRATIFS

Une première approche (sans faire usage des maths dans un premier temps) de cette attitude d'esprit (forme de jeu) est accessible à de jeunes enfants (sans qu'ils le sachent!).

 Exemple:

Imaginons deux enfants, l'un et l'autre gourmands, en présence d'un gâteau homogène, parfaitement divisible (et très bon...). Si la maman fait deux parts, il y aura immanquablement des disputes, chacun trouvant plus grosse la part de l'autre. Le seul moyen (hors dictat) d'éviter toute dispute est pour la mère d'imposer la règle suivante: l'un des enfants effectue le partage, et l'autre choisit en premier sa part. Celui qui coupe ne peut pas raisonner en tenant compte de ses seules préférences, qui le pousseraient à se couper une grosse part. Il sait en effet que l'autre pourra choisir sa part. Si donc il coupe une part plus grosse que l'autre, il risque de la retrouver dans l'assiette du voisin. Il va donc s'efforcer de couper des parts aussi égales que possibles, à ses yeux. Ainsi, quel que soit le choix de l'autre, il ne s'estimera pas maltraité. C'est cette anticipation du choix d'un autre décideur qui constitue l'originalité de la théorie de la décision et de la coopération !

Définitions:

D1. La partie de la théorie des jeux qui s'occupe de la détermination des éléments socialement préférables (au niveau du groupe plutôt que de l'individu seul en d'autres termes) de l'ensemble des issues I est souvent dite "**coopérative**" ou "**coalitionnelle**". Elle nécessite que les différentes parties puissent communiquer entre elles et... qu'elles soient rationnelles.

D2. La partie dite, au contraire, "**non-coopérative**" ou "**stratégique**" ne s'intéresse pas à la mise en oeuvre des solutions préconisées par la théorie des jeux coopératifs qui ont force de loi. Elle suppose que les différentes parties ne communiquent pas entre elles ou ne sont pas rationnelles.

Cette distinction entre jeux coopératifs et jeux non-coopératifs prête souvent à confusion. Essayons de la dissiper pour partie. Tout d'abord, cette distinction ne signifie nullement que les comportements que nous concevons intuitivement comme "coopératifs", au sens où ils induisent une part de sacrifice de nos intérêts propres au profit d'un bien jugé supérieur, ne pourront apparaître que dans le cadre des jeux coopératifs, au contraire! Les jeux stratégiques se soucient beaucoup de l'apparition endogène de tels comportements. Inversement, les jeux coopératifs sont très attentifs au respect des intérêts des individus. C'est là d'ailleurs l'une des difficultés principales qu'il leur faut affronter: si sacrifice individuel pour le bien commun il doit y avoir, qui doit se sacrifier ? Et pourquoi tel individu plutôt qu'un autre ?

Une fois défini l'ensemble I unanimement considéré comme représentant toutes les issues possibles du problème que nous cherchons à résoudre, il nous faut déterminer des critères qui permettent de sélectionner le "meilleur" état possible, compte tenu des appréciations diverses et contradictoires dont I fait l'objet par les différents citoyens en présence.

Nous savons que cette appréciation se mesure au moyen de la fonction d'utilité U_i définie sur I et prenant ses valeurs dans $\mathbb{R}$. Ainsi, si le système que nous considérons comporte N individus et si $x \in I$ est l'issue sélectionnée, $U_i(x)$ est le gain accordé par le joueur i à x .

Remarque: Si un individu i avait le pouvoir d'imposer sa volonté aux autres (quitte, au besoin, à la faire passer pour la "volonté générale"), il choisirait tout simplement l'issue x qui maximise $U_i(x)$ (c'est-à-dire son gain).

2.1. OPTIMUM DE PARETO

Un premier critère qui vient à l'esprit, et qui est dû au sociologue italien Vilfredo Pareto, est celui de l'optimalité qui porte son nom (à ne pas confondre avec la "loi de Pareto" concept complètement empirique en économie selon lequel la plupart des répartitions se font dans un rapport 20/80% (cf. chapitre de Techniques De Gestion)).

Considérons deux issues x et y , appartenant toutes deux à I , et supposons que, pour chaque individu i , nous ayons la situation suivante:

$$\forall i, U_i(y) \geq U_i(x) \quad (65.26)$$

En d'autres termes, aucun individu ne serait a priori lésé si nous substituions pour chacun l'état y à l'état x . Supposons de surcroît, qu'il existe au moins une personne j qui préfère strictement y à x tel que:

$$U_j(y) > U_j(x) \quad (65.27)$$

Dans ces conditions, nous ne voyons plus vraiment ce qui devrait retenir le législateur de choisir y plutôt que x .

Définition: Une issue i réalisable qui n'admet aucune "amélioration" est appelée un "**optimum de Pareto**" (O.P.) et est définie rigoureusement par:

$$\boxed{\forall j \in I \setminus \{i\}, \exists n \in N \quad U_n(j) < U_n(i)} \quad (65.28)$$

La "**pareto-optimalité**" est à comprendre comme une condition sine qua non, un "minimum minimorum", sans lequel le concept de solution d'un jeu coopératif que nous cherchons à élaborer devrait être automatiquement rejeté.

Remarque: Ce résultat forme rejoint donc ce que nous avons déjà écrit en début de chapitre. C'est-à-dire que si dans un jeu, un couple d'issues est tel qu'il est impossible d'améliorer le score de l'un des deux joueurs sans diminuer le score de l'autre, nous disons que ces issues sont "**pareto-optimales**" ou "**pareto-efficientes**".

3. ÉQUILIBRE DE NASH

Définition: "**L'équilibre de Nash**" (ou "équilibre" tout court) décrit donc une issue d'un jeu dans lequel aucun joueur n'a intérêt à modifier sa stratégie unilatéralement, compte tenu des stratégies des autres joueurs.

Remarque: Nous avons déjà vu de nombreux exemples avec des équilibres précédemment.

Soit un jeu à n joueurs, et:

$$s^* = (s_1^*, \dots, s_n^*) \quad (65.29)$$

une combinaison de choix stratégiques de ces n joueurs où s_i^* est le meilleur choix stratégique du joueur i et avec $s_i^* \in S_i$, l'ensemble des stratégies praticables par le joueur i . Soit $U_i(s_1^*, \dots, s_n^*)$ le gain

du joueur i lorsque s_i^* est sélectionné.

Une combinaison de choix stratégiques s^* est un équilibre de Nash si et seulement si:

$$U_i(s_1^*, \dots, s_i^*, \dots, s_n^*) \geq U_i(s_1^*, \dots, s_i, \dots, s_n^*) \quad (65.30)$$

pour tout s_i dans S_i et tout i .

Interprétation: Aucun joueur ne peut tirer un bénéfice d'une déviation de s_i^* , quelle que soit la stratégie qu'il choisisse dans son ensemble S_i . Ainsi, aucun joueur n'a intérêt à dévier, et s_i^* participe à un équilibre.

Remarque: Il peut arriver qu'un optimum de Pareto se confonde avec l'équilibre de Nash mais ce n'est pas toujours le cas (donc un équilibre de Nash n'est pas toujours un optimum de Pareto).

Définition: Quand la stratégie d'un joueur est la meilleure réponse face à toutes les stratégies possibles de ses rivaux, nous parlons alors de "**stratégie dominante**" S_i (cette stratégie domine toutes les autres stratégies du joueur). L'équilibre de ce jeu est alors appelé "**équilibre en stratégie dominante**".

In extenso, une stratégie est "**dominée**" si elle procure au joueur des gains toujours inférieurs à ceux associés à au moins une autre de ses stratégies.

Remarque: Nous pouvons nous interroger si dans un jeu non-coopératif l'équilibre de Nash (s'il existe) n'est pas tel qu'il amène de toute façon à une coopération implicite? Au fait, ce n'est pas le cas (et c'est un résultat très important) comme nous le verrons dans l'étude du fameux dilemme du prisonnier un jeu dont l'équilibre de Nash est assuré par des choix individualistes et rationnels tels qu'ils soient non coopératifs !!! Ce sera donc un exemple extrêmement important dans le cadre de l'économie de marché.

Méthode: Une manière de déterminer les équilibres d'un jeu consiste à éliminer en premier toutes les stratégies dominées puis à rechercher les équilibres dans le jeu ainsi réduit.

Exemple:

En éliminant les stratégies dominées (mêmes faiblement dominées) pour chacun des joueurs, nous tombons sur (6, 4) qui est comme nous le voyons un équilibre de Nash (car c'est celle où aucun joueur n'a intérêt à changer de stratégie).

J_1 / J_2	S_1	S_2	S_3
S_1	5, 2	4, 4	6, 4
S_2	3, 1	2, 0	5, 2

Tableau: 65.13 - Matrice avec équilibre de Nash

Le jeu suivant par contre, ne comporte pas d'équilibre de Nash. Effectivement, quel que soit le couple de stratégies envisagé, l'un des joueurs obtient toujours plus en modifiant son choix.

J_1 / J_2	S_1	S_2
S_1	1 , 0	0 , 1
S_2	0 , 1	1 , 0

Tableau: 65.14 - Matrice sans équilibre de Nash

Toutefois, pour le moment il apparaît pour le moins prématuré de prescrire aux joueurs le choix d'un équilibre; certes s'il est choisi, la situation a une certaine stabilité, mais il reste trois difficultés:

1. Nous ne sommes pas assurés de l'existence d'un couple de tactiques en équilibre (conjonction des tactiques prudentes)
2. Même en cas d'existence, nous ne sommes pas assurés de l'unicité d'un couple de tactiques en équilibre
3. Même en cas d'existence et d'unicité, nous pouvons prescrire un autre choix (!!!!)

4. UTILITÉ ESPÉRÉE

Soit le jeu non-coopératif à somme nulle suivant:

J_1 / J_2	S_1	S_2
S_1	0	2
S_2	3	1

Tableau: 65.15 - Demi-matrice d'un jeu non-coopératif à somme nulle

qui ne comporte pas d'équilibre comme nous l'avons vu plus haut. Dans ce genre de jeu, toute recommandation à un joueur de choisir une tactique plutôt qu'une autre peut lui nuire, dès lors que l'adversaire en est informé, ou peut deviner cette recommandation.

Effectivement, si J_1 pense que J_2 va choisir sa tactique 1, il a intérêt à choisir sa tactique 2 (utilité 3 contre 0). Mais alors, si J_2 pense que J_1 va choisir sa tactique 2, il a intérêt à choisir sa tactique 2 (perte 1 au lieu de 3). Alors, si J_1 pense que J_2 va choisir sa tactique 2, il a intérêt à choisir sa tactique 1 (utilité 2 contre 1). Mais alors, si J_2 pense que J_1 va choisir sa tactique 1, il a intérêt à choisir sa tactique 1 (perte 0 au lieu de 3). Et la boucle est bouclée...

En définitive, la chose qui importe avant tout dans un jeu non-coopératif c'est que la tactique d'un joueur ne puisse pas être devinée par son adversaire. Comme tout raisonnement pourrait être percé à jour, les adversaires étant parfaitement rationnels et informés, la seule solution imaginable est de s'en remettre à un processus précis, appuyé sur des probabilités affectées aux diverses tactiques possibles. Ainsi, comme nous l'avons défini plus haut, le jeu comporte un aspect à "stratégie mixte".

Naturellement, le résultat obtenu par le joueur ne peut pas être garanti de façon certaine, puisque le processus de choix de la décision fait intervenir des probabilités. Comparer des résultats revient donc à comparer des loteries. Nous imaginons la situation d'un amiral devant répondre devant un tribunal militaire de la perte d'un navire, et expliquant qu'il a pris sa décision en jouant aux dés (en supposant

une bataille sans équilibre de Nash et non-coopérative): même si parfaitement conforme aux prescriptions de la théorie des jeux, cette explication aura peine à convaincre!

4.1. CRITÈRE DE HURWITZ

Il nous faut donc introduire une utilité probabiliste (appelée aussi parfois le "**critère de Hurwitz**"). Considérons un jeu à deux stratégies propres S_1, S_2 et notons l'utilité respective:

$$U[S_1P, S_2(1-P)] \quad (65.31)$$

qui permet d'obtenir S_1 avec une probabilité P et S_2 avec une probabilité $1-P$. Cette relation s'écrit avec des notations évidentes (cf. chapitre de Probabilités):

$$E = U[S_1P, S_2(1-P)] = P \cdot U(S_1) + (1-P)U(S_2) \quad (65.32)$$

avec E que nous appellerons "**l'utilité espérée**" (en similitude avec le concept d'espérance vu dans le chapitre de Statistiques) ou "**espérance de gain anticipée**".

Nous pouvons déjà noter que, s'il existe une telle utilité (espérée), il en existe une infinité à un arbitraire près, obtenues à partir de U par une transformation affine strictement croissante, c'est-à-dire une relation de la forme:

$$V = aU + b \text{ avec } a > 0 \quad (65.33)$$

En effet, la relation:

$$U[S_1P, S_2(1-P)] = P \cdot U(S_1) + (1-P)U(S_2) \quad (65.34)$$

entraîne pour $a > 0$:

$$aU[S_1P, S_2(1-P)] = P \cdot aU(S_1) + (1-P)aU(S_2) \quad (65.35)$$

qui, additionnée terme à terme à la relation évidente (nécessaire):

$$b = Pb + (1-P)b \quad (65.36)$$

conduit bien à:

$$V[S_1P, S_2(1-P)] = P \cdot V(S_1) + (1-P)V(S_2) \quad (65.37)$$

Cela prouve entre autres ce que nous avons énoncé plus haut: nous pouvons toujours choisir une fonction d'utilité (et ce même dans une optique de stratégie pure où $P = 0$ ou $P = 1$) telle que les deltas des gains de joueurs dans les jeux à somme nulle soient égaux et opposés.

Remarque: L'utilité espérée (ou "critère de Hurwitz") se confond avec le critère du maximin lorsque $P = 0$ et du maximax lorsque $P = 1$ (voir plus loin).

Voyons de suite un exemple en considérant le jeu à somme nulle suivant:

J_1 / J_2	b_1	b_2
a_1	5	2
a_2	3	4

Tableau: 65.16 - Matrice d'un jeu à somme nulle

Nous voyons dans ce jeu qu'il n'y a pas d'équilibre de Nash (et donc pas de col). Effectivement, si J_2 pense que J_1 va décider a_1 , il a intérêt à choisir b_2 (perte de 2 au lieu de 5). Mais J_1 comprenant cela, va changer pour a_2 (gain de 4 au lieu de 2). Mais J_2 devinant cela va changer pour b_1 (perte de 3 au lieu de 4), et J_1 qui a tout compris va revenir à a_1 (gain de 5 au lieu de 3).

Considérons maintenant que le joueur J_1 va choisir un nombre compris entre 0 et 1, soit x , et prendra les décisions a_1 avec la probabilité x et a_2 avec la probabilité $1-x$. De même, le joueur J_2 va choisir un nombre compris entre 0 et 1, soit y , et prendra les décisions b_1 avec la probabilité y et b_2 avec la probabilité $1-y$.

Les résultats de ces décisions conjointes sont alors:

- 5, résultant de la conjonction de a_1, b_1 obtenu avec la probabilité xy (les décisions des deux joueurs étant indépendantes !)
- 2, obtenu avec la probabilité $x(1-y)$
- 3, obtenu avec la probabilité $y(1-x)$
- 4, obtenu avec la probabilité $(1-x)(1-y)$

L'espérance de J_1 est donc:

$$E(J_1) = 5xy + 2x(1-y) + 3y(1-x) + 4(1-x)(1-y) = x(4y-2) + 4-y \quad (65.38)$$

Remarque: Nous voyons bien que si $x=0$ (et $y=1$) alors nous tombons sur le critère du Minimax (le gain maximum des stratégies les plus pessimistes) soit $E(J_1)$ égal à 3. De même si $x=1$ (et $y=1$) alors nous tombons sur le critère du Maximax (le gain maximum des stratégies les plus optimistes).

S'il y a équilibre entre les stratégies probabilistes, J_1 n'aura aucune raison de modifier la valeur de x dans l'espoir d'augmenter $E(J_1)$. Dès lors, la dérivée par rapport à x doit être nulle telle que (maxima):

$$\frac{dE(J_1)}{dx} = 4y - 2 = 0 \Rightarrow y = \frac{1}{2} \quad (65.39)$$

Dans ces conditions:

$$E(J_1) = \frac{7}{2} \quad (65.40)$$

Pour examiner ce qui s'offre à J_2 , dont l'espérance, rappelons-le, sera dans un jeu à somme nulle nécessairement opposée à celle de J_1 , nous écrivons:

$$E(J_2) = -E(J_1) = -x(4y - 2) - 4 + y = y(1 - 4x) - 4 + 2x \quad (65.41)$$

En appliquant le même raisonnement (mais implicitement en minima):

$$\frac{dE(J_2)}{dy} = 1 - 4x = 0 \Rightarrow x = \frac{1}{4} \quad (65.42)$$

Dans ce cas:

$$E(J_2) = -\frac{7}{2} \quad (65.43)$$

Ainsi, nous avons déterminé les probabilités des stratégies qui maximisent l'espérance des gains de ce jeu non-coopératif ! En les adoptants J_1 est certain d'une espérance au moins égale à $7/2$ (puisque J_2 n'a rien à gagner à modifier sa stratégie) et J_2 est certain d'une espérance au moins égale à $-7/2$. Le nombre $7/2$ (valeur absolue) est la "**valeur du jeu**".

Définition: Si la valeur du jeu d'un jeu non-coopératif à stratégie mixte est égale pour les deux joueurs, nous disons alors qu'il s'agit d'un "**équilibre en stratégie mixte**" (aucun des joueurs n'a intérêt à dévier unilatéralement).

Ce résultat est certainement le plus remarquable jusque-là sur ce chapitre, car les jeux non-coopératifs sont les plus nombreux sur le marché.

4.2. CRITÈRE DE LAPLACE

Le critère de Laplace est un critère qui affecte la même probabilité, en l'absence d'information, pour chaque décision (équiprobabilité). Il s'agira de calculer une espérance de gain pour chaque décision compte tenu de la probabilité affectée.

Autrement dit, le critère de Laplace consiste à déterminer pour chaque projet l'espérance mathématique en affectant la même probabilité à chaque état de la nature et retenant celui dont l'espérance est la plus élevée.

Voyons de suite un exemple en considérant à nouveau le jeu de somme nulle suivant:

J_1 / J_2	b_1	b_2
a_1	5	2
a_2	3	4

Tableau: 65.17 - Matrice d'un jeu à somme nulle

En appliquant l'équiprobabilité, nous avons le tableau suivant:

J_1 / J_2	$E(b_1)$	$E(b_2)$
$E(a_1)$	$5/2+2/2=3.5, 5/2+3/2=4$	$5/2+2/2=3.5, 2/2+4/2=3$
$E(a_2)$	$3/2+4/2=3.5, 5/2+3/2=4$	$3/2+4/2=3.5, 2/2+4/2=3$

Tableau: 65.18 - Application des probabilités dans la matrice

Le jeu devient alors:

J_1 / J_2	$E(b_1)$	$E(b_2)$
$E(a_1)$	3.5 , 4	3.5 , 3
$E(a_2)$	3.5 , 4	3.5 , 3

Tableau: 65.19 - Calcul des espérances

Dans cet exemple, où l'espérance est toujours égale pour le joueur J_1 quelle que soit sa stratégie, le joueur 2 choisira la stratégie où l'espérance de sa perte est la plus faible soit b_2 . Nous avons donc ici un équilibre de Nash (sans optimum de Pareto).

Remarque: Les techniques présentées ci-dessus ne sont pas exhaustives. Les entreprises utilisent par exemple beaucoup les études statistiques de R&R (répétabilité & reproductibilité) par attributs pour analyser statistiquement avec intervalles de confiance (utilisant le test binomial exact démontré dans le chapitre de Statistiques) et sur la base d'un indicateur appelé "**Kappa de Cohen**" (voir mon livre sur Minitab) les décisions de gestionnaires ou d'employés par rapport à celles d'un expert de référence.

5. JEUX ÉVOLUTIONNAIRES

Les stratégies de l'évolution biologique, comme nous en avons fait mention au début de ce chapitre, peuvent être modélisées à l'aide de la théorie des jeux. Dans ce cadre, le biologiste est amené à définir des relations remarquables définissant une stratégie d'évolution donnée (dominance, stagnation, suicide).

Définition: Une "**stratégie évolutionnaire stable (SES)**" (ou "evolutionary stable strategy" ESS) est une stratégie adoptée par la majorité et empêchant qu'une population soit envahie par un mutant qui recourrait à une stratégie différente.

Cette stratégie s'écrit sous la forme d'une condition de stabilité telle que soient deux stratégies S_1, S_2 de deux joueurs nous ayons:

$$U(S_1, S_1) > U(S_2, S_1) \quad (65.44)$$

ou (si cette dernière n'apparaît pas) par la simultanéité des deux stratégies de non-sélection et suicide:

$$U(S_1, S_1) = U(S_2, S_1) \text{ et } U(S_1, S_2) > U(S_2, S_2) \quad (65.45)$$

- La première relation signifie qu'en aucun cas un individu n'a à changer de stratégie pour se défendre contre une évolution mutante ayant la même stratégie que lui, car toute autre lui serait défavorable.
- La deuxième relation signifie que quelle que soit la stratégie adoptée contre une stratégie mutante, il y aura stagnation.
- La troisième relation signifie que contre S_2 , toute stratégie différente de S_2 est préférable pour contrer S_2 même. Autrement dit, appliquer une stratégie S_2 différente de S_1 est suicidaire (le cas contraire ne l'est donc pas!).

Exemple:

Voyons un jeu connu Faucons (Hawk) contre Colombes (Dove).

Ce jeu vise à modéliser les rapports entre individus en compétition pour une ressource rare, c'est-à-dire dont le degré d'adaptation va être modifié à la fois par l'obtention de cette ressource et par les violences qu'ils subiront ou infligeront pour l'obtenir.

Dans leurs interactions compétitives, les organismes recourent à deux types de stratégies (comportements): la stratégie du faucon et celle de la colombe. Le faucon intensifie le conflit jusqu'à ce qu'il soit blessé ou jusqu'à ce que l'autre batte en retraite. La colombe se retire après une première démonstration de force si l'adversaire choisit d'intensifier le conflit. Lorsque deux faucons se rencontrent, l'un est blessé et l'autre emporte la ressource. Si un faucon et une colombe s'affrontent, le faucon s'empare de la ressource sans danger d'être blessé et la colombe n'obtient ni avantage ni dommage. Enfin, deux colombes se partagent à part égale la ressource.

Nous posons également les hypothèses suivantes:

H1. Les affrontements se déroulent un à un

H2. La population est infinie

H3. Les rencontres sont aléatoires

H4. Les combats sont symétriques (au sens où ni l'âge, ni la taille, ni l'expérience n'influent sur l'issue du combat)

H5. Il est impossible de savoir avant le début d'un conflit quelle stratégie un animal adoptera.

Sur la base de ces règles d'interaction (assez loin de la réalité...), il est possible de construire le tableau du jeu qui nous permettra de calculer les avantages ou les désavantages des diverses stratégies selon les circonstances.

Ainsi le tableau de jeu est le suivant:

J_1 / J_2	H	D
H	$(V-C)/2, (V-C)/2$	$V, 0$
D	$0, V$	$V/2, V/2$

Tableau: 65.20 - Matrice d'un jeu évolutionnaire

Comme il s'agit d'un jeu à somme nulle, nous pouvons le simplifier:

J_1 / J_2	H	D
H	$(V-C)/2$	V
D	0	$V/2$

Tableau: 65.21 - Demi-matrice simplifiée du jeu évolutionnaire

Nous notons ici V l'avantage qu'un organisme retire de l'obtention de la ressource. V désigne non la ressource elle-même, mais l'accroissement du degré d'adaptation qu'elle procure à l'organisme qui l'obtient. C correspond au coût payé, mise en danger ou blessure, pour acquérir la ressource.

D'abord explicitons la manière dont il faut lire ce tableau:

1. Pour la stratégie (D, D) - tout le monde est gentil avec tout le monde - le gain total des deux individus est:

$$\frac{V}{2} + \frac{V}{2} = V \quad (65.46)$$

La population restera donc constante (c'est la stagnation).

Bref, de par leurs comportements les colombes se partagent à l'amiable la valeur de la ressource.

2. Pour les stratégies $(H, D), (D, H)$ les "colombes" D sont toujours perdantes (elles ne progressent pas dans leur évolution). Leur gain est nul alors que les "faucons" auront éliminé le nombre V de colombes (d'où le gain).

3. Pour la stratégie (H, H) les "faucons" supportent une perte du type $(V - C)/2$ où C est une constante et telle que la somme des gains des faucons est normalement inférieure ou égale à V . Autrement dit:

$$\frac{(V - C)}{2} + \frac{(V - C)}{2} = V - C \leq V \quad (65.47)$$

Bref, lorsqu'un faucon en affronte un autre, il obtient en moyenne une fois sur deux la valeur de la ressource diminuée du prix encouru pour l'obtenir.

Remarque: Ce jeu peut être vu comme un jeu de guerre entre deux joueurs... l'interprétation des résultats est dès lors plus que pertinente.

Nous devons maintenant considérer deux stratégies:

1. L'étude du jeu de manière globale en stratégie pure (sans probabilités donc)
2. L'étude du jeu de manière globale en stratégie mixte (faisant intervenir les probabilités)

Commençons par la première en considérant 3 cas de figure:

$$V < C, V = C, V > C \quad (65.48)$$

- Si $V > C$ (avantage plus élevé que le coût de la stratégie), en choisissant (H, H) (qui est dès lors l'équilibre de Nash strict du jeu) nous pouvons observer que le jeu sera du type évolutionnaire stable (SES). Effectivement, nous retrouvons la relation:

$$\left[U(H, H) = \frac{(V - C)}{2} \right] > [U(D, H) = 0] \quad (65.49)$$

correspondant bien à:

$$U(S_1, S_1) > U(S_2, S_1) \quad (65.50)$$

et nous pouvons aussi observer qu'il existe aussi une stratégie faiblement dominante (pas de sélection naturelle) dans:

$$[U(H, D) = V] > \left[U(D, D) = \frac{V}{2} \right] \quad (65.51)$$

correspondant bien à:

$$U(S_1, S_2) > U(S_2, S_2) \quad (65.52)$$

Mais celle-ci ne sera pas adoptée puisque moins forte que l'équilibre de Nash.

- Si $V = C$ (avantage égal au coût), le jeu est aussi du type SES. Effectivement, (H, H) devient une stratégie faiblement dominante:

$$[U(H, H) = 0] = [U(D, H) = 0] \quad (65.53)$$

correspondant bien à $U(S_1, S_1) = U(S_2, S_1)$ et il n'y a dès lors pas d'évolution et nous pouvons aussi observer qu'il y a aussi:

$$[U(H, D) = V] > \left[U(D, D) = \frac{V}{2} \right] \quad (65.54)$$

correspondant à:

$$U(S_1, S_2) > U(S_2, S_2) \quad (65.55)$$

Puisque nous avons simultanément:

$$U(S_1, S_1) = U(S_2, S_1) \text{ et } U(S_1, S_2) > U(S_2, S_2) \quad (65.56)$$

lorsque $V = C$ le jeu est une SES.

- Si $V < C$ (avantage inférieur au coût), ni H , ni D ne sont des stratégies dominantes et nous n'avons pas de SES:

$$[U(D, H) = 0] > \left[U(H, H) = \frac{(V - C)}{2} \right] \quad (65.57)$$

et:

$$[U(H, D) = V] > \left[U(D, D) = \frac{V}{2} \right] \quad (65.58)$$

ces deux dernières relations correspondant toutes 2 à:

$$U(S_1, S_2) > U(S_2, S_2) \quad (65.59)$$

C'est plutôt embêtant... c'est une sorte de suicide collectif...

Remarque: Ces deux dernières relations nous amènent à observer que les faucons ne voudront pas forcément révéler aux autres faucons leur stratégie de prédateurs de colombes, puisque: toute stratégie vaut mieux être contrée par une autre stratégie plutôt que par elle-même. Ils préfèrent peut-être discuter entre eux ce qui amène au fait que le jeu est dès lors coopératif.

Cherchons maintenant à l'aide de l'étude en stratégie mixte ce que nous pourrions faire pour amener la dernière configuration précédente à un ESS (donc relativement à la dernière configuration $V < C$ afin de voir de plus près ce que nous pouvons faire pour éviter cela):

Nous considérons une population d'individus qui jouent donc une stratégie mixte que nous noterons pour chacun:

$$\sigma(P, 1 - P) \quad (65.60)$$

avec $P = P(H)$ et:

$$1 - P = P(D) \quad (65.61)$$

Si $P = 1$ (stratégie pure), nous aurons dès lors:

$$\sigma(P, 1 - P) = \sigma(1, 0) = \sigma_H \quad (65.62)$$

Reprenons maintenant l'étude des trois cas de figure:

$$V < C, V = C, V > C \quad (65.63)$$

1. Si $V > C$ avec σ_H et $0 \leq P < 1$ nous avons toujours:

$$U(\sigma_H, \sigma_H) > U(\sigma, \sigma_H) \quad (65.64)$$

en d'autres termes, la stratégie sera toujours du type évolutionnaire stable (SES) si σ_H est une stratégie pure et ce même si σ peut varier et s'approcher de σ_H .

2. Si $V = C$ avec σ_H et $0 \leq P < 1$ nous avons toujours:

$$U(\sigma_H, \sigma_H) = U(\sigma, \sigma_H) \text{ et } U(\sigma_H, \sigma) > U(\sigma, \sigma) = \frac{V}{2} \quad (65.65)$$

en d'autres termes, la stratégie sera toujours du type non-sélective si σ_H est une stratégie pure et ce même si σ peut varier et s'approcher de σ_H .

3. Si $V < C$ et $0 < P \leq 1$, nous laissons tomber σ_H pour ne nous intéresser qu'à la généralisation $\sigma(P, 1-P)$. Nous avons alors:

$$U(\sigma, D) > U(D, D) \quad (65.66)$$

Effectivement:

$$\left[U(\sigma, D) = PV + (1-P)\frac{V}{2} = (1+P)\frac{V}{2} \right] > \left[U(D, D) = \frac{V}{2} \right] \quad (65.67)$$

De même:

$$U(\sigma, H) > U(H, H) \quad (65.68)$$

Effectivement:

$$\left[U(\sigma, H) = P\frac{(V-C)}{2} + (1-P)0 = P\frac{(V-C)}{2} \right] > \left[U(H, H) = \frac{(V-C)}{2} \right] \quad (65.69)$$

Et nous aimerions arriver à une SES en stratégie mixte. Cela est-il possible ?

Dans le cadre d'une stratégie mixte, nous avons démontré lors de l'étude d'un jeu à somme nulle que l'équilibre mixte était donné par:

$$E(J_1) = -E(J_2) \quad (65.70)$$

Il est donc assez évident que pour un jeu qui n'est pas à somme nulle, nous ayons l'équilibre mixte qui soit donné par:

$$E(J_1) = E(J_2) \quad (65.71)$$

Dès lors, cherchons la relation entre P , C , V telle que cet équilibre soit atteint:

$$\begin{aligned}
 &\underbrace{E(J_1) = E(J_2)} \\
 &\underbrace{E(H) = E(D)} \\
 &\underbrace{U(H, \sigma) = U(D, \sigma) \Leftrightarrow U(S_1, S_1) = U(S_2, S_1)} \\
 &P \cdot U(H, H) + (1 - P)U(H, D) = P \cdot U(D, H) + (1 - P) \cdot U(D, D)
 \end{aligned} \tag{65.72}$$

En connaissant les utilités:

$$P \cdot \frac{(V - C)}{2} + (1 - P)V = P \cdot 0 + (1 - P) \cdot \frac{V}{2} \tag{65.73}$$

d'où nous tirons que l'équilibre en stratégie mixte est donné par:

$$P = \frac{V}{C} \tag{65.74}$$

et donc que l'équilibre est donné par la stratégie mixte:

$$\sigma = \left(\frac{V}{C}, 1 - \frac{V}{C} \right) \tag{65.75}$$

À quoi cette stratégie va-t-elle mener ? Eh bien simplement dans le cas suicidaire cette stratégie mixte est la meilleure réponse contre elle-même (c'est ce qu'il est possible de faire de mieux dans ce qu'il y de pire) car elle conduit aux deux conditions qui satisfont une SES.

6. ÉQUILIBRE DE COURNOT

Imaginons deux propriétaires M et N , et deux sources dont les qualités sont identiques et qui se trouvent placées de manière à alimenter concurremment le même marché de sorte que la quantité totale livrée aux commerces se compose de la somme des quantités m , n livrées par chacun des propriétaires à un prix qui est nécessairement le même pour chacun d'eux puisqu'il n'y a aucun motif de préférer une source à l'autre. Ce prix se trouve déterminé quand la somme des quantités m , n l'est elle-même, à cause de la liaison qui existe entre le prix et la demande. Admettons que le propriétaire N ait fixé arbitrairement, sans égard au prix, la quantité n qu'il entend livrer: alors le propriétaire M fixera le prix de vente, c'est-à-dire la production totale (composée de la somme des quantités m et n), c'est-à-dire encore sa production m de manière à se procurer le plus grand revenu possible.

Dans la pratique, une suite de tâtonnements et d'oscillations amènera les deux propriétaires à cette position d'équilibre, et la théorie montre que cet équilibre est stable: c'est-à-dire que si l'un ou l'autre des propriétaires, trompé sur ses intérêts véritables, vient à s'en écarter momentanément, il y sera ramené par une suite d'oscillations du genre de celles qui avaient primitivement abouti à constituer l'équilibre.

Nous allons mettre en place une situation de jeu à deux personnes. Nous poserons que le prix P est une fonction affine de la quantité totale Q produite:

$$P(Q) = P(m + n) = \alpha(m + n) \tag{65.76}$$

où α est une constante de normalisation des unités.

Nous supposons égaux et fixes les coûts marginaux de production, représentés par le nombre c , et nuls les coûts fixes, en sorte que le coût de production s'écrive respectivement $c + \alpha \cdot m$ et $c + \alpha \cdot n$ pour les deux sources.

Le modèle de Cournot pose que les deux entreprises fixent les quantités qu'elles produisent simultanément, ou, à tout le moins dans l'ignorance mutuelle de la tactique de l'autre.

Pour reconnaître un jeu sous forme normale, il ne nous reste plus qu'à identifier le gain retiré par chacun des adversaires pour tout couple (m, n) de tactique afin de pouvoir, si on le désire, construire la matrice des gains.

Le profit de M est:

$$\pi_M(m, n) = m[P(m + n) - c] = m[\alpha(m + n) - c] \quad (65.77)$$

et celui de N :

$$\pi_N(m, n) = n[P(m + n) - c] = n[\alpha(m + n) - c] \quad (65.78)$$

La recherche d'un équilibre de Nash conduit chaque entreprise à choisir sa production pour maximiser son profit et minimiser ses coûts de stockage (voir modèle de Wilson dans le chapitre de Techniques de Gestion), la production de son partenaire étant supposée connue.

Dans ce but, on annule la dérivée des deux fonctions précédentes:

$$\begin{aligned} \frac{\partial \pi_M}{\partial m} &= 0 = 2\alpha m + \alpha n - c = 0 \\ \frac{\partial \pi_N}{\partial n} &= 0 = \alpha m + 2\alpha n - c = 0 \end{aligned} \quad (65.79)$$

Système dont la résolution conduit très facilement à la détermination de:

$$m = n = \frac{1}{3} \frac{c}{\alpha} \quad (65.80)$$

(resterait à vérifier que ce sont bien des maximums, en contrôlant les dérivées de deuxième ordre et non des minimums). La situation d'équilibre du duopole intervient donc lorsque chacune des deux firmes produit un tiers du marché potentiel (son profit est alors maximum).

Le prix de vente dans le cadre d'un équilibre de Nash serait alors:

$$P(Q) = \alpha(m + n) = \frac{2}{3} c \quad (65.81)$$

Soit, une valeur inférieure au coût marginal ce qui ne présage évidemment rien de bon! Il vaudrait donc mieux que le prix ne soit pas une fonction affine de la quantité...

De ce dernier résultat nous déduisons que le profit de chaque entreprise sera donné par:

$$\begin{aligned}\pi_M(m, n) &= -\frac{m \cdot c}{3} \\ \pi_N(m, n) &= -\frac{n \cdot c}{3}\end{aligned} \quad (65.82)$$

et est in extenso négatif pour les deux (en cause le modèle de relation supposée affine).

Ces calculs sont à rapprocher du raisonnement purement économique, pour lequel chaque entreprise aimerait être seule, en monopole sur le marché. Le profit de l'entreprise M en situation de monopole serait:

$$\pi(m, 0) = m[\alpha m - c] = \alpha m^2 - mc \quad (65.83)$$

ce qui met en évidence le maximum, atteint pour (on cherche où la dérivée s'annule):

$$m = \frac{1}{2} \frac{c}{\alpha} \quad (65.84)$$

Ainsi, on voit très bien que la quantité produite en cas de monopole est plus grande qu'en cas de duopole (logique!) et que le profit ainsi que les prix sont plus élevés.

Cependant avec notre modèle affine, même si l'entreprise M est seule, son profit sera négatif...

L'idée serait maintenant, si l'on revient à nos deux entreprises, qu'un accord soit établi (cas appelé "**entente oligopole**" contre la concurrence... ce qui est interdit par la loi!), qui leur partage ce profit majoré. La parfaite symétrie des situations conduirait au partage par moitiés. Mais la difficulté vient du fait que la décision de produire:

$$\frac{1}{4} \frac{c}{\alpha} \quad (65.85)$$

n'est pas la meilleure réponse car elle incite à trahir l'accord avec l'autre. Ainsi, le meilleur équilibre est celui de Nash qui impose:

$$m = n = \frac{1}{3} \frac{c}{\alpha} \quad (65.86)$$

Lors de la mise au point d'une entente ou d'un cartel, on peut distinguer plusieurs niveaux qui dépendent du degré de précision des règles fixées par l'ensemble des entreprises.

Le premier cas est celui qu'on peut appeler "**l'entente parfaite**"; c'est l'entente qui permet de maximiser le profit total des entreprises concernées. Une condition mathématique élémentaire est que toutes les entreprises doivent fonctionner avec le même coût marginal. En effet, la maximisation du profit total d'un ensemble d'entreprises s'écrit de la manière suivante:

$$\pi_T(q_1, q_2, \dots, q_n) = Q \cdot P(Q) - \sum_{i=1}^n c_i \cdot q_i \quad (65.87)$$

où rappelons-le :

$$Q = \sum_{i=1}^n q_i \quad (65.88)$$

Ce profit est maximum quand toutes les dérivées partielles d'ordre 1 sont nulles et maximales (condition dites du "premier ordre"). Soit:

$$\frac{\partial \pi_T}{\partial q_i} = 0 \Leftrightarrow \frac{\partial \left(\sum_{i=1}^n q_i \cdot p \left(\sum_{i=1}^n q_i \right) \right)}{\partial q_i} = \frac{\partial \left(\sum_{i=1}^n c_i \cdot q_i \right)}{\partial q_i} \quad (65.89)$$

La partie de gauche de la deuxième relation exprime la variation de recette totale provoquée par une petite variation de la quantité produite par le producteur i , et la partie droite exprime la variation de coût engendrée par la même variation de q_i (coût marginal du producteur i). La recette marginale provoquée par une variation donnée de production q est la même, quel que soit le producteur qui a modifié sa production. En effet, l'influence d'une production additionnelle sur l'offre totale et sur le prix est identique, que cette production additionnelle vienne d'un producteur ou d'un autre.

Mais comme on l'a vu dans le duopole de Carnot, ce type d'égalité admet un profit total maximum à condition que toutes les entreprises de l'entente aient leur coût marginal au même niveau, correspondant à la recette marginale du marché. Cette condition d'égalité n'est certainement pas une condition facile à remplir dans la réalité des ententes...

7. CHÂÎNES DE MARKOV

Comme nous l'avons mentionné dans le chapitre de Probabilités, les chaînes de Markov sont aussi utilisées dans le domaine des techniques de décision. Nous disons alors que nous faisons une "**analyse de Markov**" (dénomination que l'on retrouve dans la norme ISO 31010). Le cas d'utilisation le plus fréquent des chaînes de Markov dans l'industrie est le milieu de la pharma-économie (analyse des coûts de traitements des malades) qui passe avant, d'après notre expérience, le domaine de la finance et de l'ingénierie.

Par exemple, la chaîne de Markov simple suivante relative à un symptôme particulier, a une étape dite "**état absorbant**" connue par tous dont il n'est pas possible de réchapper à ce jour...:

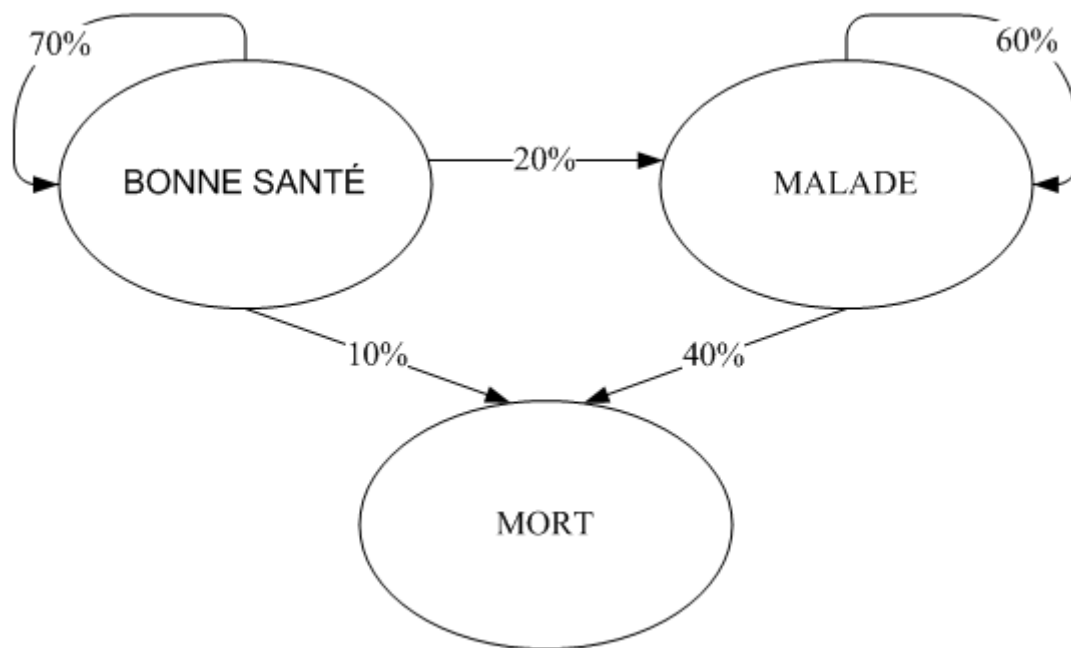


Figure: 65.11 - Chaîne de Markov avec état absorbant

Dont voici la chaîne de Markov décomposée sur 20 cycles comme le présente traditionnellement le domaine médical (cela suppose que les probabilités ne changent pas au cours du temps...):

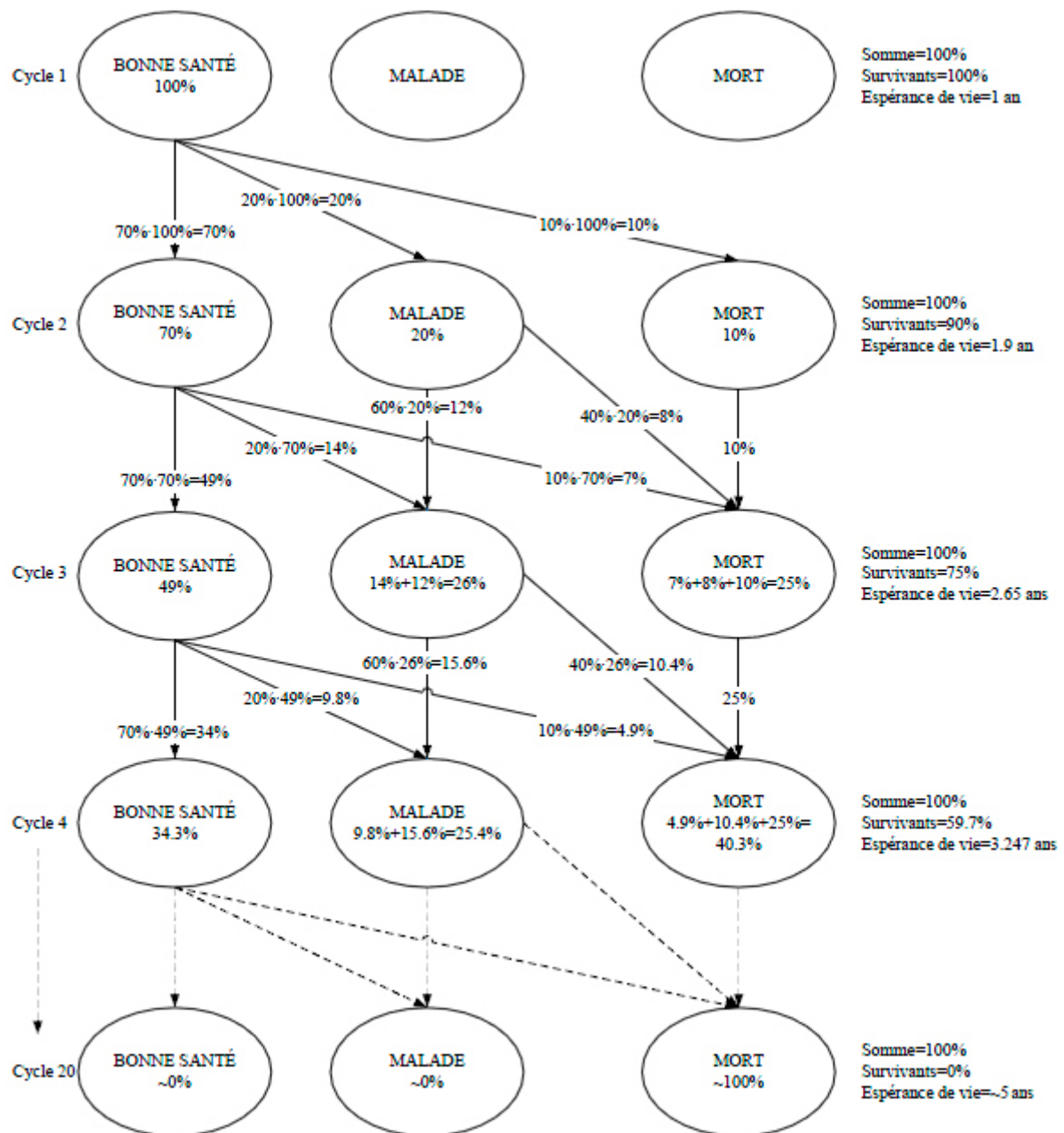


Figure: 65.12 - Chaîne de Markov avec cycles

Évidemment il est possible d'effectuer le même calcul directement sous forme matricielle... à nouveau la matrice de transition (matrice stochastique) est simple à identifier. Il s'agit de (cf. chapitre de Probabilités):

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0 & 0 & 1 \end{bmatrix} \quad (65.90)$$

Le vecteur de probabilités initiales $p(0)$ vaut bien évidemment dans les cas les plus courants de maladie...:

$$p(0) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad (65.91)$$

À chaque fois que nous multiplions la transposée de la matrice de transition par le vecteur de probabilités initiales, nous obtenons donc la probabilité d'être dans un état donné, à un cycle donné!

$$p(n+1) = P^T p(n) \quad (65.92)$$

Avec la version anglaise de Microsoft Excel 11.8346, la modélisation est assez simple à reproduire:

	A	B	C	D	E	F	G	H	I
1		Matrice Stochastique Transposée							
2		0.7	0	0					
3		0.2	0.6	0					
4		0.1	0.4	1					
5									
6		$p(0)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$	$p(6)$	$p(7)$
7		1.00	0.70	0.49	0.34	0.24	0.17	0.12	0.08
8		0.00	0.20	0.26	0.25	0.22	0.18	0.14	0.11
9		0.00	0.10	0.25	0.40	0.54	0.65	0.74	0.81
10	Espérance de vie:	1.00	0.90	0.75	0.60	0.46	0.35	0.26	0.19
11	Somme espérance de vie:	1.00	1.90	2.65	3.25	3.71	4.06	4.32	4.51

Figure: 65.13 - Calculs des vecteurs

Soit avec les formules:

	A	B	C	D
1		Matrice St		
2		0.7	0	0
3		0.2	0.6	0
4		0.1	0.4	1
5				
6		$p(0)$	$p(1)$	$p(2)$
7		1	=MMULT(\$B\$2:\$D\$4;B7:B9)	=MMULT(\$B\$2:\$D\$4;C7:C9)
8		0	=MMULT(\$B\$2:\$D\$4;B7:B9)	=MMULT(\$B\$2:\$D\$4;C7:C9)
9		0	=MMULT(\$B\$2:\$D\$4;B7:B9)	=MMULT(\$B\$2:\$D\$4;C7:C9)
10	Espérance	=B7+B8	=C7+C8	=D7+D8
11	Somme e	=B10	=B11+C10	=C11+D10

Figure: 65.14 - Calculs explicites des vecteurs

Si nous continuons ainsi jusqu'au 20ème cycle:

$p(16)$	$p(17)$	$p(18)$	$p(19)$	$p(20)$
0.00	0.00	0.00	0.00	0.00
0.01	0.00	0.00	0.00	0.00
0.99	0.99	1.00	1.00	1.00
0.01	0.01	0.00	0.00	0.00
4.98	4.98	4.99	4.99	4.99

Figure: 65.15 - Suite (...)

Il est très courant dans les entreprises de synthétiser l'évolution sous forme graphique (plus parlant pour la direction...):

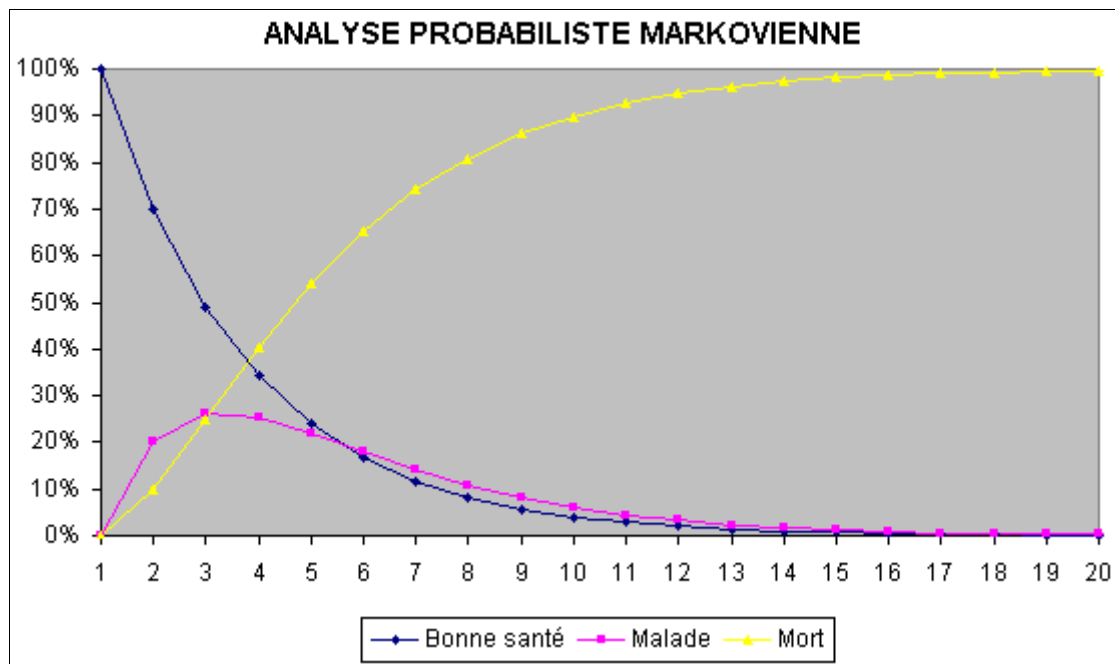
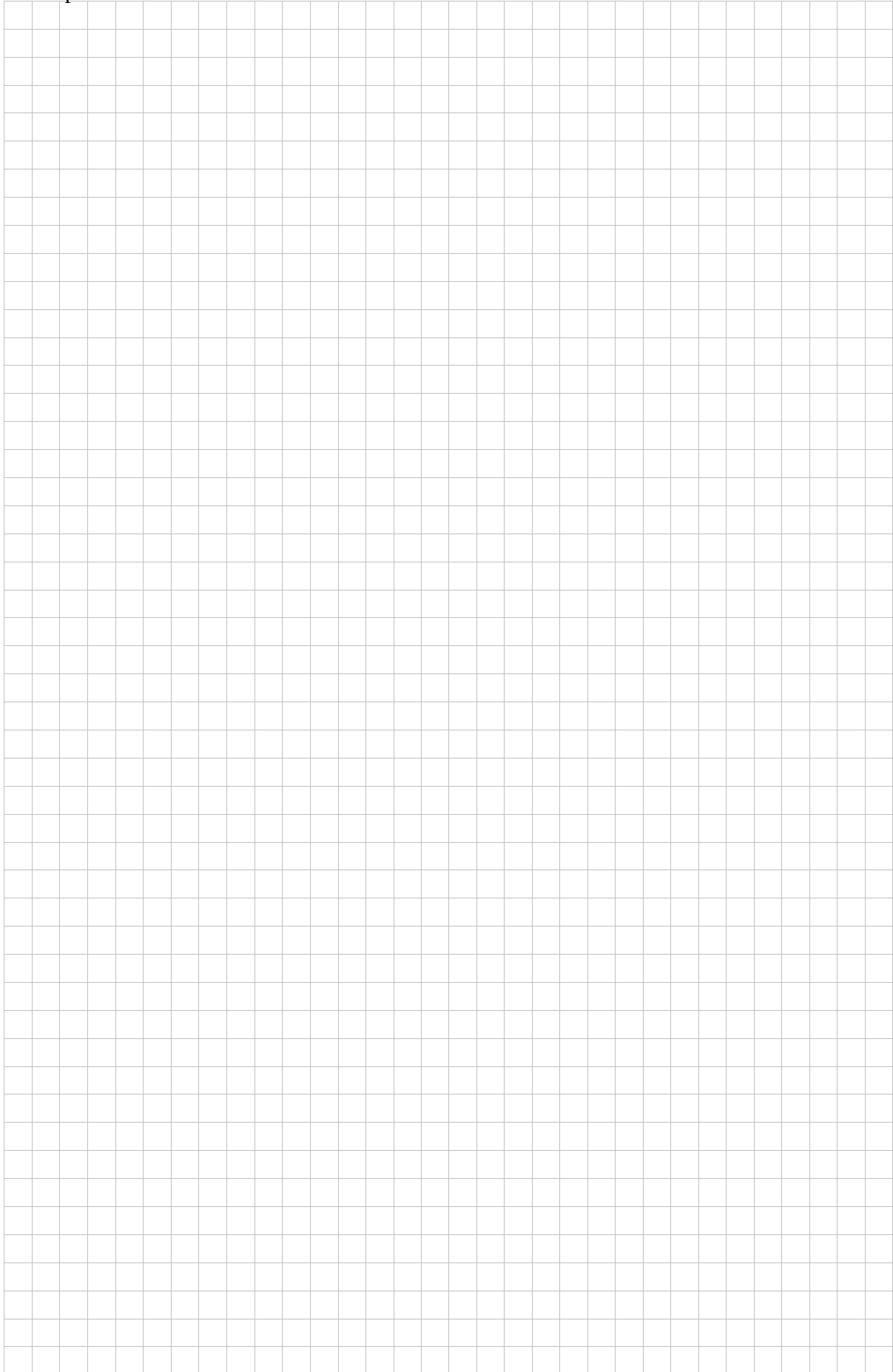


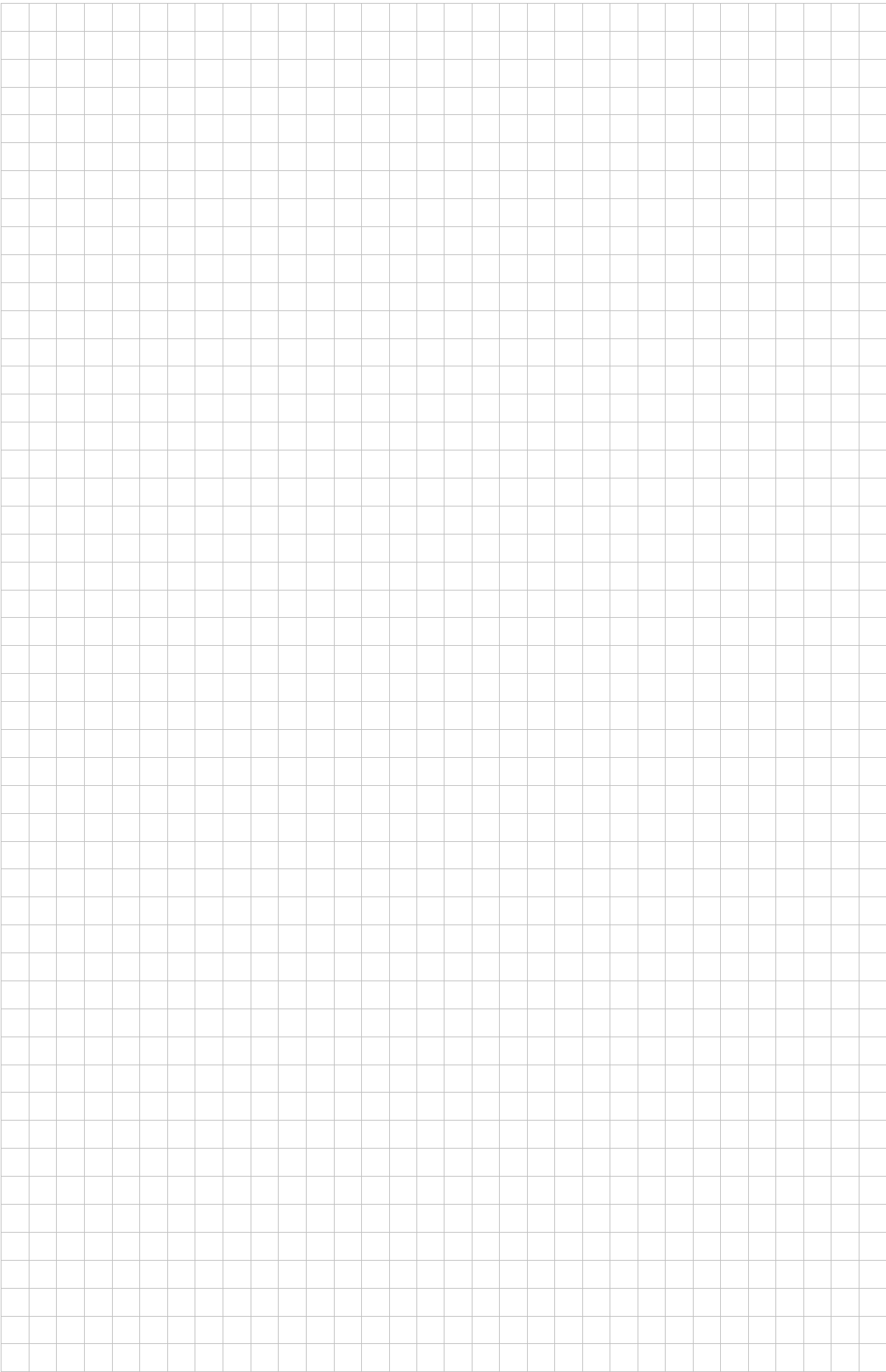
Figure: 65.16 - Évolution de la cohorte

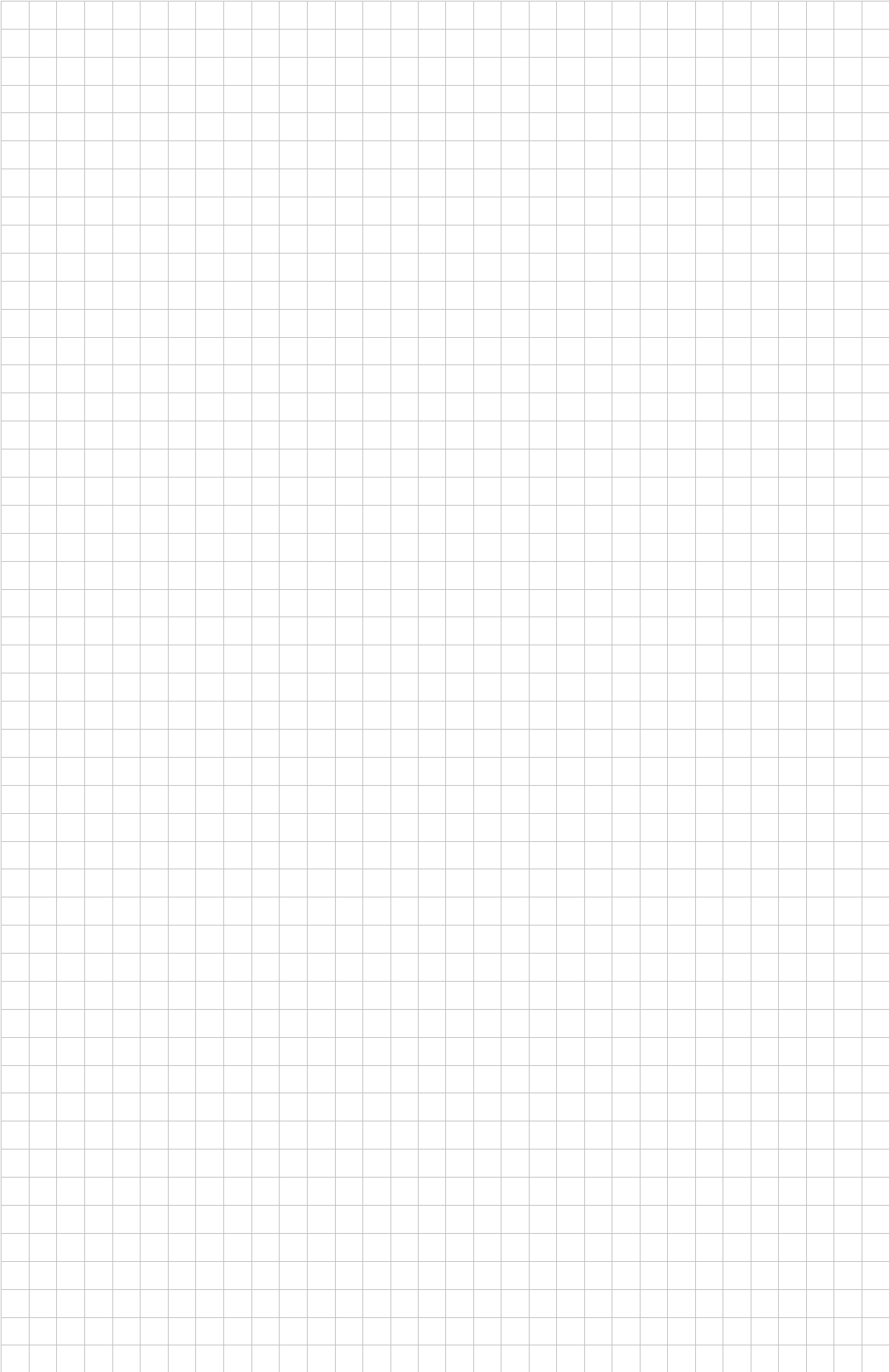
Nous voyons donc que l'état "mort" est bien un état absorbant car toute la probabilité y converge (malheureusement...). Nous voyons également que l'espérance de vie totale est de 4.99 cycles. Si nous assimilons un cycle à une année, alors l'espérance de vie est 4.99 années. Nous voyons immédiatement que la mesure stationnaire de la chaîne (cf. chapitre de Probabilités) est donc:

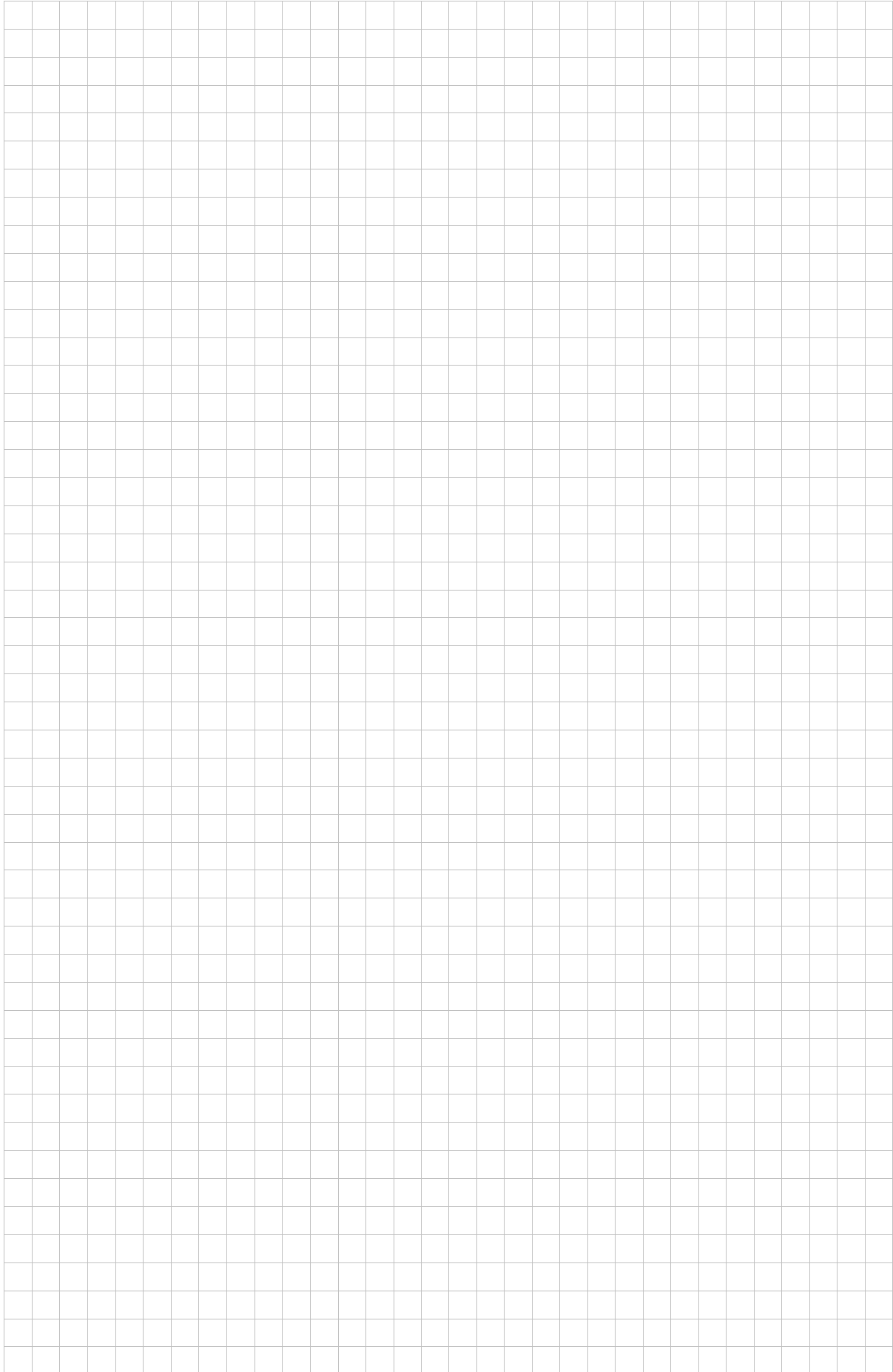
$$\pi = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad (65.93)$$

Notes personnelles:









66. ÉCONOMIE

L'économie, que nous assimilons sur ce site au domaine englobant la "théorie des biens", la "mathématique financière", "l'analyse financière", la "théorie des portefeuilles", a pour objectif de tenter de régler, de modéliser et de déterminer les origines, la dynamique et les optimums des prix de biens d'échanges ou valeurs "d'agents économiques" (acteurs du marché) en compétition rationnelle selon des modèles théoriques statistiques (simplifiés et idéalisés...) de marchés.

La profession étant majoritairement occupée par des anglo-saxons, nous indiquerons quand cela sera nécessaire les termes anglophones d'usage dans le domaine.

1. CONCEPTS

Un agent (économique) pour vivre va devoir satisfaire deux types de besoins qui peuvent exiger pour leur obtention un ou des échanges:

1. L'ensemble des "besoins primaires" (finis et dénombrables) ou physiologiques B_p
2. L'ensemble des "besoins secondaires" (qui ne sont pas vitaux et non nécessairement finis et dénombrables) B_s et qui sont subjectivement propres à tout individu (et pas que humain non plus!)

Remarque: Les besoins secondaires sont très difficiles à définir et à mesurer mais si nous raisonnons en des termes ensemblistes, nous pouvons simplement dire qu'est "besoin secondaire" tout ce qui est exclu de l'ensemble des besoins primaires $\bar{B}_p$.

Définitions:

D1. Nous disons qu'un besoin est un "besoin économique" quand il concerne un "bien rare" dont l'obtention exige un à plusieurs échanges. Ils s'opposent aux "biens libres" qui sont des biens disponibles à tous en abondance, aucun travail (typiquement...) n'étant supposé nécessaire pour en bénéficier.

La quantité importante de biens nous oblige à les classer de la manière suivante:

- C1. "Biens matériels" qui ont une réalité physique, palpable et qui peuvent être stockés.
- C2. "Biens intermédiaires" ou "services" dont la production et la consommation sont simultanées.
- C3. "Biens virtuels" qui n'ont qu'une existence mathématique et souvent limitée dans le temps.

D2. Un "marché" est un système constitué par la rencontre entre une offre et une demande qui porte sur un bien donné.

In extenso nous sommes amenés à énoncer les postulats suivants:

- P1. Le marché est assimilé à un système isolé et isotrope
- P2. Tout agent actif est en compétition

Définitions:

D1. La "**microéconomie**" est la branche de l'économie qui analyse le comportement économique au niveau d'entités individuelles telles qu'un consommateur ou une entreprise. Les consommateurs sont considérés comme des offreurs de travail et demandeurs de produits finis. Les firmes sont, quant à elles, des demandeuses de travail et des offreuses de produits.

D2. La "**macroéconomie**" est l'approche théorique qui étudie l'économie à travers les relations existant entre les grands agrégats économiques, le revenu, l'investissement, la consommation, le taux de chômage, l'inflation, etc.

1.1. MICROÉCONOMIE

Définition: La "**valeur d'échange**" d'un produit précise pour chaque bien la quantité des autres biens qui lui est équivalente. Usuellement, nous considérons que le "**prix**" (ou "**monnaie**") P est la forme monétaire de la valeur d'échange (nous reviendrons sur le concept de la monnaie plus tard).

Remarque: Le "**prix**" est un paramètre auquel s'intéresse l'économie. Tout bien matériel ou ressource humaine ainsi qu'une monnaie donnée ont un prix dont il faut déterminer la valeur (relative) soit de manière empirique soit avec des modèles mathématiques statistiques complexes.

Il existe différents types de prix dont voici un échantillon dans l'ordre d'un processus économique classique (les définitions sont propres à ce site!):

D1. Le "**prix de fabrication**" P_F d'un bien au temps t est déterminé par la somme des charges directes (mais pas forcément constantes...!) C_D de fabrication au même temps t (salaires, matières premières, machines, licences, brevets,...):

$$P_F(t) = \sum C_D(t) \quad (66.1)$$

D2. Le "**prix d'usine**" P_U d'un bien au temps t est la somme du prix de fabrication augmenté des charges indirectes C_I au même temps t (taxes, impôts, frais administratifs, frais de stockage, publicitaires, etc.). Afin de pouvoir modéliser un tant soit peu ce prix de manière théorique, nous allons supposer que le marché est à "flux tendu" ou à l'équilibre si vous préférez (nous verrons plus loin qu'il s'agit implicitement de la loi de Say). En d'autres termes, les biens sont fabriqués directement en fonction de la demande, sans stockage et sans intervalle de temps entre la mise sur le marché et la vente (c'est une approximation grossière mais nous y sommes contraints). Dès lors:

$$P_U(t) = P_F(t) + \sum C_I(t) \quad (66.2)$$

D3. Le "**prix de vente net**" P_{VN} d'un bien au temps t (ou vu de l'acheteur: le "**prix d'achat net**" P_{AN}) est le prix d'usine au même temps t augmenté de la "**marge sécuritaire**" (appelée aussi "**bénéfice brut**" ou encore ou "**marge bénéficiaire**") B_b de l'usine tel que:

$$P_{VN}(t) = P_U(t) + B_b(t) \quad (66.3)$$

Remarque: Ce bénéfice brut sera investi dans de multiples domaines par le fabricant (recherche et développement, redistribution aux investisseurs, etc.) et le solde doit permettre de se protéger contre les différentes fluctuations directes du marché, c'est-à-dire: les salaires, les taxes, les prix des matières premières.

Nous pouvons alors envisager deux cas de figure:

1. Le bénéfice brut est plus grand que la somme des charges générales et charges non prévues (il y aura donc un bénéfice net)
2. Le bénéfice brut est plus petit que les charges générales (il y aura donc un déficit ou perte nette)

De ce qui a été défini précédemment il découle trivialement que:

D4. Le "**bénéfice net**" B_n est donné par la partie de la marge sécuritaire qui était prévue pour une période et qui finalement n'a pas été utilisée par les charges imprévues C_A durant cette période telle que:

$$B_n = B_\phi - C_A \quad (66.4)$$

Remarque: Si les ventes sont supérieures aux prévisions et que des quotes-parts de charges générales et imprévues ont été comptées aux clients, nous parlons alors pour ce supplément imprévu de "**boni de suractivité**" ce qui augmente bien évidemment le bénéfice net prévu. Dans le cas contraire, nous parlons de "**coût d'inactivité partielle**", ce qui diminue bien évidemment le bénéfice net espéré.

D5. Le "**prix d'appel**" P_{AP} au temps t est le prix d'usine multiplié par un facteur F_{SA} sentimental et artistique (mode, ragots, raisons subjectives, etc.) appelé aussi parfois "**indice I.G.P.**" (pour "indice gros pipeau..."). Ce facteur peut être quantifié statistiquement à partir de l'unicité du bien, de la durée d'existence de celui-ci, du nombre d'acheteurs potentiels et ceci tant que personne n'intervient de manière à en modifier l'original après sa fabrication. Nous avons dès lors:

$$P_{AP}(t) = F_{SA}(t) \cdot P_U(t) \quad (66.5)$$

D6. Le "**prix de vente brut**" P_{VB} au temps t ou vu de l'acheteur le "**prix d'achat brut**" P_{AB} est le prix d'appel augmenté de la marge bénéficiaire du vendeur (intermédiaire entre l'usine et l'acheteur) plus les frais généraux de vente F_G . La marge du vendeur peut-être incluse dans un premier temps dans les charges directes mais les frais généraux ne sont pas déterministes mis à part dans un marché à flux tendu où il n'y en a pas et comme nous en avons fait l'hypothèse, nous avons:

$$P_{VB}(t) = P_{AP}(t) + B_\phi(t) + F_G(t) \cong P_{AP}(t) + B_\phi(t) \quad (66.6)$$

Remarque: Le prix d'achat brut est aussi parfois appelé "**prix catalogue**".

D7. Le "**prix de revient**" P_R au temps t est le prix de vente brut (ou d'achat selon le point de vue) diminué des différentes déductions D (étant une valeur négative) possibles ou obligatoires faites par le vendeur tel que:

$$P_R(t) \cong P_{VB}(t) + D(t) \quad (66.7)$$

Les agents du marché d'échange de biens admettent parfois une réduction sur le prix catalogue. Les réductions n'existent que sous deux formes connues:

1. La "**remise**" R qui est une bonification de prix accordée (valeur négative) soit à un agent demandeur qui achète un bien par fortes quantités N soit à un détaillant auquel est facturé un article de marque au prix de vente imposé par le fabricant (facteur stratégique commercial, promotion). La remise dépend donc du temps t et de la quantité N achetée ou commandée.
2. "**L'escompte**" ou "**ristourne**" E qui est une déduction au temps t consentie à l'agent demandeur pour paiement au comptant ou pour règlement anticipé ou encore pour paiement à une époque convenue (nous y reviendrons formellement lors de notre étude de l'intérêt simple en calcul actuariel plus loin).

Dans le cas le plus général qui soit nous parlerons à un temps t donné de "**prix d'exercice**" P_E (ou "**prix facturé**") auquel le bien peut être acheté ou vendu tel que:

$$P_E(t) = R(t, N) + E(t) + N \cdot \left\{ F_{SA}(t) \left[\sum C_D(t) + \sum C_I(t) + B_d(t) + F_G(t) + D(t) \right] \right\} \quad (66.8)$$

Remarque: L'ensemble des termes de ces expressions prennent généralement leurs valeurs dans $\mathbb{R}$...

Les facteurs à prendre en compte lors de l'élaboration d'une politique de prix sont synthétisés de manière non exhaustive dans le diagramme suivant:

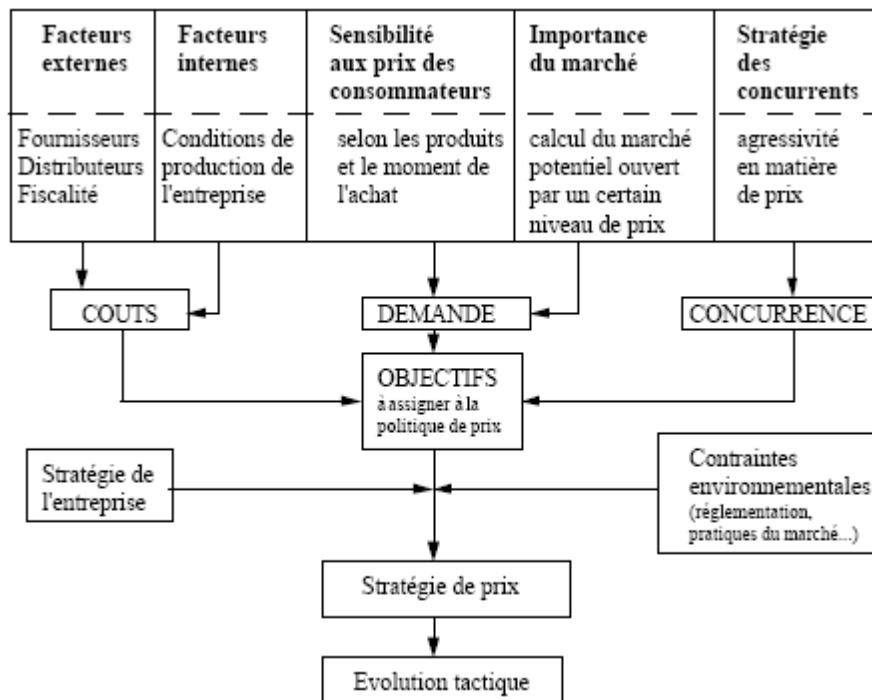


Figure: 66.1 - Résumé schématique simplifié pour l'élaboration d'une politique de prix

D1. La "**propension à consommer**" P_C est la part du revenu R d'un agent qui est consacrée à la consommation d'un montant C (consommation primaire et secondaire):

$$P_C = \frac{C}{R} \quad (66.9)$$

D2. La différence entre la dépense de consommation et le revenu est définie comme étant une "épargne" alors que les cotisations et prestations sur les revenus représentent des "transferts sociaux":

$$E = R - C \quad (66.10)$$

D3. "L'élasticité-revenu" est égale au rapport de la variation de la consommation sur la variation du revenu:

$$E_r = \frac{\Delta C}{\Delta R} \quad (66.11)$$

La notion d'élasticité-revenu permet de classer les biens de la manière suivante:

1. "**Biens inférieurs**": qui sont les biens de consommation dont l'élasticité par rapport au revenu est négative et donc dont la consommation diminue avec l'augmentation du revenu tel que $E_r < 0$ (le pain, la farine,...)
2. "**Biens supérieurs**": qui sont les biens de consommation de luxe dont l'élasticité par rapport au revenu est positive et donc dont la consommation augmente avec une augmentation du revenu tel que $E_r > 0$ (la santé, loisirs,...)
3. "**Biens normaux**": qui sont les biens neutres et dont le coefficient d'élasticité par rapport au revenu est un peu différent de 0 tel que $E_r \cong 0$.

D4. "L'élasticité-prix" est égale au rapport de la variation de la quantité de demande d'un bien sur la variation de son prix et est donnée par:

$$E_p = \frac{\Delta P}{\Delta D} \quad (66.12)$$

Remarque: Une demande est dite "**sensible au prix**" lorsque le pourcentage de variation de la quantité demandée est supérieur au pourcentage de variation de prix. Dans le cas contraire, nous parlons de demande "**rigide au prix**".

D5. Un "**investissement**" I est l'opération réalisée par un agent économique dont l'objectif est d'obtenir des biens de production en échange.

D6. La "**transaction**" T est l'échange d'une quantité de biens à un prix déterminé entre un "**vendeur**" et un "**acheteur**". Elle se conclut sur le marché dont la forme est déterminée par le nombre d'agents économiques qui y interviennent ce qui détermine la "**concurrence**".

Le tableau ci-dessous présente les différentes formes du marché:

Demandeurs	Offreurs		
	multitude	quelques-uns	un seul
multitude	concurrence parfaite	oligopole	monopole
quelques-uns	oligopsone	oligopole bilatéral	monopole contrarié
un seul	monopsone	monopsone contrarié	monopole bilatéral

Tableau: 66.1 - Différentes formes du marché de la concurrence

Une autre typologie des marchés peut être effectuée grâce à deux notions: la notion "concurrentielle" et la notion "d'état de la demande" qui se traduisent de la manière suivante:

Demande	Pression concurrentielle	
	Élevée	Faible
Stable	marché fermé	marché rigide
Instable	marché compétitif	marché ouvert

Tableau: 66.2 - Typologie des marchés

La concurrence est qualifiée de "**concurrence pure**" (CPP: concurrence pure et parfaite) si elle répond aux cinq hypothèses suivantes:

H1. Atomicité: Acheteurs et vendeurs sont nombreux au point que nul ne peut à lui seul influencer les prix.

H2. Homogénéité (postulat d'homogénéité): Les produits échangés sont identiques et substituables les uns aux autres. Ils permettent de satisfaire un même besoin.

H3. Libre entrée: Il n'existe aucune entrave à l'entrée et à la sortie de nouveaux agents.

H4. Libre déplacement: Les agents économiques peuvent se déplacer librement.

H5. Information parfaite: Tout le monde connaît en même temps et gratuitement toutes les quantités offertes et demandées par tous les agents aux prix différents.

Ces hypothèses étant données, revenons-en à nos définitions car nous n'avions pas terminé...

D7. Les "**soldes intermédiaires de gestion**" (S.I.G.) sont des parties du résultat global d'une période d'activité de marché qui sont significatives pour l'analyste financier. Il en existe de multiples dont les définitions découlent d'opérations algébriques élémentaires sur les concepts définis précédemment:

- La "**marge commerciale**" qui est la différence entre le produit des ventes de marchandises et le coût d'achat des marchandises vendues (la marge commerciale est spécifique aux activités de négoce, c'est-à-dire aux entreprises ayant une activité de distribution).

- La "**production de l'exercice**" qui est la somme des productions vendues, stockées et immobilisées (la production de l'exercice est spécifique aux activités de production, c'est-à-dire aux entreprises ayant une activité industrielle).

- La "**marge brute**" qui est la différence entre le produit retiré de la vente de l'exercice et les achats consommés de matières premières

- Le "**chiffre d'affaires**" qui est la somme des produits des ventes de marchandises et des ventes de biens et de services.

- La "**valeur ajoutée**" (V.A.) qui est définie comme la différence entre la production de l'exercice et la consommation intermédiaire par les agents (le gestionnaire la considère comme la richesse créée résultant de l'activité réelle de l'entreprise et la V.A. est comme nous l'avons vu d'importance nationale aussi car elle constitue un agrégat).

- "**L'excédent brut d'exploitation**", ou E.B.E., est le résultat de l'activité courante de l'entreprise et est défini comme étant:

$$E.B.E = V.A. + \text{Subventions d'exploitations} - \text{Impôts et taxe} - \text{Charges de personnel} \quad (66.13)$$

- Le "**résultat d'exploitation**" (R.E.) est l'enrichissement (ou l'appauvrissement) net généré par l'exploitation. Il prend en compte l'ensemble des produits et charges d'exploitation, notamment les amortissements, provisions, reprises et transferts de charges:

$$RE = E.B.E + \text{dotation aux amortissements et provisions} + \text{autres produits} - \text{autres charges} + \text{transferts de charges} \quad (66.14)$$

1.1.1. COÛT MOYEN ET MARGINAL

Introduisons ce sujet directement par un exemple:

Supposons qu'un cuisinier du dimanche (et économiste) invite ses amis à sa table et se propose de leur faire une salade de tomates. Il évalue le travail qu'il aura à faire et il chiffre ce travail en valeur monétaire. Il considère qu'une minute de son temps passée à la préparation de la salade correspond à une dépense de 1.-

Donc les données sont:

- Acheter des tomates à 1.- l'unité
- Préparer la salade en 15 minutes donc 15.-

Si chacun de ses amis est rassasié avec une seule tomate, préparer son dîner pour 5 amis (lui ne mangeant pas) lui coûtera au total:

$$C_T = 5.- + 15.- = 20.- \quad (66.15)$$

où contrairement aux calculs de la physique au sens large, l'homogénéisation des unités passe au second plan...

Mathématiquement, la "**fonction de coût total**" sera de la forme suivante si nous notons q le nombre d'amis (dans la pratique, il s'agit souvent d'une fonction polynômiale strictement croissante):

$$C_T(q) = 1 \cdot q + 15 \quad (66.16)$$

Le "**coût moyen**" ou "**fonction de coût moyen**" pour chaque invité est de 20.- divisé par 5 soit 4.-. Ce qui correspond à:

$$C_m(q) = \frac{C_T(q)}{q} \quad (66.17)$$

qui n'est donc pas une fonction nécessairement croissante.

Remarque: Selon les ouvrages (et auteurs), le coût moyen est noté avec un m minuscule ou M majuscule et le coût marginal aussi.

S'il en invite un sixième, le coût total sera de 21.-. En effet le temps de préparation restera, du moins nous le supposons..., constant. Dans ce cas, le coût marginal du sixième invité est de:

$$(21.-)-(20.-)=1.- \quad (66.18)$$

alors que le coût moyen pour l'ensemble des invités est alors de:

$$21.-/6=3.5.- \quad (66.19)$$

Nous remarquons donc dans cette situation que le coût moyen baisse indéfiniment tant que le nombre d'invités augmente.

Cet exemple permet d'illustrer les rendements d'échelle et montre que nous avons souvent intérêt à augmenter la production pour réduire le coût moyen de production (mais cela ne signifie pas pour autant que notre cuisinier du dimanche va apprécier le raisonnement lorsqu'il va arriver à sa limite physique...).

De plus, il ne s'agit cependant pas d'une règle générale! En effet, si le saladier de notre économiste ne peut contenir que 6 tomates, le 7ème invité va l'obliger à préparer un deuxième saladier. Dans ce cas, la variation du marginal sera supérieure au coût moyen préalable.

Il faut aussi savoir que dans la réalité la fonction de coût total est rarement une fonction continue (car il faut acheter des machines ou engager du personnel par palier de quantités) et de plus c'est le marché qui dicte les quantités à une usine et pas l'inverse.

Définition: Mathématiquement, le "**coût marginal**" est défini par la variation de la fonction coût total $C_T(q)$, par rapport à la quantité produite q :

$$C_M = \frac{\Delta C_T}{\Delta q} \quad (66.20)$$

ou si la quantité est dérivable (de toute façon en économie, même si la quantité n'est pas continue on en fait abstraction...):

$$C_M = \frac{dC_T}{dq} \quad (66.21)$$

Le coût marginal correspond ainsi au coût de la production d'une unité supplémentaire. En pratique, on s'intéresse plutôt au coût d'une série supplémentaire.

De plus, le lecteur remarquera que comme le coût total est une fonction strictement croissante, le coût marginal est aussi une fonction strictement croissante.

Remarque: Si le coût marginal augmente quand la quantité augmente, nous disons que les rendements sont décroissants. À l'opposé, ceux-ci sont croissants si le coût marginal est décroissant lorsque la quantité augmente. En effet, dans l'industrie notamment, on lance plutôt une série supplémentaire qu'une unité supplémentaire.

Démontrons maintenant que si le coût moyen passe par un extremum, le coût marginal lui est égal en ce point. Nous appelons cette situation particulière un "**optimum technique**".

Rappelons au préalable que si une fonction continue et dérivable $f(x)$ a un minimum (ou un maximum), sa dérivée en ce point s'annule (cf. chapitre de Calcul Différentiel Et Intégral).

Appliquons cela au coût moyen $C_m(q)$. Mais rappelons au préalable que (cf. chapitre de Calcul Différentiel Et Intégral):

$$\left(\frac{u}{v}\right)' = \frac{u'v - uv'}{v^2} \quad (66.22)$$

Nous avons donc si la dérivée du coût moyen est nulle:

$$\frac{dC_m}{dq} = 0 \Rightarrow \frac{d}{dq} \left(\frac{C_T}{q} \right) = \left(\frac{C_T}{q} \right)' = 0 \quad (66.23)$$

Il s'agit donc de la dérivée du rapport de deux fonctions. Donc:

$$\left(\frac{C_T}{q} \right)' = \frac{\frac{dC_T}{dq} q - C_T \frac{dq}{dq}}{q^2} = 0 \Rightarrow \frac{dC_T}{dq} q - C_T = C_T' q - C_T = 0 \quad (66.24)$$

D'où nous déduisons:

$$C_T' q = C_T \quad (66.25)$$

Soit:

$$C_T' = \frac{C_T}{q} \quad (66.26)$$

Ce qui nous dit déjà que dans cette situation, le coût marginal est égal au coût moyen (nous retrouvons la définition de chacun des coûts respectivement à gauche et droite de l'égalité).

En d'autres termes, là où le coût moyen atteint un optimum (minimum dans le cas qui nous intéresse), alors le coût marginal égale le coût moyen.

Exemple:

Une usine a calculé que le coût de production de q unités d'un produit peut être modélisé par la fonction suivante dans les limites des rendements des machines et de la main d'oeuvre dont elle dispose:

$$C_T(q) = 2'600 + 2 \cdot q + 0.001 \cdot q^2 \quad (66.27)$$

Calculons le coût, le coût moyen et le coût marginal de la production de 1'000, 2'000 et 3'000 unités pour l'exemple. Nous avons alors les informations suivantes:

$$\begin{aligned} C_T(q) &= 2'600 + 2 \cdot q + 0.001 \cdot q^2 \\ C_m(q) &= \frac{C_T(q)}{q} = \frac{2'600}{q} + 2 + 0.001 \cdot q \quad (66.28) \\ C_M(q) &= \frac{dC_T(q)}{dq} = 2 + 0.002 \cdot q \end{aligned}$$

Nous avons alors:

Quantité	Coût Total	Coût moyen	Coût marginal
1'000	5'600	5.60	4.00
2'000	10'600	5.30	6.00
3'000	17'600	5.87	8.00

Tableau: 66.3 - Différents coûts

Et pour que le coût moyen soit le plus faible possible, il faut que le coût marginal soit égal au coût moyen:

$$\frac{dC_T(q)}{dq} = \frac{C_T(q)}{q} \Rightarrow \frac{2'600}{q} + 2 + 0.001 \cdot q = 2 + 0.002 \cdot q \quad (66.29)$$

Nous en déduisons:

$$q = \sqrt{\frac{2'600}{0.001}} \cong 1'612 \quad (66.30)$$

En injectant cette valeur dans la fonction de coût total, nous en déduisons que le coût moyen minimum s'élève à 5.22.

Les économistes d'entreprises définissent ensuite une fonction de prix indépendante de la fonction de coût et en déduisent des recettes. Mais de mon expérience personnelle (qui n'attend que d'être éclairée), c'est n'importe quoi! D'abord parce que la fonction de prix est obligatoirement dépendante du coût, deuxièmement parce qu'en tant que consultant je n'ai jamais (j'insiste sur le "jamais") vu un économiste d'entreprise faire les calculs susmentionnés tellement la réalité est plus complexe.

Dans la réalité, on fait des simulations de Monte-Carlo (je le sais puisque j'interviens pour des grandes multinationales) qui permettent non seulement de travailler en avenir incertain et sur des étendues de quantités, mais qui peuvent aussi prendre en compte des fonctions non continues! Pour plus d'informations sur la modélisation de Monte-Carlo, le lecteur pourra se référer au chapitre de Méthodes Numériques.

1.2. MACROÉCONOMIE

Définition: Les "**agrégats**" sont des grandeurs synthétiques élaborées par les nations pour leur comptabilité nationale et qui mesurent le résultat de l'ensemble de leur économie. Les principaux agrégats sont définis par:

D1. Le "**produit intérieur brut**" (P.I.B.) qui a pour rôle de mesurer la production nationale (considérée comme isolée), c'est-à-dire de l'ensemble des valeurs des biens et services produits au cours d'une période donnée (le terme "Brut" indique que la valeur du P.I.B n'est pas déduite des différentes taxes existantes sur les productions).

D2. Le "**revenu national**" (R.N.) qui a pour rôle de mesurer l'ensemble des revenus perçus par les agents économiques.

D3. La "**consommation**" (C) qui a pour rôle de représenter la valeur des biens et services utilisés pour la satisfaction directe des besoins.

D4. La "**formation brute de capital fixe**" (F.B.C.F.) qui a pour rôle de représenter les investissements.

D5. La "**valeur ajoutée**" (V.A.J.) d'une entreprise qui a pour rôle de représenter la différence entre la valeur des biens et services produits par celle-ci avec la valeur des biens et services utilisés pour produire ces mêmes biens et services.

D6. Le "**produit national brut**" (P.N.B.) qui a pour rôle de mesurer la production nationale (comme le P.I.B.) et de prendre en compte les revenus du reste du monde. En d'autres termes, le P.N.B est le P.I.B. auquel nous soustrayons les capitaux en provenance de l'extérieur et auquel on soustrait les capitaux versés à l'extérieur.

Remarque: Inutile de parler du concept d'inflation qui ne veut rien dire et dont nous ne retrouvons de définition mathématique rigoureuse nulle part! À ce jour ce terme et le chiffre qui lui est associé ne veulent rien dire.

1.3. MODÈLE DE COBB-DOUGLAS

En 1928, Charles Cobb et Paul Douglas ont publié une étude dans laquelle apparaissait une modélisation de la croissance de l'économie américaine entre 1899 et 1922. Ils y avaient adopté une vue simplifiée de l'économie selon laquelle la quantité produite n'est fonction que de la quantité de travail réalisé et du montant des capitaux investis.

Malgré que beaucoup d'autres facteurs affectent les performances économiques, leur modèle s'est avéré remarquablement précis. La fonction qu'ils ont employée pour modéliser la production était de la forme:

$$P(L, K) = bL^\alpha K^{1-\alpha} \quad (66.31)$$

où P est la production totale (la valeur monétaire de tous les biens produits en un an), L la quantité de travail (le nombre total d'heures de travail prestées en un an) et K la quantité de capital investi (la valeur monétaire de toutes les machines, équipements et bâtiments).

Bien que ce modèle ait été appliqué en premier lieu à l'économie, on le retrouve dans de nombreux articles scientifiques de biologie ou de ressources humaines. Il nous a semblé pertinent de montrer en détail (puisque c'est le but du site) d'où vient cette étrange fonction car le démarche est relativement géniale.

D'abord, si nous avons une fonction de deux variables $P(L,K)$, la dérivée partielle $\partial P / \partial L$ indique le taux de variation de la production par rapport à la quantité de main d'oeuvre seule. C'est ce que les économistes appellent la "**production marginale par rapport au travail**" ou la "productivité marginale du travail". De même la dérivée partielle $\partial P / \partial K$ indique le taux de variation de la production par rapport au capital et est appelée la "productivité marginale du capital".

En ces termes, les hypothèses de Cobb et Douglas (à confronter aux mesures historiques et à adapter à ces mêmes mesures au cas par cas) s'énoncent comme suit:

H1. Sans travail L ou sans capital K pas de production

H2. La productivité marginale du travail est proportionnelle à la quantité produite P par unité de travail L

H3. La productivité marginale du capital est proportionnelle à la quantité produite P par unité de capital K

Vu que la production par unité de travail est le rapport P/L , l'hypothèse 2 s'écrit:

$$\frac{\partial P}{\partial L} = \alpha \frac{P}{L} \quad (66.32)$$

pour une certaine constante α . Au cas où le capital investi K est constant ($K = K_0$), l'équation aux dérivées partielles devient une équation différentielle ordinaire:

$$\frac{dP}{dL} = \alpha \frac{P}{L} \quad (66.33)$$

Ce qui nous donne:

$$\frac{1}{P} dP = \alpha \frac{1}{L} dL \quad (66.34)$$

Et donc la résolution de cette équation différentielle donne:

$$\ln(P) = \alpha \ln(L) + c^{te} \Rightarrow e^{\ln(P)} = e^{\alpha \ln(L)} e^{c^{te}} = C^{te} e^{\ln(L)^\alpha} \quad (66.35)$$

D'où:

$$P(L, K_0) = C^{te}(K_0) L^\alpha \quad (66.36)$$

car comme nous avons fixé K il va de soi que la constante dépend de K_0 .

En procédant de même pour l'hypothèse 3:

$$\frac{\partial P}{\partial K} = \beta \frac{P}{K} \quad (66.37)$$

Il vient:

$$P(L_0, K) = C^{te}(L_0) K^\beta \quad (66.38)$$

Donc nous avons au final:

$$\begin{aligned} P(L, K_0) &= C^{te}(K_0) L^\alpha \\ P(L_0, K) &= K^\beta C^{te}(L_0) \end{aligned} \quad (66.39)$$

et en identifiant terme à terme, nous avons immédiatement:

$$P(L, K) = c^{te} K^\beta L^\alpha \quad (66.40)$$

Observons ce qu'implique cette dernière relation si le capital et le travail sont tous les deux multipliés par un facteur m :

$$P(mL, mK) = c^{te} (mK)^\beta (mL)^\alpha = c^{te} m^{\alpha+\beta} K^\beta L^\alpha = m^{\alpha+\beta} P(L, K) \quad (66.41)$$

Si nous imposons:

$$\alpha + \beta = 1 \quad (66.42)$$

alors la production est elle aussi multipliée par le même facteur m . Et nous nous retrouvons dans une hypothèse de proportionnalité de la production ce qui semblait cohérent à Cobb et Douglas. Dès lors:

$$P(L, K) = c^{te} K^{1-\alpha} L^\alpha \quad (66.43)$$

Les données économiques exploitées par Cobb et Douglas sont celles de la table ci-dessous publiées par le gouvernement des États-Unis d'Amérique:

Année	P	L	K
1899	100	100	100
1900	101	105	107
1901	112	110	114
1902	122	117	122
1903	124	122	131
1904	122	121	138
1905	143	125	149
1906	152	134	163
1907	151	140	176
1908	126	123	185
1909	155	143	198
1910	159	147	208
1911	153	148	216
1912	177	155	226
1913	184	156	236
1914	169	152	244
1915	189	156	236
1916	225	183	298
1917	227	198	335
1918	223	201	366
1919	218	196	387
1920	231	194	407
1921	179	146	417
1922	240	161	431

Tableau: 66.4 - Données U.S. en % utilisées par Cobb et Douglas

Ils prirent délibérément l'année 1899 comme base, c'est-à-dire qu'ils attribuèrent le niveau 100 à chacun des facteurs et exprimèrent les valeurs des autres années en pourcentage de cette année-là.

Pour déterminer les coefficients de Cobb-Douglas, il s'agit de faire une régression linéaire par la méthode des moindres carrés (cf. chapitre de Méthodes Numériques). Si nous prenons le logarithme:

$$\begin{aligned}
 \ln(P) &= \ln(c^{te} K^{1-\alpha} L^{\alpha}) = \ln(c^{te} K^1 K^{-\alpha} L^{\alpha}) = \ln\left(c^{te} K \left(\frac{L}{K}\right)^{\alpha}\right) \\
 &= \ln(c^{te}) + \ln(K) + \ln\left(\left(\frac{L}{K}\right)^{\alpha}\right) = \ln(c^{te}) + \ln(K) + \alpha \ln\left(\frac{L}{K}\right)
 \end{aligned} \quad (66.44)$$

et après réarrangement, toujours en utilisant les propriétés des logarithmes (cf. chapitre d'Analyse Fonctionnelle):

$$\ln\left(\frac{P}{K}\right) = \ln(c^{te}) + \alpha \ln\left(\frac{L}{K}\right) \quad (66.45)$$

En posant:

$$x = \ln\left(\frac{L}{K}\right) \quad y = \ln\left(\frac{P}{K}\right) \quad (66.46)$$

Nous nous retrouvons simplement avec une fonction affine:

$$y = \alpha x + \ln(c^{te}) \quad (66.47)$$

Il suffit alors d'adapter le tableau, de l'injecter dans un tableur ou logiciel de statistique, pour obtenir (cf. chapitre de Méthodes Numériques):

$$\alpha \cong 0.75 \quad c^{te} \cong 1.01 \quad (66.48)$$

Donc:

$$P(L, K) = 1.01 K^{0.25} L^{0.75} \quad (66.49)$$

En testant par rapport aux données du tableau il vient:

$$\begin{aligned} P_{1908}(123, 185) &\cong 137 \\ P_{1910}(147, 208) &\cong 161 \\ P_{1920}(194, 407) &\cong 235 \\ P_{1922}(161, 431) &\cong 207 \end{aligned} \quad (66.50)$$

Donc nous voyons que la modèle est... un modèle... avec une certaine marge d'erreur.

2. MODÈLE MONÉTAIRE

Pour construire un modèle monétaire nous ferons l'hypothèse que "l'utilité monnaie" peut être définie a priori par trois propriétés:

P1. Unité de compte

P2. Moyen de paiement (intermédiaire d'échange)

P3. Réserve de valeur

Cette démarche de description est cependant insuffisante pour l'analyse mathématique: il faut un système explicatif complet, car, ici, nous ne faisons que constater, sans rien de plus. Il faut donc établir le lien entre la monnaie et la théorie de la valeur.

Mise à part la représentation valeur que représente la monnaie, celle-ci dérive son utilité des biens qu'elle permet d'obtenir dans l'échange. C'est ce que nous nommons "**l'utilité dérivée**".

Notons l'offre de monnaie disponible d'un marché O_u . Elle dépend donc de la quantité totale existante de monnaie Q_u moins les "**encaisses**" e conservées par les agents économiques (qui ont échangé des biens contre de la monnaie). Nous pouvons alors écrire la relation suivante nommée "**offre de monnaie selon Walras**":

$$O_u = Q_u - e \Leftrightarrow Q_u = O_u + e \quad (66.51)$$

Cette encaisse est aussi celle des ménages d'une certaine manière et est une demande réelle de biens, qui peut s'exprimer nécessairement sous forme monétaire.

Remarque: Si les gens savaient réellement comment le processus de création monétaire moderne fonctionne (celui utilisé depuis que la majorité des États à travers le monde autorisent les banques à créer de "l'argent dette" - c'est-à-dire de créer de l'argent à partir d'une promesse de remboursement de dette) le système ne tiendrait pas longtemps. Par ailleurs, le système contemporain est en réalité relativement fragile et vicieux (il va probablement droit vers le mur).

Des agents de vente, à l'occasion de la vente de leurs biens, désirent a priori encaisser une certaine somme de monnaie contre la vente de ces biens, somme notée D_u et appelée "**encaisse de monnaie désirée**".

Nous exprimons cette encaisse de monnaie désirée en "**numéraire**" et pour ce, nous introduisons alors un prix de la monnaie en numéraire. L'encaisse désirée s'écrit alors par rapport à la totalité des encaisses du marché. Le numéraire sert lui à exprimer les prix relatifs pour l'équilibre général. Il y a une encaisse désirée de la part des agents pour la réalisation de l'équilibre général. Ce sont en fait des biens réels sous forme monétaire:

$$e = p_u D_u \quad (66.52)$$

où p_u est le "**prix de la monnaie en numéraire**" (facteur variable au cours du temps et qui amène dans un marché qui n'est pas à flux tendu à faire de la spéculation). Dans un marché à flux tendu, p_u sera toujours supposé égal à l'unité. Nous pouvons alors écrire pour différents numéraires:

$$p_u D_u = \sum_{i=1}^n p_i D_i \quad (66.53)$$

Remarque: Dans un marché isotrope à monnaie unique cette relation n'aurait pas besoin d'être écrite.

L'encaisse désirée peut alors s'exprimer en utilisant la relation:

$$D_u = \frac{1}{p_u} \sum_{i=1}^n p_i D_i \quad (66.54)$$

Revenons, à O_u mais cette fois-ci vu du côté des entreprises. Elles ont besoin de monnaie pour effectuer les paiements et fonctionner (salaires, investissements, etc.) et l'encaisse désirée de l'ensemble de ces entreprises est nécessairement dans un cas idéal égale à l'ensemble de la monnaie disponible sur le marché tel que:

$$D'_u = O_u \quad (66.55)$$

puisque les entreprises vendent des biens sur les encaisses des agents (moins les marges) du marché économique.

Hypothèse: La dernière relation suppose que le prix de vente des marchandises tend à être égal à leur prix de revient

Remarque: Cette relation signifie aussi que toute l'offre est satisfaite uniquement par la demande des agents et que l'encaisse précédemment citée n'est constituée que de biens hors entreprises.

Cela correspond également à une certaine quantité de biens puisqu'il s'agit de proposer des biens pour se procurer de la monnaie (vu des entreprises). Nous pouvons donc écrire:

$$D'_u = O_u$$

$$D'_u = \left[\sum_{j=1}^n p_j D'_j \right] \frac{1}{p_u} \quad (66.56)$$

Mais comme les biens du marché (en possession des agents économiques) doivent également être renouvelés les entreprises ont finalement comme quantité de monnaie totale potentielle disponible sur le marché:

$$Q_u = D'_u + D_u$$

$$Q_u = \frac{1}{p_u} \left[\sum_{j=1}^n p_j D'_j + \sum_{i=1}^n p_i D_i \right] \quad (66.57)$$

La somme entre crochets correspond donc à l'ensemble de la monnaie disponible sur le marché sous forme de biens des ménages et des encaisses potentielles sous la restriction de biens ayant un prix de la monnaie en numéraire p_u global identique. C'est restrictif comme modèle mais suffisant dans le cadre de la détermination du prix d'un type de bien.

Nous notons alors par définition:

$$Q_u = \frac{1}{p_u} H_u \quad (66.58)$$

où:

$$H_u = \left[\sum_{j=1}^n p_j D'_j + \sum_{i=1}^n p_i D_i \right] \quad (66.59)$$

Nous avons donc in extenso:

$$H_u = p_u Q_u \quad (66.60)$$

La première relation encadrée exprime la "**théorie quantitative de la monnaie selon Walras**".

Passons à l'examen du modèle qui est fondé sur l'association des trois éléments (dont certains ont déjà été énoncés plus haut) suivants:

- La "**loi de Say**": Il ne peut y avoir de déséquilibre durable sur les marchés et la loi de l'offre et de la demande réalise une régulation spontanée et automatique de l'activité économique

- La C.P.P.: La concurrence est pure et parfaite (voir tableau plus haut)

- La "**loi de Walras**": La valeur totale des offres étant identique à la valeur totale des demandes, si l'équilibre entre offre et demande est réalisé sur $n-1$ marchés alors il est réalisé sur le n -ième marché.

Ainsi, l'objectif de Walras est de répondre à la question de savoir s'il existe un système de prix qui assure l'équilibre entre l'offre et la demande sur tous les marchés. Cette question est importante, car de sa réponse dépend la capacité du marché à assurer l'allocation des ressources de façon efficace.

Au vu de ce qui précède, le lecteur aura remarqué que le modèle de Walras considère que la monnaie est neutre en ce sens que la quantité totale de monnaie en circulation n'exerce d'influence ni sur les prix relatifs des produits les uns par rapport aux autres, ni sur le niveau de l'offre et de la demande de produits. La monnaie n'est pas souhaitée pour elle-même...

Remarque: La "**parité**" est le terme utilisé pour chercher l'équivalence des cours monétaires étrangers de différents marchés. Cette parité est dépendante (entre autres) du temps et il est important de considérer les variations de celle-ci dans le cadre du marché des biens où la monnaie n'est pas unique et le paiement non immédiat.

Nous allons maintenant mettre en évidence l'interdépendance des marchés selon Walras:

Nous supposons une économie composée de n marchés où nous avons la demande de biens notée x_i^D , et l'offre notée x_i^S et où nous avons, enfin, les prix p_i (exprimés par rapport à un autre bien) sur la marché i .

Selon la loi de Say, nous avons (équilibre entre l'offre et la demande sur tous les marchés):

$$\sum_{i=1}^n p_i x_i^D = \sum_{i=1}^n p_i x_i^S \quad (66.61)$$

L'objectif de cette loi est de montrer l'interdépendance des marchés. Pour cela, il faut faire appel à la demande excédentaire notée x_i^{XD} (différence entre l'offre et la demande). Nous avons alors (toujours de par la loi de Say):

$$x_i^{XD} = x_i^S - x_i^D \Rightarrow \sum_{i=1}^n p_i x_i^{XD} = 0 \quad (66.62)$$

Conclusion: S'il y a un déséquilibre sur un marché i , cela implique qu'il y aura un autre déséquilibre de même ampleur sur l'ensemble des autres marchés. C'est une première manière de mettre en évidence l'équilibre des marchés par l'intermédiaire de l'équilibre entre l'offre et la demande.

Remarque: Les variations des prix des monnaies en numéraire n'affectent pas l'équilibre réel. Si tous ces prix relatifs varient dans la même proportion, l'équilibre n'est pas modifié (et donc les demandes excédentaires ne devraient n'y augmenter ni diminuer).

Avant de continuer, rappelons qu'une fonction f est homogène de degré r si en multipliant tous ses termes par un même facteur k , nous obtenons:

$$f(kx) = k' f(x) \quad (66.63)$$

De cette définition il s'ensuit la propriété intuitive que dans un marché où la demande est supposée proportionnelle au prix..., les fonctions de demande sont homogènes de degré 1 telles que:

$$x_i^{XD}(\alpha p_1, \alpha p_2, \dots, \alpha p_n) = \alpha x_i^{XD}(p_1, p_2, \dots, p_n) \quad (66.64)$$

Avec ce que nous avons dit tout à l'heure, nous devrions dès lors avoir une équivalence telle que:

$$\sum_{i=1}^{n+1} p_i x_i^{XD} = 0 \Leftrightarrow \sum_{i=1}^n p_i x_i^{XD} + (p_{n+1} x_{n+1}^{XD}) = 0 \quad (66.65)$$

Démonstration:

Si tous les prix augmentent de α et qu'il y a un (nous pouvons généraliser à n) nouveau bien sur le marché dont le prix augmente de la même valeur et dont la loi de l'offre et de la demande est également proportionnelle au prix, alors:

$$\begin{aligned} \sum_{i=1}^n (\alpha p_i) x_i^{XD} + x_{n+1}^{XD}(\alpha p_1, \dots, \alpha p_{n+1}) &= \alpha \sum_{i=1}^n p_i x_i^{XD} + \alpha x_{n+1}^{XD}(p_1, \dots, p_{n+1}) \\ &= \alpha \left[\sum_{i=1}^n p_i x_i^{XD} + x_{n+1}^{XD}(p_1, \dots, p_{n+1}) \right] = \alpha \sum_{i=1}^{n+1} p_i x_i^{XD} = \alpha \cdot 0 = 0 \end{aligned} \quad (66.66)$$

□C.Q.F.D.

L'équilibre n'est donc pas affecté par la variation des prix monétaires (vous comprenez maintenant que les salaires sont un prix numéraire de la valeur monétaire du travail qui augmente(ra)it lui aussi proportionnellement aux prix des biens du marché).

Cette pseudo-démonstration (soyons objectifs...) implique donc que le $n + 1$ ème marché est totalement déterminé par les n autres.

Ici, les relations sont fondées sur des équations. Walras distingue cependant deux procédures pour assurer l'équilibre entre offre et demande:

1. Une méthode algébrique théorique. Mais... nous ne pouvons pas déterminer les besoins des individus à l'avance afin de savoir quand il y aura demande et se préparer à construire l'offre. Ce système ne fonctionne que si et seulement si les agents économiques sont raisonnables et s'accordent pour attendre
2. Une méthode empirique qui recherche la solution par des opérations d'essais/erreurs: il y a la présence d'une sorte de secrétaire de marché, le "commissaire-priseur". Ce dernier annonce des prix pour chaque type de bien qui pourrait exister: les agents économiques réagissent à ce prix, ils offrent et ils demandent en fonction du prix. Pour le bien i , il y a p_i^0 , nous avons alors D_i^0, O_i^0 . Nous comparons l'offre et la demande. En cas d'égalité, le prix est un prix d'équilibre. En cas de différence, le commissaire-priseur recommence la procédure et ainsi de suite jusqu'à ce qu'il y ait équilibre. C'est en gros cette procédure qui est utilisée dans les marchés boursiers et que l'on appelle souvent la "**main invisible des marchés**"!!!

Cependant, les équations nous montrent que nous avons besoin du prix de la monnaie en numéraire pour mesurer l'offre et la demande et il convient de se rappeler que nous avons considéré la monnaie comme une marchandise en quantité donnée fixe car le système est à l'équilibre entre offre et demande. Mais justement, les agents ne peuvent pas indéfiniment se répartir la quantité totale de monnaie si leur nombre augmente. Dès lors, pour que la demande soit possible, si elle a lieu, il faut être prêt à en injecter (ou à en disposer) sur le marché (sinon celui-ci devient immobile ce qui n'est peut-être pas favorable à long terme...). Il faut bien sûr aussi être prêt à en retirer et c'est là aussi qu'intervient une instance tel que l'État en intervenant dans l'économie pour réguler cette quantité de toutes les manières possibles (par l'intermédiaire des impôts par exemple) puisqu'elle agit directement sur les biens disponibles et déjà immobiles (achetés).

Ainsi, selon le modèle de Walras, la quantité de monnaie disponible sur le marché est donc seulement fonction du nombre d'agents économiques. Mais dès lors faut-il mettre en place un nouveau modèle pour un cadre plus général de demande de monnaie?

Au fait, cela n'est pas nécessaire. Nous savons que s'il y a équilibre général pour n biens, il y a équilibre général pour $n + 1$ biens (et par récurrence pour $n-1$ aussi) ; le dernier marché n'étant autre que celui de la monnaie. Le modèle de Walras explique dès lors pourquoi à un certain niveau de quantité de monnaie correspond un certain niveau des valeurs numéraires des biens et ce même de la monnaie.

3. THÉORIE DE L'OFFRE ET DE LA DEMANDE

C'est le deuxième des cinq modèles cités plus haut. Il nécessitera nécessairement (et cela est prévu!) une révision car fortement incomplet. Les idées présentées ci-dessous sont à ce jour à prendre avec des pincettes.

Dans notre société humaine où il existe une monnaie d'échange (de référence) et des biens persiste un problème qui consiste en la détermination de la valeur monétaire d'un bien. Pour déterminer celle-ci, il faut pouvoir connaître l'évolution de l'offre et de la demande. C'est ce à quoi nous allons nous attarder ici en commençant par des modèles simplistes et en complexifiant ceux-ci de manière croissante:

3.1. THÉORIE DE LA PRÉFÉRENCE

Avant de se lancer dans un modèle élaboré de l'offre et de la demande, il est nécessaire de cerner ce qui motive les agents économiques dans leurs choix de consommation et de modéliser leurs comportements sous le principe fondamental de rationalité.

L'agent économique sera perçu comme un individu unique disposant d'un dont il cherche à tirer le maximum de satisfaction. Ses goûts sont subjectifs même s'ils dépendent de certaines caractéristiques objectives comme l'âge ou le niveau de culture. Le niveau de satisfaction sera défini à partir d'une fonction d'utilité dont nous verrons les principes de base et la maximisation sous contrainte.

Plusieurs principes fondent l'utilité des biens et conduisent à la notion "**d'utilité marginale**", concept central dans la théorie de la préférence de l'agent économique. D'après Aristote, à l'origine du concept de valeur-utilité, l'utilité des biens dérive de la satisfaction des besoins. Condillac énonce que: "la valeur des choses est fondée sur l'usage que nous pouvons en faire". Cette idée d'une valeur fondée sur l'utilité, fondamentale chez les économistes marginalistes, s'oppose au courant théorique de la valeur-travail fondée sur la quantité de travail, directe et indirecte, incorporée dans la fabrication du bien (Adam Smith, Karl Marx).

Il faut cependant, de préférence, considérer une hypothèse importante dans ce modèle :

Hypothèse: Il existe une certaine satiété des besoins, mais elle n'est jamais totale.

Ainsi, pour un bien donné l'utilité marginale de la dernière unité consommée est donc plus faible que celle des unités précédentes mais non nulle et toujours positive c'est le "**principe de l'utilité marginale décroissante**" relative à l'unité supplémentaire consommée.

Ainsi, dans le cadre de la consommation multiple $n \in \mathbb{N}$ d'un bien unique d'utilité nominale donnée, l'utilité totale U_T (somme des utilités marginales $U_M(n)$) serait une courbe du type:

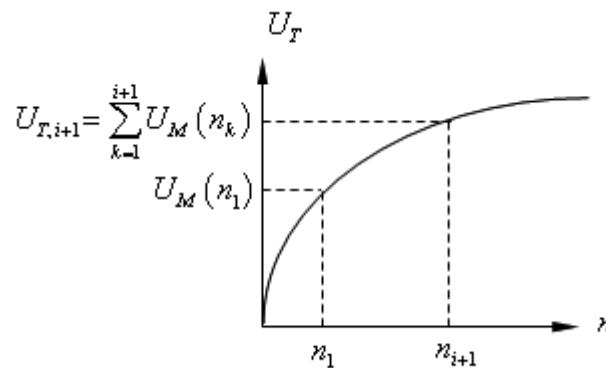


Figure: 66.2 - Utilité asymptotique

et donc l'utilité marginale est du type:

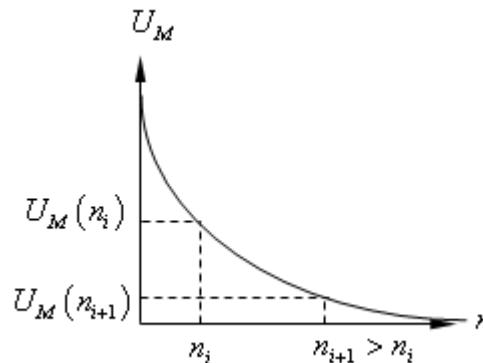


Figure: 66.3 - Utilité marginale décroissante correspondante

Ainsi, confronté à un prix donné pour chaque bien, l'agent économique compare ce prix avec les utilités marginales qu'il retire successivement de leur consommation. Il en achète tant que leur utilité dépasse le prix (surplus lié à cet achat) et cesse d'en acheter dès que l'utilité marginale tombe en dessous du prix du bien. Son intérêt est alors d'acheter d'autres produits pour lesquels il existe un surplus positif (utilité marginale de ces produits supérieure à leur prix).

Cet exemple, relatif à un bien, doit être élargi maintenant à un panier de biens pour déterminer l'utilité globale de ce panier.

Considérons pour cela un agent i dans une économie disposant de $I \in \mathbb{N}$ biens. Il peut donc en acheter au maximum I . Un panier de consommation possible correspond donc au vecteur de biens $\vec{x}_i: \vec{x}_i = (x_1, \dots, x_k)$, où les x_k représentent les quantités éventuellement nulles achetées par le consommateur. L'utilité de ce panier s'écrit $U(\vec{x}_i)$ et sera supposée additive selon:

$$U(\vec{x}_i) = \sum_k U_{r,k}(x_k) \quad (66.67)$$

c'est-à-dire la somme des utilités totales relatives aux quantités consommées de chaque bien.

Considérons maintenant un panier à deux biens, nous pouvons sans trop d'erreur émettre l'hypothèse que ces biens peuvent être divisés en fractions x_i aussi petites que nous voulons d'autres biens/composants.

Ainsi, grossièrement, nous ne travaillons plus dans $\mathbb{N}$ mais dans $\mathbb{R}_+$.

Ainsi, soit un panier de $\mathbb{R}_+^2$ d'un agent économique, nous supposons que celui-ci est tel que sa différentielle totale exacte est nulle telle que:

$$dU = \frac{\partial U}{\partial x_i} dx_i + \frac{\partial U}{\partial x_j} dx_j = 0 \quad (66.68)$$

Le rapport:

$$T.M.S. = - \frac{dx_i}{dx_j} = \frac{\frac{\partial U}{\partial x_j}}{\frac{\partial U}{\partial x_i}} \quad (66.69)$$

est défini comme le "**taux marginal de substitution**" (T.M.S.) entre les deux biens élémentaires i, j : quantité supplémentaire du bien i qu'il faut fournir à l'agent économique pour compenser exactement une diminution d'une unité du bien j .

Le comportement attribué à l'agent économique est de pouvoir classer tous ces paniers de biens possibles (vecteurs) selon une échelle de préférence, sans que celle-ci corresponde nécessairement à une évaluation chiffrée. Cette capacité de classement correspond au concept "**d'utilité ordinale**" (pouvant être ordonnée donc) et à l'utilisation d'une relation de préférence, notée $\succsim$ (préféré ou indifférent à) qui vérifie les propriétés suivantes:

- transitivité: $\vec{x}_1 \succsim \vec{x}_2$ et $\vec{x}_2 \succsim \vec{x}_3 \Rightarrow \vec{x}_1 \succsim \vec{x}_3$ (cohérence des classements successifs)
- réflexivité: $\vec{x}_1 \succsim \vec{x}_1$

Cette relation, "préordre" au sens mathématique, est utilisée dans la plupart des présentations actuelles de la théorie de la préférence. Ce préordre est complet s'il permet toujours de comparer deux paniers de biens dans l'ensemble des I biens.

Un tel préordre complet permet de définir une relation d'équivalence sur l'ensemble des biens et un ordre strict, ainsi que de représenter les préférences à partir de fonctions d'utilité:

$$\begin{aligned}
 (1) \vec{n}_1 \succsim \vec{n}_2 &\Leftrightarrow U(\vec{n}_1) \geq U(\vec{n}_2) \\
 (2) \vec{n}_1 > \vec{n}_2 &\Leftrightarrow U(\vec{n}_1) > U(\vec{n}_2) \quad (66.70) \\
 (3) \vec{n}_1 \sim \vec{n}_2 &\Leftrightarrow U(\vec{n}_1) = U(\vec{n}_2)
 \end{aligned}$$

où nous avons dans l'ordre: (1) $\vec{n}_1$ préféré ou indifférenciable à $\vec{n}_2$ (2) $\vec{n}_1$ est strictement préféré à $\vec{n}_2$ (3) $\vec{n}_1$ est indifférent à $\vec{n}_2$.

Si la fonction U est bien définie par un nombre, elle ne reflète plus une évaluation de l'utilité, mais dès lors seulement la possibilité de comparer l'ordre des utilités, relatives à des paniers de biens quelconques.

La possibilité de hiérarchiser différents paniers de biens de $\mathbb{R}^I$ permet de définir des surfaces de niveau dont l'utilité est constante, appelées "**courbes d'indifférence**" ou encore "**courbes d'iso-utilité**". Les graphiques suivants donnent bien une représentation de ces courbes dans $\mathbb{R}^2$ (panier de deux biens) et leurs principales propriétés:

Ainsi, deux paniers tels que $\vec{A} \sim \vec{B}$ dans $\mathbb{R}^2$ se traduisent graphiquement par un réseau de courbes, dont chacune est constamment décroissante:

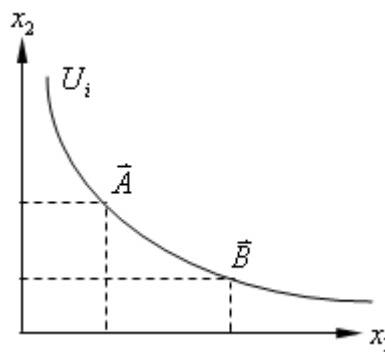


Figure: 66.4 - Exemple de courbe d'iso-utilité (d'indifférence)

Pourquoi n'avons-nous pas des droites ou autres choses? La raison en est relativement simple et le graphique suivant l'explique trivialement. Considérons l'iso-utilité (l'indifférence):

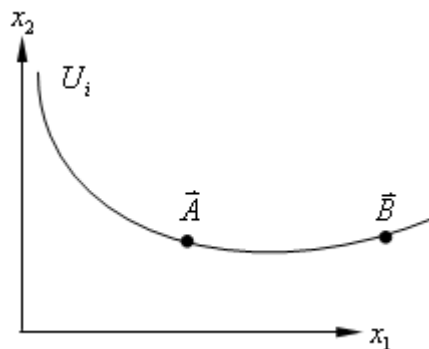


Figure: 66.5 - Contre-exemple d'iso-utilité (d'indifférence)

Ci-dessus, $\vec{B}$ domine $\vec{A}$. En effet, le panier de biens $\vec{B}$ possède plus de biens x_1, x_2 que le panier $\vec{A}$. Ces deux points ne peuvent donc être sur la même courbe d'indifférence et imposent qu'une courbe d'indifférence doit être décroissante et que c'est la seule condition (donc ce n'est pas nécessairement une droite).

Remarque: Si nous supposons que la satisfaction de l'agent économique augmente avec la taille de son panier de biens, plus une courbe d'iso-utilité est éloignée de l'origine plus elle correspond à une utilité élevée.

Les courbes d'iso-utilité (d'indifférence) ne peuvent se couper (ne peuvent être sécantes). Effectivement, soit les deux courbes d'iso-utilité (d'indifférence) ci-dessous:

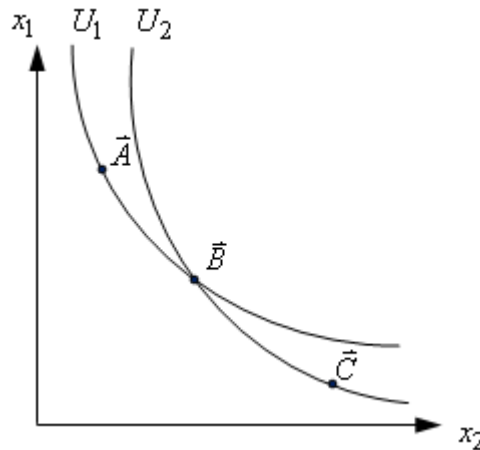


Figure: 66.6 - Autre contre-exemple d'iso-utilité (d'indifférence)

Les paniers $\vec{A}$ et $\vec{B}$ sont situés sur la même courbe d'iso-utilité U_1 alors que $\vec{B}$ et $\vec{C}$ sont, eux, situés sur la même courbe U_2 . Ainsi, nous pouvons écrire que $\vec{A} \sim \vec{B}$ et $\vec{B} \sim \vec{C}$.

D'après la transitivité des préférences, nous devrions alors avoir $\vec{A} \sim \vec{C}$. Or ces deux paniers ne sont pas équivalents puisque $\vec{A}$ et $\vec{C}$ ne sont pas situés sur la même courbe. Deux courbes d'indifférence ne peuvent donc pas se couper.

Nous sommes donc conscients qu'il existe des relations particulières entre les biens qui vont modifier nos attitudes de consommation. C'est le cas notamment des biens complémentaires et des biens de substitutions:

Définitions:

D1. Deux biens sont dits "**biens complémentaires**" si la possession de l'un et de l'autre procure une satisfaction supérieure à la somme des satisfactions des deux biens s'ils étaient pris isolément (super-additivité). Ainsi, il y a complémentarité entre des skis et un forfait sur des remontées mécaniques, entre une voiture et de l'essence. Cela peut être interprété par la courbe d'indifférence suivante:

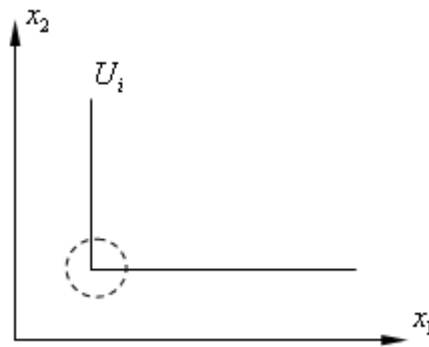


Figure: 66.7 - Principe de représentation de biens complémentaires

Effectivement, pour le couple (voiture, essence) chaque courbe a respectivement un minimum sous lequel nous ne pouvons pas descendre afin que le couple conserve son intérêt de consommation en tant que tel (il ne vaut pas la peine d'acheter une voiture si l'essence tend vers zéro).

D2. Deux biens sont dits "**biens substituables**" si nous pouvons remplacer facilement l'un par l'autre, par exemple en cas de pénurie ou de hausses de prix. Le thé et le café sont des substituts car, à défaut de l'un, nous nous reportons souvent sur l'autre. Cela est encore plus vrai pour deux marques d'une même boisson (Pepsi et Coca). La crise de la vache folle est également un bon révélateur de la substituabilité des produits carnés, avec un report de consommation sur les volailles et l'agneau. Cela peut être interprété par la courbe d'indifférence suivante:

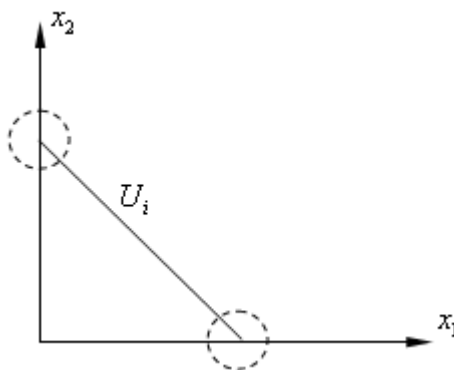


Figure: 66.8 - Principe de représentation de biens substituables

Effectivement, l'intersection avec les axes respectifs indique (exprime) justement la substitution totale possible d'un bien par l'autre dans le panier.

Voyons maintenant un exemple d'application:

Soit à calculer le T.M.S. le long de la courbe d'indifférence $U = x_1 \cdot x_2 = 100$ (il s'agit donc d'une fonction hyperbolique). Nous avons montré que le T.M.S était donné par:

$$T.M.S. = -\frac{dx_1}{dx_2} = \frac{\frac{\partial U}{\partial x_1}}{\frac{\partial U}{\partial x_2}} = \frac{\frac{\partial (x_1 \cdot x_2)}{\partial x_1}}{\frac{\partial (x_1 \cdot x_2)}{\partial x_2}} = \frac{x_2}{x_1} \quad (66.71)$$

Ainsi, pour les trois points A, B, C de coordonnées respectives:

$$A(5, 20); B(10; 10); C(20, 5) \quad (66.72)$$

nous trouvons les valeurs T.M.S. respectives:

$$TMS_A = 4, TMS_B = 1, TMS_C = 0.25 \quad (66.73)$$

Ces valeurs expriment des équivalences entre les biens 2 et 1 pour des variations marginales des quantités de ces biens. Ainsi au point A, pour conserver le niveau d'utilité de 100, le consommateur est prêt à abandonner du bien 2 pour augmenter sa consommation de bien 1 dans un rapport de 4 à 1. Au point B l'équivalence entre les deux biens est dans un rapport de 1 à 1, etc.

Remarque: Le concept de courbes d'indifférences a été développé par Vilfredo Pareto et d'autres dans la première partie du 20ème siècle. Le recours à ce concept a permis à l'analyse économique d'utiliser le concept de préférences dans la détermination des choix plutôt que le concept d'utilité qui a le problème de ne pas pouvoir être mesuré de façon objective.

Pour résumer:

Les courbes d'indifférences (iso-utilité) ont donc traditionnellement les propriétés suivantes:

P1. Les courbes d'indifférences ne se croisent pas. C'est la conséquence de l'hypothèse que les consommateurs préfèrent toujours avoir plus d'un bien plutôt que moins...

P2. Les courbes sont décroissantes. Cela découle de l'hypothèse que plus les individus consomment de moins en moins d'un produit, plus ils en demanderont un autre.

Finalement, les hypothèses sont (les trois premières hypothèses sont obligatoires, les autres sont facultatives puisque découlant des deux propriétés précédentes):

H1. Complétude: les consommateurs connaissent leurs préférences individuelles. En d'autres termes, ils peuvent choisir entre consommer le couple X ou le couple Y. Ils savent si X est préféré à Y, Y est préféré à X, ou s'ils sont indifférents entre consommer X ou Y.

H2. Transitivité: si un consommateur préfère le couple X au couple Y, et préfère le couple Y au couple Z, alors il préférera le couple X au couple Z.

H3. Continuité: cela signifie que nous pouvons choisir de consommer n'importe quelle quantité d'un bien.

H4. Non-satiété: c'est l'idée que plus d'un bien est toujours préféré à moins.

H5. Convexité: la valeur marginale qu'un individu retire de chaque bien décroît. Dans un monde à deux biens, si un consommateur a relativement plus d'un bien, il sera plus heureux avec un peu moins de ce bien et avec un peu plus de l'autre bien.

3.1.1. MODÈLE CONTRARIÉ A PERTE NETTE

Considérons maintenant, et ce indépendamment de la théorie de la préférence, un modèle à monopole contrarié et à information parfaite pour un besoin primaire. Notons $D(t)$ la demande sur le marché et $O(t)$ l'offre. Nous avons alors une demande exponentielle (en absence d'offre):

$$\frac{dD}{dt} = rD \quad (66.74)$$

et en l'absence de demande:

$$\frac{dO}{dt} = -mO \quad (66.75)$$

Les offreurs et les demandeurs sont en interaction. Pour quantifier la contribution entre groupes, nous considérons l'offre en assumant que sa valeur ou intensité est fonction de probabilité de rencontre entre le demandeur et le produit et quelle sera proportionnelle au produit du pourcentage $O \cdot D$ de l'offre de la demande.

Les effets de la découverte du produit n'ont pas les mêmes effets sur les deux groupes offreurs/demandeurs. Premièrement, bien sûr, chaque offre acquise par un demandeur est un gain net pour le premier et sera supposée comme une perte d'épargne nette pour le second. Ainsi, si l'effet des interactions est accepté, comme étant proportionnel à $O \cdot D$ les signes d'influence d'interaction différeront selon:

$$\begin{aligned} \frac{dD}{dt} &= rD - cDO \\ \frac{dO}{dt} &= -mO - bDO \end{aligned} \quad (66.76)$$

Avant d'aller plus loin, cherchons les valeurs pour lesquelles les dérivées s'annulent (ce qui nous donnera au fait le point d'équilibre entre l'offre et la demande):

$$\begin{aligned} rD' - cD'O' &= 0 \\ -mO' + bD'O' &= 0 \end{aligned} \quad (66.77)$$

D'où:

$$\begin{aligned} (r - cO')D' &= 0 \\ (bD' - m)O' &= 0 \end{aligned} \quad (66.78)$$

Une solution triviale est la "**solution d'inexistence**" (appelé aussi parfois "**solution critique**") donné par:

$$D'_e = O'_e = 0 \quad (66.79)$$

Sinon, nous avons aussi comme possible solution:

$$\begin{aligned} r = cO' &\rightarrow O' = \frac{r}{c} \\ bD' = m &\rightarrow D' = \frac{m}{b} \end{aligned} \quad (66.80)$$

Maintenant, normalisons les équations en écrivant (ainsi elles sont sans dimension):

$$\begin{aligned}x &= \frac{D}{D'} = \frac{b}{m} N \rightarrow N = \frac{m}{b} x \\y &= \frac{O}{O'} = \frac{c}{r} P \rightarrow P = \frac{r}{c} y\end{aligned}\quad (66.81)$$

avec cette normalisation, le modèle s'écrit:

$$\begin{aligned}\frac{m}{b} \frac{dx}{dt} &= \frac{mr}{b} x - \frac{cmr}{bc} xy \\ \frac{r}{c} \frac{dy}{dt} &= \frac{bmr}{bc} xy - \frac{mr}{c} y\end{aligned}\quad (66.82)$$

En réarrangeant les coefficients, le système s'écrit finalement (hors de la solution d'inexistence):

$$\begin{aligned}\frac{dx}{dt} &= rx(1-y) \\ \frac{dy}{dt} &= my(x-1)\end{aligned}\quad (66.83)$$

pour lequel les dérivées s'annulent aux points (1,1), qui sera à nouveau l'équilibre de Say.

Le tracé discret de ce système d'équations (dans lequel nous reconnaissons un terme logistique), donne avec $m = r = b = c = 1$ et les conditions initiales $(x_0, y_0) = 4, 1$:

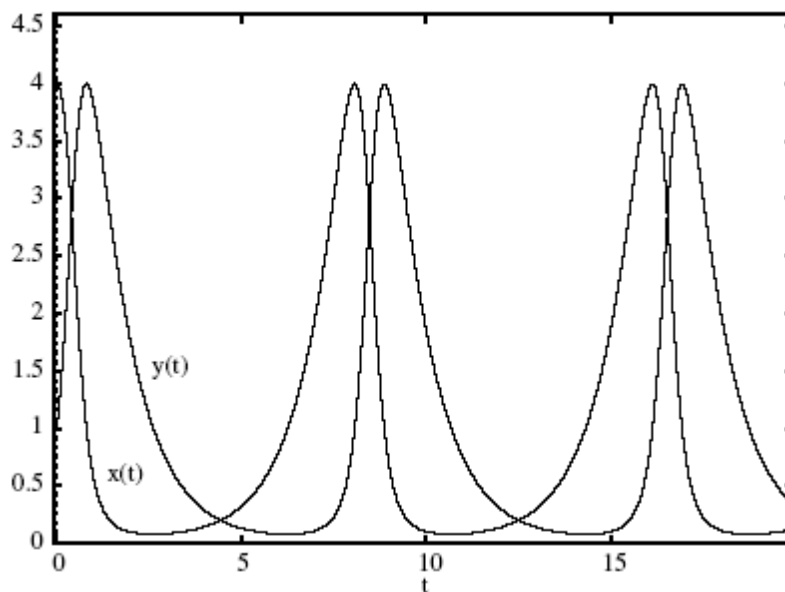


Figure: 66.9 - Cycles d'offre et demande

Nous retrouvons comme le marché semble nous le montrer, des cycles d'offre/demande (certains produits démodés reviennent à la mode) dont il faut déterminer par des statistiques, les conditions initiales afin d'en connaître la possible période. Nous remarquons que l'offre a toujours un peu de retard sur la demande dans ce modèle.

Si nous représentons l'offre et la demande non pas respectivement en fonction du temps mais en fonction de l'un et de l'autre, nous obtenons:

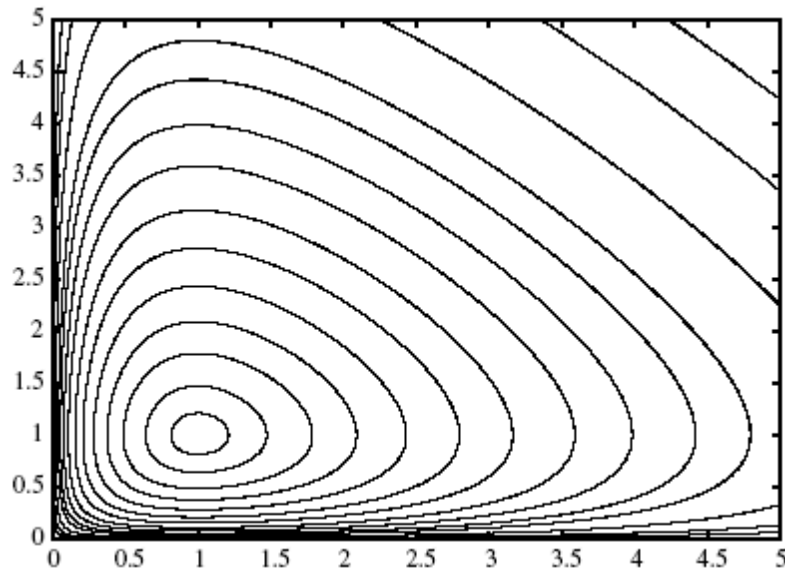


Figure: 66.10 - Représentation dans l'espace des phases des cycles offre/demande

Nous voyons ainsi (dans cette représentation de l'espace des phases) que pour des conditions initiales fixes, le système est périodique et a un point d'équilibre en

Ce qui correspond aux points où:

$$\frac{dx}{dt} = \frac{dy}{dt} = 0 \quad (66.84)$$

Finalement, nous avons deux couples de points d'équilibre (c'est trivial en regardant le système d'équation):

$$(x_1^e, y_1^e) = (0,0) \quad \text{et} \quad (x_2^e, y_2^e) = (1,1) \quad (66.85)$$

La question qui va se poser est le sens de rotation (représentation) du plan des phases. Ainsi, en représentant les directions à l'aide d'un champ de vecteurs, nous obtenons la représentation:

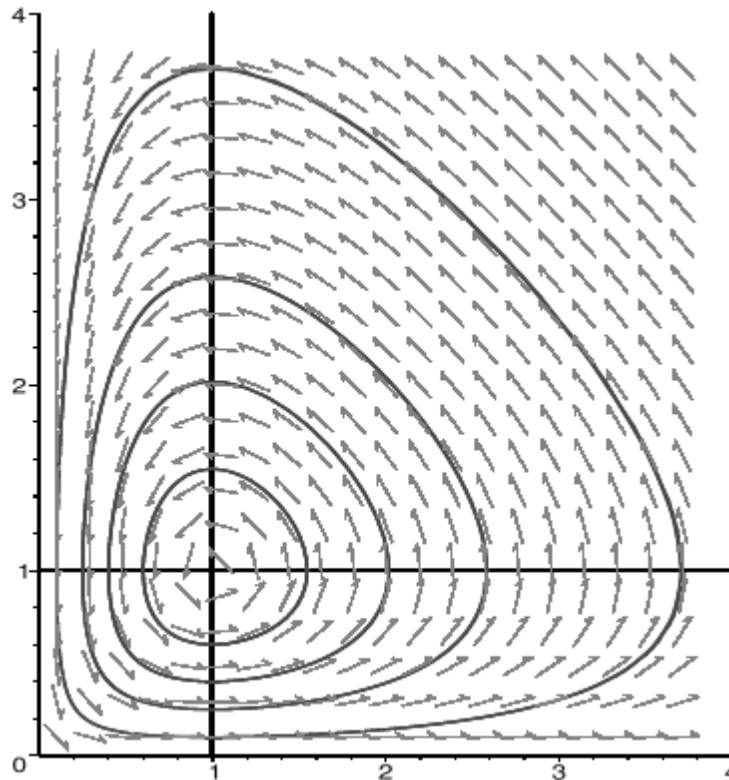


Figure: 66.11 - Sens de rotation de l'espace des phases de l'offre/demande

Pour savoir dans quelle direction nous nous dirigeons dans l'espace des phases à un moment donné, il suffit donc de connaître la dérivée dy/dx (ou réciproquement dx/dy). Ainsi nous avons:

$$\frac{dy}{dx} = \frac{dy}{dt} \frac{dt}{dx} = \frac{dy}{dt} \left(\frac{dx}{dt} \right)^{-1} = \frac{my(x-1)}{rx(1-y)} \quad (66.86)$$

Ceci dit, nous voyons bien sur le diagramme des phases dans sa forme de champ de vecteurs qu'il arrive un moment dans le cycle de ce modèle que l'offre soit très élevée pour une faible demande. Donc le modèle mathématique (théorique) explique bien ce qui peut être a priori contre-intuitif pour l'être humain.

Cependant, nous pouvons (devons) nous poser la question de ce qu'il se passe après une petite perturbation autour des points d'équilibre (ce qui est de la plus haute importance en économie).

Nous avons donc le système de Lotka-Volterra à l'équilibre:

$$\begin{aligned} \frac{dx}{dt} &= rx(1-y) \\ \frac{dy}{dt} &= my(x-1) \end{aligned} \quad (66.87)$$

En y mettant une perturbation infiniment petite, celui-ci s'écrit:

$$\begin{aligned} \frac{dx}{dt} &= r(x+x^*)(1-(y+y^*)) \\ \frac{dy}{dt} &= m(y+y^*)((x+x^*)-1) \end{aligned} \quad (66.88)$$

En négligeant les termes quadratiques, nous obtenons:

$$\begin{aligned}\frac{dx}{dt} &\cong r \left[(1-y^*)x + (1-y-y^*)x^* \right] \\ \frac{dy}{dt} &\cong -m \left[(1-x^*)y + (1-x-x^*)y^* \right]\end{aligned}\quad (66.89)$$

En nous concentrant d'abord sur l'étude proche du point (0,0) nous pouvons négliger les termes xy , nous obtenons:

$$\begin{aligned}\frac{dx}{dt} &\cong r \left[(1-y^*)x + (1-y-y^*)x^* \right] \\ \frac{dy}{dt} &\cong -m \left[(1-x^*)y + (1-x-x^*)y^* \right]\end{aligned}\quad (64.90)$$

Mais c'est encore trop compliqué. Toujours en nous considérant proche du point (0,0), nous considérons un peu abusivement mais astucieusement (sinon nous ne pourrions toujours pas résoudre le problème analytiquement) les approximations suivantes:

$$\begin{aligned}\frac{dx}{dt} &\cong r \left[x - \underbrace{y^*x}_{=0} + x^* - \underbrace{yx^*}_{=0} - \underbrace{y^*x^*}_{=0} \right] \cong r(x+x^*) \underset{x \gg x^*}{\cong} rx \\ \frac{dy}{dt} &\cong -m \left[y - \underbrace{x^*y}_{=0} + y^* - \underbrace{xy^*}_{=0} - \underbrace{x^*y^*}_{=0} \right] \cong -m(y+y^*) \underset{y \gg y^*}{\cong} -my\end{aligned}\quad (64.91)$$

Donc:

$$\begin{aligned}\frac{dx}{dt} &\cong rx \\ \frac{dy}{dt} &\cong -my\end{aligned}\quad (66.92)$$

Ce qui nous montre pour le système à l'équilibre, proche du point d'inexistence, l'offre diminue exponentiellement alors que la demande augmente elle exponentiellement. Ceci a un sens économique: quand il y a peu d'offre (respectivement de demande), la demande croît alors qu'au fur et à mesure que la demande augmente, l'offre croît et se concentre de plus en plus sur la demande (ahhhh la nature...).

Remarque: Dans la littérature nous retrouvons tantôt le signe "-" en haut ou en bas dans le système d'équations précédent. En réalité, la position du signe "-" n'a aucune importance car il s'agit juste du choix de départ dans la dynamique du système.

Proche du point d'équilibre (1,1), nous aurons (hop! nous changeons la position du signe exprès afin de d'insister sur le fait que ce n'est qu'un choix de départ!!!):

$$\begin{aligned}\frac{dx}{dt} &\cong -ry \\ \frac{dy}{dt} &\cong mx\end{aligned}\quad (66.93)$$

Pour résoudre ce système, différencions l'équation:

$$\frac{d^2 x}{dt^2} = -r \frac{dy}{dt} \quad (66.94)$$

et en y injectant dy/dt :

$$\frac{d^2 x}{dt^2} = -rmx \quad (66.95)$$

Nous obtenons donc une petite équation différentielle du deuxième ordre (cf. chapitre de Calcul Différentiel Et Intégral). Dont la solution-type est:

$$x = x_0 e^{-\lambda t} \quad (66.96)$$

En injectant cette solution dans l'équation différentielle, nous obtenons après simplification des exponentielles une simple équation du deuxième degré (cf. chapitre de Calcul Algébrique):

$$\lambda^2 x_0 + rm x_0 = 0 \quad (66.97)$$

Dont la solution est triviale:

$$\lambda_{1,2} = \pm i \sqrt{rm} \quad (66.98)$$

Ainsi, la solution générale de l'équation différentielle est la combinaison linéaire des deux solutions tel que:

$$x(t) = x_0 \left(e^{-i\sqrt{rmt}} + e^{i\sqrt{rmt}} \right) \quad (66.99)$$

Mais nous avons donc:

$$\frac{dx}{dt} = -ry \Rightarrow y(t) = -\frac{1}{r} \frac{dx}{dt} \quad (66.100)$$

Dès lors, connaissant $x(t)$ nous obtenons facilement:

$$y(t) = -x_0 \sqrt{\frac{m}{r}} \left(-ie^{-i\sqrt{rmt}} + ie^{i\sqrt{rmt}} \right) \quad (66.101)$$

Utilisons maintenant la relation d'Euler (cf. chapitre sur les Nombres):

$$e^{ix} = \cos(x) + i \sin(x) \quad (66.102)$$

Ainsi, nous avons:

$$x(t) = x_0 \left[\cos(-\sqrt{rmt}) + i \sin(-\sqrt{rmt}) + \cos(\sqrt{rmt}) + i \sin(\sqrt{rmt}) \right] \quad (66.103)$$

et comme (cf. chapitre de Trigonométrie) $\cos(x) = \cos(-x)$, $-\sin(x) = \sin(-x)$, nous avons alors:

$$x(t) = 2x_0 \cos(\sqrt{rmt}) \quad (66.104)$$

et de manière similaire, nous obtenons pour les prédateurs:

$$\begin{aligned} y(t) &= -x_0 \sqrt{\frac{m}{r}} i \left(-\left(\cos(-\sqrt{rmt}) + i \sin(-\sqrt{rmt}) \right) + \left(\cos(\sqrt{rmt}) + i \sin(\sqrt{rmt}) \right) \right) \\ &= -x_0 \sqrt{\frac{m}{r}} i \left(-\left(\cos(\sqrt{rmt}) - i \sin(\sqrt{rmt}) \right) + \cos(\sqrt{rmt}) + i \sin(\sqrt{rmt}) \right) \\ &= -x_0 \sqrt{\frac{m}{r}} i \left(-\cos(\sqrt{rmt}) + i \sin(\sqrt{rmt}) + \cos(\sqrt{rmt}) + i \sin(\sqrt{rmt}) \right) \\ &= -x_0 \sqrt{\frac{m}{r}} i \left(2i \sin(\sqrt{rmt}) \right) = 2x_0 \sqrt{\frac{m}{r}} \sin(\sqrt{rmt}) \end{aligned} \quad (64.105)$$

soit:

$$y(t) = 2x_0 \sqrt{\frac{m}{r}} \sin(\sqrt{rmt}) \quad (64.106)$$

Ainsi, autour du point d'équilibre (1,1), les perturbations suffisamment petites pour valider la linéarisation (annulation des termes quadratiques) oscillent comme des ellipses (ou cercles) dont les axes sont définis par les deux équations précédentes.

Ce modèle est cependant imparfait car il prend en compte seulement un monopole contrarié à perte nette et à information parfaite. Le fait de considérer la population constante n'est pas trop gênante mais en toute rigueur nous devrions rajouter un terme logistique dans les équations initiales. Il y a encore du travail donc...

4. CAPITALISATION ET ACTUARIAT

Définition: La "**capitalisation**" est le domaine de la mathématique financière qui permet de calculer des valeurs futures à partir de valeurs présentes, alors que le "**calcul actuariel**" permet de déterminer quelle somme doit être prêtée pour obtenir un montant fixé à l'avance.

Dans une dynamique de marché, des acteurs peuvent prêter ou emprunter un capital en contrepartie de quoi ils perçoivent ou respectivement versent un intérêt périodique. Cet intérêt se justifie par la prise de risque que prend le créancier (celui qui prête le capital) relativement au non-remboursement de la totalité ou d'une part du capital initial que doit rembourser le débiteur (celui qui doit rembourser le capital emprunté). D'une autre manière, vu au niveau du marché économique, les emprunts permettent à certains agents économiques de mettre en place des biens en pariant sur le fait que soit ceux-ci créeront l'offre soit que l'offre viendra d'elle-même mais en souhaitant devancer la concurrence.

Lorsqu'un capital est prêté (ou emprunté, c'est selon le point de vue...) dans le but d'accroître une dynamique de marché (la quantité de biens en circulation sur une durée donnée) nous parlons alors "**d'actif financier**", ceci pour faire comprendre que le capital participe à l'activité de l'économie.

Définitions:

D1. Nous appelons "**rendement d'un actif financier prêté**" le rapport relatif de progression donné par:

$$r_f A_t \% = \frac{A_{t+\Delta t} - A_t}{A_t} \cdot 100 \quad (66.107)$$

D2. Nous appelons "**rendement arithmétique d'un investissement**" le rapport relatif:

$$RSI^{Arith} = \frac{V_f - V_i}{V_i} \quad (66.108)$$

où V_i est la valeur initiale de l'investissement et V_f sa valeur finale.

Il suit de cette dernière définition que si un investissement a rapporté 5% la première année et a porté une perte nette de 2% la deuxième année, le "**RSI (Retour Sur Investissement) arithmétique moyen**" est alors de:

$$\overline{RSI}^{Arith} = \frac{1}{n} \sum_{k=1}^n RSI_k^{Arith} = \frac{RSI_1^{Arith} + RSI_2^{Arith}}{2} = \frac{5\% + (-2\%)}{2} = 1.50\% \quad (66.109)$$

Or il est faux d'utiliser la moyenne arithmétique pour ce type de situation car la somme finale obtenue après les deux années est mathématiquement de:

$$V_n^f = V_i \prod_{k=1}^n (1 + RSI_k^{Arith}) \quad (66.110)$$

ce qui donne alors en reprenant l'exemple précédent:

$$V_2^f = V_i (1 + 5\%)(1 - 2\%) \quad (66.111)$$

Donc le rendement moyen réel est par définition le "**RSI géométrique**" tel que:

$$V_i (1 + 5\%)(1 - 2\%) = V_i (1 + RSI_{Geom}) \quad (66.112)$$

c'est-à-dire qu'il s'agit simplement d'une moyenne géométrique (cf. chapitre de Statistiques). Il vient alors:

$$RSI_{Geom} = (1 + 5\%)(1 - 2\%) - 1 = 2.90\% \quad (66.113)$$

ce qui est bien évidemment nettement différent du RSI arithmétique moyen obtenu plus haut!

Remarque: Nous disons d'un actif qu'il a un "**rendement sans risques**" si la valeur future de celui-ci est parfaitement connue.

Soit un actif A_0 qui peut valoir le rendement (optimiste) futur r_{f1} avec une probabilité p_1 et la valeur (pessimiste) r_{f2} avec une probabilité p_2 ou d'autres valeurs r_{fn} avec la probabilité p_n alors l'espérance mathématique du rendement est donnée par:

$$E(r_f A_0) = \sum_{i=1}^n r_{fi} p_i \quad (66.114)$$

Que la somme monétaire soit du type actif ou non, les types de rendements applicables sont identiques et variés. Il en existe cependant de grands classiques qui ne sont pas stochastiques et connus. Pour leur étude, définissons certaines variables:

- C_0 représente le capital initial ou plus techniquement la "**valeur actuelle**" (V.A.) ou "**present value**" (P.V.) en anglais.
- C_n représente le capital final ou "**valeur capitalisée**" (V.C.) ou "**futur value**" (F.V.) en anglais après n périodes temporelles.
- $i\%$ représente le taux appelé plus techniquement "**taux effectif**"
- I_n représente "**l'intérêt**" produit au bout de n périodes (horizon) par la valeur actuelle

Rajoutons encore comme complément la relation:

$$f_c = 1 + i\% \quad (66.115)$$

appelée "**facteur de capitalisation**".

Définition: Nous définissons "**l'intérêt**" comme la rémunération d'un capital (somme d'argent) prêté ou investi pendant un certain temps. L'intérêt peut être payé en une fois ou périodiquement si la durée du prêt ou de l'investissement dure longtemps. L'intérêt peut être payable d'avance (praenumerando) ou à la fin de la période (postnumerando). L'intérêt est fonction de la durée du prêt (ou investissement), du capital emprunté (ou prêté) ainsi que du "taux" d'intérêt pratiqué. La période sur laquelle l'intérêt porte est en général l'année, mais elle peut être plus courte: semestre, trimestre, mois ou jour.

Remarque: Dans un texte, le taux d'intérêt est exprimé normalement en % mais dans les calculs financiers, il est d'usage de calculer sous forme décimale.

4.1. INTERVALLE DE DATES

Pour déterminer le montant d'un intérêt sur un prêt (ou investissement...), il est d'abord indispensable de connaître la durée de ce dernier ou les dates définissant les périodes de paiement d'une obligation (échéancier).

Le calcul des dates et des durées est donc la première étape en mathématiques actuarielles. Si certains logiciels utilisent dans le calcul de la durée l'année civile (365 jours selon le calendrier Grégorien), d'autres se basent sur l'année commerciale (360 jours), ce qui était le cas de la plupart des établissements bancaires (c'est tout à leur avantage financièrement parlant de faire le choix de cette dernière...) avant l'arrivée du calendrier target pour la zone Euro.

Remarques:

R1. Sur les marchés financiers, il existe une seule convention en matière de détermination d'un intervalle de dates pour calculer une durée: le premier jour (date de départ) est inclus dans la période. Le dernier jour (date de fin ou date d'échéance) est exclu de la période. Ainsi une période allant du 15 au 25 juin comporte 10 jours.

R2. Dans le cadre de ce site, qui se veut avoir une approche la plus rigoureuse possible des sujets traités, nous ne nous attarderons pas sur les aberrantes méthodes 30/360 allemande, européenne ou encore américaine (autant faire chaque pays de la planète alors... et se reporter à Microsoft Excel...) pour nous concentrer sur la méthode des 365 jours (système de la base exacte) qui est, et reste, le système le plus naturel de comptage à utiliser puisqu'il tient compte des mois de 28, 29, 30 ou 31 jours.

R3. Signalons qu'en ce qui concerne les carnets d'épargne, les banques se basent sur un système de "quinzaines" (moitiés d'un mois), et estiment qu'il y a donc 24 quinzaines par année.

Il nous faut dès lors dans le système de la base exacte connaître comment calculer le nombre de jours entre deux dates D_2, D_1 donné par le calcul $D_2 - D_1$ chacune de ces dates étant encryptée sous forme normalisée *aaaa-mm-jj* (année-mois-jour) ou sous forme populaire *j.m.a* (jour.mois.année).

Définitions:

D1. Le calendrier Grégorien a été défini tel qu'il ait 12 mois.

D2. Les mois de:

{ janvier(1), mars(3), mai(5), juillet(7), août(8), octobre(10), décembre(12) } (66.116)

sont des mois de 31 jours et les mois de:

{ avril(4), juin(6), septembre(9), novembre(11) } (66.117)

de 30 jours.

D3. Le mois de février est un cas particulier permettant de corriger le fait que l'année civile de 365 jours, ne corresponde pas tout à fait à la période orbitale de la Terre autour du Soleil qui est d'environ 365.25... jours. Ainsi, toutes les années qui sont multiples de 4 ou de 400 sont des années bissextiles (le mois de février à 29 jours au lieu de 28) mais les années qui sont divisibles par 100 ne sont pas bissextiles.

Exemples:

E1. 1992, 1996, 2004, 2008 sont bissextiles.

E2. Les années 1900, 2100, 2200, 2300 ne sont par contre pas bissextiles (car divisibles par 100)

E3. Les années 1600, 2000, 2400, 2800 sont bissextiles car bien que divisibles par 100, elles sont multiples de 400.

Ces définitions et exemples étant donnés, soit une date sous la forme donnée précédemment. Le nombre de jours depuis l'an 1 est:

$$D(j.m.a) = 365(a-1) + E\left[\frac{a-1}{4}\right] - E\left[\frac{a-1}{100}\right] + E\left[\frac{a-1}{400}\right] + 31(m-1) + j \quad (66.118)$$

où $E[x]$ est la partie entière de x . Cette relation se déduit logiquement de la manière suivante pour les dates où $m \leq 2$ (soit uniquement pour les deux premiers mois de l'année incluant la problématique du mois de février):

1. Nous avons $365(a-1)$ car soit a donné, le nombre de jours civils depuis l'an 1 est 365 fois le nombre d'années a diminué d'une unité puisque l'année en cours n'est pas terminée.
2. Même remarque pour les mois avec $31(m-1)$: en considérant des mois de 31 jours, le mois en cours n'étant pas pris en compte (d'où $31(m-1)$). La correction pour les mois de 28 ou 30 jours intervient sous le cas où $m > 2$ que nous allons voir un peu plus bas.
3. Logiquement, nous ajoutons j (qui contient toute l'information quant à savoir si l'année a est bissextile ou non) à la somme des deux termes précédents.
4. Les termes $E\left[\frac{a-1}{4}\right] - E\left[\frac{a-1}{100}\right] + E\left[\frac{a-1}{400}\right]$ donnent quant à eux le nombre de 29 février entre l'année 1 et a en prenant en compte les années bissextiles qui ont lieu tous les multiples de 4 et 400 ans excepté les années qui sont multiples de 100.

Si $m > 2$, nous devons utiliser la relation suivante:

$$D(j.m.a) = 365(a-1) + E\left[\frac{a}{4}\right] - E\left[\frac{a}{100}\right] + E\left[\frac{a}{400}\right] + 31(m-1) + j - E(0.42M + 2 + 0.5) \quad (66.119)$$

Cette relation se déduit toujours de la même manière que la précédente à la différence que certains termes au numérateur ne sont pas diminués d'une unité car ayant $m > 2$, il faut, pour ces termes, prendre en compte l'année en cours dans le calcul.

Le dernier terme $E(0.42M+2+0.5)$ est ici pour corriger le fait que tous les mois n'ont pas 31 jours. Pour l'obtenir, nous construisons le tableau suivant (la troisième colonne donne le décalage en jours par rapport au cas où les mois auraient tous 31 jours):

Mois	N° Mois n	Décalage d
mars	3	3
avril	4	4
mai	5	4
juin	6	5
juillet	7	5
août	8	5
septembre	9	6
octobre	10	6
novembre	11	7
décembre	12	7

Tableau: 66.5 - Décalage mensuel en jours

Une régression linéaire simple donne:

$$d(n) = 0.4242n + 2.018 \quad (66.120)$$

En prenant la valeur entière et en vérifiant bien que la fonction choisie est correcte, nous obtenons finalement bien (en prenant une précision de deux décimales):

$$E(0.42M+2+0.5)$$

Mois	N° Mois n	Décalage d	$d(n)$	$E(d(n))$
mars	3	3	3.26	3
avril	4	4	3.68	4
mai	5	4	4.1	4
juin	6	5	4.52	5
juillet	7	5	4.94	5
août	8	5	5.36	5
septembre	9	6	5.78	6
octobre	10	6	6.2	6
novembre	11	7	6.62	7
décembre	12	7	7.04	7

Tableau: 66.6 - Comparaison entre la fonction de régression et l'objectif

4.2. ÉQUIVALENCES DE TAUX

Intéressons-nous maintenant brièvement au calcul des taux avant de s'attaquer directement aux calculs des différents et nombreux types d'intérêts.

Définition: Le "**taux proportionnel**" fait apporter à un même capital, durant la même période, le même "intérêt simple" (voir la définition de l'intérêt simple plus bas) et est donc donné par la relation:

$$t_p \% = \frac{t\%}{n} \quad (66.121)$$

Si le taux proportionnel est calculé sur la base d'une année, nous parlons alors de "**taux de rendement annualisé**"; s'il est calculé sur la base d'un mois, nous parlons alors de "**taux de rendement mensualisé**".

Exemple:

Calculer le taux mensuel t_p % proportionnel (soit: le taux de rendement mensualisé) à un taux annuel t % de 12%:

$$t_p \% = \frac{12\%}{12 \text{ mois}} = 1\% \quad (66.122)$$

Définition: Le "**taux équivalent**" fait apporter à un même capital, durant la même période, le même "intérêt composé" (voir la définition de l'intérêt composé plus bas) et est donc donné par la relation:

$$\begin{aligned} (1 + t_p \%)^m &= (1 + t\%)^n \\ \Rightarrow (1 + t_p \%) &= \sqrt[n]{(1 + t\%)^n} \quad (66.123) \\ \Rightarrow t_p \% &= (1 + t\%)^{\frac{n}{m}} - 1 \end{aligned}$$

et inversement:

$$t\% = (1 + t_p \%)^{\frac{m}{n}} - 1 \quad (66.124)$$

Exemple:

Taux t % mensuel équivalent à un taux t_p % annuel de 12% (résultat tronqué au dix-millième):

$$t = (1 + 0.12)^{\frac{1}{12}} - 1 = 0.9488\% \quad (66.125)$$

la procédure inverse consisterait donc à calculer le taux annualisé et nous voyons alors qu'un taux mensuel de 1% annualisé vaudrait plus que 12%.

4.3. INTÉRÊT SIMPLE

Définition: "L'intérêt simple" est défini par la relation (voir plus haut pour la définition des notations):

$$I_n = C_0 \cdot i\% \cdot n \quad (66.126)$$

qui implique une capitalisation (valeur finale):

$$C_n = C_0 (1 + i\% \cdot n) \quad (66.127)$$

Il s'agit simplement de l'intérêt qui est calculé à chaque période sur la seule base du capital prêté ou emprunté à l'origine.

Remarques:

R1. Il est très facile à partir de la connaissance de trois des quatre paramètres de la relation précédente de retrouver la quatrième. S'agissant d'une simple équation du premier degré, nous ne nous attarderons pas sur ce genre d'exercice de style d'algèbre élémentaire.

R2. Une particularité de l'intérêt simple est d'être proportionnel à la durée du placement. Si l'intérêt par exemple sur une année est de 12%, le "taux équivalent" d'un placement identique pendant 12 mois sera de 1% par mois. Cette propriété n'est pas vraie pour l'intérêt composé comme nous le verrons de suite après.

R3. Pour les carnets d'épargne, nous avons déjà fait mention que les instituts financiers utilisent la quinzaine comme période temporelle (soit 24 périodes dans l'année composée de mois de 30 jours). Donc pour calculer l'intérêt annuel, lors de chaque quinzaine, ils prennent le solde le plus faible sur le compte au cours de la quinzaine et calculent l'intérêt simple sur un taux rapporté à 24 quinzaines par année et reporteront le résultat obtenu lors de la clôture annuelle du compte à la fin de l'année (ils sont pas fous...)

Par ailleurs, si plusieurs placements à intérêt simple sont effectués simultanément pour des durées et à des taux différents, nous pouvons être amenés à calculer le taux moyen T de l'ensemble de ces placements.

Si nous notons C_t le placement numéro t , i_t le taux d'intérêt du placement numéro t , n_t la durée du placement numéro t et k le nombre de placements, nous obtenons la moyenne arithmétique pondérée (cf. chapitre de Statistiques) comme suit:

$$T = \frac{C_1 i_1 n_1 + C_2 i_2 n_2 + \dots + C_k i_k n_k}{C_1 n_1 + C_2 n_2 + \dots + C_k n_k} = \frac{\sum_{t=1}^k C_t i_t n_t}{\sum_{t=1}^k C_t n_t} \quad (66.128)$$

4.3.1. ESCOMPTE

Toujours relativement à l'intérêt simple, nous pouvons revenir sur une notion dont nous avons parlé au début de ce chapitre: l'escompte.

Rappelons que l'escompte est une remise accordée à un acheteur par un vendeur dans le but de l'inciter à payer rapidement avant n unités (périodes) de temps (c'est donc le nombre d'unités n qui importe!). Un acheteur devrait en principe profiter de cet escompte (qui peut donc être vu comme un crédit). Dans le cas contraire, c'est comme si l'acheteur empruntait implicitement à une banque pendant une durée donnée (les n unités).

Historiquement, l'escompte est la retenue que fait un banquier sur une promesse de versement que possède un commerçant (promesse signée par un acheteur pour une certaine échéance), que le banquier verse avant l'échéance (donc comme une avance).

Voyons cela:

Notons C_0 la valeur actuelle escompte compris (déduit), C_n le montant sans escompte appelé "**valeur nominale**", n la durée rapportée à l'échelle de temps du taux d'escompte, $t\%$ le taux d'escompte et i l'intérêt implicite simple en cas de renonciation à l'escompte.

Nous avons maintenant les relations triviales suivantes:

$$C_0 = C_n - t\% C_n = C_n (1 - t\%) \quad (66.129)$$

Et rien nous empêche d'écrire que le montant sans escompte peut être obtenu par un intérêt simple de la forme:

$$C_n = C_0 (1 + n \cdot i) \quad (66.130)$$

à partir de la valeur avec escompte. L'intérêt implicite simple sur la valeur actuelle étant alors trivialement:

$$i = \frac{C_n - C_0}{nC_0} \quad (66.131)$$

Dès lors, il vient par substitution:

$$i = \frac{C_n - C_0}{nC_0} = \frac{C_n - C_n (1 - t\%)}{nC_n (1 - t\%)} = \frac{1 - (1 - t\%)}{n(1 - t\%)} = \frac{t\%}{n(1 - t\%)} \quad (66.132)$$

Nous remarquons alors qu'il suffit de connaître le taux d'escompte accordé $t\%$ (souvent donné en annuel...) ainsi que la durée de renonciation n pour déterminer le taux équivalent simple du crédit accordé.

Exemple:

Calculons le taux implicite simple i relatif à un escompte de 1% à 10 jours ou net à 30 jours:

$$i = \frac{0.01}{20 (1 - 0.01)} \cong 0.0505\% \quad (66.133)$$

Ainsi, cet escompte s'il n'est pas pris en considération, peut être vu comme un crédit à 0.505% par jour pendant 20 jours sur la somme avec escompte! Ce qui correspond sur un mois de 30 jours à un taux d'intérêt simple de 1.515%.

Cette méthode de calcul est appelée "**escompte commercial**" car les calculs se font sur la base de la valeur nominale et non de la valeur actuelle.

Cependant, historiquement, les calculs étaient plutôt du type suivant:

Exemple:

Le 19 mars, un banquier escompte $t\%$ de 6% (annuel) un contrat d'acheteur à un commerçant de 240'000.- payable le 31 mai (il y a donc 73 jours calendaires entre les deux dates). La valeur de l'escompte est alors:

$$\text{Escompte} = 240'000 \cdot \frac{6\%}{360} \cdot 73 = 2'920.- \quad (66.134)$$

Donc la valeur actuelle du contrat sera de 237'080. Le taux implicite simple sera alors donné par:

$$C_n = C_0 (1 + n \cdot i) = (C_n - C_n t\% \cdot n) (1 + n \cdot i) \Rightarrow i = \frac{\frac{1}{(1 - t\% \cdot n)} - 1}{n} = \frac{t\%}{(1 - t\% \cdot n)} \quad (66.135)$$

Soit:

$$i = \frac{t\%}{(1 - t\% \cdot n)} = \frac{\frac{6\%}{360}}{\left(1 - \frac{6\%}{360} \cdot 73\right)} \cong 0.0169\% \quad (66.136)$$

4.4. INTÉRÊT COMPOSÉ

Définition: "**L'intérêt composé**" est donné par la relation:

$$I_n = C_0 f_c^n - C_0 = C_0 (f_c^n - 1) \quad (66.137)$$

et implique que la valeur finale est donnée par:

$$C_n = C_0 f_c^n = C_0 (1 + t\%)^n \quad (66.138)$$

Nous disons donc que le taux d'intérêt est "composé" lorsqu'à la fin de chaque période l'intérêt est ajouté au capital pour le calcul de la prochaine période. Cette relation implique bien évidemment qu'il n'y a ni retrait ni dépôt pendant toutes les périodes.

Nous avons par ailleurs les relations triviales (cf. section d'Algèbre):

$$C_0 = \frac{C_n}{(1 + t\%)^n} \quad n = \frac{\ln\left(\frac{C_n}{C_0}\right)}{\ln(1 + t\%)} \quad t\% = \sqrt[n]{\frac{C_n}{C_0}} - 1 \quad (66.139)$$

où la première relation donne donc la valeur initiale si la valeur finale est connue et qu'il n'y a ni retrait ni dépôt pendant toutes les périodes.

Remarque: Les relations équivalentes dans la version française de Microsoft Excel 11.8346 pour trouver C_n, C_0, n, i sont respectivement (fonctions en français) VC(), VA(), NPM(), Taux() l'abréviation NPM signifiant "nombre paiements mensuels".

Dans le cadre des intérêts cumulés (composés), deux notions importantes sont donc la "valeur actuelle" et la "valeur finale" acquise d'un capital.

En répondant à la question: "Quel capital obtenons-nous au bout d'un certain temps en plaçant aujourd'hui une somme X sur un carnet d'épargne?", nous faisons une recherche de valeur finale ou acquise d'un capital. Nous parlons alors "**d'opération de capitalisation**".

Par contre, si nous nous demandons: "Quel capital devons-nous placer aujourd'hui sur un carnet d'épargne pour obtenir au bout d'un certain temps un capital X ?", nous faisons une recherche de valeur actuelle d'un capital. Nous parlons alors "**d'opération d'escompte**" (c'est le propre du "**calcul actuariel**").

Définition: Nous appelons f_c le "**facteur de capitalisation**" et f_e le "**facteur d'escompte**" définis par les relations:

$$f_c = 1 + i\% \quad f_e = \frac{1}{1 + i\%} \quad (66.140)$$

ce qui nous amène par ailleurs à avoir:

$$f_c = f_e^{-1} \quad (66.141)$$

Partant de la relation de capitalisation composée vue plus haut (mais donc avec la notation condensée):

$$C_n = C_0 f_c^n \quad (66.142)$$

Le capital initial C_0 peut être exprimé avec le facteur d'escompte:

$$C_0 = \frac{C_n}{f_c^n} = C_n \frac{1}{f_c^n} = C_n \left(\frac{1}{f_c^n} \right) = C_n f_e^n \quad (66.143)$$

Cela rend alors très simple le calcul d'actualisation ou de capitalisation puisqu'il s'agit de multiplier le capital final ou initial par le facteur d'escompte ou de capitalisation élevé à la puissance n .

Rappelons maintenant la relation que nous avons obtenue lors de notre présentation initiale des taux équivalents:

$$t_p \% = (1 + i\%)^{\frac{n}{m}} - 1 \quad (66.144)$$

Souvent afin de se simplifier le calcul, la personne qui cherche le taux équivalent va se contenter de poser (normaliser) $n = 1$. Ce qui l'amène à écrire:

$$t_p \% = (1 + t \%)^{\frac{1}{p}} - 1 \text{ et } t \% = (1 + t_p \%)^m - 1 \quad (66.145)$$

Vient alors une petite astuce du financier qui fait intervenir dans ses démarches de ventes les concepts de "**taux effectif**" (déjà vu!) et "**taux nominal**". Le taux nominal étant toujours inférieur au taux effectif, il permet à l'émetteur de l'emprunt d'afficher un taux inférieur à ce qu'il est réellement (pratique qui au demeurant est interdite par voie législative dans certains pays!).

Exemple:

Imaginons que les conditions d'un prêt soient les suivantes: intérêt annuel de $t_n \% = 12\%$ (taux nominal) payable par tranches mensuelle de $t_p \% = t_n \% / 12 = 1\%$. Un individu attentif se rend compte que payer 1% tous les mois dans un système d'intérêts composés ne donne pas un intérêt annuel de 12% mais de:

$$t \% = (1 + t_p \%)^p - 1 = \left(1 + \frac{t_n \% }{p}\right)^p - 1 = \left(1 + \frac{t_n \% }{12}\right)^{12} - 1 = 12.682\% \dots \quad (66.146)$$

qui est le taux effectif $t\%$! Pas forcément gagnant donc...

Maintenant, si plusieurs placements sont effectués simultanément pour des durées et à des taux différents, nous pouvons être amenés à calculer le taux moyen T de l'ensemble de ces placements.

Si nous notons C_t le placement numéro t , $f_{c,t}$ le facteur de capitalisation du placement numéro t , n_t la durée du placement numéro t , k le nombre de placements et finalement T le taux moyen de l'ensemble des placements nous pouvons à l'aide du calcul formel jusqu'au quatrième degré (voir chapitre de Calcul Algébrique) ou au-delà avec l'analyse numérique (prendre le solveur de Microsoft Excel par exemple), résoudre l'équation:

$$C_1 f_{c,1}^{n_1} + C_2 f_{c,2}^{n_2} + \dots + C_k f_{c,k}^{n_k} = C_1 (1 + T)^{n_1} + C_2 (1 + T)^{n_2} + \dots + C_k (1 + T)^{n_k} \quad (66.147)$$

Si nous faisons un changement de variables $1 + T = x$ nous avons alors à résoudre l'équation de n_k inconnues en x (tous les autres termes étant normalement connus dans l'énoncé du problème):

$$C_1 x^{n_1} + C_2 x^{n_2} + \dots + C_k x^{n_k} - C_1 f_{c,1}^{n_1} - C_2 f_{c,2}^{n_2} - \dots - C_k f_{c,k}^{n_k} = 0 \quad (66.148)$$

4.5. INTÉRÊT CONTINU

Rappelons que l'intérêt composé est défini en utilisant le taux effectif:

$$C_n = C_0 e^{n \% } = C_0 (1 + t \%)^n \quad (66.149)$$

Avec le taux nominal, nous écrivons alors:

$$C_n = C_0 \left(1 + \left[\left(1 + \frac{t_n \% }{p}\right)^p - 1\right]\right)^n \quad (66.150)$$

Nous pouvons maintenant nous demander ce qu'il adviendrait du taux effectif $t\%$ si l'intérêt était versé non pas mensuellement, ni quotidiennement, mais en continu, d'une manière instantanée (ou quasi-instantanée). Nous écrivons alors (cf. chapitre d'Analyse Fonctionnelle):

$$\lim_{p \rightarrow +\infty} \left(1 + \frac{t_n \%}{p} \right)^p - 1 = e^{t_n \%} - 1 \quad (66.151)$$

Ainsi, en cas de capitalisation continue, la fonction de capitalisation s'écrit finalement:

$$C_n = \lim_{p \rightarrow +\infty} C_0 \left(1 + \left[\left(1 + \frac{t_n \%}{p} \right)^p - 1 \right] \right)^n = C_0 \left(1 + [e^{t_n \%} - 1] \right)^n = C_0 (e^{t_n \%})^n = C_0 e^{n t_n \%} \quad (66.152)$$

Et comme:

$$(1 + t\%)^n = e^{n t \%} \Rightarrow 1 + t\% = e^{t_n \%} \quad (66.153)$$

Nous avons alors (relation relativement importante en finance des marchés):

$$t_n \% = \ln(1 + t\%) \quad (66.154)$$

Il peut être intéressant que le lecteur compare le résultat de ce calcul avec celui de l'intérêt composé classique dans un tableur. Il pourra dès lors constater que le taux continu est une approximation du taux composé qui sous-évalue toujours ce dernier (ce qui est normal puisque l'argent travaille d'autant plus que la fréquence est grande).

Une propriété utile et très utilisée en trading de l'intérêt continu est qu'il est "consistant dans le temps". Pour voir de quoi il s'agit, considérons les 3 valeurs suivantes d'un capital de portefeuille placé quelconque:

Capital	Valeur	Taux de rentabilité (simple)	Taux de rentabilité (continu)
C_{n-2}	10.00.-	-	-
C_{n-1}	14.00.-	$= (14-10)/10 = 40\%$	$= \ln\left(\frac{14}{10}\right) \cong 33.65\%$
C_n	9.00.-	$= (9-14)/14 = -35.71\%$	$= \ln\left(\frac{9}{14}\right) \cong -44.18\%$
Somme:		$= 4.29\%$	$= -10.54\%$
Comparaison:		$= (9-10)/10 = 10\%$	$= \ln\left(\frac{9}{10}\right) \cong 10.54\%$

Tableau: 66.7 - Comparaison entre la fonction de régression et l'objectif

Donc comme nous pouvons le voir, contraire au taux de renabilité simple, le taux de rentabilité continu est constant en temps car nous pouvons sommer les différents taux pour avoir la variation totale. Ce résultat provient simplement de la propriété suivante:

$$C_n = (C_{n-2} e^{t_1\%}) e^{t_2\%} = C_{n-2} e^{t_1\%} e^{t_2\%} = C_{n-2} e^{t_1\% + t_2\%} \Rightarrow \ln \left(\frac{C_n}{C_{n-2}} \right) = t_1\% + t_2\% \quad \text{Tableau: 66.8}$$

4.6. INTÉRÊT PROGRESSIF (RENTES)

Définition: Une "rente" ou "annuité" est une suite de paiements versés périodiquement à intervalles de temps réguliers et durant une période fixée d'avance à intérêt composé (typique des deuxième ou troisième piliers en Suisse).

Il suffit alors d'appliquer la relation (voir plus haut la démonstration) $C_0 = C_n f_c^n$ à chaque terme de rente versé si nous souhaitons connaître la valeur actuelle de cette rente.

Par contre, si nous souhaitons obtenir la valeur finale d'une rente, nous appliquerons à chaque terme la relation (voir plus haut la démonstration):

$$C_n = C_0 f_c^n \quad (66.155)$$

Définition: Si la rente est payable en fin de période, elle est dite "rente postnumerando". Par contre, si elle est payable en début de période, elle est dite "rente praenumerando".

Remarques:

R1. Les rentes qui sont toujours payées sont appelées "rentes certaines" et lorsque la durée est fixée d'avance, nous parlons de "rentes temporaires".

R2. Les rentes versées sur la base de la durée de vie d'un individu sont appelées "rentes viagères".

Puisque les termes sont souvent supposés constants, nous avons pour habitude de baser les calculs sur la valeur d'une unité monétaire. Ainsi, nous notons (les notations adoptées sont celles que nous trouvons dans la littérature car bien que peu pratiques, elles sont originales et jolies à regarder...):

- $\alpha_{\overline{n}|}$ la valeur actuelle d'une rente d'une unité monétaire payable postnumerando (à terme échu) pendant une durée de n périodes
- $\ddot{\alpha}_{\overline{n}|}$ la valeur actuelle d'une rente d'une unité monétaire payable praenumerando (d'avance) pendant une durée de n périodes
- $s_{\overline{n}|}$ la valeur finale d'une rente d'une unité monétaire payable postnumerando pendant une durée de n périodes
- $\ddot{s}_{\overline{n}|}$ la valeur finale d'une rente d'une unité monétaire payable praenumerando pendant une durée de n périodes

Les relations qui déterminent ces valeurs font usage des propriétés des séries géométriques et de leur somme partielle (cf. chapitre de Suites Et Séries).

4.6.1. RENTES POSTNUMERANDO

À terme constant, pour calculer la valeur finale d'une rente à échéance/postnumerando, nous pouvons donc travailler uniquement sur le facteur de capitalisation en multipliant au final par le montant de la rente.

Exemple:

Nous souhaitons calculer la valeur finale d'une rente postnumerando de 3'500.- versée durant 10 périodes et calculée au taux d'intérêt périodique de 6%.

Les versements ont lieu aux dates 1, 2,.. et 10. Le versement $C_0 = 3'500$ de la date 1 a pour valeur acquise à la date 10: $C_0 (1 + t\%)^9$. De même, le versement de la date 2 rapporte des intérêts pendant 8 ans. Sa valeur acquise date 10 est donc: $C_0 (1 + t\%)^8$ etc. Finalement le versement de la date 10 (que nous venons de déposer à la banque) a pour valeur C_0 . La valeur acquise des 10 versements est donc, en posant $f_c = 1 + t\%$ (nous démontrerons les simplifications juste après):

$$\begin{aligned} C_0 (f_c^9 + f_c^8 + \dots + f_c^1 + f_c^0) &= C_0 (f_c^9 + f_c^8 + \dots + f_c^1 + 1) = C_0 \frac{(1 - f_c^{10})}{(1 - f_c)} = C_0 \frac{f_c^{10} - 1}{f_c - 1} \\ &= C_0 \frac{f_c^{10} - 1}{t\%} = 46'132.80. - \end{aligned} \quad (66.156)$$

Donc la rente postnumerando est un versement à termes constants et à taux constant durant un nombre de périodes données conduisant à une suite géométrique simple.

Rappelons donc (encore une fois!) que:

$$C_n = C_0 f_c^n \quad (66.157)$$

Sous la forme de rente postnumerando à termes constants, nous avons alors sous forme générale:

$$C_n = C_0 f_c^0 + C_0 f_c^1 + C_0 f_c^2 + \dots + C_0 f_c^{n-1} \quad (66.158)$$

Ce qui s'écrit:

$$C_n = C_0 (f_c^0 + f_c^1 + f_c^2 + \dots + f_c^{n-1}) = C_0 s_{\overline{n}|} \quad (66.159)$$

donc:

$$s_{\overline{n}|} = \sum_{t=0}^{n-1} f_c^t \quad (66.160)$$

En fait, nous avons donc une suite géométrique de raison q (cf. chapitre de Suites Et Séries). Dès lors:

$$\begin{aligned}
 S_n &= a_0 + a_1 + \dots + a_{n-1} \\
 - qS_n &= qa_0 + qa_1 + \dots + qa_{n-1} = a_1 + a_2 + \dots + a_n \\
 \hline
 S_n - qS_n &= a_0 - a_n \Rightarrow S_n = \frac{a_0 - a_n}{(1-q)} = a_0 \frac{(1-q^n)}{(1-q)}
 \end{aligned} \quad (66.161)$$

et donc:

$$s_{\overline{n}|} = \sum_{t=0}^{n-1} f_c^t = f_c^0 \cdot \frac{(1-f_c^n)}{(1-f_c)} = 1 \frac{1-f_c^n}{1-f_c} = \frac{f_c^n - 1}{f_c - 1} \quad (66.162)$$

Finalement:

$$s_{\overline{n}|} = \frac{f_c^n - 1}{i\%} \quad (66.163)$$

Nous avons donc pour notre exemple dix périodes (et dix termes donc avec $n = 10$):

$$C_{10} = 3'500 s_{\overline{10}|} = 3'500 \frac{f_c^{10} - 1}{i\%} = 3'500 \frac{1.06^{10} - 1}{0.06} = 46'132.80.-$$

Ce capital correspond donc à la somme acquise au bout de dix périodes.

La méthode de calcul de la "valeur actuelle d'une rente à l'échéance/postnumerando" fonctionne sur le même principe mais à l'envers selon la relation démontrée plus haut:

$$C_0 = C_n f_c^n \quad (66.164)$$

Donc si les termes (montants versés) sont constants nous pouvons écrire:

$$C_0 = C_n (f_c^1 + f_c^2 + f_c^3 + \dots + f_c^n) = C_n a_{\overline{n}|} \quad (66.165)$$

donc:

$$a_{\overline{n}|} = \sum_{t=1}^n f_c^t \quad (66.166)$$

Or:

$$\begin{aligned}
 S_n &= a_1 + a_2 + \dots + a_n \\
 - qS_n &= qa_1 + qa_2 + \dots + qa_n = a_2 + a_3 + \dots + a_{n+1} \\
 \hline
 S_n - qS_n &= a_1 - a_{n+1} \Rightarrow S_n = \frac{a_1 - a_{n+1}}{(1-q)} = a_1 \frac{(1-q^n)}{(1-q)}
 \end{aligned} \quad (66.167)$$

alors:

$$a_{\overline{n}|} = \sum_{t=1}^n f_e^t = f_e^1 \cdot \frac{(1 - f_e^n)}{(1 - f_e)} = f_e \frac{1 - f_e^n}{1 - \frac{1}{f_c}} = \frac{1}{f_c} \frac{1 - f_e^n}{\frac{f_c - 1}{f_c}} = \frac{1 - f_e^n}{f_c - 1} \quad (66.168)$$

finalemt:

$$a_{\overline{n}|} = \frac{1 - f_e^n}{t^0\%} \quad (66.169)$$

Remarque: La valeur $C_0 = C_n a_{\overline{n}|}$ correspondant donc au montant qu'il faudrait placer sur un carnet d'épargne à $t\%$ afin de pouvoir y faire un retrait périodique C_n constant durant les n périodes et ainsi solder le compte.

Nous avons également les relations entre les rentes postnumerando actuelle et finale:

$$\left. \begin{aligned} a_{\overline{n}|} &= \frac{1 - f_e^n}{t^0\%} \\ s_{\overline{n}|} &= \frac{f_c^n - 1}{t^0\%} \end{aligned} \right\} a_{\overline{n}|} = f_e^n s_{\overline{n}|} \quad (66.170)$$

Nous avons également les opérations en chaîne suivantes:

$$\begin{aligned} a_{\overline{n+k}|} &= a_{\overline{n}|} + f_e^n a_{\overline{k}|} = f_e^k a_{\overline{n}|} + a_{\overline{k}|} \\ s_{\overline{n+k}|} &= s_{\overline{n}|} + f_c^n s_{\overline{k}|} = f_c^k s_{\overline{n}|} + s_{\overline{k}|} \end{aligned} \quad (66.171)$$

Remarque: Il est clair étant donné $a_{\overline{n}|}, f_e^n, t^0\%$ connus que $n = \ln(1 - t^0\% a_{\overline{n}|}) / \ln(f_e)$ et ainsi de suite pour les autres types de rente.

4.6.2. RENTES PRAENUMERANDO

La méthode de calcul de la "**valeur actuelle d'une rente à l'avance/praenumerando**" fonctionne sur le même principe que la dernière toujours en utilisant la relation $C_0 = C_n f_e^n$. mais cette fois les termes de la suite géométrique changent puisque le paiement se fait à l'avance:

$$C_0 = C_n (f_e^0 + f_e^1 + f_e^2 + \dots + f_e^{n-1}) = C_n \ddot{a}_{\overline{n}|} \quad (66.172)$$

donc:

$$\ddot{a}_{\overline{n}|} = \sum_{t=0}^{n-1} f_e^t \quad (66.173)$$

Or:

$$\begin{aligned}
 S_n &= a_0 + a_1 + \dots + a_{n-1} \\
 - \quad qS_n &= qa_0 + qa_1 + \dots + qa_{n-1} = a_1 + a_2 + \dots + a_n \\
 S_n - qS_n &= a_0 - a_n \Rightarrow S_n = \frac{a_0 - a_n}{(1-q)} = a_0 \frac{(1-q^n)}{(1-q)} \quad (66.174)
 \end{aligned}$$

alors:

$$\ddot{a}_{\overline{n}|} = \sum_{t=0}^{n-1} f_e^t = f_e^0 \cdot \frac{(1-f_e^n)}{(1-f_e)} = 1 \frac{1-f_e^n}{1-\frac{1}{f_e}} = \frac{1-f_e^n}{\frac{f_e-1}{f_e}} \quad (66.175)$$

finalemt:

$$\boxed{\ddot{a}_{\overline{n}|} = \frac{1-f_e^n}{\ell^{\%}_0 \cdot f_e}} \quad (66.176)$$

Remarque: Pour le même nombre de périodes et le même taux, nous avons $\ddot{a}_{\overline{n}|} > a_{\overline{n}|}$ car $\ddot{a}_{\overline{n}|} = f_e a_{\overline{n}|}$.

La méthode de calcul de la "valeur finale d'une rente à l'avance/praenumerando" fonctionne sur le même principe que la dernière toujours en utilisant la relation $C_n = C_0 f_e^n$. mais cette fois les termes de la suite géométrique changent puisque le paiement se fait à l'avance:

$$C_n = C_0 (f_e^1 + f_e^2 + \dots + f_e^n) = C_0 \ddot{s}_{\overline{n}|} \quad (66.177)$$

donc:

$$\ddot{s}_{\overline{n}|} = \sum_{t=1}^n f_e^t \quad (66.178)$$

Or:

$$\begin{aligned}
 S_n &= a_1 + a_1 + \dots + a_n \\
 - \quad qS_n &= qa_1 + qa_2 + \dots + qa_n = a_2 + a_3 + \dots + a_{n+1} \\
 S_n - qS_n &= a_1 - a_{n+1} \Rightarrow S_n = \frac{a_1 - a_{n+1}}{(1-q)} = a_1 \frac{(1-q^n)}{(1-q)} \quad (66.179)
 \end{aligned}$$

alors:

$$\ddot{s}_{\overline{n}|} = \sum_{t=1}^n f_e^t = f_e^1 \cdot \frac{(1-f_e^n)}{(1-f_e)} = \frac{f_e^n - 1}{\frac{f_e-1}{f_e}} = \frac{f_e^n - 1}{\frac{\ell^{\%}_0}{1+\ell^{\%}_0}} \quad (66.180)$$

finalemt en notant $d = \frac{\ell^{\%}_0}{1+\ell^{\%}_0}$ nous avons:

$$\ddot{s}_{\overline{n}|} = \frac{f_c^n - 1}{d} \quad (66.181)$$

Remarques:

R1. Avec la même notation nous avons par ailleurs la valeur actuelle de la rente praenumerando qui

s'écrit $\ddot{a}_{\overline{n}|} = \frac{1 - f_c^n}{d}$

R2. Pour le même nombre de périodes et le même taux, nous avons $\ddot{s}_{\overline{n}|} > s_{\overline{n}|}$ car $\ddot{s}_{\overline{n}|} = f_c s_{\overline{n}|}$.

4.7. ARRONDIS

Pour arrondir un nombre x au multiple de $1/n$ le plus proche la relation à utiliser est la suivante:

$$f(x, n) = \frac{E[nx + 0.5]}{n} \quad (66.182)$$

La démonstration est intuitive. Il suffit de s'imaginer l'axe des réels et de couper celui-ci en $1/n$ petits intervalles. Soit alors un nombre x donné, le nombre de ces intervalles dans x sera donné par:

$$\frac{x}{\left(\frac{1}{n}\right)} = nx \quad (66.183)$$

Enfin pour savoir quel est le nombre strictement inférieur au multiple recherché, nous prenons la valeur entière de la dernière relation et la multiplions par $1/n$ tel que:

$$f_{\text{inf}}(x, n) = E[nx] \frac{1}{n} = \frac{E[nx]}{n} \quad (66.184)$$

Si cependant, nous souhaitons avoir le nombre arrondi au multiple le plus proche, nous voyons alors qu'il faut rajouter 0.5 tel que:

$$f(x, n) = \frac{E[nx + 0.5]}{n} \quad (66.185)$$

5. EMPRUNTS

Les individus et les entreprises recourent souvent à l'emprunt (crédit) comme moyen financier. Nous allons ici définir les principaux types d'emprunts rencontrés dans la pratique ainsi que les relations qui les caractérisent.

Définition: Nous appelons "**emprunts indivis**", un emprunt ne comportant qu'un seul prêteur, en général, un établissement financier.

Les principaux points concernant les emprunts sont:

- Connaître l'état de la dette à tout moment
- Connaître le montant à rembourser à chaque période
- Connaître l'intérêt dû à chaque période

Définition: Nous appelons "**annuités**", les paiements effectués dans le cadre de l'apurement d'un emprunt. Une annuité comprend une part de remboursement R appelée aussi "**amortissement financier**" et une part d'intérêt I selon la relation:

$$\text{Annuité} = \text{Amortissement} + \text{Intérêts}$$

$$A = R + I$$

La décomposition de l'annuité en amortissements et intérêts est une notion importante non seulement en finance mais aussi en comptabilité. En effet, la part d'amortissement financier correspond à un remboursement de dette à la différence de l'intérêt qui est une charge financière.

Nous allons étudier ici trois types d'emprunts:

1. Les emprunts remboursables à échéance fixe
2. Les emprunts à remboursements constants
3. Les emprunts à annuité constante (les plus pratiqués)

Remarques:

R1. Nous considérons ici des emprunts périodiques. Le passage d'une période temporelle à une autre et le calcul d'un taux équivalent se fera selon les relations déjà démontrées plus haut.

R2. Des annuités mensuelles constantes sont appelées des "mensualités".

5.1. EMPRUNT À ÉCHÉANCE FIXE (IN FINE)

Définition: Nous parlons d'un "**emprunt à échéance fixe**" ou "**emprunt in fine**" lorsque chaque année, l'annuité comprend uniquement la part d'intérêt ! La dernière année, l'annuité comprend l'intérêt ainsi que la totalité (!) du remboursement de l'emprunt.

Remarque: Ce modèle d'amortissement est particulièrement utilisé dans les emprunts obligataires, étudiés plus loin.

Les relations suivantes permettent d'établir n'importe quel élément du "tableau d'amortissement".

Ainsi, l'état de la dette (capital emprunté) C en début d'année k est:

$$C_k = C \quad \forall k \quad (66.186)$$

Le remboursement (amortissement) R_k effectué en fin d'année k est égal à l'amortissement cumulé S_k en fin d'année k et celui-ci n'a lieu qu'à la dernière année n tel que:

$$R_k = S_k = \begin{cases} 0 & \text{si } k < n \\ C & \text{si } k = n \end{cases} \quad (66.187)$$

l'intérêt payé I_k sera constant tout au long de la durée du prêt selon un taux $t\%$ sur le capital emprunté C_k tel que:

$$I_k = t\% C_k \quad (66.188)$$

L'annuité devient alors:

$$A_k = I_k + R_k \quad (66.189)$$

Exemple:

Voyons le tableau d'amortissement d'un emprunt de 1'000.- à 10% l'an remboursé à l'échéance au bout de 4 ans. Le tableau d'amortissement correspondant sera:

Période	État de la dette	Amort.	Amort. Cumulé	Intérêt	Annuité
k	C_k	R_k	S_k	I_k	A_k
1	1'000	0	0	100	100
2	1'000	0	0	100	100
3	1'000	0	0	100	100
4	1'000	1'000	1'000	100	1'100

Tableau: 66.9 - Emprunt avec amortissement à échéance fixe

Le coût du crédit représente la somme des intérêts soit 400.-.

5.2. EMPRUNT A AMORTISSEMENT CONSTANT

Définition: Nous parlons d'un "**emprunt à amortissement constant**", lorsque montant annuel remboursé est constant, c'est-à-dire identique d'année en année (système intuitif).

Les relations suivantes permettent d'établir n'importe quel élément de ce tableau d'amortissement:

$$\begin{aligned}
 C_k &= C - k \frac{C}{n} + \frac{C}{n} = \frac{C}{n}(1 + n - k) \\
 R_k &= R = \frac{C}{n} \\
 S_k &= k \frac{C}{n} \\
 I_k &= t\% C_k \\
 A_k &= R_k + I_k
 \end{aligned}
 \tag{66.190}$$

Exemple:

Un emprunt de 1'000.- à 10% l'an est remboursé par amortissement constant en 4 ans. Établir le tableau d'amortissement et déterminer le coût du crédit. Le tableau d'amortissement correspondant sera:

Période	État de la dette	Amort.	Amort. Cumulé	Intérêt	Annuité
k	C_k	R_k	S_k	I_k	A_k
1	1'000	250	250	100	350
2	750	250	500	75	325
3	500	250	750	50	300
4	250	250	1'000	25	275

Tableau: 66.10 - Emprunt avec amortissement constant

Le coût du crédit représente la somme des intérêts soit 250.-. Donc on paie moins qu'avec le système précédent.

5.3. EMPRUNTS A ANNUITÉ CONSTANTE

C'est le cas le plus fréquent (la définition est dans le titre même). Il est utilisé par la plupart des instituts de petit crédit et de leasing. L'emprunteur connaît d'avance la somme qu'il aura à payer d'année en année. En d'autres termes, c'est comme s'il s'agissait d'un capital C que l'on doit solder en faisant à chaque période un retrait constant A : ce qui consiste à déterminer la valeur actuelle d'une rente postnumerando telle que:

$$C = Aa_{\overline{n}|} \tag{66.191}$$

Les relations suivantes permettent alors d'établir n'importe quel élément du tableau d'amortissement:

$$A = A_k = \frac{C}{a_{\overline{n}|}} \tag{66.192}$$

et puisque $C = Aa_{\overline{n}|}$, alors:

$$C_k = Aa_{\overline{n-k+1}|} \tag{66.193}$$

dès lors, lorsque $k = 1$, nous avons conformément à ce que nous attendons $C_1 = Aa_{\overline{n}|}$.

Et donc l'amortissement est de:

$$R_k = A - t\% C_k \quad (66.194)$$

L'amortissement cumulé est un peu moins évident à trouver avec le bon sens, prenons pour démonstration un amortissement A avec taux $t\%$ sur n périodes. Nous avons par définition:

$$S_k = \sum_{i=1}^k R_i \quad (66.195)$$

avec $k=2$ et $n=3$:

$$\begin{aligned} R_1 &= A - t\% C_1 \\ R_2 &= A - t\% C_2 \end{aligned} \quad (66.196)$$

d'où:

$$\begin{aligned} S_2 &= 2A - t\% (C_1 + C_2) = 2A - t\% (Aa_{\overline{2}|} + Aa_{\overline{1}|}) = A \left[2 - t\% (\alpha_{\overline{2}|} + \alpha_{\overline{1}|}) \right] \\ &= A \left[2 - t\% (f_e^1 + f_e^2 + f_e^3 + f_e^1 + f_e^2) \right] = A \left[2 - t\% (2f_e^1 + 2f_e^2 + f_e^3) \right] \\ &= A \left[2 - t\% \left(2 \frac{1}{(1+t\%)^1} + 2 \frac{1}{(1+t\%)^2} + \frac{1}{(1+t\%)^3} \right) \right] = A \left[2 - t\% \left(\frac{2(1+t\%)^2 + 2(1+t\%) + 1}{(1+t\%)^3} \right) \right] \\ &= A \frac{1}{(1+t\%)^3} \left[2(1+t\%)^3 - t\% (2(1+t\%)^2 + 2(1+t\%) + 1) \right] \\ &= A \frac{1}{(1+t\%)^3} \left[2(1+3t\%+3t\%^2+t\%^3) - t\% (5+6t\%+2t\%^2) \right] \\ &= A \frac{1}{(1+t\%)^3} [2+t\%] = Af_e^3 \left[\frac{t\%}{t\%} 2+t\% \right] = Af_e^3 \left[\frac{2t\%+t\%^2}{t\%} \right] \\ &= Af_e^3 \left[\frac{1+2t\%+t\%^2-1}{t\%} \right] = Af_e^3 \left[\frac{(1+t\%)^2-1}{t\%} \right] = Af_e^3 s_{\overline{2}|} \end{aligned} \quad (66.197)$$

Ainsi, nous avons:

$$S_k = Af_e^k s_{\overline{k}|} \quad (66.198)$$

et aussi:

$$I_k = t\% C_k \quad (66.199)$$

Exemple:

Un emprunt de 1'000.- à 10% l'an est remboursé par annuité constante en 4 ans. Établir le tableau d'amortissement et déterminer le coût du crédit. Le tableau d'amortissement correspondant sera:

Période	État de la dette	Amort.	Amort. Cumulé	Intérêt	Annuité
k	C_k	R_k	S_k	I_k	A_k
1	1'000	215	215	100	315
2	785	237	452	78	315
3	548	261	713	55	315
4	287	287	1'000	29	315

Tableau: 66.11 - Emprunt avec annuité constante

Le coût du crédit représente la somme des intérêts soit 262.-. Ce résultat pourrait s'obtenir par: $nA - C$.

Remarque: Les instituts financiers rajoutent différents frais au crédit comme les frais de dossier, les frais d'assurance, les frais de garantie, etc. L'ajout de ces frais a pour effet de faire augmenter le taux de prêt dont la valeur finale réelle (tous frais compris) est appelée "**taux effectif global**" (TEG). Comme chaque institut financier fait sa petite cuisine ou que chaque pays impose une méthode de calcul particulière (l'obligation de communiquer ce taux et la méthode de calcul associée est spécifiée dans de nombreux pays par la législation pour éviter que les consommateurs ne soient trompés), nous n'avons pas souhaité développer le sujet ici. Cependant, le lecteur intéressé pourra se référer aux exercices d'Économétrie disponibles sur le serveur d'exercices où il pourra trouver un exemple concret mais... particulier!

6. THÉORIE MODERNE DES PORTEFEUILLES

La théorie du marché des valeurs dite aussi "**théorie moderne du portefeuille**" est la théorie mathématique qui traite du choix, de la gestion et des opérations des échanges des emprunts, prêts et capitaux. Elle fait très fortement appel aux modèles statistiques et il est donc important d'avoir lu et compris le chapitre y relatif sur le site au préalable.

Il faut cependant savoir qu'en pratique, dans les instituts financiers privés ou publics, seule une infime minorité des acteurs du marché connaît, comprend et applique des modèles mathématiques et pour les autres ayant obtenu des certifications ou diplômes de formation continue (CFA, FRM, CAIA, etc.), le niveau est souvent affligeant. Effectivement la grande majorité des traders (opérateurs de marché) se limitent à l'analyse graphique de diagrammes de Bourse du type bougies en utilisant des analyses du type moyennes mobiles, limites basées sur les cartes de contrôles (cf. chapitre de Génie Industriel) avec bandes de Bollinger ou autres USL/LSL, régressions linéaires et une trentaine de techniques du même genre (le plus drôle étant les indices empiriques ayant des noms scientifiques comme "l'indice stochastique" et qui sont en fait de simples variations relatives...). La gestion financière n'est donc finalement souvent que l'application du bon sens (quand il est présent...) sur la variation des prix en fonction des quantités... en analysant des graphiques et en sachant retirer ou déplacer ses investissements au bon moment.

Cette situation s'explique aisément: les modèles théoriques actuels du 20ème siècle et du début du 21ème siècle sont incapables de prendre en compte la complexité et l'interaction dans la complexité de notre monde moderne. Vous verrez effectivement dans cette section que la majorité des modèles mathématiques étudiés dans les grandes universités supposent des cas simplifiés et idéalisés (indépendance, normalité, variance finie, choix d'indicateurs statistiques empiriques, etc.). Donc dans la situation actuelle il vaut mieux écouter un trader qui est informé des stratégies des états et entreprises qu'un mathématicien pour lequel le monde se résume à un conte de fée (mais qui néanmoins lui permet de vendre des heures de conseil à un tarif très élevé et de rassurer des clients qui veulent à tout prix des prévisions qui n'ont quasiment aucun sens).

Il faut aussi savoir que beaucoup sont persuadés que tout est écrit quelque part (croyance probablement assez ancrée dans la culture occidentale), qu'une sorte de réalité assez abstraite existe en dehors de notre monde concret et que si nous étions assez malins, nous pourrions la formaliser mathématiquement et prévoir les évolutions futures sur le long terme. Au fait le scientifique sait lui que nous avons affaire dans ce genre de domaine à un chaos déterministe du marché et que la seule manière de gérer celui-ci est de corriger au jugé avec une vague idée de ce qui va se passer. En économie les spécialistes parlent de la "**dictature des marchés**", mais c'est reconnaître en un sens que nous ne savons rien prédire! Évidemment, certains, dans les milieux de la complexité, vendent aux banquiers et à d'autres l'idée qu'ils sauront prédire les fluctuations de la Bourse... mais il suffit d'observer le passé pour voir qu'aucun modèle moderne n'aurait su le prédire. La seule chose que la mathématique peut faire dans la gestion financière, c'est d'analyser le comportement d'un actif financier idéalisé dans un cadre qui l'est lui aussi et c'est déjà pas mal et force un peu au bon sens... (pour ceux qui savent faire des maths ce qui est très loin du cas du 99% des personnes travaillant dans la finance).

En finance, les modèles mathématiques serviraient donc à mesurer et quantifier le risque des investissements. À ce titre, ils jouent le rôle d'outils d'aide à la décision pour les gestionnaires, les investisseurs et les régulateurs. Mais, à de rares exceptions près, une banque ou un fonds d'investissement ne fonde pas une décision majeure d'investissement sur une équation mathématique. La décision, pour les banques d'investissement est souvent motivée par la recherche de rentabilités toujours plus grandes et pour cela elles ne s'appuient pas sur des modèles mathématiques. Par ailleurs, les personnes dirigeantes des banques privées ou d'état sont souvent des personnes issues du monde de la politique, du droit ou de la gestion d'entreprise avec peu de compétence pour comprendre réellement le fonctionnement des marchés (voir le programme des Master en trading de nombreux pays par exemple). Par ailleurs méfiez-vous des entreprises - particulièrement des multinationales - qui recherchent des spécialistes financiers maîtrisant Microsoft Excel ou Microsoft Access. Car cela signifiera qu'elles utilisent des outils non professionnels pour faire un travail qui lui devrait pourtant l'être avec des outils adaptés (et Microsoft Excel ou Microsoft Access ne le sont pas)!!! Donc en termes d'organisation interne, vous pouvez vous assurer que ces entreprises organisent et analysent n'importe quoi, n'importe comment, avec un outil non adapté et donc que c'est le bordel général en interne.

Cela n'empêche cependant pas que pour ceux qui considèrent la mathématique comme un art (ce qui est mon cas), les modèles théoriques financiers ont une certaine élégance et permettent de comprendre rigoureusement le mécanisme de fonctionnement à défaut de pouvoir le prévoir.

Il faudrait peut-être responsabiliser aussi ce domaine d'activité en mettant en place les mesures suivantes:

- Exiger la documentation publique détaillée des composants (sous-jacents), algorithmes et modèles mathématiques des produits financiers.

- Imposer un niveau de fonds propre dynamique (réserve fractionnaire) aux instituts financiers en fonction de leurs positions sur les marchés et du risque des classes d'actifs gérés.
- Obliger les acteurs sur un marché d'opérer dans un domaine d'activité ayant un certain niveau prédéfini de connexité avec la classe d'actifs gérée.
- Limiter le nombre de transactions pour un actif financier non limité dans le temps et imposer aussi pour cette même famille d'actifs un temps minimum de non-transaction.
- Obliger les personnes actives dans ce domaine d'activité de passer des examens régulièrement pour vérifier qu'ils comprennent bien ce qu'ils font et les hypothèses de travail y relatives.
- Interdire les banques qui font du dépôt de faire également du négoce.

Il y aurait donc un grand nombre de propositions à mettre en place pour redonner son but et son éthique originelle au domaine de la gestion financière.

Définition: La "**Bourse**" ("**Stock Exchange**") est le marché public où s'échangent des titres (actions, obligations, contrats, options, etc.) dont la valeur va fluctuer relativement à la "**valeur fondamentale**" (valeur de base calculée selon des modèles théoriques) au gré de l'offre ("**ask**" en anglais) et de la demande ("**bid**" en anglais). Lorsqu'un titre est beaucoup demandé, son prix monte, et inversement, lorsque personne n'en veut. La différence entre la valeur de l'offre et de la demande est appelée "**spread**" (en anglais mais c'est aussi le terme d'usage en français).

La Bourse est une structure qui permet:

1. Pour les entreprises qui veulent investir (donc augmenter leur capital) d'obtenir des fonds afin de satisfaire la demande potentielle.
2. De rendre au plus stable l'économie en la rendant la plus dynamique et fluide possible (mais sous contrôle quand même...) dans le but qu'elle s'autorégule.

Le système cité ci-dessus fonctionne si et seulement s'il est transparent, rationnel, efficient, autorégulant et équilibré!

Remarque: Nous parlons de "**bulle spéculative**" lorsque les prix observés sur un marché boursier s'écartent trop de la valeur fondamentale des biens échangés.

Avant de commencer à s'attaquer à la mathématique pure et dure, il va nous falloir au préalable donner encore une fois un grand nombre de définitions pour s'habituer au vocabulaire usité par les analystes et ingénieurs financiers (attention c'est relativement long...).

7. ABSENCE D'OPPORTUNITÉ D'ARBITRAGE (A.O.A.)

L'une des hypothèses fondamentales des modèles financiers usuels est qu'il n'existe aucune stratégie financière permettant, pour un coût initial nul, d'acquérir une richesse certaine dans une date future. Cette hypothèse est appelée "**absence d'opportunités d'arbitrage**" (A.O.A.). Elle est justifiée théoriquement par l'unicité des prix caractérisant un marché en concurrence pure et parfaite.

Pratiquement, il existe des arbitrages mais qui disparaissent très rapidement du fait de l'existence d'arbitragistes, acteurs sur les marchés dont le rôle est de détecter ce type d'opportunités et d'en profiter. Ceux-ci créent alors une force qui tend à faire évoluer le prix de l'actif vers son prix de non-arbitrage (son prix "réel").

Remarque: Le problème, c'est que le prix "réel" de l'actif, c'est le prix vers lequel le font converger acheteurs et vendeurs et quand ceux-ci se mettent d'accord entre eux... le système se grippe...

Ainsi, si plusieurs actifs de même risque proposent des rendements différents, les investisseurs qui recherchent de nouvelles opportunités vont logiquement tourner leurs achats vers ceux dont le rendement est le plus élevé, ce comportement entraînant alors une baisse du rendement de ces actifs.

La mathématique financière reposant sur l'A.O.A. laisse ainsi ces arbitragistes gagner de l'argent et néglige ces apparitions d'opportunités qui de toute façon n'existent toujours que sur un temps supposé très bref (ce type de stratégie est mis à profit aujourd'hui avec l'informatique qui peut donner des ordres de vente et d'achat à la milliseconde près).

Un exemple royal pour illustrer ces propos est d'utiliser une version simplifiée du modèle élaboré par Cox, Ross et Rubinstein car il traduit explicitement le concept de l'A.O.A. et l'importance des modèles probabilistes.

L'exemple se base sur l'hypothèse que le marché est formé d'un actif risqué et d'un taux de placement constant r . Par exemple, une somme de 1 dollar aujourd'hui, placée au taux r , engendre un revenu certain et garanti de $1+r$ dollars au temps 1 quelle que soit l'évolution future du marché dans l'exemple considéré.

Nous commençons par étudier ce marché sur une seule période de temps telle que le temps initial sera noté $t = 0$ et l'instant final $t = 1$ (nous appelons une telle situation un "**marché monopériodique**"). Nous supposons parfaitement connaître le marché à l'instant initial. Dans notre contexte cela signifie que le prix de l'actif risqué est $S_0 > 0$ fixé et l'actif non risqué est déterminé par son rendement $r > 0$.

Quant à l'actif risqué, sa valeur à $t = 1$ n'est pas connue à l'avance. Afin de restreindre le champ des possibles, nous supposons que le rendement de cet actif ne peut prendre que deux valeurs b (bas) et h (haut) avec:

$$-1 \leq b < h \quad (66.200)$$

Ainsi, l'actif risqué au temps $t = 1$ ne peut prendre que deux valeurs positives. La valeur basse:

$$S_1 = S_0 (1+b) \quad (66.201)$$

ou la valeur haute:

$$S_1 = S_0 (1+h) \quad (66.202)$$

d'où l'appellation de "**modèle binomial**"...

Un investisseur peut ainsi acheter une quantité δ d'actif risqué et placer une somme x_0 au taux r . La valeur V_0 au temps $t = 0$ du portefeuille de composition (δ, x_0) est donc:

$$V_0 = \delta \cdot S_0 + n_0 \quad (66.203)$$

À l'instant final, nous aurons:

$$V_1 = \delta \cdot S_1 + n_0 (1+r) \quad (66.204)$$

Comme nous l'avons expliqué plus haut, S_1 peut prendre deux valeurs, il en est donc de même pour V_1 . Ce qui signifie que le revenu de ce portefeuille est incertain.

Maintenant, pour montrer le concept de l'A.O.A. passons à une application numérique en considérant la situation particulière où il est plus avantageux, et à coup sûr, d'investir dans l'actif risqué que le non risqué.

Imaginons pour cela que nous emprunions 100.- à une banque au taux sans risque de 5% et que l'unité d'actif risqué que nous souhaitons acquérir avec cet argent soit cotée aujourd'hui à $S_0 = 100$. – celle-ci pouvant prendre deux valeurs futures:

$$t = 1: \begin{cases} S_1 = 105 \rightarrow b = 5\% \\ S_2 = 108 \rightarrow h = 8\% \end{cases} \quad (66.205)$$

Nous avons alors pour notre portefeuille à l'instant initial:

$$V_0 = \delta \cdot S_0 + n_0 = 1 \cdot 100 - 100 = 0. - \quad (66.206)$$

et à l'instant final deux cas possibles:

$$V_{1,b} = \delta \cdot S_{1,b} + n_0 (1+r) = 1 \cdot 105 - 100 \cdot (1+5\%) = 0. - \quad (66.207)$$

et:

$$V_{1,h} = \delta \cdot S_{1,h} + n_0 (1+r) = 1 \cdot 108 - 100 \cdot (1+5\%) = 3. - \quad (66.208)$$

Nous voyons alors de manière triviale que si $r < b$ il existe une opportunité d'arbitrage puisqu'il devient possible de gagner de l'argent à coup sûr sans en dépenser! Pour éviter une A.O.A. dans cette situation, il faut donc que le marché s'équilibre et qu'il y ait:

$$b < r \quad (66.209)$$

Inversement, s'il est plus à coup sûr plus lucratif d'investir dans l'actif non risqué que dans l'actif risqué, le marché doit s'organiser pour éviter toute opportunité d'arbitrage de la sorte que:

$$r < h \quad (66.210)$$

Ainsi, dans les deux cas, il faut éviter à tout moment que dans le marché binomial nous ayons une A.O.A. Et cela est seulement possible si:

$$b < r < h \quad (66.211)$$

L'absence d'opportunité d'arbitrage à deux implications simples (il y en a des moins simples comme nous le verrons bien plus loin....). Considérons le cas où nous avons deux actifs respectivement de rendement i et j et que nous savons que l'actif de rendement i ne fera pas défaut dans le paiement de ses dividendes. Par contre l'actif de rendement j pourrait faire défaut avec une probabilité $1 - p$. Alors, de par l'A.O.A nous devons avoir:

$$p(1+k) = 1+i \quad (66.212)$$

Donc nous pouvons en déduire la probabilité de non défaut de paiement de l'actif:

$$p = \frac{1+i}{1+k} \quad (66.213)$$

et donc la probabilité de défaut:

$$1-p = 1 - \frac{1+i}{1+k} \quad (66.214)$$

Ce type de raisonnement permet donc aussi d'exiger un rendement k d'un émetteur d'actifs (de type obligations typiquement) connaissant par expertise/audit la probabilité de défaut de paiement/remboursement par rapport à un actif sûr à 100% de rendement i .

7.1. PORTEFEUILLES

La majorité des transactions boursières concernent le contenu des "**portefeuilles de titres**" (security portfolio) qui sont l'ensemble des titres qu'un acteur du marché peut détenir. Gérer un portefeuille consiste donc (le plus classiquement...) pour un gestionnaire à chercher un retour sur investissement (RSI) maximal pour le client tout en minimisant les risques.

Remarque: Le RSI est aussi parfois appelé "**rendement**" ou "**taux de rendement**" ou "**taux de profit**" et désigne donc un ratio financier qui mesure le montant d'argent gagné ou perdu par rapport à la somme initialement investie dans un placement (souvent sur la base d'une période annuelle). En général, ce ratio est exprimé en pourcentage plutôt qu'en valeur décimale.

Les "**titres financiers**" (financial security) se dérivent sous la forme d'actions, d'obligations, d'options de devises et de matières premières tous appelés plus généralement "**produits financiers**" ou encore "**actifs financiers**" et dont les définitions (non exhaustives) seront données ci-dessous. Lorsque certains de ces produits sont mélangés, nous parlons alors de "**produits structurés**" (typiquement l'association d'un sous-jacent avec une option).

Définitions:

D1. Pour mesurer l'évolution générale d'un marché boursier, nous utilisons des "**indices**" reflétant la moyenne arithmétique (Down Jones Index par exemple) ou la moyenne pondérée (Swiss Market Index par exemple) des cours (valeurs) d'un certain nombre de titres représentatifs. Cela permettant d'en connaître le rendement.

D2. Un "**produit dérivé**" est un produit/instrument financier, qui s'achète et se vend, et qui est toujours bâti sur la base d'un titre financier. Ce dernier est alors appelé "**actif sous-jacent**" ou "**support**" du produit dérivé. Ceux-ci peuvent donc être des actions, des obligations, des devises, ... et même des produits dérivés... Le danger avec les produits dérivés est, à force de les superposer de ne plus savoir exactement quels sont les sous-jacents.

D3. La "**volatilité**" mesure l'amplification de la variation d'un cours. Autrement dit, un titre financier dont la volatilité est élevée voit son cours varier fortement, voire de façon exagérée sur une période donnée. À l'inverse, un titre dont la volatilité est faible voit son cours varier peu et/ou de manière assez cohérente. La volatilité s'exprime souvent en pourcentage dans les modèles mathématiques simples (car il en existe plusieurs définitions dont nous en verrons certaines par la suite). Ainsi, la volatilité d'un titre sur une période donnée est basiquement définie par:

$$V = \frac{\text{Cours le plus haut} - \text{Cours le plus bas}}{\text{Cours moyen}} \quad (66.215)$$

Dans des situations complexes, la volatilité est souvent assimilée à l'écart-type et nous verrons cela plus loin.

7.1.1. ACTIONS

Définition: Les "**actions**" sont des papiers-valeurs reconnaissant par contrat des droits de propriétés sur le capital d'une entité dite "**société anonyme**" (S.A.). Ce contrat a un prix et il est échangeable sur le marché.

L'action donne à son propriétaire des droits de différentes natures tels que les droits sociaux (droit de vote aux assemblées générales, droit d'élection et d'être élu au conseil d'administration) ou patrimoniaux (droit de recevoir une part du bénéfice net, sous forme de "**dividende**" variable représenté à l'époque du papier par un petit coupon numéroté à détacher/découper), ou une part du produit de la liquidation de la société si elle venait à tomber en faillite, ainsi qu'un droit préférentiel d'acheter de nouvelles actions en cas d'augmentation du capital.



Figure: 66.12 - Exemple d'action avec dividende à l'époque du papier...

Nous distinguons les actions suivantes:

- "**L'action classique**" qui donne donc un droit de vote lors des assemblées générales, un droit à l'information (...), une droit à des dividendes.
- "**L'action privilégiée**" qui offre un privilège qui peut être une priorité lors des votes dans les assemblées générales ou une priorité lors de la distribution du dividende.
- "**L'action à dividende prioritaire**" qui donne uniquement un accès privilégié aux dividendes (donc pas de droit de vote!).
- Les "**actions à bons de souscription d'actions**" qui sont des actions qui donnent droit à leur détenteur de souscrire à de nouvelles actions à une date donnée.

Définition: Il y a plusieurs types de "**rendement boursier**" en fonction du contexte d'une discussion qui sont pour certains intuitifs et pour d'autres assez complexes. Voici trois des plus courants (nous en verrons d'autres par la suite...) à notre connaissance lorsque l'on aborde pour la première fois la mathématique financière:

- Il y a le rapport, exprimé en pourcentage et appelé "**yield**" ou "**taux de rendement**" (à ne pas confondre avec le "yield to maturity" ou "taux de rendement à maturité" que nous verrons plus loin), entre le dividende par action distribué par une société et le cours de l'action en circulation de cette société au moment du versement du dividende (certains prennent parfois la moyenne arithmétique des dividendes versés sur plusieurs périodes):

$$R_B = \frac{\text{Dividende versé}}{\text{Cours de l'action}} \quad (66.216)$$

- Il y a le rapport, exprimé en numéraire par année et appelé "**rentabilité de l'action**", de la différence entre le cours de vente de l'action majoré par les dividendes et le cours d'achat de l'action de cette société sur le nombre de périodes (exprimé souvent en années):

$$R_B = \frac{(\text{Cours de vente} + \text{Dividendes versés}) - \text{Cours d'achat}}{\text{nombre d'années}} \quad (66.217)$$

Si nous divisons le résultat de ce dernier rendement numéraire annuel par le capital initial investi, nous obtenons le rendement en pourcentage.

- Il y a le "**taux de rendement implicite**" (et son inverse le PER: "**Price Earning Ratio**") qui est le rapport entre le bénéfice net de l'entreprise et la valeur de capitalisation boursière et dont l'abréviation peut être confondue avec le taux de rendement interne (cf. chapitre Techniques de Gestion):

$$TRI = \frac{\text{Bénéfice net}}{\text{Capitalisation boursière}} \quad (66.218)$$

Le TRI est donc une sorte de rendement entre le revenu potentiel de l'action et son prix (attention : il ne serait question du rendement réel que si l'on prenait en compte le rapport dividendes versés/cours de l'action qui peut être très différent). On suppose, à tort ou à raison, que l'entreprise a un meilleur usage à long terme à faire de ses bénéfices que de les distribuer sous forme de dividendes....

Le PER (l'inverse du TRI) généralement cité dans la presse est celui calculé sur le dernier bénéfice annuel publié. Toutefois les analystes se basent souvent sur le bénéfice anticipé pour l'année en cours.

Exemples:

E1. Prenons une action remboursable achetée il y a 6 ans au prix de 10.- (capital investi). L'investisseur la vend à 12.50.- L'investisseur a reçu 3 fois un dividende de 2.20.- plus une fois un de 1.-. Les deux types de rendements donnent alors respectivement:

$$R_B = \frac{\frac{7.6}{4}}{12.5} = 15.2\% \quad (66.219)$$

$$R_B = \frac{(12.5 + 7.6) - 10}{6} = \frac{2.5}{6} + \frac{7.6}{6} = 0.41 + 1.27 = 1.68. - / an$$

et si nous divisons ce dernier par le capital initial investi (10.-), nous obtenons donc 16.8%.

E2. Le prix (frais compris) maximum que nous pouvons mettre pour acheter une action de 500.- rapportant un dividende de 12% à concurrence d'un placement identique avec un rendement de même périodicité de 5% est de 1'200.-. Effectivement, une telle action rapporterait 60.- de dividendes à chaque période (12% de 500.-). La somme qu'il faudrait placer pour avoir le même intérêt à 5% est 1'200.- (5% de 1'200.- faisant aussi 60.-).

E3. Le titre émis par une société (dont le capital est composé de 10 millions d'actions) cote 100.-, ce qui porte la valeur en Bourse de la société à 1 milliard. Le bénéfice net prévu est de 50 millions pour l'exercice en cours, soit 5.- par action. La division du bénéfice net prévu par la capitalisation donne donc un price earning ratio de 20.

Remarques:

R1. Nous différencions les "**actions au porteur**" négociables sans restrictions en Bourse des "**actions nominatives**" dont la valeur doit être négociée avec des restrictions juridiques plus ou moins complexes, car il y figure le nom de l'actionnaire qui doit être inscrit au registre des actionnaires.

R2. Lorsqu'une société anonyme veut augmenter son capital-actions, elle peut émettre des actions supplémentaires. Les nouvelles actions seront proposées aux actionnaires de la société à un cours fixe et en proportion des actions qu'ils détiennent ("**droit de souscription**") afin de ne pas les pénaliser contre une décote. Ceci leur permettra donc de maintenir le pourcentage de leur part dans le capital, ainsi que le poids de leurs droits de vote.

Pour clore cette petite introduction sur les actions, signalons qu'un modèle théorique d'évaluation actuelle (dans le sens "à la date du jour") des actions porte le nom de "**modèle de Durand**". Son idée est assez simple: le prix P_0 d'une action aujourd'hui est égale à la somme de ses cash-flows (donc de ces dividendes D_k) actualisés au taux géométrique moyen du marché (ou attendu/exigé par l'actionnaire) de la période i correspondante $t_i\%$ versés à chaque période k ainsi que de son prix de revente futur P_T actualisé lui aussi.

Soit formellement:

$$P_0 = \frac{P_T}{\prod_{i=1}^T (1+t_i\%)} + \sum_{t=1}^T \frac{D_t}{\prod_{i=1}^t (1+t_i\%)} \quad (66.220)$$

Si nous faisons l'hypothèse scolaire habituelle comme quoi le taux est toujours constant (sinon, nous utiliserons une moyenne géométrique des taux sur la totalité des périodes ou mieux encore une simulation de Monte-Carlo), nous avons alors:

$$P_0 = \frac{P_T}{(1+t\%)^T} + \sum_{t=1}^T \frac{D_t}{(1+t\%)^t} \quad (66.221)$$

Or le prix de revente en T sera égal à cette même relation et ce ainsi de suite jusqu'à l'infini, car une action n'a pas vocation à être remboursée! Nous avons alors:

$$P_0 = \frac{P_T}{(1+t\%)^\infty} + \sum_{t=1}^{\infty} \frac{D_t}{(1+t\%)^t} \quad (66.222)$$

et comme:

$$\frac{P_T}{(1+t\%)^\infty} \rightarrow 0 \quad (66.223)$$

Il reste alors:

$$P_0 = \sum_{t=1}^{\infty} \frac{D_t}{(1+t\%)^t} \quad (66.224)$$

Maintenant, considérons une simplification du modèle de Durand qui s'appelle le "**modèle de Gordon et Shapiro**" ou en anglais "**Dividend Discount Model**". Celui-ci considère qu'à chaque période, les dividendes croissent selon un même taux noté t_{cd} , d'où:

$$D_t = D_0 (1+t_{cd}\%)^t = D_1 (1+t_{cd}\%)^{t-1} \quad (66.225)$$

En appliquant ce que nous venons de montrer dans le modèle de Durand:

$$\begin{aligned} P_0 &= \sum_{t=1}^{\infty} \frac{D_t}{(1+t\%)^t} = \sum_{t=1}^{\infty} \frac{D_1 (1+t_{cd}\%)^{t-1}}{(1+t\%)^t} = \frac{D_1}{1+t\%} \sum_{t=1}^{\infty} \frac{(1+t_{cd}\%)^{t-1}}{(1+t\%)^{t-1}} \\ &= \frac{D_1}{1+t\%} \sum_{t=1}^{\infty} \left(\frac{1+t_{cd}\%}{1+t\%} \right)^{t-1} \end{aligned} \quad (66.226)$$

Posons:

$$q = \frac{1+t_{cd}\%}{1+t\%} \quad (66.227)$$

$$t-1 = n$$

Nous avons alors:

$$P_0 = \frac{D_1}{1+t\%} \sum_{t=1}^{\infty} \left(\frac{1+t_{cd}\%}{1+t\%} \right)^{t-1} = \frac{D_1}{1+t\%} \sum_{n=0}^{\infty} q^n \quad (66.228)$$

Or, nous avons démontré dans le chapitre de Suites Et Séries, que pour toute suite géométrique de raison x est donnée pour rappel par:

$$S_n = \frac{1-q^{n+1}}{1-q} \quad (66.229)$$

Donc si n tend vers l'infini et sous l'hypothèse que:

$$q = \frac{1+t_{cd}\%}{1+t\%} < 1 \quad (66.230)$$

c'est-à-dire que le taux de rendement attendu par les actionnaires est supérieur aux taux de croissance des dividendes, nous avons alors:

$$S_{\infty} = \frac{1-q^{\infty}}{1-q} = \frac{1}{1-q} \quad (66.231)$$

Dès lors:

$$P_0 = \frac{D_1}{1+t\%} \sum_{n=0}^{\infty} q^n = \frac{D_1}{1+t\%} \frac{1}{1-q} = \frac{D_1}{1+t\%} \frac{1}{1 - \frac{1+t_{cd}\%}{1+t\%}} = \frac{D_1}{1+t\%} \frac{1}{\frac{1+t\%-1-t_{cd}\%}{1+t\%}} \quad (66.232)$$

D'où au final:

$$P_0 = \frac{D_1}{t\% - t_{cd}\%} \quad (66.233)$$

Le taux de croissance des dividendes $t_{cd}\%$ est déterminé soit à partir des données historiques de l'action, soit à partir des prévisions des analystes sur les futurs dividendes.

Notons que nous trouvons cette dernière relations souvent sous les formes suivantes dans la littérature:

$$P_0 = \frac{D_0 (1+t_{cd}\%)}{t\% - t_{cd}\%} = \frac{D_0 (1+g)}{t\% - g} \quad (66.234)$$

Les praticiens utilisent parfois la dernière expression sous la forme:

$$t\% = \frac{D_1}{P_0} + t_{cd}\% \quad (66.235)$$

pour comparer le taux de rendement $t\%$ de plusieurs actions connaissant leur dividende, leur prix actuel et le taux supposé de croissance des dividendes par des prévisionnistes.

Exemple:

Nous souhaitons valoriser une action qui verse un première dividende de 5.- et dont la croissance est supposée constante (à l'infini...) à un rendement (moyen géométrique) de 14.87% en comparaison à un rendement géométrique moyen du marché de 20%. Nous avons alors le prix de cette action qui peut être estimée par:

$$P_0 = \frac{5}{0.2 - 0.1487} \cong 97.46.- \quad (66.236)$$

Signalons pour clore cette évaluation des actions une propriété importante, aussi applicable aux obligations que nous allons de suite voir, et qui est le "risque de défaut de crédit". Ce risque, qui doit être quantifié au mieux en termes de probabilités, consiste simplement dans le fait que l'émetteur de ces actions pourrait ne pas répondre à ses devoirs et verser l'argent qu'on est en droit d'attendre de lui pour la simple raison qu'il a fait faillite ou que le système économique ne s'y prête pas (crise économique majeure par exemple). L'État garantit normalement seulement une fraction du gain attendu (garantie qu'il s'agit aussi d'actualiser mathématiquement). Dès lors, il faut savoir que:

1. Le prix actualisé est toujours un peu plus faible que le simple modèle ci-dessus puisque celui-ci est lié à un événement n'ayant pas de probabilité égale à 100% (mais inférieure!).
2. Le prix actualisé est en réalité une espérance pondérée par les probabilités des scénarios économiques.
3. Tous les modèles économiques algébriques sont idéalisés et il ne faut jamais oublier cela (ils ne prennent pas en compte les phénomènes macro-économiques)!

7.1.2. OBLIGATIONS

Contrairement à l'emprunt individuel (emprunt indivis), l'emprunt dit "**emprunt obligataire**" fait appel à de nombreux prêteurs, appelés "**souscripteurs**", qui reçoivent, en échange de sommes prêtées, des titres appelés "obligations".

Définition: Les "**obligations**" sont des papiers valeurs (titres de créance) établissant par contrat des droits de créance (capital prêté) et qui rapportent un intérêt fixe (en général annuel) au titulaire (elles sont remboursables à une échéance prévue par le contrat). Ce contrat a un prix (dépendant de la date!), il est échangeable sur le marché et le débiteur est obligé de payer les intérêts. Par ailleurs si l'obligation est "**convertible**" elle donne droit au créancier d'obtenir soit le remboursement de l'obligation, soit sa conversion en actions, suivant des modalités fixées d'avance.

Remarque: Les actions et les obligations sont très différentes de par ce qu'elles représentent. Alors que l'action désigne un titre de propriété lié au capital social d'une société, cotée ou non, l'obligation est un titre de créances. L'obligation est donc basée sur les dettes d'une entreprise, d'un état, ou d'une collectivité locale alors que l'action est une part des capitaux propres d'une société par actions. Toute société composée d'actions n'est pas nécessairement cotée sur le marché boursier, et toute société n'est pas nécessairement une société d'actions. L'émission d'une obligation permet à l'émetteur de diversifier ses sources d'emprunts, et pour une action, de diversifier ses sources de financements. De plus en cas de faillite de l'émetteur, le détenteur d'une obligation est prioritaire sur l'actionnaire.

Nous distinguons principalement trois types d'obligations:

T1. "**Obligation à taux fixe**" ou "**obligation ordinaire**" qui est la plus classique des obligations. Elle procure un flux d'intérêt définitivement fixé lors de son émission selon une périodicité prédéfinie jusqu'à son échéance (ce qui est sécurisant) dont l'intérêt mathématique correspondant à ce flux est appelé "**taux de rendement à maturité**" (Yield to Maturity). Les financiers la désignent souvent sous le nom anglophone de "**plain vanilla bond**". Lorsque la durée du flux peut être considérée comme infinie, nous parlons "**d'obligation perpétuelles**". Ce n'est cependant pas un investissement sans risque comme nous le verrons dans un exemple simple plus loin.



Figure: 66.13 - Exemple d'obligation à taux fixe avec coupons à l'époque du papier..

L'obligation à taux fixe est classiquement cotée en prix ou en taux. L'obligation à taux fixe est évaluée par actualisation des flux futurs qu'elle délivre.

Enfin, nous avons la famille des "**obligations indexées**":

T2. "**Obligation à taux variable**" dont les flux d'intérêt, mais pas le prix de remboursement, sont indexés sur un taux de référence comme le taux directeur d'une banque centrale, les résultats d'une entreprise, ou autre. Le risque associé à ce taux variable est appelé "**risque de taux**". Pour le détenteur d'un portefeuille obligataire qui souhaite protéger son capital, il suffit alors d'immuniser son portefeuille contre les variations de ce taux. C'est ce que nous appelons cela la "**couverture en duration**".

Remarquons que le "**taux zéro périodique**" à échéance de n unités de période d'une obligation zéro-coupon de nominal C et de prix d'émission P peut être obtenu facilement à partir de la relation de l'intérêt composé:

$$C = P(1 + t_0 \%)^n \quad (66.237)$$

Relation que l'on retrouve fréquemment dans la littérature spécialisée anglophone sous les formes suivantes:

$$\begin{aligned} C_T &= PV(1+y)^T \\ FV &= PV(1+y)^T \end{aligned} \quad (66.238)$$

avec PV qui signifie "Present Value", FV "Future Value" et le taux est noté y pour signifier "yield".

Soit:

$$t_0 \% = \left(\frac{C}{P} \right)^{1/n} - 1 \quad (66.239)$$

Ce que l'on retrouve parfois dans la littérature plutôt sous la forme traditionnelle suivante (vive l'absence de normes...!):

$$z(T) = \left(\frac{N}{P(T)} \right)^{1/T} - 1 \quad (66.240)$$

Nous voyons que cette fonction pour un nominal donnée est un prix d'émission donné à exactement la forme de la courbe des taux de la figure précédente!

Exemple:

Le taux zéro à 4 ans correspondant à un zéro-coupon de nominal 100 est de prix d'émission 90 (en-dessous du pair) est de:

$$t_0 \% = \left(\frac{C}{P} \right)^{1/n} - 1 = \left(\frac{100}{90} \right)^{1/4} - 1 \cong 2.67\% \quad (66.241)$$

Remarque: Parmi les obligations, seules les zéro-coupon permettent d'éliminer réellement tout risque de taux entre deux dates. Une obligation à taux fixe classique génère en fait autant de risques de taux supplémentaires qu'elle est dotée de flux financiers intermédiaires: le taux de réinvestissement de chacun des coupons entre sa date de paiement et la date de remboursement final est, en fait, inconnu, même s'il est implicite dans le prix de l'obligation.

Il existe une autre manière courant et très pratique d'écrire la valeur actuelle. Rappelons que nous avons démontré bien plus haut que sous certaines conditions:

$$C_n = C_0 (1 + t\%)^n = C_0 e^{nt\%} \quad (66.242)$$

Il vient alors:

$$P = \frac{C}{(1 + t_0\%)^n} = \frac{C}{(1 + t_0\%)^n} = C e^{-nt_0\%} \quad (66.243)$$

Exemple:

Considérons une obligation zéro-coupon qui paie 100.- dans 10 ans et dont la valeur actuelle est de 55.3895.-. Ceci correspond à un taux de:

$$t_0\% = \left(\frac{C}{P} \right)^{1/n} - 1 = \left(\frac{100}{55.3895} \right)^{1/10} - 1 = 6\% \quad (66.244)$$

Et intérêt continu, cela donne (conformément à la démonstration déjà effectuée bie plus haut):

$$t_n\% = \ln(1 + t_0\%) = \ln(1 + 6\%) = 5.83\% \quad (66.245)$$

Les obligations sont caractérisées par plusieurs propriétés:

- P1. Leur "**devise**" de base qui peut fluctuer sur un marché global.
- P2. Leur "**date d'échéance**" ou "**date de maturité**" qui permettra en fonction de leur date d'émission et du type de calendrier (échancier) de connaître la valeur actualisée de l'obligation à tout moment.
- P3. Leur "**valeur nominale**", appelée le "**pair**", désigne la valeur servant de base au calcul des intérêts.
- P4. Leur "**taux d'intérêt nominal**" ou "**taux facial**" associé à la périodicité (souvent annuelle) permet de définir l'intérêt appelé "**coupon**" ou "**coupon de dividende**" appliqué sur la valeur nominale d'une obligation qui sera versée au souscripteur à la date dite "**date de jouissance**". Normalement le mode de calcul du taux d'intérêt doit être communiqué.
- P5. Leur "**prix d'émission**", "**prix de souscription**" est le prix réellement payé par le souscripteur pour devenir propriétaire d'une obligation. L'émission des obligations se fait donc au pair si la valeur nominale est égale à la somme demandée pour son acquisition. Le prix de souscription se fait en-dessous du pair si la somme demandée est inférieure au nominal (cas le plus fréquent), mais le prix peut au-dessus du nominal! La différence entre nominal et prix de souscription est appelée "**prime d'émission**".

Le prix d'émission n'est pas toujours égal à la valeur nominale afin de donner envie au prêteur d'acheter sans forcément proposer un taux d'intérêt nominal trop élevé. Ainsi, l'émission d'une obligation est un jeu subtil pour l'émetteur entre la valeur du taux facial et le prix d'émission. Ainsi, si le taux d'intérêt nominal est légèrement supérieur à celui du marché il faudra probablement avoir une prix d'émission au-dessus du pair et inversement.

P6. Leur "**prix de remboursement**" ou "**valeur de remboursement**" est la somme réellement versée à l'emprunteur lors du remboursement de l'obligation à l'échéance. Le remboursement peut être prévu au pair ou parfois au-dessus à l'échéance (in fine), par tranches, ou jamais (obligations perpétuelles). La différence entre la valeur de remboursement et le nominal est appelée "**prime de remboursement**".

Remarque: L'investisseur doit être particulièrement attentif à l'indication "**subordonné**" sur son papier d'obligation, qui signifie qu'en cas de faillite du débiteur (assimilée au "**risque de signature**"), le détenteur de l'obligation ne pourra être remboursé qu'après tous les autres créanciers... Le risque de signature peut être évité en choisissant des obligations (très) sûres comme les obligations d'état ou de sociétés renommées. Le revers de la médaille est la faiblesse des taux alors offerts qu'il faut en plus mettre en opposition avec l'inflation (sur un taux de 3% sur dix ans d'une obligation d'état qui subit une inflation de 2% il ne reste plus que 1% de rémunération par exemple).

Indiquons que les obligations à coupons multiples sont peu à la mode à cause du système d'imposition des États et aussi de par le coût des démarches administratives qu'elles génèrent. D'où le fait que les zéro-coupon sont préférées.

Exemples:

E1. Considérons un emprunt obligataire de 3'000'000.- divisé en 300 obligations de 10'000.- nominal émis en juin 2004 pour une durée de 10 ans. Souscription: 99.5% de la valeur au pair. Remboursement au pair à l'échéance. Intérêt annuel fixe: 4.5%.

Les valeurs définies plus haut s'expriment alors ainsi:

La valeur nominale C de l'obligation est donc de 10'000.-. Le nombre N d'obligations est de 300. La durée n de l'emprunt est de 10 ans et le taux facial $t\%$ est de 4.5%. Le prix d'émission est de 99.5% de 10'000.- soit $E = 9'950.-$ (en-dessous du pair!). Le coupon a donc une valeur c de 450.- et le remboursement R s'effectue au pair et vaut donc $10'000 - 4500 = 5'500.-$.

E2. Soit une obligation à taux fixe, émise au prix de 1'000.-, et versant un coupon annuel de 100.-. Le taux servi est donc de $100/1'000 = 10\%$.

Supposons que les taux du marché passent à 15%. Cela signifie qu'une nouvelle obligation, qui est émise au prix de 1'000.-, sert un coupon de 150.- (car $150/1'000 = 15\%$).

La nouvelle obligation est donc plus intéressante que l'ancienne, et tout le monde va vouloir vendre l'ancienne pour acheter la nouvelle. C'est pourquoi le prix de l'ancienne obligation va implicitement baisser, jusqu'à ce qu'il corresponde à celui d'un produit financier procurant du 15%, soit ici 666.-. Alors, nous aurons bien $100/666 = 15\%$.

De même, si les taux du marché baissent à 5%, cela signifie qu'une nouvelle obligation, qui est émise au prix de 1'000.-, sert un coupon de 50.- (car $50/1'000 = 5\%$).

La nouvelle obligation est donc moins intéressante que l'ancienne, et personne ne voudra l'acheter. C'est pourquoi le prix de l'ancienne obligation va implicitement monter, jusqu'à ce qu'il corresponde à celui d'un produit financier procurant du 5%, soit ici 2'000.-. Alors, on aura bien $100/2'000 = 5\%$.

Ainsi, le prix d'une obligation à taux fixe diminue implicitement lorsque les taux montent, et monte lorsque les taux baissent. C'est la raison pour laquelle un placement en obligations n'est pas sans risques: on peut perdre une partie du capital. En fait, la seule stratégie sans risque consiste à acheter les obligations au moment de l'émission, et à les garder jusqu'à l'échéance.

À tout moment, la valeur sur le marché d'une obligation à taux fixe ou taux variable doit donc être égale à la somme des valeurs actualisées des coupons et du remboursement auxquels elle donnera encore droit. La valeur actuelle étant calculée au taux du marché obligataire en vigueur pour des obligations du même type et de même durée.

Ainsi, la valeur actuelle d'une obligation à taux fixe (le cas étant facilement généralisable à un taux variable) doit être vue comme un capital initial dont on retire pendant n périodes restantes une certaine somme fixe, somme correspondante au prix du coupon (taux facial multiplié par le nominal):

$$c = t\% \cdot C \quad (66.246)$$

avec C la valeur nominale de l'obligation et le tout cumulé étant périodiquement soumis à l'intérêt du taux du marché $t_M\%$ (donc le "taux à maturité") constant dans le cadre d'une prise en considération d'un avenir certain.

Ainsi, la valeur actuelle d'une obligation à taux fixe est dans un premier temps constituée que de la valeur actuelle des coupons futurs (appelés aussi souvent "flux") restants pendant n périodes telle que:

$$P_c = c \cdot \alpha_n = \sum_{i=1}^n \frac{c}{(1+t_M\%)^i} = c \sum_{i=1}^n \frac{1}{(1+t_M\%)^i} = t\% C \sum_{i=1}^n \frac{1}{(1+t_M\%)^i} \quad (66.247)$$

Que l'on retrouve dans la littérature spécialisée parfois sous la forme suivante:

$$P = cF \sum_{i=1}^n \frac{1}{(1+y)^i} \quad (66.248)$$

Cette partie du prix de la valeur de l'obligation correspond donc à la somme totale nécessaire telle que l'on peut solder P_c après avoir retiré n fois (le nombre de périodes restant) la valeur c à un taux d'intérêt $t_M\%$.

Remarque: Rappelons suite à ce que nous avons démontré dans le chapitre de Suite Et Séries, que si $t_M\%$ est inférieur à 1 et que la somme tend vers l'infini, alors:

$$P_c = \frac{c}{t_M\%} = \frac{t\% C}{t_M\%} \quad (66.249)$$

Ensuite, l'obligation est constituée de la valeur du remboursement R . Bien que celle-ci soit remboursée à terme, elle peut être vue comme un capital épargné à un taux correspondant à celui du marché $t_M\%$ tel que:

$$R = C_0 (1 + t\%)^n = C_0 f_e^n \Leftrightarrow C_0 = R f_e^{-n} \quad (66.250)$$

La valeur actuelle de l'obligation concernant le remboursement est alors:

$$P_R = R f_e^{-n} \quad (66.251)$$

ce qui correspond au capital actuel pour obtenir le remboursement R après les n périodes restantes.

Ainsi, le prix total d'une obligation à taux fixe, appelé aussi "**prix obligataire de non-arbitrage**" est:

$$P = P_c + P_R = c \cdot a_n + R \cdot f_e^{-n} = c \sum_{i=1}^n \frac{1}{(1 + t_M\%)^i} + \frac{R}{(1 + t_M\%)^n} \quad (66.252)$$

c'est-à-dire la valeur actuelle des coupons futurs ainsi que la valeur actuelle du remboursement in fine. Cette relation à son importance en finance, il convient de s'en souvenir!!

La valeur d'une obligation, au sens de son cours en Bourse, peut donc différer de sa valeur nominale fixée à l'émission si les taux d'intérêts changent sur le marché d'où l'intérêt de calculer sa valeur actuelle.

Exemple:

E1. Soit à calculer le prix actuel d'une obligation, ayant des coupons annuels de 450.-, avec un remboursement au pair dans 5 ans de 10'000.-.

La valeur actuelle pour un taux du marché compris entre 0% et 100% a la caractéristique suivante:

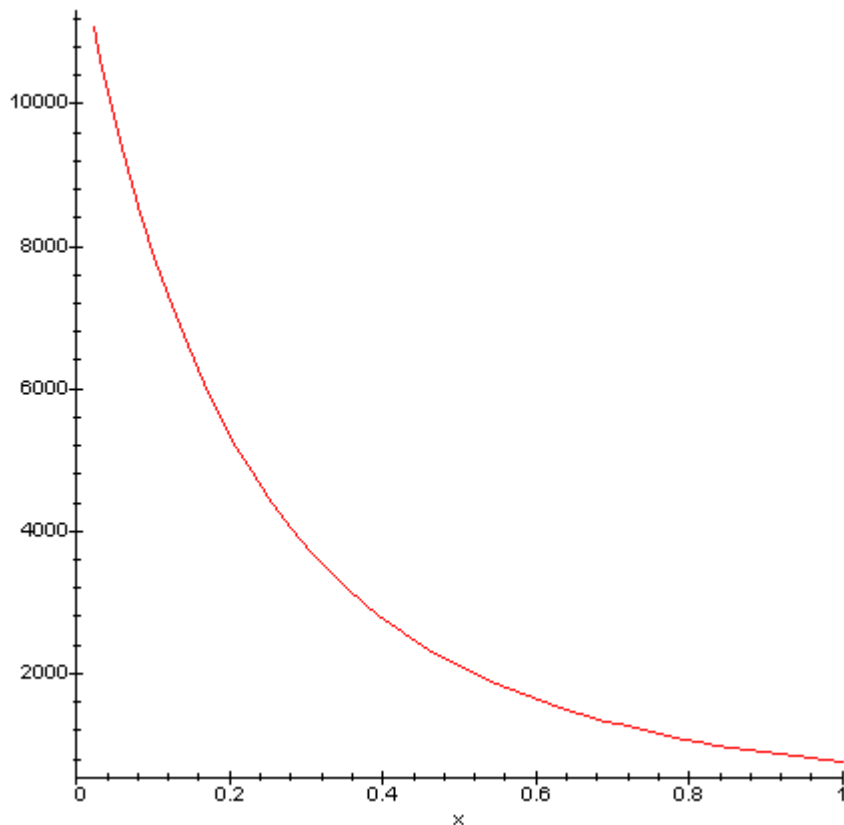


Figure: 66.15 - Valeur actuelle d'une obligation en fonction du taux du marché

E2. Aujourd'hui, nous achetons une obligation de maturité 3 ans, de montant principal 100.-, de taux de coupon 5% et de taux de rendement fixe 10%. Les flux perçus sont alors de 5, 5 et 105 au bout respectivement d'un an, deux ans et trois ans. Le prix de cette obligation est alors égale à:

$$P = 5 \cdot \sum_{i=1}^3 \frac{1}{(1+10\%)^i} + \frac{105}{(1+10\%)^3} = 87.57.- \quad (66.253)$$

Évaluer une obligation revient donc à trouver ce qu'elle devrait valoir en principe dans les conditions actuelles du marché, donc son cours potentiel, par une opération mathématique dite "**opération d'actualisation**" déterminant sa valeur actuelle théorique. Il s'agit donc, comme nous le savons déjà, d'un calcul actuariel.

L'obligataire aura évidemment pour objectif de chercher le taux du marché qui permet de faire de son investissement une action rentable. Ainsi, nous définissons le "**taux de rendement actuariel**" (TRA) x comme étant l'intérêt du marché qui permet de satisfaire les relations suivantes, en fonction de la durée restante à courir n de l'obligation.

Ainsi, à l'émission:

$$E = ca_{\overline{n}|} + Rf_e^n = \frac{c}{1+x} + \frac{c}{(1+x)^2} + \dots + \frac{c}{(1+x)^n} + \frac{R}{(1+x)^n} \quad (66.254)$$

ou à une date quelconque:

$$P = ca_{\overline{n}|} + Rf_e^n = \frac{c}{1+x} + \frac{c}{(1+x)^2} + \dots + \frac{c}{(1+x)^n} + \frac{R}{(1+x)^n} \quad (66.255)$$

Le taux de rendement actuariel d'une obligation est donc le taux x qui annule la différence entre la valeur du prix d'émission E et la valeur actuelle des flux futurs qu'elle génère. Ce taux est calculé au jour du règlement et figure obligatoirement dans les brochures d'émission. Pour l'acheteur de l'obligation, le taux actuariel représente le taux de rentabilité qu'il obtiendrait en gardant l'obligation jusqu'à son remboursement et en réinvestissant les intérêts au même taux actuariel.

Maintenant, passons à quelques éléments mathématiques des obligations zéro-coupon de nominal C , de prix d'émission

Voyons quelques autres définitions utiles relatives aux obligations:

Définitions:

D1. Le "**coupon échu**" (C.E.) d'une obligation est payé à son propriétaire sous déduction de $LA\%$ d'impôts anticipés (IA étant l'abréviation de: Impôts Anticipés). Ainsi, le calcul du coupon net annuel d'obligations à X .- (valeur monétaire) à rendement de $Y\%$ est trivialement donné par :

$$CE = X \cdot Y(1 - LA\%) \quad (66.256)$$

D2. "**L'intérêt couru**" (I.C.) est le montant de l'intérêt qui s'est accumulé depuis la dernière date de paiement de l'intérêt, mais qui n'est pas encore dû. Il est gagné par une obligation depuis sa dernière échéance et est déterminé lors d'une vente ou d'un inventaire. Son calcul est trivialement donné par une application des règles de l'intérêt simple telle que:

$$IC = \frac{X \cdot Y \cdot \Delta d}{360} \quad (66.257)$$

où Δd est bien évidemment le nombre de jours compris entre la date de la dernière échéance et la date de jouissance (l'année commerciale étant définie comme ayant 360 jours).

Remarque: Donc pour obtenir la valeur effective d'une obligation, nous ajoutons à sa valeur cotée (appelée "**prix immaculé**") l'intérêt couru depuis la dernière échéance.

D3. Par extension, si nous cherchons à calculer la valeur nette de X coupons à $Y\%$ dont la valeur nominale vaut Z avec un impôt anticipé de $IA\%$, nous calculons le "**coupon annuel net à l'échéance**" (C.A.E.) par la relation triviale:

$$CAE = X \cdot Y\% \cdot Z \cdot (1 - IA\%) \quad (66.258)$$

Contrairement au calcul de l'intérêt couru, le calcul du dividende couru est impossible. Le cours de l'action est toutefois influencé par la date plus ou moins proche du paiement du dividende.

7.1.3. BONS DE SOUSCRIPTION

Définition: Un "**bon de souscription**", également appelé "**option de souscription**" ou "**stock-option**", est un titre financier permettant (donc il n'y a pas obligation!) de souscrire pendant une période donnée, dans une proportion et à un prix fixé à l'avance (souvent une moyenne des cours de la Bourse avant l'émission des bons), à un autre titre financier sous-jacent (action, obligation, voire un autre bon...).

Le bon permet donc d'être intéressé à la hausse ou à la baisse d'une action sans avoir à y consacrer le même montant de capitaux qu'en achetant directement des actions. Ainsi, lors de l'acquisition, si le titre sous-jacent à une valeur plus élevée que sur le bon de souscription, l'acquéreur fera un bénéfice qui est appelé "**plus-value d'acquisition**". Ensuite, l'acquéreur qui possède maintenant les titres sous-jacents peut très bien vendre ceux-ci lorsque le prix est plus élevé que lorsqu'il en a fait l'acquisition et cela engendre alors un (pseudo) second bénéfice appelé "**plus-value de cession**".

Un bon de souscription peut être donc attaché à l'émission d'une action ou d'une obligation. Alors, selon les cas, nous parlons "**d'actions à bons de souscription d'actions**" (ABSA) ou "**d'obligations à bons de souscription d'actions**" (OBSA) mais également "**d'obligations à bons de souscription d'obligations**" (OBSO) ou "**d'actions à bons de souscription d'obligations**" (ABSO).

Dès l'émission de ces valeurs composées, le tout se scinde en parties: les actions ou les obligations redeviennent des titres classiques et les bons acquièrent une vie propre. Ils sont cotés séparément après l'émission.

Les "**plans de souscription**", plus connus sous le nom de "**plan de stock-options**", sont des paquets d'émission de bons de souscription (nominatifs) destinés aux employés méritant d'une entreprise et visent très souvent à renforcer l'association au développement entre cette même entreprise et ses salariés. Ainsi, ces derniers lors de l'acquisition des titres seront des actionnaires à part entière, recevant des dividendes et pouvant participer aux assemblées des actionnaires. Ce qui est censé accroître la motivation de l'employé (...).

Par ailleurs, les stock-options (données par l'entreprise), sont des actifs financiers sans risques puisqu'il n'y a aucune obligation de les appliquer et qu'ils ont été offerts... Précisons aussi que bon nombre d'entreprises annulent les bons de souscription des employés qui les quittent...

Exemple:

Le bon de la société X permet de souscrire à une action de cette société au prix de 500.- jusqu'au 30 avril 2004. Si l'action X dépasse le niveau de 525.-, le bon qui permet de se procurer une action à un coût inférieur au cours de Bourse se révèle un placement gagnant. Si l'action X vaut donc par exemple 525.- en avril 2004, le gain vaudra 25.-.

Remarque: Le développement de la liquidité sur les marchés d'actions et d'obligations a incité les établissements financiers à émettre des bons de souscription permettant de faire l'acquisition de titres financiers (sous-jacents) existants indépendamment des opérations financières de la société concernée. Sauf exception, ceux-ci ne concernent que les investisseurs et sont émis uniquement par les banques entre elles et ne permettent donc pas le financement d'entreprises (il s'agit donc de pure spéculation!). Ces bons (également cotés) sont fréquemment appelés "**warrants**" (Warrant Call ou Warrant Put) ou, plus précisément "**covered warrants**" (warrants couverts) car, dès l'émission, l'établissement financier se couvre en rachetant des titres sur le marché (les warrants sont basés sur le même principe que les options et utilisent les mêmes outils mathématiques).

D'un point de vue conceptuel, un bon est assimilable à une option d'achat (Call) vendue par une société sur des actions à émettre ou existantes (voir plus loin la définition détaillée de ce qu'est une option). Le prix d'exercice de cette option est le prix auquel le détenteur du bon peut acheter le titre financier correspondant et l'échéance de l'option est celle du bon.

Cependant, l'évaluation d'un bon présente quelques particularités par rapport à une option:

- Un bon a généralement une durée de vie longue (2 à 4 ans) et rend difficilement acceptable l'hypothèse de constance des taux d'intérêt utilisée par le modèle de Black & Scholes (voir la démonstration de ce modèle plus loin).
- Toute opération de l'entreprise émettrice qui modifie la valeur du titre sous-jacent affecte la valeur du bon. Effectivement, les entreprises ont le droit de réserve légal d'émettre un nouveau contrat pour les bons de souscription et d'en changer la valeur et la période de temps de validité!
- Si le titre sous-jacent est une obligation, son prix évoluant dans le temps et sachant que plus une obligation se rapproche de son échéance, plus sa valeur tend vers son prix de remboursement. Sa volatilité se réduit progressivement ce qui rend inapplicable le modèle de Black & Scholes qui postule la constance de la volatilité dans le temps!

Les opérateurs utilisent alors des modèles dérivés de Black & Scholes pour remédier à ces lacunes et évaluer le prix des bons de souscription.

7.1.4. CONTRATS À TERME

Ce type de contrat est symétrique, c'est-à-dire que chaque contrepartie a autant de chance de gagner que de perdre de l'argent.

Définition: Un "contrat à terme", appelé "future" sur les marchés anglo-saxons (ou "forward" / "commodity forward" lorsqu'il s'agit de contrats non standardisés négociés hors des marchés organisés) est un contrat d'achat ou de vente d'un produit financier, passé entre deux contreparties, dont toutes les caractéristiques sont fixées à l'avance: date de règlement, prix à terme, etc. Le prix conclu est appelé "prix à terme" ou aussi... "prix forward", et l'échange se fera à ce prix obligatoirement quel que soit le prix courant du marché du sous-jacent (dit aussi "spot price") à la date de livraison!

Deux types d'exécution peuvent se produire:

- Les "physical settlement" (règlement physique): le sous-jacent est effectivement échangé (ce qui est rare dans l'économie virtuelle mais devrait dans l'économie réelle être... une réalité).
- Les "cash settlement" (règlement numéraire): si le cours du sous-jacent est en-dessous du prix fixé, l'acheteur (du contrat à terme) se fournit sur le marché qu'il veut et verse la différence au vendeur et inversement.

Un contrat à terme doit faire référence au rôle de chaque intervenant (acheteur ou vendeur), une référence officielle de marché (actions, indices, obligations, marché de changes, taux d'intérêt, etc.) dont le prix au temps t est noté S_t , la date future de référence de relevé de la référence T , le prix du contrat à terme K , le montant notionnel sur lequel porte la transaction du contrat à terme N (qui est donc la quantité mais qui mathématiquement est presque toujours ramenée à l'unité), les instructions de paiement (physical ou cash settlement).

L'intérêt des contrats à terme pour les intervenants est de figer des cours dans le futur: il s'agit dans ce cas d'une opération de couverture.

A l'échéance, le gain ou "payoff" du contract est:

$$FC_T = S_T - K \quad (66.259)$$

Cette valeur correspond tout simplement aux pertes ou profits latents (on inversera le signe suivant que l'on est vendeur ou acheteur).

Exemple:

Un industriel suisse sait qu'il doit recevoir en euros une forte somme d'argent dans six mois. Pour se couvrir contre une baisse de l'euro, il achète un contrat de vente à terme, d'échéance 6 mois sur l'euro, en francs suisses (si les opérateurs financiers savent aussi que l'euro va baisser, le contrat de vente à terme aura un coût suffisamment élevé pour que l'achat du contrat ne soit pas intéressant). Notons que cette opération de couverture du risque de change peut lui être défavorable si dans six mois, le contrat cède moins que le taux de change.

Le pricing des prix de forward, dans le cadre d'une approche naïve, est assez simple dans l'hypothèse où l'actif sous-jacent ne verse pas de revenus (dividendes ou autres) que les investisseurs peuvent placer ou emprunter des capitaux sur le marché. Dès lors, le prix du forward à temps nul ne va logiquement pas être inférieur à S_0 (à moins que l'émetteur soit un très bon ami...). Le prix va être au mieux logiquement égal, pour éviter toute opportunité d'arbitrage (autre hypothèse...), actualisé au prix du sous-jacent au taux $t\%$ du rendement géométrique moyen sans risque du marché tel que si la durée du contrat est T (sous-entendu un certain nombre de périodes) nous ayons:

$$K(0) = S_0 (1+t\%)^T \quad (66.260)$$

et donc si déjà l'équivalent de t périodes de temps se sont déjà écoulées, le prix du forward sera alors donné par (en adoptant cette fois la notation traditionnelle r pour le taux du rendement géométrique moyen sans risque du marché):

$$K(t) = K_t = S_t (1+r)^{T-t} \quad (66.261)$$

Ainsi, le prix de forward K (donc le prix à terme) est logiquement plus égal ou plus grand au sous-jacent et ce afin de refléter naturellement la valeur temporelle de la monnaie.

Dans la littérature, l'unité de période de temps étant souvent prise comme étant un petit élément de temps, nous retrouvons souvent cette dernière relation sous la forme suivante en utilisant l'intérêt composé continu (ne pas oublier que rigoureusement il s'agit d'une approximation dans la pratique!):

$$K_t = S_t e^{r(T-t)} \quad (66.262)$$

Il ne faut pas oublier que le terme r a dans la pratique très souvent une certaine volatilité. Ainsi, le prix du forward a donc in extenso aussi une certaine volatilité!!

Nous n'étudierons pas plus la mathématique des opérations et de la détermination des prix ("**pricing**") des contrats à terme (futures) maintenant, car comme ils ne sont qu'un cas particulier des options que nous allons de suite voir en détail, il serait redondant de présenter deux fois tous les théorèmes alors qu'il n'y a qu'une petite modification à faire par exemple dans le modèle de Black & Scholes pour obtenir l'un à partir de l'autre.

Effectivement, comme nous allons le voir une option est un produit dérivé optionnel qui ne donne pas forcément lieu à une exécution de part son caractère optionnel par rapport à un contrat à terme ferme (futur, forward...) ou l'exécution est fixée à l'avance.

Une entreprise va préférer des futures à des options si son but est de connaître ses coûts à l'avance alors qu'elle est certaine à 100% de consommer le sous-jacent. Alors que l'achat d'une option qui demande le paiement d'une prime de risque génère alors un coût d'opportunité. Dans le cas où l'entreprise n'est pas sûre de consommer le sous-jacent, elle devra préférer l'option qui peut être exécutée ou pas, ce qui fait que le management se réserve un bon résultat quoi qu'il se passe mais alors paie la prime de risque de l'option.

Étudiant maintenant un cas intéressant de couverture du risque de la variation des prix des matières premières. Imagions pour cela qu'une entreprise à besoin d'acheter pour une somme totale S une matière première mais pour laquelle il n'existe malheureusement pas de contrats à terme (et aucun partenaire n'accepte de la faire gré à gré). Cette entreprise souhaite se protéger contre une variation haussière du prix de la matière première qui l'intéresse et alors deux stratégies s'offre à elle:

S1. Trouver des contrats à terme d'une commodity positivement corrélée à la matière première intéressée. Dès lors, en espérant que la commodity réplique bien le comportement de la matière première d'intérêt et que si son prix augmente alors l'autre augmentera très probablement aussi (si possible dans les mêmes proportions).... et l'idée est d'avoir une position d'acheteur de contrats à terme afin d'acheter la commodity à une valeur inférieure à celle des prix du marché et espérer trouver quelqu'un afin de la revendre aux plus pressants et utiliser la différence du gain pour acheter la matière première qui nous intéresse en subissant un minimum de pertes (voir même une perte nul ou même un gain!).

S2. Trouver des contrats à terme d'une commodity négativement corrélée à la matière première intéressée. Dès lors, en espérant que la commodity réplique bien de façon inverse le comportement de la matière première d'intérêt et que si son prix diminue alors l'autre augmentera très probablement aussi (si possible dans les mêmes proportions).... et l'idée est d'avoir une position de vendeurs de contrats à terme afin de vendre la commodity à une valeur supérieure à celle des prix du marché et utiliser la différence du gain pour acheter la matière première qui nous intéresse en subissant un minimum de pertes (voir même une perte nul ou même un gain!).

Mathématiquement la stratégie consiste donc à écrire dans les deux cas (nous verrons plus loin quel signe correspondant à quelle stratégie):

$$\Delta V = \Delta S \pm N \Delta F \quad (66.263)$$

où ΔS est donc la variation du prix de la matière première d'intérêt sur le marché et ΔF le gain des contrats à termes qui repliquent positivement ou négativement (suivant la stratégie S1 ou S2) l'augmentation de du prix de S . Nous avons ΔV qui est la différence de la bonne ou mauvaise réplique et qui idéalement doit tendre vers zéro. Une autre manière plus conforme de voir les choses relativement aux informations disponibles du marché est de prendre la variance de cette dernière relation:

$$\sigma_{\Delta V}^2 = \sigma_{\Delta S}^2 + N^2 \sigma_{\Delta F}^2 \pm 2N \sigma_{\Delta S, \Delta F} \quad (66.264)$$

et de chercher la valeur de N qui minimise la variance de ΔV en fonction de la corrélation des deux commodities et de leur volatilité. Nous avons alors en dérivant par rapport à N :

$$\frac{\partial \sigma_{\Delta V}^2}{\partial N} = 2N \sigma_{\Delta F}^2 \pm 2 \sigma_{\Delta S, \Delta F} = 0 \quad (66.265)$$

Ce qui donne en utilisant la relation entre la covariance et le coefficient de corrélation linéaire démontré dans le chapitre de Statistiques:

$$N^* = \pm \frac{\sigma_{\Delta S, \Delta F}}{\sigma_{\Delta F}^2} = \pm \frac{c_{\Delta S, \Delta F}}{\sigma_{\Delta F}^2} = \pm \frac{R_{\Delta S, \Delta F} \sigma_{\Delta S} \sigma_{\Delta F}}{\sigma_{\Delta F}^2} = \pm R_{\Delta S, \Delta F} \frac{\sigma_{\Delta S}}{\sigma_{\Delta F}} \quad (66.266)$$

Comme les écart-types sont positifs et que N doit obligatoirement être positif (et arrondi à l'entier le plus proche pour avoir une solution physique réaliste), il va de soi que le coefficient de corrélation doit toujours être de signe inverse au type de stratégie choisie. Raison pour laquelle finalement on a pour tradition (car cela revient toujours au même) de prendre que le signe positif tel que:

$$N^* = R_{\Delta S, \Delta F} \frac{\sigma_{\Delta S}}{\sigma_{\Delta F}} \quad (66.267)$$

et de prendre toujours la coefficient de corrélation comme étant toujours positif. Donc N^* est le nombre de contrats à terme qui quelle que soit la stratégie, minimisera la variance totale. Il est cependant d'usage de ramener cette expression telle que nous ayons non pas les variance de la variation des prix, mais de la variation des taux. Ainsi, comme nous avons:

$$\begin{aligned} \underbrace{S}_{\substack{\text{prix total} \\ \text{quantité matière} \\ \text{première}}} &= \underbrace{Q_S}_{\substack{\text{quantité} \\ \text{de matière} \\ \text{première}}} \cdot \underbrace{s}_{\substack{\text{prix à l'unité} \\ \text{de la matière} \\ \text{première}}} \\ \underbrace{F}_{\substack{\text{prix de revient} \\ \text{des contrats à} \\ \text{terme}}} &= \underbrace{Q_F}_{\substack{\text{quantité} \\ \text{de contrats} \\ \text{à terme}}} \cdot \underbrace{f}_{\substack{\text{prix à l'unité} \\ \text{d'un contrat} \\ \text{à terme}}} \end{aligned} \quad (66.268)$$

Il vient alors:

$$\begin{aligned} \sigma_{\Delta S} &= \sigma(\Delta S) = \sigma(\Delta(Q_S \cdot s)) = Q_S \sigma(\Delta s) = s Q_S \sigma\left(\frac{\Delta S}{s}\right) \\ \sigma_{\Delta F} &= \sigma(\Delta F) = \sigma(\Delta(Q_F \cdot f)) = Q_F \sigma(\Delta f) = f Q_F \sigma\left(\frac{\Delta F}{f}\right) \end{aligned} \quad (66.269)$$

Donc:

$$N^* = R_{\Delta S, \Delta F} \frac{\sigma_{\Delta S}}{\sigma_{\Delta F}} = R_{\Delta S, \Delta F} \frac{s Q_S \sigma\left(\frac{\Delta S}{s}\right)}{f Q_F \sigma\left(\frac{\Delta F}{f}\right)} = \left(R_{\Delta S, \Delta F} \frac{\sigma\left(\frac{\Delta S}{s}\right)}{\sigma\left(\frac{\Delta F}{f}\right)} \right) \frac{s Q_S}{f Q_F} \quad (66.270)$$

Ce qu'il est parfois d'usage de noter sous forme condensée:

$$N^* = \beta_{\Delta S, \Delta F} \frac{s Q_S}{f Q_F} = \beta_{SF} \frac{s Q_S}{f Q_F} \quad (66.271)$$

La variance de V devient alors en utilisant cet optimum (nous reprenons le double signe qui va à nouveau nous être utile):

$$\begin{aligned}\sigma_{\Delta V}^{*2} &= \sigma_{\Delta S}^2 + N^{*2} \sigma_{\Delta F}^2 + 2N^* \sigma_{\Delta S, \Delta F} = \sigma_{\Delta S}^2 + \left(\pm \frac{\sigma_{\Delta S, \Delta F}}{\sigma_{\Delta F}^2} \right)^2 \sigma_{\Delta F}^2 + 2 \left(\pm \frac{\sigma_{\Delta S, \Delta F}}{\sigma_{\Delta F}^2} \right) \sigma_{\Delta S, \Delta F} \\ &= \sigma_{\Delta S}^2 + \frac{\sigma_{\Delta S, \Delta F}^2}{\sigma_{\Delta F}^2} \pm 2 \frac{\sigma_{\Delta S, \Delta F}^2}{\sigma_{\Delta F}^2}\end{aligned}\quad (66.272)$$

Bien évidemment, le seul signe qui a ici un intérêt physique est le signe "-" qui va diminuer la variance globale (puisque c'est l'intérêt des deux stratégies). Nous avons alors:

$$\begin{aligned}\sigma_{\Delta V}^{*2} &= \sigma_{\Delta S}^2 - \frac{\sigma_{\Delta S, \Delta F}^2}{\sigma_{\Delta F}^2} = \sigma_{\Delta S}^2 - \frac{(R_{\Delta S, \Delta F} \sigma_{\Delta S} \sigma_{\Delta F})^2}{\sigma_{\Delta F}^2} = \sigma_{\Delta S}^2 - R_{\Delta S, \Delta F}^2 \sigma_{\Delta S}^2 \\ &= \left(\sigma \left(\frac{\Delta S}{s} \right) s Q_S \right)^2 - R_{\Delta S, \Delta F}^2 \left(\sigma \left(\frac{\Delta S}{s} \right) s Q_S \right)^2 = \left(\sigma \left(\frac{\Delta S}{s} \right) s Q_S \right)^2 (1 - R_{\Delta S, \Delta F}^2)\end{aligned}\quad (66.273)$$

Soit au final:

$$\sigma_{\Delta V}^* = \sigma \left(\frac{\Delta S}{s} \right) s Q_S \sqrt{1 - R_{\Delta S, \Delta F}^2} \quad (66.274)$$

Exemple:

Considérons qu'une compagnie d'aviation à besoin de 10'000 tonnes de carburant (dont le prix actuel est de 277.-/tonne) pour son parc de d'avions dans 3 mois et qu'elle souhaite se prémunir contre une éventuelle augmentation des cours de cette matière première transformée et qu'il n'existe pas de contrats à terme directement pour ce carburant.. L'entreprise souhaite alors se couvrir du risque en utilisant une autre commodity précise pour laquelle il existe des contrats à terme dont la quantité sous-jacente est de 42'000 tonnes (à 0.6903.-/tonne). Nous souhaiterions calculer la diminution de la variance du carburant en ayant un position d'achat de contrats à terme (donc nous faisons un pari à la baisse pour la commodity répliquante) sachant que la variance à trois mois du carburant est estimée à 21.17% et celle de la commodity répliquante de 18.59% sur évidemment la même période et que leur corrélation est en valeur absolue de 0.8242.

Nous avons alors:

$$N^* = \left(R_{\Delta S, \Delta F} \frac{\sigma \left(\frac{\Delta S}{s} \right)}{\sigma \left(\frac{\Delta F}{f} \right)} \right) \frac{s Q_S}{f Q_F} = \left(0.8243 \cdot \frac{0.2117}{0.1859} \right) \cdot \frac{277 \cdot 10'000}{0.6902 \cdot 42'000} \cong 89.7 = 90 \quad (66.275)$$

Comme nous avons pour la volatilité pure du carburant en numéraires:

$$\sigma_{\Delta S} = s Q_S \sigma \left(\frac{\Delta S}{S} \right) = 277 \cdot 10'000 \cdot 0.2117 = 586'409. - \quad (66.276)$$

Et une fois couverte par la commodity répliquante, celle-ci devient:

$$\begin{aligned} \sigma_{\Delta V}^* &= \sigma \left(\frac{\Delta S}{S} \right) s Q_S \sqrt{1 - R_{\Delta S, \Delta F}^2} = 0.2117 \cdot 277 \cdot 10'000 \sqrt{1 - 0.8243^2} \\ &\cong 331'997. - \end{aligned} \quad (66.277)$$

Donc la couverture par des contrats à terme permet de réduire la volatilité numéraire d'environ 43.38%.

7.1.5. OPTIONS

Les options sont des "**actifs conditionnels**" ("**contingent claim**"), c'est-à-dire une forme particulière d'un titre (contrat), donnant à son détenteur contre le paiement d'une somme d'argent le droit, et non l'obligation d'acheter ou de vendre une certaine quantité d'un actif financier (action ou obligation), à ou jusqu'à une date (échéance ou maturité) et à un prix fixé d'avance.

Il s'agit principalement d'un produit dérivé permettant de se couvrir des risques de variations des marchés. Par exemple, si Airbus vend un avion en dollars mais produit dans la zone euro. Le prix de vente est fixé aujourd'hui, mais la vente est réalisée à la livraison! Airbus doit alors se protéger contre le risque du taux de change (qui peut parfois être de 100% en quelques années seulement). En général, les entreprises se protègent de ces risques en achetant auprès des banques des produits dérivés comme des options.

Remarques:

R1. Nous reviendrons plus loin en détail sur les options qui sont des produits dérivés importants.

R2. La principale différence entre les options et les contrats à terme (future) réside dans le fait que les options représentent un droit d'acheter ou de vendre à l'échéance du contrat alors que les futures représentent l'obligation d'exercer le contrat à terme.

Définitions:

D1. Une "**option**" est un produit dérivé qui donne le droit, et non l'obligation, d'acheter ("**option d'achat**", appelée aussi "**Call**") ou de vendre ("**option de vente**", appelée aussi "**Put**") une quantité donnée d'un actif sous-jacent S (action, obligation, indice boursier, devise, matière première, autre produit dérivé, etc.) à un prix fixé d'avance appelé "**strike**" K ou "**prix d'exercice**" E et durant (jusqu'à) un certain temps appelé "**échéance**" ou "**maturité**" T en échange d'une "**prime**" dépendante (C pour les Call ou P pour les Put) de la valeur intrinsèque à la maturité de l'option appelée "**flux**" ou plus souvent "**pay off terminal**" (et par certains "**cible stochastique**"). La détermination de la prime avec des modèles mathématiques est ce que nous appelons le "**pricing**" de l'option. Une option est donc une sorte de "contrat d'assurance" permettant de se prémunir contre les variations des coûts en transférant le risque sur une personne prête à vendre l'option contre une prime de risque... par ailleurs souvent appelée "**prime d'assurance**".

D2. Nous parlons de "**cours spot**" ou plus simplement "**spot**" pour désigner le cours en vigueur de l'actif sous-jacent S lors d'une transaction immédiate de l'option (Call ou Put). Si le sous-jacent consiste en un taux de change de devises, nous parlons alors de "**cours du cross**" ou plus simplement "**cross**".

D3. Nous parlons de "**cours forward**" ou plus simplement "**forward**" pour désigner le cours qui sera en vigueur de l'actif sous-jacent S lors d'une transaction à maturité de l'option (Call ou Put). Nous retombons alors sur la définition d'un contrat à terme telle que vue plus haut.

D4. Si la valeur intrinsèque d'une option est positive par rapport au spot, elle est dite "**dans la monnaie**". Dans le cas de l'achat d'un Call, cela signifie que le prix d'exercice est inférieur au cours spot. Il est donc possible dès lors d'acheter moins cher que le cours du moment à la date d'exécution de l'option (qui est une date comprise entre la date d'émission et de maturité de l'option pour rappel).

D5. Si la valeur intrinsèque d'une option n'est pas avantageuse par rapport au spot, elle est dite "**en dehors de la monnaie**". Dans ce cas, le prix d'exercice est supérieur au cours du spot pour un Call par exemple. Il ne serait dès lors pas judicieux d'exercer ce Call à la date d'échéance (ni avant non plus), car cela reviendrait à acheter plus cher que le cours spot à cette date!

D6. Si la valeur intrinsèque à ce jour du sous-jacent est égale à son cours du spot à maturité, la valeur intrinsèque est nulle et la valeur de l'option est dite "**à la monnaie**" (nous verrons plus loin les implications mathématiques de cela).

Voici un tableau récapitulatif:

Call		Put	
Acheteur d'un Call (long Call)	Vendeur d'un Call (short Call)	Acheteur d'un Put (long Put)	Vendeur d'un Put (short Put)
À le droit, mais non l'obligation, d'acheter la valeur sous-jacente au prix fixé d'avance jusqu'à la date d'échéance	À l'obligation de vendre la valeur sous-jacente au prix fixé d'avance si le Call est exercé	À le droit, mais non l'obligation, de vendre la valeur sous-jacente au prix fixé d'avance jusqu'à la date d'échéance	À l'obligation d'acheter la valeur sous-jacente au prix fixé d'avance si le Put est exercé

Tableau: 66.12 - Différences entre Put et Call

Il y a donc une différence mathématique d'une énorme importance entre les options et les actions/obligations. Effectivement, ces premières ayant une date d'exercice fixée, leur dynamique de prix peut être statistiquement prédictible et ceci d'autant mieux lorsque nous sommes proche de leur date d'achat ou de leur date d'exercice (donc c'est une sorte de parabole que les financiers appellent "**smile de volatilité**"...). Leur volatilité est souvent maximale entre deux et ceci n'est pas applicable pour les actions/obligations car on ne sait jamais au niveau stratégique quand elles seront vendues ou respectivement achetées.

Remarques:

R1. L'utilité de l'existence des options peut être vue comme la création d'actifs financiers permettant d'accroître la volatilité (**écart-type** ou "**loss/gain deviation**") du marché et ainsi son équilibre.

R2. Pour des raisons évidentes, le détenteur ou acheteur d'un contrat d'option est dit être en position longue alors que sa contrepartie, l'émetteur ou vendeur du contrat, est en position courte.

R3. Si l'option peut être exercée à n'importe quel instant précédant l'échéance, nous parlons "**d'option américaine**", si l'option ne peut être exercée qu'à l'échéance, nous parlons "**d'option européenne**". Une option non exercée est considérée comme "**abandonnée**" (perdue). Ces deux familles d'options dans leur version Call/Put sont en pratique regroupées sous le terme "**d'options vanilles**" ("**plain vanilla options**" en anglais).

R4. Parallèlement aux options classiques, apparaissent depuis les années 1990, sur les marchés des options dites "**options exotiques**" caractérisées par le nom du lieu où elles ont été créées et la manière de calculer leur prix d'exercice à l'échéance comme les options asiatiques dont le prix d'exercice est fonction de la moyenne des cours du sous-jacent durant la durée de vie de l'option, ou les options parisiennes qui peuvent être annulées ou activées si le cours reste dans une certaine zone plus d'un temps donnée, etc.

Formalisons un peu plus les choses quand même... mais sans aller trop dans les détails (nous nous les gardons pour l'étude du modèle de Black & Scholes plus loin qui consiste à déterminer le montant de la prime). Ne considérons pour simplifier que des options portant sur un seul sous-jacent ne versant pas de dividendes.

Nous noterons S_t le prix (cours/taux de change) de l'actif sous-jacent de l'option au temps t et de maturité T et ferons abstraction de la différence de notation entre Puts et Calls continentaux (américains et européens).

Imaginons donc un Call, qui donne à son détenteur le droit (mais non l'obligation) d'acheter l'actif sous-jacent à tout moment entre aujourd'hui $t = 0$ et $t = T$ au prix d'exercice K fixé à l'avance. Prenons le cas pratique courant d'une option d'achat (Call) qui protège une entreprise par exemple contre la hausse du taux de change euro/dollar. Acquise aujourd'hui par l'entreprise, elle va lui conférer donc le droit (mais pas l'obligation) d'acheter 1 dollar en échange de K euros (le prix d'exercice ou strike K est une caractéristique fixe du contrat) à la date future T fixée (date de maturité).

Si le taux de change en question vaut S_t à la date t (i.e. 1 dollar = S_t euros), cette assurance revient du point de vue de l'entreprise à percevoir un montant (pay off):

$$\max(S_T - K, 0) \quad (66.278)$$

euros à la maturité T et noté traditionnellement S_T . À tout temps, deux cas se produisent dès lors pour notre acheteur du Call:

1. $S_T < K$: dans ce cas, le Call donne le droit d'acheter au prix K le sous-jacent que nous pourrions acheter moins cher sur le marché. Dans le cas de notre Call pour le change dollar/euro nous aurons donc plutôt intérêt à faire le change au taux du marché plutôt que d'exercer notre Call puisque sinon nous aurons moins d'euros pour un même dollar. Mais nous perdrons la somme déboursée (la prime) d'achat des Call.

2. $S_T > K$: le Call permet d'acheter le sous-jacent moins cher que sur le marché. Nous exercerons donc très probablement le droit (le profit étant la différence entre ces deux prix). Dans le cas de notre Call pour le change dollar/euro nous exercerons notre droit au taux plus avantageux garanti par le contrat d'option (1 dollar = K euros) avec un gain noté traditionnellement:

$$(S_T - K)^+ \quad (66.279)$$

Du point de vue de la contrepartie (vendeur du Call), dans le cas (1) elle ne verse rien à l'acheteur, et tout est oublié (le contrat expire; tout lien contractuel entre les deux parties disparaît). Dans le cas (2), le vendeur est assigné, il doit vendre à sa contrepartie l'action aux prix K . S'il ne détient pas cette action, il doit d'abord l'acheter sur le marché plus cher (au prix $S_T \geq K$). Dans les deux cas la contrepartie a encaissée par contre la prime par unité de Call.

Ainsi, dans le premier cas, l'acheteur et le vendeur ne reçoivent ni ne doivent rien. Dans le deuxième cas, tout se passe comme si l'acheteur de Call achetait l'action sur le marché et recevait au même moment la somme $S_T - K$ (pour le vendeur c'est bien évidemment l'inverse). Donc avec ces produits dérivés c'est le vendeur du Call (ou du Put) qui endosse presque tout le risque du marché et évidemment l'intérêt est grand de neutraliser ce risque en utilisant un formalisme mathématique (le modèle de Black & Scholes).

Voyons un exemple maintenant du point de vue de l'investissement (la prise de risque est flagrante dans cet exemple):

Exemple:

Imaginons le cas d'une action valant actuellement 1'000.- (peu importe la devise) et qu'elle soit supposée augmenter de 12% en une année.

Imaginons aussi qu'un investisseur ait l'alternative d'acheter l'action à 1'000.- ou d'acheter l'option Call à un prix d'exercice de 1'000.- (donc supposé égal au prix de l'action, ce qui n'est pas nécessairement toujours le cas) pour une prime de 40.- (nous verrons plus tard comment calculer les primes). Évidemment, l'investisseur peut alors pour 1'000.- acheter 25 options Call plutôt qu'une seule action.

La question est de trouver l'investissement le plus intéressant: Ainsi, une augmentation de 120.- dans le cas de l'achat d'une action représente un retour sur investissement de 12% par année, alors que l'achat d'une option Call aura un retour sur investissement de 80.- (120.- de gains sur le prix de vente moins 40.- de la prime payée) soit de 200%.

Il apparaît clairement dans cet exemple que la rentabilité d'achat d'un Call à même investissement est nettement supérieure à l'achat de l'action tant que la prime d'option ne dépasse pas un certain seuil.

Maintenant abordons de manière détaillée et par l'exemple un autre concept que nous avons déjà implicitement présenté dans les paragraphes précédents et qui nécessite toute notre attention, car il en est souvent fait mention par les analystes. Il s'agit de "**l'effet de levier**" des options qui est une arme à double tranchant, une arme atomique de destruction massive de portefeuille (si le levien n'est pas calculé à l'avance et ce même approximativement).

Lorsque nous évoquons les options, nous ne retenons souvent que le droit d'acheter ou de vendre un bien ou un instrument financier (à un prix fixé d'avance et durant un certain temps), en négligeant l'obligation correspondante du vendeur de l'option. Or, l'effet de levier qui caractérise ces instruments financiers peut rendre cette obligation dévastatrice pour le vendeur.

Pour voir de quoi il s'agit commençons par le risque des Call.

Exemple:

L'acheteur d'un Call sur une action (par exemple) limite son risque à la prime de l'option pour un gain potentiel illimité. Le vendeur du Call se trouve dans la position exactement inverse: il encaisse la prime de l'option, mais prend un risque illimité.

Prenons une action X cotée 350.- à la mi-octobre. Un investisseur parie sur la hausse du titre et achète 12.50.- (la "prime") une option Call à échéance janvier de l'année suivante au prix d'exercice de 380.-. Une représentation graphique permet de mettre aisément en relation l'évolution du titre (en abscisse) et son effet sur l'acheteur ou vendeur du Call.

Considérons le cas de l'acheteur du Call:

Tant que le cours de l'action reste en dessous de 380.- ("**valeur de levier**"), prix d'exercice, l'acheteur du Call n'aura aucun intérêt à exercer son option, qui est dite "**out of the money**". Par contre, si le cours de l'action progresse et dépasse le prix d'exercice, l'option est dite alors "**in the money**" et il devient intéressant d'exercer l'option. Lorsque le prix d'exercice de l'option est égal au prix du sous-jacent en Bourse, nous disons que l'option est "**at the money**". Dès que le cours de l'action dépasse 392.50.-, soit l'addition du prix d'exercice et de la prime de l'option à la mi-octobre ($380 + 12.50$), le détenteur du Call commence à gagner de l'argent sur son investissement initial. Si le cours du titre monte tout à coup à 500.-, soit une augmentation d'un peu plus de 30%, le gain sera beaucoup plus que proportionnel: pour 12.50.- investis, l'acheteur réalisera un bénéfice de 107.50.- soit un gain de 860%: c'est le fameux "**effet de levier**".

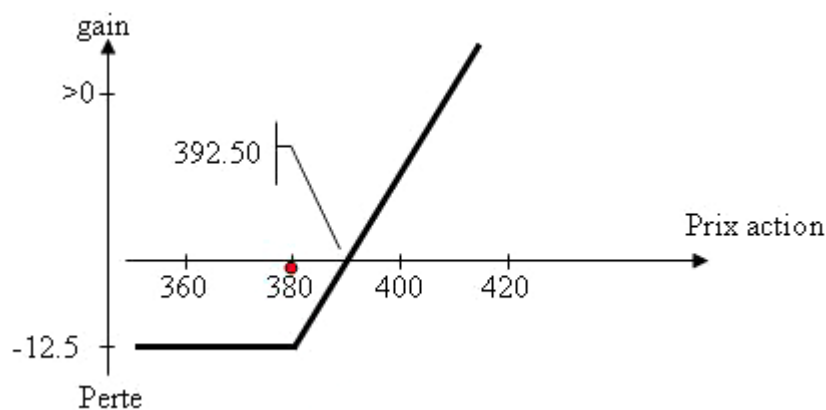


Figure: 66.16 - Effet de levier pour l'acheteur d'un Call

Considérons maintenant le cas du vendeur du Call:

Tant que l'action reste en dessous de 380.- ("**valeur de levier**"), le vendeur du Call fait un bénéfice de 12.50.-, représentant la prime de l'option. À partir de 380.-, le vendeur risque d'être obligé de livrer l'action au prix d'exercice, soit 380.-. À partir de 392.50.-, il commence à perdre de l'argent sur l'opération, puisque l'action qu'il devra sans aucun doute livrer vaudra plus chère que l'addition du prix d'exercice et de la prime encaissée. Si pour son malheur le titre monte effectivement à 500.- et qu'il ne le possède pas, il lui faudra aller le racheter en Bourse pour honorer la demande d'exercice du détenteur du Call, en perdant 107.50.- sur l'opération, soit plus de huit fois la prime encaissée au départ!!!! Avec l'effet de levier, il est donc possible de perdre beaucoup plus que la prime encaissée. Nous disons alors que nous sommes en situation "**d'appel de marge**".

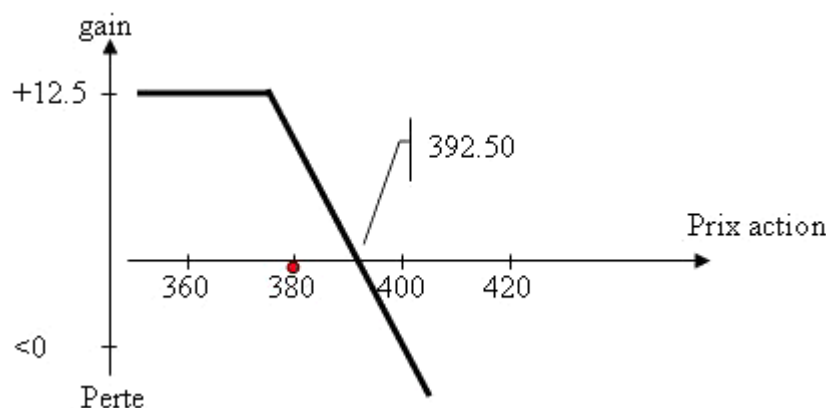


Figure: 66.17 - Effet de levier pour le vendeur d'un Call

Ainsi, si le prix du sous-jacent augmente de 3.29%, le vendeur du Call perd 100%. Nous parlons alors d'un lever de 3.29 pour 100 ou plus simplement d'un levier correspondant au rapport $100/3.29$ soit un levier d'environ 30.39 (il est en de même évidemment pour le gain!).

Maintenant intéressons-nous au risque des Put.

Exemple:

L'acheteur d'un Put limite son risque au coût de la prime de l'option pour un gain potentiel beaucoup plus important. En face de lui, le vendeur du Put se trouve dans la position exactement inverse: il encaisse la prime de l'option mais prend un risque beaucoup plus grand. Si nous prenons la même action X cotée à 350.- à la mi-octobre, nous nous trouvons cette fois avec un investisseur qui parie sur la baisse du cours de l'action. Il achète donc pour 49.50.- (la "prime") un Put d'échéance décembre au prix d'exercice de 390.-.

Considérons le cas de l'acheteur du Put:

L'acheteur du Put commence à réaliser un profit si le prix de l'action tombe en-dessous de 340.50.-, soit le prix d'exercice moins le prix de l'option ($390 - 49.50$). Entre 340.50.- et 390.- l'exercice n'est pas profitable mais permet de diminuer la perte. Au-dessus du prix at-the-money (390.-) l'exercice du Put n'offre vraiment plus aucun intérêt et nous disons alors que l'option Put est out of the money (O.T.M.).

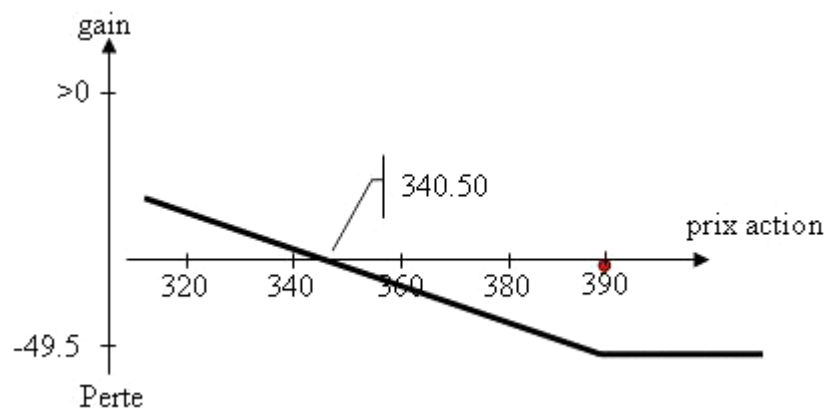


Figure: 66.18 - Effet de levier pour l'acheteur d'un Put

Considérons le cas du vendeur du Put:

Le vendeur du Put encaisse d'abord la prime de l'option soit 49.50.-. Tant que le cours se maintient au-dessus de 390.-, il est gagnant. Si le cours de l'option se situe entre 340.50.- et 390.- il perd un peu de sa prime mais reste gagnant. En-dessous de 340.50 le vendeur du Put sera obligé au moment de l'échéance de verser 390.- à l'acheteur du Put (en vendant le sous-jacent et en versant la différence d'une manière ou d'une autre). Bien évidemment si le prix du sous-jacent tombe à zéro, le vendeur du Put peut ainsi perdre jusqu'à 340.50.- de fonds propres.

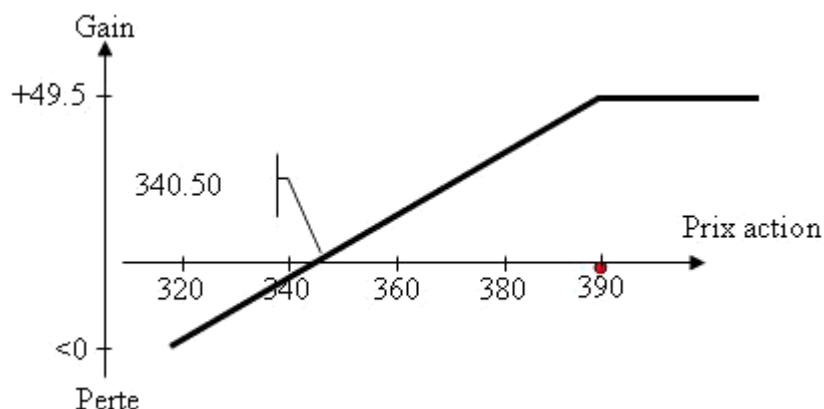


Figure: 66.19 - Effet de levier pour le vendeur d'un Put

Si nous réfléchissons un petit moment, une stratégie possible est d'acheter des Put et des Call du même sous-jacent avec le même strike et la même date d'expiration (échéance). Ce type de stratégie s'appelle un **"straddle"** ou **"straddle purchase"** ou encore **"bottom straddle"**.

Ainsi, si le cours du sous-jacent est proche du strike K à la date d'expiration T alors le straddle engendre une perte nette (car nous n'utiliserons ni le Put, ni le Call et nous aurons perdu la somme équivalant à la prime des options). En revanche, si le cours du sous-jacent est assez loin du strike à l'échéance le profit peut être très important.

C'est de la pure spéculation des banques d'affaires et cela devrait être objectivement interdit mais voyons en quand même le principe via un exemple...

Exemple:

Considérons qu'un sous-jacent qui vaut actuellement 69.- va décaler violemment dans les trois mois qui viennent. Nous pouvons mettre en place une stratégie straddle en achetant à la fois un Call et un Put de strike 70.- et de maturité 3 mois. Supposons que le Call coûte 4.- et le Put 3.-.

Dans le cas où à échéance, la valeur du sous-jacent est celle du strike K , pour le spéculateur c'est la pire situation car il aura perdu la somme des deux primes, soit 7.- car il n'avait n'a aucun intérêt à exercer le Call en utilisant son Put (puisque la différence est nulle entre les deux).

Si le cours du sous-jacent reste à 69.-, nous perdons 6.- car le Call ne vaut plus rien à maturité (du moins comme il fait perdre de l'argent il serait peu judicieux pour un pur spéculateur de l'exercer) et le Put ne vaut plus que 1.- (car on va pouvoir vendre à 70.- quelque chose qui en vaut 69.- et cela fera un petit gain de 1.-). Donc:

$$(-4 + -3) + 1 = -7 + 1 = -6.- \quad (66.280)$$

Si le cours du sous-jacent monte à 90.- (ou descend à 50.-), nous réaliserons un profit de 13.- grâce au Put car:

$$(-4 - 3) + (90 - 70) = 13.- \quad (66.281)$$

Il faut espérer pour le spéculateur qu'il pariera correctement sur une forte variation ignorée de la majeure partie des acteurs du marché, sinon quoi les primes d'options seront suffisamment élevées pour que son gain soit dès lors quand même nul.... Graphiquement le stratégie Straddle est souvent représentée sous la forme suivante (l'axe verticale est le gain, l'axe horizontal est le prix du sous-jacent):

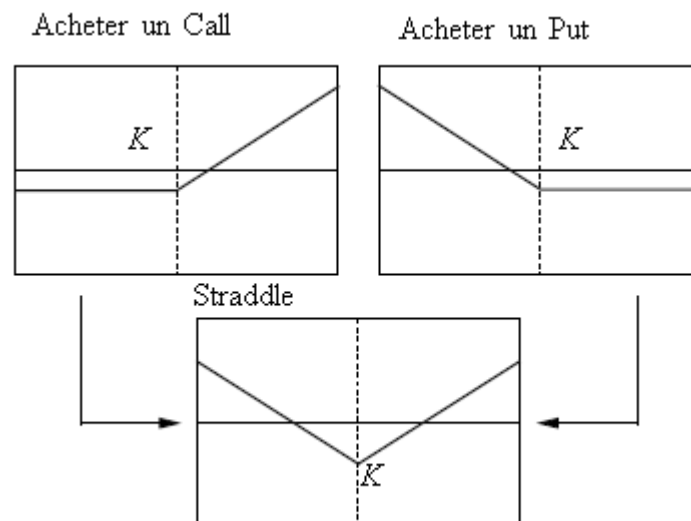


Figure: 66.20 - Stratégie Straddle

Nous pouvons résumer ces quelques propriétés sous la forme du tableau suivant:

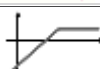
Stratégie	Anticipation du cours	Gain potentiel	Perte potentielle	Profit
Achat de Call	Hausse	"Illimité"	Limitée	
Achat de Put	Baisse	Limité	Limitée	
Vente de Call	Stabilité ou légère baisse	Limité	"Illimitée"	
Vente de Put	Stabilité ou légère hausse	Limité	Limitée	

Tableau: 66.13 - Différentes stratégies des options Call et Put

7.1.6. FONDS DE PLACEMENT

Définition: Un "**fonds de placement**" est un véhicule d'investissement (portefeuille de titres, d'actions ou d'obligations par exemple) que les établissements financiers proposent à leurs clients.

Remarque: Un "**hedge fund**" ou "**fonds couvert**" est un ensemble de produits financiers utilisés comme couverture contre les fluctuations du marché. En théorie, si la Bourse chute, le hedge fund ne descend pas et a une performance absolue. Ces types de fonds alternatifs sont cependant réservés à une clientèle fortunée et avertie.

Bien qu'un fonds de placement réunisse divers actifs financiers, les clients peuvent acheter les parts émises à une faible valeur par rapport à l'achat d'actifs individuels. Chaque part contient théoriquement une proportion de chacun des actifs se trouvant dans le fonds de placement. Elles garantissent un droit de participation à la fortune globale du fonds sans toutefois donner de droit sur les sociétés incluses dans le fonds.

Un fonds de placement peut investir les montants de diverses manières dont les plus communément pratiquées sont les papiers-valeurs (actions, obligations), papiers monétaires, valeurs immobilières, régions (pays, continents), secteurs d'activité ou encore selon des objectifs personnels. Il existe en ce début de 21ème siècle à peu près 30'000 fonds de placement à travers le monde.

Les fonds de placement rendent souvent service aux petits portefeuilles: avec des montants relativement modestes, il est possible de bénéficier d'une bonne répartition des risques et aussi de prix de gros accordés sur les transactions effectuées par les gestionnaires de fonds.

7.2. RETOURS ET TAUX D'INVESTISSEMENTS

Pour définir l'objectif poursuivi par le possesseur d'actifs financiers, nous nous référerons à la motivation économique de tout acte d'investir. Celle-ci consiste concrètement à consentir présentement à une dépense, en vue d'un accroissement de patrimoine espéré dans le futur.

De deux ou plusieurs stratégies d'investissements, la meilleure au niveau individuel est celle qui maximise le capital final de l'investisseur.

Il existe alors différents types de retour sur investissements suivant l'objet d'étude. Ainsi, nous différencions en finance (avant d'en voir les détails):

1. Les retours d'actifs financiers sur un horizon économique (return on investment) et leurs taux de rendement respectifs (rate of return).
2. Les retours sur des investissements en comparaison à un taux géométrique moyen du marché et la limite du taux de rentabilité correspondante (internal rate of return).

Ensuite, il faut considérer d'autres approches de taux de rentabilité. Outre les deux mentionnés ci-dessus les deux autres grands classiques sont (avant d'en voir les détails):

1. Le taux de retour pondéré par les capitaux investis (M.W.R.R.) qui a l'avantage par rapport au taux de rendement interne de prendre en compte les investissements faits en dehors des périodes temporelles classiques.
2. Le taux de retour pondéré dans le temps (T.W.R.R.) qui est un outil pratique pour mesurer la performance des gestionnaires de fonds car il ne prend pas en compte les flux (retraits ou investissements) des investisseurs qui sont incontrôlables.

Voyons donc un peu tout cela:

7.2.1. RETURN ON INVESTMENT

En pratique, nous définirons l'objectif de l'investisseur comme consistant à maximiser l'accroissement de sa fortune initiale, quelles que soient les modalités de cet accroissement. Cet accroissement appelé donc en anglais "**return on investment**" (R.O.I.) ou, plus brièvement, "**return**" est défini par la relation (logique) dans le cadre de la gestion d'actifs par:

$$R_t = (P_t - P_{t-1}) + C_t \quad (66.282)$$

où R_t est donc le return de l'actif financier pour la période (se terminant au temps) t , P_t le prix du marché au temps t de l'actif financier et C_t le revenu liquide attaché à la détention de l'actif financier durant la période (se terminant au temps) t .

Le revenu C_t est supposé perçu au temps t , ou, s'il est perçu entre $(t-1)$ et t , il est supposé ne pas être réinvesti avant le temps t . Le prix de marché P au temps $(t-1)$ est une valeur "**ex-coupon**" c'est-à-dire une valeur enregistrée immédiatement après (le détachement du coupon donnant droit à) la perception, au temps $(t-1)$, du revenu liquide afférant à la période $(t-2)$. Sur le plan empirique, l'hypothèse de non-réinvestissement jusqu'à la période élémentaire de temps utilisée est courte (un mois maximum), afin d'éviter des distorsions statistiques trop importantes dans le traitement des données chronologiques.

Pour faciliter les comparaisons entre investissements, nous utilisons une mesure exprimée en termes relatifs le "**taux de rentabilité**" ou "**rate of return**" défini assez logiquement par:

$$r_t = \frac{(P_t - P_{t-1}) + C_t}{P_{t-1}} \quad (66.283)$$

où r_t est le taux de rentabilité pour la période t .

Nous reviendrons lors de notre étude du modèle mathématique d'évaluation des actifs financiers sur ces outils.

7.2.2. INTERNAL RATE OF RETURN

La mise en oeuvre d'un capital financier pour permettre la réalisation d'opérations d'économie réelle (c'est-à-dire le fait de consacrer, directement ou indirectement, ce capital financier à l'acquisition ou à la constitution de moyens de production, au sens le plus large de ce terme) peut donc produire à travers le temps des retours d'argent sous la forme de flux nets de liquidités appelés "**flux net de trésorerie**" (F.N.T.) ou encore "**cash-flows**" (C.F.) (cela fait toujours mieux en anglais...).

Le calcul actuariel permet de construire formellement un critère de décision. En effet, nous définissons (logiquement mais sans toutefois être complètement réaliste) la prise de risque par le "**goodwill**" celui-ci étant donné par la relation démontrée dans le chapitre de Techniques De Gestion sous la dénomination de "valeur actuelle nette" (le terme "goodwill" étant souvent utilisé par les comptables):

$$G = -I_0 + \sum_{t=1}^T \frac{F_t}{(1 + t\%)^t} \quad (66.284)$$

Explications:

Le deuxième terme à droite de l'égalité nous est déjà connu (nous en avons démontré l'origine dans le chapitre de Technique De Gestion) mais pour rappel sous la forme:

$$C_0 = \sum_{k=1}^n C_k \cdot \frac{1}{(1 + t\%)^k} \quad (66.285)$$

Dans un contexte de certitude de l'avenir (...) il nous donne donc l'investissement initial à effectuer à un pourcentage donné constant (...) pour avoir un retour sur investissement (cash-flow) F_t à un taux d'intérêt périodique moyen géométrique $t\%$ (taux du marché) avec T étant l'horizon de l'opération (nombre de périodes), I_0 étant la dépense initiale d'investissement.

En d'autres termes, le goodwill de l'opération représente les flux excédentaires actuels obtenus après avoir remboursé la somme initiale investie sur sa durée d'utilisation et après avoir rémunéré le capital encore investi au début de chaque période au taux d'actualisation.

Si:

$$\begin{aligned} G > 0 &: \text{accepter l'investissement} \\ G < 0 &: \text{rejeter l'investissement} \\ G = 0 &: \text{niveau d'indifférence} \end{aligned} \quad (66.286)$$

A la formulation du critère de décision telle qu'elle vient d'être présentée, nombreux sont ceux, notamment les praticiens, qui préfèrent la méthode dite du "**taux interne de rentabilité**" (TRI) ou "**internal rate of return**" (I.R.R.). Celle-ci n'est en apparence qu'une variante de la première formulation. Elle consiste à calculer un taux généralement symbolisé par la lettre grecque ρ , qui annule la valeur du goodwill (il s'agit donc de déterminer le taux de rentabilité tel que la somme des flux nets de trésorerie soit égale au montant du capital investi):

$$\rho \rightarrow \sum_{t=1}^T \frac{F_t}{(1 + \rho)^t} - I_0 \quad (66.287)$$

Si:

$$\begin{aligned}\rho > t\% &: \text{accepter l'investissement} \\ \rho < t\% &: \text{rejeter l'investissement} \quad (66.288) \\ \rho = 0 &: \text{niveau d'indifférence}\end{aligned}$$

Nous voyons que le taux interne de rentabilité intervient dans le processus de décision de manière à première vue équivalente à celle dont il est utilisé dans le calcul d'une valeur actuelle nette. En outre, l'expression du résultat du calcul est indéniablement plus parlante que le montant absolu (goodwill) obtenu dans la première formulation. Nous inclinerions donc à adopter la seconde formulation si celle-ci ne présentait, à l'examen approfondi, l'inconvénient majeur que le calcul du taux interne de rentabilité comporte dans certains cas plusieurs solutions. La relation est en effet une équation polynomiale dont nous avons démontré, dans le chapitre d'Algèbre, qu'elle a autant de racines que le polynôme présente de changements de signe.

7.2.3. MONEY WEIGHTED RATE OF RETURN

Nous allons maintenant introduire un type de taux interne de rentabilité différent de celui lié au goodwill et qui s'applique mieux à la gestion de portefeuilles que le taux interne de rentabilité vu plus haut (qui rappelons-le se base sur l'hypothèse que les cash-flows sont déboursés à intervalles périodiques).

Considérons un fonds F et les informations suivantes:

1. La valeur du fonds $F_0 \in \mathbb{R}_+$ juste avant le temps 0.
2. La valeur du fonds $F_1 \in \mathbb{R}_+$ juste après le temps 1.
3. Une valeur monétaire totale nette $N \in \mathbb{R}$ investie durant la période $[0,1]$ versée en deux moitiés en début et fin de période (pour simplifier l'exemple...).

Les données qui vont nous intéresser sont les suivantes:

1. La valeur $F_0 + N/2$ qui représente la valeur totale du fonds et d'une partie de l'investissement au moment 0.
2. La valeur $(1+t\%)^{-1} \cdot N/2$ qui représente le capital qu'il aurait fallu rassembler au temps 0 pour arriver en fin de période à la valeur $N/2$ lorsque le taux du marché est à un taux $t\%$.
3. La valeur $(1+t\%)^{-1} \cdot F_1$ qui représente au temps 0 la valeur du fonds à laquelle on voudrait arriver pour arriver en fin de période à F_1 lorsque le taux du marché vaut aussi $t\%$.

La différence:

$$(1+t\%)^{-1} \cdot F_1 - (1+t\%)^{-1} \cdot \frac{N}{2} = (1+t\%)^{-1} \left(F_1 - \frac{N}{2} \right) \quad (66.289)$$

donne la valeur qu'il aurait fallu capitaliser pour obtenir la somme $\left(F_1 - \frac{N}{2} \right)$.

Ce qui est trivialement intéressant pour un investisseur est alors de connaître le taux $t\%$ tel que le premier investissement égalise son objectif annuel. C'est-à-dire:

$$F_0 + \frac{N}{2} = (1+t\%)^{-1} \left(F_1 - \frac{N}{2} \right) \quad (66.290)$$

soit:

$$F_0 + \frac{N}{2} - (1+t\%)^{-1} \left(F_1 - \frac{N}{2} \right) = 0 \quad (66.291)$$

relation qui est appelée "**relation de Hardy**".

Si cette relation se vérifie pour F_0, N connus et déterminés et un F_1 supposé, un investisseur n'aura rien à gagner ni à perdre à investir dans le fonds ou de capitaliser au taux du marché $t\%$.

Si l'équation de Hardy n'est pas non nulle, mais positive alors l'investissement dans le fonds n'est pas intéressant. Si elle est négative, il vaut alors mieux investir dans le fonds.

De l'algèbre élémentaire, nous conduit à la relation:

$$F_0 + \frac{N}{2} - (1+t\%)^{-1} \left(F_1 - \frac{N}{2} \right) = 0 \Rightarrow t\% = \frac{2I}{F_0 + F_1 - I} \quad (66.292)$$

avec:

$$I = F_1 - F_0 - N \quad (66.293)$$

Effectivement:

$$t\% = \frac{F_1 - \frac{N}{2}}{F_0 + \frac{N}{2}} - 1 = \frac{F_1 - \frac{N}{2} - F_0 - \frac{N}{2}}{F_0 + \frac{N}{2}} = \frac{2(F_1 - F_0 - N)}{2F_0 + N} = \frac{2(F_1 - F_0 - N)}{F_0 + F_1 - (F_1 - F_0 - N)} \quad (66.294)$$

Le taux $t\%$ est souvent nommé en gestion de fortune le "**Money Weighted Rate of Return**" (M.W.R.R.) ou "**Taux de Retour Pondéré par les Capitaux Investis**" (T.R.P.C.I.) et représente donc la performance d'un fonds (portefeuille) avec prise en compte des entrées et sorties de capitaux au cours de la période d'évaluation.

Exemple:

Un fonds a eu les revenus suivants pendant l'année 2006:

- Valeur au 1er Janvier 2006: 30 MFr.-
- Investissement dans le fonds pendant l'année: 18 MFr.-
- Retraits sur le fonds: 30 MFr.-

- Valeur du fonds au 31 décembre 2006: 21 MFr.-

Quel est le taux effectif (M.W.R.R.) de ce fonds en 2006 ?

Nous avons alors comme données initiales $F_0 = 30$, $F_1 = 21$, $N = 18 - 30 = -12$ ce qui donne si nous assumons les hypothèses de départ concernant N :

$$I = 21 - 30 - (-12) = 3 \quad (66.295)$$

et alors:

$$MWRR = \frac{2 \cdot 3}{21 + 30 - 3} \cong 12.5\% \quad (66.296)$$

Considérons maintenant que nous savons que les investissements ont eu lieu la 3/8ème part et les retraits la 3/4ème part de l'année.

Le M.W.R.R. est alors le taux du cash-flow:

$$c = \{(0, 30), (3/8, 18), (3/4, -30), (1, 21)\} \quad (66.297)$$

Nous devons alors trouver $t\%$ tel que:

$$30 = 21(1+t\%)^{-1} - 18(1+t\%)^{-3/8} + 30(1+t\%)^{-3/4} \quad (66.298)$$

La résolution de cette équation avec Maple 4.00b donne:

$$t\% \cong 8.875 \quad (66.299)$$

Nous voyons qu'en considérant les cash-flows et les moments où ils ont lieu (donc une analyse plus fine et rigoureuse) nous réduisons le M.W.R.R. Par ailleurs, le dernier calcul étant plus rigoureux que le premier, c'est celui que l'investisseur voudra connaître en fin d'année.

Ce taux est donc une mesure effective du taux d'accroissement du fonds, donnant l'impact du poids des cash-flows sur la valeur du fonds. Il s'agit aussi au fait d'une simple généralisation du de l'IRR (Internal Rate of Return).

Remarque: En Suisse, certaines fondations d'assurance de la LPP (Loi sur la Prévoyance Professionnelle) communiquent la M.W.R.R. annuelle ou trimestrielle sur les 5 à 10 ans passés.

7.2.4. TIME WEIGHTED RATE OF RETURN

Nous allons maintenant nous intéresser à un autre outil financier de la gestion de portefeuilles utilisé également pour juger du rendement d'un investissement.

Considérons un fonds tel que:

	Décembre 31. 2000	T1 2001	T2 2001	T3 2001	T4 2001
Valeur de début du fonds		1000	370	81	7.8
Gain ou (perte) pour le trimestre en %		10%	3%	(4%)	6%
Gain ou (perte) pour le trimestre .-		100	11.1	(3.2)	0.5
Cash-flows trimestriels entrées/(sorties)		(730)	(300)	(70)	0
Valeur du fonds	1000	370	81.1	7.8	8.3

Tableau: 66.14 - Time Weighted Rate Of Return

Le 31 décembre 2000, le fonds a une valeur de 1000.-. Durant le premier trimestre 2001 il a un retour de 10% mais nous imaginons que cette valeur est loin de celle qui était attendue alors l'investisseur retire 730.- du fonds (portefeuille basé sur le fonds). Lors du second trimestre, le fonds a gagné 3% et 300.- supplémentaires ont été retirés par l'investisseur. Lors du troisième trimestre le fonds a perdu 4% et 70.- ont été retirés. Le dernier trimestre, le fonds a gagné 6% et aucun fonds n'a été retiré.

Nous avons alors l'accroissement (retour) global sur l'ensemble de la période (année) qui est donné par:

$$TWRR = (1 + 0.1)(1 + 0.03)(1 - 0.04)(1 + 0.06) - 1 = 15.3\% \quad (66.300)$$

Nous voyons bien que cette valeur est indépendante des flux monétaires du portefeuille de l'investisseur. Nous appelons la valeur de 15.3% le "**Time Weighted Rate of Return**" (T.W.R.R.) ou "**Taux de Retour Pondéré dans le Temps**" (T.R.P.T.) sur une base trimestrielle.

En d'autres termes, le T.W.R.R. indique la performance d'un portefeuille sans prise en compte des entrées et sorties de capitaux au cours de la période d'évaluation.

Ce cas particulier peut être noté de manière générale par la relation:

$$TWRR = \prod_{i=1}^n (1 + R_i) - 1 \quad (66.301)$$

Il convient de se rappeler que si nous avions voulu calculer la moyenne du rendement du fonds par trimestre, nous aurions simplement utilisé la moyenne géométrique (cf. chapitre de Statistiques)!

Le T.W.R.R. est un outil pratique pour mesurer la performance des gestionnaires de fonds, car il ne prend pas en compte les flux (retraits ou investissements) des investisseurs qui sont incontrôlables. Ainsi, nous avons une mesure de la qualité de la dynamique des fonds indépendante du choix des investisseurs qui pourraient considérer les retraits ou investissements comme des cash-flow qui serviraient à calculer un I.R.R. qui n'aurait plus ou moins aucune signification par rapport à la dynamique du fonds.

Remarque: En Suisse, certaines fondations d'assurance de la LPP (Loi sur la Prévoyance Professionnelle) communiquent la T.W.R.R. annuelle ou trimestrielle sur les 5 à 10 ans passés mais sur une base journalière.

7.3. MODÈLE SPÉCULATIF DE BACHELIER

Après ces nombreuses définitions contextuelles, le but maintenant est d'introduire les techniques mathématiques spéculatives stochastiques de base utilisées en finance. En effet, la finance étant devenue au fil du temps un domaine de plus en plus concurrentiel, les marges sur les produits standards ont tendance à se réduire, la prime est donc donnée à l'innovation. Cette évolution a conduit à une sophistication croissante des produits financiers, faisant ainsi appel à des notions mathématiques poussées, basées principalement sur des modèles de probabilités, introduits par Louis Bachelier dans sa "Théorie de la spéculation" mais réellement utilisées que depuis 1973 grâce aux différents travaux de Black & Scholes et Merton (qui ont valu à leurs auteurs un Prix Nobel d'économie).

Regardons pour commencer quels sont les développements proposés par Louis Bachelier dans sa thèse pour déterminer l'espérance mathématique prévisionnelle et l'écart-type prévisionnel d'un actif financier (résultat que nous utiliserons dans le cadre de l'étude du modèle d'évaluation de Black & Scholes).

Désignons par $p_{x,t}$ la fonction de densité de probabilité que le cours d'un actif soit x au temps t . Dès lors, la probabilité cumulée que la valeur du cours se trouve comprise dans l'intervalle élémentaire $[x, x + dx]$ au temps t est de la forme:

$$p_{x,t} dx \quad (66.302)$$

dont l'intégrale sur l'ensemble du domaine de définition devra donner 1.

En vertu du quatrième axiome des probabilités (voir chapitre du même nom), la probabilité que le cours évolue d'une certaine valeur à une autre (chaîne de Markov temporelle à temps continu), sera égale au produit de la probabilité cumulée pour que le cours soit coté x dans un intervalle donné à l'époque t_1 , c'est-à-dire:

$$p_{x,t_1} dx \quad (66.303)$$

multipliée par la probabilité cumulée pour que, le cours étant coté x à l'époque t_1 , le cours soit coté z dans un intervalle donné à l'époque $t_1 + t_2$, c'est-à-dire, multipliée par:

$$p_{z-x,t_2} dz \quad (66.304)$$

La probabilité cherchée est donc:

$$p_{x,t_1} p_{z-x,t_2} dz dx \quad (66.305)$$

Cette écriture suppose donc que les cours sont des variables aléatoires indépendantes...

Le cours pouvant se trouver à l'époque t_1 dans tous les intervalles dx compris entre $\pm\infty$, la probabilité cumulée pour que le cours soit coté z à l'époque $t_1 + t_2$ sera:

$$\int_{-\infty}^{+\infty} p_{x,t_1} p_{z-x,t_2} dz dx \quad (66.306)$$

La probabilité de ce cours z , à l'époque $t_1 + t_2$ a aussi pour expression:

$$p_{z,t_1+t_2} dz \quad (66.307)$$

Nous avons donc:

$$p_{z,t_1+t_2} dz = \int_{-\infty}^{+\infty} p_{x,t_1} p_{z-x,t_2} dz dx \quad (66.308)$$

ou:

$$p_{z,t_1+t_2} = \int_{-\infty}^{+\infty} p_{x,t_1} p_{z-x,t_2} dx \quad (66.309)$$

telle est l'équation à laquelle doit satisfaire la fonction de distribution de probabilité p que nous recherchons. Cette équation est vérifiée, comme nous allons le voir, par la fonction:

$$p = A e^{-B^2 x^2} \quad (66.310)$$

Mais il ne faut pas oublier qu'il s'agit d'une solution particulière (raison pour laquelle ce modèle est parfois appelé "**modèle gaussien de Bachelier**")... et de plus rien ne dit que les deux variables aléatoires indépendantes suivent la même loi de probabilité...

Les deux hypothèses de construction du modèle vues jusqu'à maintenant (indépendance et distribution identique) sont souvent indiquées en finance sous l'appellation des "**hypothèses d'indépendance et de stationnarité**".

Ceci étant dit, nous devons alors bien évidemment imposer (axiomes des probabilités obligent!):

$$\int_{-\infty}^{+\infty} p dx = A \int_{-\infty}^{+\infty} e^{-B^2 x^2} dx = 1 \quad (66.311)$$

L'intégrale classique qui figure dans le deuxième terme a pour valeur (cf. chapitre de Statistiques):

$$A \int_{-\infty}^{+\infty} e^{-B^2 x^2} dx = A \frac{\sqrt{\pi}}{B} = 1 \quad (66.312)$$

nous devons donc obligatoirement avoir pour la normalisation:

$$B = A\sqrt{\pi} \quad (66.313)$$

Il en découle:

$$p = A e^{-\pi A^2 x^2} \quad (66.314)$$

En posant $x = 0$, nous obtenons $A = p_0$, c'est-à-dire que A égale la probabilité du cours coté actuellement. Il faut donc établir que la fonction:

$$p = p_0 e^{-x p_0^2 x^2} \quad (66.315)$$

où p_0 dépend du temps, satisfait bien à l'équation de condition ci-dessus.

Soient p_1, p_2 les quantités correspondant à p_0 et relatives aux temps t_1, t_2 , il faut donc prouver que l'expression:

$$\int_{-\infty}^{+\infty} p_1 e^{-x p_1^2 x^2} p_2 e^{-x p_2^2 (x-x)^2} dx \quad (66.316)$$

peut se mettre sous la forme $A e^{-B^2 x^2}$ où A, B ne dépendent que du temps! Cette intégrale devient en remarquant que z n'est pas une variable d'intégration (nous supposons qu'il est indépendant de x comme vous l'aurez compris depuis le début)

$$p_1 p_2 e^{-x p_2^2 z^2} \int_{-\infty}^{+\infty} e^{-x(p_1^2 + p_2^2)x^2 + 2x p_2^2 x z} dx \quad (66.317)$$

Nous allons maintenant changer la forme de l'intégrale (nous changeons aussi de notation pour l'exponentielle sinon cela devient illisible):

$$p_1 p_2 \exp\left(-\pi p_2^2 z^2 + \frac{\pi p_2^4 z^2}{p_1^2 + p_2^2}\right) \int_{-\infty}^{+\infty} \exp\left(-\pi\left(x\sqrt{p_1^2 + p_2^2} - \frac{p_2^2 z}{\sqrt{p_1^2 + p_2^2}}\right)^2\right) dx \quad (66.318)$$

et posons:

$$x\sqrt{p_1^2 + p_2^2} - \frac{p_2^2 z}{\sqrt{p_1^2 + p_2^2}} = u \quad (66.319)$$

Nous aurons alors:

$$\frac{p_1 p_2 \exp\left(-\pi p_2^2 z^2 + \frac{\pi p_2^4 z^2}{p_1^2 + p_2^2}\right)}{\sqrt{p_1^2 + p_2^2}} \int_{-\infty}^{+\infty} \exp(-\pi u^2) du \quad (66.320)$$

L'intégrale:

$$\int_{-\infty}^{+\infty} e^{-x u^2} du = 2 \int_0^{+\infty} e^{-x u^2} du \quad (66.321)$$

ayant pour valeur 1 (cf. chapitre de Statistiques), nous obtenons finalement:

$$\frac{p_1 p_2 \exp\left(-\pi p_2^2 z^2 + \frac{\pi p_2^4 z^2}{p_1^2 + p_2^2}\right)}{\sqrt{p_1^2 + p_2^2}} = \frac{p_1 p_2}{\sqrt{p_1^2 + p_2^2}} \exp\left(\frac{-\pi p_1^2 p_2^2}{p_1^2 + p_2^2} z^2\right) \quad (66.322)$$

Cette expression ayant la forme désirée puisque:

$$\int_{-\infty}^{+\infty} \frac{p_1 p_2}{\sqrt{p_1^2 + p_2^2}} \exp\left(-\frac{\pi p_1^2 p_2^2}{\sqrt{p_1^2 + p_2^2}} z^2\right) dz = 1 \quad (66.323)$$

Cette relation exprime le fait que la probabilité cumulée totale que la variable aléatoire z puisse prendre n'importe quelle valeur est égale à l'unité.

Nous devons en conclure que la probabilité que le titre soit coté z au temps $t_1 + t_2$ s'exprime bien par la relation:

$$p = p_0 e^{-\pi p_0^2 t^2} \quad (66.324)$$

Nous voyons que la probabilité est régie par une loi de distribution de type loi Normale centrée réduite (cf. chapitre de Statistiques)! Ceci constitue un résultat remarquable obtenu par Louis Bachelier en 1900 et qui avait été déjà spéculé par Jules Regnault au milieu du 19ème siècle.

Effectivement, Regnault compare la spéculation à un jeu de pile ou face dans lequel les deux côtés de la pièce correspondent aux deux possibilités, hausse ou baisse du cours. Sous l'hypothèse qu'à quelque moment que ce soit, il n'y a jamais plus d'avantages pour une chance que pour l'autre. Autrement dit, à chaque cotation, le cours a une chance sur deux d'augmenter et une chance sur deux de diminuer. Mais chaque spéculateur a son opinion sur la question. Sans cette diversité d'opinions, il n'y aurait par conséquent ni échanges ni variations des cours. Les opérateurs se répartissent donc en deux groupes (haussiers, baissiers) qui font des évaluations subjectives de la valeur future du cours qui comportent forcément une marge d'erreur. Cependant, pour Regnault, les erreurs des spéculateurs ne sont pas quelconques, elles obéissent à une distribution Normale. Effectivement, comme l'a démontré Laplace, si la probabilité d'erreur est petite et qu'elles sont nombreuses et indépendantes alors les résultats des erreurs suivent une loi Normale (cf. chapitre de Statistiques).

La relation antécédente nous montre que les paramètres $p_0 = f(t)$ satisfont à la relation fonctionnelle:

$$f^2(t_1 + t_2) = \frac{f^2(t_1) f^2(t_2)}{f^2(t_1) + f^2(t_2)} \quad (66.325)$$

Différentions par rapport à t_1 , puis par rapport à t_2 . Le premier membre ayant la même forme dans les deux cas, nous obtenons:

$$\begin{aligned} 2f'(t_1) &= d_{t_1} \left(\frac{f^2(t_1) f^2(t_2)}{f^2(t_1) + f^2(t_2)} \right) = \left(\frac{2f(t_1)}{f^2(t_1) + f^2(t_2)} - \frac{2f^3(t_1)}{(f^2(t_1) + f^2(t_2))^2} \right) f^2(t_2) \\ 2f'(t_2) &= d_{t_2} \left(\frac{f^2(t_1) f^2(t_2)}{f^2(t_1) + f^2(t_2)} \right) = \left(\frac{2f(t_2)}{f^2(t_1) + f^2(t_2)} - \frac{2f^3(t_2)}{(f^2(t_1) + f^2(t_2))^2} \right) f^2(t_1) \end{aligned} \quad (66.326)$$

donc après simplification:

$$\left. \begin{aligned} f'(t_1) &= \frac{f(t_1) f^4(t_2)}{(f^2(t_1) + f^2(t_2))^2} \\ f'(t_2) &= \frac{f(t_2) f^4(t_1)}{(f^2(t_1) + f^2(t_2))^2} \end{aligned} \right\} \Rightarrow \frac{f(t_1) f^4(t_2)}{f'(t_1)} = \frac{f(t_2) f^4(t_1)}{f'(t_2)} \quad (66.327)$$

Ce qui donne finalement:

$$\frac{f^3(t_2)}{f'(t_1)} = \frac{f^3(t_1)}{f'(t_2)} \quad (66.328)$$

Cette relation ayant lieu, quels que soient t_1, t_2 , la valeur commune des deux rapports est constante et nous avons donc:

$$f'(t) = c^{te} f^3(t) \quad (66.329)$$

Une fonction qui satisfait cette relation existe et est:

$$f(t) = p_0 = \frac{H}{\sqrt{t}} \quad (66.330)$$

H désignant une constante ou une fonction indépendante du temps.

Vérification:

$$f'(t) = -\frac{1}{2} \frac{H}{t^{3/2}}, f^3(t) = \frac{H^3}{t^{3/2}} \quad (66.331)$$

donc:

$$c^{te} = -\frac{1}{2H^2} \quad (66.332)$$

Nous avons donc pour expression finale de la fonction de densité de probabilité de la valeur du cours x :

$$p = \frac{H}{\sqrt{t}} e^{-\frac{\pi H^2 x^2}{t}} = N\left(0, \frac{\sqrt{t}}{H\sqrt{2\pi}}\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} = N(0, \sigma) \quad (66.333)$$

avec x (pour rappel) qui est supérieur ou égal à 0.

Le lecteur remarquera donc que pour une valeur de H et t fixées nous avons toujours ici la forme d'une loi Normale centrée (cf. chapitre de Statistiques)!! Les financiers disent alors que nous avons affaire à un "hasard sage", sous-entendu que les variations sont faibles et régulières.

7.4. ESPÉRANCE ET VARIANCE POSITIVE

Comme le cours ne peut pas être négatif, nous nous restreignons au calcul de l'espérance positive comme étant alors (cf. chapitre de Statistiques):

$$\begin{aligned}\mu &= \int_0^{+\infty} \frac{Hx}{\sqrt{t}} e^{-\frac{\pi H^2 x^2}{t}} dx = \frac{H}{\sqrt{t}} \int_0^{+\infty} x e^{-\frac{\pi H^2 x^2}{t}} dx = \frac{H}{\sqrt{t}} \left(-\frac{t}{2\pi H^2} e^{-\frac{\pi H^2 x^2}{t}} \right) \Bigg|_0^{+\infty} \\ &= \frac{H}{\sqrt{t}} \left(\left[0 - \left(-\frac{t}{2\pi H^2} \right) \right] \right) = \frac{t}{2\pi H^2} \frac{H}{\sqrt{t}} = \frac{\sqrt{t}}{2\pi H}\end{aligned}\quad (66.334)$$

en notant:

$$k = \frac{1}{2\pi H} \quad (66.335)$$

le "**coefficient d'instabilité**" (sur lequel nous ne savons rien) nous avons ainsi l'espérance positive du cours qui est au final:

$$\mu = \int_0^{+\infty} p x dx = k \sqrt{t} \quad (66.336)$$

l'espérance mathématique du cours est donc proportionnelle à la racine carrée du temps comme l'est le mouvement brownien que nous avons étudié dans le chapitre de Mécanique Statistique!!

Il découle aussi immédiatement de ce résultat que l'écart-moyen de la valeur du cours à deux instants différents consécutifs est lui aussi proportionnel à la racine carrée du temps écoulé entre les deux instants!

Nous remarquons aussi qu'à l'instant où t est égal à 0, l'espérance positive du gain est nulle, car la valeur y est connue de manière sûre (c'est ainsi qu'il faut l'interpréter).

Remarque: Le mouvement brownien est massivement employé par les professionnels, puisque les calculs de volatilité annualisée (en %/an) dont on trouve les résultats dans toute page financière de la presse quotidienne, ne sont que des conversions en racine carrée du temps des calculs de volatilité périodique (en %/mois ou %/semaine) utilisée comme base d'estimation.

Calculons maintenant la variance positive aussi:

$$Var(x) = \int_0^{+\infty} \frac{Hx^2}{\sqrt{t}} e^{-\frac{\pi H^2 x^2}{t}} dx = \frac{H}{\sqrt{t}} \int_0^{+\infty} x^2 e^{-\frac{\pi H^2 x^2}{t}} dx \quad (66.337)$$

Nous posons:

$$z = \sqrt{\frac{\pi}{t}} Hx \quad (66.338)$$

soit:

$$x = \frac{1}{H} \sqrt{\frac{t}{\pi}} z \Rightarrow x^2 = z^2 \frac{t}{H^2 \pi} \quad (66.339)$$

avec:

$$\frac{dz}{dx} = \sqrt{\frac{\pi}{t}} H \quad (66.340)$$

Il vient alors:

$$Var(x) = \int_0^{+\infty} \frac{Hx^2}{\sqrt{t}} e^{-\frac{\pi H^2 x^2}{t}} dx = \frac{H}{\sqrt{t}} \frac{1}{H} \sqrt{\frac{t}{\pi}} \frac{t}{H^2 \pi} \int_0^{+\infty} z^2 e^{-z^2} dz = \frac{t}{H^2 \pi \sqrt{\pi}} \int_0^{+\infty} z^2 e^{-z^2} dz \quad (66.341)$$

Or, dans le chapitre de Statistiques, nous avons démontré par intégration par parties (cf. chapitre de Calcul Intégral Et Différentiel) que:

$$\int_0^{+\infty} z^2 e^{-z^2} dz = \left. -\frac{ze^{-z^2}}{2} \right|_0^{+\infty} + \int_0^{+\infty} \frac{e^{-z^2}}{2} dz = \frac{1}{2} \int_0^{+\infty} e^{-z^2} dz = \frac{1}{2} \sqrt{\frac{\pi}{4}} = \frac{\sqrt{\pi}}{4} \quad (66.342)$$

Soit au final:

$$Var(x) = \frac{t}{2H^2 \pi \sqrt{\pi}} \frac{\sqrt{\pi}}{4} = \frac{t}{4H^2 \pi} \quad (66.343)$$

Donc si nous posons:

$$K = \frac{1}{2H\sqrt{\pi}} \quad (66.344)$$

Nous avons finalement:

$$\sigma = \sqrt{\int_0^{+\infty} px^2 dx} = K \sqrt{t} \quad (66.345)$$

Donc l'écart-type positif est lui aussi proportionnel à la racine carrée du temps (et le résultat serait le même si nous calculions l'écart-type total)!

Donc par stabilité de la loi Normale (cf. chapitre de Statistiques) nous avons:

$$\Delta p = N\left(\mu_2 - \mu_1, \sqrt{\sigma_2^2 - \sigma_1^2}\right) = N\left(0, \sqrt{\frac{t_2}{4\pi H^2} - \frac{t_1}{4\pi H^2}}\right) \quad (66.346)$$

Il s'ensuit immédiatement que:

$$\sigma_{\Delta p} = \frac{1}{2H\sqrt{\pi}} \sqrt{\Delta t} \quad (66.347)$$

et:

$$\mu_{\Delta p} = \frac{\sqrt{t_2 - t_1}}{2\pi H} = \frac{1}{2\pi H} \sqrt{\Delta t} \quad (66.348)$$

Donc les variations du prix du cours d'un actif financier entre deux instants successifs ont une loi de probabilité bien évidemment aussi décrite par une loi Normale centrée (découlant donc de la stabilité de cette loi) caractérisée elle aussi par une espérance positive et un écart-type positif proportionnels à la racine carrée du temps.

Ce résultat démontré mathématiquement avait été mesuré par Regnault une cinquantaine d'années auparavant (~1850) en observant que l'écart-moyen de titres obligataires français était proportionnel à la racine carrée du temps.

Ceci dit il faut accepter les limites de cette approche. Prenons par exemple les rendements journaliers de l'indice Dow Jones en 2008 et 2009. D'après les spécialistes possédant les détails de ces données, elles suivraient plutôt une loi de Student de paramètre 3 qu'une loi Normale...!

Pour donner une comparaison flagrante de la limite de ces approches rappelons (cf. chapitre de Statistiques) que la probabilité cumulée qu'une variable aléatoire suivant une loi Normale soit au-delà de 4 écart-types est de 1-99.99366% soit 0.00634%. Cela signifie, si la Bourse a 252 jours ouvrés, une certitude d'avoir une grande déviation tous les:

$$\frac{1}{\frac{0.00634\%}{252}} \cong 63 \text{ ans} \quad (66.349)$$

où nous considérons donc (cf. chapitre de Probabilités) les événements comme disjoints deux à deux.

Or la réalité montre, par exemple, que l'indice Dow Jones a eu entre 2008 et 2009 en moyenne 8 déviations au-delà de 4 écarts-types par année...et ce n'est guère qu'un peu mieux si nous faisons une approche avec la loi de Student.

Au final trois résultats majeurs sont à retenir ici sous les hypothèses fortes de normalité centrée et d'indépendance:

1. Que la fonction de distribution de probabilité que le cours d'un actif financier soit x à un instant t donné suit une loi Normale centrée....!!
2. Que l'espérance positive et l'écart-type positif de la valeur d'un actif financier sont proportionnels à la racine carrée du temps avec un facteur dont nous ne savons rien!!
3. Que l'espérance de gain est globalement nulle (ce qui rend le modèle peu réaliste mais donne déjà une base de travail).

C'est le premier modèle de base à connaître en finance (qui ne devrait plus être utilisé dans les entreprises en ce début de 21ème siècle mais qui l'est malheureusement encore en majorité...) et nous réutiliserons donc ces démarches lors de notre introduction au modèle de Black & Scholes.

Enfin, il convient de préciser que c'est un modèle théorique! Il faut donc le confronter à la pratique pour voir s'il est valide ou non. En l'occurrence l'observation des marchés financiers montre que ce n'est que hors des bulles spéculatives que les variations peuvent être modélisées par un mouvement brownien. Il faut donc chercher des modèles plus puissants et nous verrons un jour que le mouvement brownien (appelé également "processus brownien") qui est lisse (continu) et donc sans sauts brusques est un cas particulier des processus de Lévy.

7.5. MODÈLE DE DIVERSIFICATION EFFICIENTE DE MARKOWITZ

Les travaux de Markowitz en 1954 ont constitué la première tentative de théorisation de la gestion financière de portefeuilles et son modèle suggère une procédure de sélection de plusieurs titres boursiers, à partir de critères statistiques, afin d'obtenir des portefeuilles optimaux. Plus précisément, Markowitz a montré que l'investisseur cherche à optimiser ses choix en tenant compte non seulement de la rentabilité attendue de ses placements, mais aussi du risque de son portefeuille qu'il définit mathématiquement par la variance de sa rentabilité. Ainsi, le "**portefeuille efficient**" est le portefeuille le plus rentable pour un niveau de risque donné. Il est déterminé au mieux par application de méthodes de programmation quadratique (cf. chapitre de Méthodes Numériques) ou sinon de manière heuristique en les étapes suivantes:

1. Nous fixons une espérance de rentabilité et nous trouvons tous les portefeuilles de variance minimale satisfaisant l'objectif de rentabilité. Nous obtenons ainsi un ensemble de portefeuilles de variance minimale.
2. Nous gardons de ces portefeuilles celui qui pour une variance donne le rendement le plus élevé.

En procédant ainsi pour plus plusieurs valeurs de l'espérance, nous nous retrouvons avec un ou plusieurs portefeuilles efficients. Ainsi, entre deux portefeuilles (ensemble d'actifs) caractérisés par leur rendement (supposé aléatoire!), nous ferons les hypothèses suivantes:

- H1. À risque identique, nous retenons celui qui a l'espérance de rendement la plus élevée (gain maximal)
- H2. À espérance de rendement identique, nous retenons celui qui présente le risque le plus faible (aversion au risque)

Ce principe conduit à éliminer un certain nombre de portefeuilles, moins efficients que d'autres.

Passons maintenant à la théorie (un exemple pratique du modèle de Markowitz sera donné après les développements mathématiques).

Soit R_p le rendement d'un portefeuille composé de n actifs caractérisés par leur rendement respectif $R_1, R_2, \dots, R_n$. Nous posons, en outre, que chaque actif i entre pour une proportion X_i dans la composition du portefeuille P tel que:

$$\sum_{i=1}^n X_i = 1 \quad (66.350)$$

Remarque: Une part X_i d'un actif peut aussi être négative... Détenir une part négative d'un actif, c'est ce qui s'appelle en anglais le "**short-selling**" (vente à découvert). Cette technique consiste par exemple à emprunter beaucoup d'actifs (supposés surévalués sur le marché) à une banque, les vendre pour faire baisser le prix de l'actif, et faire un profit en les rachetant moins cher pour les rendre à la banque (grosso modo car c'est assez complexe au fait...).

Donc l'espérance du portefeuille est donnée par:

$$E(R_p) = E\left(\sum_{i=1}^n X_i R_i\right) = \sum_{i=1}^n X_i E(R_i) \quad (66.351) \quad (66.352)$$

où l'espérance de R_i est souvent prise comme étant simplement la moyenne arithmétique.

Maintenant, nous supposons que les return des différents actifs financiers ne fluctuent pas indépendamment les uns des autres: ils sont corrélés ou, ce qui revient au même, ont des covariances non nulles (cf. chapitre de Statistiques):

$$\text{cov}(R_i, R_j) \neq 0 \quad (66.353)$$

Dès lors, la variance du portefeuille est donnée par (cf. chapitre de Statistiques):

$$V(R_p) = \sum_{i=1}^n X_i^2 V(R_i) + 2 \sum_{i < j} X_i X_j \text{cov}(R_i, R_j) \quad (66.354)$$

Avant d'aller plus loin, précisons (car c'est important dans la pratique) que nous pouvons également écrire cette dernière relation sous forme matricielle (le lecteur peut facilement vérifier en ne prenant par exemple que deux titres que les deux écritures donnent un résultat identique) si nous notons X le vecteur des parts d'actifs et X^T le même vecteur transposé:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \quad X^T = (X_1 \quad X_2 \quad \cdots \quad X_n) \quad (66.355)$$

et finalement $c_{i,j}$ la matrice des covariances:

$$c_{i,j} = \begin{bmatrix} \text{cov}(R_1, R_1) & \text{cov}(R_1, R_2) & \cdots & \text{cov}(R_1, R_n) \\ \text{cov}(R_2, R_1) & \text{cov}(R_2, R_2) & & \vdots \\ \vdots & & \ddots & \\ \text{cov}(R_n, R_1) & \cdots & & \text{cov}(R_n, R_n) \end{bmatrix} \quad (66.356)$$

matrice qui se simplifie directement en:

$$c_{i,j} = \begin{bmatrix} \sigma_1^2 & \text{cov}(R_1, R_2) & \cdots & \text{cov}(R_1, R_n) \\ \text{cov}(R_2, R_1) & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & \\ \text{cov}(R_n, R_1) & \cdots & & \sigma_n^2 \end{bmatrix} \quad (66.357)$$

nous obtenons finalement la relation de la variance sous forme matricielle condensée:

$$V(R_p) = X^T \cdot (c_{ij} \cdot X) \quad (66.358)$$

telle que nous la voyons souvent dans la littérature spécialisée.

Pour en revenir à la forme algébrique du modèle, puisque la covariance est symétrique (cf. chapitre de Statistiques):

$$\text{cov}(R_i, R_j) = \text{cov}(R_j, R_i) \quad (66.359)$$

et que:

$$\text{cov}(R_j, R_j) = V(R_j) \quad (66.360)$$

Nous pouvons simplifier et écrire la variance:

$$V(R_p) = \sum_{i=1}^n X_i^2 V(R_i) + 2 \sum_{i < j} X_i X_j \text{cov}(R_i, R_j) \quad (66.361)$$

sous la forme algébrique suivante:

$$V(R_p) = \sum_{i=1}^n \sum_{j=1}^n X_i X_j \text{cov}(R_i, R_j) \quad (66.362)$$

telle que nous la voyons souvent dans la littérature spécialisée d'une certaine époque...

Sélectionner un portefeuille revient donc à résoudre le problème de maximisation sous contrainte suivant:

$$\begin{aligned} \min : & \text{Var}(R_p) \\ & E(R_p) = c^{te} \\ & \sum X_i = 1 \end{aligned}$$

en utilisant la programmation quadratique (cf. chapitre de Méthodes Numériques).

Dans la pratique, nous cherchons non pas un, mais tous les portefeuilles qui pour une espérance donnée minimise la variance. Nous obtenons alors une fonction de l'espérance en fonction de la variance pour les portefeuilles optimaux si nous traçons cela sur un graphique (voir plus bas). Cette fonction est souvent assimilée par les financiers (à juste titre!) à une frontière comme le précise la définition qui suit.

Définition: La frontière qui caractérise le polygone ou la courbe des contraintes s'appelle dans cette situation la "**frontière efficiente (de Markowitz)**" et dans le polygone/courbe se situent tous les portefeuilles à rejeter dits "**portefeuilles dominés**". Une autre manière de formuler ceci consiste à dire que les combinaisons (rendement, risque) de cette frontière forment un ensemble d'optima de Pareto (cf. chapitre de Théorie Des Jeux et de la Décision), c'est-à-dire que si l'un des éléments augmente, l'autre doit augmenter aussi.

Maintenant, formalisons l'optimisation comme cela était fait à l'époque où les gens devaient encore développer les algorithmes eux-mêmes...

Soit Z la fonction économique précitée:

$$Z = \Phi E(R_p) - V(R_p) \quad (66.363)$$

qui doit être maximisée sous la contrainte que $\sum_{i=1}^n X_i = 1$ et où Φ est un paramètre qui représente le degré d'aversion au risque des investisseurs (histoire aussi d'homogénéiser la relation...).

Le problème de maximisation sous contrainte consiste à déterminer le maximum de la fonction économique Z définie par:

$$Z = \Phi \sum_{i=1}^n X_i E(R_i) - \sum_{i=1}^n \sum_{j=1}^n X_i X_j \text{cov}(R_i, R_j) + \lambda \left(1 - \sum_{i=1}^n X_i \right) \quad (66.364)$$

Cette fonction de $n + 1$ variables $(X_1, X_2, \dots, X_n, \lambda)$ est maximisée si sa dérivée (partielle) par rapport à chacune de ces variables est nulle, ce qui revient à poser le système suivant:

$$\begin{cases} \frac{\partial Z}{\partial X_1} = \Phi E(R_1) - 2X_1 \text{cov}(R_1, R_1) - 2X_2 \text{cov}(R_1, R_2) - \dots - 2X_n \text{cov}(R_1, R_n) - \lambda = 0 \\ \frac{\partial Z}{\partial X_2} = \Phi E(R_2) - 2X_2 \text{cov}(R_2, R_1) - 2X_2 \text{cov}(R_2, R_2) - \dots - 2X_n \text{cov}(R_2, R_n) - \lambda = 0 \\ \dots \\ \frac{\partial Z}{\partial X_n} = \Phi E(R_n) - 2X_1 \text{cov}(R_n, R_1) - 2X_2 \text{cov}(R_n, R_2) - \dots - 2X_n \text{cov}(R_n, R_n) - \lambda = 0 \\ \frac{\partial Z}{\partial \lambda} = 1 - X_1 - X_2 - \dots - X_n = 0 \end{cases} \quad (66.365)$$

Posons:

$$\text{cov}(R_i, R_j) = \sigma_{ij} \quad (66.366)$$

Nous pouvons alors écrire:

$$\begin{cases} 2X_1\sigma_{11} + 2X_2\sigma_{12} + \dots + 2X_n\sigma_{1n} + \lambda = \Phi E(R_1) \\ 2X_2\sigma_{21} + 2X_2\sigma_{22} + \dots + 2X_n\sigma_{2n} + \lambda = \Phi E(R_2) \\ \dots \\ 2X_1\sigma_{n1} + 2X_2\sigma_{n2} + \dots + 2X_n\sigma_{nn} + \lambda = \Phi E(R_n) \\ X_1 + X_2 + \dots + X_n = 1 \end{cases} \quad (66.367)$$

soit sous forme matricielle:

$$\begin{pmatrix} 2\sigma_{11} & 2\sigma_{12} & \dots & 2\sigma_{1n} & 1 \\ 2\sigma_{21} & 2\sigma_{22} & \dots & 2\sigma_{2n} & 1 \\ \dots & \dots & \dots & \dots & \dots \\ 2\sigma_{n1} & 2\sigma_{n2} & \dots & 2\sigma_{nn} & 1 \\ 1 & 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \dots \\ X_n \\ \lambda \end{pmatrix} = \begin{pmatrix} \Phi E(R_1) \\ \Phi E(R_2) \\ \dots \\ \Phi E(R_n) \\ 1 \end{pmatrix} \quad (66.368)$$

Soit désormais:

$$A = \begin{pmatrix} 2\sigma_{11} & 2\sigma_{12} & \dots & 2\sigma_{1n} & 1 \\ 2\sigma_{21} & 2\sigma_{22} & \dots & 2\sigma_{2n} & 1 \\ \dots & \dots & \dots & \dots & \dots \\ 2\sigma_{n1} & 2\sigma_{n2} & \dots & 2\sigma_{nn} & 1 \\ 1 & 1 & \dots & 1 & 0 \end{pmatrix}, B = \begin{pmatrix} \Phi E(R_1) \\ \Phi E(R_2) \\ \dots \\ \Phi E(R_n) \\ 1 \end{pmatrix} \text{ et } X = \begin{pmatrix} X_1 \\ X_2 \\ \dots \\ X_n \\ \lambda \end{pmatrix} \quad (66.369)$$

Dans ce cas, le système d'équations à résoudre peut se résumer sous la forme matricielle:

$$A \cdot X = B \quad (66.370)$$

Par conséquent:

$$X = A^{-1} \cdot B \quad (66.371)$$

La détermination du poids de chacun des n actifs susceptibles d'entrer dans la composition d'un portefeuille passe donc par l'inversion d'une matrice carrée de $n + 1$ lignes et $n + 1$ colonnes comportant $(n^2 - n)/2$ covariances (la diagonale comportant des variances seulement et la matrice étant symétrique!). Ce qui est relativement long à calculer pour de gros portefeuilles.

Cependant, même une fois la pondération des actifs terminée, le problème lui ne l'est pas complètement. Effectivement, nous pouvons donc connaître la frontière efficiente mais le client va lui imposer une contrainte bien logique au niveau du risque nul de son portefeuille et du rapport rendement/risque maximum.

Compte tenu de la lourdeur des calculs nécessaires à l'inversion de la matrice A , Sharpe a proposé un modèle simplifié que nous verrons après un exemple pratique du modèle de Markowitz.

Exemple:

Considérons trois titres composant un portefeuille en proportions égales (que nous supposons dans des proportions égales dans le portefeuille) et les n observations de leur rendement $R_{i,j}$ saisis dans Microsoft Excel 11.8346 (la composante j pouvant être vue comme une période temporelle):

	A	B	C	D
1	Obs.	Titre 1	Titre 2	Titre 3
2	1	-0.15	0.29	0.38
3	2	0.05	0.18	0.63
4	3	-0.43	0.24	0.46
5	4	0.79	0.25	0.36
6	5	0.32	0.17	-0.57

Figure: 66.21 - 5 observations des rendements de 3 titres

Le but est donc de déterminer la frontière d'efficacité du portefeuille selon le modèle de Markowitz ainsi que la C.M.L. et la pondération des actifs (poids optimaux) qui minimise la variance pour une espérance maximum pour un portefeuille composé d'un actif sans risque d'un rendement R_f de 0.22.

Au-dessous de la table donnée précédemment nous allons créer dans Microsoft Excel le tableau contenant les proportions X_i des titres (que nous supposons équidistribuées, soit 1/3), nous afficherons la moyenne du rendement $\hat{\mu}_i$ calculée bien évidemment selon l'estimateur:

$$E(R_i) = \hat{\mu}_i = \frac{\sum_{j=1}^n R_{i,j}}{n} = \begin{cases} \text{Moyenne}(B2 : B6) \\ \text{Moyenne}(C2 : C6) \\ \text{Moyenne}(D2 : D6) \end{cases} \quad (66.372)$$

et la variance $\hat{\sigma}_i^2$ calculée pour chaque titre par l'estimateur:

$$\hat{\sigma}_i^2 = \frac{\sum_{j=1}^n (R_{i,j} - \hat{\mu}_i)^2}{n-1} = \begin{cases} \text{Var}(B2 : B6) \\ \text{Var}(C2 : C6) \\ \text{Var}(D2 : D6) \end{cases} \quad (66.373)$$

Ce qui nous donne le tableau suivant dans Microsoft Excel 11.8346:

	A	B	C	D
13		Titre 1	Titre 2	Titre 3
14	X_i	1/3	1/3	1/3
15	$\hat{\mu}_i$	0.116	0.226	0.252
16	$\hat{\sigma}_i^2$	0.21728	0.00253	0.22247

Figure: 66.22 - Proportions, estimateurs de l'espérance et de la variance des 3 titres

Soit sous forme détaillée dans la version française de Microsoft Excel 11.8346:

	A	B	C	D
13		Titre 1	Titre 2	Titre 3
14	X_i	1/3	1/3	1/3
15	$\hat{\mu}_i$	=MOYENNE(B2:B6)	=MOYENNE(C2:C6)	=MOYENNE(D2:D6)
16	$\hat{\sigma}_i^2$	=VAR(B2:B6)	=VAR(C2:C6)	=VAR(D2:D6)

Figure: 66.23 - Formules explicites pour le calcul des estimateurs de l'espérance et de la variance des 3 titres

Nous devons maintenant calculer le rendement moyen du portefeuille selon:

$$E(R_p) = \sum_{i=1}^n X_i R_i = X_1 \hat{\mu}_1 + X_2 \hat{\mu}_2 + X_3 \hat{\mu}_3 \quad (66.374)$$

Cette relation est un peu longue à saisir, et le sera davantage si nous avons un nombre bien plus important de titres.

Dans notre cas, il s'agit de faire la somme des produits terme à terme de deux plages de cellules (X_i et $\hat{\mu}_i$) ayant la même dimension (même nombre de lignes et même nombre de colonnes). Nous pouvons alors utiliser la fonction suivante dans la version française de Microsoft Excel 11.8346:

$$=\text{SOMMEPROD}(B14:D14;B15:D15)$$

Pour la variance du portefeuille, c'est un peu plus compliqué puisqu'il s'agira de calculer:

$$\begin{aligned} V(R_p) &= \sigma^2(R_p) = \sum_{i=1}^n X_i^2 V(R_i) + 2 \sum_{i < j} X_i X_j \text{cov}(R_j, R_i) \\ &= \sum_{i=1}^n X_i^2 \sigma_i^2 + 2 \sum_{i < j} X_i X_j \text{cov}(R_j, R_i) \end{aligned} \quad (66.375)$$

La relation développée dans notre cas particulier donne:

$$\begin{aligned} \sigma^2(R_p) &= X_1^2 \sigma_1^2 + X_2^2 \sigma_2^2 + X_3^2 \sigma_3^2 + 2X_1 X_2 \text{cov}(R_1, R_2) \\ &\quad + 2X_1 X_3 \text{cov}(R_1, R_3) + 2X_2 X_3 \text{cov}(R_2, R_3) \end{aligned} \quad (66.376)$$

L'astuce pour appliquer ceci dans un tableur Microsoft Excel consiste à utiliser l'algèbre linéaire et écrire cette relation sous forme matricielle comme nous l'avons démontré:

$$\sigma^2(R_p) = X^T \cdot c_{ij} \cdot X \quad (66.377)$$

Ce qui équivaut dans la version française de Microsoft Excel 11.8346 à écrire:

$$=\text{SOMMEPROD}(\text{PRODUITMAT}(B14:D14;G14:I16);B14:D14)$$

Soit sous forme matricielle explicite:

$$\sigma^2(R_p) = (X_1 \ X_2 \ X_3) \begin{bmatrix} \sigma_1^2 & \text{cov}(R_2, R_1) & \text{cov}(R_3, R_1) \\ \text{cov}(R_1, R_2) & \sigma_2^2 & \text{cov}(R_3, R_2) \\ \text{cov}(R_1, R_3) & \text{cov}(R_2, R_3) & \sigma_3^2 \end{bmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \quad (66.378)$$

En se basant sur les tableaux précédents, il est simple dans Microsoft Excel d'obtenir la matrice de covariance:

	F	G	H	I
12				
13	$c_{i,j}$	T1	T2	T3
14	T1	0.21728	-0.003376	-0.053492
15	T2	-0.003376	0.00253	0.008468
16	T3	-0.053492	0.008468	0.22247

Figure: 66.24 - Matrice de covariance des 3 titres

Soit sous forme détaillée dans la version française de Microsoft Excel 11.8346:

	F	G	H	I
12				
13	$c_{i,j}$	T1	T2	T3
14	T1	=VAR(B6:B10)	=G15	=G16
15	T2	=COVARIANCE(C6:C10;B6:B10)	=VAR(C6:C10)	=H16
16	T3	=COVARIANCE(D6:D10;B6:B10)	=COVARIANCE(D6:D10;C6:C10)	=VAR(D6:D10)

Figure: 66.25 - Formules explicites de la matrice des covariances

Rappel: La matrice des covariances est symétrique... (cf. chapitre de Statistiques).

Et pour l'espérance et la variance du portefeuille, nous aurons donc le tableau suivant:

	A	B
19		Portefeuille
20	$E(R_p)$	0.198
21		
22	$\sigma^2(R_p)$	0.038386667

Figure: 66.26 - Espérance et variance du portefeuille

en appliquant donc les relations susmentionnées:

	A	B
19		Portefeuille
20	$E(R_p)$	=SOMMEPROD(B14:D14;B15:D15)
21		
22	$\sigma^2(R_p)$	=SOMMEPROD(PRODUTMAT(B14:D14;G14:I16);B14:D14)

Figure: 66.27 - Formules explicites pour l'espérance et variance du portefeuille

Le problème maintenant est de déterminer pour un rendement du portefeuille fixé (B20), les proportions des différents titres qui minimisent le risque.

Après avoir ajouté les deux cellules B24 (rendement espéré/attendu du portefeuille) et B25 (nombre total des parts du portefeuille):

	A	B
24	Rendement attendu : $E(R_p)$	0.200
25	Sommes des poids : $\sum X_i$	1

Figure: 66.28 - Objectifs en termes de rendements et parts

Nous devons donc maintenant résoudre le problème d'optimisation non linéaire:

$$\begin{aligned} \min : & \sigma^2(R_p) \\ E(R_p) &= 0.2 \quad (66.379) \\ \sum X_i &= 1 \end{aligned}$$

et ceci ne peut que se faire (simplement) à l'aide du solveur:

	A	B	C	D	E	F	G	H	I
5	Obs	Titre 1	Titre 2	Titre 3					
6	1	-0.15	0.29	0.38					
7	2	0.05	0.18	0.63					
8	3	-0.43	0.24	0.46					
9	4	0.79	0.25	0.36					
10	5	0.32	0.17	-0.57					
11									
12									
13		Titre 1	Titre 2	Titre 3	$c_{4,j}$	T1	T2	T3	
14	X_i	1/3	1/3	1/3	T1	0.21728	-0.00338	-0.05349	
15	$\hat{\mu}_i$	0.116	0.226	0.252	T2	-0.00338	0.00253	0.00847	
16	$\hat{\sigma}_i^2$	0.21728	0.00253	0.22247	T3	-0.05349	0.00847	0.22247	
17									
18									
19		Portefeuille							
20	$E(R_p)$	0.198							
21	$\sigma^2(R_p)$	0.03836667							
22									
23									
24	Rendement attendu : $E(R_p)$	0.200							
25	Sommes des poids : $\sum X_i$	1							
26									
27									
28									
29									
30									

Set Target Cell:	\$B\$22	Solve
Equal To:	Max ☐ Min ☒ Value of: 0	Close
By Changing Cells:	\$B\$14:\$D\$14	Guess
Subject to the Constraints:	\$B\$20 = \$B\$24 \$B\$25 = 1	Add Change Delete Options Reset All Help

Figure: 66.29 - Recherche des solutions à l'aide du solveur

Ce que nous allons faire à l'aide du solveur est de chercher et reporter les solutions pour des rendements de 0.2 à 0.245 par pas de 0.05. À chaque résultat, nous noterons le numéro de l'itération, la variance du portefeuille $\sigma^2(R_p)$ et l'espérance de rendement $E(R_p)$ qui était exigée. Cela devrait donner (bon il faudrait automatiser dans l'idéal la procédure par du VBA):

Itération	Variance	Rendement
1	0.012396	0.200
2	0.008447	0.205
3	0.005515	0.210
4	0.003520	0.215
5	0.002460	0.220
6	0.002336	0.225
7	0.003149	0.230
8	0.004897	0.235
9	0.007582	0.240
10	0.011202	0.245

Figure: 66.30 - Solutions pour différentes valeurs du rendement

Ce qui donne la frontière efficiente de Markowitz suivante sous forme graphique, appelé "**plan de Markowitz**", dans Microsoft Excel 11.8346:

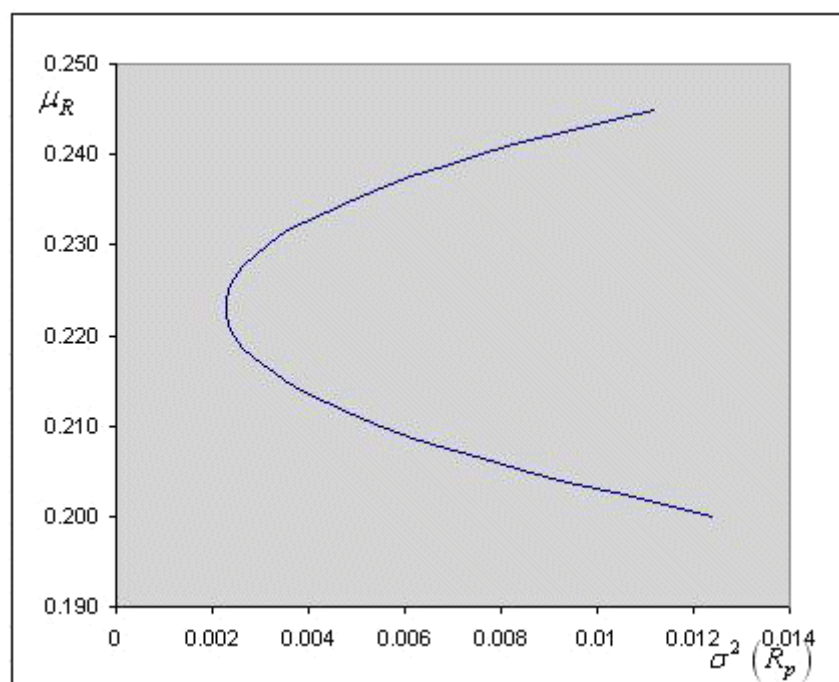


Figure: 66.31 - Plan de Markowitz correspondant aux solutions du tableau précédent

Maintenant il est aisé avec Microsoft Excel 11.8346 de déterminer l'équation de cette parabole en utilisant l'outil d'interpolation (nous sommes obligés dans Microsoft Excel 11.8346 de tourner la parabole pour cela...):

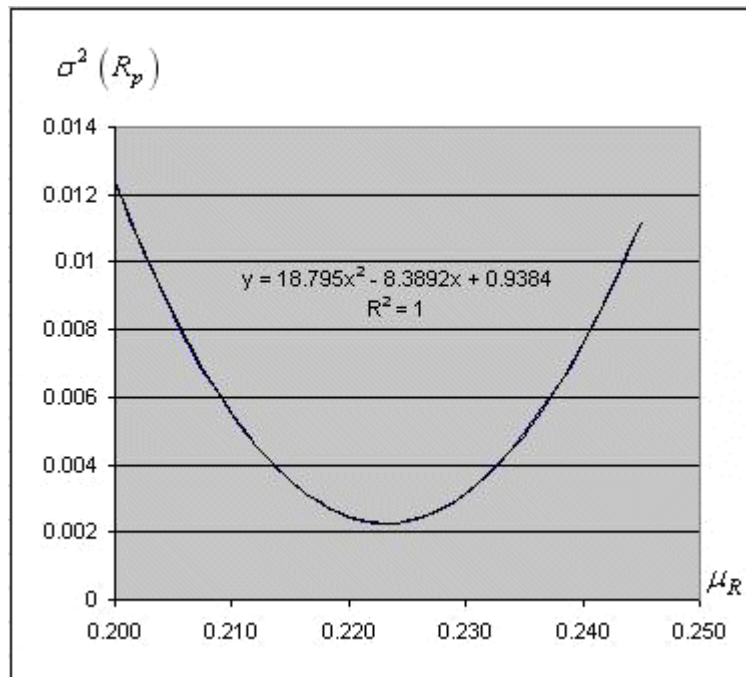


Figure: 66.32 - Rotation du plan de Markowitz pour déterminer l'équation de la parabole

Évidemment avec l'équation de cette parabole, en cherchant où sa dérivée s'annule, nous obtenons alors ce qu'il est d'usage d'appeler dans la finance le "**portefeuille global de variance minimum**" (en anglais: "Global Minimum Variance Portfolio"). Les poids des actifs de ce portefeuille sont donnés évidemment directement par les résultats affichés par le solveur.

Maintenant, nous allons déterminer la "**Capital Market Line**" abrégée C.M.L (voir le modèle du modèle des actifs financiers plus bas pour les détails) qui est la droite formée par l'ensemble des portefeuilles composés de l'actif sans risque, d'une part, et du portefeuille de marché, d'autre part. Par construction, elle associe à chaque niveau de risque, la rentabilité espérée la plus élevée.

Nous allons pour déterminer cette droite avec Microsoft Excel 11.8346 nous fixer dans un premier temps un taux de rendement sans risque que nous noterons R_f et que nous prendrons arbitrairement comme valant 0.22 (en Suisse en 2011, le rendement sans risque était estimé à 1% par exemple).

Nous avons donc la courbe de Markowitz d'équation:

$$y = 18.795x^2 - 8.3892x + 0.9384 = ax^2 + bx + c$$

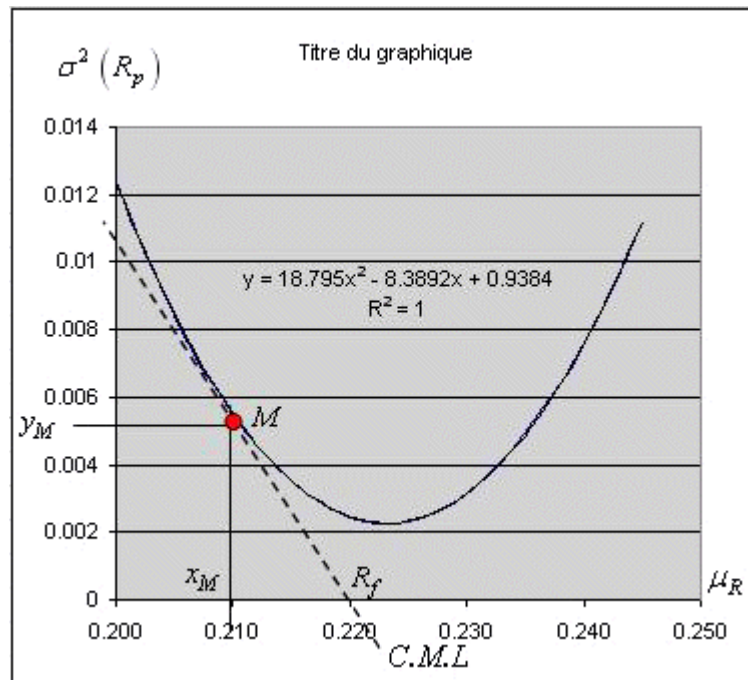


Figure: 66.33 - Mise en évidence de la Capital Market Line

et la droite du portefeuille sans actif risqué:

$$y = a'x + b' \quad (66.380)$$

avec la condition (voir sur le graphe):

$$a' = -\frac{b'}{0.22} \quad (66.381)$$

Nous avons alors deux équations à deux inconnues pour résoudre ce problème (l'intersection de la droite et de la parabole pour la première et l'égalité de la pente de la parabole et de celle de la droite au point d'intersection pour la deuxième):

$$\begin{aligned} ax_M^2 + bx_M + c &= a'x_M + b' \\ 2ax_M + b &= a' \end{aligned} \quad (66.382)$$

La deuxième équation nous donne:

$$x_M = \frac{a' - b}{2a} = \frac{\frac{-b'}{0.22} - b}{2a} \quad (66.383)$$

Injecté dans la première équation:

$$a \left(\frac{\frac{-b'}{0.22} - b}{2a} \right)^2 + b \left(\frac{\frac{-b'}{0.22} - b}{2a} \right) + c = -\frac{b'}{0.22} \left(\frac{\frac{-b'}{0.22} - b}{2a} \right) + b' \quad (66.384)$$

Si nous résolvons ce polynôme du deuxième degré nous avons deux solutions réelles (Microsoft Excel n'arrive pas à déterminer les racines de ce polynôme mais pour Maple 4.00b c'est très simple):

$$b'_1 = -0.06822748631 \quad b'_2 = 0.120763489 \quad (66.385)$$

La solution 2 est à éliminer (nous le savons en essayant de la prendre comme solution). Nous avons donc:

$$b' = -0.06822748631 \quad a' = -\frac{b'}{0.22} = 0.3101249378 \quad (66.386)$$

Ce qui donne sous forme graphique:

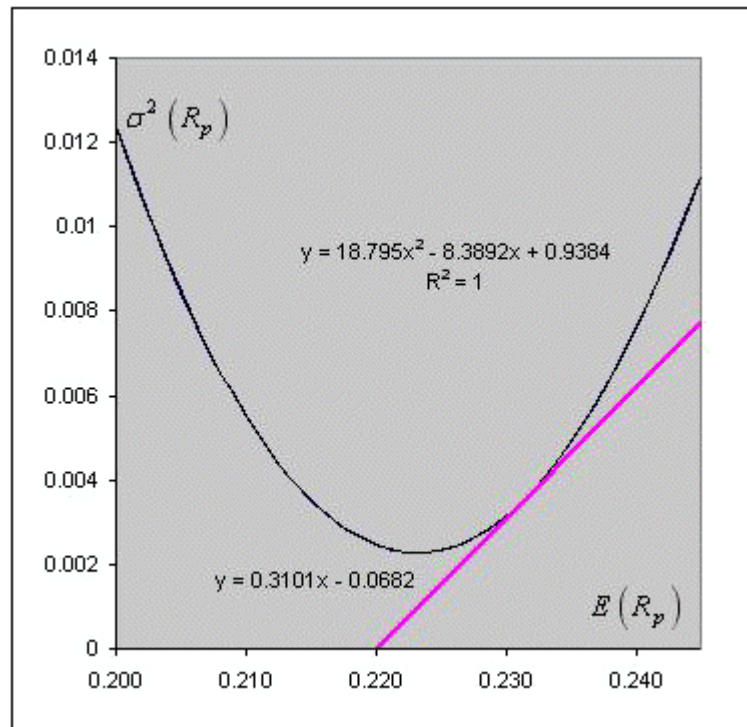


Figure: 66.34 - Représentation de la "vraie" Capital Market Line

Soit sous forme traditionnelle:

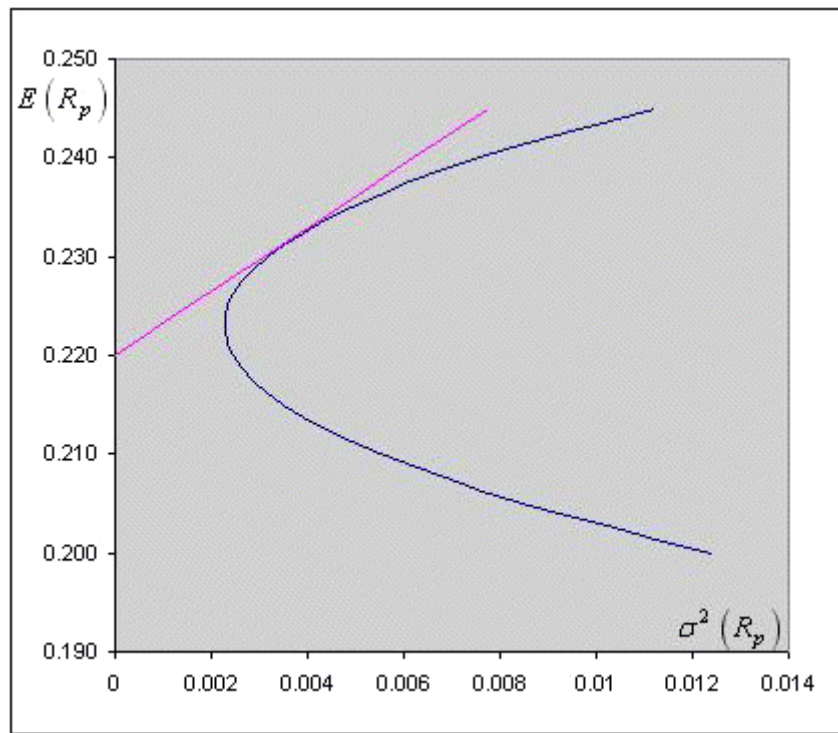


Figure: 66.35 - Petite rotation de la figure précédente

Il vient aussi immédiatement:

$$x_M = \frac{\frac{-b'}{0.22} - b}{2a} = 0.2314265746 \quad (66.387)$$

Ainsi, en réutilisant le solveur comme plus haut mais avec cette nouvelle valeur pour l'espérance, nous obtenons pour un portefeuille du marché composé d'un actif sans risque de rendement 0.22, un rendement global efficient de 0.2314276... avec la composition suivante du portefeuille donnée par le solveur:

$$\begin{aligned} X_1 &= -0.055 \\ X_2 &= 1.079 \\ X_3 &= -0.024 \end{aligned} \quad (66.388)$$

Il s'agit donc par construction du portefeuille qui, parmi l'ensemble des portefeuilles comportant uniquement des actifs risqués, maximise le rendement alors qu'il comporte un actif sans risque tout en minimisant le risque!!! Encore une fois, rappelons que ce portefeuille est appelé "portefeuille de marché" ou "portefeuille tangent".

Voilà donc un sympathique petit exemple applicatif dans un logiciel accessible à presque tout le monde!

Cependant rappelons deux hypothèses de construction de ce modèle qui font que son utilisation est limitée....:

H1. L'optimisation moyenne-variance revient à modéliser les actifs du portefeuille par des variables aléatoires pour lesquelles ces deux moments existent. Or dans la réalité, rien ne nous dit que c'est le cas.

H2. Le modèle est basé sur des rendements moyens estimés pour une période d'allocation pouvant évoluer à court-terme. Or un changement très faible de la moyenne des rendements peut avoir des conséquences démesurées sur l'allocation.

7.6. MODÈLE DE DIVERSIFICATION EFFICIENTE DE SHARPE

L'utilisation du modèle de Markowitz, tel qu'il le proposait dans son ouvrage de 1959, soulevait de nombreux problèmes dès qu'il s'agissait d'utiliser des algorithmes à partir d'une liste de base comportant un nombre élevé de valeurs. Ces problèmes étaient de deux ordres:

1. L'ampleur des matrices requérait à l'époque un calculateur de grande capacité et un temps de calcul assez long!
2. L'utilisation du modèle de base requérait que l'on connaisse dans son entièreté la matrice des covariances. Le principal problème qui se pose à ce propos réside tant dans le nombre des estimations à fournir que dans la difficulté de réaliser des estimations précises et surtout cohérentes.

Si nous voulons que l'approche proposée par Markowitz puisse entrer dans le domaine de l'application, il faut de toute évidence trouver le moyen d'alléger notablement la procédure tout en perdant le moins possible de la rigueur de la méthode.

En 1963, William Sharpe a proposé une solution dont la caractéristique essentielle consiste à faire l'hypothèse que les returns des diverses valeurs sont exclusivement liés entre eux par leur commune relation avec un facteur de base sous-jacent (indice boursier typiquement) qui permet de déterminer un coefficient appelé le "bêta" (corrélation entre le rendement d'un titre et celui du portefeuille de marché).

Cette hypothèse purement empirique appelée "**modèle à un indice**" (ou "**modèle unifactoriel**", "**modèle monofactoriel**") a revêtu par la suite une importance considérable, car elle a été, comme on le verra dans les développements ultérieurs, à la base de la théorie de la formation des prix des actifs financiers dans un univers incertain.

Remarque: Encore une fois, les développements qui vont suivre pourraient s'avérer abstraits mais... nous verrons comment résoudre l'exemple précédent fait avec Microsoft Excel pour le modèle de Markowitz mais en appliquant le modèle de Sharpe et nous pourrons ainsi même comparer visuellement les deux méthodes.

Le terme "unifactoriel" vient donc du fait qu'à la base le but du modèle de Sharpe est de définir le rendement d'un placement financier en fonction de son risque non diversifiable, assimilé au seul risque de marché (ou "risque systématique") donné par un nombre appelé "**coefficient bêta**".

Les investisseurs et gestionnaires distinguent trois sortes de risques:

1. Le "**risque spécifique**" relatif (implicite) au titre lui-même (sa variance) appelé aussi "**risque non systématique**" ou "**risque idiosyncratique**".
2. Le "**risque systématique/non diversifiable**" relatif à l'économie/marché au sens le plus large (variance du portefeuille de référence du marché).

3. Le "**risque global**" qui est en quelque sorte la somme des deux (c'est un peu plus subtil qu'une simple somme...).

Comme vous l'aurez probablement deviné, le facteur risque est difficilement quantifiable. L'élément qui aidera à le déterminer est la variation du rendement de l'actif financier par rapport à la variation du rendement du marché dans sa globalité. Un actif financier dont le cours fluctue souvent et dont la volatilité est grande présente donc certainement un risque élevé.

Voici par exemple les distributions des rendements de fonds de placements du deuxième pilier (LPP) en Suisse en fonction de la stratégie de diversification d'un portefeuille (les chiffres après LPP représentent la part en % d'actions):

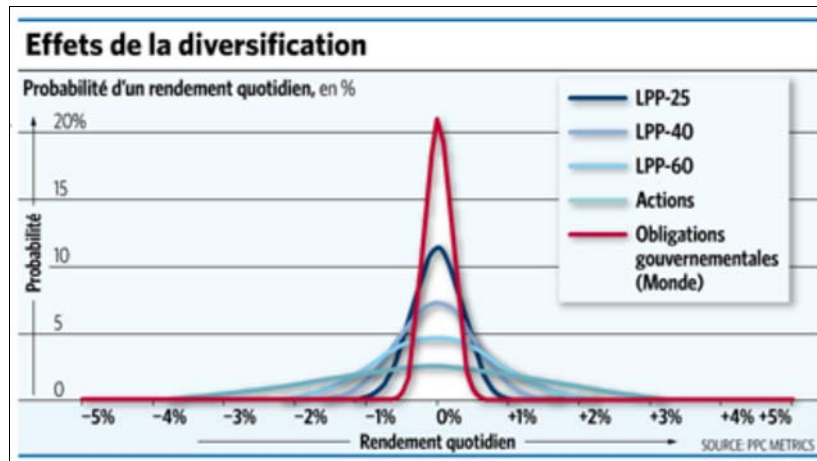


Figure: 66.36 - Effets de la diversification sur la distribution du rendement (source: PPC Metrics)

Vous remarquerez sur la figure précédente que les distributions sont centrées (donc de moyenne nulle), symétriques et que l'écart-type (volatilité) est d'autant plus grand que la part d'actions dans le portefeuille est élevée.

Définition (simpliste): Le "**coefficient bêta**" mesure la dépendance entre le rendement d'un portefeuille ou d'un actif financier et le rendement d'un indice de référence et constitue la pente d'une droite appelée "**security characteristic line**" (S.C.L.):

$$\beta = \frac{\text{variation attendue de la valeur du portefeuille}}{\text{variation du niveau de l'indice}} \quad (66.389)$$

ce coefficient est bien évidemment d'autant plus utile que l'horizon de prévision futur est éloigné et que la fréquence d'observation est petite. Ce coefficient est aussi parfois appelé "**volatilité relative**".

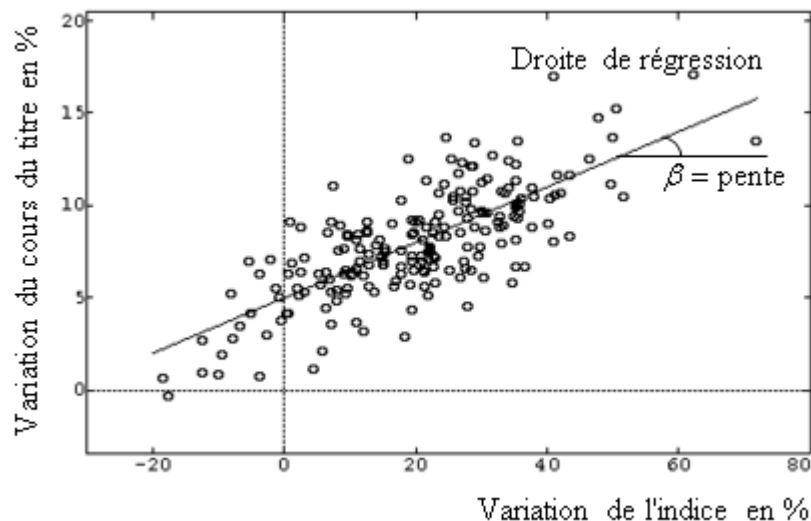


Figure: 66.37 - Principe de régression linéaire du coefficient bêta

Remarque: L'indice de référence est choisi de la manière la plus pertinente possible avec ce que cela implique... Si possible lorsque le rendement de l'indice est nul, la variation de la valeur du portefeuille ou de l'actif devrait aussi être nulle.

Une simple analyse du graphique (c'est de l'analyse fonctionnelle élémentaire) montre donc qu'un coefficient bêta égal à 1 pour un titre/actif donné signifie qu'une augmentation (respectivement: diminution) de 10 % du return des titres sur le marché pendant une certaine période se traduira par une augmentation (respectivement: diminution) de 10 % en moyenne du rendement de ce titre. Donc la volatilité de l'actif est égale à celle de l'indice.

Un bêta supérieur à 1 signifie que l'évolution du return de l'actif financier est plus volatile (ou plutôt, était volatile, puisque ce coefficient se réfère généralement à une période passée) que celle du return du marché, tandis qu'un bêta inférieur à 1 révèle l'inverse. Ainsi, un fonds ayant un bêta de 1.15 est de 15% plus volatile que l'indice. Inversement, un fonds ayant un bêta de 0.70 est 30% moins volatile que l'indice.

Donc pour résumer:

1. Un investissement ne présentant aucun risque (par rapport à l'indice de référence) afficherait donc un bêta égal à 1.
2. Un bêta inférieur à 1 indique que si le marché (indice de référence) est à la baisse, le titre sera susceptible de baisser moins que le marché.
3. Un bêta supérieur à 1 indiquera que si le marché (indice de référence) est à la hausse, le titre sera susceptible de suivre moins rapidement la tendance à la hausse.

Le concept de bêta ayant été introduit, passons maintenant à la théorie du modèle qui a donc pour objectif de simplifier celui de Markowitz en utilisant ce fameux coefficient.

Par définition, le bêta global d'un portefeuille est déterminé à partir des bêtas pondérés respectifs de chacun des titres ou bêtas sous-jacents qui le composent tel que:

$$\beta_P = \sum_{i=1}^N X_i \beta_i \quad (66.390)$$

avec β_P étant le bêta du portefeuille global, X_i la proportion du titre i dans le portefeuille P , β_i le bêta du titre i et n le nombre d'actifs financiers présents dans le portefeuille.

Sharpe suggère donc que le rendement R_i de chaque actif i à un instant t est donné par la régression linéaire (cf. chapitre de Méthodes Numériques) déterminant la "security characteristic line" vue plus haut:

$$R_i = \alpha_i + \beta_i I + \varepsilon_i \quad (66.391)$$

où:

- I est le rendement d'un indice économique donné (indice boursier, indice du produit national brut, indice des prix ou voire même le rendement du portefeuille du marché lui-même...) au temps t et est la variable expliquée de la régression (selon la terminologie utilisée dans le chapitre de Méthodes Numériques) considérée comme une variable aléatoire.

- α_i, β_i sont des estimateurs non biaisés (cf. chapitre de Statistiques) des paramètres propres à cette valeur. Le premier terme appelé en finance "**coefficient alpha**" est simplement l'ordonnée à l'origine de la régression (le rendement de l'actif lorsque le rendement de l'indice de référence est nul soit lorsque le marché a un rendement nul) et le deuxième paramètre est pour rappel simplement le bêta du titre risqué i . Le coefficient alpha n'étant pas une source de risque, il est souvent ignoré dans les calculs.

- ε_i est une variable aléatoire supposée caractérisée par une espérance nulle, une variance égale à une constante et les différents ε_i sont supposés non corrélés entre eux (covariance nulle).

Quant au niveau de l'indice I , il sera caractérisé par la relation (afin de simplifier les développements plus tard):

$$I = \alpha_{n+1} + v_{n+1} \quad (66.392)$$

où α_{n+1} est un paramètre non biaisé supplémentaire pour caractériser l'indice I et v_{n+1} une variable aléatoire caractérisée par une espérance nulle et une variance égale à une constante.

Pour résumer les points principaux, le modèle de régression linéaire simple des rendements des actifs financiers est basé sur les hypothèses majeures suivantes:

H1. Le modèle de rendement s'écrit de manière générale:

$$R_i = \alpha_i + \beta_i I + \varepsilon_i \quad (66.393)$$

en supposant que nous n'avons pas fait d'erreur sur la forme linéaire du modèle, ni sur la liste des régresseurs.

H2. Nous supposons que la perturbation de la régression est d'espérance nulle telle que (hypothèse sous-jacente d'un effet brownien!):

$$E(\varepsilon_i) = 0 \quad (66.394)$$

ce qui constitue ceci dit une hypothèse simplificatrice dangereuse mais pratique pour être utilisable (et compréhensible) par les praticiens de la finance...

H3. Pour n'importe quel échantillon de taille n , nous utilisons les estimateurs du maximum de vraisemblance (cf. chapitre de Statistiques) pour l'espérance et variance des rendements des actifs financiers du portefeuille de référence:

$$\hat{\sigma}_j = \frac{1}{n-1} \sum_{i=1}^n (R_{ji} - \hat{\mu}_j) \quad \hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n R_{ji} \quad (66.395)$$

Ces hypothèses posées, nous utilisons aussi les résultats obtenus dans le chapitre de Méthodes Numériques sur la régression linéaire pour obtenir le bêta. Nous y avons démontré qu'il existait plusieurs manières de faire une régression linéaire dont une consiste à utiliser la covariance et l'espérance. En adoptant les notations de l'économétrie, la pente de la régression peut alors s'écrire:

$$\beta_i = \frac{\text{cov}(R_i, R_I)}{\text{Var}(R_I)} = \frac{\text{cov}(R_i, R_I)}{\sigma_I^2} \quad (66.396)$$

ce qui donne la définition rigoureuse du coefficient bêta selon le modèle de Sharpe où R_i est le rendement de l'actif financier et R_I le rendement du marché (ou du portefeuille du marché/référence).

Définition (rigoureuse): Le "**coefficient bêta**" est donné par le rapport de la covariance des rendements et indices des actifs sur l'écart-type de l'indice du marché du portefeuille.

Maintenant, en considérant la même hypothèse que dans le modèle de Markowitz, le rendement R_p d'un portefeuille est défini à nouveau assez logiquement par:

$$R_p = \sum_{i=1}^n X_i R_i \quad (66.397)$$

où pour rappel X_i la proportion du titre i dans le portefeuille P .

Si les rendements ne sont pas explicitement connus dans la pratique, nous utilisons alors le modèle linéaire:

$$R_p = \sum_{i=1}^n X_i R_i = \sum_{i=1}^n X_i (\alpha_i + \beta_i I + \varepsilon_i) = \sum_{i=1}^n X_i (\alpha_i + \varepsilon_i) + \sum_{i=1}^n X_i \beta_i I \quad (66.398)$$

Dès lors en utilisant les propriétés de l'espérance:

$$E(R_p) = E \left[\sum_{i=1}^n X_i (\alpha_i + \varepsilon_i) + \sum_{i=1}^n X_i \beta_i I \right] = \sum_{i=1}^n X_i E(\alpha_i) + \sum_{i=1}^n X_i E(\varepsilon_i) + \sum_{i=1}^n X_i \beta_i E(\alpha_{n+1} + \nu_{n+1}) \quad (66.399)$$

Posons pour simplifier l'écriture que:

$$X_{n+1} = \sum_{i=1}^n X_i \beta_i \quad (66.400)$$

Dans ce cas, comme par hypothèse $E(\varepsilon_i) = 0$ et $E(\nu_{n+1}) = 0$:

$$E(R_p) = \sum_{i=1}^n X_i \alpha_i + 0 + X_{n+1} E(\alpha_{n+1} + \nu_{n+1}) = \sum_{i=1}^n X_i \alpha_i + 0 + X_{n+1} \alpha_{n+1} = \sum_{i=1}^{n+1} X_i \alpha_i \quad (66.401)$$

Finalement:

$$E(R_p) = \sum_{i=1}^{n+1} X_i \alpha_i \quad (66.402)$$

Si les rendements sont explicitement donnés et donc connus l'espérance se calculera avec:

$$E(R_p) = E\left(\sum_{i=1}^n X_i R_i\right) = \sum_{i=1}^n X_i E(R_i) \quad (66.403)$$

Cette relation est souvent appelée "**weighted average return**" ou "**rendement moyen pondéré**".

Comme le client va souvent chercher à maximiser l'espérance tout en minimisant la variance (le risque) il nous reste à déterminer cette dernière. Étant donné que maintenant nous supposons explicitement connus les rendements des actifs financiers du portefeuille et les rendements du portefeuille (indice) du marché nous avons:

$$\begin{aligned} V(R_p) &= V\left[\sum_{i=1}^n X_i R_i\right] = \sum_{i=1}^n X_i^2 V(R_i) + 2 \sum_{i < j} X_i X_j \text{cov}(R_i, R_j) \\ &= \sum_{i=1}^n X_i^2 V(R_i) + 2 \sum_{i < j} X_i X_j \text{cov}(\alpha_i + \beta_i I, \alpha_j + \beta_j I) \end{aligned}$$

Hypothèse: Si l'indice I est correctement choisi, lorsque $I = 0$ nous devons avoir $R_i = 0$ ce qui implique $\alpha_i = 0$ (c'est une hypothèse forte qui amène à avoir une approximation!).

Ainsi:

$$\begin{aligned} V(R_p) &= \sum_{i=1}^n X_i^2 V(R_i) + 2 \sum_{i < j} X_i X_j \text{cov}(\beta_i I, \beta_j I) \\ &= \sum_{i=1}^n X_i^2 V(R_i) + 2 \sum_{i < j} X_i X_j \beta_i \beta_j \text{cov}(I, I) \\ &= \sum_{i=1}^n X_i^2 V(R_i) + 2 \sum_{i < j} X_i X_j \beta_i \beta_j V(I) \end{aligned} \quad (66.404)$$

Finalement:

$$V(R_p) = \sum_{i=1}^n X_i^2 \sigma_i^2 + 2 \sigma_I^2 \sum_{i < j} X_i X_j \beta_i \beta_j \quad (66.405)$$

Ce qui donnerait donc pour un portefeuille comportant deux titres:

$$\begin{aligned}
V(R_p) &= \sum_{i=1}^2 X_i^2 \sigma_i^2 + 2\sigma_I^2 \sum_{i < j} X_i X_j \beta_i \beta_j \\
&= X_1^2 \sigma_1^2 + X_2^2 \sigma_2^2 + 2\sigma_I^2 (X_1 X_2 \beta_1 \beta_2) \\
&= X_1^2 \sigma_1^2 + X_2^2 \sigma_2^2 + 2\sigma_I^2 X_1 X_2 \beta_1 \beta_2
\end{aligned} \tag{66.406}$$

Nous pouvons condenser l'écriture de la variance en utilisant la formulation matricielle en notant d'abord respectivement le vecteur transposé et le vecteur-colonne des poids des actifs du portefeuille par:

$$X^T = [X_1 \quad X_2 \quad \dots \quad X_n] \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \tag{66.407}$$

et en en définissant la matrice des bêtas:

$$B_{i,j} = \begin{bmatrix} \frac{\sigma_1^2}{\sigma_I^2} & \beta_1 \beta_2 & \dots & \beta_1 \beta_n \\ \beta_2 \beta_1 & \frac{\sigma_2^2}{\sigma_I^2} & & \vdots \\ \vdots & & \ddots & \\ \beta_n \beta_1 & \dots & & \frac{\sigma_n^2}{\sigma_I^2} \end{bmatrix} \tag{66.408}$$

Ce qui nous donne finalement:

$$V(R_p) = \sigma_I^2 \cdot X^T \cdot (B_{i,j} \cdot X) \tag{66.409}$$

Pour un portefeuille de deux titres cette dernière relation se réduit donc à:

$$\begin{aligned}
V(R_p) &= \sigma_I^2 \cdot [X_1 \quad X_2] \cdot \left(\begin{bmatrix} \frac{\sigma_1^2}{\sigma_I^2} & \beta_1 \beta_2 \\ \beta_2 \beta_1 & \frac{\sigma_2^2}{\sigma_I^2} \end{bmatrix} \cdot \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \right) \\
&= \sigma_I^2 \cdot [X_1 \quad X_2] \begin{bmatrix} \frac{\sigma_1^2}{\sigma_I^2} X_1 + \beta_1 \beta_2 X_2 \\ \beta_2 \beta_1 X_1 + \frac{\sigma_2^2}{\sigma_I^2} X_2 \end{bmatrix} \\
&= \sigma_I^2 \left(\frac{\sigma_1^2}{\sigma_I^2} X_1^2 + \beta_1 \beta_2 X_2 X_1 + \beta_2 \beta_1 X_1 X_2 + \frac{\sigma_2^2}{\sigma_I^2} X_2^2 \right) \\
&= \sigma_1^2 X_1^2 + \sigma_2^2 X_2^2 + 2\sigma_I^2 X_1 X_2 \beta_1 \beta_2
\end{aligned} \tag{66.410}$$

Nous retrouvons donc bien la même chose que la forme algébrique.

Si nous ne connaissons pas explicitement les rendements, l'étude de la variance est un peu plus délicate. Il faut alors utiliser le modèle linéaire tel que:

$$V(R_p) = V\left[\sum_{i=1}^n X_i(\alpha_i + \varepsilon_i) + \sum_{i=1}^n X_i \beta_i I\right] = \sum_{i=1}^n X_i^2 V(\alpha_i) + \sum_{i=1}^n X_i^2 V(\varepsilon_i) + \sum_{i=1}^n X_i^2 \beta_i^2 V(\alpha_{n+1} + v_{n+1}) \quad (66.411)$$

Et toujours sous l'hypothèse déjà précisée plus haut comme quoi l'indice I est correctement choisi ce qui implique $\alpha_i = 0$.

En outre, notons:

$$Q_i = V(\varepsilon_i) \quad (66.412)$$

De plus, nous savons que:

$$Q_{n+1} = V(v_{n+1}) \quad (66.413)$$

Dès lors en négligeant la variance de l'ordonnée à l'origine:

$$V(R_p) = \sum_{i=1}^n X_i^2 Q_i + X_{n+1}^2 Q_{n+1} \quad (66.414)$$

car:

$$V(\alpha_{n+1}) = V(c^{te}) = 0 \quad (66.415)$$

Finalement:

$$V(R_p) = \sum_{i=1}^{n+1} X_i^2 Q_i \quad (66.416)$$

Dans ce contexte le problème revient toujours à maximiser la fonction économique Z :

$$Z = \Phi E(R_p) - V(R_p) + \lambda \left(1 - \sum_{i=1}^n X_i\right) \quad (66.417)$$

simplement que maintenant elle s'écrit:

$$Z = \Phi \sum_{i=1}^{n+1} X_i \alpha_i - \sum_{i=1}^{n+1} X_i^2 Q_i + \lambda \left(1 - \sum_{i=1}^n X_i\right) \quad (66.418)$$

Le calcul de chacune des dérivées partielles donne alors:

$$\begin{cases} \frac{\partial Z}{\partial X_1} = \Phi \alpha_1 - 2X_1 Q_1 - \lambda = 0 \\ \frac{\partial Z}{\partial X_2} = \Phi \alpha_2 - 2X_2 Q_2 - \lambda = 0 \\ \dots \\ \frac{\partial Z}{\partial X_n} = \Phi \alpha_n - 2X_n Q_n - \lambda = 0 \\ \frac{\partial Z}{\partial X_{n+1}} = \Phi \alpha_{n+1} - 2X_{n+1} Q_{n+1} - \lambda = 0 \\ \frac{\partial Z}{\partial \lambda} = 1 - X_1 - X_2 - \dots - X_n = 0 \end{cases} \quad (66.419)$$

soit sous forme matricielle:

$$\begin{pmatrix} 2Q_1 & 0 & \dots & 0 & 1 \\ 0 & 2Q_2 & \dots & 0 & 1 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 2Q_{n+1} & 1 \\ 1 & 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \dots \\ X_{n+1} \\ \lambda \end{pmatrix} = \begin{pmatrix} \Phi \alpha_1 \\ \Phi \alpha_2 \\ \dots \\ \Phi \alpha_{n+1} \\ 1 \end{pmatrix} \quad (66.420)$$

La résolution de ce système passe alors par l'inversion d'une matrice plus simple que celle du modèle de Markowitz mais nécessite cependant des hypothèses relativement contraignantes.

Pour finir, signalons que les financiers utilisent souvent les indicateurs de rendement pondéré par le risque, le plus répandu au niveau international étant le "**ratio de Sharpe**". Il est déterminé par le rapport entre le (pour être plus exact il s'agit de son espérance) différentiel du rendement d'un placement (actif) sans risque R_f et le rendement du marché (appelé le "**benchmark**") R_p et la déviation standard du placement au taux du marché (nous déterminerons rigoureusement l'origine de cette relation plus loin lors de notre étude du MEDAF):

$$S_p = \frac{R_p - R_f}{\sigma_p} \quad (66.421)$$

Relation qui exprime donc le niveau de rendement pur par unité de volatilité (ou par unité de risque). Pour simplifier, c'est un indicateur de la rentabilité (marginale) obtenue par unité de risque pris dans cette gestion. Il permet de répondre à la question suivante: le gestionnaire parvient-il à obtenir un rendement supérieur au référentiel, mais avec davantage de risque?

- Si le ratio est négatif, le portefeuille a moins performé que le référentiel et la situation est très mauvaise.
- Si le ratio est compris entre 0 et 0.5, le surrendement du portefeuille considéré par rapport au référentiel se fait pour une prise de risque trop élevée. Ou, le risque pris est trop élevé pour le rendement obtenu.
- Si le ratio est supérieur à 0.5, le rendement du portefeuille surperforme le référentiel pour une prise de risque ad hoc. Autrement dit, la surperformance ne se fait pas au prix d'un risque trop élevé.

Ce qui donne en développant:

$$S_p = \frac{\sum_{i=1}^n X_i (\alpha_i + \beta_i I + \varepsilon_i) - \sum_{i=1}^n X_i (\alpha_i + \varepsilon_i)}{\sqrt{\text{Var}(X_i (\alpha_i + \beta_i I + \varepsilon_i))}} = \frac{\sum_{i=1}^n \beta_i I}{\sqrt{\text{Var}(X_i (\alpha_i + \beta_i I + \varepsilon_i))}} \quad (66.422)$$

Signalons également un autre indicateur courant qui est le "**tracking error**" défini comme étant l'écart-type de l'écart de performance entre le portefeuille et le benchmark. Plus le tracking error est faible, plus le fonds ressemble à son indice de référence en termes de risque:

$$TE = \sqrt{V(R_p - R_f)} = \sigma_{R_p - R_f} \quad (66.423)$$

Ces modèles sont relativement complexes. Raison pour laquelle quelques années plus tard, Sharpe et Lintner ont créé un nouveau modèle qui leur a valu le prix Nobel d'économie et que nous allons étudier de suite après un exemple pratique de ce que nous venons de voir.

Exemple:

Considérons trois titres composants un portefeuille en proportions égales et les n observations de leurs rendements $R_{i,j}$ saisis dans Microsoft Excel 11.8346. Ces rendements seront comparés à un indice de référence I qui sera le rendement d'un portefeuille de marché de référence PF_M :

	A	B	C	D	E
5	Obs.	PF_M	Titre 1	Titre 2	Titre 3
6	1	0.45	-0.15	0.29	0.38
7	2	0.61	0.05	0.18	0.63
8	3	0.54	-0.43	0.24	0.46
9	4	0.29	0.79	0.25	0.36
10	5	0.12	0.32	0.17	-0.57

Figure: 66.38 - 3 titres avec leurs rendements comparés à un portefeuille de marché de référence

Le but est de déterminer la frontière d'efficience du portefeuille avec le modèle de Sharpe.

En détail sous forme graphique voici d'abord les bêtas (rendement de l'actif en fonction du rendement du portefeuille de marché/indice de référence) obtenus avec Microsoft Excel 11.8346:

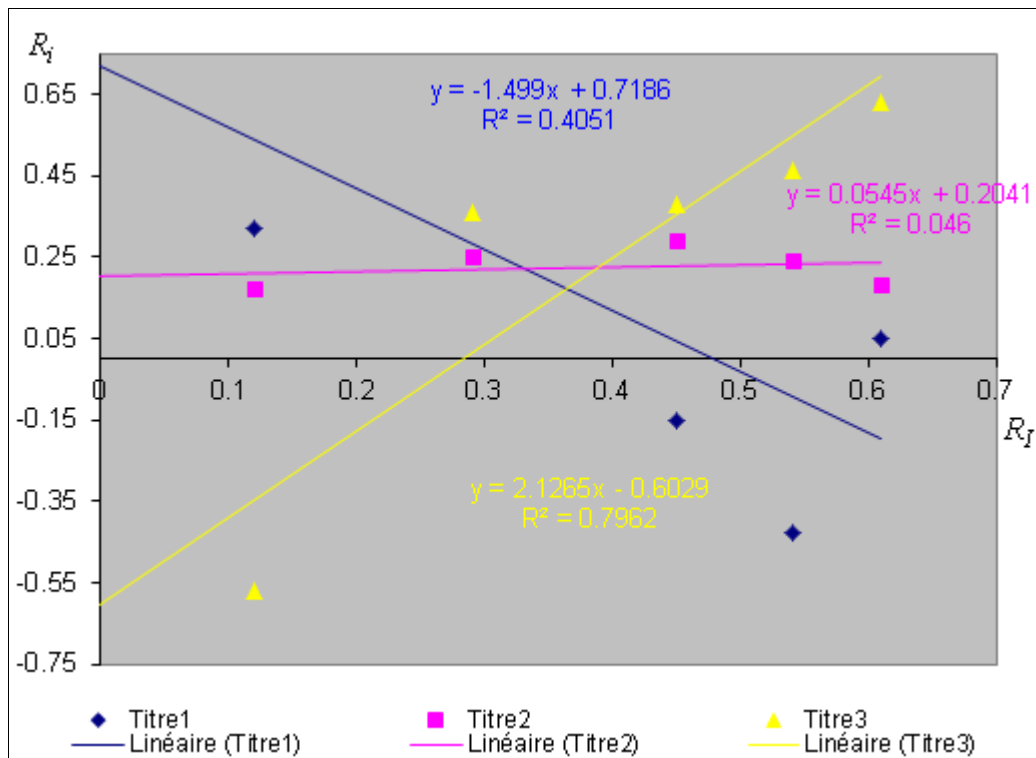


Figure: 66.39 - Régression linéaire pour le calcul des bêtas des titres

et le tableau de construction suivant pour le calcul des bêtas, la variance et l'espérance du portefeuille du marché et des différents titres:

11	Bêta: β_i	1.000	-1.499	0.055	2.126
12	Espérance: μ_i	0.402	0.116	0.226	0.252
13	Variance: σ_i^2	0.039	0.217	0.003	0.222

Figure: 66.40 - Indicateurs-clés correspondants du portefeuille de référence et des 3 titres

Voici les détails des formules (remarquez que les bêtas sont obtenus à l'aide d'une simple régression linéaire avec l'indice de référence qui est le portefeuille et les autres paramètres avec les estimateurs non biaisés):

	A	B	C
11	Bêta: β_i	=LINEST(B6:B10;\$B6:\$B10)	=LINEST(C6:C10;\$B6:\$B10)
12	Espérance: μ_i	=AVERAGE(B6:B10)	=AVERAGE(C6:C10)
13	Variance: σ_i^2	=VAR(B6:B10)	=VAR(C6:C10)

	A	D	E
11	Bêta: β_i	=LINEST(D6:D10;\$B6:\$B10)	=LINEST(E6:E10;\$B6:\$B10)
12	Espérance: μ_i	=AVERAGE(D6:D10)	=AVERAGE(E6:E10)
13	Variance: σ_i^2	=VAR(D6:D10)	=VAR(E6:E10)

Figure: 66.41 - Formules explicites pour les indicateurs précédents

L'espérance du rendement du portefeuille composé des trois titres est facile à calculer puisque nous avons leurs rendements. Donc:

$$E(R_p) = \sum_{i=1}^n X_i R_i = X_1 \hat{A}_1 + X_2 \hat{A}_1 + X_3 \hat{A}_1 \quad (66.424)$$

Ce qui donne sous Microsoft Excel 14.0.6123:

	A	B	C	D	E
15		<i>Titre 1</i>	<i>Titre 2</i>	<i>Titre 3</i>	<i>Somme</i>
16	X_i	1/3	1/3	1/3	1
17	$\hat{\mu}_i$	0.116	0.226	0.252	
18					
19		<i>Portefeuille</i>			
20	$E(R_p)$	0.198			

Figure: 66.42 - Espérance de chacun des titres et de l'ensemble du portefeuille

Soit de manière détaillée:

	A	B	C	D
15		<i>Titre 1</i>	<i>Titre 2</i>	<i>Titre 3</i>
16	X_i	1/3	1/3	1/3
17	$\hat{\mu}_i$	=C12	=D12	=E12
18				
19		<i>Portefeuille</i>		
20	$E(R_p)$	=SUMPRODUCT((B16:D16)*(B17:D17))		

Figure: 66.43 - Formules explicites pour les calculs de l'espérance

Maintenant, il nous faut calculer la variance en utilisant la relation démontrée dans la partie théorique des paragraphes précédents:

$$V(R_p) = \sigma^2(R_p) = \sigma_I^2 \cdot X^T \cdot (B_{\hat{y}} \cdot X) \quad (66.425)$$

avec pour rappel dans notre cas particulier:

$$B_{i,j} = \begin{bmatrix} \frac{\sigma_1^2}{\sigma_I^2} & \beta_1 \beta_2 & \beta_1 \beta_3 \\ \beta_2 \beta_1 & \frac{\sigma_2^2}{\sigma_I^2} & \beta_2 \beta_3 \\ \beta_3 \beta_1 & \beta_3 \beta_2 & \frac{\sigma_3^2}{\sigma_I^2} \end{bmatrix} \quad (66.426)$$

avec dans notre exemple $\sigma_I^2 = 0.039$ (cellule B13).

Soit sous forme développée pour notre exemple:

$$\begin{aligned} \sigma^2(R_p) &= X_1^2 \sigma_1^2 + X_2^2 \sigma_2^2 + X_3^2 \sigma_3^2 \\ &+ 2X_1 X_2 \sigma_I^2 \beta_1 \beta_2 + 2X_1 X_3 \sigma_I^2 \beta_1 \beta_3 + 2X_2 X_3 \sigma_I^2 \beta_2 \beta_3 \end{aligned} \quad (66.427)$$

Ce qui donne dans Microsoft Excel 11.8346 pour notre matrice des bêtas:

	A	B	C	D
23				
24	$B_{i,j}$	T1	T2	T3
25	T1	5.547102374	-0.081703338	-3.187578249
26	T2	-0.081703338	0.064590248	0.115907001
27	T3	-3.187578249	0.115907001	5.679601736

Figure: 66.44 - Matrice des bêtas

Soit sous forme développée (la matrice est symétrique):

	A	B	C	D
23				
24	$B_{i,j}$	T1	T2	T3
25	T1	=C13/B13	=C11*D11	=C11*E11
26	T2	=C11*D11	=D13/B13	=D11*E11
27	T3	=C11*E11	=D26	=E13/B13

Figure: 66.45 - Formules explicites pour la matrice des bêtas

Et finalement le couple variance/espérance du portefeuille est donné par:

	A	B
19		Portefeuille
20	$E(R_p)$	0.198
21	$\sigma^2(R_p)$	0.0217

Figure: 66.46 - Couple espérance/variance

Soit sous forme détaillée:

	A	B	C	D	E
19		Portefeuille			
20	$E(R_p)$	=SUMPRODUCT((B16:D16)*(B17:D17))			
21	$\sigma^2(R_p)$	=SUMPRODUCT(MMULT(B16:D16;B25:D27);B16:D16)*B13			

Figure: 66.47 - Formules espérance/variance

Une fois ceci fait, nous procédons comme pour la frontière de Markowitz. Nous utilisons le solveur en minimisant la variance tout en imposant une espérance et une contrainte comme quoi la somme des parts des actifs financiers est égale à l'unité:

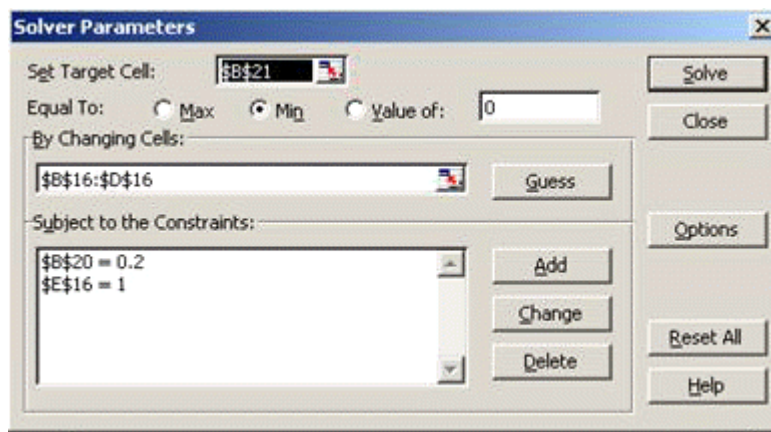


Figure: 66.48 - Paramétrages du solveur

Ce qui donne le tableau variance/rendement suivant (à comparer avec le même tableau de Markowitz):

<i>Itération</i>	<i>Variance</i>	<i>Rendement</i>
1	0.0110	0.200
2	0.0076	0.205
3	0.0051	0.210
4	0.0034	0.215
5	0.0025	0.220
6	0.0024	0.225
7	0.0032	0.230
8	0.0047	0.235
9	0.0071	0.240
10	0.0103	0.245

Figure: 66.49 - Itérations obtenues à l'aide du solveur

et le graphique suivant (comparaison directe avec Markowitz mise en évidence):

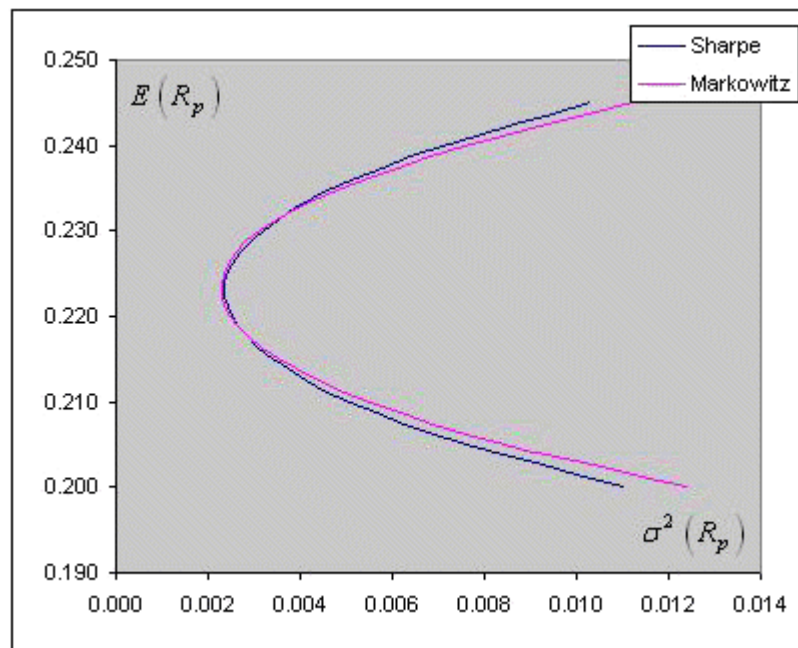


Figure: 66.50 - Comparaison entre les frontières de Sharpe et Markowitz

La suite de l'exercice (C.M.L.) se fait de la même manière que dans le modèle de Markowitz.

7.7. MODÈLE D'ÉVALUATION DES ACTIFS FINANCIERS (MEDAF)

Comme nous l'avons vu, Markowitz (1959) a développé la théorie du choix optimal d'un portefeuille par un individu sur la base de la variance du rendement espéré. Plus tard (1963), Sharpe élabore un modèle de choix d'actifs basé sur des indices de risques comme les coefficients bêta.

Sharpe, Lintner et Mossin (1965) ont ensuite étudié les conséquences de ces théories pour mettre en place une théorie extrêmement simple permettant d'évaluer les coefficients bêta, les rendements espérés et les variances d'actifs financiers d'un portefeuille à partir de données statistiques sur le marché global et de la spécificité de la composition d'un portefeuille.

Cette théorie basée encore une fois sur le problème moyenne-variance est appelée "**modèle d'évaluation des actifs financiers**" (MEDAF) ou encore "**modèle d'évaluation des actifs financiers à l'équilibre**" (MEDAFE) (en anglais: "**capital asset pricing model**" (C.A.P.M.)). C'est donc un modèle très souvent utilisé, aussi bien par les praticiens que par les académiciens, pour évaluer les rendements anticipés d'équilibre de n'importe quel actif risqué sur le marché.

Pour commencer, rappelons que nous avons vu plus haut lors de notre étude du return que le taux de rentabilité périodique (quotidien, hebdomadaire, mensuel, annuel) d'un actif se calcule comme suit:

$$r_t = \frac{(P_t - P_{t-1}) + C_t}{P_{t-1}} \quad (66.428)$$

avec P_t qui est le prix d'un actif à la fin de la période t , P_{t-1} le prix de ce même actif à la fin de la période $t-1$ et finalement C_t le flux monétaire payé par l'actif pendant la période de détention allant de $t-1$ à t .

Cette relation sert à calculer le "**rendement réalisé**" (ex post) d'un titre alors qu'en fait c'est le "**rendement espéré**" qui intéresse un investisseur donné.

À la date de la prise de la décision, le rendement que va réaliser l'investisseur en détenant un actif donné est incertain, c'est pour cette raison que l'on parle de rendement espéré: il s'agit d'un rendement que l'on cherche à évaluer et que l'on espère recevoir au cours de la prochaine période d'investissement.

Pour calculer le rendement espéré, comme nous l'avons déjà vu, il convient d'attribuer à chaque valeur possible du rendement une probabilité de réalisation, puis de calculer une moyenne pondérée de ces différentes valeurs possibles en utilisant les probabilités p_i comme pondérations:

$$E(R_i) = \mu_i = p_1 \cdot R_{i1} + p_2 \cdot R_{i2} + \dots + p_n \cdot R_{in} = \sum_{j=1}^n p_j \cdot R_{ij} \quad (66.429)$$

Or, il est clair que dans une économie donnée, l'investisseur sera tenté de détenir plusieurs actifs financiers et cherchera donc à composer des portefeuilles. Le rendement (moyen) espéré d'un portefeuille peut être calculé en utilisant la relation connue:

$$E(R_p) = X_1 \cdot E(R_1) + X_2 \cdot E(R_2) + \dots + X_n \cdot E(R_n) = \sum_{i=1}^n X_i \cdot E(R_i) \quad (66.430)$$

avec n qui est le nombre de titres inclus dans le portefeuille, $E(R_i)$ le rendement de l'actif i inclus dans le portefeuille et X_i la proportion de la richesse totale de l'investisseur investie dans l'actif i .

Le taux de rendement espéré est cependant insuffisant pour caractériser une opportunité d'investissement et il faut tenir compte également du risque, c'est-à-dire de la variabilité du rendement de cet investissement sur l'actif financier. La variance est comme nous l'avons déjà vu utilisée comme mesure du risque et donnée pour un actif financier par:

$$V(R_i) = \sigma_i^2 = p_1 \cdot [R_{i1} - E(R_i)]^2 + p_2 \cdot [R_{i2} - E(R_i)]^2 + \dots + p_n \cdot [R_{in} - E(R_i)]^2 \quad (66.431)$$

Soit:

$$\sigma_i^2 = \sum_{j=1}^n (R_{ij} - E(R_i))^2 p_j \quad (66.432)$$

Le calcul du risque d'un portefeuille fait donc intervenir deux concepts importants: la variabilité du rendement de chacun des actifs, mesurée par les variances de ces derniers, ainsi que les relations existantes entre les différents actifs composant le portefeuille.

La dépendance entre deux actifs est souvent mesurée, comme nous en avons déjà fait mention lors de notre étude des return, par la covariance ou encore le coefficient de corrélation linéaire.

La covariance entre deux actifs i et j se calcule comme suit:

$$\begin{aligned} \text{cov}(R_i, R_j) = & p_1 [R_{i1} - E(R_i)] \cdot [R_{j1} - E(R_j)] + p_2 [R_{i2} - E(R_i)] \cdot [R_{j2} - E(R_j)] + \dots + \\ & p_n [R_{in} - E(R_i)] \cdot [R_{jn} - E(R_j)] \end{aligned} \quad (66.433)$$

Soit comme nous le savons si les probabilités sont équiprobables:

$$\text{cov}(R_i, R_j) = \frac{1}{n} \sum_{k=1}^n (R_{ik} - E(R_i))(R_{jk} - E(R_j)) \quad (66.434)$$

La covariance entre les rendements de deux titres peut être positive ou négative et sa valeur n'a aucune signification économique comme nous le savons (cf. chapitre de Statistiques).

Remarque: Rappelons que nous avons vu dans le chapitre de Statistiques que, lorsque les rendements (valeurs) de deux actifs (variables aléatoires) varient dans le même sens (dans le sens contraire) leur covariance sera positive (négative).

Le coefficient de corrélation entre deux actifs i et j quant à lui se calcule comme suit (cf. chapitre de Statistiques):

$$R_{i,j} = \frac{\text{cov}(R_i, R_j)}{\sqrt{V(R_i) \cdot V(R_j)}} = \frac{c_{ij}}{\sigma_i \cdot \sigma_j} \quad (66.435)$$

Une fois les variances et covariances des différents actifs calculées, nous serons en mesure de calculer la variance de rendement d'un portefeuille contenant n actifs. Cette variance est donnée par la relation suivante (cf. chapitre de Statistiques):

$$V(R_p) = \sum_{i=1}^n \sum_{j=1}^n X_i X_j \text{cov}(R_i, R_j) = \sum_{i=1}^n X_i^2 \sigma_i^2 + \sum_{i \neq j} X_i X_j \text{cov}(R_i, R_j) \quad (66.436)$$

ou écrit autrement:

$$V(R_p) = \sum_{i=1}^n X_i^2 V(R_i) + 2 \sum_{i < j} X_i X_j \text{cov}(R_i, R_j) \quad (66.437)$$

La relation ci-dessus de la variance de rendement d'un portefeuille montre clairement que même dans le cas où les rendements des différents actifs détenus dans le portefeuille sont totalement non corrélés, la variance de ce dernier peut encore être réduite en ajoutant plus d'actifs.

Pour comprendre ceci, nous noterons que pour n actifs non corrélés, la variance se réduit à (puisque la covariance est alors nulle):

$$V(R_p) = \sigma_p^2 = \sum_{i=1}^n X_i^2 \sigma_i^2 \quad (66.438)$$

En simplifiant davantage, si toutes les variances sont supposées égales et si tous les actifs sont détenus dans les mêmes proportions ($1/n$), nous avons (cf. chapitre de Statistiques):

$$\sigma_p^2 = \sum_{i=1}^n X_i^2 \sigma_i^2 = \left(\frac{1}{n}\right)^2 n \sigma_i^2 = \frac{1}{n} \sigma_i^2 \quad (66.439)$$

Ainsi, quand n tend vers l'infini, la variance du portefeuille s'approche de zéro. Ainsi, si des risques non corrélés sont réunis en portefeuille, le risque total peut être éliminé par diversification. Dans le cas où les risques sont corrélés, la diversification ne permettra d'éliminer que les risques spécifiques aux actifs alors que le risque de marché continuera d'exister. Notons que la réduction du risque serait plus importante lorsque les différents actifs détenus sont négativement corrélés. En effet, plus le coefficient de corrélation entre les rendements des titres est petit, plus les bénéfices inhérents à la diversification sont substantiels. Dans le cas où le coefficient de corrélation est égal à 1, il n'y a aucun bénéfice lié à la diversification, puisque le risque du portefeuille sera égal à la moyenne pondérée des risques le composant. Par contre, la diversification est à son maximum lorsque le coefficient de corrélation est égal à -1. Dans cette situation, il est possible de combiner deux actifs risqués pour former un portefeuille sans risque.

D'après ce qui précède, il est clair que tout investisseur désirant former un portefeuille cherchera à détenir un ensemble d'actifs risqués qui lui permettra de recevoir un rendement donné avec un minimum de risque. En d'autres termes, il cherchera à minimiser la variance pour un niveau de rendement espéré tout en respectant une contrainte budgétaire. Nous savons que le rendement espéré et la variance de rendement d'un portefeuille contenant n actifs risqués s'écrivent comme suit:

$$E(R_p) = \sum_{i=1}^n X_i \cdot E(R_i) \quad (66.440)$$

$$Var(R_p) = \sum_{i=1}^n \sum_{j=1}^n X_i X_j \sigma_{ij}$$

Par ailleurs, nous savons qu'à partir de ces n titres, il est possible de construire une infinité de portefeuilles en faisant varier les pondérations X_i . Or, les portefeuilles les plus intéressants pour un investisseur donné sont ceux qui permettent de minimiser le risque qu'il doit supporter pour obtenir un niveau de rendement donné. Ces portefeuilles sont le résultat du problème de minimisation suivant qui est un problème d'optimisation non linéaire (cf. chapitre de Méthodes Numériques):

$$\begin{aligned} \min : & Var(R_p) \\ & E(R_p) = c^{te} \\ & \sum X_i = 1 \end{aligned} \quad (66.441)$$

que nous avons déjà vu lors de notre étude du modèle de Markowitz.

Il est donc possible de constituer une infinité de portefeuilles en faisant varier les proportions investies dans chacun des titres. La prochaine étape consiste à sélectionner, parmi l'ensemble des portefeuilles disponibles, un portefeuille donné. Pour ce faire, on doit considérer les préférences individuelles de l'investisseur.

Un investisseur rationnel ne devrait donc considérer que les portefeuilles se trouvant sur la frontière efficiente pour ses choix d'investissement. Son portefeuille optimal se situera au point de tangence entre la frontière efficiente et sa courbe d'indifférence la plus haute qu'il serait capable d'atteindre. En procédant ainsi, chaque investisseur maximisera son utilité espérée. En présence d'une économie ne contenant que des actifs risqués, la composition du portefeuille d'actifs risqués varie d'un individu à un autre.

En pratique, les investisseurs ont également la possibilité d'investir dans des actifs financiers sans risques. Nous allons donc chercher à déterminer la nouvelle frontière efficiente en tenant compte de cette nouvelle opportunité d'investissement.

Considérons alors un portefeuille qui est une combinaison de l'actif sans risque et d'un portefeuille de marché (à risque). Nous avons alors:

$$R_p = X_m R_m + (1 - X_m) R_f \quad (66.442)$$

où X_m est la fraction du portefeuille investie dans le portefeuille du marché (m) et R_f est le "taux de rendement certain" ou "taux de rendement sans risque".

Rappel: L'espérance d'une constante est égale à cette constante (cf. chapitre de Statistiques).

Nous avons donc:

$$E(R_p) = X_m E(R_m) + (1 - X_m) E(R_f) = X_m E(R_m) + (1 - X_m) R_f \quad (66.443)$$

et donc:

$$\begin{aligned}
V(R_p) &= V(X_m R_m + (1 - X_m) R_f) \\
&= E \left[\left(X_m R_m + (1 - X_m) R_f - E(X_m R_m + (1 - X_m) R_f) \right)^2 \right] \\
&= E \left[\left(X_m R_m + (1 - X_m) R_f - X_m E(R_m) - (1 - X_m) R_f \right)^2 \right] \quad (66.444) \\
&= E \left[\left(X_m R_m - X_m E(R_m) \right)^2 \right] = E \left[X_m^2 (R_m - E(R_m))^2 \right] \\
&= X_m^2 E \left[(R_m - E(R_m))^2 \right] = X_m^2 V(R_m)
\end{aligned}$$

Soit sous forme condensée:

$$\sigma_p^2 = X_m^2 \sigma_m^2 \quad (66.445)$$

La dérivée du rendement espéré par rapport à X_m nous donne:

$$\frac{\partial E(R_p)}{\partial X_m} = E(R_m) - R_f \quad (66.446)$$

La dérivée de l'écart-type par rapport à X_m nous donne:

$$\frac{\partial \sigma(R_p)}{\partial X_m} = \sigma(R_m) \quad (66.447)$$

Mettant ces deux résultats ensemble, nous avons:

$$\frac{\partial E(R_p)}{\partial \sigma(R_p)} = \frac{E(R_m) - R_f}{\sigma(R_m)} \quad (66.448)$$

Cette équation nous donne la pente de la "**capital market line**" (C.M.L.). Elle est constante (la pente!), et donc la C.M.L. est une droite. L'ordonnée à l'origine est évidemment R_f .

Puisque:

$$E(R_p) = X_m E(R_m) + (1 - X_m) R_f = X_m (E(R_m) - R_f) + R_f \quad (66.449)$$

l'équation de la C.M.L. se réduit alors à:

$$E(R_p) = \frac{\sigma(R_p)}{\sigma(R_m)} (E(R_m) - R_f) + R_f \quad (66.450)$$

Et puisque dans la finance l'intérêt est de représenter graphiquement .

$$E(R_p) = f(\sigma(R_p)) \quad (66.451)$$

Alors, il est de tradition de noter la fonction sous la forme suivante:

$$E(R_p) = \frac{E(R_m) - R_f}{\sigma(R_m)} \sigma(R_p) + R_f \quad (66.452)$$

où nous retrouvons en facteur de l'écart-type de R_p le coefficient appelé "**Sharpe ratio**" (ou ratio de Sharpe) dont nous avons parlé plus haut mais sans en démontrer la provenance.

Par construction, cette droite associe donc à chaque niveau de risque, la rentabilité espérée la plus élevée. Ainsi, étant donné le rendement d'un actif sans risque il devient facile à partir de cette équation de déterminer le point de tangence avec la frontière d'efficacité de Markowitz ou de Sharpe pour obtenir le portefeuille le plus efficace sur la base du rendement sans risque!!

Intéressons-nous maintenant à déterminer une équation pour le rendement espéré de n'importe quel actif individuel.

Considérons un nouveau portefeuille de rendement $\tilde{R}_p$ qui est une combinaison d'un actif sans risque quelconque A et du portefeuille de marché, où X_A est la fraction du portefeuille investie dans l'actif sans risque A .

Ce que nous souhaiterions évaluer est la pente de la courbe des combinaisons espérance/écart-type lorsque nous combinons le portefeuille de marché (qui contient déjà l'actif A) avec l'actif A .

Nous souhaitons évaluer la valeur de la pente de l'équation tangente à la frontière efficace telle que la pondération de l'actif sans risque A soit nulle.

Nous avons:

$$\tilde{R}_p = X_A R_A + (1 - X_A) R_m \quad (66.453)$$

Nous obtenons de suite:

$$E(\tilde{R}_p) = X_A E(R_A) + (1 - X_A) E(R_m) \quad (66.454)$$

et (cf. chapitre de Statistiques):

$$V(\tilde{R}_p) = V(R_A) + V(R_m) + 2 \text{cov}(R_A, R_m) \quad (66.455)$$

donc:

$$\sigma^2(\tilde{R}_p) = X_A^2 \sigma^2(R_A) + (1 - X_A)^2 \sigma^2(R_m) + 2 X_A (1 - X_A) \text{cov}(R_A, R_m) \quad (66.456)$$

Dérivant le rendement espéré de ce nouveau portefeuille par rapport à X_A , nous obtenons:

$$\frac{\partial E(\tilde{R}_p)}{\partial X_A} = E(R_A) - E(R_m) \quad (66.457)$$

Dérivant l'écart-type du rendement de ce nouveau portefeuille par rapport à X_A , nous obtenons:

$$\frac{\partial \sigma(\tilde{R}_p)}{\partial X_A} = \frac{1}{2} \frac{2X_A \sigma^2(R_A) - 2(1 - X_A) \sigma^2(R_m) + (2 - 4X_A) \text{cov}(R_A, R_m)}{\sqrt{X_A^2 \sigma^2(R_A) + (1 - X_A)^2 \sigma^2(R_m) + 2X_A(1 - X_A) \text{cov}(R_A, R_m)}} \quad (66.458)$$

La contribution de Sharpe et Lintner a été de dire qu'il faut évaluer ces dérivées au point où $X_A = 0$ c'est-à-dire où la pondération de l'actif A dans le nouveau portefeuille est nulle.

Ce faisant, nous obtenons, l'expression suivante pour l'écart-type du nouveau portefeuille (bien sûr, l'expression pour le rendement espéré ne change pas):

$$\left. \frac{\partial \sigma(\tilde{R}_p)}{\partial X_A} \right|_{X_A=0} = \frac{1}{2} \frac{-2\sigma^2(R_m) + 2\text{cov}(R_A, R_m)}{\sqrt{\sigma^2(R_m)}} \quad (66.459)$$

ce qui donne après simplification:

$$\left. \frac{\partial \sigma(\tilde{R}_p)}{\partial X_A} \right|_{X_A=0} = \frac{\text{cov}(R_A, R_m) - \sigma^2(R_m)}{\sigma(R_m)} \quad (66.460)$$

Avec les deux dérivées, nous pouvons obtenir une expression pour la courbe de combinaisons espérance/écart-type pour le nouveau portefeuille. Nous avons alors:

$$\left. \frac{\partial E(\tilde{R}_p)}{\partial \sigma(\tilde{R}_p)} \right|_{X_A=0} = \frac{E(R_A) - E(R_m)}{[\text{cov}(R_A, R_m) - \sigma^2(R_m)] / \sigma(R_m)} \quad (66.461)$$

Cette pente doit être égale à celle de la C.M.L. En égalisant, nous obtenons:

$$\frac{E(R_m) - R_f}{\sigma(R_m)} = \frac{E(R_A) - E(R_m)}{[\text{cov}(R_A, R_m) - \sigma^2(R_m)] / \sigma(R_m)} \quad (66.462)$$

Quelques manipulations algébriques et nous y sommes! Nous avons:

$$\frac{E(R_m) - R_f}{\sigma^2(R_m)} = \frac{E(R_A) - E(R_m)}{\text{cov}(R_A, R_m) - \sigma^2(R_m)} \quad (66.463)$$

et donc:

$$\begin{aligned} E(R_A) &= \frac{E(R_m) - R_f}{\sigma^2(R_m)} (\text{cov}(R_A, R_m) - \sigma^2(R_m)) + E(R_m) \\ &= (E(R_m) - R_f) \frac{\text{cov}(R_A, R_m)}{\sigma^2(R_m)} - E(R_m) + R_f + E(R_m) \\ &= (E(R_m) - R_f) \frac{\text{cov}(R_A, R_m)}{\sigma^2(R_m)} + R_f \end{aligned} \quad (66.464)$$

d'où:

$$E(R_A) - R_f = (E(R_m) - R_f) \frac{\text{cov}(R_A, R_m)}{\sigma^2(R_m)} \quad (66.465)$$

En posant ce que nous avons déjà vu lors de notre étude du modèle de Sharpe, c'est-à-dire le risque non diversifiable sous forme de facteur bêta:

$$\beta = \frac{\text{cov}(R_A, R_m)}{\sigma^2(R_m)} \quad (66.466)$$

c'est donc la volatilité de la rentabilité de l'actif considéré rapportée à celle du marché.

Nous avons alors:

$$E(R_A) - R_f = (E(R_m) - R_f) \beta \quad (66.467)$$

Cette expression permet donc d'exprimer le rendement excédentaire d'un actif comme le produit du rendement excédentaire du portefeuille de marché et le facteur bêta du titre.

Le rendement excédentaire d'un actif ne dépend pas directement que de sa variance, qui est souvent une mesure intuitive du risque d'un actif. Ce qui compte est son facteur bêta, qui dépend de sa covariance avec le portefeuille de marché.

Plus classiquement, la dernière relation est utilisée graphiquement sous forme de droite:

$$E(R_A) = f(\beta) = (E(R_m) - R_f) \beta + R_f \quad (66.468)$$

Cette droite est appelée la "**security market line**" (S.M.L.) elle est extrêmement importante en finance, car elle donne donc le rendement moyen d'un titre A en fonction du bêta, du rendement du marché et du taux sans risque.

Exemple:

Le taux sans risque est de 5% et la rentabilité espérée du marché est de 8%. Une action A est deux fois plus sensible aux mouvements de marché que l'indice de référence, in extenso lorsque l'indice enregistre une hausse de 1%, l'action A montre de 2%. D'après la S.M.L., la rentabilité espérée est alors de:

$$E(R_A) = f(\beta) = (E(R_m) - R_f) \beta + R_f \quad (66.469)$$

Remarque: Le lecteur vérifiera facilement que dans la relation ci-dessus, si le bêta est égal à l'unité, le rendement espéré du titre A sera égal au rendement espéré du marché. Le MEDAF est donc un outil de détermination de supplément de rentabilité que l'on doit attendre lorsqu'on s'investit dans un portefeuille ou une société donnée (société dont le rendement est estimé sur la base de ses dividendes si elle est cotée ou sinon comparée à des sociétés équivalentes qui elles seront cotées).

On la trouve aussi fréquemment sous la forme suivante:

$$E(R_A) = \lambda \beta + R_f \quad (66.470)$$

avec λ qui est appelé la "**prime par unité de risque**" ou plus simplement "**prime de risque**" (surplus de rentabilité exigé par les investisseurs lorsque ces derniers placent leur argent sur le marché plutôt que dans un actif sans risque) et l'ordonnée à l'origine est le taux d'intérêt sans risque (généralement des emprunts d'État).

Le MEDAF stipule donc que le taux de rendement espéré (ou que devrait exiger un investisseur rationnel ayant une aversion pour le risque) d'un actif risqué doit être égal au taux de rendement de l'actif sans risque, plus une prime de risque appelée parfois "**spread de crédit**". Dans ce cas, la relation entre le risque systématique et le rendement espéré demeure linéaire et seul le risque systématique doit être rémunéré par le marché puisque le risque spécifique peut être éliminé grâce à la diversification.

Il est peut-être intéressant d'expliciter les hypothèses sur lesquelles reposent mathématiquement les développements que nous avons faits. Ce sont donc les hypothèses du MEDAF dont certaines semblent difficilement acceptables. Il ne faut cependant pas oublier que la validité d'un modèle ne dépend pas du réalisme de ses hypothèses mais bien de la conformité de ses implications avec la réalité.

Nous avons pour rappel émis les hypothèses suivantes:

H1. Les investisseurs composent leurs portefeuilles en se préoccupant exclusivement de l'espérance et de la variance de rendement de ces derniers

H2. Les investisseurs ont une aversion pour le risque: ils n'aiment pas le risque

H3. Il n'y a pas de coût de transaction et les actifs sont parfaitement divisibles

H4. Ni les dividendes, ni les gains en capitaux ne sont taxés

H5. De nombreux acheteurs et vendeurs interviennent sur le marché et aucun d'entre eux ne peut avoir d'influence sur les prix.

H6. Tous les investisseurs peuvent prêter ou emprunter le montant qu'ils souhaitent au taux sans risque.

H7. Les anticipations des différents investisseurs sont homogènes

H8. La période d'investissement est la même pour tous les investisseurs

7.8. MODÈLE D'ÉVALUATION DES OPTIONS DE BLACK & SCHOLES

C'est au génie de trois célèbres mathématiciens que le marché des dérivés doit son succès, grâce à l'équation de Black & Scholes conçue dans les années 1970 (et publiée en 1973) qui permet de déterminer théoriquement la prime exacte que doit payer un client pour acquérir un Call ou un Put et la stratégie que devra suivre le vendeur de ces options pour se couvrir du risque (pour ne citer que l'exemple le plus connu). Évidemment ce modèle ne fonctionne que si les périodes temporelles considérées sont relativement courtes (de l'ordre de la semaine ou de quelques mois au mieux). Au-delà l'utilisation de ce modèle théorique particulier est une farce!

Black, Scholes et Merton sont les ancêtres d'une génération de produits dérivés sophistiqués, donnant droit de cité à tout un lexique de termes aussi exotiques que Butterflies, Rainbows, Knock-in, Knock-out, Barrières, Swaps, Calls, Puts, Baskets, Swings. Ce modèle est aussi considéré ceci dit comme un des facteurs principaux du crash boursier de 1987 par certains spécialistes...

Ce que nous aimerions dans ce qui va suivre est déterminer la valeur théorique de la prime d'une option à partir des cinq données suivantes:

1. La valeur actuelle de l'actif financier sous-jacent de l'option (déterminée par la spéculation du marché).
2. Le temps qui reste à l'option avant son échéance (choisie par la société émettrice).
3. Le prix d'exercice (strike) fixé par l'émetteur subjectivement ou après modélisation.
4. Le taux d'intérêt sans risque (supposé comme étant le taux de rendement attendu du sous-jacent).
5. La volatilité (écart-type) du prix du sous-jacent de l'option (mesurée sur le marché).

La prime de l'option ainsi déterminée sera unique et équitable pour les deux parties. Effectivement, le système des options permettrait de faire payer un prix d'option (prime) majoré par rapport aux prévisions du marché et donc de générer à coup sûr et à partir de rien un profit mais les nombreux acteurs du marché vont faire jouer la concurrence pour être au plus juste et attirer le client sur leurs options plutôt que sur celles de la concurrence.

La modélisation du cours des options (Black & Scholes) repose sur l'utilisation du calcul différentiel stochastique. Ainsi, l'approche de Black et Scholes suppose que l'évolution du cours de l'action définit un mouvement brownien géométrique (dans le sens que les mouvements possibles du prix tendent vers l'infini) et que son rendement définit un processus de Wiener généralisé (concept que nous allons définir un peu plus loin).

7.8.1. ÉQUATION DE PARITÉ CALL-PUT

Avant de nous attaquer à des calculs stochastiques un peu ardu, il est utile d'établir au préalable une équation dite de "**parité Call-Put**" qui nous servira de sorte d'équation de conservation pour vérifier la validité des résultats que nous établirons par la suite sur l'évaluation des prix des options.

L'objectif va être de répondre à la question suivante: Quelle somme M devons-nous payer maintenant pour recevoir une somme garantie E correspondant au prix "**prix d'Exercice**" (qui est la notation française du "**strike price**" K utilisée jusqu'à maintenant) à un temps futur T ?

Ainsi, nous avons vu lors de notre étude du calcul d'intérêts qu'en considérant un capital C et un intérêt r constant nous avons dans un cas de capitalisation continue:

$$C = C_0 e^{rt} \quad (66.471)$$

Dès lors, en posant $C = E$ et $C_0 = M$ nous avons:

$$E = M e^{rt} \quad (66.472)$$

d'où:

$$M = E e^{-rt} \quad (66.473)$$

Mais cette relation n'est pas tout à fait juste. Effectivement, nous devons avoir $M = E$ assuré au temps $T = t$. Dès lors nous sommes naturellement amenés à poser:

$$M = Ee^{-r(T-t)} \quad (66.474)$$

Nous allons maintenant supposer que le Call et le Put possèdent les propriétés suivantes:

P1. Même support qui vaut S à l'instant t (spot).

P2. Même échéance T

P3. Même prix d'exercice E (souvent noté K)

et les hypothèses suivantes:

H1. Il n'existe pas de coûts de transaction

H2. Le support n'est pas un instrument à terme (i.e. qu'il est payable ou livrable immédiatement)

H3. Le support ne verse pas de dividendes pendant la durée de vie de l'option (i.e. entre $[0, T]$).

H4. Les options sont européennes

En nous reposant maintenant la question initiale: Quelle somme devons-nous payer maintenant pour un portefeuille afin de recevoir une somme garantie E (prix d'exercice) à un temps futur T ?

Le portefeuille pouvant être considéré comme une boîte noire, rien ne nous empêche dès lors de répondre en écrivant avec S le prix du jour du sous-jacent, P celui du Put et C celui du Call:

$$\Pi = S + P - C = Ee^{-r(T-t)} \quad (66.475)$$

qui n'est rien d'autre que "**l'équation de parité Call-Put**" ou autrement écrite si nous adoptons la notation anglo-saxonne K au lieu de E et nous prenons comme début le temps zéro, alors:

$$S + P - C = Ke^{-rT} \quad (66.476)$$

En toute rigueur, il vaudrait cependant mieux écrire cette dernière écriture sous la forme suivante:

$$C_t - P_t = S_t - Ke^{-r(T-t)} \quad (66.477)$$

Quel que soit le réarrangement des termes dans cette dernière égalité, il doit toujours y avoir égalité sinon quoi il y a opportunité d'arbitrage. Cette relation montre aussi que la valeur d'un Call européen avec prix d'exercice K et maturité T peut être déduite de celle d'un Put européen avec le même prix d'exercice K et la même maturité T .

Voyons plus en détails suite à la demande d'un lecteur que cette dernière relation découle de l'absence d'opportunité d'arbitrage (O.A.O.):

Considérons la configuration fréquente et naturelle d'un premier agent ayant un portefeuille A contenant une option d'achat (Call) européenne de prix C et une somme Ke^{-rt} investie dans un placement sans risque. La valeur initiale de ce portefeuille sera donc:

$$P_A = C + Ke^{-rT} \quad (66.478)$$

Considérons la deuxième configuration naturellement opposée à la première d'un portefeuille contenant une option de vente (Put) européenne de prix P et le sous-jacent de prix S . La valeur initiale de ce portefeuille sera donc:

$$P_B = P + S \quad (66.479)$$

Nous avons alors deux scénarios:

S1. À l'échéance (maturité) T , si le sous-jacent a une valeur inférieure ou égale au prix d'exercice (soit $S_T \leq K$), le portefeuille A vaudra K car l'option Call ne sera très probablement pas exercée et donc sa valeur financière devient nulle mais l'actif sans risque (placé au rendement sans risque) aura lui rapporté K . Le portefeuille B quant à lui vaudra $(K - S_T) + S_T = K$ où le terme entre parenthèse représente le gain généré par le Put.

S2. À l'échéance (maturité) T , si le sous-jacent a une valeur supérieure au prix d'exercice (soit $S_T > K$) nous sommes donc une situation contraire à celle du premier scénario. Le portefeuille A vaudra alors $(S_T - K) + K = S_T$. Le portefeuille B vaudra lui $0 + S_T = S_T$ car la vente ne sera pas effectuée et l'option de vente sera alors considérée comme ayant une valeur nulle.

Pour résumer:

	si $S_T \leq K$	si $S_T > K$
P_A	$0 + K = K$	$(S_T - K) + K = S_T$
P_B	$(K - S_T) + S_T = K$	$0 + S_T = S_T$
	$A = B$	$A = B$

Tableau: 66.15 - Résumé de l'absence d'opportunité d'arbitrage

Donc les deux portefeuilles sont toujours égaux à maturité sinon quoi il y a opportunité d'arbitrage. Ce que nous pouvons écrire (l'indice T pour le cours du sous-jacent à maturité est souvent omis):

$$C + Ke^{-rT} = P + S \quad (66.480)$$

7.8.2. HYPOTHÈSE D'EFFICIENCE DU MARCHÉ

Le modèle de Black & Scholes (et beaucoup d'autres modèles financiers) se base sur le postulat que le marché est "efficient".

Définition: Un "**marché efficient**" ("Efficient Market Hypothesis" en anglais... - abrégé E.M.H) est un marché où les prix reflètent complètement toute l'information disponible. Ainsi, si le marché est efficient, il n'est pas possible de faire des profits anormaux.

Nous pouvons distinguer trois types de marchés efficients qui sont fonction du type d'information disponible:

1. L'hypothèse de marché efficient en "forme faible" qui explicite que les prix reflètent toute l'information contenue dans la série historique des prix.

2. L'hypothèse de marché efficient en "forme semi-forte" établit que les prix reflètent toute l'information publique disponible.

3. L'hypothèse de marché efficient en "forme forte" qui établit que toute l'information connue, publique et privée, est reflétée dans les prix du marché.

Plusieurs études ont essayé de tester l'hypothèse de l'efficience des marchés des actifs. Pour tester la forme faible de l'hypothèse, on a utilisé l'analyse des séries temporelles (voir plus loin) en testant spécifiquement l'hypothèse d'une marche au hasard (mouvement brownien - nous y reviendrons). Plus spécifiquement, ces tests ont essayé de tester si les accroissements des prix sont indépendants des accroissements passés. Si l'hypothèse d'une marche au hasard est rejetée, alors le marché n'est pas efficient, car les accroissements de prix passés pourraient aider à anticiper les prix futurs des actifs. L'évidence empirique soutient l'hypothèse de marché efficient en forme faible. Pour tester la forme semi-forte de l'hypothèse, on a évalué la vitesse d'ajustement des prix de marché à l'arrivée de nouvelles informations; l'évidence en faveur d'un rapide ajustement des prix de marché est dominante. La forme forte de l'hypothèse de l'efficience des marchés, consiste à tester s'il est possible de tirer profit d'une information privilégiée (information accessible à un petit groupe d'agents économiques). Étant donné qu'on ne peut pas identifier l'information non publique, un type de test de forme forte considère l'examen de la performance d'investissement des individus ou groupes qui pourraient avoir de l'information privée. Elton et Gruber (1984) signalent que l'analyse de la performance des fonds mutuels, après déduction des coûts, soutient la forme forte de l'efficience.

Ceci implique les hypothèses suivantes (pour résumer en gros):

H1. L'histoire passée du cours de l'option est complètement réfléchié dans le prix présent qui ne contient lui pas d'autres informations sur l'option

H2. Le marché incorpore immédiatement toute nouvelle information dans le prix d'une option.

Le paradoxe du postulat des marchés efficients tient à ce que si chaque investisseur pensait vraiment que le marché était parfaitement efficient, alors personne n'étudierait les sociétés, leurs bilans, etc. Il suffirait d'acheter de l'indice. En vérité, les marchés efficients dépendent d'individus actifs sur le marché parce qu'ils pensent que ce marché est "inefficient" et qu'ils peuvent faire mieux que le marché !

Ce postulat est source de beaucoup de débats dans le domaine...

Remarque: Avec les deux hypothèses précédemment énoncées, tout changement non anticipé dans le prix de l'option est appelé un "**processus de Markov**".

Rappel: Un processus de Markov est un processus dont l'évolution future ne dépend de son passé qu'à travers son état à l'instant présent. Or, le cours d'une action n'est vraisemblablement pas un processus de Markov (la "mémoire" du processus est probablement plus longue - par exemple une tendance saisonnière).

7.8.3. PROCESSUS DE WIENER

Soit Δx la variation de la valeur d'une option (ou autre actif financier volatile) sur un petit intervalle de temps noté Δt .

Nous posons que (dans le sens que la variation de l'option est similaire à la variation de la valeur du sous-jacent!):

$$\Delta x = \Delta z \quad (66.481)$$

et à l'aide de la connaissance des deux résultats majeurs du modèle de Bachelier vu plus haut nous avons donc pour les variations de la valeur de l'option une espérance positive dépendante de manière proportionnelle à la racine carrée du temps selon:

$$\Delta z = \varepsilon \sqrt{\Delta t} \quad (66.482)$$

où nous posons comme hypothèse (acceptable... car nous travaillons sur de petites variations pour rappel!) que le coefficient d'instabilité est une fonction:

$$\varepsilon \rightarrow N(0,1) \quad (66.483)$$

où rappelons-le, $N(0,1)$ est la notation de la loi Normale centrée réduite telle que nous l'avons établie dans le chapitre de Statistiques.

Remarque: Souvent dans le domaine de l'économie, nous notons WN au lieu de N en hommage à Wiener.

Ceci dit, la relation antérieure est souvent notée de manière généralisée:

$$W_t = W_{t-1} + \varepsilon \sqrt{\Delta t} = W_{t-1} + N(0,1) \sqrt{\Delta t} \quad (66.484)$$

et définie comme étant un "**mouvement brownien standard**" avec "**bruit blanc**" (loi marginale de type Normale), ou "**mouvement brownien arithmétique**", où le W est là par hommage à Wiener! Il est intéressant de remarquer que le mouvement brownien est supposé indéfiniment divisible (ce qui signifie que la période temporelle prise n'influe pas sur la loi de probabilité qui reste toujours la même... c'est une propriété fractale du mouvement brownien qui a été creusée par Mandelbrot aussi!).

Ne pas oublier pour une démonstration que nous ferons plus loin que nous en déduisons:

$$W_{t+1} - W_t = N(0,1) \sqrt{\Delta t} = N(0,1) \sqrt{(t+1) - t} = N(0,1) \quad (66.485)$$

Il est possible de produire un graphique de ce mouvement brownien dans la version anglaise de Microsoft Excel 11.8346 avec dans la colonne A le temps avec un pas Δt typique de 0.01 [s] (colonne qui sera là uniquement par souci de confort de lecture et par tradition) et dans la cellule B2 la formule suivante:

$$=B1+NORMSINV(RAND())*SQRT(0.01)$$

où B1 contient la valeur 0.

Nous obtenons alors pour 4 colonnes du même type les variations de valeurs suivantes:

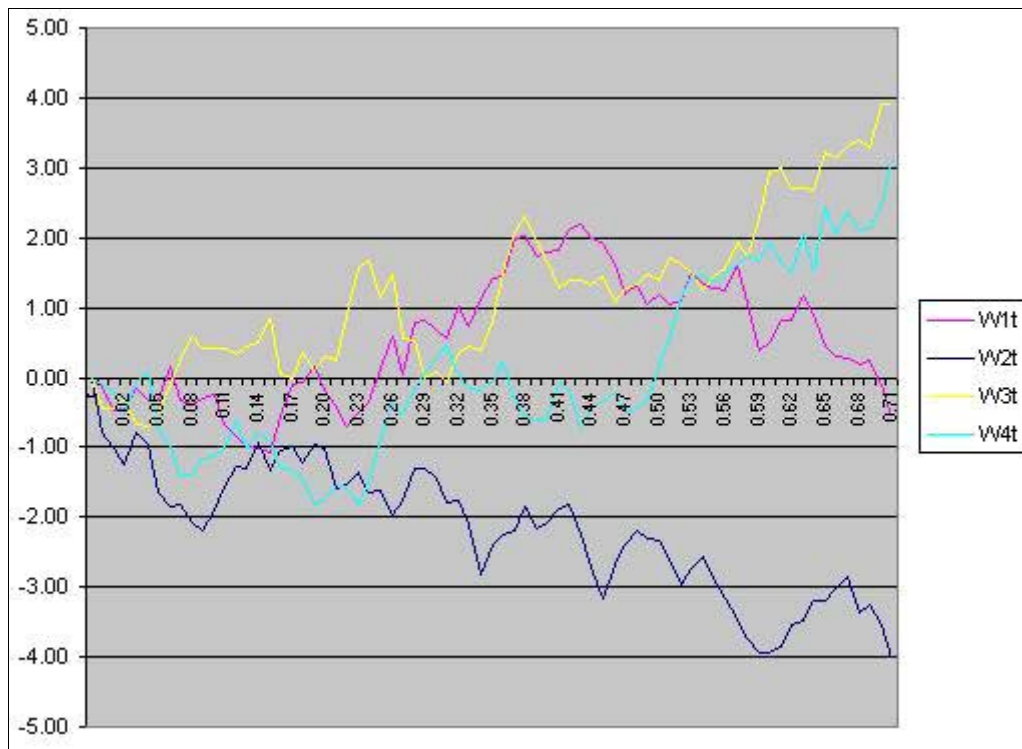


Figure: 66.51 - Exemple de 4 mouvements browniens standards

Les mouvements browniens standards ont certaines propriétés remarquables comme nous pouvons le voir: la trajectoire a tendance à alterner au-dessus et en dessous de l'axe des abscisses. Cela provient de ce que la loi Normale considérée est d'espérance nulle, autrement dit qu'il n'y a pas de tendance générale à la hausse ou à la baisse des variations (pour le vérifier faites au moins 30'000 points dans Microsoft Excel et vous verrez....).

Il est facilement possible de caractériser ΔZ à l'aide de son espérance:

$$E(\Delta Z) = E(\varepsilon \sqrt{\Delta t}) = \sqrt{\Delta t} \cdot E(\varepsilon) = \sqrt{\Delta t} \cdot 0 = 0 \quad (66.486)$$

effectivement, rappelons que pour la loi Normale centrée réduite, nous avons:

$$\mu = E(x) = 0 \quad (66.487)$$

donc nous pouvions nous attendre à ce résultat d'absence totale de tendance générale (c'était quasi-intuitif!).

Nous pouvons également caractériser ΔZ à l'aide de sa variance:

$$V(\Delta Z) = V(\varepsilon \sqrt{\Delta t}) = \Delta t \cdot V(\varepsilon) = \Delta t \cdot 1 = \Delta t \quad (66.488)$$

d'où:

$$\sigma_{\Delta Z} = \sqrt{\Delta t} \quad (66.489)$$

effectivement, rappelons que pour la loi Normale centrée réduite, nous avons:

$$\sigma_x^2 = V(x) = 1 \quad (66.490)$$

Finalement (au fait ce résultat découle de manière immédiate de la propriété de linéarité de la loi Normale):

$$\Delta x = \Delta z \rightarrow N(0, \sqrt{\Delta t}) \quad (66.491)$$

Donc lorsque nous avons:

$$\begin{aligned} E(\Delta z) &= 0 \\ E(\Delta z^2) &= E(\varepsilon \Delta t) = \Delta t E(\varepsilon) = \Delta t \end{aligned} \quad (66.492)$$

Il est d'usage de parler de "**processus de Gauss-Wiener**" ou plus simplement ou plus simplement de "**processus de Wiener**".

Pour résumer un peu les choses...

1. Nous savons avec le modèle de Bachelier que l'espérance positive et l'écart-type positif de la valeur sont proportionnels à la racine carrée du temps. Nous avons utilisé ces deux résultats ici.
2. Nous savons maintenant (sous l'hypothèse bien précise d'un coefficient de type Normal) que les variations ont une espérance (tendance) nulle et un écart-type proportionnel à racine carrée de la variation temporelle.

La propriété qui vient d'être établie reste valable pour un grand intervalle de temps noté T correspondant à n petits intervalles Δt !!! En d'autres termes:

$$T = n \cdot \Delta t \quad (66.493)$$

Dans ce contexte, il convient de remplacer Δx par:

$$x(T) - x(0) \quad (66.494)$$

Or:

$$x(T) - x(0) = \sum_{i=1}^n \Delta x_i = \sum_{i=1}^n \Delta z_i = \sum_{i=1}^n \varepsilon_i \sqrt{\Delta t} \quad (66.495)$$

Comme dans l'hypothèse d'une évolution du cours sur un petit intervalle de temps, il est possible de caractériser $x(T) - x(0)$ à l'aide de son espérance et de son écart-type:

$$\begin{aligned} E(x(T) - x(0)) &= E\left(\sum_{i=1}^n \varepsilon_i \sqrt{\Delta t}\right) = \sqrt{\Delta t} E\left(\sum_{i=1}^n \varepsilon_i\right) = \sqrt{\Delta t} \sum_{i=1}^n E(\varepsilon_i) = \sqrt{\Delta t} \cdot 0 = 0 \\ V(x(T) - x(0)) &= V\left(\sum_{i=1}^n \varepsilon_i \sqrt{\Delta t}\right) = \Delta t \cdot V\left(\sum_{i=1}^n \varepsilon_i\right) = \Delta t \sum_{i=1}^n V(\varepsilon_i) = \Delta t \sum_{i=1}^n 1 = n \cdot \Delta t = T \end{aligned} \quad (66.496)$$

ce qui est logique...

Nous retrouvons alors, pour un grand intervalle de temps T :

$$x(T) - x(0) \rightarrow N(0, \sqrt{T}) \quad (66.497)$$

que nous pouvons aussi écrire sous la forme suivante en utilisant les propriétés de la loi Normale:

$$x(T) \rightarrow N(x(0), \sqrt{T}) \quad (66.498)$$

résultat auquel nous pouvions raisonnablement nous attendre avec les hypothèses susmentionnées...

Ce dernier résultat est écrit sous la forme explicite suivante dans les tableurs:

$$x(t) = x(0) + N(0,1)\sqrt{t} \quad (66.499)$$

et nous voyons que c'est peu réaliste car cela signifierait que tout actif financier suit la même loi (quelle que soit sa volatilité....) et n'aurait aucune tendance générale à la baisse ou à la hausse. Nous verrons de suite comment améliorer cette approche.

Pour clore cette approche, remarquons que si Δt tend vers 0 (ce qui revient à considérer une subdivision du temps T en intervalles extrêmement petits) le cours subit sur la période T un nombre infiniment grand de variations. En d'autres termes, le processus d'évolution du cours de l'option est continu, ce qui conduit à remplacer Δt par dt , Δx par dx et Δz par dz .

Dans ce cas, nous obtenons:

$$dx \rightarrow N(0, \sqrt{dt}) \quad (66.500)$$

ce qui définit un "**processus de Wiener**" (nous reviendrons là-dessus lorsque nous aurons établi l'équation différentielle stochastique).

Mais évidemment ceci n'est pas vraiment conforme à la réalité comme nous l'avons déjà mentionné... Nous préférons alors ajouter un décalage constant dans le temps ce qui donne le mouvement brownien que nous allons voir maintenant.

7.8.4. MOUVEMENT BROWNIEN GÉNÉRALISÉ

Dans ce cas (généralisation un peu plus réaliste), l'évolution du cours dépend non seulement d'un processus aléatoire brownien standard (deuxième terme ci-dessous à droite de l'égalité), mais également d'un paramètre de tendance centrale, ou "**drift**" (premier terme ci-dessous à droite de l'égalité):

$$dx = a \cdot dt + b \cdot dz \quad (66.501)$$

avec toujours:

$$dz = \varepsilon \sqrt{dt} \quad (66.502)$$

et:

$$\varepsilon \rightarrow N(0,1) \quad (66.503)$$

Nous avons donc un mouvement brownien généralisé, constitué d'un mouvement brownien standard (dz représenté par une loi Normale d'espérance nulle et de variance dt comme nous l'avons vu plus haut) et d'un drift. Dans ce scénario, a et b sont imposés comme constants contrairement au cas encore plus général que nous verrons un peu plus loin.

La relation antérieure est souvent représentée dans la littérature sous la forme différentielle suivante:

$$dX_t = a dt + b dW_t = a dt + b d(N(0,1)\sqrt{dt}) \quad (66.504)$$

Donc graphiquement cela donne, en rajoutant ce drift et en prenant une valeur positive et non nulle pour a , un mouvement brownien standard qui aura tendance à alterner au-dessus et en-dessous du drift:

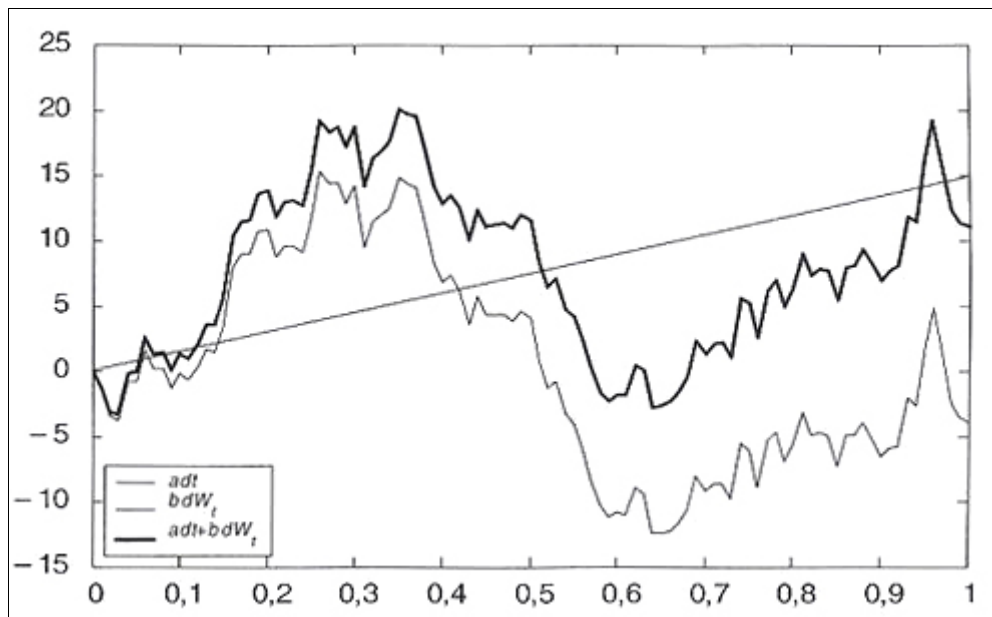


Figure: 66.52 - Exemple de mouvement brownien standard avec drift (et décomposition)

Sur un petit intervalle de temps Δt , le processus, en temps discret s'écrit bien évidemment:

$$\Delta x = a \cdot \Delta t + b \cdot \Delta z \quad (66.505)$$

Dans ce cas, nous avons:

$$E(\Delta x) = E(a \Delta t + b \Delta z) = a \cdot \Delta t + b \cdot E(\Delta z) \quad (66.506)$$

dans la mesure où seule Δz a une composante aléatoire.

Ainsi:

$$\begin{aligned} E(\Delta x) &= a \cdot \Delta t + b \cdot E(\Delta z) = a \cdot \Delta t + b \cdot E(\varepsilon \sqrt{\Delta t}) = a \Delta t + b \sqrt{\Delta t} \cdot E(\varepsilon) = a \cdot \Delta t + 0 = a \cdot \Delta t \\ V(\Delta x) &= V(a \cdot \Delta t + b \cdot \Delta z) = 0 + b^2 \cdot V(\varepsilon \sqrt{\Delta t}) = b^2 \Delta t \cdot V(\varepsilon) = b^2 \cdot \Delta t \end{aligned} \quad (66.507)$$

Finalement:

$$\Delta x \rightarrow N(a \Delta t, b \sqrt{\Delta t}) \quad (66.508)$$

En subdivisant une période T en n intervalles de temps Δt (soit $T = n \cdot \Delta t$), la variation du cours devient sur cette période T :

$$x(T) - x(0) = \sum_{i=1}^n \Delta x_i \quad (66.509)$$

Dès lors:

$$\begin{aligned} E(x(T) - x(0)) &= E\left(\sum_{i=1}^n \Delta x_i\right) = \sum_{i=1}^n E(\Delta x_i) = n \cdot a \cdot \Delta t = a \cdot T \\ V(x(T) - x(0)) &= V\left(\sum_{i=1}^n \Delta x_i\right) = \sum_{i=1}^n V(\Delta x_i) = \sum_{i=1}^n V(b\sqrt{\Delta t}) = b^2 \Delta t \sum_{i=1}^n 1 = n \cdot b^2 \Delta t = b^2 T \end{aligned} \quad (66.510)$$

Finalement:

$$x(T) - x(0) = N(a \cdot T, b \cdot \sqrt{T}) \quad (66.511)$$

Soit:

$$x(T) = x(0) + N(a \cdot T, b \cdot \sqrt{T}) \quad (66.512)$$

ou encore:

$$x(T) = N(x(0) + a \cdot T, b \cdot \sqrt{T}) \quad (66.513)$$

Il est alors aisé de comprendre pourquoi nous disons que la loi Normale régit la variable aléatoire obtenue en arrêtant un processus brownien à un instant donné: c'est une photo instantanée du mouvement brownien simple ou généralisé!

En choisissant:

$$a = x(0) \cdot \mu, b = \sigma x(0) \quad (66.514)$$

Nous avons alors la relation antérieure qui s'écrit traditionnellement sous la forme explicite suivante dans les tableurs:

$$x(t) = x(0) + \mu \cdot x(0) \cdot t + N(0,1) \sigma x(0) \sqrt{t} \quad (66.515)$$

où μ est le rendement en % de l'actif financier et σ la volatilité du rendement en %. Soit écrit autrement:

$$x(t) = N(x(0) + \mu x(0)t, \sigma x(0)\sqrt{t}) \quad (66.516)$$

Il est alors intéressant pour le financier de visualiser l'espérance en fonction de t et la valeur $x(t)$ correspondante à une probabilité cumulée de 2.5% et de 97.5% sur un graphique pour avoir une idée de l'évolution de l'intervalle de confiance à 95% de son $x(0)$. Ceci est très facile à obtenir dans Microsoft Excel et l'on tombe typiquement en jouant avec plusieurs types de graphiques dans le même diagramme sur un résultat comme celui visible ci-dessous pour un portefeuille de 500 MF avec rendement de 5% et écart-type de 20%:

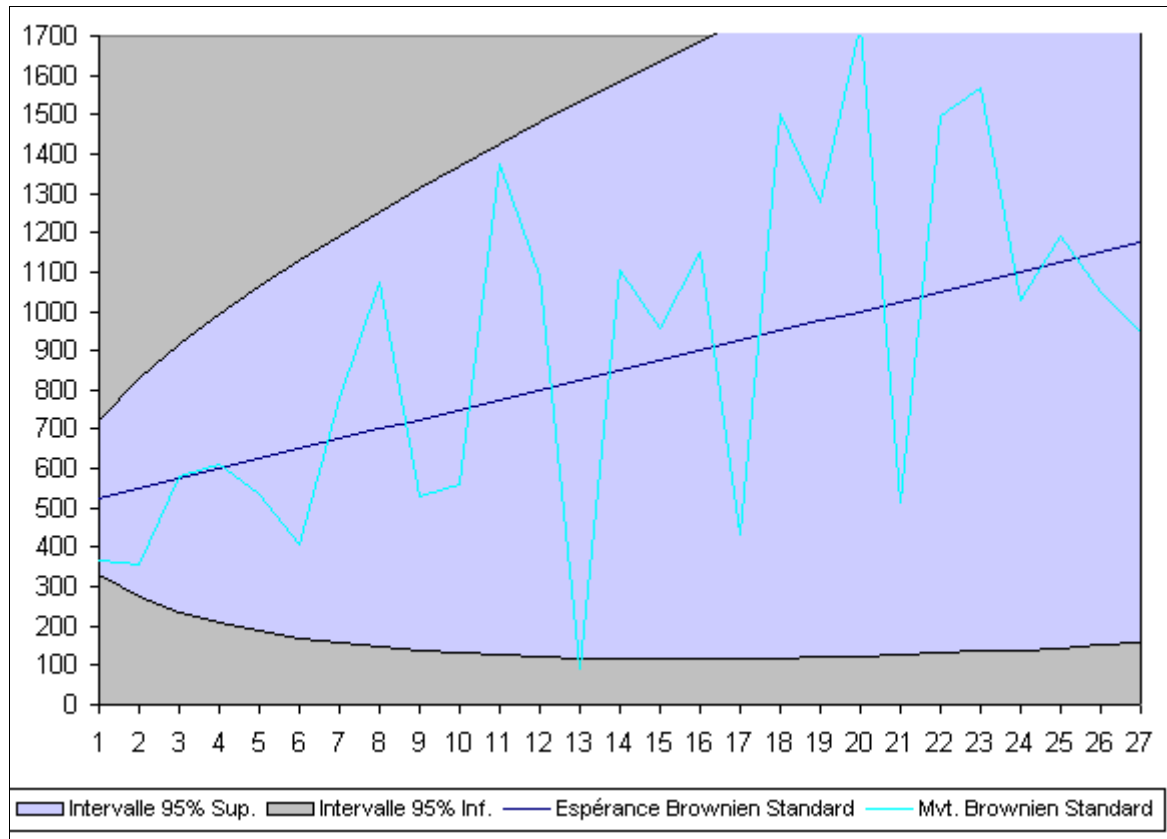


Figure: 66.53 - Tracé du mouvement brownien standard avec intervalle de confiance

Avec le tableau suivant:

	A	B	C	D	E	F
1	Mois	Espérance Brownien Standard	Mouv. Brow Standard	Ecart-type Standard	Intervalle 95%	Intervalle 95% Sup.
2	0	500.00	500.00	0.00		
3	1	525.00	336.41	100.00	329.00	721.00
4	2	550.00	620.67	141.42	272.82	827.18
5	3	575.00	842.16	173.21	235.52	914.48
6	4	600.00	565.28	200.00	208.01	991.99
7	5	625.00	701.39	223.61	186.74	1063.26
8	6	650.00	550.01	244.95	169.91	1130.09
9	7	675.00	684.45	264.58	156.44	1193.56
10	8	700.00	725.39	282.84	145.64	1254.36
11	9	725.00	1040.65	300.00	137.01	1312.99
12	10	750.00	763.32	316.23	130.20	1369.80
13	11	775.00	724.46	331.66	124.95	1425.05
14	12	800.00	703.70	346.41	121.05	1478.95
15	13	825.00	1288.31	360.56	118.32	1531.68
16	14	850.00	1283.01	374.17	116.65	1583.35
17	15	875.00	641.06	387.30	115.91	1634.09
18	16	900.00	1641.97	400.00	116.01	1683.99
19	17	925.00	1499.09	412.31	116.89	1733.11
20	18	950.00	763.09	424.26	118.46	1781.54
21	19	975.00	662.14	435.89	120.67	1829.33
22	20	1000.00	1180.59	447.21	123.48	1876.52
23	21	1025.00	1086.71	458.26	126.83	1923.17
24	22	1050.00	1076.45	469.04	130.70	1969.30
25	23	1075.00	420.13	479.58	135.03	2014.97
26	24	1100.00	1144.09	489.90	139.82	2060.18
27	25	1125.00	996.50	500.00	145.02	2104.98
28	26	1150.00	1564.97	509.90	150.61	2149.39
29	27	1175.00	1698.88	519.62	156.57	2193.43

Figure: 66.54 - Tableau des données correspondantes au tracé précédent

Utilisant donc les relations démontrées plus haut:

	A	B	C	D	E	F
	Mois	Espérance Brownien Standard	Mouv. Brow Standard	Ecart-type Standard	Intervalle 95% Inf.	Intervalle 95% Sup.
2	0	=500+500*5%*A2	=B\$2+5%*B\$2*A2+NORMSINV(RAND())*20%*B\$2*SQRT(A2)	=20%*500*SQRT(A2)		
3	1	=500+500*5%*A3	=B\$2+5%*B\$2*A3+NORMSINV(RAND())*20%*B\$2*SQRT(A3)	=20%*500*SQRT(A3)	=NORMINV(2.5%,B3,D3)	=NORMINV(97.5%,B3,D3)
4	2	=500+500*5%*A4	=B\$2+5%*B\$2*A4+NORMSINV(RAND())*20%*B\$2*SQRT(A4)	=20%*500*SQRT(A4)	=NORMINV(2.5%,B4,D4)	=NORMINV(97.5%,B4,D4)
5	3	=500+500*5%*A5	=B\$2+5%*B\$2*A5+NORMSINV(RAND())*20%*B\$2*SQRT(A5)	=20%*500*SQRT(A5)	=NORMINV(2.5%,B5,D5)	=NORMINV(97.5%,B5,D5)
6	4	=500+500*5%*A6	=B\$2+5%*B\$2*A6+NORMSINV(RAND())*20%*B\$2*SQRT(A6)	=20%*500*SQRT(A6)	=NORMINV(2.5%,B6,D6)	=NORMINV(97.5%,B6,D6)
7	5	=500+500*5%*A7	=B\$2+5%*B\$2*A7+NORMSINV(RAND())*20%*B\$2*SQRT(A7)	=20%*500*SQRT(A7)	=NORMINV(2.5%,B7,D7)	=NORMINV(97.5%,B7,D7)
8	6	=500+500*5%*A8	=B\$2+5%*B\$2*A8+NORMSINV(RAND())*20%*B\$2*SQRT(A8)	=20%*500*SQRT(A8)	=NORMINV(2.5%,B8,D8)	=NORMINV(97.5%,B8,D8)
9	7	=500+500*5%*A9	=B\$2+5%*B\$2*A9+NORMSINV(RAND())*20%*B\$2*SQRT(A9)	=20%*500*SQRT(A9)	=NORMINV(2.5%,B9,D9)	=NORMINV(97.5%,B9,D9)
10	8	=500+500*5%*A10	=B\$2+5%*B\$2*A10+NORMSINV(RAND())*20%*B\$2*SQRT(A10)	=20%*500*SQRT(A10)	=NORMINV(2.5%,B10,D10)	=NORMINV(97.5%,B10,D10)
11	9	=500+500*5%*A11	=B\$2+5%*B\$2*A11+NORMSINV(RAND())*20%*B\$2*SQRT(A11)	=20%*500*SQRT(A11)	=NORMINV(2.5%,B11,D11)	=NORMINV(97.5%,B11,D11)
12	10	=500+500*5%*A12	=B\$2+5%*B\$2*A12+NORMSINV(RAND())*20%*B\$2*SQRT(A12)	=20%*500*SQRT(A12)	=NORMINV(2.5%,B12,D12)	=NORMINV(97.5%,B12,D12)
13	11	=500+500*5%*A13	=B\$2+5%*B\$2*A13+NORMSINV(RAND())*20%*B\$2*SQRT(A13)	=20%*500*SQRT(A13)	=NORMINV(2.5%,B13,D13)	=NORMINV(97.5%,B13,D13)
14	12	=500+500*5%*A14	=B\$2+5%*B\$2*A14+NORMSINV(RAND())*20%*B\$2*SQRT(A14)	=20%*500*SQRT(A14)	=NORMINV(2.5%,B14,D14)	=NORMINV(97.5%,B14,D14)
15	13	=500+500*5%*A15	=B\$2+5%*B\$2*A15+NORMSINV(RAND())*20%*B\$2*SQRT(A15)	=20%*500*SQRT(A15)	=NORMINV(2.5%,B15,D15)	=NORMINV(97.5%,B15,D15)
16	14	=500+500*5%*A16	=B\$2+5%*B\$2*A16+NORMSINV(RAND())*20%*B\$2*SQRT(A16)	=20%*500*SQRT(A16)	=NORMINV(2.5%,B16,D16)	=NORMINV(97.5%,B16,D16)
17	15	=500+500*5%*A17	=B\$2+5%*B\$2*A17+NORMSINV(RAND())*20%*B\$2*SQRT(A17)	=20%*500*SQRT(A17)	=NORMINV(2.5%,B17,D17)	=NORMINV(97.5%,B17,D17)
18	16	=500+500*5%*A18	=B\$2+5%*B\$2*A18+NORMSINV(RAND())*20%*B\$2*SQRT(A18)	=20%*500*SQRT(A18)	=NORMINV(2.5%,B18,D18)	=NORMINV(97.5%,B18,D18)
19	17	=500+500*5%*A19	=B\$2+5%*B\$2*A19+NORMSINV(RAND())*20%*B\$2*SQRT(A19)	=20%*500*SQRT(A19)	=NORMINV(2.5%,B19,D19)	=NORMINV(97.5%,B19,D19)
20	18	=500+500*5%*A20	=B\$2+5%*B\$2*A20+NORMSINV(RAND())*20%*B\$2*SQRT(A20)	=20%*500*SQRT(A20)	=NORMINV(2.5%,B20,D20)	=NORMINV(97.5%,B20,D20)
21	19	=500+500*5%*A21	=B\$2+5%*B\$2*A21+NORMSINV(RAND())*20%*B\$2*SQRT(A21)	=20%*500*SQRT(A21)	=NORMINV(2.5%,B21,D21)	=NORMINV(97.5%,B21,D21)
22	20	=500+500*5%*A22	=B\$2+5%*B\$2*A22+NORMSINV(RAND())*20%*B\$2*SQRT(A22)	=20%*500*SQRT(A22)	=NORMINV(2.5%,B22,D22)	=NORMINV(97.5%,B22,D22)
23	21	=500+500*5%*A23	=B\$2+5%*B\$2*A23+NORMSINV(RAND())*20%*B\$2*SQRT(A23)	=20%*500*SQRT(A23)	=NORMINV(2.5%,B23,D23)	=NORMINV(97.5%,B23,D23)
24	22	=500+500*5%*A24	=B\$2+5%*B\$2*A24+NORMSINV(RAND())*20%*B\$2*SQRT(A24)	=20%*500*SQRT(A24)	=NORMINV(2.5%,B24,D24)	=NORMINV(97.5%,B24,D24)
25	23	=500+500*5%*A25	=B\$2+5%*B\$2*A25+NORMSINV(RAND())*20%*B\$2*SQRT(A25)	=20%*500*SQRT(A25)	=NORMINV(2.5%,B25,D25)	=NORMINV(97.5%,B25,D25)
26	24	=500+500*5%*A26	=B\$2+5%*B\$2*A26+NORMSINV(RAND())*20%*B\$2*SQRT(A26)	=20%*500*SQRT(A26)	=NORMINV(2.5%,B26,D26)	=NORMINV(97.5%,B26,D26)
27	25	=500+500*5%*A27	=B\$2+5%*B\$2*A27+NORMSINV(RAND())*20%*B\$2*SQRT(A27)	=20%*500*SQRT(A27)	=NORMINV(2.5%,B27,D27)	=NORMINV(97.5%,B27,D27)
28	26	=500+500*5%*A28	=B\$2+5%*B\$2*A28+NORMSINV(RAND())*20%*B\$2*SQRT(A28)	=20%*500*SQRT(A28)	=NORMINV(2.5%,B28,D28)	=NORMINV(97.5%,B28,D28)
29	27	=500+500*5%*A29	=B\$2+5%*B\$2*A29+NORMSINV(RAND())*20%*B\$2*SQRT(A29)	=20%*500*SQRT(A29)	=NORMINV(2.5%,B29,D29)	=NORMINV(97.5%,B29,D29)

Figure: 66.55 - Formules explicites pour le mouvement brownien et l'intervalle de confiance

Évidemment, dans la pratique il est possible de faire ce type de graphique avec n'importe quelle donnée comportant un drift linéaire et dont l'écart-type est connu.

7.8.5. PONT BROWNIEN

Voyons maintenant un cas d'application du mouvement brownien qui va nous être utile plus tard! En refaisant les développements vus plus haut, il vient immédiatement que:

$$x_t = x_{t-\Delta t} + \mu x_0 \Delta t + N(0,1) \sigma x_0 \sqrt{\Delta t} \quad (66.517)$$

Effectivement, vérifions cela en choisissant:

$$\Delta t = 1 \quad (66.518)$$

Nous avons:

$$x_t = x_{t-1} + \mu \cdot x_0 \cdot 1 + N(0,1)\sigma x_0 \sqrt{1} = x_{t-1} + \mu x_0 + N(0,1)\sigma x_0 \quad (66.519)$$

et donc pour $t = 1$:

$$x_1 = x_0 + \mu x_0 + N(0,1)\sigma x_0 \quad (66.520)$$

et de même pour $t = 2$:

$$\begin{aligned} x_2 &= x_1 + \mu \cdot x_0 + N(0,1)\sigma x_0 = (x_0 + \mu \cdot x_0 + N(0,1)\sigma x_0) + \mu \cdot x_0 + N(0,1)\sigma x_0 \\ &= x_0 + 2\mu \cdot x_0 + \underbrace{(N(0,1) + N(0,1))\sigma x_0}_{=N(0,\sqrt{2})} = x_0 + 2\mu \cdot x_0 + N(0,1)\sqrt{2}\sigma x_0 \end{aligned}$$

Ce qui est donc bien équivalent à:

$$x(t) = x(0) + \mu x(0)\Delta t + N(0,1)\sigma x(0)\sqrt{\Delta t} \quad (66.521)$$

pour $\Delta t = 2$.

Bref, ceci étant dit, nous souhaiterions construire un mouvement brownien qui partant au temps $t = 0$ de zéro (ce qui impose $x(0) = 0$) fluctue pendant une période T donnée et revient à zéro comme le montre les deux figures ci-dessous:

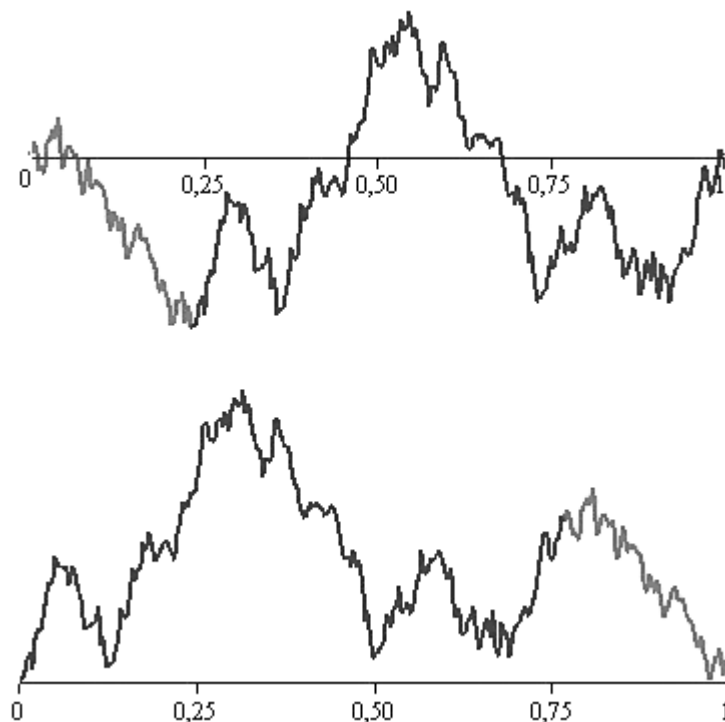


Figure: 66.56 - Deux exemples de ponts browniens

Comme il est d'usage, choisissons la notation suivante:

$$W(t) = W_{t-\Delta t} + \mu W_0 \Delta t + N(0,1) \sigma W_0 \sqrt{\Delta t} \quad (66.522)$$

Si nous voulons que le mouvement brownien s'annule au moment T , il suffit de faire une simple soustraction du type suivant:

$$BB(t) = W(t) - W(T) \quad (66.523)$$

où BB signifie "**Brownian Bridge**" ou "**pont brownien**" en français. Si nous nous arrêtons ici, nous avons bien:

$$BB(T) = W(T) - W(T) = 0 \quad (66.524)$$

mais nous avons un problème au temps $t = 0$:

$$BB(0) = W(0) - W(T) \neq 0 \quad (66.525)$$

et pour s'assurer que la différence est bien nulle, l'astuce naturelle consiste à écrire:

$$BB(t) = W(t) - tW(T) \quad (66.526)$$

Ainsi, nous avons:

$$BB(0) = W(0) - 0 \cdot W(T) = 0 \quad (66.527)$$

Mais dès lors, vient un autre problème. Effectivement:

$$BB(T) = W(T) - T \cdot W(T) \neq 0 \quad (66.528)$$

Donc, la solution finale consiste à écrire:

$$BB(t) = W(t) - \frac{t}{T} \cdot W(T) \quad (66.529)$$

Et donc:

$$BB(t) = BB(T) = 0 \quad (66.530)$$

Et nous avons:

$$E(BB(t)) = E\left(W(t) - \frac{t}{T} \cdot W(T)\right) = 0 \quad (66.531)$$

Le pont brownien:

$$BB(t) = W(t) - \frac{t}{T} \cdot W(T) \quad (66.532)$$

fonctionne pour tous les processus de Wiener, raison pour laquelle vous verrez parfois dans la littérature tantôt des ponts avec:

$$W_t = W_{t-\Delta t} + \mu W_0 \Delta t + N(0,1) \sigma W_0 \sqrt{\Delta t} \quad (66.533)$$

ou par exemple avec:

$$W_t = W_{t-\Delta t} + \mu \Delta t + N(0,1) \sigma \sqrt{\Delta t} \quad (66.534)$$

qui n'est qu'un cas particulier du précédent avec $W_0 = 1$.

7.8.6. PROCESSUS D'ITÔ

Considérons maintenant un processus brownien correspondant à une variation de x en temps continu définie par:

$$dx = a(x,t)dt + b(x,t)dz \quad (66.535)$$

a et b étant alors des fonctions des 2 variables x et t . Cette considération est ce que nous appelons un "processus d'Itô". Il s'agit donc d'une généralisation du cas précédent où a et b ne sont plus constants.

Il est possible de calculer l'espérance et la variance de dx exactement de la même façon que pour le processus de Wiener et nous obtenons très facilement par analogie:

$$E(dx) = 0 \quad V(dx) = b^2(x,t)dt \quad (66.536)$$

Par conséquent, nous pouvons écrire:

$$dx \rightarrow N(a(x,t)dt, b(x,t)\sqrt{dt}) \quad (66.537)$$

où $a(x,t)$ correspond au drift instantané et $b(x,t)$ à la variance instantanée.

Le "mouvement brownien géométrique" qui permet de définir théoriquement la meilleure prédiction d'évolution du rendement d'une option est un cas particulier de processus d'Itô (parmi tant d'autres modèles...) où nous supposons que:

$$a(x,t) = \mu \cdot x \quad \text{et} \quad b(x,t) = \sigma \cdot x \quad (66.538)$$

Dès lors nous pouvons écrire l'expression du mouvement brownien géométrique de la valeur de l'option notée:

$$dF = \mu x dt + \sigma x dz \quad (66.539)$$

souvent représentée dans la littérature aussi sous la forme suivante:

$$dX_t = \mu X_t dt + \sigma X_t dW_t \quad (66.540)$$

ou encore plus explicitement:

$$dX_t = (\mu X_t) dt + (\sigma X_t) dW_t = (\mu X_t) dt + (\sigma X_t) d(N(0,1)\sqrt{dt}) \quad (66.541)$$

L'interprétation financière de la relation:

$$dF = \mu x dt + \sigma x dz \quad (66.542)$$

devient apparente lorsque nous divisons les deux membres par x :

$$\frac{dF}{x} = \frac{F(t+dt) - F(t)}{x(t)} = \mu dt + \sigma dz$$

ce qui correspond au taux de rentabilité de l'option (ou tout autre actif de la même famille) sur une période infinitésimale dt .

Le mouvement brownien géométrique est donc a priori un bon candidat pour modéliser l'évolution du prix d'un actif financier à partir de son taux de rentabilité.

Dans la littérature spécialisée, le return (rendement) est aussi parfois noté (notation justifiée) sous la forme de l'équation différentielle stochastique (E.D.S.) suivante:

$$\boxed{\frac{dS_t}{S_t} = \mu dt + \sigma dz} \quad (66.543)$$

où S_t est bien évidemment le prix du sous-jacent appelé "**stock price**" au temps t , μ est appelé la "**dérive**" (assimilé souvent au rendement) et σ la "**volatilité**" (la volatilité du rendement). C'est la notation et le vocabulaire que nous adopterons pour la suite.

À noter que puisque nous avons:

$$\frac{d \ln(S_t)}{dS_t} = \frac{1}{S_t} \quad (66.544)$$

Nous pouvons donc aussi écrire:

$$d \ln(S_t) = \mu dt + \sigma dz \quad (66.545)$$

Au cas où $\sigma = 0$ (processus de Wiener, autrement dit le prix du sous-jacent est parfaitement connu à un temps donné et sans risques), nous nous retrouvons avec une équation différentielle (connue dans le domaine) que nous pouvons de suite résoudre:

$$\frac{dS_t}{S_t} = \mu dt \Leftrightarrow \frac{dS_t}{dt} = \mu S_t \Rightarrow S_t = S_0 e^{\mu(t-t_0)} \quad (66.546)$$

Il s'agit donc d'une exponentielle (comme l'intérêt continu que nous avons vu au début de ce chapitre). Cette relation n'étant valable que si l'intervalle de temps est donc très petit.

Nous allons voir maintenant à l'aide du "**lemme d'Itô**", qu'il est possible (ce qui n'est pas une possibilité unique!) d'établir qu'un tel processus peut définir une loi log-normale (cf. chapitre de Statistiques).

Le lemme d'Itô est établi à partir du développement de Taylor (cf. chapitre de Suites et Séries) à 2 variables x et t donné par :

$$\Delta F = \frac{1}{1!} \frac{\partial F}{\partial t} \Delta t + \frac{1}{1!} \frac{\partial F}{\partial x} \Delta x + \frac{1}{2!} \frac{\partial^2 F}{\partial t^2} \Delta t^2 + \frac{1}{2!} \frac{\partial^2 F}{\partial x^2} \Delta x^2 + \dots \quad (66.547)$$

avec $F(0) = 0$ à l'origine du mouvement brownien.

En considérant $\Delta t \ll \Delta x$, et en ne prenant les termes que jusqu'au deuxième ordre (approximation formelle périlleuse mais numériquement non obligatoire grâce à la puissance de calcul des ordinateurs), nous avons:

$$\Delta F \cong \frac{1}{1!} \frac{\partial F}{\partial t} \Delta t + \frac{1}{1!} \frac{\partial F}{\partial x} \Delta x + \frac{1}{2!} \frac{\partial^2 F}{\partial x^2} \Delta x^2 \quad (66.548)$$

Revenons maintenant à:

$$\Delta x = a \cdot \Delta t + b \cdot \Delta z = a \Delta t + b \varepsilon \sqrt{\Delta t} \quad (66.549)$$

Élevons au carré, nous obtenons:

$$\Delta x^2 = (a \Delta t + b \varepsilon \sqrt{\Delta t})^2 = a^2 \Delta t^2 + 2 \varepsilon a b \Delta t^{3/2} + b^2 \varepsilon^2 \Delta t \cong b^2 \varepsilon^2 \Delta t \quad (66.550)$$

Or:

$$E(b^2 \varepsilon^2 \Delta t) = b^2 \Delta t E(\varepsilon^2) \quad (66.551)$$

et comme nous l'avons démontré dans le chapitre de Statistiques:

$$V(X) = E(X^2) - E(X)^2 \Leftrightarrow E(X^2) = V(X) + E(X)^2 \quad (66.552)$$

Nous avons alors:

$$E(\varepsilon^2) = V(\varepsilon) + E(\varepsilon)^2 = 1 + 0 = 1 \quad (66.553)$$

Donc:

$$E(b^2 \varepsilon^2 \Delta t) = b^2 \Delta t \quad (66.554)$$

Par ailleurs:

$$V(b^2 \varepsilon^2 \Delta t) = b^4 \Delta t^2 V(\varepsilon^2) = b^4 \Delta t^2 \quad (66.555)$$

qui tendent tous deux vers 0 quand Δt tend vers 0.

Nous avons aussi en considérant une subdivision du temps en intervalles Δt extrêmement petits (approximation qui va nous être utile un peu plus loin):

$$\lim_{\Delta t \rightarrow 0} b^2 \varepsilon^2 \Delta t \cong \lim_{\Delta t \rightarrow 0} b^2 \Delta t = 0 \quad (66.556)$$

ce qui implique dans ce dernier cas de figure:

$$b^2 \varepsilon^2 \Delta t \cong b^2 \Delta t \quad (66.557)$$

donc en se plaçant en temps continu (donc un modèle continu avec intervalles de temps extrêmement petits), l'application du développement de Taylor discrète vue plus haut:

$$\Delta F \cong \frac{1}{1!} \frac{\partial F}{\partial t} \Delta t + \frac{1}{1!} \frac{\partial F}{\partial x} \Delta x + \frac{1}{2!} \frac{\partial^2 F}{\partial x^2} \Delta x^2 \quad (66.558)$$

peut alors s'écrire:

$$\begin{aligned} dF &\cong \frac{\partial F}{\partial t} dt + \frac{\partial F}{\partial x} dx + \frac{1}{2} \frac{\partial^2 F}{\partial x^2} b^2 dt = \frac{\partial F}{\partial t} dt + \frac{\partial F}{\partial x} (adt + b dz) + \frac{1}{2} \frac{\partial^2 F}{\partial x^2} b^2 dt \\ &= \left(\frac{\partial F}{\partial t} + a \frac{\partial F}{\partial x} + \frac{1}{2} b^2 \frac{\partial^2 F}{\partial x^2} \right) dt + b \frac{\partial F}{\partial x} dz \end{aligned} \quad (66.559)$$

il s'agit du lemme d'Itô également appelé "**théorème d'Itô-Doeblin**".

Remarque: Comparer la forme de la dernière égalité à la relation $dx = a(x,t)dt + b(x,t)dz$

Si nous prenons:

$$F(x) = \ln(x) \quad (66.560)$$

Dès lors:

$$\frac{\partial F(x)}{\partial t} = 0, \quad \frac{\partial F(x)}{\partial x} = \frac{1}{x}, \quad \frac{\partial^2 F(x)}{\partial x^2} = -\frac{1}{x^2} \quad (66.561)$$

Dans ce cas:

$$dF = \left(0 + \frac{a}{x} - \frac{1}{2} b^2 \frac{1}{x^2} \right) dt + b \frac{1}{x} dz \quad (66.562)$$

En revenant à l'hypothèse de mouvement brownien géométrique, nous savons que nous devons considérer que:

$$a = a(x,t) = \mu x \text{ et } b = b(x,t) = \sigma x \quad (66.563)$$

Nous avons donc:

$$dF = \left(0 + \frac{\mu x}{x} - \frac{1}{2} \frac{\sigma^2 x^2}{x^2} \right) dt + \sigma x \frac{1}{x} dz \quad (66.564)$$

et nous obtenons finalement l'équation différentielle stochastique à coefficients constants:

$$dF = \left(\mu - \frac{\sigma^2}{2} \right) dt + \sigma dz \quad (66.565)$$

Soit en reprenant la notation du début sous forme explicite:

$$d \ln(x) = \left(\mu - \frac{\sigma^2}{2} \right) dt + \sigma dz \quad (66.566)$$

ou sous une autre forme encore plus explicite:

$$d \ln(X_t) = \left(\mu - \frac{\sigma^2}{2} \right) dt + \sigma dW_t = \left(\mu - \frac{\sigma^2}{2} \right) dt + \sigma d \left(N(0,1) \sqrt{dt} \right) \quad (66.567)$$

appelée parfois "**équation intégrale de la dynamique des prix**".

Une approche triviale de la résolution de cette intégrale est (on considère toujours dans ce modèle que la volatilité est indépendante du temps...):

$$\int_t^{t+1} d \ln(X_t) = \int_t^{t+1} \left(\mu - \frac{\sigma^2}{2} \right) dt + \int_t^{t+1} \sigma dW_t = \left(\mu - \frac{\sigma^2}{2} \right) \int_t^{t+1} dt + \sigma \int_t^{t+1} dW_t \quad (66.568)$$

Soit:

$$\ln(X_{t+1}) - \ln(X_t) = \left(\mu - \frac{\sigma^2}{2} \right) ((t+1) - t) + \sigma (W_{t+1} - W_t) \quad (66.569)$$

Soit:

$$\ln \left(\frac{X_{t+1}}{X_t} \right) = \mu - \frac{\sigma^2}{2} + \sigma (W_{t+1} - W_t) = \mu - \frac{\sigma^2}{2} + \sigma N(0,1) \quad (66.570)$$

Remarques:

R1. Se rappeler que nous sommes partis de la relation $dF = \mu x dt + \sigma x dz$

R2. Les mouvements browniens ont été successivement dégagés de l'hypothèse de Normalité dans les années 1960, puis de l'hypothèse de stabilité dans les années 1980. Avec ces deux hypothèses, les mathématiciens les rangent dans la catégorie particulière des "**processus de Lévy 2-stables**".

dF définit alors un mouvement brownien géométrique avec drift particulier dont nous pouvons maintenant mesurer les paramètres (c'est ce que nous voulions obtenir). Par conséquent, les résultats que nous avons obtenus pour le mouvement brownien peuvent être récupérés et nous permettent d'écrire au final:

$$dF \rightarrow N\left(\left(\mu - \frac{\sigma^2}{2}\right)dt, \sigma\sqrt{dt}\right) \quad (66.571)$$

ce qui revient dire que dF suit une loi log-normale (cf. chapitre de Statistiques) de paramètres:

$$\left(\mu - \frac{\sigma^2}{2}\right)dt \text{ et } \sigma\sqrt{dt} \quad (66.572)$$

Allons maintenant un peu plus loin en intégrant l'élément différentiel. Nous avons donc:

$$\frac{dX_t}{X_t} = \left(\mu - \frac{\sigma^2}{2}\right)dt + \sigma d\left(N(0,1)\sqrt{dt}\right) \quad (66.573)$$

Intégrons cette dernière relation:

$$\int \frac{dX_t}{X_t} = \int \left(\mu - \frac{\sigma^2}{2}\right)dt + \int \sigma d\left(N(0,1)\sqrt{dt}\right) \quad (66.574)$$

La première primitive est simple:

$$\int \frac{dX_t}{X_t} = \ln(X_t) + c^{te} \quad (66.575)$$

La deuxième primitive est simple (pas de constante d'intégration car au temps zéro l'espérance de gain est nulle):

$$\int \left(\mu - \frac{\sigma^2}{2}\right)dt = \left(\mu - \frac{\sigma^2}{2}\right) \int dt = \left(\mu - \frac{\sigma^2}{2}\right)t \quad (66.576)$$

La troisième primitive vaut (pas de constante d'intégration car au temps zéro la valeur du gain est parfaitement connue comme valant 0):

$$\int \sigma d\left(N(0,1)\sqrt{dt}\right) = \sigma N(0,1) \int d\sqrt{dt} = \sigma N(0,1)\sqrt{t} \quad (66.577)$$

Remarque: Attention à l'abus d'écriture!!! Dans la racine il s'agit implicitement d'un dt (du pas d'analyse donc!) et non simplement d'un t . Souvenez-vous en pour la suite!!!

Nous avons donc:

$$\ln(X_t) = \left(\mu - \frac{\sigma^2}{2}\right)t + \sigma N(0,1)\sqrt{t} - c^{te} \quad (66.578)$$

Et au final (nous mettons le signe "-" dans la constante):

$$X_t = e^{\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma N(0,1)\sqrt{t}} = e^{e^{\mu}} e^{\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma N(0,1)\sqrt{t}} \quad (66.579)$$

Pour trouver la signification du premier facteur, il suffit de poser la condition initiale:

$$X_t(t=0) = X_0 \quad (66.580)$$

Nous avons alors immédiatement pour l'expression finale du brownien géométrique:

$$X_t = X_0 e^{\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma N(0,1)\sqrt{t}} \quad (66.581)$$

obtenue par P. Samuelson en 1965 et qui est parfois appelée "**modèle de Bachelier-Samuelson**" ou "**représentation log-normale de la valeur de l'option**". Ainsi, pour pricer un actif suivant ce modèle sur un horizon, t nous pouvons utiliser la fonction VBA de Monte-Carlo suivante:

```
Function PricingStock(RiskFree, StDv, StockValue, Time, Confidence)

Dim i As Integer
Dim arrPricingStock(1 To 10000) As Double
For i = 1 To 10000
    arrPricingStock(i) = StockValue * Exp((RiskFree - 0.5 * StDv ^ 2) _
        * Time + StDv * Application.NormInv(Rnd(), 0, 1))
Next i
PricingStock = Application.Percentile(arrPricingStock, 1 - Confidence)

End Function
```

Nous avons au final une formulation (sous forme de fonction de distribution probabiliste) d'une variation temporelle et du return intrinsèque d'une action qui peut être utilisée à des fins décisionnelles d'investissements sur une prévision. Mais ce modèle est quand même trop lisse en n'arrive pas à modéliser les krachs boursiers (il en est de même pour rappel avec le mouvement brownien standard) pouvant arriver sur le long terme. Raison pour laquelle certains modèles plus récents que nous n'étudierons pas ici ajoutent un processus de Poisson (discret et à événements rares par construction) à celui de Wiener.

Il existe d'autres modèles que la log-normale mais celle-ci de par sa facilité est la plus répandue. Il faut cependant encourager d'autres méthodes plus généralistes!

Pour terminer cette partie résumons par une comparaison le mouvement brownien standard et le mouvement brownien géométrique qui régissent donc la dynamique des cours lorsque les paramètres (rendement et volatilité instantanés) sont donnés en %:

$$\begin{aligned} X_t &= X_0 + X_0 \mu t + N(0,1) \sigma X_0 \sqrt{t} = X_0 + X_0 \mu t + \sigma X_0 W_t = X_0 (1 + B_s) \\ X_t &= X_0 e^{\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma N(0,1)\sqrt{t}} = X_0 e^{\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t} \end{aligned} \quad (66.582)$$

et rappelons que l'avantage du mouvement brownien géométrique est qu'il élimine (grâce à l'exponentielle) les valeurs négatives du cours que nous pouvions obtenir avec le mouvement brownien standard de Bachelier.

Il est alors intéressant pour le financier de visualiser l'espérance en fonction de T et la valeur $x(T)$ correspondante à une probabilité cumulée de 2.5% et de 97.5% sur un graphique pour avoir une idée de l'évolution de l'intervalle de confiance à 95% de son $x(0)$. Ceci est très facile à obtenir dans Microsoft Excel et l'on tombe typiquement en jouant avec plusieurs types de graphiques dans le même diagramme sur quelque chose du genre (portefeuille de 500 MF avec rendement de 5% et écart-type de 20%):

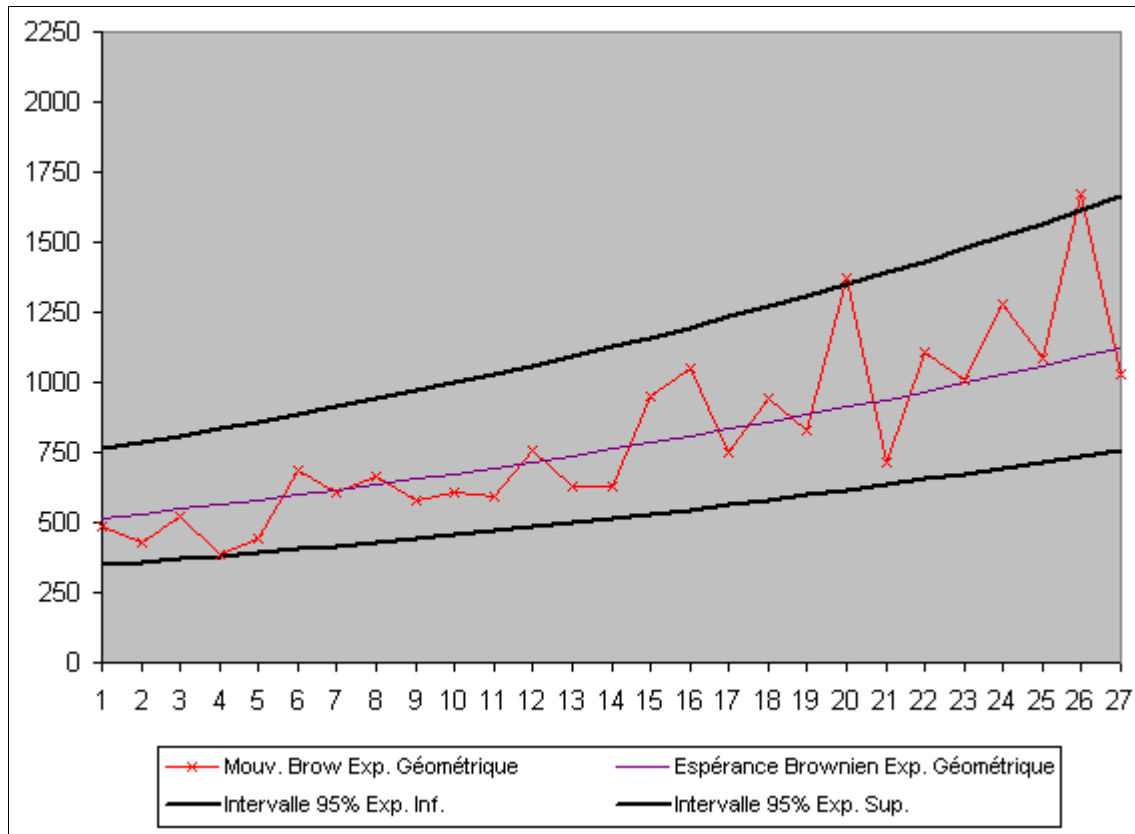


Figure: 66.57 - Tracé du mouvement brownien géométrique avec intervalle de confiance

Avec le tableau suivant:

	A	G	H	I	J
1	Mois	Espérance Brownien Exp. Géométrique	Mouv. Brow Exp. Géométrique	Intervalle 95% Exp. Inf.	Intervalle 95% Exp. Sup.
2	0	500.00			
3	1	515.23	484.24	348.14	762.50
4	2	530.92	538.26	358.75	785.72
5	3	547.09	515.09	369.67	809.65
6	4	563.75	566.32	380.93	834.31
7	5	580.92	442.14	392.53	859.71
8	6	598.61	524.75	404.49	885.90
9	7	616.84	704.80	416.80	912.88
10	8	635.62	621.09	429.50	940.68
11	9	654.98	653.62	442.58	969.33
12	10	674.93	667.10	456.06	998.85
13	11	695.48	598.87	469.94	1029.27
14	12	716.66	714.86	484.26	1060.61
15	13	738.49	809.84	499.00	1092.91
16	14	760.98	1083.04	514.20	1126.20
17	15	784.16	602.51	529.86	1160.49
18	16	808.04	674.59	546.00	1195.84
19	17	832.65	1066.80	562.63	1232.25
20	18	858.00	813.95	579.76	1269.78
21	19	884.13	897.73	597.42	1308.45
22	20	911.06	932.80	615.61	1348.30
23	21	938.81	844.90	634.36	1389.36
24	22	967.40	867.24	653.68	1431.68
25	23	996.86	890.16	673.59	1475.28
26	24	1027.22	1220.87	694.10	1520.21
27	25	1058.50	969.59	715.24	1566.50
28	26	1090.74	1325.85	737.02	1614.21
29	27	1123.95	1255.50	759.47	1663.37

Figure: 66.58 - Tableau des données correspondantes au tracé précédent

Utilisant donc les relations démontrées plus haut (la cellule \$B\$2 contient la valeur 500):

	A	G	H
1	Mois	Espérance Brownien Exp. Géométrique	Mouv. Brow Exp. Géométrique
2	0	=B\$2*EXP((5%-20% ² /2)*A2)	
3	1	=B\$2*EXP((5%-20% ² /2)*A3)	=B\$2*EXP((5%-20% ² /2)*A3)*EXP(20%*NORMSINV(RAND()))
4	2	=B\$2*EXP((5%-20% ² /2)*A4)	=B\$2*EXP((5%-20% ² /2)*A4)*EXP(20%*NORMSINV(RAND()))
5	3	=B\$2*EXP((5%-20% ² /2)*A5)	=B\$2*EXP((5%-20% ² /2)*A5)*EXP(20%*NORMSINV(RAND()))
6	4	=B\$2*EXP((5%-20% ² /2)*A6)	=B\$2*EXP((5%-20% ² /2)*A6)*EXP(20%*NORMSINV(RAND()))
7	5	=B\$2*EXP((5%-20% ² /2)*A7)	=B\$2*EXP((5%-20% ² /2)*A7)*EXP(20%*NORMSINV(RAND()))
8	6	=B\$2*EXP((5%-20% ² /2)*A8)	=B\$2*EXP((5%-20% ² /2)*A8)*EXP(20%*NORMSINV(RAND()))
9	7	=B\$2*EXP((5%-20% ² /2)*A9)	=B\$2*EXP((5%-20% ² /2)*A9)*EXP(20%*NORMSINV(RAND()))
10	8	=B\$2*EXP((5%-20% ² /2)*A10)	=B\$2*EXP((5%-20% ² /2)*A10)*EXP(20%*NORMSINV(RAND()))
11	9	=B\$2*EXP((5%-20% ² /2)*A11)	=B\$2*EXP((5%-20% ² /2)*A11)*EXP(20%*NORMSINV(RAND()))
12	10	=B\$2*EXP((5%-20% ² /2)*A12)	=B\$2*EXP((5%-20% ² /2)*A12)*EXP(20%*NORMSINV(RAND()))
13	11	=B\$2*EXP((5%-20% ² /2)*A13)	=B\$2*EXP((5%-20% ² /2)*A13)*EXP(20%*NORMSINV(RAND()))
14	12	=B\$2*EXP((5%-20% ² /2)*A14)	=B\$2*EXP((5%-20% ² /2)*A14)*EXP(20%*NORMSINV(RAND()))
15	13	=B\$2*EXP((5%-20% ² /2)*A15)	=B\$2*EXP((5%-20% ² /2)*A15)*EXP(20%*NORMSINV(RAND()))
16	14	=B\$2*EXP((5%-20% ² /2)*A16)	=B\$2*EXP((5%-20% ² /2)*A16)*EXP(20%*NORMSINV(RAND()))
17	15	=B\$2*EXP((5%-20% ² /2)*A17)	=B\$2*EXP((5%-20% ² /2)*A17)*EXP(20%*NORMSINV(RAND()))
18	16	=B\$2*EXP((5%-20% ² /2)*A18)	=B\$2*EXP((5%-20% ² /2)*A18)*EXP(20%*NORMSINV(RAND()))
19	17	=B\$2*EXP((5%-20% ² /2)*A19)	=B\$2*EXP((5%-20% ² /2)*A19)*EXP(20%*NORMSINV(RAND()))
20	18	=B\$2*EXP((5%-20% ² /2)*A20)	=B\$2*EXP((5%-20% ² /2)*A20)*EXP(20%*NORMSINV(RAND()))
21	19	=B\$2*EXP((5%-20% ² /2)*A21)	=B\$2*EXP((5%-20% ² /2)*A21)*EXP(20%*NORMSINV(RAND()))
22	20	=B\$2*EXP((5%-20% ² /2)*A22)	=B\$2*EXP((5%-20% ² /2)*A22)*EXP(20%*NORMSINV(RAND()))
23	21	=B\$2*EXP((5%-20% ² /2)*A23)	=B\$2*EXP((5%-20% ² /2)*A23)*EXP(20%*NORMSINV(RAND()))
24	22	=B\$2*EXP((5%-20% ² /2)*A24)	=B\$2*EXP((5%-20% ² /2)*A24)*EXP(20%*NORMSINV(RAND()))
25	23	=B\$2*EXP((5%-20% ² /2)*A25)	=B\$2*EXP((5%-20% ² /2)*A25)*EXP(20%*NORMSINV(RAND()))
26	24	=B\$2*EXP((5%-20% ² /2)*A26)	=B\$2*EXP((5%-20% ² /2)*A26)*EXP(20%*NORMSINV(RAND()))
27	25	=B\$2*EXP((5%-20% ² /2)*A27)	=B\$2*EXP((5%-20% ² /2)*A27)*EXP(20%*NORMSINV(RAND()))
28	26	=B\$2*EXP((5%-20% ² /2)*A28)	=B\$2*EXP((5%-20% ² /2)*A28)*EXP(20%*NORMSINV(RAND()))
29	27	=B\$2*EXP((5%-20% ² /2)*A29)	=B\$2*EXP((5%-20% ² /2)*A29)*EXP(20%*NORMSINV(RAND()))

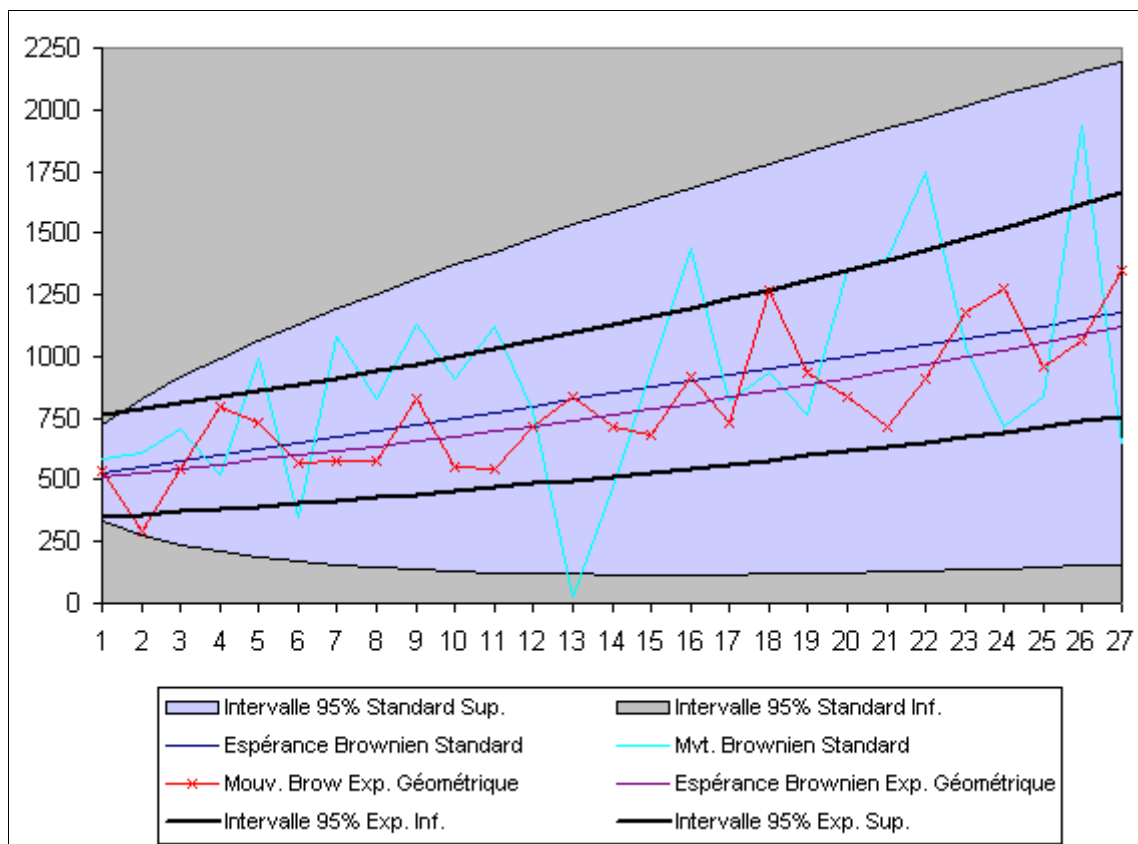
Figure: 66.59 - Formules explicites pour le mouvement brownien géométrique

et pour l'intervalle de confiance à 95%:

	I	J
1	Intervalle 95% Exp. Inf.	Intervalle 95% Exp. Sup.
2		
3	=B\$2*EXP((5%-20%*2/2)*A3)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A3)*EXP(20%*NORMSINV(97.5%))
4	=B\$2*EXP((5%-20%*2/2)*A4)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A4)*EXP(20%*NORMSINV(97.5%))
5	=B\$2*EXP((5%-20%*2/2)*A5)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A5)*EXP(20%*NORMSINV(97.5%))
6	=B\$2*EXP((5%-20%*2/2)*A6)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A6)*EXP(20%*NORMSINV(97.5%))
7	=B\$2*EXP((5%-20%*2/2)*A7)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A7)*EXP(20%*NORMSINV(97.5%))
8	=B\$2*EXP((5%-20%*2/2)*A8)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A8)*EXP(20%*NORMSINV(97.5%))
9	=B\$2*EXP((5%-20%*2/2)*A9)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A9)*EXP(20%*NORMSINV(97.5%))
10	=B\$2*EXP((5%-20%*2/2)*A10)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A10)*EXP(20%*NORMSINV(97.5%))
11	=B\$2*EXP((5%-20%*2/2)*A11)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A11)*EXP(20%*NORMSINV(97.5%))
12	=B\$2*EXP((5%-20%*2/2)*A12)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A12)*EXP(20%*NORMSINV(97.5%))
13	=B\$2*EXP((5%-20%*2/2)*A13)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A13)*EXP(20%*NORMSINV(97.5%))
14	=B\$2*EXP((5%-20%*2/2)*A14)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A14)*EXP(20%*NORMSINV(97.5%))
15	=B\$2*EXP((5%-20%*2/2)*A15)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A15)*EXP(20%*NORMSINV(97.5%))
16	=B\$2*EXP((5%-20%*2/2)*A16)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A16)*EXP(20%*NORMSINV(97.5%))
17	=B\$2*EXP((5%-20%*2/2)*A17)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A17)*EXP(20%*NORMSINV(97.5%))
18	=B\$2*EXP((5%-20%*2/2)*A18)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A18)*EXP(20%*NORMSINV(97.5%))
19	=B\$2*EXP((5%-20%*2/2)*A19)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A19)*EXP(20%*NORMSINV(97.5%))
20	=B\$2*EXP((5%-20%*2/2)*A20)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A20)*EXP(20%*NORMSINV(97.5%))
21	=B\$2*EXP((5%-20%*2/2)*A21)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A21)*EXP(20%*NORMSINV(97.5%))
22	=B\$2*EXP((5%-20%*2/2)*A22)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A22)*EXP(20%*NORMSINV(97.5%))
23	=B\$2*EXP((5%-20%*2/2)*A23)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A23)*EXP(20%*NORMSINV(97.5%))
24	=B\$2*EXP((5%-20%*2/2)*A24)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A24)*EXP(20%*NORMSINV(97.5%))
25	=B\$2*EXP((5%-20%*2/2)*A25)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A25)*EXP(20%*NORMSINV(97.5%))
26	=B\$2*EXP((5%-20%*2/2)*A26)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A26)*EXP(20%*NORMSINV(97.5%))
27	=B\$2*EXP((5%-20%*2/2)*A27)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A27)*EXP(20%*NORMSINV(97.5%))
28	=B\$2*EXP((5%-20%*2/2)*A28)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A28)*EXP(20%*NORMSINV(97.5%))
29	=B\$2*EXP((5%-20%*2/2)*A29)*EXP(20%*NORMSINV(2.5%))	=B\$2*EXP((5%-20%*2/2)*A29)*EXP(20%*NORMSINV(97.5%))

Figure: 66.60 - Formules explicites pour l'intervalle de confiance

Il est intéressant de comparer l'évolution de portefeuilles ayant les mêmes paramètres (volatilité et rendement) sur la même période de temps. Cela donne alors graphiquement:



Donc dans 6 mois, nous voyons que le portefeuille a 95% de probabilité cumulée de se situer entre 404.49 et 885.90 millions avec une espérance de 662.54 millions.

Enfin, le lecteur remarquera que l'on peut généraliser l'écriture des deux mouvements browniens (en prenant le logarithme népérien en ce qui concerne le mouvement brownien géométrique) en les écrivant sous une forme proposée par Mandelbrot en 1962:

$$X_t = X_0 + \delta t + c L_t^\alpha \quad (66.583)$$

où δ et c sont respectivement les paramètres de localisation (rentabilité moyenne) et de dispersion (volatilité non gaussienne) du processus, et où L_t^α désigne le mouvement α -stable standard de Lévy.

Le problème avec ce modèle c'est la perte de l'existence, pour certaines lois de probabilité qui marchent très bien, du deuxième moment (la variance) si important en termes de communication et d'images pour les professionnels dans les années 1970 car il leur servait d'unique mesure du risque. L'absence de variance finie constitua vraisemblablement l'une des causes les plus puissantes du rejet.

7.8.7. ÉQUATION DE BLACK & SCHOLES

Nous avons obtenu lors des développements précédents, sous la contrainte d'une loi log-normale et d'un mouvement brownien, l'équation différentielle suivante pour la marche aléatoire de la valeur de l'action:

$$dF = \left(\frac{\partial F}{\partial t} + a \frac{\partial F}{\partial x} + \frac{1}{2} b^2 \frac{\partial^2 F}{\partial x^2} \right) dt + b \frac{\partial F}{\partial x} dz = \left(\frac{\partial F}{\partial t} + \mu x \frac{\partial F}{\partial x} + \frac{1}{2} \sigma^2 x^2 \frac{\partial^2 F}{\partial x^2} \right) dt + \sigma x \frac{\partial F}{\partial x} dz \quad (66.584)$$

Soit avec les bonnes notations:

$$dF = \left(\frac{\partial F}{\partial t} + a \frac{\partial F}{\partial S} + \frac{1}{2} b^2 \frac{\partial^2 F}{\partial S^2} \right) dt + b \frac{\partial F}{\partial S} dz \underset{\substack{a=\mu x \\ b=\sigma x \\ x=S}}{=} \left(\frac{\partial F}{\partial t} + \mu S \frac{\partial F}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 F}{\partial S^2} \right) dt + \sigma S \frac{\partial F}{\partial S} dz \quad (66.585)$$

Si nous construisons maintenant un portefeuille consistant en une option et un nombre $-\Delta$ de titres sous-jacents (souvent aussi noté $-\Delta_t$ dans la littérature). La valeur du portefeuille Π est alors exprimée par:

$$\Pi = F - \Delta S \quad (66.586)$$

Le différentiel temporel du portefeuille s'écrit alors:

$$d\Pi = dF - \Delta dS \quad (66.587)$$

Vous remarquerez que nous supposons constant (et négatif) le nombre Δ durant le différentiel de temps.

Remarque: Attention!!! Certains auteurs choisissent, par convention, d'insérer les signes des termes à droite de l'égalité des deux relations précédentes. Cela ne change pas le résultat final de la solution de l'équation différentielle mais il faut savoir que cela peut se faire.

En réunissant les relations précédentes et (nous adoptons ici la notation traditionnelle usitée dans le domaine) l'équation de l'actif risqué donnée donc par:

$$\frac{dS}{S} = \sigma dz + \mu dt \quad (66.588)$$

nous obtenons:

$$d\Pi = dF - \Delta dS = \left[\left(\frac{\partial F}{\partial t} + \mu S \frac{\partial F}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 F}{\partial S^2} \right) dt + \sigma S \frac{\partial F}{\partial S} dz \right] - \Delta [S \sigma dz + S \mu dt] \quad (66.589)$$

où nous avons dans le crochet tout à droite le mouvement brownien géométrique.

Ce qui donne après réarrangement des termes l'équation différentielle du portefeuille:

$$d\Pi = \sigma S \left(\frac{\partial F}{\partial S} - \Delta \right) dz + \left(\mu S \frac{\partial F}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 F}{\partial S^2} + \frac{\partial F}{\partial t} - \mu \Delta S \right) dt \quad (66.590)$$

Considérons maintenant que Δ est lié par la relation de dépendance spéculative (dont nous prenons la valeur entière) qui élimine de la relation précédente la partie aléatoirement risquée du portefeuille (c'est le dz qui génère le risque de manière aléatoire pour rappel!) tel que:

$$\frac{\partial F}{\partial S} - \Delta = 0 \Leftrightarrow \Delta = \frac{\partial F}{\partial S} \quad (66.591)$$

Nous pouvons alors écrire:

$$d\Pi = \left(\mu S \frac{\partial F}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 F}{\partial S^2} + \frac{\partial F}{\partial t} - \mu \Delta S \right) dt \quad (66.592)$$

Or, nous avons également pour un portefeuille sans risque (puisque nous venons d'éliminer le risque aléatoire nous pouvons l'écrire ainsi):

$$d\Pi = r\Pi dt \quad (66.593)$$

noté parfois aussi dans littérature:

$$dS_t^0 = rS_t^0 dt \quad (66.594)$$

En substituant maintenant les quatre relations:

$$\Pi = F - \Delta S \quad \Delta = \frac{\partial F}{\partial S} \quad d\Pi = r\Pi dt \quad r = \mu \quad (66.595)$$

dans:

$$d\Pi = \left(\mu S \frac{\partial F}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 F}{\partial S^2} + \frac{\partial F}{\partial t} - \mu \Delta S \right) dt \quad (66.596)$$

Nous obtenons:

$$\frac{\partial F}{\partial t} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 F}{\partial S^2} + rS \frac{\partial F}{\partial S} - rF = 0 \quad (66.597)$$

qui n'est autre que "l'équation différentielle partielle (sans second membre) de Black & Scholes" (à coefficient S non constant). Cette dernière équation est plus souvent notée sous la forme relative à la référence implicite d'un Call telle que:

$$\frac{\partial C}{\partial t} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + rS \frac{\partial C}{\partial S} - rC = 0 \quad (66.598)$$

ou sous forme un peu plus condensée (suivant les auteurs...):

$$\partial_t C + \frac{1}{2} \sigma^2 S^2 \partial_S^2 C + rS \partial_S C - rC = 0 \quad (66.599)$$

et encore plus condensée... (mais là cela devient franchement abusé...):

$$C_t + \frac{1}{2} \sigma^2 S^2 C_{SS} + rSC_S - rC = 0 \quad (66.600)$$

Le lecteur aura noté que le paramètre μ (dérivation) est absent de cette équation! En d'autres termes, la valeur d'une option est indépendante de la vitesse de variation des valeurs des titres sous-jacents. Le seul paramètre qui affecte le prix de l'option est la volatilité σ du sous-jacent. Une conséquence de cela est que deux personnes ayant des opinions divergentes quant à la valeur de μ sont toujours en entente sur la valeur de l'option.

L'objectif bien évidemment est de résoudre cette équation différentielle afin de déterminer le return $F(S, t)$. Celle-ci ne se laisse par ailleurs pas résoudre en deux lignes.

Avant de nous attaquer à cette tâche quelques définitions et indications pratiques préalables concernant certains paramètres sont utiles et nécessaires (nous déterminerons leur forme explicite après la résolution de l'équation différentielle):

7.8.8. PORTEFEUILLE AUTOFINANÇANT SUR SOUS-JACENT RISQUÉ

Une stratégie de portefeuille autofinçant est une stratégie dynamique d'achat ou de vente de titres et de prêts ou d'emprunts à la banque, dont la valeur n'est pas modifiée par l'ajout ou le retrait de cash (nous aurions pu introduire ce sujet dès le début du chapitre mais nous avons jugé plus opportun de ne le faire que maintenant).

Nous supposons ici pour l'exemple que nous ne pouvons investir que dans un seul titre (placement risqué), et dans du cash (placement supposé non risqué), c'est-à-dire en plaçant ou empruntant de l'argent à une banque.

Nous désignons par S_t le prix à la date t du titre, par r_t le taux d'intérêt pour un placement entre $[t, t + dt]$ à la banque.

Soit V_t la valeur de marché, ou encore valeur liquidative, ou encore "**Mark to Market**" (M.t.M.) du portefeuille à la date t . Après renégociation, le nombre d'actions δ_t du portefeuille est constant jusqu'à la prochaine date d'échéance de gestion. Pour simplifier, nous supposons pour le moment que le gestionnaire ne prend en compte dans sa règle de décision la valeur du cours du sous-jacent qu'au moment de renégocier.

Dans un temps très court, la variation de valeur du portefeuille n'est due qu'à la variation de la valeur du sous-jacent et à l'intérêt versé par la banque sur le cash, soit, puisque le montant investi dans le cash est:

$$V_t - \delta_t \cdot S_t \quad (66.601)$$

nous avons "**l'équation d'autofinancement**":

$$dV_t = \delta_t \cdot dS_t + (V_t - \delta_t \cdot S_t) r_t dt = V_t r_t dt + \delta_t \cdot (dS_t - r_t S_t dt) \quad (66.602)$$

Ainsi, pour un vendeur de Call (par exemple...), il s'agit de trouver le coût initial V_0 et la stratégie δ_t qui permettent d'obtenir (les financiers parlent de "**réaliser l'actif financier**"):

$$(S_T - K)^+ \quad (66.603)$$

dans tous les scénarios de marché. S'il existe une telle stratégie de couverture, nous disons alors que nous avons affaire à un "**marché complet**".

7.8.9. LES GRECQUES ET AUTRES...

Dans la réalité ces modèles théoriques sont bien jolis (pour l'honneur de l'esprit humain) mais toutefois très peu réalistes. Raison pour laquelle les traders font (comme en ingénierie) appel à des indicateurs simples et parlants appelés "**les grecques**" pour prendre leurs décisions de vente ou d'achat. Cependant on peut, avec un peu de créativité, utiliser ces mêmes indicateurs dans de nombreux autres domaines (industriel, gestion de projets, chimie, etc.).

Voici la liste des plus connus:

Définitions:

D1. Le "**delta**" d'une option, qu'il est important de comprendre (ou de savoir), donné par:

$$\Delta = \frac{\partial F}{\partial S} \quad (66.604)$$

et représente le taux de changement de la valeur des options du portefeuille dépendamment des valeurs des titres sous-jacents S (mathématiquement parlant c'est donc la dérivée première de la prime de l'option sur le prix du sous-jacent). Ce terme est fondamental dans la théorie et dans la pratique et nous en ferons fréquemment usage. C'est donc une mesure dans la corrélation entre le mouvement de l'option ou autres actifs financiers et dérivés et les sous-jacents.

Considérons par exemple qu'un Call sur l'action ABC est de delta 0.25 avec un cours du support (spot) à 90.- et une prime à 5.-. Lorsque le cours de l'action ABC passe de 90.- à 91.-, la prime de l'option va augmenter alors de 1 delta, et devient alors 5.25.-. Lorsque le cours de l'action ABC passe de 90.- à 88.-, la prime de l'option va diminuer de 2 fois delta, et devient 4.50.-. Cette variation en termes de delta (nombre entier) est alors notée δ_1 .

Le delta est donc le paramètre le plus important pour un praticien qui veut se couvrir contre le risque. Effectivement, afin d'obtenir le delta global d'une position, il suffit de multiplier la valeur du delta de chaque option par sa position. Puis on fait la somme de tous ces deltas.

Par exemple, si sur un même sous-jacent nous sommes vendeur de 5 calls C_1 et acheteur de 7 Call C_2 alors notre delta global sera égal à:

$$7 \cdot \Delta_{C_2} - 5 \cdot \Delta_{C_1} \quad (66.605)$$

La valeur de ce paramètre nous informe sur la quantité de sous-jacents à acheter ou vendre (position opposée sur le sous-jacent) afin d'immuniser la valorisation de notre portefeuille aux variations du cours de ce sous-jacent. Nous disons alors qu'il s'agit d'une "**stratégie en delta-neutre**" ou de "**delta-hedging**". Plus exactement il s'agit d'une stratégie dynamique car il faut maintenir un portefeuille delta-neutre tout au long de la vie de l'option contrairement aux stratégies statiques. Par exemple, une stratégie statique pour couvrir la vente d'un Call de strike de 125.- sur 1'000 actions (prix spot 100.-), peut consister à acheter immédiatement 1'000 actions sur le marché contre 100'000.-, et attendre l'échéance. Dans cette situation, si l'acheteur exerce le Call le vendeur est en mesure de lui livrer les 1'000 actions contre 125'000.-. Mais s'il ne l'exerce pas, le vendeur reste avec 1'000 actions. De même, si le cours final s'établit par exemple à 80.-, le vendeur enregistre une perte de 20'000.-. Le delta-hedging permet d'éliminer ce risque.

Ainsi, les gestionnaires, vont entre la date à laquelle ils ont encaissé la prime (en ayant vendu un contrat d'option) et sa maturité T , tout naturellement gérer en delta-neutre au fil du temps un portefeuille autofinancé constitué de δ_t actifs sous-jacents S à chaque instant t , afin de disposer de façon certaine (donc sans risque) de la cible stochastique à la maturité. Nous parlons aussi de "**portefeuille de couverture**".

D2. Le "**thêta**" d'une option donne la sensibilité du prix de l'option par rapport à sa maturité et est donné par:

$$\theta = \frac{\partial F}{\partial t} \quad (66.606)$$

Appliqué à notre portefeuille, le thêta nous donne simplement la valeur perdue ou gagnée suite à l'écoulement d'une journée par exemple.

D3. Le "**rhô**" calcule la sensibilité du prix de l'option par rapport à la variation au taux d'intérêt géométrique moyen sans risque du marché et est donné par:

$$\rho = \frac{\partial F}{\partial r} \quad (66.607)$$

Cet indicateur semble être assez peu utilisé par les professionnels.

D4. Le "**véga**", représenté par la lettre nu minuscule car le nom véga n'est pas lui-même un nom de lettre grecque, mesure la sensibilité de l'option par rapport à la volatilité et est donné par:

$$\nu = \frac{\partial F}{\partial \sigma} \quad (66.608)$$

La volatilité est le paramètre déterminant du prix d'une option. L'impact de la variation de ce paramètre sur la valorisation de notre portefeuille est donc très important pour les traders sur options.

D5. Le "**gamma**" correspond à la dérivée du delta et est donc donné par:

$$\Gamma = \frac{\partial^2 F}{\partial S^2} \quad (66.609)$$

Une lecture possible du gamma est le sens d'évolution du delta en fonction du prix du sous-jacent. Un gamma positif indique que prix du sous-jacent et delta évoluent dans le même sens, alors qu'un gamma négatif montre le contraire.

Afin d'éviter de réajuster le delta en stratégie delta-neutre, et donc de payer les frais de transactions mais aussi les spreads inhérents, un bon moyen est d'avoir le delta le plus stable possible. Avoir le delta le plus stable signifie pas ou peu de modification de sa valeur initiale, donc avoir une stratégie "**gamma-neutre**" en plus d'avoir une stratégie "delta-neutre".

Comme la stratégie delta-neutre fait que l'on joue déjà avec le sous-jacent pour rendre le delta constant, alors on ne peut utiliser le sous-jacent pour rendre le gamma neutre. Le principe de gamma hedging d'un portefeuille contenant au moins un option, est alors de trouver une autre option sur le marché, dans des proportions qui permettent de compenser le gamma de la position originelle.

Exemple:

Considérons comme résumé un Call dont sur sous-jacent dont le prix spot est de 100.-, le strike de 100.-, la volatilité du rendement de 24% et le taux d'intérêt géométrique moyen du marché de 5%. Considérons que nous avons les valeurs suivantes:

$$\delta = 0.62 \quad \Gamma = 0.02 \quad \theta = -0.02.- \quad \nu = 0.37.- \quad \rho = 0.5.- \quad (66.610)$$

La valeur du delta nous indique donc que si le prix du sous-jacent montre de 1.-, la valeur de l'option augmentera de 0.62.-. Le gamma nous indique que si le prix du sous-jacent montre de 1.-, le delta de l'option passera de 0.62 à 0.64. La valeur du thêta nous indique en fin de journée que la valeur de l'option aura diminué de 0.02.-. La valeur du vega nous indique que si la volatilité du sous-jacent passe de 24% à 25%, la valeur de l'option augmentera de 0.37.-. Enfin la valeur du rhô nous indique si la valeur du taux moyen géométrique sans risque du marché passe de 5% à 5.5%, la valeur de l'option augmentera de 0.25.-.

D6. "**L'opérateur différentiel linéaire de Black & Scholes**" L_{BS} donné par:

$$L_{BS} = \frac{\partial}{\partial t} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2}{\partial S^2} + rS \frac{\partial}{\partial S} - r \quad (66.611)$$

aurait une interprétation financière comme mesure de la différence entre le retour d'une option (les deux premiers termes) et l'ensemble d'un portefeuille de rendement sans risque r contenant cette option (les deux derniers termes).

Bref, ceci étant dit, nous pouvons donc avoir l'écriture technique suivante de l'E.D.P. de Black & Scholes (à coefficient S non constant) en utilisant les grecques:

$$\theta + \frac{1}{2} \sigma^2 S^2 \Gamma + rS\Delta - rF = 0 \quad (66.612)$$

7.8.10. RÉOLUTION DE L'E.D.P. DE BLACK & SCHOLES

Avant de nous attaquer à la résolution de l'équation B.S. donnons déjà les solutions avec un rappel des termes (cela permettra d'avoir une idée préalable des concepts utilisés lors des développements et de plus je ne risque pas d'écrire ceux-ci avant quelques années faute de temps...):

Soient $F(S,t)$ la valeur d'une option Call $C(S,t)$ ou Put $P(S,t)$, σ la volatilité du sous-jacent, E le prix d'exercice (strike), T la date d'expiration et r l'intérêt et S le prix du sous-jacent. Nous allons résoudre l'équation différentielle de BS (à coefficient S non constant) écrite la forme suivante:

$$\frac{\partial C}{\partial t} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + rS \frac{\partial C}{\partial S} - rC = 0 \quad (66.613)$$

L'astuce (qui a menée à l'obtention d'un prix Nobel d'Économie pour rappel) est de voir qu'il s'agit d'une équation différentielle aux dérivées partielles et plus particulièrement avec des différentielles du premier ordre en t et en deuxième ordre en S . Or, il existe une équation différentielle de ce type qui est très connue en physique: l'équation de la chaleur (cf. chapitre de Thermodynamique). L'idée est alors par des changements de variables de s'y ramener. Toute la difficulté était de trouver les bons changements de variable et nos prédécesseurs se sont occupés de ce tâtonnement... pour nous (merci à eux!). Nous commençons alors par poser (ce choix permet aussi de se ramener à des conditions initiales similaires à celles de l'équation de la chaleur):

$$C(t, S) = Ku(\tau, x) \quad \tau = \sigma^2 (T - t) \quad x = \ln\left(\frac{S}{K}\right) \quad (66.614)$$

et donc explicitement:

$$C(t, S) = Ku(\tau, x) = Ku\left(\sigma^2 (T - t), \ln\left(\frac{S}{K}\right)\right) \quad (66.615)$$

Nous utiliserons plus bas le fait que:

$$\frac{\partial \tau}{\partial t} = -\sigma^2 \quad \frac{\partial x}{\partial S} = \frac{1}{S} \quad (66.616)$$

À toute fin utile, écrivons la transformée inverse (qui nous sera utile beaucoup plus tard) qui en découle:

$$u(\tau, x) = \frac{1}{K} C(t, S) = \frac{1}{K} C\left(T - \frac{\tau}{\sigma^2}, Ke^x\right) \quad (66.617)$$

Une partie de l'idée sous-jacente à ce choix de changement de variables est de transformer le fait que dans l'équation différentielle de Black & Scholes, nous connaissons les conditions finales (puisque à maturité le prix de l'option est parfaitement déterminée) alors qu'avec l'équation de la chaleur il nous faut des conditions initiales et non finales! Nous allons de suite voir que ce choix de changement de variables permet cela. Effectivement, rappelons qu'à maturité nous savons que:

$$C(T, S) = (S_T - K)^+ \quad (66.618)$$

Soit:

$$\frac{1}{K} C(T, S) = \left(\frac{S_T}{K} - 1\right)^+ \quad (66.619)$$

Comme:

$$x = \ln\left(\frac{S_T}{K}\right) \Rightarrow \frac{S_T}{K} = e^x \quad (66.620)$$

Nous avons alors:

$$\frac{1}{K} C(T, S) = (e^x - 1)^+ \quad (66.621)$$

Et le fait ce que nous permet d'avoir une condition initiale est que:

$$\frac{1}{K} C(T, S) = (e^x - 1)^+ = u(\tau, x) = u\left(\sigma^2(T-t), x\right) \Big|_{t=T} = u(0, x) \quad (66.622)$$

Pour transformer notre équation différentielle aux dérivées partielles (à coefficient S non constant):

$$\frac{\partial C}{\partial t} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + rS \frac{\partial C}{\partial S} - rC = 0 \quad (66.623)$$

avec ces nouvelles variables il nous faut calculer quelques dérivées partielles de C . Nous avons alors:

$$\begin{aligned} \frac{\partial C}{\partial t} &= \frac{\partial}{\partial t} \left[K \cdot u\left(\sigma^2(T-t), \ln\left(\frac{S}{K}\right)\right) \right] = K \frac{\partial}{\partial t} \left[u\left(\sigma^2(T-t), \ln\left(\frac{S}{K}\right)\right) \right] \\ &= K \frac{\partial}{\partial \tau} \left[u\left(\sigma^2(T-t), \ln\left(\frac{S}{K}\right)\right) \right] = -K \sigma^2 \frac{\partial u}{\partial \tau} \end{aligned} \quad (66.624)$$

et:

$$\begin{aligned}\frac{\partial C}{\partial S} &= \frac{\partial}{\partial S} \left[K \cdot u \left(\sigma^2 (T-t), \ln \left(\frac{S}{K} \right) \right) \right] = K \frac{\partial}{\partial S} \left[u \left(\sigma^2 (T-t), \ln \left(\frac{S}{K} \right) \right) \right] \\ &= K \frac{\partial}{S \partial x} \left[u \left(\sigma^2 (T-t), \ln \left(\frac{S}{K} \right) \right) \right] = \frac{K}{S} \frac{\partial u}{\partial x}\end{aligned}\quad (66.625)$$

Ainsi que:

$$\begin{aligned}\frac{\partial^2 C}{\partial S^2} &= \frac{\partial}{\partial S} \left(\frac{K}{S} \frac{\partial u}{\partial x} \right) = -\frac{K}{S^2} \frac{\partial u}{\partial x} + \frac{K}{S} \left[\frac{\partial}{\partial S} u_x \left(\sigma^2 (T-t), \ln \left(\frac{S}{K} \right) \right) \right] \\ &= -\frac{K}{S^2} \frac{\partial u}{\partial x} + \frac{K}{S} \frac{\partial}{\partial S} u_x = -\frac{K}{S^2} \frac{\partial u}{\partial x} + \frac{K}{S} \frac{\partial}{S \partial x} u_x = -\frac{K}{S^2} \frac{\partial u}{\partial x} + \frac{K}{S^2} \frac{\partial}{\partial x} u_x = -\frac{K}{S^2} \frac{\partial u}{\partial x} + \frac{K}{S^2} \frac{\partial^2 u}{\partial x^2} \\ &= \frac{K}{S^2} \left(\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial x} \right)\end{aligned}\quad (66.626)$$

En injectant cela dans l'équation différentielle, il vient:

$$\left[-K \frac{\partial u}{\partial \tau} \sigma^2 \right] + \frac{1}{2} \sigma^2 S^2 \left[\frac{K}{S^2} \left(\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial x} \right) \right] + rS \left[\frac{K}{S} \frac{\partial u}{\partial x} \right] - r[Ku] = 0 \quad (66.627)$$

On simplifie les S:

$$-K \frac{\partial u}{\partial \tau} \sigma^2 + \frac{1}{2} K \left(\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial x} \right) \sigma^2 + rK \frac{\partial u}{\partial x} - rKu = 0 \quad (66.628)$$

On multiplie des deux côtés par:

$$-\frac{1}{K \sigma^2} \quad (66.629)$$

Pour obtenir:

$$\frac{\partial u}{\partial \tau} - \frac{1}{2} \left(\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial x} \right) - \frac{r}{\sigma^2} \frac{\partial u}{\partial x} + \frac{r}{\sigma^2} u = 0 \quad (66.630)$$

Soit:

$$\frac{\partial u}{\partial \tau} = \frac{1}{2} \frac{\partial^2 u}{\partial x^2} + \left(\frac{r}{\sigma^2} - \frac{1}{2} \right) \frac{\partial u}{\partial x} - \frac{r}{\sigma^2} u \quad (66.631)$$

Posons comme il est d'usage:

$$\frac{r}{\sigma^2} = \kappa \quad (66.632)$$

Il vient alors l'équation différentielle (à coefficient constants cette fois-ci!):

$$\frac{\partial u}{\partial \tau} = \frac{1}{2} \frac{\partial^2 u}{\partial x^2} + \left(\kappa - \frac{1}{2} \right) \frac{\partial u}{\partial x} - \kappa u \quad (66.633)$$

avec pour rappel la condition initiale:

$$u(0, x) = (e^x - 1)^+ \quad (66.634)$$

La suite toute aussi astucieuse, consiste à faire encore un changement de variable en posant:

$$v(\tau, x) = e^{\alpha\tau + \beta x} \cdot u(\tau, x) \quad (66.635)$$

Nous allons déterminer les deux paramètres α, β qui permettent d'écrire l'équation différentielle partielle à coefficients constants sous forme d'équation de la chaleur!

Nous allons procéder d'abord comme plus haut en déterminant l'équivalence des différentes dérivées partielles de u . Nous avons alors:

$$\frac{\partial u}{\partial \tau} = \frac{\partial (e^{-\alpha\tau - \beta x} \cdot v(\tau, x))}{\partial \tau} = \left(-\alpha v + \frac{\partial v}{\partial \tau} \right) e^{-\alpha\tau - \beta x} \quad (66.636)$$

et:

$$\frac{\partial u}{\partial x} = \frac{\partial (e^{-\alpha\tau - \beta x} \cdot v(\tau, x))}{\partial x} = \left(-\beta v + \frac{\partial v}{\partial x} \right) e^{-\alpha\tau - \beta x} \quad (66.637)$$

Ainsi que:

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} &= \frac{\partial}{\partial x} \left(-\beta v e^{-\alpha\tau - \beta x} + \frac{\partial v}{\partial x} e^{-\alpha\tau - \beta x} \right) = \left[-\beta \frac{\partial v}{\partial x} + \beta^2 v + \frac{\partial^2 v}{\partial x^2} - \beta \frac{\partial v}{\partial x} \right] e^{-\alpha\tau - \beta x} \\ &= \left[\frac{\partial^2 v}{\partial x^2} - 2\beta \frac{\partial v}{\partial x} + \beta^2 v \right] e^{-\alpha\tau - \beta x} \end{aligned} \quad (66.638)$$

Ensuite vient une astuce très difficile à deviner (du moins a priori car je n'ai pas trouvé plus simple...). Effectivement avant de substituer ces trois dernières relations dans l'E.D., rappelons que:

$$v(\tau, x) = e^{\alpha\tau + \beta x} \cdot u(\tau, x) \quad (66.639)$$

Donc les trois dernières relations peuvent s'écrire (retrouver un mélange fonction de u et de v est très contre-intuitif):

$$\begin{aligned}\frac{\partial u}{\partial \tau} &= \left(-\alpha v + \frac{\partial v}{\partial \tau} \right) e^{-\alpha \tau - \beta x} = -\alpha u + \frac{\partial v}{\partial \tau} e^{-\alpha \tau - \beta x} \\ \frac{\partial u}{\partial x} &= \left(-\beta v + \frac{\partial v}{\partial x} \right) e^{-\alpha \tau - \beta x} = -\beta u + \frac{\partial v}{\partial x} e^{-\alpha \tau - \beta x} \\ \frac{\partial^2 u}{\partial x^2} &= \left[\frac{\partial^2 v}{\partial x^2} - 2\beta \frac{\partial v}{\partial x} + \beta^2 v \right] e^{-\alpha \tau - \beta x} = \frac{\partial^2 v}{\partial x^2} e^{-\alpha \tau - \beta x} - 2\beta \frac{\partial v}{\partial x} e^{-\alpha \tau - \beta x} + \beta^2 u\end{aligned}\quad (66.640)$$

et maintenant seulement faisons la substitution (partielle!) dans l'E.D.! Nous passons alors de:

$$\frac{\partial u}{\partial \tau} = \frac{1}{2} \frac{\partial^2 u}{\partial x^2} + \left(\kappa - \frac{1}{2} \right) \frac{\partial u}{\partial x} - \kappa u \quad (66.641)$$

à:

$$\begin{aligned}& \left[-\alpha u + e^{-\alpha \tau - \beta x} \frac{\partial v}{\partial \tau} \right] \\ &= \frac{1}{2} \left[\beta^2 u - 2\beta e^{-\alpha \tau - \beta x} \frac{\partial v}{\partial x} + e^{-\alpha \tau - \beta x} \frac{\partial^2 v}{\partial x^2} \right] + \left(\kappa - \frac{1}{2} \right) \left[-\beta u + e^{-\alpha \tau - \beta x} \frac{\partial v}{\partial x} \right] - \kappa u\end{aligned}\quad (66.642)$$

Écrivons cette dernière égalité un peu différemment en mettant en factorisant à gauche et à droit ce qui contient u :

$$\begin{aligned}& \left[-\alpha \right] u + e^{-\alpha \tau - \beta x} \frac{\partial v}{\partial \tau} = \\ & \left[\frac{1}{2} \beta^2 - \beta \left(\kappa - \frac{1}{2} \right) - \kappa \right] u - \frac{1}{2} \left[2\beta e^{-\alpha \tau - \beta x} \frac{\partial v}{\partial x} - e^{-\alpha \tau - \beta x} \frac{\partial^2 v}{\partial x^2} \right] + \left(\kappa - \frac{1}{2} \right) e^{-\alpha \tau - \beta x} \frac{\partial v}{\partial x}\end{aligned}\quad (66.643)$$

et donc en posant:

$$-\alpha = \left(\frac{1}{2} \beta^2 - \beta \left(\kappa - \frac{1}{2} \right) - \kappa \right) \quad (66.644)$$

L'équation différentielle se simplifie en:

$$e^{-\alpha \tau - \beta x} \frac{\partial v}{\partial \tau} = -\frac{1}{2} \left[2\beta e^{-\alpha \tau - \beta x} \frac{\partial v}{\partial x} - e^{-\alpha \tau - \beta x} \frac{\partial^2 v}{\partial x^2} \right] + \left(\kappa - \frac{1}{2} \right) e^{-\alpha \tau - \beta x} \frac{\partial v}{\partial x} \quad (66.645)$$

L'exponentielle peut se simplifier et il reste:

$$\frac{\partial v}{\partial \tau} = -\frac{1}{2} \left[2\beta \frac{\partial v}{\partial x} - \frac{\partial^2 v}{\partial x^2} \right] + \left(\kappa - \frac{1}{2} \right) \frac{\partial v}{\partial x} \quad (66.646)$$

Ce que l'on peut réécrire sous la forme:

$$\frac{\partial v}{\partial \tau} = \left(\left(\kappa - \frac{1}{2} \right) - \beta \right) \frac{\partial v}{\partial x} + \frac{1}{2} \frac{\partial^2 v}{\partial x^2} \quad (66.647)$$

Et si nous choisissons:

$$\left(\kappa - \frac{1}{2} \right) = \beta \quad (66.648)$$

Il nous reste alors:

$$\frac{\partial v}{\partial \tau} = \frac{1}{2} \frac{\partial^2 v}{\partial x^2} \quad (66.649)$$

ou autrement écrit (à la physicienne):

$$v_t = \frac{1}{2} v_{xx} \quad (66.650)$$

On reconnaît donc ici l'équation de la chaleur (ou "équation de diffusion") démontrée dans le chapitre de Thermodynamique mais où le coefficient de dilatation thermique vaut 1/2.

Il résulte donc de ces développements que:

$$\begin{aligned} -\alpha &= \left(\frac{1}{2} \beta^2 - \beta \left(\kappa - \frac{1}{2} \right) - \kappa \right) \\ \beta &= \left(\kappa - \frac{1}{2} \right) \end{aligned} \quad (66.651)$$

Soit en injectant la deuxième égalité dans la première et après quelques simplifications élémentaires:

$$\begin{aligned} \alpha &= -\frac{1}{2} \left[\kappa - \frac{1}{2} \right]^2 + \left[\kappa - \frac{1}{2} \right] \left(\kappa - \frac{1}{2} \right) + \kappa = \frac{1}{2} \left(\kappa - \frac{1}{2} \right)^2 + \kappa = \frac{1}{2} \left(\kappa^2 - \kappa + \frac{1}{4} \right) + \kappa \\ &= \frac{1}{2} \kappa^2 - \frac{1}{2} \kappa + \frac{1}{8} + \kappa = \frac{1}{2} \kappa^2 + \frac{1}{2} \kappa + \frac{1}{8} = \frac{1}{2} \left(\kappa + \frac{1}{2} \right)^2 \end{aligned} \quad (66.652)$$

Donc nous voyons que finalement:

$$e^{-\alpha \tau - \beta x} = e^{-\frac{1}{2} \left(\kappa + \frac{1}{2} \right)^2 \tau - \left(\kappa - \frac{1}{2} \right) x} \quad (66.653)$$

Il nous reste maintenant à transformer la condition en u :

$$u(0, x) = (e^x - 1)^+ \quad (66.654)$$

En une condition initiale en v à l'aide de:

$$v(\tau, x) = e^{\frac{1}{2}\left(\kappa + \frac{1}{2}\right)^2 \tau + \left(\kappa - \frac{1}{2}\right)x} u(\tau, x) \quad (66.655)$$

Donc il vient immédiatement:

$$\begin{aligned} v(0, x) &= e^{\frac{1}{2}\left(\kappa + \frac{1}{2}\right)^2 0 + \left(\kappa - \frac{1}{2}\right)x} u(0, x) = e^{\left(\kappa - \frac{1}{2}\right)x} \cdot (e^x - 1)^+ = \begin{pmatrix} e^{\left(\kappa - \frac{1}{2}\right)x + x} & -e^{\left(\kappa - \frac{1}{2}\right)x} \end{pmatrix}^+ \\ &= \begin{pmatrix} e^{\left(\kappa + \frac{1}{2}\right)x} & -e^{\left(\kappa - \frac{1}{2}\right)x} \end{pmatrix}^+ \end{aligned} \quad (66.656)$$

La suite (la résolution de l'équation différentielle) viendra dans un proche avenir... sauf accident (pas le temps de la rédiger...).

- Pour le Call européen (valeur de l'option d'achat de maturité T et de strike K) la solution (dont la démonstration doit encore être rédigée dans ce chapitre...) est:

$$C(S, t) = SN(d_1) + Ee^{-r(T-t)}N(d_2) \quad (66.657)$$

où $N(x)$ est donc la loi Normale centrée réduite:

$$N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}y^2} dy \quad (66.658)$$

avec:

$$d_1 = \frac{\ln\left(\frac{S}{E}\right) + \left(r + \frac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}} \quad (66.659)$$

et:

$$d_2 = d_1 - \sigma\sqrt{T-t} \quad (66.660)$$

La distribution Normale cumulée de ce paramètre représente la probabilité que l'option soit exercée dans un univers risque-neutre (en absence d'opportunité d'arbitrage pour faire simple). Multiplié par E , la valeur normale cumulée du paramètre précédent représente donc en quelque sorte l'espérance, en univers risque-neutre, de paiement du prix d'exercice. L'exponentielle se trouvant dans l'expression de $C(S,t)$ est le facteur d'actualisation.

Remarque: Si nous avons un Call ou Put à la monnaie (pour rappel cela signifie que $S = E$) alors les deux coefficients ci-dessus se simplifient puisque le logarithme népérien est alors nul.

- Pour le Put européen (valeur de l'option de vente de maturité T et strike K):

$$P(S,t) = -SN(-d_1) + Ee^{-r(T-t)}N(-d_2) = S(N(d_1) - 1) - Ee^{-r(T-t)}(N(d_2) - 1) \quad (66.661)$$

Dès lors, le "**delta du Call**" que nous avons déjà introduit plus haut est donné dans ce modèle par l'expression exacte (c'est immédiat):

$$\Delta_C = \frac{\partial C}{\partial S} = N(d_1) \quad (66.662)$$

qui est donc dans ce cas particulier une valeur positive comprise entre 0 et 1.

Le "**delta du Put**" est lui donné par (c'est aussi immédiat):

$$\Delta_P = \frac{\partial P}{\partial S} = N(d_1) - 1 \quad (66.663)$$

qui est donc toujours une valeur négative.

Il est facile de vérifier que ces solutions satisfont l'équation de parité Put-Call:

$$S + P - C = Ee^{-r(T-t)} \quad (66.664)$$

où rappelons que E est juste une autre manière traditionnelle de noter la valeur de strike K des deux options P et C .

Voici les commandes intégrées à la version anglaise de Microsoft Excel 14.0.6129 pour faire le calcul avec des valeurs d'entrées arbitraires:

	A	B	C	D	E	F	G
1	S	292.85	Prix du sous-jacent à ce jour				
2	E	300.00	Prix d'Exercice (strike) fixé à l'avance				
3	T	0.25	Maturité en années (inférieur ou égal à 0.25 comme limite théorique)				
4	r	4%	Rendement géométrique moyen annuel non risqué du marché				
5	σ	40%	Volatilité annuelle du sous-jacent				
6							
7	d_1	=(LN(B1/B2)+(B4+0.5*(B5^2))*B3)/(B5*SQRT(B3))					
8	d_2	=B7-B5*SQRT(B3)					
9							
10	$C(S, T)$	=B1:C1*NORM.S.DIST(B7;1)-B2*EXP(-B4*B3)*NORM.S.DIST(B8;1)					
11	$P(S, T)$	=-B1*NORM.S.DIST(-B7;1)+B2*EXP(-B4*B3)*NORM.S.DIST(-B8;1)					
12							
13	Delta Call	=NORM.S.DIST((LN(B1/B2)+(B4+0.5*B5^2)*B3)/(B5*SQRT(B3));1)					

Figure: 66.62 -

Application de Black & Scholes dans la version anglaise de Microsoft Excel 14.0.6129

Remarque: Il est sûr que les équations de Black & Scholes ont permis l'essor des marchés aux options, en permettant une spéculation sécurisée. Cela reste de la spéculation (les acteurs spéculent les uns par rapport aux autres sur la volatilité des actions), mais cette spéculation reste sécurisée par l'équation de couverture, qui évite que les pertes ne soient trop importantes. Il existe néanmoins des inconvénients à leur utilisation. Le plus important est sûrement l'effet d'emballement qu'elles provoquent. Supposons par exemple que vous êtes le vendeur d'une option sur l'action d'une société S. Celle-ci annonce des résultats légèrement inférieurs à ceux attendus. Son cours baisse, et c'est normal. L'équation de couverture de Black & Scholes vous recommande alors de diminuer le nombre d'actions de cette société dans votre portefeuille, ce que vous faites. Mais tous les acteurs du marché font le même raisonnement, engendrant une nouvelle baisse du cours de l'action. L'équation de couverture de Black and Scholes vous recommande de vendre encore des actions, etc.... Cela peut déclencher un véritable emballement du marché, à la baisse comme à la hausse. Ceci est accentué par le fait que bien souvent, les ordres d'achat ou de vente sont automatisés, implémentés directement dans les logiciels, et ne nécessitent plus d'interventions humaines. D'autre part, l'équation de couverture de Black & Scholes est efficace pour de petites variations de cours, mais pas pour des "déviations" brutaux et importants. Ainsi, un an à peine après avoir reçu leur prix Nobel d'Économie, Robert Merton et Myron Scholes furent impliqués dans la déconfiture du fonds d'investissement américain LTCM à l'automne 1998, à la suite de la grave crise russe de l'été 1998.

7.9. VALUE AT RISK

Les mesures du risque ont bien évolué depuis que Markowitz a avancé sa célèbre théorie de la diversification de portefeuille à la fin des années 1950, théorie qui devait révolutionner la gestion de portefeuille moderne. Le risque d'un portefeuille était alors relié à la matrice des covariances-variances comme nous l'avons démontré théoriquement et par l'exemple plus haut.

Dans les années 1960, Sharpe a proposé le modèle unifactoriel d'évaluation des actifs financiers où le bêta est le facteur explicatif principal du risque d'un portefeuille via la matrice des bêtas.

Au début des années 1990, une nouvelle mesure du risque a fait son entrée (la banque JP Morgan en serait à l'origine). En effet, on reconnaissait de plus en plus les limites des mesures traditionnelles du risque. Il fallait se donner des mesures du risque de baisse de la valeur des actifs. Pour ce faire, il fallait trouver des mesures davantage reliées à l'ensemble de la distribution des flux monétaires d'un portefeuille. C'est dans ce contexte qu'une mesure nominale du risque a été proposée: la *VaR* (notée *V@R* en gestion de projets). Dont l'idée est en gros de pouvoir dire que nous avons *X*% de probabilité cumulée de ne pas perdre plus de *Y* en numéraires dans les *N* prochains jours.

Cette nouvelle mesure a d'abord servi à quantifier le risque de marché auquel sont soumis les portefeuilles bancaires. En effet, l'Accord de Bâle a recommandé aux banques, en 1997, de détenir un montant de capital réglementaire pour pallier aux risques standards de marché. Or, ce capital est depuis lors calculé à partir de la *VaR* et celle-ci est devenue de plus en plus populaire pour évaluer le risque de portefeuilles institutionnels ou individuels (et pas que!). Il n'existe pas cependant une mesure unique de la *VaR*. En effet, elle repose sur le concept de volatilité, qui est essentiellement latent. C'est pourquoi les banques se doivent de recourir à plusieurs modèles de *VaR* de manière à définir la fourchette de leurs pertes éventuelles. Ces calculs sont d'autant plus complexes que la distribution des rendements des titres mesurés à haute fréquence s'éloigne sensiblement de la normale.

Définition: La "**Value at Risk**" (*VaR*) est la perte maximale théorique que peut subir un gestionnaire d'un portefeuille (dont la valeur est forcément implicitement variable) et pour une certaine période de temps avec une probabilité cumulée donnée (l'utilisation de la *VaR* n'est pas limitée aux instruments financiers, elle est utilisée dans beaucoup d'autres domaines de la gestion du risque en général).

Remarque: La *VaR* n'est pas réellement pertinente si elle n'est pas présentée avec d'autres indicateurs de risques tels que le ratio de Sharpe, le ratio de Treynor ou encore les coefficients des grecques (comme le bêta). Enfin, indiquons que dans la pratique la *VaR* est parfois indiquée en %.

7.9.1. VAR RELATIVE

Dans le modèle classique de la *VaR* relative (appelée aussi parfois "**VaR Paramétrique**"), nous supposons que la distribution statistique des résultats d'un portefeuille obéit à chaque instant à une loi Normale... que nous noterons par la suite:

$$X = N(\mu, \sigma^2) \quad (66.665)$$

Sous cette hypothèse de normalité, la *VaR* relative est appelée en toute rigueur "**VaR delta-normale**".

L'idée suivante est que la variable aléatoire *X* peut donc être réécrite avec une variable aléatoire Normale centrée réduite (cf. chapitre de Statistiques) en posant:

$$\varepsilon = N(0,1) \quad (66.666)$$

telle que (utilisation des propriétés de base de la loi Normale):

$$X = \mu + \varepsilon\sigma \quad (66.667)$$

et cette écriture est donc utilisée dans énormément d'autres domaines que la finance (gestion de projets, assurance qualité, logistique, etc.).

Soit α le seuil critique associé à la probabilité cumulée visée. Nous pouvons alors écrire:

$$X = \mu + \alpha\sigma \quad (66.668)$$

qui est une forme intéressante car elle reporte l'analyse du risque et de la variabilité sur l'estimation de l'écart-type seul (ce que les financiers apprécient bien...)!

Cette forme d'écriture se vérifie aisément avec la version anglaise de Microsoft Excel 11.8346 pour les sceptiques... Considérons un portefeuille P ayant un écart-type annuel de 10% et supposons que nous possédons 1'000.- en actifs de ce portefeuille. Nous avons alors à la première année:

$$= \text{NORMINV}(99\%;1000;10\%*1000)=1000+\text{NORMSINV}(99\%)*10\%*1000=1'232.6$$

Soit 99% de probabilité cumulée d'avoir un portefeuille valant entre 0 et 1'232.6.- à tout moment (on considère comme négligeable la probabilité cumulée que le portefeuille ait une valeur négative avec cette écriture).

Mais ce qui intéresse le gestionnaire n'est pas de se couvrir du risque de l'espérance (car il est nul) mais de la volatilité seule! Dans le cas précédent elle est donc de 100.- et suit une loi Normale centrée réduite. D'où la raison de définir la VaR formellement comme étant la relation mathématique qui donne un intervalle de confiance (ou une quantile selon le point de vue) de l'écart-type:

$$VaR = \alpha\sigma \quad (66.669)$$

Ainsi, pour une probabilité cumulée de 99% les logiciels nous donnent en valeur absolue (voir le traitement des intervalles de confiance dans le chapitre de Statistiques):

$$\alpha = \text{NORMSINV}(99\%) = \text{ABS}(\text{NORMSINV}(1\%)) = 2.326 \quad (66.670)$$

où par tradition les financiers prennent l'alpha (et donc la VaR) comme étant positif. D'où le fait qu'ils parlent de risque couvert à 99% alors qu'en réalité il s'agit de couvrir un risque qui a 1% de probabilité cumulée d'avoir lieu (mais strictement parlant c'est la même chose simplement que le premier est plus facile à faire comprendre à un client...!!!). Raison pour laquelle on trouve parfois aussi la VaR sous la forme suivante:

$$VaR = -\alpha\sigma \quad (66.671)$$

Signalons que le résultat du calcul peut être désigné aussi sous le nom de "**réserve fractionnaire**" (bien que cela n'ait aucun rapport avec la façon dont les États imposent une réserve fractionnaire aux banques) ou encore "**réserve mathématique**" (bien que cela n'ait aucun rapport avec la manière dont les assurances calculent la réserve mathématique).

Remarque: Le modèle RiskMetrics de J.P. Morgan/Reuters proposait en 1996 de prendre un α de 1.65 correspondant à une probabilité cumulée de 95%. Certains praticiens recommandent fortement, au vu de l'hypothèse forte et simplificatrice de Normalité..., de multiplier la VaR par un facteur 2 ou 3.

 Exemple:

Un portefeuille P de valeur 1'000.- a une volatilité annuelle de 10%. La volatilité journalière (instantanée) du rendement est alors de (nous utilisons ici la propriété du mouvement brownien standard):

$$\sigma_{R,j} = \frac{10\%}{\sqrt{252}} \cong 0.63\% \quad (66.672)$$

où 252 est le nombre de jours de Bourse dans l'année dans un pays donné. Soit en numéraires:

$$\sigma_j \cong 0.63\% \cdot 1'000 \cong 6.29 \quad (66.673)$$

La VaR relative au seuil de 99% à une journée est alors:

$$VaR_j = 2.326 \cdot 6.29 \cong 14.65 \quad (66.674)$$

De même sur un an nous aurions la VaR relative annuelle au seuil de 99% suivante:

$$VaR_A = 2.326 \cdot 1'000 \cdot 10\% = 232.6 \quad (66.675)$$

Soit une VaR relative de 23.26% (juste histoire de la donner en pourcents comme il est d'usage dans le domaine financier).

Ainsi, en ce qui concerna la VaR annuelle relative, nous avons alors 99% de probabilité cumulée de gagner 232.60.- mais aussi de les perdre! Effectivement nous avons 1% de probabilité cumulée d'avoir une perte annuelle de:

$$= \text{NORMSINV}(1\% \cdot 10\% \cdot 1000) = -232.6$$

donc il faudrait au moins un capital risque (fonds propres) de 232.6.- pour couvrir 99% des risques (couvrir cette probabilité cumulée de 1% d'être dans une mauvaise année respectivement). Nous pouvons aussi dire que nous avons 99% de probabilité cumulée de ne pas perdre plus 232.6.-. Nous retrouvons donc le même résultat numérique qu'avec l'exemple précédent.

Le lecteur remarquera que nous avons donc dans le domaine de la Bourse (ceci découle donc du mouvement brownien standard) pour passer d'un horizon temporel à un autre:

$$VaR_A = VaR_j \sqrt{t} = 2.326 \cdot \sigma_j \sqrt{252} \quad (66.676)$$

Les financiers appellent cette propriété du mouvement brownien dans le cadre de l'utilisation de la VaR la "**scaling law**". Elle est autorisée par les accords de Bâle en 1996 qui présupposent une distribution Normale et conseillent un horizon temporel de 10 à 30 jours. Nous avons vu cependant lors de notre démonstration du modèle du mouvement brownien standard que nous sous-estimons sous cette hypothèse le risque réel et que ce réflexe de changement d'échelle via la racine carrée est très critiquée par les spécialistes.

Donc la *VaR* bien qu'étant un montant placé dans les capitaux propres (provisions pour risques et charges) afin de faire face à un décrochage violent mais temporaire du marché, ne suffit pas à se protéger d'un retournement puissant et durable du marché... Raison pour laquelle les banques provisionnent souvent le minimum minimorum...

Remarque: Personnellement je préconiserai de couvrir selon la méthode Six Sigma à 99.9996% sur un horizon temporel correspondant au minimum au temps de position moyen. Mais c'est personnel...

7.9.2. VAR ABSOLUE

La mesure de *VaR* que nous venons de donner est une mesure relative car elle ne tient pas compte de la moyenne des pertes et gains futurs.

Si la volatilité est de 100.- dans l'exemple qui vient d'être donné, la *VaR* relative est donc 232.6.- Mais comme le profit moyen est généralement non nul sur une longue période de temps, nous devons la plupart du temps utiliser la mesure absolue de la *VaR* (sur une très courte période le profit étant considéré comme parfois nul, on s'en tient au calcul de la *VaR* relative).

Rappelons d'abord que suite à notre étude du modèle de Bachelier nous avons démontré que l'espérance positive de la valeur (ou rendement) ainsi que l'écart-type positif d'un portefeuille sont proportionnels à la racine carrée du temps.

Supposons que la période d'observation t soit en mois. Le rendement mensuel espéré pour le portefeuille de valeur initiale S est alors de μ (son espérance donc..!) et la variance mensuelle de son rendement de σ^2 .

Sa *VaR* relative au seuil de confiance α est donc après t mois donnée par (vous pouvez vérifier que la relation est bien homogène!):

$$VaR_r = S \cdot (\alpha \sigma \sqrt{t}) \quad (66.677)$$

comme nous avons pu le vérifier dans l'exemple précédent (donc jusqu'ici rien de nouveau...). La racine carrée du temps provient, pour rappel, du modèle de Bachelier (mouvement brownien standard).

En version de Monte-Carlo avec du VBA (car la version ci-dessus est dite "ponctuelle") cela donne:

```
Function ValueAtRiskMC(confidence, horizon, RiskFree, StdV, StockValue)
Dim i As Integer
Dim stockReturn(1 To 10000) As Double
For i = 1 To 10000
    stockReturn(i) = Exp((RiskFree - 0.5 * StdV ^ 2) _
        + StdV * Application.NormInv(Rnd(), 0, 1)) - 1
Next i

ValueAtRiskMC = (horizon) ^ 0.5 * StdV _
    * Application.Percentile(stockReturn, 1 - confidence)

ValueAtRiskMC = StockValue * ValueAtRiskMC
End Function
```

Remarque: Contrairement à ce que nous avons vu lors de notre étude des seuils/intervalles de confiances dans le chapitre de Statistiques, nous ne divisons pas par 2 l'argument de la fonction NORMALSINV() de la version anglaise de Microsoft Excel 11.8346 pour obtenir le α dans la situation ci-dessus car ce qui nous intéresse c'est seulement un côté de la courbe centrée réduite (le côté "pessimiste") et non les deux.

Si nous reprenons le même exemple que précédemment (portefeuille de 1'000.- avec 10% de volatilité annuelle). La VaR relative est donc sur une projection de 30 jours de:

$$VaR_r = S \alpha \sigma \sqrt{t} = 1000 \cdot 2.326 \cdot \frac{10\%}{\sqrt{252}} \cdot \sqrt{30} \cong 80.2 \quad (66.678)$$

Mais cette dernière relation ne tient pas compte du rendement moyen espéré μ du portefeuille dans le temps. La VaR absolue est donc obtenue en retranchant ce rendement à la VaR relative sur la même période temporelle, c'est-à-dire:

$$VaR_a = VaR_r - S \cdot \mu \cdot t = -S \left(\mu \cdot t - \alpha \sigma \sqrt{t} \right) \quad (66.679)$$

où nous faisons l'hypothèse particulière que le rendement est donc linéairement dépendant du temps (conformément à la construction semi-empirique du mouvement brownien standard). La VaR absolue est donc bien évidemment inférieure à la VaR relative de ce montant.

Mentionnons que le calcul de la VaR absolue peut être considéré comme vicieux ou ayant peu d'intérêt car il suppose que le gain obtenu grâce au rendement sera placé dans les fonds propres pour financer la VaR relative. Or, dans la majeure partie des cas les gains seront remplacés autrement.

Reprenons quand même notre exemple habituel sous cette hypothèse (portefeuille de 1'000.- avec 10% de volatilité annuelle) avec un rendement annuel de 15%. Nous avons alors:

$$VaR_a = S \alpha \sigma \sqrt{t} - S \mu t = 1'000 \cdot 2.326 \cdot 10\% - 1'000 \cdot 15\% \cong 232.6 - 150 = 82.6 \quad (66.680)$$

Concrètement, si nous finançons la VaR avec les gains alors sur une année il suffit d'avoir 82.6.- de fonds propres. Dans la pratique il peut être intéressant de savoir à partir de combien de temps les gains couvrent la totalité de la VaR. Dans ce cas il s'agit d'une simple équation du deuxième degré telle que:

$$S \alpha \sigma \sqrt{t} - S \mu t = 0 \quad (66.681)$$

et nous trouverions dans notre exemple 2.4 années. Concrètement après 2.4 années les gains auront couvert la totalité des risques selon les hypothèses de construction...

7.9.3. VAR HISTORIQUE

Une troisième manière pragmatique de calculer la VaR relative est basée sur les données historiques. Il s'agit de la manière la plus simple de faire le calcul avec la facilité d'utilisation des tableurs existant aujourd'hui.

Supposons pour l'exemple que nous ayons les cent dernières performances journalières d'un portefeuille. Les dix plus mauvaises performances journalières sont données ci-contre par ordre croissant:

Données Historiques
-19'000
-16'450
-15'000
-12'500
-11'950
-11'250
-11'050
-10'600
-10'500
-10'250
...

Tableau: 66.16 - Performances journalières triées d'un portefeuille

La VaR relative à 95% pour 1 jour consiste alors à déterminer le 5ème centile. Comme nous avons 100 échantillons, il est facile de déterminer qu'il s'agit de la 5ème valeur dans l'ordre croissant des valeurs. Donc:

$$VaR_{\tau} = 11'950. - \quad (66.682)$$

Comme nous l'avons déjà mentionné dans le chapitre de Statistiques, dans les tableurs nous utilisons la fonction CENTILE() qui n'est pas forcément calculée de la même manière d'un logiciel à l'autre.

7.9.4. VAR DE CRÉDIT

Pour introduire ludiquement la VaR de crédit Considérons trois portefeuilles, A , B et C dont les probabilités de défaut de crédit suivent une loi de Bernoulli (donc pour rappel c'est un événement binaire) de valeurs respectives 5%, 10% et 20% et dont les valeurs numéraires sont en millions de francs de 25.-, 30.- et 45.-.

Pour l'exemple qui suivra, nous ferons les hypothèses (qui sont souvent implicites même dans tout ce que l'on a vu jusqu'à maintenant) que l'exposition au risque est constante, que les portefeuilles sont indépendants et qu'en cas de défaut de crédit nous perdons tout.

Comme les trois portefeuilles sont donc indépendants, l'espérance de perte (notée " L " pour Loss) est facilement calculable à l'aide de la propriété de linéarité de l'espérance:

$$E(L) = 5\% \cdot 25 + 10\% \cdot 30 + 20\% \cdot 45 = 13.25. - \quad (66.683)$$

Comme la perte de crédit est considérée comme une variable aléatoire de Bernoulli (et que les portefeuilles sont toujours indépendants), nous avons alors toujours en millions de francs (se référer au calcul de la variance d'une variable aléatoire de type Bernoulli dans le chapitre de Statistiques):

$$\begin{aligned} V(L) &= V(L_A + L_B + L_C) = V(L_A) + V(L_B) + V(L_C) \\ &= 25 \cdot 5\% \cdot (1 - 5\%) + 30 \cdot 10\% \cdot (1 - 10\%) + 45 \cdot 20\% \cdot (1 - 20\%) = 434.688. - \end{aligned} \quad (66.684)$$

Bon ceci étant fait pour le plaisir... Comme déterminer le plus finement possible la distribution des pertes pour avoir la VaR de crédit à un seuil donné? Eh bien pour cela il va nous falloir construire le tableau de toutes les issues possibles.

Scénario perte	Perte associée	Probabilité	Probabilité cumulée
Aucun	0.-	$(1-0.05) \times (1-0.1) \times (1-0.2) = 68.40\%$	68.40%
A	25.-	$0.05 \times (1-0.1) \times (1-0.2) = 3.60\%$	72.00%
B	30.-	$(1-0.05) \times 0.1 \times (1-0.2) = 7.60\%$	79.60%
C	45.-	$(1-0.05) \times (1-0.1) \times 0.2 = 17.10\%$	96.70%
A, B	55.-	$0.05 \times 0.1 \times (1-0.2) = 0.40\%$	97.10%
A, C	70.-	$0.05 \times (1-0.1) \times 0.2 = 0.90\%$	98.00%
B, C	75.-	$(1-0.05) \times 0.1 \times 0.2 = 1.90\%$	99.00%
A, B, C	100.-	$0.05 \times 0.1 \times 0.2 = 0.1\%$	100%

Tableau: 66.17 - Scénarios de pertes et probabilités associées

Ainsi, la VaR de Crédit la plus proche de 95% est de 45 millions de francs. Puisque nous pouvons nous attendre à une perte de 13.25 millions de francs, la partie non attendue sera donc considérée comme étant de 31.75 millions de francs.

7.9.5. VAR OPÉRATIONNELLE

Nous avons déjà rencontré le concept de convolution plusieurs fois dans le chapitre de Statistiques mais pour des lois continues connues explicitement sous leur forme algébrique.... Dans la pratique il en est tout autre! Ainsi, les actuaires ont souvent à leur disposition pour une période donnée (souvent sur une année), une distribution expérimentale des fréquences des événements d'intérêt (accidents, décès ou autre) et une distribution expérimentale des coûts des événements d'intérêt. Par exemple, considérons le cas très simplifié (en taille!) suivant:

Distribution Fréquence (pertes)		Distribution des coûts de perte	
Probabilité	Fréquence	Probabilité	Perte
0.6	0	0.5	1'000.-
0.3	1	0.3	10'000.-
0.1	2	0.2	100'000.-
Espérance:	0.5	Espérance:	23'500.-

Tableau: 66.18 - Fréquence d'événements de pertes et probabilités et probabilités de perte

Si nous souhaitons à partir de ces valeurs expérimentales et discrètes déterminer la distribution empirique des coûts, nous allons devoir faire une convolution discrète dont le principe est le suivant:

Nombre de pertes	Première perte	Deuxième perte	Perte totale	Probabilité
0	0	0	0	=0
1	1'000	0	1'000	=0.3×0.6=0.150
1	10'000	0	10'000	=0.3×0.3=0.090
1	100'000	0	100'000	=0.3×0.2=0.060
2	1'000	1'000	2'000	=0.1×0.5×0.5=0.025
2	1'000	10'000	11'000	=0.1×0.5×0.3=0.015
2	1'000	100'000	101'000	=0.1×0.5×0.2=0.010
2	10'000	1'000	11'000	=0.1×0.3×0.5=0.015
2	10'000	10'000	20'000	=0.1×0.3×0.3=0.009
2	10'000	100'000	110'000	=0.1×0.3×0.2=0.006
2	100'000	1'000	101'000	=0.1×0.2×0.5=0.010
2	100'000	10'000	110'000	=0.1×0.2×0.3=0.006
2	100'000	100'000	200'000	=0.1×0.2×0.2=0.004

Tableau: 66.19 - Convolution distribution des événements de pertes et valeurs associées

L'espérance des pertes est alors égale à 11'750.-, ce qui équivaut aussi à calculer simplement (l'espérance du produit de deux variables aléatoires indépendantes étant égal au produit des espérances comme nous l'avons démontré dans le chapitre de Statistiques):

$$0.5 \cdot 23'500 = 11'750 \quad (66.685)$$

Ensuite, nous combinons et trions les pertes et probabilités:

Perte	Probabilité cumulée
0	60%
1'000	75%
2'000	77.5%
10'000	86.5%
11'000	89.5%
20'000	90.4%
100'000	96.4%
101'000	98.4%
110'000	99.6%
110'000	100%

Tableau: 66.20 - Combinaison des calculs

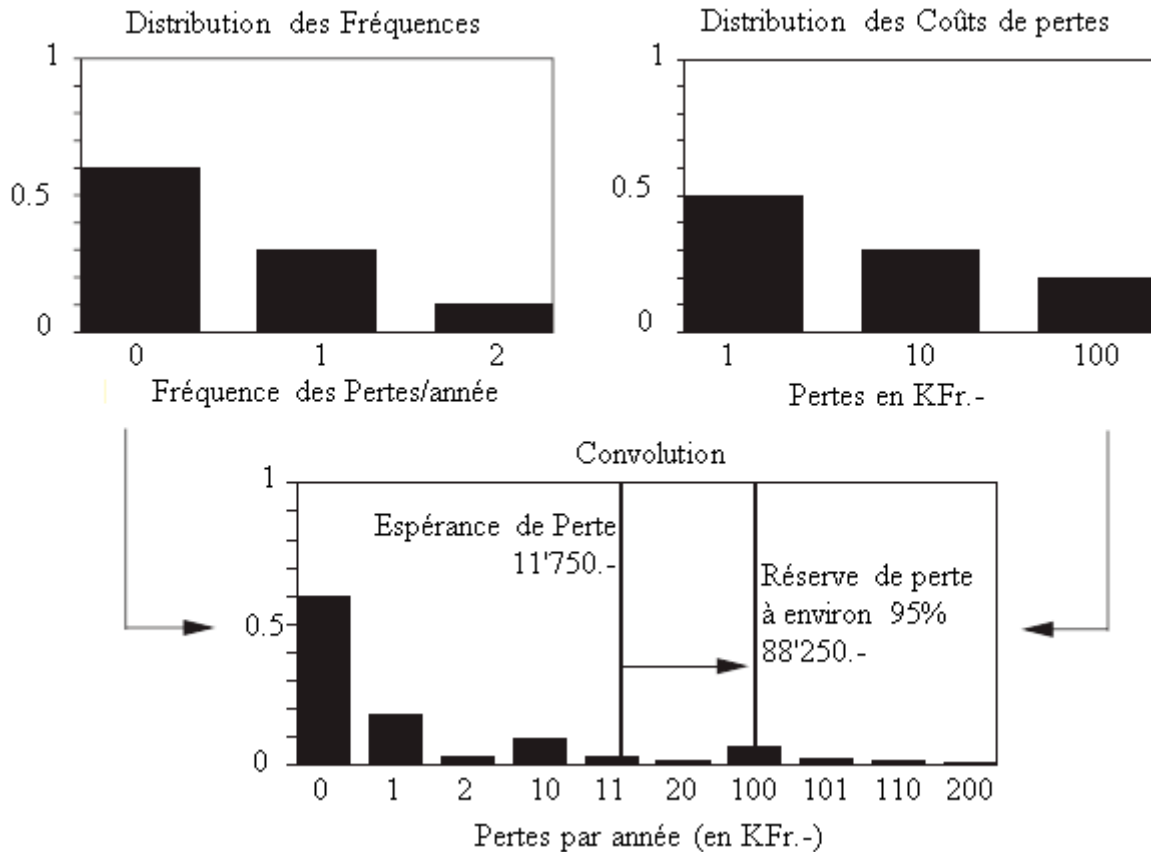
Du côté de la *VaR* opérationnelle, si on la pose à 95% (les accords de Bâle II propose plutôt de prendre 99.9%), la valeur la plus proche est alors:

$$VaR_{95} = 100'000.-$$

Dès lors, la réserve de perte non attendue est de:

$$100'000 - 11'750 = 88'250.-$$

Graphiquement nos tableaux, manipulations et calculs se résument à:



Il est au passage très intéressant de remarquer comment deux simples tables produisent une quantité de calculs en convolution relativement conséquent. Raison pour laquelle dans des cas pratiques réels, il faut faire appel à l'informatique.

7.9.6. VAR VARIANCE-COVARIANCE

La *VaR* variance-covariance est basée sur un cas plus réaliste du calcul de la *VaR* pour un portefeuille composé de plusieurs actifs financiers corrélés ou non (contrairement aux cas précédents où nous n'avions qu'un seul actif).

Considérons pour introduire ce concept un portefeuille *P1* de 5'000'000.- de volatilité journalière de 2% (soit de 100'000.-/j.) et un deuxième portefeuille *P2* de 7'000'000.- de volatilité journalière de 1% (soit de 70'000.-/j.).

Nos mesures montrent que leur coefficient de corrélation $R_{1,2}$ est de 0.6. L'écart-type global journalier est alors de (cf. chapitre de Statistiques):

$$\sigma_{1+2} = \sqrt{\sigma_1^2 + \sigma_2^2 + 2R_{1,2}\sigma_1\sigma_2} = \sqrt{100'000^2 + 70'000^2 + 2 \cdot 0.6 \cdot 100'000 \cdot 70'000} \quad (66.686)$$

$$\cong 152'643$$

Ainsi, la *VaR* relative journalière à 99% pour le portefeuille global est de (pas de scaling law à appliquer ici puisque l'écart-type est journalier et que nous voulons la *VaR* relative journalière):

$$VaR_{\gamma,1+2} = \alpha\sigma_{1+2} = 2.326 \cdot 152'634 \cong 355'000 \quad (66.687)$$

Il est intéressant de comparer la VaR relative journalière à la somme des VaR relatives des deux portefeuilles:

$$VaR_{r,1} + VaR_{r,2} = \alpha\sigma_1 + \alpha\sigma_2 = 2.326 \cdot 100'000 + 2.326 \cdot 70'000 \cong 395'420 \quad (66.688)$$

Nous avons:

$$VaR_{r,1+2} \leq VaR_{r,1} + VaR_{r,2} \quad (66.689)$$

Ceci est dû au gain de diversification!

Remarque: Un piège dans le calcul de la VaR relative variance-covariance aurait été de calculer l'écart-type global en % et ensuite de l'appliquer dans la relation du calcul de la VaR globale. Le résultat aurait dès lors été erroné!

Rappelons pour clore à quoi sert la VaR ? Mentionnons d'abord qu'elle se révèle d'une grande utilité puisqu'elle est mesurée en termes nominaux. Une fois qu'une institution financière a calculé sa VaR globale, c'est-à-dire la perte maximale qu'elle peut encourir sur l'ensemble de son bilan pour une probabilité prédéterminée, il lui est loisible de se servir de ce montant pour déterminer le capital (avoir propre) minimal qu'elle doit détenir pour ne pas s'exposer à la faillite. Si en effet elle détient un capital moindre et que la perte maximale probabiliste se produit, son avoir propre sera négatif et elle devra peut-être déposer son bilan.

La VaR est donc très utile pour une institution financière, car elle lui permet de déterminer le niveau du capital qu'elle doit maintenir pour survivre. Quand la VaR est utilisée à cette fin, nous l'appelons plus communément CaR pour "**Capital at Risk**", c'est-à-dire que le capital que doit détenir une institution financière est calculé ou évalué selon les risques auxquels elle est exposée. Plus le risque est important, plus elle devra maintenir un capital élevé. Cela apparaît bien raisonnable, car le capital détenu par une institution financière est d'abord et avant tout un filet de sécurité. Pour une banque, il vise à protéger les dépôts à son passif. La VaR se présente donc comme une mesure appropriée pour définir le capital réglementaire que doit détenir une institution financière. C'est pourquoi le Comité de Bâle, chapeauté par la Banque des Règlements Internationaux, retenait cette mesure pour calculer le capital réglementaire d'une institution de dépôts en 1995 et qui est devenue effective en janvier 1998. Celles-ci doivent maintenant calculer leur exposition au risque en recourant à la VaR et tester sa justesse en faisant des "stress tests" (confronter les calculs à des variations extrêmes) ainsi qu'à des "back testing" en vérifiant que les grandes déviations (en dehors de l'intervalle de confiance) n'ont pas lieu plus de 5 fois par année boursière (donc par comparaison aux données historiques).

7.9.7. BACK-TESTING DE LA VAR (MODÈLE BINOMIAL)

Comme mentionné précédemment, le "back-testing" est une méthode très fréquente utilisée par les traders pour contrôler a posteriori de l'efficacité d'un modèle prévisionnel ou d'une stratégie de trading.

Nous allons voir ici un modèle naïf de back-testing de la VaR (il en existe un grand nombre) mais lorsque le lecteur parcourra plus loin les modèles prévisionnels (moyenne mobile, lissage exponentiel, ARIMA, etc.) nous verrons que nous comparerons ces modèles avec les mêmes valeurs en utilisant des mesures d'erreurs (ME , MAD , MDS , MPE). Ces comparaisons sont considérées alors aussi comme une forme de "back-testing".

Donc pour en revenir à la *VaR*, il est relativement évident que lorsque l'on vient de mettre en place un modèle de *VaR* par exemple annuel et que l'on a peu d'historique il vaut mieux la ramener à une *VaR* journalière et faire le back-testing sur la base d'une plus petite unité.

Rappelons l'exemple du début où nous avons une *VaR* relative journalière et respectivement annuelle au seuil de 1% (donc couverture à 99%) de:

$$\begin{aligned} VaR_j &\cong 14.65. - \\ VaR_A &\cong 232.6. - \end{aligned} \quad (66.690)$$

Supposons que nous faisons un back-testing de ce modèle en le comparant aux données de 600 journées que nous avons en notre possession. Nous observons 9 valeurs au-dessus de 14.65.-. Que pouvons-nous en conclure???

Eh bien il faut d'abord savoir que si les N journées sont indépendantes et que nous notons par p la probabilité qu'un des journées soit hors de la *VaR* (donc valeur associée égale à 1% dans notre exemple!), alors nous avons affaire à une expérience de N épreuves de Bernoulli. Il s'agit donc d'une loi Binomiale dont la probabilité cumulée associée est pour rappel (cf. chapitre de Statistiques):

$$\sum_{k=0}^N C_k^N p^k (1-p)^{N-k} = \sum_{k=0}^N \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k} \quad (66.691)$$

Dès lors, l'espérance est déjà donnée par (c'est sympa de la calculer mais en réalité ce n'est pas très utile):

$$E(X) = Np = 600 \cdot 1\% = 6 \quad (66.692)$$

Cependant pour connaître la probabilité cumulée associée à 9 événements, nous calculons par exemples avec Microsoft Excel 14.0.6129 en anglais:

$$=BINOM.DIST(9;600;1\%;TRUE)=BINOM.DIST(9;600;1\%;1)=91.71\%$$

Soit une p -value de 8.29%. Donc que ce soit en unilatéral ou bilatéral, ce test d'hypothèse nous amène à ne pas rejeter le modèle de la *VaR*. Il ne faut pas oublier que si N est très grand et que la probabilité est petite, la loi binomiale tend vers une loi Poisson (cf. chapitre de Statistiques) qui est plus simple à manipuler (c'est la recommandation du standard CreditRisk+).

8. ANALYSE DES SÉRIES TEMPORELLES (A.S.T)

Contrairement à l'économétrie traditionnelle, le but de "l'analyse des séries temporelles" (A.S.T.) n'est pas de relier des variables entre elles, mais de s'intéresser à la dynamique d'une variable dans le temps pour découvrir certaines régularités afin de pouvoir extrapoler ou d'établir des prévisions sous réserve de l'hypothèse qu'on puisse relier une observation à celles qui l'ont précédée. Avec une analyse fine, il est même possible d'établir des prévisions "robustes" vis-à-vis de ruptures brusques et de changements non-anticipables.

Remarque: Ce domaine est un souvent appelé "Business Intelligence" dans les entreprises. Une variable analysée sous forme d'A.S.T. sera elle appelée un "Indice de Performance Clé" (IPC) et un ensemble d'IPC un "tableau de bord".

Définition: Une "**série temporelle à temps discret**" ou "**série chronologique à temps discret**" (plus rigoureusement on devrait parler de "suite"!) est une suite d'observations X_t d'une variable y à différentes dates t . Habituellement l'espace de base de t est dénombrable, de sorte que $t = 1 \dots T$. Le tout étant noté:

$$\{X_t\}_{t=1}^T \quad (66.693)$$

et parfois le Y majuscule est noté par un X (cela importe peu...).

Une série temporelle est donc toute suite d'observations correspondant à la même variable: il peut s'agit de données macroéconomiques (le PIB d'un pays, l'inflation, les exportations), microéconomiques (les ventes d'une entreprise donnée, son nombre d'employés, le revenu d'un individu, ...), financières (le CAC40, le prix d'une option d'achat ou de vente, le cours d'une action), météorologiques (la pluviosité, le nombre de jours de soleil par an...), politiques (le nombre de votants, de voix reçues par un candidat...), démographiques (la taille moyenne des habitants, leur âge...).

En pratique, tout ce qui est chiffrable et varie en fonction du temps peut être analysé relativement pertinemment sous forme de AST tant que la personne qui manipule le modèle sait ce qu'elle fait (ce qui est rare dès qu'on sort du domaine public et étatique...).

Ainsi en finance on utilise ces modèles pour modéliser les rendements mais aussi la volatilité (écart-type) et la *VaR*. Bref, il n'y pas vraiment de limites à l'application.

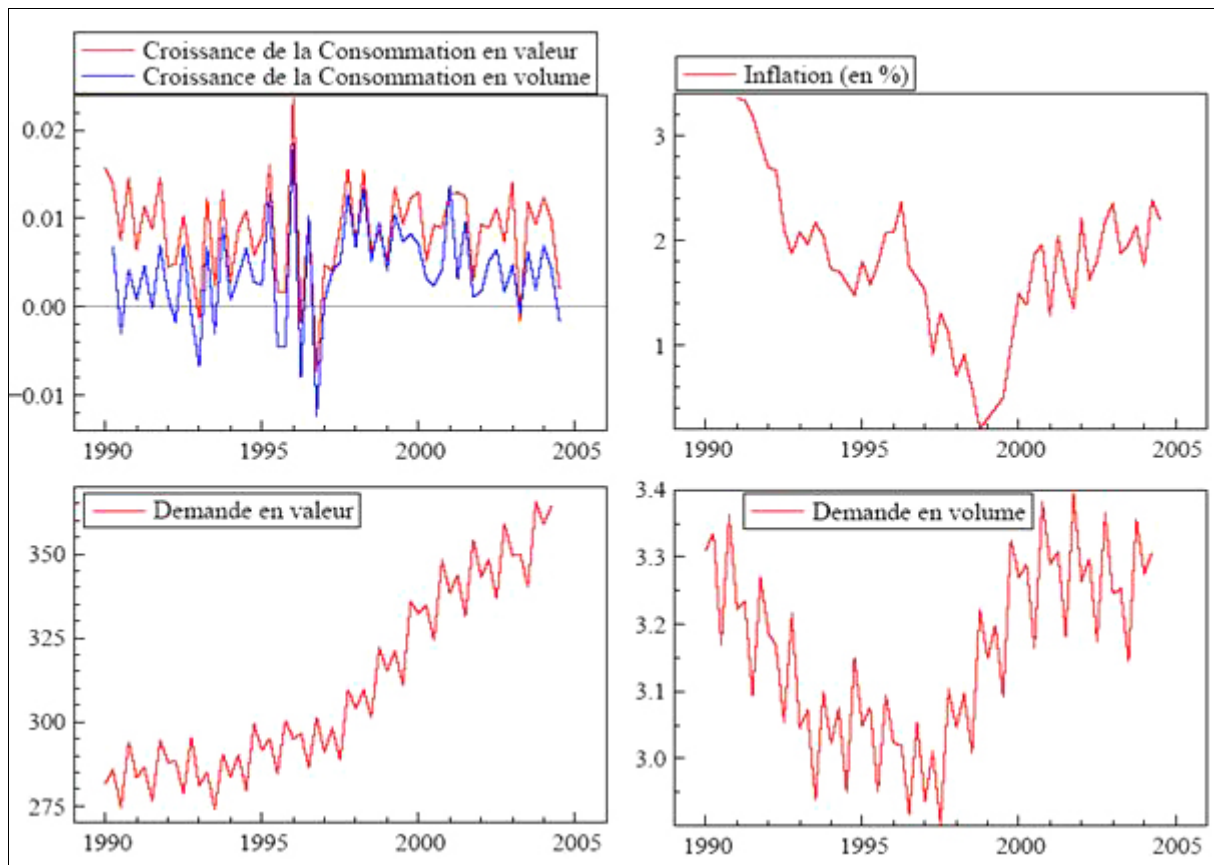


Figure: 66.63 - Quelques séries temporelles

La dimension temporelle est ici importante car il s'agit de l'analyse d'une chronique historique: des variations d'une même variable au cours du temps, afin de pouvoir en comprendre la dynamique. On représente en général les séries temporelles sur des graphiques de valeurs (ordonnées) en fonction du temps (abscisses). Une telle observation constitue un outil essentiel qui permet au modélisateur ayant un peu d'expérience de tout de suite se rendre compte des propriétés dynamiques principales, afin de savoir quel test statistique pratiquer. La figure précédente montre différentes séries temporelles à titre d'exemples.

Une série temporelle se décompose principalement en trois éléments de base:

1. La "**tendance séculaire à long terme**" ou "**trend**". Elle correspond à un mouvement conjoncturel non saisonnier qui traduit l'évolution à long terme de la variable mesurée.
2. Les "**variations saisonnières**" sont des fluctuations périodiques qui se produisent régulièrement (périodiquement: jours, mois, trimestres ou années).
3. Les "**aléas**" ou "**variations résiduelles/accidentelles**" sont des fluctuations de type aléatoires, en général de faible amplitude.

Définitions:

D1. Les séries qui oscillent autour de leur moyenne sont appelées "**séries stationnaires**" (nous verrons une définition rigoureuse un peu plus loin).

D2. Les séries qui semblent croître ou baisser sur l'ensemble de l'échantillon observé sont appelées des "**séries tendanciennes**" et leurs moyennes ne sont pas constantes.

D3. Les séries qui ne sont ni stationnaires ni tendancielle haussière ou baissière à long terme sont appelées "**séries non-stationnaires**" ou "**séries aléatoires**" ou encore "**séries stochastiques**". Tout processus dont l'évolution temporelle peut être analysée en termes de probabilité est donc dit "**processus stochastique**".

D4. Les séries qui présentent une périodicité (fluctuation) régulière sont appelées "**séries saisonnières**". Régulièrement, les séries stationnaires sont des séries tendancielles avec une composante périodique.

D5. Les séries qui semblent saisonnières ou semblent avoir une composante saisonnière mais dont la périodicité est variable, sont appelées "**séries cycliques**".

Soit y_t une valeur de la série, t_t sa composante tendancielle, s_t sa composante saisonnière et ε_t sa partie aléatoire, nous définissons:

- Un modèle additif par une expression du type:

$$y_t = t_t + s_t + \varepsilon_t \quad (66.694)$$

- Un modèle multiplicatif simple:

$$y_t = t_t (1 + s_t) + \varepsilon_t \quad (66.695)$$

- Un modèle multiplicatif complet:

$$y_t = t_t (1 + s_t)(1 + \varepsilon_t) \quad (66.696)$$

- Un modèle mixte:

$$y_t = t_t (1 + s_{1,t}) + s_{2,t} + \varepsilon_t \quad (66.697)$$

Les caractéristiques de ces graphiques sont toutes modélisables et analysables dans le cadre de l'analyse des séries temporelles. Il existe pour cela des outils plus ou moins complexes dont certains ne peuvent être mis en doute et dont d'autres sont des modèles heuristiques qu'il faut savoir manipuler et utiliser avec précaution.

Dans le texte qui va suivre, nous allons nous intéresser qu'aux modèles élémentaires accessibles sans une artillerie mathématique lourde.

L'humilité dans le métier de prévisionniste est cependant de mise pour accepter l'erreur de manière récurrente, tout en cherchant à améliorer la qualité de ses prévisions (ventes, volatilité des marchés, corrélation d'actifs, etc.) et l'identification des composantes non aléatoires des phénomènes. Il convient de se rappeler des 4 critères suivants:

1. Les prévisions sont très généralement inexactes. Il faut donc un écosystème robuste pour réagir rapidement à l'imprévu (savoir que l'on ne peut pas prévoir, ne signifie pas que l'on ne peut pas tirer profit de l'imprévisibilité).

2. Une bonne prévision est plus qu'une valeur numérique. Puisque les prévisions sont généralement inexactes, une bonne prévision doit également inclure une mesure de l'erreur anticipée pour la prévision

3. Les prévisions agrégées sont plus précises. De manière générale, nous remarquons que l'erreur faite pour la prévision de ventes d'une ligne de produits est généralement moindre que l'erreur faite dans la prévision de ventes d'un seul élément.

4. Les prévisions à long terme sont moins précises. Il n'y a pas grand-chose à en dire car c'est intuitif...

5. Les prévisions peuvent être sensibles aux conditions initiales (pensez aux modèles chaotiques de la météorologie ou de la dynamique des populations...)

Remarque: Les exemples de cas où de grands dirigeants de multinationales ont, faute de connaissances statistiques ou mal conseillés par des gestionnaires non statisticiens/non économistes, fait des prévisions de bénéfices ou ventes sur plus de 10 à 20 ans (ce qui est une aberration!!) dans les médias sont légion! Évidemment les chiffres donnés et la stratégie de développement associée ont conduit certaines de ces multinationales à la quasi-catastrophe.

Signalons que les techniques prévisionnelles quantitatives utilisent également et principalement les outils suivants (il est bien évidemment possible de les combiner) pour lesquels nous avons déjà donné des exemples dans les chapitres respectifs:

- Marche aléatoire simple (cf. chapitre d'Économie)
- Régression linéaire simple (cf. chapitre de Méthodes Numériques)
- Régression linéaire multiple (cf. chapitre de Méthodes Numériques)
- Régression logistique (cf. chapitre de Méthodes Numériques)
- Interpolation polynomiale (cf. chapitre de Méthodes Numériques)
- Linéarisation logarithmique et exponentielle (cf. chapitre de Méthodes Numériques)
- Modélisation par mouvement brownien (cf. chapitre d'Économie)
- Moyenne Mobile (cf. chapitre de Statistiques)
- Analyse de corrélations (cf. chapitre de Statistiques)
- Cartes de contrôle (cf. chapitre de Génie Industriel)
- Réseaux de neurones (cf. chapitre de Méthodes Numériques)

et ils peuvent être regroupés en deux familles principales:

1. Les modèles causes-effets basés sur plusieurs variables explicatives

2. Les modèles en série chronologique dont la seule variable explicative est le temps

et c'est toujours délicat pour un prévisionniste de choisir laquelle est la mieux adaptée à son cadre de travail et au niveau de compréhension de ses supérieurs. La partie intéressante du métier est qu'une fois le ou les modèles choisis (encore faut-il choisir parmi la centaine de modèles empiriques publiés...), nous pouvons les confronter aux données statistiques passées de l'entreprise et voir si les prévisions sont en accord avec ce qui c'était réellement passé.

Remarque: Il est préférable d'utiliser des modèles basés sur les statistiques que des modèles déterministes qui sont de piètre qualité (comme la moyenne mobile simple ou double ainsi que les lissages exponentiels et régressions linéaires simples ou multiples).

8.1. TYPES D'ERREURS

Avant de construire des modèles de prévision il faut pouvoir estimer la qualité de celle-ci (à des fins de comparaison aussi entre modèles) et pour cela il est d'usage d'utiliser plusieurs indicateurs empiriques qui sont:

- "L'erreur prévisionnelle":

$$e_t = Y_t - \hat{Y}_t \quad (66.698)$$

qui est donc la différence entre le réel et la modélisation.

- "L'écart-moyen" ou "biais de prévision" ("Mean Error" en anglais):

$$ME = \frac{1}{n} \sum_{t=1}^n e_t \quad (66.699)$$

qui va seulement nous dire s'il y a une erreur systématique positive ou négative entre la prévision et le réel (car si la valeur de ME est non nulle, le modèle est biaisé). Mais comme cette valeur est souvent très petite (les valeurs négatives équilibrant les positives dans certains modèles) nous lui préférons la mesure d'erreur suivante:

- "L'écart moyen absolu" ("Mean Absolute Deviation" ou "Mean Absolute Error" ou encore "Mean Forecast Error" en anglais):

$$MAD = MAE = MFE = \frac{1}{n} \sum_{t=1}^n |e_t| \quad (66.700)$$

qui est la moyenne des erreurs absolues faites par le modèle de prévision sur toute la durée de celle-ci, sans égard au fait que l'erreur soit une surestimation ou une sous-estimation.

- Le "carré moyen des résidus" ("Mean Square Deviation" ou "Mean Square Errors" en anglais):

$$MSD = MSE = \frac{1}{n} \sum_{t=1}^n e_t^2 \quad (66.701)$$

qui permet de pénaliser les grandes erreurs plus que les petites puisque chaque erreur est multipliée par elle-même, donnant ainsi un poids plus grand aux grandes erreurs qu'aux petites erreurs.

Le problème avec les indicateurs précédents est que l'amplitude de l'erreur dépend des amplitudes des valeurs analysées. Pour éviter cet inconvénient, il est plus intéressant de travailler avec un pourcentage (valeur relative). Ainsi, nous définissons "l'erreur relative" indiquée en pourcents ("Percentage Error" en anglais) par:

$$PE_t = \frac{Y_t - \hat{Y}_t}{Y_t} = \frac{e_t}{Y_t} \quad (66.702)$$

qui est donc la différence en pourcents entre le réel et la modélisation par rapport au réel. Il s'ensuit alors "l'erreur relative moyenne" ou "biais relatif de prévision" ("Mean Percentage Error" en anglais) toujours indiquée en pourcents:

$$MPE = \frac{1}{n} \sum_{t=1}^n PE_t \quad (66.703)$$

Pour les mêmes raisons que l'indicateur *ME* nous lui préférons la mesure d'erreur suivante.

- "L'écart relatif moyen absolu" ("Mean Absolute Percentage Error" en anglais) toujours indiqué en pourcents:

$$MAPE = \frac{1}{n} \sum_{t=1}^n |PE_t| \quad (66.704)$$

Le seul problème avec les indicateurs relatifs c'est lorsque la série temporelle analysée contient des zéros...

Enfin, pour clore cette partie concernant les indicateurs, signalons que dans le domaine du prévisionnel il faut aussi vérifier sur le long terme si un modèle théorique est viable ou non pour le remplacer par un autre. Les techniques de base pour cela consistent à faire usage des cartes de contrôles autocorrélées (cf. chapitre de Génie Industriel) où les mesures seront simplement les écarts (sans en prendre la valeur absolue!) entre les valeurs projetées et ce qui a réellement eu lieu.

8.2. DÉCOMPOSITION

Certains modèles utilisant les séries chronologiques sont des outils de prévision adéquats pour autant que la variation ait montré un patron stable dans le temps et que les conditions dans lesquelles le patron est apparu s'appliquent toujours.

Parfois, un patron n'est pas apparent lors de l'analyse des données brute; celles-ci peuvent toutefois être décomposées de manière à révéler des patrons qui facilitent la projection des données dans le futur (nous verrons la décomposition par transformée de Fourier plus loin). Voici les quatre composantes généralement reconnues (et déjà mentionnées plus haut) pour les séries chronologiques hors analyse spectrale: tendance (linéaire, exponentielle, sigmoïde, etc.), saisonnalité, cyclicité, aléatoire.

Voyons cela avec un exemple pratique utilisant (pour faire simple) la version française de Microsoft Excel 14.0.6123. Créons une composante aléatoire basée sur la valeur absolue d'une loi normale de moyenne 10 et d'écart-type 20 et ce sur une période de 32 mois avec la formule suivante dans la colonne B (de B1 à B32):

$$=ABS(LOI.NORMALE.INVERSE.N(ALEA.ENTRE.BORNES(0.1;10000)/10000;10;20))$$

où la colonne A contiendra la numérotation des périodes allant de 1 à 32 (de A1 à A32).

Ce qui donnera une simulation d'un graphique aléatoire dont nous allons fixer les valeurs. Par exemple:

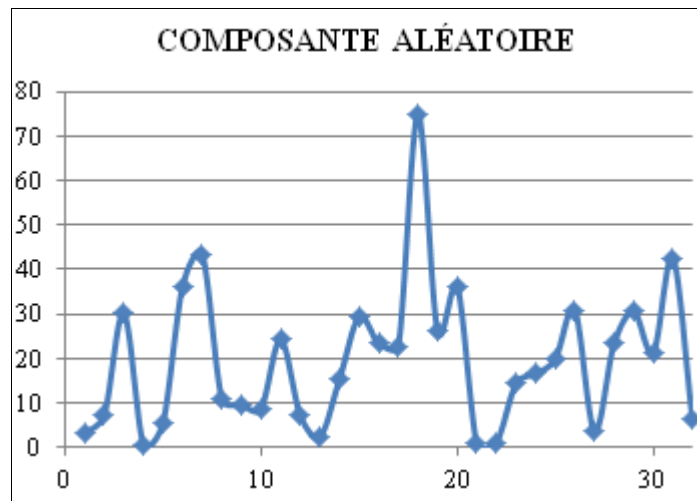


Figure: 66.64 - Tracé d'un exemple de la fonction aléatoire avec Microsoft Excel 14.0.6123

Ensuite, ajoutons une composante de tendance (drift) linéaire de pente 4 et d'ordonnée à l'origine de 20 à cette composante aléatoire de manière à avoir cette fois dans la colonne B la formule suivante:

$$=ABS(LOI.NORMALE.INVERSE.N(ALEA.ENTRE.BORNES(0.1;10000)/10000;10;20))+(4*A1+20)$$

La somme donnant graphiquement:

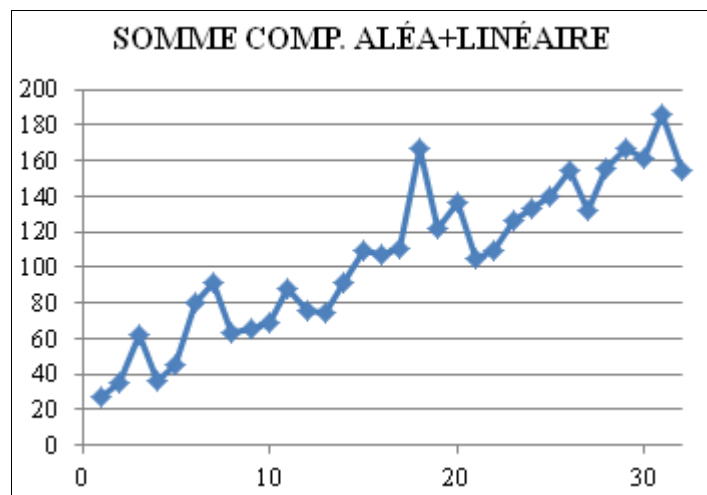


Figure: 66.65 - Tracé de la somme de la fonction aléatoire et la tendance avec Microsoft Excel 14.0.6123

Ajoutons maintenant une composante saisonnière de type sinusoïdale avec une amplitude de 50 et une période de 12 mois et sans déphasage:

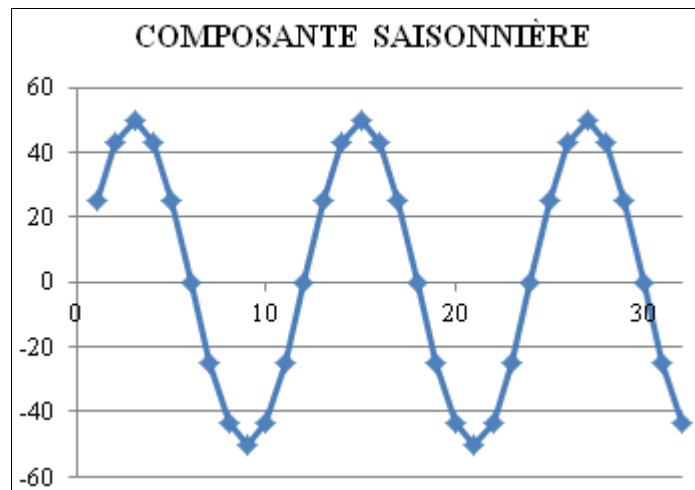


Figure: 66.66 - Tracé de la partie saisonnière seule avec Microsoft Excel 14.0.6123

La somme des trois composantes donnera dans la colonne B:

$$=ABS(LOI.NORMALE.INVERSE.N(ALEA.ENTRE.BORNES(0.1;10000)/10000;10;20))+(4*A1+20)+50*SIN(A1*2*PI()*1/12)$$

et graphiquement cela donne:

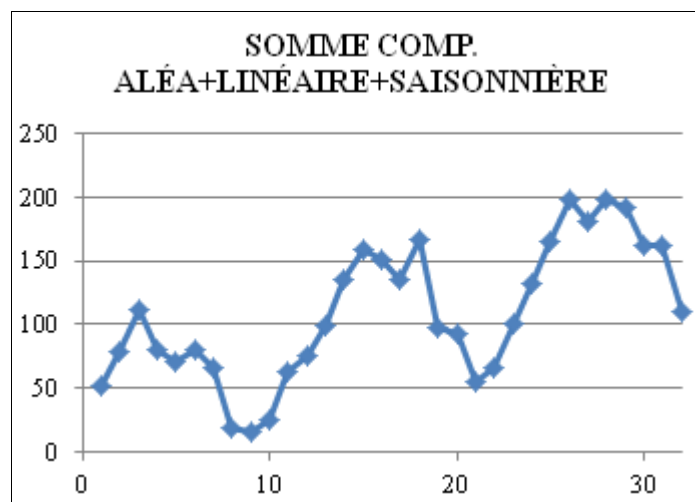


Figure: 66.67 - Tracé de la somme de la fonction aléatoire, la tendance et la saisonnalité avec Microsoft Excel 14.0.6123

et enfin ajoutons une composante cyclique qui sera encore une fois pour l'exemple une fonction sinusoïdale d'amplitude 20 et de période 64:

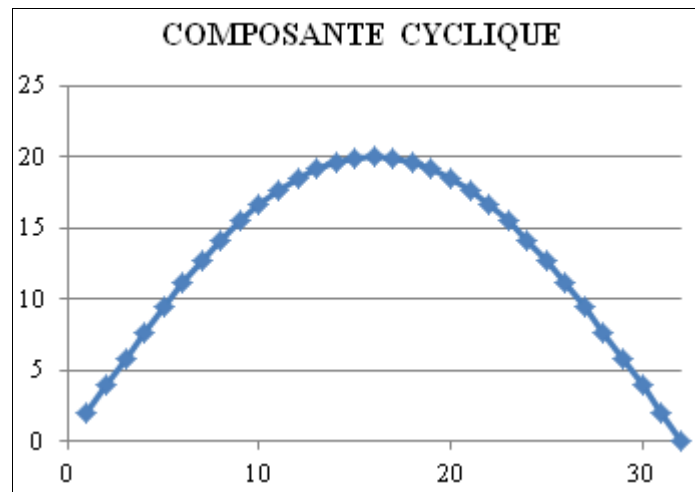


Figure: 66.68 - Tracé de la partie cyclique seule avec Microsoft Excel 14.0.6123

La somme des quatre composantes donnera dans la colonne B:

$$=ABS(LOI.NORMALE.INVERSE.N(ALEA.ENTRE.BORNES(0.1;10000)/10000;10;20))+(4*A1+20)+50*SIN(A1*2*PI()*1/12)+20*SIN(A1*2*PI()*1/64)$$

et graphiquement cela donne:

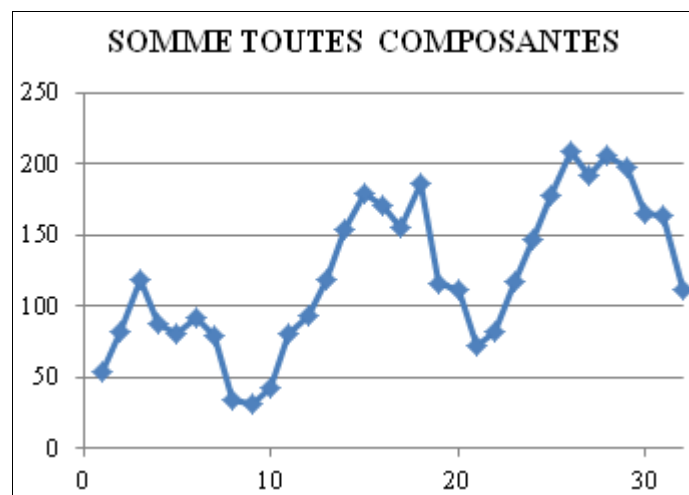


Figure: 66.69 - Somme des toutes les composantes avec Microsoft Excel 14.0.6123

Maintenant, comme nous l'avons introduit en détail dans le chapitre de Suites Et Séries, faisons une transformée de Fourier rapide (analyse spectrale) - abrégée TFR - avec l'Utilitaire d'Analyse de données intégré dans Microsoft Excel 14.0.6123. Cela nous donne alors:

	A	B	C	D
1	Mois	Somme Aléa+Lin+Sais+Cyc	Cn	Cn
2	1	54.15	3905.71268156154	3905.7127
3	2	82.43	-118.965404756721+715.838479602683i	725.6566
4	3	117.98	205.123779910172+553.607514321972i	590.3872
5	4	87.64	-517.526534512484-322.817654253957i	609.95488
6	5	79.94	-57.2052383800352-12.8708650198023i	58.6353
7	6	91.36	-99.9428098103601+120.811932420966i	156.79314
8	7	79.04	-72.0218228963377-50.0407630755668i	87.699606
9	8	33.69	-32.1476861984657-49.7390806032813i	59.223727
10	9	31.04	-157.914952894831-52.5450316445111i	166.4275
11	10	42.07	-78.2534115089101+84.5550741365412i	115.20919
12	11	80.91	-80.1174542237266+15.8369209517564i	81.667708
13	12	93.67	-80.5978065119414+74.2438612297722i	109.58174
14	13	118.54	-42.7997832944269-77.6739639261847i	88.685208
15	14	154.22	48.2069000063986+14.6517897106656i	50.384324
16	15	179.25	-48.2578811831555-34.2025740138052i	59.149296
17	16	170.61	67.2921055723724+62.0011087526294i	91.500628
18	17	155.64	-42.6708876842777	42.670888
19	18	186.39	67.2921055723718-62.0011087526298i	91.500628
20	19	116.27	-48.2578811831558+34.2025740138054i	59.149296
21	20	111.41	48.206900006399-14.6517897106662i	50.384324
22	21	72.55	-42.7997832944267+77.6739639261846i	88.685208
23	22	82.45	-80.5978065119415-74.2438612297724i	109.58174
24	23	116.80	-80.1174542237267-15.8369209517561i	81.667708
25	24	146.69	-78.2534115089099-84.555074136541i	115.20919
26	25	177.57	-157.914952894831+52.5450316445111i	166.4275
27	26	209.24	-32.1476861984655+49.7390806032808i	59.223727
28	27	190.99	-72.0218228963379+50.040763075567i	87.699606
29	28	206.25	-99.9428098103598-120.811932420965i	156.79314
30	29	197.37	-57.2052383800352+12.8708650198023i	58.6353
31	30	165.20	-517.526534512483+322.817654253958i	609.95488
32	31	163.48	205.12377991017-553.607514321973i	590.3872
33	32	110.87	-118.965404756723-715.838479602683i	725.6566

Figure: 66.70 - Tableau de la TFR avec l'Utilitaire d'analyse de Microsoft Excel 14.0.6123

Mais à la différence du chapitre de Suites Et Séries, nous allons cette fois-ci tracer le spectre de fréquence. Pour cela, nous créons d'abord une colonne avec une numérotation allant de 0 à 31 et juste à côté nous mettons la formule visible dans la capture ci-dessous

E	F
<i>n</i>	<i>Fréquences</i>
0	0
1	=E3*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
2	=E4*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
3	=E5*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
4	=E6*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
5	=E7*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
6	=E8*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
7	=E9*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
8	=E10*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
9	=E11*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
10	=E12*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
11	=E13*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
12	=E14*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
13	=E15*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
14	=E16*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
15	=E17*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
16	=E18*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
17	=E19*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
18	=E20*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
19	=E21*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
20	=E22*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
21	=E23*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
22	=E24*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
23	=E25*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
24	=E26*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
25	=E27*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
26	=E28*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
27	=E29*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
28	=E30*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
29	=E31*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
30	=E32*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32
31	=E33*(NB(\$B\$2:\$B\$33)/(\$A\$33-\$A\$2))/32

Figure: 66.71 - Formules pour obtenir les fréquences dans Microsoft Excel 14.0.6123

ce qui donne:

D	E	F	G
Cn	n	Fréquences	Périodes
3905.7127	0	-	
725.6566	1	0.03	31.00
590.3872	2	0.06	15.50
609.95488	3	0.10	10.33
58.6353	4	0.13	7.75
156.79314	5	0.16	6.20
87.699606	6	0.19	5.17
59.223727	7	0.23	4.43
166.4275	8	0.26	3.88
115.20919	9	0.29	3.44
81.667708	10	0.32	3.10
109.58174	11	0.35	2.82
88.685208	12	0.39	2.58
50.384324	13	0.42	2.38
59.149296	14	0.45	2.21
91.500628	15	0.48	2.07
42.670888	16	0.52	1.94
91.500628	17	0.55	1.82
59.149296	18	0.58	1.72
50.384324	19	0.61	1.63
88.685208	20	0.65	1.55
109.58174	21	0.68	1.48
81.667708	22	0.71	1.41
115.20919	23	0.74	1.35
166.4275	24	0.77	1.29
59.223727	25	0.81	1.24
87.699606	26	0.84	1.19
156.79314	27	0.87	1.15
58.6353	28	0.90	1.11
609.95488	29	0.94	1.07
590.3872	30	0.97	1.03
725.6566	31	1.00	1.00

Figure: 66.72 - Valeurs des fréquences de la TFR

Soit graphiquement pour les 16 premiers points (les autres étant identiques par symétrie):

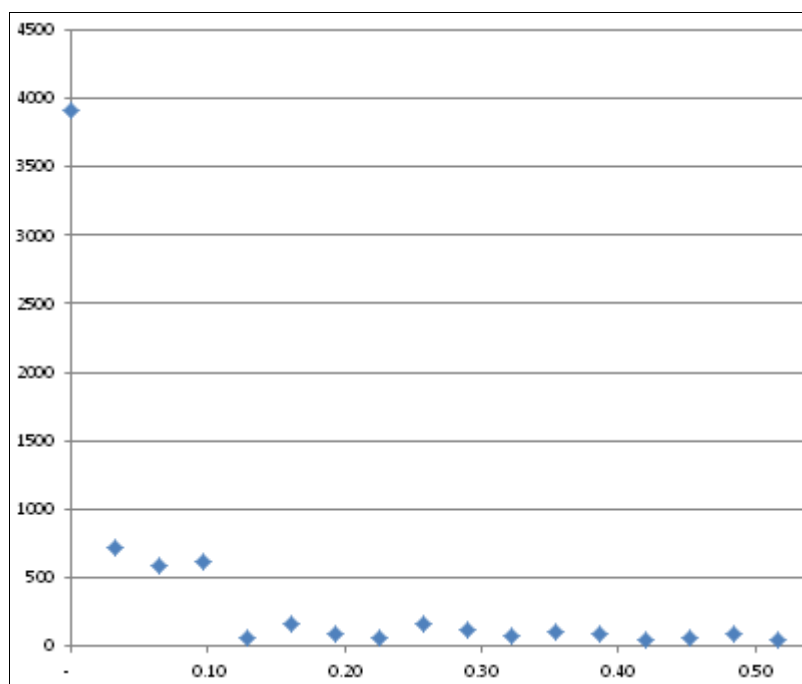


Figure: 66.73 - Tracé des pics de fréquence de la TFR avec Microsoft Excel 14.0.6123

Nous voyons donc que la transformée met en évidence une composante constante (de fréquence nulle) qui est très importante. Nous pouvons abusivement l'assimiler à la composante linéaire.

Nous voyons ensuite que trois points sortent du lot avec les périodes respectives (inverse de la fréquence): 31 mois, 15.50, 10.33. Mais nous n'allons considérer uniquement que les deux premières valeurs pour les associer respectivement au cycle (ayant la plus grande période par définition) et à la saisonnalité.

Ainsi, la transformée de Fourier nous amène à considérer une tendance constante, un phénomène cyclique ayant une période 31 mois (alors que nous avons simulé un cycle ayant une période de 64 mois...) et un phénomène saisonnier ayant une période 15 mois (alors que nous en avons simulé un avec une période de 12 mois).

Nous voyons donc bien que ce type d'analyse a ses limites mais qu'il peut être utile en tant qu'outil d'aide à la décision.

8.3. MODÈLES PRÉVISIONNELS DÉTERMINISTES

L'analyse de séries temporelles comporte deux grandes approches:

1. L'approche déterministe qui ne fait à la base aucunement appel aux outils statistiques
2. L'approche statistique qui permet d'inférer sur les prévisions

Il va de soi que la première méthode est plus simple que la deuxième mais que la deuxième inclut la première et est scientifiquement parlant plus acceptable puisqu'il va être possible de définir des intervalles de confiance.

Ces concepts généraux étant présentés nous allons passer maintenant à la présentation de quelques modèles mathématiques de régression des séries temporelles. Il faut savoir à ce sujet que:

1. Il en existe (à ma connaissance) une soixantaine
2. Certains sont déterministes, d'autres stochastiques (ces derniers étant les meilleurs)
3. Ils sont (à ma connaissance) tous empiriques
4. Un logiciel comme Microsoft Excel suffit pour modéliser la majorité d'entre eux

Voici la liste des noms des techniques les plus connues et utilisées par certaines multinationales dans la pratique: Moyenne Mobile, Single Exponential Smoothing (SES), Régression Logistique, Méthode Linéaire de Holt's, Holt's-Winters, Méthode de Brown, Stochastique Linéarisée, ARIMA, etc.

Nous allons voir ci-dessous les démonstrations de celles qui nécessitent une démarche mathématique. Les autres n'étant que des boîtes à outils, elles n'ont pas leur place sur un site web tel que Sciences.ch.

Nous allons dans ce qui suit bien évidemment nous concentrer d'abord sur les grands classiques scolaires de la première famille de modèles.

8.3.1. MOYENNE MOBILE SIMPLE (LISSAGE PAR MOYENNE MOBILE)

La technique prévisionnelle de la moyenne mobile simple d'ordre k ou "**Simple Moving Average**" en anglais (MA) est certainement la technique empirique la plus simple et aussi la plus mauvaise en termes de pouvoir prédictif. Basée sur la définition de la moyenne mobile (cf. chapitre de Statistiques) elle est définie par:

$$MA(k) = F_{t+1} = \frac{\sum_{i=t-k+1}^t Y_i}{k} \quad (66.705)$$

et constitue donc une prévision déterministe portant uniquement sur une unique période ultérieure.

Voyons un exemple avec la version anglaise de Microsoft Excel 11.8346 comportant une prévision sur l'horizon d'un mois avec des modèles de lissage par moyenne mobile dans les colonnes D et E (comme le domaine de la finance est essentiellement en anglais nous ferons toutes les captures avec la version anglophone du logiciel):

	A	B	C	D	E
1	MOYENNE MOBILE				
2					
3	Mois	Périodes	Y_i	MA (3)	MA (5)
4	Janv	1	200		
5	Févr	2	135		
6	Mars	3	195		
7	Avr	4	197	176.7	
8	Mai	5	310	175.7	
9	Juin	6	175	234.0	207.4
10	Juil	7	155	227.3	202.4
11	Août	8	130	213.3	206.4
12	Sept	9	220	153.3	193.4
13	Oct	10	277	168.3	198
14	Nov	11	235	209.0	191.4
15	Déc	12		244.0	203.4

Figure: 66.74 - Tableau de base pour l'exemple du lissage MA dans Microsoft Excel 11.8346

Avec donc les relations suivantes:

	A	B	C	D	E
1	MOYENN				
2					
3	Mois	ériode	Y_i	MA (3)	MA (5)
4	Janv	1	200		
5	Févr	2	135		
6	Mars	3	195		
7	Avr	4	197	=AVERAGE(C4:C6)	
8	Mai	5	310	=AVERAGE(C5:C7)	
9	Juin	6	175	=AVERAGE(C6:C8)	=AVERAGE(C4:C8)
10	Juil	7	155	=AVERAGE(C7:C9)	=AVERAGE(C5:C9)
11	Août	8	130	=AVERAGE(C8:C10)	=AVERAGE(C6:C10)
12	Sept	9	220	=AVERAGE(C9:C11)	=AVERAGE(C7:C11)
13	Oct	10	277	=AVERAGE(C10:C12)	=AVERAGE(C8:C12)
14	Nov	11	235	=AVERAGE(C11:C13)	=AVERAGE(C9:C13)
15	Déc	12		=AVERAGE(C12:C14)	=AVERAGE(C10:C14)

Figure: 66.75 - Formules pour obtenir les lissages MA(3) et MA(5) avec Microsoft Excel 11.8346

et le graphique correspondant:

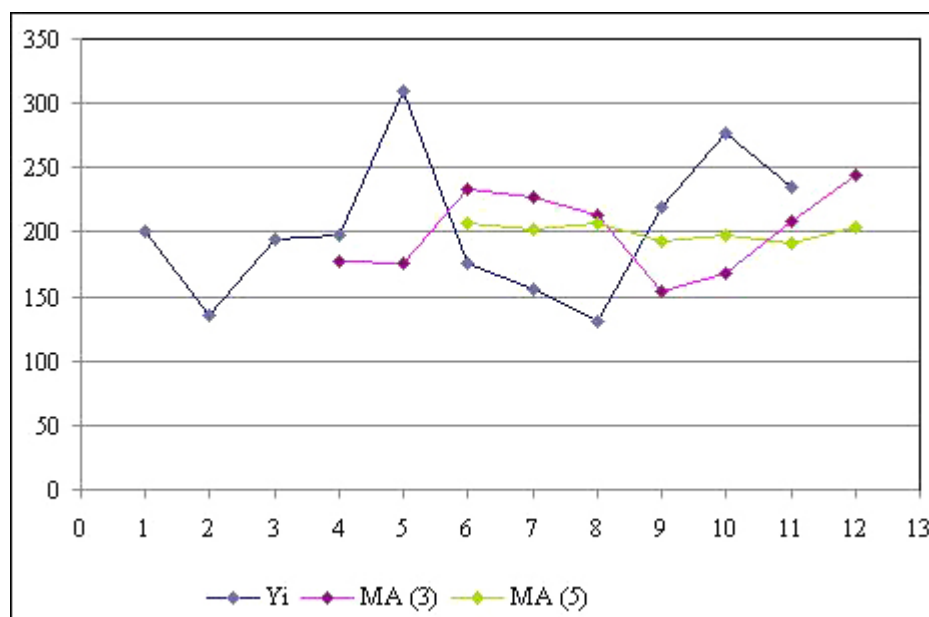


Figure: 66.76 - Tracé des observations et des deux lissages avec Microsoft Excel 11.8346

Nous n'avons pas utilisé l'outil d'Utilitaire d'analyse de Microsoft Excel 11.8346 (ce dernier intégrant un outil faisant automatiquement les calculs de la moyenne mobile) uniquement pour bien montrer au lecteur l'application explicite des relations mathématiques données plus haut.

Le choix du nombre de périodes k à utiliser dans le calcul de la moyenne mobile simple dépend beaucoup des variations attendues dans les données. Ceci peut être illustré par deux caractéristiques de la prévision:

- **Stabilité:** En faisant la moyenne de plusieurs périodes, on atténue les fluctuations aléatoires afin que la prévision soit plus stable. La stabilité est la propriété de la prévision à ne pas fluctuer de manière désordonnée. Gagner en stabilité est un avantage s'il y a beaucoup de fluctuations aléatoires dans les données. Une moyenne mobile gagnera en stabilité si un plus grand nombre de périodes est utilisé dans le calcul de la moyenne. Un gain en stabilité est désirable seulement jusqu'au point où les fluctuations aléatoires sont atténuées. Si le nombre de périodes utilisées dans le calcul de la moyenne est trop grand, la moyenne sera tellement stable qu'elle ne répondra que trop lentement aux changements non aléatoires de la demande.

- **Réactivité:** La réactivité est la propriété qu'a la prévision à s'ajuster rapidement à un changement dans le niveau moyen réel de la demande. L'utilisation d'une prévision réactive est appropriée dans le cas où les fluctuations aléatoires sont faibles. Plus le nombre de périodes utilisées dans le calcul de la moyenne mobile est petit, plus le modèle prévisionnel sera réactif.

Remarque: Signalons que certaines techniques de prévision font des moyennes mobiles de moyennes mobiles... Dans la littérature spécialisée, cela s'appelle la "**moyenne mobile double**"... Il y a encore des modèles qui appliquent des poids empiriques aux périodes passées, cela s'appelle la "**moyenne mobile pondérée**" et c'est à mon avis hors d'intérêt.

8.3.2. MODÈLE LINÉAIRE AVEC COEFFICIENTS SAISONNIERS

La technique prévisionnelle linéaire avec coefficients saisonniers est simple à mettre en place et à comprendre. Elle associe une simple régression linéaire (cf. chapitre de Méthodes Numériques) et des coefficients saisonniers calculés sur la base des valeurs passées par rapport au modèle linéaire. Cette technique permet de dégager des tendances en lissant des séries chronologiques et en faisant apparaître des variations importantes dans le temps. Les moyennes mobiles sont des moyennes généralement calculées pour des données journalières ou hebdomadaires et qui sont ensuite décalées période par période.

Voyons donc un exemple toujours avec Microsoft Excel 11.8346:

	A	B	C
1	N°	Mois	Ventes
2	1	Janvier 1987	890.00
3	2	Février 1987	1'115.00
4	3	Mars 1987	1'280.00
5	4	Avril 1987	1'328.00
6	5	Mai 1987	1'253.00
7	6	Juin 1987	1'124.00
8	7	Juillet 1987	1'192.00
9	8	Août 1987	876.00
10	9	Septembre 1987	1'401.00
11	10	Octobre 1987	1'501.00
12	11	Novembre 1987	1'065.00
13	12	Décembre 1987	976.00
14	13	Janvier 1988	895.00
15	14	Février 1988	1'179.00
16	15	Mars 1988	1'315.00
17	16	Avril 1988	1'452.00
18	17	Mai 1988	1'325.00
19	18	Juin 1988	1'202.00
20	19	Juillet 1988	1'175.00
21	20	Août 1988	933.00
22	21	Septembre 1988	1'412.00
23	22	Octobre 1988	1'528.00
24	23	Novembre 1988	1'171.00
25	24	Décembre 1988	1'012.00

Figure: 66.77 - Valeurs de départ dans Microsoft Excel 11.8346 pour l'exemple

Ce qui ressemble à:

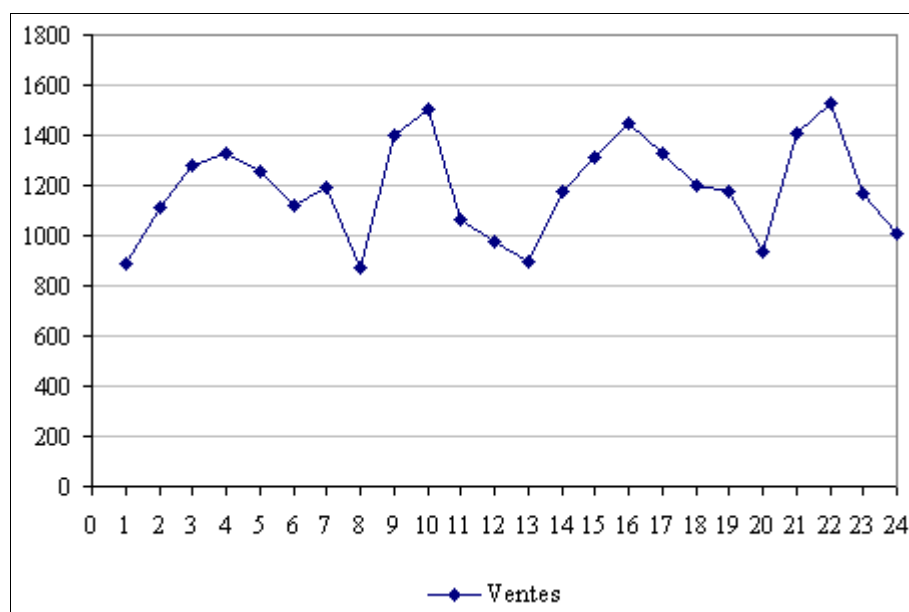


Figure: 66.78 - Graphique des ventes correspondant dans Microsoft Excel 11.8346

La première chose à faire est de déterminer l'équation de la droite de régression (ce qui permet donc in extenso de faire aussi de l'inférence!). Ce qui est simple avec la méthode des moindres carrés démontrée dans le chapitre de Méthodes Numériques et extrêmement simple à obtenir avec les outils automatiques intégrés à Microsoft Excel 11.8346:

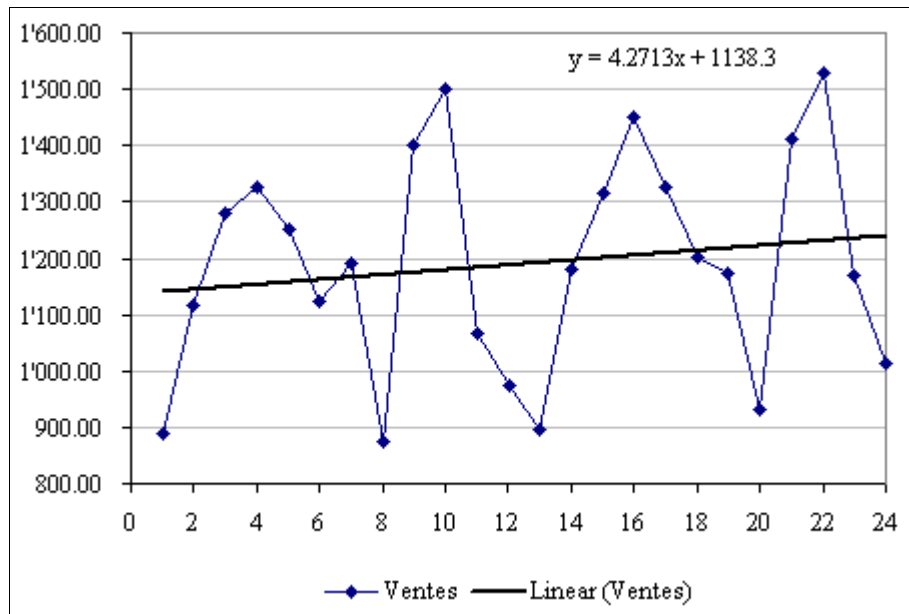


Figure: 66.79 - Régression linéaire des ventes avec équations dans Microsoft Excel 11.8346

Ce qui nous permet d'avoir deux nouvelles colonnes dans le tableau Microsoft Excel 11.8346. Une avec les valeurs du modèle linéaire et une autre avec le rapport entre les valeurs réelles et le modèle linéaire (ce qui correspond donc au coefficient saisonnier):

	A	B	C	D	E
1	N°	Mois	Ventes	Modèle Linéaire	Coeff.
2	1	Janvier 1987	890.00	1'142.57	0.78
3	2	Février 1987	1'115.00	1'146.84	0.97
4	3	Mars 1987	1'280.00	1'151.11	1.11
5	4	Avril 1987	1'328.00	1'155.39	1.15
6	5	Mai 1987	1'253.00	1'159.66	1.08
7	6	Juin 1987	1'124.00	1'163.93	0.97
8	7	Juillet 1987	1'192.00	1'168.20	1.02
9	8	Août 1987	876.00	1'172.47	0.75
10	9	Septembre 1987	1'401.00	1'176.74	1.19
11	10	Octobre 1987	1'501.00	1'181.01	1.27
12	11	Novembre 1987	1'065.00	1'185.28	0.90
13	12	Décembre 1987	976.00	1'189.56	0.82
14	13	Janvier 1988	895.00	1'193.83	0.75
15	14	Février 1988	1'179.00	1'198.10	0.98
16	15	Mars 1988	1'315.00	1'202.37	1.09
17	16	Avril 1988	1'452.00	1'206.64	1.20
18	17	Mai 1988	1'325.00	1'210.91	1.09
19	18	Juin 1988	1'202.00	1'215.18	0.99
20	19	Juillet 1988	1'175.00	1'219.45	0.96
21	20	Août 1988	933.00	1'223.73	0.76
22	21	Septembre 1988	1'412.00	1'228.00	1.15
23	22	Octobre 1988	1'528.00	1'232.27	1.24
24	23	Novembre 1988	1'171.00	1'236.54	0.95
25	24	Décembre 1988	1'012.00	1'240.81	0.82

Figure: 66.80 - Report du modèle théorique avec rapport réel/modèle dans Microsoft Excel 11.8346

Nous faisons ensuite une prévision sur la base de la régression linéaire multipliée par la moyenne des coefficients saisonniers de chaque mois. Ce que nous noterons dans le cas présent:

$$\boxed{F_t = (a \cdot t_i + b) \cdot \bar{c}_{i \bmod 12}} \quad (66.706)$$

Il s'agit donc d'une méthode déterministe permettant de faire de la prévision sur autant de périodes ultérieures que désirées avec la possibilité de faire de l'inférence statistique si besoin est (c'est le seul modèle déterministe sur lequel l'inférence est directement possible).

8.3.3. LISSAGE EXPONENTIEL SIMPLE

La prévision F_t fournie par la méthode de lissage exponentiel (EWMA pour "Exponential Weighted Moving Average" ou SES pour "Simple Exponential Smoothing" en anglais), appelée aussi "**moyenne mobile exponentielle pondérée**", avec la "**constante de lissage**" $\beta \in]0, 1[$ est donnée comme nous allons le démontrer plus loin par:

$$F_{t+1} = (1 - \beta) \sum_{i=0}^{t-1} \beta^i Y_{t-i} \quad (66.707)$$

Cette définition repose sur une idée simple (mais néanmoins totalement empirique): nous supposons que les observations influencent d'autant moins la prévision qu'elles sont éloignées de la date t pour laquelle nous faisons la prévision et que cette influence décroît exponentiellement. Nous voyons alors assez facilement dans un premier temps que:

- Plus la constante de lissage est proche de 1, plus l'influence des observations passées remonte loin dans le temps et plus la prévision est rigide, c'est-à-dire peu sensible aux fluctuations conjoncturelles.
- Au contraire, plus la constante de lissage est proche de zéro, plus la prévision est influencée par les observations récentes (prévision souple).

Démontrons dans un tout premier temps que la définition précédente s'écrit sous une autre forme moins technique qui est la plus commune dans les livres de statistiques:

$$F_{t+1} = (1 - \beta) \sum_{i=0}^{t-1} \beta^i Y_{t-i} = (1 - \beta) \sum_{i=1}^{t-1} \beta^i Y_{t-i} + (1 - \beta) Y_t \quad (66.708)$$

et nous avons aussi in extenso:

$$F_t = (1 - \beta) \sum_{i=0}^{t-2} \beta^i Y_{t-1-i} \quad (66.709)$$

Dès lors si nous réindexons la somme:

$$\begin{aligned} F_{t+1} &= (1 - \beta) \sum_{i=1}^{t-1} \beta^i Y_{t-i} + (1 - \beta) Y_t = (1 - \beta) \sum_{i=0}^{t-1-1} \beta^{i+1} Y_{t-1-i} + (1 - \beta) Y_t \\ &= (1 - \beta) \sum_{i=0}^{t-2} \beta^{i+1} Y_{t-1-i} + (1 - \beta) Y_t = \beta (1 - \beta) \sum_{i=0}^{t-2} \beta^i Y_{t-1-i} + (1 - \beta) Y_t \\ &= \beta F_t + (1 - \beta) Y_t \end{aligned} \quad (66.710)$$

La prévision F_{t+1} apparaît donc comme une moyenne pondérée entre la prévision F_t faite à la date t et la dernière observation Y_t , le poids donné à cette dernière observation étant d'autant plus fort que la constante de lissage est faible.

En pratique, le choix de la constante de lissage est souvent fondé sur des critères subjectifs de rigidité ou de souplesse de prévision. Il existe cependant une méthode plus objective qui consiste à déterminer des outils de recherche opérationnelle (cf. chapitre de Méthodes Numériques), la constante qui minimise la somme des carrés des erreurs de prévisions.

Attention!!! Dans la majorité des ouvrages de statistiques contemporains, la constante de lissage est définie comme étant:

$$\lambda = 1 - \beta \quad (66.711)$$

Il vient alors:

$$F_{t+1} = \lambda \sum_{i=0}^{t-1} (1 - \lambda)^i Y_{t-i} = (1 - \lambda) F_t + \lambda Y_t = F_t + \lambda (Y_t - F_t) \quad (66.712)$$

écriture qui amène à considérer le lissage exponentiel simple comme une bête moyenne mobile pondérée.

L'une et l'autre des écritures sont bien sûr équivalentes mais il faut sur certains logiciels faire attention de bien savoir à laquelle des deux constantes de lissage nous avons affaire... Nous allons continuer personnellement avec cette dernière forme pour la suite de nos développements.

Maintenant, repartons de la relation précédente pour l'écrire sous une autre forme en utilisant la récursivité:

$$\begin{aligned} F_{t+1} &= \lambda Y_t + (1-\lambda)F_t = \lambda Y_t + (1-\lambda)(\lambda Y_{t-1} + (1-\lambda)F_{t-1}) \\ &= \left(\sum_{i=0}^{t-1} \lambda(1-\lambda)^i Y_{t-i} \right) + (1-\lambda)^t F_1 \end{aligned} \quad (66.713)$$

Montrons maintenant que la somme des poids est égale à l'unité afin de constater qu'il s'agit bien d'une pondération. Nous avons donc:

$$\left(\sum_{i=0}^{t-1} \lambda(1-\lambda)^i \right) + (1-\lambda)^t = \lambda \left(\sum_{i=0}^{t-1} (1-\lambda)^i \right) + (1-\lambda)^t = \lambda \left(\sum_{i=0}^{t-1} x^i \right) + (1-\lambda)^t \quad (66.714)$$

Nous avons démontré dans le chapitre de Suites Et Séries que:

$$S_n = u_0 \frac{1-q^{n+1}}{1-q} \quad (66.715)$$

Il vient alors dans notre exemple:

$$S_{t-1} = u_0 \frac{1-q^t}{1-q} = \frac{1-x^t}{1-x} = \frac{1-(1-\lambda)^t}{1-(1-\lambda)} \quad (66.716)$$

Donc:

$$\begin{aligned} \lambda \left(\sum_{i=0}^{t-1} x^i \right) + (1-\lambda)^t &= \lambda \frac{1-(1-\lambda)^t}{1-(1-\lambda)} + (1-\lambda)^t \\ &= \lambda \frac{1-(1-\lambda)^t}{\lambda} + (1-\lambda)^t = 1-(1-\lambda)^t + (1-\lambda)^t = 1 \end{aligned} \quad (66.717)$$

Enfin, parlons de l'origine du nom de cette technique. Nous avons donc:

$$\begin{aligned} F_{t+1} &= (1-\lambda)F_t + \lambda Y_t = (1-\lambda)[(1-\lambda)F_{t-1} + \lambda Y_{t-1}] + \lambda Y_t \\ &= (1-\lambda)^2 F_{t-1} + \lambda(1-\lambda)Y_{t-1} + \lambda Y_t \\ &= (1-\lambda)^2 [(1-\lambda)F_{t-2} + \lambda Y_{t-2}] + \lambda(1-\lambda)Y_{t-1} + \lambda Y_t \\ &= (1-\lambda)^3 F_{t-2} + \lambda(1-\lambda)^2 Y_{t-2} + \lambda(1-\lambda)Y_{t-1} + \lambda Y_t = \dots \end{aligned} \quad (66.718)$$

Nous y retrouvons:

$$\dots, \lambda(1-\lambda)^2, \lambda(1-\lambda), \lambda \quad (66.719)$$

et comme $1-\lambda$ est strictement inférieur à l'unité, les facteurs $(1-\lambda)^n$ décroissent de manière exponentielle (il suffit de représenter graphiquement la fonction):

$$f(n) = (1-\lambda)^n \quad (66.720)$$

pour le constater avec $n \in \mathbb{N}$.

Ceci étant fait, parlons du problème des valeurs initiales. Nous avons donc:

$$F_2 = \lambda Y_1 + (1-\lambda) F_1 = \lambda Y_1 + F_1 - \lambda F_1 \quad (66.721)$$

Il est alors de convention de prendre:

$$F_1 = Y_1 \quad (66.722)$$

Ce qui nous donne:

$$F_2 = F_1 \quad (66.723)$$

Raison pour laquelle un logiciel comme Microsoft Excel 11.8346 nous retournera pour F_1 l'erreur #N/A.

Pour la suite il est nécessaire que nous étudions l'origine de l'expression du lissage exponentiel simple. L'idée à la base des mathématiciens était de réduire l'erreur entre la valeur réelle Y_{t-j} et la prévision considérée comme constante a :

$$E = \sum_{j=0}^{t-1} (Y_{t-j} - a) \quad (66.724)$$

mais le but était aussi d'affaiblir les erreurs plus petites que l'unité et d'amplifier celles qui en étaient plus grandes en élevant les écarts au carré tels que:

$$E = \sum_{j=0}^{t-1} (Y_{t-j} - a)^2 \quad (66.725)$$

et enfin de pondérer celles qui étaient éloignées dans le temps de la prévision par un coefficient tel que nous ayons à minimiser au final:

$$E = \sum_{j=0}^{t-1} \beta^j (Y_{t-j} - a)^2 \quad (66.726)$$

En prenant la dérivée de E par rapport à a , et en l'annulant il vient:

$$\frac{\partial E}{\partial a} = -2 \sum_{j=0}^{t-1} \beta^j (Y_{t-j} - a) = 0 \quad (66.727)$$

Nous obtenons alors pour $\beta \neq 1$:

$$\sum_{j=0}^{t-1} \beta^j Y_{t-j} = a \sum_{j=0}^{t-1} \beta^j = a \frac{1 - \beta^t}{1 - \beta} \quad (66.728)$$

et donc au final:

$$a = \frac{1 - \beta}{1 - \beta^t} \sum_{j=0}^{t-1} \beta^j Y_{t-j} \quad (66.729)$$

et pour t grand il vient:

$$a \cong (1 - \beta) \sum_{j=0}^{t-1} \beta^j Y_{t-j} \quad (66.730)$$

Nous retrouvons donc la relation définie au début de l'étude de ce modèle et qui était:

$$F_{t+1} = (1 - \beta) \sum_{i=0}^{t-1} \beta^i Y_{t-i} = (1 - \beta) \sum_{i=1}^t \beta^{i-1} Y_{t+1-i} = (1 - \beta) Y_t + \beta F_t \quad (66.731)$$

Il est à noter qu'il n'est pas du tout évident que la constante recherchée qui minimise l'erreur puisse être utilisée comme fonction de prévision. Cependant c'est l'usage qu'il en est fait par de nombreux praticiens... Certains spécialistes préfèrent noter ce dernier résultat sous la forme (il ne faut pas chercher à trop comparer cette notation avec celle de la relation précédente):

$$S_t = (1 - \beta) Y_t + \beta S_{t-1} = \lambda Y_t + (1 - \lambda) S_{t-1} \quad (66.732)$$

où S fait référence au "S" qu'il y a dans le mot "eStimé". Il y a même des logiciels de statistiques qui proposent à notre étonnement de faire une prévision à un temps $t+n$ en se basant sur cette méthode... ce qui nous semble un peu limite limite... (affaire à suivre).

Remarque: Le lissage exponentiel simple a été suggéré en premier par Charles C. Holt en 1957, mais la formulation (forme de la relation prévisionnelle) est attribuée à Brown. Il s'agit d'une technique que l'on retrouve implémentée dans la quasi-totalité des logiciels de statistiques.

Exemple:

Nous voulons donc appliquer:

$$F_{t+1} = (1 - \beta) Y_t + \beta F_t = \lambda Y_t + (1 - \lambda) F_t \quad (66.733)$$

Considérons le tableau suivant (la colonne C contient déjà le modèle mathématique dont nous allons détailler les formules juste après):

	A	B	C	E	F
3	Période	Y_t	SES	Écart relatif absolu	Écart absolu
4	1	143	143		
5	2	152	143.00	0.06	9.00
6	3	161	143.00	0.11	18.00
7	4	139	143.00	0.03	4.00
8	5	137	143.00	0.04	6.00
9	6	174	143.00	0.18	31.00
10	7	142	143.00	0.01	1.00
11	8	141	143.00	0.01	2.00
12	9	162	143.00	0.12	19.00
13	10	180	143.00	0.21	37.00
14	11	164	143.00	0.13	21.00
15	12	171	143.00	0.16	28.00
16	13	206	143.00	0.31	63.00
17	14	193	143.00	0.26	50.00
18	15	207	143.00	0.31	64.00
19	16	218	143.00	0.34	75.00
20	17	229	143.00	0.38	86.00
21	18	225	143.00	0.36	82.00
22	19	204	143.00	0.30	61.00
23	20	227	143.00	0.37	84.00
24	21	223	143.00	0.36	80.00
25	22	242	143.00	0.41	99.00
26	23	239	143.00	0.40	96.00
27	24	266	143.00	0.46	123.00
28			143.00		

Figure: 66.81 - Données de départ avec modèle de lissage exponentiel dans Microsoft Excel 14.0.6123

Nous n'allons pas utiliser l'outil d'Utilitaire d'analyse de Microsoft Excel 14.0.6123 (ce dernier intégrant un outil faisant automatiquement les calculs du lissage exponentiel simple) uniquement pour bien montrer au lecteur l'application explicite des relations mathématiques démontrées plus haut. De plus l'Utilitaire d'analyse de Microsoft Excel 14.0.6123 souffre d'un gros défaut: il demande la valeur de la constante de lissage plutôt que d'en calculer une valeur optimale automatiquement.

Il n'y a que trois colonnes qui ont des formules, le reste étant des saisies statiques. Nous avons:

	A	B	C	E	F
3	Période	Y_t	SES	Écart relatif absolu	Écart absolu
4	1	143	=B4		
5	2	152	=I\$3*B4+(1-I\$3)*C4	=ABS(C5-B5)/B5	=ABS(B5-C5)
6	3	161	=I\$3*B5+(1-I\$3)*C5	=ABS(C6-B6)/B6	=ABS(B6-C6)
7	4	139	=I\$3*B6+(1-I\$3)*C6	=ABS(C7-B7)/B7	=ABS(B7-C7)
8	5	137	=I\$3*B7+(1-I\$3)*C7	=ABS(C8-B8)/B8	=ABS(B8-C8)
9	6	174	=I\$3*B8+(1-I\$3)*C8	=ABS(C9-B9)/B9	=ABS(B9-C9)
10	7	142	=I\$3*B9+(1-I\$3)*C9	=ABS(C10-B10)/B10	=ABS(B10-C10)
11	8	141	=I\$3*B10+(1-I\$3)*C10	=ABS(C11-B11)/B11	=ABS(B11-C11)
12	9	162	=I\$3*B11+(1-I\$3)*C11	=ABS(C12-B12)/B12	=ABS(B12-C12)
13	10	180	=I\$3*B12+(1-I\$3)*C12	=ABS(C13-B13)/B13	=ABS(B13-C13)
14	11	164	=I\$3*B13+(1-I\$3)*C13	=ABS(C14-B14)/B14	=ABS(B14-C14)
15	12	171	=I\$3*B14+(1-I\$3)*C14	=ABS(C15-B15)/B15	=ABS(B15-C15)
16	13	206	=I\$3*B15+(1-I\$3)*C15	=ABS(C16-B16)/B16	=ABS(B16-C16)
17	14	193	=I\$3*B16+(1-I\$3)*C16	=ABS(C17-B17)/B17	=ABS(B17-C17)
18	15	207	=I\$3*B17+(1-I\$3)*C17	=ABS(C18-B18)/B18	=ABS(B18-C18)
19	16	218	=I\$3*B18+(1-I\$3)*C18	=ABS(C19-B19)/B19	=ABS(B19-C19)
20	17	229	=I\$3*B19+(1-I\$3)*C19	=ABS(C20-B20)/B20	=ABS(B20-C20)
21	18	225	=I\$3*B20+(1-I\$3)*C20	=ABS(C21-B21)/B21	=ABS(B21-C21)
22	19	204	=I\$3*B21+(1-I\$3)*C21	=ABS(C22-B22)/B22	=ABS(B22-C22)
23	20	227	=I\$3*B22+(1-I\$3)*C22	=ABS(C23-B23)/B23	=ABS(B23-C23)
24	21	223	=I\$3*B23+(1-I\$3)*C23	=ABS(C24-B24)/B24	=ABS(B24-C24)
25	22	242	=I\$3*B24+(1-I\$3)*C24	=ABS(C25-B25)/B25	=ABS(B25-C25)
26	23	239	=I\$3*B25+(1-I\$3)*C25	=ABS(C26-B26)/B26	=ABS(B26-C26)
27	24	266	=I\$3*B26+(1-I\$3)*C26	=ABS(C27-B27)/B27	=ABS(B27-C27)
28			=I\$3*B27+(1-I\$3)*C27		

Figure: 66.82 - Formules correspondantes au modèle de lissage exponentiel simple avec Microsoft Excel 14.0.6123

où le lecteur attentif aura remarqué la prévision à la ligne 28 sur un horizon de un mois.

Avec un peu plus loin dans la feuille les valeurs indispensables à l'optimisation du modèle avec le solveur que nous allons de suite voir:

	H	I
3	λ :	
4	Paramètre de lissage:	=1-I3
5		
6	MAPE:	=AVERAGE(E5:E27)
7		
8	MAD:	=AVERAGE(F5:F27)
9		
10	MSD:	=SUM((F5:F27)*(F5:F27))/COUNT(F5:F27)

Figure: 66.83 - Constante de lissage et indicateurs d'erreur pour optimiser le modèle avec Microsoft Excel 14.0.6123

Maintenant, nous allons utiliser le solveur pour minimiser soit l'erreur MAPE, soit l'erreur MAD ou enfin l'erreur MSD. Nous avons alors par exemple en minimisant l'erreur MAPE:

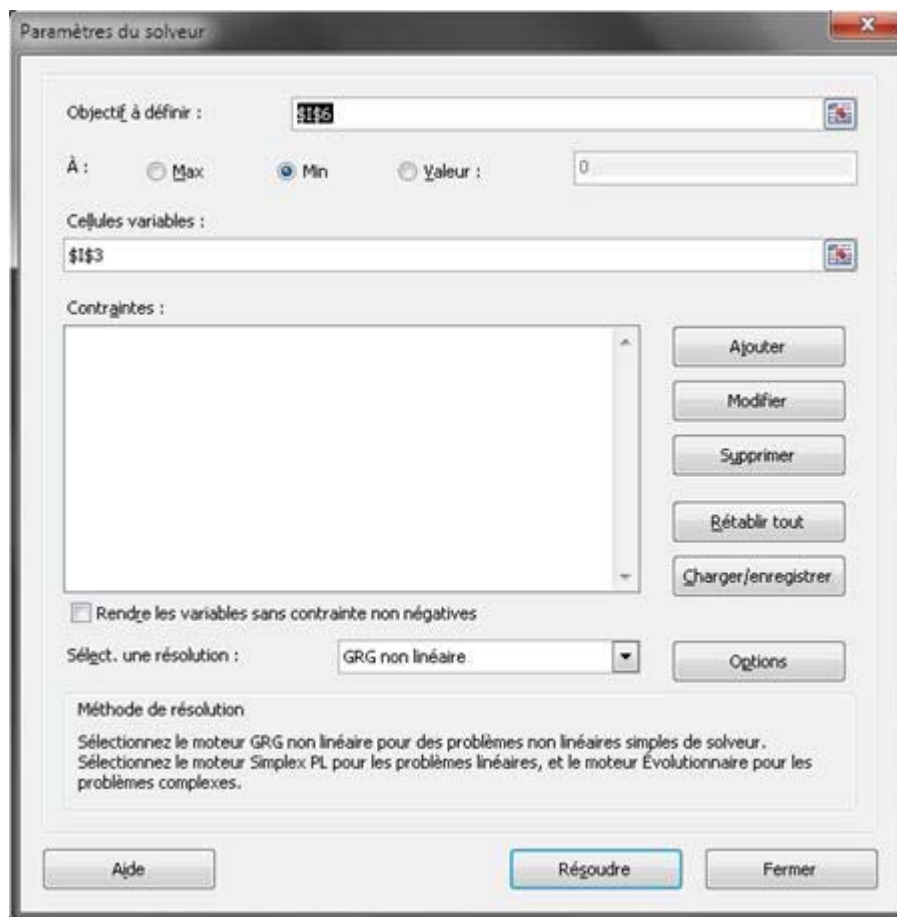


Figure: 66.84 - Paramétrage du solveur avec Microsoft Excel 14.0.6123

Ce qui donne le résultat suivant:

	H	I
3	λ :	0.64447
4	Paramètre de lissage:	0.35553
5		
6	MAPE:	7.89%
7		
8	MAD:	14.71
9		
10	MSD:	311.75

Figure: 66.85 - Solution proposée par le solveur de Microsoft Excel 14.0.6123

Soit graphiquement:

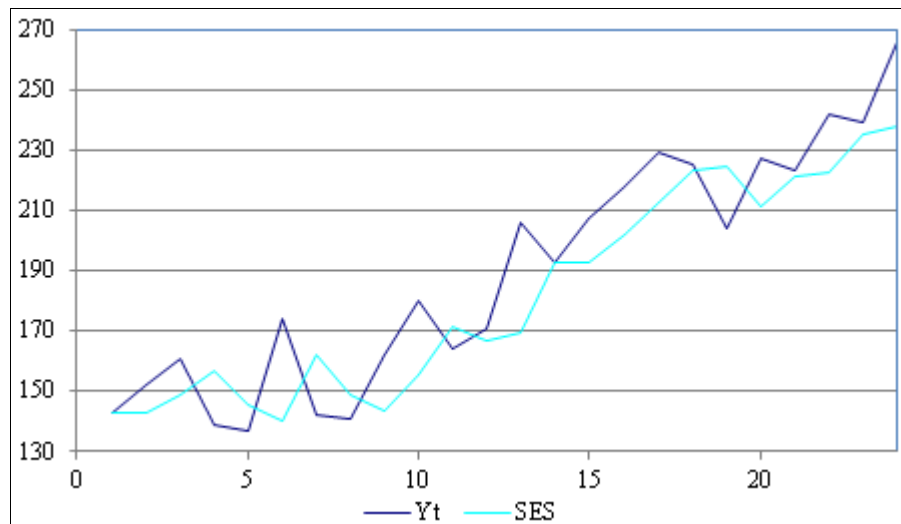


Figure: 66.86 - Tracé de la comparaison des observations et du modèle SES avec Microsoft Excel 14.0.6123 (sans la prévision)

et donc le tableau de valeurs devient:

	A	B	C	E	F
3	Période	Y_t	SES	Écart relatif absolu	Écart absolu
4	1	143	143		
5	2	152	143.00	0.06	9.00
6	3	161	148.80	0.08	12.20
7	4	139	156.66	0.13	17.66
8	5	137	145.28	0.06	8.28
9	6	174	139.94	0.20	34.06
10	7	142	161.89	0.14	19.89
11	8	141	149.07	0.06	8.07
12	9	162	143.87	0.11	18.13
13	10	180	155.55	0.14	24.45
14	11	164	171.31	0.04	7.31
15	12	171	166.60	0.03	4.40
16	13	206	169.44	0.18	36.56
17	14	193	193.00	0.00	0.00
18	15	207	193.00	0.07	14.00
19	16	218	202.02	0.07	15.98
20	17	229	212.32	0.07	16.68
21	18	225	223.07	0.01	1.93
22	19	204	224.31	0.10	20.31
23	20	227	211.22	0.07	15.78
24	21	223	221.39	0.01	1.61
25	22	242	222.43	0.08	19.57
26	23	239	235.04	0.02	3.96
27	24	266	237.59	0.11	28.41
28			255.90		

Figure: 66.87 - Tableau correspondant après optimisation avec le solveur dans Microsoft Excel 14.0.6123

et nous voyons donc la prévision à la ligne 28 du tableau qui vaut:

$$\hat{F}_{24+1} = 255.90 \quad (66.734)$$

Pour clore, indiquons que si nous faisons l'hypothèse forte que pour tout temps t :

$$Y_{t+1} = a_t + \varepsilon_t = c^{te} + N(0, \sigma_{Y-\bar{Y}}^2) \quad (66.735)$$

Alors dans le cas d'une série infinie (ou supposée comme telle...):

$$\begin{aligned} V(F_{t+1}) &= V\left((1-\beta) \sum_{i=0}^{t-1} \beta^i Y_{t-1}\right) = (1-\beta)^2 V\left(\sum_{i=0}^{t-1} \beta^i Y_{t-1}\right) \\ &= (1-\beta)^2 V\left(\sum_{i=0}^{t-1} \beta^i (c^{te} + \varepsilon_t)\right) = (1-\beta)^2 V\left(\sum_{i=0}^{t-1} \beta^i N(0, \sigma_{Y-\bar{Y}}^2)\right) \\ &= (1-\beta)^2 \sum_{i=0}^{t-1} \beta^{2i} \sigma_{Y-\bar{Y}}^2 = \sigma_{Y-\bar{Y}}^2 (1-\beta)^2 \sum_{i=0}^{t-1} \beta^{2i} \\ &= \sigma_{Y-\bar{Y}}^2 (1-\beta)^2 \sum_{i=0}^{t-1} \beta^{2i} = \sigma_{Y-\bar{Y}}^2 (1-\beta)^2 \frac{1-\beta^{2(t-1)+2}}{1-\beta^2} \\ &= \sigma_{Y-\bar{Y}}^2 (1-\beta)^2 \frac{1-\beta^{2t}}{1-\beta^2} = \sigma_{Y-\bar{Y}}^2 \frac{(1-\beta)(1-\beta^{2t})}{1+\beta} \end{aligned} \quad (66.736)$$

Or, comme $\beta \in]0, 1[$ et si t est suffisamment grand, il vient:

$$V(F_{t+1}) = \sigma_{Y-\bar{Y}}^2 \frac{(1-\beta)(1-\beta^{2t})}{1+\beta} \cong \sigma_{Y-\bar{Y}}^2 \frac{1-\beta}{1+\beta} = \sigma_{Y-\bar{Y}}^2 \frac{\lambda}{2-\lambda} \quad (66.737)$$

et considérons l'erreur:

$$e(t) = Y_t - F_t \quad (66.738)$$

Nous avons sous l'hypothèse que les deux termes soustraits sont indépendants...:

$$\begin{aligned} V(e(t)) &= V(Y_t - F_t) = V(Y_t) + V(F_{t+1}) \cong V(Y - \bar{Y}) + V(F_{t+1}) \\ &= \sigma_{Y-\bar{Y}}^2 + \sigma_{Y-\bar{Y}}^2 \frac{\lambda}{2-\lambda} = \sigma_{Y-\bar{Y}}^2 \left(1 + \frac{\lambda}{2-\lambda}\right) = \frac{2\sigma_{Y-\bar{Y}}^2}{2-\lambda} \end{aligned} \quad (66.739)$$

Il s'ensuit sous l'hypothèse de Normalité (ce qui est le cas depuis le début de ce développement) et que F_t est l'espérance de Y_t :

$$\frac{Y_t - F_t}{\sigma_{Y-\bar{Y}} \sqrt{\frac{2}{2-\lambda}}} = N(0, 1) \quad (66.740)$$

et donc sous l'hypothèse que la variance est parfaitement connue, il vient:

$$Y_t = F_t \pm Z_{\alpha/2} \sigma_{Y-\bar{Y}} \sqrt{\frac{2}{2-\lambda}} \quad (66.741)$$

Mais si l'écart-type est seulement estimé, il faut évidemment utiliser la loi de Student (cf. chapitre de Statistiques):

$$Y_t = F_t \pm T_{\alpha/2} \hat{\sigma}_{Y-\bar{Y}} \sqrt{\frac{2}{2-\lambda}} \quad (66.742)$$

Et ce sera tout en ce qui concerne le modèle de lissage exponentiel simple.

8.3.4. LISSAGE EXPONENTIEL DOUBLE À UN PARAMÈTRE (MÉTHODE DE BROWN)

L'idée du lissage exponentiel double (DEBS: "Double Exponential Brown Smoothing" en anglais) est à nouveau de réduire l'erreur totale mais cette fois-ci en utilisant non pas une simple constante mais une droite au voisinage de t telle que:

$$E = \sum_{j=0}^{t-1} \beta^j (Y_{t-j} - (a_t j + b_t))^2 \quad (66.743)$$

ce qui est écrit traditionnellement (le $-$ est absorbé dans le paramètre a et le fait que les paramètres de la droite soient recalculés à chaque t est omis afin de ne pas le confondre avec l'indice de sommation...):

$$E = \sum_{j=0}^{t-1} \beta^j (Y_{t-j} - b + a j)^2 \quad (66.744)$$

Pour minimiser nous différencions E :

$$\begin{aligned} \frac{\partial E}{\partial b} &= -2 \sum_{j=0}^{t-1} \beta^j (Y_{t-j} - b + a j) = 0 \\ \frac{\partial E}{\partial a} &= 2 \sum_{j=0}^{t-1} j \beta^j (Y_{t-j} - b + a j) = 0 \end{aligned} \quad (66.745)$$

d'où:

$$\begin{aligned} \sum_{j=0}^{t-1} \beta^j Y_{t-j} - b \sum_{j=0}^{t-1} \beta^j + a \sum_{j=0}^{t-1} j \beta^j &= 0 \\ \sum_{j=0}^{t-1} j \beta^j Y_{t-j} - b \sum_{j=0}^{t-1} j \beta^j + a \sum_{j=0}^{t-1} j^2 \beta^j &= 0 \end{aligned} \quad (66.746)$$

Simplifions tout cela en remplaçant d'abord lorsque t est grand:

$$\sum_{j=0}^{t-1} \beta^j = \frac{1 - \beta^t}{1 - \beta} = \frac{1}{1 - \beta} \quad (66.747)$$

Pour les autres, sous certaines conditions qui sont respectées dans notre cas, nous pouvons dériver terme à terme une série convergente pour obtenir la dérivée de celle-ci. Ainsi, étant donné que:

$$\frac{d}{d\beta} \sum_{j=0}^{t-1} \beta^j = \sum_{j=0}^{t-1} j \beta^{j-1} \quad (66.748)$$

et vu que dans les conditions susmentionnées:

$$\sum_{j=0}^{t-1} \beta^j = \frac{1}{1-\beta} \quad (66.749)$$

nous avons alors immédiatement:

$$\sum_{j=0}^{t-1} j \beta^{j-1} = \frac{d}{d\beta} \left(\frac{1}{1-\beta} \right) = \frac{1}{(1-\beta)^2} \quad (66.750).$$

En multipliant par β à gauche et à droite nous obtenons au final:

$$\sum_{j=0}^{t-1} j \beta^j = \frac{\beta}{(1-\beta)^2} \quad (66.751)$$

qui est la deuxième simplification recherchée.

Pour la troisième il suffit de dériver encore une fois l'égalité précédente. Nous obtenons alors:

$$\sum_{j=0}^{t-1} j^2 \beta^{j-1} = \frac{1}{(1-\beta)^2} + \frac{2\beta}{(1-\beta)^3} = \frac{(1-\beta) + 2\beta}{(1-\beta)^3} = \frac{1+\beta}{(1-\beta)^3} \quad (66.752)$$

Et nous terminons en multipliant à gauche et à droite par β :

$$\sum_{j=0}^{t-1} j^2 \beta^j = \frac{\beta(1+\beta)}{(1-\beta)^3} \quad (66.753)$$

Nous obtenons alors:

$$\begin{aligned} \sum_{j=0}^{t-1} \beta^j Y_{t-j} - b \sum_{j=0}^{t-1} \beta^j + a \sum_{j=0}^{t-1} j \beta^j &= \sum_{j=0}^{t-1} \beta^j Y_{t-j} - \frac{b}{1-\beta} + \frac{a\beta}{(1-\beta)^2} = 0 \\ \sum_{j=0}^{t-1} j \beta^j Y_{t-j} - b \sum_{j=0}^{t-1} j \beta^j + a \sum_{j=0}^{t-1} j^2 \beta^j &= \sum_{j=0}^{t-1} j \beta^j Y_{t-j} - \frac{b\beta}{(1-\beta)^2} + \frac{a\beta(1+\beta)}{(1-\beta)^3} = 0 \end{aligned} \quad (66.754)$$

Ce que nous pouvons écrire sous la forme:

$$\begin{aligned}
 (1-\beta) \sum_{j=0}^{t-1} \beta^j Y_{t-j} - b + a \frac{\beta}{1-\beta} &= 0 \\
 (1-\beta)^2 \sum_{j=0}^{t-1} j \beta^j Y_{t-j} - b \beta + a \frac{\beta(1+\beta)}{1-\beta} &= 0
 \end{aligned} \quad (66.755)$$

Définissons la série lissée:

$$S_1(t) = (1-\beta) \sum_{j=0}^{t-1} \beta^j Y_{t-j} = F_{t+1} \quad (66.756)$$

et la série doublement lissée (d'où le nom de la technique...):

$$\begin{aligned}
 S_2(t) &= (1-\beta) \sum_{j=0}^{t-1} \beta^j S_1(t-j) = (1-\beta)^2 \sum_{j=0}^{t-1} \beta^j \sum_{i=0}^{t-j-1} \beta^i X_{t-j-i} \\
 &= (1-\beta)^2 \sum_{j=0}^{t-1} \sum_{i=0}^{t-j-1} \beta^{i+j} X_{t-j-i} = F'_{t+1}
 \end{aligned} \quad (66.757)$$

Après le changement de variable $i+j=k \Leftrightarrow -i-j=-k$ nous obtenons:

$$S_2(t) = (1-\beta)^2 \sum_{j=0}^{t-1} \sum_{k=j}^{t-1} \beta^k X_{t-k} \quad (66.758)$$

L'indice de la deuxième somme s'expliquant par le fait que k est la nouvelle variable d'indexation (donc elle doit forcément être dans l'indice!) et quand i est nul nous avons:

$$i+j=k \xRightarrow{i=0} 0+j=k \Leftrightarrow k=j \quad (66.759)$$

Pour le suffixe de la somme il s'explique par le fait que pour la valeur maximale nous avons $i=t-j-1$ et donc comme $i+j=k$, il vient si nous y substituons i :

$$(t-j-1)+j=k \Leftrightarrow t-1=k \quad (66.760)$$

Ensuite, le lecteur vérifiera lui-même.... sur papier que nous pouvons réarranger les termes de la manière suivante:

$$S_2(t) = (1-\beta)^2 \sum_{j=0}^{t-1} \sum_{k=j}^{t-1} \beta^k X_{t-k} = (1-\beta)^2 \sum_{k=0}^{t-1} \sum_{j=0}^k \beta^k X_{t-k} \quad (66.761)$$

Nous avons alors:

$$\begin{aligned}
 S_2(t) &= (1-\beta)^2 \sum_{k=0}^{t-1} \sum_{j=0}^k \beta^k X_{t-k} = (1-\beta)^2 \sum_{k=0}^{t-1} (k+1) \beta^k X_{t-k} \\
 &= (1-\beta)^2 \sum_{k=0}^{t-1} k \beta^k X_{t-k} + (1-\beta)^2 \sum_{k=0}^{t-1} \beta^k X_{t-k} = (1-\beta)^2 \sum_{k=0}^{t-1} k \beta^k X_{t-k} + (1-\beta) S_1(t)
 \end{aligned} \quad (66.762)$$

En substituant ces résultats dans:

$$\begin{aligned}
 (1-\beta) \sum_{j=0}^{t-1} \beta^j Y_{t-j} - b + a \frac{\beta}{1-\beta} &= 0 \\
 (1-\beta)^2 \sum_{j=0}^{t-1} j \beta^j Y_{t-j} - b\beta + a \frac{\beta(1+\beta)}{1-\beta} &= 0
 \end{aligned}
 \tag{66.763}$$

il vient:

$$\begin{aligned}
 S_1(t) - b + a \frac{\beta}{1-\beta} &= 0 \\
 S_2(t) - (1-\beta) S_1(t) - b\beta + a \frac{\beta(1+\beta)}{1-\beta} &= 0
 \end{aligned}
 \tag{66.764}$$

Il vient alors après simplification successives:

$$\begin{aligned}
 S_1(t) - b - \frac{\beta}{1-\beta} \left[\frac{1-\beta}{\beta(1+\beta)} (S_2(t) - (1-\beta) S_1(t) - b\beta) \right] &= 0 \\
 S_1(t) - b - \frac{1}{1+\beta} (S_2(t) - (1-\beta) S_1(t) - b\beta) &= 0 \\
 b - \frac{1}{1+\beta} b\beta = S_1(t) - \frac{1}{1+\beta} (S_2(t) - (1-\beta) S_1(t)) & \\
 (1+\beta)b - b\beta = (1+\beta) S_1(t) - (S_2(t) - (1-\beta) S_1(t)) & \\
 b[(1+\beta) - \beta] = (1+\beta) S_1(t) - S_2(t) + (1-\beta) S_1(t) & \\
 b = 2S_1(t) - S_2(t) &
 \end{aligned}
 \tag{66.765}$$

et donc aussi après simplifications successives:

$$\begin{aligned}
 S_1(t) - b + a \frac{\beta}{1-\beta} &= 0 \\
 S_1(t) - (2S_1(t) - S_2(t)) + a \frac{\beta}{1-\beta} &= 0 \\
 -S_1(t) + S_2(t) + a \frac{\beta}{1-\beta} &= 0 \\
 a = \frac{1-\beta}{\beta} (S_1(t) - S_2(t)) &
 \end{aligned}
 \tag{66.766}$$

Résultats qu'il est d'usage de noter:

$$\begin{aligned}
 a &= \frac{\lambda}{1-\lambda} (S_1(t) - S_2(t)) = \gamma (S_1(t) - S_2(t)) \\
 b &= 2S_1(t) - S_2(t)
 \end{aligned}
 \tag{66.767}$$

Donc:

$$S_t = a_t j + b_t \tag{66.768}$$

est la droite constituée d'un double lissage exponentiel qui approche au mieux (l'estimé) à un temps t les données réelles et qui minimise l'erreur totale. Contrairement au lissage exponentiel simple, nous pouvons par contre maintenant faire des prévisions (cependant cela reste un usage abusif du résultat mathématique) de par la présence de la variable j .

Explicitement, cette droite s'écrit en changeant un peu les notations (ne pas oublier qu'au début nous avions implicitement injecté un signe négatif dans le terme a):

$$a_t j + b_t = \frac{\lambda}{1-\lambda} (\hat{F}_{t+1}' - F_{t+1}) j + (2F_{t+1} - \hat{F}_{t+1}') \quad (66.769)$$

Il vient donc que notre prévision s'écrit:

$$\hat{F}_{t-j}'' = \frac{\lambda}{1-\lambda} (\hat{F}_{t+1}' - F_{t+1}) j + (2F_{t+1} - \hat{F}_{t+1}') \quad (66.770)$$

Et évidemment, si nous voulons $\hat{F}_{t+1}''$, il faudra choisir $j = -1$. Dès lors:

$$\hat{F}_{t+1}'' = \frac{\lambda}{1-\lambda} (\hat{F}_{t+1}' - F_{t+1}) + (2F_{t+1} - \hat{F}_{t+1}') \quad (66.771)$$

Pour débiter les prévisions avec ce modèle, il faut un estimé de $S_1(0)$ et $S_2(0)$. Les premières prévisions seront grandement affectées par ces estimés. Il faut aussi toujours vérifier ce que les logiciels utilisent comme méthode de détermination pour ces premiers paramètres afin de la spécifier dans les rapports.

Remarque: Le lissage exponentiel de Brown est rarement intégré dans les logiciels statistiques. On lui préfère le modèle de Holt que nous allons de suite voir après et qui le contient implicitement.

Exemple:

Considérons le tableau suivant dans Microsoft Excel 14.0.6123 qui contient les données observées dans la colonne B ainsi que le modèle dans les colonnes C à F dont nous allons détailler les formules juste après:

	A	B	C	D	E	F	G	H	I
3	Période	Y_t	SES1	SES2	a	b	Projection $A + j$	Écart relatif absolu	Écart absolu
4	1	143.00	143.00	143.00	-	143.00	143.00		
5	2	152.00	143.00	143.00	-	143.00	143.00	0.00	0.00
6	3	161.00	145.55	143.00	- 1.01	148.09	148.09	0.08	12.91
7	4	139.00	149.92	143.72	- 2.45	156.12	156.12	0.12	17.12
8	5	137.00	146.83	145.48	- 0.53	148.18	148.18	0.08	11.18
9	6	174.00	144.05	145.86	0.72	142.24	142.24	0.18	31.76
10	7	142.00	152.53	145.35	- 2.83	159.70	159.70	0.12	17.70
11	8	141.00	149.55	147.38	- 0.86	151.71	151.71	0.08	10.71
12	9	162.00	147.13	147.99	0.34	146.26	146.26	0.10	15.74
13	10	180.00	151.34	147.75	- 1.42	154.93	154.93	0.14	25.07
14	11	164.00	159.45	148.76	- 4.22	170.14	170.14	0.04	6.14
15	12	171.00	160.74	151.79	- 3.53	169.69	169.69	0.01	1.31
16	13	206.00	163.64	154.32	- 3.68	172.96	172.96	0.16	33.04
17	14	193.00	175.63	156.96	- 7.37	194.30	194.30	0.01	1.30
18	15	207.00	180.55	162.24	- 7.23	198.85	198.85	0.04	8.15
19	16	218.00	188.03	167.43	- 8.14	208.64	208.64	0.04	9.36
20	17	229.00	196.52	173.26	- 9.18	219.77	219.77	0.04	9.23
21	18	225.00	205.71	179.84	- 10.21	231.58	231.58	0.03	6.58
22	19	204.00	211.17	187.16	- 9.48	235.18	235.18	0.15	31.18
23	20	227.00	209.14	193.96	- 5.99	224.32	224.32	0.01	2.68
24	21	223.00	214.20	198.26	- 6.29	230.14	230.14	0.03	7.14
25	22	242.00	216.69	202.77	- 5.50	230.61	230.61	0.05	11.39
26	23	239.00	223.85	206.71	- 6.77	241.00	241.00	0.01	2.00
27	24	266.00	228.14	211.56	- 6.55	244.72	244.72	0.08	21.28
28			238.86	216.25	- 8.92	261.46	261.46		

Figure: 66.88 - Données de départ avec modèle de lissage exponentiel double de Brown dans Microsoft Excel 14.0.6123

Indiquons les formules pour les C, D, E, F, G (H et I étant les mêmes que pour le lissage exponentiel simple). Il vient alors:

	A	B	C	D	E	F	G
3	Période	Y_t	SES1	SES2	a	b	Projection $A + j$
4	1	143	=B4	143	0	=2*C4-D4	=E4*\$L\$4+F4
5	2	152	=L\$3*B4+(1-L\$3)*C4	=L\$3*C4+(1-L\$3)*D4	0	=2*C5-D5	=E5*\$L\$4+F5
6	3	161	=L\$3*B5+(1-L\$3)*C5	=L\$3*C5+(1-L\$3)*D5	=L\$3/(1-L\$3)*(D6-C6)	=2*C6-D6	=E6*\$L\$4+F6
7	4	139	=L\$3*B6+(1-L\$3)*C6	=L\$3*C6+(1-L\$3)*D6	=L\$3/(1-L\$3)*(D7-C7)	=2*C7-D7	=E7*\$L\$4+F7
8	5	137	=L\$3*B7+(1-L\$3)*C7	=L\$3*C7+(1-L\$3)*D7	=L\$3/(1-L\$3)*(D8-C8)	=2*C8-D8	=E8*\$L\$4+F8
9	6	174	=L\$3*B8+(1-L\$3)*C8	=L\$3*C8+(1-L\$3)*D8	=L\$3/(1-L\$3)*(D9-C9)	=2*C9-D9	=E9*\$L\$4+F9
10	7	142	=L\$3*B9+(1-L\$3)*C9	=L\$3*C9+(1-L\$3)*D9	=L\$3/(1-L\$3)*(D10-C10)	=2*C10-D10	=E10*\$L\$4+F10
11	8	141	=L\$3*B10+(1-L\$3)*C10	=L\$3*C10+(1-L\$3)*D10	=L\$3/(1-L\$3)*(D11-C11)	=2*C11-D11	=E11*\$L\$4+F11
12	9	162	=L\$3*B11+(1-L\$3)*C11	=L\$3*C11+(1-L\$3)*D11	=L\$3/(1-L\$3)*(D12-C12)	=2*C12-D12	=E12*\$L\$4+F12
13	10	180	=L\$3*B12+(1-L\$3)*C12	=L\$3*C12+(1-L\$3)*D12	=L\$3/(1-L\$3)*(D13-C13)	=2*C13-D13	=E13*\$L\$4+F13
14	11	164	=L\$3*B13+(1-L\$3)*C13	=L\$3*C13+(1-L\$3)*D13	=L\$3/(1-L\$3)*(D14-C14)	=2*C14-D14	=E14*\$L\$4+F14
15	12	171	=L\$3*B14+(1-L\$3)*C14	=L\$3*C14+(1-L\$3)*D14	=L\$3/(1-L\$3)*(D15-C15)	=2*C15-D15	=E15*\$L\$4+F15
16	13	206	=L\$3*B15+(1-L\$3)*C15	=L\$3*C15+(1-L\$3)*D15	=L\$3/(1-L\$3)*(D16-C16)	=2*C16-D16	=E16*\$L\$4+F16
17	14	193	=L\$3*B16+(1-L\$3)*C16	=L\$3*C16+(1-L\$3)*D16	=L\$3/(1-L\$3)*(D17-C17)	=2*C17-D17	=E17*\$L\$4+F17
18	15	207	=L\$3*B17+(1-L\$3)*C17	=L\$3*C17+(1-L\$3)*D17	=L\$3/(1-L\$3)*(D18-C18)	=2*C18-D18	=E18*\$L\$4+F18
19	16	218	=L\$3*B18+(1-L\$3)*C18	=L\$3*C18+(1-L\$3)*D18	=L\$3/(1-L\$3)*(D19-C19)	=2*C19-D19	=E19*\$L\$4+F19
20	17	225	=L\$3*B19+(1-L\$3)*C19	=L\$3*C19+(1-L\$3)*D19	=L\$3/(1-L\$3)*(D20-C20)	=2*C20-D20	=E20*\$L\$4+F20
21	18	225	=L\$3*B20+(1-L\$3)*C20	=L\$3*C20+(1-L\$3)*D20	=L\$3/(1-L\$3)*(D21-C21)	=2*C21-D21	=E21*\$L\$4+F21
22	19	204	=L\$3*B21+(1-L\$3)*C21	=L\$3*C21+(1-L\$3)*D21	=L\$3/(1-L\$3)*(D22-C22)	=2*C22-D22	=E22*\$L\$4+F22
23	20	227	=L\$3*B22+(1-L\$3)*C22	=L\$3*C22+(1-L\$3)*D22	=L\$3/(1-L\$3)*(D23-C23)	=2*C23-D23	=E23*\$L\$4+F23
24	21	223	=L\$3*B23+(1-L\$3)*C23	=L\$3*C23+(1-L\$3)*D23	=L\$3/(1-L\$3)*(D24-C24)	=2*C24-D24	=E24*\$L\$4+F24
25	22	242	=L\$3*B24+(1-L\$3)*C24	=L\$3*C24+(1-L\$3)*D24	=L\$3/(1-L\$3)*(D25-C25)	=2*C25-D25	=E25*\$L\$4+F25
26	23	239	=L\$3*B25+(1-L\$3)*C25	=L\$3*C25+(1-L\$3)*D25	=L\$3/(1-L\$3)*(D26-C26)	=2*C26-D26	=E26*\$L\$4+F26
27	24	266	=L\$3*B26+(1-L\$3)*C26	=L\$3*C26+(1-L\$3)*D26	=L\$3/(1-L\$3)*(D27-C27)	=2*C27-D27	=E27*\$L\$4+F27
28			=L\$3*B27+(1-L\$3)*C27	=L\$3*C27+(1-L\$3)*D27	=L\$3/(1-L\$3)*(D28-C28)	=2*C28-D28	=E28*\$L\$4+F28

Figure: 66.89 - Formules correspondantes au modèle de lissage de Brown avec Microsoft Excel 14.0.6123

Avec un peu plus loin dans la feuille les valeurs indispensables à l'optimisation du modèle avec le solveur que nous allons de suite voir:

	K	L
3	λ : 0.283044871297282	
4	j (≤ 0): 0	
5		
6	MAPE: =MOYENNE(H5:H27)	
7		
8	MAD: =MOYENNE(I5:I27)	
9		
10	MSD: =SOMME((I5:I27)*(I5:I27))/NB(I5:I27)	

Figure: 66.90 - Constante de lissage et indicateurs d'erreur pour optimiser le modèle avec Microsoft Excel 14.0.6123

La procédure pour minimiser soit l'erreur MAPE, soit l'erreur MAD ou enfin l'erreur MSD avec le solveur est exactement la même que pour le lissage exponentiel simple! Nous ne la détaillerons donc pas.

Le résultat donnera graphiquement (et aussi en comparaison avec le lissage exponentiel simple):

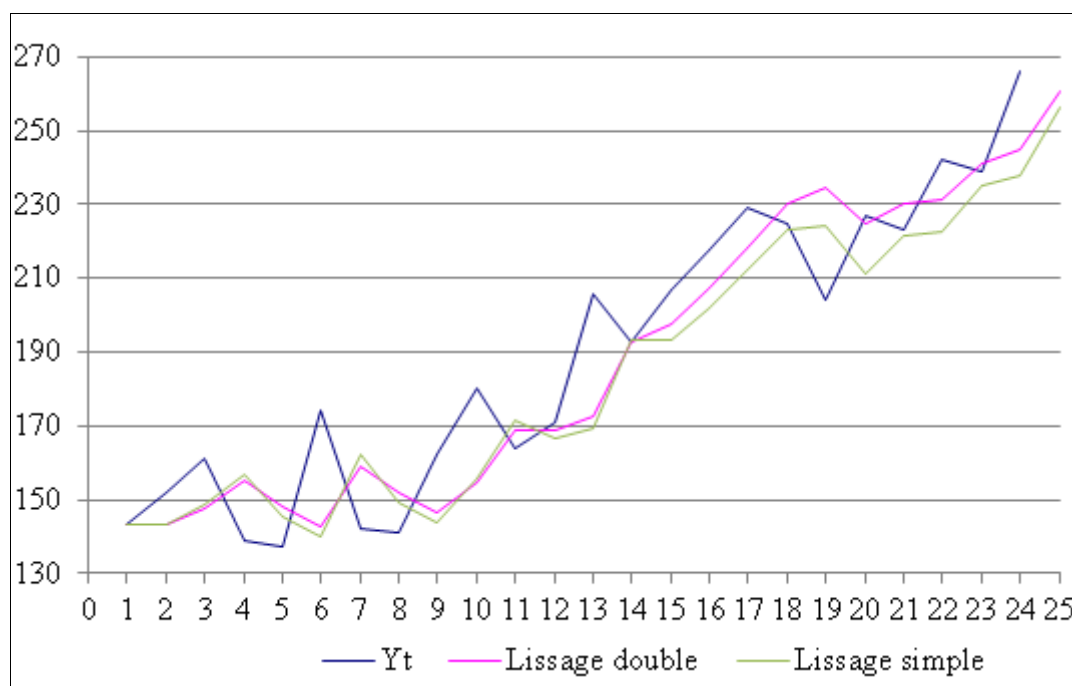


Figure: 66.91 - Tracé de la comparaison des observations et du modèle SES/DEBS avec Microsoft Excel 14.0.6123 (avec la prévision)

Nous voyons que le lissage double est meilleur que le lissage simple (son MAPE minimisé étant de 6.90% contre 7.89% pour le lissage simple). Cette comparaison est pour rappel une forme de "back-testing". La projection à la 25ème période est par contre très proche.

Ce qui est ensuite intéressant et particulier avec le lissage exponentiel double est de jouer avec la valeur de j . Ainsi, en imposant $j = -1$, nous aurons pour chaque valeur de la colonne G, sa projection à la période correspondante à sa ligne plus $|j|$ périodes. Évidemment, pour les deux premières lignes, les projections ne sont pas à prendre en considération!

Faisons un petit exemple visuel. En posant $j = -6$ (donc projection à six mois...) nous avons alors:

	A	B	G
3	Période	Y_t	Projection $A+ j $
4	1	143.00	143.00
5	2	152.00	143.00
6	3	161.00	152.88
7	4	139.00	168.04
8	5	137.00	151.62
9	6	174.00	139.65
10	7	142.00	173.58
11	8	141.00	156.87
12	9	162.00	145.69
13	10	180.00	162.35
14	11	164.00	191.10
15	12	171.00	188.33
16	13	206.00	192.82
17	14	193.00	232.13
18	15	207.00	237.33
19	16	218.00	252.41
20	17	229.00	269.48
21	18	225.00	287.20
22	19	204.00	288.34
23	20	227.00	261.50
24	21	223.00	268.19
25	22	242.00	264.55
26	23	239.00	280.53
27	24	266.00	283.20
28			311.19

Figure: 66.92 - Projection avec le modèle DEBS dans Microsoft Excel 14.0.6123

Ainsi, nous pouvons observer qu'au quatrième mois (période), le modèle donne une projection de 168.04 pour le dixième mois (période) et qu'en réalité il y a eu 180. Nous voyons un exemple d'une très mauvaise projection du modèle le sixième mois (période) où le modèle donne une projection de 139.65 alors qu'en réalité il y a eu 171...

Ensuite, ce qui est intéressant c'est de se rappeler que si nous posons $j = -6$, la 20ème période, projetée alors pour 26ème, la 21^{ème} sur la 27ème, etc. Nous obtenons alors graphiquement:

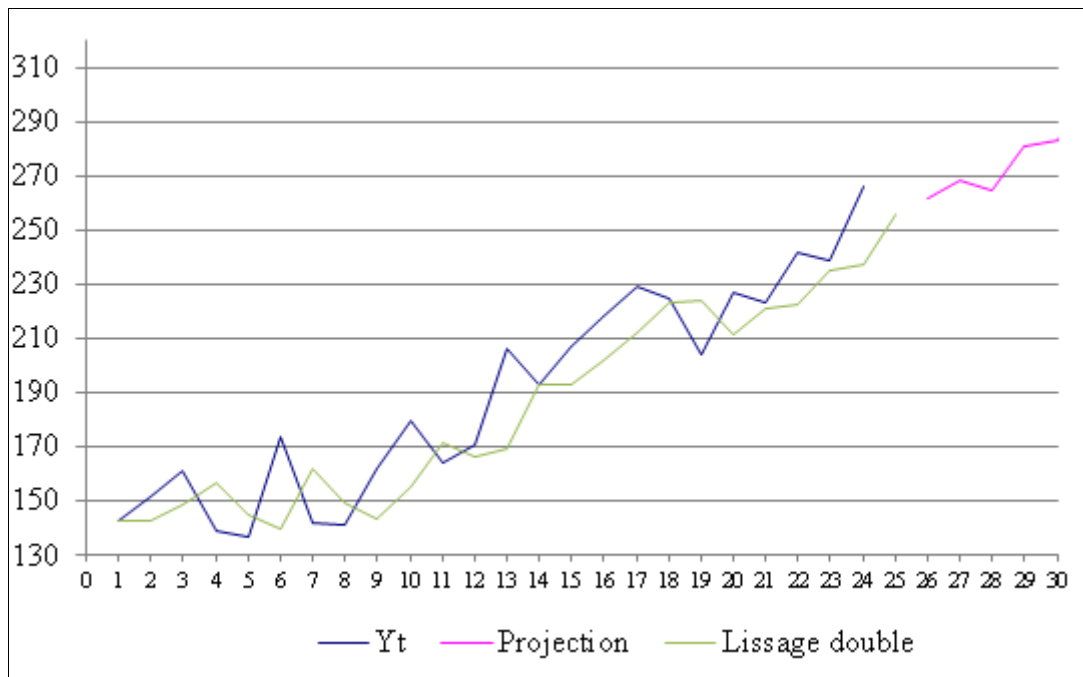


Figure: 66.93 - Projection sur un horizon de un semestre en utilisant le DEBS avec Microsoft Excel 14.0.6123

Évidemment, comme tout modèle, il faut être très prudent avec des projets à 6 mois car le modèle ne prend pas en compte les facteurs de la bourse, des catastrophes naturelles, des pandémies, des guerres, etc. Donc tant qu'il y n'a pas d'inférence statistique, il n'est pas possible de donner un intervalle de confiance pour la prévision ce qui est scientifiquement parlant pas vraiment acceptable.

8.3.5. MODÈLE DE LISSAGE EXPONENTIEL DOUBLE À 2 PARAMÈTRES DE HOLT (MODÈLE ADDITIF)

Nous allons présenter ici une technique dont la construction est empirique et qui est due à Charles C. Holt (1957). Même dans l'article d'origine de Holt il n'y a pas de démonstration rigoureuse de la construction de ce modèle. Il s'agit donc plus d'ingénierie mathématique que de mathématique pure à proprement parler (donc la présentation de ce modèle n'aurait normalement pas lieu d'être présente sur ce site Internet).

L'idée est la suivante comme pour le modèle précédent, d'avoir un modèle de la forme:

$$F_{t+j} = a_t j + b_t \quad (66.772)$$

mais avec une approche empirique qui consiste à s'imposer que:

- La pente a_t de la droite du modèle dont l'origine du repère est en t sera obtenue par lissage exponentiel simple de toutes les pentes des droites précédentes.
- L'ordonnée à l'origine b_t de la droite en t sera obtenue par lissage exponentiel simple de toutes les ordonnées à l'origine des droites précédentes.

Rappelons d'abord que nous avons obtenu pour le lissage exponentiel simple:

$$S_t = (1 - \beta)Y_t + \beta S_{t-1} = \lambda Y_t + (1 - \lambda)S_{t-1} \quad (66.773)$$

Il est clair que la pente de chaque droite de l'estimé est donnée par (voir figure ci-dessous):

$$\frac{\Delta S_t}{\Delta t} \quad (66.774)$$

et pour le dernier estimé, plus spécifique, nous aurons:

$$\frac{\Delta S_t}{\Delta t} = \frac{S_t - S_{t-1}}{t - (t-1)} = \frac{S_t - S_{t-1}}{1} = S_t - S_{t-1}$$

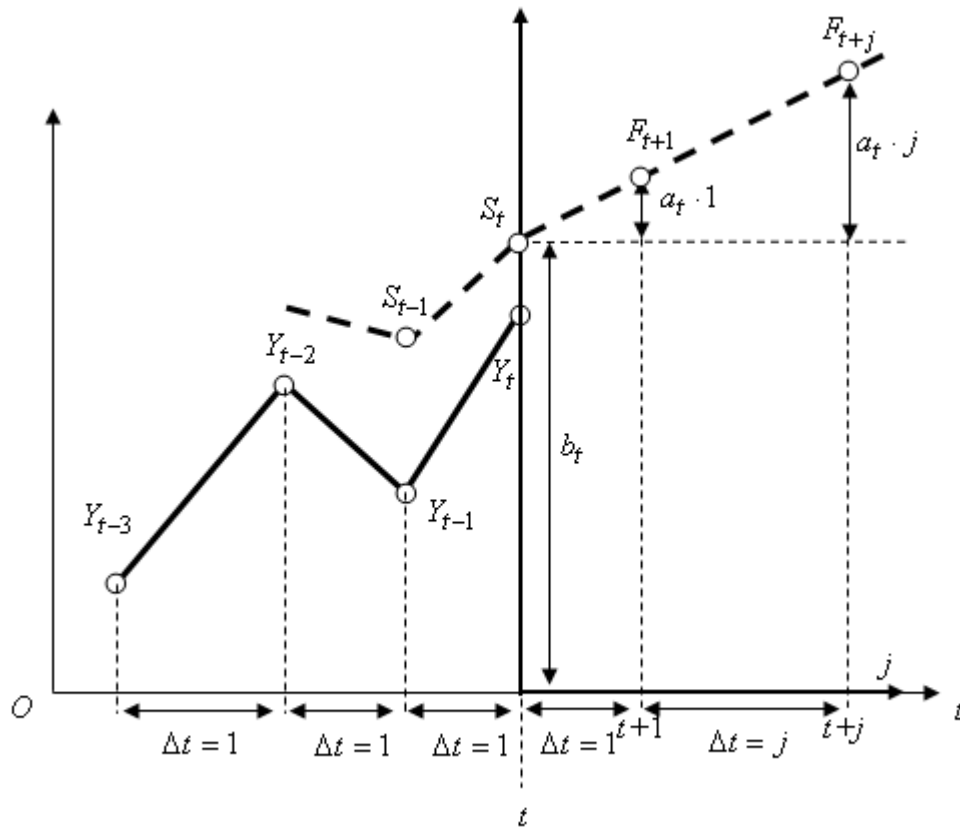


Figure: 66.94 - Schéma du principe de base de construction du modèle

et l'idée étant de faire un lissage exponentiel de la pente de toutes les droites dont la pente se calcule de la même manière, soit (à comparer à l'expression du lissage exponentiel simple):

$$a_t = \beta \frac{\Delta S_t}{\Delta t} + (1 - \beta) a_{t-1} = \beta (S_t - S_{t-1}) + (1 - \beta) a_{t-1} \quad (66.775)$$

Cette dernière expression est souvent appelée "**estimé de la pente à la période t**".

Maintenant, pour l'ordonnée à l'origine, nous remarquons sur la figure que $b_t = S_t$. Dès lors, il vient dans un premier temps que:

$$a_t = \beta \frac{\Delta S_t}{\Delta t} + (1 - \beta) a_{t-1} = \beta (b_t - b_{t-1}) + (1 - \beta) a_{t-1} \quad (66.776)$$

Une fois ce constat fait, nous choisissons un modèle où l'ordonnée à l'origine est un lissage exponentiel de toutes les ordonnées à l'origine avec sa propre constante de lissage précédente tel que:

$$b_t = S_t = \alpha Y_t + (1 - \alpha) (S_{t-1} + a_{t-1}) = \alpha Y_t + (1 - \alpha) (b_{t-1} + a_{t-1}) \quad (66.777)$$

Ce qui peut paraître parfois déroutant ici, c'est la dernière parenthèse. Au fait, techniquement, on devrait plutôt l'écrire:

$$(S_{t-1} + a_{t-1} \cdot \Delta t) = (b_{t-1} + a_{t-1} \cdot \Delta t) \quad (66.778)$$

mais comme le Δt est toujours égal à 1, nous omettons son écriture. Donc la dernière parenthèse représente l'estimé à une unité de temps précédente à laquelle on ajoute l'accroissement d'une unité de temps de la pente de l'estimé au même temps t .

L'expression:

$$S_t = \alpha Y_t + (1 - \alpha)(b_{t-1} + a_{t-1}) \quad (66.779)$$

est un peu malheureusement appelée dans la pratique "**estimé du niveau de la série à la période t** ".

Au final, nous travaillons avec les trois relations:

$$\begin{array}{l} F_{t+j} = a_t j + b_t \\ a_t = \beta(b_t - b_{t-1}) + (1 - \beta)a_{t-1} \\ b_t = \alpha Y_t + (1 - \alpha)(b_{t-1} + a_{t-1}) \end{array} \quad (66.780)$$

Pour débiter les prévisions avec ce modèle, il faut un estimé de S_0 et de a_0 . Les premières prévisions seront grandement affectées par ces estimés. Il faut aussi toujours vérifier ce que les logiciels utilisent comme méthode de détermination pour ces premiers paramètres afin de la spécifier dans les rapports.

Exemple:

Considérons le tableau suivant dans Microsoft Excel 14.0.6123 qui contient les données observées dans la colonne B ainsi que le modèle dans les colonnes C à F dont nous allons détailler les formules juste après:

	A	B	C	D	E	F	G	H
3	Période	Y_t	a_t	b_t	F_t		Écart relatif absolu	Écart absolu
4	1	143	9.00	143.00				
5	2	152	9.00	152.00	152.00		0.00	0.00
6	3	161	9.00	161.00	161.00		0.00	0.00
7	4	139	7.54	159.04	170.00		0.22	31.00
8	5	137	6.14	156.12	166.58		0.22	29.58
9	6	174	6.69	166.41	162.26		0.07	11.74
10	7	142	5.22	162.11	173.10		0.22	31.10
11	8	141	3.98	158.02	167.33		0.19	26.33
12	9	162	3.98	162.00	162.00		0.00	0.00
13	10	180	4.64	170.93	165.98		0.08	14.02
14	11	164	4.09	171.48	175.58		0.07	11.58
15	12	171	3.88	173.96	175.58		0.03	4.58
16	13	206	5.21	187.79	177.84		0.14	28.16
17	14	193	5.21	193.00	193.00		0.00	0.00
18	15	207	5.62	201.32	198.21		0.04	8.79
19	16	218	6.15	210.85	206.94		0.05	11.06
20	17	229	6.71	221.24	217.00		0.05	12.00
21	18	225	6.57	226.91	227.95		0.01	2.95
22	19	204	5.18	223.06	233.48		0.14	29.48
23	20	227	5.12	227.80	228.24		0.01	1.24
24	21	223	4.65	229.42	232.92		0.04	9.92
25	22	242	5.03	236.87	234.07		0.03	7.93
26	23	239	4.89	240.88	241.90		0.01	2.90
27	24	266	5.85	252.92	245.77		0.08	20.23
28	25				258.77 $j = 1$			
29	26				264.61 $j = 2$			
30	27				270.46 $j = 3$			
31	28				276.31 $j = 4$			
32	29				282.15 $j = 5$			
33	30				288.00 $j = 6$			

Figure: 66.95 - Données de départ avec modèle de lissage exponentiel double de Holt dans Microsoft Excel 14.0.6123

Indiquons les formules pour les C, D, E (G et H étant les mêmes que pour le lissage exponentiel simple). Il vient alors:

	A	B	C	D	E
3	Période	Y_t	a_t	b_t	F_t
4	1	143	=B5-B4	=B4	
5	2	152	=\$K\$4*(D5-D4)+(1-\$K\$4)*C4	=\$K\$3*(B5)+(1-\$K\$3)*(D4+C4)	=D4+C4
6	3	161	=\$K\$4*(D6-D5)+(1-\$K\$4)*C5	=\$K\$3*(B6)+(1-\$K\$3)*(D5+C5)	=D5+C5
7	4	139	=\$K\$4*(D7-D6)+(1-\$K\$4)*C6	=\$K\$3*(B7)+(1-\$K\$3)*(D6+C6)	=D6+C6
8	5	137	=\$K\$4*(D8-D7)+(1-\$K\$4)*C7	=\$K\$3*(B8)+(1-\$K\$3)*(D7+C7)	=D7+C7
9	6	174	=\$K\$4*(D9-D8)+(1-\$K\$4)*C8	=\$K\$3*(B9)+(1-\$K\$3)*(D8+C8)	=D8+C8
10	7	142	=\$K\$4*(D10-D9)+(1-\$K\$4)*C9	=\$K\$3*(B10)+(1-\$K\$3)*(D9+C9)	=D9+C9
11	8	141	=\$K\$4*(D11-D10)+(1-\$K\$4)*C10	=\$K\$3*(B11)+(1-\$K\$3)*(D10+C10)	=D10+C10
12	9	162	=\$K\$4*(D12-D11)+(1-\$K\$4)*C11	=\$K\$3*(B12)+(1-\$K\$3)*(D11+C11)	=D11+C11
13	10	180	=\$K\$4*(D13-D12)+(1-\$K\$4)*C12	=\$K\$3*(B13)+(1-\$K\$3)*(D12+C12)	=D12+C12
14	11	164	=\$K\$4*(D14-D13)+(1-\$K\$4)*C13	=\$K\$3*(B14)+(1-\$K\$3)*(D13+C13)	=D13+C13
15	12	171	=\$K\$4*(D15-D14)+(1-\$K\$4)*C14	=\$K\$3*(B15)+(1-\$K\$3)*(D14+C14)	=D14+C14
16	13	206	=\$K\$4*(D16-D15)+(1-\$K\$4)*C15	=\$K\$3*(B16)+(1-\$K\$3)*(D15+C15)	=D15+C15
17	14	193	=\$K\$4*(D17-D16)+(1-\$K\$4)*C16	=\$K\$3*(B17)+(1-\$K\$3)*(D16+C16)	=D16+C16
18	15	207	=\$K\$4*(D18-D17)+(1-\$K\$4)*C17	=\$K\$3*(B18)+(1-\$K\$3)*(D17+C17)	=D17+C17
19	16	218	=\$K\$4*(D19-D18)+(1-\$K\$4)*C18	=\$K\$3*(B19)+(1-\$K\$3)*(D18+C18)	=D18+C18
20	17	229	=\$K\$4*(D20-D19)+(1-\$K\$4)*C19	=\$K\$3*(B20)+(1-\$K\$3)*(D19+C19)	=D19+C19
21	18	225	=\$K\$4*(D21-D20)+(1-\$K\$4)*C20	=\$K\$3*(B21)+(1-\$K\$3)*(D20+C20)	=D20+C20
22	19	204	=\$K\$4*(D22-D21)+(1-\$K\$4)*C21	=\$K\$3*(B22)+(1-\$K\$3)*(D21+C21)	=D21+C21
23	20	227	=\$K\$4*(D23-D22)+(1-\$K\$4)*C22	=\$K\$3*(B23)+(1-\$K\$3)*(D22+C22)	=D22+C22
24	21	223	=\$K\$4*(D24-D23)+(1-\$K\$4)*C23	=\$K\$3*(B24)+(1-\$K\$3)*(D23+C23)	=D23+C23
25	22	242	=\$K\$4*(D25-D24)+(1-\$K\$4)*C24	=\$K\$3*(B25)+(1-\$K\$3)*(D24+C24)	=D24+C24
26	23	239	=\$K\$4*(D26-D25)+(1-\$K\$4)*C25	=\$K\$3*(B26)+(1-\$K\$3)*(D25+C25)	=D25+C25
27	24	266	=\$K\$4*(D27-D26)+(1-\$K\$4)*C26	=\$K\$3*(B27)+(1-\$K\$3)*(D26+C26)	=D26+C26
28	25				=D27+C27
29	26				=D27+C27*2
30	27				=D27+C27*3
31	28				=D27+C27*4
32	29				=D27+C27*5
33	30				=D27+C27*6

Figure: 66.96 - Formules correspondantes au modèle de lissage exponentiel de Holt avec Microsoft Excel 14.0.6123

Avec un peu plus loin dans la feuille les valeurs indispensables à l'optimisation du modèle avec le solveur que nous allons de suite voir:

	J	K
3	α :	0.353501814151959
4	β :	0.133684345261106
5		
6	MAPE:	=MOYENNE(G5:G27)
7		
8	MAD:	=MOYENNE(H5:H27)
9		
10	MSD:	=SOMME((H5:H27)*(H5:H27))/NB(H5:H27)

Figure: 66.97 - Constante de lissage et indicateurs d'erreur pour optimiser le modèle avec Microsoft Excel 14.0.6123

La procédure pour minimiser soit l'erreur MAPE, soit l'erreur MAD ou enfin l'erreur MSD avec le solveur est exactement la même que pour le lissage exponentiel simple! Nous ne la détaillerons donc pas.

Ce qui donne graphiquement (et aussi en comparaison avec le lissage simple et double):

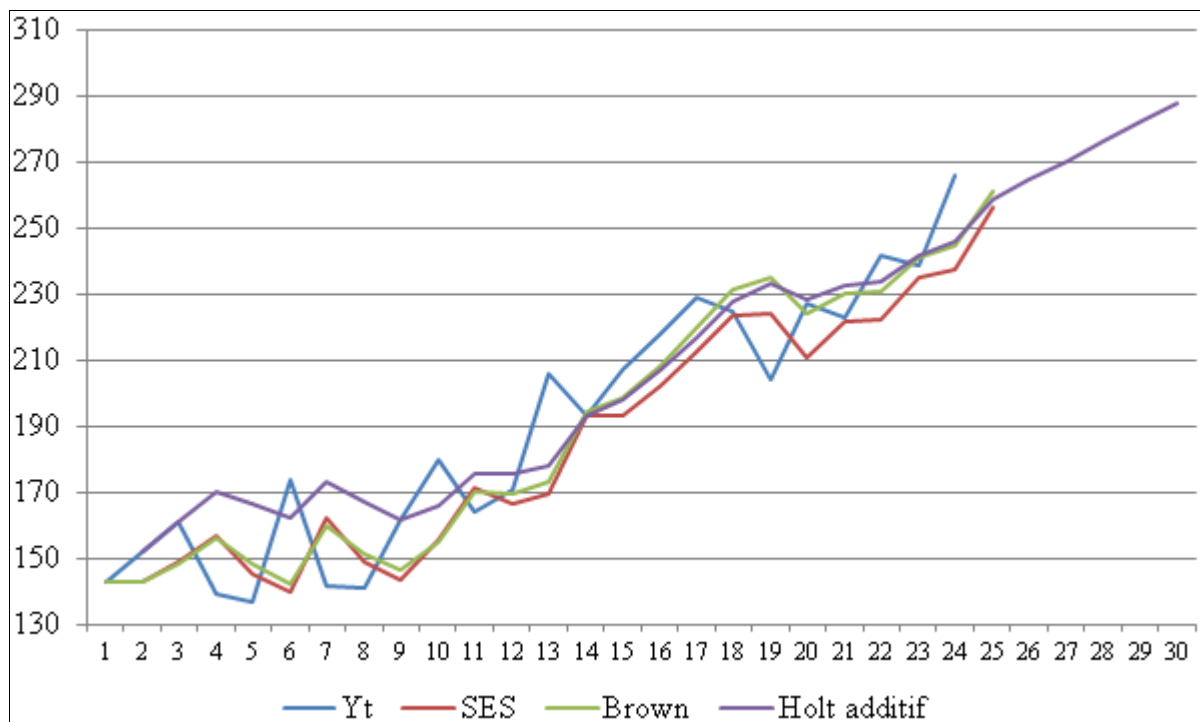


Figure: 66.98 - Projection sur un horizon de un semestre en utilisant le modèle de Holt avec Microsoft Excel 14.0.6123

Le lissage de Holt a un MAPE de 7.38% alors que celui de Brown à un MAPE de 7.12% (cette comparaison est pour rappel une forme de "back-testing"). Cependant l'avantage du modèle de Holt est qu'on peut faire des projections à plus d'une période.

Il est bien évidemment intéressant de comparer les projections à six mois entre la méthode additive de Holt et le lissage de Brown:

Brown	Holt's additif
288.34	258.77
261.50	264.61
268.19	270.46
264.55	276.31
280.53	282.15
283.20	288.00

Figure: 66.99 - Comparaison entre modèle de Brown et Holt additif

Nous voyons donc que le modèle de Holt's a presque toujours des valeurs de projection supérieures au modèle de Brown.

Il convient avant de juger ces valeurs de les comparer aux observations qui seront constatées dans l'avenir.

8.3.6. MODÈLE DE LISSAGE EXPONENTIEL TRIPLE À 3 PARAMÈTRES DE HOLT ET WINTERS (MODÈLE MULTIPLICATIF)

Nous allons présenter ici une technique dont la construction est elle aussi empirique et qui est fortement inspirée du modèle de lissage exponentiel double de Holt. Je n'ai jamais trouvé de démonstration rigoureuse de ce modèle. Il s'agit donc probablement aussi plus d'ingénierie mathématique que de mathématique pure à proprement parler mais comme c'est un classique dans les logiciels de statistiques, nous allons nous attarder dessus un petit moment.

Winters (étudiant de Holt) s'est basé en 1960 sur la méthode de son professeur pour définir un modèle qui prend en considération la composante saisonnière. L'idée consiste à utiliser trois équations de lissage: une pour le niveau de la demande (désaisonnalisée), une pour la tendance et une pour la saisonnalité.

Remarque: Il existe deux méthodes de Winters selon que la saisonnalité est additive ou multiplicative. Nous ne présentons ici que le modèle multiplicatif car c'est celui le plus couramment utilisé. Et puis nous ne pouvons pas présenter tous les modèles existants car ils sont tous empiriques et il en existe des centaines voire probablement des milliers!

Soit N , le nombre de périodes dans chaque cycle saisonnier (cycle supposé constant!). Trois équations de lissage exponentiel sont utilisées à chaque période pour actualiser les estimés de la série désaisonnalisée, les facteurs saisonniers et la tendance. Ces équations peuvent avoir différentes constantes de lissage notées traditionnellement α , β et γ .

La première équation est définie assez naturellement par:

$$b_t = \alpha \frac{Y_t}{C_{t-N}} + (1 - \alpha)(b_{t-1} + a_{t-1}) \quad (66.781)$$

où Y_t est alors le niveau actuel de la série désaisonnalisée par le facteur saisonnier approprié C_{t-N} (nous retrouvons sinon la méthode de Holt).

La deuxième équation est elle aussi naturellement définie par:

$$a_t = \beta(b_t - b_{t-1}) + (1 - \beta)a_{t-1} \quad (66.782)$$

où nous retrouvons donc la pente de la tendance identiquement au modèle de Holt.

La troisième équation est totalement nouvelle, ce qui n'empêche pas qu'elle est relativement intuitive et est donnée par:

$$C_t = \gamma \frac{Y_t}{b_t} + (1 - \gamma)C_{t-N} \quad (66.783)$$

et finalement la prévision est resaisonnalisée (c'est la forme de la prévision qui fait que nous parlons de "**modèle multiplicatif**"):

$$F_{t+j} = (a_t j + b_t)C_{t+j-N} \quad (66.784)$$

Cette dernière équation assume que $j < N$. Si $N < j < 2N$, le facteur saisonnier approprié serait C_{t+j-2N} . Si $2N < j < 3N$, le facteur saisonnier approprié serait C_{t+j-3N} et ainsi de suite.

Donc au final, nous avons le système:

$$\begin{aligned}
 F_{t+j} &= (a_t j + b_t) C_{t+j-N} \\
 a_t &= \beta (b_t - b_{t-1}) + (1 - \beta) a_{t-1} \\
 b_t &= \alpha \frac{Y_t}{C_{t-N}} + (1 - \alpha) (b_{t-1} + a_{t-1}) \\
 C_t &= \gamma \frac{Y_t}{b_t} + (1 - \gamma) C_{t-N}
 \end{aligned} \quad (66.785)$$

Pour débiter les prévisions avec ce modèle, il faut aussi un estimé des paramètres initiaux. Encore une fois, les premières prévisions seront grandement affectées par ces estimés. Il faut aussi toujours vérifier ce que les logiciels utilisent comme méthode de détermination pour ces premiers paramètres afin de la spécifier dans les rapports.

Exemple:

Considérons le tableau suivant dans Microsoft Excel 14.0.6123 qui contient les données observées dans la colonne D ainsi que le modèle dans les colonnes E à H dont nous allons détailler les formules juste après:

	A	B	C	D	E	F	G	H	I	J
3	Année	Trimestre	Période	Y_t	a_t	b_t	C_t	F_t	Écart relatif absolu	Écart absolu
4	1	1	1	143			0.961			
5	1	2	2	152			1.022			
6	1	3	3	161			1.082			
7	1	4	4	139	-0.06	148.75	0.934			
8	2	1	5	137	-0.28	146.66	0.961	142.94	0.04	5.94
9	2	2	6	174	0.55	154.21	1.022	149.59	0.14	24.41
10	2	3	7	142	-0.27	147.05	1.082	167.51	0.18	25.51
11	2	4	8	141	-0.12	148.13	0.934	137.16	0.03	3.84
12	3	1	9	162	0.59	154.72	0.961	142.28	0.12	19.72
13	3	2	10	180	1.31	162.13	1.022	158.70	0.12	21.30
14	3	3	11	164	0.90	159.54	1.082	176.90	0.08	12.90
15	3	4	12	171	1.68	167.82	0.934	149.92	0.12	21.08
16	4	1	13	206	3.23	184.17	0.961	162.95	0.21	43.05
17	4	2	14	193	3.28	187.88	1.022	191.49	0.01	1.51
18	4	3	15	207	3.28	191.19	1.082	206.90	0.00	0.10
19	4	4	16	218	4.63	207.18	0.934	181.72	0.17	36.28
20	5	1	17	229	5.54	220.46	0.961	203.62	0.11	25.38
21	5	2	18	225	5.34	224.10	1.022	230.94	0.03	5.94
22	5	3	19	204	3.92	216.02	1.082	248.33	0.22	44.33
23	5	4	20	227	4.72	227.47	0.934	205.53	0.09	21.47
24	6	1	21	223	4.71	232.12	0.961	223.21	0.00	0.21
25	6	2	22	242	4.71	236.83	1.022	242.00	0.00	0.00
26	6	3	23	239	3.99	234.75	1.082	261.43	0.09	22.43
27	6	4	24	266	5.58	253.78	0.934	223.09	0.16	42.91
28			25					249.34		
29			26					270.73		
30			27					292.81		
31			28					258.01		
32			29					265.44		
33			30					265.44		

Figure: 66.100 - Données de départ avec modèle de lissage exponentiel triple de Holt et Winters dans Microsoft Excel 14.0.6123

Indiquons les formules pour les E, F, G, H (I et J étant les mêmes que pour le lissage exponentiel simple). Il vient alors:

	D	E	F	G	H
3	γ	α	β	γ	δ
4	143			=D4/F7	
5	152			=D5/F7	
6	161			=D6/F7	
7	139	=((D8-D4)/4+(D9-D5)/4+(D10-D6)/4+(D11-D7)/4	=MOYENNE(D4:D7)	=D7/F7	
8	137	=(\$M\$4*(F8-F7)+(1-\$M\$4)*E7	=(\$M\$3*D8/G4+(1-\$M\$3)*(F7+E7)	=M5*D8/F8+(1-M5)*G4	=(F7+E7)*G4
9	174	=(\$M\$4*(F9-F8)+(1-\$M\$4)*E8	=(\$M\$3*D9/G5+(1-\$M\$3)*(F8+E8)	=R7*D9/F9+(1-R7)*G5	=(F8+E8)*G5
10	142	=(\$M\$4*(F10-F9)+(1-\$M\$4)*E9	=(\$M\$3*D10/G6+(1-\$M\$3)*(F9+E9)	=R8*D10/F10+(1-R8)*G6	=(F9+E9)*G6
11	141	=(\$M\$4*(F11-F10)+(1-\$M\$4)*E10	=(\$M\$3*D11/G7+(1-\$M\$3)*(F10+E10)	=R9*D11/F11+(1-R9)*G7	=(F10+E10)*G7
12	162	=(\$M\$4*(F12-F11)+(1-\$M\$4)*E11	=(\$M\$3*D12/G8+(1-\$M\$3)*(F11+E11)	=R10*D12/F12+(1-R10)*G8	=(F11+E11)*G8
13	180	=(\$M\$4*(F13-F12)+(1-\$M\$4)*E12	=(\$M\$3*D13/G9+(1-\$M\$3)*(F12+E12)	=R11*D13/F13+(1-R11)*G9	=(F12+E12)*G9
14	164	=(\$M\$4*(F14-F13)+(1-\$M\$4)*E13	=(\$M\$3*D14/G10+(1-\$M\$3)*(F13+E13)	=R12*D14/F14+(1-R12)*G10	=(F13+E13)*G10
15	171	=(\$M\$4*(F15-F14)+(1-\$M\$4)*E14	=(\$M\$3*D15/G11+(1-\$M\$3)*(F14+E14)	=G46*D15/F15+(1-G46)*G11	=(F14+E14)*G11
16	206	=(\$M\$4*(F16-F15)+(1-\$M\$4)*E15	=(\$M\$3*D16/G12+(1-\$M\$3)*(F15+E15)	=G47*D16/F16+(1-G47)*G12	=(F15+E15)*G12
17	133	=(\$M\$4*(F17-F16)+(1-\$M\$4)*E16	=(\$M\$3*D17/G13+(1-\$M\$3)*(F16+E16)	=G48*D17/F17+(1-G48)*G13	=(F16+E16)*G13
18	207	=(\$M\$4*(F18-F17)+(1-\$M\$4)*E17	=(\$M\$3*D18/G14+(1-\$M\$3)*(F17+E17)	=G49*D18/F18+(1-G49)*G14	=(F17+E17)*G14
19	218	=(\$M\$4*(F19-F18)+(1-\$M\$4)*E18	=(\$M\$3*D19/G15+(1-\$M\$3)*(F18+E18)	=G50*D19/F19+(1-G50)*G15	=(F18+E18)*G15
20	229	=(\$M\$4*(F20-F19)+(1-\$M\$4)*E19	=(\$M\$3*D20/G16+(1-\$M\$3)*(F19+E19)	=G51*D20/F20+(1-G51)*G16	=(F19+E19)*G16
21	225	=(\$M\$4*(F21-F20)+(1-\$M\$4)*E20	=(\$M\$3*D21/G17+(1-\$M\$3)*(F20+E20)	=G52*D21/F21+(1-G52)*G17	=(F20+E20)*G17
22	204	=(\$M\$4*(F22-F21)+(1-\$M\$4)*E21	=(\$M\$3*D22/G18+(1-\$M\$3)*(F21+E21)	=G53*D22/F22+(1-G53)*G18	=(F21+E21)*G18
23	227	=(\$M\$4*(F23-F22)+(1-\$M\$4)*E22	=(\$M\$3*D23/G19+(1-\$M\$3)*(F22+E22)	=G54*D23/F23+(1-G54)*G19	=(F22+E22)*G19
24	223	=(\$M\$4*(F24-F23)+(1-\$M\$4)*E23	=(\$M\$3*D24/G20+(1-\$M\$3)*(F23+E23)	=G55*D24/F24+(1-G55)*G20	=(F23+E23)*G20
25	242	=(\$M\$4*(F25-F24)+(1-\$M\$4)*E24	=(\$M\$3*D25/G21+(1-\$M\$3)*(F24+E24)	=G56*D25/F25+(1-G56)*G21	=(F24+E24)*G21
26	233	=(\$M\$4*(F26-F25)+(1-\$M\$4)*E25	=(\$M\$3*D26/G22+(1-\$M\$3)*(F25+E25)	=G57*D26/F26+(1-G57)*G22	=(F25+E25)*G22
27	266	=(\$M\$4*(F27-F26)+(1-\$M\$4)*E26	=(\$M\$3*D27/G23+(1-\$M\$3)*(F26+E26)	=G58*D27/F27+(1-G58)*G23	=(F26+E26)*G23
28					=(F27+E27)*G24
29					=(F27+E27)*G25
30					=(F27+E27)*G26
31					=(F27+E27)*G27
32					=(F27+E27)*G24
33					=(F27+E27)*G24

Figure: 66.101 - Formules correspondantes au modèle de lissage exponentiel de Holt et Winters avec Microsoft Excel 14.0.6123

Avec un peu plus loin dans la feuille les valeurs indispensables à l'optimisation du modèle avec le solveur que nous allons de suite voir:

	L	M
3	α :	0.327469638750293
4	β :	0.105758348520427
5	γ :	0
6		
7	MAPE:	=MOYENNE(I8:I27)
8		
9	MAD:	=MOYENNE(J8:J27)
10		
11	MSD:	=SOMME((J8:J27)*(J8:J27))/NB(J8:J27)

Figure: 66.102 - Constante de lissage et indicateurs d'erreur pour optimiser le modèle avec Microsoft Excel 14.0.6123

La procédure pour minimiser soit l'erreur MAPE, soit l'erreur MAD ou enfin l'erreur MSD avec le solveur est exactement la même que pour le lissage exponentiel simple! Nous ne la détaillerons donc pas.

Ce qui donne graphiquement (et aussi en comparaison avec le lissage simple, double et le modèle additif de Holt):

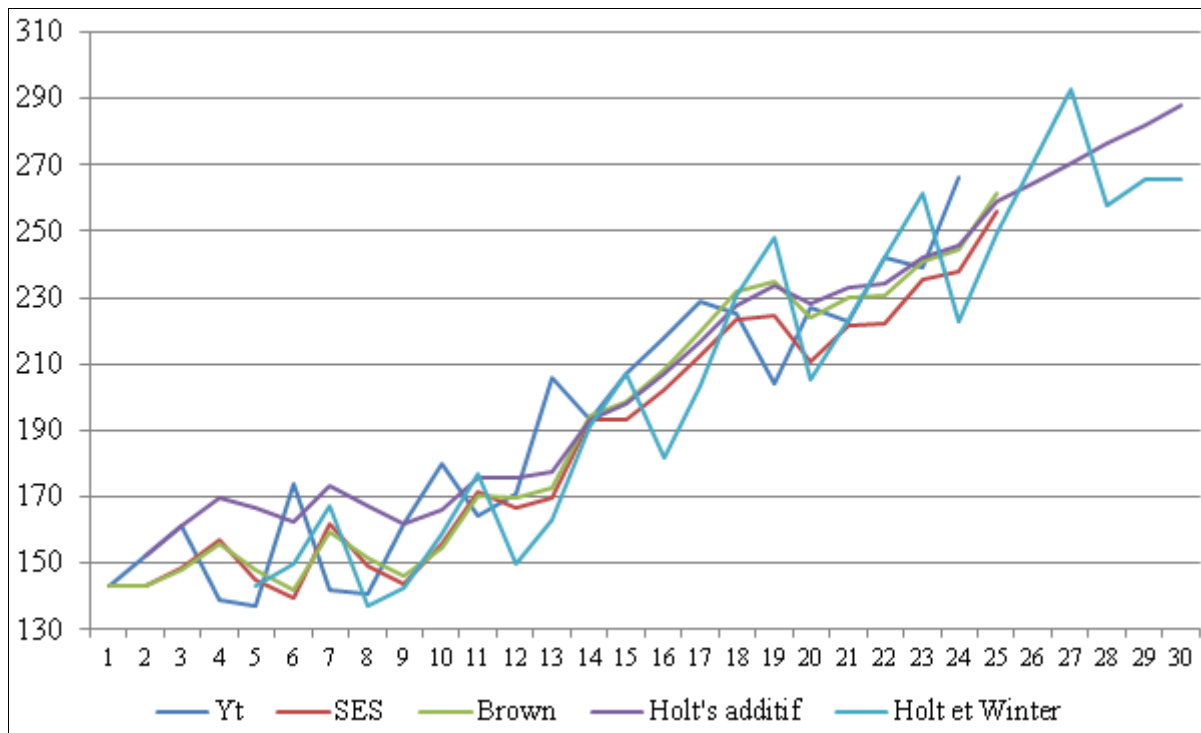


Figure: 66.103 - Projection sur un horizon de un semestre en utilisant le modèle de Holt et Winters avec Microsoft Excel 14.0.6123

Le lissage de Holt avait donc un MAPE de 7.38% et celui de Brown à un MAPE de 7.12%. Le lissage saisonnier de Holt et Winters a un MAPE de 9.61%. Cependant l'avantage du modèle de Holt et Winters est qu'on peut faire des projections à plus d'une période avec cycles ainsi que nous pouvons le voir sur le graphique.

Il est bien évidemment intéressant de comparer les projections à 6 mois entre la méthode additive de Holt et le lissage de Brown:

Brown	Holt's additif	Holt's-Winter
288.34	258.77	249.34
261.50	264.61	270.73
268.19	270.46	292.81
264.55	276.31	258.01
280.53	282.15	265.44
283.20	288.00	265.44

Figure: 66.104 - Comparaison entre modèle de Brown, et Holt additif et Holt et Winters

Donc nous voyons encore une fois qu'en fonction de la technique choisie, les différences sont relativement importantes.

8.3.7. RÉGRESSION LOGISTIQUE

Il arrive toujours dans les entreprises que dans l'analyse d'un produit ou d'un service, que celui-ci voie son nombre de ventes croître, ensuite passer par un point d'inflexion et ensuite aller vers une asymptote pour diminuer à nouveau par la suite avec une caractéristique similaire.

Le modèle logistique permettant de simuler un tel comportement dans le cadre de l'analyse des séries temporelles (à ne pas confondre avec la régression logistique vue dans le chapitre de Méthodes Numériques!) est défini ainsi:

$$y_t = \frac{y_{\max}}{1 + br^t} \quad (66.786)$$

Il s'inspire de nombreux modèles que nous retrouvons en physique et où $y_{\max}$ est le seuil de saturation (asymptote horizontale) qui peut être déterminé suite à un audit du marché et son % de pénétration.

Remarque: Il faut aussi savoir que ce modèle est bien meilleur que celui utilisé par le lissage exponentiel compris dans l'Utilitaire d'Analyse de Microsoft Excel 14.0.6123 (même si comme nous l'avons vu mathématiquement, ils ne fonctionnent absolument pas sur les mêmes bases). Une simple observation comparative des résultats obtenus suffit à s'en rendre compte.

b et r sont eux deux paramètres du modèle tels que:

$$\begin{aligned} b &> 0 \\ 0 < r < 1 \end{aligned} \quad (66.787)$$

le point d'inflexion est toujours donné par le cumul de 50% du seuil de saturation. Le résultat est alors une courbe en S du type suivant:

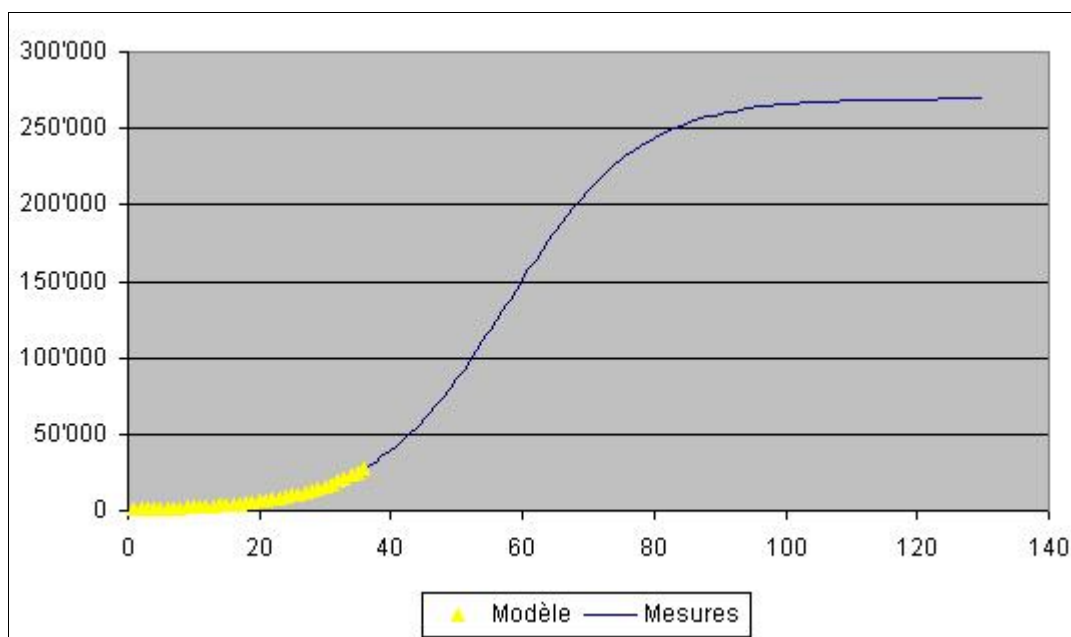


Figure: 66.105 - Comparaison mesures avec modèle logistique

où en jaune a été représenté les données actuelles d'une entreprise et en bleu le modèle théorique prévisionnel associé.

Pour déterminer l'équation de la courbe logistique, nous pouvons utiliser directement les solveurs de certains logiciels. Mais ceux-ci ont parfois besoin d'avoir des données de départ proches de la valeur théorique. Nous allons donc d'abord montrer comment ces valeurs de départ peuvent être déterminées avec un exemple.

Considérons le tableau suivant fait avec Microsoft Excel 11.8346 (les ventes sont en centaines de milliers d'unités):

	A	B
1	Semaines	Ventes
2	1	44.7
3	2	61
4	3	81.3
5	4	105.8
6	5	134
7	6	165.6
8	7	200.1
9	8	236.9
10	9	275.4
11	10	315
12	11	355.2
13	12	395.3
14	13	434.8
15	14	473.3
16	15	510.1
17	16	544.9
18	17	577.3
19	18	607
20	19	633.9

Figure: 66.106 - Ventes en fonction des semaines

et le graphique associé:

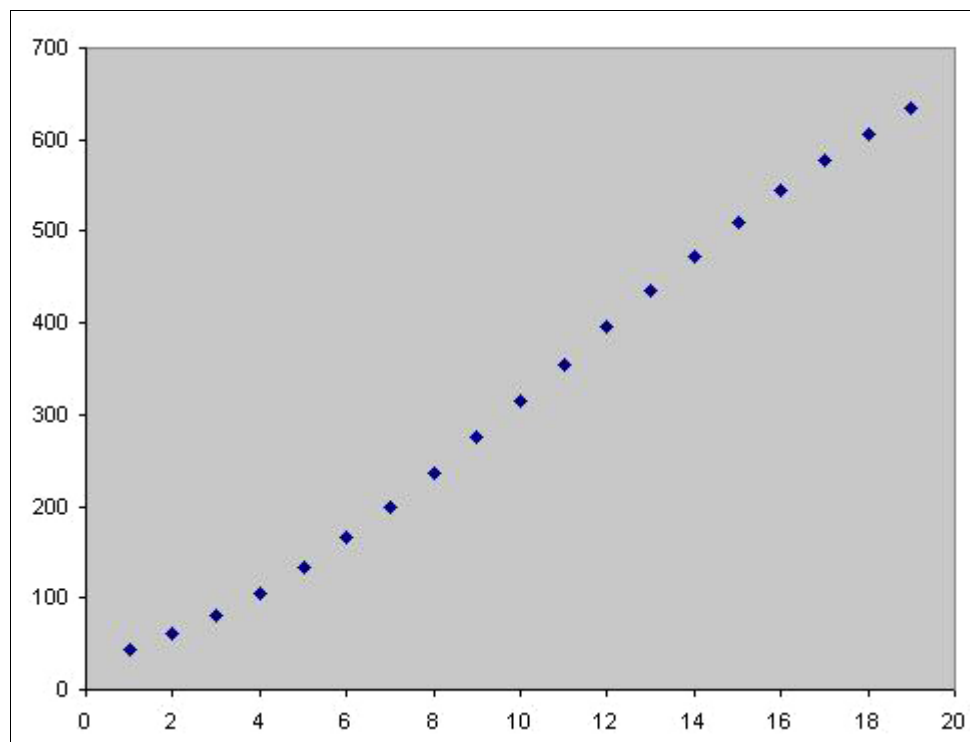


Figure: 66.107 - Tracé des mesures précédentes

qui pourrait être jugé comme linéaire suivant le moment auquel débute l'analyse descriptive des ventes dans l'entreprise.

Pour déterminer le modèle théorique, nous allons linéariser l'équation logistique en utilisant un seuil hypothétique (objectifs de ventes du marché) de 800.

Donc:

$$y_t = \frac{y_{\max}}{1 + br^t} \Rightarrow 1 + br^t = \frac{y_{\max}}{y_t} \Rightarrow br^t = \frac{y_{\max}}{y_t} - 1 \Rightarrow \ln\left(\frac{y_{\max}}{y_t} - 1\right) = \ln(b) + t \ln(r) \quad (66.788)$$

Soit à calculer la nouvelle variable à expliquer:

$$y'_t = \ln\left(\frac{y_{\max}}{y_t} - 1\right) = \ln\left(\frac{800}{y_t} - 1\right) \quad (66.789)$$

et nous écrirons le modèle linéaire correspondant:

$$Y_t = a_1 t + a_0 + \varepsilon_t \quad (66.790)$$

avec donc:

$$\begin{aligned} a_1 &= \ln(r) \Rightarrow r = e^{a_1} \\ a_0 &= \ln(b) \Rightarrow b = e^{a_0} \end{aligned} \quad (66.791)$$

Soit:

$$y_t = \frac{y_{\max}}{1 + e^{a_0} e^{a_1 t}} = \frac{y_{\max}}{1 + e^{a_1 t + a_0}} \quad (66.792)$$

Dans notre exemple, la régression linéaire (cf. chapitre de Méthodes Numériques) donne:

$$Y_t = -0.224t + 2.7908 \quad (66.793)$$

soit:

Nous avons alors immédiatement:

$$y_t \cong \frac{800}{1 + 16.294 \cdot 0.7988^t} \quad (66.794)$$

Soit sous forme graphique:

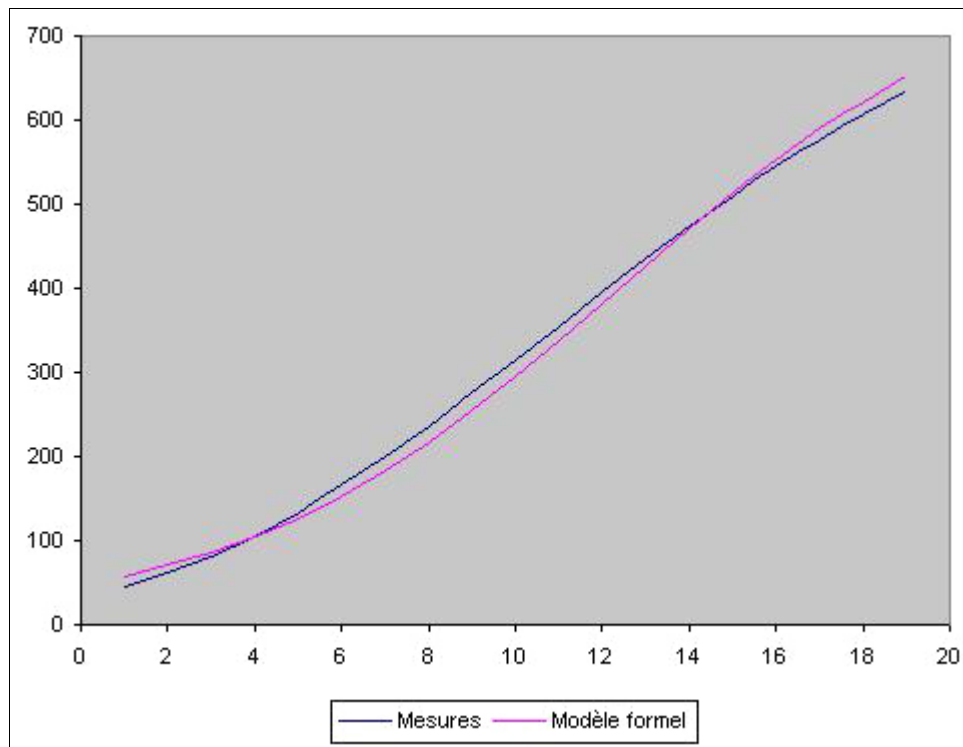


Figure: 66.108 - Comparaison mesures avec modèle logistique

avec ce modèle formel, nous avons une somme des carrés des écarts entre les mesures et le modèle (cf. chapitre de Statistiques) de:

$$SSE \cong 3520.828 \quad (66.795)$$

Maintenant, entrons ces données dans Microsoft Excel 14.0.6123 sous la forme suivante:

	A	B	C	D	E	F
1	Semaines	Ventes	Modèle	Différence^2	$y_{\max}$:	800.00
2	1	44.70	57.08	153.31	r :	0.7988
3	2	61.00	70.20	84.67	b :	16.2940
4	3	81.30	85.99	21.97	SSE :	3'520.829
5	4	105.80	104.81	0.97		
6	5	134.00	127.03	48.61		
7	6	165.60	152.91	160.92		
8	7	200.10	182.64	304.68		
9	8	236.90	216.22	427.48		
10	9	275.40	253.44	482.11		
11	10	315.00	293.84	447.61		
12	11	355.20	336.72	341.66		
13	12	395.30	381.13	200.68		
14	13	434.80	426.02	77.03		
15	14	473.30	470.26	9.22		
16	15	510.10	512.80	7.28		
17	16	544.90	552.73	61.34		
18	17	577.30	589.39	146.24		
19	18	607.00	622.37	236.10		
20	19	633.90	651.48	308.94		
21				3520.82858		

Figure: 66.109 - Comparaison mesures et valeurs du modèle théorique

avec la structure suivante:

	A	B	C	D	E	F
1	Semaines	Ventes	Modèle	Différence^2	$y_{\max}$:	800
2	1	44.7	$=\$F\$1/(1+\$F\$3*\$F\$2^A2)$	$=(B2-C2)^2$	r :	0.79875
3	2	61	$=\$F\$1/(1+\$F\$3*\$F\$2^A3)$	$=(B3-C3)^2$	b :	16.2940
4	3	81.3	$=\$F\$1/(1+\$F\$3*\$F\$2^A4)$	$=(B4-C4)^2$	SSE :	$=D21$
5	4	105.8	$=\$F\$1/(1+\$F\$3*\$F\$2^A5)$	$=(B5-C5)^2$		
6	5	134	$=\$F\$1/(1+\$F\$3*\$F\$2^A6)$	$=(B6-C6)^2$		
7	6	165.6	$=\$F\$1/(1+\$F\$3*\$F\$2^A7)$	$=(B7-C7)^2$		
8	7	200.1	$=\$F\$1/(1+\$F\$3*\$F\$2^A8)$	$=(B8-C8)^2$		
9	8	236.9	$=\$F\$1/(1+\$F\$3*\$F\$2^A9)$	$=(B9-C9)^2$		
10	9	275.4	$=\$F\$1/(1+\$F\$3*\$F\$2^A10)$	$=(B10-C10)^2$		
11	10	315	$=\$F\$1/(1+\$F\$3*\$F\$2^A11)$	$=(B11-C11)^2$		
12	11	355.2	$=\$F\$1/(1+\$F\$3*\$F\$2^A12)$	$=(B12-C12)^2$		
13	12	395.3	$=\$F\$1/(1+\$F\$3*\$F\$2^A13)$	$=(B13-C13)^2$		
14	13	434.8	$=\$F\$1/(1+\$F\$3*\$F\$2^A14)$	$=(B14-C14)^2$		
15	14	473.3	$=\$F\$1/(1+\$F\$3*\$F\$2^A15)$	$=(B15-C15)^2$		
16	15	510.1	$=\$F\$1/(1+\$F\$3*\$F\$2^A16)$	$=(B16-C16)^2$		
17	16	544.9	$=\$F\$1/(1+\$F\$3*\$F\$2^A17)$	$=(B17-C17)^2$		
18	17	577.3	$=\$F\$1/(1+\$F\$3*\$F\$2^A18)$	$=(B18-C18)^2$		
19	18	607	$=\$F\$1/(1+\$F\$3*\$F\$2^A19)$	$=(B19-C19)^2$		
20	19	633.9	$=\$F\$1/(1+\$F\$3*\$F\$2^A20)$	$=(B20-C20)^2$		
21				$=SOMME(D2:D20)$		

Figure: 66.110 - Formules explicites pour le calcul du modèle théorique

Si nous lançons le solveur de Microsoft Excel 14.0.6123 avec les paramètres suivants (nous avons mis 0.0001 comme plus petite valeur puisque le solveur ne propose pas de relation d'ordre stricte):

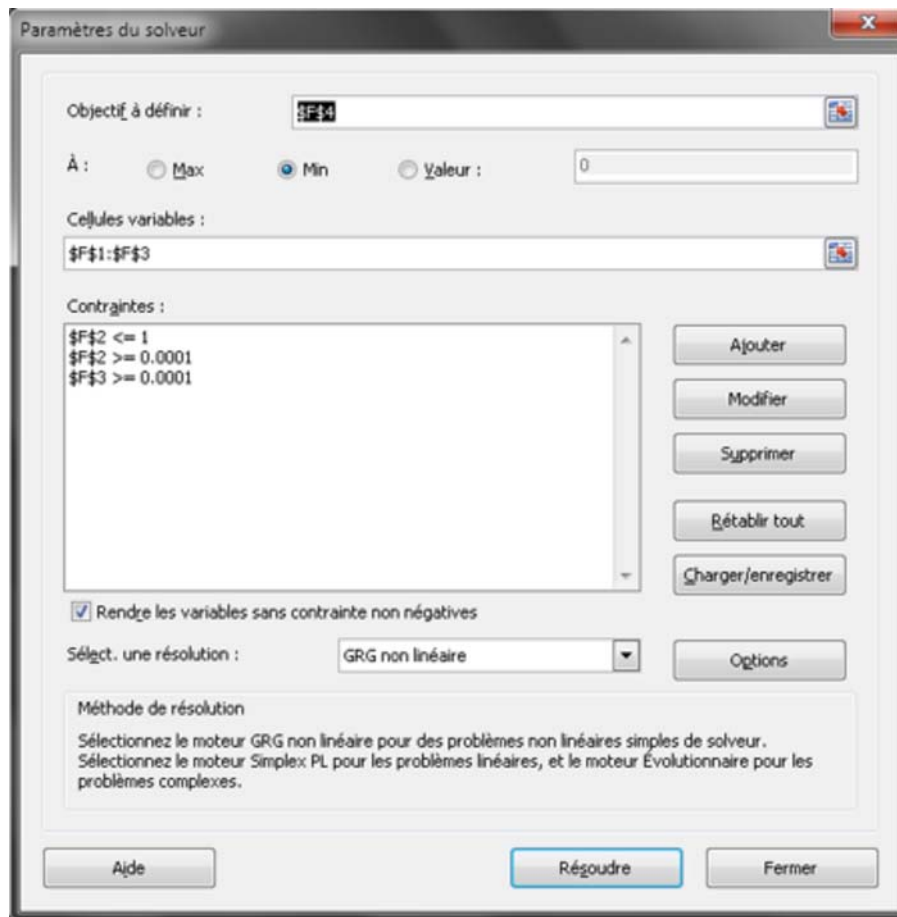


Figure: 66.111 - Paramètres du solveur pour minimiser SSE

Ce qui donne:

E	F
$y_{\max}$:	717.94
r :	0.7849
b :	14.7744
SSE :	698.511

Figure: 66.112 - Résultat du solveur

Soit:

$$y_t \cong \frac{717.94}{1 + 14.7744 \cdot 0.7849^t} \quad (66.796)$$

avec:

$$SSE = 698.511 \quad (66.797)$$

soit nettement inférieur à notre approche utilisant la régression linéaire et donc meilleur. Effectivement voyons le tableau de résultat:

	A	B	C
1	Semaines	Ventes	Modèle
2	1	44.70	57.00
3	2	61.00	71.07
4	3	81.30	88.16
5	4	105.80	108.66
6	5	134.00	132.93
7	6	165.60	161.18
8	7	200.10	193.46
9	8	236.90	229.53
10	9	275.40	268.88
11	10	315.00	310.68
12	11	355.20	353.86
13	12	395.30	397.19
14	13	434.80	439.42
15	14	473.30	479.43
16	15	510.10	516.33
17	16	544.90	549.53
18	17	577.30	578.73
19	18	607.00	603.92
20	19	633.90	625.28

Figure: 66.113 - Valeurs données par le modèle obtenu par optimisation

et graphiquement cela donne:

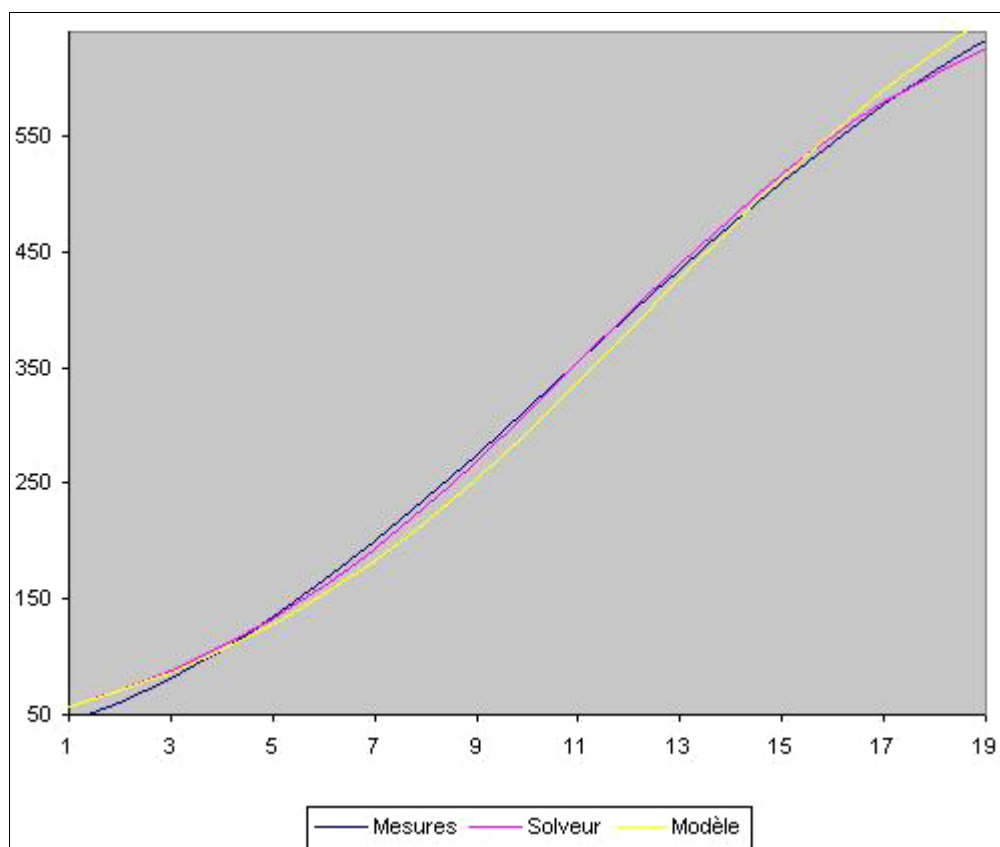


Figure: 66.114 - Comparaison entre mesures, modèle par régression et modèle par optimisation

Nous voyons nettement que le modèle du solveur (modèle numérique) est meilleur que le modèle formel donné par une régression linéaire et il est aussi meilleur comme déjà mentionné que le lissage exponentiel proposé par l'Utilitaire d'analyse de Microsoft Excel 11.8346!

Pour en finir avec ces modèles de prévisions, rappelons que ceux-ci doivent être suivis régulièrement pour s'assurer que le modèle et les paramètres utilisés sont toujours appropriés. Comme le lecteur l'aura peut-être deviné, un modèle de prévision ne doit pas avoir de biais. La somme des erreurs doit donc voisiner zéro, avec parfois des surestimations et parfois des sous-estimations. Lorsque le modèle a tendance à toujours surestimer (ou sous-estimer), ce dernier a un biais et il doit être révisé.

Afin de détecter que le modèle est biaisé, on fera typiquement le suivi de la différence entre la prévision et l'observation sur une carte de contrôle (cf. chapitre de Génie Industriel).

Il faut aussi pour conclure rappeler qu'il existe une différence parfois importante entre les valeurs observées (ventes effectives) et les observations qu'il y aurait eu réellement si l'offre était infinie. Effectivement, les ventes observées ne prennent pas en compte les ventes ratées parce que faute de stock dans une des milliers de succursales que possède votre entreprise, un client est allé voir chez la concurrence. Ainsi, les ventes observées sont toujours un "minimum" faute des ventes loupées!

Remarque: Indiquons qu'il est aussi trivialement possible de faire la différence de tous les points consécutifs dans le temps d'une série temporelle et ensuite de faire un histogramme pour déterminer la loi de probabilité des fluctuations, ce qui permet de faire de l'inférence statistique avec toutes les précautions nécessaires.

8.4. MODÈLES AUTORÉGRESSIFS

Dans l'étude d'une série chronologique, il est naturel de penser que la valeur de la série à la date t peut dépendre en partie des valeurs prises aux dates précédentes:

$$X_t = f(X_{t-1}, X_{t-2}, \dots) \quad (66.798)$$

ce qui évidemment est une écriture limitée à une série univariée (nous n'étudierons par les séries multivariées), discrète, et dont la fréquence d'échantillonnage est constante (au contraire que ce qui se fait dans certains domaines des transactions financières).

Nous avons vu dans le chapitre de Probabilité que lorsque nous nous tenons uniquement à une dépendance avec l'instant $t-1$, nous parlons alors de "chaîne de Markov", ou plus rigoureusement de "**chaîne de Markov d'ordre 1**". Avant de continuer, ouvrons un paranthèse sur l'intérêt des chaînes de Markov (rappel orienté finance)... Considérons que nous avons des valeurs boursières à une fréquence fixe (par exemple journalière ou horaire... peu importe!). Il est alors très aisé de calcul la proportion de Augmentations (A) versus les Diminutions (D) des valeurs d'intérêt. Ainsi, en prenant en compte l'ensemble de l'historique, nous aurons typiquement un tableau du genre (les valeurs proviennent des indices du S&P 500 entre 1947 et 2007):

Proportion Augmentations (A)	Proportions Diminutions (D)	Total
0.474	0.526	1

Tableau: 66.21 - Proportion de variations A/D sur l'ensemble d'un historique

Mais bon cela ne sera pas vraiment à grand chose... Si nous compliquons un peu l'analyse en nous demandant quelle proportion nous avons deux augmentations successives (AA), diminutions successives (DD), ou alternance (DA) et (AD) dans l'ensemble de l'historique cela double le travail et pourrait nous donner typiquement un tableau du genre:

$X_i =$	(A)	(D)	Total
$X_{i-1} = (D)$	0.519	0.481	1
(A)	0.433	0.567	1

Tableau: 66.22 - Proportion de variations A/D sur l'ensemble par rapport à la période précédente

Les valeurs ci-dessus peuvent donc bien évidemment être vues comme les probabilités conditionnelles:

$$P(X_i | X_{i-1}) \quad (66.799)$$

Et si maintenant nous faisons une analyse non pas que sur la période précédente mais sur les deux dernières, cela signifie que nous calculons les proportions des séquences (DDD), (DDA), (DAD), (DAA), (ADD), (ADA), (AAD), (AAA) et cela double encore une fois le nombre de calculs:

		$X_i =$		
X_{i-2}	X_{i-1}	(A)	(D)	Total
(D)	(D)	0.501	0.499	1
(D)	(A)	0.412	0.588	1
(A)	(D)	0.539	0.461	1
(A)	(A)	0.449	0.551	1

Tableau: 66.23 - Proportion de variations A/D sur l'ensemble par rapport à 2 périodes précédentes

Les valeurs ci-dessus peuvent donc bien évidemment être vues comme les probabilités conditionnelles:

$$P(X_i | X_{i-1}, X_{i-2}) \quad (66.800)$$

Nous voyons donc qu'en ajoutant à chaque fois une période précédente supplémentaire à l'analyse, nous doublons le nombre de calculs. Ainsi, en prenant les 20 dernières périodes, cela nous amène à près de 1 millions de probabilités conditionnelles (proportions). Nous comprenons alors tout l'intérêt de nous restreindre à un faible nombre de valeurs antérieures ou à une seule valeur comme le font les chaînes de Markov. Ainsi, la probabilité que la séquence des 4 prochains jours soit AADA sera sans hypothèse simplificatrice:

$$P(AADA) = P(X_1 = A) \cdot P(X_2 = A | X_1 = A) \cdot P(X_3 = D | X_2 = A, X_1 = A) \cdot P(X_4 = A | X_3 = D, X_2 = A, X_1 = A) \quad (66.801)$$

mais se réduira dans l'hypothèse d'une chaîne de Markov (c'est-à-dire une dépendance seulement avec la période précédente) à:

$$\begin{aligned} P(AADA) \\ = P(X_1 = A) \cdot P(X_2 = A | X_1 = A) \cdot P(X_3 = D | X_2 = A) \cdot P(X_4 = A | X_3 = D) \quad (66.802) \\ = 0.526 \cdot 0.567 \cdot 0.433 \cdot 0.481 \cong 0.062 = 6.2\% \end{aligned}$$

Savoir si le modèle de probabilités conditionnelles totales ou la version simplifié de Markov est un vaste débat. L'hypothèse de la marche aléatoire (implicitement: le modèle de Markov) est défendue par une majorité d'économistes... mais cette hypothèse a bien évidemment aussi un certain nombre de contradicteurs.

Il n'est généralement pas nécessaire de prendre en compte tout le passé de la série ou seulement le dernier événement et nous pouvons le plus souvent nous limiter à p valeurs:

$$X_t = f(X_{t-1}, X_{t-2}, \dots, X_{t-p}) + \varepsilon_t \quad (66.803)$$

où ε_t est un bruit blanc $N(0, \sigma^2)$ ou processus de Wiener noté parfois WN. Plus rigoureusement, et pour des raisons d'application de techniques statistiques, un bruit blanc est défini par:

$$\begin{aligned} E(\varepsilon_t) &= 0 \\ E(\varepsilon_t^2) &= \sigma^2 \\ E(\varepsilon_t \varepsilon_\tau) &= 0 \quad \forall t \neq \tau \end{aligned} \quad (66.804)$$

La seule différence par rapport au mouvement brownien standard c'est que comme nous le verrons plus loin, il y a ici la présence d'un facteur d'inertie noté traditionnellement α dans les cas simples qui influence fortement la dynamique du processus tel que:

$$X_t = \alpha X_{t-1} + \varepsilon_t \quad (66.805)$$

Effectivement, comme il est très facile de le faire dans Microsoft Excel conformément à la procédure indiquée lors de notre étude des processus de Wiener:

Voici différents tracés de la série temporelle en fonction de quelques valeurs du facteur d'inertie:

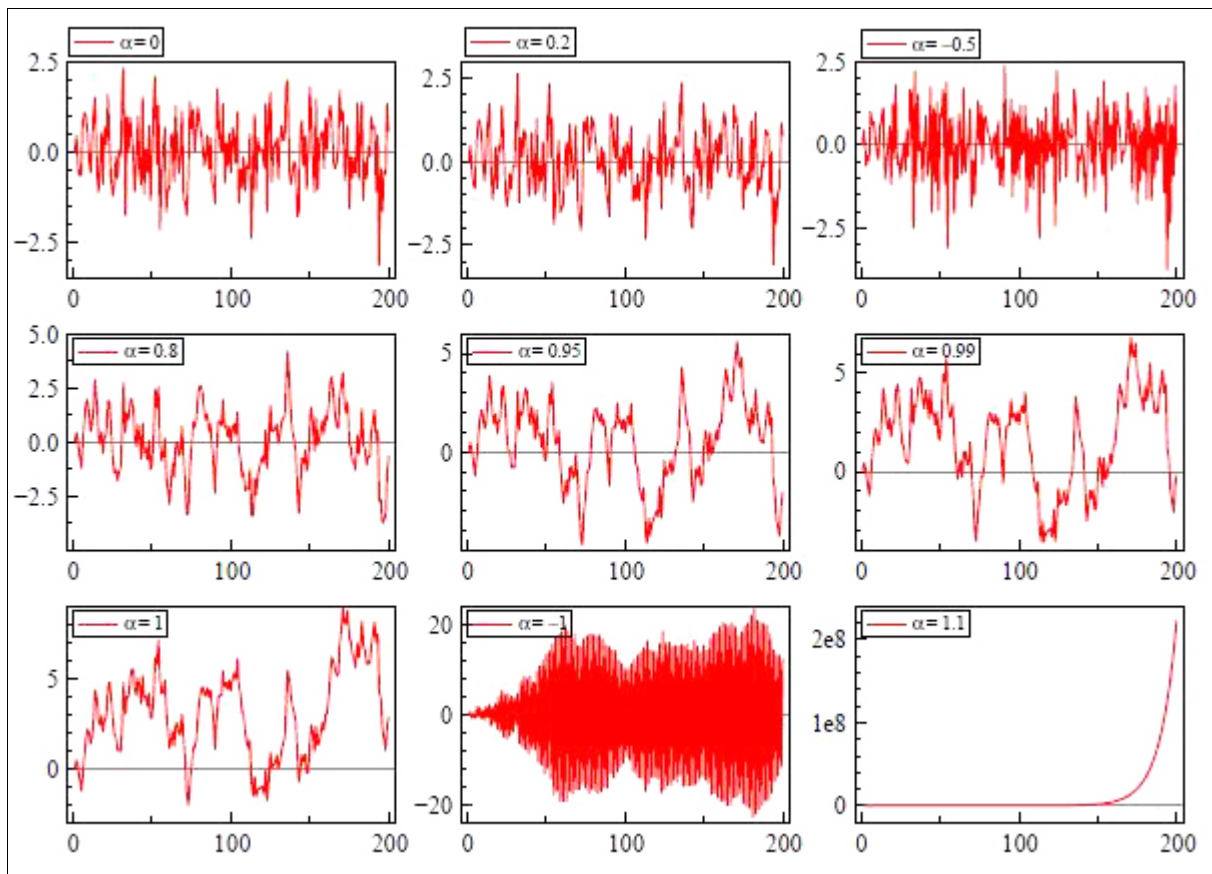


Figure: 66.115 - Processus autorégressif d'ordre 1

Si nous considérons X_t comme une variable aléatoire spécifique (ayant une fonction de densité donnée) que nous noterions X et X_{t-h} comme une autre variable aléatoire spécifique (ayant une fonction de densité donnée) que nous noterions Y , alors rien ne nous empêche étant connues les fonctions de densité de chacune de ces variables, de calculer leur covariance:

$$\gamma_t(h) = c_{X,Y} = \text{cov}(X,Y) = \text{cov}(X_t, X_{t-h}) = E(X_t \cdot X_{t-h}) - E(X_t)E(X_{t-h}) \quad (66.806)$$

Par exemple, dans la pratique nous connaissons souvent les espérances μ des deux variables aléatoires aux deux moments différents ainsi que quelques-unes des valeurs de leurs distributions sous-jacentes (réalisations aléatoires). Alors, il devient aisé de calculer leur covariance. Mais ce n'est pas un indicateur vraiment utile. Rien ne nous empêche en supposant une relation linéaire d'utiliser le coefficient de corrélation linéaire:

$$R_{X,Y} = \frac{c_{X,Y}}{\sqrt{V(X)V(Y)}} = \frac{c_{X,Y}}{\sqrt{\text{cov}(X,X)\text{cov}(Y,Y)}} = \frac{c_{X,Y}}{\sqrt{c_{X,X}c_{Y,Y}}} \quad (66.807)$$

Mais qui se note alors traditionnellement et trivialement dans le cas des séries temporelles:

$$\rho_t(h) = \frac{\gamma_t(h)}{\sqrt{\gamma_t(0)\gamma_{t-h}(0)}} \quad (66.808)$$

et est appelé "**coefficient d'autocorrélation**" souvent abrégé dans les logiciels de statistiques ACF.

Personnellement je préfère réécrire la relation précédente sous la forme explicite suivante (la covariance est explicitée):

$$\rho_{X,Y,t} = \frac{E[(X_t - \mu_{X,t})(Y_t - \mu_{Y,t})]}{\sigma_X \sigma_Y} \quad (66.809)$$

Il faut savoir que lorsque nous faisons de l'analyse de séries temporelles, dans la pratique, nous comparons souvent une suite avec elle-même mais comportant un décalage temporel h . L'autocorrélation d'ordre h est donc la corrélation entre la série et elle-même retardée de h relevés.

Ainsi, puisque les deux suites sont dépendantes et ont les mêmes caractéristiques statistiques, nous écrivons dans ce cas particulier dit de "**processus stationnaire du second ordre**" la relation suivante appelée "**coefficient d'autocorrélation empirique**":

$$\begin{aligned} \rho(h) &= \frac{E[(X_t - \mu_X)(X_{t-h} - \mu_X)]}{\sigma_X \sigma_X} = \frac{\frac{1}{n-h} \sum_{t=h+1}^n (X_t - \mu_X)(X_{t-h} - \mu_X)}{\frac{1}{n-1} \sum_{t=1}^n (X_t - \mu_X)^2} \\ &= \frac{\sum_{t=h+1}^n (X_t - \mu_X)(X_{t-h} - \mu_X)}{(n-h) \sigma_{X,t}^2} \end{aligned} \quad (66.810)$$

où évidemment il ne faut pas oublier que dans la pratique l'espérance et l'écart-type ne sont que des estimateurs bien que cela ne soit pas explicite dans la définition ci-dessus puisque le domaine de l'analyse temporelle n'indique que très rarement les petites chapeaux au-dessus des estimateurs (contrairement à ce que nous avons fait dans le chapitre de Statistiques).

Évidemment dans le cas de l'analyse de séries temporelles où nous soupçonnons une cyclicité, la valeur la plus grande de ρ pour un h donné indique la fréquence (respectivement la périodicité) de la cyclicité sous-jacente à la série ce qui peut aider par exemple pour le choix du type de moyenne mobile (voir plus bas).

Définition: Nous appelons "**corrélogramme**", le diagramme représentant les coefficients d'autocorrélation d'ordre 1, 2, 3, ..., h , ... de la série.

Remarque: Il existe différentes variantes de cette relation dans certains logiciels. Il peut arriver d'obtenir des petites différences numériques par rapport à l'utilisation de la relation ci-dessus.

Voici par exemple une famille de séries temporelles:

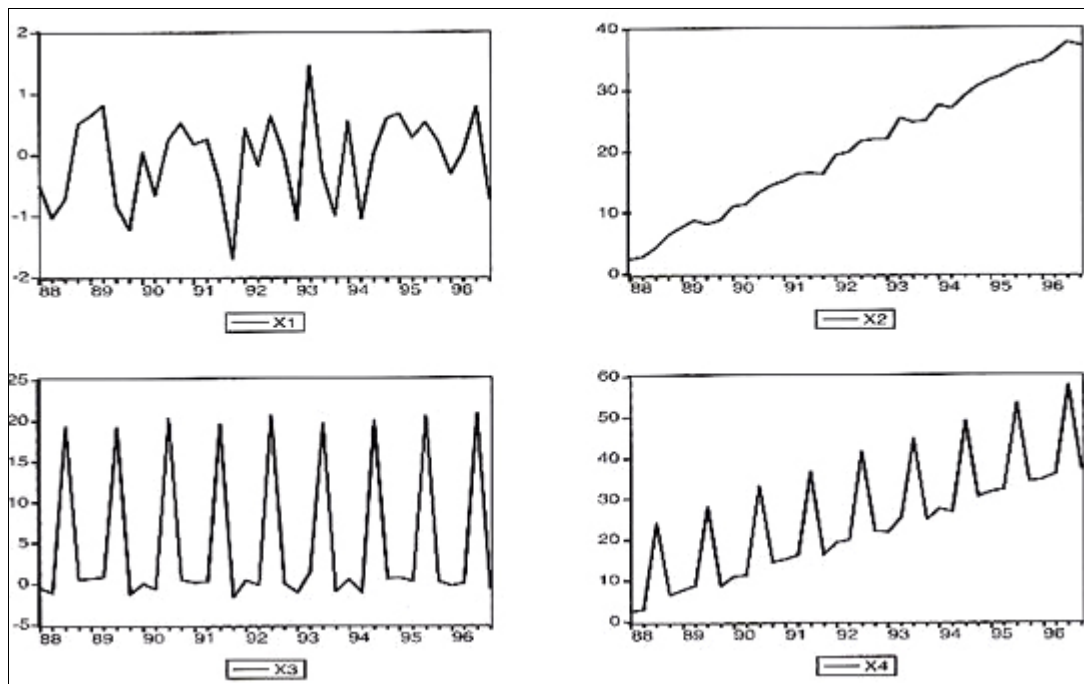


Figure: 66.116 - Quelques exemples de séries temporelles

avec leur corrélogramme correspondant avec bien évidemment pour différentes valeurs de h en abscisse et la valeur de ρ en ordonnée:

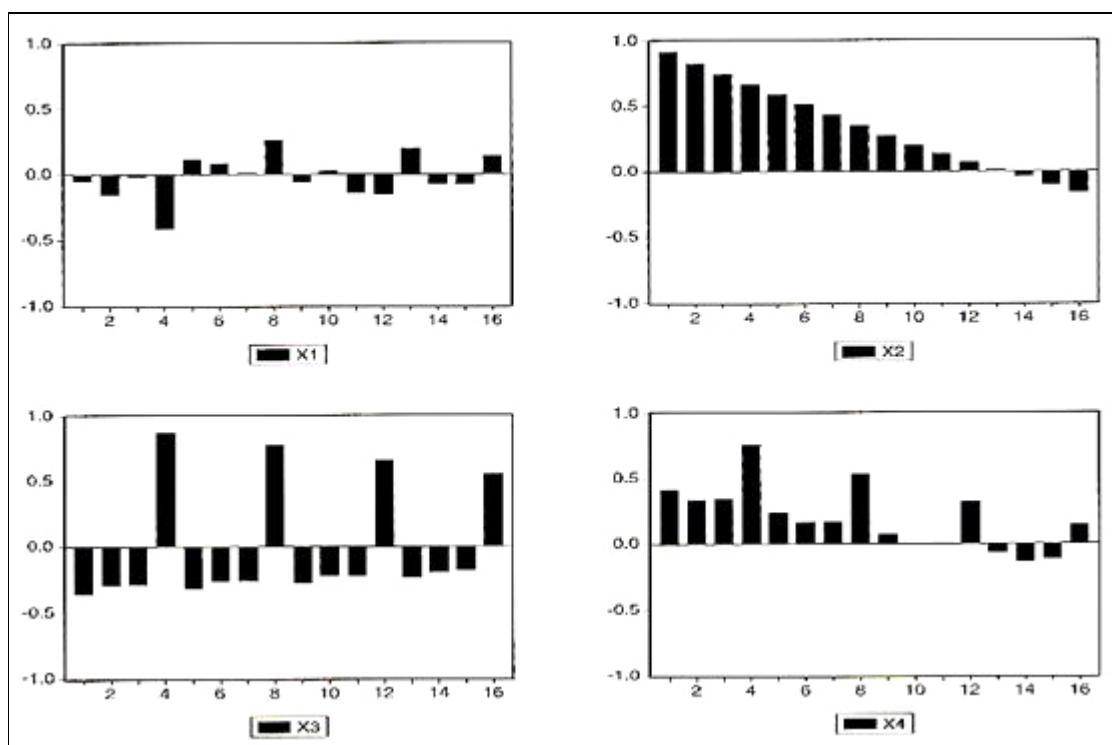


Figure: 66.117 - Corrélogrammes correspondants

Faisons un exemple pratique de calcul de corrélogramme avec Microsoft Excel 14.0.6123 et des données fictives:

	A	B	C	D	E
3	Ventes	Moyenne		ρ	h
4	144	159.15		0.269596705	1
5	151	159.15		0.751831577	2
6	134	159.15		0.159770276	3
7	151	159.15		0.668773644	4
8	145	159.15		-0.032360218	5
9	145	159.15			
10	141	159.15			
11	166	159.15			
12	151	159.15			
13	164	159.15			
14	151	159.15			
15	176	159.15			
16	170	159.15			
17	180	159.15			
18	156	159.15			
19	187	159.15			
20	166	159.15			
21	182	159.15			
22	154	159.15			
23	169	159.15			

Figure: 66.118 - Données avec moyenne et coefficients d'autocorrélation

Soit pour la partie qui nous intéresse:

	D	E
3	ρ	h
4	=SOMME(((A4:A22)-(B4:B22))*((A5:A23)-(B4:B22)))/ECARTYPE(\$A\$4:\$A\$23)^2/(NB(\$A\$4:\$A\$23)-E4)	1
5	=SOMME(((A4:A21)-(B4:B21))*((A6:A23)-(B6:B23)))/ECARTYPE(\$A\$4:\$A\$23)^2/(NB(\$A\$4:\$A\$23)-E5)	2
6	=SOMME(((A4:A20)-(B4:B20))*((A7:A23)-(B7:B23)))/ECARTYPE(\$A\$4:\$A\$23)^2/(NB(\$A\$4:\$A\$23)-E6)	3
7	=SOMME(((A4:A19)-(B4:B19))*((A8:A23)-(B8:B23)))/ECARTYPE(\$A\$4:\$A\$23)^2/(NB(\$A\$4:\$A\$23)-E7)	4
8	=SOMME(((A4:A18)-(B4:B18))*((A9:A23)-(B9:B23)))/ECARTYPE(\$A\$4:\$A\$23)^2/(NB(\$A\$4:\$A\$23)-E7)	5

Figure: 66.119 - Calculs des coefficients d'autocorrélation

et donc le corrélogramme correspondant:

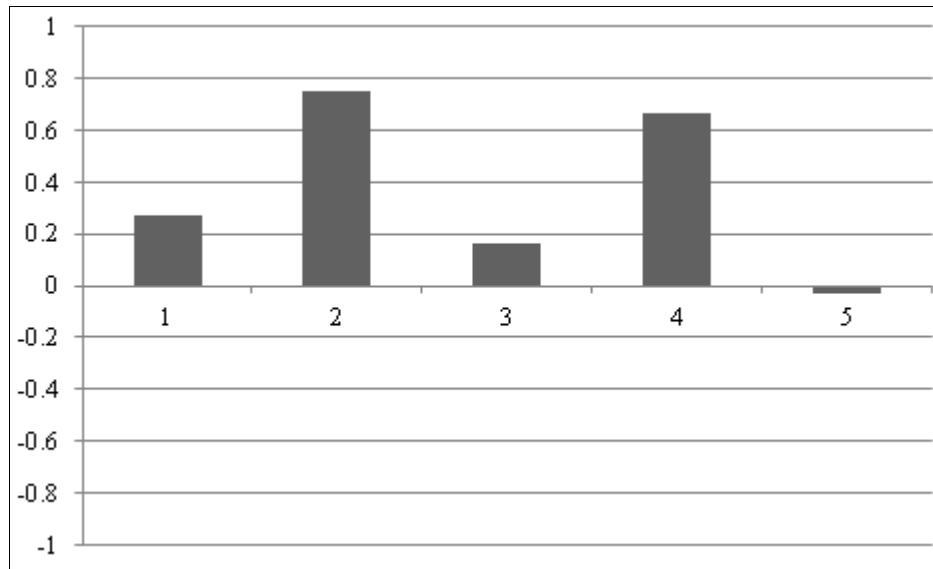


Figure: 66.120 - Corrélogramme correspondant à l'exemple

Définition: Une série temporelle est dite "**stationnaire au sens faible**" ou plus simplement un "**processus stationnaire faible**" si le premier (espérance) et second (variance) moment existent, sont finis et constants dans le temps (donc ne dépendant pas du temps):

$$\begin{aligned} E(X_i) &= \mu, \forall i \in 1 \dots T \\ V(X_i) &= \sigma^2, \forall i \in 1 \dots T \end{aligned} \quad (66.811)$$

La première condition (constance de l'espérance) élimine donc toute tendance. Si la fonction de densité sous-jacente à chaque y_i est une loi Normale, alors nous parlons de "**processus Gaussien**". Dans le cas contraire, nous dirons sur ce site qu'elle est non stationnaire.

Remarque: Certains auteurs rajoutent à la définition des processus stationnaires faibles que ceux-ci doivent avoir une covariance qui existe, soit finie et indépendante des instants choisis, ou plus exactement ne dépende que de l'écart entre les instants choisis, c'est-à-dire:

$$\text{cov}(X_{t+k}, X_t) = f(k)$$

Considérons un cas important dans de nombreuses entreprises sur certaines périodes plus ou moins longues. Soit:

$$\bar{X}_T = \frac{1}{T} \sum_{t=1}^T X_t \quad (66.812)$$

la moyenne temporelle. Si $\bar{X}_T$ converge en probabilité vers μ quand $T \rightarrow \infty$ nous disons que le processus est "**ergodique pour la moyenne**". Donc quand $T \rightarrow \infty$:

$$\bar{X}_T = \frac{1}{T} \sum_{t=1}^T X_t \rightarrow 0 \quad (66.813)$$

8.4.1. PROCESSUS AUTORÉGRESSIFS AR(P)

Présentons le concept de processus autorégressif en introduisant le cas particulier de processus autorégressif d'ordre 1, noté AR(1) et défini par la relation (à une constante additive près):

$$X_t = \varphi X_{t-1} + \varepsilon_t \quad (66.814)$$

Il s'agit donc d'une chaîne de Markov d'ordre 1.

Nous disons que le processus autorégressif AR(1) est stationnaire si nous avons à toute date t :

$$E(X_t) = \mu \quad V(X_t) = \sigma_X^2 \quad (66.815)$$

En toute généralité, nous avons:

$$\begin{aligned} \mu &= E(X_t) = \varphi E(X_{t-1}) + E(\varepsilon_t) = \varphi E(X_{t-1}) \\ \sigma_X^2 &= V(X_t) = \varphi^2 V(X_{t-1}) + V(\varepsilon_t) = \varphi^2 V(X_{t-1}) + \sigma^2 \end{aligned} \quad (66.816)$$

et donc la stationnarité implique obligatoirement que (se rappeler que la constante additive a été posée comme nulle dans la définition ci-dessus, dans le cas contraire cela changer les conclusions ci-dessous):

$$\begin{cases} \mu = \varphi\mu \\ \sigma_X^2 = \varphi^2 \sigma_X^2 + \sigma^2 \end{cases} \Rightarrow \begin{cases} \mu = 0 \text{ ou } \varphi = 1 \\ (1 - \varphi^2) \sigma_X^2 = \sigma^2 \end{cases} \Rightarrow \begin{cases} \mu = 0 \\ |\varphi| < 1 \end{cases} \quad (66.817)$$

Donc la stationnarité implique in extenso les deux résultats utiles:

$$\begin{aligned} E(X_t) &= \varphi E(X_{t-1}) = 1 \cdot E(X_{t-1}) = E(X_{t-1}) \\ \sigma_X^2 &= \frac{\sigma^2}{1 - \varphi^2} \end{aligned} \quad (66.818)$$

Il est intéressant de remarquer que la variance est indépendante du temps et donc qu'elle ne diverge pas. Cependant, il ne faut pas considérer cela comme forcément positif car il n'est pas très cohérent dans la pratique que la volatilité soit constante (processus homoscedastique). Raison pour laquelle nous aborderons des modèles plus complexes dont la volatilité est dépendante du temps.

Voyons ce que l'hypothèse de stationnarité implique et pour cela, nous allons d'abord avoir besoin de l'égalité suivante:

$$\begin{aligned} X_t &= \varphi X_{t-1} + \varepsilon_t = \varphi(\varphi X_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \varphi^2 X_{t-2} + \varphi \varepsilon_{t-1} + \varepsilon_t \\ &= \varphi^h X_{t-h} + \sum_{k=0}^{h-1} \varphi^k \varepsilon_{t-k} \end{aligned} \quad (66.819)$$

Donc in extenso:

$$\begin{aligned}
 \gamma_t(h) &= \text{cov}(X_t, X_{t-h}) = E(X_t \cdot X_{t-h}) - E(X_t)E(X_{t-h}) = E(X_t \cdot X_{t-h}) \\
 &= E\left(\left(\varphi^h X_{t-h} + \sum_{k=0}^{h-1} \varphi^k \varepsilon_{t-k}\right) \cdot X_{t-h}\right) = E(\varphi^h X_{t-h}^2) = \varphi^h E(X_{t-h}^2) \quad (66.820)
 \end{aligned}$$

Et rappelons la relation de Huyghens (cf. chapitre de Statistiques):

$$V(X) = E(X^2) - E(X)^2 \quad (66.821)$$

Il vient alors:

$$\gamma_t(h) = \varphi^h E(X_{t-h}^2) = \varphi^h \left[V(X_{t-h}) + E(X_{t-h})^2 \right] = \varphi^h V(X_{t-h}) = \varphi^h \sigma_X^2 \quad (66.822)$$

Donc l'autocorrélation vaut:

$$\rho_t(h) = \frac{\gamma_t(h)}{\sqrt{\gamma_t(0)\gamma_{t-h}(0)}} = \frac{\varphi^h \sigma_X^2}{\sqrt{\varphi^0 \sigma_X^2 \varphi^0 \sigma_X^2}} = \varphi^h \quad (66.823)$$

Puisque la stationnarité implique que $\varphi \ll 1$, les corrélogrammes (ACF) des processus AR(p) auront des allures décroissantes amorties: le processus oublie en quelque sorte les valeurs passées.

Nous avons aussi pour un processus AR(1) stationnaire avec constante (offset):

$$X_t = \varphi_0 + \varphi_1 X_{t-1} + \varepsilon_t \quad (66.824)$$

l'espérance suivante:

$$\begin{aligned}
 E(X_t) &= E(c^{te} + \varphi_1 X_{t-1} + \varepsilon_t) = c^{te} + E(\varphi_1 X_{t-1}) + E(\varepsilon_t) = c^{te} + \varphi_1 E(X_{t-1}) + 0 \\
 &= c^{te} + \varphi_1 E(X_{t-1}) + 0 = c^{te} + \varphi_1 E(X_t) \quad (66.825)
 \end{aligned}$$

$$\Rightarrow E(X_t) = \frac{\varphi_0}{1 - \varphi_1}$$

Donc si la constante est nulle nous retrouvons bien une espérance nulle comme démontré plus haut. La variance aura la même expression que celle déjà démontrée plus haut, c'est-à-dire:

$$\sigma_X^2 = \frac{\sigma^2}{1 - \varphi^2} \quad (66.826)$$

puisque l'ajout d'une constante ne change rien à la variance de la série. Nous avons donc obtenu ci-dessus l'espérance et la variance (inconditionnelles) d'un processus AR(1).

Un processus AR(2) est un processus autorégressif qui vérifie une relation de la forme (à une constante additive près):

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \varepsilon_t \quad (66.827)$$

Donc dans un tel modèle, l'influence du passé se manifeste par une régression linéaire sur les deux valeurs antérieures. Selon les valeurs de φ_1 et φ_2 il n'est pas toujours possible de trouver un processus stationnaire vérifiant cette dernière relation.

En général, un processus AR(p) est un processus qui dépend linéairement des p valeurs antérieures (à une constante additive près):

$$X_t = \varphi_1 X_{t-1} + \dots + \varphi_p X_{t-p} + \varepsilon_t = \sum_{k=1}^p \varphi_k X_{t-k} + \varepsilon_t \quad (66.828)$$

8.4.2. PROCESSUS AUTORÉGRESSIFS MA(Q)

D'abord une mise en garde. Il ne faut pas confondre ce processus stationnaire avec les moyennes mobiles. En effet, le terme n'est pas très bien choisi car s'il y a bien mobilité (ce qui est la moindre des choses de la part d'un processus), il est abusif de parler de moyenne...

Nous sommes en fait dans le cadre d'une chronique de prévisions qui montrent des erreurs autocorrélées mais rien d'autre (au fait cela servira pour un mélange un peu plus loin).

Présentons le concept de processus autorégressif de moyenne mobile en introduisant le cas particulier de processus autorégressif de moyenne mobile d'ordre 1, noté MA(1) et défini par la relation:

$$X_t = \varepsilon_t + \theta \varepsilon_{t-1} \quad (66.829)$$

Pour un tel processus, nous avons par définition du bruit blanc:

$$\begin{aligned} E(X_t) &= E(\varepsilon_t) + \theta E(\varepsilon_{t-1}) = 0 + \theta \cdot 0 = 0 \\ V(X_t) &= V(\varepsilon_t) + \theta^2 V(\varepsilon_{t-1}) + 2 \text{cov}(\varepsilon_t, \varepsilon_{t-1}) = \sigma^2 + \theta^2 \sigma^2 = \sigma^2 (1 + \theta^2) \end{aligned} \quad (66.830)$$

Donc les deux moments sont indépendants du temps est nous avons affaire à un processus stationnaire faible. Cependant, il ne faut pas considérer cela comme forcément positif car il n'est pas très cohérent dans la pratique que la volatilité soit constante (processus homoscédastique). Raison pour laquelle nous aborderons des modèles plus complexes dont la volatilité est dépendante du temps.

En ce qui concerne l'autocorrélation, nous avons d'abord (ne pas oublier de revoir les trois conditions qui définissent le bruit blanc et particulièrement la troisième):

$$\begin{aligned} \gamma_t(1) &= \text{cov}(X_t, X_{t-1}) = \text{cov}(\varepsilon_t + \theta \varepsilon_{t-1}, \varepsilon_{t-1} + \theta \varepsilon_{t-2}) = E(X_t \cdot X_{t-1}) - E(X_t) E(X_{t-1}) \\ &= E(X_t \cdot X_{t-1}) = E(\varepsilon_t \varepsilon_{t-1}) + \theta E(\varepsilon_t \varepsilon_{t-2}) + \theta E(\varepsilon_{t-1}^2) + \theta^2 E(\varepsilon_{t-1} \varepsilon_{t-2}) = \theta E(\varepsilon_{t-1}^2) \\ &= \theta \left[V(\varepsilon_{t-1}) + E(\varepsilon_{t-1})^2 \right] = \theta \sigma^2 \end{aligned} \quad (66.831)$$

Le coefficient d'autocorrélation d'ordre 1 vaut donc:

$$\begin{aligned}\rho_t(1) &= \frac{\gamma_t(1)}{\sqrt{\gamma_t(0)\gamma_{t-1}(0)}} = \frac{\text{cov}(X_t, X_{t-1})}{\sqrt{V(X_t)V(X_{t-1})}} = \frac{\theta\sigma^2}{\sqrt{\sigma^2(1+\theta^2)\sigma^2(1+\theta^2)}} \\ &= \frac{\theta\sigma^2}{\sigma^2(1+\theta^2)} = \frac{\theta}{1+\theta^2}\end{aligned}\quad (66.832)$$

Pour $h > 1$, nous avons immédiatement en faisant le même développement que:

$$\gamma_t(h) = \text{cov}(X_t, X_{t-h}) = \text{cov}(\varepsilon_t + \theta\varepsilon_{t-1}, \varepsilon_{t-h} + \theta\varepsilon_{t-h-1}) = 0 \quad (66.833)$$

donc les coefficients d'autocorrélations d'ordre supérieur à 1 sont tous nuls!

Un processus MA(2) est un processus autorégressif qui vérifie une relation de la forme:

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \quad (66.834)$$

Et nous pouvons réitérer les mêmes développements que pour MA(1) avec enthousiasme... pour arriver aussi à la conclusion que l'autocorrélation (ACF) est nulle pour $\forall h > 2$.

En général, un processus MA(q) est un processus qui dépend linéairement des q valeurs antérieures:

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} \quad (66.835)$$

et donc pour un tel processus on peut admettre sans faire la démonstration générale que:

$$\begin{aligned}E(X_t) &= 0 \\ V(X_t) &= \sigma^2 \sum_{i=0}^q \theta_i^2\end{aligned}\quad (66.836)$$

et que l'autocorrélation est nulle pour $\forall h > q$. Ainsi, on peut facilement reconnaître un processus MA(q) car son ACF présente un "cut-off" à partir de q .

8.4.3. PROCESSUS ARMA

Nous pouvons bien évidemment envisager de combiner les deux modèles précédents (MA(q) et AR(p)) en introduisant:

- Une dépendance du processus vis-à-vis de son passé avec un processus AR(p)
- Un effet retard des aléas avec un processus MA(q)

Un tel modèle, appelé "**autorégressif-moyenne mobile non saisonnier**" ou "**ARMA non saisonnier**", est caractérisé par le paramètre p de la partie autorégressive et le paramètre q de la partie moyenne mobile.

Ainsi, un processus ARMA(p, q) non saisonnier vérifie la relation:

$$X_t = \varphi_1 X_{t-1} + \dots + \varphi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} \quad (66.837)$$

Le principal avantage des processus ARMA non saisonniers tient au fait qu'ils permettent de fournir des prévisions pour des échéances éloignées (du moins plus éloignée dans le temps que la prochaine date). Effectivement, rappelons que pour les processus AR(1), nous avons obtenu:

$$E(X_t) = \varphi E(X_{t-1}) \quad (66.838)$$

Nous avons alors:

$$\hat{X}_t = E(X_t | X_{t-1}) = \varphi X_{t-1} \quad (66.839)$$

Soit in extenso:

$$\hat{X}_{t+1} = E(X_{t+1} | X_t) = \varphi X_t \quad (66.840)$$

Les prévisions s'obtiennent ensuite de manière récursive:

$$\hat{X}_{t+2} = E(X_{t+2} | X_t) = \varphi E(X_{t+1} | X_t) = \varphi^2 X_t \quad (66.841)$$

Et donc d'une façon plus générale:

$$\hat{X}_{t+k} = \varphi^k X_t \quad (66.842)$$

et donc lors k devient très grand, nous avons alors:

$$\hat{X}_{t+k} \rightarrow 0 \quad (66.843)$$

Autrement dit, pour un ordre k élevée, le modèle AR(1) se contente de fournir comme espérance conditionnelle la prévision la moyenne (historique) du processus. Quand l'espérance conditionnelle est nulle, nous disons alors dans le domaine des séries temporelles que "**l'espérance est orthogonale à tout passé**".

8.4.4. PROCESSUS ARIMA

Les modèles de séries chronologiques vus précédemment sont en réalité que des cas particuliers d'un modèle général empirique que l'on appelle les modèles "**ARIMA**" pour "**AutoRegressive Integrated Moving Average**" connus aussi sous le nom de "**technique de Box-Jenkins**".

Cette technique présuppose que la série chronologique est générée par un processus aléatoire (brownien ou autre) et que les valeurs futures sont liées aux erreurs passées.

Cependant ce modèle a une limitation importante, et le fait que l'on retrouve à travers cette généralisation les lissages exponentiels signifie que ceux-ci ont aussi la même limitation: la série chronologique doit être stationnaire. Donc pour rappel, cela signifie que les moments comme l'espérance ou la variance et l'autocorrélation doivent être constants dans le temps.

Un modèle ARIMA est souvent classifié sous la notation:

$$\text{ARIMA}(p, d, q)$$

où p est le nombre de termes autorégressifs, d le nombre de différences non-saisonnalières et q le nombre de termes d'erreurs dans l'équation de prédiction. Voici quelques exemples:

- Moyenne: ARIMA(0,0,0) avec constante
- Naïf: ARIMA(0,1,0)
- Drift: ARIMA(0,1,0) avec constante
- Lissage exponentiel simple: ARIMA(0,1,1)
- Lissage exponentiel double de Holt: ARIMA(0,2,2)
- Lissage exponentiel de Holt-Winters (multiplicatif): SARIMA(0,1, $m+1$)(0,1,0)

Revenons un peu sur les techniques de projection (prédiction) des modèles ARIMA. Commençons par un cas particulièrement simple (c'est à notre avis plus pédagogique) en considérant le processus noté AR(1) et défini pour rappel par la relation (à une constante additive près):

$$X_t = \varphi X_{t-1} + \varepsilon_t \quad (66.844)$$

Lorsque nous faisons l'estimation ponctuelle d'une projection de ce modèle, sur l unités de temps ultérieures, il est d'usage de noter la valeur ponctuelle estimée par différentes notations d'usage:

$$X_t(l) = X_{t+l} = \varphi X_{t+l-1} + \varepsilon_{t+l} \quad (66.845)$$

où φ est bien évidemment supposé avoir été déterminé à l'aide des valeurs historiques connues.

Cependant, l'estimation ponctuelle projetée est... justement une estimation!!!! Il nous faudrait donc connaître son espérance. L'hypothèse naïve la plus simple est alors d'utiliser l'espérance conditionnelle de la valeur projetée, qui nous avait mené (voir juste plus haut) au résultat:

$$\hat{X}_{t+l} = \varphi^l X_t \quad (66.846)$$

qui tend donc vers 0 quand l tend vers l'infini comme nous l'avons déjà mentionné. Mais n'oubliez pas que ce modèle AR(1) est défini à une constante additive près (qui est souvent la moyenne arithmétique ou la médiane des valeurs connues de la série temporelle), donc dans le cas général cela tend vers une constante.

Au passage, observons que nous avons explicitement l'erreur de prévision qui est donnée par:

$$e_t(l) = X_{t+l} - \hat{X}_t(l) = X_{t+l} - \varphi^l X_t \quad (66.847)$$

Et rappelons que nous avons démontré plus haut que pour les AR(1):

$$X_t = \varphi^h X_{t-h} + \sum_{k=0}^{h-1} \varphi^k \varepsilon_{t-k} \quad (66.848)$$

Et considérons h (respectivement l) grand et $|\varphi| < 1$, nous avons alors:

$$X_t \cong \sum_{k=0}^{\infty} \varphi^k \varepsilon_{t-k} \quad (66.849)$$

Adapté à notre contexte de notation:

$$X_{t+l} \cong \sum_{k=0}^{\infty} \varphi^k \varepsilon_{t+l-k} \quad (66.850)$$

Soit, un processus $MA(\infty)$. Alors:

$$\begin{aligned} e_t(l) &= X_{t+l} - \hat{X}_t(l) = X_{t+l} - \varphi^l X_t \underset{\substack{|\varphi| < 1 \\ l \rightarrow \infty}}{\cong} \sum_{k=0}^{\infty} \varphi^k \varepsilon_{t+l-k} - \varphi^l \sum_{k=0}^{\infty} \varphi^k \varepsilon_{t-k} \\ &= \varepsilon_{t+l} + \varphi \varepsilon_{t+l-1} + \dots + \varphi^{l-1} \varepsilon_{t+1} + \varphi^l \varepsilon_t + \varphi^{l+1} \varepsilon_{t+1} + \dots - \varphi^l \varepsilon_t - \varphi^{l+1} \varepsilon_{t-1} - \dots \\ &= \varepsilon_{t+l} + \varphi \varepsilon_{t+l-1} + \varphi^2 \varepsilon_{t+l-2} + \dots + \varphi^{l-1} \varepsilon_{t+1} \end{aligned} \quad (66.851)$$

Nous remarquons donc que l'erreur de projection $e_t(l)$ est un processus du type $MA(l-1)$. Il vient alors de façon immédiate de ce dernier développement:

$$\begin{aligned} E(e_t(l)) &= 0 \\ V(e_t(l)) &= \sigma^2 \left(1 + \varphi^2 + \varphi^4 + \dots + \varphi^{2(l-1)} \right) = \sigma^2 \frac{1 - \varphi^{2l}}{1 - \varphi^2} \end{aligned} \quad (66.852)$$

où la somme des carrés a été simplifiée en utilisant un résultat démontré dans le chapitre de Suites Et Séries. Nous voyons ainsi que la variance de l'erreur de prévision ne diverge pas (du moins dans le cas d'un AR(1)) et tend vers une valeur finie (ce qui n'est pas le cas de tous les processus) quand l tend vers l'infini à condition que $|\varphi| < 1$. La difficulté dans la pratique étant de choisir une technique pour déterminer σ^2 (il en existe une quantité: adéquation, Monte-Carlo, Boostapping, bayésienne, etc.). Raison pour laquelle il ne faut jamais se limiter à l'utilisation d'un seul logiciel statistique pour faire des projections mais au moins à trois différents.

Pourquoi revenir là-dessus penserez-vous peut-être??? Pour la simple raison qu'il nous faut un intervalle statistique de prédiction. Ainsi, étant donné que nous avons un estimateur ponctuel X_{t+l} qui est la réalisation de la variable aléatoire de la valeur projetée et son espérance associée $\hat{X}_{t+l}$, nous pouvons écrire:

$$X_{t+l} = \hat{X}_t(l) + e_t(l) \quad (66.853)$$

Et faire l'hypothèse à la vue des développement précédents que l'erreur de prévision suit une loi Normale centrée:

$$\left. \begin{array}{l} E(e_t(l)) = 0 \\ V(e_t(l)) = \sigma^2 \frac{1 - \phi^{2l}}{1 - \phi^2} \end{array} \right\} \Rightarrow \text{Hypothèse : } e_t(l) = N\left(0, \sigma^2 \frac{1 - \phi^{2l}}{1 - \phi^2}\right) \quad (66.854)$$

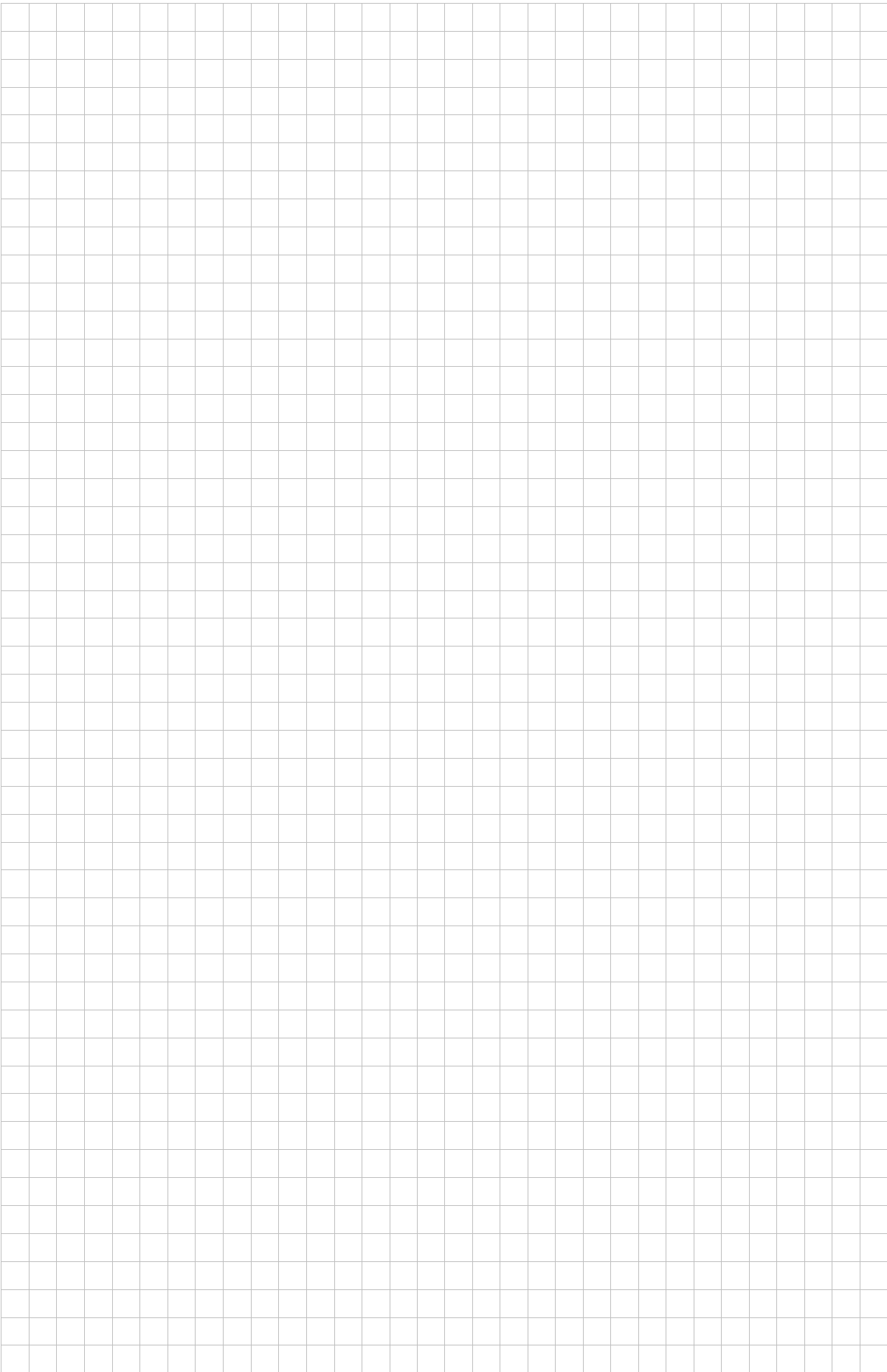
Nous avons alors à un seuil de confiance donné si σ^2 est parfaitement déterminé:

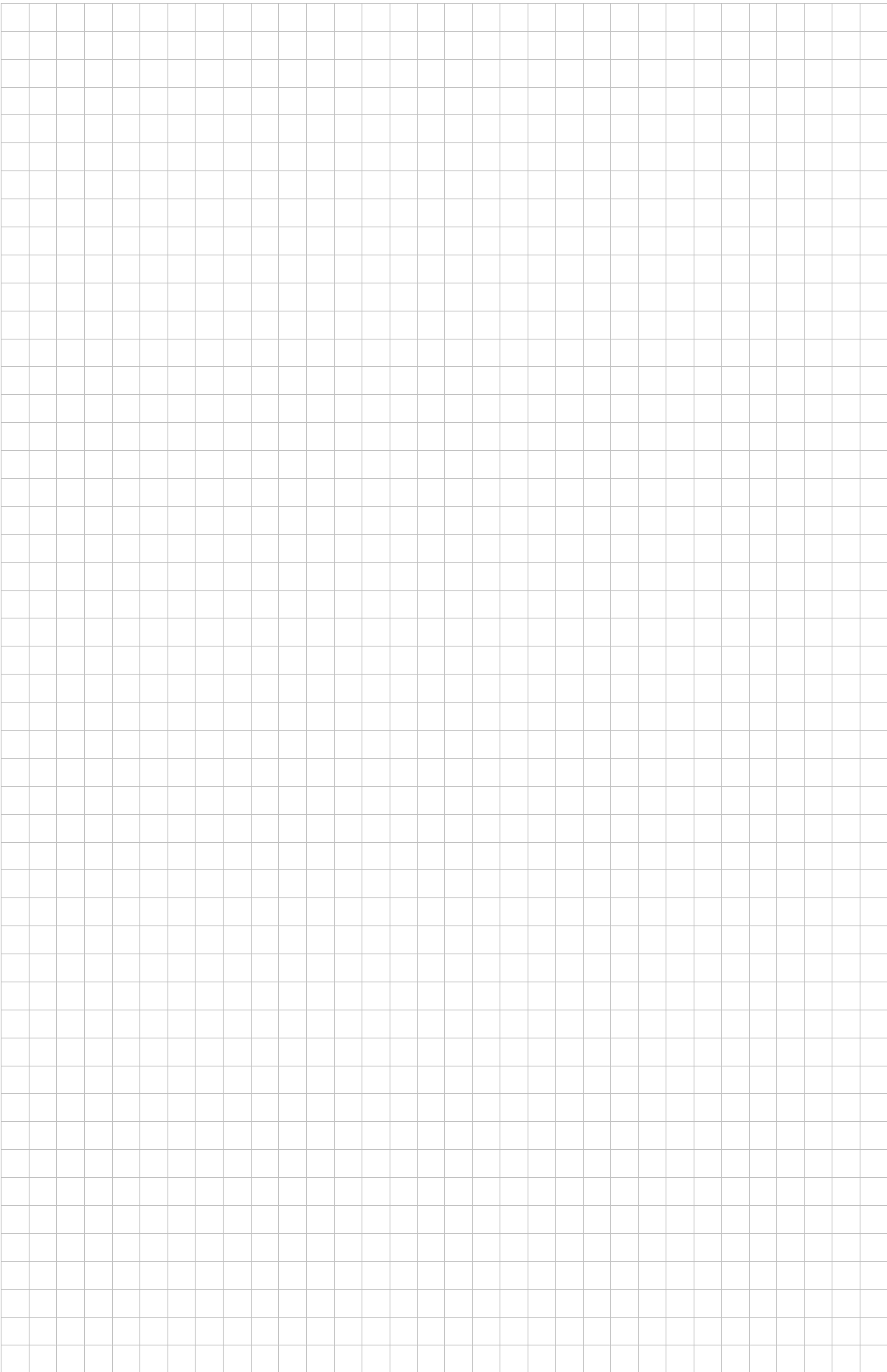
$$\hat{X}_t(l) - Z_{\alpha/2} \sigma \sqrt{\frac{1 - \phi^{2l}}{1 - \phi^2}} \leq X_{t+l} \leq \hat{X}_t(l) + Z_{\alpha/2} \sigma \sqrt{\frac{1 - \phi^{2l}}{1 - \phi^2}} \quad (66.855)$$

et si nous n'avons qu'un estimateur de σ^2 alors nous prendrons (cf. chapitre de Statistiques):

$$\hat{X}_t(l) - T_{\alpha/2} \hat{\sigma} \sqrt{\frac{1 - \phi^{2l}}{1 - \phi^2}} \leq X_{t+l} \leq \hat{X}_t(l) + T_{\alpha/2} \hat{\sigma} \sqrt{\frac{1 - \phi^{2l}}{1 - \phi^2}} \quad (66.856)$$

La pratique de ces modèles de projections tendrait à montrer qu'ils sous-estiment le risque (la variance). Donc plutôt que de prendre un intervalle à 95% il vaut peut-être mieux prendre un intervalle plus élevé en fonction du retour sur expérience.







67. TECHNIQUES DE GESTION

L'objectif de ce chapitre est d'introduire les principales techniques mathématiques de production et de gestion, de maintenance et de qualité dont l'utilisation est devenue indispensable aux ingénieurs, gestionnaires et cadres des entreprises modernes (ceux qui le peuvent font de la science, les autres de la méthodologie...) et qui constituent le minimum-minimum de la connaissance du métier de la gestion en général. Ce sont des techniques qui datent du début du 20ème siècle et qui ont d'abord été utilisées par l'armée américaine pour la majorité pour ensuite se voir petit à petit implémentées dans les plus grandes entreprises mondiales.

Par ailleurs, c'est un excellent chapitre pour la culture générale du physicien ou du mathématicien... et un acquis pour l'ingénieur!

Beaucoup de techniques ne seront cependant pas présentées ici car ayant déjà été démontrées dans d'autres chapitres du site. Effectivement, les techniques de gestion complexes et critiques font un énorme usage des statistiques et probabilités, de la théorie de la décision, de la théorie des graphes (chaînes de Markov ou non, graphes complets pour la gestion de la communication), de l'analyse financière et des algorithmes d'optimisation. Des chapitres entiers y étant déjà consacrés sur ce site, il serait redondant d'y revenir!

Remarque: Nous parlons souvent de 3M (Méthodes Mathématiques de Management) ou chez les anglophones de "scientific management" pour décrire l'ensemble des outils mathématiques appliqués à la gestion (il existe d'ailleurs un cursus de formation continue à ce sujet d'une cinquantaine de jours) ou encore de SciProM pour "Scientific Project Management). Un terme anglophone courant et qui devient à la mode en Europe pour décrire ce domaine d'application est aussi le "**decisioneering**" faisant référence au fait que ce sont des outils d'aide à la décision pour les ingénieurs.

Indiquons avant de commencer que malgré leur importance intrinsèque les techniques quantitatives de gestion (incluant aussi les techniques de recherche opérationnelle, modélisation par Monte-Carlo et le Bootstrapping vues dans le chapitre d'Informatique Théorique ainsi que Six Sigma vu dans le chapitre de Génie Industriel et la Théorie de la Décision) sont encore peu utilisées dans le monde industriel, soit à cause du manque de formation des décideurs (excepté dans le domaine de la finance), soit par le manque de pertinence de l'outil ou sa difficulté de mise en œuvre. Les principales craintes émises par les décideurs quant à l'application de modèles mathématiques dans l'entreprise sont:

1. Les outils sont trop abstraits: Dans cette situation, je rappelle toujours que le concret, c'est de l'abstrait auquel on est habitué!
2. Une prise en compte limitée des facteurs: Pour les questions stratégiques, la réponse pure et parfaite d'une solution mathématique semble rarement applicable de facto. Même si les techniques quantitatives intègrent beaucoup de facteurs, si certains aspects sont relativement faciles à modéliser au sens mathématique du terme (le coût, la rentabilité, la distance, la durée, la cadence, par exemple), d'autres éléments sont en revanche plus difficiles à modéliser: contraintes légales, volonté commerciale de faire barrage à un concurrent, importance des relations avec les élus, climat social, etc. Le poids de ces éléments dans la décision est pourtant important, parfois déterminant.

3. Un investissement important: L'outil mathématique lui-même exige un niveau élevé de connaissances mathématiques, une bonne aptitude à modéliser les problèmes et décrire les facteurs. Ces contraintes sont consommatrices de beaucoup de temps et d'une certaine somme d'argent (que ce soit par développement interne, qui consomme des ressources - ou par développement externe, qui consomme de l'argent). Il est alors nécessaire de trouver un équilibre entre l'investissement nécessaire et les retombées prévues.

4. Pour des événements peu fréquents: L'entreprise ne bénéficie pas de l'effet d'expérience et donc d'une fois sur l'autre, le problème concerne un service différent, ou les responsables ont changé entre deux études. Il est donc difficile d'entretenir les compétences à l'intérieur de l'entreprise. Le décideur devra prendre ces différents aspects en compte lorsqu'il décidera ou non de mettre en œuvre des techniques quantitatives dans son entreprise.

Nous pouvons aussi argumenter que ce qui fait que les techniques scientifiques de gestion (TQG) ou les méthodes mathématiques de management (3M) ne sont guère répandues est que:

1. Dans un environnement peu exigeant le surcroît de puissance qu'offrent les TQG/3M ne justifie pas forcément les efforts nécessaires à leur apprentissage.

2. Les TQG/3M sont exigeantes en raisonnements et rigoureuses, elles ne peuvent donc convenir aux personnes qui ne connaissent ni les uns ni les autres...

Remarque: Méfiez-vous des entreprises - particulièrement des multinationales - qui recherchent des spécialistes en gestion de projets ou en logistique maîtrisant Microsoft Excel ou Microsoft Access. Car cela signifiera qu'elles utilisent des outils non professionnels pour faire un travail qui lui devrait pourtant l'être avec des outils adaptés (et Microsoft Excel ou Microsoft Access ne le sont pas)!!! Donc en termes d'organisation interne, vous pouvez vous assurer que ces entreprises organisent et analysent n'importe quoi, n'importe comment, avec un outil non adapté et donc que c'est le bordel général en interne.

Nous avons choisi de présenter les techniques quantitatives de gestion les plus classiques dans un ordre croissant de difficulté (niveau de technicité). Cela nous semble à ce jour le choix pédagogique le plus pertinent.

Redonnons un résumé schématique de la méthode scientifique vue dans le chapitre d'Introduction et qui s'applique assez bien au niveau du monde des affaires en ce qui concerne le management scientifique:

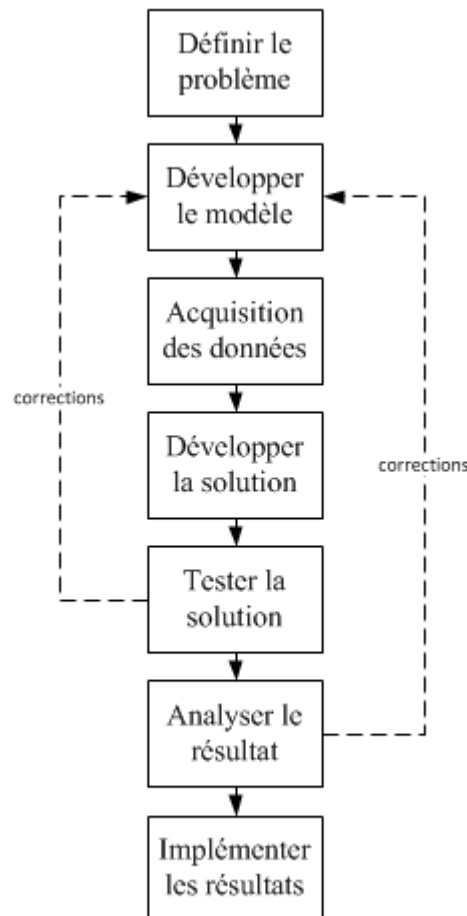


Figure: 67.1 - Processus simplifié type

Les sujets simples sont donnés sans exemples pratiques alors que ceux qui dépassent le niveau de l'école obligatoire sont accompagnés d'exemples pratiques plus ou moins détaillés.

1. ANALYSE DU SEUIL DE RENTABILITÉ (POINT MORT)

L'analyse du seuil de rentabilité ("break-even analysis" en anglais), appelé aussi "analyse du point mort", est l'ensemble des calculs permettant de déterminer le niveau d'activité minimum à partir duquel une organisation devient rentable pour elle-même par ses économies d'échelle, c'est-à-dire qu'elle cesse de perdre de l'argent sur cette activité. Évidemment, cette technique permet aussi de comparer plusieurs stratégies entre elles.

Le point mort est donc assez intuitivement point d'intersection entre la courbe du chiffre d'affaires (car le chiffre d'affaires est normalement une fonction de plusieurs variables) et la courbe des charges (elle aussi une fonction de plusieurs variables) nécessaires pour produire ce chiffre d'affaires. En d'autres termes, il s'agit des coordonnées correspondantes à l'égalité des deux fonctions (dans la pratique on le représente rarement graphiquement car les variables sont trop nombreuses).

Normalement cette technique doit permettre à un gestionnaire de déterminer des informations comme:

- Prédire le volume de ventes pour couvrir tout juste les coûts (point mort)
- Quel doit être le prix par unité pour couvrir tout juste les coûts (point mort)

- Quel est le niveau des coûts fixes qui équilibre tout juste les gains (point mort)
- Comment les prix influent sur le volume de ventes correspondant à l'équilibre (point mort)

L'analyse du point mort est basée sur l'hypothèse que tous les coûts relatifs à la fabrication d'un bien peuvent être divisés en deux catégories: coûts variables et coûts fixes.

Le coût total T_c est défini par la relation:

$$T_c = F + cQ \quad (67.1)$$

où F représente les coûts fixes, c le coût par unité (supposé indépendant de la quantité) de biens produits et Q la quantité correspondante.

L'hypothèse que le coût par unité soit indépendant de la quantité permet est de simplifier le calcul pour avoir le coût total qui est une simple fonction linéaire, mais cette hypothèse n'est plus vraiment nécessaire aujourd'hui avec les tableurs.

Si nous supposons que toutes les unités produites sont vendues, le revenu total T_r est défini par:

$$T_r = pQ \quad (67.2)$$

où p est le revenu par unité vendue.

Le point mort est alors l'égalité entre les deux coûts, ce qui signifie:

$$pQ = F + cQ \quad (67.3)$$

Ainsi, nous pourrions alors par exemple après quelques opérations algébriques élémentaires, déterminer la quantité qui permet d'arriver au point mort:

$$Q = \frac{F}{p - c} \quad (67.4)$$

Ainsi, connaissant F (coûts fixes), p (revenus par unité produite) et c (coûts par unité produite), il est possible de déterminer la quantité Q à vendre pour arriver au point mort.

Signalons qu'un point qui intéresse beaucoup les gestionnaires c'est l'étude de la variation de certains paramètres en fonction des autres variables. Nous appelons cela "**l'analyse de la sensibilité**". Évidemment dans le cas linéaire, cette analyse est assez triviale.

Remarques: Depuis le début des années 1980, il existe des logiciels très simples à utiliser comme @Risk permettant de faire des simulations de Monte-Carlo pour déterminer le point-mort et ce même avec des solveurs qui permettent de résoudre des systèmes non linéaires et faire de l'analyse de la sensibilité très poussée.

2. DIAGRAMME DE PARETO

Le diagramme de Pareto, comme nous allons le voir, est un moyen (parmi d'autres) simple pour classer les phénomènes par ordre d'importance. Mais la distribution de Pareto est aussi souvent utilisée en entreprise comme base de simulation stochastique pour des variables aléatoires représentant des investissements chiffrés de projets (à peu près aussi souvent utilisée que la distribution triangulaire, la distribution Bêta ou log-Normale dans l'industrie moderne).

Rappelons alors qu'une variable aléatoire est dite par définition suivre une loi de Pareto si sa fonction de répartition est donnée par (cf. chapitre de Statistiques):

$$P(X \leq x) = 1 - \left(\frac{x_m}{x} \right)^k \quad (67.5)$$

avec $k \in \mathbb{R}_+$ et $x \geq x_m \geq 0$ (donc $x > 0$) et pour l'espérance (moyenne):

$$\mu = \frac{k \cdot x_m}{k - 1} \quad (67.6)$$

et la variance:

$$\sigma^2 = V(X) = \frac{k \cdot x_m^2}{k - 2} - \left(\frac{k \cdot x_m}{k - 1} \right)^2 = \frac{k x_m^2}{(k - 1)^2 (k - 2)} \quad (67.7)$$

Pour illustrer ce type de représentation, nous supposerons qu'une étude de réorganisation du réseau commercial de l'entreprise LAMBDA conduise le directeur commercial à s'intéresser à la répartition des 55'074 bons de commande (valeur en bas à droite du tableau) reçus pendant une année donnée selon la ville où sont domiciliés les clients (nous avons retenu ici les 200 premières villes du pays imaginaire d'où les 200 lignes réparties en 5 colonnes).

Pour calculer les paramètres de la loi de Pareto, le lecteur devra se reporter au chapitre de Génie Industriel, car nous n'allons ici avoir qu'une approche qualitative comme le font les gestionnaires, l'analyse quantitative étant réservée majoritairement aux ingénieurs.

Ces données sont reproduites dans le tableau ci-dessous:

n_i	N_i	n_i	N_i	n_i	N_i	n_i	N_i	n_i	N_i
8965	8965	240	40027	125	46378	88	50916	55	53752
4555	13520	236	40263	124	46862	87	51003	54	53806
3069	16589	224	40487	123	46985	87	51090	54	53860
2336	18925	223	40710	123	47108	86	51176	54	53914
1871	20796	220	40930	121	47229	85	51261	54	53968
1595	22391	215	41145	121	47350	85	51346	50	54018
1327	23718	213	41357	118	47468	84	51430	50	54068
1184	24902	212	41570	118	47586	84	51514	48	54116
1085	25987	202	41772	118	47704	83	51597	48	54164
934	26921	202	41972	117	47821	81	51678	47	54211
843	27764	200	42172	115	47936	79	51757	44	54255
746	28510	189	42363	114	48050	78	51835	43	54298
722	29232	190	42553	112	48162	78	51913	42	54340
710	29943	188	42741	112	48274	77	51990	42	54382
647	30589	185	42926	108	48382	77	52067	41	54423
631	31221	179	43105	108	48490	77	52144	40	54463
607	31828	173	43278	107	48597	74	52218	40	54503
539	32367	170	43448	106	48703	72	52290	39	54542
502	32868	161	43609	105	48808	72	52362	38	54580
462	33330	160	43769	102	48910	72	52434	37	54617
459	33790	158	43927	102	49012	71	52505	37	54654
422	34212	153	44080	102	49114	71	52576	36	54690
372	34584	153	44233	101	49215	71	52647	36	54726
362	34945	149	44382	100	49315	69	52716	34	54760
355	35300	149	44531	100	49415	68	52784	33	54793
347	35647	148	44679	99	49514	67	52851	32	54825
342	35989	147	44826	98	49612	67	52918	31	54856
338	36327	147	44973	98	49710	65	52983	28	54884
329	36656	146	45119	97	49807	64	53047	25	54909
314	36970	146	45265	97	49904	63	53110	22	54931
313	37283	146	45411	96	50000	63	53173	20	54951
310	37593	142	45553	95	50095	63	53236	20	54971
310	37903	140	45693	95	50190	60	53296	20	54991
300	38203	137	45830	94	50284	60	53356	19	55010
284	38755	136	45966	92	50376	58	53414	15	55025
268	38755	135	46101	92	50468	57	53471	13	55038
267	39022	131	46232	91	50559	57	53539	11	55049
265	39287	128	46360	91	50650	57	53528	11	55060
251	39538	127	46487	90	50740	56	53641	8	55068
249	39787	126	46613	88	50828	56	53697	6	55074

Tableau: 67.1 - Dataset pour analyse de Pareto

où les 200 valeurs n_i (nombre de bons de commande en provenance d'une ville i) ont été classées par valeurs décroissantes et cumulées dans une colonne N_i .

La première ville se caractérise par 8'965 bons de commande (ce qui correspond à 16.28% du total des bons), la seconde par 4'555, ce qui fait que les deux premières villes ont passé 13'520 bons de commande (ce qui correspond à 24.55% des bons), les trois premières villes, ont passé 16'589 bons de commande (ce qui correspond à 30.12%), etc. Une autre façon de décrire ce phénomène consiste à dire: 0.5% des villes (classées par valeur décroissante du critère) ont passé 16.28% des bons, 1% des villes ont passé 24.55% des bons, etc.

Ces calculs, sont partiellement présentés dans le tableau ci-dessous:

$i\%$	$N_i\%$	$i\%$	$N_i\%$	$i\%$	$N_i\%$	$i\%$	$N_i\%$	$i\%$	$N_i\%$
0.5	16.28	20.5	72.68	40.5	84.86	60.6	92.45	80.5	97.6
1	24.55	21	73.11	41	85.09	61	92.61	81	97.7
1.5	30.12	21.5	73.51	41.5	85.32	61.5	92.77	81.5	97.8
2	34.36	22	73.92	42	85.54	62	92.92	82	97.89
...	...	...	...	...	...	...	...	...	...
19	71.33	39	84.18	59	91.97	79	97.3	99	99.7
19.5	71.79	39.5	84.41	59.5	92.13	79.5	97.4	99.5	99.99
20	72.24	40	84.64	60	92.29	80	97.5	100	100

Tableau: 67.2 - Données cumulées et normalisées pour analyse de Pareto

traduits graphiquement (selon les traditions d'usage) ci-dessous sous la forme d'un diagramme de Pareto:

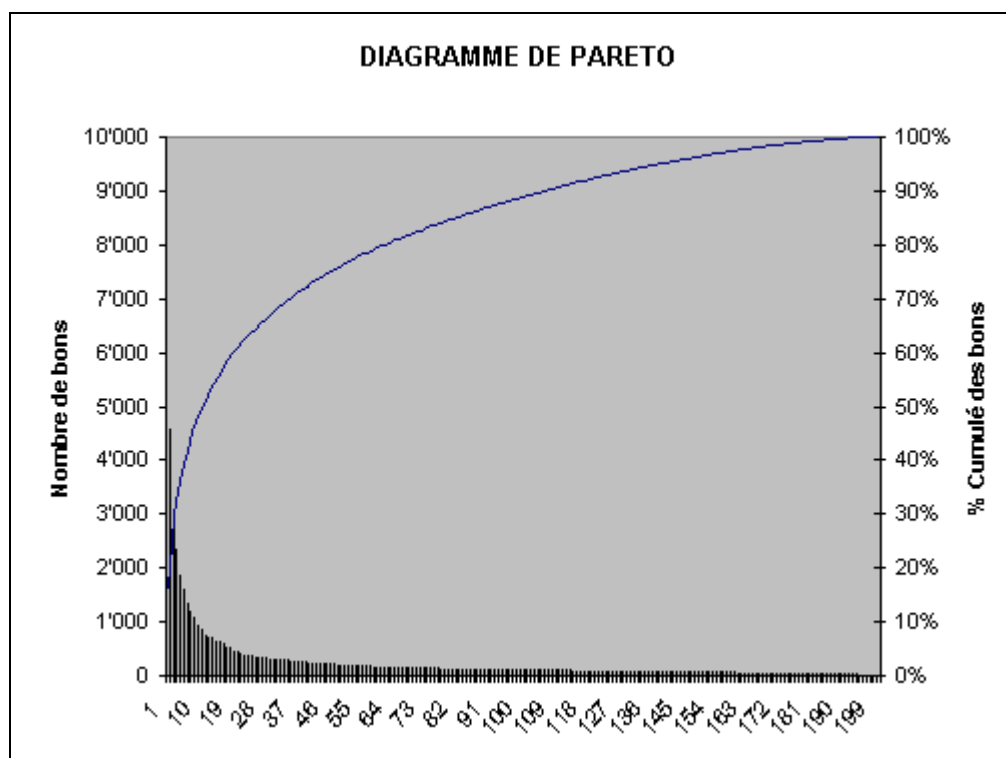


Figure: 67.2 - Diagramme de Pareto

ou de manière plus pertinente sous la forme d'un diagramme de Lorenz (l'abscisse des numéros des villes est remplacée par le %cumulé de villes comme dans le tableau de données ci-dessus):

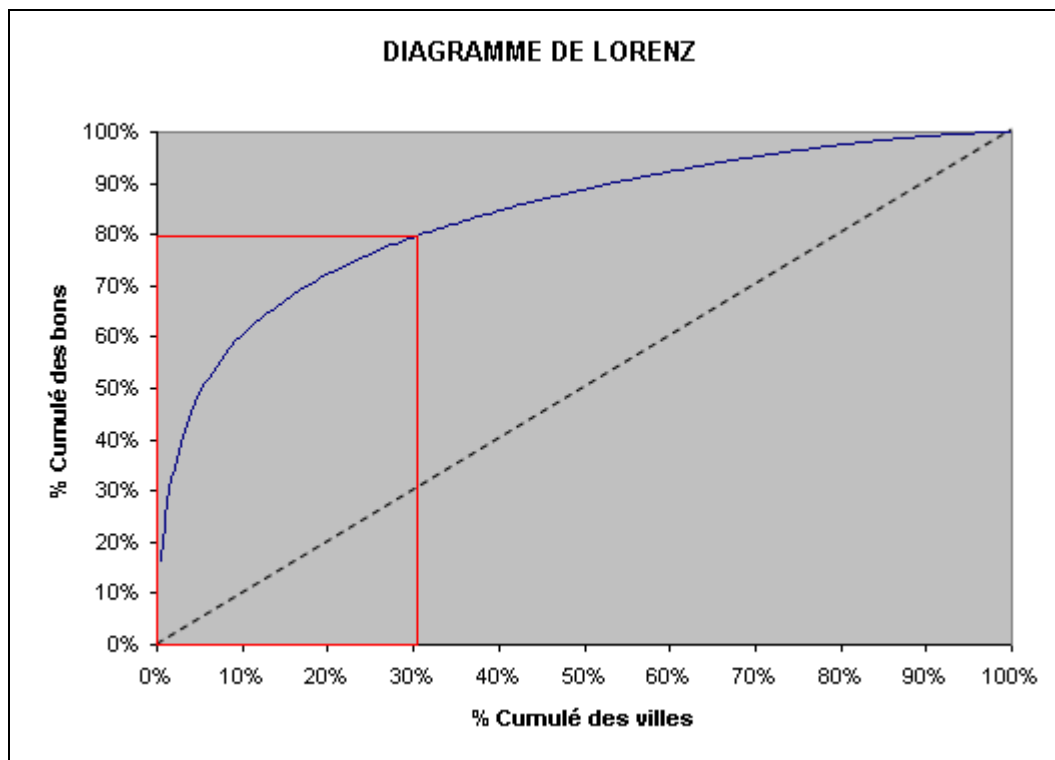


Figure: 67.3 - Diagramme de Lorenz

Cette analyse met en évidence qu'un nombre faible d'éléments est à l'origine de la majeure partie du phénomène étudié (par exemple, ici 31% des villes génèrent 80% des bons de commande). Ceci explique qu'elle soit l'une des grandes techniques utilisées dans le domaine de la "**gestion de la qualité totale**" (Total Quality Management) ET dans l'analyse de l'importance des causes d'un problème de qualité. Elle est également utilisée par les gestionnaires pour structurer l'organisation, en particulier pour différencier les processus en fonction de caractéristique de la demande (suivi différencié de la clientèle selon son importance, par exemple) et lorsqu'elle est utilisée pour définir 3 classes, cette technique prend souvent le nom de "**Méthode ABC**" (cf. chapitre de Génie Industriel).

La ligne en pointillés de ce graphique correspond, elle a ce que nous aurions observé en cas d'équirépartition du phénomène étudié, c'est-à-dire si chaque ville s'était caractérisée par le même nombre de bons de commande.

De façon générale, plus une courbe de Pareto (ou de Lorenz) se rapproche de la droite d'égalité parfaite et plus la répartition de la masse considérée au sein de la population est égalitaire. En effet, dans ce cas, la masse (des salaires, de la richesse, du revenu, etc.) est peu concentrée sur quelques-uns.

Remarque: La présentation de cette analyse a été faite en classant les observations par valeurs décroissantes, mais nous aurions pu tout aussi bien partir d'un classement par valeurs croissantes et, dans ce dernier cas, la courbe obtenue aurait été symétrique, le centre de symétrie étant le point de coordonnées (0.5,0.5).

Dans un environnement industriel, les points d'amélioration potentiels sont quasi innombrables. On pourrait même améliorer indéfiniment, tout et n'importe quoi. Il ne faut cependant pas perdre de vue que l'amélioration coûte aussi et qu'en contrepartie elle devrait apporter de la valeur ajoutée, ou au moins apurer des pertes.

2.1. INDICE DE GINI

Le phénomène étudié est d'autant moins équitablement réparti que la courbe s'éloigne de la droite d'équirépartition. Les économistes, gestionnaires ou responsables de départements d'entreprise utilisent parfois (pour leur tableau de bord de performance) un indicateur synthétique pour mesurer ce phénomène et son évolution, "l'indice de Gini" (appelé aussi "coefficient de Gini").

Ce coefficient est défini par le rapport:

$$G = \frac{A}{A+B} \quad (67.8)$$

où les surfaces A et B se rapportent à la figure ci-après.

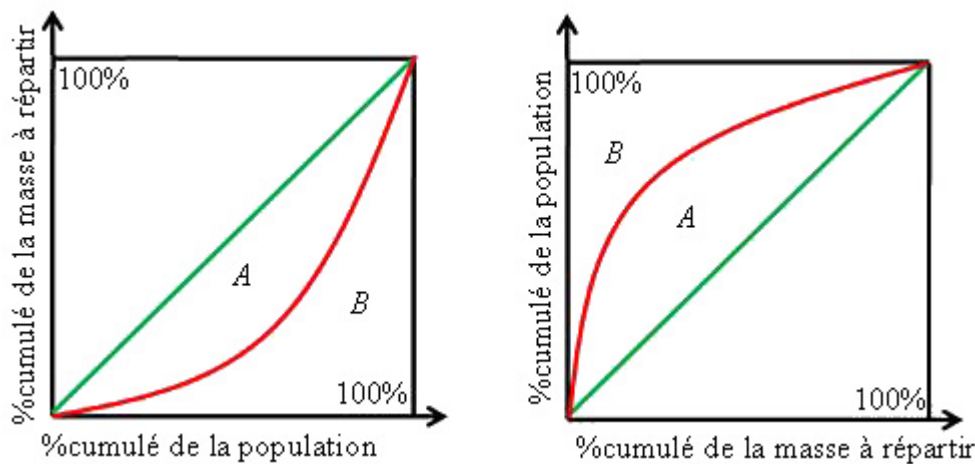


Figure: 67.4 - Les deux situations les plus courantes de l'indice de Gini

Le coefficient de Gini est donc un nombre réel compris entre zéro et 1. En cas d'égalité parfaite, il est égal à zéro (car $A=0$). En cas d'inégalité totale il est égal à 1, car $B=0$. Par conséquent, à mesure que G augmente de 0 à 1, l'inégalité de la répartition augmente.

Sachant que la courbe de Lorenz est 1×1 on voit que la surface $A+B$ est égale à la moitié de cette surface. Nous avons donc:

$$A+B = \frac{1}{2} \quad (67.9)$$

Nous pouvons de ce fait écrire:

$$G = \frac{A}{A+B} = \frac{A}{\frac{1}{2}} = 2A \quad (67.10)$$

De plus comme:

$$A+B = \frac{1}{2} \Rightarrow A = \frac{1}{2} - B \quad (67.11)$$

Finalement, nous pouvons écrire que:

$$G = \frac{A}{A+B} = 2A = 1 - 2B \quad (67.12)$$

En utilisant la méthode de calcul d'intégrale des trapèzes (cf. chapitre de Méthodes Numériques) nous avons l'aire B qui est donnée par:

$$\begin{aligned} G &= 1 - 2 \cdot \left(\frac{A_G + A_D}{2} \right) = 1 - (A_G + A_D) = 1 - h \left(\sum_{i=0}^{n-1} f(x_{i+1}) + \sum_{i=0}^{n-1} f(x_i) \right) \\ &= 1 - h \sum_{i=0}^{n-1} (f(x_{i+1}) + f(x_i)) = 1 - h \sum_{i=0}^{n-1} (y_{i+1} + y_i) \end{aligned} \quad (67.13)$$

où pour rappel h est la longueur des intervalles supposés tous égaux (ce qui est toujours le cas dans le contexte des analyses de Lorenz) et si la configuration du tableau de données est telle que nous avons le graphique suivant (ce qui est plutôt rare...):

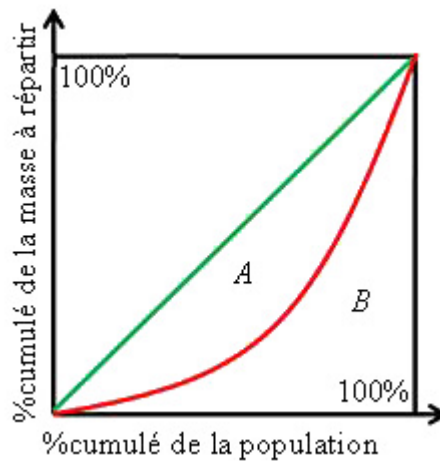


Figure: 67.5 - Choix pour le calcul de l'indice de Gini

Traditionnellement nous trouvons cette dernière relation sous la forme:

$$G = 1 - \frac{1}{n} \sum_{i=1}^n (y_{i+1} + y_i) \quad (67.14)$$

où n représente le nombre d'unités statistiques (la population).

Dans la situation où le graphique est sous la forme (cas qui est nettement le plus fréquent en entreprise et correspond par hasard à notre exemple!!):

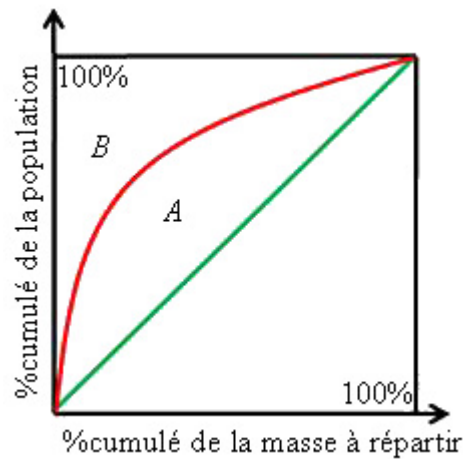


Figure: 67.6 - Deuxième cas pour le calcul de l'indice de Gini

La méthode des trapèzes nous amène alors à écrire:

$$\begin{aligned}
 G &= 1 - 2 \cdot \left(1 - \frac{A_G + A_D}{2} \right) = 1 - 2 + (A_G + A_D) = -1 + h \left(\sum_{i=0}^{n-1} f(x_{i+1}) + \sum_{i=0}^{n-1} f(x_i) \right) \\
 &= -1 + h \sum_{i=0}^{n-1} (f(x_{i+1}) + f(x_i)) = -1 + h \sum_{i=0}^{n-1} (y_{i+1} + y_i)
 \end{aligned} \quad (67.15)$$

Traditionnellement nous trouvons cette dernière relation sous la forme:

$$G = -1 + \frac{1}{n} \sum_{i=1}^n (y_{i+1} + y_i) \quad (67.16)$$

Dans le cas qui nous sert d'exemple depuis le début avec le tableau:

$i\%$	$N_i\%$	$i\%$	$N_i\%$	$i\%$	$N_i\%$	$i\%$	$N_i\%$	$i\%$	$N_i\%$
0.5	16.28	20.5	72.68	40.5	84.86	60.6	92.45	80.5	97.6
1	24.55	21	73.11	41	85.09	61	92.61	81	97.7
1.5	30.12	21.5	73.51	41.5	85.32	61.5	92.77	81.5	97.8
2	34.36	22	73.92	42	85.54	62	92.92	82	97.89
...	...	...	...	...	...	...	...	...	...
19	71.33	39	84.18	59	91.97	79	97.3	99	99.7
19.5	71.79	39.5	84.41	59.5	92.13	79.5	97.4	99.5	99.99
20	72.24	40	84.64	60	92.29	80	97.5	100	100

Tableau: 67.3 - Données cumulées et normalisées pour analyse de Pareto

et en adoptant les notations en conséquence, cela donne:

$$G = -1 + \frac{1}{n} \sum_{i=1}^n (y_{i+1} + y_i) = -1 + \frac{1}{200} \sum_{i=1}^{200} (N_{i+1}\% + N_i\%) = -1 + \frac{335.10}{200} \cong 0.675 \quad (67.17)$$

3. PERT PROBABILISTE

En restant toujours dans le cadre des statistiques et probabilités relativement aux techniques mathématiques de gestion, il existe une loi empirique en gestion de projet (domaine que nous supposons connu par le lecteur car trivial) très utilisée dans le cadre des "**PERT probabilistes**" (Program Evaluation and Review Technic) et incluse dans nombres de logiciels spécialisés afin de faire des prédictions de coûts et de temps (cf. chapitre de Théorie Des Graphes). Nous parlons alors de "**Probabilistic Network Evaluation Technic**" (P.N.E.T.).

Cette loi, appelée "**loi bêta**" ou encore "**loi de Pert**" est souvent présentée sous la forme suivante dans les ouvrages et sans démonstration (...):

$$t_R(T_i) = \frac{t_O(T_i) + 4 \cdot t_V(T_i) + t_P(T_i)}{6} \quad (67.18)$$

et donne la durée probable (espérée) t_R d'une tâche élémentaire (non décomposable en sous-tâches) où nous avons t_O, t_V, t_P qui sont respectivement les durées optimiste, vraisemblable et pessimiste de la tâche T_i . Nous allons démontrer plus bas la provenance de cette relation et en quoi elle est partiellement fausse!

Remarque: Cette approche classique date de 1962 et est due à C.E. Clark. Elle serait basée sur une analyse de bases de données de tâches de projets. Il faut par contre vérifier systématiquement quel type de loi suivent les tâches dans votre entreprise et ne pas appliquer bêtement et simplement la loi bêta.

Ses principes sont les suivants:

La durée de chaque tâche du projet est considérée comme aléatoire et la distribution Bêta est systématiquement utilisée; les paramètres de cette loi que nous allons démontrer sont déterminés moyennant une hypothèse de calcul assez forte, à partir des valeurs extrêmes a et b que la durée d'exécution peut prendre, et du mode M_0 .

Il suffit donc de poser les trois questions suivantes:

1. Quelle est la durée minimale physiquement possible?
2. Quelle est la durée maximale au pire?

pour obtenir respectivement les paramètres a, b, M_0 , qui permettent ensuite de calculer la moyenne (espérance) et la variance de cette durée aléatoire.

Ensuite, nous déterminons le chemin critique du projet (par la méthode des potentiels métra supposée connue par le lecteur), en se plaçant en univers certain et en utilisant les durées moyennes (espérées) obtenues avec la loi Bêta, ce qui permet de trouver le(s) chemin(s) critique(s).

Ensuite, nous nous plaçons en univers aléatoire et la durée du projet est considérée comme la somme des durées des tâches du chemin critique précédemment identifié. Nous utilisons alors le théorème de la limite centrale vu dans le chapitre de Statistiques (rappelons que ce théorème établit, sous des conditions généralement respectées, que la variable aléatoire constituée par une somme de n variables aléatoires indépendantes suit approximativement une loi Normale, quelles que soient les lois d'origine, dès que n est assez grand) pour approximer la loi de distribution de probabilités de la durée d'exécution du projet.

L'espérance mathématique (ainsi que la variance) de cette loi Normale se calcule comme la somme des espérances mathématiques (ou des variances) de chaque durée des tâches du chemin critique (cf. chapitre de Statistiques) tel que:

$$E\left(\sum_{i=1}^k X_i\right) = \sum_{i=1}^k E(X_i) \quad (67.19)$$

et dans le cas particulier où les variables sont linéairement indépendantes, la covariance étant nulle (cf. chapitre de Statistiques) nous avons aussi:

$$V\left(\sum_{i=1}^k X_i\right) = \sum_{i=1}^k V(X_i) \quad (67.20)$$

Rappelons que nous avons vu dans les chapitres de Statistiques et Calcul Différentiel Et Intégral que:

$$B(p, q) = \frac{\Gamma(p) \cdot \Gamma(q)}{\Gamma(p+q)} = \int_0^1 t^{p-1} (1-t)^{q-1} dt \quad (67.21)$$

et

$$\Gamma(x+1) = x \cdot \Gamma(x) \quad (67.22)$$

En gestion nous remarquons que souvent, les durées des tâches suivent une loi que nous appelons "**loi bêta de première espèce**" (cf. chapitre de Statistiques) donnée par:

$$P_{0,1}(x) = \frac{x^\alpha (1-x)^\gamma}{\int_0^1 t^\alpha (1-t)^\gamma dt} 1_{[0,1]} \quad (67.23)$$

Pour un intervalle $[a, b]$ quelconque dans lequel x est compris, nous obtenons la forme plus générale:

$$P_{a,b}(x) = \frac{(x-a)^\alpha (b-x)^\gamma}{(b-a)^{\alpha+\gamma+1} \int_0^1 t^\alpha (1-t)^\gamma dt} 1_{[a,b]} \quad (67.24)$$

Vérifions que nous ayons bien:

$$\int_a^b P_{a,b}(x) dx = \frac{1}{(b-a)^{\alpha+\gamma+1} \int_0^1 t^\alpha (1-t)^\gamma dt} \int_a^b (x-a)^\alpha (b-x)^\gamma dx = 1 \quad (67.25)$$

Par le changement de variable:

$$u = \frac{x-a}{b-a} \Rightarrow du = \frac{1}{b-a} dx \text{ et } 1-u = 1 - \frac{x-a}{b-a} = \frac{b-x}{b-a} \quad (67.26)$$

nous obtenons:

$$\begin{aligned} \int_a^b P_{a,b}(x) dx &= \frac{\int_0^1 ((b-a)u)^\alpha (b-x)^\gamma (b-a) du}{(b-a)^{\alpha+\gamma+1} \int_0^1 t^\alpha (1-t)^\gamma dt} = \frac{\int_0^1 u^\alpha (b-x)^\gamma du}{(b-a)^\gamma \int_0^1 t^\alpha (1-t)^\gamma dt} = \\ &= \frac{\int_0^1 u^\alpha \left(\frac{b-x}{b-a} \right)^\gamma du}{\int_0^1 t^\alpha (1-t)^\gamma dt} = \frac{1}{\int_0^1 t^\alpha (1-t)^\gamma dt} \int_0^1 u^\alpha (1-u)^\gamma du = 1 \end{aligned} \quad (67.27)$$

Déterminons maintenant l'espérance:

$$\mu = \int_a^b x \cdot P_{a,b}(x) dx = \frac{1}{(b-a)^{\alpha+\gamma+1} \int_0^1 t^\alpha (1-t)^\gamma dt} \int_a^b x \cdot (x-a)^\alpha (b-x)^\gamma dx \quad (67.28)$$

Toujours avec le même changement de variable, nous obtenons:

$$\begin{aligned} \mu &= \int_a^b x \cdot P_{a,b}(x) dx = \frac{1}{\int_0^1 t^\alpha (1-t)^\gamma dt} \int_0^1 (a + u(b-a)) \cdot u^\alpha (1-u)^\gamma du \\ &= a + \frac{(b-a)}{\int_0^1 t^\alpha (1-t)^\gamma dt} \int_0^1 u^{\alpha+1} (1-u)^\gamma du = a + (b-a) \cdot \frac{B(\alpha+2, \gamma+1)}{B(\alpha+1, \gamma+1)} \end{aligned} \quad (67.29)$$

Or:

$$\begin{aligned} \frac{B(\alpha+2, \gamma+1)}{B(\alpha+1, \gamma+1)} &= \frac{\Gamma(\alpha+2) \cdot \Gamma(\gamma+1)}{\Gamma(\alpha+\gamma+3)} \cdot \frac{\Gamma(\alpha+\gamma+2)}{\Gamma(\alpha+1) \cdot \Gamma(\gamma+1)} \\ &= \frac{\Gamma(\alpha+2)}{\Gamma(\alpha+\gamma+3)} \cdot \frac{\Gamma(\alpha+\gamma+2)}{\Gamma(\alpha+1)} = \frac{(\alpha+1)(\alpha)\Gamma(\alpha)}{\Gamma(\alpha+\gamma+3)} \cdot \frac{\Gamma(\alpha+\gamma+2)}{\alpha\Gamma(\alpha)} \\ &= \frac{(\alpha+1)(\alpha)\Gamma(\alpha)}{(\alpha+\gamma+2)\Gamma(\alpha+\gamma+2)} \cdot \frac{\Gamma(\alpha+\gamma+2)}{\alpha\Gamma(\alpha)} = \frac{\alpha+1}{\alpha+\gamma+2} \end{aligned} \quad (67.30)$$

Donc:

$$\mu = a + (b-a) \cdot \frac{\alpha+1}{\alpha+\gamma+2} \quad (67.31)$$

Calculons maintenant la variance en utilisant la formule d'Huyghens démontrée dans le chapitre de Statistiques:

$$V(X) = E(X^2) - E(X)^2 \quad (67.32)$$

Calculons d'abord $E(X^2)$.

$$E(X^2) = \int_a^b x^2 \cdot P_{a,b}(x) dx = \frac{1}{(b-a)^{\alpha+\gamma+1} \int_0^1 t^\alpha (1-t)^\gamma dt} \int_a^b x^2 \cdot (x-a)^\alpha (b-x)^\gamma dx \quad (67.33)$$

Toujours par le même changement de variable, nous obtenons:

$$\begin{aligned} E(X^2) &= \frac{1}{\int_0^1 t^\alpha (1-t)^\gamma dt} \int_0^1 (a^2 + u^2 (b-a)^2 + 2a \cdot u(b-a)) u^\alpha (1-u)^\gamma du \\ &= a^2 + \frac{(b-a)^2}{B(\alpha+1, \gamma+1)} B(\alpha+3, \gamma+1) + \frac{2a(b-a)}{\int_0^1 t^\alpha (1-t)^\gamma dt} \int_0^1 u^{\alpha+1} (1-u)^\gamma du \\ &\quad \underbrace{\hspace{10em}}_{= 2a(b-a) \frac{\alpha+1}{\alpha+\gamma+2} \text{ (cf. calculs précédents)}} \\ &= a^2 + 2a(b-a) \frac{\alpha+1}{\alpha+\gamma+2} + (b-a)^2 \cdot \frac{B(\alpha+3, \gamma+1)}{B(\alpha+1, \gamma+1)} \end{aligned} \quad (67.34)$$

Or:

$$\frac{B(\alpha+3, \gamma+1)}{B(\alpha+1, \gamma+1)} = \frac{\Gamma(\alpha+3) \cdot \Gamma(\gamma+1)}{\Gamma(\alpha+\gamma+4)} \cdot \frac{\Gamma(\alpha+\gamma+2)}{\Gamma(\alpha+1) \cdot \Gamma(\gamma+1)} = \frac{(\alpha+2)(\alpha+1)}{(\alpha+\gamma+3)(\alpha+\gamma+2)} \quad (67.35)$$

Donc:

$$E(X^2) = a^2 + 2a(b-a) \frac{\alpha+1}{\alpha+\gamma+2} + (b-a)^2 \cdot \frac{(\alpha+2)(\alpha+1)}{(\alpha+\gamma+3)(\alpha+\gamma+2)} \quad (67.36)$$

Pour finir:

$$\begin{aligned} V(X) &= \sigma^2 = E(X^2) - E(X)^2 = E(X^2) - \mu^2 \\ &= a^2 + 2a(b-a) \frac{\alpha+1}{\alpha+\gamma+2} + (b-a)^2 \cdot \frac{\alpha+2}{\alpha+\gamma+3} \cdot \frac{\alpha+1}{\alpha+\gamma+2} - \left(a + (b-a) \cdot \frac{\alpha+1}{\alpha+\gamma+2} \right)^2 \\ &= (b-a)^2 \cdot \frac{\alpha+1}{\alpha+\gamma+2} \cdot \left(\frac{\alpha+2}{\alpha+\gamma+3} - \frac{\alpha+1}{\alpha+\gamma+2} \right) = \frac{(b-a)^2 (\alpha+1)(\gamma+1)}{(\alpha+\gamma+2)^2 (\alpha+\gamma+3)} \end{aligned} \quad (67.37)$$

Calculons maintenant la valeur modale M_0 de cette loi de distribution (pour rappel du chapitre de Statistiques il s'agit par définition du maximum global de la fonction de distribution):

$$P_{a,b}(x) = \frac{(x-a)^\alpha (b-x)^\gamma}{(b-a)^{\alpha+\gamma+1} \int_0^1 t^\alpha (1-t)^\gamma dt} 1_{[a,b]} \quad (67.38)$$

Il suffit pour le calculer de résoudre l'équation:

$$\frac{dP_{a,b}(x)}{dx} = 0 \quad (67.39)$$

Après dérivation, nous obtenons:

$$\alpha(x-a)^{\alpha-1}(b-x)^\gamma - \gamma(x-a)^\alpha(b-x)^{\gamma-1} = 0 \quad (67.40)$$

en divisant par $(x-a)^{\alpha-1}(b-x)^{\gamma-1}$ nous avons:

$$\alpha(b-x) - \gamma(x-a) = 0 \quad (67.41)$$

c'est-à-dire:

$$x = M_0 = \frac{\alpha b + \gamma a}{\alpha + \gamma} \quad (67.42)$$

Maintenant, le lecteur aura remarqué que la valeur a est la valeur la plus petite et la b la plus grande. Entre deux il y a donc le mode M_0 .

En gestion de projets, cela correspond respectivement aux durées optimiste t_o , pessimiste t_p d'une tâche.

Ensuite, nous imposons une hypothèse assez forte:

$$\alpha = 2 + \sqrt{2}, \gamma = 2 - \sqrt{2} \text{ ou } \alpha = 2 - \sqrt{2}, \gamma = 2 + \sqrt{2} \quad (67.43)$$

Ce qui implique que nous ayons:

$$M_0 = \frac{\alpha b + \gamma a}{\alpha + \gamma} = \frac{(2 + \sqrt{2})b + (2 - \sqrt{2})a}{4} = \frac{2b + \sqrt{2}b + 2a - \sqrt{2}a}{4} = \frac{2(a+b) + \sqrt{2}(b-a)}{4} \quad (67.44)$$

ainsi que:

$$V(X) = \frac{(b-a)^2 (\alpha+1)(\gamma+1)}{(\alpha+\gamma+2)^2 (\alpha+\gamma+3)} = \frac{(b-a)^2 (4+2-2+2+1)}{7 \cdot 6^2} = \frac{(b-a)^2}{6^2} \quad (67.45)$$

Et finalement:

$$\begin{aligned}
E(X) &= a + (b-a) \cdot \frac{\alpha+1}{\alpha+\gamma+2} = \frac{6a + (b-a)(2+\sqrt{2}+1)}{6} \\
&= \frac{6a + 2b + \sqrt{2}b + b - 2a - \sqrt{2}a - a}{6} = \frac{5a + a + b + b + \sqrt{2}b + b - 2a - \sqrt{2}a - a}{6} \quad (67.46) \\
&= \frac{a+b}{6} + \frac{2a+2b+(b-a)\sqrt{2}}{6} = \frac{a+b}{6} + \frac{4M_0}{6} = \frac{a+4M_0+b}{6}
\end{aligned}$$

Ce qui nous donne sous forme synthétique:

$$\begin{aligned}
M_0 &= \frac{2(a+b) + \sqrt{2}(b-a)}{4} \\
\sigma(X) &= \frac{b-a}{6} \\
E(X) &= \frac{a+4M_0+b}{6}
\end{aligned} \quad (67.47)$$

Remarque: Les deux dernières expressions de la variance et de l'espérance sont celles que vous pouvez trouver dans n'importe quel livre de gestion de projets (sans démonstration bien sûr...) comme par exemple le fameux PMBOK. Malheureusement, il y a une erreur dans cet ouvrage de référence de gestion de projets à ma connaissance (elle a été confirmée et corrigée plus tard). Effectivement, la valeur modale y est imposée par le chef de projet alors qu'en toute rigueur elle doit être calculée avec la première des trois relations ci-dessus à partir de la durée pessimiste et de la durée optimiste de la tâche!!!

Indiquons qu'exactement les mêmes développements effectués jusqu'ici sont valables avec la formulation suivante de la loi bêta:

$$P_{a,b} = \frac{(x-a)^{\alpha-1} (x-a)^{\gamma-1}}{(b-a)^{\alpha+\gamma+1} \int_0^1 t^{\alpha-1} (1-t)^{\gamma-1}} = \frac{(x-a)^{\alpha-1} (x-a)^{\gamma-1}}{(b-a)^{\alpha+\gamma+1} B(\alpha, \gamma)} \quad (67.48)$$

cette dernière formulation étant celle disponible dans les logiciels ou tableurs comme la version française de Microsoft Excel 11.8346 par exemple en utilisant la fonction LOI.BETA() et pour sa réciproque par LOI.BETA.INVERSE().

Nous obtenons alors pour l'espérance:

$$E(X) = a + (b-a) \frac{B(\alpha+1, \gamma)}{B(\alpha, \gamma)} = a + (b-a) \frac{\alpha}{\alpha+\gamma} \quad (67.49)$$

et pour la variance:

$$V(X) = \frac{(b-a)^2 \alpha \gamma}{(\alpha+\gamma)^2 (\alpha+\gamma+1)} \quad (67.50)$$

et le mode:

$$M_0 = \frac{\alpha b - b + \gamma a - a}{\alpha + \gamma - 2} \quad (67.51)$$

Avec:

$$\alpha = 3 + \sqrt{2}, \gamma = 3 - \sqrt{2} \quad (67.52)$$

Sinon pour le reste, les résultats (expressions) restent exactement les mêmes!

Exemple:

Tracé de la fonction de distribution et répartition pour la fonction de bêta de paramètres $(t_0, t_V, t_P) = (0, 1, 3)$:

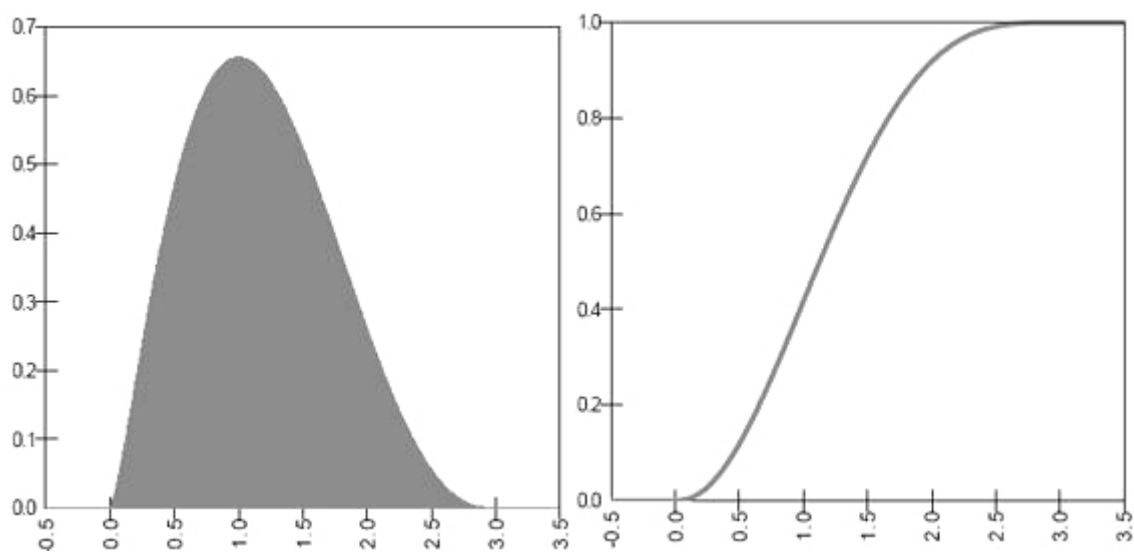


Figure: 67.7 - Fonction de distribution et de répartition de la loi bêta

Nous définissons aussi le "**risque d'action**" par le rapport (dont l'interprétation est laissée aux responsables de projet et aux clients...):

$$R(T_i) = \frac{t_P(T_i) - t_0(T_i)}{t_P(T_i)} \quad (67.53)$$

4. PROCESSUS SIX SIGMA (LEAN)

Comme nous l'avons vu dans le chapitre de Statistiques, la loi Normale est une fonction en forme de "cloche" donnée par:

$$P(k, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(k-\mu)^2}{2\sigma^2}} \quad (67.54)$$

dont les écarts-types sont utilisés pour donner l'intervalle de probabilité cumulée de se situer dans ces bornes centrées sur la moyenne comme représenté ci-dessous:

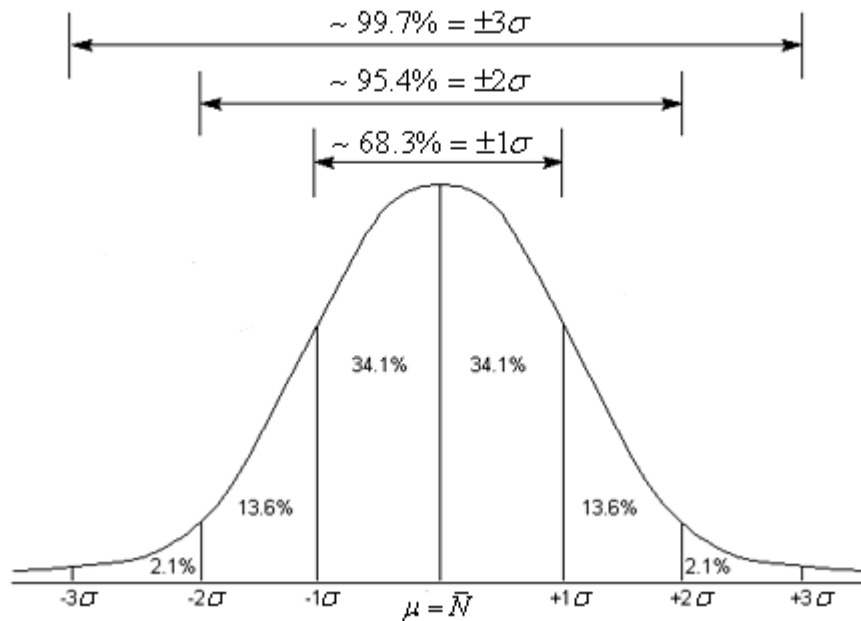


Figure: 67.8 - Rappel des intervalles sigma de la loi Normale

Ceci étant rappelé, nous avons également présenté dans le chapitre de Génie Industriel les probabilités conjointes dans le cadre de Six Sigma pour une chaîne de processus P connectés en série.

Au fait les processus mentionnés ne sont pas forcément des processus industriels mais peuvent être assimilés sous des hypothèses identiques à des processus quelconques (administratifs, procédures, workflows, etc.).

Nous avons vu que la probabilité conjointe (ou cumulée) est appelée dans Six Sigma "**Rolled Troughput Yield**" (R.T.Y.) ou "**Rendement Global Combiné**" (R.G.C.) et est donnée par (cf. chapitre de Probabilités) :

$$RTY = P\left(\bigcap_i P_i\right) = \prod_i P_i \quad (67.55)$$

Par exemple, l'application de la relation précédente donne pour un processus série en 4 étapes dont la fiabilité est de 90% chaque (sur des structures plus complexes, rappelons que nous parlons parfois "d'arbres de probabilités pondérés"):

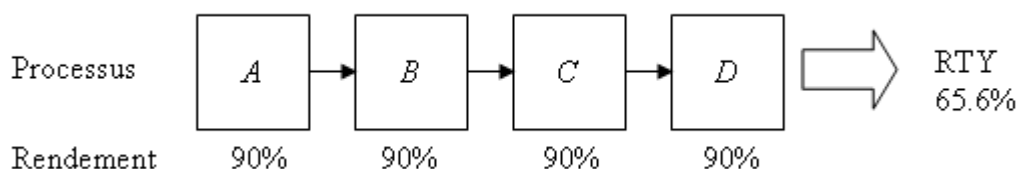


Figure: 67.9 - Processus série

Nous nous retrouvons donc au final avec une fiabilité de 65.6% soit une probabilité cumulée de défaut pour l'ensemble du processus de 34.4%.

Remarque: Le lecteur attentif aura remarqué que le système série est toujours moins fiable que sa composante la moins fiable!!

Redonnons le tableau au pire selon Six Sigma, soit le tableau en procédé non centré avec une déviation de la moyenne de $+1.5\sigma$ (donc à droite mais on pourrait prendre à gauche et les résultats sont les mêmes) par rapport à la cible et d'écart-type unitaire avec USL et LSL symétriques (ce qui restreint toujours le champ d'application):

C_p	C_{pk}	Défauts (PPM)	Niveau de qualité Sigma	Critère
0.5	0	501350	1.5	Mauvais
0.6	0.1	382572	1.8	
0.7	0.2	274412	2.1	
0.8	0.3	184108	2.4	
0.9	0.4	115083	2.7	
1	0.5	66810	3	
1.1	0.6	35931	3.3	
1.2	0.7	17865	3.6	
1.3	0.8	8198	3.9	Limite
1.4	0.9	3467	4.2	
1.5	1	1350	4.5	
1.6	1.1	483	4.8	
1.7	1.2	159	5.1	
1.8	1.3	48	5.4	
1.9	1.4	13	5.7	
2	1.5	3.4	6	Excellent

Tableau: 67.4 - Défauts et niveau de qualité Sigma en procédé décentré

où nous avons démontré dans le chapitre de Génie Industriel que les valeurs PPM étaient données par:

$$\begin{aligned}
 PPM_{Tot} &= \left(1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{USL} e^{-\frac{(x-\mu)^2}{2}} dx \right) \cdot 10^6 + \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{LSL} e^{-\frac{(x-\mu)^2}{2}} dx \right) \cdot 10^6 \\
 &= \left(1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{3C_p} e^{-\frac{(x-1.5\sigma)^2}{2}} dx \right) \cdot 10^6 + \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-3(C_p+1)} e^{-\frac{(x-1.5\sigma)^2}{2}} dx \right) \cdot 10^6 \quad (67.56) \\
 &= \left(1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{3C_p} e^{-\frac{(x-1.5)^2}{2}} dx \right) \cdot 10^6 + \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-3(C_p+1)} e^{-\frac{(x-1.5)^2}{2}} dx \right) \cdot 10^6
 \end{aligned}$$

Ce qui donne pour un niveau de qualité de 3 sigmas:

```
>evalf((1-1/sqrt(2*Pi)*int(exp(-(x-1.5)^2/2),x=-infinity..(1*3))))*1E6+evalf((1/sqrt(2*Pi)*int(exp(-(x-1.5)^2/2),x=-infinity..(-3*(1+1)))))*1E6;
```

soit une valeur de $\sim 66'810$ c'est-à-dire la valeur de la 6ème ligne. Ce qui correspond à $\sim 6.68\%$ en termes de probabilité cumulée de non-qualité ($66'810$ divisé par 1 million et mis en pourcents) et donc respectivement à une probabilité cumulée de $\sim 93.32\%$ de qualité.

Nous avons le tableau suivant qui peut résumer certaines valeurs importantes du tableau précédent en utilisant la commande Maple 4.00b donnée précédemment:

	$\pm 3\sigma$	$\pm 4\sigma$	$\pm 5\sigma$	$\pm 6\sigma$
Qualité%	93.32	99.38	99.98	99.9996
C_p	1	1.33	1.68	2
Jugement	Mauvais	Limite	Bon	Excellent

Tableau: 67.5 - Valeurs importantes de maîtrise du procédé

Sous l'hypothèse que chaque étape d'un processus série suit la même loi avec les mêmes moments et les mêmes déviations par rapport à la cible nous avons alors pour le RTY:

Étapes/ Qualité%	$\pm 3\sigma$	$\pm 4\sigma$	$\pm 5\sigma$	$\pm 6\sigma$
1	93.32	99.38	99.98	99.9996
7	61.63	95.73	99.84	99.9976
10	50.08	93.96	99.77	99.9966
20	25.08	88.29	99.54	99.9932
40	6.29	77.94	99.07	99.9864
60	1.58	68.81	98.61	99.9796
80	0.40	60.75	98.16	99.9728
100	0.10	53.64	97.70	99.966
...	...	...	...	...

Tableau: 67.6 - Niveau de qualité d'un processus/procédé ayant un nombre fini d'étapes

où chaque ligne représente le Rolled Troughput Yield (RTY) donc calculé avec:

$$RTY = P\left(\bigcap_i P_i\right) = \prod_i P_i \quad (67.57)$$

où i est le nombre d'étapes du processus (1, 7, 10, 20, 40) et où les probabilités sont données par la première ligne (1). Ainsi, avec un niveau de qualité de $\pm 3\sigma$ et une déviation à la cible de $\pm 1.5\sigma$ nous avons pour un processus de 20 étapes identiquement distribuées:

$$RTY = \prod_i P_i = \prod_{20} 93.32\% = 25.08\% \quad (67.58)$$

ce qui est lamentable! Un procédé industriel en 20 étapes d'un produit complet devrait avoir au minimum un niveau de qualité de 4 sigmas pour être acceptable en termes de rendement et de qualité

Ainsi, l'objectif du Lean Six Sigma dans une entreprise sera d'augmenter le niveau de qualité avec un *RTY* maximum pour un nombre donné d'étapes d'un processus.

5. MODÈLE STATISTIQUE DE CONTRÔLE DES SALAIRES

Basé sur les us et coutumes en statistiques, le contrôle statistique des salaires (C.S.S.) peut se justifier puisque l'outil et le principe de base est appliqué dans tous les domaines de l'économie (finance, médecine, ingénierie, qualité, biologie, projets, logistique, commerce, etc.)

L'idée empirique, fort simple, est d'analyser la distribution statistique des salaires dans une population (nation, entreprise, département, niveau hiérarchique ou autre...) et de considérer tout ce qui est au-delà de 99.9996% de probabilité cumulée comme une anomalie. Au contraire, ce qui est à l'intérieur de l'intervalle sera justifié conformément aux standards d'usage et ne nécessitera donc ni débat, ni explications hasardeuses.

Déjà, une première règle empirique évidente, est que la variable aléatoire des salaires doit être catégorisée si sa distribution statistique est multimodale. Ce qui est fréquent dans certaines entreprises ou institutions. Il convient ensuite de choisir adéquatement les catégories qui seront fréquemment des niveaux hiérarchiques normalisés au sein de l'institution (ressources, responsables d'équipes, responsables de projets, directeurs de projets, direction...).

Une fois ceci fait, il doit rester une distribution statistique unimodale dont la variable aléatoire est strictement positive (les salaires négatifs étant rares...).

Par exemple, pour l'ensemble des salaires bruts annuels fixes en fin de carrière dans le canton de Genève (Suisse) en 2004 (source: Office Cantonal Genevois du personnel), nous avons la variable aléatoire qui suit une loi log-logistique de paramètres:

$$\begin{aligned} E &= 109'157. - \text{CHF} \\ \sigma &= 54'238. - \text{CHF} \\ M_e &= 96'338. - \text{CHF} \\ Mod &= 85'224. - \text{CHF} \end{aligned} \quad (67.59)$$

Nous avons ainsi:

Probabilité cumulée (ou centile)	Valeur
99%	300'000
99.9%	535'000
99.99%	930'000
99.999%	1'600'000
99.9996%	2'030'000
99.9999%	2'850'000
99.99999%	5'000'000
99.999999%	8'800'000
99.9999998%	13'040'000

Tableau: 67.7 - Centile de la répartition des salaires avec valeur correspondante

Ainsi, si nous nous basons sur les us et coutumes en statistiques de la gestion de portefeuilles et du risque dans les bons instituts financiers, un salaire annuel de plus de 300'000.- CHF à Genève, tous secteurs et hiérarchies confondus, est une anomalie...

Si nous nous basons sur les us et coutumes en statistiques de la gestion du risque et de la qualité dans l'industrie des procédés de fabrication massive, un salaire annuel de plus de 2'030'000.- CHF à Genève, tous secteurs et hiérarchies confondus, est une anomalie...

Si nous nous basons sur les us et coutumes des statistiques les plus pointues en termes de qualité et du contrôle du risque, un salaire annuel de plus de ~13'040'000.- CHF à Genève, tous secteurs et hiérarchies confondus, est une anomalie...

Ainsi, il serait contradictoire que les personnes dirigeantes qui imposent des critères statistiques à leurs collaborateurs en ce qui concerne les services et la production ne les respectent pas eux-mêmes en ce qui concerne leur salaire ou n'impose pas des règles identiques au département des ressources humaines (dont le concept de qualité statistique est souvent inconnu...).

Remarque: Le même raisonnement statistique peut s'appliquer pour la part variable du salaire.

6. GESTION DES STOCKS

L'enjeu de la gestion des stocks et approvisionnement est important: mettre en place des processus qui optimisent la fonction économique, sous contrainte d'une disponibilité en théorie sans faille. Tels sont les objectifs du gestionnaire de stocks. Cela suppose de disposer d'une visibilité sur ses stocks et des méthodologies appropriées aux différentes situations.

Une production sans stock est quasi inconcevable vu les nombreuses fonctions que remplissent les stocks. En effet, la constitution de stocks est nécessaire s'il y a:

- Non-coïncidence dans le temps ou l'espace de la production et de la consommation: le stock est indispensable dans ce cas, car il est impossible de produire là et quand la demande se manifeste. Les exemples classiques sont la fabrication de jouets ou la confiserie pour la non-coïncidence dans le temps, et les supermarchés pour la non-coïncidence dans l'espace.
- Incertitude sur le niveau de demande ou sur le prix: s'il y a incertitude sur la quantité demandée, on va constituer un stock de sécurité qui permet de faire face à une pointe de demande (en prenant soin d'éviter l'effet "**coup de fouet**"). S'il y a incertitude sur le prix, on va constituer un stock de spéculation. Par exemple, les compagnies pétrolières achètent plus que nécessaire en pétrole brut lorsque le prix de celui-ci est relativement bas sur le marché.
- Risque de problèmes en chaîne: il s'agit ici d'éviter qu'une panne à un poste ne se répercute sur toute la chaîne d'approvisionnement. Un retard d'exécution au poste précédent ou une grève des transports n'arrêtera pas immédiatement l'ensemble du processus de production s'il y a des stocks tampons.
- Présence de coûts de lancement: dans ce cas, travailler par lots permet une économie d'échelle sur les coûts de lancement de production mais, en revanche, provoque une augmentation des coûts de possession du stock.

Le contrôle du stock et de l'approvisionnement d'une entreprise est donc aussi fondamental dans la vie de celle-ci. Afin de réduire (ou optimiser c'est selon...) un maximum les coûts divers qui tournent autour du stockage, il faut faire encore une fois appel à des connaissances en statistiques mathématiques comme nous allons le voir de suite.

Remarque: L'application de ce type d'outils ne s'adressent pas vraiment aux P.M.E. de moins de 50 employés produisant de petites pièces de manière irrégulière mais plutôt à des multinationales produisant en énorme quantité ou en faible quantité des objets de consommation de taille non négligeable et de manière régulière. Par ailleurs comme auteur de ces pages, je me suis renseigné dans de nombreuses entreprises (dont des multinationales!) et je n'ai trouvé encore aucun logisticien utilisant dans la pratique les modèles mathématiques qui vont être présentés ci-après.

Dans un premier temps, nous allons établir comment déterminer le stock initial nécessaire à une entreprise en se basant sur des données statistiques et ce à partir de modèles simples ensuite de quoi nous ferons de même avec les modèles de réapprovisionnement dont la démarche d'approche est un peu différente et permet comme pour la première d'arriver à des résultats très satisfaisants à grande échelle.

Les modèles que nous allons construire permettront ainsi:

1. De réguler les aléas des flux de fournitures
2. De permettre la production par lots (réduit les coûts de production)
3. De faire face à des demandes saisonnières

Des stocks supplémentaires pouvant engranger des "**coûts d'intérêt**" (capital immobilisé), des "**coûts d'obsolescence**" (les produits deviennent entre temps obsolètes), des "**coûts de stockage**", des "**coûts d'assurances**" (protection contre les accidents pouvant subvenir sur les produits) et de nombreux autres...

Nous distinguons dans le domaine classiquement trois types de stocks:

1. Le "**stock minimum**", appelé encore "**stock tampon**" ou "**stock d'alarme**" ou "**point de commande**" ou "**seuil de réapprovisionnement**" est la quantité de stock à partir de laquelle on lance une nouvelle commande pour augmenter les réserves.
2. Le "**stock de sécurité**" qui permet de répondre aux aléas les plus fréquents liés à la consommation et à la livraison.
3. Le "**stock d'alerte**", appelé encore "**stock critique**" qui est le niveau de stock pour lequel on déclenche une commande au risque de connaître une rupture. Par construction le stock d'alerte est donc la somme du stock de sécurité et du stock minimum.

On considère souvent trois stratégies de gestion de stocks:

1. La "**gestion de stock par point de commande**": l'approvisionnement du stock est déclenché lorsque l'on observe que le stock descend en-dessous du point de commande.
2. La "**gestion calendaire**": l'approvisionnement du stock est déclenché à intervalles réguliers T , par exemple, chaque jour ou chaque semaine.

3. La "**gestion calendaire conditionnelle**": l'approvisionnement du stock est déclenché à intervalles réguliers T , mais uniquement lorsque l'on observe que le stock descend en-dessous du point de commande.

Un stock est constitué pour satisfaire une demande future. En cas de demande aléatoire, il peut y avoir non coïncidence entre la demande et le stock. Deux cas sont évidemment possibles:

1. Une demande supérieure au stock: nous parlons alors de "**rupture de stock**"
2. Une demande inférieure au stock: nous parlons alors de "**stock résiduel**"

Le critère de gestion généralement retenu en gestion des stocks est celui de la minimisation des coûts. Nous noterons cette fonction par la lettre C , suivie, entre parenthèses (ou en indice), de la ou des variables de commande du système.

6.1. STOCKS EN AVENIR INCERTAIN

Commençons notre étude par le cas le plus simple qui suppose que la consommation est statistiquement régulière dans le temps et sous contrôle. Il s'ensuit (cf. chapitres de Statistiques et de Génie Industriel) que la consommation périodique suit alors une loi de Gauss.

Prenons un exemple concret puisque la théorie a déjà été étudiée en long et en large dans le chapitre de Statistiques.

Considérons un article dont la demande quotidienne suit approximativement une loi normale de paramètres:

$$\mu = 50, \sigma = 20 \quad (67.60)$$

Le stock disponible au moment de la commande est de 500 unités et le délai de réapprovisionnement de 5 jours. Nous souhaiterions savoir quelle est la probabilité cumulée d'être au-dessus ou égal à la rupture de stock ainsi que la probabilité cumulée d'être au-dessus de la consommation quotidienne supposée?

Nous avons pour le premier point la consommation moyenne sur 5 jours qui est en utilisant la propriété de stabilité de la loi normale:

$$\begin{aligned} \mu &= 5 \cdot 90 = 450 \\ \sigma &= \sqrt{20^2 + 20^2 + 20^2 + 20^2 + 20^2} = 20 \cdot \sqrt{5} = 44.72 \end{aligned} \quad (67.61)$$

Nous avons alors:

$$P(X \geq 450) = 13.17\% \quad (67.62)$$

soit en utilisant la version française de Microsoft Excel 11.8346:

$$=1-LOI.NORMALE(500;450;44.72;1)=13.17\%$$

Et pour la consommation quotidienne, il vient simplement:

$$=1-LOI.NORMALE(500;450;44.72;1)=30.85\%$$

Évidemment, dans la pratique la réalité est beaucoup moins simple et de plus, la loi normale peut avoir une faible probabilité de ventes négatives, ce qui est peu réaliste... Mais c'est déjà un premier modèle qui peut être confronté à la réalité pour voir sa puissance prédictive.

6.2. STOCK INITIAL OPTIMAL EN GESTION CALENDRAIRE ET À ROTATION NULLE

Imaginons de suite un scénario afin de développer un autre modèle statistique mais un tout petit peu plus élaboré que le précédent (inspiré de l'ouvrage *Gestion de la Production* de Vincent Giard). Nous nous limiterons cependant ici à une loi statistique discrète et nous referons ensuite un exemple similaire avec une loi continue.

Considérons pour cela que l'entreprise MAC est la spécialiste d'un certain produit dont le coût direct de fabrication par unité est de 25.- et le prix de vente 60.- par unité (prix supposés constants et donc indépendants de la quantité et sans escompte).

La vente quotidienne de ce produit est, en semaine, de 2.5 en moyenne et le relevé des demandes pendant trois mois laisse supposer que celle-ci suit une loi discrète de type de Poisson (cette fois-ci le support est au moins positif contrairement à l'exemple d'avant...), c'est-à-dire que nous avons une distribution de probabilités du nombre X de ces produits au cours d'une journée. Nous tronquerons à $x = 10$ la réalisation de cette variable aléatoire car la probabilité de ventes supérieures à 10 sera supposée comme nulle.

x	$P(X)$	Probabilité cumulée
0	0.0821	0.0821
1	0.2052	0.2873
2	0.2565	0.5438
3	0.2138	0.7576
4	0.1336	0.8912
5	0.0668	0.958
6	0.0278	0.9858
7	0.0099	0.9957
8	0.0031	0.9988
9	0.0009	0.9997
10	0.0003	1

Tableau: 67.8 - Probabilités cumulées des ventes

Nous avons alors le tableau ci-dessus qui nous montre que la quantité la plus souvent vendue à un agent économique est à peu près de 2 (valeur modale) et le calcul de l'espérance nous donne pour ce tableau:

$$\bar{x} = E(X) = 2.5$$

résultat qui peut aussi bien être interprété comme la moyenne des ventes quotidiennes que comme la moyenne du niveau de rupture des stocks.

Nous supposons que le stock est à flux tendu. En d'autres termes, d'un jour à l'autre, aucune unité n'est reportée pour les ventes du lendemain car il n'est plus censé y en avoir (ce que nous appelons la "**gestion des stocks par rotation nulle**"). La question dès lors est de savoir, le tableau ci-dessus étant donné, combien de produits mettre en fabrication (ou commander) chaque jour pour maximiser le bénéfice et minimiser les pertes.

Dès lors, dans l'optique retenue de minimisation du coût de possession C_P (appelé souvent "**coût unitaire**" de possession) associé aux invendus est de 25.- par unité, tandis que le coût de rupture C_R (appelé souvent "**coût unitaire de rupture**") est égal au manque à gagner consécutif à la vente ratée, c'est-à-dire le prix de vente 60.- soustrait des 25.- soit, un coût de rupture de 35.- par unité.

Remarque: Le coût de rupture n'est pas assimilé qu'à la perte engendrée par une vente ratée dans la pratique mais aussi au coût d'une solution de remplacement temporaire proposée par le vendeur à son client en attente de fournir le bon produit. En interne dans les entreprises, le coût de rupture entraîne souvent un chômage technique des postes en aval lorsqu'une pièce de la chaîne de fabrication n'est plus disponible.

Une gestion rationnelle doit permettre de calculer le stock initial S (autrement dit le nombre de produits à commander ou à fabriquer pour la journée) qui minimise l'indicateur de "**coût de gestion**" $C(S)$ défini comme étant la somme du coût de possession associé au stock moyen des invendus $I_P(S)$ (appelé souvent "**stock moyen possédé**"), et du coût de rupture associé au stock moyen de ventes ratées $I_R(S)$ (appelé souvent "**stock moyen des invendus**"):

$$C(S) = C_P I_P(S) + C_R I_R(S) \quad (67.63)$$

Cette relation est cependant une simplification de la relation:

$$C(S) = C_P I_P(S) + C_R I_R(S) + C_C I_C(S) \quad (67.64)$$

où le dernier terme est le coût de commande.

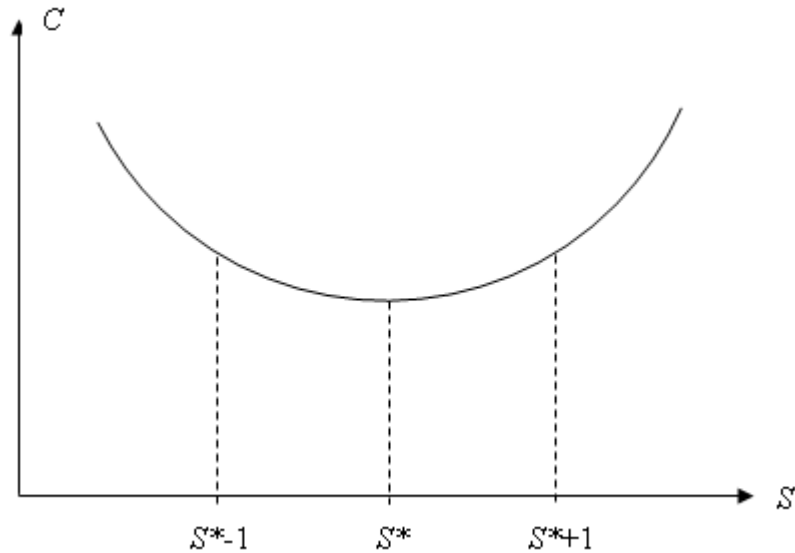
Minimiser la fonction de coût $C(S)$ revient du point de vue mathématique à chercher un extremum (minimum) de la fonction de coût de gestion tel que pour la valeur optimale S^* de l'approvisionnement initial le coût $C(S^*)$ soit immédiatement inférieur à $C(S^* \pm 1)$ c'est-à-dire inférieur par la droite et par la gauche. En d'autres termes:

$$\begin{aligned} C(S^*) &< C(S^* + 1) \\ C(S^*) &< C(S^* - 1) \end{aligned} \quad (67.65)$$

ou après réarrangement:

$$\begin{aligned} C(S^*) - C(S^* + 1) &< 0 \\ C(S^*) - C(S^* - 1) &< 0 \end{aligned} \quad (67.66)$$

inégalités que le schéma ci-dessous peut aider à mieux comprendre:

Figure: 67.10 - Exemple arbitraire d'extremum de la fonction $C(S)$

À partir de maintenant la question est de savoir comment procéder pour déterminer S^* . Au fait l'idée est subtile mais simple tant qu'elle est bien exposée et réfléchie.

Reprenons la distribution de probabilités de la loi de demande quotidienne et supposons que nous voulions calculer les ruptures moyennes (donc l'espérance) pour un niveau donné du stock initial S :

$$I_R(S = 4) = E(x - 4) \quad (67.67)$$

et:

$$I_R(S = 5) = E(x - 5) \quad (67.68)$$

associées aux stocks initiaux respectifs par rapport à la distribution donnée (bien évidemment x doit être prix comme étant supérieur à S). Soit en toute généralité:

$$I_R(S) = \sum_{x=S+1}^n (x - S) P(X) \quad (67.69)$$

Évidemment, si la loi de probabilité est continue (ou du moins approximée comme telle), cette dernière relation s'écrit:

$$I_R(S) = \int_S^{\infty} (x - S) f(x) dx \quad (67.70)$$

L'idée est d'alors d'écrire la distribution de densité de probabilité par rapport à la quantité manquante de stock et non plus vendue:

x	$P(X)$	$x-4$	$(x-4)P(X)$	$x-5$	$(x-5)P(X)$
0	0.0821	-	-	-	-
1	0.2052	-	-	-	-
2	0.2565	-	-	-	-
3	0.2138	-	-	-	-
4	0.1336	-	-	-	-
5	0.0668	1	0.0668	-	-
6	0.0278	2	0.0556	1	0.0278
7	0.0099	3	0.0297	2	0.0198
8	0.0031	4	0.0124	3	0.0093
9	0.0009	5	0.0045	4	0.0036
10	0.0003	6	0.0018	5	0.0015
Σ	1	-	$I_R(4) = 0.1708$	-	$I_R(5) = 0.062$

Tableau: 67.9 - Distribution de densité de probabilité par rapport à la quantité manquante

Il ressort du tableau précédent que le fait de faire passer le stock initial S de 4 à 5, diminue la rupture moyenne en la faisant passer de 0.1708 à 0.062. Mais de ce résultat, nous ne pouvons rien faire pour l'instant car à notre niveau actuel du développement, cela signifierait qu'en prenant un stock initial de 10, nous aurions une rupture moyenne nulle (... ce qui n'avance pas à grande chose...) et que si nous prenons aucun stock initial, nous aurions une rupture de stock totale...

Mais cependant nous pouvons tirer un résultat intermédiaire intéressant. Effectivement regardons la manière dont varie la différence de la rupture moyenne (résultat facilement généralisable - nous pouvons faire la démonstration sur demande au besoin):

$$\begin{aligned}
 I_R(S=4) - I_R(S=5) &= \sum_{x=S+1}^{\infty} (x-S)P(X=x) - \sum_{x=S+1}^{\infty} (x-S)P(X=x) \\
 &= \sum_{x=5}^{10} (x-4)P(X=x) - \sum_{x=6}^{10} (x-5)P(X=x) \\
 &= 1P(X=5) + (2-1)P(X=6) + (3-2)P(X=7) + (4-3)P(X=8) \\
 &\quad + (5-4)P(X=9) + (6-5)P(X=10) \\
 &= P(X=5) + P(X=6) + P(X=7) + P(X=8) + P(X=9) + P(X=10) \\
 &= \sum_{x=5}^{10} P(X=x) = P(X > 4)
 \end{aligned} \tag{67.71}$$

Le résultat étant identique si la loi de probabilité est continue!

Autrement dit (soyez bien attentif!!!), la diminution de rupture moyenne occasionnée en augmentant d'une unité un stock préalablement dimensionné à $S=4$, est égale à la probabilité cumulée que la demande soit strictement supérieure à celle du stock initial $P(X > 4)$.

En d'autres termes, au cas où cela ne serait pas clair, le fait d'augmenter le stock initial diminue certes la rupture moyenne, mais impose en contrepartie qu'il y ait moins d'acheteurs qui risquent de satisfaire l'offre et les seuls qui le peuvent sont ceux qui correspondent à la probabilité cumulée $P(X > 4)$.

Finalement, nous pouvons écrire:

$$I_R(S+1) - I_R(S) = -P(X > S) \quad (67.72)$$

Le tableau ci-dessous représente la diminution de rupture moyenne que nous obtenons en accroissant d'une unité le stock (et respectivement la probabilité de clients capables de consommer le stock...):

x	$P(X)$	$I_r(x) - I_r(x+1)$
0	0.0821	0.9179
1	0.2052	0.7127
2	0.2565	0.4562
3	0.2138	0.2424
4	0.1336	0.1088
5	0.0668	0.0420
6	0.0278	0.0142
7	0.0099	0.0043
8	0.0031	0.0012
9	0.0009	0.0003
10	0.0003	0

Tableau: 67.10 - Rupture moyenne en accroissant d'une unité le stock

Maintenant regardons les invendus $I_P(S)$ en tant que fonction aléatoire du stock initial S . Leur espérance est bien évidemment donnée par (servez-vous des tableaux au besoin pour comprendre):

$$I_P(S) = \sum_{x=0}^S (S-x)P(X) \quad (67.73)$$

Évidemment, si la loi de probabilité est continue (ou du moins approximée comme telle), cette dernière relation s'écrit:

$$I_P(S) = \int_0^S (S-x)f(x)dx \quad (67.74)$$

Ce que nous pouvons écrire:

$$\begin{aligned}
I_p(S) &= E(S - X) = \sum_{x=0}^S (S - x)P(X) = \sum_{x=0}^{+\infty} (S - x)P(X) - \sum_{x=S+1}^{+\infty} (S - x)P(X) \\
&= S \sum_{x=0}^{+\infty} P(X) - \sum_{x=0}^{+\infty} xP(X) - \left[- \sum_{x=S+1}^{+\infty} (x - S)P(X) \right] \\
&= S - \sum_{x=0}^{+\infty} xP(X) - \left[- \sum_{x=S+1}^{+\infty} (x - S)P(X) \right] = S - \sum_{x=0}^{+\infty} xP(X) - [-I_R(S)] \\
&= S - \sum_{x=0}^{+\infty} xP(X) + I_R(S) = S - \bar{x} + I_R(S)
\end{aligned} \tag{67.75}$$

d'où:

$$I_p(S) = S - \bar{x} + I_R(S) \tag{67.76}$$

qui est donc le stock résiduel de fin de période. C'est donc un résultat remarquable qui va nous permettre de déterminer S^* seulement à partir de $I_R(S)$. Nous pouvons aussi à nouveau considérer comme intuitif que le résultat est identique si la loi de probabilité est continue.

Cette dernière relation peut également s'écrire après réarrangement:

$$\bar{x} - I_R(S) = S - I_p(S) \tag{67.77}$$

où le terme de gauche représente la demande moyenne satisfaite et le terme de droite l'offre moyenne utilisée. Cette relation est donc une relation particulière d'équilibre entre une offre et une demande.

Nous pouvons par ailleurs vérifier cela à partir des tables ci-dessous en mettant un exemple particulier en évidence:

$$I_p(S = 4) = 4 - 2.5 + 0.1708 = 1.6708$$

x	$4 - x$	$(4 - x)P(X)$
0	4	0.3284
1	3	0.6156
2	2	0.5130
3	1	0.2138
4	0	-
5	-	-
6	-	-
7	-	-
8	-	-
9	-	-
10	-	-
		$I_P(4) = 1.6708$

Tableau: 67.11 - Espérance des invendus

Finalement nous pouvons écrire une expression de $C(S)$, fonction de la seule rupture moyenne:

$$C(S) = C_P(S - \bar{x} + I_R(S)) + C_R I_R(S) \quad (67.78)$$

ou ce qui s'écrit après réarrangement:

$$C(S) = C_P(S - \bar{x}) + (C_R + C_P) I_R(S) \quad (67.79)$$

Il s'ensuit que:

$$\begin{aligned} & C(S+1) - C(S) \\ &= [C_P(S+1) - C_P \bar{x} + (C_R + C_P) I_R(S+1)] - [C_P(S) - C_P \bar{x} + (C_R + C_P) I_R(S)] \quad (67.80) \\ &= C_P + (C_R + C_P) [I_R(S+1) - I_R(S)] \end{aligned}$$

Ce qui donne avec les résultats obtenus plus haut:

$$C(S+1) - C(S) = C_P + (C_R + C_P)(-P(X > S)) \quad (67.81)$$

Dans ces conditions, les relations:

$$\left. \begin{aligned} C(S^*) - C(S^* + 1) &< 0 \\ C(S^*) - C(S^* - 1) &< 0 \end{aligned} \right\} \Leftrightarrow \left\{ \begin{aligned} C(S^* + 1) - C(S^*) &> 0 \\ C(S^*) - C(S^* - 1) &< 0 \end{aligned} \right. \quad (67.82)$$

Deviennent:

$$\begin{cases} C_P + (C_R + C_P)(-P(X > S^*)) > 0 \\ C_P + (C_R + C_P)(-P(X > S^* - 1)) < 0 \end{cases} \quad (67.83)$$

d'où:

$$\begin{aligned} \frac{C_P}{C_P + C_R} &> P(X > S^*) \\ \frac{C_P}{C_P + C_R} &< P(X > S^* - 1) \end{aligned} \quad (67.84)$$

d'où S^* est optimal si:

$$P(X > S^*) < \frac{C_P}{C_P + C_R} < P(X > S^* - 1) \quad (67.85)$$

Dans notre exemple numérique, nous avons:

$$\frac{C_P}{C_P + C_R} = 0.417 \quad (67.86)$$

avec:

$$P(X > 2) = 0.4562, P(X > 3) = 0.2424 \quad (67.87)$$

d'où le stock optimal initial journalier appelé aussi "**stock minimum**" à mettre en production (ou à commander) pour minimiser le coût de gestion:

$$S^* = 3 \quad (67.88)$$

ce qui n'est ni 2 ni 2.5!!! Mais c'est un chiffre qui est bien évidemment très proche de la valeur modale de la distribution initiale.

Il faut noter que cependant dans la pratique, les employés qui font du controlling logistique dans les multinationales n'ont aucune idée des prix, marges et coûts qui ont lieu chez leurs clients alors que ceux-ci leur demande conseil (ils les paient même pour cela!). Alors souvent on prend simplement la valeur modale ou la valeur correspondant à une couverture de 95% de de la demande périodique (cela devient assez arbitraire). Dans le cas de notre petite exercice, la valeur modale du stock initiale arrondi à l'entier supérieur le plus proche serait 2 et la valeur à 95% de 5. Par prudence il y a cependant une tendance à prendre plutôt la valeur modale.

6.3. MODÈLES DE WILSON

Il existe plusieurs modèles d'optimisation de gestion de stocks (Statistique, Wilson, ABC, 20/80...). Parmi ceux-ci, nous avons souhaité nous arrêter sur les "**modèles de Wilson**" avec ses trois variantes qui sont les plus connues (mais pas forcément les plus réalistes...).

Nous considérons donc trois situations:

- L'entreprise se réapprovisionne par lots à l'extérieur et il n'y a alors pas d'inertie des stocks et elle relance une commande lorsque le stock de sécurité est atteint. Nous parlons alors de "**modèle de Wilson avec réapprovisionnement**".

- L'entreprise se réapprovisionne par une production interne qui implique alors une certaine inertie pour revenir au niveau de stock désiré. Nous parlons alors de "**modèle de Wilson sans réapprovisionnement**" (dont la totalité des résultats se déduit très vite du premier modèle sans refaire tous les développements).

- L'entreprise se réapprovisionne par lots à l'extérieur et il n'y a alors pas d'inertie des stocks et elle relance une commande lorsque le stock de sécurité est atteint mais avec une inertie et demande à ses clients d'attendre. Nous parlons alors de "**modèle de Wilson avec attente**".

Remarque: Ce modèle appelé également "**modèle du lot économique**" permet de déterminer la fréquence et la quantité optimale de réapprovisionnement pour un magasin, une usine... Elle est couramment employée par les services logistiques de grandes structures. Elle a en fait été introduite dès 1913...

Le but est de déterminer la stratégie qu'il faut adopter pour que le total périodique (annuel, mensuel, hebdomadaire, journalier, ...) des commandes ou fabrications de pièces minimise le total des coûts d'acquisition et de possession de stocks pour l'entreprise. Nous parlons aussi quelquefois de "**gestion à flux tendu**".

L'existence de stocks au sein de l'entreprise amène le gestionnaire à se poser la question du niveau optimal de ces derniers, en évitant deux écueils principaux:

1. Le "**surstockage**", source de coûts pour l'entreprise (coût du stockage physique, manutention, locaux et surfaces utilisées, coûts annexes, assurances de gardiennage, coût des capitaux immobilisés).
2. Le "**sous-stockage**" qui risque d'aboutir à des ruptures de stocks préjudiciables à l'activité de production ou à l'activité commerciale de l'entreprise (arrêt de la production, perte de ventes, perte de clientèle, etc.).

Ainsi, les différents modèles de gestion des stocks ont pour objectif de minimiser le coût de gestion dans ce système de contraintes en déterminant la fréquence de réapprovisionnement et la quantité associée.

Voyons d'abord une approche purement qualitative. Pour chaque référence, les quantités en stock évoluent dans le temps par exemple sous une forme:

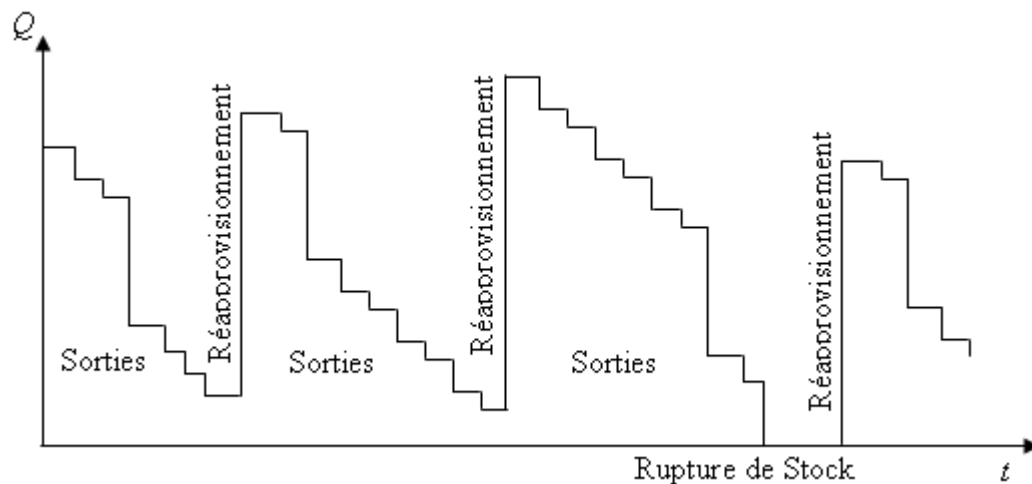


Figure: 67.11 - Première stratégie de réapprovisionnement

En simplifiant nous obtenons un graphique dit "en dents de scie":

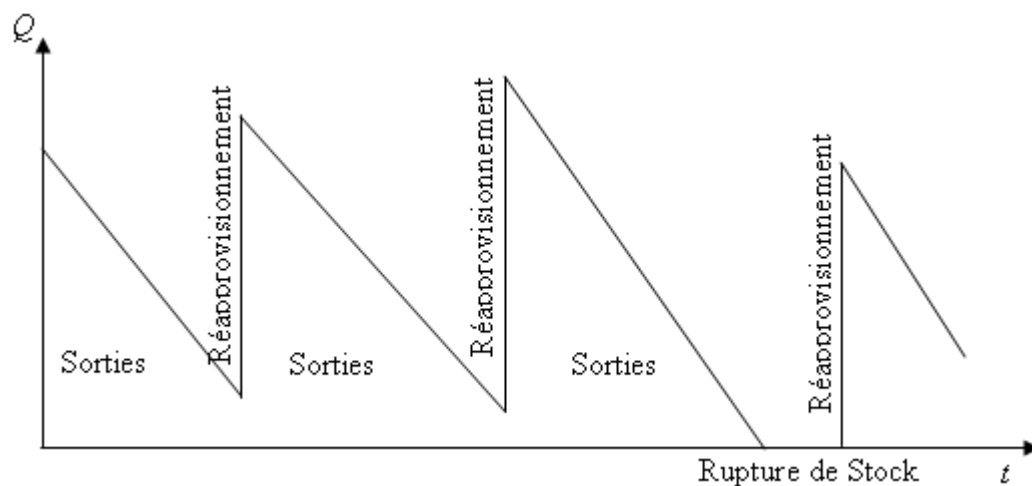


Figure: 67.12 - Diagramme simplifié correspondant en dents de scie

où nous sommes donc dans une situation typique de "**gestion de stocks à rotation non nulle**" (c'est-à-dire lorsque l'inventu peut être vendu à une période ultérieure.)

Pour éviter la rupture de stock, il faut bien évidemment faire en sorte que l'entrée d'une commande se fasse, au plus tard, lorsque la quantité en stock devient nulle:

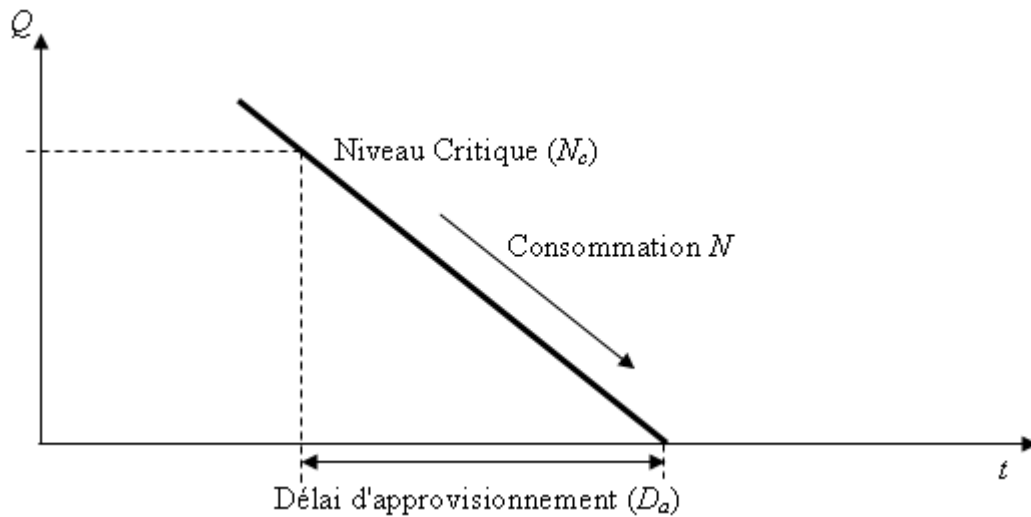


Figure: 67.13 - Principe visuel de la stratégie

Si nous considérons une consommation constante d'une quantité N par unité de temps (jour, mois, année,...) et que nous connaissons à l'avance le délai d'approvisionnement (en jour, mois, année...) alors si le tout est mis en des unités équivalentes (journalières par exemple) nous avons le niveau critique qui est donné par:

$$N_c = N \cdot D_a \quad (67.89)$$

qui est aussi assimilé à la terminologie justifiée de "**point de commande**" puisqu'il s'agit de la quantité que nous avons en stock à partir de laquelle il faut lancer une nouvelle commande d'approvisionnement:

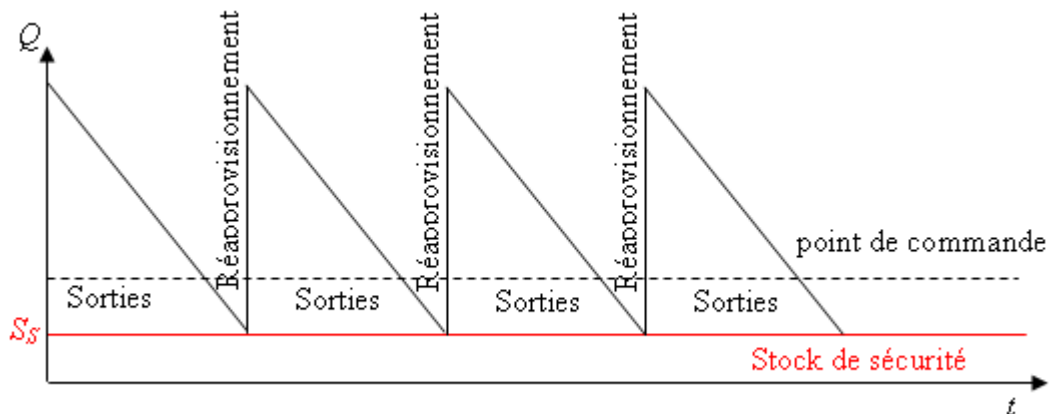


Figure: 67.14 - Représentation du point de commande (diagramme en dents de scie)

Ainsi, si la consommation est de 10 unités par jour et que le délai d'approvisionnement est de 15 jours, le niveau critique est alors de 150 unités.

Pour éviter les aléas (grève, transports, variation de consommation, remplacements,...) nous envisageons un stock de sécurité S_s :

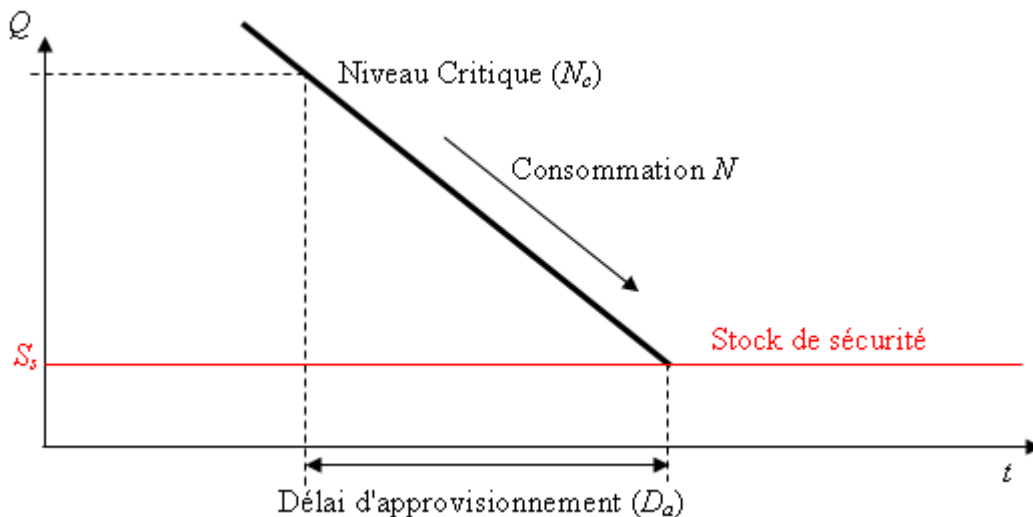


Figure: 67.15 - Représentation du stock de sécurité sur le diagramme en dents de scie

Nous avons alors pour le niveau critique:

$$N_c = S_s + N \cdot D_a \quad (67.90)$$

Voyons maintenant l'influence du nombre de livraisons sur le coût de stockage (puisque plus le taux de détention est gros, plus les coûts de stockage sont élevés). Supposons pour cela que le marché consomme 100 unités par mois et ce de manière régulière. Dans le cas d'un seul approvisionnement annuel, la consommation est représentée par le graphique ci-dessous (aucun stock de sécurité afin de simplifier l'exemple):

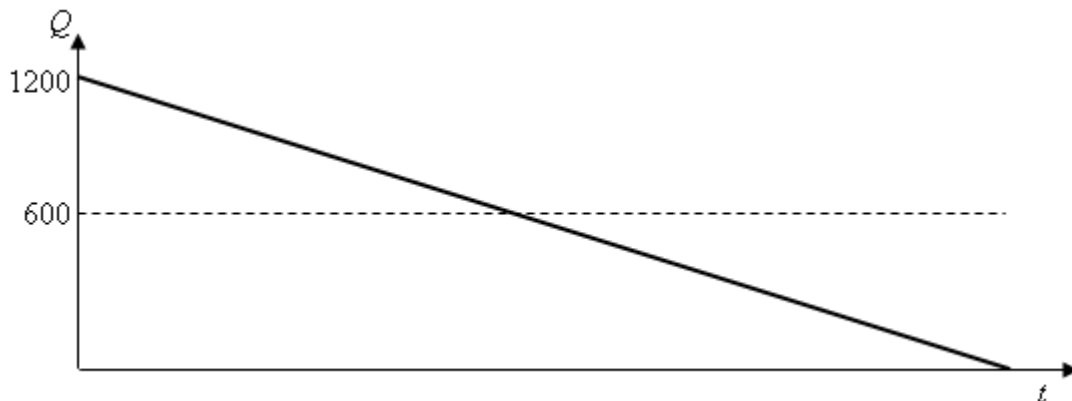


Figure: 67.16 - Représentation de l'effet d'un seul approvisionnement

où nous voyons immédiatement que le stock moyen est de 600. Ce stock moyen est obtenu par simple moyenne arithmétique ou simplement en utilisant la définition de la moyenne intégrale de la fonction de consommation:

$$\mu_f = \frac{1}{b-a} \int_a^b f(t) dt = \frac{1}{12-0} \int_0^{12} (-100 \cdot t + 1200) dt = 600 \quad (67.91)$$

Soit trivialement la moitié du stock maximum (ou autrement vu: la surface du triangle rectangle divisé par T ...).

Et si nous divisons la gestion en deux approvisionnements, nous avons alors:

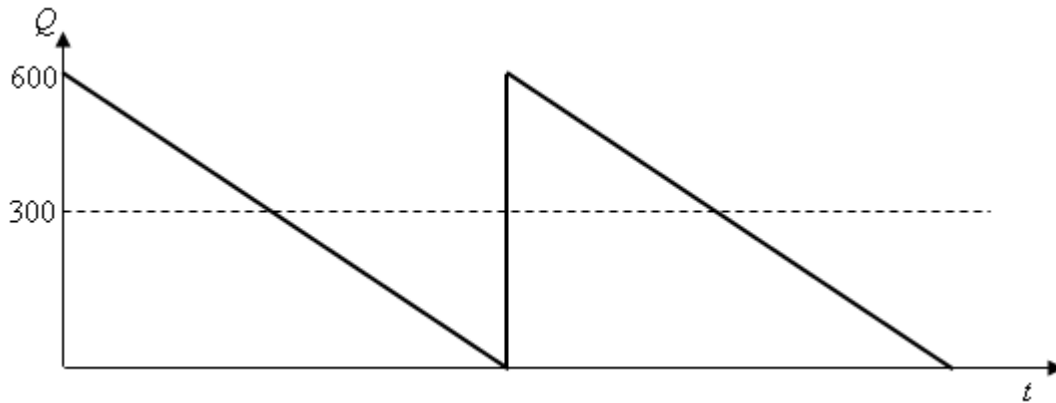


Figure: 67.17 - Représentation de l'effet de deux approvisionnements

soit un stock moyen deux fois inférieur et dès lors un coût de stockage moyen deux fois moindre. Mais bien évidemment il faut y associer le coût d'approvisionnement. C'est à ce niveau de complexité qu'intervient justement la formalisation mathématique de Wilson.

Chaque commande d'achat ou ordre de fabrication coûte donc à l'entreprise. Le "**coût de lancement**" ou "**coût de passation**" des commandes ou lancements de fabrications représente tous les frais liés (administratifs, réglages machines, préparation, communications,...) au fait de passer une commande (ou une fabrication) et est supposé être proportionnel à la quantité. Ces coûts sont déterminés à l'aide de la comptabilité analytique.

Ainsi, le coût d'une commande est obtenu en divisant le coût total de fonctionnement du service achat par une grandeur significative et pertinente. Par exemple, le nombre de commandes passées (ou ordres de fabrications) annuellement par exemple. Le coût d'un lancement en fabrication lui sera obtenu en divisant le coût total de fonctionnement du service ordonnancement, auquel, il faut, ajouter les coûts de réglage des machines et des préséries, par le nombre de lancements de fabrication.

Ces valeurs dépendent essentiellement de l'entreprise, de ses choix en matière de comptabilité analytique. Il est difficile de définir une fourchette de valeurs standards. Bon nombre d'entreprises ne savent pas à combien leur revient une commande ou un lancement de fabrication (et bon nombre ne savent tout simplement pas faire une analyse...).

Le coût de possession du stock est constitué des charges liées au stockage physique mais également de la non-rémunération des capitaux immobilisés dans le stock (voire du coût des capitaux empruntés pour financer le stock). Pour cette dernière raison, ce coût est considéré comme étant proportionnel à la valeur du stock moyen et à la durée de détention de ce stock.

Le taux de possession annuel $t\%$ est le coût de possession ramené à une unité monétaire de matériel stocké. Il est obtenu en divisant le coût total des frais annuels de possession par le stock moyen annuel.

Ces frais couvrent l'intérêt du capital immobilisé, les coûts de magasinage (loyer et entretien des locaux, assurances, frais de personnel et de manutention, gardiennage, etc.), les détériorations du matériel, les risques d'obsolescence.

Ce taux oscille habituellement entre 15 et 35% de la valeur marchande stockée dans les entreprises, suivant le type des articles et la qualité de leur gestion des stocks.

Wilson a établi une relation basée sur un modèle mathématique simplificateur dans lequel nous considérons que la demande est stable sans tenir compte des évolutions de prix, des risques de rupture et des variations dans le temps des coûts de commande et de lancement (nous sommes alors en "avenir certain").

6.3.1. MODÈLE DE WILSON AVEC RÉAPPROVISIONNEMENT

Les hypothèses très simplificatrices de ce modèle sont les suivantes:

- H1. La demande périodique est connue et certaine (déterministe)
- H2. Les quantités commandées sont constantes à chaque période
- H3. La pénurie (ruptures de stock) a lieu en fin de période (donc il n'y a jamais de défaut de stock)
- H4. Le délai de production est constant et l'approvisionnement supposé instantané
- H5. Les coûts (stocks, articles, passation,...) sont invariables dans le temps
- H7. Le coût de possession est proportionnel à la valeur
- H8. L'horizon de planification est infini

et selon les auteurs cette liste d'hypothèses varie plus ou moins (certaines hypothèses étant implicites ou relativement triviales).

Remarques:

- R1. Nous supposons que la gestion du stock s'effectue sur une période temporelle donnée.
- R2. D'après ces hypothèses, nous concluons qu'il y aura le même niveau de commande à lancer chaque fois, que le coût total de pénurie est nul.

Nous noterons:

- N la quantité correspondant à une demande ou respectivement à des pièces consommées par période.
- Q la quantité d'approvisionnements ou de pièces lancées en fabrication en une seule fois pendant ce même temps (taille des lots).
- P_u le prix unitaire d'achat de la pièce (supposé constant).
- S_s le stock de sécurité envisagé pour cette pièce (supposé constant) pour répondre aux aléas par période.
- t le taux de coût de possession en % (supposé constant) et parfois appelé "taux de détention".
- C_L le coût d'approvisionnement/acquisition par commande ou de lancement de fabrication.

Nous définissons de par la même occasion, le "**coût unitaire de stockage**" calculé souvent (mais pas uniquement!) sur la base du prix unitaire d'achat d'une pièce:

$$C_s = t\% \cdot P_u \quad (67.92)$$

Propositions:

P1. Le rapport:

$$\frac{N}{Q} \quad (67.93)$$

donne "**l'inertie des stocks**" ou qui peut être vu de manière plus explicite comme étant le "**nombre périodique de lançements**" (ou la cadence de réapprovisionnement) pour satisfaire la demande. L'inertie des stocks a donc la dimension d'un inverse du temps.

P2. Le "**coût d'inertie**" ou respectivement le "**coût d'acquisition**", ou encore "**coût de lancement**" est donc par unité de période de:

$$\frac{N}{Q} \cdot C_L \quad (67.94)$$

Ce dernier est donc supposé proportionnel à la consommation! Ce qui est important ceci dit est de remarquer que le coût de lancement est inversement proportionnel à la quantité Q et donc qu'il tend vers zéro lorsque Q tend vers l'infini. Ceci dit, normalement on aura dans la majorité des cas théoriques:

$$Q \leq N \quad (67.95)$$

P3. Le stock moyen dans l'entreprise dans l'hypothèse d'une consommation (décroissance linéaire du stock) et d'un niveau de sécurité constants dans le temps est trivialement par période:

$$S_s + \frac{Q}{2} \quad (67.96)$$

Le "**coût périodique de possession**", appelé encore "**coût de possession**" ou "**coût de gestion**" ou "**coût de stockage**" ou "**coût de détention**"..., est alors:

$$C_p = t\% \cdot P_u \left(S_s + \frac{Q}{2} \right) = C_s \left(S_s + \frac{Q}{2} \right) \quad (67.97)$$

Il s'agit donc de la fonction d'une droite (dont l'ordonnée à l'origine est non-nulle si le stock de sécurité est non nul) si nous considérons que uniquement Q y est variable. Il est important de remarquer que ce coût ne prend pas en compte les concepts de remise de volume faite par les commerciaux...

Ces propositions nous amènent donc à l'équation du "**coût total d'approvisionnement**", appelé aussi "**coût total de stockage**" que nous allons chercher à minimiser:

$$C_t = \left(\frac{N}{Q} \cdot C_L \right) + \left(\frac{Q}{2} + S_s \right) \cdot t\% \cdot P_u \quad (67.98)$$

et qui donne une "courbe (hyperbolique) des coûts cumulés" du type:

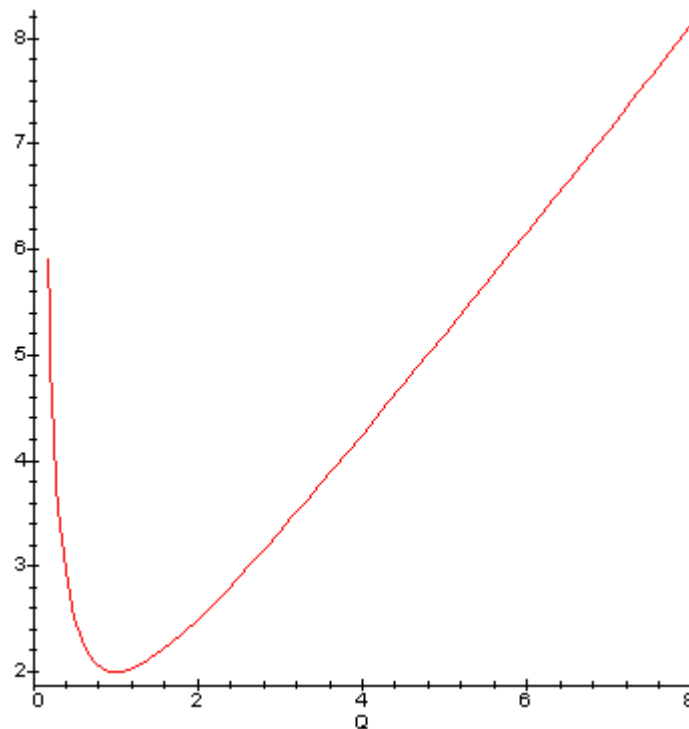


Figure: 67.18 - Courbe des coûts cumulés

Trouver la quantité économique Q_e , c'est trouver la valeur de Q pour laquelle le coût total est minimal (en d'autres termes c'est chercher la taille du lot qui permette de rentabiliser au mieux tous les coûts de stockage, fabrication, etc.), c'est-à-dire mathématiquement la valeur Q_e pour laquelle la dérivée du coût total par rapport à la quantité est nulle:

$$\frac{dC_t}{dQ} = -\frac{N}{Q^2} C_L + \frac{1}{2} t\% P_u = 0 \quad (67.99)$$

D'où la "relation de Wilson" (après un calcul élémentaire), appelée aussi simplement "formule de Wilson", pour le "lot/quantité économique optimal":

$$Q_e = \sqrt{\frac{2NC_L}{t\% \cdot P_u}} = \sqrt{\frac{2NC_L}{C_s}} \quad (67.100)$$

qui est donc pour rappel, le taille du lot qui minimise le coût total (indépendante du stock de sécurité qui s'annule lors de la dérivée). Nous constatons que la quantité économique est proportionnelle à la racine carée de la consommation. Donc grossièrement une article vendu 100 fois plus par une société nécessitera qu'un stock moyen 10 fois supérieur.

Bien évidemment une fois connue la quantité économique, il devient facile de calculer le "coût de gestion optimal par période" $C_{t,e}$ en injectant Q_e dans la relation obtenue plus haut:

$$C_t = \left(\frac{N}{Q} \cdot C_L \right) + \left(\frac{Q}{2} + S_s \right) \cdot t\% \cdot P_u \quad (67.101)$$

ainsi que la "**cadence ou fréquence optimale de réapprovisionnement**" puisque donnée par le rapport:

$$\frac{N}{Q} = \frac{N}{\sqrt{\frac{2NC_L}{C_S}}} = \sqrt{\frac{C_S N}{2C_L}} = \sqrt{\frac{N \cdot P_u \cdot t\%}{2C_L}} \quad (67.102)$$

Nous avons alors après un peu d'arithmétique élémentaire le coût de gestion optimal par période qui est donné par:

$$\begin{aligned} C_{t,e} &= \left(\sqrt{\frac{C_S N}{2C_L}} C_L \right) + \left(\frac{\sqrt{\frac{2NC_L}{C_S}}}{2} + S_s \right) t\% P_u = \sqrt{\frac{C_S C_L N}{2}} + \frac{\sqrt{\frac{2NC_L}{C_S}}}{2} C_S + S_s C_S \\ &= \sqrt{\frac{C_S C_L N}{2}} + \sqrt{\frac{NC_L C_S}{2}} + S_s C_S = 2\sqrt{\frac{NC_L C_S}{2}} + S_s C_S = \sqrt{2NC_L C_S} + S_s C_S \end{aligned} \quad (67.103)$$

Comme souvent, les gestionnaires aiment bien faire de l'analyse de la sensibilité (nous en avons déjà fait mention lors de notre étude du seuil de rentabilité). Dans le domaine de la gestion de stock il est de tradition de faire l'analyse de la sensibilité de la variation relative du coût total de stockage par rapport au coût de gestion optimale par période:

$$\frac{C_t}{C_{t,e}} = \frac{\frac{N}{Q} C_L + \frac{Q}{2} t\% \cdot P_u + S_s t\% \cdot P_u}{\sqrt{2NC_L C_S} + S_s C_S} = \frac{\frac{N}{Q} C_L + \frac{Q}{2} C_S + S_s C_S}{\sqrt{2NC_L C_S} + S_s C_S} \quad (67.104)$$

Mais comme le $S_s C_S$ (coût du stock de sécurité) est supposé indépendant de la quantité, il est d'usage de ne considérer que la variation relative indépendamment de ce terme telle que:

$$\frac{C_t}{C_{t,e}} = \frac{\frac{N}{Q} C_L + \frac{Q}{2} C_S}{\sqrt{2NC_L C_S}} \quad (67.105)$$

Enfin, il est d'usage de simplifier la forme analytique de cette dernière relation sous la forme suivante après quelques manipulations algébriques élémentaires:

$$\begin{aligned}
\frac{C_t}{C_{t,e}} &= \frac{\frac{N}{Q}C_L + \frac{Q}{2}C_S}{\sqrt{2NC_L C_S}} = \frac{NC_L}{Q\sqrt{2NC_L C_S}} + \frac{QC_S}{2\sqrt{2NC_L C_S}} \\
&= \frac{NC_L}{Q\sqrt{2NC_L C_S}} \frac{\sqrt{2NC_L C_S}}{\sqrt{2NC_L C_S}} + \frac{Q}{2\sqrt{\frac{2NC_L C_S}{C_S^2}}} \\
&= \frac{\sqrt{2NC_L C_S}}{2QC_S} + \frac{Q}{2\sqrt{\frac{2NC_L}{C_S}}} = \frac{\sqrt{2NC_L C_S}}{2QC_S} + \frac{Q}{2Q_e} = \frac{1}{2Q} \sqrt{\frac{2NC_L C_S}{C_S^2}} + \frac{Q}{2Q_e} \\
&= \frac{1}{2Q} \sqrt{\frac{2NC_L}{C_S}} + \frac{Q}{2Q_e} = \frac{1}{2Q} Q_e + \frac{Q}{2Q_e} = \frac{1}{2} \left(\frac{Q_e}{Q} + \frac{Q}{Q_e} \right)
\end{aligned} \tag{67.106}$$

c'est donc un résultat assez esthétique... que le lecteur remarquera au passage comme étant toujours positif!

Il est évident que nous tirons aussi de la relation de la fréquence optimale de réapprovisionnement la "**période optimale de réapprovisionnement**" donnée alors par (l'inverse de la fréquence par définition):

$$T_e = \sqrt{\frac{2C_L}{C_S N}} = \sqrt{\frac{2C_L}{N \cdot P_u \cdot t\%}} \tag{67.107}$$

Si nous reportons sur un graphique les fonctions:

- coût de lancement (approvisionnement) en fonction des quantités
- coût de possession en fonction des quantités
- coûts totaux en fonction des quantités

alors la quantité économique se trouve à l'intersection des deux courbes, lancement et possession (lorsque le coût de possession est égal au coût d'acquisition), ou au point d'inflexion de la courbe cumulée. Dans la pratique toutefois, il est possible de commander exactement la quantité économique, on choisira une taille de lot répondant aux diverses contraintes et comprise dans la "**zone économique**":

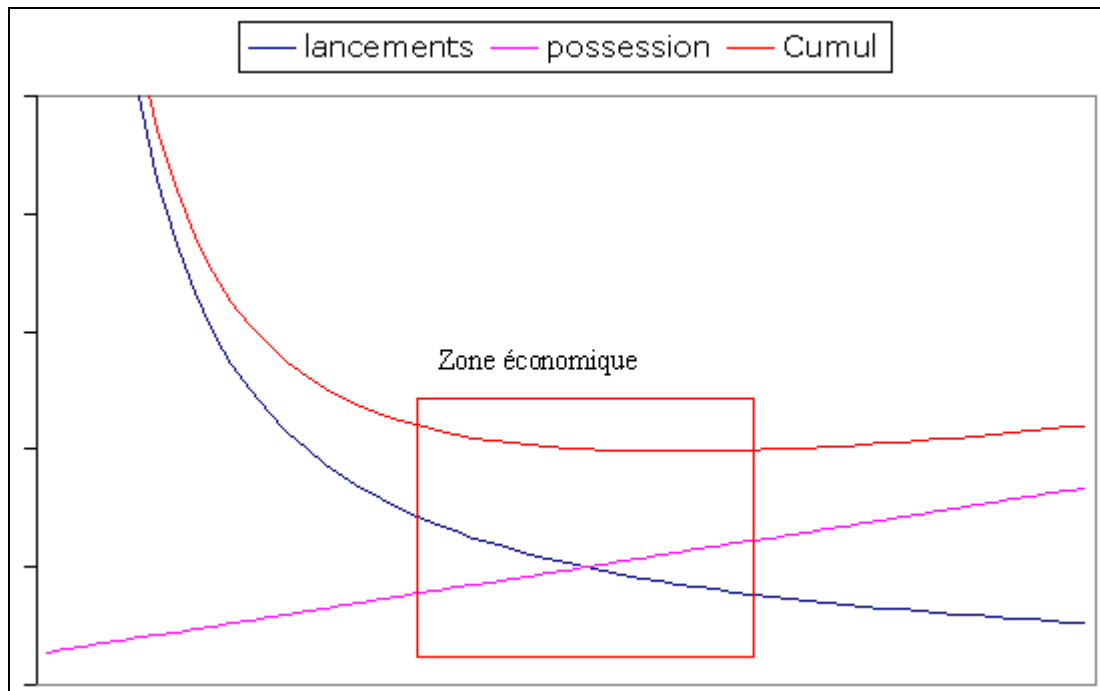


Figure: 67.19 - Représentation de la zone économique

Évidemment dans certaines entreprises un objectif est plutôt d'essayer de diminuer le coût de stockage afin d'atteindre l'équivalent de la demande comme quantité économique. Nous avons alors avec un peu d'algèbre élémentaire:

$$Q_e = \sqrt{\frac{2NC_L}{C_s}} \Rightarrow Q_e^2 = N^2 = \frac{2NC_L}{C_s} \quad (67.108)$$

soit au final le coût de stockage optimal si la quantité d'approvisionnement Q est imposée:

$$C_{S,e} = \frac{2C_L}{N} \quad (67.109)$$

Donc pour résumer, il faut savoir que l'on considère souvent que l'ensemble du modèle de Wilson se résume aux 5 relations ci-dessous:

$$\begin{aligned} C_{t,e} &= \sqrt{2NC_L C_S} + S_s C_S \\ Q_e &= \sqrt{\frac{2NC_L}{C_S}} \\ T_e &= \sqrt{\frac{2C_L}{C_S N}} \\ C_{S,e} &= \frac{2C_L}{N} \\ \frac{C_t}{C_{t,e}} &\cong \frac{1}{2} \left(\frac{Q_e}{Q} + \frac{Q}{Q_e} \right) \end{aligned} \quad (67.110)$$

Il existe un autre type de cas de figure qu'il faut étudier. Si l'on commande en quantités plus importantes en bénéficiant ainsi d'une remise, on augmente certes les coûts de possession mais on réduit théoriquement le nombre de commandes annuelles.

L'objectif pour le gestionnaire est bien sûr de vérifier mathématiquement que la remise consentie par son fournisseur n'entraîne pas de coûts induits supérieurs à la remise (ce serait une preuve d'incompétence du gestionnaire!).

Pour ce faire, il faut ramener tous les coûts à une pièce telle que le coût total unitaire s'écrive:

$$C_u = \left(\frac{C_L}{Q} \right) + \frac{\left(\frac{Q}{2} + S_s \right) \cdot t\% \cdot P_u}{N} + P_u \quad (67.111)$$

cette relation est importante, car elle détermine la valeur de la remise pour que cette dernière soit intéressante.

Pour connaître le seuil de remise R pour une quantité donnée, on remplace dans la relation précédente, Q par la quantité visée Q' et P_u par $P_u(1 - R)$, R étant la remise.

Nous résolvons alors l'équation et nous obtenons:

$$C_u \geq \left(\frac{C_L}{Q'} \right) + \frac{\left(\frac{Q'}{2} + S_s \right) \cdot t\% \cdot P_u(1 - R)}{N} + P_u(1 - R) \quad (67.112)$$

Nous déterminerons donc la valeur limite de R sous laquelle la remise ne compense pas les coûts internes.

Dans la pratique nous ne pouvons commander exactement la quantité optimale Q_e , notamment du fait des unités d'achats imposées par les fournisseurs (quantités minimales, conditionnements, etc.). Il est donc plus judicieux de s'intéresser à la "zone économique", constituée par la partie inférieure (le ventre) de la courbe des coûts totaux.

Du fait des hypothèses simplificatrices, le modèle de Wilson ne peut fournir au mieux qu'un ordre de grandeur si consommation et/ou prix sont sujets à variations (puisque'elle est extrêmement dépendante des deux paramètres subjectifs: coûts de stockages et lancements).

Le recours aux lancements de fabrication économiques est "anti-flexible" par essence. Ce genre de politique amène fréquemment des conséquences importantes, risque de gonfler le stock de produits finis, reportant les coûts et pertes en aval du processus.

Cependant, le modèle de Wilson a ceci d'intéressant qu'il peut s'appliquer également assez bien à des ressources humaines.

Exemple:

L'entreprise MAC utilise un article X330 pour lequel la consommation prévisionnelle de l'année N devrait être de 4'000 articles. Les données sont les suivantes:

- Le coût unitaire de l'article X330 est de $P_u = 8$ (peu importe le numéraire)

- Le coût de passation/lancement d'une commande est de $C_L = 100$
- Le taux de possession du stock est de $t\% = 10\%$ par an
- Le stock de sécurité est de 250 unités

Le fournisseur de cet article, pour inciter ses clients à augmenter l'importance de leurs commandes, propose à l'entreprise les conditions suivantes:

C1. Quantités commandées inférieures à 2'000 unités: prix unitaire $P_u = 8$

C2. Quantités commandées comprises entre 2'000 et 3'500 unités: remise de 2%.

C3. Quantités commandées supérieures à 3'500 unités: remise de 3 %.

Travail à faire: Dire quelle solution l'entreprise doit adopter.

Le prix varie donc en fonction de la quantité tel qu'étant donnée une quantité choisie, la remise s'applique d'une façon équivalente à tous les articles (nous parlons alors de "**remise uniforme**").

D'après l'énoncé et en fonction de Q la quantité d'approvisionnement, nous savons que:

$$1. \text{ Si } Q < 2'000 \Rightarrow P_u = P_1 = 8$$

$$2. \text{ Si } 2'000 < Q < 3'500 \Rightarrow P_u = P_2 = 8 - 2\% \cdot 8 = 7.84$$

$$3. \text{ Si } Q \geq 3'500 \Rightarrow P_u = P_3 = 8 - 3\% \cdot 8 = 7.76$$

En fonction de la relation de Wilson du lot économique, nous allons calculer la quantité économique Q_e pour le prix le plus avantageux à savoir P_3 :

$$Q_e = \sqrt{\frac{2NC_L}{t\% \cdot P_u}} = \sqrt{\frac{2 \cdot 4'000 \cdot 100}{7.76 \cdot 10\%}} = 1015'35 \quad (67.113)$$

Mais pour avoir droit à P_3 il faut commander au minimum 3'500 articles il y a donc contradiction et cette solution est donc hors zone. Des calculs identiques (que nous laissons le soin de faire avec la calculatrice quand même...) montrent que seul le lot économique Q_e de P_1 correspond à la contrainte $Q < 2'000$.

Dès lors, la période optimale de réapprovisionnement correspondante sera:

$$T_e = \sqrt{\frac{2C_L}{C_S N}} = \sqrt{\frac{2 \cdot 100}{(10\% \cdot 8) \cdot 4'000}} = 0.25 = 91.25 [j.] \quad (67.114)$$

et le coût de gestion optimal par période:

$$C_{t,e} = \sqrt{2NC_L C_S} + S_s C_S = \sqrt{2 \cdot 4'000 \cdot 100 \cdot (10\% \cdot 8)} + 250 \cdot (10\% \cdot 8) = 1'000.- \quad (67.115)$$

Attention! Pour ce dernier calcul, certains auteurs incluent dans la coût du stock de sécurité $S_s C_s$ la quantité N . Il s'agit d'un choix plutôt discutable et qui nécessiterait d'être précisé.

Enfin, l'impact sur le coût total de gestion si nous augmentons ou diminuons la quantité de plus ou moins 30% par rapport à la quantité économique est donné par:

$$\begin{aligned}\frac{C_t}{C_{t,e}} &= \frac{1}{2} \left(\frac{Q_e}{Q} + \frac{Q}{Q_e} \right) = \frac{1}{2} \left(\frac{1}{1.3} + \frac{1.3}{1} \right) \cong 1.0346 \\ \frac{C_t}{C_{t,e}} &= \frac{1}{2} \left(\frac{Q_e}{Q} + \frac{Q}{Q_e} \right) = \frac{1}{2} \left(\frac{1}{0.7} + \frac{0.7}{1} \right) \cong 1.0643\end{aligned}\quad (67.116)$$

Donc évidemment, cela ne fait que d'augmenter les coûts et comme nous l'avons vu sur le graphique plus haut, la courbe n'étant pas symétrique, les deux résultats ne sont pas égaux.

6.3.2. MODÈLE DE WILSON SANS RÉAPPROVISIONNEMENT

Dans le modèle précédent, nous avons donc supposé une gestion de stocks en dents de scie de la forme suivante:

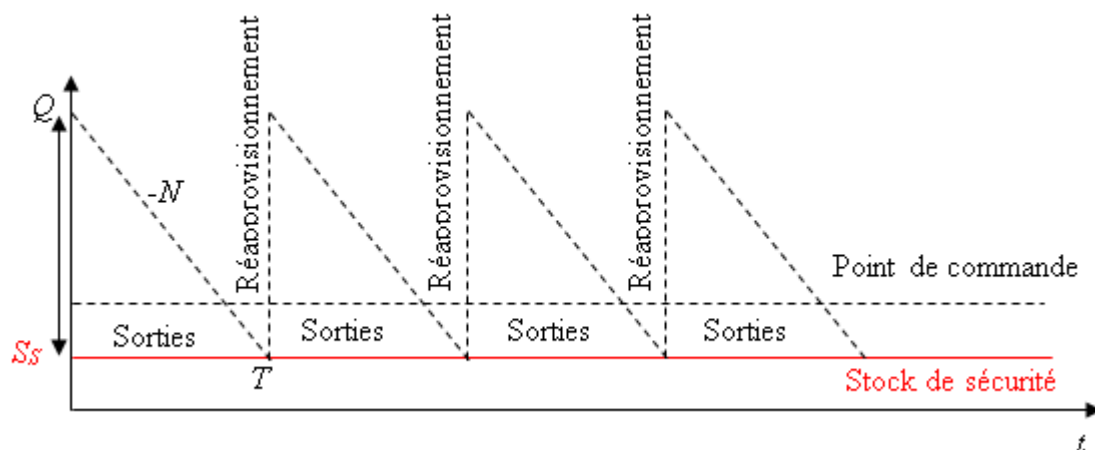


Figure: 67.20 - Représentation schématique du modèle de Wilson avec réapprovisionnement

Dans le modèle qui va nous intéresser maintenant, nous allons considérer une gestion de stock avec une inertie du réapprovisionnement souvent due à une production à l'interne sous la forme suivante:

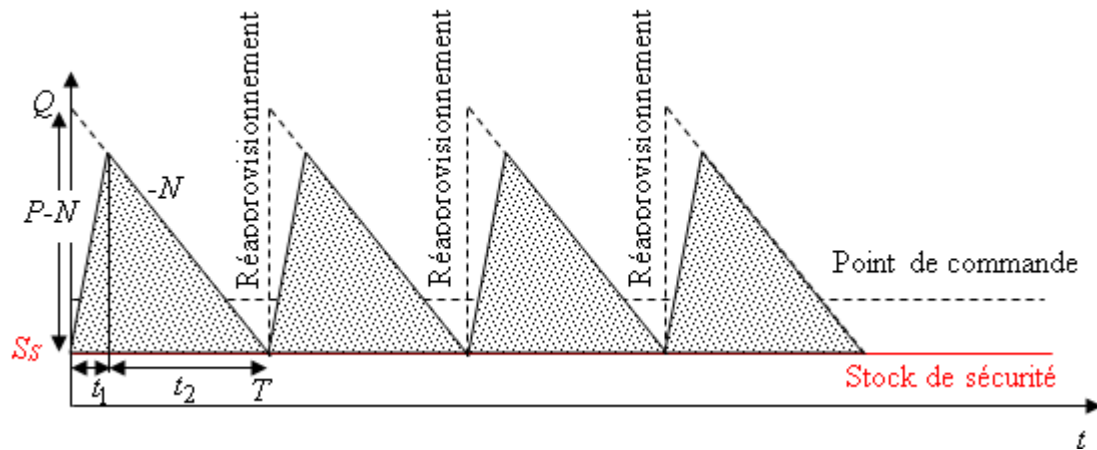


Figure: 67.21 - Représentation schématique du modèle de Wilson sans réapprovisionnement

Donc dans la figure ci-dessus, nous pouvons observer que la demande de production se fait au point de commande et que pendant la production au rythme de P pièces par unité de temps pendant un temps t_1 , la consommation N par période T (rapportée à l'unité de temps de la production), tape déjà dans les stocks générés.

Devons-nous donc refaire tous les calculs du modèle précédent? La réponse est: non pas en totalité! Car si nous réfléchissons un tout petit peu en observant la fonction hyperbolique du coût total de stockage que nous avons démontrée plus haut:

$$C_t = \left(\frac{N}{Q} \cdot C_L \right) + \left(\frac{Q}{2} + S_s \right) \cdot t^{\circ}_0 \cdot P_u \quad (67.117)$$

Il est évident que le premier terme dû au réapprovisionnement n'a aucune raison de changer. Le troisième terme, dû au stock de sécurité, n'a aucune raison de changer non plus! Donc il ne reste que le deuxième terme représenté par le coût de stockage moyen qui est forcément différent!

Donc le stockage moyen n'est plus:

$$\frac{Q}{2} \quad (67.118)$$

Nous pourrions calculer facilement la surface moyenne du triangle en utilisant la formule de Héron pour le calcul de la surface d'un triangle quelconque démontré dans le chapitre de Formes Géométriques. Mais le problème dans le cas présent, c'est que nous voulons faire apparaître Q dans le stock moyen. Nous ne pourrions alors utiliser la formule de Héron. L'idée va être la suivante (se référer à la figure précédente):

Lorsque l'entreprise n'a plus de stock, elle en lance pendant un temps t_1 la production à un rythme de P éléments par unité temporelle jusqu'à atteindre en cumul la quantité objectif Q . Nous avons alors:

$$Q = P \cdot t_1 \quad (67.119)$$

Mais en même temps qu'elle produit (et après aussi), l'entreprise consomme N éléments par unité temporelle. Le stock réel généré sera alors au temps t_1 (sous la condition évident que P est supérieur à N) l'inventaire maximum:

$$Inv_{\max} = P \cdot t_1 - N \cdot t_1 = (P - N) \cdot t_1 \quad (67.120)$$

Cet inventaire maximum correspond donc à la hauteur de notre triangle quelconque (toujours se référer à la figure précédente) mais nous souhaitons y faire apparaître le stocke objectif Q . Donc, pour cela, il n'y a maintenant rien de plus simple en utilisant la relation antérieure:

$$Inv_{\max} = (P - N) \cdot t_1 = (P - N) \cdot \frac{Q}{P} = \left(1 - \frac{N}{P}\right) Q \quad (67.121)$$

Donc le stock moyen devient:

$$\frac{1}{2} \left(1 - \frac{N}{P}\right) Q \quad (67.122)$$

Donc nous voyons que la seule différence entre les deux modèles est le facteur:

$$\left(1 - \frac{N}{P}\right) \quad (67.123)$$

Dès lors, dans relations obtenues dans le modèle de Wilson avec réapprovisionnement (du moins les plus importantes) et qui étaient pour rappel:

$$\begin{aligned} C_{t,e} &= \sqrt{2NC_L C_S} + S_s C_S \\ Q_e &= \sqrt{\frac{2NC_L}{C_S}} \\ T_e &= \sqrt{\frac{2C_L}{C_S N}} \\ C_{S,e} &= \frac{2C_L}{N} \\ \frac{C_t}{C_{t,e}} &\cong \frac{1}{2} \left(\frac{Q_e}{Q} + \frac{Q}{Q_e} \right) \end{aligned} \quad (67.124)$$

deviennent (la dernière relation ne change pas):

$$\begin{aligned}
 C_{t,e} &= \sqrt{2NC_L C_S \left(1 - \frac{N}{P}\right) + S_s C_S} \\
 Q_e &= \sqrt{\frac{2NC_L}{\left(1 - \frac{N}{P}\right) C_S}} \\
 T_e &= \sqrt{\frac{2C_L}{\left(1 - \frac{N}{P}\right) C_S N}} \\
 C_{S,e} &= \frac{2C_L}{N \left(1 - \frac{N}{P}\right)}
 \end{aligned}
 \tag{67.125}$$

L'adaptation des relations du coût périodique de possession et du coût total d'approvisionnement sont quant à eux immédiats.

Donc à la différence du modèle avec approvisionnement, nous voyons que seulement une nouvelle variable P apparaît est qui est donc la vitesse de production (éléments produits par unité de temps) pour permettre une production de N éléments par unité de temps. Ce que lecteur ne doit pas oublier lors de cas d'application numérique, c'est de rapporter P et N à la même unité de temps.

6.3.3. MODÈLE DE WILSON AVEC RÉAPPROVISIONNEMENT ET RUPTURE

Dans les grands classiques d'études scolaires (le reste étant ensuite une combinaisons cas des différents scénarios basé sur le même raisonnement), voyons le modèle de Wilson avec réapprovisionnement (instantané) mais où la rupture et la vente à découvert (en contrepartie d'une pénalité) est autorisé (donc il n'y a pas de stock de sécurité!). Nous représenterons ce scénario par la figure en dents-de-scie suivante:

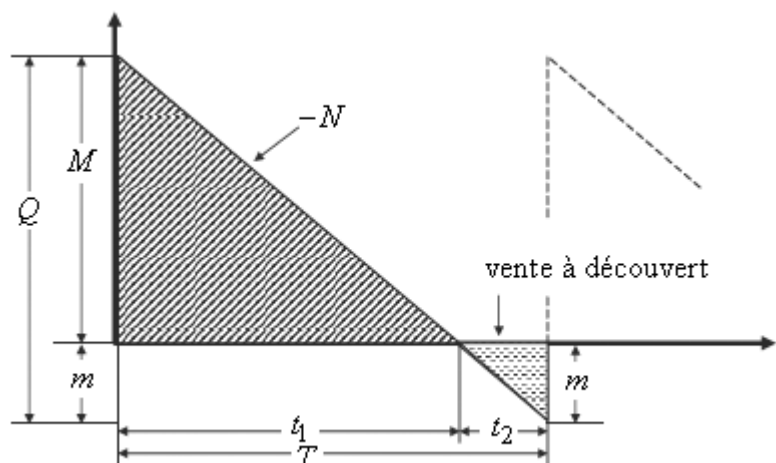


Figure: 67.22 - Représentation schématique du modèle de Wilson avec réapprovisionnement et rupture

Nous considérerons C_R comme étant la perte financière (coût) par unité de pièce pendant la période à découvert (ne pas encaisser l'argent tout de suite fait que nous en perdons rien que par le fait que nous ne pouvons pas le placer sur des fonds à rendement

Pour ce cas là, nous n'avons pas trouvé de manière simple d'éviter de refaire les calculs des relations du modèle de Wilson avec réapprovisionnement (bon si quelqu'un a une astuce nous sommes preneurs). Cependant il s'agit toujours d'algèbre élémentaire donc le lecteur ne devrait souffrir d'aucune difficulté face à ce genre d'exercice de style.

Nous avons d'abord trivialement en se référant à la figure ci-dessus:

$$\begin{aligned} T &= \frac{Q}{N} \\ t_1 &= \frac{M}{N} \\ t_2 &= T - t_1 = \frac{Q}{N} - \frac{M}{N} = \frac{Q - M}{N} \\ m &= Q - M \end{aligned} \quad (67.126)$$

D'abord, le coût de stockage moyen est trivialement de:

$$\frac{M \cdot t_1}{2} C_S = \frac{M \cdot \left(\frac{M}{N}\right)}{2} C_S = \frac{M^2}{2N} C_S \quad (67.127)$$

Le coût moyen de vente en rupture est lui aussi donné trivialement par:

$$\frac{m \cdot t_2}{2} C_R = \frac{(Q - M) \cdot \left(\frac{Q - M}{N}\right)}{2} C_R = \frac{(Q - M)^2}{2N} C_R \quad (67.128)$$

Le coût moyen total sur une période est alors de:

$$C = \frac{M^2}{2N} C_S + \frac{(Q - M)^2}{2N} C_R \quad (67.129)$$

Rapportons maintenant cela au temps d'une période:

$$\frac{C}{T} = \frac{C}{\left(\frac{Q}{N}\right)} = \frac{M^2}{2Q} C_S + \frac{(Q - M)^2}{2Q} C_R \quad (67.130)$$

Le coût total moyen par période est alors de:

$$\begin{aligned}
 C_t &= \frac{N}{Q} C_L + \frac{M^2}{2Q} C_S + \frac{(Q-M)^2}{2Q} C_R \\
 &= \frac{N}{Q} C_L + \frac{M^2 (C_S + C_R)}{2Q} + \frac{QC_R}{2} - MC_R
 \end{aligned} \quad (67.131)$$

Le problème ici c'est que Q dépend de M et inversement. Nous allons donc devoir faire usage des dérivées partielles pour fixer tantôt l'un et tantôt l'autre. Donc pour trouver la quantité économique, nous allons arbitrairement d'abord déterminer quelle est la quantité M qui va minimiser le coût total moyen par période (ceci afin de se débarrasser par anticipation de la présence de la variable M dans l'expression de la quantité économique). Nous avons alors en prenant la dérivée partielle (donc sous-entendu que nous maintenons toutefois Q constant):

$$\frac{\partial C_t}{\partial M} = \frac{M(C_S + C_R)}{Q} - C_R = 0 \quad (67.132)$$

Soit:

$$M_e = \frac{C_R Q}{C_S + C_R} \quad (67.133)$$

Maintenant, pour la quantité économique, nous avons respectivement en prenant la dérivée partielle (dons sous-entendu que nous fixons M):

$$\frac{\partial C_t}{\partial Q} = -\frac{N}{Q^2} C_L - \frac{M^2 (C_S + C_R)}{2Q^2} + \frac{C_R}{2} = 0 \quad (67.134)$$

En y injectant le résultat antérieur il vient:

$$\begin{aligned}
 &-\frac{NC_L}{Q^2} - \frac{\left(\frac{QC_R}{C_S + C_R}\right)^2 (C_S + C_R)}{2Q^2} + \frac{C_R}{2} \\
 &= -\frac{NC_L}{Q^2} - \frac{C_R^2}{2(C_S + C_R)} + \frac{C_R}{2} = 0
 \end{aligned} \quad (67.135)$$

Soit après quelques manipulations algébrique élémentaires:

$$\frac{2NC_L}{Q^2} = \frac{C_R C_S}{C_S + C_R} \quad (67.136)$$

Ce qui donne finalement:

$$Q_e = \sqrt{2NC_L \frac{C_S + C_R}{C_R C_S}} = \sqrt{\frac{2NC_L}{C_S}} \sqrt{\frac{C_S + C_R}{C_R}} \quad (67.137)$$

Nous avons alors explicitement:

$$M_e = \frac{C_R}{C_S + C_R} Q_e = \frac{C_R}{C_S + C_R} \sqrt{\frac{2NC_L}{C_S}} \sqrt{\frac{C_S + C_R}{C_R}} = \sqrt{\frac{2NC_L}{C_S}} \sqrt{\frac{C_R}{C_S + C_R}} \quad (67.138)$$

La période optimale de réapprovisionnement est alors:

$$T_e = \frac{Q_e}{N} = \frac{1}{N} \sqrt{\frac{2NC_L}{C_S}} \sqrt{\frac{C_S + C_R}{C_R}} = \sqrt{\frac{2C_L}{NC_S}} \sqrt{\frac{C_S + C_R}{C_R}} \quad (67.139)$$

La quantité m économique de rupture donnant le déclenchement de lancement de restockage est alors:

$$\begin{aligned} m_e = Q_e - M_e &= \sqrt{\frac{2NC_L}{C_S}} \sqrt{\frac{C_S + C_R}{C_R}} - \sqrt{\frac{2NC_L}{C_S}} \sqrt{\frac{C_R}{C_S + C_R}} \\ &= \sqrt{\frac{2NC_L}{C_S}} \left(\sqrt{\frac{C_S + C_R}{C_R}} - \sqrt{\frac{C_R}{C_S + C_R}} \right) \end{aligned} \quad (67.140)$$

Une difficulté pratique de ce modèle est de déterminer C_R . L'idée est alors de l'estimer sur la durée maximale d'attente supposée du client ayant fait une commande lors d'une rupture. En partant de:

$$\begin{aligned} t_{2,e} = t_{\text{limite}} &= \frac{Q_e - M_e}{N} = \sqrt{\frac{2C_L}{NC_S}} \left(\sqrt{\frac{C_S + C_R}{C_R}} - \sqrt{\frac{C_R}{C_S + C_R}} \right) \\ &= \sqrt{\frac{2C_L}{NC_S}} \left(\sqrt{\frac{(C_S + C_R)(C_S + C_R)}{C_R(C_S + C_R)}} - \sqrt{\frac{C_R C_R}{(C_S + C_R)C_R}} \right) \\ &= \sqrt{\frac{2C_L}{NC_S}} \left(\sqrt{\frac{(C_S + C_R)^2}{C_R(C_S + C_R)}} - \sqrt{\frac{C_R^2}{(C_S + C_R)C_R}} \right) \\ &= \sqrt{\frac{2C_L}{NC_S C_R (C_S + C_R)}} \left(\sqrt{(C_S + C_R)^2} - \sqrt{C_R^2} \right) \\ &= \sqrt{\frac{2C_L}{NC_S C_R (C_S + C_R)}} (C_S + C_R - C_R) = \sqrt{\frac{2C_L}{NC_S C_R (C_S + C_R)}} C_S \\ &= \sqrt{\frac{2C_L C_S}{NC_R (C_S + C_R)}} \end{aligned} \quad (67.141)$$

Et en mettant au carré dans l'idée d'extraire C_R :

$$t_{\text{limite}}^2 = \frac{2C_L C_S}{NC_R (C_S + C_R)} \Rightarrow C_R (C_S + C_R) = \frac{2C_L C_S}{Nt_{\text{limite}}^2} \quad (67.142)$$

$$\Rightarrow C_R^2 + C_R C_S = \frac{2C_L C_S}{Nt_{\text{limite}}^2}$$

Nous avons donc finalement une équation du deuxième degré:

$$C_R^2 + C_R C_S - \frac{2C_L C_S}{Nt_{\text{limite}}^2} = 0 \quad (67.143)$$

Dont les deux racines réelles sont:

$$C_{R,2} = \frac{-C_R \pm \sqrt{C_R^2 + 4 \frac{2C_L C_S}{Nt_{\text{limite}}^2}}}{2} = -\frac{C_R}{2} \pm \sqrt{\left(\frac{C_R}{2}\right)^2 + \frac{2C_L C_S}{Nt_{\text{limite}}^2}} \quad (67.144)$$

La racine qu'il faut garder est celle qui donne un coût de rupture positif. Soit:

$$C_R = -\frac{C_R}{2} + \sqrt{\left(\frac{C_R}{2}\right)^2 + \frac{2C_L C_S}{Nt_{\text{limite}}^2}} \quad (67.145)$$

Avant de passer à l'exemple pratique, le résumé se réduit à:

$$\begin{aligned} Q_e &= \sqrt{\frac{2NC_L}{C_S}} \sqrt{\frac{C_S + C_R}{C_R}} \\ M_e &= \sqrt{\frac{2NC_L}{C_S}} \sqrt{\frac{C_R}{C_S + C_R}} \\ T_e &= \sqrt{\frac{2C_L}{NC_S}} \sqrt{\frac{C_S + C_R}{C_R}} \\ m_e &= \sqrt{\frac{2NC_L}{C_S}} \left(\sqrt{\frac{C_S + C_R}{C_R}} - \sqrt{\frac{C_R}{C_S + C_R}} \right) \\ t_{\text{limite}} &= \sqrt{\frac{2C_L C_S}{NC_R (C_S + C_R)}} \\ C_R &= -\frac{C_R}{2} + \sqrt{\left(\frac{C_R}{2}\right)^2 + \frac{2C_L C_S}{Nt_{\text{limite}}^2}} \end{aligned} \quad (67.146)$$

 **Exemple:**

L'entreprise MAC livre un article X330. Dans le but de ne pas perdre ses clients, les gestionnaires décident d'un temps de rupture optimal (sous-entendu: maximum) de 2 jours. Nous avons les données suivantes:

$$C_L = 200 \left[\frac{\text{€}}{\text{restockage}} \right], C_S = 1 \left[\frac{\text{€}}{\text{unité} \times \text{jour}} \right], t_{\text{limite}} = 2 \text{ [jour]}, N = 5 \left[\frac{\text{unités}}{\text{jour}} \right] \quad (67.147)$$

Ce qui donne:

$$C_R = 4 \left[\frac{\text{unités}}{\text{jour}} \right], Q_e = 50 \text{ [unités]}, M_e = 40 \text{ [unités]}, m_e = 10 \text{ [unités]}, T_e = 10 \text{ [jour]} \quad (67.148)$$

7. ANALYSE DE LA SENSIBILITÉ

Dans le domaine de la gestion (projets, finance, statistique), l'analyse de la sensibilité est une technique simple basée sur l'utilisation du coefficient de corrélation linéaire (de Pearson) démontré dans le chapitre de Statistiques dont une écriture possible est:

$$R_{x,y} = \frac{\sum_{k=1}^n x_k y_k - \frac{1}{n} \sum_{k=1}^n x_k \sum_{k=1}^n y_k}{\sqrt{\left[\sum_{k=1}^n x_k^2 - \frac{1}{n} \left(\sum_{k=1}^n x_k \right)^2 \right] \left[\sum_{k=1}^n y_k^2 - \frac{1}{n} \left(\sum_{k=1}^n y_k \right)^2 \right]}} \quad (67.149)$$

et une écriture parfaitement équivalente démontrée dans le chapitre de Méthodes Numériques:

$$R_{x,y} = \frac{\sum_k (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_k (x_k - \bar{x})^2 \sum_k (y_k - \bar{y})^2}} \quad (67.150)$$

avec pour rappel:

$$-1 < R_{x,y} < +1 \quad (67.151)$$

L'idée est que si nous avons une variable principale qui dépend de manière sous-jacente de plusieurs variables alors nous calculons le coefficient de corrélation (ou le coefficient de détermination qui est simplement le carré du coefficient de corrélation) de la variable principale avec chacune des variables sous-jacentes.

Prenons un exemple simple et particulier au monde de la gestion de projets. Imaginons un projet dont le planning est basé un chemin critique basé uniquement de deux tâches A et B. La première tâche A s'étale sur une durée dont la loi est Normale de paramètres:

$$N(\mu, \sigma) = N(8, 1) \quad (67.152)$$

soit une durée espérée de 8 jours et un écart-type de 1 jour et la tâche B s'étale sur une durée suivant une loi bêta (selon recommandation du PMI...) de paramètres (5,12), soit une durée optimiste de 5 jours et une durée pessimiste de 12 jours:

$$P_{a,b} = \frac{(x-a)^{\alpha-1} (x-a)^{\gamma-1}}{(b-a)^{\alpha+\gamma-1} B(\alpha, \gamma)} = \frac{(x-5)^{3+\sqrt{2}-1} (x-5)^{3-\sqrt{2}-1}}{(12-5)^{3+\sqrt{2}+3-\sqrt{2}-1} B(3+\sqrt{2}, 3-\sqrt{2})}$$

$$= \frac{(x-5)^{2+\sqrt{2}} (x-5)^{2-\sqrt{2}}}{7^7 B(3+\sqrt{2}, 3-\sqrt{2})} \quad (67.153)$$

Faisons pour l'exemple une simulation de Monte-Carlo à 1'000 itérations avec la version anglaise de Microsoft Excel 11.8346 (puisque la majorité des gestionnaires travaillent avec la version anglaise):

	B	C	D
1	A	B	Somme
2	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C2+B2
3	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C3+B3
4	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C4+B4
5	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C5+B5
6	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C6+B6
7	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C7+B7
8	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C8+B8
9	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C9+B9
10	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C10+B10
11	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C11+B11
12	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C12+B12
13	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C13+B13
14	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C14+B14
15	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C15+B15
16	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C16+B16
17	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C17+B17
18	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C18+B18
19	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C19+B19
20	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C20+B20
21	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C21+B21
22	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C22+B22
23	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C23+B23
24	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C24+B24
25	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C25+B25
26	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C26+B26
27	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C27+B27
28	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C28+B28
29	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C29+B29
30	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C30+B30
31	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C31+B31
32	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C32+B32

Figure: 67.23 - Simulation de deux petites tâches respectivement avec une distribution Normale et bêta

Ensuite, nous calculons toujours avec la version anglaise de Microsoft Excel 11.8346 le coefficient de corrélation:

	F	G
1	Coeff. C/A	Coeff. C/B
2	=PEARSON(B2:B1001;D2:D1001)	=PEARSON(C2:C1001;D2:D1001)

Figure: 67.24 - Calcul du coefficient de corrélation entre les sommes et les tâches

ce qui donne:

	F	G
1	Coeff. C/A	Coeff. C/B
2	0.64	0.75

Figure: 67.25 - Corrélation correspondante

Soit sous la forme d'un graphique de type "**tornado chart**" comme il est d'usage dans le domaine de l'analyse de la sensibilité:

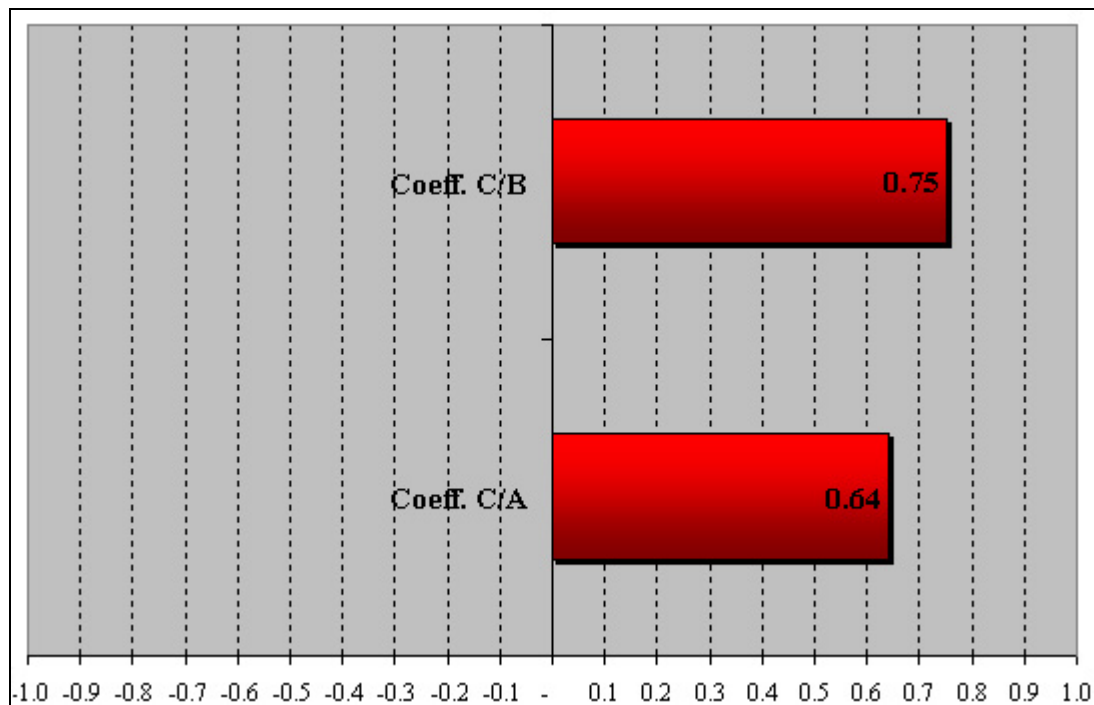


Figure: 67.26 - Tornado chart obtenu avec Microsoft Excel 11.8346

Il est dangereux d'utiliser le coefficient de corrélation (de Pearson) lorsque la relation entre la variable principale et les variables sous-jacentes n'est pas linéaire! Cependant, les logiciels peuvent identifier de quel autre type de relation il s'agit (exponentielle, logarithmique, puissance, polynomiale) et faire une transformation pour ramener l'équation à celle d'une droite. Cependant, cette manipulation peut aussi amener à des erreurs grossières dans certains cas. La majorité des logiciels spécialisés utilisent alors la définition empirique de la régression que nous avons vue dans le chapitre de Méthodes Numériques:

$$R^2_{xy} = 1 - \frac{SCR}{SCT} \quad (67.154)$$

et utilisent alors la recherche opérationnelle (cf. chapitre de Méthodes Numériques) pour déterminer les paramètres de l'équation qui minimisent le R^2 selon la relation précédente.

Un logiciel spécialisé pour obtenir très rapidement ce genre d'analyse sur des situations complexes est @Risk de Palisade qui donne lui le graphique suivant:

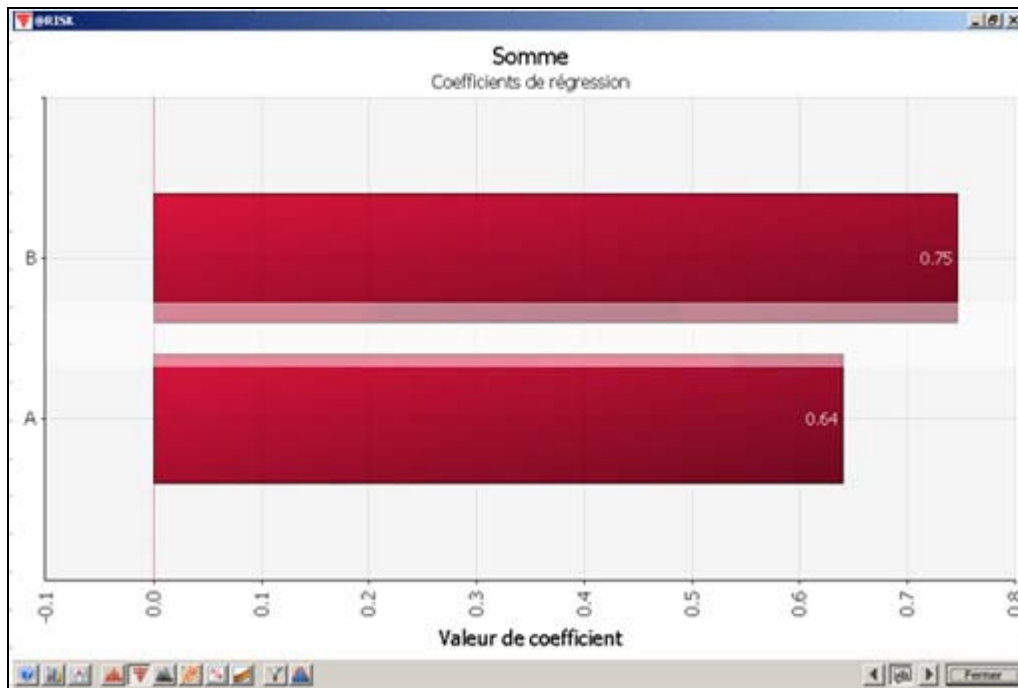


Figure: 67.27 - Tornado chart de la même simulation obtenu avec @Risk

Enfin, quand nous avons un très grand nombre de variables sous-jacentes, il peut être intéressant de regrouper les coefficients de corrélation par intervalles avec un graphique du type suivant (exemple pris de la bourse) où l'ordonnée représente le % du nombre de variables:

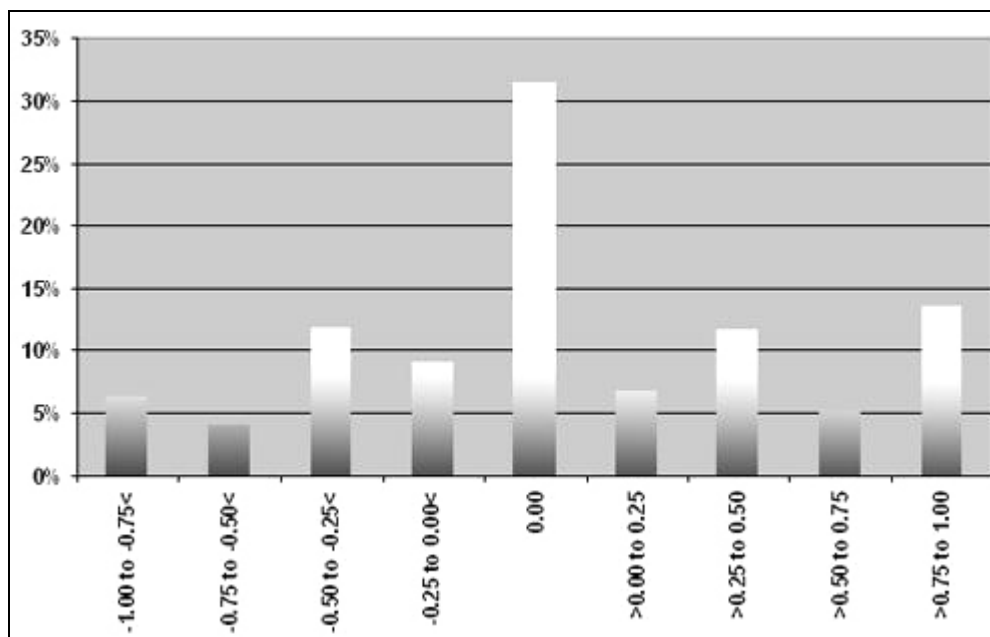


Figure: 67.28 - Analyse des groupes de corrélation de variables

Ainsi, par exemple, ci-dessus 32% des variables ont un coefficient de corrélation nul avec la variable principale (qui est l'indice de performance du portefeuille concerné) et 14% ont un coefficient de corrélation compris entre +0.75 et +1.

Il est aussi courant de représenter l'influence en % de chaque variable sur la variable principale en faisant la moyenne des rapports entre les deux. Ce qui donne le graphique suivant si nous en avons un grand nombre:

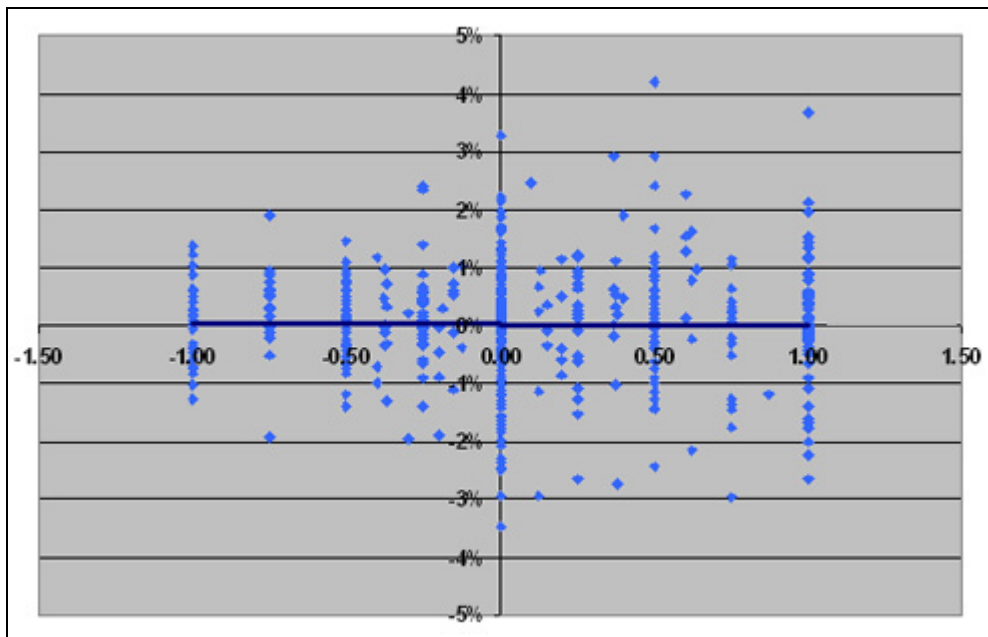


Figure: 67.29 - % d'influence sur la variable principale en fonction de la corrélation

Ainsi, par exemple, nous pouvons observer sur le graphique ci-dessus qu'une variable a une influence moyenne de +4% sur l'indice du portefeuille avec une corrélation de +0.5.

Bref... libre cours pour faire le graphique que l'on veut en fonction des besoins de son entreprise!

8. BIENS D'ÉQUIPEMENT

Les installations, les biens d'équipement subissent une dépréciation progressive due à l'usure ou à l'obsolescence. Cette baisse de valeur, enregistrée comme une charge en comptabilité, est appelée "**amortissement comptable**". Il ne faut pas confondre l'amortissement financier vu dans le chapitre d'Économie qui correspond au remboursement d'une dette et l'amortissement comptable qui est une diminution de valeur des moyens de production.

Certains types de biens ont une perte de valeur assez uniforme dans le temps contrairement à d'autres qui se déprécient plus rapidement les premières années. Nous allons présenter ici quelques-unes des méthodes comptables utilisées en pratique qui décrivent l'un ou l'autre de ces phénomènes.

8.1. AMORTISSEMENT LINÉAIRE

Définition: Nous parlons d'un "**amortissement linéaire**" d'un bien lorsque sa valeur d'immobilisation est diminuée (amortissement) d'un montant périodique (annuel dans la comptabilité) constant durant sa durée de vie:

Ainsi, si nous notons A_k le montant du k -ème amortissement et V_0 la valeur initiale du bien d'équipement et V_n sa valeur finale souhaitée (qui dans la plupart des cas sera nulle), nous avons:

$$A = A_k = \frac{V_0 - V_n}{n} \quad (67.155)$$

Il est possible d'obtenir la valeur d'amortissement en utilisant la fonction AMORLIN() de la version française de Microsoft Excel 11.8346.

Le taux d'amortissement équivalent constant basé sur la valeur d'achat et résiduelle est alors donné évidemment par:

$$t\% = \frac{\frac{V_0 - V_n}{n}}{V_0} \quad (67.156)$$

À remarquer que ce taux constant ainsi que la valeur finale sont parfois imposés par la législation dans certains pays (comme c'est le cas en Suisse par exemple) jusqu'à une valeur résiduelle nulle ou non nulle elle aussi définie par la législation!

8.2. AMORTISSEMENT ARITHMÉTIQUE DÉGRESSIF

Définition: Lorsque la valeur d'immobilisation (amortissement) d'un bien décroît inversement à l'ordre des périodes (années en comptabilité) nous parlons alors "**d'amortissement arithmétique dégressif**" (donc pas de taux d'amortissement imposé par l'État possible!).

Par exemple, un bien d'une durée de vie de 4 ans, sera amortissable de 4/10 la première année, 3/10 la seconde, 2/10 la troisième et 1/10 la dernière. La base commune "10" (dans cet exemple) étant la somme arithmétique 1+2+3+4 afin que la totalité des fractions soit égale à l'unité. Il s'agit d'une règle purement fiscale américaine " Sum-of-Years Digits" (SYD).

Remarque: Il ne faut pas confondre "l'amortissement arithmétique dégressif" avec "l'amortissement dégressif" qui consiste à appliquer un coefficient multiplicateur (édicte par le fisc) au pourcentage d'amortissement linéaire correspondant et que nous ne traiterons pas sur ce site (sauf demande explicite d'un lecteur). Il ne peut être utilisé que pour des biens neufs et ne concerne pas tous les types d'immobilisation. Ce taux d'amortissement dégressif s'applique chaque année sur la valeur comptable résiduelle du bien.

Soit A_k le k -ème amortissement et V_0 la valeur initiale du bien d'équipement et V_n sa valeur finale souhaitée (qui dans la plupart des cas sera nulle) nous avons alors:

$$\begin{aligned} A_1 &= \frac{4}{10}(V_0 - V_n) \\ A_2 &= \frac{3}{10}(V_0 - V_n) \\ A_3 &= \frac{2}{10}(V_0 - V_n) \\ A_4 &= \frac{1}{10}(V_0 - V_n) \end{aligned} \quad (67.157)$$

ce qui peut s'écrire:

$$A_k = \frac{n - k + 1}{1 + 2 + 3 + 4}(V_0 - V_n) = \frac{n - k + 1}{S_n}(V_0 - V_n) \quad (67.158)$$

et comme nous l'avons démontré dans le chapitre des Suites Et Séries:

$$S_n = \frac{n(n+1)}{2} \quad (67.159)$$

ce qui nous amène à écrire:

$$A_k = \frac{2(V_0 - V_n)}{n(n+1)}(n-k+1) \quad (67.160)$$

Il est possible d'obtenir l'amortissement à une période k en utilisant la fonction SYD() de la version française de Microsoft Excel 11.8346.

8.3. AMORTISSEMENT GÉOMÉTRIQUE DÉGRESSIF

Définition: Lorsque la valeur d'immobilisation d'un bien décroît selon un taux d'amortissement constant basé sur la valeur résiduelle de la période précédente (donc en quelque sorte comme un intérêt composé à taux négatif), nous parlons alors "**d'amortissement géométrique dégressif simple**".

Ainsi, la valeur du bien après n années est définie par:

$$V_n = V_0 (1 - t\%)^n \quad (67.161)$$

avec donc:

$$t\% = 1 - \sqrt[n]{\frac{V_n}{V_0}} \quad (67.162)$$

Remarque: Nous constatons que la valeur de $t\%$ étant comprise dans l'intervalle entre $[0,1[$ la limite de $(1 - t\%)^n$ quand n tend vers l'infini n'est jamais nulle. Ainsi, la valeur résiduelle ne le sera jamais non plus !

Sachant par définition de cet amortissement que:

$$A_k = V_{k-1} t\% \quad (67.163)$$

nous obtenons:

$$A_k = V_0 (1 - t\%)^{k-1} t\% \quad (67.164)$$

En injectant l'expression du taux dans la relation précédente, nous obtenons:

$$\begin{aligned} A_k &= V_0 \left(1 - \left(1 - \sqrt[n]{\frac{V_n}{V_0}} \right) \right)^{k-1} \left(1 - \sqrt[n]{\frac{V_n}{V_0}} \right) = V_0 \left(\sqrt[n]{\frac{V_n}{V_0}} \right)^{k-1} \left(1 - \sqrt[n]{\frac{V_n}{V_0}} \right) = V_0 \left(\frac{V_n}{V_0} \right)^{\frac{k-1}{n}} \left(1 - \sqrt[n]{\frac{V_n}{V_0}} \right) \\ &= V_0 \left(\left(\frac{V_n}{V_0} \right)^{\frac{k-1}{n}} - \left(\frac{V_n}{V_0} \right)^{\frac{1}{n}} \left(\frac{V_n}{V_0} \right)^{\frac{k-1}{n}} \right) = V_0 \left(\left(\frac{V_n}{V_0} \right)^{\frac{k-1}{n}} - \left(\frac{V_n}{V_0} \right)^{\frac{k}{n}} \right) \end{aligned} \quad (67.165)$$

Nous remarquons donc que les valeurs d'amortissement ne nécessitent pas de connaître le taux de manière explicite. Il suffit de connaître la valeur finale et initiale. C'est justement pour cela que la fonction DB() de la version française de Microsoft Excel 11.8346 ne demande pas le taux d'amortissement.

À remarquer que ce taux constant ainsi que la valeur finale sont parfois imposés par la législation dans certains pays (comme c'est le cas en Suisse par exemple) jusqu'à une valeur résiduelle nulle ou non nulle elle aussi définie par la législation!

Remarque: L'amortissement géométrique dégressif convient particulièrement aux biens ayant une très forte dépréciation les premières années.

8.4. CHOIX D'INVESTISSEMENTS

Par définition, un investissement est l'acquisition ou le développement d'un bien (quelle que soit sa forme, matérielle ou non) par une entreprise, une collectivité ou un individu. Un investissement implique dans le cadre économique simple:

1. Une dépense immédiate, payable en une ou plusieurs fois
2. Des entrées futures, appelées "**cash-flows**"
3. Une valeur résiduelle

Il existe plusieurs critères et techniques pour les choix d'investissements, que nous présenterons ci-après, qui permettent d'opter pour un investissement A ou B: celui de la "**valeur actuelle nette**" (VAN), celui du "**taux interne de rentabilité**" (TRI) ou encore celui de "**délai de récupération**" (DR) appelé aussi "**délai de recouvrement**".

Remarque: Il ne faut pas oublier aussi que les techniques de décisions (cf. chapitre de Théorie Des Jeux Et De La Décision) ont une énorme importance lorsque les sommes considérées atteignent des valeurs non négligeables.

8.4.1. VALEUR ACTUELLE NETTE

Comme nous l'avons spécifié plus haut, un investissement implique trois points. Ce qui intéresse bien évidemment l'investisseur, c'est qu'en valeur actuelle, l'investissement rapporte plus que ce qui est dépensé.

Voyons une situation-type: Une entreprise souhaite acquérir une nouvelle machine valant 6'000.-, ce qui devrait permettre d'abaisser les coûts de production de 1'000.- durant 5 ans. Nous estimons que dans 5 ans, la valeur résiduelle de cette machine sera de 3'000.-. Doit-on acheter cette machine si cet investissement peut être financé par un emprunt à 10%?

Quelles informations avons-nous ici?

1. La dépense immédiate $V_0 = 6'000$
2. La valeur finale ou résiduelle du bien d'équipement après 5 ans $V_n = 3'000$

3. Les cash-flows de chaque année $C_k = 1'000$ (qui sont constants sur toute la période dans cet exemple)
4. Le taux d'intérêt (taux géométrique moyen du marché) de l'emprunt correspondant $t\% = 10\%$

Quelles informations, ou questions intéressantes, financièrement parlant, pouvons-nous poser par rapport aux données ci-dessus? :

1. Quel serait le capital initial qui au taux du marché nous permettrait de retirer 1'000.- par année pendant 5 ans (jusqu'à ce que l'on solde le compte)? :

$$\frac{C_1}{1+t\%} + \frac{C_2}{(1+t\%)^2} + \dots + \frac{C_n}{(1+t\%)^n} \quad (67.166)$$

ce qui s'écrit si $C_k = C$ (nous retrouvons la relation de la rente certaine postnumerando vue dans le chapitre d'Économie):

$$\frac{C}{1+t\%} + \frac{C}{(1+t\%)^2} + \dots + \frac{C}{(1+t\%)^n} = C a_{\overline{n}|} \quad (67.167)$$

Dans notre exemple cette somme (après un petit calcul) revient à environ 3'790.-.

En d'autres termes, il nous suffirait de mettre en épargne 3'790.- pendant les mêmes 5 ans, pour en retirer 1'000.- par année jusqu'à solder le compte. Donc pour l'instant, un investissement de 3'790 pour économiser (gagner) 1'000.- par année semble beaucoup plus favorable qu'en dépenser 6'000.- pour le même retour, sur la même durée...

Déjà là, nous pouvons dire que l'achat de la machine est défavorable.

Mais il ne faut pas oublier aussi un deuxième facteur... la valeur résiduelle de notre machine!!!

2. Quel serait le capital initial qui au taux du marché nous rapporterait une valeur équivalente à la valeur résiduelle de notre machine (c'est une valeur immobilisée au même titre qu'une épargne, donc nous pouvons nous intéresser à ce qu'il adviendrait si cette somme provenait d'une épargne) ?

$$\frac{V_n}{(1+t\%)^n} \quad (67.168)$$

Dans notre exemple, cette somme revient environ à 1'862.-

En d'autres termes, il nous suffirait de mettre en épargne 1'862.- pendant les mêmes 5 ans, pour obtenir une somme égale à la valeur résiduelle de notre machine. Et alors?

Eh bien la conclusion est assez simple.... La somme:

$$\sum_{k=1}^n \frac{C_k}{(1+t\%)^k} + \frac{V_n}{(1+t\%)^n} \quad (67.169)$$

représente le retour sur la base d'une épargne initiale pour obtenir, par rapport aux informations de valeur résiduelle et de cash-flow, la somme finale équivalente à l'achat de notre machine. Or, dans cet exemple, cette somme nous donne environ 5'652.-.

Ce résultat est important car il est à comparer avec l'investissement que nous voulions faire initialement. Deux options s'offrent donc à nous:

1. Acheter la machine à 6'000.- avec les cash-flows et les valeurs résiduelles que nous connaissons
2. Épargner 5'652.- pendant la même période, avec les mêmes cash-flows pour nous retrouver avec une épargne finale qui devrait être équivalente à celle de la valeur résiduelle de notre machine.

Or, que pouvons-nous conclure de notre exemple? Eh, bien simplement que nous nous trouvons dans un cas défavorable si:

$$\sum_{k=1}^n \frac{C_k}{(1+t\%)^k} + \frac{V_n}{(1+t\%)^n} < V_0 \Leftrightarrow \sum_{k=1}^n \frac{C_k}{(1+t\%)^k} + \frac{V_n}{(1+t\%)^n} - V_0 < 0 \quad (67.170)$$

En d'autres termes, pour le financier (ou chef de projet), le calcul intéressant à faire est le suivant:

$$VAN = \sum_{k=1}^n \frac{C_k}{(1+t\%)^k} + \frac{V_n}{(1+t\%)^n} - V_0 \quad (67.171)$$

qui:

1. Lorsqu'il est négatif correspond à un investissement qu'il vaut mieux éviter
2. Lorsqu'il est nul est un investissement indécidable
3. Lorsqu'il est positif est un bon investissement

Dans certains cas scolaires et livres de gestion de projets, la relation de la VAN ci-dessus est simplifiée, car ils supposent que seul un investissement... est fait en début de projet. Ainsi, nous avons:

$$VAN = \sum_{k=1}^n \frac{C_k}{(1+t\%)^k} - V_0 \quad (67.172)$$

Enfin, précisons que le rapport:

$$\frac{1}{(1+t\%)^n} \quad (67.173)$$

est souvent appelé "**coefficient d'actualisation**" de la n -période.

Ainsi, la VAN est utilisée comme critère décisionnel dans les grandes entreprises pour chiffrer l'apport spécifique d'un projet au résultat financier de l'entreprise, en tenant compte du coût du capital utilisé via le taux d'actualisation.

Donc dans le cas d'un choix entre plusieurs investissements, nous choisirons celui dont la VAN est la plus grande. Si les cash-flows sont non déterministes il faudra alors calculer l'espérance et la variance de la VAN. Les spécialistes abrègent souvent le calcul de l'espérance de la VAN par l'abréviation VANe pour "**valeur actuelle nette espérée**" ou plus fréquemment en anglais "**expected net present value**" (eN.P.V.) ou encore "**risk-adjusted net present value**" (rN.P.V.). Le lecteur pourra trouver un exemple de calcul de VANe dans le serveur d'exercice (sinon, sur demande, nous intégrerons directement l'exemple sur la présente page).

Remarque: La VAN est aussi souvent appelée "**quasi-rente actualisée**" ou encore en anglais "**net present value**" (N.P.V.). Nous trouvons également souvent la dénomination "**méthode des cash-flows actualisés**".

Si nous considérons en avenir certain avec un investissement initial unique de 10'000.- avec un taux d'actualisation constant de 10% et un cash-flow parfaitement périodique de 3'250.-, 3'750.-, 4'250.-, 4'750.- nous avons la représentation tabulaire traditionnelle:

Année	Action	Cash-Flow	C.A.	C.F.A.
0	Investissement	-10'000	1	-10'000
1	Entrée	3'250	0.909	2'955
2	Entrée	3'750	0.826	3'099
3	Entrée	4'250	0.751	3'193
4	Entrée	4'750	0.683	3'244
SOMME	F.N.T:	6'000	F.N.T.A:	2'491

Tableau: 67.12 - Actualisation détaillée sous forme comptable

où dans le tableau ci-dessus F.N.T. signifie "**Fond Net de Trésorerie**", F.N.T.A. "**Fond Net de Trésorerie Actualisé**", C.A. "**Coefficient d'Actualisation**" et C.F.A. "**Cash-Flow Actualisé**".

Ainsi dans ce tableau, l'investissement rapporte 2'491.- de plus qu'une opération de placement à 10% après 4 ans en avenir certain (univers déterministe).

8.4.2. TAUX DE RENTABILITÉ INTERNE

Définition (technique): Le "**taux de rentabilité interne**" (TRI), appelé aussi parfois "**taux limite de rentabilité**" est le taux d'actualisation $t\%$ pour lequel la valeur actualisée des rentrées nettes de fonds résultant d'un projet d'investissement est égale à la valeur actualisée des décaissements requis pour réaliser cet investissement.

En d'autres termes, cela revient à se demander quel est le taux moyen géométrique du marché pour lequel la V.A.N. du projet est nulle. Soit à satisfaire la relation:

$$\begin{aligned}
 V.A.N. &= \sum_{k=1}^n \frac{C_k}{(1+t\%)^k} + \frac{V_n}{(1+t\%)^n} - V_0 = 0 \\
 \sum_{k=1}^n \frac{C_k}{(1+t\%)^k} + \frac{V_n}{(1+t\%)^n} &= V_0
 \end{aligned}
 \tag{67.174}$$

qui ne peut que se calculer rapidement avec des outils informatiques (l'outil Cible dans Microsoft Excel 11.8346 par exemple).

Donc entre deux investissements, nous choisissons dans les entreprises celui dont le TRI est le plus élevé et satisfait aux contraintes internes.

Ce type de calcul s'applique donc sur le retour sur projets contre investissements sur le marché et non pas au retour sur projets contre exploitation. Ainsi, il s'agit d'un outil calculatoire d'aide à la décision purement financier et non industriel ou commercial.

8.4.3. DÉLAI DE RÉCUPÉRATION ET D'AMORTISSEMENT

Le "**délai de récupération**" (DR), appelé aussi "**délai de recouvrement**" ou "**pay back**" (en anglais) est un autre indicateur pour l'aide à la décision dans le cadre des choix d'investissements de projets et plus simple à l'utilisation (et à la compréhension) que la VAN.

Cet indicateur a pour simple et seul objectif de montrer quand, dans le temps, l'investissement sera remboursé. En d'autres termes, il indique le nombre de périodes nécessaires pour que la somme des cash-flows couvre l'investissement initial. C'est une information très simple à déterminer qui revient à chercher le plus petit entier p tel que:

$$V_0 \leq C_1 + C_2 + \dots + C_p \quad (67.175)$$

ou encore:

$$V_0 \leq \sum_{k=1}^p C_k \quad (67.176)$$

En d'autres termes, c'est le moment d'un projet où les cash-flows équilibrent l'investissement initial.

Définition: Le "**délai d'amortissement**" est le nombre de périodes nécessaires tel que:

$$V_0 \leq \frac{C_1}{1+t\%} + \frac{C_2}{(1+t\%)^2} + \dots + \frac{C_p}{(1+t\%)^p} \quad (67.177)$$

ou encore:

$$V_0 \leq \sum_{k=1}^p \frac{C_k}{(1+t\%)^k} \quad (67.178)$$

9. THÉORIES DES FILES D'ATTENTE

Les théories des files d'attente sont des outils extrêmement puissants et vastes (une présentation complète nécessite au bas mot 300 pages A4) permettant de prendre en compte et de modéliser les goulots d'étranglement dans les processus des entreprises soit au niveau de la logistique, des centrales téléphoniques, des requêtes sur les serveurs, des caisses de grands magasins ou encore dans les toilettes des grands stades sportifs (...) en fonction des hypothèses et contraintes de départ.

Généralement, les clients voient évidemment dans l'attente une activité sans valeur ajoutée et, s'ils attendent trop longtemps, ils associent cette perte de temps à une mauvaise qualité de service. De la même façon, au sein de l'entreprise, des employés inoccupés ou des équipements inutilisés représentent des activités sans valeur ajoutée. Pour éviter ces situations, la majorité des entreprises ont mis en place des processus d'amélioration continue dont le but ultime est l'élimination de toute forme de gaspillage, notamment l'attente. Tous ces exemples révèlent l'importance de l'analyse des files d'attente. A remarquer que si une entreprise ou administration n'a pas les moyens financiers ou intellectuels de faire de la théorie des files d'attente, le respect minimum envers les clients est au moins d'indiquer à l'aide d'un affichage peu onéreux, le temps d'attente ou de mettre à disposition des moyens de distractions (télévision, journaux ou magazines)!

Système	Clients	Service
Comptoir-réception	Personnes	Réceptionnistes
Atelier de réparation	Machines	Techniciens
Garages	Camions	Mécaniciens
Hôpital	Patients	Infirmiers
Ordinateur	Tâches	Processeurs, disques, rubans
Aéroport	Appareils	Pistes d'envols
Réseau routier	Véhicules	Feux de circulation
Atelier de fabrication	Tâches	Machines/Ouvriers
Téléphone	Appels	Échangeurs
Transport en commun	Voyageurs	Autobus/Métros
Buanderie	Lignes	Laveuses/Sécheuses
...	...	...

Tableau: 67.13 - Exemples de files d'attente typiques

Ces théories se révèlent notamment utiles pour justifier des investissements, des embauches ou des achats d'équipements. De façon plus générale, elle est une partie intégrante des techniques mathématiques de gestion lorsqu'il est nécessaire de rechercher un optimum économique entre des coûts d'attente et des coûts de service d'un système.

La problématique type dans les entreprises peut s'exprimer ainsi:

- Quel est le nombre optimal de stations/terminaux à mettre en service permettant de traiter la demande tout en évitant une file d'attente trop importante et le départ de certains clients?
- Quel est le temps d'attente moyen d'un client/employé devant la station/terminal?
- Quel est le nombre moyen de clients en attente dans la file?

Ces questions permettent d'exprimer des objectifs en matière de qualité et de niveau de service:

1. Un temps d'attente moyen ou médian dans le service ou dans la file à ne pas dépasser
2. Une probabilité d'attente maximale
3. Un nombre de clients moyen en attente donné

En pratique, lorsque le client est externe à l'entreprise, le coût d'attente est difficile à évaluer, car il s'agit d'un impact plutôt que d'un coût pouvant être comptabilisé. Cependant, on peut considérer les temps d'attente comme un critère de mesure du niveau de service. Le gestionnaire décide du temps d'attente acceptable et il met en place la capacité susceptible de fournir ce niveau de service.

Lorsque le client est interne à l'entreprise, nous pouvons établir directement certains coûts se rapportant au temps d'attente des clients (machines). Par ailleurs, il ne faut pas conclure trop rapidement que pour l'entreprise, le coût du temps d'attente d'un employé qui attend est égal à son salaire durant le temps d'attente. Cela impliquerait que la baisse nette des gains de l'entreprise, du fait de l'inactivité d'un employé, est égale au salaire de ce dernier, ce qui, a priori, n'est pas évident. L'employé, qu'il travaille ou qu'il attende, reçoit le même salaire. Par contre, sa contribution aux gains de l'entreprise est réellement perdue, car la productivité baisse. Quand un opérateur de machine est inactif parce qu'il attend, sa force productive (qui peut comprendre, outre son salaire, une proportion des coûts fixes de l'entreprise) est perdue. En d'autres termes, il faut tenir compte non pas de la ressource physique en attente, mais plutôt de la valeur (coût) de toutes les ressources économiques inactives, et évaluer ensuite la perte de profit à partir de la perte de productivité. L'objectif de l'analyse des files d'attente est de trouver un compromis entre le coût associé à la capacité de service et le coût d'attente des clients.

Ces théories font donc appel à des méthodes statistiques et algébriques que nous avons étudiées dans les chapitres de Statistiques et de Théorie Des Graphes. Elles n'en sont alors que plus passionnantes.

Pour présenter le sujet, plutôt que de faire une généralisation abstraite, nous avons choisi de développer la théorie autour d'un exemple concret et classique qui est le télétrafic. Une généralisation à tout autre cas d'étude se faisant ensuite relativement facilement par analogie.

Considérons donc une centrale téléphonique regroupant les lignes d'un ensemble d'immeubles d'une ville et ne possédant pas autant de lignes allant vers le réseau que de lignes allant vers les différents particuliers qu'il dessert.

Nous pouvons donc légitimement nous demander de combien de lignes nous avons besoin pour desservir tous ces abonnés.

Pour dimensionner son réseau, un opérateur va devoir calculer le nombre de ressources à mettre en oeuvre pour qu'avec une probabilité extrêmement proche de 1, un usager qui décroche son téléphone puisse disposer d'un circuit. Pour cela, il va falloir développer quelques relations de probabilité de blocage. Ces relations vont demander une modélisation statistique des instants de début et de fin d'appels ainsi que des durées de ces appels. Les paragraphes qui suivent vont donc introduire les lois de probabilités utilisées pour ces dimensionnements.

Enfin, avant de commencer, nous souhaitons mettre à disposition le tableau récapitulatif ci-dessous des notations les plus importantes que le lecteur découvrira au fur et à mesure de sa lecture et auquel il pourra se reporter en cas de confusion:

Variables	Information	Unité
λ	Flux d'arrivée de clients dans la file d'attente (communication), appelé également "taux moyen d'arrivée des appels", ou encore "fréquence moyenne d'arrivées". L'inverse λ^{-1} donne le temps moyen entre arrivées (appels) dans la file d'attente.	$[h^{-1}]$
μ	Flux de départ de clients (communication) correspondant au taux de traitement. L'inverse μ^{-1} donne le temps moyen d'attente pendant le service (donc une fois arrivé en fin de file d'attente) appelé aussi "temps moyen de service".	$[h^{-1}]$
A ou ρ	Taux d'utilisation du service (par unité de serveur). Assimilé au concept de trafic (un peu abusivement) ou de charge. Est égal au rapport λ / μ et doit être strictement inférieur à 1 pour éviter l'engorgement.	-
C	Nombre de clients au total dans le système	-
C_Q	Nombre de clients en attente dans la queue	-
C_s	Nombre de clients en service (traitement)	-
T	Temps d'attente dans le système	$[s]$
T_Q	Temps d'attente dans la queue	$[s]$
p_k	Probabilité d'avoir k clients dans le système	-

Tableau: 67.14 - Notations conventionnelles des files d'attente

et précisons que les paramètres à prendre en compte sont souvent:

- Le type de processus d'arrivée des clients (données)
- La distribution statistique du temps de service (traitement)
- Le nombre de caisses (serveurs)
- La capacité du système (souvent supposée infinie dans la pratique...)
- La taille de la population
- Le discipline du service

9.1. MODÉLISATION DES DURÉES D'ARRIVÉES M/M/...

Dans ce modèle $M/M/...$ dont nous expliquerons l'origine de la notation plus tard, nous considérons des appels qui débuteraient de manière aléatoire. Prenons ensuite un intervalle de temps t et divisons cet intervalle en n sous-intervalles de durée t/n .

Nous choisissons n suffisamment grand pour que les hypothèses suivantes soient respectées:

H1. Une seule arrivée d'appel peut survenir dans un intervalle t/n

H2. Les instants d'arrivée d'appels sont indépendants les uns des autres (le taux d'arrivée n'est pas influencé par le nombre d'appels provenant de la population). Ce qui présuppose une population infinie.

H3. La probabilité qu'un appel arrive dans un sous-intervalle donné est proportionnelle à un facteur constant près à la durée du sous-intervalle.

Nous écrivons alors la probabilité de un appel p_1 dans un sous-intervalle (1) de la manière suivante:

$$p_1(1) = \lambda \frac{t}{n} \quad (67.179)$$

où le 1 en indice du p représente donc l'analyse sur 1 appel, le 1 entre parenthèses le fait que l'analyse se fait sur 1 sous-intervalle et enfin le terme λ représente le coefficient de proportionnalité entre la probabilité et la durée t/n du sous-intervalle.

L'hypothèse de départ consistant à considérer comme nulle la probabilité d'avoir plusieurs appels dans un sous-intervalle s'écrit alors:

$$p_2(1) + p_3(1) + \dots + p_n(1) + \dots = \sum_{k=2}^{+\infty} p_k(1) = 0 \quad (67.180)$$

La probabilité de n'avoir aucun appel durant un sous-intervalle de temps t/n s'écrit donc:

$$p_0(1) = 1 - \sum_{k=1}^{+\infty} p_k(1) \quad (67.181)$$

En développant, nous obtenons:

$$p_0(1) = 1 - p_1(1) - \sum_{k=2}^{+\infty} p_k(1) \quad (67.182)$$

et en utilisant la propriété énoncée juste au-dessus:

$$p_0(1) = 1 - p_1(1) \quad (67.183)$$

La probabilité d'avoir k arrivées d'appels durant n intervalles de temps s'obtient alors en considérant le nombre de manières de choisir k intervalles parmi n ... (puisque'il ne peut y avoir plus d'un appel par intervalle).

Pour chacune de ces solutions, nous aurons alors forcément k intervalles avec une arrivée d'appel et $n-k$ intervalles avec aucune arrivée d'appel. Nous avons vu dans le chapitre de Statistique que la loi qui permettait d'obtenir la probabilité de choisir un certain arrangement d'issues binaires parmi un nombre total d'issues était la loi de Bernoulli donnée par:

$$P = p^k q^{n-k} \quad (67.184)$$

Il vient donc dans notre cas de figure que la probabilité d'une des solutions sera:

$$p_1(1)^k p_0(1)^{n-k} \quad (67.185)$$

La probabilité globale s'obtient en sommant les probabilités de tous les cas, ce qui nous donne la loi binomiale (cf. chapitre de Statistiques):

$$p_k(n) = \frac{n!}{k!(n-k)!} p_1(1)^k p_0(1)^{n-k} = C_k^n p_1(1)^k p_0(1)^{n-k} \quad (67.186)$$

Ou encore, en remplaçant les probabilités par leurs valeurs en fonction de λ, t, n :

$$p_k(n) = C_k^n \left(\lambda \frac{t}{n} \right)^k \left(1 - \lambda \frac{t}{n} \right)^{n-k} \quad (67.187)$$

La limite de la probabilité $p_k(n)$ lorsque n tend vers l'infini va être égale à la probabilité d'avoir k arrivées d'appel durant un intervalle de temps t . Nous notons p_k cette probabilité:

$$p_k = \lim_{n \rightarrow +\infty} p_k(n) \quad (67.188)$$

En reprenant alors les différents termes de l'expression de $p_k(n)$, il vient:

$$\left(1 - \lambda \frac{t}{n} \right)^{n-k} = \left(e^{\ln \left(1 - \lambda \frac{t}{n} \right)} \right)^{n-k} = e^{(n-k) \ln \left(1 - \lambda \frac{t}{n} \right)} \quad (67.189)$$

En utilisant les développements de Taylor (cf. chapitre de Suites Et Séries):

$$\ln(1+x) = \sum_{n=1}^{+\infty} \frac{(-1)^{n+1}}{n} x^n \quad (67.190)$$

Soit en ne prenant que le premier terme, c'est-à-dire en considérant x très petit:

$$\ln(1+x) \cong x \quad (67.191)$$

Donc:

$$e^{(n-k) \ln \left(1 - \lambda \frac{t}{n} \right)} \underset{n \rightarrow \infty}{\cong} e^{(n-k) \left(-\lambda \frac{t}{n} \right)} = e^{-\lambda t - \frac{k}{n} \lambda t} \underset{n \rightarrow \infty}{\cong} e^{-\lambda t} \quad (67.192)$$

et pour la dernière partie:

$$\begin{aligned} C_k^n \left(\lambda \frac{t}{n} \right)^k &= \frac{n!}{k!(n-k)!} \frac{(\lambda t)^k}{n^k} = \frac{(\lambda t)^k}{k!} \frac{n!}{(n-k)! n^k} \\ &= \frac{(\lambda t)^k}{k!} \frac{n(n-1)(n-2)\dots(n-k+1)}{n^k} \underset{n \rightarrow +\infty}{\cong} \frac{(\lambda t)^k}{k!} \end{aligned} \quad (67.193)$$

d'où après regroupement:

$$p_k \equiv \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad (67.194)$$

Cette fonction de distribution est donc extrêmement importante, car elle représente la probabilité d'observer k arrivées d'appels dans un intervalle de durée t (ou le nombre de clients qui se trouvent devant une caisse dans un intervalle de durée t) et il s'agit donc d'une distribution de Poisson (cf. chapitre de Statistiques). Dans la pratique, il faut donc s'assurer avant d'utiliser les relations qui vont suivre que cette distribution soit bien respectée (avec un test du Khi-deux typiquement).

Considérons maintenant la fonction de répartition, appelée "**fonction de répartition d'Erlang**", qui pour un intervalle de temps t donné donne la probabilité cumulée d'avoir un nombre k d'arrivées supérieur ou égal à n :

$$F(t, k) = \sum_{k=n}^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad (67.195)$$

Si nous dérivons cette fonction par rapport à t cela donne la "**fonction de distribution d'Erlang**":

$$\begin{aligned} \frac{dF(t, k)}{dt} &= \frac{d}{dt} \sum_{k=n}^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t} = \sum_{k=n}^{\infty} \frac{k \lambda (\lambda t)^{k-1}}{k!} e^{-\lambda t} - \sum_{k=n}^{\infty} \frac{(\lambda t)^k}{k!} \lambda e^{-\lambda t} \\ &= \sum_{k=n}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} \lambda e^{-\lambda t} - \sum_{k=n}^{\infty} \frac{(\lambda t)^k}{k!} \lambda e^{-\lambda t} = \lambda e^{-\lambda t} \left(\sum_{k=n}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} - \sum_{k=n}^{\infty} \frac{(\lambda t)^k}{k!} \right) \\ &= \frac{(\lambda t)^{n-1}}{(n-1)!} \lambda e^{-\lambda t} \end{aligned} \quad (67.196)$$

et donc nous retrouvons une forme particulière de la fonction de distribution de la loi Gamma vue dans le chapitre de Statistiques. Rappelons au passage l'équivalence déjà présentée dans le chapitre de Statistiques dans la version française de Microsoft Excel 14.0.6123 entre la loi de Poisson, la loi du Khi-2 et la loi Gamma:

$$\begin{aligned} &= \text{LOI.POISSON.N}(x \in \mathbb{N}; \mu; \text{VRAD}) \\ &= 1 - \text{LOI.KHIDEUX.N}(2 * \mu; 2 * (x + 1); \text{VRAD}) \quad (67.197) \\ &= 1 - \text{LOI.GAMMA.N}(\mu; x + 1; \text{VRAD}) \end{aligned}$$

Il s'ensuit par analogie avec la forme générale de la loi de Poisson que le paramètre λ est le taux moyen d'arrivée (taux moyen d'apparition) d'appels par unité de temps (ou en anglais "**Poisson Arrivals See Time Average**": PASTA...). Typiquement il s'agira d'un nombre moyen d'appels par seconde (voir les estimateurs de la loi de Poisson dans le chapitre de Statistiques).

Ainsi, nous avons pour espérance et variance (cf. chapitre de Statistiques) du nombre d'appels:

$$E[k] = \lambda \cdot t \quad \text{Var}(k) = \lambda \cdot t \quad (67.198)$$

 Exemple:

Une TPE souhaitant mettre en place une hotline estime qu'au début elle recevra par journée de 8 heures, 4 appels téléphoniques (soit une probabilité de 1 chance sur 2 d'avoir un appel par heure et donc un taux moyen λ de 0.5 appels par heure). Alors la probabilité qu'elle reçoive exactement 4 appels (k) par jour et au moins 4 appels (k) par jour selon le modèle théorique de la théorie des files d'attente est de:

$$p_k (k = 4) \cong \frac{(0.5 \cdot 8)^4}{4!} e^{-0.5 \cdot 8} = \text{POISSON}(4; 0.5 \times 8; \text{FAUX}) = 19.55\% \quad (67.199)$$

$$p_k (k \leq 4) \cong \sum_{k=0}^4 \frac{(0.5 \cdot 8)^k}{k!} e^{-0.5 \cdot 8} = \text{POISSON}(4; 0.5 \times 8; \text{VRAI}) = 62.88\%$$

où nous avons utilisé la fonction POISSON() intégrée dans la version française de Microsoft Excel 14.0.6123.

Maintenant, introduisons la variable aléatoire τ représentant le temps séparant deux arrivées d'appel. Nous définissons pour cela la probabilité $A(t)$ qui est la probabilité que le temps τ soit inférieur ou égal à une valeur t :

$$A(t) = P(\tau \leq t) \quad (67.200)$$

Nous avons donc:

$$A(t) = 1 - P(\tau > t) \quad (67.201)$$

Or, $P(\tau > t)$ représente la probabilité cumulée qu'il n'y ait aucune arrivée d'appels durant un temps t . Cette probabilité cumulée a justement été établie plus haut:

$$P(\tau > t) = p_0 = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t} \quad (67.202)$$

Nous en déduisons:

$$A(t) = 1 - e^{-\lambda t} \quad (67.203)$$

Il s'agit donc de la fonction de répartition d'une loi exponentielle! Nous pouvons aussi introduire la densité de probabilité de la variable aléatoire τ . Nous obtenons ainsi:

$$a(t) = \frac{\partial A(t)}{\partial t} = \lambda e^{-\lambda t} \quad (67.204)$$

Remarque: Rappelons que dans le chapitre de Statistiques, nous avons souvent fait la démarche inverse. C'est-à-dire compte tenu d'une densité de probabilité $a(t)$ nous cherchions la fonction de répartition $A(t)$ via une intégration sur le domaine de définition de la variable aléatoire.

La densité de probabilité permet donc de calculer la durée moyenne entre deux arrivées d'appel:

$$\bar{\tau} = E(\tau) = \int_0^{+\infty} t \cdot a(t) dt = \int_0^{+\infty} \lambda t e^{-\lambda t} dt \quad (67.205)$$

En intégrant par parties, il vient:

$$E(\tau) = \lambda t \cdot \frac{-1}{\lambda} e^{-\lambda t} \Big|_0^{+\infty} + \int_0^{+\infty} \lambda \frac{-1}{\lambda} e^{-\lambda t} dt = - \int_0^{+\infty} e^{-\lambda t} dt = \frac{1}{\lambda} \quad (67.206)$$

Nous obtenons ainsi, que pour un taux d'arrivée de λ appels par seconde, le temps moyen entre appel est égal à $1/\lambda$ (résultat relativement logique mais encore fallait-il le démontrer rigoureusement).

Effectivement, si nous avons λ qui vaut 2 appels par heure, le temps moyen d'arrivée est bien de 0.5 heures (1/2) entre appels.

Supposons maintenant qu'aucun appel ne soit arrivé jusqu'à un temps t_0 et que nous souhaitions calculer la probabilité qu'un appel arrive durant une durée t après le temps t_0 .

Nous devons donc calculer la probabilité d'avoir une durée entre deux appels inférieure à $t + t_0$ tout en étant supérieure à t_0 .

Cette probabilité s'écrit $P(\tau \leq t + t_0 \mid \tau > t_0)$. En utilisant la formule de Bayes (cf. chapitre de Probabilités):

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} \quad (67.207)$$

mais avec les notations idoines, il vient:

$$P(\tau \leq t + t_0 \mid \tau > t_0) = \frac{P(t_0 < \tau \leq t + t_0)}{P(\tau > t_0)} \quad (67.208)$$

Cette probabilité peut encore s'écrire:

$$P(\tau \leq t + t_0 \mid \tau > t_0) = \frac{P(\tau \leq t + t_0) - P(\tau \leq t_0)}{P(\tau > t_0)} = \frac{P(\tau \leq t + t_0) - P(\tau \leq t_0)}{1 - P(\tau \leq t_0)} \quad (67.209)$$

En reprenant les expressions des différentes probabilités:

$$\begin{aligned} P(\tau \leq t + t_0 \mid \tau > t_0) &= \frac{\left(1 - e^{-\lambda(t+t_0)}\right) - \left(1 - e^{-\lambda t_0}\right)}{1 - \left(1 - e^{-\lambda t_0}\right)} = \frac{-e^{-\lambda(t+t_0)} + e^{-\lambda t_0}}{e^{-\lambda t_0}} \\ &= \frac{-e^{-\lambda t} e^{-\lambda t_0} + e^{-\lambda t_0}}{e^{-\lambda t_0}} = -e^{-\lambda t} + 1 = 1 - e^{-\lambda t} \end{aligned} \quad (67.210)$$

Nous voyons donc que la probabilité d'apparition d'un appel durant un temps t après une durée t_0 pendant laquelle aucun n'est arrivé est la même que la probabilité d'apparition d'un appel pendant une durée t , indépendamment de ce qui a pu arriver avant. Nous considérons donc que le phénomène (la loi exponentielle) est sans mémoire.

9.2. MODÉLISATION DES DURÉES DE SERVICE M/M/...

Dans ce modèle $M/M/...$ dont nous expliquerons encore une fois l'origine de la notation plus tard, nous allons étudier les lois de probabilité qui modélisent les durées des appels (sous-entendu: en service une fois la fin de la file d'attente atteinte). Pour cela, nous procédons comme précédemment.

Nous considérons donc un intervalle de temps de durée t que nous décomposons en n sous-intervalles de durée t/n . Nous choisissons n de sorte que les hypothèses suivantes restent justifiées:

H1. La probabilité qu'un appel se termine durant un sous-intervalle est proportionnelle à la durée du sous-intervalle.

Nous noterons:

$$\frac{\mu \cdot t}{n} \quad (67.211)$$

cette probabilité, expression dans laquelle μ représente le coefficient de proportionnalité.

H2. La probabilité qu'un appel se termine durant un sous-intervalle est indépendant du sous-intervalle considéré.

Nous introduisons alors la variable aléatoire θ représentant la durée d'un appel et la probabilité $H(t)$ que la durée d'un appel soit inférieure ou égale à t :

$$H(t) = P(\theta \leq t) \quad (67.212)$$

La probabilité qu'un appel ayant débuté à $t = 0$ ne se termine pas avant t s'écrit alors:

$$P(\theta > t) = 1 - H(t) \quad (67.213)$$

Cette probabilité est égale à la probabilité que l'appel ne se termine dans aucun des n sous-intervalles de durée t/n :

$$1 - H(t) = \left(1 - \mu \frac{t}{n}\right)^n \quad (67.214)$$

En faisant tendre n vers l'infini, nous obtenons:

$$1 - H(t) = \lim_{n \rightarrow +\infty} \left(1 - \mu \frac{t}{n}\right)^n = \lim_{n \rightarrow +\infty} e^{n \ln \left(1 - \mu \frac{t}{n}\right)} \cong \lim_{n \rightarrow +\infty} e^{n \left(-\mu \frac{t}{n}\right)} = e^{-\mu t} \quad (67.215)$$

Nous obtenons donc l'expression de la probabilité qu'un appel ait une durée inférieure ou égale à t :

$$H(t) = 1 - e^{-\mu t} \quad (67.216)$$

Nous pouvons en déduire la densité de probabilité associée, notée $h(t)$:

$$h(t) = \frac{\partial H(t)}{\partial t} = \mu e^{-\mu t} \quad (67.217)$$

qui correspond donc à une loi exponentielle (le temps qu'un client emploie pour être servi à une caisse - durée de service - ou pour que son appel soit traité une fois en fin de file d'attente suit donc une loi exponentielle!). Dans la pratique, il faut donc s'assurer avant d'utiliser les relations qui vont suivre que cette distribution soit bien respectée (avec un test du Khi-deux typiquement).

De la même manière que dans les paragraphes précédents, la durée moyenne d'appel (laps de temps moyen entre deux fins d'appels en service) s'obtient en calculant (toujours la même intégration par parties que plus haut):

$$\bar{\theta} = E(\theta) = \int_0^{+\infty} t h(t) dt = \frac{1}{\mu} \quad (67.218)$$

En conclusion, nous avons μ appels qui cessent par seconde et nous avons une durée moyenne d'appel en service égale à:

$$\frac{1}{\mu} \quad (67.219)$$

Effectivement, si nous avons par exemple 2 appels en service qui cessent par heure, cela nous donne une durée moyenne de 0.5 heure (1/2) de service par appel.

Le rapport:

$$A = \frac{\lambda}{\mu} \quad (67.220)$$

représente donc le nombre d'appels qui apparaissent dans la file d'attente sur le nombre d'appels en service qui se terminent pendant un intervalle de temps (temps de service moyen), c'est-à-dire qui représente en fait tout simplement le trafic (ou autrement dit: l'intensité de trafic) ou d'un autre point de vue l'utilisation moyenne du service dont l'unité est le "**Erlang**" (nous rappellerons cette définition plusieurs fois par la suite...).

Exemple:

Dans un magasin, on compte 240 clients par heure et il faut 28 secondes en moyenne pour traiter un client (temps de service moyen). Sachant que la durée de service suit une loi exponentielle et la distribution des arrivées une loi de Poisson, quelle est l'intensité du trafic et le taux d'occupation des caisses si le magasin n'en a que deux?

Le trafic est donc donné par le nombre de clients par heure divisé par le nombre de clients traités par heure. Comme il faut 28 secondes pour en traiter un et qu'il y a 3'600 secondes dans une heure, nous avons l'intensité de trafic suivante:

$$A = \frac{240}{\frac{3'600}{28}} = 1.866 \text{ Erlangs} \quad (67.221)$$

La charge moyenne (ou taux d'occupation) par caisse, sachant qu'il y en a deux, est donc de:

$$\frac{1.866}{2} = 0.93 = 93\% \quad (67.222)$$

C'est cette dernière valeur qui sera prise au final comme valeur du trafic A par station pour les calculs ultérieurs.

Les probabilités d'apparition d'appels et de fin d'appels qui ont été développées dans les paragraphes précédents permettent de modéliser le processus complet d'apparition et de fin d'appels.

9.3. NOTATION DE KENDALL

Une notation a été développée par Kendall pour représenter les files d'attente suite aux développements intensifs (et nombreux) des modèles mathématiques les concernant. La forme réduite de cette notation est:

$$A/B/C$$

où A représente le processus d'arrivée des clients dans le système (loi de distribution), B représente la distribution du temps de service des clients du système (loi de distribution) et C le nombre de serveurs du système.

Par exemple, la notation $M/M/1$ signifie que les clients arrivent dans le système selon une loi de Poisson modélisée par une chaîne de Markov (cf. chapitre de Probabilités), que le temps de traitement suit une distribution de type exponentiel (modélisé par une chaîne de Markov aussi) et le système constitué d'un seul serveur selon le principe du premier arrivé premier servi dans une file d'attente à population infinie et régime permanent. Ce qui correspond respectivement aux trois relations suivantes que nous avons démontrées plus haut pour la probabilité d'arrivée de k appels dans un temps donné:

$$p_k \cong \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad (67.223)$$

la probabilité que le temps de traitement (temps de service) soit égale à une certaine valeur:

$$h(t) = \mu e^{-\mu t} \quad (67.224)$$

et la probabilité d'avoir k clients (communications):

$$p_k = A^k (1 - A) \quad (67.225)$$

La lettre M est utilisée pour indiquer que les processus employés sont du type markovien (sous-entendu exponentiels).

En général, nous utilisons la notation:

$$A/B/C/d:e$$

d représentant le nombre maximum de clients pouvant être présents simultanément dans le système. Ce nombre entier varie entre 1 et l'infini. Quand la capacité mémoire de la file est considérée comme illimitée, ce paramètre est souvent omis. Quant à e , il représente la discipline de service. Par exemple:

- FIFO pour premier arrivé, premier servi (First In-First Out appelé aussi FCFS pour First Come-First Serve). Tous les modèles théoriques que nous allons développer ci-dessous seront de type FIFO!
- LIFO pour dernier arrivé premier servi (Last In-First Out appelé aussi LCFS pour Last Come-First Serve)
- SJF pour servir le travail le plus court d'abord (Shortest Job First)
- SRO pour servir en ordre aléatoire (Service in Random Order)

et encore d'autres...

Ce dernier paramètre est omis si la discipline est FIFO. Ainsi, $M/M/1$ s'écrit sans rien omettre:

$$M / M / 1 / \infty : \text{FIFO}$$

On trouve également des variantes dans la littérature de la définition ci-dessus ce qui n'aide pas toujours à la lecture...

9.4. MODÉLISATION DES ARRIVÉES ET DÉPARTS $M/M/1$

À chaque instant un certain nombre d'appels vont apparaître et d'autres vont se terminer. Nous pouvons donc modéliser l'état où l'on se trouve à un instant donné comme une chaîne d'états.

Chaque état représente le nombre d'appels en cours. Nous concevons donc bien que si, à un instant donné, il y a k appels, nous ne pouvons passer que dans deux états adjacents selon nos hypothèses: $k-1$ et $k+1$.

Nous reconnaissons alors une chaîne de Markov (cf. chapitre de Probabilités). La probabilité de passer d'un état i à un état j pendant un temps dt sera donc notée:

$$p_{ij}(dt) \quad (67.226)$$

Nous introduisons alors les probabilités de transition d'état suivantes:

- Étant dans l'état k , la probabilité $p_{k,k+1}(dt)$ pour passer à l'état $k+1$ durant un intervalle de temps dt sera notée $\lambda_k dt$
- Étant dans l'état k , la probabilité $p_{k,k-1}(dt)$ pour passer à l'état $k-1$ durant un intervalle de temps dt sera notée $\mu_k dt$
- Étant dans l'état $k+1$, la probabilité $p_{k+1,k}(dt)$ pour passer à l'état k durant un intervalle de temps dt sera notée $\mu_{k+1} dt$
- Étant dans l'état $k-1$, la probabilité $p_{k-1,k}(dt)$ pour passer à l'état k durant un intervalle de temps dt sera notée $\lambda_{k-1} dt$

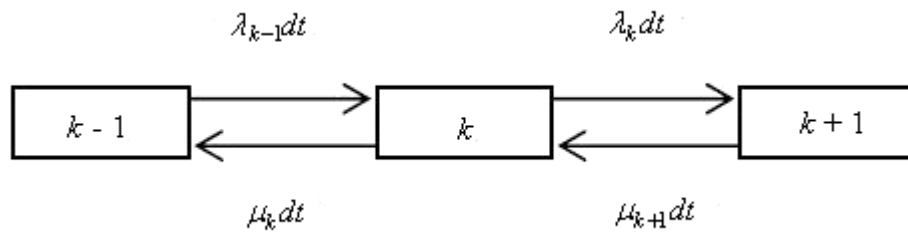


Figure: 67.30 - Schéma de principe

Les grandeurs λ_k et μ_k sont des taux d'arrivée (apparition) et de départ (fin) d'appels du même type que ceux utilisés lors des paragraphes précédents. La seule différence tient au fait que ces taux ont en indice l'état où se trouve le système.

Nous pouvons alors introduire la probabilité d'état, c'est-à-dire la probabilité d'être dans un état k à un instant t . Notons pour cela $p_k(t)$ cette probabilité (à rapprocher de la notation $p_j(n)$ utilisée pour les chaînes de Markov à temps discret dans le chapitre de Probabilités).

La variation de cette probabilité durant un intervalle de temps dt est alors égale à la probabilité de rejoindre cet état en venant d'un état $k-1$ ou d'un état $k+1$ moins la probabilité de quitter cet état pour aller vers un état $k-1$ ou vers un état $k+1$.

Ce qui s'écrit:

$$\frac{dp_k(t)}{dt} = \lambda_{k-1}p_{k-1}(t) + \mu_{k+1}p_{k+1}(t) - (\lambda_k + \mu_k)p_k(t) \quad (67.227)$$

En supposant le système stable, c'est-à-dire en supposant qu'il se stabilise sur des probabilités d'état fixes lorsque le temps tend vers l'infini, nous pouvons écrire que:

$$\lim_{t \rightarrow +\infty} \frac{dp_k(t)}{dt} = 0 \quad (67.228)$$

Nous pouvons alors noter $p_k = p_k(t)$ d'où finalement:

$$\lambda_{k-1}p_{k-1} + \mu_{k+1}p_{k+1} - (\lambda_k + \mu_k)p_k = 0 \quad (67.229)$$

Nous aurions pu introduire cette dernière relation d'une autre manière: Elle exprime simplement le fait que la probabilité de partir d'un état est égale à celle pour y arriver (c'est peut-être plus simple ainsi).

Cette relation est vérifiée pour tout $k \geq 0$ avec les conditions mathématiques suivantes (car sinon ces termes n'ont aucun sens mathématique):

$$p_{-1} = 0, \lambda_{-1} = 0 \quad (67.230)$$

et la condition logique réelle suivante (des appels non encore existants ne peuvent finir...):

$$\mu_0 = 0 \quad (67.231)$$

Remarque: Insistons sur le fait que la stabilité des probabilités signifie qu'il y a une probabilité égale de quitter l'état p_k que de le rejoindre.

En écrivant le système d'équation précédent, nous trouvons:

$$\begin{aligned}
 p_1 &= \frac{\lambda_0}{\mu_1} p_0 \\
 p_2 &= \frac{1}{\mu_2} \left[(\lambda_1 + \mu_1) \frac{\lambda_0}{\mu_1} p_0 - \lambda_0 p_0 \right] = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} p_0 \\
 p_3 &= \frac{1}{\mu_3} \left[(\lambda_2 + \mu_2) \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} p_0 - \lambda_1 \frac{\lambda_0}{\mu_1} p_0 \right] = \frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1} p_0 \\
 &\dots
 \end{aligned} \tag{67.232}$$

Nous trouvons alors assez facilement la forme générale:

$$p_k = \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) p_0 \tag{67.233}$$

Le système se trouvant obligatoirement dans un des états nous avons la relation suivante qui doit obligatoirement être respectée:

$$\sum_{k=0}^{+\infty} p_k = 1 \tag{67.234}$$

En remplaçant avec la relation antérieure:

$$\sum_{k=0}^{+\infty} \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) p_0 = 1 \tag{67.235}$$

Ce qui donne aussi:

$$p_0 + \sum_{k=1}^{+\infty} \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) p_0 = 1 \tag{67.236}$$

et donc:

$$p_0 = \frac{1}{1 + \sum_{k=1}^{+\infty} \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right)} \tag{67.237}$$

Si nous considérons maintenant un système avec une seule ligne et de capacité infinie (régime permanent), les grandeurs λ_k (taux d'arrivée des appels) et μ_k (taux de départ des appels) auront des valeurs identiques pour tout k . C'est-à-dire que nous considérons que le taux d'arrivée ainsi que le taux de départ sont constants quelle que soit la position dans laquelle on se trouve dans la file d'attente. Nous avons alors la dernière relation qui se simplifie:

$$p_0 = \frac{1}{1 + \sum_{k=1}^{+\infty} \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right)} = \frac{1}{1 + \sum_{k=1}^{+\infty} \left(\frac{\lambda}{\mu} \right)^k} \quad (67.238)$$

En utilisant le résultat démontré dans le chapitre sur les Suites Et Séries (série de Gauss) nous avons sous des conditions précises nécessaires de convergence (A doit être strictement plus petit que 1):

$$p_0 = \frac{1}{1 + \sum_{k=1}^{+\infty} \left(\frac{\lambda}{\mu} \right)^k} = \frac{1}{\sum_{k=0}^{+\infty} \frac{1}{\left(\frac{\mu}{\lambda} \right)^k}} = \frac{1}{\frac{1}{1 - \frac{\mu}{\lambda}}} = \frac{1}{\frac{\lambda}{1 - \mu}} = 1 - A \quad (67.239)$$

Puisque k représente le nombre d'appels, cette dernière relation donne la probabilité d'avoir 0 appel (clients) pour un trafic permanent donné A sur la ligne.

En utilisant:

$$p_k = \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) p_0 \quad (67.240)$$

nous avons alors en toute généralité pour ce type de système la probabilité d'avoir k appels en régime permanent qui est donné par:

$$p_k = \left(\frac{\lambda}{\mu} \right)^k (1 - A) = A^k (1 - A) \quad (67.241)$$

Il en vient que l'espérance du nombre de communications (clients) dans le système (dans la queue + en service) est alors par définition de l'espérance:

$$E(C) = \sum_{k=0}^{+\infty} k \cdot p_k = \sum_{k=0}^{+\infty} k A^k (1 - A) = (1 - A) \sum_{k=0}^{+\infty} k A^k \quad (67.242)$$

puisque p_k représente la probabilité pour qu'il y ait à tout instant k appels dans le système (file d'attente + en service). Or, nous avons vu dans le chapitre sur les Suites et Séries que:

$$\sum_{i=0}^n q^i = \frac{1 - q^{n+1}}{1 - q} \quad (67.243)$$

et si q est strictement inférieur à 1 et que n tend vers l'infini, nous avons immédiatement:

$$\sum_{i=0}^{+\infty} q^i = \frac{1}{1-q} \quad (67.244)$$

Si nous dérivons cette dernière relation:

$$\frac{d}{dq} \sum_{i=0}^{+\infty} q^i = \sum_{i=0}^{+\infty} i q^{i-1} = \frac{1}{(1-q)^2} \quad (67.245)$$

et en multipliant par q il vient alors:

$$q \frac{d}{dq} \sum_{i=0}^{+\infty} i q^{i-1} = \sum_{i=0}^{+\infty} i q^i = \frac{q}{(1-q)^2} \quad (67.246)$$

Il vient alors au final pour l'espérance du nombre de clients dans le système:

$$E(C) = (1-A) \sum_{k=0}^{+\infty} k A^k = (1-A) \sum_{k=1}^{+\infty} k A^k = (1-A) \frac{A}{(1-A)^2} = \frac{A}{1-A} \quad (67.247)$$

Si nous souhaitons connaître l'espérance du nombre de clients en attente dans la queue uniquement, il faut bien comprendre qu'à chaque instant nous avons donc une probabilité p_k qu'il y ait k clients dans le système mais comme 1 parmi ceux-ci est alors toujours en service (donc hors de la queue d'attente) il reste que nous en avons toujours $k-1$ réellement en attente. Donc:

$$\begin{aligned} E(C_Q) &= \sum_{k=1}^{+\infty} (k-1) p_k = \sum_{k=1}^{+\infty} k p_k - \sum_{k=1}^{+\infty} p_k = (1-A) \sum_{k=1}^{+\infty} k A^k - (1-p_0) \\ &= \frac{A}{1-A} - A = \frac{A}{1-A} - \frac{A(1-A)}{1-A} = \frac{A-A+A^2}{1-A} = \frac{A^2}{1-A} \end{aligned} \quad (67.248)$$

Connaissant le nombre de communications (ou de clients) que nous avons sur l'unique ligne dans tout le système, nous avons alors l'espérance du temps d'attente (temps d'attente moyen), notée $E(T)$ qui sera donnée par le rapport de l'espérance du nombre de communications (clients) en régime permanent sur la ligne $E(C)$ par le taux d'arrivée des appels.

$$E(T) = \frac{E(C)}{\lambda} \quad (67.249)$$

Ce résultat, appelé "**relation de Little**" (ce dernier ayant démontré rigoureusement que la relation est valable pour n'importe quel type de file d'attente), est intuitif dans le cas présent. Effectivement, prenons un appel type au hasard. Quand il arrive dans le système, il va se trouver statistiquement face à $E(C)$ en train d'attendre. Quand il quittera le système, il y aura été un temps moyen $E(T)$. Donc pendant ce temps moyen, $\lambda E(T)$ appels seront arrivés derrière lui dans le système. En régime permanent, le nombre d'appels laissés derrière lors du départ doit égaler le nombre d'arrivées. D'où l'égalité $\lambda E(T) = E(C)$ dont on déduit alors immédiatement la relation de Little.

Nous avons alors pour l'espérance du temps d'attente dans le système:

$$E(T) = \frac{E(C)}{\lambda} = \frac{\frac{A}{1-A}}{\lambda} = \frac{A}{\lambda(1-A)} = \frac{\frac{\lambda}{\mu}}{\lambda\left(1-\frac{\lambda}{\mu}\right)} = \frac{1}{\frac{\mu}{\lambda}\left(1-\frac{\lambda}{\mu}\right)} = \frac{1}{\mu-\lambda} \quad (67.250)$$

Pour déterminer le temps d'attente dans la queue seule, il suffit de soustraire le temps de traitement/service de par la propriété de linéarité de la moyenne (durée d'appel moyenne en service):

$$E(T_Q) = E(T) - \frac{1}{\mu} = \frac{1}{\mu-\lambda} - \frac{1}{\mu} = \frac{1}{\mu} \left(\frac{1}{1-A} - 1 \right) = \frac{1}{\mu} \left(\frac{1-1+A}{1-A} \right) = \frac{1}{\mu} \left(\frac{A}{1-A} \right) \quad (67.251)$$

Pour résumer, car il y a beaucoup de paramètres et résultats, nous avons donc pour une file d'attente de type $M/M/1$ (selon la notation de Kendall):

Information	$M/M/1$
Probabilité système vide	$1-A$
Probabilité d'attente	A
Nombre moyen de clients dans le système	$E(C) = \frac{A}{1-A}$
Nombre moyen de clients en attente	$E(C_Q) = \frac{A^2}{1-A}$
Nombre moyen de clients en service	A
Temps moyen de séjour dans le système	$E(T) = \frac{E(C)}{\lambda} = \frac{1}{\mu-\lambda}$
Temps moyen d'attente dans la queue	$E(T_Q) = \frac{1}{\mu} \left(\frac{A}{1-A} \right)$
Condition d'atteinte de l'équilibre	$\frac{\lambda}{\mu} < 1$
Probabilité d'avoir k clients	$p_k = A^k (1-A)$

Tableau: 67.15 - Résumé des relations importantes d'une file $M/M/1$

Nous voyons que certaines relations divergent vers l'infini lorsque le trafic permanent A (rapport des entrées sur) tend vers l'unité. Raison pour laquelle nous avons imposé plus haut que ce paramètre soit strictement inférieur à l'unité.

Il faudrait pour chaque type possible de file d'attente faire les démonstrations détaillées des relations correspondantes ce qui est long et laborieux (c'est un métier/spécialisation à part entière et il existe des ouvrages de plus de 400 pages sur le sujet).

Exemple:

Supposons que l'on dispose d'une machine à commande numérique traitant des pièces une à la fois. Supposons que $\lambda = 8$ pièces/heure (nombre de pièces arrivant en moyenne par heure) et que $\mu = 10$ pièces/heure (nombre de pièces sortant en moyenne par heure). Nous avons alors:

$$A = \frac{\lambda}{\mu} = \frac{8}{10} = 80\% \quad (67.252)$$

ce qui correspond au trafic ou taux d'occupation de la machine. Donc il y a 20% de probabilité pour que le système soit vide et 80% de probabilité pour qu'il y ait une attente.

$$E(C) = \frac{A}{1-A} = \frac{0.8}{1-0.8} = 4 \quad (67.253)$$

ce qui correspond donc au nombre moyen de pièces dans le système (machine + en attente).

$$E(C_Q) = \frac{A^2}{1-A} = \frac{0.8^2}{1-0.8} = 3.2 \quad (67.254)$$

ce qui correspond donc au nombre moyen de pièces en attente en dehors de la machine.

$$E(T) = \frac{1}{\mu - \lambda} = \frac{1}{10 - 8} = 0.5 \quad (67.255)$$

ce qui correspond à un temps moyen de séjour de 30 minutes dans le système.

$$E(T_Q) = \frac{1}{\mu} \left(\frac{A}{1-A} \right) = \frac{1}{10} \left(\frac{0.8}{1-0.8} \right) = 0.4 \quad (67.256)$$

ce qui correspond à une attente moyenne de 24 minutes dans la file d'attente.

Et la probabilité qu'il y ait 5 pièces dans le système (exécution + attente):

$$p_k = A^k (1-A) = 0.8^5 (1-0.8) \cong 6.5\% \quad (67.257)$$

9.5. PROBABILITÉ DE MISE EN ATTENTE M/M/k/k (FORMULE D'ERLANG-B)

Nous allons nous intéresser ici à un système disposant de N canaux de communication (chaque canal est censé supporter un débit de un appel avec réponse immédiate). Si les N canaux sont occupés, les appels qui arrivent sont considérés comme perdus (absence de tonalité par exemple). Nous parlons alors de blocage ou ruine du système. Il s'agit donc d'une file d'attente limitée de type $M/M/k/k$ selon la notation de Kendall, appelée également "**système à perte**".

Nous allons chercher à estimer cette probabilité de blocage en fonction du nombre de canaux disponibles et du trafic.

Compte tenu de ce qui a été énoncé sur le caractère sans mémoire du processus d'arrivée d'appels, nous pouvons considérer que la probabilité:

$$p_k = \lambda_k dt \quad (67.258)$$

d'avoir k appels à l'état k est indépendante de l'état du système tel que:

$$p_k = p = \lambda dt \quad (67.259)$$

Ainsi, à chaque état k du système la loi de probabilité de type Poisson est valable. La différence de traitement c'est que plutôt que de considérer des états, nous allons considérer qu'un canal de communication peut être envisagé comme un état propre.

Pour la probabilité de fin d'appel, nous avons par contre:

$$p_k = \mu_k dt = k \mu dt \quad (67.260)$$

Effectivement, cette probabilité traduit juste le fait que si k appels sont en cours chacun a une probabilité μdt de se terminer, d'où la somme qui donne $k \mu dt$. Nous avons alors:

$$\mu_k = k \cdot \mu \quad (67.261)$$

Ainsi, en injectant ces relations dans:

$$p_0 = \frac{1}{1 + \sum_{k=1}^{+\infty} \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right)} \quad (67.262)$$

il vient:

$$p_0 = \frac{1}{1 + \sum_{k=1}^N \left(\prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} \right)} = \frac{1}{1 + \sum_{k=1}^N \left(\left(\frac{\lambda}{\mu} \right)^k \frac{1}{k!} \right)} \quad (67.263)$$

En introduisant alors (qui doit être strictement inférieure à 1 si nous voulons que les développements suivants convergent vers une valeur finie: chaîne de Markov ergodique):

$$A = \frac{\lambda}{\mu} \quad (67.264)$$

qui représente pour rappel le nombre d'appels qui apparaissent sur le nombre d'appels qui se terminent pendant un intervalle de temps (temps de service moyen), ce qui représente en fait tout simplement le trafic (ou autrement dit: l'intensité de trafic), il vient alors:

$$p_0 = \frac{1}{1 + \sum_{k=1}^N \left(\frac{A^k}{k!} \right)} \quad (67.265)$$

ou encore en introduisant le 1 dans la sommation:

$$p_0 = \frac{1}{\sum_{k=0}^N \left(\frac{A^k}{k!} \right)} \quad (67.266)$$

Puisque k représente le nombre d'appels, cette dernière relation donne la probabilité d'avoir 0 appel (clients) pour un trafic permanent donné A dans le système.

En reportant p_0 dans l'expression suivante de p_k (probabilité d'être dans l'état k donc...) obtenue plus haut:

$$p_k = \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) p_0 \quad (67.267)$$

il vient:

$$p_k = \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) p_0 = \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) \frac{1}{\sum_{k=0}^N \left(\frac{A^k}{k!} \right)} \quad (67.268)$$

et en considérant le caractère sans mémoire, nous avons la relation:

$$\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} = \prod_{i=0}^{k-1} \frac{\lambda}{(i+1) \cdot \mu} = \prod_{i=1}^k \frac{\lambda}{i \cdot \mu} = \frac{A^k}{k!} \quad (67.269)$$

Il vient:

$$p_k = \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) p_0 = \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) \frac{1}{\sum_{k=0}^N \left(\frac{A^k}{k!} \right)} = \frac{\frac{A^k}{k!}}{\sum_{k=0}^N \left(\frac{A^k}{k!} \right)} \quad (67.270)$$

où les k peuvent malheureusement porter à confusion. Il convient de faire un peu le ménage. Au numérateur, le k fait référence au nombre de canaux (serveurs, lignes, opérateurs ou terminaux) et au dénominateur le N aussi. Il convient donc de réécrire cela de manière plus convenable:

$$p_k = \frac{\frac{A^k}{k!}}{\sum_{k=0}^N \left(\frac{A^k}{k!} \right)} \quad (67.271)$$

qui donne donc la probabilité de mise en attente (et donc de saturation/blocage) d'un système disposant de N canaux à capacité finie selon le principe du premier/arrivé premier servi (FIFO: First In/First Out) et pour un trafic A (exprimé donc en "Erlang") et dans lequel les communications sont perdues si mises en attente.

Cette relation est parfois notée dans les ouvrages spécialisés sous la forme:

$$\pi_k = \frac{\frac{\rho^k}{k!}}{\sum_{i=0}^k \left(\frac{\rho^i}{i!} \right)} \quad (67.272)$$

Cette relation est très importante en théorie des files d'attente et porte le nom de "**formule d'Erlang-B**".

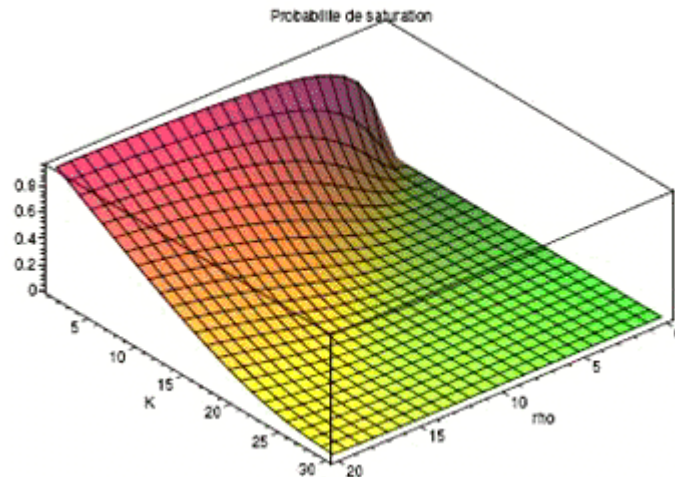


Figure: 67.31 - Probabilité de saturation (plot de la fonction Erlang-B)

Elle est à la base du dimensionnement des réseaux à commutation de circuits. En effet, le problème du dimensionnement d'un commutateur de circuits est le suivant: Étant donné le trafic en nombre de communications par unité de temps A en Erlang, trouver le nombre d'unités de service k tel que la probabilité qu'un appel arrive dans un système devenu bloquant soit inférieur à une certaine valeur.

La probabilité obtenue représente alors la qualité de service offerte par le réseau du point de vue de l'usage. Quant au trafic A , il est estimé en fonction du nombre de postes téléphoniques existants et/ou à venir sur la base d'une activité moyenne par poste et par application (téléphone, télécopieur, terminal, serveur informatique, etc.).

Exemple:

E1. Quelle est la probabilité de saturation d'une hotline (dont la durée de service suit une loi exponentielle et la distribution des arrivées suit une loi de Poisson) sachant que le trafic A de la ligne est estimé à 2 Erlang (1 appel par heure pour 1 appel traité par ½ heure - donc rapport de 2 sur 1) pour une seule ligne téléphonique ($N=1$) en utilisant le modèle d'Erlang-B?:

$$p_k = \frac{\frac{2^1}{1!}}{\frac{2^0}{0!} + \frac{2^1}{1!}} \cong 67\% \quad (67.273)$$

E2. Dans une entreprise, on a dénombré aux heures de pointe 200 appels d'une durée moyenne de 6 minutes à l'heure (temps de service moyen). Quelle est la probabilité de saturation avec 20 opérateurs (sachant que la durée de service suit une loi exponentielle et la distribution des arrivées une loi de Poisson)?

La plus grosse difficulté ici est de calculer le trafic! Il y a donc 200 appels par heure avec 10 appels traités seulement par heure (puisque 6 minutes par appel dans une heure de 60 minutes font 10 appels). Le trafic A est donc de $200/10$ soit 20 Erlang. En appliquant alors la relation précédente, nous avons:

$$p_k = \frac{\frac{20^{20}}{20!}}{\sum_{k=0}^{20} \frac{20^k}{k!}} \cong 15.89\% \quad (67.274)$$

Dans l'industrie, on admet un taux de saturation de 0.01%. En jouant avec un tableur comme Microsoft Excel et l'outil Valeur cible, nous trouvons rapidement que N doit alors être égal à 30.

9.6. PROBABILITÉ DE MISE EN ATTENTE $M/M/K/\infty$ (FORMULE D'ERLANG-C)

Considérons maintenant un système pour lequel les appels peuvent être mis en d'attente avant d'être servis lorsque les k serveurs sont bloqués. Il s'agit donc d'une file d'attente à capacité illimitée de type $M/M/k/\infty$ selon la notation de Kendall.

Avec ce système, nous avons toujours:

$$p_k = p = \lambda dt \quad (67.275)$$

mais pour la probabilité de fin d'appel l'analyse devient plus subtile. D'abord il y a la probabilité que les appels qui se trouvent sur les canaux N disponibles cessent et qui est donnée par:

$$\mu_k dt = k \mu dt \quad \forall 0 \leq k \leq N-1 \quad (67.276)$$

Mais dès que le nombre d'appels est plus grand que le nombre de canaux de communication disponible, la probabilité que cessent les appels est:

$$\mu_k dt = N \mu dt \quad \forall k \geq N \quad (67.277)$$

Ce qui exprime que quel que soit le nombre d'appels, N ont la probabilité d'être mis en attente dès que k (le nombre d'appels) est supérieur ou égal à N .

Ainsi, pour résumer:

$$\mu_k dt = \begin{cases} k \mu dt & \forall 0 \leq k \leq N-1 \\ N \mu dt & \forall k \geq N \end{cases} \quad (67.278)$$

En utilisant:

$$p_k = \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) p_0 \quad (67.279)$$

Nous obtenons par décomposition du terme produit:

$$p_k = \left(\prod_{i=0}^{N-1} \frac{\lambda}{(1+i) \mu_{i=N}} \prod_{i=N}^{k-1} \frac{\lambda}{N \mu} \right) p_0 = \left(\frac{A^N}{N!} \frac{A^{k-N}}{N^{k-N}} \right) p_0 \quad (67.280)$$

D'où finalement:

$$p_k = \begin{cases} \frac{A^N}{N!} p_0 & \forall 0 \leq k \leq N-1 \\ \frac{A^N}{N!} \frac{A^{k-N}}{N^{k-N}} p_0 & \forall k \geq N \end{cases} \quad (67.281)$$

En utilisant l'expression de p_0 :

$$p_0 = \frac{1}{1 + \sum_{k=1}^{+\infty} \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right)} \quad (67.282)$$

Nous pouvons décomposer la sommation:

$$p_0 = \frac{1}{1 + \sum_{k=1}^{N-1} \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) + \sum_{k=N}^{+\infty} \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right)} \quad (67.283)$$

et décomposer le deuxième produit:

$$\begin{aligned} p_0 &= \frac{1}{1 + \sum_{k=1}^{N-1} \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) + \sum_{k=N}^{+\infty} \left(\prod_{i=0}^{N-1} \frac{\lambda}{(i+1) \mu} \prod_{i=N}^{k-1} \frac{\lambda}{N \mu} \right)} \\ &= \frac{1}{1 + \sum_{k=1}^{N-1} \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) + \sum_{k=N}^{+\infty} \left(\frac{A^N}{N!} \frac{A^{k-N}}{N^{k-N}} \right)} \\ &= \frac{1}{1 + \sum_{k=1}^{N-1} \frac{A^k}{k!} + \frac{A^N}{N!} \sum_{k=N}^{+\infty} \left(\frac{A^{k-N}}{N^{k-N}} \right)} \quad (67.284) \\ &= \frac{1}{\sum_{k=0}^{N-1} \frac{A^k}{k!} + \frac{A^N}{N!} \sum_{k=N}^{+\infty} \left(\frac{A^{k-N}}{N^{k-N}} \right)} \\ &= \frac{1}{\sum_{k=0}^{N-1} \frac{A^k}{k!} + \frac{A^N}{N!} \sum_{k=0}^{+\infty} \left(\frac{A}{N} \right)^k} \end{aligned}$$

Sous l'hypothèse que:

$$A/N < 1 \quad (67.285)$$

La somme:

$$\sum_{k=0}^{+\infty} \left(\frac{A}{N} \right)^k \quad (67.286)$$

peut être simplifiée. Effectivement, en posant:

$$N/A = x \quad (67.287)$$

il vient:

$$\sum_{k=0}^{+\infty} \frac{1}{x^k} \quad (67.288)$$

et comme nous l'avons montré lors de notre étude de la fonction Zeta (cf. chapitre de Suites Et Séries) cette somme peut s'écrire sous la forme:

$$\sum_{k=0}^{+\infty} \frac{1}{x^k} = \frac{1}{1 - \frac{1}{x}} \quad (67.289)$$

Donc:

$$\sum_{k=0}^{+\infty} \left(\frac{A}{N} \right)^k = \frac{1}{1 - \frac{A}{N}} \quad (67.290)$$

Donc finalement:

$$p_0 = \frac{1}{\sum_{k=0}^{N-1} \frac{A^k}{k!} + \frac{A^N}{N!} \frac{1}{1 - \frac{A}{N}}} \quad (67.291)$$

Puisque k représente le nombre d'appels, cette dernière relation donne la probabilité d'avoir 0 appel (clients) pour un trafic permanent donné A dans le système.

Nous avons donc pour $\forall k \geq N$:

$$p_k = \frac{A^N}{N!} \frac{A^{k-N}}{N^{k-N}} p_0 = \frac{\frac{A^N}{N!} \frac{A^{k-N}}{N^{k-N}}}{\sum_{k=0}^{N-1} \frac{A^k}{k!} + \frac{A^N}{N!} \frac{1}{1 - \frac{A}{N}}} = \frac{\frac{A^N}{N!} \frac{A^{k-N}}{N^{k-N}}}{\sum_{k=0}^{N-1} \frac{A^k}{k!} + \frac{A^N}{N!} \frac{N}{N-A}} \quad (67.292)$$

La probabilité cumulée de mise en file d'attente se note $C(N, A)$. Elle est égale à:

$$\begin{aligned}
 C(N, A) &= \sum_{k=N}^{+\infty} p_k = \sum_{k=N}^{+\infty} \frac{A^N}{N!} \frac{A^{k-N}}{N^{k-N}} p_0 = \frac{\sum_{k=N}^{+\infty} \frac{A^N}{N!} \frac{A^{k-N}}{N^{k-N}}}{\sum_{k=0}^{N-1} \frac{A^k}{k!} + \frac{A^N}{N!} \frac{N}{N-A}} \\
 &= \frac{\frac{A^N}{N!} \sum_{k=N}^{+\infty} \frac{A^{k-N}}{N^{k-N}}}{\sum_{k=0}^{N-1} \frac{A^k}{k!} + \frac{A^N}{N!} \frac{N}{N-A}} = \frac{\frac{A^N}{N!} \sum_{k=0}^{+\infty} \left(\frac{A}{N}\right)^k}{\sum_{k=0}^{N-1} \frac{A^k}{k!} + \frac{A^N}{N!} \frac{N}{N-A}}
 \end{aligned} \quad (67.293)$$

en procédant exactement comme dans les paragraphes précédents, nous avons finalement:

$$C(N, A) = \frac{\frac{A^N}{N!} \frac{N}{N-A}}{\sum_{k=0}^{N-1} \frac{A^k}{k!} + \frac{A^N}{N!} \frac{N}{N-A}} \quad (67.294)$$

qui donne donc la probabilité de mise en attente (et donc de saturation/blocage - c'est-à-dire que tous les serveurs sont bloqués) dès l'arrivée dans un système disposant de N canaux à capacité infinie selon le principe du premier arrivé/premier servi (FIFO: First In/First Out) et pour un trafic A (exprimé donc en "Erlang") et dans lequel les communications peuvent être mises en attente (à l'opposé du modèle Erlang-B). Cette dernière relation est appelée "**formule d'Erlang-C**". Traditionnellement on note:

$$C(N, A) = P(W) \quad (67.295)$$

le "W" venant de l'anglais "Wait" (attendre).

Le modèle proposé ci-dessus est bien évidemment critiquable, car en réalité la capacité de la file d'attente est finie et certains clients abandonnent lorsque l'attente est trop longue.

Exemple:

Si nous prenons un taux d'arrivée de 1 appel par minute et une durée moyenne de service de 5 minutes, nous avons alors:

$$A = \frac{\lambda}{\mu} = \lambda \frac{1}{\mu} = 1 \cdot 5 = 5 \quad (67.296)$$

Si nous prenons un nombre N de 7 serveurs, nous avons:

$$C(N, A) = \frac{\frac{A^N}{N!} \frac{N}{N-A}}{\sum_{k=0}^{N-1} \frac{A^k}{k!} + \frac{A^N}{N!} \frac{N}{N-A}} = \frac{\frac{5^7}{7!} \frac{7}{7-5}}{\sum_{k=0}^6 \frac{5^k}{k!} + \frac{5^7}{7!} \frac{7}{7-5}} \cong 0.3241 \quad (67.297)$$

Donc 32.41% de probabilité cumulée d'être mis en attente. Ce qui un peu beaucoup... (une règle empirique consiste à chercher le nombre N de serveurs afin que cette dernière valeur descende en dessous des 20%).

10. ASSURANCES

L'assurance est une opération pour laquelle une personne (l'assureur) groupe en mutualité d'autres personnes (les assurés) afin de les mettre en situation de s'indemniser mutuellement des pertes éventuelles (les sinistres) auxquelles les expose la réalisation de certains risques, au moyen des sommes (primes ou cotisations) versées par chaque assuré à une masse commune gérée par l'assureur.

Le dénominateur commun de la majorité des lois définit le contrat d'assurance comme un contrat en vertu duquel, moyennant le paiement d'une prime fixe ou variable, une partie, l'assureur, s'engage envers une autre partie, le preneur d'assurance, à fournir une prestation stipulée dans le contrat au cas où surviendrait un événement incertain que selon le cas, l'assuré ou le bénéficiaire, a intérêt à ne pas voir se réaliser.

Évidemment, l'opération d'assurance a pour effet le transfert (total ou partiel) des conséquences financières du risque subi par l'assuré vers une société d'assurance. Dès lors, à la souscription du contrat, l'assureur et l'assuré conviennent:

- D'un événement ou d'une liste d'événements, repris dans la police d'assurance, et garantis par l'assureur.
- D'une prime payée par l'assuré à l'assureur.

Les dépenses prises en charge par l'assureur peuvent correspondre:

- Soit à des indemnités à verser à des tiers, au titre de la responsabilité (civile, professionnelle ou autre) de l'assuré.
- Soit à la réparation des dommages subis par ce dernier.

Les hypothèses d'usage de l'assurance sont les suivantes:

H1. D'un point de vue juridique, un contrat d'assurance est un "**contrat aléatoire**" valide uniquement pour couvrir des risques ayant une composante aléatoire.

H2. La "règle du jeu" du risque doit être stable dans un laps de temps considéré comme long (au moins quelques années).

H3. La perte maximale possible ne doit pas être trop importante par rapport à la marge de solvabilité de l'assureur.

H4. La prime moyenne du risque doit être identifiable et quantifiable selon des variables statistiques explicatives bien choisies afin éventuellement de permettre une segmentation de la gestion des risques.

H5. Les risques doivent être indépendants (et s'ils sont identiques distribués en termes de probabilités et de pondération c'est mieux...) et démontrables comme étant tels significativement en utilisant les outils statistiques.

H6. Il doit exister un marché dans le sens que l'offre et la demande d'assurance doivent arriver à un prix d'équilibre (en quelque sorte l'équivalent de l'absence d'opportunité d'arbitrage en finance).

H7. L'espérance mathématique est considérée comme le prix de la prime pure juste à faire payer aux assurés.

Définitions:

D1. La "**prime pure**" est dans le domaine de l'assurance choisie comme étant l'espérance mathématique de la charge; elle correspond à la prime minimale que peut demander un assureur pour ne pas, statistiquement, faire ruine de façon certaine.

D2. Le "**chargement de sécurité**" est le montant qui vient s'ajouter à la prime pure et permettant à l'assureur de pouvoir résister à la volatilité des remboursements.

D3. Le "**chargement de frais de gestion**" est le montant qui vient s'ajouter aux deux précédents et est lié au fonctionnement de leur société, de la gestion des contrats, du recouvrement des primes, du placement des actifs (prime technique de base), des taxes...

D4. Le "**chargement de frais commerciaux**" vient s'ajouter aux trois précédents et est lié à l'acquisition des contrats (commissions des intermédiaires, frais des réseaux commerciaux, publicité).

D5. Au final, l'ensemble des coûts se retrouve dans la "**prime commerciale**" qu'est celle communiquée au client.

Nous pouvons déjà en conclure que dans le cas d'un système d'assurance étatique obligatoire ou facultatif, les frais de gestion et commerciaux seront toujours inférieurs, sous l'hypothèse d'une méthode managériale égale, à un système privé.

10.1. CALCUL DE PRIME

L'assureur ne connaît donc pas exactement le montant des sinistres qui va survenir. En tarifiant les contrats au niveau de la prime pure (et en supposant une distribution des pertes symétriques), l'assureur perd de l'argent une année sur deux. En l'absence de fonds propres, cette situation conduirait immédiatement à la faillite.

Pour se protéger, l'assureur ajoute donc à sa prime un chargement de sécurité. De nombreuses méthodes permettant de le déterminer sont possibles, aucune n'ayant à ce jour supplanté largement les autres:

- Chargement proportionnel à la prime pure. Le coefficient de proportionnalité reflète l'idée que l'assureur a de la volatilité du risque.
- Chargement dépendant de l'écart-type des pertes. Cette méthode est une légère formalisation de la précédente. Elle pose problème, car elle introduira un chargement de sécurité qui dépendra des cas de gains (perte réelle inférieure à la prime pure)
- Chargement dépendant d'un certain quantile des pertes (par exemple le troisième quartile). Un tel chargement permet de garantir que la prime sera suffisante dans un nombre de cas déterminé à l'avance, mais ne donne aucune information sur les cas de pertes techniques.

Considérons le cas où la population serait hétérogène, in extenso deux classes $\mathcal{S}_A, \mathcal{S}_B$ de risque coexistent dans la population de poids respectifs n_A, n_B avec bien évidemment:

$$\frac{n_A}{n_A + n_B} + \frac{n_B}{n_A + n_B} = \frac{n_A}{n} + \frac{n_B}{n} = \frac{n_A + n_B}{n} = 100\% \quad (67.298)$$

Considérons une assurance maladie avec deux catégories d'assurés (jeunes en bonne santé A /seniors à risque B).

Pour simplifier l'exemple à l'extrême, imaginons que l'assureur sait, à l'aide de ses statistiques internes, qu'un individu du groupe A va coûter à l'assurance une somme définie par une loi de distribution statistique que nous noterons dans le cadre de ce chapitre:

$$F_X^A(x) \rightarrow \mathbb{P}(X_A \leq x) \quad (67.299)$$

Soit à tout coût x est associée une certaine probabilité cumulée donnée par la fonction F_X^A . De même pour le groupe B:

$$F_X^B(x) \rightarrow \mathbb{P}(X_B \leq x) \quad (67.300)$$

Si nous prenons alors au hasard un individu dans le portefeuille, si θ désigne le groupe (notation traditionnelle en assurance), l'assureur devrait donc réclamer en primes pures pour le groupe A:

$$E[S_A] = E[S | \theta = A] = \mu_A \quad (67.301)$$

soit l'espérance de la fonction F_X^A .

Idem pour le groupe B:

$$E[S_B] = E[S | \theta = B] = \mu_B \quad (67.302)$$

Et pour un individu pris au hasard dans les deux groupes:

$$\mu_S = E[S] = E[E(S | \theta)] = \frac{n_A}{n} \mu_A + \frac{n_B}{n} \mu_B \quad (67.303)$$

Ce dernier cas étant appelé "**mécanisme de solidarité**". Ainsi, les bons risques paient pour les mauvais risques...

La variance de la prime pure sera elle donnée par la relation démontrée dans le chapitre de Statistiques lors de l'étude des propriétés de la variance (variance de deux séries statistiques):

$$\sigma_S^2 = \frac{n_A \sigma_A^2 + n_B \sigma_B^2}{n_A + n_B} + \frac{n_A n_B}{(n_A + n_B)^2} (\mu_A - \mu_B)^2 \quad (67.304)$$

où nous voyons que le terme:

$$(\mu_A - \mu_B)^2 \geq 0 \quad (67.305)$$

au numérateur correspond à l'homogénéité des deux groupes. Et donc que l'hétérogénéité fait accroître la variabilité de la prime pure dans le mécanisme de solidarité.

Imaginons maintenant le cas où:

$$\mu_A < \mu_B \quad (67.306)$$

et une assurance privée (1) qui pratique le mécanisme de solidarité et une autre assurance (2) qui ne le pratique pas. Dans le cas présent, nous devons distinguer deux situations d'un point de vue économique:

Si tous les assurés sont rationnels (sous-entendus un peu... égoïstes) alors:

- Les jeunes en bonne santé représentant le groupe A vont aller chez l'assurance privée (2) qui a segmenté les risques et permet donc aux jeunes de payer moins cher.
- Les seniors en moins bonne santé représentant le groupe B vont aller chez l'assurance privée (1) qui n'a pas segmenté les risques, mais qui par idéologie a appliqué le principe de solidarité.

La conclusion est que l'assurance privée (1) va rapidement faire faillite car:

- Les bons risques ne compensent plus le rabais de solidarité par égoïsme
- Nous sommes sur un marché concurrentiel où les assurances ne sont pas idéologiques...

Ce constat s'appelle le "**problème d'anti-sélection**" ou de "**sélection inverse**" basée sur l'approche G. Akerlof, prix Nobel d'Économie en 2001.

La conclusion est que la privatisation des assurances basées sur un principe étatique (législation) de solidarité ne peut pas fonctionner sans engendrer des coûts supplémentaires aux primes pures à cause d'une anti-sélection périodique qui engendre un flux constant d'assurés d'une assurance à une autre et donc engendre des coûts administratifs et informatiques phénoménaux! Donc en tout point la suppression de la concurrence reste meilleure en terme de solidarité mais par contre pas en termes d'emplois pour les salariés des assurances (qui se trouveraient alors en grande majorité au chômage...).

Ainsi, la suppression de la concurrence pour des services d'assurances qui ont toutes des prestations de qualité équivalentes (comme c'est le cas pour l'assurance maladie en Suisse par exemple...) élimine le principe d'anti-sélection, applique de manière concrète le mécanisme de solidarité voulu par l'état et enfin diminue les coûts administratifs dus aux va-et-vient des assurés et des développements d'outils informatiques de gestion maison coûtant des millions à chaque assurance et qui engendrent des coûts qui au final sont répercutés sur le prix de la prime commerciale!

À ceci, il faut rajouter que pour diminuer la volatilité globale (écart-type global), une assurance devrait en théorie segmenter les risques à l'infini ce qui en fait un système non viable pour certains domaines particuliers de l'assurance.

Mais il faut se rappeler que cette conclusion n'est valable que si les prestations des assurances sont identiques (ou quasi similaires) sur un marché donné!

Signalons également un souci récurrent dans le domaine des assurances, appelé "aléa moral", qui se base sur le constat que les personnes qui s'assurent ont tendance à être moins prudentes que les personnes qui ne s'assurent pas. En d'autres termes, l'assurance génère du risque.

10.2. PRISE EN COMPTE DE L'EXPERIENCE

Pour l'instant, pour déterminer une prime d'assurance, nous avons noté qu'il était possible d'intégrer des variables exogènes (sexe, âge, enfants, puissance du véhicule, nationalité, environnement, etc.).

Mais un point important à ne pas négliger est le retour d'expérience des sinistres d'un assuré. Voyons en un exemple concret.

Supposons que le nombre de sinistres sur un an, pour un assuré donné, suive une loi de Poisson (loi des événements rares) donnée pour rappel par (cf. chapitre de Statistiques):

$$P_k = \frac{\mu^k}{k!} \cdot e^{-\mu} \quad (67.307)$$

ce qui se note dans le domaine des assurances:

$$P_\theta = \frac{\mu^\theta}{\theta!} e^{-\mu} \quad (67.308)$$

Supposons que la population des assurés est séparée en trois classes de risques suivant chacune une distribution de Poisson telle que:

$$\begin{aligned} P(\theta = 1) &= \frac{1^\theta}{\theta!} e^{-1} \rightarrow 70\% \\ P(\theta = 2) &= \frac{2^\theta}{\theta!} e^{-2} \rightarrow 20\% \\ P(\theta = 3) &= \frac{3^\theta}{\theta!} e^{-3} \rightarrow 10\% \end{aligned} \quad (67.309)$$

En d'autres termes, il y a 70% de la population totale qui suit une loi de Poisson d'espérance 1 (classe de bons risques), 20% qui suit une loi de Poissons d'espérance 2 (classe de risques moyens), et 10% qui suit une loi de Poisson d'espérance 3 (classe de mauvais risques).

Évidemment l'espérance globale d'un individu du portefeuille est alors de l'espérance de la "**distribution mixte**":

$$\mu = E(\theta) = 70\% \cdot 1 + 20\% \cdot 2 + 10\% \cdot 3 = 1.4 \quad (67.310)$$

Supposons que les coûts d'un incident soient fixes et de type indemnitaires de 1'000.-. Nous pouvons alors nous demander quelle devrait être la prime pour un assuré, sachant que la première année, il a eu 2 sinistres. Ce qui revient à se demander, quel est le nombre d'accidents qu'il risque d'avoir la deuxième année.

Si nous notons X_1 le nombre d'accidents de la première année, et X_2 le nombre d'accidents a posteriori connaissant X_1 nous nous retrouvons donc avec un problème de probabilité conditionnelle (une démarche bayésienne autrement dit...) conforme à ce que nous avons étudié dans le chapitre de Probabilités:

$$P(X_2 | X_1 = 2) = \frac{P(X_1 \cap X_2)}{P(X_1)} = \frac{P(X_1 = 2 \cap X_2)}{P(X_1 = 2)} \quad (67.311)$$

où nous avons adopté conformément à la tradition dans le domaine la notion de "|" pour signaler la probabilité conditionnelle.

Mais nous ne pouvons pas calculer le numérateur, car nous ne savons pas à l'avance quelle sera la valeur de X_2 . Nous allons donc calculer l'espérance conditionnelle espérée afin de contourner ce problème:

$$E(X_2 | X_1 = 2) = \frac{E(X_1 \cap X_2)}{E(X_1)} = \frac{E(X_1 = 2 \cap X_2)}{E(X_1 = 2)} \quad (67.312)$$

Nous avons d'abord:

$$E(X_1 = 2) = 70\% \cdot \frac{1^2}{2!} e^{-1} + 20\% \cdot \frac{2^2}{2!} e^{-2} + 10\% \cdot \frac{3^2}{2!} e^{-3} = 0.205 \quad (67.313)$$

Ce dernier calcul (espérance de la distribution mixte) étant noté dans le domaine de l'assurance de la manière suivante:

$$E(X_1 = 2) = \sum P(\Theta = \theta) P(X_1 = 2 | \Theta = \theta) = \sum P(\Theta = \theta) \frac{\theta^2}{2!} e^{-\theta} \quad (67.314)$$

Si nous considérons les deux variables aléatoires comme indépendantes:

$$E(X_1 = 2 \cap X_2) = E((X_1 = 2)(X_2)) = E(X_1 = 2)E(X_2) = 0.205 \cdot E(X_2) \quad (67.315)$$

Il vient alors immédiatement que:

$$E(X_2 | X_1 = 2) = \frac{E(X_1 \cap X_2)}{E(X_1)} = \frac{E(X_1 = 2 \cap X_2)}{E(X_1 = 2)} = E(X_2) \quad (67.316)$$

Nous retombons donc sur une valeur connue correspondant à la prime pure:

$$\begin{aligned} E(X_2) &= 70\% \cdot \theta_1 + 20\% \cdot \theta_2 + 10\% \cdot \theta_3 \\ &= 70\% \cdot 1 + 20\% \cdot 2 + 10\% \cdot 3 = 1.4 = E(\theta) \end{aligned} \quad (67.317)$$

identique au calcul de:

$$E(X_1) = 70\% \cdot \theta_1 + 20\% \cdot \theta_2 + 10\% \cdot \theta_3 = 1.4 = E(\theta) \quad (67.318)$$

Donc les variables aléatoires X_1, X_2 sont confondues avec celle de la classe de risque!

Ce système n'est alors bien évidemment pas conforme au bonus-malus. Nous devons alors considérer que les deux variables aléatoires ne sont pas indépendantes (au fait c'est la définition même du bonus-malus!).

Ainsi, si les deux variables aléatoires ne sont pas indépendantes, nous avons:

$$\begin{aligned} E(X_1 = 2 \cap X_2) &= E\left(\frac{\theta^2}{2!} e^{-\theta} \cap \theta\right) = \sum P(\Theta = \theta) \left(\frac{\theta^2}{2!} e^{-\theta} \cdot \theta\right) \\ &= 70\% \cdot \frac{1^2}{2!} e^{-1} \cdot 1 + 20\% \cdot \frac{2^2}{2!} e^{-2} \cdot 2 + 10\% \cdot \frac{3^2}{2!} e^{-3} \cdot 3 = 0.304 \end{aligned} \quad (67.319)$$

Et alors:

$$E(X_2 | X_1 = 2) = \frac{E(X_1 = 2 \cap X_2)}{E(X_1 = 2)} = \frac{0.304}{0.205} = 1.484 \quad (67.320)$$

À comparer avec la prime pure sans bonus-malus de 1.4!

Ce qui est délicat avec cette méthode, c'est que lorsque l'on cumule les années... cette approche bayésienne devient pénible. Effectivement, imaginons que nous souhaiterions déterminer la prime pure de la troisième année sachant que la deuxième année, l'assuré a eu 1 accident. Nous avons alors l'espérance conditionnelle:

$$E(X_3 | X_1 = 2, X_2 = 2) = \frac{E((X_1 = 2, X_2 = 1) \cap X_3)}{E(X_1 = 2, X_2 = 1)} \quad (67.321)$$

que nous laissons le soin au lecteur de développer...

Nous remarquons de ce qui a été vu ci-dessus que l'espérance d'une distribution mixte définie par:

$$V(X) = Vp_1 f_{X_1}(x) + \dots + p_n f_{X_n}(x) = \sum_{i=1}^n p_i f_{X_i}(x) \quad (67.322)$$

est donc trivialement:

$$E(X) = \mu = \sum_{i=1}^n p_i \mu_i \quad (67.323)$$

10.3. FACTEUR D'ACTUALISATION D'UNE ASSURANCE RETRAITE

Recherchons maintenant quelque chose qui n'a absolument rien à voir avec ce que nous venons de calculer et qui est le "**facteur d'actualisation**" par lequel nous devons diviser un solde d'épargne au début de la retraite, pour répartir ce montant sur chaque année durant la retraite, selon l'espérance de vie ainsi que les taux d'intérêt et d'inflation. L'approche ici sera donc de simplifier en se questionnant sur quel niveau d'épargne il sera nécessaire d'avoir accumulé au début de la retraite pour obtenir une rente de 1.- (réelle) chaque année durant la retraite. C'est à partir de ce niveau d'épargne nécessaire pour recevoir 1.- qu'il sera possible d'utiliser n'importe quel solde d'épargne pour dériver quelle rente sera disponible à l'épargnant.

En premier lieu, nous allons définir les paramètres utilisés:

- m Nombre de périodes à la retraite (années ou mois)
- S_k^* Solde d'épargne à la période k
- p Prestation périodique (que nous supposons ici annuelle)
- i Taux périodique d'inflation (supposé ici constant et annuel)
- R Taux de rendement moyen géométrique du marché (supposé constant et annuel)
- α_m Facteur d'actualisation à l'âge m

Ainsi, comme mentionné plus haut, nous sommes à la recherche du niveau d'épargne initial, S_0 , qui permettra d'obtenir une prestation périodiques (donc annuelle dans le cas présent), p , égale à 1.- indexée sur l'inflation i .

En commençant à partir du solde initial $S_0 = S$, en présumant que les prestations sont versées au début de l'année (*praenumerando*), nous pouvons dériver le solde restant à la fin de la première période de retraite ainsi:

$$S_1 = (S - p) \cdot (1 + R) \quad (67.324)$$

Plus précisément, le retraité aura encaissé un montant p pour la première période, laissant le résiduel être exposé au rendement moyen géométrique du marché R .

Ensuite, d'une manière similaire et en considérant que les prestations sont indexées sur l'inflation i , nous pouvons calculer ce solde en fin de deuxième période et de toute période k subséquente:

$$S_2 = [(S - p)(1 + R) - (1 + i)p](1 + R) \quad (67.325)$$

ou:

$$S_2 = S(1 + R)^2 - p(1 + R)^2 - p(1 + R)(1 + i) \quad (67.326)$$

ou encore:

$$\begin{aligned} S_3 &= [S(1 + R)^2 - p(1 + R)^2 - p(1 + R)(1 + i) - p(1 + i)^2](1 + R) \\ &= S(1 + R)^3 - p(1 + R)^3 - p(1 + R)^2(1 + i) - p(1 + R)(1 + i)^2 \\ &= S(1 + R)^3 - p(1 + R)(1 + i)^2 - p(1 + R)^2(1 + i) - p(1 + R)^3 \end{aligned} \quad (67.327)$$

et ainsi de suite... Il vient alors:

$$S_k = S(1 + R)^k - p \sum_{j=1}^k (1 + R)^j (1 + i)^{k-j} \quad (67.328)$$

ou autrement écrit:

$$S_k = S(1 + R)^k - p(1 + i)^k \sum_{j=1}^k \left(\frac{1 + R}{1 + i} \right)^j = S(1 + R)^k - p(1 + i)^k \sum_{j=1}^k q^j \quad (67.329)$$

Ainsi nous avons, à un terme près, une série géométrique de raison q connue dans le dernier terme. Nous avons effectivement démontré dans le chapitre de Suites Et Séries que pour j allant de 0 à n celle-ci vaut :

$$S_n = u_0 \frac{1 - q^{n+1}}{1 - q} \quad (67.330)$$

et donc:

$$\sum_{j=1}^k \left(\frac{1+R}{1+i} \right)^j = \left(\frac{1+R}{1+i} \right) \sum_{j=0}^{k-1} \left(\frac{1+R}{1+i} \right)^j = \left(\frac{1+R}{1+i} \right) \frac{1 - \left(\frac{1+R}{1+i} \right)^k}{1 - \frac{1+R}{1+i}} = \frac{1 - \left(\frac{1+R}{1+i} \right)^k}{\frac{1+i}{1+R} - 1} \quad (67.331)$$

Il nous est donc possible de réécrire, pour la dernière période de retraite m , le solde de la manière suivante:

$$S_m = S(1+R)^m - p(1+i)^m \frac{1 - \left(\frac{1+R}{1+i} \right)^m}{\frac{1+i}{1+R} - 1} \quad (67.332)$$

Remarquons que dans le cas peu probable et particulier où le rendement moyen géométrique du marché est égal à l'inflation, c'est-à-dire si $R = i$, nous avons:

$$\begin{aligned} S_m &= S(1+R)^m - p(1+i)^m \sum_{j=1}^m \left(\frac{1+R}{1+i} \right)^j = S(1+R)^m - p(1+R)^m \sum_{j=1}^m \left(\frac{1+R}{1+R} \right)^j \\ &= S(1+R)^m - p(1+R)^m \sum_{j=1}^m 1^j = S(1+R)^m - p(1+R)^m m = (1+R)^m [S - p \cdot m] \end{aligned} \quad (67.333)$$

À ce stade, la raison pour laquelle le terme S_m est un paramètre important est qu'il s'agit du solde du retraité à la fin de sa période de retraite, c'est-à-dire au décès (mort). Il sera ici présumé que le retraité ne planifie pas de legs et donc que

$$S_m = 0 \quad (67.334)$$

Conséquemment, pour $R \neq i$, cela revient à trouver S pour lequel:

$$S_m = S(1+R)^m - p(1+i)^m \frac{1 - \left(\frac{1+R}{1+i} \right)^m}{\frac{1+i}{1+R} - 1} = 0 \quad (67.335)$$

Nous pouvons alors établir dans ce cas que pour $R \neq i$:

$$S = p \frac{(1+i)^m}{(1+R)^m} \frac{1 - \left(\frac{1+R}{1+i} \right)^m}{\frac{1+i}{1+R} - 1} = p \left(\frac{1+i}{1+R} \right)^m \frac{1 - \left(\frac{1+R}{1+i} \right)^m}{\frac{1+i}{1+R} - 1} \quad (67.336)$$

Et dans le cas peu probable et particulier où le rendement moyen géométrique du marché est égal à l'inflation, c'est-à-dire si $R = i$, nous avons:

$$S_m = (1+R)^m [S - p \cdot m] = 0 \quad (67.337)$$

et nous déduisons un résultat logique:

$$S = p \cdot m \quad (67.338)$$

Pour la suite, revenons au cas réaliste où $R \neq i$ et donc avec:

$$S = p \left(\frac{1+i}{1+R} \right)^m \frac{1 - \left(\frac{1+R}{1+i} \right)^m}{\frac{1+i}{1+R} - 1} \quad (67.339)$$

et notons:

$$b = \frac{1+R}{1+i} = 1+r \quad (67.340)$$

La relation antérieure s'écrit alors sous une forme un peu plus condensée:

$$S = p \frac{1}{b^m} \frac{1-b^m}{b-1} \quad (67.341)$$

où:

$$a_m = \frac{1}{b^m} \frac{1-b^m}{b-1} \quad (67.342)$$

est donc le facteur d'actualisation. De plus, nous avons la relation suivante donnant l'espérance d'être en vie à l'âge a (âge correspondant à l'âge supposé de décès dans le cas présent) démontrée dans le chapitre de Dynamique Des Populations:

$$m = e(a,s) - a + 1 \quad (67.343)$$

avec pour rappel la fonction $e(a,s)$ représentant l'espérance de vie conditionnelle à l'âge a et pour le sexe s .

10.4. ASSURANCES DE RENTES

Nous allons maintenant étudier une rente qui n'est pas versée d'une manière certaine mais d'une manière "viagère", c'est-à-dire qui dépend de la survie de l'assuré. Il va donc falloir c'est multiplier chaque terme de rente par la probabilité qu'il lui soit versé.

Évidemment, dans la majorité des cas, ce qui va nous intéresser sera la "valeur actuelle" des rentes viagères afin de savoir ce que devra payer l'assureur (et donc ce qu'il devra avoir capitalisé). Il existe deux types de rentes viagères courantes. Nous allons voir en quoi consiste chacune. Mais avant cela, rappelons que nous avons vu dans le chapitre de Dynamique Des Populations que la probabilité d'être en vie à une $x + 1$ sachant qu'on est en vie à un âge x est donné par:

$${}_1p_x = \frac{S_{x+1}}{S_x} < 1 \quad (67.344)$$

et rappelons (suite à la demande d'un lecteur) que la valeur actuelle d'un capital C_0 à taux constant payable dans 1 année est donné par (cf. chapitre d'Économie):

$$C_0 = \frac{C_1}{(1+t\%)^1} \quad (67.345)$$

Ce qui donne donc la somme à placer à un taux $t\%$ pour avoir après un an le capital C_1 .

En combinant probabilité de survie et valeur actuelle, nous pouvons définir dans le cas d'une rente la prime unique de la prestation comme étant:

$$PU = C_0 \cdot {}_1p_x \quad (67.346)$$

Si l'assuré est en vie à l'échéance, c'est l'assuré qui gagne au jeu (en réalité ils gagnent très souvent les deux...), dans le cas contraire c'est l'assureur (qui alors lui dans ce cas là est vraiment très gagnant!).

Bien évidemment dans le cas ci-dessus, extrêmement simplifié, il n'y a qu'un seul versement de la part de l'assureur. Mais dans la réalité il peut y en avoir plusieurs (c'est ce que nous allons étudier), le versement de l'assuré peut ne pas être sous la forme d'une prime unique, le taux moyen géométrique du marché peut ne pas être constant (ce dernier cas étant traité normalement avec des simulations de Monte-Carlo).

Donc même si ce que nous allons voir est très simplifié par rapport à ce qui se fait dans les assurances par les actuaires, cela donne déjà une piste de recherche pour des modèles plus élaborés.

10.4.1. RENTE VIAGÈRE TEMPORAIRE

La rente viagère temporaire consiste pour l'assureur à verser des termes de rentes, tant que l'assuré est en vie mais au maximum durant n années après qu'il ait atteint un âge x donné mais tout de suite après que l'assuré ait payé sa prime unique (qui correspond justement à l'âge x en question). L'objectif est donc de déterminer le capital (valeur actuelle) que doit avoir amassé l'assureur au moment du commencement du paiement de ces rentes (ce capital pouvant aussi être demandé sous forme de paiements multiples plutôt que sous la forme d'une prime unique).

Évidemment ce cas est peu réaliste car peu de personnes vont payer une prime unique (grosse somme d'argent) à un moment x de leur vie pour commencer tout de suite à recevoir des rentes. Mais c'est pédagogiquement intéressant avant d'aborder le type de rente viagère différée et qui est un peu plus réaliste.

La rente viagère temporaire praenumberando (in extenso: la valeur actuelle de la rente viagère unitaire payable praenumberando tant que l'assuré est en vie mais au maximum pendant n années) est donnée par la relation suivante qui découle des développements effectués dans le chapitre d'Économie à la différence que l'on y trouve des coefficients de probabilité de survie:

$$\ddot{a}_{x:n} = 1 + f_e \frac{S_{x+1}}{S_x} + f_e^2 \frac{S_{x+2}}{S_x} + \dots + f_e^{n-1} \frac{S_{x+n-1}}{S_x} = \sum_{t=0}^{n-1} f_e^t \frac{S_{x+t}}{S_x} \quad (67.347)$$

où pour rappel (cf. chapitre d'Économie), nous avons le facteur d'escompte (dans le cas présent nous devrions plutôt utiliser le terme de "facteur de capitalisation") qui est:

$$f_e = \frac{1}{1+t\%} \quad (67.348)$$

La rente viagère temporaire postnumerando (valeur actuelle d'une rente viagère unitaire payable postnumerando tant que l'assuré est en vie mais au maximum pendant n années) est elle donnée par:

$$a_{x:n} = f_e \frac{S_{x+1}}{S_x} + f_e^2 \frac{S_{x+2}}{S_x} + \dots + f_e^n \frac{S_{x+n}}{S_x} = \sum_{t=1}^n f_e^t \frac{S_{x+t}}{S_x} \quad (67.349)$$

Exemple:

Nous souhaitons verser une rente de 1'000.- durant 3 ans à un homme de 40 ans. La prime unique qu'il devra payer si ni l'intérêt, ni la mortalité n'interviennent dans les calculs, ensuite si seulement l'intérêt est considéré dans les calculs et enfin en prenant en compte l'intérêt et la mortalité est donnée respectivement par les trois relations suivantes:

$$\begin{aligned} PU_{(1)} &= 1'000 + 1'000 + 1'000 = 1'000 \cdot 3 \\ PU_{(2)} &= 1'000 \cdot f_e + 1'000 \cdot f_e^2 + 1'000 \cdot f_e^3 = 1'000 \cdot a_{\overline{3}|} \leq PU_{(1)} \\ PU_{(3)} &= 1'000 \cdot f_e \cdot \frac{S_{41}}{S_{40}} + 1'000 \cdot f_e^2 \cdot \frac{S_{42}}{S_{40}} + 1'000 \cdot f_e^3 \cdot \frac{S_{43}}{S_{40}} = 1'000 \cdot a_{\overline{40:3}|} < PU_{(2)} \end{aligned} \quad (67.350)$$

10.4.2. RENTE VIAGÈRE DIFFÉRÉE

Les rentes viagères différées constituent la majorité des contrats d'assurance de rentes. L'assuré s'acquitte toujours d'une prime unique, pour recevoir une rente s'il est en vie à un certain âge (à l'âge de sa retraite par exemple), donc commencer à recevoir la rente k années après avoir payé la prime unique, mais cette fois-ci jusqu'à son décès.

La rente viagère différée praenumberando (valeur actuelle d'une rente viagère unitaire différée de k années et payable praenumberando jusqu'à décès de l'assuré) est donnée par:

$${}_k\ddot{a}_x = f_e^k \frac{S_{x+k}}{S_x} + f_e^{k+1} \frac{S_{x+k+1}}{S_x} + \dots + f_e^{\omega-x} \frac{S_{\omega}}{S_x} = \sum_{t=k}^{\omega-x} f_e^t \frac{S_{x+t}}{S_x} \quad (67.351)$$

où ω est la dernière valeur (numéro de ligne) de la table de mortalité. Le lecteur remarquera peut-être que si nous posons $k = 1$ (sans différé) et $\omega = x + n$, nous retombons sur une rente viagère temporaire postnumerando $a_{x:n}$.

Donc normalement pour un même âge x et une même durée de rente supposée, la valeur actuelle de la rente viagère différée (prime unique) est inférieure à celle de rente viagère temporaire et ce d'autant plus petit que le différé est grand.

La rente viagère différée postnumerando (valeur actuelle d'une rente viagère unitaire différée de k années et payable postnumerando jusqu'à décès de l'assuré) est donnée par:

$${}_k|a_x = f_e^{k+1} \frac{S_{x+k+1}}{S_x} + f_e^{k+2} \frac{S_{x+k+2}}{S_x} + \dots + f_e^{\omega-x} \frac{S_{\omega}}{S_x} = \sum_{t=k+1}^{\omega-x} f_e^t \frac{S_{x+t}}{S_x} \quad (67.352)$$

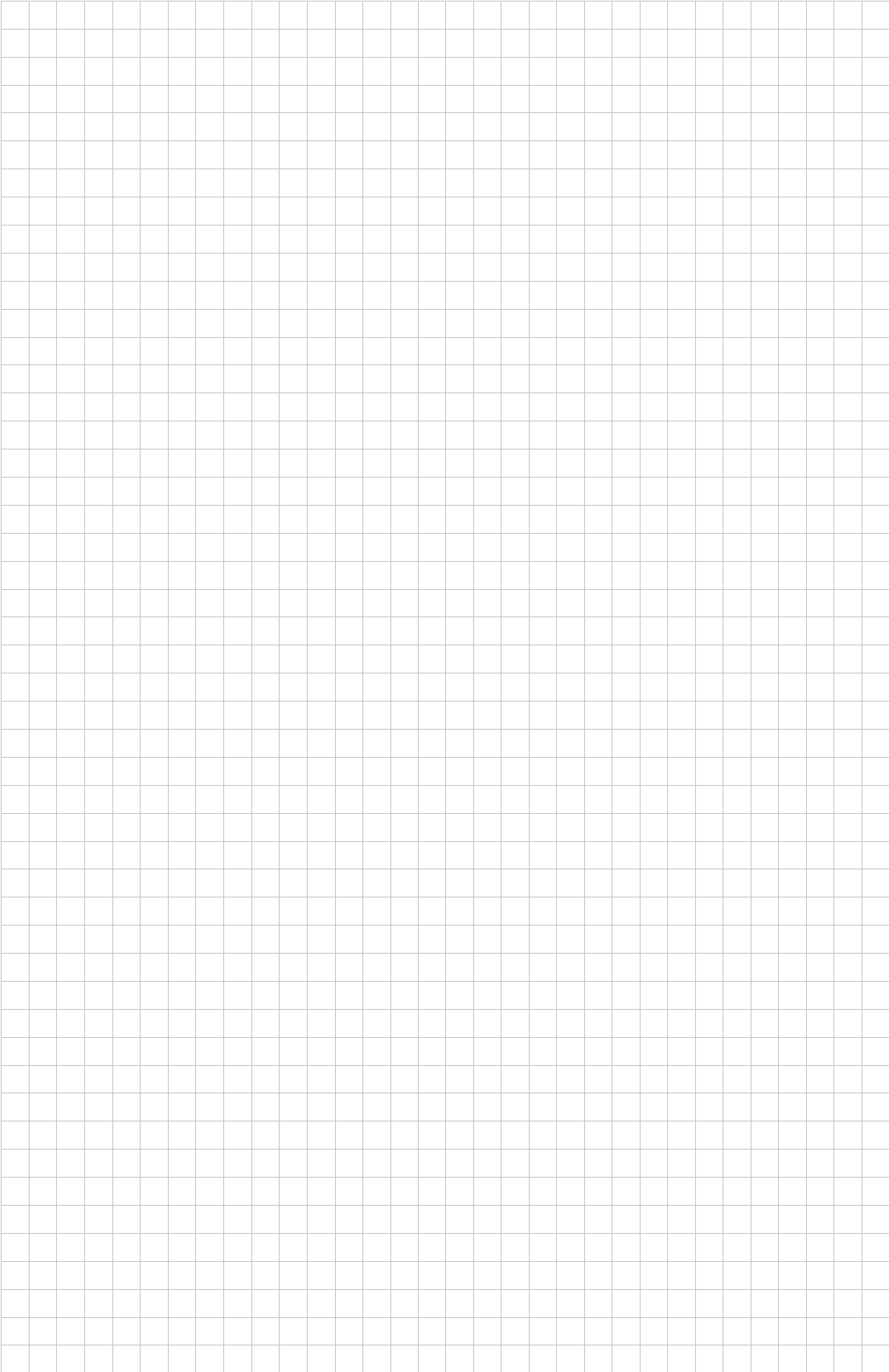
Évidemment, les deux rentes ci-dessus ont la même fin car étant donné que c'est jusqu'à la mort, c'est difficile de payer en postnumerando de la mort...

Signalons encore qu'il est possible mathématiquement de créer une rente temporaire ET différée mais c'est rare et donc nous n'en ferons pas l'étude ici.

Remarque: Contrairement aux rentes certaines, il n'existe pas ici de formule simplifiée permettant un calcul rapide de ces rentes basées sur les tables de mortalité. L'utilisation d'un tableur ou la programmation de ces fonctions permet de calculer facilement les valeurs actuelles des prestations viagères. À l'époque où l'informatique n'avait pas cours, on construisait des tables appelées "nombre de commutations" qui permettaient d'obtenir les informations voulues.







68. MUSIQUE MATHÉMATIQUE (ACOUSTIQUE)

La théorie de la musique est l'ensemble des aspects théoriques d'un système musical particulier. Il existe, non pas une, mais une infinité de théories musicales, chaque type de musique possédant la sienne. Tout système musical repose en effet sur un certain nombre d'usages, plus ou moins contraignants, susceptibles de faire l'objet d'une théorisation, orale ou écrite.

Une théorie de la musique possède fréquemment un point de départ religieux, philosophique ou magique ; d'autres fois, un point de départ arithmétique ou scientifique (acoustique). C'est à cette dernière que nous nous intéresserons ici évidemment...

Nous allons pour commencer considérer dans ce chapitre les ondes élastiques dans un gaz, résultant des variations de pression dans le gaz. Le son constitue l'exemple le plus important de ce type d'ondes qui fait partie de notre environnement quotidien.

1. ONDES SONORES LONGITUDINALES

Dans les milieux élastiques: les gaz et les liquides, des ondes sonores longitudinales se propagent suivant le mécanisme suivant (pour les solides il s'agit d'ondes transversales déjà étudiées dans le chapitre de Mécanique des Milieux Continus): une couche du milieu se déplace dans le sens de propagation de l'onde (d'où le nom de longitudinal) et comprime la couche suivante, laquelle sous l'effet de la pression avance et comprime la couche suivante et ainsi de suite. Cela fonctionne aussi pour une couche qui recule: la pression sur la couche suivante diminue ce qui a pour effet de faire reculer la couche suivante laquelle diminue à son tour la pression sur la suivante, etc.

La vitesse à laquelle se déplacent ces ondes longitudinales dépend comme souvent des caractéristiques, du milieu. En général, elle est plus faible dans les gaz que dans les liquides et plus faible dans les liquides que dans les solides. Par exemple $\sim 340 \left[m \cdot s^{-1} \right]$ dans l'air, $\sim 1'500 \left[m \cdot s^{-1} \right]$ dans l'eau et $\sim 5'000 \left[m \cdot s^{-1} \right]$ pour les ondes transversales dans l'acier.

En ce qui concerne les fréquences, il n'y a presque pas de limites. On peut générer des ondes sonores à des fractions de Hz jusqu'à des centaines de MHz. Par référence à la gamme de fréquences audibles à l'être humain, nous appelons "infrasons" des fréquences inférieures à 20 [Hz] et "ultrasons" des fréquences supérieures à 20 [KHz].

En général, les ondes sont produites par une source de dimensions limitées et, à partir de cette source, se propagent dans toutes les directions. Dans des milieux isotropes le front d'onde d'une perturbation est sphérique. Nous éviterons d'introduire des coordonnées sphériques en nous limitant à traiter des parties de fronts d'onde suffisamment éloignées de la source et suffisamment petites devant la distance à la source pour que nous puissions assimiler le morceau de sphère à une surface plane. Autrement dit, nous ne traiterons que les ondes planes longitudinales.

Il existe aussi une autre différence importante entre les ondes élastiques longitudinales dans un gaz ou liquide et les ondes élastiques transversales dans un barreau solide. Les gaz sont très compressibles, et si des fluctuations de pression s'établissent dans un gaz, sa densité subira le même type de fluctuation que la pression.

Considérons les ondes se propageant à l'intérieur d'un tuyau ou tube cylindrique (horizontal) de section S . Notons P_0 et ρ_0 la pression et la masse volumique à l'équilibre du gaz. Dans ces conditions d'équilibre, P_0 et ρ_0 sont les mêmes dans tout le volume gazeux du cylindre, c'est-à-dire indépendants de x .

Si la pression du gaz est perturbée par exemple à l'une des deux ouvertures de cylindre creux, un élément de volume de celui-ci Sdx sera intuitivement mis en mouvement parce que les pressions P et P' sur les deux faces S , S' de ce petit volume seront différentes et produiront donc une force résultante.

Remarque: Même si elles ont une très grande vitesse, dans un gaz les molécules subissent des chocs très fréquents les uns avec les autres. Elles parcourent au fait moins d'un micron en moyenne (libre parcours moyen), dans les conditions normales, avant d'en heurter une autre.

Il en résulte un déplacement de la section S d'une quantité ξ et de la section S' d'une quantité différente ξ' nécessairement différente car l'équilibre de pression n'aura pas eu le temps de se faire.

Ainsi, l'élément de volume au début a une largeur dx mais après la variation de pression, il aura une largeur si les variations de pression sont petites en première approximation:

$$dx + (\xi - \xi') = dx + d\xi \quad (68.1)$$

Cependant, en raison de la variation du volume, il y a également à présent une variation de densité due à la compressibilité du gaz. La masse contenue dans le volume non perturbé est initialement:

$$M_0 = \rho_0 S dx \quad (68.2)$$

Si ρ est la masse volumique du gaz perturbé, la masse du volume perturbé vaut au final:

$$M = \rho S (dx + d\xi) \quad (68.3)$$

La conservation de la matière demande que ces deux masses soient égales, c'est-à-dire que:

$$\rho S (dx + d\xi) = \rho_0 S dx \quad (68.4)$$

ou:

$$\rho \left(1 + \frac{d\xi}{dx} \right) = \rho_0 \quad (68.5)$$

En résolvant en ρ nous obtenons:

$$\rho = \frac{\rho_0}{1 + \frac{d\xi}{dx}} \quad (68.6)$$

Comme nous considérons les variations de pression petites par rapport à la pression ambiante, $\frac{d\xi}{dx}$ est petit, nous pouvons remplacer:

$$\rho = \rho_0 \frac{1}{1 + \frac{d\xi}{dx}} \quad (68.7)$$

par son développement limité de Taylor (cf. chapitres de Suites Et Séries):

$$\rho = \rho_0 \left(1 - \frac{d\xi}{dx} + \left(\frac{d\xi}{dx} \right)^2 - \dots \right) \cong \rho_0 \left(1 - \frac{d\xi}{dx} \right) \quad (68.8)$$

Soit:

$$\frac{\rho - \rho_0}{\rho_0} = - \frac{d\xi}{dx} \quad (68.9)$$

En admettant maintenant que la pression est uniquement reliée à la masse volumique (pour faire simple) nous pouvons écrire:

$$P = f(\rho) \quad (68.10)$$

En utilisant la forme générale du développement de Taylor (cf. chapitre de Suites et Séries):

$$f(x) = \sum_{k=0}^n \frac{1}{k!} \frac{d^k f(x_0)}{dx^k} (x - x_0)^k \quad (68.11)$$

Nous avons alors:

$$P = \sum_{k=0}^n \frac{1}{k!} \frac{d^k f(\rho_0)}{d\rho^k} (\rho - \rho_0)^k \cong P_0 + (\rho - \rho_0) \frac{dP(\rho_0)}{d\rho} \quad (68.12)$$

La quantité:

$$\frac{1}{\kappa} = \rho_0 \left(\frac{dP}{d\rho} \right)_{\rho=\rho_0} \quad (68.13)$$

plus souvent notée:

$$\kappa = \frac{1}{\rho_0} \left(\frac{d\rho}{dP} \right)_{\rho=\rho_0} \quad (68.14)$$

est appelée "**coefficient de compressibilité**" ou plus techniquement "**coefficient de compressibilité isotherme**".

Rappelons maintenant que nous avons démontré dans le chapitre de Génie Météo à partir de la loi de Laplace lors de notre étude du modèle atmosphérique adiabatique que:

$$\frac{dP}{d\rho} = \gamma \frac{P}{\rho_0} \quad (68.15)$$

où pour rappel, γ est le "**coefficient de Laplace**", appelé aussi "**coefficient adiabatique**" défini par:

$$\gamma = \frac{C_P}{C_V} \quad (33.16)$$

Alors nous obtenons (relation que nous utiliserons plus loin):

$$\frac{1}{\kappa} = \rho_0 \gamma \frac{P}{\rho_0} \Rightarrow \frac{1}{\kappa} = \gamma P \quad (68.17)$$

Le coefficient de compressibilité isotherme s'écrira bien évidemment dans le cas général:

$$\kappa = \frac{1}{\rho} \left(\frac{d\rho}{dP} \right) \quad (68.18)$$

Au passage, remarquons qu'en passant de la masse volumique au volume, il vient:

$$\kappa = \frac{1}{\rho} \left(\frac{d\rho}{dP} \right) = \frac{1}{\left(\frac{m}{V} \right)} \left(\frac{d \left(\frac{m}{V} \right)}{dP} \right) = V \frac{d \left(\frac{1}{V} \right)}{dP} = V \frac{-\frac{1}{V^2} dV}{dP} = -\frac{1}{V} \left(\frac{dV}{dP} \right) \quad (68.19)$$

souvent notée χ_T dans le domaine de la Thermodynamique.

Revenons aux deux relations:

$$\frac{1}{\kappa} = \rho_0 \left(\frac{dP}{d\rho} \right)_{\rho=\rho_0} \quad \text{et} \quad \frac{\rho - \rho_0}{\rho_0} = -\frac{d\xi}{dx} \quad (68.20) \quad (68.21)$$

et injectons les dans le développement de Taylor fait plus haut. Nous avons alors:

$$P \cong P_0 + \frac{(\rho - \rho_0)}{\rho_0} \frac{1}{\kappa} = P_0 + \frac{-\rho_0 \frac{d\xi}{dx}}{\rho_0} \frac{1}{\kappa} = P_0 - \frac{d\xi}{dx} \frac{1}{\kappa} \quad (68.22)$$

Cette expression relie la pression en tout point du gaz à la déformation au même point.

Nous avons ensuite besoin de l'équation du mouvement de l'élément de volume. La masse de l'élément est

$\rho_0 S dx$ et son accélération $\frac{d^2 \xi}{dt^2}$.

Nous avons naturellement en termes de force (le signe moins indiquant que la force qui varie la pression s'oppose à la pression initiale dans le cylindre):

$$-S dP = (\rho_0 S dx) \frac{d^2 \xi}{dt^2} \quad (68.23)$$

soit:

$$\frac{dP}{dx} = -\rho_0 \frac{d^2 \xi}{dt^2} \quad (68.24)$$

Dans ce problème, il y a deux champs, le champ des déplacements ξ et le champ des pressions P . Nous pouvons les combiner de la manière suivante en prenant:

$$P \cong P_0 - \frac{d\xi}{dx} \frac{1}{\kappa} \quad (68.25)$$

et en dérivant par rapport à x et en nous rappelant que P_0 est indépendante de la position dans le gaz. Nous avons alors:

$$\frac{dP}{dx} \cong -\frac{1}{\kappa} \frac{d^2 \xi}{dx^2} \quad (68.26)$$

Ce que nous pouvons combiner avec:

$$\frac{dP}{dx} = -\rho_0 \frac{d^2 \xi}{dt^2} \quad (68.27)$$

Nous retrouvons donc au final une équation d'onde de la forme (rappel) d'une équation de Poisson (plus particulièrement il s'agit d'une équation de d'Alembert):

$$\boxed{\frac{d^2 \xi}{dt^2} = \frac{1}{\kappa \rho_0} \frac{d^2 \xi}{dx^2}} \quad (68.28)$$

Nous obtenons donc une relation similaire à ce que nous avons obtenu dans le chapitre de Mécanique Ondulatoire pour les ondes mécaniques ou dans le chapitre d'Électrodynamique dans l'équation de propagation des ondes. Nous en concluons que le déplacement dû à une perturbation de pression dans un gaz se propage à la vitesse:

$$v = \sqrt{\frac{1}{\kappa \rho_0}} = \sqrt{\frac{\gamma P}{\rho_0}} = \sqrt{\frac{\gamma}{\rho_0} \frac{n k T}{V}} = \sqrt{\frac{\gamma N k T}{M}} = \sqrt{\frac{\gamma n R T}{M}} = \sqrt{\frac{\gamma R T}{M_m}} \quad (68.29)$$

où pour rappel la constante des gaz parfaits vaut $8.314 \left[J \cdot K^{-1} \cdot mol^{-1} \right]$.

La relation:

$$\boxed{v = \sqrt{\frac{\gamma P}{\rho_0}}} \quad (68.30)$$

est parfois appelée "**relation de Newton-Laplace**".

Si nous considérons l'air comme un gaz parfait diatomique alors (cf. chapitre de Thermodynamique) nous avons $\gamma \cong 7/5 = 1.4 \dots$ de masse molaire $29 \cdot 10^3 \left[\text{kg} \cdot \text{mol}^{-1} \right]$ (moyenne pondérée des masses molaires de N_2 et O_2) il vient à une température de 300 [K]:

$$v = 347 \left[\text{m} \cdot \text{s}^{-1} \right] \quad (68.31)$$

Ce qui est en parfait accord avec l'expérience! Par contre, lorsque l'on dit qu'un avion vole à Mach 2 (le Mach est le rapport entre la vitesse d'un avion et celle du son), on ne connaît pas vraiment la vitesse du son (ni de l'avion) ni la température.

Or, nous avons aussi le résultat suivant (cf. chapitre de Mécanique des Milieux Continus):

$$\kappa = \frac{3(1-2\eta)}{E} \quad (68.32)$$

Or, nous avons démontré dans le chapitre de Génie Mécanique que le coefficient de Poisson η respectait:

$$\eta < \frac{1}{2} \quad (68.33)$$

En prenant l'approximation que pour un gaz $\eta \cong \frac{1}{3} \dots$ nous avons alors:

$$\kappa = \frac{1}{E} \quad (68.34)$$

et donc:

$$v = \sqrt{\frac{E}{\rho_0}} \quad (68.35)$$

La vitesse du son est alors donnée par le même type d'expression pour les fluides ou les solides! Ainsi, la propagation d'une déformation longitudinale dans un solide est donnée par:

$$\frac{d^2 \xi}{dt^2} = \frac{E}{\rho_0} \frac{d^2 \xi}{dx^2} \quad (68.36)$$

et la déformation transversale (cf. chapitre de Mécanique Des Milieux Continus):

$$\frac{\partial^2 y}{\partial t^2} = \frac{G}{\rho_0} \frac{\partial^2 y}{\partial x^2} \quad (68.37)$$

Nous avons aussi:

$$\begin{aligned}\frac{\partial^2}{\partial t^2} P &\cong \frac{\partial^2}{\partial t^2} \left(P_0 - \frac{1}{\kappa} \frac{d\xi}{dx} \right) = -\frac{1}{\kappa} \frac{\partial^2}{\partial t^2} \left(\frac{d\xi}{dx} \right) \\ \frac{\partial^2}{\partial x^2} P &= \frac{\partial}{\partial x} \left(-\rho_0 \frac{d^2 \xi}{dt^2} \right) = -\rho_0 \frac{\partial}{\partial x} \left(\frac{d^2 \xi}{dt^2} \right) = -\rho_0 \frac{\partial^2}{\partial x^2} \left(\frac{\partial \xi}{\partial x} \right)\end{aligned}\quad (68.38)$$

En divisant ces deux égalités membre à membre, nous obtenons:

$$\frac{\frac{\partial^2 P}{\partial t^2}}{\frac{\partial^2 P}{\partial x^2}} = \frac{-\frac{1}{\kappa} \frac{\partial^2}{\partial t^2} \left(\frac{d\xi}{dx} \right)}{-\rho_0 \frac{\partial^2}{\partial x^2} \left(\frac{\partial \xi}{\partial x} \right)} \quad (68.39)$$

Soit après simplification:

$$\boxed{\frac{\partial^2 P}{\partial t^2} = \frac{1}{\kappa \rho_0} \frac{\partial^2 P}{\partial x^2}} \quad (68.40)$$

Pour l'usage en laboratoire, ce modèle d'onde sonore est plus pratique que le précédent, car nous mesurons plus facilement des variations de pression dans un liquide ou un gaz, que des déplacements de molécules. Il est intéressant de savoir alors que toutes les analyses que nous avons faits avec l'équation d'onde des cordes (cf. chapitre de Mécanique Ondulatoire) ou des ondes (cf. chapitre d'Électrodynamique) peut alors s'appliquer aussi à l'équation ci-dessus (aspect énergétique, solutions particulières, diminution de l'amplitude en fonction de la distance à la source, etc.).

Le mouvement des ondes dans les gaz est un processus adiabatique (cf. chapitre de Thermodynamique) donc il n'y a aucun échange d'énergie sous forme de chaleur par élément de volume du gaz.

1.1. PUISSANCE TRANSPORTÉE PAR UNE ONDE SONORE

Nous avons vu plus haut que:

$$\frac{\rho - \rho_0}{\rho_0} = -\frac{d\xi}{dx} \quad (68.41)$$

Notons les variations de densité dues à l'onde sonore par:

$$\rho_s = -\rho_0 \frac{d\xi}{dx} \quad (68.42)$$

Nous ne ferons le calcul de la puissance transportée que pour le cas des ondes sinusoïdales. Dans ce cas ξ sera de la forme:

$$\xi = \xi_0 \cos(\omega t - kx) \quad (68.43)$$

Comme ξ est connu, nous pouvons calculer la pression correspondante en utilisant:

$$P \cong P_0 - \frac{d\xi}{dx} \frac{1}{\kappa} \quad (68.44)$$

En introduisant les variations de pression dues à l'onde sonore:

$$P_s = - \frac{1}{\kappa} \frac{d\xi}{dx} \quad (68.45)$$

et en utilisant:

$$v = \sqrt{\frac{1}{\kappa \rho_0}} \quad (68.46)$$

Il vient:

$$P_s = -\rho_0 v^2 \frac{d\xi}{dx} \quad (68.47)$$

Nous avons alors:

$$P_s = -\rho_0 v^2 \frac{d(\xi_0 \cos(\omega t - kx))}{dx} = -\rho_0 k v^2 \xi_0 \sin(\omega t - kx) = -P_0 \sin(\omega t - kx) \quad (68.48)$$

Pour calculer la puissance transportée par une onde sonore, nous allons calculer le travail effectué, pendant une période, sur une surface S située sur un plan perpendiculaire à l'axe des x et de coordonnées x .

La force exercée sur cette surface sera:

$$F = S \cdot P_s \quad (68.49)$$

Le travail effectué quand cette surface se déplace de $d\xi$ sera:

$$dW = F d\xi = S \cdot P_s \cdot d\xi \quad (68.50)$$

Comme nous faisons un calcul pour les déplacements ξ d'une couche dont la position d'équilibre x ne varie pas, seule la variable t varie:

$$d\xi = \frac{\partial \xi}{\partial t} dt = -\omega \xi_0 \sin(\omega t - kx) dt \quad (68.51)$$

En remplaçant nous obtenons:

$$\begin{aligned} dW &= S \cdot P_s \cdot d\xi = -S \cdot P_s \cdot \omega \xi_0 \sin(\omega t - kx) dt \\ &= -S \cdot (-\rho_0 k v^2 \xi_0 \sin(\omega t - kx)) \cdot \omega \xi_0 \sin(\omega t - kx) dt \\ &= S \rho_0 k v^2 \omega \xi_0^2 \sin^2(\omega t - kx) dt \end{aligned} \quad (68.52)$$

Le travail exercé pendant une période:

$$T = \frac{2\pi}{\omega} \quad (68.53)$$

sera:

$$W = S \rho_0 k v^2 \omega \xi_0^2 \int_0^{\frac{2\pi}{\omega}} \sin^2(\omega t - kx) dt \quad (68.54)$$

Nous avons déjà démontré l'expression de ce type d'intégrale dans le chapitre de Calcul Différentiel Et Intégral:

$$\int \sin^2(\omega t - kx) dt = \frac{1}{\omega} \int \sin^2(y) dy = \frac{1}{2\omega} (-\cos(y) \sin(y) + y) \quad (68.55)$$

d'où:

$$\begin{aligned} \int_0^{\frac{2\pi}{\omega}} \sin^2(\omega t - kx) dt &= \left(\frac{1}{2\omega} (-\cos(\omega t - kx) \sin(\omega t - kx) + (\omega t - kx)) \right) \Bigg|_0^{\frac{2\pi}{\omega}} \\ &= \frac{1}{2\omega} \left(-\cos\left(\omega \frac{2\pi}{\omega} - kx\right) \sin\left(\omega \frac{2\pi}{\omega} - kx\right) + \left(\omega \frac{2\pi}{\omega} - kx\right) \right) \\ &\quad - \frac{1}{2\omega} (-\cos(-kx) \sin(-kx) + (-kx)) \\ &= \frac{1}{2\omega} (\cos(kx) \sin(kx) + 2\pi - kx) - \frac{1}{2\omega} (\cos(kx) \sin(kx) - kx) \\ &= \frac{\pi}{\omega} \end{aligned} \quad (68.56)$$

En remplaçant il vient:

$$W = S \rho_0 k v^2 \omega \xi_0^2 \frac{\pi}{\omega} = S \rho_0 k v^2 \pi \xi_0^2 \quad (68.57)$$

Pour calculer la puissance, il faut diviser ce travail par le temps dans lequel il a été effectué:

$$T = \frac{2\pi}{\omega} \quad (68.58)$$

et pour calculer la puissance transmise par unité de surface, il faut diviser par la surface:

$$\frac{P}{S} = \frac{S \rho_0 k v^2 \pi \xi_0^2}{\frac{2\pi}{\omega} \cdot S} = \frac{\rho_0 k v^2 \omega}{2} \xi_0^2 \quad (68.59)$$

Si nous remplaçons k par:

$$k = \frac{\omega}{v} \quad (68.60)$$

Nous obtenons:

$$\frac{P}{S} = \frac{\rho_0 v \omega^2}{2} \xi_0^2 \quad (68.61)$$

Avant de continuer, signalons qu'il existe plusieurs façons de mesurer l'amplitude d'un son, et par extension, d'un signal quelconque de nature ondulatoire:

- l'amplitude moyenne (la valeur moyenne arithmétique du signal positif)
- l'amplitude efficace (amplitude continue équivalente en puissance)
- l'amplitude crête (maximale positive)
- l'amplitude crête à crête (l'écart maximal d'amplitude positive et négative)

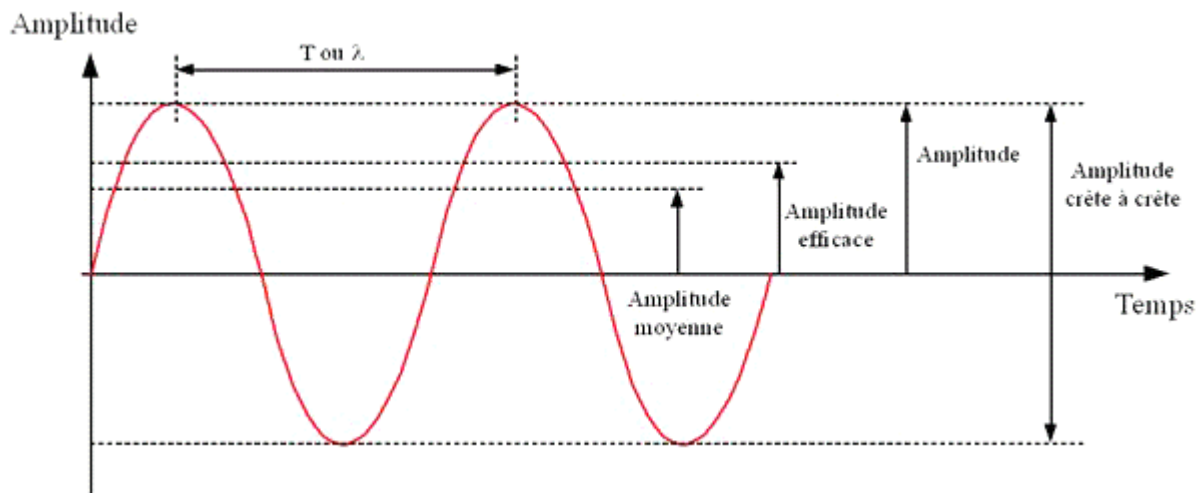


Figure: 68.1 - Types d'amplitudes

Dans la pratique, l'amplitude moyenne présente peu d'intérêt et n'est pas utilisée. En revanche, la valeur efficace ou RMS, pour Root Mean Square en anglais, soit la racine carrée de la valeur quadratique moyenne du signal est universellement adoptée pour mesurer la valeur des tensions alternatives, dans le cadre général autant qu'en acoustique. Un amplificateur qui est donné pour 10 watts RMS fera 14 watts en crête et 28 watts en crête à crête (aussi noté cc). Les mesures de puissance crête à crête sont assez souvent appelées "watts musicaux" par les vendeurs de matériel audiovisuel, car les chiffres sont plus flatteurs.

L'unité de mesure de l'amplitude dépend de la grandeur physique mesurée:

- Pour une corde vibrante, c'est une distance.
- Pour une onde sonore, c'est la pression de l'air, ou des mouvements du diaphragme
- Pour le rayonnement électromagnétique, l'amplitude correspond au champ électrique.
- Pour un signal électrique, cela correspond à la valeur maximale.

Pour les régimes sinusoïdaux (ou plus généralement périodiques!), il est facile de démontrer que quel que soit le domaine de la physique la valeur efficace (racine carrée de la moyenne du carré du signal) est la valeur de crête divisée par la racine carrée de deux:

$$\xi_{\text{eff}} = \sqrt{\frac{1}{T} \int_0^T \xi^2(t) dt} = \sqrt{\frac{1}{2\pi} \int_0^{2\pi} \xi^2(\theta) d\theta} = \sqrt{\frac{1}{2\pi} \xi_0^2 \int_0^{2\pi} \cos^2(\theta) d\theta} = \sqrt{\frac{1}{2\pi} \xi_0^2 \pi} = \frac{\xi_0}{\sqrt{2}} \quad (68.62)$$

Dons en valeur efficace (il s'agit uniquement d'un choix arbitraire et rien d'autre il faut juste savoir de quoi on parle par la suite!!), la relation antérieure s'écrit:

$$\frac{P}{S} = \rho_0 v \omega^2 \xi_{\text{eff}}^2 = \rho_0 k v^2 \omega \xi_{\text{eff}}^2 \quad (68.63)$$

Comme nous avons l'amplitude de la pression qui est égale à:

$$P_0 = \rho_0 k v^2 \xi_0 = \rho_0 \omega v \xi_0 \quad (68.64)$$

et sa valeur efficace par:

$$P_{\text{eff}} = \rho_0 k v^2 \xi_{\text{eff}} = \rho_0 \omega v \xi_{\text{eff}} \quad (68.65)$$

Finalement, nous avons pour la puissance (que nous représenterons par la suite par un P majuscule stylisé afin de ne pas confondre avec la pression):

$$\frac{\mathcal{P}}{S} = \rho_0 v \omega^2 \xi_{\text{eff}}^2 = \frac{1}{\rho_0 v} \left(\rho_0 v \omega \xi_{\text{eff}} \right)^2 = \frac{1}{\rho_0 v} P_{\text{eff}}^2 \quad (68.66)$$

1.2. MESURE DE L'INTENSITÉ DU SON

Pour caractériser le son, on a inventé une unité de mesure: le "**Bel**" et son sous-multiple le "**décibel**" qui vaut 1/10 de Bel. Cette intensité acoustique, assimilée à la "**sonie**" ou "**force sonore**", a été définie à partir de la pression sonore efficace (le P_{eff} donné plus haut) et de la pression sonore d'une onde (à une fréquence précise!) à la limite du seuil auditif d'une petite élite de l'humanité (environ 10%).

Cette pression efficace de référence vaut:

$$P_{\text{eff},\text{réf}} = 2 \cdot 10^{-5} \text{ [Pa]} \quad (68.67)$$

L'intensité sonore d'une onde dont la pression acoustique efficace est P_{eff} vaut par définition et par convention:

$$I_s = \log_{10} \left(\frac{P_{\text{eff}}}{P_{\text{eff},\text{réf}}} \right)^2 = 2 \log_{10} \left(\frac{P_{\text{eff}}}{P_{\text{eff},\text{réf}}} \right) = 2 \log_{10} \left(\frac{P_{\text{eff}}}{2 \cdot 10^{-5}} \right) \text{ [Bels]} \quad (68.68)$$

et est souvent notée simplement S dans la littérature spécialisée pour signaler qu'il s'agit de la "**sensation physiologique**" de l'oreille humaine. Comment est-on arrivé à cette relation? Eh bien en constatant sur un panel d'individus que la sensation physiologique de l'oreille n'était pas linéaire mais que les variations absolues de la sensation sont proportionnelles aux variations relatives de pression telles que:

$$dI_s = c^{te} \frac{dP}{P} \quad (68.69)$$

Soit après intégration:

$$I_s - I_{s,0} = c^{te} \ln \left(\frac{P}{P_0} \right) \quad (68.70)$$

où par convention la sensibilité minimale est posée comme étant nulle et pour des raisons historiques on prend les logarithmes décimaux au lieu des népériens. Il vient alors:

$$I_s = c^{te} \log_{10} \left(\frac{P}{P_0} \right) \quad (68.71)$$

Dans certaines sources, nous trouvons la définition du Bel à partir d'une intensité de référence de (à 1 [KHz]):

$$10^{-12} \left[W \cdot m^{-2} \right] \quad (68.72)$$

Il vient alors (la définition est équivalente puisqu'il s'agit d'un rapport et que la puissance est directement proportionnelle au carré de la pression comme nous l'avons démontré plus haut):

$$I_s = \log_{10} \left(\frac{P / S}{P_{réf}} \right)^2 = \log_{10} \left(\frac{P_{eff}}{10^{-12}} \right)^2 [Bels] \quad (68.73)$$

En réalité, le Bel est rarement utilisé et on lui préfère le décibel (la plupart des êtres humains peuvent à ce jour déceler une différence d'intensité entre deux sons dont l'intensité diffère de 1 [dB]):

$$I_s = 10 \cdot \log_{10} \left(\frac{P_{eff}}{P_{eff,réf}} \right)^2 = 20 \cdot \log_{10} \left(\frac{P_{eff}}{2 \cdot 10^{-5}} \right) [déciBels] \quad (68.74)$$

Le décibel est donc par construction un dixième de Bel... d'où son nom...

La raison de ce choix logarithmique est que la sensation auditive est aussi logarithmique: on a la même impression d'augmentation du son quand celui-ci passe de 1 à 10 que quand il passe de 10 à 100.

Pour donner une idée de la valeur d'un décibel, voici un tableau de valeurs typiques (pour une fréquence et une pression de référence donnée):

Décibels	Source
160	Dommages au tympan
140	Seuil de la douleur
120	Seuil de la gêne
100	Atelier de machines lourdes
85	Automobile
75	Usine moyenne
60	Conversation normale
40	Bruit des spectateurs au cinéma
20	Studio de radiodiffusion
10	Chambre anéchoïque
0	Seuil de l'audition

Tableau: 68.1 - Amplitudes relatives typiques de différentes sources sonores

Comme nous l'avons déjà dit la pression efficace sonore minimum qui provoque une sensation auditive est de:

$$2 \cdot 10^{-5} [Pa] \quad (68.75)$$

c'est-à-dire environ 10^{-10} fois plus petite que la pression atmosphérique. Il faut tout de même que la fréquence de cette onde se situe autour de 3 [KHz], là où la sensibilité est maximum.

Il est très intéressant de calculer à quelle amplitude efficace de déplacement cela correspond. Nous utilisons alors la relation démontrée plus haut:

$$P_{eff} = \rho_0 \omega v \xi_{eff} \quad (68.76)$$

Donc pour le cas limite et à 1 [KHz], aux conditions normales de température et de pression, nous aurons un déplacement de (valeurs prises dans les tables de la commission romande de mathématique):

$$\xi_{eff} = \frac{P_{eff}}{\rho_0 \omega v} = \frac{2 \cdot 10^{-5}}{1.293 \cdot 2\pi \cdot 1000 \cdot 331} \cong 7.43 \cdot 10^{-12} [m] \quad (68.77)$$

et la valeur crête sera donc de:

$$\xi_0 = \xi_{eff} \cdot \sqrt{2} \cong 1 \cdot 10^{-11} [m] \quad (68.78)$$

Cette valeur est inférieure au rayon des atomes selon le modèle de Dalton. La nature nous a doté d'un organe d'une sensibilité exquise!

Il est facile de calculer à quelle valeur de puissance par unité de surface correspond une onde sonore à la limite de l'audition. En utilisant la relation démontrée plus haut, nous obtenons:

$$\frac{\mathcal{P}}{S} = \frac{1}{\rho_0 v} P_{eff}^2 = \frac{1}{1.293 \cdot 331} (2 \cdot 10^{-5})^2 = 1 [pW \cdot m^{-2}] \quad (68.79)$$

Comme la surface de la section du canal auditif fait moins d'un cm^2 , la puissance qui arrive au tympan est inférieure à 10^{-16} [W] .

2. ONDES SPHÉRIQUES

Dans la réalité les ondes sonores sont générées par des sources d'étendue finie. À des distances plus faibles ou comparables à l'étendue de la source, la forme du front d'onde qui s'éloigne de la source peut être très compliquée (imaginez le front d'onde d'un tonnerre provoqué par un éclair tarabiscoté). Mais, loin de la source, elle est vue comme un objet ponctuel et le front d'onde devient de plus en plus sphérique à mesure que l'on s'éloigne de la source.

Si nous regardons un petit morceau de la sphère, la petite surface sphérique sera très peu incurvée et nous pourrions l'assimiler à une surface plane sans faire trop d'erreur. Nous retrouverons les ondes planes... mais pas tout à fait!

La différence est que l'énergie transportée par l'onde sonore est distribuée dans une surface qui s'agrandit comme R^2 (où R est la distance à la source). Donc la puissance par unité de surface doit diminuer comme $1/R^2$ (exactement selon le même développement mathématique que celui pour les ondes électromagnétiques vues dans le chapitre d'Électrodynamique) et comme la puissance par unité de surface est proportionnelle à ξ_{eff}^2 ou à P_{eff}^2 , cela implique que dans une onde sphérique, aussi bien le déplacement ξ_{eff} , que la pression sonore P_{eff} , doivent diminuer comme $1/R$.

Dans le cas des ondes sinusoïdales, il faudra modifier l'équation d'onde:

$$\xi = \xi_0 \cos(\omega t - kx) \quad (68.80)$$

en ajoutant un coefficient $1/R$. Si nous appelons r la distance entre la source et la position d'observation, et que, à la place de mesurer la position avec x nous le faisons avec r :

$$\xi = \xi_0 \frac{\cos(\omega t - kr)}{r} \quad (68.81)$$

et de même pour la pression sonore:

$$P_s = -P_0 \frac{\sin(\omega t - kr)}{r} \quad (68.82)$$

3. EFFET DOPPLER

Dans ce qui a été étudié jusqu'ici, la source S et l'observateur O de l'onde ont toujours été considérés comme immobiles l'un par rapport à l'autre. L'effet Doppler explique les variations de fréquences de l'onde observée lorsque S et O sont en mouvement relatif.

3.1. SOURCE FIXE-OBSERVATEUR EN MOUVEMENT

Considérons donc le premier cas où la source est fixe et l'observateur en mouvement rectiligne et uniforme:

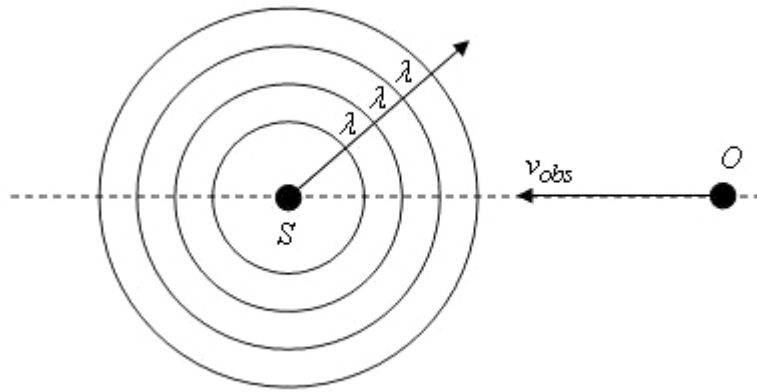


Figure: 68.2 - Schéma de principe d'une source fixe et d'un observateur en approche

Soient:

$$f = \frac{1}{T} \quad (68.83)$$

la fréquence de la source S et v_{phase} la vitesse de propagation de l'onde dans le milieu, supposé immobile.

Si O reste immobile, il voit arriver les fronts d'onde à la vitesse v_{phase} et il n'y a pas grand-chose à dire. Si O se dirige vers S à vitesse v_{obs} , il reçoit les fronts d'onde à la vitesse:

$$v_{phase} + v_{obs} \quad (68.84)$$

et franchit les distances λ en des temps égaux:

$$\Delta t = \frac{\lambda}{v_{phase} + v_{obs}} \quad (68.85)$$

Pendant chaque intervalle Δt il perçoit un cycle complet de l'onde. Il en déduit que Δt est la période T' d'une onde de fréquence f' :

$$f'_{approche} = \frac{1}{T'} = \frac{v_{phase} + v_{obs}}{\lambda} = \frac{v_{phase} + v_{obs}}{v_{phase} \cdot T} = \frac{1}{T} \left(1 + \frac{v_{obs}}{v_{phase}} \right) \quad (68.86)$$

Donc au final:

$$f'_{approche} = f \left(1 + \frac{v_{obs}}{v_{phase}} \right) > f \quad (68.87)$$

Le son devient donc plus aigu lorsqu'on approche une source.

Si O s'éloigne de S à vitesse v_{obs} , il reçoit les fronts d'onde à la vitesse:

$$v_{phase} - v_{obs} \quad (68.88)$$

et les mêmes développements nous amènent alors à:

$$f'_{\text{éloignement}} = f \left(1 - \frac{v_{\text{obs}}}{v_{\text{phase}}} \right) \leq f \quad (68.89)$$

La limite de validité de cette relation étant pour $v_{\text{obs}} \leq v_{\text{phase}}$. Le son devient donc plus grave lorsqu'on s'éloigne d'une source.

Remarque: Si la vitesse de l'observateur est orientée de manière quelconque, par rapport à la ligne $\overline{OS}$, seule la composante radiale est à prendre en compte.

3.2. SOURCE EN MOUVEMENT-OBSERVATEUR FIXE

Il n'est pas forcément intuitif de voir que le système n'est pas symétrique. Mais en prenant un cas particulier on se rend vite compte que ça ne l'est effectivement pas. En effet, si le récepteur fuit l'émetteur à une vitesse supérieure à v_{phase} , il ne recevra jamais d'onde, alors que si l'émetteur fuit un récepteur immobile, celui-ci recevra toujours une onde. On ne peut pas inverser le rôle de l'émetteur et du récepteur.

Dans le cas classique, il y a donc dissymétrie dans le décalage fréquentiel selon que l'émetteur ou le récepteur est en mouvement.

Considérons donc la situation suivante:

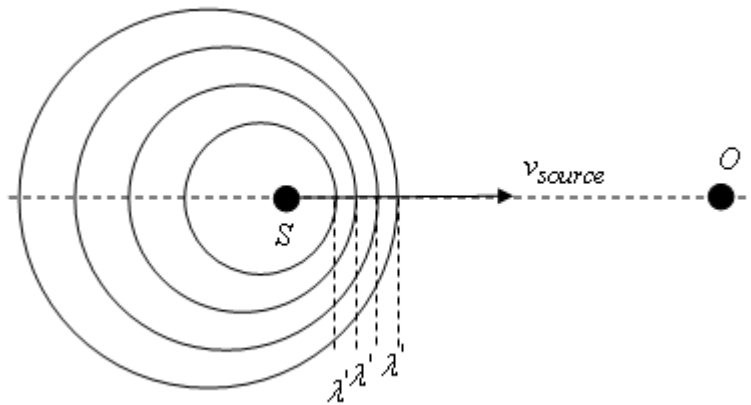


Figure: 68.3 - Schéma de principe d'un observateur fixe et d'une source en approche

Soient:

$$f = \frac{1}{T} \quad (68.90)$$

la fréquence de la source S en mouvement rectiligne et uniforme et v_{phase} la vitesse de propagation de l'onde dans le milieu et O l'observateur supposé immobile.

Si la source se déplace à vitesse v_{source} , elle parcourt une distance:

$$d = v_{source} \cdot T \quad (68.91)$$

entre chaque cycle. La longueur d'onde vue par l'observateur varie donc de:

$$d = \Delta\lambda = v_{source} \cdot T \quad (68.92)$$

Pour l'observateur, elle diminue telle que:

$$\lambda' = \lambda - v_{source} \cdot T = \lambda - \frac{v_{source}}{f} \quad (68.93)$$

Nous avons alors:

$$f'_{approche} = \frac{v_{phase}}{\lambda'} = \frac{v_{phase}}{\lambda - \frac{v_{source}}{f}} = \frac{v_{phase}}{\frac{v_{phase}}{f} - \frac{v_{source}}{f}} = f \left(\frac{v_{phase}}{v_{phase} - v_{source}} \right) \quad (68.94)$$

Donc après simplification:

$$f'_{approche} = f \left(\frac{1}{1 - \frac{v_{source}}{v_{phase}}} \right) > f \quad (68.95)$$

La limite de validité de cette relation étant pour $v_{source} \leq v_{phase}$.

Il vient donc:

$$f'_{approche} > f \quad (68.96)$$

Le son devient donc plus aigu lorsque la source s'approche (donc même sensation que lorsque c'est l'observateur qui s'approche).

De même, lorsque la source s'éloigne de l'observateur, nous avons immédiatement:

$$f'_{éloignement} = f \left(\frac{1}{1 + \frac{v_{source}}{v_{phase}}} \right) < f \quad (68.97)$$

Le son devient donc plus grave lorsque la source s'éloigne (donc même sensation que lorsque c'est l'observateur qui s'éloigne).

Remarque: Dans le cas d'ondes électromagnétiques, la vitesse de l'onde est la vitesse de la lumière qui ne dépend pas du référentiel. Nous devons alors traiter le problème dans le cadre de la relativité restreinte et on s'attend alors à trouver un effet parfaitement symétrique puisqu'on ne peut pas distinguer entre vitesse de l'émetteur et vitesse du récepteur, seule comptant la vitesse relative entre les deux (voir les développements dans le chapitre de Relativité Restreinte).

3.3. OBSERVATEUR ET SOURCE EN MOUVEMENT

Si les deux (source et observateur) sont en mouvement rectiligne sur une droite commune, les deux effets s'ajoutent. Nous avons alors lorsqu'ils se rapprochent:

$$\left. \begin{aligned} f'_{\text{approche}} &= f \left(1 + \frac{v_{\text{obs}}}{v_{\text{phase}}} \right) \\ f'_{\text{approche}} &= f \left(\frac{1}{1 - \frac{v_{\text{source}}}{v_{\text{phase}}}} \right) \end{aligned} \right\} f'_{\text{approche}} = f \left(1 + \frac{v_{\text{obs}}}{v_{\text{phase}}} \right) \left(\frac{1}{1 - \frac{v_{\text{source}}}{v_{\text{phase}}}} \right) \quad (68.98)$$

Et s'ils s'éloignent:

$$\left. \begin{aligned} f'_{\text{éloignement}} &= f \left(1 - \frac{v_{\text{obs}}}{v_{\text{phase}}} \right) \\ f'_{\text{éloignement}} &= f \left(\frac{1}{1 + \frac{v_{\text{source}}}{v_{\text{phase}}}} \right) \end{aligned} \right\} f'_{\text{éloignement}} = f \left(1 - \frac{v_{\text{obs}}}{v_{\text{phase}}} \right) \left(\frac{1}{1 + \frac{v_{\text{source}}}{v_{\text{phase}}}} \right) \quad (68.99)$$

4. ONDES DE CHOC

Les ondes de choc se produisent dans un gaz pour des perturbations très intenses. Par exemple, l'onde du choc du Concorde (bang supersonique) fait au niveau du sol une surpression d'environ 100 [Pa]. Le son est très fort, de l'ordre de 140 [dB], mais la surpression n'est que d'un millièème de la pression atmosphérique! Les ondes à la sortie des armes à feu ont des surpressions de l'ordre de quelques 10^5 [Pa].

Nous avons démontré que la vitesse des ondes dans un gaz dépendait de la température mais si nous prenons en compte que les ondes sonores produisent des variations de pression, celles-ci produisent donc elles-mêmes des variations de température (cf. chapitre de Thermodynamique). Donc, une augmentation de pression augmente la température ce qui augmente la vitesse. Pour le cas d'une sinusoïde, les sommets de pression se retrouvent à voyager plus vite que les creux... La sinusoïde se déforme, la pente entre les sommets et les creux de devant augmente ce qui a tendance à créer un front d'onde abrupt entre la partie chaude, derrière le front d'onde et la partie froide juste devant. C'est ce que l'on appelle "**onde de choc**". Cette onde de choc se propage à la vitesse qui correspond à sa température et qui est supérieure à celle du son normal.

À mesure que le front d'onde se propage, son amplitude diminue et à la fin, l'onde de choc devient une onde sonore normale. Ce qui est remarquable est que les ondes de choc, aussi bien celles qui sont en surpression (plus chaudes et plus rapides) que celles qui sont en dépression (plus froides et plus lentes) fusionnent avec les ondes qu'elles rattrapent ou qui les rattrapent. En effet, une onde normale qui se fait rattraper par une onde de choc se retrouve à voyager dans une zone chaude dont la vitesse est celle de l'onde qui l'a rattrapée.

5. GAMMES MUSICALES

Lorsque nous faisons vibrer une corde de guitare, de violon ou autre instrument à cordes dont les extrémités sont fixes (onde stationnaire), nous avons démontré dans le chapitre de Mécanique Ondulatoire que nous sommes dans des conditions dites de Dirichlet.

Nous avons obtenu après de nombreux calculs, le résultat suivant:

$$\omega_n = \sqrt{\frac{F}{\rho_l}} \left(\frac{n\pi}{L} \right) \quad (68.100)$$

Soit:

$$f_n = \frac{n}{2L} \sqrt{\frac{F}{\rho_l}} \propto \frac{1}{L} \quad (68.101)$$

donc la fréquence (qui est un des facteurs de la sensation de la "**hauteur**" d'un son chez l'être humain) est bien inversement proportionnelle à la longueur L d'une corde fixée aux extrémités (comme l'aurait déduit Pythagore bien avant que la théorie soit établie).

La relation précédente le met en évidence; plus la partie vibrante est longue (plus n est petit), plus le son sera bas (basse fréquence) et plus elle sera courte (n grand), plus le son sera aigu (haute fréquence).

Rappelons que pour $n = 1$, nous parlons de "fréquence fondamentale" et pour $n \geq 2$ de "fréquence harmonique". Ainsi, $n = 2$ est la première harmonique, $n = 3$ la deuxième harmonique, etc.

Définitions:

D1. Nous disons que deux fréquences (ou deux ondes périodiques) sont "**à l'unisson**" l'une de l'autre si le rapport de leurs fréquences est égal à l'unité:

$$Unisson: \frac{f''}{f'} = 1 \quad (68.102)$$

D2. Nous disons que deux fréquences (ou deux ondes périodiques) sont "**à l'octave**" l'une de l'autre si le rapport de leurs fréquences est égal à deux:

$$Octave: \frac{f''}{f'} = 2 \quad (68.103)$$

Remarquons que la première harmonique correspond à l'octave de la fondamentale:

$$\frac{f_1}{f_0} = 2$$

Ainsi, certaines chanteuses sont capables de jouer sur 8 octaves (c'est-à-dire un rapport de fréquences de 16). Ainsi, dans le cas d'un instrument à cordes avec conditions de Dirichlet cela équivaut bien évidemment à pincer la corde en son milieu (pour les guitares cela correspond à appuyer la corde sur la frette de la 12^{ème} case par rapport à la fréquence de la corde à vide pour avoir une octave).

D3. Nous disons que deux fréquences (ou deux ondes périodiques) sont "**à la quinte**" l'une de l'autre si le rapport de leurs fréquences est égal à 3/2:

$$\text{Quinte : } \frac{f''}{f'} = \frac{3}{2} \quad (68.104)$$

Hormis l'octave, cet intervalle est le plus simple de tous, et depuis toujours, dans la musique occidentale, il a été considéré comme l'intervalle consonant par excellence, donc le plus remarquable à l'oreille. Pour cette raison, la quinte a joué un rôle essentiel dans l'établissement des gammes musicales, la gamme de Pythagore (gamme pythagoricienne) étant même exclusivement construite sur cet intervalle particulier.

Nous remarquons que nous arrivons à mettre au plus un nombre entier de 2 quintes dans une octave (au-delà nous dépassons la valeur d'une octave).

D4. Nous disons que deux fréquences (ou deux ondes périodiques) sont "**à la quarte**" l'un de l'autre si le rapport de leurs fréquences est égal à 4/3:

$$\text{Quarte : } \frac{f''}{f'} = \frac{4}{3} \quad (68.105)$$

D5. Le rapport des fréquences de deux sons (ou deux ondes périodiques) éloignés l'un de l'autre d'un "**ton**" peut être de 9/8. Il s'agit alors du "**ton majeur**":

$$\text{Ton : } \frac{f''}{f'} = \frac{9}{8} \quad (68.106)$$

D6. Le "**ton mineur**" a pour rapport de fréquences 10/9.

Lorsqu'un instrument à cordes a ses vibrations à vides qui sont dans un rapport de fréquences correspondant à la quinte, nous disons que l'instrument est "accordé de quinte en quinte".

Remarque: L'octave, la quinte, la quarte et le ton sont appelés en musique des "intervalles".

Évidemment, si nous souhaitons faire de la musique, il faut choisir un repère. Comme la gamme de fréquences est infinie, il a été choisi une fréquence à laquelle l'oreille humaine actuelle est la plus sensible: 440 [Hz] que l'on appelle le "**la**" et qui est noté (A) chez les Anglo-saxons.

Au solfège, les étudiants occidentaux apprennent la suite empirique suivante de 7 notes et dont l'origine écrite est purement *mnémotechnique* (on retrouve cette suite sur les accordeurs électroniques avec le système Anglo-saxon qui utilise certaines lettres de l'alphabet):

do (C) - ré (D) - mi (E) - fa (F) - sol (G) - la (A) - si (B) - do (C)



où les lignes de la symbolique solfégique ci-dessus représentent une quinte dans la gamme pythagoricienne. Entre le premier *do* et le deuxième *do*, nous avons donc un intervalle d'une octave. Par ailleurs, ce nombre de notes à l'intérieur de l'intervalle d'octave est souvent insuffisant pour faire de la bonne musique et des artistes ont ajouté d'autres notes par altération des fréquences de ces 7 notes, en les diésant (multiplication de la fréquence par $25/24$) ou en les bémolisant (multiplication par $1-25/24$).

Signalons qu'il existe une dizaine de manières de découper l'octave solfégique et qui sont presque toutes encore utilisée au début du 21ème siècle (elles sont toutes complètement empiriques donc nous n'entrerons pas dans les détails):

- La "gamme naturelle" (appelée aussi "gamme harmonique")
- La "gamme tempérée"
- La "gamme pythagoricienne"
- La "gamme chromatique"
- La "gamme de Zarlino" (appelée aussi "gamme des physiciens")

etc.

La méthode et l'ordre d'apprentissage susmentionné est une aberration mathématique (mais on est plus à ça près en musique... qui est un art empirique).

6. OSCILLATEUR HARMONIQUE

Considérons le système mécanique de la figure ci-dessous constitué d'un ressort de constante k , d'une masse m et d'un amortisseur visqueux de constante c . Ce système est souvent assimilé à la mécanique d'un haut-parleur (ou d'un micro) où l'air sert d'amortisseur visqueux et le champ magnétique de l'aimant excitateur de force d'application ou encore d'un amortisseur de voiture:

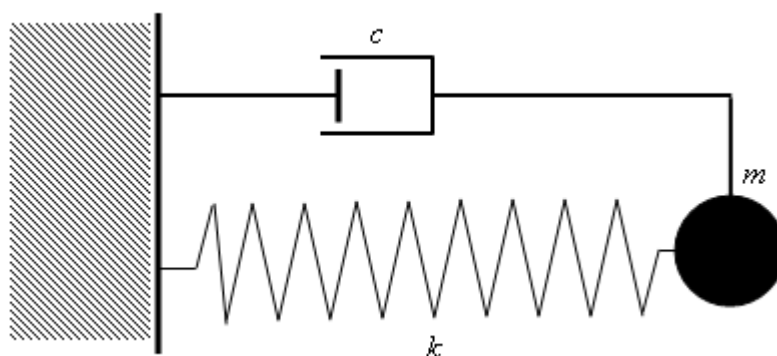


Figure: 68.4 - Schéma de principe de l'oscillateur harmonique

Nous avons démontré dans le chapitre de Génie Mécanique que pour un ressort de type hélicoïdal nous avions:

$$F = \left(\frac{G \cdot d^4}{8 \cdot n \cdot D^3} \right) x = k \cdot x \quad (68.107)$$

et c'est l'expression de la force avec laquelle le ressort s'opposera à une force extérieure de compression ou de traction. À l'équilibre, nous avons forcément la somme des forces qui est nulle:

$$\vec{F}_{ext} + \vec{F}_{ressort} = 0 \quad (68.108)$$

soit:

$$m\ddot{x} + k \cdot x = 0 \quad (68.109)$$

Nous avons démontré dans le chapitre de Mécanique Des Milieux Continus que la loi de Stokes dans le cadre d'un cylindre était donnée par:

$$F = 6\pi\mu Rv = c \cdot v \quad (68.110)$$

Nous supposons que notre amortisseur visqueux est basé sur une loi du même type où la force de résistance est proportionnelle à la vitesse. À l'équilibre, nous avons forcément la somme des forces qui est nulle:

$$\vec{F}_{ext} + \vec{F}_{ressort} + \vec{F}_{amortisseurs} = 0 \quad (68.111)$$

soit:

$$m\ddot{x} + c \cdot \dot{x} + k \cdot x = 0 \quad (68.112)$$

Nous allons simplifier cette expression:

$$\frac{m}{k} \ddot{x} + \frac{c}{k} \cdot \dot{x} + x = 0 \quad (68.113)$$

Nous nous retrouvons donc avec une équation différentielle exactement identique à celle d'un circuit RLC série en régime constant (cf. chapitre de Génie Électrique):

$$LC \cdot \ddot{q} + RC \cdot \dot{q} + q = 0 \quad (68.114)$$

Nous avons donc de la même manière, la viscosité critique:

$$c_c = 2\sqrt{m \cdot k} \quad (68.115)$$

au lieu de la résistance critique du circuit RLC:

$$R_c = 2\sqrt{\frac{L}{C}} \quad (68.116)$$

qui va déterminer si le système est en régime critique:

$$C = C_c \quad (68.117)$$

en régime apériodique (hypercritique):

$$C > C_c \quad (68.118)$$

ou pseudo-périodique (oscillations amorties) avec:

$$C < C_c \quad (68.119)$$

Dans le cas des voitures ou haut-parleurs, il est souvent intéressant de s'arranger pour être en mode pseudo-périodique. Nous retrouvons alors exactement les mêmes résultats que le circuit RLC avec oscillations amorties puisque l'équation différentielle est la même.

Les spécialistes de l'acoustique ou de la mécanique s'arrangent alors pour s'approprier le concept de "facteur d'amortissement" du circuit RLC:

$$\alpha = \frac{R}{2L} \quad (68.120)$$

qui s'écrira dans le cas présent par identification des termes:

$$\lambda = \frac{c}{2m} \quad (68.121)$$

Avec la pulsation propre:

$$\omega = \sqrt{\frac{1}{LC}} \quad (68.122)$$

qui s'écrit dans le cas présent:

$$\omega = \sqrt{\frac{k}{m}} \quad (68.123)$$

L'oscillateur amorti perd de l'énergie au cours du temps. Calculons cette perte. Nous avons (cf. chapitres de Mécanique Classique et Génie Mécanique):

$$E = E_c + E_p = \frac{1}{2} m \dot{x}^2 + \frac{1}{2} k x^2 \quad (68.124)$$

d'où:

$$\frac{dE}{dt} = m \dot{x} \ddot{x} + k x \dot{x} = (m \ddot{x} + k x) \dot{x} \quad (68.125)$$

Or, selon l'équation du mouvement:

$$m \ddot{x} + c \cdot \dot{x} + k \cdot x = 0 \Rightarrow m \ddot{x} + k \cdot x = -c \cdot \dot{x} \quad (68.126)$$

Il vient alors:

$$\frac{dE}{dt} = -c \cdot \dot{x}^2 \quad (68.127)$$

La perte d'énergie (ou puissance dissipée) est proportionnelle au carré de la vitesse instantanée.

Nous avons démontré dans le chapitre de Génie Électrique pour le circuit RLC qu'en régime d'oscillations amorties, nous avons:

$$\begin{aligned} q &= \frac{q_0}{\cos \phi} e^{-\frac{R}{2L}t} \cos(\omega t - \phi) \\ i &= -\frac{\omega q_0}{\cos^2 \phi} e^{-\frac{R}{2L}t} \sin(\omega t) \end{aligned} \quad (68.128)$$

Ce qui devient ici:

$$\begin{aligned} x &= \frac{x_0}{\cos \phi} e^{-\frac{R}{2L}t} \cos(\omega t - \phi) \\ \dot{x} &= -\frac{\omega x_0}{\cos^2 \phi} e^{-\frac{R}{2L}t} \sin(\omega t) \end{aligned} \quad (68.129)$$

Dès lors:

$$\frac{dE}{dt} = -c \cdot \dot{x}^2 = c \left(\frac{\omega x_0}{\cos^2 \phi} \sin(\omega t) \right)^2 e^{-\frac{R}{L}t} = c x^2 \sin^2(\omega t) e^{-\frac{R}{L}t} \quad (68.130)$$

6.1. L'OSCILLATEUR FORCÉ

Considérons toujours le même l'oscillateur mais excité cette fois-ci par une force harmonique (donc en régime forcé) tel que:

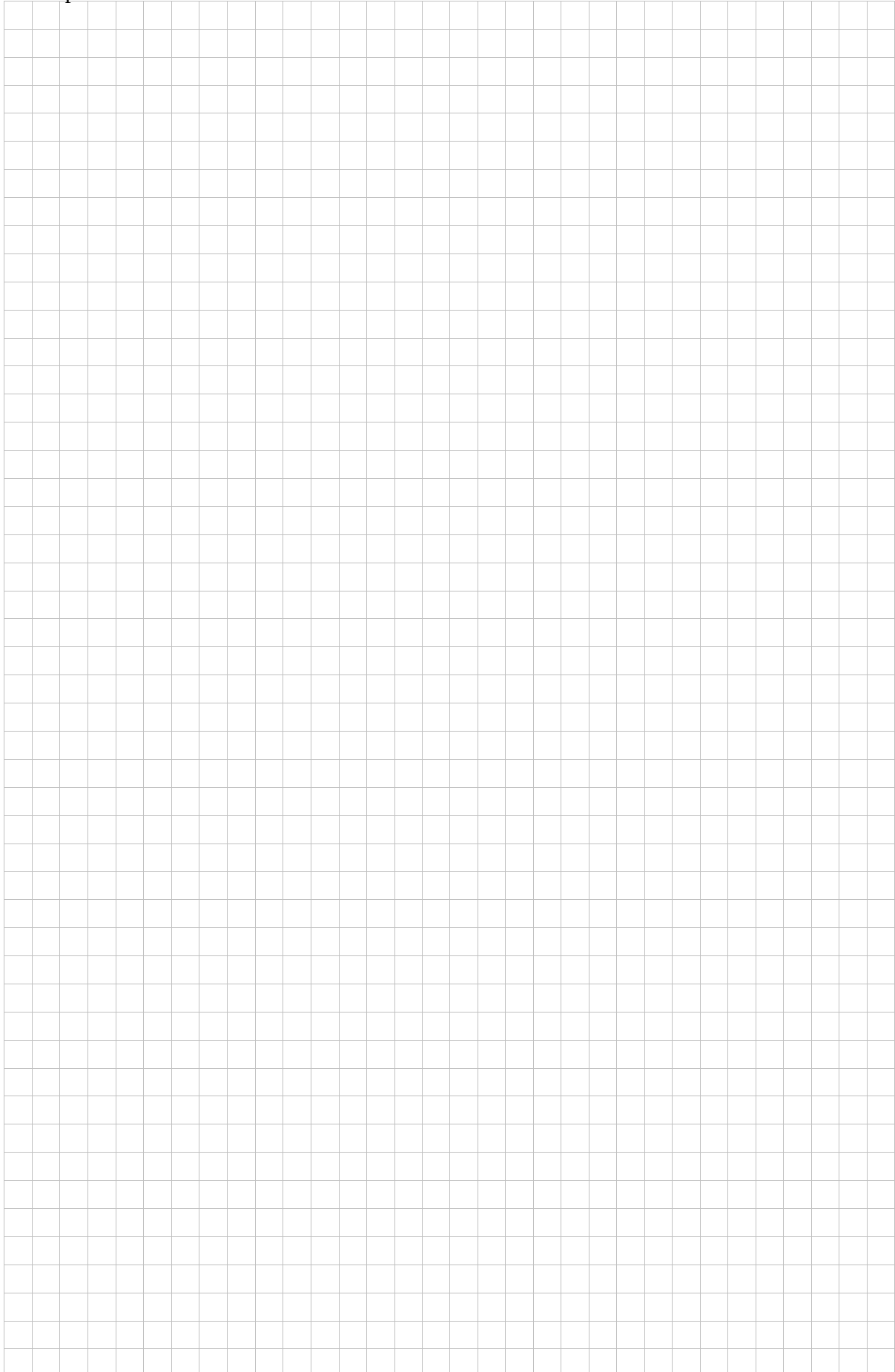
$$m\ddot{x} + c \cdot \dot{x} + k \cdot x = F_0 \sin(\omega t) \quad (68.131)$$

Encore une fois, il s'agit exactement du même type d'équation différentielle que pour l'oscillateur RLC en régime forcé étudié dans le chapitre de Génie Électrique pour lequel nous avons:

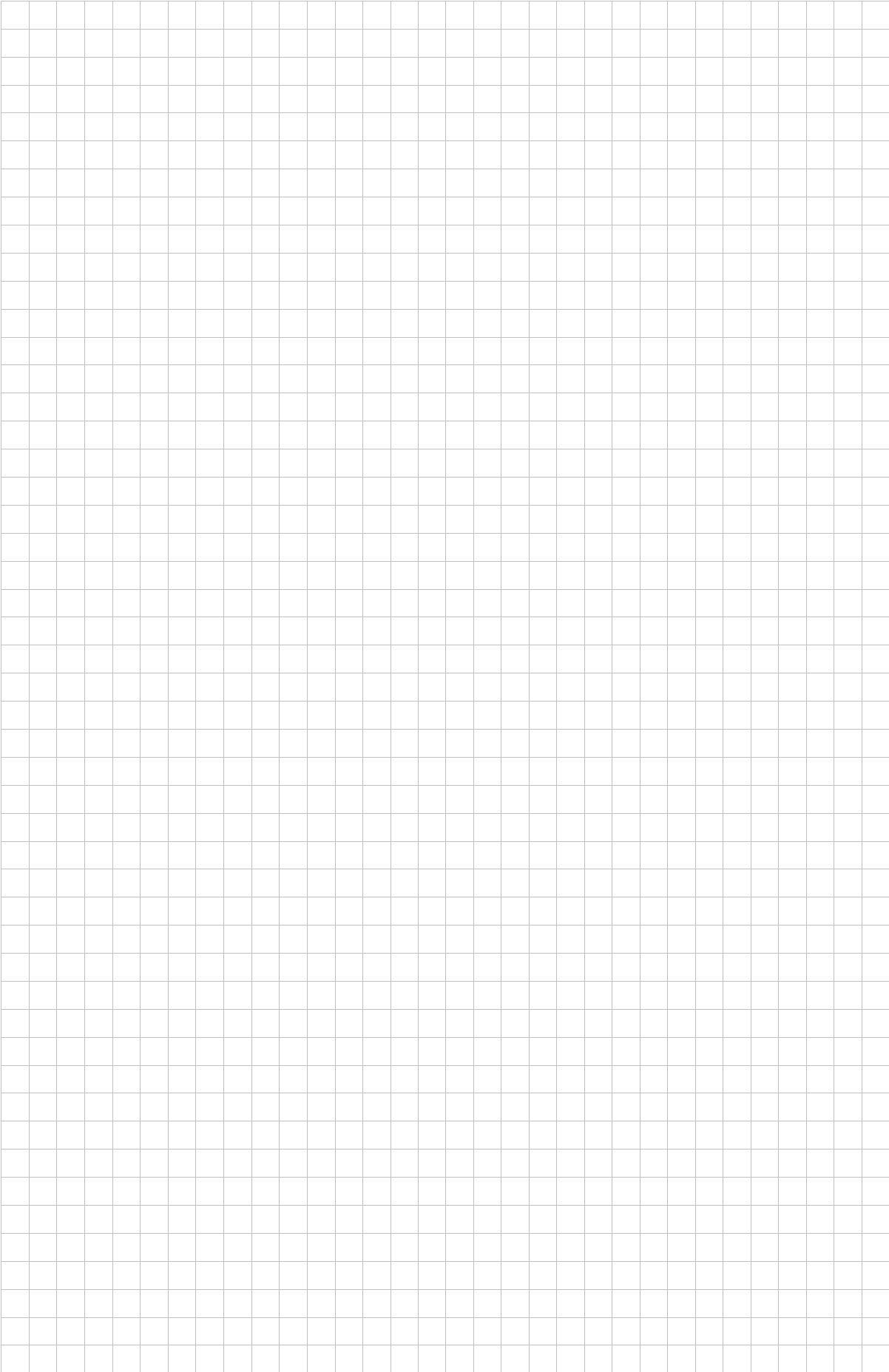
$$L \frac{d^2 q}{dt^2} + R \frac{dq}{dt} + \frac{1}{C} q = U_0 \sin(\omega t) \quad (68.132)$$

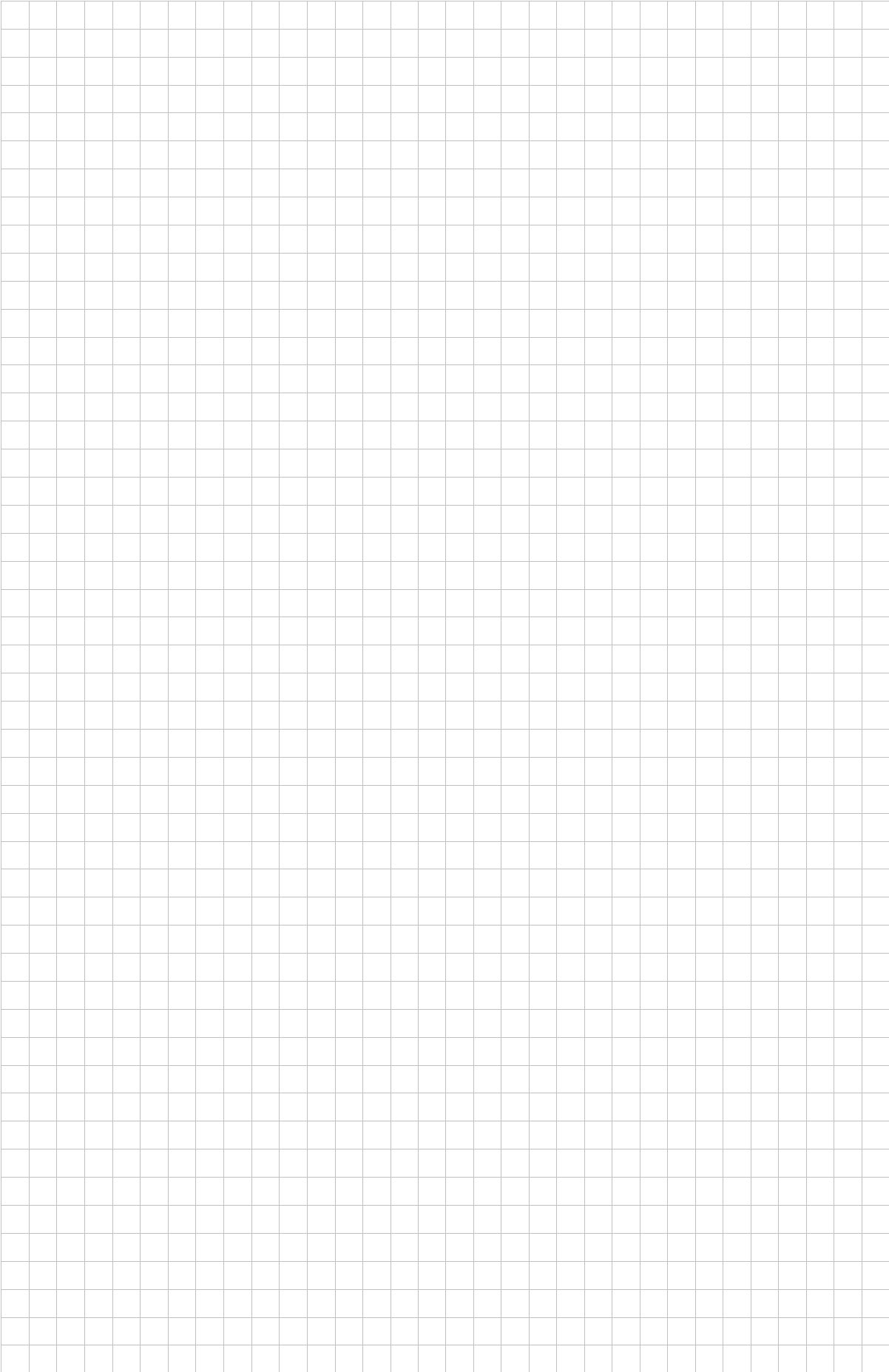
Il s'ensuit les mêmes résultats et conclusions que le lecteur pourra trouver dans le chapitre de Génie Électrique.

Notes personnelles:









XII

Ingénierie

69. GÉNIE MARIN & MÉTÉO

La météorologie est l'étude des phénomènes atmosphériques tels les nuages, les dépressions et les précipitations pour comprendre comment ils se forment et évoluent. C'est une discipline qui traite principalement de la mécanique des fluides appliquée à l'air mais qui fait usage de différentes autres branches de la physique et de la chimie. Elle permet donc d'établir des prévisions météorologiques en s'appuyant sur des modèles mathématiques à court comme à relativement long terme. Elle est également appliquée pour la prévision de la qualité de l'air, pour les changements climatiques et pour l'étude dans plusieurs domaines de l'activité humaine (construction, trafic aérien, etc.)

La météorologie est liée à une grande quantité de variables dont il serait très difficile de faire une liste même non exhaustive... Cependant sur notre planète Terre, un facteur important à ne pas négliger est celui constitué par les surfaces océaniques et leur dynamique intrinsèque dont nous tenterons de présenter au travers d'une étude mathématique sommaire quelques-unes de leurs propriétés.

Ce chapitre est une introduction générale de base au domaine des applications techniques de la thermodynamique et de la mécanique des fluides. Il permettra au lecteur de se familiariser avec le langage et certaines méthodes de calculs fondamentales utilisées par les ingénieurs de cette branche. Bien sûr, il faut compléter cette étude par des travaux pratiques en laboratoire.

1. HORIZON VISUEL

Étudions d'abord un petit sujet sympathique faisant souvent débat lors des vacances ou plus sérieusement... Dans certains logiciels de météorologie, il est demandé de saisir la distance de l'horizon visuel lors de mesures de température et pression... or celle-ci est difficile à déterminer par très beau temps lorsque nous sommes en hauteur.

Pour cela, considérons la Terre de rayon R et un point de perspective de hauteur h par rapport au niveau de la mer que nous noterons A . La question est de savoir à quelle distance se trouve le point C donné par définition par la tangente AC qui est simplement la ligne d'horizon.

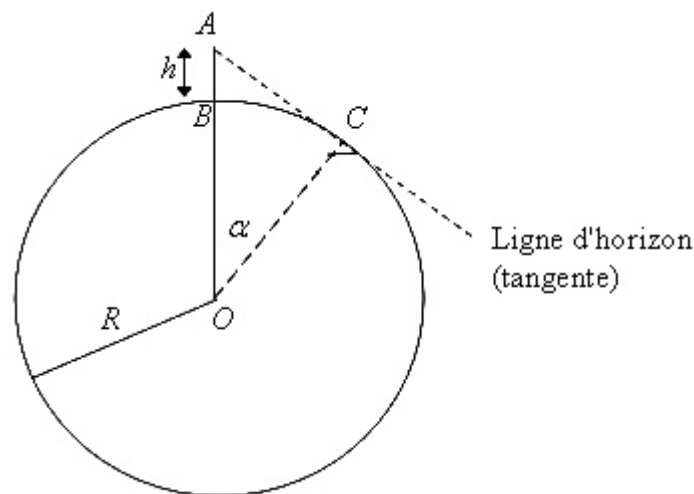


Figure: 69.1 - Configuration de l'expérience...

Le lecteur observera déjà que l'étude va principalement faire appel à de la trigonométrie et de la géométrie élémentaire.

L'angle $\widehat{OCA}$ est un angle droit. En effet, une droite tangente en un point d'un cercle est perpendiculaire au rayon en ce point. Le triangle OCA est donc rectangle en C .

Nous avons donc:

$$\cos(\alpha) = \frac{OC}{OA} = \frac{OC}{OB + BA} \quad (69.1)$$

Or, nous avons $OB = OC = R$. D'où nous déduisons:

$$\cos(\alpha) = \frac{R}{R+h} \Rightarrow \alpha = \arccos\left(\frac{R}{R+h}\right) \quad (69.2)$$

La distance AC est la distance à vol d'oiseau entre le point de vue (belvédère) et le bateau que nous observons sur l'horizon. La distance qui nous intéresse cependant ici est BC : c'est la distance que nous devrions parcourir à l'altitude 0 pour rejoindre l'autre bateau.

Dans la suite, nous poserons $BC = d$.

Lorsque l'angle α varie de 0° à 360° (tour complet), nous décrivons toute la circonférence de la Terre, c'est-à-dire $2\pi R$ puisque la Terre est supposée être ronde.

Utilisation de la règle de trois:

Si un angle de 360° correspond à une distance de longueur $2\pi R$ alors un angle de α correspond à une distance:

$$d = \frac{2\pi R}{360} \alpha = \frac{\pi R}{180} \alpha \quad (69.3)$$

Or, nous avons vu précédemment que:

$$\alpha = \arccos\left(\frac{R}{R+h}\right) \quad (69.4)$$

D'où, finalement:

$$d = \frac{\pi R}{180} \alpha = \frac{\pi R}{180} \arccos\left(\frac{R}{R+h}\right) \quad (69.5)$$

Avec $R = 6'378 \text{ [km]}$, nous trouvons (h doit être exprimé en kilomètres):

$$d \cong 111.32 \arccos\left(\frac{6'378}{6'378+h}\right) \quad (69.6)$$

Nous avons alors dans le vide, dans un paysage sans obstacles... la table suivante:

Altitude h [m]	Distance de l'horizon d [km]
5	8
10	11.3
50	25.3
100	35.7
200	50.5
400	71.4
600	87.5
800	101
1000	113
2000	159.7
3000	195.6
4000	225.8
5000	252.5
10000	357

Tableau: 69.1 - Horizon visuel en fonction de l'altitude

Remarque: Si nous ne tenons donc pas compte de la réfraction atmosphérique, nous constatons qu'il faudrait une altitude de l'ordre de plusieurs kilomètres pour voir au-delà de 200 [km] de distance. Pourtant, sans aller très loin, depuis les hauteurs de Nice (Alpes-Maritimes), il est possible d'observer la pointe du Cap Corse qui se trouve à environ 220 km du continent!!! La réfraction atmosphérique joue donc un rôle dans ce phénomène.

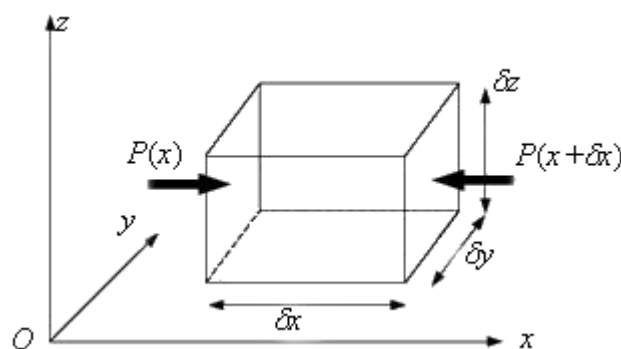
2. DIRECTION DES VENTS

Nous allons démontrer mathématiquement maintenant quelque chose de tout à fait intuitif: que les vents se déplacent des hautes vers les basses pressions (c'est bête comme ça mais il faut tout de même le montrer).

Nous savons (cf. chapitre de Mécanique Des Milieux Continus) que la force de pression s'exerçant sur une surface S est normale à cette surface et vaut sous forme scalaire $P \cdot S$.

Pour une parcelle d'air de volume $dV = \delta x \delta y \delta z$ la force de pression totale selon la direction x vaut alors:

$$F_p = P(x) \cdot \delta y \delta z - P(x + \delta x) \delta y \delta z \quad (69.7)$$



De plus, nous avons (cf. chapitre de Calcul Différentiel Et Intégral):

$$\frac{P(x + \delta x) - P(x)}{\delta x} = \frac{\partial P}{\partial x} \quad (69.8)$$

Donc:

$$F_p = P(x) \cdot \delta y \delta z - P(x + \delta x) \delta y \delta z = -\frac{\partial P}{\partial x} \delta x \delta y \delta z \quad (69.9)$$

La force de pression massique est donc:

$$f_p = \frac{F_p}{m} = \frac{F_p}{\rho \delta x \delta y \delta z} = -\frac{1}{\rho} \frac{\partial P}{\partial x} \quad (69.10)$$

Nous pouvons faire le même calcul selon y. Finalement, la force de pression horizontale massique sera donnée par:

$$\vec{f}_p = -\frac{1}{\rho} \begin{pmatrix} \frac{\partial P}{\partial x} \\ \frac{\partial P}{\partial y} \end{pmatrix} = -\frac{1}{\rho} \vec{\nabla} P \quad (69.11)$$

Ainsi, la force de pression (massique ou non) est opposée au gradient horizontal. La transmission de l'information (de la force) se fait à la vitesse du son pour cette équation (ce qui explique la vitesse des appels d'air dans votre maison ou appartement et la force pouvant faire claquer les portes ou fenêtre).

Elle est donc:

- Dirigée des hautes vers les basses pressions, perpendiculaire aux isobares
- Inversement proportionnelle à l'écartement des isobares.

Si nous relevons les valeurs de la pression atmosphérique en différents points du globe et que nous relient entre eux les points de pression identique, nous obtenons une série de courbes, appelées "**isobares**". Le vent est directement déterminé par ce relief atmosphérique, puisque c'est un déplacement d'air entre des hautes vers les basses pressions.

La vitesse du vent est donc fixée par le gradient de pression: autrement dit, si la pression atmosphérique varie rapidement avec la distance, le vent soufflera fort, tandis qu'il sera faible dans un "marais" barométrique où cette pression reste quasiment inchangée sur de grandes distances. En résumé, plus les isobares sont rapprochées, plus le vent soufflera fort.

Les isobares sont traditionnellement indiquées par un pas de 5 millibars sur les cartes météo tel que le montre l'exemple ci-dessous:

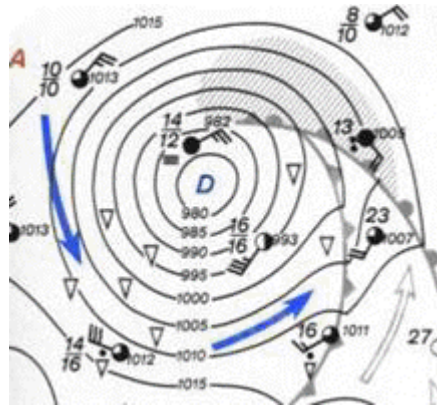


Figure: 69.2 - Représentation typique des isobares (sur une dépression)

Ensuite, les météorologues ont défini empiriquement (c'est sympathique pour la culture générale) une unité de mesure des vents qui n'est qu'une correspondance entre la force du vent et la distance séparant 2 isobares (5 en 5 [mb]):

Distance entre isobares [km]	Unité [Beaufort]	Vitesse [m/s]
600 (brise légère)	2	1.6-3.3
500 (brise moyenne)	4	3.4-5.4
400 (brise fraîche)	5	8-10.7
300 (vent fort)	6	10.8-13.8
200 (grand vent)	7	13.9-17.1
100 (tempête)	9	20.8-24.4

Tableau: 69.2 - Distance entre isobares et vents

3. MODÈLE ATMOSPHÉRIQUE EXPONENTIEL

Considérons que l'atmosphère est un fluide parfait dans un champ de gravité. Alors à partir de la relation du théorème de Bernoulli suivante, démontrée dans le chapitre de Mécanique Des Milieux Continus (fluide statique):

$$P = -\rho gz + c^{te} \quad (69.12)$$

Il vient alors:

$$\frac{\partial P}{\partial z} = -\rho g \quad (69.13)$$

Ainsi, pour connaître la variation de pression avec l'altitude dans l'atmosphère ou la profondeur dans l'océan, nous avons pris comme hypothèse "**l'équilibre hydrostatique**", soit que la variation de pression avec la hauteur/profondeur est proportionnelle à la gravité et à la densité du fluide.

Ceci n'est bien évidemment pas valide dans le cas de mouvements rapides de convection, comme dans les orages, mais se vérifie assez bien dans les mouvements plus lents et à grande échelle: l'échelle synoptique.

Nous allons alors combiner cette dernière relation avec une équation d'état, par exemple celle du gaz parfait à la température T et de densité ρ dont les particules constituantes ont pour masse m . Nous avons donc l'équation des gaz parfaits (cf. chapitre de Mécanique Des Milieux Continus):

$$P \cdot V = nRT \Rightarrow P = \frac{nRT}{V} = \frac{\rho kT}{m} \Rightarrow \rho = \frac{mP}{kT} \quad (69.14)$$

avec pour rappel P est la pression exprimée en Pascals, V le volume exprimé en mètres cubes, R qui est la constante des gaz parfaits, T la température exprimée en Kelvins, n le nombre de moles, k est la constante de Boltzmann, ρ la densité de particules, m la masse totale des particules.

Dans le cas isotherme (par exemple dans la stratosphère Terrestre, au-dessus de 10 [km] d'altitude où la température est quasi constante autour de -55 degrés Celsius), l'intégration s'effectue facilement:

$$\frac{\partial P}{\partial z} = -\rho g \Rightarrow \frac{\partial P}{\partial z} = -\frac{m \cdot P}{k \cdot T} g \quad (69.15)$$

Donc, à une pression donnée, le gradient vertical de pression est inversement proportionnel à la température.

Considérons maintenant la relation suivante:

$$\frac{dP}{P} = -\frac{m \cdot g}{k \cdot T} dz \Rightarrow \ln(P) = -\frac{m \cdot g}{k \cdot T} z + c^{te} \quad (69.16)$$

En utilisant l'exponentielle:

$$P = c^{te} \exp\left(-\frac{mg}{kT} z\right) = P_0 e^{-\frac{mg}{kT} z} \quad (69.17)$$

La pression décroît donc exponentiellement avec l'altitude. P_0 étant la pression au niveau du sol.

Revenons aussi à la relation:

$$\frac{\partial P}{\partial z} = -\frac{m \cdot P}{k \cdot T} g \quad (69.18)$$

Elle peut bien évidemment aussi s'écrire sous la forme:

$$\frac{\partial z}{\partial P} = -\frac{k}{m \cdot P \cdot g} T \quad (69.19)$$

qui nous dit que la distance z entre les surfaces isobares est directement proportionnelle à la température.

Voyons également une autre approche courante. Repartons pour cela de la relation démontrée plus haut mais pour une masse m de 1 kilo:

$$\frac{dP}{dz} = -\frac{g}{kT} P \quad (69.20)$$

et notons cette relation sous la forme suivante:

$$\frac{d_z P(z)}{P(z)} = -\frac{g}{kT(z)} \quad (69.21)$$

Rappelons que (cf. chapitre de Calcul Différentiel Et Intégral):

$$d_x \ln(f(x)) = \frac{1}{f(x)} d_x f(x) \quad (69.22)$$

Donc:

$$d_z \ln(P(z)) = - \frac{g}{kT(z)} \quad (69.23)$$

et supposons maintenant que la variation de température est linéaire dans l'atmosphère (ce qui n'est pas loin de la vérité pour les 10 à 20 premiers kilomètres de l'atmosphère) tel que:

$$T(z) = T_0 + \lambda \cdot z \quad (69.24)$$

avec $\lambda < 0$ qui est le gradient de température en $[\text{°K/m}]$.

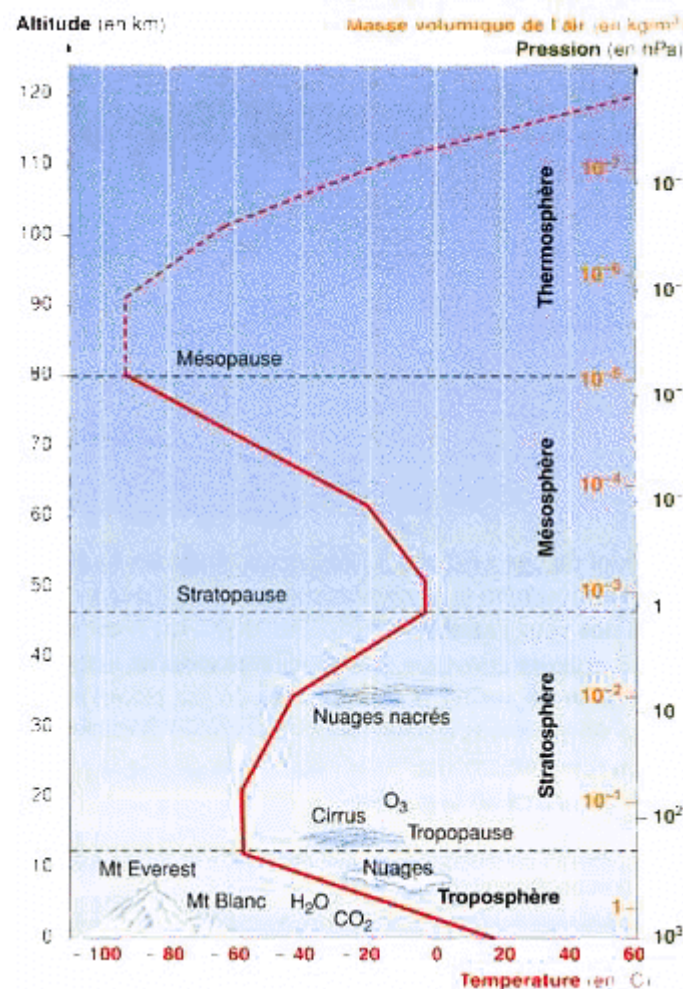


Figure: 69.3 - Profil type de la température et de la pression sur Terre (fin 20ème siècle)

Nous avons alors:

$$d_z \ln(P(z)) = - \frac{g}{k(T_0 + \lambda \cdot z)} = \frac{-g/k}{T_0 + \lambda \cdot z} \quad (69.25)$$

Soit:

$$\int_{z_0}^{z_1} d \ln(P(z)) = \int_{z_0}^{z_1} \frac{-g/k}{T_0 + \lambda \cdot z} dz \quad (69.26)$$

Ce qui donne:

$$\ln(P(z_1)) - \ln(P(z_0)) = -\frac{g}{k \cdot \lambda} \ln(T_0 + \lambda \cdot z) \Big|_{z_0}^{z_1} \quad (69.27)$$

Après simplification:

$$\ln\left(\frac{P(z_1)}{P(z_0)}\right) = -\frac{g}{k \cdot \lambda} \ln\left(\frac{T_0 + \lambda \cdot z_1}{T_0 + \lambda \cdot z_0}\right) \quad (69.28)$$

Soit:

$$P(z_1) = P(z_0) \cdot \left(\frac{T_0 + \lambda \cdot z_1}{T_0 + \lambda \cdot z_0}\right)^{-\frac{g}{k \cdot \lambda}} \quad (69.29)$$

Soit écrit de manière plus esthétique:

$$P(z) = P_0 \cdot \left(\frac{T_0 + \lambda \cdot z}{T_0 + \lambda \cdot z_0}\right)^{-\frac{g}{k \cdot \lambda}} \quad (69.30)$$

Un bon exemple d'application courant de cette relation est les planeurs et les deltaplanes. Ceux-ci attendent de la météo que celle-ci leur communique l'altitude de l'isotherme du zéro degré lors de ses bulletins. Ils en déduisent alors le gradient de température par mètre. Pour ces sportifs, une bonne condition est d'avoir un gradient de 1 [°C] pour 100 mètres. Il est dès lors aisé avec la relation précédente de calculer la pression à une altitude de 2'000 mètres et d'en déduire le gradient de pression qui leur permet d'utiliser certains ascendants pour leurs exercices de voltige.

4. MODÈLE ATMOSPHÉRIQUE ADIABATIQUE

Le gradient thermique adiabatique est, dans l'atmosphère terrestre, la variation de température de l'air avec l'altitude (autrement dit le gradient de la température de l'air), qui ne dépend que de la pression atmosphérique, c'est-à-dire:

- Sans considération d'échange de chaleur avec l'environnement (autres masses d'air, relief)
- Sans considération de condensation (formation de nuages) ni de précipitation.

Ce concept a une grande importance en météorologie, ainsi qu'en navigation aérienne et maritime.

Nous avons démontré dans le chapitre de Thermodynamique la loi de Laplace (satisfaite que sous certaines conditions!):

$$PV^\gamma = c^{te} \quad (69.31)$$

avec le coefficient de Laplace:

$$\gamma = \frac{C_P}{C_V} \quad (69.32)$$

Soit sous forme massique:

$$\frac{P}{\rho^\gamma} = c^{te} \quad (69.33)$$

Nous pouvons en prendre le logarithme:

$$\ln\left(\frac{P}{\rho^\gamma}\right) = \ln(c^{te}) \Rightarrow \ln(P) - \gamma \ln(\rho) = c^{te} \Rightarrow \ln(P) = \gamma \ln(\rho) + c^{te} \quad (69.34)$$

Or en prenant la différentielle logarithmique:

$$\begin{aligned} \frac{d(\ln(P))}{dP} &= \frac{1}{P} \Rightarrow d(\ln(P)) = \frac{dP}{P} \\ \frac{d(\gamma \ln(\rho))}{d\rho} &= \gamma \frac{1}{\rho} \Rightarrow d(\gamma \ln(\rho)) = \gamma \frac{d\rho}{\rho} \quad (69.35) \\ d(c^{te}) &= 0 \end{aligned}$$

Nous avons alors (relation que nous retrouverons dans le chapitre de Musique Mathématiques):

$$\frac{dP}{P} = \gamma \frac{d\rho}{\rho} \quad (69.36)$$

Nous aurions aussi pu trouver ce résultat directement à partir de la relation démontrée dans le chapitre de Thermodynamique:

$$\gamma \frac{dV}{V} + \frac{dP}{P} = 0 \Rightarrow \frac{dP}{P} = -\gamma \frac{dV}{V} = -\gamma \frac{d\left(\frac{m}{\rho}\right)}{\frac{m}{\rho}} = -\rho \gamma d\left(\frac{1}{\rho}\right) = \gamma \frac{d\rho}{\rho} \quad (69.37)$$

En prenant aussi la différentielle logarithmique de loi des gaz parfaits où n est le nombre de moles:

$$PV = nRT \quad (69.38)$$

Mais sous la forme massique pour une mole:

$$\rho = \frac{M_m P}{RT} \quad (69.39)$$

où M_m est donc la masse molaire, nous avons:

$$\ln(\rho) = \ln\left(\frac{M_m P}{RT}\right) = \ln(M_m P) - \ln(RT) \quad (69.40)$$

Soit:

$$\frac{d\rho}{\rho} = \frac{dP}{P} - \frac{dT}{T} \quad (69.41)$$

Nous obtenons alors:

$$\frac{1}{\gamma} \frac{dP}{P} = \frac{dP}{P} - \frac{dT}{T} \Rightarrow \frac{dT}{T} = \frac{dP}{P} - \frac{1}{\gamma} \frac{dP}{P} \Rightarrow \frac{dT}{T} = \frac{\gamma - 1}{\gamma} \frac{dP}{P} \quad (69.42)$$

Soit:

$$\frac{dP}{P} = \frac{\gamma}{\gamma - 1} \frac{dT}{T} \quad (69.43)$$

Utilisons la relation démontrée plus haut:

$$\frac{dP}{dz} = -\frac{mg}{kT} P \Rightarrow \frac{dP}{P} = -\frac{mg}{kT} dz \quad (69.44)$$

Il vient alors:

$$dT = \frac{1 - \gamma}{\gamma} \frac{mg}{k} dz \quad (69.45)$$

Nous avons donc une atmosphère à gradient thermique constant et négatif (la température diminue de manière linéaire avec l'altitude):

$$\kappa = \frac{1 - \gamma}{\gamma} \frac{mg}{k} = \frac{1 - \gamma}{\gamma} \frac{M_m g}{R} \quad (69.46)$$

La dernière forme utilisant la masse molaire étant plus pratique car elle permet de caractériser le milieu étudié. À remarquer que c'est un modèle qui semble très très bien marcher pour une altitude de 0 à 90 [km] sur la planète Vénus mais nettement moins bien pour la planète Terre.

Nous avons alors par exemple pour la Terre (donc tout en sachant que le modèle n'est pas très bien adapté) $g = 9.81 \left[m \cdot s^{-2} \right]$, $R = 8.314 \left[JK^{-1} mol^{-1} \right]$ et le coefficient adiabatique pour l'air $\gamma = 1.4$, et sa masse molaire $M_m = 28.96 \cdot 10^{-3} \left[Kg \cdot mol^{-1} \right]$.

Soit:

$$\kappa = -.00976 \left[K \cdot m^{-1} \right] \cong -1 \left[K \cdot hm^{-1} \right] \quad (69.47)$$

ce qui correspond à l'idée courante (un degré pour 100 mètres).

5. ÉQUATION HYPSONOMÉTRIQUE

Nous avons donc pour l'équilibre hydrostatique:

$$\frac{\partial P}{\partial z} = -\frac{m \cdot P}{k \cdot T} g \quad (69.48)$$

où n est le nombre de moles par kilogramme.

Nous pouvons intégrer cette relation si nous connaissons T en fonction de P ou z . La mesure directe de P dans la pratique est plus facile (les altimètres simples sont en fait des baromètres).

Nous pouvons alors séparer les variables:

$$k T \frac{dP}{P} = -mgdz \quad (69.49)$$

En intégrant entre deux niveaux a et b :

$$k \int_a^b T d(\ln(P)) = - \int_a^b mgdz \quad (69.50)$$

Puisque:

$$\frac{d(\ln(P))}{dP} = \frac{1}{P} \Rightarrow d(\ln(P)) = \frac{dP}{P} \quad (69.51)$$

Ensuite pour continuer nous allons utiliser une astuce. Nous allons définir la température moyenne par la relation:

$$\langle T \rangle = \frac{\int_a^b T d(\ln(P))}{\int_a^b d(\ln(P))} \quad (69.52)$$

Ce qui nous permet alors d'écrire:

$$k \int_a^b T d(\ln(P)) = k \int_a^b T d(\ln(P)) \frac{\int_a^b d(\ln(P))}{\int_a^b d(\ln(P))} = k \langle T \rangle \int_a^b d(\ln(P)) = - \int_a^b mgdz \quad (69.53)$$

Soit:

$$mg(z_a - z_b) = k \langle T \rangle \ln \left(\frac{P_b}{P_a} \right) \quad (69.54)$$

où autrement écrit:

$$z_a - z_b = \frac{k}{mg} \langle T \rangle \ln \left(\frac{P_b}{P_a} \right) \quad (69.55)$$

Les deux relations encadrées sont appelées chacune respectivement "**équation hypsométrique**" (du grec "hypso" pour "hauteur"):

$$\frac{P_b}{P_a} = e^{-\frac{mg(z_a - z_b)}{k \langle T \rangle}} \quad (69.56)$$

Remarque: En météorologie, z_b est posé comme étant égal à 0 au niveau de la mer où la pression P_b est supposée connue. Ainsi, nous avons trois paramètres libres. En en connaissant deux sur les trois, il est facile de déterminer le troisième.

À l'armée, par exemple, les ballons-sondes donnent la température et la hauteur du ballon et les militaires au sol mesurent z_b et P_b . Ensuite toutes ces informations sont communiquées aux chars d'assaut qui peuvent calculer la pression P_a à différentes hauteurs et donc l'influence de celle-ci sur la trajectoire de leurs obus... via la différence de force.

6. BALLON-SONDE

Un joli petit exemple intéressant d'application des mathématiques appliquées au génie météo est l'étude des fameux ballons-sondes et particulièrement la caractéristique de leur volume en fonction de l'altitude qui est souvent sujet à débat dans des groupes de discussion lorsque personne n'y formalise le problème une bonne fois pour toute. Vous aurez donc compris que c'est ce que nous allons étudier ici et surtout nous allons tenter de déterminer le diamètre théorique de ceux-ci à une altitude donnée.

L'énoncé du problème souvent débattu est le suivant:

Un ballon-sonde en PVC (Polyvinyl-Chloride) de masse m sert à emmener à haute altitude un appareillage en vue d'effectuer des mesures. L'enveloppe du ballon contient n moles de gaz parfait d'hydrogène H_2 ayant donc une masse molaire de:

$$M_m(H_2) = 2 [g \cdot mol^{-1}] \quad (69.57)$$

L'atmosphère sera assimilée à un gaz parfait, de masse molaire moyenne:

$$M_m(air) \cong 29 [g \cdot mol^{-1}] \quad (69.58)$$

aux C.N.T.P. (Conditions Normales de Température et de Pression).

Nous voulons d'abord chercher quelle est la force ascensionnelle F_x ressentie par le ballon?

Ensuite, nous voulons évaluer la quantité de matière minimale n_0 assurant le décollage de celui-ci pour une masse totale (ballon compris!) de 2.6 [Kg]. puis le volume V_0 correspondant, à l'altitude de départ.

Rappelons deux choses pour résoudre déjà ce premier point:

1. Tout corps plongé dans un liquide (ou un gaz) subit une force vers le haut égale au poids du volume qu'il déplace (force d'Archimède) selon la relation démontrée dans le chapitre de Mécanique des milieux continus:

$$F_A = \rho \cdot g \cdot V = F_z \quad (69.59)$$

2. Tout gaz parfait ayant une masse en grammes égale à la masse molaire occupe selon la loi des gaz parfaits un volume de 22.4 [L] à 273.15 [K] et à une pression de 1 [atm] comme nous l'avons démontré dans le chapitre de Chimie Thermique. Ce qui donne aux C.N.T.P.:

$$V = nR \frac{T}{P} = 1 \cdot 8.314 \cdot \frac{273.15}{1.01325 \cdot 10^5} \cong 22 \text{ [L]} \quad (69.60)$$

Donc pour que le ballon flotte à hauteur constante (sans monter mais sans tomber aussi...) avec juste la quantité n_0 d'hydrogène suffisante, il faut donc d'après le principe d'Archimède que le volume d'air qu'il déplace ait un poids égal à la masse totale du ballon et de la sonde, soit 2.6 [Kg] dans notre cas!

Donc puisque 22 [L] d'air pèsent environ 29 grammes, il faut que le volume soit tel qu'il déplace 2.6 [Kg] d'air. Soit en faisant une simple règle de trois:

$$\frac{2.6 \text{ [Kg]}}{0.029 \text{ [Kg]}} \cdot 22 \text{ [L]} \cong 1'972 \text{ [L]} = 1.972 \text{ [m}^3\text{]} = V_0 \quad (69.61)$$

Donc si le ballon est sphérique, cela donne un rayon de:

$$V_0 = \frac{4}{3} \pi R_0^3 \Rightarrow R_0 = \sqrt[3]{\frac{3}{4\pi} V_0} \cong 0.78 \text{ [m]} \quad (69.62)$$

Soit un diamètre d'environ 1.56 [m] au sol. Ce qui est conforme à la réalité!

Il nous faut encore déterminer le nombre de moles d'hydrogène. Il vient immédiatement:

$$n_0 = \frac{2.6}{0.029} \cong 90 \text{ [moles]} \quad (69.63)$$

Maintenant que nous connaissons le nombre de moles dans le ballon, si nous connaissons la température et la pression à une hauteur de 22'000 [m] (altitude type d'un petit ballon-sonde) il ne nous reste qu'à appliquer la loi des gaz parfaits pour connaître le volume à cette altitude donné alors par la relation démontrée dans le chapitre de Mécanique des Milieux Continus:

$$V = n_0 R \frac{T}{P} \quad (69.64)$$

ainsi à 22'000 [m] d'altitude, nous avons environ selon les tables disponibles sur Internet:

$$\begin{aligned} T &\cong 218 [K] \\ P &\cong 4 \cdot 10^3 [Pa] \end{aligned} \quad (69.65)$$

Mais au fait le rayonnement solaire est environ 30% plus élevé à cette altitude et le ballon est considéré comme un système adiabatique (sans échange de chaleur) et ne restitue donc pas la puissance emmagasinée à l'environnement extérieur. Nous considérons alors que la température est au moins 30% plus élevée ce qui nous donne comme chiffres:

$$\begin{aligned} T &\cong 283 [K] \\ P &\cong 4 \cdot 10^3 [Pa] \end{aligned} \quad (69.66)$$

Nous avons donc:

$$V = n_0 R \frac{T}{P} \cong 90 \cdot 8.314 \cdot \frac{283}{5.4 \cdot 10^3} \cong 53 [m^3] \quad (69.67)$$

Nous aurions pu également utiliser (hypothèse adiabatique oblige!) la relation de Boyle-Mariotte (cf. chapitre de Mécanique Des Milieux Continus) pour arriver au même résultat:

$$P_0 \cdot V_0 = P \cdot V \Rightarrow V = \frac{P_0 \cdot V_0}{P} \quad (69.68)$$

Ce qui donne un rayon d'environ 2.33 [m] (diamètre donc d'environ 4.6 [m]) au lieu des 0.78 [m] au sol! Soit une augmentation du diamètre d'environ 300%. Il est plus important cependant de s'intéresser à l'augmentation de la surface pour déterminer les forces de contraintes élastiques sur le PVC. Nous avons donc avant:

$$S_0 = 4\pi R_0^2 = 4\pi \cdot 0.78^2 \cong 7 [m^2] \quad (69.69)$$

et après:

$$S = 4\pi R^2 = 4\pi \cdot 2.33^2 \cong 68 [m^2] \quad (69.70)$$

Soit une augmentation de la surface d'environ 1'000% alors qu'un élastomère-type (dont le PVC fait partie) ne résiste pas à une variation relative de 500%!!! Il est donc beaucoup plus aisé de comprendre sous le point de vue de la surface, pourquoi le ballon ne résiste pas à une augmentation du diamètre d'environ 300%.

Par ailleurs si nous appliquons de manière un peu abusive la loi de Hook au ballon (cf. chapitre de Mécanique des Milieux Continus), avec le module de Young du PVC qui est compris entre (source Wikipédia):

$$2.9 \cdot 10^7 - 3.3 \cdot 10^7 [N \cdot m^{-2}] \quad (69.71)$$

nous avons:

$$\frac{F}{S} = E \cdot \varepsilon = E \cdot \frac{|\Delta R|}{R_0} \cong 3.1 \cdot 10^7 \frac{2.12 - 0.73}{0.73} = 5.9 \cdot 10^7 [N \cdot m^{-2}] = 59 [MPa] \quad (69.72)$$

Ce qui est conforme aux données des tables numériques qui donnent une valeur limite inférieure élastique de 50 [MPa] et une limite supérieure de 80 [MPa] pour le PVC (source Wikipédia). Il nous est alors possible de calculer la hauteur minimale et maximale théorique que le ballon peut atteindre.

Ainsi pour la hauteur minimale, nous écrivons d'abord:

$$|\Delta R| = \frac{\frac{F}{S}}{E} R_0 = \frac{5 \cdot 10^7}{3.1 \cdot 10^7} 0.73 = 1.17 [m] \quad (69.73)$$

Ce qui correspond alors à un rayon final de:

$$R = R_0 + 1.38 = 1.9 [m] \quad (69.74)$$

Ce qui correspond à un volume de:

$$V = \frac{4}{3} \pi R^3 = 28.7 [m^3] \quad (69.75)$$

En appliquant Boyle-Mariotte:

$$P = \frac{P_0 V_0}{V} = \frac{1.01325 \cdot 10^5 \cdot 2.151}{28.7} = 7.5 \cdot 10^3 [Pa] \quad (69.76)$$

qui est une pression correspondante à une hauteur d'environ 18'000 [m] selon les mesures expérimentales (www.engineeringtoolbox.com) et c'est effectivement une hauteur à laquelle certains rares ballons en PVC éclatent.

Maintenant faisons de même avec la hauteur maximale:

$$|\Delta R| = \frac{\frac{F}{S}}{E} R_0 = \frac{8 \cdot 10^7}{3.1 \cdot 10^7} 0.73 = 1.88 [m] \quad (69.77)$$

Ce qui correspond alors à un rayon final de:

$$R = R_0 + 1.88 = 2.61 [m] \quad (69.78)$$

Ce qui correspond à un volume de:

$$V = \frac{4}{3} \pi R^3 = 74.5 [m^3] \quad (69.79)$$

En appliquant Boyle-Mariotte:

$$P = \frac{P_0 V_0}{V} = \frac{1.01325 \cdot 10^5 \cdot 2.151}{74.5} = 2.9 \cdot 10^3 [Pa] \quad (69.80)$$

qui est une pression correspondant à une hauteur d'environ 24'000 [m] selon les mesures expérimentales et c'est effectivement une hauteur à laquelle les ballons en PVC les plus hauts éclatent.

7. CYCLOGENÈSE ET ANTICYCLOGENÈSE

L'essentiel de la masse atmosphérique est contenu dans les 20 premiers kilomètres d'altitude, si bien que la météorologie à grande échelle se déroule sur une mince coquille sphérique (assimilable à de la mécanique des fluides bidimensionnelle).

Le moteur de la circulation atmosphérique dans les tropiques est le réchauffement solaire. À cause de l'inclinaison de 23.5 degrés de l'axe de rotation de la Terre, le Soleil n'est jamais plus qu'à quelques dixièmes de degré du zénith à midi tout au long de l'année dans les tropiques, ce qui donne un maximum de réchauffement autour de l'équateur géographique.

Il faut donc distinguer la circulation au voisinage des tropiques, caractérisée par de forts mouvements verticaux, dus aux convections thermiques, et la circulation des latitudes moyennes, faites quasiment que de mouvements horizontaux:

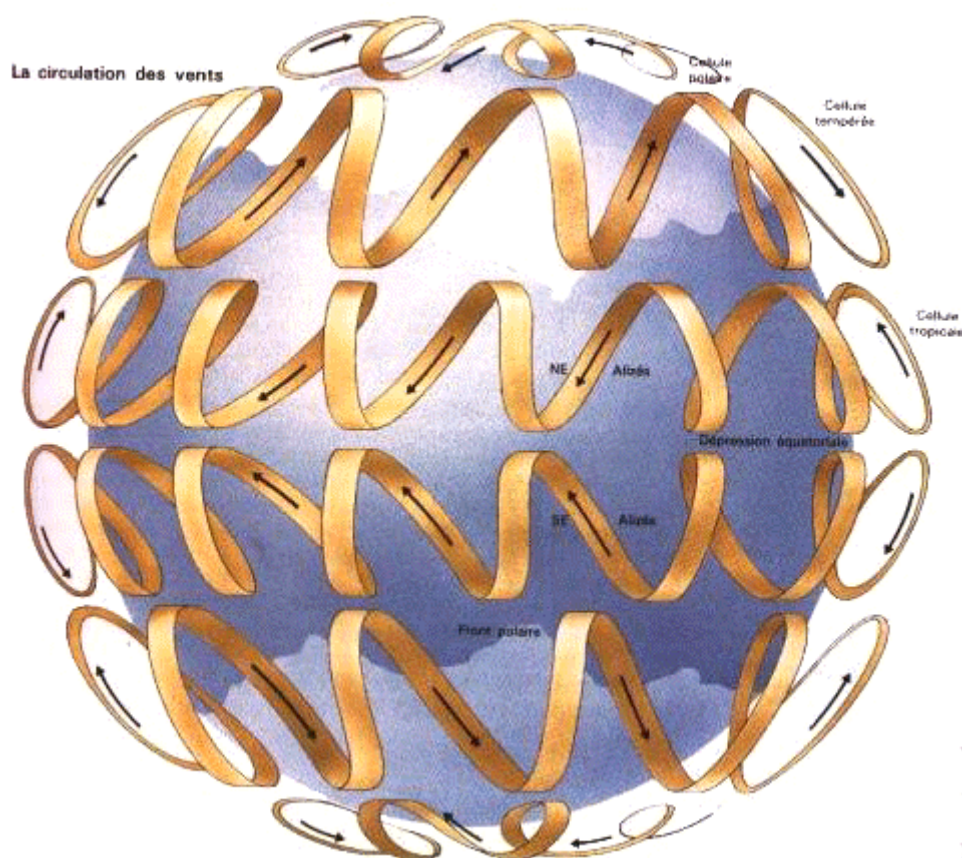


Figure: 69.4 - Schéma simplifié de la circulation des vents sur Terre

Supposons un moment que nous arrêtons complètement le mouvement de l'air dans l'atmosphère relativement à la surface de la planète, et que nous le laissions ensuite recommencer à tourner d'Ouest en Est (de gauche à droite sur les images) partir du repos. La force du gradient de pression pousse l'air à se mouvoir des régions de haute pression vers les régions de basse pression (appel du vide). Ces mouvements de convections sont appelés des "**cellules de Hadley**".

Toutefois, dès que le mouvement s'amorce la force de Coriolis (due à la rotation de la Terre) dévie donc les vents Nord-Sud en direction de l'Ouest et les vents Sud-Nord vers l'Est pour un observateur se situant au pôle Nord. Nous observons dès lors la formation de cyclones tournants dans le sens contraire des aiguilles d'une montre dans l'hémisphère Nord et inversement dans l'hémisphère Sud (à cause de la direction du vecteur $\vec{\omega}$ dans cette partie de l'hémisphère).

Plus la vitesse de l'air augmente, plus la force de Coriolis augmente de concert en accentuant la déviation. Éventuellement la force de Coriolis atteint une valeur égale et opposée à celle de la force du gradient de pression, produisant ainsi un écoulement d'une vitesse constante (sans accélération), parallèle aux isobares définissant ainsi la limite géométrique de la cellule de Hadley. C'est ce que nous appelons "l'équilibre géostrophique". En pratique, l'écoulement en dehors des tropiques est presque toujours en quasi-équilibre géostrophique.

En l'absence d'observations de vent, les météorologues peuvent estimer la force du vent en un point donné en mesurant sur une carte d'analyse météo le gradient de pression et la latitude. L'approximation géostrophique est purement diagnostique. Elle n'a pas de valeur prédictive, car son équation ne contient aucun terme de changement.

Dans les tropiques, où la force de Coriolis est de plus en plus faible jusqu'à être nulle à l'équateur, ce sont d'autres forces, comme la force centrifuge, qui viennent équilibrer la force de gradient de pression.

C'est ce que nous allons démontrer ici mathématiquement à l'aide de la mécanique des milieux continus (fluides) et la mécanique classique (voir chapitres correspondants).

Nous savons que dans notre système intervient donc les forces de pression (gradient), les forces centrifuges, les forces de pesanteur (gravité). Forces auxquelles il ne faut pas oublier d'ajouter la force (implicitement: l'accélération) de Coriolis interne au système (sous-entendu le cyclone) de pulsation $\vec{\omega}$ (cf. chapitre de Mécanique Classique):

$$\vec{a}_c = 2\vec{\omega} \times \vec{v} \quad (69.81)$$

et la force (implicitement: l'accélération) de Coriolis par unité de masse de fluide (la raison de ce choix d'unité paraîtra évidente quelques paragraphes plus loin) relativement à la pulsation $\vec{\Omega}$ de la Terre:

$$\vec{a}_C = 2\vec{\Omega} \times \vec{v} \quad (69.82)$$

Ainsi, comme nous le savons (cf. chapitre de Mécanique Classique), la force de Coriolis va tendre à dévier tout mouvement descendant vers la droite (Est) dans l'hémisphère Nord et tout mouvement montant vers la gauche (Ouest) dans l'hémisphère Sud (selon que l'on se place dans la direction du fluide en mouvement selon la figure précédente).

C'est ainsi que l'air à la base des cellules de Hadley, voyageant à basse altitude du tropique vers l'équateur sera dévié vers l'Ouest pour donner les Alizés de vents d'Est.

Nous avons par ailleurs démontré dans le chapitre de mécanique des milieux continus une forme particulière de l'équation d'Euler de 2ème forme qui était:

$$\partial_t \vec{v} + \vec{\nabla} \left(\frac{v^2}{2} + \frac{p}{\rho} + U \right) + 2\vec{\omega} \times \vec{v} = \vec{0} \quad (69.83)$$

Remaniée, cette relation s'écrit aussi:

$$\vec{\nabla} p = \rho \left(-\vec{\nabla} U - \left(\partial_t \vec{v} + \vec{\nabla} \frac{v^2}{2} + 2\vec{\omega} \times \vec{v} \right) \right) \quad (69.84)$$

Or, nous avons aussi démontré que:

$$-\vec{\nabla} U = \vec{g} \quad (69.85)$$

Il vient dans le référentiel Terrestre:

$$\vec{\nabla} p = \rho \left(\vec{g} - \left(\partial_t \vec{v} + \vec{\nabla} \frac{v^2}{2} + 2\vec{\omega} \times \vec{v} \right) \right) \quad (69.86)$$

Il s'agit donc de l'équation définissant la pression à l'intérieur du fluide considéré comme isolé. À cette relation, il faut donc encore soustraire les forces de pression de Coriolis dues au référentiel géocentrique pour obtenir la dynamique du système "cyclone":

$$\vec{F}_C = \rho 2\vec{\Omega} \times \vec{v} \quad (69.87)$$

Ce qui donne finalement:

$$\vec{\nabla} p = \rho \left(\vec{g} - \left(\partial_t \vec{v} + \vec{\nabla} \frac{v^2}{2} + 2\vec{\omega} \times \vec{v} + 2\vec{\Omega} \times \vec{v} \right) \right) \quad (69.88)$$

Soit sous forme condensée traditionnelle:

$$\vec{\nabla} p = \rho (\vec{g} - (\vec{a} + \vec{a}_C)) \quad (69.89)$$

Représentons maintenant la Terre dans une tranche Nord-Sud:

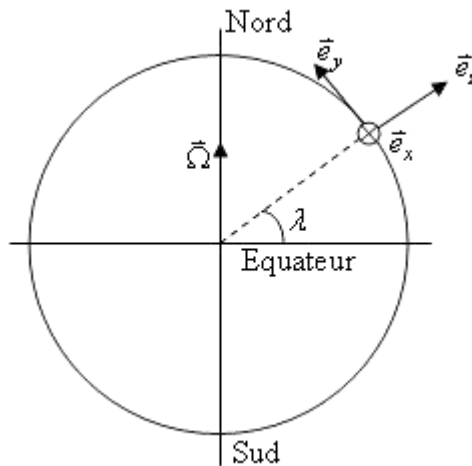


Figure: 69.5 - Tranche Nord-Sud de la Terre

Si nous agrandissons la repère lié au cyclone et y translatons le vecteur pulsation de la Terre, nous avons:

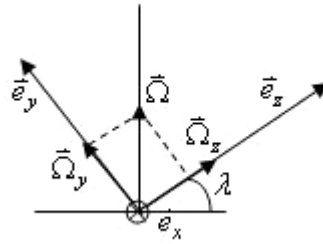


Figure: 69.6 - Repère lié au cyclone avec la pulsation plane

Soit:

$$\vec{\Omega} = \begin{pmatrix} \Omega_x \\ \Omega_y \\ \Omega_z \end{pmatrix} = \begin{pmatrix} 0 \\ \|\vec{\Omega}\| \cos(\lambda) \\ \Omega_z \cos(\pi/2 - \lambda) \end{pmatrix} = \begin{pmatrix} 0 \\ \Omega \cos \lambda \\ \Omega \sin \lambda \end{pmatrix} \quad (69.90)$$

Nous avons donc:

$$2\vec{\Omega} \times \vec{v} = 2 \begin{pmatrix} 0 \\ \Omega \cos \lambda \\ \Omega \sin \lambda \end{pmatrix} \times \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} = 2\Omega \begin{pmatrix} v_z \cos \lambda - v_y \sin \lambda \\ v_x \sin \lambda \\ v_x \cos \lambda \end{pmatrix} \quad (69.91)$$

Comme nous étudions les mouvements (quasi) horizontaux dans l'atmosphère à cette latitude, nous pouvons considérer que les particules de fluide sont assujetties à demeurer dans le plan horizontal $(\vec{e}_x, \vec{e}_y)$. Les composantes de la force de Coriolis pour un mouvement plan sont alors ($v_z = 0$):

$$\vec{a}_C = 2\vec{\Omega} \times \vec{v} = 2\Omega \begin{pmatrix} v_z \cos \lambda - v_y \sin \lambda \\ v_x \sin \lambda \\ v_x \cos \lambda \end{pmatrix} = 2\Omega \sin \lambda \begin{pmatrix} -v_y \\ v_x \\ 0 \end{pmatrix} = f \begin{pmatrix} -v_y \\ v_x \\ 0 \end{pmatrix} \quad (69.92)$$

où f est appelé "**paramètre de Coriolis**". Donc la force de Coriolis en océanographie et en météorologie est traditionnellement notée:

$$F_C = \rho f v \quad (69.93)$$

Le nombre f , positif dans l'hémisphère Nord, négatif dans l'hémisphère Sud, varie de 0 à 1.458 aux pôles alors que la force est de l'ordre du millième de Newton pour les masses de fluide (courants océaniques) et du même ordre de grandeur (car la vitesse compense la faible densité) pour les gaz (courants atmosphériques).

Nous appliquons maintenant l'approximation de l'équilibre géostrophique, c'est-à-dire que nous considérons que l'air est animé d'un mouvement rectiligne uniforme (vent géostrophique), en d'autres termes, nous négligeons l'action de la force centrifuge due à la rotation du tourbillon devant celle de la force de Coriolis due à la rotation de la Terre, ce qui revient à supposer que:

$$f \cdot v \gg R \cdot \omega^2 \quad (69.94)$$

avec R étant le rayon du tourbillon et ϖ sa pulsation. Puisque (cf. chapitre de Mécanique Classique):

$$\varpi = v / R \quad (69.95)$$

cette dernière inégalité devient:

$$\frac{v}{f \cdot R} \ll 1 \quad (69.96)$$

où:

$$R_o = \frac{v}{f \cdot R} \quad (69.97)$$

est appelé le "**nombre de Rossby**" et n'a pas de dimensions.

Remarque: Pour les moyennes latitudes ($\lambda = 45^\circ$), l'expérience et les mesures donnent $f = 10^{-4} [m/s]$ et $v = 10 [m/s]$. La valeur limite pour laquelle $R_o = 1$ est $R = 100 [km]$. Pour une échelle supérieure, comme c'est le cas pour les cyclones où $R \cong 100 [km]$, nous sommes donc proche de l'équilibre géostrophique. Pour une échelle inférieure, Coriolis est négligeable et le vent est accéléré des hautes vers les basses pressions.

Le nombre de Rossby représente donc le rapport entre les forces d'inerties et les forces dues à la rotation qui caractérisent le mouvement d'un fluide dans un repère tournant.

Ainsi, nous pouvons faire la différence entre un écoulement géophysique à fort nombre de Rossby ou à faible nombre de Rossby. Si le nombre de Rossby est très supérieur à l'unité, alors les forces de Coriolis dues par exemple à la rotation terrestre sont négligeables devant l'inertie de l'écoulement. Dans le cas contraire d'un nombre de Rossby très inférieur à l'unité, les forces de Coriolis dominent le mouvement du fluide.

Ainsi, si on se rapproche de l'équateur, f tendant vers 0 le nombre de Rossby devient très grand et aux pôles il devient respectivement très faible.

Dans le cadre de cette approximation, notre équation d'Euler peut alors s'écrire sous la forme:

$$\vec{\nabla} p \cong \rho(\vec{g} - \vec{a}_c) = \rho(\vec{g} - 2\vec{\Omega} \times \vec{v}) \quad (69.98)$$

et puisque nous nous intéressons qu'au plan horizontal de l'atmosphère cela se simplifie encore plus sous la forme:

$$\vec{\nabla}_{x,y} p \cong -\rho 2(\vec{\Omega} \times \vec{v})_{x,y} = -\rho f \begin{pmatrix} -v_y \\ v_x \end{pmatrix} \quad (69.99)$$

Soit totalement sous forme vectorielle développée et en reprenant la majuscule P pour la pression comme il est d'usage en météorologie:

$$\begin{pmatrix} \frac{\partial P}{\partial x} \\ \frac{\partial P}{\partial y} \end{pmatrix} = -\rho f \begin{pmatrix} -v_y \\ v_x \end{pmatrix} \quad (69.100)$$

Il vient ainsi que:

$$\begin{pmatrix} -v_y \\ v_x \end{pmatrix} = -\frac{1}{\rho f} \begin{pmatrix} \frac{\partial P}{\partial x} \\ \frac{\partial P}{\partial y} \end{pmatrix} = \frac{1}{\rho f} \begin{pmatrix} -\frac{\partial P}{\partial x} \\ -\frac{\partial P}{\partial y} \end{pmatrix} \quad (69.101)$$

soit:

$$\begin{pmatrix} v_y \\ v_x \end{pmatrix} = \frac{1}{\rho f} \begin{pmatrix} \frac{\partial P}{\partial x} \\ -\frac{\partial P}{\partial y} \end{pmatrix} \quad (69.102)$$

Donc sous forme conventionnelle:

$$\vec{v}_{xy} = \begin{pmatrix} v_x \\ v_y \end{pmatrix} = \frac{1}{\rho f} \begin{pmatrix} -\frac{\partial P}{\partial y} \\ \frac{\partial P}{\partial x} \end{pmatrix} = \frac{1}{\rho f} \vec{e}_z \times (\vec{\nabla} P) \quad (69.103)$$

La norme étant donnée par:

$$v = \|\vec{v}_{xy}\| = \begin{pmatrix} v_x \\ v_y \end{pmatrix} = \sqrt{\left(-\frac{1}{\rho f} \frac{\partial P}{\partial y}\right)^2 + \left(-\frac{1}{\rho f} \frac{\partial P}{\partial x}\right)^2} = \frac{1}{\rho f} \sqrt{\left(\frac{\partial P}{\partial y}\right)^2 + \left(\frac{\partial P}{\partial x}\right)^2} \quad (69.104)$$

Soit:

$$\boxed{v = \frac{1}{\rho f} \frac{dP}{dR}} \quad (69.105)$$

relation qui est appelée "**équation des vents (géostrophiques)**".

Quatre scénarios sont à considérer:

1. Nous sommes dans l'hémisphère Nord et donc f est positif. Supposons que dP/dR soit positif, la pression augmente alors en s'éloignant du centre du tourbillon (qui lui est donc un minimum de basse pression). Dès lors v est positif et nous avons un tourbillon appelé "dépression" dans l'hémisphère Nord. Ainsi, le fluide (le vent) souffle autour de la dépression dans le sens antihoraire (vers l'Ouest) dans l'hémisphère Nord.

Définitions:

D1. Une "**dépression**" (ou "**basse pression**") est une zone où la pression atmosphérique diminue horizontalement vers un centre de basse pression, c'est-à-dire un minimum local de pression.

D2. Les systèmes atmosphériques intenses à circulation autour d'un centre fermé de basse pression (comme un aspirateur cela attire les nuages d'où le fait que les cyclones sont visibles sur des photos satellites) reçoivent systématiquement le terme plus général de "**cyclone**" ou de "**cyclone tropical**".

Remarque: Nous associons les dépressions au mauvais temps, car la dynamique qui entoure une dépression présuppose l'existence de courants ascendants (peuvent difficilement entrer dans le sol donc la seule voie d'échappement est le haut!) qui provoquent des nuages et de la précipitation. De plus, le gradient de pression autour d'une dépression peut engendrer de forts vents.

2. Nous sommes toujours dans l'hémisphère Nord et donc f est positif. Supposons que dP/dR soit cette fois négatif, la pression diminue alors en s'éloignant du centre du tourbillon (qui lui est donc un maximum de haute pression). Dès lors v est négatif et nous avons un tourbillon appelé "haute pression" dans l'hémisphère Nord. Ainsi, le fluide (le vent) souffle autour de la haute pression dans le sens horaire (vers l'Est) dans l'hémisphère Nord.

Définitions:

D1. Une "**haute pression**" est une zone où la pression atmosphérique augmente horizontalement vers un centre de haute pression, c'est-à-dire un maximum local de pression.

D2. Les systèmes atmosphériques intenses à circulation autour d'un centre fermé de haute pression (comme un ventilateur cela rejette et disperse les nuages d'où le fait que les anticyclones ne sont pas visibles de manière simple sur les photos satellites) reçoivent systématiquement le terme plus général de "**anticyclone**".

Remarque: Les anticyclones généralement apportent du beau temps et des ciels clairs. La dynamique atmosphérique fait en sorte que l'air aux altitudes moyennes y est relativement chaud et sec, et donc sans nuages.

3. Nous sommes toujours dans l'hémisphère Sud et donc f est négatif. Supposons que dP/dR soit positif, la pression augmente alors en s'éloignant du centre du tourbillon. Dès lors v est négatif et nous avons un tourbillon appelé "haute pression" (ou "anticyclone") dans l'hémisphère Sud. Ainsi, le fluide (le vent) souffle autour de la haute pression mais dans le sens antihoraire (vers l'Ouest) dans l'hémisphère Sud.

4. Nous sommes toujours dans l'hémisphère Sud et donc f est négatif. Supposons que dP/dR soit négatif, la pression diminue alors en s'éloignant du centre du tourbillon. Dès lors v est positif et nous avons un tourbillon appelé "basse pression" (ou "cyclone") dans l'hémisphère Sud. Ainsi, le fluide (le vent) souffle autour de la basse pression dans le sens horaire (vers l'Est) dans l'hémisphère Sud.

Remarque: Il est donc possible de dire de manière générale sur les grandes dimensions que le vent arrive des hautes pressions (Anticyclone) pour se diriger vers les basses pressions (Dépression).

Voici un exemple d'une image de dépression (anticyclone) et haute pression (cyclone) dans l'hémisphère Nord tel que le représentent les professionnels de la météorologie:

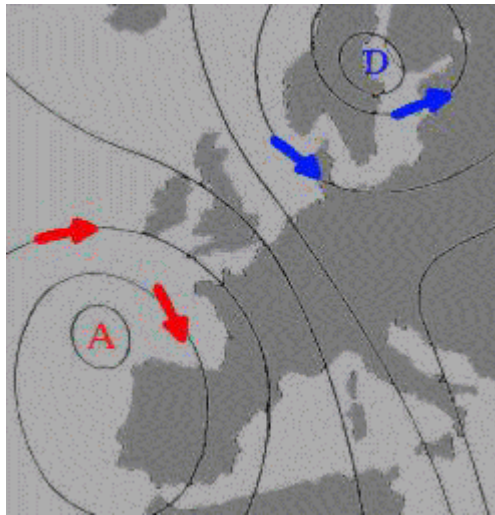


Figure: 69.7 - Basse pression et haute pression

Nous pouvons effectivement observer que la dépression (D) tourne dans le sens antihoraire et la haute pression (A) dans le sens horaire (et inversement dans l'hémisphère Sud).

Remarque: L'air au centre d'un anticyclone (A) descend vers la surface, subissant une compression et par conséquent un échauffement. Pour une dépression (D), c'est le phénomène inverse qui se produit.

8. MARÉES

Parmi les phénomènes de la nature, la marée est l'un des plus majestueux par son ampleur et par sa puissance, l'un des plus surprenants par sa régularité et par la discrétion de ses causes. On comprend sans peine non seulement qu'il se soit imposé à l'attention des navigateurs mais encore qu'il ait, depuis la plus lointaine antiquité, suscité les recherches des savants les plus émérites.

Pour aborder le sujet des marées de manière simple, nous pouvons partir d'un constat logique: Si l'attraction lunaire était identique en chaque point de la Terre, il n'y aurait pas de marées. Il faut donc aborder l'étude des marées sur les différences de forces. L'influence de la Lune sur la marée est appelée "**composante diurne**".

8.1. PREMIÈRE APPROCHE

Considérons pour l'étude naïve du phénomène une masse d'eau m à l'équateur et aux pôles. Nous allons calculer la force d'attraction sur cette masse par rapport au centre de la Terre et en prenant en compte l'influence de la Lune de masse M_L .

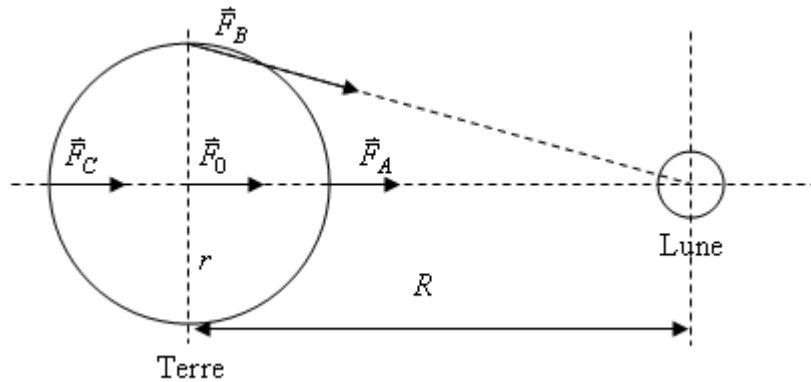


Figure: 69.8 - Configuration Terre-Lune pour l'étude des marées

Commençons par calculer la force F_A à l'équateur au point le plus proche de la Lune relativement à la figure ci-dessus.

Nous avons alors:

$$\begin{aligned}\Delta F_L &= F_A - F_0 = G \frac{mM_L}{(R-r)^2} - G \frac{mM_L}{R^2} = GmM_L \frac{R^2 - (R-r)^2}{R^2(R-r)^2} \\ &= GmM_L \frac{R^2 - R^2 + 2Rr - r^2}{R^4 - 2R^3r + R^2r^2} \cong GmM_L \frac{2Rr}{R^4} = 2GmM_L \frac{r}{R^3} = 2f\end{aligned}\quad (69.106)$$

en considérant que $R \gg r$ et en notant la "force de marée statique":

$$f = GmM_L \frac{r}{R^3} \quad (69.107)$$

Une application numérique pour une masse m de 1 [Kg] donne:

$$f = 1 \cdot 10^{-6} [m \cdot s^{-2}] \quad (69.108)$$

Sous forme vectorielle nous avons bien évidemment:

$$\Delta \vec{F}_L = 2GmM_L \frac{r}{R^2} \frac{\vec{R}}{\|\vec{R}\|} \quad (69.109)$$

Comme la distance Terre-Lune atteint environ 60 rayons terrestres, l'intensité de l'accélération varie à peu près linéairement (...) le long de la portion terrestre d'une droite passant par le centre de la Lune. C'est notamment le cas pour le segment qui relie les deux points antipodaux A et C de la figure ci-dessus. Nous pouvons donc écrire, O désignant le centre de la Terre:

$$\begin{aligned}\vec{F}_A &\cong \vec{F}_0 + \Delta \vec{F}_L \\ \vec{F}_C &\cong \vec{F}_0 - \Delta \vec{F}_L\end{aligned}\quad (69.110)$$

Nous devons maintenant séparer les deux contributions de la Lune.

- La force $\vec{F}_0$ qui s'applique sur le centre de masse G est donc uniforme à la planète par construction. C'est cette force qui est responsable de la révolution de notre planète autour du centre de masse commun aux deux astres.

- Le terme résiduel $\pm \Delta \vec{F}_L$ se superpose et prend des valeurs opposées aux antipodes. Elle est responsable des marées (en première approximation dans ce modèle simpliste).

Ainsi, la force due à la Lune est de signe opposé sur l'horizontale. Nous avons alors deux marées (lunaires) par jour à des lieux antipodaux:

- Celle de la Lune qui attire (de ce côté-ci de la Terre)
- Celle de la Lune qui repousse (du côté opposé de la Terre)

Soit sous forme schématique (sans aucun respect des proportions réelles):

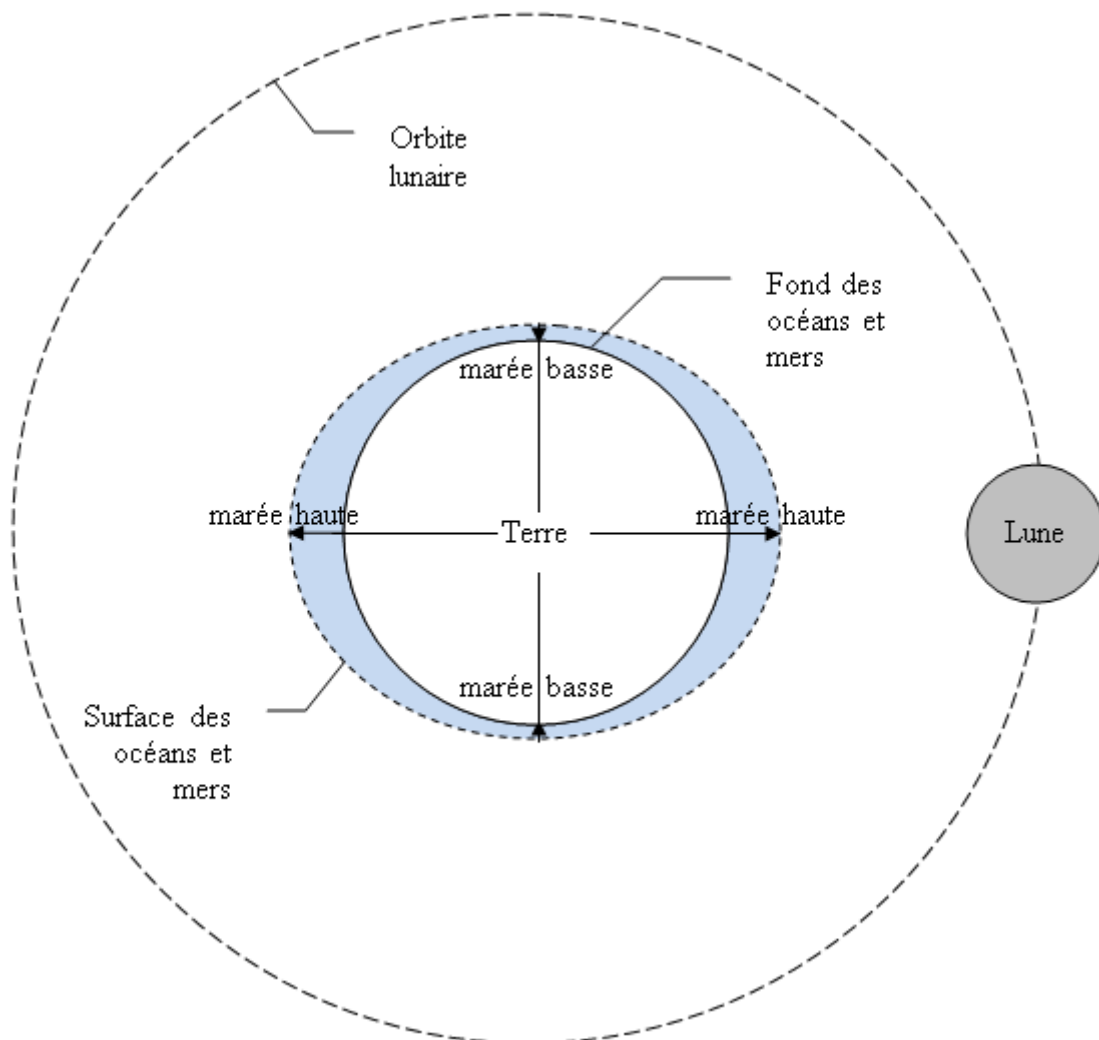


Figure: 69.9 - Lieux antipodaux des marées

Si l'on considérait la surface de la Terre comme parfaitement sphérique et recouverte d'eau, elle prendrait alors la forme d'une ellipsoïde dont l'axe serait dirigé vers l'astre générant la marée. Nous observerions alors des marées dont les pleines et basses mers auraient lieu deux fois par jour et toujours à la même heure. Nous appelons cette situation la "**marée statique**" et le modèle correspondant "**modèle statique des marées**".

Il convient de préciser que le jeu subtil entre la rotation de la Terre et la Lune produit des frottements gigantesques au niveau des masses d'eau qui ont pour effet de ralentir la vitesse de rotation de la Terre d'environ 1 seconde tous les mille ans.

8.2. DEUXIÈME APPROCHE

Pour la deuxième approche, qui vaut la peine d'être vue pour la culture générale et aussi parce qu'elle présente un autre aspect intéressant de l'explication de l'attraction entre deux astres, considérons le schéma suivant:

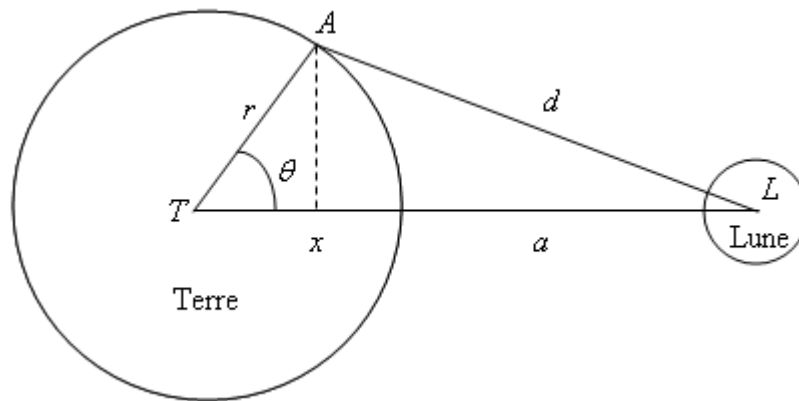


Figure: 69.10 - Approche schématique de l'approche

Nous avons démontré dans le chapitre de Trigonométrie le théorème du cosinus (formule d'Al-Kashi) qui nous donne:

$$d^2 = a^2 + r^2 - 2ar \cos \theta \quad (69.111)$$

d'où le potentiel gravitationnel (cf. chapitre d'Astronomie):

$$\varphi = \frac{GM}{d} = \frac{GM}{\sqrt{a^2 + r^2 - 2ar \cos \theta}} = \frac{GM}{a} \frac{1}{\sqrt{1 + \frac{r^2}{a^2} - \frac{2r \cos \theta}{a}}} \quad (69.112)$$

Mais:

$$\frac{r}{a} \ll 1 \quad (69.113)$$

et le potentiel gravitationnel est de la forme:

$$f(x) = \frac{1}{\sqrt{1+x}} \quad (69.114)$$

que nous pouvons développer en série de Maclaurin (cf. chapitre Suites Et Séries) jusqu'à l'ordre 2:

$$\begin{aligned}
 f(x) &= \frac{1}{\sqrt{1+x}} \underset{x_0=0}{\cong} \frac{1}{0!} f(x_0) x^0 + \frac{1}{1!} f'(x_0) x^1 + \frac{1}{2!} f''(x_0) x^2 \\
 &= \frac{1}{0!} \frac{1}{(1+x_0)^{1/2}} - \frac{1}{1!} \frac{1}{2} \frac{1}{(1+x_0)^{3/2}} x + \frac{1}{2!} \frac{3}{4} \frac{1}{(1+x_0)^{5/2}} x^2 \\
 &= \frac{1}{0!} - \frac{1}{1!} \frac{1}{2} x + \frac{1}{2!} \frac{3}{4} x^2 = 1 - \frac{1}{2} x + \frac{3}{8} x^2
 \end{aligned} \quad (69.115)$$

Le potentiel gravitationnel devient alors:

$$\begin{aligned}
 \varphi &= \frac{GM}{a} \frac{1}{\sqrt{1 + \frac{r^2}{a^2} - \frac{2r \cos \theta}{a}}} = \frac{GM}{a} \left[1 - \frac{1}{2} \left(\frac{r^2}{a^2} - \frac{2r \cos \theta}{a} \right) + \frac{3}{8} \left(\frac{r^2}{a^2} - \frac{2r \cos \theta}{a} \right)^2 \right] \\
 &= \frac{GM}{a} \left[1 - \frac{1}{2} \frac{r^2}{a^2} + \frac{r \cos \theta}{a} + \frac{3}{8} \left(\frac{r^4}{a^4} - 2 \frac{2r \cos \theta}{a} \frac{r^2}{a^2} + \frac{4r^2 \cos^2 \theta}{a^2} \right) \right]
 \end{aligned} \quad (69.116)$$

Si nous ne gardons que les termes de puissance 1 et 2 en r/a , il reste:

$$\begin{aligned}
 \varphi &\cong \frac{GM}{a} \left[1 + \frac{r}{a} \cos \theta - \frac{1}{2} \frac{r^2}{a^2} + \frac{3}{8} \frac{4r^2 \cos^2 \theta}{a^2} \right] = \frac{GM}{a} \left[1 + \frac{r}{a} \cos \theta - \frac{1}{2} \frac{r^2}{a^2} + \frac{3}{2} \frac{r^2 \cos^2 \theta}{a^2} \right] \\
 &= \frac{GM}{a} \left[1 + \frac{r}{a} \cos \theta + \frac{1}{2} \frac{r^2}{a^2} (3 \cos^2 \theta - 1) \right]
 \end{aligned} \quad (69.117)$$

Le premier terme du potentiel est :

$$\varphi_1 = \frac{GM}{a} \quad (69.118)$$

C'est le potentiel pour $r = 0$, c'est-à-dire le potentiel créé par la Lune au centre de la Terre.

Ce terme ne contient aucune variable, il est constant et donc son gradient est nul, il ne donne lieu à aucune force puisque:

$$\vec{F} = -m \vec{\nabla} \cdot \varphi_1 = 0 \quad (69.119)$$

Le deuxième terme:

$$\varphi_2 = \frac{GM}{a^2} r \cos \theta \quad (69.120)$$

contient les deux variables r et θ . Son gradient ne sera pas nul. Il va générer une force de gravitation que nous allons calculer, mais en utilisant une astuce.

Puisque:

$$x = r \cos \theta \quad (69.121)$$

Il vient:

$$\varphi_2 = \frac{GM_L}{a^2} x \quad (69.122)$$

Dès lors, le gradient en coordonnées cartésiennes se réduit pour la masse dm se situant au point A à:

$$d\vec{F} = -dm \vec{\nabla} \cdot \varphi_2 = -dm \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{pmatrix} \varphi_2 = -dm \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{pmatrix} \frac{GM_L}{a^2} x = -dm \begin{pmatrix} \frac{GM_L}{a^2} \\ 0 \end{pmatrix} \quad (69.123)$$

Donc tous les éléments de la Terre subissent de la part de la Lune des forces parallèles (selon l'axe X seulement) dirigées vers la Lune. La masse totale de la Terre est la somme de toutes ces masses et la force totale que subit la Terre de la part de la Lune est la somme des forces élémentaires. Donc:

$$\vec{F} = \sum -dm \vec{\nabla} \cdot \varphi_2 = \sum -dm \begin{pmatrix} \frac{GM_L}{a^2} \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{GM_T M_L}{a^2} \\ 0 \end{pmatrix} \quad (69.124)$$

et donc la force totale est seulement selon l'axe X et est donnée par:

$$F_x = \frac{GM_T M_L}{a^2} \quad (69.125)$$

Cette force est la même que si toute la masse de la Terre était concentrée au centre T

et si toute la masse de la Lune était concentrée au centre L . La Lune subit de la part de la Terre la même force de sens contraire, c'est le principe des actions mutuelles. C'est cette dernière qui oblige la Lune à tourner autour de la Terre.

Le troisième terme est celui responsable des marées:

$$\varphi_3 = \frac{GM}{a} \frac{1}{2} \frac{r^2}{a^2} (3 \cos^2 \theta - 1) \quad (69.126)$$

La force dérivant du gradient en coordonnées polaires étant donnée par:

$$\vec{F} = -m \begin{pmatrix} \frac{\partial}{\partial r} \\ \frac{1}{r} \frac{\partial}{\partial \theta} \end{pmatrix} \cdot \varphi_3 = -m \begin{pmatrix} \frac{\partial}{\partial r} \\ \frac{1}{r} \frac{\partial}{\partial \theta} \end{pmatrix} \cdot \frac{GM}{a} \frac{1}{2} \frac{r^2}{a^2} (3 \cos^2 \theta - 1) \quad (69.127)$$

Donc, la composante radiale de la force est:

$$F_r = -m \frac{\partial}{\partial r} \frac{GM}{a} \frac{1}{2} \frac{r^2}{a^2} (3 \cos^2 \theta - 1) = -\frac{GmMr}{a^3} (3 \cos^2 \theta - 1) \quad (69.128)$$

la composante orthoradiale:

$$\begin{aligned}
 F_{\theta} &= -m \frac{1}{r} \frac{\partial}{\partial \theta} \cdot \frac{GM}{a} \frac{1}{2} \frac{r^2}{a^2} (3 \cos^2 \theta - 1) = -\frac{GmMr}{2a^3} \frac{\partial}{\partial \theta} (3 \cos^2 \theta - 1) \\
 &= -\frac{3GmMr}{2a^3} \frac{\partial}{\partial \theta} (\cos^2 \theta) = -\frac{3GmMr}{2a^3} (-2 \cos \theta \sin \theta) = \frac{3GmMr}{a^3} \cos \theta \sin \theta
 \end{aligned} \quad (69.129)$$

Donc:

$$\vec{F} = \begin{pmatrix} F_r \\ F_{\theta} \end{pmatrix} = \begin{pmatrix} -\frac{GmMr}{a^3} (3 \cos^2 \theta - 1) \\ \frac{3GmMr}{a^3} \cos \theta \sin \theta \end{pmatrix} \quad (69.130)$$

Pour une masse m d'eau, la force de marée est:

$$\vec{F} = -m \vec{\nabla} \cdot \varphi_3 \quad (69.131)$$

Pour $\theta = 0, \frac{\pi}{2}, \pi$, la composante orthoradiale F_{θ} s'annule. Nous avons alors pour ces trois angles:

$$\vec{F} = \begin{pmatrix} F_r \\ F_{\theta} \end{pmatrix} = \begin{pmatrix} -\frac{GmMr}{a^3} (3 \cos^2 \theta - 1) \\ 0 \end{pmatrix} \quad (69.132)$$

Pour $\theta = 0$ la force est uniquement radiale:

$$F_r = -\frac{GmMr}{a^3} (3 \cos^2 \theta - 1) = -\frac{GmMr}{a^3} (3 \cdot (1)^2 - 1) = -\frac{2GmMr}{a^3} \quad (69.133)$$

Le signe de cette force est positif et il est dirigé vers la Lune. Pour $\theta = \pi$ la force est aussi uniquement radiale et on retombe sur la même expression:

$$F_r = -\frac{GmMr}{a^3} (3 \cos^2 \theta - 1) = -\frac{GmMr}{a^3} (3 \cdot (-1)^2 - 1) = -\frac{2GmMr}{a^3} \quad (69.134)$$

Donc, l'amplitude est la même pour l'angle $\theta = 0 = \pi$.

Par contre, si le lecteur se rappelle que pour revenir en composantes cartésiennes nous avons:

$$\begin{aligned}
 x &= r \cos \theta \\
 y &= r \sin \theta
 \end{aligned} \quad (69.135)$$

Nous voyons alors que pour $\theta = 0$, la force est orientée dans le sens positif de l'axe X puisque x sera positif (mais nulle selon Y). Il y a donc une marée dans la direction de la Lune (intuitif)

Nous voyons aussi que pour $\theta = \pi$, la force est orientée dans le sens négatif de l'axe X puisque x sera négatif (mais nulle selon Y). Il y a donc une marée dans la direction opposée à la Lune (contre intuitif).

Pour $\theta = \pi/2, 3\pi/2$, la composante x est nulle et la composante radiale sera:

$$F_r = -\frac{GmMr}{a^3}(3\cos^2\theta - 1) = \frac{GmMr}{a^3} \quad (69.136)$$

et elle est dirigée vers le centre de la Terre puisque:

$$\begin{aligned} x &= r \cos \theta \\ y &= r \sin \theta \end{aligned} \quad (69.137)$$

Nous voyons alors que pour $\theta = \pi/2$, la force est orientée dans le sens positif de l'axe Y puisque y sera positif (mais nulle selon X). Nous voyons aussi que pour $\theta = 3\pi/2$, la force est orientée dans le sens négatif de l'axe Y puisque y sera négatif (mais nulle selon x).

En réalité, les marées sont beaucoup plus complexes que le modèle ci-dessus (excentricité de l'orbite lunaire, superposition de la marée diurne, orbite lunaire, alignement Lune-Soleil, inclinaison du plan de l'orbite de la Lune, équinoxes, etc.). Voici une superbe animation de l'élévation de la surface des océans en mètre, sur 1 cycle de marée, calculée à partir d'un modèle plus élaboré:

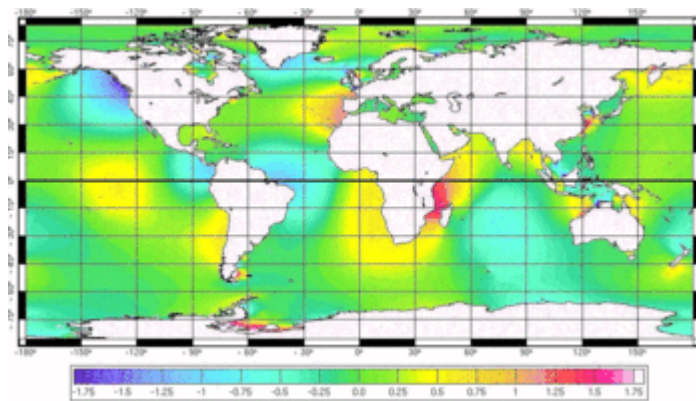


Figure: 69.11 - Complexité réelle des marées sur Terre (source: Wikipédia)

Remarque: Le phénomène est donc dû à la déformation de la surface des océans par suite des attractions combinées des corps célestes. Ce mouvement peut même détruire l'astre qui le subit: si la force de marée l'emporte sur la force de gravitation de ses constituants, l'astre se désagrège. Cette limite où les forces de marées l'emportent sur la force gravitationnelle s'appelle "limite de Roche" (cf. chapitre d'Astronomie).

Outre la marée diurne due à l'attraction de la Lune, il faut compter en plus sur une marée due à la force centrifuge du mouvement de la Terre et de la Lune autour de leur centre de masse (mais cela dépend des latitudes, du relief et de plein d'autres paramètres objectivement car sur certains endroits de la planète il n'y a qu'une seule marée par jour). Effectivement, la Terre et la Lune tournent autour du centre de masse qui définit l'orbite de couple Terre-Lune (les échelles ne sont pas respectées):

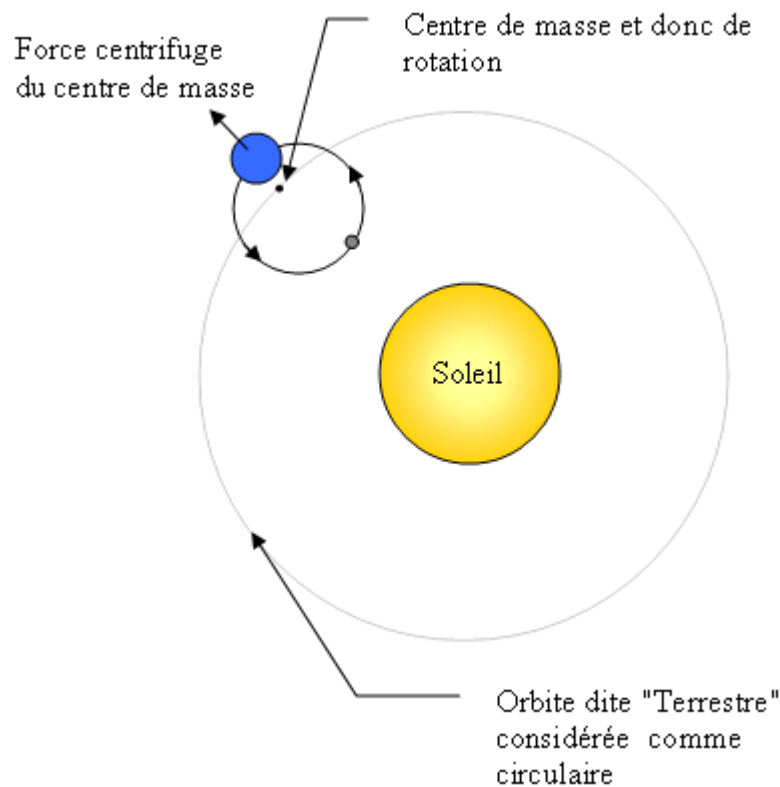


Figure: 69.12 - Principe des marées superposées aux marées diurne

et nous ferons l'impasse sur les marées d'équinoxes et autres... jusqu'à ce jour...

9. ÉQUATION DE LORENZ

La "**convection libre**" ou "**convection naturelle**" est le régime d'écoulement obtenu lorsque nous chauffons un fluide sans qu'il n'y ait d'écoulement extérieur imposé. C'est le cas des mouvements de convections de l'atmosphère (gaz chauds dans gaz froids), des mouvements de convections de la roche en fusion responsable de la tectonique des plaques, des mouvements de l'eau chaude sous pression dans les geysers et de bien d'autres phénomènes...

Ces écoulements sont inexplicables si nous ne couplons pas les équations de la dynamique et de la thermodynamique!

Nous allons dans ce contexte établir le fameux système des équations de Lorenz au prix cependant de nombreuses approximations et hypothèses afin de simplifier au maximum les calculs et les outils mathématiques utilisés (car à l'époque du développement du modèle les ordinateurs n'étaient pas ce qu'ils sont aujourd'hui).

Nous montrerons ainsi dans le cadre de la convection (une des dynamiques importante de notre atmosphère) que les équations qui déterminent certains paramètres du mouvement sont très sensibles aux conditions initiales ce qui a pour but de montrer la difficulté de la prévision à plus ou moins long terme avec des modèles théoriques déterministes (raison pour laquelle en météorologie il est fait usage de nos jours de la méthode des éléments finis).

A priori, la densité ρ est fonction de la température et de la pression par la loi d'état des gaz parfaits (cf. chapitre de Mécanique Des Milieux Continus). Il est donc naturel de penser que si nous chauffons une paroi, la température du fluide environnant augmente par diffusion. La stratification de pression s'en trouve changée, le gradient de pression crée le mouvement.

Dans tous les chapitres du site, nous avons jusqu'à présent négligé toute variation de ρ . Mais le découplage n'est plus valable ici puisque c'est le chauffage qui provoque le mouvement. Nous allons donc permettre une variation de la densité avec le chauffage en supposant cependant que cette perturbation est petite. Il faut donc réintroduire une variation de ρ autour d'une position d'équilibre: le repos. En revanche la viscosité sera supposée constante.

Soit donc un fluide au repos et à la température T_∞ au loin, il est en présence d'une paroi chauffée à la température T_p . Pour obtenir la dépendance de ρ , rappelons les coefficients thermo-élastiques classiques (cf. chapitre de Thermodynamique):

- Coefficient de compressibilité (ou de dilatation suivant l'écriture en termes de densité) isobare:

$$\alpha = \frac{1}{V} \frac{\partial V}{\partial T} = \frac{1}{M \rho} \frac{\partial M \rho}{\partial T} = \frac{1}{\rho} \frac{\partial \rho}{\partial T} \quad (69.138)$$

- Coefficient de compressibilité isotherme:

$$\chi = -\frac{1}{V} \frac{\partial V}{\partial P} = -\frac{1}{M \rho} \frac{\partial M \rho}{\partial P} = -\frac{1}{\rho} \frac{\partial \rho}{\partial P} \quad (69.139)$$

En admettant maintenant que la densité est principalement reliée à la température (pour faire simple) nous pouvons écrire (cette hypothèse marche bien pour les fluides mais pas trop... pour les gaz!!):

$$\rho = f(T) \quad (69.140)$$

En utilisant la forme générale du développement de Taylor (cf. chapitre de Suites et Séries):

$$f(x) = \sum_{k=0}^n \frac{1}{k!} \frac{d^k f(x_0)}{dx^k} (x - x_0)^k \quad (69.141)$$

Nous avons alors c'est une approche à la façon ingénieur...:

$$\begin{aligned} \rho(T) &= \sum_{k=0}^n \frac{1}{k!} \frac{d^k f(x_0)}{dx^k} (T - T_\infty)^k \cong \rho_\infty + (T - T_\infty) \frac{d\rho(T_\infty)}{dT} = \rho_\infty + \alpha \rho_\infty (T - T_\infty) \\ &= \rho_\infty (1 + \alpha(T - T_\infty)) = \rho_\infty (1 + \bar{T} \alpha(T_p - T_\infty)) = \rho_\infty (1 + \bar{T} \varepsilon) \end{aligned} \quad (69.142)$$

Soit:

$$\boxed{\rho = \rho_\infty (1 + \bar{T} \varepsilon)} \quad (69.143)$$

où $\bar{T}$ est donc un coefficient sans dimensions (comme ε ...) plus facilement mesurable expérimentalement.

L'équation de continuité (cf. chapitre de Thermodynamique) ou de bilan de masse:

$$\frac{\partial \rho(T)}{\partial t} - \vec{\nabla} \circ (\rho(T) \cdot \vec{v}) = 0 \quad (69.144)$$

devient alors:

$$\frac{\partial \rho_{\infty} (1 + \bar{T} \varepsilon)}{\partial t} - \vec{\nabla} \circ (\rho_{\infty} (1 + \bar{T} \varepsilon) \cdot \vec{v}) = \rho_{\infty} \varepsilon \frac{\partial \bar{T}}{\partial t} - \rho_{\infty} (1 + \bar{T} \varepsilon) \vec{\nabla} \circ \vec{v} = 0 \quad (69.145)$$

au premier ordre en ε . De plus, nous avons montré dans le chapitre de Mécanique des Milieux Continus que si le fluide est incompressible:

$$\vec{\nabla} \circ \vec{v} = 0 \quad (69.146)$$

Retenons qu'en première approximation le fluide est incompressible. Il ne reste alors que:

$$\rho_{\infty} \varepsilon \frac{\partial \bar{T}}{\partial t} = 0 \quad (69.147)$$

Comme nous souhaitons étudier un écoulement en présence de gravité, il serait judicieux de poser:

$$P = p + \rho_{\infty} g z \quad (69.148)$$

et donc de ne s'intéresser qu'aux variations autour de la position d'équilibre hydrostatique (δp est sans dimensions!). Nous avons démontré toujours dans le même chapitre de Mécanique des Milieux continus que dans le cas du fluide incompressible avec viscosité, l'équation d'Euler de 1^{ère} forme (équation du mouvement):

$$\rho d_t \vec{v} = -\vec{\nabla} P + \mu_d \vec{\nabla}^2 \vec{v} + \rho \vec{f} \quad (69.149)$$

Intéressons-nous dans un premier temps aux deux termes:

$$-\vec{\nabla} P + \rho \vec{f} \quad (69.150)$$

qui s'écrivent selon l'axe Z:

$$-\frac{\partial}{\partial z} P + \rho f_z = -\frac{\partial}{\partial z} P + \rho g = 0 \quad (69.151)$$

Lorsqu'il y a mouvement, la projection suivant Z fait donc apparaître:

$$-\frac{\partial}{\partial z} P + \rho g \quad (69.152)$$

que nous récrivons alors:

$$-\frac{\partial}{\partial z} (p + \rho_{\infty} g z) + \rho g = -\rho_{\infty} g \frac{\partial}{\partial z} z - \frac{\partial}{\partial z} p + \rho g = -\rho_{\infty} g - \frac{\partial}{\partial z} p + \rho g \quad (69.153)$$

Soit:

$$-\frac{\partial}{\partial z} p + (\rho - \rho_\infty) g \quad (69.154)$$

puisque:

$$(\rho - \rho_\infty) = \rho_\infty \varepsilon \bar{T} \quad (69.155)$$

Il vient:

$$-\frac{\partial}{\partial z} p + \rho_\infty \varepsilon \bar{T} g \quad (69.156)$$

Il reste donc une force de flottabilité dirigée vers le haut.

La variation de la densité en fonction de la température dans le produit $\rho d_t \vec{v}$ de la relation:

$$\rho d_t \vec{v} = -\vec{\nabla} P + \rho \vec{f} \quad (69.157)$$

sera négligée ($\rho d_t \vec{v} \cong \rho_\infty d_t \vec{v}$) car nous nous restreindrons au cas où la vitesse est petite. Nous avons alors en réintroduisant la viscosité...:

$$\rho_\infty d_t \vec{v} = -\vec{\nabla} p + \rho_\infty \varepsilon \bar{T} g \vec{e}_z + \mu_d \vec{\nabla}^2 \vec{v} \quad (69.158)$$

et nous avons la dérivée particulaire (cf. chapitre de Mécanique des Milieux Continus):

$$d_t \vec{v} = (\vec{v} \circ \vec{\nabla}) \vec{v} + \partial_t \vec{v} \quad (69.159)$$

soit aussi une autre relation utile:

$$d_t = \vec{v} \circ \vec{\nabla} + \partial_t \quad (69.160)$$

Nous avons alors comme expression de la densité de force:

$$\rho_\infty (\vec{v} \circ \vec{\nabla}) \vec{v} + \partial_t \vec{v} = -\vec{\nabla} p + \rho_\infty \varepsilon \bar{T} g \vec{e}_z + \mu_d \vec{\nabla}^2 \vec{v} \quad (69.161)$$

Pour continuer, nous allons chercher à déterminer la loi d'énergie de l'équation de comportement démontrée dans le chapitre de Mécanique Des Milieux Continus:

$$\sigma_{ij} = -p \delta_{ij} - \frac{2}{3} \mu_d (\vec{\nabla} \circ \vec{v}) \delta_{ij} + 2\mu_d \theta_{ij} \quad (69.162)$$

pour qu'elle rende également compte de la relation entre les contraintes et les caractéristiques thermodynamiques du fluide, comme le flux de chaleur et la température. Nous allons le faire en caractérisant la diffusion de l'énergie dans le milieu due aux effets (supposés découplés) de la viscosité du fluide et de la conduction thermique du fluide.

Nous réécrivons cette relation avec de nouvelles constantes et une autre notation pour la divergence:

$$\sigma_{ij} = -p\delta_{ij} + \lambda \operatorname{div}(\vec{v})\delta_{ij} + 2\mu_d\theta_{ij} \quad (69.163)$$

où μ_d, λ sont dans ce contexte appelés les "**coefficients de Lamé**".

Nous avons aussi démontré dans le chapitre de Mécanique des Milieux Continus la relation:

$$\rho \partial_t \vec{v} = \vec{\nabla}(\vec{\sigma}) + \rho \vec{f} \quad (69.164)$$

Soit:

$$\rho \left((\vec{v} \circ \vec{\nabla}) \vec{v} + \partial_t \vec{v} \right) = \vec{\nabla}(\vec{\sigma}) + \rho \vec{f} \quad (69.165)$$

Ce qui donne:

$$\rho \partial_t \vec{v} + \rho (\vec{v} \circ \vec{\nabla}) \vec{v} = \vec{\nabla}(\vec{\sigma}) + \rho \vec{g} \quad (69.166)$$

Notons l'énergie totale comme:

$$E_{\text{tot}} = E_c + U = \int_D \frac{1}{2} \rho \vec{v}^2 dV + \int_D \rho e dV = \rho \int_D \left(\frac{1}{2} \vec{v}^2 + e \right) dV \quad (69.167)$$

où e est l'énergie interne massique du fluide (rapportée donc à une unité de masse de fluide). Or la variation instantanée d'énergie interne du fluide est égale à l'apport d'une puissance mécanique et de l'apport de chaleur (selon ce qui a été vu dans le chapitre de Thermodynamique):

$$\frac{d}{dt} \rho \int_D \left(\frac{1}{2} \vec{v}^2 + e \right) dV = \frac{dW}{dt} + \frac{dQ}{dt} = P + \dot{Q} \quad (69.168)$$

où P donne la puissance des efforts extérieurs au système donnée forcément par la force du champ de potentiel environnant et des forces mécaniques données par le tenseur des contraintes uniquement (nous sommes toujours dans la situation d'un fluide parfait...). Soit:

$$P = \int_D \rho \vec{g} \circ \vec{v} dV + \int_D (\sigma \vec{v}) d\vec{S} \quad (69.169)$$

et en utilisant le théorème d'Ostrogradsky (cf. chapitre de Calcul Vectoriel):

$$\begin{aligned} P &= \int_D \rho (\vec{g} \circ \vec{v}) dV + \int_D (\vec{\nabla} \circ \sigma \vec{v}) dV = \int_D \rho (\vec{g} \circ \vec{v}) dV + \int_D \operatorname{div}(\sigma \vec{v}) dV \\ &= \int_D [\rho \vec{g} \circ \vec{v} + \operatorname{div}(\sigma \vec{v})] dV = \int_D \left[\rho g_i v_i + \frac{\partial (\sigma_{ij} v_j)}{\partial x_i} \right] dV \end{aligned} \quad (69.170)$$

ce qui a bien les unités d'une puissance et nous avons bien:

$$\begin{aligned}
\operatorname{div}(\sigma \vec{v}) &= \vec{\nabla} \circ \left(\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \right) = \vec{\nabla} \circ \left(\begin{bmatrix} \sigma_{11}v_1 + \sigma_{12}v_2 + \sigma_{13}v_3 \\ \sigma_{21}v_1 + \sigma_{22}v_2 + \sigma_{23}v_3 \\ \sigma_{31}v_1 + \sigma_{32}v_2 + \sigma_{33}v_3 \end{bmatrix} \right) \\
&= \frac{\partial(\sigma_{11}v_1 + \sigma_{12}v_2 + \sigma_{13}v_3)}{\partial x_1} + \frac{\partial(\sigma_{21}v_1 + \sigma_{22}v_2 + \sigma_{23}v_3)}{\partial x_2} + \frac{\partial(\sigma_{31}v_1 + \sigma_{32}v_2 + \sigma_{33}v_3)}{\partial x_3} \quad (69.171) \\
&= \frac{\partial(\sigma_{ij}v_j)}{\partial x_i}
\end{aligned}$$

Pour la puissance chaleur $\dot{Q}$ c'est très facile aussi grâce aux développements que nous avons fait dans le chapitre de Thermodynamique où nous avons obtenu l'équation de la chaleur:

$$\frac{\partial q}{\partial t} + \vec{\nabla} \circ \vec{\Phi}_q = \dot{q} + \vec{\nabla} \circ \vec{\Phi}_q = 0 \quad (69.172)$$

Soit:

$$\dot{Q} = \int_D \dot{q} dV = - \int_D \vec{\nabla} \circ \vec{\Phi}_q dV \quad (69.173)$$

Nous avons finalement:

$$\begin{aligned}
\frac{d}{dt} \rho \int_D \left(\frac{1}{2} \vec{v}^2 + e \right) dV &= \int_D [\rho g \vec{v} + \operatorname{div}(\sigma \vec{v})] dV - \int_D \vec{\nabla} \circ \vec{\Phi}_q dV \\
&= \int_D [\rho g \vec{v} + \operatorname{div}(\sigma \vec{v}) - \vec{\nabla} \circ \vec{\Phi}_q] dV \quad (69.174)
\end{aligned}$$

Donc tout cela nous donne alors l'équation de l'énergie d'un fluide:

$$\rho \frac{d}{dt} \left(\frac{1}{2} \vec{v}^2 + e \right) = \rho g \vec{v} + \operatorname{div}(\sigma \vec{v}) - \vec{\nabla} \circ \vec{\Phi}_q \quad (69.175)$$

Soit:

$$\rho \frac{de}{dt} + \frac{1}{2} \rho \frac{d\vec{v}^2}{dt} = \rho g \vec{v} + \operatorname{div}(\sigma \vec{v}) - \vec{\nabla} \circ \vec{\Phi}_q \quad (69.176)$$

et comme (cf. chapitre de Thermodynamique) le flux de chaleur suit la loi de Fourier:

$$\vec{\Phi}_q = -\kappa \cdot \vec{\nabla} \cdot T \quad (69.177)$$

Nous avons alors:

$$\rho \frac{de}{dt} + \frac{1}{2} \rho \frac{d\vec{v}^2}{dt} = \rho g \vec{v} + \operatorname{div}(\sigma \vec{v}) - \vec{\nabla} \circ (-\kappa \vec{\nabla} T) \quad (69.178)$$

Soit en utilisant la définition du laplacien d'un champ scalaire:

$$\begin{aligned}
 \rho \frac{de}{dt} + \frac{1}{2} \rho \frac{d\vec{v}^2}{dt} &= \rho g \vec{v} + \operatorname{div}(\sigma \vec{v}) + \kappa \vec{\nabla}^2 T \\
 \rho \frac{de}{dt} + \frac{1}{2} \rho \frac{d(v_i v_i)}{dt} &= \rho g_i v_i + \frac{\partial \sigma_{ij} v_j}{\partial x_i} + \kappa \frac{\partial T}{\partial^2 x_i}
 \end{aligned} \quad [1]$$

(69.179)

Maintenant, en faisant le produit scalaire de:

$$\rho d_t \vec{v} = \vec{\nabla}(\vec{\sigma}) + \rho \vec{g} \quad (69.180)$$

avec la vitesse $\vec{v}$ nous obtenons le bilan de l'énergie cinétique:

$$\begin{aligned}
 \rho \vec{v} \circ d_t \vec{v} &= \vec{v} \circ \vec{\nabla}(\vec{\sigma}) + \rho \vec{v} \circ \vec{g} \\
 \rho v_i d_t v_i &= \rho d_t \left(\frac{v_i v_i}{2} \right) = \frac{1}{2} \rho \frac{d(v_i v_i)}{dt} = v_i \frac{\partial \sigma_{ij}}{\partial x_i} + \rho v_i g_i
 \end{aligned} \quad [2]$$

(69.181)

En soustrayant [2] de [1], nous obtenons une relation locale de l'énergie interne spécifique e :

$$\begin{aligned}
 \rho \frac{de}{dt} &= -\vec{v} \circ \vec{\nabla}(\vec{\sigma}) + \operatorname{div}(\sigma \vec{v}) + \kappa \vec{\nabla}^2 T \\
 \rho \frac{de}{dt} &= -v_i \frac{\partial \sigma_{ij}}{\partial x_j} + \frac{\partial(\sigma_{ij} v_j)}{\partial x_i} + \kappa \frac{\partial T}{\partial^2 x_j}
 \end{aligned} \quad (69.182)$$

Or, nous avons aussi (dérivation d'un produit):

$$\frac{\partial(\sigma_{ij} v_j)}{\partial x_i} = \sigma_{ij} \frac{\partial v_j}{\partial x_i} + v_i \frac{\partial \sigma_{ij}}{\partial x_j} \quad (69.183)$$

Soit:

$$\frac{\partial(\sigma_{ij} v_j)}{\partial x_i} - v_i \frac{\partial \sigma_{ij}}{\partial x_j} = \sigma_{ij} \frac{\partial v_j}{\partial x_i} \quad (69.184)$$

Effectivement:

$$\begin{aligned}
\frac{\partial(\sigma_{ij}v_j)}{\partial x_i} &= \frac{\partial(\sigma_{11}v_1 + \sigma_{12}v_2 + \sigma_{13}v_3)}{\partial x_1} + \frac{\partial(\sigma_{21}v_1 + \sigma_{22}v_2 + \sigma_{23}v_3)}{\partial x_2} + \frac{\partial(\sigma_{31}v_1 + \sigma_{32}v_2 + \sigma_{33}v_3)}{\partial x_3} \\
&= v_1 \frac{\partial \sigma_{11}}{\partial x_1} + \sigma_{11} \frac{\partial v_1}{\partial x_1} + v_2 \frac{\partial \sigma_{12}}{\partial x_1} + \sigma_{12} \frac{\partial v_2}{\partial x_1} + v_3 \frac{\partial \sigma_{13}}{\partial x_1} + \sigma_{13} \frac{\partial v_3}{\partial x_1} \\
&+ v_1 \frac{\partial \sigma_{21}}{\partial x_2} + \sigma_{21} \frac{\partial v_1}{\partial x_2} + v_2 \frac{\partial \sigma_{22}}{\partial x_2} + \sigma_{22} \frac{\partial v_2}{\partial x_2} + v_3 \frac{\partial \sigma_{23}}{\partial x_2} + \sigma_{23} \frac{\partial v_3}{\partial x_2} \\
&+ v_1 \frac{\partial \sigma_{31}}{\partial x_3} + \sigma_{31} \frac{\partial v_1}{\partial x_3} + v_2 \frac{\partial \sigma_{32}}{\partial x_3} + \sigma_{32} \frac{\partial v_2}{\partial x_3} + v_3 \frac{\partial \sigma_{33}}{\partial x_3} + \sigma_{33} \frac{\partial v_3}{\partial x_3}
\end{aligned} \quad (69.185)$$

Nous avons donc:

$$\rho \frac{de}{dt} = \sigma_{ij} \frac{\partial v_i}{\partial x_j} + \kappa \frac{\partial T}{\partial^2 x_j} \quad (69.186)$$

et comme le tenseur σ est symétrique:

$$\sigma_{ij} \frac{\partial v_i}{\partial x_j} = \frac{1}{2} \sigma_{ij} \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) = \sigma_{ij} \frac{1}{2} \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) = \sigma_{ij} d_{ij} \quad (69.187)$$

Nous avons donc:

$$\rho \frac{de}{dt} = \sigma_{ij} d_{ij} + \kappa \frac{\partial T}{\partial^2 x_j} \quad (69.188)$$

Ce qui est parfois noté:

$$\rho \frac{de}{dt} = \sigma : D + \kappa \nabla^2 T \quad (69.189)$$

où $\overline{\overline{D}}$ est appelé "**tenseur des taux de déformation**" et $\overline{\overline{\sigma}} : \overline{\overline{D}}$ représente le produit doublement contracté du tenseur des contraintes et du tenseur des taux de déformation.

Nous avons montré dans le chapitre de Mécanique des Milieux Continus:

$$\sigma = -p \delta_{ij} - \frac{2}{3} \mu_d (\vec{\nabla} \circ \vec{v}) \delta_{ij} + 2\mu_d \theta_{ij} = -p \delta_{ij} + 2\mu_d \left(\theta_{ij} - \frac{1}{3} (\vec{\nabla} \circ \vec{v}) \delta_{ij} \right) \quad (69.190)$$

où:

$$\tau = 2\mu_d \left(\theta_{ij} - \frac{1}{3} (\vec{\nabla} \circ \vec{v}) \delta_{ij} \right) \quad (69.191)$$

Ainsi, il est simple de différencier forces normales et forces tangentielles. Bref pour en revenir à l'équation de l'énergie:

$$\rho \frac{de}{dt} = \sigma : D + \kappa \nabla^2 T \quad (69.192)$$

Nous avons donc:

$$\rho \frac{de}{dt} = -p \delta_{ij} D + 2\mu_d \left(\theta_{ij} - \frac{1}{3} (\vec{\nabla} \circ \vec{v}) \delta_{ij} \right) D + \kappa \nabla^2 T \quad (69.193)$$

Mais dans notre cas:

$$\theta_{ij} = D \quad (69.194)$$

Soit:

$$\rho \frac{de}{dt} = -p \delta_{ij} d_{ij} + 2\mu_d \left(d_{ij} - \frac{1}{3} (\vec{\nabla} \circ \vec{v}) \delta_{ij} \right) d_{ij} + \kappa \nabla^2 T \quad (69.195)$$

Mais nous avons:

$$\delta_{ij} d_{ij} = d_{ii} = \text{Tr}(D) = \frac{1}{2} \left(\frac{\partial v_i}{\partial x_i} + \frac{\partial v_i}{\partial x_i} \right) = \frac{\partial v_i}{\partial x_i} = \vec{\nabla} \circ \vec{v} \quad (69.196)$$

Nous avons donc:

$$\rho \frac{de}{dt} = -p (\vec{\nabla} \circ \vec{v}) + 2\mu_d d_{ij} d_{ij} + \frac{1}{3} \mu_d (\vec{\nabla} \circ \vec{v})^2 + \kappa \nabla^2 T \quad (69.197)$$

Soit sous forme technique et condensée:

$$\rho \frac{de}{dt} = -p (\vec{\nabla} \circ \vec{v}) + 2\mu_d DD + \lambda (\vec{\nabla} \circ \vec{v})^2 + \kappa \nabla^2 T \quad (69.198)$$

Il est clair qu'au niveau de l'entropie, nous avons:

$$S = \int_D \rho s dV \quad (69.199)$$

Nous avons aussi:

$$dU = TdS - pdV \Rightarrow dS = \frac{dU}{T} + \frac{pdV}{T} \quad (69.200)$$

Soit réduit au rapport massique:

$$ds = \frac{de}{T} + \frac{pdV}{T} \Rightarrow \rho T ds = \rho de + \rho p_m dV \quad (69.201)$$

La variation temporelle donnant:

$$\rho T \frac{ds}{dt} = \rho \frac{de}{dt} + p_m \frac{dm}{dt} \quad (69.202)$$

Or, nous avons l'équation de continuité (cf. chapitre de Thermodynamique):

$$\frac{dm}{dt} - \vec{\nabla} \circ (m\vec{v}) = 0 \Rightarrow \frac{dm}{dt} = \vec{\nabla} \circ (m\vec{v}) \quad (69.203)$$

Ce qui nous donne finalement:

$$\rho T \frac{ds}{dt} = \rho \frac{de}{dt} + p \vec{\nabla} \circ \vec{v} \quad (69.204)$$

ou autrement écrit:

$$\rho \frac{de}{dt} = \rho T \frac{ds}{dt} - p \vec{\nabla} \circ \vec{v} \quad (69.205)$$

Injectée dans:

$$\rho \frac{de}{dt} = -p (\vec{\nabla} \circ \vec{v}) + 2\mu_d d_{ij} d_{ij} + \lambda (\vec{\nabla} \circ \vec{v})^2 + \kappa \nabla^2 T \quad (69.206)$$

Cela donne:

$$\rho T \frac{ds}{dt} = 2\mu_d d_{ij} d_{ij} + \lambda (\vec{\nabla} \circ \vec{v})^2 + \kappa \nabla^2 T \quad (69.207)$$

Si nous considérons le gradient de vitesse comme étant très faible (quasi-statique) nous pouvons alors écrire l'approximation:

$$\rho T \frac{ds}{dt} \cong 2\mu_d (d_{ij})^2 + \kappa \nabla^2 T \quad (69.208)$$

Maintenant donnons l'expression de l'entropie (différentielle totale exacte) en fonction des paramètres de température et de pression seuls:

$$dS = \left(\frac{\partial S}{\partial T} \right)_P dT + \left(\frac{\partial S}{\partial P} \right)_T dP \quad (69.209)$$

soit sous forme massique:

$$ds = \left(\frac{\partial s}{\partial T} \right)_P dT + \left(\frac{\partial s}{\partial P} \right)_T dP \quad (69.210)$$

Or nous avons démontré dans le chapitre de Thermodynamique la relation suivante:

$$T \left(\frac{\partial S}{\partial T} \right)_P = C_P = \left(\frac{\partial H}{\partial T} \right)_P \quad (69.211)$$

soit sous forme massique:

$$T \left(\frac{\partial s}{\partial T} \right)_P = c_P = \left(\frac{\partial h}{\partial T} \right)_P \quad (69.212)$$

Ce qui nous donne:

$$ds = \frac{c_P}{T} dT + \left(\frac{\partial s}{\partial P} \right)_T dP \quad (69.213)$$

Or, nous avons démontré avec dans le chapitre de Thermodynamique une des relations de Maxwell:

$$\left(\frac{\partial V}{\partial T} \right)_P = - \left(\frac{\partial S}{\partial P} \right)_T \quad (69.214)$$

soit sous forme massique:

$$\left(\frac{\partial \rho^{-1}}{\partial T} \right)_P = - \left(\frac{\partial s}{\partial P} \right)_T \quad (69.215)$$

d'où:

$$ds = \frac{c_P}{T} dT - \left(\frac{\partial \rho^{-1}}{\partial T} \right)_P dP \quad (69.216)$$

Soit:

$$\rho T ds = \rho c_P dT - \rho T \left(\frac{\partial \rho^{-1}}{\partial T} \right)_P dP = \rho c_P dT - \frac{T}{\rho} \left(\frac{\partial \rho}{\partial T} \right)_P dP \quad (69.217)$$

Soit notre relation:

$$\rho T \frac{ds}{dt} \cong 2\mu_d (d_{ij})^2 + \kappa \nabla^2 T \quad (69.218)$$

peut alors s'écrire:

$$\rho c_P \frac{dT}{dt} - \frac{T}{\rho} \left(\frac{\partial \rho}{\partial T} \right)_P \frac{dP}{dt} \cong 2\mu_d (d_{ij})^2 + \kappa \nabla^2 T \quad (69.219)$$

Si nous admettons que la variation de la densité avec la température est faible, nous avons alors à l'échelle atmosphérique :

$$\rho_\infty c_P \frac{dT}{dt} \cong 2\mu_d (d_{ij})^2 + \kappa \nabla^2 T \quad (69.220)$$

et en se rappelant que:

$$d_t = \vec{v} \circ \vec{\nabla} + \partial_t \quad (69.221)$$

Il vient finalement:

$$\rho_{\infty} c_P (\vec{v} \circ \vec{\nabla} + \partial_t) T \cong 2\mu_d (d_{\vec{v}})^2 + \kappa \nabla^2 T \quad (69.222)$$

Nous avons maintenant deux équations importantes:

$$\begin{aligned} \rho_{\infty} ((\vec{v} \circ \vec{\nabla}) \vec{v} + \partial_t \vec{v}) &= -\vec{\nabla} p + \rho_{\infty} \varepsilon \bar{T} \vec{g} + \mu_d \vec{\nabla}^2 \vec{v} \\ \rho_{\infty} c_P (\vec{v} \circ \vec{\nabla} + \partial_t) T &\cong 2\mu_d (d_{\vec{v}})^2 + \kappa \nabla^2 T \end{aligned} \quad (69.223)$$

Soit:

$$\begin{aligned} (\vec{v} \circ \vec{\nabla}) \vec{v} + \partial_t \vec{v} &= -\frac{1}{\rho_{\infty}} \vec{\nabla} p + \varepsilon \bar{T} \vec{g} + \frac{\mu_d}{\rho_{\infty}} \vec{\nabla}^2 \vec{v} \\ \rho_{\infty} c_P (\vec{v} \circ \vec{\nabla} + \partial_t) T &\cong 2\mu_d (d_{\vec{v}})^2 + \kappa \nabla^2 T \end{aligned} \quad (69.224)$$

Examinons maintenant rapidement le problème de Rayleigh-Bénard qui consiste en deux plaques limitant un fluide une étant plus chauffée que l'autre.

Nous pouvons alors observer des rouleaux longitudinaux parallèles dans un film de fluide visqueux (huile de silicone) maintenus entre deux plaques à une température chaude en bas et froide en haut. Voici une photo de ces rouleaux vus de côtés:

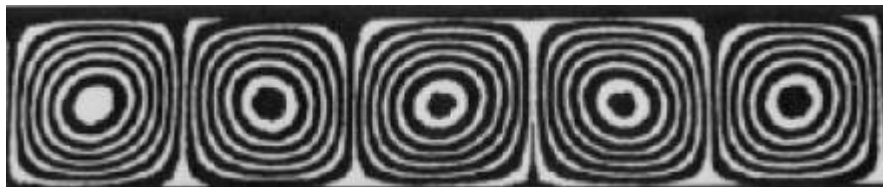


Figure: 69.13 - Rouleaux de convection de Rayleigh-Bénard (instabilité)

vue de dessus:

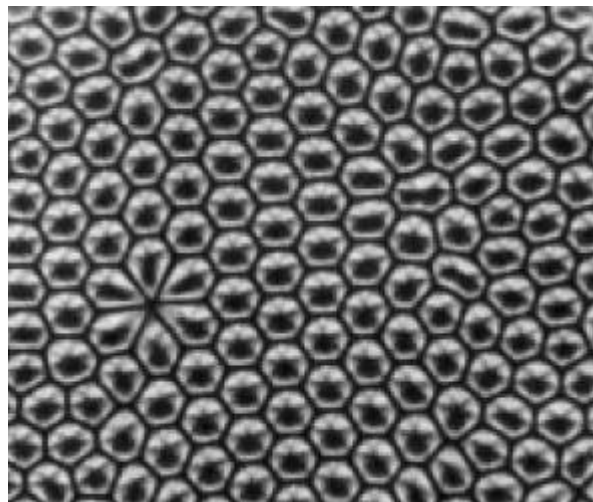


Figure: 69.14 - Instabilités de Rayleigh-Bénard vues du dessus

Il s'agit d'un problème de convection naturelle: le fluide chauffé en bas se dilate et remonte entraînée par la force d'Archimède, arrivé en haut il se refroidit et retombe. C'est ce mouvement qu'il faut expliquer qui est similaire à celui de l'atmosphère terrestre

Nous remarquons également que les mouvements de convection se font approximativement selon un tore (voir la photo vue de côté). Nous pouvons tirer parti de cette symétrie pour simplifier l'analyse.

Considérons donc une boucle verticale de fluide circulant à vitesse constante (donc sans trop de turbulences...):

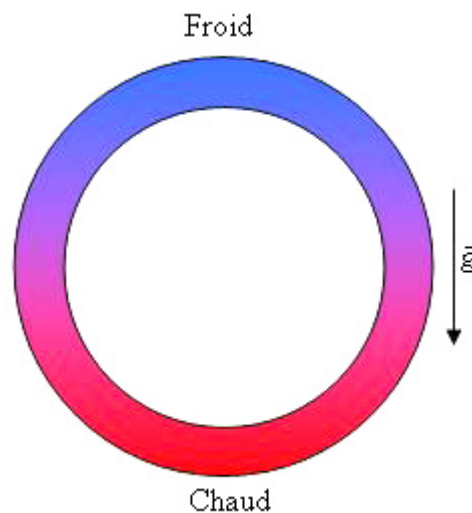


Figure: 69.15 - Boucle verticale de fluide avec gradient

La configuration sera imposée comme étant la suivante:

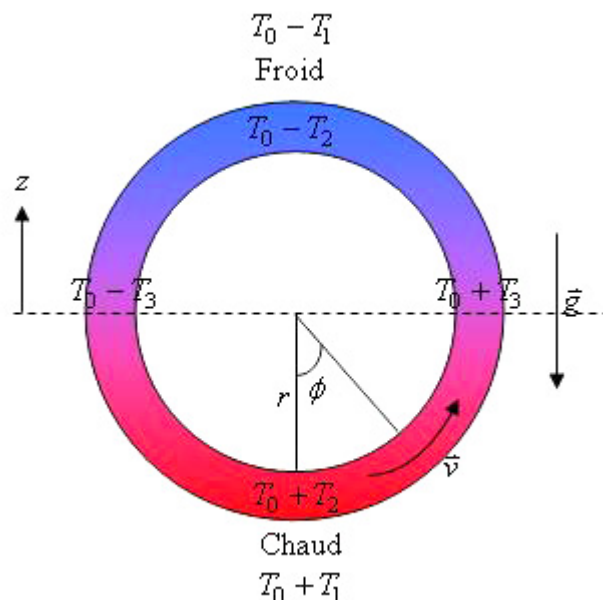


Figure: 69.16 - Configuration imposée pour le modèle théorique

où T_0 est la température moyenne du fluide (attention: ne pas oublier que ce n'est pas une grandeur extensive!) et où nous avons indiqué respectivement les températures à l'intérieur du tore et à l'extérieur (soit de l'environnement) qui peuvent toutes varier en fonction du temps.

Nous voyons que la différence de température est de $2T_2$ entre le haut et le bas et de $2T_3$ entre la droite et la gauche.

Nous posons que la température varie linéairement avec la hauteur (ce qui bien évidemment est faux dans un modèle atmosphérique...):

$$T_E = T_0 - T_1 \frac{z}{r} = T_0 + T_1 \cos \phi \quad (69.225)$$

Nous remarquons qu'il possible de paramétrer la température le long de l'intérieur du tore avec la relation suivante (équation paramétrique du cercle):

$$\Delta T = T - T_0 = T_2 \cos \phi + T_3 \sin \phi \quad (69.226)$$

Nous avons alors conformément au schéma:

$$\begin{aligned} T(0) &= T_2 \cos 0 + T_3 \sin 0 = T_2 + T_0 \\ T(\pi) &= T_2 \cos \pi + T_3 \sin \pi = -T_2 + T_0 \quad (69.227) \\ &\dots \end{aligned}$$

Ceci étant posé, revenons à:

$$(\vec{v} \circ \vec{\nabla}) \vec{v} + \partial_t \vec{v} = -\frac{1}{\rho_\infty} \vec{\nabla} p + \varepsilon \vec{T} \vec{g} + \frac{\mu_d}{\rho_\infty} \vec{\nabla}^2 \vec{v} \quad (69.228)$$

Nous allons passer ce système en coordonnées polaire correspondant le mieux à la géométrie de notre problème. Rappelons d'abord que dans terme:

$$(\vec{v} \circ \vec{\nabla}) \vec{v}' \quad (69.229)$$

l'opérateur différentiel $\vec{\nabla}$ est la divergence. Or, nous avons démontré dans le chapitre de Calcul Vectoriel que celui-ci s'écrivait alors en coordonnées polaires:

$$\vec{\nabla} = \left(\frac{1}{r} \frac{\partial}{\partial r}, \frac{1}{r} \frac{\partial}{\partial \phi} \right) \quad (69.230)$$

Or, si nous nous basons sur l'hypothèse que dans le volume du tore, la vitesse ne varie ni en fonction de l'angle, ni à l'intérieur du tore (donc ne varie pas selon le rayon r) alors en coordonnées polaires:

$$(\vec{v} \circ \vec{\nabla}) \vec{v}' = 0 \quad (69.231)$$

Nous avons alors:

$$\partial_t \vec{v} = -\frac{1}{\rho_\infty} \vec{\nabla} p + \varepsilon \vec{T} \vec{g} + \frac{\mu_d}{\rho_\infty} \vec{\nabla}^2 \vec{v} \quad (69.232)$$

Nous allons réduire l'analyse à une seule dimension qui sera celle comme quoi le phénomène ne dépend que de l'angle. Nous avons alors en coordonnées polaires et en explicitant tous les termes:

$$\partial_t v = -\frac{1}{\rho_\infty r} \frac{\partial p}{\partial \phi} + \varepsilon g (T - T_0) \sin \phi - \frac{\mu_d}{\rho_\infty} \frac{1}{r^2} \frac{\partial^2}{\partial \phi^2} v \quad (69.233)$$

où nous avons fait par la même occasion la projection selon l'axe z tel que:

$$\vec{g} = g \vec{e}_z = g \sin \phi \quad (69.234)$$

Le coefficient différentiel du dernier terme va nous embêter. Nous le remplaçons par un coefficient que nous supposons constant et qui s'oppose au mouvement tel que nous ayons:

$$\partial_t v = -\frac{1}{\rho_\infty r} \frac{\partial p}{\partial \phi} + g \varepsilon (T - T_0) \sin \phi - \Gamma v \quad (69.235)$$

Ou plus explicitement:

$$\partial_t v = -\frac{1}{\rho_\infty r} \frac{\partial p}{\partial \phi} + g \varepsilon (T_2 \cos \phi + T_3 \sin \phi) \sin \phi - \Gamma v \quad (69.236)$$

Nous intégrons maintenant l'ensemble sur l'entier de la boucle en fonction de ϕ . Nous avons alors:

$$\int_0^{2\pi} \partial_t v' d\phi = -\frac{1}{\rho_\infty r} \underbrace{\int_0^{2\pi} \frac{\partial p}{\partial \phi} d\phi}_{=0} + g \varepsilon \int_0^{2\pi} (T_2 \cos \phi \sin \phi + T_3 \sin \phi) d\phi - \Gamma \int_0^{2\pi} v d\phi \quad (69.237)$$

Nous avons alors le terme pression qui disparaît, car il n'y a pas de gradient de pression au long de la boucle. Ainsi:

$$2\pi \partial_t v = \varepsilon g \int_0^{2\pi} (T_2 \cos \phi \sin \phi + T_3 \sin \phi) d\phi - 2\pi \Gamma v \quad (69.238)$$

Nous avons ensuite (cf. chapitre de Calcul Différentiel Et Intégral):

$$\int_0^{2\pi} \cos \phi \sin \phi d\phi = \frac{1}{2} \sin^2 \phi \Big|_0^{2\pi} = 0 \quad (69.239)$$

et:

$$\int_0^{2\pi} \sin^2 \phi d\phi = \pi \quad (69.240)$$

Il nous reste donc:

$$2\pi \partial_t v = \varepsilon g \pi T_3 - 2\pi \Gamma v \quad (69.241)$$

Soit:

$$\partial_t v = \varepsilon g \frac{T_3}{2} - \Gamma v \quad (69.242)$$

Nous voyons dans cette équation que le mouvement est piloté par la différence de température horizontale T_3 .

Maintenant, revenons sur:

$$\partial_t T + \vec{v} \circ \vec{\nabla} T \cong \frac{2\mu_d}{\rho_\infty c_p} (d_{\vec{v}})^2 + \frac{\kappa}{\rho_\infty c_p} \nabla^2 T \quad (69.243)$$

Si nous négligeons les forces tangentielles à l'intérieur de fluide, nous avons alors:

$$\partial_t T + \vec{v} \circ \vec{\nabla} T \cong D \nabla^2 T \quad (69.244)$$

où D est le coefficient de diffusion thermique (cf. chapitre de Thermodynamique).

En coordonnées polaires cela se réduit à:

$$\partial_t T + v \frac{1}{r} \frac{\partial}{\partial \phi} T = D \frac{1}{r} \frac{\partial^2}{\partial \phi^2} T \quad (69.245)$$

et nous allons aussi faire une autre approximation:

$$\partial_t T + \frac{v}{r} \frac{\partial T}{\partial \phi} \cong K T \quad (69.246)$$

Et nous avons les deux relations:

$$\begin{aligned} T_E &= T_0 + T_1 \cos \phi \\ T - T_0 &= T_2 \cos \phi + T_3 \sin \phi \end{aligned} \quad (69.247)$$

En soustrayant:

$$T_E - T = (T_1 - T_2) \cos \phi - T_3 \sin \phi \quad (69.248)$$

Soit:

$$\partial_t T + \frac{v}{r} \frac{\partial T}{\partial \phi} \cong K [(T_1 - T_2) \cos \phi - T_3 \sin \phi] \quad (69.249)$$

et encore:

$$\partial_t (T_0 + T_2 \cos \phi + T_3 \sin \phi) + \frac{v}{r} \frac{\partial}{\partial \phi} (T_0 + T_2 \cos \phi + T_3 \sin \phi) \cong K (T_1 - T_2) \cos \phi - K T_3 \sin \phi \quad (69.250)$$

Soit:

$$\frac{dT_2}{dt} \cos \phi + \frac{dT_3}{dt} \sin \phi + \frac{v}{r} T_2 \frac{\partial \cos \phi}{\partial \phi} + \frac{v}{r} T_3 \frac{\partial \sin \phi}{\partial \phi} \cong K(T_1 - T_2) \cos \phi - KT_3 \sin \phi \quad (69.251)$$

Après dérivation:

$$\frac{dT_2}{dt} \cos \phi + \frac{dT_3}{dt} \sin \phi - \frac{v}{r} T_2 \sin \phi + \frac{v}{r} T_3 \cos \phi \cong K(T_1 - T_2) \cos \phi - KT_3 \sin \phi \quad (69.252)$$

Nous regroupons les termes:

$$\left[\frac{dT_2}{dt} + \frac{vT_3}{r} \right] \cos \phi + \left[\frac{dT_3}{dt} - \frac{vT_2}{r} \right] \sin \phi \cong K(T_1 - T_2) \cos \phi - KT_3 \sin \phi \quad (69.253)$$

Nous avons alors les trois équations différentielles suivantes qui gouvernent la dynamique du système:

$$\begin{aligned} \frac{dT_2}{dt} + \frac{vT_3}{r} &\cong K(T_1 - T_2) \\ \frac{dT_3}{dt} - \frac{vT_2}{r} &= -KT_3 \\ \frac{dv}{dt} &= \varepsilon g \frac{T_3}{2} - \Gamma v \end{aligned} \quad (69.254)$$

Nous terminons les multiples simplifications en posant...:

$$T_4(t) = T_1 - T_2(t)$$

Ce qui nous donne:

$$\begin{aligned} \frac{dT_2}{dt} &= \frac{d(T_1 - T_4(t))}{dt} = -\frac{dT_4}{dt} \cong KT_4 - \frac{vT_3}{r} \\ \frac{dT_3}{dt} &= -KT_3 + \frac{vT_2}{r} = -KT_3 + \frac{vT_1}{r} - \frac{vT_4}{r} \\ \frac{dv}{dt} &= \varepsilon g \frac{T_3}{2} - \Gamma v \end{aligned} \quad (69.255)$$

En remettant cela au propre:

$$\boxed{\begin{aligned} \frac{dv}{dt} &= -\Gamma v + \varepsilon g \frac{T_3}{2} \\ \frac{dT_3}{dt} &= -KT_3 + \frac{vT_1}{r} - \frac{vT_4}{r} \\ \frac{dT_4}{dt} &\cong -KT_4 + \frac{vT_3}{r} \end{aligned}} \quad (69.256)$$

Maintenant, introduisons les variables sans dimensions suivantes:

$$X = \frac{v}{rK} \quad Y = \frac{g\varepsilon T_3}{2r\Gamma K} \quad Z = \frac{g\varepsilon T_4}{2r\Gamma K} \quad t' = tK \quad (69.257)$$

où nous pouvons assimiler:

- X à la vitesse adimensionnelle
- Y à la différence adimensionnelle de température entre courants ascendants et descendants
- Z à la déviation adimensionnelle de l'équilibre de convection.

Nous avons alors effectivement:

$$\begin{aligned}\frac{d\left(\frac{v}{rK}\right)}{dtK} &= -\frac{\Gamma}{K} \frac{v}{rK} + \frac{\Gamma}{K} \frac{g\varepsilon T_3}{2r\Gamma K} \\ \frac{d\left(\frac{g\varepsilon T_3}{2r\Gamma K}\right)}{dtK} &= -\frac{g\varepsilon T_3}{2r\Gamma K} + \frac{g\varepsilon T_1}{2r\Gamma K} \frac{v}{rK} - \frac{g\varepsilon T_4}{2r\Gamma K} \frac{v}{rK} \quad (69.258) \\ \frac{d\left(\frac{g\varepsilon T_4}{2r\Gamma K}\right)}{dtK} &\cong -\frac{g\varepsilon T_4}{2r\Gamma K} + \frac{g\varepsilon T_3}{2r\Gamma K} \frac{v}{rK}\end{aligned}$$

Soit:

$$\begin{aligned}\frac{dX}{dt'} &= -\text{Pr } X + \text{Pr } Y \\ \frac{dY}{dt'} &= -Y + \text{Re } X - XZ \quad (69.259) \\ \frac{dZ}{dt'} &= -Z + YX\end{aligned}$$

De manière encore plus condensée et traditionnelle:

$$\begin{cases} X' = \text{Pr}(Y - X) \\ Y' = rX - Y - XZ \\ Z' = -Z + YX \end{cases} \quad (69.260)$$

où nous avons:

$$\text{Pr} = \frac{\Gamma}{K} \cong \frac{\frac{\mu_d}{\rho_\infty} \frac{1}{r} \frac{\partial^2}{\partial \phi^2}}{k \frac{1}{r} \frac{\partial^2}{\partial \phi^2}} = \frac{\mu_d}{\rho_\infty k} = \frac{\mu_d}{\rho_\infty \frac{\kappa}{\rho_\infty c_p}} = \frac{\mu_d c_p}{\kappa} \quad (69.261)$$

ce qui correspond au "**nombre de Prandtl**" et:

$$\text{Re} = \frac{g \varepsilon T_1}{2r \Gamma K} = \frac{g \varepsilon T_1}{2r \frac{\mu_d}{\rho_\infty} D \left(\frac{1}{r} \frac{\partial^2}{\partial \phi^2} \right)^2} = \frac{g \varepsilon \frac{T_1}{2} r}{\mu_c D \left(\frac{\partial^2}{\partial \phi^2} \right)^2} = \frac{g \varepsilon \frac{T_1}{2} L^3}{\mu_c D} \quad (69.262)$$

ce qui est assimilé au "**nombre de Rayleigh**".

Ce système de trois équations est essentiellement le même que celui du célèbre système de Lorenz. À une différence près, le système de Lorenz (réel) contient un facteur b dans la dernière équation (ce qui change peu de toute façon le résultat puisque l'on obtient quand même un attracteur étrange au bout du compte comme nous allons de suite le voir):

$$\begin{aligned} \frac{dX}{dt'} &= -\text{Pr } X + \text{Pr } Y \\ \frac{dY}{dt'} &= -Y + rX - XZ \\ \frac{dZ}{dt'} &= -bZ + YX \end{aligned} \quad (69.263)$$

Pr , Re et b sont strictement positifs, et on pose souvent $\text{Pr} = 10, r = 28, b = 8/3$ où le nombre de Prandtl correspond à la valeur de l'eau.

Les équations de Lorenz décrivent les phénomènes de convection d'un fluide idéal à deux dimensions, dans un réservoir chauffé par le bas.

Nous voyons par cette démonstration que contrairement aux dires non démontrés sur Internet que:

1. Le système n'est de loin pas simple mathématiquement et est très approximatif
2. Qu'il existe des systèmes vraiment plus simples et eux aussi chaotiques (cf. chapitre de Dynamique des Populations).

L'intérêt des équations de Lorenz réside cependant dans la sensibilité aux conditions initiales et à la convergence des variables adimensionnelles.

Voyons un exemple avec Maple 4.00b:

```
>with(DEtools):
>lorenz:=diff(x(t),t) = 10*(y(t)-x(t)),diff(y(t),t) = 28*x(t)-y(t)-x(t)*z(t),diff(z(t),t) = x(t)*y(t)-8/3*z(t);
>DEplot3d({lorenz}, [x(t),y(t),z(t)], t=0..100, stepsize=0.01, [[x(0)=10, y(0)=10, z(0)=10]], orientation=
[-35,75], linecolor = t, thickness = 1);
```

Ce qui donne pour les 100 premières unités de temps adimensionnel:

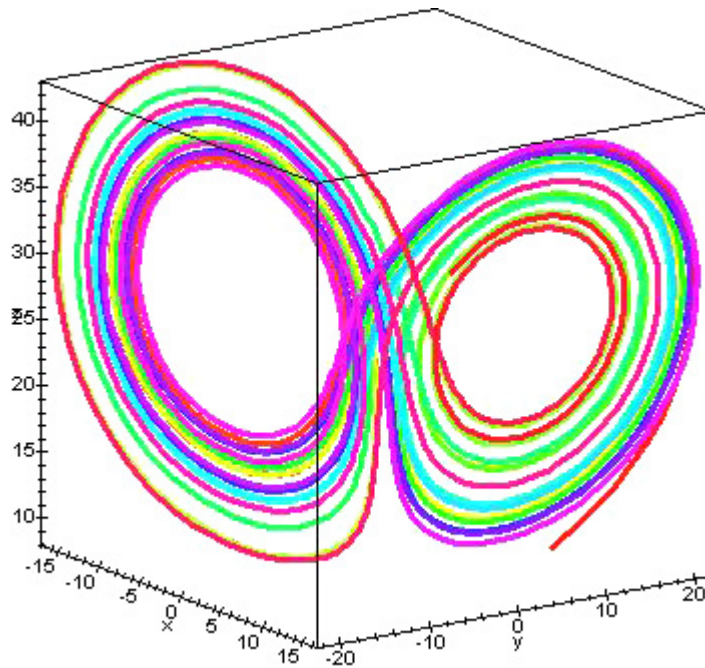


Figure: 69.17 - Espace des phases du système des équations de Lorenz

Ou pour les dix premières unités de temps adimensionnel:

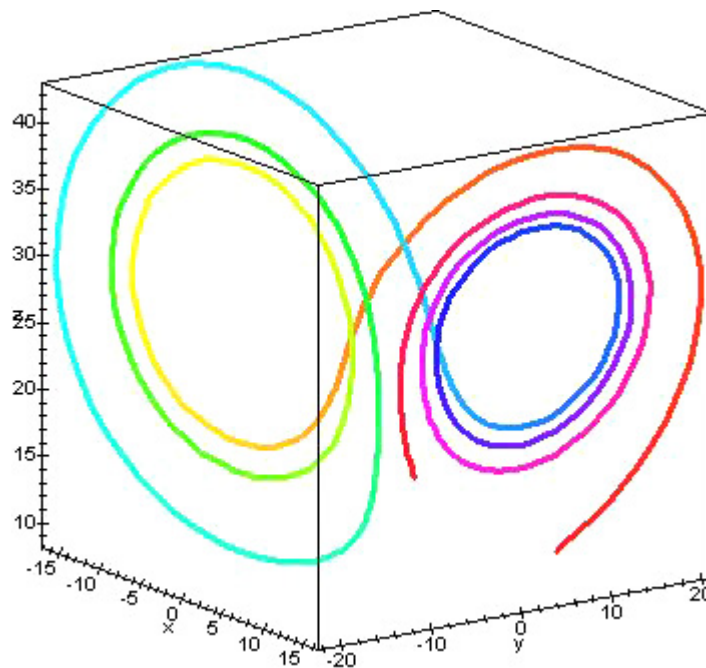


Figure: 69.18 - Espace des phases du système des équations de Lorenz

Bon jusque-là on s'en rend compte que les paramètres adimensionnels tournent autour de deux points que nous appelons les "**attracteurs étranges**".

Définition: Dans l'étude des systèmes dynamiques, un attracteur (ou ensemble-limite) est un ensemble, une courbe ou un espace vers lequel un système évolue de façon irréversible en l'absence de perturbations.

Maintenant toujours pour les mêmes valeurs du temps adimensionnel, nous prenons $x(0) = 10.1$, soit un changement relativement faible des conditions initiales. Nous avons alors:

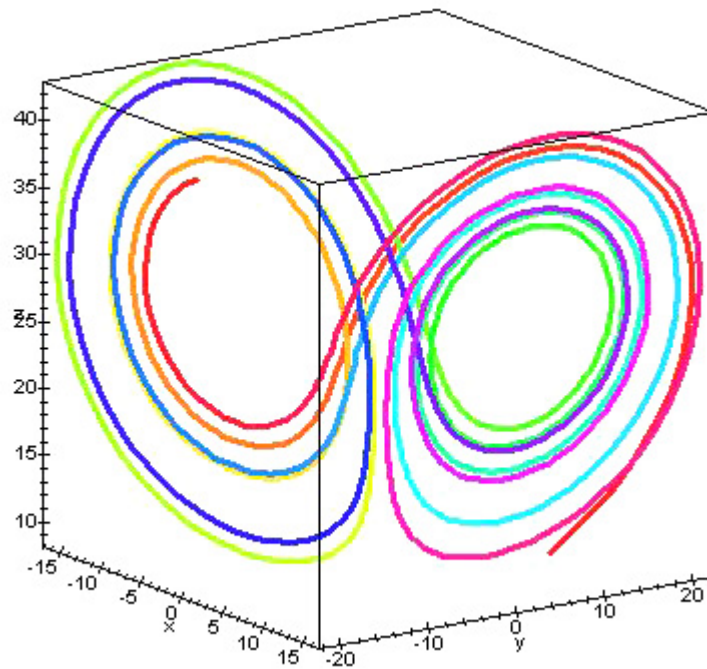


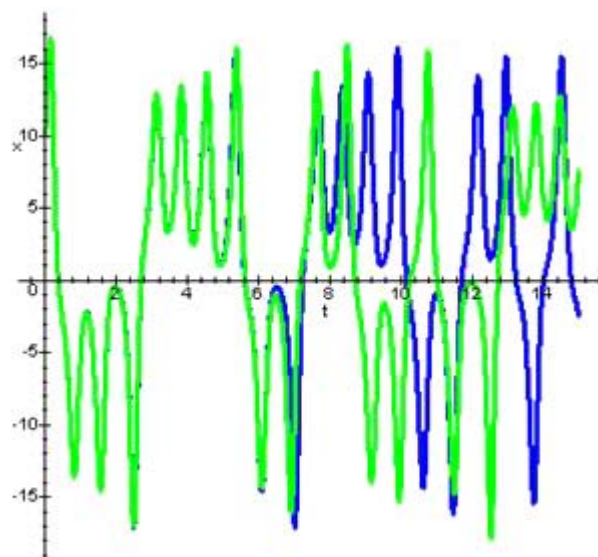
Figure: 69.19 - Petite variation dans les conditions initiales

Nous remarquons donc que le phénomène n'est plus vraiment semblable.

Considérons par exemple la variable x en prenant comme condition initiale $x(0) = 10, y(0) = 10, z(0) = 10$ puis $x(0) = 10, y(0) = 10.1, z(0) = 10$ soit une légère variation de 0.01 sur la valeur de $y(0)$.

Soit dans Maple 4.00b:

```
>DEplot({lorenz}, [x(t), y(t), z(t)], t=0..15, stepsize = 0.01, [[x(0)=10, y(0)=10, z(0)=10],[x(0)=10, y(0)=10.01, z(0)=10]], scene = [t,x], linecolor = [blue,green], thickness = 1);
```

Figure: 69.20 - Analyse de la variable adimensionnée x pour une faible variation des C.I.

Nous voyons que le système se décale assez rapidement du modèle initial alors qu'au début il reste identique mais la forme globale reste.

Autre chose... suivant les paramètres le système peut converger. Effectivement, en changeant le facteur 28 par la valeur 22 nous avons par exemple (convergence à gauche):

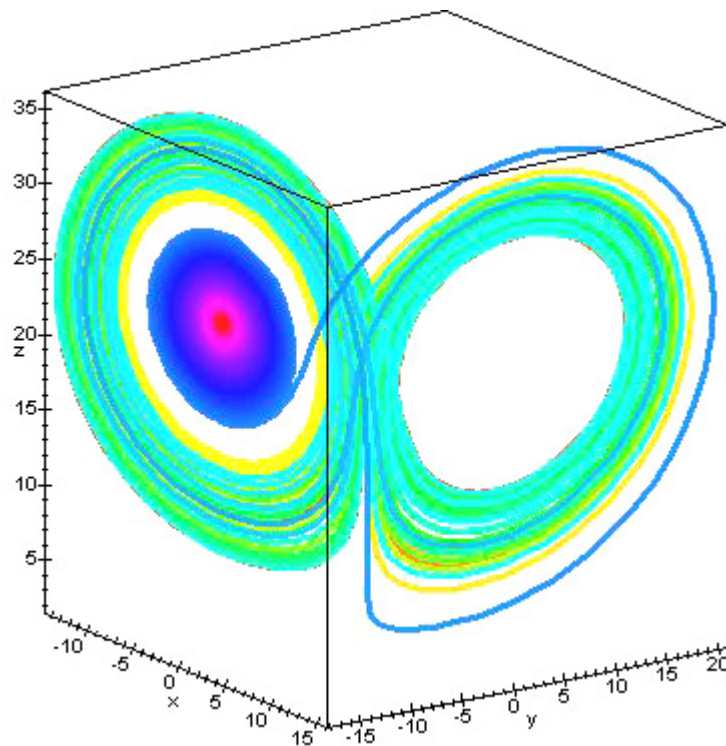


Figure: 69.21 - Convergence du système

ou avec la valeur 19 le résultat est encore plus trivial:

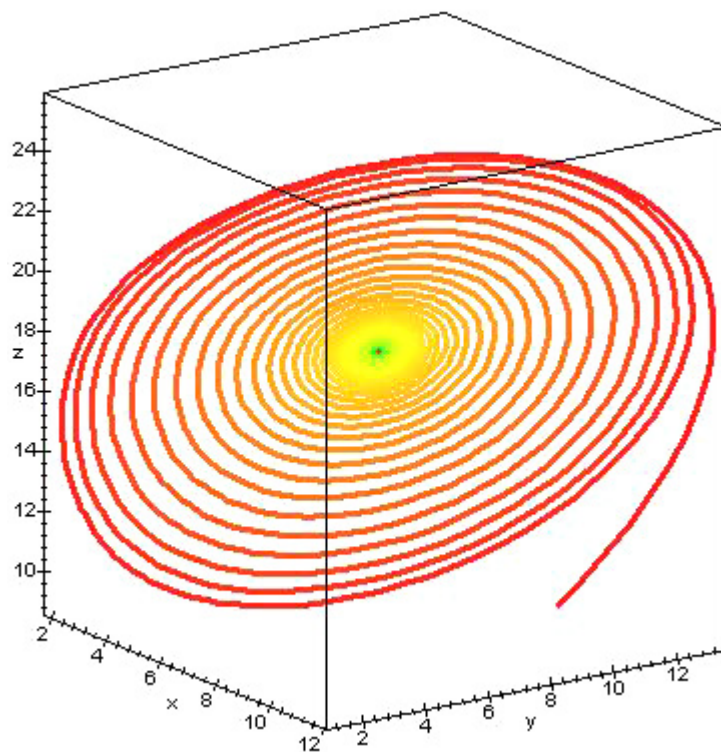


Figure: 69.22 - Convergence directe

ou encore avec une valeur proche de 1:

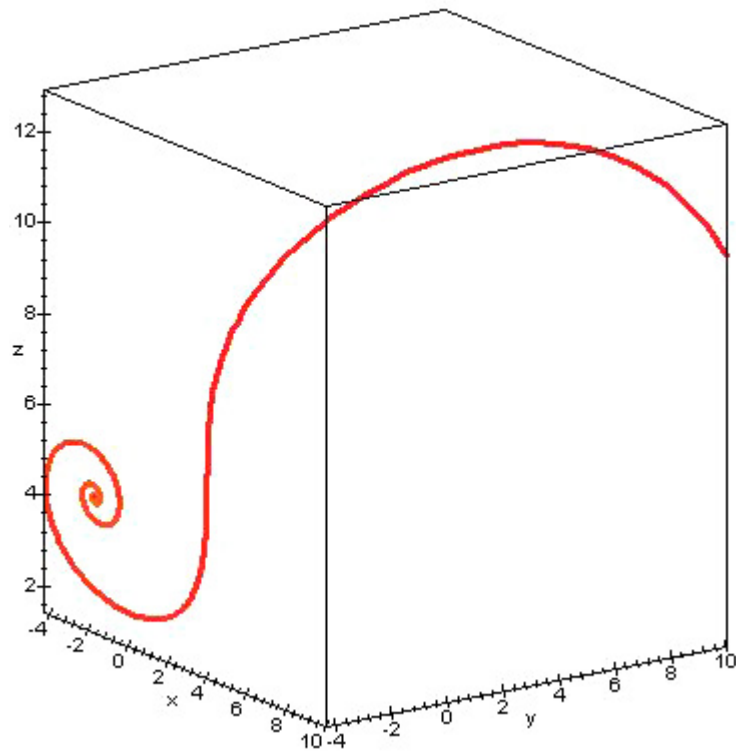


Figure: 69.23 - Encore plus rapide

On remarque un dernier cas intéressant, c'est que si le nombre de Prandtl vaut 1 alors le système est stable:

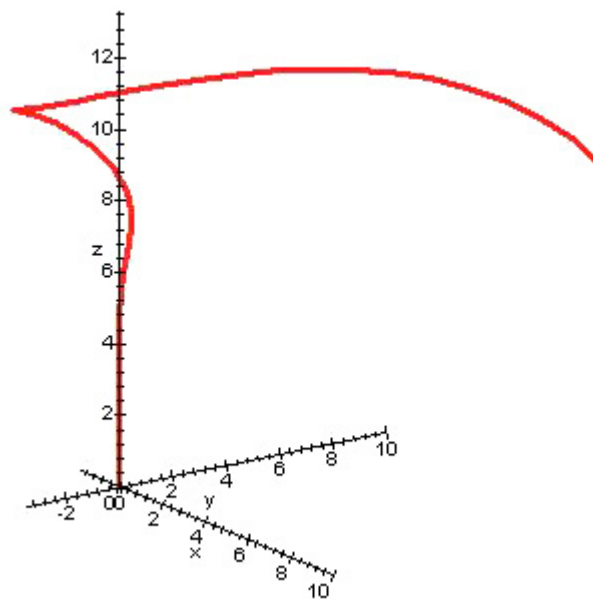


Figure: 69.24 - Stabilité du système

Cette sensibilité aux conditions initiales, ainsi que la forme de l'attracteur étrange de Lorenz a amené les météorologues à faire une métaphore avec la phrase suivante: le battement d'ailes d'un papillon au Brésil peut-il provoquer une tornade au Texas? (en taisant la dissipation de l'erreur dû aux échelles considérées...).

D'où la dénomination par la suite de "**effet Papillon**" pour l'étude de l'attracteur de Lorenz appelé aussi dans le domaine le "**papillon de Lorenz**".

10. VAGUES

Quand on se retrouve sur un bateau, le mouvement de l'eau rend beaucoup de personnes malades. C'est aussi le cas quand on se trouve dans la galère de calculer ce type de mouvement... En effet, les vagues dans l'eau ne sont ni des ondes transversales, ni des ondes longitudinales. Elles sont un peu des deux. Les particules d'eau décrivent des cercles ou des ellipses, dans un plan vertical parallèle au déplacement de la vague. Le gros problème est qu'une des conditions limites se trouve à la surface de l'eau et que l'on ne sait pas où se trouve cette surface. C'est précisément une des inconnues.

Le problème est assez ardu pour que ce calcul ne figure pas dans les ouvrages de physique générale ou spécialisés francophones. Nous avons alors tenté de l'inclure ici en simplifiant énormément et en ne gardant que les cas les plus simples. Le calcul qui suit peut être pris comme exemple de la façon comme les physiciens se simplifient la vie en faisant des approximations et aussi comme exemple magistral d'un phénomène banal mais complexe à modéliser.

Nous ne calculerons pas l'équation d'onde. Nous devons calculer le mouvement des particules depuis la surface jusqu'au fond de l'eau, puisque le mouvement s'étend jusqu'au fond. Mais comme les particules décrivent des ellipses, il faut au moins deux variables par point: il nous faut un champ vectoriel. Nous calculerons donc le champ vectoriel des vitesses des particules, mais nous ne le calculerons pas directement. Ce que nous allons trouver, c'est un potentiel scalaire de vitesses dont le gradient sera le champ vectoriel de vitesses.

Un dernier mot sur les ondes que nous allons calculer. Ces ondes s'appellent "**ondes de gravité**", parce que c'est le poids de l'eau qui sert de force de restitution pour faire descendre les sommets et remonter les creux.

Je tiens à remercier Monsieur Louis Peralta pour sa maîtrise mathématique et sa pédagogie sans qui ces développements passionnants n'auraient jamais pu être présentés sur le site web. Je lui suis redevable, avec son accord, d'un copier/coller de 99% de son support de cours.

Nous allons faire des calculs dans le cas le plus simple: des ondes dans un canal rectiligne de section rectangulaire et de largeur et profondeur constante. Ceci réduit les dimensions du problème à deux (puisque la troisième à un profil identique de vague quel que soit l'endroit où l'on se positionne en largeur): la profondeur et la direction d'avance des vagues. Bien entendu, nous négligerons la friction sur le fond ou les parois du canal et... bien d'autres choses.

Le système de coordonnées choisi est alors:

- x horizontal, positif dans le sens d'avancement des vagues (sens de la longueur du canal)
- y horizontal dans le sens perpendiculaire à l'avancement des vagues (sens de la largeur du canal). Encore une fois, aucune variable ne dépendra de y pour les raisons déjà susmentionnées
- z vertical et positif vers le haut. Le zéro étant choisi à la surface de l'eau non perturbée. Sous l'eau, z sera donc négatif. La profondeur du canal étant h , le fond de l'eau se trouvera alors à $z = -h$

Maintenant, rappelons que nous avons démontré après de très longs développements dans le chapitre de Mécanique Des Milieux Continus, que si le fluide est incompressible (divergence du champ vectoriel des vitesses est donc nul $\vec{\nabla} \cdot \vec{v} = 0$) et que la viscosité dynamique l'est aussi, nous avons alors l'équation d'Euler de 1^{ère} forme:

$$\rho d_t \vec{v} = -\vec{\nabla} \cdot \vec{p} + \rho \vec{f} = -\vec{\nabla} \cdot \vec{p} + \rho U = -\vec{\nabla} \cdot \vec{p} - \rho g z \quad (69.264)$$

car U est pour rappel le potentiel scalaire du champ gravitationnel.

Ce que nous écrirons explicitement sous la forme suivante dans le cas de notre étude des vagues:

$$d_t \vec{v} = -\frac{1}{\rho} (\vec{\nabla} \cdot \vec{p} + \rho U) \quad (69.265)$$

Avant de continuer, nous allons changer la forme du terme $d_t \vec{v}$ qui est une dérivée totale et non partielle. Examinons une des deux composantes (x ou z) de ce terme. Prenons la composante x du vecteur vitesse et utilisons ce que nous avons vu dans le chapitre de Calcul Différentiel Et Intégral (nous avons omis la composante y puisque le système n'en dépend pas):

$$df(x, y) = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy \quad (69.266)$$

Nous avons dans notre situation:

$$dv_x(x, z, t) = \frac{\partial v_x}{\partial t} dt + \frac{\partial v_x}{\partial x} dx + \frac{\partial v_x}{\partial z} dz \quad (69.267)$$

Il vient alors:

$$\frac{dv_x}{dt} = \frac{\partial v_x}{\partial t} \frac{dt}{dt} + \frac{\partial v_x}{\partial x} \frac{dx}{dt} + \frac{\partial v_x}{\partial z} \frac{dz}{dt} = \frac{\partial v_x}{\partial t} + \frac{\partial v_x}{\partial x} v_x + \frac{\partial v_x}{\partial z} v_z \quad (69.268)$$

Le premier terme $\partial_t v$ est un terme linéaire, par contre les trois autres sont des termes quadratiques (car ils comportent le produit d'une vitesse par la dérivée d'une vitesse) qui sont une infection dans tous les calculs formels. Nous allons donc limiter nos calculs aux cas où ces termes quadratiques sont négligeables devant les termes linéaires (ce qui signifie implicitement la vitesse doit varier très peu sinon quoi cette hypothèse n'est plus valable). Avec cette restriction, nous avons alors:

$$\frac{d\vec{v}}{dt} \cong \frac{\partial \vec{v}}{\partial t} \quad (69.269)$$

Nous pouvons alors écrire:

$$\partial_t \vec{v} = -\frac{1}{\rho} (\vec{\nabla} \cdot \vec{p} + \rho U) \quad (69.270)$$

Nous allons rajouter une restriction supplémentaire à nos calculs. Cette restriction consiste à imposer que le rotationnel de la vitesse est nul:

$$\vec{\nabla} \times \vec{v} = \vec{0} \quad (69.271)$$

Cela revient à dire, comme nous l'avons vu dans le chapitre de Mécanique Des Milieux Continus, qu'il n'y a pas de tourbillons dans l'eau. Nous avons déjà imposé que la vitesse était nulle dans le sens des y , ce qui empêche les tourbillons verticaux (axe de rotation vertical). Maintenant nous éliminons la possibilité de tourbillons avec l'axe de rotation horizontal. Ce n'est pas, en fait, une trop grande restriction, nous savons qu'un objet abandonné aux vagues ne tourne pas sur lui-même à moins d'être pris dans des rouleaux. Dans les cas que nous allons calculer, les amplitudes sont limitées, nous sommes dans le cas linéaire et, surtout, pas dans celui des rouleaux.

Nous avons démontré dans le chapitre de Calcul Vectoriel que lorsque le rotationnel d'une variable vectorielle est nul, la variable peut être exprimée comme le gradient d'un potentiel scalaire:

$$\text{rot}[\overrightarrow{\text{grad}}(f)] = \vec{\nabla} \times (\vec{\nabla} f) = \vec{0} \quad (69.272)$$

Il vient alors que nous pouvons écrire:

$$\vec{v} = \vec{\nabla} \phi \quad (69.273)$$

Nous avons alors:

$$\partial_t (\vec{\nabla} \phi) = -\frac{1}{\rho} (\vec{\nabla} \cdot p + \rho U) \quad (69.274)$$

Ce qui peut s'écrire:

$$\vec{\nabla} \cdot \left(\partial_t \phi + \frac{1}{\rho} p + gz \right) = 0 \quad (69.275)$$

Cette équation tient lieu d'équation d'onde, même si elle ne décrit pas directement la position ou la vitesse des particules. Une fois résolue l'équation et la fonction $\phi(x, z, t)$ trouvée, nous pourrions déduire les vitesses:

$$\begin{aligned} v_x &= \frac{\partial \phi}{\partial x} \\ v_z &= \frac{\partial \phi}{\partial z} \end{aligned} \quad (69.276)$$

Il ne faudra pas oublier que puisque la divergence du champ vectoriel des vitesses est nul:

$$\vec{\nabla} \circ \vec{v} = \vec{\nabla} \cdot \vec{v} = 0 = \vec{\nabla} \cdot (\vec{\nabla} \phi) = \vec{\nabla}^2 \phi = \Delta \phi \quad (69.277)$$

donc le laplacien scalaire de ϕ est nul:

$$\Delta \phi = \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial z^2} = 0 \quad (69.278)$$

Soit:

$$\frac{\partial^2 \phi}{\partial x^2} = -\frac{\partial^2 \phi}{\partial z^2} \quad (69.279)$$

avec $-h < z < 0$.

La solution doit satisfaire des conditions limites. La plus évidente est que la vitesse verticale doit être nulle au fond de l'eau.

$$v_z = \frac{\partial \phi}{\partial z} = 0 \quad (69.280)$$

Pour trouver les conditions limites à la surface, nous allons définir la variable $\xi(x, t)$ qui est égale au déplacement vertical d'une particule d'eau située à la surface de l'eau dont la position d'équilibre (en absence de vagues) est x (à ne pas confondre avec l'autre x !). La hauteur de la surface de l'eau, mesurée par référence au niveau de l'eau sans vague, sera donnée donc par $\xi(x, t)$. Ainsi, la vitesse verticale v_z à la surface de l'eau sera:

$$v_z = \frac{\partial \phi}{\partial z} = \frac{\partial \xi}{\partial t} \quad (69.281)$$

C'est une astuce très subtile et pas évidente du tout à anticiper!

À la surface, nous pouvons considérer que la pression p est toujours égale à la pression atmosphérique pour toutes les positions en x et en ξ . Dès lors, à priori rien ne changera si la pression atmosphérique change, et il est plus commode de la considérer nulle. Avec ceci, l'équation:

$$\vec{\nabla} \cdot \left(\partial_t \phi + \frac{1}{\rho} p + gz \right) = 0 \quad (69.282)$$

Devient:

$$\vec{\nabla} \cdot (\partial_t \phi + gz) = 0 \quad (69.283)$$

Dont nous déduisons:

$$\partial_t \phi = -gz \quad (69.284)$$

Si nous nous plaçons à la surface, ce qui correspond à $z = 0$, la relation précédente devient:

$$\partial_t \phi = -g\xi \quad (69.285)$$

et en dérivant cette dernière relation par rapport au temps, nous obtenons:

$$\frac{\partial^2 \phi}{\partial t^2} = -g \frac{\partial \xi}{\partial t} \quad (69.286)$$

puis en utilisant la relation démontrée juste plus haut:

$$v_z \Big|_{z=0} = \frac{\partial \phi}{\partial z} = \frac{\partial \xi}{\partial t} \quad (69.287)$$

Il vient:

$$\frac{\partial^2 \phi}{\partial t^2} = -g \frac{\partial \phi}{\partial z} \quad (69.288)$$

Résumons les conditions que le potentiel scalaire ϕ doit satisfaire:

$$\begin{aligned} \frac{\partial^2 \phi}{\partial x^2} &= -\frac{\partial^2 \phi}{\partial z^2} & -h < z < 0 \\ \frac{\partial \phi}{\partial z} &= 0 & z = -h \\ \frac{\partial^2 \phi}{\partial t^2} &= -g \frac{\partial \phi}{\partial z} & z = 0 \end{aligned} \quad (69.289)$$

Nous allons chercher seulement des solutions du type séparable (cf. chapitre de Calcul Différentiel Et Intégral). C'est-à-dire que ϕ , qui est une fonction de x , z et t pourra s'écrire comme le produit de trois fonctions qui dépendent, chacune, d'une seule variable. De plus, nous travaillerons seulement en régime sinusoïdal (une intuition... de l'observation du réel) et en utilisant le formalisme des phaseurs (cf. chapitre de Mécanique Ondulatoire). Dès lors, nous avons:

$$\phi(x, z, t) = \chi(x)\psi(z)e^{i\omega t} \quad (69.290)$$

Comme d'habitude, nous retiendrons seulement la partie réelle des phaseurs des solutions à la fin. Calculons les deuxièmes dérivées partielles par rapport à x et z :

$$\begin{aligned} \frac{\partial^2 \phi}{\partial x^2} &= \chi''\psi e^{i\omega t} \\ \frac{\partial^2 \phi}{\partial z^2} &= \chi\psi'' e^{i\omega t} \end{aligned} \quad (69.291)$$

L'équation suivante:

$$\frac{\partial^2 \phi}{\partial x^2} = -\frac{\partial^2 \phi}{\partial z^2} \quad -h < z < 0 \quad (69.292)$$

nous dit que sous la surface, ces deux expressions sont égales au signe près:

$$\chi''\psi e^{i\omega t} = -\chi\psi'' e^{i\omega t} \quad (69.293)$$

d'où nous déduisons:

$$\frac{\chi''}{\chi} = -\frac{\psi''}{\psi} = \lambda \quad (69.294)$$

λ ne peut être qu'une constante, car les deux côtés de l'expression dépendent de variables indépendantes (ce qui est très astucieux comme observation!). Cette constante peut prendre, en principe, n'importe quelle valeur réelle ou complexe.

Ce qui nous donne le système de deux équations différentielles indépendantes du deuxième ordre:

$$\begin{aligned}\chi'' - \lambda\chi &= 0 \\ \psi'' + \lambda\psi &= 0\end{aligned}\quad (69.295)$$

Il s'agit donc d'un système d'équations différentielles linéaire d'ordre 2 indépendantes, quasiment identique à celui résolu dans le chapitre de Thermodynamique lorsque nous avons étudié l'équation de la chaleur.

Ceci étant dit, nous ne connaissons pas le signe de la constante λ . Ce que nous savons d'après ce que nous avons étudié dans le chapitre de Calcul Différentiel Et Intégral c'est que si b est positif alors la solution de l'équation différentielle en z :

$$\psi'' + \lambda\psi = 0 \quad (69.296)$$

aura des solutions harmoniques. Ce qui serait très surprenant... et contre intuitif avec ce que nous observons dans la réalité. Donc nous allons imposer λ négatif et dès lors nous avons vu dans le chapitre de Calcul Différentiel Et Intégral que nous aurons en adoptant à peu près les mêmes écritures:

$$\begin{aligned}\chi &= Ae^{(\alpha+i\beta)(x+\tau')} + Be^{(\alpha-i\beta)(x+\tau')} \\ \psi &= Ce^{k_1(z+\tau)} + De^{k_2(z+\delta)}\end{aligned}\quad (69.297)$$

où nous avons le discriminant de la deuxième équation différentielle (cf. chapitre de Calcul Différentiel Et Intégral) qui vaut:

$$k_{1,2} = \frac{-a \pm \sqrt{a^2 - 4b}}{2} = \pm\sqrt{\lambda} \quad (69.298)$$

Donc en d'autres termes:

$$k_1 = -k_2 = k \quad (69.299)$$

Il vient alors:

$$\begin{aligned}\chi &= Ae^{(\alpha+i\beta)(x+\tau')} + Be^{(\alpha-i\beta)(x+\tau')} \\ \psi &= Ce^{k(z+\tau)} + De^{-k(z+\delta)}\end{aligned}\quad (69.300)$$

et le discriminant de la première équation différentielle sera purement complexe:

$$\alpha \pm i\beta = \pm\sqrt{\lambda} \quad (69.301)$$

et donc peut s'écrire:

$$\pm ik \quad (69.302)$$

Notre système se simplifie alors encore plus:

$$\begin{aligned}\chi &= Ae^{ik(x+\tau')} + Be^{-ik(x+\tau')} \\ \psi &= Ce^{k(z+\tau)} + De^{-k(z+\delta)}\end{aligned}\quad (69.303)$$

Si nous considérons l'avance (déphasage) nulle nous avons:

$$\begin{aligned}\chi &= A'e^{ikx} + B'e^{-ikx} \\ \psi &= Ce^{kz} + De^{-kz}\end{aligned}\quad (69.304)$$

Puisque:

$$\phi(x, z, t) = \chi(x)\psi(z)e^{i\omega t} \quad (69.305)$$

Nous avons:

$$\phi(x, z, t) = (Ce^{kz} + De^{-kz}) \left(A'e^{i(\omega t + kx)} + B'e^{i(\omega t - kx)} \right) \quad (69.306)$$

Donc dans la deuxième parenthèse, nous avons simplement deux vagues qui vont dans des sens opposés. Gardons-en qu'une des deux pour simplifier l'analyse (peu importe laquelle!):

$$\phi(x, z, t) = (Ce^{kz} + De^{-kz}) \left(A'e^{i(\omega t + kx)} \right) \quad (69.307)$$

Pour déterminer les constantes, nous allons utiliser la condition limite donnée plus haut:

$$v_z \Big|_{z=-h} = \frac{\partial \phi}{\partial z} = 0 \quad (69.308)$$

ce qui nous donne dans le cas présent:

$$k(Ce^{-kh} - De^{kh}) \left(A'e^{i(\omega t + kx)} \right) = 0 \quad (69.309)$$

Alors évidemment nous n'allons pas simplifier totalement de manière à avoir K étant nul sinon cela ne correspondrait pas avec la réalité. Par contre, nous allons simplifier de la manière suivante:

$$k(Ce^{-kh} - De^{kh}) = 0 \quad (69.310)$$

ce qui impose dans le cas non trivial que:

$$Ce^{-kh} = De^{kh} \quad (69.311)$$

Nous pouvons donc choisir ce que nous voulons, mais si nous souhaitons nous simplifier la vie en anticipant ce qui va venir... il vaudrait mieux prendre le choix suivant:

$$C e^{-kh} = D e^{kh} = \frac{1}{2} \quad (69.312)$$

Dès lors:

$$\begin{aligned} C &= \frac{1}{2} e^{kh} \\ D &= \frac{1}{2} e^{-kh} \end{aligned} \quad (69.313)$$

Ainsi, il vient:

$$\phi(x, z, t) = (C e^{kz} + D e^{-kz}) (A' e^{i(\omega t + kx)}) = \left(\frac{e^{k(z+h)} + e^{-k(z+h)}}{2} \right) (A' e^{i(\omega t + kx)}) \quad (69.314)$$

En utilisant les fonctions hyperboliques (cf. chapitre de Trigonométrie) nous avons:

$$\phi(x, z, t) = \left(\frac{e^{k(z+h)} + e^{-k(z+h)}}{2} \right) (A' e^{i(\omega t + kx)}) = \cosh(k(z+h)) (A' e^{i(\omega t + kx)}) \quad (69.315)$$

Nous allons continuer à tenter de déterminer les constantes en utilisant la deuxième condition limite:

$$\frac{\partial^2 \phi}{\partial z^2} = -g \frac{\partial \phi}{\partial z} \quad z = 0 \quad (69.316)$$

Il vient alors (pour la dérivée du cosinus hyperbolique voir le chapitre de Calcul Différentiel Et Intégral):

$$-\omega^2 A' \cosh(k(z+h)) e^{i(\omega t + kx)} = -g k A' \sinh(k(z+h)) e^{i(\omega t + kx)} \quad (69.317)$$

et en posant donc $z = 0$:

$$-\omega^2 A' \cosh(kh) e^{i(\omega t + kx)} = -g k A' \sinh(kh) e^{i(\omega t + kx)} \quad (69.318)$$

et en simplifiant:

$$-\omega^2 \cosh(kh) = -g k \sinh(kh) \quad (69.319)$$

ou autrement écrit:

$$\tanh(kh) = \frac{\omega^2 h}{g} \frac{1}{kh} \quad (69.320)$$

En se rappelant que (cf. chapitre de Mécanique Ondulatoire):

$$k = \frac{\omega}{v_{phase}} \quad (69.321)$$

La relation antéprécédente n'a pas à ma connaissance de solutions permettant d'exprimer directement la vitesse de phase à partir d'une expression analytique (elle est donc transcendante) fonction de la profondeur h et de la fréquence (implicitement la pulsation) de la vague sans passer un développement de Maclaurin (cf. chapitre de Suites Et Séries) de la tangente hyperbolique pour de petites valeurs de kh telles que:

$$\tanh(kh) \cong kh \Rightarrow kh \cong \frac{\omega^2 h}{g} \frac{1}{kh} \quad (69.322)$$

d'où nous tirons donc lorsque la profondeur et k aussi sont petits en faisant une simplification élémentaire de la relation précédente:

$$v_{phase} \cong \sqrt{gh} \quad (69.323)$$

Relation qui est appelée "**formule de Lagrange**". Par exemple, pour un tsunami, les longueurs d'onde sont immenses (de l'ordre de la centaine de kilomètres) donc le produit hk est petit (de l'ordre de 0.25 dans les océans).

Si au contraire la profondeur h est très grande ainsi que k nous avons alors en développant en série de Taylor (cf. chapitre de Suites Et Séries):

$$\tanh(kh) \cong 1 \quad (69.324)$$

d'où nous tirons alors aussi après une simplification élémentaire:

$$v_{phase} = \sqrt{\frac{g}{k}} \quad (69.325)$$

Relation qui est appelée "**formule de Newton**".

Cependant nous pouvons résoudre numériquement la relation:

$$\tanh(kh) = \frac{\omega^2 h}{g} \frac{1}{kh} \quad (69.326)$$

et nous avons alors:

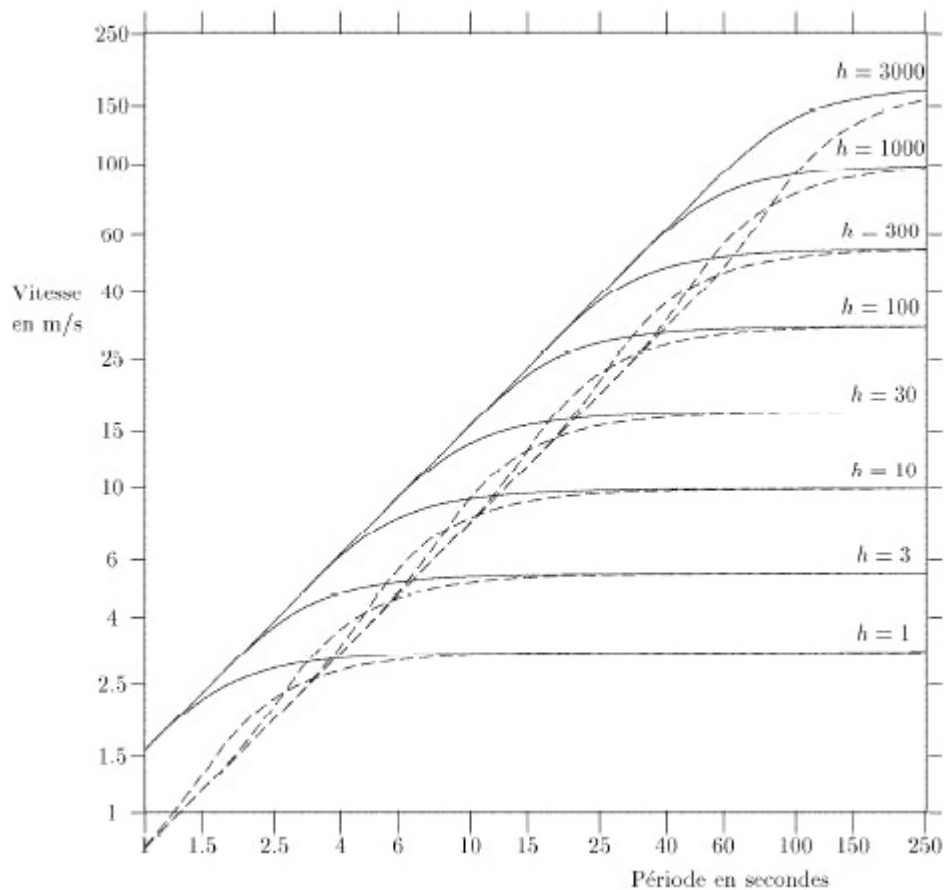


Figure: 69.25 - Espace des phases du système des équations de Lorenz Source: LPFR

Nous pouvons en observant ce graphique comprendre pourquoi les vagues se brisent (flux supercritique). Effectivement, puisque la vitesse de phase pour une période donnée augmente en fonction de la profondeur h , la partie supérieure de la vague va plus vite que la partie inférieure et finit donc pas s'effondrer faute de support. La partie supercritique de la vague est destructrice dans le cas des tsunamis car elle l'eau est accélérée dans sa chute par la gravité.

Il en découle aussi que les plus grandes vagues rattrapent les petites et que leurs amplitudes se superposent sur une certaine distance. Nous déterminerons plus loin la fonction reliant l'amplitude de la vague à la distance.

Revenons maintenant sur:

$$\phi(x, z, t) = \cosh(k(z+h)) \left(A' e^{i(\omega t + kx)} \right) \quad (69.327)$$

en ne conservant que la partie réelle:

$$\phi(x, z, t) = A' \cosh(k(z+h)) \cos(\omega t + kx) \quad (69.328)$$

Rappelons que nous avons vu au début:

$$v_z = \frac{\partial \phi}{\partial z} \quad (69.329)$$

Pour obtenir Z il faut donc dériver la relation antérieure par rapport à z , puis intégrer le résultat par rapport au temps.

La dérivée donne donc:

$$v_z = \frac{dz}{dt} = \frac{\partial \phi}{\partial z} = kA' \sinh(k(z+h)) \cos(\omega t + kx) \quad (69.330)$$

Soit:

$$dz = kA' \sinh(k(z+h)) \cos(\omega t + kx) dt \quad (69.331)$$

La primitive donne:

$$\int dz = \int kA' \sinh(k(z+h)) \cos(\omega t + kx) dt = kA' \sinh(k(z+h)) \int \cos(\omega t + kx) dt \quad (69.332)$$

Soit:

$$Z = \frac{k}{\omega} A' \sinh(k(z+h)) \sin(\omega t + kx) \quad (69.333)$$

Or nous avons vu au début de notre étude qu'en nous plaçant en $z = 0$ nous avons donc en surface:

$$\xi = \frac{k}{\omega} A' \sinh(kh) \sin(\omega t + kx) \quad (69.334)$$

et en notant R l'amplitude de la vague:

$$\xi = R \sin(\omega t + kx) \quad (69.335)$$

Avec ceci le déplacement vertical Z devient:

$$Z = R \frac{\sinh(k(z+h))}{\sinh(kh)} \sin(\omega t + kx) \quad (69.336)$$

et en se rappelant que le potentiel de vitesses est donné par:

$$\phi(x, z, t) = A' \cosh(k(z+h)) \cos(\omega t + kx) \quad (69.337)$$

Nous avons avec cette nouvelle constante R :

$$\phi(x, z, t) = R \frac{\omega \cosh(k(z+h))}{k \sinh(kh)} \cos(\omega t + kx) \quad (69.338)$$

La vitesse horizontale d'une particule d'eau sera alors:

$$v_x = \frac{\partial \phi}{\partial x} = \frac{dx}{dt} = -R\omega \frac{\cosh(k(z+h))}{\sinh(kh)} \sin(\omega t + kx) \quad (69.339)$$

et le déplacement horizontal sera l'intégrale de la relation précédente par rapport au temps:

$$\int dx = \int -R\omega \frac{\cosh(k(z+h))}{\sinh(kh)} \sin(\omega t + kx) dt = -R\omega \frac{\cosh(k(z+h))}{\sinh(kh)} \int \sin(\omega t + kx) dt \quad (69.340)$$

Soit:

$$X = R \frac{\cosh(k(z+h))}{\sinh(kh)} \cos(\omega t + kx) \quad (69.341)$$

et à la surface en $z = 0$:

$$X = R \frac{\cosh(kh)}{\sinh(kh)} \cos(\omega t + kx) \quad (69.342)$$

10.1. PROFONDEUR DES VAGUES

Il est maintenant intéressant de calculer la profondeur de pénétration de la vague. En effet, nous savons par vécu que les vagues se font sentir de moins en moins à mesure que l'on s'enfonce dans l'eau.

Rappelons que nous avons donc:

$$Z = R \frac{\sinh(k(z+h))}{\sinh(kh)} \sin(\omega t + kx) \quad (69.343)$$

qui correspond au déplacement vertical de la vague. Si:

$$kh = \frac{2\pi}{\lambda} h \gg 1 \quad (69.344)$$

où nous considérerons que la longueur d'onde des vagues est petite et que la profondeur h du milieu dans lequel elles se propagent est grande.

Nous avons alors:

$$\sinh(kh) = \frac{e^{kh} - e^{-kh}}{2} \cong \frac{e^{kh}}{2} \quad (69.345)$$

et:

$$\sinh[k(z+h)] = \frac{e^{k(z+h)} - e^{-k(z+h)}}{2} = \frac{e^{k(z+h)} - e^{-kz} e^{-kh}}{2} \cong \frac{e^{k(z+h)}}{2} \quad (69.346)$$

et nous voyons que pour que cette approximation soit satisfaite, il faut que:

$$e^{-kz} e^{-kh} \ll 1 \quad (69.347)$$

Donc pour que la suite des calculs soit valable, nous imposerons de petites valeurs de z (donc près de la surface).

Dès lors:

$$\begin{aligned} Z &\cong R \frac{\frac{e^{k(z+h)}}{2}}{\frac{e^{kh}}{2}} \sin(\omega t + kx) = R \frac{e^{k(z+h)}}{e^{kh}} \sin(\omega t + kx) = R e^{kz} \sin(\omega t + kx) \\ &= \frac{k}{\omega} A' \sinh(kh) e^{kz} \sin(\omega t + kx) \cong \frac{k}{2\omega} A' e^{2kz+kh} \sin(\omega t + kx) \end{aligned} \quad (69.348)$$

Sans oublier que $-h < z < 0$ (donc z est négatif pour rappel!). L'amplitude des mouvements diminue donc exponentiellement dans ces conditions avec la profondeur. Le lecteur pourra par contre vérifier que pour certaines valeurs de kh petites, nous avons:

$$\frac{\sinh(k(z+h))}{\sinh(kh)} \rightarrow -\infty \quad (69.349)$$

ce qui montre les limites du modèle avec cette approche.

10.2. AMPLITUDE DES VAGUES

Pour l'amplitude du mouvement horizontal la situation est la même. C'est-à-dire pour:

$$kh = \frac{2\pi}{\lambda} h \gg 1 \quad (69.350)$$

où nous considérerons que la longueur d'onde des vagues est petite et que la profondeur h du milieu dans lequel elles se propagent est grande.

Comme nous avons:

$$X = R \frac{\cosh(k(z+h))}{\sinh(kh)} \cos(\omega t + kx) \quad (69.351)$$

il vient alors:

$$\begin{aligned} X &= R \frac{\cosh(k(z+h))}{\sinh(kh)} \cos(\omega t + kx) = R \frac{\frac{e^{k(z+h)} + e^{-k(z+h)}}{2}}{\frac{e^{kh} - e^{-kh}}{2}} \cos(\omega t + kx) \\ &= R \frac{e^{k(z+h)}}{e^{kh}} \cos(\omega t + kx) \cong R e^{kz} \cos(\omega t + kx) = \frac{k}{\omega} A' \sinh(kh) e^{kz} \cos(\omega t + kx) \\ &\cong \frac{k}{2\omega} A' e^{k(z+h)} \cos(\omega t + kx) \end{aligned} \quad (69.352)$$

où pour rappel, nous avons toujours $-h < z < 0$ et que ce modèle n'est valable que près de la surface.

Donc pour ces deux approximations, nous avons alors lorsque z est petit en valeur absolue par rapport à h (donc en eau peu profonde):

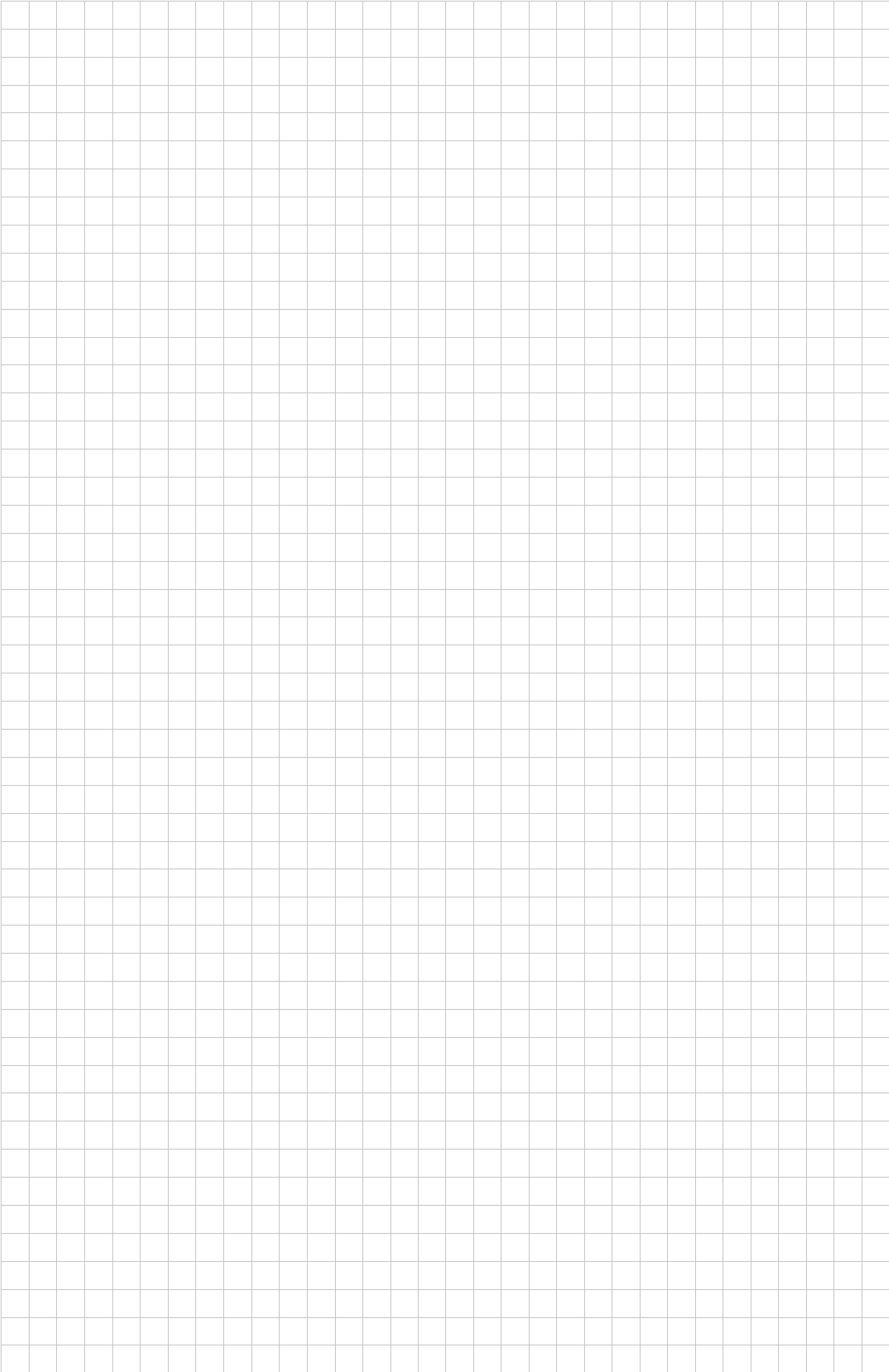
$$\begin{aligned} Z &\cong \frac{k}{2\omega} A' e^{kh} \sin(\omega t + kx) \\ X &\cong \frac{k}{2\omega} A' e^{kh} \cos(\omega t + kx) \end{aligned} \quad (69.353)$$

Ainsi, nous pouvons observer que pour le cas de $kh \gg 1$ et en eau peu profonde (pour z petit), les particules d'eau décrivent un mouvement circulaire (puisque les deux composantes Z et X ont même amplitude).

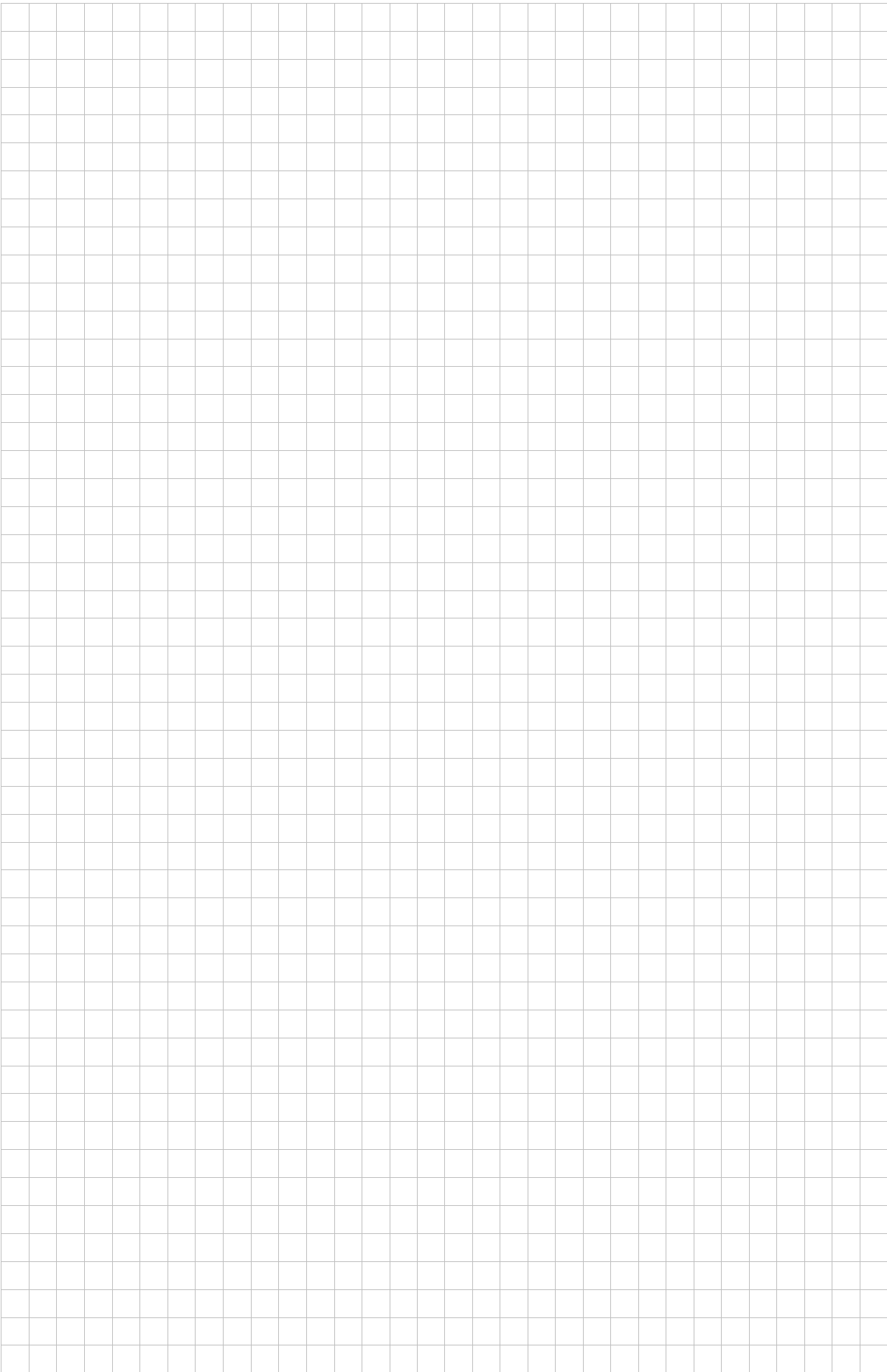
Par contre, dans le cas général, nous avons:

$$\begin{aligned} Z &= R \frac{\sinh(k(z+h))}{\sinh(kh)} \sin(\omega t + kx) \\ X &= R \frac{\cosh(k(z+h))}{\sinh(kh)} \cos(\omega t + kx) \end{aligned} \quad (69.354)$$

Les amplitudes ne sont dès lors plus les mêmes et les mouvements sont alors elliptiques.







70. GÉNIE MÉCANIQUE

Le Génie Mécanique désigne l'ensemble des connaissances liées à la mécanique, au sens physique (sciences des mouvements) et au sens technique (étude des mécanismes). Ce champ de connaissances va de la conception d'un produit mécanique au recyclage de ce dernier en passant, bien sûr par la fabrication, la maintenance, etc.

Ses applications sont très importantes dans de nombreux domaines de la vie de tous les jours que ce soit pour la fabrication de machines, de jouets, d'appareils électro-ménagers ou encore d'immeubles ou de toutes sortes de moyens de transports... et la liste est encore longue...

Encore une fois, nous allons nous concentrer ici uniquement sur la formalisation mathématique de cas pratiques d'applications courantes! Donc ce chapitre est seulement une introduction générale des applications techniques de la mécanique et doit être complété par des travaux pratiques en laboratoire ainsi que par la lecture du chapitre de Thermodynamique où l'équation d'état des solides y est traitée.

Remarque: Il serait prétentieux de prétendre dans le présent chapitre vouloir faire aussi bien et complet que les PDF gratuits d'*Éléments de machine* et de *Résistance des Matériaux* de Nicolet Gaston Raymond qui sont inégalables en matière de contenu et de qualité à ce jour (même comparés aux ouvrages payants sur le sujet!). Il est donc fortement recommandé au lecteur de s'y référer s'il veut une information complète sur le génie mécanique (voir la section de téléchargement du site).

1. ENGRENAGES

Un engrenage est un système mécanique composé de deux roues dentées engrenées servant à la transmission du mouvement de rotation entre elles ou à la propulsion d'un fluide (nous parlons alors de pompe à engrenages):

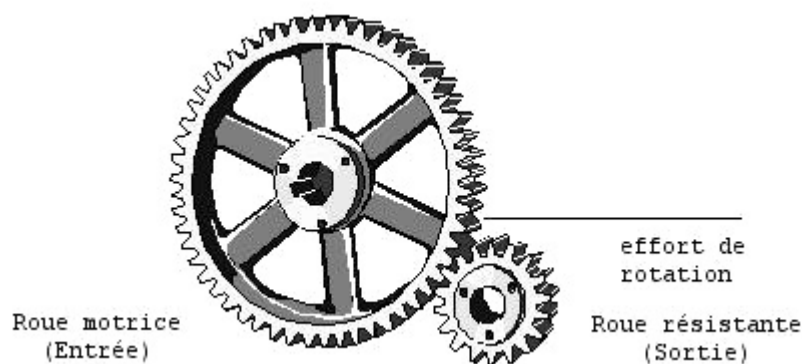


Figure: 70.1 - Système d'engrenage à deux roues dentées circulaires

L'inventeur de la roue dentée est non moins que le célèbre mathématicien, ingénieur et physicien Archimède.

Nous trouvons les engrenages absolument partout dans notre quotidien: voitures, vélos, montres, chaises réglables, etc. Nous conseillons par ailleurs vivement au lecteur de penser aux engrenages (plateaux et pignons liés par une chaîne de transmission) de son vélo pour interpréter les résultats qui vont suivre.

Il m'a semblé donc important de présenter très brièvement comment calculer le pas de denture d'une roue cylindrique en fonction d'une autre pour le plus connu des engrenages qui est "l'engrenage cylindrique" comme celui représenté ci-dessous:

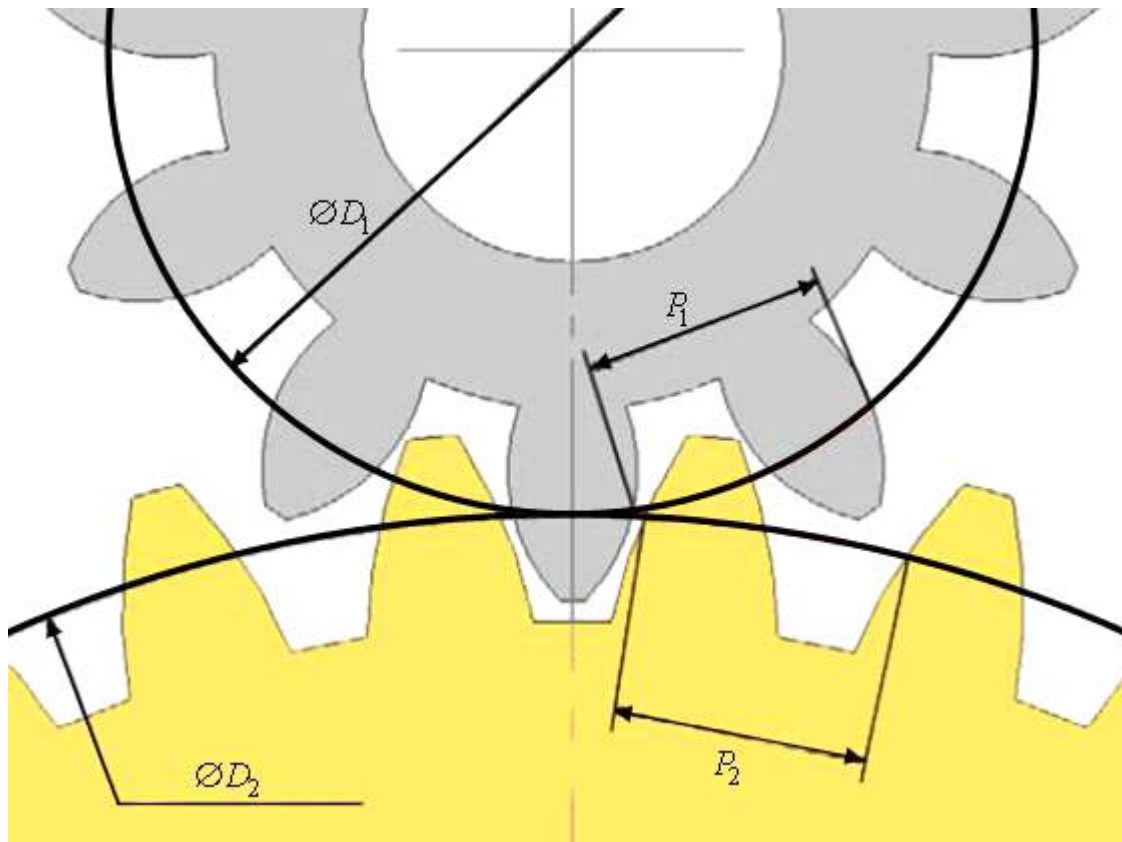


Figure: 70.2 - Engrenage cylindrique avec grammaire

Pour assurer l'entraînement, nous devons donc faire en sorte que les dents d'une des roues cylindriques de diamètre D_1 s'intercalent bien entre les dents d'une autre roue de diamètre D_2 . Pour cela, il nous faut introduire la notion de "pas" de la denture de chaque roue que nous noterons respectivement P_1 et P_2 .

Pour cela nous allons utiliser l'hypothèse que le point de friction des dents est assimilable aux diamètres (représentés ci-dessous par des cercles noirs) qu'auraient les roues cylindriques s'il n'y avait pas de glissement (et donc pas besoin de dents...) de la même façon que l'engrenage. Ces cercles sont appelés "cercles primitifs", ou "cylindres primitifs" ou encore "diamètres primitifs".

Le pas des dents va s'exprimer donc en fonction de la circonférence et du nombre de dents sur chaque roue. Si nous notons Z_1 le nombre de dents d'une roue cylindrique 1 et Z_2 le nombre de dents d'une roue cylindrique 2, nous avons alors les pas qui valent respectivement:

$$\begin{aligned} P_1 &= \frac{2\pi \cdot R_1}{Z_1} = \frac{\pi D_1}{Z_1} \\ P_2 &= \frac{2\pi \cdot R_2}{Z_2} = \frac{\pi D_2}{Z_2} \end{aligned} \quad (70.1)$$

et comme pour que l'engrenage fonctionne il faut que les deux pas soient égaux:

$$P_1 = P_2 \quad (70.2)$$

Nous obtenons alors:

$$\frac{D_1}{Z_1} = \frac{D_2}{Z_2} = m \quad (70.3)$$

où m est appelé le "**module de denture**". Nous retiendrons donc que:

$$D_1 = m \cdot Z_1 \quad D_2 = m \cdot Z_2 \quad (70.4)$$

Nous observons aussi de par les relations ci-dessus que le pas est donc proportionnel au module de denture (le choix d'un gros module donne un nombre de dents faible et le choix d'un petit module un grand nombre de dents).

Le choix du module d'un engrenage s'effectue (parmi des valeurs normalisées) à partir de critères de résistance à la rupture ou de résistance à l'usure.

1.1. RAPPORTS DE TRANSMISSION

Le "**rapport de transmission**" appelé aussi "**rapport de réduction**" d'un engrenage ou d'un système de poulies très important dans le domaine de la mécanique et est défini par:

$$r = \frac{\omega_s}{\omega_e} = \frac{\frac{d\theta_s}{dt}}{\frac{d\theta_e}{dt}} = \frac{d\theta_s}{d\theta_e} \quad (70.5)$$

et donc si la vitesse de rotation (pulsation) est constante (voir le chapitre de Mécanique Classique pour les calculs détaillés de la cinématique du mouvement circulaire) et ne varie plus entre le démarrage et le fonctionnement nominal, il vient:

$$r = \frac{\omega_s}{\omega_e} = \frac{\theta_s}{\theta_e} \quad (70.6)$$

Il s'agit d'une technique très utilisée dans les voitures (en général presque dans tous les moteurs):



Figure: 70.3 - Moteur avec quelques poulies et courroies de transmission

aussi bien dans presque toutes les montres mécaniques:

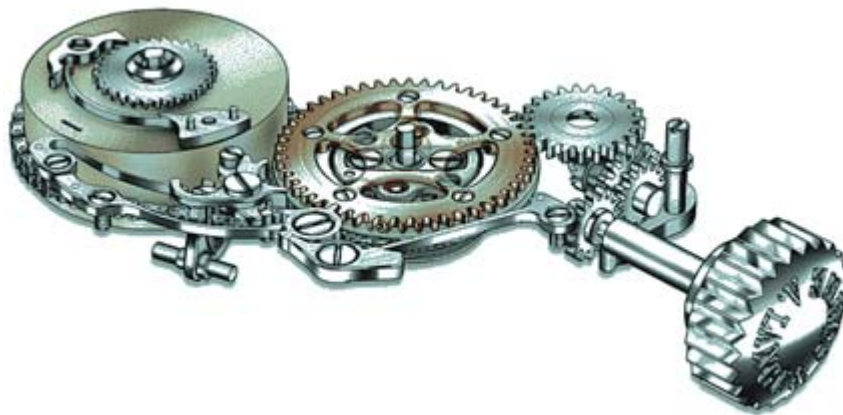


Figure: 70.4 - Système d'engrenage avec rapport de transmission particulier d'une montre

Nous avons aussi bien évidemment:

$$r = \frac{\frac{2\pi}{T_s}}{\frac{2\pi}{T_e}} = \frac{T_e}{T_s} \quad (70.7)$$

Donc, le rapport des vitesses de rotation (pulsation) entre l'arbre de sortie et l'arbre d'entrée est égal au rapport des angles parcourus entre l'arbre de sortie et l'arbre d'entrée.

Nous avons de même:

$$r = \frac{\omega_s}{\omega_e} = \frac{\frac{v_s}{R_s}}{\frac{v_e}{R_e}} \quad (70.8)$$

et comme au point de contact il y a frottement sans glissement alors les deux vitesses sont égales et donc au final:

$$r = \frac{\omega_s}{\omega_e} = \frac{\theta_s}{\theta_e} = \frac{R_e}{R_s} = \frac{\frac{mZ_e}{2}}{\frac{mZ_s}{2}} = \frac{Z_e}{Z_s} = \frac{D_e}{D_s} \quad (70.9)$$

L'égalité:

$$r = \frac{\omega_s}{\omega_e} = \frac{Z_e}{Z_s} \quad (70.10)$$

est souvent appelée aussi "**rapport de réduction**". Il est bien évidemment important de considérer un rapport de réduction sous forme fractionnaire comme une chose qui existe. Ainsi, pour un r donné il faut chercher la fraction la plus proche.

Il paraît clair que si l'engrenage est réducteur de vitesse alors:

$$r = \frac{\omega_s}{\omega_e} = \frac{\theta_s}{\theta_e} = \frac{R_e}{R_s} = \frac{Z_e}{Z_s} = \frac{D_e}{D_s} < 1 \quad (70.11)$$

Raison pour laquelle en vélo, lorsque nous sommes en descente et que la roue motrice tourne très rapidement, nous allons favoriser un grand plateau et un petit pignon (Z_e très grand par rapport à Z_s). Ainsi, avec à 60 [km/h] en descente, une roue de vélo de cours de rayon d'environ 35 [mm] aura une vitesse de rotation:

$$\omega_s \cong 47.6 \left[rad \cdot s^{-1} \right] \cong 7 \left[tours \cdot s^{-1} \right] \quad (70.12)$$

or il n'est pas concevable de tourner les jambes aussi vite pour aller encore plus vite... Donc avec un rapport plateaux/pignons de 53/12, nous aurons:

$$\omega_e = \omega_s \frac{Z_s}{Z_e} \cong 47 \frac{12}{53} \cong 9.75 \left[rad \cdot s^{-1} \right] = 1.55 \left[tours \cdot s^{-1} \right] \quad (70.13)$$

ce qui est déjà beaucoup plus acceptable... pour aller encore plus vite.

Maintenant observons ce qui se passe au niveau du moment de force (qu'il est dur par exemple en vélo de grimper une côte avec un grand plateau et un petit pignon et nous allons voir pourquoi il vaut mieux favoriser un petit plateau avec un grand pignon).

Revenons sur notre schéma en représentant un moment de force constant dont le module M_1 est appliqué sur un pignon à l'engrenage d'entrée de rayon R_1 selon (cf. chapitre de Mécanique Classique):

$$\|\vec{M}_1\| = R_1 \times \vec{F}_1 = R_1 \cdot \|\vec{F}_1\| \cdot \sin(\alpha) = R_1 \cdot \|\vec{F}_1\| = \frac{D_1}{2} \cdot \|\vec{F}_1\| \quad (70.14)$$

La force $\vec{F}_1$ tangentielle sur l'engrenage d'entrée est la même que celle qui fait tourner l'engrenage de sortie

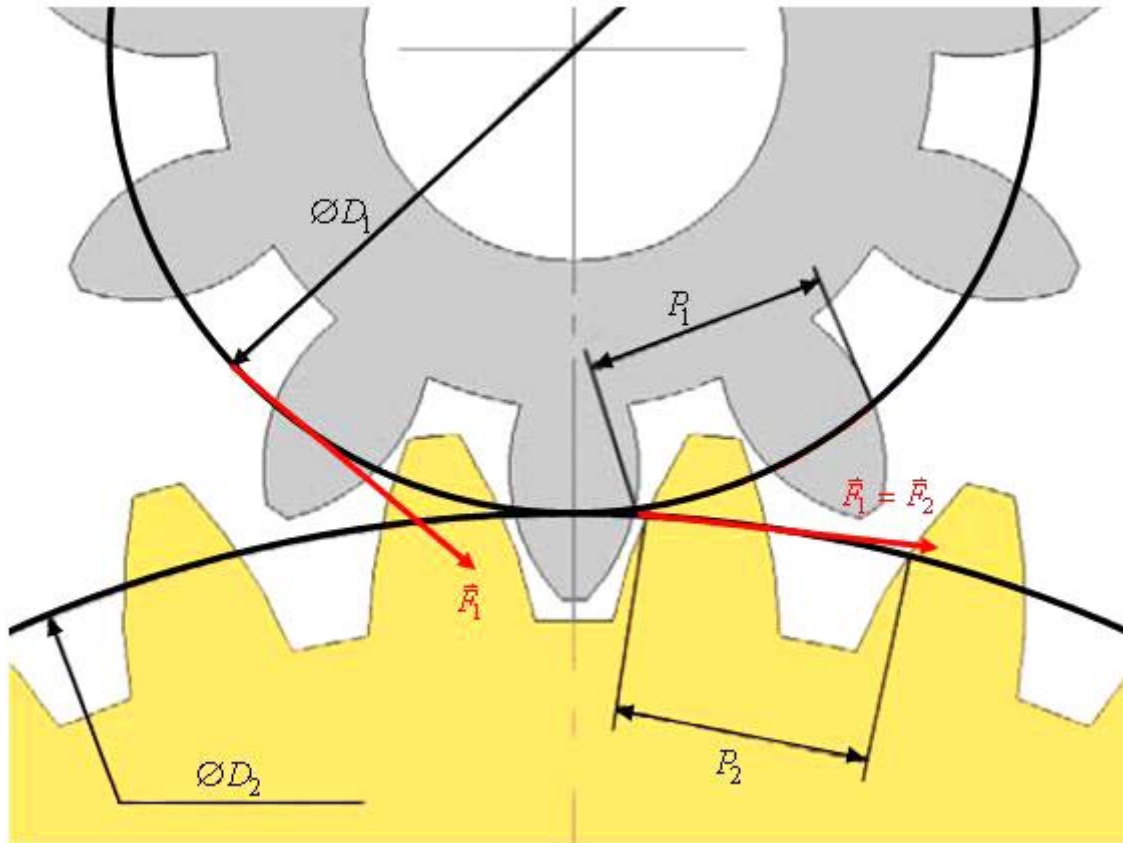


Figure: 70.5 - Engrenage cylindrique avec forces sur pont de contact

Nous avons alors pour les deux engrenages, en l'absence de force de frottement et en considérant que la force au point de contact des deux engrenages est perpendiculaire aux deux rayons respectifs, les relations suivantes :

$$\begin{aligned} F_1 \cdot R_1 &= M_1 \\ F_2 \cdot R_2 &= F_1 \cdot R_2 = M_2 \end{aligned} \quad (70.15)$$

Nous égalisons alors:

$$F_1 = \frac{M_1}{R_1} = \frac{M_2}{R_2} \quad (70.16)$$

et donc:

$$\frac{M_1}{M_2} = \frac{R_1}{R_2} \quad (70.17)$$

et donc:

$$r = \frac{\omega_s}{\omega_e} = \frac{R_e}{R_s} = \frac{Z_e}{Z_s} = \frac{M_e}{M_s} \quad (70.18)$$

Ainsi, pour reprendre notre exemple du vélo... si nous souhaitons appliquer le plus grand moment de force possible à la roue arrière en minimisant le moment de force que nous fournissons sur le pédalier alors la meilleure stratégie selon la relation déterminée ci-dessus est de mettre le plus petit plateau possible associé au plus grand pignon (l'idéal serait d'avoir un plateau plus petit que le plus grand des pignons pour avoir un rapport de transmission supérieur à l'unité).

Attention! Le principe des poulies/engrenages avec rapport de transmission est un fantastique démultiplicateur de force mais en aucun cas il démultiplie le travail!

Remarque: La norme ISO 1122-1 donne la définition inverse et note le rapport de transmission i au lieu de r :

$$i = \frac{\omega_e}{\omega_s} = \frac{\theta_e}{\theta_s} = \frac{M_s}{M_e} \quad (70.19)$$

1.2. ASSOCIATION D'ENGRENAGES

Pour des raisons de contraintes géométriques ou mécaniques, il est parfois nécessaire de construire des étages d'engrenages comme le montre par exemple le train de rouage à 4 roues de la figure ci-dessous (nous allons voir un exemple une fois les notions de base introduites):

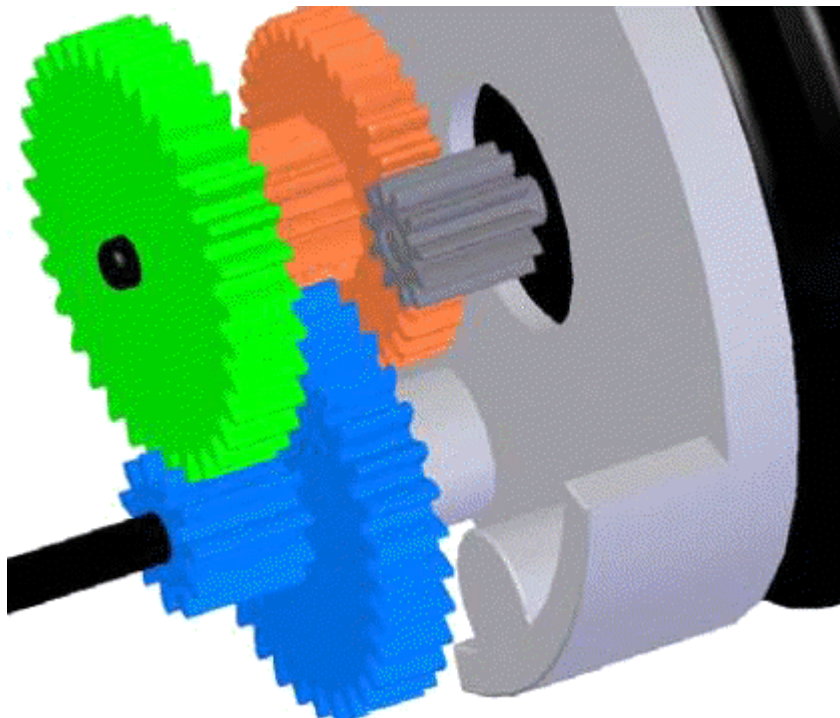


Figure: 70.6 - Exemple de train d'engrenage en 3D

Soit sous forme schématique technique (non conforme aux normes VSM Suisses mais pratique):

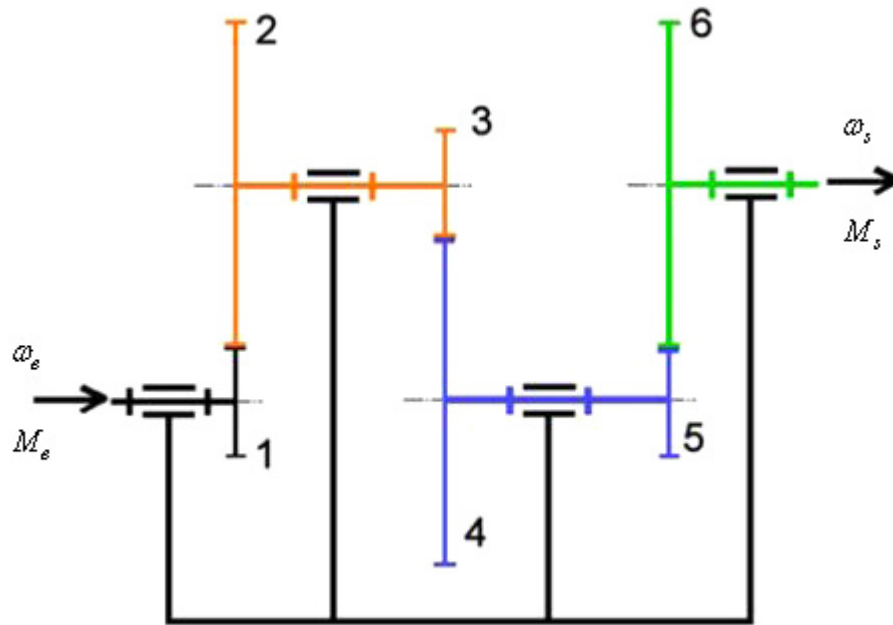


Figure: 70.7 - Train d'engrenage 2D équivalent

Le rapport de transmission global sera alors donné pour la vitesse de rotation par:

$$r = r_{1-6} = \frac{\omega_s}{\omega_e} = \frac{\omega_6}{\omega_1} = \frac{\omega_2}{\omega_1} \cdot \frac{\omega_4}{\omega_3} \cdot \frac{\omega_6}{\omega_5} = \frac{\prod \omega_{menées}}{\prod \omega_{menantes}} \quad (70.20)$$

où la quatrième égalité se simplifie parce que dans le cas particulier ci-dessus:

$$\omega_5 = \omega_4 \text{ et } \omega_3 = \omega_2 \quad (70.21)$$

Nous pouvons aussi exprimer la transmission totale en termes de diamètres. Puisque nous avons démontré que:

$$r = \frac{\omega_s}{\omega_e} = \frac{Z_e}{Z_s} \quad (70.22)$$

il vient alors immédiatement:

$$r = r_{1-6} = \frac{Z_1}{Z_2} \cdot \frac{Z_3}{Z_4} \cdot \frac{Z_5}{Z_6} = \frac{\prod Z_{menantes}}{\prod Z_{menées}} \quad (70.23)$$

expression que nous ne pouvons pas simplifier!

Ainsi, dans le cadre d'une transmission d'une montre à complication astronomique, le rapport de transmission a été obtenu en déterminant la fraction rationnelle 1802/217. Cela étant très difficile à mettre en oeuvre avec uniquement deux engrenages, nous allons construire simplement à l'aide de trois axes et quatre roues (à 7, 31, 34 et 53 dents) le même rapport – qui heureusement n'est pas irréductible - de la façon suivante:

$$r = \frac{1802}{217} = \frac{34 \cdot 53}{7 \cdot 31} = \frac{34}{7} \cdot \frac{53}{31} = \frac{\prod Z_{menantes}}{\prod Z_{menées}} \quad (70.24)$$

Exemple:

Nous souhaitons avec un train à 4 roues sur 2 axes, entraîné par l'axe des heures (qui fera un tour en 12 heures), réaliser un rapport r de transmission très précis, pour réaliser l'entraînement d'une complication de montre incluant une phase de lune par le classique disque portant 2 lunes et tournant derrière un masque.

Le nombre maximum de dents par roue ne devra pas excéder 300 dents et ne descendra pas en dessous de 7. La précision sera la meilleure permise dans les limites de ces nombres de dents.

Nous prendrons pour la lunaison 29 jours 12 heures 44 minutes et 2.9 secondes, soit 2'551'442.9 secondes.

Le calcul du rapport recherché est relativement simple. Le disque à deux lunes fait un tour complet en 2 lunaisons, soit en 5'102'885.8 secondes. La roue du garde-temps fait un tour en 12 heures, soit 43'200 secondes.

Pour un tour du disque à 2 lunes (axe mené) l'axe de la roue 12h (axe menant) fait :

$$r = \frac{T_e}{T_s} = \frac{43'200}{5'102'885.8} \cong 0.008465798 \quad (70.25)$$

et en arrondissant comme ci-dessous, nous avons une erreur d'environ 3 dix-millième de seconde par tranche de 12 heures. Soit environ 0.2 secondes par année de retard. Ce qui est acceptable pour une montre mécanique.

En arrondissant à :

$$r = \frac{T_e}{T_s} = \frac{43'200}{5'102'885.8} \cong 0.00846579 \quad (70.26)$$

nous avons une erreur d'environ 30 secondes par année ce qui reste toujours acceptable pour une montre mécanique. Si nous enlevons encore une décimale, nous avons alors un retard de 6 minutes par année, en enlevant encore une décimale supplémentaire, nous aurions un retard de 49 minutes (ce qui reste toujours acceptable pour nombre de montres mécaniques). Par contre au-delà ce n'est plus acceptable!

Maintenant, pour trouver la fraction rationnelle décomposable la plus proche de ce chiffre il existe de nombreuses méthodes empiriques (par tâtonnement ou en utilisant un arbre de Brocot) et des tables mais la moins pire... pour moi... est celle utilisant les fractions continues (cf. chapitre de Théorie des Nombres) quand elle est applicable...

Rappelons que nous avons vu démontré que:

$$\begin{aligned}
 \frac{a}{b} &= q_1 + \frac{1}{b/r_1} = q_1 + \frac{1}{q_2 + \frac{1}{r_1/r_2}} \\
 &\dots \\
 &= q_1 + \frac{1}{q_2 + \frac{1}{q_3 + \frac{1}{\ddots \frac{1}{q_{n-1} + \frac{1}{q_n}}}}} \quad (70.27)
 \end{aligned}$$

en considérant:

$$\begin{aligned}
 \frac{a}{b} &= q_1 + \frac{r_1}{b} \\
 \frac{b}{r_1} &= q_2 + \frac{r_2}{r_1} \\
 \frac{r_1}{r_2} &= q_3 + \frac{r_3}{r_2} \quad (70.28) \\
 &\dots \\
 \frac{r_{n-2}}{r_{n-1}} &= q_n
 \end{aligned}$$

Si nous notons x le rapport a/b alors les relations ci-dessous nous donnent que q_1 est la partie entière de x , q_2 la partie entière de b/r_1 soit de $b/((x - q_1) \cdot b) = 1/(x - q_1)$ et ainsi de suite

Nous avons alors dans notre cas:

$$\begin{aligned}
& 0.008465798 \\
& \cong 0 + \frac{1}{\frac{1}{0.008465798}} = 0 + \frac{1}{118.1223553881...} \\
& = 0 + \frac{1}{118 + \frac{1}{0.1223553881...}} = 0 + \frac{1}{118 + 8.17291347....} \\
& = 0 + \frac{1}{118 + \frac{1}{8 + \frac{1}{0.17291347....}}} = 0 + \frac{1}{118 + \frac{1}{8 + 5.78323935....}} \\
& = 0 + \frac{1}{118 + \frac{1}{8 + \frac{1}{5 + \frac{1}{0.78323935....}}}} = 0 + \frac{1}{118 + \frac{1}{8 + \frac{1}{5 + 1.2767489....}}} \\
& = 0 + \frac{1}{118 + \frac{1}{8 + \frac{1}{5 + \frac{1}{1 + \dots}}}}
\end{aligned} \tag{70.29}$$

Soit après une mise au dénominateur commun:

$$0 + \frac{1}{118 + \frac{1}{8 + \frac{1}{5 + 1}}} = \frac{1}{118 + \frac{1}{8 + \frac{1}{6}}} = \frac{1}{118 + \frac{1}{\frac{49}{6}}} = \frac{1}{118 + \frac{6}{49}} = \frac{49}{5788} \tag{70.30}$$

Si la différence (erreur) entre cette fraction et la valeur exacte est acceptable (il suffit de calculer l'erreur de temps que cela provoque après une année de fonctionnement de la montre par exemple), nous nous arrêtons ici. Mais il y a un deuxième critère... le numérateur et le dénominateur doivent être décomposables de manière satisfaisante relativement aux exigences du nombre de dents. Or dans le cas présent, le dénominateur (5788) n'est pas décomposable de manière satisfaisante par rapport à nos contraintes (faites la décomposition avec la commande ifactor de Maple 4.00b et vous verrez!).

Si nous continuons à développer notre fraction continue, nous ne trouverons pas de solution acceptable avant que le numérateur ou le dénominateur dépasse la valeur maximale admise de 300^2 .

Dès lors, nous revenons par exemple à la fraction (bon ce n'est pas une approximation acceptable car dans la réalité le dérèglement de la montre se fera trop vite):

$$\frac{49}{5788} \tag{70.31}$$

et nous trichons un peu par essais successifs jusqu'à trouver un bon rapport:

$$\frac{49}{5792} \quad (70.32)$$

Ensuite, comme nous avons un train à quatre roues, nous devons avoir:

$$r \cong \frac{\prod Z_{menardes}}{\prod Z_{menées}} = \frac{Z_1}{Z_2} \cdot \frac{Z_3}{Z_4} \quad (70.33)$$

et nous prenons alors (toujours en nous aidant de la commande ifactor de Maple 4.00b):

$$r \cong \frac{Z_1}{Z_2} \cdot \frac{Z_3}{Z_4} = \frac{7}{32} \cdot \frac{7}{181} \quad (70.34)$$

2. RÉSISTANCE DES MATÉRIAUX (R.D.M.)

La résistance des matériaux (R.D.M. se dit "ResDem" pour les intimes...) est, comme tous les autres chapitres de ce site, un domaine extrêmement vaste dont le niveau de détail et la complexité des calculs peut exploser. Nous allons dans les paragraphes qui suivent nous attarder sur l'essentiel que l'ingénieur (en entreprise) doit savoir. Les développements sont simplifiés à l'extrême pour des cas particuliers triviaux (barres et poutres rectilignes). Dans la réalité, il faut utiliser le calcul tensoriel, les plans d'expérience ou la modélisation informatique avec les MEF (méthodes des éléments finis).

Avant de commencer à étudier quelques cas concrets simples faisons quelques rappels des démonstrations issues du chapitre de Mécanique des Milieux Continus:

Le solide considéré comme rigide n'existe pas, ce n'est qu'une approximation commode. L'expérience montre en effet qu'un solide est toujours légèrement déformable sous l'effet de forces extérieures.

Les relations entre déformations et tensions sont en général compliquées par suite de l'anisotropie des réseaux cristallins. Cependant, les solides n'étant généralement pas des monocristaux mais des substances polycristallines, constituées d'assemblages de microcristaux associés au hasard, ils peuvent être considérés comme isotropes.

Ensuite, il convient de considérer globalement les hypothèses suivantes relativement aux développements qui vont suivre:

H1. La matière est homogène, c'est-à-dire pour rappel de même constitution physique et de même structure dans tout le volume de la pièce.

H2. La matière est isotrope, c'est-à-dire pour rappel que ses propriétés mécaniques sont les mêmes en tout point du corps.

H3. La matière est parfaitement élastique, c'est-à-dire pour rappel qu'après élimination des efforts extérieurs, la pièce reprend immédiatement ses dimensions primitives (au contraire de la limite plastique!).

H4. Les déformations (déplacements des points de la ligne caractéristiques) sont petites par rapport aux dimensions des objets étudiés.

H5. Toute section droite (sections transversales) avant déformation reste droite après déformation (hypotheses de Navier-Bernoulli).

H6. Les résultats obtenus en R.D.M. ne s'appliquent valablement qu'à une distance suffisamment éloignée de la région d'application des efforts concentrés (Hypothèse de Barré de Saint Venant).

H7. Dans le domaine élastique, la matière obéit à la loi de proportionnalité et donc les déformations sont liées par la loi de Hooke démontrée dans le chapitre de Mécanique Des Milieux Continus. Cette loi linéaire permet d'appliquer le principe de superposition des forces et des déformations à la résistance.

Nous avons vus dans le chapitre de Mécanique des Milieux Continus que la loi de Hooke stipule, lorsque les déformations sont réversibles, qu'il y a proportionnalité entre tension et déformation (une des variantes de formulation de la loi de Hooke):

$$F = k |\Delta L| = \frac{E \cdot S}{L_0} |\Delta L| = E \cdot S \cdot \varepsilon = E \cdot S \cdot \varepsilon \quad (70.35)$$

ou:

$$\frac{F}{S} = \sigma = E \cdot \varepsilon \quad (70.36)$$

où E est le module de Young, ε la déformation normale et σ la contrainte normale. Indiquons que le rapport:

$$\frac{E \cdot S}{L_0} \quad (70.37)$$

est souvent appelée "**raideur de la barre**" dans la littérature spécialisée et est souvent notée k .

Nous avons également démontré dans le chapitre de Mécanique des Milieux Continus que la contrainte tangentielle était donnée par:

$$\tau = \frac{E}{(1+2\eta)} \gamma = G \cdot \gamma \quad (70.38)$$

où G est le module de cisaillement, γ est l'angle de déformation et η le coefficient de Poisson, nombre sans dimensions. Nous avons donc une relation entre le module d'élasticité et de rigidité valable dans le cas des petites déformations.

Nous avons vu également dans le même chapitre que pour un solide ou un liquide soumis à une surpression isotrope uniforme nous avons:

$$\frac{\Delta V}{V_0} = \sigma \frac{3(1-2\eta)}{E} = \sigma \cdot \kappa = \kappa \cdot \Delta P \quad (70.39)$$

Le coefficient de compressibilité κ est donc un nombre positif, par conséquent en utilisant la relation précédente, nous avons:

$$\kappa > 0 \Rightarrow (1-2\eta) > 0 \Rightarrow 2\eta < 1 \quad (70.40)$$

et vient alors un résultat connu:

$$\eta < \frac{1}{2} \quad (70.41)$$

Donc le coefficient de Poisson ne peut pas être plus grand que $\frac{1}{2}$ et il peut être négatif (dans ce dernier cas nous parlons alors de matériaux auxétiques).

Enfin, rappelons que nous avons vu dans le chapitre de Mécanique des Milieux Continus que la contraction unitaire selon l'axe z était donnée lors d'une traction selon l'axe x par:

$$\varepsilon_z = -\eta \frac{\sigma_x}{E} = -\eta \varepsilon_x \quad (70.42)$$

Soit autrement écrit (en se concentrant sur le plan XZ):

$$\frac{|\Delta L_z|}{L_{0z}} = -\eta \frac{|\Delta L_x|}{L_{0x}} \quad (70.43)$$

Soit:

$$\eta = -\frac{\frac{|\Delta L_z|}{L_{0z}}}{\frac{|\Delta L_x|}{L_{0x}}} = \frac{(L_z - L_{0z})/L_{0z}}{(L_x - L_{0x})/L_{0x}} \quad (70.44)$$

Et c'est ce que montre la figure ci-dessous:

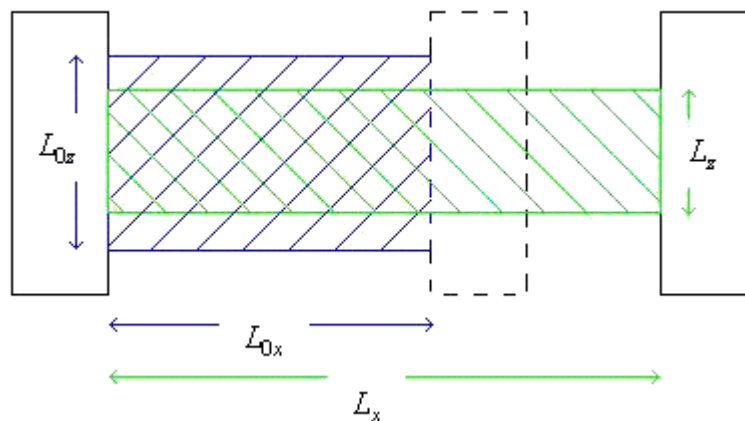


Figure: 70.8 - Traction sur une pièce test

Nous avons également démontré dans le chapitre de Mécanique des Milieux Continus la relation suivante lors de notre étude du module de flexion:

$$M = \frac{E \cdot I}{R} = \frac{E \cdot \int y^2 dS}{R} = \frac{-\sigma I}{y} \quad (70.45)$$

qui exprime le moment de flexion pour une poutre sous un effort M (moment de force), la travée d'écrivant alors un arc de cercle de rayon de courbure R et où I caractérise la "rigidité de forme" du matériau ayant une aire transversale donnée. C'est une relation très importante dans de nombreux domaines de la construction (navale, automobile, architecture, etc.).

Remarque: I est appelé le "moment d'inertie statique" ou "moment quadratique" comme nous l'avons déjà spécifié dans le chapitre de mécanique des milieux continus.

2.1. MOMENTS QUADRATIQUES

Voyons les trois moments d'inerties statiques I_x classiques du domaine de la RDM car souvent rencontrés dans la pratique (construction).

Remarque: La théorie des moments d'inertie est pour rappel présentée dans le chapitre de Mécanique Classique. Dans le chapitre sur les Formes Géométriques nous avons démontré en détails les moments d'inerties des volumes les plus courants.

1. Moment d'inertie statique transversal de la plaque rectangulaire de côté b et hauteur h :

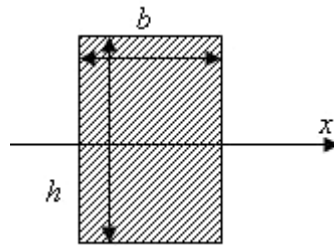


Figure: 70.9 - Plaque rectangulaire et axe d'inertie

Le domaine occupé par la plaque est donné par:

$$S = \{(x, y) \in \mathbb{R}^2 \mid -b/2 \leq x \leq b/2, -h/2 \leq y \leq h/2\} \quad (70.46)$$

Nous avons alors:

$$I_x = \int_{-h/2}^{h/2} \int_{-b/2}^{b/2} y^2 dx dy = \int_{-h/2}^{h/2} y^2 x \Big|_{-b/2}^{b/2} dy = \int_{-h/2}^{h/2} b y^2 dy = \frac{b}{3} y^3 \Big|_{-h/2}^{h/2} = \frac{bh^3}{12} \quad (70.47)$$

2. Moment d'inertie statique transversal d'un disque de diamètre:

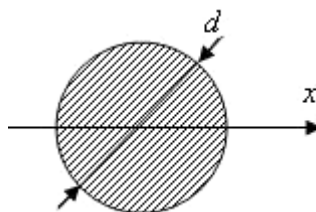


Figure: 70.10 - Disque et axe d'inertie

Ici le domaine d'intégration est:

$$S = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq d^2 / 4\} \quad (70.48)$$

où d est le diamètre du disque.

Nous avons toujours:

$$I_x = \int_S y^2 dx dy \quad (70.49)$$

Pour calculer cette intégrale, nous utilisons les coordonnées polaires:

$$\begin{cases} x = r \cos \theta \\ y = r \sin \theta \end{cases} \quad (70.50)$$

Le jacobien de la transformation est égal à r (cf. chapitre de Calcul Différentiel Et Intégral). Nous obtenons:

$$I_x = \int_0^{d/2} \int_0^{2\pi} r^3 (\sin \theta)^2 d\theta dr = \pi \int_0^{d/2} r^3 dr = \frac{\pi d^4}{64} \quad (70.51)$$

3. Moment d'inertie statique d'une couronne de diamètre extérieur D et diamètre intérieur d :

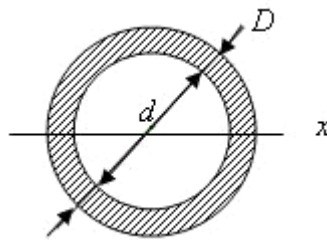


Figure: 70.11 - Couronne et axe d'inertie

Ici le domaine d'intégration est:

$$S = \{(x, y) \in \mathbb{R}^2 \mid d^2 / 4 \leq x^2 + y^2 \leq D^2 / 4\} \quad (70.52)$$

où D et d sont respectivement les diamètres du grand et du petit disque.

Si nous notons S_1 le domaine du grand disque et S_2 celui du petit disque alors:

$$I_x = \int_{S_1} y^2 dx dy - \int_{S_2} y^2 dx dy = \frac{\pi(D^4 - d^4)}{64} \quad (70.53)$$

en utilisant le moment d'inertie statique du disque.

Pour résumer, nous avons donc:

$$\begin{aligned} I_x &= \int_S y^2 ds = \iint_S y^2 dx dy \\ I_y &= \int_S x^2 ds = \iint_S x^2 dx dy \end{aligned} \quad (70.54)$$

et enfin il existe aussi le moment quadratique polaire de S par rapport à un point O :

$$I_0 = \int_S r^2 ds = \iint_S (x^2 + y^2) dx dy = I_x + I_y \quad (70.55)$$

Il est donc aisé dans des cas simples de connaître le moment d'inertie polaire et celui-ci est très utile dans le cadre de l'étude de la torsion.

Il découle de ces outils que plus les éléments de la section sont situés loin de l'axe, plus le moment quadratique sera important et plus (nous le démontrerons dans ce qui suit) les flèches seront faibles.

2.2. ÉQUATION DE LA LIGNE ÉLASTIQUE

Pour cet exemple de cas d'école mais très utilisé dans la pratique nous allons d'abord devoir obtenir mathématiquement la forme géométrique que prend la fibre neutre d'une poutre soumise à des efforts de flexion.

Remarque: Si toutes les réactions d'un système sollicité se trouvent à partir des équations d'équilibre statique, le problème est dit "**statiquement déterminé**" ou "**isostatique**". Lorsque les équations de la statique ne permettent pas de trouver l'équilibre d'un système, le système est dit "**hyperstatique**" ou "**statiquement indéterminé**". Nous disons aussi que tant qu'il y a une "liberté de mouvement" malgré n points de liaison, nous sommes en isostatique.

En faisant l'hypothèse que les déformations sont faibles et que le poids de la poutre est faible devant la force qui plie la poutre, nous pouvons faire le schéma suivant:

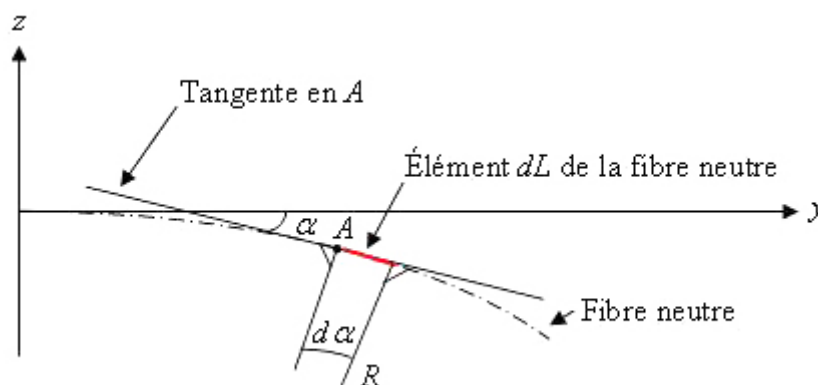


Figure: 70.12 - Poutre (simplifiée...) soumise à un effort

Par définition de la dérivée et en vertu de l'hypothèse des faibles déformations (cela fonctionne donc quand même bien jusqu'à 45°...):

$$\frac{dz}{dy} = \tan \alpha \cong \alpha \quad (70.56)$$

Soit en dérivant encore une fois:

$$\frac{d^2 z}{dy^2} = \frac{d\alpha}{dy} \quad (70.57)$$

D'autre part, la figure montre que (cf. chapitre de Trigonométrie):

$$dL = R d\alpha \Rightarrow \frac{1}{R} = \frac{d\alpha}{dL} \quad (70.58)$$

Mais du fait que la courbe de la fibre neutre s'écarte peu de l'axe y (déformations faibles), nous pouvons écrire:

$$dy \cong dL \quad (70.59)$$

Donc:

$$\frac{1}{R} \cong \frac{d\alpha}{dy} \quad (70.60)$$

Ainsi, nous pouvons écrire en utilisant les relations obtenues plus haut:

$$\boxed{\frac{d^2 z}{dy^2} = \frac{1}{R} = \frac{M}{E \cdot I}} \quad (70.61)$$

qui est donc l'équation différentielle donnant $z = f(y)$, appelée "équation de la ligne élastique".

Une autre approche courante mais moins intuitive pour obtenir la même relation est de partir de:

$$M = \frac{E \cdot I}{R} \quad (70.62)$$

et de rappeler que le rayon de courbure R est donné par (cf. chapitre de Géométrie Différentielle):

$$R = \frac{1}{C} = \frac{\left(1 + \left(\frac{dy}{dx}\right)^2\right)^{3/2}}{\left|\frac{d^2 y}{dx^2}\right|} = \frac{(1 + y'^2)^{3/2}}{|y''|} \quad (25.63)$$

Soit en adaptant l'écriture à notre contexte:

$$M = \frac{E \cdot I}{R} = \frac{E \cdot I}{\frac{(1 + z'^2)^{3/2}}{|z''|}} \quad (25.64)$$

et en négligeant la dérivée première pour les faibles déformations nous retrouvons bien:

$$M = \frac{E \cdot I}{R} = \frac{E \cdot I}{\frac{(1+z'^2)^{3/2}}{|z''|}} \cong E \cdot I |z''| = E \cdot I \frac{d^2 z}{dy^2} \quad (25.65)$$

Exemples:

E1. Poutre encastrée que d'un seul côté dite "poutre en porte à faux" (cas classique dans la construction et les habitations) avec charge concentrée (ponctuelle) à l'extrémité:

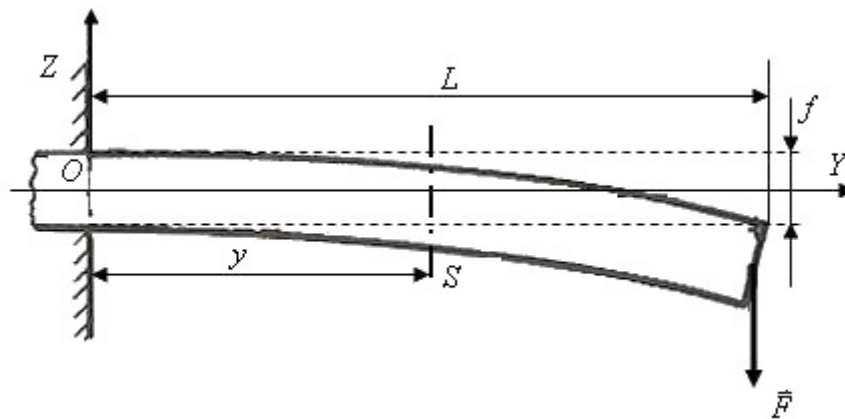


Figure: 70.13 - Poutre encastrée d'un seul côté avec force ponctuelle

Dans la section S quelconque, le moment de force (de flexion) vaut donc:

$$M(y) = F \cdot (L - y) = \frac{E \cdot I}{R} \Rightarrow \frac{F \cdot (L - y)}{E \cdot I} = \frac{1}{R} \quad (70.66)$$

D'autre part:

$$\frac{d^2 z}{dy^2} = \frac{1}{R} \quad (70.67)$$

En éliminant R entre ces deux relations, il reste:

$$\frac{d^2 z}{dy^2} = \frac{F}{E \cdot I} (L - y) \quad (70.68)$$

La figure montre que les conditions aux limites sont:

$$y = 0 \Rightarrow \begin{cases} z = 0 \\ \frac{dz}{dy} = 0 \end{cases} \quad (70.69)$$

Nous tirons après intégration:

$$\begin{aligned}\frac{dz}{dy} &= \frac{F}{E \cdot I} \left(Ly - \frac{1}{2} y^2 \right) \\ z &= \frac{F}{E \cdot I} \left(L \frac{1}{2} y^2 - \frac{1}{2} \frac{1}{3} y^3 \right)\end{aligned}\quad (70.70)$$

Soit:

$$z = \frac{F \cdot L}{E \cdot I} \cdot \frac{y^2}{2} - \frac{F}{2 \cdot E \cdot I} \cdot \frac{y^3}{3} \quad (70.71)$$

Si $y = L$ la déformation est maximale et z prend donc la valeur maximale f appelée la "flèche". Il s'ensuit:

$$z = f = \frac{F \cdot L}{E \cdot I} \cdot \frac{L^2}{2} - \frac{F}{2 \cdot E \cdot I} \cdot \frac{L^3}{3} = \frac{F \cdot L^3}{3 \cdot E \cdot I} \quad (70.72)$$

Soit la déformation verticale maximale de l'extrémité pour une poutre encastree que d'un côté est finalement:

$$f = \frac{F \cdot L^3}{3 \cdot E \cdot I} \quad (70.73)$$

Toutes les données de cette relation nous sont connues (force, longueur, module de Young, inertie statique) et il est alors possible de déterminer si la barre va casser ou non car il suffit d'appliquer la relation démontrée plus haut:

$$M = \frac{-\sigma \cdot I}{y} \Rightarrow |\sigma| = \frac{y \cdot M}{I} = \frac{f \cdot M}{I} \quad (70.74)$$

Et donc grâce à la relation:

$$\sigma = \frac{f \cdot M}{I} \quad (70.75)$$

et sachant expérimentalement à partir de quelle valeur expérimentale de σ le matériau casse on saura quand la barre cassera (du moins approximativement!) en connaissant la flèche f , le moment fléchissant M et le moment quadratique I .

Nous avons donc un résultat qui va nous être utile par la suite:

$$df = F \frac{1}{E \cdot I} y^2 dy \quad (70.76)$$

et en intégrant de 0 à L nous retrouvons la flèche de notre poutre précédente!

E2. La poutre soutenue (appelée aussi "poutre à 2 appuis simples" ou "poutre isostatique") est l'exemple le plus classique en construction et donc en architecture (et ceux qui ont joué pendant l'enfance à mettre des lattes de bois pour passer par dessus une petite rivière). Il s'agit d'une poutre homogène, de section constante, reposant sur deux appuis libres à ses extrémités et soumis à une charge F en son centre:

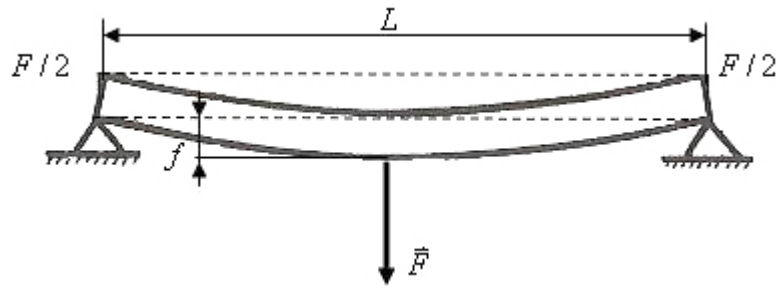


Figure: 70.14 - Poutre soutenue avec force ponctuelle centrée

Nous pouvons donc considérer que tout se passe comme si nous avions $F/2$ aux deux extrémités de deux poutres de longueur $L/2$ (ainsi la somme vaut bien F , c'est-à-dire le moment fléchissant). Remarquons que nous négligeons le poids de la poutre devant F , mais F peut être tout simplement le poids de la poutre! En utilisant la dernière relation de l'exemple précédent, nous avons:

$$df = \frac{F}{2} \frac{1}{E \cdot I} y^2 dy \quad (70.77)$$

Soit:

$$f = \frac{F}{2} \frac{1}{E \cdot I} \int_0^{L/2} y^2 dy = \frac{F \cdot L^3}{48 \cdot E \cdot I} \quad (70.78)$$

Soit la déformation verticale maximale d'une poutre posée des deux côtés est finalement (en changeant un peu la notation):

$$f = \frac{F \cdot L^3}{48 \cdot E \cdot I} \quad (70.79)$$

Ainsi, pour une même longueur de poutre, à F identique la flèche est donc 16 fois moindre que pour une poutre encastree! Il était intuitif qu'elle soit moins élevée pour une force identique mais difficile de deviner qu'elle le serait d'un facteur 16!

C'est cette relation qui est aussi utilisée pour les poutres IPN (fameuses en construction!).

E3. Poutre encastree que d'un seul côté dite "poutre en porte à faux" (cas classique dans la construction et les habitations) avec charge linéique constante w :

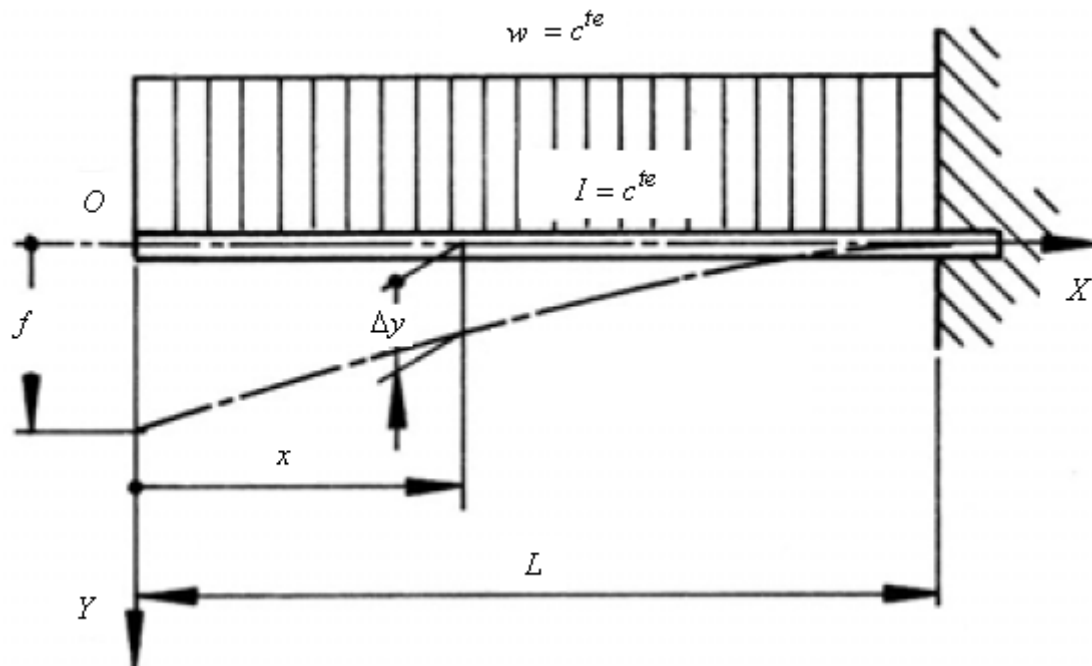


Figure: 70.15 - Poutre encastrée d'un seul côté avec force linéique constante

Le développement est simple mais certaines simplifications sont astucieuses afin d'obtenir un résultat élégant.

Nous partons toujours de l'équation de la ligne élastique en adoptant les écritures à la configuration choisie:

$$\frac{d^2 y}{dx^2} = \frac{M}{EI} \quad (70.80)$$

Soit explicitement:

$$\frac{d^2 y}{dx^2} = \frac{\int F dx}{EI} = \frac{\int \left(\int q dx \right) dx}{EI} = \frac{\int w x dx}{EI} = \frac{w \frac{x^2}{2} + c^{te}}{EI} \quad (70.81)$$

Comme le moment de force M est nul en $x = 0$, nous avons la constante qui est nulle. Dès lors:

$$\frac{d^2 y}{dx^2} = \frac{w \frac{x^2}{2}}{EI} \quad (70.82)$$

En intégrant encore une fois, il vient:

$$y' = \frac{dy}{dx} = \frac{w}{EI} \int \frac{x^2}{2} dx = \frac{w}{EI} \frac{x^3}{6} + c^{te} \quad (70.83)$$

Comme pour $x = L$ par hypothèse la déformation est nulle, nous avons la constante qui est alors donnée par:

$$y' = \frac{w}{EI} \frac{L^3}{6} + c^{te} = 0 \Rightarrow c^{te} = -\frac{wL^3}{6EI} \quad (70.84)$$

Soit:

$$y' = \frac{w}{EI} \frac{x^3}{6} - \frac{wL^3}{6EI} = \frac{wL^3}{6EI} \left(\left(\frac{x}{L} \right)^3 - 1 \right) \quad (70.85)$$

En intégrant encore une dernière fois:

$$y = \frac{w}{6EI} \int (x^3 - L^3) dx = \frac{w}{6EI} \left(\frac{x^4}{4} - L^3 x \right) + c^{te} \quad (70.86)$$

Et comme en $x = L$ nous avons y qui est aussi nul, il vient pour la constante:

$$y = \frac{w}{6EI} \left(\frac{L^4}{4} - L^3 L \right) + c^{te} = 0 \Rightarrow c^{te} = \frac{wL^4}{8EI} \quad (70.87)$$

Soit:

$$y = \frac{w}{6EI} \left(\frac{x^4}{4} - L^3 x \right) + \frac{wL^4}{8EI} \quad (70.88)$$

et comme la flèche est de toute façon en $x = 0$, nous avons alors:

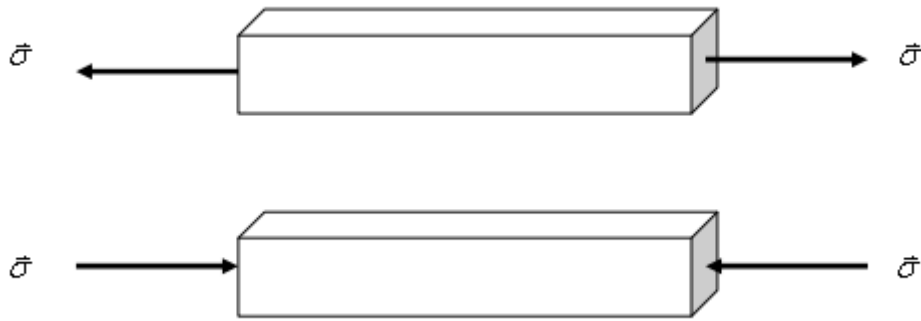
$$f = \frac{wL^4}{8EI} \quad (70.89)$$

2.2.1. ÉQUATION DES POUTRES (EULER-BERNOULLI)

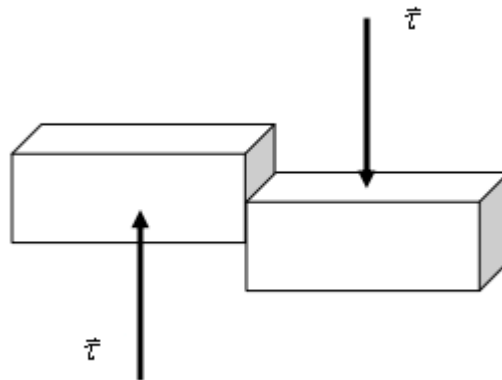
Considérons maintenant le cas d'une poutre encastree des deux côtés (cas encore plus courant que les deux exemples précédents!). L'analyse va être un peu plus difficile et il va nous falloir introduire plusieurs concepts.

Une poutre en pratique doit résister aux efforts suivants:

- Tension ou compression:



- Cisaillement (effort tranchant):



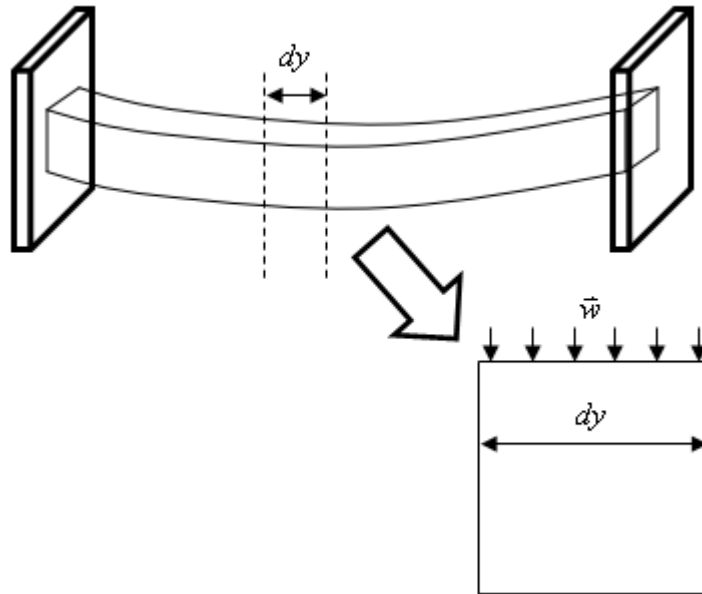
- Flexion (effort fléchissant):



Si une poutre est en équilibre, alors les efforts internes doivent satisfaire en tout point:

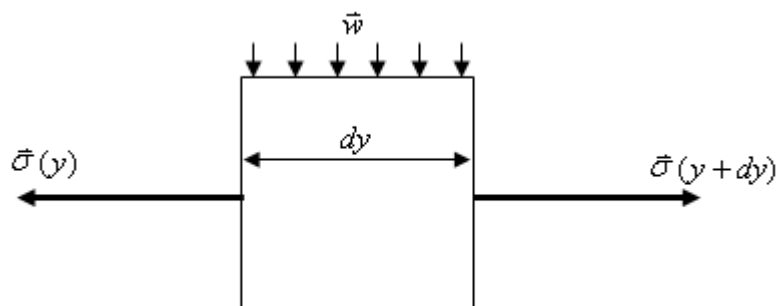
$$\begin{aligned}
 \sum \vec{F}_{\text{horizontales}} &= \sum \vec{F}_{\text{Tension/Compression}} = \vec{0} \\
 \sum \vec{F}_{\text{verticales}} &= \sum \vec{F}_{\text{Cisaillement}} = \vec{0} \\
 \sum \vec{M} &= \sum \vec{M}_{\text{Flexion}} = \vec{0}
 \end{aligned}
 \quad (70.90)$$

Considérons maintenant une poutre encastree à ses deux extrémités (poutre bi-encastree) et prenons en un tranche de longueur infinitésimale dy telle que localement sa courbure soit nulle. La poutre sera supposée soumise à une force (poids) uniforme sur toute sa longueur (force qui peut aussi être assimilée à son propre poids comme nous l'avons déjà précisé plus haut). Il est alors d'usage de noter $\vec{w}$ la force par unité de longueur (poids total divisé par la longueur) qui est bien évidemment une charge linéique:



Si la poutre est à l'équilibre une fois déformée (faiblement ou beaucoup déformée cela importe peu!) alors les sommes des forces de tension, compression, cisaillement et flexion doivent être nulles en chaque point comme nous l'avons déjà dit! Cela ne veut cependant pas dire qu'en chaque point de la ligne élastique les valeurs de chacun des forces est égale! Bien au contraire. Il y a bien évidemment des différences (sinon quoi il n'y aurait pas déformation).

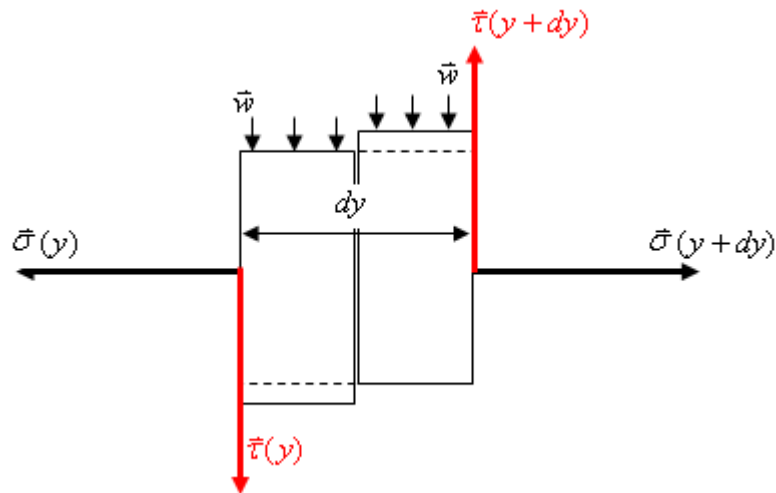
Faisons alors d'abord la somme des forces locales de tension et de compression (horizontales) de l'élément de longueur dy . Nous avons alors schématiquement:



Soit algébriquement (le variationnel pouvant être négatif ou positif peu importe!):

$$\sum \vec{F}_{horizontales} = \vec{\sigma}(y+dy) - \vec{\sigma}(y) = \vec{\sigma} + d\vec{\sigma} - \vec{\sigma} = d\vec{\sigma} = \vec{0} \quad (70.91)$$

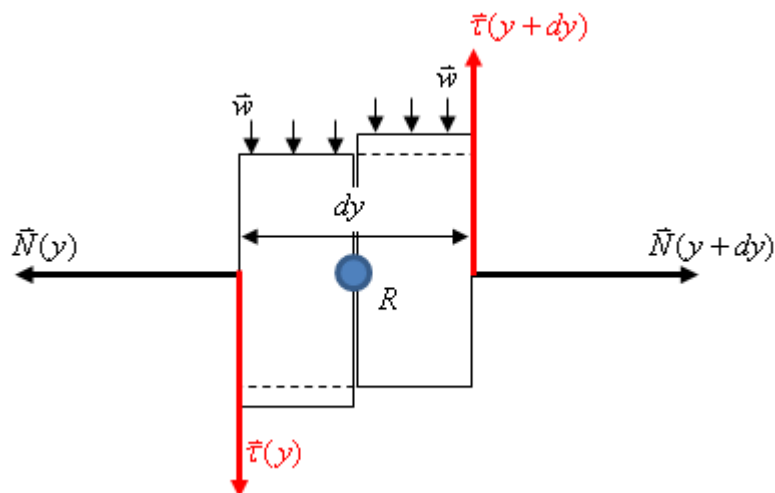
Si maintenant nous nous occupons des forces verticales à la source du cisaillement. Nous avons alors schématiquement:



Soit algébriquement (le variationnel pouvant être négatif ou positif peu importe!):

$$\sum \vec{F}_{\text{verticales}} = \vec{T}(y+dy) - \vec{T}(y) = \vec{T} + d\vec{T} - \vec{T} - \vec{w}dy = d\vec{T} - \vec{w}dy = \vec{0} \quad (70.92)$$

Et finalement pour les moments de flexion la somme est aussi forcément nulle à l'équilibre. Cependant contrairement aux deux sommes algébriques précédentes où nous pouvions utiliser seulement le différentiel, nous devons pour les moments de flexion choisir un point R de référence puisque pour rappel... le moment de force est par définition le produit d'une force par une distance. Nous choisirons donc naturellement le centre de gravité:



La charge linéaire uniforme $\vec{w}$ sur la longueur dy génère une de force à mi-distance de:

$$\vec{F}_{\vec{w}} = (\vec{w}(y)dy) \frac{1}{2} dy = \frac{1}{2} \vec{w}(y) dy^2 \quad (70.93)$$

mais comme elle est confondue aux choix de notre repère, alors son moment de force est nul!

Nous avons alors algébriquement:

$$\begin{aligned}
\sum \vec{M} &= \vec{M}(y) + \vec{M}(-y) + \vec{\tau}(y) \frac{dy}{2} + \vec{\tau}(-y) \frac{dy}{2} \\
&= \underbrace{\vec{M} + d\vec{M}}_{=\vec{M}(y)} - \underbrace{\vec{M}}_{=\vec{M}(-y)} + \underbrace{(\vec{\tau} + d\vec{\tau})}_{=\vec{\tau}(y)} \frac{dy}{2} + \underbrace{\vec{\tau}}_{=\vec{\tau}(-y)} \frac{dy}{2} = d\vec{M} + \vec{\tau} \frac{dy}{2} + d\vec{\tau} \frac{dy}{2} + \vec{\tau} \frac{dy}{2} \quad (70.94) \\
&= d\vec{M} + \vec{\tau} dy + d\vec{\tau} \frac{dy}{2} = \vec{0}
\end{aligned}$$

et si nous négligeons les différentiels d'ordre deux:

$$d\vec{M} + \vec{\tau} dy = \vec{0} \quad (70.95)$$

Nous avons donc au final:

$$\left. \begin{aligned} d\vec{\sigma} &= \vec{0} \\ d\vec{\tau} - \vec{w} dy &= \vec{0} \\ d\vec{M} + \vec{\tau} dy &= \vec{0} \end{aligned} \right\} \Leftrightarrow \begin{cases} \frac{d\vec{\sigma}}{dy} = \vec{0} \\ \frac{d\vec{\tau}}{dy} = \vec{w} \\ \frac{d\vec{M}}{dy} = -\vec{\tau} \end{cases} \quad (70.96)$$

Pour déterminer le moment de fléchissement à partir de la charge linéaire (ce qui intéresse en priorité le praticien), nous dérivons deux fois la troisième relation et faisons une substitution:

$$\frac{d}{dy} \left(\frac{d\vec{M}}{dy} \right) = -\frac{d}{dy} \vec{\tau} = -\vec{w} \quad (70.97)$$

Soit:

$$\frac{d^2 \vec{M}}{dy^2} = -\vec{w} \quad (70.98)$$

Le problème avec cette dernière relation est la connaissance des moments. Il faudrait nous débarrasser absolument de ce terme. Ce que nous connaissons facilement c'est la fonction de déformation et nous avons démontré plus haut que:

$$\frac{d^2 z}{dy^2} = \frac{1}{R} = \frac{M}{E \cdot I} \quad (70.99)$$

Il vient alors immédiatement en substituant la relation précédente dans l'antéprécédente:

$$EI \frac{d^4 z}{dy^4} = -w(y) \quad (70.100)$$

Il s'agit de la relation la plus important de la théorie des poutres car elle permet en connaissance la charge linéaire de déterminer la fonction de déformation ou inversement! Elle est tellement important qu'on l'appelle "**équation des poutres**" ou encore en honneur à ceux qui l'ont déterminé: "**équation d'Euler-Bernoulli**".

Comme il s'agit d'une équation différentielle d'ordre 4 qui va générer quatre constantes à chaque intégration, il nous faudra alors 4 conditions initiales pour la résoudre complètement.

Exemple:

Nous souhaiterions calculer la déformation d'une poutre fixée des deux côtés et chargée uniformément connaissant sa longueur L , son module d'élasticité E , son moment d'inertie I . Nous partons alors (nous changeons les notations pour montrer que selon les ouvrages les axes pris peuvent être notés différemment):

$$EI \frac{d^4 y}{dx^4} = -w(x) \quad (70.101)$$

avec les conditions initiales:

$$y(x=0) = 0, \quad y(x=L) = 0, \quad \frac{dy}{dx}(x=0) = 0, \quad \frac{dy}{dx}(x=L) = 0 \quad (70.102)$$

Et intégrant à répétition, nous avons:

$$\begin{aligned} \frac{d^3 y}{dx^3} &= \frac{-wx}{EI} + C_1 \Rightarrow \frac{d^2 y}{dx^2} = \frac{-wx^2}{2EI} + C_1 x + C_2 \Rightarrow \frac{dy}{dx} = \frac{-wx^3}{6EI} + \frac{C_1 x^2}{2} + C_2 x + C_3 \\ \Rightarrow y(x) &= \frac{-wx^4}{24EI} + \frac{C_1 x^3}{6} + \frac{C_2 x^2}{2} + C_3 x + C_4 \end{aligned} \quad (70.103)$$

Des deux conditions initiales:

$$y(x=0) = 0, \quad \frac{dy}{dx}(x=0) = 0 \quad (70.104)$$

Il en découle que:

$$C_3 = C_4 = 0 \quad (70.105)$$

Avec les deux autres conditions restantes, nous obtenons le système suivant qui doit nous permettre de déterminer les deux constantes restantes:

$$\begin{cases} \frac{-wL^4}{24EI} + \frac{C_1 L^3}{6} + \frac{C_2 L^2}{2} = 0 \\ \frac{-wL^3}{6EI} + \frac{C_1 L^2}{2} + C_2 L = 0 \end{cases} \quad (70.106)$$

Après il s'agit simplement de résoudre un simple système linéaire (cf. chapitre d'Algèbre Linéaire):

$$\begin{cases} \frac{-wL^4}{24EI} + \frac{C_1L^3}{6} + \frac{C_2L^2}{2} = 0 \\ \frac{-wL^3}{6EI} + \frac{C_1L^2}{2} + C_2L = 0 \end{cases} \cdot \begin{cases} \frac{24}{L^2} \\ \frac{6}{L} \end{cases} \Rightarrow \begin{cases} \frac{-wL^2}{EI} + 4C_1L + 12C_2 = 0 \\ \frac{-2wL^2}{EI} + 6C_1L + 12C_2 = 0 \end{cases} \Rightarrow \begin{cases} C_1 = \frac{wL}{2EI} \\ C_2 = \frac{-wL^2}{12EI} \end{cases} \quad (70.107)$$

Donc une poutre bi-encastree idéale soumise à une charge uniforme et décrite par:

$$y(x) = \frac{-wx^4}{24EI} + \frac{wLx^3}{12EI} - \frac{wL^2x^2}{24EI} = \frac{-wx^2}{24EI}(x^2 - 2Lx + L^2) = \frac{-wx^2}{24EI}(x - L)^2 \quad (70.108)$$

Pour déterminer la flèche, il nous faut donc chercher le point x où cette relation a un optimum. Nous avons alors:

$$f'(x) = [x^2(x - L)^2]' = 2x(x - L)^2 + 2x^2(x - L) = 2x(x - L)(x - L + x) = 2x(x - L)(2x - L) = 0 \quad (70.109)$$

Il s'ensuit que la flèche à un maximum en $x = L/2$. En injectant cela dans $y(x)$ nous obtenons la fameuse relation souvent donnée dans la littérature mais rarement démontrée:

$$f = \frac{w}{24EI} \cdot \frac{L^4}{16} = \frac{wL^4}{384EI} \quad (70.110)$$

Il s'ensuit que la flèche d'une poutre est proportionnelle à la puissance quatrième de la longueur de la poutre! Une si forte dépendance impose des limitations significatives dans les constructions civiles basées sur des poutres.

Remarque: La charge linéique constante, soit sur la longueur totale de la poutre, soit par tronçons successifs, est une sollicitation fréquente dans les pièces à axe horizontal. Elle peut provenir du poids propre de la pièce à section constante ou d'une charge provoquée par une poussée extérieure (gravité par exemple).

2.2.2. ÉNERGIE POTENTIELLE ÉLASTIQUE

Après un survol rapide des diverses déformations élastiques des pièces sollicitées par les efforts fondamentaux, nous allons établir ici l'expression générale de l'énergie élastique accumulée dans une barre de forme quelconque sollicité par des efforts extérieurs.

Rappelons pour cette étude que nous pouvons écrire la loi de Hooke (cf. chapitre de Mécanique Des Milieux Continus) sous la forme suivante:

$$|\Delta L| = \frac{F \cdot L_0}{E \cdot S} = \frac{F}{k} \quad (70.111)$$

et l'énergie potentielle élastique d'un ressort démontrée dans le chapitre de Mécanique Classique est donnée par:

$$E_{\text{Elastique}} = \frac{1}{2} kx^2 \quad (70.112)$$

ou en déplacement relatif:

$$E_{\text{Elastique}} = \frac{1}{2} k \Delta x^2 \quad (70.113)$$

Dans le domaine de traction (ou compression) longitudinale des barres, il est d'usage de considérer la barre comme un ressort (...) et alors d'utiliser la constante de raideur de la loi de Hook... en espérant que cela colle à l'expérience:

$$E_{\text{Elastique}} = \frac{1}{2} k \Delta x^2 = \frac{1}{2} \frac{E \cdot S}{L_0} \Delta x^2 = \frac{E \cdot S}{2 L_0} \Delta L^2 \quad (70.114)$$

où conformément à l'usage, nous notons le déplacement longitudinal L au lieu de x . En injectant la loi de Hook (oui... ça tourne un peu en rond...):

$$E_{\text{Elastique}} = \frac{E \cdot S}{2 L_0} \left(\frac{F \cdot L_0}{E \cdot S} \right)^2 = \frac{F^2 \cdot L_0}{2 \cdot E \cdot S} \quad (70.115)$$

Il suffit alors de diviser par la longueur de la barre pour avoir l'énergie linéique élastique:

$$E_{\text{Elastique,linéique}} = \frac{F^2}{2 \cdot E \cdot S} \quad (70.116)$$

L'énergie volumique, notée traditionnellement comme en thermodynamique par une minuscule, est alors:

$$e = \frac{F^2 \cdot L_0}{2 \cdot E \cdot S \cdot V} = \frac{F^2 \cdot L_0}{2 \cdot E \cdot S \cdot S \cdot L_0} = \frac{F^2}{2 \cdot E \cdot S^2} = \left(\frac{F}{S} \right)^2 \frac{1}{2E} = \frac{\sigma^2}{2E} \quad (70.117)$$

2.3. TORSION

Rappelons au lecteur d'abord une étude faite dans le chapitre de Mécanique Classique sur le pendule de torsion où certains éléments avaient volontairement tus. Étudions cela plus en détails car très utile pour les arbres de transmission ou les ressorts dans la vie de tous les jours.

Considérons maintenant un fil cylindrique fixé en sa base et soumis à un moment de torsion $\vec{M}$. Sous l'effet de ce moment de torsion, la face supérieure du fil est décalée d'un angle θ par rapport à la face inférieure, la matière subissant une tension de torsion (ou cisaillement τ):

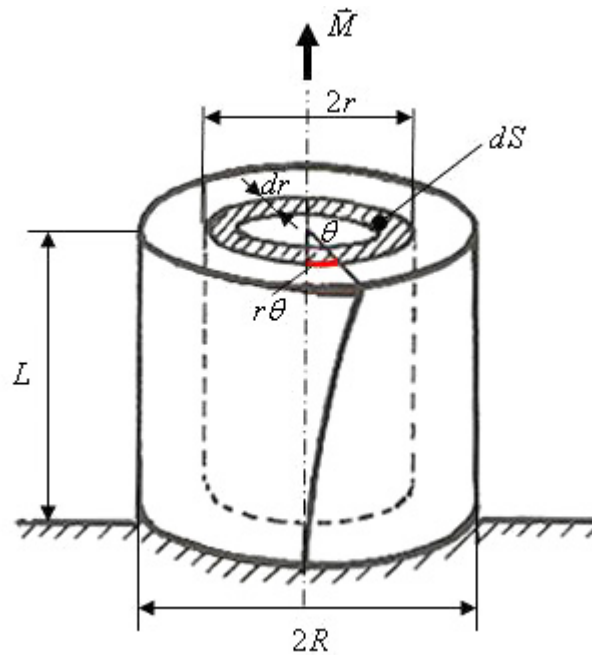


Figure: 70.16 - Fil sous torsion

Imaginons à l'intérieur du fil un tube élémentaire de rayon r , d'épaisseur dr , et observons l'effet de la torsion sur ce tube déroulé (cela nous permettra une approche approximative du phénomène intéressé):

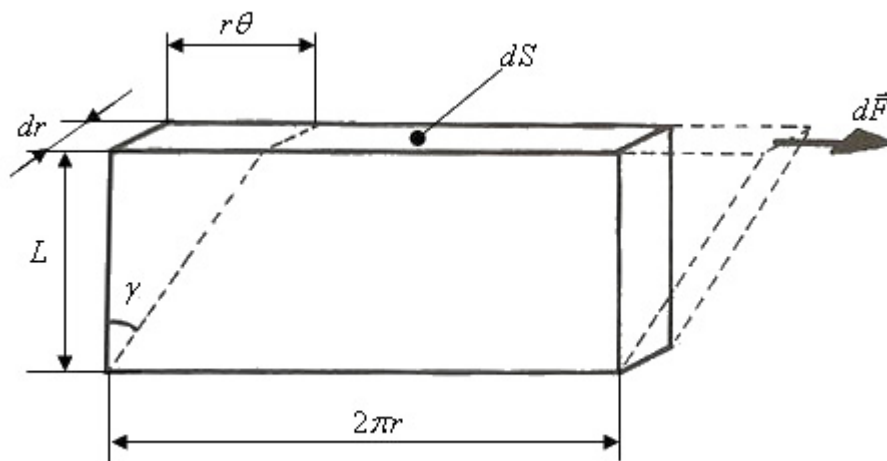


Figure: 70.17 - Extraction un élément du fil sous torsion

Cherchons une relation entre le moment de torsion $\vec{M}$ et l'angle de torsion θ .

Pour le tube déroulé, appliquons les relations du cisaillement:

$$G \cdot \gamma = \tau = \frac{dF}{dS} \Rightarrow dF = \gamma G dS \quad (70.118)$$

or la figure montre que (les déformations étant faibles) au premier ordre en série de Taylor (cf. chapitre sur les Suites et Séries):

$$\gamma = \arctan\left(\frac{r \cdot \theta}{L}\right) \cong \frac{r \cdot \theta}{L} \quad (70.119)$$

d'où:

$$d\vec{F} = \frac{r \cdot \theta}{L} G dS \quad (70.120)$$

Le moment élémentaire dû à cette force est par définition du moment de force:

$$d\vec{M} = \vec{r} \times d\vec{F} \quad (70.121)$$

Soit puisque $\vec{r}$ et $d\vec{F}$ sont perpendiculaires:

$$dM = \frac{\theta \cdot G}{L} r^2 dS \quad (70.122)$$

Le moment total vaut alors:

$$M = \int_{r=0}^{r=R} dM = \frac{\theta \cdot G}{L} \int_{r=0}^{r=R} r^2 dS = \frac{\theta \cdot G}{L} I_0 \quad (70.123)$$

donc:

$$M = \left(\frac{G \cdot I_0}{L} \right) \theta = - \left(- \frac{G \cdot I_0}{L} \right) \theta = -k\theta \quad (70.124)$$

Nous retrouvons donc la relation du pendule de torsion que nous avons posé lors de notre étude du pendule de torsion dans le chapitre de Mécanique Classique avec comme différence que cette fois la constante k , la "**constante de torsion**" est explicite!!!!

Le numérateur de la constante k est appelé dans le domaine dans la résistance des matériaux "**rigidité torsionnelle**" et la constante k elle-même est souvent appelée "**raideur de l'arbre**" au lieu de "constante de torsion". Dans la pratique on cherche surtout à trouver la valeur numérique de l'expression suivante:

$$\theta = \frac{M \cdot L}{G \cdot I_0} \quad (70.125)$$

puisque cela donnera l'amplitude angulaire de la torsion.

Voyons donc une application très importante au ressort de compression de type hélicoïdal (l'approche est approximative à nouveau à défaut de mieux...) travaillant en torsion.

D'abord il faut bien se rendre compte que lorsqu'une force est appliquée au ressort, les extrémités vont tourner d'un angle θ alpha faible (torsion) correspondant au parcours d'une distance x qui elle-même correspond au rétrécissement du ressort (ben oui! il faut bien que cette longueur soit reprise quelque part).

Soit alors un ressort de rayon extérieur R (soit de diamètre D), de module de cisaillement G , avec un diamètre de corps d (diamètre du cylindre plié dont est composé le ressort):

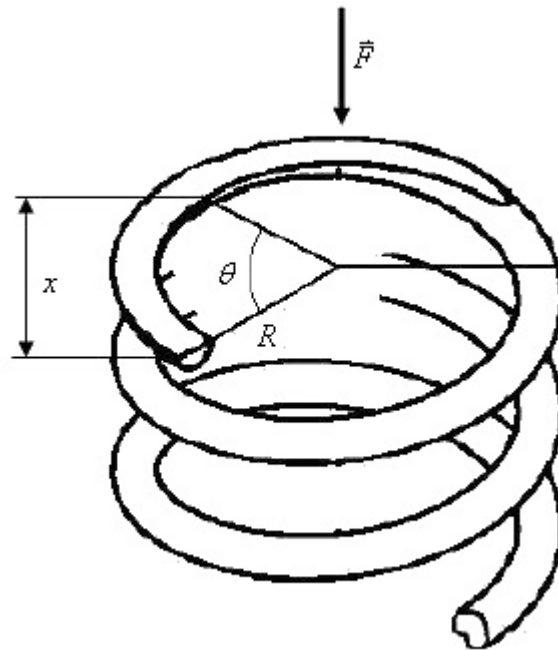


Figure: 70.18 - Ressort spiral sous effort

Pour l'analyse nous aurons besoin simplement de mélanger plusieurs de relations démontrées jusqu'à maintenant. En premier lieu l'angle de torsion d'une poutre de longueur L (longueur du ressort en l'occurrence!):

$$\theta = \frac{M \cdot L}{G \cdot I_0} \quad (70.126)$$

Avec:

$$I_0 = I_x + I_y = \frac{\pi d^4}{64} + \frac{\pi d^4}{64} = \frac{\pi d^4}{32} \quad (70.127)$$

et:

$$L = n \cdot \pi \cdot D \quad (70.128)$$

Par ailleurs, le moment de torsion s'écrit:

$$M = F \cdot R = F \cdot \frac{D}{2} \quad (70.129)$$

Nous arrivons donc à:

$$\theta = \frac{M \cdot L}{G \cdot I_0} = \frac{F \cdot \frac{D}{2} \cdot n \cdot \pi \cdot D}{G \cdot \frac{\pi d^4}{32}} = \frac{16 \cdot F \cdot n \cdot D^2}{G \cdot d^4} \quad (70.130)$$

Remarque: Le rapport M / θ , au même titre que pour l'arbre, est appelé la "**raideur du ressort**".

Le déplacement (déformation) x vaut lui (cf. chapitre de Trigonométrie):

$$x = R \cdot \theta = \frac{D}{2} \theta \quad (70.131)$$

Nous arrivons finalement à:

$$x = \frac{D}{2} \frac{16 \cdot F \cdot n \cdot D^2}{G \cdot d^4} = \frac{8 \cdot F \cdot n \cdot D^3}{G \cdot d^4} \quad (70.132)$$

ce qui nous amène à la relation mondialement connue dans le monde dans la R.D.M. en ce qui concerne les ressorts:

$$F = \left(\frac{G \cdot d^4}{8 \cdot n \cdot D^3} \right) x = k \cdot x \quad (70.133)$$

où k est la constante de "raideur du ressort"!! Si maintenant, nous utilisons l'expression de l'énergie potentielle élastique d'un ressort démontrée dans le chapitre de Mécanique Classique:

$$E_{\text{Elastique}} = \frac{1}{2} k x^2 \quad (70.134) \quad (70.135)$$

nous pouvons alors déterminer l'énergie qu'un ressort spiral peut absorber.

2.4. FLAMBAGE

Nous terminons cette étude de la R.D.M. avec le flambage (cas d'étude classique en construction et mécanique) qui consiste à déterminer (dans un cas particulier simple) la force minimale F_0 à partir de laquelle une barre de longueur L , de module de Young E fixée à ses deux extrémités peut plier (avec un rayon R) jusqu'à casser sans qu'il y ait besoin de trop augmenter la force F_0 (il s'agit donc à nouveau d'une valeur d'indication!).

Dans l'étude de ce phénomène, nous considérons que dès que la barre commence à plier nous avons alors F_0 (et nous ne sommes alors plus très loin de la force permettant de la casser).

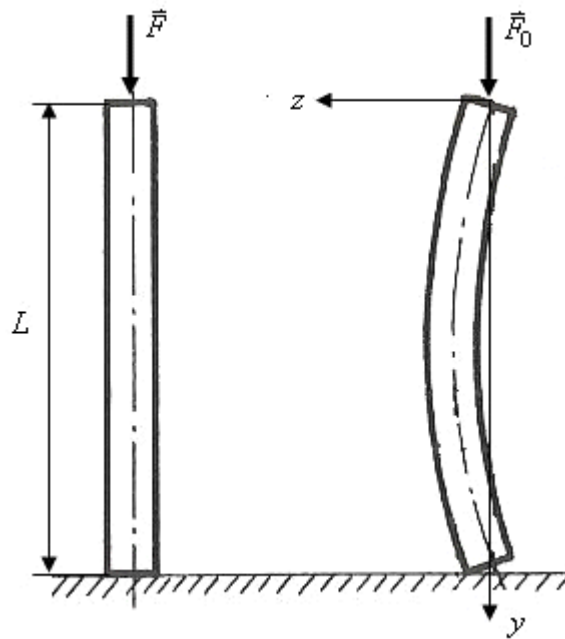


Figure: 70.19 - Exemple de flambage

Lorsque la barre commence à plier nous avons alors une force $\vec{F}_0$ qui s'applique à chaque élément de volume de la barre mais comme ceux-ci ne sont pas distribués de la même manière selon l'axe z , ils ne créent pas le même moment de force!

À l'équilibre de la force de flambement, la barre soumet un moment de rappel. Nous avons alors:

$$M(z) = -z \cdot F \quad (70.136)$$

En exprimant le moment de flexion M au moyen de la relation:

$$M = -\frac{E \cdot I}{R} \quad (70.137)$$

Il vient:

$$z \cdot F_0 + \frac{E \cdot I}{R} = 0 \quad (70.138)$$

En utilisant l'équation de la ligne élastique et en substituant, nous obtenons:

$$z \cdot F_0 + E \cdot I \frac{d^2 z}{dy^2} = 0 \quad (70.139)$$

soit:

$$\frac{d^2 z}{dy^2} + \frac{F_0}{E \cdot I} z = \frac{d^2 z}{dy^2} + \Omega z = 0 \quad (70.140)$$

qui est l'équation différentielle de flambage permettant de calculer la force de flambage avec les conditions initiales:

$$\begin{aligned} y = 0 &\Rightarrow z = 0 \\ y = L &\Rightarrow z = 0 \end{aligned} \quad (70.141)$$

Indiquons que la relation:

$$z \cdot F_0 + E \cdot I \frac{d^2 z}{dy^2} = 0 \quad (70.142)$$

est très souvent écrite sous la forme suivante dans la littérature:

$$\boxed{\frac{d^2 z}{dy^2} = -\frac{z F_0}{EI} = \frac{M}{EI}} \quad (70.143)$$

La résolution de l'équation différentielle du second ordre:

$$\frac{d^2 z}{dy^2} + \frac{F_0}{E \cdot I} z = \frac{d^2 z}{dy^2} + \Omega z = 0 \quad (70.144)$$

est relativement aisée (cf. chapitre de Calcul Différentiel Et Intégral) puisque l'équation caractéristique est:

$$K^2 + \frac{F_0}{E \cdot I} = 0 \Rightarrow K_{1,2} = \pm i \sqrt{\frac{F_0}{E \cdot I}} \quad (70.145)$$

Nous avons alors la solution homogène:

$$\begin{aligned} z_h &= A e^{i \sqrt{\frac{F_0}{E \cdot I}} y} + B e^{-i \sqrt{\frac{F_0}{E \cdot I}} y} \\ &= A \cos \left(\sqrt{\frac{F_0}{E \cdot I}} y \right) + A i \sin \left(\sqrt{\frac{F_0}{E \cdot I}} y \right) + B \cos \left(\sqrt{\frac{F_0}{E \cdot I}} y \right) - B i \sin \left(\sqrt{\frac{F_0}{E \cdot I}} y \right) \end{aligned} \quad (70.146)$$

La condition $y = 0 \Rightarrow z = 0$ impose:

$$z_h = A + B = 0 \Rightarrow B = -A \quad (70.147)$$

Il vient alors:

$$z = 2 A i \sin \left(\sqrt{\frac{F_0}{E \cdot I}} y \right) \quad (70.148)$$

La deuxième condition $y = 0 \Rightarrow z = L$ impose:

$$z = 2 A i \sin \left(\sqrt{\frac{F_0}{E \cdot I}} L \right) = 0 \quad (70.149)$$

Donc il vient immédiatement que:

$$\sqrt{\frac{F_0}{E \cdot I}} L = k\pi \quad (70.150)$$

avec $k = 1$ (car k valant zéro n'est pas une solution physique possible et k entier supérieur à 1 signifierait que la barre plie sur plusieurs périodes ce qui n'est pas le cas puisqu'elle le fait seulement sur une demi-période comme le montrait la figure plus haut). Soit:

$$F_0 = \frac{\pi^2 \cdot E \cdot I}{L^2} \quad (70.151)$$

Cette relation est parfois appelée "**formule d'Euler**" (à ne pas confondre avec la formule du même nom en théorie des graphes) et la charge limite la "**charge ou force critique d'Euler**" pour une poutre parfaitement encastrée aux extrémités. L'ensemble de l'étude étant le "**flambage d'Euler**".

2.5. TRACTION

Considérons maintenant le cas d'une barre suspendue seulement à son propre poids. La surface de sa section circulaire est S et h la hauteur totale de cette barre. Le module de Young du matériau est noté E (cf. chapitre de Mécanique des Milieux Continus) et ρ sa masse volumique.

Il est facile de constater qu'une section située à une altitude z supporte le poids du morceau de barre situé sous elle:

$$F(z) = \rho g S (h - z) \quad (70.152)$$

La contrainte n'est alors pas constante dans la barre:

$$\sigma(z) = \frac{F(z)}{S} = \rho g (h - z) \quad (70.153)$$

et la déformation non plus:

$$\varepsilon(z) = \frac{\sigma(z)}{E} = \frac{\rho g}{E} (h - z) \quad (70.154)$$

z étant l'abscisse sur la barre, la déformation inhomogène est liée au déplacement par la relation:

$$\varepsilon(z) = \frac{du(z)}{dz} \quad (70.155)$$

Après intégration, nous obtenons la forme générale du déplacement:

$$du(z) = \varepsilon(z) dz = \frac{\rho g}{E} (h - z) dz \Rightarrow u(z) = \frac{\rho g}{E} \left(hz - \frac{z^2}{2} \right) + c^{te} \quad (70.156)$$

où la constante est à déterminer en utilisant les éventuelles conditions de liaison aux extrémités de la barre. Si l'extrémité supérieure est encastrée, le déplacement y est donc nul:

$$u(z = 0) = 0 \Rightarrow c^{te} = 0 \quad (70.157)$$

Le déplacement en tout point de la barre s'exprime donc:

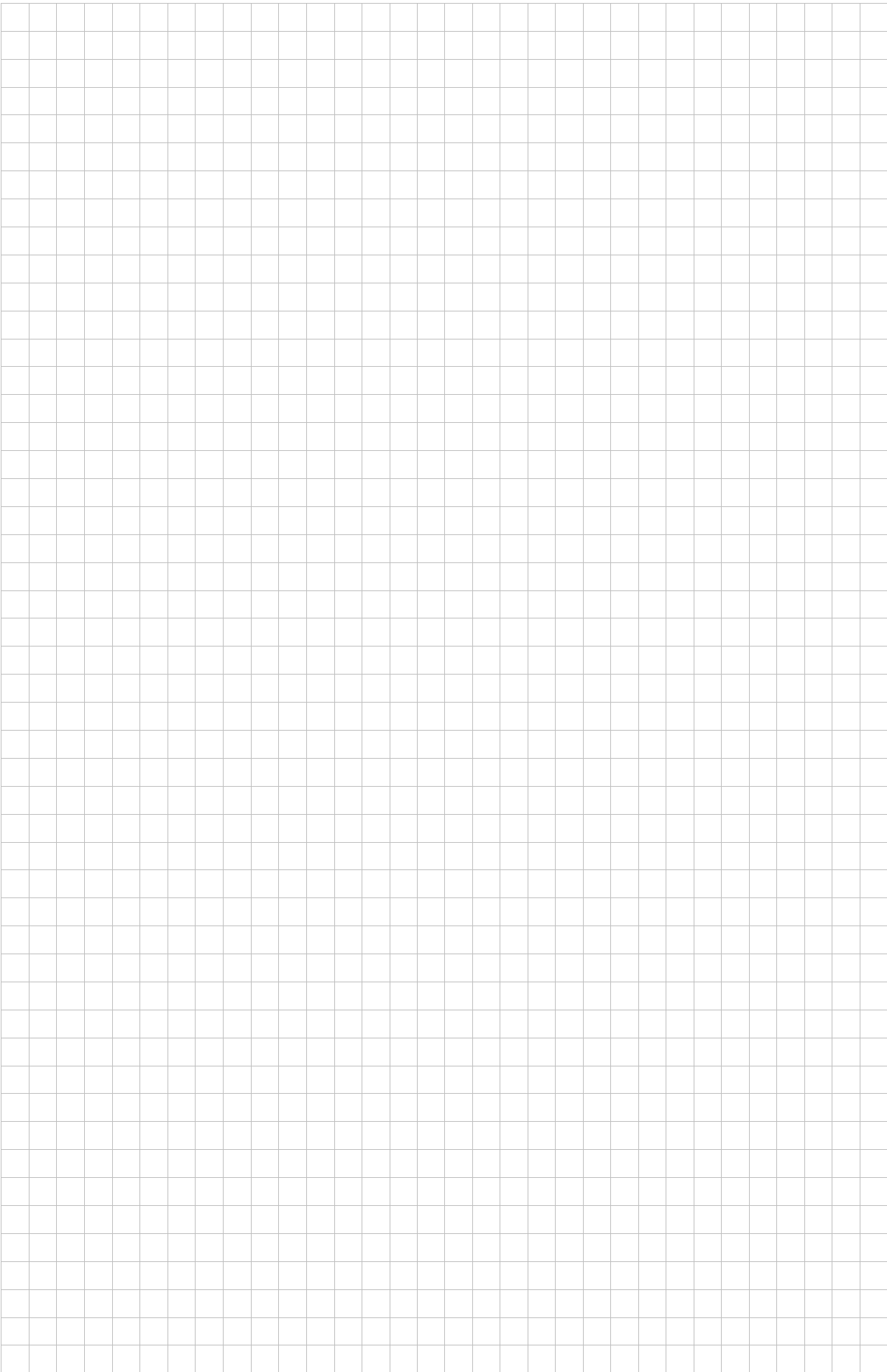
$$u(z) = \frac{\rho g}{E} \left(hz - \frac{z^2}{2} \right) \quad (70.158)$$

L'allongement de la barre est l'écart en déplacement entre les deux extrémités de la barre:

$$\Delta h = u(h) - u(0) \quad (70.159)$$

Nous avons alors trivialement:

$$\Delta h = \frac{\rho g}{E} \left(h^2 - \frac{h^2}{2} \right) = \frac{\rho g}{E} \frac{h^2}{2} \quad (70.160)$$







71. GÉNIE ÉLECTRIQUE

Nous allons voir dans ce chapitre l'étude - sous son aspect mathématique - générale des circuits, montages, puces, et machines, émetteur/récepteurs que l'ingénieur doit savoir formaliser, analyser, comprendre, fabriquer et simuler à la suite de ses études. Pour cette raison, nous avons choisi sur ce site de nous concentrer sur des cas mathématiques d'applications pratiques (utiles!) dans la vie de tous les jours en mentionnant les éventuels pièges et dangers du montage.

Nous traiterons ici dans l'électronique analogique, ensuite de l'électronique de puissance (électrotechnique), de l'électronique numérique ainsi de que la physique des semi-conducteurs pour bien comprendre les fondements de certains éléments.

Le génie électrique est donc une hiérarchie de modèles. C'est la seule façon d'aborder la conception de systèmes complexes. En principe, le fonctionnement de nombreux dispositifs peut, certes, toujours se ramener à l'application des équations de Maxwell (cf. chapitre d'Électrodynamique), mais il est cependant humainement impossible de comprendre la conception de certains systèmes en ce cantonnant à un niveau aussi théorique.

Il est alors d'usage dans l'industrie de hiérarchiser l'analyse en 5 niveaux:

N0. Physique du solide: Ce modèle est essentiel pour l'analyse des propriétés électriques et magnétiques de la matière. Il s'appuie sur les lois de la physique quantique et mène essentiellement à la description de bandes d'énergie et au calcul de leur degré d'occupation. Ce modèle explique par exemple les propriétés fondamentales des semi-conducteurs.

N1. Électromagnétisme: Ce modèle est essentiel pour l'analyse des dispositifs travaillant aux hyperfréquences et celle des dispositifs électromagnétiques. Il s'appuie sur les relations de Maxwell et fait appel à la théorie mathématique des équations aux dérivées partielles. Ce modèle ne permet plus d'analyser l'influence d'un atome car les objets étudiés sont à un niveau plus macroscopique, décrits par leurs dimensions, leur permittivité, leur conductibilité, etc.

N2. Théorie des circuits: Ce modèle est essentiel pour l'analyse des dispositifs électroniques dans le cas très courant où les dimensions du dispositif sont largement inférieures à la longueur d'onde du phénomène étudié. Ce modèle s'appuie sur les lemmes de Kirchhoff et la définition d'une demi-douzaine d'éléments discrets, résistance, capacité, inductance, cours, etc... Il n'y a plus de géométrie dans un tel modèle mais seulement une topologie. On peut calculer le courant et la tension, grandeurs scalaires, alors que les champs n'ont plus de sens. Les techniques mathématiques sont celles des équations différentielles ordinaires, transformations de Laplace, calcul complexe et matriciel, etc...

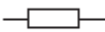
N3. Schémas fonctionnels: À ce niveau, nous ne tenons plus compte de courants ou de tensions, in a fortiori de la géométrie du système. Celui-ci est constitué par la connexion de blocs remplissant des fonctions caractérisées par des relations entre grandeurs de sortie et d'entrée.

N4. Systèmes: À ce niveau, nous schématisons en bloc fonctionnel un ensemble de blocs du niveau 3. Un ordinateur est par exemple une interconnexion de différents systèmes logiques réalisant chacun une fonction particulière.

N5. Logiciel: À partir de ce niveau, l'ingénieur n'ajoute plus de dispositifs supplémentaires, ne les combine plus en des systèmes plus vastes, mais il programme la machine. Les méthodes théoriques se rapprochent alors plus souvent de la linguistique que de la mathématique.

Ce chapitre doit permettre de se familiariser avec le langage et certaines méthodes de calculs utilisées par les ingénieurs de cette branche pour les 4 premiers niveaux. Cependant en toute rigueur toute formation dans ce domaine doit être complétée par des travaux pratiques en laboratoire!

Nous avons tenté de respecter aux mieux dans les schémas la série de normes NF EN 60617 pour les symboles des composants électriques. Comme ces derniers sont très nombreux, nous proposons ci-dessous un tableau uniquement des composants utilisés ou mentionnés jusqu'à maintenant dans ce chapitre. Ce tableau évoluera donc au fil du temps:

Nom	Symbole
Courant alternatif	
Polarité positive	
Polarité négative	
Source idéale de tension	
Liaison	
Points de liaison	
Liaison en T	
Résistance	
Condensateur	
Inductance	
Contact à fermeture, à retour automatique	
Inductance à noyau magnétique	
Transformateur à deux enroulements	

1. COURANT ALTERNATIF VS COURANT CONTINU

Le lecteur va remarquer que tout au long de ce chapitre du site, nous allons principalement travailler avec du courant alternatif. Il nous semble important d'expliquer l'origine de cet attrait du monde industriel contemporain pour le courant alternatif avant d'aller plus loin.

Au fait, l'origine de cet attrait est relativement simple:

Quand des centrales électriques virent le jour, surtout dans les régions éloignées des centres urbains, il fallut transporter l'énergie électrique produite sur des longues distances. Mais les câbles qui transportent l'électricité ont une certaine résistance et cela posa un problème majeur.

Effectivement, une ville moyenne peut largement avoir besoin d'une puissance d'environ 10 [MW]. Si cette quantité devait être transportée sous une tension modeste d'environ 100 [V], comme $P = UI$ (cf. chapitre d'Électrocinétique), le courant devait être énorme: 100'000 [A] !

Mais l'effet Joule dans le cuivre de 1 [cm] de diamètre a une résistance linéaire de $R \cong 0.4 [\Omega / km]$. Avec un courant de 100'000 [A], la perte d'énergie par effet Joule serait d'environ (en négligeant la chute de potentiel):

$$P = R \cdot I^2 = 4'000 [MW / km] \quad (71.1)$$

... on voit très vite le problème!

Au prix de 0.1 [\$/kW · h], cela représentant un coût (perte) d'environ:

$$C = 0.4 \cdot 10^6 [\$ \cdot h^{-1} km^{-1}] \quad (71.2)$$

hummm...!

Il n'y avait d'autres choix économiques que de baisser le courant. En clair, si la tension atteignait 10^5 [V], la même puissance pourrait être transportée efficacement par 100 [A]. Ainsi, en élevant la tension par un facteur 1000, nous pouvons réduire le courant d'un facteur 1000 aussi, et donc la perte Joule par un facteur 10^6 .

Comme il existait déjà un dispositif simple pour élever et abaisser la tension alternative (les transformateurs) sans aucun dispositif comparable pour la tension continue (du moins à l'époque), la course a été gagnée par les adeptes du courant alternatif.

Il faut rajouter également comme deuxième intérêt que certains composants électriques linéaires (voir plus loin) n'ont pas beaucoup d'intérêt en courant continu... nous y reviendrons!

Voyons un montage simple pour générer du courant alternatif monophasé:

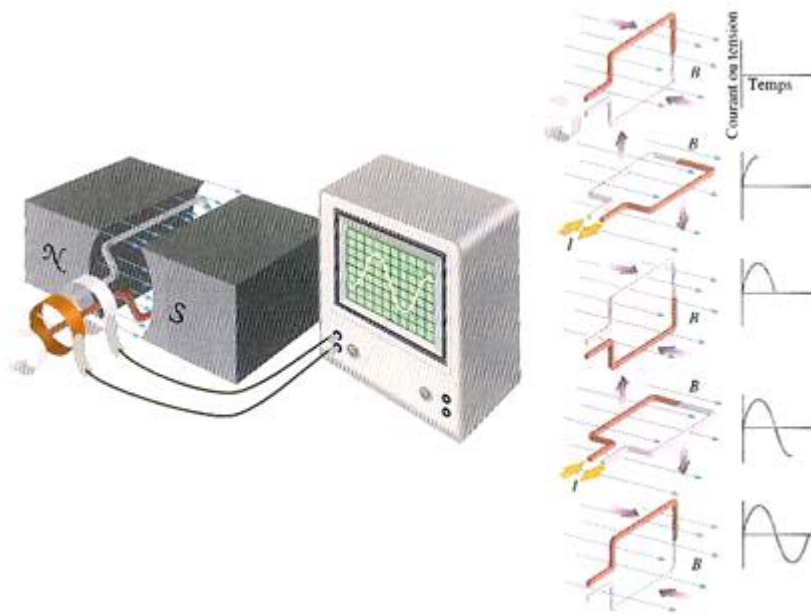


Figure: 71.1 - Exemple de générateur de courant alternatif

La tension (respectivement le courant) est déterminée par la loi de Faraday démontrée dans le chapitre d'Électrocinétique:

$$e = - \frac{d\Phi}{dt} \quad (71.3)$$

qui donne donc la force électromotrice (ou tension dans le cas d'un générateur sans résistance...) induite.

Nous avons bien évidemment dans la situation ci-dessus si l'aimant est permanent et la longueur de la spire carrée est L :

$$e = - \frac{d\Phi}{dt} = - \frac{\vec{B} \cdot d\vec{S}}{dt} = -BL^2 \frac{d \cos \theta}{dt} = -BS \frac{d \cos(\omega t)}{dt} = BS\omega \sin(\omega t) \quad (71.4)$$

Nous voyons déjà que pour obtenir une certaine force électromotrice, il sera préférable de jouer avec la fréquence de rotation plutôt qu'avec la surface ou l'intensité du champ magnétique... ou encore d'augmenter le nombre de spires par un montage permettant d'arriver à la relation suivante:

$$e = NBS\omega \sin(\omega t) \quad (71.5)$$

Il convient d'indiquer pour les sceptiques qu'il y a bien conservation de l'énergie dans ce système! Effectivement, l'énergie nécessaire à faire tourner la spire sera celle en partie utilisée par le système (et c'est pourquoi les barrages font tourner des turbines avec de l'eau et les centrales nucléaires avec de la vapeur et les éoliennes avec du vent...).

Évidemment le cas inverse comme un courant alternatif injecté dans la spire la fera tourner dans le champ magnétique. Donc dans une situation, nous avons un générateur électrique et dans le cas inverse un moteur électrique.

Il est possible avec un appareillage similaire de produire une tension continue à peu près fiable de la manière suivante appelée "**dynamo**":

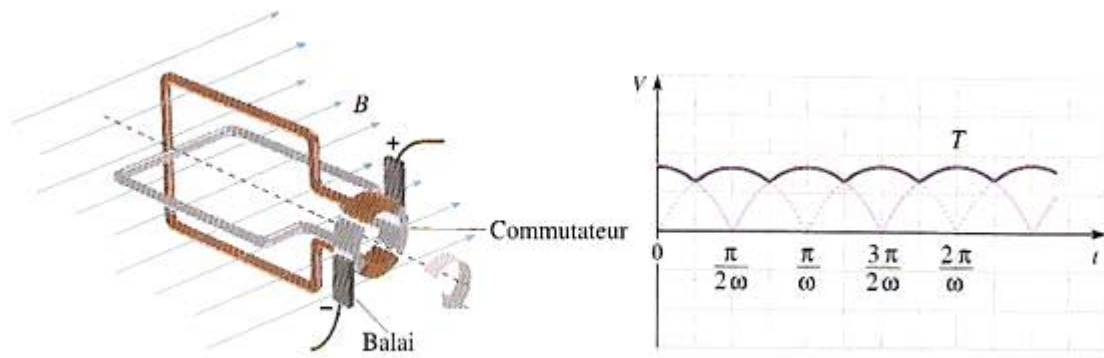


Figure: 71.2 - Exemple de générateur de courant continu

Le générateur simple donné en premier avec quelques aimants en fer à cheval produisant le champ magnétique était très utilisé au début de l'ère de la technologie électrique mais à des tensions élevées (plusieurs [kV]) et des courants internes élevés aussi (plus de 50 [A]). Les balais métalliques et les bagues collectrices produisaient alors des étincelles et s'abîmaient relativement vite. Actuellement, cette machine à armature tournante n'est plus très utilisée.

Pour éviter les difficultés associées aux tensions dépassant environ 600 [V], nous faisons aujourd'hui tourner des électroaimants autour d'une armature immobile. Le courant qui alimente les électroaimants tournants (qui peuvent aussi être des aimants permanents), est relativement faible et il ne pose aucun problème au fonctionnement des bagues et balais. Cette configuration est appelée alors un "**alternateur**" (l'exemple le plus commun étant la vieille lampe à vélo abusivement appelée dynamo alors que ce n'est pas une!).

Avec des composants linéaires électriques il est aussi possible au besoin de redresser la tension (nous verrons cela beaucoup plus loin). Nous avons alors une dynamo!

1.1. PUISSANCE MOYENNE

Nous avons défini dans le chapitre d'Électrocinétique que la puissance joule était donnée par en régime constant pour un conducteur par:

$$P = U \cdot I = R \cdot I^2 \quad (71.6)$$

En régime alternatif périodique et à basse fréquence (afin de considérer la résistance comme constante) nous avons dans le cas purement résistif:

$$p(t) = R \cdot i(t)^2 \quad (71.7)$$

Il vient alors (cf. chapitre de Calcul Différentiel Et Intégral) pour la valeur moyenne d'un signal périodique de période T :

$$\bar{P} = \frac{1}{T} \int_{t_0}^{t_0+T} u(t) \cdot i(t) dt = R \left(\frac{1}{T} \int_{t_0}^{t_0+T} i(t)^2 dt \right) \quad (71.8)$$

Le terme entre parenthèses peut donc être comparé à la valeur qu'aurait un courant continu produit la même puissance joule. Donc:

$$\bar{p} = R \cdot I^2 = R \left(\frac{1}{T} \int_{t_0}^{t_0+T} i(t)^2 dt \right) \quad (71.9)$$

Nous appelons ce courant équivalent la "**valeur efficace du courant**" ou "**courant root mean square**" (abrégé "**courant RMS**") et la notons:

$$I_{\text{eff}} = \sqrt{\frac{1}{T} \int_{t_0}^{t_0+T} i(t)^2 dt} \quad (71.10)$$

les valeurs RMS étant ce que mesurent les multimètres. Pour les régimes sinusoïdaux, nous avons alors:

$$I_{\text{eff}} = \sqrt{\frac{1}{T} \int_{t_0}^{t_0+T} i(t)^2 dt} = \sqrt{\frac{1}{2\pi} \int_0^{2\pi} i(\theta)^2 d\theta} = \sqrt{\frac{1}{2\pi} I_0^2 \int_0^{2\pi} \cos^2(\theta) d\theta} = \sqrt{\frac{1}{2\pi} I_0^2 \pi} = \frac{I_0}{\sqrt{2}} \quad (71.11)$$

Dès lors nous avons:

$$p = u(t) \cdot i(t) = u(t) \frac{I_0}{\sqrt{2}} \quad (71.12)$$

Et, par la même méthode, en souhaitant calculer la puissance moyenne, nous obtenons en régime sinusoïdal:

$$U_{\text{eff}} = \frac{U_0}{\sqrt{2}} \quad (71.13)$$

Donc pour une tension et un courant ayant même phase, nous pouvons écrire:

$$\bar{p} = U_{\text{eff}} I_{\text{eff}} = \frac{U_0}{\sqrt{2}} \frac{I_0}{\sqrt{2}} = \frac{U_0 I_0}{2} \quad (71.14)$$

En toute rigueur il nous faut généraliser cette dernière relation pour des situations où le courant et la tension sont déphasées d'un angle (ou temps) φ .

Remarque: Malheureusement de nombreux ouvrages d'électronique ou mêmes d'articles de recherche ont défini une puissance qui n'a aucun sens physique et qui est appelée la "**puissance RMS**" et mesure des.... "**Watts RMS**" définie par analogie de la manière suivante:

$$P_{\text{eff}} = \sqrt{\frac{1}{T} \int_{t_0}^{t_0+T} p(t)^2 dt} \quad (71.15)$$

Même si ce terme est utilisé par les annonceurs et certains rédacteurs, il n'a pas sa place dans une publication technique digne de ce nom. Il apparaît souvent pour donner un semblant d'expertise technique...

Considérons maintenant une autre approche de la puissance cette fois-ci du point de vue instantané dans un cas plus général que purement résistif. Dans le cas général, nous avons en régime permanent sinusoïdal le droit d'écrire la tension et le courant avec des cosinus sous la forme suivante (forme qui implicitement prend aussi en compte le déphasage du courant et de la tension):

$$\begin{aligned} u(t) &= U_0 \cos(\omega t + \alpha) \\ i(t) &= I_0 \cos(\omega t + \beta) \end{aligned} \quad (71.16)$$

Nous avons alors en utilisant les relations remarquables trigonométriques:

$$\begin{aligned} p(t) &= u(t)i(t) = I_0 U_0 \cos(\omega t + \alpha) \cos(\omega t + \beta) \\ &= U_0 I_0 \left(\frac{\cos(\alpha - \beta) + \cos(2\omega t + \alpha + \beta)}{2} \right) = \frac{U_0 I_0}{2} \cos(\varphi) + \frac{U_0 I_0}{2} \cos(2\omega t + \alpha + \beta) \quad (71.17) \\ &= U_{\text{eff}} I_{\text{eff}} \cos(\varphi) + U_{\text{eff}} I_{\text{eff}} \cos(2\omega t + \alpha + \beta) \end{aligned}$$

où nous avons choisi de poser:

$$\varphi = \alpha - \beta \quad (71.18)$$

tel que ce terme soit positif ou nul (nous pouvons toujours faire le choix de soustraire l'un des deux termes à l'autre afin d'avoir une valeur de φ strictement positive ou nulle sans que cela change les développements et conclusions qui vont suivre!) et où nous avons utilisé les résultats obtenus précédemment:

$$U_{\text{eff}} = \frac{U_0}{\sqrt{2}}, I_{\text{eff}} = \frac{I_0}{\sqrt{2}} \quad (71.19)$$

Ainsi, dans l'expression:

$$p(t) = U_{\text{eff}} I_{\text{eff}} \cos(\varphi) + U_{\text{eff}} I_{\text{eff}} \cos(2\omega t + \alpha + \beta) \quad (71.20)$$

la puissance instantanée comprend donc une composant constante (premier terme) et une composante pulsée et de fréquence double de celle du courant et de la tension. En utilisant:

$$\varphi = \alpha - \beta \Rightarrow \beta = \alpha - \varphi \quad (71.21)$$

ainsi que la relation trigonométrique remarquable suivante (cf. chapitre de Trigonométrie):

$$\cos(2\omega t + \alpha + \beta) = \cos(2\omega t + 2\alpha - \varphi) = \cos(\varphi) \cos(2\omega t + 2\alpha) + \sin(\varphi) \sin(2\omega t + 2\alpha) \quad (71.22)$$

Nous avons alors:

$$\begin{aligned}
 p(t) &= U_{\text{eff}} I_{\text{eff}} \cos(\varphi) + U_{\text{eff}} I_{\text{eff}} [\cos(\varphi) \cos(2\omega t + 2\alpha) + \sin(\varphi) \sin(2\omega t + 2\alpha)] \\
 &= U_{\text{eff}} I_{\text{eff}} \cos(\varphi) [1 + \cos(2\omega t + 2\alpha)] + U_{\text{eff}} I_{\text{eff}} \sin(\varphi) \sin(2\omega t + 2\alpha)
 \end{aligned} \quad (71.23)$$

Cette dernière relation met en évidence que la puissance instantanée peut toujours être ramenée en régime sinusoïdal permanent à la somme de deux termes où :

- le premier est une composante pulsée, toujours positive (par construction), qui oscille autour de la valeur moyenne $U_{\text{eff}} I_{\text{eff}} \cos(\varphi)$ et qui traduit un échange d'énergie unidirectionnel entre une source et une charge.

- le deuxième est une composante alternative qui varie sinusoïdalement avec une amplitude $U_{\text{eff}} I_{\text{eff}} \sin(\varphi)$ et une valeur moyenne nulle. Elle est donc alternativement positive et négative et traduit un échange oscillatoire d'énergie entre une source et une charge.

Lorsque $\varphi = 0$ (charge purement résistive parfaite), nous avons alors :

$$\cos(\varphi) = 1, \sin(\varphi) = 0 \quad (71.24)$$

La valeur moyenne $U_{\text{eff}} I_{\text{eff}} \cos(\varphi)$ est alors maximum et égal à $U_{\text{eff}} I_{\text{eff}}$, alors que le deuxième terme (la composante alternative) est nul.

Par contre, lorsque $\varphi = \pm \pi/2$ (charge purement réactive comme inductance ou capacité idéales), alors :

$$\cos(\varphi) = 0, \sin(\varphi) = \pm 1 \quad (71.25)$$

et dans ce cas, la puissance instantanée se réduit à la seule composante alternative. De ceci il découle les définitions suivantes :

Définitions :

D1. Nous appelons "**puissance active**" le terme :

$$U_{\text{eff}} I_{\text{eff}} \cos(\varphi) \quad (71.26)$$

mesurable par un wattmètre et qui correspond à une fourniture réelle d'énergie convertible en travail ou en chaleur et qui est donc maximale dans le cas d'une charge purement résistive (idéale) et nulle en cas de charge purement réactive (idéale).

D2. Nous appelons "**puissance réactive**" le terme :

$$U_{\text{eff}} I_{\text{eff}} \sin(\varphi) \quad (71.27)$$

qui n'est pas mesurable par un wattmètre puisque nul par alternation.

D3. Nous appelons "**puissance apparente**" le produit :

$$U_{eff} I_{eff} \quad (71.28)$$

qui est en apparence une puissance, mais ne fournit pas nécessairement un travail, d'où son nom... Par tradition c'est la puissance apparente qui est indiquée sur des grosses installations industrielles. On peut la reconnaître car les unités sont souvent indiquées en volts-ampères VA plutôt qu'en watts W (car il ne s'agit pas totalement d'une puissance).

2. TRANSFORMATEUR

Un transformateur électrique statique est une machine électrique permettant de modifier les valeurs de tension et d'intensité du courant délivrées par une source d'énergie électrique alternative, en un système de tension et de courant de valeurs différentes, mais de même fréquence et de même forme.

Rappelons que nous avons démontré dans le chapitre d'Électrocinétique que:

$$e = -N \frac{d\Phi}{dt} = -L \frac{di}{dt} \quad (71.29)$$

Si nous arrivons d'une manière ou d'une autre à faire passer le flux d'un premier solénoïde, appelé alors "enroulement primaire", ayant:

$$e_1 = -N_1 \frac{d\Phi}{dt} = -L_1 \frac{di}{dt} \quad (71.30)$$

à un deuxième solénoïde, appelé alors "enroulement secondaire", et ce sans aucune perte (ou du moins une perte négligeable) tel que:

$$e_2 = -N_2 \frac{d\Phi}{dt} = -L_2 \frac{di}{dt} \quad (71.31)$$

Il vient alors par identification terme à terme:

$$\frac{e_1}{e_2} = \frac{N_1}{N_2} \quad (71.32)$$

et si les résistances internes sont négligeables la force électromotrice induite est égale alors à la tension aux bornes, il vient alors:

$$\frac{U_1}{U_2} = \frac{N_1}{N_2} \quad (71.33)$$

et si toute l'énergie du champ magnétique est transmise dans l'enroulement secondaire, nous avons:

$$P_1 = P_2 \Leftrightarrow U_1 I_1 = U_2 I_2 \quad (71.34)$$

d'où:

$$\frac{U_1}{U_2} = \frac{I_2}{I_1} = \frac{N_1}{N_2} \quad (71.35)$$

Ainsi, le rapport du nombre de spires primaires sur le nombre de spires secondaires détermine totalement le rapport de transformation du transformateur qui peut être utilisé alors comme station de transformation de la tension en montant ou abaissant celle-ci en fonction du nombre de spires de l'enroulement secondaire. Il convient aussi de noter qu'un transformateur qui augmente la tension diminue simultanément le courant et inversement.

Un instrument historique et pédagogique typique permettant de faire cela est le transformateur monophasé à noyau ferromagnétique:

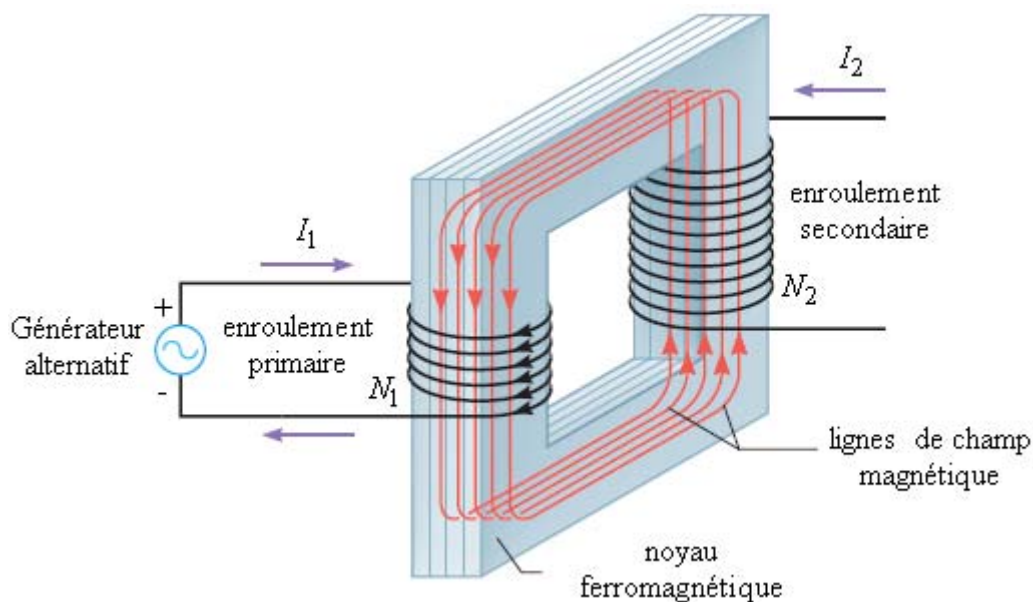


Figure: 71.3 - Exemple type de transformateur monophasé à noyau ferromagnétique

Dans la pratique, les transformateurs monophasés bas de gamme ont les enroulements qui sont concentriques pour minimiser les fuites de flux. Un isolant est inséré entre le circuit primaire et le secondaire:

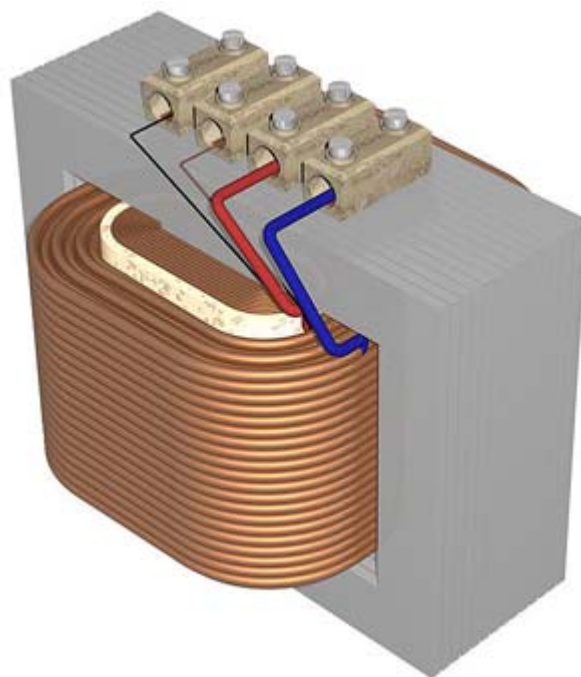


Figure: 71.4 - Transformateur monophasé concentrique type (source Wikipédia)

Les transformateurs de par leur possibilité d'augmenter la tension et de réduire le courant (économie au niveau de la perte Joule comme nous l'avons déjà mentionné) jouent un rôle important dans la transmission de l'électricité des infrastructures domestiques (basse tension: BT, moyenne tension: MT, haute tension HT). Ainsi nous retrouvons les transformateurs élévateurs (TE) et les transformateurs abaisseurs de tension (AT) dans la vie de tous les jours:

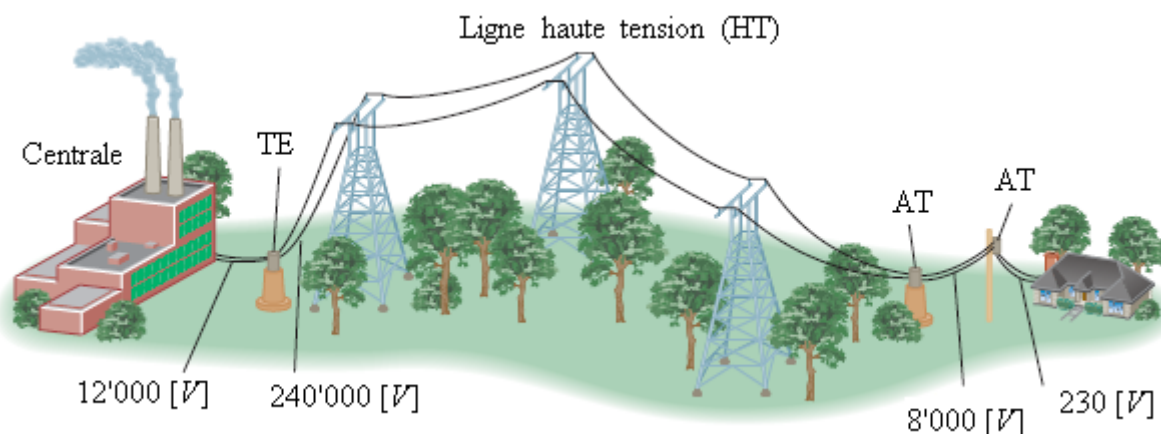


Figure: 71.5 - Installation civile typique de transformation de tension

3. CIRCUITS LINÉAIRES EN RÉGIME CONSTANT

Nous allons voir ici des circuits composés d'éléments simples comme une résistance, une capacité et une impédance. Ces circuits, dont l'équation représentative est une équation différentielle linéaire, sont appelés "**circuits linéaires**". Par ailleurs, ils sont un excellent exemple pour voir la lourdeur des développements en utilisant des représentations mathématiques classiques à l'opposé d'autres techniques beaucoup plus souples et puissantes (représentation par phaseurs et transformées de Laplace).

3.1. CIRCUIT RC SÉRIE

Tout circuit possédant un condensateur possède également une résistance, ne serait-ce que celle des fils de connexion. De tels circuits RC série sont très courants et parfois d'une grande importance (stimulateur cardiaque par exemple). Effectivement, quand nous fermons un circuit qui ne contient que des résistances (circuit purement résistif), le courant monte à sa valeur nominale dans un temps extrêmement bref, de sorte que nous pouvons considérer que le courant et la tension sont constants avec une excellente approximation. Ainsi, le régime permanent s'établit après un régime transitoire très bref. Au contraire, dans un circuit RC série, tension et courant prennent un certain temps pour atteindre leurs valeurs nominales. Cette dépendance en fonction du temps a de multiples applications et permet de produire toute une gamme de signaux modulables dans le temps en fonction des besoins.

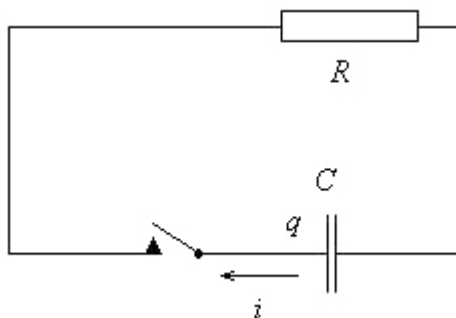


Figure: 71.6 - Circuit RC série

Nous supposons qu'initialement le condensateur est chargé et qu'il ne circule aucun courant (interrupteur ouvert):

$$q(t=0) = q_0 \text{ et } i(t=0) = 0 \quad (71.36)$$

Quand nous fermons l'interrupteur les électrons partent du condensateur C . Nous avons alors aux bornes de la résistance:

$$u_R = R \cdot i = R \frac{dq}{dt} \quad (71.37)$$

aux bornes du condensateur:

$$u_C = \frac{1}{C} q \quad (71.38)$$

L'équation du circuit est alors:

$$R \frac{dq}{dt} + \frac{1}{C} q = 0 \quad (71.39)$$

Soit:

$$\frac{1}{q} dq = -\frac{1}{RC} dt \quad (71.40)$$

Équation différentielle triviale dont la solution est avec les conditions initiales:

$$\ln\left(\frac{Q}{Q_0}\right) = -\frac{t}{RC} \quad (71.41)$$

Soit:

$$q = Q_0 e^{-t/RC} \quad (71.42)$$

Le courant a alors la forme suivante:

$$i = -\frac{Q_0}{RC} e^{-t/RC} \quad (71.43)$$

Il s'agit donc d'un système où le courant décroît exponentiellement et ce d'autant plus vite que le facteur RC appelé "**constante de temps**" est petit. Nous voyons ce genre de système lorsque la lumière à l'intérieur d'une voiture s'éteint document après fermeture des portes.

Lorsque l'on met ce régime sous une tension permanente U_0 , nous avons alors une équation de la forme:

$$R \frac{dq}{dt} + \frac{1}{C} q = U_0 \quad (71.44)$$

Une solution particulière évidente est alors:

$$q = Q_0 = C U_0 \quad (71.45)$$

Nous avons alors la solution générale:

$$q = -Q_0 e^{-t/RC} + C U_0 \quad (71.46)$$

soit pour le courant:

$$i = \frac{dq}{dt} = \frac{Q_0}{RC} e^{-t/RC} \quad (71.47)$$

Et pour la tension aux bornes du condensateur:

$$u_C = \frac{1}{C} q = -\frac{Q_0}{C} e^{-t/RC} + U_0 = -U_0 e^{-t/RC} + U_0 = U_0 \left(1 - e^{-t/RC}\right) \quad (71.48)$$

ce qui représente donc la tension aux bornes du condensateur lors de la charge. Donc au final nous avons les deux relations:

$$\begin{aligned} i &= \frac{Q_0}{RC} e^{-t/RC} \\ u_C &= U_0 (1 - e^{-t/RC}) \end{aligned} \quad (71.49)$$

Ainsi, en posant $U_0 = u_e$ nous avons respectivement à la fermeture et ouverture du circuit:

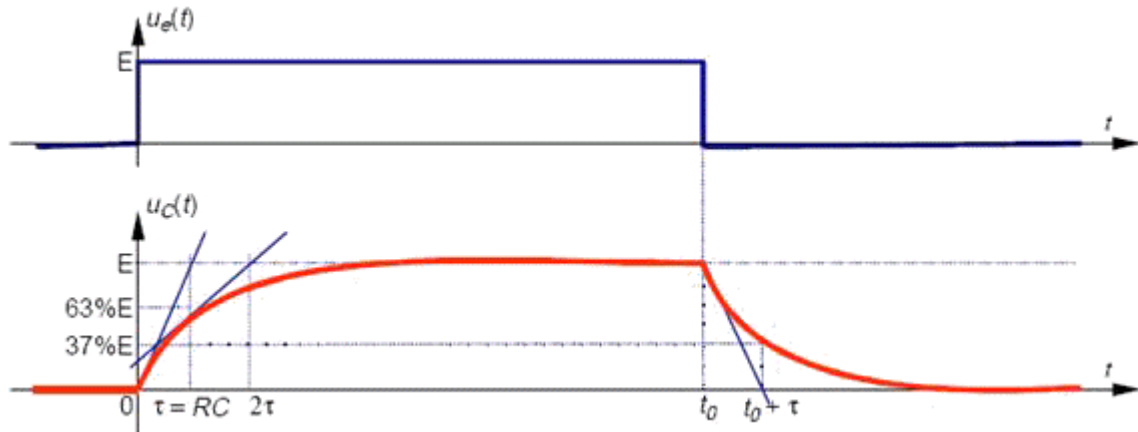


Figure: 71.7 - Charge et décharge du condensateur lors de l'ouverture/fermeture de l'interrupteur

Étudions maintenant l'aspect énergétique qui est important en ingénierie puisque la consommation ou perte de puissance (rendement) est un facteur de vente important dans certaines applications!

Pour cela reprenons la relation:

$$R \frac{dq}{dt} + \frac{1}{C} q = R \frac{dq}{dt} + u_C = U_0 \quad (71.50)$$

et multiplions par i :

$$Ri^2 + u_C i = U_0 i \quad (71.51)$$

Ce que nous écrivons:

$$p_R + p_C = p_G \quad (71.52)$$

Puisque:

$$u_C = \frac{1}{C} q \Rightarrow \frac{du_C}{dt} = \frac{1}{C} \frac{dq}{dt} = \frac{1}{C} i \Rightarrow i = C \frac{du_C}{dt} \quad (71.53)$$

où nous voyons donc que dès que le régime transitoire de charge ou décharge est terminé, la tension aux bornes du condensateur étant nulle alors le courant est nul.

Nous avons alors:

$$U_0 i = Ri^2 + u_c C \frac{du_c}{dt} = Ri^2 + \frac{d}{dt} \left(\frac{1}{2} C u_c^2 \right) \quad (71.54)$$

où le premier terme est la puissance fournie par le générateur au circuit, le second est le terme d'effet Joule et le troisième est la puissance stockée dans le condensateur:

L'énergie fournie par le générateur se retrouve stockée dans le condensateur et dissipée par la résistance par effet Joule.

Ce qui est le plus important c'est de faire un bilan sur toute la durée de charge du condensateur pour signaler la puissance dissipée dans les caractéristiques de vente (cela passe mieux que de mettre des équations...). Pour cela, il suffit d'intégrer la relation précédente de 0 à l'infini pour obtenir l'énergie dissipée.

Le premier terme à gauche de l'égalité donne:

$$\begin{aligned} \int_0^{+\infty} U_0 i dt &= U_0 \int_0^{+\infty} i dt = U_0 \int_0^{+\infty} C \frac{du_c}{dt} dt = U_0 C \int_0^{+\infty} du_c \\ &= U_0 C (u_c(+\infty) - u_c(0)) = U_0 C (U_0 - 0) = C U_0^2 \end{aligned} \quad (71.55)$$

Le second terme s'intègre en utilisant $i(t)$:

$$\begin{aligned} \int_0^{+\infty} Ri^2 dt &= \int_0^{+\infty} R \left(\frac{Q_0}{RC} e^{-t/RC} \right)^2 dt = \int_0^{+\infty} R \left(\frac{1}{R^2} \left(\frac{Q_0}{C} \right)^2 e^{-2t/RC} \right) dt = \frac{U_0^2}{R} \int_0^{+\infty} e^{-2t/RC} dt \\ &= \frac{U_0^2}{R} \left(-\frac{RC}{2} e^{-2t/RC} \right) \Big|_0^{+\infty} = -\frac{U_0^2 C}{2} \left(e^{-2t/RC} \right) \Big|_0^{+\infty} = -\frac{U_0^2 C}{2} (0 - 1) = \frac{1}{2} U_0^2 C \end{aligned} \quad (71.56)$$

Le troisième terme s'intègre immédiatement puisque nous avons déjà la primitive:

$$\int_0^{+\infty} \frac{d}{dt} \left(\frac{1}{2} C u_c^2 \right) dt = \int_0^{+\infty} d \left(\frac{1}{2} C u_c^2 \right) = \frac{1}{2} C U_0^2 \quad (71.57)$$

Nous obtenons finalement:

$$C U_0^2 = \frac{1}{2} U_0^2 C + \frac{1}{2} U_0^2 C \quad (71.58)$$

Ainsi, pour des durées grandes, la moitié de l'énergie fournie par le générateur est dissipée par effet Joule et l'autre stockée dans le condensateur.

3.2. CIRCUIT RL SÉRIE

Considérons le circuit suivant:

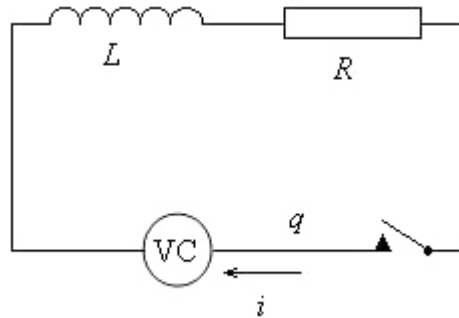


Figure: 71.8 - Circuit RL série

Quand nous fermons l'interrupteur, nous avons alors aux bornes de la résistance:

$$u_R = R \cdot i \quad (71.59)$$

et aux bornes de l'inductance:

$$u_L = L \frac{di}{dt} \quad (71.60)$$

et U_0 aux bornes du générateur de tension continue.

L'équation du circuit est alors:

$$L \frac{di}{dt} + Ri = U_0 \quad (71.61)$$

Soit:

$$\frac{L}{dt} = \frac{U_0 - Ri}{di} \quad (71.62)$$

En inversant:

$$\frac{1}{L} dt = \frac{1}{U_0 - Ri} di \quad (71.63)$$

Faisons un changement de variable:

$$U_0 - Ri = x \Rightarrow \frac{dx}{di} = -R \quad (71.64)$$

Alors:

$$\int_0^t \frac{1}{L} dt = -\frac{1}{R} \int_{U_0}^{U_0 - Ri} \frac{1}{x} dx \quad (71.65)$$

Il vient alors après intégration:

$$\frac{1}{L}t = -\frac{1}{R} \ln(x) \Big|_{U_0}^{U_0 - Ri} = -\frac{1}{R} [\ln(U_0 - Ri) - \ln(U_0)] = -\frac{1}{R} \ln \left(\frac{U_0 - Ri}{U_0} \right) \quad (71.66)$$

Multiplions les deux membres par $-R$, puis prenons l'exponentielle des deux membres:

$$-\frac{R}{L}t = \ln \left(\frac{U_0 - Ri}{U_0} \right) \quad (71.67)$$

et:

$$\frac{U_0 - Ri}{U_0} = e^{-\frac{R}{L}t} \quad (71.68)$$

Alors:

$$i = \frac{U_0}{R} \left(1 - e^{-\frac{R}{L}t} \right) = \frac{U_0}{R} \left(1 - e^{-t/\tau} \right) \quad (71.69)$$

où nous avons la constante de temps définie par:

$$\tau = L / R \quad (71.70)$$

Ainsi, à la fermeture de l'interrupteur, le courant croît de manière exponentielle avec une asymptote à la valeur U_0 / R . Donc contrairement à au circuit RC, le courant tend vers une valeur fixe non nulle lorsque t tend vers l'infini!!!

Nous avons donc:

$$u_L(t) = L \frac{d}{dt} \frac{U_0}{R} \left(1 - e^{-t/\tau} \right) = \frac{LU_0}{R} \frac{d}{dt} \left(1 - e^{-t/\tau} \right) = \frac{LU_0}{R} \frac{d}{dt} \left(-e^{-\frac{R}{L}t} \right) = U_0 e^{-\frac{R}{L}t} \quad (71.71)$$

Donc au final nous avons les deux relations:

$$\boxed{\begin{aligned} i &= \frac{U_0}{R} \left(1 - e^{-t/\tau} \right) \\ u_L &= U_0 e^{-\frac{R}{L}t} \end{aligned}} \quad (71.72)$$

Étudions maintenant l'aspect énergétique qui est important en ingénierie puisque la consommation ou perte de puissance (rendement) est un facteur de vente important dans certaines applications! Comme pour le circuit RC, nous allons directement faire un bilan sur toute la durée du régime transitoire pour signaler la puissance dissipée dans les caractéristiques de vente (cela passe mieux que de mettre des équations...)

$$L \frac{di}{dt} + Ri = U_0 \quad (71.73)$$

Multiplions les termes de l'équation différentielle par $i(t)$:

$$L \frac{di}{dt} i + Ri^2 = \frac{1}{2} \frac{d}{dt} (Li^2) + Ri^2 = U_0 i \quad (71.74)$$

Ce que nous écrivons:

$$p_L(t) + p_R(t) = p_G(t) \quad (71.75)$$

Pour calculer l'énergie dissipée, nous procédons de la même manière que pour le circuit RC série. Nous avons après intégration:

$$\frac{1}{2} L (i^2(+\infty) - i^2(0)) + \int_0^{+\infty} Ri^2 dt = \int_0^{+\infty} U_0 i dt \quad (71.76)$$

soit:

$$\begin{aligned} \frac{1}{2} Li^2 + R \int_0^{+\infty} \left(\frac{U_0}{R} (1 - e^{-t/\tau}) \right)^2 dt &= U_0 \int_0^{+\infty} \left(\frac{U_0}{R} (1 - e^{-t/\tau}) \right) dt \\ \frac{1}{2} Li^2 + \frac{U_0^2}{R} \int_0^{+\infty} \left(1 - e^{-\frac{R}{L}t} \right)^2 dt &= \frac{U_0^2}{R} \int_0^{+\infty} \left(1 - e^{-\frac{R}{L}t} \right) dt \end{aligned} \quad (71.77)$$

Contrairement au cas du circuit RC, nous ne pouvons intégrer ci-dessus avec les bornes données à cause du "1-" qui est devant l'exponentielle car celui-ci fait tendre la puissance consommée vers l'infini ce qui est logique, contrairement au circuit RC qui finit lui par se bloquer une fois la capacité chargée (le courant i tendant vers zéro très vite).

Donc soit, nous intégrons seulement jusqu'à un t temps limite suffisamment grand par rapport aux valeurs des éléments passifs (deux ou trois τ), soit nous nous intéressons à titre purement indicatif à la valeur instantanée de la puissance. Nous avons alors:

$$\begin{aligned} p_G(t) &= U_0 \cdot i = U_0 \cdot \frac{U_0}{R} (1 - e^{-t/\tau}) = \frac{U_0^2}{R} (1 - e^{-t/\tau}) \\ p_R(t) &= R \cdot i^2 = R \cdot \frac{U_0^2}{R^2} (1 - e^{-t/\tau})^2 = \frac{U_0^2}{R} (1 - e^{-t/\tau})^2 \\ p_L(t) &= u_L(t) \cdot i = U_0 e^{-\frac{R}{L}t} \frac{U_0}{R} (1 - e^{-t/\tau}) = \frac{U_0^2}{R} e^{-t/\tau} (1 - e^{-t/\tau}) \end{aligned} \quad (71.78)$$

Et donc à la fin du régime transitoire quand $t \rightarrow +\infty$:

$$\begin{aligned}
 p_G(t) &= \frac{U_0^2}{R} \\
 p_R(t) &= \frac{U_0^2}{R} \quad (71.79) \\
 p_L(t) &= 0
 \end{aligned}$$

donc en régime stable la résistance est le seul élément dissipatif d'énergie dans le circuit et il suffit de multiplier alors la puissance dissipée par l'intervalle de temps désiré pour avoir une estimation de l'énergie dissipée.

3.3. CIRCUIT RLC SÉRIE

Un fil électrique (une antenne par exemple) n'est pas un conducteur parfait. En réalité il peut être assimilé à une résistance, une capacité et une inductance interne en série. Si nous prenons le cas par exemple des générateurs, souvent on ne considère que la résistance interne comme non négligeable et celle-ci fait bien évidemment diminuer la tension nominale du générateur d'un facteur en première approximation proportionnel au courant qui le traverse.

Pour étudier le comportement d'un tel élément souvent appelé "**circuit RLC**" nous le représentons d'abord sous la forme suivante:

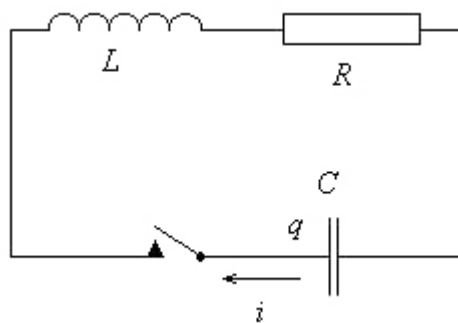


Figure: 71.9 - Circuit RL série

Nous supposons qu'initialement le condensateur est chargé et qu'il ne circule aucun courant (interrupteur ouvert):

$$q(t=0) = q_0 \text{ et } i(t=0) = 0 \quad (71.80)$$

Quand nous fermons l'interrupteur les électrons partent du condensateur C . Nous avons alors aux bornes de la résistance:

$$u_R = Ri = R \frac{dq}{dt} \quad (71.81)$$

aux bornes du condensateur:

$$u_C = \frac{1}{C} q \quad (71.82)$$

aux bornes de la bobine:

$$u_L = L \frac{di}{dt} = L \frac{d^2 q}{dt^2} \quad (71.83)$$

La somme des différences de potentiel du circuit est égale à la différence initiale de potentiel d'où:

$$L \frac{d^2 q}{dt^2} + R \frac{dq}{dt} + \frac{1}{C} q = 0 \quad (71.84)$$

ou autrement écrit:

$$LC \frac{d^2 q}{dt^2} + RC \frac{dq}{dt} + q = 0 \quad (71.85)$$

Il s'agit d'une équation différentielle linéaire du second ordre très connue en physique (nous la retrouvons à l'identique dans d'autres domaines avec juste des constantes différentes). Pour la résoudre il faut chercher les racines de l'équation caractéristique associée (cf. chapitre de Calcul Différentiel Et Intégral):

$$LCx^2 + RCx + 1 = 0 \quad (71.86)$$

Celle-ci a pour discriminant:

$$\Delta = R^2 C^2 - 4LC \quad (71.87)$$

La valeur de la résistance pour laquelle ce discriminant est nul est appelée "**résistance critique**":

$$R_c = 2\sqrt{\frac{L}{C}} \quad (71.88)$$

Nous pouvons ceci dit écrire le discriminant sous la forme suivante:

$$\Delta = C^2 \left(R^2 - 4\frac{L}{C} \right) = C^2 \left(R^2 - R_c^2 \right) \quad (71.89)$$

Les solutions de l'équation différentielle sont différentes selon le nombre et le type des racines de l'équation caractéristique.

3.3.1. RÉGIME CRITIQUE

Il s'agit du cas où $R = R_c$. L'équation caractéristique admet alors une racine double réelle puisque:

$$\Delta = C^2 \left(R^2 - 4\frac{L}{C} \right) = C^2 \left(R^2 - R_c^2 \right) = 0 \quad (71.90)$$

Nous avons alors:

$$x_{1,2} = \frac{-RC \pm \sqrt{\Delta}}{2LC} = \frac{-R_c}{2L} = -\frac{1}{\sqrt{LC}} = -\frac{1}{\tau} = K \quad (71.91)$$

avec:

$$\tau = \sqrt{LC} \quad (71.92)$$

L'équation différentielle admet alors une solution du type suivant lorsque le discriminant est nul (cf. chapitre de Calcul Différentiel et Intégral):

$$q = (C_1 t + C_2) e^{-\frac{t}{\sqrt{LC}}} \quad (71.93)$$

en omettant le retard.

Ce qui donne pour l'intensité:

$$i = \frac{dq}{dt} = \left(C_1 - C_1 \frac{t}{\sqrt{LC}} - \frac{C_2}{\sqrt{LC}} \right) e^{-\frac{t}{\sqrt{LC}}} \quad (71.94)$$

Les constantes sont définies par les conditions initiales:

$$\begin{cases} q(t=0) = C_2 = q_0 \\ i(t=0) = \left(C_1 - \frac{C_2}{\sqrt{LC}} \right) = 0 \Rightarrow C_1 = \frac{C_2}{\sqrt{LC}} = \frac{q_0}{\sqrt{LC}} \end{cases} \quad (71.95)$$

Nous obtenons donc pour la solution globale:

$$\begin{aligned} i &= q_0 \left(1 + \frac{t}{\tau} \right) e^{-t/\tau} \\ u_L &= \frac{q_0 t}{\tau^2} e^{-t/\tau} \end{aligned} \quad (71.96)$$

Donc au final, nous avons les deux relations:

$$\boxed{\begin{aligned} i &= q_0 \left(1 + \frac{t}{\tau} \right) e^{-t/\tau} \\ u_L &= \frac{q_0 t}{\tau^2} e^{-t/\tau} \end{aligned}} \quad (71.97)$$

Les figures suivantes illustrent l'allure de l'évolution temporelle de la charge du condensateur et de l'intensité au travers de l'inductance. L'intensité est maximale pour $t = \tau$ dans l'inductance:

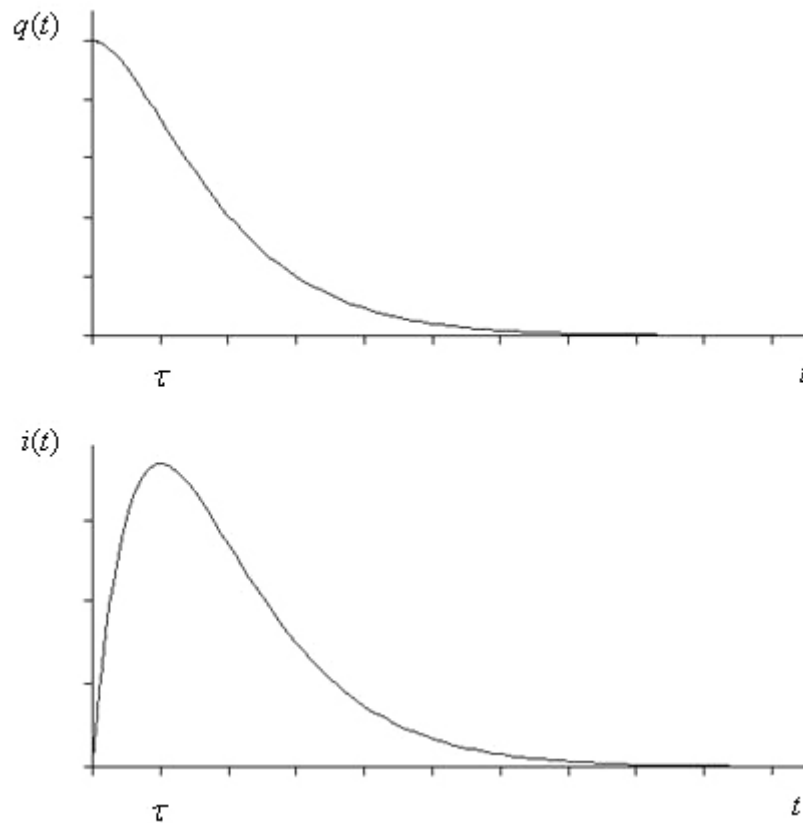


Figure: 71.10 - Comportement du courant dans le circuit lors de la décharge du condensateur

3.3.2. RÉGIME APÉRIODIQUE (OU HYPERCRITIQUE)

Il s'agit du cas où $R > R_C$. L'équation caractéristique admet alors deux racines réelles distinctes:

$$\Delta = C^2 \left(R^2 - 4 \frac{L}{C} \right) > 0 \quad (71.98)$$

Soit:

$$\begin{aligned} K_{1,2} &= \frac{-RC \pm \sqrt{C^2 \left(R^2 - 4 \frac{L}{C} \right)}}{2LC} = \frac{-RC \pm C \sqrt{R^2 - 4 \frac{L}{C}}}{2LC} = \frac{-R \pm \sqrt{R^2 - 4 \frac{L}{C}}}{2L} \\ &= \frac{-R \pm \sqrt{R^2 - R_C^2}}{2L} \end{aligned} \quad (71.99)$$

Les deux racines sont de même signe, car en utilisant les relations de Viète (cf. chapitre de Calcul Algébrique) nous avons:

$$K_1 K_2 = \frac{1}{LC} > 0 \quad (71.100)$$

Les deux racines sont donc obligatoirement négatives. Nous notons leurs valeurs absolues:

Ces deux racines sont donc négatives. Nous notons leurs valeurs absolues:

$$|K_{1,2}| = \frac{R \pm \sqrt{R^2 - R_C^2}}{2L} \quad (71.101)$$

qui vérifie donc:

$$K_1 K_2 = \frac{1}{LC} = \frac{1}{\tau^2} \quad (71.102)$$

Nous avons vu dans le chapitre de Calcul Différentiel Et Intégral qu'à ce moment la solution (sans déphasage) est de la forme:

$$q = Ae^{K_1 t} + Be^{K_2 t} \quad (71.103)$$

Ce qui donne pour l'intensité:

$$i = AK_1 e^{K_1 t} + BK_2 e^{K_2 t} \quad (71.104)$$

Les constantes A et B sont définies par les conditions initiales:

$$\begin{cases} q(t=0) = A + B = q_0 \\ i(t=0) = K_1 A + K_2 B = 0 \end{cases} \quad (71.105)$$

Ce qui nous donne:

$$\begin{aligned} A - \frac{K_1}{K_2} A = q_0 &\Rightarrow A \left(1 - \frac{K_1}{K_2}\right) = q_0 \Rightarrow A = \frac{K_2 q_0}{K_2 - K_1} \\ -\frac{K_2}{K_1} B + B = q_0 &\Rightarrow B \left(1 - \frac{K_2}{K_1}\right) = q_0 \Rightarrow B = \frac{K_1 q_0}{K_1 - K_2} \end{aligned} \quad (71.106)$$

Soit sous forme conventionnelle:

$$A = -\frac{K_2 q_0}{K_1 - K_2} \quad B = \frac{K_1 q_0}{K_1 - K_2} \quad (71.107)$$

Soit en reportant dans les expressions de la charge et de l'intensité:

$$\boxed{\begin{aligned} q &= \frac{q_0}{K_1 - K_2} (K_1 e^{K_1 t} - K_2 e^{K_2 t}) \\ i &= \frac{K_1 K_2 q_0}{K_1 - K_2} (e^{K_1 t} - e^{K_2 t}) \end{aligned}} \quad (71.108)$$

Les figures suivantes illustrent l'évolution temporelle de ces fonctions (se rappeler que les racines sont négatives!).

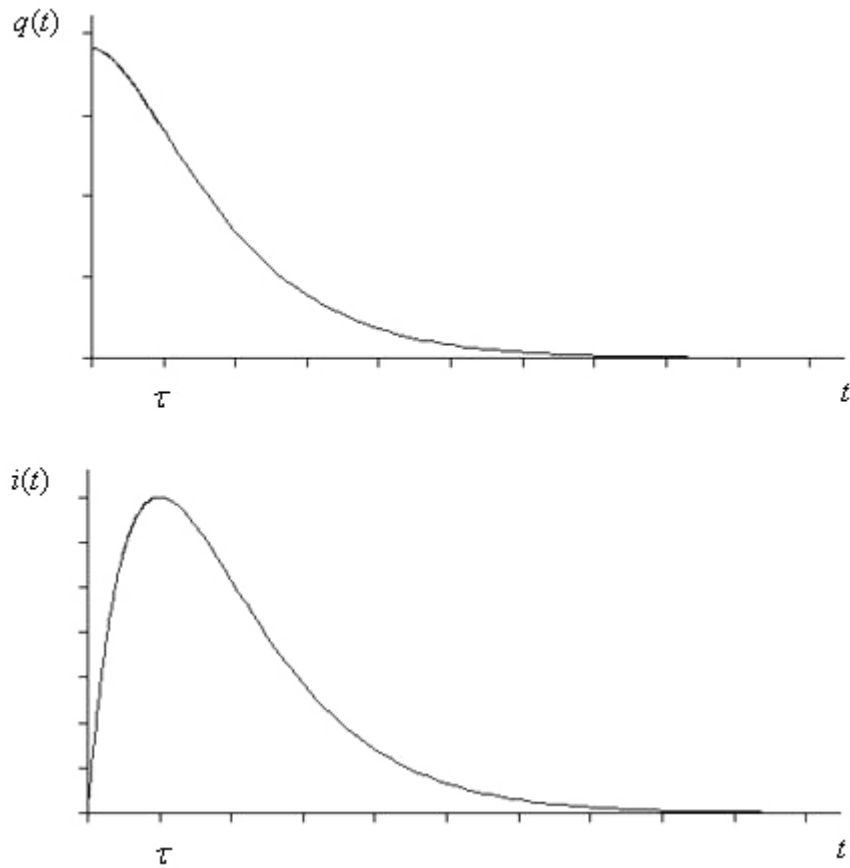


Figure: 71.11 - Comportement du courant dans le circuit lors de la décharge du condensateur

3.3.3. RÉGIME PSEUDO-PÉRIODIQUE (OU DES OSCILLATIONS AMORTIES)

Il s'agit du cas où $R < R_C$. L'équation caractéristique admet alors deux racines complexes conjuguées:

$$K_{1,2} = \frac{-R \pm \sqrt{R^2 - R_C^2}}{2L} = \frac{-R \pm \sqrt{-(R_C^2 - R^2)}}{2L} = \frac{-R \pm \sqrt{i^2 (R_C^2 - R^2)}}{2L} = \frac{-R \pm i \sqrt{R_C^2 - R^2}}{2L} \quad (71.109)$$

qui sont assimilées à la résistance du circuit. Nous l'appelons "**impédance complexe**".

Nous allons voir que contrairement à l'intuition de l'époque, les racines complexes ont une signification physique réelle.

Notons pour cela α et ω les valeurs absolues des parties réelle et imaginaire de ces racines:

$$K_{1,2} = -\alpha \pm i\omega \quad (71.110)$$

avec:

$$\alpha = K_{1,2} = \frac{R}{2L} \quad (71.111)$$

et:

$$\omega = \frac{\sqrt{R_C^2 - R^2}}{2L} = \sqrt{\frac{R_C^2}{4L^2} - \frac{R^2}{4L^2}} = \sqrt{\frac{1}{LC} - \alpha^2} = \sqrt{\frac{1}{\tau^2} - \alpha^2} \quad (71.112)$$

Nous avons vu dans le chapitre de Calcul Différentiel Et Intégral que la solution de l'équation différentielle s'écrit alors:

$$q = Ae^{K_1 t} + Be^{K_2 t} = e^{\alpha x} (A' \cos(\omega t) + B' \sin(\omega t)) = C' e^{-\alpha t} \cos(\omega t - \phi) \quad (71.113)$$

Ce qui nous donne pour l'intensité:

$$\begin{aligned} i &= \frac{d}{dt} C' e^{-\alpha t} \cos(\omega t - \phi) = -\alpha C' e^{-\alpha t} \cos(\omega t - \phi) - C' e^{-\alpha t} \sin(\omega t - \phi) \\ &= -C' e^{-\alpha t} [\alpha \cos(\omega t - \phi) + \omega \sin(\omega t - \phi)] \end{aligned} \quad (71.114)$$

Les constantes C' et ϕ sont déterminées par les conditions initiales:

$$\begin{aligned} q(t=0) &= C' \cos(\phi) = q_0 \\ i(t=0) &= -C' [\alpha \cos(-\phi) + \omega \sin(-\phi)] = 0 \end{aligned} \quad (71.115)$$

Ce qui nous donne:

$$C' = \frac{q_0}{\cos \phi} \quad \text{et} \quad \tan(-\phi) = \frac{\alpha}{\omega} \quad (71.116)$$

Soit:

$$C' = \frac{q_0}{\cos \phi} \quad \text{et} \quad \tan \phi = \frac{\sin \phi}{\cos \phi} = -\frac{\alpha}{\omega} \quad (71.117)$$

Soit en reportant dans les expressions de la charge q et du courant i :

$$q = \frac{q_0}{\cos \phi} e^{-\alpha t} \cos(\omega t - \phi) \quad (71.118)$$

et:

$$\begin{aligned} i &= -\frac{q_0}{\cos \phi} e^{-\alpha t} [\alpha \cos(\omega t - \phi) + \omega \sin(\omega t - \phi)] \\ &= -\frac{q_0}{\cos \phi} e^{-\alpha t} \sqrt{\alpha^2 + \omega^2} \left[\frac{\alpha}{\sqrt{\alpha^2 + \omega^2}} \cos(\omega t - \phi) + \frac{\omega}{\sqrt{\alpha^2 + \omega^2}} \sin(\omega t - \phi) \right] \end{aligned} \quad (71.119)$$

Cherchons à simplifier un peu cette dernière égalité. D'abord nous avons démontré juste avant que:

$$\alpha^2 = \omega^2 \tan^2 \phi \quad (71.120)$$

d'où:

$$\begin{aligned}
 \sqrt{\alpha^2 + \omega^2} &= \sqrt{\omega^2 \tan^2 \phi + \omega^2} = \omega \sqrt{\tan^2 \phi + 1} = \omega \sqrt{\frac{\sin^2 \phi}{\cos^2 \phi} + 1} \\
 &= \frac{\omega}{\cos \phi} \sqrt{\sin^2 \phi + \cos^2 \phi} = \frac{\omega}{\cos \phi}
 \end{aligned} \quad (71.121)$$

Ce qui nous donne:

$$\begin{aligned}
 i &= -\frac{q_0}{\cos \phi} e^{-\alpha t} [\alpha \cos(\omega t - \phi) + \omega \sin(\omega t - \phi)] \\
 &= -\frac{\omega q_0}{\cos^2 \phi} e^{-\alpha t} \left[\frac{\alpha}{\sqrt{\alpha^2 + \omega^2}} \cos(\omega t - \phi) + \frac{\omega}{\sqrt{\alpha^2 + \omega^2}} \sin(\omega t - \phi) \right]
 \end{aligned} \quad (71.122)$$

Or, nous avons aussi:

$$\omega \frac{1}{\sqrt{\alpha^2 + \omega^2}} = \omega \frac{\cos \phi}{\omega} = \cos \phi \quad (71.123)$$

et:

$$\alpha \frac{1}{\sqrt{\alpha^2 + \omega^2}} = \alpha \frac{\cos \phi}{\omega} = \cos \phi \frac{\alpha}{\omega} = -\cos \phi \frac{\sin \phi}{\cos \phi} = -\sin \phi \quad (71.124)$$

Nous avons alors:

$$\begin{aligned}
 i &= -\frac{\omega q_0}{\cos^2 \phi} e^{-\alpha t} [-\sin \phi \cos(\omega t - \phi) + \cos \phi \sin(\omega t - \phi)] \\
 &= \frac{\omega q_0}{\cos^2 \phi} e^{-\alpha t} [\sin \phi \cos(\omega t - \phi) - \cos \phi \sin(\omega t - \phi)] \\
 &= \frac{\omega q_0}{\cos^2 \phi} e^{-\alpha t} \sin(-\omega t) = -\frac{\omega q_0}{\cos^2 \phi} e^{-\alpha t} \sin(\omega t)
 \end{aligned} \quad (71.125)$$

Nous avons au final les deux relations suivantes:

$$\begin{aligned}
 q &= \frac{q_0}{\cos \phi} e^{-\frac{R}{2L} t} \cos(\omega t - \phi) \\
 i &= -\frac{\omega q_0}{\cos^2 \phi} e^{-\frac{R}{2L} t} \sin(\omega t)
 \end{aligned} \quad (71.126)$$

Donc un tracé de i dans l'inductance et de q de la capacité donnera:

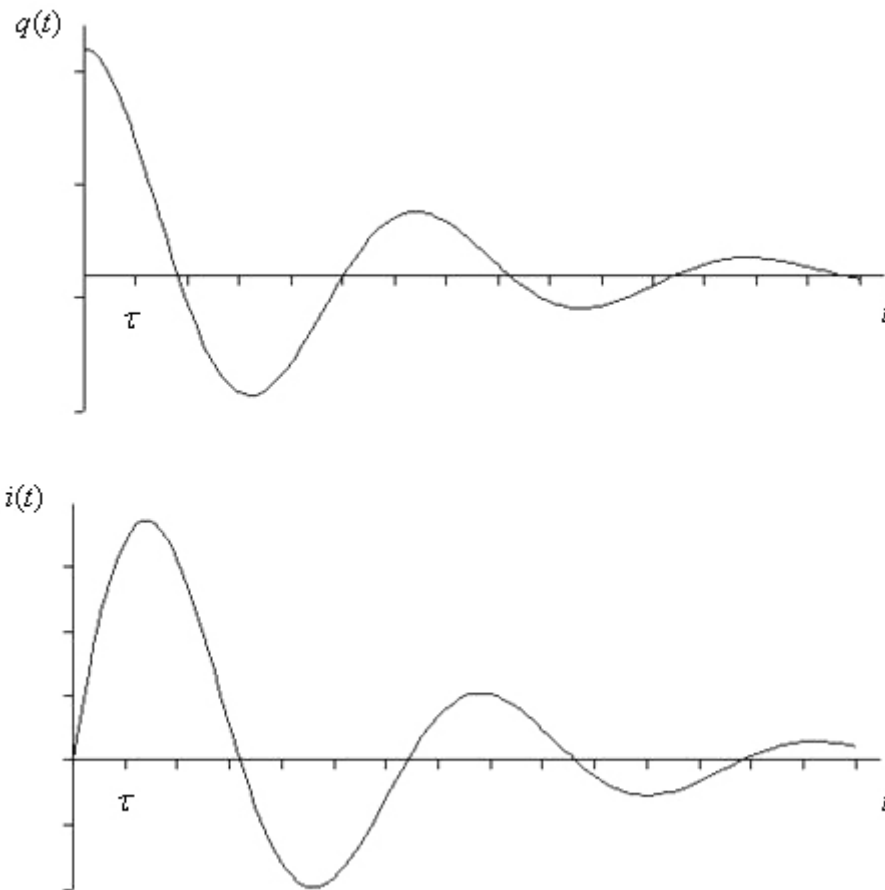


Figure: 71.12 - Comportement du courant dans le circuit lors de la décharge du condensateur

où:

$$\alpha = \frac{R}{2L} \quad (71.127)$$

est le "**facteur d'amortissement**". Si nous voulons avoir de belles oscillations peu amorties, il y a intérêt à avoir ce terme le plus petit possible donc une valeur de R petite.

Lorsque R est nul nous avons alors:

$$i(t) = -\omega q_0 \sin(\omega t) \quad (71.128)$$

avec donc:

$$\omega = \frac{\sqrt{R_C^2 - R^2}}{2L} = \frac{R_C}{2L} = \frac{2\sqrt{\frac{L}{C}}}{2L} = \sqrt{\frac{1}{LC}} \quad (71.129)$$

que nous appelons la "**pulsation propre**". Soit une période de:

$$T = \frac{2\pi}{\omega} = \frac{2\pi}{\sqrt{\frac{1}{LC}}} = 2\pi\sqrt{LC} \quad (71.130)$$

Il faut donc jouer alors avec C ou L pour obtenir la période désirée dans le cas où la résistance est nulle. Signalons également que cette situation particulière est appelée "**oscillateur harmonique**".

Enfin, de par les résultats obtenus, nous avons donc la généralisation des circuits RC, RL ou LC série.

Maintenant, supposons que dans le circuit nous posions une alimentation continue en série dans le circuit. Nous avons alors:

$$L \frac{d^2 q}{dt^2} + R \frac{dq}{dt} + \frac{1}{C} q = U_0 \quad (71.131)$$

L'équation différentielle linéaire à coefficients constants a maintenant un second membre (constant dans ce cas). Nous trouvons alors immédiatement une solution particulière qu'il suffit ensuite d'ajouter à toutes les solutions que nous avons obtenues précédemment.

Une solution particulière est donc:

$$q = C U_0 \quad (71.132)$$

Donc:

$$L \frac{d^2 q}{dt^2} + R \frac{dq}{dt} + \frac{1}{C} q = L \cdot 0 + R \cdot 0 + \frac{1}{C} q = \frac{1}{C} q = U_0 \quad (71.133)$$

Soit:

$$q = \frac{U_0}{C} \quad (71.134)$$

Cette solution particulière qui est à ajouter aux solutions précédentes, n'a aucune influence sur les équations du courant (sa dérivée étant nulle). Par contre, elle décale sur les graphiques le tracé de $q(t)$ vers le haut. Voilà donc l'effet qu'il y a à rajouter une source de tension constante (comme une simple pile par exemple).

4. CIRCUITS LINÉAIRES EN RÉGIME FORCÉ

L'objectif sera pour commencer, d'étudier le comportement d'un circuit linéaire RLC série excité par un générateur de tension sinusoïdale puisqu'il est une généralisation des circuits RL ou RC (il suffit d'annuler L ou C respectivement pour tomber sur les solutions d'un circuit RC ou RL).

Nous avons alors:

$$\boxed{L \frac{d^2 q}{dt^2} + R \frac{dq}{dt} + \frac{1}{C} q = U_0 \sin(\omega t)} \quad (71.135)$$

qui est une équation différentielle de forme connue et donc importante dans le domaine de l'acoustique car nous l'y retrouvons identiquement pour l'étude des haut-parleurs.

Nous pourrions très bien rajouter un déphasage au terme sinus à droite de l'égalité (arbitraire de phase). Cela ne changerait rien aux développements qui vont suivre et rappelons aussi que le cosinus n'est qu'un sinus avec un déphasage bien précis!

Enfin, le plus important, c'est que si nous trouvons une solution particulière à l'E.D. ci-dessus, alors puisque l'amplitude et la pulsation peuvent prendre n'importe quelle valeur à un déphasage arbitraire près alors nous avons donc une infinité de solutions particulières. Et comme nous avons démontré lors de notre étude des équations différentielles que la somme de solutions particulières est aussi solution alors cela signifie qu'une excitation obtenue avec une série de Fourier aura aussi une solution et en passant à la limite nous avons une transformée de Fourier!

Donc passons à notre étude. Pour cela, dérivons cette relation par rapport à t :

$$L \frac{d^2 i}{dt^2} + R \frac{di}{dt} + \frac{1}{C} i = U_0 \omega \cos(\omega t) \quad (71.136)$$

Cherchons alors une solution particulière de la forme:

$$\begin{aligned} i &= a \cos(\omega t) + b \sin(\omega t) \\ \frac{di}{dt} &= -a\omega \sin(\omega t) + b\omega \cos(\omega t) \\ \frac{d^2 i}{dt^2} &= -a\omega^2 \cos(\omega t) - b\omega^2 \sin(\omega t) \end{aligned} \quad (71.137)$$

Nous remarquons que cette proposition de solution est en tout point identique à la fondamentale d'une série de Fourier dont le terme a_0 est nul (qui est la moyenne du signal ou la composante continue si elle existe)!

Puis injectons ces relations dans:

$$L \frac{d^2 i}{dt^2} + R \frac{di}{dt} + \frac{1}{C} i = U_0 \omega \cos(\omega t) \quad (71.138)$$

en regroupant les termes trigonométriques de même nature:

$$\left(-a\omega^2 L + b\omega R + \frac{1}{C} a \right) \cos(\omega t) + \left(-b\omega^2 L - a\omega R + \frac{1}{C} b \right) \sin(\omega t) = U_0 \omega \cos(\omega t) \quad (71.139)$$

Ce qui est équivalent à:

$$\left(-a\omega^2 L + b\omega R + \frac{1}{C} a \right) \cos(\omega t) + \left(-b\omega^2 L - a\omega R + \frac{1}{C} b \right) \sin(\omega t) = U_0 \omega \cos(\omega t) + 0 \cdot \sin(\omega t) \quad (71.140)$$

d'où en identifiant les termes:

$$\begin{aligned} -a\omega^2 L + b\omega R + \frac{1}{C} a &= U_0 \omega \\ -b\omega^2 L - a\omega R + \frac{1}{C} b &= 0 \end{aligned} \quad (71.141)$$

Nous pouvons factoriser:

$$\begin{aligned} -a\left(\omega^2 L - \frac{1}{C}\right) + b\omega R &= U_0 \omega \\ -a\omega R - b\left(\omega^2 L - \frac{1}{C}\right) &= 0 \end{aligned} \quad (71.142)$$

Et en simplifiant par ω :

$$\begin{aligned} -a\left(\omega L - \frac{1}{C\omega}\right) + bR &= U_0 \\ -aR - b\left(\omega L - \frac{1}{C\omega}\right) &= 0 \end{aligned} \quad (71.143)$$

et en changeant de signe la deuxième ligne:

$$\begin{aligned} -a\left(\omega L - \frac{1}{C\omega}\right) + bR &= U_0 \\ aR + b\left(\omega L - \frac{1}{C\omega}\right) &= 0 \end{aligned} \quad (71.144)$$

Il s'agit donc d'un système de deux équations à deux inconnues a , b que nous résolvons en posant:

$$X = \omega L - \frac{1}{\omega C} = X_L - X_C \quad (71.145)$$

où nous retrouvons donc la réactance inductive et la réactance capacitive introduits dans le chapitre d'Électrocinétique.

Ce qui nous donne immédiatement:

$$\begin{aligned} -aX + bR &= U_0 \\ aR + bX &= 0 \end{aligned} \quad (71.146)$$

d'où:

$$-aX + \frac{-aR}{X} R = U_0 \Rightarrow a\left(\frac{X^2 + R^2}{X}\right) = -U_0 \Rightarrow a = \frac{-XU_0}{X^2 + R^2} \quad (71.147)$$

et donc:

$$\left(\frac{-XU_0}{X^2 + R^2}\right) R + bX = 0 \Rightarrow b = \frac{U_0 R}{X^2 + R^2} \quad (71.148)$$

Nous posons de plus traditionnellement que (nous verrons plus tard que cette expression pourra être assimilée au concept d'impédance par analogie avec la norme d'un vecteur):

$$Z = \sqrt{X^2 + R^2} \quad (71.149)$$

Ce qui donne la solution particulière suivante:

$$i(t) = \frac{-XU_0}{X^2 + R^2} \cos(\omega t) + \frac{U_0 R}{X^2 + R^2} \sin(\omega t) = \frac{U_0}{Z} \left(\frac{R}{Z} \sin(\omega t) - \frac{X}{Z} \cos(\omega t) \right) \quad (71.150)$$

à l'arbitraire de phase près.

Il est possible de trouver θ tel que:

$$\sin \theta = \frac{X}{Z} \quad \cos \theta = \frac{R}{Z} \quad (71.151)$$

Ou autrement écrit (ainsi on voit mieux que l'on balaie toutes les valeurs possibles hors singularités):

$$\frac{\sin \theta}{\cos \theta} = \frac{X}{R} \quad (71.152)$$

Nous avons alors en utilisant les relations trigonométriques remarquables (cf. chapitre de Trigonométrie):

$$i = \frac{U_0}{Z} \left(\cos(\theta) \sin(\omega t) - \frac{X}{Z} \sin(\theta) \cos(\omega t) \right) = \frac{U_0}{Z} \sin(\omega t - \theta) \quad (71.153)$$

θ est donc la phase du courant, soit l'avance ou le retard du courant sur la tension. Si $\theta = 0$ alors nous avons:

$$\sin(0) = 0 = \frac{X}{Z} \Rightarrow X=0 \quad \cos(0) = 1 = \frac{R}{Z} \Rightarrow R = Z = \sqrt{X^2 + R^2} = \sqrt{0^2 + R^2} = R \quad (71.154)$$

et donc:

$$i = \frac{U_0}{R} \sin(\omega t) \quad (71.155)$$

nous disons alors qu'il y a résonance du circuit avec donc:

$$X = L\omega - \frac{1}{C\omega} = 0 \Rightarrow \omega = \frac{1}{\sqrt{LC}} = \frac{2\pi}{R} \quad (71.156)$$

Soit lorsque la réactance inductive est égale à la réactance capacitive.

4.1. FILTRE PASSE-BAS PASSIF

Considérons le cas où L est nul. Nous avons alors:

$$X = L\omega - \frac{1}{C\omega} = \frac{1}{C\omega} \quad (71.157)$$

Donc:

$$Z = \sqrt{X^2 + R^2} = \sqrt{\left(\frac{1}{C\omega}\right)^2 + R^2} = \frac{1}{C\omega} \sqrt{1 + C\omega R^2} \quad (71.158)$$

D'où:

$$i = \frac{dq}{dt} = \frac{C\omega}{\sqrt{1 + (C\omega R)^2}} U_0 \sin(\omega t - \theta) \quad (71.159)$$

Soit:

$$\begin{aligned} q &= \int \frac{C\omega U_0}{\sqrt{1 + (C\omega R)^2}} \sin(\omega t - \theta) dt = \frac{C\omega U_0}{\sqrt{1 + (C\omega R)^2}} \int \sin(\omega t - \theta) dt \\ &= \frac{C\omega U_0}{\sqrt{1 + (C\omega R)^2}} \left(-\frac{\cos(\omega t - \theta)}{\omega} \right) = -\frac{CU_0}{\sqrt{1 + (C\omega R)^2}} \cos(\omega t - \theta) \end{aligned} \quad (71.160)$$

Nous avons alors aux bornes du condensateur:

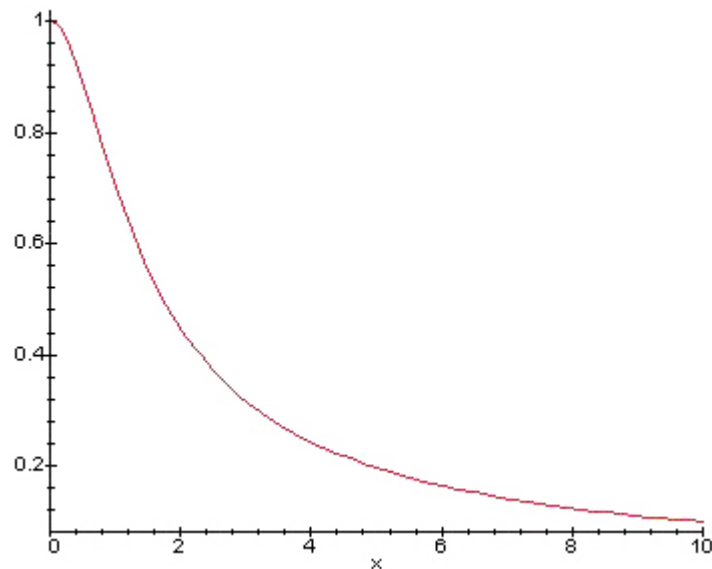
$$u_C = \frac{1}{C} q = -\frac{1}{\sqrt{1 + (C\omega R)^2}} U_0 \cos(\omega t - \theta) \quad (71.161)$$

Nous voyons donc que la tension aux bornes du condensateur fait office de ce que nous appelons un "**filtre passe-bas**". C'est-à-dire que l'amplitude de la tension aux bornes du condensateur par rapport à la tension d'excitation du circuit sera amoindrie et ce d'autant plus que la fréquence sera grande.

Ce genre d'outil est très pratique pour par exemple éliminer les harmoniques à haute fréquence d'un signal périodique obtenu par série de Fourier ou pour nettoyer un bruit à haute fréquence. On peut également utiliser des filtres passe-bas en cascade afin de réaliser des analyseurs de spectre.

Voici le tracé du facteur:

$$\frac{1}{\sqrt{1 + (C\omega R)^2}} \quad (71.162)$$



Nous voyons bien qu'aux basses fréquences (à gauche) l'amplitude est conservée (le filtre passe-bas laisse donc passer les basses fréquences). Au-delà, le signal est coupé.

Le rapport:

$$\frac{u_C}{U_0} = \frac{1}{\sqrt{1 + (C\omega R)^2}} \quad (71.163)$$

est souvent exprimé en décibels soit par définition en utilisant la mesure:

$$dB = 20 \log \left(\frac{u_C}{U_0} \right) \quad (71.164)$$

et porte alors le nom de "**fonction de transfert**" du filtre.

4.2. FILTRE PASSE-HAUT PASSIF

Pour ce qui est de la tension aux bornes de la résistance, nous avons:

$$u_R = R \cdot i_0 = \frac{RC\omega}{\sqrt{1 + (C\omega R)^2}} U_0 \sin(\omega t - \theta) \quad (71.165)$$

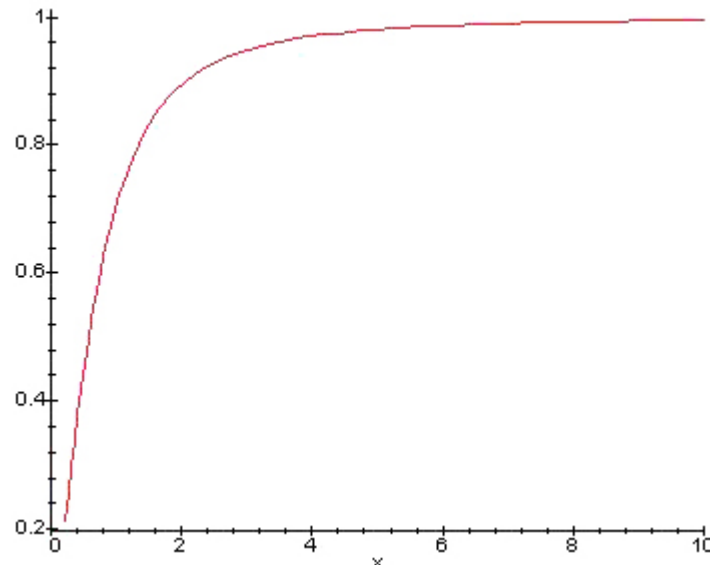
ce qui est traditionnellement remanié sous la forme suivante:

$$u_R = \frac{1}{\sqrt{1 + \frac{1}{(C\omega R)^2}}} U_0 \sin(\omega t - \theta) \quad (71.166)$$

Nous voyons donc que la tension aux bornes de la résistance fait office de ce que nous appelons un "**filtre passe-haut**". C'est-à-dire que l'amplitude de la tension aux bornes de la résistance par rapport à la tension d'excitation du circuit sera amoindrie et ce d'autant plus que la fréquence sera faible.

Voici le tracé du facteur:

$$\frac{1}{\sqrt{1 + \frac{1}{(C\omega R)^2}}} \quad (71.167)$$



Nous voyons bien qu'aux basses fréquences (à droite) l'amplitude est conservée (le filtre passe-haut laisse donc passer les hautes fréquences). Au-delà le signal est coupé.

Le rapport:

$$\frac{u_R}{U_0} = \frac{1}{\sqrt{1 + \frac{1}{(C\omega R)^2}}} \quad (71.168)$$

est souvent exprimé en décibels soit par définition en utilisant la mesure:

$$dB = 20 \log \left(\frac{u_R}{U_0} \right) \quad (71.169)$$

et porte alors aussi le nom de "**fonction de transfert**" du filtre.

Avec différents types de filtres assemblés nous pouvons ainsi supprimer (mais jamais complètement) des plages de fréquences. Nous parlons alors de filtre coupe-bandes. C'est la technique utilisée par exemple pour la réception d'une certaine radio ou chaîne de télévision se trouvant dans une plage de fréquence bien précise. Ou encore en musique électronique pour atténuer sons graves ou aigus. Ou encore pour séparer le signal ADSL de la voix d'une ligne téléphonique.

Un filtre passif se caractérise donc par l'usage exclusif de composants passifs linéaires (résistances, condensateurs, bobines couplées ou non). Par conséquent, leur gain (rapport de puissance entre la sortie et l'entrée) ne peut excéder l'unité. Ils ne peuvent donc qu'atténuer en partie des signaux, mais pas les amplifier, car cela nécessiterait un apport d'énergie (ce qui est le rôle des "filtres actifs").

4.3. INTÉGRATEUR ET DÉRIVATEUR

Nous avons donc aux bornes du condensateur:

$$\begin{aligned}
 u_C &= \frac{1}{C} \int i_0 dt = \frac{\omega}{\sqrt{1 + (C\omega R)^2}} \int U_0 \sin(\omega t - \theta) dt = \frac{1}{\sqrt{\frac{1 + (C\omega R)^2}{\omega^2}}} \int U_0 \sin(\omega t - \theta) dt \\
 &= \frac{1}{\sqrt{\frac{1}{\omega^2} + (CR)^2}} \int U_0 \sin(\omega t - \theta) dt
 \end{aligned} \tag{71.170}$$

Maintenant, si $\omega \gg \frac{1}{RC}$, nous avons:

$$u_C \cong \frac{1}{CR} \int U_0 \sin(\omega t - \theta) dt \tag{71.171}$$

Si nous faisons en sorte que $\theta = 0$ nous devons avoir:

$$X = L\omega - \frac{1}{C\omega} = 0 \tag{71.172}$$

Soit:

$$L = \frac{1}{C} \tag{71.173}$$

Dès lors:

$$u_C \cong \frac{1}{CR} \int U_0 \sin(\omega t) dt \tag{71.174}$$

Le circuit est alors ce que nous appelons assez logiquement... un "intégrateur".

Regardons maintenant du côté de la résistance:

$$u_R = R \cdot i_0 \tag{71.175}$$

Or, nous avons:

$$u_C = \frac{1}{C} \int i_0 dt \Rightarrow \frac{du_C}{dt} = \frac{1}{C} i_0 \Rightarrow i_0 = C \frac{du_C}{dt} \tag{71.176}$$

Donc:

$$u_R = RC \frac{du_C}{dt} \tag{71.177}$$

Comme:

$$u_C = -\frac{1}{\sqrt{1+(C\omega R)^2}} U_0 \cos(\omega t - \theta) \quad (71.178)$$

Nous avons alors:

$$u_R = -\frac{RC}{\sqrt{1+(C\omega R)^2}} \frac{d}{dt} U_0 \cos(\omega t - \theta) = -\frac{1}{\sqrt{\frac{1}{(RC)^2} + \omega^2}} \frac{d}{dt} U_0 \cos(\omega t - \theta) \quad (71.179)$$

Si $\omega \ll \frac{1}{RC}$ alors:

$$u_R = -RC \frac{d}{dt} U_0 \cos(\omega t - \theta) \quad (71.180)$$

Si nous nous arrangeons pour avoir:

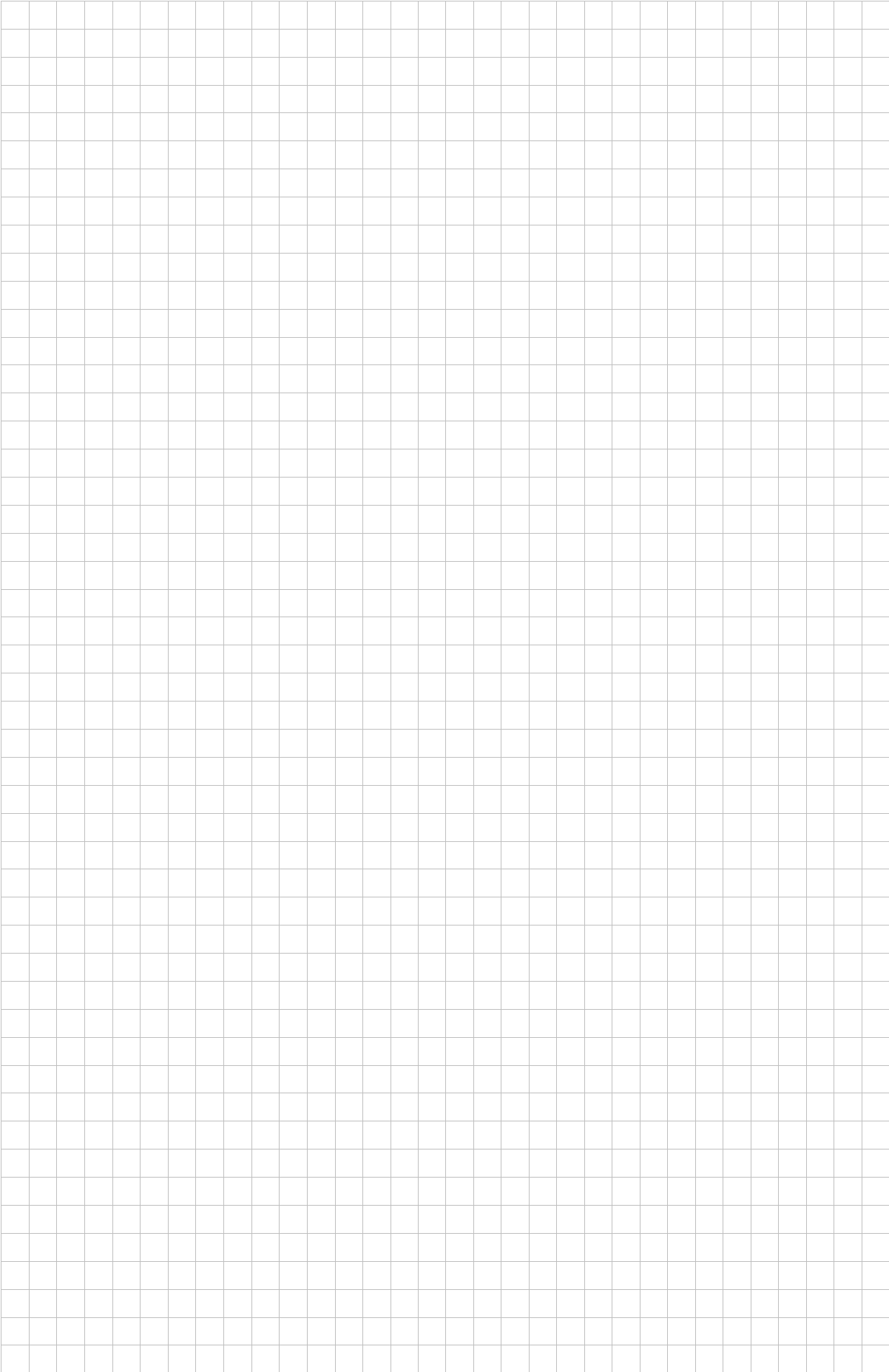
$$\theta = \frac{\pi}{2} \quad (71.181)$$

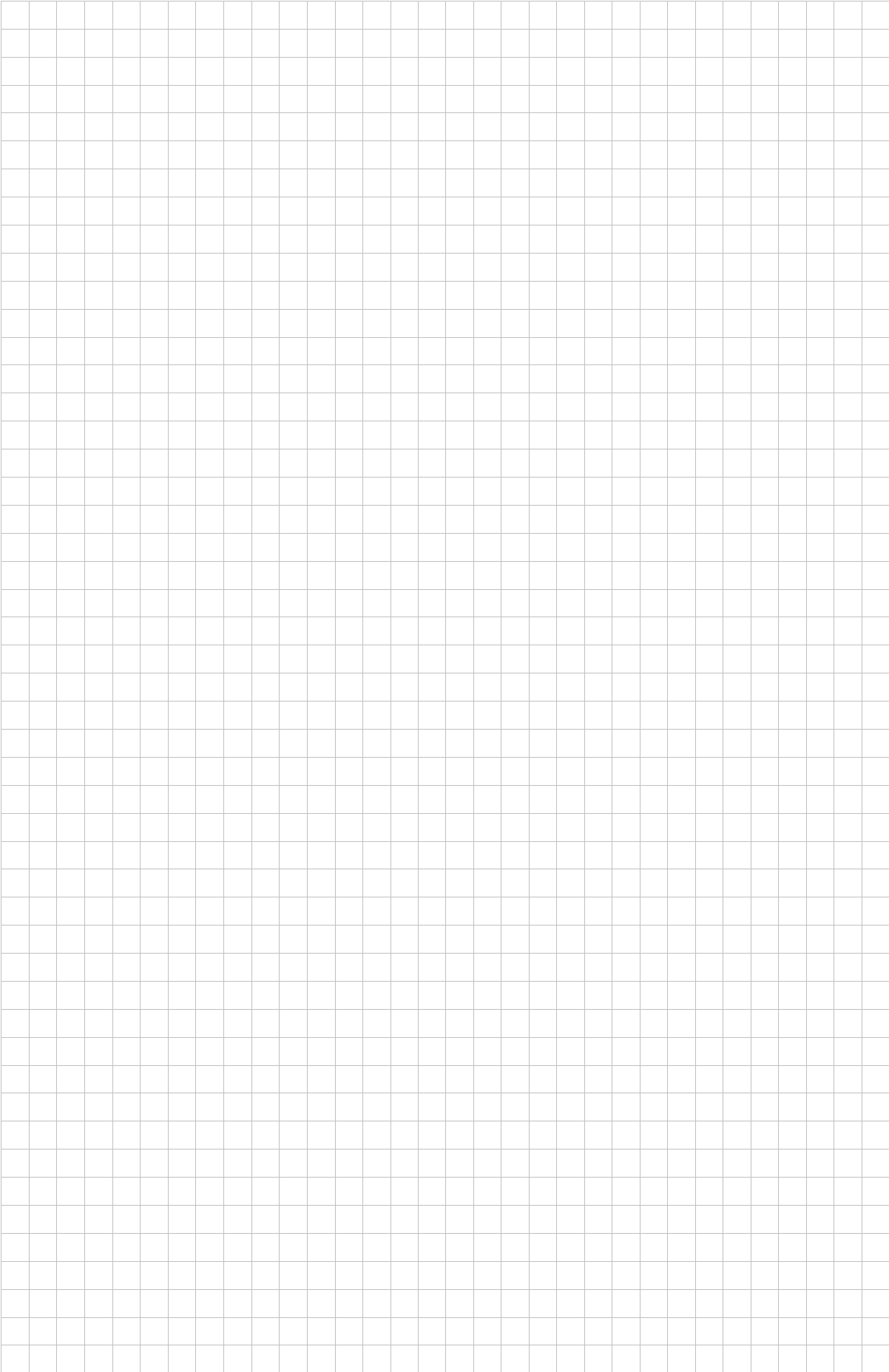
alors:

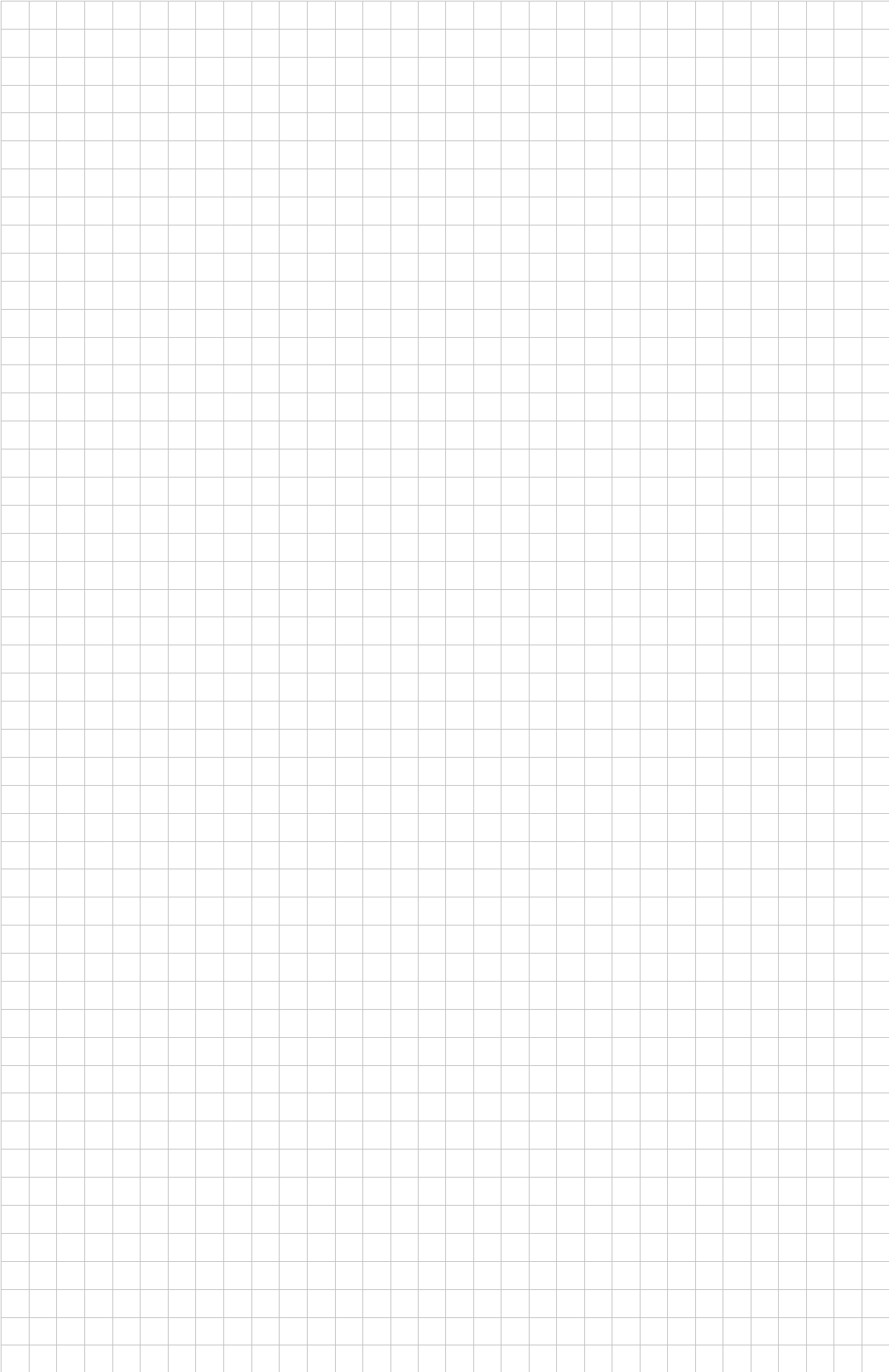
$$u_R = -RC \frac{d}{dt} U_0 \cos\left(\omega t - \frac{\pi}{2}\right) = -RC \frac{d}{dt} U_0 \cos\left(\omega t - \frac{\pi}{2}\right) = RC \frac{d}{dt} U_0 \sin(\omega t) \quad (71.182)$$

Le circuit est alors ce que nous appelons assez logiquement... un "dérivateur".

L'utilité d'un circuit intégrateur est par exemple de transformer un signal périodique en une constante (puisque la moyenne temporelle d'un signal périodique ayant un offset ne sera jamais nulle).







72. GÉNIE CIVIL

Le Génie Civil représente l'ensemble des techniques concernant les constructions civiles et les outils qui y sont associés. Les ingénieurs civils s'occupent eux de la conception, de la réalisation, de l'exploitation et de la réhabilitation d'ouvrages de construction et d'infrastructures urbaines dont ils assurent la gestion afin de répondre aux besoins de la société, tout en assurant la sécurité du public et la protection de l'environnement en théorie. Très variées et intéressantes, leurs réalisations se répartissent principalement dans cinq grands domaines d'intervention: structures, géotechnique, hydraulique, transport, et environnement. Comme à l'habitude sur ce site, nous allons nous concentrer uniquement sur la formalisation mathématique de cas très courants et qui ont une application pratique dans la vie de tous les jours

Signalons qu'en génie civil, il est parfois fait usage du calcul des surfaces minimales. Ceci est déjà traité par un exemple concret dans le chapitre de Mécanique Analytique. En ce qui concerne les efforts des poutres à la chaleur, ceci est déjà traité aussi dans le chapitre de Thermodynamique.

Nous allons nous limiter dans ce chapitre à des calculs démontrables de manière précise et non à des formules expérimentales afin d'avoir une introduction générale aux techniques du génie civil et de se familiariser avec le langage et certaines méthodes de calcul utilisées par les ingénieurs dans ce domaine. Dans la pratique, les ingénieurs en génie civil font usage d'un paquet de normes avec formules incluant des coefficients ou basées sur des tables ou encore des cartes géographiques. Par exemple, en Suisse, les ingénieurs en génie civil se réfèrent aux normes SIA (Société suisse des Ingénieurs et des Architectes) qui contiennent entre autres ces fameuses formules. L'approche théorique est évidemment insuffisante et toute formation dans ce domaine doit être complétée par des travaux pratiques en laboratoire.

Remarque: Il serait prétentieux de prétendre dans le présent chapitre vouloir faire aussi bien et complet que le PDF gratuit de *Statique* de Nicolet Gaston Raymond qui est inégalable en matière de contenu et de qualité à ce jour (même comparés aux ouvrages payants sur le sujet!). Il est donc fortement recommandé au lecteur de s'y référer s'il veut une information complète sur le génie civil (voir la section de téléchargement du site).

1. STATIQUE

Le génie civil utilise énormément la statique pour la construction. Nous n'allons donc pas récrire ici tout le chapitre de Mécanique Classique avec le principe fondamental de la statistique et tout ce qui va avec, ni l'analyse statique des poutres déjà effectuée dans le chapitre de Génie Mécanique, mais seulement présenter quelques aspects applicatifs du principe.

Pour commencer ce chapitre, intéressons-nous au moins au plus petit cas non trivial que l'on rencontre dans la pratique. Nous souhaiterions poser contre un mur un objet massif et nous souhaiterions savoir la force que devra supporter ce mur. Nous pouvons représenter cela basiquement par le schéma suivant:

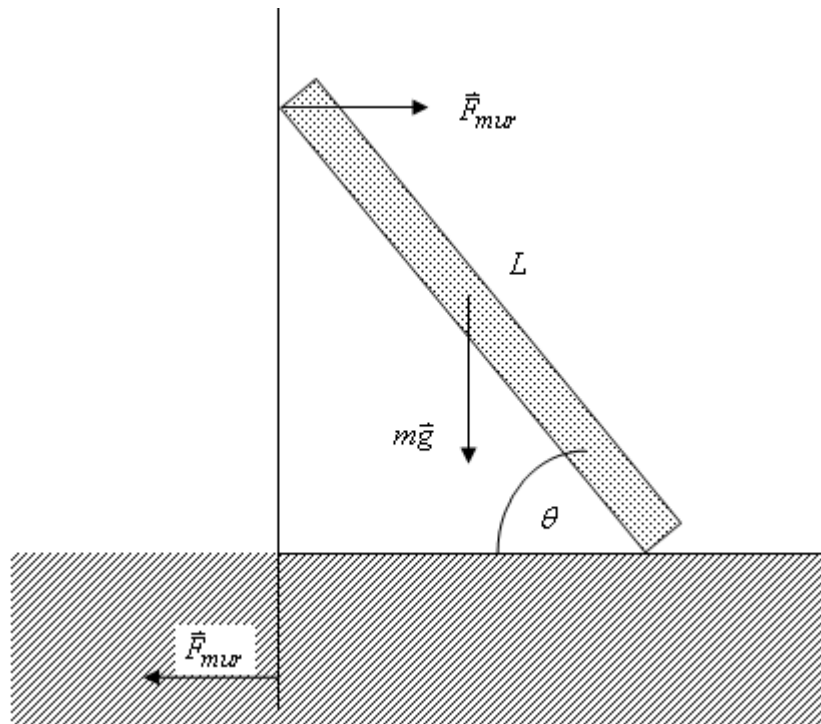


Figure: 72.1 - Objet massif contre un mur

Pour que le mur tienne, il faut (mais n'est pas une condition suffisante, elle est juste nécessaire) que le moment de force de la gravité égale le moment de force du mur. Nous avons alors:

$$mg \frac{L}{2} \cos \theta = F_{mur} L \sin \theta \quad (72.1)$$

d'où:

$$F_{mur} = \frac{1}{2} \frac{mg}{\tan \theta} \quad (72.2)$$

In extenso, il faut que les fondations du mur soient aptes à contrer le moment de force opposé.

2. POULIES

Dans le cadre de la statique des forces, il y a un exemple industriel qui est fameux et que nous croisons quasiment toutes les semaines en marchant ou en roulant devant des chantiers (grues), des gares (tendeurs), des ports (bateaux) ou en allant dans des salles de fitness ou garages: la poulie! Son origine est encore une idée d'Archimède (paraît-il...) qui l'appliqua pour le déplacement de grosses masses nécessaires dans divers chantiers de son époque. La relation avec le génie civil est donc toute justifiée! Voyons donc cela de plus près (exceptionnellement il y a très peu d'équations).

Considérons la situation suivante appelée "**poulie simple fixe**" avec une masse de 10 [Kg] (soit une force de 100 [N] avec la gravité terrestre arrondie à la dizaine la plus proche) accrochée une corde glissée dans la gouttière d'une poulie:

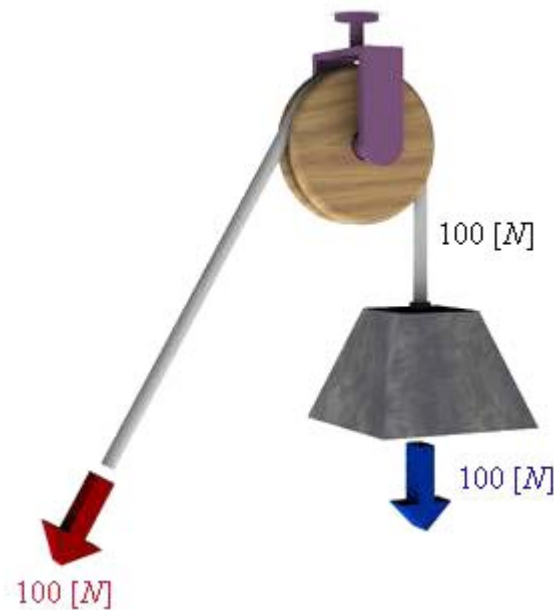
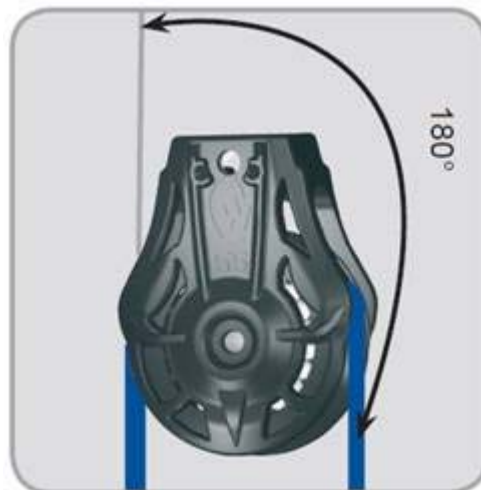


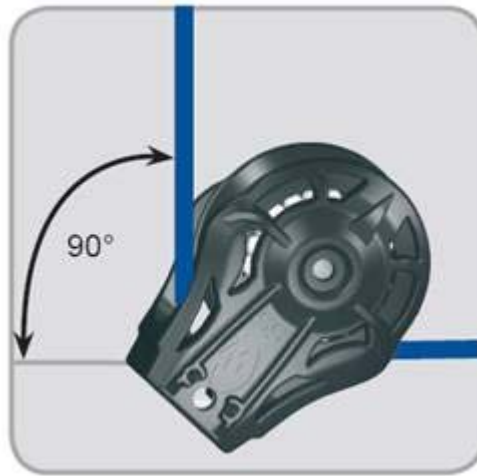
Figure: 72.2 - Poulie simple fixe(source: Wikipédia)

Une poulie simple fixe n'a l'avantage mécanique que de pouvoir exercer la force dans une direction différente à celle du déplacement, la force qui doit être appliquée est la même que celle qui est requise pour déplacer l'objet sans la poulie!

Le point d'ancrage de la poulie doit lui supporter la force nécessaire au déplacement de l'objet plus la force de traction, soit environ deux fois cette force au pire. Sinon le charge totale que doit supporter le point d'ancrage est fonction de "**l'angle de tire**" du cordage (compris entre 90° et 180°) bien évidemment:

Figure: 72.3 - Angle de tire à 180°

Pour un angle de 180° le coefficient de charge est de 200%. Une charge de 10 [Kg] sur le cordage représente une charge de 20 [Kg] sur la poulie.

Figure: 72.4 - Angle de tire à 90°

Pour un angle de 90° , le coefficient de charge est de 140%. Une charge de 10 [Kg] représente une charge de 14 [Kg] sur la poulie.

Considérons maintenant une situation où nous fixons une extrémité de la corde au support et de tirer avec l'autre extrémité, pour déplacer à la fois la poulie et la charge de 10 [Kg]. Cette configuration est appelée "**poulie simple mobile**" ou "**poulie inversée**" (la légende veut dire que c'est ce système qu'Archimède utilisait pour tirer un bateau):

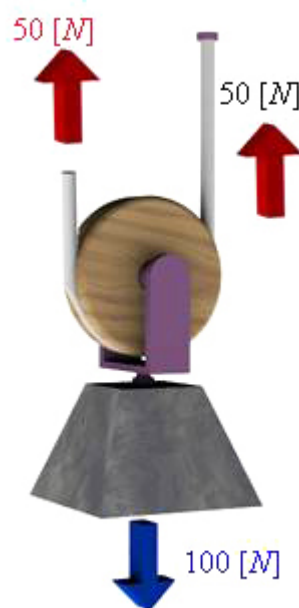


Figure: 72.5 - Poulie simple mobile (source: Wikipédia)

Au fait dans ce système (mis en place à la verticale ou à l'horizontale peu importe!) c'est comme s'il y avait deux individus qui se partageaient l'effort du déplacement: le mur et la partie libre de la corde (celle qui est tirée).

La poulie simple mobile permet donc de réduire la force nécessaire au déplacement de moitié (le point d'ancrage supportant l'autre moitié) et en rajoutant ainsi des poulies mobiles, nous continuons à diviser l'action à exercer! C'est bête mais il fallait y penser!

Ce système par contre nécessite un déplacement de l'extrémité de corde tirée du double de la distance du déplacement de la charge et ce indépendamment du rayon de la poulie.

Indiquons aussi que plus la poulie à un rayon grand, plus le moment de force le sera aussi! Donc dans le cas de très lourdes charges nous privilégieront de grands rayons pour les poulies si le système qui tire ne peut fournir qu'une faible force.

Une configuration plus réaliste (car nous allons rarement nous placer au-dessus du point d'ancrage pour tirer la corde et en plus le système précédent est peu stable mécaniquement parlant...) de la poulie libre présentée ci-dessus est la suivante:

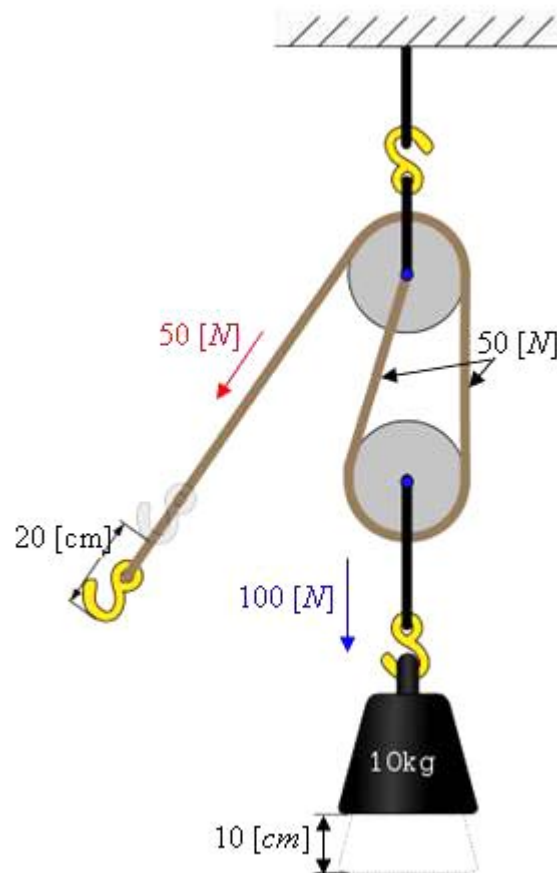


Figure: 72.6 - Poulie simple fixe et mobile mixte (source: Wikipédia)

Évidemment, quand nous représentons des systèmes comme ceux-ci dans les cas scolaires, nous négligeons de manière simplificatrice la masse des poulies elles-mêmes qu'il faudrait en toute rigueur prendre en compte!

Quand nous utilisons des systèmes de plusieurs poulies qui travaillent ensemble, nous disons que nous avons une configuration de "**poulies composées**". La configuration de ce type la plus commune est le "**palan**": les poulies sont distribuées en deux groupes (ou moufle), l'un fixe, l'autre mobile:

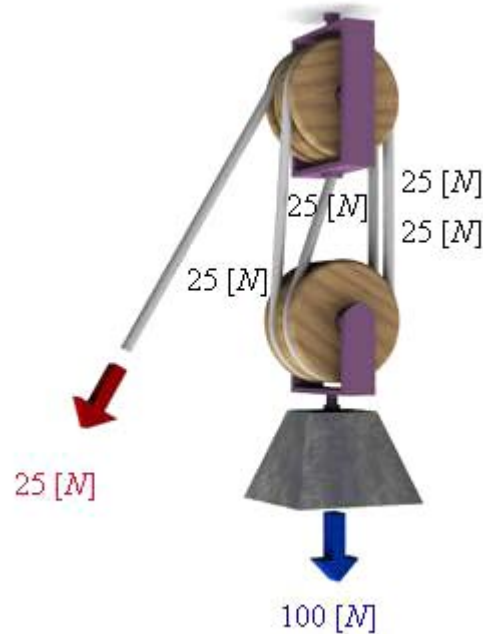


Figure: 72.7 - Système de poulies combinées fixes et mobiles (source: Wikipédia)

Dans chaque groupe nous installons un nombre arbitraire de poulies qui démultiplient donc d'un même facteur la charge initiale. La charge est bien évidemment unie au groupe mobile.

Nous avons donc 25 [N] au bout de la corde. Le lecteur peut donc chercher à s'amuser à trouver les 4 points d'accroches dans l'illustration précédente et les deux poulies qui divisent chacune par deux la force nécessaire... Si jamais voici la même configuration mais représentée sous une forme "dépliée":

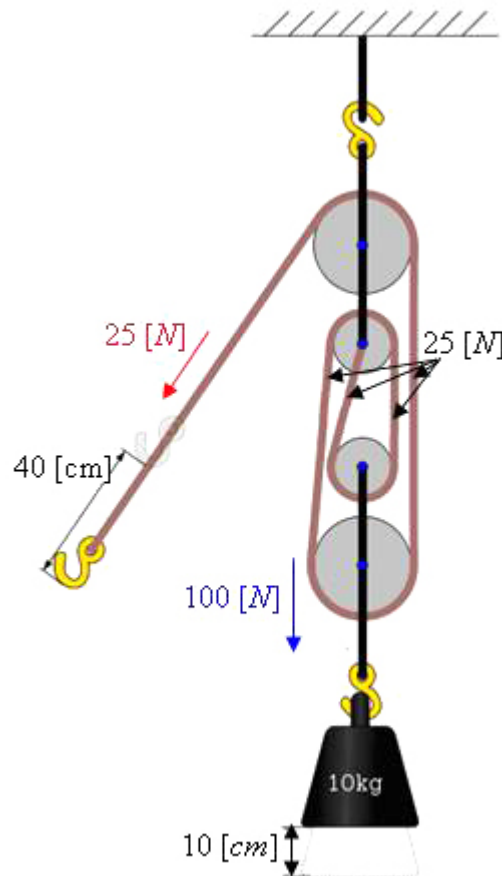


Figure: 72.8 - Système déplié de poulies combinées fixes et mobiles (source: Wikipédia)

Nous voyons déjà que la grosse poulie supérieure ne sert à rien excepté à changer la direction de la force de tirage. Au fait, les deux poulies qui servent à diviser chacune la force par deux sont les deux inférieures, le reste n'étant là que par commodité pour le mouvement de la corde.

Rappelons au lecteur qu'en réalité il faudrait prendre en compte les frottements de la corde sur la poulie (science de la tribologie) et que nous avons démontré dans le chapitre de Mécanique Classique que la force (tension) réelle d'une extrémité par rapport à l'autre était donnée par:

$$T_2 = T_1 e^{\mu_s \beta} \quad (30.3)$$

et donc que la force (tension) réelle utile pour soulever un poids était donnée par:

$$T_1 = \frac{T_2}{e^{\mu_s \beta}} < T_2 \quad (30.4)$$

Donc la force (tension) fourni pour soulever l'objet étant T_2 , la force (tension réelle) soulevant le poids étant T_1 , la différence donne la force (tension) faisant tourner la poulie par l'intermédiaire du frottement. Nous avons alors:

$$T_P = T_2 - T_1 = T_1 e^{\mu_s \beta} - T_1 = T_1 (e^{\mu_s \beta} - 1) \quad (30.5)$$

Le moment de force de la poulie de rayon R est alors:

$$M = R \cdot T_P = R \cdot T_1 \left(e^{\mu_s \beta} - 1 \right) = \frac{v}{\omega} T_1 \left(e^{\mu_s \beta} - 1 \right) \quad (30.6)$$

Voyons une application connue des poulies dans certaines gares ferroviaires:

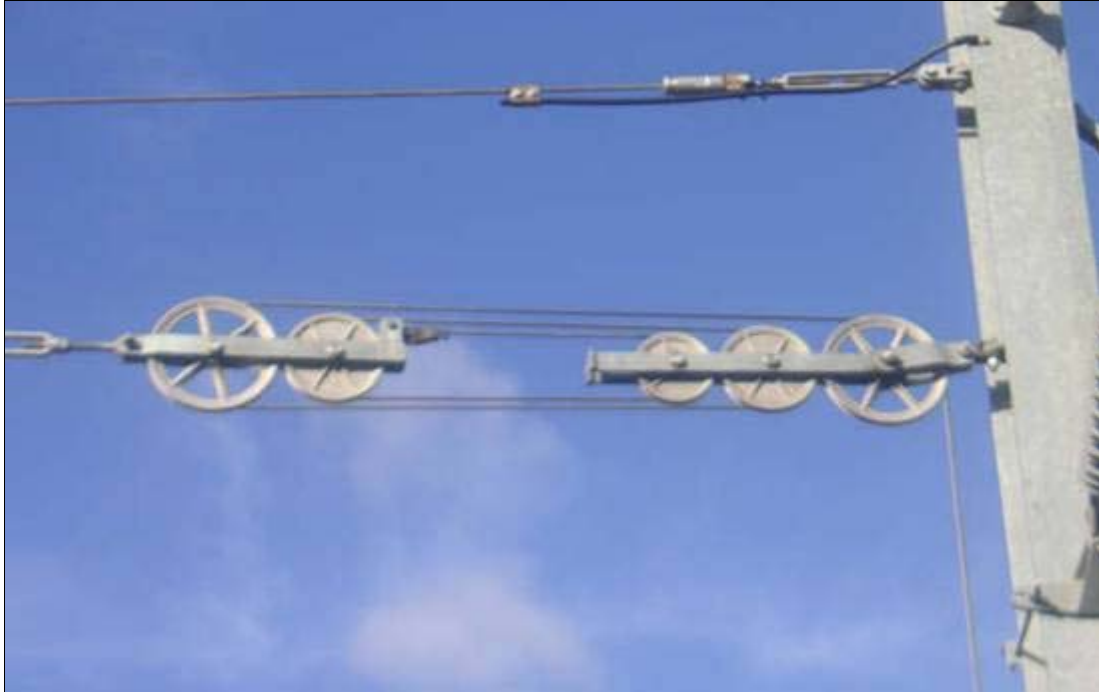


Figure: 72.9 - Système réel de poulies combinées fixes et mobiles

Il s'agit d'un palan pour tendre les câbles électriques avec un contrepoids non visible sur la photographie (en bas à droite) qui assure une certaine force donc une certaine tension. L'avantage de ce système est qu'il permet de rajouter des poids au fur et à mesure que le câble se détend et ceux-ci sont alors quatre fois plus élevés au niveau du câble électrique grâce aux deux poulies mobiles (à gauche). Les poulies à droite ne sont là que par commodité pour le mouvement de la corde et la direction de tirage en ce qui concerne la poulie à l'extrémité droite.

Dans le cas d'un levage horizontal ou vertical, il est facile de déterminer le rapport de démultiplication D . Effectivement, si nous considérons F la force nécessaire pour soulever l'objet d'une hauteur h en tirant la corde sur une longueur d et F_g la force de gravitation sur la masse tirée, nous avons alors en négligeant les frottements et le poids des poulies mobiles:

$$D = \frac{F}{F_g} = \frac{h}{d} \quad (72.7)$$

Enfin, remarquons qu'il est possible de jouer avec le rayon de la poulie de déviation pour diminuer la force à fournir tout en gardant constant le moment de force (nous parlons alors de "**palan différentiel**") mais au final l'énergie dépensée restera toujours la même pour soulever un objet à une même hauteur (et il faudra tirer la corde encore plus pour soulever la charge à la même hauteur).

3. SPIRALE DE CORNU

La clothoïde est une courbe transcendante plane dont la courbure est proportionnelle à l'abscisse curviligne. Elle est également appelée "**spirale de Cornu**", en référence à Alfred Cornu, le physicien français qui l'a inventée. Plus rarement, elle peut apparaître sous le nom de radioïde aux arcs, spirale d'Euler ou spirale de Fresnel.

Cette forme est également adaptée aux fins de courbes dans les tracés des chemins de fer parce qu'un véhicule suivant ce tracé à une vitesse constante subit une accélération angulaire constante, ce qui réduit à la fois les efforts sur les rails et l'inconfort des passagers dans les voitures.

Enfin, les sabots montés sur les pylônes de téléphériques, et qui supportent le câble porteur, adoptent cette forme. De fait, il est possible de faire circuler la cabine à sa vitesse maximale sur le pylône, sans incommoder les passagers.

De même cette courbe est utilisée pour les boucles verticales ou loopings dans les montagnes russes pour le confort des passagers, afin que l'accélération verticale subie soit continue



Figure: 72.10 - Quelques photos de spirales de Cornu routières

Lorsqu'un véhicule aborde une courbe circulaire,, il va subir une force $\vec{F} = m\vec{a}$ perpendiculaire à sa direction (force centrifuge) donc de norme (cf. chapitre de Mécanique Classique):

$$F = m \frac{v^2}{R} \quad (72.8)$$

dès le début de son entrée dans la courbe. Cet effet est problématique car pour une voiture de poids moyen sur une autoroute, la force centrifuge peut égaler la force de pesanteur (lorsque sa vitesse est dans les valeurs légales!).

Ainsi, l'accélération passe brutalement de 0 à v^2 / R aussi, les ingénieurs relèvent les courbes pour améliorer l'adhérence, mais il est aussi possible d'essayer de trouver des courbes pour lesquelles l'accélération sera plus progressive. Par exemple si la courbure C donnée par (cf. chapitre de Géométrie Différentielle):

$$C = \frac{1}{R} \quad (72.9)$$

est proportionnelle au trajet s (abscisse curviligne) parcouru dans la courbe, nous aurons au début de la courbe $C = 0$ donc l'accélération sera nulle

Ce que nous cherchons est alors des courbes telles que:

$$C = k \cdot s \quad (72.10)$$

Pour cela, rappelons que nous pouvons aussi écrire naturellement pour un cercle, la courbure sous la forme:

$$C = \frac{1}{R} = \frac{d\theta}{ds} \quad (72.11)$$

Effectivement, si nous tournons d'un angle $d\theta$ alors nous nous déplaçons d'une longueur $ds = R d\theta$ (cf. chapitre de Trigonométrie).

Nous avons donc la relation:

$$k \cdot s = \frac{d\theta}{ds} \quad (72.12)$$

Soit:

$$d\theta = k s ds \quad (72.13)$$

d'où:

$$\int_0^\theta d\theta = \int_0^s k s ds \Rightarrow \theta = \frac{1}{2} k s^2 = K s^2 \quad (72.14)$$

De plus rappelons que l'équation paramétrique du cercle est:

$$\begin{cases} x = R \cos \theta \\ y = R \sin \theta \end{cases} \quad (72.15)$$

Nous avons donc:

$$\begin{cases} dx = dR \cos \theta + R d(\cos \theta) = R \cos \theta d\theta = R \cos \theta ds / R \\ dy = dR \sin \theta + R d(\sin \theta) = R \sin \theta d\theta = R \sin \theta ds / R \end{cases} \quad (72.16)$$

Soit:

$$\begin{cases} \frac{dx}{ds} = \cos \theta \\ \frac{dy}{ds} = \sin \theta \end{cases} \quad (72.17)$$

Nous pouvons donc maintenant écrire:

$$\begin{cases} dx = \cos \theta ds = \cos(Ks^2) ds \\ dy = \sin \theta ds = \sin(Ks^2) ds \end{cases} \quad (72.18)$$

Soit:

$$\begin{cases} x = \int_0^s \cos(Kt^2) dt \\ y = \int_0^s \sin(Kt^2) dt \end{cases} \quad (72.19)$$

avec un petit changement de variables:

$$u^2 = Kt^2 \Rightarrow u = \sqrt{K}t \Rightarrow \frac{du}{dt} = \sqrt{K} \Rightarrow dt = \frac{du}{\sqrt{K}} \quad (72.20)$$

il vient:

$$\begin{cases} x = \frac{1}{\sqrt{K}} \int_0^s \cos(u^2) du \\ y = \frac{1}{\sqrt{K}} \int_0^s \sin(u^2) du \end{cases} \quad (72.21)$$

en prenant $x_0 = y_0 = 0$ (nous pouvons toujours faire une translation par la suite).

Les deux intégrales s'appellent des "**intégrales de Fresnel**" et ne sont pas calculables directement. Nous pouvons cependant les exprimer sous forme de développement de Taylor (cf. chapitre Suites Et Séries) sous la forme:

$$\begin{aligned} \sqrt{K}x &= s - \frac{s^5}{5 \cdot 2!} + \frac{s^9}{9 \cdot 4!} - \frac{s^{13}}{13 \cdot 6!} + \dots \\ \sqrt{K}y &= \frac{s^3}{3} - \frac{s^7}{7 \cdot 3!} + \frac{s^{11}}{11 \cdot 5!} - \dots \end{aligned} \quad (72.22)$$

Le plot de l'intégrale de Fresnel donne dans Maple 4.00b:

```
>plot([FresnelC(t),FresnelS(t),t=-5..5]);
```

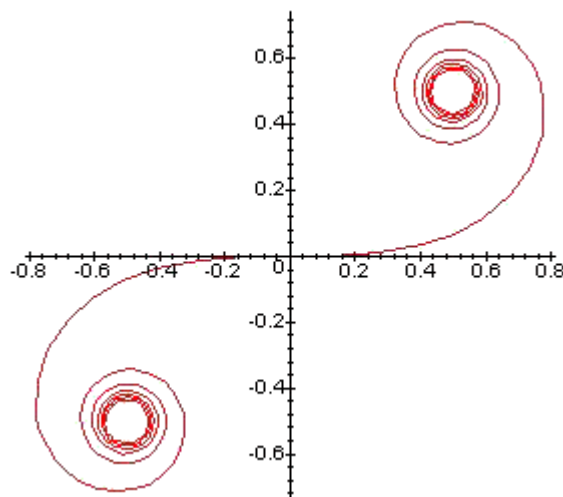


Figure: 72.11 - Plot de l'intégrale de Fresnel sous Maple 4.00b

En zoomant sur la partie qui nous intéresse:

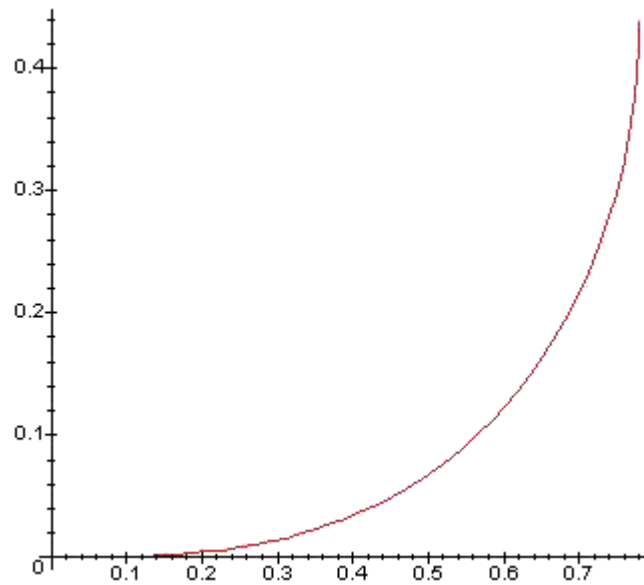


Figure: 72.12 - Zoom sur l'origine de l'intégrale de Fresnel

La même chose à une constante près en utilisant la série de Taylor présentée antérieurement:

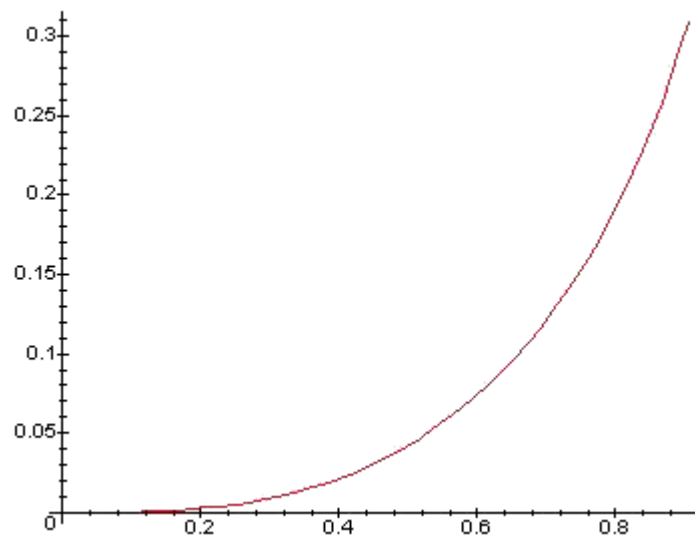


Figure: 72.13 - Développement en série de Taylor équivalent

Les bureaux d'ingénieurs utilisent des logiciels spéciaux intégrant des spirales des clothoïdes dans des environnements 2D ou 3D sur la base de relevés topographiques fait par des géomaticiens.

4. CÂBLES SUSPENDUS

Galilée, fut sans doute le premier à s'intéresser à la chaînette qu'il prit pour un arc de parabole. Jean Bernoulli, Huygens et Leibniz trouvèrent (indépendamment) en réponse au défi lancé par Jakob Bernoulli, sa véritable nature en 1691: engendrée par un cosinus hyperbolique.



Figure: 72.14 - Câbles suspendus des lignes HT (source: chronomaths)



Figure: 72.15 - Câble suspendu pour gazoduc par-dessus une rivière

4.1. CÂBLE SUSPENDU LIBRE (CHAÎNETTE)

Considérons (source: ChronoMath) pour l'étude un câble homogène libre, flexible, attaché en deux points A et B. Dans sa position d'équilibre, le câble pend dans un plan vertical et semble prendre une forme parabolique. En fait, pas vraiment...

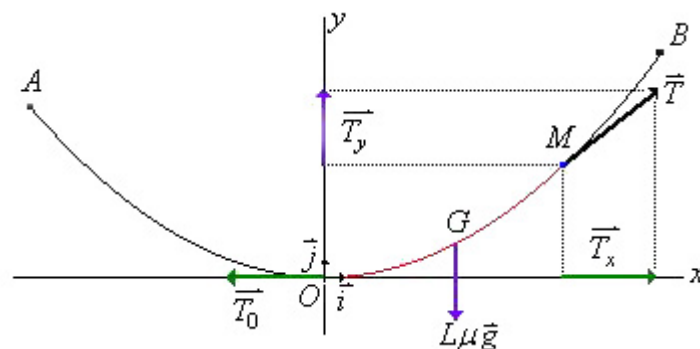


Figure: 72.16 - Configuration pour l'étude des câbles suspendus

Créons dans ce plan un repère orthonormé $(O, \vec{i}, \vec{j})$, où O désigne le point le plus bas du câble et notons $\vec{g}$ le champ de pesanteur à son endroit.

Appelons $\vec{T}_0$ la tension (force) au point O faisant éche à la tension en M de sorte que la portion de câble $[OM]$ de longueur L , soumise à son poids linéique $\mu \cdot \vec{g}$ au point G , soit en équilibre au sens statique:

$$\vec{T} + \vec{T}_0 + \mu L \vec{g} = \vec{0} \quad (72.23)$$

Remarque: Le produit $\mu \cdot \vec{g}$ de la masse linéaire avec la force gravitationnelle est aussi très souvent noté w comme nous l'avons fait dans le chapitre de Génie Mécanique pour représenter la charge linéique.

Projetons sur les axes de coordonnées en notant α l'angle $\widehat{i\vec{T}}$.

Nous avons alors les décompositions suivantes:

$$\begin{aligned} \vec{T} &= T_x \vec{i} + T_y \vec{j} = (\|\vec{T}\| \cos \alpha) \vec{i} + (\|\vec{T}\| \sin \alpha) \vec{j} \\ \vec{T}_0 &= -\|\vec{T}_0\| \vec{i} \\ \mu L \vec{g} &= -\mu L \|\vec{g}\| \vec{j} \end{aligned} \quad (72.24)$$

Nous pouvons alors écrire le système:

$$\begin{aligned} (\|\vec{T}\| \cos \alpha) \vec{i} - \|\vec{T}_0\| \vec{i} &= \vec{0} \\ (\|\vec{T}\| \sin \alpha) \vec{j} - \mu L \|\vec{g}\| \vec{j} &= \vec{0} \end{aligned} \quad (72.25)$$

Soit après simplification:

$$\begin{aligned} \|\vec{T}\| \cos \alpha - \|\vec{T}_0\| &= 0 \\ \|\vec{T}\| \sin \alpha - \mu L \|\vec{g}\| &= 0 \end{aligned} \quad (72.26)$$

Soit:

$$\begin{aligned} \|\vec{T}\| \cos \alpha &= \|\vec{T}_0\| \\ \|\vec{T}\| \sin \alpha &= \mu L \|\vec{g}\| \end{aligned} \quad (72.27)$$

En calculant le rapport:

$$\tan \alpha = \frac{\mu L \|\vec{g}\|}{\|\vec{T}_0\|} = L \frac{\mu \|\vec{g}\|}{\|\vec{T}_0\|} = \frac{L}{k} \quad (72.28)$$

Pour obtenir une équation différentielle différentions... (là c'est subtil...):

$$d(\tan \alpha) = \frac{dL}{k} \quad (72.29)$$

où dL est donc l'abscisse curviligne du câble (souvent notée ds dans la littérature conformément à ce qui se fait en géométrie différentielle).

Ensuite:

$$d(\tan \alpha) = \frac{\sqrt{dx^2 + dy^2}}{k} = dx \frac{\sqrt{1+y'^2}}{k} \quad (72.30)$$

Mais la tangente c'est aussi la dérivée de la fonction décrivant la chaînette. Donc:

$$d(\tan \alpha) = d(y') = dx \frac{\sqrt{1+y'^2}}{k} \quad (72.31)$$

Il vient alors:

$$\frac{y''}{\sqrt{1+y'^2}} = \frac{1}{k} \quad (72.32)$$

Suite à l'intervention d'un lecteur, nous allons proposer deux manières de résoudre cette équation différentielle. La première est celle d'origine et elle est un peu compliquée et la deuxième (disponible beaucoup plus bas) est celle proposée par un lecteur qui est peut-être plus élégante et simple.

Première approche:

Posons $u = y'$ et cherchons la primitive du membre de gauche dans un premier temps (celle du membre droite étant évidente). Les calculs faits dans le chapitre de Calcul Différentiel Et Intégral dans la détermination des primitives usuelles nous donnent:

$$\int \frac{u'}{\sqrt{1+u^2}} du = u' \int \frac{1}{\sqrt{1+u^2}} du = \ln(u + \sqrt{1+u^2}) + c^{te} \quad (72.33)$$

Nous avons donc:

$$\ln(y' + \sqrt{1+y'^2}) = \frac{x}{k} + c^{te} \quad (72.34)$$

en passant à l'exponentielle:

$$y' + \sqrt{1+y'^2} = e^{\frac{x}{k} + c^{te}} \quad (72.35)$$

en remarquant que dans notre problème en $x = 0$ nous avons bien $y' = 0$.

Pour trouver y' nous utilisons une astuce: Nous savons que la fonction est symétrique. Donc si nous remplaçons x par $-x$ la tangente change aussi de signe et passe de y' à $-y'$:

$$-y' + \sqrt{1 + (-y')^2} = e^{\frac{-x}{k} + c^*} \quad (72.36)$$

En soustrayant:

$$y' + \sqrt{1 + y'^2} = e^{\frac{x}{k} + c^*} \text{ et } -y' + \sqrt{1 + y'^2} = e^{\frac{-x}{k} + c^*} \quad (72.37)$$

il vient:

$$2y' = e^{\frac{x}{k} + c^*} - e^{\frac{-x}{k} + c^*} \Rightarrow y' = \frac{e^{\frac{x}{k} + c^*} - e^{\frac{-x}{k} + c^*}}{2} \quad (72.38)$$

Donc après intégration:

$$y = k \frac{e^{\frac{x}{k} + c^*} - e^{\frac{-x}{k} + c^*}}{2} = k \cosh\left(\frac{x}{k} + c_1^{te}\right) + c_2^{te} \quad (72.39)$$

où l'expression après la première égalité est appelée "**courbe caténaire**". Il vient au final:

$$y = \frac{T_0}{\mu \cdot g} \cosh\left(x \frac{\mu \cdot g}{T_0} + c_1^{te}\right) + c_2^{te} \quad (72.40)$$

où les constante seront déterminées par les conditions initiales.

Nous voyons bien avec Maple 4.00b la différence entre une parabole et la chaînette:

```
> plot([x^2, cosh(x)], x=-4...4);
```

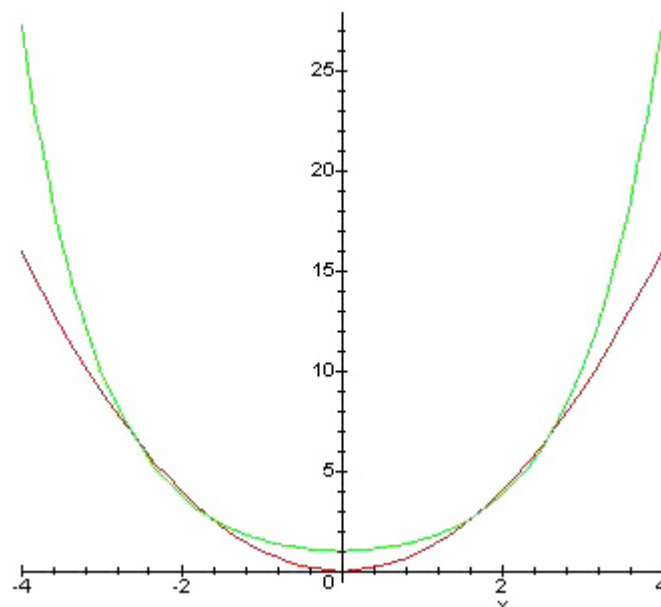


Figure: 72.17 - Plot entre une chaînette et parabole avec Maple 4.00b

Considérons maintenant deux points dans le plan (x_1, y_1) , (x_2, y_2) et déterminons l'équation de la chaînette de longueur L ayant ces deux points comme extrémités.

Nous avons les deux équations:

$$\begin{cases} y_1 = k \cosh\left(\frac{x_1}{k} + c_1^{te}\right) + c_2^{te} \\ y_2 = k \cosh\left(\frac{x_2}{k} + c_1^{te}\right) + c_2^{te} \end{cases} \quad (72.41)$$

Nous obtenons une troisième équation à l'aide de la longueur L qui est connue. En effet (cf. chapitre de Mécanique Analytique):

$$\begin{aligned} L &= \int_{x_1}^{x_2} \sqrt{1 + y'^2} dx = \int_{x_1}^{x_2} \sqrt{1 + \sinh^2\left(\frac{x}{k} + c_1^{te}\right)} dx = \int_{x_1}^{x_2} \cosh\left(\frac{x}{k} + c_1^{te}\right) dx \\ &= k \left(\sinh\left(\frac{x_2}{k} + c_1^{te}\right) - \sinh\left(\frac{x_1}{k} + c_1^{te}\right) \right) \end{aligned} \quad (72.42)$$

où nous avons toujours:

$$k = \frac{T_0}{\mu \cdot g} \quad (72.43)$$

Ainsi, nous obtenons un système non linéaire de trois équations à trois inconnues (k, c_1^{te}, c_2^{te}) :

$$\begin{cases} y_1 = k \cosh\left(\frac{x_1}{k} + c_1^{te}\right) + c_2^{te} \\ y_2 = k \cosh\left(\frac{x_2}{k} + c_1^{te}\right) + c_2^{te} \\ L = k \left(\sinh\left(\frac{x_2}{k} + c_1^{te}\right) - \sinh\left(\frac{x_1}{k} + c_1^{te}\right) \right) \end{cases} \quad (72.44)$$

Déterminons à titre d'exemple la chaînette de longueur 38 cm passant par les points $(-9, 0)$, $(9, 10)$.

Il faut alors résoudre le système suivant:

$$\begin{cases} 0 = k \cosh\left(\frac{-9}{k} + c_1^{te}\right) + c_2^{te} \\ 10 = k \cosh\left(\frac{9}{k} + c_1^{te}\right) + c_2^{te} \\ 38 = k \left(\sinh\left(\frac{9}{k} + c_1^{te}\right) - \sinh\left(\frac{-9}{k} + c_1^{te}\right) \right) \end{cases} \quad (72.45)$$

Voici les commandes Maple 4.00b qui nous permettent d'obtenir le résultat.

```
>e1:=0=k*cosh(-9/k+c1)+c2;
>e2:=10=k*cosh(9/k+c1)+c2;
>e3:=38=k*(sinh(9/k+c1)-sinh(-9/k+c1));
>fsolve({e1,e2,e3},{k,c1,c2},{k=0..infinity});
```

Maple donne:

$$k = 4.073758798, c1 = .2694982504, c2 = -14.46356329$$

Graphiquement nous avons alors:

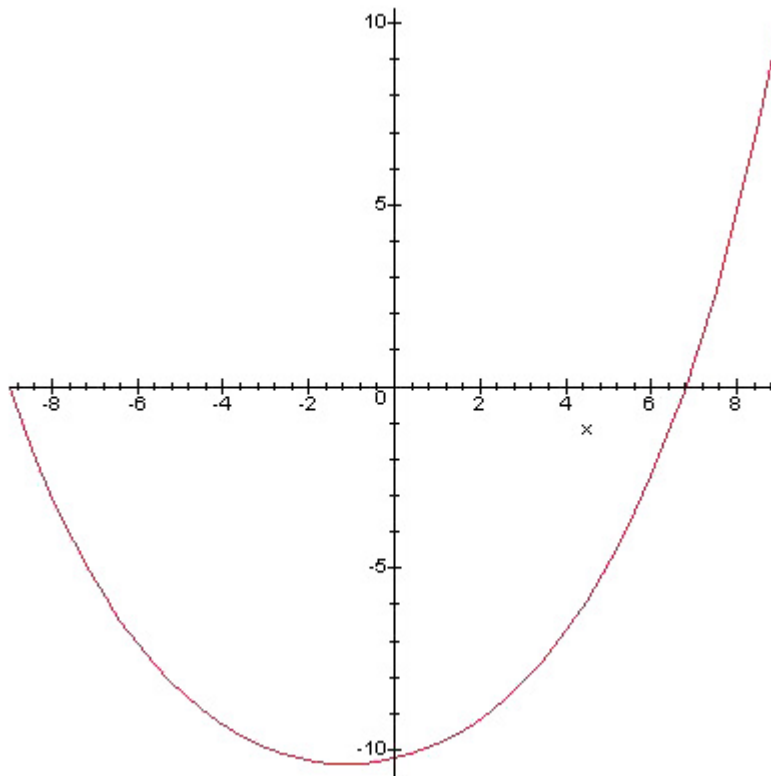


Figure: 72.18 - Tracé d'une petite chaînette avec Maple 4.00b

Deuxième approche:

Pour cette deuxième approche de résolution de l'équation différentielle, nous allons garder la notation proposée par le lecteur qui nous a contacté. Nous repartons de l'équation différentielle:

$$\frac{d^2 y}{dx^2} = \frac{1}{k} \sqrt{1 + \left(\frac{dy}{dx} \right)^2} \quad (72.46)$$

Nous faisons le changement de variable:

$$u = \frac{dy}{dx} \quad (72.47)$$

Il vient alors:

$$\frac{du}{dx} = \frac{1}{k} \sqrt{1+u^2} \quad (72.48)$$

Puis:

$$\frac{du}{\sqrt{1+u^2}} = \frac{dx}{k} \quad (72.49)$$

L'intégration donne conformément à la primitive usuelle démontrée dans le chapitre de Calcul Différentiel Et Intégral:

$$\operatorname{arcsinh}(u) = \frac{x}{k} + c^{te} \quad (72.50)$$

Et la condition:

$$\frac{dy}{dx} = u = 0 \quad (72.51)$$

en $x = 0$ (point le plus bas de la chaînette) impose que la constante d'intégration est nulle et donc que les points d'accroche sont par symétrie situés à même hauteur. Nous avons alors:

$$\operatorname{arcsinh}(u) = \frac{x}{k} \Rightarrow u = \frac{dy}{dx} = \sinh\left(\frac{x}{k}\right) \quad (72.52)$$

Soit:

$$y = k \cosh\left(\frac{x}{k}\right) + c^{te} \quad (72.53)$$

à comparer avec la méthode précédente où nous avons obtenu:

$$y = \frac{T_0}{\mu \cdot g} \cosh\left(x \frac{\mu \cdot g}{T_0} + c_1^{te}\right) + c_2^{te} \quad (72.54)$$

La constante d'intégration est déterminée par les points d'attache de la corde distants de D où nous avons $y = 0$ en $x = -D/2$ et en $x = D/2$ (donc les extrémités sont sur la même horizontale). Nous avons alors:

$$y(x) = k \left[\cosh\left(\frac{x}{k}\right) - \cosh\left(\frac{D}{2k}\right) \right] \quad (72.55)$$

C'est sous cette dernière forme que l'équation de la chaînette a été obtenue indépendamment par Leibniz, Huyghens et Bernoulli à la fin du 17ème siècle.

On peut alors maintenant facilement calculer la déviation maximale (le "flèche" comme disent les ingénieurs en génie civil que nous noterons f comme dans le chapitre de Génie Mécanique) par rapport à l'horizontale passant par les points d'attache:

$$y(0) = f = k \left[1 - \cosh \left(\frac{D}{2k} \right) \right] \quad (72.56)$$

Enfin il peut être intéressant de calcul la longueur du câble à l'équilibre. Pour cela, rappelons que nous avons démontré plus haut que:

$$\begin{aligned} L &= \int_{x_1}^{x_2} \sqrt{1 + y'^2} dx = \int_{x_1}^{x_2} \sqrt{1 + \sinh^2 \left(\frac{x}{k} + c_1^{te} \right)} dx = \int_{x_1}^{x_2} \cosh \left(\frac{x}{k} + c_1^{te} \right) dx \\ &= k \left(\sinh \left(\frac{x_2}{k} + c_1^{te} \right) - \sinh \left(\frac{x_1}{k} + c_1^{te} \right) \right) \end{aligned} \quad (72.57)$$

Dans le cas où les attaches sont à même hauteur, la chaînette est symétrique et nous avons alors:

$$L = 2 \cdot \int_0^{D/2} \cosh \left(\frac{x}{k} + c_1^{te} \right) dx = 2k \left(\sinh \left(\frac{1}{k} \frac{D}{2} \right) - \sinh \left(c_1^{te} \right) \right) \quad (72.58)$$

Reste à déterminer la constante! Si D est nul alors L doit être nul. La constante vaut alors zéro et il reste:

$$L = 2k \sinh \left(\frac{1}{k} \frac{D}{2} \right) \quad (72.59)$$

Dans le cas des lignes de chemin de fer électrifiées, nous pallions à la flèche rédhibitoire par un câble porteur principal de la caténaire: le câble supérieur (ci-dessous à droite) subit une flèche acceptée, ce qui diminue les tensions entre les pylônes. La caténaire reste ainsi bien linéaire grâce aux accroches auxiliaires multiples à un câble auxiliaire.



Figure: 72.19 - Chaînette des Chemins de Fer (source: Chronomaths)

Sinon signalons que nous retrouvons aussi les chaînettes dans tous les endroits de la vie de tous les jours où un câble est suspendu entre deux points sur une même horizontale.

Exemple:

Considérons un câble suspendu avec les données suivantes:

$$\begin{aligned}\mu &= 0.5 \left[\text{Kg} \cdot \text{m}^{-1} \right] \\ g &= 9.81 \left[\text{m} \cdot \text{s}^{-2} \right] \\ D &= 20 \left[\text{m} \right] \\ T_0 &= 45 \left[\text{N} \right]\end{aligned} \quad (72.60)$$

Nous avons puisque la charge linéique du câble est constante (en reprenant la notation du chapitre de Génie Mécanique au passage):

$$k = \frac{T_0}{\mu \cdot g} = \frac{T_0}{w_0} \cong 9.17 \left[\text{m} \right] \quad (72.61)$$

et donc:

$$y(0) = f = k \left[1 - \cosh \left(\frac{D}{2k} \right) \right] = 9.17 \left[1 - \cosh \left(\frac{20}{2 \cdot 9.17} \right) \right] \cong -6 \left[\text{m} \right] \quad (72.62)$$

Ainsi la flèche du câble est 6 mètres en-dessous de l'horizontale des deux accroches.

4.2. CÂBLE SUSPENDU PORTEUR (PONT SUSPENDU)

Considérons le pont suspendu suivant portant une charge linéique constante, donc la flèche h est imposée par l'architecte, ainsi que la distance D entre les 2 piliers:

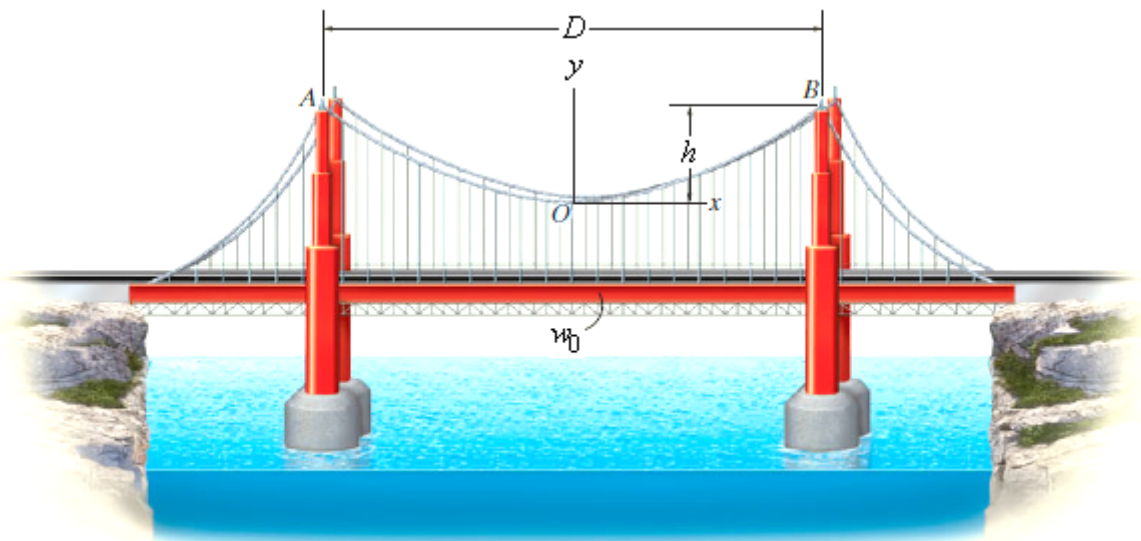


Figure: 72.20 - Pont suspendu (source: ISBN 0-13-814929-1)

La subtilité de cas d'étude réside dans le fait que la charge linéique n'est plus dans le câble lui-même mais dans la travée du pont qui est de masse linéique bien supérieur. Nous ne pouvons plus alors utiliser les analyses faites plus haut. Le développement fait pour le câble suspendu normal reste valide jusqu'à la relation:

$$d(\tan \alpha) = \frac{dL}{k} \quad (72.63)$$

Maintenant, il faut se rappeler que nous avons aussi:

$$d(\tan \alpha) = d\left(\frac{dy}{dx}\right) = \frac{dL}{k} \quad (72.64)$$

en réarrangeant et en dérivant encore par dx :

$$\frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{1}{k} \frac{dL}{dx} \quad (72.65)$$

Sur le pont cependant chaque portion dL du câble est négligeable devant la portion dx du pont. Ceci marque toute la différence entre le pont suspendu et la simple chaînette tel que nous devons remplacer dL par dx (c'est assez subtile avec cette démarche mais il y a plusieurs approches possible à ce développement). Nous avons alors la relation précédente qui devient:

$$\frac{d^2 y}{dx^2} = \frac{1}{k} \frac{dx}{dx} = \frac{1}{k} \quad (72.66)$$

Soit après réarrangement:

$$d^2 y = \left(\frac{1}{k} dx\right) dx \quad (72.67)$$

En intégrant une première fois il vient alors:

$$dy = \left(\frac{1}{k} dx\right) x + C^{te} \quad (72.68)$$

et une deuxième et dernière fois:

$$y = \frac{1}{2k} x^2 + C_1 x + C_2 \quad (72.69)$$

Les constantes sont déterminées par les conditions initiales. L'endroit où nous avons posé le repère impose que $y = 0$ en $x = 0$ et que $dy/dx = 0$ en $x = 0$, nous avons les deux constante qui sont nulles et il reste alors:

$$y = \frac{1}{2k} x^2 = \frac{\mu \cdot g}{2T_0} x^2 = \frac{w_0}{2T_0} x^2 \quad (72.70)$$

Il s'agit donc de l'équation d'une parabole et non plus d'une chaînette! Un petit souci est que nous ne connaissons pas la tension dans le câble. Il faudrait nous en débarrasser. Or du fait que l'architecte impose que la flèche soit h en $x = D/2$, il vient alors:

$$T_0 = \frac{w_0 D^2}{8h} \quad (72.71)$$

et donc:

$$y = \frac{4h}{D^2} x^2 \quad (72.72)$$

Encore une fois cela reste une parabole et ce indépendamment du poids porté! Cela étant dû au fait qu'il soit uniforme.

Il nous reste alors à déterminer la longueur du câble. Pour cela nous reprenons la relation déjà rappelée plus haut (démontrée dans le chapitre de Mécanique Analytique ou de Formes Géométriques):

$$L = \int_{x_1}^{x_2} \sqrt{1 + y'^2} dx \quad (72.73)$$

Par symétrie de la fonction du pont suspendu nous pouvons écrire:

$$\begin{aligned} L &= \int_{x_1}^{x_2} \sqrt{1 + y'^2} dx = 2 \int_0^{D/2} \sqrt{1 + \left(\frac{4h}{D^2} x^2\right)^2} dx = 2 \int_0^{D/2} \sqrt{1 + \left(\frac{8h}{D^2} x\right)^2} dx \\ &= 2 \frac{8h}{D^2} \int_0^{D/2} \sqrt{\left(\frac{D^2}{8h}\right)^2 + x^2} dx \end{aligned} \quad (72.74)$$

Dans le chapitre de Calcul Différentiel Et Intégral lors de nos démonstrations des primitives usuelles, nous avons démontré que:

$$\int \sqrt{a^2 + x^2} dx = \frac{a^2}{2} \ln \left(x + \sqrt{a^2 + x^2} \right) + \frac{x}{2} \sqrt{a^2 + x^2} \quad (72.75)$$

Il vient alors:

$$L = \frac{16h}{D^2} \int \sqrt{\left(\frac{D^2}{8h}\right)^2 + x^2} dx = \frac{16h}{D^2} \left[\frac{1}{2} \left(\frac{D^2}{8h}\right)^2 \ln \left(x + \sqrt{\left(\frac{D^2}{8h}\right)^2 + x^2} \right) + \frac{x}{2} \sqrt{\left(\frac{D^2}{8h}\right)^2 + x^2} \right]_0^{D/2} \quad (72.76)$$

Exemple:

Considérons les caractéristiques du Golden Bridge de San Francisco. Sa flèche h est d'environ 230 mètres et sa portée principale est de 1280 [m]. Nous avons alors:

$$L = \frac{16h}{D^2} \left[\frac{1}{2} \left(\frac{D^2}{8h} \right)^2 \ln \left(x + \sqrt{\left(\frac{D^2}{8h} \right)^2 + x^2} \right) + \frac{x}{2} \sqrt{\left(\frac{D^2}{8h} \right)^2 + x^2} \right]_{\substack{D/2 \\ 0}} \cong 1383 \text{ [m]} \quad (72.77)$$

4.3. CÂBLE TRÈS TENDU

Rappelons que dans le cas général nous avons obtenus la relation suivante:

$$\vec{T} + \vec{T}_0 + \mu L \vec{g} = \vec{0} \quad (72.78)$$

Prenons-en la différentielle:

$$d\vec{T} + \underbrace{d\vec{T}_0}_{=\vec{0}} + \mu \vec{g} dL = \vec{0} \quad (72.79)$$

Soit pour la composante y :

$$dT_y + \mu g dL = 0 \quad (72.80)$$

Mais nous avons aussi:

$$\frac{T_y}{T_x} = \frac{dy}{dx} = \tan \alpha \quad (72.81)$$

Donc injecté dans la relation précédente cela donne:

$$T_x d\left(\frac{dy}{dx}\right) + \mu g dL = 0 \quad (72.82)$$

Soit:

$$T_x d\left(\frac{dy}{dx}\right) = -\mu g dL \quad (72.83)$$

Et sous l'hypothèse que le câble est très tendu:

$$dL \cong dx \quad (72.84)$$

Il vient alors:

$$T_x d\left(\frac{dy}{dx}\right) = -\mu g dx \quad (72.85)$$

Après une première intégration il vient:

$$T_x \frac{dy}{dx} = -\mu g x + c^{te} \quad (72.86)$$

Avant d'aller plus loin remarquons que la constante est facile à déterminer puisque en $x = L/2$, la dérivée doit être nulle. Nous avons alors:

$$T_x \frac{dy}{dx} = -\mu g x + \mu g \frac{L}{2} \quad (72.87)$$

Après une deuxième intégration (où la constante d'intégration est nulle):

$$T_x y = -\frac{\mu g}{2} (xL - x^2) \quad (72.88)$$

Nous avons alors pour la déformation de la corde:

$$y = -\frac{\mu g}{T_x 2} (xL - x^2) \quad (72.89)$$

Ainsi, au milieu de la corde la flèche est alors de:

$$f = -\frac{\mu g}{T_x 2} \left(\frac{L}{2} L - \frac{L^2}{4} \right) = -\frac{\mu g L^2}{8 T_x} \quad (72.90)$$

Ou respectivement si on connaît la flèche et qu'on cherche à déterminer la tension:

$$T_x = -\frac{\mu g L^2}{8 f} \quad (72.91)$$

et nous pourrions aussi calculer la longueur de la corde en utilisant la même technique que pour le pont suspendu.

5. BARRAGES

Considérons le barrage de hauteur z , de longueur L et stockant de l'eau de densité ρ ci-dessous:

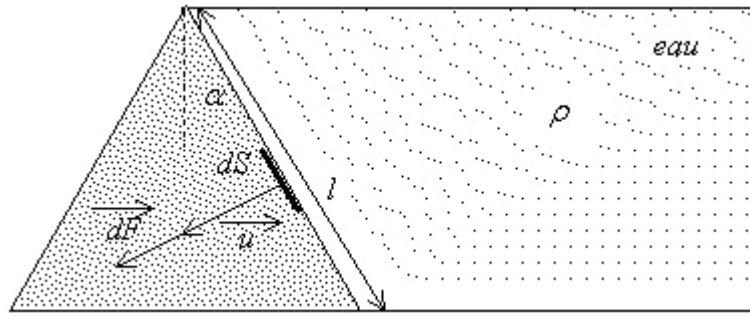


Figure: 72.21 - Approche simplifiée pour l'étude de pression sur un barrage

Nous avons vu dans le chapitre de Mécanique des Milieux Continus que la pression hydrostatique était donnée par:

$$P = P_0 + h \cdot \rho \cdot g \quad (72.92)$$

mais dans cette situation, nous avons évidemment:

$$P = P_0 + (h - z) \cdot \rho \cdot g \quad (72.93)$$

Ainsi lorsque nous nous plaçons à la surface de l'eau en $z = h$:

$$P = P_0 \quad (72.94)$$

soit la pression de l'air à la surface du lac de barrage.

Sur un élément de surface d'aire dS il s'exerce une force élémentaire:

$$\overrightarrow{dF} = P \cdot dS \cdot \vec{u} = P \cdot L dl \cdot \vec{u} \quad (72.95)$$

Or:

$$\cos \alpha = \frac{z}{l} \Rightarrow dl = \frac{dz}{\cos \alpha} \quad (72.96)$$

Ainsi:

$$\overrightarrow{dF} = P \cdot L dl \cdot \vec{u} = (P_0 + (h - z) \cdot \rho \cdot g) \cdot L \frac{dz}{\cos \alpha} \cdot \vec{u} \quad (72.97)$$

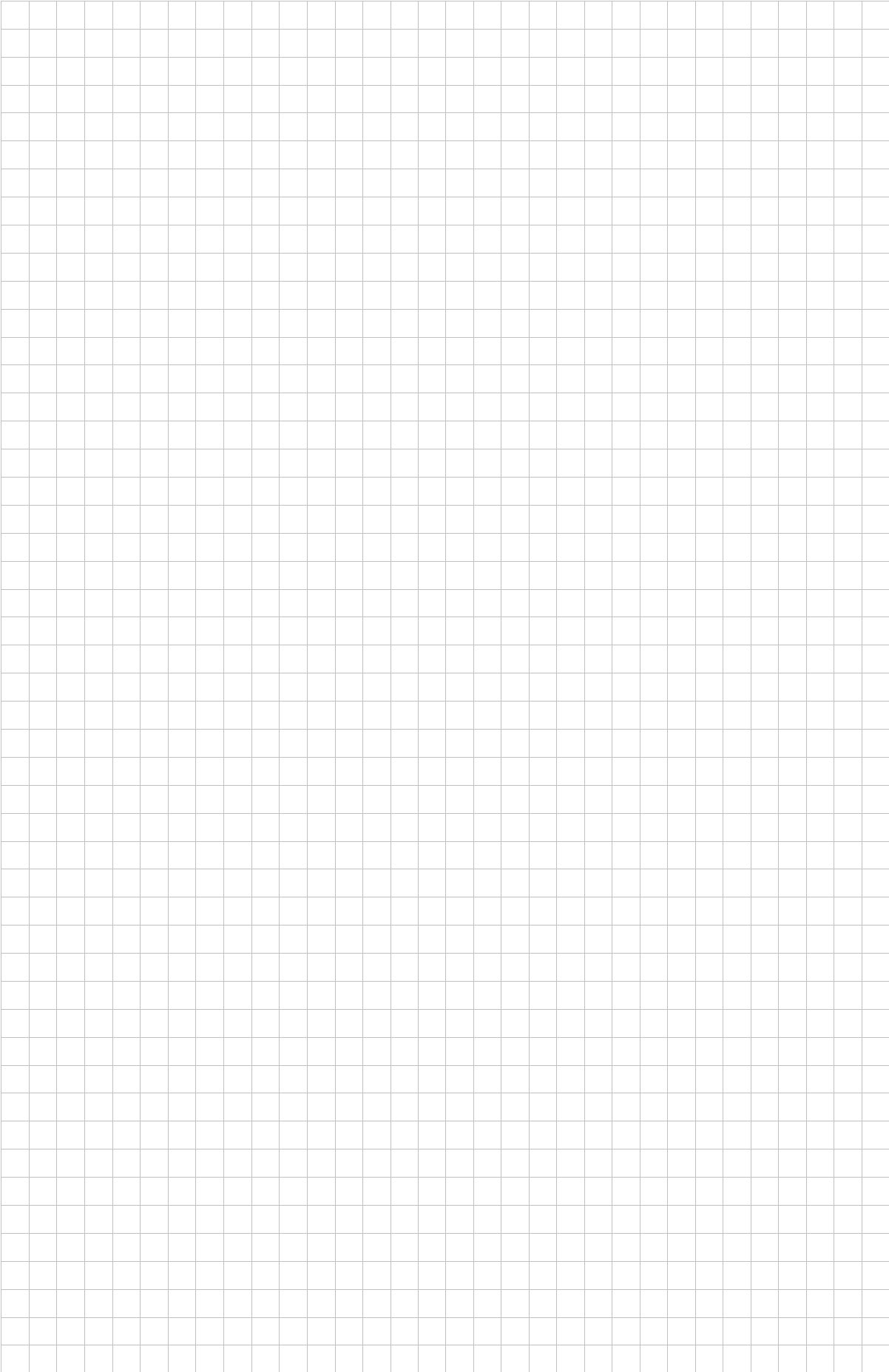
d'où après intégration

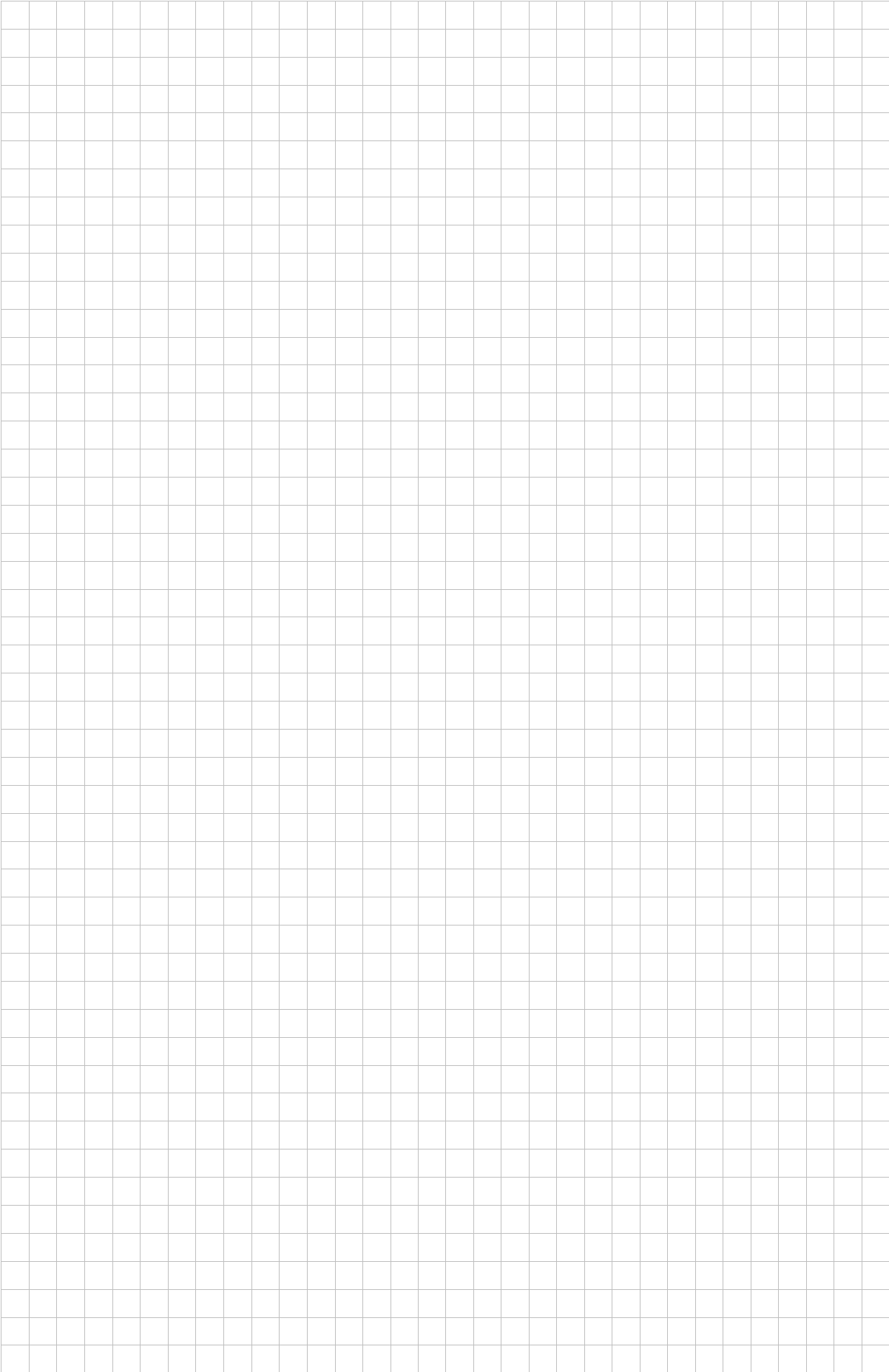
$$F = P \cdot L dl = \int_0^h (P_0 + (h - z) \cdot \rho \cdot g) \cdot \frac{L}{\cos \alpha} dz = \frac{Lh}{\cos \alpha} \left(P_0 + \frac{1}{2} \rho g h \right) \quad (72.98)$$

Il s'agit donc de la force exercée sur la face immergée. La force sur la face émergée (à gauche sur l'illustration) est simplement donnée en posant $\rho = 0$. Nous avons donc:

$$F = \frac{Lh}{\cos \alpha} P_0 \quad (72.99)$$

Remarque: En moyenne à vide et à plein un barrage se déplacerait de 80 cm.







73. GÉNIE AÉROSPATIAL

Dans ce chapitre, nous allons voir quelques cas pratiques mathématiques utiles et simples de ce que nous avons étudié dans les chapitres de Géométrie Analytique, Mécanique Classique, d'Astronomie dans le cadre du Génie Aérospatial qui est donc la discipline scientifique qui rassemble les techniques de l'aéronautique de l'ingénieur (déplacement dans l'atmosphère, utilisant des avions ou des hélicoptère par exemple) et celles de l'astronautique (déplacements spatiaux, c'est-à-dire trajets hors atmosphère et interplanétaires, en utilisant des navettes spatiales ainsi que des fusées).

Le lecteur découvrira certainement qu'il s'agit au fait que d'exemples que nous trouvons souvent dans les livres scolaires en tant qu'exercices

Nous ne reviendrons pas ici sur la théorie des cônes (très importante pour la mise en orbite de satellites) vue dans le chapitre de Géométrie Analytique, ni la théorie du Gyroscope très utile pour orienter/forcer l'axe de rotation (spin) des satellites et déjà vue dans le chapitre de Mécanique Classique (par contre nous ne pourrions pas utiliser la théorie balistique se trouvant dans ce même chapitre puisque nous y avons supposé la vitesse initiale constante...) ou la théorie des points de Lagrange utile parfois pour mettre en orbite des satellites sur des astres lointains (ce qui ne veut pas dire que nous n'utiliserons pas les résultats théoriques des sujets que nous avons étudié). De plus par hypothèse nous considérerons les corps dans un mouvement non relativiste ce qui est à ce jour le cas le plus fréquent...

Dans ce chapitre nous négligerons en réalité pas mal de choses comme les frottements, les vibrations les cas à plus de deux corps (astres...) et autres nombreux facteurs. Nous ne verrons pas non plus les bricolages propres à l'ingénierie spatiale (comme le fait que certains satellites ont des poids attachés par des câbles réglables pour augmenter ou diminuer leur moment cinétique gyroscopique).

Remarque: Certains détails vous sembleront négligeables mais il faut savoir que le prix au kilo de l'ancement se situe entre 20'000 et 30'000\$ sans compter les assurances... (du moins en ce début de 20ème siècle). Donc tout ce qui peut être optimisé sans accroître le risque doit pouvoir l'être!

Avant de commencer sur les sujets purement mathématiques, abordons quelques notions purement pragmatiques.

Le fait de placer le décollage proche de l'équateur sera d'une aide non négligeable puisque la vitesse tangentielle (horizontale) de la rotation de la Terre est donnée par (application de la cinématique circulaire triviale étudiée dans le chapitre de Mécanique Classique):

$$v = \omega \cdot R \cong 465 \text{ [m} \cdot \text{s}^{-1}] \quad (73.1)$$

Soit un peu plus de la vitesse du son au sol dans des conditions normales de température et de pression et qui représente environ 5% de la première vitesse cosmique que nous traiterons juste un peu après (c'est déjà ça de pris pour économiser du carburant!). Et puis évidemment, comme la Terre tourne sur elle-même d'Ouest en Est il vaut mieux faire partir la fusée vers l'Est plutôt que tout droit ou même pire encore... vers l'Ouest (signalons au passage que pour des raisons de sécurité des agents au sol ce serait encore mieux si elle part à l'Est au-dessus d'un océan).

Il faut aussi noter que nous cherchons bien sûr à acquérir de la vitesse verticale pour quitter rapidement les couches denses de l'atmosphère mais nous voulons aussi communiquer au lanceur une composante horizontale de la vitesse: c'est-à-dire celle-ci qui permet la mise en orbite des satellites (d'où une raison supplémentaire pour tirer vers l'Est).

Quant à la différence de l'accélération gravitationnelle elle est un facteur aussi à prendre en compte. Effectivement certains sites spécialisés donnent:

$$\begin{aligned} g_{\text{équateur}} &= 9.78039 \left[m \cdot s^{-2} \right] \\ g_{\text{pôle nord}} &= 9.83217 \left[m \cdot s^{-2} \right] \end{aligned} \quad (73.2)$$

Ainsi, une différence d'environ 0.527% ce qui sur le poids d'une navette spatiale n'est pas négligeable pour une fois encore économiser du carburant.

1. VITESSE COSMOLOGIQUES

La "**première vitesse cosmique**" ou "**vitesse de satellisation minimale**" représente la vitesse minimale à laquelle doit être portée un corps minimale pour se trouver en orbite basse autour de la. Elle est déterminée par la relation qui équilibre force centrifuge et force centripète (cf. chapitre de Mécanique Classique):

$$\frac{v^2}{R} = \frac{GM}{R^2} \quad (73.3)$$

d'où nous déduisons trivialement:

$$v = \sqrt{\frac{GM}{R}} \quad (73.4)$$

En réalité dans nos temps actuels (...) une orbite échappe à l'usure de l'atmosphère terrestre que si elle est à une altitude supérieure à 200 kilomètres. Nous avons alors (vitesse qui correspondrait à peu près à celle de l'ISS d'après plusieurs recoupements):

$$v = \sqrt{\frac{GM}{R}} \cong 7'900 \left[m \cdot s^{-1} \right] \cong 28'500 \left[km \cdot h^{-1} \right] \quad (73.5)$$

La "**deuxième vitesse cosmique**" correspond à la vitesse de libération d'un corps quittant la Terre. C'est la vitesse minimale au-delà de laquelle un corps peut s'éloigner définitivement de la Terre, en tout cas tant que l'on néglige la présence du Soleil et de notre Galaxie. Nous avons déjà démontré la relation y relative dans le chapitre de Mécanique classique donc inutile d'y revenir. Rappelons jusque que la relation obtenue était:

$$v = \sqrt{\frac{2GM}{R}} \quad (73.6)$$

Ce qui donnait pour la Terre (en prenant le rayon moyen au niveau de la mer...):

$$v = \sqrt{\frac{2Gm}{R}} \cong 11.2 \cdot 10^3 \left[m \cdot s^{-1} \right] \cong 40'000 \left[km \cdot h^{-1} \right] \quad (73.7)$$

Du fait de l'atmosphère terrestre, il est difficile (et peu utile) d'amener un objet proche de sa surface à cette vitesse, celle-ci étant située trop avant dans le régime hypersonique pour être réalisable par la plupart des systèmes de propulsion. En outre, elle provoquerait une destruction de la plupart des objets par friction ou compression atmosphérique. En pratique, un objet serait tout d'abord placé en orbite terrestre circulaire basse puis accéléré à partir de cette altitude puisque les frottements y sont quasiment nuls, la poussée a alors un très bon rendement. Par ailleurs, la table ci-dessous fournies par la NASA atteste de cette stratégie (vous pouvez remarquer qu'à partir de 105 kilomètres d'altitude le gain de vitesse est très efficace):

Temps [s]	Altitude [m]	Vitesse [m/s]	Accélération [m/s ²]
0	-8	0	2.45
20	1'244	139	18.62
40	5'377	298	16.37
60	11'617	433	19.40
80	19'872	685	24.50
100	31'412	1'026	24.01
120	44'726	1'279	8.72
140	57'396	1'373	9.70
160	67'893	1'490	10.19
180	77'485	1'634	10.68
200	85'662	1'800	11.17
220	92'481	1'986	11.86
240	98'004	2'191	12.45
260	10'301	2'417	13.23
280	105'321	2'651	13.92
300	107'449	2'915	14.90
320	108'619	3'203	15.97
340	108'942	3'516	17.15
360	108'543	3'860	18.62
380	107'690	4'216	20.29
400	106'539	4'630	22.34
420	105'142	5'092	24.89
440	103'775	5'612	28.03
460	102'807	6'184	29.01
480	102'552	6'760	29.30
500	103'297	7'327	29.01
520	105'069	7'581	0.10

Tableau: 73.1 - Données ascensionnelles de la navette Discovery (source: NASA)

2. ÉQUATION FONDAMENTALE DE LA PROPULSION

Un lanceur spatial a pour mission de placer une charge en orbite, pour cela il doit fonctionner dans l'atmosphère et le vide. Les principes utilisés sont ceux de l'action et de la réaction de Newton, et de la conservation de la quantité de mouvement: grossièrement nous pouvons affirmer qu'une fusée accélère en éjectant des gaz à grande vitesse (le ballon baudruche qui se dégonfle donne une bonne idée du phénomène). Après une étude mécanique simple, nous pouvons obtenir l'équation fondamentale de la propulsion.

En considérant les différentes parties mobiles et indépendantes de la fusée (en fait la structure principale séparément des gaz éjectés), nous pouvons affirmer à l'aide du principe d'action et réaction (cf. chapitre de Mécanique classique) pour un système donné, la somme des forces extérieures est:

$$\sum \vec{F}_{ext} = m\vec{a} = \frac{d\vec{p}}{dt} \quad (73.8)$$

Ainsi, en prenant le système fusée et gaz ensembles, nous avons:

$$\sum \frac{d\vec{p}_i}{dt} = \vec{0} \quad (73.9)$$

Soit que:

$$\sum \vec{p}_i = c^{te} \quad (73.10)$$

Le principe de la propulsion s'énonce alors grâce à la conservation de la quantité de mouvement, que nous avons établie.

Considérons une fusée de masse m et de vitesse v_f , la vitesse d'éjection des gaz étant v_g . Au temps t , nous avons:

$$p = mv_f \quad (73.11)$$

au temps $t+dt$ nous avons:

$$p = (m - dm)(v_f + dv_f) + dm(v_f + dv_f - v_g) \quad (73.12)$$

mais le dm est une perte de masse donc il faut changer son signe sinon la relation précédente ne correspond pas à l'interprétation de la réalité. Ainsi:

$$p = (m + dm)(v_f + dv_f) - dm(v_f + dv_f - v_g)$$

d'après le principe de conservation de la quantité de mouvement:

$$mv_f = (m + dm)(v_f + dv_f) - dm(v_f + dv_f - v_g) \quad (73.1)$$

d'où après simplification:

$$mdv_f = -dmv_g \quad (73.2)$$

d'où:

$$dv_f = -v_g \frac{dm}{m} \quad (73.3)$$

ce qui par intégration donne "**l'équation fondamentale de la propulsion**" ou le "**delta-v**" comme on dit souvent dans la littérature américaine et ce pour une fusée (non-relativiste...) hors champ de gravité et dans le vide à vitesse d'éjection constante des gaz...:

$$\Delta v_f = v_g \ln \left(\frac{m_{initiale}}{m_{finale}} \right) \quad (73.4)$$

La différence entre masse initiale et masse finale étant souvent appelée "**poids mort**" dans le domaine.

Nous comprenons alors aisément pourquoi les fusées sont composées de plusieurs éléments de propulsion dont elles se séparent. Cela leur permet d'accroître leur vitesse finale en se débarrassant de la masse des réservoirs qu'elles emportent initialement avec elles.

Nous en déduisons donc qu'un lanceur accélère d'autant plus que la vitesse des gaz est grande, que la poussée dépend de la quantité de gaz fournis et de leur vitesse, et le rapport des masses initiale et finale doit être maximum pour favoriser la propulsion, c'est-à-dire que la structure du système est voulue négligeable (la masse finale est alors minimale).

Remarquons que nous avons par extension aussi la relation suivante entre l'accélération de la fusée et le débit de masse éjecté:

$$a_f = \frac{dv_f}{dt} = -\frac{v_g}{m} \frac{dm}{dt} \quad (73.5)$$

Cependant, dans un champ gravitationnel (bien évidemment il faut que l'accélération de la fusée soit supérieur à celle de la gravité à tout moment...) il faut ajouter le terme qui freine la fusée (la "**perte par pesanteur**") et qui donnera donc l'expression:

$$v = v_g \ln \left(\frac{m_{initiale}}{m(t)} \right) - g \cdot t \quad (73.6)$$

si nous supposons la gravité g comme constante pendant la phase d'accélération principale. Cette dernière relation serait parfois appelée "**formule de Tsiolkovski**". En notant par D_e le débit massique des propulseurs, nous retrouvons parfois cette dernière relation sous la forme suivante où le temps n'intervient plus explicitement:

$$v = v_g \ln \left(\frac{m_{initiale}}{m_{initiale} - D_e \cdot t} \right) - g \cdot t = v_g \ln \left(\frac{m_{initiale}}{m(t)} \right) - g \cdot \frac{(m_{initiale} - m(t))}{D_e} \quad (73.7)$$

puisque:

$$t = \frac{m_{initiale} - m(t)}{D_e} \quad (73.8)$$

Cependant, il faudrait aussi prendre en compte la variation de la gravité en fonction de la distance comme nous l'avons démontré dans le chapitre d'Astronomie. Nous avons alors:

$$v = v_g \ln \left(\frac{m_{initiale}}{m(t)} \right) - g_o \left(\frac{R}{r} \right)^2 t \quad (73.9)$$

Donc si nous supposons la vitesse d'éjection des gaz constante et la trajectoire en ligne droite depuis le corps d'attraction nous avons donc que plus le temps passe plus la vitesse de la fusée augmente à cause de la diminution de sa masse (bien évidemment elle finira par être constante) mais en même temps moins l'influence de la gravité est grande.

La valeur théorique de l'altitude atteinte, même si son expression est très simple à déterminer, est tellement fautive que cela ne vaut même pas la peine d'en faire mention.

Les navettes ne vont donc pas à la vitesse de libération avec qu'un seul étage de propulsion et même... elles n'ont souvent comme seul objectif d'aller seulement à la vitesse de mise en orbite basse... (première vitesse cosmique). La navette n'est donc dans la pratique pas libérée de l'attraction terrestre, loin de là!

Exemple:

Calculons la vitesse finale de la première phase de lancement d'Ariane 5 en supposant la gravité constante, en connaissant la vitesse d'échappement des gaz (supposée constante), la masse totale initiale de la fusée, la masse éjectée et le débit massique (supposé aussi constant...). Nous avons alors:

$$v \cong 3'600 \ln \left(\frac{700'000}{260'000} \right) - 9.7 \cdot \frac{(700'000 - 260'000)}{3'900} \cong 2'470 \text{ [m} \cdot \text{s}^{-1}] \quad (73.10)$$

avec:

$$t = \frac{m_{initiale} - m(t)}{D_e} = \frac{700'000 - 260'000}{3'900} \cong 112 \text{ [s]} \quad (73.11)$$

Alors que la valeur réelle est comprise entre 2'000 et 2'800 mètres par seconde d'après le recoupement de plusieurs sites Internet et vidéos du centre de contrôle de lancement d'Ariane (donc nous sommes très loin de la première vitesse cosmique!). Si nous ajoutions la variation de la gravité avec l'altitude le résultat calculé serait encore plus proche des 2'800 mètres par seconde (pour information la séparation des boosters de la première phase se ferait environ à 80 kilomètres d'altitude d'après le recoupement de plusieurs sources environ 132 secondes après l'allumage).

Cela reste cependant un ordre de grandeur convenable (et de plus similaire aux décollages des navettes américaines) compte tenu de la non prise en compte des frottements de l'air (qui pour rappel à ce que nous avons démontré dans le chapitre de Mécanique Des Milieux Continus sont proportionnels au carré de la vitesse dans le cas subsonique pour le moins...!!!).

Déterminons maintenant la distance atteinte après un temps t donné dans le cas de l'approximation à champ de pesanteur constant. Nous avons alors bien évidemment dans un premier temps (nous changeons un peu la notation afin de la condenser)

$$\begin{aligned} v(t) = \frac{dz(t)}{dt} &= v_g \ln \left(\frac{m_0}{m_0 - D_e t} \right) - g \cdot t = v_g \ln \left(\frac{1}{1 - \frac{D_e}{m_0} t} \right) - g \cdot t = v_g \ln \left(\frac{1}{1 - \frac{D_e}{m_0} t} \right) - g \cdot t \\ &= -v_g \ln \left(1 - \frac{D_e}{m_0} t \right) - g \cdot t \end{aligned} \quad (73.12)$$

Soit:

$$z(t) = -v_g \int \ln \left(1 - \frac{D_e}{m_0} t \right) dt - g \int t dt \quad (73.13)$$

Nous allons faire le changement de variable suivant:

$$u = 1 - \frac{D_e}{m_0} t \quad (73.14)$$

Soit:

$$du = -\frac{D_e}{m_0} dt \quad (73.15)$$

Ce qui nous donne (cf. chapitre de Calcul Différentiel Et Intégral):

$$\begin{aligned} z(t) &= -v_g \int \frac{m_0}{D_e} \ln(u) du - g \int t dt = \frac{v_g m_0}{D_e} \int \ln(u) du - g \int t dt \\ &= \frac{v_g m_0}{D_e} (u \ln(u) - u) - \frac{1}{2} g t^2 + c^{\text{te}} \\ &= \frac{v_g m_0}{D_e} \left(\left(1 - \frac{D_e}{m_0} t \right) \ln \left(1 - \frac{D_e}{m_0} t \right) - \left(1 - \frac{D_e}{m_0} t \right) \right) - \frac{1}{2} g t^2 + c^{\text{te}} \\ &= \frac{v_g m_0}{D_e} \left(1 - \frac{D_e}{m_0} t \right) \left[\ln \left(1 - \frac{D_e}{m_0} t \right) - 1 \right] - \frac{1}{2} g t^2 + c^{\text{te}} \end{aligned} \quad (73.16)$$

Déterminons la constante d'intégration avec le fait qu'au temps $t = 0$, nous devons avoir $z(t)$ qui est nul. Il vient alors immédiatement que:

$$c^{\text{te}} = \frac{v_g m_0}{D_e} \quad (73.17)$$

Soit au final:

$$z(t) = \frac{v_g m_0}{D_e} \left(1 - \frac{D_e}{m_0} t \right) \left[\ln \left(1 - \frac{D_e}{m_0} t \right) - 1 \right] - \frac{1}{2} g t^2 + \frac{v_g m_0}{D_e} \quad (73.18)$$

où nous pouvons à nouveau nous débarrasser de la variable explicite du temps en réutilisant le fait que:

$$t = \frac{m_{initiale} - m(t)}{D_e} \quad (73.19)$$

mais dont nous changeons la notation sous la forme plus condensée et simple suivante:

$$t = \frac{m_0 - m(t)}{D_e} \quad (73.20)$$

Ce qui nous donne:

$$z(t) = \frac{v_g m(t)}{D_e} \left[\ln \left(\frac{m(t)}{m_0} \right) - 1 \right] - \frac{1}{2} g \left(\frac{m_0 - m(t)}{D_e} \right)^2 + \frac{v_g m_0}{D_e} \quad (73.21)$$

Exemple:

Calculons la hauteur de la première phase de lancement d'Ariane 5 en supposant la gravité constante avec les mêmes données numériques que l'exemple précédent. Cela donne alors:

$$\begin{aligned} z(t) &\cong \frac{3'600 \cdot 260'000}{3'900} \left[\ln \left(\frac{260'000}{700'000} \right) - 1 \right] - \frac{1}{2} 9.7 \left(\frac{700'000 - 260'000}{3'900} \right)^2 + \frac{3'600 \cdot 260'000}{3'900} \\ &\cong 106'725 \cong 100 \text{ [km]} \end{aligned} \quad (73.22)$$

Ce qui d'après le recoupement de plusieurs site Internet et vidéos du centre de contrôle de lancement d'Ariane ne serait pas trop faux car les propulseurs principaux (EAP: étages accélérateurs à poudre) seraient largués à environ entre 70 et 125 kilomètres d'altitude (entre 132 secondes et 205 secondes après leur allumage) et même la coiffe qui protège les satellites ainsi que l'étage à propulsion (EPC: étage principal cryotechnique) sont parfois largués juste quelques secondes après (cela fait toujours quelques tonnes de moins!).

Donc comme nous l'avons vu avec les calculs de l'exemple précédent, il reste encore environ 100 kilomètres avant d'atteindre l'orbite basse et sa vitesse cosmique correspondante à l'aide de l'étage à propergols stockables (EPS).

Remarquons que si nous faisons décoller la fusée avec une accélération égale à celle de la pesanteur (en supposant de plus cette dernière constante), nous obtiendrions une hauteur égale à:

$$z(t) = \frac{1}{2} g t^2 = \frac{1}{2} \cdot 9.7 \cdot (112)^2 \cong 60'838 \text{ [m]} \cong 60 \text{ [km]} \quad (73.23)$$

Donc nous pouvons approximativement en conclure que la poussée d'Ariane 5 est significativement supérieur à celle opposée de la pesanteur.

Nous pouvons aussi calculer la variation de l'intensité de la pesanteur à l'altitude de 100 kilomètres pour voir si sa variation est significative ou non en % de celle au sol:

$$\frac{\Delta g}{g_0} = \left(\frac{R}{R+z} \right)^2 \cong \left(\frac{6'378'137}{6'378'137 + 100'000} \right)^2 \cong 97\% \quad (73.24)$$

donc nous voyons que cela fait 3% ce qui est non négligeable.

Par ailleurs, voici un graphique soit disant tiré d'un ouvrage d'Arianespace montrant l'accélération horizontale du lanceur Ariane 5 en multiple de la pesanteur locale au pas de tir:

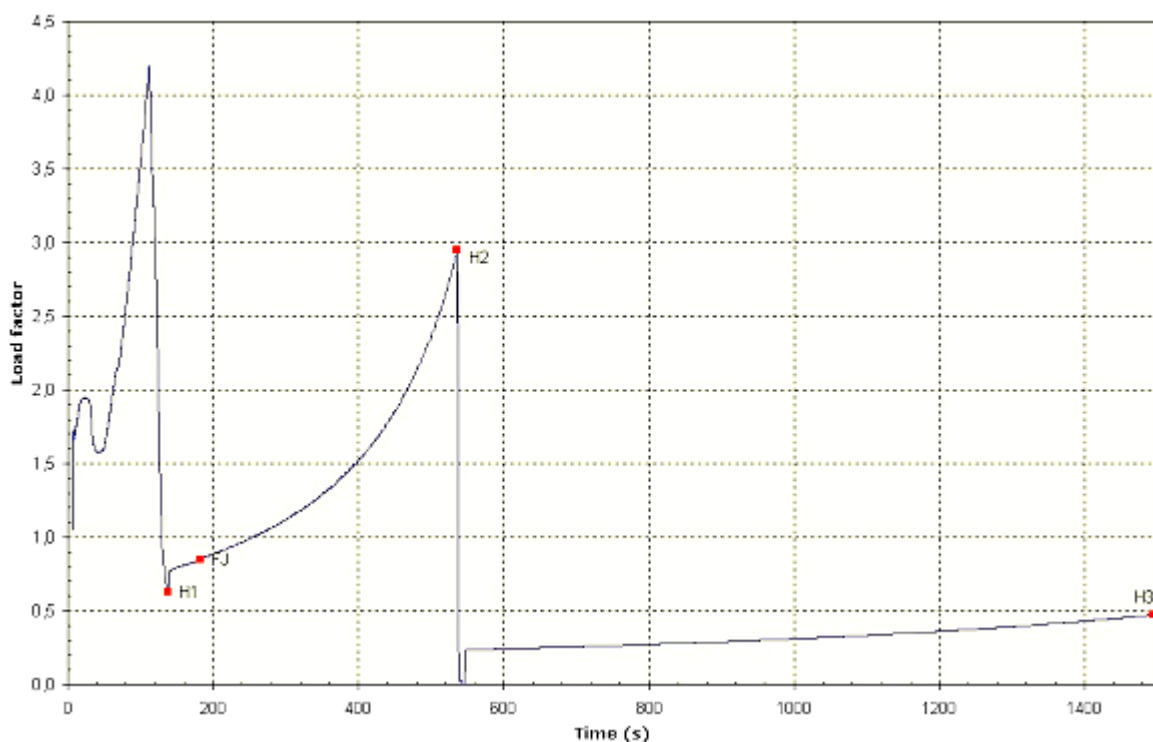


Figure: 73.1 - Profil de l'accélération horizontale d'Ariane 5 (source: ?)

Supposons maintenant qu'en se basant sur la relation:

$$t = \frac{m_0 - m(t)}{D_e} \quad (73.25)$$

nous savons la masse totale éjectable limitée. Nous aurons alors aussi le temps de combustion qui sera limité. Notons cela:

$$t_f = \frac{m_0 - m_f}{D_e} \quad (73.26)$$

En utilisant les relations obtenus plus haut dans le cas à champ gravitationnel constant, nous avons alors bien évidemment:

$$\begin{aligned}
 v_f &= v_g \ln \left(\frac{m_0}{m_f} \right) - g \frac{(m_0 - m_f)}{D_e} \\
 z_f &= \frac{v_g m_f}{D_e} \left[\ln \left(\frac{m_f}{m_0} \right) - 1 \right] - \frac{1}{2} g \left(\frac{m_0 - m_f}{D_e} \right)^2 + \frac{v_g m_0}{D_e}
 \end{aligned}
 \quad (73.27)$$

Après avoir atteint la vitesse finale, la vitesse de la fusée sera donnée par la relation classique de la cinématique rectiligne (cf. chapitre de Mécanique Classique):

$$v = v_f - g(t - t_f) \quad (73.28)$$

et la hauteur aussi:

$$z = z_f + v_f(t - t_f) - \frac{1}{2} g(t - t_f)^2 \quad (73.29)$$

En y injectant les relations obtenues plus haut, nous avons d'abord pour la vitesse:

$$\begin{aligned}
 v &= v_f - g(t - t_f) = \left[v_g \ln \left(\frac{m_0}{m_f} \right) - g \frac{(m_0 - m_f)}{D_e} \right] - g \left(t - \frac{m_0 - m_f}{D_e} \right) \\
 &= v_g \ln \left(\frac{m_0}{m_f} \right) - gt
 \end{aligned}
 \quad (73.30)$$

et pour la hauteur:

$$\begin{aligned}
z &= z_f + v_f(t - t_f) - \frac{1}{2}g(t - t_f)^2 \\
&= \left[\frac{v_g m_f}{D_e} \left[\ln\left(\frac{m_f}{m_0}\right) - 1 \right] - \frac{1}{2}gt_f^2 + \frac{v_g m_0}{D_e} \right] + \left[v_g \ln\left(\frac{m_0}{m_f}\right) - gt_f \right] (t - t_f) \\
&\quad - \frac{1}{2}g(t - t_f)^2 \\
&= \frac{v_g m_f}{D_e} \left[\ln\left(\frac{m_f}{m_0}\right) - 1 \right] - \frac{1}{2}gt_f^2 + \frac{v_g m_0}{D_e} + v_g \ln\left(\frac{m_0}{m_f}\right)t - gt_f t - v_g t_f \ln\left(\frac{m_0}{m_f}\right) + gt_f^2 \\
&\quad - \frac{1}{2}g(t^2 - 2tt_f + t_f^2) \\
&= \frac{v_g m_f}{D_e} \left[\ln\left(\frac{m_f}{m_0}\right) - 1 \right] + \frac{v_g m_0}{D_e} + v_g t \ln\left(\frac{m_0}{m_f}\right) - v_g \left(\frac{m_0 - m_f}{D_e} \right) \ln\left(\frac{m_0}{m_f}\right) - \frac{1}{2}gt^2 \\
&= \frac{v_g m_f}{D_e} \left[\ln\left(\frac{m_f}{m_0}\right) - 1 \right] + \frac{v_g m_0}{D_e} + v_g t \ln\left(\frac{m_0}{m_f}\right) + \frac{v_g m_f}{D_e} \ln\left(\frac{m_0}{m_f}\right) - \frac{v_g m_0}{D_e} \ln\left(\frac{m_0}{m_f}\right) - \frac{1}{2}gt^2 \\
&= \frac{v_g m_f}{D_e} \left[\ln\left(\frac{m_f}{m_0}\right) + \ln\left(\frac{m_0}{m_f}\right) - 1 \right] + \frac{v_g m_0}{D_e} + v_g t \ln\left(\frac{m_0}{m_f}\right) - \frac{v_g m_0}{D_e} \ln\left(\frac{m_0}{m_f}\right) - \frac{1}{2}gt^2 \\
&= -\frac{v_g m_f}{D_e} + \frac{v_g m_0}{D_e} + v_g t \ln\left(\frac{m_0}{m_f}\right) - \frac{v_g m_0}{D_e} \ln\left(\frac{m_0}{m_f}\right) - \frac{1}{2}gt^2 \\
&= v_g \frac{m_0 - m_f}{D_e} + v_g t \ln\left(\frac{m_0}{m_f}\right) - \frac{v_g m_0}{D_e} \ln\left(\frac{m_0}{m_f}\right) - \frac{1}{2}gt^2
\end{aligned} \tag{73.31}$$

Mais lorsque la hauteur maximale est atteinte, la vitesse est nulle, nous avons alors:

$$v = v_g \ln\left(\frac{m_0}{m_f}\right) - gt \xrightarrow{v=0} t = \frac{v_g}{g} \ln\left(\frac{m_0}{m_f}\right) \tag{73.32}$$

Soit injecté dans la relation précédente cela donne:

$$\begin{aligned}
z &= v_g \frac{m_0 - m_f}{D_e} + v_g \left(\frac{v_g}{g} \ln\left(\frac{m_0}{m_f}\right) \right) \ln\left(\frac{m_0}{m_f}\right) - \frac{v_g m_0}{D_e} \ln\left(\frac{m_0}{m_f}\right) - \frac{1}{2}g \left(\frac{v_g}{g} \ln\left(\frac{m_0}{m_f}\right) \right)^2 \\
&= v_g \frac{m_0 - m_f}{D_e} + \frac{1}{g} \left(v_g \ln\left(\frac{m_0}{m_f}\right) \right)^2 - \frac{v_g m_0}{D_e} \ln\left(\frac{m_0}{m_f}\right) - \frac{1}{2g} \left(v_g \ln\left(\frac{m_0}{m_f}\right) \right)^2 \\
&= v_g \frac{m_0 - m_f}{D_e} + \frac{1}{2g} \left(v_g \ln\left(\frac{m_0}{m_f}\right) \right)^2 - \frac{v_g m_0}{D_e} \ln\left(\frac{m_0}{m_f}\right)
\end{aligned} \tag{73.33}$$

Les deux premiers termes sont positifs. Les troisième et dernier terme est négatif. Dès lors, nous voyons que pour maximiser la hauteur atteinte, le mieux est de faire tendre le débit massique vers l'infini (soit: donner toute l'impulsion dès le début!). Dès lors, nous avons:

$$z_{\max} \underset{D_e \rightarrow +\infty}{=} \frac{1}{2g} \left(v_g \ln \left(\frac{m_0}{m_f} \right) \right)^2 \quad (73.34)$$

Exemple:

Calculons la hauteur maximale de la première phase de lancement d'Ariane 5 en supposant la gravité constante avec les mêmes données numériques que l'exemple précédent. Cela donne alors:

$$z_{\max} \underset{D_e \rightarrow +\infty}{=} \frac{1}{2 \cdot 9.86} \left(3'600 \cdot \ln \left(\frac{700'000}{260'000} \right) \right)^2 \cong 655'274 \text{ [m]} \cong 650 \text{ [km]} \quad (73.35)$$

Donc il s'agit de l'altitude à partir de laquelle la fusée aurait une vitesse ascensionnelle nulle et commencerait à retomber. Comme nous pouvons le constater, c'est largement au-dessus de l'orbite basse mais aussi très en-dessous de l'orbite géostationnaire comme nous allons le calculer par la suite.

3. ORBITE GÉOSTATIONNAIRE

L'orbite géostationnaire est une orbite située à 35'786 kilomètres d'altitude au-dessus de l'équateur de la Terre, dans le plan équatorial et d'une excentricité orbitale nulle. C'est un cas particulier de l'orbite géosynchrone (dans le cas contraire la période orbitale correspond toujours à la durée de la révolution de la Terre mais l'orbite s'écarte également au Nord et au Sud de l'équateur en décrivant un analemma dans le ciel lorsqu'il est observé depuis un point fixe de la surface de la Terre).

Cette caractéristique est particulièrement importante pour les satellites de télécommunications ou de diffusion de télévision. La position du satellite semblant immobile, un équipement de réception muni d'une antenne fixe pointant dans la direction du satellite géostationnaire suffira pour capter ses émissions. Cette orbite est également utilisée pour l'observation de la Terre depuis une position fixe dans l'espace comme c'est le cas pour les satellites météorologiques géostationnaires.

Les satellites géostationnaires sont donc nécessairement situés à la verticale ou au zénith d'un point de l'équateur ou, en d'autres termes, situés dans le plan équatorial de la Terre.

Pour calculer la position de l'orbite géostationnaire, nous allons d'abord utiliser la seconde loi de Newton (cf. chapitre de Mécanique Classique):

$$F = m \cdot a \quad (73.36)$$

et nous avons démontré dans le chapitre de Mécanique Classique que lorsque le mouvement est circulaire, nous avons:

$$a = \frac{v^2}{R} \quad (73.37)$$

Et nous allons utiliser la loi de la gravitation présentée dans le chapitre de Mécanique Classique (et démontrée dans le chapitre de Relativité Générale):

$$F = \frac{GMm}{R^2} \quad (73.38)$$

Nous allons utiliser ces relations avec la masse de la Terre $M_T = 5.9736 \cdot 10^{24} [kg]$, m_s la masse du satellite, le rayon de la Terre $R_T = 6'378.14 [km]$ à l'équateur en moyenne, h la hauteur du satellite par rapport au sol et v la vitesse du satellite.

Sur l'orbite géostationnaire, il y a donc équilibre entre les forces de gravitation et la force centrifuge du satellite et la force d'attraction gravitationnelle de la planète:

$$\frac{GMm}{R^2} = m \frac{v^2}{R} \quad (73.39)$$

En adoptant les notations citées précédemment cela donne donc:

$$\frac{GM_T m_s}{(R_T + h)^2} = m_s \frac{v^2}{R_T + h} \quad (73.40)$$

Nous voyons que la masse du satellite se simplifie. Donc l'orbite géostationnaire en est indépendante! Il est important aussi de noter que puisque la vitesse v est après simplification indépendante la masse du satellite pour une orbite circulaire, alors un astronaute à l'intérieur de celui-ci aura la même vitesse et sera donc en apesanteur à l'intérieur de celui-ci (il en va de même pour tous les objets proches à l'intérieur ou à l'extérieur du satellite).

La vitesse pour une trajectoire circulaire est le rapport de la circonférence du cercle sur la période de temps nécessaire pour le parcourir en entier. Nous avons donc:

$$v = \frac{2\pi(R_T + h)}{T} \quad (73.41)$$

Donc T étant égal par définition dans ce contexte particule d'orbite géostationnaire à la durée de la journée sur Terre, nous avons en prenant les tables disponibles:

$$T = 23.93419 \cdot 60 \cdot 60 \cong 86163.084 [s] \quad (73.42)$$

Retournons à la relation antérieure pour la simplifier:

$$\frac{GM_T}{R_T + h} = v^2 \quad (73.43)$$

et y injectant la relation explicite de la vitesse:

$$\frac{GM_T}{R_T + h} = \left(\frac{2\pi(R_T + h)}{T} \right)^2 \quad (73.44)$$

Soit:

$$\left(\frac{GM_T T^2}{4\pi^2} \right)^{1/3} = R_T + h \quad (73.45)$$

Il vient enfin:

$$h = \left(\frac{GM_T T^2}{4\pi^2} \right)^{1/3} - R_T \cong 35'784 \text{ [km]} \quad (73.46)$$

et la vitesse du satellite:

$$v = \frac{2\pi(R_T + h)}{T} \cong 3074 \text{ [m} \cdot \text{s}^{-1}] \quad (73.47)$$

Remarquons qu'en utilisant les relations précédents, si le rayon de l'orbite d'un satellite non géostationnaire nous est donné ainsi que la masse de la Terre, nous pouvons alors aussi déterminer sa période de révolution sans avoir à connaître sa vitesse.

La force centrifuge est souvent utilisée dans le calcul de génie spatial. Par exemple pour calculer la vitesse de rotation que devrait avoir une laboratoire circulaire spatial de rayon r pour simuler la gravité terrestre comme ci-dessous:

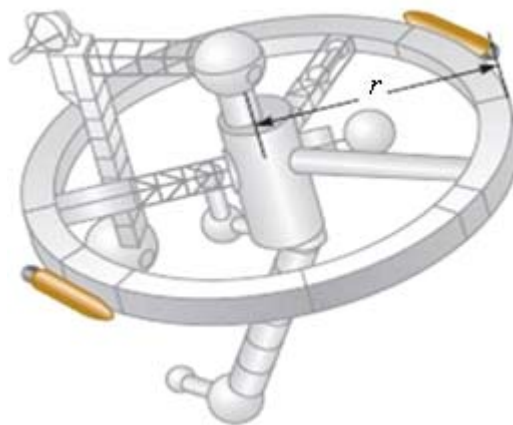


Figure: 73.2 - Station spatiale simulant gravité

nous appliquons simplement à nouveau l'analyse de la force centrifuge:

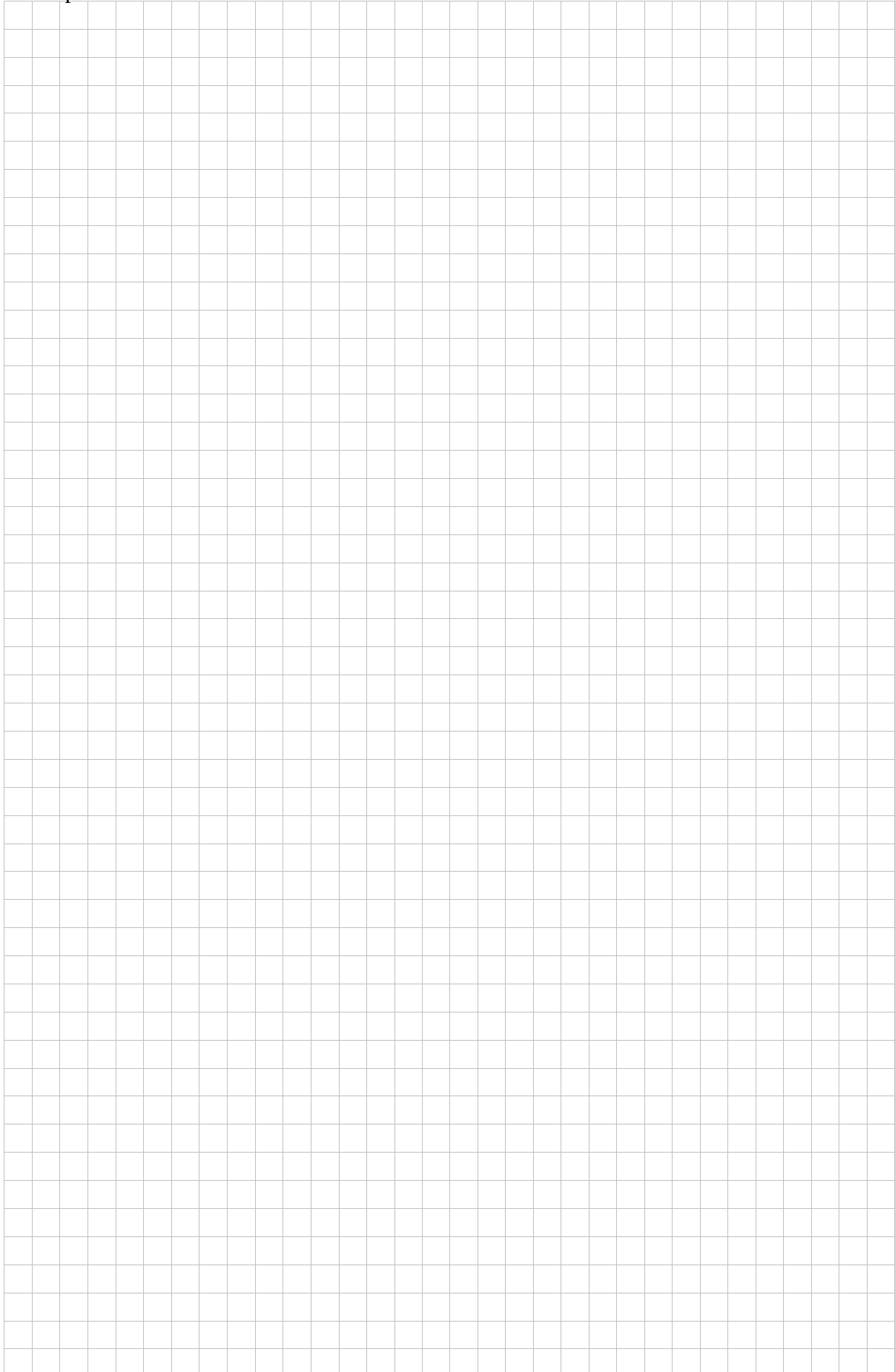
$$mg = m \frac{v^2}{r} \quad (73.48)$$

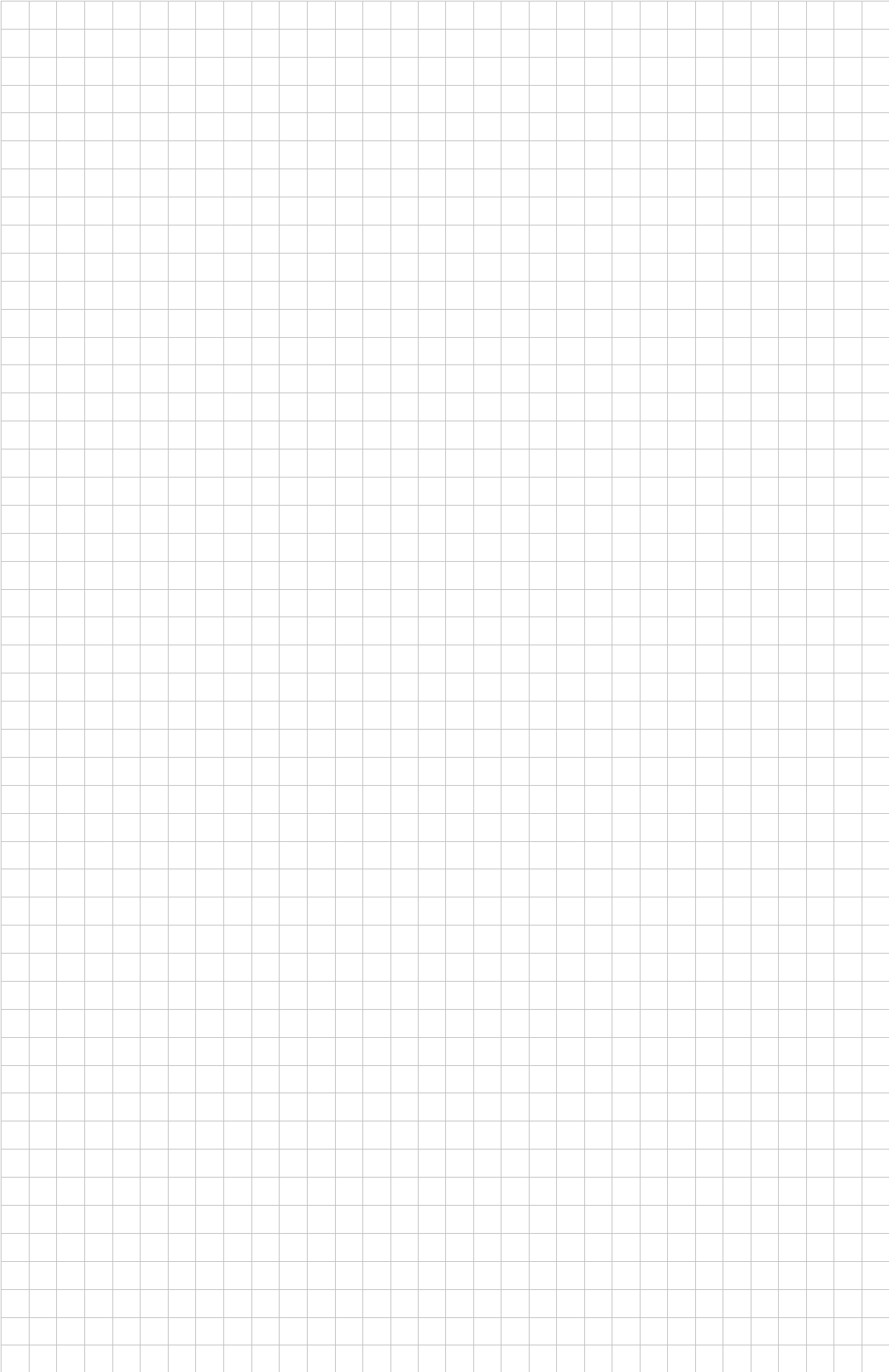
et nous voyons de suite que la masse se simplifie pour obtenir:

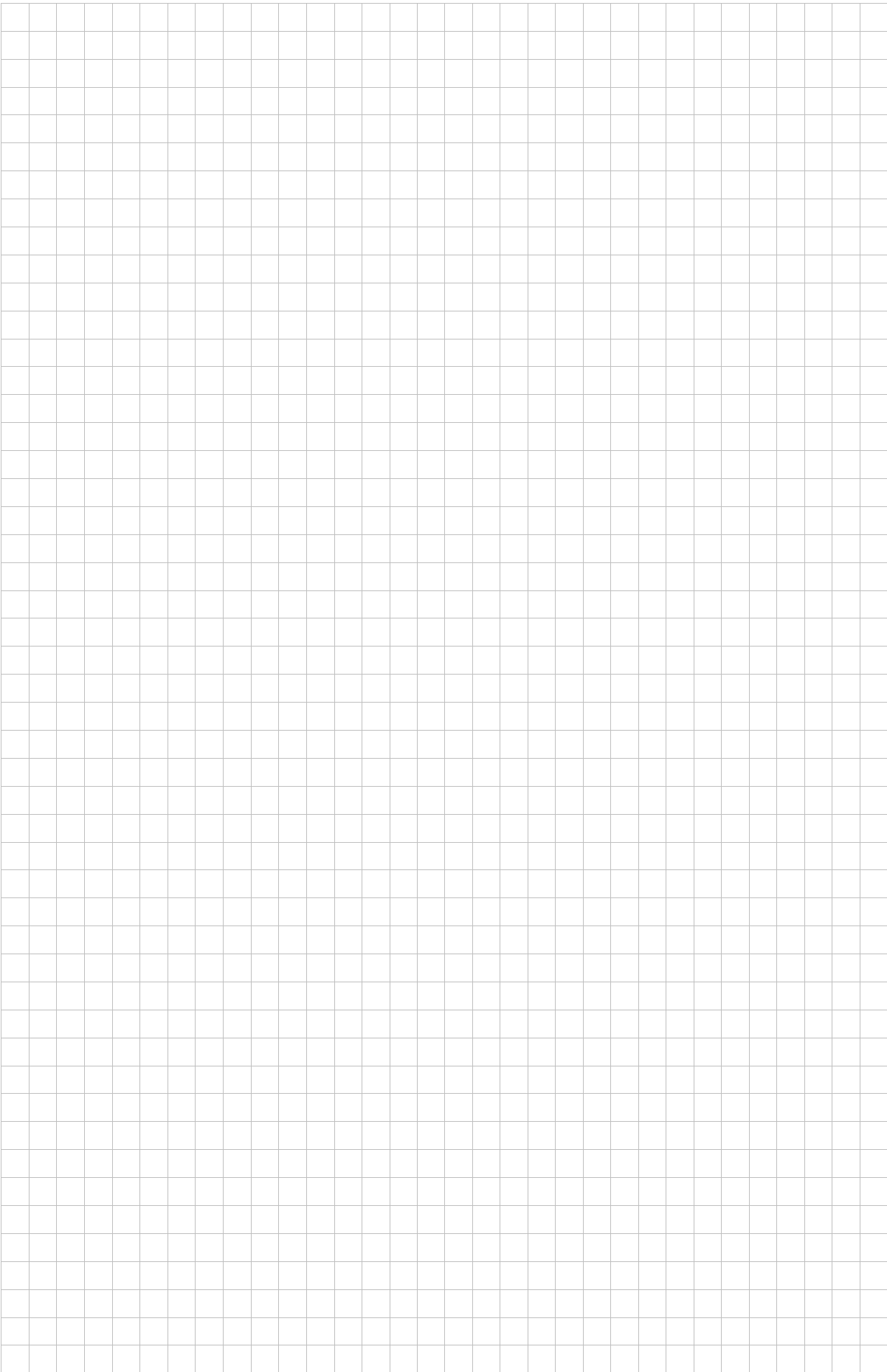
$$v = \sqrt{g \cdot r} \quad (73.49)$$

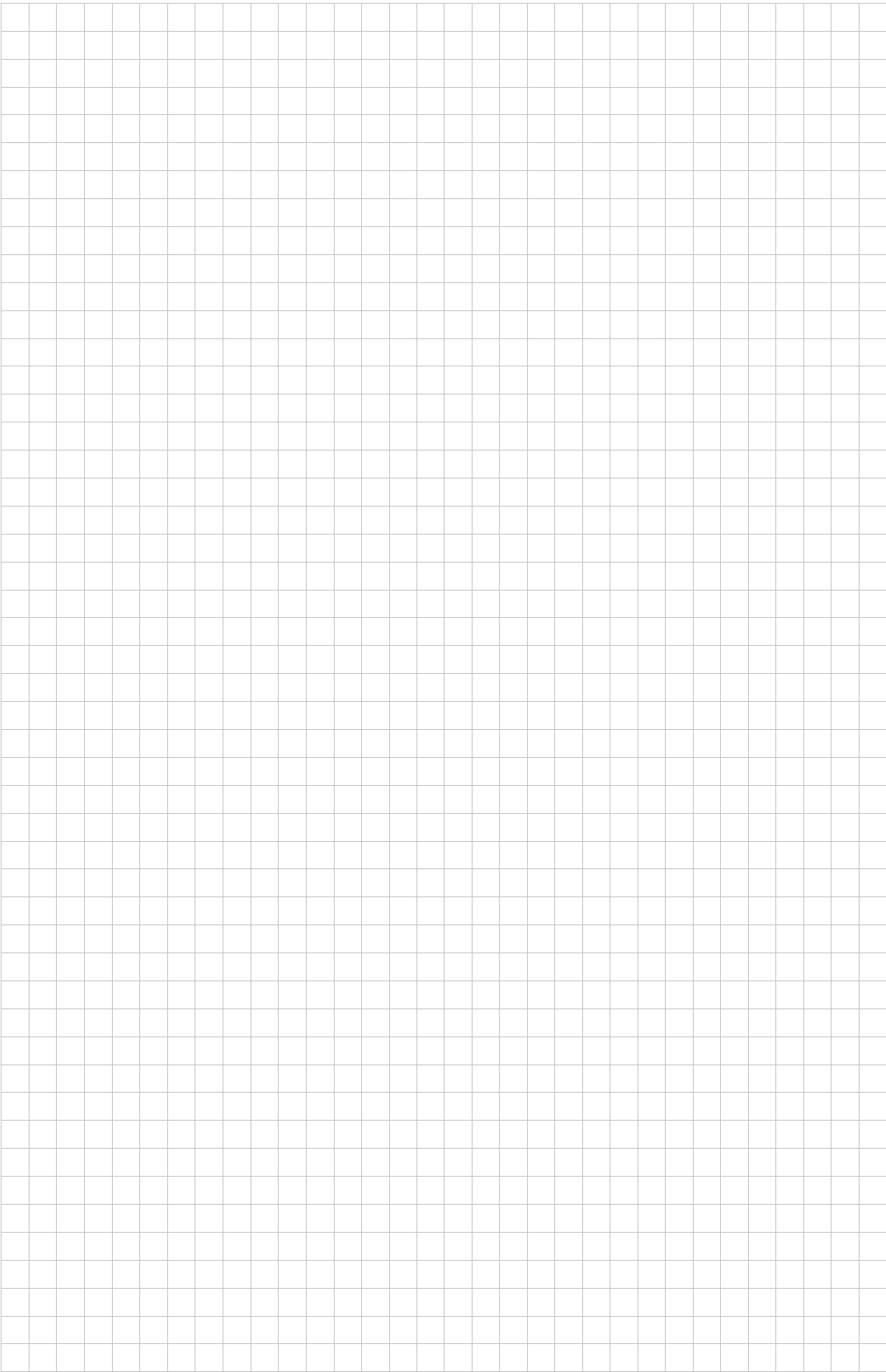
Le reste (trajectoires des sondes spatiales) viendra quand j'aurais du temps à disposition...

Notes personnelles:









74. GÉNIE INDUSTRIEL

Le Génie Industriel englobe la conception, l'amélioration et l'installation de systèmes. Il utilise les connaissances provenant des sciences mathématiques, physiques et sociales, ainsi que les principes et méthodes propres à l'art de l'ingénieur, dans le but de spécifier, prédire et évaluer les résultats découlant de ces systèmes.

Nous pouvons résumer tous les domaines qui touchent au génie industriel (et pas seulement... cela peut s'appliquer avec adaptation ad hoc à l'administration) par l'objectif d'optimiser et contrôler les performances globales de l'entreprise (coûts, délais, qualité) car:

On ne peut améliorer que ce que l'on mesure!

Remarquons que certaines techniques de génie industriel ont déjà été abordées dans d'autres chapitres comme les techniques de gestion quantitatives, l'optimisation (recherche opérationnelle), l'analyse financière, l'analyse des files d'attente, etc. et que ce domaine englobe le "**génie qualité**".

Dans ce chapitre, nous traiterons uniquement des aspects théoriques minimaux du SQC (Statistical Quality Control) relatifs au contrôle statistique de la qualité (dont c'est le métier du "qualiticien") dans le cadre de la fabrication et de la mise en production de biens ou de services et qui constitue le minimum-minimum de la connaissance de tout ingénieur qualité actif dans une organisation quelconque (industrielle ou administrative) sous peine d'avoir aucune crédibilité! Par ailleurs méfiez vous des entreprises - particulièrement des multinationales - qui recherchent des spécialistes qualité maîtrisant Microsoft Excel ou Microsoft Access. Car cela signifiera qu'elles utilisent des outils non professionnels pour faire un travail qui lui devrait pourtant l'être avec des outils adaptés (et Microsoft Excel ou Microsoft Access ne le sont pas)!!! Donc en termes d'organisation interne, vous pouvez vous assurer que ces entreprises organisent et analysent n'importe quoi, n'importe comment, avec un outil non adapté et donc que c'est le bordel général en interne.

Selon l'utilisation, nous distinguons trois domaines principaux qui dans l'ordre conventionnel sont:

1. Contrôle statistique de processus, surveillance de fabrication ou réglage de qualité (Statistical Process Control, SPC). Il s'agit de la surveillance d'un processus de fabrication pendant la production de produits de masse, pour découvrir des différences de qualité et pour pouvoir intervenir et conduire directement.

L'ingénieur doit obligatoirement consulter la norme ISO/TR 13425:2006 *Lignes directrices pour la sélection des méthodes statistiques dans la normalisation et la spécification* ainsi que la norme ISO 8258 *Cartes de contrôle de Shewhart* et enfin ISO/TR 18532 *Lignes directrices pour l'application des méthodes statistiques à la qualité et à la normalisation industrielle* avant de mettre en place des outils SPC au sein de son entreprise.

2. Contrôle de réception ou examen d'échantillon de réception (Acceptance Sampling, AC). Il s'agit du contrôle d'entrée, d'un contrôle pendant la production et d'un contrôle final des produits dans une entreprise (ou usine) sans influence directe sur la production. Ainsi, le montant de rebut produit est mesuré. Le contrôle initial sert aussi à refuser la marchandise arrivante. Elle n'influence par conséquent la production que de manière indirecte.

L'ingénieur doit obligatoirement consulter la famille de normes ISO 3591 *Règles d'échantillonnage pour les contrôles par mesures et par attributs* avant de mettre en place des outils de contrôle de réception au sein de son entreprise.

3. Maintenance préventive et contrôle du vieillissement et de la défaillance et impacts critiques (Analyse des Modes de Défaillances, de leurs Effets et de leur Criticité, AMDEC). Il s'agit principalement de calculer la durée de vie de composants ou de machines afin de prévoir des remplacements à l'avance et les actions y relatives à mener pour éviter les situations critiques humaines ou financières.

L'ingénieur doit obligatoirement consulter les normes CEI 61649 *Analyse de Weibull* et NF EN 13306 *Terminologie de la maintenance* avant de mettre en place des outils de maintenance préventive au sein de son entreprise.

Ces trois domaines utilisant les statistiques en général, l'ingénieur devra toujours se référer à la famille de normes ISO 3534 *Vocabulaire et symboles*, ISO 3534-1 *Probabilité et termes statistiques généraux*, ISO 3534-2 *Maîtrise statistique de la qualité*, ISO 3534-3 *Plans d'expérience*.

Indiquons que depuis la fin du 20ème siècle, il est à la mode de regrouper les deux premiers points dans une méthodologie de travail appelée "Six Sigma" que nous allons aborder immédiatement. Enfin, signalons que dans la pratique, pour avoir un intérêt de la direction d'une entreprise, il faut toujours trouver une relation quantitative entre non-qualité et les coûts pour pouvoir faire bouger les choses...

1. SIX SIGMA

Deux objets ne sont jamais rigoureusement identiques. Quelles que soient les techniques utilisées pour fabriquer ces objets, si précis soient les outils, il existe une variabilité dans tout processus de production. L'objectif de tout industriel est que cette variabilité naturelle demeure dans des bornes acceptables. C'est une préoccupation majeure dans l'amélioration de la qualité industrielle.

Un des outils utilisés pour tendre vers cette qualité est la Maîtrise Statistique des Processus (MSP). Si vous produisez un certain type d'objets, et si vous souhaitez conserver vos clients pour pérenniser votre entreprise, vous devez vous assurer que les lots que vous leur livrez sont conformes à ce qui a été convenu entre vous, le plus souvent par contrat. Tout industriel sérieux effectue des contrôles sur les lots produits pour en vérifier la qualité, qu'il en soit le producteur ou bien qu'il les réceptionne. Diverses techniques statistiques liées aux prélèvements d'échantillons sont alors utilisées pour éviter, dans la plupart des cas, de vérifier un à un tous les objets contenus dans un lot. Ce contrôle d'échantillons prélevés dans des lots est indispensable si les contrôles à effectuer détruisent l'objet fabriqué, comme lors d'une analyse de la dose de composant actif contenue dans un comprimé. Il existe cependant des cas où l'on préfère vérifier tous les objets (il est par exemple souhaitable que les freins d'une voiture fonctionnent et un contrôle du freinage sur un échantillon dans la production d'un lot d'automobiles ne garantit pas que tous les véhicules freinent correctement...

Lorsqu'un lot est contrôlé, il est conforme ou il ne l'est pas. S'il est conforme, on le livre (fournisseur) ou on l'accepte (client). S'il n'est pas conforme, on peut le détruire, en vérifier un à un tous les éléments et ne détruire que ceux qui ne sont pas conformes, etc. Toutes les solutions pour traiter les lots non conformes sont onéreuses. Si le lot n'est pas conforme, le mal est fait. La MSP se fixe pour objectif d'éviter de produire des lots non conformes en surveillant la production et en intervenant dès que des anomalies sont constatées. Une bonne MSP permet de supprimer un nombre important de contrôles in fine des lots produits en mettant en place des outils statistiques de surveillance des processus de fabrication.

Pour résumer, la MSP consiste donc à contrôler le procédé en cours de fabrication et à agir sur le procédé plutôt que sur le produit si des problèmes sont détectés. Cette approche tente donc de remonter la chaîne de production le plus haut possible pour prévenir l'apparition de produit défectueux. Nous parlerons dans ce cas en particulier de contrôle de procédés.

Six Sigma est à l'origine une démarche qualité limitée dans un premier temps aux techniques de "**maîtrise statistique des procédés**" (M.S.P.) appelée aussi "**statistiques des processus qualité**" (S.P.Q. ou S.P.C. en anglais pour Statistical Process Control). Le lecteur francophone pourra se référer à la norme AFNOR X06-030 ou différentes normes ISO dont nous ferons référence plus tard dans le texte.

C'est une méthodologie de travail utile pour satisfaire les clients dont l'idée est de délivrer des produits/services de qualité, sachant que la qualité est inversement proportionnelle à la variabilité. Par ailleurs, l'introduction de la qualité doit être optimisée afin de ne pas trop augmenter les coûts. Le jeu subtil entre ces deux paramètres (qualité/coûts) et leur optimisation conjointe est souvent associé au terme de "**Lean management**". Si nous y intégrons Six Sigma, nous parlons alors de "**Lean Six Sigma**".

Six Sigma intègre tous les aspects de la maîtrise de la variabilité en entreprise que ce soit au niveau de la production, des services, de l'organisation ou de la gestion (management). D'où son intérêt! Par ailleurs, dans Six Sigma un défaut doit être paradoxalement la bienvenue, car c'est une source de progrès d'un problème initialement caché. Il faut ensuite se poser plusieurs fois la question "Pourquoi?" (traditionnellement 5 fois) afin de bien remonter à la source de celui-ci.

Nous distinguons deux types de variabilité dans la pratique:

- La "**variabilité inhérente**" au processus (et peu modifiable) qui induit la notion de distribution des mesures (le plus souvent admise par les entreprises comme étant une loi Normale).
- La "**variabilité externe**" qui induit le plus souvent un biais (déviation) dans les distributions dans le temps.

Les processus de fabrication dans l'industrie de pointe ayant une forte tendance à devenir terriblement complexes, il faut noter que les composants de base utilisés pour chaque produit ne sont pas toujours de qualité ou de performance égale. Et si de surcroît, les procédures de fabrication sont difficiles à établir, la dérive sera inévitablement au rendez-vous.

Que ce soit pour l'une ou l'autre raison, au final bon nombre de produits seront en dehors de la normale et s'écarteront ainsi de la fourchette correspondant à la qualité acceptable pour le client. Cette dérive est fort coûteuse pour l'entreprise, la gestion des rebuts, des retouches ou des retours clients pour non-conformité générant des coûts conséquents amputant sérieusement les bénéfices espérés.

Comme nous allons le voir dans ce qui suit, une définition possible assez juste de Six Sigma est: la résolution de problèmes basée sur l'exploitation de données. C'est donc une méthode scientifique de gestion.

1.1. CONTRÔLE QUALITÉ

Dans le cadre des études qualité en entreprise, nous renonçons souvent à un contrôle à 100% à cause du prix que cela engendrerait. Nous procédons alors à une prise d'échantillons. Ceux-ci doivent bien évidemment être représentatifs, c'est-à-dire quelconques et d'égales chances (in extenso le mélange est bon).

Le but de la prise d'échantillons étant bien évidemment la probabilité du taux de défaillance réel du lot complet sur la base des défaillances constatées sur l'échantillonnage.

Rappelons avant d'aller plus loin que nous avons vu dans le chapitre de Statistiques la loi hypergéométrique (et son interprétation) donnée pour rappel par (cf. chapitre de Statistiques):

$$H(k) = \frac{C_k^m C_{p-k}^{n-m}}{C_p^n} \quad (74.1)$$

où la notation du coefficient binomial est conforme à celle définie et choisie dans le chapitre de Probabilités (donc non-conforme à la norme ISO 31-11).

Lors d'un échantillonnage, nous avons normalement un paquet de n éléments dont nous en tirons p . Au lieu de prendre m (nombre entier!) comme le nombre d'éléments défectueux, nous allons implicitement le définir comme étant égal à:

$$m = p_d \cdot n \quad (74.2)$$

où p_d est la probabilité (supposée connue ou imposée...) qu'une pièce soit défectueuse. Ainsi, nous avons pour probabilité de trouver k pièces défectueuses dans un échantillon de p pièces parmi n :

$$H(k) = \frac{C_k^{p_d \cdot n} C_{p-k}^{n-p_d \cdot n}}{C_p^n} = \frac{C_k^{p_d \cdot n} C_{p-k}^{n(1-p_d)}}{C_p^n} \quad (74.3)$$

La probabilité cumulée de trouver k pièces défectueuses (entre 0 et k en d'autres termes) se calcule alors avec la distribution hypergéométrique cumulative:

$$\sum_{x=0}^k H(x) = H(0) + H(1) + \dots + H(k) = \sum_{x=0}^k \frac{C_x^{p_d \cdot n} C_{p-x}^{n(1-p_d)}}{C_p^n} \quad (74.4)$$

Exemple:

Dans un lot n de 100 machines, nous admettons au maximum que 3 soient défectueuses (soit que $p_d = 3\%$). Nous procédons à un échantillonnage p à chaque sortie de commande de 20 machines.

Nous voulons savoir dans un premier temps qu'elle est la probabilité que dans cet échantillonnage p , trois machines soient défectueuses, et dans un deuxième temps quel est le nombre de machines défectueuses maximum autorisé dans cet échantillonnage p qui nous dirait avec 90% de certitude que le lot de n machines en contienne que 3 défectueuses.

x	$H(x)$	$\sum_{x=0}^k H(x)$
0	0.508	0.508
1	0.391	0.899
2	0.094	0.993
3	0.007	1.000

Tableau: 74.1 - Application de la loi hypergéométrique

Ainsi, la probabilité de tirer en une série de tirages trois machines défectueuses dans l'échantillon de 20 est de 0.7% et le nombre de pièces défectueuses maximum autorisé dans cet échantillon de 20 qui nous permet avec au moins 90% de certitude d'avoir 3 défectueuses est de 1 pièce défectueuse trouvée (probabilité cumulée)!

Les valeurs $H(x)$ peuvent être calculées facilement avec la version française de Microsoft Excel 11.8346. Par exemple, la première valeur est obtenue grâce à la fonction:

$$=LOI.HYPERGEOMETRIQUE(0;20;3;100)=0.508$$

1.2. DÉFAUTS/ERREURS

Intéressons-nous donc à exposer pour la culture générale un exemple pratique et particulier de ce qui n'est qu'une application simple de la théorie des statistiques et probabilités. Pour comprendre l'importance de la qualité et du concept "zéro défauts", considérons l'exemple général suivant de l'inventeur de la méthode:

Considérons que le montage d'une voiture de tourisme comprend 2'500 opérations et que chaque opération est parfaite 99 fois sur 100. A priori, réussir une opération dans 99 % des cas est le signe d'une maîtrise quasi parfaite de la qualité. Mais en fait, la perfection de l'ensemble suppose que 2'500 fois de suite, les opérations soient parfaitement réalisées. Si la production quotidienne se monte à 2'000 unités, sur les 5 millions d'opérations effectuées quotidiennement dans notre usine, il y a 1% de défauts de montage soit 50'000, et en moyenne 25 défauts par voiture, ce qui est difficilement acceptable. Aussi, admettons que cette usine imaginaire soit dotée d'un service de contrôle intervenant à la fin du montage de façon systématique. Cela représente un coût considérable en heures de travail de contrôle. Si les défauts peuvent être corrigés, il faudra faire des retouches, remplacer des pièces peut-être, et travailler dans des conditions imprévues pour corriger les défauts. S'ils sont trop importants, ces défauts rendent les produits inutilisables, et les rebuts sont extrêmement coûteux. Pire encore, si le service de contrôle voit 99% des défauts, il en subsiste 500 quotidiennement, et cela suppose des retours et réparations coûteuses ainsi qu'une détérioration significative de l'image de marque selon les performances de la concurrence.

Cet exemple faisant office de cas d'école, imaginons pour la suite une entreprise fabricant trois copies d'un même produit sortant d'une même chaîne, chaque copie étant composée de huit éléments.

Remarque: Nous pouvons tout aussi bien imaginer une société de services développant (fabriquant) trois copies d'un logiciel (produit) sortant d'une même équipe de développement (chaîne), chacun composé d'un nombre égal de modules (éléments).

Supposons que le produit $P1$ a un défaut, le produit $P2$ zéro défauts et le produit $P3$ deux défauts.

Ici, Six Sigma suppose implicitement que les défauts sont des variables indépendantes, ce qui est relativement rare dans les chaînes de fabrication machines mais plus courant dans les chaînes dans lesquelles des humains sont les intervenants. Cependant, nous pouvons considérer lors de l'application de la SPC sur des machines qu'un échantillonnage du temps dans le processus de mesure équivaut à avoir une variable aléatoire!!

Remarques:

R1. Dans le cadre de l'exemple du logiciel pris plus haut, l'indépendance est peu probable si nous ne prenons pas un exemple dans lequel les modules sont personnalisés selon les besoins du client.

R2. L'inconstance des résultats de production de certaines machines dont les réglages bougent pendant le fonctionnement... (ce qui est courant), voir que la matière première change de qualité pendant la production (ce qui est aussi courant!) posent donc de gros problèmes d'application des méthodes SPC.

La moyenne arithmétique des défauts nommée dans le standard Six Sigma "**Defects Per Unit**" (D.P.U.) est alors défini par:

$$DPU = \frac{\text{nombre total de défauts observés}}{\text{nombre total d'unités de production observées}} \quad (74.5)$$

et donne dans notre exemple:

$$DPU = \frac{3}{3} = 1 \quad (74.6)$$

ce qui signifie en moyenne que chaque produit a un défaut de conception ou fabrication. Attention! Cette valeur n'est pas une probabilité pour les simples raisons qu'elle peut d'abord être supérieure à 1 et qu'ensuite elle a comme dimension des [défauts]/[produits].

De même, l'analyse peut être faite au niveau du nombre total d'éléments défectueux possibles qui composent le produit tel que nous sommes amenés naturellement à définir selon le standard Six Sigma le "**Defects per Unit Opportunity**" (D.P.O.):

$$DPO = \frac{DPU}{\text{nombre total d'éléments par produits}} \quad (74.7)$$

ainsi, dans notre exemple, nous avons:

$$DPO = \frac{DPU}{m} = \frac{DPU}{8} = \frac{1}{8} = 0.125 \quad (74.8)$$

et ceci peut être vu comme la probabilité d'avoir un défaut par élément de produit puisque c'est une valeur sans dimensions:

$$DPO = \frac{DPU}{m} = \frac{\left[\frac{\text{défauts}}{\text{unité}} \right]}{\text{total} \left[\frac{\text{défauts}}{\text{unité}} \right]} \quad (74.9)$$

Par extension, nous pouvons argumenter que 87.5% des éléments d'une unité n'ont pas de défauts et comme Six Sigma aime bien travailler avec des exemples de l'ordre du million (c'est plus impressionnant) nous avons alors les "**Defects Per Million Opportunities**" (D.P.M.O.) qui devient:

$$DPMO = DPO \cdot 1'000'000 = \frac{DPU}{m} \cdot 10^6 \quad (74.10)$$

ce qui dans notre exemple donne:

$$DPMO = 125'000 \quad (74.11)$$

Comme la probabilité D qu'un élément d'une pièce soit non défectueux est de 87.5% (soit 12.5% de taux de rebus) alors, par l'axiome des probabilités conjointes (cf. chapitre de Probabilités), la probabilité qu'un produit dans son ensemble soit non défectueux est de:

$$P\left(\bigcap_{i=1..8} D_i\right) = \prod_{i=1}^8 (1 - DPO) = (1 - DPO)^8 \quad (74.12)$$

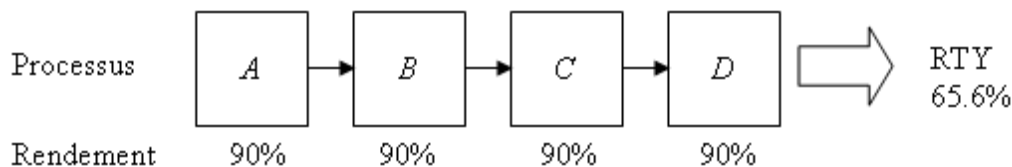
ce qui dans notre exemple donne:

$$P\left(\bigcap_{i=1..8} D_i\right) = 0.875^8 = 0.3436 = 34.36\% \quad (74.13)$$

ce qui n'est pas excellent...

Remarque: Dans Six Sigma, les probabilités conjointes sont aussi naturellement utilisées pour calculer la probabilité conjointe de produits non défectueux dans une chaîne de processus de production P connectés en série. Cette probabilité conjointe est appelée dans Six Sigma "**Rolled Troughput Yield**" (R.T.Y.) ou "**Rendement Global Combiné**" (R.G.C.) et vaut:

$$RTY = P\left(\bigcap_i P_i\right) = \prod_i P_i \quad (74.14)$$



Ce type de calcul étant très utilisé par les logisticiens qui nomment le résultat "**taux de disponibilité**" ainsi que par les chefs de projets pour la durée d'une phase d'un projet lorsqu'ils considèrent les durées des tâches comme indépendantes (sur des structures plus complexes, on parle parfois "d'arbres de probabilités pondérés").

Ainsi, dans une chaîne industrielle basée sur l'exemple précédent pour avoir une quantité Q bien définie de produits (supposés utiliser qu'un seul composant de chaque étape) au bout de la chaîne, il faudra à l'étape A prévoir:

$$A = \frac{Q}{RTY} = \frac{Q}{65.6\%} = 152.42\% \quad (74.15)$$

soit 52.42% de composants A de plus que prévu. Il faudra prévoir à l'étape B:

$$B = \frac{Q}{90\% \cdot 90\% \cdot 90\%} = \frac{Q}{72.9\%} = 137.17\% \quad (74.16)$$

soit 37.17% de composants de plus. Et ainsi de suite...

Rappelons maintenant que la densité de probabilité d'avoir k fois l'événement p et $N-k$ fois l'événement q dans n'importe quel arrangement (ou ordre) est donné par (cf. chapitre de Statistiques):

$$P(n, k) = C_k^n \cdot P = \frac{n!}{k!(n-k)!} p^k q^{n-k} \quad (74.17)$$

et est appelée la loi binomiale ayant pour espérance et écart-type (cf. chapitre de Statistiques):

$$\mu = n \cdot p \quad \sigma = \sqrt{n \cdot p \cdot q} \quad (74.18)$$

Ainsi, dans le standard Six Sigma, nous pouvons appliquer la loi binomiale pour connaître quelle est la probabilité d'avoir zéro éléments défectueux et 8 autres en bon état de marche sur un produit de la chaîne de fabrication de notre exemple (si tous les éléments ont la même probabilité de tomber en panne...):

$$P(8,0) = \frac{8!}{0!(8-0)!} \cdot 0.125^0 \cdot 0.875^8 = 0.125^0 \cdot 0.875^8 = 0.3436 = 34.36\% \quad (74.19)$$

et nous retombons bien évidemment sur la valeur obtenue avec les probabilités conjointes avec:

$$\begin{aligned} \mu &= 8 \cdot 0.125 = 1 = DPU \\ \sigma &= \sqrt{8 \cdot 0.125 \cdot 0.875} = 0.935 \end{aligned} \quad (74.20)$$

Ou la probabilité d'avoir un élément défectueux et sept autres en bon état sur un produit de la chaîne de fabrication:

$$P(8,1) = \frac{8!}{1!(8-1)!} \cdot 0.125^1 \cdot 0.875^7 = \frac{8!}{1!7!} 0.125^1 \cdot 0.875^7 = 0.3436 = 39.26\% \quad (74.21)$$

nous voyons que la loi binomiale nous donne 39.26% de probabilité d'avoir un élément défectueux sur 8 dans un produit.

Par ailleurs, dans le chapitre de Statistiques, nous avons démontré que lorsque la probabilité p est très faible et tend vers zéro mais que toutefois la valeur moyenne $n \cdot p$ tend vers une valeur fixe si n tend vers l'infini, la loi binomiale de moyenne $\mu = n \cdot p$ avec k épreuves était alors donnée par une loi de Poisson:

$$P_k = P(X = k) = p(k) = \frac{\mu^k}{k!} e^{-\mu} \quad (74.22)$$

avec:

$$\begin{aligned} \mu &= n \cdot p \\ \sigma &= \sqrt{\mu} \end{aligned} \quad (74.23)$$

Remarque: Dans un cadre pratique, il est fait usage de l'estimateur de maximum de vraisemblance de la loi exponentielle pour déterminer la moyenne et l'écart-type ci-dessus (cf. chapitre de Statistiques).

Ce que Six Sigma note naturellement:

$$p(k) = \frac{DPU^k}{k!} e^{-DPU} \quad (74.24)$$

avec:

$$\begin{aligned} \mu &= DPU \\ \sigma &= \sqrt{DPU} \end{aligned} \quad (74.25)$$

Ainsi, dans notre exemple, il est intéressant de regarder la valeur obtenue (qui sera forcément différente étant donné que nous sommes loin d'avoir une infinité d'individus et que p est loin d'être petit) en appliquant une telle loi continue (la loi continue la plus proche de la loi binomiale en fait):

$$p(k) = \frac{\mu^k}{k!} e^{-\mu} = \frac{\mu^k}{k!} e^{-n \cdot p} = \frac{\mu^0}{0!} e^{-0.125 \cdot 8} = \frac{1}{0!} e^{-1} = 0.3679 = 36.79\% \quad (74.26)$$

avec:

$$\mu = 1 \quad \sigma = 1 \quad (74.27)$$

ce qui est un résultat encore plus mauvais qu'avec la loi binomiale pour nos produits.

Cependant, si p est fixé au départ, la moyenne $\mu = n \cdot p$ tend également vers l'infini théoriquement dans la loi de Poisson de plus l'écart-type $n \cdot p \cdot q$ tend également vers l'infini.

Si nous voulons calculer la limite de la distribution binomiale, il s'agira donc de faire un changement d'origine qui stabilise la moyenne, en 0 par exemple, et un changement d'unité qui stabilise l'écart, à 1 par exemple. Ce calcul ayant déjà été fait dans le chapitre de Statistique, nous savons que le résultat est la loi Normale:

$$P(k, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(k-\mu)^2}{2\sigma^2}} \quad (74.28)$$

Ainsi, dans notre exemple, nous avons $\mu = 1$ et l'écart-type est donné par l'estimateur sans biais de l'écart-type (cf. chapitre de Statistique):

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2} \quad (74.29)$$

ce qui dans notre exemple donne $\sigma = 1$.

Pour déterminer la probabilité nous calculons la valeur numérique de la loi de Gauss-Laplace (loi Normale) pour $k = 0, \sigma = 1, \mu = 1$:

$$P(0, 1, 1) = \frac{1}{1 \cdot \sqrt{2\pi}} e^{-\frac{(0-1)^2}{2 \cdot 1^2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}} = 0.2419 = 24.19\% \quad (74.30)$$

Ainsi, en appliquant la loi Normale, nous avons 24.19% de chance de tirer au premier coup un produit défectueux. Cet écart par rapport aux autres méthodes s'expliquant simplement par les hypothèses de départ (nombre d'individus fini, probabilité faible, etc.)

Remarque: Ceux qui penseraient utiliser la loi triangulaire (cf. chapitres de Statistiques) doivent tout de suite l'oublier. Effectivement, comme en qualité la valeur optimiste sera le zéro par définition, la probabilité que le nombre de défauts soit égal à 0 sera immédiatement de zéro.

1.3. INDICES DE CAPABILITÉ

Six Sigma (et aussi la série de normes ISO 22514) définit plusieurs indices permettant de mesurer pendant le processus de fabrication la capabilité de contrôle dans le cas d'un grand nombre de mesures de défauts répartis souvent selon une loi de Gauss-Laplace (loi Normale).

Basiquement, si nous nous imaginons dans une entreprise, responsable de la qualité d'usinage d'une nouvelle machine, d'une nouvelle série de pièces, nous allons être confrontés aux deux situations suivantes:

1. Au début de la production, il peut y avoir de gros écarts de qualité dus à des défauts de la machine ou de réglages importants mal initialisés. Ce sont des défauts qui vont souvent être rapidement corrigés (sur le court terme). Dès lors pendant cette période de grosses corrections, nous faisons des contrôles par lot (entre chaque grosse correction) et chacun sera considéré comme une variable aléatoire indépendante et identiquement distribuée (selon une loi Normale) mais de moyenne et écart-type bien évidemment différent.

2. Une fois les gros défauts corrigés, nous n'allons avoir en théorie plus que des défauts minimes très difficilement contrôlables et ce même sur long terme. Alors l'analyse statistique ne se fait plus forcément par lot de pièces mais par pièces et l'ensemble des pièces sur le long terme est considéré comme un unique lot à chaque fois.

Ces deux scénarios mettent en évidence que nous n'effectuons alors logiquement pas les mêmes analyses en début de production et ensuite sur le long terme. Raison pour laquelle en SPC nous définissons plusieurs indices (dont les notations sont propre à ce site Internet car elles changent selon les normes) dont 2 principaux qui sont:

D1. Nous appelons "**capabilité potentielle du procédé court terme**" ou "**indice dispersion court terme**" le rapport entre l'étendue de contrôle E de la distribution des valeurs et la qualité de Six Sigma (6 sigma) lorsque le processus est centré (c'est-à-dire sous contrôle statistique - les caractéristiques du produit fabriqué varient peu) tel que:

$$C_p = \frac{E}{6\sigma} \quad (74.31)$$

ce qui s'écrit aussi:

$$C_p = \frac{USL - LSL}{6\sigma_{CT}} \quad (74.32)$$

où USL est la limite supérieure de contrôle/tolérance ou "**Upper Specification Level**" (USL) de la distribution et LSL la limite inférieure ou "**Lower Specification Level**" (LSL) que nous imposons souvent (mais pas toujours!) dans l'industrie comme à distances égales par rapport à la moyenne μ théorique souhaitée. La capabilité ci-dessus est donc un indice que nous chercherons à maximiser (puisque l'écart-type au dénominateur doit être minimisé!).

Ce rapport est utile dans l'industrie dans le sens où l'étendue E (qui est importante, car elle représente la dispersion/variation du processus) est assimilée à la "voix du client" (ses exigences) et le 6 sigma au dénominateur au comportement réel du procédé/processus censé inclure quasiment toutes les issues possibles. Il vaut donc mieux espérer que ce rapport soit au pire égal à l'unité!

Voici typiquement un exemple en gestion de projets où, lorsque le client ne paie pas pour une modélisation fine du risque fine (le mandataire accepte alors par contrat une variation des délais et coûts qui peut dépasser les 50%), on tombe sur ce type de distribution :

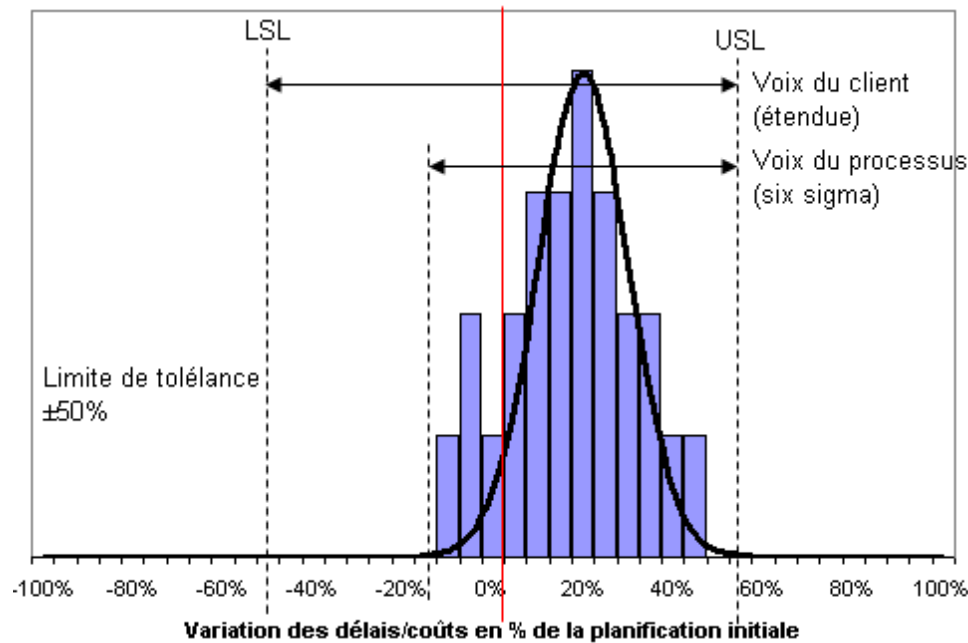


Figure: 74.1 - Tracé typique d'un plot de contrôle avec étendues et limites

Remarque: En MSP, l'étendue E est souvent notée IT , signifiant "**intervalle de tolérance**".

L'écart-type au dénominateur étant donné par la relation démontrée dans le chapitre de Statistiques dans le cas de k variables aléatoires indépendantes et identiquement distribuées selon une loi Normale (mais d'écart-type et moyenne non-identique):

$$\sigma_{CT} = \sqrt{\frac{\sum_{i=1}^k \sigma_i^2}{k}} \quad (74.33)$$

où CT est l'abréviation de "court terme" (abréviation souvent non précisée dans la pratique car supposée connue dans le contexte). Cet écart-type est bien évidemment le meilleur pour le premier scénario dont nous avons fait mention plus haut. Car entre chaque grosse correction, les lots sont considérés comme indépendants et ne peuvent pas être analysés comme un seul et unique lot (ce serait une aberration!).

Avant de continuer, voyons qu'il nous est possible de construire facilement un test d'hypothèse pour l'indicateur C_p . Effectivement, en se rappelant que nous avons démontré dans le chapitre de Statistique que:

$$\frac{(n-1)S^2}{\chi_{\alpha/2, n-1}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-\alpha/2, n-1}^2} \quad (7.34)$$

Il est alors immédiat que:

$$\frac{(n-1)S^2}{\chi_{\alpha/2, n-1}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-\alpha/2, n-1}^2} \quad (7.35)$$

et in extenso:

$$\frac{1}{S} \sqrt{\frac{\chi_{\alpha/2, n-1}^2}{(n-1)}} \leq \frac{1}{\sigma} \leq \frac{1}{S} \sqrt{\frac{\chi_{1-\alpha/2, n-1}^2}{(n-1)}} \quad (7.36)$$

Et dès lors:

$$\frac{E}{S} \sqrt{\frac{\chi_{\alpha/2, n-1}^2}{(n-1)}} \leq \frac{E}{\sigma} \leq \frac{E}{S} \sqrt{\frac{\chi_{1-\alpha/2, n-1}^2}{(n-1)}} \quad (7.37)$$

Et en adaptant l'écriture au domaine du génie industriel, il vient:

$$\frac{E}{\hat{\sigma}_{CT}} \sqrt{\frac{\chi_{\alpha/2, n-1}^2}{(n-1)}} \leq \frac{E}{\sigma_{CT}} \leq \frac{E}{\hat{\sigma}_{CT}} \sqrt{\frac{\chi_{1-\alpha/2, n-1}^2}{(n-1)}} \quad (7.38)$$

ou encore:

$$\hat{C}_P \sqrt{\frac{\chi_{\alpha/2, n-1}^2}{(n-1)}} \leq C_P \leq \hat{C}_P \sqrt{\frac{\chi_{1-\alpha/2, n-1}^2}{(n-1)}} \quad (7.39)$$

et nous pouvons appliquer le même raisonnement pour tous les types d'indicateurs du même genre que nous verrons par la suite!

Remarque: Dans le cadre de notre étude plus loin des cartes de contrôle, nous verrons qu'il est possible d'utiliser des expressions particulières pour l'écart-type lorsque nous travaillons avec des échantillons de mesures. Ces expressions seront basées pour l'une sur la loi du Khi-deux et pour l'autre les statistiques d'ordre.

Attention cependant! Comme souvent dans la situation court terme (lors de la correction des grosses sources d'erreurs donc) les lots de tests sont petits, même très petits, afin de diminuer les coûts en production. Dès lors l'écart-type se trouvant sous la racine (qui est l'estimateur de maximum de vraisemblance de la loi Normale) n'a pas une valeur vraiment correcte... Il est alors bon d'utiliser soit d'autres méthodes de calcul assez empiriques comme le font de nombreux logiciels, soit de calculer un intervalle de confiance de l'indice de capabilité en calculant l'intervalle de confiance de l'écart-type court terme comme nous l'avons vu dans le chapitre de Statistiques.

D2. Nous appelons "**Performance globale du procédé long terme**" le rapport entre l'étendue de contrôle E de la distribution des valeurs et la qualité de Six Sigma (6 sigma) lorsque le processus est centré tel que:

$$P_p = \frac{E}{6\sigma} \quad (74.40)$$

ce qui s'écrit aussi:

$$P_p = \frac{USL - LSL}{6\sigma_{LT}} \quad (74.41)$$

L'écart-type au dénominateur étant donné cette fois par le cas où nous considérons tous les gros défauts corrigés et le processus stable afin de considérer toutes les pièces fabriquées comme un seul et unique lot de contrôle:

$$\sigma_{LT} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \quad (74.42)$$

où LT est l'abréviation de "long terme" (abréviation souvent non précisée dans la pratique car supposée connue dans le contexte). Cet écart-type est bien évidemment le meilleur pour le deuxième scénario dont nous avons fait mention plus haut. Car les variations étant maintenant, par hypothèse, toutes petites, l'ensemble de la fabrication peut être supposé comme étant un seul et unique lot de contrôle sur le long terme (bon cela n'empêche pas qu'il faut parfois nettoyer les valeurs extrêmes qui peuvent se produire).

Le tolérancement des caractéristiques est donc très important pour l'obtention de la qualité et de la fiabilité des produits assemblés. Traditionnellement, une tolérance s'exprime sous la forme d'un bipoint [Min,Max]. Une caractéristique est alors déclarée conforme si elle se situe dans les tolérances.

Le problème du tolérancement consiste à tenter de concilier la fixation des limites de variabilité acceptable les plus larges possibles pour diminuer les coûts de production et d'assurer un niveau de qualité optimal sur le produit fini.

Deux approches tentent de résoudre ce problème:

1. Le tolérancement au pire des cas garantit l'assemblage dans toutes les situations à partir du moment où les caractéristiques élémentaires sont dans les tolérances.
2. Le tolérancement statistique tient compte de la faible probabilité d'assemblages d'extrêmes entre eux et permet d'élargir de façon importante les tolérances pour diminuer les coûts et c'est donc à celui-ci que nous allons nous intéresser ici comme vous l'aurez compris.

Un processus est dit "**limite capable**" (soit limite stable par rapport aux exigences du client en d'autres termes) s'il le ratio donné ci-dessus (en choisissant 6 fois l'écart-type) est supérieur à 1. Mais dans l'industrie, on préfère prendre en réalité la valeur de ~ 1.33 dans le cas d'une distribution Normale des données.

Bien évidemment, la valeur σ de l'écart-type peut être calculée en utilisant les estimateurs de maximum de vraisemblance avec ou sans biais vus dans le chapitre de Statistiques mais il ne s'agit en aucun cas dans la réalité pratique de l'écart-type théorique mais uniquement d'un estimateur! Par ailleurs, nous verrons plus loin qu'en fonction de l'écart-type utilisé, les notations changent!

Remarque: En entreprise, il faut faire attention car l'instrument de mesure rajoute son propre écart-type (erreur) sur celui de la production.

Comme nous l'avons démontré au chapitre de Statistiques, l'erreur-type (écart-type de la moyenne) est:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (74.43)$$

Dans la méthodologie Six Sigma, nous prenons alors souvent pour les processus à long terme et sous contrôle:

$$\begin{aligned} LCL &= \mu - \frac{3\sigma}{\sqrt{n}} \\ UCL &= \mu + \frac{3\sigma}{\sqrt{n}} \end{aligned} \quad (74.44)$$

quand nous analysons des cartes de contrôle dont les variables aléatoires sont des échantillons de n variables aléatoires indépendantes et identiquement distribuées et que les limites n'ont pas été imposées par un client ou par une politique interne ou des contraintes techniques! Bien évidemment, il faut bien être conscient que UCL et LCL n'ont pas la même expression dans des cas plus complexes et donc pour des distributions autres que la loi Normale!

Par ailleurs, l'expression précédente diffère aussi pour les processus à court terme, car l'exemple donné ci-dessus est pour un cas de mesures sur le long terme uniquement pour rappel!

Le lecteur remarquera que nous avons maintenant:

$$C_p = \frac{USL - LSL}{UCL - LCL} \quad (74.45)$$

Normalement, au sein des entreprises, l'étendue de contrôle est fixe (le numérateur) et donc quand la valeur de l'écart-type est grande (plus de variations, moins de contrôles) la valeur de l'indice est faible et lorsque l'écart-type est faible (moins de variations, plus de contrôles) la valeur de l'indice est élevée.

Comme le montrent les deux exemples ci-dessous:

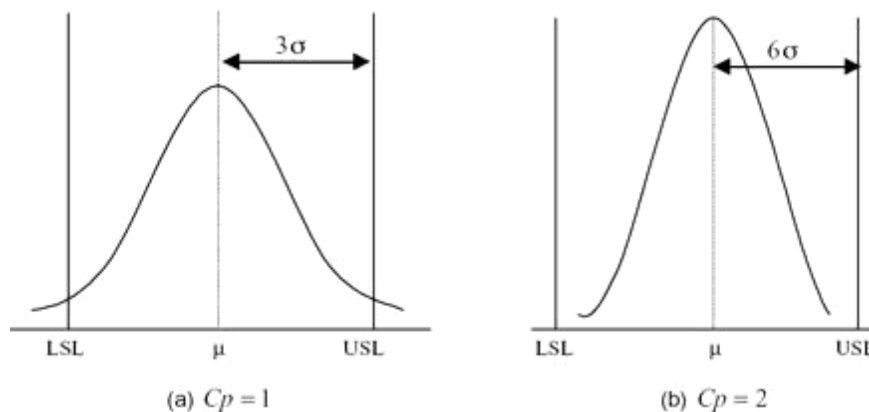


Figure: 74.2 - Deux plots de mesures sous contrôle statistique avec une capacité différente

L'indice C_p impose que la moyenne (l'objectif) est centrée entre LSL et USL . Dès lors, la moyenne est confondue avec ce que nous appelons la "cible" T du processus.

Mais la moyenne μ dans la réalité peut être décalée par rapport à l'objectif T initial qui doit lui toujours (dans l'usage courant) être à distance égale entre USL et LSL comme le montre la figure ci-dessous dans le cas particulier d'une loi Normale:

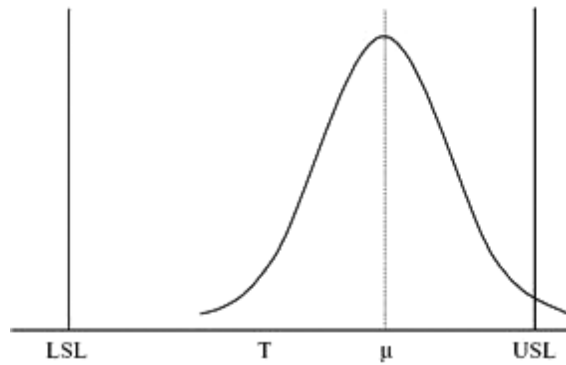


Figure: 74.3 - Mesures sous contrôle statistique décalé par rapport à la cible

Mais ce n'est pas forcément le cas dans la réalité où les ingénieurs (quelque soit leur domaine d'application) peuvent choisir des LSL et USL asymétriques par rapport à la moyenne ne serait-ce que parce que la loi n'est pas toujours Normale (typiquement le cas en gestion de projets...)! D'où la définition suivante:

D2. Nous appelons alors "**Capabilité potentielle décentrée court terme du procédé**" (dans le cas décentré) ou "**Process Capability Index (within)**" la relation:

$$C_{pk} = C_p (1 - k) \quad (74.46)$$

avec:

$$k = \frac{|T - \mu|}{\frac{1}{2}(USL - LSL)} = \frac{|T - \mu|}{3\sigma C_p} \quad (74.47)$$

où $k \geq 0$ est appelé le "**dégré de biais**" ou "**indice de position**" et T le "**target**" donné naturellement par:

$$T = \frac{USL + LSL}{2} \quad (74.48)$$

qui donne le milieu de la distribution relativement au bi-point $[LSL, USL]$ imposé (ne pas oublier que l'écart-type au dénominateur de la relation antérieure est l'écart-type court terme!).

Au fait cet indicateur de capabilité de contrôle peut sembler très artificiel, mais il ne l'est pas totalement... Effectivement, il y a quelques valeurs remarquables (celles qui intéressent l'ingénieur) qui permettent de se faire une bonne idée de ce qu'il se passe avec celui-ci:

1. Si la moyenne et la cible sont confondues, nous avons alors:

$$T = \mu \quad (74.49)$$

nous nous retrouvons donc avec $k = 0$ et donc $C_{pk} = C_p$ et le critère de jugement de la valeur de l'indice sera basé sur l'indice de capabilité centré court terme.

2. Si faute d'un mauvais contrôle du processus nous avons:

$$|T - \mu| > 0.5 \cdot (USL - LSL) \quad (74.50)$$

alors la moyenne μ est soit au-dessus de USL soit en dessous de LSL ce qui a pour conséquence d'avoir $k > 1$ et donc $C_{pk} < 0$.

3. Si nous avons:

$$0 < |T - \mu| < 0.5 \cdot (USL - LSL) \quad (74.51)$$

alors la moyenne μ est comprise entre les valeurs USL et LSL ce qui a pour conséquence d'avoir $0 < k < 1$ et donc $0 < C_{pk} < C_p$.

4. Si nous avons:

$$|T - \mu| = 0.5 \cdot (USL - LSL) \quad (74.52)$$

alors cela signifie simplement que la moyenne est confondue avec USL ou LSL et nous avons alors $k = 1$ et $C_{pk} = 0$.

Comme l'interprétation reste cependant délicate et difficile, nous construisons les indices de capacité unilatéraux "**Upper Capability Index CPU**" et "**Lower Capability Index CPL**" donnés par:

$$CPU = \frac{USL - \mu}{3\sigma} \quad CPL = \frac{\mu - LSL}{3\sigma} \quad (74.53)$$

que nous chercherons bien évidemment aussi à maximiser. Voyons d'où viennent ces deux valeurs et comment les utiliser:

Démonstration:

D'abord, nous avons besoin de deux formulations particulières du degré de biais k .

Si:

$$\frac{USL + LSL}{2} < \mu \quad (74.54)$$

alors nous pouvons nous débarrasser de la valeur absolue:

$$\begin{aligned} k &= \frac{|T - \mu|}{\frac{1}{2}(USL - LSL)} = \frac{\left| \frac{1}{2}(USL + LSL) - \mu \right|}{\frac{1}{2}(USL - LSL)} = \frac{\mu - \frac{1}{2}(USL + LSL)}{\frac{1}{2}(USL - LSL)} \\ &= \frac{2\mu}{USL - LSL} - \frac{USL + LSL}{USL - LSL} \end{aligned} \quad (74.55)$$

Si:

$$\frac{USL + LSL}{2} > \mu \quad (74.56)$$

alors nous pouvons nous débarrasser de la valeur absolue:

$$\begin{aligned}
 k &= \frac{|T - \mu|}{\frac{1}{2}(USL - LSL)} = \frac{\left| \frac{1}{2}(USL + LSL) - \mu \right|}{\frac{1}{2}(USL - LSL)} = \frac{\frac{1}{2}(USL + LSL) - \mu}{\frac{1}{2}(USL - LSL)} \quad (74.57) \\
 &= \frac{USL + LSL}{USL - LSL} - \frac{2\mu}{USL - LSL}
 \end{aligned}$$

Nous avons alors lorsque $T > \mu$:

$$\begin{aligned}
 C_p(1-k) &= C_p \left[1 - \left(\frac{USL + LSL}{USL - LSL} - \frac{2\mu}{USL - LSL} \right) \right] \\
 &= C_p \left[1 - \left(\frac{USL + LSL}{USL - LSL} - \frac{2\mu}{USL - LSL} \right) \right] \\
 &= \frac{USL - LSL}{6\sigma} \left[1 - \left(\frac{USL + LSL}{USL - LSL} - \frac{2\mu}{USL - LSL} \right) \right] \quad (74.58) \\
 &= \left(\frac{USL - LSL}{6\sigma} - \frac{USL - LSL}{6\sigma} \frac{USL + LSL}{USL - LSL} + \frac{USL - LSL}{6\sigma} \frac{2\mu}{USL - LSL} \right) \\
 &= \left(\frac{USL - LSL}{6\sigma} - \frac{USL + LSL}{6\sigma} + \frac{2\mu}{6\sigma} \right) = \left(-\frac{2LSL}{6\sigma} + \frac{2\mu}{6\sigma} \right) = \left(\frac{-LSL + \mu}{3\sigma} \right) \\
 &= \frac{\mu - LSL}{3\sigma} = \frac{1}{3} Z_{\Pi}
 \end{aligned}$$

et respectivement lorsque $T < \mu$:

$$\begin{aligned}
 C_p(1-k) &= C_p \left[1 - \left(\frac{2\mu}{USL - LSL} - \frac{USL + LSL}{USL - LSL} \right) \right] \\
 &= C_p \left[1 - \left(\frac{2\mu}{USL - LSL} - \frac{USL + LSL}{USL - LSL} \right) \right] \\
 &= \frac{USL - LSL}{6\sigma} \left[1 - \left(\frac{2\mu}{USL - LSL} - \frac{USL + LSL}{USL - LSL} \right) \right] \quad (74.59) \\
 &= \left(\frac{USL - LSL}{6\sigma} - \frac{USL - LSL}{6\sigma} \frac{2\mu}{USL - LSL} + \frac{USL - LSL}{6\sigma} \frac{USL + LSL}{USL - LSL} \right) \\
 &= \left(\frac{USL - LSL}{6\sigma} - \frac{2\mu}{6\sigma} + \frac{USL + LSL}{6\sigma} \right) = \left(\frac{2USL}{6\sigma} - \frac{2\mu}{6\sigma} \right) \\
 &= \frac{USL - \mu}{3\sigma} = \frac{1}{3} Z_{ul}
 \end{aligned}$$

□ C.Q.F.D.

À long terme, dans certaines entreprises, il est intéressant de savoir qu'elles sont les plus mauvaises valeurs prises par les indices *CPU* et *CPL* (c'est le cas dans le domaine de la production mais pas forcément de la gestion de projets)

Les plus mauvaises valeurs étant trivialement les plus petites, nous prenons souvent (avec quelques unes des notations différentes que l'on peut trouver dans la littérature spécialisée...):

$$C_{pk} = \min \{CPU, CPL\} = \min \{C_{pu}, C_{pl}\} = \min \left\{ \frac{1}{3} Z_{ul}, \frac{1}{3} Z_{ll} \right\} \quad (74.60)$$

Voici par exemple un diagramme d'analyse de la capacité généré par le logiciel Minitab 15.1.1 (en anglais) avec les différents facteurs susmentionnés sur un échantillon de 68 données suivant une loi Normale (un test de normalité a été fait avant):

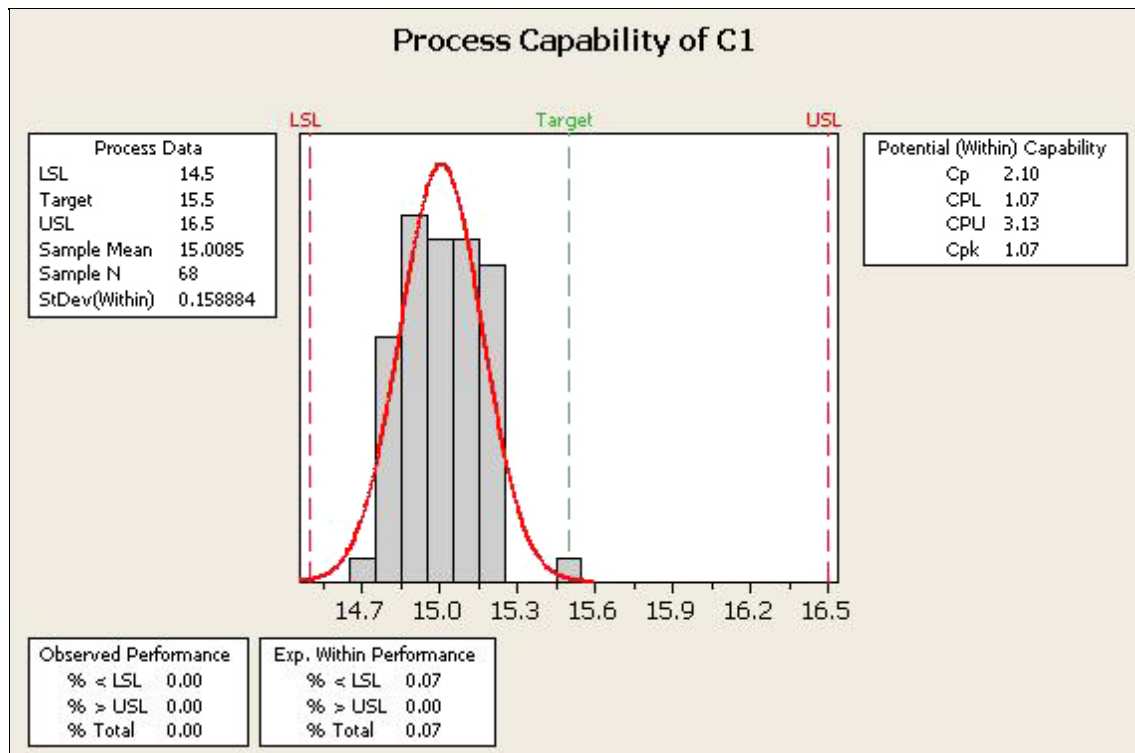


Figure: 74.4 - Diagramme de capacité généré avec Minitab 15.1.1

Deux lectures typiques sont possibles (nous expliquerons la partie inférieure gauche du graphique plus loin):

1. En production: Le processus est capable (valeur >1.33) mais avec une (trop) forte déviation vers le gauche par rapport à la cible définie, ce qui n'est pas bon (*CPL* ayant la valeur la plus petite) et doit être corrigé.

2. En gestion de projets: Les tâches redondantes sont sous contrôle (valeur >1.33) mais avec une forte déviation vers le gauche, ce qui peut être bon si notre objectif est de prendre de l'avance par rapport au planifié (rien à corriger).

Il faut vraiment prendre garde au fait que dans la réalité, il n'est pas toujours possible de prendre la loi Normale, or tous les exemples donnés ci-dessus sont basés sur cette hypothèse simplificatrice.

Toujours le cadre de la gestion de la qualité en production, la figure ci-dessous représente bien la réalité dans le cadre d'un processus court ou long terme:

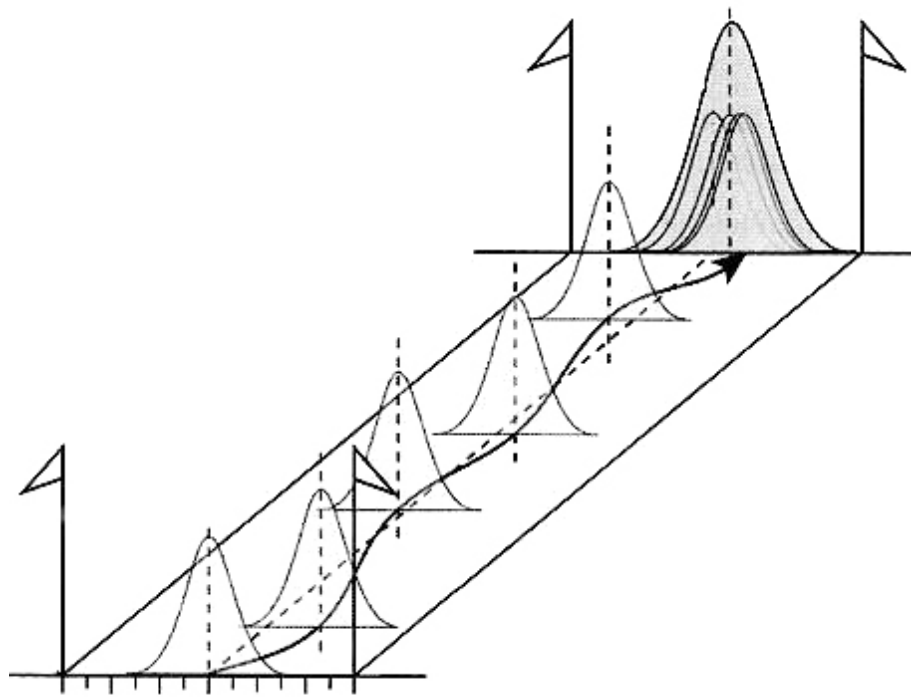


Figure: 74.5 - Processus court/long terme (source: MSP/SPC de Maurice Pillet)

Chaque petite gaussienne en gris clair, représente une analyse de lots assimilé au concept de "**dispersion instantanée**". Nous voyons bien que leurs moyennes ne cessent de bouger pendant la période de mesures (que cette variation soit grande ou très faible) et c'est ce que nous appelons la "**dispersion globale**". Le but dans les organisations (industries ou administrations) est de faire en sorte que la variabilité instantanée ou globale soit limitée au maximum.

Or la relation définissant C_p supposait, comme nous l'avons mentionné, que le processus est sous contrôle centré (donc toutes les gaussiennes sont alignées) et sur une optique court terme.

De même, la relation définissant C_{pk} supposait, comme nous l'avons mentionné, que le processus est sous contrôle, sur une optique court terme et décentré par choix (ou à cause du fait que la loi n'est pas Normale).

Par contre, si le processus n'est pas centré parce qu'il n'est pas sous contrôle alors qu'il devrait l'être, la variable aléatoire mesurée est la somme de la variation aléatoire des réglages X de la machine et des variations aléatoires non contrôlables des contraintes des pièces Y .

L'écart-type total est alors, si les deux variables aléatoires suivent une loi normale, la racine carrée de la somme des écarts-types (cf. chapitre de Statistiques):

$$\sigma = \sqrt{\sigma_X^2 + \sigma_Y^2} = \sqrt{\sigma_{CT}^2 + \sigma_Y^2} \quad (74.61)$$

Or, si nous n'avons qu'une seule mesure, il vient en prenant l'estimateur biaisé (c'est un peu n'importe quoi de l'utiliser dans ce cas-là mais bon...):

$$\sigma_Y^2 = \frac{1}{n} (Y - \mu_Y)^2 = (Y - \mu_Y)^2 \quad (74.62)$$

Or dans le cas d'étude qui nous intéresse Y représente la moyenne expérimentale (mesurée) du processus qu'on cherche à mettre sous contrôle. Cette moyenne est notée traditionnellement m dans le domaine.

Ensuite, μ_Y n'étant pas connu on prend ce qu'il devrait être: c'est la cible T du processus. Ainsi, nous introduisons un nouvel indice appelé "**Capabilité potentielle décentrée moyenne court terme du procédé**":

$$C_{pm} = \frac{USL - LSL}{6\sigma} \cong \frac{USL - LSL}{6\sqrt{\sigma^2 + (m - T)^2}} \quad (74.63)$$

où encore une fois il faut se rappeler que l'écart-type dans la racine au dénominateur est l'écart-type court terme!

Nous voyons immédiatement que plus C_{pm} est proche de C_p mieux c'est (dans les domaines de production du moins).

Nous avons donc finalement les trois indices de capabilités court terme centré et non centré les plus courants (nous avons délibérément choisi d'uniformiser les notations et de mettre le maximum d'infos dans celles-ci):

$$C_p = \frac{USL - LSL}{\sigma_{CT}} \quad C_{pk} = \left\{ \frac{USL - \mu}{3\sigma_{CT}}, \frac{\mu - LSL}{3\sigma_{CT}} \right\} \quad C_{pm} = \frac{USL - LSL}{6\sqrt{\sigma_{CT}^2 + (\mu - T)^2}} \quad (74.64)$$

De même nous avons aussi les trois indices de capabilités long terme centré et non centré les plus courants (nous avons délibérément choisi d'uniformiser les notations et de mettre le maximum d'infos dans celles-ci):

$$P_p = \frac{USL - LSL}{\sigma_{LT}} \quad P_{pk} = \left\{ \frac{USL - \mu}{3\sigma_{LT}}, \frac{\mu - LSL}{3\sigma_{LT}} \right\} \quad P_{pm} = \frac{USL - LSL}{6\sqrt{\sigma_{LT}^2 + (\mu - T)^2}} \quad (74.65)$$

Enfin, indiquons que bien que ce soit pas très pertinent, il arrive parfois que certains ingénieurs fassent les deux analyses (court terme + long terme) en même temps sur la même base de données de mesures.

Cependant, pour faire de l'analyse objective sur les indices de capabilité vus jusqu'à maintenant, il faudrait d'abord que les instruments de mesure soient eux-mêmes capables... ce que nous appelons souvent les "**méthodes R&R**" (Répétabilité, Reproductibilité).

Le principe de base (car le principe avancé consiste à faire une ANOVA à deux facteurs avec répétition) consiste alors à évaluer la dispersion courte terme ou respectivement long terme de l'instrument de mesure afin de calculer une "**capabilité de processus de contrôle**" définie par:

$$C_{pc} = \frac{E}{6\sigma_{\text{instrument}}} \quad (74.66)$$

Dans les cas classiques, nous déclarons le moyen de contrôle capable pour un suivi MSP lorsque cette capabilité est supérieure à 4 et nous allons de suite voir pourquoi. Rappelons pour cela d'abord que:

$$C_p = \frac{E}{6\sigma} \quad (74.67)$$

Mais la variance observée est au fait la somme de la "vraie" variance et de celle de l'instrument telle que:

$$C_{p,\text{observé}} = \frac{E}{6\sigma} = \frac{E}{6\sqrt{\sigma_{\text{vrai}}^2 + \sigma_{\text{instrument}}^2}} \quad (74.68)$$

Or nous avons:

$$C_{p,\text{vrai}} = \frac{E}{6\sigma_{\text{vrai}}} \text{ et } C_{pc,\text{vrai}} = \frac{E}{6\sigma_{\text{instrument}}} \quad (74.69)$$

En mettant le tout au carré, nous en déduisons:

$$\begin{aligned} C_{p,\text{observé}}^2 &= \frac{E^2}{36(\sigma_{\text{vrai}}^2 + \sigma_{\text{instrument}}^2)} = \frac{E^2}{36\left(\frac{E^2}{36C_{p,\text{vrai}}^2} + \frac{E^2}{36C_{pc,\text{instrument}}^2}\right)} \\ &= \frac{1}{\left(\frac{1}{C_{p,\text{vrai}}^2} + \frac{1}{C_{pc,\text{instrument}}^2}\right)} = \frac{1}{\left(\frac{C_{pc}^2 + C_{p,\text{vrai}}^2}{C_{p,\text{vrai}}^2 \cdot C_{pc}^2}\right)} \end{aligned} \quad (74.70)$$

D'où:

$$C_{p,\text{observé}}^2 = \frac{C_{p,\text{vrai}}^2 \cdot C_{pc}^2}{C_{pc}^2 + C_{p,\text{vrai}}^2} \quad (74.71)$$

Ce qui nous donne:

$$C_{p,\text{observé}}^2 (C_{pc}^2 + C_{p,\text{vrai}}^2) = C_{p,\text{vrai}}^2 \cdot C_{pc}^2 \Rightarrow C_{p,\text{observé}}^2 C_{pc}^2 = C_{p,\text{vrai}}^2 (C_{pc}^2 - C_{p,\text{observé}}^2) \quad (74.72)$$

Soit:

$$C_{p,\text{vrai}} = \frac{C_{p,\text{observé}} C_{pc}}{\sqrt{C_{pc}^2 - C_{p,\text{observé}}^2}} = \frac{C_{p,\text{observé}}}{\sqrt{1 - \frac{C_{p,\text{observé}}^2}{C_{pc}^2}}} \quad (74.73)$$

Ce qui se traduit par le graphique de la figure suivante qui montre bien l'intérêt d'un C_{pc} au moins égal à 4!

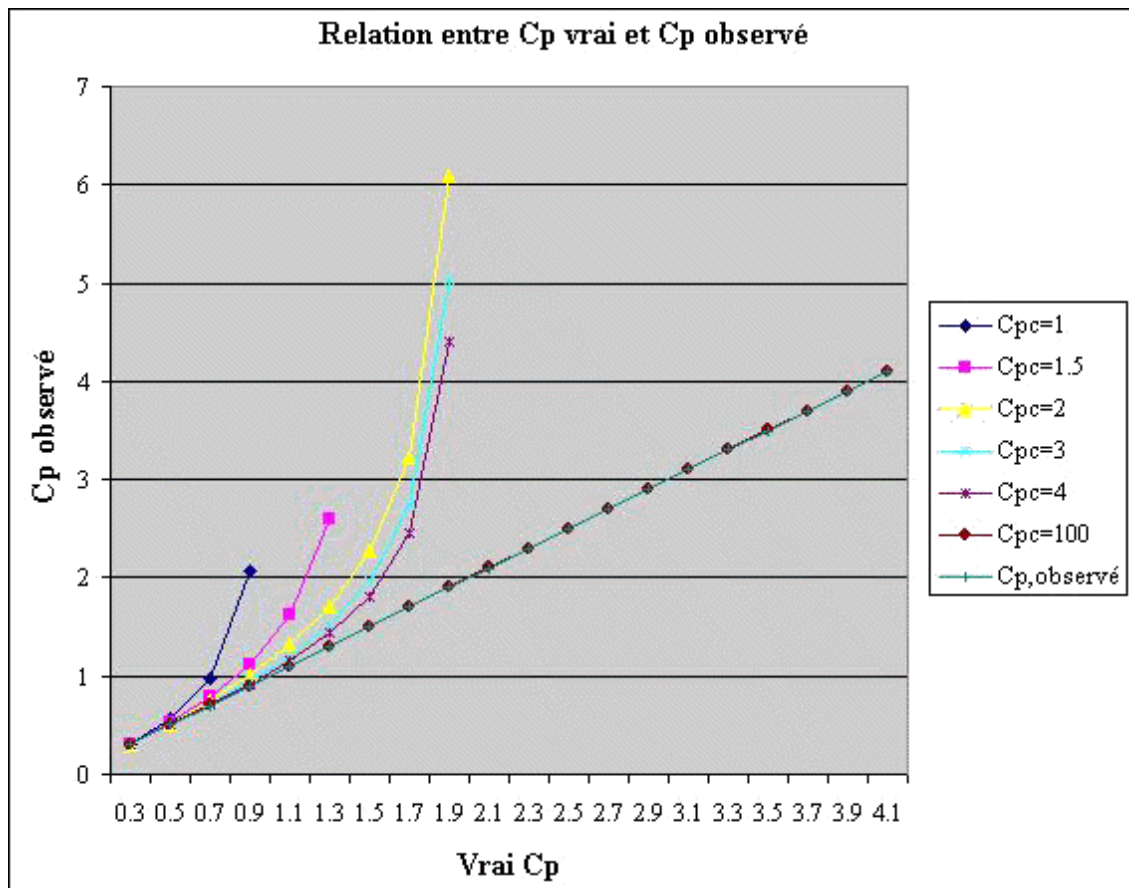


Figure: 74.6 - Relation entre capabilité vraie et mesurées

Dans la pratique, signalons que pour déterminer C_{pc} on se sert d'une pièce étalon mesurée par interférométrie LASER et s'assurer ensuite que tous les essais répétés de mesure se fassent sur les deux mêmes points de mesure.

Une fois ceci fait, on effectue plusieurs mesures de la cote étalon et on prend l'écart-type de ces mesures. Ce qui donnera le $\sigma_{\text{instrument}}$.

L'étendue E est elle imposée par le client ou par des ingénieurs internes à l'entreprise. Elle sera souvent prise comme étant au plus dixième de l'unité de tolérance d'une pièce.

Par exemple, si nous avons un diamètre intérieur de $36 \pm 1 [\mu m]$ (étendue de tolérance de 2 microns ce qui est déjà du haut de gamme niveau précision, car à notre époque le standard se situe plutôt autour des 3!), notre appareil devra alors avoir selon la règle précédemment citée une étendue de 0.2 microns... Il est alors aisé de déterminer qu'elle devra être l'écart-type maximum de l'instrument si on se fixe une capacité de processus de contrôle de 4 (et encore... 4 c'est grossier!).

Certains ingénieurs apprécient de savoir à combien d'éléments en millions d'unités produites (parties par million: PPM) seront considérées comme défectueuses relativement.

Le calcul est alors aisé puisque l'ingénieur a à sa disposition au moins les informations suivantes:

$$\sigma, \mu, USL, LSL, T \quad (74.74)$$

et que les données suivent une loi Normale alors il est immédiat que (cf. chapitre de Statistiques):

$$PPM_{LL}^{CT} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{LL} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \cdot 10^6 \quad (74.75)$$

et:

$$PPM_{UL}^{CT} = \left(1 - \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{UL} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \right) \cdot 10^6 \quad (74.76)$$

valeurs très aisées à obtenir avec n'importe quel tableur comme Microsoft Excel par exemple.

Nous avons alors

$$PPM_{Tot}^{CT} = PPM_{Tot} = PPM_{UL} + PPM_{LL} \quad (74.77)$$

il en est de même pour la capacité long terme (il suffit de prendre alors l'expression correspondante de l'écart-type).

1.4. NIVEAUX DE QUALITÉS

Signalons un point important relativement à Six Sigma. Au fait, objectivement, l'idée de cette méthode est, certes, de faire de la SPC (entre autres, mais ça ce n'est pas nouveau) mais surtout de garantir au client selon la tradition couramment admise avec un écart-type ayant une borne supérieure de $\sigma = 1$ avec une déviation à la moyenne (en valeur absolue) de 1.5σ par rapport à la cible, ce qui garantit au plus 3.4 PPM (c'est-à-dire 3.4 rejets par million).

Remarque: Ce choix empirique vient de la mise en pratique de la méthode Six Sigma par son créateur (Bill Smith). Il a observé dans son entreprise (Motorola) que sous contrôle statistique, il avait quasiment systématiquement une déviation comprise entre 1.2 et 1.8σ à la moyenne pour tous ses procédés industriels.

Voyons d'où vient cette dernière valeur à l'aide des deux tableaux suivants:

1. D'abord construisons un tableau de type idéal qui présente des données d'un procédé court terme (mais les calculs sont parfaitement identiques pour du long terme) centré sur la cible (de cible nulle ici, ce qui est un cas typique), de moyenne nulle (donc sur la cible et alors donc $C_p = C_{pk}$) et d'écart-type unitaire avec USL et LSL symétriques (ce qui restreint par contre le champ d'application):

C_p	C_{pk}	Défauts (PPM)	Niveau de qualité Sigma	Critère
0.5	0.5	133614	1.5	Mauvais
0.6	0.6	71861	1.8	
0.7	0.7	35729	2.1	
0.8	0.8	16395	2.4	
0.9	0.9	6934	2.7	
1	1	2700	3	
1.1	1.1	967	3.3	
1.2	1.2	318	3.6	
1.3	1.3	96	3.9	Limite
1.4	1.4	27	4.2	
1.5	1.5	6.8	4.5	
1.6	1.6	1.6	4.8	
1.7	1.7	0.34	5.1	
1.8	1.8	0.067	5.4	
1.9	1.9	0.012	5.7	
2	2	0.002	6	Excellent

Tableau: 74.2 - Capabilité, Niveau de qualité Sigma et P.P.M. dans procédé centré

où toutes les données sont obtenues à l'aide des relations suivantes à partir de l'indice de capabilité potentielle uniquement:

$$\begin{aligned}
 CPU &= \frac{USL - (T + \mu)}{3\sigma} = \frac{USL - (T + 0)}{3\sigma} = \frac{USL - 0}{3\sigma} = \frac{USL}{3\sigma} = \frac{USL}{3} \\
 CPL &= \frac{(T + \mu) - LSL}{3\sigma} = \frac{(T + 0) - LSL}{3\sigma} = \frac{0 - LSL}{3\sigma} = -\frac{LSL}{3}
 \end{aligned} \quad (74.78)$$

si l'écart-type est réduit (ce qui peut toujours être fait et ne change point la justesse des résultats!). Et puisque dans le tableau ci-dessus LSL et USL sont symétriques par rapport à la cible:

$$C_{pk} = C_p = \min\{CPU, CPL\} \Rightarrow LSL = -USL = -3C_p \quad (74.79)$$

et les PPM sont conformément à ce que nous avons vu juste avant donnés par:

$$PPM_{Tot} = \left(1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{USL} e^{-\frac{x^2}{2}} dx \right) \cdot 10^6 + \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{LSL} e^{-\frac{x^2}{2}} dx \right) \cdot 10^6 \quad (74.80)$$

et donc puisque dans l'exemple ci-dessus LSL et USL sont symétrique par rapport à la cible cela se simplifie en:

$$PPM_{Tot} = \left(1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{3C_p} e^{-\frac{x^2}{2}} dx \right) \cdot 10^6 \cdot 2 \quad (74.81)$$

où, par exemple, la valeur du *PPM* donnée à la ligne "Limite" est obtenue avec Maple 4.00b à l'aide de la commande:

```
>evalf((1-1/sqrt(2*Pi)*int(exp(-x^2/2),x=-infinity..3.9))*2)*1E6;
```

ou avec la version anglaise de Microsoft Excel 11.8346:

```
=(1-NORMDIST(3.9;0;1;1))*1E6
```

Rappelons que le "**niveau de qualité sigma**" noté σ_q est au fait donné à l'aide du tableau suivant que nous avons construit dans le chapitre de Statistiques:

Niveau de qualité Sigma	Taux de non-défection assuré en %	Taux de défection en parties par million
$1\sigma_q$	68.26894	317'311
$2\sigma_q$	95.4499	45'500
$3\sigma_q$	99.73002	2'700
$4\sigma_q$	99.99366	63.4
$5\sigma_q$	99.999943	0.57
$6\sigma_q$	99.9999998	0.002

Tableau: 74.3 - Capabilité, Taux de non-défection en % et PPM

et pour lequel nous avons donné la commande Maple 4.00b pour obtenir les valeurs qui sont valables pour tout écart-type et toute espérance!

2. Maintenant construisons le tableau au pire selon Six Sigma, soit un tableau en procédé non centré (c'est-à-dire où $C_p = C_{pk}$ n'est pas satisfait) avec une déviation de la moyenne de $+1.5\sigma$ (donc à droite mais on pourrait prendre à gauche et les résultats seraient les mêmes) par rapport à la cible et d'écart-type unitaire avec *USL* et *LSL* symétriques (ce qui restreint toujours le champ d'application):

C_p	C_{pk}	Défauts (PPM)	Niveau de qualité Sigma	Critère
0.5	0	501350	1.5	Mauvais
0.6	0.1	382572	1.8	
0.7	0.2	27412	2.1	
0.8	0.3	184108	2.4	
0.9	0.4	115083	2.7	
1	0.5	66810	3	
1.1	0.6	35931	3.3	
1.2	0.7	17865	3.6	
1.3	0.8	8198	3.9	Limite
1.4	0.9	3467	4.2	
1.5	1	1350	4.5	
1.6	1.1	483	4.8	
1.7	1.2	159	5.1	
1.8	1.3	48	5.4	
1.9	1.4	13	5.7	
2	1.5	3.4	6	Excellent

Tableau: 74.4 - Capabilité, Niveau de qualité Sigma et P.P.M. dans procédé décentré

où toutes les données sont obtenues à l'aide des relations suivantes à partir de l'indice de capabilité potentielle uniquement:

$$\begin{aligned}
 CPU &= \frac{USL - (T + 1.5\sigma)}{3\sigma} = \frac{USL - T}{3\sigma} - 0.5 = \frac{USL}{3} - 0.5 = C_p - 0.5 \\
 CPL &= \frac{(T + 1.5\sigma) - LSL}{3\sigma} = \frac{T - LSL}{3\sigma} + 0.5 = -\frac{LSL}{3} + 0.5 = C_p + 0.5
 \end{aligned}
 \tag{74.82}$$

et donc:

$$C_{pk} = \min\{CPU, CPL\} = C_p - 0.5 \tag{74.83}$$

d'où:

$$\begin{aligned}
 USL &= 3 \cdot C_p \\
 LSL &= -3 \cdot (C_p + 1)
 \end{aligned}
 \tag{74.84}$$

et les PPM sont conformément à ce que nous avons vu juste avant donnés par:

$$\begin{aligned}
 PPM_{Tot} &= \left(1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{USL} e^{-\frac{(x-\mu)^2}{2}} dx \right) \cdot 10^6 + \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{LSL} e^{-\frac{(x-\mu)^2}{2}} dx \right) \cdot 10^6 \\
 &= \left(1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{3C_p} e^{-\frac{(x-1.5\sigma)^2}{2}} dx \right) \cdot 10^6 + \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-3(C_p+1)} e^{-\frac{(x-1.5\sigma)^2}{2}} dx \right) \cdot 10^6 \quad (74.85) \\
 &= \left(1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{3C_p} e^{-\frac{(x-1.5)^2}{2}} dx \right) \cdot 10^6 + \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-3(C_p+1)} e^{-\frac{(x-1.5)^2}{2}} dx \right) \cdot 10^6
 \end{aligned}$$

où la ligne "Limite" du tableau précédent est par exemple obtenue avec Maple 4.00b à l'aide de la commande:

```
>evalf((1-1/sqrt(2*Pi)*int(exp(-(x-1.5)^2/2),x=-infinity..(1.3*3))))*1E6+evalf((1/sqrt(2*Pi)*int(exp(-(x-1.5)^2/2),x=-infinity..-(3*(1.3+1)))))*1E6;
```

ou avec la version anglaise de Microsoft Excel 11.8346:

```
=(((1-NORMDIST(3*1.3;1.5;1;1))+NORMDIST(-3*(1.3+1);1.5;1;1)))*1E6
```

On comprend enfin en voyant cette fameuse ligne "Limite", pourquoi un procédé sous-contrôle est dit "limite capable" avec un indice de capabilité potentielle de 1.33 étant donné le nombre de PPM!

Donc, le but dans la pratique c'est bien évidemment d'être dans la situation du premier tableau avec pour valeur correspondante dans ce premier tableau à un niveau de qualité sigma de $\sim 4.7\sigma$ pour avoir l'équivalent des 3.4 PPM du deuxième tableau (car il est plus facile de centrer un procédé que de contrôler ses écarts).

Toute l'importance des valeurs calculées ci-dessous est dans l'application de procédés de fabrication à n -étapes en série (considérés sous la dénomination de "processus"). Cette application sera présentée dans le chapitre sur les Techniques de Gestion.

Exemple:

Faisons un résumé de tout cela en considérant une nouvelle petite production de $n = 50$ pièces par lot de 10 (afin d'ajuster en cours de production). La mesure de côtes de 5 pièces chaque heure pendant 10 heures avec une tolérance de 10 ± 0.07 soit en termes de centièmes une étendue de:

$$E = 2 \cdot 7 = 14 \quad (74.86)$$

et une cible de $T = 0$ (en termes d'écarts). Nous avons les données suivantes:

	1	2	3	4	5	6	7	8	9	10
1	-2	-4	-1	0	4	0	3	0	1	-1
2	0	-3	0	-2	1	-2	0	1	-1	2
3	-1	0	-3	-1	0	0	-1	-1	3	1
4	1	1	-2	2	2	0	1	0	4	0
5	-1	-1	-3	0	0	3	3	2	1	0
μ	-0.6	-1.4	-1.8	-0.2	1.4	0.2	1.2	0.4	1.6	0.5
σ	1.14	2.07	1.30	1.48	1.67	1.79	1.79	1.14	1.95	1.14

Tableau: 74.5 - Application d'analyse de maîtrise statistique des procédés

Nous voyons immédiatement que le processus de fabrication a été non stationnaire pendant cette première production, il faudra donc apporter des corrections à l'avenir:

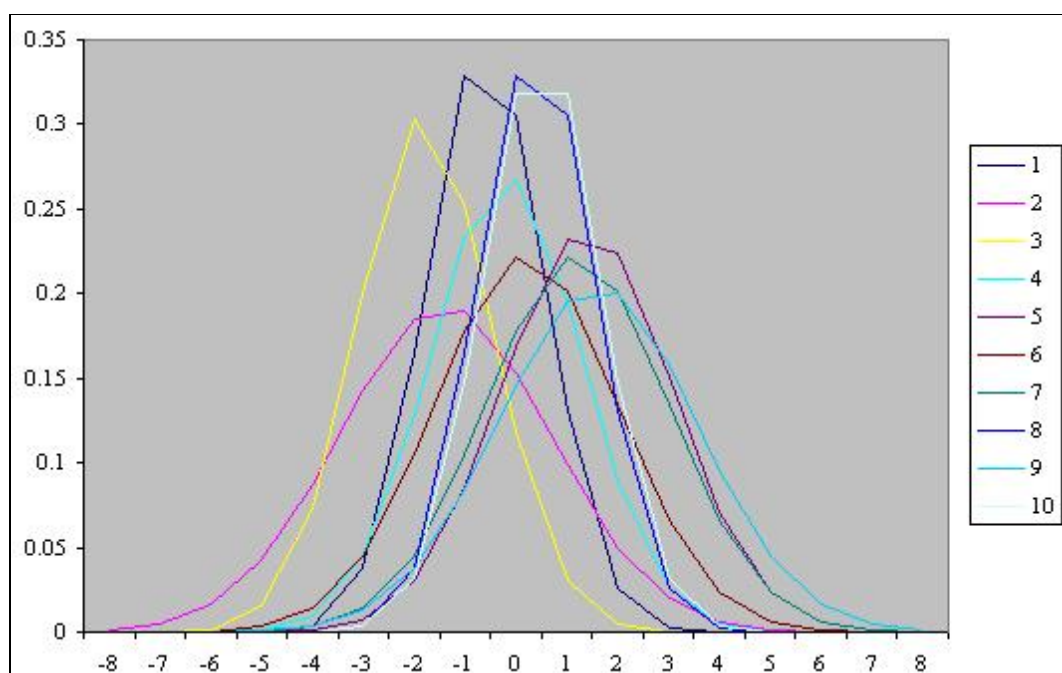


Figure: 74.7 - Preuve par un tracé que le processus est non stationnaire

ou sous forme de carte de contrôle (comme je les aime) avec la représentation d'un écart-type de 3σ (ce qui est suffisant pour des petites quantités des pièces bon marché à fabriquer):

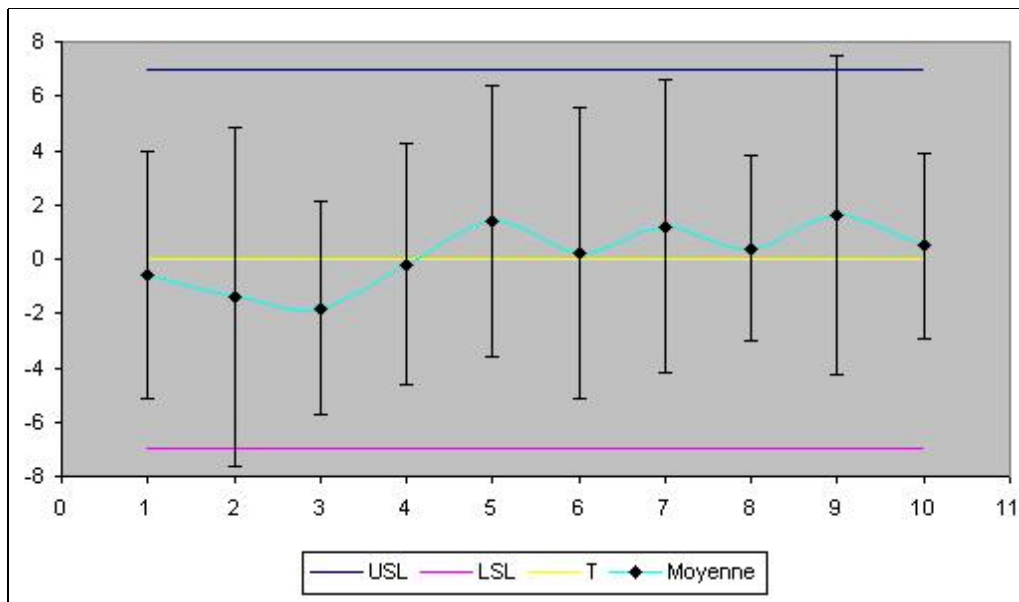


Figure: 74.8 - Petite carte de contrôle empirique

Donc, on devine quand même que le processus est limite...

Remarque: Une chose intéressante c'est que l'on peut analyser aussi ce graphique en utilisant les outils mathématiques de l'analyse des séries temporelles (cf. chapitre d'Économie).

D'abord, si nous voulons faire une étude statistique pertinente des différentes données ci-dessus nous pouvons calculer la moyenne générale des écarts qui sous l'hypothèse d'une distribution Normale est la moyenne arithmétique (cf. chapitre de Statistiques):

$$\bar{X} = 0.12 \quad (74.87)$$

Ensuite l'écart-type des données de toutes les pièces est de:

$$S_{LT} = 1.80 \quad (74.88)$$

en utilisant l'estimateur de maximum de vraisemblance de la variance de la loi Normale:

$$S_{LT} = \sqrt{\frac{\sum_{i,j} (X_{ij} - \bar{X})^2}{n}} \quad (74.89)$$

nous avons une valeur supérieur à 1 ce qui est donc non-conforme à ce que Six Sigma exige dans son niveau de qualité.

Donc, l'erreur-standard (l'estimateur de l'écart-type de la moyenne) est de:

$$\sigma_{\mu} = \frac{S_{LT}}{\sqrt{n}} = 0.25 \quad (74.90)$$

Donc, l'intervalle de confiance à 95% de la moyenne est de (cf. chapitre de Statistiques):

$$\bar{X} - 1.96 \frac{S_{LT}}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{S_{LT}}{\sqrt{n}} \quad (74.91)$$

Soit dans notre cas:

$$-0.379 \leq \mu \leq 0.619 \quad (74.92)$$

Et l'inférence statistique avec notre écart-type long terme utilisant le test d'hypothèse bilatéral du χ^2 donne (cf. chapitre de Statistiques):

$$\frac{(n-1)S_{LT}^2}{\chi_{\alpha/2}^2(n-1)} \leq \sigma_{LT}^2 \leq \frac{(n-1)S_{LT}^2}{\chi_{1-\alpha/2}^2(n-1)} \quad (74.93)$$

Ce qui nous donne dans notre cas:

$$2.268 \leq \sigma_{LT}^2 \leq 5.047 \quad (74.94)$$

soit:

$$1.506 \leq \sigma_{LT} \leq 2.246 \quad (74.95)$$

Nous remarquons alors que sur une analyse long terme, nous avons les intervalles:

$$\begin{aligned} -0.644 &\leq \mu \leq 0.884 \\ 1.506 &\leq \sigma_{LT} \leq 2.246 \end{aligned} \quad (74.96)$$

Calculons maintenant la performance globale du procédé long terme (si supposé centré donc!). Nous avons:

$$P_p = \frac{USL - LSL}{6S_{LT}} = \frac{E}{6S_{LT}} = \frac{14}{6 \cdot 1.80} = 1.294 \quad (74.97)$$

Mais avec un instrument ayant un P_{pc} de 4, cela correspond réellement à:

$$P_{p,vrai} = \frac{P_{p,observé}}{\sqrt{1 - \frac{P_{p,observé}}{P_{pc}}}} = 1.574 \quad (74.98)$$

De plus, indiquons que comme nous savons faire un calcul d'intervalle de confiance pour σ_{LT} (voir le calcul fait précédemment), il est alors aisé d'en avoir un pour P_p aussi!

Si l'analyse de la performance globale du procédé long terme est non centrée (ce qui est le cas ici) nous utilisons donc:

$$P_{pk} = \min \left\{ \frac{USL - \bar{X}}{3S_{LT}}, \frac{\bar{X} - LSL}{3S_{LT}} \right\} = \min \left\{ \frac{7 - 0.12}{3 \cdot 1.80}, \frac{0.12 - (-7)}{3 \cdot 1.80} \right\} = \min \{1.271, 1.316\} = 1.271 \quad (74.99)$$

et nous savons encore une fois qu'à cause de l'instrument, cette valeur est un peu sous-évaluée! Nous avons bien évidemment:

$$P_{pk} \neq P_p \quad (74.100)$$

donc le processus n'est pas centré (on s'en doutait...). Alors, il faut calculer la capacité potentielle décentrée moyenne long terme du procédé P_{pm} selon les relations déterminées plus haut:

$$P_{pm} \cong 1.291 \quad (74.101)$$

Bref, que ce soit de la valeur de P_{pk} , P_p ou P_{pm} , nous voyons que les valeurs sont toutes limites capables (c'est-à-dire que la valeur est supérieure à 1 - voir définition plus haut pour un rappel de ce que signifie "limite capable").

Si nous faisons alors nos calculs de PPM selon les relations obtenues plus haut avec la valeur de S_{LT} et de $\bar{X}$ obtenues, nous avons alors:

$$PPM_{tot}^{LT} = PPM_{UL} + PPM_{LL} \cong 114 \quad (74.102)$$

Ensuite, dire que ce chiffre est bon ou mauvais est difficile, car il nous manque l'information de savoir quel est le coût de production, le coût de revient et de réparation d'un produit et le tout est lui-même dépendant de la quantité totale fabriquée! Mais nous pouvons utiliser aussi le modèle de Taguchi pour connaître la valeur des paramètres (moments) calculés qu'il serait préférable de ne pas dépasser!

Calculons maintenant les indices de capacité court terme! Pour cela, il nous faut l'estimateur de la moyenne de l'ensemble en considérant chaque individu comme une variable aléatoire. Nous savons (cf. chapitre de Statistiques) que cette moyenne est aussi la moyenne arithmétique dans le cas d'une loi Normale et elle est strictement égale à celle que l'on calcule en considérant l'ensemble des individus comme une seule et unique variable aléatoire. Donc, il vient que:

$$\bar{X} = 0.12 \quad (74.103)$$

En ce qui concerne l'écart-type par contre ce n'est pas pareil. Mais nous savons (cf. chapitre de Statistiques) que la loi Normal est stable par la somme. Par exemple, nous avons démontré qu'étant données deux variables aléatoires indépendantes et distribuées selon une loi Normale (en imaginant que chaque variable représente deux de nos dix échantillons), nous avons pour leur somme.:

$$X+Y = N(\mu_{X+Y}, \sigma_{X+Y}) = N(\mu_X + \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2}) \quad (74.104)$$

Or nous avons aussi démontré dans le chapitre de Statistiques que de par la propriété de linéarité de l'espérance, nous avons:

$$E(Z) = E\left(\frac{X+Y}{2}\right) = \frac{1}{2} E(X+Y) = \frac{1}{2}(\mu_X + \mu_Y) = \frac{\mu_X + \mu_Y}{2} \quad (74.105)$$

ce qui est conforme à notre remarque précédente pour la variance:

$$V(Z) = V\left(\frac{X+Y}{2}\right) = \frac{1}{4}V(X+Y) = \frac{1}{4}\sqrt{\sigma_X^2 + \sigma_Y^2} = \sqrt{\frac{\sigma_X^2 + \sigma_Y^2}{2}} \quad (74.106)$$

Donc in extenso:

$$S_{CT} = \sqrt{\frac{\sum_{i=1}^k S_i^2}{k}} \quad (74.107)$$

et dans notre cas particulier:

$$\sigma_{CT} = \sqrt{\frac{\sum_{i=1}^5 \sigma_i^2}{5}} = 1.583 \quad (74.108)$$

nous avons une valeur supérieur à 1, ce qui est donc non-conforme à ce que Six Sigma exige dans son niveau de qualité.

Donc l'erreur-standard (l'estimateur de l'écart-type de la moyenne) est de:

$$\sigma_{\mu} = \frac{S_{CT}}{\sqrt{k}} = 0.707 \quad (74.109)$$

Donc l'intervalle de confiance à 95% de la moyenne est de (cf. chapitre de Statistiques):

$$\bar{X} - 1.96 \frac{S_{CT}}{\sqrt{k}} \leq \mu \leq \bar{X} + 1.96 \frac{S_{CT}}{\sqrt{k}} \quad (74.110)$$

Soit dans notre cas:

$$-1.267 \leq \mu \leq 1.507 \quad (74.111)$$

Nous remarquons donc qu'en court terme, l'intervalle est beaucoup plus large qu'en long terme, ce qui est normal étant donné la faible valeur de k (qui vaut donc 5 dans notre exemple).

Et l'inférence statistique avec notre écart-type long terme utilisant le test d'hypothèse bilatéral du χ^2 donne (cf. chapitre de Statistiques):

$$\frac{(n-1)S_{CT}^2}{\chi_{\alpha/2}(n-1)} \leq \sigma_{CT}^2 \leq \frac{(n-1)S_{CT}^2}{\chi_{1-\alpha/2}(n-1)} \quad (74.112)$$

Ce qui nous donne dans notre cas:

$$0.899 \leq \sigma_{CT}^2 \leq 20.692 \quad (74.113)$$

soit:

$$0.948 \leq \sigma_{CT} \leq 4.549 \quad (74.114)$$

Nous remarquons alors que sur une analyse long terme avec les intervalles:

$$\begin{aligned} -1.267 \leq \mu \leq 1.507 \\ 0.948 \leq \sigma_{CT} \leq 4.549 \end{aligned} \quad (74.115)$$

Les variations peuvent donc être énormes avec une probabilité cumulée de 95% et il faudra prendre garde dans un cas pratique d'apporter des réglages au plus vite afin de diminuer au maximum les moments!

Calculons maintenant la capabilité potentielle du procédé court terme (si supposé centré donc!). Nous avons:

$$C_p = \frac{USL - LSL}{6\sigma_{CT}} = \frac{E}{6\sigma_{CT}} = \frac{14}{6 \cdot 1.58} = 1.473 \quad (74.116)$$

Donc, nous avons:

$$C_p > P_p \quad (74.117)$$

ce qui est normal, car si les mesures que nous avons étaiement vraiment faites sur une longue période alors ce serait très problématique alors que sur une courte période, c'est déjà un peu plus normal. D'où cette relation d'ordre entre les deux indices.

Mais avec un instrument ayant un C_{pc} de 4, cela correspond réellement à:

$$C_{p,vrai} = \frac{C_{p,observé}}{\sqrt{1 - \frac{C_{p,observé}}{C_{pc}}}} = 1.854 \quad (74.118)$$

De plus, indiquons que comme nous savons faire un calcul d'intervalle de confiance pour σ_{CT} (voir le calcul fait précédemment), il est alors aisé d'en avoir un pour C_p aussi!

Si l'analyse de la capabilité potentielle du procédé court terme est non centrée (ce qui est le cas ici) nous utilisons donc:

$$C_{pk} = \min \left\{ \frac{USL - \bar{X}}{3\sigma_{CT}}, \frac{\bar{X} - LSL}{3\sigma_{CT}} \right\} = \min \left\{ \frac{7 - 0.12}{3 \cdot 1.583}, \frac{0.12 - (-7)}{3 \cdot 1.583} \right\} = \min \{1.44, 1.49\} = 1.44 \quad (74.119)$$

et nous savons encore une fois qu'à cause de l'instrument, cette valeur est un peu sous-évaluée! Nous avons bien évidemment:

$$C_{pk} \neq C_p \quad (74.120)$$

donc le processus n'est pas centré (on s'en doutait...). Alors, il faut calculer la capabilité potentielle décentrée moyenne court terme du procédé C_{pm} selon les relations déterminées plus haut:

$$C_{pm} = \frac{USL - LSL}{6\sigma} = \frac{7 - (-7)}{6\sigma_{LT}} = \frac{15}{6 \cdot 1.8} \cong 1.296 \quad (74.121)$$

Bref, que ce soit de la valeur de C_{pk} , C_p ou C_{pm} , nous voyons que les valeurs sont toutes limites capables.

Si nous faisons alors nos calculs de PPM selon les relations obtenues plus haut avec la valeur de S_{CT} et de $\bar{X}$ obtenues, nous avons alors:

$$PPM_{tot}^{CT} = PPM_{UL} + PPM_{LL} \cong 10 \quad (74.122)$$

Ensuite, dire que ce chiffre est bon ou mauvais est difficile, car il nous manque l'information de savoir quel est le coût de production, le coût de revient et de réparation d'un produit et le tout est lui-même dépendant de la quantité totale fabriquée! Mais nous pouvons utiliser aussi le modèle de Taguchi pour connaître la valeur des paramètres (moments) calculés qu'il serait préférable de ne pas dépasser!

Pour clore cette partie, voici la sortie d'un logiciel comme Minitab 15.1.1 dans lequel nous retrouvons tous les calculs effectués ci-dessus plus des cartes de contrôle dont nous ferons les démonstrations détaillées plus loin (se référer au chapitre de Méthodes Numériques pour les test AD d'Anderson-Darling):

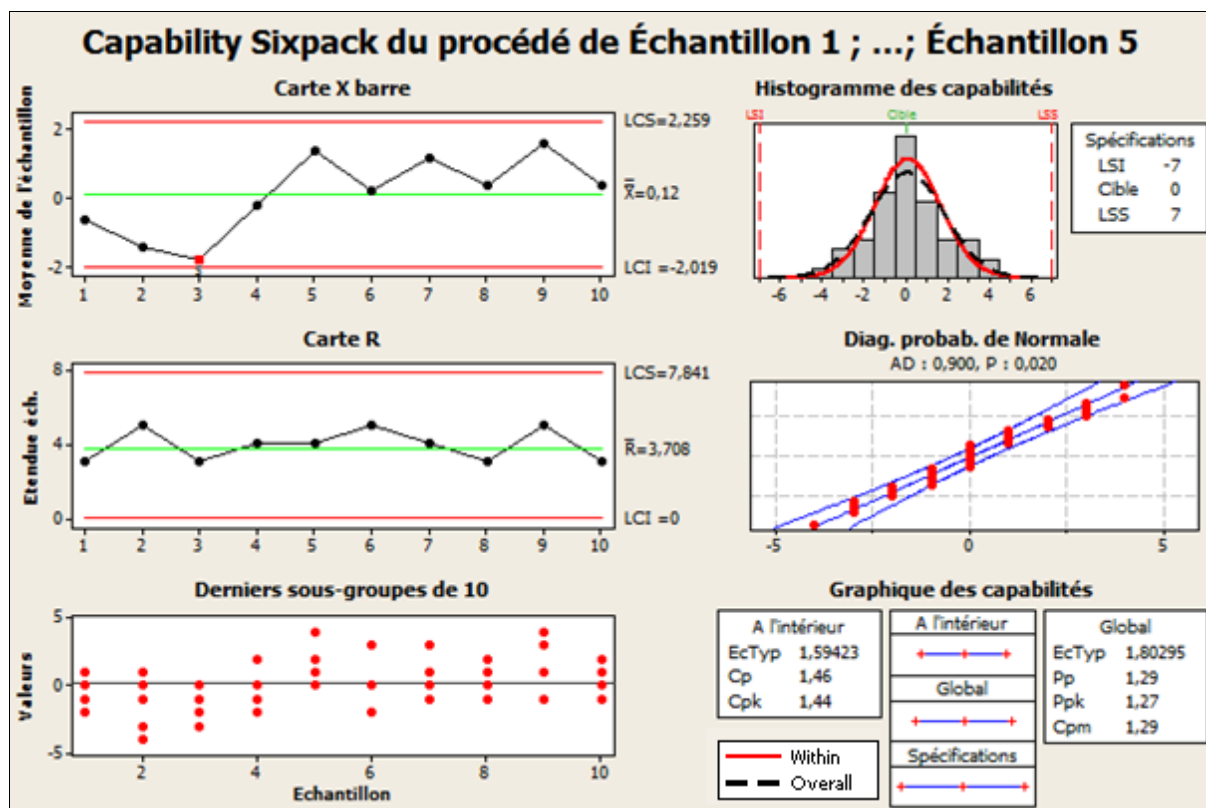


Figure: 74.9 - Analyse de capabilité Six Pack de Minitab 15.1.1

1.5. MODÈLE DE TAGUCHI

Dans le cadre des SPC (Statistical Process Control), il est intéressant pour un industriel d'estimer les pertes financières générées par les écarts à la cible (attention on peut appliquer également cette approche dans d'autres domaines que l'industrie!)

Nous pouvons avoir une estimation relativement simple et satisfaisante de ses pertes (coûts) sous les hypothèses suivantes:

H1. Le processus est sous contrôle (écart-type constant) et suit une loi de densité symétrique décroissante à gauche et à droite par rapport à la cible (qui peut être une côte, un nombre d'erreurs par périodes, etc.)

H2. Le coût est nul lorsque la production (ou le travail) est centrée sur la cible (minimum).

H3. Le coût augmente de manière identique lorsque la production se décentre sur la gauche et sur la droite (ce qui n'est pas le cas dans le domaine de l'administration par exemple). La fonction de coût passe donc selon H2 et H3 par un minimum sur la cible.

Dès lors, si nous notons Y le décentrage par rapport à la cible T et L la perte financière ("loss" en anglais d'où le L). Nous avons:

$$L = f(Y) \quad (74.123)$$

Même si nous ne connaissons pas la forme de cette fonction, nous pouvons l'écrire sous forme de développement de Taylor (cf. chapitre de Suites et Séries) autour de T tel que :

$$L_n(Y) = \sum_{k=0}^n \frac{1}{k!} \frac{d^k L(T)}{dY^k} (Y-T)^k \quad (74.124)$$

Si nous développons au troisième ordre:

$$\begin{aligned} L_4(Y) &\cong \frac{1}{0!} \frac{d^0 L(T)}{dY^0} (Y-T)^0 + \frac{1}{1!} \frac{d^1 L(T)}{dY^1} (Y-T)^1 + \frac{1}{2!} \frac{d^2 L(T)}{dY^2} (Y-T)^2 \\ &= L(T) + \frac{dL(T)}{dY^1} (Y-T) + \frac{1}{2!} \frac{d^2 L(T)}{dY^2} (Y-T)^2 \end{aligned} \quad (74.125)$$

Or par l'hypothèse H2, nous avons $L(T)$ qui est nul. Il reste alors:

$$L(Y) \cong \frac{dL(T)}{dY^1} (Y-T) + \frac{1}{2!} \frac{d^2 L(T)}{dY^2} (Y-T)^2 \quad (74.126)$$

et comme par H3, la dérivée de la fonction $L(Y)$ est nulle en T puisqu'il s'agit d'un minimum alors:

$$L(Y) \cong \frac{1}{2!} \frac{d^2 L(T)}{dY^2} (Y-T)^2 \quad (74.127)$$

Ce qui est noté en SPC:

$$L(Y) \cong k(Y-T)^2 \quad (74.128)$$

et est appelée "**fonction de perte de Taguchi (centrée)**" ou plus simplement "**fonction perte de qualité (centrée)**".

Bon c'est bien joli d'avoir cette relation mais comment doit-on l'utiliser?

Au fait, c'est relativement simple. Sous les hypothèses mentionnées plus haut, si nous avons en production des mesures de défauts (côtes, retards, pannes, bug, etc.) alors il suffit de calculer leur moyenne arithmétique μ (estimateur de la moyenne d'une loi Normale) et ensuite de savoir le coût financier ou horaire L que cela a engendré pour l'entreprise ou l'institution (parfois cette moyenne est calculée sur la base d'un unique échantillon...).

Donc, la relation précédente devient:

$$L \cong k(\mu - T)^2 \quad (74.129)$$

avec L et μ connus.

Et comme T est donné par les exigences du client ou du contexte alors il est aisé d'obtenir le facteur k :

$$k \cong \frac{L}{(\mu - T)^2} \quad (74.130)$$

qui est au fait mathématiquement parlant le point d'inflexion de la fonction mathématique L .

Cette dernière relation est parfois notée:

$$k \cong \frac{\Delta}{m^2} \quad (74.131)$$

Une fois que nous avons k avec une bonne estimation, il est possible de connaître L pour toute valeur Y et ainsi nous pouvons calculer en production le coût d'une déviation quelconque par rapport à la cible.

Exemple:

Considérons une alimentation pour une chaîne stéréo pour laquelle T vaut 110 [V]. Si la tension sort des 110 ± 20 [V] alors la stéréo tombe en panne et doit être réparée. Supposons que le coût de réparation est (tous frais directs et indirects compris!) de 100.-. Alors, le coût associé pour une valeur donnée de la tension est:

$$L(y) = \frac{100}{20^2} (y - 110)^2 = 0.25 (y - 110)^2 \quad (74.132)$$

Voyons maintenant une manière élégante de calculer le coût moyen de Taguchi (perte unitaire moyenne). Nous avons bien évidemment dans une chaîne de production sur plusieurs pièces d'une même famille:

$$\begin{aligned} \bar{L} &= [L(y_1) + L(y_2) + \dots + L(y_n)] = \frac{k}{n} [(y_1 - T)^2 + (y_2 - T)^2 + \dots + (y_n - T)^2] \\ &= \frac{k}{n} [X_1^2 + X_2^2 + \dots + X_n^2] \end{aligned} \quad (74.133)$$

où les X_i^2 sont des variables aléatoires normales (gaussiennes par hypothèse car le procédé de fabrication est sous contrôle statistiques). Or, nous avons démontré dans le chapitre de Statistiques lors de notre étude de l'intervalle de confiance sur la variance avec moyenne empirique connue que:

$$\chi^2(n) = \sum_{i=1}^n (Y_i - \mu)^2 = \frac{1}{\sigma^2} \left[\sum_{i=1}^n (Y_i - \bar{Y})^2 + n(\bar{Y} - \mu)^2 \right] \quad (74.134)$$

Donc:

$$\sigma^2 \chi^2(n) = \sum_{i=1}^n (Y_i - \bar{Y})^2 + n(\bar{Y} - \mu)^2 \quad (74.135)$$

Dès lors:

$$\begin{aligned} \bar{L} = \sigma^2 \chi^2(n) &= \frac{k}{n} [X_1^2 + X_2^2 + \dots + X_n^2] = \frac{k}{n} \left[\sum_{i=1}^n (Y_i - \bar{Y})^2 + n(\bar{Y} - \mu)^2 \right] \\ &= k \left[\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2 + \frac{1}{n} n(\bar{Y} - \mu)^2 \right] = k \left[\hat{\sigma}^2 + (\bar{Y} - T)^2 \right] \end{aligned} \quad (74.136)$$

Cette dernière expression présente l'avantage de montrer très clairement que pour minimiser la perte, il faut agir sur la dispersion et l'ajustement de la moyenne sur la valeur nominale.

Or nous avons démontré dans le chapitre de Statistiques que (il est important dans les présents développements que nous utilisons les notations qui distinguent les différents estimateurs!):

$$\hat{\sigma}^2 = \frac{n-1}{n} \sigma^2 \quad (74.137)$$

Donc:

$$\bar{L} = k \left[\frac{n-1}{n} \sigma^2 + (\bar{Y} - T)^2 \right] = k \left[\frac{n-1}{n} \sigma^2 + (\hat{\mu} - T)^2 \right] \quad (74.138)$$

Et si n est grand, nous avons alors pour un lot de produits:

$$\boxed{\bar{L} = k \left[\sigma^2 + (\hat{\mu} - T)^2 \right]} \quad (74.139)$$

où le premier terme entre crochets représente donc la variance de Y autour de sa propre moyenne et le deuxième terme la déviation de Y par rapport à la cible T .

Indiquons que l'on retrouve cette dernière relation dans la littérature et dans les logiciels souvent sous la forme suivante:

$$L = k \left[S^2 + (\bar{y} - y_N)^2 \right] \quad (74.140)$$

où l'indice N signifie "nominal". Évidemment, quand la cible T (ou la valeur nominale) est prise comme étant nulle, les relations se simplifient encore d'avantage.

Maintenant, rappelons les propriétés suivantes de l'espérance et de la variance (cf. chapitre de Statistiques):

$$\begin{aligned} E(\alpha X) &= \alpha E(X) \\ \sigma_{\alpha X} &= \alpha \sigma_X \end{aligned} \quad (74.141)$$

Dès, lors si le processus est décentré et que nous devons corriger la variable aléatoire pour la recentrer, il vient intuitivement que ce facteur sera alors:

$$\alpha = \frac{y_N}{\bar{y}} \quad (74.142)$$

Dès lors, il vient immédiatement:

$$L = k \left[S^2 \left(\frac{y_N}{\bar{y}} \right)^2 + \left(\bar{y} \frac{y_N}{\bar{y}} - y_N \right)^2 \right] = k y_N^2 \left(\frac{S^2}{\bar{y}^2} \right) \quad (74.143)$$

Donc pour minimiser L , sachant que les autres termes sont imposés, il faut minimiser le rapport:

$$\frac{S^2}{\bar{y}^2} \quad (74.144)$$

ou, ce qui revient au même, maximiser le rapport inverse (qui reste sans dimensions):

$$\frac{\bar{y}^2}{S^2} \quad (74.145)$$

Ce nombre étant souvent très grand, il est d'usage dans la littérature et dans les logiciel de statistiques d'en prendre le logarithme en base 10 et de multiplier par 10 (c'est une idée inspirée de la physique acoustique). Nous avons alors ce que nous appelons le rapport signal/bruit (en anglais: signal/noise), donné en décibels par:

$$S / N = 10 \log_{10} \left(\frac{\bar{y}^2}{S^2} \right) \quad [dB] \quad (74.146)$$

pour ce que les ingénieurs appellent "**le nominal est le meilleur**".

Relation que l'on retrouve telle quelle dans un logiciel comme Minitab lors de l'analyse de plans d'expérience de Taguchi (voir plus loin ce que sont les plans d'expérience). Certains très rares logiciels proposent également le cas particulier suivant:

$$S / N = 10 \log_{10} \left(\frac{\bar{y}^2}{S^2} \right) = 10 \log_{10} (\bar{y}^2) - 10 \log_{10} (S^2) \underset{\bar{y}=1}{=} -10 \log_{10} (S^2) \quad (74.147)$$

Pour la suite revenons à la relation:

$$\begin{aligned} \bar{L} &= [L(y_1) + L(y_2) + \dots + L(y_n)] = \frac{k}{n} [(y_1 - T)^2 + (y_2 - T)^2 + \dots + (y_n - T)^2] \\ &= \frac{k}{n} [X_1^2 + X_2^2 + \dots + X_n^2] \end{aligned} \quad (74.148)$$

Les ingénieurs parlent de minimiser la fonction perte de qualité quand T (la cible) doit être nul et donc qu'il faut minimiser tous les y_i . Donc la relation précédente se réduit finalement à:

$$\bar{L} = \frac{k}{n} \sum_i y_i^2 \quad (74.149)$$

Ce que les ingénieurs aiment parfois bien résumer sous la forme suivante (alors qu'à mon avis prendre le logarithme en base 10 ainsi que de mettre un signe "-" dans ce cas ne se justifie vraiment pas... bien au contraire):

$$-10 \log \left(\frac{1}{n} \sum_i y_i^2 \right) \quad (74.150)$$

et même parfois il disent que c'est quand même un rapport signal/bruit... alors que de toute évidence il n'y a aucun bruit dans ce cas particulier qui est appelé: "**le plus petit est le meilleur**" (PPEM). C'est exactement sous cette forme que l'on peut trouver cet objectif écrit dans le logiciel Minitab.

Dans le cadre d'un paramètre à maximiser il est d'usage de dire qu'on cherche alors plutôt à minimiser:

$$\bar{L} = \frac{k}{n} \left[\left(\frac{1}{y_1} - \frac{1}{T} \right)^2 + \left(\frac{1}{y_2} - \frac{1}{T} \right)^2 + \dots + \left(\frac{1}{y_n} - \frac{1}{T} \right)^2 \right] \quad (74.151)$$

car vu que T tend vers l'infini il serait difficile de faire quoi que ce soit mathématiquement parlant.

Donc justement vu que T tend vers l'infini, cette dernière relation se réduit à:

$$\bar{L} = \frac{k}{n} \sum_i \frac{1}{y_i^2} \quad (74.152)$$

Ce que les ingénieurs aiment bien résumer parfois sous la forme suivante (ici le logarithme en base 10 se justifie!):

$$-10 \log \left(\frac{1}{n} \sum_i \frac{1}{y_i^2} \right) \quad (74.153)$$

et même parfois il disent que c'est quand même un rapport signal/bruit... alors que de toute évidence il n'y a aucun bruit dans ce cas particulier qui est appelé: "**le plus grand est le meilleur**" (PGLM). C'est exactement sous cette forme que l'on peut trouver cet objectif écrit dans le logiciel Minitab.

Il est cependant, en étant malin, possible d'introduire le bruit dans la relation:

$$\frac{1}{n} \sum_i \frac{1}{y_i^2} \quad (74.154)$$

Pour cela écrivons:

$$\frac{1}{n} \sum_i \frac{1}{y_i^2} = \frac{1}{n} \sum_i \frac{1}{(\mu + y_i - \mu)^2} = \frac{1}{n} \sum_i \frac{1}{\left(1 + \frac{y_i - \mu}{\mu}\right)^2 \mu^2} = \frac{1}{n\mu^2} \sum_i \frac{1}{\left(1 + \frac{y_i - \mu}{\mu}\right)^2} \quad (74.155)$$

Posons:

$$x = \frac{y_i - \mu}{\mu} < 1 \quad (74.156)$$

Alors dans ce cas, en utilisant la série de Taylor démontrée dans le chapitre de Suites Et Séries:

$$\frac{1}{(1+x)^2} \cong 1 - 2x + 3x^2 \quad (74.157)$$

Nous avons alors:

$$\frac{1}{n\mu^2} \sum_i \frac{1}{\left(1 + \frac{y_i - \mu}{\mu}\right)^2} \cong \frac{1}{n\mu^2} \sum_i \left(1 - 2\frac{y_i - \mu}{\mu} + 3\left(\frac{y_i - \mu}{\mu}\right)^2\right) \quad (74.158)$$

et sous contrôle statistique nous avons:

$$\sum_i \frac{y_i - \mu}{\mu} \rightarrow 0 \quad (74.159)$$

et en supposant que nous travaillons sur l'ensemble de la population de pièces, nous avons:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \mu)^2 \quad (74.160)$$

Il vient alors:

$$\frac{1}{n\mu^2} \sum_i \left(1 - 3\left(\frac{y_i - \mu}{\mu}\right)^2\right) = \frac{1}{n\mu^2} \left(n - 3\frac{n\sigma^2}{\mu^2}\right) = \frac{1}{\mu^2} \left(1 - 3\frac{\sigma^2}{\mu^2}\right) \quad (74.161)$$

Donc au final en remplaçant un peu abusivement par les estimateurs, il vient:

$$S / N = -10 \log_{10} \left(\frac{1}{\bar{y}^2} \left(1 + 3 \frac{S^2}{\bar{y}^2} \right) \right) \quad (74.162)$$

2. MAINTENANCE PRÉVENTIVE

L'évolution des techniques de production vers une plus grande robotisation des systèmes techniques plus complexes a augmenté l'importance de la fiabilité des machines de production. Aussi, un arrêt imprévu coûte cher à une entreprise. De même, dans l'industrie aéronautique et spatiale, les problèmes de fiabilité, de maintenabilité, de disponibilité sont capitaux. La maintenance garantit le niveau de fiabilité pour l'ensemble des composantes (mécaniques, électromécaniques et informatiques).

L'existence d'un service de maintenance a pour raison le maintien des équipements (systèmes) et aussi la diminution des pannes. En effet, ces dernières coûtent cher, elles occasionnent:

- Des coûts d'intervention, de réparation
- Des coûts de non-qualité du produit
- Des coûts indirects tels que des frais fixes, pertes de production, la marge bénéficiaire perdue...

De ce fait, il faut tout mettre en oeuvre pour éviter la panne, agir rapidement lorsqu'elle survient afin d'augmenter la disponibilité du matériel. Pour ce faire, il faut modéliser la vie des équipements. L'ensemble des méthodes et techniques relatives à ses problématiques sont habituellement classifiées sous le nom de "**Analyse des Modes de Défaillance, des Effets et de leur Criticité**" AMDEC.

Nous distinguons principalement deux classes de systèmes: les systèmes non réparables (satellites, bien de consommations à faibles coûts, etc.) et les systèmes réparables (machines de production, moyens de transports, etc.) où les approches théoriques sont différentes. Pour la deuxième catégorie, il est possible d'utiliser aussi les chaînes de Markov (où les états représenteront le nombre de composants fonctionnels ou en panne d'un système conformément à la norme ISO 31010), les réseaux de Petri ou la simulation par Monte-Carlo.

L'idée est dans les textes qui vont suivre de faire un petit point sur ces méthodes, d'en rechercher l'efficacité et de permettre aux praticiens ingénieurs ou techniciens de mieux appréhender ces problèmes. Une large place sera faite au modèle de Weibull, qui a une implication importante dans le domaine.

2.1. OBSOLESCENCE PROGRAMMÉE

L'obsolescence programmée va à l'encontre de la maintenance préventive dont le but premier est de fournir un produit de qualité et respectueux de son environnement en connaissance précise de la durée de vie du produit et du risque quantifiable de désuétude.

L'idée de l'obsolescence programmée est de réduire la durée de vie ou d'utilisation d'un produit afin d'en augmenter le taux de remplacement par le consommateur (stratégie parfois motivée par les politiques...). La demande ainsi induite aurait pour but (non démontré sur le très long terme...) de profiter au producteur, ou à ses concurrents. Le secteur bénéficie alors d'une production plus importante, stimulant les gains de productivité (économies d'échelle) et le progrès technique (qui accélère l'obsolescence des produits antérieurs) et maintient l'emploi d'une démographie toujours croissante est qui est donc problématique par définition.

Le lecteur aura donc compris que je suis farouchement contre ces méthodes et que tout ingénieur et tout scientifique travaillant pour une entreprise se doit de refuser par déontologie ce type de demande par la hiérarchie quitte à dénoncer ces méthodes anonymement sur Internet ou des associations de consommateurs.

2.2. ESTIMATEURS EMPIRIQUES

Dans le cadre de l'étude de fiabilité non accélérée (le vieillissement accéléré est un sujet trop complexe pour être abordé sur ce site), nous sommes amenés à définir certaines variables dont voici la liste:

- N_0 sera le nombre d'éléments bons à t_0 (instant initial)
- N_i le nombre d'éléments bons à t_i
- n_i le nombre d'éléments défaillant entre t_i et t_{i+1} noté aussi ΔN_i
- Δt_i l'intervalle de temps observé égal à $t_{i+1} - t_i$.

Définitions:

D1. Nous définissons le "**taux de défaillance par tranche temporelle**" Δt par la relation:

$$\hat{\lambda}(t_i) = \frac{n_i}{N_i \Delta t_i} \quad (74.163)$$

qui s'interprète donc comme étant le nombre d'éléments défectueux par rapport au nombre d'éléments survivants sur une tranche de temps donnée (il s'agit donc d'un pourcentage de non-conformités relatives par tranche temporelle).

Cette dernière relation est aussi parfois appelée "**hazard ratio**" (HR) ou "**survie relative**" et souvent notée $\hat{h}(t_i)$.

D2. Nous définissons la "**fonction de défaillance**" par la relation (densité de probabilité de défaillances à l'instant t_i):

$$\hat{f}(t_i) \Delta t_i = \frac{n_i}{N_0} \quad (74.164)$$

en remarquant bien que le dénominateur n'est pas le même que celui qui définit le taux de défaillance par tranche!

Cette fonction s'interprète donc comme étant le % d'éléments défectueux dans la tranche de temps étudiée par rapport au nombre total d'éléments initialement testés. Il s'agit de l'indicateur qui intéresse le plus souvent l'ingénieur!

D3. Nous définissons naturellement la "**fonction de défaillance cumulée**" par:

$$\hat{F}(t_i) = \sum_i \hat{f}(t_i) \Delta t_i = \frac{\sum_i n_i}{N_0} = \frac{N_0 - N_i}{N_0} = 1 - \frac{N_i}{N_0} \quad (74.165)$$

que tend vers 1 lorsque le temps tend vers l'infini.

Cette fonction s'interprète donc comme étant le % d'éléments défectueux cumulé par rapport au nombre total d'éléments initialement testés.

D4. Nous définissons in extenso la "**fonction de fiabilité**" par (il s'agit du deuxième terme de la précédente relation):

$$\hat{R}(t_i) = \frac{N_i}{N_0} = 1 - \hat{F}(t_i) \quad (74.166)$$

Son nom provient par l'interprétation du rapport dans le cadre de la définition de la fonction de défaillance cumulée.

Il faut se souvenir que R provient de l'anglais "**reliability**" qui signifie "fiabilité" alors que le F en anglais signifie "**failure**" qui signifie en français "panne".

Par suite nous avons aussi:

$$\frac{\hat{f}(t_i)}{\hat{R}(t_i)} = \hat{\lambda}(t_i) \Rightarrow \hat{\lambda}(t_i) \Delta t_i = \frac{\hat{f}(t_i)}{\hat{R}(t_i)} \Delta t_i \quad (74.167)$$

Cette dernière relation servant au calcul des lois de fiabilité!

Puisque:

$$\hat{f}(t_i) \Delta t_i \leq 1 \text{ et } \lim_{i \rightarrow +\infty} \sum_i \hat{f}(t_i) \Delta t_i = 1 \quad (74.168)$$

la fonction de défaillance peut être vue comme une probabilité, ce qui nous conduit à définir naturellement son espérance:

$$E(T) = \sum t_i \hat{f}(t_i) \Delta t_i = \sum t_i \frac{n_i}{N_0} \quad (74.169)$$

Relation très utile dans la pratique qui donne en théorie le pourcentage moyen d'éléments en panne à l'instant t_i .

 **Exemple:**

Nous avons relevé sur un lot de 37 moteurs d'un type donné les défaillances suivantes répertoriées par tranches (données par des clients ou mesurées en interne sur des bancs d'essais):

0 à 1'000 h.	1'000 à 2'000 h.	2'000 à 3'000 h.	3'000 à 4'000 h.	4'000 à 5'000 h.	5'000 à 6'000 h.
1	4	7	12	11	2

Tableau: 74.6 - Défaillances des moteurs par tranches d'effort

Il faut que nous estimions la valeur de la fonction de fiabilité $\hat{R}(t_i)$, la fonction de défaillance $\hat{f}(t_i)$ et la défaillance par tranche $\hat{\lambda}(t_i)$. Les calculs sont élémentaires et nous obtenons le tableau suivant:

Intervalle d'observation	Nombre de défaillances dans l'intervalle	Survivants	Cumul des défaillants	$\hat{R}(t_i)$	$\hat{f}(t_i)$	$\hat{\lambda}(t_i)$ $10^{-6} [s^{-1}]$
0	-	37	0	100%	0	-
0 à 1'000 h.	1	36	1	97%	2.7%	27
1'000 à 2'000 h.	4	32	5	86%	10.8%	111
2'000 à 3'000 h.	7	25	12	67%	18.9%	218
3'000 à 4'000 h.	12	13	24	35.1%	32.4%	480
4'000 à 5'000 h.	11	2	35	5.4%	5.4%	846
5'000 à 6'000 h.	2	0	37	0%	-	1

Tableau: 74.7 - Analyse des défaillances des moteurs par tranches d'effort

Nous voyons ci-dessus par exemple que le taux de défaillance n'est pas constant bien évidemment!

Le taux de panne $\hat{f}(t_i)$ serait de l'ordre de 2.5 à 3% par an en 2013 pour les réfrigérateurs, lave-vaisselles, etc. Pour le matériel audio-visuel il tournerait autour des 1.5 à 2% par an.

Concernant les taux de défaillance, les ingénieurs reconnaissent souvent trois tranches d'analyses suivant que certains objets étudiés soient jeunes, en fonctionnement normal ou considéré en vieillissement.

On considère alors assez intuitivement (et parfois grossièrement) que le taux de défaillance suit une courbe en baignoire comme représenté ci-dessous (dans les tables techniques c'est souvent le taux de défaillance en fonctionnement normal qui est donné):

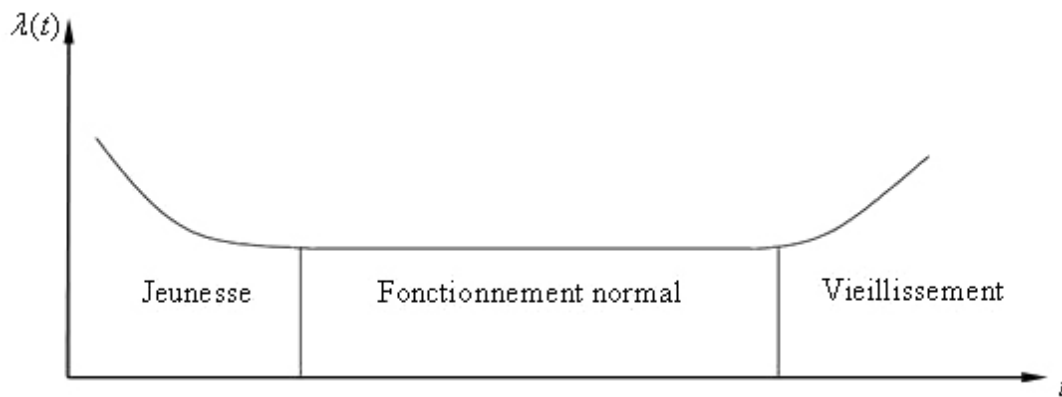


Figure: 74.10 - Baignoire de la défaillance

alors que si vous observez le tableau précédent, le taux de défaillance ne suit pas du tout une courbe en baignoire (c'est donc un contre-exemple).

Les ingénieurs en fiabilité découpent souvent la baignoire en trois parties visibles ci-dessus, mais sous la dénomination technique suivante:

- D.F.R. pour "**Decreasing Failure Rate**": les composants jeunes ayant des problèmes de fabrication non identifiés lors de procédé sont éliminés du lot ce qui a pour effet de diminuer le taux de défaillance. La loi de Weibull est relativement bien adaptée pour modéliser cette phase.
- C.F.R. pour "**Constant Failure Rate**": les composants sont dans un état stationnaire.
- I.F.R. pour "**Increasing Failure Rate**" les composants sont en fin de vie et leur taux de défaillance augmente. La loi de Weibull est à nouveau relativement bien adaptée pour modéliser cette phase.

La loi de Poisson est relativement bien adaptée pour modéliser les arrivées de pannes, et la loi exponentielle pour modéliser le temps entre pannes successives. Ce constat découle de démonstrations mathématiques disponibles dans le chapitre sur les Techniques De Gestion (théorie des files d'attentes). Donc, cela signifie que pannes sont souvent considérées comme indépendantes. Dès lors, la loi sur le nombre de pannes dans une période donnée est alors une loi "sans mémoire" et on peut modéliser par une loi de Poisson (démonstration aussi faite dans le chapitre sur les Techniques De Gestion).

Remarque: Contrairement à ce que pas mal de théoriciens pensent..., les logiciels informatiques grand public ont aussi leur taux de défaillance qui suit une courbe en baignoire. Effectivement, au début il y a des bugs non détectés qui font que la défaillance va diminuer au fur et à mesure de leur détection et leur correction. Ensuite, à cause des mises à jour fréquentes de l'environnement qui ont tendance à rajouter d'autres problèmes (service pack), le taux de défaillance se maintient à peu près constant. Enfin, avec le temps, l'évolution des technologies environnantes (framework) rendent l'appliquatif obsolète et des fonctions ne répondent ou n'agissent plus correctement ce qui fait à nouveau augmenter le taux de défaillance.

Vis-à-vis de l'efficacité de la rénovation, indiquons qu'elles peuvent (en simplifiant) très fréquemment se ranger en trois catégories:

1. "**As good as new**": C'est de la maintenance préventive dans le sens que nous changeons une pièce lorsque sa durée de vie l'amène à un taux de défaillance que nous considérons comme trop élevé et dont la rupture non anticipée coûtera plus cher que sa non-anticipation.
2. "**As bad as old**": C'est de la maintenance déficiente dans le sens que nous changeons une pièce que lorsqu'elle est arrivée à rupture, ce qui engendre majoritairement des coûts d'arrêts plus élevés que la maintenance préventive qui consiste elle à anticiper au plus juste la casse.
3. "**Restauration partielle**": C'est de la maintenance préventive minimale dans le sens que nous réparons la pièce défaillante plutôt que de la remplacer par une nouvelle. À nouveau le problème du coût doit être calculé en faisant un audit des besoins et des délais de l'entreprise.

Revenons-en à d'autres définitions au passage à la limite du continu.

Nous savons donc que le "**taux de défaillance instantané**" aura pour unité l'inverse du temps tel que $[\lambda] = \varepsilon^{-1}$. Ce taux est, dans le cadre de notre étude, pas nécessairement constant dans le temps comme nous avons pu le constater!

Soit $R(t)$ le pourcentage cumulé d'objets analysés toujours en état de bon fonctionnement d'un échantillon testé au temps t . Le nombre d'objets tombant en panne durant le temps infinitésimal dt est donc égal à:

$$\lambda(t) \cdot R(t) dt \quad (74.170)$$

ce qui correspond donc à la diminution du stock initial en bon fonctionnement au temps t .

Nous pouvons alors écrire la relation:

$$dF(t) = \lambda(t) \cdot R(t) dt \quad (74.171)$$

soit:

$$\frac{dF(t)}{R(t)} = \frac{F(t+dt) - F(t)}{R(t)} = \frac{dF(t)}{1 - F(t)} = \lambda(t) \cdot dt \quad (74.172)$$

Que l'on retrouve souvent dans la littérature sous la forme suivante:

$$\lambda(t) = h(t) = \frac{dF(t)}{dt R(t)} = \frac{f(t)}{R(t)} \quad (74.173)$$

Définition: La "**probabilité conditionnelle de défaillance**" (rien à voir avec les probabilités conditionnelles cependant...) entre t et $t + dt$ est définie par:

$$\lambda(t) dt = \frac{F(t+dt) - F(t)}{R(t)} = \frac{dF(t)}{1 - F(t)} \quad (74.174)$$

où $F(t)$ et $R(t)$ sont respectivement la fonction cumulée de défaillance (probabilité cumulée de tomber en panne au temps t) et la fonction fiabilité appelée également "**fonction de survie**". $R(t)$ valant 1 au temps 0 et... 0 après un temps infini comme nous l'avons déjà vu avant!

Dès lors, $\lambda(t)dt$ peut s'interpréter comme la défaillance instantanée! En analyse de survie, le facteur de dt est nommé aussi parfois "**fonction de risque**".

Remarque: Par la même démarche intellectuelle, plutôt que de définir une fonction de défaillance $F(t)$ et de survie $R(t)$ avec sa fonction de risque, nous pouvons définir une fonction de réparabilité avec sa fonction de $M(t)$ qui serait alors une "**fonction de maintenabilité**".

Si nous intégrons (attention u représente maintenant le temps!):

$$\begin{aligned}\int_0^t \lambda(u) du &= \int_0^t \frac{dF(u)}{1-F(u)} \\ -\int_0^t \lambda(u) du &= -\int_0^t \frac{dF(u)}{1-F(u)} \quad (74.175) \\ -\int_0^t \lambda(u) du &= \ln(1-F(u)) \Big|_0^t\end{aligned}$$

Comme $F(t_0 = 0) = 0$ nous avons:

$$-\int_0^t \lambda(u) du = \ln(1-F(t)) \Rightarrow \exp\left(-\int_0^t \lambda(u) du\right) = 1-F(t) \quad (74.176)$$

d'où:

$$\begin{aligned}R(t) &= e^{-\int_0^t \lambda(u) du} \\ F(t) &= 1 - e^{-\int_0^t \lambda(u) du}\end{aligned} \quad (74.177)$$

Par ailleurs, puisque nous avons vu que $\hat{f}(t_i) = \hat{\lambda}(t_i) \hat{R}(t_i)$, nous avons alors la "**fonction de densité/répartition de défaillance instantanée**":

$$f(t) = \lambda(t) R(t) = \lambda(t) e^{-\int_0^t \lambda(u) du} \quad (74.178)$$

Nous pouvons obtenir cette relation et interprétation sous une autre manière:

$$f(t) = \frac{dF(t)}{dt} = \lambda(t) e^{-\int_0^t \lambda(u) du} \quad (74.179)$$

où nous retrouvons donc $F(t)$ la fonction de probabilité cumulée de défaillance. Évidemment pour déterminer la loi $f(t)$, nous utilisons les outils statistiques d'ajustements habituels (cf. chapitre de Statistiques).

Nous avons alors la relation très importante (voir la plus importante!) dans la pratique qui relie la loi de densité de la défaillance instantanée et la loi de fiabilité:

$$f(t) = \frac{dF(t)}{dt} = \frac{d(1-R(t))}{dt} = -\frac{dR(t)}{dt} \quad (74.180)$$

Nous avons ci-dessus les trois expressions les plus générales liant les lois de fiabilité et le taux instantané de défaillance. Ajoutons encore que nous avons la fonction de hasard instantanée suivante qui en découle:

$$\lambda(t) = h(t) = \frac{f(t)}{R(t)} = \frac{-\frac{dR(t)}{dt}}{R(t)} = \frac{-R'(t)}{R(t)} = \frac{d}{dt}(-\ln(R(t))) = \frac{d}{dt}(-\ln(1-F(t))) \quad (74.181)$$

Il s'ensuit alors immédiatement que:

$$\begin{aligned} H(t) &= \int_0^t h(t) dt = \int_0^t \frac{d}{dt}(-\ln(1-F(t))) dt = -\int_0^t d \ln(1-F(t)) = -\ln(1-F(t)) \Big|_0^t \\ &= -\ln(1-F(t)) = -\ln(R(t)) \end{aligned} \quad (74.182)$$

Puisque $f(t)$ est la fonction de densité de défaillance, l'espérance mathématique de la défaillance est alors donnée par:

$$E(t) = \int_0^{+\infty} t f(t) dt \quad (74.183)$$

Ainsi, si la répartition des pannes est équiprobable (fonction de densité uniforme), ce qui est plutôt rare, la moitié des équipements seront hors service à $E(t)$.

Remarque: En observant 100'000 disques durs, des ingénieurs de Google auraient observé en moyenne 8% de pertes par an! Donc un taux de perte plus élevé que celui qu'aurait annoncé des fabricants qui serait d'environ 300'000 heures! Le taux de perte est plus élevé les 3 premières années! Mais peut-être que les disques qui survivent vivent plus longtemps!

Nous avons aussi en utilisant l'intégration par parties et la relation précédente:

$$\begin{aligned} E(t) &= \int_0^{+\infty} t \frac{dF}{dt} dt = \int_0^{+\infty} t dF = \int_0^{+\infty} t d(1-R(t)) = -\int_0^{+\infty} t dR(t) \\ &= -t dR(t) \Big|_0^{+\infty} + \int_0^{+\infty} R(t) dt = \int_0^{+\infty} R(t) dt \end{aligned}$$

D'où une autre manière de l'exprimer:

$$E(t) = \int_0^{+\infty} R(t) dt \quad (74.184)$$

Nous avons donc la relation:

$$E(t) = \int_0^{+\infty} t f(t) dt = \int_0^{+\infty} R(t) dt \quad (74.185)$$

Avant de continuer donnons quelques définitions d'indicateurs de maintenance dont les plus importantes sont dans la norme européenne de maintenance NF EN 13306:2010 (je précise car on trouve sur le web le pire et le meilleur concernant les définitions de ces indicateurs...):

- La **T.B.M** ou "**Time Between Maintenance**" est le temps contractuel entre deux visites ou contrôle de maintenance (non défini dans la norme).
- La **M.O.T.B.F** ou "**Mean Operating Time Between Failures**" applicable uniquement pour des éléments réparables est l'espérance mathématique du temps opérationnel d'un système entre la fin de la première défaillance et le début de la prochaine (défini dans la norme). Dans l'industrie la M.O.T.B.F. est aussi notée **M.U.T** pour "Mean Up Time".
- La **M.T.B.F** ou "**Mean Time Between Failures**" applicable uniquement pour des éléments réparables est l'espérance mathématique du temps opérationnel d'un système entre le début de deux défaillances (défini dans la norme). Cet indicateur inclut donc les temps de non-fonctionnement. Dans le domaine des services, la **M.T.B.F** est parfois notée M.T.B.S.A pour "Mean Time Between System Accidents". Dans le domaine des lignes de production la M.T.B.F. est notée **M.T.B.D** pour "Mean Time Between Defects" (on arrête pas la production parce que l'on a détecté un défaut conformément à ce que nous verrons les de notre étude des cartes de contrôle). Dans le domaine de l'informatique la M.T.B.F. est notée **M.T.B.E** pour "Mean Time Between Errors" (on arrête rarement un logiciel ou un site web parce que l'on y a détecté une erreur!).
- La **M.T.T.F** ou "**Mean Time To Failure**" applicable pour des éléments réparables ou non est l'espérance mathématique du temps opérationnel d'un système jusqu'à sa première panne (non défini dans la norme). Au fait c'est ce dernier que l'on calcule le plus souvent dans les laboratoires de tests.
- La **M.R.T** ou "**Mean Repair Time**" applicable uniquement pour des éléments réparables est l'espérance mathématique du temps de réparation (défini dans la norme).
- La **M.T.T.R** ou "**Mean Time To Restore**" applicable uniquement pour des éléments réparables est l'espérance mathématique du temps de redémarrage (défini dans la norme). Dans le domaine des services, la M.T.T.R. est parfois notée **M.T.T.R.S** pour "Mean Time To Restore Service".
- La **M.D.T** ou "**Mean Down Time**" est égale à l'espérance mathématique du temps pendant lequel le système n'est pas opérationnel. Cet indicateur (non défini dans la norme) inclut donc la M.R.T. Il en découle que la M.T.B.F est égale à la somme de la M.O.T.B.F. (M.U.T.) et la M.D.T.
- La **M.T.T.D** ou "**Mean Time To Detection**" est égale à l'espérance mathématique du temps pendant lequel le système n'est pas été détecté comme étant non opérationnel. Cet indicateur (non défini dans la norme), au même titre que la M.R.T., est inclus dans la M.D.T.
- La **M.T.T.A** ou "**Mean Time to Answer**" est égale à l'espérance mathématique du temps pendant lequel l'utilisateur a attendu une réponse ou intervention du fabricant après avoir signalé une panne. Cet indicateur (non défini dans la norme), au même titre que la M.R.T. et la M.T.T.D., est inclus dans la M.D.T.

- Le **T.R.S.** ou "**Taux de Rendement Synthétique**" dans le domaine de la maintenance est défini traditionnellement comme étant simplement le rapport entre le temps de fonctionnement opérationnel et le temps attendu de fonctionnement opérationnel (non défini dans la norme).

- Le **S.A.** ou "**Service Ability**" qui dans le domaine de la maintenance est souvent donné par les deux relations suivantes et dont le but est qu'il soit le plus proche possible de 1 (c'est-à-dire de faire tendre vers zéro la M.D.T.):

$$SA = \frac{MTBF - MDT}{MTBF} = \frac{MOTBF}{MOTBF + MDT} \quad (74.186)$$

et on peut en définir encore beaucoup d'autres en dehors de la norme...

Voici un schéma qui résume peut-être un peu mieux tout cela:

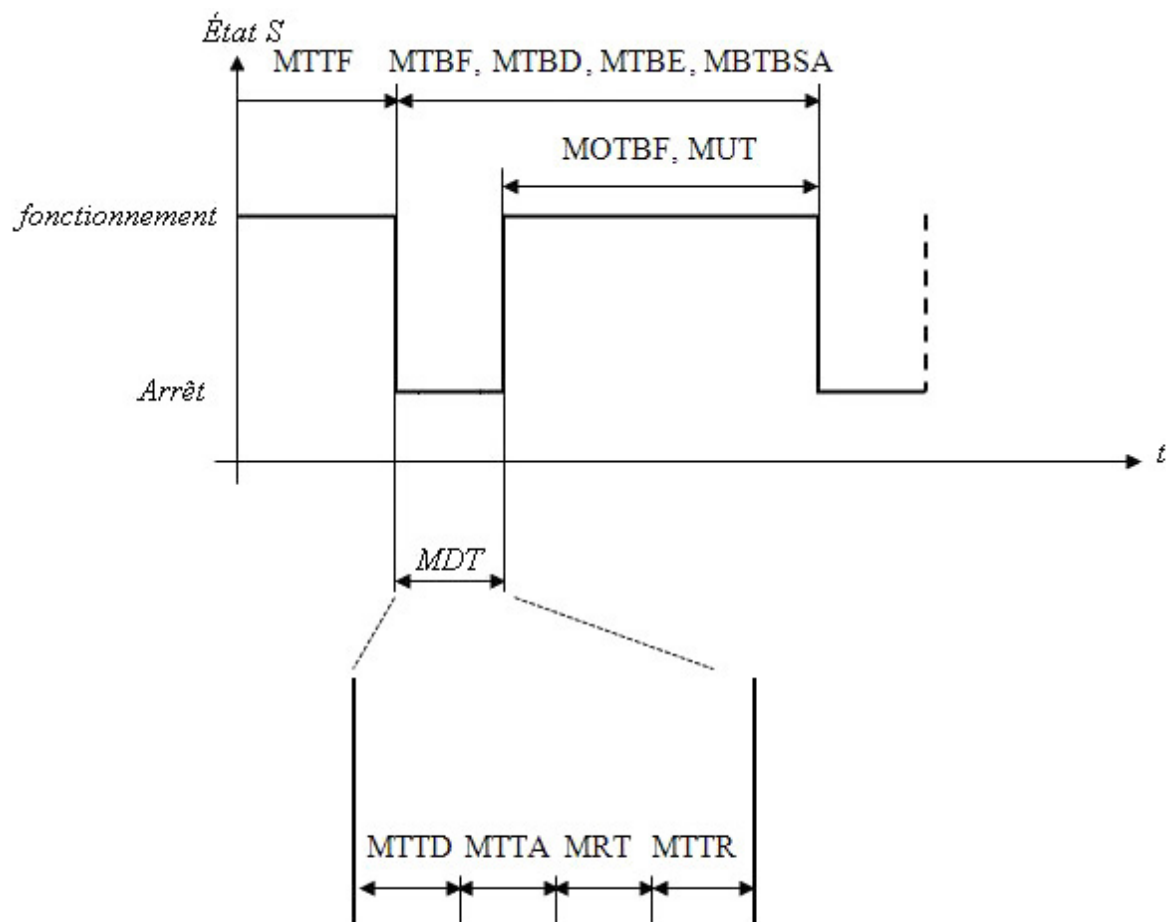


Figure: 74.11 - Résumé des indicateurs de fiabilité les plus importants

Remarques:

R1. Pour chacun de ces indicateurs on indiquera s'il a été pris en compte des causes extérieures ou non.

R2. Si la taille du lot testée est très petite, il est d'usage d'utiliser la moyenne arithmétique comme estimateur de l'espérance. Évidemment cela peut amener à de grossières erreurs...

R3. Il est bien évidemment possible à partir des indicateurs définis de calculer si une machine (élément) est apte à garantir le service souhaité (comme par exemple un nombre de pièces par année à fabriquer!). Donc non seulement ces indicateurs peuvent être mesurés, mais ils peuvent donc aussi être utilisés pour vérifier l'atteinte d'un objectif souhaité!

R4. En ce tout début de 21ème siècle les pays devraient en toute rigueur légiférer pour obliger tous les industriels à communiquer la M.T.T.F. de leurs produits afin que le consommateur puisse mieux faire son choix à l'achat et comparer les valeurs par rapport à la garantie fournie! Malheureusement, ce n'est pas le cas et cela permettrait de mettre en évidence une mauvaise tradition actuelle dans l'industrie des produits grand public qui est de fabriquer des composants dont la durée de vie tourne autour des 200'000 heures afin d'assurer aux industriels un renouvellement de leur marché.

Exemple:

Une machine est censée travailler 24 heures sur 24, 7 jours sur 7. Elle a tourné jusqu'à aujourd'hui pendant 5'020 heures mais avec 2 arrêts d'un total de 20 heures (donc inclus dans les 5'020 [h]). Donnez les indicateurs classiques relativement au peu d'informations communiquées:

$$\begin{aligned}
 MTBF &= \frac{5020}{2} = 2510 \text{ [h]} \\
 MDT &= \frac{20}{2} = 10 \text{ [h]} \\
 MOTBF = MUT &= \frac{5020 - 20}{2} = 2500 \text{ [h]} \quad (74.187) \\
 SA &= \frac{MOTBF}{MOTBF + MDT} = \frac{2500}{2500 + 10} \cong 99.60\%
 \end{aligned}$$

La classification des systèmes en terme de disponibilité conduit communément à 7 classes de non prise en compte (système disponible 90% du temps, et donc indisponible plus d'un mois par an) à ultra disponible (disponible 99.99999% du temps et donc indisponible seulement 3 secondes par an): ces différentes classes correspondent au nombre de 9 dans le pourcentage de temps durant lequel les systèmes de la classe sont disponibles (une année comporte 525'600 minutes):.

Type	Indisponibilité (minutes par an)	Pourcentage disponibilité	Classe
non géré	50'000 (~35 jours)	90%	1
géré	5'000 (~3.5 jours)	99%	2
bien géré	500 (~8 heures)	99.9%	3
tolérance fautive	50	99.99%	4
haute disponibilité	5	99.999%	5
très haute disponibilité	0.5	99.9999%	6
très grande haute disponibilité	0.05	99.99999%	7

Tableau: 74.8 - Classes de défaillances AMDEC

L'usage de ces paramètres dans le cadre de fiabilité font dire que nous avons une "**approche en valeurs moyennes**".

Attention cependant à une chose! Ce n'est pas parce qu'un événement à une probabilité plus petite ou infiniment plus petite qu'une autre qu'il en est moins important!! Effectivement il faut bien évidemment pendre en compte (de façon certes un peu empirique) la gravité de la défaillance. Ainsi, une centrale nucléaire à beau être peut-être 1'000 fois plus sûre qu'une central thermique en termes de probabilités... il n'en est rien en termes de gravité lorsqu'un défaillance majeure se produit! Effectivement dans le cas d'une central thermique il peut y avoir au pire 500 personnes touchées alors qu'avec une centrale nucléaire c'est une autre histoire...

Signalons enfin un cas simple: Certains composants (électroniques typiquement) ont dans leur période de maturité un taux de défaillance constant. La loi de probabilité cumulée de la défaillance qui en découle s'en déduit alors immédiatement puisque la fonction de hasard instantanée est donnée par $\lambda(t) = \lambda$:

$$F(t) = 1 - \exp\left(-\int_0^t \lambda(t) dt\right) = 1 - e^{-\lambda t} \quad (74.188)$$

Indiquons que nous retrouvons très souvent cette dernière relation sous la forme suivante dans la littérature:

$$1 - F(t) = R(t) = e^{-\lambda t} \quad (74.189)$$

Les fonction de hasard et de survie sont alors graphiquement de la forme suivante:

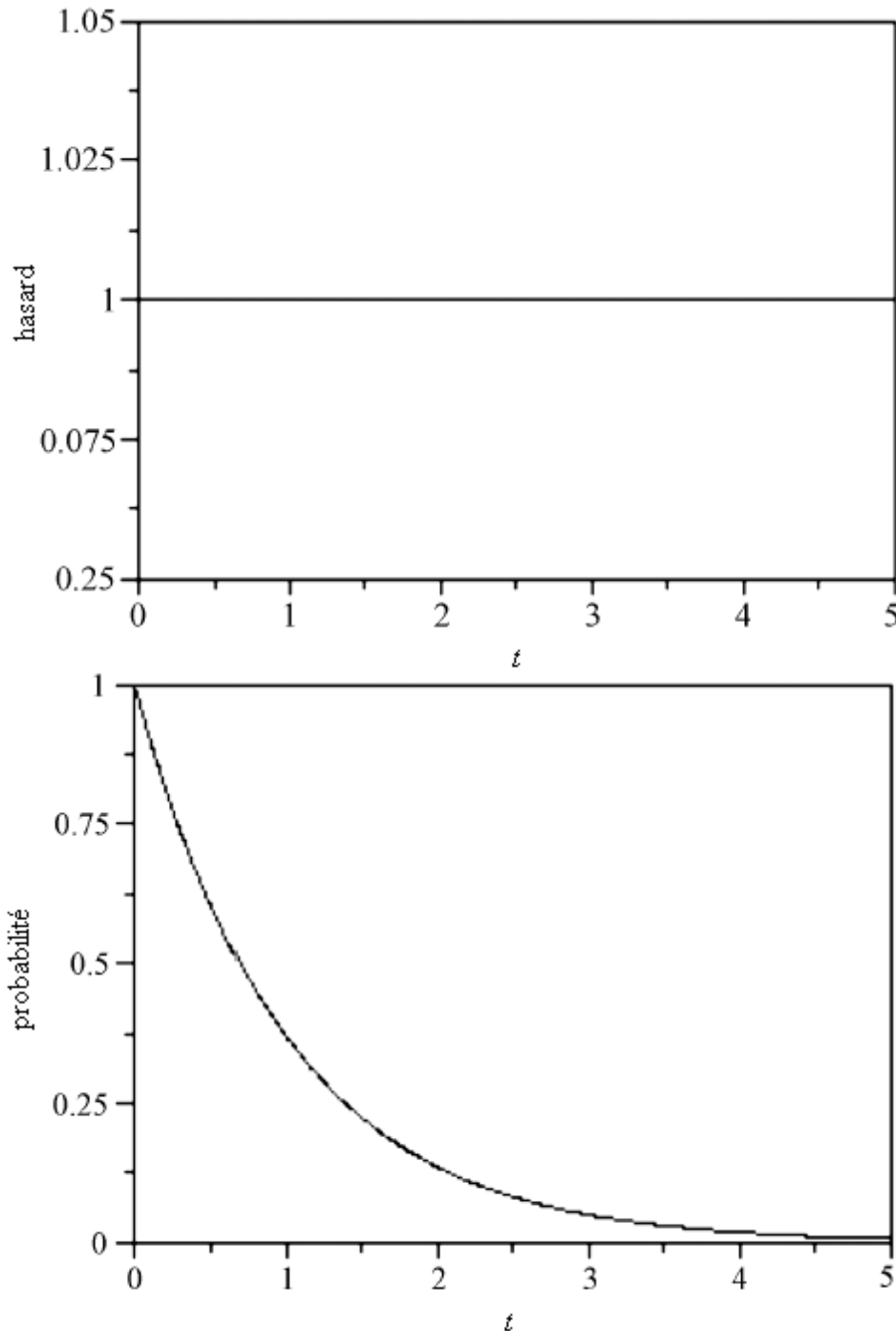


Figure: 74.12 - Plot de la fonction de hasard et de survie de la loi exponentielle

La fonction de densité des éléments défaillants au temps t est alors:

$$f(t) = \frac{dF(t)}{dt} = \lambda e^{-\lambda t} \quad (74.190)$$

Elle suit donc une loi exponentielle! Cette loi et ses moments nous sont connus (cf. chapitre de Statistiques). Il devient alors facile déterminer le M.T.T.F. et son écart-type (inverse du taux de défaillance) ainsi qu'un intervalle de confiance.

Par ailleurs, si nous calculons la fiabilité $R(t)$ au temps correspondant à la M.T.T.F. (inverse du taux de défaillance dans le cas de la loi exponentielle) nous obtiendrons toujours une probabilité cumulée de 36.8% (donc en gros une chance sur trois de fonctionner à ce moment là et 2 chances sur 3 de tomber en panne) et non de 50% comme on peut le penser intuitivement (ce qui est le cas seulement si la loi de probabilité est symétrique).

Étant donné que les tables techniques de fiabilité dans l'industrie supposent presque toujours le taux de fiabilité constant alors nous comprenons mieux l'importance de celle-ci (nous en aurons un exemple lors de notre présentation des topologies de systèmes).

Rappelons aussi que nous avons vu dans le chapitre de Probabilités que si nous avons un ensemble d'événements indépendants (mécanismes indépendants) avec une probabilité donnée, alors les probabilités calculées sur l'ensemble dans sa totalité implique la multiplication des probabilités.

Donc dans un mécanisme ayant des pièces indépendantes mais essentielles (système dit en "série") la fonction de densité de fiabilité globale sera égale à la multiplication des lois de probabilités cumulées $R(t)$ et ce qui équivaut donc dans le cas d'une loi exponentielle d'avoir une seule loi exponentielle dont le taux de défaillance global est égal à la somme des taux de défaillance des différentes pièces.

Un autre exemple: En mécanique, où le phénomène d'usure est à l'origine de la défaillance, le taux de défaillance est souvent du type linéaire (il augmente donc de manière constante avec le temps et est non nul lors du premier allumage de l'appareil):

$$\lambda(t) = at + b \quad (74.191)$$

Alors:

$$F(t) = 1 - \exp\left(-\int_0^t (at + b)dt\right) = 1 - \exp\left(-\frac{at^2}{2} - bt\right) \quad (74.192)$$

Soit:

$$f(t) = \frac{dF(t)}{dt} = (at + b) \exp\left(-\frac{at^2}{2} - bt\right) \quad (74.193)$$

Comme (pour un matériel dont le temps de réparation est négligeable):

$$MTBF = \int_0^{+\infty} R(t)dt = \int_0^{+\infty} \exp\left(-\frac{at^2}{2} - bt\right)dt \quad (74.194)$$

cette intégrale ne peut se calculer que par une méthode numérique. Dès lors on préfère prendre des lois approchées à l'aide d'ajustement du Khi-deux, comme par exemple la loi de Weibull, qui est un peu la poubelle à tout mettre dans le domaine...

Il faut savoir que dans les banques de données de fiabilité communes et gratuitement disponibles sur le marché sont normalement donnés: la dénomination du matériel ou du composant, la MTBF, le taux de défaillance moyen, le patrimoine statistique, un coefficient multiplicatif du taux de défaillances dépendant de l'environnement ou des contraintes d'utilisation.

2.3. MODÈLE DE WEIBULL

Encore une fois, les techniques de maintenances utilisent les probabilités et statistiques donc nous renvoyons le lecteur au chapitre du même nom. Cependant, il existe dans le domaine de la maintenance (et pas seulement) une fonction de densité de probabilité très utilisée appelée "**loi de Weibull**".

Elle serait complètement empirique et est définie par:

$$f(x) = \frac{\beta}{\eta} \left(\frac{x-\gamma}{\eta} \right)^{\beta-1} \exp \left(- \left(\frac{x-\gamma}{\eta} \right)^{\beta} \right) \quad (74.195)$$

où $f(x) \geq 0, x \geq 0$ avec $\beta > 0, \eta > 0, -\infty < \gamma < +\infty$ qui sont respectivement appelés "**paramètre d'échelle**" η , "**paramètre de forme**" β et "**paramètre d'emplacement**" γ .

Remarque: Les ingénieurs dans les entreprises doivent obligatoirement se référer à la norme CEI 61649 *Analyse de Weibull* pour en faire usage de manière standardisée!

La loi de Weibull est aussi souvent notée sous la forme suivante en posant $x := x - \gamma, \eta := \beta, \beta := \alpha$ (dans ce cas le paramètre d'emplacement est implicitement posé comme étant nul):

$$f(x) = \frac{\alpha}{\beta} \frac{x^{\alpha-1}}{\beta^{\alpha-1}} \exp \left(- \frac{x^{\alpha}}{\beta^{\alpha}} \right) = \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1} \exp \left(- \left(\frac{x}{\beta} \right)^{\alpha} \right) \quad (74.196)$$

Elle peut être calculée dans la version française de Microsoft Excel 11.8346 sous cette dernière forme en utilisant la fonction LOI.WEIBULL().

Remarque: Elle peut être vue comme une généralisation de la fonction de distribution exponentielle (nous allons voir plus loin pourquoi!) avec l'avantage qu'il est possible de jouer avec les trois paramètres de manière à obtenir presque n'importe quoi.

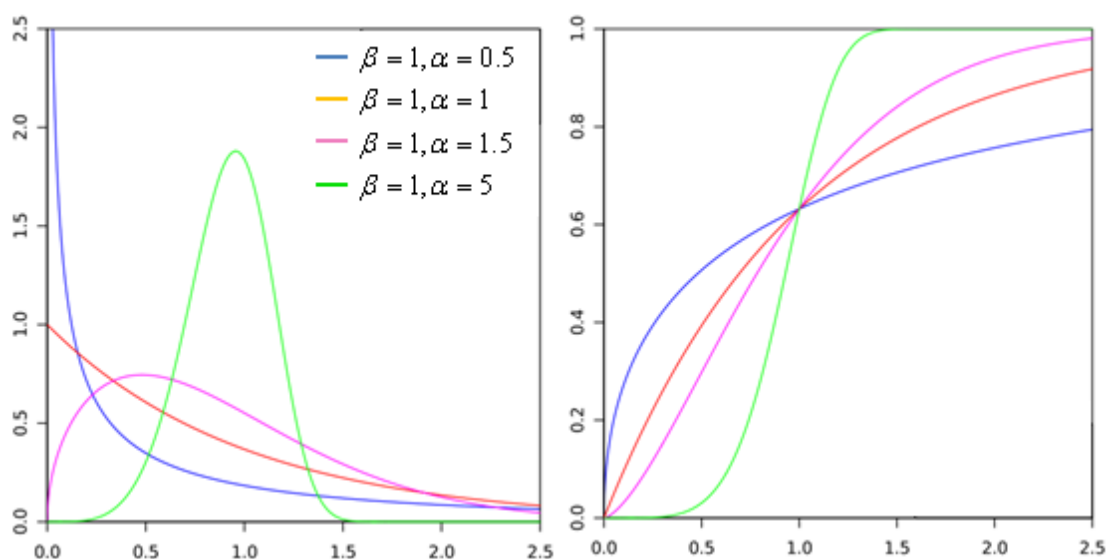


Figure: 74.13 - Fonctions de densité et de distribution de Weibull pour différentes valeurs de alpha (source: Wikipédia)

En annulant le paramètre d'emplacement γ , nous obtenons la "**distribution de Weibull à deux paramètres**":

$$f(x) = \frac{\beta}{\eta} \left(\frac{x}{\eta} \right)^{\beta-1} \exp \left(- \left(\frac{x}{\eta} \right)^{\beta} \right) \quad (74.197)$$

qui a une application pratique importante et dont nous avons calculé les estimateurs des paramètres dans le chapitre de Statistiques.

En posant encore une fois $\gamma = 0$ et en assumant que $\beta = C = c^{te}$ nous avons la "**distribution de Weibull à un paramètre**" (c'est un peu idiot comme définition mais bon...):

$$f(x) = \frac{C}{\eta} \left(\frac{x}{\eta} \right)^{C-1} \exp \left(- \left(\frac{x}{\eta} \right)^C \right) \quad (74.198)$$

où le seul paramètre inconnu est le paramètre d'échelle η . Nous assumons que le paramètre β est connu a priori d'expériences passées sur des échantillons identiques. Remarquons que la fonction de distribution de Weibull à un paramètre est la dérivée de la fonction de répartition (probabilité cumulée) suivante (il y a trois dérivées intérieures successives pour retomber sur la fonction de distribution... ce qui constitue un bon exemple de ce que nous avons vu dans le chapitre de Calcul Différentiel Et Intégral):

$$F(x) = 1 - \exp \left(- \left(\frac{x}{\eta} \right)^{\beta} \right) \quad (74.199)$$

Il s'ensuit alors dans l'ingénierie de la fiabilité que dans ce cas particulier, nous avons alors:

$$R(t) = 1 - F(t) = 1 - \left(1 - \exp \left(- \left(\frac{t}{\eta} \right)^{\beta} \right) \right) = \exp \left(- \left(\frac{t}{\eta} \right)^{\beta} \right) \quad (74.200)$$

Nous retrouvons également parfois cette relation dans la littérature sous la forme:

$$\begin{aligned} f(x) &= C \frac{1}{\eta} \left(\frac{1}{\eta} \right)^{C-1} x^{C-1} \exp \left(- \left(\frac{x}{\eta} \right)^C \right) = C \left(\frac{1}{\eta} \right)^C x^{C-1} \exp \left(- \left(\frac{1}{\eta} \right)^C x^C \right) \\ &= a \cdot b \cdot x^{a-1} \exp(-b \cdot x^C) \end{aligned} \quad (74.201)$$

Remarquons que si β est égal à l'unité, nous retrouvons alors une loi exponentielle d'espérance η :

$$f(x) = \frac{\beta}{\eta} \left(\frac{x}{\eta} \right)^{\beta-1} \exp \left(- \left(\frac{x}{\eta} \right)^{\beta} \right) = \frac{1}{\eta} e^{-\frac{x}{\eta}} \quad (74.202)$$

La MTBF est alors donnée par l'espérance de la loi de Weibull avec γ non nul:

$$E(X) = \int_{\gamma}^{+\infty} x \cdot f(x) dx = \int_{\gamma}^{+\infty} x \frac{\beta}{\eta} \left(\frac{x-\gamma}{\eta} \right)^{\beta-1} \exp \left(- \left(\frac{x-\gamma}{\eta} \right)^{\beta} \right) dx \quad (74.203)$$

Posons:

$$t = \left(\frac{x-\gamma}{\eta} \right)^{\beta} \Rightarrow dt = \frac{\beta}{x-\gamma} \left(\frac{x-\gamma}{\eta} \right)^{\beta} dx = \frac{\beta}{x-\gamma} t dx \Rightarrow dx = \frac{x-\gamma}{\beta} t^{-1} dt \quad (74.204)$$

avec:

$$t = \left(\frac{x-\gamma}{\eta} \right)^{\beta} \Rightarrow t^{1/\beta} \eta + \gamma = x \quad (74.205)$$

Ce qui donne:

$$\begin{aligned} E(X) &= \int_0^{+\infty} x \frac{\beta}{\eta} t \left(\frac{x-\gamma}{\eta} \right)^{-1} e^{-t} \left(\frac{x-\gamma}{\beta} t^{-1} \right) dt = \int_0^{+\infty} x e^{-t} dt = \int_0^{+\infty} (t^{1/\beta} \eta + \gamma) e^{-t} dt = \\ &= \int_0^{+\infty} t^{1/\beta} \eta e^{-t} dt + \int_0^{+\infty} \gamma e^{-t} dt = \eta \int_0^{+\infty} t^{1/\beta} e^{-t} dt + \gamma \end{aligned} \quad (74.206)$$

La première intégrale nous est déjà connue, nous l'avons déjà vue dans le chapitre de Calcul Différentiel. Il s'agit de la fonction gamma d'Euler:

$$\Gamma_{-1}(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt = x! \Rightarrow \Gamma_{-1} \left(\frac{1+\beta}{\beta} \right) = \int_0^{+\infty} t^{1/\beta} e^{-t} dt \quad (74.207)$$

Nous avons finalement:

$$E(X) = MTBF = \eta \int_0^{+\infty} t^{1/\beta} e^{-t} dt + \gamma = \gamma + \eta \Gamma_{-1} \left(\frac{\beta+1}{\beta} \right) \quad (74.208)$$

En annulant paramètre d'emplacement γ il vient le cas courant en fiabilité:

$$E(X) = \eta \Gamma_{-1} \left(1 + \frac{1}{\beta} \right) \quad (74.209)$$

Remarque: Si par le plus grand des hasards $\beta^{-1} \in \mathbb{N}$, alors comme nous l'avons démontré lors de notre étude de la fonction Gamma d'Euler:

$$E(X) = MTBF = \gamma + \eta \Gamma_{\circ} \left(\frac{1}{\beta} \right) = \gamma + \eta \left(\frac{1}{\beta} \right)! \quad (74.210)$$

Dans le cas où $\beta^{-1} \in \mathbb{R}$ il faut faire appel aux tables numériques obtenues par les algorithmes vus dans le chapitre de Méthodes Numériques.

De même:

$$E(X^2) = \int_0^{+\infty} x^2 f(x) dx = \int_0^{+\infty} (t^{1/\beta} \eta + \gamma)^2 e^{-t} dt = \eta^2 \Gamma_{-1} \left(\frac{\beta+2}{\beta} \right) + 2\eta \gamma \Gamma_{-1} \left(\frac{\beta+1}{\beta} \right) + \gamma^2 \quad (74.211)$$

Finalement lorsque paramètre d'emplacement est nul, il vient:

$$\sigma^2 = Var(X) = E(X^2) - E(X)^2 = \eta^2 \left(\Gamma_{-1} \left(\frac{\beta+2}{\beta} \right) - \Gamma_{-1} \left(\frac{\beta+1}{\beta} \right)^2 \right) \quad (74.212)$$

Remarque: Certains spécialistes, lorsqu'il est fait usage des deux moments d'ordre deux dans l'analyse probabiliste de la défaillance, parlent parfois "**d'approche en moyenne quadratique**"...

Pour clore, signalons qu'il est d'usage de considérer 5 situations importantes et pertinentes de la loi de Weibull en fonction des valeurs de ses paramètres

1. La fonction de densité de survie ressemble à une décroissance exponentielle et dès lors, la fonction de risque (taux de défaillance) est élevée au départ et diminue au fil du temps (première partie en forme de "baignoire" de la fonction de risque). Il s'agit typiquement d'une situation où les défaillances sont précoces car les défaillances se produisent dans la période initiale de la vie du produit.
2. La fonction de densité de survie ressemble à une décroissance exponentielle. Dès lors, le risque de défaillance est constant au cours de la vie de la pièce (deuxième partie en forme de "baignoire" de la fonction de risque). Il s'agit typiquement d'une situation où les défaillances sont aléatoires et sources de multiples causes. Dans cette situation, la loi de Weibull modélise la phase de "vie utile" du produit.
3. La fonction de densité de survie s'élève jusqu'à un sommet puis décroît avec une asymétrie assez prononcée (densité forte à gauche) et dès lors, la fonction de risque (taux de défaillance) augmente rapidement au début et finit par se stabiliser. Il s'agit typiquement d'une situation où l'usure initiale est très rapide.
4. La fonction de densité de survie s'élève jusqu'à un sommet puis décroît avec faible asymétrie assez (densité un peu plus élevée à gauche) et dès lors, la fonction de risque (taux de défaillance) augmente de manière linéaire. Il s'agit typiquement d'une situation où le risque de défaillance est dû à l'usure et augmente de façon constante au cours de la durée de vie du produit.

5. La fonction de densité survie ressemble à une gaussienne large et la fonction de risque (taux de défaillance) croît rapidement (croissance de type exponentielle). Il s'agit typiquement de la situation où l'on arrive à la fin de la vie d'un ensemble de pièces (troisième partie en forme de "baignoire" de la fonction de risque).

Voici quelques exemples types de fonction de risque (hasard) de Weibull:

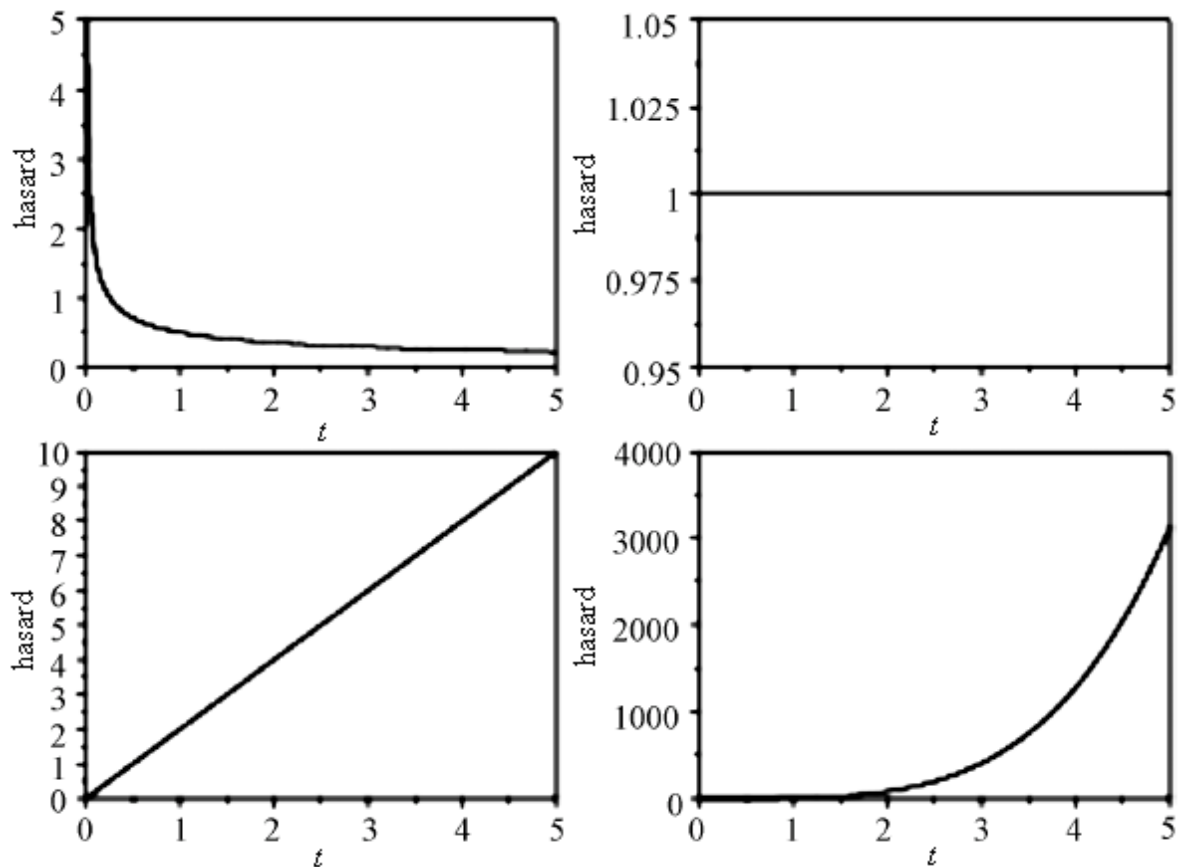


Figure: 74.14 - Plot de 4 fonctions de hasard de Weibull pour différentes valeurs des paramètres

avec les fonctions de survie correspondantes:

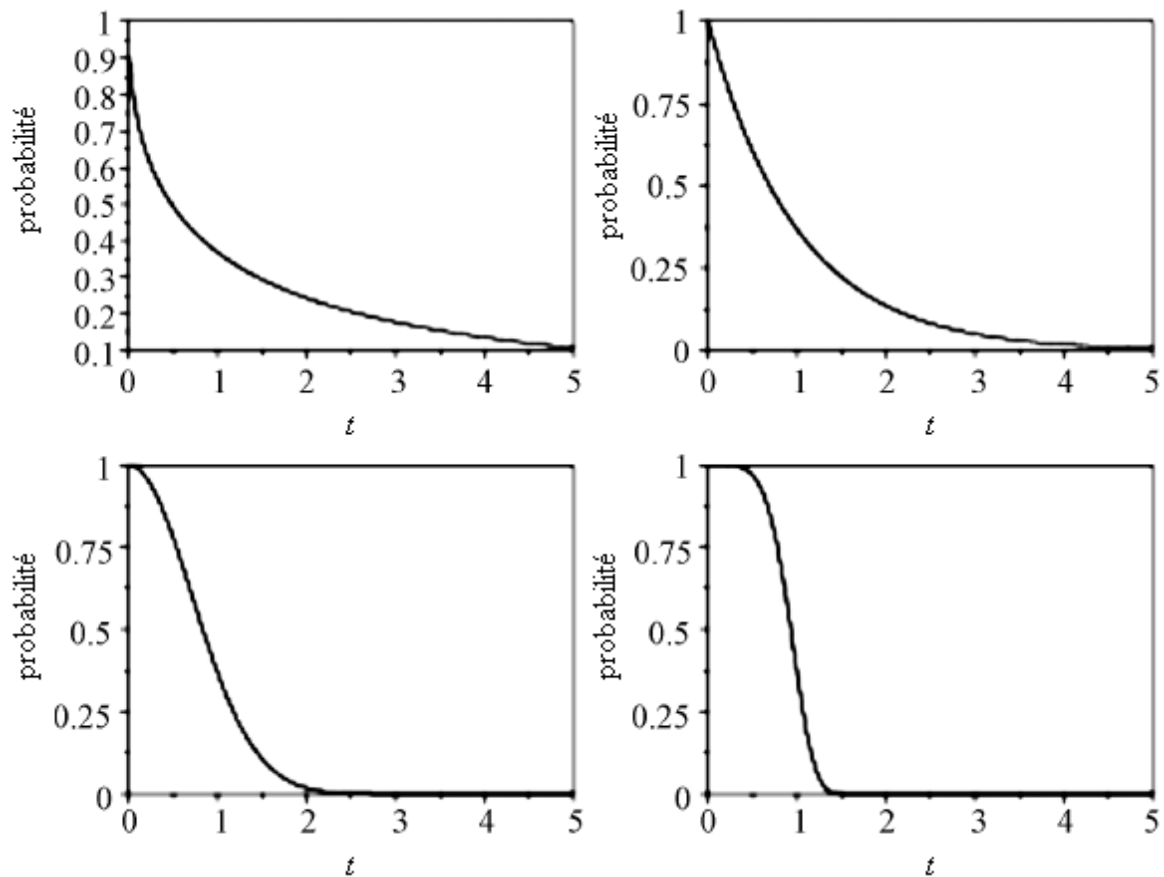


Figure: 74.15 - Plot de 4 fonctions de survie de Weibull pour différentes valeurs des paramètres

Exemple:

Un constructeur de voiture connaît la distribution du temps jusqu'à la première panne d'un de ses véhicules dans des conditions moyennes de conduite. La fonction de répartition de Weibull obtenue de par les essais est fonction de Weibull à deux paramètres donnée par:

$$f(x) = 1 - \exp\left(-\left(\frac{x}{\eta}\right)^\beta\right) = 1 - \exp\left(-\left(7.7 \cdot 10^{-6} x\right)^{1.22}\right) \quad (74.213)$$

où x est la distance parcourue en kilomètres des véhicules tests.

Si nous voulons trouver la garantie en kilomètres englobant 5% des véhicules, nous devons alors résoudre l'équation:

$$1 - \exp\left(-\left(7.7 \cdot 10^{-6} x\right)^{1.22}\right) = 5\% \quad (74.214)$$

Soit avec un peu d'algèbre élémentaire:

$$t = \frac{[-\ln(95\%)]^{1/1.22}}{7.7 \cdot 10^{-6}} \cong 11'381 \text{ [km]} \quad (74.215)$$

2.4. TOPOLOGIE DES SYSTÈMES

Lorsque nous travaillons avec des systèmes réels non réparables (mécaniques, électroniques ou autres), nous sommes confrontés à des contraintes différentes suivant le type de montage que nous avons. L'étude de ce type de systèmes est aussi appelée "**Reliability Block Diagram**".

Les deux hypothèses principales de l'étude de ces systèmes sont:

H1. La panne d'un composant est indépendante des autres

H2. Pas de pannes arrivant conjointement

Nous reconnaissons 5 topologies principales dont chaque composant est représenté par un bloc:

T1. Topologie série:

Si un seul composant est défectueux le système ne fonctionne plus (les exemples sont tellement nombreux et simples à trouver que nous n'en citerons pas).

Alors dans l'hypothèse où la panne d'un composant est indépendante des autres, la probabilité cumulée de défaillance est égale au produit des probabilités cumulées (cf. chapitre de Probabilités).

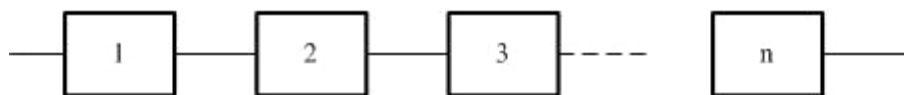


Figure: 74.16 - Topologie série

Dans le cadre des arbres de défaillance probabilistes (voir plus bas), cette topologie correspond à une porte OU (OR) avec n entrées car il suffit qu'un des blocs soit en panne pour que la sortie ne fonctionne plus:

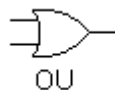


Figure: 74.17 - Porte OU (OR)

Comme souvent il est fait usage de la loi exponentielle, la multiplication de plusieurs termes de probabilités cumulées étant relativement longue à écrire, nous lui préférons l'utilisation de la probabilité de fiabilité cumulée R .

Ainsi, la fiabilité (probabilité de fonctionnement) d'un système série est donnée par:

$$R_s = 1 - F_s = \prod_{i=1}^n R_i \quad (74.216)$$

Ce qui nous amène bien à une valeur nulle pour la fiabilité si au minimum un bloc a une fiabilité nulle.

Ce qui se note traditionnellement dans le domaine:

$$\phi(x) = \min(x_1, x_2, \dots, x_n) = \prod_{i=1}^n x_i \quad (74.217)$$

Remarque: Le lecteur attentif aura remarqué que le système série est toujours moins fiable que sa composante la moins fiable!

Attention!! Dans le cas des composantes électroniques, le taux de défaillance est souvent considéré comme constant par souci de simplification et la fonction de densité est alors celui de la loi exponentielle:

$$f(t) = \lambda e^{-\lambda t} \quad (74.218)$$

Or, nous avons démontré que dans le cas d'un système non réparable:

$$E(t) = \int_0^{+\infty} t f(t) dt = MTBF \quad (74.219)$$

Et comme nous avons démontré dans le chapitre de Statistiques que l'espérance de la loi exponentielle est:

$$E(t) = \int_0^{+\infty} t \lambda e^{-\lambda t} dt = \frac{1}{\lambda} = MTBF \quad (74.220)$$

Nous avons aussi démontré plus haut que:

$$f(t) = \frac{dF}{dt} = -\frac{dR(t)}{dt} \quad (74.221)$$

Donc:

$$R(t) = e^{-\lambda t} = e^{-\frac{1}{MTBF}t} \quad (74.222)$$

Alors pour un système série à taux de défaillance constant:

$$R_S = \prod_i e^{-\lambda_i t} = \prod_i e^{-\frac{1}{MTBF_i}t} = e^{-\left(\sum_i \frac{1}{MTBF_i}\right)t} \quad (74.223)$$

Ainsi, dans ce cas particulier:

$$MTBF_S = \left(\sum_{i=1}^n \frac{1}{MTBF_i} \right)^{-1} \quad (74.224)$$

ou autrement écrit:

$$\lambda_S = \sum_{i=1}^n \lambda_i \quad (74.225)$$

et... le problème sur Internet c'est que les pages web qui parlent de systèmes à topologie série (ou parallèle) font implicitement usage d'une loi exponentielle, d'où parfois de grosses erreurs de la part des praticiens qui utilisent cette dernière relation sans avoir étudié les détails mathématiques sous-jacents.

T2. Topologie parallèle:

Contrairement au système précédent, le système continue à fonctionner si au moins un composant fonctionne (typiquement les systèmes de redondance dans les avions, les fusées ou les centrales nucléaires).

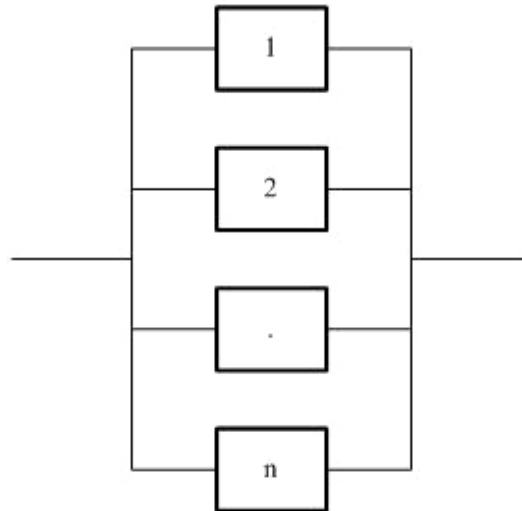


Figure: 74.18 - Topologie parallèle

Dans le cadre des arbres de défaillance probabilistes (cf. chapitre Techniques De Gestion), cette topologie correspond à une porte ET (AND) avec n entrées car il faut que tous les blocs soient en panne pour que la sortie ne fonctionne plus:



Figure: 74.19 - Porte ET (AND)

En d'autres termes, il s'arrête de fonctionner si et seulement si:

$$\prod_{i=1}^n F_i = \prod_{i=1}^n (1 - R_i) = 1 \quad (74.226)$$

Donc, il vient immédiatement que:

$$R_S = 1 - \prod_{i=1}^n F_i \quad (74.227)$$

Remarque: En gestion de projets ce type de structure parallèle est typiquement utilisée lorsque l'on souhaite connaître la fiabilité totale de plusieurs fournisseurs de pièces identiques (comme cela si l'un d'eux à des retards ou autres problèmes, les autres doivent permettre d'atténuer l'impact négatif) en connaissant ou en estimant bien évidemment la fiabilité de chacun...

Observons (utile pour plus tard) que dans le cas très particulier où la fiabilité de tous les éléments est égale, nous avons:

$$R_S = 1 - \prod_{i=1}^n F_i = 1 - F^n = 1 - (1 - R)^n \quad (74.228)$$

Nous avons donc pour le système parallèle dont les composantes sont à taux de défaillance constant et tous identiques (pour simplifier l'étude):

$$R_S = 1 - \prod_{i=1}^n F_i = 1 - \prod_{i=1}^n (1 - R_i) = 1 - \prod_{i=1}^n (1 - e^{-\lambda t}) \quad (74.229)$$

Nous avons alors:

$$MTBF_S = \int_0^{+\infty} \left[1 - \prod_{i=1}^n (1 - e^{-\lambda t}) \right] dt = \int_0^{+\infty} \left[1 - (1 - e^{-\lambda t})^n \right] dt \quad (74.230)$$

En posant:

$$u = 1 - e^{-\lambda t} \Rightarrow du = \lambda e^{-\lambda t} dt = \lambda(1 - u) dt \quad (74.231)$$

et en remplaçant dans l'expression antérieure, nous avons:

$$MTBF_S = \int_0^1 \frac{1 - u^n}{\lambda(1 - u)} du = \frac{1}{\lambda} \int_0^1 \frac{u^n - 1}{u - 1} du \quad (74.232)$$

Il n'est pas possible d'intégrer cette dernière relation de manière formelle, mais si on compare pour différentes valeurs de n fixées alors nous voyons très vite que:

$$MTBF_S = \frac{1}{\lambda} \sum_{i=1}^n \frac{1}{i} \quad (74.233)$$

ou autrement écrit:

$$\lambda_S = \lambda \left(\sum_{i=1}^n \frac{1}{i} \right)^{-1} \quad (74.234)$$

Faisons quand même le développement détaillé pour un système parallèle à deux composantes dont le taux de défaillance est constant et non identique. Pour cela, nous partons de la relation démontrée plus haut:

$$R_S = 1 - \prod_{i=1}^n F_i = 1 - \prod_{i=1}^n (1 - R_i) \quad (74.235)$$

Soit dans le cas où n vaut 2, nous avons:

$$R_S = 1 - (1 - R_1)(1 - R_2) = R_1 + R_2 - R_1 R_2 \quad (74.236)$$

et donc:

$$\begin{aligned}
 MTBF_S &= \int_0^{+\infty} \left[1 - \prod_{i=1}^2 (1 - R_i) \right] dt = \int_0^{+\infty} (R_1 + R_2 - R_1 R_2) dt \\
 &= \int_0^{+\infty} e^{-\lambda_1 t} dt + \int_0^{+\infty} e^{-\lambda_2 t} dt - \int_0^{+\infty} e^{-\lambda_1 t} e^{-\lambda_2 t} dt = \int_0^{+\infty} e^{-\lambda_1 t} dt + \int_0^{+\infty} e^{-\lambda_2 t} dt - \int_0^{+\infty} e^{-(\lambda_1 + \lambda_2)t} dt \quad (74.237) \\
 &= -\frac{1}{\lambda_1} e^{-\lambda_1 t} \Big|_0^{+\infty} - \frac{1}{\lambda_2} e^{-\lambda_2 t} \Big|_0^{+\infty} + \frac{1}{\lambda_1 + \lambda_2} e^{-(\lambda_1 + \lambda_2)t} \Big|_0^{+\infty} = \frac{1}{\lambda_1} + \frac{1}{\lambda_2} - \frac{1}{\lambda_1 + \lambda_2}
 \end{aligned}$$

Donc si $\lambda_1 = \lambda_2$, nous avons:

$$MTBF_S = \frac{1}{\lambda} + \frac{1}{\lambda} - \frac{1}{2\lambda} = \frac{2}{\lambda} - \frac{1}{2\lambda} = \frac{3}{2\lambda} \quad (74.238)$$

Soit:

$$\lambda_S = \frac{2}{3} \lambda \quad (74.239)$$

Avant de continuer avec d'autres topologies, voici un exemple d'un petit devoir scolaire (avec solution) mélangeant topologie parallèle et et série qui montre que cela ne s'applique pas seulement à des machines.

Exemple:

Un laborantin doit quotidiennement effectuer 3 tests différents sur un échantillon, donné : Les tests A , B et C dans l'ordre donné! Nous savons que les laborantins ratent le test A une fois sur quatre mais peuvent le recommencer 2 fois en cas d'échec (donc ils peuvent le faire 3 fois au total), ils réussissent le B dans 96% des cas et peuvent aussi le recommencer une fois en cas d'échec (donc ils peuvent le faire 2 fois au total), il ratent le test C une fois sur deux et ne peuvent pas le recommencer. Quelle est la probabilité qu'un laborantin réussisse les 3 tests de suite?

Pour résoudre cet exercice il faut voir chacun des tests reproductible comme un processus parallèle et l'ensemble des tests comme un processus série. Nous avons alors la probabilité totale qui est:

$$\begin{aligned}
 R_S &= \underbrace{\left(1 - \prod_{i=1}^3 F_{A,i} \right)}_{R_{S,A,parallel}} \underbrace{\left(1 - \prod_{i=1}^2 F_{B,i} \right)}_{R_{S,B,parallel}} (1 - F_C) = \underbrace{\left(1 - \left(\frac{1}{4} \right)^3 \right) \left(1 - (4\%)^2 \right) (1 - 50\%)}_{R_A \cdot R_B \cdot R_C} = 41.34\% \quad (74.240)
 \end{aligned}$$

T3. Topologie k/n:

Ce système fonctionne si k parmi n composants de même loi de fiabilité R fonctionnent. C'est typiquement le cas des disques RAID en informatique pour lesquels il en faut plus d'un qui fonctionne toujours à la fin et ce nombre est déterminé par le volume de données. Il s'agit aussi principe de fonctionnement des système de vote à redondance dans les avions où nous avons un système 2/3 appelé "**triplex**".

Nous avons alors la représentation schématique suivante:

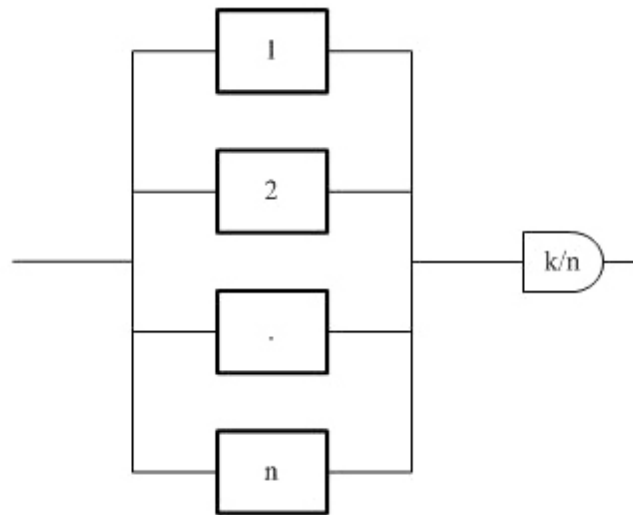
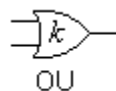


Figure: 74.20 - Topologie k/n

Dans le cadre des arbres de défaillance probabilistes (cf. chapitre Techniques De Gestion), cette topologie correspond à une porte k -OU (Voting OR) avec n entrées car il faut au moins que k blocs soient en panne pour que la sortie ne fonctionne plus:

Figure: 74.21 - Porte k -OU (Voting OR)

Donc chercher la loi de probabilité cumulée de fiabilité revient à se poser la question de la probabilité cumulée d'avoir k éléments qui fonctionnent parmi n qui sont non distinguables.

Cela revient donc à utiliser la loi binomiale (cf. chapitre de Statistiques) pour laquelle nous avons démontré que la probabilité cumulée était donnée alors par:

$$R_S = \sum_{x=k}^n C_x^n R^x (1-R)^{n-x} \quad (74.241)$$

Cette relation est extrêmement importante dans la pratique de l'ingénierie de la fiabilité!!! Nous remarquons ainsi que pour k égalant n nous retrouvons la fiabilité d'un système série dont les éléments ont tous la même fiabilité:

$$R_S = \sum_{x=n}^n C_x^n R^x (1-R)^{n-x} = C_n^n R^n (1-R)^{n-n} = \frac{n!}{n!(n-n)!} R^n (1-R)^{n-n} = R^n \quad (74.242)$$

Et pour k égalant 1, nous retrouvons la fiabilité d'un système parallèle dont les éléments ont tous la même fiabilité (revoir les propriétés du coefficient binomial dans le chapitre de Calcul Algébrique si nécessaire):

$$\begin{aligned} R_S &= \sum_{x=1}^n C_x^n R^x (1-R)^{n-x} = \sum_{x=0}^n C_x^n R^x (1-R)^{n-x} - C_0^n R^0 (1-R)^{n-0} \\ &= (R + (1-R))^n - C_0^n (1-R)^n = 1^n - (1-R)^n = 1 - (1-R)^n \end{aligned} \quad (74.243)$$

Dans le cas du triplex, nous avons donc:

$$\begin{aligned}
 R_S &= \sum_{x=2}^3 C_x^3 R^x (1-R)^{3-x} = \sum_{x=2}^3 \frac{3!}{x!(3-x)!} R^x (1-R)^{3-x} \\
 &= \frac{3!}{2!(3-2)!} R^2 (1-R)^{3-2} + \frac{3!}{3!(3-3)!} R^3 (1-R)^{3-3} \\
 &= \frac{3!}{2!1!} R^2 (1-R)^1 + \frac{3!}{3!0!} R^3 (1-R)^0 = 3R^2 (1-R) + R^3 = 3R^2 - 3R^3 + R^3 \\
 &= 3R^2 - 2R^3
 \end{aligned} \tag{74.244}$$

Par exemple, si la fiabilité est donnée par une loi exponentielle où nous avons déjà démontré que:

$$R(t) = e^{-\lambda t} = e^{-\frac{1}{MTBF}t} \tag{74.245}$$

Nous avons alors:

$$MTBF = \int_0^{+\infty} (3R^2(t) - 2R^3(t)) dt = \int_0^{+\infty} (3e^{-2\lambda t} - 2e^{-3\lambda t}) dt = \frac{5}{6\lambda} < \frac{1}{\lambda} \tag{74.246}$$

Donc la MTBF globale du système 2/3 est plus grande qu'un système simple ce qui évidemment est le but recherché.

T4. Topologie série/parallèle et parallèle/série à configuration symétrique:

Ce sont simplement des compositions simples des deux premiers systèmes étudiées précédemment.

Nous avons d'abord le système série/parallèle:

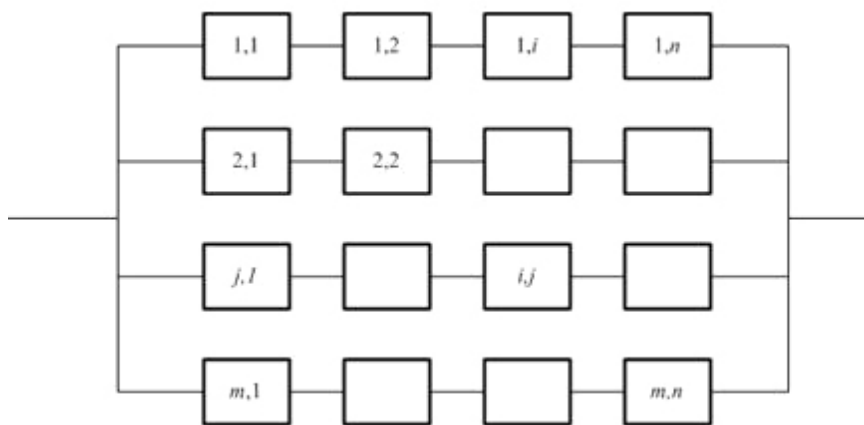


Figure: 74.22 - Topologie série/parallèle symétrique

Or, comme les systèmes séries sont donnés par:

$$R_S = 1 - F_S = \prod_{i=1}^n R_i \tag{74.247}$$

et les parallèles par:

$$R_S = 1 - \prod_{i=1}^n F_i = 1 - \prod_{i=1}^n (1 - R_i) \quad (74.248)$$

la composition des deux donne trivialement dans notre cas ci-dessus:

$$R_S = 1 - \prod_{j=1}^m (1 - R_j) = 1 - \prod_{j=1}^m \left(1 - \prod_{i=1}^n R_{i,j} \right) \quad (74.249)$$

Et nous avons dans la famille aussi le système parallèle/série:

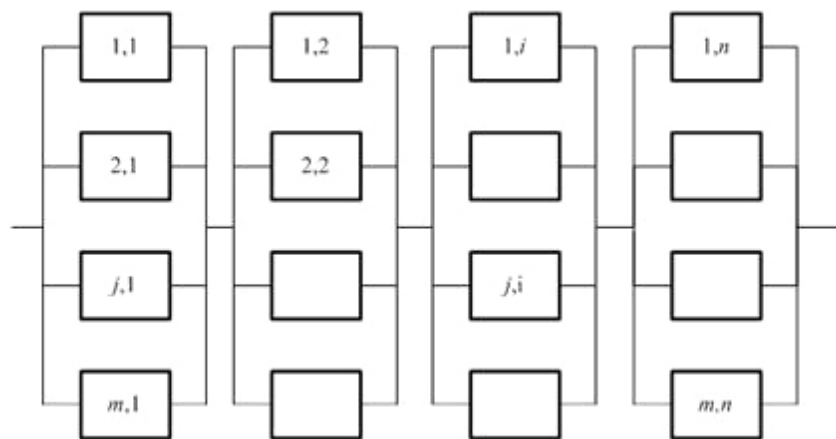


Figure: 74.23 - Topologie parallèle/série symétrique

où en utilisant exactement la même démarche que précédemment nous avons:

$$R_S = \prod_{i=1}^n R_i = \prod_{i=1}^n \left(1 - \prod_{j=1}^m (1 - R_{i,j}) \right) \quad (74.250)$$

T5. Topologie complexe:

Au fait, il ne s'agit pas vraiment de systèmes complexes mais ils nécessitent simplement une petite maîtrise des axiomes de probabilités. L'exemple particulier qui va nous intéresser est le suivant (typiquement filtre RLC en cascade):

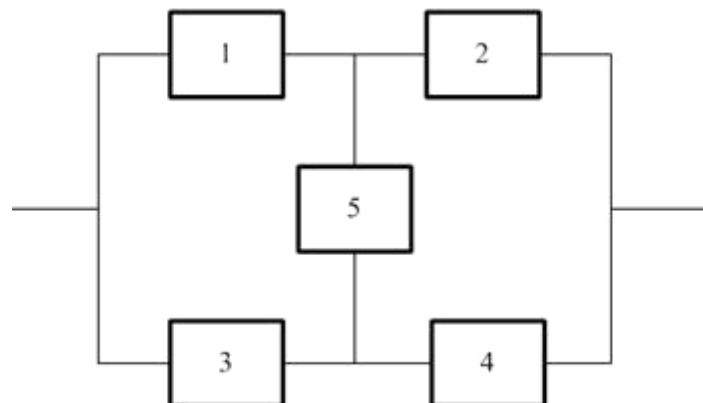


Figure: 74.24 - Topologie complexe

dénoté "réseau avec bridge".

Et nous devinons que ce qui rend le système complexe est le composant 5. Nous pouvons alors considérer une première approche qui est de décomposer le problème.

Le système par rapport au composant 5 sera soit dans l'état:

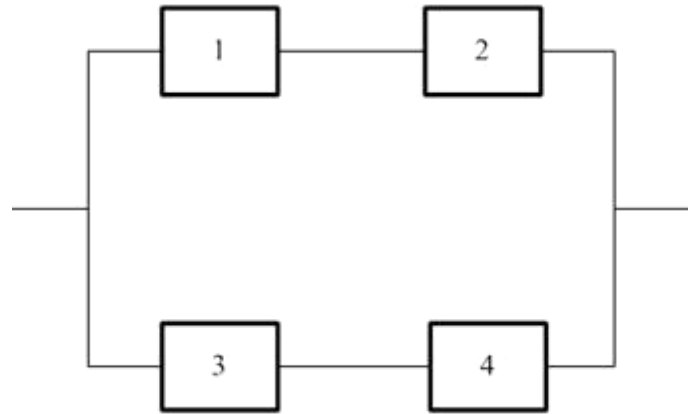


Figure: 74.25 - Décomposition du système (première étape)

s'il est défectueux avec une loi de densité probabilité:

$$F_5 = 1 - R_5 \quad (74.251)$$

et ayant lui-même une fiabilité de:

$$R_{S1} = 1 - \prod_{j=1}^2 \left(1 - \prod_{i=1}^2 R_{i,j} \right) = 1 - (1 - R_1 R_2)(1 - R_3 R_4) \quad (74.252)$$

selon nos résultats précédents du système complexe série/parallèle.

Soit dans l'état suivant s'il est fonctionnel avec une loi de densité probabilité R_5 :

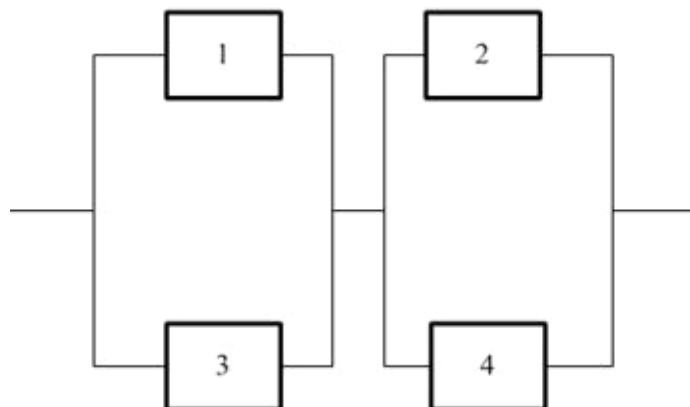


Figure: 74.26 - Décomposition du système (deuxième et dernière étape)

et ayant lui-même une fiabilité de:

$$R_{S2} = \prod_{i=1}^2 \left(1 - \prod_{j=1}^2 (1 - R_{i,j}) \right) = (1 - (1 - R_1)(1 - R_3))(1 - (1 - R_2)(1 - R_4)) \quad (74.253)$$

selon nos résultats précédents du système complexe parallèle/série.

Comme le système ne peut pas être dans les deux états en même temps, nous avons affaire à une probabilité disjointe (cf. chapitre de Probabilités) soit la somme des densités auxquelles nous devons associer les autres composants. Dès lors nous avons:

$$R_S = (1 - R_5) [1 - (1 - R_1 R_2)(1 - R_3 R_4)] + R_5 [(1 - (1 - R_1)(1 - R_3))(1 - (1 - R_2)(1 - R_4))] \quad (74.254)$$

Nous pouvons alors à l'aide de l'ensemble des 5 topologies précédentes faire une évaluation de la fiabilité d'un système quelconque!

Une autre stratégie pour les systèmes complexes consiste à les décomposer en système séries ou parallèles simples. Si cela n'est pas possible, nous pouvons calculer la fiabilité de chaque configuration du système qui est considérée comme fonctionnant et faire ensuite la somme des probabilités de fonctionnement de chaque configuration.

Faisons un exemple en considérant le cas particulier suivant:

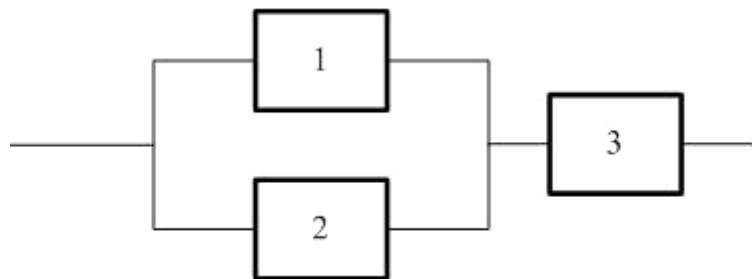


Figure: 74.27 - Cas particulier de système supposé comme complexe

Et considérons la table de vérité suivante avec les 2^n configurations possibles:

État n°	Bloc 1	Bloc 2	Bloc 3	Sortie
1	0	0	0	0
2	1	0	0	0
3	0	1	0	0
4	1	0	1	1
5	0	1	1	1
6	1	1	1	1
7	1	1	0	0
8	0	0	1	0

Tableau: 74.9 - Table de vérité d'un système de maintenance préventive

Le principe (il faut le savoir...) consiste à dire qu'un état UP (valant donc: 1) est affecté à la valeur R_i du bloc i concerné et que l'état DOWN (valant donc: 0) est affecté à la valeur $1 - R_i$ du bloc i . Chaque valeur sera multipliée avec les autres pour obtenir la fiabilité totale du système à un état donné.

Ainsi, dans l'exemple précédent, nous avons donc que 3 états qui permettent au système de fonctionner. Calculons leur fiabilité respective. Les états n°4 et n°5 donnent donc par définition:

$$R_{S6} = R_1 R_2 R_3 \quad (74.255)$$

L'état n°6 donne lui par définition:

$$R_{S6} = R_1 R_2 R_3 \quad (74.256)$$

et nous sommes le tout comme indiqué ultérieurement:

$$\begin{aligned} R_S &= R_{S4} + R_{S5} + R_{S6} = R_1 R_3 (1 - R_2) + R_2 R_3 (1 - R_1) + R_1 R_2 R_3 \\ &= R_1 R_3 - R_1 R_3 R_2 + R_2 R_3 - R_2 R_3 R_1 + R_1 R_2 R_3 = R_1 R_3 - R_1 R_3 R_2 + R_2 R_3 \quad (74.257) \\ &= R_1 R_3 + R_2 R_3 - R_1 R_3 R_2 \end{aligned}$$

Et nous pouvons vérifier que cette approche est effectivement correcte en prenant la relation générale d'un tel système démontrée plus haut:

$$\begin{aligned} R_S &= \prod_{i=1}^n R_i = \prod_{i=1}^n \left(1 - \prod_{j=1}^m (1 - R_{i,j}) \right) = \left(1 - \prod_{j=1}^2 (1 - R_{1,j}) \right) \left(1 - \prod_{j=1}^1 (1 - R_{2,j}) \right) \\ &= \left(1 - \prod_{j=1}^2 (1 - R_{1,j}) \right) (1 - (1 - R_{2,1})) = \left(1 - \prod_{j=1}^2 (1 - R_{1,j}) \right) R_3 \quad (74.258) \\ &= (1 - (1 - R_1)(1 - R_2)) R_3 = R_3 - R_3 (1 - R_1)(1 - R_2) = R_3 - (R_3 - R_3 R_1)(1 - R_2) \\ &= R_3 - (R_3 - R_3 R_1 - R_3 R_2 + R_3 R_1 R_2) = R_3 R_1 + R_3 R_2 - R_3 R_1 R_2 \end{aligned}$$

ce qui montre que nous avons bien le même résultat et que l'approche par décomposition fonctionne aussi.

Enfin, signalons pour terminer que lorsque nous incluons dans les systèmes des éléments qui permettent de tolérer ou d'accepter certaines erreurs, nous parlons alors de "**tolérance aux fautes**" et nous en distinguons principalement de trois types:

- Redondance active: Dans ce cas tous les composants redondants fonctionnent en même temps.
- Redondance passive: Un seul élément redondant fonctionne, les autres sont en attente, ce qui a pour avantage de diminuer ou de supprimer le vieillissement des éléments redondants, mais en contrepartie nécessite l'insertion d'un composant de détection de panne et de commutation.
- Redondance majoritaire: Cette redondance concerne surtout des signaux. Le signal de sortie sera celui de la majorité des composants redondants.

2.4.1. ARBRES DE DÉFAILLANCES PROBABILISTES

Un arbre de défaillances (aussi appelée arbre de pannes ou arbre de fautes) ou "fault tree analysis" en anglais est une technique d'ingénierie très utilisée dans les études de sécurité et de fiabilité des systèmes consistant à représenter graphiquement les combinaisons possibles d'événements qui permettent la réalisation d'un événement indésirable prédéfini. Une telle représentation graphique met donc en évidence les relations de cause à effet. Cette technique est complétée par un traitement mathématique qui permet la combinaison de défaillances simples ainsi que de leur probabilité d'apparition. Elle permet ainsi de quantifier la probabilité d'occurrence d'un événement indésirable.

Le principe est très simple tant que l'on fait appel à des probabilités binaires, par contre dès qu'il faut utiliser des fonctions de distributions continues (cas le plus fréquent dans la pratique), il faut passer par des simulations de Monte-Carlo (cf. chapitre de Méthodes Numériques) avec typiquement des logiciels comme Weibull++.

Considérons l'arbre de défaillance ci-dessous qui comporte donc 5 événements primitifs et 3 événements combinés avec les probabilités des événements primitives supposés connus:

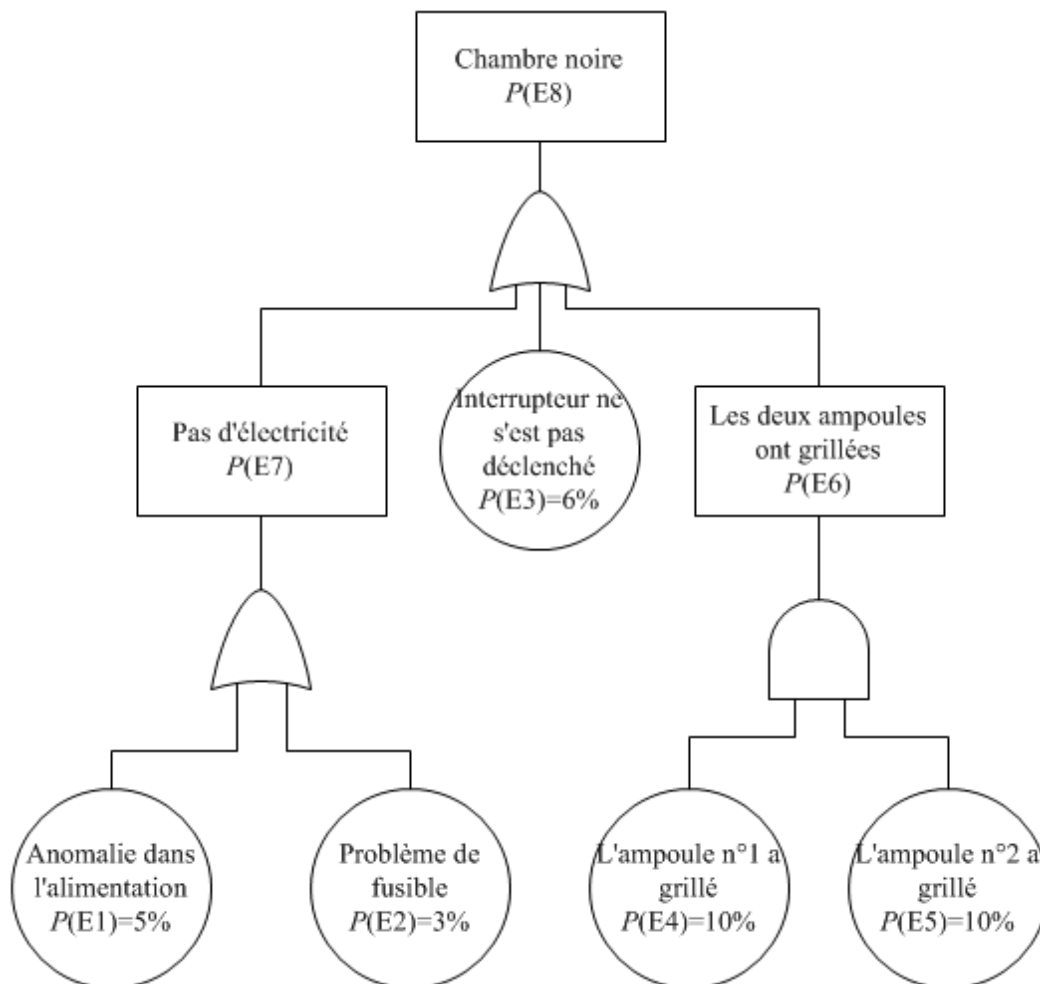


Figure: 74.28 - Arbre de défaillance simple à probabilités binaires et portes ET/OU seules

Nous avons alors pour l'événement E6 la probabilité suivante puisqu'il s'agit d'une porte ET (assimilable à un système parallèle):

$$P(E_6) = F_6 = \prod_{i=4}^5 F_i = P(E_4) P(E_5) = 0.1 \cdot 0.1 = 0.01 = 1\% \quad (74.259)$$

et pour l'événement E7 puisqu'il s'agit d'une porte OU (assimilable à un système série), nous avons.

$$\begin{aligned} P(E_7) = F_7 &= 1 - \prod_{i=1}^2 (1 - F_i) = 1 - (1 - P(E_1))(1 - P(E_2)) \\ &= 1 - (1 - 0.05) \cdot (1 - 0.03) = 0.0785 = 7.85\% \end{aligned} \quad (74.260)$$

Et donc au final:

$$\begin{aligned} P(E_8) = F_8 &= 1 - (1 - F_3)(1 - F_6)(1 - F_7) = 1 - (1 - P(E_3))(1 - P(E_6))(1 - P(E_7)) \\ &= 1 - (1 - 0.06)(1 - 0.0785)(1 - 0.01) = 0.1424 = 14.24\% \end{aligned} \quad (74.261)$$

Ainsi, la probabilité d'avoir la chambre de travail sans lumière est de 14.24%.

2.5. MÉTHODE DU MAXIMUM DE VRAISEMBLANCE

Nous avons étudié dans le chapitre de Statistiques que l'utilisation de la technique du maximum de vraisemblance permettait d'estimer la valeur maximale ou minimale des paramètres d'une loi statistiques sous les hypothèses d'indépendance des événements.

Cette technique, bien qu'empirique et tout à fait discutable, peut être utilisée dans le domaine de la fiabilité. Comme la théorie a déjà été fait dans le chapitre de Statistiques, voyons de suite des cas pratiques.

Exemples:

E1. Considérons un système dont les intervalles de temps entre deux pannes est distribué selon un loi exponentielle:

$$f(t) = \frac{dF(t)}{dt} = \lambda e^{-\lambda t} \quad (74.262)$$

de paramètre λ inconnu et où nous considérerons (comme toujours...) que chaque panne est indépendante. Des observations nous ont donné le nombre de mois suivants entre chaque panne: 10, 5, 11, 12, 15.

En utilisant alors le maximum de vraisemblance qui consiste donc à maximiser la probabilité totale d'événements indépendants (donc il faut multiplier), nous avons:

$$L_5(\lambda) = \lambda e^{-10\lambda} \cdot \lambda e^{-5\lambda} \cdot \lambda e^{-11\lambda} \cdot \lambda e^{-12\lambda} \cdot \lambda e^{-15\lambda} = \lambda^5 e^{-53\lambda} \quad (74.263)$$

Donc déterminer le taux de défaillance qui maximise la vraisemblance consiste à calculer quand la dérivée s'annule:

$$\frac{dL_5(\lambda)}{d\lambda} = \frac{d\lambda^5 e^{-53\lambda}}{d\lambda} = (5\lambda^4 - 53\lambda^5) e^{-53\lambda} = \lambda^4 (5 - 53\lambda) e^{-53\lambda} = 0 \quad (74.264)$$

Nous avons alors deux racines triviales qui sont:

$$\begin{aligned}\lambda_1 &= 0 \\ \lambda_2 &= \frac{5}{53}\end{aligned}\quad (74.265)$$

et évidemment nous allons conserver la deuxième racine comme taux de défaillance de notre système.

E2. Nous considérons les mêmes données qu'avant mais avec cette fois-ci une loi de Weibull avec la notation simplifiée suivante:

$$f(x) = \frac{\alpha}{\beta^\alpha} x^{\alpha-1} \exp\left(-\left(\frac{x}{\beta}\right)^\alpha\right) = \alpha \lambda x^{\alpha-1} \exp(-\lambda x^\alpha) \quad (74.266)$$

et donc:

$$\begin{aligned}L(\alpha, \lambda) &= f(10)f(5)f(11)f(12)f(15) \\ &= \left[\alpha \lambda 10^{\alpha-1} \exp(-\lambda 10^\alpha)\right] \left[\alpha \lambda 5^{\alpha-1} \exp(-\lambda 5^\alpha)\right] \left[\alpha \lambda 11^{\alpha-1} \exp(-\lambda 11^\alpha)\right] \\ &\quad \left[\alpha \lambda 12^{\alpha-1} \exp(-\lambda 12^\alpha)\right] \left[\alpha \lambda 15^{\alpha-1} \exp(-\lambda 15^\alpha)\right] \\ &= \alpha^5 \lambda^5 10^{\alpha-1} 5^{\alpha-1} 11^{\alpha-1} 12^{\alpha-1} 15^{\alpha-1} \exp(-\lambda (10^\alpha + 5^\alpha + 11^\alpha + 12^\alpha + 15^\alpha))\end{aligned}\quad (74.267)$$

Pour chercher à maximiser ce genre de fonction formellement il n'existe pas de méthodes connues (du moins à ma connaissance...). Par contre, un tableur comme MS Excel avec son solveur permet de trouver les deux paramètres de cette loi de Weibull en trois clics!

E3. Nous supposons que nous mettons sous observation 10 éléments de fabrication identique. Parmi les 10, nous avons observé un arrêt de 4 d'entre eux après 700, 800, 900 et 950 heures. Les 6 éléments restants seront considérés comme dépassant les 1'000 heures mais nous arrêtons l'observation à partir de cette valeur. En supposant que la loi de défaillance est à nouveau exponentielle, nous avons alors (l'idée est subtile mais simple):

$$\begin{aligned}L_{10}(\lambda) &= f(700)f(800)f(900)f(950) \\ &= \lambda e^{-700\lambda} \lambda e^{-800\lambda} \lambda e^{-900\lambda} \lambda e^{-950\lambda} \left(1 - (1 - e^{-1'000\lambda})\right)^6 \\ &= \lambda^4 e^{-3'350\lambda} e^{-6'000\lambda} = \lambda^4 e^{-9'350\lambda}\end{aligned}\quad (74.268)$$

Il vient alors:

$$\frac{dL_{10}(\lambda)}{d\lambda} = \frac{d\lambda^4 e^{-9'350\lambda}}{d\lambda} = (4\lambda^3 - 9'350\lambda^4) e^{-9'350\lambda} = \lambda^3 (4 - 9'350\lambda) e^{-9'350\lambda} = 0 \quad (74.269)$$

Nous avons alors deux racines triviales qui sont:

$$\lambda_1 = 0$$

$$\lambda_2 = \frac{4}{9'350} \quad (74.270)$$

et évidemment nous allons conserver la deuxième racine comme taux de défaillance de notre système.

2.6. MODÈLE DE SURVIE DE KAPLAN-MEIER

Dans des domaines comme l'industrie, la médecine ou en biologie, nous intéressons souvent aux:

- Durée de survie après un événement grave
- Durée de rémission après un traitement ou une opération
- Durée d'un symptôme après une anomalie
- Durée d'une infection sans symptôme

Nous cherchons très souvent à distinguer au moins "l'événement d'intérêt":

- Arrêt du système après l'événement grave
- Fin de la rémission
- Fin d'un symptôme après anomalie
- Début d'un symptôme lors d'une infection

de la variable à expliquer "durée avant l'apparition de l'événement d'intérêt":

- Temps écoulé avant l'arrêt du système
- Temps écoulé avant la fin de la rémission
- Temps écoulé avant la fin du symptôme
- Temps écoulé sans symptôme

Définitions:

D1. Nous appelons "rémission", la diminution d'une maladie ou d'un dysfonctionnement de façon temporaire.

D2. La "durée de survie" ou "durée de vie" T désigne le temps qui s'écoule depuis un instant initial (début d'un traitement, diagnostic, panne, etc.) jusqu'à la survenue d'un événement d'intérêt final (décès du patient, rechute, rémission, guérison, réparation, etc.). Nous disons que l'objet de l'étude "survit au temps t " si, à cet instant l'évènement d'intérêt final n'a pas encore eu lieu.

Remarque: Bien que ce type d'étude puisse être associé à la maintenance préventive, les statisticiens l'associent plutôt au domaine de "l'analyse de la survie".

Nous nous intéresserons dans le cas présent à un cadre particulier mais qui peut être facilement généralisé:

- Cohorte/Essai clinique où nous étudions la durée de survie de chaque patient (machine).
- Tous les patients (machines) n'ont pas la même durée d'observation (instants différents d'entrée dans l'étude).
- Nous avons des informations sur le temps de survie de chaque patient (machine) mais nous ne savons pas quand il arrive exactement.

Des deux derniers points, nous concluons que le temps de survie peut être censuré. Donc les techniques statistiques habituelles ne s'appliquent pas directement car les données censurées demandent un traitement particulier (bien évidemment, nous enlevons les données censurées nous perdons de l'information). Il va sans dire que ce type de situation est extrêmement fréquent et donc l'étude de l'estimation de Kaplan-Meier est très importante.

Définition: La durée T est dite censurée si la durée n'est pas intégralement observée. Les différents types de censures sont:

- La censure de type I: fixée à droite. Dans cette situation, la durée n'est pas observable au-delà d'une durée maximale, fixe appelée "**censure fixe**" et imposée. Donc, soit nous avons l'opportunité d'observer la vraie durée de l'événement d'intérêt pour l'élément s'il a lieu avant la censure fixe, soit nous nous contenterons de la durée de la censure fixe s'il l'événement d'intérêt n'a pas eu lieu avant.
- La censure de type I: fixée à gauche. Dans cette situation, la durée d'étude n'est pas observable avant une date connue appelée "**censure d'attente**" et imposée. Donc soit, l'événement d'intérêt a lieu au moment de la censure d'attente, soit après. S'il y a lieu après, nous considérerons la durée entre la date de la censure et la date de l'événement d'intérêt.
- La censure de type II: attente. Dans cette situation, nous observons les durées de vie de n individus jusqu'à ce que m individus aient vu l'événement se produire (décédés). La durée considérée est donc celle du début de l'expérience jusqu'à l'événement d'intérêt pour le m -ème.
- La censure de type III: aléatoire à gauche. Dans cette situation, nous ne connaissons pas quand l'événement d'intérêt a eu lieu (car nous avons commencé à observer le sujet d'étude trop tard). Nous ne pouvons alors pas traiter des "durées" dans le sens mesurable du terme et il nous faudra nous limiter à un simple comptage.
- La censure de type III: aléatoire à droite. Dans cette situation, nous ne connaissons pas quand l'événement aura lieu (car nous avons arrêté d'observer le sujet avant qu'il y ait lieu pour une raison quelconque). Nous ne pouvons alors pas traiter des "durées" dans le sens mesurable du terme et il nous faudra nous limiter à un simple comptage.
- La censure de type IV: aléatoire par intervalle. Dans cette situation, nous avons un mélange de la censure aléatoire à gauche et à droite. C'est-à-dire que pour certains sujets d'étude, nous ne savons pas quand l'événement d'intérêt a commencé, et pour d'autres, nous ne savons pas quand il aura lieu (s'il a lieu....).

Dans l'industrie des machines, nous savons souvent affaire à la censure de type I: fixée droite. Dans le domaine médical, lors d'essais cliniques, nous avons souvent affaire à une censure aléatoire à droite. Dans le cas des pandémies, nous avons affaire à des censure de type aléatoire à gauche.

Supposons que l'étude soit un essai clinique portant sur deux groupes de patients recevant deux types de traitements. Deux questions importantes se posent aux médecins:

Q1. L'un des deux traitements est-il plus efficace que l'autre en matière d'amélioration de la survie des patients?

Q2. Peut-on mettre en évidence des facteurs pronostiques, c'est-à-dire qui améliorent/détériorient la survie?

Pour répondre à la question Q1 nous pouvons mettre en place des méthodes statistiques qui vont permettre de comparer les deux groupes de patients qui reçoivent les deux types de traitement.

Pour répondre à la question Q2 nous pouvons proposer un modèle qui relie la durée de survie des patients à des variables explicatives et mettre en évidence des facteurs pronostiques.

Accompagnons la théorie directement d'un exemple et considérons le tableau suivant où deux cohortes de patients (nous passons de l'ingénierie mécanique à l'ingénierie humaine...) de même taille initiale atteints de leucémie aiguë testent un médicament (6-MP) contre un placebo de façon bien évidemment aveugle.

Nous avons les durées de rémission suivantes pour 21 patients (le tableau indique donc le nombre de semaines pendant lesquelles un patient est considéré comme guéri après traitement avant de rechuter):

6-mercaptopurine (6-MP)	Placebo
6	1
6	1
6	2
6+	2
7	3
9+	4
10	4
10+	5
11+	5
13	8
16	8
17+	8
19+	8
20+	11
22	11
23	12
25	12
32+	15
32+	17
34+	22
35+	23

Tableau: 74.10 - Analyse de survie de deux cohortes avec censure à droite

Le signe + correspond à des patients qui ont quitté l'étude à la semaine considérée. Ils sont donc censurés. Par exemple, le 4ème patient est perdu de vue pour une raison quelconque au bout de 6 semaines de traitement avec le 6-MP: il a donc une durée de rémission supérieure à 6 semaines. Donc dans l'étude 6-MP, il y a 21 patients avec et 12 données censurées.

Remarque: Le modèle théorique suppose que la censure est indépendant du temps de survie (censure non informative). Mais si la censure est due à l'arrêt du traitement, l'hypothèse d'indépendance n'est pas valide.

Pour le groupe placebo il est simple de faire une courbe de survie. Il suffit de produire le tableau suivant (pour les semaines omises, on impose bien évidemment le nombre de rémission comme constantes):

Semaine i	Nombre de rémission à la semaine i	Proportion (probabilité) de rémission à la semaine i
0	21	100%
1	19	19/21=90%
2	17	17/21=81%
3	16	16/21=76%
4	14	14/21=67%
5	12	12/21=57%
8	8	8/21=38%
11	6	6/21=29%
12	4	4/21=19%
15	3	3/21=14%
17	2	2/21=10%
22	1	1/20=0.05%
23	0	0

Donc si les données ne sont pas censurées, la survie $S(t)$ peut être estimée par la proportion d'individus ayant survécu à l'instant t , qu'il est d'usage d'écrire sous la forme mathématique suivante:

$$\hat{S}_k(t) = P_k(T_k > t) = \frac{1}{n} \sum_{i=1}^k 1_{T_i > t} \quad (74.271)$$

L'idée est donc d'estimer:

$$P(T > t) = P(\text{probabilité de survivre jusqu'au temps } t) \quad (74.272)$$

par la la proportion de patients ayant survécu jusqu'au temps t .

Si les données sont censurées, l'estimation de la fonction de survie nécessite des outils spécifiques. Kaplan et Meier ont proposé dans ce cas particulier le calcul suivant:

$$\begin{aligned}
 &P(\text{\textit{être en rémission à la } i\text{-ème semaine}}) \\
 &= P(\text{\textit{ne pas rechuter à la } i\text{-ème semaine}}) \cdot P(\text{\textit{être en rémission à la } (i-1)\text{-ème semaine}}) \\
 &\quad (74.273)
 \end{aligned}$$

Voyons cela sous une forme un peu plus mathématique. Si nous notons $X(1) \leq X(2) \leq \dots \leq X(n)$ les instants (ordonnées) où un évènement s'est produit (décès ou censure), nous avons alors:

$$P(T > X(k)) = P(T > X(k) | T > X(k-1)) \cdot P(T > X(k-1)) \quad (74.274)$$

Avec bien évidemment:

$$P(T > 0) = 1 \quad (74.275)$$

Nous estimons:

$$\hat{P}_k = \left(1 - \frac{d_k}{R_k}\right) \quad (74.276)$$

où d_k est le nombre de décès observé au temps correspondant à l'évènement $X(k)$ et R_k est le nombre d'individus à risque (exposés au risque de décès) juste avant $X(k)$.

Nous définissons l'estimateur de Kaplan-Meier pour tout $X(0) \leq t < X(k)$:

$$\hat{S}_{KM}(t) = \prod_{i=1}^k \hat{P}_k = \prod_{i=1}^k \left(1 - \frac{d_k}{R_k}\right)$$

Durée de rémission (non censurées) observées	Sujets en rémission au début de k	Probabilité de ne pas rechuter à k sachant qu'on est en rémission à $k-1$	Probabilité de survie de Kaplan-Meier
0	21	21/21=100%	100%
6	21	18/21=85.7%	100%*85.7%=85.7%
7	17	16/17=94.1%	85.7%*94.1%=80.7%
10	15	14/15=93.3%	80.7%*93.3%=75.3%
13	12	11/12=91.7%	75.3%*91.7%=69%
16	11	10/11=90.9%	69%*90.9%=62.7%
22	7	6/7=85.7%	62.7%*85.7%=53.8%
23	6	5/6=83.3%	53.8%*83.3%=44.8%

Nous retrouvons donc les mêmes valeurs que celles données par exemple par le logiciel de statistiques Minitab 15.1.1.

2.7. MÉTHODE ABC

Dans une entreprise, les tâches sont nombreuses et les équipes d'entretien et de maintenance sont systématiquement réduites à leur minimum par les temps qui courent. De plus, les technologies les plus évoluées en matière de maintenance coûtent cher ou peuvent coûter très cher, et ne doivent pas être appliquées sans discernement.

Il convient par conséquent de s'organiser de façon efficace et rationnelle. L'analyse ABC (on devrait dire en toute rigueur "objectif ABC"), utilisant implicitement la loi de probabilité cumulée de Pareto (cf. chapitre de Statistiques), permet d'y remédier relativement bien. Ainsi, un classement des coûts par rapport aux types de panne donne des priorités sur les interventions à mener (cette méthode empirique est aussi utilisée dans de nombreux autres domaines dont un qui est très connu: la gestion de stocks).

L'idée consiste dans un premier temps comme pour l'analyse de Pareto (cf. chapitre de Techniques de Gestion) de classer les pannes par ordre croissant de coût de maintenance (ou de coût d'impact en cas de panne) chaque panne se rapportant un système simple ou complexe et à établir un graphique faisant correspondre les pourcentages de coûts cumulés aux pourcentages de type de panne cumulés.

Ensuite, nous distinguons par tradition dans l'industrie trois zones:

- Zone A: Dans la majorité des cas, environ 20% des pannes représentent 80% des coûts et il s'agit donc de la zone prioritaire.
- Zone B: Dans cette tranche, les 30% de pannes suivantes coûtent 15% supplémentaires.
- Zone C: 50% des pannes restantes ne revient qu'à 5% des coûts.

Le domaine du web participatif considère lui des zones (objectifs) de respectivement 90-9-1 pourcents donc on trouve vraiment de tout et n'importe quoi...

Voyons un exemple en considérant que les données suivantes ont été recueillies et que nous aimerions faire une analyse du % de machines sur lesquelles il faudrait que nous nous concentrions pour diminuer le coût d'heures de pannes d'environ 80% (dans la réalité on s'intéressera plus au % du coût financier!).

N° de Machine	Heures d'arrêt	Nb. de pannes
Machine n°1	100	4
Machine n°2	32	15
Machine n°3	50	4
Machine n°4	19	14
Machine n°5	4	3
Machine n°6	30	8
Machine n°7	40	12
Machine n°8	80	2
Machine n°9	55	3
Machine n°10	150	5
Machine n°11	160	4
Machine n°12	5	3
Machine n°13	10	8
Machine n°14	20	8

Tableau: 74.11 - Analyse ABC sur pannes machines

Ensuite, dans un tableur comme Microsoft Excel ou autre... nous pouvons aisément établir le tableau suivant:

Machine	Arrêt [h.]	Cumul [h.]	%Cumulé	Pannes	Cumul pannes	%Cumulé
11	160	160	21.19%	4	4	4.30%
10	150	310	41.06%	5	9	9.68%
1	100	410	54.30%	4	13	13.98%
8	80	490	64.90%	2	15	16.13%
9	55	545	72.19%	3	18	19.35%
3	50	595	78.81%	4	22	23.66%
7	40	635	84.11%	12	34	36.56%
2	32	667	88.34%	15	49	52.69%
6	30	697	92.32%	8	57	61.29%
14	20	717	94.97%	8	65	69.89%
4	19	736	97.48%	14	79	84.95%
13	10	746	98.81%	8	87	93.55%
12	5	751	99.47%	3	90	96.77%
5	4	755	100.00%	3	93	100.00%

Tableau: 74.12 - Normalisation des données pour analyse ABC

Nous avons alors graphiquement:

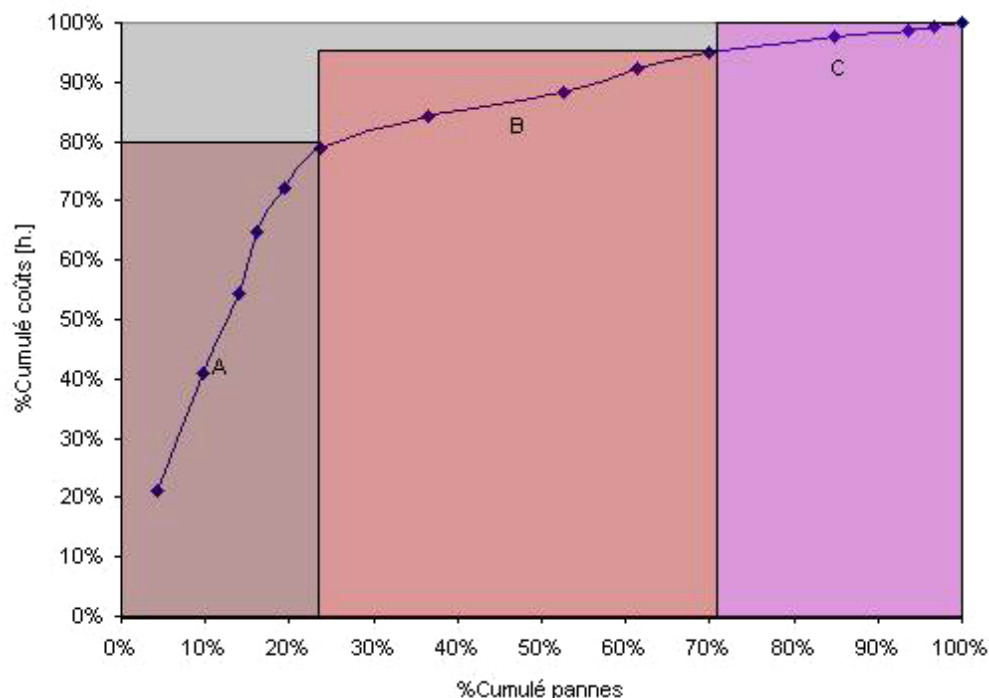


Figure: 74.29 - Graphique de la méthode ABC

où les zones A, B et C sont arrondies à des points existants. Ainsi, la zone critique A compte les machines 11, 10, 1, 8, 9 et 3. Une amélioration de la fiabilité de ces machines peut donc procurer jusqu'à 78.8% de gain de temps sur les pannes.

Maintenant déterminons les paramètres de la loi de répartition de Pareto:

$$P(X \leq x) = P_i = 1 - \left(\frac{x_m}{x_i} \right)^k \quad (74.277)$$

Nous devons déterminer x_m et k les autres paramètres nous sont donnés par nos mesures (le tableau).

Nous pouvons jouer de la manière suivante:

$$P_i - P_{i-1} = - \left(\frac{x_m}{x_i} \right)^k + \left(\frac{x_m}{x_{i-1}} \right)^k = x_m^k \left(\frac{1}{x_{i-1}} - \frac{1}{x_i} \right)^k \quad (74.278)$$

d'où:

$$\log(P_i - P_{i-1}) = \log \left(x_m^k \left(\frac{1}{x_{i-1}} - \frac{1}{x_i} \right)^k \right) = k \log(x_m) + k \log \left(\frac{1}{x_{i-1}} - \frac{1}{x_i} \right) \quad (74.279)$$

Donc à l'aide de:

$$\log(\Delta P) = k \log \left(\frac{1}{x_{i-1}} - \frac{1}{x_i} \right) + k \log(x_m) \quad (74.280)$$

on doit pouvoir déterminer les deux paramètres recherchés en considérant l'expression comme l'équation d'une droite dont k est la pente et $k \log(x_m)$ l'ordonnée à l'origine:

$$\log(\Delta P) = ax + b \quad (74.281)$$

Une régression linéaire simple (cf. chapitre de Méthodes Numériques) nous donne:

$$k = 0.6189 \text{ et } k \log(x_m) = -1.2234 \quad (74.282)$$

donc:

$$x_m = 10^{\frac{-1.2234}{k}} = 0.0105 \quad (74.283)$$

Nous avons donc au final:

$$P(X \leq x) = P_i = 1 - \left(\frac{x_m}{x_i} \right)^k = 1 - \left(\frac{0.0105}{x_i} \right)^{0.6189} \quad (74.284)$$

Ce qui donne alors le tableau suivant (les x_i de la loi de Pareto sont les %Cumulé de panne):

Machine	%Cumulé coûts [h.]	%Cumulé panne	%Cumulé loi Pareto
11	21.19%	4.30%	61.76%
10	41.06%	9.68%	78.01%
1	54.30%	13.98%	82.89%
8	64.90%	16.13%	84.48%
9	72.19%	19.35%	86.29%
3	78.81%	23.66%	88.05%
7	84.11%	36.56%	91.12%
2	88.34%	52.69%	93.08%
6	92.32%	61.29%	93.75%
14	94.97%	69.89%	94.29%
4	97.48%	84.95%	95%
13	98.81%	93.55%	95.32%
12	99.47%	96.77%	95.43%
5	100.00%	100.00%	95.53%

Tableau: 74.13 - Comparaisons données expérimentales/théoriques

Nous pouvons alors obtenir la vraie courbe de Pareto correspondante facilement dans Microsoft Excel 11.8346:

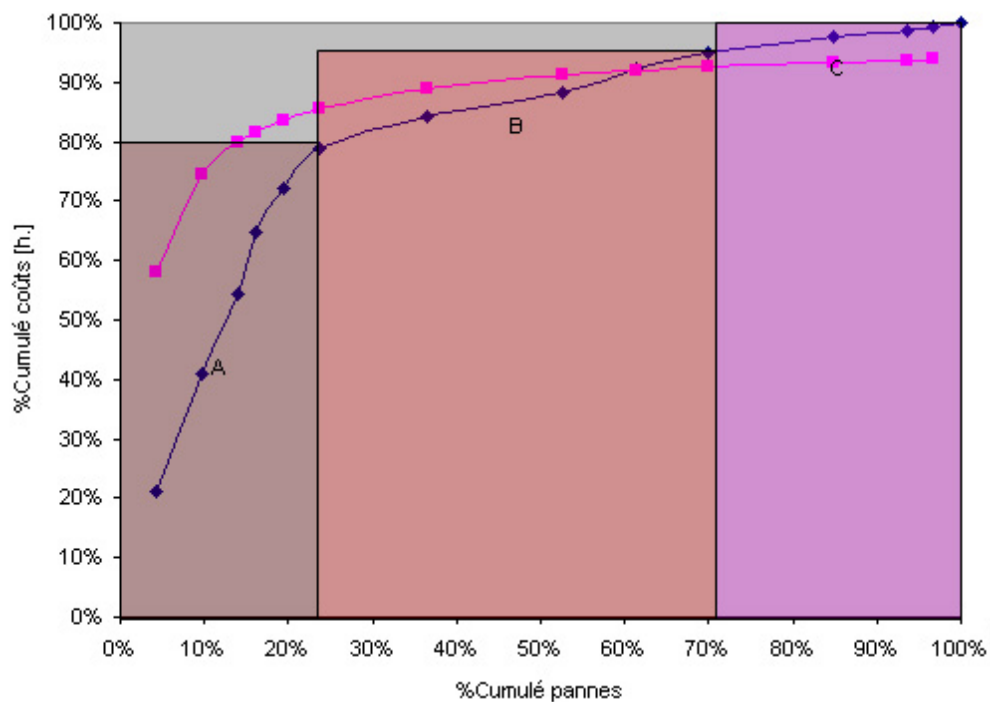


Figure: 74.30 - Graphique de la méthode ABC associé à une courbe de Pareto

La différence entre la courbe expérimentale et la théorique est non négligeable... Comme k est inférieur à 1, alors comme nous l'avons vu démontré dans le chapitre de Statistiques, la loi de Pareto n'a ni espérance, ni variance.

Il faut faire attention au fait que dans le domaine de la gestion, la loi de Pareto est utilisée un peu à tort et à travers alors qu'une autre loi de probabilité pourrait être beaucoup plus adaptée.

Par ailleurs, un simple test d'ajustement du Khi-deux (cf. chapitre de Statistiques) nous montre qu'il faut rejeter l'approximation par une loi de Pareto. Par ailleurs, des logiciels spécialisés comme @Risk rejettent l'approximation par Pareto au-delà des 20 meilleurs ajustements, le meilleur ajustement étant selon ce même logiciel une loi log-normale.

3. PLANS D'EXPÉRIENCE

Le comportement des produits industriels est généralement fonction de nombreux phénomènes, souvent dépendants les uns des autres. Pour prévoir ce comportement, le produit et les phénomènes sont modélisés, et des simulations sont effectuées. La pertinence des résultats des simulations dépend évidemment de la qualité des modèles.

En particulier, dans le cadre de la conception ou reconception d'un produit, les modèles font généralement intervenir un certain nombre de grandeurs physiques (paramètres) que l'on s'autorise à modifier. Le comportement des produits industriels est généralement fonction de nombreux phénomènes, souvent dépendants les uns des autres. Pour prévoir ce comportement, le produit et les phénomènes sont modélisés, et des simulations sont effectuées.

Or, ces essais sont coûteux, et ce d'autant plus que le nombre de paramètres à faire varier est important. En effet, la modification d'un paramètre peut par exemple exiger un démontage et un remontage du produit, ou bien la fabrication de plusieurs prototypes différents (cas d'une pièce produite en série), ou encore l'interruption de la production pour changer d'outil (cas d'un processus de fabrication)... Le coût d'une étude expérimentale dépend donc du nombre et de l'ordre des essais effectués.

L'idée consiste alors à sélectionner et ordonner les essais afin d'identifier, à moindres coûts, les effets des paramètres sur la réponse du produit. Il s'agit de méthodes statistiques faisant appel à des notions mathématiques simples le plus souvent. La mise en oeuvre de ces méthodes comporte trois étapes (mais mieux vaut se référer à la norme ISO 3534-3:1999 sur le sujet pour être plus rigoureux!):

1. Postuler un modèle de comportement du système (avec des coefficients pouvant être inconnus)
2. Définir un protocole d'expérience, c'est-à-dire une série d'essais ("**runs**" en anglais) permettant d'identifier les coefficients du modèle
3. Faire les essais, identifier les coefficients ("**constrasts**" en anglais) et conclure.

Les plans d'expériences ("**Design of Experiment**" (D.O.E.) en anglais abrégés PEX en français) permettent d'organiser au mieux les essais qui accompagnent une recherche scientifique ou des études industrielles. Ils sont applicables à de nombreuses disciplines et à toutes les industries à partir du moment où l'on recherche le lien qui existe entre une grandeur d'intérêt, y (quantité de rebus, défauts, détections, amplitude, etc.), et des variables x_i dans un but d'optimisation. Raison pour laquelle il existe des logiciels pour les traiter comme Minitab ou principalement JMP pour ne citer que les plus connus.

Indiquons également que les plans d'expérience sont un des piliers de la chimiométrie (outils mathématiques, en particulier statistiques, pour obtenir le maximum d'informations à partir des données chimiques).

Au cours des dernières années du 20ème siècle, l'application des plans d'expérience s'est développée, particulièrement en raison du fait reconnu que ceux-ci sont essentiels pour l'amélioration de la qualité des biens et des services. Bien que la maîtrise statistique de la qualité, les solutions managériales, les inspections, et autres outils de qualité remplissent également cette fonction, le plan d'expérience représente la méthodologie par excellence dans le cas d'un environnement de paramètres complexes, variables et interactifs. D'un point de vue historique, les plans d'expérience ont évolué et se sont développés dans le secteur de l'agriculture. La médecine a également bénéficié d'une longue histoire de plans d'expérience élaborés avec soin. Actuellement, les environnements industriels témoignent de bénéfices considérables de la méthodologie, en raison de la facilité d'initiation des efforts (logiciels d'application conviviaux), d'une meilleure formation, de défenseurs influents, et des nombreux succès obtenus grâce aux plans d'expérience.

Il existe trois grandes familles de plans d'expériences:

1. Les "**plans de criblages**": dont l'objectif est de découvrir les facteurs les plus influents sur une réponse donnée en un minimum d'expériences. C'est la plus simple des familles car proche de l'intuition expérimentale (elle est parfois considérée comme une sous-famille de la deuxième famille car elle fait abstraction des interactions des facteurs et se réduit donc à un modèle purement additif).
2. Les "**plans de modélisation**": dont l'objectif est de trouver la relation mathématique qui lie les réponses mesurées aux variables associées aux facteurs soit via une démarche mathématique analytique ou purement matricielle. Les plans factoriels complets et fractionnaires (2 niveaux par facteurs avec modèles linéaires) ainsi que les plans pour surfaces de réponse (au moins 3 niveaux par facteurs avec modèles du second degré) font partie de cette famille.
3. Les "**plans de mélange**": dont l'objectif est le même que la deuxième famille, mais où les facteurs ne sont pas indépendants et sont contraints (par exemple leur somme ou leur rapport doit être égale à une certaine constante).

La technique des plans d'expérience a pour objectif d'être scientifiquement plus avantageux que les techniques consistant à changer seulement un facteur à la fois (en laissant donc fixes les autres facteurs) car cette dernière ne permet pas de faire une étude statistique approfondie des erreurs et (surtout!) masque complètement les interactions.

Le principe général de la stratégie des plans d'expérience se base sur le fait que l'étude d'un phénomène peut, le plus souvent, être schématisé de la manière suivante: nous nous intéressons à une grandeur, y qui dépend d'un grand nombre de variables, $x_1, x_2, \dots, x_n$ (et leur ordre n'a pas d'influence... ce qui est par contre problématique en chimie...).

La modélisation mathématique consiste à trouver une fonction f telle que:

$$y = f(x_1, x_2, \dots) \quad (74.285)$$

qui prenne en compte l'influence de chaque facteur seul ou des facteurs combinés (interactions). Cette fonction est donc déterministe (la réponse dépend uniquement des facteurs sans aucune incertitude possible, ce qui revient à ignorer les bruits tels que les erreurs de mesure) et invariant (le comportement n'évolue pas au cours du temps).

Une méthode classique d'étude consiste en la mesure de la réponse y pour plusieurs valeurs de la variable x_i tout en laissant fixe la valeur des $(n-1)$ autres variables. Nous itérons alors cette méthode pour chacune des variables.

Ainsi, par exemple, si nous avons 8 variables et si l'on décide de donner 2 valeurs expérimentales à chacune d'elles, nous sommes conduits à effectuer $2^8 = 256$ expériences.

Ce nombre élevé dépasse les limites de faisabilité tant en temps qu'en coût. Il faut donc réduire le nombre d'expériences à effectuer sans pour autant perdre sur la qualité des résultats recherchés.

L'utilisation d'un plan d'expérience donne alors une stratégie dans le choix des méthodes d'expérimentation. Le succès des plans d'expériences dans la recherche et l'industrie est lié au besoin de compétitivité des entreprises: ils permettent une amélioration de la qualité et une réduction des coûts.

Remarque: La méthode des plans d'expériences a été mise au point au début du siècle, dans les années 1920, par Ronald A. Fisher (celui du Test de Fisher). Elle a pris un essor considérable avec le développement de l'informatique et la puissance de calcul qui l'accompagne.

Le traitement des résultats se fait enfin à l'aide de la régression linéaire simple ou multiple et l'analyse de variance.

Avec les plans d'expériences, le but est donc d'obtenir le maximum de renseignements (mais pas tous!) avec le minimum d'expériences (et donc le minimum de coût) dans le but de modéliser ou d'optimiser des phénomènes étudiés.

Un expérimentateur qui lance une étude s'intéresse à une grandeur qu'il mesure à chaque essai. Cette grandeur s'appelle la "**réponse**", c'est la grandeur d'intérêt. La valeur de cette grandeur dépend de plusieurs variables. Au lieu du terme "variable" on utilisera le mot "**facteur**". La valeur donnée à un facteur pour réaliser un essai est appelée "**niveau**". Lorsque l'on étudie l'influence d'un facteur, en général, on limite ses variations entre deux bornes (oui faut bien s'avoir s'arrêter un jour et être raisonnable...) appelées respectivement: "**niveau bas**" et "**niveau haut**".

Bien évidemment, lorsque nous avons plusieurs facteurs $x_1, x_2, \dots, x_n$ ceux-ci représentent un point dans $\mathbb{R}^n$ appelé alors "**espace expérimental**".

L'ensemble de toutes les valeurs que peut prendre le facteur entre le niveau bas et le niveau haut, s'appelle le "**domaine de variation du facteur**" ou plus simplement le "**domaine du facteur**". Nous avons l'habitude de noter le niveau bas par -1 et le niveau haut par +1 quand tous les facteurs n'ont que deux niveaux, sinon est d'usage de noter les niveaux de 1 à n (nombre de niveaux). Bien évidemment, en fonction de la notation choisie, il faut adapter les expressions du modèle mathématique en conséquence.

Donc par exemple pour un facteur ayant un domaine de variation compris entre un niveau haut de 20 [°C] correspond à +1 et un niveau bas de 5 [°C] correspondant à -1 nous devons à la fin de notre étude transformer toutes les valeurs expérimentales en "**unités centrées réduites**" dans lesquelles doivent être utilisées les x_i .

Ainsi, nous avons deux points d'entrée (20,5) et deux de sorties (+1,-1). Toute valeur intermédiaire est donnée simplement par l'équation de la droite:

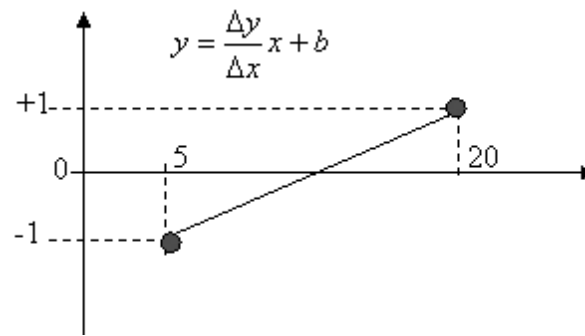


Figure: 74.31 - Principe de construction des unités centrées réduites

La pente est donc triviale à obtenir..., pour déterminer b , nous avons simplement une équation à une inconnue:

$$+1 = \frac{2}{(20-15)} 20 + b \Rightarrow b = +1 - \frac{2}{5} 20 = -1.\bar{6} \quad (74.286)$$

ou (ce qui revient au même):

$$-1 = \frac{2}{(20-15)} 5 + b \Rightarrow b = -1 - \frac{2}{5} 5 = -1.\bar{6} \quad (74.287)$$

Donc le passage de variables non normalisées, notées x , aux normalisées, notées X , s'écrit alors:

$$X = \frac{\Delta y}{\Delta x} x + b \quad (74.288)$$

soit dans un cas de sorties (+1,-1):

$$X = \frac{2}{\Delta x} x + \left(1 - \frac{2}{\Delta x}\right) x_{\max} \quad (74.289)$$

et inversement...:

$$x = (X - b) \frac{\Delta x}{\Delta y} \quad (74.290)$$

soit dans un cas de sorties (+1,-1):

$$x = \left(X - \left(1 - \frac{2}{\Delta x}\right) x_{\max} \right) \frac{\Delta x}{2} \quad (74.291)$$

Le regroupement des domaines de tous les facteurs définit le "**domaine d'étude**". Ce domaine d'étude est la zone de l'espace expérimental choisie par l'expérimentateur pour faire ses essais. Une étude, c'est-à-dire plusieurs expériences bien définies, est représentée par des points répartis dans le domaine d'étude.

Par exemple, pour deux facteurs, le domaine d'étude est une surface et l'espace expérimental est $\mathbb{R}^2$:

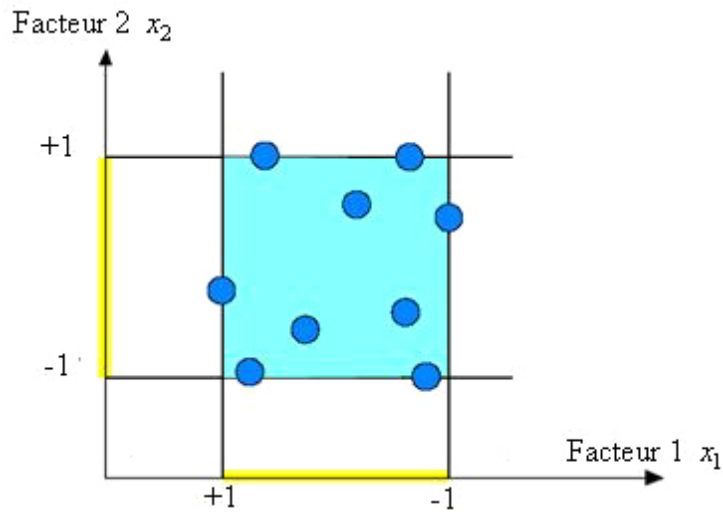


Figure: 74.32 - Exemple à deux facteurs

Les niveaux x_i centrés réduits représentent les coordonnées d'un point expérimental et y est la valeur de la réponse en ce point. On définit un axe orthogonal à l'espace expérimental et on l'attribue à la réponse. La représentation géométrique du plan d'expériences et de la réponse nécessite un espace ayant une dimension de plus que l'espace expérimental. Un plan à deux facteurs utilise un espace à trois dimensions pour être représenté: une dimension pour la réponse, deux dimensions pour les facteurs.

Lorsque nous avons déterminé le modèle mathématique, à chaque point du domaine d'étude correspond un ensemble de réponses qui se localisent sur une surface appelée la "**surface de réponse**" (raison pour laquelle ce domaine d'étude est parfois appelé: "plans d'expérience pour l'estimation de surfaces de réponse") par exemple avec deux facteurs:

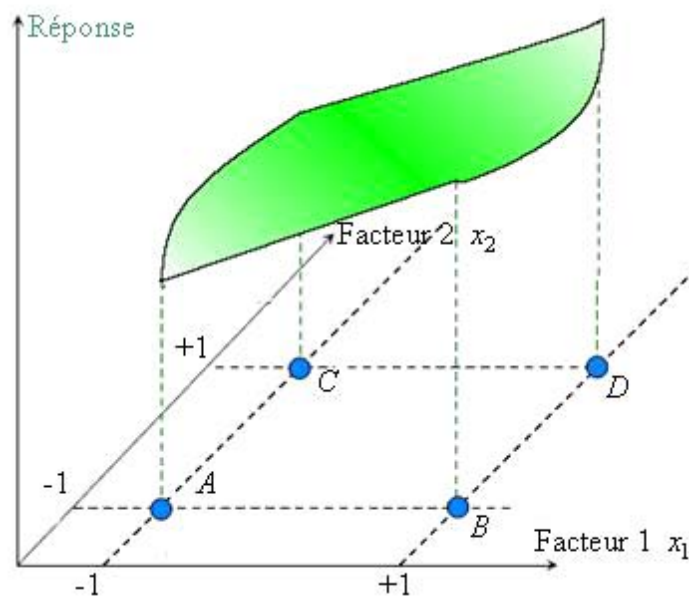


Figure: 74.33 - Deux facteurs avec la surface de réponse

Le nombre et l'emplacement des points d'expériences est le problème fondamental des plans d'expériences. Nous cherchons à obtenir la meilleure précision possible sur la surface de réponse tout en limitant le nombre d'expériences! L'ingénieur va donc rechercher une fonction mathématique continue qui relie la réponse aux facteurs.

Pour cela, nous simplifions le problème en se rappelant (cf. chapitre de Suites et Séries) que toute fonction quelle que soit son nombre de variables, peut être approchée en une somme de série de puissance en un point donné

Nous prenons alors un développement limité de la série de Taylor:

$$f(x, y) = f(x_0, y_0) + \frac{1}{1!} \frac{\partial f(x_0, y_0)}{\partial x} (x - x_0) + \frac{1}{1!} \frac{\partial f(x_0, y_0)}{\partial y} (y - y_0) + \frac{1}{2!} \left[\frac{\partial^2 f(x_0, y_0)}{\partial x^2} (x - x_0)^2 + 2 \frac{\partial^2 f(x_0, y_0)}{\partial x \partial y} (x - x_0)(y - y_0) + \frac{\partial^2 f(x_0, y_0)}{\partial y^2} (y - y_0)^2 \right] + \dots \quad (74.292)$$

Soit autour de $x_0 = 0, y_0 = 0$ (ce que nous pouvons nous permettre car nous prenons toujours des unités centrées réduites comme vues plus haut!), nous avons la série de Maclaurin au deuxième ordre et en changeant de notation pour $i = 1..2$:

$$\begin{aligned} y &= f(x_1, x_2) \\ &\cong f(0, 0) + \frac{1}{1!} \frac{\partial f}{\partial x_1}(0) x_1 + \frac{1}{1!} \frac{\partial f}{\partial x_2}(0) x_2 + \frac{1}{2!} \left[\frac{\partial^2 f}{\partial x_1^2}(0) x_1^2 + \frac{\partial^2 f}{\partial x_1 \partial x_2} x_1 x_2 + \frac{\partial^2 f}{\partial x_2 \partial x_1} x_2 x_1 + \frac{\partial^2 f}{\partial x_2^2}(0) x_2^2 \right] \\ &= a_0 + \sum a_i x_i + \sum a_{ij} x_i x_j = a_0 + \sum a_i x_i + \sum_{i \neq j} a_{ij} x_i x_j + \sum_i a_{ii} x_i x_i \end{aligned} \quad (74.293)$$

où y est donc la réponse et les x_i les facteurs et les a_0, a_i, a_{ij}, a_{ii} sont les coefficients du modèle mathématique adopté a priori. Ils ne sont pas connus et doivent être calculés à partir des résultats des expériences.

L'intérêt de modéliser la réponse par un polynôme est de pouvoir calculer ensuite toutes les réponses du domaine d'étude sans être obligé de faire les expériences en passant par un modèle appelé "**modèle postulé**" ou "**modèle a priori**".

Deux compléments doivent être apportés au modèle précédemment décrit:

1. Le premier complément est le "**manque d'ajustement**". Cette expression traduit le fait que le modèle a priori est fort probablement différent (ne serait-ce que par l'approximation de l'approche) du modèle réel qui régit le phénomène étudié. Il y a un écart entre ces deux modèles. Cet écart est le manque d'ajustement ("**lack of fit**" en anglais).
2. Le second complément est la prise en compte de la nature aléatoire de la réponse (sans que cette dernière soit toutefois stochastique sinon il faut utiliser d'autres outils!). En effet, si l'on mesure plusieurs fois une réponse en un même point expérimental, on n'obtient pas exactement le même résultat. Les résultats sont dispersés. Les dispersions ainsi constatées sont appelées "**erreurs expérimentales**".

Ces deux écarts, manque d'ajustement et erreur expérimentale, sont souvent réunis dans un seul écart, notée e .

Le modèle utilisé par l'expérimentateur s'écrit alors au deuxième ordre et au premier degré:

$$y \cong a_0 + \sum a_i x_i + \sum_{i \neq j} a_{ij} x_i x_j + e \quad (74.294)$$

et est appelée parfois "**modèle synergique**".

Remarques:

R1. Si nous prenons en compte les termes du deuxième degré, nous parlons alors de "**modèle quadratique complet**".

R2. Si nous arrêtons le développement au premier ordre et au premier degré (sans interactions), nous parlons alors de "**modèle affine**".

Dans la pratique nous notons cette dernière relation (nous enlevons le terme d'erreur):

$$y \cong a_0 + a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2 \quad (74.295)$$

où nous avons la notation abusive pour ce qui est d'usage d'appeler dans le domaine le "**terme rectangle**":

$$a_{12} := a_{21} + a_{12} \quad (74.296)$$

Dès lors, il vient que pour 2 facteurs, l'expression contient 4 termes, pour 3 facteurs elle contient 8 termes, pour 4 facteurs, 16 termes, etc. Il s'agit donc toujours de puissances de 2.

Ce modèle sans erreur est souvent appelé "**modèle contrôlé avec interactions**" (linéaire d'ordre 2). Par exemple, une surface de réponse associée à une relation du type précédente est avec Maple 4.00b:

```
>plot3d(5+3*x1+2*x2+4*x1*x2,x1=-10..10,x2=-10..10,view=
[-10..10,-10..10,-10..10],contours=10,style=PATCHCONTOUR,axes=frame,numpoints=10000);
```

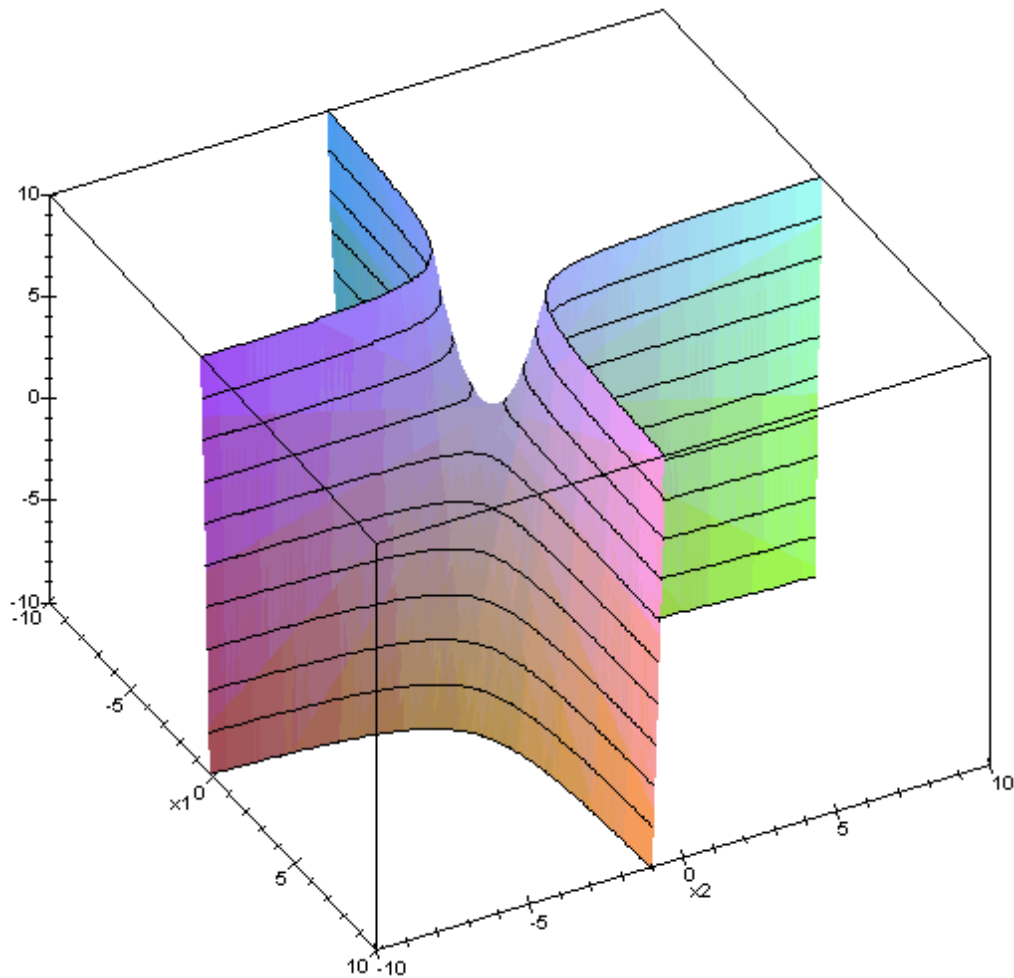


Figure: 74.34 - Exemple générique d'interaction pour deux facteurs (paraboloïde hyperbolique)

où nous avons représenté les isoclines (lignes dont les réponses de la fonction sont égales quelle que soient les valeurs des variables d'entrée). Ce qui permet de rechercher les combinaisons les moins coûteuses pour un résultat égal!

Évidemment, si nous supprimons les termes d'interactions, nous avons simplement un plan:

```
>plot3d(5+3*x1+2*x2,x1=-10..10,x2=-10..10,view=
[-10..10,-10..10,-10..10],contours=10,style=PATCHCONTOUR,axes=frame,numpoints=10000);
```

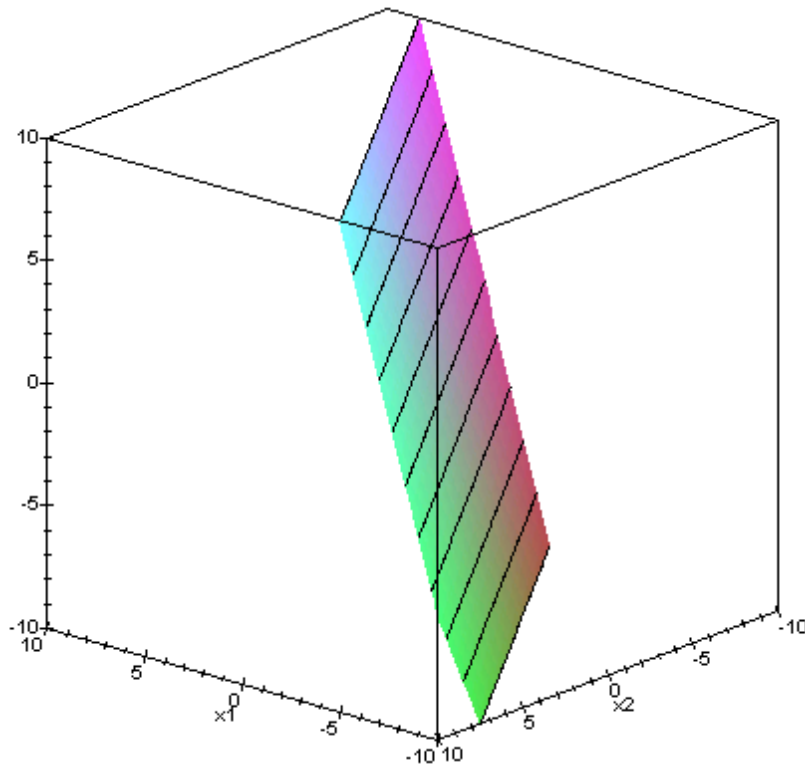


Figure: 74.35 - Exemple générique de modèle additif pour deux facteurs (plan)

Évidemment, plus le degré du polynôme utilisé est élevé plus nous avons, théoriquement, de chances d'avoir un modèle proche de la réalité. Mais les polynômes de degré élevé réclament beaucoup de points expérimentaux et leur validité peut vite diverger en dehors du domaine expérimental. Si l'étude l'exige, nous préférons utiliser des fonctions mathématiques particulières pour mieux ajuster le modèle aux résultats expérimentaux.

Cependant, en pratique, les interactions d'ordre élevé ont souvent une influence très faible sur la réponse (bon cette affirmation dépend fortement du domaine d'activité...!). Il est donc possible de ne pas les inclure dans le modèle, ce qui conduit à faire moins d'essais. Ce principe est utilisé dans la construction de nombreux plans d'expériences, comme nous le verrons dans la partie suivante. Dans de nombreuses applications, on obtient des résultats tout à fait satisfaisants en se limitant aux interactions doubles.

Pourquoi nous satisfaisons-nous de cette relation approchée de quatre termes? Pour la simple raison que:

1. La réponse peut être non nulle lorsque tous les facteurs sont nuls (c'est le premier coefficient a_0)
2. La réponse dépend trivialement (intuitivement) de la somme des effets du premier et deuxième facteurs x_1, x_2 de manière indépendante (coefficients a_1, a_2).
3. La réponse dépend aussi des interactions entre les deux facteurs x_1, x_2 (coefficients a_{12}).

Chaque point expérimental dont les x_i sont données permet alors d'obtenir une valeur de la réponse y . Cette réponse est modélisée par un polynôme dont les coefficients sont les inconnues qu'il faut déterminer.

3.1. PLANS FACTORIELS COMPLETS

Donc dans un plan d'expérience de 2 facteurs à 2 niveaux, nous avons besoin d'au moins (et au plus pour des raisons de coûts!) de 4 mesures pour avoir un système de quatre équations à quatre inconnues qui sont les coefficients $\alpha_0, \alpha_1, \alpha_2, \alpha_{12}$.

Remarque: Pour une étude de 2 facteurs à 3 niveaux, nous ne pouvons plus prendre le modèle linéaire. Il nous faut alors prendre les termes quadratiques du développement de Taylor.

Puisque pour chacun des facteurs nous devons nous fixer un niveau bas et un niveau haut pour pouvoir travailler raisonnablement... alors si nous avons deux facteurs, nous avons un espace expérimental défini par 4 points {(haut, haut), (bas, bas), (haut, bas), (bas, haut)}, correspondant aux 2 fois 2 niveaux (2^2), qui nous suffit pour obtenir alors nos 4 équations à 4 inconnues et alors de déterminer les 4 coefficients.

Ainsi, les points à prendre pour notre expérience correspondent naturellement aux sommets (géométriquement parlant) de l'espace expérimental.

Nous avons alors le système d'équations (rappelons que cette approche ne fonctionne que pour des facteurs ayant 2 niveaux, le cas à trois niveaux étant traité bien plus loin avec un exemple):

$$\begin{aligned} y_1 &\cong \alpha_0 + \alpha_1 x_{1,-1} + \alpha_2 x_{2,-1} + \alpha_{12} x_{1,-1} x_{2,-1} \\ y_2 &\cong \alpha_0 + \alpha_1 x_{1,+1} + \alpha_2 x_{2,-1} + \alpha_{12} x_{1,+1} x_{2,-1} \\ y_3 &\cong \alpha_0 + \alpha_1 x_{1,-1} + \alpha_2 x_{2,+1} + \alpha_{12} x_{1,-1} x_{2,+1} \\ y_4 &\cong \alpha_0 + \alpha_1 x_{1,+1} + \alpha_2 x_{2,+1} + \alpha_{12} x_{1,+1} x_{2,+1} \end{aligned} \quad (74.297)$$

ou explicitement écrit:

$$\begin{aligned} y_1 &\cong \alpha_0 + \alpha_1(-1) + \alpha_2(-1) + \alpha_{12}(-1)(-1) = \alpha_0 - \alpha_1 - \alpha_2 + \alpha_{12} \\ y_2 &\cong \alpha_0 + \alpha_1(+1) + \alpha_2(-1) + \alpha_{12}(+1)(-1) = \alpha_0 + \alpha_1 - \alpha_2 - \alpha_{12} \\ y_3 &\cong \alpha_0 + \alpha_1(-1) + \alpha_2(+1) + \alpha_{12}(-1)(+1) = \alpha_0 - \alpha_1 + \alpha_2 - \alpha_{12} \\ y_4 &\cong \alpha_0 + \alpha_1(+1) + \alpha_2(+1) + \alpha_{12}(+1)(+1) = \alpha_0 + \alpha_1 + \alpha_2 + \alpha_{12} \end{aligned} \quad (74.298)$$

Comme, nous nous intéressons aux coefficients (et que les variables sont connues!), il s'agit en fait simplement de résoudre un système linéaire (cf. chapitre d'Algèbre Linéaire).

Si les variables n'étaient pas codées, nous aurions par exemple pour un plan factoriel complet à deux niveaux avec une variable (vitesse du véhicule) ayant pour valeur basse 40 [km/h] et pour valeur haute 50 [km/h], ainsi qu'une deuxième variable (pression de pneus) ayant pour valeur basse 1.5 [Pa] et pour valeur haute 3 [Pa] le système suivant à résoudre sachant que nous avons obtenu respectivement pour chaque combinaison les distances parcourues suivantes 32'700, 32'680, 31'710 33'220 [km]:

$$\begin{aligned}
 a_0 \cdot 1 + a_1(40) + a_2(1.5) + a_{12}(60)(1.5) &\cong 32'700 \\
 a_0 \cdot 1 + a_1(50) + a_2(1.5) + a_{12}(50)(1.5) &\cong 32'680 \\
 a_0 \cdot 1 + a_1(40) + a_2(3) + a_{12}(40)(3) &\cong 31'710 \\
 a_0 \cdot 1 + a_1(50) + a_2(3) + a_{12}(50)(3) &\cong 33'220
 \end{aligned}
 \quad (74.299)$$

Soit sous forme matricielle:

$$\begin{bmatrix} 1 & 40 & 1.5 & 60 \\ 1 & 50 & 1.5 & 75 \\ 1 & 40 & 3 & 120 \\ 1 & 50 & 3 & 150 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_{12} \end{bmatrix} = \begin{bmatrix} 32'700 \\ 32'680 \\ 31'710 \\ 33'220 \end{bmatrix} \quad (74.300)$$

En résolvant ce système à la main (cf. chapitre d'Algèbre Linéaire), ou avec un logiciel de calcul (tableur ou logiciel de statistiques), nous obtenons:

	A	B	C	D	E
1	Constante	Vitesse	Pression	Vitesse*Pression	Réponse
2	1	40.00	1.50	60	32700
3	1	50.00	1.50	75	32680
4	1	40.00	3.00	120	31710
5	1	50.00	3.00	150	33220
6					
7					Coefficient
8	10	-8	-5	4	39890
9	-0.2	0.2	0.1	-0.1	-155
10	-3.333333	2.666667	3.333333333	-2.666666667	-4740
11	0.066667	-0.06667	-0.066666667	0.066666667	102

Figure: 74.36 - Mise en équation et résolution implicite sous la version anglaise de Microsoft Excel 14.0.6123

soit explicitement:

	A	B	C	D	E
1	Constante	Vitesse	Pression	Vitesse*Pression	Réponse
2	1	40	1.5	=B2*C2	32700
3	1	50	1.5	=B3*C3	32680
4	1	40	3	=B4*C4	31710
5	1	50	3	=B5*C5	33220
6					
7					Coefficient
8	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MMULT(A8:D11,E2:E5)
9	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MMULT(A8:D11,E2:E5)
10	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MMULT(A8:D11,E2:E5)
11	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MMULT(A8:D11,E2:E5)

Figure: 74.37 - Mise en équation et résolution explicite sous la version anglaise de Microsoft Excel 14.0.6123

Donc la fonction est finalement:

$$y \cong 39'890 - 155x_1 - 4740x_2 + 102x_{12} \quad (74.301)$$

soit exactement les mêmes coefficients que ce que nous donne un logiciel spécialisé comme Minitab 15.1.1.

Pour en revenir à notre système avec les variable codées, il vient alors en résolvant le système algébriquement (relations valables que si et seulement si les variables sont codées!):

$$\begin{aligned}a_0 &= \frac{1}{4}(y_1 + y_2 + y_3 + y_4) \\a_1 &= \frac{1}{4}(-y_1 + y_2 - y_3 + y_4) \\a_2 &= \frac{1}{4}(-y_1 - y_2 + y_3 + y_4) \\a_{12} &= \frac{1}{4}(y_1 - y_2 - y_3 + y_4)\end{aligned}\quad (74.302)$$

qui peut s'écrire sous la forme matricielle suivante:

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_{12} \end{bmatrix} = \frac{1}{4} \begin{bmatrix} +1 & +1 & +1 & +1 \\ -1 & +1 & -1 & +1 \\ -1 & -1 & +1 & +1 \\ +1 & -1 & -1 & +1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \quad (74.303)$$

Ce qui s'écrit de manière générale pour des modèles linéaires du deuxième ordre sous la forme générale (cf. chapitre d'Algèbre Linéaire):

$$A = \frac{1}{n} XY \quad (74.304)$$

La matrice X contenant 2^n lignes est appelée "**plan factoriel complet (PFC) 2^n avec interactions**" (le terme "factoriel" venant du fait que tous les facteurs varient simultanément).

La matrice X dans la pratique est appelée "**matrice d'expérimentation**" ou encore "**matrice des effets**" et est souvent représentée de la manière suivante dans le cas particulier précédent:

Essai n°	Repos	Facteur 1	Facteur 2	Facteur 12	Réponse
1	+1	-1	-1	+1	y_1
2	+1	+1	-1	-1	y_2
3	+1	-1	+1	-1	y_3
4	+1	+1	+1	+1	y_4

Tableau: 74.14 - Matrice d'expérimentation

Mais l'on voit tout de suite que dans la pratique, la deuxième colonne (*Repos*) est inutile, car elle vaut toujours +1 et elle est implicitement implémentée dans les logiciels.

Il en est de même pour la cinquième colonne (*Facteur 12*), car elle se déduit automatiquement de la troisième et quatrième colonnes (c'est la multiplication des termes ligne par ligne... ce que certains appellent la "**multiplication de Box**").

Le lecteur remarquera aussi qu'en passant d'une colonne à l'autre ou d'une ligne à l'autre, il y a toujours deux facteurs qui changent de niveau! Par contre, si nous nous concentrons uniquement sur les colonnes *Facteur 1* et *Facteur 2*, nous voyons qu'en passant d'une ligne à l'autre, il n'y a qu'un facteur qui change à la fois.

Remarque: Observez donc que la première colonne vaut toujours +1 et il y a toujours aussi une ligne avec uniquement des +1!

Ainsi, dans la pratique (logiciels) et dans de nombreux ouvrages, on représente à juste titre uniquement le tableau suivant (ce qui masque le fait que nous avons affaire à une matrice carrée):

Essai n°	Facteur 1	Facteur 2	Réponse
1	-1	-1	y_1
2	+1	-1	y_2
3	-1	+1	y_3
4	+1	+1	y_4

Tableau: 74.15 - Matrice d'expérimentation simplifiée

pour un plan d'expérience de 2 facteurs à 2 niveaux avec interactions en modèle linéaire (sans erreur) nous avons encore plus extrême ("**notation de Yates**")... en termes d'écriture:

Essai n°	Facteur 1	Facteur 2	Réponse
1	-	-	y_1
2	+	-	y_2
3	-	+	y_3
4	+	+	y_4

Tableau: 74.16 - Matrice d'expérimentation avec notation de Yates

Nous voyons que dans cette forme d'écriture qu'outre le fait que les deux colonnes *Facteur 1* et *Facteur 2* sont orthogonales (la norme ISO 3534-3:1999 parle de "contraste orthogonal"), elles sont aussi "balancées", dans le sens qu'il y a autant de + et de - dans chacune des colonnes.

Remarque: Par défaut, la majorité des logiciels randomise l'ordre des essais du plan (qu'il soit complet, fractionnaire, factoriel ou non). En règle générale, il est recommandé de randomiser l'ordre des essais pour atténuer les effets des facteurs qui ne sont pas inclus dans l'étude, notamment les effets liés au temps. Cependant, dans certains cas, la randomisation ne produit pas un ordre des essais intéressant et peut être même dangereux car elle peut masquer certaines sources d'erreurs systématiques qui n'avaient pas été identifiées avant l'expérience. Par exemple, dans les applications industrielles, la modification des niveaux de facteurs peut s'avérer difficile ou coûteuse. Il est également possible qu'une fois la modification effectuée, le retour du système à un état constant prenne beaucoup de temps. En pareil cas, il peut être souhaitable de ne pas randomiser le plan afin de limiter au maximum les modifications de niveau.

Les praticiens apprécient de calculer la moyenne des réponses et l'effet moyen d'un facteur donné puisque le système est linéaire. Ainsi, au niveau +1 pour le facteur x_1 , en nous basant toujours sur le système linéaire:

$$\begin{aligned} y_1 &\cong \alpha_0 + \alpha_1 x_{1,-1} + \alpha_2 x_{2,-1} + \alpha_{12} x_{1,-1} x_{2,-1} \\ y_2 &\cong \alpha_0 + \alpha_1 x_{1,+1} + \alpha_2 x_{2,-1} + \alpha_{12} x_{1,+1} x_{2,-1} \\ y_3 &\cong \alpha_0 + \alpha_1 x_{1,-1} + \alpha_2 x_{2,+1} + \alpha_{12} x_{1,-1} x_{2,+1} \\ y_4 &\cong \alpha_0 + \alpha_1 x_{1,+1} + \alpha_2 x_{2,+1} + \alpha_{12} x_{1,+1} x_{2,+1} \end{aligned} \quad (74.305)$$

nous pouvons alors construire et définir la "**moyenne des réponses**" donnée par:

$$y_{x_1,+} = \frac{1}{2}(y_2 + y_4) \quad (74.306)$$

et il vient de même au niveau -1 toujours pour le même facteur:

$$y_{x_1,-} = \frac{1}{2}(y_1 + y_3) \quad (74.307)$$

Nous avons alors "**l'effet global**" du facteur x_1 qui sera donné par :

$$E_{G_{x_1,+}} = y_{x_1,+} - y_{x_1,-} \quad (74.308)$$

et "**effet moyen**" du facteur x_1 qui est alors défini par la demi-différence entre la moyenne des réponses au niveau haut du facteur x_1 et la moyenne des réponses au niveau bas du même facteur:

$$E_{M_{x_1,+}} = \frac{1}{2}(y_{x_1,+} - y_{x_1,-}) = \frac{1}{2} \left(\frac{y_2 + y_4}{2} - \frac{y_1 + y_3}{2} \right) \quad (74.309)$$

or après quelques simplification d'algèbre élémentaire, il vient très rapidement que:

$$E_{M_{x_1,+}} = \alpha_1 \quad (74.310)$$

Il est évident que si l'effet global (et in extenso l'effet moyen) est non nul, nous pouvons nous douter qu'il y a une interaction entre les facteurs puisque la variation d'amplitude de la réponse n'est pas identique en fonction de la valeur du niveau de l'autre facteur (voir l'étude de l'analyse de la variance à deux facteurs dans le chapitre de Statistiques pour plus de détails). Évidemment, dans la pratique, l'étude des interactions se fera en tout rigueur comme souvent pour l'ANOVA: avec des graphiques.

La moyenne de toutes les réponses donne donc la valeur de la réponse au centre du domaine expérimental:

$$\bar{y} = \frac{1}{4}(y_1 + y_2 + y_3 + y_4) \quad (74.311)$$

En faisant un peu d'algèbre élémentaire, que le plan soit écrit sous forme normalisée et centrée ou non, cette dernière expression se réduit à:

$$\bar{y} = a_0 \quad (74.312)$$

Évidemment, nous pouvons résumer ce cas simple sous forme graphique si cela peut aider le lecteur:

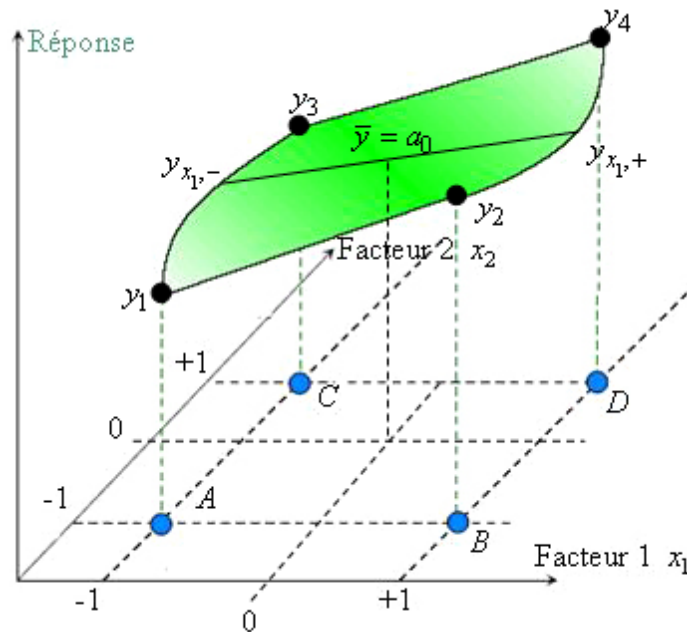


Figure: 74.38 - Deux facteurs avec la surface de réponse

et avec la figure ci-dessus, la variations n'étant pas très parallèles lorsqu'un facteur est fixé, nous pouvons supposer qu'il y a interaction entre les facteurs qui nécessiteront l'utilisation d'une ANOVA.

Lorsque le nombre de facteurs est grand, il n'est pas toujours facile pour tout le monde de poser rapidement les facteurs +, - sans l'aide d'un logiciel. Alors, il existe une petite marche à suivre appelée "**algorithme de Yates**" ou "**algorithme de Yates et Hunter**" qui permet vite d'arriver au résultat voulu pour les plans factoriels (dont les facteurs ont deux niveaux) dont le nombre de facteurs est une puissance de 2:

D'abord, nous commençons toutes les colonnes par -1 et nous alternons les -1 et les +1 toutes les 2^{j-1} lignes pour la j -ème colonne.

Remarques:

R1. Si le type de tableau précédent contient des valeurs codées, nous parlons de "**plan d'expérience**" sinon avec les unités physiques habituelles nous parlons de "**tableau d'expérimentation**".

R2. Dans le cas des tableaux codés, il est d'usage d'indiquer sous le tableau un deuxième tableau avec les correspondances entre les unités codées et les unités physiques.

Insistons sur une chose importante: C'est que si nous avons 3 facteurs à 2 niveaux, alors nous avons 2^3 possibilités d'expériences (soit 8). Or, huit correspond exactement au nombre de coefficients que nous avons également dans le modèle linéaire avec interactions d'une réponse à trois variables:

$$y \cong \underline{a}_0 + \underline{a}_1 x_1 + \underline{a}_2 x_2 + \underline{a}_3 x_3 + \underline{a}_{12} x_1 x_2 + \underline{a}_{13} x_1 x_3 + \underline{a}_{23} x_2 x_3 + \underline{a}_{123} x_1 x_2 x_3 \quad (74.313)$$

ce qui correspond aussi aux termes seulement linéaires et sous forme condensée du développement en série de Maclaurin d'une fonction f de trois variables.

Et ainsi de suite... pour n facteurs à deux niveaux. C'est la raison pour laquelle les plans factoriels complets linéaires 2^n sont traditionnellement les plus utilisés car ils sont mathématiquement intuitifs et simples à démontrer.

Par ailleurs, il est important de remarquer que tous ces plans linéaires complets approximés au deuxième ordre sont sous forme matricielles des matrices carrées $M_{n,n}$ orthogonales et donc bien évidemment inversibles (cf. chapitre d'Algèbre Linéaire)!

Cependant les matrices précédentes ne satisfont pas la relation suivante vue dans le chapitre d'Algèbre Linéaire (produit scalaires des colonnes-bases avec elles-mêmes et de vecteurs orthogonaux):

$$X^T X = \mathbb{1} \quad (74.314)$$

mais ont pour particularité pour tout plan d'expérience complet de satisfaire la relation:

$$X^T X = n \mathbb{1} \quad (74.315)$$

Donc contrairement aux matrices orthogonales qui par définition ont toutes les colonnes (ou lignes) qui forment une base orthonormée (norme unitaire), les matrices des plans d'expérience ont pour différence de ne pas avoir les normes de la base orthogonale à l'unité.

Ainsi, nous définissons la matrice A dont les coefficients sont tous des +1 ou des -1 ET satisfaisant la relation précédente comme étant une "**matrice de Hadamard**". Cette dernière a par ailleurs pour propriété d'exister que pour les ordres 1, 2, 4, 8, 12, 16, 20, 24, ... C'est-à-dire uniquement pour les ordres multiple de 4 (si nous omettons l'ordre 1 et 2).

Démonstration:

Sachant que les cas d'ordre 1 et 2 sont triviaux et que les cas impairs sont à éliminer intuitivement immédiatement (essayez et vous verrez très vite...), faisons la démonstration pour $n \geq 4$.

Puisque toutes les colonnes doivent être obligatoirement orthogonales (pour que la matrice soit inversible et donc le système résoluble), nous pouvons toujours écrire le problème sous la forme (forme particulière pour n valant 4 mais facilement généralisable):

$$\begin{aligned} +x + y + z + w &= n \\ +x + y - z - w &= 0 \\ +x - y + z - w &= 0 \\ +x - y - z + w &= 0 \end{aligned} \quad (74.316)$$

et si nous notons n comme étant l'ordre de la matrice. Alors nous avons par sommation de toutes les lignes:

$$4x = n \quad (74.317)$$

donc n doit être divisible par 4 pour $n \geq 4$ pour que toutes les colonnes soient orthogonales et donc que la matrice soit de Hadamard sachant que x, y, z, w ont pour valeur 1.

Pour les lecteurs qui doutent concernant les valeurs paires: 6, 10, 14, 18, etc. Essayez de trouver une matrice de Hadamard à la main dont tous les vecteurs sont orthogonaux deux à deux... vous verrez très vite que vous allez coincer.

□ C.Q.F.D.

Il s'ensuit alors trivialement la relation suivante (avec une notation très abusive car on omet la notation de la matrice unité):

$$A = \frac{1}{n} XY = \left(X^T X \right) XY \quad (74.318)$$

Cependant, nous verrons un tout petit peu plus loin lors de notre étude des plans factoriels fractionnaires que les matrices d'Hadamard d'ordre 1, 4, 8, 16, 32, ..., 2^n ne sont quasiment jamais utilisées sous l'expression de "matrices d'Hadamard" car elles se confondent avec les plans factoriels fractionnaires.

Par contre, il est intéressant de constater que nous avons des plans d'expérience possibles avec 12, 20, 24, 28, ... essais qui ont un intérêt certain lorsque le nombre de facteurs est supérieur ou égal 4 (certains logiciels comme JMP ne les proposent pas cependant si le nombre de facteurs est inférieur à cinq). Ces plans sont appelés "**plans de Plackett-Burman**" et parfois "**plans irréguliers de Plackett-Burman**" ou encore plus rarement "**plans non-géométriques de Plackett-Burman**" (un logiciel comme Minitab propose cependant arbitrairement - et l'indique explicitement ainsi dans l'aide du logiciel - un plan de Plackett-Burman d'ordre 32 et des plans de Plackett-Burman même pour 2 ou 3 facteurs).

Plackett et Burman ont cherché avec des algorithmes et par tâtonnement l'expression des matrices d'expériences correspondant aux plans qui portent leur nom et qui contiennent donc 12, 20, 24, 28, etc. essais. Ils ont proposé une astuce fort utile pour que les praticiens puissent faire usage de ces plans sans logiciels. Faisons l'exemple pour commencer avec une matrice de Hadamard d'ordre 12 (sans s'occuper pour l'instant du nombre de facteurs que nous allons utiliser). Les tables de Plackett-Burman (ou les logiciels) donnent uniquement la première ligne:

+ + - + + - - - + -

Ensuite, nous construisons la table d'expérience suivante en suivant l'algorithme proposé par Plackett et Burman (les logiciels utilisent le même):

- Première étape, nous mettons la première ligne sous forme de colonne dans un tableau de 12 par 12:

+											
+											
-											
+											
+											
+											
-											
-											
-											
+											
-											

Tableau: 74.17 - Première étape de construction du plan de Plackett- Burman

- Deuxième étape, nous déduisons la deuxième colonne à partir de cette première colonne en décalant les signes d'un cran vers la bas, le dernier signe "-" étant, lui, remonté au sommet de la deuxième colonne (il s'agit donc d'une permutation circulaire):

+	-										
+	+										
-	+										
+	-										
+	+										
+	+										
-	+										
-	-										
-	-										
+	-										
-	+										

Tableau: 74.18 - Deuxième étape de construction du plan de Plackett-Burman

et ainsi de suite jusqu'à la 11ème colonne:

+	-	+	-	-	-	+	+	+	-	+	
+	+	-	+	-	-	-	+	+	+	-	
-	+	+	-	+	-	-	-	+	+	+	
+	-	+	+	-	+	-	-	-	+	+	
+	+	-	+	+	-	+	-	-	-	+	
+	+	+	-	+	+	-	+	-	-	-	
-	+	+	+	-	+	+	-	+	-	-	
-	-	+	+	+	-	+	+	-	+	-	
-	-	-	+	+	+	-	+	+	-	+	
+	-	-	-	+	+	+	-	+	+	-	
-	+	-	-	-	+	+	+	-	+	+	

Tableau: 74.19 - 3ème à 11ème étape de construction du plan de Plackett-Burman

- Dernière étape, nous ajoutons une rangée et une colonne de signes "-":

+	-	+	-	-	-	+	+	+	-	+	-
+	+	-	+	-	-	-	+	+	+	-	-
-	+	+	-	+	-	-	-	+	+	+	-
+	-	+	+	-	+	-	-	-	+	+	-
+	+	-	+	+	-	+	-	-	-	+	-
+	+	+	-	+	+	-	+	-	-	-	-
-	+	+	+	-	+	+	-	+	-	-	-
-	-	+	+	+	-	+	+	-	+	-	-
-	-	-	+	+	+	-	+	+	-	+	-
+	-	-	-	+	+	+	-	+	+	-	-
-	+	-	-	-	+	+	+	-	+	+	-
-	-	-	-	-	-	-	-	-	-	-	-

Tableau: 74.20 - Dernière étape de construction du plan de Plackett-Burman

Le lecteur pourra facilement vérifier que chaque ligne ou chaque colonne prises deux à deux sont bien orthogonales.

Maintenant se pose à la question de savoir si nous choisissons d'associer par exemple ce plan d'expérience à un plan à 4 facteurs (sous-entendu centrés réduits bien évidemment!):

$$\begin{aligned}
 y \cong & a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 \\
 & + a_{12} x_1 x_2 + a_{13} x_1 x_3 + a_{14} x_1 x_4 + a_{23} x_2 x_3 + a_{24} x_2 x_4 + a_{34} x_3 x_4 \\
 & + a_{123} x_1 x_2 x_3 + a_{124} x_1 x_2 x_4 + a_{134} x_1 x_3 x_4 + a_{234} x_2 x_3 x_4 \\
 & + a_{1234} x_1 x_2 x_3 x_4
 \end{aligned}
 \quad (74.319)$$

pour réduire le nombre d'essais de 16 à 12 à quelle colonne devons nous associer à quoi? Ou encore en allant plus loin: pour combien de facteurs pouvons nous associer ce type de plan à 12 essais pour des facteurs à deux niveaux?

Eh bien il y a deux grandes religions (maisons) suivant les ingénieurs.

- La première consiste à dire que les plans de Plackett-Burman d'ordre n ne doivent être utilisés que pour l'étude des effets principaux (donc aucune interaction ne sera prise en compte) de $n-1$ facteurs. Ainsi, un plan de Plackett & Burmann d'ordre 12, sera réservé pour une étude de 11 facteurs et uniquement de leurs effets principaux (modèle linéaire purement additif).

- La deuxième consiste à dire que les plans de Plackett- Burman contiennent un concept que nous verrons plus loin et qui s'appelle les "alias" (les plans de Plackett-Burman sont tous des plans de résolution III au deuxième ordre). Par conséquent, il faut utiliser ce type de plan que lorsque nous sommes prêts à considérer dans un premier temps que les interactions à deux facteurs sont très négligeables.

La complexité des alias avec les plans de Plackett- Burmann fait que dans la pratique, ils sont finalement plutôt utilisés selon la première religion... par les débutants et selon la seconde par les consultants.

Remarque: Un logiciel comme Minitab 15.1.1 n'affiche pas du tous les alias utilisés lors de l'application de plans de Placket et Burmann.

Pour clore cette partie, résumons un constat simple:

Plan	Facteurs	Interactions	Somme
2^2	2	1	3
2^3	3	4	7
2^4	4	11	15
2^5	5	26	31
2^6	6	57	63
2^7	7	120	127
...	...	...	...

Tableau: 74.21 - Types de plans avec facteurs & interactions

Donc en utilisant les plans factoriels complets, l'utilisateur est sûr d'avoir la procédure expérimentale optimale puisque ces plans sont basés sur des matrices d'Hadamard et qu'il a été démontré que nous ne pouvions pas faire mieux.

3.2. PLANS FACTORIELS FRACTIONNAIRES

En pratique, les plans complets ne sont utilisables que sur des systèmes avec très peu de facteurs, ou lorsque chaque essai prend très peu de temps. Lorsque n est plus grand ou égal 3 alors les coûts des expériences peut très vite devenir onéreux.

Le plus petit cas où il est intéressant d'optimiser le nombre d'essais et celui consistant en 3 facteurs à 2 niveaux. Nous avons alors l'équation et le tableau d'expérience suivant:

$$y \cong a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_{12} x_1 x_2 + a_{13} x_1 x_3 + a_{23} x_2 x_3 + a_{123} x_1 x_2 x_3$$

Essai n°	Facteur 1	Facteur 2	Facteur 3	Réponse
1	-	-	-	y_1
2	+	-	-	y_2
3	-	+	-	y_3
4	+	+	-	y_4
5	-	-	+	y_5
6	+	-	+	y_6
7	-	+	+	y_7
8	+	+	+	y_8

Tableau: 74.22 - Plan d'expérience à 3 facteurs complet sous forme de Yates

Soit sous forme de tableau d'expérience complet:

Essai n°	Repos	F 1	F 2	F 3	F 12	F 13	F 23	F 123	Réponse
1	+	-	-	-	+	+	+	-	y_1
2	+	+	-	-	-	-	+	+	y_2
3	+	-	+	-	-	+	-	+	y_3
4	+	+	+	-	+	-	-	-	y_4
5	+	-	-	+	+	-	-	+	y_5
6	+	+	-	+	-	+	-	-	y_6
7	+	-	+	+	-	-	+	-	y_7
8	+	+	+	+	+	+	+	+	y_8

Tableau: 74.23 - Plan d'expérience à 3 facteurs et interactions complet sous forme de Yates

ou de matrice d'expérience complète:

$$\begin{bmatrix}
 +1 & -1 & -1 & -1 & +1 & +1 & +1 & -1 \\
 +1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \\
 +1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 \\
 +1 & +1 & +1 & -1 & +1 & -1 & -1 & -1 \\
 +1 & -1 & -1 & +1 & +1 & -1 & -1 & +1 \\
 +1 & +1 & -1 & +1 & -1 & +1 & -1 & -1 \\
 +1 & -1 & +1 & +1 & -1 & -1 & +1 & -1 \\
 +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1
 \end{bmatrix} \quad (74.320)$$

où à nouveau il est facile de contrôler que toutes les colonnes sont orthogonales et balancées (même nombre de + ou de - dans chaque colonne ou autrement dit la somme de leurs colonnes est nulle) et que la matrice est bien de type Hadamard. Ainsi, un plan factoriel complet pour 3 facteurs implique 8 essais.

Nous pouvons également représenter tout cela sous forme graphique:

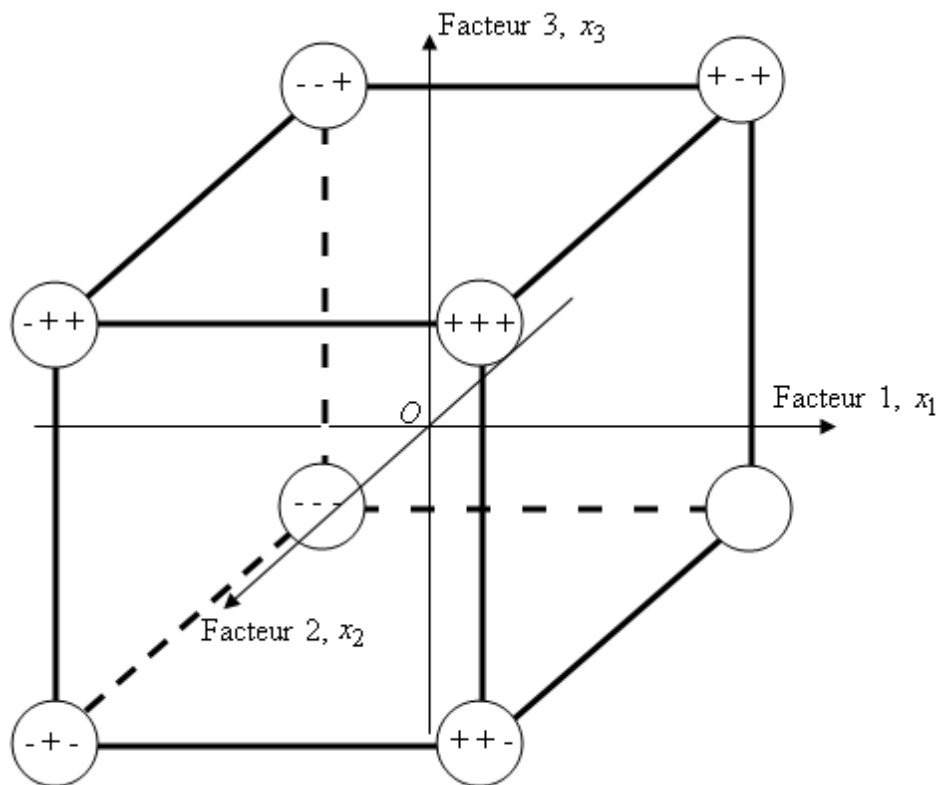


Figure: 74.39 - Représentation traditionnelle d'un plan factoriel fractionnaire complet à 3 facteurs

Les plans réduits (plans factoriels fractionnaires) ou "**plans de screening**" (selon la norme ISO 3534-3:1999), consistant à sélectionner certaines combinaisons, ont donc été proposés. Ils permettent naturellement de réduire les coûts, mais diminuent également l'information disponible sur le comportement du système! Il faut donc s'assurer de la pertinence de la sélection par rapport au modèle à identifier.

Une première méthode élémentaire est par exemple de faire l'hypothèse qu'il n'y a aucune interaction. Dès lors, notre fonction se réduit à un modèle purement additif:

$$y \cong a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 \quad (74.321)$$

et pour résoudre ce système, il nous suffit d'avoir 4 essais. Nous passons ainsi de 8 à 4 essais en supposant qu'il n'y a pas d'interactions et nous nous retrouvons avec un simple plan de criblage.

Pour réduire les coûts d'expérimentation tout en gardant les interactions implicitement, nous pouvons jouer avec la mathématique. D'abord reprenons le problème actuel avec un plan factoriel de 3 facteurs complet sous forme explicite:

$$\begin{aligned}
 +a_0 - a_1 - a_2 - a_3 + a_{12} + a_{13} + a_{23} - a_{123} &= y_1 \\
 +a_0 + a_1 - a_2 - a_3 - a_{12} - a_{13} + a_{23} + a_{123} &= y_2 \\
 +a_0 - a_1 + a_2 - a_3 - a_{12} + a_{13} - a_{23} + a_{123} &= y_3 \\
 +a_0 + a_1 + a_2 - a_3 + a_{12} - a_{13} - a_{23} - a_{123} &= y_4 \\
 +a_0 - a_1 - a_2 + a_3 + a_{12} - a_{13} - a_{23} + a_{123} &= y_5 \\
 +a_0 + a_1 - a_2 + a_3 - a_{12} + a_{13} - a_{23} - a_{123} &= y_6 \\
 +a_0 - a_1 + a_2 + a_3 - a_{12} - a_{13} + a_{23} - a_{123} &= y_7 \\
 +a_0 + a_1 + a_2 + a_3 + a_{12} + a_{13} + a_{23} + a_{123} &= y_8
 \end{aligned} \quad (74.322)$$

Pouvons-nous en réduire l'écriture afin de minimiser le nombre d'expériences à faire? La réponse est Oui mais en contre partie nous allons perdre la mesure des effets purs (nous parlons alors parfois de "confusion").

L'écriture inférieure la plus proche est une matrice de Hadamard d'ordre 4. Ce qui signifie bien évidemment que nous ne devons conserver 4 lignes sur les 8 et que celles-ci doivent rester orthogonales, balancées et satisfaire la relation:

$$X^T X = n \mathbb{I} \quad (74.323)$$

L'idée, appelée "**méthode de Box et Hunter**" et qui ne marche que pour les plans avec des facteurs ayant deux niveaux, est alors dans un premier temps de rassembler les facteurs d'influence (en indices) tel que (développements similaires pour tout n):

$$\begin{aligned} &+(a_0 - a_{123}) + (-a_1 + a_{23}) + (-a_2 + a_{13}) + (-a_3 + a_{12}) = y_1 \\ &+(a_0 + a_{123}) + (+a_1 + a_{23}) + (-a_2 - a_{13}) + (-a_3 - a_{12}) = y_2 \\ &+(a_0 + a_{123}) + (-a_1 - a_{23}) + (+a_2 + a_{13}) + (-a_3 - a_{12}) = y_3 \\ &+(a_0 - a_{123}) + (+a_1 - a_{23}) + (+a_2 - a_{13}) + (-a_3 + a_{12}) = y_4 \\ &+(a_0 + a_{123}) + (-a_1 - a_{23}) + (-a_2 - a_{13}) + (+a_3 + a_{12}) = y_5 \\ &+(a_0 - a_{123}) + (+a_1 - a_{23}) + (-a_2 + a_{13}) + (+a_3 - a_{12}) = y_6 \\ &+(a_0 - a_{123}) + (-a_1 - a_{23}) + (+a_2 - a_{13}) + (+a_3 - a_{12}) = y_7 \\ &+(a_0 + a_{123}) + (+a_1 + a_{23}) + (+a_2 + a_{13}) + (+a_3 + a_{12}) = y_8 \end{aligned} \quad (74.324)$$

Le choix du regroupement est aussi fait en sorte que les coefficients des d'interactions supposés négligeables de par la normalisation centrée réduite se retrouvent avec un coefficient d'un facteur principal supposé comme non négligeable. Dès lors, il est aussi logique que dans chaque regroupement, on ne retrouve jamais en indice le numéro du facteur principal.

Nous disons alors que nous avons une structure d'alias (la norme ISO 3534-3:1999 appelle cela une "concomitance" lorsque c'est l'expérimentateur qui force le regroupement et "alias" si la confusion est due à la nature de l'expérience) du type:

$$0+123;1+23;2+13;3+12$$

et si nous notons cela comme le font de nombreux logiciels de statistiques, cela donne (c'est exactement les alias que donne un logiciel comme Minitab 15.1.1):

$$I+ABC;A+BC;B+AC;C+AB$$

Écrivons cela de la manière suivante:

$$\begin{aligned}
(a_0 - a_{123}) - (a_1 - a_{23}) - (a_2 - a_{13}) - (a_3 - a_{12}) &= y_1 \\
(a_0 + a_{123}) + (a_1 + a_{23}) - (a_2 + a_{13}) - (a_3 + a_{12}) &= y_2 \\
(a_0 + a_{123}) - (a_1 + a_{23}) + (a_2 + a_{13}) - (a_3 + a_{12}) &= y_3 \\
(a_0 - a_{123}) + (a_1 - a_{23}) + (a_2 - a_{13}) - (a_3 - a_{12}) &= y_4 \\
(a_0 + a_{123}) - (a_1 + a_{23}) - (a_2 + a_{13}) + (a_3 + a_{12}) &= y_5 \\
(a_0 - a_{123}) + (a_1 - a_{23}) - (a_2 - a_{13}) + (a_3 - a_{12}) &= y_6 \\
(a_0 - a_{123}) - (a_1 - a_{23}) + (a_2 - a_{13}) + (a_3 - a_{12}) &= y_7 \\
(a_0 + a_{123}) + (a_1 + a_{23}) + (a_2 + a_{13}) + (a_3 + a_{12}) &= y_8
\end{aligned} \tag{74.325}$$

Changeons de notation:

$$\begin{aligned}
l_0^- - l_1^- - l_2^- - l_3^- &= y_1 \\
l_0^+ + l_1^+ - l_2^+ - l_3^+ &= y_2 \\
l_0^+ - l_1^+ + l_2^+ - l_3^+ &= y_3 \\
l_0^- + l_1^- + l_2^- - l_3^- &= y_4 \\
l_0^+ - l_1^+ - l_2^+ + l_3^+ &= y_5 \\
l_0^- + l_1^- - l_2^- + l_3^- &= y_6 \\
l_0^- - l_1^- + l_2^- + l_3^- &= y_7 \\
l_0^+ + l_1^+ + l_2^+ + l_3^+ &= y_8
\end{aligned} \tag{74.326}$$

Tout naturellement, si nous considérons cette nouvelle notation comme des variables propres, ce système unique se sépare maintenant en deux sous-systèmes (les regroupements étant respectivement appelés des "**contrastes**" dans le domaine mais cela n'est pas conforme à la définition de la norme ISO 3534-3) pour être résoluble:

$$\begin{aligned}
l_0^- - l_1^- - l_2^- - l_3^- &= y_1 & l_0^+ + l_1^+ - l_2^+ - l_3^+ &= y_2 \\
l_0^- + l_1^- + l_2^- - l_3^- &= y_4 & l_0^+ - l_1^+ + l_2^+ - l_3^+ &= y_3 \\
l_0^- + l_1^- - l_2^- + l_3^- &= y_6 & l_0^+ - l_1^+ - l_2^+ + l_3^+ &= y_5 \\
l_0^- - l_1^- + l_2^- + l_3^- &= y_7 & l_0^+ + l_1^+ + l_2^+ + l_3^+ &= y_8
\end{aligned} \quad \text{et} \tag{74.327}$$

ce qui permet de diviser le nombre d'essais par deux par rapport à un plan complet. En résolvant un de ces deux systèmes, nous disons que les interactions sont "**aliasées**" avec les effets purs en négatif ou en positif (dans le cas présent, certains effets principaux sont confondus avec des interactions à deux facteurs).

Il est ensuite de tradition de garder que le système aliasé positivement:

$$\begin{aligned}
l_0^+ + l_1^+ - l_2^+ - l_3^+ &= y_2 \\
l_0^+ - l_1^+ + l_2^+ - l_3^+ &= y_3 \\
l_0^+ - l_1^+ - l_2^+ + l_3^+ &= y_5 \\
l_0^+ + l_1^+ + l_2^+ + l_3^+ &= y_8
\end{aligned} \tag{74.328}$$

car si les interactions sont nulles, nous retrouvons à l'identique la matrice d'expérience d'un plan factoriel complet 2^2 ! La démarche conduit donc à ne sélectionner que les essais 2, 3, 5 et 8, ce qui permet de diviser le nombre d'essais par deux par rapport à un plan complet. Ainsi, un plan factoriel fractionnaire d'une expérience à 3 facteurs peut être réduit à 4 essais avec cette méthode. Le plan factoriel fractionnaire ci-dessus sera représenté naturellement par la matrice d'expérience suivante:

Essai n°	Facteur 1	Facteur 2	Facteur 3	Réponse
2	+	-	-	y_2
3	-	+	-	y_3
5	-	-	+	y_5
8	+	+	+	y_8

Tableau: 74.24 - Plan d'expérience fractionnaire à 3 facteurs complet sous forme de Yates

Il y a cependant un problème : Même si l'interaction triple est réellement nulle, il peut rester jusqu'à 7 autres coefficients dans le modèle, alors que l'on ne dispose que de 4 résultats d'essais pour les identifier. Autrement dit, à moins que l'on sache a priori qu'au moins 3 de ces coefficients sont nuls (afin de se ramener à quatre équations avec 4 inconnues), on n'obtiendra au mieux que des relations entre les coefficients et l'identification rigoureuse sera impossible. Ainsi, il n'est pas possible de réduire indéfiniment le coût d'une étude expérimentale sans en dégrader la robustesse.

Il est important d'observer que dans la plan factoriel fractionnaire ci-dessus, le troisième facteur est confondu (peut être assimilé) avec l'interaction 12 des facteurs 1 et 2. Nous appelons cela "**l'alias initiale**" ou le "**générateur**" et nous pouvons remarquer en réitérant les calculs pour des plans factoriels à 4, 5, 6 ... facteurs que les générateurs permettent d'identifier immédiatement les essais à préserver (par ailleurs vous pouvez contrôler que lorsque les logiciels Minitab ou JMP vous imposent les générateurs, ils prennent par défaut les théoriques). Par exemple, le générateur du tableau ci-dessus s'écrira selon la tradition:

$$C=AB$$

Nous venons donc de voir qu'un plan $2^{(3-1)}$ présente le sérieux inconvénient de confondre un facteur principal avec une interaction d'ordre 2. Un plan $2^{(4-1)}$ présente seulement l'inconvénient de confondre un facteur principal avec une interaction d'ordre 3 et deux interactions d'ordre 2 entre elles. Un plan $2^{(5-1)}$ présente des inconvénients encore moindres. C'est pour cette raison que la théorie des plans factoriels utilise la notion de résolution. Plus la résolution est grande, plus le plan est précis.

Définitions:

D1. Lorsqu'aucun effet principal ne possède d'alias avec un autre effet principal, mais les effets principaux possèdent des alias avec des interactions à 2 facteurs, nous parlons de "**plan de résolution III**". D'un point de vue pratique les plans de résolution III ont surtout pour but de permettre des recherches exploratoires car ils permettent d'explorer un grand nombre de facteurs avec une économie certaine. À défaut d'obtenir un modèle suffisamment précis ils permettent éventuellement d'éliminer un grand nombre de facteurs dans un premier temps.

D2. Lorsqu'aucun effet principal ne possède d'alias avec un autre effet principal ou une autre interaction à 2 facteurs, mais certaines des interactions à 2 facteurs possèdent des alias avec d'autres interactions à 2 facteurs et des effets principaux possèdent des alias avec des interactions à 3 facteurs, nous parlons de "**plan de résolution IV**".

D3. Lorsqu'aucun effet principal ou aucune interaction à 2 facteurs ne possède d'alias avec un autre effet principal ou une autre interaction à 2 facteurs, mais des interactions à 2 facteurs possèdent des alias avec des interactions à 3 facteurs et des effets principaux possèdent des alias avec des interactions à 4 facteurs, nous parlons de "**plan de résolution V**".

et ainsi de suite...

Ce concept de résolution est important. Nous le retrouvons dans des logiciels comme Minitab 15.1.1 lors de sélection de plans factoriels fractionnaires:

	Factors														
Runs	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
4	Full	III													
8		Full	IV	II	II	III									
16			Full	V	IV	IV	IV	II	III	III	II	II	III	II	
32				Full	VI	IV	IV	IV	IV	IV	IV	IV	IV	IV	
64					Full	VII	V	IV	IV	IV	IV	IV	IV	IV	
128						Full	VIII	VI	V	V	IV	IV	IV	IV	

Figure: 74.40 - Affichage des choix de résolutions de plans factoriels fractionnaires dans Minitab 15.1.1

ou encore avec le logiciel Design Experts (partie du tableau mais qui est plus explicite que Minitab 15.1.1):

	3	4	5	6	7
4	2 ³⁻¹ III				
8	2 ³	2 ⁴⁻¹ IV	2 ⁵⁻² III	2 ⁶⁻³ III	2 ⁷⁻⁴ III
16		2 ⁴	2 ⁵⁻¹ V	2 ⁶⁻² IV	2 ⁷⁻³ IV

Figure: 74.41 - Affichage des choix de résolutions de plans factoriels fractionnaires dans Design Experts

Ainsi, le lecteur pourra observer qu'un plan complet de 5 facteurs à 32 essais peut être réduit à 16 essais en rassemblant les facteurs d'influence par paires ou 2-uplets (d'où la division par deux du nombre d'essais), ou encore à 8 essais en rassemblant les facteurs d'influence par 4-uplets.

Ensuite, c'est à l'expérimentateur de bien connaître son analyse et de savoir si:

1. Parmi les facteurs aliésés s'il y a des interactions ou non!

2. Dans le facteur aliasé l'influence forte sur la réponse vient de l'interaction ou de l'effet pur seul!

Une fois les coefficients déterminés, sous l'hypothèse que chacun des facteurs ou interactions est indépendant (hypothèse limite acceptable...) certains ingénieurs font une analyse de la variance de la droite de régression obtenue au final ou déterminent le coefficient de corrélation afin de déterminer si l'approximation linéaire du modèle est acceptable dans le domaine d'étude et d'application.

Concernant les générateurs des plans factoriels fractionnaires voici un petit tableau récapitulatif non exhaustif:

Facteurs	Essais	Résolution	Réduction	Générateurs	Alias
3	4	III	$2^{(3-1)}$	C=AB	I+ABC, A+BC, B+AC, C+AB
4	8	IV	$2^{(4-1)}$	D=ABC	I+ABCD, A+BCD, B+ACD, C+ABD, D+ABC, AB+CD, AC+BD, AD+BC
5	8	III	$2^{(5-2)}$	D=AB, E=AC	I+ABD+ACE+BCDE A+BD+CE+ABCDE B+AD+CDE+ABCE C+AE+BDE+ABCD D+AB+BCE+ACDE E+AC+BCD+ABDE BC+DE+ABE+ACD BE+CD+ABC+ADE
	16	V	$2^{(5-1)}$	E=ABCD	A+BCDE B+ACDE C+ABDE D+ABCE E+ABCD AB+CDE AC+BDE AD+BCE AE+BCD BC+ADE BD+ACE BE+ACD CD+ABE CE+ABD DE+ABC
...					

Tableau: 74.25 - Quelques plans factoriels fractionnaires avec générateurs et alias

Bien évidemment, utiliser des plans fractionnaires est un pari économique (et temporel). Si les conclusions sont claires, nous avons alors gagné du temps et réduit notre effort. Mais il arrive que l'on perde le pari. Nous pouvons dès lors utiliser un "plan complémentnaire" qui consiste à ajouter de manière pertinente suffisamment de lignes au plan initial pour désaliaser les coefficients désirés sur la base du choix du générateur d'alias.

Avant de passer à un autre type de plan, revenons juste sur le plan factoriel fractionnaire à 3 facteurs donc avec 4 essais pour des raisons pédagogiques. Considérons que nous avons fait ces 4 essais et que pour chacun, nous avons obtenu une mesure donnée dans la figure ci-dessous:

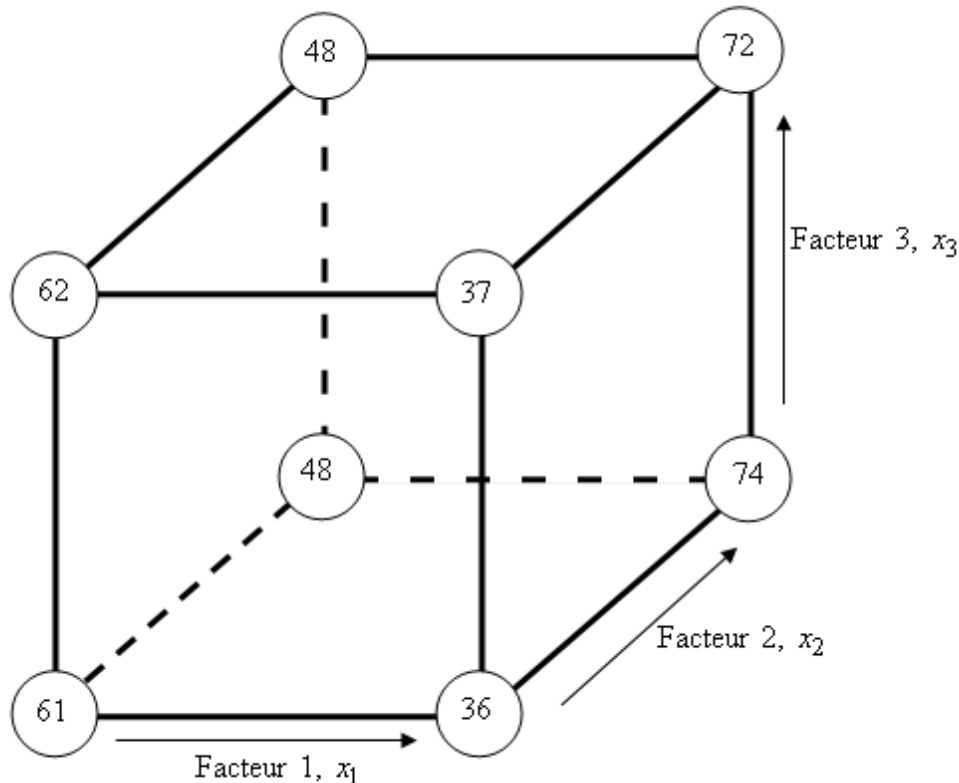


Figure: 74.42 - Représentation d'un plan factoriel fractionnaire réel à 3 facteurs

Nous voulons montrer au lecteur non pas comment déterminer le calcul des coefficients (résolution d'un simple système linéaire et puis il y aura un exemple à ce propos un peu plus loin) mais comment calculer les effets dans le cadre de ce cas particulier.

Ainsi, l'effet de x_1 est:

$$\frac{72 + 37 + 74 + 36}{4} - \frac{48 + 62 + 48 + 61}{4} = \frac{24 + (-25) + 26 + (-25)}{4} = 0 \quad (74.329)$$

et celui de x_2 est:

$$\frac{48 + 72 + 48 + 74}{4} - \frac{62 + 37 + 61 + 36}{4} = \frac{-14 + 35 + (-13) + 38}{4} = 11.5 \quad (74.330)$$

et celui de x_3 est:

$$\frac{48 + 72 + 62 + 37}{4} - \frac{48 + 74 + 61 + 36}{4} = \frac{0 + (-2) + 1 + 1}{4} = 0 \quad (74.331)$$

L'effet des interaction est un peu subtile dans le cas de 3 facteurs en utilisant la figure ci-dessous (nous verrons dans l'exemple plus bas qu'avec un tableau c'est beaucoup plus intuitif). Ainsi, nous avons pour l'interaction $x_1 x_2$:

$$\frac{72+74+62+61}{4} - \frac{48+37+48+36}{4} = 25 \quad (74.332)$$

et pour $x_1 x_3$:

$$\frac{37+72+48+61}{4} - \frac{48+62+74+36}{4} = -0.5 \quad (74.333)$$

et pour $x_2 x_3$:

$$\frac{48+72+61+36}{4} - \frac{62+37+48+74}{4} = -1 \quad (74.334)$$

Pour l'interaction triple $x_1 x_2 x_3$ avec la figure ci-dessous ce n'est pas non plus trivial (cela l'est un peu plus avec un table comme nous le verrons dans l'exemple plus bas). Nous avons:

$$\frac{72+62+48+36}{4} - \frac{48+37+74+61}{4} = -0.5 \quad (74.335)$$

3.3. PLANS ET NOMENCLATURE DE TAGUCHI

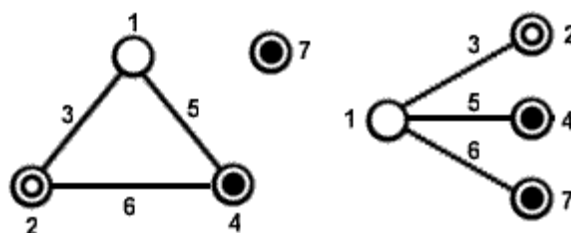
Les plans de Taguchi ne sont qu'une technique particulière pour retrouver des plans factoriels ou multifactoriels complets ou fractionnaires (à partir d'une matrice, d'une table triangulaire et d'un graphe linéaire. Avec les logiciels qui génèrent automatiquement les tables, cette technique est devenue un peu désuète mais elle avait l'avantage à l'époque de son utilisation de proposer une liste de plus d'une centaine de tables avec des facteurs multiples à 2 ou plusieurs niveaux avec ou sans interactions. Voyons en quelques exemples classiques pour la culture générale (car cela fait bien de les connaître et en plus c'est joli).

Taguchi proposa d'organiser les expériences selon des tables et des graphes qu'il appelle $L_4, L_8, \dots$ où L signifie "Level" et correspond aux nombres d'essais d'une expérience associés à des graphes permettant d'identifier les interactions (les tables de Taguchi incluent certains plans de Plackett-Burman).

Commençons l'exemple avec une table que nous connaissons bien, la table L_8 . Chez Taguchi cette table peut être lue comme la table d'un plan factoriel fractionnaire de 7 facteurs (donc sans interactions) où comme la table d'un plan factoriel complet pour 3 facteurs. Raison pour laquelle est notée $L_8(2^7)$ tantôt $L_8(2^3)$:

Essais	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

Tableau: 74.26 - Table L8 avec notation factorielle d'usage et graphes linéaires correspondants



Nous allons voir si vraiment elle diffère de ce que nous connaissons déjà. D'abord, chaque facteur ne prend aussi que deux niveaux dans ce tableau, nous pouvons donc remplacer la notation de Taguchi avec notre notation d'usage pour les plans factoriels en remplaçant les 1 par des "+" et les 2 par des "-":

Essais	1	2	3	4	5	6	7
1	+	+	+	+	+	+	+
2	+	+	+	-	-	-	-
3	+	-	-	+	+	-	-
4	+	-	-	-	-	+	+
5	-	+	-	+	-	+	-
6	-	+	-	-	+	-	+
7	-	-	+	+	-	-	+
8	-	-	+	-	+	+	-

Tableau: 74.27 - Table L8 avec notation factorielle d'usage

Nous voyons déjà beaucoup mieux que toutes les lignes et colonnes sont orthogonales prises deux à deux (matrice d'Hadamard). Nous ajoutons une colonne de "+":

Essais	1	2	3	4	5	6	7	Repos
1	+	+	+	+	+	+	+	+
2	+	+	+	-	-	-	-	+
3	+	-	-	+	+	-	-	+
4	+	-	-	-	-	+	+	+
5	-	+	-	+	-	+	-	+
6	-	+	-	-	+	-	+	+
7	-	-	+	+	-	-	+	+
8	-	-	+	-	+	+	-	+

Tableau: 74.28 - Table L8 avec notation factorielle d'usage avec colonne supplémentaire manquante

et nous remarquons que nous retrouvons le plan factoriel complet de 3 facteurs à deux niveaux si nous renotons la ligne de titre des colonnes de la manière suivante:

Essais	F3	F2	F23	F1	F13	F12	F123	Repos
1	+	+	+	+	+	+	+	+
2	+	+	+	-	-	-	-	+
3	+	-	-	+	+	-	-	+
4	+	-	-	-	-	+	+	+
5	-	+	-	+	-	+	-	+
6	-	+	-	-	+	-	+	+
7	-	-	+	+	-	-	+	+
8	-	-	+	-	+	+	-	+

Tableau: 74.29 - Tableau précédent avec notation traditionnelle pour la ligne de titre

et que cela correspond aussi au plan factoriel fractionnaire d'un plan d'expérience de 7 facteurs à deux niveaux (sans aucune interaction).

La question que se posera bien évidemment le lecteur, c'est comment aurions nous fait pour identifier quel facteur appartenait à quelle colonne dans le cadre d'un plan factoriel complet à 3 facteurs si nous ne connaissions pas le tableau établi avec les techniques vues plus haut? Eh bien en utilisant le graphe linéaire suivant:

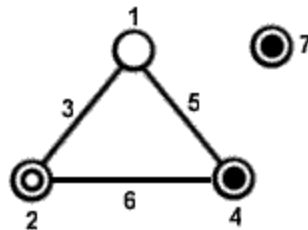


Figure: 74.43 - Graphe linéaire de Taguchi L8

qui indique les facteurs principaux dans les sommets du graphe (donc les colonnes 1, 2 et 4 sont des facteurs principaux). La colonne 3 est l'interaction entre les colonnes 1 et 2 (d'où le fait qu'elle se trouve sur l'arête qui sépare les deux sommets), la colonne 5 est l'interaction entre les colonnes 1 et 4 (d'où le fait qu'elle se trouve sur l'arête qui sépare les deux sommets), et la colonne 6 est l'interaction entre les colonnes 2 et 4 (d'où le fait qu'elle se trouve sur l'arête qui sépare les deux sommets). Le fait que le 7 soit à l'extérieur, c'est parce qu'une triple interaction ne peut pas être représentée avec cette technique de graphe planaire. On utilise alors les symboles des sommets pour signifier que c'est la superposition du sommet 1, 2 et 4 (cercle+anneau+disque).

Le deuxième graphe linéaire associé à cette table permet de mettre en évidence un autre usage possible:

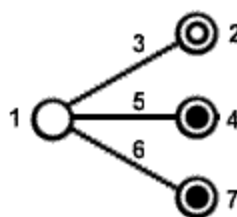


Figure: 74.44 - Graphe linéaire de Taguchi L8

C'est-à-dire d'en faire usage pour une analyse de 4 facteurs (toujours à deux modalités dans le cas présent) représentées par les colonnes 1, 2, 4, 7 (le lecteur pourra effectivement vérifier que ces quatre colonnes correspondent bien à un plan factoriel fractionnaire avec pour générateur 12 pour 4 facteurs soit à la main, soit avec un logiciel) dont 3 interactions (1 et 2; 1 et 4; 1 et 7).

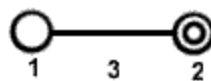
Remarque: Il faut faire attention car il y a donc par exemple une table de Taguchi notée donc L_8 pour 3 facteurs à 2 niveaux complète et 7 facteurs à 2 niveaux sans interactions comme nous venons de le voir (donc avec 7 colonnes), mais il y a aussi des tables L_8 pour 5 facteurs décomposés en 4 facteurs à 3 niveaux et 1 facteur à 4 niveaux (sans interactions) mais seulement avec 5 colonnes. Raison pour laquelle les livres listant les tables de Taguchi spécifient normalement explicitement le cadre d'application des tables.

Voyons quelques autres tables (une infime partie de la liste complète)

- Table $L_4(2^3)$:

Essais	F1	F2	F3
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1

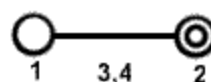
Tableau: 74.30 - Table L4 de Taguchi



- Table $L_9(3^4)$:

Essais	F1	F2	F3	F4
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	2	1	2	3
5	2	2	3	1
6	2	3	1	2
7	3	1	3	2
8	3	2	1	3
9	3	3	2	1

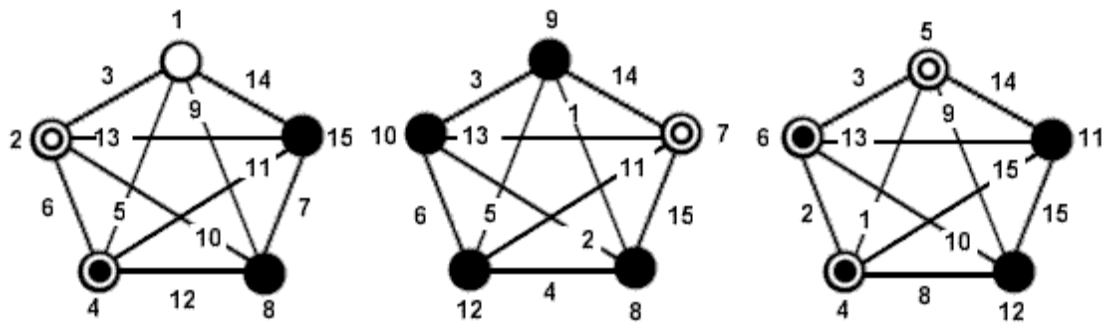
Tableau: 74.31 - Table L9 de Taguchi



- Table $L_{16}(2^{15})$:

Essais	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2
3	1	1	1	2	2	2	2	1	1	1	1	2	2	2	2
4	1	1	1	2	2	2	2	2	2	2	2	1	1	1	1
5	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2
6	1	2	2	1	1	2	2	2	2	1	1	2	2	1	1
7	1	2	2	2	2	1	1	1	1	2	2	2	2	1	1
8	1	2	2	2	2	1	1	2	2	1	1	1	1	2	2
9	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
10	2	1	2	1	2	1	2	2	1	2	1	2	1	2	1
11	2	1	2	2	1	2	1	1	2	1	2	2	1	2	1
12	2	1	2	2	1	2	1	2	1	2	1	1	2	1	2
13	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1
14	2	2	1	1	2	2	1	2	1	1	2	2	1	1	2
15	2	2	1	2	1	1	2	1	2	2	1	2	1	1	2
16	2	2	1	2	1	1	2	2	1	1	2	1	2	2	1

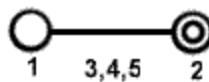
Tableau: 74.32 - Table L16 de Taguchi



- Table $L_{16}(4^5)$:

Essais	F1	F2	F3	F4	F5
1	1	1	1	1	1
2	1	2	2	2	2
3	1	3	3	3	3
4	1	4	4	4	4
5	2	1	2	3	4
6	2	2	1	4	3
7	2	3	4	1	2
8	2	4	3	2	1
9	3	1	3	4	2
10	3	2	4	3	1
11	3	3	1	2	4
12	3	4	2	1	3
13	4	1	4	2	3
14	4	2	3	1	4
15	4	3	2	4	1
16	4	4	1	3	2

Tableau: 74.33 - Table L16 de Taguchi



etc.

Les tables de Taguchi accompagnées de leurs graphes sont donc des plans factoriels fractionnaires ou complets avec tous leurs avantages. Le mérite de Taguchi est d'avoir essayé de simplifier l'utilisation des plans factoriels pour les rendre accessibles à un grand nombre d'expérimentateurs.

Enfin résumons ce que nous avons vu jusqu'à maintenant avec les noms respectifs (à vérifier car il n'y pas de norme claire à notre connaissance concernant ces définitions) dans l'ordre de généralisation:

- les plans qui contiennent uniquement des facteurs à deux niveaux, sous le nom de "**plans factoriels**" qu'ils soient avec un modèle linéaire ou non-linéaire, additif ou non.
- les plans factoriels qui permettent uniquement de déterminer les coefficients des facteurs principaux (donc modèle additif: sans interactions) sous le nom de "**plans de Koshal**" ou "**plans de criblage**" (appelés aussi "plans un à la fois" car comme la technique du "un à la fois" il ne permettent d'analyser aucune interaction mais la comparaison s'arrête là!).
- les plans factoriels où toutes les interactions d'ordre trois et supérieures sont négligées sous le nom de "**plans de Rechtschaffner**".
- les plans factoriels où les coefficients ont été aliasés tout en gardant les interactions sous le nom de "**plans factoriels fractionnaires**", ou encore de "**plans de Box et Hunter**" ou encore sous le nom de "**plans basés sur la matrice d'Hadamard**". Avec les plans factoriels fractionnaires, il convient de préciser aussi la résolution.

- les plans factoriels dont l'ordre est un multiple de quatre mais pas une puissance de deux (donc 12, 16, 20, 24, etc.) sont appelés "**plans de Plackett-Burman**".
- les plans avec un nombre quelconque de facteurs, d'interactions d'ordre quelconque et à un nombre de niveaux quelconque sous le nom de "**plans de Fisher**" (en hommage au créateur d'origine du concept de plan d'expérience).
- les plans qui contiennent autant d'essais que de coefficients à déterminer sous le nom de "**plans saturés**".
- les plans qui contiennent moins d'essais que de coefficients à déterminer sous le nom de "**plans sursaturés**".

Un aspect mérite encore d'être précisé: c'est la vérification de la validité du modèle mathématique du premier degré. Or aucun de ces plans ne prévoit un tel test de validité utilisant des statistiques élaborées. C'est pourquoi il est préconisé de toujours ajouter au moins un point expérimental au contre du domaine expérimental. La valeur de la réponse en ce point sera comparée à la valeur déduite des autres points expérimentaux grâce au modèle mathématique. Si les deux valeurs sont semblables, le modèle mathématique sera adopté, si elles ne le sont pas nous devons rejeter ce modèle et compléter les résultats déjà obtenus par des expériences permettant de passer au second degré.

Enfin, pour terminer, faisons un exemple particulier et relativement complet d'un plan factoriel complet général avec des facteurs à 3 niveaux.

Exemple:

Dans le cadre de l'étude de pneus, le critère retenu sera la longévité (nombre de kilomètres parcourus avant que le pneu soit usé) et les facteurs pourront être l'usage (ville ou route), la vitesse moyenne (40 [km/h] ou 50 [km/h]) et la pression de gonflage (1.5 ou 2 ou 2.5 [kg]).

La deuxième étape est la création du plan d'expérience. Dans un plan factoriel complet, nous croisons les différents niveaux de tous les facteurs de façon à tester chaque combinaison. Par exemple pour les pneus, nous serons amenés à réaliser $2 \times 2 \times 3 = 12$ expériences au minimum si nous voulons travailler avec un plan complet. Par ailleurs, il est plus sain de répéter les expériences (pour analyser la dispersion des valeurs), donc d'en faire 24, 36, etc.

Nous noterons comme à l'habitude y la variable étudiée et dans le cas présent x_1, x_2, x_3 les variables facteurs. Nous recherchons un modèle simple sous la forme:

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \dots \quad (74.336)$$

Comme chacun des facteurs peut avoir plusieurs niveau, nous aurons par exemple la valeur y au niveau 2 du facteur 1 et au niveau 1 du facteur 2 et au niveau 3 du facteur 3 qui sera noté:

$$y = a_0 + a_1 x_{1,2} + a_2 x_{2,1} + a_3 x_{3,3} \quad (74.337)$$

Le plan factoriel complet est l'exposé des types d'expériences suivant les niveaux. Par exemple, pour l'étude des pneus nous aurons:

- Pour le facteur 1 (x_1) lieu: ville = 1, route = 2

- Pour le facteur 2 (x_2) vitesse: 40 [km/h] = 1, 50 [km/h] = 2

- Pour le facteur 3 (x_3) pression de pneus: 1.5 [kg] = 1, 2 [kg] = 2, 2.5 [kg] = 3

La matrice d'expérimentation sera alors (c'est exactement la même que celle que l'on obtient en général avec Minitab un plan factoriel complet pour 3 facteurs avec 2 à deux niveaux et 1 un trois niveaux sans randomisation):

Essai N°	x_1	x_2	x_3
1	1	1	1
2	1	1	2
3	1	1	3
4	1	2	1
5	1	2	2
6	1	2	3
7	2	1	1
8	2	1	2
9	2	1	3
10	2	2	1
11	2	2	2
12	2	2	3

Tableau: 74.34 - Matrice d'expérimentation (ou tableau factoriel/plan d'expérience)

À ce tableau, nous pourrions associer les résultats des expériences. Par exemple, pour les pneus, nous avons fait trois essais (trois "réplications") pour chacune des conditions de niveaux et nous avons obtenu:

N° Essai	x_1	x_2	x_3	Valeurs	de	y
1	1	1	1	32'700	32'750	32'960
2	1	1	2	33'430	33'360	32'910
3	1	1	3	31'710	32'100	32'220
4	1	2	1	32'680	32'270	33'130
5	1	2	2	34'070	33'100	33'610
6	1	2	3	33'220	33'700	33'285
7	2	1	1	33'180	32'160	32'640
8	2	1	2	34'430	34'280	34'460
9	2	1	3	33'570	33'300	32'570
10	2	2	1	33'270	33'080	32'415
11	2	2	2	33'440	33'570	34'204
12	2	2	3	32'840	33'210	32'470

Tableau: 74.35 - Matrice d'expérimentation avec valeurs expérimentales

À l'aide de ces valeurs, nous pouvons calculer les effets des différents facteurs. Nous partons du principe que leur moyenne est nulle pour un facteur donné. Donc la valeur moyenne arithmétique globale (qui peut être calculée comme la moyenne arithmétique si et seulement si les variables sont codées):

$$\bar{y} = 33'119.3 \quad (74.338)$$

est le coefficient constant du modèle:

$$x_0 = 33'119.3 \quad (74.339)$$

L'effet de x_1 au niveau 1 est la variation de cette moyenne quand on ne considère que les cas où x_1 est au niveau 1. La moyenne arithmétique devient 32'955.8, donc l'effet est:

$$32'955.8 - 33'119.3 = -163.4 \quad (74.340)$$

Nous noterons cela:

$$x_{1,1} = -163.4 \quad (74.341)$$

Au niveau 2 de x_1 , la moyenne devient 33'282.7 donc l'effet de x_1 au niveau 2 vaut:

$$x_{1,2} = 33'282.7 - 33'119.3 = 163.4 \quad (74.342)$$

Nous remarquons que la somme des effets de x_1 est nulle, car c'est la somme des écarts à la moyenne de sous-populations de même taille (donc la moyenne globale est la moyenne arithmétique des moyennes des niveaux).

Exactement de la même façon, nous trouvons:

$$\begin{aligned} x_{2,1} &= -78.7 & x_{2,2} &= 78.7 \\ x_{3,1} &= -349.7 & x_{3,2} &= 619.4 & x_{3,3} &= -269.7 \end{aligned} \quad (74.343)$$

et nous retrouvons aussi que la somme des effets sur une variable est nulle.

Nous modélisons ces résultats sous la forme utilisant la moyenne de réponse:

$$y = 33'119.3 + x_1 \begin{bmatrix} -163.4 \\ 163.4 \end{bmatrix} + x_2 \begin{bmatrix} -78.7 \\ 78.7 \end{bmatrix} + x_3 \begin{bmatrix} -349.7 \\ 619.4 \\ -269.7 \end{bmatrix} \quad (74.344)$$

qui donne le modèle trouvé. Fonction que nous pouvons récrire sous la forme d'effet moyen (la forme la plus intéressante mathématiquement parlant):

$$y = 33'119.3 + 163.4x_1 + 78.7x_2 + 619.4x_3 \quad (74.345)$$

Résultat intermédiaire à comparer avec une régression linéaire simple qui est (résultat parfaitement identique entre Microsoft Excel 14.0.6117 et Minitab 15.1.1):

$$y \cong 32313.8 + 326.9x_1 + 157.4x_2 + 40x_3 \quad (74.346)$$

mais qui n'a pas de sens puisque la régression linéaire n'est pas construite pour des facteurs à plus de deux niveaux.

De ces résultats, nous tirons déjà la meilleure stratégie d'utilisation des pneus: sur route, à 50 [km/h], et avec un pneu gonflé à 2 [kg], nous pouvons espérer faire:

$$y = x_0 + x_{1,2} + x_{2,2} + x_{3,2} = 33'119.3 + 163.4 + 78.7 + 619.4 \cong 33'981 \quad (74.347)$$

Remarquons tout de suite que dans certains cas, l'expérimentation a fait mieux, même en ville ou à 40 [km/h]. Ceci relève de la dispersion évidente des résultats. Mais alors, les effets trouvés ne viennent-ils pas seulement de cette dispersion? D'autre part nous avons étudié les effets indépendamment les uns des autres. Mais il se pourrait qu'ils se renforcent. Nous allons étudier ces questions dans la suite.

Reprenons nos 36 expériences, en comparant les valeurs obtenues et les prévisions du modèle:

x_1	x_2	x_3	Mesuré	Modèle	Écart	x_1	x_2	x_3	Mesuré	Modèle	Écart
1	1	1	32'700	32'527.5	172.5	2	1	1	32'160	32'854.3	-694.3
1	1	2	33'430	33'496.6	-66.6	2	1	2	34'280	33'823.4	456.6
1	1	3	31'710	32'607.5	-897.5	2	1	3	33'300	32'934.3	365.7
1	2	1	32'680	32'684.9	-4.9	2	2	1	33'080	33'011.7	68.3
1	2	2	34'070	33'654	416	2	2	2	33'570	33'980.8	-410.8
1	2	3	33'220	32'764.9	455.1	2	2	3	33'210	33'091.7	118.3
2	1	1	33'180	32'854.3	325.7	1	1	1	32'960	32'527.5	432.5
2	1	2	34'430	33'823.4	606.6	1	1	2	32'910	33'496.6	-586.6
2	1	3	33'570	32'934.3	635.7	1	1	3	32'220	32'607.5	-387.5
2	2	1	33'270	33'011.7	258.3	1	2	1	33'130	32'684.9	445.1
2	2	2	33'440	33'980.8	-540.8	1	2	2	33'610	33'654	-44
2	2	3	32'840	33'091.7	-251.7	1	2	3	33'285	32'764.9	520.1
1	1	1	32'750	32'527.5	222.5	2	1	1	32'640	32'854.3	-214.3
1	1	2	33'360	33'496.6	-136.6	2	1	2	34'460	33'823.4	636.6
1	1	3	32'100	32'607.5	-507.5	2	1	3	32'570	32'934.3	-364.3
1	2	1	32'270	32'684.9	-414.9	2	2	1	32'415	33'011.7	-596.7
1	2	2	33'100	33'654	-554	2	2	2	34'204	33'980.8	223.2
1	2	3	33'700	32'764.9	935.1	2	2	3	32'470	33'091.7	-621.7

Tableau: 74.36 - Matrice d'expérience comparant expérience et modèle

Nous constatons que les écarts au modèle sont parfois importants. Cela peut venir d'une dispersion aléatoire naturelle des valeurs, ou bien d'un modèle inadapté. Cela fait penser à un modèle inadapté, ou bien cela pourrait être dû à ce que des facteurs conjugués ont plus d'effet que séparés. Les effets positifs ou négatifs de 2 facteurs peuvent faire plus que s'additionner (modèle purement additif à rejeter peut-être).

Il y a donc probablement des interactions (les chimistes parlent de "potentialisation"). Pour voir s'il y en a des interactions d'ordre deux, un bon moyen c'est d'étudier tous les facteurs par paires (et faire abstraction de l'existence des autres). Ainsi, si nous commençons avec les facteurs (lieu) x_1 et (vitesse) x_2 en faisant abstraction de la pression des pneus, nous aurons (l'explication se trouve après la première ligne):

Niveaux	Moyenne expérimentale	Effet total	Somme des effets	Effet de l'interaction
$x_{1,1} x_{2,1}$	32'682.2	-437.1	-242.2	-194.9
$x_{1,1} x_{2,2}$	33'229.4	110.2	-84.7	194.9
$x_{1,2} x_{2,1}$	33'398.9	279.6	84.7	194.9
$x_{1,2} x_{2,2}$	33'166.6	47.3	242.2	-194.9

Tableau: 74.37 - Tableau d'interaction facteurs 1 et 2

Donc, 32'682.2 est la moyenne arithmétique des expériences faites au niveau $x_{1,1}$ (Ville) et $x_{2,1}$ (40 [km/h]). En soustrayant la moyenne (toujours 33'119.3), Nous obtenons un effet de -437.1, alors que l'effet de x_1 (-163.4) plus l'effet de x_2 (-78.7) ne donne que -242.2.

Il y a donc un surplus de -194.9, dû en grande partie à l'interaction $x_{1,1} x_{2,1}$ (notez que c'est la même pour $x_{1,2} x_{2,2}$, et l'opposée pour $x_{1,1} x_{2,2}$ et $x_{1,2} x_{2,1}$). Notons tout de suite que l'interaction $x_{1,1} x_{2,1}$ a un effet bien supérieur aux effets de x_1 et de x_2 !

Voyons les deux autres interactions possibles:

Niveaux	Moyenne expérimentale	Effet total	Somme des effets	Effet de l'interaction
$x_{1,1} x_{3,1}$	32'748.3	-370.9	-513.1	142.2
$x_{1,1} x_{3,2}$	33'413.3	294.1	455.9	-161.9
$x_{1,1} x_{3,3}$	32'705.8	-413.4	-433.1	19.7
$x_{1,2} x_{3,1}$	32'790.8	-328.4	-186.3	-142.2
$x_{1,2} x_{3,2}$	34'064.0	944.7	782.8	161.9
$x_{1,2} x_{3,3}$	32'993.3	-125.9	-106.3	-19.7

Tableau: 74.38 - Tableau d'interaction facteurs 1 et 3

Niveaux	Moyenne expérimentale	Effet total	Somme des effets	Effet de l'interaction
$x_{2,1} x_{3,1}$	32'731.7	-387.6	-428.4	40.8
$x_{2,1} x_{3,2}$	33'811.7	692.4	540.7	151.7
$x_{2,1} x_{3,3}$	32'578.3	-540.9	-348.4	-192.5
$x_{2,2} x_{3,1}$	32'807.5	-311.8	-271.0	-40.8
$x_{2,2} x_{3,2}$	33'665.7	546.4	698.1	-151.7
$x_{2,2} x_{3,3}$	33'120.8	1.6	-191.0	192.5

Tableau: 74.39 - Tableau d'interaction facteurs 2 et 3

Nous pouvons représenter les interactions différemment, en croisant les facteurs:

	$x_{2,1}$	$x_{2,2}$
$x_{1,1}$	-194.9	194.9
$x_{1,2}$	194.9	-194.9

	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$
$x_{1,1}$	142.2	-161.9	19.7
$x_{1,2}$	-142.2	161.9	-19.7

	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$
$x_{2,1}$	40.8	151.7	-192.5
$x_{2,2}$	-40.8	-151.7	192.5

Nous remarquons alors que chaque ligne ou chaque colonne a un total nul. Nous voyons aussi que, dans ce cas, il est difficile de négliger les interactions, et que x_2 (par exemple) intervient plus par ses interactions que par son effet propre!

Nous pouvons encore essayer de vérifier si l'interaction des trois facteurs a un effet notable. L'expérience ayant été répétée 3 fois à chaque niveau (nous parlons alors de "trois blocs"), nous pouvons calculer la moyenne de y , et lui soustraire le modèle avec les interactions calculé ci-dessus:

$$y = 33'119.3 + x_1 \begin{bmatrix} -163.4 \\ 163.4 \end{bmatrix} + x_2 \begin{bmatrix} -78.7 \\ 78.7 \end{bmatrix} + x_3 \begin{bmatrix} -349.7 \\ 619.4 \\ -269.7 \end{bmatrix} + x_1 x_2 \begin{bmatrix} -194.9 \\ 194.9 \\ 194.9 \\ -194.9 \end{bmatrix} + x_1 x_3 \begin{bmatrix} 142.2 \\ -161.9 \\ 19.7 \\ -142.2 \\ 161.9 \\ -19.7 \end{bmatrix} + x_2 x_3 \begin{bmatrix} 40.8 \\ 151.7 \\ -192.5 \\ -40.8 \\ -151.7 \\ 192.5 \end{bmatrix} \quad (74.348)$$

qu'il est d'usage d'écrire par souci de simplification évidente:

$$y = 33'119.3 + x_1 \begin{bmatrix} -163.4 \\ 163.4 \end{bmatrix} + x_2 \begin{bmatrix} -78.7 \\ 78.7 \end{bmatrix} + x_3 \begin{bmatrix} -349.7 \\ 619.4 \\ -269.7 \end{bmatrix} + x_1 x_2 \begin{bmatrix} -194.9 \\ 194.9 \end{bmatrix} + x_1 x_3 \begin{bmatrix} 142.2 \\ -161.9 \\ 19.7 \end{bmatrix} + x_2 x_3 \begin{bmatrix} 40.8 \\ 151.7 \\ -192.5 \end{bmatrix} \quad (74.349)$$

Remarque: Il suffit de savoir que pour les coefficients d'interaction les effets manquants (non signalés) sont opposés en signe.

Certains logiciels de statistiques vont plus loin dans la simplification d'écriture (toujours dans la même idée que les autres coefficients sont opposés) et écrivent:

$$y = 33'119.3 + x_1 [-163.4] + x_2 [-78.7] + x_3 \begin{bmatrix} -349.7 \\ 619.4 \\ -269.7 \end{bmatrix} + x_1 x_2 [-194.9] + x_1 x_3 \begin{bmatrix} 142.2 \\ -161.9 \\ 19.7 \end{bmatrix} + x_2 x_3 \begin{bmatrix} 40.8 \\ 151.7 \\ -192.5 \end{bmatrix} \quad (74.350)$$

et d'autres logiciels (comme Minitab 15.1.1 par exemple) vont encore plus loin en écrivant:

$$y = 33'119.3 + x_1 [-163.4] + x_2 [-78.7] + x_3 \begin{bmatrix} -349.7 \\ 619.4 \end{bmatrix} + x_1 x_2 [-194.9] + x_1 x_3 \begin{bmatrix} 142.2 \\ -161.9 \end{bmatrix} + x_2 x_3 \begin{bmatrix} 40.8 \\ 151.7 \end{bmatrix} \quad (74.351)$$

car les valeurs manquantes s'obtiennent par addition et en inversant le signe (encore faut-il le savoir...). Nous ne pouvons donc pas à cause de la présence du facteur à 3 niveaux réduire rigoureusement le modèle mathématique à une simple fonction. Raison pour laquelle, nous nous retrouvons avec la symbolique visible dans la relation précédente.

Nous pouvons alors résumer ces résultats sous la forme:

N° Essai	x_1	x_2	x_3	Valeurs	de	y	Moyenne	Modèle	Écart
1	1	1	1	32'700	32'750	32'960	32'803.3	32'515.6	287.8
2	1	1	2	33'430	33'360	32'910	33'233.3	33'291.5	-58.1
3	1	1	3	31'710	32'100	32'220	32'010.0	32'239.8	-229.8
4	1	2	1	32'680	32'270	33'130	32'693.3	32'981.2	-287.8
5	1	2	2	34'070	33'100	33'610	33'593.3	33'535.3	58.1
6	1	2	3	33'220	33'700	33'285	33'401.7	33'172.0	229.7
7	2	1	1	33'180	32'160	32'640	32'660.0	32'947.8	-287.8
8	2	1	2	34'430	34'280	34'460	34'390.0	34'331.9	58.1
9	2	1	3	33'570	33'300	32'570	33'146.7	32'917.0	229.7
10	2	2	1	33'270	33'080	32'415	32'921.7	32'633.8	287.9
11	2	2	2	33'440	33'570	34'204	33'738.0	33'796.1	-58.1
12	2	2	3	32'840	33'210	32'470	32'840.0	33'069.6	-229.6

Tableau: 74.40 - Matrice d'expérience avec mesures, moyenne, modèle et écart

Nous nous apercevons alors que les valeurs du modèle s'écartent toujours assez nettement des valeurs moyennes, ce qui peut s'interpréter soit par le fait que les effets ne sont pas linéaires, soit par une interaction entre les trois facteurs.

Il conviendrait alors de refaire les calculs effectués plus haut avec l'interaction des trois facteurs. Mais comme c'est toujours le même principe, nous le ferons que si un lecteur ne nous le demande explicitement.

Mais il existe encore une cause d'erreur dans le modèle, c'est de tenir compte d'effets qui n'interviennent pas en réalité. Pour être sûr que l'effet calculé sur une variable ou sur une interaction est réel, nous utiliserons pour commencer l'analyse de la variance à une voie (ANOVA) que nous avons étudié en détails dans le chapitre de Statistiques. Pour cela, il est important d'avoir répété les expériences, de façon à pouvoir mettre en évidence la dispersion due aux facteurs non contrôlés.

Par exemple, le facteur x_2 a un effet plutôt faible. A-t-il une influence véritable? Pour cela, nous séparons les 36 expériences en deux échantillons (niveau $x_{2,1}$ et niveau $x_{2,2}$), et nous calculons les sommes de carrés et les degrés de liberté.

sommes de carrés	ddl	variances	F	
contrôlée	223'098.78	1	calculé:	0.485
résiduelle	15'645'288.44	34	limite (5%):	4.13
totale	15'868'387.22	35	test réussi	

Tableau: 74.41 - Tableau d'ANOVA

Ce qui donne avec Minitab 15.1.1 (nous pouvons y constater que le modèle n'est très probablement pas linéaire):

ANOVA à un facteur contrôlé : 40 km/h; 50 km/h

Source	DL	Somme des carrés	CM	F	P
Facteur	1	223099	223099	0.48	0.491
Erreur	34	15645288	460156		
Total	35	15868387			

S = 678.3 R carré = 1.41 % R carré (ajust) = 0.00 %

				Limites de confiance = 95 % distinctes pour la moyenne en fonction de l'écart type regroupé			
Niveau	N	Moyenne	EcTyp	+-----+-----+-----+-----+			
40 km/h	18	33041	797	{-----*-----}			
50 km/h	18	33198	533	{-----*-----}			
				+-----+-----+-----+-----+			
				32750	33000	33250	33500

Ecart type regroupé = 678

Le test ayant réussi, nous concluons que le facteur x_2 n'est sans doute pas influent (car le test avait comme hypothèse: le facteur contrôlé n'est pas influent). Il en est de même pour le facteur x_1 :

sommes de carrés		ddl	variances	F	
contrôlée	961'707.1	1	961'707	calculé:	2.19
résiduelle	14'906'680.1	34	438'432	limite (5%):	4.13
totale	15'868'387.2	35		test réussi	

Tableau: 74.42 - Tableau d'ANOVA

Ce qui donne avec Minitab 15.1.1 (nous pouvons aussi y constater que le modèle n'est très probablement pas linéaire):

ANOVA à un facteur contrôlé : Ville; Route

Source	DL	Somme des carrés	CM	F	P
Facteur	1	961707	961707	2.19	0.148
Erreur	34	14906680	438432		
Total	35	15868387			

S = 662.1 R carré = 6.06 % R carré (ajust) = 3.30 %

				Limites de confiance = 95 % distinctes pour la moyenne en fonction de l'écart type regroupé			
Niveau	N	Moyenne	EcTyp	+-----+-----+-----+-----+			
Ville	18	32956	610	{-----*-----}			
Route	18	33283	711	{-----*-----}			
				+-----+-----+-----+-----+			
				32750	33000	33250	33500

Ecart type regroupé = 662

Par contre, le facteur x_3 est, lui, bel et bien influent:

sommes de carrés		ddl	variances	F	
contrôlée	6'943'966,7	2	3'471'983	calculé:	12.84
résiduelle	8'924'420,5	33	270'437	limite (5%):	3.28
totale	15'868'387.2	35		test échoué	

Tableau: 74.43 - Tableau d'ANOVA

Ce qui donne avec Minitab 15.1.1 (nous pouvons y constater que le modèle n'est très probablement pas linéaire):

ANOVA à un facteur contrôlé : 1.5 kg; 2 kg; 2.5 kg

Source	DL	Somme des carrés	CM	F	P
Facteur	2	6943967	3471983	12.84	0.000
Erreur	33	8924421	270437		
Total	35	15868387			

S = 520.0 R carré = 43.76 % R carré (ajust) = 40.35 %

				Limites de confiance = 95 % distinctes pour la moyenne en fonction de l'écart type regroupé			
Niveau	N	Moyenne	EcTyp	-----+-----+-----+-----+			
1.5 kg	12	32770	364	({-----*-----})			
2 kg	12	33739	530	({-----*-----})			
2.5 kg	12	32850	631	({-----*-----})			
				-----+-----+-----+-----+			
				32800	33200	33600	34000

Ecart type regroupé = 520

Rappelons en effet que l'échec du test correspond à la mise en évidence d'un facteur influent, celui qui est contrôlé. Alors, le modèle de base à considérer (on laisse de côté les interactions) n'est plus :

$$y = 33'119.3 + x_1 \begin{bmatrix} -163.4 \\ 163.4 \end{bmatrix} + x_2 \begin{bmatrix} -78.7 \\ 78.7 \end{bmatrix} + x_3 \begin{bmatrix} -349.7 \\ 619.4 \\ -269.7 \end{bmatrix} \quad (74.352)$$

mais:

$$y = 33'119.3 + x_3 \begin{bmatrix} -349.7 \\ 619.4 \\ -269.7 \end{bmatrix} \quad (74.353)$$

Attention!!! Le fait que x_1 et x_2 ne soient pas influents n'a pas comme conséquence que les interactions entre x_1 et x_2 , ou x_1 et x_3 , ou même x_1x_2 et x_3 ne le sont pas. Par exemple, ici, x_1x_2 n'est pas influent (le test de Fischer-Snedecor réussit), mais au niveau 1 de x_3 , x_1x_2 l'est (au sens $x_1 = x_2$ contre $x_1 \neq x_2$).

Voyons cela avec Minitab 15.1.1:

ANOVA : Response en fonction de Lieu; Vitesse; Pression

Facteur	Type	Niveaux	Valeurs
Lieu	fixe	2	Route; Ville
Vitesse	fixe	2	40; 50
Pression	fixe	3	1.5; 2; 2.5

Analyse de la variance pour Response

Source	DL	Somme des carrés	CM	F	P
Lieu	1	961707	961707	6.78	0.016
Vitesse	1	223099	223099	1.57	0.222
Pression	2	6943967	3471983	24.49	0.000
Lieu*Vitesse	1	1367340	1367340	9.65	0.005
Vitesse*Pression	2	741020	370510	2.61	0.094
Lieu*Pression	2	561782	280891	1.98	0.160
Lieu*Vitesse*Pression	2	1667622	833811	5.88	0.008
Erreur	24	3401851	141744		
Total	35	15868387			

S = 376.489 R carré = 78.56 % R carré (ajust) = 68.74 %

Nous voyons qu'en faisant une ANOVA multifactorielle, que finalement le facteur x_1 est statistiquement significatif (p -value inférieure à 5%), la vitesse par contre ne l'est pas (p -value supérieure à 5%), la pression a une influence très significative (p -value nulle). L'interaction x_1x_2 (Lieu*Vitesse) est significative (p -value inférieure à 5%). Par contre, les interactions x_2x_3 (Vitesse*Pression) et x_1x_3 (Lieu*Pression) ne sont pas significatives (p -value supérieure à 5%). La triple interaction (Lieu*Vitesse*Pression) est elle significative!

4. CONTRÔLE DE RÉCEPTION

Le "**contrôle de réception**" ou "**contrôle statistique des lots**" est un domaine extrêmement important, vaste et mathématiquement technique du contrôle statistique de la qualité dans l'industrie (alimentaire, pharma, biens de grande consommation) et des sociétés de services (prestations à répétition, contrôles des processus, sondages). Son but, selon la norme NF X06-021, est de permettre l'application à un lot contrôle (de produits ou prestations) de l'une ou l'autre des décisions suivantes: Acceptation ou Rejet à différentes étapes d'une production/développement pour vérifier si un lot est recevable pour les étapes suivantes ou pour vérifier le produit fini avant envoi au client.

Remarque: Nous allons ici présenter que les bases car les cas particuliers nécessitent un bagage mathématique considérable! C'est par ailleurs pour cette raison que de nombreuses entreprises (PME ou multinationales) qui utilisent des étudiants stagiaires ou employés non mathématiciens pour définir les contrôles qualité afin d'économiser de l'argent se retrouvent par la suite avec des problèmes considérables au niveau du respect des lois et des normes! Cependant, ce que nous allons présenter ici constitue le minimum-minimum de la connaissance de tout responsable ou ingénieur qualité ou responsable du contrôle de la réception fournisseurs actif dans une organisation quelconque (industrielle ou administrative) sous peine d'avoir aucune crédibilité! Par ailleurs méfiez vous des entreprises - particulièrement des multinationales - qui recherchent des spécialistes qualité maîtrisant Microsoft Excel ou Microsoft Access. Car cela signifiera qu'elles utilisent des outils non professionnels pour faire un travail qui lui devrait pourtant l'être avec des outils adaptés (et Microsoft Excel ou Microsoft Access ne le sont pas)!!! Donc en termes d'organisation interne, vous pouvez vous assurer que ces entreprises organisent et analysent n'importe quoi, n'importe comment, avec un outil non adapté et donc que c'est le bordel général en interne.

Le but du contrôle de réception n'est pas d'estimer la qualité d'un lot mais uniquement de l'accepter ou de le rejeter. Le contrôle de réception n'est pas non plus un substitut aux méthodes de contrôle de procédé comme les cartes de contrôle ou l'analyse de capacité. Une utilisation dynamique de ces derniers outils durant la fabrication aura pour conséquence de réduire et, dans certains cas, d'éliminer les besoins d'effectuer du contrôle de réception

Il existe principalement trois approches pour évaluer un lot:

- Soit accepter le lot sans inspection (utile dans des situations où une relation de confiance absolue règne entre le client et le fournisseur).
- Soit prélever un échantillon (échantillonnage) et en tirer des conclusions sur le lot complet (utilité dans la majorité des cas car peu coûteux et motive fortement le fournisseur à améliorer sa qualité mais il y a un risque à rejeter de bons lots ou d'en accepter des mauvais). Par exemple PSA Mulhouse contrôle 5% des véhicules qui sortent des ces usines chaque jour.
- Soit inspecter la totalité des individus du lot (utile quand le fournisseur ne peut pas prouver qu'il maîtrise son procédé et est capable de produire dans les spécifications ou quand la caractéristique suivie est critique).

Pour que le résultat d'un contrôle d'un lot par échantillonnage soit fiable, il est trivialement primordial que, lors du prélèvement d'un échantillon d'individus, aucune préférence ne soit accordée à certains individus du lot. Idéalement, chaque individu du lot doit avoir la même probabilité d'être pris dans l'échantillon.

La ou les caractéristiques contrôlées peuvent être:

- Qualitatives (attributs): l'aspect d'un produit, la présence ou l'absence d'un caractère non-conforme, le non respect d'un contrôle de calibre, un résultat atteint ou non. Les individus sont alors directement classés en conformes ou non-conformes repérés par le nombre de défauts qu'ils comportent.
- Quantitatives (variables ou mesures): caractéristique mesurée sur une échelle continue: volume, longueur, poids,...

Signalons enfin qu'il y a deux façons de classer les plans de réception. La première famille consiste à définir les trois catégories suivantes:

- le "**contrôle de la proportion d'individus non-conformes par comptage**" ou "**contrôle par attribut**": Une ou plusieurs caractéristiques de type qualitatif ou quantitatif sont contrôlées sur chaque individu dans le but de le classer en conforme ou non-conforme suivant certains critères. La décision est prise d'après le nombre d'individus non-conformes trouvés dans le (ou les) échantillons prélevés.

- le "**contrôle de la proportion d'individus non-conformes par mesurage**" ou "**contrôle par variable**". Une caractéristique est mesurée sur chaque individu d'un échantillon et la décision est prise en fonction de la moyenne et de la dispersion de la caractéristique calculée sur l'ensemble des individus prélevés.

Nous pouvons également classer les plans de réception selon le nombre d'échantillons prélevés.

- Un "**plan d'échantillonnage simple**" consiste à prendre n individus dans le lot et de prendre une décision sur la base de ceux-ci.

- Un "**plan d'échantillonnage double**" consiste à prélever un premier échantillon dans le lot et d'en tirer une des trois conclusions possibles: (1) accepter le lot, (2) rejeter le lot, (3) rééchantillonner. Si un second échantillon est prélevé, les informations des deux échantillons sont rassemblées pour décider d'accepter ou de rejeter le lot (si nous répétons cette stratégie, nous parlons alors de "**plan d'échantillonnage multiple**").

- Un "**plan d'échantillonnage progressif ou séquentiel**" est le cas extrême du plan d'échantillonnage multiple. Les unités sont prélevées une à une et après chaque prélèvement une décision est prise: (1) accepter, (2) rejeter, (3) prélever une nouvelle unité.

Exemple:

Un producteur de boissons gazeuses désire contrôler une propriété des bouteilles qui lui sont fournies. Pour cela, il utilise un plan de réception simple avec contrôle de la proportion d'individus non-conformes par mesurage quand un lot de bouteilles lui est livré:

1. Prélever un certain nombre n d'unités (bouteilles) du lot (échantillonnage simple)
2. Mesurer à quelle pression chacune des n bouteilles éclate (contrôle destructif)
3. Calculer des indicateurs statistiques des valeurs observées.
4. Suivre la règle de décision d'acceptation ou de rejet que nous devons déterminer.

Remarque: Comme les cartes de contrôle, l'application d'un plan d'échantillonnage est un test d'hypothèse. Il peut mener à des décisions correctes ou fausses. Les risques encourus sont ceux d'accepter un lot qui n'est pas conforme ou de rejeter un lot qui est conforme.

Évidemment l'échantillonnage comporte des avantages et désavantages. Voyons la liste des avantages les plus communs:

A1. Cela coûte moins cher car on ne contrôle pas tout...

A2. Il y a moins de manipulations des produits donc moins de déchets

A3. Il y a besoin de mobiliser moins de personnes ou de machines de contrôle

A4. Cela réduit considérablement les erreurs d'inspection

A5. Le retour d'un lot complet juste à cause d'un échantillon non accepté a tendance à plus motiver le fournisseur à faire de la qualité...

Mais cela comporte aussi des désavantages comme:

D1. Le risque d'accepter de mauvais lots et d'en rejeter des bons...

D2. Requière plus de compétences des employés et des contrats bien écrits...

4.1. PLAN D'ÉCHANTILLONNAGE PAR MESURE SIMPLE POUR UNE UNIQUE TOLÉRANCE AVEC ÉCART-TYPE CONNU

Pour ce type de plan d'échantillonnage, nous supposons que l'échantillon n est beaucoup plus petit que le lot total (d'un facteur 10 au moins) et que l'écart-type est connu!

La caractéristique X mesurée pour chaque individu de l'échantillon est supposée suivre une loi Normale d'espérance et d'écart-type et de variance identique et le calcul de ces deux derniers paramètres est basé sur l'utilisation de la moyenne empirique (estimateur de maximum de vraisemblance de l'espérance) et de l'estimateur sans biais de l'écart-type (estimateur de maximum de vraisemblance de l'écart-type). Cette hypothèse implique que le produit du fournisseur à une fabrication sous contrôle statistique.

Exemple:

Dans le cadre de notre exemple, le producteur de boissons gazeuses, il reçoit des lots de 10'000 bouteilles (donc l'échantillon satisfera la condition susmentionnée).

Nous noterons:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \text{ et } \bar{X} = N\left(\mu, \frac{\sigma^2}{\sqrt{n}}\right) \quad (74.354)$$

Nous conviendrons que le client et le fournisseur se sont entendus sur des limites de spécification (limites de tolérance). Nous noterons comme à l'habitude ces limites USL et LSL .

Soit D la proportion (inconnue) de produits non-conformes (défectueux), elle est donnée naturellement dans le domaine de la maîtrise statistique des procédés par:

$$D = P(X < LSL) + P(X > USL) = P\left(Z < \frac{LSL - \mu}{\sigma}\right) + P\left(Z > \frac{USL - \mu}{\sigma}\right) \quad (74.355)$$

Dans la pratique, afin d'économiser de l'argent dans le cadre des tests de qualité nous essayons de réduire (reformuler) tout problème de contrôle à un unique calibre. Au fait le problème peut toujours avec ou sans reformulation se réduire à avoir une proportion de non-conforme en dessous de LSL ou au-dessus de USL pour rejeter le lot si la fabrication du fournisseur est bien sous contrôle statistique. D'où le fait que dans la pratique les développements mathématiques ne sont faits que par rapport à une seule borne que nous noterons alors LSL pour "**limite de spécification du lot**".

Ainsi, nous réduisons les problèmes à (ou nous nous arrangeons à avoir les problèmes qui se réduisent à cette formulation):

$$D = P(X < LSL) \text{ ou } D = P(X > LSL) \quad (74.356)$$

Exemple:

Dans le cadre de notre exemple, le producteur de boissons gazeuses, il s'agit de la résistance à la pression d'un lot de 10'000 bouteilles que nous souhaiterions contrôler. Le producteur s'est mis d'accord avec son fournisseur sur une limite de spécification inférieure pour cette caractéristique, LSL valant 22.5, limite qu'une majorité des bouteilles doit dépasser. Ainsi:

$$D = P(X < LSL) = P(22.5 < LSL) \quad (74.357)$$

Nous supposons que des données passées ont permis d'estimer précisément l'écart-type et que l'on peut donc le considérer comme connu et valant 1.5 et qu'il est sous contrôle statistique.

Car évidemment si la proportion de non-conforme est supérieure (respectivement inférieure) à ce que définit le contrat de qualité dans une borne, elle le sera aussi à la borne opposée dans des proportions identiques puisque le processus de fabrication du fournisseur est par hypothèse sous contrôle statistique.

Avant de partir plus loin dans les développements mathématiques, il y a une chose importante qu'il faut se rappeler:

Si nous nous imaginons être un fournisseur, il est évident que nous souhaiterions avoir un niveau de qualité tel que la probabilité cumulée que le client rejette à tort un lot soit faible (car en tant que fournisseur nous serons contractuellement obligés de reprendre le lot même si un deuxième contrôle ultérieur complet s'avère qu'il a tort!). Pour cela, le fournisseur définit une valeur D_f du pourcentage de produit non-conforme ("**niveau de qualité acceptable**": NQA) en dessous de laquelle il juge que le lot ne peut être refusé que très rarement. De plus, il définit également α comme étant la probabilité cumulée maximale de se voir refuser un lot qui a une proportion de non-conforme plus petite ou égale à D_f (il s'agit donc de la probabilité cumulée avec laquelle un client rejetterait à tort un lot). Ceci se note:

$$P(\text{Refuser un lot} | D \leq D_f) \leq \alpha \quad (74.358)$$

α est appelé bien évidemment... "**risque du fournisseur**" et est en général de l'ordre de 0.1 à 10%.

Enfin si nous nous imaginons être le client, il est évident que nous souhaiterions que la probabilité cumulée d'accepter à tort un lot qui ne respecte pas le contrat qualité soit faible pour des raisons évidentes de coûts. Pour cela, le client de son côté définit une valeur D_c du pourcentage de produit non-conforme ("**niveau de qualité limite**": NQL) au-delà de laquelle il juge que le lot ne peut être accepté que très rarement. Il définit également β comme étant la probabilité cumulée maximale de devoir accepter un lot qui a une proportion de non-conforme plus grande ou égale à D_c (il s'agit donc de la probabilité cumulée avec laquelle on accepterait à tort un lot du fournisseur). Ceci se note:

$$P(\text{Accepter un lot} | D \geq D_c) \leq \beta \quad (74.359)$$

β est appelé bien évidemment... "**risque du client**" et est en général aussi de l'ordre de 5 à 10%.

Pour résumer la situation peut-être illustrée par le tableau suivant (qui est en tout poit similaire au tableau des erreurs de Type I et II vu dans le chapitre de Statistiques):

	Le lot est bon (NQA)	Le lot est mauvais (NQL)
Refus	Risque α	Décision correcte
Acceptation	Décision correcte	Risque β

Exemple:

Dans le cadre de notre exemple - le producteur de boissons gazeuses - qui reçoit des lots de 10'000 bouteilles dont la résistance à la pression (LSL) doit être supérieure à 22.5 avec un écart-type sous contrôle statistique et valant 1.5, le fournisseur et le client ont chacun fixé les risques qu'ils sont prêts à courir:

- Le fournisseur désire qu'un lot avec moins de 1% de bouteilles non-conformes soit refusé par erreur par le client dans au plus 5% des cas. Soit:

$$P(\text{Refuser un lot} | D \leq D_f) \leq \alpha \Rightarrow P(\text{Refuser un lot} | D \leq 1\%) \leq 5\% \quad (74.360)$$

- Le client désire qu'un lot avec plus de 5% de bouteilles non-conformes soit accepté à tort dans moins de 10% des cas. Soit:

$$P(\text{Accepter un lot} | D \geq D_o) \leq \beta \Rightarrow P(\text{Accepter un lot} | D \geq 5\%) \leq 10\% \quad (74.361)$$

Remarque: La règle de décision d'acceptation ou de rejet du lot pour un contrôle par mesure est en générale basée sur la moyenne arithmétique de la caractéristique estimée sur base de l'échantillon plutôt que sur base d'une estimation de taux de non-conforme.

Pour la suite, quand nous écrivons:

$$D = P(X < LSL) \quad (74.362)$$

la particularité c'est que dans la pratique D est imposé σ est donné. Le reste est soit à calculer, soit à éliminer (effectivement le praticien doit normalement uniquement parler en termes de probabilité cumulées et proportions et facteurs pour définir sa politique de qualité et LSL n'est ni une probabilité cumulée, ni une proportion, ni un facteur indépendant de la chose que l'on analyse!). Dès lors nous écrivons:

$$D = P(X < LSL) = P\left(Z < \frac{LSL - \mu}{\sigma}\right) \quad (74.363)$$

avec donc la variable centrée réduite habituelle:

$$Z = \frac{X - \mu}{\sigma} \quad (74.364)$$

Évidemment nous en tirons que la valeur limite est:

$$\mu = LSL - Z_D \sigma \quad (74.365)$$

où Z_D est donc le percentile de niveau D de la distribution centrale normale réduite (comme à l'habitude). Malheureusement, nous avons toujours LSL qui est là. Pouvons-nous nous en soustraire? Pour cela, nous avons dans l'hypothèse que la mesure est sous contrôle statistique:

$$\bar{X} = N\left(\mu, \frac{\sigma^2}{n}\right) \quad (74.366)$$

La probabilité cumulée P_D d'accepter (nous aurions pu le faire pour le rejet mais il faut choisir un des deux pour l'exemple...) le lot quand le taux de non-conforme est D s'en déduit dans le cadre de l'utilisation de la moyenne de la mesure:

$$P_D = P(\bar{X} \geq LSL + k\sigma) \quad (74.367)$$

dans le cas particulier où la mesure ne doit pas être inférieure à une certaine valeur donnée (comme c'est le cas pour l'exemple des bouteilles susmentionné!).

Nous avons alors:

$$\begin{aligned} P_D &= P(\bar{X} \geq LSL + k\sigma) = P\left(\frac{\bar{X} - \mu_D}{\sigma/\sqrt{n}} \geq \frac{LSL + k\sigma - \mu_D}{\sigma/\sqrt{n}}\right) = P\left(Z \geq \frac{LSL + k\sigma - \mu_D}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z \geq \frac{LSL + k\sigma - (LSL - Z_D\sigma)}{\sigma/\sqrt{n}}\right) = P\left(Z \geq \frac{k\sigma + Z_D\sigma}{\sigma/\sqrt{n}}\right) = P(Z \geq \sqrt{n}(k + Z_D)) \end{aligned} \quad (74.368)$$

Nous avons alors un résultat très intéressant! La probabilité cumulée d'accepter le lot quand le taux de non-conforme est D est alors dans le cas particulier où il s'agit d'une mesure en dessous de laquelle il ne faut pas aller:

$$P_D = P(Z \geq \sqrt{n}(k + Z_D)) \quad (74.369)$$

et de le rejeter dans le cas d'une mesure au-delà de laquelle il ne faut pas aller:

$$P_D = P(Z \leq \sqrt{n}(k + Z_D)) \quad (74.370)$$

et ce qui est bien c'est que le problème se réduit à un percentile Z_D aisé à choisir (norme ou choix politique), d'un nombre d'individus n , et d'un facteur k donc que des éléments indépendants de la mesure elle-même et alors utilisable dans n'importe quel métier pour n'importe quel type d'objet (et en plus nous avons éliminé l'écart-type)!!

Évidemment, le fournisseur et le client doivent se mettre d'accord pour un plan d'échantillonnage afin d'établir des tests qualité correspondant aux attentes de l'un et de l'autre. Ce que nous savons jusqu'à maintenant c'est qu'ils s'imposent respectivement un α , β ce qui leur permet de calculer respectivement immédiatement les $Z_{D,\alpha}$, $Z_{D,\beta}$. Il faut cependant calculer le nombre d'individus n et le facteur k qui satisfont les exigences respectives.

Pour cela, rappelons que si α est donc le risque du fournisseur (probabilité cumulée que le client rejette à tort le lot) et β le risque du client (probabilité cumulée d'accepter à tort un lot du fournisseur) et:

$$P_D = P\left(Z \geq \sqrt{n}(k + Z_D)\right) \quad (74.371)$$

la probabilité cumulée d'accepter un lot avec limite inférieure, alors nous devons résoudre le système suivant:

$$\begin{aligned} 1 - \alpha &= P\left(Z \geq \sqrt{n}(k + Z_{D,\alpha})\right) \\ \beta &= P\left(Z \geq \sqrt{n}(k + Z_{D,\beta})\right) \end{aligned} \quad (74.372)$$

ou:

$$\begin{aligned} \alpha &= P\left(Z \leq \sqrt{n}(k + Z_{D,\alpha})\right) \\ 1 - \beta &= P\left(Z \leq \sqrt{n}(k + Z_{D,\beta})\right) \end{aligned} \quad (74.373)$$

pour que le fournisseur et le client aient le plan d'échantillonnage qui correspondent aux exigences de chacun!

Nous allons choisir le premier (car il faut bien en choisir un...). Celui-ci peut d'ailleurs se récrire:

$$\begin{aligned} 1 - \alpha &= P\left(Z \geq \sqrt{n}(k + Z_{D,\alpha})\right) = 1 - P\left(Z \leq \sqrt{n}(k + Z_{D,\alpha})\right) \\ \beta &= P\left(Z \geq \sqrt{n}(k + Z_{D,\beta})\right) = 1 - P\left(Z \leq \sqrt{n}(k + Z_{D,\beta})\right) \end{aligned} \quad (74.374)$$

Il vient alors:

$$\begin{aligned} \alpha &= P\left(Z \leq \sqrt{n}(k + Z_{D,\alpha})\right) \\ 1 - \beta &= P\left(Z \leq \sqrt{n}(k + Z_{D,\beta})\right) \end{aligned} \quad (74.375)$$

Ce qui peut s'écrire:

$$\begin{aligned} Z_\alpha &= \sqrt{n}(k + Z_{D,\alpha}) \\ Z_{1-\beta} &= \sqrt{n}(k + Z_{D,\beta}) \end{aligned} \quad (74.376)$$

Soit:

$$\frac{Z_{\alpha}}{\sqrt{n}} - Z_{D,\alpha} = \frac{Z_{1-\beta}}{\sqrt{n}} - Z_{D,\beta} = k \quad (74.377)$$

d'où nous tirons finalement n :

$$\begin{aligned} \frac{Z_{\alpha}}{\sqrt{n}} - \frac{Z_{1-\beta}}{\sqrt{n}} &= Z_{D,\alpha} - Z_{D,\beta} \Leftrightarrow \frac{1}{\sqrt{n}} = \frac{Z_{D,\alpha} - Z_{D,\beta}}{Z_{\alpha} - Z_{1-\beta}} \\ \Leftrightarrow \sqrt{n} &= \frac{Z_{\alpha} - Z_{1-\beta}}{Z_{D,\alpha} - Z_{D,\beta}} \end{aligned} \quad (74.378)$$

d'où au final:

$$n = \left(\frac{Z_{\alpha} - Z_{1-\beta}}{Z_{D,\alpha} - Z_{D,\beta}} \right)^2 \quad (74.379)$$

et k :

$$\begin{aligned} k &= \frac{Z_{\alpha}}{\sqrt{n}} - Z_{D,\alpha} = \frac{Z_{\alpha}}{\frac{Z_{\alpha} - Z_{1-\beta}}{Z_{D,\alpha} - Z_{D,\beta}}} - Z_{D,\alpha} = \frac{Z_{\alpha} (Z_{D,\alpha} - Z_{D,\beta})}{Z_{\alpha} - Z_{1-\beta}} - Z_{D,\alpha} \\ &= \frac{Z_{\alpha} (Z_{D,\beta} - Z_{D,\alpha})}{Z_{1-\beta} - Z_{\alpha}} - Z_{D,\alpha} = \frac{Z_{\alpha} (Z_{D,\beta} - Z_{D,\alpha}) - Z_{D,\alpha} (Z_{1-\beta} - Z_{\alpha})}{Z_{1-\beta} - Z_{\alpha}} \\ &= \frac{Z_{\alpha} Z_{D,\beta} - Z_{\alpha} Z_{D,\alpha} - Z_{D,\alpha} Z_{1-\beta} + Z_{D,\alpha} Z_{\alpha}}{Z_{1-\beta} - Z_{\alpha}} \end{aligned} \quad (74.380)$$

d'où au final:

$$k = \frac{Z_{\alpha} Z_{D,\beta} - Z_{D,\alpha} Z_{1-\beta}}{Z_{1-\beta} - Z_{\alpha}} \quad (74.381)$$

Exemples:

E1. Dans le cadre de notre exemple - le producteur de boissons gazeuses - qui reçoit des lots de 10'000 bouteilles dont la résistance à la pression (LSL) doit être supérieure à 22.5 avec un écart-type sous contrôle statistique et valant 1.5, le fournisseur et le client ont chacun fixé les risques qu'ils sont prêts à courir:

$$\begin{aligned} P(\text{Refuser un lot} | D \leq D_f) &\leq \alpha \Rightarrow P(\text{Refuser un lot} | D \leq 1\%) \leq 5\% \\ P(\text{Accepter un lot} | D \geq D_e) &\leq \beta \Rightarrow P(\text{Accepter un lot} | D \geq 5\%) \leq 10\% \end{aligned} \quad (74.382)$$

Nous avons alors en utilisant la version anglaise de Microsoft Excel 11.8346 (j'ai préféré faire ce choix car le nom des formules y est beaucoup... beaucoup plus court qu'en français):

$$\begin{aligned}
 n &= \left(\frac{Z_{\alpha} - Z_{1-\beta}}{Z_{D,\alpha} - Z_{D,\beta}} \right)^2 = \left(\frac{\text{NORMSINV}(5\%) - \text{NORMSINV}(1-10\%)}{\text{NORMSINV}(1\%) - \text{NORMSINV}(5\%)} \right)^2 \\
 &= \left(\frac{-1.645 - 1.282}{-2.326 - (-1.645)} \right)^2 = \left(\frac{-2.926}{-0.618} \right)^2 = 4.294^2 = 18.439 \cong 18
 \end{aligned} \quad (74.383)$$

et (calculé avec les mêmes fonctions Microsoft Excel que nous omettrons donc de préciser):

$$k = \frac{Z_{\alpha} Z_{D,\beta} - Z_{D,\alpha} Z_{1-\beta}}{Z_{1-\beta} - Z_{\alpha}} \cong 1.943 \quad (74.384)$$

Le plan d'échantillonnage consiste alors à prélever et analyser un échantillon de taille 18, calculer la moyenne $\bar{X}$ et appliquer la règle de décision suivante:

$$\begin{aligned}
 \text{Acceptation du lot si: } \bar{X} &\geq LSL + k\sigma = 22.5 + 1.943 \cdot 1.5 \cong 25.4 \\
 \text{Rejet du lot si: } \bar{X} &< LSL + k\sigma \cong 25.4
 \end{aligned} \quad (74.385)$$

Supposons qu'un échantillon de taille 18 soit prélevé et donne les résultats suivants:

25.6, 26.0, 23.0, 27.0, 27.5, 29.0, 28.5, 26.0, 25.6

25.8, 26.5, 28.8, 27.3, 25.2, 27.1, 29.8, 26.5, 27.8

Nous avons alors:

$$\bar{X} = 26.8 \geq 25.4 \quad (74.386)$$

et donc le lot est accepté. Il faut savoir que beaucoup de praticiens préfèrent le calcul suivant utilisant un "indice de qualité", noté Q , utilisant simplement la relation suivante:

$$Q = \frac{\bar{X} - LSL}{\sigma} = \frac{26.8 - 22.5}{1.5} \cong 2.86 > k \quad (74.387)$$

donc le lot est accepté.

Remarque: Il est quand même conseillé dans la pratique de calculer l'estimateur non biaisé de l'écart-type de l'échantillon pour vérifier qu'il ne s'éloigne pas trop de l'écart-type du contrat (même s'il ne peut pas être identique pour de petits échantillons)!

E2. Un deuxième calcul-type fait par les praticiens (souvent par pure curiosité ou parfois pour faire un rapport de la stratégie de la qualité à la direction) qui découle du premier exemple est le suivant: Quelle est la probabilité cumulée que l'on rejette avec n et k donnés, un lot dont la proportion de non-conformes D serait de 5%? Dès lors, nous utilisons la relation démontrée suivante:

$$P_D = P\left(Z \leq \sqrt{n}(k + Z_D)\right) \quad (74.388)$$

Soit avec les valeurs précédentes:

$$P_D = P\left(Z \leq \sqrt{18}(1.943 + \text{NORMSINV}(5\%))\right) \cong 89.70\% \quad (74.389)$$

et la probabilité cumulée qu'on l'accepte est alors de 1-89.70% soit 10.29%.

4.1.1. CALCUL DES PARAMÈTRES PAR UTILISATION DE LA NORME AF-X06-023

La norme AFNOR X06-023 propose une méthode quelque peu différente pour déterminer k et n , non directement basée sur les risques du client et du fournisseur mais sur un paramètre appelé "**niveau de qualité acceptable**" (NQA.). En outre, cette norme tient également compte de la taille du lot à tester (considérée plus haut comme infinie) et du niveau de contrôle désiré.

Remarque: La norme ISO 3951:1981 correspondante donne à ma connaissance des valeurs très différentes de celle de l'AFNOR pour des raisons qui m'échappent (j'espère que c'est juste parce que je n'ai pas compris quelque chose...).

Définition: Le NQA., est le pourcentage d'individus non-conformes qui ne doit pas être dépassé pour qu'une production, contrôlée sur une série de lots, puisse être considérée comme satisfaisante.

En d'autres termes, quand une production est stable avec un pourcentage d'individus non-conformes au plus égal au NQA., la grande majorité des lots présentés au contrôle doivent être acceptés par le client. Donc en général le NQA. est proche - mais pas égal! - de D_f (pour rappel: la proportion de non-conforme tel qu'il y ait une probabilité cumulée α que le client rejette à tort le lot).

La méthode proposée est du type "recette de cuisine" mais est imposée par certaines normes européennes donc les industriels n'ont pas d'autre choix que de l'appliquer. Mais, finalement, les valeurs numériques sont relativement proches de celles obtenues avec les relations démontrées précédemment. Au fait leur vraie utilité, c'est qu'elles sont simples à mettre en oeuvre dans des situations beaucoup plus complexes que celles que nous venons d'étudier.

En fonction du contexte, la norme propose d'appliquer différents niveaux de contrôle:

- Le contrôle normal qui est à adopter quand nous sommes en relation de confiance étroite avec le fournisseur et avons un oeil en temps réel sur la qualité de sa production à l'aide d'outils informatiques de surveillance à distance.
- Le contrôle renforcé qui est plus strict que le contrôle normal (n est plus grand) et a pour but de mieux protéger le client contre le risque de devoir accepter des lots à tort. Ce type de contrôle doit être effectué temporairement quand il y a de sérieuses raisons de penser que la qualité de la production n'est pas (ou plus) ce qu'elle doit être jusqu'à un retour à la normale (suspension de livraisons entre temps).
- Enfin, le contrôle réduit qui est le plus économique et qui peut être appliqué quand sur la base de lots précédemment contrôlés et acceptés, nous pouvons croire que le fournisseur maîtrise particulièrement la qualité de son procédé.

Exemples:

E1. Dans le cadre de notre exemple - le producteur de boissons gazeuses - qui reçoit des lots de 10'000 bouteilles dont la résistance à la pression (LSL) doit être supérieure à 22.5 avec un écart-type sous contrôle statistique et valant 1.5, le fournisseur et le client ont chacun fixé les risques qu'ils sont prêts à courir:

$$\begin{aligned} P(\text{Refuser un lot} | D \leq D_f) &\leq \alpha \Rightarrow P(\text{Refuser un lot} | D \leq 1\%) \leq 5\% \\ P(\text{Accepter un lot} | D \geq D_c) &\leq \beta \Rightarrow P(\text{Accepter un lot} | D \geq 5\%) \leq 10\% \end{aligned} \quad (74.390)$$

Nous devons alors en utilisant cette norme définir le NQA. Comme mentionné dans la définition, en général le NQA. est proche de D_f (mais pas égal!). Donc le NQA. sera posé comme valant 1%.

La norme exige le choix d'un niveau de contrôle. Nous prendrons celui de type II. Une table disponible dans la norme nous dit que nous devons utiliser le code L pour la prochaine étape. Enfin, la dernière table nous donne donc pour un NQA. de 1% et un code L la cellule suivante:

1.06%
$n=25$ $k=1.97$
4.28%

ce qui correspond à:

D_f
n k
D_c

La norme indique donc qu'un échantillon de taille 25 doit être utilisé et un k de 1.97! Ce plan assure des risques un peu moins grands que le plan calculé plus haut mais est plus coûteux, car nécessite plus d'échantillons.

E2. Les praticiens lors de l'utilisation de cette norme souhaitent souvent pouvoir calculer le risque fournisseur et client associé au NQA choisi et paramètres correspondants donnés par la norme. Alors en utilisant:

$$k = \frac{Z_\alpha}{\sqrt{n}} - Z_{D,\alpha} \quad (74.391)$$

Il vient:

$$Z_\alpha = \sqrt{n} (k + Z_{D,\alpha}) \quad (74.392)$$

Soit dans le cadre de notre exemple et en utilisant toujours la version anglaise de Microsoft Excel 11.8346:

$$Z_\alpha = \sqrt{25} (1.97 + \text{NORMSINV}(1.06\%)) \cong -1.67 \quad (74.393)$$

Soit:

$$\alpha = \text{NORMSINV}(-1.67) = 4.73\% \quad (74.394)$$

à comparer aux 5%. De même comme:

$$k = \frac{Z_{1-\beta}}{\sqrt{n}} - Z_{D,\beta} \quad (74.395)$$

alors:

$$Z_{1-\beta} = \sqrt{n} (k + Z_{D,\beta}) \quad (74.396)$$

Soit dans le cadre de notre exemple et en utilisant toujours la version anglaise de Microsoft Excel 11.8346:

$$Z_{1-\beta} = \sqrt{25} (1.97 + \text{NORMSINV}(4.28\%)) \cong 1.25 \quad (74.397)$$

il vient:

$$1 - \beta = \text{NORMSINV}(1.25) = 89.52\% \quad (74.398)$$

et au final:

$$\beta = 1 - 89.52\% = 10.48\% \quad (74.399)$$

à comparer aux 10%.

4.2. PLAN D'ÉCHANTILLONNAGE SIMPLE PAR ATTRIBUTS

Les plans par attributs sont plus simples que les plans par échantillonnage, mais ils sont souvent plus coûteux (en nombre de prélèvements) que les plans pour mesures et ne sont donc à conseiller que quand nous ne pouvons pas résumer l'information sur la qualité d'une unité par des mesures ou que les coûts associés aux analyses sont trop élevés par rapport à une simple classification des unités en conforme ou non-conforme.

Le "**plan d'échantillonnage simple par attributs**" est une procédure de contrôle de réception consistant à prélever au hasard un échantillon de taille n donnée dans le lot. Les unités prélevées (de manière homogène!) sont ensuite inspectées et classées comme conformes ou non-conformes. Soit X le nombre de composants non-conformes trouvés dans l'échantillon.

Si $X \leq A$, la qualité du lot est jugée bonne et le lot est accepté (A). Si $X \geq R$ avec dans les plans simples $R=A+1$ la qualité du lot est jugée mauvaise et le lot n'est pas accepté.

Exemple:

Des bouteilles par lot de taille $N=10'000$ sont reçues par un client. La procédure de contrôle utilise le plan empirique par attributs suivants: on prélève un échantillon de taille $n=200$. Le critère d'acceptation A est $X \leq 1$ et le critère de rejet R est $X \geq 2$.

Mentionnons avant de partir dans les détails des calculs de la stratégie d'un plan d'échantillonnage simple par attributs les principes des plans d'échantillonnage double et multiple pour la culture générale:

Le "**plan d'échantillonnage double par attributs**" est une procédure de contrôle de réception ayant deux stades de décision. Elle consiste à prélever d'abord un premier échantillon de taille inférieure à celui que l'on prélèverait pour un plan d'échantillonnage simple. Le premier stade de décision est basé sur les informations fournies par ce premier échantillon. Si la qualité du premier échantillon est jugée suffisamment bonne, le lot est accepté. Si elle est jugée suffisamment mauvaise, le lot n'est pas accepté. En revanche si la qualité du premier échantillon est jugée intermédiaire, un second échantillon est prélevé et inspecté dans le but d'accepter ou de refuser le lot. Les critères de décision du second stade sont basés sur les informations recueillies sur les deux échantillons.

Exemple:

Des bouteilles par lot de taille $N=10'000$ sont reçues par un client. La procédure de contrôle utilise le plan empirique par attributs suivants: on prélève un premier échantillon de taille $n_1 = 125$. Le critère d'acceptation A au premier stade est $X_1 = 0$ et le critère de rejet R est $X_1 \geq 2$. Si la qualité est non satisfaisante, nous prenons un deuxième échantillon de taille $n_2 = 125$ avec un critère d'acceptation A au deuxième stade de $X_2 \leq 1$ et le critère de rejet R est $X_2 \geq 2$.

Nous reviendrons sur un exemple beaucoup plus détaillé de double échantillonnage un peu plus loin!!

Le "**plan d'échantillonnage multiple par attributs**" est simplement une généralisation du plan d'échantillonnage double par attributs en ayant D étapes de décision au lieu de 1 pour le plan simple ou 2 pour le plan double...

Revenons à nos développements mathématiques d'un plan d'échantillonnage simple par attributs. Si n est la taille du lot, p la taille de l'échantillon et m le nombre total de défectueux dans le lot alors la probabilité d'avoir k défectueux parmi p est (toujours avec la notation du coefficient binomial non-conforme à la norme ISO 31-11):

$$P(n, p, m, k) = \frac{C_k^m C_{p-k}^{n-m}}{C_p^n} \quad (74.400)$$

Afin d'être conforme à la tradition du domaine de l'échantillonnage, faisons un premier changement de notation. Notons n la taille de l'échantillon et N la taille du lot. Il vient alors:

$$P(N, n, m, k) = \frac{C_k^m C_{n-k}^{N-m}}{C_n^N} \quad (74.401)$$

Faisons un deuxième changement de notation. Notons p la proportion de non-conformes dans le lot. Il vient alors la notation traditionnelle d'usage:

$$P(N, n, p, k) = \frac{C_k^{Np} C_{n-k}^{N-Np}}{C_n^N} \quad (74.402)$$

Ainsi, la probabilité cumulée d'acceptation d'un lot de qualité p est donnée par:

$$P_a = P(X \leq A) = \sum_{k=0}^A P(N, n, p, X = k) = \sum_{k=0}^A \frac{C_k^{Np} C_{n-k}^{N-Np}}{C_n^N} \quad (74.403)$$

 Exemple:

Quelle est la probabilité cumulée de tirer 5 éléments non-conformes (critère d'acceptation) d'un lot de 10'000 bouteilles (N) dont nous avons pris un prélèvement de 200 individus (n), dont la proportion de non-conformes est connue comme étant de 5% (p). Alors avec Microsoft Excel 14.0.6123 nous obtenons comme probabilité cumulée d'acceptation:

$$P_a = P(X \leq A) = \sum_{k=0}^5 \frac{C_k^{10'000 \cdot 5\%} C_{200-k}^{10'000-10'000 \cdot 5\%}}{C_{200}^{10'000}} \quad (74.404)$$

$$= \text{HYPGEOM.DIST}(5; 200; 10000 \cdot 5\%; 10000; 1) = 6.05\%$$

Remarque: Une erreur (bêtise) classique dans les entreprises consiste à penser que le critère d'acceptation est proportionnel à la taille du lot. En d'autres termes que si pour 10'000 bouteilles, le fait de tester un échantillon de 200 pièces (2%) donne une probabilité cumulée d'acceptation de ~6%, alors sur 1'000 bouteilles, un échantillon de 20 pièces (2%) donne la même probabilité cumulée d'acceptation. Au fait, avec Microsoft Excel 14.0.6123, nous avons alors non plus 6.05% mais alors 99.7%... la différence est donc considérable!!

Si nous rappelons (voir la partie concernant les plans d'échantillonnage par mesure) que α est donc le risque du fournisseur (probabilité cumulée que le client rejette à tort le lot de qualité dont la proportion de non-conformes est p_α) et β le risque du client (probabilité cumulée d'accepter à tort un lot du fournisseur de qualité dont la proportion de non-conformes est p_β) alors pour que le fournisseur et le client aient le plan d'échantillonnage qui correspondent aux exigences de chacun nous devons résoudre le système suivant:

$$1 - \alpha = \sum_{k=0}^A \frac{C_k^{Np_\alpha} C_{n-k}^{N-Np_\alpha}}{C_n^N} \quad (74.405)$$

$$\beta = \sum_{k=0}^A \frac{C_k^{Np_\beta} C_{n-k}^{N-Np_\beta}}{C_n^N}$$

À ma connaissance, il n'est pas possible de résoudre ce système analytiquement. Les techniques de recherche opérationnelle utilisant la méthode du simplexe, la méthode des gradients ou de Newton échouent lamentablement à trouver une solution à ce système (ce qui est normal). Par contre, avec les algorithmes évolutionnaires (outil disponible dans Microsoft Excel 14.0.6123) on peut trouver des solutions acceptables, mais les paramétrages ne sont pas aisés.

Dès lors, en supposant que la taille des lots soumise à l'inspection est grande par rapport à la taille de l'échantillon, et en exploitant l'approximation binomiale de la loi hypergéométrique (approximation démontrée dans le chapitre de Statistiques), nous avons:

$$P_a = P(X \leq A) = \sum_{k=0}^A P(n, p, X = k) = \sum_{k=0}^A C_k^n p^k (1-p)^{n-k} = \sum_{k=0}^A C_k^n p^k q^{n-k} \quad (74.406)$$

L'avantage de cette approximation est immense! La taille totale N du lot n'intervient plus dans le problème!

 Exemple:

E1. Quelle est la probabilité cumulée de tirer 5 éléments non-conformes (critère d'acceptation) d'un lot de 10'000 bouteilles (N) dont nous avons pris un prélèvement de 200 individus (n), dont la proportion de non-conformes est connue comme étant de 5% (p). Alors avec la version anglaise de Microsoft Excel 14.0.6123 nous obtenons:

$$P_a = P(X \leq A) = \sum_{k=0}^A C_k^n p^k q^{n-k} = \sum_{k=0}^5 C_k^{200} 0.05^k 0.95^{200-k} \quad (74.407)$$

$$= \text{BINOMDIST}(5; 200; 5\%; 1) = 6.23\%$$

E2. Cette dernière relation est aussi utilisée en ingénierie de la fiabilité. Effectivement, supposons que nous souhaitons savoir quel est le nombre d'éléments n que nous devons prendre pour faire un plan de démonstration de fiabilité montrant avec 90% de probabilité cumulée que le nombre de A défaillances (non-conformes) est nul sachant que leur fiabilité est de 80%. Alors cela revient à trouver n tel que nous soyons au plus proche de:

$$P_a = P(X \leq 0) = \sum_{k=0}^A C_k^n p^k q^{n-k} = \sum_{k=0}^A C_k^n (1-R)^k R^{n-k} = \sum_{k=0}^0 C_k^n 0.2^k 0.8^{n-k} \quad (74.408)$$

$$= R^n = 0.8^n = 1 - 90\% = 10\%$$

Donc la valeur la plus proche n pour cela est de prendre 11 éléments (c'est la valeur que retourne le logiciel de référence de la fiabilité qu'est Weibull++). On se retrouve alors avec:

$$= \text{BINOMDIST}(0; 11; 20\%; 1) = 8.59\%$$

Dans le cadre du domaine de la fiabilité est il d'usage d'appeler cette approche "**plan de démonstration de test de la fiabilité binomial non-paramétrique**"....

E3. Voyons encore un dernier exemple relatif à la fiabilité. Rappelons que nous avons démontré plus haut que pour la loi de Weibull à un paramètre, nous avons:

$$R(t) = \exp \left(- \left(\frac{t}{\eta} \right)^\beta \right) \quad (74.409)$$

Et supposons que des tests précédents nous ont montré que à un temps t égal à 2'000 jours, nous avons une fiabilité de 80% et que nous savons que le paramètre de forme comme β vaut 2. Nous souhaiterions savoir quelle taille d'échantillons nous devons prendre pour qu'un plan de test de la fiabilité de 1'500 jours montre avec au plus 90% de probabilité cumulée au maximum une non-conformité (panne). Pour cela, nous allons d'abord déterminer le paramètre d'échelle:

$$R(t) = \exp \left(- \left(\frac{t}{\eta} \right)^\beta \right) \Rightarrow \eta = \frac{t}{(-\ln(R(t)))^{1/\beta}} = \frac{2'000}{(-\ln(R(2'000)))^{1/2}} \quad (74.410)$$

$$= \frac{2'000}{(-\ln(R(0.8)))^{1/2}} \cong 4'233.87$$

Il vient alors que pour ce même test à 1'500 jours, nous avons une fiabilité de:

$$R(1'500) = \exp\left(-\left(\frac{1'500}{4'233.87}\right)^2\right) \cong 0.882 \quad (74.411)$$

En injectant dans la loi binomiale, nous avons alors:

$$P_a = P(X \leq 1) = \sum_{k=0}^A C_k^n p^k q^{n-k} = \sum_{k=0}^1 C_k^n (1-0.882)^k 0.882^{n-k} = 1-90\% = 10\% \quad (74.412)$$

Donc la valeur la plus proche n pour cela est de prendre 32 éléments (c'est la valeur que retourne le logiciel de référence de la fiabilité qu'est Weibull++). On se retrouve alors avec:

$$= \text{BINOMDIST}(1;32;1-88.2\%;1)=10.50\%$$

Dans le cadre du domaine de la fiabilité est il d'usage d'appeler cette approche "**plan de démonstration de test de la fiabilité binomial paramétrique**"....

Remarque: Là aussi... une erreur (bêtise) classique dans les entreprises consiste à penser que le critère d'acceptation est proportionnel à la taille du lot. En d'autres termes que si pour 10'000 bouteilles, le fait de tester un échantillon de 200 pièces (2%) donne une probabilité cumulée d'acceptation de ~6%, alors sur 1'000 bouteilles, un échantillon de 20 pièces (2%) donne la même probabilité cumulée d'acceptation. Au fait, avec Microsoft Excel 14.0.6123, nous avons non plus pas 6.23% mais alors 99.7%... la différence est donc considérable!!

Pour en revenir à la théorie.... le système que nous avons alors avec la loi Hypergéométrique peut alors s'écrire:

$$\begin{aligned} 1-\alpha &\cong \sum_{k=0}^A C_k^n p_\alpha^k (1-p_\alpha)^{n-k} = \sum_{k=0}^A C_k^n p_\alpha^k q_\alpha^{n-k} \\ \beta &\cong \sum_{k=0}^A C_k^n p_\beta^k (1-p_\beta)^{n-k} = \sum_{k=0}^A C_k^n p_\beta^k q_\beta^{n-k} \end{aligned} \quad (74.413)$$

Mais encore une fois, il n'est pas possible (à ma connaissance) de résoudre le système précédent analytiquement. En 1967, H. R. Larson a alors créé un abaque de la loi binomiale cumulative fréquemment désigné sous le terme de "**nomographe binomial de Larson**" (je n'en ai jamais trouvé ni vu un pour la loi hypergéométrique malheureusement...) qui permet de résoudre approximativement le problème de la détermination de la valeur de A et de n pour le plan d'échantillonnage:

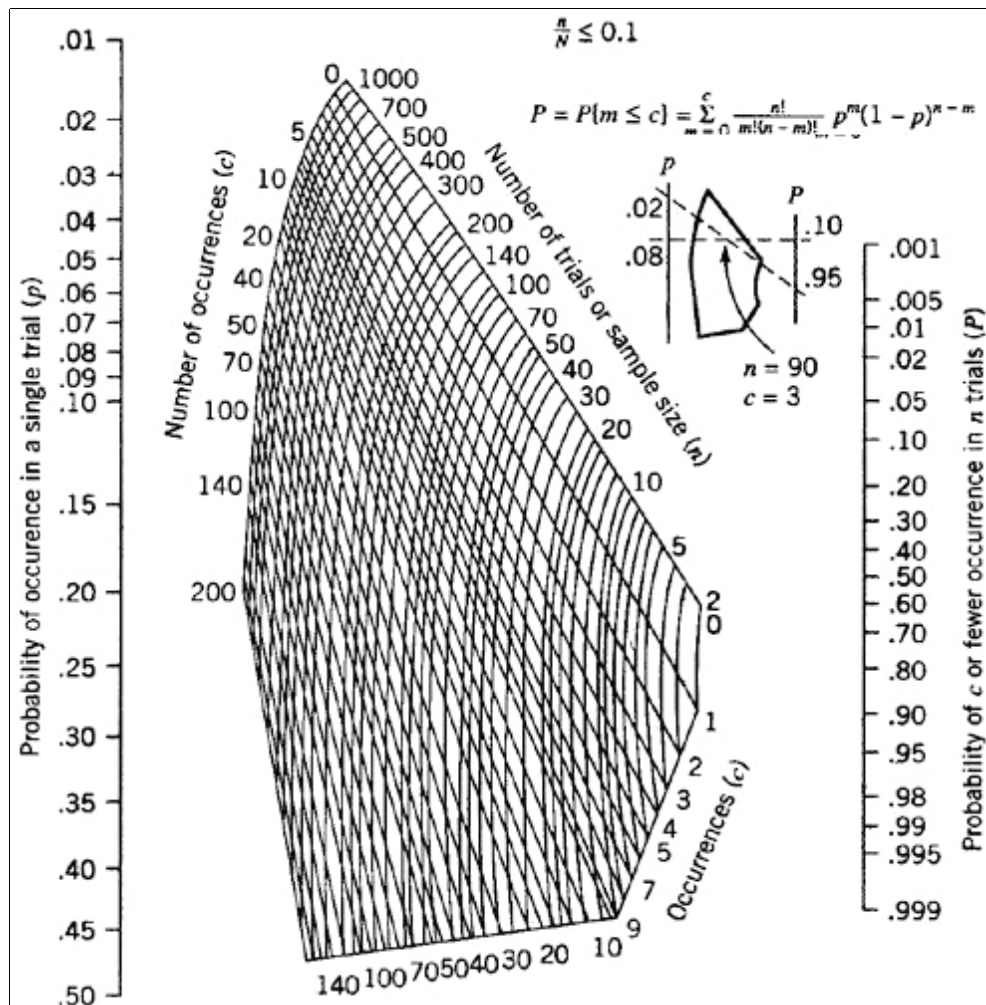


Figure: 74.45 - Source: Montgomery, D. C. (1996), Introduction to Statistical Quality Control (3e ed.), Ed. Wiley

Exemple:

Un lot de bouteilles est livré sous forme de lots de 10'000 unités correspondant à N . Nous cherchons à mettre en place un plan de contrôle de la réception par attributs avec la probabilité cumulée (risque) α de 1% que le client rejette à tort le lot avec moins de 2.5% (p_α) de non-conformes. De son côté le client souhaite une probabilité cumulée (risque) β de 10% d'accepter à tort un lot avec plus de 5% (p_β) de non-conformes (au fait les exigences sont indépendantes de la taille du N du lot).

Le but est de déterminer donc la valeur de A et de n à l'aide du nomographe. Pour cela, comme indiqué sur le nomographe, il faut tracer deux droites:

- Points de départs des deux droites: 0.99 (100%-1%) et 0.1 (10%) de l'axe de droite.
- Points d'arrivée des deux droites: 0.025 (2.5%) et 0.05 (5%) de l'axe de gauche.

et l'intersection des deux droites donnera les paramètres recherchés. Soit dans le cas présent:

$$\begin{aligned} A &\cong 30 \\ n &\cong 700 \end{aligned} \quad (74.414)$$

Nous voyons aussi que si nous injectons ces deux valeurs dans:

$$\begin{aligned}
 1 - \alpha &\cong \sum_{k=0}^A C_k^n p_{\alpha}^k (1 - p_{\alpha})^{n-k} = \sum_{k=0}^A C_k^n p_{\alpha}^k q_{\alpha}^{n-k} \\
 \beta &\cong \sum_{k=0}^A C_k^n p_{\beta}^k (1 - p_{\beta})^{n-k} = \sum_{k=0}^A C_k^n p_{\beta}^k q_{\beta}^{n-k}
 \end{aligned}
 \quad (74.415)$$

avec Microsoft Excel, nous sommes en réalité loin du compte! Donc l'erreur du nomographe peut amener à donner l'impression de grossières approximations! Mais changez dans Microsoft Excel la valeur 30 par 27 et la valeur 700 par 695 et vous verrez en réalité ce que vous pouvez en conclure...

Nous voyons aussi via cet exemple qu'un plan d'échantillonnage par attribut est effectivement aussi beaucoup plus coûteux qu'un plan par mesurage (puisque nous avons repris le même exemple avec les mêmes paramètres que dans la partie des plans d'échantillonnage avec mesures).

Le lecteur aura aussi remarqué que N n'influence pas sur la valeur de A et n lorsque nous utilisons le nomographe. Donc si nous utilisons des outils informatiques pour chercher la solution, nous avons pour contrainte que n sera compris entre 0 et 1'000 et A entre 0 et 150.

Remarque: Enfin, indiquons que certains utilisent l'approximation Poissonienne de la loi binomiale (lorsque la probabilité de non-conformité est très faible que la taille du lot est très grande et même parfois quand ces critères sont loin d'être respectés...).

Revenons maintenant sur le double échantillonnage en utilisant l'approximation binomiale qui était pour rappel:

$$P_a = P(X \leq A) = \sum_{k=0}^A P(n, p, X = k) = \sum_{k=0}^A C_k^n p^k (1 - p)^{n-k} = \sum_{k=0}^A C_k^n p^k q^{n-k} \quad (74.416)$$

et considérons un plan à double échantillonnage avec:

$$\begin{aligned}
 n_1 &= 50 \\
 n_2 &= 100
 \end{aligned}
 \quad (74.417)$$

et où la proportion de non-conformes est connue comme étant de 5%. Alors, dans ce cas la probabilité cumulée d'avoir par exemple au moins 1 non-conforme au premier échantillonnage et correspondant donc à la probabilité d'acceptation du lot est de:

$$P_a^1 = P(X_1 \leq 1) = \sum_{k_1=0}^1 \frac{50!}{k_1! (50 - k_1)!} 0.05^{k_1} (1 - 0.05)^{50-k_1} \cong 0.279 \quad (74.418)$$

Donc jusque là rien de spécial! Maintenant pour obtenir la probabilité d'acceptation du second lot, nous devons lister toutes les manières dont le second échantillonnage. Supposons pour l'exemple que le lot est accepté qu si et seulement si le nombre cumulé de non-conformes sur les deux échantillonnages est compris entre 2 et 3 unités (choix arbitraire mais simplifié pour l'exemple).

Nous avons alors les combinaisons suivantes possibles de tirage entre les deux échantillonnages: $\{0,2\}\{0,3\}, \{1,1\}, \{1,2\}, \{3,0\}, \{2,1\}, \{2,0\}$. Mais comme nous acceptons le lot si le premier échantillonnage donne un nombre de non-conformes inférieur ou égal à 1, il reste alors seulement les combinaisons suivantes $\{3,0\}, \{2,1\}, \{2,0\}$ à considérer en double échantillonnage. Donc il nous faut calculer les probabilités suivantes:

$$\begin{aligned}
 P_a &= P_a^1 + P_a^2 = 0.279 + [P(X_1 = 2, X_2 \leq 1) + P(X_1 = 3, X_2 = 0)] \\
 &\cong 0.279 + [P(X_1 = 2)P(X_2 \leq 1) + P(X_1 = 3)P(X_2 = 0)] \\
 &= 0.279 + \left[\frac{50!}{2!48!} 0.05^2 0.95^{48} \sum_{k_2=0}^1 \frac{100!}{k_2!(100-k_2)!} 0.05^{k_2} 0.95^{100-k_2} \right. \\
 &\quad \left. + \frac{50!}{3!47!} 0.05^3 0.95^{47} \frac{100!}{0!100!} 0.05^0 0.95^{100} \right] \quad (74.419) \\
 &\cong 0.279 + [0.261 \cdot 0.037 + 0.220 \cdot 0.0059] = 0.279 + [0.0097 + 0.001] \\
 &= 0.279 + 0.0107 = 0.2897
 \end{aligned}$$

4.2.1. CALCUL DES PARAMÈTRES PAR UTILISATION DE LA NORME ISO 2859-1

La norme ISO 2859 (qui découle du standard MIL 105E) est consacrée aux règles d'échantillonnage pour les contrôles par attributs. Au même titre que la norme AFNOR X06-023 elle est basée le concept de NQA. (pourcentage d'individus non-conformes qui ne doit pas être dépassé pour qu'une production, contrôlée sur une série de lots, puisse être considérée comme satisfaisante).

Selon cette norme, il faut que le NQA soit une des 26 valeurs recommandées par la norme. Nous choisirons pour niveau d'inspection le numéro II (ils sont décrits dans la norme). Par rapport à l'effectif choisi, comme pour la norme AFNOR, ISO 2589 donne un code (lettre) qui va être utilisée ensuite dans une table. Enfin, en fonction du type de plan (simple, double ou triple) on trouve les paramètres A et n .

Exemples:

E1. Dans le cadre de notre exemple - le producteur de boissons gazeuses - qui reçoit des lots de 10'000 bouteilles, le niveau de qualité NQA est donc de 1%. Comme pour l'exemple avec la norme AFNOR, nous prendrons un contrôle de type II basé sur un plan simple.

Une table disponible dans la norme nous dit que nous devons utiliser le code L pour la prochaine étape. Ensuite, pour un contrôle simple, la norme nous dit d'utiliser la table IIA.

Nous y lisons alors que n doit être égal à 200 et que A vaut 5. Nous constatons donc que la valeur obtenue est très différente de celle calculée théoriquement. Honnêtement la raison de cette différence m'est inconnue...

E2. Les praticiens lors de l'utilisation de cette norme souhaitent souvent pouvoir calculer le risque fournisseur et client associé aux facteurs n et A donnés par la norme relativement au NQA. choisi pour un lot d'une qualité donnée. Alors comme dans le cas présent, 200/10'000 est plus petit que 10%, nous pouvons utiliser l'approximation binomiale de la loi hypergéométrique. Supposons que nous souhaitons donc calculer le risque fournisseur pour le de l'exemple précédent (celui avec l'utilisation du nomographe):

$$1 - \alpha \cong \sum_{k=0}^A C_k^n p_{\alpha}^k (1 - p_{\alpha})^{n-k} = \sum_{k=0}^5 C_k^{200} 0.025_{\alpha}^k (1 - 0.025)^{200-k} \quad (74.420)$$

$$= \text{BINOMDIST}(5; 200; 2.5\%; 1) = 61.59\%$$

d'où un risque fournisseur de 100%-61.59% soit 38.40% ce qui est considérable. Ainsi, ce plan d'échantillonnage obtenu via les normes est adapté à des proportions de non-conformes de beaucoup inférieures à 2.5% (au fait inférieur à ~1.25%).

Pour le risque client, nous procédons de la même manière:

$$\beta \cong \sum_{k=0}^A C_k^n p_{\beta}^k (1 - p_{\beta})^{n-k} = \sum_{k=0}^5 C_k^{200} 0.05_{\beta}^k (1 - 0.05)^{200-k} \quad (74.421)$$

$$= \text{BINOMDIST}(5; 200; 5\%; 1) = 6.23\%$$

ce qui est par contre plus acceptable. Au fait nous constatons que la norme est plus adaptée aux critères du client que du fournisseur.

Remarque: Le logiciel QuickControl Pro de la société Logystem SA par exemple gère l'échantillonnage selon les normes ISO 2859 et 3951.

4.2.2. COURBE D'EFFICACITÉ

La courbe qui représente les probabilités d'acceptation d'un lot en fonction de la qualité effective (proportion de pièces défectueuses) du lot s'appelle "**courbe d'efficacité du plan d'échantillonnage**".

Cette courbe s'obtient donc en traçant sur un graphique la probabilité cumulée d'acceptation du lot que nous notons sur ce type de graphique, en fonction de la proportion présumée de pièces non-conformes (que nous notons p) dans le lot.

 **Exemple:** Pour un contrôle de qualité par attributs, nous appliquons le plan de contrôle basé sur le prélèvement de 200 individus, le lot sera accepté si le nombre d'éléments défectueux est inférieur ou égal à 5. Nous souhaitons tracer la probabilité d'accepter un lot où le pourcentage de défectueux est p , pour p variant de 0 à 7% par pas de 0.1%.

Nous avons alors avec la version française de Microsoft Excel 14.0.6123 pour les premières lignes:

	A	B		A	B
1	Qualité Effective	Prob. Cumulée	1	Qualité Effective	Prob. Cumulée
2	0.00%	100%	2	0	=LOIBINOMIALE.N(5;80;A2;1)
3	0.20%	100%	3	0.002	=LOIBINOMIALE.N(5;200;A3;1)
4	0.20%	100%	4	0.002	=LOIBINOMIALE.N(5;200;A4;1)
5	0.30%	100%	5	0.003	=LOIBINOMIALE.N(5;200;A5;1)
6	0.40%	100%	6	0.004	=LOIBINOMIALE.N(5;200;A6;1)
7	0.50%	100%	7	0.005	=LOIBINOMIALE.N(5;200;A7;1)
8	0.60%	100%	8	0.006	=LOIBINOMIALE.N(5;200;A8;1)
9	0.70%	100%	9	0.007	=LOIBINOMIALE.N(5;200;A9;1)
10	0.80%	99%	10	0.008	=LOIBINOMIALE.N(5;200;A10;1)
11	0.90%	99%	11	0.009	=LOIBINOMIALE.N(5;200;A11;1)
12	1.00%	98%	12	0.01	=LOIBINOMIALE.N(5;200;A12;1)
13	1.10%	98%	13	0.011	=LOIBINOMIALE.N(5;200;A13;1)
14	1.20%	97%	14	0.012	=LOIBINOMIALE.N(5;200;A14;1)
15	1.30%	95%	15	0.013	=LOIBINOMIALE.N(5;200;A15;1)
16	1.40%	94%	16	0.014	=LOIBINOMIALE.N(5;200;A16;1)
17	1.50%	92%	17	0.015	=LOIBINOMIALE.N(5;200;A17;1)
18	1.60%	90%	18	0.016	=LOIBINOMIALE.N(5;200;A18;1)
19	1.70%	87%	19	0.017	=LOIBINOMIALE.N(5;200;A19;1)
20	1.80%	85%	20	0.018	=LOIBINOMIALE.N(5;200;A20;1)
21	1.90%	82%	21	0.019	=LOIBINOMIALE.N(5;200;A21;1)
22	2.00%	79%	22	0.02	=LOIBINOMIALE.N(5;200;A22;1)
23	2.10%	75%	23	0.021	=LOIBINOMIALE.N(5;200;A23;1)
24	2.20%	72%	24	0.022	=LOIBINOMIALE.N(5;200;A24;1)
25	2.30%	69%	25	0.023	=LOIBINOMIALE.N(5;200;A25;1)
26	2.40%	65%	26	0.024	=LOIBINOMIALE.N(5;200;A26;1)
27	2.50%	62%	27	0.025	=LOIBINOMIALE.N(5;200;A27;1)
28	2.60%	58%	28	0.026	=LOIBINOMIALE.N(5;200;A28;1)
29	2.70%	55%	29	0.027	=LOIBINOMIALE.N(5;200;A29;1)
30	2.80%	51%	30	0.028	=LOIBINOMIALE.N(5;200;A30;1)
31	2.90%	48%	31	0.029	=LOIBINOMIALE.N(5;200;A31;1)
32	3.00%	44%	32	0.03	=LOIBINOMIALE.N(5;200;A32;1)
33	3.10%	41%	33	0.031	=LOIBINOMIALE.N(5;200;A33;1)
34	3.20%	38%	34	0.032	=LOIBINOMIALE.N(5;200;A34;1)

Figure: 74.46 - Données pour une courbe d'efficacité obtenue avec Microsoft Excel 14.0.6123

Ce qui donne graphiquement:

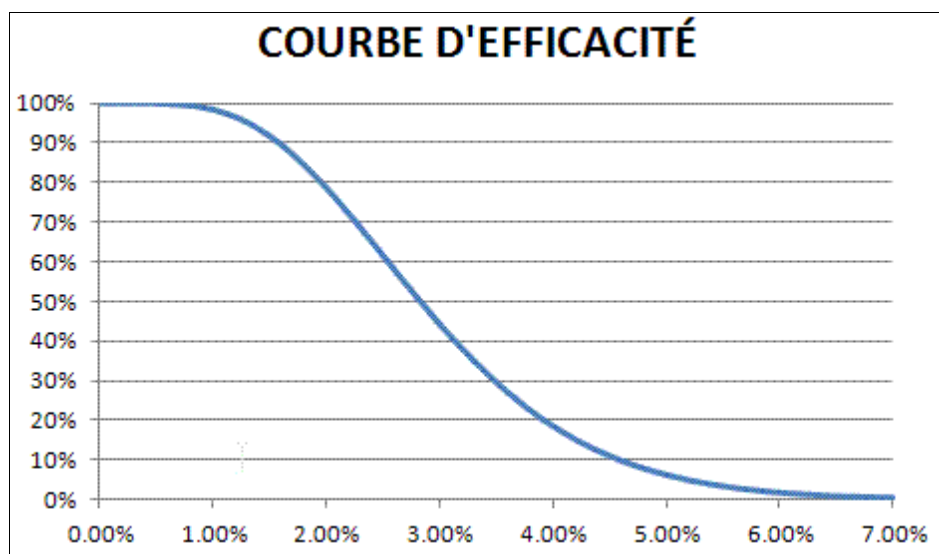


Figure: 74.47 - Graphique correspondant aux valeurs calculées dans Microsoft Excel 14.0.6123

Une manière élégante de représenter une courbe d'efficacité est la suivante:

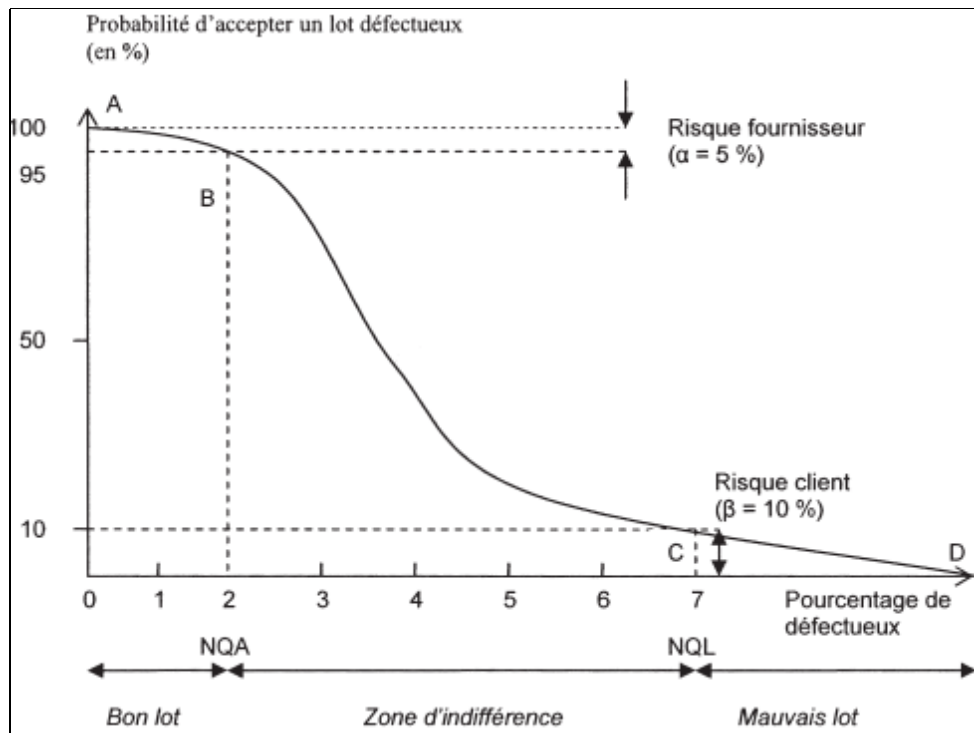


Figure: 74.48 - Concept général d'une courbe d'efficacité (source: ISBN 978-2-297-01111-2)

La courbe d'efficacité montre que la probabilité d'accepter un lot défectueux diminue lorsque le pourcentage de produits défectueux dans le lot augmente: lorsque la qualité est bonne, la probabilité d'accepter le lot est élevée ; lorsqu'elle est mauvaise, la probabilité d'accepter le lot est faible. La courbe se caractérise par différents points:

- Au point A, le pourcentage de produits défectueux du lot est de 0 %, il y a 100 % de chances que le client accepte le lot. Le fournisseur ne court aucun risque
- Au point B, le pourcentage de défautueux effectif est de 2 %. Ce pourcentage correspond au Niveau de qualité acceptable (NQA) défini d'un commun accord entre le client et le fournisseur. Mais, comme le contrôle est effectué à l'aide d'échantillons aléatoires, il se peut que le lot soit rejeté par le client. Le risque α pour le fournisseur est de 5 %. Réciproquement, il y a une probabilité de 95 % ($1-\alpha$) que le client accepte le lot.
- Au point C, le pourcentage de défautueux effectif est de 7 % correspondant au Niveau de qualité limite (NQL). La probabilité d'acceptation d'un lot de ce niveau de qualité est égale au risque client β . Il y a exactement à ce point 10 % de chance d'accepter un lot mauvais, ce qui signifie que le client va conserver un nombre de défautueux supérieur à ce qu'il souhaite.
- Au point D, le pourcentage de produits défectueux effectif du lot est de 100 % ; il y a 0 % de chance d'accepter un tel lot.

Ce type de courbe d'efficacité basé sur de gros lots échantillonnés (dixit l'utilisation de la loi binomiale) est appelée dans le cas de l'utilisation de la loi binomiale: "**courbe d'efficacité de type B**" (CE-B). Si évidemment la taille de l'échantillon est significative par rapport à la taille du lot nous utiliserons la loi hypergéométrique et nous parlerons alors de: "**courbe d'efficacité de type A**" (CE-A).

5. CARTES DE CONTRÔLES (CC) DE LA QUALITÉ

Une carte de contrôle (appelée aussi parfois "**carte séquentielle**") est une représentation empirique plane de la variation d'une mesure précise d'un processus ou d'un procédé où l'axe vertical représente l'indicateur quantitatif choisi et l'axe horizontal le temps (c'est donc une sous-famille des "**run charts**").

Cette méthode de contrôle qualité non destructive et dite "on-line" (car l'acquisition peut se faire en temps réel) permettant de mettre en évidence des causes spéciales de variations et d'optimiser les fréquences de réglages (étalonnage) aurait été créée par l'ingénieur physicien Walter A. Shewhart alors employé chez Bell Labs dans les années 1920 et s'est imposée en tant que standard international dès 1991 sous la norme ISO 8258. C'est d'ailleurs aussi Shewhart qui inspira le non moins connu W. Edwards Deming (statisticien) dans le domaine de la gestion de projets et de la qualité (tous les maîtres dans le domaine de la gestion de projets de pointe ont une formation scientifique...).

Remarque: Le sujet des cartes de contrôle est immensément vaste et il y a une littérature anglophone très abondante sur le sujet. Nous souhaitons ici ne donner que les fondamentaux (donc uniquement les cartes de base!) afin de montrer une application pratique des statistiques d'ordre et des statistiques des valeurs extrêmes développées en détail dans le chapitre de Statistiques. Pour placer un ordre de grandeur, la majorité des logiciels spécialisés proposent entre 20 et 30 cartes de contrôle différentes!

En début d'utilisation de cartes de contrôle, il est conseillé de mettre des cartes sur toutes les caractéristiques mesurables importantes du produit. Les cartes obtenues révèlent en général rapidement quelles cartes sont nécessaires ou inutiles. Les cartes inutiles ou inadaptées seront retirées et d'autres cartes éventuellement ajoutées.

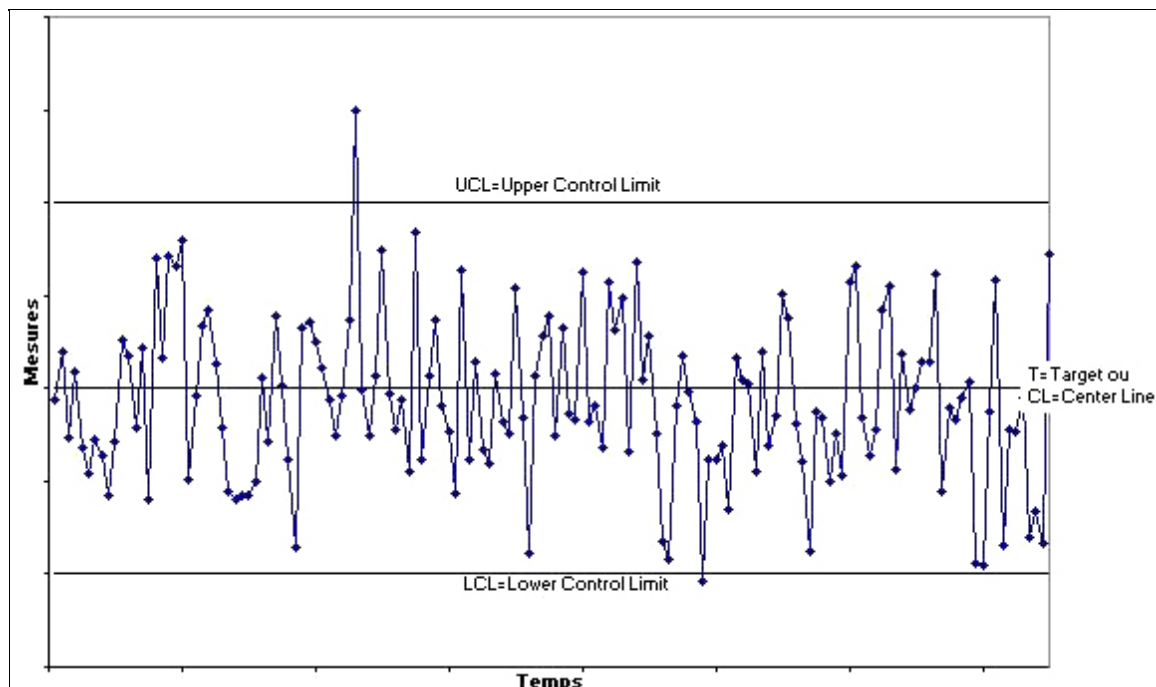


Figure: 74.49 - Principe d'une carte de contrôle

Les limites peuvent être soit imposées/spécifiées par un client et dès lors nous les notons *USL* et *LSL*, sinon elles sont calculées en interne et dès lors nous les notons *UCL* (Upper Control Limit) et *LCL* (Lower Control Limit). Pour améliorer l'efficacité d'une carte de contrôle, nous ajoutons parfois sur la carte des limites plus étroites appelées "**limites de surveillance**" (dans la réalité nous représentons quasiment systématiquement les limites calculées ET imposées par la politique qualité sur la carte de contrôle).

Remarque: Dans la pratique je recommande fortement de représenter à la fois *CL*, *UCL*, *LCL* et comme dans la réalité il y normalement toujours des limites et objectifs qui sont imposés contractuellement par le client, de représenter aussi la cible *T* et *UCL* et *LCL* sur le même graphique quand cela est possible!!!

La carte de contrôle est donc un outil de mesure permettant de visualiser la variation et donc de déterminer le moment où apparaît une cause assignable entraînant une dérive d'un processus ou procédé de fabrication/administratif ou encore une variation d'une valeur financière sur les marchés boursiers qui nécessite une action de correction rapide pour diminuer les coûts de non-qualité. Ainsi, dans l'idéal le processus sera arrêté ou corrigé au bon moment, c'est-à-dire avant qu'il ne produise trop de livrables/services non conformes (hors de l'intervalle de tolérance) et le surcontrôle (réajustement inutile pour cause de variabilité commune) sera évité.

Le type de carte de contrôle correspond au type de caractéristique qui fait l'objet d'un contrôle. Toutefois l'aspect économique peut être un facteur important dans le choix de la carte de contrôle à mettre en oeuvre. Si nous nous intéressons qu'à vérifier si une pièce est conforme ou non à certaines normes, une carte par attribut (voir plus loin) sera utilisée, mais nécessitera néanmoins une taille d'échantillon élevée. Le contrôle de caractéristiques quantitatives nécessite l'emploi de cartes de contrôle par mesure. Ce type de contrôle est plus efficace et plus précis mais également plus dispendieux puisqu'il requiert l'emploi d'instruments de mesure qui doivent être vérifiés régulièrement.

Avant de partir dans les détails mathématiques signalons que les cartes de contrôle parmi les plus utilisées sont les cartes de contrôle par mesure de la moyenne $\bar{\bar{x}}$ en association avec celle de l'étendue *R* ("range" en anglais) avec limites de Shewhart (quand nous parlons de "limites de Shewart", nous faisons implicitement référence au fait que la limite supérieure de contrôle calculée *UCL* et la limite inférieure de contrôle calculée *LCL* sont empiriquement à $\pm 3\sigma$ de la moyenne que la distribution soit symétrique ou non!). Ces deux cartes sont établies et interprétées la majorité du temps ensemble car les deux paramètres sont indépendants et complémentaires et simples à comprendre (la majorité des employés dans les entreprises ne sachant pas ce qu'est un écart-type). La valeur moyenne peut varier sans que la dispersion ne varie et inversement.

Remarques:

R1. Dans le cas d'une carte de contrôle avec limites de Shewhart, les limites de surveillance sont traditionnellement placées à $\pm 2\sigma$ (à l'opposée des limites de contrôle qui sont donc placées à $\pm 3\sigma$).

R2. Dans le cas d'une distribution Normale, $\pm 3\sigma$ correspond à un intervalle de confiance de 99.73% comme nous l'avons vu plus haut dans le présent chapitre et dans celui de Statistiques. Cela correspondant donc à un alpha d'environ:

$$\alpha = 100\% - 99.73 = 0.27\%$$

La majeure partie de l'industrie prend donc $\pm 3\sigma$ comme limites, mais en toute rigueur les distributions de certaines cartes ne sont pas symétriques. Il convient alors de prendre des bornes correspondant à une probabilité cumulée de $\alpha/2$ ce qui peut se calculer assez aisément avec de nombreuses distributions (mais pas toutes!) en ayant juste à disposition un tableur bon marché. Le choix de la valeur de α dépend de la politique qualité de l'entreprise et a l'avantage d'être un indicateur beaucoup plus précis que les limites de Shewhart (ces dernières étant erronées lorsque les distributions sont asymétriques).

R3. Shewhart aurait proposé qu'il y ait au moins 25 prélèvements (échantillons) de 4 individus pour que la validité de certaines cartes de contrôle (que nous verrons plus loin) commence à être acceptable.

R4. Comme nous le verrons, la quasi-totalité des cartes de contrôle aux mesures supposent une distribution Normale et que les observations sont indépendantes (entre sous-groupes et à l'intérieur des sous-groupes).

Par exemple (exemple qui sera donc détaillé mathématiquement un peu plus bas), pour les procédés déjà stables car lancés depuis un bout de temps, nous effectuons plusieurs observations individuelles sur plusieurs sous-groupes numérotés à une fréquence de temps donnée (toutes les heures, trois fois par jour ...). Sur chaque sous-groupe k chronologique, nous effectuons n observations. Nous reportons sur la carte de moyenne $\bar{\bar{X}}$ la moyenne du sous-groupe en fonction de son numéro chronologique qui sera reporté sur l'axe horizontal des cartes de contrôle. En raison du théorème de la limite centrale (TCL) démontré dans le chapitre de Statistiques, la moyenne des valeurs sur la carte de contrôle suit une loi Normale (donc symétrique!) que les observations des sous-groupes soient normalement distribuées ou non!!! Cette hypothèse est valable même pour des échantillons de petite taille et pour un processus de fabrication sous-contrôle, ce qui est fréquent en contrôle qualité. Une production sera dite stable, si la tendance et la dispersion sont statistiquement constantes dans le temps (variables identiquement distribuées et indépendantes).

5.1. RÈGLES EMPIRIQUES DE LA WECO

L'interprétation d'une carte de contrôle aussi puissante et élégante qu'elle soit mathématiquement n'est pas aisée et requière expérience et savoir faire pour savoir s'il faut procéder à une intervention corrective ou non. La Western Electric (WECO) aurait publié, en 1956, un ouvrage avec des règles empiriques simples permettent de prendre des décisions plus simplement (le ministère français de la santé a fait de même 50 ans après... mais avec des règles qui diffèrent un peu).

La WECO s'est basée à l'époque sur l'hypothèse que les distributions statistiques sont toujours symétriques (même hypothèse que Shewhart) et a donc adopté dans l'énoncé des règles des limites de contrôle qui sont des multiples k entiers de l'écart-type. Évidemment, rien n'empêche le spécialiste de la qualité d'adapter ces règles avec des limites probabilistes correspondant à un intervalle de confiance (la majorité des logiciels spécialisés permettent de choisir les règles de la WECO à appliquer pour la détection automatiquement de non-conformités).

Dans le cas des cartes de contrôle mesurées, parmi ces règles édictées voici celles qui sont applicables afin d'identifier si un processus ou procédé est défaillant et qui ont été complétées avec les années par d'autres spécialistes (nous avons représenté ci-dessous quelques cas de certaines règles dans une unique carte de contrôle pour des raisons pédagogiques évidentes):

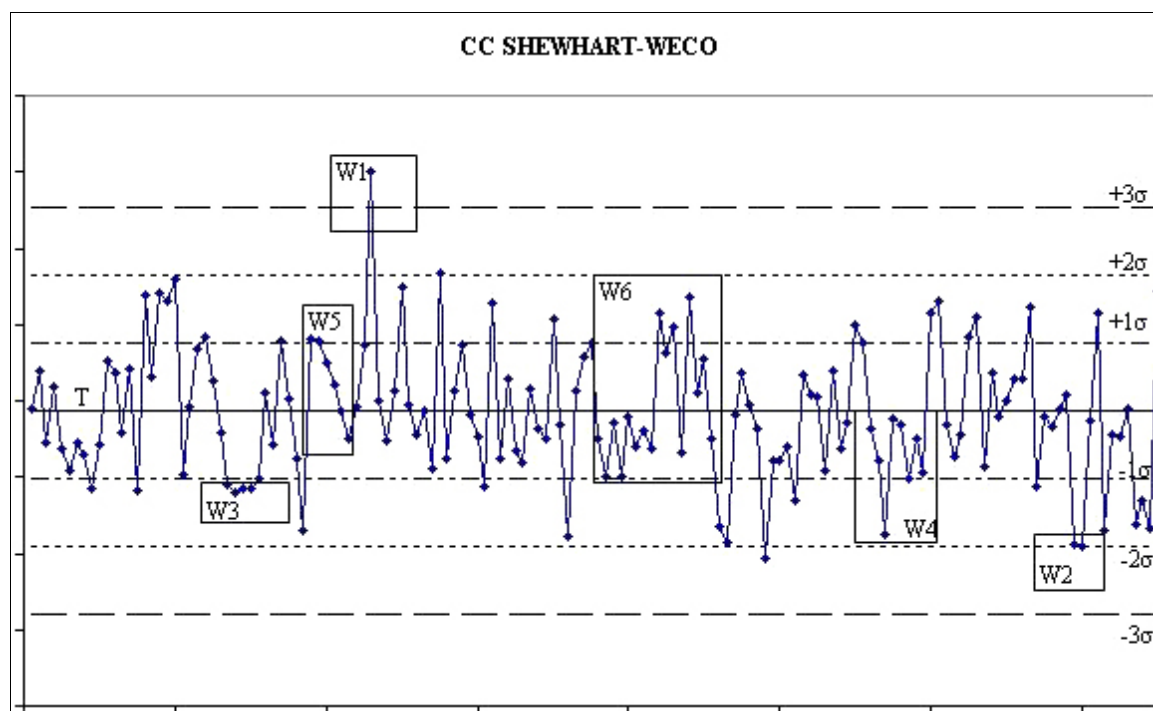


Figure: 74.50 - Carte de contrôle générique avec quelques points de la WECO

W1. Un point de mesure est au-delà des *USL*, *LSL* spécifiés par le client ou au-delà des *UCL*, *LCL* probabilistes ou empiriques correspondant à $\pm 3\sigma$.

W2. Deux points consécutifs de mesure sont au-delà *UCL*, *LCL* probabilistes ou empiriques correspondant à $\pm 2\sigma$.

W3. Quatre points consécutifs de mesure sont au-delà *UCL*, *LCL* probabilistes ou empiriques correspondant à $\pm 1\sigma$ (souvent appelés "**limites d'alerte**").

W4. Huit points successifs tombent d'un même côté de la moyenne (ce chiffre est empirique et sujet à débats...).

W5. Six points consécutifs sont sur une tendance haussière ou respectivement baissière.

W6. Quatorze points consécutifs sont sur une alternance systématique de hausse/baisse.

W7. Deux points consécutifs au-delà des *UCL*, *LCL* probabilistes ou empiriques correspondant à $\pm 2\sigma$.

W8. Quatre points consécutifs hors des limites à $\pm 1\sigma$ et du même côté de la ligne centrale.

W9. Huit points consécutifs dans les limites de $\pm 1\sigma$.

Remarque: Un logiciel comme Minitab intègre par exemple 8 de ses règles pour mettre en évidence une anomalie. Pour plus d'informations, le lecteur pourra se reporter au e-book que j'ai écrit sur ce logiciel.

Si le processus/procédé est hors de contrôle, des actions correctives sont donc mises en place. Les causes de variation peuvent être aléatoires ou déterministes. Si les causes sont seulement dues au hasard, elles sont alors appelées "**causes aléatoires**" (Shewhart) ou "**causes communes**" (Deming) et sont donc assimilées à la dispersion instantanée dont nous avons fait mention plus haut. Pas toutes les variations sont dues au hasard évidemment, certaines sont spécifiques et identifiables de manière déterministe et certaine. Dans ce dernier cas, les variations sont appelées "**variations assignables**" (Shewhart) ou "**causes spéciales**" (Deming) et sont assimilées à la dispersion globale (dont nous avons aussi déjà fait mention plus haut). Corriger les causes spéciales est évidemment bien plus simple que de corriger les causes aléatoires. L'objectif principal des cartes de contrôle est bien évidemment d'identifier les causes spéciales...

Voilà en ce qui concerne les us et coutumes d'interprétation des points de mesure. Maintenant, ce prélude non mathématique effectué, voyons les différents types de cartes de contrôle (CC) courantes avec un exemple concret pour chacune.

5.2. ÉCHANTILLONNAGE

L'usage des cartes de contrôle pose au début deux difficultés majeures:

- Le choix de la bonne carte de contrôle
- Le choix du nombre d'échantillons et de prélèvements (fréquence de l'échantillonnage)

Nous souhaiterions ici discuter du deuxième point. Il faut d'abord observer que c'est bien évidemment la cadence d'exécution et la qualité du fonctionnement qui va nous donner la réponse. Cette qualité est appréciée par la moyenne du nombre d'interventions de corrections observées pendant un laps déterminé dans le passé. Plus le nombre d'actions correctives a été grand, moins la qualité du fonctionnement du procédé/processus est bonne et plus il faudra faire de prélèvements.

Le choix du nombre d'individus dans un prélèvement (échantillonnage) n'est lui pas un problème, car il découlera du calcul statistique de l'intervalle de confiance de l'indicateur statistique que l'on s'impose et des propriétés de loi de distribution ainsi que des hypothèses de construction de la carte de contrôle (voir plus loin dans le texte qui va suivre).

La difficulté est donc de quantifier la fréquence de prélèvement. Il existe pour ce point deux méthodes empiriques principales:

- Choisir une fréquence telle que les actions correctives sont au moins quatre fois plus faible que la fréquence de prélèvement (ce critère est donc indépendant du nombre d'individus par prélèvements donc parfois peu applicable).

- Nous considérons T comme étant la durée de vie du procédé/processus entre chaque changement/modification. Nous notons comme à l'habitude n le nombre d'individus par prélèvement (échantillons) nécessaires pour garantir de l'intervalle de confiance imposé de l'indicateur statistique (pour l'analyse de la dispersion instantanée). Nous notons N le nombre d'exécutions du procédé/processus pendant la période T . Nous avons donc au maximum N/n prélèvements possibles. Enfin, ce qui décidera du nombre de prélèvements, sera la précision de l'exigence de l'inférence statistique sur un indicateur statistique pour la dispersion globale.

Dans la totalité des cartes de contrôle que nous allons présenter ci-dessous, nous avons présenté des exemples avec des prélèvements à taille constante (donc les limites des cartes de contrôle sont presque toujours constantes à quelques exceptions près). Effectivement, si la taille de ceux-ci vient à changer, c'est que le rythme de production (ou la taille des lots de livraison) a changé et donc les paramètres de fabrication aussi. Il conviendrait alors de recommencer une nouvelle carte! Il en est de même après une réingénierie du procédé/processus ou changements des paramètres du procédé/processus comme le montre par exemple l'image ci-dessous:

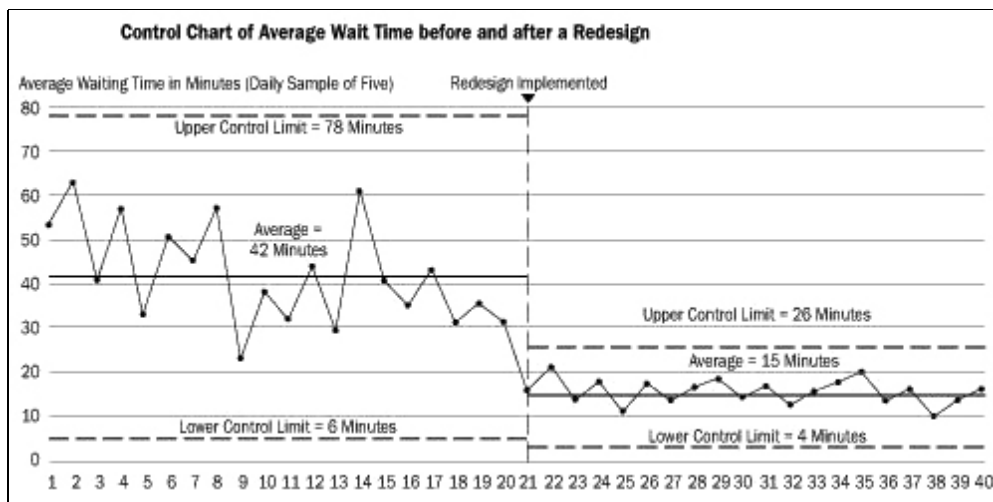


Figure: 74.51 - Exemple de carte de contrôle avant et après mesure corrective

5.3. CARTES DE CONTRÔLES QUALITATIVES (AUX ATTRIBUTS)

Pour mesurer des variables qualitatives (% de défectueux, % de pannes, ...), nous nous servons de cartes aux attributs p (binomial), np ou c pour contrôler les attributs dans le temps.

Les cartes de contrôle par attributs sont très simples d'utilisation et d'interprétation, mais ont plusieurs inconvénients comme la forte dissymétrie des répartitions, l'absence de limite inférieure pour les faibles tailles d'échantillons, la faible efficacité de détection des détériorations.

5.3.1. CC-P

La carte de contrôle de type p (proportion) est utilisée lorsqu'il s'agit de travailler avec des ratios, proportions ou pourcentages de conformité ou non-conformité d'échantillons d'un prélèvement.

De bons exemples de carte de contrôle de type p sont: l'inspection des produits d'une ligne de production ou de la réception de lots d'un fournisseur ou encore contrôle du dysfonctionnement d'un certain nombre d'appareils, le respect de délais de livraison ou des spécifications de produits.

Avantages: Très simple à construire car ne nécessite aucune connaissance statistique poussée pour être calculée.

Problèmes: En réalité c'est une carte dont les limites de contrôle peuvent être asymétriques si l'application pratique est à la limite de l'acceptable lorsque les hypothèses théoriques de construction ne sont pas respectées. Une difficulté parfois rencontrée par les utilisateurs de cette carte est de déterminer la taille du lot et du niveau d'acceptation (A) ou rejet (R) équivalant à l' USL en se basant sur les développements (et normes) des plans d'échantillonnages disponibles un peu plus loin dans le présent chapitre. Enfin, une dernière difficulté est que certains employés maîtrisent mal le concept de pourcentages.

La loi probabilité utilisée dans ce contexte est la loi binomiale (cf. chapitre de Statistiques) où p représentera la proportion en % de non-conformités et q (qui vaut donc $1-p$) représentera la proportion en % d'éléments conformes.

Nous avons vu dans le chapitre de Statistiques qu'il nous était possible d'écrire formellement, sous certaines hypothèses, uniquement l'intervalle de confiance suivant:

$$\hat{p} - Z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \leq p \leq \hat{p} + Z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \quad (74.422)$$

Donc nous pouvons alors calculer des probabilités de lots non-conformes par excès ou par défaut en utilisant le fait que p suit une loi Normale de paramètres:

$$N(\mu, \sigma^2) \cong N\left(\hat{p}, \frac{\hat{p}(1 - \hat{p})}{n}\right) \quad (74.423)$$

Bref, nous n'avons pas d'autres choix que de baser la carte de contrôle là-dessus. Dès lors, la première étape dans la création d'une carte de contrôle de type p est de calculer la proportion de non-conformité, du moins son estimateur, pour chaque lot i :

$$\hat{p}_i = \frac{k_i}{N_i} \quad (74.424)$$

où k_i est le nombre d'éléments non-conformes et N_i le nombre total d'éléments du lot contrôlé i .

La proportion moyenne d'éléments non-conformes, qui correspondra à la ligne centrale (CL) de la carte de contrôle sera donnée par:

$$\bar{\hat{p}} = CL = \frac{\sum_{i=1}^n k_i}{\sum_{i=1}^n N_i} \quad (74.425)$$

et si les N_i sont tous égaux, cela se simplifie en:

$$\bar{\hat{p}} = CL = \frac{\sum_{i=1}^n k_i}{n \cdot N} = \frac{1}{n} \sum_{i=1}^n \frac{k_i}{N} = \frac{1}{n} \sum_{i=1}^n p_i \quad (74.426)$$

où n représente donc le nombre de lots contrôlés. Nous avons alors:

USL_p = imposé par client/direction

$$UCL_p = \bar{\hat{p}} + 3\sqrt{\frac{\bar{\hat{p}}(1-\bar{\hat{p}})}{N_i}} \quad (74.427)$$

$$CL_p = \bar{\hat{p}}$$

$$LCL_p = \bar{\hat{p}} - 3\sqrt{\frac{\bar{\hat{p}}(1-\bar{\hat{p}})}{N_i}}$$

et si la taille des lots sont tous identiques:

USL_p = imposé par client/direction

$$UCL_p = \bar{\hat{p}} + 3\sqrt{\frac{\bar{\hat{p}}(1-\bar{\hat{p}})}{N}} \quad (74.428)$$

$$CL_p = \bar{\hat{p}}$$

$$LCL_p = \bar{\hat{p}} - 3\sqrt{\frac{\bar{\hat{p}}(1-\bar{\hat{p}})}{N}}$$

Conformément à l'étude que nous ferons plus tard des plans d'échantillonnages, il ne peut pas y avoir de limite inférieure imposée par le client/direction (LSL) ni d'objectif (T) dans ce domaine d'application.

Ainsi, avec le tableau suivant où la taille des lots (cas particulier!) est identique:

	A	B	C	D	E	F	G	H	I	J
1	N° de Lot	Taille lot	Non-conformes		Proportion	CL	Sigma	LCL	UCL	USL
2	1	50	1		0.020	0.042	0.028	0	0.127	0.050
3	2	50	2		0.040	0.042	0.028	0	0.127	0.050
4	3	50	5		0.100	0.042	0.028	0	0.127	0.050
5	4	50	2		0.040	0.042	0.028	0	0.127	0.050
6	5	50	2		0.040	0.042	0.028	0	0.127	0.050
7	6	50	3		0.060	0.042	0.028	0	0.127	0.050
8	7	50	4		0.080	0.042	0.028	0	0.127	0.050
9	8	50	3		0.060	0.042	0.028	0	0.127	0.050
10	9	50	1		0.020	0.042	0.028	0	0.127	0.050
11	10	50	1		0.020	0.042	0.028	0	0.127	0.050
12	11	50	2		0.040	0.042	0.028	0	0.127	0.050
13	12	50	1		0.020	0.042	0.028	0	0.127	0.050
14	13	50	2		0.040	0.042	0.028	0	0.127	0.050
15	14	50	1		0.020	0.042	0.028	0	0.127	0.050
16	15	50	2		0.040	0.042	0.028	0	0.127	0.050
17	16	50	3		0.060	0.042	0.028	0	0.127	0.050
18	17	50	3		0.060	0.042	0.028	0	0.127	0.050
19	18	50	1		0.020	0.042	0.028	0	0.127	0.050
20	19	50	2		0.040	0.042	0.028	0	0.127	0.050
21	20	50	0		0	0.042	0.028	0	0.127	0.050
22	21	50	3		0.060	0.042	0.028	0	0.127	0.050
23	22	50	2		0.040	0.042	0.028	0	0.127	0.050

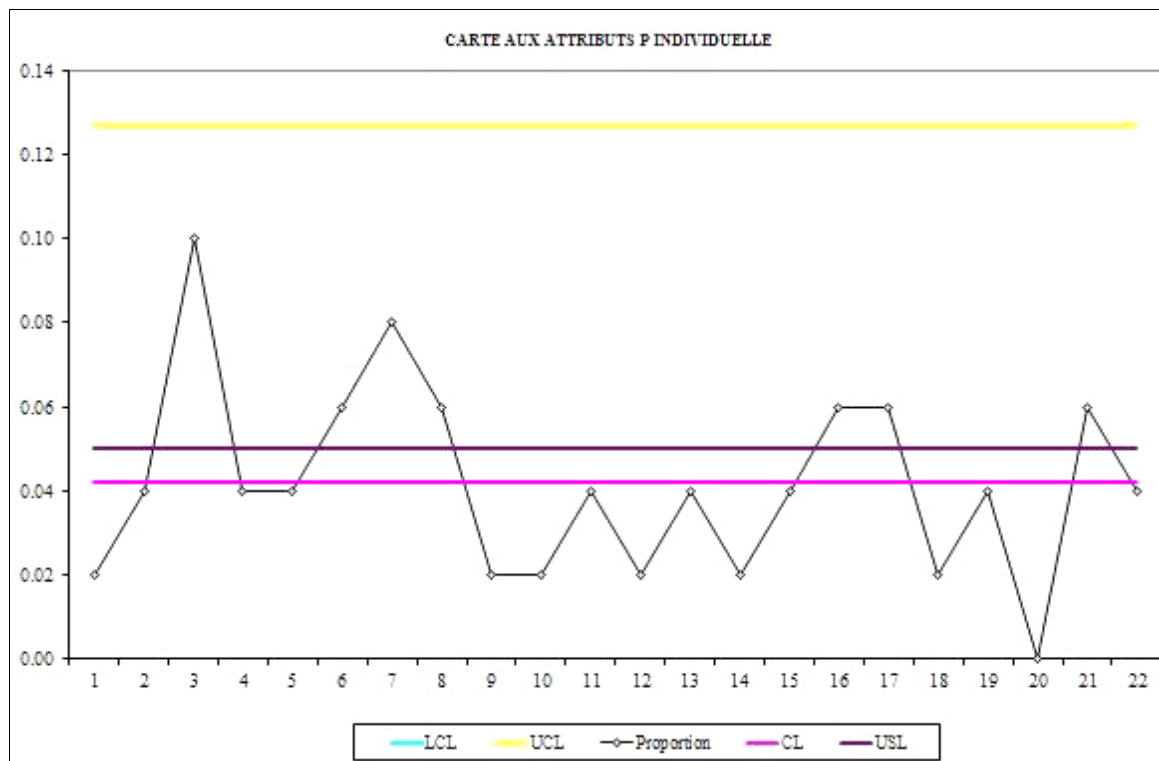
Figure: 74.52 - Tableaux Microsoft Excel 14.0.6123 avec lots de taille identique et non-conformités

Soit:

	E	F	G	H	I
1	Proportion	CL	Sigma	LCL	UCL
2	=C2/B2	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F2*(1-F2)/B2)	=SI((F2-3*G2)<0;0;F2-3*G2)	=F2+3*G2
3	=C3/B3	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F3*(1-F3)/B3)	=SI((F3-3*G3)<0;0;F3-3*G3)	=F3+3*G3
4	=C4/B4	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F4*(1-F4)/B4)	=SI((F4-3*G4)<0;0;F4-3*G4)	=F4+3*G4
5	=C5/B5	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F5*(1-F5)/B5)	=SI((F5-3*G5)<0;0;F5-3*G5)	=F5+3*G5
6	=C6/B6	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F6*(1-F6)/B6)	=SI((F6-3*G6)<0;0;F6-3*G6)	=F6+3*G6
7	=C7/B7	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F7*(1-F7)/B7)	=SI((F7-3*G7)<0;0;F7-3*G7)	=F7+3*G7
8	=C8/B8	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F8*(1-F8)/B8)	=SI((F8-3*G8)<0;0;F8-3*G8)	=F8+3*G8
9	=C9/B9	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F9*(1-F9)/B9)	=SI((F9-3*G9)<0;0;F9-3*G9)	=F9+3*G9
10	=C10/B10	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F10*(1-F10)/B10)	=SI((F10-3*G10)<0;0;F10-3*G10)	=F10+3*G10
11	=C11/B11	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F11*(1-F11)/B11)	=SI((F11-3*G11)<0;0;F11-3*G11)	=F11+3*G11
12	=C12/B12	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F12*(1-F12)/B12)	=SI((F12-3*G12)<0;0;F12-3*G12)	=F12+3*G12
13	=C13/B13	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F13*(1-F13)/B13)	=SI((F13-3*G13)<0;0;F13-3*G13)	=F13+3*G13
14	=C14/B14	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F14*(1-F14)/B14)	=SI((F14-3*G14)<0;0;F14-3*G14)	=F14+3*G14
15	=C15/B15	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F15*(1-F15)/B15)	=SI((F15-3*G15)<0;0;F15-3*G15)	=F15+3*G15
16	=C16/B16	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F16*(1-F16)/B16)	=SI((F16-3*G16)<0;0;F16-3*G16)	=F16+3*G16
17	=C17/B17	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F17*(1-F17)/B17)	=SI((F17-3*G17)<0;0;F17-3*G17)	=F17+3*G17
18	=C18/B18	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F18*(1-F18)/B18)	=SI((F18-3*G18)<0;0;F18-3*G18)	=F18+3*G18
19	=C19/B19	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F19*(1-F19)/B19)	=SI((F19-3*G19)<0;0;F19-3*G19)	=F19+3*G19
20	=C20/B20	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F20*(1-F20)/B20)	=SI((F20-3*G20)<0;0;F20-3*G20)	=F20+3*G20
21	=C21/B21	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F21*(1-F21)/B21)	=SI((F21-3*G21)<0;0;F21-3*G21)	=F21+3*G21
22	=C22/B22	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F22*(1-F22)/B22)	=SI((F22-3*G22)<0;0;F22-3*G22)	=F22+3*G22
23	=C23/B23	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F23*(1-F23)/B23)	=SI((F23-3*G23)<0;0;F23-3*G23)	=F23+3*G23

Figure: 74.53 - Formules Microsoft Excel 14.0.6123 correspondantes au tableau précédent

Ce qui donne la carte de contrôle suivante:

Figure: 74.54 - Carte de contrôle aux attributs p individuelle basée sur le tableau Microsoft Excel 14.0.6123 précédent

où nous avons pris comme exemple les limites imposées par la clientèle/direction:

$$USL_p = 5.00\% \quad (74.429)$$

Remarque: Cette carte de contrôle n'est donc pas adaptée aux tirages exhaustifs puisqu'il faudrait alors utiliser la loi hypergéométrique. Malheureusement, comme il n'existe pas de relation analytique pour l'intervalle de confiance de la proportion d'une loi hypergéométrique, il n'est pas possible (à ce jour... et à ma connaissance...) de construire une carte de contrôle de ce type.

La carte- p s'interprète comme une autre carte de Shewhart en utilisant les règles de la WECCO.

5.3.2. CC-NP

La carte- np est similaire à la carte- p précédente, mais représente le nombre d'unités non-conformes plutôt que la proportion. Cette carte est plus simple à interpréter que la carte- p puisque pas tous les employés arrivent à se représenter le concept de pourcentage, mais par contre il est très difficile de comparer en toute objectivité deux points dont la taille des lots à l'origine n'est pas la même.

Avantages: Très simple à construire car ne nécessite aucune connaissance statistique poussée pour être calculée. Pour certains employés dont le concept de pourcentage est mal maîtrisé, elle est plus simple à lire qu'une carte- np .

Problèmes: Encore une fois c'est une carte dont les limites de contrôle peuvent être asymétriques si l'application pratique est à la limite de l'acceptable lorsque les hypothèses théoriques de construction ne sont pas respectées. Une difficulté parfois rencontrée par les utilisateurs de cette carte et de déterminer la taille du lot et du niveau d'acceptation (A) ou rejet (R) équivalant à l' USL en se basant sur les développements (et normes) des plans d'échantillonnages disponibles un peu plus loin dans le présent chapitre. Il est aussi difficile avec cette carte de comparer en toute objectivité deux points dont la taille des lots à l'origine n'est pas la même.

Nous avons alors en toute généralité:

$$\begin{aligned}
 USL_{np} &= \text{imposé par client/direction} \\
 UCL_{np} &= n_i \bar{\hat{p}} + 3 \sqrt{\frac{N_i \bar{\hat{p}} N_i (1 - \bar{\hat{p}})}{N_i}} = n_i \bar{\hat{p}} + 3 \sqrt{N_i \bar{\hat{p}} (1 - \bar{\hat{p}})} \\
 CL_{np} &= n_i \bar{\hat{p}} \\
 LCL_{np} &= n_i \bar{\hat{p}} - 3 \sqrt{\frac{N_i \bar{\hat{p}} N_i (1 - \bar{\hat{p}})}{N_i}} = n_i \bar{\hat{p}} - 3 \sqrt{N_i \bar{\hat{p}} (1 - \bar{\hat{p}})}
 \end{aligned} \tag{74.430}$$

Mais puisqu'il est très difficile (contrairement à la carte- p) de comparer en toute objectivité avec cette carte deux points lorsque la taille des n'est pas identique, certains spécialistes imposent – ou recommandent – que la taille des lots soit la même tel que:

$$\begin{aligned}
 USL_{np} &= \text{imposé par client/direction} \\
 UCL_{np} &= N \bar{\hat{p}} + 3 \sqrt{N \bar{\hat{p}} (1 - \bar{\hat{p}})} \\
 CL_{np} &= N \bar{\hat{p}} \\
 LCL_{np} &= N \bar{\hat{p}} - 3 \sqrt{N \bar{\hat{p}} (1 - \bar{\hat{p}})}
 \end{aligned} \tag{74.431}$$

Conformément à l'étude que nous ferons plus tard des plans d'échantillonnages, il ne peut pas y avoir de limite inférieure imposée par le client/direction (LSL) ni d'objectif (T) dans ce domaine d'application.

Ainsi, avec le tableau suivant où la taille des lots (cas particulier!) est identique:

	A	B	C	D	E	F	G	H	I	J	K
1	N° de Lot	Taille lot	Non-conformes		Non-Conformes	P Moyen	CL	Sigma	LCL	UCL	USL
2	1	50	1		1	0.042	2.091	1.415	0	6.337	2.500
3	2	50	2		2	0.042	2.091	1.415	0	6.337	2.500
4	3	50	5		5	0.042	2.091	1.415	0	6.337	2.500
5	4	50	2		2	0.042	2.091	1.415	0	6.337	2.500
6	5	50	2		2	0.042	2.091	1.415	0	6.337	2.500
7	6	50	3		3	0.042	2.091	1.415	0	6.337	2.500
8	7	50	4		4	0.042	2.091	1.415	0	6.337	2.500
9	8	50	3		3	0.042	2.091	1.415	0	6.337	2.500
10	9	50	1		1	0.042	2.091	1.415	0	6.337	2.500
11	10	50	1		1	0.042	2.091	1.415	0	6.337	2.500
12	11	50	2		2	0.042	2.091	1.415	0	6.337	2.500
13	12	50	1		1	0.042	2.091	1.415	0	6.337	2.500
14	13	50	2		2	0.042	2.091	1.415	0	6.337	2.500
15	14	50	1		1	0.042	2.091	1.415	0	6.337	2.500
16	15	50	2		2	0.042	2.091	1.415	0	6.337	2.500
17	16	50	3		3	0.042	2.091	1.415	0	6.337	2.500
18	17	50	3		3	0.042	2.091	1.415	0	6.337	2.500
19	18	50	1		1	0.042	2.091	1.415	0	6.337	2.500
20	19	50	2		2	0.042	2.091	1.415	0	6.337	2.500
21	20	50	0		-	0.042	2.091	1.415	0	6.337	2.500
22	21	50	3		3	0.042	2.091	1.415	0	6.337	2.500
23	22	50	2		2	0.042	2.091	1.415	0	6.337	2.500

Figure: 74.55 - Tableaux Microsoft Excel 14.0.6123 avec lots de taille identique et non-conformités

Soit:

	E	F	G	H	I	J	K
1	Non-Conformes	P Moyen	CL	Sigma	LCL	UCL	USL
2	=C2	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B2*F2	=RACINE(B2*F2*(1-F2))	=S((G2-3*H2)<0;0;G2-3*H2)	=G2+3*H2	=B2*0.05
3	=C3	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B3*F3	=RACINE(B3*F3*(1-F3))	=S((G3-3*H3)<0;0;G3-3*H3)	=G3+3*H3	=B3*0.05
4	=C4	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B4*F4	=RACINE(B4*F4*(1-F4))	=S((G4-3*H4)<0;0;G4-3*H4)	=G4+3*H4	=B4*0.05
5	=C5	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B5*F5	=RACINE(B5*F5*(1-F5))	=S((G5-3*H5)<0;0;G5-3*H5)	=G5+3*H5	=B5*0.05
6	=C6	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B6*F6	=RACINE(B6*F6*(1-F6))	=S((G6-3*H6)<0;0;G6-3*H6)	=G6+3*H6	=B6*0.05
7	=C7	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B7*F7	=RACINE(B7*F7*(1-F7))	=S((G7-3*H7)<0;0;G7-3*H7)	=G7+3*H7	=B7*0.05
8	=C8	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B8*F8	=RACINE(B8*F8*(1-F8))	=S((G8-3*H8)<0;0;G8-3*H8)	=G8+3*H8	=B8*0.05
9	=C9	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B9*F9	=RACINE(B9*F9*(1-F9))	=S((G9-3*H9)<0;0;G9-3*H9)	=G9+3*H9	=B9*0.05
10	=C10	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B10*F10	=RACINE(B10*F10*(1-F10))	=S((G10-3*H10)<0;0;G10-3*H10)	=G10+3*H10	=B10*0.05
11	=C11	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B11*F11	=RACINE(B11*F11*(1-F11))	=S((G11-3*H11)<0;0;G11-3*H11)	=G11+3*H11	=B11*0.05
12	=C12	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B12*F12	=RACINE(B12*F12*(1-F12))	=S((G12-3*H12)<0;0;G12-3*H12)	=G12+3*H12	=B12*0.05
13	=C13	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B13*F13	=RACINE(B13*F13*(1-F13))	=S((G13-3*H13)<0;0;G13-3*H13)	=G13+3*H13	=B13*0.05
14	=C14	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B14*F14	=RACINE(B14*F14*(1-F14))	=S((G14-3*H14)<0;0;G14-3*H14)	=G14+3*H14	=B14*0.05
15	=C15	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B15*F15	=RACINE(B15*F15*(1-F15))	=S((G15-3*H15)<0;0;G15-3*H15)	=G15+3*H15	=B15*0.05
16	=C16	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B16*F16	=RACINE(B16*F16*(1-F16))	=S((G16-3*H16)<0;0;G16-3*H16)	=G16+3*H16	=B16*0.05
17	=C17	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B17*F17	=RACINE(B17*F17*(1-F17))	=S((G17-3*H17)<0;0;G17-3*H17)	=G17+3*H17	=B17*0.05
18	=C18	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B18*F18	=RACINE(B18*F18*(1-F18))	=S((G18-3*H18)<0;0;G18-3*H18)	=G18+3*H18	=B18*0.05
19	=C19	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B19*F19	=RACINE(B19*F19*(1-F19))	=S((G19-3*H19)<0;0;G19-3*H19)	=G19+3*H19	=B19*0.05
20	=C20	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B20*F20	=RACINE(B20*F20*(1-F20))	=S((G20-3*H20)<0;0;G20-3*H20)	=G20+3*H20	=B20*0.05
21	=C21	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B21*F21	=RACINE(B21*F21*(1-F21))	=S((G21-3*H21)<0;0;G21-3*H21)	=G21+3*H21	=B21*0.05
22	=C22	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B22*F22	=RACINE(B22*F22*(1-F22))	=S((G22-3*H22)<0;0;G22-3*H22)	=G22+3*H22	=B22*0.05
23	=C23	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=B23*F23	=RACINE(B23*F23*(1-F23))	=S((G23-3*H23)<0;0;G23-3*H23)	=G23+3*H23	=B23*0.05

Figure: 74.56 - Formules Microsoft Excel 14.0.6123 correspondantes au tableau précédent

Ce qui donne la carte de contrôle suivante:

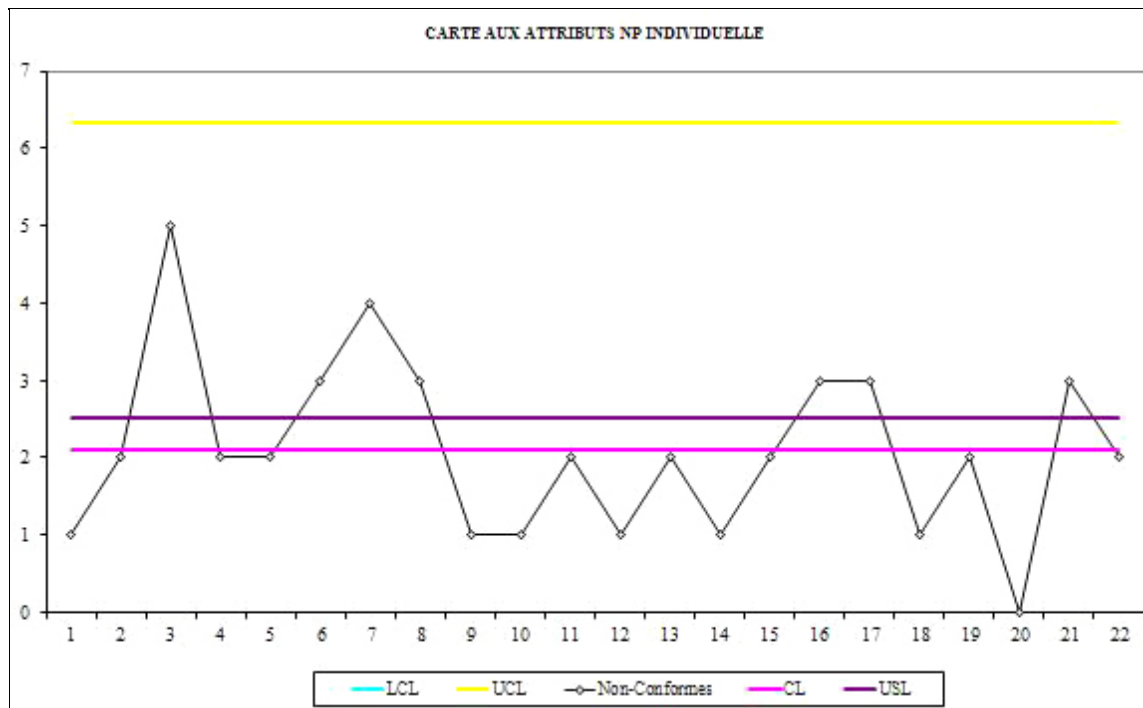


Figure: 74.57 - Carte de contrôle aux attributs np individuelle basée sur le tableau Microsoft Excel 14.0.6123 précédent

où nous avons pris comme exemple les limites imposées par la clientèle/direction:

$$USL_{np} = 5.00\% \cdot N = 5.00\% \cdot 50 = 2.5 \quad (74.432)$$

Les mêmes remarques que pour la carte- p s'appliquent pour la carte- np .

5.3.3. CC-C

Dans le chapitre de Statistiques, nous avons démontré que lorsque la probabilité p est très faible et tend vers zéro, mais que toutefois la valeur moyenne $n \cdot p$ tend vers une valeur fixe si n tend vers l'infini, la loi binomiale de moyenne $\mu = n \cdot p$ avec k épreuves était alors donnée par une loi de Poisson:

$$P(X = k) = \frac{\mu^k}{k!} e^{-\mu} \quad (74.433)$$

avec:

$$\begin{aligned} \mu &= \mu \\ \sigma &= \sqrt{\mu} \end{aligned} \quad (74.434)$$

En pratique, certains remplacent la loi binomiale par une loi de Poisson dès que $n > 30$ et $np < 5$ ou lorsque $n > 50$ et $p < 0.1$... Dans la pratique je recommande personnellement l'usage de l'approximation lorsque $n > 1000$ et $p < 0.01$.

Soit

$$\begin{aligned}
 USL_c &= \text{imposé par client/direction} \\
 UCL_c &= \mu + 3\sqrt{\mu} \\
 CL_c &= \mu \\
 LCL_c &= \mu - 3\sqrt{\mu}
 \end{aligned}
 \tag{74.435}$$

ce qui se note traditionnellement dans le domaine:

$$\begin{aligned}
 USL_c &= \text{imposé par client/direction} \\
 UCL_c &= \bar{c} + 3\sqrt{\bar{c}} \\
 CL_c &= \bar{c} \\
 LCL_c &= \bar{c} - 3\sqrt{\bar{c}}
 \end{aligned}
 \tag{74.436}$$

où $\bar{c}$ est simplement la moyenne du comptage (c'est de ce mot que provient le "c" du nom de la carte) de non-conformités:

$$\bar{c} = \frac{1}{n} \sum_{i=1}^n c_i \tag{74.437}$$

Conformément à l'étude que nous ferons plus tard des plans d'échantillonnages, il ne peut pas y avoir de limite inférieure imposée par le client/direction (LSL) ni d'objectif (T) dans ce domaine d'application.

Avantages: Très simple à construire car ne nécessite aucune connaissance statistique poussée pour être calculée.

Problèmes: Encore une fois c'est une carte dont les limites de contrôle peuvent être asymétriques si l'application pratique est à la limite de l'acceptable lorsque les hypothèses théoriques de construction ne sont pas respectées. Une difficulté parfois rencontrée par les utilisateurs de cette carte est de déterminer la taille du lot et du niveau d'acceptation (A) ou rejet (R) équivalant à l' USL en se basant sur les développements (et normes) des plans d'échantillonnages disponibles un peu plus loin dans le présent chapitre. Enfin, une dernière difficulté est que certains employés maîtrisent mal le concept de pourcentages.

Ainsi, avec le tableau suivant où la taille des lots (cas particulier!) est identique (les non-conformes ne peuvent pas être exprimés en % de la taille du lot puisque la loi de Poisson est une loi discrète!):

	A	B	C	D	E	F	G	H	I
1	N° de Lot	Non-conformes		Non-Conformes	CL	Sigma	LCL	UCL	USL
2	1	1		1	2.031	1.446	0	6.429	2.500
3	2	2		2	2.031	1.446	0	6.429	2.500
4	3	5		5	2.031	1.446	0	6.429	2.500
5	4	2		2	2.031	1.446	0	6.429	2.500
6	5	2		2	2.031	1.446	0	6.429	2.500
7	6	3		3	2.031	1.446	0	6.429	2.500
8	7	4		4	2.031	1.446	0	6.429	2.500
9	8	3		3	2.031	1.446	0	6.429	2.500
10	9	1		1	2.031	1.446	0	6.429	2.500
11	10	1		1	2.031	1.446	0	6.429	2.500
12	11	2		2	2.031	1.446	0	6.429	2.500
13	12	1		1	2.031	1.446	0	6.429	2.500
14	13	2		2	2.031	1.446	0	6.429	2.500
15	14	1		1	2.031	1.446	0	6.429	2.500
16	15	2		2	2.031	1.446	0	6.429	2.500
17	16	3		3	2.031	1.446	0	6.429	2.500
18	17	3		3	2.031	1.446	0	6.429	2.500
19	18	1		1	2.031	1.446	0	6.429	2.500
20	19	2		2	2.031	1.446	0	6.429	2.500
21	20	0		-	2.031	1.446	0	6.429	2.500
22	21	3		3	2.031	1.446	0	6.429	2.500
23	22	2		2	2.031	1.446	0	6.429	2.500

Figure: 74.58 - Tableaux Microsoft Excel 14.0.6123 avec lots de taille identique et non-conformités

Soit:

	D	E	F	G	H
1	Non-Conformes	CL	Sigma	LCL	UCL
2	=B2	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E2)	=SI((E2-3*F2)<0;0;E2-3*F2)	=E2+3*F2
3	=B3	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E3)	=SI((E3-3*F3)<0;0;E3-3*F3)	=E3+3*F3
4	=B4	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E4)	=SI((E4-3*F4)<0;0;E4-3*F4)	=E4+3*F4
5	=B5	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E5)	=SI((E5-3*F5)<0;0;E5-3*F5)	=E5+3*F5
6	=B6	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E6)	=SI((E6-3*F6)<0;0;E6-3*F6)	=E6+3*F6
7	=B7	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E7)	=SI((E7-3*F7)<0;0;E7-3*F7)	=E7+3*F7
8	=B8	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E8)	=SI((E8-3*F8)<0;0;E8-3*F8)	=E8+3*F8
9	=B9	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E9)	=SI((E9-3*F9)<0;0;E9-3*F9)	=E9+3*F9
10	=B10	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E10)	=SI((E10-3*F10)<0;0;E10-3*F10)	=E10+3*F10
11	=B11	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E11)	=SI((E11-3*F11)<0;0;E11-3*F11)	=E11+3*F11
12	=B12	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E12)	=SI((E12-3*F12)<0;0;E12-3*F12)	=E12+3*F12
13	=B13	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E13)	=SI((E13-3*F13)<0;0;E13-3*F13)	=E13+3*F13
14	=B14	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E14)	=SI((E14-3*F14)<0;0;E14-3*F14)	=E14+3*F14
15	=B15	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E15)	=SI((E15-3*F15)<0;0;E15-3*F15)	=E15+3*F15
16	=B16	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E16)	=SI((E16-3*F16)<0;0;E16-3*F16)	=E16+3*F16
17	=B17	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E17)	=SI((E17-3*F17)<0;0;E17-3*F17)	=E17+3*F17
18	=B18	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E18)	=SI((E18-3*F18)<0;0;E18-3*F18)	=E18+3*F18
19	=B19	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E19)	=SI((E19-3*F19)<0;0;E19-3*F19)	=E19+3*F19
20	=B20	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E20)	=SI((E20-3*F20)<0;0;E20-3*F20)	=E20+3*F20
21	=B21	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E21)	=SI((E21-3*F21)<0;0;E21-3*F21)	=E21+3*F21
22	=B22	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E22)	=SI((E22-3*F22)<0;0;E22-3*F22)	=E22+3*F22
23	=B23	=MOYENNE(\$B\$2:\$B\$23)	=RACINE(E23)	=SI((E23-3*F23)<0;0;E23-3*F23)	=E23+3*F23

Figure: 74.59 - Formules Microsoft Excel 14.0.6123 correspondantes au tableau précédent

Ce qui donne la carte de contrôle suivante:

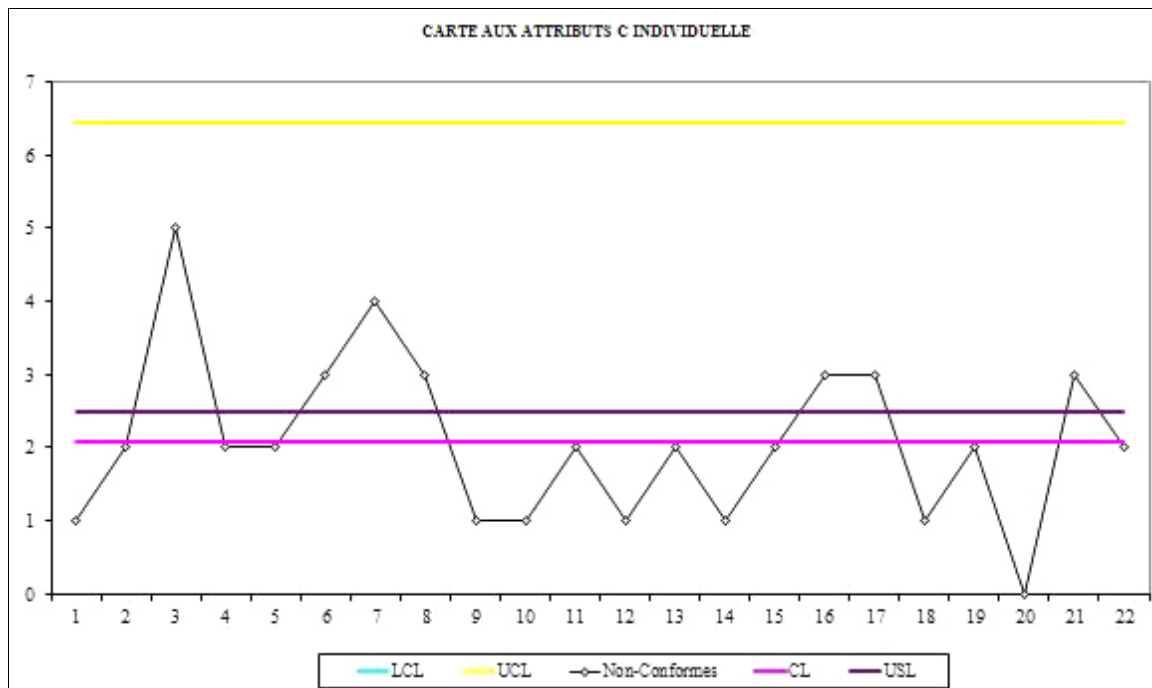


Figure: 74.60 - Carte de contrôle aux attributs C individuelle basée sur le tableau Microsoft Excel 14.0.6123 précédent

où nous avons pris comme exemple les limites imposées par la clientèle/direction:

$$USL_c = 2.5 \quad (74.438)$$

La carte-c s'interprète donc comme une autre carte de Shewhart en utilisant les règles de la WECO.

Remarque: Cette carte s'applique très bien lorsque chaque point représente le nombre de non-conformités de matériel assemblé de nombreux composants (plus d'une dizaine de milliers souvent) dont la probabilité d'être défectueux est très faible.

5.3.4. CC-U

La carte-*u* est similaire à la carte *c* à la différence qu'elle a pour rôle de normaliser les données à l'unité afin d'avoir une carte-*c* basée sur une loi de Poisson avec des pourcentages.

Pour cela, nous utilisons les mêmes hypothèses de travail que celles démontrées dans le chapitre de Statistiques concernant la loi de Poisson:

$$\begin{aligned} \mu &= \mu = N\hat{p} \\ \sigma &= \sqrt{\mu} = \sqrt{N\hat{p}} \end{aligned} \quad (74.439)$$

Ce qui fait que la carte-*c* s'écrit alors:

$$\begin{aligned} USL_u &= \text{imposé par client/direction} \\ UCL_u &= N_i \hat{\bar{p}} + 3\sqrt{N_i \hat{\bar{p}}} \\ CL_u &= N_i \hat{\bar{p}} \\ LCL_u &= N_i \hat{\bar{p}} - 3\sqrt{N_i \hat{\bar{p}}} \end{aligned} \quad (74.440)$$

Si nous divisons la variable aléatoire k par la taille du lot (ou nombre de composants) nous avons alors de par la propriété de l'espérance et de la variance:

$$\begin{aligned}
 USL_u &= \text{imposé par client/direction} \\
 UCL_u &= \bar{\hat{p}} + 3\sqrt{\frac{\hat{p}}{N_i}} \\
 CL_u &= \bar{\hat{p}} \\
 LCL_u &= \bar{\hat{p}} - 3\sqrt{\frac{\hat{p}}{N_i}}
 \end{aligned} \tag{74.441}$$

ce qui se note traditionnellement:

$$\begin{aligned}
 USL_u &= \text{imposé par client/direction} \\
 UCL_u &= \bar{u} + 3\sqrt{\frac{\bar{u}}{N_i}} \\
 CL_u &= \bar{u} \\
 LCL_u &= \bar{u} - 3\sqrt{\frac{\bar{u}}{N_i}}
 \end{aligned} \tag{74.442}$$

Avec donc:

$$\bar{u} = CL_u = \frac{\sum_{i=1}^n c_i}{\sum_{i=1}^n N_i} \tag{74.443}$$

Ainsi, avec le tableau suivant dans le cas particulier (mais courant) où tous les lots ont la même taille ou le même nombre de composants:

	A	B	C	D	E	F	G	H	I	J
1	N° de Lot	Taille Lot	Non-conformes		Non-Conformes	CL	Sigma	LCL	UCL	USL
2	1	1000	10		0.010	0.0203	0.0046	0.0072	0.0346	0.010
3	2	1000	20		0.020	0.0203	0.0046	0.0072	0.0346	0.010
4	3	1000	50		0.050	0.0203	0.0046	0.0072	0.0346	0.010
5	4	1000	20		0.020	0.0203	0.0046	0.0072	0.0346	0.010
6	5	1000	20		0.020	0.0203	0.0046	0.0072	0.0346	0.010
7	6	1000	30		0.030	0.0203	0.0046	0.0072	0.0346	0.010
8	7	1000	40		0.040	0.0203	0.0046	0.0072	0.0346	0.010
9	8	1000	30		0.030	0.0203	0.0046	0.0072	0.0346	0.010
10	9	1000	10		0.010	0.0203	0.0046	0.0072	0.0346	0.010
11	10	1000	10		0.010	0.0203	0.0046	0.0072	0.0346	0.010
12	11	1000	20		0.020	0.0203	0.0046	0.0072	0.0346	0.010
13	12	1000	10		0.010	0.0203	0.0046	0.0072	0.0346	0.010
14	13	1000	20		0.020	0.0203	0.0046	0.0072	0.0346	0.010
15	14	1000	10		0.010	0.0203	0.0046	0.0072	0.0346	0.010
16	15	1000	20		0.020	0.0203	0.0046	0.0072	0.0346	0.010
17	16	1000	30		0.030	0.0203	0.0046	0.0072	0.0346	0.010
18	17	1000	30		0.030	0.0203	0.0046	0.0072	0.0346	0.010
19	18	1000	10		0.010	0.0203	0.0046	0.0072	0.0346	0.010
20	19	1000	20		0.020	0.0203	0.0046	0.0072	0.0346	0.010
21	20	1000	0		-	0.0203	0.0046	0.0072	0.0346	0.010
22	21	1000	30		0.030	0.0203	0.0046	0.0072	0.0346	0.010
23	22	1000	20		0.020	0.0203	0.0046	0.0072	0.0346	0.010

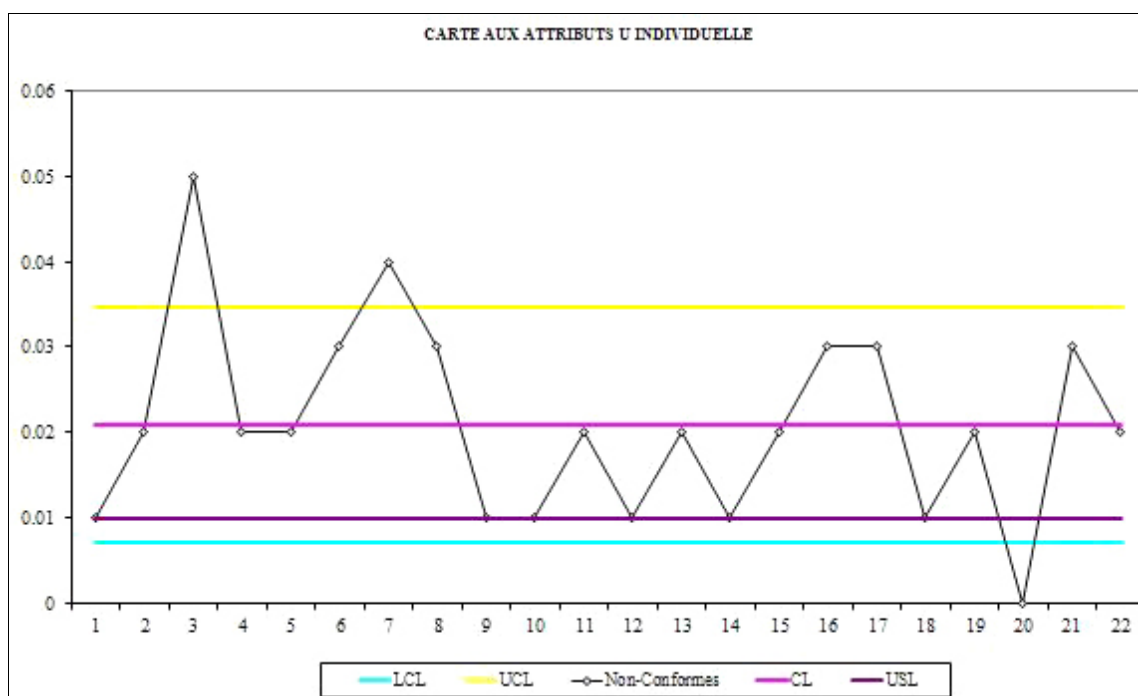
Figure: 74.61 - Tableaux Microsoft Excel 14.0.6123 avec lots de taille identique et non-conformités

Soit:

	E	F	G	H	I
1	Non-Conformes	CL	Sigma	LCL	UCL
2	=C2/B2	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F2/B2)	=SI((F2-3*G2)<0;0;F2-3*G2)	=F2+3*G2
3	=C3/B3	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F3/B3)	=SI((F3-3*G3)<0;0;F3-3*G3)	=F3+3*G3
4	=C4/B4	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F4/B4)	=SI((F4-3*G4)<0;0;F4-3*G4)	=F4+3*G4
5	=C5/B5	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F5/B5)	=SI((F5-3*G5)<0;0;F5-3*G5)	=F5+3*G5
6	=C6/B6	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F6/B6)	=SI((F6-3*G6)<0;0;F6-3*G6)	=F6+3*G6
7	=C7/B7	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F7/B7)	=SI((F7-3*G7)<0;0;F7-3*G7)	=F7+3*G7
8	=C8/B8	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F8/B8)	=SI((F8-3*G8)<0;0;F8-3*G8)	=F8+3*G8
9	=C9/B9	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F9/B9)	=SI((F9-3*G9)<0;0;F9-3*G9)	=F9+3*G9
10	=C10/B10	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F10/B10)	=SI((F10-3*G10)<0;0;F10-3*G10)	=F10+3*G10
11	=C11/B11	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F11/B11)	=SI((F11-3*G11)<0;0;F11-3*G11)	=F11+3*G11
12	=C12/B12	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F12/B12)	=SI((F12-3*G12)<0;0;F12-3*G12)	=F12+3*G12
13	=C13/B13	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F13/B13)	=SI((F13-3*G13)<0;0;F13-3*G13)	=F13+3*G13
14	=C14/B14	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F14/B14)	=SI((F14-3*G14)<0;0;F14-3*G14)	=F14+3*G14
15	=C15/B15	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F15/B15)	=SI((F15-3*G15)<0;0;F15-3*G15)	=F15+3*G15
16	=C16/B16	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F16/B16)	=SI((F16-3*G16)<0;0;F16-3*G16)	=F16+3*G16
17	=C17/B17	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F17/B17)	=SI((F17-3*G17)<0;0;F17-3*G17)	=F17+3*G17
18	=C18/B18	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F18/B18)	=SI((F18-3*G18)<0;0;F18-3*G18)	=F18+3*G18
19	=C19/B19	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F19/B19)	=SI((F19-3*G19)<0;0;F19-3*G19)	=F19+3*G19
20	=C20/B20	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F20/B20)	=SI((F20-3*G20)<0;0;F20-3*G20)	=F20+3*G20
21	=C21/B21	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F21/B21)	=SI((F21-3*G21)<0;0;F21-3*G21)	=F21+3*G21
22	=C22/B22	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F22/B22)	=SI((F22-3*G22)<0;0;F22-3*G22)	=F22+3*G22
23	=C23/B23	=SOMME(\$C\$2:\$C\$23)/SOMME(\$B\$2:\$B\$23)	=RACINE(F23/B23)	=SI((F23-3*G23)<0;0;F23-3*G23)	=F23+3*G23

Figure: 74.62 - Formules Microsoft Excel 14.0.6123 correspondantes au tableau précédent

Ce qui donne la carte de contrôle suivante:

Figure: 74.63 - Carte de contrôle aux attributs U individuelle basée sur le tableau Microsoft Excel 14.0.6123 précédent

où nous avons pris comme exemple les limites imposées par la clientèle/direction:

$$USL = 2.5 \quad (74.444)$$

La carte-c s'interprète comme une autre carte de Shewhart en utilisant les règles de la WECO.

Remarque: Cette carte s'applique très bien lorsque chaque point représente le nombre de non-conformités de matériel assemblé de nombreux composants (plus d'une dizaine de milliers souvent) dont la probabilité d'être défectueux est très faible.

5.4. CARTES DE CONTRÔLES QUANTITATIVES (AUX MESURES)

Les variables quantitatives sont donc des mesures continues de type poids, longueur, épaisseur, température, diamètre...

Contrairement aux cartes de contrôle autocorrélées et aux cartes de contrôle qualitatives (aux attributs), les cartes de contrôle quantitatives (aux mesures) sont souvent - pour ne pas dire toujours - présentées par paire:

- Une carte de contrôle représentant l'analyse/évolution de la tendance statistique centrale (souvent l'espérance, car il est relativement facile d'en construire un intervalle de confiance).
- Une carte de contrôle représentant l'analyse/évolution de la dispersion via l'étendue ou l'écart-type.

Les limites de contrôles sont souvent prises (faute de temps de réflexion dans les entreprises) comme des limites de Shewhart et non des limites probabilistes.

5.4.1. CC A VALEURS INDIVIDUELLES AVEC LIMITES IMPOSÉES

C'est la carte de contrôle la plus simple et la plus utilisée car ne nécessite aucune connaissance et hypothèse statistique particulière. Elle consiste uniquement à reporter les mesures effectuées et à apporter des actions correctives quand cela semble nécessaire.

Avantages: Simple à construire car ne nécessite aucune connaissance ou hypothèse statistique particulière. Permet de vérifier de manière ponctuelle si une production satisfait les contraintes spécifiées par les clients.

Problèmes: Elle peut représenter des déviations aléatoires temporaires qui pourront faire penser à tort à un problème du procédé/processus et amèneront à des actions correctives inutiles (fausses alarmes). Elle ne permet pas non plus d'identifier si le procédé/processus est sous contrôle statistique ou pas.

Considérons par exemple le tableau suivant:

	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	→									N°	Mesures	T	USL	LSL
2	↓	3.3343	10.0136	3.3732	10.0088	3.3685	3.3544		1	3.3343	10.0000	10.1000	3.3000	
3		3.3721	3.3643	3.3426	3.3712	10.0259	10.0177		2	10.0136	10.0000	10.1000	3.3000	
4		3.3788	10.0213	3.3407	10.0696	10.0161	10.0709		3	3.3732	10.0000	10.1000	3.3000	
5		10.0658	10.08	3.3514	3.3358	10.0338	10.0422		4	10.0088	10.0000	10.1000	3.3000	
6		10.0136	3.3731	3.3443	3.3403	3.3424	3.3421		5	3.3685	10.0000	10.1000	3.3000	
7		3.3435	10.0057	3.3715	10.0388	10.0019	3.3616		6	3.3544	10.0000	10.1000	3.3000	
8		3.3144	10.0321	10.0361	10.0253	10.0109	3.3335		7	3.3721	10.0000	10.1000	3.3000	
9		3.3748	3.3362	10.0363	10.15	3.3336	3.3741		8	3.3643	10.0000	10.1000	3.3000	
10		10.0063	10.0749	3.3371	3.378	3.3334	3.3555		9	3.3426	10.0000	10.1000	3.3000	
11		10.084	3.3619	10.0063	10.0364	3.3307	3.3762		10	3.3712	10.0000	10.1000	3.3000	
12		3.3436	10.0638	3.3617	10.0138	3.3672	3.3531		11	10.0253	10.0000	10.1000	3.3000	
13		10.0074	3.3817	3.3753	10.0541	3.3845	3.3107		12	10.0177	10.0000	10.1000	3.3000	
14		10.0063	10.0282	10.0385	3.3747	10.0323	3.386		13	3.3788	10.0000	10.1000	3.3000	
15		3.3836	10.063	3.3824	3.3303	3.3684	10.0575		14	10.0213	10.0000	10.1000	3.3000	
16		10.0317	10.0485	3.3659	10.0675	10.0047	10.0277		15	3.3407	10.0000	10.1000	3.3000	
17		3.3756	3.3174	3.3076	3.3303	10.018	3.3379		16	10.0696	10.0000	10.1000	3.3000	
18		3.3818	3.3362	3.3614	3.3612	3.3634	3.3349		17	10.0161	10.0000	10.1000	3.3000	
19		10.0168	10.0043	10.0029	3.3543	10.0139	3.3631		18	10.0709	10.0000	10.1000	3.3000	
20		3.3857	10.0503	10.0381	3.3811	3.3603	3.3123		19	10.0658	10.0000	10.1000	3.3000	
21		3.3878	3.3841	3.3436	3.3754	3.3532	10.0576		20	10.0800	10.0000	10.1000	3.3000	
22		10.0657	3.3838	3.3636	3.3779	10.0423	10.0547		21	3.3514	10.0000	10.1000	3.3000	
23		3.3564	10.0184	3.383	3.3333	10.0146	10.0145		22	3.3358	10.0000	10.1000	3.3000	
24		10.0611	3.3443	3.3839	3.3833	3.3347	10.0033		23	10.0338	10.0000	10.1000	3.3000	
25		3.3054	3.3043	3.387	10.058	3.3156	3.3782		24	10.0422	10.0000	10.1000	3.3000	
26		3.3768	3.3354	3.3136	3.3341	3.3162	10.0721		25	10.0136	10.0000	10.1000	3.3000	
27									26	3.3731	10.0000	10.1000	3.3000	
28									27	3.3443	10.0000	10.1000	3.3000	
29									28	3.3403	10.0000	10.1000	3.3000	
30									29	3.3424	10.0000	10.1000	3.3000	
31									30	3.3421	10.0000	10.1000	3.3000	
32									31	3.3435	10.0000	10.1000	3.3000	
33									32	10.0057	10.0000	10.1000	3.3000	
34									33	3.3715	10.0000	10.1000	3.3000	
35									34	10.0388	10.0000	10.1000	3.3000	
36									35	10.0019	10.0000	10.1000	3.3000	
37									36	3.3616	10.0000	10.1000	3.3000	
38									37	3.3144	10.0000	10.1000	3.3000	
39									38	10.0321	10.0000	10.1000	3.3000	
40									39	10.0361	10.0000	10.1000	3.3000	
41									40	10.0253	10.0000	10.1000	3.3000	
42									41	10.0109	10.0000	10.1000	3.3000	
43									42	3.3335	10.0000	10.1000	3.3000	
44									43	3.3748	10.0000	10.1000	3.3000	
45									44	3.3362	10.0000	10.1000	3.3000	
46									45	10.0363	10.0000	10.1000	3.3000	
47									46	10.1500	10.0000	10.1000	3.3000	

Figure: 74.64 - Tableaux Microsoft Excel 14.0.6123 avec mesures individuelles

où nous avons pris en ce qui concerne les limites (nous supposons qu'elles imposent un écart-type de 0.10):

$$USL = 10 + 3\sigma = 10 + 3 \cdot 0.0\bar{3} = 10.10$$

$$T = 10 \quad (74.445)$$

$$LSL = 10 - 3\sigma = 10 - 3 \cdot 0.0\bar{3} = 9.90$$

Cela qui nous donne la carte de contrôle suivante:

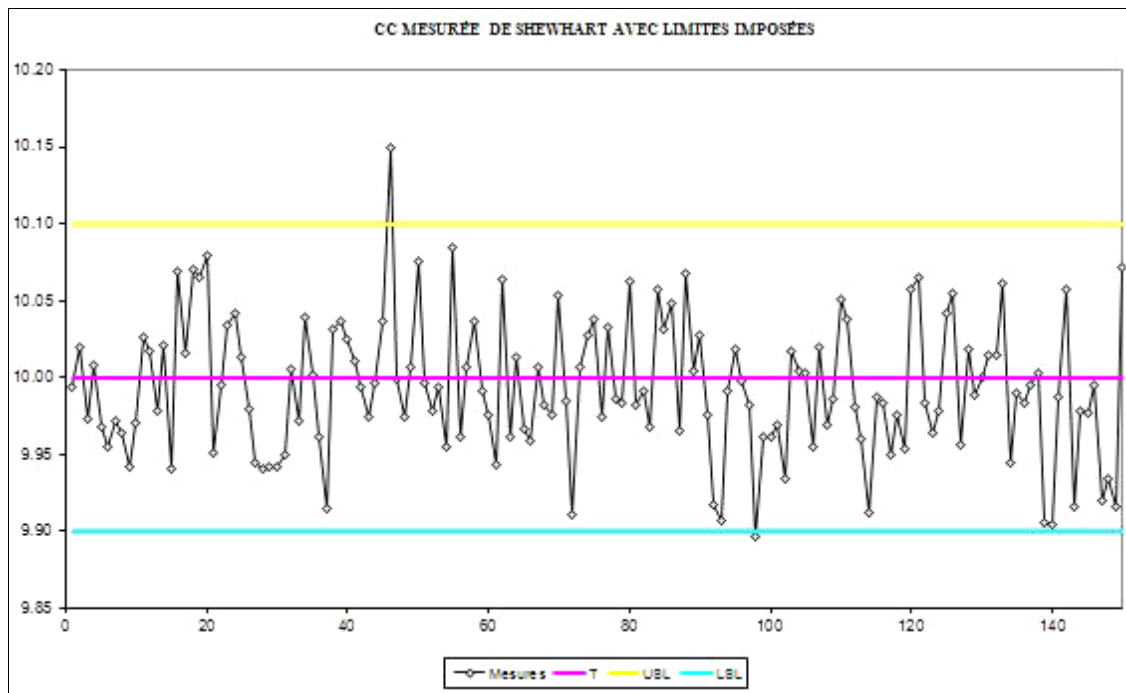


Figure: 74.65 - Carte de contrôle mesurée de Shewhart avec limites imposées

Bien que cette carte soit la plus utilisée dans la pratique (car les ingénieurs ont des directives de qualité avec des tolérances bien précises à respecter), son problème reste cependant qu'elle ne nous indique pas vraiment si le procédé ou processus semble se stabiliser dans le temps (d'ailleurs ce n'est pas son rôle!) et si le contrôle statistique est dans les limites du client (elle nous montre seulement les statistiques ponctuelles ce qui est peu robuste!). Nous allons partiellement y remédier avec la prochaine carte.

5.4.2. CC À VALEURS INDIVIDUELLES AVEC MOYENNE ET ÉCART-TYPE COURT TERME

Lorsque nous commençons les premières observations (mesures) le procédé ou processus est rarement stable et il est hors de question de s'amuser à faire des quantités de pièces d'essais juste pour faire une carte de contrôle de la moyenne dans ces conditions.

Avantage: Met en évidence l'évolution dans le temps des indicateurs de statistiques de positions et de dispersion et permet d'observer la stabilisation du procédé/processus pendant la phase de prototypage (souvent nommée "Phase I" dans la littérature spécialisée).

Problèmes: Ils sont les mêmes que pour la carte de contrôle précédente. Il faut y rajouter le fait que nous supposons les données identiquement distribuées et indépendantes selon une loi de probabilité symétrique, ce qui n'était pas le cas précédemment.

Nous restons donc avec une carte de contrôle mesurée à laquelle nous ajoutons la moyenne arithmétique et l'écart-type ponctuels recalculés à chaque nouvel élément. En d'autres termes, comme nous n'avons aussi pas assez d'échantillons pour connaître la loi de probabilité et qu'il est hors de question de faire des déchets pour cela et de former les gens pendant des semaines aux statistiques, nous rajoutons à la carte de contrôle précédente les indicateurs de limite de contrôle de Shewhart fixes suivants:

USL = imposé par client/direction

$$UCL_{\sigma_{ct}} = \bar{X} + 3\sigma_X$$

$$CL_{\sigma_{ct}} = E(X) = \bar{X}$$

T = objectif imposé par client/direction (74.446)

$$LCL_{\sigma_{ct}} = \bar{X} - 3\sigma_X$$

LSL = imposé par client/direction

où CL signifie "**Center Line**" et avec l'estimateur de l'espérance et l'estimateur non biaisé de l'écart-type (définitions générales de la moyenne et de l'écart-type vues dans le chapitre de Statistiques):

$$E(X) = \bar{X} = \frac{1}{k} \sum_{i=1}^k X_i$$

$$\sigma_X = \sqrt{\frac{1}{k-1} \sum_{i=1}^k (X_i - \bar{X})^2} \quad (74.447)$$

sans y associer aucun intervalle de confiance (puisque nous n'avons justement pas assez d'échantillons ou que nous ne souhaitons pas demander aux employés de passer du temps à déterminer la loi de probabilité suivie). D'où le terme de "carte de contrôle avec moyenne et écart-type de la moyenne ponctuelle". Cette terminologie permet d'insister sur le fait que les indicateurs seront indiqués, mais sans en connaître l'erreur par manque d'échantillons ou simplement parce qu'il est irréaliste de demander aux employés concernés d'identifier la loi de probabilité y relative et de faire le calcul d'intervalle de confiance correspondant.

Remarquons que nous pouvons alors calculer des probabilités de pièces non-conformes par excès ou par défaut en utilisant le fait que X suit une loi Normale de paramètres:

$$N(\mu, \sigma^2) \cong N(\bar{X}, \sigma_X^2) \quad (74.448)$$

Ainsi, en reprenant le tableau suivant ayant 150 (k) mesures:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	→								N	Mesures	T	CL	UCL	LCI	USL	LSL
2	↓	9.994	10.020	9.973	10.009	9.969	9.954		1	9.9943	10.0000	9.9943	9.9943	9.9943	10.1000	9.9000
3		9.972	9.964	9.943	9.971	10.026	10.018		2	10.0196	10.0000	10.0070	10.0449	9.9690	10.1000	9.9000
4		9.979	10.021	9.941	10.070	10.016	10.071		3	9.9732	10.0000	9.9957	10.0526	9.9388	10.1000	9.9000
5		10.066	10.080	9.951	9.996	10.034	10.042		4	10.0088	10.0000	9.9990	10.0511	9.9468	10.1000	9.9000
6		10.014	9.979	9.945	9.940	9.942	9.942		5	9.9685	10.0000	9.9929	10.0521	9.9336	10.1000	9.9000
7		9.950	10.006	9.972	10.039	10.002	9.962		6	9.9544	10.0000	9.9865	10.0556	9.9173	10.1000	9.9000
8		9.914	10.032	10.036	10.025	10.011	9.994		7	9.9721	10.0000	9.9844	10.0502	9.9187	10.1000	9.9000
9		9.975	9.996	10.037	10.150	10.000	9.974		8	9.9643	10.0000	9.9819	10.0466	9.9172	10.1000	9.9000
10		10.006	10.075	9.997	9.978	9.993	9.956		9	9.9426	10.0000	9.9775	10.0489	9.9062	10.1000	9.9000
11		10.084	9.962	10.006	10.036	9.991	9.976		10	9.9712	10.0000	9.9769	10.0448	9.9090	10.1000	9.9000
12		9.944	10.064	9.962	10.014	9.967	9.959		11	10.0253	10.0000	9.9814	10.0587	9.9040	10.1000	9.9000
13		10.007	9.982	9.975	10.054	9.985	9.911		12	10.0177	10.0000	9.9844	10.0643	9.9045	10.1000	9.9000
14		10.007	10.028	10.039	9.975	10.032	9.986		13	9.9788	10.0000	9.9840	10.0609	9.9070	10.1000	9.9000
15		9.984	10.063	9.982	9.991	9.968	10.058		14	10.0213	10.0000	9.9866	10.0662	9.9071	10.1000	9.9000
16		10.032	10.049	9.966	10.068	10.005	10.028		15	9.9407	10.0000	9.9836	10.0677	9.9394	10.1000	9.9000
17		9.976	9.917	9.908	9.991	10.018	9.998		16	10.0696	10.0000	9.9889	10.0916	9.8862	10.1000	9.9000
18		9.982	9.896	9.961	9.961	9.969	9.935		17	10.0161	10.0000	9.9905	10.0920	9.8891	10.1000	9.9000
19		10.017	10.004	10.003	9.955	10.020	9.969		18	10.0709	10.0000	9.9950	10.1080	9.8820	10.1000	9.9000
20		9.986	10.050	10.038	9.981	9.960	9.912		19	10.0658	10.0000	9.9987	10.1185	9.8789	10.1000	9.9000
21		9.988	9.984	9.950	9.975	9.953	10.058		20	10.0800	10.0000	10.0028	10.1311	9.8745	10.1000	9.9000
22		10.066	9.984	9.964	9.978	10.042	10.055		21	9.9514	10.0000	10.0003	10.1298	9.8709	10.1000	9.9000
23		9.956	10.018	9.989	10.000	10.015	10.015		22	9.9958	10.0000	10.0001	10.1266	9.8737	10.1000	9.9000
24		10.061	9.944	9.990	9.983	9.995	10.003		23	10.0338	10.0000	10.0016	10.1270	9.8762	10.1000	9.9000
25		9.905	9.905	9.987	10.058	9.916	9.978		24	10.0422	10.0000	10.0033	10.1284	9.8781	10.1000	9.9000
26		9.977	9.995	9.920	9.934	9.916	10.072		25	10.0136	10.0000	10.0037	10.1265	9.8809	10.1000	9.9000
27									26	9.9791	10.0000	10.0028	10.1240	9.8815	10.1000	9.9000
28									27	9.9443	10.0000	10.0006	10.1240	9.8772	10.1000	9.9000
29									28	9.9403	10.0000	9.9985	10.1242	9.8727	10.1000	9.9000
30									29	9.9424	10.0000	9.9965	10.1238	9.8692	10.1000	9.9000
31									30	9.9421	10.0000	9.9947	10.1233	9.8662	10.1000	9.9000
32									31	9.9435	10.0000	9.9933	10.1220	9.8645	10.1000	9.9000
33									32	10.0057	10.0000	9.9936	10.1205	9.8668	10.1000	9.9000
34									33	9.9715	10.0000	9.9930	10.1184	9.8675	10.1000	9.9000
35									34	10.0388	10.0000	9.9943	10.1201	9.8686	10.1000	9.9000
36									35	10.0019	10.0000	9.9945	10.1185	9.8706	10.1000	9.9000
37									36	9.9616	10.0000	9.9936	10.1169	9.8703	10.1000	9.9000
38									37	9.9144	10.0000	9.9915	10.1191	9.8639	10.1000	9.9000
39									38	10.0321	10.0000	9.9926	10.1200	9.8651	10.1000	9.9000
40									39	10.0361	10.0000	9.9937	10.1211	9.8662	10.1000	9.9000
41									40	10.0253	10.0000	9.9945	10.1212	9.8677	10.1000	9.9000
42									41	10.0109	10.0000	9.9949	10.1203	9.8695	10.1000	9.9000
43									42	9.9935	10.0000	9.9948	10.1187	9.8709	10.1000	9.9000
44									43	9.9748	10.0000	9.9944	10.1171	9.8716	10.1000	9.9000
45									44	9.9962	10.0000	9.9944	10.1158	9.8730	10.1000	9.9000
46									45	10.0369	10.0000	9.9953	10.1168	9.8739	10.1000	9.9000
47									46	10.1500	10.0000	9.9987	10.1366	9.8608	10.1000	9.9000

Figure: 74.66 - Tableaux Microsoft Excel 14.0.6123 avec mesures individuelles

Soit:

	I	J	K	L	M	N
1	N°	Mesures	T	CL	UCL	LCL
2	1	9.3943	10	=MOYENNE(\$J\$2:J2)	=L2+3*ECARTYPEP(\$J\$2:J2)	=L2-3*ECARTYPEP(\$J\$2:J2)
3	2	10.0196	10	=MOYENNE(\$J\$2:J3)	=L3+3*ECARTYPEP(\$J\$2:J3)	=L3-3*ECARTYPEP(\$J\$2:J3)
4	3	9.3732	10	=MOYENNE(\$J\$2:J4)	=L4+3*ECARTYPEP(\$J\$2:J4)	=L4-3*ECARTYPEP(\$J\$2:J4)
5	4	10.0088	10	=MOYENNE(\$J\$2:J5)	=L5+3*ECARTYPEP(\$J\$2:J5)	=L5-3*ECARTYPEP(\$J\$2:J5)
6	5	9.3685	10	=MOYENNE(\$J\$2:J6)	=L6+3*ECARTYPEP(\$J\$2:J6)	=L6-3*ECARTYPEP(\$J\$2:J6)
7	6	9.3544	10	=MOYENNE(\$J\$2:J7)	=L7+3*ECARTYPEP(\$J\$2:J7)	=L7-3*ECARTYPEP(\$J\$2:J7)
8	7	9.3721	10	=MOYENNE(\$J\$2:J8)	=L8+3*ECARTYPEP(\$J\$2:J8)	=L8-3*ECARTYPEP(\$J\$2:J8)
9	8	9.3643	10	=MOYENNE(\$J\$2:J9)	=L9+3*ECARTYPEP(\$J\$2:J9)	=L9-3*ECARTYPEP(\$J\$2:J9)
10	9	9.3426	10	=MOYENNE(\$J\$2:J10)	=L10+3*ECARTYPEP(\$J\$2:J10)	=L10-3*ECARTYPEP(\$J\$2:J10)
11	10	9.3712	10	=MOYENNE(\$J\$2:J11)	=L11+3*ECARTYPEP(\$J\$2:J11)	=L11-3*ECARTYPEP(\$J\$2:J11)
12	11	10.0259	10	=MOYENNE(\$J\$2:J12)	=L12+3*ECARTYPEP(\$J\$2:J12)	=L12-3*ECARTYPEP(\$J\$2:J12)
13	12	10.0177	10	=MOYENNE(\$J\$2:J13)	=L13+3*ECARTYPEP(\$J\$2:J13)	=L13-3*ECARTYPEP(\$J\$2:J13)
14	13	9.3788	10	=MOYENNE(\$J\$2:J14)	=L14+3*ECARTYPEP(\$J\$2:J14)	=L14-3*ECARTYPEP(\$J\$2:J14)
15	14	10.0213	10	=MOYENNE(\$J\$2:J15)	=L15+3*ECARTYPEP(\$J\$2:J15)	=L15-3*ECARTYPEP(\$J\$2:J15)
16	15	9.3407	10	=MOYENNE(\$J\$2:J16)	=L16+3*ECARTYPEP(\$J\$2:J16)	=L16-3*ECARTYPEP(\$J\$2:J16)
17	16	10.0636	10	=MOYENNE(\$J\$2:J17)	=L17+3*ECARTYPEP(\$J\$2:J17)	=L17-3*ECARTYPEP(\$J\$2:J17)
18	17	10.0161	10	=MOYENNE(\$J\$2:J18)	=L18+3*ECARTYPEP(\$J\$2:J18)	=L18-3*ECARTYPEP(\$J\$2:J18)
19	18	10.0709	10	=MOYENNE(\$J\$2:J19)	=L19+3*ECARTYPEP(\$J\$2:J19)	=L19-3*ECARTYPEP(\$J\$2:J19)
20	19	10.0658	10	=MOYENNE(\$J\$2:J20)	=L20+3*ECARTYPEP(\$J\$2:J20)	=L20-3*ECARTYPEP(\$J\$2:J20)
21	20	10.08	10	=MOYENNE(\$J\$2:J21)	=L21+3*ECARTYPEP(\$J\$2:J21)	=L21-3*ECARTYPEP(\$J\$2:J21)
22	21	9.3514	10	=MOYENNE(\$J\$2:J22)	=L22+3*ECARTYPEP(\$J\$2:J22)	=L22-3*ECARTYPEP(\$J\$2:J22)
23	22	9.3958	10	=MOYENNE(\$J\$2:J23)	=L23+3*ECARTYPEP(\$J\$2:J23)	=L23-3*ECARTYPEP(\$J\$2:J23)
24	23	10.0338	10	=MOYENNE(\$J\$2:J24)	=L24+3*ECARTYPEP(\$J\$2:J24)	=L24-3*ECARTYPEP(\$J\$2:J24)
25	24	10.0422	10	=MOYENNE(\$J\$2:J25)	=L25+3*ECARTYPEP(\$J\$2:J25)	=L25-3*ECARTYPEP(\$J\$2:J25)
26	25	10.0136	10	=MOYENNE(\$J\$2:J26)	=L26+3*ECARTYPEP(\$J\$2:J26)	=L26-3*ECARTYPEP(\$J\$2:J26)
27	26	9.3731	10	=MOYENNE(\$J\$2:J27)	=L27+3*ECARTYPEP(\$J\$2:J27)	=L27-3*ECARTYPEP(\$J\$2:J27)
28	27	9.3449	10	=MOYENNE(\$J\$2:J28)	=L28+3*ECARTYPEP(\$J\$2:J28)	=L28-3*ECARTYPEP(\$J\$2:J28)
29	28	9.3403	10	=MOYENNE(\$J\$2:J29)	=L29+3*ECARTYPEP(\$J\$2:J29)	=L29-3*ECARTYPEP(\$J\$2:J29)
30	29	9.3424	10	=MOYENNE(\$J\$2:J30)	=L30+3*ECARTYPEP(\$J\$2:J30)	=L30-3*ECARTYPEP(\$J\$2:J30)
31	30	9.3421	10	=MOYENNE(\$J\$2:J31)	=L31+3*ECARTYPEP(\$J\$2:J31)	=L31-3*ECARTYPEP(\$J\$2:J31)
32	31	9.3435	10	=MOYENNE(\$J\$2:J32)	=L32+3*ECARTYPEP(\$J\$2:J32)	=L32-3*ECARTYPEP(\$J\$2:J32)
33	32	10.0057	10	=MOYENNE(\$J\$2:J33)	=L33+3*ECARTYPEP(\$J\$2:J33)	=L33-3*ECARTYPEP(\$J\$2:J33)
34	33	9.3715	10	=MOYENNE(\$J\$2:J34)	=L34+3*ECARTYPEP(\$J\$2:J34)	=L34-3*ECARTYPEP(\$J\$2:J34)
35	34	10.0388	10	=MOYENNE(\$J\$2:J35)	=L35+3*ECARTYPEP(\$J\$2:J35)	=L35-3*ECARTYPEP(\$J\$2:J35)
36	35	10.0019	10	=MOYENNE(\$J\$2:J36)	=L36+3*ECARTYPEP(\$J\$2:J36)	=L36-3*ECARTYPEP(\$J\$2:J36)
37	36	9.3616	10	=MOYENNE(\$J\$2:J37)	=L37+3*ECARTYPEP(\$J\$2:J37)	=L37-3*ECARTYPEP(\$J\$2:J37)
38	37	9.3144	10	=MOYENNE(\$J\$2:J38)	=L38+3*ECARTYPEP(\$J\$2:J38)	=L38-3*ECARTYPEP(\$J\$2:J38)
39	38	10.0321	10	=MOYENNE(\$J\$2:J39)	=L39+3*ECARTYPEP(\$J\$2:J39)	=L39-3*ECARTYPEP(\$J\$2:J39)
40	39	10.0361	10	=MOYENNE(\$J\$2:J40)	=L40+3*ECARTYPEP(\$J\$2:J40)	=L40-3*ECARTYPEP(\$J\$2:J40)
41	40	10.0253	10	=MOYENNE(\$J\$2:J41)	=L41+3*ECARTYPEP(\$J\$2:J41)	=L41-3*ECARTYPEP(\$J\$2:J41)
42	41	10.0109	10	=MOYENNE(\$J\$2:J42)	=L42+3*ECARTYPEP(\$J\$2:J42)	=L42-3*ECARTYPEP(\$J\$2:J42)
43	42	9.3935	10	=MOYENNE(\$J\$2:J43)	=L43+3*ECARTYPEP(\$J\$2:J43)	=L43-3*ECARTYPEP(\$J\$2:J43)
44	43	9.3748	10	=MOYENNE(\$J\$2:J44)	=L44+3*ECARTYPEP(\$J\$2:J44)	=L44-3*ECARTYPEP(\$J\$2:J44)
45	44	9.3962	10	=MOYENNE(\$J\$2:J45)	=L45+3*ECARTYPEP(\$J\$2:J45)	=L45-3*ECARTYPEP(\$J\$2:J45)
46	45	10.0369	10	=MOYENNE(\$J\$2:J46)	=L46+3*ECARTYPEP(\$J\$2:J46)	=L46-3*ECARTYPEP(\$J\$2:J46)
47	46	10.15	10	=MOYENNE(\$J\$2:J47)	=L47+3*ECARTYPEP(\$J\$2:J47)	=L47-3*ECARTYPEP(\$J\$2:J47)
48	47	9.3936	10	=MOYENNE(\$J\$2:J48)	=L48+3*ECARTYPEP(\$J\$2:J48)	=L48-3*ECARTYPEP(\$J\$2:J48)

Figure: 74.67 - Formules Microsoft Excel 14.0.6123 correspondantes au tableau précédent

Nous avons alors la carte de contrôle suivante:

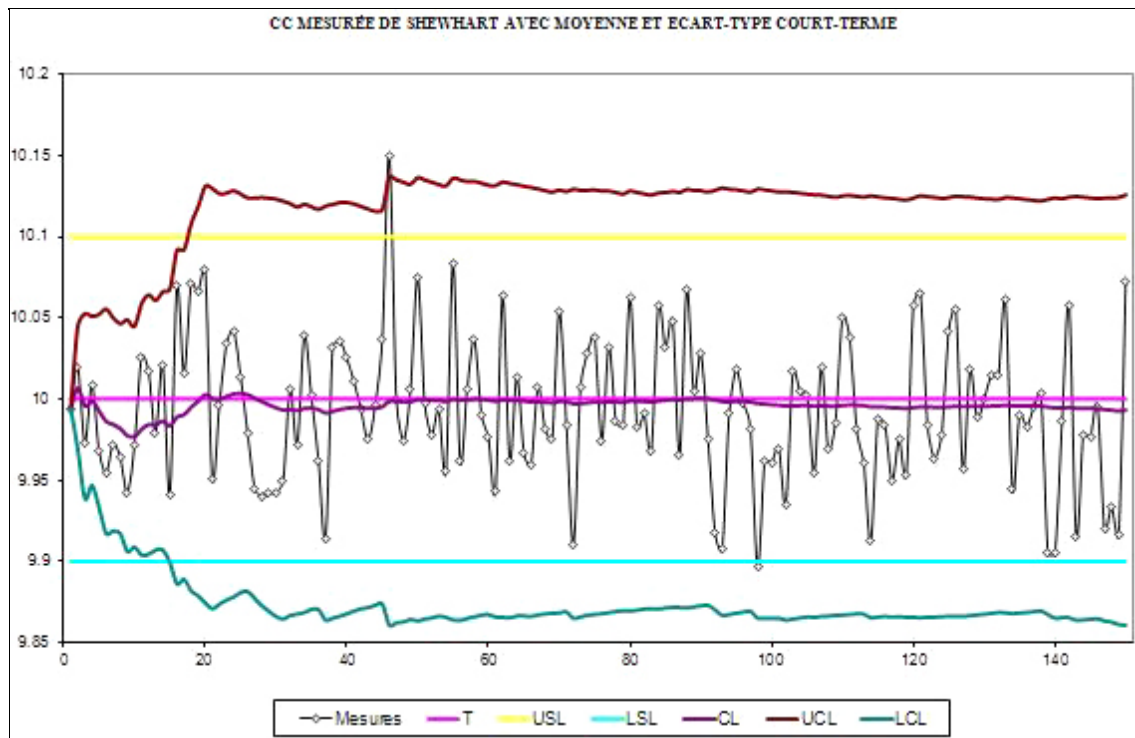


Figure: 74.68 - Carte de contrôle mesurée de Shewhart avec moyenne et écart-type court-terme

où nous avons pris en ce qui concerne les limites imposées (nous supposons qu'elles imposent un écart-type de 0.10):

$$USL = 10 + 3\sigma = 10 + 3 \cdot 0.0\bar{3} = 10.10$$

$$T = 10 \quad (74.449)$$

$$LSL = 10 - 3\sigma = 10 - 3 \cdot 0.0\bar{3} = 9.90$$

Nous voyons par exemple avec la carte de contrôle ci-dessus que la mesure CL (donc l'estimateur de la moyenne) se stabilise proche de la cible T entre le début et la fin. Il en va de même pour UCL et LCL (car l'estimateur de l'écart-type se stabilise) par contre selon l'usage en ingénierie les résultats ne sont pas bons puisque l' UCL et LCL sont au-delà des spécifications imposées par le client! À remarquer que cette carte de contrôle aurait selon les règles de la WECO un unique point au-delà des $\pm 3\sigma$ (nous indiquons cela pour des raisons de comparaison avec les prochaines cartes).

Remarque: Sur toutes les cartes sans aucune exception, les spécifications du client doivent être indiquées en plus des limites calculées!

À ce niveau, nous imaginerons que nous restons dans l'incapacité de déterminer la loi de probabilité des données (problème d'échantillons ou de compétences) et donc in extenso tout aussi incapables de calculer des intervalles de confiance pour la moyenne ou l'écart-type!

La prochaine carte de contrôle est censée y remédier.

5.4.3. CC AVEC MOYENNES BASÉES SUR L'ERREUR-STANDARD

Une fois le processus ou procédé stabilisé, nous considérons que la moyenne et l'écart-type se stabilisent et que même si la loi de probabilité que suit un groupe de k échantillons n'est pas identique d'un jour à l'autre, de par le théorème central limite, l'ensemble des moyennes suit elle une loi Normale. Dès lors il peut être plus intéressant de représenter la variation des moyennes dans le temps et non plus simplement les mesures simples.

Avantages: Masque (lisse) les variations aléatoires non systématiques et permet donc souvent d'éviter des actions correctives non nécessaires (fausses alertes).

Problèmes: Cette carte de contrôle suppose que les prélèvements sont identiquement distribués de par les hypothèses de base de sa construction mathématique. De plus, dans le cadre des limites fixes de Shewhart, la loi de distribution est supposée symétrique. Le plus gros problème de cette carte est aussi que le nombre d'individus par échantillon doit être suffisamment grand pour que l'estimateur sans biais de l'écart-type ne soit pas trop éloigné de la réalité, ce qui fait qu'on ne la trouve pas dans les logiciels (problème qui fait que, souvent, elle est remplacée par une autre carte que nous verrons un peu plus loin et qu).

Nous considérons alors le tableau suivant de 25 prélèvements (que nous identifierons dans le cas présent par la variable k) de 6 individus chacun (que nous identifierons par la variable n) prises d'une production continue (sans arrêt machine!):

	A	B	C	D	E	F	G
4		Ech. 1	Ech. 2	Ech. 3	Ech. 4	Ech. 5	Ech. 6
5	Jour 1	9.9343	10.0136	9.9732	10.0088	9.9685	9.9544
6	Jour 2	9.9721	9.9643	9.9426	9.9712	10.0259	10.0177
7	Jour 3	9.9788	10.0213	9.9407	10.0636	10.0161	10.0709
8	Jour 4	10.0658	10.08	9.9514	9.9358	10.0338	10.0422
9	Jour 5	10.0136	9.9791	9.9443	9.9403	9.9424	9.9421
10	Jour 6	9.9435	10.0057	9.9715	10.0388	10.0019	9.9616
11	Jour 7	9.9144	10.0321	10.0361	10.0253	10.0109	9.9935
12	Jour 8	9.9748	9.9962	10.0363	10.15	9.9936	9.9741
13	Jour 9	10.0063	10.0749	9.9971	9.978	9.9934	9.9555
14	Jour 10	10.084	9.9619	10.0063	10.0364	9.9907	9.9762
15	Jour 11	9.9436	10.0638	9.9617	10.0138	9.9672	9.9591
16	Jour 12	10.0074	9.9817	9.9753	10.0541	9.9845	9.9107
17	Jour 13	10.0069	10.0282	10.0385	9.9747	10.0323	9.986
18	Jour 14	9.9836	10.063	9.9824	9.9909	9.9684	10.0575
19	Jour 15	10.0317	10.0485	9.9659	10.0675	10.0047	10.0277
20	Jour 16	9.9756	9.9174	9.9076	9.9909	10.018	9.9979
21	Jour 17	9.9818	9.9962	9.9614	9.9612	9.9634	9.9349
22	Jour 18	10.0168	10.0043	10.0029	9.9549	10.0199	9.9691
23	Jour 19	9.9857	10.0503	10.0381	9.9811	9.9603	9.9123
24	Jour 20	9.9878	9.9841	9.9496	9.9754	9.9532	10.0576
25	Jour 21	10.0657	9.9838	9.9636	9.9779	10.0423	10.0547
26	Jour 22	9.9564	10.0184	9.989	9.9999	10.0146	10.0145
27	Jour 23	10.0611	9.9443	9.9899	9.9833	9.9947	10.0033
28	Jour 24	9.9054	9.9049	9.987	10.058	9.9156	9.9782
29	Jour 25	9.9768	9.9954	9.9196	9.9341	9.9162	10.0721

Figure: 74.69 - Mesures journalières sans arrêt de la machine

où chaque jour, nous prenons que 6 individus de l'ensemble des éléments à notre disposition. Sur la carte de contrôle, nous représenterons alors chaque point comme étant la moyenne de ces 6 individus.

Nous avons démontré dans le chapitre de Statistique que l'écart-type de la moyenne (erreur-type) d'une série de n variables aléatoires indépendantes et identiquement distribuées était:

$$\sigma_{\bar{x}} = \frac{\sigma_X}{\sqrt{n}} \quad (74.450)$$

Nous avons alors les indicateurs de limite de contrôle de Shewhart suivants (évidemment comme nous utilisons les estimateurs de la moyenne et de l'écart-type, plus le nombre de prélèvements et d'individus est grand, plus les limites sont asymptotiquement précises):

$$\begin{aligned}
 LSL &= \text{imposé par client/direction} \\
 UCL_{\bar{X}} &= \bar{\bar{X}} + 3\sigma_{\bar{X}} = \bar{\bar{X}} + 3 \frac{\sigma_X}{\sqrt{n_i}} \\
 CL_{\bar{X}} &= E(\bar{X}) = \bar{\bar{X}} \\
 T &= \text{objectif imposé par client/direction} \\
 LCL_{\bar{X}} &= \bar{\bar{X}} - 3\sigma_{\bar{X}} = \bar{\bar{X}} - 3 \frac{\sigma_X}{\sqrt{n_i}} \\
 LSL &= \text{imposé par client/direction}
 \end{aligned} \tag{74.451}$$

Avec:

$$\bar{\bar{X}} = \frac{1}{k} \sum_{i=1}^k \left(\frac{1}{n_i} \sum_{j=1}^{n_i} X_{i,j} \right) \tag{74.452}$$

et si tous les n_i sont identique:

$$\bar{\bar{X}} = \frac{1}{k} \sum_{i=1}^k \left(\frac{1}{n} \sum_{j=1}^n X_{i,j} \right) = \frac{1}{nk} \sum_{i=1}^k \left(\sum_{j=1}^n X_{i,j} \right) = \frac{1}{N} \sum_{i=1}^k \left(\sum_{j=1}^n X_{i,j} \right) \tag{74.453}$$

Remarquons que nous pouvons alors calculer des probabilités de sous-groupes non-conformes par excès ou par défaut en utilisant le fait que $\bar{X}$ suit une loi Normale de paramètres:

$$N(\mu, \sigma^2) \cong N\left(\bar{\bar{X}}, \frac{\sigma_X^2}{n}\right) \tag{74.454}$$

Nous obtenons alors le tableau suivant:

	I	J	K	L	M	N	O	P
1	N°	Moyenne Mesures	CL	UCL	LCL	USL	LSL	T
2	1	3.3865	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
3	2	3.3823	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
4	3	10.0162	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
5	4	10.0282	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
6	5	3.3604	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
7	6	3.3682	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
8	7	10.0021	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
9	8	10.0219	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
10	9	10.0009	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
11	10	10.0093	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
12	11	3.3843	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
13	12	3.3856	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
14	13	10.0111	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
15	14	10.0076	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
16	15	10.0243	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
17	16	3.3673	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
18	17	3.3508	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
19	18	3.3347	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
20	19	3.3880	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
21	20	3.3846	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
22	21	10.0147	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
23	22	3.3988	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
24	23	3.3361	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
25	24	3.3582	3.3333	10.0476	3.3330	10.04082483	3.353175171	10
26	25	3.3630	3.3333	10.0476	3.3330	10.04082483	3.353175171	10

Figure: 74.70 - Indicateurs correspondants aux mesures obtenues

Soit:

	I	J	K	L	M	N
1		Moyenne mesures:	=MOYENNE(\$B5:\$G5)			
2		Ecart-Type:	=ECARTYPE(\$B5:\$G23)			
3						
4	N°	Moyenne Mesures	CL	T	UCL	LCL
5	1	=MOYENNE(\$B5:\$G5)	=MOYENNE(\$B5:\$G23)	10	=K5+3*\$K\$2/RACINE(NBVAL(B5:G5))	=K5-3*\$K\$2/RACINE(NBVAL(B5:G5))
6	2	=MOYENNE(\$B6:\$G6)	=MOYENNE(\$B5:\$G23)	10	=K6+3*\$K\$2/RACINE(NBVAL(B6:G6))	=K6-3*\$K\$2/RACINE(NBVAL(B6:G6))
7	3	=MOYENNE(\$B7:\$G7)	=MOYENNE(\$B5:\$G23)	10	=K7+3*\$K\$2/RACINE(NBVAL(B7:G7))	=K7-3*\$K\$2/RACINE(NBVAL(B7:G7))
8	4	=MOYENNE(\$B8:\$G8)	=MOYENNE(\$B5:\$G23)	10	=K8+3*\$K\$2/RACINE(NBVAL(B8:G8))	=K8-3*\$K\$2/RACINE(NBVAL(B8:G8))
9	5	=MOYENNE(\$B9:\$G9)	=MOYENNE(\$B5:\$G23)	10	=K9+3*\$K\$2/RACINE(NBVAL(B9:G9))	=K9-3*\$K\$2/RACINE(NBVAL(B9:G9))
10	6	=MOYENNE(\$B10:\$G10)	=MOYENNE(\$B5:\$G23)	10	=K10+3*\$K\$2/RACINE(NBVAL(B10:G10))	=K10-3*\$K\$2/RACINE(NBVAL(B10:G10))
11	7	=MOYENNE(\$B11:\$G11)	=MOYENNE(\$B5:\$G23)	10	=K11+3*\$K\$2/RACINE(NBVAL(B11:G11))	=K11-3*\$K\$2/RACINE(NBVAL(B11:G11))
12	8	=MOYENNE(\$B12:\$G12)	=MOYENNE(\$B5:\$G23)	10	=K12+3*\$K\$2/RACINE(NBVAL(B12:G12))	=K12-3*\$K\$2/RACINE(NBVAL(B12:G12))
13	9	=MOYENNE(\$B13:\$G13)	=MOYENNE(\$B5:\$G23)	10	=K13+3*\$K\$2/RACINE(NBVAL(B13:G13))	=K13-3*\$K\$2/RACINE(NBVAL(B13:G13))
14	10	=MOYENNE(\$B14:\$G14)	=MOYENNE(\$B5:\$G23)	10	=K14+3*\$K\$2/RACINE(NBVAL(B14:G14))	=K14-3*\$K\$2/RACINE(NBVAL(B14:G14))
15	11	=MOYENNE(\$B15:\$G15)	=MOYENNE(\$B5:\$G23)	10	=K15+3*\$K\$2/RACINE(NBVAL(B15:G15))	=K15-3*\$K\$2/RACINE(NBVAL(B15:G15))
16	12	=MOYENNE(\$B16:\$G16)	=MOYENNE(\$B5:\$G23)	10	=K16+3*\$K\$2/RACINE(NBVAL(B16:G16))	=K16-3*\$K\$2/RACINE(NBVAL(B16:G16))
17	13	=MOYENNE(\$B17:\$G17)	=MOYENNE(\$B5:\$G23)	10	=K17+3*\$K\$2/RACINE(NBVAL(B17:G17))	=K17-3*\$K\$2/RACINE(NBVAL(B17:G17))
18	14	=MOYENNE(\$B18:\$G18)	=MOYENNE(\$B5:\$G23)	10	=K18+3*\$K\$2/RACINE(NBVAL(B18:G18))	=K18-3*\$K\$2/RACINE(NBVAL(B18:G18))
19	15	=MOYENNE(\$B19:\$G19)	=MOYENNE(\$B5:\$G23)	10	=K19+3*\$K\$2/RACINE(NBVAL(B19:G19))	=K19-3*\$K\$2/RACINE(NBVAL(B19:G19))
20	16	=MOYENNE(\$B20:\$G20)	=MOYENNE(\$B5:\$G23)	10	=K20+3*\$K\$2/RACINE(NBVAL(B20:G20))	=K20-3*\$K\$2/RACINE(NBVAL(B20:G20))
21	17	=MOYENNE(\$B21:\$G21)	=MOYENNE(\$B5:\$G23)	10	=K21+3*\$K\$2/RACINE(NBVAL(B21:G21))	=K21-3*\$K\$2/RACINE(NBVAL(B21:G21))
22	18	=MOYENNE(\$B22:\$G22)	=MOYENNE(\$B5:\$G23)	10	=K22+3*\$K\$2/RACINE(NBVAL(B22:G22))	=K22-3*\$K\$2/RACINE(NBVAL(B22:G22))
23	19	=MOYENNE(\$B23:\$G23)	=MOYENNE(\$B5:\$G23)	10	=K23+3*\$K\$2/RACINE(NBVAL(B23:G23))	=K23-3*\$K\$2/RACINE(NBVAL(B23:G23))
24	20	=MOYENNE(\$B24:\$G24)	=MOYENNE(\$B5:\$G23)	10	=K24+3*\$K\$2/RACINE(NBVAL(B24:G24))	=K24-3*\$K\$2/RACINE(NBVAL(B24:G24))
25	21	=MOYENNE(\$B25:\$G25)	=MOYENNE(\$B5:\$G23)	10	=K25+3*\$K\$2/RACINE(NBVAL(B25:G25))	=K25-3*\$K\$2/RACINE(NBVAL(B25:G25))
26	22	=MOYENNE(\$B26:\$G26)	=MOYENNE(\$B5:\$G23)	10	=K26+3*\$K\$2/RACINE(NBVAL(B26:G26))	=K26-3*\$K\$2/RACINE(NBVAL(B26:G26))
27	23	=MOYENNE(\$B27:\$G27)	=MOYENNE(\$B5:\$G23)	10	=K27+3*\$K\$2/RACINE(NBVAL(B27:G27))	=K27-3*\$K\$2/RACINE(NBVAL(B27:G27))
28	24	=MOYENNE(\$B28:\$G28)	=MOYENNE(\$B5:\$G23)	10	=K28+3*\$K\$2/RACINE(NBVAL(B28:G28))	=K28-3*\$K\$2/RACINE(NBVAL(B28:G28))
29	25	=MOYENNE(\$B29:\$G29)	=MOYENNE(\$B5:\$G23)	10	=K29+3*\$K\$2/RACINE(NBVAL(B29:G29))	=K29-3*\$K\$2/RACINE(NBVAL(B29:G29))

Figure: 74.71 - Formules Microsoft Excel 14.0.6123 correspondantes au tableau précédent

Nous avons alors la carte de contrôle suivante:

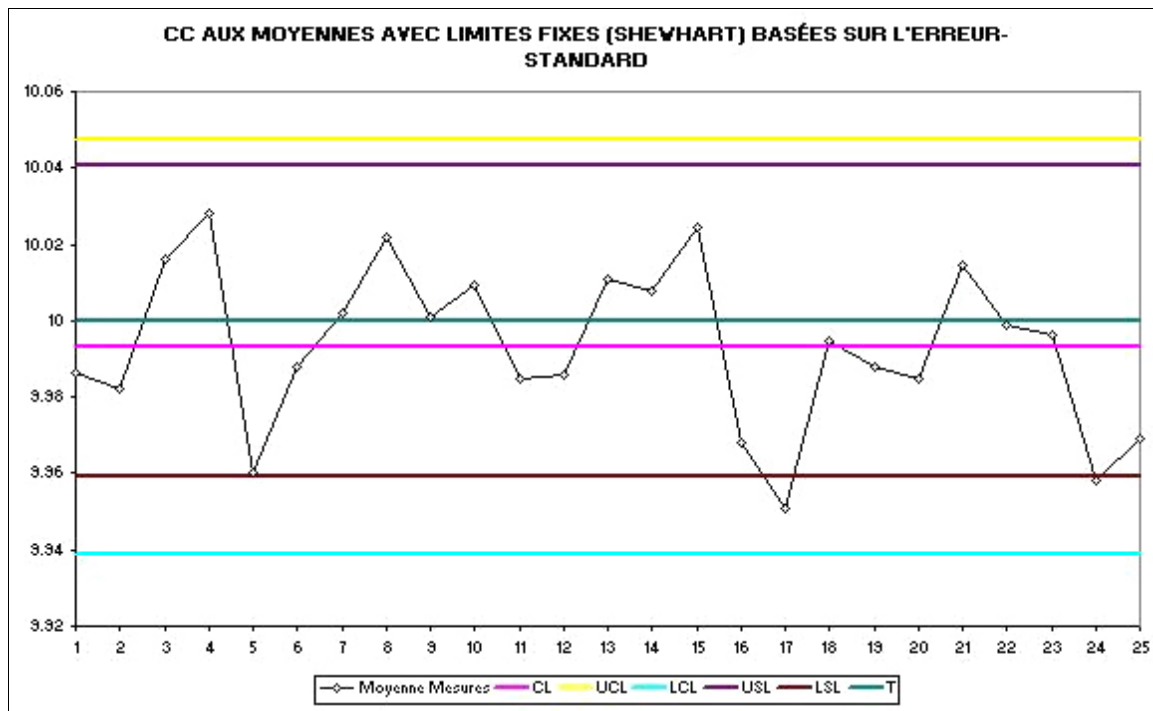


Figure: 74.72 - Carte de contrôle X barre/barre - ER

où nous avons pris en ce qui concerne les limites imposées:

$$\begin{aligned}
 USL &= 10 + 3 \cdot \frac{0.0\bar{3}}{\sqrt{6}} \cong 10.040 \\
 T &= 10 \\
 LSL &= 10 - 3 \cdot \frac{0.0\bar{3}}{\sqrt{6}} \cong 9.959
 \end{aligned}
 \quad (74.455)$$

À remarquer que cette carte de contrôle aurait selon les règles de la WECO six points au-delà des $\pm 3\sigma$ (nous indiquons cela pour des raisons de comparaison avec la carte précédente et les prochaines cartes). Il s'agit donc d'une carte plus sensible aux variations que la précédente. Il faut remarquer aussi que les limites UCL et LCL sont plus larges que les USL et LSL imposées.

Avec des limites de $\pm 3\sigma_{\bar{x}}$ et selon l'hypothèse de données normalement distribuées, nous avons en utilisant la symétrie de la loi Normale et en notant comme à l'habitude Φ la fonction de probabilité cumulée de la loi Normale centrée réduite:

$$\begin{aligned}
 1 - P(LCL_{\bar{x}} \leq \bar{X} \leq UCL_{\bar{x}}) &= 2P(LCL_{\bar{x}} \leq \bar{X}) = 2\left(1 - P(\bar{X} \leq UCL_{\bar{x}})\right) \\
 &= 2\Phi(-3) = 2(1 - \Phi(3)) = 2 \cdot 0.01350 = 0.0027 \cong \frac{1}{370}
 \end{aligned}
 \quad (74.456)$$

Donc nous aurons statistiquement 1 point (pièce/action) sur le graphique sur 370 en dehors des $\pm 3\sigma_{\bar{x}}$ par unité de temps de mesure (donc si chaque point sur la carte est fait par fréquence de 1 heure nous avons alors une ARL de 0.0027/heure) uniquement pour des raisons statistiques et qui seront donc de fausses alarmes.

Nous appelons ce dernier chiffre l'ARL pour "**Average Run Length**" (certains utilisent l'appellation française P.O.M. pour "**Période Opérationnelle Moyenne**"). Lorsqu'il est rapporté à une unité de temps particulière, nous parlons alors de ATS pour "**Average Time to Signal**". Évidemment, par extension, l'ajout des règles empiriques de la WECO ajoutent des fausses alarmes et cela a pour effet négatif immédiat de diminuer l'ARL.

L'ARL est aussi importante dans la pratique, car elle donne la taille d'intervalle de pièces au-delà de laquelle il ne faut pas aller pour échantillonner une production! On fera en sorte dans la pratique (quand cela est possible) de prendre une valeur inférieure proche de l'ARL. pour de grandes productions mais si possible pas au-delà!

Dans la pratique, la valeur de 0.0027 que nous avons obtenu ci-dessus, est appelée "l'erreur de type I" de la carte de contrôle (fausse alarme). Évidemment il y a aussi les erreurs de type II (absence d'alarme alors qu'il devrait y en avoir une). Mais comme je n'ai jamais vu qui que ce soit utiliser cela dans la pratique.

Remarque: Comme nous l'avons déjà mentionné, parfois plusieurs niveaux de contrôles sont ajoutés sur les cartes, ainsi nous pouvons observer du $\pm 3\sigma_{\bar{x}}$ et en même temps du $\pm 2\sigma_{\bar{x}}$. Aux U.S. le facteur 2 ou 3 serait majoritairement utilisé que la distribution soit symétrique ou non... alors que dans les pays européens, nous utilisons des limites probabilistes qui ont une bien meilleure fiabilité lorsque les données ne sont pas distribuées selon des fonctions symétriques (comme c'est le cas pour les cartes bien connues utilisant l'étendue R que nous verrons plus tard).

Rappelons que le problème de cette carte est que $\sigma_{\bar{x}}$ est asymptotiquement correcte que si n_i est grand. Alors que dans la réalité ce n'est de loin pas le cas puisque souvent cette valeur est inférieure à 10. Dès lors nous verrons plus loin une carte de la moyenne dont les limites fixes sont calculées sur la base de l'étendue R et non pas de $\sigma_{\bar{x}}$. Cependant, dans le chapitre de Statistiques, nous avons démontré que lorsque l'écart-type est estimé, nous avons alors:

$$\bar{X} - \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \quad (74.457)$$

Nous pouvons donc utiliser la carte de Shewhart aux limites probabilistes suivante:

USL = imposé par client/direction

$$UCL_{\bar{X}_p} = \bar{\bar{X}} + 3S_{\bar{x}} = \bar{\bar{X}} + T_{\alpha/2}(n_i - 1) \frac{S_X}{\sqrt{n}}$$

$$CL_{\bar{X}_p} = E(\bar{X}) = \bar{\bar{X}}$$

(74.458)

T = objectif imposé par client/direction

$$LCL_{\bar{X}_p} = \bar{\bar{X}} - 3S_{\bar{x}} = \bar{\bar{X}} - T_{\alpha/2}(n_i - 1) \frac{S_X}{\sqrt{n}}$$

LSL = imposé par client/direction

Enfin, remarquons que rien ne nous empêche de faire de l'inférence statistique sur la moyenne en utilisant la relation démontrée dans le chapitre de Statistiques:

$$\bar{X} - Z_{\alpha/2} \frac{\sigma_X}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha} \frac{\sigma_X}{\sqrt{n}} \quad (74.459)$$

en sachant que dans le domaine du contrôle qualité il est fortement souhaitable que pour un α donné (généralement de l'ordre de 5%), la valeur zéro soit comprise dans l'intervalle:

$$\left[\bar{X} - Z_{\alpha/2} \frac{\sigma_X}{\sqrt{n}}, \bar{X} + Z_{\alpha} \frac{\sigma_X}{\sqrt{n}} \right] \quad (74.460)$$

ce qui est assimilé alors à un test d'hypothèse avec $H_0 : \mu = 0$.

5.4.4. CC AVEC ÉCARTS-TYPES (S BARRE - S)

Toujours dans le cadre d'un processus ou procédé sous contrôle, il peut être intéressant en plus de communiquer la carte de contrôle des moyennes (avec limites fixes) précédemment présentée une carte représentant les volatilités dans le temps en considérant à nouveau une distribution Normale des variables aléatoires. Effectivement, rappelons que la loi Normale est totalement définie par ses paramètres μ, σ . D'où la raison d'avoir un oeil sur ses deux là!

Avantages: Complète très bien la carte de contrôle précédente et permet de visualiser si la volatilité du procédé se stabilise. Représente bien la variation à l'intérieur même des données de chaque prélèvement (dispersion instantanée) que nous pourrions détecter.

Problèmes: Cette carte de contrôle est utilisée seulement lorsque la phase de prototypage est terminée (souvent nommée "Phase II" dans la littérature spécialisée) car elle est basée sur une hypothèse forte incontournable qui est que les données sont normalement distribuées. De plus, les limites de contrôles de Shewhart sont peu adaptées dans le cas présent, car comme nous allons le démontrer en détails, la loi de distribution de la variance n'est pas une loi symétrique alors que Shewhart suppose que c'est le cas.

L'idée est de représenter:

$$\begin{aligned} UCL_S &= \bar{S} + 3\sigma_S \\ CL_S &= E(S) = \bar{S} \\ LCL_S &= \bar{S} - 3\sigma_S \end{aligned} \quad (74.461)$$

toujours avec l'estimateur de l'espérance (k étant le nombre de prélèvements pour rappel...):

$$\bar{S} = \frac{1}{k} \sum_{i=1}^k S_i \quad (74.462)$$

et toujours avec l'estimateur de maximum de vraisemblance non biaisé de l'écart-type noté S dans le domaine de la qualité (à l'opposée de $\hat{\sigma}$ dans le domaine des Statistiques pures):

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \quad (74.463)$$

où n correspond donc bien évidemment au nombre d'individus par prélèvement (échantillon).

Pour calculer l'espérance et la variance σ_S , il faut considérer S comme une variable aléatoire distribuée normalement (c'est l'hypothèse forte en incontournable de cette carte de contrôle dont nous avons fait mention dans les problèmes!). Nous avons alors démontré dans le chapitre de Statistiques que nous avons:

$$\chi^2(n-1) = \frac{(n-1)S^2}{\sigma^2} \quad (74.464)$$

que nous noterons par la suite t comme étant la variable aléatoire correspondante. Pour la suite, nous allons considérer $(n-1)$ et σ comme de simples coefficients constants de S . Donc la variable aléatoire t sera en réalité fonction implicite de S tel que:

$$t = \frac{(n-1)}{\sigma^2} S^2 \quad (74.465)$$

Nous nous intéressons maintenant au calcul de l'espérance de S . Il est alors plus convenable de prendre la racine carrée de t tel que:

$$\sqrt{t} = \frac{\sqrt{n-1}}{\sigma} S \quad (74.466)$$

Il vient alors:

$$E(S) = \mu_S = E\left(\frac{\sigma}{\sqrt{n-1}} \sqrt{t}\right) = \frac{\sigma}{\sqrt{n-1}} E(\sqrt{t}) = \int_0^{+\infty} \sqrt{t} f(t) dt \quad (74.467)$$

Rappelons que nous avons vu dans le chapitre de Statistiques que la loi du Khi-deux était définie par la fonction de densité de probabilité suivante:

$$P_k(x) = \frac{1}{2^{k/2} \cdot \Gamma(k/2)} x^{k/2-1} e^{-x/2} \cdot 1_{[0,+\infty[} \quad (74.468)$$

Il vient alors:

$$E(S) = \frac{\sigma}{\sqrt{n-1}} \int_0^{+\infty} \sqrt{t} \frac{t^{\frac{n-1}{2}-1} e^{-t/2}}{2^{\frac{n-1}{2}} \Gamma\left(\frac{n-1}{2}\right)} dt \quad (74.469)$$

Posons pour la suite:

$$n-1 = k \quad (74.470)$$

Nous avons alors:

$$\begin{aligned}
E(S) &= \frac{\sigma}{\sqrt{k}} \int_0^{+\infty} \sqrt{t} \frac{t^{\frac{k}{2}-1} e^{-t/2}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} dt = \frac{\sigma}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right) \sqrt{k}} \int_0^{+\infty} \sqrt{t} t^{\frac{k}{2}-1} e^{-t/2} dt \\
&= \frac{\sigma}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right) \sqrt{k}} \int_0^{+\infty} t^{\frac{k}{2}-1+\frac{1}{2}} e^{-t/2} dt = \frac{\sigma}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right) \sqrt{k}} \int_0^{+\infty} t^{\frac{k-1}{2}} e^{-t/2} dt \\
&= \frac{\sigma}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right) \sqrt{k}} \frac{\sqrt{2} \Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\left(\frac{1}{2}\right)^k}} = \sigma \frac{\sqrt{2}}{2^{\frac{k}{2}} \sqrt{\left(\frac{1}{2}\right)^k} \sqrt{k}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \quad (74.471) \\
&= \sigma \frac{\sqrt{2}}{\sqrt{k}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} = \sigma \frac{\sqrt{2}}{\sqrt{n-1}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \\
&= \sqrt{\frac{2}{n-1}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sigma = c_4 \sigma
\end{aligned}$$

Nous avons donc:

$$E(S) = \bar{S} = \sigma c_4 = \sigma \sqrt{\frac{2}{n-1}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \quad (74.472)$$

Pour terminer sur ce point, il faut savoir qu'il est de tradition de noter c_4 de la manière suivante:

$$c_4(n_i) = \sqrt{\frac{2}{n_i-1}} \frac{\Gamma\left(\frac{n_i}{2}\right)}{\Gamma\left(\frac{n_i-1}{2}\right)} \quad (74.473)$$

et que nous en trouvons des valeurs dans les tables alors que c'est inutile si nous possédons un tableur comme la version anglophone de Microsoft Excel 14.0.6123 en écrivant la formule suivante (n_i serait écrit dans la cellule A1):

$$=\text{SQRT}(2/(A1-1))*\text{EXP}(\text{GAMMALN}(A1/2))/\text{EXP}(\text{GAMMALN}((A1-1)/2))$$

ou dans Calc de OpenOffice (la formule y étant plus simple):

$$=\text{RACINE}(2/(A1-1))*\text{GAMMA}(A1/2)/\text{GAMMA}((A1-1)/2)$$

Calculons maintenant la variance de S en utilisant la relation de Huyghens démontrée dans le chapitre de Statistiques:

$$V(S) = E(S^2) - E(S)^2 = E(S^2) - c_4^2(n_i) \sigma^2 \quad (74.474)$$

et nous avons:

$$\begin{aligned} E(S^2) &= \frac{\sigma^2}{k} \int_0^{+\infty} t \frac{t^{\frac{k}{2}-1} e^{-t/2}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} dt = \frac{\sigma^2}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right) k} \int_0^{+\infty} t^{\frac{k}{2}} e^{-t/2} dt \\ &= \frac{\sigma^2}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right) k} \int_0^{+\infty} t^{\frac{k}{2}} e^{-t/2} dt = \frac{\sigma^2}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right) k} \frac{\Gamma\left(\frac{k}{2}\right) k}{\sqrt{\left(\frac{1}{2}\right)^k}} = \sigma^2 \end{aligned} \quad (74.475)$$

Il vient alors:

$$V(S) = E(S^2) - E(S)^2 = \sigma^2 - c_4^2(n_i) \sigma^2 = \sigma^2 (1 - c_4^2(n_i)) \quad (74.476)$$

et:

$$\sigma_S = \sigma \sqrt{1 - c_4^2(n_i)} \quad (74.477)$$

Nous avons alors:

$$\begin{aligned} USL &= \text{imposé par client/direction} \\ UCL_S &= \bar{S} + 3\sigma_S = \bar{S} + 3\sigma \sqrt{1 - c_4^2(n_i)} \\ CL_S &= E(S) = \bar{S} = \sigma c_4 \\ T &= \text{objectif imposé par client/direction} \\ LCL_S &= \bar{S} - 3\sigma_S = \bar{S} - 3\sigma \sqrt{1 - c_4^2(n_i)} \\ LSL &= \text{imposé par client/direction} \end{aligned} \quad (74.478)$$

soit au final:

$$\begin{aligned} USL &= \text{imposé par client/direction} \\ UCL_S &= \bar{S} + 3 \frac{\bar{S}}{c_4(n_i)} \sqrt{1 - c_4^2(n_i)} = \bar{S} + 3 \frac{\bar{S}}{h_i} \\ CL_S &= \bar{S} \\ T &= \text{objectif imposé par client/direction} \\ LCL_S &= \bar{S} - 3 \frac{\bar{S}}{c_4(n_i)} \sqrt{1 - c_4^2(n_i)} = \bar{S} - 3 \frac{\bar{S}}{h_i(n_i)} \\ LSL &= \text{imposé par client/direction} \end{aligned} \quad (74.479)$$

ce qui est plus traditionnellement condensé sous la forme:

$$\begin{aligned}
 USL &= \text{imposé par client/direction} \\
 UCL_S &= \bar{S} \left(1 + \frac{3}{h_i} \right) = \bar{S} \cdot B_4(n_i) \\
 CL_S &= \bar{S} \\
 T &= \text{objectif imposé par client/direction} \\
 LCL_S &= \bar{S} \left(1 - \frac{3}{h_i} \right) = \bar{S} \cdot B_3(n_i) \\
 LSL &= \text{imposé par client/direction}
 \end{aligned} \tag{74.480}$$

Les coefficients c_4, B_3, B_4 (tous fonctions de n_i) sont tabulés dans des tables en fonction du nombre n_i d'individus par prélèvement (échantillon), mais encore une fois ils sont très simples à calculer dans un tableur comme Microsoft Excel avec la relation explicite donnée plus haut. Et nous avons donc:

$$\bar{S} = \frac{1}{k} \sum_{i=1}^k S_i \tag{74.481}$$

où k est le nombre de prélèvements.

Ainsi, si nous considérons toujours le tableau suivant de 25 prélèvements (k) de 6 individus (n_i):

	A	B	C	D	E	F	G
5		Ech. 1	Ech. 2	Ech. 3	Ech. 4	Ech. 5	Ech. 6
6	Jour 1	9.994	10.020	9.973	10.009	9.969	9.954
7	Jour 2	9.972	9.964	9.943	9.971	10.026	10.018
8	Jour 3	9.979	10.021	9.941	10.070	10.016	10.071
9	Jour 4	10.066	10.080	9.951	9.996	10.034	10.042
10	Jour 5	10.014	9.979	9.945	9.940	9.942	9.942
11	Jour 6	9.950	10.006	9.972	10.039	10.002	9.962
12	Jour 7	9.914	10.032	10.036	10.025	10.011	9.994
13	Jour 8	9.975	9.996	10.037	10.150	10.000	9.974
14	Jour 9	10.006	10.075	9.997	9.978	9.993	9.956
15	Jour 10	10.084	9.962	10.006	10.036	9.991	9.976
16	Jour 11	9.944	10.064	9.962	10.014	9.967	9.959
17	Jour 12	10.007	9.982	9.975	10.054	9.985	9.911
18	Jour 13	10.007	10.028	10.039	9.975	10.032	9.986
19	Jour 14	9.984	10.063	9.982	9.991	9.968	10.058
20	Jour 15	10.032	10.049	9.966	10.068	10.005	10.028
21	Jour 16	9.976	9.917	9.908	9.991	10.018	9.998
22	Jour 17	9.982	9.896	9.961	9.961	9.969	9.935
23	Jour 18	10.017	10.004	10.003	9.955	10.020	9.969
24	Jour 19	9.986	10.050	10.038	9.981	9.960	9.912
25	Jour 20	9.988	9.984	9.950	9.975	9.953	10.058
26	Jour 21	10.066	9.984	9.964	9.978	10.042	10.055
27	Jour 22	9.956	10.018	9.989	10.000	10.015	10.015
28	Jour 23	10.061	9.944	9.990	9.983	9.995	10.003
29	Jour 24	9.905	9.905	9.987	10.058	9.916	9.978
30	Jour 25	9.977	9.995	9.920	9.934	9.916	10.072

Figure: 74.73 - Mesures journalières avec 6 individus sans arrêt de la machine

Nous pouvons calculer, ou trouver dans les tables que:

$$c_4(n_i) = c_4(6) = 0.915, B_3(n_i) = B_3(6) = 0.030, B_4(n_i) = B_4(6) = 1.970 \tag{74.482}$$

Ce qui donne:

	I	J	K	L	M	N	O	P
1	c4	0.952						
2	b3	0.030						
3	b4	1.970						
4								
5	N°	S	CL	UCL	LCL	USL	LSL	T
6	1	0.0252	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
7	2	0.0325	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
8	3	0.0509	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
9	4	0.0475	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
10	5	0.0299	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
11	6	0.0333	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
12	7	0.0457	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
13	8	0.0668	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
14	9	0.0404	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
15	10	0.0447	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
16	11	0.0454	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
17	12	0.0467	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
18	13	0.0263	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
19	14	0.0414	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
20	15	0.0355	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
21	16	0.0451	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
22	17	0.0309	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
23	18	0.0265	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
24	19	0.0509	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
25	20	0.0391	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
26	21	0.0445	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
27	22	0.0236	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
28	23	0.0379	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
29	24	0.0610	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
30	25	0.0597	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333

Figure: 74.74 - Calculs des indicateurs correspondants

Soit:

	I	J	K	L	M	N	O	P
1	c	=RACINE(2/(NBVAL(B6:G6)-1))*EXP(LNGAMMA(NBVAL(B6:G6)/2))/EXP(LNGAMMA((NBVAL(B6:G6)-1)/2))						
2	b3	=1-3/J!*RACINE(1-J!*2)						
3	b4	=1-3/J!*RACINE(1-J!*2)						
4								
5	N°	S	CL	UCL	LCL	USL	LSL	T
6	1	=ECARTYPE(B6:G6)	=MOYENNE(\$J\$6:\$J\$30)	=K6*\$J\$3	=K6*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
7	2	=ECARTYPE(B7:G7)	=MOYENNE(\$J\$6:\$J\$30)	=K7*\$J\$3	=K7*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
8	3	=ECARTYPE(B8:G8)	=MOYENNE(\$J\$6:\$J\$30)	=K8*\$J\$3	=K8*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
9	4	=ECARTYPE(B9:G9)	=MOYENNE(\$J\$6:\$J\$30)	=K9*\$J\$3	=K9*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
10	5	=ECARTYPE(B10:G10)	=MOYENNE(\$J\$6:\$J\$30)	=K10*\$J\$3	=K10*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
11	6	=ECARTYPE(B11:G11)	=MOYENNE(\$J\$6:\$J\$30)	=K11*\$J\$3	=K11*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
12	7	=ECARTYPE(B12:G12)	=MOYENNE(\$J\$6:\$J\$30)	=K12*\$J\$3	=K12*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
13	8	=ECARTYPE(B13:G13)	=MOYENNE(\$J\$6:\$J\$30)	=K13*\$J\$3	=K13*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
14	9	=ECARTYPE(B14:G14)	=MOYENNE(\$J\$6:\$J\$30)	=K14*\$J\$3	=K14*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
15	10	=ECARTYPE(B15:G15)	=MOYENNE(\$J\$6:\$J\$30)	=K15*\$J\$3	=K15*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
16	11	=ECARTYPE(B16:G16)	=MOYENNE(\$J\$6:\$J\$30)	=K16*\$J\$3	=K16*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
17	12	=ECARTYPE(B17:G17)	=MOYENNE(\$J\$6:\$J\$30)	=K17*\$J\$3	=K17*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
18	13	=ECARTYPE(B18:G18)	=MOYENNE(\$J\$6:\$J\$30)	=K18*\$J\$3	=K18*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
19	14	=ECARTYPE(B19:G19)	=MOYENNE(\$J\$6:\$J\$30)	=K19*\$J\$3	=K19*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
20	15	=ECARTYPE(B20:G20)	=MOYENNE(\$J\$6:\$J\$30)	=K20*\$J\$3	=K20*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
21	16	=ECARTYPE(B21:G21)	=MOYENNE(\$J\$6:\$J\$30)	=K21*\$J\$3	=K21*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
22	17	=ECARTYPE(B22:G22)	=MOYENNE(\$J\$6:\$J\$30)	=K22*\$J\$3	=K22*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
23	18	=ECARTYPE(B23:G23)	=MOYENNE(\$J\$6:\$J\$30)	=K23*\$J\$3	=K23*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
24	19	=ECARTYPE(B24:G24)	=MOYENNE(\$J\$6:\$J\$30)	=K24*\$J\$3	=K24*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
25	20	=ECARTYPE(B25:G25)	=MOYENNE(\$J\$6:\$J\$30)	=K25*\$J\$3	=K25*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
26	21	=ECARTYPE(B26:G26)	=MOYENNE(\$J\$6:\$J\$30)	=K26*\$J\$3	=K26*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
27	22	=ECARTYPE(B27:G27)	=MOYENNE(\$J\$6:\$J\$30)	=K27*\$J\$3	=K27*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
28	23	=ECARTYPE(B28:G28)	=MOYENNE(\$J\$6:\$J\$30)	=K28*\$J\$3	=K28*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
29	24	=ECARTYPE(B29:G29)	=MOYENNE(\$J\$6:\$J\$30)	=K29*\$J\$3	=K29*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3
30	25	=ECARTYPE(B30:G30)	=MOYENNE(\$J\$6:\$J\$30)	=K30*\$J\$3	=K30*\$J\$2	=0.1/3*\$J\$3	=0.1/3*\$J\$2	=0.1/3

Figure: 74.75 - Formules Microsoft Excel 14.0.6123 explicites du tableau précédent

Nous avons alors la carte de contrôle suivante:

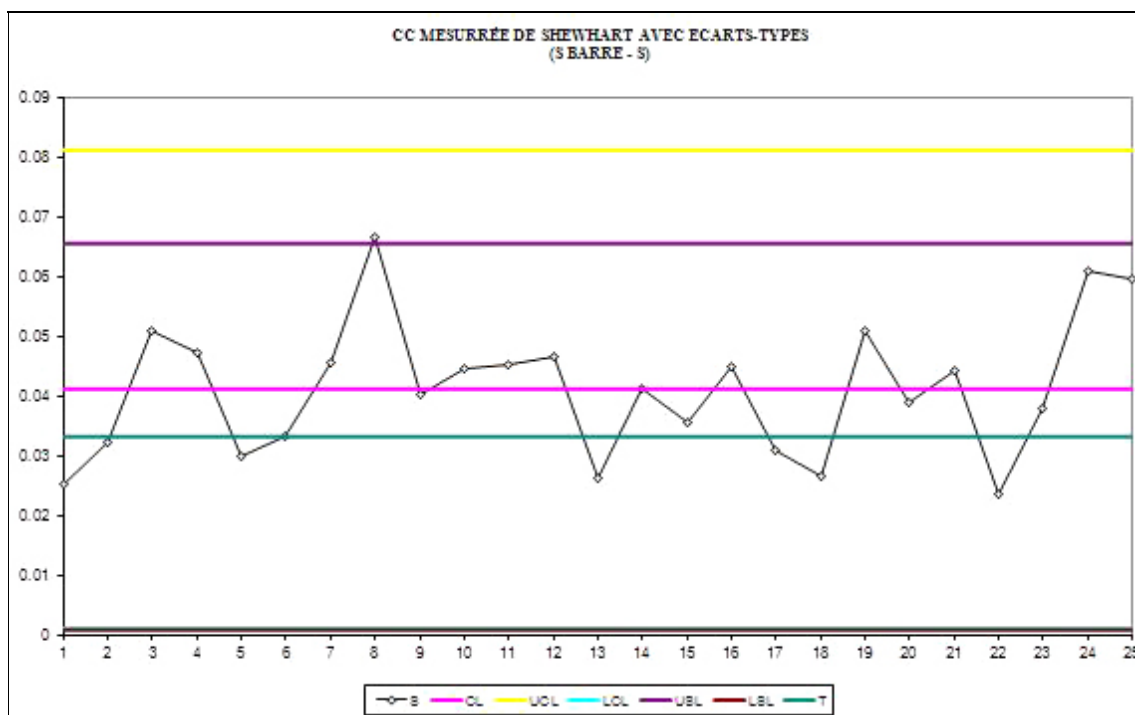


Figure: 74.76 - Carte de contrôle S Barre - S

où nous avons pris pour le client (qui avait imposé un USL et LSL de 10 ± 0.1):

$$\begin{aligned}
 USL &= 0.0\bar{3} \cdot B_4(n_i) = 0.0\bar{3} \cdot B_4(6) = 0.0657 \\
 T &= 0.0\bar{3} \\
 LSL &= 0.0\bar{3} \cdot B_3(n_i) = 0.0\bar{3} \cdot B_3(6) = 0.001
 \end{aligned}
 \tag{74.483}$$

À remarquer que cette carte de contrôle aurait selon les règles de la WECO zéro points au-delà des $\pm 3\sigma$ (nous indiquons cela pour des raisons de comparaison avec les cartes précédentes et les prochaines). Elle est donc moins sensible que les autres cartes. Il faut remarquer aussi que les limites UCL et LCL sont plus larges que les USL et LSL imposées.

Pour rappel, le problème majeur de cette carte de contrôle est qu'elle n'est juste que si et seulement si les données sont normalement distribuées et que Shewart suppose la loi de S symétrique alors que ce n'est pas le cas. Dès lors, les responsables qualités associent alors avec la carte de la moyenne une autre carte que nous allons présenter de suite.

5.4.5. CC AVEC MOYENNES BASÉES SUR L'ÉCART-TYPE (X BARRE - S)

Lors de notre étude de la carte de contrôle aux moyennes avec les limites fixes basées sur l'écart-type de la moyenne, nous avons bien mis en évidence que cette dernière souffrait de l'handicap que $\sigma_{\bar{X}}$ est trop approximatif dans la pratique (car n est souvent trop petit).

Dès lors, un moyen de contourner cela est d'utiliser les résultats obtenus lors de l'étude de la carte de contrôle précédente.

Avantages: Corriger le problème de la carte de contrôle aux moyennes lorsque le nombre d'individus par prélèvements (échantillons) est trop petit. Représente bien la variation du niveau moyen des données (tendance centrale).

Problèmes: La démonstration mathématique de l'origine des limites de contrôles est ardue (voir chapitre de Statistiques pour le détail des démonstrations). Un autre problème c'est que cette carte ne fonctionne que si et seulement si les données sont identiquement distribuées et le sont selon une loi Normale! Par ailleurs puisque les limites sont basées sur l'estimateur de l'écart-type, il faudrait un n qui dans l'idée ne soit pas en dessous de la dizaine...

Rappelons d'abord que nous avons:

$$\begin{aligned}
 USL &= \text{imposé par client/direction} \\
 UCL_{\bar{X}} &= \bar{\bar{X}} + 3\sigma_{\bar{X}} = \bar{\bar{X}} + 3 \frac{\sigma_{\bar{X}}}{\sqrt{n_i}} \\
 CL_{\bar{X}} &= E(\bar{X}) = \bar{\bar{X}} \\
 T &= \text{objectif imposé par client/direction} \\
 LCL_{\bar{X}} &= \bar{\bar{X}} - 3\sigma_{\bar{X}} = \bar{\bar{X}} - 3 \frac{\sigma_{\bar{X}}}{\sqrt{n_i}} \\
 LSL &= \text{imposé par client/direction}
 \end{aligned} \tag{74.484}$$

et nous avons démontré plus haut que:

$$\bar{S} = \sigma_X c_4(n_i) \tag{74.485}$$

Nous avons donc:

USL = imposé par client/direction

$$UCL_{\bar{X}(s)} = \bar{\bar{X}} + 3 \frac{\sigma_X}{\sqrt{n_i}} = \bar{\bar{X}} + 3 \frac{\bar{S}}{c_4(n_i)\sqrt{n_i}}$$

$$CL_{\bar{X}(s)} = \bar{\bar{X}}$$

(74.486)

T = objectif imposé par client/direction

$$LCL_{\bar{X}(s)} = \bar{\bar{X}} - 3 \frac{\sigma_X}{\sqrt{n_i}} = \bar{\bar{X}} - 3 \frac{\bar{S}}{c_4(n_i)\sqrt{n_i}}$$

LSL = imposé par client/direction

et que nous retrouvons parfois dans la littérature sous la forme:

USL = imposé par client/direction

$$UCL_{\bar{X}(s)} = \bar{\bar{X}} + 3 \frac{\bar{S}}{c_4(n_i)\sqrt{n_i}} = \bar{\bar{X}} + A_3(n_i)\bar{S}$$

$$CL_{\bar{X}(s)} = \bar{\bar{X}}$$

(74.487)

T = objectif imposé par client/direction

$$LCL_{\bar{X}(s)} = \bar{\bar{X}} - 3 \frac{\bar{S}}{c_4(n_i)\sqrt{n_i}} = \bar{\bar{X}} - A_3(n_i)\bar{S}$$

LSL = imposé par client/direction

Remarquons que nous pouvons alors calculer des probabilités de sous-groupes non-conformes par excès ou par défaut en utilisant le fait que $\bar{X}$ suit une loi Normale de paramètres:

$$N(\mu, \sigma^2) \cong N\left(\bar{\bar{X}}, \frac{\bar{S}^2}{c_4^2(n)n}\right) \quad (74.488)$$

dont il découle immédiatement que X suit une loi Normale de paramètres:

$$N(\mu, \sigma^2) \cong N\left(X, \frac{\bar{S}^2}{c_4^2(n)}\right) \quad (74.489)$$

Ainsi, si nous considérons toujours le tableau suivant de 25 prélèvements (k) de 6 individus (n_i):

	A	B	C	D	E	F	G
4		Ech. 1	Ech. 2	Ech. 3	Ech. 4	Ech. 5	Ech. 6
5	Jour 1	9.9943	10.0196	9.9732	10.0088	9.9685	9.9544
6	Jour 2	9.9721	9.9643	9.9426	9.9712	10.0259	10.0177
7	Jour 3	9.9788	10.0213	9.9407	10.0696	10.0161	10.0709
8	Jour 4	10.0658	10.0800	9.9514	9.9958	10.0338	10.0422
9	Jour 5	10.0136	9.9791	9.9449	9.9403	9.9424	9.9421
10	Jour 6	9.9495	10.0057	9.9715	10.0388	10.0019	9.9616
11	Jour 7	9.9144	10.0321	10.0361	10.0253	10.0109	9.9935
12	Jour 8	9.9748	9.9962	10.0369	10.1500	9.9996	9.9741
13	Jour 9	10.0063	10.0749	9.9971	9.9780	9.9934	9.9555
14	Jour 10	10.0840	9.9619	10.0063	10.0364	9.9907	9.9762
15	Jour 11	9.9436	10.0638	9.9617	10.0138	9.9672	9.9591
16	Jour 12	10.0074	9.9817	9.9753	10.0541	9.9845	9.9107
17	Jour 13	10.0069	10.0282	10.0385	9.9747	10.0323	9.9860
18	Jour 14	9.9836	10.0630	9.9824	9.9909	9.9684	10.0575
19	Jour 15	10.0317	10.0485	9.9659	10.0675	10.0047	10.0277
20	Jour 16	9.9756	9.9174	9.9076	9.9909	10.0180	9.9979
21	Jour 17	9.9818	9.9962	9.9614	9.9612	9.9694	9.9349
22	Jour 18	10.0168	10.0043	10.0029	9.9549	10.0199	9.9691
23	Jour 19	9.9857	10.0503	10.0381	9.9811	9.9603	9.9123
24	Jour 20	9.9878	9.9841	9.9496	9.9754	9.9532	10.0576
25	Jour 21	10.0657	9.9838	9.9636	9.9779	10.0423	10.0547
26	Jour 22	9.9564	10.0184	9.9890	9.9999	10.0146	10.0145
27	Jour 23	10.0611	9.9443	9.9899	9.9833	9.9947	10.0033
28	Jour 24	9.9054	9.9049	9.9870	10.0580	9.9156	9.9782
29	Jour 25	9.9768	9.9954	9.9196	9.9341	9.9162	10.0721

Figure: 74.77 - Mesures journalières avec 6 individus sans arrêt de la machine

Nous pouvons calculer, ou trouver dans les tables que:

$$c_4(n_i) = c_4(6) = 0.9515, A_3(n_i) = A_3(6) = 0.630 \quad (74.490)$$

Nous avons alors le tableau suivant:

	I	J	K	L	M	N	O	P	Q
1	c4	0.952							
2	a3	1.287							
3									
4	N°	Moyennes	CL	UCL	LCL	S	USL	LSL	T
5	1	9.9865	9.9933	10.0464	9.9402	0.0252	10.0429	9.9571	10.0000
6	2	9.9823	9.9933	10.0464	9.9402	0.0325	10.0429	9.9571	10.0000
7	3	10.0162	9.9933	10.0464	9.9402	0.0509	10.0429	9.9571	10.0000
8	4	10.0282	9.9933	10.0464	9.9402	0.0475	10.0429	9.9571	10.0000
9	5	9.9604	9.9933	10.0464	9.9402	0.0299	10.0429	9.9571	10.0000
10	6	9.9882	9.9933	10.0464	9.9402	0.0333	10.0429	9.9571	10.0000
11	7	10.0021	9.9933	10.0464	9.9402	0.0457	10.0429	9.9571	10.0000
12	8	10.0219	9.9933	10.0464	9.9402	0.0668	10.0429	9.9571	10.0000
13	9	10.0009	9.9933	10.0464	9.9402	0.0404	10.0429	9.9571	10.0000
14	10	10.0093	9.9933	10.0464	9.9402	0.0447	10.0429	9.9571	10.0000
15	11	9.9849	9.9933	10.0464	9.9402	0.0454	10.0429	9.9571	10.0000
16	12	9.9856	9.9933	10.0464	9.9402	0.0467	10.0429	9.9571	10.0000
17	13	10.0111	9.9933	10.0464	9.9402	0.0263	10.0429	9.9571	10.0000
18	14	10.0076	9.9933	10.0464	9.9402	0.0414	10.0429	9.9571	10.0000
19	15	10.0243	9.9933	10.0464	9.9402	0.0355	10.0429	9.9571	10.0000
20	16	9.9679	9.9933	10.0464	9.9402	0.0451	10.0429	9.9571	10.0000
21	17	9.9508	9.9933	10.0464	9.9402	0.0309	10.0429	9.9571	10.0000
22	18	9.9947	9.9933	10.0464	9.9402	0.0265	10.0429	9.9571	10.0000
23	19	9.9880	9.9933	10.0464	9.9402	0.0509	10.0429	9.9571	10.0000
24	20	9.9846	9.9933	10.0464	9.9402	0.0391	10.0429	9.9571	10.0000
25	21	10.0147	9.9933	10.0464	9.9402	0.0445	10.0429	9.9571	10.0000
26	22	9.9988	9.9933	10.0464	9.9402	0.0236	10.0429	9.9571	10.0000
27	23	9.9961	9.9933	10.0464	9.9402	0.0379	10.0429	9.9571	10.0000
28	24	9.9582	9.9933	10.0464	9.9402	0.0610	10.0429	9.9571	10.0000
29	25	9.9690	9.9933	10.0464	9.9402	0.0597	10.0429	9.9571	10.0000

Figure: 74.78 - Indicateurs correspondants aux mesures obtenues

Soit:

	I	J	K	L	M	N
1	c4	=RACINE(2/(NBVAL(B5:G5)-1))*EXP(LNGAMMA(NBVAL(B5:G5)/2))/EXP(LNGAMMA((NBVAL(B5:G5)-1)/2))				
2	a3	=3/(J*I*RACINE(NBVAL(B5:G5)))				
3						
4	N	Moyennes	CL	S	UCL	LCL
5	1	=MOYENNE(B5:G5)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B5:G5)	=\$K5+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K5-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
6	2	=MOYENNE(B6:G6)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B6:G6)	=\$K6+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K6-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
7	3	=MOYENNE(B7:G7)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B7:G7)	=\$K7+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K7-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
8	4	=MOYENNE(B8:G8)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B8:G8)	=\$K8+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K8-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
9	5	=MOYENNE(B9:G9)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B9:G9)	=\$K9+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K9-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
10	6	=MOYENNE(B10:G10)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B10:G10)	=\$K10+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K10-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
11	7	=MOYENNE(B11:G11)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B11:G11)	=\$K11+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K11-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
12	8	=MOYENNE(B12:G12)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B12:G12)	=\$K12+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K12-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
13	9	=MOYENNE(B13:G13)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B13:G13)	=\$K13+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K13-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
14	10	=MOYENNE(B14:G14)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B14:G14)	=\$K14+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K14-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
15	11	=MOYENNE(B15:G15)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B15:G15)	=\$K15+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K15-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
16	12	=MOYENNE(B16:G16)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B16:G16)	=\$K16+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K16-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
17	13	=MOYENNE(B17:G17)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B17:G17)	=\$K17+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K17-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
18	14	=MOYENNE(B18:G18)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B18:G18)	=\$K18+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K18-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
19	15	=MOYENNE(B19:G19)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B19:G19)	=\$K19+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K19-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
20	16	=MOYENNE(B20:G20)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B20:G20)	=\$K20+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K20-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
21	17	=MOYENNE(B21:G21)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B21:G21)	=\$K21+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K21-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
22	18	=MOYENNE(B22:G22)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B22:G22)	=\$K22+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K22-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
23	19	=MOYENNE(B23:G23)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B23:G23)	=\$K23+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K23-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
24	20	=MOYENNE(B24:G24)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B24:G24)	=\$K24+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K24-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
25	21	=MOYENNE(B25:G25)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B25:G25)	=\$K25+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K25-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
26	22	=MOYENNE(B26:G26)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B26:G26)	=\$K26+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K26-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
27	23	=MOYENNE(B27:G27)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B27:G27)	=\$K27+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K27-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
28	24	=MOYENNE(B28:G28)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B28:G28)	=\$K28+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K28-MOYENNE(\$L\$5:\$L\$29)*\$J\$2
29	25	=MOYENNE(B29:G29)	=MOYENNE(\$J\$5:\$J\$29)	=ECARTYPE(B29:G29)	=\$K29+MOYENNE(\$L\$5:\$L\$29)*\$J\$2	=\$K29-MOYENNE(\$L\$5:\$L\$29)*\$J\$2

Figure: 74.79 - Formules Microsoft Excel 14.0.6123 explicites du tableau précédent

Nous avons alors la carte de contrôle suivante:

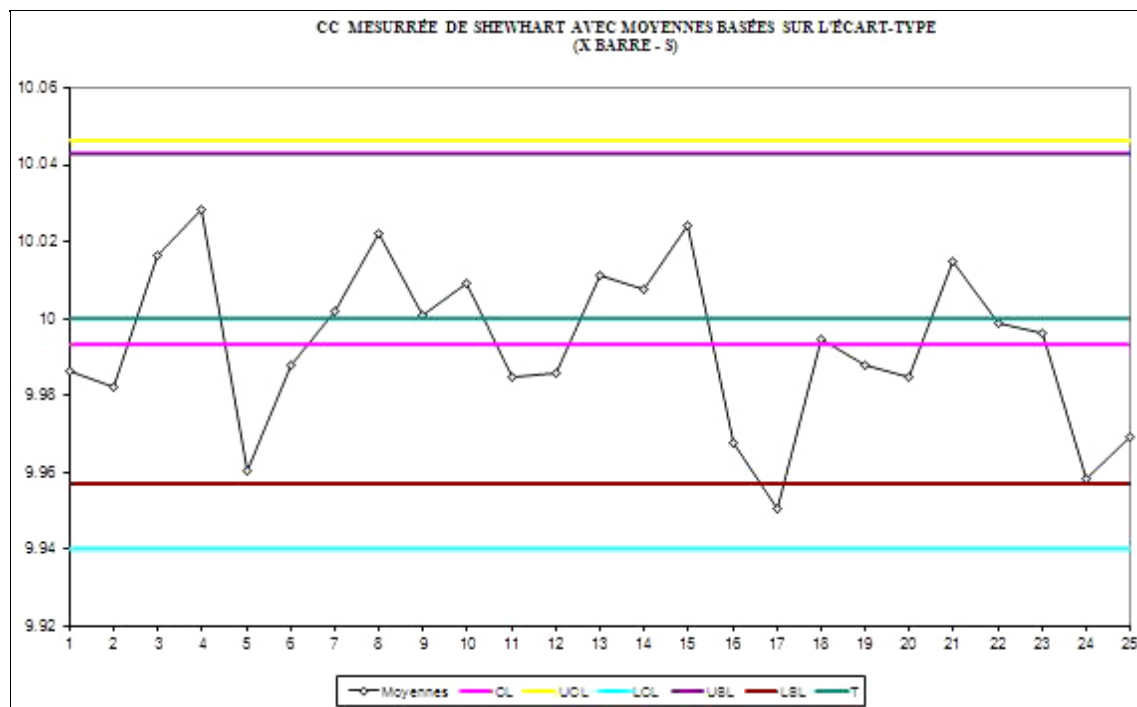


Figure: 74.80 - Carte de contrôle X barre - S

où nous avons pris pour le client (qui avait imposé un USL et LSL de 10 ± 0.1):

$$\begin{aligned}
 USL &= 10 + 0.03\bar{A}_3(n_i) = 10 + 0.03\bar{A}_3(6) = 10.021 \\
 T &= 10 \\
 LSL &= 10 - 0.03\bar{A}_3(n_i) = 10 - 0.03\bar{A}_3(6) = 9.979
 \end{aligned}
 \tag{74.491}$$

À remarquer que cette carte de contrôle aurait selon les règles de la WECO six points au-delà des $\pm 3\sigma$ (nous indiquons cela pour des raisons de comparaison avec les cartes précédentes et les prochaines). Elle est donc plus sensible que la carte précédente et tout aussi sensible que la carte aux moyennes basée sur l'erreur-standard. À remarquer aussi que les limites UCL et LCL sont plus larges que les limites USL et LSL imposées.

5.4.6. CC AVEC ÉTENDUES (R BARRE - R)

La carte de contrôle aux étendues est donc souvent associée à la carte de contrôle de la moyenne. Elle a pour objectif de remplacer avantageusement la carte de contrôle de l'écart-type S lorsque les données ne sont pas normalement distribuées, car sa construction mathématique permet si on le désire de choisir la loi de probabilité sous-jacente. Ainsi, nous utilisons cette nouvelle carte lorsque $n_i \leq 10$ (le nombre d'individus par prélèvement), alors que pour la carte de l'écart-type il faut plutôt que $n_i \geq 10$.

Avantages: Complète très bien la carte de contrôle de la moyenne et remplace avantageusement la carte de l'écart-type lorsque le nombre de prélèvements est peu élevé. Représente bien la variation à l'intérieur même des données de chaque prélèvement (dispersion instantanée) que nous pourrions détecter.

Problèmes: La démonstration mathématique de l'origine des limites de contrôles est ardue (voir chapitre de Statistiques pour le détail des démonstrations) et la détermination des coefficients nécessite l'utilisation de calculs par la méthode de Monte-Carlo (cf. chapitre de Méthodes numériques). Un autre problème, c'est que bien que la loi de probabilité de l'étendue puisse être choisie, elle est souvent donnée dans les livres spécialisés que pour la loi Normale...

L'idée est de représenter (il est impossible que le client fixe une limite USL , LSL ou une cible T en ce qui concerne l'étendue d'où le fait que ces paramètres ne soient cette fois-ci pas mentionnés):

$$\begin{aligned} UCL_R &= \bar{R} + 3\sigma_R \\ CL_R &= E(R) = \bar{R} \\ LCL_R &= \bar{R} - 3\sigma_R \end{aligned} \quad (74.492)$$

et cette fois-ci nous n'allons pas utiliser les estimateurs de l'espérance et de l'écart-type car pour $n \leq 10$ ils ne convergent pas assez vite. Dès lors, nous allons utiliser les résultats démontrés dans le chapitre de Statistiques lors de notre étude des statistiques d'ordre. Nous avons obtenu comme estimateur de l'écart-type:

$$\sigma_R = \frac{d_3(n)\bar{R}}{d_2(n)} = \frac{d_3\bar{R}}{d_2} \quad (74.493)$$

avec:

$$\bar{R} = \frac{1}{k} \sum_{i=1}^k R_i \quad (74.494)$$

ou selon les logiciels (mais cette dernière est plus logique quand on y réfléchit):

$$\bar{R} = \frac{1}{k-1} \sum_{i=1}^k R_i \quad (74.495)$$

Dès lors:

$$\begin{aligned}
 UCL_R &= \bar{R} + 3\sigma_R = \bar{R} + 3 \frac{d_3(n_i)\bar{R}}{d_2(n_i)} = \bar{R} \left(1 + 3 \frac{d_3(n_i)}{d_2(n_i)} \right) \\
 CL_R &= E(R) = \bar{R} \\
 LCL_R &= \bar{R} - 3\sigma_R = \bar{R} - 3 \frac{d_3(n_i)\bar{R}}{d_2(n_i)} = \bar{R} \left(1 - 3 \frac{d_3(n_i)}{d_2(n_i)} \right)
 \end{aligned}
 \tag{74.496}$$

ce qui se note traditionnellement:

$$\begin{aligned}
 UCL_R &= \bar{R} + 3\sigma_R = \bar{R} + 3 \frac{d_3(n_i)\bar{R}}{d_2(n_i)} = \bar{R} \cdot D_4(n_i) \\
 CL_R &= E(R) = \bar{R} \\
 LCL_R &= \bar{R} - 3\sigma_R = \bar{R} - 3 \frac{d_3(n_i)\bar{R}}{d_2(n_i)} = \bar{R} \cdot D_3(n_i)
 \end{aligned}
 \tag{74.497}$$

Les coefficients D_3, D_4 sont disponibles dans la littérature (ou sur le web) sous forme de tables en fonction du nombre n_i d'individus par prélèvements et souvent (malheureusement) que pour une distribution normale (voir l'exemple dans le chapitre de Statistiques).

Ainsi, si nous considérons toujours le tableau suivant de 25 prélèvements (k) de 6 individus (n):

	A	B	C	D	E	F	G
4		Ech. 1	Ech. 2	Ech. 3	Ech. 4	Ech. 5	Ech. 6
5	Jour 1	9.3943	10.0136	9.3732	10.0088	9.3685	9.3544
6	Jour 2	9.3721	9.3643	9.3426	9.3712	10.0259	10.0177
7	Jour 3	9.3788	10.0213	9.3407	10.0636	10.0161	10.0709
8	Jour 4	10.0658	10.0800	9.3514	9.3958	10.0338	10.0422
9	Jour 5	10.0136	9.3791	9.3449	9.3403	9.3424	9.3421
10	Jour 6	9.3495	10.0057	9.3715	10.0388	10.0019	9.3616
11	Jour 7	9.3144	10.0321	10.0361	10.0253	10.0109	9.3935
12	Jour 8	9.3748	9.3962	10.0369	10.1500	9.3996	9.3741
13	Jour 9	10.0063	10.0749	9.3971	9.3780	9.3934	9.3555
14	Jour 10	10.0840	9.3619	10.0063	10.0364	9.3907	9.3762
15	Jour 11	9.3436	10.0638	9.3617	10.0138	9.3672	9.3531
16	Jour 12	10.0074	9.3817	9.3753	10.0541	9.3845	9.3107
17	Jour 13	10.0069	10.0282	10.0385	9.3747	10.0323	9.3860
18	Jour 14	9.3836	10.0630	9.3824	9.3909	9.3684	10.0575
19	Jour 15	10.0317	10.0485	9.3659	10.0675	10.0047	10.0277
20	Jour 16	9.3756	9.3174	9.3076	9.3909	10.0180	9.3979
21	Jour 17	9.3818	9.3962	9.3614	9.3612	9.3634	9.3349
22	Jour 18	10.0168	10.0043	10.0029	9.3549	10.0199	9.3631
23	Jour 19	9.3857	10.0503	10.0381	9.3811	9.3603	9.3123
24	Jour 20	9.3878	9.3841	9.3496	9.3754	9.3532	10.0576
25	Jour 21	10.0657	9.3838	9.3636	9.3779	10.0423	10.0547
26	Jour 22	9.3564	10.0184	9.3890	9.3999	10.0146	10.0145
27	Jour 23	10.0611	9.3443	9.3899	9.3833	9.3947	10.0033
28	Jour 24	9.3054	9.3049	9.3870	10.0580	9.3156	9.3782
29	Jour 25	9.3768	9.3954	9.3196	9.3341	9.3162	10.0721

Figure: 74.81 - Mesures journalières avec 6 individus sans arrêt de la machine

Puisque $n = 6$ nous trouvons dans les tables:

$$D_3 = 0, D_4 = 2.004 \tag{74.498}$$

Nous avons alors le tableau suivant:

	J	K	L	M	N
1	D3	0			
2	D4	2.004			
3					
4	N°	R	CL	UCL	LCL
5	1	0.0652	0.1032	0.2188	-
6	2	0.0833	0.1032	0.2188	-
7	3	0.1302	0.1032	0.2188	-
8	4	0.1286	0.1032	0.2188	-
9	5	0.0733	0.1032	0.2188	-
10	6	0.0893	0.1032	0.2188	-
11	7	0.1217	0.1032	0.2188	-
12	8	0.1759	0.1032	0.2188	-
13	9	0.1194	0.1032	0.2188	-
14	10	0.1221	0.1032	0.2188	-
15	11	0.1202	0.1032	0.2188	-
16	12	0.1434	0.1032	0.2188	-
17	13	0.0638	0.1032	0.2188	-
18	14	0.0346	0.1032	0.2188	-
19	15	0.1016	0.1032	0.2188	-
20	16	0.1104	0.1032	0.2188	-
21	17	0.0856	0.1032	0.2188	-
22	18	0.0650	0.1032	0.2188	-
23	19	0.1380	0.1032	0.2188	-
24	20	0.1080	0.1032	0.2188	-
25	21	0.1021	0.1032	0.2188	-
26	22	0.0620	0.1032	0.2188	-
27	23	0.1168	0.1032	0.2188	-
28	24	0.1531	0.1032	0.2188	-
29	25	0.1559	0.1032	0.2188	-

Figure: 74.82 - Indicateurs correspondants aux mesures obtenues

Soit:

	I	J	K	L	M
1	D3	0			
2	D4	2.004			
3					
4	N°	R	CL	UCL	LCL
5	1	= (MAX(B5:G5)-MIN(B5:G5))	=MOYENNE(\$J\$5:\$J\$29)	= \$K5*\$J\$2	= \$K5*\$J\$1
6	2	= (MAX(B6:G6)-MIN(B6:G6))	=MOYENNE(\$J\$5:\$J\$29)	= \$K6*\$J\$2	= \$K6*\$J\$1
7	3	= (MAX(B7:G7)-MIN(B7:G7))	=MOYENNE(\$J\$5:\$J\$29)	= \$K7*\$J\$2	= \$K7*\$J\$1
8	4	= (MAX(B8:G8)-MIN(B8:G8))	=MOYENNE(\$J\$5:\$J\$29)	= \$K8*\$J\$2	= \$K8*\$J\$1
9	5	= (MAX(B9:G9)-MIN(B9:G9))	=MOYENNE(\$J\$5:\$J\$29)	= \$K9*\$J\$2	= \$K9*\$J\$1
10	6	= (MAX(B10:G10)-MIN(B10:G10))	=MOYENNE(\$J\$5:\$J\$29)	= \$K10*\$J\$2	= \$K10*\$J\$1
11	7	= (MAX(B11:G11)-MIN(B11:G11))	=MOYENNE(\$J\$5:\$J\$29)	= \$K11*\$J\$2	= \$K11*\$J\$1
12	8	= (MAX(B12:G12)-MIN(B12:G12))	=MOYENNE(\$J\$5:\$J\$29)	= \$K12*\$J\$2	= \$K12*\$J\$1
13	9	= (MAX(B13:G13)-MIN(B13:G13))	=MOYENNE(\$J\$5:\$J\$29)	= \$K13*\$J\$2	= \$K13*\$J\$1
14	10	= (MAX(B14:G14)-MIN(B14:G14))	=MOYENNE(\$J\$5:\$J\$29)	= \$K14*\$J\$2	= \$K14*\$J\$1
15	11	= (MAX(B15:G15)-MIN(B15:G15))	=MOYENNE(\$J\$5:\$J\$29)	= \$K15*\$J\$2	= \$K15*\$J\$1
16	12	= (MAX(B16:G16)-MIN(B16:G16))	=MOYENNE(\$J\$5:\$J\$29)	= \$K16*\$J\$2	= \$K16*\$J\$1
17	13	= (MAX(B17:G17)-MIN(B17:G17))	=MOYENNE(\$J\$5:\$J\$29)	= \$K17*\$J\$2	= \$K17*\$J\$1
18	14	= (MAX(B18:G18)-MIN(B18:G18))	=MOYENNE(\$J\$5:\$J\$29)	= \$K18*\$J\$2	= \$K18*\$J\$1
19	15	= (MAX(B19:G19)-MIN(B19:G19))	=MOYENNE(\$J\$5:\$J\$29)	= \$K19*\$J\$2	= \$K19*\$J\$1
20	16	= (MAX(B20:G20)-MIN(B20:G20))	=MOYENNE(\$J\$5:\$J\$29)	= \$K20*\$J\$2	= \$K20*\$J\$1
21	17	= (MAX(B21:G21)-MIN(B21:G21))	=MOYENNE(\$J\$5:\$J\$29)	= \$K21*\$J\$2	= \$K21*\$J\$1
22	18	= (MAX(B22:G22)-MIN(B22:G22))	=MOYENNE(\$J\$5:\$J\$29)	= \$K22*\$J\$2	= \$K22*\$J\$1
23	19	= (MAX(B23:G23)-MIN(B23:G23))	=MOYENNE(\$J\$5:\$J\$29)	= \$K23*\$J\$2	= \$K23*\$J\$1
24	20	= (MAX(B24:G24)-MIN(B24:G24))	=MOYENNE(\$J\$5:\$J\$29)	= \$K24*\$J\$2	= \$K24*\$J\$1
25	21	= (MAX(B25:G25)-MIN(B25:G25))	=MOYENNE(\$J\$5:\$J\$29)	= \$K25*\$J\$2	= \$K25*\$J\$1
26	22	= (MAX(B26:G26)-MIN(B26:G26))	=MOYENNE(\$J\$5:\$J\$29)	= \$K26*\$J\$2	= \$K26*\$J\$1
27	23	= (MAX(B27:G27)-MIN(B27:G27))	=MOYENNE(\$J\$5:\$J\$29)	= \$K27*\$J\$2	= \$K27*\$J\$1
28	24	= (MAX(B28:G28)-MIN(B28:G28))	=MOYENNE(\$J\$5:\$J\$29)	= \$K28*\$J\$2	= \$K28*\$J\$1
29	25	= (MAX(B29:G29)-MIN(B29:G29))	=MOYENNE(\$J\$5:\$J\$29)	= \$K29*\$J\$2	= \$K29*\$J\$1

Figure: 74.83 - Formules Microsoft Excel 14.0.6123 explicites du tableau précédent

Nous avons alors la carte de contrôle suivante:

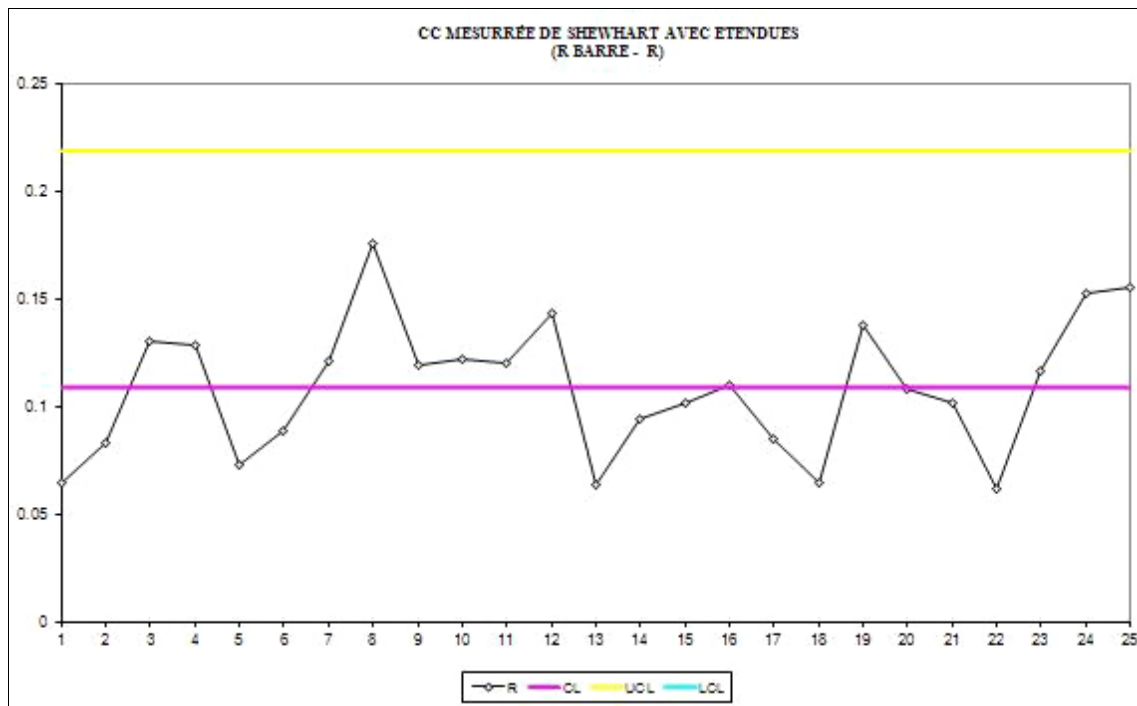


Figure: 74.84 - Carte de contrôle R barre - R

qui (est-il nécessaire de le rappeler?) sous-estime donc le nombre de détections d'erreurs puisque la loi de probabilité de l'étendue n'est en réalité pas symétrique (donc l'application des limites fixes de Shewhart n'est pas vraiment adaptée...) et parce que la quasi-totalité des tables supposent quand même les données distribuées normalement aussi.

À remarquer que cette carte de contrôle aurait selon les règles de la WECO zéro points au-delà des $\pm 3\sigma$ (nous indiquons cela pour des raisons de comparaison avec les cartes précédentes et les prochaines).

5.4.7. CC AVEC MOYENNES BASÉES SUR L'ÉTENDUE (X BARRE-R)

Lors de notre étude de la carte de contrôle aux moyennes avec les limites fixes basées sur l'écart-type de la moyenne, nous avons bien mis en évidence que cette dernière souffrait de l'handicap que $\sigma_{\bar{x}}$ est trop approximatif dans la pratique (car n est trop petit).

Dès lors, un moyen de contourner cela est d'utiliser les résultats obtenus lors de l'étude de la carte de contrôle précédente.

Avantages: Corriger le problème de la carte de contrôle aux moyennes avec l'écart-type lorsque le nombre d'individus par prélèvements (échantillons) est trop petit. Représente bien la variation du niveau moyen des données (tendance centrale).

Problèmes: La démonstration mathématique de l'origine des limites de contrôles est ardue (voir chapitre de Statistiques pour le détail des démonstrations) et la détermination des coefficients nécessite l'utilisation de calculs par la méthode de Monte-Carlo (cf. chapitre de Méthodes numériques). Un autre problème, c'est que bien que la loi de probabilité de l'étendue puisse être choisie, elle est souvent donnée dans les livres spécialisés que pour la loi Normale

Remarque: Cette carte de contrôle est très utilisée dans le cadre de l'étude de la reproductibilité et de la répétabilité (Étude R&R). Raison pour laquelle on la retrouve dans les logiciels de maîtrise statistique des processus automatiquement lorsque l'on lance une analyse R&R (typiquement ce que fait le logiciel Minitab).

Rappelons d'abord que nous avons:

$$\begin{aligned}
 USL &= \text{imposé par client/direction} \\
 UCL_{\bar{X}} &= \bar{\bar{X}} + 3\sigma_{\bar{X}} = \bar{\bar{X}} + 3 \frac{\sigma_X}{\sqrt{n_i}} \\
 CL_{\bar{X}} &= E(\bar{X}) = \bar{\bar{X}} \\
 T &= \text{objectif imposé par client/direction} \\
 LCL_{\bar{X}} &= \bar{\bar{X}} - 3\sigma_{\bar{X}} = \bar{\bar{X}} - 3 \frac{\sigma_X}{\sqrt{n_i}} \\
 LSL &= \text{imposé par client/direction}
 \end{aligned} \tag{74.499}$$

et nous avons démontré dans le chapitre de Statistique dans le cadre de notre étude des statistiques d'ordre que:

$$\bar{R} = \sigma_X d_2 \tag{74.500}$$

Il vient alors (il est impossible que le client fixe une limite USL , LSL ou une cible T en ce qui concerne l'étendue d'où le fait que ces paramètres ne soient cette fois-ci pas mentionnés):

$$\begin{aligned}
 UCL_{\bar{X}(R)} &= \bar{\bar{X}} + 3 \frac{\sigma_X}{\sqrt{n}} = \bar{\bar{X}} + 3 \frac{\bar{R}}{d_2(n_i)\sqrt{n_i}} \\
 CL_{\bar{X}(R)} &= E(\bar{X}) = \bar{\bar{X}} \\
 LCL_{\bar{X}(R)} &= \bar{\bar{X}} - 3 \frac{\sigma_X}{\sqrt{n}} = \bar{\bar{X}} - 3 \frac{\bar{R}}{d_2(n_i)\sqrt{n_i}}
 \end{aligned} \tag{74.501}$$

toujours avec:

$$\bar{R} = \frac{1}{k} \sum_{i=1}^k R_i \tag{74.502}$$

et l'ensemble se note traditionnellement dans le domaine:

$$\begin{aligned}
 UCL_{\bar{X}(R)} &= \bar{\bar{X}} + 3 \frac{\bar{R}}{d_2(n_i)\sqrt{n_i}} = \bar{\bar{X}} + A_2(n_i)\bar{R} \\
 CL_{\bar{X}(R)} &= E(\bar{X}) = \bar{\bar{X}} \\
 LCL_{\bar{X}(R)} &= \bar{\bar{X}} - 3 \frac{\bar{R}}{d_2(n_i)\sqrt{n_i}} = \bar{\bar{X}} - A_2(n_i)\bar{R}
 \end{aligned} \tag{74.503}$$

Remarquons que nous pouvons alors calculer des probabilités de sous-groupes non-conformes par excès ou par défaut en utilisant le fait que $\bar{X}$ suit une loi Normale de paramètres:

$$N(\mu, \sigma^2) \cong N\left(\bar{X}, \frac{\bar{R}^2}{d_2^2(n)n}\right) \quad (74.504)$$

dont il découle immédiatement que X suit une loi Normale de paramètres:

$$N(\mu, \sigma^2) \cong N\left(\bar{X}, \frac{\bar{R}^2}{d_2^2(n)}\right) \quad (74.505)$$

Ainsi, si nous considérons toujours le tableau suivant de 25 prélèvements (k) de 6 individus (n):

	A	B	C	D	E	F	G
3		Ech. 1	Ech. 2	Ech. 3	Ech. 4	Ech. 5	Ech. 6
4	Jour 1	9.9943	10.0196	9.9732	10.0088	9.9685	9.9544
5	Jour 2	9.9721	9.9643	9.9426	9.9712	10.0259	10.0177
6	Jour 3	9.9788	10.0213	9.9407	10.0696	10.0161	10.0709
7	Jour 4	10.0658	10.0800	9.9514	9.9958	10.0338	10.0422
8	Jour 5	10.0136	9.9791	9.9449	9.9403	9.9424	9.9421
9	Jour 6	9.9495	10.0057	9.9715	10.0388	10.0019	9.9616
10	Jour 7	9.9144	10.0321	10.0361	10.0253	10.0109	9.9935
11	Jour 8	9.9748	9.9962	10.0369	10.1500	9.9996	9.9741
12	Jour 9	10.0063	10.0749	9.9971	9.9780	9.9934	9.9555
13	Jour 10	10.0840	9.9619	10.0063	10.0364	9.9907	9.9762
14	Jour 11	9.9436	10.0638	9.9617	10.0138	9.9672	9.9591
15	Jour 12	10.0074	9.9817	9.9753	10.0541	9.9845	9.9107
16	Jour 13	10.0069	10.0282	10.0385	9.9747	10.0323	9.9860
17	Jour 14	9.9836	10.0630	9.9824	9.9909	9.9684	10.0575
18	Jour 15	10.0317	10.0485	9.9659	10.0675	10.0047	10.0277
19	Jour 16	9.9756	9.9174	9.9076	9.9909	10.0180	9.9979
20	Jour 17	9.9818	9.9962	9.9614	9.9612	9.9694	9.9349
21	Jour 18	10.0168	10.0043	10.0029	9.9549	10.0199	9.9691
22	Jour 19	9.9857	10.0503	10.0381	9.9811	9.9603	9.9123
23	Jour 20	9.9878	9.9841	9.9496	9.9754	9.9532	10.0576
24	Jour 21	10.0657	9.9838	9.9636	9.9779	10.0423	10.0547
25	Jour 22	9.9564	10.0184	9.9890	9.9999	10.0146	10.0145
26	Jour 23	10.0611	9.9443	9.9899	9.9833	9.9947	10.0033
27	Jour 24	9.9054	9.9049	9.9870	10.0580	9.9156	9.9782
28	Jour 25	9.9768	9.9954	9.9196	9.9341	9.9162	10.0721

Figure: 74.85 - Mesures journalières avec 6 individus sans arrêt de la machine

Puisque $n = 6$ nous trouvons dans les tables:

$$A_2(n_i) = A_2(6) = 0.483 \quad (74.506)$$

Nous avons alors le tableau suivant:

	I	J	K	L	M	N
1	A2	0.483				
2						
3	N°	Moyennes	CL	UCL	LCL	R
4	1	9.9865	9.9933	10.0460	9.9406	0.0652
5	2	9.9823	9.9933	10.0460	9.9406	0.0833
6	3	10.0162	9.9933	10.0460	9.9406	0.1302
7	4	10.0282	9.9933	10.0460	9.9406	0.1286
8	5	9.9604	9.9933	10.0460	9.9406	0.0733
9	6	9.9882	9.9933	10.0460	9.9406	0.0893
10	7	10.0021	9.9933	10.0460	9.9406	0.1217
11	8	10.0219	9.9933	10.0460	9.9406	0.1759
12	9	10.0009	9.9933	10.0460	9.9406	0.1194
13	10	10.0093	9.9933	10.0460	9.9406	0.1221
14	11	9.9849	9.9933	10.0460	9.9406	0.1202
15	12	9.9856	9.9933	10.0460	9.9406	0.1434
16	13	10.0111	9.9933	10.0460	9.9406	0.0638
17	14	10.0076	9.9933	10.0460	9.9406	0.0946
18	15	10.0243	9.9933	10.0460	9.9406	0.1016
19	16	9.9679	9.9933	10.0460	9.9406	0.1104
20	17	9.9508	9.9933	10.0460	9.9406	0.0856
21	18	9.9947	9.9933	10.0460	9.9406	0.0650
22	19	9.9880	9.9933	10.0460	9.9406	0.1380
23	20	9.9846	9.9933	10.0460	9.9406	0.1080
24	21	10.0147	9.9933	10.0460	9.9406	0.1021
25	22	9.9988	9.9933	10.0460	9.9406	0.0620
26	23	9.9961	9.9933	10.0460	9.9406	0.1168
27	24	9.9582	9.9933	10.0460	9.9406	0.1531
28	25	9.9690	9.9933	10.0460	9.9406	0.1559

Figure: 74.86 - Indicateurs correspondants aux mesures obtenues

Soit:

	I	J	K	L	M	N
1	A2	0.483				
2						
3	N°	Moyennes	CL	R	UCL	LCL
4	1	=MOYENNE(B4:G4)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B4:G4)-MIN(B4:G4))	=K4+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K4-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
5	2	=MOYENNE(B5:G5)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B5:G5)-MIN(B5:G5))	=K5+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K5-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
6	3	=MOYENNE(B6:G6)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B6:G6)-MIN(B6:G6))	=K6+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K6-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
7	4	=MOYENNE(B7:G7)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B7:G7)-MIN(B7:G7))	=K7+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K7-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
8	5	=MOYENNE(B8:G8)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B8:G8)-MIN(B8:G8))	=K8+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K8-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
9	6	=MOYENNE(B9:G9)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B9:G9)-MIN(B9:G9))	=K9+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K9-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
10	7	=MOYENNE(B10:G10)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B10:G10)-MIN(B10:G10))	=K10+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K10-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
11	8	=MOYENNE(B11:G11)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B11:G11)-MIN(B11:G11))	=K11+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K11-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
12	9	=MOYENNE(B12:G12)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B12:G12)-MIN(B12:G12))	=K12+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K12-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
13	10	=MOYENNE(B13:G13)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B13:G13)-MIN(B13:G13))	=K13+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K13-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
14	11	=MOYENNE(B14:G14)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B14:G14)-MIN(B14:G14))	=K14+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K14-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
15	12	=MOYENNE(B15:G15)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B15:G15)-MIN(B15:G15))	=K15+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K15-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
16	13	=MOYENNE(B16:G16)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B16:G16)-MIN(B16:G16))	=K16+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K16-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
17	14	=MOYENNE(B17:G17)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B17:G17)-MIN(B17:G17))	=K17+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K17-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
18	15	=MOYENNE(B18:G18)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B18:G18)-MIN(B18:G18))	=K18+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K18-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
19	16	=MOYENNE(B19:G19)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B19:G19)-MIN(B19:G19))	=K19+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K19-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
20	17	=MOYENNE(B20:G20)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B20:G20)-MIN(B20:G20))	=K20+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K20-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
21	18	=MOYENNE(B21:G21)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B21:G21)-MIN(B21:G21))	=K21+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K21-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
22	19	=MOYENNE(B22:G22)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B22:G22)-MIN(B22:G22))	=K22+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K22-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
23	20	=MOYENNE(B23:G23)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B23:G23)-MIN(B23:G23))	=K23+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K23-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
24	21	=MOYENNE(B24:G24)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B24:G24)-MIN(B24:G24))	=K24+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K24-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
25	22	=MOYENNE(B25:G25)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B25:G25)-MIN(B25:G25))	=K25+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K25-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
26	23	=MOYENNE(B26:G26)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B26:G26)-MIN(B26:G26))	=K26+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K26-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
27	24	=MOYENNE(B27:G27)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B27:G27)-MIN(B27:G27))	=K27+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K27-\$J\$1*MOYENNE(\$L\$4:\$L\$28)
28	25	=MOYENNE(B28:G28)	=MOYENNE(\$B\$4:\$G\$28)	=(MAX(B28:G28)-MIN(B28:G28))	=K28+\$J\$1*MOYENNE(\$L\$4:\$L\$28)	=K28-\$J\$1*MOYENNE(\$L\$4:\$L\$28)

Figure: 74.87 - Formules Microsoft Excel 14.0.6123 explicites du tableau précédent

Nous avons alors la carte de contrôle suivante:

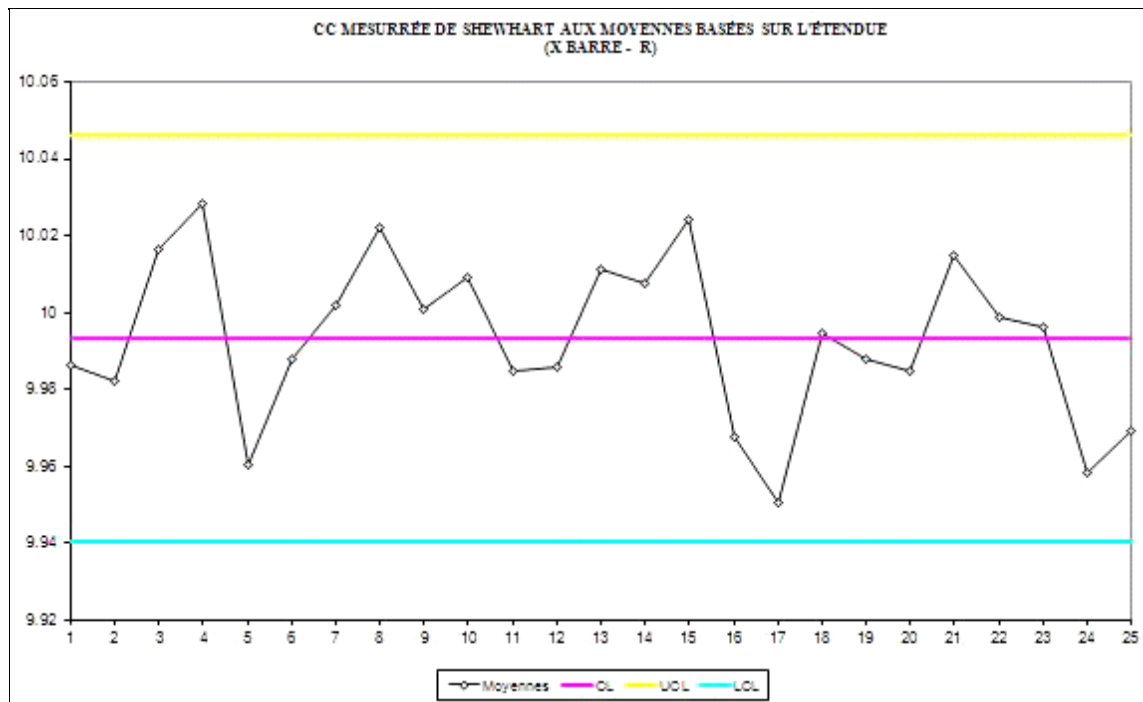


Figure: 74.88 - Carte de contrôle X barre - R

À remarquer que cette carte de contrôle aurait selon les règles de la WECO zéro points au-delà des $\pm 3\sigma$ (nous indiquons cela pour des raisons de comparaison avec les cartes précédentes et les prochaines).

La question est alors de savoir laquelle est la plus juste... Et c'est très délicat, car les limites de Shewhart comme nous l'avons déjà dit de nombreuses fois sont peu réalistes et supposent une distribution symétrique. Or, les calculs mathématiques détaillés du coefficient $A_2(n)$ montrent que les limites ne devraient pas être symétriques. De plus, ce dernier coefficient est basé majoritairement sur une hypothèse de normalité de la distribution des mesures, ce qui lui ajoute une faiblesse supplémentaire. Cependant, si nous comparons à la toute première carte de contrôle avec l'écart-type court terme qui n'avait qu'une seule alarme, la carte de contrôle ci-dessus est plus réaliste que celle avec l'erreur standard qui affiche trop de fausses alertes.

Ces difficultés que nous retrouvons à toutes les cartes de contrôle vues jusqu'à maintenant montrent qu'en fin de compte il faut mieux faire une analyse fréquentielle des mesures, déterminer la loi de probabilité et avec des outils spécialisés basés sur la méthode de Monte-Carlo, calculer les limites supérieures et inférieures pour un $\alpha/2$ donné.

5.5. CARTES DE CONTRÔLES QUANTITATIVES AUTOCORRÉLÉES (AUX MESURES)

La famille des cartes de contrôle autocorrélées est définie par le fait que soit les points sur la carte, soit les limites de contrôles sont calculées sur la base d'un certain nombre de points qui précèdent.

Ces cartes ont pour avantage de nécessiter moins de données que les autres et d'amplifier le phénomène de déviation et donc de détecter les anomalies plus rapidement.

Les règles de la WECO ne s'appliquent pas aux cartes autocorrélées. Si un point est à l'extérieur des limites calculées, il faut une action de correction.

5.5.1. CC A VALEURS INDIVIDUELLES BASÉES SUR L'ÉTENDUE MOBILE (I-EM X BARRE)

En phase de prototypage, nous ne pouvons pas nous permettre de mettre en place le concept de prélèvement. De plus, la carte de contrôle mesurée avec moyenne et écart-type court terme ne permet pas de faire de l'inférence statistique, car la loi de distribution est supposée inconnue car le procédé non stable.

Une solution intermédiaire consiste à considérer alors que le procédé est stable (hypothèse de normalité des données) et de calculer les limites à partir d'un nombre minimal d'individus qu'il est possible d'avoir pour faire une statistique, c'est-à-dire: deux!

Cette situation est donc bien adaptée aux phases de prototypage ou dans le cas de procédés/processus à cadence lente ou dont les mesures (pièces) sont très onéreuses.

Avantages: Complète les premières cartes à valeur individuelle que nous avons vues tout au début pour des situations où il est difficile d'avoir un grand nombre d'individus pour des raisons de coûts. Cette carte de contrôle est parfois considérée comme le couteau suisse du contrôle qualité... et elle est de plus en plus utilisée puisqu'elle correspond à la politique de la production "Just in Time" (JIT) de petites séries (pour la diminution des coûts).

Problèmes: La démonstration mathématique de l'origine des limites de contrôles est ardue (voir chapitre de Statistiques pour le détail des démonstrations) et la détermination des coefficients nécessite l'utilisation de calculs par la méthode de Monte-Carlo (cf. chapitre de Méthodes numériques). Un autre problème, c'est que bien que la loi de probabilité de l'étendue puisse être choisie, elle est souvent donnée dans les livres spécialisés que pour la loi Normale.

Afin de détourner le problème d'une estimation acceptable de l'écart-type, l'idée consiste à considérer deux mesures successives x_{i-1}, x_i comme les variables aléatoires d'une statistique d'ordre dont nous calculons l'étendue en valeur absolue:

$$MR_i = R_i = |x_i - x_{i-1}| = \max(x_{i-1}, x_i) - \min(x_{i-1}, x_i) \quad (74.507)$$

où MR signifie "**Moving Range**" (cette carte de contrôle est appelée parfois "carte-MR(2)"). Et nous avons alors:

$$\overline{MR} = \bar{R} = \frac{1}{n-1} \sum_{i=2}^n |x_i - x_{i-1}| \quad (74.508)$$

puisque nous avons forcément $n-1$ étendues mobiles.

Nous reprenons alors les définitions de la carte de contrôle mesurée avec moyenne et écart-type court terme:

$$\begin{aligned} USL &= \text{imposé par client/direction} \\ UCL_{\bar{X}} &= \bar{\bar{X}} + 3\sigma_{\bar{X}} \\ CL_{\bar{X}} &= E(\bar{X}) = \bar{\bar{X}} \\ T &= \text{objectif imposé par client/direction} \\ LCL_{\bar{X}} &= \bar{\bar{X}} - 3\sigma_{\bar{X}} \\ LSL &= \text{imposé par client/direction} \end{aligned} \quad (74.509)$$

et nous pouvons nous débarrasser du problème de l'estimation de σ_X en utilisant la relation démontrée dans le chapitre de Statistique dans le cadre de l'étude des statistiques d'ordre:

$$\bar{R} = \sigma_X d_2(n_i) \quad (74.510)$$

Ce qui nous amène alors à (il est impossible que le client fixe une limite USL , LSL ou une cible T en ce qui concerne l'étendue d'où le fait que ces paramètres ne soient cette fois-ci pas mentionnés):

$$\begin{aligned} UCL_{\bar{X}(MR)} &= \bar{X} + 3 \frac{\overline{MR}}{d_2(n_i)} \\ CL_{\bar{X}(MR)} &= \bar{X} \\ LCL_{\bar{X}(MR)} &= \bar{X} - 3 \frac{\overline{MR}}{d_2(n_i)} \end{aligned} \quad (74.511)$$

ce qui est parfois noté:

$$\begin{aligned} UCL_{\bar{X}(MR)} &= \bar{X} + E_2 \overline{MR} \\ CL_{\bar{X}(MR)} &= \bar{X} \\ LCL_{\bar{X}(MR)} &= \bar{X} - E_2 \overline{MR} \end{aligned} \quad (74.512)$$

Nous devons prendre dans le cas présent la valeur d_2 pour $n_i = 2$. Les tables nous donnent:

$$d_2(n_i) = d_2(2) = 1.128 \quad (74.513)$$

ce qui nous donne:

$$\begin{aligned} UCL_{\bar{X}(MR)} &= \bar{X} + 2.66 \cdot \overline{MR} \\ CL_{\bar{X}(MR)} &= \bar{X} \\ LCL_{\bar{X}(MR)} &= \bar{X} - 2.66 \cdot \overline{MR} \end{aligned} \quad (74.514)$$

Remarquons que nous pouvons alors calculer des probabilités de sous-groupes non-conformes par excès ou par défaut en utilisant le fait que X suit une loi Normale de paramètres:

$$N(\mu, \sigma^2) \cong N\left(X, \frac{\overline{MR}}{d_2(n)}\right) \quad (74.515)$$

dont il découle immédiatement que X suit une loi Normale de paramètres:

$$N(\mu, \sigma^2) \cong N\left(X, \frac{\bar{S}^2}{c_4^2(n)}\right) \quad (74.516)$$

Reprenons les données suivantes:

	A	B	C	D	E	F	G
3	→						
4	↓	9.9943	10.0196	9.9732	10.0088	9.9685	9.9544
5		9.9721	9.9643	9.9426	9.9712	10.0259	10.0177
6		9.9788	10.0213	9.9407	10.0696	10.0161	10.0709
7		10.0658	10.08	9.9514	9.9958	10.0338	10.0422
8		10.0136	9.9791	9.9449	9.9403	9.9424	9.9421
9		9.9495	10.0057	9.9715	10.0388	10.0019	9.9616
10		9.9144	10.0321	10.0361	10.0253	10.0109	9.9935
11		9.9748	9.9962	10.0369	10.15	9.9996	9.9741
12		10.0063	10.0749	9.9971	9.978	9.9934	9.9555
13		10.084	9.9619	10.0063	10.0364	9.9907	9.9762
14		9.9436	10.0638	9.9617	10.0138	9.9672	9.9591
15		10.0074	9.9817	9.9753	10.0541	9.9845	9.9107
16		10.0069	10.0282	10.0385	9.9747	10.0323	9.986
17		9.9836	10.063	9.9824	9.9909	9.9684	10.0575
18		10.0317	10.0485	9.9659	10.0675	10.0047	10.0277
19		9.9756	9.9174	9.9076	9.9909	10.018	9.9979
20		9.9818	9.8962	9.9614	9.9612	9.9694	9.9349
21		10.0168	10.0043	10.0029	9.9549	10.0199	9.9691
22		9.9857	10.0503	10.0381	9.9811	9.9603	9.9123
23		9.9878	9.9841	9.9496	9.9754	9.9532	10.0576
24		10.0657	9.9838	9.9636	9.9779	10.0423	10.0547
25		9.9564	10.0184	9.989	9.9999	10.0146	10.0145
26		10.0611	9.9443	9.9899	9.9833	9.9947	10.0033
27		9.9054	9.9049	9.987	10.058	9.9156	9.9782
28		9.9768	9.9954	9.9196	9.9341	9.9162	10.0721

Figure: 74.89 - Mesures journalières sans arrêt de la machine

Nous avons alors:

	I	J	K	L	M	N	O	P	Q
1	d2	1.128							
2									
3	N°	Mesures	CL	Etendue Mobile	Moyenne étendue mobile	LCL étendue mobile	UCL étendue mobile	LCL	UCL
4	1	3.9943	3.9933		0.0447	-	0.1635	3.87	10.11
5	2	10.0196	3.9933	0.0253	0.0447	-	0.1635	3.87	10.11
6	3	9.9732	3.9933	0.0464	0.0447	-	0.1635	3.87	10.11
7	4	10.0088	3.9933	0.0356	0.0447	-	0.1635	3.87	10.11
8	5	9.9685	3.9933	0.0403	0.0447	-	0.1635	3.87	10.11
9	6	9.9544	3.9933	0.0141	0.0447	-	0.1635	3.87	10.11
10	7	9.9721	3.9933	0.0177	0.0447	-	0.1635	3.87	10.11
11	8	9.9643	3.9933	0.0078	0.0447	-	0.1635	3.87	10.11
12	9	9.9426	3.9933	0.0217	0.0447	-	0.1635	3.87	10.11
13	10	9.9712	3.9933	0.0286	0.0447	-	0.1635	3.87	10.11
14	11	10.0259	3.9933	0.0547	0.0447	-	0.1635	3.87	10.11
15	12	10.0177	3.9933	0.0082	0.0447	-	0.1635	3.87	10.11
16	13	9.9788	3.9933	0.0389	0.0447	-	0.1635	3.87	10.11
17	14	10.0213	3.9933	0.0425	0.0447	-	0.1635	3.87	10.11
18	15	9.9407	3.9933	0.0806	0.0447	-	0.1635	3.87	10.11
19	16	10.0696	3.9933	0.1289	0.0447	-	0.1635	3.87	10.11
20	17	10.0161	3.9933	0.0535	0.0447	-	0.1635	3.87	10.11
21	18	10.0709	3.9933	0.0548	0.0447	-	0.1635	3.87	10.11
22	19	10.0658	3.9933	0.0051	0.0447	-	0.1635	3.87	10.11
23	20	10.0800	3.9933	0.0142	0.0447	-	0.1635	3.87	10.11
24	21	9.9514	3.9933	0.1286	0.0447	-	0.1635	3.87	10.11
25	22	9.9958	3.9933	0.0444	0.0447	-	0.1635	3.87	10.11
26	23	10.0338	3.9933	0.0380	0.0447	-	0.1635	3.87	10.11
27	24	10.0422	3.9933	0.0084	0.0447	-	0.1635	3.87	10.11
28	25	10.0136	3.9933	0.0286	0.0447	-	0.1635	3.87	10.11
29	26	9.9791	3.9933	0.0345	0.0447	-	0.1635	3.87	10.11
30	27	9.9449	3.9933	0.0342	0.0447	-	0.1635	3.87	10.11
31	28	9.9403	3.9933	0.0046	0.0447	-	0.1635	3.87	10.11
32	29	9.9424	3.9933	0.0021	0.0447	-	0.1635	3.87	10.11
33	30	9.9421	3.9933	0.0003	0.0447	-	0.1635	3.87	10.11
34	31	9.9435	3.9933	0.0074	0.0447	-	0.1635	3.87	10.11
35	32	10.0057	3.9933	0.0562	0.0447	-	0.1635	3.87	10.11
36	33	9.9715	3.9933	0.0342	0.0447	-	0.1635	3.87	10.11
37	34	10.0388	3.9933	0.0673	0.0447	-	0.1635	3.87	10.11
38	35	10.0019	3.9933	0.0369	0.0447	-	0.1635	3.87	10.11
39	36	9.9616	3.9933	0.0403	0.0447	-	0.1635	3.87	10.11
40	37	9.9144	3.9933	0.0472	0.0447	-	0.1635	3.87	10.11
41	38	10.0321	3.9933	0.1177	0.0447	-	0.1635	3.87	10.11
42	39	10.0361	3.9933	0.0040	0.0447	-	0.1635	3.87	10.11
43	40	10.0253	3.9933	0.0108	0.0447	-	0.1635	3.87	10.11
44	41	10.0109	3.9933	0.0144	0.0447	-	0.1635	3.87	10.11
45	42	9.9935	3.9933	0.0174	0.0447	-	0.1635	3.87	10.11
46	43	9.9748	3.9933	0.0187	0.0447	-	0.1635	3.87	10.11
47	44	9.9962	3.9933	0.0214	0.0447	-	0.1635	3.87	10.11
48	45	10.0369	3.9933	0.0407	0.0447	-	0.1635	3.87	10.11

Figure: 74.90 - Indicateurs correspondants aux mesures obtenues

Soit (nous avons coupé la capture d'écran en deux parties faute de place):

	I	J	K	L	M	N
1	d2	1.128				
2						
3	N°	Mesures CL	Etendue Mobile	Moyenne étendue mobile	LCL étendue mobile	
4	1	3.9943	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M4-3/\$J\$1*M4<0;0;M4-3/\$J\$1*M4)	
5	2	10.0196	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M5-3/\$J\$1*M5<0;0;M5-3/\$J\$1*M5)	
6	3	9.9732	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M6-3/\$J\$1*M6<0;0;M6-3/\$J\$1*M6)	
7	4	10.0088	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M7-3/\$J\$1*M7<0;0;M7-3/\$J\$1*M7)	
8	5	9.9685	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M8-3/\$J\$1*M8<0;0;M8-3/\$J\$1*M8)	
9	6	9.9544	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M9-3/\$J\$1*M9<0;0;M9-3/\$J\$1*M9)	
10	7	9.9721	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M10-3/\$J\$1*M10<0;0;M10-3/\$J\$1*M10)	
11	8	9.9643	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M11-3/\$J\$1*M11<0;0;M11-3/\$J\$1*M11)	
12	9	9.9426	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M12-3/\$J\$1*M12<0;0;M12-3/\$J\$1*M12)	
13	10	9.9712	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M13-3/\$J\$1*M13<0;0;M13-3/\$J\$1*M13)	
14	11	10.0259	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M14-3/\$J\$1*M14<0;0;M14-3/\$J\$1*M14)	
15	12	10.0177	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M15-3/\$J\$1*M15<0;0;M15-3/\$J\$1*M15)	
16	13	9.9788	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M16-3/\$J\$1*M16<0;0;M16-3/\$J\$1*M16)	
17	14	10.0213	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M17-3/\$J\$1*M17<0;0;M17-3/\$J\$1*M17)	
18	15	9.9407	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M18-3/\$J\$1*M18<0;0;M18-3/\$J\$1*M18)	
19	16	10.0696	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M19-3/\$J\$1*M19<0;0;M19-3/\$J\$1*M19)	
20	17	10.0161	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M20-3/\$J\$1*M20<0;0;M20-3/\$J\$1*M20)	
21	18	10.0709	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M21-3/\$J\$1*M21<0;0;M21-3/\$J\$1*M21)	
22	19	10.0658	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M22-3/\$J\$1*M22<0;0;M22-3/\$J\$1*M22)	
23	20	10.08	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M23-3/\$J\$1*M23<0;0;M23-3/\$J\$1*M23)	
24	21	9.9514	=MOYENNE(\$J\$4:\$J\$153)	=SOMME(\$L\$4:\$L\$153)/(NB(\$L\$4:\$L\$153)-1)	=SI((M24-3/\$J\$1*M24<0;0;M24-3/\$J\$1*M24)	

Figure: 74.91 - Formules Microsoft Excel 14.0.6123 explicites du tableau précédent

et:

	O	P	Q
1			
2			
3	UCL étendue mobile	LCL	UCL
4	=M4+3/\$J\$1*M4	=K4-3/\$J\$1*M4	=K4+3/\$J\$1*M4
5	=M5+3/\$J\$1*M5	=K5-3/\$J\$1*M5	=K5+3/\$J\$1*M5
6	=M6+3/\$J\$1*M6	=K6-3/\$J\$1*M6	=K6+3/\$J\$1*M6
7	=M7+3/\$J\$1*M7	=K7-3/\$J\$1*M7	=K7+3/\$J\$1*M7
8	=M8+3/\$J\$1*M8	=K8-3/\$J\$1*M8	=K8+3/\$J\$1*M8
9	=M9+3/\$J\$1*M9	=K9-3/\$J\$1*M9	=K9+3/\$J\$1*M9
10	=M10+3/\$J\$1*M10	=K10-3/\$J\$1*M10	=K10+3/\$J\$1*M10
11	=M11+3/\$J\$1*M11	=K11-3/\$J\$1*M11	=K11+3/\$J\$1*M11
12	=M12+3/\$J\$1*M12	=K12-3/\$J\$1*M12	=K12+3/\$J\$1*M12
13	=M13+3/\$J\$1*M13	=K13-3/\$J\$1*M13	=K13+3/\$J\$1*M13
14	=M14+3/\$J\$1*M14	=K14-3/\$J\$1*M14	=K14+3/\$J\$1*M14
15	=M15+3/\$J\$1*M15	=K15-3/\$J\$1*M15	=K15+3/\$J\$1*M15
16	=M16+3/\$J\$1*M16	=K16-3/\$J\$1*M16	=K16+3/\$J\$1*M16
17	=M17+3/\$J\$1*M17	=K17-3/\$J\$1*M17	=K17+3/\$J\$1*M17
18	=M18+3/\$J\$1*M18	=K18-3/\$J\$1*M18	=K18+3/\$J\$1*M18
19	=M19+3/\$J\$1*M19	=K19-3/\$J\$1*M19	=K19+3/\$J\$1*M19
20	=M20+3/\$J\$1*M20	=K20-3/\$J\$1*M20	=K20+3/\$J\$1*M20
21	=M21+3/\$J\$1*M21	=K21-3/\$J\$1*M21	=K21+3/\$J\$1*M21
22	=M22+3/\$J\$1*M22	=K22-3/\$J\$1*M22	=K22+3/\$J\$1*M22
23	=M23+3/\$J\$1*M23	=K23-3/\$J\$1*M23	=K23+3/\$J\$1*M23
24	=M24+3/\$J\$1*M24	=K24-3/\$J\$1*M24	=K24+3/\$J\$1*M24

Figure: 74.92 - Suite des formules Microsoft Excel 14.0.6123 explicites du tableau antérieur

Nous avons alors le graphique:

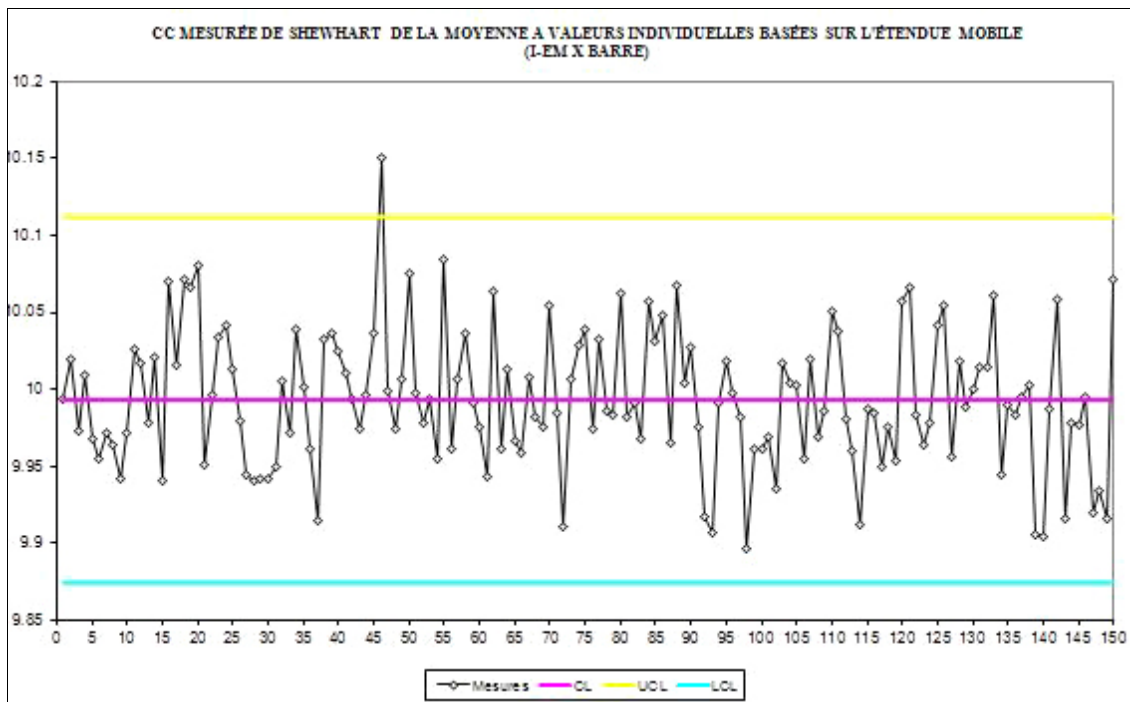


Figure: 74.93 - Carte de contrôle I-EM X barre

À remarquer que cette carte de contrôle aurait selon les règles de la WECO zéro points au-delà des $\pm 3\sigma$ (nous indiquons cela pour des raisons de comparaison avec les cartes précédentes et les prochaines).

5.5.2. CC A VALEURS INDIVIDUELLES AVEC ÉTENDUE MOBILE (I-EM EM BARRE)

La carte précédente utilisant la tendance centrale doit être complétée (au même titre que toutes les autres cartes précédentes à tendance centrale) avec une carte affichant la tendance de la dispersion (les avantages et problèmes de cette carte de contrôle sont les mêmes que la précédente).

Nous reprenons alors les définitions d'une carte de contrôle mesurée pour obtenir:

$$\begin{aligned} UCL_{MR} &= \overline{MR} + 3\sigma_{MR} \\ CL_{MR} &= \overline{MR} \\ LCL_{MR} &= \overline{MR} - 3\sigma_{MR} \end{aligned} \quad (74.517)$$

Encore une fois, l'écart-type de l'étendue mobile (EM) se construit de la même manière que l'écart-type de l'étendue. Nous avons alors:

$$\sigma_{MR} = \frac{d_3(n_i)\overline{MR}}{d_2(n_i)} \quad (74.518)$$

Dès lors il vient (il est impossible que le client fixe une limite USL , LSL ou une cible T en ce qui concerne l'étendue d'où le fait que ces paramètres ne soient cette fois-ci pas mentionnés):

$$\begin{aligned}
 UCL_{MR} &= \overline{MR} + 3\sigma_{MR} = \overline{MR} + 3 \frac{d_3(n_i)\overline{MR}}{d_2(n_i)} = \overline{MR} \left(1 + 3 \frac{d_3(n_i)}{d_2(n_i)} \right) \\
 CL_{MR} &= \overline{MR} \\
 LCL_{MR} &= \overline{MR} - 3\sigma_{MR} = \overline{MR} - 3 \frac{d_3(n_i)\overline{MR}}{d_2(n_i)} = \overline{MR} \left(1 - 3 \frac{d_3(n_i)}{d_2(n_i)} \right)
 \end{aligned}
 \tag{74.519}$$

ce qui se note au même titre que la carte de contrôle à l'étendue:

$$\begin{aligned}
 UCL_{MR} &= \overline{MR} \cdot D_4(n_i) \\
 CL_{MR} &= \overline{MR} \\
 LCL_{MR} &= \overline{MR} \cdot D_3(n_i)
 \end{aligned}
 \tag{74.520}$$

Reprenons le tableau de données suivant:

	A	B	C	D	E	F	G
4	→						
5	↓	9.9943	10.0196	9.9732	10.0088	9.9685	9.9544
6		9.9721	9.9643	9.9426	9.9712	10.0259	10.0177
7		9.9788	10.0213	9.9407	10.0696	10.0161	10.0709
8		10.0658	10.08	9.9514	9.9958	10.0338	10.0422
9		10.0136	9.9791	9.9449	9.9403	9.9424	9.9421
10		9.9495	10.0057	9.9715	10.0388	10.0019	9.9616
11		9.9144	10.0321	10.0361	10.0253	10.0109	9.9935
12		9.9748	9.9962	10.0369	10.15	9.9996	9.9741
13		10.0063	10.0749	9.9971	9.978	9.9934	9.9555
14		10.084	9.9619	10.0063	10.0364	9.9907	9.9762
15		9.9436	10.0638	9.9617	10.0138	9.9672	9.9591
16		10.0074	9.9817	9.9753	10.0541	9.9845	9.9107
17		10.0069	10.0282	10.0385	9.9747	10.0323	9.986
18		9.9836	10.063	9.9824	9.9909	9.9684	10.0575
19		10.0317	10.0485	9.9659	10.0675	10.0047	10.0277
20		9.9756	9.9174	9.9076	9.9909	10.018	9.9979
21		9.9818	9.9962	9.9614	9.9612	9.9694	9.9349
22		10.0168	10.0043	10.0029	9.9549	10.0199	9.9691
23		9.9857	10.0503	10.0381	9.9811	9.9603	9.9123
24		9.9878	9.9841	9.9496	9.9754	9.9532	10.0576
25		10.0657	9.9838	9.9636	9.9779	10.0423	10.0547
26		9.9564	10.0184	9.989	9.9999	10.0146	10.0145
27		10.0611	9.9443	9.9899	9.9833	9.9947	10.0033
28		9.9054	9.9049	9.987	10.058	9.9156	9.9782
29		9.9768	9.9954	9.9196	9.9341	9.9162	10.0721

Figure: 74.94 - Mesures journalières sans arrêt de la machine

Puisque la carte est avec des valeurs individuelles, avec l'étendue mobile nous avons $n = 2$. Nous trouvons dans les tables:

$$D_3(n_i) = D_3(2) = 0, D_4(n_i) = D_4(2) = 3.267 \tag{74.521}$$

Ce qui nous donne:

	I	J	K	L	M	N	O
1	D3	0					
2	D4	3.267					
3							
4	N°	Mesure	Moyenne	Etendue Mob	Moyenne étendue mobil	LCL	UCL
5	1	9.9943	9.9933		0.0447	-	0.1459
6	2	10.0196	9.9933	0.0253	0.0447	-	0.1459
7	3	9.9732	9.9933	0.0464	0.0447	-	0.1459
8	4	10.0088	9.9933	0.0356	0.0447	-	0.1459
9	5	9.9685	9.9933	0.0403	0.0447	-	0.1459
10	6	9.9544	9.9933	0.0141	0.0447	-	0.1459
11	7	9.9721	9.9933	0.0177	0.0447	-	0.1459
12	8	9.9643	9.9933	0.0078	0.0447	-	0.1459
13	9	9.9426	9.9933	0.0217	0.0447	-	0.1459
14	10	9.9712	9.9933	0.0286	0.0447	-	0.1459
15	11	10.0259	9.9933	0.0547	0.0447	-	0.1459
16	12	10.0177	9.9933	0.0082	0.0447	-	0.1459
17	13	9.9788	9.9933	0.0389	0.0447	-	0.1459
18	14	10.0213	9.9933	0.0425	0.0447	-	0.1459
19	15	9.9407	9.9933	0.0806	0.0447	-	0.1459
20	16	10.0696	9.9933	0.1289	0.0447	-	0.1459
21	17	10.0161	9.9933	0.0535	0.0447	-	0.1459
22	18	10.0709	9.9933	0.0548	0.0447	-	0.1459
23	19	10.0658	9.9933	0.0051	0.0447	-	0.1459
24	20	10.0800	9.9933	0.0142	0.0447	-	0.1459
25	21	9.9514	9.9933	0.1286	0.0447	-	0.1459
26	22	9.9958	9.9933	0.0444	0.0447	-	0.1459
27	23	10.0338	9.9933	0.0380	0.0447	-	0.1459
28	24	10.0422	9.9933	0.0084	0.0447	-	0.1459
29	25	10.0136	9.9933	0.0286	0.0447	-	0.1459
30	26	9.9791	9.9933	0.0345	0.0447	-	0.1459
31	27	9.9449	9.9933	0.0342	0.0447	-	0.1459
32	28	9.9403	9.9933	0.0046	0.0447	-	0.1459
33	29	9.9424	9.9933	0.0021	0.0447	-	0.1459
34	30	9.9421	9.9933	0.0003	0.0447	-	0.1459
35	31	9.9495	9.9933	0.0074	0.0447	-	0.1459
36	32	10.0057	9.9933	0.0562	0.0447	-	0.1459
37	33	9.9715	9.9933	0.0342	0.0447	-	0.1459
38	34	10.0388	9.9933	0.0673	0.0447	-	0.1459

Figure: 74.95 - Indicateurs correspondants aux mesures obtenues

Soit:

	I	J	K	L	M	N	O
1	D3	0					
2	D4	3.267					
3							
4	N°	Mesure Moyenne	Etendue Mol	Moyenne étendue mobile (CL)	LCL	UCL	
5	1	9.9943 =MOYENNE(\$J\$5:\$J\$154)		=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M5*\$J\$1	=M5*\$J\$2	
6	2	10.0196 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J5-J6)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M6*\$J\$1	=M6*\$J\$2	
7	3	9.9732 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J6-J7)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M7*\$J\$1	=M7*\$J\$2	
8	4	10.0088 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J7-J8)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M8*\$J\$1	=M8*\$J\$2	
9	5	9.9685 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J8-J9)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M9*\$J\$1	=M9*\$J\$2	
10	6	9.9544 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J9-J10)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M10*\$J\$1	=M10*\$J\$2	
11	7	9.9721 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J10-J11)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M11*\$J\$1	=M11*\$J\$2	
12	8	9.9643 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J11-J12)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M12*\$J\$1	=M12*\$J\$2	
13	9	9.9426 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J12-J13)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M13*\$J\$1	=M13*\$J\$2	
14	10	9.9712 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J13-J14)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M14*\$J\$1	=M14*\$J\$2	
15	11	10.0259 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J14-J15)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M15*\$J\$1	=M15*\$J\$2	
16	12	10.0177 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J15-J16)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M16*\$J\$1	=M16*\$J\$2	
17	13	9.9788 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J16-J17)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M17*\$J\$1	=M17*\$J\$2	
18	14	10.0213 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J17-J18)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M18*\$J\$1	=M18*\$J\$2	
19	15	9.9407 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J18-J19)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M19*\$J\$1	=M19*\$J\$2	
20	16	10.0696 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J19-J20)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M20*\$J\$1	=M20*\$J\$2	
21	17	10.0161 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J20-J21)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M21*\$J\$1	=M21*\$J\$2	
22	18	10.0709 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J21-J22)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M22*\$J\$1	=M22*\$J\$2	
23	19	10.0658 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J22-J23)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M23*\$J\$1	=M23*\$J\$2	
24	20	10.08 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J23-J24)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M24*\$J\$1	=M24*\$J\$2	
25	21	9.9514 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J24-J25)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M25*\$J\$1	=M25*\$J\$2	
26	22	9.9958 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J25-J26)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M26*\$J\$1	=M26*\$J\$2	
27	23	10.0338 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J26-J27)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M27*\$J\$1	=M27*\$J\$2	
28	24	10.0422 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J27-J28)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M28*\$J\$1	=M28*\$J\$2	
29	25	10.0136 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J28-J29)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M29*\$J\$1	=M29*\$J\$2	
30	26	9.9791 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J29-J30)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M30*\$J\$1	=M30*\$J\$2	
31	27	9.9449 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J30-J31)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M31*\$J\$1	=M31*\$J\$2	
32	28	9.9403 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J31-J32)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M32*\$J\$1	=M32*\$J\$2	
33	29	9.9424 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J32-J33)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M33*\$J\$1	=M33*\$J\$2	
34	30	9.9421 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J33-J34)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M34*\$J\$1	=M34*\$J\$2	
35	31	9.9495 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J34-J35)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M35*\$J\$1	=M35*\$J\$2	
36	32	10.0057 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J35-J36)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M36*\$J\$1	=M36*\$J\$2	
37	33	9.9715 =MOYENNE(\$J\$5:\$J\$154)	=ABS(J36-J37)	=SOMME(\$L\$5:\$L\$154)/(NB(\$I\$5:\$I\$154)-1)	=M37*\$J\$1	=M37*\$J\$2	

Figure: 74.96 - Formules Microsoft Excel 14.0.6123 explicites du tableau précédent

Nous avons alors la carte de contrôle suivante:

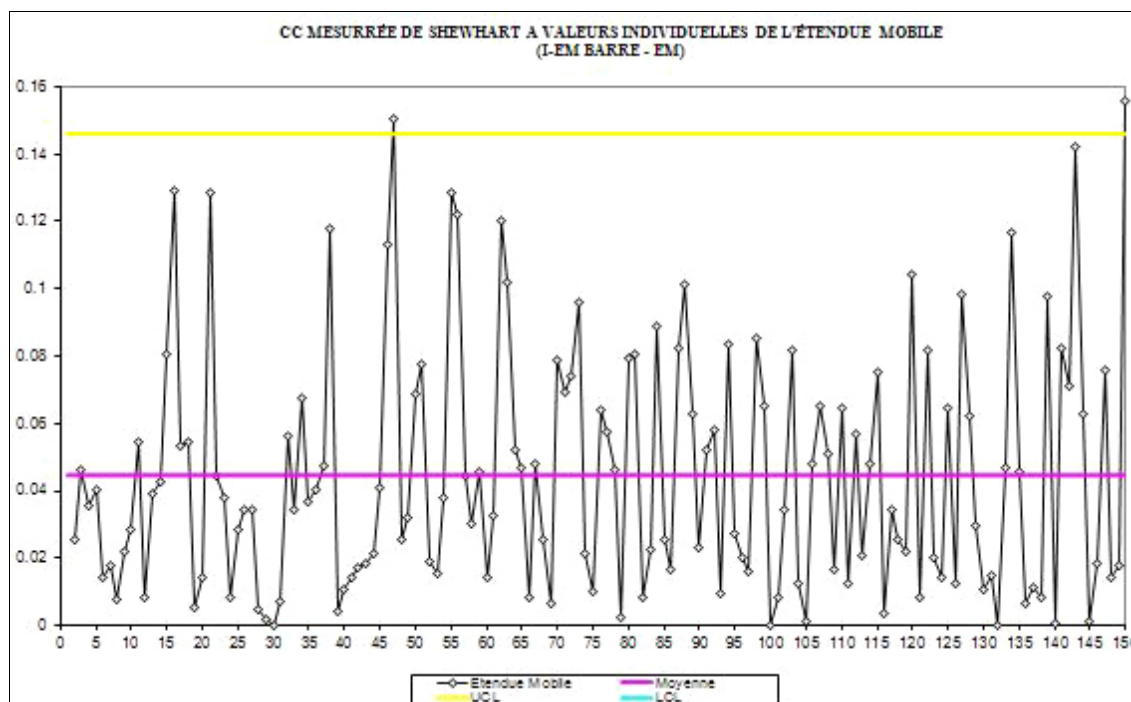


Figure: 74.97 - Carte de contrôle I-EM barre - EM

Nous voyons que la carte de contrôle avec étendue mobile semble beaucoup plus sensible que n'importe quelle carte. Mais le problème subsiste quant à savoir si l'observation que nous faisons de cette carte est correcte dans le sens où l'étendue mobile ne suit pas une loi de distribution symétrique contrairement à ce que présuppose les limites fixes de Shewhart.

Il faut aussi savoir que certaines règles de la WECO (les règles concernant les séquences) ne s'appliquent pas sur ce type de carte puisque les données y sont autocorrélées.

5.5.3. CC MOYENNES MOBILES (MA)

Nous supposons toujours que la caractéristique suit une loi Normale et que l'on prélève des échantillons de taille constante n . La moyenne mobile de durée h à l'instant t , montée M_t , est définie par la moyenne mobile (moving average) vue dans le chapitre de Statistiques:

$$M_t = \frac{\bar{X}_t + \bar{X}_{t-1} + \dots + \bar{X}_{t-h+1}}{h} \quad (74.522)$$

Le paramètre h est appelé aussi horizon. Plus grande sera la valeur de h meilleure est l'efficacité de la carte pour la détection des petits dérèglages.

Lorsque $t \geq h$, en supposant l'indépendance entre les échantillons, nous obtenons :

$$\begin{aligned} V(M_t) &= V\left(\frac{\bar{X}_t + \bar{X}_{t-1} + \dots + \bar{X}_{t-h+1}}{h}\right) = \frac{1}{h^2} V(\bar{X}_t + \bar{X}_{t-1} + \dots + \bar{X}_{t-h+1}) \\ &= \frac{1}{h^2} \sum_{i=t-h+1}^t V(\bar{X}_i) = \frac{1}{h^2} \sum_{i=t-h+1}^t \frac{\sigma^2}{n} = \frac{1}{h^2} h \frac{\sigma^2}{n} = \frac{\sigma^2}{hn} \end{aligned} \quad (74.523)$$

Et si tous les échantillons ont la même taille (si échantillons il y a...):

$$V(M_t) = \frac{1}{h^2} \sum_{i=t-h+1}^t V(\bar{X}_i) = \frac{1}{h^2} \sum_{i=t-h+1}^t \frac{\sigma^2}{n} = \frac{1}{h^2} h \frac{\sigma^2}{n} = \frac{\sigma^2}{hn} \quad (74.524)$$

Soit:

$$\sigma_{M_t} = \frac{\sigma}{\sqrt{hn}} \quad (74.525)$$

Et:

$$\begin{aligned} E(M_t) &= E\left(\frac{\bar{X}_t + \bar{X}_{t-1} + \dots + \bar{X}_{t-h+1}}{h}\right) = \frac{1}{h} E(\bar{X}_t + \bar{X}_{t-1} + \dots + \bar{X}_{t-h+1}) \\ &= \frac{1}{h} h \cdot \bar{X} = \bar{X} \end{aligned} \quad (74.526)$$

Donc:

$$\begin{aligned}
 UCL_{MA} &= M_t + 3 \cdot \frac{\sigma}{\sqrt{hn}} \\
 CL_{MA} &= \bar{\bar{X}} \\
 LCL_{MA} &= M_t - 3 \cdot \frac{\sigma}{\sqrt{hn}}
 \end{aligned}
 \quad (74.527)$$

Avantages: Le fait que l'on lisse les données avec une moyenne mobile ayant pour objectif d'adoucir les fortes variations rend cette carte très bien adaptée aussi aux petites productions raison pour laquelle on l'utilise principalement pour des mesures individuelles dont le nombre se situe au moins aux alentours de 10 (ceci dit elle est utilisable aussi pour des mesures groupées). Par ailleurs dans l'exemple qui suivra, nous nous limiterons à cas de figure particulier! Nous voyons aussi que les limites de la carte seront d'autant plus petites que h est grand. Donc la carte peut être très sensible aux petites variations.

Problèmes: L'écart-type est supposé connu ou du moins calculé suffisamment précis pour que l'estimateur converge vers la valeur théorique. Tous les échantillons doivent avoir une taille identique n si nous utilisons l'écriture développée ci-dessus (sinon les logiciels implémentent le cas où cette condition n'est pas satisfaite).

Prenons le tableau de données suivant:

	A
4	Pièces
5	9.9943
6	10.0196
7	9.9732
8	10.0088
9	9.9685
10	9.9544
11	9.9721
12	9.9643
13	9.9426
14	9.9712
15	10.0259
16	10.0177
17	9.9788
18	10.0213
19	9.9407
20	10.0696
21	10.0161
22	10.0709
23	10.0658

Figure: 74.98 - Mesures individuelles sans arrêt de la machine

Puisque la carte est avec des valeurs individuelles, nous avons $n = 1$. Et le système, en utilisant l'étendue mobile pour l'estimation de l'écart-type, se réduit donc à:

$$\begin{aligned}
 UCL_{MA} &= \bar{\bar{X}} + 3 \cdot \frac{\sigma}{\sqrt{h}} = \bar{\bar{X}} + 3 \frac{\overline{MR}}{d_2(n_i)\sqrt{h}} \\
 CL_{MA} &= \bar{\bar{X}} \\
 LCL_{MA} &= \bar{\bar{X}} - 3 \cdot \frac{\sigma}{\sqrt{h}} = \bar{\bar{X}} - 3 \frac{\overline{MR}}{d_2(n_i)\sqrt{h}}
 \end{aligned}
 \quad (74.528)$$

donc outre les dénominateurs avec la racine de h , les paramètres CL , UCL et LCL sont les mêmes que la carte de contrôle I-EM $\bar{X}$ barre.

Nous devons prendre dans le cas présent la valeur d_2 pour $n_i = 2$. Les tables nous donnent:

$$d_2(n_i) = d_2(2) = 1.128 \quad (74.529)$$

Ce qui nous donne:

	C	D	E	F	G	H	I
1	d2	1.128					
2	h	4					
3							
4	N°	Mt	CL	Etendue Mobile	Moyenne étendue mobile	LCL	UCL
5	1		9.9987		0.0392	9.9467	10.0508
6	2		9.9987	0.0253	0.0392	9.9467	10.0508
7	3		9.9987	0.0464	0.0392	9.9467	10.0508
8	4	9.998975	9.9987	0.0356	0.0392	9.9467	10.0508
9	5	9.992525	9.9987	0.0403	0.0392	9.9467	10.0508
10	6	9.976225	9.9987	0.0141	0.0392	9.9467	10.0508
11	7	9.97595	9.9987	0.0177	0.0392	9.9467	10.0508
12	8	9.964825	9.9987	0.0078	0.0392	9.9467	10.0508
13	9	9.95835	9.9987	0.0217	0.0392	9.9467	10.0508
14	10	9.96255	9.9987	0.0286	0.0392	9.9467	10.0508
15	11	9.976	9.9987	0.0547	0.0392	9.9467	10.0508
16	12	9.98935	9.9987	0.0082	0.0392	9.9467	10.0508
17	13	9.9984	9.9987	0.0389	0.0392	9.9467	10.0508
18	14	10.010925	9.9987	0.0425	0.0392	9.9467	10.0508
19	15	9.989625	9.9987	0.0806	0.0392	9.9467	10.0508
20	16	10.0026	9.9987	0.1289	0.0392	9.9467	10.0508
21	17	10.011925	9.9987	0.0535	0.0392	9.9467	10.0508
22	18	10.024325	9.9987	0.0548	0.0392	9.9467	10.0508
23	19	10.0556	9.9987	0.0051	0.0392	9.9467	10.0508

Figure: 74.99 - Indicateurs correspondants aux mesures obtenues

Soit (nous avons coupé la capture d'écran en deux parties faute de place):

	C	D	E	F	G
1	d2	1.128			
2	h	4			
3					
4	N°	Mt	CL	Etendue Mob	Moyenne étendue mobile
5	1		=MOYENNE(\$A\$5:\$A\$23)		=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
6	2		=MOYENNE(\$A\$5:\$A\$23)	=ABS(A5-A6)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
7	3		=MOYENNE(\$A\$5:\$A\$23)	=ABS(A6-A7)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
8	4	=SOMME(A5:A8)/4	=MOYENNE(\$A\$5:\$A\$23)	=ABS(A7-A8)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
9	5	=SOMME(A6:A9)/4	=MOYENNE(\$A\$5:\$A\$23)	=ABS(A8-A9)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
10	6	=SOMME(A7:A10)/4	=MOYENNE(\$A\$5:\$A\$23)	=ABS(A9-A10)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
11	7	=SOMME(A8:A11)/4	=MOYENNE(\$A\$5:\$A\$23)	=ABS(A10-A11)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
12	8	=SOMME(A9:A12)/4	=MOYENNE(\$A\$5:\$A\$23)	=ABS(A11-A12)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
13	9	=SOMME(A10:A13)/4	=MOYENNE(\$A\$5:\$A\$23)	=ABS(A12-A13)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
14	10	=SOMME(A11:A14)/4	=MOYENNE(\$A\$5:\$A\$23)	=ABS(A13-A14)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
15	11	=SOMME(A12:A15)/4	=MOYENNE(\$A\$5:\$A\$23)	=ABS(A14-A15)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
16	12	=SOMME(A13:A16)/4	=MOYENNE(\$A\$5:\$A\$23)	=ABS(A15-A16)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
17	13	=SOMME(A14:A17)/4	=MOYENNE(\$A\$5:\$A\$23)	=ABS(A16-A17)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
18	14	=SOMME(A15:A18)/4	=MOYENNE(\$A\$5:\$A\$23)	=ABS(A17-A18)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
19	15	=SOMME(A16:A19)/4	=MOYENNE(\$A\$5:\$A\$23)	=ABS(A18-A19)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
20	16	=SOMME(A17:A20)/4	=MOYENNE(\$A\$5:\$A\$23)	=ABS(A19-A20)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
21	17	=SOMME(A18:A21)/4	=MOYENNE(\$A\$5:\$A\$23)	=ABS(A20-A21)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
22	18	=SOMME(A19:A22)/4	=MOYENNE(\$A\$5:\$A\$23)	=ABS(A21-A22)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)
23	19	=SOMME(A20:A23)/4	=MOYENNE(\$A\$5:\$A\$23)	=ABS(A22-A23)	=SOMME(\$F\$5:\$F\$23)/(NB(\$C\$5:\$C\$23)-1)

Figure: 74.100 - Formules Microsoft Excel 14.0.6123 explicites du tableau précédent (première partie)

et:

	H	I
1		
2		
3		
4	LCL	UCL
5	=E5-3*G5/(\$D\$1*RACINE(\$D\$2))	=E5+3*G5/(\$D\$1*RACINE(\$D\$2))
6	=E6-3*G6/(\$D\$1*RACINE(\$D\$2))	=E6+3*G6/(\$D\$1*RACINE(\$D\$2))
7	=E7-3*G7/(\$D\$1*RACINE(\$D\$2))	=E7+3*G7/(\$D\$1*RACINE(\$D\$2))
8	=E8-3*G8/(\$D\$1*RACINE(\$D\$2))	=E8+3*G8/(\$D\$1*RACINE(\$D\$2))
9	=E9-3*G9/(\$D\$1*RACINE(\$D\$2))	=E9+3*G9/(\$D\$1*RACINE(\$D\$2))
10	=E10-3*G10/(\$D\$1*RACINE(\$D\$2))	=E10+3*G10/(\$D\$1*RACINE(\$D\$2))
11	=E11-3*G11/(\$D\$1*RACINE(\$D\$2))	=E11+3*G11/(\$D\$1*RACINE(\$D\$2))
12	=E12-3*G12/(\$D\$1*RACINE(\$D\$2))	=E12+3*G12/(\$D\$1*RACINE(\$D\$2))
13	=E13-3*G13/(\$D\$1*RACINE(\$D\$2))	=E13+3*G13/(\$D\$1*RACINE(\$D\$2))
14	=E14-3*G14/(\$D\$1*RACINE(\$D\$2))	=E14+3*G14/(\$D\$1*RACINE(\$D\$2))
15	=E15-3*G15/(\$D\$1*RACINE(\$D\$2))	=E15+3*G15/(\$D\$1*RACINE(\$D\$2))
16	=E16-3*G16/(\$D\$1*RACINE(\$D\$2))	=E16+3*G16/(\$D\$1*RACINE(\$D\$2))
17	=E17-3*G17/(\$D\$1*RACINE(\$D\$2))	=E17+3*G17/(\$D\$1*RACINE(\$D\$2))
18	=E18-3*G18/(\$D\$1*RACINE(\$D\$2))	=E18+3*G18/(\$D\$1*RACINE(\$D\$2))
19	=E19-3*G19/(\$D\$1*RACINE(\$D\$2))	=E19+3*G19/(\$D\$1*RACINE(\$D\$2))
20	=E20-3*G20/(\$D\$1*RACINE(\$D\$2))	=E20+3*G20/(\$D\$1*RACINE(\$D\$2))
21	=E21-3*G21/(\$D\$1*RACINE(\$D\$2))	=E21+3*G21/(\$D\$1*RACINE(\$D\$2))
22	=E22-3*G22/(\$D\$1*RACINE(\$D\$2))	=E22+3*G22/(\$D\$1*RACINE(\$D\$2))
23	=E23-3*G23/(\$D\$1*RACINE(\$D\$2))	=E23+3*G23/(\$D\$1*RACINE(\$D\$2))

Figure: 74.101 - Formules Microsoft Excel 14.0.6123 explicites du tableau précédent (deuxième partie)

Nous avons alors la carte de contrôle suivante:

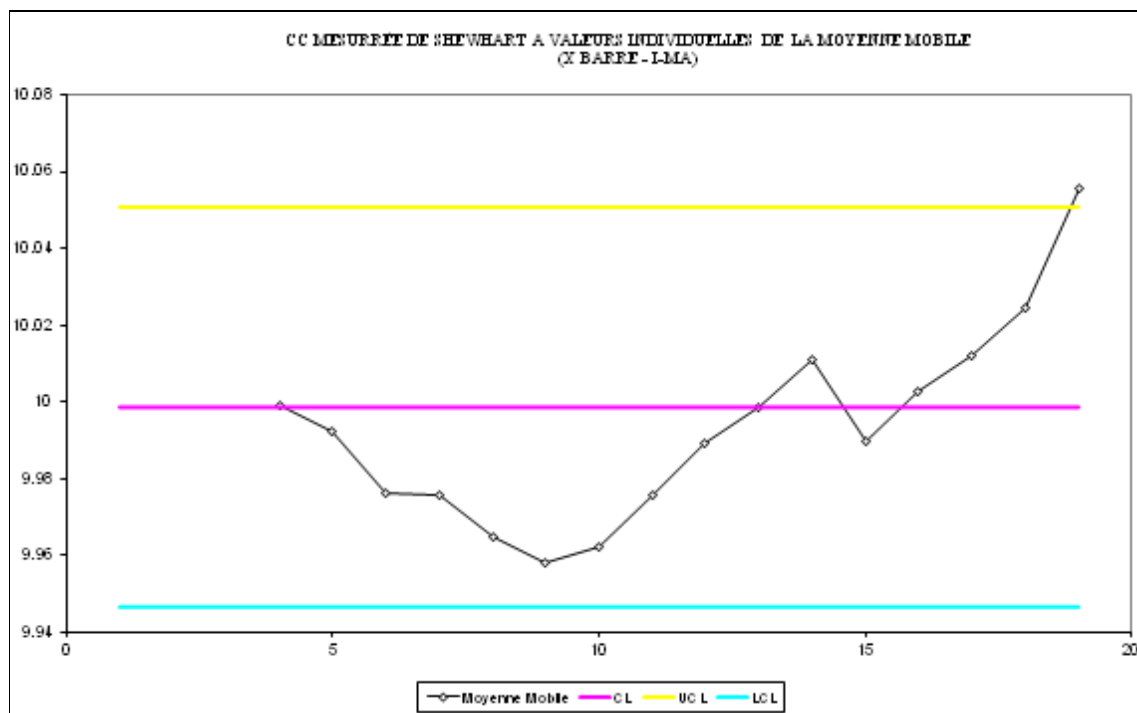


Figure: 74.102 - Carte de contrôle à valeurs individuelles de la moyenne mobile (X barre - I-MA)

5.5.4. CC CUSUM AVEC V-MASQUE EMPIRIQUE

Un des principaux inconvénients des cartes de type Shewhart est de ne baser l'analyse que sur les dernières informations recueillies (hypothèse d'indépendance des mesures). Elles ignorent les informations relatives aux tendances du procédé contenues dans les dernières estimations. Il faut savoir que de nombreux instituts de normalisation qualité dans le monde (AFNOR ou ASQ) considèrent les cartes de Shewhart comme complètement obsolètes (à tort ou à raison?) et préconisent les cartes de type CUSUM car utilisent l'autocorrélation et l'amplification des variations.

Dans les années 1950 des cartes de contrôle appelées donc CUSUM (CUMulated Sum: somme cumulée) ou "cartes de Page-Hinkley" (en honneur à leurs inventeurs) ont été introduites par des statisticiens pour détecter de petites variations. Même si elles sont mathématiquement plus optimales, elles sont difficiles à configurer correctement et très sensibles aux paramètres initiaux.

Le principe des cartes CUSUM (car il en existe plusieurs pour une loi de distribution donnée) est de sommer les écarts entre les estimations de la position du procédé/processus et la cible. Lorsque le cumul des écarts positifs ou le cumul des écarts négatifs dépasse une certaine valeur, nous concluons à un décentrage du procédé.

Nous posons:

$$C_n = \sum_{i=1}^n (X_i - \mu) \text{ et } S_n = \frac{C_n}{\sigma} \quad (74.530)$$

où souvent la première égalité est aussi notée:

$$C_n = \sum_{i=1}^n (X_i - \mu) = (X_n - \mu) + \sum_{i=1}^{n-1} (X_i - \mu) = (X_n - \mu) + C_{n-1} \quad (74.531)$$

Dès lors il vient une écriture très connue dans la littérature:

$$S_n = \frac{C_n}{\sigma} = \frac{(X_n - \mu) + C_{n-1}}{\sigma} = \frac{(X_n - \mu)}{\sigma} + S_{n-1} = Z_n + S_{n-1} \quad (74.532)$$

En général, la carte de CUSUM se comporte comme suit:

- Si le procédé est sous-contrôle, la somme cumulée fluctue aléatoirement autour de zéro.
- Si la moyenne subit une dérive positive de la moyenne, la statistique C_n ou S_n croît rapidement par accumulation d'un biais systématique.

Remarques:

R1. Dans de nombreux ouvrages et de très nombreux cas applicatifs, nous n'utilisons pas S_n (variable centrée et réduite) car la variance est inconnue et nous nous restreignons qu'à l'usage de C_n . Les développements mathématiques qui suivent alors s'adaptent très facilement.

R2. La même approche peut être faite avec des moyennes de prélèvements selon:

$$C_n = \sum_{i=1}^n (\bar{X}_i - \mu_{\bar{X}}) \text{ et } S_n = \frac{C_n}{\sigma_{\bar{X}}}$$

R3. Il existe de nombreuses versions différentes de carte CUSUM et même pour la carte CUSUM avec V-Masque il en existe de multiples sous-familles. Pour être honnête c'est selon mon opinion un sacré foutoir...

Pour l'exemple, reprenons le tableau de données suivant:

	A	B	C	D	E	F	G
1		Ech. 1	Ech. 2	Ech. 3	Ech. 4	Ech. 5	Ech. 6
2	Jour 1	9.9943	10.0196	9.9732	10.0088	9.9685	9.9544
3	Jour 2	9.9721	9.9643	9.9426	9.9712	10.0259	10.0177
4	Jour 3	9.9788	10.0213	9.9407	10.0696	10.0161	10.0709
5	Jour 4	10.0658	10.08	9.9514	9.9958	10.0338	10.0422
6	Jour 5	10.0136	9.9791	9.9449	9.9403	9.9424	9.9421
7	Jour 6	9.9495	10.0057	9.9715	10.0388	10.0019	9.9616
8	Jour 7	9.9144	10.0321	10.0361	10.0253	10.0109	9.9935
9	Jour 8	9.9748	9.9962	10.0369	10.15	9.9996	9.9741
10	Jour 9	10.0063	10.0749	9.9971	9.978	9.9934	9.9555
11	Jour 10	10.084	9.9619	10.0063	10.0364	9.9907	9.9762
12	Jour 11	9.9436	10.0638	9.9617	10.0138	9.9672	9.9591
13	Jour 12	10.0074	9.9817	9.9753	10.0541	9.9845	9.9107
14	Jour 13	10.0069	10.0282	10.0385	9.9747	10.0323	9.986
15	Jour 14	9.9836	10.063	9.9824	9.9909	9.9684	10.0575
16	Jour 15	10.0317	10.0485	9.9659	10.0675	10.0047	10.0277
17	Jour 16	9.9756	9.9174	9.9076	9.9909	10.018	9.9979
18	Jour 17	9.9818	9.8962	9.9614	9.9612	9.9694	9.9349
19	Jour 18	10.0168	10.0043	10.0029	9.9549	10.0199	9.9691
20	Jour 19	9.9857	10.0503	10.0381	9.9811	9.9603	9.9123
21	Jour 20	9.9878	9.9841	9.9496	9.9754	9.9532	10.0576
22	Jour 21	10.0657	9.9838	9.9636	9.9779	10.0423	10.0547
23	Jour 22	9.9564	10.0184	9.989	9.9999	10.0146	10.0145
24	Jour 23	10.0611	9.9443	9.9899	9.9833	9.9947	10.0033
25	Jour 24	9.9054	9.9049	9.987	10.058	9.9156	9.9782
26	Jour 25	9.9768	9.9954	9.9196	9.9341	9.9162	10.0721

Figure: 74.103 - Mesures journalières avec 6 individus sans arrêt de la machine

et nous calculons la somme cumulée en considérant toujours une cible de $\mu = T = 10$:

	I	J	K
1	Points	Moyenne-Cible	Ci
2	1	-0.0135	-0.0135
3	2	-0.0177	-0.0312
4	3	0.0162	-0.0150
5	4	0.0282	0.0132
6	5	-0.0396	-0.0264
7	6	-0.0118	-0.0383
8	7	0.0020	-0.0362
9	8	0.0219	-0.0143
10	9	0.0009	-0.0134
11	10	0.0092	-0.0042
12	11	-0.0151	-0.0193
13	12	-0.0144	-0.0337
14	13	0.0111	-0.0226
15	14	0.0076	-0.0150
16	15	0.0243	0.0094
17	16	-0.0321	-0.0227
18	17	-0.0432	-0.0719
19	18	-0.0053	-0.0773
20	19	-0.0120	-0.0893
21	20	-0.0154	-0.1047
22	21	0.0147	-0.0900
23	22	-0.0012	-0.0912
24	23	-0.0039	-0.0951
25	24	-0.0418	-0.1369
26	25	-0.0310	-0.1679

Figure: 74.104 - Indicateurs correspondants aux mesures obtenues

Soit:

	I	J	K	L	M
1	Points	Moyenne-Cible	C _i		Objectif
2	1	=MOYENNE(B2:G2)-\$M\$2	=J2		10
3	2	=MOYENNE(B3:G3)-\$M\$2	=K2+J3		h
4	3	=MOYENNE(B4:G4)-\$M\$2	=K3+J4		1
5	4	=MOYENNE(B5:G5)-\$M\$2	=K4+J5		d
6	5	=MOYENNE(B6:G6)-\$M\$2	=K5+J6		0.05
7	6	=MOYENNE(B7:G7)-\$M\$2	=K6+J7		K
8	7	=MOYENNE(B8:G8)-\$M\$2	=K7+J8		=M6/2
9	8	=MOYENNE(B9:G9)-\$M\$2	=K8+J9		
10	9	=MOYENNE(B10:G10)-\$M\$2	=K9+J10		
11	10	=MOYENNE(B11:G11)-\$M\$2	=K10+J11		
12	11	=MOYENNE(B12:G12)-\$M\$2	=K11+J12		
13	12	=MOYENNE(B13:G13)-\$M\$2	=K12+J13		
14	13	=MOYENNE(B14:G14)-\$M\$2	=K13+J14		
15	14	=MOYENNE(B15:G15)-\$M\$2	=K14+J15		
16	15	=MOYENNE(B16:G16)-\$M\$2	=K15+J16		
17	16	=MOYENNE(B17:G17)-\$M\$2	=K16+J17		
18	17	=MOYENNE(B18:G18)-\$M\$2	=K17+J18		
19	18	=MOYENNE(B19:G19)-\$M\$2	=K18+J19		
20	19	=MOYENNE(B20:G20)-\$M\$2	=K19+J20		
21	20	=MOYENNE(B21:G21)-\$M\$2	=K20+J21		
22	21	=MOYENNE(B22:G22)-\$M\$2	=K21+J22		
23	22	=MOYENNE(B23:G23)-\$M\$2	=K22+J23		
24	23	=MOYENNE(B24:G24)-\$M\$2	=K23+J24		
25	24	=MOYENNE(B25:G25)-\$M\$2	=K24+J25		
26	25	=MOYENNE(B26:G26)-\$M\$2	=K25+J26		

Figure: 74.105 - Formules Microsoft Excel 14.0.6123 explicites du tableau précédent

Nous obtenons alors:

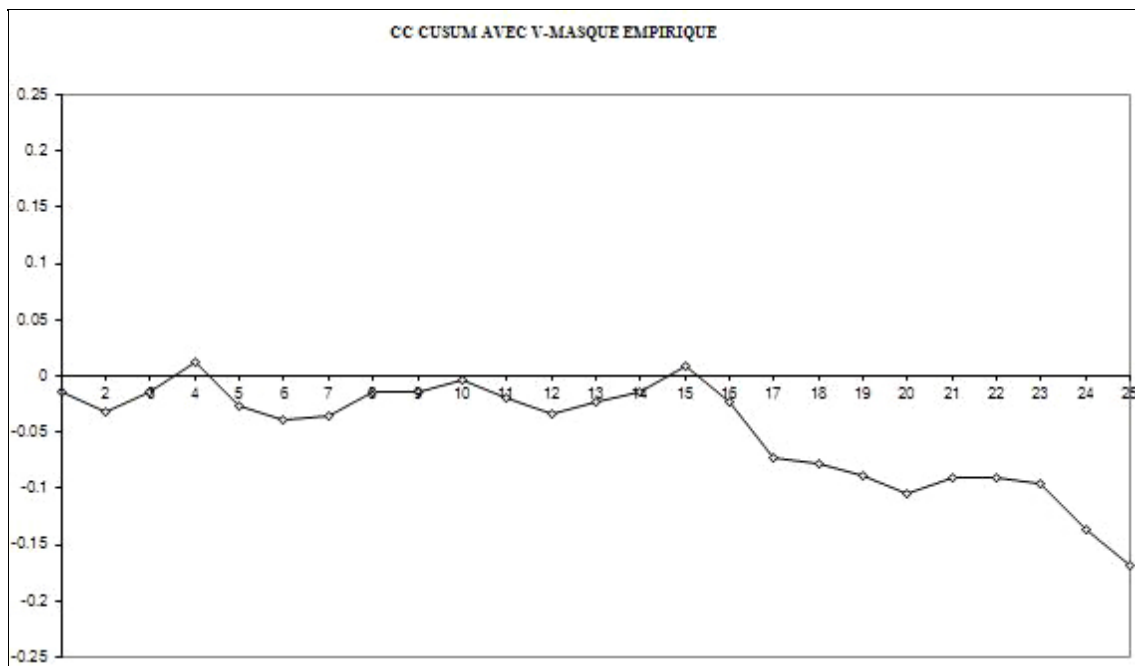


Figure: 74.106 - Carte de contrôle CUSUM pour l'instant sans le V-Masque empirique

Nous voyons alors très vite avec cette carte que la somme cumulée semble décroître. Ceci indique que la vraie qualité moyenne est inférieure à la cible (vous pouvez essayer avec 9.994 comme cible afin de voir la sensibilité).

Il faut maintenant considérer deux cas:

H1. Il vient sous l'hypothèse de normalité (processus sous contrôle) et par la propriété de stabilité de la loi Normale que pour une variable aléatoire centrée réduite:

$$S_i = N(\mu, \sigma) = N(0, \sqrt{i}) \quad (74.533)$$

ou sinon:

$$C_n = N(\mu, \sigma) = N(0, \sigma\sqrt{i}) \quad (74.534)$$

Dès lors, sous cette hypothèse, nous avons les sommes cumulées $(S_i)_{i=1\dots n}$ qui sont réparties autour d'une droite de régression de pente 0 (puisque la moyenne des sommes de variables aléatoires dont la moyenne est nulle est... nulle) et les distributions ont leur écart-type qui est proportionnel à la racine de i (respectivement la variance qui est proportionnelle à i)

H2. Si le processus est hors contrôle (dispersion globale), il existe un instant $t < n$ tel que $X_i = N(\mu, \sigma)$ pour $i \leq r-1$ et $X_i = N(\mu + \delta, \sigma)$ pour $i \geq r$ avec $\delta > 0$ (cette dernière valeur étant souvent imposée par les contraintes de qualité et nommée "**translation moyenne importante**"). Dès lors, nous avons:

$$S_i = N(0, \sqrt{i}) \quad (74.535)$$

ou:

$$C_n = N(0, \sigma\sqrt{i}) \quad (74.536)$$

pour $i = 1, \dots, r-1$ et:

$$\begin{aligned} S_k - S_{r-1} &= N\left(\frac{\delta}{\sigma}, \sqrt{k}\right) - N(0, \sqrt{r-1}) = N\left(\frac{\delta}{\sigma} + 0, \sqrt{(\sqrt{k})^2 + (\sqrt{r-1})^2}\right) \\ &= N\left(\frac{\delta}{\sigma} + 0, \sqrt{k + (r-1)}\right) \end{aligned} \quad (74.537)$$

pour $k \geq r$. Ceci étant précisé, il va de soi que les sommes cumulées:

$$S_k = N\left(\frac{\delta}{\sigma}, \sqrt{k}\right) \quad (74.538)$$

sont réparties autour d'une droite de régression de pente δ/σ à partir de l'instant r . Respectivement, nous avons:

$$C_k = N(\delta, \sigma\sqrt{k}) \quad (74.539)$$

Pour continuer, l'idée est de discriminer l'hypothèse H1 d'une droite de régression de pente nulle et l'hypothèse H2 d'une droite de régression de pente δ/σ pour S_k en fixant pour frontière empirique la pente moitié (le juste milieu quoi...):

$$K_s = \frac{\delta}{2\sigma} \quad (74.540)$$

respectivement:

$$K_C = \frac{\delta}{2} \quad (74.541)$$

et un décalage lui aussi empirique $h \geq 0$ avec $h \in \mathbb{N}$. À l'étape n nous rejetons H_1 , si pour un entier $m \leq n$ nous avons:

$$S_m \geq (n-m)K_S + S_1 + hK_S \quad (74.542)$$

Il bien évidemment aussi possible de faire le rejet avec une représentation géométrique (cela a l'avantage d'aller beaucoup plus vite et de faire le contrôle d'un seul coup pour n'importe quel m plus petit ou égal à n). L'idée est alors de tracer une droite de pente K partant du point d'abscisse $n+h$. S'il existe un point (m, S_m) avec $m < n$ en dessous de la droite, nous tirons alors l'alarme.

Le lecteur aura remarqué que dans le cas présent, nous considérons δ strictement positif et donc in extenso la pente K aussi. Évidemment ce n'est pas réaliste et dès lors il faut utiliser une droite de pente K positive et une autre droite de pente $-K$ négative. La zone comprise entre les deux droites avec une abscisse $m \leq n$ est appelée un "**V-masque**" et le contrôle est assimilé logiquement alors à un test bilatéral. Ainsi, s'il existe un point (m, S_m) avec $m < n$ en dehors du V-masque nous tirons l'alarme.

Faisons un exemple avec le graphique précédent. Considérons les valeurs empiriques suivantes des paramètres:

$$\delta = 0.05 \text{ avec } h = 1 \quad (74.543)$$

Soit:

$$K_C = 0.025 \quad (74.544)$$

Il vient alors:

	Q	P	Q
1	N	Masque Y inf 25	Masque Y sup 25
2	1	-0.793	0.457
3	2	-0.768	0.432
4	3	-0.743	0.407
5	4	-0.718	0.382
6	5	-0.693	0.357
7	6	-0.668	0.332
8	7	-0.643	0.307
9	8	-0.618	0.282
10	9	-0.593	0.257
11	10	-0.568	0.232
12	11	-0.543	0.207
13	12	-0.518	0.182
14	13	-0.493	0.157
15	14	-0.468	0.132
16	15	-0.443	0.107
17	16	-0.418	0.082
18	17	-0.393	0.057
19	18	-0.368	0.032
20	19	-0.343	0.007
21	20	-0.318	-0.018
22	21	-0.293	-0.043
23	22	-0.268	-0.068
24	23	-0.243	-0.093
25	24	-0.218	-0.118
26	25	-0.193	-0.143

Figure: 74.107 - Indicateurs correspondants aux mesures obtenues

Soit:

	Q	P	Q
1	N	Masque V inf 25	Masque V sup 25
2	1	=(\$O\$26-Q2)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q2)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
3	2	=(\$O\$26-Q3)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q3)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
4	3	=(\$O\$26-Q4)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q4)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
5	4	=(\$O\$26-Q5)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q5)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
6	5	=(\$O\$26-Q6)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q6)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
7	6	=(\$O\$26-Q7)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q7)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
8	7	=(\$O\$26-Q8)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q8)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
9	8	=(\$O\$26-Q9)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q9)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
10	9	=(\$O\$26-Q10)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q10)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
11	10	=(\$O\$26-Q11)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q11)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
12	11	=(\$O\$26-Q12)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q12)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
13	12	=(\$O\$26-Q13)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q13)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
14	13	=(\$O\$26-Q14)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q14)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
15	14	=(\$O\$26-Q15)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q15)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
16	15	=(\$O\$26-Q16)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q16)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
17	16	=(\$O\$26-Q17)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q17)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
18	17	=(\$O\$26-Q18)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q18)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
19	18	=(\$O\$26-Q19)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q19)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
20	19	=(\$O\$26-Q20)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q20)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
21	20	=(\$O\$26-Q21)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q21)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
22	21	=(\$O\$26-Q22)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q22)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
23	22	=(\$O\$26-Q23)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q23)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
24	23	=(\$O\$26-Q24)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q24)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
25	24	=(\$O\$26-Q25)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q25)*\$M\$8+\$K\$26+\$M\$4*\$M\$8
26	25	=(\$O\$26-Q26)*\$M\$8+\$K\$26-\$M\$4*\$M\$8	=(\$O\$26-Q26)*\$M\$8+\$K\$26+\$M\$4*\$M\$8

Figure: 74.108 - Formules Microsoft Excel 14.0.6123 explicites du tableau précédent

Nous avons alors par exemple au point 26 correspondant à 25+h:

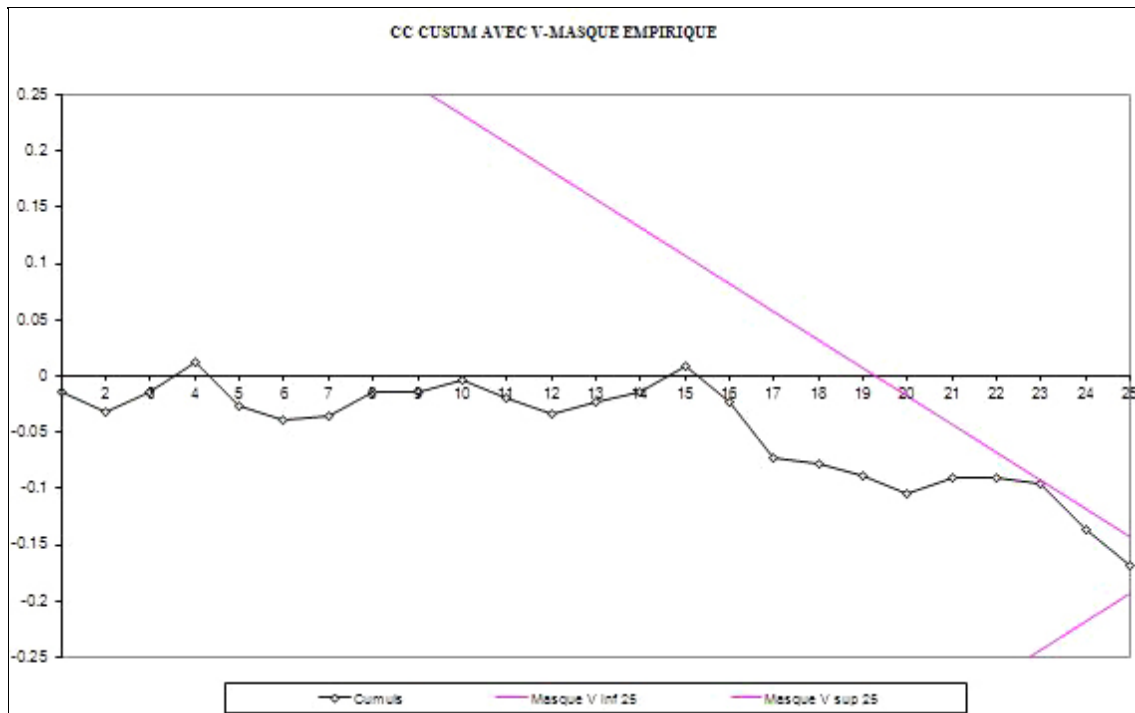


Figure: 74.109 - Carte de contrôle CUSUM avec le V-Masque empirique

Le lecteur observera facilement qu'avec le choix effectué, nous n'aurions aucune alarme avec les paramètres empiriques choisis. Il conviendrait alors dans la pratique de diminuer la valeur de δ .

Pour résumer, les avantages de la carte de CUSUM par rapport aux cartes de Shewhart sont les suivantes:

- Elles sont plus efficaces pour détecter de petites dérives

- La dérive du procédé apparaît visuellement sur ces cartes
- Il est en général facile de détecter en quel point le procédé a commencé à dériver

Les désavantages de ces cartes sont les suivantes:

- Elles peuvent être lentes pour détecter de grandes dérives
- Elles ne sont pas très intéressantes pour analyser des données passées et rechercher des cycles ou allures caractéristiques dans la distribution étudiée.
- Elles sont moins facilement acceptées par des opérateurs car moins intuitives et ne représentent pas directement la caractéristique étudiée.

Pour les raisons discutées ci-dessus, il est conseillé de représenter en parallèle la carte de Shewhart et la carte CUSUM.

Remarque: Il existe encore d'autres cartes de CUSUM de type V-masque. Donc attention avec les logiciels à savoir ce que l'on fait!!!

5.5.5. CC EWMA (PONDÉRATION EXPONENTIELLE AVEC MOYENNE MOBILE) AVEC LIMITES FIXES

Une alternative intéressante aux cartes CUSUM dans le cadre de l'autocorrélation et la détection de faibles dérèglages sont les cartes de contrôle EWMA (Exponentially Weighed Moving Average) qui ont leur origine dans l'analyse des séries temporelles (cf. chapitre d'Économie).

Remarque: Certains utilisent aussi des cartes de contrôle avec les techniques empiriques de la moyenne mobile simple ou double comme étudiées dans le chapitre d'Économie lorsque les données sont autocorrélées.

Une carte EWMA est facile à mettre en place et n'est pas trop sensible aux changements de paramètres et la non-normalité des données, elle reste toutefois tout aussi performante que les cartes CUSUM pour détecter les petites variations.

Par définition, la statistique reportée sur la carte de contrôle EWMA se calcule par la relation suivante (remarquez que les indices des différents termes ne sont pas les mêmes que le lissage exponentiel d'une série temporelle comme vu dans le chapitre d'Économie puisque l'objectif n'est pas de faire de la prévision!!):

$$F_t = \lambda \bar{X}_t + (1 - \lambda) F_{t-1} \quad \forall \lambda \in]0.6; 0.8[\quad (74.545)$$

avec la particularité que pour cette carte de contrôle certains logiciels (comme Minitab) prennent:

$$F_0 = \bar{X}_1 \text{ ou } F_0 = \bar{X} \quad (74.546)$$

La constante λ (constante de lissage) détermine le poids que nous souhaitons affecter aux dernières mesures. Plus cette valeur est petite, plus la carte est sensible aux dérives subites.

Démontrons que cette dernière relation peut s'écrire sous la forme:

$$F_t = \left(\sum_{i=0}^{t-1} \lambda(1-\lambda)^i \bar{X}_{t-i} \right) + (1-\lambda)^t F_0 \quad (74.547)$$

Effectivement:

$$\begin{aligned} F_t &= \lambda \bar{X}_t + (1-\lambda) F_{t-1} = \lambda \bar{X}_t + (1-\lambda) (\lambda \bar{X}_{t-1} + (1-\lambda) F_{t-2}) \\ &= \lambda \bar{X}_t + (1-\lambda) (\lambda \bar{X}_{t-1} + (1-\lambda) (\lambda \bar{X}_{t-2} + (1-\lambda) F_{t-3})) \\ &= \left(\sum_{i=0}^{t-1} \lambda(1-\lambda)^i \bar{X}_{t-i} \right) + (1-\lambda)^t F_0 \end{aligned} \quad (74.548)$$

Supposons pour la suite que:

$$X_i = N(\mu, \sigma^2) \quad (74.549)$$

Nous avons alors:

$$V(F_t) = V \left(\left(\sum_{i=0}^{t-1} \lambda(1-\lambda)^i \bar{X}_{t-i} \right) + (1-\lambda)^t F_0 \right) \quad (74.550)$$

et en se rappelant d'une des propriétés de la variance:

$$V(a \cdot \bar{X}) = a^2 V(\bar{X}) \quad (74.551)$$

Il vient:

$$\begin{aligned} V(F_t) &= \left(\sum_{i=0}^{t-1} \lambda^2 (1-\lambda)^{2i} V(\bar{X}_{t-i}) \right) + (1-\lambda)^{2t} V(F_0) \\ &= \left(\sum_{i=0}^{t-1} \lambda^2 (1-\lambda)^{2i} V(\bar{X}_{t-i}) \right) + (1-\lambda)^{2t} \cdot 0 \\ &= \sum_{i=0}^{t-1} \lambda^2 (1-\lambda)^{2i} V(\bar{X}_{t-i}) = \lambda^2 \sum_{i=0}^{t-1} (1-\lambda)^{2i} \sigma^2 \\ &= \lambda^2 \sigma^2 \sum_{i=0}^{t-1} (1-\lambda)^{2i} \end{aligned} \quad (74.552)$$

Faisons un changement de variable:

$$x = (1-\lambda)^2 \quad (74.553)$$

Il vient alors en utilisant le résultat démontré dans le cadre de l'étude des suites arithmétiques du chapitre sur les Suites et Séries:

$$\begin{aligned}
V(F_t) &= \lambda^2 \sigma^2 \sum_{i=0}^{t-1} x^i = \lambda^2 \sigma^2 \frac{1 - x^{(t-1)+1}}{1 - x} = \lambda^2 \sigma^2 \frac{1 - (1-\lambda)^{2t}}{1 - (1-\lambda)^2} \\
&= \lambda^2 \sigma^2 \frac{1 - (1-\lambda)^{2t}}{1 - 1 + 2\lambda + \lambda^2} = \sigma^2 \frac{\lambda^2}{2\lambda + \lambda^2} (1 - (1-\lambda)^{2t}) \\
&= \sigma^2 \frac{\lambda}{2 + \lambda} (1 - (1-\lambda)^{2t})
\end{aligned} \quad (74.554)$$

quand t est suffisamment grand, la variance se réduit alors à (approximation que la majorité des logiciels - comme Minitab - ne font pas et donc les limites ne sont pas constantes en fonction de t):

$$V(F_t) = \sigma_{F_t}^2 \cong \sigma^2 \frac{\lambda}{2 + \lambda} \quad (74.555)$$

Les limites de contrôle de la carte EWMA se construisent alors aussi à $\pm 3\sigma$ avec:

$$\sigma_{EWMA} = \sigma \sqrt{\frac{\lambda}{2 + \lambda}} \quad (74.556)$$

Nous avons donc lorsqu'un objectif est fixé ainsi qu'un écart-type (il est impossible que le client fixe une limite USL , LSL ou une cible T en ce qui concerne la pondération exponentielle d'où le fait que ces paramètres ne soient cette fois-ci pas mentionnés):

$$\begin{aligned}
UCL_{EWMA} &= T + 3\sigma \sqrt{\frac{\lambda}{2 + \lambda}} \\
CL_{EWMA} &= T \\
LCL_{EWMA} &= T - 3\sigma \sqrt{\frac{\lambda}{2 + \lambda}}
\end{aligned} \quad (74.557)$$

que nous retrouvons sous la forme suivante lorsqu'il s'agit de travailler avec des prélèvements d'échantillons:

$$\begin{aligned}
UCL_{EWMA} &= \bar{\bar{X}} + 3 \frac{\sigma}{\sqrt{n_i}} \sqrt{\frac{\lambda}{2 + \lambda}} = \bar{\bar{X}} + 3\sigma \sqrt{\frac{\lambda}{n_i (2 + \lambda)}} \\
CL_{EWMA} &= \bar{\bar{X}} \\
LCL_{EWMA} &= \bar{\bar{X}} - 3 \frac{\sigma}{\sqrt{n_i}} \sqrt{\frac{\lambda}{2 + \lambda}} = \bar{\bar{X}} - 3\sigma \sqrt{\frac{\lambda}{n_i (2 + \lambda)}}
\end{aligned} \quad (74.558)$$

Certains logiciels utilisent pour le calcul de l'écart-type la relation suivante que nous avons démontrée plus haut:

$$\sigma = \frac{\bar{R}}{d_2(n_i)} \quad (74.559)$$

et certains autres (comme Minitab):

$$\sigma = \frac{\overline{MR}}{d_2(2)} \quad (74.560)$$

cette dernière option ayant l'avantage de fonctionner si la taille des échantillons est unitaire (nous parlons alors bien évidemment de "carte EWMA aux valeurs individuelles"). Rappelons effectivement que nous avons déjà mentionné que la carte EWMA est mal adaptée aux grandes séries.

Il convient de remarquer que quelle que soit l'option choisie, comme nous avons:

$$\sqrt{\frac{\lambda}{2+\lambda}} < 1 \quad (74.561)$$

les limites de contrôles seront donc toujours nettement inférieures à ceux d'une carte à l'étendue ou à étendue mobile. C'est aussi une raison pour laquelle nous lisons les données!

Faisons un exemple pour chaque variante de calcul de l'écart-type puisque chacune est d'égale importance dans la pratique. Pour cela, commençons par la carte EWMA avec:

$$\begin{aligned} UCL_{EWMA} &= \bar{\bar{X}} + 3 \frac{\bar{R}}{d_2(n_i)} \sqrt{\frac{\lambda}{n_i(2+\lambda)}} \\ CL_{EWMA} &= \bar{\bar{X}} \\ LCL_{EWMA} &= \bar{\bar{X}} - 3 \frac{\bar{R}}{d_2(n_i)} \sqrt{\frac{\lambda}{n_i(2+\lambda)}} \end{aligned} \quad (74.562)$$

et en prenant les données du tableau suivant:

	A	B	C	D	E	F	G
4		Ech. 1	Ech. 2	Ech. 3	Ech. 4	Ech. 5	Ech. 6
5	Jour 1	9.9943	10.0196	9.9732	10.0088	9.9685	9.9544
6	Jour 2	9.9721	9.9643	9.9426	9.9712	10.0259	10.0177
7	Jour 3	9.9788	10.0213	9.9407	10.0696	10.0161	10.0709
8	Jour 4	10.0658	10.08	9.9514	9.9958	10.0338	10.0422
9	Jour 5	10.0136	9.9791	9.9449	9.9403	9.9424	9.9421
10	Jour 6	9.9495	10.0057	9.9715	10.0388	10.0019	9.9616
11	Jour 7	9.9144	10.0321	10.0361	10.0253	10.0109	9.9935
12	Jour 8	9.9748	9.9962	10.0369	10.15	9.9996	9.9741
13	Jour 9	10.0063	10.0749	9.9971	9.978	9.9934	9.9555
14	Jour 10	10.084	9.9619	10.0063	10.0364	9.9907	9.9762
15	Jour 11	9.9436	10.0638	9.9617	10.0138	9.9672	9.9591
16	Jour 12	10.0074	9.9817	9.9753	10.0541	9.9845	9.9107
17	Jour 13	10.0069	10.0282	10.0385	9.9747	10.0323	9.986
18	Jour 14	9.9836	10.063	9.9824	9.9909	9.9684	10.0575

Figure: 74.110 - Mesures journalières avec 6 individus sans arrêt de la machine

Nous obtenons:

	I	J	K	L	M	N	O
1	λ	0.6000					
2	d2	2.534					
3							
4	N°	Mesures	Range	EWMA	CL	LCL	UCL
5	0	9.9865	0.0652	9.98647	9.9989	9.9735	10.0243
6	1	9.9823	0.0833	9.98397	9.9989	9.9735	10.0243
7	2	10.0162	0.1302	10.00333	9.9989	9.9735	10.0243
8	3	10.0282	0.1286	10.01823	9.9989	9.9735	10.0243
9	4	9.9604	0.0733	9.98353	9.9989	9.9735	10.0243
10	5	9.9882	0.0893	9.98631	9.9989	9.9735	10.0243
11	6	10.0021	0.1217	9.99576	9.9989	9.9735	10.0243
12	7	10.0219	0.1759	10.01146	9.9989	9.9735	10.0243
13	8	10.0009	0.1194	10.00510	9.9989	9.9735	10.0243
14	9	10.0093	0.1221	10.00759	9.9989	9.9735	10.0243
15	10	9.9849	0.1202	9.99396	9.9989	9.9735	10.0243
16	11	9.9856	0.1434	9.98895	9.9989	9.9735	10.0243
17	12	10.0111	0.0638	10.00224	9.9989	9.9735	10.0243
18	13	10.0076	0.0946	10.00548	9.9989	9.9735	10.0243

Figure: 74.111 - Indicateurs correspondants aux mesures obtenues

Soit (nous avons coupé la capture d'écran en deux parties faute de place):

	I	J	K	L	M
1	λ	0.6			
2	d2	2.534			
3					
4	N°	Mesures	Range	EWMA	CL
5	0	=MOYENNE(B5:G5)	=MAX(B5:G5)-MIN(B5:G5)	=J5	=MOYENNE(\$J\$5:\$J\$18)
6	1	=MOYENNE(B6:G6)	=MAX(B6:G6)-MIN(B6:G6)	=J\$1*J6+(1-J\$1)*L5	=MOYENNE(\$J\$5:\$J\$18)
7	2	=MOYENNE(B7:G7)	=MAX(B7:G7)-MIN(B7:G7)	=J\$1*J7+(1-J\$1)*L6	=MOYENNE(\$J\$5:\$J\$18)
8	3	=MOYENNE(B8:G8)	=MAX(B8:G8)-MIN(B8:G8)	=J\$1*J8+(1-J\$1)*L7	=MOYENNE(\$J\$5:\$J\$18)
9	4	=MOYENNE(B9:G9)	=MAX(B9:G9)-MIN(B9:G9)	=J\$1*J9+(1-J\$1)*L8	=MOYENNE(\$J\$5:\$J\$18)
10	5	=MOYENNE(B10:G10)	=MAX(B10:G10)-MIN(B10:G10)	=J\$1*J10+(1-J\$1)*L9	=MOYENNE(\$J\$5:\$J\$18)
11	6	=MOYENNE(B11:G11)	=MAX(B11:G11)-MIN(B11:G11)	=J\$1*J11+(1-J\$1)*L10	=MOYENNE(\$J\$5:\$J\$18)
12	7	=MOYENNE(B12:G12)	=MAX(B12:G12)-MIN(B12:G12)	=J\$1*J12+(1-J\$1)*L11	=MOYENNE(\$J\$5:\$J\$18)
13	8	=MOYENNE(B13:G13)	=MAX(B13:G13)-MIN(B13:G13)	=J\$1*J13+(1-J\$1)*L12	=MOYENNE(\$J\$5:\$J\$18)
14	9	=MOYENNE(B14:G14)	=MAX(B14:G14)-MIN(B14:G14)	=J\$1*J14+(1-J\$1)*L13	=MOYENNE(\$J\$5:\$J\$18)
15	10	=MOYENNE(B15:G15)	=MAX(B15:G15)-MIN(B15:G15)	=J\$1*J15+(1-J\$1)*L14	=MOYENNE(\$J\$5:\$J\$18)
16	11	=MOYENNE(B16:G16)	=MAX(B16:G16)-MIN(B16:G16)	=J\$1*J16+(1-J\$1)*L15	=MOYENNE(\$J\$5:\$J\$18)
17	12	=MOYENNE(B17:G17)	=MAX(B17:G17)-MIN(B17:G17)	=J\$1*J17+(1-J\$1)*L16	=MOYENNE(\$J\$5:\$J\$18)
18	13	=MOYENNE(B18:G18)	=MAX(B18:G18)-MIN(B18:G18)	=J\$1*J18+(1-J\$1)*L17	=MOYENNE(\$J\$5:\$J\$18)

Figure: 74.112 - Formules Microsoft Excel 14.0.6123 explicites du tableau précédent (première partie)

et:

	H	O
1		
2		
3		
4	LCL	UCL
5	=M5-3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B5:G5)*(2+\$J\$1)))	=M5+3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B5:G5)*(2+\$J\$1)))
6	=M6-3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B6:G6)*(2+\$J\$1)))	=M6+3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B6:G6)*(2+\$J\$1)))
7	=M7-3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B7:G7)*(2+\$J\$1)))	=M7+3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B7:G7)*(2+\$J\$1)))
8	=M8-3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B8:G8)*(2+\$J\$1)))	=M8+3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B8:G8)*(2+\$J\$1)))
9	=M9-3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B9:G9)*(2+\$J\$1)))	=M9+3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B9:G9)*(2+\$J\$1)))
10	=M10-3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B10:G10)*(2+\$J\$1)))	=M10+3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B10:G10)*(2+\$J\$1)))
11	=M11-3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B11:G11)*(2+\$J\$1)))	=M11+3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B11:G11)*(2+\$J\$1)))
12	=M12-3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B12:G12)*(2+\$J\$1)))	=M12+3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B12:G12)*(2+\$J\$1)))
13	=M13-3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B13:G13)*(2+\$J\$1)))	=M13+3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B13:G13)*(2+\$J\$1)))
14	=M14-3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B14:G14)*(2+\$J\$1)))	=M14+3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B14:G14)*(2+\$J\$1)))
15	=M15-3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B15:G15)*(2+\$J\$1)))	=M15+3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B15:G15)*(2+\$J\$1)))
16	=M16-3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B16:G16)*(2+\$J\$1)))	=M16+3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B16:G16)*(2+\$J\$1)))
17	=M17-3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B17:G17)*(2+\$J\$1)))	=M17+3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B17:G17)*(2+\$J\$1)))
18	=M18-3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B18:G18)*(2+\$J\$1)))	=M18+3*(MOYENNE(\$K\$5:\$K\$18)/\$J\$2)*RACINE(\$J\$1/(NB(B18:G18)*(2+\$J\$1)))

Figure: 74.113 - Formules Microsoft Excel 14.0.6123 explicites du tableau précédent (deuxième partie)

Nous avons alors la carte de contrôle suivante:

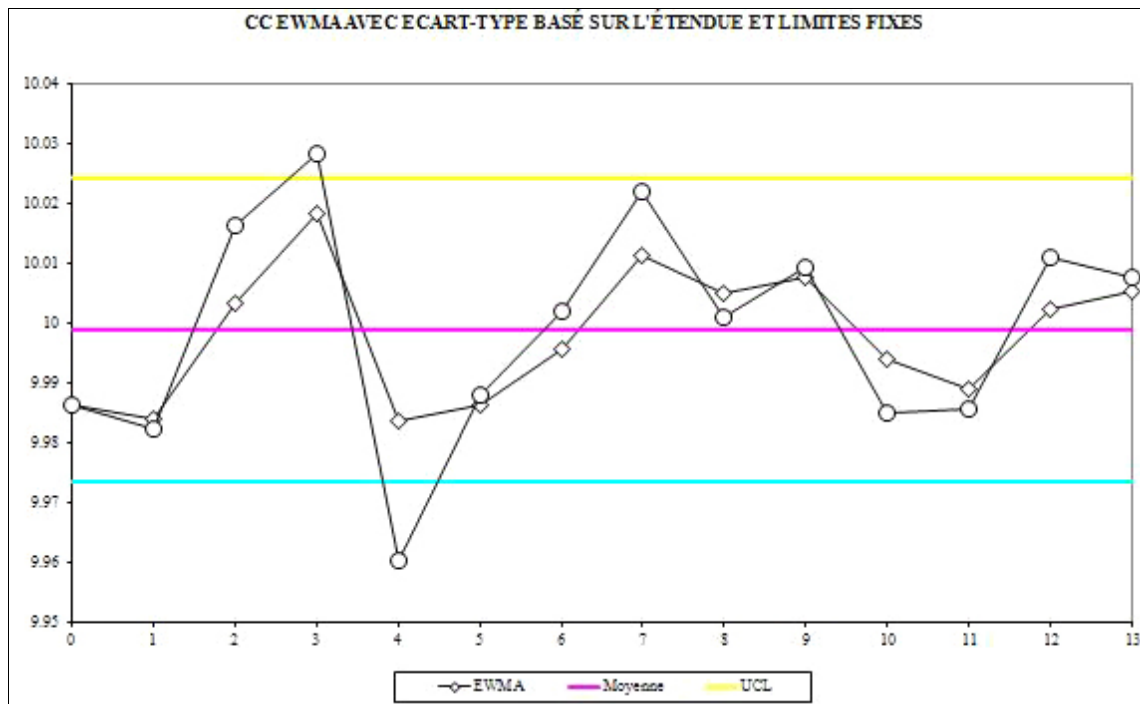


Figure: 74.114 - Carte de contrôle EWMA avec écart-type basé sur l'étendue et limites fixées

Pour le dernier exemple, nous prenons le cas fameux de la carte EWMA avec valeurs individuelles et:

$$\begin{aligned}
 UCL_{EWMA} &= \bar{\bar{X}} + 3 \frac{\overline{MR}}{d_2(2)} \sqrt{\frac{\lambda}{n_i(2+\lambda)}} \\
 CL_{EWMA} &= \bar{\bar{X}} \\
 LCL_{EWMA} &= \bar{\bar{X}} - 3 \frac{\overline{MR}}{d_2(2)} \sqrt{\frac{\lambda}{n_i(2+\lambda)}}
 \end{aligned} \quad (74.563)$$

avec les données suivantes:

	A
4	Pièces
5	9.9943
6	10.0196
7	9.9732
8	10.0088
9	9.9685
10	9.9544
11	9.9721
12	9.9643

Figure: 74.115 - Mesures individuelles sans arrêt de la machine

en prenant donc (attention le n dans la relation qui suit n'est pas le même que celui qui est dans la racine contenant la constante de lissage!):

$$\overline{MR} = \bar{R} = \frac{1}{n-1} \sum_{i=2}^n |x_i - x_{i-1}| \quad (74.564)$$

et:

$$d_2(2) = 1.128 \quad (74.565)$$

ainsi que:

$$\bar{F}_0 = \bar{X} \quad (74.566)$$

Nous avons alors avec une constante de lissage de 0.6:

	C	D	E	F	G	H
1	λ	0.60				
2	d_2	1.128				
3						
4	N°	Etendue Mobile	EWMA	CL	LCL	UCL
5	0		9.9893	9.9819	9.9477	10.0161
6	1	0.0253	10.0075	9.9819	9.9477	10.0161
7	2	0.0464	9.9869	9.9819	9.9477	10.0161
8	3	0.0356	10.0000	9.9819	9.9477	10.0161
9	4	0.0403	9.9811	9.9819	9.9477	10.0161
10	5	0.0141	9.9651	9.9819	9.9477	10.0161
11	6	0.0177	9.9693	9.9819	9.9477	10.0161
12	7	0.0078	9.9663	9.9819	9.9477	10.0161

Figure: 74.116 - Indicateurs correspondants aux mesures obtenues

Soit (nous avons coupé la capture d'écran en deux parties faute de place):

	C	D	E	F
1	λ	0.6		
2	d_2	1.128		
3				
4	N°	Etendue Mob	EWMA	CL
5	0		= \$D\$1*\$A5*(1-\$D\$1)*MOYENNE(\$A\$5:\$A\$12)	=MOYENNE(\$A\$5:\$A\$12)
6	1	=ABS(A6-A5)	= \$D\$1*\$A6*(1-\$D\$1)*E5	=MOYENNE(\$A\$5:\$A\$12)
7	2	=ABS(A7-A6)	= \$D\$1*\$A7*(1-\$D\$1)*E6	=MOYENNE(\$A\$5:\$A\$12)
8	3	=ABS(A8-A7)	= \$D\$1*\$A8*(1-\$D\$1)*E7	=MOYENNE(\$A\$5:\$A\$12)
9	4	=ABS(A9-A8)	= \$D\$1*\$A9*(1-\$D\$1)*E8	=MOYENNE(\$A\$5:\$A\$12)
10	5	=ABS(A10-A9)	= \$D\$1*\$A10*(1-\$D\$1)*E9	=MOYENNE(\$A\$5:\$A\$12)
11	6	=ABS(A11-A10)	= \$D\$1*\$A11*(1-\$D\$1)*E10	=MOYENNE(\$A\$5:\$A\$12)
12	7	=ABS(A12-A11)	= \$D\$1*\$A12*(1-\$D\$1)*E11	=MOYENNE(\$A\$5:\$A\$12)

Figure: 74.117 - Formules Microsoft Excel 14.0.6123 explicites du tableau précédent (première partie)

et:

	G	H
1		
2		
3		
4	LCL	UCL
5	=F5-3*MOYENNE(\$D\$5:\$D\$12)/\$D\$2*RACINE(\$D\$1/(2*\$D\$1))	=F5+3*MOYENNE(\$D\$5:\$D\$12)/\$D\$2*RACINE(\$D\$1/(2*\$D\$1))
6	=F6-3*MOYENNE(\$D\$5:\$D\$12)/\$D\$2*RACINE(\$D\$1/(2*\$D\$1))	=F6+3*MOYENNE(\$D\$5:\$D\$12)/\$D\$2*RACINE(\$D\$1/(2*\$D\$1))
7	=F7-3*MOYENNE(\$D\$5:\$D\$12)/\$D\$2*RACINE(\$D\$1/(2*\$D\$1))	=F7+3*MOYENNE(\$D\$5:\$D\$12)/\$D\$2*RACINE(\$D\$1/(2*\$D\$1))
8	=F8-3*MOYENNE(\$D\$5:\$D\$12)/\$D\$2*RACINE(\$D\$1/(2*\$D\$1))	=F8+3*MOYENNE(\$D\$5:\$D\$12)/\$D\$2*RACINE(\$D\$1/(2*\$D\$1))
9	=F9-3*MOYENNE(\$D\$5:\$D\$12)/\$D\$2*RACINE(\$D\$1/(2*\$D\$1))	=F9+3*MOYENNE(\$D\$5:\$D\$12)/\$D\$2*RACINE(\$D\$1/(2*\$D\$1))
10	=F10-3*MOYENNE(\$D\$5:\$D\$12)/\$D\$2*RACINE(\$D\$1/(2*\$D\$1))	=F10+3*MOYENNE(\$D\$5:\$D\$12)/\$D\$2*RACINE(\$D\$1/(2*\$D\$1))
11	=F11-3*MOYENNE(\$D\$5:\$D\$12)/\$D\$2*RACINE(\$D\$1/(2*\$D\$1))	=F11+3*MOYENNE(\$D\$5:\$D\$12)/\$D\$2*RACINE(\$D\$1/(2*\$D\$1))
12	=F12-3*MOYENNE(\$D\$5:\$D\$12)/\$D\$2*RACINE(\$D\$1/(2*\$D\$1))	=F12+3*MOYENNE(\$D\$5:\$D\$12)/\$D\$2*RACINE(\$D\$1/(2*\$D\$1))

Figure: 74.118 - Formules Microsoft Excel 14.0.6123 explicites du tableau précédent (deuxième partie)

Nous avons alors la carte de contrôle suivante:

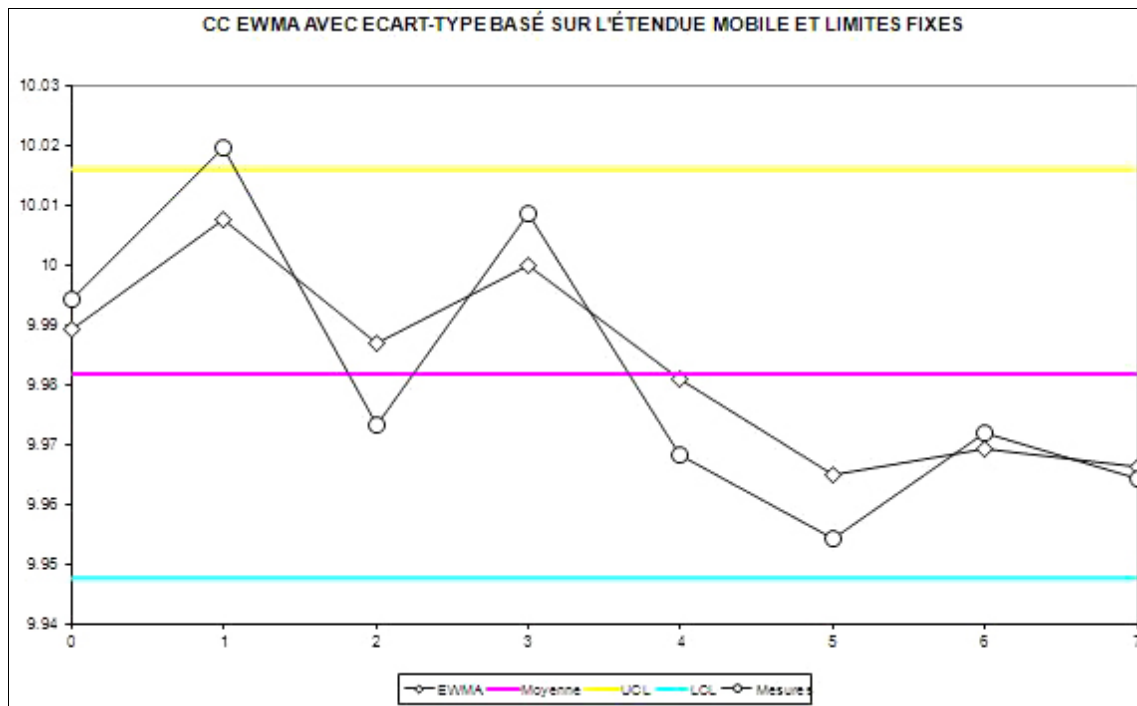


Figure: 74.119 - Carte de contrôle EWMA avec écart-type basé sur l'étendue et limites fixées

Comme pour les autres cartes de contrôle, la carte EWMA doit être accompagnée d'une carte de contrôle pour le suivi de la variabilité, donc soit une carte avec étendues aux limites fixes ou une carte aux écarts-types avec limites fixes.

5.6. CARTE DE CONTRÔLE DES FRÉQUENCES AVEC LIMITES PROBABILISTES

Le suivi des fréquences d'apparition de défauts/accidents/anomalies/pannes est une technique très simple à mettre en pratique et très parlante pour des personnes ayant peu voire pas de connaissances en statistiques. De plus c'est une technique adaptée aux cas où les événements contrôlés sont très rares.

L'idée est alors de ne plus suivre le nombre d'événements spéciaux, mais le temps T entre deux apparitions des événements surveillés. Le temps T se mesure alors soit en nombre de pièces produites, soit en nombre d'accidents, anomalies, en jours, etc.

Par exemple:

	A	B
	N° d'apparition de défaut	Nombre de pièces produites jusqu'au défaut
1		
2	1	1500
3	2	166
4	3	2501
5	4	238
6	5	1392
7	6	4686
8	7	128
9	8	31
10	9	2566
11	10	591
12	11	470
13	12	502
14	13	4147
15	14	1131
16	15	884
17	16	1864
18	17	665
19	18	3364
20	19	881
21	20	4726
22	21	517
23	22	5787
24	23	1250
25	24	2869
26	25	7087
27	26	18653
28	27	7772
29	28	1683

Figure: 74.120 - Mesures

Ce qui donne sous forme graphique:

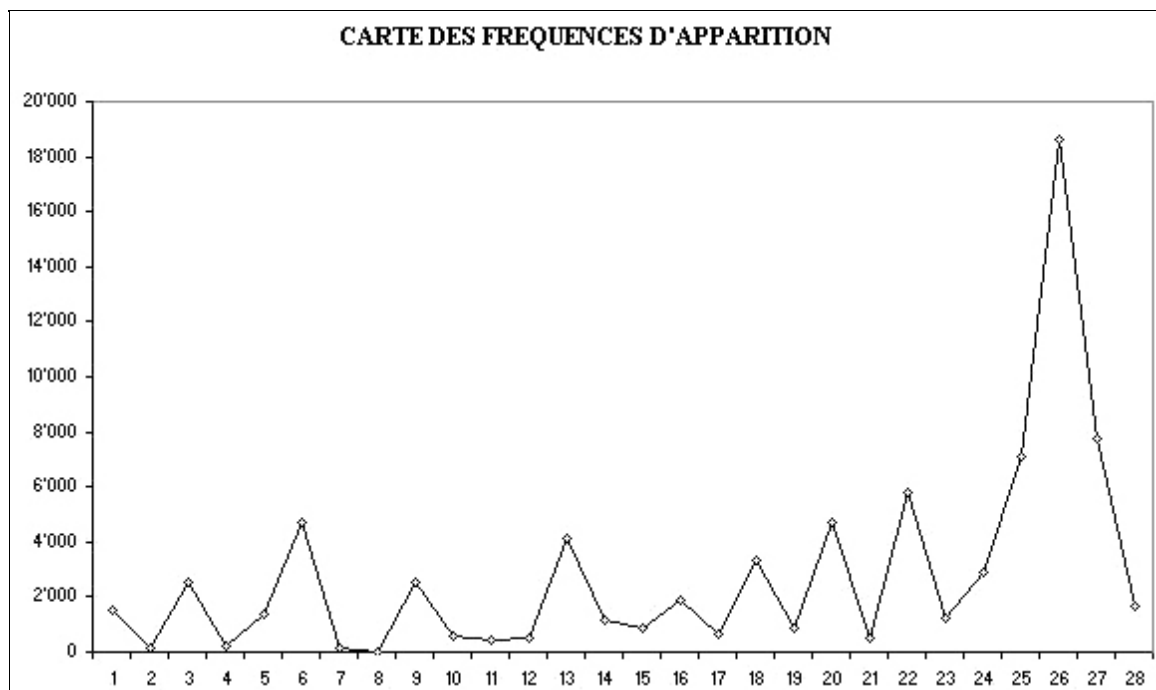


Figure: 74.121 - Carte de contrôle des fréquences

Bon c'est bien joli, mais nous il nous faut une *UCL* et *LCL* pour que cela soit utile! Afin d'y arriver nous allons utiliser la théorie des files d'attente avec les mêmes hypothèses et les mêmes contrôles préalables (cf. chapitre Techniques de Gestion). Rappelons d'abord que nous y avons démontré que la probabilité d'observer k événements dans un intervalle de durée t (ou un intervalle de nombre de pièces t), sous certaines hypothèses bien précises (!!!), suivait une distribution de Poisson:

$$p_k \cong \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad (74.567)$$

où le paramètre λ est le taux moyen d'arrivée (taux moyen d'apparition) d'événements par unité de temps (ou en anglais "**Poisson Arrivals See Time Average**": PASTA...) supposé constant sur toute la durée:

$$\lambda = \frac{1}{\bar{T}} \quad (74.568)$$

et où nous avons pour espérance et variance (cf. chapitre de Statistiques) du nombre d'événements:

$$E[k] = \lambda \cdot t \quad Var(k) = \lambda \cdot t \quad (74.569)$$

Il s'agit d'une loi de probabilité à support discret, ce qui correspond bien à notre cas d'études donc la loi de Poisson peut être un bon candidat (mais il y en a d'autres, comme la loi Géométrique par exemple).

Mais nous avons aussi démontré que la variable aléatoire τ représentant le temps (ou le nombre de pièces) séparant deux arrivées d'événements non désirés (ou désirés...) était donnée par la fonction de répartition basé sur la loi exponentielle:

$$A(\tau \leq t) = 1 - e^{-\lambda t} \quad (74.570)$$

toujours avec:

$$\lambda = \frac{1}{\bar{T}} \quad (74.571)$$

où:

$$\mu = \sigma = \frac{1}{\lambda} = \bar{T} \quad (74.572)$$

Donc, nous pourrions prendre les limites de Shewhart suivantes:

$$\begin{aligned} USL &= \text{imposé par client/direction} \\ UCL &= \bar{T} + 3\bar{T} = 4\bar{T} \\ CL &= \bar{T} \\ LCL &= \bar{T} - 3\bar{T} \equiv 0 \end{aligned} \quad (74.573)$$

Mais il est plus rigoureux (comme pour toutes les autres cartes) de prendre des limites probabilistes. Ainsi, si nous voulons fixer les limites avec un risque bilatéral habituel dans le domaine de $\alpha/2$ de 0.5% nous calculons alors pour la limite inférieure:

$$\frac{\alpha}{2} = 1 - e^{-\lambda x} \Rightarrow x = \frac{\ln\left(1 - \frac{\alpha}{2}\right)}{-\lambda} = -\ln\left(1 - \frac{0.25\%}{2}\right) \bar{T} \cong 0.012 \bar{T} \quad (74.574)$$

et pour la limite supérieure:

$$1 - \frac{\alpha}{2} = 1 - e^{-\lambda x} \Rightarrow x = \frac{\ln\left(\frac{\alpha}{2}\right)}{-\lambda} = -\ln\left(\frac{0.25\%}{2}\right) \bar{T} \cong 4.382 \bar{T} \quad (74.575)$$

Nous avons alors:

	A	B	C	D
1	N° d'apparition de défaut	Nombre de pièces produites jusqu'au défaut	UCL	LCL
2	1	1500	12214.98	33.45
3	2	166	12214.98	33.45
4	3	2501	12214.98	33.45
5	4	238	12214.98	33.45
6	5	1392	12214.98	33.45
7	6	4686	12214.98	33.45
8	7	128	12214.98	33.45
9	8	31	12214.98	33.45
10	9	2566	12214.98	33.45
11	10	591	12214.98	33.45
12	11	470	12214.98	33.45
13	12	502	12214.98	33.45
14	13	4147	12214.98	33.45
15	14	1131	12214.98	33.45
16	15	884	12214.98	33.45
17	16	1864	12214.98	33.45
18	17	665	12214.98	33.45
19	18	3364	12214.98	33.45
20	19	881	12214.98	33.45
21	20	4726	12214.98	33.45
22	21	517	12214.98	33.45
23	22	5787	12214.98	33.45
24	23	1250	12214.98	33.45
25	24	2869	12214.98	33.45
26	25	7087	12214.98	33.45
27	26	18653	12214.98	33.45
28	27	7772	12214.98	33.45
29	28	1683	12214.98	33.45

Figure: 74.122 - Indicateurs correspondants aux mesures obtenues

Soit:

	C	D
1	UCL	LCL
2	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
3	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
4	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
5	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
6	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
7	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
8	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
9	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
10	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
11	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
12	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
13	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
14	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
15	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
16	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
17	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
18	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
19	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
20	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
21	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
22	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
23	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
24	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
25	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
26	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
27	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
28	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)
29	=4.382*MOYENNE(\$B\$2:\$B\$29)	=0.012*MOYENNE(\$B\$2:\$B\$29)

Figure: 74.123 - Formules Microsoft Excel 14.0.6123 explicites du tableau précédent

Nous avons alors la carte de contrôle suivante (nous n'y avons exceptionnellement pas représenté *USL*):

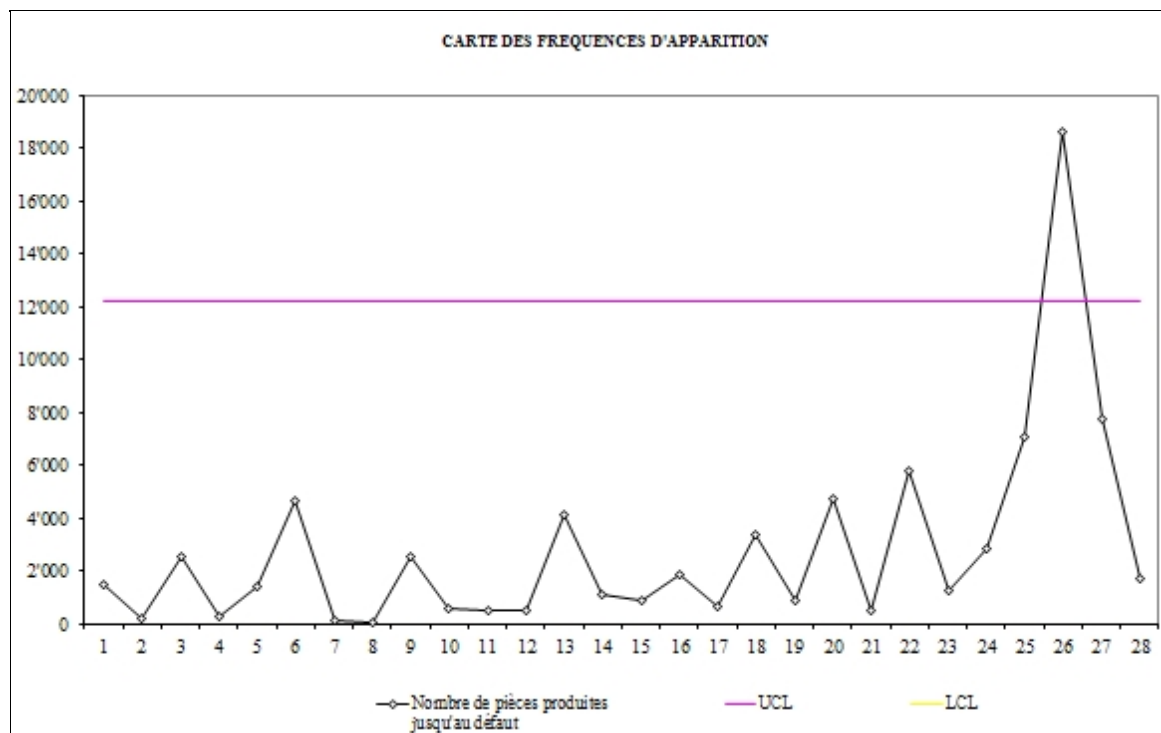


Figure: 74.124 - Carte de contrôle des fréquences limites

Il ne faut régulièrement (et donc aussi au tout début) vérifier que la distribution des événements suit bien une distribution de Poisson avant d'appliquer cette carte de contrôle!

5.7. CARTE DE CONTRÔLE DES ÉVÉNEMENTS RARES (CARTE-G)

Nous avons démontré dans le chapitre de Statistique que l'espérance d'avoir R réussites avant le E -ème échec sachant que la probabilité d'avoir un échec est p était donnée par l'espérance de la loi binomiale négative:

$$E(R) = \frac{E}{p} \quad (74.576)$$

et nous avons obtenu pour la variance:

$$V(R) = E \frac{q}{p^2} \quad (74.577)$$

Bien évidemment, si nous posons E comme étant le premier échec, l'espérance se réduit à celle de la loi géométrique:

$$E(R) = \frac{1}{p} \text{ et } V(R) = \frac{q}{p^2} = \frac{1-p}{p^2} \quad (74.578)$$

Nous avons démontré dans le chapitre de Statistiques que l'estimateur du paramètre de la loi géométrique était donné par:

$$\hat{p} = \frac{k}{N} \quad (74.579)$$

Il vient alors:

$$E(R) = \frac{N}{k} = \bar{g} \quad (74.580)$$

et:

$$V(R) = \frac{1 - \left(\frac{k}{N}\right)}{\left(\frac{k}{N}\right)^2} = \left(\frac{N}{k}\right)^2 \left(1 - \left(\frac{k}{N}\right)\right) = \left(\frac{N}{k}\right)^2 - \frac{N}{k} = \frac{N}{k} \left(\frac{N}{k} - 1\right) = \bar{g}(\bar{g} - 1) \quad (74.581)$$

et donc:

$$\hat{\sigma}_R = \sqrt{\bar{g}(\bar{g} - 1)} \quad (74.582)$$

Dès lors, nous avons:

$$\begin{aligned} USL &= \text{Imposé par la direction} \\ UCL &= \bar{g} + 3\sqrt{\bar{g}(\bar{g} - 1)} \\ CL &= \bar{g} \\ LCL &= \bar{g} - 3\sqrt{\bar{g}(\bar{g} - 1)} \equiv 0 \end{aligned} \quad (74.583)$$

où si LCL est négatif, nous le posons comme étant égal à zéro.

Voyons un exemple en considérant la table suivante dans Microsoft Excel 14.0.6123:

	A	B	C	D	E
1	Date	Jours entre événements	g	LCL	UCL
2	05.02.2009		39.24	0	155.43
3	03.04.2009	57	39.24	0	155.43
4	25.04.2009	22	39.24	0	155.43
5	06.05.2009	11	39.24	0	155.43
6	15.09.2009	132	39.24	0	155.43
7	21.09.2009	6	39.24	0	155.43
8	07.10.2009	16	39.24	0	155.43
9	22.11.2009	46	39.24	0	155.43
10	01.12.2009	9	39.24	0	155.43
11	05.01.2010	35	39.24	0	155.43
12	30.01.2010	25	39.24	0	155.43
13	07.04.2010	67	39.24	0	155.43
14	06.04.2010	58	39.24	0	155.43
15	24.06.2010	20	39.24	0	155.43
16	05.07.2010	11	39.24	0	155.43
17	03.08.2010	49	39.24	0	155.43
18	08.10.2010	46	39.24	0	155.43
19	04.12.2010	57	39.24	0	155.43

Figure: 74.125 - Indicateurs correspondants aux mesures obtenues

Avec les formules suivantes:

C	D	E
g	LCL	UCL
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C2-3*RACINE(C2*(C2-1))<0;0;C2-3*RACINE(C2*(C2-1)))	=C2+3*RACINE(C2*(C2-1))
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C3-3*RACINE(C3*(C3-1))<0;0;C3-3*RACINE(C3*(C3-1)))	=C3+3*RACINE(C3*(C3-1))
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C4-3*RACINE(C4*(C4-1))<0;0;C4-3*RACINE(C4*(C4-1)))	=C4+3*RACINE(C4*(C4-1))
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C5-3*RACINE(C5*(C5-1))<0;0;C5-3*RACINE(C5*(C5-1)))	=C5+3*RACINE(C5*(C5-1))
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C6-3*RACINE(C6*(C6-1))<0;0;C6-3*RACINE(C6*(C6-1)))	=C6+3*RACINE(C6*(C6-1))
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C7-3*RACINE(C7*(C7-1))<0;0;C7-3*RACINE(C7*(C7-1)))	=C7+3*RACINE(C7*(C7-1))
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C8-3*RACINE(C8*(C8-1))<0;0;C8-3*RACINE(C8*(C8-1)))	=C8+3*RACINE(C8*(C8-1))
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C9-3*RACINE(C9*(C9-1))<0;0;C9-3*RACINE(C9*(C9-1)))	=C9+3*RACINE(C9*(C9-1))
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C10-3*RACINE(C10*(C10-1))<0;0;C10-3*RACINE(C10*(C10-1)))	=C10+3*RACINE(C10*(C10-1))
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C11-3*RACINE(C11*(C11-1))<0;0;C11-3*RACINE(C11*(C11-1)))	=C11+3*RACINE(C11*(C11-1))
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C12-3*RACINE(C12*(C12-1))<0;0;C12-3*RACINE(C12*(C12-1)))	=C12+3*RACINE(C12*(C12-1))
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C13-3*RACINE(C13*(C13-1))<0;0;C13-3*RACINE(C13*(C13-1)))	=C13+3*RACINE(C13*(C13-1))
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C14-3*RACINE(C14*(C14-1))<0;0;C14-3*RACINE(C14*(C14-1)))	=C14+3*RACINE(C14*(C14-1))
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C15-3*RACINE(C15*(C15-1))<0;0;C15-3*RACINE(C15*(C15-1)))	=C15+3*RACINE(C15*(C15-1))
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C16-3*RACINE(C16*(C16-1))<0;0;C16-3*RACINE(C16*(C16-1)))	=C16+3*RACINE(C16*(C16-1))
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C17-3*RACINE(C17*(C17-1))<0;0;C17-3*RACINE(C17*(C17-1)))	=C17+3*RACINE(C17*(C17-1))
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C18-3*RACINE(C18*(C18-1))<0;0;C18-3*RACINE(C18*(C18-1)))	=C18+3*RACINE(C18*(C18-1))
=SOMME(\$B\$2:\$B\$19)/NB(\$A\$3:\$A\$19)	=SI(C19-3*RACINE(C19*(C19-1))<0;0;C19-3*RACINE(C19*(C19-1)))	=C19+3*RACINE(C19*(C19-1))

Figure: 74.126 - Formules Microsoft Excel 14.0.6123 explicites du tableau précédent

Ce qui donne sous forme graphique:

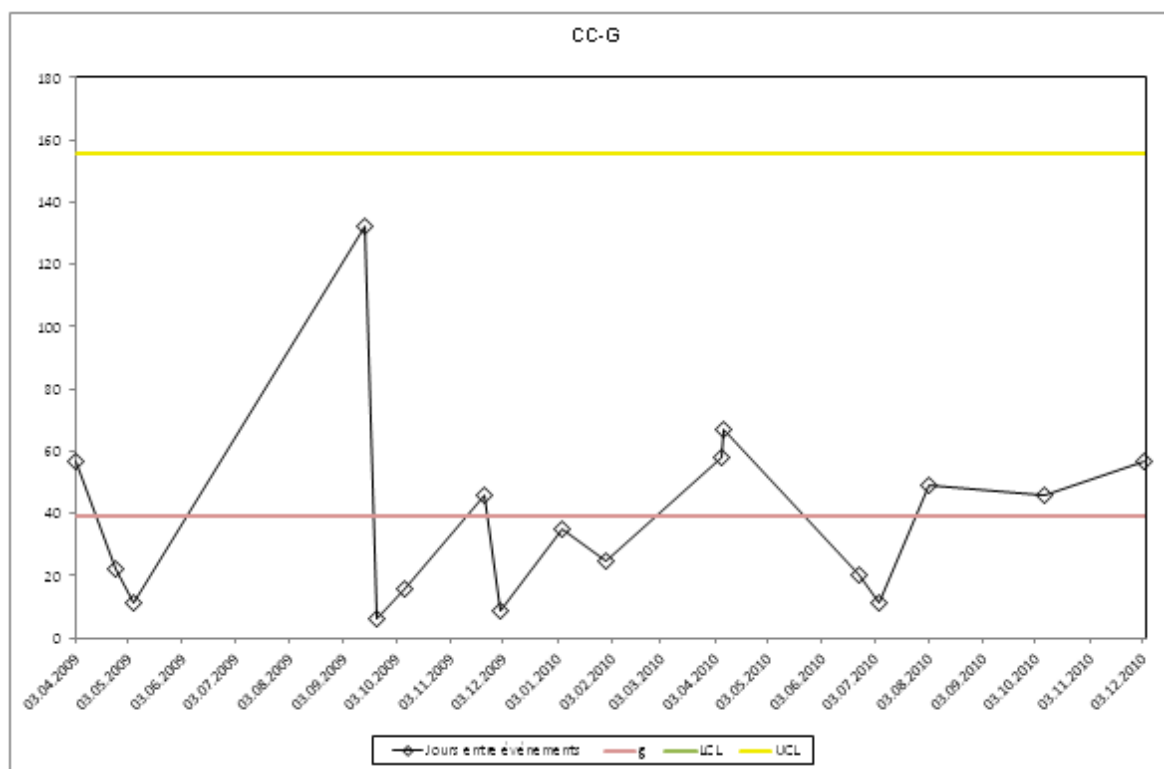
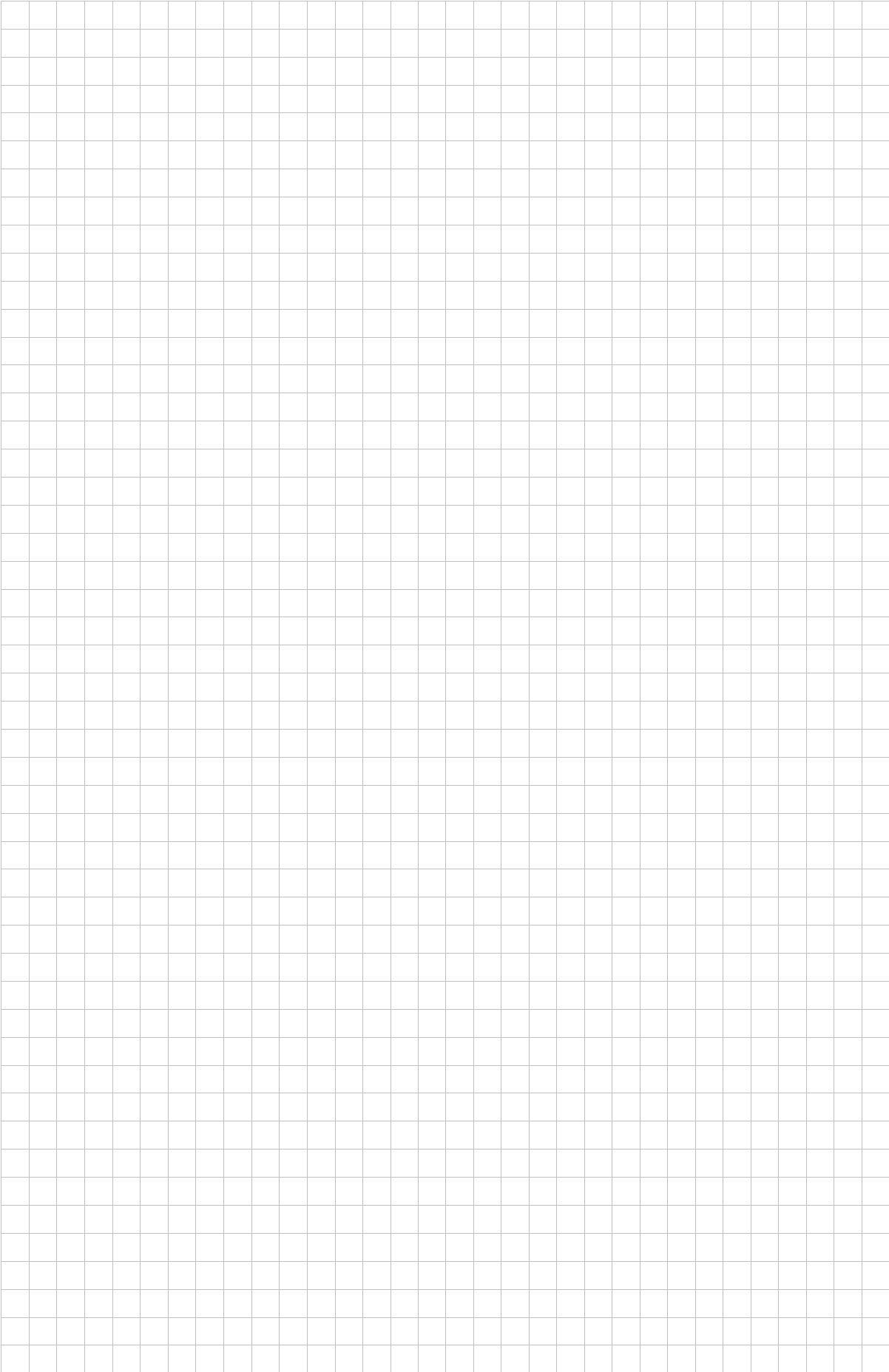
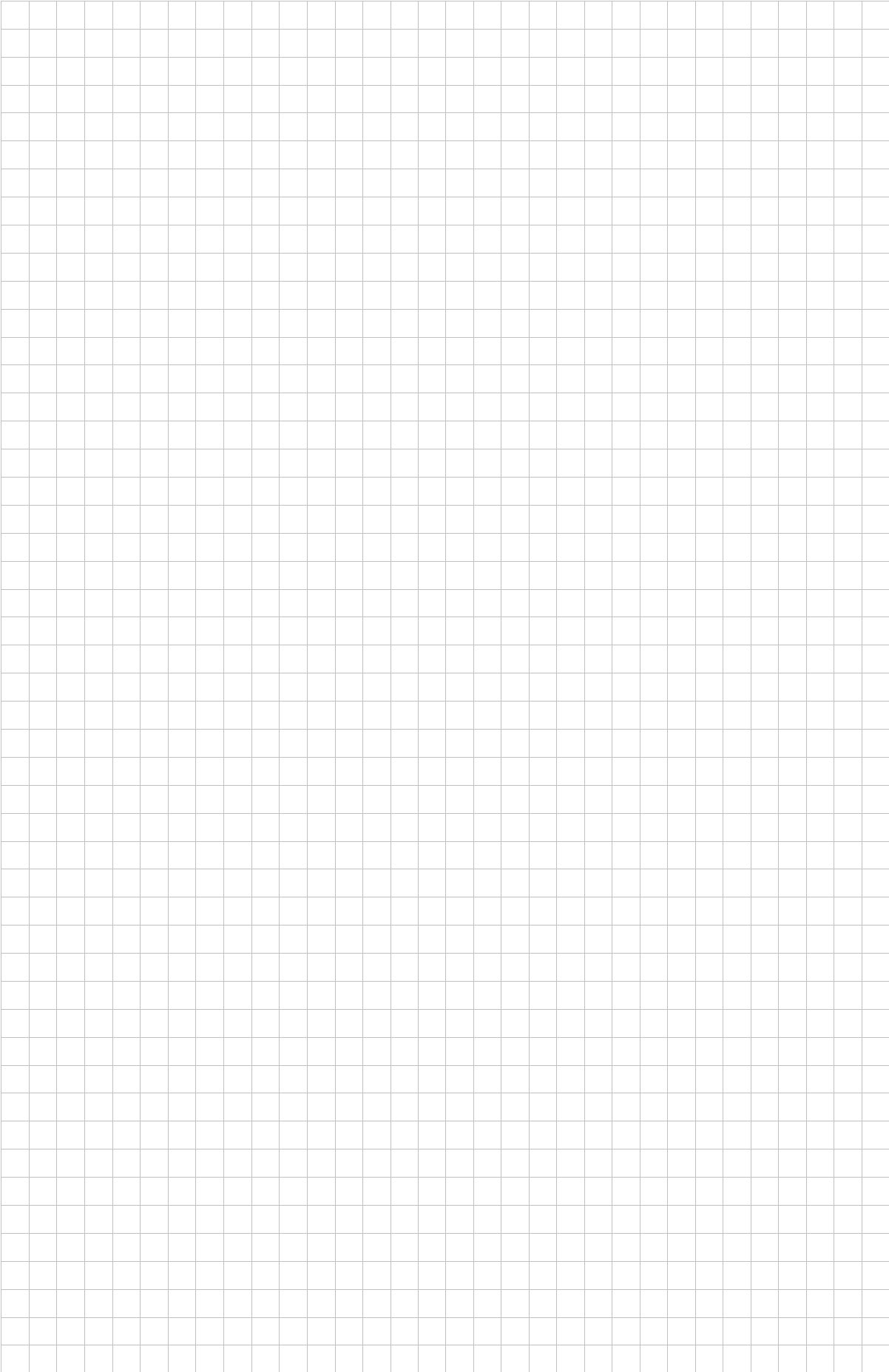


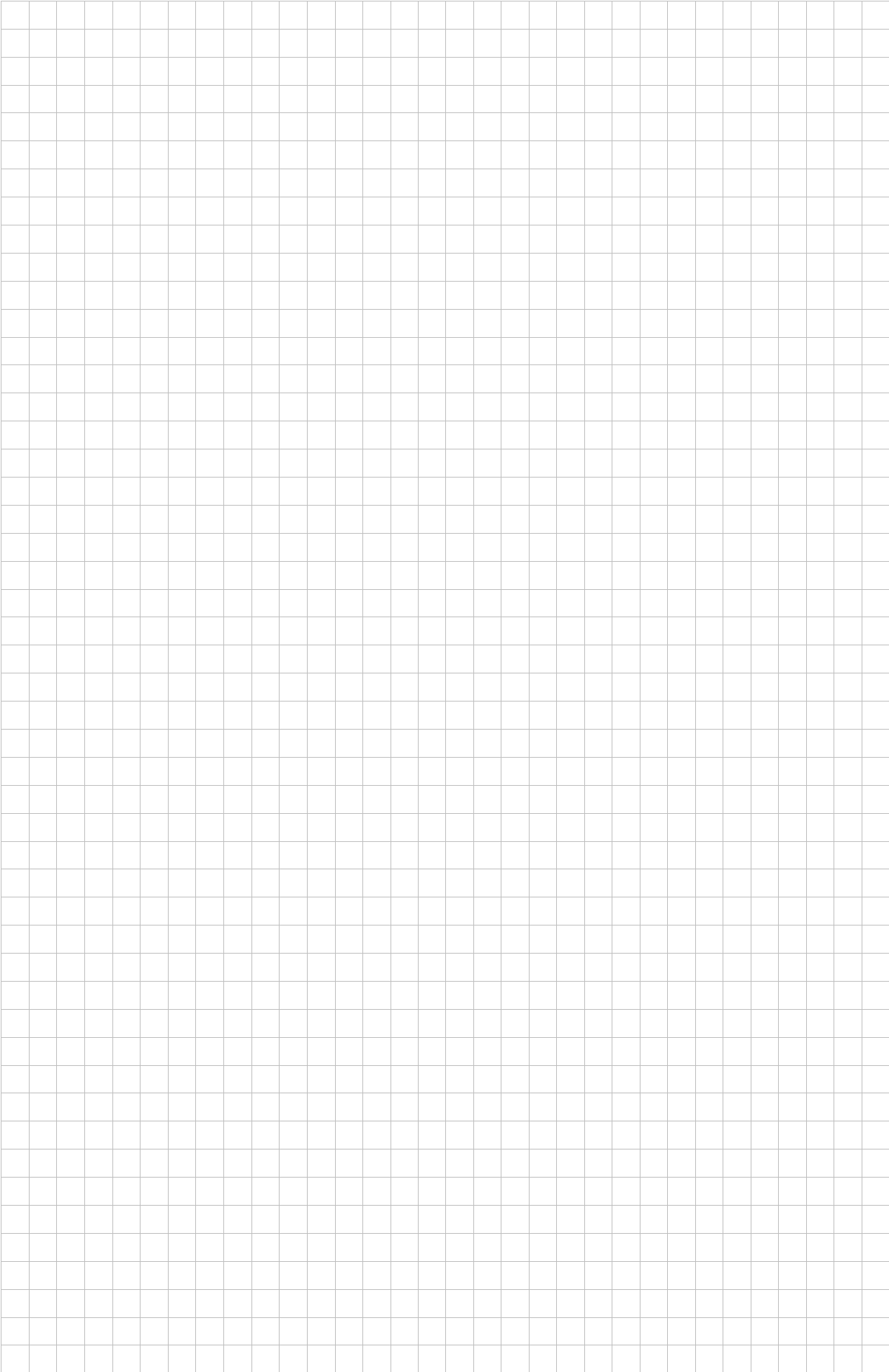
Figure: 74.127 - Carte de contrôle G

Cependant, il convient de se rappeler que la loi Géométrique (ou la loi Binomiale négative) n'est pas nécessairement symétrique. Raison pour laquelle, dans la pratique, il vaut mieux utiliser la médiane pour la CL (médiane que l'on déterminera en faisant une simple simulation de la loi et en prenant la valeur la plus proche de 50%) et pour les limites de contrôle inférieure LCL et supérieure UCL utiliser des limites probabilistes avec les mêmes simulations.

Remarque: Pour un logiciel comme Minitab, la limite de contrôle inférieure de la carte des événements rare est définie au percentile 0.0013499 de la distribution géométrique (en prenant la valeur décimale la plus proche par interpolation linéaire). La limite de contrôle supérieure est définie au percentile 0.99865 (en prenant aussi la valeur décimale la plus proche par interpolation linéaire). La ligne centrale est définie au percentile 0.5 (en prenant la valeur décimale la plus proche par interpolation linéaire).







75. GÉNIE LOGICIEL

Le Génie Logiciel sera défini dans ce chapitre comme l'ensemble des techniques ou approches de résolution de problèmes utilisant fortement des mathématiques de niveau universitaire et que nous retrouvons (ou trouvions...) fréquemment implémentés dans de nombreuses solutions informatiques.

Le lecteur retrouvera dans les textes qui vont suivre des outils mathématiques dont les détails ont tous été démontrés dans d'autres chapitres, mais dont les mises en applications informatiques n'avaient pas leur place dans la partie purement théorique.

Remarque: Si le lecteur cherche les méthodes mathématiques pour simuler des objets 3D sur un écran, il devra se reporter au chapitre de Géométrie Projective.

1. ALGORITHME PAGERANK DE GOOGLE

Lors de sa création en 1998 Google (moteur de recherche actuellement très connu des internautes) a dominé et domine encore en 2011 le marché des moteurs de recherche sur Internet. Son point fort initial est basé sur l'idée très simple de trier les résultats par ordre de pertinence, ce que ne faisaient pas les autres grands moteurs de recherche avant lui.

L'idée principale du principe de fonctionnement initial de Google est intéressante à présenter sur ce site, car elle est un exemple pratique de concepts purement mathématiques que nous avons présentés jusqu'ici (théorie des graphes et chaînes de Markov dans les chapitre respectifs de Théorie des Graphes et de Probabilités) et qui ont créé l'une des plus grandes entreprises internationales dans les nouvelles technologies dans les années 2000 (Google gagne des milliards avec l'algèbre linéaire...)! Comme c'est souvent le cas dans l'économie de marché, ce sont d'abord des recherches théoriques (mathématiques ou physiques) qui ont permis de développer de nouveaux outils!

Remarques:

R1. Évidemment l'algorithme de base a changé depuis 1998 car on peut y rajouter beaucoup de paramètres mathématiques (pondération par une distance par exemple) ou empiriques (visites, clics, publicités ...).

R2. Cet algorithme n'est pas utilisé que dans le domaine du web. Effectivement, il peut être utilisé pour mettre en évidence n'importe quel objet en relation multiple avec d'autres objets (par exemple des variables dans un projet qui s'influencent les unes des autres en ayant des relations directes ou indirectes: une analyse structurelle).

Pour trouver une information dans le tas amorphe et non vraiment structuré qu'est le web (vu le peu de gens qui respectent correctement les standards émis par le W3C...), l'utilisateur lancera une recherche par mots-clés. Ceci nécessite bien évidemment une certaine préparation pour être efficace: le moteur de recherche copie préalablement les pages web en mémoire locale et trie les mots (ou composition de mots) par ordre alphabétique en utilisant les algorithmes d'index empiriques traditionnels du domaine de

l'informatique (arbres-b, arbres-b+ ou autres encore...). Le résultat est un annuaire de mots-clés avec leurs pages web associées.

Afin d'analyser comment Google identifie les pages web les plus pertinentes, nous allons négliger le contenu des pages et ne compter que les liens entre elles. Ce que nous obtenons alors est la structure d'un graphe (non nécessairement connexe!).

L'image suivante montre un exemple miniature (il y aurait $25 \cdot 10^9$ sites web d'après Google début 2007):

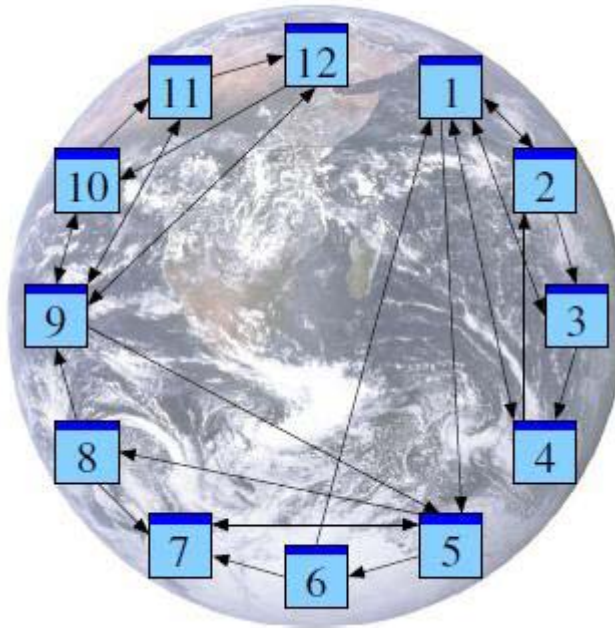


Figure: 75.1 - Principe schématique du graphe à étudier

Pour la suite, nous noterons les pages web par:

$$P_1, P_2, P_3, \dots, P_n \quad (75.1)$$

et nous écrirons:

$$j \rightarrow i \quad (75.2)$$

si la page P_j cite directement la page P_i (citation directement connexe). Ainsi, dans l'image précédente, nous avons un lien $1 \rightarrow 5$, par exemple, mais pas de lien $5 \rightarrow 1$.

Bien évidemment, un lien $j \rightarrow i$ est une recommandation de la page P_j d'aller lire la page P_i . C'est ainsi un vote de P_j en faveur de l'autorité de la page P_i (nous comprenons alors aisément pourquoi l'échange de liens entre sites n'ayant aucun rapport contextuel ait été en quelque sorte bannie par Google).

Analysons l'image précédente sous un aspect plus conforme à la théorie des graphes, dont la hiérarchie devra être justifiée par la suite:

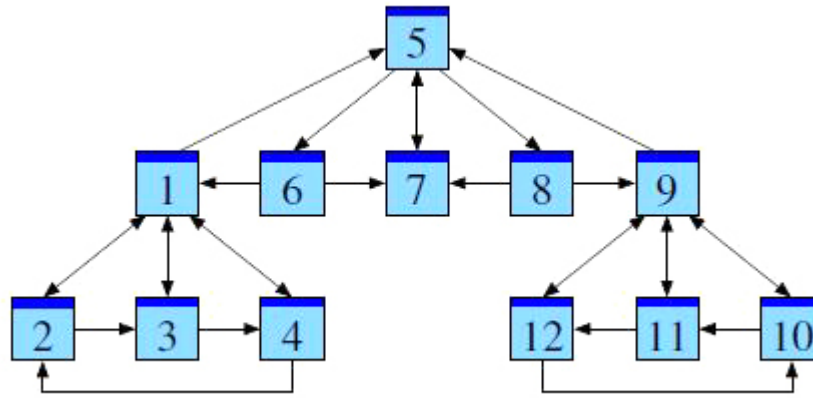


Figure: 75.2 - Graphe précédent réarrangé

Pour des raisons purement pédagogiques, nous allons introduire trois modèles théoriques de ranking de pages web. Le dernier modèle étant le plus proche de celui implémenté par Google en 1998 et les deux premiers modèles permettant de préparer le terrain.

1.1. COMPTAGE PONDÉRÉ

Il est fortement plausible que si une page est importante, elle reçoit beaucoup de liens. Avec un peu de naïveté, on peut penser que la réciproque est vraie (une page qui reçoit beaucoup de liens est importante). Ainsi, nous pourrions définir l'importance μ_i d'une page P_i comme étant le nombre de liens $j \rightarrow i$ directement connexes. Sous forme mathématique, nous écrirons cela ainsi:

$$\mu_i = \sum_{j \rightarrow i} 1 \quad (75.3)$$

Autrement dit, μ_i est égal au nombre de votes pour la page P_i , où chaque vote contribue par la même valeur 1. Cela est, certes, facile à calculer, mais ne correspond souvent pas à l'importance ressentie par l'utilisateur.

Dans notre exemple, nous avons ainsi:

$$\mu_1 = \mu_9 = 4 \quad (75.4)$$

devant:

$$\mu_5 = \mu_7 = 3 \quad (75.5)$$

ce qui est pire, c'est que ce comptage naïf est trop facile à manipuler en ajoutant des pages (d'un autre nom de domaine internet) sans intérêt recommandant une page quelconque.

Remarque: Donc le poids maximum est égal in extenso à $n-1$

Certaines pages émettent beaucoup de liens. Ceux-ci semblent donc moins spécifiques (spécialisés) et leurs poids sera dès lors plus faible. Nous partageons donc le vote de la page P_j en l_j parts égales, où l_j dénote le nombre de liens émis. Ainsi, nous pouvons définir une mesure plus fine:

$$\mu_i = \sum_{j \rightarrow i} \frac{1}{l_j} \quad (75.6)$$

Autrement dit, μ_i compte le nombre de votes pondérés pour la page P_i . C'est facile à calculer, mais nous restons toujours avec le problème susmentionné.

Dans notre exemple, nous avons ainsi:

$$\begin{aligned} \mu_1 &= \sum_{j \rightarrow 1} \frac{1}{l_j} = \frac{1}{l_2} + \frac{1}{l_3} + \frac{1}{l_4} + \frac{1}{l_6} = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 2 \\ \mu_9 &= \sum_{j \rightarrow 9} \frac{1}{l_j} = \frac{1}{l_8} + \frac{1}{l_{10}} + \frac{1}{l_{11}} + \frac{1}{l_{12}} = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 2 \end{aligned} \quad (75.7)$$

et:

$$\mu_5 = \sum_{j \rightarrow 5} \frac{1}{l_j} = \frac{1}{l_1} + \frac{1}{l_7} + \frac{1}{l_9} = \frac{1}{4} + \frac{1}{1} + \frac{1}{4} = 1.5 \quad (75.8)$$

et:

$$\mu_7 = \sum_{j \rightarrow 7} \frac{1}{l_j} = \frac{1}{l_5} + \frac{1}{l_6} + \frac{1}{l_8} = \frac{1}{3} + \frac{1}{2} + \frac{1}{2} = \frac{4}{3} \quad (75.9)$$

Remarque: Si toutes les pages étaient directement connexes, nous aurions alors $\mu_i = 1$ pour tout i .

1.2. COMPTAGE RÉCURSIF

Heuristiquement, une page P_i paraît importante si beaucoup de pages importantes la citent. Ceci nous mène à définir l'importance μ_i de manière récursive comme suit:

$$\mu_i = \sum_{j \rightarrow i} \frac{1}{l_j} \mu_j \quad (75.10)$$

Il s'agit donc d'un système d'équations linéaires homogène à n inconnues, où dans l'exemple choisi $n = 12$. Plus traditionnellement, le système pourra être noté:

$$\left(\sum_{j \rightarrow i} \frac{1}{l_j} \mu_j \right) - \mu_i = 0 \quad (75.11)$$

ou plus explicitement (cf. chapitre d'Algèbre Linéaire):

$$\begin{bmatrix}
 -1 & l_2^{-1} & l_3^{-1} & l_4^{-1} & l_5^{-1} & l_6^{-1} & l_7^{-1} & l_8^{-1} & l_9^{-1} & l_{10}^{-1} & l_{11}^{-1} & l_{12}^{-1} \\
 l_1^{-1} & -1 & l_3^{-1} & l_4^{-1} & l_5^{-1} & l_6^{-1} & l_7^{-1} & l_8^{-1} & l_9^{-1} & l_{10}^{-1} & l_{11}^{-1} & l_{12}^{-1} \\
 l_1^{-1} & l_2^{-1} & -1 & l_4^{-1} & l_5^{-1} & l_6^{-1} & l_7^{-1} & l_8^{-1} & l_9^{-1} & l_{10}^{-1} & l_{11}^{-1} & l_{12}^{-1} \\
 l_1^{-1} & l_2^{-1} & l_3^{-1} & -1 & l_5^{-1} & l_6^{-1} & l_7^{-1} & l_8^{-1} & l_9^{-1} & l_{10}^{-1} & l_{11}^{-1} & l_{12}^{-1} \\
 \vdots & \ddots & & & -1 & & & & & & & \vdots \\
 & & & & & -1 & & & & & & \\
 & & & & & & -1 & & & & & \\
 \vdots & & & & & & & -1 & & \ddots & & \vdots \\
 l_1^{-1} & l_2^{-1} & l_3^{-1} & l_4^{-1} & l_5^{-1} & l_6^{-1} & l_7^{-1} & l_8^{-1} & -1 & l_{10}^{-1} & l_{11}^{-1} & l_{12}^{-1} \\
 l_1^{-1} & l_2^{-1} & l_3^{-1} & l_4^{-1} & l_5^{-1} & l_6^{-1} & l_7^{-1} & l_8^{-1} & l_9^{-1} & -1 & l_{11}^{-1} & l_{12}^{-1} \\
 l_1^{-1} & l_2^{-1} & l_3^{-1} & l_4^{-1} & l_5^{-1} & l_6^{-1} & l_7^{-1} & l_8^{-1} & l_9^{-1} & l_{10}^{-1} & -1 & l_{12}^{-1} \\
 l_1^{-1} & l_2^{-1} & l_3^{-1} & l_4^{-1} & l_5^{-1} & l_6^{-1} & l_7^{-1} & l_8^{-1} & l_9^{-1} & l_{10}^{-1} & l_{11}^{-1} & -1
 \end{bmatrix}
 \begin{pmatrix}
 \mu_1 \\
 \mu_2 \\
 \mu_3 \\
 \mu_4 \\
 \mu_5 \\
 \mu_6 \\
 \mu_7 \\
 \mu_8 \\
 \mu_9 \\
 \mu_{10} \\
 \mu_{11} \\
 \mu_{12}
 \end{pmatrix}
 =
 \begin{pmatrix}
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0
 \end{pmatrix}
 \quad (75.12)$$

Ce qui donne avec les valeurs de notre exemple:

$$\begin{bmatrix}
 -1 & 2^{-1} & 2^{-1} & 2^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 \\
 4^{-1} & -1 & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 4^{-1} & 2^{-1} & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 4^{-1} & 0 & 2^{-1} & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 4^{-1} & 0 & 0 & 0 & -1 & 0 & 1^{-1} & 0 & 4^{-1} & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 3^{-1} & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 3^{-1} & 2^{-1} & -1 & 2^{-1} & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 3^{-1} & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2^{-1} & -1 & 2^{-1} & 2^{-1} & 2^{-1} \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & -1 & 0 & 2^{-1} \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 2^{-1} & -1 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 0 & 2^{-1} & -1
 \end{bmatrix}
 \begin{pmatrix}
 \mu_1 \\
 \mu_2 \\
 \mu_3 \\
 \mu_4 \\
 \mu_5 \\
 \mu_6 \\
 \mu_7 \\
 \mu_8 \\
 \mu_9 \\
 \mu_{10} \\
 \mu_{11} \\
 \mu_{12}
 \end{pmatrix}
 =
 \begin{pmatrix}
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0
 \end{pmatrix}
 \quad (75.13)$$

Outre la solution triviale (vecteur nul), ce système à une infinité de solutions car son déterminant est nul (cf. chapitre d'Algèbre Linéaire), ce que l'on peut facilement vérifier avec la fonction MDETERM() de la version anglaise de Microsoft Excel 11.8346.

Notons maintenant le système précédent sous la forme:

$$\begin{bmatrix}
 0 & 2^{-1} & 2^{-1} & 2^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 \\
 4^{-1} & 0 & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 4^{-1} & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 4^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 4^{-1} & 0 & 0 & 0 & 0 & 0 & 1^{-1} & 0 & 4^{-1} & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 3^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 3^{-1} & 2^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 3^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2^{-1} & 0 & 2^{-1} & 2^{-1} & 2^{-1} \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 0 & 0 & 2^{-1} \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 2^{-1} & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 0 & 2^{-1} & 0
 \end{bmatrix}
 \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ \mu_9 \\ \mu_{10} \\ \mu_{11} \\ \mu_{12} \end{pmatrix}
 =
 \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ \mu_9 \\ \mu_{10} \\ \mu_{11} \\ \mu_{12} \end{pmatrix} \quad (75.14)$$

Nous remarquons deux choses avant de continuer:

1. Dans la matrice ci-dessus, chaque colonne est telle que la somme de ses valeurs égale l'unité. Il s'agit donc de la transposée d'une matrice stochastique d'une chaîne de Markov (cf. chapitre de Probabilités).
2. L'importance μ_i est positive ou nulle. Donc le vecteur $\vec{\mu}$ est strictement positif et il s'agit du vecteur stochastique de la chaîne de Markov (cf. chapitre de Probabilités) dont la somme des composantes doit être égale à l'unité.

Nous écrirons alors ce système sous forme condensée:

$$P^T \mu = \mu \quad (75.15)$$

Ainsi, 1 est la valeur propre et μ le vecteur propre de l'application P . Nous savons de par notre étude des chaînes de Markov que cette relation n'est vérifiée que pour la mesure invariante notée traditionnellement π dans le domaine d'étude des chaînes de Markov:

$$P^T \pi = \pi \quad (75.16)$$

Pour déterminer la mesure, il faut d'abord réécrire le réseau sous le format d'application matricielle d'une chaîne de Markov (cf. chapitre de Probabilités):

$$p(n+1) = P^T p(n) \quad (75.17)$$

Avec Microsoft Excel 11.8346, la modélisation est assez simple à reproduire pour déterminer la mesure invariante avec un point de départ quelconque (distribution quelconque) dans le graphe:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Matrice Stochastique Transposée												
2		0.00	0.50	0.50	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
3		0.25	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4		0.25	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5		0.25	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6		0.25	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.25	0.00	0.00	0.00
7		0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8		0.00	0.00	0.00	0.00	0.33	0.50	0.00	0.50	0.00	0.00	0.00	0.00
9		0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.50	0.50	0.50
11		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.50
12		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.50	0.00	0.00
13		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.50	0.00
14													
15		$p(0)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$	$p(6)$	$p(7)$	$p(8)$	$p(9)$	$p(10)$	$p(11)$
16		1.00	0.00	0.38	0.23	0.23	0.23	0.21	0.20	0.20	0.19	0.18	0.18
17		0.00	0.25	0.13	0.16	0.14	0.13	0.12	0.11	0.11	0.10	0.10	0.10
18		0.00	0.25	0.13	0.16	0.14	0.13	0.12	0.11	0.11	0.10	0.10	0.10
19		0.00	0.25	0.13	0.16	0.14	0.13	0.12	0.11	0.11	0.10	0.10	0.10
20		0.00	0.25	0.00	0.18	0.15	0.12	0.18	0.15	0.16	0.17	0.16	0.17
21		0.00	0.00	0.08	0.00	0.06	0.05	0.04	0.06	0.05	0.05	0.06	0.05
22		0.00	0.00	0.08	0.08	0.06	0.11	0.09	0.10	0.11	0.10	0.11	0.11
23		0.00	0.00	0.08	0.00	0.06	0.05	0.04	0.06	0.05	0.05	0.06	0.05
24		0.00	0.00	0.00	0.04	0.00	0.05	0.03	0.04	0.05	0.05	0.06	0.06
25		0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.03
26		0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.03
27		0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.03

Figure: 75.3 - Détermination de la mesure invariante avec Microsoft Excel 11.8346

avec les formules suivantes:

	A	B	C	D
1		Matrice Stochastique Transposée		
2		0	0.5	0.5
3		0.25	0	0
4		0.25	0.5	0
5		0.25	0	0.5
6		0.25	0	0
7		0	0	0
8		0	0	0
9		0	0	0
10		0	0	0
11		0	0	0
12		0	0	0
13		0	0	0
14				
15		$p(0)$	$p(1)$	$p(2)$
16		1	=MMULT(\$B\$2:\$M\$13,B16:B27)	=MMULT(\$B\$2:\$M\$13,C16:C27)
17		0	=MMULT(\$B\$2:\$M\$13,B16:B27)	=MMULT(\$B\$2:\$M\$13,C16:C27)
18		0	=MMULT(\$B\$2:\$M\$13,B16:B27)	=MMULT(\$B\$2:\$M\$13,C16:C27)
19		0	=MMULT(\$B\$2:\$M\$13,B16:B27)	=MMULT(\$B\$2:\$M\$13,C16:C27)
20		0	=MMULT(\$B\$2:\$M\$13,B16:B27)	=MMULT(\$B\$2:\$M\$13,C16:C27)
21		0	=MMULT(\$B\$2:\$M\$13,B16:B27)	=MMULT(\$B\$2:\$M\$13,C16:C27)
22		0	=MMULT(\$B\$2:\$M\$13,B16:B27)	=MMULT(\$B\$2:\$M\$13,C16:C27)
23		0	=MMULT(\$B\$2:\$M\$13,B16:B27)	=MMULT(\$B\$2:\$M\$13,C16:C27)
24		0	=MMULT(\$B\$2:\$M\$13,B16:B27)	=MMULT(\$B\$2:\$M\$13,C16:C27)
25		0	=MMULT(\$B\$2:\$M\$13,B16:B27)	=MMULT(\$B\$2:\$M\$13,C16:C27)
26		0	=MMULT(\$B\$2:\$M\$13,B16:B27)	=MMULT(\$B\$2:\$M\$13,C16:C27)
27		0	=MMULT(\$B\$2:\$M\$13,B16:B27)	=MMULT(\$B\$2:\$M\$13,C16:C27)

Figure: 75.4 - Formules Microsoft Excel 11.8346 (version anglaise) correspondantes au tableau précédent

Le système converge assez vite, après la 30ème étape, nous avons déjà une convergence à la deuxième décimale:

	AI	AJ	AK
15	$p(33)$	$p(34)$	$p(35)$
16	0.12	0.12	0.12
17	0.06	0.06	0.06
18	0.06	0.06	0.06
19	0.06	0.06	0.06
20	0.18	0.18	0.18
21	0.06	0.06	0.06
22	0.12	0.12	0.12
23	0.06	0.06	0.06
24	0.11	0.11	0.11
25	0.05	0.05	0.06
26	0.05	0.05	0.06
27	0.05	0.05	0.06

Figure: 75.5 - Convergence à la 35ème itération

Donc finalement la mesure invariante est:

$$\pi = \mu = \begin{pmatrix} 0.12 \\ 0.06 \\ 0.06 \\ 0.06 \\ 0.18 \\ 0.06 \\ 0.12 \\ 0.06 \\ 0.11 \\ 0.06 \\ 0.06 \\ 0.06 \end{pmatrix} \quad (75.18)$$

quelle que soit la distribution du vecteur de départ. Propriété que l'on appelle "ergodique" dans le domaine des chaînes de Markov.

Étant donné que, nous avons pour tout $\lambda \in \mathbb{R}^+$ le système suivant qui est toujours vérifié:

$$P^T (\lambda \cdot \mu) = (\lambda \cdot \mu) \quad (75.19)$$

Alors, il est d'usage dans l'étude de l'algorithme de Google d'écrire le vecteur stochastique avec des valeurs entières (chaque valeur du vecteur stochastique a été divisée par la plus petite probabilité qui était 0.06) correspondant au fameux "**PageRank Value**" (au début de l'existence de Google car actuellement ce calcul est beaucoup plus subtil) tel que:

$$\pi = \mu = \begin{pmatrix} 2 \\ 1 \\ 1 \\ 1 \\ 3 \\ 1 \\ 2 \\ 1 \\ 2 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad (75.20)$$

Cela nous donne donc la répartition convergente par page d'une cohorte de 17 internautes (sommes des valeurs du vecteur stochastique).

Les pages où μ_i est grand sont les plus fréquentées (les plus populaires) en termes de probabilités à l'équilibre. Dans la quête de classer les pages web, c'est encore un argument pour utiliser la mesure μ comme indicateur empirique.

1.3. ÉTATS ABSORBANTS

Cependant, si notre graphe (chaîne) contient une page (ou un groupe de pages) sans issue alors celle-ci absorbe toute la probabilité, car notre cohorte tombera tôt ou tard sur cette page et ne pourra plus en sortir (état absorbant). Nous représentons cela sous la forme graphique suivante si, par exemple, nous ajoutons une page P_{13} qui est absorbante:

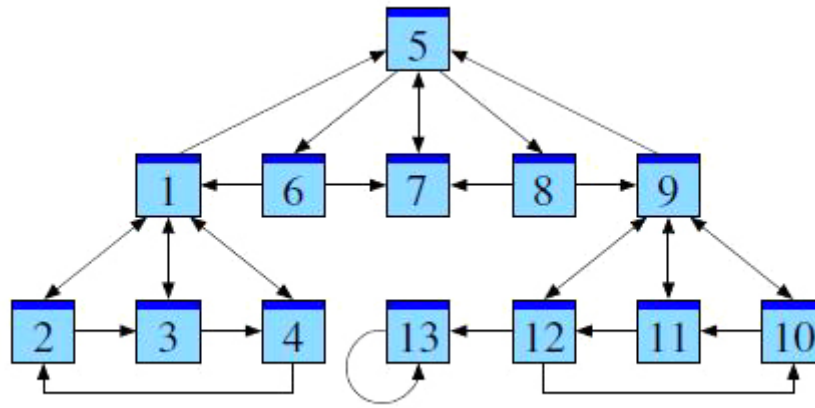


Figure: 75.6 - Graphe avec un état absorbant

avec la matrice de transition associée:

$$\begin{bmatrix}
 0 & 2^{-1} & 2^{-1} & 2^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 4^{-1} & 0 & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 4^{-1} & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 4^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 4^{-1} & 0 & 0 & 0 & 0 & 0 & 1^{-1} & 0 & 4^{-1} & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 3^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 3^{-1} & 2^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 3^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2^{-1} & 0 & 2^{-1} & 2^{-1} & 3^{-1} & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 0 & 0 & 3^{-1} & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 2^{-1} & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 0 & 2^{-1} & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3^{-1} & 1
 \end{bmatrix} \quad (75.21)$$

Notre modèle n'est donc pas encore satisfaisant. Nous pourrions construire alors une matrice qui permet d'échapper aux états absorbants en posant que toute colonne qui ne contient qu'un seul 1 avec que des zéros se voit téléporter le visiteur sur toutes les pages avec un poids égal. Ce qui donnerait puisque nous avons 13 pages:

$$\begin{bmatrix}
 0 & 2^{-1} & 2^{-1} & 2^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 13^{-1} \\
 4^{-1} & 0 & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 13^{-1} \\
 4^{-1} & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 13^{-1} \\
 4^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 13^{-1} \\
 4^{-1} & 0 & 0 & 0 & 0 & 0 & 1^{-1} & 0 & 4^{-1} & 0 & 0 & 0 & 13^{-1} \\
 0 & 0 & 0 & 0 & 3^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 13^{-1} \\
 0 & 0 & 0 & 0 & 3^{-1} & 2^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 13^{-1} \\
 0 & 0 & 0 & 0 & 3^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 13^{-1} \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2^{-1} & 0 & 2^{-1} & 2^{-1} & 3^{-1} & 13^{-1} \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 0 & 0 & 3^{-1} & 13^{-1} \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 2^{-1} & 0 & 0 & 13^{-1} \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 0 & 2^{-1} & 0 & 13^{-1} \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3^{-1} & 13^{-1}
 \end{bmatrix} \quad (75.22)$$

Mais ceci n'est pas très optimal, car cela charge la matrice donc c'est très gourmand en termes de mémoires et calculs.

Pour échapper aux états absorbants, Google utilise un modèle empirique plus raffiné.

Nous prenons l'application initiale qui est une application linéaire:

$$f: \mathbb{R}^n \rightarrow \mathbb{R}^n \quad \mu \rightarrow P^T \mu \quad (75.23)$$

où pour rappel la somme des composantes de μ vaut toujours 1. Nous réécrivons cela sous la forme:

$$f: \mathbb{R}^n \rightarrow \mathbb{R}^n \quad \mu \rightarrow 1 \cdot P^T \mu \quad (75.24)$$

où le 1 représente le fait qu'il y a 100% de probabilité d'avoir l'état $P^T \mu$. Maintenant, l'astuce consiste à écrire:

$$f: \mathbb{R}^n \rightarrow \mathbb{R}^n \quad \mu \rightarrow \varepsilon \cdot c + (1-c) \cdot P^T \mu \quad (75.25)$$

avec:

$$\varepsilon = \begin{pmatrix} 1 \\ \frac{1}{n} \\ \vdots \\ \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{pmatrix} \quad (75.26)$$

ce qui signifie que l'internaute a à tout moment une probabilité $(1-c)$ d'être dans l'état $P^T \mu$ et une probabilité c d'être dans l'état ε , c'est-à-dire de se trouver avec une probabilité équiprobable en n'importe quel point du graphe. Enfin, nous sommes les deux pour obtenir à nouveau un vecteur stochastique (dont la somme des composantes vaut 1).

Effectivement, si nous prenons par exemple 10% pour la valeur de c , nous avons la somme des composantes du vecteur:

$$(1-c) \cdot P^T \mu \quad (75.27)$$

qui vaudra 0.9 et respectivement la somme des composantes du vecteur $\varepsilon \cdot c$ qui vaudra 0.1 (puisque la somme des composantes de ε vaut aussi 1 par construction).

Donc pour résumer, avec une probabilité fixée c le moteur d'indexation (ou l'internaute) abandonne la page actuelle P_j et recommence sur une des n pages du web, celle-ci étant choisie, afin de ne privilégier personne. Sinon, avec une probabilité $1-c$ le moteur suit un des liens de la page P_j , choisi de manière équiprobable. Le seul point délicat reste à calibrer la valeur de c qui vaut donc entre 0 et 1. S'il vaut 0 nous retombons sur le modèle initial avec le problème des états absorbants, s'il vaut 1 nous avons un modèle qui donne une note égale à chaque page. Il faut donc plutôt choisir c avec des valeurs faibles proches de 0.

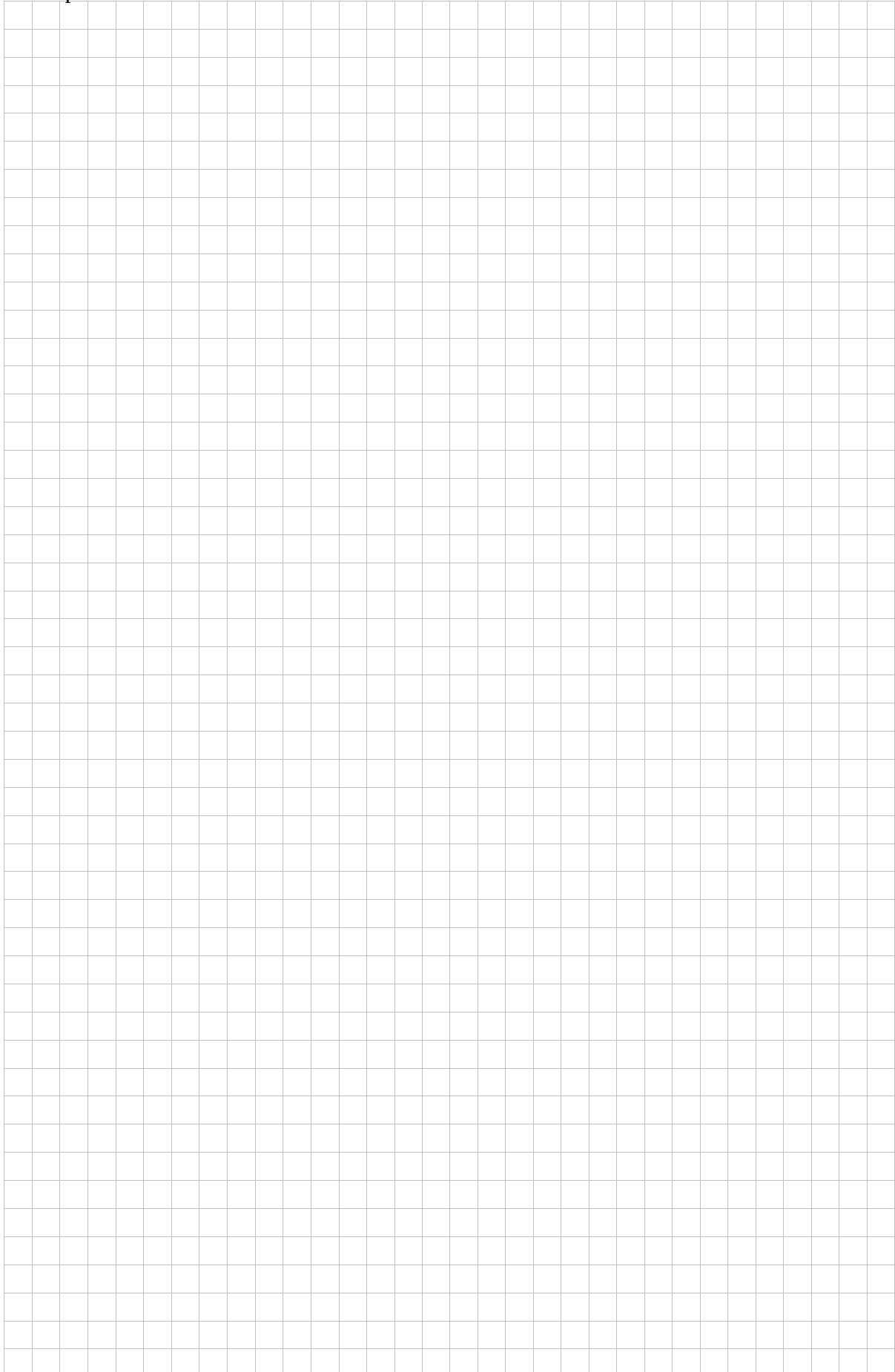
Cette astuce de téléportation évite de se faire piéger par une page absorbante et garantit d'arriver n'importe où dans le graphe, indépendamment des questions de connectivité.

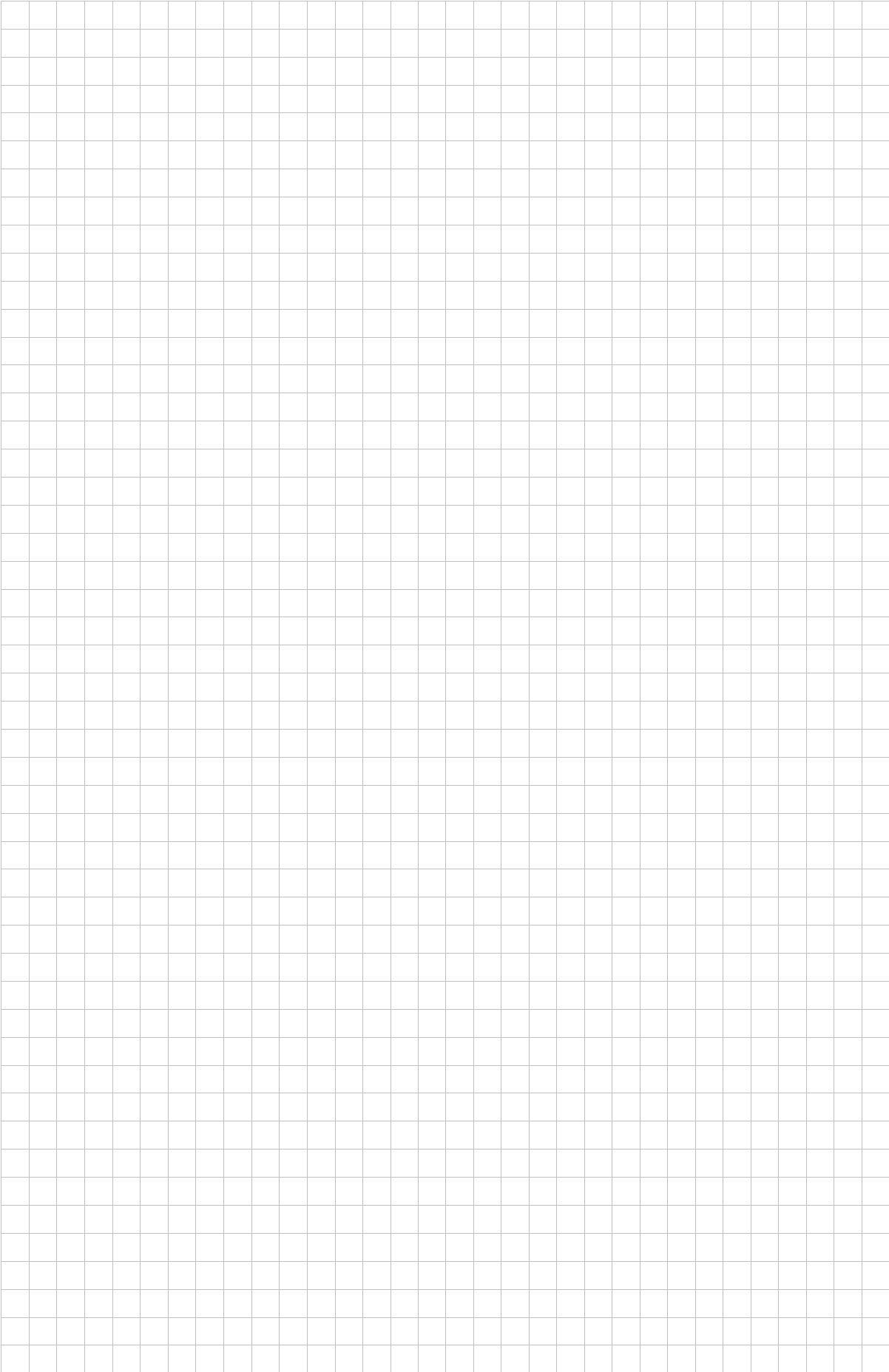
Il convient de remettre le tout sous forme de composantes, nous avons donc l'écriture (dont le deuxième terme converge):

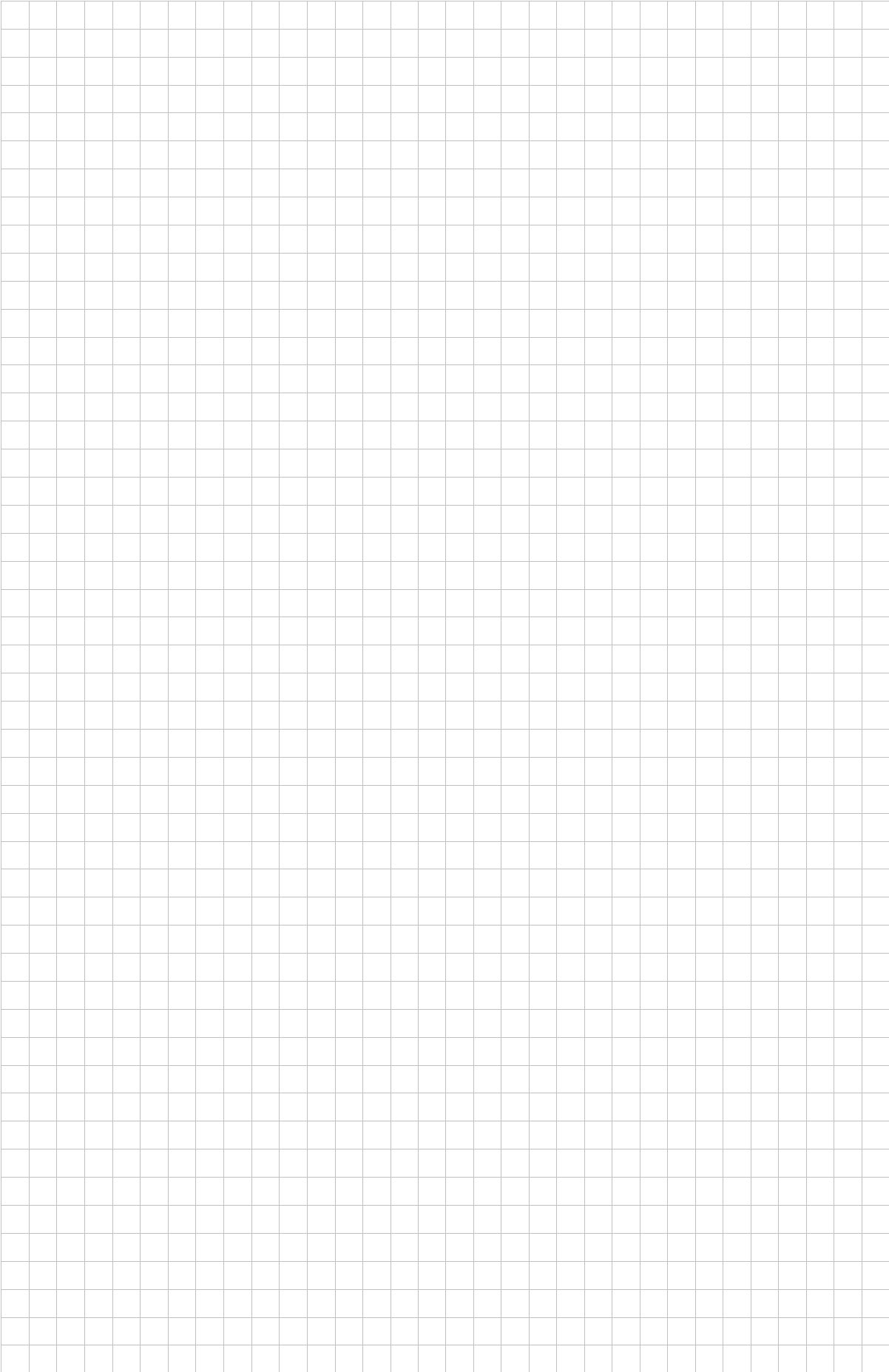
$$\mu_i = \frac{1}{n}c + \sum_{j \rightarrow i} \frac{1-c}{l_j} \mu_j \quad (75.28)$$

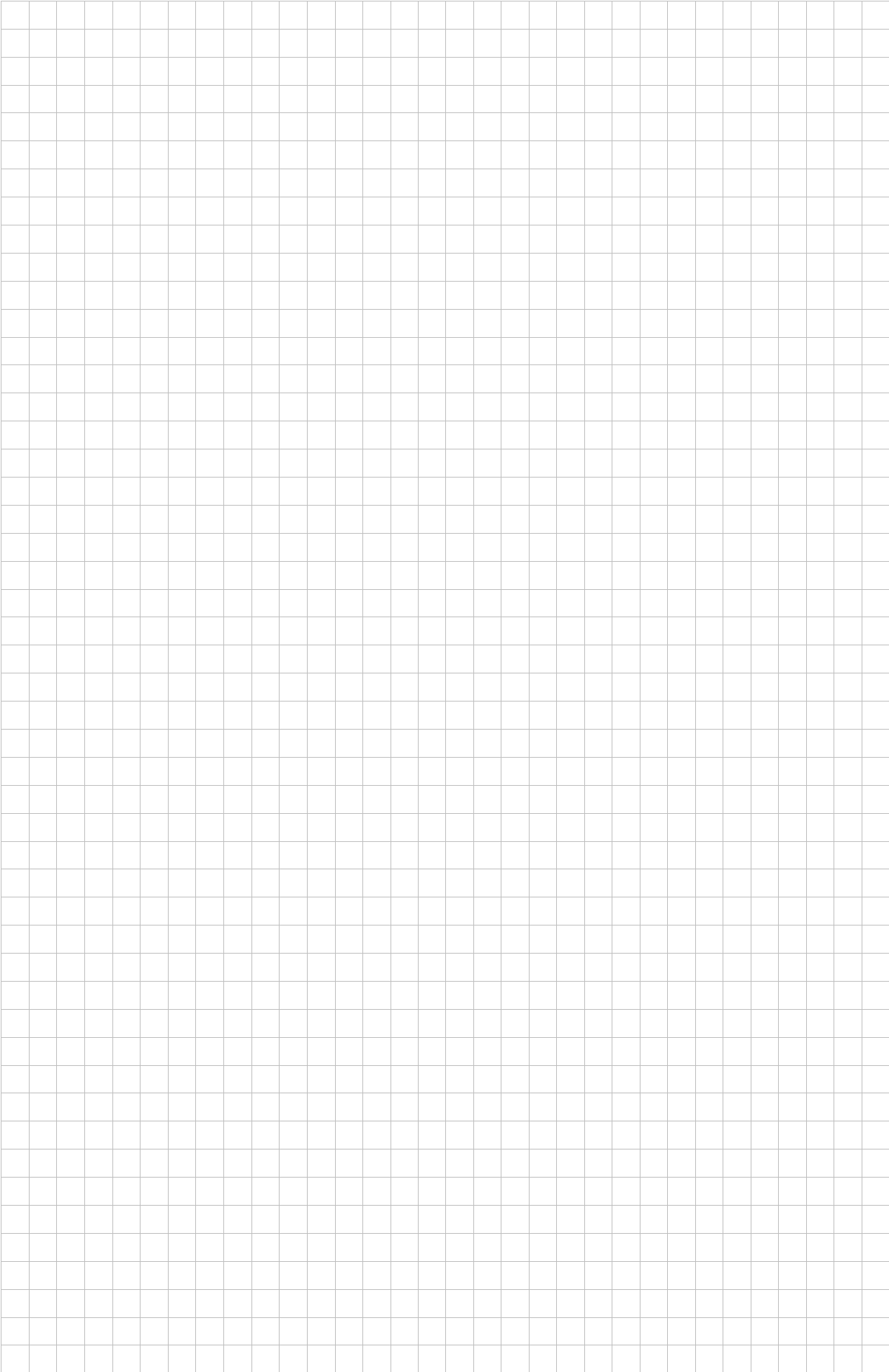
qui a l'avantage énorme de pouvoir calculer la valeur de n'importe quelle composante μ_i du vecteur stochastique de n'importe quelle étape d'une promenade aléatoire sur le graphe et ce sans faire appel à aucun moment à la manipulation d'une immense matrice qui serait trop gourmande en termes de capacités de mémoire!!

Notes personnelles:









BIOGRAPHIES

Cette chapitre présente une poignée d'hommes qui postulent une étrange renommée. Selon les règles de l'histoire que l'on enseigne à l'école primaire, ils n'existent pas, ils n'ont commandé aucune armée, ils n'ont envoyé personne à la mort, ils n'ont dirigé aucun empire et ils n'ont eu qu'une part minimale dans les grandes décisions historiques. Certains ont acquis quelque célébrité, mais aucun ne fut jamais un héros national. Pourtant, leur oeuvre a davantage influencé le cours de l'histoire que bien des actes accomplis par des hommes d'État auréolés d'une gloire très supérieure. Elle a produit plus de bouleversements que le va-et-vient des armées en bataille par-dessus les frontières, elle a fait plus pour le bonheur ou le malheur que les édits des rois et des assemblées. Car leur oeuvre, c'est d'avoir façonné l'esprit de l'homme.

Qui propage ses idées, manie un pouvoir bien supérieur à celui de l'épée ou du sceptre: aussi ont-ils façonné et dirigé le monde. Pour la plupart, ils n'ont pas levé le moindre petit doigt pour agir physiquement; ils ont travaillé essentiellement en intellectuels, dans le silence et l'oubli, sans se soucier outre mesure du monde environnant. Mais, dans leur sillage, des empires se sont écroulés et des continents disloqués, des régimes politiques se sont soit renforcés, soit érodés, les classes se sont dressées les unes contre les autres, ainsi que les nations. Non pas sous l'effet d'un noir complot, mais de par la puissance extraordinaire de leurs idées. Qui sont ces hommes (femmes) ?: des savants, économistes, chimistes, biologistes, mathématiciens, physiciens, informaticiens, ingénieurs,...

Les biographies ci-dessous des scientifiques les plus connus à travers le monde et cités dans les différents chapitre du site sont triées par ordre alphabétique et presque tous les textes sont de simples copier/coller simplifiés de [Wikipédia](#). Si vous souhaitez que nous rajoutions une entrée, il vous suffit de nous envoyer par [courriel](#) le nom et prénom de la personne concernée et la raison pour laquelle vous aimeriez la voir figurer dans la liste ci-dessous. Nous étudierons ensuite la proposition et prendrons la décision qui s'impose.

Nous rendons aussi hommage aux centaines de milliers de scientifiques (savants), ingénieurs, philosophes, artisans, mécènes, artistes, amateurs éclairés connus ou anonymes dont la collaboration a permis à travers les millénaires l'évolution de la science et de la condition humaine!

Les tailles des biographies ne sont pas proportionnelles au nombre d'articles publiés ou découvertes effectuées, mais à la quantité d'informations trouvées sur ces personnages sur l'Internet ou dans la littérature. La liste n'est aussi pas exhaustive, mais son objectif est de rendre hommage et de se remémorer les grands hommes qui ont fait des sciences pures et exactes ce qu'elles sont aujourd'hui et qui ont consacré une partie ou l'entier de leur vie à la science: l'art le plus contraint.

Attention! En physique (aussi bien qu'en mathématique) une théorie, une équation, une constante ou autre porte rarement le nom de son vrai inventeur. Ce fait est largement connu chez les scientifiques et est souvent source de moqueries de la communauté.



Al-Biruni, Muhammad Ibn Ahmad Abul-Rayhan (973-1048) est un mathématicien, un astronome, un physicien, un érudit, un encyclopédiste, un philosophe, un astrologue, un voyageur, un historien, un pharmacologue et un précepteur, originaire de la Perse, qui contribua grandement aux domaines des mathématiques, philosophie, médecine et des sciences. Il est connu pour sa théorie sur la rotation de la Terre autour de son axe et autour du Soleil, et ceci bien avant Copernic. Il s'attacha notamment à calculer la marche du Soleil (apogée), corrigea certaines données de Ptolémée. Excellent mathématicien, Al-Biruni développa de nouvelles équations inconnues de ses prédécesseurs. Il calcula également le méridien local et les coordonnées des localités. Mais le tableau ne serait pas complet si l'on oubliait de mentionner que six siècles avant Galilée, Al Biruni mettait déjà en avant une Terre qui tournait autour de son axe. Avec l'aide d'un astrolabe, de la mer et d'une montagne avoisinante, il calcula la circonférence de la Terre en résolvant une équation complexe pour son époque. Le principal d'apport d'Al-Biruni aux mathématiques résiderait dans ses travaux en trigonométrie (calculs des valeurs des fonctions trigonométriques qui n'étaient pas encore définies en tant que tel à l'époque).



Alembert, Jean le Rond (1717-1783), enfant naturel d'un commissaire d'artillerie, abandonné sur les marches de la chapelle parisienne de Saint-Jean-Le-Rond, le futur grand philosophe, mathématicien et physicien est recueilli par un vitrier qui recevra secrètement une pension pour subvenir à l'éducation du jeune garçon qui étudiera brillamment le droit, la médecine et la mathématique. Suite à la publication de divers mémoires (sur le calcul intégral, sur la réfraction des corps solides), d'Alembert entre à l'Académie des sciences (1741). On lui doit le célèbre principe de la quantité de mouvement, dit "principe de d'Alembert" dans son *Traité de dynamique* (1743). En astronomie, il est l'auteur (1749) d'un traité sur la précession des équinoxes qu'il explique au moyen de la théorie de la gravitation universelle de Newton et d'une solution partielle au problème des trois corps. D'Alembert établit aussi une théorie mathématique des cordes vibrantes en étudiant la nature composite du son (harmoniques).



Ampère, André-Marie (1775-1836) à 18 ans, il connaît la majeure partie des oeuvres mathématiques de son temps. Mathématicien de premier ordre, il montre comment l'on doit utiliser cette science, qu'il considérait comme une branche de la philosophie, à l'étude des découvertes des faits physiques pour en donner une relation définitive. En quelques semaines, Ampère établit les bases de toute une science à laquelle il donne le nom d'électromagnétisme. Il cherche à comprendre le magnétisme des aimants et en tire une hypothèse de "courants particuliers" (orbites électroniques et orientation du spin aujourd'hui). Il établit également l'égalité du nombre de molécules dans des volumes égaux de Gaz de natures différentes, mais mesurés dans des conditions identiques de température et de pression (observation expérimentale de Gay-Lussac).



Archimède, De Syracuse (287-212 av. J.-C.), mathématicien et ingénieur grec célèbre à la fois comme mécanicien théoricien et comme constructeur de machines. Archimède de Syracuse eut une production mathématique exceptionnelle, dont une partie nous est parvenue dans des traités comme *Sur la sphère et le cylindre*; *la Mesure du cercle*; *la Quadrature de la parabole*; *Des spirales*; *Des conoïdes et sphéroïdes*; *la Méthode*, *Des corps flottants*... C'est à partir de ses travaux mécaniques que les principales anecdotes le mettant en scène, comme celle du levier ou du bain, vont se constituer. La célèbre maxime: «Donnez-moi une place où me tenir et je mettrai la terre en mouvement» est un écho populaire de la contribution archimédienne à la statique, exposée dans le traité des *Équilibres*. Archimède démontre la loi du levier, introduit la notion fondamentale de centre de gravité, et détermine ces barycentres pour les principales figures géométriques planes. Il en est de même pour l'anecdote d'Archimède, jaillissant nu de son bain, en criant «Eurêka», parce qu'il venait, dit-on, de trouver le moyen de résoudre le problème que lui avait posé le roi Hiéron. En fait, le récit est une mise en scène spectaculaire de la découverte du principe fondamental de l'hydrostatique (communément appelé depuis "principe d'Archimède"). En géométrie, l'oeuvre d'Archimède développe celle d'Eudoxe de Cnide telle que nous la connaissons par le livre XII des *Éléments* d'Euclide: il s'agit de comparer les mesures des figures planes et solides, en particulier des figures curvilignes. Ainsi, Archimède démontre que le volume du cylindre circonscrit à une sphère est égal à une fois et demi le volume de celle-ci et que la surface latérale du cylindre est égale à celle de la sphère ou quatre fois la surface d'un grand cercle. Donc, si l'on sait calculer la surface du cercle, on connaîtra celle de la sphère, du cylindre, son volume et celui de la sphère, etc. Son résultat le plus célèbre et le plus simple concerne le cercle. Archimède ramène sa quadrature à un autre problème: la rectification de sa circonférence, c'est-à-dire «trouver une ligne droite égale qui lui soit égale», problème qu'il résout à l'aide d'une courbe géométrique que l'on appelle désormais "spirale d'Archimède". En outre, il calcule des valeurs approchées du rapport circonférence/diamètre (ce que nous appelons le nombre "Pi").



Avogadro, Amedeo (1776-1856), Fils d'un magistrat de Turin, Amadeo Avogadro commence par suivre la voie paternelle. Il passe une licence de droit en 1795 et s'inscrit au barreau de sa ville natale. Mais son goût pour la physique et la mathématique, auxquelles il n'a cessé de s'intéresser en solitaire, le pousse à entamer sur le tard des études scientifiques. En 1809, il fait une communication à l'Académie royale de Turin ; le succès qu'il remporte grâce à elle lui permet d'obtenir un poste de professeur au Collège royal de Verceil. En 1820, l'Université de Turin crée pour lui une chaire de physique qu'il gardera jusqu'à la fin de sa vie. C'est en étudiant les lois régissant la compression et la dilatation des gaz qu'Avogadro énonce, en 1811 l'hypothèse restée célèbre sous le nom de "loi d'Avogadro". Reposant sur la théorie atomique de Dalton et la loi de Gay-Lussac sur les rapports volumiques, la théorie d'Avogadro indique que deux volumes égaux de gaz différents, dans les mêmes conditions de température et de pression, contiennent le même nombre de molécules. Sous son apparente simplicité, cette loi comporte des implications importantes ; grâce à elle, il devient possible de déterminer la masse molaire d'un gaz à partir de celle d'un autre. Mais les

chimistes de l'époque, plus intéressés par les expériences, boudent quelque peu les études théoriques d'Avogadro qui ne seront d'ailleurs reconnues que 50 ans plus tard. Le nom d'Avogadro reste également lié à celui du "nombre d'Avogadro" indiquant le nombre de molécules contenues dans une seule mole.



Bachelier, Louis (1870-1946) est né au Havre dans une famille de négociants. Il apparaît à sa majorité sur les listes électorales du Havre en 1892 comme représentant de commerce à la même adresse professionnelle que son père. Après avoir effectué son service militaire, à l'âge de 22 ans, il reprend ses études à la faculté des sciences de Paris. Elles sont couronnées par une licence ès sciences en 1895 (mention passable) et par la soutenance en 1900 de sa non moins fameuse et méconnue thèse de doctorat en mathématiques. Bien que cette thèse soit considérée aujourd'hui comme un travail précurseur en théorie des probabilités et en théorie financière, elle ne vaut à l'époque à son auteur qu'une mention honorable. De 1913 à 1914 Bachelier dispensa un cours libre de théorie des probabilités appliquées à la mécanique, la balistique et la biométrie. Il fut également chargé de conférences supplémentaires sur la mathématique générale de 1913 à 1914. Ce n'est qu'après la guerre de 1914-1918 qu'il obtient un premier poste de chargé de cours à la faculté des sciences de Besançon. Après divers remplacements à Dijon puis à Rennes, il revient à Besançon en 1927 comme professeur titulaire de la chaire de calcul différentiel et intégral, poste qu'il occupe jusqu'à sa retraite en 1937. Louis Bachelier a, parmi ses nombreux travaux, été le premier à avoir introduit la continuité dans les problèmes de probabilité en prenant le temps comme variable. En particulier, il a élaboré une théorie mathématique du mouvement brownien 5 ans avant Albert Einstein. Il est également bien avant Norbert Wiener, le premier à avoir défini la fonction du mouvement brownien et donné un grand nombre de ses propriétés.



Banach, Stefan (1892-1945) mathématicien polonais qui a posé les bases de l'analyse fonctionnelle. Né à Cracovie en 1892, en Autriche-Hongrie (actuellement ville polonaise). Banach fit ses études secondaires à Cracovie; il se révéla particulièrement brillant en mathématiques et en sciences naturelles, mais son désintérêt pour les autres matières l'empêcha d'obtenir les meilleures mentions. La vie (au moins mathématique) de Banach va basculer au printemps 1916, quand il rencontre Steinhaus à Cracovie. Avec Otto Nikodym, ils décident de fonder une société mathématique. La recherche mathématique de Banach commence là. Son premier article est cosigné avec Steinhaus. Steinhaus lui avait parlé d'une propriété qu'il ne parvenait pas à démontrer, et après quelques jours de réflexion, Banach exhiba un contre-exemple. Il est difficile de dire ce qu'il serait advenu de l'activité mathématique de Banach sans la rencontre avec Steinhaus, mais toujours est-il qu'il entama à compter de celle-ci une recherche intense et fructueuse. Banach retourne à Lvov en 1920 où un poste d'assistant lui est proposé. Il soutient sa thèse en 1922, et c'est dans cette thèse qu'apparaît pour la première fois la notion d'espace de Banach, qu'y sont démontrés les théorèmes fondamentaux sur ces objets, où on y évoque la topologie faible... Bref, cette thèse marque la naissance de l'analyse fonctionnelle. En 1929, il fonde avec Steinhaus la revue mathématique *Studia Math*, consacrée au développement de l'analyse fonctionnelle, et en 1939 il est élu président de la société mathématique

de Pologne. En 1945, peu avant la fin de la Seconde Guerre Mondiale, il décède d'un long cancer. De nombreux théorèmes sont associés au nom de Banach, qu'il les ait démontrés lui-même, ou qu'ils fassent référence à ces idées. Citons entre autres: le théorème de Hahn-Banach de prolongement des formes linéaires continues, le théorème de Banach-Steinhaus, de Banach-Alaoglu, le théorème du point fixe de Banach, ainsi que le paradoxe de Banach-Tarski.



Bell, John (1928-1990) fut dès la plus petite enfance attiré par les livres traitant des sciences. À cause de problèmes financiers familiaux, il ne put poursuivre immédiatement des études académiques. Il travailla donc pendant une année en tant que technicien au département de physique de l'université de Queen's à Belfast avant de devenir étudiant en 1945 dans ce même département. Il sortit premier de sa promotion en mathématiques-physique. Bell trouva dans les années 1960 une nouvelle inspiration dans les bases de la théorie quantique, une matière supposée épuisée par les résultats de la discussion de Bohr-Einstein 30 ans plus tôt, et ignorée par pratiquement tous ceux qui ont employé la théorie quantique entre-temps. Effectivement, Bell était intrigué par les incertitudes quantiques de Heisenberg et voulait creuser le sujet en montrant que la discussion de tels concepts comme le "réalisme", le "déterminisme" et la "localité" pouvaient être affiliés dans un rapport mathématique rigoureux: "les inégalités de Bell" vérifiables expérimentalement. Bell poussa très loin les doutes qu'il avait sur les principes d'incertitudes au point qu'il en irrita même son professeur (Sloane) qui lui fit remarquer que maintenant il allait un peu trop loin! Bell attendit son travail de thèse pour développer ses idées. Malheureusement, à nouveau à cause de problèmes financiers, il dut repousser ses recherches à plus tard et joindre le centre anglais de recherche atomique à Harwell. Pendant sa carrière, il épousa une femme (Mary Bell) qui l'aida dans le développement de ses travaux sur les principes fondamentaux de la théorie quantique. C'est, en 1951, avec Rudolf Peierls que Bell développa sa célèbre théorie C.P.T. (*Charge, Parity, Time*). Malheureusement, pour Bell, Gerhard Lüders et Wolfgang Pauli arrivèrent au même résultat dans la même période et c'est à eux que furent attribués les crédits de la découverte. Les développements théoriques de Bell sont à l'origine de la cryptographie et de la théorie de l'information quantique. L'attention à la théorie quantique de l'information a énormément augmenté au cours des dernières années, et le sujet semble sûr d'être l'un des secteurs scientifiques dont la croissance sera la plus importante au 21^{ème} siècle. Un autre travail de première importance de Bell en 1969 fut la participation au développement de "l'anomalie A.B.J." (Adler-Bell-Jackiw) dans la théorie quantique des champs. Ces trois physiciens montrèrent que le modèle algébrique standard contenait une erreur. Effectivement, la quantification du modèle des champs brise une symétrie. Bell fut nommé pour le prix Nobel, qu'il aurait certainement obtenu s'il n'était pas décédé d'une attaque cérébrale en 1990.



Bernoulli, Daniel (1700-1782), est un savant suisse qui découvrit les principes de base du comportement d'un fluide (c'est le fils de Jean Bernoulli et le neveu de Jacques Bernoulli.). Il cultiva à la fois les sciences mathématiques et les sciences naturelles, enseigna les mathématiques, l'anatomie, la botanique et la physique. Ami de Leonhard Euler, il travaille avec lui dans plusieurs domaines des mathématiques et de la physique (il partagea avec lui dix fois le prix annuel de l'Académie des sciences de Paris), qu'il s'en fit une sorte de revenu. Les différents problèmes qu'il tente de résoudre (théorie de l'élasticité, mécanisme des marées) le conduisent à s'intéresser et développer des outils mathématiques tels que les équations différentielles ou les séries. Il collabore également avec Jean le Rond d'Alembert dans l'étude des cordes vibrantes. Il étudia l'écoulement des fluides (1738) et formula le principe (le fameux théorème de Bernoulli) selon lequel la pression exercée par un fluide est inversement proportionnelle à sa vitesse d'écoulement. Il utilisa des concepts atomistes pour ébaucher la première théorie cinétique des gaz, en exprimant leur comportement en termes de probabilités sous des conditions particulières de pression et de température. On peut le considérer comme l'un des fondateurs de l'hydrodynamique.



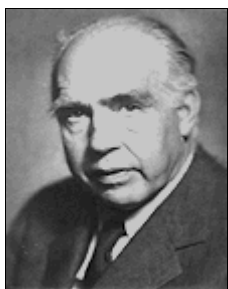
Bernoulli, Jacques (1654-1705), est un mathématicien et physicien suisse, frère de Jean Bernoulli et oncle de Daniel Bernoulli et Nicolas Bernoulli. Né à Bâle en 1654, il rencontre Robert Boyle et Robert Hooke lors d'un voyage en Angleterre en 1676. Après cela, il se consacre à la physique et aux mathématiques. Il enseigne à l'université de Bâle à partir de 1682, devenant professeur de mathématiques en 1687. Il mérita par ses travaux et ses découvertes d'être nommé associé de l'Académie des sciences de Paris (1699) et de celle de Berlin (1701). Sa correspondance avec Gottfried Wilhelm Leibniz le conduisit à étudier le calcul infinitésimal en collaboration avec son frère Jean. Il fut un des premiers à comprendre et à appliquer le calcul différentiel et intégral, proposé par Leibniz, découvrit les propriétés des nombres dits depuis "nombres de Bernoulli" et donna la solution de problèmes regardés jusque-là comme insolubles. Il pose les principes du calcul des probabilités et introduit les nombres de Bernoulli dans un ouvrage publié après sa mort en 1713.



Bernoulli, Jean (1667-1748), est un mathématicien et physicien suisse. Frère de Jacques Bernoulli et père de Daniel et Nicolas Bernoulli, il professa les mathématiques à Groningue (1695), puis à Bâle, après la mort de Jacques Bernoulli (1705), et devint associé des Académies de Paris, de Londres, de Berlin et de Saint-Pétersbourg. Formé par son frère Jacques Bernoulli, il avait longtemps travaillé de concert avec lui à développer les conséquences du nouveau calcul infinitésimal inventé par Gottfried Leibniz ; mais il s'établit ensuite entre eux, à l'occasion de la résolution de quelques problèmes, une rivalité qui dégénéra en inimitié. Il a aussi contribué dans beaucoup de secteurs aux mathématiques y compris le problème d'une particule se déplaçant dans un champ de gravité. Il trouva l'équation de la chaînette en 1690 et développa le calcul exponentiel en 1691. Il eut aussi la gloire de former Leonhard Euler. Il vint à Paris en 1690, et se lia avec les savants les plus distingués, particulièrement avec de L'Hôpital. Jean Bernoulli est devenu membre de la Royal Society en 1712.



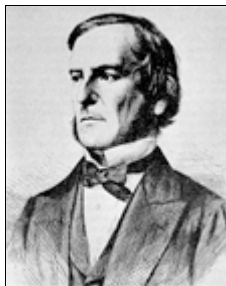
Bessel, Friedrich (1784-1846) Né à Minden en Westphalie, Bessel commença à travailler très jeune comme commis. Attiré par la navigation maritime, il s'intéressa aux observations nautiques, construisant lui-même son sextant et étudiant l'astronomie à ses heures de liberté. Il calcula la trajectoire de la comète de Halley, résultat qui fut immédiatement publié et lui permit d'obtenir, en 1806, un emploi d'assistant à l'observatoire de Lilienthal. En 1810, il devint directeur du nouvel observatoire de Königsberg, tout en poursuivant des études mathématiques. Il dut enseigner la mathématique à ses étudiants en astronomie jusqu'en 1825 (date à laquelle Jacobi vint enseigner cette matière à Königsberg). Toute sa vie fut consacrée à l'astronomie (il écrivit plus de 350 articles) et, peu avant sa mort, il commença l'étude du mouvement d'Uranus, problème qui devait aboutir à la découverte de Neptune. En mathématiques, Bessel est connu pour avoir introduit les fonctions qui portent son nom, les utilisant pour la première fois, en 1817, lors de l'étude d'un problème de Kepler, et les employant plus complètement 7 ans plus tard pour étudier les perturbations planétaires.



Bohr, Niels Henrik David (1885-1962), physicien danois, prix Nobel en 1922, pour sa contribution à la physique nucléaire et à la compréhension de la structure atomique. La théorie de Bohr sur la structure atomique, pour laquelle il reçut le prix Nobel de physique en 1922, fut publiée entre 1913 et 1915. Son travail s'inspira du modèle nucléaire de l'atome de Rutherford, dans lequel l'atome est considéré comme un noyau compact entouré d'un essaim d'électrons. Le modèle pose en principe que l'atome n'émet de rayonnement électromagnétique que lorsqu'un électron se déplace d'un niveau quantique à un autre. Ce modèle contribua énormément aux développements ultérieurs de la physique atomique théorique.



Boltzmann, Ludwig (1844-1906), physicien autrichien qui contribua à établir les bases de la mécanique statistique. Ayant fait ses études à Vienne et à Oxford, il enseigna la physique dans différentes universités allemandes et autrichiennes pendant plus de 40 ans. Développant la théorie cinétique des gaz, notamment à partir des travaux de Maxwell, il établit que la seconde loi de la thermodynamique pouvait être obtenue sur la base de l'analyse statistique. Calculant le nombre de particules dotées d'une énergie donnée, il établit la statistique dite de Maxwell-Boltzmann. Il exprima l'entropie S d'un système en fonction de la probabilité W de son état (par l'intermédiaire de sa fameuse équation de transport à partir de laquelle démontra que l'entropie ne pouvait qu'augmenter au cours du temps... résultat qui était jusque là admis expérimentalement mais sans preuve théorique). Il put aussi établir de manière théorique la loi de Stefan relative au rayonnement d'un corps noir. Mais il lui fallut expliquer comment les principes mécaniques, où les phénomènes sont réversibles, pouvaient engendrer des lois thermodynamiques décrivant des phénomènes marqués par l'irréversibilité. Il avança l'idée que les évolutions irréversibles, quoiqu'elles ne soient que des possibilités parmi d'autres, sont si probables que ce sont pratiquement toujours elles qui se produisent.



Boole, Georges (1815-1864) Mathématicien et logicien anglais, Boole est le créateur de la logique symbolique. Né à Lincoln et fils d'un petit commerçant, il reçut ses premières leçons de mathématiques de son père, qui lui apprit aussi à fabriquer des instruments d'optique. En dehors des conseils de son père et de quelques années passées dans les écoles locales, Boole est un autodidacte. Quand les affaires de son père déclinèrent, il fut obligé de travailler pour aider sa famille et, dès 16 ans, il enseigna dans des écoles de village ; à 20 ans, il ouvrit sa propre école à Lincoln. Pendant ses loisirs, il étudiait la mathématique à l'Institut de mécanique, créé vers cette époque ; c'est là qu'il se familiarisa avec les *Principia* de Newton, la Mécanique céleste de Laplace et la Mécanique analytique de Lagrange et qu'il commença à résoudre des problèmes d'algèbre supérieure. Boole soumit au nouveau *Cambridge Mathematical Journal* une série d'articles originaux dont le premier est *Recherches sur la théorie des transformations analytiques* ; ces articles portaient sur les équations différentielles et sur les invariants par transformation linéaire. En 1844, il étudie les liens entre l'algèbre et le calcul infinitésimal dans un important mémoire publié dans les Transactions de la Royal Society, qui lui décerne une médaille cette même année pour sa contribution à l'analyse (c'est-à-dire l'utilisation de l'algèbre dans l'étude des infiniment petits et grands). Développant de nouvelles idées sur la méthode en logique et confiant dans le symbolisme qu'il avait élaboré à partir de ses recherches mathématiques, il publie, en 1847, un opuscule, *Mathematical Analysis of Logic*, dans lequel il soutient que la logique doit être rattachée aux mathématiques et non à la philosophie. Bien qu'il n'eût aucun titre universitaire, Boole fut, sur la base de ses publications, nommé en 1849 professeur au Queen's College à Cork, en Irlande. Avec Boole, en 1847 et en 1854, commence l'algèbre de la logique, c'est-à-dire ce que l'on appelle de nos jours l'algèbre de Boole. Dans son ouvrage de 1854, Boole énonce complètement sa nouvelle méthode symbolique d'inférence logique, qui permet, étant donnés des propositions contenant un certain nombre de termes, d'en tirer, par traitement symbolique des prémisses, des conclusions qui étaient logiquement contenues dans les prémisses. Il rechercha aussi une méthode générale en calcul des probabilités, qui aurait permis, à partir des probabilités connues d'un système d'événements donnés, de déterminer la probabilité de tout autre événement relié logiquement aux événements donnés.



Borel, Émile (1871-1956), reçu major à l'X et à ULM, il choisit cette dernière et se consacre aux mathématiques. Il fonda l'institut Henri Poincaré et fut élu député de l'Aveyron et maire de Saint-Affrique. Il étudie les mesures d'ensembles et notamment, définit les ensembles de mesure nulle et l'ensemble des boréliens, sur lequel on peut définir une mesure. Il se tourne ensuite vers les probabilités et la physique-mathématique. Borel est également considéré comme un mathématicien constructiviste. Il fut à l'origine de la théorie des jeux stratégiques et de la cybernétique que développeront von Neumann et Morgenstern. Son élève Henri Lebesgue utilisera ses résultats en topologie et théorie de la mesure pour sa théorie de l'intégration.



Born, Max (1882-1970) né à Breslau est décédé à Göttingen est un physicien allemand, puis britannique. Initialement il suit ses études au collège de König-Wilhelm, et poursuit à l'Université de Breslau suivie par celle d'Heidelberg et de Zürich. Pendant les études pour son doctorat il entra en contact avec des mathématiciens comme Klein, Hilbert, Minkowski, Runge, Schwarzschild. En 1921, il est nommé professeur de physique théorique à Göttingen. Il émigre en Écosse en 1933 et devient citoyen britannique en 1939. Physicien théoricien remarquable, il est principalement connu pour son importante contribution à la physique quantique: développement (1925) de la mécanique quantique matricielle introduite par Werner Heisenberg et, surtout, il sera le premier à donner au carré du module de la fonction d'onde la signification d'une densité de probabilité de présence. Il fut également un pionnier dans la théorie quantique des solides (conditions de Born-Von Karmann) et dans l'électrodynamique non linéaire de Born-Infeld. Il est lauréat de la moitié du prix Nobel de physique de 1954 (l'autre moitié a été remise à Walther Bothe) pour ses recherches fondamentales en mécanique quantique, particulièrement pour son interprétation statistique de la fonction d'onde. La Royal Society lui décerne la Médaille Hughes en 1950.



Bose, Satyendranath (1894-1974) Mathématicien et physicien indien, connu pour ses contributions à la théorie quantique. Né à Calcutta, Bose a fait ses études au Presidency College de Calcutta. En 1924, il propose une description statistique des systèmes quantiques, reprise par Albert Einstein, et qui n'impose aucune restriction sur la distribution en énergie des particules du système. Cette description est connue sous le nom de "statistique de Bose-Einstein", par opposition à la "statistique de Fermi-Dirac". Appliquée à la théorie du rayonnement du corps noir, cette nouvelle statistique conduit à la distribution de Planck et permet de traiter ce rayonnement comme un gaz de photons. Dans le domaine de la physique des particules élémentaires, la statistique de Bose-Einstein impose à la fonction d'onde des particules (dans l'équation de Schrödinger) d'être parfaitement symétrique pour l'ensemble des variables d'espace et de spin. Les particules obéissant à cette statistique (photons, mésons π , etc.) sont appelées des bosons. Professeur de physique aux universités de Calcutta et de Dacca, Satyendranath Bose a été nommé, en 1958, professeur national des Indes.



de Broglie, Louis Victor (1892-1987), physicien français et lauréat du prix Nobel, qui apporta une contribution essentielle à la théorie quantique avec ses études de la radiation électromagnétique. Né à Dieppe, Louis de Broglie fit ses études à Paris. Il essaya de cerner la nature dualiste de la matière et de l'énergie et proposa l'association d'une onde à toute particule. Il proposa ainsi directement comment il était possible d'obtenir les règles de quantification du modèle d'atome de Bohr et Sommerfeld en exigeant qu'un nombre entier d'ondes soit logé sur une orbite stationnaire. Sa découverte de la nature ondulatoire des électrons (1924) lui valut le prix Nobel de physique en 1929 sans qu'il proposa toutefois une équation d'onde décrivant les phénomènes quantiques (ce que fera Schrödinger). Il fut élu à l'Académie des sciences en 1933 et à l'Académie française en 1943. Il fut nommé professeur de physique théorique à l'université de Paris (1928), secrétaire

perpétuel de l'Académie des sciences (1942), et conseiller au Commissariat à l'énergie atomique (1945).



Brouwer, Luitzen Egbertus Jan (1881-1966) fut un grand mathématicien néerlandais du début du 20^e siècle. Né d'un père proviseur, il réalisa des études secondaires très brillantes, et très rapides. À l'université d'Amsterdam, il fut formé par Korteweg, qui est connu pour des contributions en mathématiques appliquées. Il soutient son doctorat en 1904. De 1909 à 1913, Brouwer s'intéresse à la topologie, et découvre la majeure partie des théorèmes auxquels son nom est resté attaché, dont son fameux théorème du point fixe. Pour beaucoup, Brouwer est le père de la topologie moderne. En 1912, il obtient grâce aux recommandations de Hilbert, une chaire à l'Université d'Amsterdam. Il y enseigne la théorie des ensembles, celle des fonctions, et l'axiomatique. Plus tard, il refusera de rejoindre Hilbert à Göttingen. Pendant la première Guerre mondiale sa santé se fragilisa et il s'éloigna quelques temps des champs de la recherche scientifique. Quand il y revint, ce fut pour se consacrer à ses premières amours (sa thèse portait déjà sur ce sujet): les fondements des mathématiques. Brouwer est le fer de lance avec Poincaré des mathématiques intuitionnistes, par opposition au logicisme de Russell et Frege, et au formalisme de Hilbert. En particulier, pour Brouwer, un théorème d'existence ne peut être vrai que si on peut exhiber un processus, même formel, de construction. Cela le conduit notamment à rejeter la loi du tiers-exclu, qui dit qu'une propriété est ou vraie, ou fausse! Les preuves ainsi obtenues sont souvent plus longues, mais Brouwer fut capable de réécrire des traités de théorie des ensembles, de théorie de la mesure, et de théorie des fonctions en se conformant aux règles de l'intuitionnisme. Bizarrement, Brouwer n'enseigna jamais la topologie. C'est probablement dû au fait que les théorèmes que lui-même avait prouvés ne rentraient plus dans le cadre qu'il s'était fixé. Selon les témoignages de quelques-uns de ses étudiants, il était un personnage vraiment étrange, fou amoureux de sa philosophie, et un professeur auquel il ne fallait surtout pas poser de questions!



Cantor, Georg (1845-1918) se révèle être un étudiant brillant, notamment dans les matières manuelles. Malgré les injonctions de son père, qui rêve d'en faire un ingénieur, il part en 1862 à Berlin étudier la mathématique, où ses maîtres sont Weierstrass et Kronecker. Il soutient son doctorat en 1867 (sur la théorie des nombres). Les premières recherches postdoctorales de Cantor sont consacrées à la décomposition des fonctions en sommes de séries trigonométriques (les célèbres séries de Fourier) et particulièrement à l'unicité de cette décomposition. Afin de résoudre complètement ce difficile problème, il est amené à introduire et à étudier des ensembles dits "ensembles exceptionnels". Cela le conduit à définir en 1872 très précisément ce qu'est un nombre réel, comme limite d'une suite de nombres rationnels; parallèlement, son ami Dedekind donne la même année une autre définition de la droite des réels, à partir des coupures. Cantor et Dedekind constatent à cette occasion qu'il y a beaucoup plus de réels que de rationnels, mais il n'y a pas jusque-là de définition mathématique à ce "beaucoup plus". En 1874, dans le prestigieux *Journal de Crelle*, Cantor donne une définition du nombre d'éléments d'un ensemble infini qui prolonge naturellement celle du

cardinal d'un ensemble infini, qui prolonge celle du cardinal d'un ensemble fini. Il en découle, jusqu'en 1897, une succession de découvertes étranges: il y a autant d'entiers pairs que d'entiers tout court, autant de points sur un segment que dans un carré, beaucoup plus de nombres transcendants que de nombres rationnels. Cette hiérarchie dans les ensembles infinis conduit progressivement Cantor à définir des nouveaux nombres, les ordinaux transfinis, et à définir une arithmétique sur ces nombres. Les travaux de Cantor ont eu beaucoup d'influence au 20^{ème} siècle. On citera d'abord, en 1903, un paradoxe soulevé par Russell dans la théorie naïve des ensembles: si A est l'ensemble de tous les ensembles qui ne sont pas éléments d'eux-mêmes, A est-il contenu dans A ? Les logiciens surmonteront cette difficulté conceptuelle, sans rien changer des conclusions de Cantor. Citons aussi le problème de l'hypothèse du continu. Un des derniers axes de recherche de Cantor était d'estimer le nombre d'éléments de la droite réelle. Plus précisément, Cantor souhaitait prouver l'absence de tout ensemble dont le cardinal soit strictement compris entre le cardinal des entiers et celui des réels. C'est ce que l'on appelle "l'hypothèse du continu". Tous les travaux de Cantor et de ses successeurs pour confirmer ou infirmer l'hypothèse du continu furent vains, et pour cause: en 1963, le logicien américain Cohen prouva que, dans une théorie standard des ensembles, l'hypothèse du continu est indécidable. On peut très bien supposer qu'elle est vraie ou qu'elle est fausse sans obtenir de contradiction dans la théorie.



Carnot, Nicolas Léonard Sadi (1796-1832), physicien et ingénieur militaire français, considéré comme le créateur de la science thermodynamique. Fils aîné de Lazare Carnot, surnommé "le Grand Carnot", Sadi fit ses études à l'École polytechnique. En 1824, il décrivit sa conception du moteur à chaleur idéal, appelé "moteur Carnot", dans lequel toute l'énergie disponible est utilisée. Il découvrit que la chaleur ne pouvait passer d'un corps froid à un corps plus chaud, et que le rendement d'un moteur dépendait de la quantité de chaleur qu'il était capable d'utiliser. Cette découverte, ou cycle de Carnot, est à la base de la seconde loi de la thermodynamique.

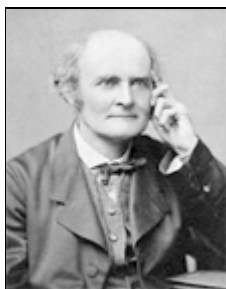


Cartan, Élie (1869-1951) fit ses études primaires à l'École de Dolomieu, puis au collège de Vienne et au lycée de Grenoble. Il suivit au lycée Jeanson-de-Sailly la préparation à l'École Normale Supérieure, où il entra en 1888. Il y suivit notamment les enseignements de Poincaré, Picard et de Hermite. Les premiers travaux d'Élie Cartan qui devaient déboucher sur sa thèse soutenue en 1894, portent sur les groupes de Lie simples complexes, où il reprend, corrige et développe les résultats de structure et de classification obtenus par W. Killing. Élie Cartan obtient un poste de lecteur à l'Université de Montpellier de 1894 à 1896, puis à la Faculté des sciences de Lyon de 1896 à 1903. La même année, il est nommé professeur à la Faculté des sciences de Nancy, où il restera jusqu'en 1909. Il donne en même temps des cours à l'École d'Électrotechnique et de Mécanique Appliquée. Il rédige deux grands articles sur une généralisation en dimension infinie des groupes de Lie simples. Il élabore la méthode du "repère mobile", et la théorie des formes extérieures qui devaient influencer le développement ultérieur de la géométrie différentielle. En 1909, il quitte Nancy pour venir enseigner à la Sorbonne, où il est nommé professeur en 1912. Il assure

par ailleurs un enseignement à l'École de Physique et Chimie de Paris. En 1914, il résout le problème de la classification des groupes de Lie simples réels, et détermine les représentations de dimension finie de ces groupes. Pendant la guerre, il sert comme sergent dans l'hôpital aménagé dans les locaux de l'École Normale Supérieure, tout en continuant ses travaux mathématiques. Son oeuvre mathématique ultérieure est considérable, avec près de 200 publications et de nombreux ouvrages. Parmi les thèmes abordés, mentionnons l'étude des systèmes de Pfaff, la théorie de la déformation, l'étude des variétés à courbure constante négative, la théorie de la gravitation d'Einstein, la théorie des connexions affines, les groupes d'holonomie, les espaces riemanniens symétriques, les spineurs. Il est aussi l'auteur de plusieurs articles sur l'histoire de la géométrie. Il prit sa retraite en 1940.



Cauchy, Augustin-Louis (1789-1857). C'est à Cherbourg que Cauchy commence ses recherches mathématiques sur les polyèdres, et ses premiers résultats sont prometteurs. Mais, fatigué par le cumul de la charge d'ingénieur et des longues veillées de recherche, Cauchy connaît un état dépressif qui s'éternise et le pousse à retourner vivre chez ses parents. À Paris, il cherche une situation en adéquation avec sa volonté de faire de la recherche en mathématique pure. En 1815, il achève un brillant mémoire où il démontre un célèbre théorème de Fermat sur les nombres polygonaux. Ceci fera beaucoup pour sa notoriété, et en 1816, il accède à l'Académie des Sciences, en remplacement de Carnot et Monge touchés par l'épuration. Le cours d'analyse que Cauchy professe à l'École Polytechnique est décrié tant par ses élèves que par ses collègues des autres matières. Pourtant, c'est ce cours publié en 1821 et 1823, qui devait devenir la référence de l'analyse au 19ème siècle. en mettant en avant la rigueur, et plus seulement l'intuition. C'est la première fois que de vraies définitions de limites, de continuité, de convergence de suites, de séries, sont énoncées. Cette rigueur reste toutefois encore relative, puisque Cauchy "prouve" que la limite d'une série de fonctions continues est continue, ce qui est faux. Il est vrai que Cauchy ne dispose pas encore d'une définition claire et précise des nombres réels. C'est l'époque aussi où Cauchy réalise des travaux profonds sur les fonctions d'une variable complexe (établissant par exemple l'expression des résidus), ainsi que des avancées dans la théorie des groupes finis. Cauchy ne fut jamais le chef d'une école de mathématiciens, et il se comporta parfois maladroitement avec de jeunes chercheurs comme Abel ou Galois, dont il sous-estime, ou même perd, des mémoires de première importance. Ses relations avec ses collègues ne sont en général pas très faciles.



Cayley, Arthur (1821-1895), né à Richmond (Surrey), manifesta très tôt de vives dispositions pour la mathématique. Cependant, malgré le grand intérêt de ses premières publications, il ne put s'imposer comme mathématicien ; il décida de faire des études de droit et devint avocat en 1849. Pendant 14 ans, il exerça ce métier tout en s'adonnant à des recherches scientifiques. En 1863, Cayley est nommé professeur à Cambridge et peut enfin se consacrer entièrement aux mathématiques. Dans l'ensemble de l'oeuvre de Cayley, notamment dans ses travaux de jeunesse, est sensible l'influence des fondateurs de l'école algébrique anglaise qui avaient formulé le programme de l'algèbre moderne en accordant une

priorité marquée à l'approche formelle des problèmes. Mathématicien lettré et créateur, Cayley, dans le sillage de l'école anglaise, sut élaborer de nouvelles et fructueuses théories. La richesse de l'approche de Cayley apparaît dès ses premiers travaux sur la théorie des groupes (1854). Cayley, abordant les travaux de Galois, Gauss et Cauchy avec les méthodes des algébristes anglais, donne une définition des groupes abstraits ce qui le conduisit à la notion d'isomorphisme. L'étude des systèmes d'équations linéaires conduisit Cayley à celle des déterminants. Dans ses premiers travaux, il établit de nombreuses règles de calcul sur les déterminants, y compris la relation de multiplication des déterminants qui figurait déjà dans les travaux de Cauchy, Binet et Jacobi. À côté d'études originales sur les déterminants, on y rencontre la notion de tableau rectangulaire représentant les coefficients d'un système d'équations linéaires ou les coefficients d'une transformation linéaire. Cayley étudie les matrices rectangulaires à coefficients réels ou complexes ; il introduit les opérations sur les matrices et décrit leurs propriétés, y compris le caractère non commutatif de la multiplication. Il s'agit là sans doute de la première apparition de l'algèbre linéaire. Quelques années plus tard, Cayley étudiera aussi les systèmes non associatifs et publiera des résultats d'algèbre multilinéaire. Cayley a consacré un grand nombre de ses publications aux problèmes de la géométrie et à l'étude des courbes et des surfaces algébriques. À 22 ans, il émettait l'idée de la géométrie à n dimensions, idée qui fut formulée aussi, presque simultanément, mais sous une forme un peu différente, par Grassman. Cayley ne revint que beaucoup plus tard (en 1870) sur l'espace à n dimensions, mais sa méthode algébrique contribua aux importantes découvertes qui eurent lieu dans les autres domaines de la géométrie. C'est ainsi que, dans le *Sixth Memoir on Quantics* de 1859, il introduit la métrique projective, subordonnant ainsi la géométrie métrique à la géométrie projective ; il démontre alors que les notions fondamentales de la géométrie métrique (angles et distances) sont les invariants et les covariants de certaines transformations linéaires de la quadrique absolue.



Chandrasekhar, Subrahmanyan (1910-1995) obtint à l'âge de 23 son doctorat au Trinity College de l'université de Cambridge. Spécialiste en astrophysique Chandrasekhar fit progresser de façon décisive la connaissance de l'évolution hydrodynamique et hydromagnétique des transferts d'énergie par rayonnement sans oublier les effets quantiques et relativistes dans les évolutions des étoiles. Sa contribution majeure dans ce domaine est la transformation des étoiles en naines blanches et au-delà d'un astre d'une masse supérieure à la limite de Chandrasekhar (1.44 celle du soleil), l'effondrement de l'astre en une étoile à neutrons. Les objets plus massifs donnant eux des trous noirs.



Clairaut, Alexis-Claude (1713-1765) était un membre de l'Académie française des sciences et fut l'un des mathématiciens et physiciens les plus renommés du 18ème siècle. À l'âge de 10 ans, il connaissait le calcul infinitésimal, à 12 ans, il soumettait sa première étude à l'académie des sciences et à 18 ans, il publia un livre contenant des extensions importantes à la géométrie qui lui ont valu l'admission à l'académie en 1731. Clairaut fut l'un des scientifiques qui accompagnaient Maupertuis en Laponie pour acquérir les données nécessaires pour la détermination de la forme de la Terre. En 1743, il publia sa *Théorie de la figure*

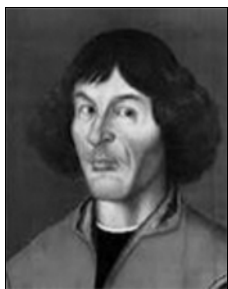
de la Terre, qui calculait plus précisément que l'avait fait Newton, la forme qu'adopte un corps en rotation due à la gravitation naturelle de ses parties. En 1760, il publia sa *Théorie du mouvement des comètes*, qui prédit avec précision la date à laquelle la comète de Halley sera arrivée au point le plus proche du soleil.



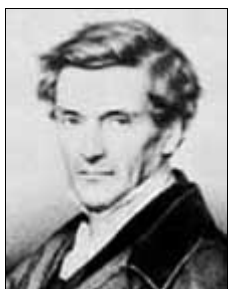
Cohen, Paul Joseph (1934-2007) est un mathématicien et logicien américain. En 1963, Cohen a découvert une nouvelle construction de modèles, appelée "forcing", qui joue désormais un rôle fondamental dans la théorie des ensembles et dans la théorie des modèles. Il a aussi construit des modèles de la théorie des ensembles (supposée consistante) dans lesquels l'axiome du choix et l'hypothèse du continu ne sont pas vérifiés, ce qui, compte tenu de l'oeuvre antérieure de Kurt Gödel, établit que l'axiome du choix et l'hypothèse du continu sont indépendants des systèmes usuels de la théorie des ensembles. Ce travail a valu à Cohen, en 1966, la médaille Fields de l'Union Mathématique Internationale. Il est également l'auteur de travaux intéressants en analyse classique.



Connes, Alain (1947-) est né en 1947 à Draguignan. Ancien élève à l'École normale supérieure, il a reçu, en 1980, le prix Ampère, l'un des plus importants décernés par l'Académie des sciences. Il a été élu membre de cette académie, dont il a été le benjamin, en 1981. Les premiers travaux d'Alain Connes s'inscrivent directement dans la tradition de John von Neumann et de ses continuateurs immédiats. Le développement de la physique quantique vers les années vingt avait mis à l'ordre du jour l'étude d'espaces non plus à trois dimensions, comme celui où nous croyons vivre, ni à quatre, comme en relativité einsteinienne, mais à une infinité de dimensions (les espaces de Hilbert). L'un des outils essentiels de la physique quantique est la notion d'opérateur dans un tel espace, notion généralisant celle de rotation d'un espace euclidien. La théorie des algèbres d'opérateurs a débuté vers 1930 par les travaux de von Neumann, qui a montré l'importance d'un certain type d'algèbres d'opérateurs, appelées aujourd'hui "algèbres de von Neumann", et qui a établi pour ces algèbres un théorème de décomposition en facteurs premiers assez analogue au théorème de décomposition bien connu pour les nombres entiers usuels. Dès l'origine, les facteurs avaient été classés en trois types: facteurs de type I, II, III. On a eu assez tôt une bonne compréhension des facteurs de type I et pas mal d'informations sur ceux de type II, mais les facteurs de type III sont restés pendant longtemps beaucoup plus mystérieux. Même les exemples étaient rares et von Neumann disait, à propos de ce cas: "C'est le plus réfractaire de tous, et les outils pour l'étudier nous font défaut, au moins pour l'instant". La première réussite de Connes, qui lui a d'emblée valu la renommée internationale, a été une percée spectaculaire vers l'élucidation de la structure des facteurs de type III ; on peut dire qu'il est le premier à avoir acquis une connaissance concrète de ces objets, jusque-là assez énigmatiques, pris dans leur ensemble. Très grosso modo, les résultats de Connes ramènent l'étude des facteurs de type III à celle des facteurs de type II et de leurs automorphismes. L'oeuvre d'Alain Connes est celle d'un mathématicien très complet, capable de résoudre des problèmes difficiles, légués par le passé, mais aussi de transformer entièrement une discipline par l'introduction d'idées nouvelles, d'une grande originalité.



Copernic, Nicolas (1473-1543), Étudiant à l'université de Cracovie à partir de 1491, il se rend ensuite en Italie pour y suivre des cours de droit canon à l'université de Bologne. Il suit également les cours d'astronomie de Domenico Maria Novara, un des premiers scientifiques à remettre en cause les enseignements de Ptolémée. En 1500, il enseigne la mathématique à Rome, avant de retourner pour un an à Frauenburg où son oncle l'a nommé chanoine en 1497. Ayant obtenu l'autorisation de poursuivre ses études en Italie, il s'inscrit aux facultés de droit et de médecine de Padoue et obtient son doctorat en droit canon à Ferrare en 1503. Enfin, il retourne à Frauenburg où il fait construire un observatoire et entame ses recherches en astronomie. Il y demeurera jusqu'à sa mort. La cosmologie de l'époque est alors basée sur le système géocentrique de Ptolémée. La Terre se trouve immobile au centre de plusieurs sphères concentriques qui portent la Lune, Mercure, Vénus, le Soleil, Mars, Jupiter, Saturne et enfin les étoiles. Mais ce système ne convient pas à Copernic, qu'il trouve compliqué et bancal. Il consulte alors les auteurs de l'Antiquité (Cicéron, Aristarque de Samos, etc.) et constate que certains d'entre eux envisagent la rotation des planètes, dont la Terre, autour du Soleil, considéré comme fixe. Copernic démontre alors que la combinaison des mouvements de la Terre et des planètes explique parfaitement le mouvement apparent des planètes (dans le sens direct et rétrograde). De plus, il établit que leurs changements de diamètre apparent apparaissent comme une conséquence de leur révolution autour du Soleil. Ses recherches se poursuivront pendant 36 ans et il démontrera que la Lune est un satellite de la Terre et que l'axe de la Terre n'est pas fixe. Son oeuvre maîtresse *De Revolutionibus orbium coelestium* est publiée en 1543 à Nuremberg et Copernic n'en reçoit les premiers exemplaires que quelques heures avant sa mort. Dans la dédicace qu'il fait au Pape Paul III, il présente son système comme une pure hypothèse, évitant ainsi la vindicte de l'Église. Adopté un siècle après sa mort après avoir été violemment rejeté, le système copernicien apporta une profonde révolution dans la conception du monde et plus généralement dans la pensée scientifique.



Coriolis, Gaspard (1792-1843), ingénieur et mathématicien français qui mit en évidence les forces centrifuges composées, dites "forces de Coriolis". Cet ingénieur des Ponts et Chaussées est l'auteur d'importants travaux en mécanique. En 1835, il démontra que l'accélération d'un mobile dans un référentiel en rotation est soumis à une complémentaire (force de Coriolis) perpendiculaire au sens de déplacement du mobile dans ce référentiel. Bien que de faible intensité à la surface de la Terre, cette force, produite par la rotation de la planète, influence la direction des courants marins et aériens. Elle produit une déviation vers l'est et explique, par exemple, le mouvement circulaire des ouragans.



Coulomb, Charles Augustin de (1736-1806), physicien français, pionnier de la théorie de l'électricité. Né à Angoulême, il servit comme ingénieur militaire pour la France aux Antilles, mais se retira à Blois à la révolution française, pour continuer ses recherches sur le magnétisme, le frottement et l'électricité. En 1777, il inventa la balance de torsion qui permet de mesurer la force de l'attraction magnétique et électrique. Grâce à cette invention, Coulomb fut capable de formuler le principe, maintenant connu sous le nom de loi de Coulomb, qui gouverne l'interaction entre les charges électriques. En 1779, Coulomb publia le traité *Théorie des machines simples*, une analyse du frottement dans les machines. Après la révolution, Coulomb quitta sa retraite et aida le nouveau gouvernement à concevoir un système métrique pour les poids et mesures. L'unité utilisée pour exprimer la quantité de charge électrique, le "Coulomb", tient son nom du physicien



Cournot, Antoine Augustin (1801-1877) étudia au collège de Gray de 1809 à 1816. Il obtint des prix d'excellence de mathématiques. Il entre en 1820 au collège Royal de Besançon et obtient le prix d'honneur de mathématiques spéciales. Avec deux mémoires et deux traductions de traités divers de mathématiques, il se fait remarquer par Poisson, qui le fait nommer en 1834 professeur d'analyse et de mécanique à la faculté des sciences de Lyon. Augustin Cournot est un savant, c'est-à-dire un homme de savoir étendu à tous les domaines de la science, un savant philosophe mais, qui par sa modestie, n'a pas connu la célébrité. Cournot fut d'abord un professeur et un vulgarisateur d'une grande clarté. Trois ouvrages mathématiques le distinguent: *Traité élémentaire de la théorie des fonctions et du calcul infinitésimal* (1841); *Exposition de la théorie des chances et des probabilités* (1843) ; *De l'origine et des limites de la correspondance entre l'algèbre et la géométrie* (1847). Mais le génie de Cournot se situe dans l'introduction des probabilités en économie. Il est le précurseur des théories modernes en économie, reprises ensuite par Léon Walras qui dans sa notice autobiographique achevée en 1904, ainsi que dans plusieurs lettres, a rappelé le rôle primordial qu'ont joué dans le développement de sa pensée, d'une part, l'oeuvre d'Antoine Augustin Cournot et d'autre part, celle de son père, l'économiste et philosophe Auguste Walras qui fut le condisciple d'Augustin Cournot à l'École Normale.



Clausius, Rudolf (1822-1888) est l'un des plus grands physiciens du 19ème siècle. Il est connu principalement pour sa contribution à l'étude de la thermodynamique. Le premier, ce savant allemand formula ce que l'on a coutume d'appeler le "deuxième principe" et proposa une définition claire de l'entropie. Il est aussi l'un des principaux créateurs de la théorie cinétique des gaz. Né à Köslin, en Poméranie, Clausius fréquenta les universités de Berlin, puis de Halle dont il sortit diplômé en 1848. Professeur jusqu'à sa mort, il fut titulaire de la chaire de physique de l'École royale d'artillerie et du génie à Berlin (1850-1855), puis, simultanément, à l'université et à l'École polytechnique de Zürich (1855-1867), ensuite à l'université de Würzburg (1867-1869), enfin à celle de Bonn, de 1869 à sa mort. Sa première publication, en 1850 dans les *Annalen der Physik* de Poggendorff, attira largement l'attention. Il cherchait à y concilier l'idée de l'équivalence entre le travail et la chaleur. Clausius fit remarquer que l'hypothèse

de la conservation de la chaleur dans le processus de transfert n'était pas une partie essentielle de la théorie de Carnot. Il établit en fait que, dans une machine idéale, la quantité de chaleur prise à la chaudière doit toujours être supérieure à celle qui est cédée au condenseur, et ce d'une quantité exactement équivalente au travail fourni. Cette importante synthèse effectuée, Clausius, dans la même publication, énonça ce que nous appelons aujourd'hui le deuxième principe de la thermodynamique. C'était la généralisation de la nécessité, déjà établie par Carnot, de la présence, non seulement d'un corps chaud (la chaudière), mais aussi d'un corps froid (le condenseur) pour qu'un travail soit fourni par une machine à vapeur. En 1854, Clausius, poussant plus avant les vues exprimées dès 1850, proposa le premier énoncé clair du concept de l'entropie. Il cherchait à mesurer l'aptitude de l'énergie calorifique de n'importe quel système réel non idéal à fournir du travail. Dans le cas de la conduction thermique le long d'un barreau solide, par exemple, la chaleur passe de l'extrémité chaude à l'extrémité froide sans fournir aucun travail, bien que ce transfert s'accompagne d'une diminution de l'aptitude de l'extrémité chaude à servir par la suite de source potentielle de travail. Cette diminution survient parce qu'à la fin du processus l'énergie calorifique est détenue par un corps situé à une température inférieure à celle de l'état initial. Elle n'a donc pas été perdue, mais seulement dégradée puisque, d'après le deuxième principe de la thermodynamique, on ne peut retrouver la température initiale qu'avec l'aide d'un travail extérieur. Les dernières contributions majeures de Clausius à la science datent de 1857 et 1858 et sont relatives à la théorie cinétique des gaz. Bien qu'il ne soit pas le premier à avoir conçu cette dernière, déjà proposée et discutée par Joule et Krönig notamment, il prend rang avec Maxwell parmi ses fondateurs. Il introduisit le concept du libre parcours moyen et établit l'importante distinction entre l'énergie de translation et l'énergie interne d'une particule de gaz. De plus, on lui reconnaît généralement le mérite d'avoir, par ses travaux théoriques, jeté un pont entre la théorie atomique et la thermodynamique.



Curie, Pierre (1859-1906) est considéré comme un des pionniers de la chimie/physique sur la radioactivité. C'est même lors d'une thèse publiée en 1898 que le terme radioactivité fut employé pour la première fois par sa femme Marie et lui. L'éducation de Pierre commença à un très jeune âge par son père, qui était médecin général. Les Curie avaient l'habitude de fréquenter la campagne et les environs de Paris les dimanches ; Pierre, lors de ses promenades, apprit rapidement tous les noms de plantes et d'animaux. Étant donné que l'école n'était pas obligatoire à cette époque (pas avant 1881 où la loi Ferry l'a rendue obligatoire), Pierre reçut son éducation à la maison, en compagnie de sa mère, ensuite avec son frère et par après, avec des précepteurs et finalement, seul. À l'âge de 14 ans, l'éducation de Pierre fut confiée à M. Bazille qui lui enseigna la mathématique élémentaire et spéciale, ceci développa énormément les capacités mentales de Pierre qui avait clairement un intérêt pour la mathématique. À l'âge de 16 ans, il fut reçu bachelier en sciences. En 1877, il obtint la licence en sciences physiques de l'école de pharmacie. Dans les années qui suivront, il étudiera les cristaux et le magnétisme, ce qui le mènera éventuellement à la découverte de la piézo-électricité. En 1877, il prit un poste comme préparateur où il fut payé la somme de 1200 francs par année. Il devint par après démonstrateur

d'expériences de physique pour les laboratoires jusqu'en 1882 où il devint directeur de tous les travaux pratiques aux écoles de physique et de chimie industrielle. Pierre épousa sa femme Marie Sklodowska en 1895 et ils eurent ensemble deux enfants, Irène et Ève. Pierre Curie gagna en 1903, avec sa femme, le prix Nobel de physique pour leurs travaux sur les substances radioactives et leurs découvertes de deux nouveaux éléments: le radium et le polonium.



Dalton, John (1766-1844), chimiste et physicien britannique, qui développa la théorie atomique sur laquelle fut fondée la science physique moderne. Dalton commença en 1787 une série d'observations météorologiques qu'il poursuivit pendant 57 ans, accumulant quelque deux cent mille observations et mesures du temps dans la région de Manchester. L'intérêt de Dalton pour la météorologie le conduisit à étudier différents phénomènes ainsi que les instruments utilisés pour les mesurer. Il fut le premier à prouver la validité de l'idée selon laquelle la pluie est précipitée par une baisse de température, non par un changement de la pression atmosphérique. Dalton arriva à sa théorie atomique par une étude des propriétés physiques de l'air atmosphérique et des autres gaz. Au cours de ses recherches, il découvrit la loi des pressions partielles des gaz mélangés, souvent connue comme la "loi de Dalton", selon laquelle la pression totale exercée par un mélange de gaz est égale à la somme des pressions individuelles qu'exercerait chacun des gaz s'il occupait seul le volume entier.



Da Vinci, Leonardo (1452-1519), est un peintre, sculpteur, architecte et homme de sciences italien. Homme d'esprit universel, à la fois artiste, scientifique, inventeur et philosophe, Léonard incarna l'esprit universaliste de la Renaissance et demeure l'un des grands hommes de cette époque. À 5 ans, son père ayant noté ses dons pour le dessin, le place comme apprenti dans l'atelier de Verrocchio, à Florence. Il entre à 20 ans à la Guilde des peintres, et débute sa carrière de peintre par des oeuvres immédiatement remarquables comme *La vierge à l'oeillet*, ou *L'Annonciation* (1473). Il améliore la technique du sfumato (impression de brume) à un point de raffinement jamais atteint avant lui. En 1481, le monastère de San Donato lui commande *L'Adoration des Mages*, mais Léonard, vexé de pas être choisi pour la décoration de la chapelle Sixtine à Rome, ne terminera jamais ce tableau et quitte Florence pour Milan. Après la réalisation de *La Vierge aux rochers*, pour la chapelle San Francesco Grande, et celle de la statue équestre de Francesco Sforza, il trouve la gloire dans toute l'Italie. En 1495, les Dominicains de Sainte-Marie-des-Grâces lui commandent *La Cène*. En 1498, il réalise le plafond du palais Sforza. De cette époque, datent aussi *La Joconde* et *La Bataille d'Anghiari*. Léonard réalise aussi une grande quantité d'études sur la zoologie, la botanique, l'anatomie, la géologie. Il imagine de multiples appareils et machines, dont la première machine volante, qui resteront au stade de dessins. Plus qu'en tant que scientifique proprement dit, Léonard de Vinci a impressionné ses contemporains et les générations suivantes par son approche méthodique du savoir, du savoir apprendre, du savoir observer, du savoir analyser. La démarche qu'il déploya dans l'ensemble des activités qu'il abordait, aussi bien en art qu'en technique (les deux ne se distinguant d'ailleurs pas dans son esprit), procédait d'une accumulation préalable d'observations détaillées, de savoirs disséminés ça et là, qui tendait vers un surpassement de ce qui existait déjà, avec la perfection

pour objectif. Bon nombre des croquis, notes et traités de Léonard de Vinci ne sont pas à proprement parler des trouvailles originales, mais sont le résultat de recherches effectuées dans un souci encyclopédique, avant l'heure. En 1516, il rejoint la cour de François Ier, où il participe à des projets d'urbanisme. De Léonard de Vinci, subsistent aujourd'hui 7'000 notes et dessins, et quarante oeuvres attestées, dont huit ont disparu.



Dantzig, George Bernard (1914-2005), était un mathématicien américain, inventeur du fameux algorithme du simplexe en optimisation linéaire. Son père, Tobias, est un mathématicien russe qui avait étudié avec Poincaré à Paris. Il a épousé une collègue de la Sorbonne, Anja Ourisson, et le couple a émigré aux États-Unis. Il est l'acteur principal d'une histoire fameuse en mathématique. Dans l'un de ses cours de doctorat à l'Université de Berkeley, le professeur Jerzy Neyman a proposé deux problèmes dits ouverts en statistiques. Un problème ouvert est un problème qui bien qu'ayant été formulé, n'a pas encore été résolu. De tels problèmes sont d'une difficulté importante et demandent des recherches pouvant s'étaler sur plusieurs années. Dantzig était en retard et croyait qu'il s'agissait de devoirs. Sans prendre plusieurs années mais bien quelques jours, il les a résolus. Il a reçu son doctorat de Berkeley en 1946. Six ans plus tard, il était engagé pour faire de la recherche mathématique à la RAND Corporation, où il implante l'algorithme du simplexe dans les ordinateurs. En 1960, l'Université de Berkeley l'engage pour enseigner en informatique, pour éventuellement devenir le responsable du centre de recherche opérationnelle. Six ans plus tard, il occupe un poste similaire à l'université Stanford, poste qu'il occupe jusqu'à sa retraite pendant les années 1990. En plus de ses travaux sur l'algorithme du simplexe et l'optimisation linéaire, il a aussi travaillé sur les méthodes de décomposition des problèmes de grande taille, l'analyse de sensibilité, les méthodes de résolution matricielles avec pivot, l'optimisation non linéaire et l'optimisation stochastique.



Descartes, René (1596-1650), philosophe, scientifique et mathématicien français, fondateur du rationalisme moderne. Né à La Haye, d'un père conseiller au parlement de Rennes, Descartes reçut, de 1607 à 1614, l'enseignement, décisif pour lui, des pères jésuites du Collège royal de La Flèche. Cette expérience le conduisit à proposer une refondation des sciences, critiquant l'absence de fondement de l'enseignement professé. Il reçut une formation de juriste en 1616 puis entra dans la carrière militaire en 1618, entreprit des voyages, mêla vie scientifique et vie mondaine, avant de se consacrer pleinement à la philosophie. Il passa sa vie entre la France et les Pays-Bas, fuyant les villes, fréquentant les bibliothèques et rencontrant les esprits les plus illustres de son temps, notamment Bérulle, Fermat, Gassendi, Hobbes et Pascal. Il mourut d'une pneumonie à Stockholm, léguant à la postérité une oeuvre entourée de légendes et imprégnée d'un esprit nouveau.



Dirac, Paul Adrien Maurice (1902-1984). Né à Bristol, Dirac fait ses études aux universités de Bristol et de Cambridge. En 1926, pour son doctorat (la première thèse au monde ayant pour sujet la "mécanique quantique"), il introduit un formalisme général pour la physique quantique peu après Heisenberg mais indépendamment (il y retrouve la non-commutativité des opérateurs position et quantité de mouvement). En 1928, il élabore une théorie relativiste pour décrire les propriétés de l'électron. Celle-ci le conduit à postuler l'existence d'une particule identique à l'électron dans tous ses aspects mais de charge opposée, c'est-à-dire positive et devant s'annihiler en même temps que l'électron négatif lors d'une collision avec celui-ci. La théorie de Dirac est confirmée en 1932 quand le physicien américain Carl Anderson découvre le positron. Dirac contribue aussi, avec Fermi, au développement de la statistique dite de Fermi-Dirac, décrivant le comportement collectif des particules de spin demi-entier. En 1933, Dirac partage le prix Nobel de physique avec le physicien autrichien Erwin Schrödinger. En 1939, il devient membre de la Société royale. Il est professeur de mathématiques à Cambridge de 1932 à 1968, professeur de physique à l'université d'État de Floride de 1971 jusqu'à sa mort, et membre de l'Institute of Advanced Studies périodiquement entre 1934 et 1959.

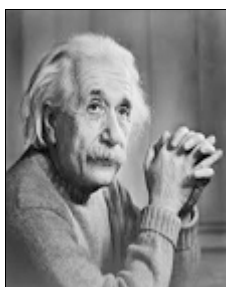


Dirichlet(-Lejeune), Peter-Gustav (1805-1859) est né à Düren (Allemagne). Dirichlet est un élève brillant, qui achève ses études secondaires à 16 ans. Devant la faible qualité des formations universitaires allemandes à cette époque, Dirichlet décide de partir étudier à Paris, emportant avec lui les *Disquisitiones Arithmeticae* de Gauss comme une bible. Dans la capitale française, sa situation personnelle est facilitée par le général Foy, un ancien grand général des campagnes napoléoniennes, dont il devient le précepteur des enfants, et qui se montrera bienveillant avec lui. Dirichlet rencontre alors quelques-uns des plus grands mathématiciens, dont Legendre, Poisson, Laplace et Fourier. Ce dernier surtout impressionnera beaucoup Dirichlet, et sera à l'origine de l'intérêt qu'il portera aux séries trigonométriques et à la physique-mathématique. C'est à Paris que Dirichlet rédige sa première contribution d'importance aux mathématiques, étant à l'initiative en 1825 de la preuve du cas $n=5$ dans le grand théorème de Fermat, preuve achevée par Legendre dans la foulée. Fin 1825, le général Foy décède, et Dirichlet décide de retourner en Allemagne. Il enseigne d'abord à l'université de Breslau, au lycée militaire de Berlin, puis à l'université de Berlin à partir de 1829, où il restera 27 ans durant. Parmi ses élèves, on retiendra les noms de Kronecker et Riemann. Dirichlet est décrit comme un bon professeur, mais non exempt de défauts. Il donne l'apparence de quelqu'un de sale, toujours affublé d'un cigare et d'un café, visiblement peu préoccupé de l'image qu'il donne. On dit aussi de lui qu'il était très souvent en retard. En 1848, son maître et ami Karl Jacobi est diagnostiqué comme étant malade du diabète. Dirichlet l'accompagne dans un voyage de 18 mois en Italie. De retour en Allemagne, Dirichlet commence à être lassé des lourdes charges d'enseignement qu'il doit assumer. À la mort de Gauss, il prend sa succession à Göttingen. C'est malheureusement pour peu de temps, car lui-même s'éteint en 1859 des suites d'un malaise cardiaque. L'éventail des travaux de Dirichlet illustre la profondeur de la culture mathématique allemande au début de son âge d'or. On lui doit le premier énoncé d'une condition suffisante de convergence d'une série de Fourier (dans le cas des

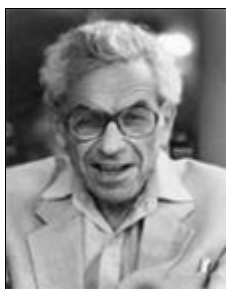
fonctions continues par morceaux), le théorème de la progression arithmétique, le prolongement des fonctions harmoniques définies sur la frontière d'un ouvert et toute une classe d'équations aux dérivées partielles porte le nom de "problème de Dirichlet". Nous lui devons aussi de très nombreuses contributions en arithmétique, où il existe le théorème des unités de Dirichlet, les séries de Dirichlet, etc.



Doppler, Christian (1803-1853), est un mathématicien et physicien autrichien, célèbre pour sa découverte de l'effet Doppler. Après avoir étudié à l'Université de Vienne, Doppler devient assistant-professeur dans cet établissement en 1829. Ce poste n'étant pas renouvelé, il envisage un temps une émigration vers les États-Unis. Il renonce à quitter son pays après avoir été nommé à Prague en 1837, puis à l'École polytechnique de Vienne en 1849. En 1850, il fonde l'Institut de Physique de l'Université de Vienne dont il est seul professeur et le premier directeur. Atteint d'une affection pulmonaire, la tuberculose, il quitte ses fonctions en 1852. Son travail scientifique est varié : optique, astronomie, électricité... Sa publication la plus célèbre a été présentée en 1842 à l'Académie royale des sciences de Bohême et a pour titre *Sur la lumière colorée des étoiles doubles et d'autres étoiles du ciel*, utilisant l'effet Doppler. Ses calculs étaient erronés, le décalage réel de la fréquence lumineuse étant trop faible pour pouvoir être détecté à l'époque. En 1846, Doppler publie une correction de son travail initial où il tient compte des vitesses relatives de la source de lumière et de l'observateur.

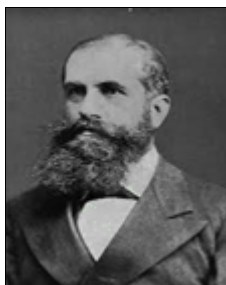


Einstein, Albert (1879-1955), né à Ulm et mort à Princeton, est un physicien théoricien qui fut successivement allemand, puis apatride (1896), suisse (1901), et enfin sous la double nationalité helvético-américaine (1940). Il publie sa théorie de la relativité restreinte en 1905, et une théorie de la gravitation dite relativité générale en 1915. Il contribue largement au développement de la mécanique quantique et de la cosmologie, et reçoit le prix Nobel de physique de 1921 pour son explication de l'effet photoélectrique. Son travail est notamment connu pour l'équation d'équivalence qui établit une équivalence entre la matière et l'énergie d'un système. Il est aussi connu pour son hypothèse audacieuse sur la nature corpusculaire de la lumière. Mais il a également contribué au développement de nombre d'autres théories (physique quantique y comprise). En 1905, Einstein obtint son doctorat de l'université de Zurich pour une thèse théorique sur les dimensions des molécules. Il publia également trois articles théoriques d'une importance capitale pour le développement de la physique du XXe siècle. Dans le premier de ces articles, sur le mouvement brownien, il fit des prédictions importantes sur le mouvement des particules distribuées aléatoirement dans un fluide. Pendant le reste de sa vie, Einstein consacra énormément de temps à généraliser encore plus sa théorie de la relativité générale. Il visait une théorie de champ unifié, qui ne fut pas complètement couronnée de succès, et fit de nombreuses tentatives pour décrire l'interaction électromagnétique et l'interaction gravitationnelle dans un modèle commun.



Erdős, Paul (1913-1996) est le plus prolifique des mathématiciens du 20ème siècle, avec environ 1'500 articles publiés (il faut remonter à Euler pour obtenir un tel volume). Plus que quelqu'un qui bâtissait des théories, il résolvait des problèmes, le plus souvent avec élégance et simplicité. Erdős est né à Budapest. Ses deux parents étaient professeurs de mathématiques dans le secondaire. Alors que Erdős était âgé d'à peine un an, son père fut fait prisonnier par les Russes et déporté en Sibérie. Ces événements ont contribué au développement d'une relation très forte mère/fils, qui influera beaucoup sur le cours de la vie de Paul Erdős. C'est à l'âge de 19 ans, alors qu'il commence ses études à l'université et qu'il se fait connaître des milieux mathématiques. Il publie en effet une nouvelle démonstration du postulat de Bertrand, qui affirme qu'il existe un nombre premier entre n et $2n$, pour tout n . Deux ans plus tard, il obtient son doctorat (à 21 ans), puis s'en va faire un post-doc à Manchester. Comme Erdős est d'origine juive, il ne peut retourner en Hongrie à la fin des années 30, et il émigre aux États-Unis. Après quelques visites en Europe aux rescapés de sa famille après l'Holocauste, il a des problèmes aux États-Unis avec le MacCarthysme, et il se voit interdit de séjour sur le territoire américain. Erdős est donc contraint de poser ses valises en Israël. Avec ses 1'500 articles, les contributions de Erdős aux mathématiques sont nombreuses: en théorie des nombres, en combinatoire, en mathématiques discrètes, il fut un maître. Erdős avait une exceptionnelle aptitude à s'entourer des mathématiciens les plus compétents pour résoudre ses conjectures. Il en résulte que Erdős a eu beaucoup de collaborateurs: 500 mathématiciens environ ont écrit un article en commun avec lui. Les mathématiciens se sont amusés à définir un nombre de Erdős: tout mathématicien qui a publié un papier en commun avec Erdős a un nombre de Erdős égal à 1. Toute personne qui a publié un article en commun avec une personne qui a un nombre de Erdős égal à 1 a un nombre de Erdős égal à 2. Et ainsi de suite... Albert Einstein est l'un d'entre eux: son nombre de Erdős est 2. Pourtant, parmi toutes ces collaborations, une au moins a mal tourné, et c'est d'autant plus regrettable qu'elle concerne le plus grand succès d'Erdős. À la fin du 19ème siècle, Hadamard et de La Vallée Poussin avaient démontré le théorème des nombres premiers, à savoir que le nombre de nombres premiers inférieurs ou égaux à n est équivalent, quand n est grand, à $n/\ln(n)$. Leur démonstration est particulièrement rude! En 1949, Atle Selberg trouve une inégalité qu'il pense pouvoir être une étape importante vers une démonstration élémentaire du théorème des nombres premiers. Elle est présentée à Erdős, qui trouve la clef manquante pour boucler la preuve. Un article coécrit de plus aurait sans doute été la solution la plus appropriée pour mesurer les apports de chacun. Mais, à la suite d'un malentendu lié à l'envoi de cartes postales triomphales d'Erdős, Selberg craint qu'Erdős ne tire la couverture à lui. Il publie seul une preuve complète. Il recevra la médaille Fields en 1950, alors qu'Erdős devra se contenter du prix Wolf en 1984. La vie d'Erdős fut vraiment étrange. Il n'avait pas de maison, pas d'épouse, les contingences matérielles étaient pénibles pour lui. Il voyageait en solitaire, accompagné de deux valises qui portaient toutes ses affaires, allant d'université en université, habitant à l'hôtel ou chez un ami mathématicien... Il est par ailleurs l'auteur de nombreux "erdosismes", comme cette phrase célèbre: "un mathématicien est une machine à transformer le café en théorème". Lui-même dopé à toutes sortes d'amphétamines! Jusqu'à la fin de sa vie, Erdős ne ralentira pas son activité mathématique. Mourir signifiait pour lui

arrêter de faire des mathématiques. Il décède à Varsovie, en plein congrès.



Erlang, Agner Krarup (1878-1979), était un mathématicien danois ayant beaucoup travaillé sur la théorie des files d'attente, et la gestion des réseaux téléphoniques. Erlang s'est attelé, sur la base notamment des travaux de Poisson dont la loi des événements rares a trouvé toute sa dimension appliquée aux réseaux de télécommunications, à l'élaboration d'un modèle mathématique pour le dimensionnement des réseaux télécoms sur une approche statistique afin de parvenir à des coûts d'exploitation de nature à permettre un marché de masse.

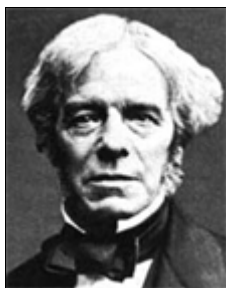


Euclide (3^e siècle av. J.-C.) On ne sait que très peu de choses sur la vie d'Euclide. Il semble qu'il ait enseigné la mathématique à Alexandrie à la demande de Ptolémée Ier. Il apparaîtrait donc comme le fondateur de la célèbre École d'Alexandrie qui influença les travaux d'Archimède. En revanche, les théories d'Euclide sont connues et constituent une référence dans l'histoire des mathématiques. L'oeuvre maîtresse d'Euclide est incontestablement les *Éléments*. Cet ouvrage représente une synthèse remarquable de résultats mathématiques et a marqué de son empreinte la discipline tout entière. Il est composé de treize livres. Les quatre premiers traitent de géométrie dans le plan avec les définitions du point, de la droite et de la surface. Ils exposent également le calcul d'aires de différents polygones. Le livre V contient les premières notions d'analyse. Le sixième aborde la similitude des figures et donne la résolution des équations du second degré à l'aide de constructions géométriques. Les livres VII, VIII, et IX portent sur l'arithmétique. Le dixième étudie les nombres irrationnels et enfin les trois derniers abordent la géométrie dans l'espace. Euclide a, en outre, rédigé des ouvrages sur l'analyse géométrique, l'optique et l'astronomie. Représentation parfaite de l'exposé scientifique, les *Éléments* sont composés de différentes propositions classées en deux groupes: les hypothèses et les axiomes. Parmi les 5 axiomes, on trouve le célèbre postulat d'Euclide: "par tout point du plan passe une et une seule droite parallèle à une autre droite." Cet axiome constitue le fondement de la géométrie euclidienne, en opposition aux géométries non-euclidiennes apparues quelque 2000 ans plus tard.



Euler, Leonhard (1707-1783), mathématicien suisse, physicien, ingénieur et philosophe, est l'un des fondateurs des méthodes de calcul différentiel et intégral. Leonhard Euler naquit à Bâle¹, de Paul Euler, pasteur des Églises réformées et de Marguerite Brucker, fille de pasteur. Il eut deux jeunes sœurs du nom d'Anna Maria et de Maria Magdalena⁶. Peu de temps après la naissance de Leonhard, la famille Euler déménagea de Bâle pour rejoindre la ville de Riehen, où Euler passa la plupart de son enfance. Paul Euler était un ami de la famille Bernoulli. Jean Bernoulli, alors considéré comme le principal mathématicien européen, pourrait être celui ayant eu la plus grande influence sur le jeune Leonhard. L'éducation officielle d'Euler commença tôt à Bâle, où il fut envoyé vivre avec sa grand-mère maternelle. À l'âge de 13 ans, il s'inscrivit à l'université de Bâle, et en 1723, obtint son Master of Philosophy grâce à une dissertation qui comparait la philosophie de Descartes à celle de Newton. À cette époque, il recevait tous les samedis après-midi des leçons de Jean Bernoulli, qui découvrit rapidement chez son

nouvel élève un incroyable talent pour les mathématiques. Euler commença alors à étudier la théologie, le grec et l'hébreu à la demande de son père, afin de devenir un pasteur, mais Jean Bernoulli convainquit Paul Euler que Leonhard était destiné à devenir un grand mathématicien. Euler fut le premier à traiter de manière analytique et complète l'algèbre, la théorie des équations, la trigonométrie et la géométrie analytique. Dans ce travail, il traita le sujet du développement des séries de fonctions et formula la règle selon laquelle seules les séries infinies convergentes pouvaient être correctement évaluées. Il discuta aussi des surfaces à trois dimensions et prouva que les sections coniques sont représentées par l'équation générale du second degré à deux dimensions. D'autres travaux traitent du calcul, dont le calcul des variations, la théorie des nombres, les nombres imaginaires et transcendants, l'algèbre déterminée et indéterminée, la théorie des graphes. Euler apporta ses contributions dans les domaines de l'astronomie, de la mécanique analytique (calcul variationnel), l'hydrodynamique, l'optique et l'acoustique. Euler est considéré comme un éminent mathématicien du 18^{ème} siècle et l'un des plus grands et des plus prolifiques de tous les temps et a introduit une grande partie des notations encore utilisées en ce début de 21^{ème} siècle (symboles pour la somme, fonction, logarithme, exponentiel, etc.).



Faraday, Michael (1791-1867) était un scientifique anglais qui a contribué aux domaines de l'électromagnétisme et l'électrochimie. Le jeune Faraday, qui était le troisième des 4 enfants, n'ayant que l'éducation scolaire la plus élémentaire, a dû se former en autodidacte. À 14 ans, il devient apprenti chez un relieur local. Au cours de ses 7 années d'apprentissage il a pu lire beaucoup de livres. A cette époque, il a également développé un intérêt pour la science, en particulier pour l'électricité. Faraday a été particulièrement inspiré par le livre *Conversations on Chemistry* de Jane Marcet. Ses principales découvertes comprennent l'induction électromagnétique, le diamagnétisme et l'électrolyse. C'est par ses recherches sur le champ magnétique autour d'un conducteur parcouru par un courant continu que Faraday a établi la base pour le concept de champ électromagnétique en physique. Faraday a également établi que le magnétisme pourrait affecter rayons de lumière et qu'il y avait une relation sous-jacente entre les deux phénomènes. Il a découvert le principe même de l'induction électromagnétique, en même temps que Joseph Henry, diamagnétisme, et les lois de l'électrolyse. Ses inventions de dispositifs électromagnétiques rotatifs ont formé la base de la technologie des moteurs électriques, et c'est en grande partie grâce à ses efforts que l'électricité est devenue pratique pour une utilisation dans la technologie.



Fermat, Pierre de (1601-1665), mathématicien français, auteur d'un célèbre théorème sans démonstration en arithmétique et surnommé le "prince des amateurs". Il fut à l'origine du principe de Fermat (optique) et avec son ami Blaise Pascal de celui du calcul des probabilités. Il créa également la théorie des nombres et fit dans ce domaine différentes découvertes. Ainsi, certains le considèrent comme le père de la théorie moderne. Il devança le calcul différentiel par ses travaux sur le calcul infinitésimal. Il laissa à la postérité le soin de démontrer un théorème (le fameux "grand théorème de Fermat") sur lequel les mathématiciens s'acharnent depuis plus de trois siècles. Ce n'est qu'en 1993 que le

chercheur britannique Andrew Wiles en proposa une démonstration.



Fermi, Enrico (1901-1954), physicien italien, connu pour la réalisation de la première réaction nucléaire contrôlée. Très jeune, Enrico Fermi fait preuve d'une mémoire exceptionnelle et d'une grande intelligence, qui lui permettent d'exceller dans les études. Enrico, profondément marqué par le décès d'un de ses très jeune frère, se jette alors dans l'étude de la physique pour surmonter sa douleur. Bon élève, il se passionne très vite pour la physique et les mathématiques et commence à étudier divers ouvrages qu'il achète et qui traitent de mécanique, d'optique, d'astronomie et d'acoustique. Un ami de son père, l'ingénieur Adolfo Amidei, qui prend conscience des qualités hors du commun du jeune Fermi, lui prête divers ouvrages traitant de mathématiques. Ainsi, à 17 ans, Enrico Fermi maîtrise la géométrie analytique, la géométrie projective, le calcul infinitésimal, le calcul intégral et la mécanique rationnelle. À partir de 1918, Fermi étudie à l'Université de Pise au sein de l'École normale supérieure de Pise. Comme à son habitude, il étudie seul divers problèmes de physique mathématique et consulte des ouvrages de Poincaré, de Poisson ou d'Appell. À partir de 1919, il s'intéresse aux nouvelles théories comme la relativité ou la physique atomique, ainsi il acquiert une grande connaissance de théories telles que la relativité restreinte, la théorie du corps noir ou encore le modèle de l'hydrogène de Bohr. Ainsi Enrico Fermi, le seul à l'université au fait de ces théories, en arrive, sur l'insistance de ses professeurs, à donner des conférences où il expose aux professeurs et aux assistants les dernières découvertes de physique atomique. En 1922, après 4 ans passés à l'université, Enrico Fermi publie son premier article qui traite de la relativité générale. Dans une communauté scientifique italienne hostile aux travaux d'Einstein, il est l'un des rares avec Levi-Civita à défendre la théorie de la relativité. En 1922, Fermi obtient son diplôme de fin d'études après avoir présenté un mémoire sur la diffraction des rayons X. Il fréquente ensuite divers physiciens de haut rang dans l'Italie de l'époque, avant de devenir, pendant 2 ans, conférencier à l'université de Florence. En 1926, il devient professeur de physique théorique à l'université La Sapienza de Rome. C'est durant cette période qu'il développe la théorie statistique quantique que l'on appellera plus tard la statistique de Fermi-Dirac. À partir de 1932, il se tourne plus précisément vers la physique nucléaire, et c'est cette même année qu'il rédige un article sur la radioactivité bêta. En 1934, il développe sa théorie sur l'émission de rayonnement bêta en y incluant le neutron postulé en 1930 par Wolfgang Pauli, qu'il rebaptise neutrino (le nom neutron étant déjà utilisé pour une autre particule), et s'oriente vers la création d'isotopes radioactifs artificiels par bombardement de neutrons lents (raison pour laquelle il reçut le prix Nobel en 1938).



Feynman, Richard Phillips (1918-1988), est né à Far Rockaway dans le Queens, quartier de New York (États-Unis) de parents d'origine polonaise et russe. Son père, qui l'encourageait à poser des questions et à remettre en cause les choses communément admises, l'a durablement influencé. De sa mère, il tient un solide sens de l'humour qui ne l'a jamais quitté. Feynman est l'un des physiciens les plus influents de la seconde moitié du 20^{ème} siècle, en raison notamment de ses travaux sur l'électrodynamique quantique relativiste, les quarks et l'hélium

superfluide. Durant sa dernière année à l'école secondaire de Far Rockaway, Feynman remporta le championnat de mathématiques de l'université de New York. Il reçut donc une bourse pour étudier au Massachusetts Institute of Technology (MIT) où il reçut son baccalauréat en 1939 après s'être orienté d'abord en électronique, puis en mathématiques, et enfin avoir assisté à tous les cours de physique offerts y compris pendant sa seconde année un cours de physique théorique réservé aux étudiants de maîtrise. Feynman obtient un score remarquable aux examens d'entrée de l'Université de Princeton en mathématiques et en physique, mais il eut une note très faible dans la partie littéraire de l'examen. Durant ses études à l'Institute for Advanced Study de Princeton (IAS) (créé depuis peu et dirigé par Albert Einstein), Feynman travailla sous la direction de John Wheeler sur le principe de moindre action appliqué à la mécanique quantique. Il établit ici les bases des diagrammes de Feynman et de l'approche de la mécanique quantique par les intégrales de chemin. Il obtint son doctorat en 1942. Il reformula entièrement la mécanique quantique à l'aide de son intégrale de chemin qui généralise le principe de moindre action de la mécanique classique et inventa les diagrammes qui portent son nom et qui sont désormais largement utilisés en théorie quantique des champs (dont l'électrodynamique quantique fait partie). Musicien, pédagogue remarquable, rédacteur de nombreux ouvrages de vulgarisation, il a aussi été impliqué dans le développement de la bombe atomique américaine. Après la Seconde Guerre mondiale, il enseigna à l'université Cornell puis au Caltech où il effectua des travaux fondamentaux notamment dans la théorie de la superfluidité et des quarks. Sin-Itiro Tomonaga, Julian Schwinger et lui sont colauréats du prix Nobel de physique de 1965 pour leurs travaux en électrodynamique quantique.



Fisher, Ronald Aymler (1890-1962) né à Londres était un biologiste et statisticien britannique, qui a énormément contribué à fonder les statistiques modernes. Ses travaux sur les statistiques lui valurent la médaille Darwin en 1948, la médaille Copley en 1955 et la médaille d'argent Darwin-Wallace en 1958. Dans le domaine des statistiques, il a introduit de nombreux concepts clés tels que le maximum de vraisemblance, l'information de Fisher et l'analyse de la variance (ANOVA). Il est considéré comme un grand précurseur de Shannon. Il est également un des fondateurs de la génétique moderne et un grand continuateur de Darwin, en particulier grâce à son utilisation des méthodes statistiques, incontournables dans la génétique des populations. Il a ainsi contribué à la formalisation mathématique du principe de sélection naturelle. Il est d'abord attiré par la physique et obtient en 1912 une licence d'astronomie à l'université de Cambridge. De 1915 à 1919, il enseigne les mathématiques à Londres dans des écoles privées. En 1919, il est engagé à la station expérimentale de Rothamsted pour analyser l'effet des précipitations sur le rendement du blé où il travaille jusqu'en 1933. Dans son article de 1922 *On the mathematical foundations of theoretical statistics*, il définit une quinzaine de notions fondamentales en statistiques dont la notion de convergence, d'efficacité, de vraisemblance et de statistique suffisante. Il propose l'estimateur du maximum de vraisemblance en 1922 après avoir présenté une première version en 1912. Il introduit aussi en 1924 l'analyse de la variance. En 1925 il publie des innovations en séries temporelles et en analyse des corrélations multiples.



Foucault, Léon (1819-1868), physicien français célèbre pour sa démonstration du mouvement de la Terre par la rotation du plan d'oscillation du pendule. Né à Paris, il travailla avec le physicien français Armand Fizeau sur la détermination de la vitesse de la lumière. Foucault prouva, de façon indépendante, que la vitesse de la lumière dans l'air était plus élevée que dans l'eau. En 1851, il fit une démonstration spectaculaire de la rotation de la Terre en suspendant un pendule à un long câble attaché à la coupole du Panthéon à Paris. Le mouvement du pendule démontra la rotation de la Terre sur son axe. En 1855 il découvre que la force nécessaire à la rotation d'un disque de cuivre augmente quand il doit tourner avec sa jante entre les pôles d'un aimant, le disque chauffant dans le même temps du fait des "courants de Foucault" induits dans le métal. Il conçut également une méthode de mesure de la courbure des miroirs de télescopes. Il développa d'autres instruments dont un prisme polarisateur et une forme de gyroscope qui est à la base du gyrocompas moderne.



Fourier, Joseph (1768-1830), physicien et mathématicien français connu pour la découverte des séries trigonométriques et des transformées qui portent son nom. Fourier est orphelin de père et de mère à 10 ans. L'organiste d'Auxerre, Joseph Pallais, le fait entrer dans le pensionnat qu'il dirige. Recommandé par l'évêque d'Auxerre, il fait ses études à l'École militaire d'Auxerre tenue alors par les Bénédictins de la Congrégation de Saint-Maur. Destiné à l'état monastique, il préfère s'adonner aux sciences pour lesquelles il remporte la plupart des premiers prix. Élève brillant, il y est promu professeur dès l'âge de 16 ans et peut dès lors commencer ses recherches personnelles. Il intègre l'École normale supérieure à 26 ans, où il a entre autres comme professeurs Joseph-Louis Lagrange, Gaspard Monge et Pierre-Simon de Laplace, auquel il succède à la chaire à Polytechnique en 1797. Fourier a contribué à la résolution numérique des équations et à la diffusion de la chaleur dont une des lois porte son nom. Ses travaux ont une implication directe dans la convergence des séries et leur somme infinie. Il participa, avec Monge, à la campagne d'Égypte en tant qu'observateur scientifique. Anobli sous Napoléon, il fut professeur à l'École Polytechnique, secrétaire de l'institut d'Égypte et préfet de l'Isère. Il fut aussi élu à l'Académie des sciences et à l'Académie française. On le considère comme l'un des fondateurs, avec le français Poisson et le suisse Daniel Bernoulli, de ce que l'on appelle aujourd'hui la physique-mathématique.

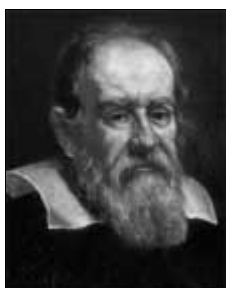


Fraunhofer, Joseph von (1787-1826), opticien et physicien allemand, né à Straubing. Fraunhofer apporta de nombreuses améliorations à la fabrication du verre optique, au meulage et au polissage des lentilles et à la construction des télescopes et d'autres instruments d'optique. Joseph Fraunhofer était le onzième enfant d'un souffleur de verre. Il avait 11 ans à la mort de ses parents: aussi son tuteur l'envoya-t-il à Munich en apprentissage pour 6 ans afin qu'il apprenne la miroiterie. C'est là, qu'en 1801, il faillit trouver la mort dans l'effondrement de l'atelier. À la fin de son apprentissage en 1806, il eut la possibilité de poursuivre une formation d'opticien dans l'Institut de Mécanique Reichenbach. Les ateliers furent transférés en 1807 à Benediktbeuern, et Fraunhofer y fut nommé contremaître. Là, il mit au point de nouvelles machines à polir les miroirs et de

nouveaux types de verres optiques (le verre flint achromatique), qui apportèrent une amélioration décisive à la qualité des lentilles. Fraunhofer inventa aussi de nombreux instruments scientifiques. Son nom est associé à des lignes fixes et noires dans le spectre solaire, appelées les "lignes Fraunhofer", qu'il fut le premier à décrire en détail. Ses recherches dans le domaine de la réfraction et de la dispersion de la lumière aboutirent à l'invention du spectroscope et au développement de la spectroscopie.



Fresnel, Augustin Jean (1788-1827), physicien français, fondateur de l'optique moderne, il proposa une explication de tous les phénomènes optiques dans le cadre de la théorie ondulatoire de la lumière. Il commença par réaliser de nombreuses expériences sur les interférences lumineuses, pour lesquelles il forgea la notion de longueur d'onde, et calcula les intégrales dites de Fresnel. Il fut le premier à prouver que deux faisceaux de lumière polarisés dans des plans différents n'ont aucun effet d'interférence. Il déduisit très justement de cette expérience que le mouvement ondulatoire de la lumière polarisée est transversal et non longitudinal (comme celui du son) ainsi qu'on le croyait avant lui. En outre, il fut le premier à produire une lumière polarisée circulaire. Pour expliquer la propagation des ondes lumineuses, Fresnel eut recours à la notion d'éther, malheureusement contradictoire avec d'autres expériences. Cette théorie sera abandonnée avec la relativité, mais les relations dites de Fresnel sur la réfraction sont toujours utilisées. Dans le domaine de l'optique appliquée, Fresnel conçut la lentille à échelons utilisée pour accroître le pouvoir éclairant des phares. De son vivant, les travaux scientifiques de Fresnel n'étaient connus que d'un petit groupe de scientifiques et certains de ses articles ne furent publiés qu'après sa mort.



Galileo, Galilée (1564-1642), physicien et astronome italien né à Pise à l'origine de la révolution scientifique du 17^{ème} siècle. Ses théories ainsi que celles de l'astronome allemand Johannes Kepler servirent de fondement aux travaux du physicien britannique sir Isaac Newton sur la loi de l'attraction universelle. Sa principale contribution à l'astronomie fut l'amélioration considérable (quand la technique marchait...) de la lunette astronomique (ce qui lui a permis de procéder à des observations qui ont bouleversé la discipline) et la découverte des taches solaires, des montagnes et des vallées lunaires, des quatre plus grands satellites de Jupiter et des phases de Vénus. En physique, il découvrit la loi de la chute des corps et les mouvements paraboliques des projectiles. Ses études sur les oscillations du pendule pesant l'ont amené à inventer le pulsomètre. Cet appareil permettait d'aider à la mesure du pouls et fournissait un étalon de temps, qui n'existait pas à l'époque. Il débute aussi ses études sur la chute des corps. Dans l'histoire de la culture, Galilée est le symbole de la bataille livrée contre les autorités religieuses pour la liberté de la recherche (il avait cependant très bonne réputation et de très bonnes relations auprès des instances religieuses ce qui a aidé...). Dans le domaine des mathématiques et de la physique, il a contribué à faire avancer les connaissances relatives à la cinématique et la dynamique, jetant ainsi les fondements des sciences mécaniques. Il est de ce fait considéré comme le fondateur de la physique moderne.



Galois, Évariste (1811-1832) est un mathématicien français, qui a donné son nom à une branche des mathématiques: la théorie de Galois. Sa vie est tellement mythique qu'il est parfois difficile de démêler le mythe et la réalité. Dès 1827-1828, la fureur des mathématiques domine. Galois lit Legendre, Lagrange, Euler, Gauss, Jacobi. Le professeur, Louis-Paul-Émile Richard, admire le génie mathématique de son élève et garde les copies qu'il confiera à un autre de ses élèves: Charles Hermite. C'est l'époque où il publie son premier article dans les *Annales mathématiques* de Joseph Gergonne (il démontre un théorème sur les fractions continues périodiques). Il rédige aussi un premier mémoire sur la théorie des équations, envoyé à l'Académie des Sciences, perdu par Cauchy. Il échoue au concours d'entrée à Polytechnique. On raconte qu'il a jeté le chiffon à effacer la craie à la tête de son examinateur devant la stupidité des questions posées. Sur les conseils de son professeur, Galois entre à l'École Préparatoire (future École Normale). Il rédige le résultat de ses recherches dans un mémoire - *Conditions pour qu'une équation soit résoluble par radicaux* - afin de concourir au grand prix de mathématiques de l'Académie des Sciences. Fourier emporte le manuscrit chez lui et meurt peu après: le manuscrit est perdu, et le grand prix est décerné à Abel (mort l'année précédente), et à Jacobi. Pour des raisons politiques, Galois se retrouve en prison, où il y continue ses travaux. Libéré en 1832, il s'éprend en 1832 d'une femme, avec qui il rompt la même année. On ne sait trop pourquoi, mais un duel semble en résulter quelques jours plus tard. La nuit précédente, le 29 mai, Galois rassemble ses dernières découvertes dans une splendide lettre adressée à son ami Auguste Chevalier. De cette lettre naquit la légende selon laquelle Galois fit ses découvertes majeures en une seule nuit, pris par la fièvre de la mort. La matinée du 30 mai, Galois, abandonné, grièvement blessé, est relevé par un paysan et conduit à l'Hôpital Cochin. Il meurt le 31 mai 1832 dans les bras de son jeune frère et est enterré dans la fosse commune du cimetière de Montparnasse. Les travaux de Galois sont redécouverts une dizaine d'années plus tard par Liouville, qui en 1843 annonce à l'Académie des Sciences qu'il vient de trouver dans les papiers de Galois une solution aussi exacte que profonde au problème de la résolubilité par radicaux. Ce n'est qu'en 1846 qu'il publie les textes sans y joindre de commentaires. À partir de 1850, les écrits de Galois sont enfin accessibles par les meilleurs mathématiciens.

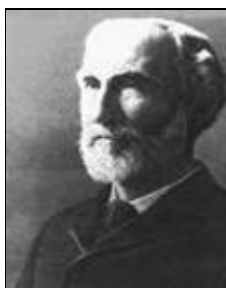


Gamow, George (1904-1968) est un physicien théorique, astronome, cosmologiste et auteur/vulgarisateur scientifique américano-russe né à Odessa en Ukraine. Gamow vient en 1928 à Göttingen, où il utilise la physique quantique pour faire une théorie quantique de la radioactivité alpha. Deux mois plus tard, il rejoint Niels Bohr à Copenhague. Il émet l'idée d'un noyau atomique se comportant comme un fluide nucléaire, modèle repris presque une décennie plus tard par Bohr. En 1929, il obtient une nouvelle bourse et il rejoint Ernest Rutherford à l'université de Cambridge. Il développe l'idée de l'effet tunnel afin de faire interagir des protons pour obtenir des noyaux de numéro atomique plus élevé. Il y rencontre John Cockcroft, qui construit peu après le premier accélérateur de particules, parvenant ainsi à valider le modèle de Gamow en réussissant une transmutation du lithium. Professeur à Washington en 1934, Gamow collabore avec Edward Teller pour formuler la théorie de l'émission bêta (1936). S'intéressant ensuite à l'astrophysique, Gamow et Teller donnent un

modèle de la structure interne des étoiles géantes rouges (1942). En 1954, c'est vers la biochimie qu'il se tourne, proposant le concept de code génétique déterminé par l'ordre des composants de l'ADN. En 1956, il est nommé professeur de physique à Boulder (Colorado).

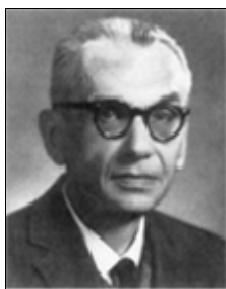


Gauss, Carl Friedrich (1777-1855), mathématicien allemand, qui a apporté des contributions essentielles à la plupart des branches des sciences exactes et appliquées. À l'âge de 17 ans, il essaya de trouver une solution au problème classique de construction d'un polygone à sept côtés, à la règle et au compas. Il réussit à prouver l'impossibilité de cette construction et poursuivit sa démarche en donnant des méthodes de construction de polygones à 17, 257, et 65'537 côtés. Plus généralement, il prouva que la construction, à la règle et au compas, d'un polygone régulier à nombre impair de côtés n'était possible que si le nombre de côtés est un des nombres premiers 3, 5, 17, 257, et 65'537, ou un produit de ces nombres. Pour sa thèse de doctorat, il démontra que toute équation algébrique a au moins une racine. Ce théorème, dont la démonstration avait résisté aux mathématiciens les plus célèbres, est encore appelé le théorème fondamental de l'algèbre ou théorème de d'Alembert-Gauss. Gauss tourna ensuite son attention vers le domaine de l'astronomie pour lequel il élaborait également une nouvelle méthode de calcul des orbites des corps célestes, en développant une théorie des erreurs d'observation connue sous le nom de méthode des moindres carrés. En probabilités, son nom est attaché à la loi normale (dite aussi loi de Laplace-Gauss), dont la répartition est décrite par la fameuse courbe en cloche ou courbe de Gauss. On lui doit aussi des travaux en géodésie. Avec le physicien allemand Wilhelm Eduard Weber, Gauss fit, à partir de 1831, des recherches approfondies dans le domaine du magnétisme et de l'électricité. Il fit aussi des recherches en optique, en particulier sur les systèmes de lentilles. Pour revenir aux mathématiques, il fut le premier, en étudiant la série hypergéométrique, à donner des conditions rigoureuses de convergence d'une série. Il étudia des généralisations fructueuses de la loi de réciprocité quadratique et dégagait leurs liens avec la théorie des fonctions elliptiques. Son mémoire de 1828 sur la théorie intrinsèque des surfaces fut le point de départ d'une théorie générale des espaces courbes (travaux de Riemann et de ses successeurs). Signalons aussi l'étude arithmétique des entiers de Gauss (de la forme $a+ib$) qui repose sur une présentation géométrique des nombres complexes comme points du plan.



Gibbs, Josiah Willard (1839-1903), physicien et mathématicien américain né et décédé à New Haven dans le Connecticut (après y avoir passé presque toute son existence en célibataire). Issu d'une famille de lettrés, il poursuit des études de latin et de physique, puis il entreprend une carrière de professeur de physique-mathématique au Yale College. Il séjourne successivement à Paris, à Berlin où il suit les leçons de Heinrich Gustav Magnus et à Heidelberg où il rencontre Gustav Kirchhoff et Herman Ludwig Helmholtz. Il laisse le souvenir d'un savant d'une modestie proverbiale et d'une extraordinaire puissance d'investigation scientifique. Son oeuvre remarquablement compacte fut d'abord peu connue. Aujourd'hui, elle est considérée comme un monument au sein des contributions scientifiques du 19ème siècle. Les deux principales publications datent de 1877 et de 1902. La première s'intitule *On the Equilibrium of Heterogeneous Substances* et est

comparée, en importance, à la chimie pondérale créée par Antoine Laurent Lavoisier. La seconde, jugée plus originale encore, est intitulée *Elementary Principles in Statistical Mechanics*, et est comparée, pour son génie, à la mécanique analytique de Joseph Louis Lagrange. Bien que les exposés de Gibbs se distinguent par une exceptionnelle clarté, et la façon dont l'idée essentielle y est toujours soigneusement dégagée, le premier des deux mémoires n'a guère retenu tout d'abord l'attention des chimistes de son époque, peu accoutumés au langage rigoureux des sciences exactes. La richesse des méthodes thermodynamiques sur lesquelles il s'appuie en a fait cependant une base unifiée de la théorie physico-chimique des états d'équilibre et de leur stabilité. La plupart des lois qui se rapportent à cette discipline, et qui portèrent d'abord d'autres noms, furent redécouvertes ultérieurement au sein de ce premier mémoire. Il en est ainsi, par exemple, de la loi des phases donnant la variance des systèmes en équilibre, longtemps attribuée à Bakkuis Roozeboom (également des lois dites "loi de Van't Hoff" et aussi "loi de Le Chatelier"), relatives aux déplacements d'équilibre à température constante et à pression constante. Il en est encore de même, des critères de stabilité de l'équilibre, dont le théorème de modération dit "théorème de Braun et Le Chatelier". En bref, la plupart des propriétés qui relèvent à présent de la thermodynamique chimique des états d'équilibre, telles que la pression osmotique, l'influence de la tension superficielle, celle des déformations élastiques, la loi relative à l'entropie des mélanges gazeux et le paradoxe de Gibbs associé, ont ce même mémoire pour origine. Gibbs conduit à développer, dans deux communications antérieures à la précédente, un exposé complet des diagrammes et des surfaces thermodynamiques qui contribua largement à la diffusion de leur emploi auprès des praticiens. La théorie de Gibbs utilise pour la première fois la notion d'ensemble ainsi que la distinction entre un ensemble canonique et un ensemble microcanonique de même qu'entre un grand et un petit ensemble. Elle introduit aussi le concept d'espace des phases, caractérisé par les coordonnées et les quantités de mouvement de chaque élément. Elle établit, à partir de l'équation de Liouville, la loi de conservation de l'élément d'extension en phase, ainsi que celle de densité et de probabilité de l'état statistique. Il réalise finalement un accord formel mais remarquable avec les lois macroscopiques de la thermodynamique, régissant le comportement des milieux matériels en équilibre. Les développements actuels de la mécanique statistique constituent encore, sur plus d'un point, des prolongements de la méthode de Gibbs. Il définit pour les réactions chimiques deux quantités très utiles, à savoir l'enthalpie qui représente la chaleur d'une réaction à pression constante, et l'enthalpie libre qui détermine si oui ou non une réaction peut procéder de façon spontanée à température et pression constante. Cette dernière quantité est maintenant nommée énergie de Gibbs en son honneur (ou comme anglicisme énergie libre de Gibbs). L'emploi du point pour désigner un produit scalaire, celui de la croix de Saint-André pour un produit vectoriel et l'adoption des opérateurs vectoriels différentiels del et nabla proviendraient de Gibbs.



Gödel, Kurt (1906-1978) est le mathématicien et logicien austro-américain, qui de tout le 20ème siècle, a le plus révolutionné les fondements logiques des mathématiques. Il était un homme tellement obsédé par la logique qu'on raconte que, alors qu'il cherchait à obtenir sa naturalisation américaine, il osa démontrer devant le juge la contradiction de certains articles de la constitution des États-Unis. Sa thèse, et surtout un article publié en 1931 sous le titre *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme* (sur l'indécidabilité formelle des *Principia Mathematica* et de systèmes équivalents), donneront à Gödel une réputation internationale. Gödel met fin aux espoirs de Hilbert d'axiomatiser totalement la mathématique, et de n'en faire qu'une suite de déductions mécaniques ne laissant aucune place à l'intuition. Ainsi, Gödel montre qu'il existe des propositions vraies sur les nombres entiers, mais que l'on ne sait pas démontrer. Il montre même que, si on ajoute d'autres axiomes, on trouvera toujours des propositions vraies indécidables (qu'on ne sait pas démontrer). Il prouve notamment que l'hypothèse du continu et l'axiome du choix ne sont pas en contradiction avec les autres axiomes de la théorie des ensembles. Puis il s'oriente vers la relativité, étant en relation directe à Princeton avec son ami Einstein. Il est notamment connu des physiciens pour avoir démontré que le voyage vers le passé est possible dans le cadre des équations de la relativité générale.

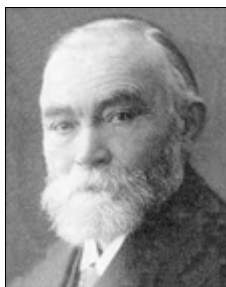


Göpper-Meyer, Maria (1906-1972) est une physicienne américaine d'origine allemande, prix Nobel en 1963, pour son étude de la structure nucléaire. Elle était mariée à un physicien, le spécialiste de la physique du solide Joseph Mayer (1904-1983). Mais, dans ce couple, chacun travaillait de son côté et dans sa spécialité. Goeppert-Mayer obtint son doctorat à l'université de Göttingen, en Allemagne. Elle enseigna dans de nombreuses institutions avant de rentrer à l'université de Californie à San Diego, en 1960. En 1963, elle partagea avec H.D.Jensen et E.Wigner le prix Nobel de physique, et fut citée par le comité Nobel pour son oeuvre indépendante à la fin des années 1940. Elle démontra que le noyau atomique possède un nombre de neutrons et de protons bien définis: elle introduisit un modèle structural du noyau atomique en couches. Ce modèle développé en détail à partir de 1948 supposait que la forte interaction entre le mouvement de rotation intrinsèque (quantifié par le spin) des nucléons et leur mouvement orbital était responsable de la structure des niveaux d'énergie des noyaux. De nombreuses conséquences déduites de cette hypothèse se révélèrent vérifiées par les mesures expérimentales. Quelques années plus tard, James Rainwater, Aage Bohr et Ben R. Mottelson (tous trois Prix Nobel de physique 1975) complétaient la théorie en tenant compte du couplage entre les mouvements des nucléons de la couche externe et le mouvement collectif du coeur nucléaire.



Gosset, William Sealy (1876-1937) connu sous le pseudonyme de "Student" c'est un statisticien anglais. Employé de la brasserie Guinness pour stabiliser le goût de la bière, il a ainsi inventé le test de Student utilisé de manière standard dans de très nombreux domaines de l'industrie ou de l'économie. Il a aussi déterminé en 1908 l'origine de la distribution expérimentale qu'il obtenait dans le cadre de son travail et après avoir suivi un cours de statistique avec Karl Pearson, il obtint son fameux résultat qu'il publia sous le pseudonyme de Student avec la loi qui porte

son nom et son test.



Gottlob, Frege Friedrich Ludwig (1848-1925) Mathématicien et philosophe allemand, initiateur de la logique moderne. Frege est né à Wismar en 1848, et fit ses études aux universités de Iéna et de Göttingen, où il obtint son doctorat de philosophie en 1873. De 1879 à 1917, il fut professeur à la faculté de philosophie d'Iéna. Ses travaux concernent notamment la logique mathématique et ses applications. Confronté à l'ambiguïté du langage ordinaire et à l'imperfection des systèmes logiques disponibles, il inventa de nombreuses notations symboliques, comme les quantificateurs et les variables, posant alors les bases de la logique mathématique moderne. Il est ainsi le premier à avoir présenté une théorie cohérente du calcul des prédicats et du calcul des propositions. Il fut aussi le premier à faire dériver l'arithmétique de la logique. Il définit ainsi notamment la suite des nombres entiers à partir de l'ensemble vide, en appliquant quelques règles simples.

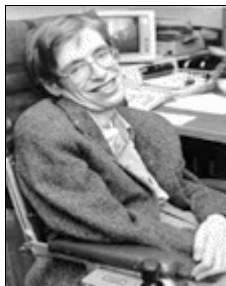


Grothendieck, Alexander (1928-) est né à Berlin d'un père anarchiste russe, tué par les nazis, et d'une mère femme de lettres, réfugiée en France. Il passe sa licence à la faculté des sciences de Montpellier, puis passe une année en 1948-1949 à l'École Normale Supérieure à Paris, avant de migrer en 1949 à l'université de Nancy. Il y devient l'élève, en analyse fonctionnelle, de Schwartz et Dieudonné. Ce dernier le trouve un peu prétentieux, et lui propose de travailler sur des questions que ni Schwartz, ni lui n'ont su résoudre. Voilà ce qu'en dit Schwartz dans son autobiographie: "Dieudonné, avec l'agressivité (toujours passagère), dont il était capable, lui passa un savon mémorable, arguant qu'on ne devait pas travailler de cette manière, en généralisant pour le plaisir de généraliser. [...] L'article s'achevait sur 14 questions, des problèmes que nous n'avions pas su résoudre, Dieudonné et moi. Dieudonné lui [Grothendieck] proposa de réfléchir à certains d'entre eux qu'il choisirait. Nous ne le revîmes plus pendant quelques semaines. Lorsqu'il avait réapparu, il avait trouvé la solution de la moitié d'entre eux!". Rapidement, Grothendieck rédige sa thèse intitulée *Produits tensoriels topologiques et espaces nucléaires*, et devient le spécialiste mondial de la théorie des espaces vectoriels topologiques. Il devient aussi membre du célèbre groupe Bourbaki auprès de ses aînés. Au début des années 1960, il obtient une charge au tout récent Institut des Hautes Études Scientifiques (IHES), et son centre d'intérêt s'oriente vers la géométrie algébrique. Il y réalise des travaux gigantesques, qui lui valent la médaille Fields en 1966. Toutefois, Grothendieck refuse de se rendre en URSS pour la recevoir, afin de protester contre la répression de l'insurrection hongroise en 1956. On la lui remet plus tard, mais il l'offre au Viêt-nam, afin qu'il utilise son or. Il y enseigne d'ailleurs plusieurs semaines sous les bombardements américains. Vers la fin des années 60, Grothendieck, qui a perdu l'habitude de rédiger (Dieudonné a rédigé des années durant son séminaire), devient de moins en moins clair. Il ne pardonnera jamais aux autres mathématiciens de ne pas le comprendre et de dénaturer ainsi ses idées. Si ses relations avec la communauté mathématique n'avaient jamais été faciles (il travaillait énormément en solitaire, ses journées faisaient 27 ou 28 heures, de sorte que parfois il lui arrivait de se décaler - Il méprisait légèrement Dieudonné, séquelle du premier coup de gueule de ce dernier - ses prises de becs

avec Weil causèrent son départ de Bourbaki...), elles sont plus tendues que jamais... Il abandonne peu à peu la mathématique et quitte l'IHES après une dispute interne sur des financements militaires, pour se retirer dans sa maison de l'Hérault, où il se consacre à la méditation et à l'écologie. Il écrit vers 1985 une sorte d'autobiographie, *Récoltes et semailles*, qui ne trouve pas d'éditeur. Ceux qui ont pu la lire sont unanimes pour dire qu'elle contenait de nombreuses attaques contre la communauté des mathématiciens.

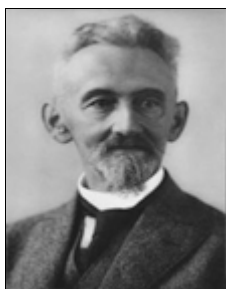


Hamilton, William Rowan (1805-1865) est un mathématicien, physicien et astronome irlandais (né et mort à Dublin) qui fut l'objet de son vivant des plus grands honneurs, on l'appelait le "Lagrange irlandais", et même le "Newton irlandais", et pourtant son oeuvre était peu connue et rarement étudiée. Il est connu pour sa découverte des quaternions, mais il contribua aussi au développement de l'optique, de la dynamique et de l'algèbre. Ses recherches se révélèrent importantes pour le développement de la mécanique quantique. Les travaux mathématiques de Hamilton incluent l'étude de l'optique géométrique, l'adaptation des méthodes dynamiques aux systèmes optiques, l'application des quaternions et des vecteurs aux problèmes de mécanique et géométries, les possibilités de résolution des équations polynomiales, notamment l'équation générale du 5 degré, les opérateurs linéaires, dont il prouve un résultat concernant ces opérateurs dans l'espace des quaternions et qui est un cas spécial du théorème de Cayley-Hamilton. Sa carrière scientifique fut prédestinée par des études à Trinity College, à Dublin, où, à l'âge de 19 ans, il terminait un travail remarquable sur l'optique. À 23 ans, il devint professeur d'astronomie à Dublin et astronome royal à l'observatoire de Dunsink. Il restera toute sa vie fidèle à Dublin et à son observatoire. Hamilton s'efforce de donner aux principes fondamentaux de la mécanique une forme simple permettant d'édifier toute une théorie déductive. Pour cela, il modifie les principes de variations antérieurs, notamment le principe de moindre action, et introduit ce qu'on appelle de nos jours le "principe de Hamilton". Indiquons enfin qu'on lui doit la forme dite "canonique" des équations de la dynamique qui n'apporte rien de nouveau à la physique mais fournit une méthode plus puissante pour résoudre les équations du mouvement. Dans ses travaux des années 1832 à 1835 Hamilton attache une grande importance à l'interprétation géométrique des nombres complexes, et c'est à partir de là qu'il cherche un calcul algébrique qui s'interpréterait dans l'espace à trois dimensions. Il n'arrive à ce but qu'en 1843, en construisant les quaternions. Dans les années qui suivent cette découverte, il se consacre à son développement et à sa diffusion, en lui trouvant des applications à divers domaines des mathématiques et de la physique. Les quaternions de Hamilton constituent un des premiers systèmes de vecteurs et ont, par leurs conséquences théoriques, beaucoup contribué à l'élaboration de l'algèbre et de la physique quantique du 20ème siècle.



Hawking, Stephen (1942-) né à Oxford, est un physicien théoricien et cosmologiste britannique. Au même titre qu'Albert Einstein, Hawking n'aurait pas été particulièrement brillant à la petite école, mais son goût pour les sciences physiques le mène à l'université d'Oxford, un lieu d'après lui d'ennui relatif d'où il sort avec les honneurs. Après avoir obtenu son diplôme B.A. à Oxford en 1962, il est resté pour étudier l'astronomie. Il a décidé d'arrêter quand il trouva que l'étude des taches solaires ne l'attirait pas et qu'il était plus intéressé par la théorie que par l'observation. Il a quitté Oxford, avec les honneurs, pour Trinity Hall où il a participé à l'étude de l'astronomie théorique et la cosmologie théorique.

L'université de Cambridge est un tout autre monde: d'un côté, Hawking y débute son passionnant doctorat sur la relativité générale, de l'autre, sa maladie se déclare. Malgré cette difficulté, l'étude des singularités, concept physique et astronomique récent, permet au chercheur de développer différentes théories, qui le mèneront du Big Bang aux trous noirs. En premier lieu, Roger Penrose et lui construisent la structure mathématique répondant à la question d'une singularité comme origine de l'Univers. Ensuite, à partir des années 1970, Hawking approfondit ses recherches sur les densités infinies locales, et ses études sur les trous noirs ont fait progresser bien d'autres domaines. Enfin, la théorie du tout, visant à unifier les quatre forces physiques, est au centre des recherches actuelles de Hawking. Le but est de démontrer que l'Univers peut être décrit par un modèle mathématique stable, déterminé par les lois physiques connues, en vertu du principe de croissance finie mais non bornée, modèle auquel Hawking a donné beaucoup de crédit. Son handicap lourd ne saurait expliquer à lui seul le grand succès de ses recherches ; Hawking a cherché à vulgariser son travail, et son livre *Une brève histoire du temps* est l'un des plus grands succès de littérature scientifique. En 2001, paraît son deuxième ouvrage, *L'univers dans une coquille de noix* qui vulgarise le dernier état de ses réflexions, en abordant la supergravité et la supersymétrie, la théorie quantique et théorie-M, l'holographie et la dualité, la théorie des supercordes et des p-branes... Il s'interroge également sur la possibilité de voyager dans le temps et sur l'existence d'univers multiples.



Hausdorff, Felix (1868-1942) La renommée du mathématicien allemand Felix Hausdorff repose surtout sur son ouvrage *Grundzüge der Mengenlehre* (1914), qui en fit le fondateur de la topologie et de la théorie des espaces métriques. Né à Breslau dans une famille de marchands aisés, Hausdorff fit ses études secondaires à Leipzig, puis étudia la mathématique et l'astronomie à Leipzig, Fribourg-en-Brisgau et Berlin. En 1891, il obtint son doctorat à Leipzig et y enseigna de 1896 à 1902. Durant toute cette époque, Hausdorff, tout en publiant plusieurs mémoires d'astronomie, d'optique et de mathématiques, s'intéressa surtout à la philosophie, la littérature et l'art. De 1910 à 1935, il était professeur de mathématiques à l'université de Bonn, à l'exception des années 1913-1921, où il enseignait à Greifswald. Depuis sa retraite forcée, en 1935, les travaux de Hausdorff ne furent plus publiés en Allemagne. Juif, Hausdorff risqua le camp de concentration et, lorsqu'en 1942 l'internement devint imminent, il se suicida à Bonn, avec sa femme et sa belle-soeur. Les contributions de Hausdorff au développement des mathématiques se situent dans plusieurs domaines. Son étude approfondie des séries déboucha sur la démonstration de théorèmes sur les méthodes de sommation et les coefficients de Fourier (1921). Considérant les

propriétés d'ensembles numériques, il introduisit une classe importante de mesures. Il a étudié, en théorie générale des ensembles, les ensembles partiellement ordonnés et a obtenu plusieurs théorèmes sur les ensembles ordonnés (1906-1909). En théorie descriptive des ensembles, il a démontré le théorème sur la cardinalité des ensembles boréliens (1916). Outre des résultats isolés mais profonds en topologie et en théorie des ensembles, Hausdorff a surtout, par ses *Grundzüge der Mengenlehre*, posé les fondements d'une discipline.. Hausdorff développe une théorie des espaces topologiques et métriques englobant parfaitement les résultats antérieurs. Il choisit de construire sa théorie des espaces abstraits sur la notion de voisinage. Il ajouta bon nombre de résultats nouveaux à la théorie des espaces métriques, dont le plus profond est le théorème affirmant que chaque espace métrique peut être étendu d'une manière unique à un espace métrique complet. Hausdorff était un professeur méthodique, mais ses cours, au contenu riche et rigoureusement structuré, passèrent au-dessus du niveau de ses auditeurs.



Heaviside, Oliver (1850-1925) est né à la ville de Camden à Londres en Angleterre et mort à Torquay dans le Devon en Angleterre. C'est là qu'il a vécu les 25 dernières années de sa vie. Il est issu d'une famille assez pauvre. Il a attrapé la scarlatine quand il était un enfant en bas âge, ce qui a affecté son audition, il est resté partiellement sourd. Ce qui a eu un impact sur sa vie rendant son enfance difficile surtout au niveau des relations avec les autres enfants. Il a compensé par la timidité et le sarcasme. Cependant, malgré tout, son rendement académique était plutôt élevé. On peut même dire qu'à 16 ans, c'était un étudiant supérieur, mais il a échoué dans la géométrie d'Euclide. Il a détesté devoir déduire un fait d'autres. Le primat de la preuve rigoureuse en arithmétique, idée fortement détestée par Heaviside en fit le sujet où il était le plus faible. Bien qu'ayant interrompu ces études à 16 ans, il a continué à s'instruire par lui-même. Il a appris le code Morse, étudié l'électricité et d'autres langues en particulier le Danois et l'Allemand. Il était autodidacte. En 1868, après avoir quitté ses études, Heaviside est allé au Danemark et il est devenu opérateur de télégraphe. Il a progressé rapidement dans sa profession et il est revenu en Angleterre en 1871. C'est son travail qui l'a incité à étudier l'électricité. Il a donc lu le nouveau traité de Maxwell sur l'électricité et le magnétisme. Après avoir lu ce traité, il a apporté des changements à sa vie. Il a arrêté de travailler et il s'est enfermé dans une chambre de la maison familiale pour travailler sur la théorie de Maxwell. Heaviside a réduit la théorie de Maxwell et c'est à partir de ce moment que la théorie électrique a pris sa forme moderne. Maxwell avait écrit vingt équations à vingt variables. Heaviside réduit ces vingt équations en les remplaçant par quatre équations à deux variables. Aujourd'hui, nous appelons ces équations: "Les quatre équations de Maxwell", oubliant qu'elles sont en fait les équations de Heaviside. Cependant, c'est Hertz qui a obtenu le crédit pour cela, mais il admet que ses idées lui sont venues de Heaviside.



Heisenberg, Werner Karl (1901-1976), né à Wurtzbourg et décédé à Munich était un physicien allemand. Il fut le fondateur des concepts théoriques rigoureux de la mécanique quantique. Il est lauréat du prix Nobel de physique de 1932. Il fréquente le prestigieux Maximiliangymnasium où Max Planck avait étudié 40 ans plus tôt. A l'âge de 12 ans, il se mit à apprendre le calcul intégral et plus tard, passionné par les mathématiques, il suivit en auditeur libre plusieurs cours de l'Université de Munich, notamment sur les méthodes mathématiques de la physique moderne. Il accomplit ses études de physique dans le délai record de 3 ans, et soutint sa thèse (qu'il faillit louper à cause de lacunes en physique expérimentale élémentaire) sous la direction d'Arnold Sommerfeld avec lequel il élaborait une théorie expliquant l'effet Zeeman anormal à l'âge de 20 ans qui attira l'attention des grands physiciens européens (il était considéré aussi brillant que Pauli qui lui-même était déjà considéré comme plus génial qu'Einstein). Dès 1924 il devenait l'assistant de Max Born à Göttingen puis il travailla avec Niels Bohr à Copenhague. C'est au cours des années suivantes qu'avec Max Born et Pascual Jordan, il jeta les bases théoriques de la mécanique quantique. Heisenberg fut recruté en 1927 comme professeur à l'Université de Leipzig âgé seulement de 26 ans. Il fit de cet établissement l'un des hauts-lieux de la physique théorique (et en particulier de la physique nucléaire) en Europe. Il développa la première formalisation de la mécanique quantique, en 1925, en même temps qu'Erwin Schrödinger. Toutefois le formalisme mathématique était différent. Heisenberg adopta une formalisation matricielle complexe (alors qu'il ne savait pas ce qu'était une matrice comme la majorité des physiciens de son temps...) qui faisait naturellement émerger la non commutativité alors que Schrödinger utilisa une approche par les équations différentielles (simple équation d'ordre). Pour cette raison, on crut d'abord que les deux théories étaient distinctes, mais l'année suivante, Schrödinger établit l'équivalence mathématique des deux formulations. Son principe d'incertitude, découvert en 1927, affirme que la détermination de certains couples de valeurs, par exemple la position et la quantité de mouvement, ne peut se faire avec une précision infinie. À partir de 1929, il travailla avec Wolfgang Pauli à l'élaboration de la théorie quantique des champs. Après la découverte du neutron par James Chadwick en 1932, Heisenberg proposa le modèle proton-neutron du noyau atomique, et s'en servit pour expliquer le spin nucléaire des isotopes.



Helmholtz, Hermann Ludwig Ferdinand von (1821-1894) né à Potsdam et mort à Berlin. Il n'est guère de domaines des sciences de la nature auxquels Helmholtz n'ait consacré quelque recherche. On pourrait répéter à son endroit, ce qu'il disait lui-même de Friedrich von Humboldt dans sa célèbre conférence inaugurale du colloque scientifique d'Innsbruck (sur le but et les progrès de la science de la Nature, 1869): "Il avait réussi à dominer toutes les sciences de la nature à son époque et à pénétrer jusqu'en chacune de leurs spécialités." Même si Helmholtz ajoute que dans la seconde moitié du 19ème siècle ce savoir encyclopédique est désormais impossible, et qu'il faut se résigner à besogner dans un secteur étroitement délimité, il suffit de jeter un regard sur l'ensemble de ses travaux pour constater qu'il s'est préoccupé de matières aussi différentes que la thermodynamique, l'hydrodynamique, l'électrodynamique et la théorie de l'électricité, la physique météorologique, la physiologie, et plus particulièrement la

théorie de l'acoustique et l'optique physiologique. Pourvu de dons remarquables pour la vulgarisation des résultats scientifiques les plus récents, il écrivit de nombreux articles et prononça maintes conférences où les exposés scientifiques populaires voisinent avec des préoccupations esthétiques ou philosophiques. Son nom reste surtout attaché à la formulation du principe de la conservation de l'énergie, qui fait de lui l'un des pères de l'énergétique, même si certaines de ses assertions peuvent sembler d'un mécanisme intransigeant et ont pu le faire considérer par certains comme le dernier tenant de la physique galiléenne. Son nom est lié également à quelques inventions notoires comme celle de l'ophtalmoscope ou des résonateurs sphériques. Sur la fin de sa vie, Helmholtz reconnâtra l'importance et l'universalité d'un autre principe physique, le principe de moindre action, qu'il appliquera, en particulier, à l'électrodynamique.



Henry, Joseph (1797-1878) était un scientifique américain né à New-York et décédé à Washington. Ses parents étaient pauvres, et le père d'Henry est mort alors qu'il était encore jeune. Pour le reste de son enfance, Henry a vécu avec sa grand-mère à New York. Il a fréquenté une école qui allait plus tard être nommée Joseph Henry Elementary School, à son honneur. Après l'école, il a travaillé dans un magasin général, et à l'âge de 13 ans, il est devenu apprenti horloger et orfèvre. Son intérêt pour la science s'est éveillé à l'âge de 16 ans par un livre de conférences sur des sujets scientifiques intitulé *Popular Lectures on Experimental Philosophy*. En 1819, il entra à l'Académie Albany, où il a suivi des cours gratuits. Il était si pauvre qu'il devait subvenir à ses besoins avec des postes d'enseignement et de tutorat privé. Il avait l'intention d'étudier la médecine, mais en 1824 il a été nommé assistant ingénieur de l'inspection nationale des routes en cours de construction entre la rivière Hudson et le lac Érié. Dès lors, il a été inspiré d'une carrière soit dans le génie civil ou mécanique. Henry a tellement excellé dans ses études (à tel point, qu'il aidait souvent ses professeurs à enseigner la science) que, en 1826, il est nommé professeur de mathématiques et de philosophie naturelle à l'Académie Albany. Certains de ses travaux de recherche les plus importants ont été réalisés pendant qu'il était à ce poste. Sa curiosité sur le magnétisme terrestre l'a amené à faire des expériences avec le magnétisme en général. Il a été le premier à enrouler un fil isolé étroitement autour d'un noyau de fer dans le but de faire un électro-aimant puissant. Pendant la construction des électro-aimants, Henry a découvert le phénomène électromagnétique d'auto-induction. Il a également découvert l'inductance mutuelle indépendamment de Michael Faraday, mais puisque Faraday a publié ses résultats en premier, il est devenu le découvreur officiellement reconnu du phénomène. En utilisant son principe électromagnétique nouvellement développé, Henry en 1831 a créé l'une des premières machines à utiliser l'électromagnétisme pour le mouvement. Ce fut le premier ancêtre du moteur à courant continu. L'unité SI de l'inductance, la Henry, est nommé en son honneur.



Hermite, Charles (1822-1901), né à Dieuze, il publia ses premiers travaux alors qu'il était encore élève à l'École polytechnique, et à 30 ans, il était déjà considéré comme un des meilleurs mathématiciens de son temps. Il fut successivement professeur à l'École polytechnique, au Collège de France et enfin à la Sorbonne à partir de 1869 où son enseignement et sa volumineuse correspondance eurent une influence considérable. Il vécut à Paris jusqu'à sa mort. Il avait été élu membre de l'Académie des sciences à 34 ans. En algèbre, Hermite prit une part active aux premiers développements de la théorie des invariants, inaugurée par Arthur Cayley et James Joseph Sylvester, il acheva, entre autres, la détermination des invariants des formes binaires du 5 degré, commencée par Sylvester, et découvrit la loi de réciprocité entre covariants de formes binaires de degrés différents. On lui doit aussi un procédé d'interpolation améliorant la méthode de Lagrange en tenant compte des valeurs des dérivées premières, et la découverte de la famille de polynômes orthogonaux qui portent son nom. Les travaux d'analyse d'Hermite portent la marque de son tempérament d'algébriste. Son sujet de prédilection pendant toute sa vie a été la théorie des fonctions elliptiques et des fonctions abéliennes, dont il aimait particulièrement explorer les liens cachés avec l'algèbre et la théorie des nombres. Un de ses résultats qui frappa le plus ses contemporains est la résolution de l'équation du 5 degré à l'aide des fonctions elliptiques. Sa virtuosité dans les calculs des fonctions lui permit d'obtenir directement les remarquables relations sur les nombres de classes d'idéaux des corps quadratiques, que Kronecker avait déduites de la multiplication complexe. Il fut un des pionniers dans l'étude des fonctions abéliennes, où il développa la théorie de la transformation et rencontra à cette occasion pour la première fois le groupe symplectique. Enfin, le plus célèbre des mémoires d'Hermite est celui où, en 1872, il démontra la transcendance du nombre e ; il y avait été conduit par ses recherches sur les fractions continuées algébriques, et sa méthode est restée presque la seule dont on dispose encore aujourd'hui pour aborder les problèmes de transcendance.



Hertz, Heinrich Rudolf (1857-1894), physicien allemand né à Hambourg et décédé à Bonn. Il fit ses études à l'Université de Berlin. En 1879, il est l'élève de Gustav Kirchhoff et Hermann von Helmholtz à l'Institut de physique de Berlin. Il devient maître de conférence à l'université de Kiel en 1883 où il effectue des recherches sur l'électromagnétisme. De 1885 à 1889, à l'origine de la télégraphie sans fil, il fut professeur de physique à l'École technique de Karlsruhe, et, à partir de 1889, enseigna la physique à l'Université de Bonn. Hertz clarifia et étendit la théorie électromagnétique de la lumière proposée par le physicien anglais James Maxwell, en 1884. Il prouva que l'électricité pouvait être transmise par des ondes électromagnétiques qui se déplacent à la vitesse de la lumière et possèdent de nombreuses autres propriétés de la lumière. Ses expérimentations avec ces ondes aboutirent au développement du télégraphe sans fil et de la radio. L'unité de fréquence, une période par seconde, fut dénommée le "Hertz"



Hilbert, David (1862-1943), né à Königsberg et décédé à Göttingen fut un étudiant de Lindemann sous la supervision duquel il obtint sa thèse en 1885 et eut pour camarade Herman Minkowski, avec qui il resta lié par une profonde amitié. Bien que les intérêts mathématiques de Hilbert furent vastes, il préféra travailler à un sujet à la fois. Ses principaux domaines d'intérêts furent: jusqu'en 1892, la théorie algébrique des invariants; de 1892 à 1899 la théorie algébrique des nombres; de 1899 à 1905, le calcul des variations; de 1901 à 1912, les équations intégrales; de 1912 à 1917, les fondements mathématiques de la physique. Vers 1910, Hilbert soutient les efforts d'Emmy Noether, mathématicienne de premier ordre, qui souhaite enseigner à l'université de Göttingen. Pour déjouer le système établi contre les femmes, Hilbert prête son nom à Noether qui peut ainsi annoncer l'horaire de ses cours sans entacher la réputation de l'université. De 1917 jusqu'à la fin de sa vie, il s'occupa de la logique mathématique. Il donna une impulsion décisive à l'essor des recherches sur les fondements des mathématiques. Au congrès international des mathématiques de 1900, Hilbert présenta une liste de vingt-trois problèmes dont plusieurs ne sont pas encore résolus aujourd'hui. Il a adopté et défendu avec vigueur les idées de Georg Cantor en théorie des ensembles et sur les nombres transfinis. Il est aussi connu comme l'un des fondateurs de la théorie de la démonstration, de la logique mathématique et a clairement distingué les mathématiques des métamathématiques. Il est considéré par plusieurs comme le plus grand mathématicien du 20ème siècle.



Hoyle, Fred (1915-2001) Né à Bingley, dans le Yorkshire et décédé à Bournemouth, Hoyle étudie la mathématique et la physique théorique à Cambridge de 1933 à 1939. Lorsque les hostilités éclatent, il s'engage dans la Royal Navy pour travailler au développement du radar au centre de recherche de Witley. Il y rencontre les deux physiciens Hermann Bondi et Thomas Gold. Tous trois passionnés de cosmologie, ils considèrent avec scepticisme le modèle standard de l'Univers de l'époque (d'un point de vue philosophique il leur est inacceptable). À l'époque, le modèle standard achoppait de plus à une difficulté sérieuse: d'après les estimations de Hubble, l'âge de l'Univers devait être d'environ 2 milliards d'années ; or, les données géologiques conduisaient à un âge de la Terre d'au moins 4 milliards d'années. Pendant la guerre, et dans les quelques années qui suivent la fin des hostilités, Hoyle publie plusieurs études sur la théorie de l'accrétion et sur la théorie de la structure stellaire, en particulier pour les étoiles géantes et les naines blanches. La guerre terminée, les trois hommes retournent à Cambridge, où Hoyle obtient une chaire de mathématiques. En 1948, ils exposent leur théorie dans deux articles, l'un de Bondi et Gold, l'autre de Hoyle. En 1963, le premier quasar est découvert. Sa luminosité intrinsèque est très supérieure à celle de tout autre objet céleste connu: il est cent fois plus lumineux que n'importe quelle galaxie! En 1962, Hoyle et William A. Fowler avaient proposé une théorie qui pouvait rendre compte de la luminosité énorme des quasars ; il s'agissait de la théorie des étoiles supermassives. Des considérations théoriques permettent de démontrer que des étoiles normales de masses supérieures à environ 60 masses solaires seraient le siège d'instabilités violentes dues à la pression de radiation et à la génération de l'énergie nucléaire. Cette hypothèse est corroborée par le fait que l'on n'observe pas d'étoiles normales au-delà de la limite d'instabilité. En dépit de cet argument, Hoyle et

Fowler proposaient le concept d'étoile supermassive, étoile qui serait supportée presque entièrement par la pression de radiation. Ainsi, pour atteindre la luminosité caractéristique d'un quasar, l'étoile supermassive doit avoir une masse de l'ordre de 100 millions de masses solaires. Lorsque la densité devient suffisamment élevée, une étoile supermassive de moins de 1 million de masses solaires explose, tandis qu'une étoile plus massive subit un effondrement cataclysmique et forme des trous noirs supermassifs. Ces deux possibilités sont très importantes pour comprendre les quasars, et elles ont été étudiées par de nombreux chercheurs. Une autre explication du phénomène quasar, suggérée pour la première fois par Donald Lynden-Bell, suppose l'accrétion de matière dans un trou noir supermassif situé au centre d'une galaxie (hypothèse actuellement adoptée par consensus de la communauté scientifique).



Huygens, Christian (1629-1695), astronome, mathématicien et physicien néerlandais. Ses découvertes scientifiques nombreuses et originales lui valurent une large reconnaissance et les honneurs parmi les personnalités scientifiques du 17^{ème} siècle. Avec son *Traité de la lumière* (1690), il est à l'origine de la théorie ondulatoire de la lumière (qui plus tard prit son nom): chaque point d'ondes en mouvement est lui-même source de nouvelles ondes. Il se penche très vite, dès 1652 sur les règles exposées par Descartes dans les *Principes de la philosophie*. Prenant appui sur la conservation cartésienne de la quantité de mouvement mv , il a l'idée de résoudre algébriquement le problème du choc en comparant les quantités mv^2 qui ne sont introduites que pour le bien du calcul, sans signification physique particulière. Découvrant alors que ces quantités se conservent avant et après le choc, il peut écrire les règles dans le cas général, ce que Descartes n'avait pu faire incluant donc conservation de la quantité de mouvement et de l'énergie cinétique. En 1655, il inventa une méthode de meulage et de polissage des lentilles d'optique. La définition plus fine ainsi obtenue lui permit de découvrir un satellite de Saturne et de fournir la première description précise des anneaux de Saturne. La nécessité de disposer d'une mesure exacte du temps pour l'observation du ciel l'amena à appliquer les lois du pendule composé pour régler les mouvements des horloges et montres. En 1656, il conçut une lunette de télescope qui porte son nom. Entre 1658 et 1659, Huygens travaille à la théorie du pendule oscillant. Il a en effet l'idée de réguler des horloges au moyen d'un pendule, afin de rendre la mesure du temps plus précise. Il découvre la formule de l'isochronisme rigoureux en 1659: lorsque l'extrémité du pendule parcourt un arc de cycloïde, la période d'oscillation est constante quelle que soit l'amplitude. Dans *Horologium oscillatorium* (1673), il détermina la véritable relation existant entre la longueur d'un pendule et la durée d'oscillation, et présenta ses théories sur la force centrifuge des mouvements circulaires, qui aidèrent le physicien anglais Isaac Newton à formuler les lois de la gravité. En 1673, Huygens et son jeune assistant Denis Papin, mettent en évidence le principe des moteurs à combustion interne, qui conduiront au xix^{ème} siècle à l'invention de l'automobile. En 1678, il découvrit la polarisation de la lumière par double réfraction sur la calcite.



Ibn, Al Haytham (965-1039) est un mathématicien, un philosophe et un physicien Arabe. Il est l'un des pères de la physique quantitative et de l'optique moderne, le pionnier de la méthode scientifique moderne et le fondateur de la physique expérimentale et certains, pour ces raisons, l'ont décrit comme le premier scientifique. Al Haytham commença sa carrière de scientifique dans sa ville natale de Bassorah. Il fut cependant convoqué par le calife Hakim qui voulait maîtriser les inondations du Nil qui frappaient l'Égypte année après année. Après avoir mené une expédition en plein désert pour remonter jusqu'à la source du fameux fleuve, Alhazen se rendit compte que ce projet était pratiquement impossible. De retour au Caire, il craignait que le calife qui était furieux de son échec ne se vengeât et décida donc de feindre la folie. Le calife se borna à l'assigner à résidence. Alhazen profita de ce loisir forcé pour écrire plusieurs livres sur des sujets variés comme l'astronomie, la médecine, la mathématique, la méthode scientifique et l'optique. Le nombre exact de ses écrits n'est pas connu avec certitude mais on parle d'un nombre entre 80 et 200. Peu de ces ouvrages, en effet, ont survécu jusqu'à nos jours. Quelques-uns d'entre eux, ceux sur la cosmologie et ses traités sur l'optique notamment, n'ont survécu que grâce à leur traduction latine. La plupart de ses recherches concernaient l'optique géométrique et physiologique. Contrairement à une croyance populaire, il a été le premier à expliquer pourquoi le soleil et la lune semblent plus gros (on a cru longtemps que c'était Ptolémée), il établit aussi que la lumière de la lune vient du soleil. C'est aussi lui qui a contredit Ptolémée sur le fait que l'oeil émettrait de la lumière. Selon lui, si l'oeil était conçu de cette façon on pourrait voir la nuit. Il a compris que la lumière du soleil était diffusée par les objets et ensuite entrainée dans l'oeil. En astronomie il a tenté de mesurer la hauteur de l'atmosphère et a trouvé que le phénomène du crépuscule est dû à un phénomène de réfraction. Il parla également de l'attraction des masses et on croit qu'il connaissait l'accélération gravitationnelle. Alhazen a devancé de quelques siècles plusieurs découvertes faites par des scientifiques occidentaux pendant la Renaissance. Il fut un des premiers à se servir d'une méthode d'analyse scientifique et influença grandement des scientifiques comme Roger Bacon et Kepler.



Jacobi, Carl (1804-1851) né à Potsdam et décédé à Berlin fut, avec N. H. Abel, le fondateur de la théorie des fonctions elliptiques dont il donna de nombreuses applications aux branches les plus diverses des mathématiques. On lui doit également des exposés de mécanique théorique où il reprend les résultats de W. R. Hamilton, et des applications de la théorie des équations différentielles à la dynamique. À son entrée au gymnase, en 1816, Jacobi avait déjà achevé le cycle des études secondaires et, assez réfractaire à l'enseignement traditionnel, il étudia directement les oeuvres des grands mathématiciens, particulièrement celles d'Euler et de Lagrange. Inscrit en 1821 à l'université de Berlin, il y apprit la philologie et la mathématique, à laquelle il se consacra bientôt uniquement. En 1825 il était docteur en philosophie avec une thèse où il démontrait ou généralisait certaines formules de Lagrange. Il enseigna à Berlin pendant une année environ, puis à Königsberg où il fut transféré par décision ministérielle. Fin 1827, il fut nommé professeur extraordinaire à l'université de cette ville où il entra en contact avec l'astronome Friedrich Wilhelm Bessel. Pensionné par le gouvernement de Prusse, il fut, après un voyage en Italie, en 1843, nommé académicien à Berlin,

dispensé de tout enseignement mais autorisé à traiter, à l'Université, tout sujet qui lui conviendrait. Présenté comme candidat aux élections de mai 1848, il fut persécuté un temps pour ses opinions libérales. Jacobi consacra de nombreux travaux à la transformation des intégrales et apporta une contribution essentielle à la théorie des équations différentielles et des équations aux dérivées partielles. C'est à cela que se rattachent ses apports au calcul des variations, à la dynamique des solides et à la mécanique céleste –problème des trois corps, perturbations des mouvements planétaires. L'algèbre lui doit d'importantes recherches sur les formes quadratiques et une exposition devenue classique de la théorie des déterminants, prélude au mémoire sur les déterminants fonctionnels appelés de nos jours "jacobiens". Il perfectionne la théorie de l'élimination et enseigne à représenter les racines d'une équation algébrique par des intégrales définies ou par des séries. Il étudie les points communs aux courbes et aux surfaces algébriques, et trouve directement le nombre des tangentes doubles d'une courbe plane, établi déjà par J. Plücker en utilisant la dualité.



Jordan, Camille (1838-1921) né à Lyon et décédé à Paris fut le spécialiste indiscuté de la théorie des groupes pendant toute la fin du 19ème siècle et on lui doit de très nombreux résultats, tant sur les groupes finis que sur les groupes dits classiques, dont il fut le premier à mesurer toute l'importance. Ses cours d'analyse contribuèrent au développement de la théorie des fonctions de variable réelle. En 1855, à 17 ans, il est reçu premier à l'École polytechnique et sort de l'École des Mines en 1861. Il sera, du moins en titre, ingénieur chargé de la surveillance des carrières de Paris jusqu'en 1885, ce qui n'empêchera pas une intense activité de recherche mathématique. Nommé examinateur à l'École polytechnique en 1873, puis professeur en 1876, il entre à l'Académie des sciences en 1881 puis succède à Joseph Liouville au Collège de France deux années plus tard. De 1885 à 1921, il assume la direction du *Journal de mathématiques pures et appliquées* fondé par Liouville. Malgré les efforts de Liouville, l'oeuvre d'Évariste Galois était restée à peu près totalement inconnue du monde des mathématiques (seul Leopold Kronecker avait utilisé certains de ses résultats), et c'est à Jordan, avec son *Traité des substitutions et des équations algébriques*, publié à Paris en 1870, que l'on doit le premier exposé systématique de théorie des groupes, enrichi de dix années de recherches personnelles. Il s'y limite aux groupes finis, plus précisément aux groupes de permutations, et introduit de nombreux concepts nouveaux. Dans des mémoires ultérieurs, Jordan étudie en détail, essentiellement du point de vue des facteurs de composition, le groupe linéaire et les groupes orthogonaux et symplectiques sur un corps premier. Les études de Jordan sur le groupe linéaire font intervenir des considérations sur la réduction des matrices, et, en particulier, la forme dite "forme de Jordan". Indiquons enfin les efforts de Jordan pour déterminer tous les groupes résolubles finis en réponse au problème, posé par Niels Henrik Abel, de rechercher toutes les équations de degré donné résolubles par radicaux. En plus des résultats donnés ci-dessus relatifs au groupe linéaire, on doit à Jordan un exposé complet de la géométrie euclidienne réelle à n dimensions par des méthodes entièrement analytiques. L'enseignement de Jordan à l'École polytechnique, puis au Collège de France, l'amène à préciser de nombreuses notions de la théorie des fonctions de variable réelle et son *Cours d'analyse de l'École polytechnique*, dont la première édition date de 1880, contribuera à

former des générations de mathématiciens. On lui doit aussi la notion de fonction à variation bornée, qui lui permet de donner une définition correcte de la longueur d'une courbe et d'obtenir sous sa forme générale le théorème de convergence des séries de Fourier ; mais le résultat le plus célèbre est celui qui affirme qu'une courbe fermée simple (dite, de nos jours, "courbe de Jordan") sépare le plan en deux régions. Signalons enfin, pour terminer, que Jordan, précurseur de Poincaré, a écrit plusieurs mémoires d'*Analysis situs*, c'est-à-dire sur la topologie combinatoire. On lui doit une démonstration, devenue classique, du théorème d'Euler sur les polyèdres et le fait que deux surfaces de même genre sont applicables l'une sur l'autre (ce qui, comme l'a montré Poincaré, n'est pas vrai en général pour les hypersurfaces).



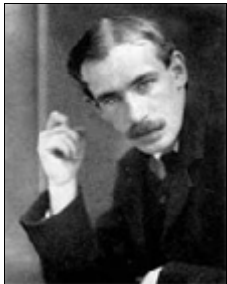
Jordan, Pascual (1902 - 1980) né à Hanovre et décédé à Hamburg est un physicien théoricien allemand, professeur à l'université de Göttingen. Jordan passa dès 1921 une partie de ses études à l'université technique de Hanovre où il étudia un mélange de zoologie, de mathématique et physique. En 1923 il se spécialisa lors de son entrée à l'université de Göttingen, qui était alors à son zénith tant du point de vue de la mathématique que de la physique. À Göttingen, Jordan devint assistant de Richard Courant et surtout de Max Born qui l'influença grandement. Il contribua ainsi de façon décisive à la fondation de la mécanique quantique et de la théorie quantique des champs. En raison de son affiliation au parti nazi, il fut cependant isolé de la communauté physicienne. En 1925, avec Max Born, Jordan écrit la relation de commutation canonique entre la quantité de mouvement et la position. Dans ce même article, il propose également l'idée qu'il faut aussi quantifier le champ électromagnétique, ouvrant la voie à la théorie quantique des champs. En 1925 également, avec Max Born et Werner Heisenberg, Jordan développe la formulation matricielle de Heisenberg de la mécanique quantique. Ils introduisent les transformations canoniques, la théorie des perturbations, le traitement des systèmes dégénérés, et enfin la fameuse relation de commutation canonique des composantes du moment cinétique.



Joule, James Prescott (1818-1889), physicien britannique, né à Salford, dans le Lancashire et décédé à Sale. Il fut l'un des plus grands physiciens de son époque. Joule est célèbre pour ses travaux de recherche en électricité et en thermodynamique. Au cours de ses recherches sur la chaleur émise dans un circuit électrique, il formula la loi, connue sous le nom de loi de Joule, sur la chaleur électrique, qui indique que la quantité de chaleur produite chaque seconde dans un conducteur par le passage du courant électrique est proportionnelle à la résistance électrique du conducteur et au carré du courant électrique. Joule a vérifié expérimentalement la loi de la conservation de l'énergie dans son étude sur la transformation de l'énergie mécanique en énergie thermique (relation entre joules et calories: il faut 1 calorie, soit 4.18 joules pour augmenter 1 gramme d'eau de 1 degré).



Kepler, Johannes (1571-1630), astronome et physicien allemand, célèbre pour sa formulation et sa vérification des trois lois du mouvement planétaire. Ces lois sont maintenant connues sous le nom de "**lois de Kepler**". Son principal traité contient les formulations de deux des lois du mouvement planétaire. La première stipule que les planètes se déplacent selon des orbites elliptiques avec le Soleil comme foyer et la seconde, ou "**loi des aires**", énonce que la ligne imaginaire que l'on tracerait entre le Soleil et une planète balaie des aires identiques d'une ellipse pendant des intervalles de temps égaux; en d'autres termes, plus la planète se rapproche du Soleil, plus elle se déplace rapidement. Un autre traité contient une autre découverte sur le mouvement planétaire: le cube de la distance entre une planète et le Soleil divisé par la période orbitale de cette planète au carré est une constante et est la même pour toutes les planètes. Le mathématicien et physicien anglais Isaac Newton se reposa fortement sur les théories et les observations de Kepler pour formuler sa théorie de la force gravitationnelle. Kepler apporta également sa contribution dans le domaine de l'optique et développa en mathématiques un système infinitésimal qui fut le précurseur du calcul infinitésimal.



Keynes, John Maynard (1883-1946) était un économiste britannique. Il est le fondateur du "**keynésianisme**", doctrine économique qui encourage l'intervention de l'État au sein de l'économie, pour assurer le plein-emploi. Keynes est né dans une famille d'universitaires. À 7 ans, il entra à Perse School. Deux ans plus tard, il entra en classe préparatoire à St Faith's. Avec les années, il se montra très prometteur et en 1894, il termina premier de sa classe et reçut un prix pour la première fois en mathématiques. Un an plus tard, il intègre l'Eton College où il brille et gagne, en 1899 et en 1900, le prix de mathématiques. En 1901, il finit premier en mathématiques, histoire et anglais. En 1902, il gagne sa place pour le King's college de Cambridge. Keynes est sans aucun doute une importante figure de l'histoire de la science économique qu'il révolutionna avec son oeuvre principale, la *Théorie générale de l'emploi, de l'intérêt et de la monnaie* parue en 1936. L'ouvrage est considéré comme le traité de science sociale le plus influent du 20ème siècle dans la mesure où il a rapidement et continuellement modifié la façon dont le monde a considéré l'économie et le rôle du pouvoir politique dans la société. Certains estiment qu'aucun autre ouvrage n'a eu une telle importance depuis en Europe, bien que l'ouvrage de Friedrich Hayek qui lui valut son Prix Nobel, *The Road to Serfdom*, fasse la démonstration fulgurante des limites de la théorie keynésienne. Avec la *Théorie générale*, Keynes a développé une théorie qui pouvait expliquer le niveau de la production et par conséquent de l'emploi ; le facteur déterminant étant la demande. Parmi les concepts révolutionnaires apportés par Keynes, on retiendra surtout: ceux de l'équilibre de sous-emploi où le chômage est possible pour un niveau donné de la demande effective, l'absence de mécanisme de régulation par les prix afin de résorber le chômage, une théorie de la monnaie fondée sur la préférence pour la liquidité, l'introduction de l'incertain et des prévisions, la notion d'efficacité marginale de l'investissement brisant la loi de Say (et renversant donc le lien de causalité épargne-investissement). Ces concepts accréditent la possibilité de politiques interventionnistes pour éliminer les récessions et freiner les surchauffes économiques. L'ensemble de ces concepts constitue ce qu'on appelle aujourd'hui la macroéconomie.



Kirchhoff, Gustav Robert (1824-1887) né à Königsberg (aujourd'hui Kaliningrad en Russie) et décédé à Berlin, Kirchhoff étudie la physique-mathématique auprès de Franz Neumann. Après un doctorat en 1847, il devient conférencier à l'université de Berlin avant d'obtenir, en 1850, le poste de professeur de physique extraordinaire à l'université de Breslau. C'est là qu'il fait la connaissance du chimiste Robert Wilhelm Bunsen, avec qui il sera amené à travailler de nombreuses années. Leur collaboration se poursuivra en effet au-delà de 1854, date à laquelle Kirchhoff est nommé à professeur de physique à l'université de Heidelberg. Élu vice-recteur de cette même université en 1865, il finit par accepter une chaire de physique théorique à Berlin en 1875. Kirchhoff est encore étudiant lorsqu'il commence à s'intéresser aux problèmes liés à l'électricité. En 1845, il établit la notion de potentiel électrique et énonce les lois de réseaux qui portent son nom (loi des noeuds, loi des mailles). Il généralise la loi d'Ohm sur le courant électrique à des conducteurs à trois dimensions et, plus tard, montre que le passage du courant à travers un conducteur se fait à la vitesse de la lumière. Sa rencontre avec Bunsen aboutit à la naissance de la spectroscopie. Ensemble, les deux chercheurs découvrent le caractère spécifique du spectre de la lumière émise par chaque corps chimique. Grâce à ce nouvel outil d'analyse, ils dépistent deux éléments encore inconnus: le césium (1860) et le rubidium (1861). La mise au point du spectroscope à prisme, pour analyser la lumière de substances en combustion, permet également à Kirchhoff d'établir la loi du rayonnement: le rapport des pouvoirs d'émission et d'absorption d'un corps, indépendant des propriétés de ce corps, est fonction de la température et de la longueur d'onde. Le pouvoir d'émission est ainsi proportionnel à celui du "corps noir", défini par Kirchhoff (1862) comme le corps parfaitement absorbant. Cette loi, qui explique notamment la présence des raies sombres d'absorption (dites "raies de Fraunhofer") dans le spectre de rayonnement solaire, marque le début d'une nouvelle ère en astrophysique et annonce l'avènement de la théorie des quanta de Planck.



Klein, Félix (1849-1925) fit ses études à Bonn, à Göttingen et à Berlin. En 1872, il devint professeur de mathématiques à l'université d'Erlangen, où son cours inaugural fut l'énoncé des grandes lignes de son fameux programme d'Erlangen. Il enseigna ensuite à Munich (1875-1880), puis à l'université de Leipzig (1880-1886) et enfin à Göttingen (1886-1913). À partir de 1872, il édita les *Mathematische Annalen* de Göttingen et fonda, en 1895, la grande *Encyclopédie mathématique*, dont il supervisa la rédaction jusqu'à sa mort, à Göttingen. Il fut le chef incontesté de l'école mathématique allemande, et son influence fut très grande (il donna de nombreuses conférences à l'étranger, dont les États-Unis), notamment sur le développement de la géométrie, grâce à son programme d'Erlangen. Avec le texte, publié dans son ouvrage *Gesammelte mathematische Abhandlungen* (1921-1923), Klein donne une définition de la géométrie englobant aussi bien la géométrie classique (c'est-à-dire euclidienne) que la géométrie projective, les géométries non euclidiennes, etc., mettant fin aux controverses stériles entre partisans de la géométrie synthétique et ceux de la géométrie analytique. Pour lui, une géométrie est l'étude des propriétés invariantes par un groupe donné de transformations: ainsi les théorèmes de géométrie classique sont

l'expression de relations entre invariants du groupe des similitudes ; ceux de la géométrie projective entre covariants du groupe projectif. On doit aussi à Klein d'importants travaux sur l'équation différentielle hypergéométrique, sur les fonctions abéliennes, sur le groupe de l'icosaèdre régulier (*Lectures on the Icosahedron*, 1914), sur les fonctions elliptiques, à partir desquelles il dégage la notion de fonction modulaire (*Vorlesungen über die Theorie der automorphen Funktionen*, 1897-1902).

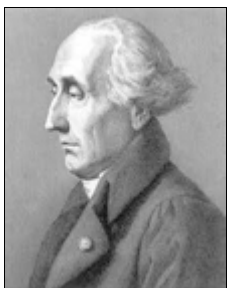


Kolmogorov, Andreï (1903-1987) est un mathématicien russe dont les apports en mathématiques sont considérables. Kolmogorov est né à Tambov. Sa mère célibataire mourut à sa naissance et il fut élevé par sa tante avec les économies de son grand-père. On pense que son père fut tué lors de la Guerre civile russe. Kolmogorov fut scolarisé à l'école du village de sa tante, et ses premiers efforts littéraires et articles mathématiques furent imprimés dans le journal de l'école. Adolescent, il conçut des machines à mouvement perpétuel, cachant tellement bien leurs défauts intrinsèques que ses professeurs d'enseignement secondaire n'arrivaient pas à les découvrir. En 1910, sa tante l'adopta et ils déménagèrent à Moscou, où il intégra un Gymnasium et y obtint son diplôme en 1920. Après avoir terminé ses études secondaires, il suit les cours à l'Université de Moscou et à l'institut Mendeleïev. Il étudie non seulement les mathématiques, mais aussi l'histoire russe et la métallurgie. En 1922, Kolmogorov publie ses premiers résultats concernant la théorie des ensembles et, en 1923 il publie ses travaux sur la théorie de l'intégration, sur l'analyse de Fourier et pour la première fois sur la théorie des probabilités et commence à devenir connu à l'étranger. Après la fin de ses études supérieures en 1925, il commence son doctorat auprès de Nikolaï Louzine, qu'il termine en 1929. En 1931, il reçoit une chaire de professeur à l'Université de Moscou. En 1933, paraît en allemand son manuel des *Fondements de la théorie des probabilités* dans lequel il présente son axiomatisation du calcul des probabilités ainsi qu'une manière adaptée à traiter les processus stochastiques. La même année, il devient directeur de l'Institut de mathématiques de l'Université de Moscou. En 1934, il publie son travail sur la cohomologie et obtient, grâce à cette thèse, le titre de docteur en mathématique et en physique. Il obtient des récompenses des autorités soviétiques, comme l'Ordre de la science socialiste (1940), le prix Staline (1941) et plusieurs fois le prix Lénine. En 1941, il élabore une théorie fameuse de la turbulence des fluides. En 1953 et 1954, il décrit la théorie KAM (Kolmogorov-Arnold-Moser) de stabilité des systèmes dynamiques (un système mécanique complexe exactement soluble est stable si on le perturbe seulement un tout petit peu). Il introduit également la notion d'entropie métrique pour les systèmes dynamiques mesurés. En 1955, il devient docteur honoris causa de la Sorbonne. En 1962, il reçoit le Prix Balzan pour les mathématiques.



Kronecker, Leopold (1823-1891) était un mathématicien allemand qui nous apparaît comme l'un des plus grands arithméticiens du 19^{ème} siècle et l'un des fondateurs de la grande théorie des nombres algébriques. Ses travaux sur le corps de classes dans un cas particulier ont préparé ceux de Hilbert. Né à Liegnitz, dans une famille de riches commerçants, Kronecker suivit au gymnase les cours d'Ernst Kummer, qu'il devait retrouver plus tard comme professeur à l'université de Breslau, puis comme collègue à Berlin, et qui, avec Peter Gustav Lejeune-

Dirichlet, devait avoir l'influence la plus profonde sur le développement de sa pensée. Après avoir soutenu, en 1845, une thèse très originale sur les unités des corps cyclotomiques, il s'occupa pendant plusieurs années des affaires familiales, et ne put se livrer entièrement de nouveau aux recherches mathématiques qu'à partir de 1853. Élu, en 1860, membre de l'Académie des sciences de Berlin, il donna, à partir de cette époque, des cours libres à cette université, où il fut nommé professeur titulaire en 1883 et où il acheva sa vie. Bien que maniant avec virtuosité toutes les ressources de l'analyse (comme le montrent ses travaux sur les fonctions elliptiques, les séries de Dirichlet ou encore sa formule intégrale donnant le nombre des racines d'un système d'équations dans un espace à n dimensions), Kronecker est avant tout un algébriste et un arithméticien. Vers la fin de sa vie, il professait une doctrine tendant à rejeter l'infini actuel des mathématiques en ne gardant comme valable que ce qui pouvait être uniquement fondé sur les nombres entiers (ses polémiques avec Cantor sont restées célèbres). En algèbre, Kronecker fut l'un des animateurs les plus actifs du groupe de mathématiciens qui, dans les années 1860-1890, achevèrent de mettre sur pied l'algèbre linéaire et multilinéaire inaugurée par Arthur Cayley et Hermann Grassmann aux alentours de 1845. C'est ainsi qu'il reprit et compléta les travaux de Karl Weierstrass et fut l'un des premiers à comprendre et à utiliser les travaux d'Évariste Galois (publiés en 1846).

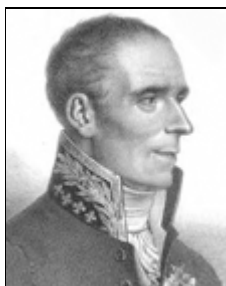


Lagrange, Joseph Louis de (1736-1813), né à Turin et décédé à Paris, était comme l'un des plus grands mathématicien et astronome du 18^{ème} siècle. Élève brillant issu d'un milieu aisé, il étudie au collège de Turin. Il prend goût pour les mathématiques par hasard à l'âge de 17 ans après la lecture d'un mémoire de Edmund Halley portant sur les applications de l'algèbre en optique. Le sujet l'intéresse au plus haut point. Dès lors, il se passionne pour les mathématiques qu'il étudie seul et assidûment. Il devient rapidement un mathématicien confirmé et ses premiers résultats ne se font pas attendre. Dans une lettre adressée à Leonhard Euler il jette les bases du calcul variationnel. Cet échange est le début d'une longue correspondance entre les deux hommes. Lagrange a alors 19 ans et enseigne à l'école d'artillerie de Turin où il fut nommé en 1755. Il fonde en 1758 l'Académie des Sciences de Turin qui publiera ses premiers résultats sur l'application du calcul variationnel à des problèmes de mécanique (propagation du son, corde vibrante...). En 1764, ses travaux sur les librations de la Lune (petites variations de son orbite) sont récompensés par le Grand Prix de l'Académie des Sciences de Paris. Il introduisit de nouvelles méthodes pour le calcul des variations et pour l'étude des équations différentielles, qui lui permirent de donner un exposé systématique de la mécanique dans son célèbre ouvrage *Mécanique analytique* (1788). Il travailla sur la théorie additive des nombres. On lui doit le théorème sur la décomposition d'un entier en quatre carrés. Dans l'étude des équations algébriques, il introduisit des concepts qui conduiront à la théorie des groupes développée plus tard par Abel et Galois. En physique, en précisant le principe de moindre action, avec le calcul des variations, vers 1756, il invente la fonction de Lagrange, qui vérifie les équations de Lagrange, puis développe la mécanique analytique, vers 1788, pour laquelle il introduit les multiplicateurs de Lagrange. Il entreprend aussi des recherches importantes sur le problème des trois corps en astronomie, un de ses résultats étant la mise en évidence des points de

libration (dits points de Lagrange) (1772).



Langevin, Paul (1872-1946) Physicien français né et décédé à Paris. Très jeune, Langevin manifeste des dons exceptionnels. Encouragé par ses instituteurs, il parcourt rapidement les divers échelons de l'enseignement obligatoire avant d'entrer à 16 ans à l'École supérieure de physique et de chimie industrielle de la Ville de Paris. Langevin y suit les cours et l'enseignement de laboratoire de Pierre Curie, avec lequel il se lie d'amitié. À sa sortie de l'École, il renonce à la carrière d'ingénieur et décide, sur les conseils de Pierre Curie, de se consacrer à la recherche et à l'enseignement. Aussi, se présente-t-il à l'École Normale Supérieure où il est reçu premier en 1894. En 1897, il bénéficie d'une bourse pour aller travailler un an au Cavendish Laboratory de Cambridge, haut lieu de la science européenne où se trouvent alors E. Rutherford et J. J. Thomson. De retour en France, il soutient sa thèse en 1902, est nommé professeur suppléant, puis professeur au Collège de France. En 1904, il succède à Pierre Curie à l'École de physique et de chimie, dont il devient directeur en 1925. L'oeuvre de Langevin se situe dans cette longue période de transition qui, de 1900 à 1930, mène de la physique classique à la physique moderne, dominée par la théorie de la relativité et la théorie quantique. Ses premiers travaux (sur l'ionisation des gaz) l'amènent à élaborer en 1905 son modèle théorique majeur, qui devrait par la suite servir de base à de nombreuses autres explications des propriétés macroscopiques de la matière, dans lequel les électrons à l'intérieur des atomes décrivent des orbites fermées, conférant ainsi aux atomes des propriétés analogues à celles de petits aimants. En 1906 il aboutit au résultat étonnant selon lequel l'inertie serait une propriété de l'énergie..., du moins dans le cas de l'électron. Ce n'est que quelques mois plus tard qu'il lira le mémoire d'Einstein sur la théorie de la relativité restreinte à laquelle il va consacrer son enseignement dans ses cours au Collège de France. Langevin est aussi à l'origine des fameux congrès Solvay qui, dès 1911, réunirent périodiquement tous les grands noms de la physique, et où furent largement discutés les concepts de la théorie quantique. C'est d'ailleurs grâce à Langevin que les travaux de son élève Louis de Broglie sur la mécanique ondulatoire connurent la diffusion qu'ils méritaient: d'abord étonné, Langevin fut très vite convaincu de la justesse des idées de De Broglie et inscrivit immédiatement la nouvelle mécanique ondulatoire au programme de son cours au Collège de France. Fidèle à l'idéal de clarté pédagogique qui fut toujours le sien, Langevin a par ailleurs effectué, sur les concepts encore en gestation de la théorie quantique, un travail d'analyse et de refonte épistémologiques dont on mesure aujourd'hui l'importance.



Laplace, Pierre Simon (1749-1827) Né à Beaumont-en-Auge et décédé à Paris, fils de cultivateur, Laplace s'initia aux mathématiques à l'École militaire de cette petite ville. Il y commença son enseignement. Il doit cette éducation à ses voisins aisés qui avaient détecté son intelligence exceptionnelle. À 18 ans, il arrive à Paris avec une lettre de recommandation pour rencontrer le mathématicien d'Alembert, mais ce dernier refuse de rencontrer l'inconnu. Mais Laplace insiste: il envoie à d'Alembert un article qu'il a écrit sur la mécanique classique. D'Alembert en est si impressionné qu'il est tout heureux de patronner Laplace. Il lui obtient un poste d'enseignement en mathématiques. L'oeuvre la plus importante de Laplace

concerne le calcul des probabilités, les équations différentielles (laplacien) et la mécanique céleste. Il établit aussi, grâce à ses travaux avec Lavoisier entre 1782 et 1784 la relation des transformations adiabatiques d'un gaz, ainsi que deux lois fondamentales de l'électromagnétisme. En mécanique, c'est avec le mathématicien Joseph-Louis de Lagrange, que Laplace résume ses travaux et réunit ceux de Newton, Halley, Clairaut, d'Alembert et Euler, concernant la gravitation universelle (particulièrement le problème de stabilité du système solaire), dans les 5 volumes de sa *Mécanique céleste* (1798-1825). On rapporte (mais c'est très probablement une légende) que, feuilletant la *Mécanique céleste*, Napoléon fit remarquer à Laplace qu'il n'y était nulle part fait mention de Dieu. "Je n'ai pas eu besoin de cette hypothèse", rétorqua le savant qui n'était par ailleurs pas modeste (se considérant - probablement à juste titre - comme le meilleur mathématicien de sa génération). Il aussi un des premiers scientifiques à concevoir l'existence des trous noirs et la notion de collapsus gravitationnel.



Laurent, Pierre Alphonse (1813-1854) est un mathématicien français né à Paris et qui s'est rendu célèbre pour la découverte de la Série de Laurent dans le domaine de l'analyse complexe et qui a un grand impact dans le calcul de certaines intégrales en physique. Il est entré à l'École Polytechnique de Paris en 1830. Laurent a été diplômé en 1832 comme un des meilleurs élèves de l'année et entra dans le corps d'ingénierie comme lieutenant. Pendant la gestion de ses projets de développement du port du Havre, Laurent écrivait sa première publication mathématique sur les séries de Laurent. Cette recherche était contenue dans un mémoire soumis au Grand prix de l'Académie des sciences en 1843, mais sa candidature étant trop tardive, l'article n'a jamais été inscrit au prix. Cependant, Cauchy fit une référence dans ses travaux au papier de Laurent 3 mois plus tard. Le même problème se réitéra pour une autre publication importante de Laurent quelques mois plus tard. Après ces événements, Laurent, déçu changea de domaine de recherche pour se concentrer sur la physique (mathématique appliquée). Cauchy lui propose un poste vacant à l'Académie des Sciences en 1846, mais sa candidature ne fut pas retenue. Laurent est décédé à Paris, à l'âge de 41 ans. Ses écrits n'ont été publiés qu'après son décès.



Lavoisier, Antoine Laurent de (1743-1794), chimiste français, considéré comme le fondateur de la chimie moderne. Lavoisier naquit à Paris et fit ses études au collège Mazarin. Il fut élu membre de l'Académie des sciences en 1768. Il occupa de nombreux postes, y compris celui de directeur des Poudreries nationales en 1776, de membre de la Commission pour l'établissement du nouveau système de poids et mesures en 1790 et de secrétaire de la Trésorerie en 1791. Il tenta d'introduire des réformes dans le système monétaire et fiscal français, ainsi que dans le système agricole. Lavoisier fut l'un des premiers à réaliser des expériences chimiques réellement quantitatives. Il montra qu'en dépit du changement d'état de la matière au cours d'une réaction chimique, la quantité de matière restait constante entre le début et la fin de chaque réaction. Ces expérimentations ont fourni des preuves en faveur de la loi de la conservation de la matière. Lavoisier fit également des recherches sur la composition de l'eau, dont il appela les composants oxygène et hydrogène. L'une des plus importantes expériences de Lavoisier concerna la nature de la combustion (ou brûlage). Il démontra ainsi que

le processus de combustion implique la présence d'oxygène. Il démontra également le rôle de l'oxygène dans la respiration chez les animaux et chez les végétaux. Les explications de Lavoisier sur la combustion remplacèrent la doctrine du phlogistique. Celle-ci postulait en effet qu'une substance se dégageait, le "phlogiston", lorsque la matière se consumait. Étant l'un des vingt-huit fermiers généraux, Lavoisier est bêtement stigmatisé comme traître par les révolutionnaires en 1794 et guillotiné lors de la Terreur à Paris le 8 mai 1794, à l'âge de 50 ans, en même temps que l'ensemble de ses collègues.

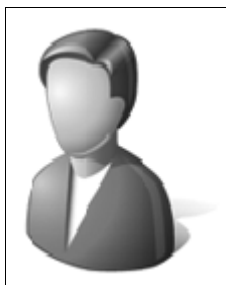


Lebesgue, Henri Léon (1875-1941) né à Beauvais est décédé à Paris est un ancien élève de l'E.N.S., il eut Émile Borel comme professeur (à qui l'on doit les premiers travaux importants en théorie de la mesure). Après quelques années au lycée de Nancy, Lebesgue enseignera à Rennes. C'est pendant cette période qu'il se fera connaître par son élégante théorie de la mesure. Professeur à la Sorbonne puis au collège de France, il sera élu à l'Académie des sciences en 1922. Par sa théorie des fonctions mesurables (1901) s'appuyant sur les tribus boréliennes (du nom du mathématicien Émile Borel), Lebesgue a profondément remanié et généralisé le calcul intégral. Sa théorie de l'intégration (1902-1904) répond aux besoins des physiciens en permettant la recherche et l'existence de primitives pour des fonctions "irrégulières". On lui doit aussi la transformée de Fourier établie dans la fin des années 30. Il est nommé professeur à la Sorbonne en 1910, puis au Collège de France en 1921. Il donne également des cours à l'École supérieure de physique et de chimie industrielles de la ville de Paris de 1927 à 1937 et à l'École normale supérieure de Sèvres.



Lee, Tsung-Dao (1926-) Né à Shanghai (Chine), Lee Tsung-Dao est le fils d'un homme d'affaires. La guerre sino-japonaise de 1937-1945 lui fit quitter l'université Kweichow dans la province du Zhejiang, pour rejoindre celle de Kunming, dans le Yunnan, où il rencontra Yang Chen-Ning, dont il sera longtemps l'ami et le collaborateur. Une bourse du gouvernement chinois lui permit de terminer ses études à l'université de Chicago (États-Unis), où il soutint sa thèse sur le contenu en hydrogène des naines blanches, en 1950. Membre de l'Institute for Advanced Study de Princeton (New Jersey) de 1951 à 1953, il devint bientôt, à 29 ans, le plus jeune professeur de l'université Columbia, à New York. En 1956, les physiciens étaient en butte à une énigme surgie du dépouillement des données fournies par l'accélérateur de particules du laboratoire national de Brookhaven, près de New York: deux particules, appelées "tau" et "thêta", semblaient avoir même masse et mêmes interactions nucléaires, mais différaient par leurs produits de désintégration. Lee et Yang proposèrent qu'elles n'étaient qu'une seule particule, maintenant notée "K0", et que l'interaction faible responsable de leur désintégration ne respectait pas la symétrie de parité. Ils en conclurent qu'il était indispensable de soumettre à vérification expérimentale le fait que l'interaction faible distingue la droite de la gauche. Six mois suffirent à l'équipe du National Bureau of Standards de Washington, mobilisée par la physicienne chinoise Wu Chien-Shiung, pour montrer que des noyaux radioactifs de cobalt 60 polarisés émettaient plus d'électrons dans une direction que dans la direction opposée. Confirmée rapidement par plusieurs autres groupes expérimentaux, cette violation de la symétrie miroir valut à Lee Tsung-Dao et à Yang Chen-Ning de se partager

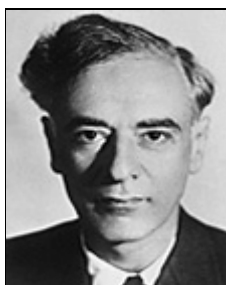
le prix Nobel de physique 1957.



Legendre, Adrien Marie (1752-1833) mathématicien français né à Paris et décédé à Auteil. Il occupe la chaire de Mathématiques de l'école militaire de Paris de 1775 à 1780. En 1783, il devient membre de l'académie des sciences. En 1787, il est nommé commissaire chargé des opérations géodésiques. Les centres d'intérêts de Legendre étaient variés: analyse, théorie des nombres, géométrie, statistiques (méthodes des moindres carrés) et mécanique (transformation de Legendre en mécanique analytique et thermodynamique). Environ un siècle avant que l'on en obtienne les preuves, il conjectura le théorème des nombres premiers ainsi que la loi de réciprocité quadratique. Toute sa vie, il s'intéressa aux intégrales elliptiques, dont les travaux allaient finalement donner naissance aux courbes elliptiques, sujet très étudié par les mathématiciens contemporains. Il laisse en héritage à la communauté mathématique du 19ème siècle un traité de géométrie élémentaire, qui s'avère très précieux dans le monde de l'enseignement..



Leibniz, Gottfried Wilhelm (1646-1716) né à Leipzig et décédé à Hanovre était philosophe, juriste et mathématicien considéré comme un des plus brillants esprits du 17ème siècle. Fils d'un jurisconsulte il obtient son baccalauréat en philosophie ancienne en 1663 et écrira un peu plus tard une théorie des probabilités en droit. Il entre ensuite à l'université de Leipzig et en 1666 obtient son doctorat en droit. En 1669, il devient conseiller à la Chancellerie de l'électorat de Mayence. Il est envoyé en 1672 à Paris, en mission diplomatique dit-on, pour convaincre Louis XIV de porter ses conquêtes vers l'Égypte plutôt que l'Allemagne. Il y reste jusqu'en 1676 et y rencontre les grands savants de l'époque. C'est pendant cette période que Leibniz travail sur ses travaux scientifiques. En 1676 il est nommé bibliothécaire du Brunswick-Lunebourg et s'y occupe aussi de mathématique, de physique, de religion et de diplomatie. Leibniz contribua aux mathématiques en découvrant, en 1675, les principes fondamentaux du calcul infinitésimal. Cette découverte fut réalisée indépendamment des découvertes de Newton, qui inventa son système de calcul en 1666. Le système de Leibniz fut publié en 1684, celui de Newton en 1687, date à laquelle la méthode de notation imaginée par Leibniz fut adoptée et on le considère aussi comme un pionnier du développement de la logique mathématique.



Landau, Lev Davidovich (1908-1968) est né en Azerbaïdjan et est le fils d'un ingénieur et médecin. Il est décédé à Moscou. Après avoir achevé ses études au Département de Physique de l'Université de Léninegrad à l'âge 19 ans, il commence sa carrière scientifique à l'Institut Physico-technique de Léninegrad. De 1932 à 1937 il est le chef du Département Théorique de l'Institut Physico-technique Ukrainien à Kharkov et dès 1937 il est nommé chef du Département Théorique de l'Institut pour les Problèmes Physiques de l'Académie des Sciences de l'URSS à Moscou. Le travail de Landau couvre toutes les branches de physique théorique, aux limites de la mécanique liquide à la théorie des champs quantiques. Une grande partie de ses papiers se réfère à la théorie de l'état condensé. Ils ont commencé en 1936 par une formulation d'une théorie générale des transitions de phase du deuxième ordre. Après la découverte de Kapitsa, en 1938, de la

superfluidité de l'hélium liquide, Landau a engagé la vaste recherche qui l'a mené à la construction de la théorie complète des liquides quantiques aux températures très basses. Parmi ses écrits, couvrant une vaste gamme de thèmes liés aux phénomènes physiques, on relève plus de cent articles et de nombreux livres, dont le célèbre *Cours de physique théorique*, publié en 1943 avec E.M.Lifchitz. Landau a dominé toute la physique théorique de 1930 à 1965. Il avait créé un ensemble d'examens de physique théorique, appelé le "Minimum théorique" que les étudiants ou chercheurs confirmés devaient passer pour entrer dans son groupe de recherche, examen qui incluait des problèmes dans toutes les branches des mathématiques.



Levi-Civita, Tullio (1873-1941) Né à Padoue et décédé à Rome. Il fut diplômé en 1892 de la faculté de mathématiques de l'Université de Padoue. En 1894, il obtint un diplôme d'enseignement au Collège d'enseignement de la faculté des sciences de Pavie. En 1898, il fut nommé à la tête de la chaire de mécanique analytique et céleste de Padoue et il y rencontra Libera Trevisani, une de ses élèves, avec qui il se maria en 1914. Il resta à Padoue jusqu'en 1918, puis fut nommé à la chaire d'analyse supérieure à l'université de Rome, où il prit 2 ans plus tard la chaire de mécanique. Avant tout physicien, ses travaux s'orientent principalement vers l'électromagnétisme et les théories de Lorentz et de Maxwell. En 1900, Ricci et lui publièrent *La Théorie des tenseurs dans les méthodes de calcul différentiel et leurs applications* qu'Einstein utilisa afin de mieux maîtriser le calcul tensoriel, un outil-clef pour Einstein dans le développement de sa théorie de la relativité générale. Levi-Civita discuta aussi d'une série de problèmes à propos du champ gravitationnel statique dans sa correspondance avec Einstein entre les années 1915-1917. Leur correspondance tournait autour de la formulation variationnelle des équations de champs gravitationnelles et leurs propriétés covariantes, et la définition de l'énergie gravitationnelle et de l'existence d'ondes gravitationnelles. En 1933 Levi-Civita contribua aussi aux équations de la mécanique quantique de Dirac.



Lie, Sophus (1842-1899), mathématicien norvégien Lie fit ses études à l'université de Christiania. Il donna des leçons particulières pour gagner sa vie, et passa avec Klein l'hiver 1869-1870 à Berlin, l'été 1870 à Paris. En 1872, une chaire de mathématiques fut créée pour lui à Christiania, et en 1886, il succéda à Klein à Leipzig. Outre des travaux en géométrie projective de l'espace, on retient surtout de Lie l'étude de structures algébriques nouvelles qu'il applique à la géométrie, jusqu'à la création de toutes pièces de la théorie des groupes et algèbres qui portent son nom. Dans la notion de groupe et d'algèbre de Lie, interviennent des propriétés de continuité (groupe topologique), annonçant la nouvelle branche importante des mathématiques que sera la topologie. Les travaux de Lie, dans ce domaine, seront principalement poursuivis par Élie Cartan.



Lindemann von, Ferdinand (1852-1939) né à Hanovre et décédé à Munich a été le premier mathématicien à démontrer la transcendance de π . Quand Ferdinand était âgé de 2 ans se déplaça à Schwerin où il passa ses années d'enfance et sa scolarité primaire. Comme il était de pratique à cette époque en Allemagne pendant la seconde moitié du 19^{ème}, Lindemann se déplaça fréquemment d'une université à l'autre. Il commença ses études à Göttingen en 1870 et y fut grandement influence par Clebsch. Plus tard, Lindemann qui avait établi des très bonnes relations avec Clebsch rédigea à nouveau les notes de géométrie de ce dernier après son décès pour leur publication en 1876. Ensuite, Lindemann étudia à Erlangen à Munich où il effectua son travail de doctorat sous la direction de Klein sur les géométries non-euclidiennes et ses applications à la physique. Après avoir obtenu son doctorat, Lindemann fit des visites importantes à des centres de mathématiques anglais et français. En Angleterre, il visita Oxford, Cambridge et Londres, alors qu'en France, il passa la majeure partie de son temps à Paris où il fut grandement influencé par Chasles, Bertrand, Jordan et Hermite. Lorsqu'il retourna en Allemagne, Lindemann travailla sur des publications permettant sa réintégration et sa reconnaissance dans le domaine scientifique Allemand. Ce fut enfin en 1877 qu'il fut nommé professeur extraordinaire à l'université de Würzburg et professeur ordinaire à l'université de Freiburg en 1879. Le principal travail de Lindemann porta sur la géométrie et l'analyse et il est particulièrement connu pour la fameuse preuve de transcendance. En 1873, alors que Lindemann venait d'avoir obtenu son doctorat, Hermite démontra la transcendance du nombre d'Euler. Peu de temps après, Lindemann rencontra Hermite à Paris et discuta des méthodes utilisées pour la démonstration. Ainsi, utilisant un raisonnement similaire, Lindemann démontra en 1882 la transcendance de π (sur la base que le nombre d'Euler est lui-même transcendant).

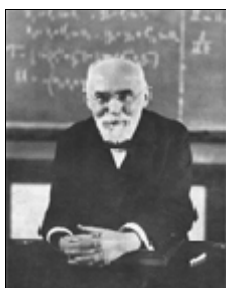


Liouville, Joseph (1809-1882) né à St-Omer et décédé à Paris fut un artisan des mathématiques déployant une activité considérable dans l'enseignement et la diffusion des idées mathématiques de son temps. Il est le fondateur du *Journal de mathématiques pures et appliquées* appelé traditionnellement "Journal de Liouville". Ses principaux travaux portent sur l'analyse et on lui doit un important théorème sur l'approximation des irrationnels algébriques. L'élection de Joseph Liouville à l'Assemblée constituante de 1848 est seule à rompre l'unité d'une carrière toute scientifique: sorti de l'École polytechnique en 1827, il y revenait en 1833 comme répétiteur puis professeur d'analyse. Dès sa 31^{ème} année, il était élu à l'Académie des sciences, dans la section d'astronomie. Il fut un des meilleurs professeurs de son temps, et ses cours, à Polytechnique et au Collège de France, prirent une grande part de son activité. Liouville fonda le *Journal de mathématiques pures et appliquées* en 1836 et le géra pendant 39 ans. Ses tâches d'académicien et d'éditeur lui ôtèrent la liberté d'esprit nécessaire à une recherche approfondie ce pour quoi il se plaint. Mais il mit à profit l'une et l'autre tâche pour aider plusieurs jeunes mathématiciens de grand avenir, par exemple C. Hermite et C. Jordan, par des rapports élogieux devant l'Académie, ou par la publication de leurs travaux dans son journal. Quant à lui, il y publia surtout de courtes notes sur un grand nombre de questions: analyse, arithmétique, géométrie, mécanique, astronomie. Il partage avec A. Cauchy le mérite d'avoir soumis l'analyse à une règle de rigueur souvent transgressée au 18^{ème} siècle, et ce mérite est d'autant

plus grand que le langage mathématique de son temps n'aidait guère à la rigueur.



Lobatchevski, Nikolai Ivanovitch (1792-1856) Mathématicien russe né à Nijni-Novgorod et décédé à Kazan. Lobatchevski étudia à l'université de Kazan, où il enseigna à partir de 1812 et occupa la chaire de mathématiques pures de 1822 à 1846. Sous l'influence de Gauss et de Laplace, ses premiers travaux sont: *Théorie du mouvement elliptique des corps célestes* et *De la solution de l'équation algébrique complexe simple*. Mais ses principales recherches concernent la géométrie. Son premier ouvrage, *Géométrie* (1823), jugé trop révolutionnaire (il utilisait le système métrique), ne pourra être publié de son vivant. En 1826, Lobatchevski expose devant ses collègues de l'université un mémoire qui montre qu'il fut l'un des premiers mathématiciens à être convaincu de la possibilité d'une géométrie différente de celle d'Euclide. Malgré le scepticisme de ses collègues, il continue l'étude de cette nouvelle géométrie (où le postulat d'Euclide est remplacé par le postulat suivant, dit "postulat de Lobatchevski": par tout point extérieur à une droite, il passe une infinité de parallèles à cette droite) et consacre sa vie de mathématicien à essayer de convaincre le monde scientifique. Il publie successivement *Éléments de géométrie* (1829), *Nouveaux Éléments de géométrie* avec la théorie complète des parallèles (1838) et *Pangéométrie* (1855). Mais la pleine reconnaissance de la valeur de ses travaux ne viendra qu'après sa mort (lorsque Eugenio Beltrami, en 1868, construira un modèle de la géométrie de Lobatchevski: la pseudo-sphère). En plus de ses recherches mathématiques, Lobatchevski fut l'animateur de l'université de Kazan: recteur de 1827 à 1846, il eut la charge de la bibliothèque de l'université, mit en place son observatoire, organisa son musée et dirigea la construction de nouveaux locaux universitaires.



Lorentz, Hendrik (1853-1928) né à Arnhem et décédé à Haarlem (Pays-Bas) a amélioré la théorie électromagnétique de Maxwell dans sa thèse doctorale sur la théorie de la réflexion et la réfraction de la lumière qu'il présenta en 1875. Il a été nommé professeur de physique-mathématique à l'université de Leyde en 1878. Il est resté dans cet établissement jusqu'en 1912 où Ehrenfest a été nommé à sa place. Lorentz est ensuite nommé directeur de recherche à l'institut de Teyler, Haarlem. Il a maintenu une position honorifique à Leyde, où il a continué à donner quelques cours. Avant que l'existence des électrons ait été prouvée, Lorentz a proposé que les vagues de lumière étaient dues aux oscillations d'une charge électrique dans l'atome. Lorentz a développé sa théorie mathématique de l'électron pour lequel il a reçu le prix Nobel en 1902. Le prix Nobel a été attribué conjointement à Lorentz et à Pieter Zeeman, un étudiant de Lorentz. Zeeman avait vérifié expérimentalement le travail théorique de Lorentz sur la structure atomique, démontrant l'effet d'un champ magnétique fort sur les oscillations en mesurant le changement de la longueur d'onde de la lumière produite. Lorentz est également célèbre pour son travail sur la contraction de Fitzgerald-Lorentz, qui est une contraction dans la longueur d'un objet aux vitesses relativistes. Les transformations de Lorentz, qu'il a présentées en 1904, forment la base de la théorie de relativité restreinte d'Einstein qui à l'époque était appelée "théorie de la relativité d'Einstein-Lorentz". Elles décrivent l'augmentation de la masse, du rapetissement de la longueur, et de la dilatation de temps d'un corps se déplaçant aux vitesses proches de celle de la lumière. Lorentz était président de la première

conférence de Solvay tenue à Bruxelles en automne de 1911. Cette conférence avait pour sujet les deux approches de la théorie atomique, à savoir la théorie classique et la physique quantique. Cependant, Lorentz n'a jamais entièrement accepté la théorie quantique et a toujours espéré qu'il serait possible de l'incorporer de nouveau dans l'approche classique.



Lucas, Édouard (1842-1891) est un arithméticien français né à Amiens et décédé à Paris. Enfant issu d'une famille très modeste, il reçoit une bourse et réussit le concours d'entrée à l'École Normale Supérieure, en 1861. À la sortie de l'École, il devient astronome adjoint à l'Observatoire de Paris, puis après la guerre franco-prussienne, il obtient une chaire de Mathématiques Spéciales à Moulins, de 1872 à 1876. Puis il occupe une chaire à Paris, d'abord au lycée Charlemagne à Paris, puis au déjà très prestigieux lycée Saint-Louis. Ses travaux mathématiques concernent la géométrie euclidienne non élémentaire (celle des transformations, en particulier la géométrie projective vue à travers ses homographies), et surtout la théorie des nombres. Sa principale contribution est celle faite aux tests de primalité. Tombée dans un oubli relatif en France (où la théorie algébrique des nombres est reléguée au second plan, en attendant Weil), l'oeuvre de Lucas est reprise et enrichie par les anglo-saxons, et notamment par Lehmer, qui améliorera son test de primalité et prouvera totalement certains résultats de Lucas, pour obtenir le test de Lucas-Lehmer, qui est encore celui qui est utilisé à la fin du 20ème siècle pour battre des records de grands nombres premiers. Ces travaux prennent une importance particulière depuis que l'avènement de l'informatique rend la cryptographie avide de très grands nombres premiers. Lucas est aussi connu pour être l'inventeur de nombreuses récréations mathématiques. La plus répandue d'entre elles est le problème des tours de Hanoi, qu'il publia sous le nom de Claus de Siam, professeur au collège de Li-Sou-Tsiam, anagramme de Lucas d'Amiens, professeur à Saint-Louis.



Malthus, Thomas Robert (1766-1834) né à Guildford et décédé à Bath est un pasteur anglican, qui s'inquiéta de la croissance trop importante de la population en Angleterre, au début de la révolution industrielle (de 1750 à 1900). Sa crainte tournait autour de l'idée que la progression démographique est plus rapide que l'augmentation des ressources, d'où une paupérisation de la population. Les anciens régulateurs démographiques (les guerres et les épidémies) ne jouant plus leurs rôles, il imagine de nouveaux obstacles, comme la limitation de la taille des familles et le recul de l'âge du mariage. Ces propositions ne sont appliquées à ce jour, toutes les deux, qu'en Chine, qui est en effet obligée de limiter sévèrement sa démographie. Les prévisions sinistres de Malthus sont dans la réalité mises à mal, car il n'imaginait pas une si grande augmentation des ressources et des rendements agricoles (révolution verte: chimie appliqué à l'agronomie ce qui n'est pas pour autant bénéfique...); les nouveaux moyens d'échanges internationaux de biens de subsistance (contribuant à la pollution des océans au passage...); le fait que le trop plein d'individus émigrerait vers les États-Unis ou les colonies. En revanche, si les prévisions de Malthus ne sont pas au rendez-vous, sa théorie garde tous ses droits. Il est exact que la population est en croissance dans certains pays (Arabie Saoudite: 6 enfants par femme) il est aussi exact (et heureux) que les progrès de l'hygiène et de la médecine augmentent la taille de la population, il est exact que

les ressources renouvelables sur Terre sont limitées, in fine par l'énergie solaire que reçoit celle-ci, qui elle-même détermine la biomasse, sauf découverte scientifique majeure... et dans ces conditions, la mathématique est formelle: il ne sera pas possible à la population Terrestre d'augmenter indéfiniment, et la régulation devra intervenir à un moment ou à un autre, et d'une manière ou d'une autre!



Marconi, Guglielmo (1874-1937) né à Rome et décédé à Rome est un physicien, inventeur et homme d'affaires italien. Il est, avec Karl Ferdinand Braun, colauréat du prix Nobel de physique de 1909 en reconnaissance de leurs contributions au développement de la télégraphie sans fil (on peut considérer qu'il est l'origine des appareils de transmission/réception d'ondes électromagnétiques et donc de la radio et télévision hertzienne). Marconi dans une famille aisée, second fils de Giuseppe Marconi, un propriétaire italien, et d'une mère irlandaise, Annie Jameson, petite-fille du fondateur de la Distillerie Jameson Whiskey. Il a fait ses études à Bologne dans le laboratoire d'Augusto Righi, à Florence, à l'Institut Cavallero et, plus tard, à Livourne. Il fait en 1885 des expériences sur les ondes découvertes par Heinrich Rudolf Hertz 7 ans auparavant. Il reproduit le matériel utilisé par Hertz en l'améliorant avec un cohéreur de Branly pour augmenter la sensibilité et l'antenne de Alexandre Popov. Après ses toutes premières expériences en Italie, il réalise dans les Alpes suisses à Salvan une liaison de 1.5 km durant l'été 1895. L'année d'après, faute d'être suivi par ses compatriotes, il part pour l'Angleterre, poursuit ses expériences et dépose un brevet. En 1897 il effectue la première communication en morse à plus de 13 km entre Lavernock (Pays de Galles) et Brean (Angleterre) par-dessus le Canal de Bristol. L'année d'après, il ouvre la première usine de radios au monde, à Chelmsford, Angleterre. Au début du 20ème siècle le nom Marconi est (malheureusement) plutôt connu comme étant le propriétaire du groupe de cinéma Pathé (qui s'appelle en réalité Pathé-Marconi).



Mandelbrot, Benoît (1924-2010) né à Varsovie et décédé à Cambridge. Sa famille a quitté la Pologne pour Paris afin de fuir la menace hitlérienne. C'est à Paris qu'il fut initié aux mathématiques par deux oncles dont un était professeur au Collège de France. L'invasion allemande force la famille à se réfugier ensuite à Brive-la-Gaillarde. Après avoir fréquenté le lycée Edmond-Perrier de Tulle, il poursuit ses études au lycée du Parc, à Lyon. Après avoir quitté l'École polytechnique (promotion 1944), où il a suivi les cours d'un spécialiste du calcul des probabilités (Paul Lévy), il s'intéresse aux phénomènes d'information, les idées de Claude Shannon étant alors en plein essor. Mandelbrot fit ses études en France et aux États-Unis et obtint son doctorat de mathématiques à l'université de Paris en 1952. Il enseigna l'économie à l'université Harvard, l'ingénierie à Yale, la physiologie à la faculté de médecine et la mathématique à Paris et à Genève. À partir de 1958, il travailla pour IBM au centre de recherche Thomas B. Watson à New York sur la transmission optimale dans les milieux bruités, tout en poursuivant son travail sur des objets étranges jusque-là assez négligés par les mathématiciens : les objets à complexité récursivement définie, comme la courbe de Von Koch, auxquels il pressent une unité: la géométrie fractale. La géométrie fractale se distingue par son approche plus abstraite de la dimension qu'elle ne

l'est dans la géométrie traditionnelle. Elle trouve de plus en plus d'applications dans différents domaines de la science et de la technologie.

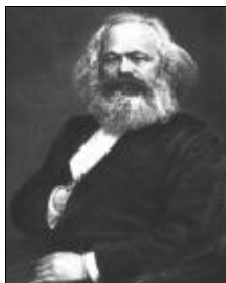


Markov, Andreï Andreïevitch (1856-1922) Mathématicien russe spécialiste de la théorie des nombres, de la théorie des probabilités et de l'analyse mathématique né à Riazan et décédé à Petrograd. Issu d'une la famille d'un petit fonctionnaire du gouvernement, il fait ses études à l'université de Saint-Pétersbourg et reçoit une médaille d'or pour son mémoire *De l'intégration des équations différentielles par la méthode des fractions continues* (1878). Professeur à l'université de Saint-Pétersbourg en 1886, il devient membre de l'Académie des sciences en 1896. Les recherches de Markov continuent l'oeuvre de ses devanciers de l'école mathématique pétersbourgeoise: P. L. Tchebychev, E. I. Zolotarev et A. N. Korkin. Sa thèse *Des formes quadratiques bilinéaires de déterminant positif* (1880) inaugure ses travaux dans le domaine de la théorie des nombres. En analyse, ses recherches concernent les fractions continues, les limites d'intégrales, la convergence des séries et la théorie de l'approximation. On lui doit une solution simple de la détermination de la limite supérieure de la dérivée d'un polynôme (inégalité de Markov). Après 1910, se tournant vers la théorie des probabilités, il démontre de façon rigoureuse, sous des conditions assez générales, le théorème central limite relatif à une somme de variables aléatoires indépendantes. Cherchant à généraliser ce théorème aux variables aléatoires dépendantes, il est amené à considérer la notion importante d'événements en chaînes, appelés depuis chaînes de Markov, et il établit une série de lois, fondement de la théorie des processus de Markov. Il étend plusieurs résultats classiques concernant des événements indépendants à certains types de chaînes. Ses travaux sont à l'origine de la théorie moderne des processus stochastiques. Markov s'intéressait aussi aux applications de la théorie des probabilités, et il a justifié de façon probabiliste la méthode des moindres carrés.



Markowitz, Harry Maurice (1927-) né à Chicago, professeur à la City University de New York est connu pour avoir développé la théorie dite du "choix des portefeuilles pour le placement des fortunes". Markowitz ne se doutait pas que son article de jeunesse publié en 1952 dans le *Journal of Finance*, puis développé dans un livre paru en 1959, *Portfolio Selection: Efficient diversification*, jetterait les bases de la théorie moderne du portefeuille et de son utilisation par un grand nombre de praticiens. Plus précisément, Markowitz a montré que l'investisseur cherche à optimiser ses choix en tenant compte non seulement de la rentabilité attendue de ses placements, mais aussi du risque de son portefeuille qu'il définit mathématiquement par la variance de sa rentabilité. Appliquant des théorèmes classiques du calcul statistique et des techniques probabilistes, il a ainsi démontré qu'un portefeuille composé de plusieurs titres est toujours moins risqué qu'un portefeuille composé d'un seul titre, quand bien même il s'agirait du moins risqué d'entre eux. La mise en oeuvre du modèle de Markowitz a très vite posé des problèmes d'ordre pratique. Alors que le volume des statistiques nécessaires au calcul augmentait rapidement avec le nombre de titres retenus (avec 100 titres, le nombre de statistiques nécessaires était de 3'150, mais il passait à 20'300 pour 200 titres et à 125'750 pour 300 titres !), la collecte des informations et leur traitement devenaient presque impossibles avec les

ordinateurs disponibles dans les années 1960, entraînant de surcroît des coûts de traitement prohibitifs. C'est la raison pour laquelle William F. Sharpe cherchera une méthode de sélection des portefeuilles efficient plus simple. Markowitz et Sharpe seront alors reconnus comme les pères fondateurs de la gestion de portefeuilles et du corps doctrinal sur lequel elle se fonde. Le prix Nobel de sciences économiques leur sera décerné ainsi qu'à Merton Miller en 1990.



Marx, Karl (1818-1883) né à Trèves et décédé à Londres entra à l'Université de Bonn puis à celle de Berlin, après avoir terminé le Lycée de Trèves. Il étudia à Berlin le droit, mais surtout l'histoire et la philosophie. Marx a ensuite contribué à parachever les trois principaux courants d'idées du 19ème siècle: la philosophie classique allemande, l'économie politique classique anglaise et le socialisme français. La théorie sociale de Marx a pour objectif de dévoiler la loi économique de la société capitaliste où ce qui domine est la production des marchandises en recherchant l'origine de la forme monétaire de la valeur. Ainsi pour Marx, l'argent (en tant que produit suprême du développement de l'échange et de la production marchande) estompe et dissimule le caractère et le lien social du travail individuel. À un certain degré du développement de la production des marchandises, l'argent se transforme aussi en capital. Ainsi, la séquence de la circulation des marchandises était: M (marchandise) - A (argent) - M (marchandise), c'est-à-dire vente d'une marchandise pour l'achat d'une autre. La séquence générale du capital est par contre A-M-A, c'est-à-dire l'achat pour la vente (avec un profit). C'est cet accroissement de la valeur primitive de l'argent, donc sa transformation en capital, que Marx appelle "plus-value" et qui ne peut provenir de la circulation des marchandises, car celle-ci ne connaît que l'échange d'équivalants; elle ne peut provenir non plus d'une majoration des prix, étant donné que les pertes et les profits réciproques des acheteurs et des vendeurs s'équilibrent à grande échelle. Pour obtenir de la plus-value il faut selon Marx une marchandise dont le processus de consommation fût en même temps un processus de création de valeur. Or, cette marchandise est la force de travail humaine. Le possesseur d'argent achète la force de travail à sa valeur, déterminée, comme celle de toute autre marchandise, par le temps de travail socialement nécessaire à sa production. Ayant acheté la force de travail, le possesseur d'argent est en droit de la consommer, c'est-à-dire de l'obliger à travailler toute la journée, disons, 8 heures. Or, en 5 heures (temps de travail nécessaire), l'ouvrier crée un produit qui couvre les frais de son entretien, et, pendant les 3 autres heures (temps de travail supplémentaire), il crée un produit supplémentaire, non rétribué par le capitaliste, et qui est la plus-value. Aussi, pour exprimer le degré d'exploitation de la force de travail par le capital, faut-il comparer la plus-value non pas au coût total de production, mais uniquement au coût variable de la main d'oeuvre humaine.



Maxwell, James Clerk (1831-1879) né à Edimbourg et décédé à Glenlair. Brillant élève au collège, James Clerk Maxwell poursuit des études de mathématiques à l'université de Cambridge. Il obtient une chaire de philosophie naturelle à Aberdeen à l'âge de 25 ans. Puis, de 1860 à 1865, il occupe le poste de professeur au King's College de Londres. À la suite de ces 5 années d'enseignement, il décide de se retirer dans sa propriété de Glenlair, en Écosse. Il y restera 5 autres années qu'il emploiera à étudier. En 1871, Maxwell est nommé

directeur du laboratoire Cavendish que vient de fonder le duc du Devonshire. Il n'aura alors de cesse de le développer afin qu'il devienne le centre de formation scientifique le plus illustre. Dès le début de sa carrière, Maxwell s'intéresse à la dynamique des gaz. Après avoir prouvé mathématiquement que les anneaux de Saturne sont constitués de particules distinctes, il étudie la répartition des vitesses des molécules gazeuses (conforme à loi de Gauss). En 1860, il montre que l'énergie cinétique de ces molécules ne dépend que de leur nature. Mais ce sont ses recherches en électromagnétisme qui font de Maxwell un des savants les plus célèbres du 19ème siècle. En se basant sur les travaux de Faraday, il introduit dès 1862 la notion de champ. Puis, il montre qu'un champ magnétique peut être créé par la variation d'un champ électrique (Faraday avait alors découvert l'induction, phénomène par lequel la variation d'un champ électrique crée un champ magnétique). Son enseignement purement mathématique va alors lui permettre d'élaborer les célèbres équations différentielles décrivant la nature des champs électromagnétiques dans l'espace et le temps. Il les expose dans son *Traité d'électricité et de magnétisme* publié en 1873. Maxwell, en élaborant les théories de l'électromagnétisme, a également défini la lumière en tant qu'onde électromagnétique, ouvrant ainsi la voie aux recherches d'autres physiciens comme Heinrich Rudolph Hertz.



McFadden, Daniel (1937-) né à Raleigh est un économétricien ayant reçu en 2000, avec James Heckman, le prix Nobel d'économie pour son apport aux théories et méthodes de l'analyse des choix discrets. Il obtient un Bachelor de Science en physique à l'âge de 19 ans à l'Université du Minnesota puis un doctorat de philosophie en sciences du comportement (économie) 5 ans plus tard en 1962. En 1964, il intègre l'Université de Berkeley et focalise ses recherches sur les comportements de choix, et sur les liens entre la théorie économique et les mesures économiques. En 1975, il est récompensé par la médaille John Bates Clark. En 1977, il se rend au Massachusetts Institute of Technology, mais retourne à Berkeley en 1991, car le MIT n'avait pas de département de statistiques. Après son retour, il fonde le laboratoire d'économétrie, qui est dédié à l'informatique statistique appliquée à l'économie. McFadden a développé en microéconométrie des théories et des méthodes d'analyse des comportements par choix discrets (par exemple, les données sur les métiers et les lieux de résidence des individus) et est connu pour être à l'origine du coefficient pseudo- R de la régression logistique probit. À partir de sa théorie économique sur les choix discrets, McFadden a développé de nouvelles méthodes statistiques qui ont eu une influence décisive sur la recherche théorique, mais qui sont aussi largement utilisées par le marketing.



Mendel, Johann Gregor (1822-1884) né à Heinzendorf et décédé à Brno, a été moine dans le monastère de Brno (en Moravie). Mendel est communément reconnu comme un botaniste père fondateur de la génétique. Il est à l'origine de ce qui est aujourd'hui appelé les lois de Mendel, qui définissent la manière dont les gènes se transmettent de génération en génération. Mendel naît dans une famille de paysans. Doué pour les études, mais de tendance dépressive qui lui vaudra de multiples indispositions dans la suite de sa carrière, le jeune garçon est

très vite remarqué par le curé du village qui décide de l'envoyer poursuivre ses études loin de chez lui. Mendel part en 1851 pour suivre les cours, en tant qu'auditeur libre, de l'Institut de physique de Christian Doppler. Il y étudie, en plus des matières obligatoires: la botanique, la physiologie végétale, l'entomologie et la paléontologie. Durant 2 années, il acquiert les bases méthodologiques qui lui permettront de réaliser plus tard ses expériences. Au cours de son séjour à Vienne, Mendel est amené à s'intéresser aux théories de Franz Unger, professeur de physiologie végétale. Celui-ci préconise l'étude expérimentale pour comprendre l'apparition des caractères nouveaux chez les végétaux au cours de générations successives. Il espère ainsi résoudre le problème que pose l'hybridation chez les végétaux. De retour au monastère, Mendel installe un jardin expérimental dans la cour et dans la serre, en accord avec son abbé, et met sur pied un plan d'expériences visant à comprendre les lois de l'origine et de la formation des hybrides. Il choisit pour cela le pois qui a l'avantage d'être facilement cultivé avec de nombreuses variétés décrites. En 1865, il expose à la Société des sciences naturelles de Brunn et publie en 1866 les résultats de ses études. Après dix années de travaux minutieux, Mendel a ainsi posé les bases théoriques de la génétique et de l'hérédité moderne. Son travail ne va pas susciter d'enthousiasme auprès de ses contemporains qui ont du mal à comprendre la formalisation mathématique de ses expériences. Très peu de scientifiques de son temps vont citer son travail et Mendel ne reçoit guère de réponses auprès des différents correspondants à qui il envoie un tiré-à-part. Parmi ces derniers, seul Karl Wilhelm von Nägeli, professeur de botanique à Munich, lui écrit, doutant d'ailleurs de certaines de ses conclusions. En 1868, Mendel est élu supérieur de son couvent à la mort de l'abbé.



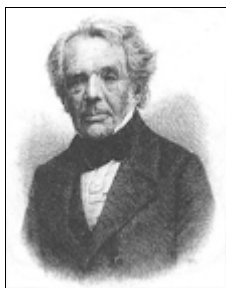
Mendeleïev, Dmitri Ivanovitch (1834-1907), né à Tobolsk et décédé à Saint-Pétersbourg fut chimiste russe surtout connu pour sa classification périodique des éléments publiée en 1869. Il montra en effet que les propriétés chimiques des éléments dépendaient directement de leur poids atomique et qu'elles étaient des fonctions périodiques de ce poids. Il entre à l'âge de 14 ans au lycée de Tobolsk, après la mort de son père. En 1849, la famille devenue pauvre s'installe à Saint-Pétersbourg et Mendeleïev entre à l'université en 1850. Après avoir reçu son diplôme, il contracte la tuberculose ce qui l'oblige à se déplacer dans la péninsule criméenne près de la Mer Noire en 1855, où il devient responsable des sciences du lycée local. Il revient complètement guéri à Saint-Pétersbourg en 1856 où il étudie la chimie et fut diplômé en 1856. À 25 ans, il vient travailler à Heidelberg avec des savants comme Robert Bunsen et Gustav Kirchhoff. À Heidelberg, il rencontra le chimiste italien Stanislao Cannizzaro, dont les idées sur le poids atomique influencèrent sa réflexion. Mendeleïev retourna à Saint-Pétersbourg et enseigna la chimie à l'Institut technique en 1863. En 1867, il est nommé professeur de chimie minérale à l'Université de Saint-Pétersbourg. Il fut nommé professeur de chimie générale à l'université de Saint-Pétersbourg en 1866.



Merton, Robert Cox (1944-) a reçu le prix Nobel d'économie en 1997, en même temps que son compatriote Myron Scholes, pour avoir élaboré la méthode d'évaluation des instruments financiers dérivés. Cette méthode d'évaluation a certainement accéléré la croissance rapide des marchés des instruments financiers dérivés depuis les années 1980 et permis l'amélioration de la gestion des risques attachés à ces nouveaux produits financiers. Merton a sans conteste contribué à ouvrir une voie nouvelle dans le champ des sciences économiques et fortement influencé les deux autres lauréats. Né en 1944 à New York, il quitte le California Institute of Technology avec un mastère en mathématiques appliquées. Il obtient par la suite un doctorat en sciences économiques au Massachusetts Institute of Technology (M.I.T.) de Cambridge, sous la direction de Paul Samuelson (Prix Nobel d'économie 1970) et se spécialise dans les problèmes d'application des méthodes probabilistes à l'évolution aléatoire des cours des actifs financiers. En 1988, il occupe la chaire George Fischer Backer de professeur en Business Administration à la Harvard Business School de Cambridge. Le travail novateur de Merton date du début des années 1970, période pendant laquelle il élabore une méthode originale de calcul de la valeur des instruments dérivés. L'échec de sa méthode appliquée à la gestion d'un fonds de placement à risques américain (Long-Term Capital Management) en 1998, a quelque peu terni sa réputation de spécialiste de la finance internationale. Mais Merton avait lui-même déclaré à une chaîne de télévision américaine, au lendemain de l'attribution de sa récompense que c'est une mauvaise interprétation de penser que l'on peut éliminer les risques simplement parce qu'on les comprend et qu'on les mesure.



Minkowski, Hermann (1864-1909) né à Alexotas et décédé à Göttingen, est un physicien-mathématicien qui a étudié aux universités de Berlin et de Königsberg. Il fait des études secondaires au lycée de Königsberg où il se fait remarquer par ses résultats en mathématiques et reçoit son doctorat en 1885 dans la même ville. Il a ensuite enseigné dans plusieurs universités, à Bonn, Königsberg et à Zurich. À Zurich, Einstein était un des étudiants dans plusieurs des cours qu'il a donnés. Minkowski a accepté une chaire en 1902 à l'université de Göttingen, où il est resté pour le reste de sa vie. À Göttingen, il apprit la physique-mathématique de Hilbert, il a participé à une conférence sur la théorie de l'électron en 1905 et appris les derniers résultats dans la théorie dans l'électrodynamique. En 1907, Minkowski s'est rendu compte que le travail de Lorentz et d'Einstein pourrait mieux être compris dans un espace non-euclidien. Il a considéré l'espace et le temps, qui a été autrefois pensé pour être indépendant, d'être couplé ensemble dans un continuum d'espace-temps quadridimensionnel. Minkowski a établi un traitement quadridimensionnel de l'électrodynamique. Ce continuum d'espace-temps a fourni un cadre pour tout le travail mathématique postérieur dans la relativité. Ces idées ont été employées par Albert Einstein en développant la théorie générale de relativité. Minkowski était principalement intéressé par la mathématique pure et a passé beaucoup de son temps en étudiant les formes quadratiques et les fractions continues. Son travail le plus original était cependant sa *Géométrie des nombres*. Cette étude a mené à des travaux sur les corps convexes et aux questions au sujet des problèmes d'emballage (les manières dans lesquelles des figures d'une forme donnée peuvent être placées dans une autre figure donnée)



Möbius, August Ferdinand (1790-1868) mathématicien et astronome allemand né à Schulpforta et mort à Leipzig. Möbius fit ses études à Leipzig, à Göttingen (sous la direction de Gauss) et à Halle. En 1815, il devint professeur d'astronomie à Leipzig, puis directeur de l'observatoire de cette ville, après en avoir dirigé la construction. On lui doit plusieurs ouvrages d'astronomie théorique, notamment *De computandis occultationibus fixarum per planetas* (1815). Ses travaux mathématiques concernent principalement la géométrie et furent, pour la plupart, publiés dans le *Journal des mathématiques pures et appliquées* de Crelle, de 1828 à 1858, comme compléments à son ouvrage fondamental *Der barycentrische Calcul* (1827). En introduisant un nouveau système de coordonnées, Möbius y étudie les transformations géométriques, principalement la transformation projective. Son ouvrage eut une très grande importance dans le développement de la géométrie projective. Étudiant la statique sous l'angle de la géométrie, Möbius développa également la théorie des complexes linéaires de droites (*Lehrbuch der Statik*, 1837). On peut considérer Möbius comme un des pionniers de la topologie, avec la découverte, publiée dans un mémoire à l'Académie des sciences française, du fameux "ruban de Möbius", surface n'ayant qu'un seul côté.



Monge, Gaspard (1746-1818) né à Beaune et décédé à Paris Monge est fils d'un marchand forain. Il suit d'abord le collège de Beaune, puis va ensuite collège de Lyon, où il enseigne dès l'âge de 16 ans les sciences physiques. Un officier du génie, qui avait vu un plan de la ville de Beaune fait par Monge à l'aide de nouvelles méthodes d'observation et de construction graphique, le recommande au commandant de l'École militaire de Mézières. Mais il ne peut y être admis à cause de son origine roturière et n'est accepté que dans une annexe technique de l'école. Ses talents scientifiques sont reconnus lorsqu'un jour il dresse le plan de fortifications à l'aide d'une méthode bien plus rapide que les méthodes connues jusque-là. Il est alors admis à l'École militaire comme professeur de mathématiques et continue ses recherches, arrivant à la méthode générale de représentation géométrique connue depuis lors sous le nom de géométrie descriptive. Mais ses découvertes, considérées comme secret militaire de grande valeur, ne peuvent être publiées. En 1780, il vient à Paris enseigner l'hydrodynamique. Il entre aussitôt à l'Académie des sciences, où il fait une communication sur les lignes de courbure tracées sur une surface (problème déjà étudié par Euler en 1760). En 1786, il publie son célèbre *Traité élémentaire de la statique* et fonde peu après l'École polytechnique, où il aura l'occasion de donner des leçons de géométrie descriptive et de publier ses travaux jusque-là inconnus. Chargé de mission en Italie, Monge rencontre Bonaparte et se charge du recrutement des savants pour l'expédition d'Égypte. Revenu en France, il reprend son enseignement à l'École polytechnique, devient sénateur et est anobli. Mais la Restauration le privera de tous ses titres, le rayera de la liste des membres de l'Institut et lui enlèvera son poste d'enseignant. En 1889, ses cendres ont été transférées au Panthéon. Toutes ses recherches mêlent étroitement la géométrie pure, l'analyse infinitésimale et la géométrie analytique, lui permettant, par exemple, de lier chaque famille de surfaces à une équation aux dérivées partielles et, par là, de trouver les solutions d'équations différentielles à l'aide de sa théorie des surfaces. L'influence de Monge s'exerça par son enseignement oral, et la

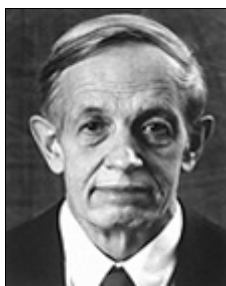
plupart des mathématiciens français du 19^{ème} siècle ont été ses élèves.



Napier, John (1550-1617) né et décédé à Merchiston est un théologien, physicien, astronome et mathématicien. Comme c'était la pratique courante pour les membres de la noblesse à l'époque, John Napier n'est pas entré à l'école avant l'âge de 13 ans. Il n'est pas resté à l'école très longtemps, cependant. On croit qu'il a quitté l'école en Ecosse et a peut-être voyagé en Europe continentale afin de mieux poursuivre ses études. En 1571, Napier, âgé de 21 ans, est retourné en Écosse, et a acheté un château à Gartness en 1574. A la mort de son père en 1608, Napier et sa famille ont emménagé dans Merchiston Castle à Edimbourg, où il résida le reste de sa vie. La mathématique n'était pas son activité principale mais il ne manquait pas d'idées pour simplifier les calculs. Il établit quelques formules de trigonométrie sphérique, popularisa l'usage du point pour la notation anglo-saxonne des nombres décimaux mais surtout inventa les logarithmes. Son objectif était de simplifier les calculs trigonométriques nécessaires en astronomie. Il s'attacha à définir le logarithme d'un sinus en s'appuyant sur des considérations mécaniques de points mobiles et sur le lien entre les progressions arithmétique et géométrique.



Navier, Henri (1785-1836) né à Dijon et décédé à Paris était un ingénieur, mathématicien et économiste surtout connu pour ses travaux sur l'hydrodynamique. Henri devient orphelin à 9 ans, après la mort de son père, avocat réputé et ancien député durant la Révolution. Son oncle, ingénieur du Corps des Ponts et Chaussées s'occupe de son éducation à Paris, le considère comme son fils avant de l'adopter avec sa femme, également proche parente du jeune Henri. L'ingénieur cantonnier le pousse à se présenter à l'École Polytechnique. Bien qu'étant parmi les derniers reçus en 1802, il y réussit sa scolarité et son classement lui permet d'intégrer le corps des ponts et chaussées. Il est nommé ingénieur ordinaire des Ponts et Chaussées en 1808. Plus tard, il deviendra inspecteur divisionnaire de ce corps et, semble-t-il quelque temps, inspecteur général à l'instar de son oncle. De 1819 à 1835, il assure le cours de mécanique appliquée de l'École nationale des ponts et chaussées (il y est titularisé en 1830 à la suite de la retraite d'Eisenmann). Au début des années 1820, il explore avec Augustin-Louis Cauchy les facettes de la théorie mathématique de l'élasticité, ce qui lui permet de proposer des équations sur le mouvement des fluides newtoniens.



Nash, John (1928-) né en Virginie Occidentale, fils de John Nash Sr., ingénieur, et Virginia Martin, enseignante. Jeune, il passait beaucoup de temps à lire et à faire des expériences dans sa chambre qu'il avait convertie en laboratoire. De 1945 à 1948, Nash a étudié au Carnegie Institute of Technology à Pittsburgh, dans l'intention de devenir ingénieur comme son père. À la place, il y a développé une passion durable pour les mathématiques, et en particulier pour la théorie des nombres, les équations diophantiennes, la mécanique quantique et la théorie de la relativité. Il fut admis en troisième cycle, à 20 ans, dans toutes les universités qu'il avait sollicitées: Harvard, Princeton... Il choisit d'aller à Princeton. Ayant un intérêt pour l'économie, Nash se mit à étudier la théorie des jeux, domaine qu'avait défriché John von Neumann, un des grands noms de Princeton, un peu

plus d'une décennie auparavant. C'est sur ce sujet qu'il décida de faire sa thèse et qu'il obtint le prix Nobel d'économie en 1994. Durant l'été 1950, Nash fut employé comme consultant à la RAND, institut top-secret qui employait de la matière grise pour mettre au point diverses stratégies de statu quo soit de victoire, en cas de conflit faisant appel à l'arme nucléaire. À la suite, Nash se mit à étudier les variétés lisses compactes, ce qui fit l'objet d'un papier. Il devint ensuite assistant au M.I.T à la rentrée 1951-52, âgé de 23 ans seulement. Il avait vraiment un tempérament de problem-solver et releva ainsi le pari de résoudre une question de Warren Ambros: est-il possible de plonger une variété riemannienne quelconque dans un espace euclidien? Nash trouve une méthode fondamentale originale pour y arriver. Nash devint malade après quelques problèmes privés et professionnels, mais il attribua sa maladie à sa tentative de résoudre les contradictions de la physique quantique. D'autant plus que peu de temps auparavant, il avait réalisé des travaux sur les EDP elliptiques non linéaires qui lui valurent beaucoup d'admiration autour de lui, mais dont il dut finalement partager la paternité avec un jeune italien qui avait énoncé, indépendamment et quelques semaines avant lui, des résultats similaires: ceci leur valu de ne pas obtenir la médaille Fields en 1958...



Newton, Isaac (1642-1727), mathématicien et physicien anglais, considéré comme l'un des plus grands scientifiques de l'histoire. Newton naît dans le Lincolnshire (Angleterre), de parents paysans et décédera à Londres. À 5 ans, il fréquente l'école primaire de Skillington, puis à 12 ans celle de Grantham. Il y reste quatre années jusqu'à ce que sa mère le rappelle à Woolsthorpe pour qu'il devienne fermier et qu'il apprenne à administrer son domaine. Pourtant, sa mère, s'apercevant que son fils était plus doué pour la mécanique que pour le bétail, l'autorisa à retourner à l'école pour peut-être pouvoir entrer un jour à l'université. À 17 ans, Newton tombe amoureux d'une camarade de classe, mademoiselle Storey. On l'autorise à la fréquenter et même à se fiancer avec elle, mais il doit terminer ses études avant de se marier. Finalement, le mariage ne se fit pas et Newton restera alors célibataire toute sa vie. À 18 ans, il entre alors au Trinity College de Cambridge (il y restera 7 ans), où il se fait remarquer par son maître, Isaac Barrow. Il a également comme professeur Henry More qui l'influencera dans sa conception de l'espace absolu. À Cambridge, il étudie l'arithmétique, la géométrie dans les *Éléments d'Euclide* et la trigonométrie, mais s'intéresse particulièrement à l'astronomie, à l'alchimie et à la théologie. Il devient à 25 ans bachelier des arts, mais est contraint de suspendre ses études pendant deux années suite à l'apparition de la peste qui s'est abattue sur la ville en 1665 ; il retourne alors dans sa région natale. C'est à cette période que Newton progresse fortement en mathématiques, physique et surtout en optique. Il laissa d'importantes contributions à de nombreuses branches de la science. Ses découvertes et théories furent à la base d'une grande partie des progrès scientifiques réalisés après lui. Newton fut l'un des inventeurs de la branche des mathématiques appelée calcul infinitésimal (l'autre inventeur fut le mathématicien allemand Gottfried Wilhelm Leibniz). Il éclaircit également les mystères de la lumière et de l'optique, formula trois lois sur le mouvement et en déduisit la loi de la gravitation universelle en se basant sur les lois de Kepler. Il parvint au raisonnement selon lequel la lumière est un mélange de différents rayons de

couleurs différentes, et qu'en raison des phénomènes de réflexion et de réfraction, ses couleurs apparaissent en composants séparés. Newton démontra sa théorie des couleurs en faisant passer de la lumière au travers d'un prisme, qui scinde le faisceau lumineux en couleurs séparées. En 1696, il quitte Cambridge pour devenir d'abord gardien de la Royal Mint puis maître de la monnaie dès l'année suivante. En 1699, il est nommé membre du conseil de la Royal Society et y est élu président en 1703. Il garde cette place jusqu'à sa mort.



Neumann Von, John (1903-1957), mathématicien né à Budapest et décédé à Washington. Neumann est un enfant prodige: à 6 ans, il converse avec son père en grec ancien et peut mentalement faire la division d'un nombre à huit chiffres. Une anecdote rapporte qu'à 8 ans, il a déjà lu les 44 volumes de l'histoire universelle de la bibliothèque familiale et qu'il les a entièrement mémorisés : doté d'une mémoire absolue, il sera capable de citer de mémoire des pages entières de livres lus des années auparavant. Il entre au lycée à Budapest en 1911. C'est âgé d'à peine 23 ans qu'il reçoit son doctorat en mathématiques (avec des mineures en physique expérimentale et en chimie) de l'université de Budapest. En parallèle, il obtient un diplôme en génie chimique de l'École polytechnique fédérale de Zurich (à la demande de son père, désireux que son fils s'investisse dans un secteur plus rémunérateur que les mathématiques). Il est intéressant de noter que von Neumann n'a mis les pieds dans ces deux universités que pour les examens. Entre 1926 et 1930, il est privatdozent à Berlin et à Hambourg. Il travaille également à Göttingen avec Robert Oppenheimer sous la direction de David Hilbert. Durant cette période, l'une des plus fécondes de sa vie, il côtoie également Werner Heisenberg et Kurt Gödel. En 1930, von Neumann est professeur-invité à l'université de Princeton. Puis, de 1933 à sa mort en 1957, il est professeur de mathématiques à la faculté de l'Institute for Advanced Study qui vient d'être créée. Il y rejoint donc Albert Einstein et Kurt Gödel. Neumann émigra aux États-Unis en 1933 pour rejoindre l'Institut de Recherche Avancée à Princeton. Il rédigea un important ouvrage sur la mathématique appliquée et effectua un travail majeur dans l'axiomatisation de la physique quantique (c'est lui qui réalisa système quantique peut être considéré comme un point dans un espace de Hilbert et qui introduisit les opérateurs linéaires). Il participa durant la deuxième guerre mondiale au développement théorique de la bombe atomique et à l'étude des ondes de choc. Ces travaux mathématiques sur les calculs ultra-rapides pour les simulations de la bombe H, contribua de façon non négligeable au développement de l'informatique (il est aussi à l'origine des méthodes de Monte-Carlo). Il contribua également à la théorie des jeux où certains de ces résultats eurent une grande influence sur l'économie.



Niels, Abel (1802-1829) était mathématicien norvégien né à Frindø et décédé à Froland. Son père était un éminent homme politique norvégien, mais à la fin de sa vie, il tomba en disgrâce, et quand il mourut en 1820, c'est Abel qui dut supporter toute la charge de la famille. Son père, éduqua lui-même Abel jusqu'en 1815, puis l'envoya au collège paroissial d'Oslo. Dans ce lycée, le latin, le grec et la religion étaient enseignés à l'ancienne, avec punitions et châtiments corporels. La situation évolua en 1817 à la suite du renvoi d'un professeur consécutif au décès d'un élève: le lycée recruta un jeune enseignant ouvert aux idées nouvelles et instruit

de mathématiques qui en découvrant l'intérêt de Niels pour les mathématiques, lui trouva une bourse pour l'université. Grâce à l'aide financière de ses professeurs, il parvient donc toutefois à poursuivre ses études et à faire ses premières découvertes. Mais ses mémoires sont perdus par Cauchy et mésestimés par Gauss. Après son doctorat, Abel ne parvint pas à trouver un poste et ses conditions de vie devinrent de plus en plus précaires et sa santé se fragilisa: il fut ainsi atteint de la tuberculose. Malgré des déplacements à Paris et à Berlin, ses travaux ne sont toujours pas parvenus à leur juste valeur. Dans ses dernières semaines, il n'a plus assez de force pour quitter son lit. Il décède à même pas 27 ans, alors qu'un ami venait juste de lui trouver un poste à Berlin. C'est Jacobi qui comprendra tout le génie de ce jeune mathématicien. Abel avait notamment démontré, à l'âge de 19 ans, l'impossibilité de résoudre par radicaux les équations algébriques de degré 5, ce que son contemporain Galois généralisera à tout degré. À titre posthume, Abel recevra en 1830 le grand prix de Mathématiques de l'Institut de France.



Noether, Emmy (1882 -1935) née à Erlangen et décédée à Princeton. Emmy envisage d'abord d'enseigner le français et l'anglais après avoir passé les examens requis, mais étudie finalement les mathématiques à l'université d'Erlangen où son père donne des conférences. Durant le semestre d'hiver 1903-1904, elle étudie à l'université de Göttingen et assiste aux cours de l'astronome Karl Schwarzschild et des mathématiciens Hermann Minkowski, Felix Klein et David Hilbert. Après avoir achevé sa thèse en 1907 elle travaille bénévolement à l'Institut de Mathématiques d'Erlangen pendant 7 ans. En 1915, elle est invitée par David Hilbert et Felix Klein à rejoindre le très renommé département de mathématiques de l'université de Göttingen jusqu'en 1933. En 1935, elle est opérée en raison d'un kyste ovarien et, malgré des signes de rétablissement, meurt 4 jours plus tard à l'âge de 53 ans. Elle reste dans l'histoire des mathématiques comme la fondatrice principale de l'Algèbre abstraite, ou algèbre moderne, qui est une des branches essentielles des mathématiques contemporaines. Cette algèbre abstraite prend de la hauteur par rapport aux calculs menés dans divers ensembles, munis de diverses opérations, et montre ce que ces calculs ont en commun. En physique, le théorème de Noether explique le lien fondamental entre la symétrie et les lois de conservation. Ses idées ont contribué aussi au progrès de la physique, en particulier dans la théorie de la Relativité. Malgré toutes ses qualités, elle eut des difficultés à mener une carrière normale de professeur d'université, car elle était une femme, dans un milieu exclusivement masculin. Elle bénéficia cependant de l'estime et de l'appui de David Hilbert, d'Albert Einstein et de Felix Klein.



Oppenheimer, J. Robert (1904-1967), physicien américain né à New-York et décédé à Princeton. Il a été le directeur scientifique du projet Manhattan et y dirigea donc la mise au point des premières bombes atomiques. Entré à Harvard avec une année de retard à cause d'une attaque de colite ulcéreuse, il profite de cette période pour se rendre, avec son ancien professeur d'anglais, au Nouveau-Mexique. Il y devint amateur de promenades à cheval ainsi que des montagnes et plateaux de cette région. À son retour, il obtient son diplôme de chimie en 3 ans. Percy Bridgman lui fait découvrir la physique expérimentale. C'est durant ses études au laboratoire Cavendish d'Ernest Rutherford à Cambridge, qu'il réalise qu'il maîtrise mieux la théorie que la pratique en raison de sa maladresse. En

1926, il poursuit ses études sous la direction de Max Born à l'Université de Göttingen, et obtient son doctorat à l'âge de 22 ans. À Göttingen, il publie des articles sur la théorie quantique. En 1927, il retourne à Harvard puis l'année suivante à l'Institut de technologie de Californie. Il est aussi connu pour sa contribution à la théorie quantique et à la théorie de la relativité, et pour ses études sur les rayons cosmiques, les positrons et les étoiles à neutrons. Il fait des recherches importantes en astrophysique, en physique nucléaire, et en spectroscopie. Il découvre alors l'approximation de Born-Oppenheimer.



Ostrogradsky, Mikhail Vassilievitch (1801-1862) était un physicien et mathématicien ukrainien. Il commença ses études de mathématiques à l'université de Kharkov, et les continua ensuite à Paris où il fut en contact étroit avec les célèbres mathématiciens français Cauchy, Binet, Fourier et Poisson. De retour dans sa patrie, il enseigna à l'école des cadets de la Marine, à l'Académie du génie Nicolas et à l'école d'Artillerie de Saint-Pétersbourg. Il est célèbre en particulier pour avoir établi le théorème de flux-divergence, qui permet d'exprimer l'intégrale sur un volume (ou intégrale triple) de la divergence d'un champ vectoriel comme l'intégrale de surface (intégrale double étendue à la superficie qui entoure ce volume) du flux défini par ce champ. Il fut élu à l'Académie américaine des arts et des sciences en 1834, à l'Académie des sciences de Turin en 1841, et à l'Académie des sciences de Rome en 1853. Enfin il fut élu membre correspondant de l'Académie des sciences de Paris en 1856. Les travaux scientifiques d'Ostrogradski sont dans le droit fil des principes professés à cette époque à l'École polytechnique dans les domaines de l'analyse, et des mathématiques appliquées. En physique mathématique, il imagina une synthèse grandiose qui embrasserait l'hydromécanique, la théorie de l'élasticité, la théorie de la chaleur, et la théorie de l'électricité dans le cadre d'une seule méthode homogène.



Pareto, Vilfredo (1848-1923), économiste et sociologue italien, dont la contribution la plus célèbre à la théorie économique est la définition du concept d'optimum économique. Né à Paris d'un père italien en exil et d'une mère française, il retourna en Italie à l'âge de 10 ans. Il fit ses études à l'université de Turin et devint ingénieur. En 1893, il fut nommé à la chaire d'économie politique de l'université de Lausanne (il décéda à Céligny en Suisse), où il succédait à Léon Walras. Parmi ses travaux figure l'analyse des anticipations des agents économiques. Celles-ci, n'étant pas indépendantes les unes des autres, peuvent susciter des mouvements d'opinion pessimistes qui génèrent des crises. Pareto est également le père de la notion d'optimum. L'économie est à un optimum lorsqu'on ne peut améliorer la situation d'un agent sans détériorer celle d'au moins un autre agent. Ce concept est très utilisé en économie, car il permet de prendre en compte la non-additivité des utilités des différents agents. La concurrence permet d'atteindre l'optimum au sens de Pareto. Pareto a également intégré les courbes d'indifférence (formalisées par Francis Edgeworth) à la logique walrassienne d'équilibre général. Le travail sociologique de Pareto fut plus discuté. Dans le *Traité de sociologie générale*, paru en 1916, il présenta sa théorie des élites, selon laquelle le pouvoir d'État est dans toutes les sociétés l'objet d'un combat entre les seules élites. Cette thèse discréditait les démocraties, et contribua implicitement au développement du fascisme alors montant en Italie.



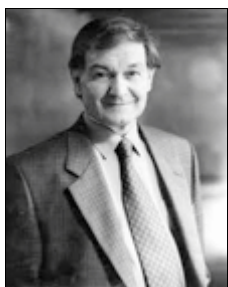
Pascal, Blaise (1623-1662), mathématicien, physicien, théologien, mystique, philosophe, moraliste et polémiste né à Clermont et décédé à Paris. Enfant précoce (à 11 ans, il compose un court *Traité des sons* des corps vibrants et aurait démontré la 32^e proposition du I^{er} livre d'Euclide, à 16 ans il écrit un traité sur les *Coniques*), il est éduqué par son père qui était mathématicien. Les tout premiers travaux de Pascal concernent les sciences naturelles et appliquées. Il contribua de manière importante à l'étude des fluides. Il a clarifié les concepts de pression et de vide, en étendant le travail de Torricelli. L'étendue des domaines d'intérêt et du génie de Pascal est impressionnante: inventeur de la machine à calculer, concepteur des premiers transports en commun en France, artisan de l'assèchement des marais poitevins il fut également l'un des plus brillants prosateurs de la langue française et l'une des plus grandes figures du 17^{ème} siècle français.



Pauli, Wolfgang (1900-1958) était un physicien autrichien, né à Vienne et décédé à Zürich, connu pour sa définition du principe d'exclusion en mécanique quantique, ce qui lui valut le prix Nobel de physique de 1945. Pauli est né d'un père professeur des universités et d'une mère journaliste et juriste. Au lycée à Vienne, Pauli était considéré comme un enfant prodige en mathématiques. À partir de 1919, il commence ses études de physique à l'université de Munich avec pour professeur Arnold Sommerfeld. Depuis 1898, Sommerfeld était en charge d'écrire le cinquième volume de la *Enzyklopädie der mathematischen Wissenschaften* (20'000 pages), consacré à la physique. Il requiert dans un premier temps la collaboration d'Albert Einstein pour rédiger l'article sur la relativité, mais ce dernier refuse. Sommerfeld fait alors appel à Pauli, dont la relativité était la spécialité lors de son inscription aux cours de Sommerfeld. C'est ainsi qu'à 21 ans, Pauli publie son article de synthèse des théories de la relativité restreinte et de la relativité générale. En 1921, il obtient son doctorat avec pour sujet l'atome d'hydrogène qui montra clairement la limite du modèle de l'atome de Bohr, auquel il travaillera en tant qu'assistant de Max Born à Göttingen entre 1921 et 1922. Pendant les années 1922 et 1923, il travailla aux côtés de Niels Bohr à Copenhague. Entre 1923 et 1928, il enseigna à Hambourg avant de partir à l'ETH de Zurich, où il obtint un poste de professeur de physique théorique. À partir de 1935, il est parti pour les États-Unis, où il occupe des postes de professeur invité, notamment à l'Institute for Advanced Study à Princeton durant les années 1935-1936, mais aussi à l'Université du Michigan, en 1931 et 1941, et l'Université Purdue, en 1942. En 1946, il obtient la citoyenneté américaine, mais revient la même année à l'ETH de Zurich, où une place de professeur lui avait été gardée. En 1949, il devient citoyen suisse. Dans les années 1950, il retourne régulièrement à Princeton afin de donner des cours en tant que professeur invité. Dans les dernières années de sa vie, il participa à la fondation du CERN. Il meurt d'un ulcère gastro-duodéal.



Pearson, Karl (1857-1936) est un mathématicien britannique né à Londres et décédé à Surrey fondateur des statistiques modernes. L'analyse statistique connaît un grand développement à la fin du 19^{ème} siècle au Royaume-Uni et Karl Pearson domine ses contemporains par l'étendue et la variété de ses contributions bien que s'étant intéressés aux statistiques seulement à partir de l'âge de 33 ans. Il développe des méthodes d'analyse pour l'étude de la sélection naturelle et de l'eugénisme dont il est un ardent promoteur. Ses principales contributions sont la création du test du d'indépendance du Khi-deux destiné à estimer si les écarts observés dans un ensemble de variables par rapport aux valeurs théoriques peuvent être attribués ou non à un échantillonnage au hasard et la définition du coefficient de corrélation. Il reçoit la médaille Darwin (biologie) en 1898. Pearson était aussi consultant dans les entreprises. Il a entre autres donné des cours à William S. Gosset qui introduisit la loi de Student en 1910. Il est l'un des fondateurs de la revue *Biometrika* dont il a été l'éditeur pendant 36 ans et qu'il a hissé au rang de meilleure revue de statistiques mathématiques.

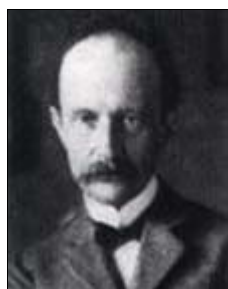


Penrose, Roger (1931-) physicien et mathématicien né à Colchester. Penrose obtient son diplôme en mathématiques de l'Université du Collège de Londres et son doctorat à l'université de Cambridge avec une thèse sur les méthodes tensorielles en géométrie algébrique. Entre 1964 et 1973, il enseigne la mathématique au Birkbeck College de Londres et rencontre le célèbre physicien Stephen W. Hawking avec lequel il travailla sur une théorie de l'origine de l'Univers en apportant sa contribution mathématique à la théorie de la relativité générale appliquée à la cosmologie et à l'étude des trous noirs. En 1965, à Cambridge, il prouve que des singularités gravitationnelles peuvent être formées à partir de l'effondrement gravitationnel d'étoiles massives en fin de vie. En 1971, Penrose découvre les réseaux de spin qui devaient plus tard former la géométrie de l'espace-temps dans la théorie quantique à boucles. Professeur à Oxford, il reçut, avec Hawking, le prix Wolf 1988 pour la physique.



Picard, Charles-Émile (1856-1941) né et décédé à Paris fait ses études classiques au lycée de Vanves dès 1864, puis au lycée Napoléon (futur lycée Henry IV) de 1868 à 1874 où il se révèle excellent élève, mais peu attiré par la mathématique. Il obtient cette année-là le baccalauréat ès lettres puis l'année suivante le baccalauréat ès sciences. Il est reçu second à l'École Polytechnique, et premier à l'École Normale Supérieure. Finalement, passionné par les sciences, il opte pour cette dernière, où il prépare l'Agrégation qu'il réussit en 1877. Après divers postes d'assistant à Paris et Toulouse, il devient en 1881 Maître de Conférences à l'École Normale Supérieure. Son nom est déjà célèbre dans le cercle des mathématiciens, car il a démontré un théorème important sur les singularités des fonctions holomorphe qui lui vaudra une nomination pour devenir membre de l'Académie des Sciences. Il est cependant trop jeune, et son élection est reportée en 1889. En 1885, Picard devient professeur à la Sorbonne, où il occupe la chaire de calcul différentiel. Là encore, son jeune âge est une gêne (il faut avoir au minimum 30 ans pour occuper un tel poste) et il faut utiliser une procédure astucieuse pour contourner la législation. Plus tard, Picard occupera la chaire d'analyse et d'algèbre, et il exercera aussi à l'École Centrale des Arts et

Manufacture (1894-1937): il y forme à la mécanique plus de 10'000 ingénieurs, et est, selon Hadamard, un excellent professeur. Les travaux de Picard sont ardues, et ouvrent la voie à de nouvelles recherches. Il est le premier à utiliser le théorème du point fixe dans une méthode d'approximations successives qui permet de résoudre des équations aux dérivées partielles. On lui doit également des travaux en géométrie algébrique, comme des recherches plus appliquées sur l'élasticité ou la chaleur. Il est aussi l'un des premiers défenseurs des théories d'Einstein. Son *Traité d'Analyse* constitua longtemps une référence, et Picard fut aussi philosophe et historien des sciences. Parmi les distinctions que Picard a reçues, citons qu'il présida le congrès International des mathématiciens, qu'il fut élu membre de l'Académie Française en 1924, et qu'il reçut la médaille d'or Mittag-Leffler en 1937.



Planck, Max (1858-1947) était un physicien allemand né à Kiel et décédé à Göttingen considéré comme le fondateur de la physique quantique. Après avoir obtenu son baccalauréat à 17 ans à Munich où son père enseigne, Max Planck poursuit ses études de physique à Berlin. Passionné par la thermodynamique, il soutient une thèse de doctorat sur le second principe de la thermodynamique et la notion d'entropie 1879, notion qui restera le moteur explicatif de la majorité de ses recherches. L'année suivante, il devient maître de conférence à l'université de Munich puis obtient la chaire de physique de l'université de Kiel en 1885. Quatre ans plus tard, il est professeur de physique à l'université de Berlin, poste qu'il occupera pendant près de 40 ans. En 1930 il prend la direction de l'Institut Kaiser Wilhelm pour la recherche scientifique qui portera bientôt son nom. Amorcées par sa thèse de doctorat, les recherches de Planck en thermodynamique se portent rapidement sur le corps noir. Entité purement théorique, le corps noir absorbe toutes les radiations qu'il reçoit (le noir de carbone, en absorbant 97% du rayonnement, se rapproche de cet idéal). Pour tenter d'expliquer ce phénomène, Planck élabore une nouvelle théorie. Il émet l'hypothèse que l'énergie d'un rayonnement ne peut être émise ou absorbée par la matière que par quantités finies, les quanta. Il montre alors que ces "paquets d'énergie" ont pour valeur $h\nu$, où ν est la fréquence du rayonnement et h une constante universelle (la "constante de Planck"). En exposant sa théorie à la Société allemande de physique en 1900, à Berlin, Planck ne sait pas encore qu'il vient d'inventer une nouvelle branche de la physique: la physique quantique. Sa découverte entraînera alors la création du modèle de l'atome par Niels Bohr, l'élaboration de la mécanique ondulatoire par Louis de Broglie, l'explication du phénomène photoélectrique par Albert Einstein ou encore la découverte du principe d'incertitude par Werner Heisenberg. Considéré comme l'un des plus célèbres physiciens, Planck recevra le prix Nobel en 1918.



Poincaré, Henri (1854-1912), mathématicien et physicien français né à Nancy et décédé à Paris dont on a dit qu'il était le dernier savant susceptible de connaître la totalité des mathématiques de son temps. Élève d'exception au Lycée Impérial de Nancy, il obtient en 1871, le baccalauréat ès lettres, mention Bien, et la même année son baccalauréat ès sciences. Il se classe premier au concours d'entrée à l'École polytechnique en 1873, puis à l'École des Mines de Paris, comme ingénieur du Corps des Mines, en 1875. Il est licencié ès sciences en 1876. Nommé ingénieur des mines de troisième classe en 1879 à Vesoul, il obtient, la même année le doctorat ès sciences mathématiques à la Faculté des sciences de Paris, et devient chargé de cours d'analyse à la faculté des sciences de Caen. Les premiers travaux de Poincaré portent sur les fonctions automorphes ou fuchsiennes, la théorie qualitative des équations différentielles et la théorie des fonctions. Dans une série de 6 articles publiés à partir de 1894, il est le créateur de la topologie algébrique, science en pleine expansion au 20ème siècle et dans laquelle plusieurs conjectures dues à Poincaré restent ouvertes. Il s'est vivement intéressé à la mécanique céleste: *Les Méthodes nouvelles de la mécanique céleste*, trois volumes parus entre 1892 et 1899, annoncent les recherches modernes sur les systèmes dynamiques et le chaos. En physique-mathématique, il dégagait les propriétés du groupe de Poincaré-Lorenz, qui allaient quelques mois plus tard conduire à l'article fondamental d'Einstein sur la relativité restreinte.



Poisson, Siméon Denis (1781-1840) était un mathématicien français dont les travaux ont porté sur les intégrales définies, la théorie électromagnétique et le calcul des probabilités. Sa famille le força à faire des études de médecine qu'il abandonna, en 1798, pour aller étudier la mathématique à l'École polytechnique, où il fut l'élève de Laplace et Lagrange, qui devinrent l'un et l'autre ses amis. Il enseigna à l'École polytechnique à partir de 1802 et en 1808, il fut nommé astronome du Bureau des longitudes et, à sa création, en 1809, professeur à la Faculté des sciences. Les travaux les plus importants de Poisson portent sur les applications des mathématiques à la physique et à la mécanique. Son *Traité de mécanique* a été l'ouvrage de référence en mécanique pendant de nombreuses années. Un mémoire, publié en 1812, contient les lois les plus usuelles de l'électrostatique et la théorie selon laquelle l'électricité est constituée de deux fluides dont les éléments semblables se repoussent, tandis que les éléments différents s'attirent. En mathématiques pures, il a publié une série d'articles sur les intégrales définies, et ses recherches sur les séries de Fourier ont annoncé celles de Dirichlet et de Riemann sur ce sujet. C'est dans l'ouvrage *Recherches sur la probabilité des jugements...* (1837), qui est un livre important sur le calcul des probabilités, qu'apparaît pour la première fois la distribution de Poisson (ou "loi de Poisson"). Obtenue initialement comme une approximation de la loi binomiale de Bernoulli, elle est devenue fondamentale dans de très nombreux problèmes. Les autres publications de Poisson comprennent *Théorie nouvelle de l'action* (1831) et *Théorie mathématique de la chaleur* (1835). Le nom de Poisson est attaché à de nombreuses notions mathématiques et physiques (intégrale et équation de Poisson en théorie du potentiel, crochets de Poisson dans la théorie des équations différentielles, rapport de Poisson en élasticité et constante de Poisson en électricité).



Ramanujan, Srivanasa (1887-1920) est né à Erode, un petit village situé 400 km au sud de Madras, dans une famille pauvre de la caste des Brahmanes. Il passe son enfance dans la ville de Kumbakonam, où son père exerce le métier de comptable chez un drapier. À partir de l'âge de 5 ans, il fréquente différentes écoles primaires avant de pouvoir intégrer la Town High School en 1898. En 1900, il commence à développer ses propres mathématiques en se basant sur son premier livre de mathématiques, *La Trigonométrie plane*. Il définit seul une méthode pour résoudre les équations du 3^e degré, puis du 4^e, puis il tente aussi de résoudre celles du 5^e degré, ignorant qu'elles ne peuvent être résolues par les radicaux. On est alors en 1902 et c'est à cette époque que Ramanujan se procure le second (et dernier!) livre dans lequel il puisera ses connaissances mathématiques de bases, *Synopsis of elementary results in pure mathematics*, compilation d'environ 6'000 théorèmes et autres formules par G.S. Carr. Ce livre étant essentiellement un livre de résultats, la plupart sans démonstrations, influencera le style futur de Ramanujan, qui n'a laissé que très peu de preuves de ses propres résultats. À 17 ans, sa démarche est déjà celle d'un chercheur en mathématiques. Comme ses résultats scolaires sont bons, il reçoit une bourse lui permettant d'entrer au Government College de Kumbakonam en 1904. Cependant, il consacre trop de temps à ses recherches en mathématiques et néglige les autres matières, ce qui lui vaut la suppression de cette bourse l'année suivante. Sans argent, il part, à l'insu de ses parents, pour la ville de Vizagapatnam où il poursuit ses travaux sur les séries hypergéométriques et les relations entre intégrales et séries. En 1906, il retourne à nouveau au lycée, à Madras cette fois-ci, avec l'idée de passer un examen lui permettant d'entrer à l'université. Il assiste quelques mois aux cours puis tombe malade. Au cours de l'examen, il réussit seulement en mathématiques et échoue partout ailleurs, ce qui lui interdit l'entrée à l'université de Madras. Dans les années qui suivent, il continue alors de développer seul ses idées, sans aucune aide extérieure et sans connaissance des thèmes de recherche possibles, en dehors de ceux découlant des notions abordées dans le livre de Carr. Ramanujan étudie ainsi les fractions continues et les séries divergentes en 1908. Il tombe alors de nouveau très malade et doit subir, en 1909, une opération dont il aura du mal à se remettre. Il commence alors de poser et de résoudre des problèmes mathématiques dans le journal de la Société Indienne de Mathématiques (SIM). En 1910, il développe des relations sur les équations modulaires elliptiques. Un an plus tard, la publication d'un article brillant sur les nombres de Bernoulli dans ce même journal lui vaut la reconnaissance de son travail par ses pairs. Bien qu'il ne possède aucun diplôme universitaire, il acquiert la réputation de génie des mathématiques dans la région de Madras. La même année, il rencontre le fondateur de la SIM, qui lui permet d'obtenir un emploi temporaire chez un comptable de Madras et lui conseille de contacter Ramachandra Rao, un mécène membre de la SIM. Grâce à cette lettre, Ramanujan obtient le poste et commence son travail en 1912. Il a alors la chance d'être entouré de personnes ayant une formation en mathématiques et qui s'intéressent à son travail. Le chef comptable du port de Madras est un mathématicien qui publie un article sur le travail de Ramanujan en 1913, *On the distribution of primes*. D'autre part, un professeur du Madras Engineering College est intéressé par les capacités de Ramanujan. Ayant lui-même fait ses études à

Londres, il écrit à un de ses professeurs de mathématiques, à qui il envoie de quelques résultats de Ramanujan. L'Université de Madras allouera plus tard une bourse à Ramanujan en 1913 et en 1914, Hardy le fait venir au Trinity College, à Cambridge. C'est le début d'une extraordinaire collaboration entre les deux hommes. En 1916, il obtient le titre de docteur de l'université de Cambridge, malgré qu'il ne possède pas les diplômes requis pour préparer une thèse. En 1918, Ramanujan est élu membre de la Cambridge Philosophical Society. Trois jours plus tard, probablement le plus grand honneur de toute sa carrière, son nom apparaît sur la liste des élections des membres de la "Royal Society of London". Il a été proposé par une liste impressionnante d'éminents mathématiciens. Son élection a effectivement lieu en 1918 et il est également élu membre du Trinity College pour 6 ans. Ramanujan repart pour l'Inde en 1919. Cependant, son état de santé déjà très mauvais ne cesse de se dégrader. Il meurt l'année suivante probablement à cause de graves carences alimentaires. Ramanujan a laissé derrière lui un grand nombre de cahiers non-publiés (les fameux *Carnets de Ramanujan*), remplis de théorèmes que les mathématiciens continuent d'étudier. Aujourd'hui, ses travaux ont bien sûr des applications en physique théorique.



Ricci-Curbastro, Gregorio (1853-1925) né à Lugo et décédé à Boulogne était un mathématicien spécialiste de la géométrie différentielle et l'un des pères du calcul tensoriel. Après des études de philosophie et de mathématiques, Ricci soutient sa thèse de doctorat à l'université de Pise. En 1880, il sera nommé professeur de physique-mathématique à l'université de Padoue. Levi-Civita fut son élève et contribua avec Ricci à l'élaboration de son calcul différentiel absolu (1900) visant à expliciter en mécanique, dans des espaces abstraits (variétés différentiables), des relations indépendantes du système de coordonnées utilisé, inhérentes au phénomène étudié (invariants différentiels). Associée à la géométrie différentielle de Gauss et de Riemann, le célèbre physicien Albert Einstein trouva, dans cette nouvelle approche de la mécanique qu'il nomma "calcul tensoriel" (1916), les outils mathématiques nécessaires à sa théorie de la relativité générale.



Riemann, Georg Friedrich Bernhard (1826-1866) était un mathématicien allemand. Au lycée, Riemann étudie la Bible intensivement, mais il est distrait par les mathématiques. Il essaie même de prouver, mathématiquement, l'exactitude de la Genèse. Ses professeurs sont surpris par ses capacités à résoudre des problèmes complexes en mathématique. En 1846, grâce à l'argent de sa famille, il commence à étudier la philosophie et la théologie pour devenir prêtre afin de financer sa famille. En 1847, son père l'autorise à étudier les mathématiques. Il étudie d'abord à l'université de Göttingen où il rencontre Carl Friedrich Gauss, puis à l'université de Berlin, où il a entre autres comme professeurs Jacobi, Steiner et Dirichlet. Dans sa thèse, présentée en 1851 sous la direction de Gauss, Riemann met au point la théorie des fonctions d'une variable complexe. En 1854 il donne un exposé qui jette les bases de la géométrie différentielle. Il y introduit la bonne façon d'étendre à n dimensions les résultats de Gauss lui-même sur les surfaces. Cette présentation a profondément changé la conception de la notion de géométrie, notamment en ouvrant la voie aux géométries non euclidiennes et à la théorie de la relativité générale. On lui doit également d'importants travaux sur les intégrales, poursuivant ceux de Cauchy, qui ont donné entre autres ce qu'on

appelle aujourd'hui les intégrales de Riemann. Intéressé par la dynamique des gaz, il jette les bases de l'analyse des équations aux dérivées partielles de type hyperbolique. Il succéda à Dirichlet sur la chaire de Gauss en 1859. À 39 ans, il fut emporté par la tuberculose.



Salam, Abdus (1926- 1996), physicien pakistanais, a reçu le prix Nobel de physique en 1979 pour ses travaux sur l'interaction électrofaible, synthèse de l'électromagnétisme et de l'interaction faible. Né à Jhang Sadar, il étudie au Government College à Lahore. À l'âge de 14 ans, Salam obtint les meilleurs notes jamais enregistrées pour l'examen d'entrée à l'Université du Punjab. Persécuté par la majorité musulmane de son pays pour son appartenance religieuse (ahmadiste), il doit fuir son pays. Réfugié en Grande-Bretagne, il obtient en 1952, un doctorat en mathématiques et en physique de l'université de Cambridge. Sa thèse doctorale fut une étude fondamentale en électrodynamique quantique. Ses travaux le rendirent célèbre internationalement. Il retourna au Government College de Lahore en tant que professeur de mathématiques, garda cette fonction de 1951 à 1954, puis retourna ensuite à Cambridge en tant que conférencier en Mathématiques. Il enseigne dans ces établissements, puis en 1957, est nommé professeur de physique théorique à l'Imperial College de Londres. Il y demeura jusqu'à sa retraite. En 1959, il devient le plus jeune membre de la Royal Society à l'âge de 33 ans. Durant les années 1960, Salam joua un rôle important dans l'établissement de l'agence de recherche nucléaire du Pakistan, et de l'agence de recherche spatiale du Pakistan, de laquelle il fut le Directeur fondateur. En 1964, il devient directeur du Centre international de physique théorique de Trieste, nouvellement créé. Cette même année, il est lauréat de la Médaille Hughes. En 1967, avec le physicien Steven Weinberg, Salam propose une théorie permettant d'unifier les interactions électromagnétiques et faibles entre particules élémentaires, théorie qui sera confirmée par l'expérience. Salam sera ainsi le premier musulman à obtenir le prix Nobel de physique en 1979, conjointement aux physiciens Sheldon Lee Glashow et Weinberg.



Say, Jean-Baptiste (1767- 1832) était un économiste, journaliste et industriel français né à Lyon et décédé à Paris. Il est issu d'une famille de négociants nîmois ayant émigré à Amsterdam puis à Genève. C'est au cours d'un voyage en Grande-Bretagne, où la révolution industrielle est en cours, qu'il adoptera les idées libérales et en particulier les théories d'Adam Smith dont il sera un ardent défenseur de retour en France. En 1789, il publie la brochure: *la Liberté de la presse*. En 1792, il participe aux campagnes militaires de la Révolution française en Champagne. D'abord employé dans une banque, il dirigea ensuite une filature de coton à Auchy-lès-Hesdin, dans le Pas-de-Calais. Ses nombreux ouvrages d'économie politique firent qu'il fut nommé professeur au Conservatoire national des arts et métiers en 1821, puis au Collège de France en 1830. La "loi de Say", ou "loi des débouchés", stipule que plus les producteurs sont nombreux et les productions multiples, plus les débouchés sont faciles, variés et vastes. Dans une économie où la concurrence est libre et parfaite, les crises de surproduction sont impossibles. Il ne peut y avoir de déséquilibre global dans les économies de marché et de libre entreprise, il y a un équilibre spontané des flux économiques (production, consommation, épargne, investissement). Cette loi est parfois réduite

à tort à la formule: toute offre crée sa propre demande. Un meilleur résumé de cette approche serait: on dépense que l'argent qu'on a gagné. L'économie de l'offre, dans la tradition de Say, s'oppose à l'économie de la demande, qui est celle de Malthus et plus tard de Keynes.



Schaefer, Milner Baily (1912-1970) né au Wyoming et décédé à San Diego a étudié à l'Université de Washington où il avait obtenu un baccalauréat en science en 1935. Dès l'obtention du baccalauréat, il a travaillé au département des pêches de l'état de Washington à Seattle. De 1937 à 1942, il a travaillé à la Commission des pêches du saumon du Pacifique à Westminster, Colombie-Britannique. Il a servi dans la marine durant la guerre et par la suite, il a occupé divers postes en tant que biologiste spécialiste des pêches. Après avoir terminé son doctorat à l'Université de Washington, en 1950, Schaefer est devenu directeur des enquêtes de l'IATTC (Inter-American Tropical Tuna Commission), une commission internationale des pêches. Pendant les dix années qui ont suivi, il a travaillé sur la théorie de la dynamique de la pêche et mis au point un modèle de population des espèces marines qui est connu sous le nom de "modèle de Schaefer". Au cours des années 1950, Schaefer est devenu de plus en plus impliqué dans plusieurs comités, groupes et organisations concernés par les ressources marines, en particulier la pêche et tous les aspects de l'océanographie. Durant cette période, il a donné des cours sur la dynamique et l'exploitation des populations de poissons. En 1962, il a démissionné de son poste de directeur des enquêtes à l'IATTC pour occuper le poste de directeur de l'Institut des Ressources marines de l'Université de Californie tout en agissant comme conseiller scientifique pour l'IATTC.

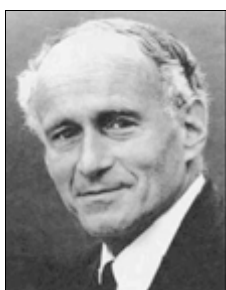


Scholes, Myron (1941-) né à Ontario, il présente son doctorat en 1969 à l'université de Chicago. Il occupe en 1988 la chaire Frank E. Buck de professeur de finance au Graduate School of Business de l'université Stanford (Californie) où il dirige également des recherches pour l'Institution Hoover. Il a reçu le prix Nobel d'économie en 1997 pour avoir élaboré, avec Fischer Black, une méthode d'évaluation des instruments financiers dérivés (résultat mathématique novateur pour estimer les risques liés aux options sur actions) ayant ouvert de nouveaux horizons au champ des évaluations économiques. Le baccalauréat de Myron Scholes, Robert Merton, a joué un rôle très important dans l'élaboration de cette méthode d'évaluation ainsi que dans les applications qu'elle a permises pour améliorer la gestion des risques attachés aux nouveaux produits financiers. Déjà en 1900, Louis Bachelier, présentait à la Sorbonne une thèse de doctorat au titre visionnaire: *Théorie de la spéculation*. Dans les années 1960, des auteurs tels James Boness et Paul Samuelson (Prix Nobel d'économie en 1970) proposaient des modèles pour déterminer les prix d'équilibre des options. Leurs hypothèses ne se sont pas révélées suffisamment réalistes pour entraîner des applications, mais des améliorations apportées à ces modèles au début des années 1970 ont permis d'obtenir des résultats plus satisfaisants. C'est en 1973 que Scholes et Black mettent leurs compétences en commun et proposent la première version de la formule de calcul du prix des options qui leur vaudra le prix Nobel. Si Myron Scholes et Fischer Black ont eu l'intuition fondamentale de la démonstration, ils ont pris pour base de recherche le modèle d'équilibre des actifs financiers (ou

Capital Asset Pricing Model, dit C.A.P.M.) de leur compatriote William Sharpe récompensé à ce titre par le jury du Nobel en 1990 (les deux autres lauréats étaient Harry Markowitz et Merton Miller).



Schrödinger, Erwin (1887-1961) Né et décédé à Vienne il entre au Gymnase de cette même ville en 1898. Presque depuis son premier jour de classe jusqu'à son départ du lycée 8 ans plus tard, Schrödinger fut un excellent élève. Il était toujours premier de sa classe en travaillant dur chez lui entre les quatre murs de son bureau personnel.. Il poursuit ses études à l'université d'Iéna. En 1920, il est nommé professeur à la Haute École technique de Stuttgart puis à l'université de Breslau l'année suivante. En 1927, il succède à Max Planck à l'université de Berlin. Israélite, il quitte le pays à l'avènement du national-socialisme pour se rendre à Oxford où il obtient une chaire en 1933. Sept ans plus tard, il devient professeur de physique théorique à Dublin à l'Institut des hautes études de l'État libre d'Irlande. Il ne rentrera en Autriche qu'en 1956. Schrödinger comme son contemporain Albert Einstein avait horreur d'apprendre par coeur et d'être forcé de retenir des faits inutiles. Les premiers travaux de Schrödinger portent sur l'étude des couleurs et la théorie des quanta. Mais il est avant tout reconnu pour ses recherches en mécanique ondulatoire, discipline développée par le Français Louis de Broglie. L'équation de Schrödinger, élaborée en 1926, permet de calculer la fonction d'onde d'une particule se déplaçant dans un champ. En établissant cette équation de propagation, il donne à la mécanique quantique un outil intuitif aujourd'hui indispensable (au contraire de l'approche matricielle et abstraite de Heisenberg) qu'Einstein qualifia d'idée de Génie. Avec celle de Werner Heisenberg, la théorie de Schrödinger constitue ainsi la base de la mécanique quantique. En 1933, Schrödinger partage le prix Nobel de physique avec Paul Dirac pour leur contribution au développement de cette nouvelle discipline. Schrödinger essaiera également d'appliquer sa théorie à la biologie et à la génétique dans ses ouvrages *What is life* (1944) et *Science and Humanism* (1951).



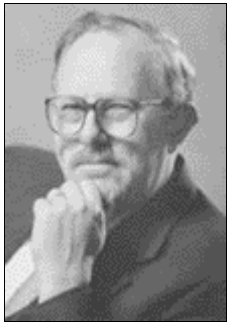
Schwartz, Laurent (1915-2002) Mathématicien français né et décédé à Paris. Ses travaux sont principalement relatifs à l'analyse. Ancien élève de l'École normale supérieure, Laurent Schwartz a enseigné de 1959 à 1960 et de 1963 à 1983 à l'École polytechnique. En 1975, il est élu membre de l'Académie des sciences. Sa thèse (1943) porte sur l'approximation et l'étude des sommes d'exponentielles. La théorie des distributions, dont l'idée initiale remonte à 1945, lui a valu la médaille Fields en 1950. Le langage et les notations de Schwartz pour les distributions ont été adoptées par les mathématiciens et constituent le cadre naturel de la théorie des équations aux dérivées partielles. De 1959 à 1962, Schwartz se consacre à la physique théorique: l'emploi des distributions lui permet une formulation mathématique correcte de la théorie des particules élémentaires. Il a aussi effectué des recherches sur les mesures de Radon et sur les espaces topologiques quelconques ; il a écrit diverses publications sur les probabilités cylindriques et les désintégrations de mesures.



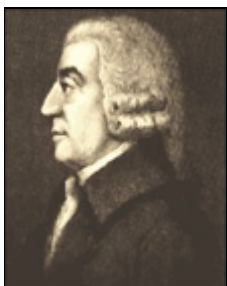
Schwarzschild, Karl (1873-1916) était un astronome mathématicien et physicien né à Francfort et décédé à Postdam qui prédit l'existence des trous noirs. Sa curiosité pour les étoiles se manifesta dès ses premières années scolaires, lorsqu'il construisit un petit télescope. Témoin de cet intérêt, son père le présenta à un ami mathématicien qui avait un observatoire privé. Schwarzschild apprit à utiliser un télescope et étudia des mathématiques plus avancées qu'à l'école. Il devint célèbre avec ses deux premiers articles sur la théorie des orbites publiés à l'âge de 16 ans alors qu'il était encore au collège. Il étudia à l'université de Strasbourg, puis de Munich, et obtint son doctorat à l'âge de 23 ans pour des travaux sur les théories de Henri Poincaré. Il fut alors engagé en tant qu'assistant à l'Observatoire Kuffner à Ottakring. Il se consacra principalement à la photométrie: il accomplit un travail de pionnier pour améliorer les plaques photographiques et implanter leur utilisation en astronomie, ainsi que dans l'étude spectrale des étoiles. De 1901 à 1909, il officia comme professeur au prestigieux institut de Göttingen, où il eut l'occasion de travailler avec des personnalités telles que David Hilbert et Hermann Minkowski. Il occupa ensuite un poste à l'Observatoire d'astrophysique de Potsdam en 1909. Schwarzschild est surtout connu pour ses contributions théoriques, tant en physique du Soleil qu'en relativité générale, ou en cinématique stellaire, ainsi que dans divers domaines de l'astrophysique. En 1916, il détermina une grandeur, dite "rayon de Schwarzschild", dans le cadre de la théorie de la relativité, énoncée peu de temps avant par Albert Einstein. Lorsqu'une étoile suffisamment massive explose en supernova, la contraction gravitationnelle produit ce que l'on appelle un "trou noir": rien, pas même la lumière, ne peut sortir de ce champ de gravitation intense. Lorsque le rayon d'une masse gazeuse devient inférieur au "rayon de Schwarzschild" pour cette masse, elle s'effondre en trou noir.



Shannon, Claude Elwood (1916-2001) né au Michigan et décédé au Massachusetts était un mathématicien spécialiste en mathématiques appliquées et ingénieur électricien, qui développa la théorie de la communication, aujourd'hui connue sous le nom de la "théorie de l'information". Shannon suivit les cours de l'université du Michigan et obtint en 1940 son doctorat de l'institut de technologie du Massachusetts, de la faculté duquel il devint un membre, en 1956, après avoir travaillé aux laboratoires de téléphone Bell. En 1949, Shannon publia la *Théorie mathématique de la communication*, un article dans lequel il présenta son concept initial pour une théorie unificatrice de la transmission et du traitement des informations. Les informations, selon cette théorie, incluent toutes formes de messages transmis, y compris ceux envoyés le long des canaux nerveux des organismes vivants. La théorie de l'information est aujourd'hui importante dans de nombreux domaines.



Sharpe, William Forsyth (1934-) est un économiste né à Boston. L'Académie royale des sciences de Suède a décerné en 1990 le prix Nobel de sciences économiques à 3 professeurs américains: Harry Markowitz, Merton Miller et William Sharpe. Même si les travaux récompensés étaient déjà anciens et se situent pour l'essentiel entre 1950 et 1970, l'Académie a jugé que les lauréats étaient des novateurs dans le domaine de la théorie de l'économie financière et du financement des entreprises. Ils ont en effet tous contribué à faire sortir de l'ombre de quelques universités américaines, une nouvelle discipline: la finance. C'était la première fois que l'Académie royale de Suède récompensait des travaux traitant des marchés boursiers et de la gestion de portefeuilles plutôt que des grands équilibres économiques. William Sharpe, de l'université Stanford, fut récompensé pour son modèle d'équilibre des actifs financiers et pour ses travaux sur la théorie de la formation des prix des avoirs financiers. Il s'est aussi engagé dans ses recherches dans la voie ouverte par Harry Markowitz. Ce dernier avait en effet élaboré une procédure complexe de sélection des titres boursiers afin d'optimiser un portefeuille de placements. Mais la mise en oeuvre de ce modèle a très vite posé des problèmes d'ordre pratique, au point que la collecte des informations nécessaires et leur traitement devenaient presque impossibles avec les ordinateurs disponibles dans les années 1960. C'est la raison pour laquelle William Sharpe se mit à chercher une méthode de sélection des portefeuilles efficaces plus simple. Il découvre que les variations de la rentabilité de chaque titre sont liées, linéairement, à la variation du marché dans son ensemble, mesurée par l'indice du marché concerné (par exemple l'indice Standard & Poor 500 aux États-Unis, ou le C.A.C. 40 en France). Le nombre de statistiques nécessaires s'en est trouvé fortement réduit: 302 statistiques au lieu de 3'150 dans le modèle Markowitz pour 100 titres, 602 au lieu de 20'300 pour 200 titres et 10'002 au lieu de 125'750 pour 300 titres, le calcul fut aussitôt facilité. C'est à partir de ce concept, simple en apparence, que Sharpe découvre ensuite le fameux coefficient Bêta reliant la rentabilité d'un titre à celle de l'indice du marché et constituant une mesure du risque associé à la volatilité du marché. Au-delà de leur apport pratique, les travaux de Sharpe ont contribué de façon décisive à la formulation d'une théorie de la formation des cours des actifs financiers plus connue sous le nom de "modèle C.A.P." (Capital Asset Pricing) ou, en français, de "Modèle d'équilibre des actifs financiers (MEDAF)".



Smith, Adam (1723-1790), économiste et philosophe écossais, est né à Kirkcaldy et décédé à Edimbourg, en Écosse. Il étudia aux universités de Glasgow et Oxford. De 1748 à 1751, il enseigna la rhétorique et les belles-lettres à Edimbourg. Durant cette période, il se lia avec le philosophe David Hume, dont la pensée exerça une grande influence sur les conceptions de Smith en matière d'éthique et d'économie. Smith fut nommé professeur de logique en 1751 puis professeur de philosophie morale en 1752 à l'université de Glasgow. Plus tard, il rassembla les cours d'éthique qu'il dispensait et les publia dans sa première oeuvre maîtresse intitulée *Theory of Moral Sentiments*, en 1759. En 1763, il démissionna de son poste de professeur pour accompagner le duc de Buccleuch dans un voyage de 18 mois en France et en Suisse, en qualité de précepteur. De 1766 à 1776, il vécut à Kirkcaldy où il travailla à son ouvrage fondamental, la *The Wealth of Nations*. Smith fut ensuite nommé commissaire des douanes à

Édimbourg en 1778, poste qu'il occupa jusqu'à sa mort. En 1787, il fut également nommé recteur de l'université de Glasgow. Son célèbre traité *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776), première étude tentant de décrire la nature du capital et le développement historique de l'industrie et des échanges entre les pays européens, lui valut d'être considéré comme le père de la science économique moderne. La *The Wealth of Nations* constitue le premier essai traitant de l'histoire de la science économique qui considère l'économie politique comme une discipline autonome, distincte de la science politique, de l'éthique et de la jurisprudence. Smith y propose une analyse du processus de production et de répartition de la richesse, et démontre que les sources principales de tout revenu, c'est-à-dire les formes fondamentales dans lesquelles la richesse est distribuée, sont les rentes, les salaires et les profits. *The Wealth of Nations* affirme contre les physiocrates le principe selon lequel le travail est la source de toute richesse, et présente le développement de l'industrie comme une source d'accroissement de la production. Pour Smith, théoricien du capitalisme libéral, le progrès économique et moral procède de la concurrence, la production et les échanges de biens ne pouvant être stimulés, et en conséquence le niveau de vie général amélioré, que lorsque les gouvernements régulent et contrôlent au minimum les activités industrielles et commerciales individuelles. Pour décrire cette situation, il parle d'un ordre naturel réglé par la "main invisible", qui fait naturellement converger la somme des intérêts individuels vers l'intérêt général. En conséquence, trop d'intervention de l'État dans ce contexte de libre concurrence ne pourrait être que néfaste.



Sommerfeld, Arnold (1868-1951) était un physicien allemand né à Königsberg et décédé à Münch. Il étudia les mathématiques et les sciences naturelles à l'université de Königsberg où il reçut son doctorat en 1891. Il occupa successivement les chaires de mathématiques à Clausthal (1897), de mathématiques appliquées à Aix-la-Chapelle (1900) et de physique théorique à Munich (1906-1931). En 1897, il commença, avec C. F. Klein, un traité en 4 volumes sur le gyroscope, qu'il mit 13 ans à terminer et, à la même époque, fit également des recherches dans d'autres domaines de physique appliquée et d'ingénierie, comme la friction, la lubrification et la radio. On lui doit une amélioration du modèle de Bohr (1916) introduisant des orbites elliptiques et des corrections relativistes. Ce nouveau modèle, qui implique une dépendance de l'énergie vis-à-vis du deuxième nombre quantique, permet d'expliquer la structure fine des raies spectrales émises par les atomes. Sommerfeld introduisit d'ailleurs la fameuse "constante de structure fine". Il s'intéressa également après Drude et Lorenz au modèle des électrons libres qui explique certaines propriétés des métaux, en particulier la conduction, en considérant un comportement quantique des électrons. Il participa ainsi aux développements de la théorie des bandes en physique du solide, formulant en 1928 l'idée selon laquelle les électrons occupent des états quantifiés dans la matière.



Stokes, George Gabriel (1819-1903) était un mathématicien et physicien né en Irlande et décédé à Cambridge. En 1841, il reçoit son diplôme avec mention d'honneur de l'université de Cambridge et entame une carrière de chercheur. Influencé par son ancien professeur, il se consacre à l'étude des fluides visqueux. Il publie en 1845 le résultat de ses travaux sur les mouvements des fluides dans sa thèse *On the theories of the internal friction of fluids in motion*. Son approche mathématique décrivant l'écoulement d'un fluide newtonien incompressible dans un espace tridimensionnelle, en ajoutant une force de viscosité à partir des équations d'Euler (Principes généraux du mouvement des fluides, 1755), est à l'origine des équations de Navier-Stokes. L'ensemble de ses recherches est synthétisé par son traité *Report on recent research in Hydrodynamics*, paru en 1846, texte fondateur de l'hydrodynamique. Il devient dès 1849 professeur à la chaire de mathématique de cette même université. Élu en 1851 à la Royal Society, il en sera le président de 1885 à 1890. Les trois derniers postes cités avaient été occupés par Isaac Newton. Il est lauréat du prix Smith en 1841, de la médaille Rumford en 1852 et de la médaille Copley en 1893.



Stefan, Josef (1835-1893) était un physicien autrichien né à Sankt Peter près de Klagenfurt et décédé à Vienne. Les travaux originaux de Stefan comprennent la théorie cinétique des gaz, l'hydrodynamique et surtout la théorie du rayonnement. Après des études à l'université de Vienne où il obtient son doctorat en 1858, nommé Privatdozent de physique-mathématique, il devient professeur de physique en 1863, puis directeur de l'Institut de physique (1866). Membre de l'Académie des sciences de Vienne, il en est le secrétaire à partir de 1875. Avant les travaux de Stefan, G. R. Kirchhoff avait déjà décrit les propriétés du "corps parfaitement noir", susceptible d'absorber la totalité du rayonnement incident et d'émettre un spectre étendu de longueurs d'ondes. Stefan démontre empiriquement en 1879 que l'intensité du rayonnement du corps noir est proportionnelle à la quatrième puissance de sa température absolue, relation connue depuis sous le nom de "loi de Stefan-Boltzmann", Boltzmann l'ayant déduite en 1884 de considérations thermodynamiques. Cette loi constitue l'une des premières étapes importantes qui ont conduit à l'interprétation du rayonnement du corps noir et à la théorie quantique du rayonnement.



Sturm, Charles François (1803-1855) Après avoir été étudiant à l'université de Genève (sa ville natale), Sturm se rend, pour être précepteur dans la famille De Broglie, à Paris, où il fréquente les plus grands savants de l'époque et où il se fixe définitivement à partir de 1825. Il détermine en 1826 la vitesse de propagation du son dans l'eau, ce qui lui vaut, l'année suivante, le grand prix de mathématiques proposé pour le meilleur mémoire sur la compressibilité des liquides. En 1829, il énonce le célèbre théorème qui porte son nom, essentiel pour l'étude des propriétés des racines d'une équation algébrique et qui précise le nombre de racines réelles d'une équation numérique comprises entre deux limites données. Il publie la démonstration de ce théorème en 1835. À partir de 1830, en liaison avec son ami Liouville, il aborde le problème de la théorie générale des oscillations et étudie des équations différentielles du second ordre (problème de Sturm-Liouville) dans plusieurs articles, dont *Sur les équations différentielles linéaires du second ordre* (1836) et *Sur une classe d'équations à différences partielles*

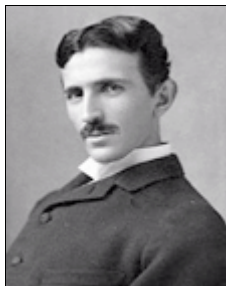
(1836). Les méthodes employées seront à l'origine de nombreux travaux et découvertes mathématiques. Il est élu en 1836 à l'Académie des sciences et travaille à l'École polytechnique. Succédant à Poisson, il enseigne, à partir de 1840, à la faculté des sciences de Paris (chaire de mécanique). Ses *Cours d'analyse de l'École polytechnique* (1857-1863) et ses *Cours de mécanique de l'École polytechnique* (1861) seront publiés après son décès à Paris.



Taylor, Brook (1685-1731), mathématicien anglais né à Edmonton et décédé à Londres, célèbre pour ses contributions au développement du calcul infinitésimal. Taylor fit ses études au collège Saint John, à Cambridge. Il obtint, en 1708, une remarquable solution du problème du *centre d'oscillation*, qui pourtant demeura inédite jusqu'en 1714 lorsque son droit de priorité lui fut contesté par Jean Bernoulli. L'ouvrage de Taylor, *Methodus incrementorum directa et inversa* (1715), ajoute aux mathématiques supérieures un nouveau chapitre, que l'on appelle de nos jours le "calcul des différences finies". Entre autres applications ingénieuses, il s'en sert pour déterminer la forme du mouvement d'une corde vibrante en le réduisant avec succès aux principes de la mécanique. Le même ouvrage contient la célèbre formule connue sous le nom de "théorème de Taylor", dont l'importance n'apparut qu'en 1772, quand Louis de Lagrange réalisa sa puissance et en fit le principe fondamental du calcul différentiel. Dans son essai *Linear Perspective*, Taylor pose les principes de l'art sous une forme originale et plus générale qu'aucun de ses prédécesseurs. Mais l'ouvrage souffrit de la confusion et du manque de clarté qui affectaient la plupart de ses écrits. Taylor fut élu membre de la Royal Society en 1712. Il siégea la même année au comité chargé de régler les querelles de priorité entre Newton et Leibniz et fut secrétaire de la société de 1714 à 1718. À partir de 1715, ses recherches prirent une orientation philosophique et religieuse.



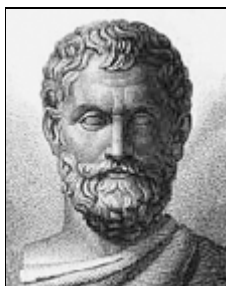
Teller, Edward (1908-2003) est un physicien nucléaire né à Budapest et décédé à Stanford. Il quitte Budapest en 1926 pour aller à Karlsruhe (Allemagne), afin d'étudier la chimie, mais très vite une affinité se créa avec la nouvelle théorie de la physique quantique ce qui l'amènera à étudier à l'université de Leipzig où il obtiendra son doctorat à l'âge de 22 ans. Teller obtint ce titre sous la direction de Werner Heisenberg qui participa plus tard activement dans le camp des nationalistes allemands lors de la seconde guerre mondiale. En 1935, Teller s'expatria aux États-Unis et ses compétences dans la physique de pointe l'amènèrent à se faire beaucoup de relations et une très bonne réputation dans la communauté scientifique. Il fut ainsi nommé professeur dans de nombreuses universités américaines et travailla en 1942 au projet Manhattan où il mena des travaux très importants qui permirent de créer la première bombe nucléaire à fission. Le travail effectué, Teller soutint la continuité du travail pour la recherche d'une bombe thermonucléaire par peur de l'avancée des Russes dans ce domaine (Teller était anticommuniste et très bon ami de Landau qui se fit arrêter par la police communiste). Teller réussit à convaincre l'administration américaine à financer les recherches pour une bombe à hydrogène et mena les travaux avec succès qui fait qu'on le considère aujourd'hui comme le père de la bombe H.



Tesla, Nikola (1856-1943) était un inventeur et ingénieur serbe de génie dans le domaine de l'électricité décédé à New York. Il est souvent considéré comme l'un des plus grands scientifiques dans l'Histoire de la technologie, pour avoir déposé plus de 900 brevets (qui sont pour la plupart repris au compte de Thomas Edison) traitant de nouvelles méthodes pour aborder la conversion de l'énergie. En 1875, il entre à l'École polytechnique de Graz, en Autriche, où il étudie la mathématique, la physique et la mécanique. Une bourse lui est attribuée par l'administration des Confins militaires (Vojna Krajina), le mettant à l'abri des problèmes d'argent. Ceci ne l'empêche cependant pas de travailler avec acharnement pour assimiler le programme des deux premières années d'études en un an. L'année suivante, la suppression des Confins militaires retire toute aide financière à Tesla, hormis celle, très maigre, que peut lui apporter son père, ce qui ne lui permet pas d'achever sa seconde année d'études. On lui doit le moteur électrique asynchrone, l'alternateur polyphasé, le montage triphasé en étoile, la commutatrice. Tesla découvre le principe de la réflexion des ondes sur les objets en 1900, il étudie et publie, malgré des problèmes financiers, les bases de ce qui deviendra presque trois décennies plus tard le radar.



Thom, René (1923-2002) Mathématicien français auteur d'importants travaux en topologie différentielle. Né à Montbéliard et décédé à Bures-sur-Yvette, Thom fut élève de l'École normale supérieure. En 1958, il a reçu la médaille Fields pour sa théorie du cobordisme (relation d'équivalence entre variétés différentielles compactes). Dans une communication au colloque de Strasbourg (1951), Thom établit que, si les zéros d'un idéal polynomial forment une variété, c'est une variété bordante, et sa thèse, *Espaces fibrés en sphères et carrés de Steenrod* (1951), contient déjà en germe les principales méthodes cobordistes. C'est dans le dernier chapitre d'un mémoire de 1954 (*Quelques Propriétés globales des variétés différentiables*) que la théorie du cobordisme est exposée pour la première fois. Après 1955, Thom a surtout étudié les espaces feuilletés et les ensembles et morphismes stratifiés. On lui doit des résultats sur les approximations des transformations différentiables et leurs singularités, les comparaisons de structures différentiables sur une variété triangulée et une théorie de Morse pour les variétés feuilletées. Il est également l'un des premiers à avoir utilisé les techniques de "chirurgie" des variétés. Depuis 1969, Thom s'est consacré aux applications de la topologie aux phénomènes de la vie. Pour décrire la naissance et l'évolution des formes, il a élaboré une mathématique spécifique: sa théorie des catastrophes est une théorie des singularités de certaines équations différentielles. Concrètement, elle permet, à partir de phénomènes observés, de remonter à leurs causes inconnues, au moins partiellement. Thom a donné un exposé de ses travaux dans l'ouvrage *Stabilité structurelle et morphogénèse* (1973).



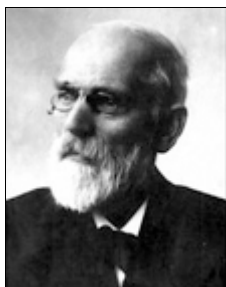
Thalès de Milet (~624 av. J.-C.- ~524 av. J.-C.) est le premier mathématicien dont l'histoire ait retenu le nom. Il est né à Milet, en Asie mineure, sur les côtes méditerranéennes de l'actuelle Turquie. Plus qu'un simple mathématicien, Thalès était un savant universel, curieux de tout, astronome et philosophe, très observateur. On ne démontrait pas ce qu'on avançait à l'époque de Thalès, on ne faisait que remarquer certaines propriétés. Mais la façon qu'avait Thalès de réfléchir, d'analyser des situations, d'en rechercher les causes font de lui le

précurseur des scientifiques (il s'en tenait à l'observation et à l'expérimentation). Une de ses grandes interrogations était l'eau, et les causes de la pluie. Il avait remarqué que l'air se transformait en pluie, et il en cherchait désespérément les réponses. Thalès a formulé plusieurs propriétés géométriques qu'il tenait peut-être des Égyptiens et dont les premières traces de démonstration connues sont bien ultérieures mais, ce faisant, il pose les premiers jalons du raisonnement sur des figures idéales grâce auquel il obtint plusieurs résultats connus sous le nom de "théorèmes de Thalès". Mais le fait d'armes de Thalès est sans conteste la prévision d'une éclipse du soleil, probablement celle du 8 mai 585 avant notre ère. On lui doit notamment la première connaissance de l'électricité, grâce à deux expériences. Il remarqua d'abord que l'ambre avait la propriété d'attirer les matériaux légers. Une autre expérience réalisée en Magnésie..., vers -600, lui permet de mettre en évidence les propriétés d'aimantation de l'oxyde de fer.



Turing, Alan (1912-1954) Par ses travaux théoriques dans les domaines de la logique et des probabilités, Turing est considéré, sinon comme le fondateur des ordinateurs, en tout cas, comme l'un des pères spirituels de l'intelligence artificielle. Né à Paddington (Londres) Turing connaît une scolarité sans éclat malgré un esprit brillant et de nettes dispositions pour les sciences. En 1928, à la Sherborne School où il est entré 2 ans plus tôt, il fait une rencontre qui provoque en lui un déclic et l'amène à s'intéresser réellement à la science et plus exactement aux mathématiques. De 1931 à 1934, Alan Turing est étudiant en mathématiques au King's College de l'Université de Cambridge. Au cours de cette période, il prend connaissance des travaux de John von Neumann sur la mécanique quantique. Stimulé par ces recherches, il se lance dans l'étude de problèmes de probabilités et de logique. C'est aussi au King's College qu'il rencontre des théoriciens de l'économie comme John Keynes. Diplôme en poche, il apprend à l'été 1936 les avancées de Max Newman concernant l'élaboration d'une théorie mathématique sur l'incomplétude de Gödel et la question de la décidabilité de Hilbert. Si pour beaucoup de propositions, il est facile de trouver un algorithme, qu'en est-il de celles pour lesquelles l'algorithme, pas assez rigoureux, est insuffisant à valider la proposition? Doit-on en déduire qu'elles ne peuvent être validées? C'est désormais dans ce sens que vont s'orienter les recherches de Turing. En 1936 il reçoit le prix Smith pour ses travaux sur les probabilités et le concept de la "Machine de Turing". Ce concept constitue la base de toutes les théories sur les automates et plus généralement celle de la théorie de la calculabilité. Il s'agit en fait de formaliser le principe d'algorithme, représenté par une succession d'instructions –agissant en séquence sur des données d'entrée –susceptible de fournir un résultat. Cette formalisation oblige Turing à développer la notion de calculabilité et à déterminer des classes de problèmes "décidables". Cela le conduit à introduire une nouvelle classe de fonctions: les "fonctions calculables au sens de Turing". Au cours de son doctorat à l'Université de Princeton, de 1936 à 1938, Turing conçoit l'idée de la construction d'un ordinateur. De retour à Cambridge, il poursuit ses études mathématiques et s'intéresse à la fonction zêta de Riemann. La seconde Guerre Mondiale lui offre bientôt l'opportunité de mettre en pratique ses théories. C'est au département des communications du Ministère des affaires étrangères britannique qu'il se retrouve confronté au secret d'Enigma, nom de code de la machine utilisée par la marine allemande pour communiquer

avec leurs sous-marins. Le cryptage utilisé par les Nazis échappait toujours aux modes d'investigation classiques. Mais avec la collaboration de W. G. Welchman, Turing réussit à percer le code en appliquant sa nouvelle méthode et, de façon indirecte, contribue ainsi à la victoire de la bataille de l'Atlantique. La guerre achevée, Turing intègre le National Physical Laboratory de Grande-Bretagne où il entreprend, en concurrence avec les projets américains, de créer le premier ordinateur. Les avancées technologiques lui laissent entrevoir la réalisation de cet objectif dans un avenir proche. En 1948, grâce à Newman, il obtient un poste de chargé de cours en mathématiques à l'Université de Manchester qu'il occupera jusqu'à la fin de sa vie. Deux ans plus tard, il participe avec Frederic Williams et Tom Kilburn à la réalisation d'un calculateur électronique, le Mark I, et conçoit à cette occasion un manuel de programmation. Dans la foulée, il publie *Can a machine think ?* dans lequel il fait la synthèse des bases mathématiques et conceptuelles de l'ordinateur électronique programmable et résume sa philosophie de la "machine intelligente". Il énonce également le célèbre "Test de Turing" qui se résume à une expérience dans laquelle un homme tient une conversation avec une machine. Comment dans ce cas, un observateur, par l'unique analyse des messages échangés, pourra-t-il distinguer l'homme de la machine? Turing était convaincu que tout n'était qu'un problème d'information et que le développement des technologies permettrait d'ici 50 ans aux machines de tenir en échec l'être humain au moins cinq minutes. Turing se suicida par empoisonnement au cyanure à cause des pressions homophobes qu'il subissait au Royaume-Uni.



Van Der Waals, Johannes Diderik (1837-1923) Physicien néerlandais né à Leyde et décédé à Amsterdam. Van Der Waals fut tout d'abord instituteur dès l'âge de 20 ans avant de devenir, à la suite d'efforts solitaires, professeur dans l'enseignement moyen (1863). Il fréquenta les cours de l'université de Leyde de 1862 à 1865 et enseigna la physique à Deventer et à La Haye (1866). En 1873, il fut reçu docteur par l'université de Leyde, après la défense d'une dissertation intitulée: *Over de continuïteit van den gas en vloeistofoestand* qui contient la présentation de l'équation d'état qui porte son nom et conduit à des résultats beaucoup plus satisfaisants que l'équation classique des gaz parfaits au voisinage de la zone de liquéfaction. Cette étude contribua d'une façon décisive à accréditer l'idée de l'existence de forces intermoléculaires d'attraction et à déterminer le rôle du volume d'encombrement moléculaire dans le comportement des gaz à haute pression, deux concepts encore mal assurés à l'époque. Le succès rapide de la nouvelle théorie est illustré par les multiples traductions de la dissertation originale qui suivirent sa présentation. On sait à présent que l'équation de Van der Waals est encore imparfaite et qu'il serait téméraire de vouloir lui conserver le nom "d'équation des gaz réels" qui lui fut naguère attribué. En effet, des équations d'état encore mieux appropriées permettent d'atteindre aujourd'hui une approximation plus complète qui sont en général déduites de considérations de cinétique moléculaire fondées sur le théorème du viriel des forces. De 1877 à 1907, date de sa retraite, Van der Waals occupa la chaire de physique à l'université d'Amsterdam. C'est pendant cette période qu'il fit connaître sa loi dite "loi des états correspondants" (1880). Cette équation d'état unique pour tous les corps purs contribua largement, elle aussi, à sa renommée, car elle servit par la suite de guide aux essais préalables à la liquéfaction de l'hydrogène et de l'hélium.

D'un autre point de vue, cette contribution de Van der Waals est également considérée comme l'une des premières tentatives pour exprimer des lois de la physique en fonction de variables réduites. Parmi les autres travaux de Van der Waals, citons une contribution à la théorie moléculaire des mélanges binaires et l'étude de la capillarité. Le prix Nobel de physique lui a été décerné en 1910 pour ses travaux concernant l'équation de l'état d'agrégation des gaz et des liquides.



Viète, François (1540-1603) né à Fontenay-le-Comte et décédé à Paris. Viète est célèbre aujourd'hui en tant qu'inventeur de l'algèbre moderne. Or, à son époque, il était plus connu comme maître des requêtes et conseiller privé d'Henri IV que comme mathématicien. Toute sa vie est en effet marquée par cette dualité d'une carrière politique brillante et d'un ardent travail de cabinet sur les plus hauts problèmes posés par la mathématique de son siècle. Son oeuvre scientifique a beaucoup souffert de ses nombreuses occupations politiques et du peu de temps qu'elles lui laissaient. Il reste néanmoins que la contribution de Viète au développement des mathématiques à la fin du 16ème siècle est fort importante. Elle se caractérise par l'introduction systématique de la représentation littérale dans les problèmes algébriques, tant pour les inconnues que pour les quantités connues, ce qui présente le principal avantage de traiter le cas général et non les cas particuliers et de s'intéresser à la structure des problèmes plutôt qu'à leur expression. Dans sa jeunesse Viète est l'élève des franciscains, au collège des Cordeliers. Il poursuit ses études de droit à la faculté de Poitiers et entra dans la vie active comme avocat. Il est nommé conseiller au parlement de Bretagne en 1573, il y séjourne en fait assez peu, occupé qu'il est par ses travaux mathématiques et les missions confidentielles que lui confie le roi. On retrouve ensuite sa trace à Paris en 1579 où il publie le *Canon mathematicus*, accompagné du *Liber singularis*. Nommé maître des requêtes de l'hôtel du roi en 1580, il est démis de sa fonction en 1585, à la suite de conflits de personnes. En 1589, il est à Tours et prépare la publication de son oeuvre scientifique. Il s'occupe également de cryptographie statistique pour le compte du roi. Il regagne Paris avec ce dernier et est nommé conseiller privé. Viète décédera après une assez longue période de déclin du à la maladie.



Walras, Léon (1834-1910) était un économiste français né à Évreux et décédé à Clarens. Il est le fils d'Auguste Walras, un économiste français dont la pensée influencera beaucoup celle de son fils, dans le domaine de la réforme sociale en général et foncière en particulier. Il étudie au collège de Caen en 1844, puis au lycée de Douai en 1850. Il est diplômé bachelier-ès-lettres en 1851 et bachelier-ès-sciences en 1853. La même année, il n'est pas déclaré admissible à l'École polytechnique et ce aussi lors d'un second essai. En 1854, il est reçu élève externe à l'École des Mines de Paris, mais il n'a pas d'intérêt pour la formation d'ingénieur et il abandonne cette école. Nommé professeur à l'université de Lausanne, Walras dénonça à partir des années 1870, les théories économiques libérales alors enseignées dans les universités, qu'il jugeait incapables de rendre compte des problèmes économiques de son temps. Dans ses *Éléments d'économie politique pure* (1874), sa critique vise en particulier les théories de la valeur travail et de la rente foncière mais à travers lui c'est tout l'héritage classique qu'il remet en cause (notamment celui d'Adam Smith). Influencé par le mathématicien Antoine

Cournot, il est l'un des premiers à introduire de manière systématique le calcul mathématique en économie. Walras place l'entreprise au coeur de l'économie et s'intéresse à son action dans le cadre d'une concurrence entre agents, ainsi que dans celui d'une interdépendance de tous les marchés économiques: les marchés des produits (biens et services) et ceux des facteurs de production (notamment la terre, le travail et les capitaux). Il se demande comment se fixent les prix et les quantités de façon simultanée, et pose le problème de l'équilibre général, c'est-à-dire de la stabilité des équilibres sur tous les marchés. L'attention portée à cette question caractérise les membres de l'École de Lausanne, en particulier le successeur de Walras, Vilfredo Pareto. Avec l'Autrichien Carl Menger et le Britannique Stanley Jevons, qu'il ne connaissait pas au moment où il s'engageait sur cette voie, il est considéré comme l'un des fondateurs du courant néoclassique et du marginalisme.



Weber, Wilhelm (1804-1891), physicien allemand né à Wittenber et décédé à Göttingen qui se spécialisa en électrodynamique. Weber écrivit, en 1824, un traité sur le mouvement ondulatoire avec son frère aîné, Ernst Heinrich Weber, anatomiste réputé, et étudia, avec son frère cadet Eduard Friedrich Weber le mécanisme de la marche (1836). À Göttingen, il collabora avec Carl Friedrich Gauss pour l'étude du géomagnétisme, et il relia leurs laboratoires par un télégraphe électrique: ce fut l'une des premières transmissions par télégraphe que l'on connaisse. Sa réalisation majeure fut celle qu'il mena à Leipzig, avec F.W.G. Kohlrausch: il détermina le rapport des unités de charge électrostatiques et électrodynamiques (la constante de Weber) qui se révéla être l'équivalent d'une vitesse, et fut utilisé plus tard par James Clerk Maxwell pour renforcer sa théorie sur l'électromagnétisme.



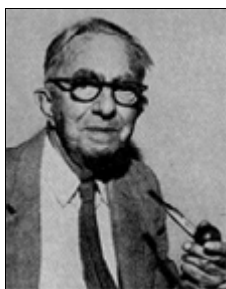
Weierstrass, Karl Theodor Wilhelm (1815-1897) était un mathématicien allemand, qui donna à la théorie des fonctions sa forme moderne en précisant en particulier le formalisme des limites et est considéré à ce titre comme le père de l'analyse moderne. Né à Ostenfelde, il fit ses études à Bonn et à Münster où il fut instituteur. C'est là qu'il s'intéressa aux mathématiques, et plus particulièrement à l'étude des fonctions elliptiques. Pendant de nombreuses années, Weierstrass travailla dans l'ombre pour établir sa théorie des fonctions de variable complexe, qui repose sur les développements en série entière. En 1854, il publia un mémoire sur les intégrales abéliennes et sur l'inversion des intégrales hyperelliptiques, qui établit sa réputation comme mathématicien et lui valut un doctorat honoraire de l'université de Königsberg. Nommé professeur à l'université de Berlin, il enseigna de 1864 à sa mort. Il a peu publié de son vivant et sa réputation est venue principalement de l'influence de ses cours à Berlin. Ceux-ci furent suivis par de nombreux mathématiciens et établirent la théorie des fonctions sur des bases de rigueur auxquelles son nom reste attaché, la "rigueur weierstrassienne". Il est aussi connu pour avoir rendu public un exemple de fonction continue nulle part dérivable (fonction de Weierstrass).



Weyl, Hermann (1885-1955) est un des mathématiciens les plus influents du 20^{ème} siècle, l'un des premiers à combiner la relativité générale avec les lois de l'électromagnétisme. Ses recherches en mathématiques portèrent essentiellement sur la topologie et la géométrie. Il effectua des recherches en mécanique quantique et en théorie des nombres. Né à Elmshorn à proximité de Hambourg en Allemagne, Weyl étudia de 1904 à 1908 à Göttingen et à Munich, principalement intéressé par la mathématique et la physique. Son doctorat fut soutenu à Göttingen sous la direction de Hilbert et Minkowski. En 1910, il obtint un poste d'enseignant comme lecteur privé à Göttingen. Il enseigna la mathématique à l'École polytechnique fédérale de Zürich en Suisse en 1913. C'est à Princeton qu'il travailla avec Einstein. Weyl rechercha une unification de la gravitation et de l'électromagnétisme. Cette recherche donna des explications de la violation de la non-conservation de la parité, une caractéristique des interactions faibles. Weyl continua à travailler à l'IAS jusqu'à sa retraite en 1952 ; il mourut à Zurich. En 1918, il introduit la notion de jauge, première étape de ce qui deviendra la théorie de jauge. En réalité, sa vision était une tentative non réussie de modéliser les champs électromagnétiques et gravitationnels comme des propriétés géométriques de l'espace-temps. Ces travaux se révélèrent fondamentaux pour comprendre la symétrie des lois de la mécanique quantique. Il en posa les bases, donnant naissance aux spineurs, devenus familiers autour des années 1930.



Weinberg, Steven (1933-) Né à New York il débuta ses études à New York même puis à l'université Cornell (dans l'État de New York) et soutint, en 1957 à Princeton, sa thèse sur les effets de l'interaction forte dans les processus dominés par l'interaction faible. Chercheur à l'université de Californie à Berkeley de 1959 à 1966, il s'intéressa à de multiples problèmes en théorie quantique des champs, en physique des particules et en astrophysique. Professeur à Harvard à partir de 1973, il contribua de façon décisive à la compréhension moderne des interactions fondamentales. Il rejoignit l'université du Texas à Austin en 1982. L'unification des forces fondamentales a sous-tendu les efforts des physiciens modernes depuis Newton, Maxwell et Einstein qui, après avoir uni l'espace et le temps, tenta, mais en vain, d'englober en une seule théorie gravitation et électromagnétisme. La découverte, au début du 20^{ème} siècle, des deux forces nucléaires, les interactions faible et forte, donna un nouvel élan à ces tentatives. En 1967, Weinberg et le physicien pakistanais Abdus Salam proposèrent, indépendamment, que l'électromagnétisme et l'interaction nucléaire faible soient issus d'une même interaction électrofaible, dont la symétrie de jauge est spontanément brisée et dont le vecteur est un triplet de bosons massifs et le photon. Quelques années plus tard des expériences au CERN de Genève apportaient les premières confirmations du modèle de Weinberg-Salam. Le prix Nobel de physique 1979 (partagé avec l'Américain Sheldon Lee Glashow, pour l'importance de ses travaux de précurseur) récompensa les deux auteurs de ce qu'on appelle maintenant le "modèle standard" des interactions électrofaibles. Pédagogue, Weinberg est l'auteur de plusieurs cours de physique de haut niveau, tant sur la gravitation que sur la théorie des champs. Vulgarisateur de talent, son livre *Les Trois Premières Minutes de l'Univers* fut un succès mondial.



Wilcoxon, Frank (1892-1965) était un chimiste et statisticien connu pour le développement de tests statistiques très répandus! Frank Wilcoxon est né de parents américains à County Cork en Irlande. Il a grandi à Catskill, New York mais a suivi une partie de sa scolarité en Angleterre. En 1917, il est diplômé du collège militaire de Pennsylvanie avec une licence. Après la première guerre mondiale, il débute des cours de maîtrise à l'université de Rutgers où il obtint sa maîtrise en chimie en 1922 en ensuite à l'université de Cornell où il obtint son doctorat en chimie physique en 1924. Wilcoxon commença sa carrière de chercheur à l'institut Boyce Thompson en 1925 et y resta jusqu'en 1941. Ensuite, il prit un poste dans la compagnie Atlas Powder où mis en place et dirigea le laboratoire de contrôle avant de rejoindre la compagnie chimique American Cyanamid en 1943. Pendant cette période il développa un intérêt pour la statistique inférentielle à travers les lectures des textes de R.A. Fisher de 1925. Il prit sa retraite en 1957. Pendant sa carrière, Wilcoxon publia 70 articles, le plus connu étant celui contenant les deux tests statistiques qui portent ce nom: le test de la somme des rangs de Wilcoxon et le test de la somme des rangs signés de Wilcoxon. Il s'agit d'alternatives non paramétriques aux tests-t de Student. Over his career Wilcoxon published over 70 papers. His most well-known paper contained the two new statistical tests that still bear his name, the Wilcoxon rank-sum test and the Wilcoxon signed-rank test. Wilcoxon mourut après une brève maladie



Witten, Edward (1951-) Mathématicien et physicien américain, lauréat de la médaille Fields en 1990. Né à Baltimore (Maryland), Edward Witten fait ses études supérieures à l'université Brandeis à Waltham (Massachusetts), puis à l'université de Princeton (New Jersey), où il soutient sa thèse de doctorat en physique en 1974. Chercheur à l'université Harvard de 1976 à 1980, il enseigne ensuite à l'université de Princeton, puis devient membre de l'Institute for Advanced Study de Princeton en 1987. Après des travaux en physique théorique des particules élémentaires, Witten axe ses recherches sur la physique-mathématique et contribue en particulier de façon déterminante au développement des théories des supercordes dans l'espoir que celles-ci pourraient émerger vers une compréhension de l'interaction gravitationnelle au niveau quantique. En mathématiques, il a contribué à l'étude de la théorie de Morse, démontrant les inégalités classiques de Morse en reliant les points critiques à l'homologie. En 1987, il démontre une suite infinie de théorèmes de rigidité sur l'espace des solutions d'équations différentielles, telles que l'équation de Rarita-Schwinger, rencontrées en physique. En théorie des noeuds, il a montré en 1989 qu'on peut interpréter les invariants de noeuds de Vaughan Jones comme des intégrales de Feynman pour une théorie de jauge tridimensionnelle. Il a, de plus, exploré les relations entre la théorie quantique des champs et la topologie différentielle des variétés bi- ou tridimensionnelles. Les progrès récents dans la compréhension des modèles bidimensionnels de la gravitation sont largement dus à l'influence des idées originales de Witten.



Yang, Chen-Ning (1922-) Professeur à l'université chinoise de Hong Kong et à l'Université de Tsinghua à Pékin, professeur émérite de l'Université de New York à Stony Brook, Yang est l'un des plus grands physiciens théoriciens de la seconde moitié du 20ème siècle. Il obtient son Master of Science à l'Université de Tsinghua en 1944. Il s'inscrit en 1946 à l'Université de Chicago que Fermi venait de rejoindre. Plus tard, il décide de se consacrer à la physique théorique et, en 1949, il soutient sa thèse avec un travail sur la phénoménologie des réactions nucléaires. Sa carrière débute à l'Institute for Advanced Studies à Princeton en 1949. En 1965, il refuse de succéder à Oppenheimer comme directeur, mais il décide en 1966 de sortir de sa tour d'ivoire et finit par accepter la chaire Einstein et le poste de directeur de l'Institut de physique théorique de la toute nouvelle Université de New York à Stony Brook. À partir de 1971 il s'engage très activement dans le rétablissement des relations scientifiques entre la Chine et les États-Unis et s'implique dans la création de nouveaux instituts de recherche, en particulier à Nankin. Les contributions de Yang se caractérisent par leur profondeur, par l'ampleur et la variété de leur spectre, de la phénoménologie des particules à la théorie quantique des champs, en passant par la mécanique statistique ainsi que par différentes incursions en physique de la matière condensée. Ses travaux sur la brisure de la symétrie par réflexion d'espace (ou violation de la parité) dans les interactions faibles constituent un exemple parfait d'analyse phénoménologique d'une expérience en contradiction avec les idées reçues, à savoir l'absence d'une orientation privilégiée de l'espace dans les lois de la physique. Son grand mérite porte sur deux points: d'une part, il met en évidence le fait que l'hypothèse en question n'avait pas été testée pour les interactions faibles et, d'autre part, il a imaginé tout un ensemble de tests nouveaux pour l'invariance par réflexion d'espace. Ce bond en avant de la théorie des interactions faibles a permis d'aboutir, avec l'introduction des champs de Yang-Mills, au modèle standard électrofaible. L'idée de Yang fut de généraliser l'invariance de jauge aux groupes des rotations dans un espace abstrait à 3 dimensions censé décrire les degrés de liberté interne des champs de matière. Les champs de Yang-Mills s'imposèrent comme outil fondamental pour la construction d'une théorie prédictive de l'ensemble des interactions faibles, fortes et électromagnétiques, événement décisif qui engagea la révolution de la physique des années 1970. L'ensemble de ses travaux ont eu un impact considérable en physique théorique. Près de 20 ans après la publication de son article avec Mills, Yang a donné une reformulation précise de la théorie des champs de Yang-Mills dans le cadre rigoureux des espaces fibrés. L'analogie avec la théorie de la gravitation devient ainsi apparente et les notions de courbure et de transport parallèle s'introduisent naturellement. Des solutions particulières des équations de Yang-Mills, comme celle découverte par Gerard't Hooft, sont utilisées par les mathématiciens pour explorer les propriétés des variétés différentielles à quatre dimensions. Yang a reçu de nombreux prix scientifiques, dont le prix Nobel de physique en 1957 qu'il a partagé avec Tsung-Dao Lee. Ce prix prestigieux leur a été accordé pour leurs travaux sur les lois de la parité dans le domaine des particules élémentaires. Ces travaux fondamentaux sont particulièrement importants parce qu'ils ont montré que la symétrie droite-gauche des particules élémentaires, universellement admise à l'époque, était tout simplement incorrecte, ce qui fut ensuite prouvé expérimentalement.

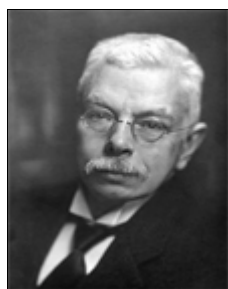


Yukawa Hideki (1907-1981), physicien japonais, né et décédé à Tokyo, il était le cinquième de sept enfants qui devinrent tous des universitaires renommés. Il fut très vite porté vers la mathématique et la philosophie. Admis au département de physique de l'université de Kyoto en 1926, grand lecteur, Yukawa se passionna vite pour les nouvelles conceptions philosophiques accompagnant la relativité et la théorie des quanta, conceptions qu'il avait découvertes en particulier dans les ouvrages de Max Planck. En marge de ses études, il eut connaissance des développements contemporains de la physique quantique qui aboutirent à sa formulation bien établie vers la fin des années 1920. Il obtint son diplôme à l'université de Kyoto en 1929 et commença, dès lors, des recherches personnelles dans la double direction de la physique quantique relativiste et de la physique nucléaire qui se dessinait alors. Il s'attacha tout d'abord au problème de la liaison nucléaire électron-proton, le neutron étant une particule encore inconnue, puis à la théorie quantique des champs. Tout en enseignant la physique quantique, Yukawa poursuivait ses recherches sur les problèmes de la physique des noyaux. En 1934, il s'attaqua au problème de la force nucléaire, que la théorie de Fermi était impuissante à résoudre. Il reprit une idée qu'il avait déjà considérée lors de ses premiers travaux, celle d'une force d'échange, transmise entre le neutron et le proton par une particule nouvelle associée à un champ nouveau, dont il se proposait de déduire les propriétés à partir de la force nucléaire. C'est en octobre 1934 qu'il découvrit la solution, en obtenant une relation entre la masse de cette particule d'échange hypothétique et la portée de l'action des forces nucléaires. La particule de Yukawa, le méson, devait avoir une masse valant 200 fois celle de l'électron. Il fallait supposer que ces mésons étaient de spin entier ou nul, qu'ils obéissaient à la statistique de Bose-Einstein et qu'ils étaient pourvus de charges positive et négative. Ce travail n'attira pas l'attention jusqu'au jour où d'autres chercheurs annoncèrent la découverte d'une particule nouvelle dans le rayonnement cosmique, ayant la masse prévue par Yukawa. Il apparut toutefois que l'interaction de ce méson avec la matière était trop peu intense pour qu'il puisse être la particule d'échange des forces nucléaires. La théorie des deux mésons pallia la difficulté. Il avait découvert entre-temps le mécanisme de désintégration du noyau par capture d'un électron orbital, en appliquant la théorie de Fermi. Il fut le premier Japonais à recevoir le prix Nobel de Physique, en 1949, pour sa théorie mésique des forces nucléaires. Yukawa fonda l'institut de recherches de physique fondamentale de l'université de Kyoto et le dirigea jusqu'à sa retraite, en 1970. Il ne se cantonna pas dans une activité de physicien: il écrivit des essais sur la créativité scientifique et milita en faveur de la paix, signant l'appel d'Albert Einstein et de Bertrand Russell contre l'utilisation des armes atomiques.



Young, Thomas (1773-1829), physicien, médecin et égyptologue britannique né à Milverton et décédé à Londres, surtout connu pour ses découvertes en optique (phénomènes d'interférence), en élasticité des matériaux et en médecine (explication de la vision colorée). À l'âge de 14 ans il se débrouille déjà dans plus d'une dizaine de langues étrangères. Young commence à étudier la médecine en 1792 à Londres, part en 1794 pour Édimbourg, puis un an plus tard pour Göttingen, où il obtient le titre de docteur en physique en 1796. En 1799, il

commença à pratiquer la médecine à Londres. À partir de 1802, et jusqu'à sa mort, il occupa le poste de secrétaire de la Royal Society. En 1811, Young fut nommé à l'hôpital Saint-George de Londres. Il fit partie de plusieurs commissions scientifiques officielles et, à partir de 1818, il fut nommé secrétaire du Bureau des longitudes et éditeur de l'*Almanach nautique*. En optique, Young découvrit le phénomène des interférences, et contribua ainsi à établir le caractère ondulatoire de la lumière. Il fut le premier à décrire et à mesurer l'astigmatisme et à trouver une explication physiologique à la sensation de couleur. Young est également connu pour ses travaux sur les théories de la capillarité et de l'élasticité. Il contribua également au déchiffrement des hiéroglyphes inscrits sur la pierre de Rosette. Ses écrits comportent d'importants travaux en médecine, en égyptologie et en physique.



Zeeman, Pieter (1865-1943) était un physicien né Zonnemaire et décédé à Amsterdam. Il commença à s'intéresser très jeune à la science. En 1883 lors d'aurores boréales visibles aux Pays-Bas, Zeeman, alors étudiant au collège, fit une description et un dessin détaillé du phénomène qui fut publié dans la revue *Nature*. Après avoir passé ses examens d'entrée en 1885, il étudia la physique à l'université de Leiden sous la direction de Hendrik Lorentz. En 1890, avant même de terminer sa thèse, il devint l'assistant de Lorentz. Cela lui permit de participer à un programme de recherche sur l'effet Kerr. En 1893 il soumit sa thèse sur l'effet Kerr, la réflexion de la lumière polarisée sur une surface magnétisée. Après avoir obtenu son doctorat il partit pour un semestre à l'institut F. Kohlrausch à Strasbourg. En 1895, après son retour de Strasbourg, Zeeman devint Privatdozent en mathématiques et physique à Leiden. En 1896, trois ans après avoir soumis sa thèse sur l'effet Kerr, il désobéit aux ordres directs de ses supérieurs et utilisa l'équipement du laboratoire pour mesurer la séparation des lignes du spectre sous un champ magnétique intense. Il a été licencié pour ses efforts... mais il fut récompensé plus tard: il obtint le prix Nobel de Physique en 1902 pour sa découverte ce qui est connu aujourd'hui sous le nom d'effet Zeeman. En plus de son travail de thèse, dans l'étude de l'effet d'un champ magnétique sur une source de lumière. Grâce à sa découverte, Zeeman se vit offrir un poste d'assistant professeur à Amsterdam en 1897. En 1900 s'ensuivit la place de professeur à l'université d'Amsterdam. En 1902 avec son mentor Lorentz il se vit attribuer le Prix Nobel de physique pour l'effet Zeeman. Cinq années plus tard, en 1908, il succéda à Van der Waals comme professeur à temps plein et directeur à l'Institut de Physique à Amsterdam. Il se retira encore en tant que professeur en 1935.

CHRONOLOGIE

En arrivant à la 3'000ème page A4 de rédaction du site et à l'occasion de la 3ème édition du PDF associé, il nous a semblé pertinent de tenter de donner un premier jet de la chronologie de la majorité des sujets traités ou mentionnés sur l'ensemble du site Internet. Cela permet d'avoir un meilleur recul sur les outils utilisés et aussi de rendre hommage à nos illustres prédécesseurs auxquels nous sommes redevables de notre confort de vie, notre longévité, notre maîtrise de l'environnement (pas forcément son respect...) et de sa compréhension (même si pour certaines dates il n'est pas possible de vérifier s'il s'agit de légendes ou de faits réels...).

S'il manque des dates importantes (mais uniquement sur des sujets traités de près ou de loin dans les différents chapitres du site!) ou que vous identifiez des coquilles, n'hésitez pas à nous le faire savoir, il s'agit d'un premier jet et donc la chronologie ne peut qu'être améliorée (de plus l'histoire n'est pas une science exacte!).

Pour plus d'informations le lecteur pourra se référer à Wikipédia qui a atteint des sommets au niveau de la quantité de biographies disponibles et de leurs détails (avec vérification des sources!).

Ceci est l'histoire de comment l'histoire a fait la science et comment la science est entrée dans l'histoire, et comment les idées qui en ont émergé ont façonné notre monde contemporain.

+1994

Les travaux du mathématicien **Andrew Wiles** (plus de 10 ans de recherche!) font tomber le dernier théorème de Fermat.

+1983

Les physiciens **Carlo Rubbia**, **Simon van der Meer**, et l'équipe UA-1 du CERN découvrent les bosons W et Z qui concrétisent l'unification de l'interaction nucléaire faible et de la force électromagnétique.

+1978

Les mathématiciens cryptologues **Ronald Rivest**, **Adi Shamir** et **Leonard Adelman** proposent une procédure de cryptage à clé publique nommée "R.S.A." basée sur la difficulté de factorisation en nombres premiers.

+1973

Le mathématicien **Fischer Black** et l'économiste **Myron S. Scholes** publient le modèle d'évaluation du prix des actifs financiers.

+1965

Détection par les astrophysiciens **Arno Penzias** et **Robert Wilson** des micro-ondes cosmiques dues au rayonnement de fond de ciel prévu par la théorie du physicien **Robert Dicke**.

+1964

L'astrophysicien **Irwin Shapiro** prédit un décalage temporel pour la propagation d'une onde électromagnétique comme test pour la relativité générale. La même année, le physicien **John Stewart Bell** montre que toute théorie à variables locales cachées doit satisfaire l'inégalité de Bell.

+1963

Le mathématicien et météorologiste **Edward Lorenz** trouve ce qui est probablement le premier attracteur étrange et qui ouvre la voie à la théorie du chaos.

+1962

Le mathématicien **Benoît Mandelbrot** découvre les fractales par hasard dans le cadre de l'analyse de signaux aux laboratoires Bell situés aux États-Unis où il se servira des ordinateurs pour répéter les motifs graphiques à l'infini et dont le principe est à la base de la théorie des fractales. La même année, l'économiste **William Forsyth Sharpe** publie le modèle CAPM (Capital Asset Pricing Model).

+1960

Le physicien **Abdus Salam** postule l'existence des bosons W et Z pour expliquer la désintégration bêta et l'apparition d'un nouveau boson Z qui n'avait jamais été observé auparavant. La même année, les physiciens **Ali Javan** et **Theodore Maiman** inventent chacun un type de LASER particulier. Les mathématiciens et ingénieurs **Irving Reed** et **Gustave Solomon** présentent le code correcteur d'erreur de Reed-Solomon.

+1959

Les physiciens **Yakir Aharonov** et **David Bohm** prédisent l'effet Aharonov-Bohm et l'année suivante le physicien **Robert G. Chambers** confirme expérimentalement l'effet.

+1956

Les physiciens **Clyde Cowan** et **Fred Reines** observent le neutrino hypothétisé 25 ans plus tôt par le physicien **Wolfgang Pauli**.

+1955

Les physiciens **Owen Chamberlain**, **Emilio Gino Segrè**, **Clyde Wiegand** et **Thomas Ypsilantis** découvrent l'antiproton.

+1954

L'économiste **Harry Markowitz** publie sa thèse consacrée au modèle de diversification efficiente des portefeuilles d'actifs financiers. La même année, les physiciens **John Bell** et le duo **Wolfgang Pauli** et **Gerhart Lüders** développent la théorie C.P.T. qui analyse la symétrie des lois physiques pour les transformations impliquant de manière simultanée la charge, la parité et le temps. Le physicien **Charles Hard Townes** développe le MASER. Le biologiste **Milner Baily Schaefer** publie son modèle d'équilibre des populations.

+1952

Le mathématicien **George Bernard Dantzig** développe l'algorithme du Simplexe en recherche opérationnelle.

+1951

Le théorème CPT apparaît pour la première fois de manière implicite dans le travail du physicien **Julian Schwinger** afin de prouver la corrélation entre spin et statistique.

+1950

Les physiciens **Johannes Hans Daniel Jensen** et **Maria Goeppert-Mayer** développent le modèle en couches du noyau nucléaire. La même année, le mathématicien et économiste **John Forbes Nash** développe le concept de jeux non coopératifs et généralise la notion de minimax pour les jeux à somme nulle ; l'ingénieur **David Huffman** trouve l'algorithme permettant de compresser n'importe quel type de séries de symboles.

+1949

Le mathématicien et ingénieur électricien **Claude Shannon** publie l'article contenant sa théorie de l'information qui deviendra le socle de base d'un certain nombre de théories physiques, statistiques et de méthodes numériques.

+1948

La physicienne **Maria Göpper-Meyer** développe un modèle théorique pour la structure du noyau atomique avec succès.

+1947

Les physiciens **Cecil Powell**, **Cesare Mansueto Giulio Latte**, et **Giuseppe Occhialini** découvrent le méson pi par l'étude des rayons cosmiques. La même année les physiciens **John Bardeen**, **Walter H. Brattain** et **William Schockley** inventent les transistors à semi-conducteur dans les laboratoires de la compagnie téléphonique Bell aux États-Unis qui sera à l'origine de la révolution informatique.

+ 1946

Le physicien et chimiste **Willard Frank Libby** développe et découvre la possibilité de datation au Carbone 14.

+1945

Le test de Trinity, le premier essai concluant de détonation d'une bombe nucléaire par le physicien **Robert Oppenheimer** et son équipe au Nouveau Mexique.

+1944

Le mathématicien **John von Neumann** développe les premières bases de la théorie mathématique des jeux.

+1943

Le physicien **Sin-Itiro Tomonaga** publie l'article qui pose les bases physiques de l'électrodynamique quantique.

+1942

Le physicien **Enrico Fermi** conduit la première réaction en chaîne contrôlée.

+1941

Le physicien **Ernst Stueckelberg** interprète les positrons comme des électrons d'énergie positive qui remontent le temps.

+1940

Le physicien **Edward Teller** entrevoit la possibilité d'utiliser l'énorme quantité de chaleur dégagée par l'explosion d'une bombe à fission pour déclencher le processus de fusion nucléaire. Il s'agit de la démarche considérée comme la découverte de la fusion nucléaire. La même année, le physicien **Donald William Kerst** met au point le premier bêtatron.

+1939

Les chimistes **Otto Hahn** et **Fritz Strassmann** bombardent des sels d'uranium par des neutrons et découvrent que du baryum est produit par l'expérience (découverte de la fission nucléaire). La même année, la physicienne **Lise Meitner** et le physicien **Otto Robert Frisch** déterminent qu'une fission nucléaire est intervenue lors de l'expérience de Hahn-Strassman.

+1938

Le chimiste et physicien Isidor Isaac Rabi et ses collègues étudient l'effet de champs magnétiques puissants sur des molécules amenant au développement la résonance magnétique nucléaire (RMN). La même année, les physiciens **Hans Bethe** et **Carl von Weizsäcker** proposent une théorie nucléaire des étoiles.

+1937

Les physiciens **Seth Neddermeyer**, **Carl Anderson**, **Jabez Curry Street** et **E.C. Stevenson** découvrent les muons dans les traces laissées par des rayons cosmiques dans une chambre à bulles. La même année le mathématicien **John Von Neumann** développe la méthode de Monte-Carlo pour des méthodes numériques diverses et le physicien **Niels Bohr** développe de la goutte liquide du noyau nucléaire.

+1936

Les physiciens **Georges Gamow** et **Edward Teller** collaborent pour formuler la théorie de l'émission radioactive bêta.

+1935

Le physicien **Hideki Yukawa** présente la théorie de l'interaction forte et prédit l'existence des mésons. La même année, l'astrophysicien et mathématicien **Chandrasekhar Subrahmanyan** communique ses résultats sur l'effondrement d'étoiles en naines blanches et au-delà de 1.44 masses solaires en étoile à neutrons. L'article d'**Albert Einstein**, **Boris Podolsky** et **Nathan Rosen** sur le paradoxe EPR est publié dans Physical Review et remet en question la non localité de l'interprétation de Copenhague.

+1934

Le physicien **Pavel Alekseyevich Cherenkov** étudie l'émission de lumière quand des particules relativistes traversent un milieu amorphe. La même année le physicien **Enrico Fermi** suggère de bombarder des atomes d'uranium par des neutrons pour obtenir un élément à 93 protons, formule la théorie de la désintégration bêta et le physicien **Leó Szilárd** réalise qu'une réaction nucléaire en chaîne est possible. Les physiciens **Irène Joliot-Curie** et **Frédéric Joliot** bombardent des atomes d'aluminium avec des particules alpha et créent artificiellement du phosphore-30 radioactif.

+1933

Le mathématicien **Andreï Nikolaïevitch Kolmogorov** publie un ouvrage contenant une base solide d'axiomes des probabilités.

+1932

Le physicien **Carl David Anderson** découvre le positron. La même année le physicien **Werner Heisenberg** présente le modèle proton-neutron du noyau et l'utilise pour expliquer les isotopes. Le physicien **James Chadwick** découvre le neutron et les physiciens **John Cockcroft** et **Ernest Walton** cassent le noyau du lithium et du bore par bombardement de protons.

+1931

Le physicien **Wolfgang Pauli** met en avant l'hypothèse du neutrino pour expliquer l'apparente violation du principe de conservation de l'énergie au cours de la désintégration bêta. La même année le mathématicien et logicien **Kurt Gödel** démontre qu'un système peut être à la fois cohérent et complet (théorème d'incomplétude) et que si le système est cohérent, alors la cohérence des axiomes ne peut pas être prouvée au sein même du système. Le physicien **Ernest Lawrence** invente le premier cyclotron.

+1930

Le physicien **Fritz London** explique que les forces de van der Waals sont dues à l'interaction des moments dipolaires des molécules. La même année le physicien **Paul Dirac** présente la théorie électron-trou et l'économiste **John Maynard Keynes** publie son *Traité sur la monnaie*.

+1929

L'astronome **Edwin Hubble** en étudiant le décalage vers le rouge émet l'hypothèse que l'Univers n'est pas statique. La même année, le physicien **Robert Van de Graaff** invente le premier accélérateur de particules connu aujourd'hui sous le nom "d'accélérateur de Van de Graaf".

+1928

Le physicien **Paul Dirac** établit son équation d'onde relativiste pour l'électron qui généralise et améliore l'équation relativiste sans spin de Klein-Gordon. La même année, les physiciens **Friedrich Hund** et **Robert S. Mulliken** présentent le concept d'orbitale moléculaire et le physicien et cosmologiste **Georges Gamow** développe le modèle théorique quantique de la désintégration radioactive alpha par effet tunnel.

+1927

Le physicien **Werner Heisenberg** établit le principe d'incertitude, d'après lequel la position et la quantité de mouvement d'une particule ne peuvent être connues simultanément avec précision indirectement, en développant une base théorique nouvelle pour la mécanique quantique. La même année les physiciens **Walter Heitler** et **Fritz London** présentent la théorie quantique de la liaison chimique établie à partir de la molécule d'hydrogène et le physicien **Max Born** interprète de manière probabiliste la fonction d'onde de Schrödinger et avec l'aide du physicien **Robert Oppenheimer** présente l'approximation de Born-Oppenheimer. Les physiciens **Clinton Joseph Davisson**, **Lester Germer** et **George Paget Thomson** confirment la nature ondulatoire des électrons en les diffractant. Le physicien **Paul Ehrenfest** démontre le théorème fameux en physique quantique qui porte aujourd'hui son nom.

+1926

Le physicien **Erwin Schrödinger** établit son équation d'onde qui définit la mécanique quantique sous forme analytique en développant les idées de De Broglie dans une théorie de la mécanique ondulatoire et prouve que la formation matricielle et ondulatoire de la mécanique quantique sont identiques. La même année les physiciens **Oskar Klein** et **Walter Gordon** établissent l'équation de la mécanique quantique relativiste pour les particules sans spin et **Paul Dirac** définit la statistique de Fermi-Dirac. Dans le domaine de la dynamique des populations le mathématicien et physicien **Vito Volterra** publie l'équation différentielle non linéaire modélisant l'équilibre proie/prédateur.

+1925

Le physicien Pierre Auger découvre l'effet Auger (2 ans après Lise Meitner) et la même année les physiciens **George Uhlenbeck** et **Samuel Goudsmit** postulent et mettent en évidence l'existence du spin de l'électron. La même année, le physicien **Wolfgang Pauli** établit par nécessité le principe d'exclusion quantique. Les physiciens **Werner Heisenberg**, **Max Born** et **Pascual Jordan** développent une version matricielle de la mécanique quantique.

+1924

Le physicien **John Lennard-Jones** propose une description semi empirique des forces d'interaction inter-atomiques et la même année les physiciens **Satyendranath Bose** et **Albert Einstein** définissent la statistique de Bose-Einstein. Dans le domaine des statistiques, le statisticien **Ronald Fisher** définit des notions modernes majeures en statistiques

+1923

L'astronome **Edwin Hubble** estime la distance entre la Terre et les galaxies spirales, prouvant qu'elles sont très éloignées de la Voie lactée et la même année le physicien **Louis de Broglie** suggère la dualité onde-particule à partir de la théorie des quanta et de l'équivalence masse-énergie d'Einstein et la physicienne **Lise Meitner** découvre l'effet Auger.

+1922

Le physicien **Arthur Compton** étudie la dispersion de photons X par des électrons et l'astrophysicien **Alexandre Friedmann** développe les modèles d'Univers non statiques.

+1921

Le physicien **Alfred Landé** définit le rapport gyromagnétique et la même année les physiciens **Otto Stern** et **Walter Gerlach** démontrent expérimentalement que le moment intrinsèque de l'électron est quantifié. Le mathématicien et physicien **Theodor Franz Eduard Kaluza** démontre qu'une version à 5 dimensions des équations des champs d'Einstein permet d'unifier la gravitation et l'électromagnétisme.

+1920

L'astronome **Vesto Melvin Slipher** met en évidence le phénomène de décalage vers le rouge du spectre des galaxies. La même année, le physicien **Arnold Sommerfeld** introduit un quatrième nombre quantique au modèle original de l'atome de Bohr.

+1919

Le physicien **Ernest Rutherford** réalise la première désintégration artificielle d'un atome en bombardant de l'azote avec des particules alpha. La même année, la physicienne et mathématicienne **Amali Emmy Noether** développe son théorème sur les invariants différentiels dans le calcul des variations, l'un des plus importants théorèmes mathématiques jamais prouvé dans l'orientation du développement de la physique moderne.

+1918

L'astronome **Harlow Shapley** fait la première estimation précise de la taille de notre galaxie et de la position du Soleil dans celle-ci. La même année, le physicien **Hermann Weyl** introduit la notion de jauge, première étape de ce qui deviendra la théorie de jauge.

+1917

Le physicien **Albert Einstein** présente l'idée d'émission stimulée de radiation qui servira à la base de la fabrication des LASER. La même année, le physicien **Arnold Sommerfeld** introduit un troisième nombre quantique au modèle original de l'atome de Bohr.

+1916

Le physicien **Albert Einstein** développe sa théorie de la relativité générale et montre comment la matière joue sur l'espace-temps pour produire des effets gravitationnels. Il s'agit de la première théorie dite "indépendante de fond". La même année les physiciens **Gilbert Lewis** et **Irving Langmuir** présentent le modèle en couches électroniques pour expliquer les liaisons chimiques et le physicien **Arnold Sommerfeld** introduit la relativité dans son modèle de 1915 et cette correction relativiste permet d'expliquer les valeurs observées par des spectrographes de très haute résolution et donc de la scission des raies spectrales qu'on appelle "structure fine" et introduit par la même occasion un deuxième nombre quantique décrivant les orbites elliptiques. Le physicien **Karl Schwarzschild** trouve une solution mathématique des équations d'Einstein, il l'applique aux étoiles à neutrons, et aux trous noirs.

+1915

Le physicien **Arnold Sommerfeld** affine le modèle atomique du physicien **Niels Bohr** en introduisant des orbites elliptiques pour expliquer la structure fine des raies de l'atome d'hydrogène. Ce nouveau modèle n'expliquait cependant pas l'étendue des spectres observés de l'atome d'hydrogène. La même année le physicien **Albert Einstein** calcule avec une grande précision la trajectoire de Mercure à l'aide de la relativité générale et la courbure des rayons lumineux dans le champ de gravitation du Soleil. La fin de cette même année il soumet l'article qui décrit les équations de champ de la gravitation, qui vont être la base de la théorie de la relativité générale.

+1914

Le physicien **Ernest Rutherford** montre que les noyaux atomiques chargés positivement contiennent des protons. La même année le physicien **Albert Einstein** et le mathématicien **Marcel Grossmann** publient un article sur le calcul tensoriel, et plus particulièrement sur les tenseurs de Riemann-Christoffel et de Ricci (plus généralement sur l'analyse tensorielle et la géométrie différentielle) et le physicien **Peter Debye** développe un modèle du comportement de la capacité thermique des solides en fonction de la température. Le mathématicien **Felix Hausdorff** introduit les concepts de distance de Hausdorff et de dimension de Hausdorff.

+1913

Le physicien **Niels Bohr** présente le modèle quantique en couches circulaires de l'atome et la même année le physicien **Robert Millikan** mesure la charge électrique fondamentale. La même année, les physiciens **William Henry Bragg** et **William Lawrence Bragg** trouvent la condition de Bragg pour la réflexion des rayons X forts et le physicien **Henry Moseley** montre que le numéro atomique est le vrai critère de discrimination des éléments. Dans le domaine des mathématiques, le mathématicien **Élie Cartan** annonce sa découverte des spineurs. Le physicien **Johannes Stark** démontre qu'un champ électrique va dédoubler les raies de Balmer du spectre de l'hydrogène.

+1912

Le physicien **Max von Laue** propose l'utilisation de réseaux cristallins pour diffracter les rayons X et la même année les physiciens **Walter Friedrich** et **Paul Knipping** diffractent les rayons X par du sulfure de zinc. La même année, le physicien **Ernest Rutherford** propose l'utilisation de la radioactivité comme moyen de datation.

+1911

Le physicien **Ernest Rutherford** découvre le noyau atomique en bombardant une fine feuille d'or de particules alpha. Certaines particules rebondissent sur le noyau des atomes d'or. La même année, le physicien et chimiste **Jean Perrin** prouve l'existence des atomes et des molécules et le physicien **Heike Kammerlingh Onnes** découvre la supraconductivité.

+1909

Les physiciens **Hans Geiger** et **Ernest Marsden** découvrent que des particules alpha peuvent être fortement déviées par des feuilles métalliques minces. La même année, les physiciens **Ernest Rutherford** et **Thomas Royds** démontrent que les particules alpha sont des atomes d'hélium ionisés deux fois. Dans le domaine des mathématiques appliquées, le mathématicien **Agner Krarup Erlang** publie le premier article sur la théorie des files d'attente.

+1908

Le statisticien **William Sealy Gosset** publie un article proposant une nouvelle distribution statistique et un nouveau test statistique appelés respectivement loi de Student et test-t de Student. La même année, le mathématicien **Ernst Friedrich Ferdinand Zermelo** propose une amélioration des axiomes de la théorie des ensembles.

+1907

Le physicien **Albert Einstein** déduit l'expression de la fameuse équivalence entre la masse et l'énergie et la même année il établit l'expression de la capacité calorifique de solides cristallins et calcule le décalage vers le rouge gravitationnel. Le mathématicien et physicien **Hermann Minkowski** réunit l'espace-temps dans une structure mathématique unifiée.

+1906

Le physicien **Walther Nernst** présente une formulation du troisième principe de la thermodynamique. La même année, le mathématicien **Andreï Markov** publie les premiers travaux sur les chaînes d'événements qui porteront plus tard son nom et qui ont occupé une place importante dans la physique quantique de l'époque.

+1905

Le physicien **Albert Einstein** explique l'effet photoélectrique par l'existence de quantum et la même année il explique mathématiquement le mouvement Brownien comme un résultat du mouvement aléatoire des molécules et il publie ses recherches sur sa théorie de la relativité restreinte qui démontre l'équivalence de la masse et de l'énergie. Le physicien **Paul Langevin** publie sa théorie sur la susceptibilité des matériaux paramagnétiques.

+1904

Le physicien **Antoon Lorentz** découvre la contraction du temps dans le sens du déplacement des corps relativement à la constante de la vitesse de la lumière et propose les transformations des équations des forces électromagnétiques. La même année, le physicien **Hantaro Nagaoka** propose le modèle théorique saturnien de l'atome où les électrons tournent autour d'un noyau positif massif comme pour les anneaux de Saturne.

+1903

La radioactivité est expliquée en termes de fission des atomes par le physicien **Ernest Rutherford** et par le radiochimiste **Frederick Soddy**.

+1902

Le physicien **Philipp Lenard** observe que l'effet photoélectrique ne dépend pas de la puissance du faisceau lumineux mais de sa fréquence. La même année le chimiste **Theodor Svedberg** propose que les fluctuations du bombardement moléculaire créent le mouvement brownien et le logicien **Bertrand Russell** énonce le paradoxe "ultime" mettant à mal la théorie naïve des ensembles.

+1901

Les mathématiciens **Gregorio Ricci-Curbastro** et son assistant **Tullio Levi-Civita** développent le calcul tensoriel.

+1900

Le physicien **Max Planck** suggère que la lumière peut être émise en fréquence discrètes en généralisant la loi de rayonnement du corps noir. La même année, le physicien **Johannes Rydberg** affine l'expression mathématique concernant les longueurs d'ondes des raies de l'hydrogène de Balmer et le physicien **Paul Villard** découvre les rayons gamma en étudiant la désintégration de l'uranium. Dans le domaine des mathématiques appliquées, le mathématicien **Louis Bachelier** développe le modèle de mouvement brownien appliqué à la théorie des jeux et de la spéculation qui sera le pilier des outils financiers quantitatifs du 20ème siècle. La même année, le mathématicien **Karl Pearson** définit la distribution statistique du Khi-2 et en étudie des propriétés importantes pour l'inférence statistique. Le physicien **Paul Drude** adapte la théorie cinétique des gaz aux électrons des métaux et obtient un modèle qui porte encore aujourd'hui son nom.

+1899

Le physicien **Ernest Rutherford** découvre que les rayonnements émis par l'uranium sont composés de particules alpha chargées positivement et de particules bêta chargées négativement. La même année, le mathématicien **David Hilbert** remplace les cinq axiomes usuels de la géométrie euclidienne par 21 axiomes pour éliminer les faiblesses de la géométrie d'Euclide.

+1898

Le mathématicien **David Hilbert** donne une première approche du corps de classes. La même année, les physiciens **Marie et Pierre Curie** isolent et étudient le radium et le polonium et le physicien **Wilhelm Carl Werner Wien** identifie une nouvelle particule de charge positive approximativement égale à la masse de l'hydrogène qu'il va appeler le "proton". L'ingénieur **Alfred-Marie Liénard** calcule, le champ électromagnétique produit par une charge ponctuelle, animée d'un mouvement absolument quelconque. Les expressions mathématiques qui ont été établies de façon indépendante, mais 2 ans plus tard, par le physicien **Emil Wiechert**.

+1897

Le physicien **Joseph John Thomson** mesure le rapport charge/masse de certaines particules négatives créées par des rayons cathodiques. Il mesure leur charge, et il en conclut que leur masse est environ 2'000 fois inférieure à celle de l'hydrogène. Ces particules sont plus tard appelées "électrons", expression suggérée par le physicien **George Johnstone Stoney**. Les téléviseurs et autres écrans cathodiques sont des versions améliorées du dispositif de Thomson.

+1896

Le physicien **Henri Becquerel** découvre la radioactivité de l'uranium et la même année le physicien **Pieter Zeeman** étudie la décomposition des raies D du sodium quand il est chauffé dans un champ magnétique intense et découvre que les raies spectrales d'une source de lumière soumise à un champ magnétique possèdent plusieurs composantes, chacune d'elles présentant une certaine polarisation. Pour expliquer le phénomène, il faut alors ajouter un nombre quantique supplémentaire appelé "nombre quantique magnétique". La même année, le physicien **Wilhelm Carl Werner Wien** établit la loi qui porte son nom pour l'énergie émise par le corps noir.

+1895

Le physicien **Wihelm Röntgen** découvre les rayons X et la même année le physicien et inventeur **Guglielmo Marconi** réalise dans les Alpes suisses à Salvan la première transmission sans fil "longue distance" de 1.5 kilomètre.

+1893

Le mathématicien, physicien, philosophe et ingénieur **Henri Poincaré** publie ses études sur le problème des trois corps et initie l'étude qualitative des systèmes d'équations différentielles et de la théorie du chaos. La même année le mathématicien **Georg Cantor** développe la théorie des ensembles transfinis et la proposition de l'ingénieur **Nikola Tesla** d'utiliser le courant alternatif au lieu de continu est adoptée par le premier État américain.

+1892

Le physicien autodidacte **Oliver Heaviside** réduit les 8 équations de l'électrodynamique de Maxwell à 4 équations différentielles.

+1891

Le mathématicien **Giuseppe Peano** introduit le symbole d'appartenance ainsi qu'une première version de l'écriture des quantificateurs. Leur forme définitive sera donnée par le mathématicien **David Hilbert**. Il donne plus de 40'000 définitions dans une langue qu'il veut universelle.

+1890

L'étude systématique des groupes s'amplifie avec les mathématiciens **Sophus Lie**, **Issai Schur** et **Élie Cartan**. Ce dernier introduit la notion de groupe algébrique.

+1889

Le mathématicien **Giuseppe Peano** postule 5 propriétés des nombres entiers comme axiomes dans l'idée de faire avec les entiers ce que Euclide à fait pour la géométrie.

+1888

L'anthropologue, explorateur, géographe, inventeur, météorologue, proto-généticien, psychométricien et statisticien... **Francis Galton** définit statistiquement le concept de coefficient de corrélation. La même année, le mathématicien **Richard Dedekind** propose la définition d'un ensemble fini.

+1887

Les physiciens **Albert Michelson** et **Edward Morley** mesurent la vitesse de la lumière pour vérifier l'hypothèse de l'éther que leurs résultats expérimentaux rejettent et la même année le physicien **Heinrich Hertz** découvre l'effet photoélectrique et fait des expériences sur les ondes électromagnétiques (production et réception).

+1886

Le mathématicien et physicien **Olivier Heaviside** introduit la manipulation d'opération différentiels comme des entités algébriques ce qui amènera aux transformées de Laplace plus tard.

+1885

Le chimiste **Johan Balmer** trouve l'expression mathématique qui donne la longueur d'onde des différentes raies du spectre de l'hydrogène.

+1884

Le physicien et chimiste **Willard Gibbs** pose la notation encore d'usage au début du 21ème siècle pour le produit scalaire et vectoriel ainsi que pour les opérateurs vectoriels différentiels dans ses ouvrages de cours pour le calcul vectoriel. La même année, le physicien **Ludwig Boltzmann** développe la loi de Stefan-Boltzmann du flux de rayonnement du corps noir à partir de considérations thermodynamiques et le physicien **John Henry Poynting** introduit le vecteur qui porte encore aujourd'hui son nom.

+1882

Le mathématicien **Ferdinand von Lindemann** prouve la transcendance de π . La même année, l'astronome, mathématicien, économiste et statisticien **Simon Newcomb** observe un excès de 43" par siècle dans précession de l'orbite de Mercure. Là où la mécanique de Newton explique correctement la période de précession des autres planètes, elle échoue à expliquer celle de Mercure.

+1886

Le mathématicien et physicien **Olivier Heaviside** réduit les équations de Maxwell de 8 à 4 équations et définit la fonction escalier qui porte encore son nom aujourd'hui (fonction d'Heaviside).

+1879

Le physicien et mathématicien **Joseph Stefan** publie la loi de Stefan qui indique que la puissance émise dans l'ensemble du domaine spectral est proportionnelle à la puissance quatrième de la température absolue d'une étoile et à la surface de celle-ci. Pour une même température superficielle, une étoile est d'ailleurs plus lumineuse plus elle est grosse. La même année, le physicien **Edwin Herbert Hall** découvre qu'un courant électrique traversant un matériau baignant dans un champ magnétique engendre une tension perpendiculaire au déplacement initial courant électrique.

+1878

Le mathématicien et philosophe **William Kingdon Clifford** introduit l'opérateur divergence.

+1877

Le physicien et chimiste **Willard Gibbs** définit pour les réactions chimiques deux quantités très utiles, à savoir l'enthalpie qui représente la chaleur d'une réaction à pression constante, et l'enthalpie libre qui détermine si oui ou non une réaction peut procéder de façon spontanée à température et pression constante.

+1876

Le mathématicien et philosophe **William Kingdon Clifford** suggère que le mouvement de la matière peut être dû à des changements dans la géométrie de l'espace.

+1874

Le physicien **Lord Kelvin** formalise le seconde principe de la thermodynamique. La même année, l'économiste mathématicien **Léon Walras** publie son *Éléments d'économie politique pure*.

+1873

Le mathématicien **Georg Cantor** jette les bases de la théorie des ensembles et des cardinaux et montre que les nombres algébriques sont en fait dénombrables. La même année le physicien **James Clerk Maxwell** démontre que la lumière est un phénomène électromagnétique et réduit les équations de l'électrodynamique à 8 au lieu de 20 (au passage il introduit l'opérateur rotationnel) et le physicien **Johannes van der Waals** introduit l'idée qu'il existe des forces faibles attractives entre les molécules. Le mathématicien **Charles Hermite** prouve la transcendance de e .

+1872

Le mathématicien **Karl Weierstrass** expose à l'Académie royale des sciences de Berlin l'exemple d'une fonction continue partout et dérivable nulle part.

+1871

Le chimiste **Dmitri Mendeleïev** examine son tableau périodique et prédit l'existence du gallium, du scandium et du germanium. La même année le physicien **James Clerk Maxwell** établit les relations thermodynamiques de Maxwell.

+1870

Le physicien **Rudolph Clausius** prouve le théorème (scalaire) du Viriel.

+1869

Le chimiste **Dmitri Mendeleïev** propose le tableau périodique des éléments qui porte encore aujourd'hui son nom.

+1867

L'historien, journaliste, philosophe, économiste, sociologue Karl Marx publie *Le Capital*.

+1866

Le physicien **James Clerk Maxwell** formule, indépendamment du physicien **Ludwig Boltzmann**, la théorie cinétique des gaz dite de Maxwell-Boltzmann. La même année, le moine et botaniste **Johann Gregor Mendel** formule les lois statistiques de l'hybridation et le mathématicien and the **Leopold Kronecker** utilise pour la première fois le symbole qui porte encore aujourd'hui son nom.

+1865

Le physicien **James Clerk Maxwell** publie pour la première fois les équations de l'électrodynamique sous la forme des 20 équations à 20 inconnues à l'aide de quaternions.

+1862

Le physicien **Gustav Kirchhoff** élabore le concept de Corps Noir capable d'absorber et d'émettre des rayonnements à toutes les fréquences et établit que l'énergie émise dépend seulement de la fréquence de la radiation émise et de la température du corps noir.

+1859

Le physicien **James Clerk Maxwell** découvre la loi de distribution des vitesses moléculaires. La même année l'astronome **Urbain Le Verrier** signale une anomalie dans le mouvement de Mercure que la théorie de Newton ne prévoit pas et le physicien **Gustav Kirchhoff** avec le chimiste **Robert Wilhelm Bunsen** développent la spectroscopie par prisme.

+1858

Le mathématicien et avocat **Arthur Cayley** dégage la notion d'espace vectoriel, la notion de matrice et en expose l'usage en faisant emploi de la multiplication des matrices ainsi que des déterminants ; il réécrit les systèmes d'équations linéaires sous forme de matrices. Ces travaux sont souvent considérés comme l'émergence de l'algèbre linéaire.

+1855

Le physicien et astronome **Léon Foucault** découvre que la force nécessaire à la rotation d'un disque de cuivre augmente quand il doit tourner avec sa jante entre les pôles d'un aimant, le disque chauffant dans le même temps du fait des "courants de Foucault" induits dans le métal.

+1854

Le mathématicien **George Boole** publie son système de logique symbolique, aujourd'hui connu sous le nom d'algèbre de Boole. La même année, le mathématicien **Arthur Cayley** montre que les quaternions peuvent être utilisés pour représenter les rotations dans l'espace et **Georg Friedrich Bernhard Riemann** donne une nouvelle définition de l'intégrale et jette les bases de la géométrie différentielle.

+1852

Les physiciens **James Joule** et **William Thomson Kelvin** démontrent qu'un gaz en extension se refroidit rapidement.

+1851

Le physicien et astronome **Léon Foucault** fait une démonstration spectaculaire de la rotation de la Terre en suspendant un pendule à un long câble attaché à la coupole du Panthéon à Paris. La même année, le mathématicien **Georg Friedrich Bernhard Riemann** publie des travaux sur les premières fonctions d'une variable complexe.

+1850

Les mathématiciens **Arthur Cayley** et **James Joseph Sylvester** introduisent le terme de matrices et la même année, le mathématicien **Richard Dedekind** introduit le terme de corps. Le physicien **Rudolf Clausius** développe la théorie mécanique de la chaleur et formule le second principe de la thermodynamique. Le physicien **George Stokes** prouve le théorème de Stokes.

+1849

Le mathématicien et astronome **Edouard Roche** trouve le rayon limite de destruction de création par forces de marées et d'un corps qui tient par sa seule gravité utilise ce résultat pour expliquer pourquoi les anneaux de Saturne ne se condensent en un satellite.

+1848

Le physicien **William Thomson Kelvin** découvre le point absolu 0 de température et définit sa propre unité de mesure. La même année le physicien et astronome **Hyppolite Fizeau** transpose les résultats de **Christian Doppler** à la lumière qui, comme le son, possède un caractère ondulatoire (effet Doppler-Fizeau) et met en évidence le décalage vers le rouge et vers le bleu.

+1847

Le physicien **James Joule** trouve expérimentalement l'équivalent mécanique de la chaleur et la même année, le physiologiste et physicien **Hermann Helmholtz** formalise le concept de conservation de l'énergie.

+1845

Le physicien **Gustav Kirchhoff** définit le concept de potentiel électrique et énonce les lois de réseaux qui portent son nom (loi des noeuds, loi des mailles). La même année, le physicien **George Stokes** publie ce qui sera les fondements des équations de Navier-Stokes en mécanique des fluides et le physicien **Michael Faraday** découvre que la propagation de la lumière dans un matériau peut être influencé par des champs magnétiques externes.

+1844

Le mathématicien **Joseph Liouville** montre l'existence d'une infinité de nombres transcendants.

+1843

Le physicien, mathématicien et astronome **William Rowan Hamilton** définit des espaces de vecteurs complexes (quaternions). La notion d'espace vectoriel sera clairement définie par le mathématicien et astronome **August Ferdinand Möbius** et par le mathématicien et linguiste **Giuseppe Peano** 40 ans plus tard. La même année, le mathématicien **Laurent Pierre Alphonse** publie son mémoire sur ce qui deviendra plus tard les séries qui portent son nom en analyse complexe.

+1842

Le principe de la conservation de l'énergie est formulé par le physicien **Julius Von Mayer** qui a calculé la quantité de travail pouvant être obtenue par transformation de la chaleur en énergie, soit l'équivalent mécanique de la calorie. La même année, le physicien **Christian Doppler** découvre l'effet acoustique qui porte son nom (variation de la fréquence avec le mouvement relatif).

+1841

Le mathématicien **Karl Weierstrass** découvre mais ne publie pas ce que nous appelons aujourd'hui les séries de Laurent. La même année, le mathématicien **Carl Gustav Jacob Jacobi** introduit les matrices jacobiniennes et réintroduit la notation de la dérivée partielle proposée initialement par le mathématicien **André-Marie Legendre**.

+1838

L'astronome et mathématicien **Friedrich Bessel** mesure que la distance qui nous sépare de l'étoile 61 Cygni est de 96 billions de kilomètres.

+1835

Le mathématicien **Carl Friedrich Gauss** donne une construction rigoureuse des nombres complexes et le mathématicien **Augustin Louis Cauchy** établit une première théorie des déterminants. La même année, le mathématicien et ingénieur **Gaspard Coriolis** démontre que l'accélération d'un mobile dans un référentiel en rotation est soumise à une force complémentaire perpendiculaire au sens de déplacement du mobile dans ce référentiel.

+1834

L'ingénieur et physicien **Émile Clapeyron** présente une formulation de la seconde loi de la thermodynamique. La même année, le physicien **Heinrich Lenz** établit la loi d'induction électromagnétique.

+1832

Le physicien **Michael Faraday** établit la théorie fondamentale de l'électrolyse.

+1831

Le physicien **Michael Faraday** découvre l'induction électromagnétique, à savoir l'obtention d'un courant électrique à partir de la variation d'un champ magnétique (principe de la dynamo: le contraire de l'expérience d'**Ørsted**). La même année le mathématicien, astronome et physicien **Carl Friedrich Gauss** énonce deux des quatre équations de Maxwell.

+1829

Le mathématicien **Évariste Galois** présente la première ébauche de son travail sur les équations résolubles qui sera à l'origine de l'approche ensembliste de la résolution d'équation algébrique par radicaux.

+1828

Le physicien **George Green** prouve le théorème de Green

+1827

Le botaniste **Robert Brown** découvre le mouvement brownien des particules de pollen et de colorant dans l'eau ; la même année le physicien **Georg Ohm** établit la loi de la résistance électrique et le physicien et mathématicien **André Ampère** découvre les lois qui lient les forces magnétiques au courant électrique. Le physicien, mathématicien et astronome **William Rowan Hamilton** présente la théorie d'une unique fonction qui unifie la mécanique, optique et mathématiques et qui aida à établir la théorie ondulatoire de la lumière.

+1825

Le mathématicien **Augustin-Louis Cauchy** présente le théorème de Cauchy et l'intégration curviligne générale et introduit le théorème des résidus. Le scientifique et inventeur **William Sturgeon** invente le premier électroaimant.

+1824

Le physicien et ingénieur **Sadi Carnot** analyse scientifiquement l'efficacité des moteurs à vapeur (cycle de Carnot) en montrant que leur rendement est limité et pose le second principe de la thermodynamique.

+1823

Le physicien et chimiste **Michael Faraday** présente une série de papiers sur la liquéfaction des gaz.

+1822

Le mathématicien **Jean-Victor Poncelet** fonde la géométrie projective. La même année, le physicien et mathématicien **Joseph Fourier** présente formellement l'utilisation des dimensions (unités) pour des quantités physiques.

+1821

Le principe de la dynamo est décrit pour la première fois par le physicien et chimiste **Michael Faraday**. La même année le physicien **John Herapath** propose que la chaleur n'est en réalité que l'effet d'agitation et donc de mouvement de corps élémentaires.

+1820

Le physicien **Hans Christian Ørsted** découvre et prouve les effets magnétiques du courant électrique. La même année, les physiciens **Jean-Baptiste Biot** et **Félix Savart** déterminent dans le domaine du magnétisme la fameuse loi qui porte leur nom.

+1818

Le mathématicien, géomètre et physicien **Siméon Poisson** calcule la tache lumineuse au centre de l'ombre d'un obstacle circulaire opaque.

+1817

En étudiant la polarisation de la lumière, le physicien **Augustin Fresnel** montre que cette dernière est un mouvement ondulatoire transversal et non longitudinal et montre aussi que la diffraction et l'interférence peuvent être expliquées si l'on considère la lumière comme une onde. La même année, l'astronome **Friedrich Bessel** publie des travaux faisant usage des fameuses fonctions qui portent son nom.

+1816

Le mathématicien **Joseph Diaz Gergonne** introduit le symbole marquant l'inclusion dans la théorie des ensembles.

+1814

Le physicien et opticien **Joseph Von Fraunhofer** étudie pour la première fois les raies d'absorption du spectre solaire et ce au moyen du spectroscopie dont il fût l'inventeur.

+1812

Le mathématicien, astronome et physicien **Pierre-Simon de Laplace** publie un ouvrage majeur sur la théorie des probabilités (incluant aussi la méthode des moindres carrés) dont il est considéré comme l'un des fondateurs.

+1811

Le chimiste **Amaedo Avogadro** avance l'hypothèse selon laquelle des volumes égaux de gaz différents contiennent le même nombre de molécules, dans des conditions identiques de température et de pression.

+1810

Le mathématicien, astronome et physicien **Carl Friedrich Gauss** découvre les concepts de base de la géométrie non-euclidienne mais ne publiera jamais ses travaux à ce sujet. La même année, le physicien et mathématicien **Joseph Fourier** modélise l'évolution de la température au travers de séries trigonométriques.

+1809

Le mathématicien, astronome et physicien **Carl Friedrich Gauss** développe la méthode des moindres carrés indépendamment du mathématicien **André-Marie Legendre**. La même année, le mathématicien, astronome et physicien **Pierre-Simon de Laplace** démontre la forme générale du théorème central limite. L'ingénieur, physicien et mathématicien **Etienne Malus** publie la loi de Malus.

+1808

Le physicien et chimiste **John Dalton** propose ce qui est considéré comme la première théorie de l'atome.

+1806

Le mathématicien **Jean Robert Argand** publie la première représentation plane des nombres complexes et utilise des mesures algébriques.

+1805

Le mathématicien **André-Marie Legendre** développe la méthode des moindres carrés.

+1803

Le physicien et chimiste **John Dalton** a l'idée originale de considérer que chaque élément chimique est constitué d'atomes différents. Une combinaison chimique s'explique alors par l'union de ces atomes en proportions fixes et les masses atomiques relatives devenaient calculables à partir de faits expérimentaux. La même année, l'économiste, journaliste et industriel **Jean-Baptiste Say** publie son *Traité d'économie politique*.

+1802

Le physicien **Thomas Young** prouve la nature ondulatoire de la lumière par une expérience importante qui montre l'interférence des ondes. La même année, le chimiste et physicien **Louis Joseph Gay-Lussac** découvre la fameuse loi sur les gaz qui relie volume et température d'un gaz réel, loi qui porte son nom (loi de Gay-Lussac).

+1801

Le chimiste et physicien **John Dalton**, découvre la loi de la somme des pressions partielles des gaz qui porte encore aujourd'hui son nom.

+1800

Le chimiste **William Nicholson** et le chirurgien **Anthony Carlisle** utilisent l'électrolyse pour séparer l'eau en hydrogène et oxygène. La même année, l'astronome **William Herschel** découvre le rayonnement infrarouge et le physicien **Allessandro Volta** invente la première pile électrique.

+1799

Le mathématicien **Gaspard Monge** publie son ouvrage de géométrie descriptive. Il en est considéré comme l'inventeur.

+1798

Le mathématicien **Carl Friedrich Gauss** donne une démonstration rigoureuse du théorème de d'Alembert (théorème fondamental de l'algèbre). La même année, le physicien **Benjamin Thompson** a l'idée que la chaleur est une forme d'énergie et le physicien et chimiste **Henry Cavendish** mesure la valeur de la constante gravitationnelle. L'économiste **Thomas Malthus** énonce sa loi de population.

+1797

Le mathématicien **Caspar Wessel** associe des vecteurs aux nombres complexes et étudie l'interprétation géométrique des opérations sur les nombres complexes.

+1793

L'assemblée nationale de la République Française instaure le système métrique.

+1789

Le physicien et chimiste **Antoine Lavoisier** énoncé le principe de conservation de la masse.

+1787

Le physicien, chimiste et inventeur **Jacques Alexandre César Charles** détermine expérimentalement que le volume d'une masse fixe d'un gaz à pression constante est proportionnel à la température.

+1786

L'astronome **William Herschel** fait une description précise de notre galaxie.

+1785

Le physicien **Charles-Augustin Coulomb** prouve que les forces entre charges électriques et entre aimants s'exercent en raison inverse du carré de la distance.

+1781

Le chimiste et physicien **Joseph Priestley** crée de l'eau par combustion d'hydrogène et d'oxygène ce qui démontre que l'eau n'est pas un élément fondamental comme on le pensait depuis Aristote.

+1778

Les physiciens et chimistes **Carl Scheele** et **Antoine Lavoisier** découvrent que l'air est composé essentiellement d'azote et d'oxygène.

+1777

Le physicien et mathématicien **Leonhard Euler** introduit la lettre *i* pour la partie imaginaire des nombres complexes.

+1776

Le philosophe moraliste et pionnier de l'économie politique **Adam Smith** publie son *Une enquête sur la nature et les causes de la richesse des nations*.

+1774

Le mathématicien, astronome et physicien **Pierre-Simon de Laplace** explicite l'intégrale d'Euler. La même année, le théologien, pasteur dissident, philosophe naturel, pédagogue et théoricien de la politique **Joseph Priestley** fit sa principale découverte, celle de l'oxygène.

+1772

Le mathématicien, mécanicien et astronome **Joseph-Louis Lagrange** étudie le problème des trois corps et découvre les points de libration appelés aujourd'hui "points de Lagrange".

+1766

Le physicien et chimiste **Henry Cavendish** découvre et étudie l'hydrogène.

+1764

Les chaleurs latente et spécifique sont décrites pour la première fois par le physicien et chimiste **Joseph Black**. Il est également le premier à distinguer nettement température et quantité de mouvement. La même année, le physicien et mathématicien **Leonhard Euler** examine l'équation aux dérivées partielles pour la vibration d'un tambour circulaire et trouve l'une des fonctions de Bessel en tant que solution.

+1763

Un article posthume de mathématicien et pasteur **Thomas Bayes** met en évidence qu'il a découvert ce qui est appelée encore aujourd'hui le "théorème de Bayes".

+1756

Le mathématicien, mécanicien et astronome **Joseph-Louis Lagrange** développe la mécanique analytique en se basant sur son invention du calcul variationnel.

+1755

Le physicien et mathématicien **Leonhard Euler** introduit la lettre grecque sigma majuscule (S) pour le symbole de la somme.

+1750

Le mathématicien **Gabriel Cramer** établit la règle de Cramer pour la résolution de systèmes linéaires.

+1749

L'astronome et physicien **Jean le Rond D'Alembert** développe le premier modèle de précession des équinoxes basé sur la théorie de la gravitation de Newton et donne une piste de solution pour le problème des trois corps.

+1746

L'encyclopédiste **Jean le Rond D'Alembert** donne la première preuve (acceptable mais qui sera corrigée plus tard) du théorème fondamental de l'algèbre. L'année d'après (1747) il publie l'équation des cordes vibrantes qui a été le premier exemple de l'équation des ondes. Cela fait de D'Alembert, l'un des fondateurs de la physique mathématique.

+1744

Le philosophe, mathématicien, physicien, astronome et naturaliste **Pierre Louis Moreau de Maupertuis** énonce le principe de moindre action qui sera formalisé mathématiquement 22 ans plus tard par le mathématicien, mécanicien et astronome **Joseph-Louis Lagrange**. La même année, le physicien et mathématicien **Leonhard Euler** montre l'existence des nombres transcendants.

+1742

L'astronome **Anders Celsius** définit sa propre unité de mesure de la température.

+1739

Le physicien et mathématicien **Leonhard Euler** résout les équations différentielles linéaires homogènes à coefficients constants.

+1738

Le médecin, physicien et mathématicien **Daniel Bernoulli** publie un ouvrage sur l'hydrodynamique introduisant la théorie cinétique des gaz et le fameux théorème de Bernoulli (équilibre des pressions).

+1737

Le physicien et mathématicien **Leonhard Euler** résout le problème de théorie des graphes relatif aux ponts de Königsberg. La résolution de ce problème est considérée comme le premier théorème de la théorie des graphes. Il établit par la même occasion la "formule d'Euler" liant le nombre de sommets, d'arêtes et de faces d'un polyèdre convexe, et donc d'un graphe planaire.

+1736

L'inventeur **Jonathan Hulls** dépose le premier brevet d'un bateau propulsé par une machine à vapeur.

+1734

Le physicien et mathématicien **Leonhard Euler** introduit la notation $f(x)$ pour une fonction appliquée à l'argument x .

+1733

Le mathématicien **Geralamo Saccheri** étudie ce que serait la géométrie si les 5ème postulat d'Euclide était faux.

+1729

Le teinturier et amateur de physique et d'astronomie **Stephen Gray** est le premier à découvrir la transmission de l'électricité dans des matériaux qu'il appellera des "conducteurs".

+1727

Le physicien et mathématicien **Leonhard Euler** introduit la notation moderne des fonctions trigonométriques et la lettre e pour la base du logarithme naturel (également connue occasionnellement sous le nom de "nombre d'Euler").

+1724

Le mathématicien **Abraham De Moivre** étudie les statistiques de mortalité et fonde la théorie des annuités sur la vie.

+1715

Le mathématicien **Brook Taylor** publie les outils permettant de faire des intégrations par parties ainsi que des développements en série de fonctions (les fameuses séries de Taylor).

+1714

Le mathématicien **Brook Taylor** dérive la fréquence fondamentale de vibration d'une corde tendue en fonction de sa tension et de sa densité linéique en résolvant une équation différentielle ordinaire.

+1713

Le mathématicien et physicien **Jacques Bernoulli** publie les principes rigoureux de base du calcul des probabilités et statistique.

+1705

L'astronome **Edmund (ou Edmond) Halley** prédit avec une erreur quasi négligeable à l'aide du calcul que la comète qui est passée près de la Terre en 1682 repassera en 1758.

+1704

Le physicien et mathématicien **Isaac Newton** constate expérimentalement que la lumière blanche se compose de multiples couleurs. Il suppose également qu'un rayon lumineux est formé de corpuscules.

+1696

Le mathématicien et physicien **Jacques Bernoulli** pose clairement le problème de la courbe brachistochrone (qui appartient à la famille des courbes cycloïdes) et en propose une solution. La même année, le mathématicien **Guillaume de L'Hopital** présente sa règle pour l'examen des formes indéterminées.

+1693

L'astronome et ingénieur **Edmund Halley** découvre la relation qui lie la focale d'une lentille avec la distance de l'image à son axe et de l'objet réel à son axe. La même année, il construit la première table statistique de mortalité liant le taux de décès à l'âge.

+1691

Le philosophe et mathématicien **Gottfried Leibniz** découvre une technique de séparatoïn de variables pour les équations différentielles ordinaires.

+1690

La théorie ondulatoire de la lumière est avancée par le physicien et astronome **Christian Huygens** ; la même année le physicien et mathématicien **Jean Bernoulli** développe le calcul exponentiel et trouve l'équation de la chaînette. La même année, le mathématicien et physicien **Jacques Bernoulli** (frère de Jean Bernoulli) développe le calcul intégral.

+1687

Le physicien et mathématicien **Isaac Newton** publie un ouvrage où il explique la force de gravité et les orbites des planètes. Il énonce également les trois lois de la dynamique. Il s'agit de la première révolution scientifique (avant la relativité restreinte/générale et la physique quantique).

+1685

Le philosophe et mathématicien **Gottfried Leibniz** résout les systèmes linéaires en usant sans justification théorique de matrices et de déterminants.

+1682

Le physicien et mathématicien **Isaac Newton** établit la loi de la gravitation qui porte aujourd'hui son nom.

+1679

Le philosophe et mathématicien **Gottfried Leibniz** introduit l'arithmétique binaire et met au point une machine à calculer qui effectue les 4 opérations. La même année, le physicien, mathématicien et inventeur **Denis Papin** montre expérimentalement l'influence de la pression atmosphérique sur le point d'ébullition de l'eau.

+1678

Le mathématicien, astronome et physicien **Christian Huygens** postule le son principe du front d'onde.

+1676

Le physicien **Robert Hooke** énonce que l'étirement d'un ressort est proportionnel à la tension.

+1675

L'astronome **Olaüs Röemer** réalise des mesures précises de la vitesse de la lumière. La même année, le physicien et astronome **Isaac Newton** invente un algorithme pour calculer les racines de fonctions.

+1673

Le philosophe et mathématicien **Gottfried Leibniz** invente son calcul différentiel.

+1669

Le mathématicien, astronome et physicien **Christian Huygens** publie des résultats concernant l'observation de la conservation d'énergie cinétique devenant ainsi in extenso le découvreur du concept d'énergie cinétique.

+1668

Le physicien et astronome **Isaac Newton** réalise le premier télescope à réflexion et la même année le mathématicien **John Wallis** suggère la loi de conservation de la quantité de mouvement.

+1665

Le physicien **Isaac Newton** formule les trois lois de la mécanique. Il jette les bases du calcul différentiel, ces techniques lui permettant à partir de l'expression d'une force inverse au carré de la distance, de retrouver la forme générale des lois de Kepler.

+1661

Le fondateur de la démographie statistique **John Graunt** publie la première table de mortalité ; la même année le physicien et chimiste **Robert Boyle** détermine les lois de compressibilité des gaz portant aujourd'hui son nom attaché parfois à celui du physicien **Edme Mariotte** qui redécouvrit quelques années après les mêmes lois.

+1659

Le mathématicien, astronome et physicien **Christian Huygens** découvre la formule de l'isochronisme rigoureux (lorsque l'extrémité du pendule parcourt un arc de cycloïde, la période d'oscillation est constante quelle que soit l'amplitude).

+1658

Le mathématicien, astronome et physicien **Christian Huygens** découvre expérimentalement que les balles placées n'importe où sur une cycloïde renversée atteignent le point le plus bas de la cycloïde dans le même temps et ainsi montre expérimentalement que l'isochronisme de la cycloïde.

+1657

Le juriste et mathématicien **Pierre de Fermat** énonce son "principe de Fermat" en optique comme quoi la lumière se propage d'un point à un autre sur des trajectoires telles que la durée du parcours soit localement minimale.

+1655

Le mathématicien, astronome et physicien **Christian Huygens** est le premier à utiliser le concept d'espérance en probabilités.

+1654

Le mathématicien, physicien, inventeur, philosophe, moraliste et théologien **Blaise Pascal** et le juriste et mathématicien **Pierre de Fermat** créent la théorie des probabilités.

+1644

Le physicien et mathématicien **Evangelista Torricelli** a l'idée de substituer du mercure à de l'eau dans l'expérience dite de Torricelli de mise en évidence du "grosso-vido"; suivront plus tard les travaux du mathématicien, physicien, inventeur, philosophe, moraliste et théologien **Pascal Blaise** (expérience du Puy de Dôme).

+1638

Le mathématicien, géomètre, physicien et astronome **Galileo Galilée** publie la relation mathématique définissant la période du pendule simple.

+1637

Le philosophe et mathématicien **René Descartes** renomme les inconnues x , y , z et les paramètres a , b , c et étend l'usage de l'algèbre aux longueurs et au plan, créant avec le juriste et mathématicien **Pierre de Fermat** la géométrie analytique. La même année, toujours **René Descartes**, détermine quantitativement les angles primaires et secondaires des arc-en-ciels respectivement à l'inclinaison du Soleil.

+1631

Le mathématicien **Thomas Harriot** introduit, dans une publication posthume, les symboles $>$ et $<$. La même année le mathématicien et théologien **William Oughtred** donne pour la première fois le symbole multiplié.

+1629

L'avocat et mathématicien **Pierre de Fermat** développe un calcul différentiel rudimentaire.

+1624

Invention du premier thermomètre (dont les graduations ne sont bien évidemment pas normalisées...) par le médecin **Santorio Santorio**.

+1621

L'astronome et physicien **Willebrord Snell** découvre que l'angle de réfraction de la lumière est déterminé par le sinus de l'angle formé par la lumière incidente avec la normale au dioptr.

+1620

L'ingénieur **Francis Thomas Bacon** défend la méthode expérimentale et mène de nombreuses observations sur la chaleur. Il suggère que la chaleur est reliée au mouvement.

+1619

L'astronome et mathématicien **Johannes Kepler** a fini de publier les trois lois relatives au mouvement des planètes.

+1611

L'astronome et mathématicien **Johannes Kepler** découvre la réflexion interne totale, la loi de réfraction aux petits angles de réfraction et l'optique des lentilles minces.

+1614

Le mathématicien **John Napier** (en français John **Neper**) invente les logarithmes, qui ramènent les opérations de multiplication et de division à de simples additions ou soustractions.

+1608

L'opticien **Hans Lippershey** invente le télescope qui sera utilisé et amélioré (avec une qualité aléatoire) l'année suivante par le mathématicien, géomètre, physicien et astronome **Galileo Galilée** pour confirmer les théories de Copernic.

+1603

Le mathématicien et astronome **Thomas Harriot** détermine la façon de calculer qualitativement la surface d'un triangle sphérique.

+1591

Le mathématicien **François Viète** ouvre une nouvelle période de l'algèbre en faisant opérer les calculs sur des lettres, utilisant les voyelles pour désigner les inconnues et les consonnes pour les paramètres. Par ailleurs, il donne le développement du binôme de Newton.

+1590

L'astronome **Galileo Galilei** démontre expérimentalement que tous les corps en chute libre ont une accélération identique. La même année, les opticiens **Hans** et **Zachariaus Janssen** créent le premier microscope en associant plusieurs lentilles ce qui définira les débuts de la biologie et la médecine scientifique.

+1586

L'ingénieur et physicien **Simon Stevin** démontra la méthode du parallélogramme des forces et découvre que la pression d'un liquide sur le fond d'un récipient est indépendante de sa forme, et aussi de la surface du fond; elle dépend seulement de la hauteur d'eau dans le récipient. Il donna aussi la mesure de la pression sur n'importe quelle portion du côté d'un récipient.

+1576

L'astronome **Tycho Brahé** observe une nouvelle étoile dans la constellation de Cassiopée et construit un observatoire dans l'île de Hveen.

+1572

Le mathématicien **Rafaele Bombelli** donne une formulation des nombres complexes et les règles de calculs effectifs. Il introduisit les termes più di meno (pdm) et meno di meno (mdm) pour représenter $+i$ et $-i$.

+1548

Le mathématicien et physicien **Simon Stevin** écrit les puissances du dixième cernées d'un exposant. Il donne la première écriture des vecteurs.

+1545

Le mathématicien **Ludovico Ferrari** donne la solution des équations de degré 4.

+1543

L'ouvrage de l'astronome **Nicolas Copernic** résumant 26 ans de recherches et d'observations est publié et met clairement en avant que le système héliocentrique de Ptolémée n'est pas valide.

+1536

Le mathématicien **Niccolò Fontana** développe la science de la balistique.

+1530

Le mathématicien et physicien **Robert Recorde** introduit le signe $=$ et le mathématicien **Michael Stifel** développe une première forme de notation algébrique.

+1525

Le mathématicien **Christoff Rudolff** introduit la notation des racines carrées.

+1510

Le peintre, graveur et mathématicien **Albert Dürer** développe les bases de la géométrie descriptive et de la perspective.

+1500

Le mathématicien italien **Scipione del Ferro** parvient pour la première fois à une résolution algébrique d'un grand type d'équations du troisième degré.

+1490

Le peintre, sculpteur, architecte, musicien, mathématicien, inventeur, anatomiste, géologue, cartographe, botaniste et écrivain... **Leonardo da Vinci** décrit le phénomène de capillarité.

+1420

Le mathématicien et astronome **Jamshīd Al- Kāshī** calcule et observe les éclipses solaires de 1406, 1407 et 1408. Il serait également le premier à utiliser la notation décimale en arithmétique et dans les chiffres arabes.

+1400

Le mathématicien et astronome **Jamshīd Al-Kāshī** développe une première forme de la méthode de Newton Regula falsi.

+1269

Le savant **Pierre de Maricourt** invente les expressions de la magnétique "pôle nord" et "pôle sud" et il a été le premier qui a écrit que les pôles opposés d'un aimant s'attirent

+1268

Le philosophe, savant et alchimiste **Roger Bacon** publie des propositions pour réformer l'école, argumentant que pour étudier la nature, l'utilisation des observations des mesures est la seule base rigoureuse de l'expérimentation et de la vérification tout en affirmant ainsi la nécessité des mathématiques.

+1200

Le mathématicien **Jordan de Nemore** introduit la notation des inconnues par des symboles.

+1121

L'astronome, physicien, biologiste, chimiste, mathématicien et philosophe **Abu al-Fath Khāzini** publie un livre dans lequel il propose que la gravité et l'énergie potentielle gravitationnelle varient selon la distance du centre de la Terre. Il fait aussi une distinction entre la force, la masse et le poids. Il invente aussi plusieurs instruments scientifiques, y compris une balance romaine et une balance hydrostatique. Il introduit aussi des méthodes scientifiques expérimentales à la statique et la dynamique, les unifie dans la science de la mécanique et combine l'hydrostatique avec la dynamique pour créer l'hydrodynamique.

+1114

Le mathématicien **Bhaskara** offre un résumé complet de la mathématique hindoue, telle qu'elle s'est développée du 5ème au 7ème siècle de notre ère. Ainsi il reconnaît les racines carrées négatives, résout des équations quadratiques à plusieurs inconnues, des équations d'ordre supérieur comme celles de Fermat ainsi que les équations du second degré générales. Il a également été un pionnier dans le principe du calcul différentiel de près de 500 ans relativement à Newton et Leibniz.

+1100

Le philosophe et physicien **Hibat Allah Abu'l-Barakat al-Baghdaadi** est le premier à nier l'idée d'Aristote selon laquelle une force constante produit un mouvement uniforme ce qui préfigure la seconde loi de Newton sur le mouvement. Comme Newton, il a décrit l'accélération comme étant la variation de vitesse.

+1037

Le mathématicien, physicien et philosophe **Ibn al-Haytham** est conscient de la magnitude de l'accélération due à la gravité. Il découvre la loi de l'inertie, connue aujourd'hui comme la première loi du mouvement de Newton.

+1030

Le philosophe, écrivain, médecin et scientifique **Abu 'Ali al-Husayn Ibn Abd Allah Ibn Sina** (connu en Occident sous le nom de **Avicenne**) note que si la perception de la lumière est due à l'émission d'une sorte de particules par une source lumineuse, la vitesse de la lumière doit être finie. Il a également fourni une explication sophistiquée pour le phénomène de l'arc en ciel. Le mathématicien, astronome, physicien, érudit, encyclopédiste, philosophe, astrologue, voyageur, historien, pharmacologue **Abū al-Rayhān Muhammad ibn Ahmad Al-Biruni**, et plus tard l'astronome **Abu al'Fath Khāzini**, ont été les premiers à appliquer des méthodes scientifiques expérimentales dans la mécanique, en particulier les domaines de la statique et la dynamique, pour déterminer le poids spécifique, tels que ceux basés sur la théorie des équilibres et de pesage.

+1021

Le philosophe, mathématicien et physicien **Ibn al-Haytham** considéré comme le père de l'optique et un pionnier de la méthode scientifique explique correctement la lumière et la vision, et introduit la méthode scientifique expérimentale, jetant les bases de la physique expérimentale. Il discute aussi de la psychologie expérimentale et décrit les divers instruments d'optique comme la chambre noire.

+1019

Le mathématicien, astronome et physicien **Abū Rayhān Al-Biruni** a observé et décrit l'éclipse solaire du 8 avril 1019, ainsi que l'éclipse lunaire du 17 Septembre 1019, en détail ; il a donné la localisation exacte des étoiles lors de l'éclipse lunaire. Il invente l'astrolabe orthographique et le planisphère.

+1010

Le mathématicien **Al-Sijistani** invente le Zuraqi, un astrolabe unique conçu pour un modèle planétaire héliocentrique dans lequel la Terre est en mouvement plutôt que le ciel.

+1000

Le mathématicien, physicien et astronome **Abu Sahl al-Qouhi** découvre que la lourdeur des corps varie en fonction de leur distance du centre de la Terre, et résout des équations plus élevées que le deuxième degré. Durant le même décennie, le mathématicien et ingénieur **Al'Karkhi** écrit un livre contenant les premières preuves connues par induction mathématique. Il l'utilise pour démontrer le théorème du binôme, le triangle de Pascal, et la somme des cubes intégrales.

+996

L'astrolabe mécanique orienté, comportant 8 roues dentées est inventé par le mathématicien, astronome et physicien **Abū Rayhān Al-Biruni** qui est aussi l'auteur de travaux sur la sommation de séries et la combinatoire.

+964

Le mathématicien, philosophe et physicien **Abd al-Rahman al-Soufi** explique le pouvoir grossissant des lentilles et fut un des premiers à se servir d'une méthode d'analyse scientifique qui influencera grandement de futurs scientifiques.

+953

Le mathématicien et ingénieur **Al-Karkhi** définit différents monômes et donne des règles pour les produits de n'importe quels deux d'entre eux. Il a aussi découvert le théorème du binôme pour des exposants entiers.

+952

Le mathématicien **Abu'l-Hasan al-Uqlidisi** modifie les méthodes de calcul pour le système numérique indien pour le rendre possible aux plumes et à l'utilisation du papier. Jusque-là, faire des calculs avec les chiffres indiens nécessitait l'emploi d'une planche.

+900

La première référence à un tube d'observation se trouve dans l'oeuvre de l'astronome et mathématicien **Al-Battani**, et la première description exacte du tube d'observation a été donnée par le mathématicien, astronome, physicien, érudit, encyclopédiste, philosophe, astrologue, voyageur, historien, pharmacologue **Al-Biruni**, dans une section de son travail dédiée à vérifier la présence du nouveau croissant de Lune à l'horizon. Bien que ces tubes d'observations préliminaires n'aient pas de lentilles, ils ont permis à un observateur de se concentrer sur une partie du ciel en éliminant les interférences lumineuses. Ces tubes d'observation ont été adoptés ultérieurement en Europe latine, où ils ont influencé le développement du télescope.

+880

L'astronome et mathématicien **Al-Battani** découvre le mouvement de l'apogée du Soleil, calcule les valeurs de la précession des équinoxes et l'inclinaison de l'axe terrestre. Il est à l'origine de la définition de la fonction trigonométrique tangente et cotangente.

+820

Le mot "algèbre" naît. Le mathématicien, géographe, astrologue et astronome **Muhammad ibn Musa Al'Khwarizmi** est souvent considéré comme le père de l'algèbre médiévale, car il dégage celle-ci de l'emprise géométrique. On lui doit aussi le quadrant, l'instrument mural, le quadrant des sinus qui a été utilisé pour résoudre les problèmes trigonométriques et faire des observations astronomiques.

+800

Des astronomes inventent le cadran solaire universel et le cadran horaire universel à Bagdad.

+780

L'alchimiste **Jabir Ibn Hayyan** introduit la méthode scientifique expérimentale pour la chimie, ainsi que les appareils de laboratoire tels que l'alambic et des processus tels que la distillation pure, la liquéfaction, la cristallisation, et la filtration. Il a également inventé plus de vingt types d'appareils de laboratoire, ce qui a entraîné la découverte de plusieurs substances chimiques. Il a également développé des recettes pour le verre coloré.

+773

Les chiffres arabes (adaptés de l'Inde) font leur première apparition en Europe.

+628

Le mathématicien **Brahmagupta**, donne des règles pour résoudre les équations linéaires et quadratiques. Il découvre que les équations du second degré ont deux racines, dont les négatives et les irrationnelles.

+550

Les mathématiciens hindous donnent une représentation du chiffre zéro dans un système de notation de position.

+499

Le mathématicien **Âryabhat** obtient le nombre complet de solutions d'un système d'équations linéaires par des méthodes équivalentes aux méthodes modernes, et décrit la solution générale de telles équations. Il donne également des solutions d'équations différentielles.

+275

Le mathématicien **Diophante d'Alexandrie** considéré comme le père de l'algèbre étudie les équations à variables rationnelles (incluant donc les équations du second degré) et les équations diophantiennes.

+121

Année correspondant au plus ancien document faisant mention de la pierre magnétique.

+100

L'ingénieur, mécanicien et mathématicien **Héron d'Alexandrie** redécouvre (après les chinois) le concept de force. Il invente aussi un système d'engrenages pour soulever des poids utilisant la puissance de la vapeur. Il fait la première description du sextant (mais ne l'a toutefois pas inventé). Son contemporain astronome **Claude Ptolémée** invente le sextant et décrit l'astrolabe (peut-être inventé par l'astronome, géographe et mathématicien **Hipparque**) et étudie la réfraction et la réflexion. Pendant le même siècle le mathématicien et philosophe **Nicomaque de Gêrasedéfinit ainsi les nombres pairs et impairs, les nombres premiers et composés, les nombres parfaits.**

+80

L'érudit **Wang Ch'ung** réalise la première boussole aimantée sur un plateau de laiton.

-87

Année correspondant à la datation de la machine d'Anticythère, considérée comme le premier calculateur analogique et la première machine à engrenages (une trentaine!) antique permettant de calculer des positions astronomiques complexes. Le soin et l'adresse avec lesquels cette machine fut réalisée, ainsi que les capacités nécessaires en mécanique et en astronomie remettent en question les connaissances historiques sur les sciences grecques avant sa découverte. En effet, aucun objet de même âge et de même complexité n'était connu dans le monde et il faut attendre près d'un millénaire pour voir apparaître des mécanismes comparables! Le physicien, mathématicien et ingénieur **Archimède de Syracuse** en est l'hypothétique créateur.

-134

L'astronome, géographe et mathématicien **Hipparque de Nicéedécouvre la précession des équinoxes.**

-150

L'astronome, géographe et mathématicien **Hipparque** est souvent désigné comme le créateur de la trigonométrie et des tables numériques correspondantes. Il calcule le premier la période de révolution du Soleil autour de la Terre (mais les résultats numériques sont en réalité ceux de la rotation de la Terre autour du Soleil) et élabore la théorie des excentriques et des épicycles.

-200

Dans le siècle, les chinois auraient inventé l'écluse, le gouvernail, le principe de la machine à vapeur plusieurs centaines d'années avant les occidentaux! Pendant ce siècle, le scientifique et ingénieur **Philon de Byzance** rédige des traités sur les leviers, la pneumatique, les automates, la traction et les clepsydres.

-225

Le géomètre et astronome **Apollonius de Perge** publie la première étude sur les coniques donnant à l'ellipse, à la parabole et à l'hyperbole les noms que nous leur connaissons. On lui attribue en outre l'hypothèse des orbites excentriques pour expliquer le mouvement apparent des planètes et la variation de vitesse de la Lune.

-250

Le mathématicien, physicien et ingénieur **Archimède de Syracuse** étudie des machines simples tel que le levier, la fameuse vis permettant entre autre de pomper l'eau ("vis d'Archimède") et découvre le principe d'Archimède qui explique la flottaison. Dans la même décennie, l'astronome, géographe, philosophe et mathématicien **Ératosthène de Cyrène** calcule le diamètre de la Terre à l'aide d'un gnomon et de son ombre et démontra l'inclinaison de l'écliptique.

-260

Le mathématicien, physicien et ingénieur **Archimedes of Syracuse** calcule π avec une précision de deux décimales utilisant des polygones inscrits et circonscrits et calcule la surface sous un segment de parabole.

-281

L'astronome et mathématicien **Aristarque de Samos** fait l'hypothèse que le Soleil occupe le centre du système solaire et utilise la trigonométrie pour estimer le rayon de la Lune et sa distance à la Terre en utilisant l'ombre de la Terre pendant une éclipse Lunaire.

-300

Le mathématicien et géomètre **Euclide** publie ses *Éléments*, où il réorganise toute la connaissance de la géométrie incluant des démonstrations logiques, la construction des 5 solides platoniciens. Dans son *Optica*, il nota que la lumière va en ligne droite et décrit la loi de réflexion.

-310

Le savant **Autolycus de Pitane** définit le mouvement uniforme tel qu'un objet parcourt une quantité égale de distance en un temps égal.

-370

Le philosophe **Aristote** développe la logique avec une théorie naïve des propositions, des quantités et du raisonnement par inférence.

-388

Le philosophe et astronome **Héraclide du Pont** fait l'hypothèse de la rotation de la Terre sur elle-même afin d'expliquer le mouvement apparent des étoiles au cours de la nuit (mais toujours dans un cadre géocentrique) et suggère que chaque planète est un corps comme la Terre.

-400

L'École Stoïcienne développe les propositions composées et les connecteurs logiques: "implication", "et", "ou" et les inférences "Modus ponens" et "Modus tollens".

-430

Le philosophe **Démocrite d'Abdère** avance l'idée selon laquelle la matière se compose de particules identiques et minuscules qu'il appelle des "atomes". En réalité il s'agit plutôt d'une vulgarisation des idées de son maître, le philosophe **Leucippe de Milet** élaborées 10 ans avant.

-500

Les philosophes **Leucippe** et **Démocrite** sont les fondateurs de l'atomisme.

-540

Le philosophe et mathématicien **Pythagore** étudie la géométrie propositionnelle et la vibration de la corde de la lyre.

-600

Le philosophe **Empédocle** plaide pour une décomposition du monde en quatre éléments fondamentaux: l'eau, la terre, l'air et le feu. Le même siècle, le mathématicien **Thalès de Milet** met en évidence l'électrostatique en frottant un morceau d'ambre, prédit une éclipse et développe la géométrie du triangle.

-800

Les Assyriens utilisent des clepsydres et les Chinois tracent les mouvements planétaires pour leur calendrier.

-1500

Les indiens développent la théorie des 4 éléments (eau, air, terre feu).

-1700

Le mathématicien **Apastamba** résout les équations linéaires générales et utilise les systèmes d'équations diophantiennes comportant jusqu'à 5 inconnues. Le même siècle, les mathématiciens égyptien utilisent des fractions simples.

-1800

Les scribes Babyloniens recherchent la solution d'une équation quadratique.

-2000

Les prêtres babyloniens font les premiers relevés des observations célestes.

-2300

Les astronomes chinois font les premières observations du ciel.

-2500

Les Mésopotamiens imaginent un système de numérotation de position composé de symboles dont la valeur est fonction de leur rang à l'intérieur d'un nombre.

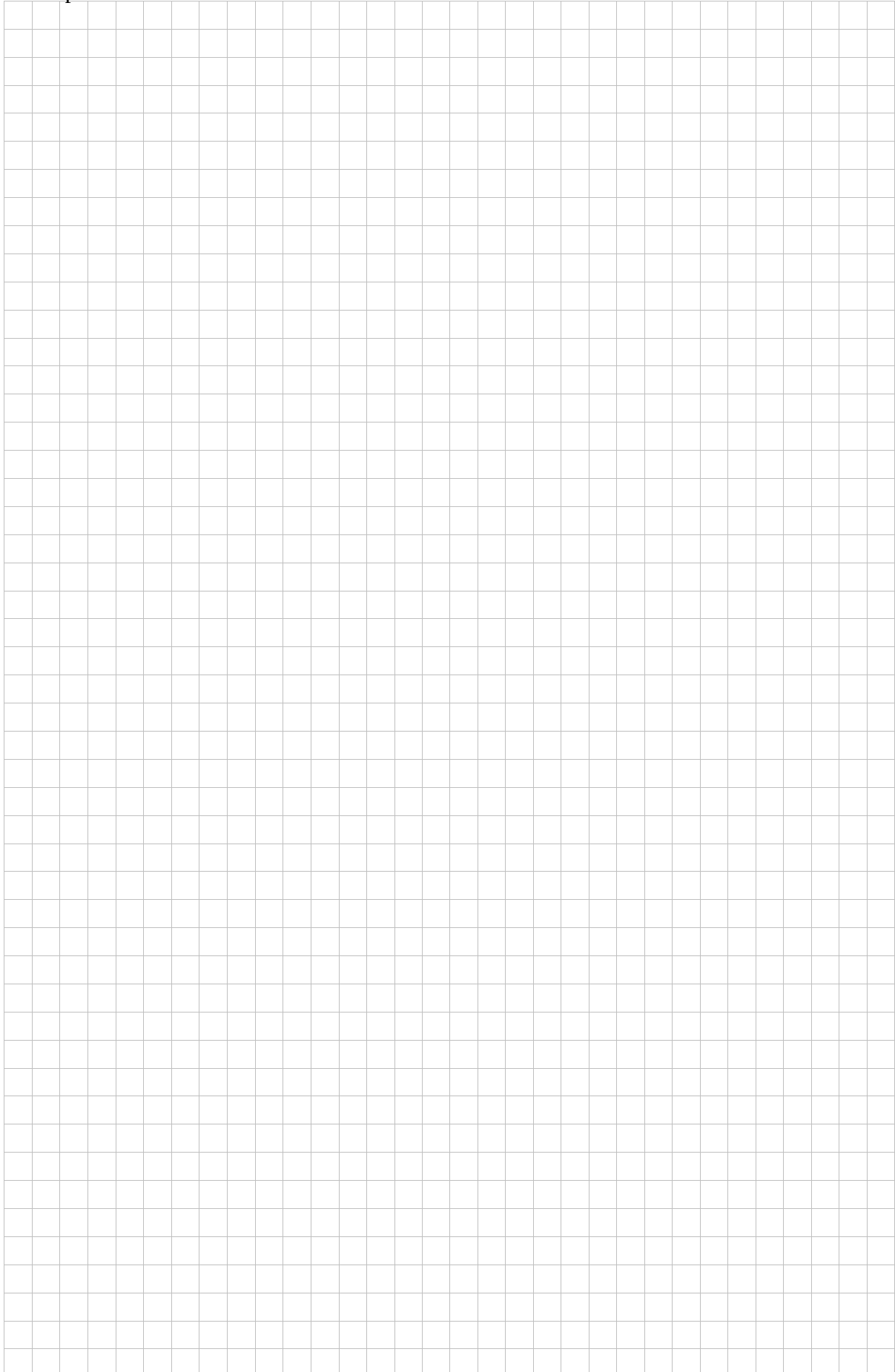
-3000

Les Chinois et les Babyloniens inventent l'abaque (boulter) première forme de machine à additionner. Des notions de géométrie pour l'arpentage sont développées (calcul de l'hypoténuse).

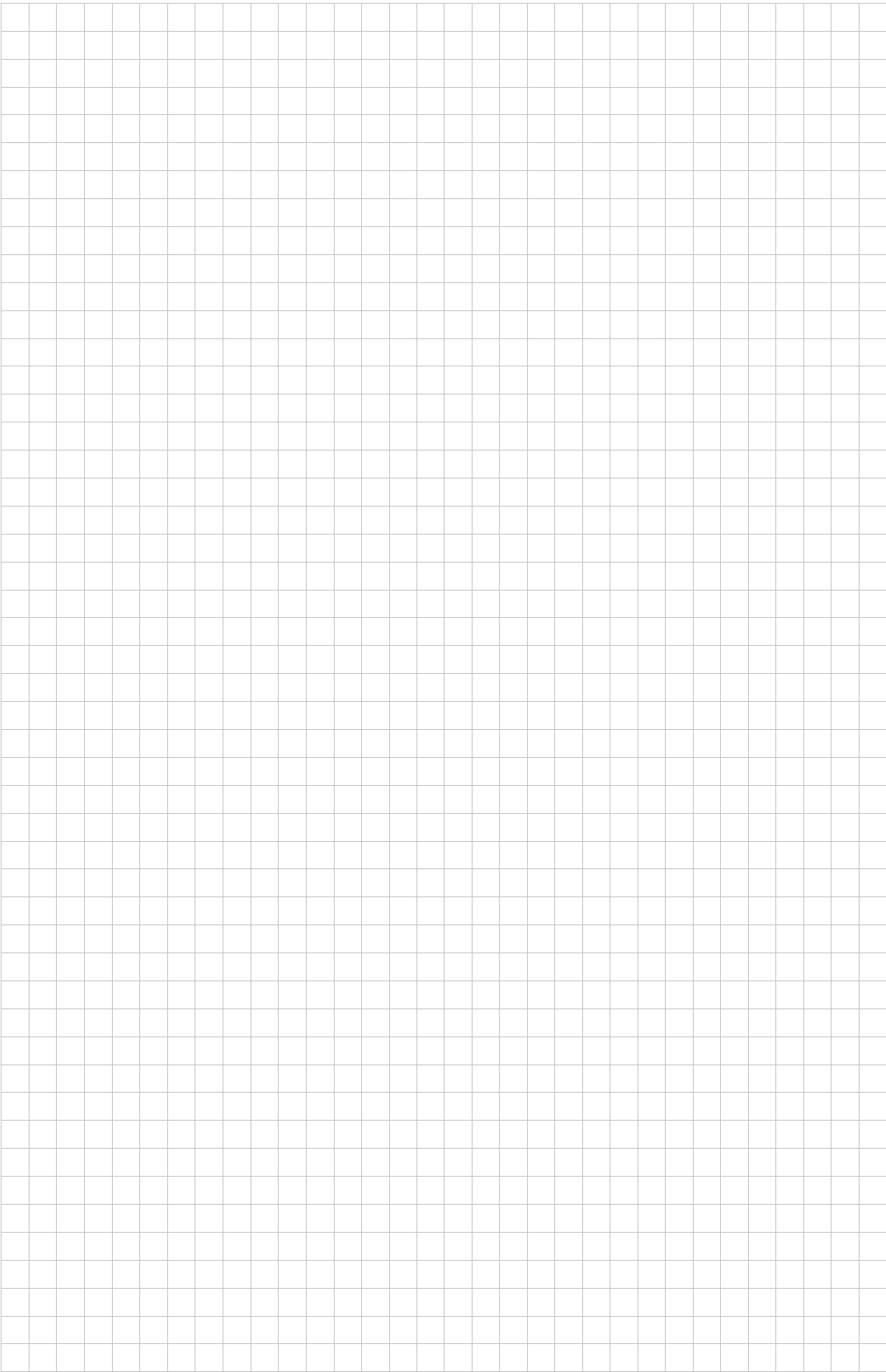
-3200

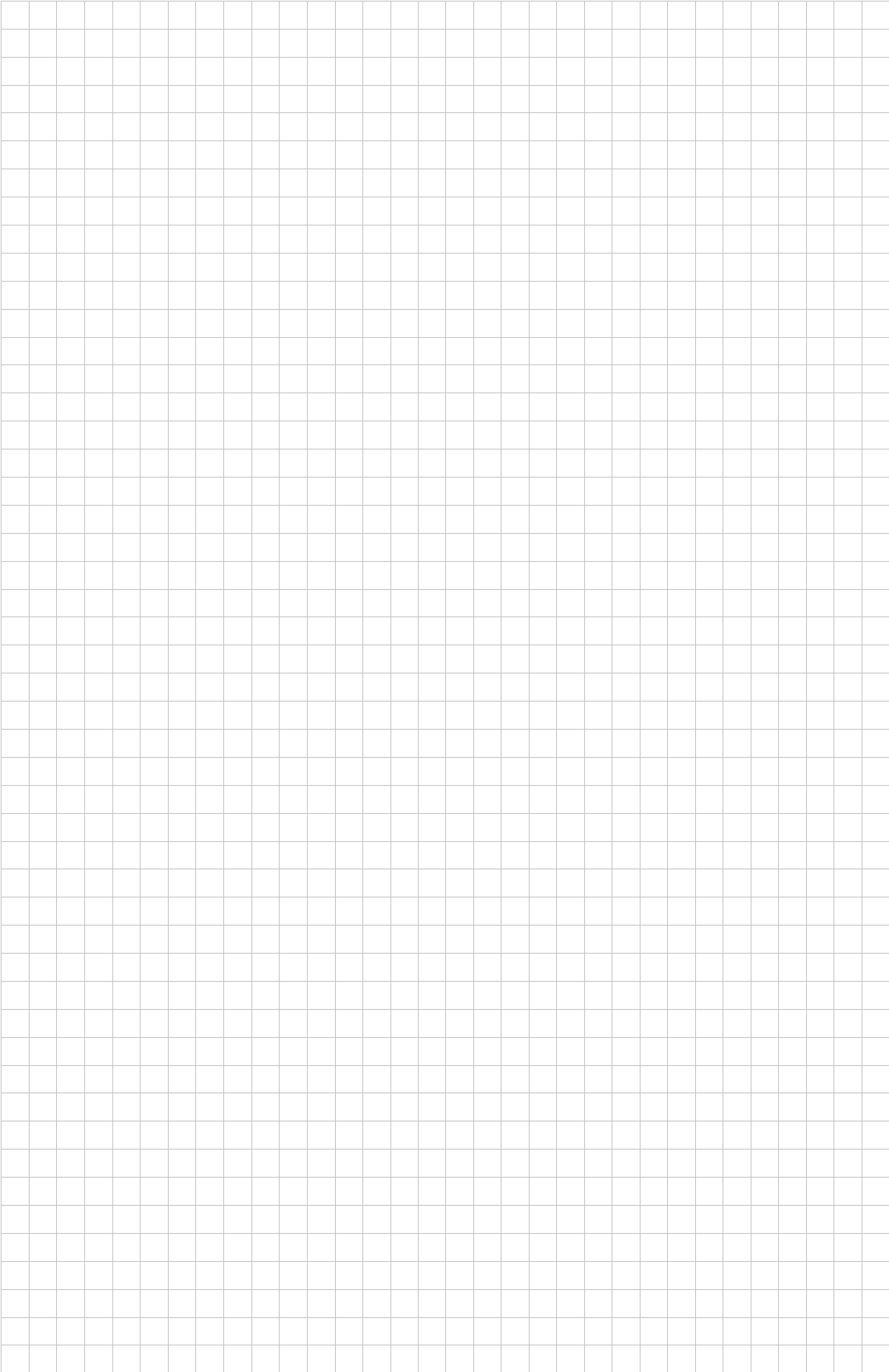
Datation des premières roues trouvées.

Notes personnelles:









Pour se détendre après avoir lu quelques centaines de pages de développements mathématiques, rien ne vaut parfois un peu d'humour scientifique...



Cette page est transmise avec des électrons 100% recyclés

HUMOUR

1. SITUATIONS

Lors d'un grand jeu télévisé, les trois concurrents se trouvent être un ingénieur, un physicien et un mathématicien. Ils ont une épreuve à réaliser. Cette épreuve consiste à construire une clôture tout autour d'un troupeau de moutons en utilisant aussi peu de matériel que possible.

- L'ingénieur: Regroupe le troupeau dans un cercle, puis décide de construire une barrière tout autour.
- Le physicien: Construit une clôture d'un diamètre infini et tente de relier les bouts de la clôture entre eux jusqu'au moment où tout le troupeau peut tenir dans le cercle.
- Le mathématicien: Voyant ceci, construit une clôture autour de lui-même et se définit comme y étant à l'extérieur.

Deux hommes se déplaçant en ballon sont perdus dans le désert. Ils aperçoivent un individu en train de méditer à l'ombre d'un arbre.

- "Où sommes-nous, s'il vous plaît ?" lui demandent-ils.

Après un long moment de réflexion, l'homme leur répond:

- "Dans un ballon."
- "Merci, monsieur le mathématicien."

L'homme demande étonné:

- "Comment avez-vous su que j'étais mathématicien?"
- "Pour trois raisons (répond l'aéronaute). Premièrement, vous avez beaucoup réfléchi avant de nous répondre. Deuxièmement, votre réponse est très exacte. Troisièmement, elle ne sert absolument à rien."

Réactions devant un feu

- L'ingénieur: Il prend un seau d'eau et éteint le feu
- Le physicien: Après avoir mesuré la hauteur de la flamme, la pression, la température etc..., il calcule la quantité minimale d'eau pour éteindre la flamme
- Réaction du mathématicien: "D'après le corollaire I. IV. II. du théorème I. IV. I., il existe une solution

pour éteindre ce feu"

Un médecin, un légiste et un mathématicien discutent des mérites comparés d'une épouse et d'une maîtresse.

- Le légiste: "Il vaut mieux avoir une maîtresse. En cas de divorce, une épouse pose de nombreux problèmes légaux."

- Le médecin: "Il vaut mieux avoir une épouse, car le sentiment de sécurité réduit le stress, et c'est bon pour la santé."

- Le mathématicien: "Vous avez tous les deux torts. Le mieux est d'avoir les deux. Quand votre femme vous croît chez votre maîtresse, et votre maîtresse chez votre femme, vous pouvez faire des maths."

Un grand homme d'affaires engage un mathématicien, un informaticien et un physicien afin de pouvoir gagner à tous les Tiercés.

Le mathématicien le premier s'attaque à la tâche, il calcule des matrices à n'en plus finir, pose des axiomes à tout bout de champs et après de longues semaines de Lemmes, théorèmes et conjectures, il conclut que le problème est formellement irrésolvable.

Ensuite, l'informaticien heureux d'avoir vu le mathématicien en échec, s'approche de son Cray III et après avoir écrit quantités d'algorithmes en C++ et introduit tous les paramètres et conditions initiales annonce joyeusement qu'il faudra juste quelques centaines d'années pour calculer le résultat de chaque Tiercé...

Le physicien, le sourire aux lèvres, informe ses éminents collègues qu'il a la solution. Il s'approche d'un tableau noir et tout en dessinant une sphère commence par dire: "Approximons le cheval par une sphère parfaite..."

Lors d'un entretien d'embauche, un chef d'entreprise reçoit quatre ingénieurs: un ayant fait l'École polytechnique, le second HEC, le troisième informaticien, et le dernier sortant de l'université. Celui-ci explique aux quatre candidats qu'en définitive, pour faire marcher une entreprise, il suffit de savoir compter.

Il s'adresse donc au premier d'entre eux, le polytechnicien, et lui dit: "allez-y, comptez..."

- Le polytechnicien: "un... deux... un... deux..."

L'homme étonné s'adresse ensuite à l'ingénieur sortant d'HEC: "À vous! Comptez..."

- L'ingénieur sortant d'HEC: "un KiloFranc... deux KiloFrancs, trois KF..."

Il se retourne ensuite, inquiet, vers l'informaticien: "À vous! Comptez..."

- L'informaticien: "0... 1... 0... 1..."

Désespéré, le chef d'entreprise s'adresse au dernier candidat sortant de faculté: "Allez-y, comptez..."

- Le jeune homme commence: "1... 2... 3... 4... 5... 6... 7..."

Le chef d'entreprise rassuré: "continuez, continuez..."

"8... 9... 10... valet... dame.. roi... " !!

Un concours met au défi un mathématicien, un physicien et un informaticien de démontrer que tous les impairs ne sont pas premiers.

Le mathématicien en premier s'attaque à la tâche et énonce:

"3 impair et premier
5 impair et premier
7 impair et premier
9=3x3 impair et non premier

Donc, la proposition initiale est fausse. Fin de la démonstration"

Ensuite, le physicien s'énonce à son tour:

"3 impair et premier
5 impair et premier
7 impair et premier
9 est impair mais non premier
11 impair et premier
13 impair et premier

Donc 9 doit être une erreur expérimentale. La proposition est donc fausse"

L'informaticien...:

"3 impair et premier
5 impair et premier
7 impair et premier
9 impair et non premier
9 impair et non premier
9 impair et non premier
9 impair et non premier
....."

- Finalement, le chimiste:

"3 impair et premier
5 impair et premier
7 impair et premier
9 impair et premier
11 impair et premier

13 impair et premier
15 impair et premier...."

Un mathématicien, un ingénieur et un physicien se voient remettre une balle en gomme rouge afin d'en déterminer le volume.

- Le mathématicien: Mesure le diamètre et évalue le résultat de la triple intégrale.
 - Le physicien: Remplis un tonneau d'eau, dépose la balle dans l'eau et mesure le déplacement total de volume.
 - L'ingénieur: Observe le modèle et le numéro de série dans sa table des "balles en gomme rouge"
-

Cette histoire est une légende urbaine fort sympathique:

J'ai reçu un coup de fil d'un collègue à propos d'un étudiant. Il estimait qu'il devait lui donner un zéro à une question de physique, alors que l'étudiant réclamait un 20. Le professeur et l'étudiant se mirent d'accord pour choisir un arbitre impartial et je fus choisi. Je lus la question de l'examen: "Montrez comment il est possible de déterminer la hauteur d'un building à l'aide d'un baromètre". L'étudiant avait répondu: "On prend le baromètre en haut du building, on lui attache une corde, on le fait glisser jusqu'au sol, ensuite on le remonte et on calcule la longueur de la corde. La longueur de la corde donne la hauteur du building."

L'étudiant avait raison vu qu'il avait répondu juste et complètement à la question. D'un autre côté, je ne pouvais pas lui mettre ses points: dans ce cas, il aurait reçu son grade de physique alors qu'il ne m'avait pas montré de connaissances en physique. J'ai proposé de donner une autre chance à l'étudiant en lui donnant six minutes pour répondre à la question avec l'avertissement que pour la réponse, il devait utiliser ses connaissances en physique. Après cinq minutes, il n'avait encore rien écrit. Je lui ai demandé s'il voulait abandonner, mais il répondit qu'il avait beaucoup de réponses pour ce problème et qu'il cherchait la meilleure d'entre elles. Je me suis excusé de l'avoir interrompu et lui ai demandé de continuer. Dans la minute qui suivit, il se hâta pour me répondre: "On place le baromètre à la hauteur du toit. On le laisse tomber en calculant son temps de chute avec un chronomètre. Ensuite en utilisant la bonne formule connue par tous, on trouve la hauteur du building". À ce moment, j'ai demandé à mon collègue s'il voulait abandonner. Il me répondit par l'affirmative et donna presque 20 à l'étudiant. En quittant son bureau, j'ai rappelé l'étudiant, car il avait dit qu'il avait plusieurs solutions à ce problème. "Hé bien, dit-il, il y a plusieurs façons de calculer la hauteur d'un building avec un baromètre. Par exemple, on le place dehors lorsqu'il y a du soleil. On calcule la hauteur du baromètre, la longueur de son ombre et la longueur de l'ombre du building. Ensuite, avec un simple calcul de proportion, on trouve la hauteur du building."

Bien, lui répondis-je, et les autres? À quoi l'élève répondit: "Il y a une méthode assez basique que vous allez apprécier. On monte les étages avec un baromètre et en même temps on marque la longueur du baromètre sur le mur. En comptant le nombre de traits, on a la hauteur du building en longueur de baromètre. C'est une méthode très directe. Bien sûr, si vous voulez une méthode plus sophistiquée, vous pouvez pendre le baromètre à une corde, le faire balancer comme un pendule et déterminer la valeur de g au niveau de la rue et au niveau du toit. À partir de la différence de g la hauteur de building peut être calculée. De la même façon, on l'attache à une grande corde et en étant sur le toit, on le laisse

descendre jusqu'à peu près le niveau de la rue. On le fait balancer comme un pendule et on calcule la hauteur du building à partir de sa période de balancement."

Finalement, l'élève conclut: "Il y a encore d'autres façons de résoudre ce problème. Probablement la meilleure est d'aller au sous-sol, frapper à la porte du concierge et lui dire: "J'ai pour vous un superbe baromètre si vous me dites quelle est la hauteur du building."

J'ai ensuite demandé à l'étudiant s'il connaissait la réponse que j'attendais. Il a admis que oui, mais qu'il en avait marre du collège et des professeurs qui essayaient de lui apprendre comment il devait penser.

Pour l'anecdote, l'étudiant était Niels Bohr (prix Nobel de Physique en 1923) et l'arbitre Rutherford (prix Nobel de Chimie en 1908).

Dans un cours de Russell portant sur le fait que d'une proposition fausse, toute proposition peut être déduite, un étudiant lui posa la question suivante:

- "Prétendez-vous que de $2 + 2 = 5$, il s'ensuit que vous êtes le pape ? "
- "Oui", fit Russell
- "Et pourriez-vous le prouver !", demanda l'étudiant sceptique
- "Certainement", répliqua Russell, qui proposa sur le champ la démonstration suivante.

(1) Supposons que $2 + 2 = 5$

(2) Soustrayons 2 de chaque membre de l'égalité, nous obtenons $2 = 3$

(3) Par symétrie, $3 = 2$

(4) Soustrayant 1 de chaque côté, il vient $2 = 1$

Maintenant le pape et moi sommes deux. Puisque $2 = 1$, le pape et moi sommes un. Par suite, je suis le pape.

Sur ce ...

What is "pi"?

- Mathematician: "Pi is the ratio of the circumference of a circle to its diameter."
- Engineer: "Pi is about $22/7$."
- Physicist: "Pi is 3.14159 plus or minus 0.000005"
- Computer Programmer: "Pi is 3.141592653589 in double precision."
- Nutritionist: "You're one track math-minded fellows, Pie is a healthy and delicious dessert!"

An astronomer, a physicist, a mathematician and a computer scientist (it is said) were holidaying in Scotland. Glancing from a train window, they observed a black sheep in the middle of a field.

- "How interesting", observed the astronomer, "all scottish sheep are black!"
- To which the physicist responded: "No, no! Some Scottish sheep are black!"
- The mathematician gazed heavenward in supplication, and then intoned: "In Scotland there exists at least one field, containing at least one sheep, at least one side of which is black."
- The computer scientist: "Oh, no! A special case!"

A mathematician, a biologist and a physicist are sitting in a street cafe watching people going in and coming out of the house on the other side of the street.

First they see two people going into the house. Time passes. After a while they notice three persons coming out of the house.

- The physicist: "The measurement wasn't accurate."
- The biologists: "They have reproduced".
- The mathematician: "If now exactly one person enters the house then it will be empty again."

A mathematician and an engineer attend a lecture by a physicist. The topic concerns Kaluza-Klein theories involving physical processes that occur in spaces with dimensions of 9, 12 and even higher. The mathematician is sitting, clearly enjoying the lecture, while the engineer is frowning and looking generally confused and puzzled. By the end the engineer has a terrible headache. At the end, the mathematician comments about the wonderful lecture.

- The engineer says: "How do you understand this stuff?"
- Mathematician: "I just visualize the process."
- Engineer: "How can you visualize something that occurs in 9-dimensional space?"
- Mathematician: "Easy, first visualize it in N-dimensional space, then let N go to 9."

Two mathematicians are in a bar. The first one says to the second that the average person knows very little about basic mathematics. The second one disagrees, and claims that most people can cope with a reasonable amount of math.

The first mathematician goes off to the washroom, and in his absence the second calls over the waitress. He tells her that in a few minutes, after his friend has returned, he will call her over and ask her a question. All she has to do is answer "one third x cubed."

She repeats "one thir -- dex cue"?

He repeats "one third x cubed".

She asks, "one thir dex cuebd?"

"Yes, that's right," he says.

So she agrees, and goes off mumbling to herself, "one thir dex cuebd..."

The first guy returns and the second proposes a bet to prove his point, that most people do know something about basic math. He says he will ask the blonde waitress an integral, and the first laughingly agrees. The second man calls over the waitress and asks "what is the integral of x squared?".

The waitress says "one third x cubed" and while walking away, turns back and says over her shoulder "plus a constant!"

Paroles de profs:

- Il reste des lambeaux de l'argument
- Le signe '-' devant le potentiel vous interpelle ? Et si je mets un '+' est-ce que cela vous convient mieux? Oui ! Alors mettons un '+'...
- C'est une courbe de phase ça ? Non, c'est une jambe de chien !
- On va le tuer et on fera croire à un suicide
- Vous pouvez dire que c'est prévisible puisque c'est imprévisible !
- Je compte sur vous pour comprendre un petit peu à fond la suite
- Nous allons étudier cela maintenant un peu plus tard
- Je vais vous présenter des résultats dépendant des équations de Maxwell que vous n'avez pas encore vues... de toute façon, au point où on en est...
- Si vous ne comprenez pas, c'est normal... le contraire serait d'ailleurs étonnant
- La relativité générale ne sert à rien ! C'est pas avec ça qu'on met des satellites en orbite!
- Question à cinquante centimes
- De temps en temps, il faut être absurde
- Quand on n'a rien à se mettre sous la dent, on prend le théorème de Gauss
- On ne voit pas par quelle vertu du Saint-Esprit il deviendrait neutre ?!
- Vous comprendrez quand vous serez grand...
- Prenons l'exemple d'une banque: monsieur y fait un dépôt et madame y fait un retrait... enfin comme d'habitude quoi !
- ... plus le parallélisme sera parallèle...
- Effacez-moi ces gris-gris résidus de calculs antérieurs
- Effacez le membre de gauche. J'ai dit DE GAUCHE ! Où est votre droite ? C'est bien... eh bien la gauche, c'est de l'autre côté !
- Voyons, qui sera la prochaine victime ?
- Le '+' se reconnaît, et les électrons ne s'y trompent pas, à sa belle couleur rouge !
- Le but du jeu est de vérifier que le critère ne dit pas de conneries
- Si vous n'arrivez pas à faire cela, je vous rassure: c'est foutu pour l'examen
- Si vous n'arrivez pas à faire cela, faites-le !
- Vous être très forts pour trouver des choses fausses
- Ce genre de démonstration, ça me plaît. Vous, ça vous fait faire des cauchemars
- Je ne résoudrai pas ceci, car cela risquerait de heurter votre sensibilité !

- Tout ce que les profs adorent mérite votre méfiance !
- Vous ajoutez des patates et des cochons, c'est sans dimension...
- On peut chercher la dérivée de l'impulsion de Dirac, c'est pas ça qui va nous sauter à la tête
- C'est bête mais Riemann, c'est comme ça
- Regardez l'équation que je viens d'effacer
- Juridiquement parlant, le courant est dans ce sens
- Il est hors de question que je perde mon temps à vous résoudre cette plaisanterie de basse volée !
- Si la fille a compris alors vous devez avoir tous compris
- Ah oui, vous vous moquez de moi. En fait, vous avez raison de vous moquer des profs car nous, on n'hésite pas à se moquer de vous
- Attention, très très important: je vous propose d'en rêver la nuit
- J'ai l'impression du jouer du stradivarius devant des vaches
- La main de dieu répartit le champ perpendiculairement à la surface
- Le miracle de la disparition de l'harmonique n'aura pas lieu aujourd'hui
- Si vous vous réveillez la nuit, dites-vous que échantillonner dans le domaine des temps, c'est périodiser dans le domaine des fréquences... puis rendormez-vous
- Les courbes gauches ne sont pas droites
- Vous voyez, quelques fois mes résultats sont corrects
- On note "Q" la sortie, c'est évident...
- Il va falloir une entourloupette pour réussir à récupérer cette variable
- Je suis un 68000; M. le directeur passe dans le couloir, frappe à la porte: on ne fait pas mieux comme interruption !
- Attention: un, deux, trois, à vos cerveaux !
- Dans la steppe de l'automatique, nous arrivons à la partie aride des mathématiques
- Quel intérêt ? Eh bien aucun. C'est une figure de rhétorique pédagogique
- Si vous avez quelques souvenirs sur les graphes de fluence que jadis nous traçâmes...
- Les électrons dans du métal très chaud, c'est la place de la Concorde à 5h du soir, donc ça ne conduit pas non plus
- Les cercles, c'est ce qui a le moins de coins !
- Si vous mettez vos doigts dans une prise, ce n'est pas un nombre complexe qui sort
- Voir la démonstration dans vos lectures journalières... donc je m'en dispense
- Et puis vient ensuite le père Carnot: il ramène sa fraise pour ne pas dire grand chose
- Quelquefois, les gens qui ont un peu de culture connaissent ça... c'est raté pour aujourd'hui !
- On n'appellera pas cela "rendement", car on peut obtenir des résultats supérieurs à 1, ce qui pourrait troubler quelques esprits faibles
- ... si vous avez par exemple de la glace à 20°C...
- Moi, si on m'avait envoyé au tableau, j'aurais pas écrit cela
- Vous trouverez la réponse sur 36 15 Archimède
- J'utilise une méthode qui date de quelques siècles... comme moi d'ailleurs !
- Posez-moi des questions... j'aimerais qu'on me pose des questions! ... d'autres questions ? Je vais être obligé de prendre la liste et de dire: "toi, tu as une question à poser"
- L'ordonnancement implémenté dépend de l'instanciation du garbage collector s'il est synchronisé sur le multithreading préemptif de l'OS.
- vous n'êtes que des boîtes qui reçoivent des entrées et crachent des sorties

Sherlock Holmes et le Dr Watson sont au camping. Après un bon repas et une bonne bouteille de vin, ils gagnent leur sac de couchage dans la tente et s'endorment.

Quelques heures plus tard, Holmes se réveille et aussitôt secoue son compagnon:

- "Watson, regardez le ciel et dites-moi ce que vous voyez."

- "Je vois des millions et des millions d'étoiles."

- "Et qu'est-ce que vous en concluez ?"

- "Du point de vue astronomique, répond Watson, cela me dit qu'il y a des millions de galaxies et potentiellement des milliards de planètes. Du point de vue astrologique, j'observe que Saturne est en Lion. Du point de vue chronologique, j'en déduis qu'il est environ 3 heures 15. Du point de vue théologique, je vois que Dieu est tout-puissant et que nous sommes petits et insignifiants. Du point de vue météorologique, je pense que nous aurons une belle journée demain. Et vous, Holmes ?"

Sherlock Holmes reste pensif une minute puis déclare:

- "Watson, vous êtes un âne. On nous a fauché la tente."

An engineer, a mathematician, and a physicist are staying for the night in a hotel. Fortunately for this joke, a small fire breaks out in each room.

The physicist awakes, sees the fire, makes some careful observations, and on the back of the hotel's wine list does some quick calculations. Grabbing the fire extinguisher, he puts out the fire with one, short, well placed burst, and then crawls back into bed and goes back to sleep.

The engineer awakes, sees the fire, makes some careful observations, and on the back of the hotel's room service list (pizza menu) does some quick calculations. Grabbing the fire extinguisher (and adding a factor of safety of 5), he puts out the fire by hosing down the entire room several times over, and then crawls into his soggy bed and goes back to sleep.

The mathematician awakes, sees the fire, makes some careful observations, and on a blackboard installed in the room, does some quick calculations. Jubilant, he exclaims "A solution exists!", and crawls into his dry bed and goes back to sleep.

How do you know that the driver driving toward you is a physicist?

He has a red sticker on his bumper, saying: "If this sticker is blue, you are driving too fast."

A Princeton plasma physicist is at the beach when he discovers a ancient looking oil lantern sticking out of the sand. He rubs the sand off with a towel and a genie pops out. The genie offers to grant him one wish. The physicist retrieves a map of the world from his car and circles the Middle East and tells the genie, "I wish you to bring peace in this region".

After 10 long minutes of deliberation, the genie replies, "Gee, there are lots of problems there with Lebanon, Iraq, Israel, and all those other places. This is awfully embarrassing. I've never had to do this before, but I'm just going to have to ask you for another wish. This one is just too much for me".

Taken aback, the physicist thinks a bit and asks, "I wish that the Princeton tokamak would achieve scientific fusion energy break-even."

After another deliberation the genie asks, "Could I see that map again?"

2. MATHÉMATIQUES

Un professeur de maths explique à une blonde les limites. Il fait avec elle l'exercice suivant:

$$\lim_{x \rightarrow 8} \frac{1}{x-8} = \infty$$

À la fin de l'exercice, il demande à la blonde si elle a compris. "Oh oui monsieur j'ai tout compris!". N'y croyant qu'à moitié, il lui pose l'exercice suivant:

Soit à calculer:

$$\lim_{x \rightarrow 5} \frac{1}{x-5}$$

Et la blonde de résoudre:

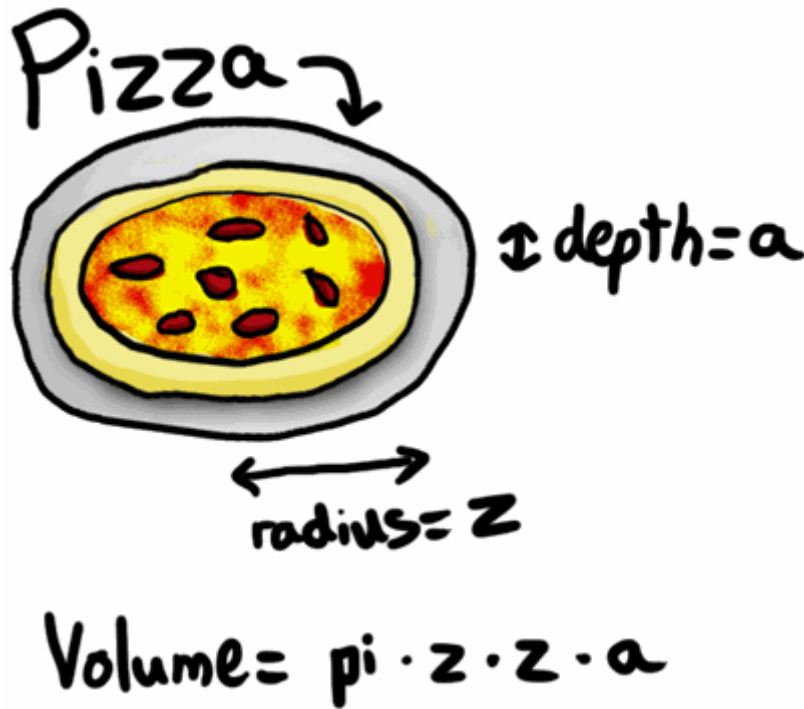
$$\lim_{x \rightarrow 8} \frac{1}{x-8} = \infty \text{ alors } \lim_{x \rightarrow 5} \frac{1}{x-5} = \infty$$

Évolution de l'enseignement des Mathématiques (...)

- Enseignement 1960: Un paysan vend un sac de pommes de terre pour 100 Frs. Ses frais de production s'élèvent au 4/5 du prix de vente. Quel est son bénéfice ?
- Enseignement traditionnel 1970: Un paysan vend un sac de pommes de terre pour 100 Frs. Ses frais de production s'élèvent au 4/5 du prix de vente, c'est-à-dire à 80 Frs. Quel est son bénéfice ?
- Enseignement moderne 1970: Un paysan échange un ensemble P de pommes de terre contre un ensemble M de monnaie. Le cardinal de l'ensemble M est égal à 100, et chaque élément PFM vaut 1 Fr. Dessinez 100 gros points représentant les éléments de l'ensemble M . L'ensemble F des frais de production comprend 20 gros points de moins que l'ensemble M . Représentez l'ensemble F comme sous-ensemble de l'ensemble M et donnez la réponse à la question suivante: Quel est le cardinal de l'ensemble B des bénéfices ? (à dessiner en rouge)
- Enseignement rénové 1980: Un agriculteur vend un sac de pommes de terre pour 100 Frs. Les frais de production s'élèvent à 80 Frs et le bénéfice est de 20 Frs. Devoir: souligne les mots "Pommes de terre" et discutes-en avec ton voisin.
- Enseignement réformé 1980: Un peizan kapitalist privilégié sanrichi injustement de 20 Frs sur un sac de patat, analiz le tekst et recherche les fotes de contenu, de gramère, >d'orthografe, de ponctuassion et ensuite dit ce que tu panse de set manière de s'enrichir.
- Enseignement start-up 1999: Un producteur de l'espace agricole câblé consulte une data bank qui display le day-rate de la patate. Il load son progiciel de computation fiable et détermine le cash-flow sur

écran bit-map (sous WMil avec config floppy et DD 40Go). Dessine avec ta souris le contour intégré 3D du sac de pommes de terre. Puis logues-toi au network par le www.blue-potatoe.com et suis les indications du menu.

- Enseignement 2010: Qu'est-ce qu'un paysan ?



Un mathématicien à son ami:

- "Es-tu fidèle?"
- "Oui, à isomorphisme près"

Comment les mathématiciens le font:

- Les théoriciens des nombres l'ont fait en premier
- Nous savons que les analystes réels le font continûment, mais pour les spécialistes de théorie des ensembles, ce n'est qu'une hypothèse
- Les analystes complexes le font entièrement mais avec conformisme
- Les algébristes le font avec détermination et sans discrimination
- Les topologistes le font ouvertement, mais compactement
- Les topologistes différentiels et algébriques le font avec variété

- Les spécialistes de combinatoire le font discrètement
- Les statisticiens le font soit presque toujours, soit presque jamais
- Les théoriciens de la mesure le font presque partout
- Les logiciens le font avec consistance
- Les géomètres le font au foyer mais avec courbures et torsions
- Les théoriciens des groupes le font simplement et fidèlement
- Les théoriciens des anneaux le font avec intégrité
- Les théoriciens des corps le font en inversé
- Les spécialistes de programmation linéaire maximisent la performance et minimisent les efforts
- Markov avait besoin de chaînes pour le faire, et Noether d'anneaux
- Euler le faisait en cercle, tandis que Bernoulli le faisait en spirale ou en huit
- Möbius le faisait toujours du même côté
- Gauss le faisait normalement, Lebesgue, avec mesure, et Cauchy le faisait complètement, au contraire de Gödel
- Fermat a essayé de le faire dans la marge, mais il n'y avait pas assez de place
- On pense que Riemann et Goldbach l'on fait, mais on n'est encore jamais arrivé à le prouver

Qu'est-ce qu'un homme complexe dit à une femme réelle ?

Réponse: "viens danser !" (il faut lire "dans C", c'est-à-dire l'ensemble des complexes)

Les fonctions logarithme et exponentielle sont au restaurant. Quand viendra l'addition, qui payera ?

Réponse: Exponentielle, car logarithme né paie rien.

Plus tard dans la nuit, Logarithme et Exponentielle rentrent chez eux un peu bourrés. Logarithme demande: Est-ce que je prends le volant ?

Exponentielle répond: Je préfère que ce soit moi qui conduise. Au cas où on dérive...

Deux suites de Cauchy veulent sortir en boîte. Elles arrivent devant une boîte où se déroule la soirée "No Limit". Elles décident de rentrer, mais le vigile les arrête en leur disant: "Désolé, c'est complet!". (dans un espace complet, une suite de Cauchy est convergente par définition, donc elle a une limite.)

A mathematician went insane and believed that he was the differentiation operator. His friends had him placed in a mental hospital until he got better. All day he would go around frightening the other patients by staring at them and saying: "I differentiate you!"

One day he met a new patient; and true to form he stared at him and said "I differentiate you!", but for once, his victim's expression didn't change. Surprised, the mathematician marshaled his energies, stared fiercely at the new patient and said loudly "I differentiate you!", but still the other man had no reaction. Finally, in frustration, the mathematician screamed out "I DIFFERENTIATE YOU!"

The new patient calmly looked up and said: "You can differentiate me all you like: I'm e to the x ."

Ce que les mathématiciens disent et ce qu'il faut comprendre:

- Trivial: Si je dois vous montrer ceci, vous êtes dans la mauvaise classe
- On peut trivialement montrer: On a pas besoin de plus de 4 heures pour le démontrer
- Contrôlez vous-même: C'est la partie difficile de la démonstration donc vous pouvez le faire sur votre temps libre
- Similairement: Au moins une ligne de la démonstration est identique à la précédente
- Procédons formellement: Nous allons manipuler des symboles avec des règles bien prédéfinies sans rien comprendre au sens réel du résultat.
- Nous nous dispenserons de la démonstration: Faites-moi confiance, c'est vrai!
- Le lecteur montrera facilement: Ça m'ennuie de montrer que...
- Nous conseillons vivement au lecteur de faire les exercices indiqués: Comme je les ai pas faits, vous pourriez me les corriger
- J'ai montré ce résultat dans un papier antérieur: Je ne sais plus diable comment on fait pour prouver ce truc-là
- On généralise facilement: La généralisation dépasse mon niveau
- D'après une propriété bien connue: Par 10 personnes au monde...

Tout le monde connaît le "Théorème du Salaire" qui établit que les ingénieurs et les scientifiques ne peuvent JAMAIS gagner autant que les hommes d'affaires et les commerciaux. Ce théorème peut enfin se démontrer par la résolution d'une équation mathématique simple.

Notre équation s'appuie sur deux postulats très connus:

P1. La Connaissance c'est la Puissance

P2. Le Temps c'est de l'Argent

Tout ingénieur sait ensuite que:

$$Puissance = \frac{Travail}{Temps}$$

Puisque:

$$Connaissance = Puissance$$

et que:

$$Temps = Argent$$

Nous avons donc:

$$Connaissance = \frac{Travail}{Argent}$$

Nous obtenons alors facilement:

$$Argent = \frac{Travail}{Connaissance}$$

Ainsi quand la Connaissance tend vers zéro, l'Argent tend vers l'infini quelle que soit la valeur attribuée à Travail, cette valeur peut être très faible. À l'inverse quand la Connaissance tend vers l'infini, l'Argent tend alors vers zéro, même si la valeur Travail est élevée.

D'où la conclusion évidente suivante: Moins vous en connaissez, plus vous gagnez d'argent.

PS: Ceux d'entre vous qui ont eu quelques difficultés de compréhension doivent être les mieux rémunérés...

C.Q.F.D.

Dans le même genre voici le "Théorème du misogyne":

D'abord, nous déclarons que les filles sont des variables factorisables en quantités de temps et d'argent telle que:

$$Filles = Temps \cdot Argent$$

Comme nous le savons tous "le temps c'est de l'argent!". Donc:

$$Temps = Argent$$

et parce que "l'argent est la racine du mal..." et que le mal est un synonyme du ce qui est mauvais:

$$Argent = \sqrt{Mauvais}$$

Donc:

$$Filles = (\sqrt{Mauvais})^2$$

Nous sommes donc forcés de conclure:

$$Filles = Mauvais$$

C.Q.F.D.

The Top Ten Excuses For Not Doing Your Math Homework

#10. Galileo didn't know calculus; what do I need it for?

#9. "A math addict stole my homework. When they arrested him, they discovered Mr. Pleacher had been his teacher."

#8. I'm taking physics and the homework in there seemed to involve math, so I thought I could just do that instead.

#7. I have the proof, but there isn't room to write it in the margin.

#6. I have a solar powered calculator and it was cloudy.

#5. I was watching the World Series and got tied up trying to prove that it converged.

#4. I could only get arbitrarily close to my textbook.

I couldn't actually reach it.

(I reached half way, and then half of that, and then ...)

#3. I couldn't figure out whether
i am the square root of negative one or
i is the square root of negative one.

#2. It was Einstein's birthday and pi day
and we had this big celebration!
(This only works for March 14)

#1. I accidentally divided by zero and my paper burst into flames.

Que fait:

$$\frac{2ab}{2Fr.16}$$

(lire "2 abbés sur 2 françaises)

Solution:

$$2bb\frac{a}{e}$$

(lire "2 bébés assurés")

Que fait:

$$\frac{cheval}{oiseau}$$

Comme:

$$\frac{cheval}{oiseau} = \frac{vache \cdot l}{\beta \cdot l}$$

(lire "vache + l" (contient toutes les lettres de "cheval") divisé par "bête à ailes")

Mais:

$$\frac{vache \cdot l}{\beta \cdot l} = \frac{\beta \cdot \pi \cdot l}{\beta \cdot l}$$

(lire "bête à pie + l" divisé par "bête à ailes")

On simplifie:

$$\frac{\cancel{\beta} \cdot \pi \cdot \cancel{l}}{\cancel{\beta} \cdot \cancel{l}} = \pi$$

Ceci démontre que Pi est incommensurable (soit irrationnel) car il n'y a aucune commune mesure entre un cheval et un oiseau...

À démontrer:

$$\frac{ROSSINI}{SOLSIDO} = 1$$

$$\frac{ROS \ SI \ NI}{SOL \ SI \ DO} = \frac{ROS \ NI}{SOL \ DO}$$

mais $NI = DO$ (NI vaut Do c'est-à-dire "niveau d'eau") donc:

$$\frac{ROS}{SOL}$$

Mais SOL fait RINO (Solferino: c'est lors de cette bataille qu'Henri Dunant en voyant l'horreur a décidé de créer la Croix-Rouge) donc:

$$\frac{ROS}{RI \ NO} = 1$$

et comme RINO c'est ROS (rhinocéros) donc RINO = ROS alors:

$$\frac{ROS}{ROS} = 1$$

C.Q.F.D.

Chaque futur ingénieur apprend à inscrire la somme de deux chiffres rationnels, par exemple:

$$1 + 1 = 2$$

Cette forme est cependant assez banale et indique des lacunes dans votre éducation.

En premier semestre, on apprend que:

$$1 = \ln(e)$$

et ensuite que:

$$1 = \sin^2(p) + \cos^2(p)$$

Tout le monde sait aussi que:

$$2 = \sum_{n=0}^{\infty} \left(\frac{1}{2} \right)^n$$

et que donc l'équation:

$$1 + 1 = 2$$

peut être écrite plus simplement:

$$\ln(e) + \sin^2(p) + \cos^2(p) = \sum_{n=0}^{\infty} \left(\frac{1}{2} \right)^n$$

(il faut admettre que l'aspect est bien plus clair et plus scientifique)

D'autre part, il est évident que:

$$1 = \cosh(q) \sqrt{1 - \tanh^2(q)}$$

et aussi:

$$e = \lim_{z \rightarrow \infty} \left(1 + \frac{1}{z} \right)^z$$

il en résulte que:

$$\ln(\varnothing) + \sin^2(p) + \cos^2(p) = \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^n$$

peut être réécrite comme suit:

$$\ln \left(\lim_{z \rightarrow \infty} \left(1 + \frac{1}{z} \right)^z \right) + \sin^2(p) + \cos^2(p) = \sum_{n=0}^{\infty} \left(\frac{\cosh(q) \sqrt{1 - \tanh^2(q)}}{2} \right)^n$$

Il faut également se rappeler que:

$$0! = 1$$

et que l'exposant inverse de l'exposant opposé est égal à l'exposant opposé de l'exposant inverse. En supposant un espace à n dimension, on sait que:

$$(X^T)^{-1} - (X^{-1})^{-T} = 0$$

Si l'on prend encore la matrice comme la métrique d'un espace canonique orthogonal et orienté:

$$(\mathcal{G}_{ij}^T)^{-1} - (\mathcal{G}_{ij}^{-1})^{-T} = 0$$

nous obtenons logiquement:

$$\left((\mathcal{G}_{ij}^T)^{-1} - (\mathcal{G}_{ij}^{-1})^{-T} \right)! = 1$$

nous obtenons ainsi une expression simple et claire pour tout le monde de $1 + 1 = 2$:

$$\ln \left(\lim_{z \rightarrow \infty} \left(\left((\mathcal{G}_{ij}^T)^{-1} - (\mathcal{G}_{ij}^{-1})^{-T} \right)! + \frac{1}{z} \right)^z \right) + \sin^2(p) + \cos^2(p) = \sum_{n=0}^{\infty} \left(\frac{\cosh(q) \sqrt{1 - \tanh^2(q)}}{2} \right)^n$$

Il est donc évident que cette équation est bien plus compréhensible que:

$$1 + 1 = 2$$

Il serait possible de montrer plusieurs autres développements de cette simple expression et nous le ferons à partir du moment où vous commencerez à comprendre les principes simples de la méthode précédente.

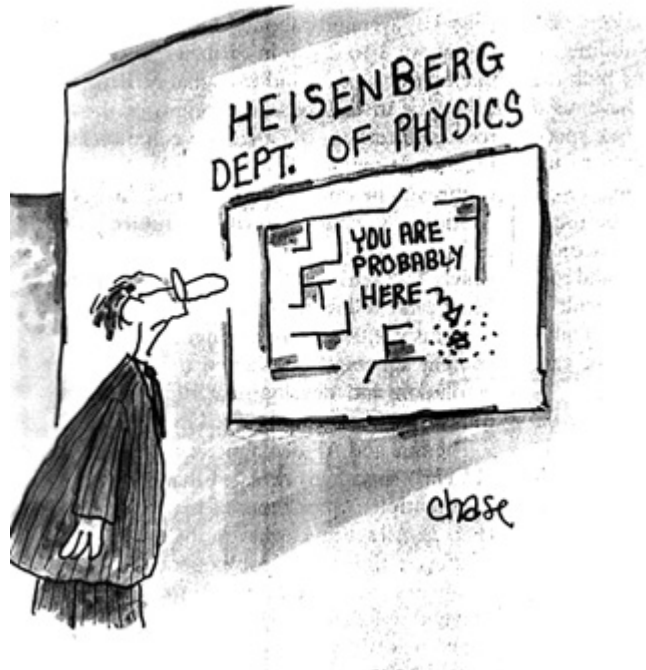
Un catcheur, un physicien et un mathématicien sont sujet à une expérience: on les enferme dans une pièce avec chacun une boîte d'épinards, fermée, et sans ouvre-boîte. Au bout de 24 heures, on va voir ce qu'ils sont devenus.

- Le catcheur a réussi à ouvrir sa boîte: « Et bien, j'ai simplement violemment projeté la boîte contre le mur. L'impact a été tel qu'elle s'est ouverte », explique t-il.
- Le physicien a également réussi à ouvrir sa boîte: « J'ai observé le solide, et distingué ses points de rupture. J'ai alors effectué une pression de manière à exercer une force maximale sur ceux-ci, et la

boîte s'est tout naturellement ouverte. »

- Le mathématicien, enfin, est retrouvé prostré dans un coin de la pièce, la sueur ruisselant sur son visage, et sa boîte de conserve, fermée, entre les pieds: « Admettons que la boîte est ouverte... Admettons que... »

3. PHYSIQUE



-En théorie, il n'y a pas de différences entre la théorie et la pratique. En pratique, si!

-La théorie, c'est quand on sait tout, mais que rien ne marche. La pratique, c'est quand tout marche, mais qu'on ne sait pas pourquoi. En informatique, la théorie et la pratique sont réunies: rien ne marche et on ne sait pas pourquoi!

Matter is fundamentally lazy:- It always takes the path of least effort

Matter is fundamentally stupid:- It tries every other path first.

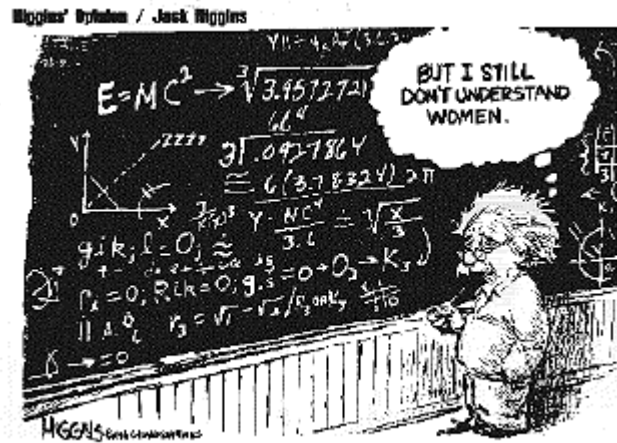
That is the heart of physics - The rest is details.

Ce sont deux atomes qui se rencontrent.

L'un dit à l'autre: "Merde, j'ai perdu un électron!"

L'autre: "T'es sûr?"

Et le premier répond: "POSITIVEMENT !!"



You enter the laboratory and see an experiment. How will you know which class is it?

- If it's green and wiggles, it's biology.
- If it stinks, it's chemistry.
- If it doesn't work, it's physics.

Theorem: a cat has nine tails.

Proof: No cat has eight tails. A cat has one tail more than no cat. Therefore, a cat has nine tails.



Un physicien étudiant la physique quantique, c'est quelqu'un qui ne voyant pas très bien, cherche dans une chambre obscure, un chat noir, qui probablement n'existe pas.

An engineer, a physicist, a mathematician, and a mystic were asked to name the greatest invention of all times.

- The engineer chose fire, which gave humanity power over matter.
- The physicist chose the wheel, which gave humanity the power over space.
- The mathematician chose the alphabet, which gave humanity power over symbols.

- The mystic chose the thermos bottle.

"Why a thermos bottle?" the others asked.

- The mystic: "Yes, because the thermos keeps hot liquids hot in winter and cold liquids cold in summer."

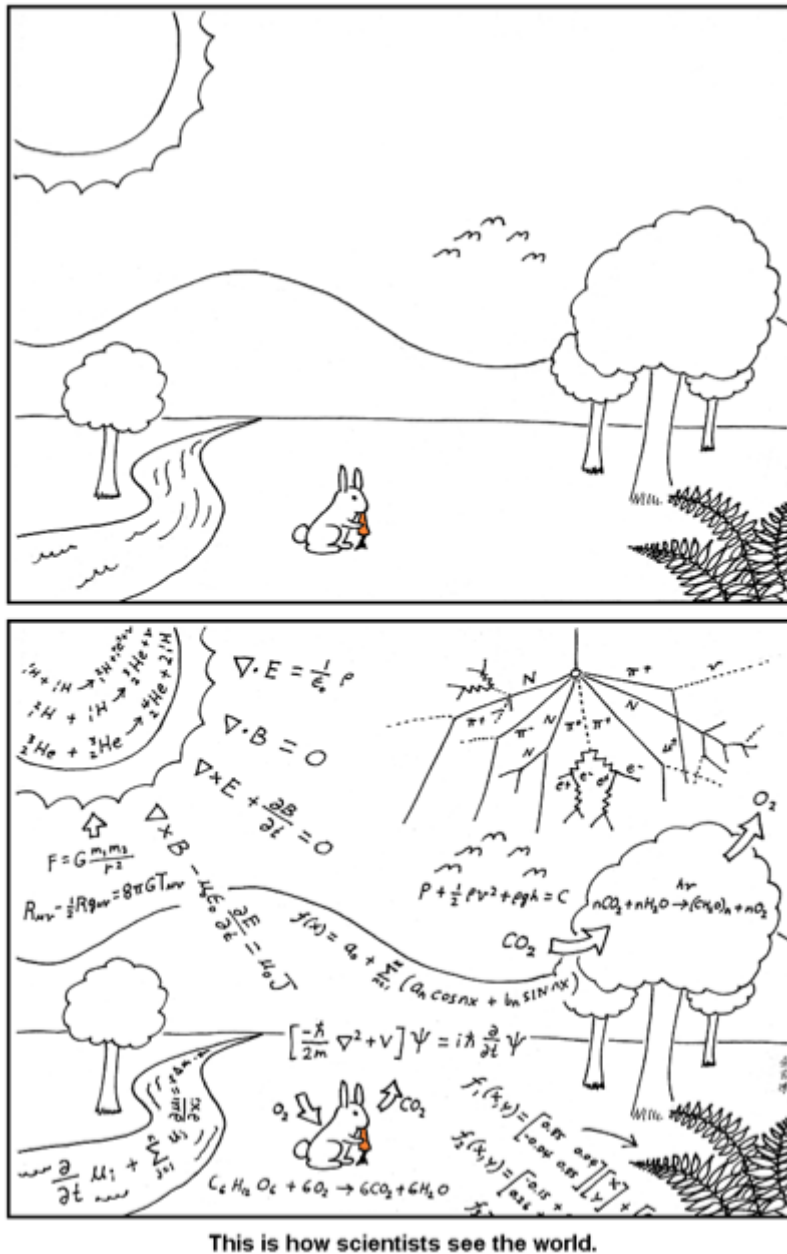
- "Yes... so what?" the others asked.

- The mystic: "Think about it." said the mystic reverently. "That little bottle.. how does it know?"

Physics professor has been doing an experiment, and has worked out an empirical equation that seems to explain his data. He asks the math professor to look at it.

A week later, the math professor says the equation is invalid. By then, the physics professor has used his equation to predict the results of further experiments, and he is getting excellent results, so he asks the math professor to look again.

Another week goes by, and they meet once more. The math professor tells the physics professor the equation does work, "But only in the trivial case where the numbers are real and positive."



This is how scientists see the world.

Heisenberg was driving down the highway whereupon he was pulled over by a policeman.

The policeman asked:

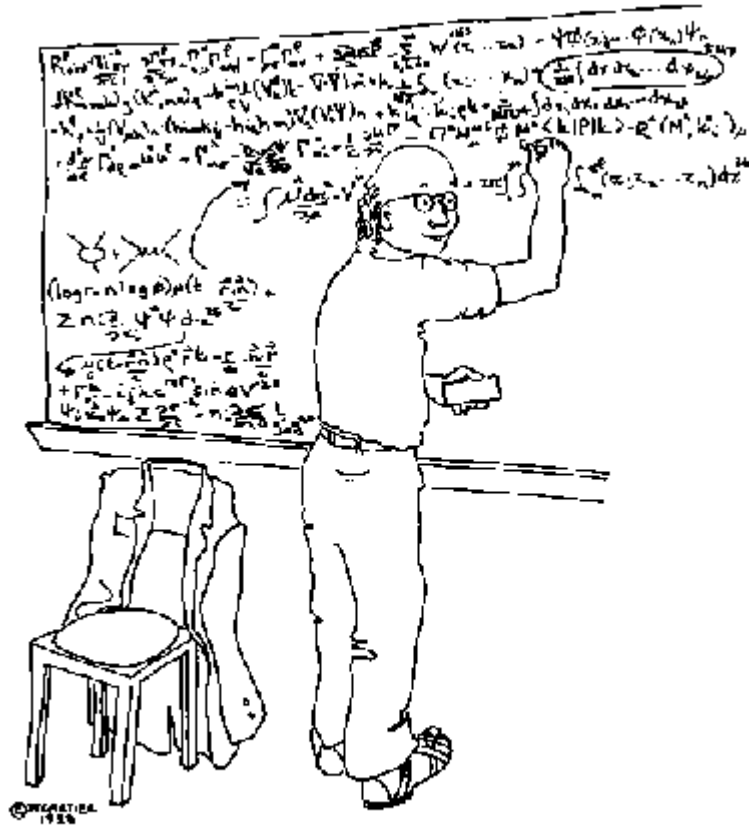
"Do you know how fast you were going back there?"

Heisenberg replied:

"No, but I know where I am."

What's the difference between an auto mechanic and a quantum mechanic?

The quantum mechanic can sometimes get the car inside the garage without opening the door.



"At this point we notice that this equation is beautifully simplified if we assume that space-time has 92 dimensions."

Why God Never Received Tenure at any University

1. He had only one major publication.
2. It was in Hebrew.
3. It had no references.
4. It wasn't published in a refereed journal.
5. Some even doubt he wrote it himself.
6. It may be true that he created the world, but what has he done since then?
7. His cooperative efforts have been quite limited.
8. The scientific community has had a hard time replicating his results.
9. He never applied to the Ethics Board for permission to use human subjects.
10. When one experiment went awry he tried to cover it up by drowning the subjects.
11. When subjects didn't behave as predicted, he deleted them from the sample .

12. He rarely came to class, just told students to read the Book.
 13. Some say he had his son teach the class.
 14. He expelled his first two students for learning.
 15. Although there were only ten requirements, most students failed his tests.
 16. His office hours were infrequent and usually held on a mountaintop.
-

In the beginning there was Aristotle
And objects at rest tended to remain at rest
And objects in motion tended to come to rest
And God saw that it was boring, although very restful.

Then God created Newton
And objects at rest tended to remain at rest
And objects in motion tended to remain in motion
And energy was conserved, and momentum was conserved,
And matter was conserved
And God saw that it was conservative.

Then God created Einstein
And everything was relative
And fast things became short
And straight things became curved
And the universe was filled with inertial frames
And God saw that it was relatively general
but some of it was especially relative.

Then God created Bohr
And there was the principle
And the principle was quantum
And all things were quantified
But some things were still relative
And God saw that it was confusing.

Then God was going to create Furgeson
And Furgeson would have unified
And he would have fielded a theory
And all would have been one.
But it was the seventh day
And God rested
And objects at rest tend to remain at rest.



The Physicist's Bill of Rights

We hold these postulates to be intuitively obvious, that all physicists are born equal, to a first approximation, and are endowed by their creator with certain discrete privileges, among them a mean rest life, n degrees of freedom, and the following rights which are invariant under all linear transformations:

1. To approximate all problems to ideal cases.
2. To use order of magnitude calculations whenever deemed necessary (i.e. whenever one can get away with it).
3. To use the rigorous method of "squinting" for solving problems more complex than the addition of positive real integers.
4. To dismiss all functions which diverge as "nasty" and "unphysical".
5. To invoke the uncertainty principle when confronted by confused mathematicians, chemists, engineers, psychologists, dramatists, and other lower scientists.
6. When pressed by non-physicists for an explanation of (4) to mumble in a sneering tone of voice something about physically naive mathematicians.
7. To equate two sides of an equation which are dimensionally inconsistent, with a suitable comment to the effect of, "Well, we are interested in the order of magnitude anyway".
8. To the extensive use of "bastard notations" where conventional mathematics will not work.
9. To invent fictitious forces to delude the general public.
10. To justify shaky reasoning on the basis that it gives the right answer.
11. To cleverly choose convenient initial conditions, using the principle of general triviality.
12. To use plausible arguments in place of proofs, and thenceforth refer to these arguments as proofs.
13. To take on faith any principle which seems right but cannot be proved.

You can at least take this simple Real Scientist Quiz to find out if you're cut out for the life of a true

scientist.

The Real Scientist Quiz

1. At Christmas time, you:

- a. Take a couple of days off to spend time with your family.
- b. Leave early on Christmas eve so you can pick up a few presents for the family.
- c. Only work half a day, spending the rest of the day at home working on your grant application.

2. Your spouse wants to discuss plans for the family vacation with your kids. You:

- a. Propose to go camping so you can explain the wonders of nature to your kids
- b. Propose to go to another city so you can spend the day in your friend's lab while your spouse takes the kids sightseeing.
- c. Ask your spouse, "We have kids?"

3. At a scientific meeting on an island in the South Pacific, no talks are scheduled in the afternoon. During this free time, you:

- a. Follow the local custom and sunbathe on the beach in the nude.
- b. Sit on the beach fully clothed, unaware of the nude sunbathers, and discuss science with your colleagues.
- c. Sit in your hotel room with the drapes closed, and work on your manuscript.

4. The nurse at school calls to tell you that your second-grade child has chicken pox. You:

- a. Immediately drop what you're doing and rush to school to pick up your sick kid.
- b. Immediately drop what you're doing and begin trying to find a cure for chicken pox.
- c. Ask the nurse for directions to the school, and the names of your kids.

5. Beings from outer space visit Earth, and you are the first human they meet. To show their friendship, they present you with a highly advanced device that is capable of prolonging life, ending human suffering, and curing disease. You:

- a. Present it to the United Nations.
- b. Apply for a patent.
- c. Break it open to see how it works.

6. What is the longest amount of time that you have worked without a vacation (excluding scientific meetings)?

- a. Six months.
- b. Two years.
- c. I took a weekend off about 10 years ago.

7. What are your hobbies?

- a. Sports, music, and dance, because they allow the analytical parts of my brain to relax.
- b. Cooking, because it's quite a lot like science.
- c. Reading back issues of scientific journals cover to cover.

8. Your best friend is:

- a. A member of your college fraternity.
- b. A member of your immediate family.
- c. A member of a gene family.

Score:

Give yourself one point for every question you answered with an "a," 5 points for every "b," and 50 points for every "c." If you took the test three times and averaged your score, give yourself 100 extra points. If you calculated the standard error of the mean, give yourself 500 points.

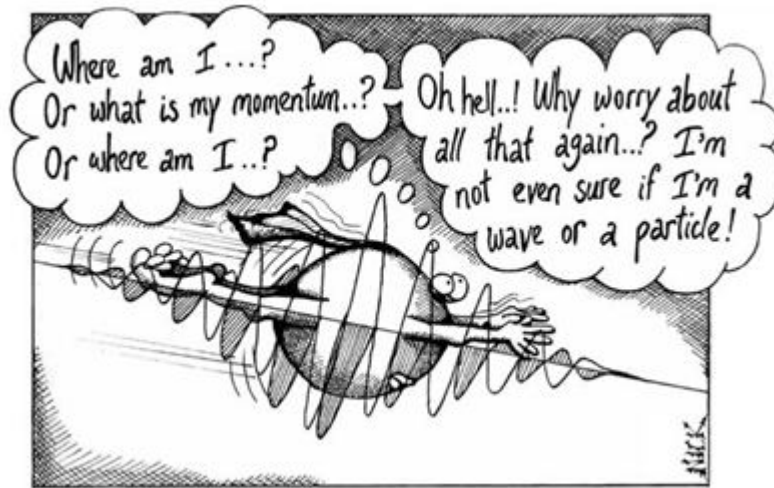
If you scored less than 10, you are normal. Scores of 11-50 indicate an obsessed scientist. If you scored more than 50, you are in need of help and should consider joining Scientists Anonymous; if you scored greater than 500, you should forget Scientists Anonymous and get back to work since you are beyond help, and may actually succeed as a scientist.



Comment il faut comprendre certaines phrases dans les publications des physiciens:

- Il est bien établi que...: Je ne me suis pas donné la peine de lire les références, mais...
- Ceci est de grande importance théorique: Ceci est important pour moi.

- Quoiqu'il n'ait pas été possible de donner une réponse définitive: L'expérience échoua, mais il me semble tout de même pouvoir en tirer une publication.
 - La technique utilisée fut particulièrement adéquate...: Le copain du labo d'à côté avait déjà mis la technique au point.
 - 3 échantillons furent choisis pur une étude exhaustive: Les résultats obtenus à partir des autres échantillons n'ont rien donné de cohérent .
 - Manipulé avec la plus grande précaution durant toute l'expérimentation: Ne fut pas jeté à l'égout.
 - La concordance avec la théorie est excellente: elle est passable.
 - La concordance avec la théorie est bonne: elle est faible.
 - La concordance avec la théorie est satisfaisante: elle est douteuse.
 - La concordance avec la théorie est passable: elle est totalement imaginaire.
 - Il est généralement admis que...: 2 collègues pensent comme moi
 - Il est admis que: Je crois que.
 - Il est évident que des travaux complémentaires seront utiles: Je n'ai rien compris.
 - Voici quelques résultats typiques: Voici les meilleurs résultats.
 - Significatif dans un intervalle de confiance de...: non significatif.
 - Les réactifs utilisés furent synthétisés au laboratoire selon des techniques standardisées: Les réactifs furent achetés chez...
 - Malheureusement, les bases quantitatives permettant de tirer profit des résultats n'ont pas encore été formulées: Personne n'est arrivé à comprendre quoi que ce soit à ce qui a été observé.
 - Nous remercions X pour sa précieuse collaboration et Y pour les discussions fructueuses: X a fait le travail et Y m'a expliqué ce que signifiaient les résultats.
-



4. STATISTIQUES

3 statisticiens sortent faire du tir sur cible statique. Le premier statisticien fait feu et tire trop à gauche, le deuxième fait feu, mais tire symétriquement trop à droite. Enfin, le dernier ne tire pas, mais s'exprime triomphalement: "En moyenne nous l'avons eu!"

Patient: "Vais-je survivre à cette opération délicate?"

Chirurgien: "Oui, je suis absolument sûr que vous y survivrez."

Patient: "Comment pouvez-vous être si sûr?"

Chirurgien: "9 patients sur 10 meurent lors de cette opération et hier mon neuvième patient est mort. aujourd'hui"

10 bonnes raisons pour s'orienter dans le domaine de la statistique:

1. Estimer des paramètres est plus simple que de se battre dans la vraie vie
2. Les statisticiens sont des gens reconnus
3. Vous apprendrez l'alphabet grec en entier
4. La probabilité que vous obteniez un job dans ce domaine est > 0.9999
5. Si vous êtes virés, vous pourrez toujours vous reconvertir à l'ingénierie
6. Vous faites ce travail dans la confiance, la régularité et la variabilité
7. Vous êtes normal et le reste du monde est faux
8. La ligne de régression paraît meilleure que la ligne du chômage
9. Vous n'avez jamais besoin d'être exact - seulement approximatif

10. Personne ne comprend ce que vous faites, alors vous avez toujours raison

5. ÉCONOMIE

Pour aider à comprendre le jargon spécifique du marketing..... et éviter de paraître ridicule dans une soirée (pot de départ par exemple...):

1. Michel est à une soirée et voit une nana très attirante. Il s'approche d'elle et lui dit: "Je suis un très bon coup". C'est ce qu'on appelle du "marketing direct"
2. Michel est à une soirée avec un groupe de copains et il voit une nana très attirante. Un de ses amis s'approche d'elle et lui dit: "Tu vois ce garçon là-bas, c'est un très bon coup". C'est ce qu'on appelle de la "publicité."
3. Michel est à une soirée et il voit une nana très attirante. Il lui demande son numéro de téléphone. Le lendemain, il l'appelle et lui dit: "Je suis un très bon coup". C'est ce qu'on appelle du "télémarketing."
4. Michel est à une soirée et il voit une nana très attirante. Il la reconnaît, s'approche d'elle, lui rafraîchit la mémoire et lui dit: "Tu te souviens que je suis un très bon coup ?". C'est ce qu'on appelle du "Customer Relationship Management (CRM)"
5. Michel est à une soirée et il voit une nana très attirante. Il se lève, s'arrange un peu, s'approche d'elle et lui sert un verre. Il lui ouvre la porte lorsqu'elle part, ramasse son sac lorsqu'il tombe, lui offre une cigarette et lui dit: "Je suis un très bon coup". C'est ce qu'on appelle des "relations publiques" ou "public relations" (PR)
6. Michel est à une soirée et il voit une nana très attirante. Il invite à danser toutes ses copines, leur offre à boire et les fait rire ostensiblement par ses plaisanteries très spirituelles. La belle nana l'aborde et lui dit: "J'ai l'impression que tu es un très bon coup". C'est ce qu'on appelle du "lobbying".
7. Michel est à une soirée et il voit une nana très attirante. Elle s'approche de lui et lui dit: "J'ai entendu dire que tu es un très bon coup". C'est ce qu'on appelle le "pouvoir de la marque".
8. Michel est à une soirée et voit une super belle nana. Il la mate avec ses potes, fait des réflexions très fines, se bourre la gueule, ne fait rien du tout et rentre bredouille. C'est ce qu'on appelle la "réalité du marché"...

6. CHIMIE

On vient de découvrir un nouvel élément chimique:

ÉLÉMENT NUMÉRO: 115

NOM: Femme

SYMBOLE: Fm

MASSE ATOMIQUE: Acceptable 60 kg mais des isotopes connus de 40 à 250 kg

OCCURRENCE: Très abondant de par le monde

PROPRIÉTÉS PHYSIQUES:

- Entre en ébullition pour un rien et gèle sans raison
- Conductivité thermique: faible surtout aux extrémités inférieures
- Coefficient de dilatation: augmente avec les années
- Cède aux pressions appliquées aux points sensibles
- Structure moléculaire: parfaite 90/60/90, existe aux USA sous forme croissante 60/90/120 et dans les pays nordiques sous forme dite plate 50/50/50

PROPRIÉTÉS CHIMIQUES:

- Très grande affinité pour l'or, l'argent, le platine et tous les métaux nobles. Absorbe de grandes quantités de substances onéreuses
- Peut exploser spontanément sans avertissement
- Insoluble dans les liquides mais présente une activité grandement augmentée par saturation dans l'alcool
- Réactivité très variable selon les périodes de la journée
- Grande aptitude aux changements d'humeur et à la jalousie.
- Sensible aux élections qui lui transmettent parfois la migraine

UTILISATIONS COURANTES:

- Hautement décorative surtout dans les voitures de sport
- Puissant agent nettoyant
- Aide efficace pour la relaxation et la détente

TEST:

Tourne au vert si placée à côté d'un spécimen de meilleure qualité

PRÉCAUTIONS D'EMPLOI:

- Hautement dangereuse si placée entre des mains non expertes
- Il est illégal d'en posséder plus d'un spécimen, mais il est possible d'en entretenir plusieurs à des endroits différents tant que les différents spécimens n'entrent pas en contact (risque d'explosion)

ATTENTION:

Certains chercheurs d'Amérique du Sud ont découvert le moyen d'en fabriquer artificiellement, présentées généralement sous les marques "Travelo" ou "Dragqueen". Ne consommer que le produit générique

On vient de découvrir un nouvel élément chimique:

ÉLÉMENT NUMÉRO: 116

NOM: Homme

SYMBOLE: Hm

ANALYSE QUANTITATIVE: Mesuré à 17 cm, bien que quelques isotopes existent en 25, 20, 13 et même 10 cm.

DÉCOUVREUR: Ève (découvert par accident un jour où elle avait envie de côtelettes)

LIEU D'EXTRACTION: Se trouve en grandes quantités en présence d'un gisement de Fm très pur

PROPRIÉTÉS PHYSIQUES

- Surface recouverte de poils, raides par endroits, doux dans d'autres
- Bout quand on l'agite, se glace quand on le met en présence de la logique et du bon sens, se liquéfie quand on le traite comme un dieu
- Devient exécrable lorsqu'on le mélange à n'importe quel alcool
- Peut être la cause de maux de tête (ou des maux d'autres parties du corps); à manipuler avec précaution
- Diminue son entropie directement après sa réaction avec l'élément Fm (état se manifestant par des ronflements... zzzzz)
- Augmente sa masse considérablement en vieillissant, perd de ses capacités réactionnelles.
- Se déshydrate rapidement par temps sec.
- Rarement trouvé à l'état pur après 14 ans
- possède souvent un attachement inexplicable à sa roche mère, rendant l'extraction difficile
- Si on le met sous pression, devient trop dur et improductif; n'est productif que si l'on utilise la subtilité, les subterfuges, et la flatterie

PROPRIÉTÉS CHIMIQUES

- Tendance très forte à réagir avec l'élément Fm, même si la réaction est parfois endothermique
- Réputé être le meilleur catalyseur pour les réactions de transformation de l'élément Fm
- Possède la faculté d'entrer en réaction avec à peu près n'importe quoi
- En cas de réaction importante, l'aspect de l'élément change pour virer au rouge cramoisi.
- S'il est saturé en alcool, il devient inerte et repoussant pour la plupart des éléments

- Ne convient pas pour les tâches ménagères et les opérations de nettoyage
- Ne convient pas non plus pour les tâches familiales
- Est neutre en ce qui concerne la courtoisie et l'impartialité

USAGES COURANTS

- Transport de choses lourdes, chauffeur, dîners gratuits au restaurant...
- Usage possible pour les activités sexuelles

TESTS

- Les spécimens les plus purs ne sont pas synonymes de pureté, et ceux qui ont déjà servi, encore moins

DANGERS

- La réaction avec un autre élément Hm est extrêmement violente si l'élément Fm est le catalyseur

La question suivante a réellement été posée en ces termes à l'université de chimie de Washington:

L'Enfer est-il exothermique (dégage-t-il de la chaleur) ou endothermique (absorbe-t-il de la chaleur) ? Appuyez votre réponse avec une preuve.

La plupart des étudiants écrivirent comme preuve de leurs théories la loi de Boyle (Les gaz se réchauffent quand ils sont comprimés et se refroidissent quand ils se décompressent) ou une variante.

Un étudiant, toutefois, a écrit ce qui suit:

Premièrement, nous avons besoin de savoir comment la masse de l'Enfer évolue dans le temps. Ce qui signifie aussi que nous avons besoin de connaître le rythme auquel les âmes vont en Enfer et le rythme auquel elles en sortent. Je pense que nous pouvons sans crainte affirmer qu'une fois qu'une âme est en Enfer, elle n'en sortira plus. Par conséquent, aucune âme ne sort des enfers.

Pour ce qui est des nombreuses âmes qui vont en Enfer, examinons les différentes religions qui existent de par le monde aujourd'hui. Certaines d'entre elles décrètent que si vous n'êtes pas membre de leur religion, vous irez en Enfer. Depuis qu'il y a plus d'une religion de cette sorte et depuis que les gens ne pratiquent qu'une seule religion, nous pouvons en déduire que presque tout le monde et toutes les âmes vont en Enfer.

Avec le rythme des naissances et des morts qui sont ce qu'ils sont, nous pouvons nous attendre à ce que le nombre des âmes en Enfer augmente de façon exponentielle.

Maintenant occupons-nous du rythme d'évolution du volume de l'Enfer, parce que la loi de Boyle prédit que pour que la température et la pression restent les mêmes, le volume de l'Enfer doit s'agrandir proportionnellement aux âmes qui s'ajoutent.

Ceci nous donne deux possibilités:

1. Si l'Enfer croît à un rythme plus lent que celui des âmes qui arrivent en Enfer, alors la température et la pression s'accroissent jusqu'à ce que l'Enfer craque de partout.
2. Bien sûr, si l'Enfer s'agrandit à un rythme plus rapide que le nombre d'âmes en Enfer s'accroît, alors la pression et la température baissent jusqu'à ce que l'Enfer gèle tout entier.

A chemist walks into a pharmacy and asks the pharmacist: "Do you have any acetylsalicylic acid?"

- "You mean aspirin?" asked the pharmacist.

- "That's it, I can never remember that word."

A physicist, biologist and a chemist were going to the ocean for the first time.

The physicist saw the ocean and was fascinated by the waves. He said he wanted to do some research on the fluid dynamics of the waves and walked into the ocean. Obviously he was drowned and never returned.

The biologist said he wanted to do research on the flora and fauna inside the ocean and walked inside the ocean. He too, never returned.

The chemist waited for a long time and afterwards, wrote the observation: "The physicist and the biologist are soluble in ocean water".

CLASSIFICATION OF CHEMISTRY

Physical Chemistry: The pitiful attempt to apply $y=mx+b$ to everything in the universe.

Organic Chemistry: The practice of transmuting vile substances into publications.

Inorganic Chemistry: That which is left over after the organic, analytical, and physical chemists get through picking over the periodic table.

Chemical Engineering: The practice of doing for a profit what an organic chemist only does for fun.

Free radicals have revolutionized chemistry.

Chemist's last words:

- And now the tasting test...

- And now shake it a bit...

- In which glass was my mineral water?

- Why does that stuff burn with a green flame?!?
- And now the detonating gas problem.
- This is a completely safe experimental setup.
- Now you can take the protection window away...
- Where do all those holes in my kettle come from?
- And now a cigarette...

7. INGÉNIERIE

Des scientifiques de la NASA ont développé un fusil spécialement conçu pour projeter des poulets morts dans les pare-brises des avions de ligne, des jets militaires et des navettes spatiales. Le but étant de vérifier les conséquences de possibles collisions avec des volatiles et d'adapter les matériaux des pare-brises en conséquence.

Les ingénieurs britanniques ayant entendus parler de ce genre de fusil ont de suite développés un identique pour effectuer le même type de tests sur leurs trains à très grande vitesse. Lorsque le premier test fut effectué, le poulet traversa le pare-brise, la console de commande, la motorisation pour finir sa course dans le mur opposé.

Horriifiés et étonnés par ces résultats, les ingénieurs britanniques les communiquèrent à leurs confrères américains afin d'obtenir des suggestions de leur part. La NASA leur répondit en une seule phrase: "Décongelez le poulet avant!"

Comprendre les ingénieurs

- Tentative N° 1

Deux élèves ingénieurs marchent le long de leur campus lorsque l'un des deux dit à l'autre, admiratif: "Où est-ce que tu as trouvé ce vélo ?" Le second lui répond: "Ben en fait, alors que je marchais, hier, et que j'étais dans mes pensées, je croise une super nana en vélo qui s'arrête devant moi, pose son vélo par terre, se déshabille entièrement et me dit: "Prends ce que tu veux... ". J'ai donc choisi son vélo.

Le premier opine et lui dit: "Tu as raison, les vêtements auraient certainement été trop serrés.

- Tentative N° 2

Pour une personne optimiste, le verre est à moitié plein.

Pour une personne pessimiste, il est à moitié vide.

Pour l'ingénieur, il est deux fois plus grand que nécessaire.

- Tentative N° 3

Un pasteur, un médecin et un ingénieur jouent au golf. Ils attendent après un groupe de golfeurs particulièrement lents. Au bout d'un moment, l'ingénieur explose et dit: "Mais qu'est-ce qu'ils fichent? ça fait bien un quart d'heure qu'on attend là !" Le docteur intervient, exaspéré lui aussi: "Je ne sais pas,

mais je n'ai jamais vu des gens s'y prendre aussi mal !" Le pasteur dit alors: "Attendez, voilà quelqu'un du golf. On n'a qu'à le lui demander. Dites-moi, il y a un problème avec le groupe de devant. Ils sont plutôt lents, non ?" L'autre répond: "Ah oui, c'est un groupe de pompiers aveugles. Ils ont perdu la vue en tentant de sauver le golf des flammes l'année dernière, alors depuis, on les laisse jouer gratuitement". Le groupe reste silencieux un moment, et le pasteur dit: "C'est si triste. Je vais faire une prière spécialement pour eux ce soir". Le médecin ajoute: "Bonne idée. Et moi, je vais contacter un copain chercheur ophtalmologiste pour voir ce qu'il peut faire". Et l'ingénieur: "Mais putain ! Pourquoi ils jouent pas la nuit ?"

- Tentative N° 4

Un ingénieur traversait la rue lorsqu'une grenouille l'appela et lui dit: "Si tu m'embrasses, je me transformerai en une magnifique princesse". Il se baissa, ramassa la grenouille et la mit dans sa poche. La grenouille lui dit alors: "Si tu m'embrasses, je me transformerai en une magnifique princesse et je resterai à tes côtés pendant une semaine". L'ingénieur sortit la grenouille de sa poche, lui fit un sourire et la replaça dans sa poche. La grenouille se mit alors à crier: "Si tu m'embrasses, je me transformerai en une magnifique princesse, je resterai à tes côtés pendant une semaine et je ferai TOUT ce que tu veux". Encore une fois, l'ingénieur sortit la grenouille de sa poche, lui sourit et la remit dans sa poche. La grenouille lui demanda alors: "Quoi, qu'est-ce qu'il y a ? Je te dis que je suis une magnifique princesse, que je resterai à tes côtés pendant une semaine et que je ferai tout ce que tu veux. Alors pourquoi tu ne m'embrasses pas ?" L'ingénieur répondit: "Regarde-moi, je suis un ingénieur. J'ai pas le temps d'avoir une petite amie. Par contre, une grenouille qui parle, ça, c'est cool !"

- Tentative N° 5

Un journaliste interviewe un paysan corse:

- "Dites-moi, comment faites-vous pour tracer les routes ici? "

- "beh, on lâche un âne et on regarde par où il passe dans la montagne....et c'est là qu'on fait passer la route".

- "et si vous n'avez pas d'âne? "

- "ah....beh on prend un ingénieur.... "

Lors de la course à la conquête spatiale dans les années 1960, la NASA décida pour le besoin de ses astronautes de développer un stylo à bille fonctionnant dans un système sans gravité.

Après un temps considérable de recherche et de développements, le stylo "Astronaute" fut développé pour un coût de 1 million de dollars. Le stylo fonctionna très bien mais trouva un faible intérêt de la part du public sur Terre.

L'Union Sovétique, se trouvant confronté au même problème, utilisa elle un crayon...

The great mathematician John Von Neumann was consulted by a group who was building a rocket ship to send into outer space. When he saw the incomplete structure, he asked, "Where did you get the plans for this ship?"

He was told, "We have our own staff of engineers."

He disdainfully replied: "Engineers! Why, I have complete sewn up the whole mathematical theory of rocketry. See my paper of 1952."

Well, the group consulted the 1952 paper, completely scrapped their 10 million dollar structure, and rebuilt the rocket exactly according to Von Neumann's plans. The minute they launched it, the entire structure blew up. They angrily called Von Neumann back and said: "We followed your instructions to the letter. Yet when we started it, it blew up! Why?"

Von Neumann replied, "Ah, yes! That is technically known as the blow-up problem - I treated that in my paper of 1954."

Dans un laboratoire d'électronique:

-Dis-donc, c'est quoi le semi-remorque sur le parking devant?

- Le semi-remorque?

- Ben oui, le chauffeur dit que tu es au courant...

-Ah oui!! J'ai commandé un condensateur de 1 Farad.

The Dictionary: what engineers say and what they mean by it

- Major Technological Breakthrough: Back to the drawing board.

- Developed after years of intensive research: It was discovered by accident.

- The designs are well within allowable limits: We just made it, stretching a point or two.

- Test results were extremely gratifying: It works, and are we surprised!

- Customer satisfaction is believed assured: We are so far behind schedule that the customer was happy to get anything at all.

- Close project coordination: We should have asked someone else.

- Project slightly behind original schedule due to unforeseen difficulties: We are working on something else.

- The design will be finalized in the next reporting period: We haven't started this job yet, but we've got to say something.

- A number of different approaches are being tried: We don't know where we're going, but we're moving.

- Extensive effort is being applied on a fresh approach to the problem: We just hired three new guys; we'll let them kick it around for a while.

- Preliminary operational tests are inconclusive: The darn thing blew up when we threw the switch.
- The entire concept will have to be abandoned: The only guy who understood the thing quit.
- Modifications are underway to correct certain minor difficulties: We threw the whole thing out and are starting from scratch.
- Essentially complete: Half done.
- We predict...: We hope to God!
- Drawing release is lagging: Not a single drawing exists.
- Risk is high, but acceptable: 100 to 1 odds, or with 10 times the budget and 10 times the manpower, we may have a 50/50 chance.
- Serious, but not insurmountable, problems: It will take a miracle. God should be the program manager.
- Not well defined: Nobody has thought about it.
- Requires further analysis and management attention: Totally out of control.
- The project is designed for high availability: Malfunctions will be blamed on the operators mistakes.
- This project has low maintenance requirements: We wouldn't let the technicians change a light bulb, much less fool around with our baby.
- The software is being developed without excessive process overhead: The documentation will be written in clear and lucid Chinese.
- The delivery is scheduled for the last quarter of next year: This leaves us plenty of time to decide who to blame for it being late.

-
- How many first year engineering students does it take to change a light bulb?: None. That's a second year subject.
 - How many second year engineering students does it take to change a light bulb?: One, but the rest of the class copies the report
 - How many third year engineering students does it take to change a light bulb?: Will this question be in the final examination?
 - How many civil engineers does it take to change a light bulb?: Two. One to do it and one to steady the chandelier
 - How many electrical engineers does it take to change a light bulb?: None. They simply redefine darkness as the industry standard
 - How many computer engineers does it take to change a light bulb?: Why bother? The socket will be obsolete in six months anyway

- How many mechanical engineers does it take to change a light bulb?: Five. One to decide which way the bulb ought to turn, one to calculate the force required, one to design a tool with which to turn the bulb, one to design a comfortable - but functional - hand grip, and one to use all this equipment.
- How many nuclear engineers does it take to change a light bulb?: Seven. One to install the new bulb and six to figure out what to do with the old one for the next 10,000 years.

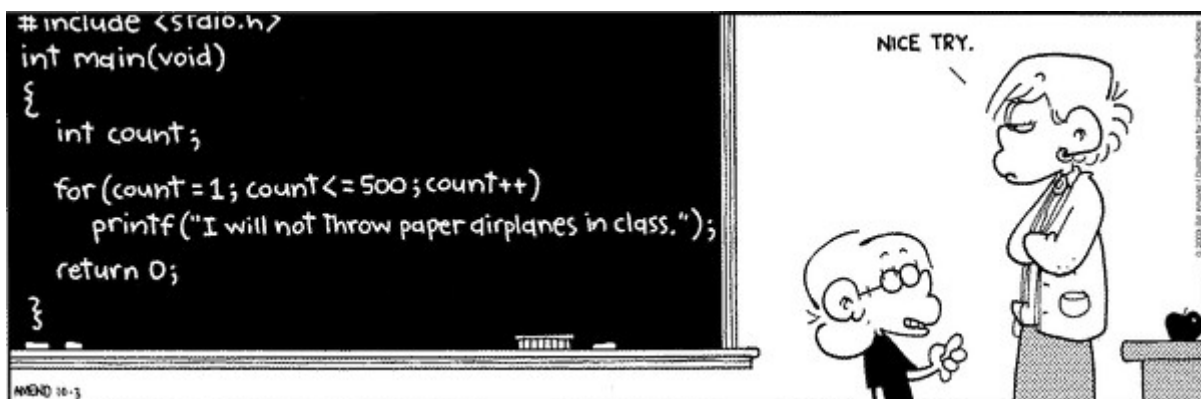
Cela se passe à Moscou,: un couple de touristes demande son chemin, sur un pont, à un quidam, russe, qui se trouve être ingénieur (trouver un pont moscovite connu, et la première rue qui le croise sur la berge...).

Le gars leur dit: "vous traversez, et, d'ici 50 m, vous tournez à droite..."

Remerciements de la part des touristes... , puis ils partent. Alors le gars leur court derrière:

"Attendez, attendez ! Je viens de me souvenir que le pont fait 70 m. Si vous tournez à droite au bout de 50 m, comme je vous l'ai dit, vous tomberez à l'eau."

8. INFORMATIQUE



Comment devenir un bon Hacker ?

Bon, si tu veux être un vrai hacker, il va te falloir Linux.

Là, tu as 2 solutions:

1. Tu es un sale bourgeois capitaliste et tu l'achètes 150 balles à la FNAC
2. Tu es un vrai trou du cul, et là tu le downloades par le Net

Évidemment, tu es un vrai trou du cul donc tu ouvres ton tit client FTP et tu te tapes tranquillement les 20 ou 25 heures de download pour une Slack ou une Debian. Evite la Red Hat, ça fait trop grand public, toi t'es un mec uNdERgrOuNd maintenant, c'est normal, t'es Hacker.

Bon, tu as ton Linux, maintenant c'est bon oublie-le. Pas la peine de se casser le cul à apprendre un nouvel OS dont tu ne te serviras jamais parce que "XWing vs Tie Fighter" tourne pas dessus. La meilleure solution consiste carrément à niquer lilo, comme ça tu es sûr que tu ne booteras que sous Windows95. C'est une solution élégante que de nombreux trous du cul semblent avoir choisie. Pour ça, ouvre une session DOS par Windows et tape fdisk /mbr. Ca va effacer lilo qui était installé sur le MBR de ton disque dur, comme ça tu n'auras plus à te soucier de Linux.

L'essentiel est de l'avoir, pas de savoir s'en servir.

"Ouais mais comment je peux prouver aux gens que j'ai Linux et passer pour un gros rebelle ?"

C'est une question bien naturelle. J'ai pensé à toi petit looser et voici une série de phrases qu'il faut balancer à propos de Linux:

- "Linux c'est trop puissant, t'es complètement libre par rapport à ces OS de fachos genre Windaube. De toute façon, MS c'est trop ripoux."
- "Bah si t'es un débutant, va pas sous Linux, c'est fait pour les eLiTeS ce truc, toi reste sous winfuck."
- "Dis, tu sais pas où je pourrais trouver la libc5.4.36 ? Parce que chez moi la 5.4.35 est incompatible avec les modifs que j'ai faites au kernel."
- "Ça pue netscape, moi ça me dumpe des cores de 10 mégas dès que le lance, je préfère Lynx au moins c'est pas prise de gueule c'est mieux le mode texte."
- "Ralala le bouffon que c'est lui ! Il s'est installé une Red Hat !! Tain c'est de la daube les Red Hat c'est nul y'a que la Debian qui est bien, au moins tu sais ce que tu fais t'es le master de ton system nan vraiment c'est ripoux Red Hat."

Avec ce genre de petites phrases, tu te retrouveras très vite classé dans la catégorie "OK c'est un trou du cul, mais un trou du cul sous Linux", ce qui est la première étape pour être un vrai hacker. Maintenant que tout le monde sait que tu as ton Linux, il faut passer au stade suivant, celui du pro des réseaux, genre le mec qui maîtrise ICMP à mort. C'est la deuxième étape de ton long périple.

Ici, il faut mettre la main au porte-monnaie. Direction la FNAC, tu achètes n'importe quel bouquin sur Unix et sur les réseaux. L'essentiel est que le titre soit compliqué. Un petit "Protocole rlogin sur réseau Ethernet en sous-adressage" sera du meilleur effet. N'hésite pas, dès que tu ne comprends même pas le titre, il faut acheter le bouquin: c'est pas pour lire, c'est pour impressionner tes autres potes trous du cul.

Une bonne méthode consiste à acheter un bouquin genre "TCP/IP volume 43" et de prendre des mots

au hasard à apprendre par coeur: rai socket, sur-adressage, FDDI, telnet par exemple. Ensuite, tu les ressors dans une phrase, même hors contexte c'est pas grave personne n'ira vérifier ce que ça veut dire. Par exemple, il ne faut pas hésiter à balancer un "Le Telnet, ça prends combien de rai sockets en sur-adressage sur un FDDI" sur un bon gros channel de cowboyz, ça impressionne toujours, et personne n'ira te dire que ça n'a aucun sens, ne t'inquiète pas.

Dispose ensuite ces bouquins dans ta chambre, avec les titres les plus compliqués aux endroits les plus visibles. Corne quelques pages pour faire plus vrai. Prends aussi quelques feuilles et dessine des schémas bidons de réseau, ou met des trucs genre 123.44.5.34 root/lydia pour faire croire que tu te chopes de password comme un ouf. Faut se la jouer à mort, ne jamais hésiter à en rajouter, scanne-toi une photo de Mitnick et accroche la au-dessus de ton lit, ou met des autocollants à tête de mort sur ton UC pour bien dire que maintenant, t'es un voyou, un mec dangereux.

Pour compléter le tout et vraiment passer pour un hacker, il ne faut pas hésiter à dire des conneries du genre "Je ne suis qu'un assoiffé de connaissances". OK, tu quadruples ta seconde, mais bon c'est pas grave, tu aimes quand même apprendre, c'est ta grande passion et tu as beaucoup de volonté. Précise bien que jamais tu ne causes de dégâts aux très nombreuses machines que tu pénètres, dis que tu fais juste ça "pour le challenge intellectuel". Oui, là, il faudra te forcer pour ne pas exploser de rire, mais entraîne-toi devant ta glace avant.

Quand on est un mec dangereux comme toi, on doit se réunir avec d'autres bandits pour mettre en péril la sûreté de l'État. Pour ça, il existe LE rendez-vous de toute la racaille, c'est le "Meet 2600". Tous les mois, tu iras dans un MacDo de Paris, Place d'Italie, et là tu rencontreras des grands monsieurs, des mecs qui ont rebooté tout Internet avec un prog en Visual Basic et qui ont des coupes de cheveux de rebelles de la société.

Bon, tu n'y apprendras pas grand chose, les loosers qui viennent là-bas se branlent entre eux en se disant "Ouais, on est des hAcKeRz, on est sans pitié, on est des vrais durs, oh zut il est déjà 18 heures faut que je rentre ma mère va me cogner sinon". Tu pourras quand même avoir un vrai frisson en t'imaginant que le MacDo est truffé de caméras et de micros, et que tous les employés sont des agents de la DST qui écoutent des conversations aussi dangereuses que:

- Trouduku1: il est à combien le Whooper ?
- Trouduku2: euh MacDo fait des Whoopers maintenant?
- Trouduku1: bah ouais ils en ont toujours fait nan ?

La communauté des hAcKeRz aime bien aussi les raves. Ça fait partie du trip "rebel no future fuck da society, on gobe des extas on écoute de la musique de daube mais on s'en fout c'est super parce que c'est interdit ". N'hésite pas à te rendre là-bas, ça fait incontestablement partie de la culture du paumé que d'aller jouer les chauds dans ces soirées.

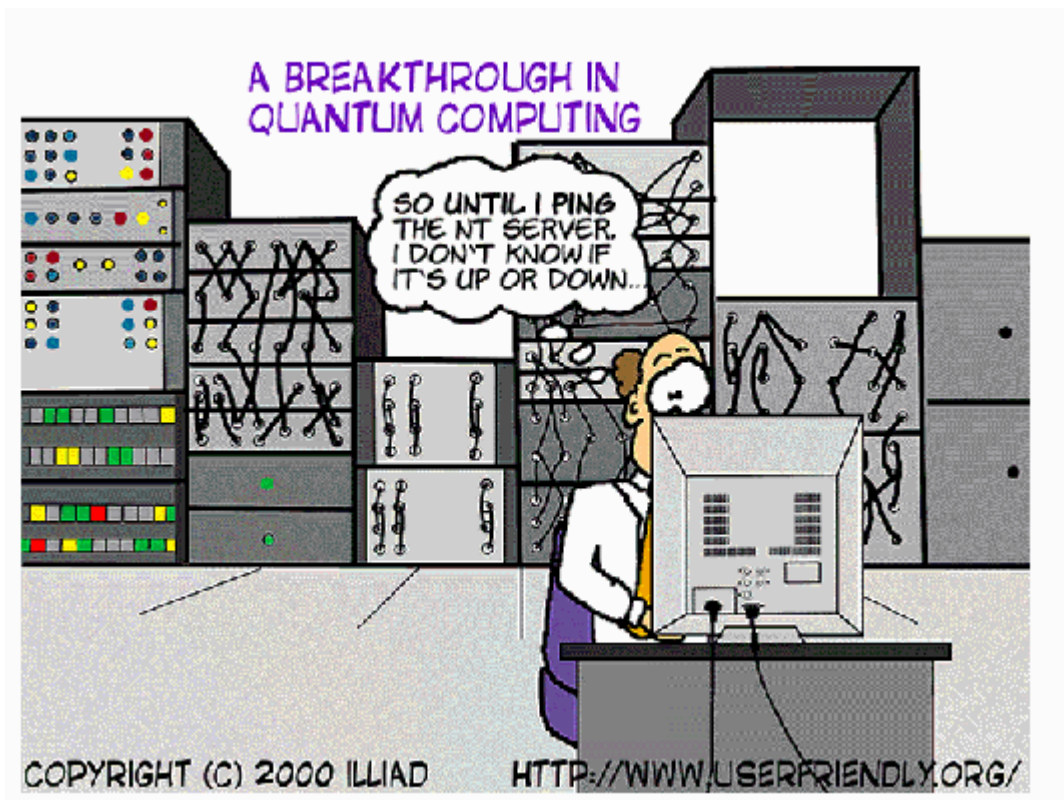
Toi, t'es un vrai trou du cul qui hacke, et tu entends bien répandre ton savoir pour former d'autres minables comme toi. Pour ça, il existe les e-zine. On peut citer les plus connus comme NoWay ou NoRoute ou le pire cotoie le meilleur (et c'est dommage pour le meilleur...) mais aussi des vraies merdes qui mériteraient d'être plus connues, comme l'excellent Core-Dump qui est une véritable farandole de guignolos expliquant des trucs archi connus dans un français que mon chat comprends mieux que moi.

Évidemment, tu n'as pas lu les bouquins sur Unix, tu n'as jamais hacké la moindre machine de ta vie donc tu ne sais pas quoi écrire. Rassure-toi, tu n'es pas le seul dans ce cas. La meilleure méthode est de pondre un article sur le rap, à raconter sa dernière rave ou à pomper Phrack sans rien comprendre. Là

encore, si tu pompes Phrack, n'hésite pas à carrément corriger le mec ou à rajouter des trucs pour faire plus compliqué, personne n'ira vérifier, donc vas-y lâche-toi t'es un assoiffé de connaissances, oublies pas.

Maintenant, c'est clair, tu es un vrai hacker, une racaille de l'IRC, un loubard d'Internet, tu fais peur à toutes les agences gouvernementales et IBM veut t'embaucher pour sécuriser leur réseau parce que cette pédale d'Henri leur a encore collé un virus d'Internet. Il va donc falloir, au quotidien, se comporter comme un hacker, un vrai, un dur, c'est à dire avec un esprit hacker et un langage de hacker.

Un hacker, ça vit avant tout sur IRC. Une fois que tes amis et ta famille auront bien vu que tu as changé, que tu n'es plus le même homme, il va falloir répandre aussi la nouvelle sur IRC et te faire des nouveaux amis qui seront comme toi des trous du cul. Fini les #coquelicots ou les #amitié_fr, maintenant tu devras aller dans les bas fond de l'IRC, le cyber-Bronx, nuke-city, là où seuls les vrais cogneurs réussissent à se faire une place dans cet univers de violence. Pour ça, tu vas devoir passer du stade hacker trou du cul à celui de trou du cul sur IRC qui se la pète Mitnick, à savoir le c0wb0y.



//--- cherchez votre niveau en programmation dans l'échelle suivante
//--- l'objectif est d'écrire un programme affichant "hello world" à l'écran

Terminale:

```
10 PRINT "HELLO WORLD"  
20 END
```

1ère année du gymnase:

```
program HELLO(input, output)
begin
writeln('HELLO WORLD')
end
```

2ème année du gymnase:

```
defun HELLO
print
cons HELLO (list WORLD)
```

Fraîchement sorti de l'institut de technologies:

```
#include void main(void)
char *message[] = "HELLO ", "WORLD" ;
int i;
for i = 0; i #include class string
private:
int size;
char *ptr;
public:
string(): size(0), ptr(new char(' 0'))
string(const string &s): size(s.size)
ptr = new char[size + 1];
strcpy(ptr, s.ptr);
~string()
delete [] ptr;
friend ostream &operator main()
char *tmp;
int i=0;
/* on y va bourin */
tmp=(char *)malloc(1024*sizeof(char));
while (tmp[i]="HELLO WORLD"[i++]);
/* Ooopps y'a une infusion ! */
i=(int)tmp[8];
tmp[8]=tmp[9];
tmp[9]=(char)i;
printf("%s n",tmp);
```

Apprenti Hacker:

```
#!/usr/local/bin/perl
$msg="HELLO, WORLD. n";
if ($#ARGV >= 0)
while(defined($arg=shift(@ARGV)))
$outfilename = $arg;
open(FILE, ">". $outfilename) || die "Can't write
$arg:$! n";
print (FILE $msg);
```

```
close(FILE) || die "Can't close $arg: $! n";
else
print ($msg);
1;
```

Hacker expérimenté:

```
#include #define S "HELLO, WORLD n"
main() exit(sprintf(S) == strlen(S)? 0: 1);
```

Hacker très expérimenté:

```
% cc -o a.out ~/src/misc/bv/bv.c
% a.out
```

Gourou des Hackers:

```
% cat
Hello, World
^^D
```

Directeur junior:

```
10 PRINT "HELLO WORLD"
20 END
```

Directeur:

```
mail -s "HELLO, WORLD."
Pierre, pourrais-tu m'écrire un programme qui écrit "HELLO, WORLD." à l'écran?
J'en ai besoin pour demain.
^^D
```

Président directeur général:

```
% zmail Francis
J'ai besoin d'un programme "HELLO, WORLD." Pour cette après-midi.
Président Directeur Général
% letter
letter: Command not found.
% mail
To: ^X ^F ^C
% help mail
help: Command not found.
% damn!
!: Event unrecognized
% logout
```

Ce que disent les ingénieurs en informatique... et ce qu'il faut comprendre:

- "Nous allons inscrire ce projet au planning": On s'en occupera si on a rien d'autre à faire
- "C'est un programme complètement nouveau!": C'est pas du tout compatible avec l'ancienne version
- "Ce programme ne nécessite aucune maintenance": C'est impossible à déboguer
- "Ce programme ne nécessite que peu de maintenance": C'est quasiment impossible à déboguer
- "Nous respecterons les standards": On a toujours fait comme ça et ce n'est pas aujourd'hui qu'on va changer
- "Nous tenons à respecter les standards": Vous n'allez pas remettre en cause tout ce qu'on vient de faire
- "La nouvelle version de ce programme est 100% compatible avec la précédente": On n'a touché à rien
- "Différentes approches ont été tentées": On essaie encore de deviner ce qui se passe.
- "On approche d'une solution": On s'est réunis pour prendre un café...
- "Les tests préliminaires n'ont pas été franchement concluants": Ce satané programme a planté dès qu'on a lancé
- "Il va falloir abandonner le concept en son entier": La seule personne qui comprenait quelque chose vient de démissionner
- "On prépare un rapport complet, selon une approche entièrement nouvelle": On vient juste d'engager trois bleus sortis de l'école
- "C'est une avancée technologique majeure": On n'arrive toujours pas à comprendre pourquoi ça ne marche pas
- "C'est le résultat d'années de développement": On a enfin réussi à faire fonctionner un bout du programme...
- "C'est en cours": On est tellement dans le pétrin que c'est sans espoir
- "Faites-nous part de vos réflexions": On écouterait ce que vous avez à dire tant que ça ne remet pas en cause ce qui est déjà fait, ou ce que nous avons décidé de faire
- "Nous allons y jeter un coup d'oeil": Laissez tomber! On a déjà assez de problèmes comme ça...
- "Je n'ai pas reçu votre e-mail": Ça fait des lustres que je n'ai pas vérifié ma messagerie...

9. GÉNÉRALITÉS



Un homme, dans la nacelle d'une montgolfière ne sait plus où il se trouve. Il descend et aperçoit une femme au sol. Il descend encore plus bas et l'interpelle:

- Excusez-moi! Pouvez-vous m'aider? J'avais promis à un ami de le rencontrer et j'ai déjà une heure de retard, car je ne sais plus où je me trouve...

La femme au sol répond:

- Vous êtes dans la nacelle d'un ballon à air chaud à environ 10 m du sol. Vous vous trouvez exactement à 49°,28' et 11" Nord et 8°,25' et 58" Est .

- Vous devez être ingénieur dit l'aérostier.

- Effectivement», répond la femme, comment avez-vous deviné?

- Eh bien, dit l'aérostier, tout ce que vous m'avez dit à l'air techniquement parfaitement correct, mais je n'ai pas la moindre idée de ce que je peux faire de vos informations et en fait je ne sais toujours pas où je me trouve. Pour parler ouvertement, vous ne m'avez été d'aucune aide. Pire, vous avez encore retardé mon voyage!

- La femme lui répond: Vous devez être un manager!

- C'est exact, répond l'homme avec fierté, mais comment avez-vous deviné???

- Eh bien, dit la femme, vous ne savez ni où vous êtes, ni où vous allez. Vous avez atteint votre position actuelle en chauffant et en brassant une énorme quantité d'air. Vous avez fait une promesse sans avoir la moindre idée comment vous pourriez la tenir et vous comptez maintenant sur les gens situés en dessous de vous pour qu'ils résolvent votre problème. Votre situation avant et après notre rencontre n'a pas changé, mais comme par hasard, c'est moi maintenant qui à vos yeux en suis responsable...

Un homme va un samedi à un mariage en Corse dans un petit village. Il est en retard et il conduit aussi vite que possible sur des routes sinueuses. Soudain, après un virage, il doit s'arrêter net, un troupeau d'ovins occupe toute la route. Le berger est là et fait lentement avancer son troupeau. L'automobiliste use du klaxon de son véhicule plusieurs fois sans le moindre effet. Au bout de quelques minutes, le conducteur apostrophe le berger et lui dit:

- Je suis en retard, je vais à un mariage qui sera suivi d'un méchoui, si je vous dis combien de moutons vous avez, m'en céderez-vous un ?
- Volontiers dit le berger, je ne suis pas à un près. Le conducteur prend sa calculatrice et au bout d'une minute annonce: 1233.
- Vous avez gagné dit le berger, choisissez votre animal.

Le conducteur en désigne alors un. Le berger dit alors:

- Si je trouve quelle est votre profession, me rendrez-vous ma bête?
- Bien sûr dit le conducteur, je vous écoute.
- Vous êtes haut fonctionnaire et vous avez fait l'ENA ou une autre grande école de ce genre.
- Vous avez raison dit le conducteur, mais comment avez-vous deviné ?

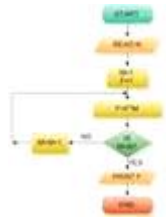
Le berger:

- Je vous prie de me rendre mon chien.
-



Ouvrier

+



Processus

=



Ingénieur



Ingénieur

+



Sociabilité

=



Marketing



Marketing

-



Vérité

=



Commercial



Commercial

-



Cerveau

=



Manager



Manager

+



Ego

=



Project Manager



Project Manager

-



Humour

=



Ressources Humaines

RÉFÉRENCES

REMERCIEMENTS

Dans l'ordre alphabétique, merci à:

- *Guglielmetti Rafael*, passionné du développement web et ancien étudiant en mathématiques à l'E.P.F.L. (École Polytechnique Fédérale de Lausanne) en Suisse, pour le développement et l'amélioration du code PHP du forum.
- *Harmel Léon* (†2012), ingénieur diplômé en électromécanique avec une spécialisation en électronique et automatique, responsable au laboratoire de recherches physiques chez A.C.E.C à Charleroi, pour l'apport de la documentation qui a servi à la rédaction des chapitres de physique quantique corpusculaire, ondulatoire, de physique quantique des champs et de calcul spinoriel et relativité générale.
- *Isoz Vincent*, ingénieur diplômé en physique HES (Bachelor of Science) de l'école d'ingénieurs de Genève (CH) avec une spécialisation en génie nucléaire. Responsable et créateur/développeur du site Sciences.ch et rédacteur de la plus grande partie du contenu.
- *Ricchiuto Ruben*, ingénieur diplômé en physique HES (Bachelor of Science) de l'école d'ingénieurs de Genève (CH) et licencié (bachelor) en maths de l'université de Genève (CH), pour sa précieuse aide et disponibilité en physique des plasmas, électromagnétisme, physique quantique, statistiques, topologie, analyse et chimie quantique et encore beaucoup d'autres domaines touchant aux mathématiques pures et informatiques.
- *Habitués des forums Les Mathematiques.net et Futura-Sciences.com*, pour leur précieuse aide dans de nombreux domaines de la mathématique et de la physique. Les débats et discussions qui y ont lieu permettent de constamment améliorer l'aspect pédagogique et accessible du présent site.
- *Wikipédia* et *PlanetMath* à qui je suis redevable de nombreux emprunts presque au mot par mot (et c'est réciproque...).

Et encore merci à tous les internautes, webmasters et professeurs anonymes pour leurs sites et documents de qualité mis à disposition gratuitement et anonymement sur la toile et aux intervenants réguliers du forum. J'ai parfois récupéré mot à mot leurs explications qui ne nécessitaient ni compléments ni corrections.

Pour terminer, je souhaiterais remercier ici les quelques collègues et clients qui ont bien voulu me faire part de leurs remarques pour améliorer le contenu de ce site. Il est cependant certain qu'il est encore perfectible sur de nombreux points.



RÉFÉRENCES BIBLIOGRAPHIQUES

Puisque le présent travail constitue un ouvrage de référence, nous avons pris le parti d'exclure du texte tout renvoi bibliographique. Nous avons donc simplement rassemblé une bibliographie finale allégée sans prétention d'exhaustivité et selon un classement d'utilisation.

Les ouvrages listés ci-dessous constituent ce que je considère comme le meilleur des références payantes (donc sans compter les PDF's gratuits) qui ont été consultées à ce jour pour la préparation de ce site Internet et qui leur est redevable de nombreux emprunts de grande qualité. Leur lecture ne pourra donc être que profitable pour tous ceux qui désirent améliorer, approfondir et élargir leurs connaissances.

De nombreuses références ne peuvent cependant pas ou plus être trouvées sur le marché et il faut aussi que le lecteur se rappelle que chaque entrée de la table faisant elle-même référence à des dizaines d'autres ouvrages, il est alors impossible de donner une liste exhaustive objective (ce serait un cercle vicieux) de tous les ouvrages de qualité existant dans les domaines traités par Sciences.ch.

Et la liste globale:

★ Support de qualité tant du point de vue contenu que contenant

Σ Contenu avec développements et démonstrations

📖 Magazine ou journal

😊 Ouvrage sympathique de vulgarisation scientifique

❤ Coup de coeur personnel

🗨 Débats et discussions scientifiques

- [001] ★ Σ *Encyclopédie Larousse Méthodique II*, Dépôt légal 1965-3^e N°3177
ISBN-10: -
- [002] ★ Σ 😊 ❤ *Physique Générale*, Alonso/Finn, Éditions InterÉditions U
ISBN-10 volume I: 2729601376 (538 pages) - Imprimé en 1997
ISBN-10 volume II: 2729601368 (678 pages) - Imprimé en 1997
- [003] ★ Σ *La relativité générale*, M. Ludvigsen, Éditions Dunod
ISBN-10: 210004896 (241 pages) - Imprimé en 2005
- [004] ★ 😊 *Physique*, Section Génie Nucléaire, J.-M. Thiébaud, Service d'impression EIG
ISBN-10: -
- [005] ★ 😊 *Physique nucléaire*, P. Triscone, Service d'impression EIG
ISBN-10: -
- [006] Σ *Plasma diagnostics*, O. Auciello + D.L. Flamm, Éditions Academic Press
ISBN-10: 0120676354 (456 pages) - Imprimé en 1989
- [007] ★ Σ *Physique quantique*, J. Laplace, Service d'impression EIG
ISBN-10: -
- [008] ★ Σ *Optique géométrique*, J. Laplace, Service d'impression EIG
ISBN-10: -
- [009] ★ Σ *Élasticité et mécanique des fluides*, J. Laplace, Service d'impression EIG
ISBN-10: -

- [010] ★ C.R.M., Commission Romande de Mathématique, Physique et Chimie, Éditions du Tricorne
ISBN: 2829300607 (282 pages) - Imprimé en 2000
- [011] 🧮 *Physique des plasmas*, J.-L. Delcroix + A. Bers, InterÉditions
ISBN-10: 2868833691 (499 pages) - Imprimé en 1994
- [012] ★ 🧮 😊 *Électricité et Magnétisme*, Resnick/Halliday, Éditions ERPI
ISBN-10: 2761300196 (299 pages) - Imprimé en 2003
- [013] ★ 📖 😊 ❤️ *La nature et le principe de moindre action*, Les cahiers de Science&Vie N°68 (Magazine), Avril 2002
- [014] ★ 🧮 *Probabilités et statistiques*, J. Istas, Éditions Dunod
ISBN-10: 2729879889 (166 pages) - Imprimé en 2000
- [015] ★ 🧮 ❤️ *Calcul différentiel et intégral*, N. Piskunov, Édition Mir (Moscou)
ISBN-10: (511 pages) - Imprimé en 1972
- [016] ★ 🧮 ❤️ *Thermodynamique et énergétique*, L. Borel, Éditions Presses polytechniques et universitaires romandes
ISBN-10: 2880742145 (814 pages) - Imprimé en 2005
- [017] 😊 *Calculs commerciaux et bancaires*, J.-Ch. Corbaz + D. Goetschi, Collection C.C.L
ISBN ?
- [018] *Théorie Financière*, R. Cobbault, Éditions Economica
- [019] *Leçons d'algèbre*, Ch. Briot, Éditions Paris
ISBN-10: - (314 pages) - Imprimé en 1874
- [020] ★ 📖 😊 ❤️ *Newton*, Les cahiers de Science&Vie N°13 (Magazine)
- [021] ★ 🧮 ❤️ *Algèbre linéaire*, R. Cairoli, Éditions Presses polytechniques et universitaires romandes
ISBN-10: 2880741874 (325 pages) - Imprimé en 1993
- [022] ★ 🧮 ❤️ *Le calcul tensoriel en physique*, J. Hladik + P.-E. Hladik, Éditions Dunod
ISBN-10: 2100040715 (172 pages) - Imprimé en 1993
- [023] 😊 *Acoustique et électroacoustique*, cours de A. Pitter, Service d'impression EIG
ISBN-10: -
- [024] ★ *Formulaire Technique Gieck*, K. + R. Gieck
ISBN-10: 3920379187 (656 pages) - Imprimé en 1997
- [025] ★ 🧮 ❤️ *Introduction à la logique (théorie de la démonstration)*, R. David + K. Nour + C. Raffalli, Éditions Dunod
ISBN-10: 2100048929 (352 pages) - Imprimé en 2004
- [026] ★ *Introducing special relativity*, W.S.C Williams, Éditions Taylor&Francis,
ISBN-10: 0415277620 (249 pages) - Imprimé en 2002
- [027] ★ 😊 ❤️ *Les Trous Noirs*, M. Froissart, Éditions Taylor&Francis
- [028] ★ 🧮 😊 ❤️ *Physique Générale*, F. Rothen, Éditions Presses polytechniques et universitaires romandes
ISBN-10: 2880743966 (862 pages) - Imprimé en 1999
- [029] ★ 🧮 ❤️ *Cours de mécanique quantique*, Y. Ayant + E. Belorizky, Éditions Dunod
ISBN-10: 2100047426 (339 pages) - Imprimé en 2000
- [030] 🧮 *Introduction to Bessel function*, F. Bowman, Éditions Dover publications
ISBN-10: 486604624 (134 pages) - Imprimé en 1958

- [031]    *Physique*, E. Hecht, Éditions DeBoeck-Université
ISBN-10: 2744500186 (1328 pages) - Imprimé en 1999
- [032]   *Introduction à la théorie des Nombres*, Éditions Modulo, J.-M. De Koninck + A. Mercier
ISBN-10: 2891135008 (254 pages) - Imprimé en 1954
- [033]    *Calcul différentiel et intégral*, J. Douchet + B. Zwahlen, Éditions Presses polytechniques et romandes
ISBN-10 volume I: 2880741963 (241 pages) - Imprimé en 2007
ISBN-10 volume II: 2880742579 (167 pages) - Imprimé en 2007
- [034]  *Introduction à la théorie des ensembles*, P.R. Halmos, Éditions Jacques Gabay
ISBN-10: 876471264 (127 pages) - Réimprimé en 1997
- [035]   *L'art du secret*, Dossiers pour la Science (Magazine)
- [036]    *Physique statistique*, B. Diu + C. Guthmann + D. Lederer + B. Roulet Éditions Hermann
ISBN-10: 2705660658 (1001 pages) - Imprimé en 1989
- [037] *Standard Mathematical Tables and Formulas*, D. Zwillinger, Éditions CRC Press
ISBN: 084932479 (812 pages)
- [038]   *Compléments de Mathématiques*, A. Angot, Éditions Masson
ISBN-10: - (869 pages)
- [039]    *Éléments de théorie des quanta et de mécanique ondulatoire*, Louis de Broglie, Éditions Gauthier-Villars
ISBN-10: - (302 pages) - Réimprimé en 1953
- [040]    *Physique atomique*, E. Chpolski, Éditions Mir
ISBN-10: - (Tome 1: 540 pages + Tome 2: 413 pages) - Imprimé en 1977
- [041]    *Physique atomique*, B. Cagnac + J.-C. Pebay-Peyroula, Éditions Dunod
ISBN-10: 2040004246 (295 pages) - Imprimé en 2005
- [042]   *Exercices de physique*, Le Morvan, Éditions Vuibert
ISBN-10: - (251 pages) - Imprimé en 1973
- [043]   *Mécanique analytique*, A. Charlier + A. Bérard + M.-F. Charlier, Éditions Ellipses
ISBN-10: 272988940 (237 pages)- Imprimé en 1989
- [044] *Quantum Mechanics*, F. Mandl, Éditions John Wiley & Sons
ISBN-10: 0471931551 (301 pages) - Imprimé en 1992
- [045]  *Physik für Hochschulanänger*, H. Wegener, Éditions B.G. Teubner Suttgart
ISBN-10: 351913053 (498 pages)
- [046]    *Toute la physique*, H. Stöcker + F. Jundt + G. Guillaume, Éditions Dunod
ISBN-10: 2100039423 (1180 pages) - Imprimé en 2007
- [047]  *Mécanique des milieux continus*, G. Duvaut, Éditions Dunod
ISBN-10: 2100039326 (292 pages) - Imprimé en 1998
- [048]   *Matière à pensée*, J.-P. Changeux + A. Connes, Édition Odible Jacob
ISBN-10: 2738108156 (266 pages) - Imprimé en 1989
- [049]   *L'Univers élégant*, B. Green, Éditions Robbert Laffont
ISBN-10: 2221090659 (656 pages) - Imprimé en 2005
- [050]   *L'infini dans la paume de la main*, M. Ricard + T.X. Thuan, Édition NiL
ISBN-10: 2841111741 (399 pages) - Imprimé en 2002

- [051] 📖❤️ *Le chaos et l'harmonie*, T.X. Thuan, Éditions France Loisirs
ISBN-10: 2744120448 (478 pages) - Imprimé en 1998
- [052] 📖❤️ *La plus belle histoire du monde*, H. Reeves + J. de Rosnay + Y. Coppens + D. Simonnet, Éditions France Loisirs
ISBN-10: 2744104248 (186 pages) - Imprimé en 2001
- [053] ⭐😊❤️ *Algèbre*, E. W. Swokowski + J. A. Cole
ISBN-10: 2606006755 (600 pages) - Imprimé en 1999
- [054] *The mathematics of financial derivatives*, P. Wilmott + S. Howison, Éditions Cambridge
ISBN-10: 0521497892 (317 pages) - Imprimé en 1997
- [055] 📊 *Théorie des graphes*, O. Cogis + C. Robert, Éditions Vuibert
ISBN 2711753212
- [056] ⭐📊❤️ *Les spineurs en physique*, J. Hladik, Éditions Masson
ISBN-10: 2225853134 (190 pages) - Imprimé en 1996
- [057] ⭐📊❤️ *Quantum Field Theory*, A. Lahiri + P. B. Pal, Éditions CRC Press
ISBN 0849309778
- [058] ⭐❤️ *Chimie Générale*, G. Gesier + G. Delpin, Éditions du Tricorne
ISBN-10: - (315 pages) - Imprimé en 1993
- [059] ⭐📊❤️ *Chimie Générale*, R. Didier, Éditions Lavoisier
ISBN-10: 274300181 (1071 pages) - Imprimé en 2004
- [060] ⭐📊 *Théorie de la spéculation*, L. Bachelier, Éditions Jacques Gabay
ISBN-10: 2876471299 (148 pages) - Réimprimé en 1995
- [061] ⭐😊 *Initiation à la théorie des jeux*, J.-L. Boursin, Éditions Montchrestien
ISBN-10: 227511069 (188 pages) - Imprimé en 1998
- [062] *La maintenance (mathématiques et méthodes)*, P. Lyonnet, Éditions TEC&DOC
ISBN-10: 2743004193 (393 pages) - Imprimé en 2000
- [063] *Gestion de la production*, V. Giard, Éditions Economica (2ème édition)
ISBN-10: 2717815279 (1068 pages) - Imprimé en 1988
- [064] 📊 *Méthodes numériques*, Arrigo + Marchino + Wilhem
Service d'impression EIG
ISBN-10: -
- [065] 📊⭐❤️ *A first course in String Theory*, B. Zwiebach, Éditions Cambridge Press
ISBN-10: 052183143 (558 pages) - Imprimé en 2004
- [066] ⭐😊❤️ *Mathématiques financières & Actuarielles*, J.-P. Favre, Éditions Digilex
ISBN-10: 297001081X (269 pages) - Imprimé en 2004
- [067] ⭐😊 *Statistiques descriptives pour les gestionnaires*, V. Giard, Éditions Economica
ISBN-10: 2717828893 (111 pages) - Imprimé en 1995
- [068] 📊⭐ *Cours de calcul tensoriel appliqué*, M. Denis-Papin + A. Kaufmann, Éditions Albin Michel
ASIN: B0000DSGRM (388 pages) - Imprimé en 1961
- [069] 😊 *La modélisation du risque et les simulations de Monte-Carlo*, H. Thiriez, Éditions Economica
ISBN-10: 2717848223 (221 pages) - Imprimé en 2003

- [070]    *Dynamics of the Atmosphere (A course in Theoretical Meteorology)*
W. Zdunkowski + A. Bott, Éditions Cambridge
ISBN-10: 052100666X (738 pages)- Imprimé en 2003
- [071]    *Cosmologie*, E. Elbaz, Éditions Ellipses
ISBN-10: 2729892117 (268 pages) - Imprimé en 1998
- [072]    *Analyse I, Analyse II, Analyse III*, L. Schwartz, Édition Hermann
ISBN-10: 2705661611 (1291 pages) - Imprimé en 1997
- [073]   *Comprendre (enfin) les nouveaux instruments financiers*, P. Novellos
ISBN-10: - (55 pages) - Imprimé en 1998
- [074]     *Promenades mathématiques, histoire fondements, applications*, F. Laroche, Éditions Ellipses
ISBN-10: 2729814175 (448 pages) - Imprimé en 2003
- [075]    *Explosion démographique*, A. Jacquard, Éditions Dominos
ISBN-10: 208035163 (127 pages) - Imprimé en 2006
- [076]   *IFS et L-Systèmes*, V. Rezzonico, C. Hebeisen sous la responsabilité du Prof. Buser, printemps 2000-2001, EPFL.
ISBN-10: -
- [077]    *Six Sigma Statistics*, Issa Bass, Éditions McGraw Hill
ISBN-10: 0071496467 (374 pages) - Imprimé en 2007
- [078]    *Finance des marchés, Techniques quantitatives et applications pratiques*
S. Bossu, P. Henrotte, Éditions Dunod
ISBN-13: 9782100518128 (274 pages) - Imprimé en 2008
- [079]   *Pratiquer les plans d'expérience*, J. Goupy, Éditions Dunod
ISBN-10: 2100042173 (551 pages) - Imprimé en 2005
- [080]    *L'analyse au fil de l'histoire*, E. Hairer, G. Wanner, Éditions Scopus
ISBN-10: 3540674632 (371 pages) - Imprimé en 2001
- [081]   *Physique des semiconducteurs et des composants électriques*, H. Mathieu, Éditions Dunod
ISBN-10: 2100486330 (826 pages) - Imprimé en 2004
- [082]    *Six Sigma et Minitab*, ARIAQ, QSB Consulting, Éditions Quentin Brook
ISBN-13: 9780954681340 (250 pages) - Imprimé en 2008
- [083]  *Recherche opérationnelle et management des entreprise*, D. Thiel, Éditions Economica
ISBN-10: 2717819304 (190 pages) - Imprimé en 1990
- [084]    *Contrôle de la qualité (MSP, Analyse des performances, Contrôle de réception)*, L. Jaupi, Éditions Dunod
ISBN-10: 2100042645 (282 pages) - Imprimé en 2002
- [085]   *Histoire de la science*, M. Daumas, Éditions les Pléiades
ASIN: B0046Y2G3A (1905 pages) - Imprimé en 1957
- [086]    *Analyse, Concepts et Contextes Volume 1. Fonctions d'une variable*
J. Stewart, Éditions DeBoeck - 3ème édition
ISBN-10: 2804150305 (750 pages) - Imprimé en 2006
- [087]    *Guide de localisation des astres*, C. Gentili, Éditions EDP Sciences, 1ère édition
ISBN-13: 9782759800599 (286 pages) - Imprimé en 2008

- [088]    *Chimie quantique*, T. Tung Nguyen-Dang, Département de chimie (FSG), Université Laval, Québec, A-2005
PDF (157 pages) - 2005
- [089] *Pratique industrielle des plans d'expérience*, Éditions AFNOR, J. Alexis + P. Alexis,
ISBN-10: 2124650386 (276 pages) - Imprimé en 1999
- [090]  *A simple approach to clustering in Microsoft Excel 2007*, International Journal of Computer Applications (0975-887), Volume 11 - N°7, December 2010, H. Aravind + C. Rajgopal + K.P. Soman (7 pages)
- [091]     *Quantum: Einstein, Bohr, and the Great Debate about the Nature of Reality*, Éditions W. W. Norton & Company, M. Kumar
ISBN-10: 0393339882 (448 pages) - Imprimé en 2011
- [092]     *Le grand roman de la physique quantique*, Éditions JC Lattès, M. Kumar
ISBN-10: 2709624656 (530 pages) - Imprimé en 2011
- [093]     *Statics and dynamics* (12ème Édition), Éditions Pearson Education, R. C. Hibbeler
ISBN-10: 0138149291 (1403 pages) - Imprimé en 2010
- [094]     *Physics* (9ème Édition), Éditions John Wiley & Sons, Inc, John D. Cutnell + Kenneth W. Johnson
ISBN-13: 9780470879528 (1086 pages) - Imprimé en 2012
- [095]     *Quantum Mechanics: Classical Results, Modern Systems, and Visualized Examples* (2ème Édition), Éditions Oxford University Press, Richard W. Robinett
ISBN-10: 0198530978 (722 pages)- Imprimé en 2006
- [096]     *Quantum Mechanics with Basic Field Theory*, Éditions Cambridge University Press, Bipin R. Desai
ISBN-13: 978051691348 (860 pages)- Imprimé en 2009



D.V.D.

Nous conseillons pour les enfants et les adultes (professeurs) les documentaires et films suivants qui relativement aux mathématiques et à la physique nous semblent intéressants et parfois sympathiques.

- *Tous sur orbite!* (2004/2 DVD's), Nicolas Gessner, Montparnasse
- *Will Hunting* (1997), avec Robin Williams et Matt Damon un film de Gus Van Sant, TF1 Vidéo
- *Il était une fois les découvreurs* (1994/6 DVD's), Albert Barillé, Sony Music
- *Un homme d'exception* (2001), avec Russell Crow un film de Ron Howard, Universal Pictures
- *L'amour en équations* (1994), avec Tim Robbins et Meg Ryan, Paramount Pictures

LIENS INTERNET

Cette annexe contient des liens répartis dans 7 catégories tous ayant trait aux sciences et que je trouve intéressants. Je tiens à préciser qu'en aucun cas j'ai été rétribué sous quelque forme que ce soit pour l'ajout de liens dans la liste ci-dessous! S'il y avait par ailleurs, trois liens à mettre en évidence parmi tous ce serait [Google](#), [Wikipedia](#) et [YouTube](#) dont nous sommes redevables de nombreux emprunts!

Légendes:

- ★ Site de qualité tant du point de vue contenu que contenant
- Σ Contenu avec développements et démonstrations
- 📄 Site avec forum
- 📁 Softwares, sharewares, freewares scientifiques à télécharger
- 📖 Ouvrages, publications, magazines, journaux scientifiques (à visualiser ou télécharger)
- 😊 Site sympa
- ❤ Coup de coeur personnel
- 👁 À voir absolument

1. SCIENCES EXACTES

Certainement la meilleure encyclopédie (condensée) sur la mathématique et la physique (il s'agit également du site officiel du logiciel Mathematica). Le visiteur pourra trouver de nombreux articles plus ou moins développés dans de nombreux domaines.

<http://scienceworld.wolfram.com>



ChronoMath, petite chronologie des mathématiques, est un document pédagogique en perpétuel renouvellement que Serge Mehl fit connaître dès 1988 en tant que conseiller pédagogique de mathématiques en Afrique francophone où il fut coopérant pendant de nombreuses années. Plus de 450 mathématiciens (et leurs travaux) sont passés en revue...!!! À voir!

<http://www.chronomath.com>



Ce site propose une actualité des mathématiques, une encyclopédie avec parties dictionnaire, des biographies et formulaires, ainsi que différents dossiers sur des sujets mathématiques divers et un forum. Ce site est particulièrement impressionnant relativement à la quantité d'informations proposée en téléchargement.

<http://www.bibmath.net>



Un bon site traitant de quelques sujets pertinents en physique (synophysique). Le design pourrait être revu en ce qui concerne la navigation utilisant des cadres... mais il faut tout de même privilégier la qualité du contenu et celle-ci prédomine largement.

<http://www.sciences.univ-nantes.fr/physique/perso/blanquet/frame3.htm>



Nombreux contenus PDFs sur l'algèbre, l'analyse et la géométrie.

<http://c.caignaert.free.fr>



2'000 pages de documents de maths (cours, exercices corrigés, devoirs) au lycee (première, terminale, BAC), classes prepas, études supérieures et l'oral de l'agregation interne. Forum en LaTeX.

<http://perso.wanadoo.fr/gilles.costantini/>



arXiv est une archive de prépublications électroniques d'articles scientifiques. Le laps de temps s'écoulant entre le moment où un chercheur termine un projet et le moment où son travail est publié dans un journal peut être de l'ordre d'un an. À l'échelle du temps de la recherche c'est une durée longue. La mise en place de l'arXiv a donc été un moyen de pallier à ce problème temporel et de coût.

<http://arxiv.org>



L'archive ouverte pluridisciplinaire HAL est destinée au dépôt et à la diffusion gratuite d'articles scientifiques de niveau recherche, publiés ou non, et de thèses, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés. On y trouve même des livres scientifiques publiés récemment par des maisons d'éditions françaises prestigieuses.

<http://hal.archives-ouvertes.fr/>

2. ÉDITIONS - MAGAZINES



On peut considérer le site proposé ici comme l'équivalent du précédent, mais pour les ingénieurs francophones. Il est cependant qualitativement meilleur et on y trouve une quantité de dossiers toute aussi grande mais plus homogène. On regrettera juste peut-être l'accès à certains éléments qui n'est pas toujours aisé du premier coup.

<http://www.techniques-ingenieur.fr>



Absolument excellent et à visiter! De nombreux cours complets de l'École polytechnique (France) publiés sont disponibles gratuitement au téléchargement au format PDF (est-ce que cela va durer?). Environ 1000 ressources pédagogiques sont disponibles au total dont la qualité est très variable, mais dont la pertinence et la rareté des sujets est toujours égale.

<http://catalogue.polytechnique.fr>



Le site en soi n'est pas excellent (ce qui est fort dommage) mais le magazine pour lequel il vous propose de vous abonner (contre paiement) vaut largement le détour pour les passionnées de mathématiques et de leur actualité.

<http://tangente.poleditions.com/>



Le site en soi n'est pas excellent non plus (ce qui est fort dommage aussi) mais certains des ouvrages proposés sont simplement historiques!!!

<http://urss.ru>



Excellent site avec une ressource considérable de documentation électronique traitant des mathématiques uniquement. Le site contient des liens vers les pages des auteurs des documents (sinon quoi il faudrait un serveur considérable...).

<http://mathslinker.chez-alice.fr/>



Ce site contient une bibliothèque en ligne de publications de grands mathématiciens du 20ème et 19ème siècle. À voir absolument!

<http://matwbn.icm.edu.pl/wyzukiwarka.php>



Les presses polytechniques et universitaires romandes constituent une des références mondiales dans l'édition d'ouvrages francophones sur la physique théorique et ses applications. Le choix est cependant assez voir trop pauvre.

<http://ppur.epfl.ch>



Reprints d'oeuvres fondamentales concernant les mathématiques, la physique, l'histoire et la philosophie des Sciences.

<http://www.gabay.com>



Excellent site contenant de nombreux ouvrages francophones et anglophones de qualité. Propose au fait des ouvrages de plusieurs maisons d'édition dont Dunod, Springer, etc. Il vaut la peine d'aller jeter un coup d'oeil dans les sections "Mathématique" et "Physique" y'a du bon...

<http://www.eyrolles.com>



Éditions Dunod - ouvrages scientifiques contemporains (niveau 1-3 cycle et +). Les ouvrages de cette maison d'édition sont pour la plupart excellents. Personnellement ils sont mes préférés, car souvent les développements y sont très détaillés.

<http://www.dunod.com>



Éditions Springer - ouvrages scientifiques (niveau postdoc). Le site est assez difficile à utiliser, car les choix sont disposés un peu n'importe comment mais sinon les ouvrages proposés sont très techniques et de haut niveau... très haut niveau...

<http://www.springer.de>



Le programme NUMDAM, piloté par la Cellule MathDoc (UMS 5638 CNRS - UJF) pour le compte du CNRS, propose la numérisation rétrospective des fonds mathématiques publiés en France.

<http://www.numdam.org>



Éditions Wrox - ouvrages informatiques de développement (niveau expert). D'après ce que je sais (et j'en pense de même), cette maison d'édition est la référence mondiale dans le domaine des ouvrages traitant des langages de programmation.

<http://www.wrox.com>



Éditions Cambridge University Press - ouvrages scientifiques (niveaux postdoc). Un peu l'équivalent des éditions Springer mais le site est mieux fait.

<http://uk.cambridge.org>



Thèses en ligne (mis à disposition par le CNRS). Bof... on peut rarement tirer quelque chose de pertinent d'une thèse, car il manque presque toujours les détails des calculs mais bon... au moins cela permet de contrôler si vous ne refaites pas les choses à double.

<http://tel.ccsd.cnrs.fr>



Bibliothèque Numérique de France (gratuité d'ouvrages scientifiques dont les droits d'auteur sont tombés après 75 années). Le système de téléchargement est peu convivial, mais on peut tomber sur des ouvrages assez pertinents mais vieux... Les PDF font souvent plus de 10 MB donc attention à ceux qui ont du petit débit.

<http://gallica.bnf.fr>



Physical Review Letters search machine (articles scientifiques). Il faut s'inscrire et payer, mais on a rien sans rien... Il faut admettre que le niveau de détail des articles proposés n'est pas très élevé mais là n'est pas l'objectif pour les spécialistes.

<http://prola.aps.org/search>



Très bonne maison d'édition proposant des ouvrages très spécialisés dans le domaine de l'ingénierie électronique, électrotechnique, civile, informatique, etc. On appréciera particulièrement la possibilité de s'abonner gratuitement à leur catalogue. Voir la partie "Hermès Sciences" et "Tech & Doc2 plus particulièrement...

<http://www.lavoisier.fr> (www.editions-hermes.fr + www.tec-et-doc.com)



Maison d'édition américaine proposant des ouvrages de haut niveau pour les sciences physiques et mathématiques (voir les sections correspondantes dans leur site).

<http://www.worldscibooks.com>

3. ASSOCIATIONS



myScience.ch donne une vue d'ensemble des sciences, de la recherche, des universités, des entreprises et d'autres centres de recherche en Suisse. Le site fournit des informations pratiques sur l'emploi, le financement et la vie en Suisse ainsi que des nouvelles scientifiques aux chercheurs, scientifiques, universitaires et à toute personne intéressée par les sciences. Il a une proportion plus grande d'articles en allemand... même dans la version francophone...

<http://www.myscience.ch>



Union des rationalistes- L'Union Rationaliste a été fondée en 1930 pour répandre dans le grand public l'esprit et les méthodes de la science. Il faudrait créer une telle union dans chaque pays...

<http://www.union-rationaliste.org>



American Mathematical Society - Excellent site contenant des centaines de ressources et des ouvrages de haute qualité et de haut niveau à télécharger gratuitement.

<http://www.ams.org>



American Physical Society

<http://www.aps.org>



European Physical Society (possibilité de s'abonner au magazine europysicsnews et plein d'autres choses)

<http://www.eps.org>

4. EMPLOIS



Voici une liste de sites d'offres d'emplois destinée particulièrement aux physiciens et mathématiciens suisses... (cependant un lien est destiné uniquement au territoire U.S.!). Donc souvent les descriptifs sont en allemand (...) et les offres relatives au domaine bancaire (qui représente ~10-15% de l'emploi en Suisse d'après ce que je sais...). Certaines offres sont reliées à des sites plus généraux comme le fameux jobup.ch bien connu en Suisse.

www.telejob.ch

www.jobs.myscience.ch

www.sciencejobs.com

www.math-jobs.ch

www.analyticrecruiting.com

5. TÉLÉVISION-RADIO



Site internet de la chaîne web Télésavoirs.com - visionnement d'interviews et reportages ayant trait aux sciences en différé.

<http://www.telesavoirs.com>



Canal-U: la web TV des universités, de nombreuses ressources (vidéos enseignement supérieur, conférences en ligne, e-learning).

<http://www.canal-u.education.fr>



Cité-Sciences: visionnement de vidéo-conférences, ou écoute d'enregistrements audio sur des sujets de physique et astronomie (+ d'autres) vulgarisés. (requière RealOne Player)

<http://www.cite-sciences.fr>



Sur Canal Académie, les académiciens, spécialistes en physique, mathématique, sciences humaines, philosophie, sociologie, droit et jurisprudence, économie, politique et finances, histoire, géographie et démographie et sciences politiques vous font partager leurs réflexions sur l'actualité et les évolutions de la société.

<http://www.canalacademie.com>

6. DIVERS SCIENCES



Site comportant de nombreuses fiches de cours et surtout quantités d'exercices résolus intéressants. Malheureusement, le site est devenu payant avec accès pendant une période limitée... le prix étant peu élevé il peut être quand même pertinent de déboursier la somme ad hoc.

<http://www.web-sciences.com>



Il s'agit ici d'un portail e-learning pluridisciplinaire en accès libre offrant une sélection de plus de 1200 liens vers de cours en ligne en langue française de grande universités ou écoles d'ingénieurs. Il y a du bon comme du moins bon mais on tombe quand même souvent sur des pépites qui valent le détour.

<http://eunomie.u-bourgogne.fr/elearning/>



Futura-Sciences est un site à vocation de vulgarisation sur les sciences pures et exactes avec des informations quotidiennes sur les sciences et technologies, de nombreux dossiers dans toutes les thématiques et des forums de sciences de débats et de discussions ... (le forum de physique y est particulièrement bien fréquenté.

<http://www.futura-sciences.com>



Astrosurf est un portail de liens pour les astronomes amateurs francophones. Il y est aussi proposé des forums d'astronomie, des petites annonces, l'hébergement gratuit de sites d'astronomie, des éphémérides et tout sur les clubs et associations d'astronomie francophones. On y trouve en particulier le site fameux de Thierry Lombry ([ici](#)).

<http://www.astrosurf.com>



Le CEA est simplement le Commissariat Français à l'Énergie Atomique. Le site propose des dossiers et news intéressantes sur la physique de pointe. On regrettera cependant le manque d'un flux RSS.

<http://www.cea.fr>



Voici un site qui, si j'étais encore enfant, aurait fait mon bonheur... et le malheur de mes parentes. Ce n'est pas tant contenu qui est intéressant mais surtout la boutique en ligne (eStore) qui propose une bonne centaine de gadgets et jeux ludo-éducatifs pour les passionnées de sciences. Attention au porte-monnaie quand même!

<http://www.physlink.com>



Un petit bonheur ce site sur la géophysique. Il est très bien illustré et les explications sont claires et concises. On regrettera peut-être l'absence d'une compilation des articles dans un PDF unique, mais à ce jour je n'ai pas vu mieux sur le sujet au niveau de la qualité du contenu, de l'accès et de la présentation des sujets.

<http://www.geopedia.fr>

7. LOGICIELS



Site très intéressant avec beaucoup d'outils à télécharger pour les physiciens, mathématiciens et ingénieurs. À voir absolument!!!!

<http://www.mathtools.net>



Spécialiste du logiciel scientifique, Integral Software est distributeur officiel de Maple, Origin, Mathcad et de nombreux logiciels pour l'édition et le calcul scientifiques dans de nombreux domaines.

<http://www.intesoft.com>



Minitab est le logiciel de référence pour les ingénieurs travaillant dans l'industrie et appliquant la maîtrise statistique des procédés (voir chapitre de Génie Industriel) dans le cadre de leur travail ou faisant tout autre étude statistique dans le cadre de la R&D.

<http://www.minitab.com>



Isograph est une excellente suite de logiciels pour les ingénieurs travaillant dans l'industrie et appliquant les techniques de la maintenance préventive (fiabilité, AMDEC) et d'assistance à la décision (voir chapitre de Théorie Des Jeux Et De La Décision).

<http://www.isograph-software.com>



JMP est pour moi le meilleur logiciel connu pour les plans d'expérience à ce jour. Son approche est pédagogique, structurée et il existe de très bonnes documentations avec les démonstrations mathématiques de la démarche utilisée par le logiciel.

<http://www.jmp.com>



Site internet officiel du logiciel de calcul formel Maple (propose beaucoup de documentation et de plugins à télécharger).

<http://www.mapleapps.com>



TANAGRA est un puissant logiciel gratuit de datamining destiné à l'enseignement et à la recherche. Il implémente une série de méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'analyse de données, de l'apprentissage automatique et des bases de données.

<http://eric.univ-lyon2.fr/~ricco/tanagra/>



Site internet officiel du logiciel de calcul et simulation numérique Matlab.

<http://www.mathworks.com>



Site internet officiel de la société Statistica. Une référence mondiale dans le domaine!

<http://www.statsoft.com>



COMSOL Multiphysics (anciennement FEMLAB) est un excellent environnement interactif pour la modélisation d'applications industrielles et scientifiques basées sur les équations aux dérivées partielles (EDP) utilisant les méthodes par éléments finis. À acheter pour les ingénieurs et chercheurs (pour ceux qui ont les moyens d'acheter la licence bien sûr.)!

<http://www.comsol.fr>



Crystal Ball est un add-in de MS Excel pour la simulation probabiliste (de Monte Carlo et Latin Hypercube). Avec plus de 100'000 licences à ce jour, c'est à la fois l'add-in d'Excel le plus vendu au monde et l'outil de simulation le plus vendu au monde pour ce logiciel. Son succès est dû à la combinaison de sa grande convivialité et de la richesse de ses fonctionnalités dans le domaine de la recherche opérationnelle, de l'optimisation et de la simulation.

<http://www.decisioneering.com>



Palisade @Risk est une suite d'add-in pour MS Excel et MS Project pour la simulation probabiliste (de Monte Carlo et Latin Hypercube). Son intégration conjointe avec MS Project et MS Excel ainsi que ses modules de calculs basés sur les algorithmes génétiques et réseaux de neurones et son module de théorie de la décision en fait un outil très convoité pas les hauts dirigeants et ingénieurs des grandes entreprises dans le domaine de la finance, de la qualité et de la production.

<http://www.palisade-europe.com>



Le logiciel Decision Analysis par TreeAge (DATA) permet à l'utilisateur de construire, d'analyser et de distribuer des arbres d'analyse, des modèles de Markov et des diagrammes d'influence. Les modèles DATA intègrent visuellement les aspects quantitatifs et qualitatifs reliés aux décisions d'entreprise afin de fournir un outil d'assistance dans les projets lors d'analyse de risques complexes.

<http://www.treeage.com>



Site internet officiel du logiciel MathType (logiciel utilisé pour le présent site).

<http://www.dessci.com>



Site internet officiel du logiciel de calcul formel et de simulation numérique Mathematica.

<http://www.wolfram.com>



Scilab (contraction de Scientific Laboratory) est un logiciel libre, développé à l'INRIA Rocquencourt. C'est un environnement de calcul numérique qui permet d'effectuer rapidement toutes les résolutions et représentations graphiques couramment rencontrées en mathématiques appliquées.

<http://www.scilab.org>



Site officiel du fameux logiciel de géométrie Cabri (référence dans le domaine).

<http://www.cabri.com>



Site d'un prof de math ayant développé un très bon petit logiciel gratuit pour faire de nombreux plots pertinents.

<http://perso.wanadoo.fr/patrice.rabiller/SineQuaNon>



Excellent logiciel de conception et de modélisation (2D, 3D) de molécules avec accès à certaines propriétés physiques.

<http://www.acdlabs.com>

CITATIONS

Il n'y a pas de vérité dans le monde du droit, en morale, en religion, en art, en philosophie, en économie [...]. Il n'y a que des préférences, des goûts et des aversions. Il n'y a ni démonstration, ni vérification mais uniquement la cohérence des argumentations. C'est donc perdre son temps que de prétendre fonder ou disqualifier en vérité une opinion. Epistémon

[...] pour la masse c'est le plaisir qui semble être le bien, tandis que pour les plus subtils, c'est la compréhension. Socrate

*La méthode de l'hypothèse en Science théorique suppose que contrairement aux sciences expérimentales, que les soi-disant données sensorielles sont totalement trompeuses. Rien de ce que l'on peut voir, entendre, sentir, toucher goûter n'est ce qu'il a l'air d'être. Formuler une telle idée n'est cependant que la première étape du travail de la méthode de l'hypothèse. Il faut avoir une croyance passionnée - dans un sens proche de la véritable signification du mot **Foi** - en la réalité d'une simple idée de ce genre. Ce n'est que si elle est guidée par une telle passion, par un amour de l'idée, qu'une personne peut trouver la motivation nécessaire pour la pousser davantage. C. F. Gauss*

Plus grande est la fierté d'un homme moins grande est sa valeur. F. Nietzsche

Le but de la dispute ou de la discussion ne doit pas être la victoire mais l'amélioration. J. Joubert

Celui qui reconnaît consciemment ses limites est le plus proche de la perfection. J. W. von Goethe

Ce n'est pas dans la science qu'est le bonheur, mais dans l'acquisition de la science. E. A. Poe

On ne pactise pas avec les difficultés, ou l'on est vaincu par elles ou on les vainc. J. Rivière

Qu'est-ce que l'originalité ? C'est voir quelque chose qui n'a pas encore de nom, qui ne peut encore être nommé, bien que cela soit sous les yeux de tous. Tels sont les hommes habituellement qu'il leur faut d'abord un nom pour qu'une chose leur soit visible. F. Nietzsche

Les idées révolutionnaires sont souvent simples. Il suffit d'y penser et ça c'est moins simple. Inconnu

Examinez la validité de mes enseignements, comme vous examineriez la pureté d'un pépite d'or, en la frottant contre une pierre, en la martelant ou en la faisant fondre. N'acceptez pas simplement ce que j'enseigne par simple respect pour moi. Bouddha

Tout ce qu'on accumule à la fin se disperse, et ce qu'on construit finit par s'effondrer, ce qui est assemblé pour finir se sépare, et ce qui est vivant disparaît dans la mort. Bouddha

La science peut fonctionner sans la spiritualité, la spiritualité peut exister sans la science. Mais l'homme, pour être complet, a besoin des deux... M. Ricard

Ferme les yeux et tu verras J. Joubert

Le génie crée, l'esprit arrange Duc de Levis

Seules deux choses sont infinies : l'Univers et la bêtise humaine. Mais je ne suis pas sûr pour l'Univers A. Einstein

Science sans conscience n'est que ruine de l'âme... (Rabelais). Conscience sans science ne va pas très loin... M. Weber

La vie n'est bonne qu'à deux choses: découvrir les mathématiques et enseigner les mathématiques S. D. Poisson

La personnalité créatrice doit penser et juger par elle-même, car le progrès moral de la société dépend exclusivement de son indépendance. Sinon la société est inexorablement vouée à l'échec, comme l'être humain privé de la possibilité de communiquer. A. Einstein

Mais c'est la personne humaine, libre, créatrice et sensible qui façonne le beau et le sublime, alors que les masses restent entraînées dans une ronde infernale d'imbécillité et d'abrutissement. A. Einstein

La politique est éphémère mais une équation est éternelle. A. Einstein

«Connais-toi toi-même», voilà toute la science. C'est seulement quand la connaissance des choses sera achevée que l'homme se connaîtra lui-même. Car les choses ne sont que les limites de l'homme. F. Nietzsche

... le péché fondamental des religions: faire des adeptes qui ne posent plus de questions. L'attitude scientifique est exactement à l'opposé. A. Jacquard

Je me suis bien souvent trouvé en présence de personnes... qui manifestaient leur incrédulité devant le manque de culture des scientifiques. Une ou deux fois, il m'est arrivé, me sentant attaqué, de leur demander si elles pouvaient m'énoncer le deuxième principe de la thermodynamique. Leur réponse fut ... négative. Et pourtant, ce que je leur demandais était l'équivalent pour la science de: avez-vous déjà lu une ligne de Shakespeare? P. Heinz

Ceux qui ont fréquenté l'université savent qu'on y raconte beaucoup de sottises; et le dire tout haut suscitera des protestations qui ne viendront que confirmer l'affirmation. B. Arcand

Qui a entrevu l'univers, qui a entrevu les ardents desseins de l'univers ne peut plus penser à un homme, à ses banales félicités ou à ses bonheurs médiocres, même si c'est lui cet homme. J. L. Borges

La musique nous donne accès au coeur du monde. Quand j'écoute Mozart, Schubert ou Wagner, je sens monter en moi un irrésistible sentiment d'exaltation et de reconnaissance pour l'univers qui a engendré la vie et la musique. H. Reeves

Il faudrait être plus savant que tous les savants pour savoir que tous les savants sont des imbéciles. Inconnu

Le grand oeuvre s'accomplira par la science, non par la démocratie. E. Renan

Il faut n'appeler Science que l'ensemble des recettes qui réussissent toujours. Tout le reste est littérature. P. Valéry

C'est manquer son but que d'amuser et de plaire quand on peut instruire et toucher. D. Diderot

Dieu a créé les nombres entiers, le reste est l'oeuvre de l'homme... L. Kronecker

Le concret, c'est de l'abstrait auquel on est habitué. Inconnu

Le secret du succès est de faire de sa vocation ses vacances. M. Twain

Dis-moi quelles sont tes fréquentations et je te dirai qui tu es. Inconnu

À quoi servent les mathématiques ? : "C'est pour l'honneur de l'esprit humain". C. G. Jacobi

Les religions sont un poison, la tolérance est leur cheval de bataille, l'humain leur jouet et la bêtise leur meilleure alliée mais les sciences leur salut. Inconnu

Placez votre main sur un poêle une minute et ça vous semble durer une heure. Asseyez-vous auprès d'une jolie fille une heure et ça vous semble durer une minute. C'est ça la relativité. A. Einstein

Le plus solide avantage des sciences, mais un avantage dont il n'est pas facile de faire sentir le prix, c'est qu'elles purifient et raniment un organe de l'âme aveuglé et comme éteint par les autres occupations de la vie, organe dont la conservation est mille fois plus précieuse que celle des yeux du corps, puisque c'est par celui-là seul qu'on aperçoit la vérité. Platon

Celui qui chemine sur la voie de la science, Dieu chemine avec lui sur la voie du Paradis. Coran

Certains sont menés par l'amour et la richesse, d'autres guidés aveuglement par la soif insensée de puissance et de domination, mais l'homme le plus noble se consacre à la découverte du sens et du but de la vie. Il cherche à découvrir les secrets de la nature. C'est celui que j'appelle "un philosophe" car, bien qu'aucun homme ne soit sage à tous égards, il peut aimer la sagesse comme clef des secrets de la nature. Pythagore

Si je devais choisir un pays où mourir, je choisirai la Suisse. Car là-bas, tout se passe 20 ans plus tard. A. Einstein

L'imaginaire couvre tout l'univers de l'inconscient. H. Baron

La science n'est rien de plus qu'un raffinement de la pensée ordinaire. A. Einstein

Moi. Je serais Dieu, je démissionnerais... pour que plus personne ne se batte en mon nom. Le Chat

La violence est le dernier refuge de l'incompétence I. Asimov

La plus belle chose que nous puissions éprouver, c'est le mystère des choses. A. Einstein

Si vous ne pouvez expliquer un concept à un enfant de six ans, c'est que vous ne le comprenez pas complètement. A. Einstein

L'école devrait toujours avoir pour but de donner à ses élèves une personnalité harmonieuse, et non de les former en spécialiste. A. Einstein

C'est la théorie qui décide de ce que nous pouvons observer. A. Einstein

L'enseignement devrait être ainsi : celui qui le reçoit le recueille comme un don inestimable mais jamais comme une contrainte pénible. A. Einstein

L'impossible recule devant celui qui avance Inconnu

... sinon que la nature seule, et non pas le tumulte des foules, peut nous initier à la vérité; que pour scruter les mystères de la science il est bon de se retirer dans la solitude et de développer son intelligence par la réflexion ? E. Reclus

La différence entre la pratique et la théorie ? C'est qu'en théorie il n'y a pas différences... Inconnu

Les personnes de faible intelligence parlent des gens, les personnes de moyenne intelligence parlent d'actualité et les personnes de haute intelligence parlent d'idées Inconnu

C'est donc par l'étude des mathématiques, et seulement par elle, que l'on peut se faire une idée juste et approfondie de ce que c'est qu'une science. A. Comte

Les masses médias rendent les masses médiocres ! Aguigui Mouna

C'est quand la science et la technologie atteignent leurs limites que l'homme peut construire d'aussi belles choses que la nature peut le faire Inconnu

La réalisation d'un monde meilleur passe par la foi en soi. La foi en soi passe par la connaissance de soi. F. Hatem

Un chercheur doit savoir sécher une heure, un jour, ou toute une vie. Il sèche beaucoup plus qu'il ne trouve, il se pose une série de questions, tâtonne, avance pas à pas. C'est très difficile ; puis, à un moment donné, une certaine illumination vient. Elle est souvent très brusque, mais elle est le résultat d'une accumulation énorme de réflexions infructueuses. L. Schwartz

Il n'y a pas de pire orgueil que la fausse modestie Proverbe chinois

La tradition est maître de l'ignorance Inconnu

Dans la vie, il y a 10 types de personnes: ceux qui comprennent le binaire et ceux qui ne le comprennent pas. Inconnu

Le paradoxe est [...] le nom que les imbéciles donnent à la vérité J. Moréas

Aucune investigation humaine ne peut être qualifiée de science véritable si elle ne peut être démontrée mathématiquement L. da Vinci

La seule manière d'apprendre les mathématiques, c'est de faire des mathématiques P. Halmos

Est rigoureuse toute démonstration, qui, chez tout lecteur suffisamment instruit et préparé, suscite un état d'évidence qui entraîne l'adhésion. R. Thom

En mathématiques, les noms sont arbitraires. Libre à chacun d'appeler un opérateur autoadjoint un "éléphant" et une décomposition spectrale une "trompe". On peut alors démontrer un théorème suivant lequel "un éléphant à une trompe". Mais on n'a pas le droit de laisser croire que ce résultat a quelque chose à voir avec de gros animaux gris. G. Sussman

L'imagination est plus importante que la connaissance A. Einstein

Quand j'étais jeune, j'espérais démontrer l'hypothèse de Riemann. Quand je suis devenu un peu plus vieux, j'ai encore eu l'espoir de pouvoir lire et comprendre une démonstration de l'hypothèse de Riemann. Maintenant, je me contenterais bien d'apprendre qu'il en existe une démonstration. S. Weil

Pour être théoricien, il faut à la fois avoir étudié les mathématiques et posséder beaucoup d'intuition. On ne saurait sous-estimer le rôle de l'intuition et de l'imagination dans les sciences. Les étudiants qui réussissent à tous les examens ne sont pas nécessairement les chercheurs les plus créatifs. Lors d'un examen, il faut résoudre un problème spécifique mais, dans le monde de la recherche théorique, le véritable problème consiste à découvrir quel est le problème. Alors seulement peut-on le formuler de manière précise et le résoudre grâce à des techniques mathématiques appropriées. Poser la bonne question requiert beaucoup d'imagination. P. Heinz

Si l'esprit d'un homme s'égare, faites-lui étudier les mathématiques car dans les démonstrations, pour peu qu'il s'écarte, il sera obligé de recommencer. F. Bacon

Je ne peux pas imaginer un Dieu qui récompense et punit l'objet de sa création. Je ne peux pas me figurer un Dieu qui réglerait sa volonté sur l'expérience de la mienne. Je ne veux pas et je ne peux pas concevoir un être qui survivrait à la mort de son corps. Si de pareilles idées se développent en un esprit, je le juge faible, craintif et stupidement égoïste. A. Einstein

Le but de la physique n'est pas de découvrir ce qu'est la nature, mais ce que l'on peut dire sur elle N. Bohr

Ceux qui disent que le soleil tourne autour de la terre ont le mérite de croire ce qu'ils voient. Ceux qui disent que la terre tourne autour du soleil répètent ce qu'ils ont entendu dire parce qu'ils croient que c'est mieux que ce qu'ils voient. Et puis il y a ceux qui disent que cela dépend du référentiel où l'on se place... Inconnu

Au début, toutes les pensées sont vouées à l'amour. Plus tard, tout l'amour est voué à la pensée A. Einstein

L'hypothèse du continu en théorie des ensembles fut d'abord résolue partiellement en 1938 par un des plus grands logiciens du 20ème siècle, Kurt Gödel, qui démontra que l'hypothèse de Cantor n'était pas réfutable, c'est-à-dire qu'on ne pourrait jamais démontrer qu'elle était fausse. Puis en 1963, le mathématicien Paul Cohen boucla la boucle. Il démontra qu'on ne pourrait jamais non plus démontrer qu'elle était vraie !!! Nous ne pouvons conclure à juste raison que Cantor avait perdu la raison à chercher à démontrer un problème qui ne pouvait pas l'être. Sciences.ch

La mathématique est la seule science qui permet à l'être humain de quitter le concret abstrait pour aller à l'abstrait concret. Isoz

Un bon philosophe est au moins à moitié un mathématicien, un bon mathématicien est au moins à moitié un philosophe. Gottlob Frege

Je ne doute pas, bien qu'on doive s'attendre à des changements progressifs en doctrine physique, les doctrines actuelles sont probablement plus proche de la vérité que n'importe quelle théorie rivale existante. La science n'est jamais tout à fait correcte, mais elle est rarement tout à fait fausse et a, en général, plus de chances d'être correcte que les théories non scientifiques. Il est, par conséquent, rationnel de l'accepter à titre provisoire. Bertrand Russel

Devant un livre, nous ne devons pas nous demander ce qu'il dit mais ce qu'il veut dire... Umberto Eco

La science est la seule lumière : plus de science fera une meilleure morale, une meilleure société, un art meilleur. Condorcet

La science n'a pas de patrie. Louis Pasteur

La physique théorique, c'est les mathématiques sans la rigueur, et la physique sans l'expérience.

J.M.Souriau

L'étude des mathématiques en comprimant la sensibilité et l'imagination, rend quelquefois l'explosion des passions terrible Msgnr. Dupanloup

L'étude des mathématiques est comme le Nil, qui commence en modestie et finit en magnificence. Charles Caleb Colton

La science comme l'amour est aveugle, voilà pourquoi elle se plaît à procéder par tâtonnements. Jean O'Neil

Les hautes mathématiques sont la musique de la pensée. Georges Steiner

La science dans ses résultats est plus magique que la magie: c'est une magie à preuves! Jean Marc Adiaffi

Un physicien moderne étudie la physique quantique les lundis, mercredis et vendredis, et médite sur la théorie de la relativité gravitationnelle les mardis, jeudis et samedis. Le dimanche, il prie...pour que quelqu'un trouve la corrélation entre les deux. Norbert Wiener

La rigueur vient toujours à bout de l'obstacle. L. Da Vinci

En essayant continuellement on finit par réussir, donc plus ça rate plus on a de chance que ça marche. Jacques Rouxel

S'il n'y a pas de solution c'est qu'il n'y a pas de problème. Jacques Rouxel

Je suis désolé de dire que la matière que j'aimais le moins était les mathématiques. J'y ai repensé. Je pense que la cause de cela était le fait que les mathématiques ne laissent pas de place à l'argumentation. Si vous faites une erreur, c'est tout ce qu'il y a à en dire. Malcolm X

Comment puis-je accepter cette doctrine qui exalte comme une bible, au-dessus de toute critique, un manuel démodé dont je sais qu'il est non seulement scientifiquement erroné, mais encore inintéressant et inapplicable dans le monde moderne ? J.M. Keynes

Il m'a fallu soixante ans pour faire cette peinture en 6 minutes A. Renoir

Ce qui se conçoit clairement s'énonce de même. Inconnu

L'humanité a connu trois vexations : Première vexation, Nicolas Copernic révèle que la Terre n'est pas située au centre de l'Univers mais qu'elle tourne autour d'un soleil, et que celui-ci est probablement lui-même dans la périphérie d'un système plus vaste. Deuxième vexation, Charles Darwin annonce que l'homme n'est pas une créature au-dessus des autres, mais juste un animal parmi les autres animaux. Troisième vexation, Sigmund Freud déclare que l'homme croit créer de l'art, conquérir des territoires, inventer des sciences, élaborer des systèmes philosophiques ou politiques parce qu'il est motivé par des ambitions supérieures qui le transcendent, alors qu'en fait, il n'est motivé que par son envie de séduire des partenaires sexuels. E. Wells

La science est asymptote de la vérité, elle l'approche sans cesse et ne la touche jamais. Victor Hugo

La physique s'occupe des causes invisibles des choses visibles. Inconnu

Lorsque l'on a éliminé l'impossible, ce qui reste – aussi improbable que ce soit – doit être la vérité.
Sherlock Holmes

Tout regard se transforme en une observation, toute observation en une réflexion, toute réflexion en une appréhension, et ainsi, nous pouvons dire qu'à chaque regard attentif, nous théorisons le monde. J.W. Von Goethe

Le physicien et le mathématicien ne peuvent pas descendre de l'Homme; ils doivent certainement posséder leur propre branche dans l'évolution, une qui mène nul part et partout à la fois. La plus fascinante de toutes! Inconnu

La science est l'art de donner des mauvaises réponses à des problèmes auxquels autrement de pires réponses seraient données! Inconnu

Toute étude faite par l'homme ne peut être élevée au rang de science que si elle peut être mise sous forme mathématique. Leonardo Da Vinci

La science ne sert guère qu'à nous donner une idée de l'étendue de notre ignorance. Félicité de Lamennais

Toute science commence comme philosophie et se termine en art. Will Durant

La science efface l'ignorance de hier et révèle celle de demain. Inconnu

*Beaucoup de professeurs succombent à la tentation de feindre une totale maîtrise de leur propre action, comme si la moindre faille pouvait nuire à leur réputation et à la légitimité de leur message. On peut, au contraire, soutenir qu'un professeur crédible parvient à faire partager la part d'hésitation et d'incertitude qui l'habite, non parce qu'il est incompetent, mais parce que l'action compétente n'est jamais infaillible, ni dénuée de doutes. Pour y parvenir, il faut évidemment avoir et partager la conviction que la compétence n'est pas synonyme de perfection, mais de réflexivité et de régulation!*Philippe Perrenoud

L'étude des mathématiques est comme le Nil, qui commence en modestie et finit en magnificence. Charles Caleb Colton

Celui qui trouve sans chercher est celui qui a longtemps cherché sans trouver. Gaston Bachelard

L'ignorance n'a jamais fait de mal ; on ne s'égare point parce qu'on ne sait pas, mais parce qu'on croit savoir. Rousseau

L'intelligence a été inventée il y a très longtemps par un type vachement malin. La connerie... c'est autre chose: C'est une invention collective. Le Chat

Tous les modèles sont faux, mais certains sont utiles. George Box

Un jour il y eut un immense incendie de forêt. Tous les animaux terrifiés, atterrés, observaient impuissants le désastre. Seul un petit colibri s'activait, allant chercher quelques gouttes avec son bec pour les jeter sur le feu. Après un moment, le Toucan, agacé par cette agitation dérisoire, lui dit : « Colibri ! Tu n'es pas fou?! Tu crois que c'est avec ces gouttes d'eau que tu vas éteindre le feu?! ». Et le colibri lui répondit :«Non, mais je fais ma part». Pierre Rabhi

Un jour j'irai vivre en Théorie, car en Théorie tout se passe bien. Inconnu

Il ne peut y avoir que deux solutions: soit j'ai raison, soit vous avez tort. Inconnu

L'ennui dans ce monde, c'est que les idiots sont sûrs d'eux et les gens sensés pleins de doutes. Bertrand Russell

La connaissance est le seul bien qui s'accroît quand on le partage. Sacha Boudjema

Le succès est un mauvais professeur. Il pousse les gens intelligents à se croire infailibles. Bill Gates

Le scientifique évolue dans l'ombre, sans chercher à acquérir les honneurs. Le sentiment du devoir accompli est sa seule récompense et sa seule fierté. Vincent Isoz

La science commence par le doute... Inconnu

Il faut viser les étoiles pour atteindre la cime des arbres. Bertrand Piccard

Ceux qui le peuvent font de la science, les autres font de la méthodologie. Paul Samuelson

Le mathématicien, en règle générale, est plutôt brut de décoffrage quand il s'agit de dire des choses désagréables. L'élégance, il la garde pour la démonstration réussie. Inconnu

Experience teaches nothing without theory, but theory without experience is mere intellectual play. Immanuel Kant

La science est le miroir de Dieu Descartes

Concevoir une idée est noble, la concrétiser est servile Leonardo Da Vinci

Une part d'autisme est essentielle pour le succès des sciences et de l'art Hans Asperger

Ce que l'homme ne veut pas apprendre par la sagesse, il l'apprendra par la souffrance. Inconnu

Tout mathématicien digne de ce nom a ressenti, même si ce n'est que quelquefois, l'état d'exaltation lucide dans lequel une pensée succède à une autre comme par miracle... Contrairement au plaisir sexuel, ce sentiment peut durer pendant plusieurs heures, voire plusieurs jours. André Weil

L'agressivité est le pouvoir des faibles. Inconnu

Une nouvelle vérité scientifique ne triomphe pas en convainquant ses adversaires et en leur faisant voir la lumière, mais plutôt parce que ses adversaires finissent par mourir, et que grandit une nouvelle génération à qui cette vérité est familière. Max Planck

La vraie logique du monde est celle du calcul des probabilités. James Maxwell

Les individus ne devraient pas être responsables des actes du gouvernement de la nation à laquelle ils se trouvent appartenir. Albert Einstein

Ce n'est ni l'espèce la plus forte, ni la plus intelligente qui survit mais celle qui s'adapte le mieux au changement. Charles Darwin

Le pouvoir des nombres est d'autant plus respecté que l'on n'y comprend rien. Voltaire

À PROPOS DU RESPONSABLE

**Vincent ISOZ**

Année de naissance: 1978

Domicilié à ce jour à Lausanne (Suisse)

Ingénieur HES-SO en Physique Appliquée (2001)

Équivalence Internationale: Bachelor of Science

Consultant en mathématiques appliquées dans le tutorat de hauts potentiels

**Actuellement:**

Consultant indépendant et auteur de plusieurs livres électroniques dans les domaines suivants:

- maîtrise statistique des processus/procédés (méthodes paramétriques et non paramétriques)
- modélisation prévisionnelle/décisionnelle avancée (arbres de décisions, chaînes de Markov)
- recherche opérationnelle (simplexe, algorithmes génétiques, algorithme GRG)
- data mining (réseaux de neurones, ACP, AFC, régressions, scoring, clustering, etc.)
- modélisation du risque en gestion de projets et finance d'entreprise (monte-carlo, bootstrapping)
- gestion de projets (modèles et best practices théoriques EFQM+Six Sigma, MS Office Project)
- ISO 9001:2008, 5807:1985, 10015:1999, 31000+31010:2009, 8258:1991, ISO 10017:2003
- Adobe Photoshop et Illustrator
- 12 applications de la suite Microsoft Office System (Project, Visio, SharePoint, Access, InfoPath, etc.)

Études scolaires:

- Travail de diplôme effectué à l'École Polytechnique Fédérale de Lausanne (Suisse) au Département de Physique Expérimentale en physique de la matière condensée.
- Travail de projet (TPE) sur les sondes Langmuir à l'École d'Ingénieurs de Genève pour les plasmas faiblement et fortement ionisés

Les connaissances générales de l'ingénieur HES en Génie Nucléaire lui permettent de collaborer avec des milieux professionnels différents (physiciens, chimistes, médecins, électroniciens, mécaniciens, informaticiens, financiers, etc.) pour fournir l'assistance théorique et technique nécessaire à la réalisation de la partie expérimentale de projets.

Le plan d'étude en physique est composé d'une formation scientifique en mathématiques, physique générale, physique et chimie nucléaire, radioprotection, acoustique, électrotechnique, électronique, microtechnique, analyse numérique et informatique ainsi que de physique des matériaux.

PROFIL PROFESSIONNEL

Actuellement, très régulièrement consultant dans de nombreuses entreprises et administrations nationales et internationales dans les domaines suivants:

- Mathématique appliquée: Statistiques générales, Maîtrise et optimisation statistique des processus et procédés, Modélisation financière, Ingénierie de la Fiabilité, Décisionnel

- Gestion de projets, gestion de la qualité, gestion de stocks, gestion documentaire, gestion de la formation, gestion du risque
- Développement (langages utilisés dans le cadre professionnel): XML, XSL, XSD, ADA, C++, C#, PHP, ASP 3.0, (X)HTML/DHTML, Javascript, VBA, VBScript, SQL, RibbonX, XPath, CSS
- Multimédia: Macromedia Flash, Macromedia Dreamweaver, Adobe Photoshop, Adobe Illustrator, Adobe Acrobat
- Bureautique: Word, Excel, Outlook, Visio, PowerPoint, Publisher, FrontPage, InfoPath, Project, Project Server, SharePoint, Communicator (Lync), Mind Manager, Camtasia
- Mathématiques: MathType, Matlab, Maple, SPSS, R, Minitab, Tanagra et LabView (mais il y a bien longtemps que je n'ai plus utilisé ce dernier)

LANGUES

Français: langue maternelle

Anglais: connaissances orales et écrites (pratiqué dans le cadre professionnel)

Allemand: bonnes connaissances orales et écrites (pratiqué dans le cadre professionnel)

Italien: en cours d'apprentissage...

CONNAISSANCES INFORMATIQUES

Compétences multiples en informatique développées en autodidacte sur plateformes Microsoft, Apple, Unix et Linux et dispense de cours sur les systèmes et logiciels (certaines notions ne sont plus utilisées depuis... pas mal de temps):

MS Windows 3.x/9x/2000/XP (Home, Pro), IIS, Apache, SQL Server, Oracle, MS-DOS (QBasic), MS Word, MS Excel, MS Publisher, MS FrontPage, MS InfoPath, MS OneNote, MS PowerPoint, MS Project (Server), MS Visio, MS Access, MS Outlook, SharePoint, Lotus Notes, AutoCAD, Maple, Matlab, LabView, MM Dreamweaver, Swish, Swift, MM Flash, Adb Photoshop, Adb Illustrator, Adb Acrobat, 3D Studio Max, Installer Vise, Crystal Reports, MiniTab, @Risk Palissade, Mindjet Mind Manager,...

Langage machine, assembleur 680xx, C++, VBA, VB.Net, C#, ASP 3, ASP.Net, ADO .Net, Ada, QBASIC, BASIC, Pascal, HTML, XHTML, DHTML, CSS, PHP, VBScript, JavaScript, XML, XSL, XPath, XSD, RibbonX, SQL, Pascal, MySQL, ActionScript, AJAX, SMIL,...

Notions de fonctionnement des réseaux informatiques (administrateur informatique premier niveau).

LOISIRS

Études et développements en mathématique appliquée, philosophie, informatique (web, 3D, développement, graphisme), économie, décisionnel, modélisation du risque, actualité, guitare électrique,

Sciences.ch

Éléments de mathématiques appliquées

Le but de ce site est de présenter à celui qui aborde l'étude de la mathématique appliquée, ses concepts fondamentaux et de le faire avec certain niveau de rigueur (les démonstrations sont complètes ou du moins poussées au point où nous pouvons raisonnablement les juger telles), de détail et cohérence dans le respect des écritures et ceci avec un maximum de pédagogie. Ce site est aussi né du désir de présenter certaines idées scientifiques et de comprendre modestement comment elles influent sur notre mode de vie, notre manière de penser et de travailler. Effectivement, les questions relatives à la science prennent une importance grandissante dans notre culture contemporaine et représentent un enjeu majeur en termes d'éthique et de citoyenneté.

Ce site ne se veut pas être cependant un roman qui se lirait d'un bout à l'autre. Il se veut être un ouvrage de référence (une fois que toutes les erreurs auront été corrigées et les textes complétés...) permettant, lorsque des questions simples se posent, de trouver des réponses rapidement à l'aide des technologies informatiques et ce gratuitement. Ce site ne peut (et ne veut) également se prétendre à remplacer un enseignement scolaire structuré avec un professeur et la pratique par de nombreux exercices. Il peut être cependant vu comme un formulaire ou un relativement bon complément théorique à la préparation de divers examens.

Le point de vue qui y est adopté est celui de l'ingénieur pragmatique, soucieux d'étudier la mathématique, physique classique, économétrie, analyse numérique, statistique, mécanique relativiste, physique quantique, mathématique sociale, chimie, etc. en allant à l'essentiel et sans perdre de temps dans un formalisme et vocabulaire extravagant et inapplicable dans l'industrie moderne. De ce point de vue les concepts et les méthodes qui y sont présentées sont simplement quelques-uns des outils mathématico-physiques types (le minimum minimorum dans le domaine). Le spécialiste expert n'y trouvera rien de nouveau et l'étudiant qui serait intéressé par une théorie particulière doit savoir que tout sujet est malheureusement beaucoup plus vaste que tout ce qui peut être abordé ici à ce jour... L'ambition n'est pas celle que l'on peut avoir pour des étudiants d'un cursus de mathématiques pour qui l'acquisition d'outils est importante. Cela rend possible un style assez peu formel où l'on souhaite moins donner des preuves complètes que de faire comprendre de façon intuitive les objets présentés, de les voir et de se les approprier.

Nous dédions ce site à tous ceux pour qui chaque réponse est une question.

Vincent Isoz (Consultant Physicien Indépendant - Responsable Sciences.ch)

